Chapter 4
Prediction and Variable Selection When $n >> p$

This chapter considers variable selection when $n >> p$ and prediction intervals that can work if $n > p$ or $p > n$. Prediction regions and prediction intervals applied to a bootstrap sample can result in confidence regions and confidence intervals. The bootstrap confidence regions will be used for inference after variable selection.

4.1 Variable Selection

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large. Consider the 1D regression model where $Y \perp\!\!\!\perp \boldsymbol{x}|SP$ where $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. See Chapters 1 and 10. A *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{4.1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{\beta}_O.$$

Suppose that S is a subset of I and that model (4.1) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ and the sample correlation $\text{corr}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{x}_{I,i}^T\boldsymbol{\beta}_I) = 1.0$ for the population model if $S \subseteq I$. The estimated sufficient predictor (ESP) is $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$, and *a submodel I is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$.

Definition 4.1. The model $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}$ that uses all of the predictors is called the *full model.* A model $Y \perp\!\!\!\perp \boldsymbol{x}_I|\boldsymbol{x}_I^T\boldsymbol{\beta}_I$ that uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel.* The **full model is always a submodel**. The full model has *sufficient predictor* $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$.

Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection. The relaxed lasso or relaxed elastic net estimator fits the regression method, such as a GLM or Cox (1972) proportional hazards regression, to the predictors than had nonzero lasso or elastic net coefficients. See Chapters 5 and 10.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to β_1 is always in the model, and $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, 0)^T$. Then the $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that always contain 1 are $I_1 = \{1\}$, $S = I_2 = \{1, 2\}$, $I_3 = \{1, 3\}$, $I_4 = \{1, 4\}$, $I_5 = \{1, 2, 3\}$, $I_6 = \{1, 2, 4\}$, $I_7 = \{1, 3, 4\}$, and $I_8 = \{1, 2, 3, 4\}$. There are $2^{p-a_S} = 4$ subsets I_2, I_5, I_6, and I_8 such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$ and $\boldsymbol{x}_{I_7} = (x_1, x_3, x_4)^T$.

Underfitting occurs if submodel I does not contain S. Following, for example, Pelawa Watagoda (2019), let $\boldsymbol{X} = [\boldsymbol{X}_I \quad \boldsymbol{X}_O]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. Then $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{X}_O\boldsymbol{\beta}_O$, and $\hat{\boldsymbol{\beta}}_I = (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y}$. Assuming the usual MLR model, $\text{Cov}(\hat{\boldsymbol{\beta}}_I) = \text{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}^T = \sigma^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$. Now $E(\hat{\boldsymbol{\beta}}_I) = E(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T(\boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{X}_O\boldsymbol{\beta}_O) =$

$$\boldsymbol{\beta}_I + (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{X}_O\boldsymbol{\beta}_O = \boldsymbol{\beta}_I + \boldsymbol{A}\boldsymbol{X}_O\boldsymbol{\beta}_O.$$

If $S \subseteq I$, then $\boldsymbol{\beta}_O = \mathbf{0}$, but if underfitting occurs then the bias vector $\boldsymbol{A}\boldsymbol{X}_O\boldsymbol{\beta}_O$ can be large.

4.1.1 OLS Variable Selection

Simpler models are easier to explain and use than more complicated models, and there are several other important reasons to perform variable selection. For example, an OLS MLR model with unnecessary predictors has $\sum_{i=1}^n V(\hat{Y}_i)$ that is too large. If (4.1) holds, $S \subseteq I$, $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{\beta}_I$ is a $j \times 1$ vector with $j > a_S$, then

$$\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Ii}) = \frac{\sigma^2 j}{n} > \frac{\sigma^2 a_S}{n} = \frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Si}). \tag{4.2}$$

In particular, the full model has $j = p$. Hence having unnecessary predictors decreases the precision for prediction. Fitting unnecessary predictors is sometimes called *fitting noise* or *overfitting*. As an extreme case, suppose that the full model contains $p = n$ predictors, including a constant, so that the hat matrix $\boldsymbol{H} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. Then $\hat{Y} = Y$ so that $\mathrm{VAR}(\hat{Y}|\boldsymbol{x}) = \mathrm{VAR}(Y)$. A model I underfits if it does not include all of the predictors in S. A model I does not underfit if $S \subseteq I$.

To see that (4.2) holds, assume that the full model includes all p possible terms so the full model may overfit but does not underfit. Then $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$ and $\mathrm{Cov}(\hat{\boldsymbol{Y}}) = \sigma^2 \boldsymbol{HIH}^T = \sigma^2 \boldsymbol{H}$. Thus

$$\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_i) = \frac{1}{n} tr(\sigma^2 \boldsymbol{H}) = \frac{\sigma^2}{n} tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}) = \frac{\sigma^2 p}{n}$$

where $tr(\boldsymbol{A})$ is the trace operation. Replacing p by j and a_S and replacing $\boldsymbol{H}$ by $\boldsymbol{H}_I$ and $\boldsymbol{H}_S$ implies Equation (4.2). Hence if only a_S parameters are needed and $p >> a_S$, then serious overfitting occurs and increases $\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_i)$.

Two important summaries for submodel I are $R^2(I)$, the proportion of the variability of Y explained by the nontrivial predictors in the model, and $MSE(I) = \hat{\sigma}_I^2$, the estimated error variance. See Definitions 1.17 and 1.18. Suppose that model I contains k predictors, including a constant. Since adding predictors does not decrease R^2, the adjusted $R_A^2(I)$ is often used, where

$$R_A^2(I) = 1 - (1 - R^2(I))\frac{n}{n-k} = 1 - MSE(I)\frac{n}{SST}.$$

See Seber and Lee (2003, pp. 400-401). Hence the model with the maximum $R_A^2(I)$ is also the model with the minimum $MSE(I)$.

For multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p - k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}\left[\frac{SSE(I)}{SSE} - 1\right]$$

where SSE is the error sum of squares from the full model, and SSE(I) is the error sum of squares from the candidate submodel. An extremely important criterion for variable selection is the C_p criterion.

Definition 4.2.

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model.

Note that when H_0 is true, $(p-k)(F_I-1)+k \xrightarrow{D} \chi^2_{p-k} + 2k - p$ for a large class of iid error distributions. Minimizing $C_p(I)$ is equivalent to minimizing $MSE\ [C_p(I)] = SSE(I) + (2k-n)MSE = \boldsymbol{r}^T(I)\boldsymbol{r}(I) + (2k-n)MSE$. The following theorem helps explain why C_p is a useful criterion and suggests that for subsets I with k terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting. Olive and Hawkins (2005) show that this interpretation of C_p can be generalized to 1D regression models with a linear predictor $\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$, such as generalized linear models. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of Y on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ and $\hat{Y}_{I,i} = \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ where $i = 1, ..., n$.

Theorem 4.1. Suppose that a numerical variable selection method suggests several submodels with k predictors, including a constant, where $2 \leq k \leq p$.

a) The model I that minimizes $C_p(I)$ maximizes $\mathrm{corr}(r, r_I)$.

b) $C_p(I) \leq 2k$ implies that $\mathrm{corr(r, r_I)} \geq \sqrt{1 - \frac{\mathrm{p}}{\mathrm{n}}}$.

c) As $\mathrm{corr}(r, r_I) \to 1$,

$$\mathrm{corr}(\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \boldsymbol{x}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \mathrm{corr(ESP, ESP(I))} = \mathrm{corr}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Proof. These results are a corollary of Theorem 4.2 below. □

Remark 4.1. Consider the model I_i that deletes the predictor x_i. Then the model has $k = p - 1$ predictors including the constant, and the test statistic is t_i where

$$t_i^2 = F_{I_i}.$$

Using Definition 4.2 and $C_p(I_{full}) = p$, it can be shown that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

Using the screen $C_p(I) \leq \min(2k, p)$ suggests that the predictor x_i should not be deleted if

$$|t_i| > \sqrt{2} \approx 1.414.$$

If $|t_i| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases. The literature suggests using the $C_p(I) \leq k$ screen, but this screen eliminates too many potentially useful submodels.

More generally, it can be shown that $C_p(I) \leq 2k$ iff

$$F_I \leq \frac{p}{p-k}.$$

Now k is the number of terms in the model I including a constant while $p-k$ is the number of terms set to 0. As $k \to 0$, the partial F test will reject Ho: $\boldsymbol{\beta}_O = \mathbf{0}$ (i.e. say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p-k$ is very small, then the partial F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p-k$ predictors.

Definition 4.3. The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus r_i. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus Y_i. An *EE plot* is a plot of ESP(I) versus ESP. For MLR, the EE and FF plots are equivalent.

Six graphs will be used to compare the full model and the candidate submodel: the FF plot, RR plot, the response plots from the full and submodel, and the residual plots from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (4.1) holds and that a good estimator (such as OLS) for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

Application 4.1. To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the r_i and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset I is good if the plotted points cluster tightly about the identity line in *both plots*. In particular, the OLS line and the identity line should "nearly coincide" so that it is difficult to tell that the two lines intersect at the origin in the RR plot.

To verify that the six plots are useful for assessing variable selection, the following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively.

A plot can be very useful if the OLS line can be compared to a reference line and if the OLS slope is related to some quantity of interest. Suppose that a plot of w versus z places w on the horizontal axis and z on the vertical axis. Then denote the OLS line by $\hat{z} = a + bw$. The following theorem shows that

the plotted points in the FF, RR, and response plots will cluster about the identity line. Notice that the theorem is a property of OLS and holds even if the data does not follow an MLR model. Let $\text{corr}(x, y)$ denote the correlation between x and y.

Theorem 4.2. Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.
Response Plot: i) If $w = \hat{Y}_I$ and $z = Y$ then the OLS line is the identity line.
ii) If $w = Y$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I)$ and intercept $a = \overline{Y}(1 - R^2(I))$ where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ and $R^2(I)$ is the coefficient of multiple determination from the candidate model.
FF or EE Plot: iii) If $w = \hat{Y}_I$ and $z = \hat{Y}$ then the OLS line is the identity line. Note that $ESP(I) = \hat{Y}_I$ and $ESP = \hat{Y}$.
iv) If $w = \hat{Y}$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.
RR Plot: v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.
vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [\text{corr}(r, r_I)]^2$ and

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

Proof: Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{Y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing z on w is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

$$b = \frac{\sum(w_i - \overline{w})(z_i - \overline{z})}{\sum(w_i - \overline{w})^2} = \frac{SD(z)}{SD(w)}\text{corr}(z, w).$$

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing z on w is equal to one if and only if the OLS slope from regressing w on z is equal to $[\text{corr}(z, w)]^2$.

i) The slope $b = 1$ if $\sum \hat{Y}_{I,i} Y_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

ii) By (*), the slope

$$b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I) = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = SSR(I)/SSTO.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i}\hat{Y}_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{H}_I\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_I\boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T\hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)}[\text{corr}(\hat{Y}, \hat{Y}_I)].$$

Hence

$$\text{corr}(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}\text{corr}(\hat{Y}, \hat{Y}_I) = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T\boldsymbol{r}_I/\boldsymbol{r}^T\boldsymbol{r}$. Since $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) = \boldsymbol{I} - \boldsymbol{H}$, the numerator $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{r}^T\boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}}[\text{corr}(r, r_I)].$$

Hence

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[\text{corr}(r, r_I)] = [\text{corr}(r, r_I)]^2.$$

Algebra shows that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}. \quad \square$$

Remark 4.2. Let I_{min} be the model than minimizes $C_p(I)$ among the models I generated from the variable selection method such as forward se-

lection. Assuming the the full model I_p is one of the models generated, then $C_p(I_{min}) \leq C_p(I_p) = p$, and $\text{corr}(r, r_{I_{min}}) \to 1$ as $n \to \infty$ by Theorem 4.2 vi). Referring to Equation (4.1), if $P(S \subseteq I_{min})$ does not go to 1 as $n \to \infty$, then the above correlation would not go to one. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$.

A standard model selection procedure will often be needed to suggest models. For example, forward selection or backward elimination could be used. If $p < 30$, Furnival and Wilson (1974) provide a technique for selecting a few candidate subsets after examining all possible subsets.

Remark 4.3. Daniel and Wood (1980, p. 85) suggest using Mallows' graphical method for screening subsets by plotting k versus $C_p(I)$ for models close to or under the $C_p = k$ line. Theorem 4.2 vi) implies that if $C_p(I) \leq k$ or $F_I < 1$, then $\text{corr}(r, r_I)$ and $\text{corr}(ESP, ESP(I))$ both go to 1.0 as $n \to \infty$. Hence models I that satisfy the $C_p(I) \leq k$ screen will contain the true model S with high probability when n is large. This result does not guarantee that the true model S will satisfy the screen, but overfit is likely. Let d be a lower bound on $\text{corr}(r, r_I)$. Theorem 4.2 vi) implies that if

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(r, r_I) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

$$d \equiv d_n = \sqrt{1 - \frac{p}{n}}.$$

To avoid excluding too many good submodels, consider models I with $C_p(I) \leq \min(2k, p)$. Models under both the $C_p = k$ line and the $C_p = 2k$ line are of interest.

Rule of thumb 4.1. a) After using a numerical method such as forward selection or backward elimination, let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model. Do not use more predictors than model I_I to avoid overfitting.

b) Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined.

c) Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit: important predictors are deleted from the model. Underfit is especially likely to occur if a predictor with one degree of freedom is deleted (if the $c - 1$ indicator variables corresponding to a factor are deleted, then

the factor has $c-1$ degrees of freedom) and the jump in C_p is large, greater than 4, say.

d) If there are no models I with fewer predictors than I_I such that $C_p(I) \leq \min(2k, p)$, then model I_I is a good candidate for the best subset found by the numerical procedure.

Forward selection forms a sequence of submodels $I_1, ..., I_p$ where I_j uses j predictors including the constant. Let I_1 use $x_1^* = x_1 \equiv 1$: the model has a constant but no nontrivial predictors. To form I_2, consider all models I with two predictors including x_1^*. Compute $SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_2 minimize $SSE(I)$ for the $p-1$ models I that contain x_1^* and one other predictor. Denote the predictors in I_2 by x_1^*, x_2^*. In general, to form I_j consider all models I with j predictors including variables $x_1^*, ..., x_{j-1}^*$. Compute $SSE(I)$ and let I_j minimize $SSE(I)$ for the $p-j+1$ models I that contain $x_1^*, ..., x_{j-1}^*$ and one other predictor not already selected. Denote the predictors in I_j by $x_1^*, ..., x_j^*$. Continue in this manner for $j = 2, ..., M = p$.

Backward elimination also forms a sequence of submodels $I_1, ..., I_p$ where I_j uses j predictors including the constant. Let I_p be the full model. To form I_{p-1} consider all models I with $p-1$ predictors including the constant. Compute $SSE(I)$, and let I_{p-1} minimize $Q_{p-1}(I)$ for the $p-1$ models I that exclude one of the predictors $x_2, ..., x_p$. Denote the predictors in I_{p-1} by $x_1^*, x_2^*, ..., x_{p-1}^*$. In general, to form I_j consider all models I with j predictors including variables $x_1^*, ..., x_{j+1}^*$. Compute $SSE(I)$, and let I_j minimize $SSE(I)$ for the $p-j+1$ models I that exclude one of the predictors $x_2^*, ..., x_{j+1}^*$. Denote the predictors in I_j by $x_1^*, ..., x_j^*$. Continue in this manner for $j = p = M, p-1, ..., 2, 1$ where I_1 uses $x_1^* = x_1 \equiv 1$.

Several criterion produce the same sequence of models if forward selection or backward elimination are used, including $MSE(I)$, $C_p(I)$, $R_A^2(I)$, $AIC(I)$, $BIC(I)$, and $EBIC(I)$. This result holds since if the number of predictors k in the model I is fixed, the criterion is equivalent to minimizing $SSE(I)$ plus a constant. The constants differ so the model I_{min} that minimizes the criterion often differ. Heuristically, backward elimination tries to delete the variable that will increase C_p the least while forward selection tries to add the variable that will decrease C_p the most.

When there is a sequence of M submodels, the final submodel I_d needs to be selected with a_d terms, including a constant. Let the candidate model I contain a terms, including a constant, and let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors. Then there are many criteria used to select the final submodel I_d. For a given data set, the quantities p, n, and $\hat{\sigma}^2$ act as constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset I_{min} that minimizes the criterion.

Let criteria $C_S(I)$ have the form

$$C_S(I) = SSE(I) + aK_n\hat{\sigma}^2.$$

These criteria need a good estimator of σ^2 and n/p large. See Shibata (1984). The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. See Jones (1946) and Mallows (1973) for C_p. It can be shown that $C_p(I) = AIC_S(I)$ is equivalent to the $C_P(I)$ criterion of Definition 4.2. Typically $\hat{\sigma}^2$ is the OLS full model MSE when n/p is large.

The following criteria also need n/p large. AIC is due to Akaike (1973), AIC_C is due to Hurvich and Tsai (1989), and BIC to Schwarz (1978) and Akaike (1977, 1978). Also see Burnham and Anderson (2004).

$$AIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2a,$$

$$AIC_C(I) = n \log\left(\frac{SSE(I)}{n}\right) + \frac{2a(a+1)}{n-a-1},$$

$$\text{and } BIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + a \log(n).$$

Forward selection with C_p and AIC often gives useful results if $n \geq 5p$ and if the final model has $n \geq 10a_d$. For $p < n < 5p$, forward selection with C_p and AIC tends to pick the full model (which overfits since $n < 5p$) too often, especially if $\hat{\sigma}^2 = MSE$. The Hurvich and Tsai (1989, 1991) AIC_C criterion can be useful if $n \geq \max(2p, 10a_d)$.

The EBIC criterion given in Luo and Chen (2013) may be useful when n/p is not large. Let $0 \leq \gamma \leq 1$ and $|I| = a \leq \min(n, p)$ if $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$. We may use $a \leq \min(n/5, p)$. Then $EBIC(I) =$

$$n \log\left(\frac{SSE(I)}{n}\right) + a \log(n) + 2\gamma \log\left[\binom{p}{a}\right] = BIC(I) + 2\gamma \log\left[\binom{p}{a}\right].$$

This criterion can give good results if $p = p_n = O(n^k)$ and $\gamma > 1 - 1/(2k)$. Hence we will use $\gamma = 1$. Then minimizing $EBIC(I)$ is equivalent to minimizing $BIC(I) - 2\log[(p-a)!] - 2\log(a!)$ since $\log(p!)$ is a constant.

The above criteria can be applied to forward selection and relaxed lasso. The C_p criterion can also be applied to lasso. See Efron and Hastie (2016, pp. 221, 231).

Now suppose $p = 6$ and S in Equation (4.1) corresponds to $x_1 \equiv 1, x_2$, and x_3. Suppose the data set is such that underfitting (omitting a predictor in S) does not occur. Then there are eight possible submodels that contain S: i) x_1, x_2, x_3; ii) x_1, x_2, x_3, x_4; iii) x_1, x_2, x_3, x_5; iv) x_1, x_2, x_3, x_6; v) x_1, x_2, x_3, x_4, x_5; vi) x_1, x_2, x_3, x_4, x_6; vii) x_1, x_2, x_3, x_5, x_6; and the full model viii) $x_1, x_2, x_3, x_4, x_5, x_6$. The possible submodel sizes are $k = 3, 4, 5$, or 6. Since the variable selection criteria for forward selection described above minimize the MSE given that $x_1^*, ..., x_{k-1}^*$ are in the model, the $MSE(I_k)$ are too small and underestimate σ^2. Also the model I_{min} fits the data a bit too well. Suppose $I_{min} = I_d$. Compared to selecting a model I_k before examining

the data, the residuals $r_i(I_{min})$ are too small in magnitude, the $|\hat{Y}_{I_{min},i} - Y_i|$ are too small, and $MSE(I_{min})$ is too small. Hence using $I_{min} = I_d$ as the full model for inference does not work. In particular, the partial F test statistic F_R in Theorem 2.27, using I_d as the full model, is too large since the MSE is too small. Thus the partial F test rejects H_0 too often. Similarly, the confidence intervals for β_i are too short, and hypothesis tests reject $H_0 : \beta_i = 0$ too often when H_0 is true. The fact that the selected model I_{min} from variable selection cannot be used as the full model for classical inference is known as **selection bias**. Also see Hurvich and Tsai (1990).

This chapter offers two remedies: i) use the large sample theory of $\hat{\boldsymbol{\beta}}_{I_{min},0}$ (defined two paragraphs below) and the bootstrap for inference after variable selection, and ii) use data splitting for inference after variable selection.

4.2 Large Sample Theory for Some Variable Selection Estimators

Large sample theory is often tractable if the optimization problem is convex. The optimization problem for variable selection is not convex, so new tools are needed. Tibshirani et al. (2018) and Leeb and Pötscher (2006, 2008) note that we can not find the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_{min}} - \boldsymbol{\beta}_I)$ after variable selection. One reason is that with positive probability, $\hat{\boldsymbol{\beta}}_{I_{min}}$ does not have the same dimension as $\boldsymbol{\beta}_I$ if AIC or C_p is used. Hence $\boldsymbol{Z}_n$ is not defined with positive probability.

The large sample theory for OLS variable selection estimators such as forward selection and lasso variable selection in this section is due to Pelawa Watagoda and Olive (2019, 2020). Rathnayake and Olive (2020) extend this theory to many other variable selection estimators such as generalized linear models. Charkhi and Claeskens (2018) have a related result for forward selection with AIC when the iid errors are $N(0, \sigma^2)$. Assume p is fixed, and $n \to \infty$. Suppose that model (4.1) holds. Assume the maximum leverage

$$\max_{i=1,\ldots,n} \boldsymbol{x}_{iI_j}^T (\boldsymbol{X}_{I_j}^T \boldsymbol{X}_{I_j})^{-1} \boldsymbol{x}_{iI_j} \to 0$$

in probability as $n \to \infty$ for each I_j with $S \subseteq I_j$ where the dimension of I_j is a_j. For the OLS model with $S \subseteq I_j$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j = \sigma^2 \boldsymbol{W}_j$ and $(\boldsymbol{X}_{I_j}^T \boldsymbol{X}_{I_j})/n \xrightarrow{P} \boldsymbol{W}_j^{-1}$ by the LS CLT Theorem 2.26. Then

$$\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{4.3}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model.

For MLR, $\boldsymbol{V}_{j,0} = \sigma^2 \boldsymbol{W}_{j,0}$. For example, if $p = 3$ and model I_j uses a constant $x_1 \equiv 1$ and x_3 with

$$\boldsymbol{V}_j = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \quad \text{then} \quad \boldsymbol{V}_{\mathrm{j},0} = \begin{bmatrix} V_{11} & 0 & V_{12} \\ 0 & 0 & 0 \\ V_{21} & 0 & V_{22} \end{bmatrix}.$$

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. Use zero padding to form the $p \times 1$ variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. In the following definition, if each subset contains at least one variable, then there are $J = 2^p - 1$ subsets.

Definition 4.4. The *variable selection estimator* $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$, and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets.

Definition 4.5. Let $\hat{\boldsymbol{\beta}}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities equal to π_{kn}. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with same probabilities π_{kn} of the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the I_k are randomly selected.

The large sample distribution of $\hat{\boldsymbol{\beta}}_{MIX}$ is simpler than that of $\hat{\boldsymbol{\beta}}_{VS}$, and is useful for explaining the large sample distribution of $\hat{\boldsymbol{\beta}}_{VS}$. For how to bootstrap $\hat{\boldsymbol{\beta}}_{MIX}$, see Rathnayake and Olive (2020). For mixture distributions, see Section 1.6.

The first assumption in Theorem 4.3 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the variable selection estimator corresponding to I_{min} underfits with probability going to zero, and the assumption holds under regularity conditions if BIC or AIC is used. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). For multiple linear regression with Mallows (1973) C_p or AIC, see Li (1987), Nishii (1984), and Shao (1993). For a shrinkage estimator that does variable selection, let $\hat{\boldsymbol{\beta}}_{I_{min}}$ be the OLS estimator applied to a constant and the variables with nonzero shrinkage estimator coefficients. If the shrinkage estimator is a consistent estimator of $\boldsymbol{\beta}$, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. See Zhao and Yu (2006, p. 2554). Hence Theorem 4.3c) proves that the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ if lasso and elastic net are consistent. Also see Theorem 4.4 and Remark 4.5. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 4.3 is reasonable by (4.3) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection.

Theorem 4.3. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive

π_k by π_j. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{4.4}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{4.5}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

c) The estimator $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{d,0})$ where SEL is VS or MIX.

Proof. a) Since $\boldsymbol{u}_n$ has a mixture distribution of the $\boldsymbol{u}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{u}_n$ is $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \to \infty$.

b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$.

c) The result follows since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).

d) If $\pi_d = 1$, there is no selection bias, asymptotically. The result also follows by Pötscher (1991, Lemma 1). □

The following subscript notation is useful. Subscripts before the MIX are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. Let $\hat{\boldsymbol{\beta}}_{i,MIX} = \hat{\beta}_i$. Similarly, if $I = \{i_1, ..., i_a\}$, then $\hat{\boldsymbol{\beta}}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T$. Subscripts after MIX denote the ith vector from a sample $\hat{\boldsymbol{\beta}}_{MIX,1}, ..., \hat{\boldsymbol{\beta}}_{MIX,B}$. Similar notation is used for other estimators such as $\hat{\boldsymbol{\beta}}_{VS}$. The subscript 0 is still used for zero padding. We may use $FULL$ to denote the full model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$.

Typically the mixture distribution is not asymptotically normal unless a $\pi_d = 1$ (e.g. if S is the full model), or if for each π_j, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This special case occurs for $\hat{\boldsymbol{\beta}}_{S,MIX}$ if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ where the asymptotic covariance matrix $\boldsymbol{V}$ is diagonal and nonsingular. Then $\hat{\boldsymbol{\beta}}_{S,MIX}$ and $\hat{\boldsymbol{\beta}}_{S,FULL}$ have the same multivariate normal limiting distribution. For several criteria, this result should hold for $\hat{\boldsymbol{\beta}}_{VS}$ since asymptotically, $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{A}\boldsymbol{\beta})$ is selecting from the $\boldsymbol{A}\boldsymbol{u}_j$ which have the same distribution. Then the confidence regions applied to $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL} = \boldsymbol{B}\hat{\boldsymbol{\beta}}^*_{S,SEL}$ should have similar volume and cutoffs where SEL is MIX, VS, or $FULL$.

Theorem 4.3 can be used to justify prediction intervals after variable selection. See Pelawa Watagoda and Olive (2020). Theorem 4.3d) is useful for *variable selection consistency* and the *oracle property* where $\pi_d = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Claeskens and Hjort (2008, pp. 101-114) and Fan and Li (2001) for references. A necessary condition for $P(I_{min} = S) \to 1$ is that S is one of the models considered with probability going to one. This condition holds under strong regularity conditions for fast methods. See Wieczorek (2018) for forward selection and Hastie et al. (2015, pp. 295-302) for lasso, where the predictors need a "near orthogonality" condition.

Remark 4.4. If $A_1, A_2, ..., A_k$ are pairwise disjoint and if $\cup_{i=1}^{k} A_i = S$, then the collection of sets $A_1, A_2, ..., A_k$ is a *partition* of S. Then the *Law of Total Probability* states that if $A_1, A_2, ..., A_k$ form a partition of S such that $P(A_i) > 0$ for $i = 1, ..., k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

Let sets $A_{k+1}, ..., A_m$ satisfy $P(A_i) = 0$ for $i = k+1, ..., m$. Define $P(B|A_j) = 0$ if $P(A_j = 0$. Then a Generalized Law of Total Probability is

$$P(B) = \sum_{j=1}^{m} P(B \cap A_j) = \sum_{j=1}^{m} P(B|A_j)P(A_j),$$

and will be used in the following paragraph.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $W = W_{VS} = k$ if $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ where $P(W_{VS} = k) = \pi_{kn}$ for $k = 1, ..., J$. Then $(\hat{\boldsymbol{\beta}}_{VS:n}, W_{VS:n}) = (\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$ has a joint distribution where the sample size n is usually suppressed. Note that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. Define $P(B|A_k)P(A_k) = 0$ if $P(A_k) = 0$. Let $\hat{\boldsymbol{\beta}}^{C}_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(W_{VS} = k)$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(W_{VS} = k) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^{C}_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \le t_1, ..., z_p \le t_p)$ by $P(\boldsymbol{z} \le \boldsymbol{t})$. Then

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]\pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}]\pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities π_{kn}, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}.

Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the MLE with AIC. Here $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\mathbf{0}$. Hence $E(\boldsymbol{w}_j) = 0$, and $\mathrm{Cov}(\boldsymbol{w}_j) = \boldsymbol{\Sigma}_j$ exits. Referring to Definitions 4.4 and 4.5, note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$ of Equation (4.3). The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$ does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ to that of $\boldsymbol{w}_{jn}$. Similarly, selection bias does change the distribution of the selected $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_j$. The reasonable Theorem 4.4 assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild.

Theorem 4.4, Variable Selection CLT. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{4.6}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$. Thus $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

Proof. Since $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{w}_n$ is $F_{\boldsymbol{w}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{w}_j}(\boldsymbol{t})$ as $n \to \infty$. □

Remark 4.5. If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959). By both this result and Theorems 4.3 and 4.4, the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent.

Mixture distributions are useful for variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_j,0}$. Review mixture distributions from Section 1.6. The following theorem is due to Pelawa Watagoda and Olive (2019a). Note that the cdf of T_n is $F_{T_n}(\boldsymbol{z}) = \sum_j \pi_{jn} F_{T_{jn}}(\boldsymbol{z})$ where $F_{T_{jn}}(\boldsymbol{z})$ is the cdf of T_{jn}.

Theorem 4.5, Mixture Distribution CLT. Suppose the $g \times 1$ statistic T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$ where

$\sum_j \pi_{jn} = 1$, $\pi_{jn} \to \pi_j$ as $n \to \infty$, and $\boldsymbol{u}_{jn} = \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$ with $E(\boldsymbol{u}_j) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}_j) = \boldsymbol{\Sigma}_j$. Then

$$\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{4.7}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ and $F_{\boldsymbol{u}_j}(\boldsymbol{z})$ is the cdf of $\boldsymbol{u}_j$. Thus, $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{\Sigma}_j$.

Proof: Note that T_n has a mixture distribution of the T_{jn} with probabilities π_{jn}. Hence $\sqrt{n}(T_n - \boldsymbol{\theta})$ has a mixture distribution of the $\boldsymbol{u}_{jn} = \sqrt{n}(T_{jn} - \boldsymbol{\theta})$, and the cdf of $\sqrt{n}(T_n - \boldsymbol{\theta})$ is $\sum_j \pi_{jn} F_{\boldsymbol{u}_{jn}}(\boldsymbol{z}) \to \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ at continuity points $\boldsymbol{z}$ of the $F_{\boldsymbol{u}_j}$. □

Remark 4.6. Another variable selection model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T \boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then submodel I underfits if no $S_i \subseteq I$. A necessary condition for an estimator to be consistent is P(no $S_i \subseteq I_{min}$) $\to 0$ as $n \to \infty$. Then in Theorem 4.4, we can replace $P(S \subseteq I_{min}) \to 1$ by P(no $S_i \subseteq I_{min}$) $\to 0$ as $n \to \infty$.

4.3 Prediction Intervals

Prediction intervals for regression and prediction regions for multivariate regression are important topics. Inference after variable selection will consider bootstrap hypothesis testing. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. The prediction intervals and regions are based on samples of size n, while the bootstrap sample size is $B = B_n$. Hence this section and the following section are important.

Definition 4.6. Consider predicting a future test value Y_f given a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ and training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

If $Y_f|\boldsymbol{x}_f$ has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as $n \to \infty$. The interpretation of a 100 $(1-\delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where the k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI,

then the probability that $Y_{fi} \in PI_i$ for j of the PIs approximately follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated. This section will describe three nonparametric PIs for the additive error regression model, $Y = m(\boldsymbol{x}) + e$, that work well for a large class of unknown zero mean error distributions.

First we will consider the location model, $Y_i = \mu + e_i$, where $Y_1, ..., Y_n, Y_f$ are iid and there are no vectors of predictors $\boldsymbol{x}_i$ and $\boldsymbol{x}_f$. Let $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$ be the order statistics of n iid random variables $Z_1, ..., Z_n$. Let a future random variable Z_f be such that $Z_1, ..., Z_n, Z_f$ are iid. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1 - \delta/2) \rceil$ where $\lceil x \rceil$ is the smallest integer $\geq x$. For example, $\lceil 7.7 \rceil = 8$. Then a common nonparametric large sample $100(1-\delta)\%$ prediction interval for Z_f is

$$[Z_{(k_1)}, Z_{(k_2)}] \tag{4.8}$$

where $0 < \delta < 1$. See Frey (2013) for references.

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. With the Z_i and $Z_{(i)}$ as in the above paragraph, let the shortest closed interval containing at least c of the Z_i be

$$\text{shorth(c)} = [\text{Z}_{(\text{s})}, \text{Z}_{(\text{s+c}-1)}]. \tag{4.9}$$

Let

$$k_n = \lceil n(1 - \delta) \rceil. \tag{4.10}$$

Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth$(c)$ estimator as the large sample $100(1 - \delta)\%$ PI where

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{4.11}$$

An interesting fact is that the maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions where such a distribution has pdf $f(y) = 1/(\theta_2 - \theta_1)$ for $\theta_1 \leq y \leq \theta_2$ where $f(y) = 0$, otherwise, and $\theta_1 < \theta_2$.

A problem with the prediction intervals that cover $\approx 100(1 - \delta)\%$ of the training data cases Y_i (such as (4.8) using $c = k_n$ given by (4.9)), is that they have coverage lower than the nominal coverage of $1 - \delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. For iid data, Frey (2013) used (4.10) to correct for undercoverage.

Example 4.1. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778

                  13 = 89 - 76

                      33 = 111 - 78

                          689 = 778 - 89
shorth(3) = [76,89]
```

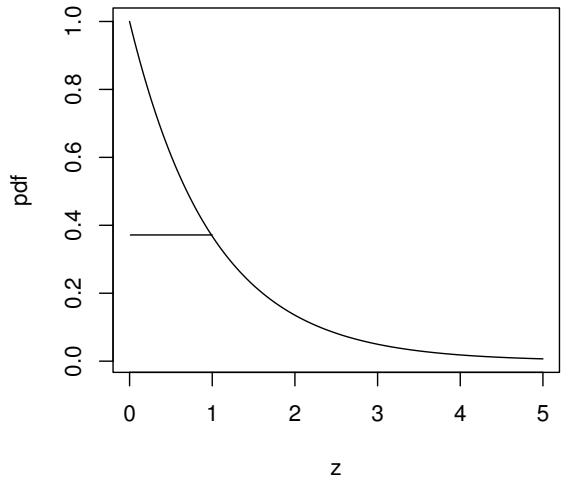

Fig. 4.1 The 36.8% Highest Density Region is [0,1]

Remark. 4.7. The large sample $100(1-\delta)\%$ shorth PI (4.10) may or may not be asymptotically optimal if the $100(1-\delta)\%$ population shorth is $[L_s, U_s]$ and $F(x)$ is not strictly increasing in intervals $(L_s - \delta, L_s + \delta)$ and $(U_s - \delta, U_s + \delta)$ for some $\delta > 0$. To see the issue, suppose Y has probability mass function (pmf) $p(0) = 0.4$, $p(1) = 0.3$, $p(2) = 0.2$, $p(3) = 0.06$, and $p(4) = 0.04$. Then the 90% population shorth is [0,2] and the $100(1-\delta)\%$

population shorth is [0,3] for $(1-\delta) \in (0.9, 0.96]$. Let $W_i = I(Y_i \leq x) = 1$ if $Y_i \leq x$ and 0, otherwise. The empirical cdf

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_{(i)} \leq x)$$

is the sample proportion of $Y_i \leq x$. If $Y_1, ..., Y_n$ are iid, then for fixed x, $n\hat{F}_n(x) \sim binomial(n, F(x))$. Thus $\hat{F}_n(x) \sim AN(F(x), F(x)(1-F(x))/n)$. For the Y with the above pmf, $\hat{F}_n(2) \xrightarrow{P} 0.9$ as $n \to \infty$ with $P(\hat{F}_n(2) < 0.9) \to 0.5$ and $P(\hat{F}_n(2) \geq 0.9) \to 0.5$ as $n \to \infty$. Hence the large sample 90% PI (4.10) will be [0,2] or [0,3] with probabilities $\to 0.5$ as $n \to \infty$ with expected asymptotic length of 2.5 and expected asymptotic coverage converging to 0.93. However, the large sample $100(1-\delta)\%$ PI (4.10) converges to [0,3] and is asymptotically optimal with asymptotic coverage 0.96 for $(1-\delta) \in (0.9, 0.96)$.

For a random variable Y, the $100(1-\delta)\%$ highest density region is a union of $k \geq 1$ disjoint intervals such that the mass within the intervals $\geq 1-\delta$ and the sum of the k interval lengths is as small as possible. Suppose that $f(z)$ is a unimodal pdf that has interval support, and that the pdf $f(z)$ of Y decreases rapidly as z moves away from the mode. Let $[a, b]$ be the shortest interval such that $F_Y(b) - F_Y(a) = 1-\delta$ where the cdf $F_Y(z) = P(Y \leq z)$. Then the interval $[a, b]$ is the $100(1-\delta)$ highest density region. To find the $100(1-\delta)\%$ highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $[a_1, b_1], ..., [a_k, b_k]$ for some $k \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is equal to $1-\delta$. As an example, let $f(z) = e^{-z}$ for $z > 0$. See Figure 4.1 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest density region. The shorth PI estimates the highest density interval which is the highest density region for a distribution with a unimodal pdf. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

The additive error regression model is $Y = m(\boldsymbol{x}) + e$ where $m(\boldsymbol{x})$ is a real valued function and the e_i are iid, often with zero mean and constant variance $V(e) = \sigma^2$. The large sample theory for prediction intervals is simple for this model, and variable selection models for the multiple linear regression model have this form with $m(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ if $S \subseteq I$. Let the residuals $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$ for $i = 1, ..., n$. Assume $\hat{m}(\boldsymbol{x})$ is a consistent estimator of $m(\boldsymbol{x})$ such that the sample percentiles $[\hat{L}_n(r), \hat{U}_n(r)]$ of the residuals are consistent estimators of the population percentiles $[L, U]$ of the error distribution where $P(e \in [L, U]) = 1-\delta$. Let $\hat{Y}_f = \hat{m}(\boldsymbol{x}_f)$. Then $P(Y_f \in [\hat{Y}_f + \hat{L}_n(r), \hat{Y}_f + \hat{U}_n(r)] \to P(Y_f \in [m(\boldsymbol{x}_f)+L, m(\boldsymbol{x}_f)+U]) = P(e \in [L, U]) = 1-\delta$ as $n \to \infty$. Three common choices are a) $P(e \leq U) = 1-\delta/2$ and $P(e \leq L) = \delta/2$, b)

$P(e^2 \leq U^2) = P(|e| \leq U) = P(-U \leq e \leq U) = 1 - \delta$ with $L = -U$, and c) the population shorth is the shortest interval (with length $U - L$) such that $P[e \in [L, U]) = 1 - \delta$. The PI c) is asymptotically optimal while a) and b) are asymptotically optimal on the class of symmetric zero mean unimodal error distributions. The split conformal PI (4.16), described below, estimates $[-U, U]$ in b).

Prediction intervals based on the shorth of the residuals need a correction factor for good coverage since the residuals tend to underestimate the errors in magnitude. With the exception of ridge regression, let d be the number of "variables" used by the method. For MLR, forward selection, lasso, and relaxed lasso use variables $x_1^*, ..., x_d^*$ while PCR and PLS use variables that are linear combinations of the predictors $V_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ for $j = 1, ..., d$. (We could let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used.) See Chapter 5 for more about these estimators. See Hong et al. (2018) for why classical prediction intervals after variable selection fail to work.

For n/p large and $d = p$, Olive (2013a) developed prediction intervals for models of the form $Y_i = m(\boldsymbol{x}_i) + e_i$, and variable selection models for MLR have this form, as noted by Olive (2018). Pelawa Watagoda and Olive (2019b) gave two prediction intervals that can be useful even if n/p is not large. These PIs will be defined below. The first PI modifies the Olive (2013a) PI that can only be computed if $n > p$. Olive (2007, 2017a, 2017b, 2018) used similar correction factors for several prediction intervals and prediction regions with $d = p$. We want $n \geq 10d$ so that the model does not overfit.

If the OLS model I has d predictors, and $S \subseteq I$, then

$$E(MSE(I)) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-d}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right)$$

and $MSE(I)$ is a $\sqrt{n}$ consistent estimator of σ^2 for many error distributions by Su and Cook (2012). Also see Freedman (1981). For a wide range of regression models, extrapolation occurs if the leverage $h_f = \boldsymbol{x}_{I,f}^T (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{x}_{I,f} > 2d/n$: if $\boldsymbol{x}_{I,f}$ is too far from the data $\boldsymbol{x}_{I,1}, ..., \boldsymbol{x}_{I,n}$, then the model may not hold and prediction can be arbitrarily bad. These results suggests that

$$\sqrt{\frac{n}{n-d}}\sqrt{(1+h_f)} \;\; r_i \approx \sqrt{\frac{n+2d}{n-d}} \;\; r_i \approx e_i.$$

In simulations for prediction intervals and prediction regions with $n = 20d$, the maximum simulated undercoverage was near 5% if q_n in (4.11) is changed to $q_n = 1 - \delta$.

Next we give the correction factor and the first prediction interval. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.} \tag{4.12}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let

$$c = \lceil nq_n \rceil, \tag{4.13}$$

and let

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \tag{4.14}$$

if $d \leq 8n/9$, and

$$b_n = 5\left(1 + \frac{15}{n}\right),$$

otherwise. As d gets close to n, the model overfits and the coverage will be less than the nominal. The piecewise formula for b_n allows the prediction interval to be computed even if $d \geq n$. Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then the first 100 $(1 - \delta)\%$ large sample PI for Y_f is

$$[\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}]. \tag{4.15}$$

The second PI randomly divides the data into two half sets H and V where H has $n_H = \lceil n/2 \rceil$ of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $\hat{m}_H(\boldsymbol{x})$ is computed using the training data set H. Then the validation residuals $v_j = Y_{i_j} - \hat{m}_H(\boldsymbol{x}_{i_j})$ are computed for the $j = 1, ..., n_V$ cases in the validation set V. Find the Frey PI $[v_{(s)}, v_{(s+c-1)}]$ of the validation residuals (replacing n in (4.10) by $n_V = n - n_H$). Then the second new $100(1 - \delta)\%$ large sample PI for Y_f is

$$[\hat{m}_H(\boldsymbol{x}_f) + v_{(s)}, \hat{m}_H(\boldsymbol{x}_f) + v_{(s+c-1)}]. \tag{4.16}$$

Remark 4.8. Note that correction factors $b_n \to 1$ are used in large sample confidence intervals and tests if the limiting distribution is N(0,1) or χ^2_p, but a t_{d_n} or pF_{p,d_n} cutoff is used: $t_{d_n,1-\delta}/z_{1-\delta} \to 1$ and $pF_{p,d_n,1-\delta}/\chi^2_{p,1-\delta} \to 1$ if $d_n \to \infty$ as $n \to 1$. Using correction factors for large sample confidence intervals, tests, prediction intervals, prediction regions, and bootstrap confidence regions improves the performance for moderate sample size n.

Remark 4.9. For a good fitting model, residuals r_i tend to be smaller in magnitude than the errors e_i, while validation residuals v_i tend to be larger in magnitude than the e_i. Thus the Frey correction factor can be used for PI (4.15) while PI (4.14) needs a stronger correction factor.

We can also motivate PI (4.15) by modifying the justification for the Lei et al. (2018) split conformal prediction interval

$$[\hat{m}_H(\boldsymbol{x}_f) - a_q, \hat{m}_H(\boldsymbol{x}_f) + a_q] \tag{4.17}$$

where a_q is the $100(1-\alpha)$th quantile of the absolute validation residuals. PI (4.15) is a modification of the split conformal PI that is asymptotically optimal. Suppose $(Y_i, \boldsymbol{x}_i)$ are iid for $i = 1, ..., n, n+1$ where $(Y_f, \boldsymbol{x}_f) = (Y_{n+1}, \boldsymbol{x}_{n+1})$. Compute $\hat{m}_H(\boldsymbol{x})$ from the cases in H. For example, get $\hat{\boldsymbol{\beta}}_H$ from the cases in H. Consider the validation residuals v_i for $i = 1, ..., n_V$ and the validation residual v_{n_V+1} for case $(Y_f, \boldsymbol{x}_f)$. Since these $n_V + 1$ cases are iid, the probability that v_t has rank j for $j = 1, ..., n_V + 1$ is $1/(n_V + 1)$ for each t, i.e., the ranks follow the discrete uniform distribution. Let $t = n_V + 1$ and let the $v_{(j)}$ be the ordered residuals using $j = 1, ..., n_V$. That is, get the order statistics without using the unknown validation residual v_{n_V+1}. Then $v_{(i)}$ has rank i if $v_{(i)} < v_{n_V+1}$ but rank $i+1$ if $v_{(i)} > v_{n_V+1}$. Thus

$$P(Y_f \in [\hat{m}_H(\boldsymbol{x}_f)+v_{(k)}, \hat{m}_H(\boldsymbol{x}_f)+v_{(k+b-1)}]) = P(v_{(k)} \leq v_{n_V+1} \leq v_{(k+b-1)}) \geq$$

$P(v_{n_V+1}$ has rank between $k+1$ and $k+b-1$ and there are no tied ranks) $\geq (b-1)/(n_V+1) \approx 1-\delta$ if $b = \lceil (n_V+1)(1-\delta) \rceil + 1$ and $k+b-1 \leq n_V$. This probability statement holds for a fixed k such as $k = \lceil n_V \ \delta/2 \rceil$. The statement is not true when the shorth(b) estimator is used since the shortest interval using $k = s$ can have s change with the data set. That is, s is not fixed. Hence if PI's were made from J independent data sets, the PI's with fixed k would contain Y_f about $J(1-\delta)$ times, but this value would be smaller for the shorth(b) prediction intervals where s can change with the data set. The above argument works if the estimator $\hat{m}(\boldsymbol{x})$ is "symmetric in the data," which is satisfied for multiple linear regression estimators.

The PIs (4.14) to (4.16) can be used with $\hat{m}(\boldsymbol{x}) = \hat{Y}_f = \boldsymbol{x}_{I_d}^T \hat{\boldsymbol{\beta}}_{I_d}$ where I_d denotes the index of predictors selected from the model or variable selection method. If $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, the PIs (4.14) and (4.15) are asymptotically optimal for a large class of error distributions while the split conformal PI (4.16) needs the error distribution to be unimodal and symmetric for asymptotic optimality. Since $\hat{m}_H$ uses $n/2$ cases, $\hat{m}_H$ has about half the efficiency of $\hat{m}$. When $p \geq n$, the regularity conditions for consistent estimators are strong. For example, EBIC and lasso can have $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then forward selection with EBIC and relaxed lasso can produce consistent estimators. PLS can be $\sqrt{n}$ consistent. See Chapter 5 for the large sample for many MLR estimators.

None of the three prediction intervals (4.14), (4.15), and (4.16) dominates the other two. Recall that $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector in (4.1). If a good fitting method, such as lasso or forward selection with EBIC, is used, and $1.5a_S \leq n \leq 5a_S$, then PI (4.14) can be much shorter than PIs (4.15) and (4.16). For n/d large, PIs (4.14) and (4.15) can be shorter than PI (4.16) if the error distribution is not unimodal and symmetric; however, PI (4.16) is often shorter if n/d is not large since the sample shorth converges to the population shorth rather slowly. Grübel (1982) shows that for iid data, the length and center the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval. For a

unimodal and symmetric error distribution, the three PIs are asymptotically equivalent, but PI (4.16) can be the shortest PI due to different correction factors.

If the estimator is poor, the split conformal PI (4.16) and PI (4.15) can have coverage closer to the nominal coverage than PI (4.14). For example, if $\hat{m}$ interpolates the data and $\hat{m}_H$ interpolates the training data from H, then the validation residuals will be huge. Hence PI (4.15) will be long compared to PI (4.16).

Asymptotically optimal PIs estimate the population shorth of the zero mean error distribution. Hence PIs that use the shorth of the residuals, such as PIs (4.14) and (4.15), are the only easily computed asymptotically optimal PIs for a wide range of consistent estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for the multiple linear regression model. If the error distribution is $e \sim EXP(1) - 1$, then the asymptotic length of the 95% PI (4.14) or (4.15) is 2.966 while that of the split conformal PI is $2(1.966) = 3.992$. For more about these PIs applied to MLR models, see Section 5.10 and Pelawa Watagoda and Olive (2019b).

4.4 Prediction Regions

Consider predicting a $p \times 1$ future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Much as confidence regions and intervals give a measure of precision for the point estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, prediction regions and intervals give a measure of precision of the point estimator $T = \hat{\boldsymbol{x}}_f$ of the future random vector $\boldsymbol{x}_f$.

Definition 4.7. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. A prediction region is *asymptotically optimal* if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

If $\boldsymbol{x}_f$ has a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \to 1 - \delta$ as $n \to \infty$. A PI is a prediction region where $p = 1$. Highest density regions are usually hard to estimate for p not much larger than four, but many elliptically contoured distributions with a nonsingular population covariance matrix, including the multivariate normal distribution, have highest density regions that can be estimated by the nonparametric prediction region (4.24). For more about highest density regions, see Olive (2017b, pp. 148-155) and Hyndman (1996).

For multivariate data, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. Let the observed training data be collected in an $n \times p$ matrix $\boldsymbol{W}$. Let the $p \times 1$ column vector $T = T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator.

Definition 4.8. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 4.9. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be the data where $\boldsymbol{x}_i$ is a $p \times 1$ vector. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T \boldsymbol{1}$$

where $\boldsymbol{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T =$

$$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{W}.$$

Hence if the *centering matrix* $\boldsymbol{G} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{G}\boldsymbol{W}$.

See Definition 1.24 for the population mean and population covariance matrix. Definition 2.18 also defined a sample covariance matrix. The Ma-

halanobis distance in Definition 4.9 is a random variable that estimates the population Mahalanobis distance of Definition 1.38.

Definition 4.9. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (4.18)$$

for each point $\boldsymbol{x}_i$. Notice that D_i^2 is a random variable (scalar valued). Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$. Then

$$D_{\boldsymbol{x}}^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T).$$

Hence D_i^2 uses $\boldsymbol{x} = \boldsymbol{x}_i$.

Let the $p \times 1$ location vector be $\boldsymbol{\mu}$, often the population mean, and let the $p \times p$ dispersion matrix be $\boldsymbol{\Sigma}$, often the population covariance matrix. Notice that if $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance from Definition 1.38 is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \quad (4.19)$$

and that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0, 1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}^2\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\}. \quad (4.20)$$

If n is large, we can use $c = k_n = \lceil n(1 - \delta) \rceil$. If n is not large, using $c = U_n$ where U_n decreases to k_n, can improve small sample performance. U_n will be defined in the paragraph below Equation (4.23). Olive (2013a) showed that (4.19) is a large sample $100(1 - \delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist. Note that the result follows since if $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{S}$ are nonsingular, then the Mahalanobis distance is a continuous function of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Let $\boldsymbol{\mu} = E(\boldsymbol{x})$ and $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cumulative distribution function of D.

A problem with the prediction regions that cover $\approx 100(1 - \delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (4.19) for $c = k_n$), is that they have coverage lower than the nominal coverage of $1 - \delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data.

Increasing c will improve the coverage for moderate samples. Also see Remark 4.8. Empirically for many distributions, for $n \approx 20p$, the prediction region (4.19) applied to iid data using $c = k_n = \lceil n(1-\delta) \rceil$ tended to have undercoverage as high as 5%. The undercoverage decreases rapidly as n increases. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n), \quad \text{otherwise.} \tag{4.21}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Using

$$c = \lceil n q_n \rceil \tag{4.22}$$

in (4.19) decreased the undercoverage. Note that Equations (4.11) and (4.12) are similar to Equations (4.20) and (4.21), but replace p by d.

If $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_p(1).$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (at continuity points $D_{1-\delta}$ of the cdf of $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_m$.

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. The classical estimator and RMVN estimator from Section 7.1 satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \le h\} \tag{4.23}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \tag{4.24}$$

A future observation (random vector) $\boldsymbol{x}_f$ is in the region (4.22) if $D_{\boldsymbol{x}_f} \le h$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then (4.22) is a large sample $100(1-\delta)\%$ prediction region if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i where q_n is defined above (4.21). If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, then prediction region (4.24) is asymptotically optimal for a large class of elliptically contoured distributions since the volume of (4.24) converges in probability to the volume of the highest density region. (These distributions have a highest density region which is a hyperellipsoid determined by a population Mahalanobis distance. See Section 1.7.)

The Olive (2013a) nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151). Refer to the above paragraph for $D_{(U_n)}$.

Definition 4.10. The large sample $100(1-\delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \tag{4.25}$$

while the large sample $100(1-\delta)\%$ *classical prediction region* is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}. \tag{4.26}$$

If p is small, Mahalanobis distances tend to be right skewed with a population shorth that discards the right tail. For $p = 1$ and $n \geq 20$, the finite sample correction factors c/n for c given by (4.10) and (4.21) do not differ by much more than 3% for $0.01 \leq \delta \leq 0.5$. See Figure 4.2 where ol = (Eq. 4.21)$/n$ is plotted versus fr = (Eq. 4.10)$/n$ for $n = 20, 21, ..., 500$. The top plot is for $\delta = 0.01$, while the bottom plot is for $\delta = 0.3$. The identity line is added to each plot as a visual aid. The value of n increases from 20 to 500 from the right of the plot to the left of the plot. Examining the axes of each plot shows that the correction factors do not differ greatly. *R* code to create Figure 4.2 is shown below.

```
cmar <- par("mar"); par(mfrow = c(2, 1))
par(mar=c(4.0,4.0,2.0,0.5))
frey(0.01); frey(0.3)
par(mfrow = c(1, 1)); par(mar=cmar)
```

Remark 4.10. The nonparametric prediction region (4.24) is useful if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with a nonsingular covariance matrix, and the sample size n is large enough. The distribution could be continuous, discrete, or a mixture. The asymptotic coverage is $1 - \delta$ if D has a pdf, although prediction regions with smaller volume may exist. If the $100(1-\delta)$th percentile $D_{1-\delta}$ of D is not a continuity point of the distribution of D, then the asymptotic coverage tends to be $\geq 1 - \delta$ since a sample percentile with cutoff q_n that decreases to $1-\delta$ is used and a closed region is used. Often D has a continuous distribution and hence has no discontinuity points for $0 < \delta < 1$. (If there is a jump in the distribution from 0.9 to 0.96 at discontinuity point a, and the nominal coverage is 0.95, we want 0.96 coverage instead of 0.9. So we want the sample percentile to decrease to a.) The nonparametric prediction region (4.24) contains U_n of the training data cases $\boldsymbol{x}_i$ provided that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. For many distributions, the coverage started to be close to $1 - \delta$ for $n \geq 10p$ where the coverage is the simulated percentage of times that the prediction region contained $\boldsymbol{x}_f$.

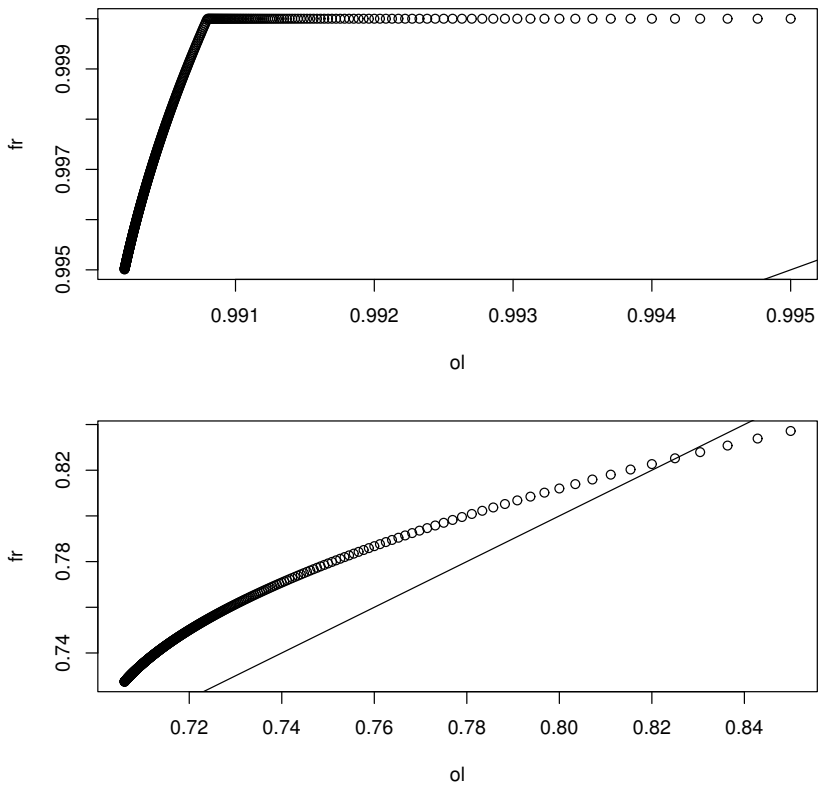

Fig. 4.2 Correction Factor Comparison when $\delta = 0.01$ (Top Plot) and $\delta = 0.3$ (Bottom Plot)

Remark 4.11. The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. Using (4.23), the ratio of the volumes of regions (4.25) and (4.24) is

$$\left(\frac{\chi^2_{p,1-\delta}}{D^2_{(U_n)}}\right)^{p/2},$$

which can become close to zero rapidly as p gets large if the $\boldsymbol{x}_i$ are not from the light tailed multivariate normal distribution. For example, suppose $\chi^2_{4,0.5} \approx 3.33$ and $D^2_{(U_n)} \approx D^2_{\boldsymbol{x},0.5} = 6$. Then the ratio is $(3.33/6)^2 \approx 0.308$. Hence if the data is not multivariate normal, severe undercoverage can occur if the classical prediction region is used, and the undercoverage tends to get worse as the dimension p increases. The coverage need not to go to 0, since by the multivariate Chebyshev's inequality, $P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \gamma) \geq 1 - p/\gamma > 0$

for $\gamma > p$ where the population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x})$. See Budny (2014), Chen (2011), and Navarro (2014, 2016). Using $\gamma = h^2 = p/\delta$ in (4.22) usually results in prediction regions with volume and coverage that is too large.

Remark 4.12. The nonparametric prediction region (4.24) starts to have good coverage for $n \geq 10p$ for a large class of distributions. Olive (2013a) suggests $n \geq 50p$ may be needed for the prediction region to have a good volume. Of course for any n there are error distributions that will have severe undercoverage.

For the multivariate lognormal distribution with $n = 20p$, the large sample nonparametric 95% prediction region (4.24) had coverages 0.970, 0.959, and 0.964 for $p = 100, 200$, and 500. Some R code is below.

```
nruns=1000 #lognormal, p = 100, n = 20p = 2000
count<-0
for(i in 1:nruns){
x <- exp(matrix(rnorm(200000),ncol=100,nrow=2000))
xff <- exp(as.vector(rnorm(100)))
count <- count + predrgn(x,xf=xff)$inr}
count #970/1000, may take a few minutes
```

Notice that for the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $c \approx 100q_n\%$ of the n cases are in the prediction regions for $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1 - \delta/2$ or $q_n = 1 - \delta + 0.05$ for $n \leq 20p$ and $q_n \to 1 - \delta$ as $n \to \infty$. If $q_n \equiv 1 - \delta$ and $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ where $d > 0$ and $\boldsymbol{\Sigma}$ is nonsingular, then (4.22) with $h = D_{(U_n)}$ is a large sample prediction region, but taking q_n given by (4.20) improves the finite sample performance of the prediction region. Taking $q_n \equiv 1 - \delta$ does not take into account variability of $(T, \boldsymbol{C})$, and for $n = 20p$ the resulting prediction region tended to have undercoverage as high as $\min(0.05, \delta/2)$. Using (4.20) helped reduce undercoverage for small $n \geq 20p$ due to the unknown variability of $(T, \boldsymbol{C})$.

4.5 Bootstrapping Hypothesis Tests and Confidence Regions

This section shows that, under regularity conditions, applying the nonparametric prediction region of Section 4.4 to a bootstrap sample results in a confidence region. The volume of a confidence region $\to 0$ as $n \to 0$, while

the volume of a prediction region goes to that of a population region that would contain a new $\boldsymbol{x}_f$ with probability $1-\delta$. The nominal coverage is $100(1-\delta)$. If the actual coverage $100(1-\delta_n) > 100(1-\delta)$, then the region is *conservative*. If $100(1-\delta_n) < 100(1-\delta)$, then the region is *liberal*. A region that is 5% conservative is considered "much better" than a region that is 5% liberal.

When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations $(2SD(\overline{Y}_n) = 2\sigma/\sqrt{n})$ of $\theta = \mu$ is about 95%. Hence the probability that θ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\theta - 1.96S/\sqrt{n}, \theta + 1.96S/\sqrt{n}]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if θ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}]$ is a large sample 95% confidence interval for the population mean θ. Note that the lengths of the two intervals are the same. Where the interval is centered, at the parameter θ or the statistic $\overline{Y}_n$, determines whether the interval is a prediction or a confidence interval. See Theorem 4.7 for a similar relationship between confidence regions and prediction regions.

Definition 4.11. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

If $\mathcal{A}_n$ is based on a squared Mahalanobis distance D^2 with a limiting distribution that has a pdf, we often want $P(\boldsymbol{\theta} \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$.

There are several methods for obtaining a bootstrap sample $T_1^*, ..., T_B^*$ where the sample size n is suppressed: $T_i^* = T_{in}^*$. The parametric bootstrap, nonparametric bootstrap, and residual bootstrap will be used. Applying prediction region (4.24) to the bootstrap sample will result in a confidence region for $\boldsymbol{\theta}$. When $g = 1$, applying the shorth PI (4.10) or PI (4.7) to the bootstrap sample results in a confidence interval for θ. Section 4.5.2 will help clarify ideas.

When $g = 1$, a confidence interval is a special case of a confidence region. One sided confidence intervals give a lower or upper confidence bound for θ. A large sample $100(1-\delta)\%$ lower confidence interval $(-\infty, U_n]$ uses an upper confidence bound U_n and is in the lower tail of the distribution of $\hat{\theta}$. A large sample $100(1-\delta)\%$ upper confidence interval $[L_n, \infty)$ uses a lower confidence bound L_n and is in the upper tail of the distribution of $\hat{\theta}$. These CIs can be useful if $\theta \in [a, b]$ and $\theta = a$ or $\theta = b$ is of interest for a hypothesis test. For example, $[a, b] = [0, 1]$ if $\theta = \rho^2$, the squared population correlation. Then use $[0, U_n]$ and $[L_n, 1]$ as CIs, e.g. if we expect $\theta = 0$ we might test $H_0 : \theta \leq 0.05$ versus $H_0 : \theta > 0.05$, and fail to reject H_0 if $U_n < 0.05$. See Section 4.5.4 for an illustration. Again we often want the probability to converge to $1-\delta$ if the confidence interval is based on a statistic with an asymptotic distribution that has a pdf.

Definition 4.12. The interval $[L_n, U_n]$ is a large sample $100(1-\delta)\%$ *confidence interval* for θ if $P(L_n \leq \theta \leq U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. The interval $(-\infty, U_n]$ is a large sample $100(1-\delta)\%$ *lower confidence interval* for θ if $P(\theta \leq U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. The interval $[L_n, \infty)$ is large sample $100(1-\delta)\%$ *upper confidence interval* for θ if $P(\theta \geq L_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

Next we discuss bootstrap confidence intervals that are obtained by applying prediction intervals (4.7) and (4.10) to the bootstrap sample. Some additional bootstrap CIs are obtained from bootstrap confidence regions from Section 4.5.2 when $g = 1$. See Efron (1982) and Chen (2016) for the percentile method CI. Let T_n be an estimator of a parameter θ such as $T_n = \overline{Z} = \sum_{i=1}^{n} Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for T_n. Let $T_{(1)}^*, ..., T_{(B)}^*$ be the order statistics of the the bootstrap sample. The CI (4.26) is obtained by applying PI (4.7) to the bootstrap sample with B used instead of n. Hence (4.26) is also a large sample prediction interval for a future value of T_f^* if the T_i^* are iid from the empirical distribution discussed in Section 4.5.1.

Definition 4.13. The bootstrap percentile method large sample $100(1-\delta)\%$ confidence interval for θ is an interval $[T_{(k_L)}^*, T_{(K_U)}^*]$ containing $\approx \lceil B(1-\delta) \rceil$ of the T_i^*. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is

$$[T_{(k_1)}^*, T_{(k_2)}^*]. \tag{4.27}$$

The large sample $100(1-\delta)\%$ *lower percentile method* CI for θ is $(-\infty, T_{(\lceil B(1-\delta) \rceil)}^*]$. The large sample $100(1-\delta)\%$ *upper percentile method* CI for θ is $[T_{(\lceil B\delta \rceil)}^*, \infty)$.

Definition 4.14. The large sample $100(1-\delta)\%$ *lower shorth* CI for θ is $(-\infty, T_{(c)}^*]$, while the large sample $100(1-\delta)\%$ *upper shorth* CI for θ is $[T_{(B-c+1)}^*, \infty)$. The large sample $100(1-\delta)\%$ *shorth*(c) *CI* uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length. Here

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{4.28}$$

Applied to a bootstrap sample, the Frey shorth interval can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. See Remark 4.16 for some theory for bootstrap CIs such as (4.26) and (4.27).

4.5.1 The Bootstrap

This subsection illustrates the nonparametric bootstrap with some examples. Suppose a statistic T_n is computed from a data set of n cases. The nonparametric bootstrap draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

Definition 4.15. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 4.2. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 4.15. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \frac{n-1}{n}\boldsymbol{S} \equiv \frac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \quad \square$$

Example 4.3. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *nonparametric bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 4.4. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic $\mathrm{Cov}(T_n)$, for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T_{1n}^*, ..., T_{Bn}^*$ is formed where $T_{jn}^* = t(\boldsymbol{w}_{j1}^*, ..., \boldsymbol{w}_{jn}^*)$. Section 4.5.3 will show that $T_{1n}^* - T_n, ..., T_{Bn}^* - T_n$ is pseudodata for $T_{1n} - \boldsymbol{\theta}, ..., T_{Bn} - \boldsymbol{\theta}$ when n is large in that $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T^* - T_n) \xrightarrow{D} \boldsymbol{u}$.

Example 4.5. Suppose there is training data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ for the model $\boldsymbol{y}_i = m(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$, and it is desired to predict a future test value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and the training data. The model can be fit and the residual vectors formed. One method for obtaining a prediction region for $\boldsymbol{y}_f$ is to form the pseudodata $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$, and apply the nonparametric

prediction region (4.24) to the pseudodata. See Section 8.3 and Olive (2017b, 2018). The residual bootstrap could also be used to make a bootstrap sample $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_1^*, ..., \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_B^*$ where the $\hat{\boldsymbol{\epsilon}}_j^*$ are selected with replacement from the residual vectors for $j = 1, ..., B$. As $B \to \infty$, the bootstrap sample will take on the n values $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ (the pseudodata) with probabilities converging to $1/n$ for $i = 1, ..., n$.

Suppose there is a statistic T_n that is a $g \times 1$ vector. Let

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* \quad \text{and} \quad \boldsymbol{S}_T^* = \frac{1}{B-1}\sum_{i=1}^{B}(T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \tag{4.29}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$ where $T_i^* = T_{i,n}^*$. Fix n, and let $E(T_{i,n}^*) = \boldsymbol{\theta}_n$ and $\mathrm{Cov}(T_{i,n}^*) = \boldsymbol{\Sigma}_n$.

We will often assume that $\mathrm{Cov}(T_n) = \boldsymbol{\Sigma}_T$, and $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ where $\boldsymbol{\Sigma}_A > 0$ is positive definite and nonsingular. Often $n\hat{\boldsymbol{\Sigma}}_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. For example, using least squares and the residual bootstrap for the multiple linear regression model, $\boldsymbol{\Sigma}_n = \dfrac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, $T_n = \boldsymbol{\theta}_n = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, $\hat{\boldsymbol{\Sigma}}_T = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\boldsymbol{\Sigma}_A = \sigma^2 \lim_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1}$. See Example 4.6 in Section 4.6.

Suppose the $T_i^* = T_{i,n}^*$ are iid from some distribution with cdf $\tilde{F}_n$. For example, if $T_{i,n}^* = t(F_n^*)$ where iid samples from F_n are used, then $\tilde{F}_n$ is the cdf of $t(F_n^*)$. With respect to $\tilde{F}_n$, both $\boldsymbol{\theta}_n$ and $\boldsymbol{\Sigma}_n$ are parameters, but with respect to F, $\boldsymbol{\theta}_n$ is a random vector and $\boldsymbol{\Sigma}_n$ is a random matrix. For fixed n, by the multivariate central limit theorem,

$$\sqrt{B}(\overline{T}^* - \boldsymbol{\theta}_n) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_n) \quad \text{and} \quad \mathrm{B}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n})^\mathrm{T}[\boldsymbol{S}_\mathrm{T}^*]^{-1}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n}) \xrightarrow{\mathrm{D}} \chi^2_\mathrm{r}$$

as $B \to \infty$.

Remark 4.13. For Examples 4.2, 4.5, and 4.6, the bootstrap works but is expensive compared to alternative methods. For Example 4.2, fix n, then $\overline{T}^* \xrightarrow{P} \boldsymbol{\theta}_n = \overline{\boldsymbol{x}}$ and $\boldsymbol{S}_T^* \xrightarrow{P} (n-1)\boldsymbol{S}/n$ as $B \to \infty$, but using $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ makes more sense. For Example 4.5, use the pseudodata instead of the residual bootstrap. For Example 4.6, using $\hat{\boldsymbol{\beta}}$ and the classical estimated covariance matrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ makes more sense than using the bootstrap. For these three examples, it is known how the bootstrap sample behaves as $B \to \infty$. The bootstrap can be very useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, but it not known how to estimate $\boldsymbol{\Sigma}_A$ without using a resampling method like the bootstrap. The bootstrap may be useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, but the limiting distribution (the distribution of $\boldsymbol{u}$) is unknown.

4.5.2 Bootstrap Confidence Regions for Hypothesis Testing

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ is a set $\mathcal{A}_n = \mathcal{A}_{n,B}$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_{n,B})$ is eventually bounded below by $1-\delta$ as $n, B \to \infty$. The B is often suppressed. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\boldsymbol{A}$ be a full rank $g \times p$ constant matrix. For variable selection, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The statistic $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is the variable selection estimator padded with zeroes. See Section 4.2. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$. See Equation (4.28). See Theorem 2.25 for why $d_n F_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \leq \chi^2_{g,1-\delta}) = 1-\delta$ if $X \sim \chi^2_g$, and $P(X \leq F_{g,d_n,1-\delta}) = 1-\delta$ if $X \sim F_{g,d_n}$. Let $k_B = \lceil B(1-\delta) \rceil$.

Definition 4.16. a) The standard bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{4.30}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$. b) The Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(k_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_A/n) \leq D^2_{(k_B,T)}\} \tag{4.31}$$

where the cutoff $D^2_{(k_B,T)}$ is the $100k_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (T_i^* - T_n) = n(T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A]^{-1} (T_i^* - T_n)$.

Confidence region (4.29) needs $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$ and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A > 0$ as $n, B \to \infty$. See Machado and Parente (2005) for regularity conditions for this assumption. Bickel and Ren (2001) have interesting sufficient conditions for (4.30) to be a confidence region when $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of positive definite $\boldsymbol{\Sigma}_A$. Let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, and the bootstrapped statistic $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and

$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from $F_n =$ "F." Thus if n is fixed, and a sample of size m is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{u}$.

The following three confidence regions will be used for inference after variable selection. The Olive (2017ab, 2018) prediction region method applies prediction region (4.24) to the bootstrap sample. Olive (2017ab, 2018) also gave the modified Bickel and Ren confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$. The hybrid confidence region is due to Pelawa Watagoda and Olive (2019a). Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{4.32}$$

If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. Let $D_{(U_B)}$ be the $100q_B$th sample quantile of the D_i. Use (4.31) as a correction factor for finite $B \geq 50p$.

Definition 4.17. a) The prediction region method large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - \overline{T}^*) \leq D_{(U_B)}^2\} =$

$$\{\boldsymbol{w} : D_{\boldsymbol{w}}^2(\overline{T}^*, \boldsymbol{S}_T^*) \leq D_{(U_B)}^2\} \tag{4.33}$$

where $D_{(U_B)}^2$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D_{(U_B)}^2$. (This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi_{g,1-\delta}^2$ cutoff by $D_{(U_B)}^2$.) b) The modified Bickel and Ren (2001) large sample $100(1 - \delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D_{(U_B,T)}^2\} =$

$$\{\boldsymbol{w} : D_{\boldsymbol{w}}^2(T_n, \boldsymbol{S}_T^*) \leq D_{(U_B,T)}^2\} \tag{4.34}$$

where the cutoff $D_{(U_B,T)}^2$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D_{(U_B,T)}^2$. c) Shift region (4.32) to have center T_n, or equivalently, change the cutoff of region (4.33) to $D_{(U_B)}^2$ to get the hybrid large sample $100(1 - \delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D_{(U_B)}^2\} =$

$$\{\boldsymbol{w} : D_{\boldsymbol{w}}^2(T_n, \boldsymbol{S}_T^*) \leq D_{(U_B)}^2\}. \tag{4.35}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D_{(U_B)}^2$.

Hyperellipsoids (4.32) and (4.34) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (4.32) and (4.33) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g. \tag{4.36}$$

The volume of confidence region (4.33) tends to be greater than that of (4.32) since the T_i^* are closer to $\overline{T}^*$ than T_n on average.

If $g = 1$, then a hyperellipsoid is an interval, and confidence intervals are special cases of confidence regions. Suppose the parameter of interest is θ, and there is a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic T_n is an estimator of θ based on a sample of size n. The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^*. Let $a_i = |T_i^* - \overline{T}^*|$. Let $\overline{T}^*$ and S_T^{2*} be the sample mean and variance of the T_i^*. Then the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T}^*)^2 / S_T^{*2} \leq D_{(U_B)}^2$ is equivalent to $\theta \in [\overline{T}^* - S_T^* D_{(U_B)}, \overline{T}^* + S_T^* D_{(U_B)}] = [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*. Hence the prediction region method is a special case of the percentile method if $g = 1$. See Definition 4.13. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The CI corresponding to (4.33) is defined similarly, and $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$ is the CI for (4.34). Note that the three CIs corresponding to (4.32)–(4.34) can be computed without finding S_T^* or $D_{(U_B)}$ even if $S_T^* = 0$. The Frey (2013) shorth(c) CI (4.27) computed from the T_i^* can be much shorter than the Efron (2014) or prediction region method confidence intervals. See Remark 4.16 for some theory for bootstrap CIs.

Remark 4.14. From Example 4.6, $\text{Cov}(\hat{\boldsymbol{\beta}}^*) = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n} \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ where $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ starts to give good estimates of $\text{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_T$ for many error distributions if $n \geq 10p$ and $T = \hat{\boldsymbol{\beta}}$. For the residual bootstrap with large B, note that $\boldsymbol{S}_T^* \approx 0.95 \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 20p$ and $\boldsymbol{S}_T^* \approx 0.99 \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 100p$. Hence we may need $n >> p$ before the $\boldsymbol{S}_T^*$ is a good estimator of $\text{Cov}(T) = \boldsymbol{\Sigma}_T$. The distribution of $\sqrt{n}(T_n - \boldsymbol{\theta})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T}^*)$, but n may need to be large before the approximation is good.

Suppose the bootstrap sample mean $\overline{T}^*$ estimates $\boldsymbol{\theta}$, and the bootstrap sample covariance matrix $\boldsymbol{S}_T^*$ estimates $c_n \widehat{\text{Cov}}(T_n) \approx c_n \boldsymbol{\Sigma}_T$ where c_n increases to 1 as $n \to \infty$. Then $\boldsymbol{S}_T^*$ is not a good estimator of $\widehat{\text{Cov}}(T_n)$ until $c_n \approx 1$ ($n \geq 100p$ for OLS $\hat{\boldsymbol{\beta}}$), but the squared Mahalanobis distance $D_{\boldsymbol{w}}^{2*}(\overline{T}^*, \boldsymbol{S}_T^*) \approx D_{\boldsymbol{w}}^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T)/c_n$ and $D_{(U_B)}^{2*} \approx D_{1-\delta}^2/c_n$. Hence the prediction region method has a cutoff $D_{(U_B)}^{2*}$ that estimates the cutoff $D_{1-\delta}^2/c_n$. Thus the prediction region method may give good results for much smaller n than

a bootstrap method that uses a $\chi^2_{g,1-\delta}$ cutoff when a cutoff $\chi^2_{g,1-\delta}/c_n$ should be used for moderate n.

Remark 4.15. For bootstrapping the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, we will often want $n \geq 20p$ and $B \geq \max(100, n, 50p)$. If T_n is $g \times 1$, we might replace p by g or replace p by d if d is the model degrees of freedom. Sometimes much larger n is needed to avoid undercoverage. We want $B \geq 50g$ so that $\boldsymbol{S}_T^*$ is a good estimator of $Cov(T_n^*)$. Prediction region theory uses correction factors like (4.21) and (4.10) to compensate for finite n. The bootstrap confidence regions (4.32)–(4.34) and the shorth CI use the correction factors (4.31) and (4.27) to compensate for finite $B \geq 50g$. Note that the correction factors make the volume of the confidence region larger as B decreases. Hence a test with larger B will have more power.

4.5.3 Theory for Bootstrap Confidence Regions

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. This section gives some theory for bootstrap confidence regions and for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Breiman (1996), Yang (2003), and Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator. Since (4.33) is a large sample confidence region by Bickel and Ren (2001), (4.32) and (4.34) are too, provided $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$.

If i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, then under regularity conditions, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} \mathrm{Cov}(\boldsymbol{u})$.

Suppose i) and ii) hold with $E(\boldsymbol{u}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Let $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{u}$ where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. Fix B. Then the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right)$$

where $\boldsymbol{z} \sim AN_g(\mathbf{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$, and iii) and iv) hold. If B is fixed and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, then

$$\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim N_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right) \text{ and } \sqrt{\mathrm{B}}\sqrt{\mathrm{n}}(\overline{\mathrm{T}}^* - \mathrm{T_n}) \xrightarrow{\mathrm{D}} \mathrm{N_g}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}).$$

Hence the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$. Since iii) and iv) hold, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma_u}$ is nonsingular. Olive (2017b: § 5.3.3, 2018) proved that the prediction region method gives a large sample confidence region under the much stronger conditions of v) and $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma_u})$, but the above Pelawa Watagoda and Olive (2019a) proof is simpler.

Remark 4.16. Note that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where U has a unimodal probability density function symmetric about zero, then the confidence intervals from the three confidence regions (4.32)–(4.34), the shorth confidence interval (4.27), and the "usual" percentile method confidence interval (4.26) are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically).

Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}_T^*$ are nonsingular $g \times g$ matrices, and T_n is an estimator of $\boldsymbol{\theta}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{4.37}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2}\boldsymbol{u} = \boldsymbol{z},$$

$$n\ (T_n - \boldsymbol{\theta})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{z}^T\boldsymbol{z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\theta})^T\ [\boldsymbol{S}_T^*]^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} D^2 \tag{4.38}$$

as $n, B \to \infty$. Assume the cumulative distribution function of D^2 is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \leq D^2_{1-\delta}) = 1-\delta$. If the distribution of D^2 is known, then we could use the large sample confidence region (4.29) $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\}$. Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_g$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$.

Remark 4.17. Under reasonable conditions, i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Then

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D_{\boldsymbol{\theta}}^2(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D_{\boldsymbol{\theta}}^2(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D_{T_i^*}^2(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - T_n)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - T_n),$$

are well behaved. If $(n\boldsymbol{S}_T^*)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_T^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T \boldsymbol{\Sigma}_T^{-1} \boldsymbol{u}$. If $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned" then $D_j^2 \approx \boldsymbol{u}^T (n\boldsymbol{S}_T^*)^{-1} \boldsymbol{u}$ for large n, and the confidence regions (4.32), (4.33), and (4.34) will have coverage near $1 - \delta$. The regularity conditions for (4.32)–(4.34) are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

The following Pelawa Watagoda and Olive (2019a) theorem is very useful. Let $D_{(U_B)}^2$ be the cutoff for the nonparametric prediction region (4.24) computed from the $D_i^2(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. Hence n is replaced by B. Since T_n depends on the sample size n, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved ("not too ill conditioned") for each $n \geq 20g$, say. This condition is weaker than $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Note that $T_i = T_{in}$.

Theorem 4.7: Geometric Argument. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D_{\boldsymbol{w}}^2(\overline{T}, \boldsymbol{S}_T) \leq D_{(U_B)}^2\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D_{\boldsymbol{w}}^2(T_n, \boldsymbol{S}_T) \leq D_{(U_B)}^2\}$ is a large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ where T_n is a randomly selected T_i.

Proof. The region R_c centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 1.30 and 1.31, and see Example 1.16.) For fixed B, the average of these random vectors is

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B} \sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right)$$

by Theorem 1.33. Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Thus R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. □

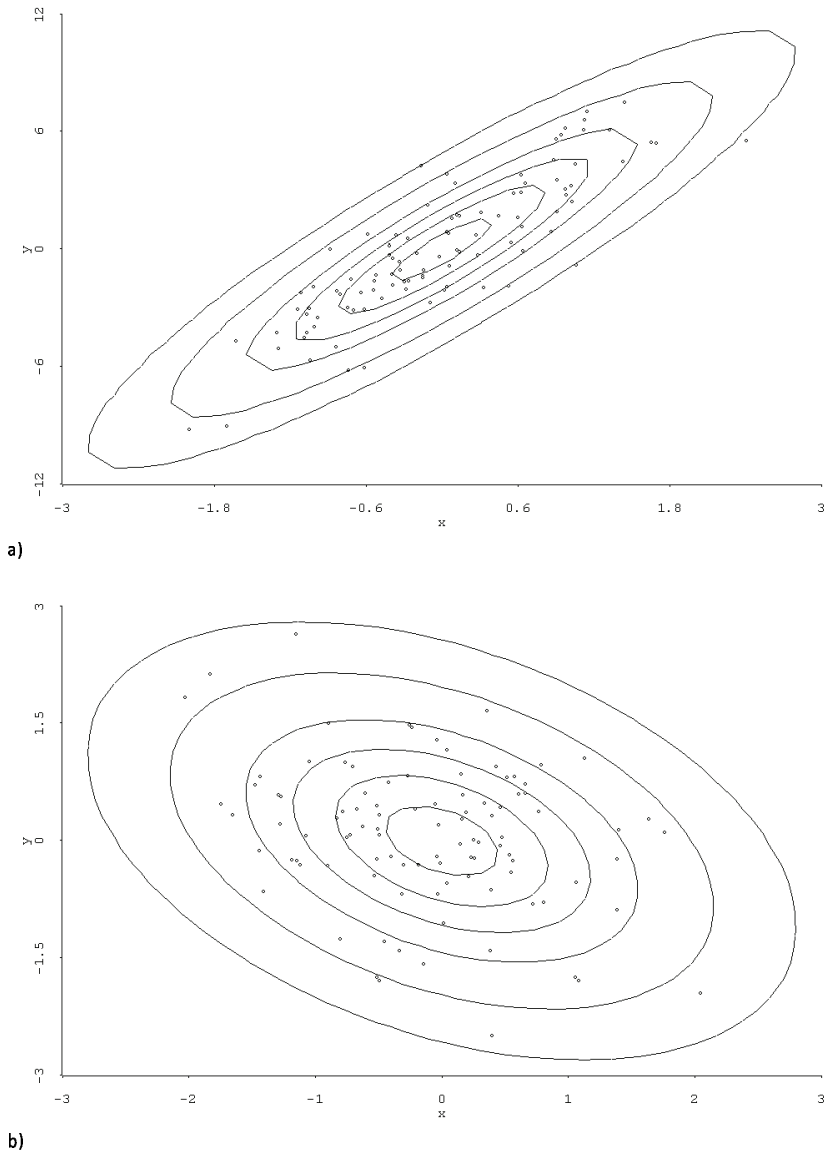

Fig. 4.3 Confidence Regions for 2 Statistics with MVN Distributions

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u}$, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at T_n. The nonparametric confidence region (4.32) applies the prediction region to the bootstrap. Then the hybrid region (4.34) centers that region at T_n. Hence (4.34) is a confidence region by the geometric argument, and (4.32) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$. Since the T_i^* are closer to $\overline{T}^*$ than T_n on average, $D^2_{(U_B,T)}$ tends to be greater than $D^2_{(U_B)}$. Hence the coverage and volume of (4.33) tend to be at least as large as the coverage and volume of (4.34).

The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at T_n, while the hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T \boldsymbol{C}^{-1} (\overline{T} - T_n) = (T_n - \overline{T})^T \boldsymbol{C}^{-1} (T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \leq D^2_{(U_B)}$.

The prediction region method will often simulate well even if B is rather small. If the ellipses are centered at T_n or $\overline{T}^*$, Figure 4.3 shows confidence regions if the plotted points are $T_1^*, ..., T_B^*$ where the T_i^* are approximately multivariate normal. If the ellipses are centered at $\overline{T}$, Figure 4.3 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. Then the plotted points are iid $T_1, ..., T_B$. If $nCov(T) \xrightarrow{P} \boldsymbol{\Sigma}_A$, and the T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. By Theorem 4.7, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths of this covering region are about one tenth of those in Figure 4.3. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence B does not need to be large provided that n and B are large enough so that $S_T^* \approx \text{Cov}(T^*) \approx \boldsymbol{\Sigma}_A/n$. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jg$ where $J = 20$ or 50. For small g, using $B = 1000$ often led to good simulations, but $B = \max(50g, 100)$ may work well.

Remark 4.18. Remark 4.14 suggests that even if the statistic T_n is asymptotically normal so the Mahalanobis distances are asymptotically χ^2_g, the pre-

diction region method can give better results for moderate n by using the cutoff $D^2_{(U_B)}$ instead of the cutoff $\chi^2_{g,1-\delta}$. Theorem 4.7 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when n is moderate by using $D^2_{(U_n)}$. If n is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when B is moderate, say $B \geq Jg$ where $J = 20$ or 50. See Remark 4.15. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jg$. The price to pay is that the prediction region method confidence region is inflated to have better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B.

4.5.4 Bootstrapping the Population Coefficient of Multiple Determination

This subsection illustrates a case where the shorth(c) bootstrap CI fails, but the lower shorth CI can be useful. See Definition 4.14.

The multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$$

for $i = 1, ..., n$. See Definition 1.17 for the *coefficient of multiple determination*

$$R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSTO}}$$

where $\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})$ is the sample correlation of Y_i and $\hat{Y}_i$.

Assume that the variance of the errors is σ_e^2 and that the variance of Y is σ_Y^2. Let the linear combination $L = \sum_{i=2}^p x_i\beta_i$ where $Y = \beta_1 + \sum_{i=2}^p x_i\beta_i + e = \beta_1 + L + e$. Let the variance of L be σ_L^2. Then

$$R^2 = 1 - \frac{\sum_{i=1}^n r_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \xrightarrow{P} \tau^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2} = 1 - \frac{\sigma_e^2}{\sigma_e^2 + \sigma_L^2}.$$

Here we assume that e is independent of the predictors $x_2, ..., x_p$. Hence e is independent of L and the variance $\sigma_Y^2 = V(L+e) = V(L) + V(e) = \sigma_L^2 + \sigma_e^2$.

One of the sufficient conditions for the shorth(c) interval to be a large sample CI for θ is $\sqrt{n}(T - \theta) \xrightarrow{D} N(0, \sigma^2)$. If the function $t(\theta)$ has an inverse, and $\sqrt{n}(t(T) - t(\theta)) \xrightarrow{D} N(0, v^2)$, then the above condition typically holds by the delta method. See Remark 4.16.

For $T = R^2$ and $\theta = \tau^2$, the test statistic F_0 for testing $H_0 : \beta_2 = \cdots = \beta_p = 0$ in the Anova F test has $(p-1)F_0 \xrightarrow{D} \chi^2_{p-1}$ for a large class of error distributions when H_0 is true, where

$$F_0 = \frac{R^2}{1-R^2} \quad \frac{n-p}{p-1}$$

if the MLR model has a constant. If H_0 is false, then F_0 has an asymptotic scaled noncentral χ^2 distribution. These results suggest that the large sample distribution of $\sqrt{n}(R^2 - \tau^2)$ may not be $N(0, \sigma^2)$ if H_0 is false so $\tau^2 > 0$. If $\tau^2 = 0$, we may have $\sqrt{n}(R^2 - 0) \xrightarrow{D} N(0, 0)$, the point mass at 0. Hence the shorth CI may not be a large sample CI for τ^2. The lower shorth CI should be useful for testing $H_0 : \tau^2 = 0$ versus $H_A : \tau^2 > a$ where $0 < a \leq 1$ since the coverage is 1 and the length of the CI converges to 0. So reject H_0 if a is not in the CI.

The simulation simulated iid data $\boldsymbol{w}$ with $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{w}$ and $\boldsymbol{A}_{ij} = \psi$ for $i \neq j$ and $\boldsymbol{A}_{ii} = 1$ where $0 \leq \psi < 1$ and $\boldsymbol{u} = (x_2, ..., x_p)^T$. Hence $\text{Cor}(x_i, x_j) = \rho = [2\psi + (p-3)\psi^2]/[1 + (p-2)\psi^2]$ for $i \neq j$. If $\psi = 1/\sqrt{kp}$, then $\rho \to 1/(k+1)$ as $p \to \infty$ where $k > 0$. We used $\boldsymbol{w} \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1})$. If ψ is high or if p is large with $\psi \geq 0.5$, then the data are clustered tightly about the line with direction $\boldsymbol{1} = (1, ..., 1)^T$, and there is a dominant principal component with eigenvector $\boldsymbol{1}$ and eigenvalue λ_1. We used $\psi = 0, 1/\sqrt{p}$, and 0.9. Then $\rho = 0, \rho \to 0.5$, or $\rho \to 1$ as $p \to \infty$.

We also used $V(x_2) = \cdots = V(x_p) = \sigma_x^2$. If $p > 2$, then $Cov(x_i, x_j) = \rho\sigma_x^2$ for $i \neq j$ and $Cov(x_i, x_j) = V(x_i) = \sigma_x^2$ for $i = j$. Then $V(Y) = \sigma_Y^2 = \sigma_L^2 + \sigma_e^2$ where

$$\sigma_L^2 = V(L) = V(\sum_{i=2}^{p} \beta_i x_i) = Cov(\sum_{i=2}^{p} \beta_i x_i, \sum_{j=2}^{p} \beta_j x_j) = \sum_{i=2}^{p}\sum_{j=2}^{p} \beta_i \beta_j Cov(x_i, x_j)$$

$$= \sum_{i=2}^{p} \beta_i^2 \sigma_x^2 + 2\rho\sigma_x^2 \sum_{i=2}^{p} \sum_{j=i+1}^{p} \beta_i \beta_j.$$

The simulations took $\beta_i \equiv 0$ or $\beta_i \equiv 1$ for $i = 2, ..., p$. For the latter case,

$$\sigma_L^2 = V(L) = (p-1)\sigma_x^2 + 2\rho\sigma_x^2 p(p-1)/2.$$

The zero mean errors e_i were from 5 distributions: i) N(0,1), ii) t_3, iii) $EXP(1) - 1$, iv) uniform$(-1, 1)$, and v) $(1-\epsilon)N(0,1) + \epsilon N(0, (1+s)^2)$ with $\epsilon = 0.1$ and $s = 9$ in the simulation. Then $Y = 1 + bx_2 + bx_3 + \cdots + bx_p + e$ with $b = 0$ or $b = 1$.

Remark 4.19. Suppose the simulation uses K runs and $W_i = 1$ if μ is in the ith CI, and $W_i = 0$ otherwise, for $i = 1, ..., K$. Then the W_i are iid binomial(1,$1 - \delta_n$) where $\rho_n = 1 - \delta_n$ is the true coverage of the CI when the sample size is n. Let $\hat{\rho}_n = \overline{W}$. Since $\sum_{i=1}^{K} W_i \sim$ binomial(K, ρ_n), the standard error $SE(\overline{W}) = \sqrt{\rho_n(1-\rho_n)/K}$. For $K = 5000$ and ρ_n near 0.9, we have $3SE(\overline{W}) \approx 0.01$. Hence an observed coverage of $\hat{\rho}_n$ within 0.01 of the nominal coverage $1 - \delta$ suggests that there is no reason to doubt that the nominal CI coverage is different from the observed coverage. So for a large sample

95% CI, we want the observed coverage to be between 0.94 and 0.96. Also a difference of 0.01 is not large. Coverage slightly higher than the nominal coverage is better than coverage slightly lower than the nominal coverage.

Bootstrapping confidence intervals for quantities like ρ^2 and τ^2 is notoriously difficult. If $\beta_2 = \cdots = \beta_p = 0$, then $\sigma_L^2 = 0$ and $\tau^2 = 0$. However, the probability that $R_i^{2*} > 0 = 1$. Hence the usual two sided bootstrap percentile and shorth intervals for τ^2 will never contain 0. The one sided bootstrap CI $[0, T^*_{(c)}]$ always contains 0, and is useful if the length of the CI goes to 0 as $n \to \infty$. In the table below, $\beta_i = b$ for i = 2, ..., p. If $b = 0$, then $\tau^2 = 0$.

The simulation for the table used 5000 runs with the bootstrap sample size $B = 1000$. When $n = 400$, the shorth(c) CI never contains $\tau^2 = 0$ and the average length of the CI is 0.035. See *ccov* and *clen*. The lower shorth CI always contained $\tau^2 = 0$ with $lcov = 1$, and the average CI length was $llen = 0.036$. The upper shorth CI never contains $\tau^2 = 0$, and the average length is near 1.

Table 4.1 Bootstrapping τ^2 with R^2 and $B = 1000$

etype	n	p	b	ψ	τ^2	ccov	clen	lcov	llen	ucov	ulen
1	100	4	0	0	0	0	0.135	1	0.137	0	0.990
1	200	4	0	0	0	0	0.0693	1	0.0702	0	0.995
1	400	4	0	0	0	0	0.0354	1	0.0358	0	0.988

Three *linmodpack* functions were used in the simulation. The function `shorthLU` gets the shorth(c) CI, the lower shorth CI, and the upper shorth CI. The function `Rsqboot` bootstraps R^2, while the function `Rsqbootsim` does the simulation. Some R code for the first line of Table 4.1 is below where $b = cc$.

```
Rsqbootsim(n=100,p=4,BB=1000,nruns=5000,type=1,psi=0,
cc=0)
$rho
[1] 0
$sigesq
[1] 1
$sigLsq
[1] 0
$poprsq
[1] 0
$cicov
[1] 0
$avelen
[1] 0.1348881
$lcicov
```

```
[1] 1
$lavelen
[1] 0.13688
$ucicov
[1] 0
$uavelen
[1] 0.9896608
```

4.6 Bootstrapping Variable Selection

This section considers bootstrapping the MLR variable selection model. Rathnayake and Olive (2020) shows how to bootstrap variable selection for many other regression models. This section will explain why the bootstrap confidence regions (4.32), (4.33), and (4.34) give useful results. Much of the theory in Section 4.5.3 does not apply to the variable selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, because T_n is not smooth since T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$.

Obtaining the bootstrap samples for $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ is simple. Generate $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ that would be used to produce $\hat{\boldsymbol{\beta}}^*$ if the full model estimator $\hat{\boldsymbol{\beta}}$ was being bootstrapped. Instead of computing $\hat{\boldsymbol{\beta}}^*$, compute the variable selection estimator $\hat{\boldsymbol{\beta}}^*_{VS,1} = \hat{\boldsymbol{\beta}}^{*C}_{I_{k_1},0}$. Then generate another $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ and compute $\hat{\boldsymbol{\beta}}^*_{MIX,1} = \hat{\boldsymbol{\beta}}^*_{I_{k_1},0}$ (using the same subset I_{k_1}). This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$. Let the selection probabilities for the bootstrap variable selection estimator be ρ_{kn}. Then this bootstrap procedure bootstraps both $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ with $\pi_{kn} = \rho_{kn}$.

The key idea is to show that the bootstrap data cloud is slightly more variable than the iid data cloud, so confidence region (4.33) applied to the bootstrap data cloud has coverage bounded below by $(1-\delta)$ for large enough n and B.

For the bootstrap, suppose that T^*_i is equal to T^*_{ij} with probability ρ_{jn} for $j = 1, ..., J$ where $\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. Let B_{jn} count the number of times $T^*_i = T^*_{ij}$ in the bootstrap sample. Then the bootstrap sample $T^*_1, ..., T^*_B$ can be written as

$$T^*_{1,1}, ..., T^*_{B_{1n},1}, ..., T^*_{1,J}, ..., T^*_{B_{Jn},J}$$

where the B_{jn} follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. Denote $T^*_{1j}, ..., T^*_{B_{jn},j}$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}^*_j$ and sample covariance matrix $\boldsymbol{S}^*_{T,j}$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T_{ij}^* = \sum_j \hat{\rho}_{jn}\overline{T}_j^*.$$

Similarly, we can define the jth component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$.

Let $T_n = \hat{\boldsymbol{\beta}}_{MIX}$ and $T_{ij} = \hat{\boldsymbol{\beta}}_{I_j,0}$. If $S \subseteq I_j$, assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j}^* - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$. Then by Equation (4.3),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \text{ and } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\beta}}_{\mathrm{I_j},0}^* - \hat{\boldsymbol{\beta}}_{\mathrm{I_j},0}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{V}_{\mathrm{j},0}). \quad (4.39)$$

This result means that the component clouds have the same variability asymptotically. The iid data component clouds are all centered at $\boldsymbol{\beta}$. If the bootstrap data component clouds were all centered at the same value $\tilde{\boldsymbol{\beta}}$, then the bootstrap cloud would be like an iid data cloud shifted to be centered at $\tilde{\boldsymbol{\beta}}$, and (4.33) would be a confidence region for $\boldsymbol{\theta} = \boldsymbol{\beta}$. Instead, the bootstrap data component clouds are shifted slightly from a common center, and are each centered at a $\hat{\boldsymbol{\beta}}_{I_j,0}$. Geometrically, the shifting of the bootstrap component data clouds makes the bootstrap data cloud similar but more variable than the iid data cloud asymptotically (we want $n \geq 20p$), and centering the bootstrap data cloud at T_n results in the confidence region (4.33) having slightly higher asymptotic coverage than applying (4.33) to the iid data cloud. Also, (4.33) tends to have higher coverage than (4.34) since the cutoff for (4.33) tends to be larger than the cutoff for (4.34). Region (4.32) has the same volume as region (4.34), but tends to have higher coverage since empirically, the bagging estimator $\overline{T}^*$ tends to estimate $\boldsymbol{\theta}$ at least as well as T_n for a mixture distribution. A similar argument holds if $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$, $T_{ij} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$.

To see that T^* has more variability than T_n, asymptotically, look at Figure 4.3. Imagine that n is huge and the $J = 6$ ellipsoids are 99.9% covering regions for the component data clouds corresponding to T_{jn} for $j = 1, ..., J$. Separating the clouds slightly, without rotation, increases the variability of the overall data cloud. The bootstrap distribution of T^* corresponds to the separated clouds. The shape of the overall data cloud does not change much, but the volume does increase.

In the simulations for $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta}_S = \boldsymbol{\theta}_0$ with $n \geq 20p$, the coverage tended to get close to $1 - \delta$ for $B \geq \max(200, 50p)$ so that $\boldsymbol{S}_T^*$ is a good estimator of $\mathrm{Cov}(T^*)$. In the simulations where S is not the full model, inference with backward elimination with I_{min} using AIC was often more precise than inference with the full model if $n \geq 20p$ and $B \geq 50p$.

The matrix $\boldsymbol{S}_T^*$ can be singular due to one or more columns of zeros in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model. A simple remedy is to add d bootstrap samples of the

full model estimator $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}^*_{FULL}$ to the bootstrap sample. For example, take $d = \lceil cB \rceil$ with $c = 0.01$. A confidence interval $[L_n, U_n]$ can be computed without $\boldsymbol{S}^*_T$ for (4.32), (4.33), and (4.34). Using the confidence interval $[\max(L_n, T^*_{(1)}), \min(U_n, T^*_{(B)})]$ can give a shorter covering region.

Undercoverage can occur if bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

The bootstrap component clouds for $\hat{\boldsymbol{\beta}}^*_{VS}$ are again separated compared to the iid clouds for $\hat{\boldsymbol{\beta}}_{VS}$, which are centered about $\boldsymbol{\beta}$. Heuristically, most of the selection bias is due to predictors in E, not to the predictors in S. Hence $\hat{\boldsymbol{\beta}}^*_{S,VS}$ is roughly similar to $\hat{\boldsymbol{\beta}}^*_{S,MIX}$. Typically the distributions of $\hat{\boldsymbol{\beta}}^*_{E,VS}$ and $\hat{\boldsymbol{\beta}}^*_{E,MIX}$ are not similar, but use the same zero padding. In simulations, confidence regions for $\hat{\boldsymbol{\beta}}_{VS}$ tended to have less undercoverage than confidence regions for $\hat{\boldsymbol{\beta}}^*_{MIX}$.

4.6.1 The Parametric Bootstrap

The parametric bootstrap generates $\boldsymbol{Y}^*_j = (Y^*_i)$ from a parametric distribution. Then regress $\boldsymbol{Y}^*_j$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}^*_j$ for $j = 1, ..., B$. Consider the parametric bootstrap for the MLR model with $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_n \boldsymbol{I}) \sim N_n(\boldsymbol{HY}, \hat{\sigma}^2_n \boldsymbol{I})$ where **we are not assuming** that the $e_i \sim N(0, \sigma^2)$, and

$$\hat{\sigma}^2_n = MSE = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2$$

where the residuals are from the full OLS model. Then MSE is a $\sqrt{n}$ consistent estimator of σ^2 under mild conditions by Su and Cook (2012). Hence

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{e}^*$$

where the e^*_i are iid $N(0, MSE)$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$.

Thus $\hat{\boldsymbol{\beta}}^*_I = (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \boldsymbol{Y}^* \sim N_{a_I}(\hat{\boldsymbol{\beta}}_I, \hat{\sigma}^2_n (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1})$ since $E(\hat{\boldsymbol{\beta}}^*_I) = (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \boldsymbol{HY} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{H}\boldsymbol{X}_I = \boldsymbol{X}_I$, and $\text{Cov}(\hat{\boldsymbol{\beta}}^*_I) = \hat{\sigma}^2_n (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \sim N_{a_I}(\boldsymbol{0}, n\hat{\sigma}^2_n (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$ if $S \subseteq I$.

4.6.2 The Residual Bootstrap

The *residual bootstrap* is often useful for additive error regression models of the form $Y_i = m(\boldsymbol{x}_i) + e_i = \hat{m}(\boldsymbol{x}_i) + r_i = \hat{Y}_i + r_i$ for $i = 1, ..., n$ where the ith residual $r_i = Y_i - \hat{Y}_i$. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{r} = (r_1, ..., r_n)^T$, and let $\boldsymbol{X}$ be an $n \times p$ matrix with ith row $\boldsymbol{x}_i^T$. Then the fitted values $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$, and the residuals are obtained by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here the errors e_i are iid, and it would be useful to be able to generate B iid samples $e_{1j}, ..., e_{nj}$ from the distribution of e_i where $j = 1, ..., B$. If the $m(\boldsymbol{x}_i)$ were known, then we could form a vector $\boldsymbol{Y}_j$ where the ith element $Y_{ij} = m(\boldsymbol{x}_i) + e_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j$ on $\boldsymbol{X}$. Instead, draw samples $r_{1j}^*, ..., r_{nj}^*$ with replacement from the residuals, then form a vector $\boldsymbol{Y}_j^*$ where the ith element $Y_{ij}^* = \hat{m}(\boldsymbol{x}_i) + r_{ij}^*$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$. If the residuals do not sum to 0, it is often useful to replace r_i by $\epsilon_i = r_i - \overline{r}$, and r_{ij}^* by ϵ_{ij}^*.

Example 4.6. For multiple linear regression, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$, $\boldsymbol{r}$, and $\hat{\boldsymbol{Y}}$ with ith element $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. For $j = 1, ..., B$, regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to form $\hat{\boldsymbol{\beta}}_{1,n}^*, ..., \hat{\boldsymbol{\beta}}_{B,n}^*$ using the residual bootstrap.

Now examine the OLS model. Let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ be the fitted values from the OLS full model. Let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the OLS full model residuals. Following Freedman (1981) and Efron (1982, p. 36),

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

follows a standard linear model where the elements r_i^W of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. Hence

$$E(r_i^W) = \frac{1}{n}\sum_{i=1}^{n} r_i = 0, \quad V(r_i^W) = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{n-p}{n}MSE,$$

$$E(\boldsymbol{r}^W) = \boldsymbol{0}, \text{ and } \mathrm{Cov}(\boldsymbol{Y}^*) = \mathrm{Cov}(\boldsymbol{r}^{\mathrm{W}}) = \sigma_{\mathrm{n}}^2\boldsymbol{I}_{\mathrm{n}}.$$

Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma_n^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n}MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and $E(\hat{\boldsymbol{\beta}}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

For the OLS estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$ is $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The sample covariance matrix of the $\hat{\boldsymbol{\beta}}^*$ is estimating $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*)$ as $B \to \infty$. Hence the residual bootstrap standard error $SE(\hat{\beta}_i^*) \approx \sqrt{\frac{n-p}{n}}\, SE(\hat{\beta}_i)$ for $i = 1, ..., p$ where

$\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. The LS CLT Theorem 2.26 says

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \lim_{n\to\infty} n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS})) \sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{W})$$

where $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{W}$. Since $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model, it may not be surprising that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \lim_{n\to\infty} n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}^*)) \sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{W}).$$

See Freedman (1981).

For the above residual bootstrap, $\hat{\boldsymbol{\beta}}^*_{I_j} = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*_{I_j}) = \sigma^2_n(\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}^*_{I_j}) = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ since $\boldsymbol{H}\boldsymbol{X}_{I_j} = \boldsymbol{X}_{I_j}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

Thus for $S \subseteq I$ and the residual bootstrap using residuals from the full OLS model, $E(\hat{\boldsymbol{\beta}}^*_I) = \hat{\boldsymbol{\beta}}_I$ and $n\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*_I) = n[(n-p)/n]\hat{\sigma}^2_n(\boldsymbol{X}^T_I\boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{V}_I$ as $n \to \infty$ with $\hat{\sigma}^2_n = MSE$. Hence $\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I \xrightarrow{P} \mathbf{0}$ as $n \to \infty$ by Lai et al (1979). Note that $\hat{\boldsymbol{\beta}}^*_I = \hat{\boldsymbol{\beta}}^*_{I,n}$ and $\hat{\boldsymbol{\beta}}_I = \hat{\boldsymbol{\beta}}_{I,n}$ depend on n.

Remark 4.20. The Cauchy Schwartz inequality says $|\boldsymbol{a}^T\boldsymbol{b}| \le \|\boldsymbol{a}\| \ \ \|\boldsymbol{b}\|$. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_P(1)$ is bounded in probability. This will occur if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, e.g. if $\hat{\boldsymbol{\beta}}$ is the OLS estimator. Then

$$|r_i - e_i| = |Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - (Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})| = |\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|.$$

Hence

$$\sqrt{n}\max_{i=1,...,n}|r_i - e_i| \le (\max_{i=1,...,n}\|\boldsymbol{x}_i\|) \ \ \|\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\| = O_P(1)$$

since $\max\|\boldsymbol{x}_i\| = O_P(1)$ or there is extrapolation. Hence OLS residuals behave well if the zero mean error distribution of the iid e_i has a finite variance σ^2.

Remark 4.21. Note that both the residual bootstrap and parametric bootstrap for OLS are robust to the unknown error distribution of the iid e_i. For the residual bootstrap with $S \subseteq I$ where I is not the full model, it may not be true that $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}_I)$ as $n, B \to \infty$. For the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the e_i are iid from a distribution that does not depend on n, and $\boldsymbol{\beta}_E = \mathbf{0}$. For $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$, the distribution of the r^W_i depends on n and $\hat{\boldsymbol{\beta}}_E \neq \mathbf{0}$ although $\sqrt{n}\hat{\boldsymbol{\beta}}_E = O_P(1)$.

4.6.3 The Nonparametric Bootstrap

The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, the pairwise bootstrap, and the pairs bootstrap) draws a sample of n cases $(Y_i^*, \boldsymbol{x}_i^*)$ with replacement from the n cases $(Y_i, \boldsymbol{x}_i)$, and regresses the Y_i^* on the $\boldsymbol{x}_i^*$ to get $\hat{\boldsymbol{\beta}}^*_{VS,1}$, and then draws another sample to get $\hat{\boldsymbol{\beta}}^*_{MIX,1}$. This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$.

Then for the full model,

$$\boldsymbol{Y}^* = \boldsymbol{X}^* \hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

and for a submodel I,

$$\boldsymbol{Y}^* = \boldsymbol{X}_I^* \hat{\boldsymbol{\beta}}_{I,OLS} + \boldsymbol{r}_I^W.$$

Freedman (1981) showed that under regularity conditions for the OLS MLR model, $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{W}) \sim N_p(\boldsymbol{0}, \boldsymbol{V})$. Hence if $S \subseteq I_j$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$. (Treat I_j as if I_j is the full model.)

One set of regularity conditions is that the MLR model holds, and if $\boldsymbol{x}_i = (1 \ \boldsymbol{u}_i^T)^T$, then the $\boldsymbol{w}_i = (Y_i \ \boldsymbol{u}_i^T)^T$ are iid from some population with a nonsingular covariance matrix. Since cases are sampled with replacement, we have $Y_i^* = \boldsymbol{x}_i^{*T}\boldsymbol{\beta} + e_i^*$ for $i = 1, ..., n$. In matrix form $\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{e}^*$, but $\boldsymbol{X}^*$ is a random matrix and the e_i^* are not iid from the distribution of the e_i since the e_i^* are "sampled with replacement" from the unknown $e_1, ..., e_n$.

The nonparametric bootstrap uses $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ where the $\boldsymbol{w}_i^*$ are sampled with replacement from $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. By Example 4.2, $E(\boldsymbol{w}^*) = \overline{\boldsymbol{w}}$, and

$$\text{Cov}(\boldsymbol{w}^*) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \begin{bmatrix} \tilde{S}_Y^2 & \tilde{\boldsymbol{\Sigma}}_{Y\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} & \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} \end{bmatrix}.$$

Note that $\hat{\boldsymbol{\beta}}$ is a constant with respect to the bootstrap distribution. Assume all inverse matrices exist. Then by Theorem 2.20,

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^* \end{bmatrix} = \begin{bmatrix} \overline{Y}^* - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^{*T}\overline{\boldsymbol{u}}^* \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1^*}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}^* \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \overline{Y} - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^T\overline{\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}} \end{bmatrix} = \hat{\boldsymbol{\beta}}$$

as $B \to \infty$. This result suggests that the nonparametric bootstrap for OLS MLR might work under milder regularity conditions than the $\boldsymbol{w}_i$ being iid from some population with a nonsingular covariance matrix.

4.6.4 Bootstrapping OLS Variable Selection

Undercoverage can occur if the bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

To see the effect of zero padding, consider $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_O = \mathbf{0}$ where $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$ and $O \subseteq E$ in (4.1) so that H_0 is true. Suppose a nominal 95% confidence region is used and $U_B = 0.96$. Hence the confidence region (4.32) or (4.33) covers at least 96% of the bootstrap sample. If $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for more than 4% of the $\hat{\boldsymbol{\beta}}^*_{O,1}, ..., \hat{\boldsymbol{\beta}}^*_{O,B}$, then $\mathbf{0}$ is in the confidence region and the bootstrap test fails to reject H_0. If this occurs for each run in the simulation, then the observed coverage will be 100%.

Now suppose $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for $j = 1, ..., B$. Then $\boldsymbol{S}^*_T$ is singular, but the singleton set $\{\mathbf{0}\}$ is the large sample $100(1-\delta)\%$ confidence region (4.32), (4.33), or (4.34) for $\boldsymbol{\beta}_O$ and $\delta \in (0, 1)$, and the pvalue for $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ is one. (This result holds since $\{\mathbf{0}\}$ contains 100% of the $\hat{\boldsymbol{\beta}}^*_{O,j}$ in the bootstrap sample.) For large sample theory tests, the pvalue estimates the population pvalue. Let I denote the other predictors in the model so $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. For the I_{min} model from forward selection, there may be strong evidence that $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_I$ is in the model if the "100%" confidence region is $\{\mathbf{0}\}$, $n \geq 20p$, $B \geq 50p$, and the error distribution is unimodal and not highly skewed. (Since the pvalue is one, this technique may be useful for data snooping: applying OLS theory to submodel I may have negligible selection bias.)

Remark 4.22. The assumption $\rho_{jn} \to \pi_j$ as $n \to \infty$ seems to be the most reasonable for the residual bootstrap since $|r_i - e_i| \to 0$ fast by Remark 4.20. The assumption may not hold for the parametric bootstrap of Section 4.6.1 if the e_i are not iid $N(0, \sigma^2)$. Another way to look at the bootstrap confidence region for OLS variable selection estimators is to consider the estimator $T_{2,n}$ that chooses I_j with probability equal to the observed bootstrap proportion $\hat{\rho}_{jn}$. The bootstrap sample $T^*_1, ..., T^*_B$ tends to be slightly more variable than an iid sample $T_{2,1}, ..., T_{2,B}$, and the geometric argument suggests that the large sample coverage of the nominal $100(1-\delta)\%$ confidence region will be at least as large as the nominal coverage $100(1-\delta)\%$.

Remark 4.23. Note that there are several important variable selection models, including the model given by Equation (4.1) where $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Another model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T\boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then there are $K \geq 2$ competing "true" nonnested submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. For example, suppose the $K = 2$ models have predictors x_1, x_2, x_3 for S_1 and x_1, x_2, x_4 for S_2. Then x_3 and x_4 are likely to be selected and omitted often by forward selection for the B bootstrap samples. Hence omitting all predictors x_i that have a $\beta^*_{ij} = 0$ for at least one of the bootstrap samples $j = 1, ..., B$ could

result in underfitting, e.g. using just x_1 and x_2 in the above $K = 2$ example. If n and B are large enough, the singleton set $\{\mathbf{0}\}$ could still be the "100%" confidence region for a vector $\boldsymbol{\beta}_O$. See Remark 4.6.

Suppose the predictors x_i have been standardized. Then another important regression model has the β_i taper off rapidly, but no coefficients are equal to zero. For example, $\beta_i = e^{-i}$ for $i = 1, ..., p$.

Example 4.7. Cook and Weisberg (1999, pp. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. Let the response variable be the logarithm $\log(M)$ of the *muscle mass*, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log(W)$ of the *shell width* W, the logarithm $\log(S)$ of the *shell mass* S, and a constant. Inference for the full model is shown below along with the shorth(c) nominal 95% confidence intervals for β_i computed using the nonparametric and residual bootstraps. As expected, the residual bootstrap intervals are close to the classical least squares confidence intervals $\approx \hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$.

```
      large sample full model inference
      Est.    SE  t   Pr(>|t|)   nparboot      resboot
int -1.249 0.838 -1.49 0.14 [-2.93,-0.093][-3.045,0.473]
L   -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
logW 0.130 0.374  0.35 0.73 [-0.457,0.829][-0.703,0.890]
H    0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.016]
logS 0.640 0.169  3.80 0.00 [ 0.244,1.040][ 0.336,1.012]
output and shorth intervals for the min Cp submodel FS
      Est.     SE     95% shorth CI   95% shorth CI
int   -0.9573  0.1519 [-3.294, 0.495] [-2.769, 0.460]
L      0              [-0.005, 0.004] [-0.004, 0.004]
logW   0              [ 0.000, 1.024] [-0.595, 0.869]
H      0.0072  0.0047 [ 0.000, 0.016] [ 0.000, 0.016]
logS   0.6530  0.1160 [ 0.322, 0.901] [ 0.324, 0.913]
               for forward selection for all subsets
```

The minimum C_p model from all subsets variable selection and forward selection both used a constant, H, and $\log(S)$. The shorth(c) nominal 95% confidence intervals for β_i using the residual bootstrap are shown. Note that the intervals for H are right skewed and contain 0 when closed intervals are used instead of open intervals. Some least squares output is shown, but should only be used for inference if the model was selected before looking at the data.

It was expected that $\log(S)$ may be the only predictor needed, along with a constant, since $\log(S)$ and $\log(M)$ are both log(mass) measurements and likely highly correlated. Hence we want to test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the I_{min} model selected by all subsets variable selection. (Of course this test would be easy to do with the full model using least squares theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_3, \beta_4)^T = \mathbf{0}$. Using the prediction region method with the

full model gave an interval [0,2.930] with $D_{\mathbf{0}} = 1.641$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} variable selection model had $[0, D_{(U_B)}] = [0, 3.293]$ while $D_{\mathbf{0}} = 1.134$. So fail to reject H_0.

Then we redid the bootstrap with the full model and forward selection. The full model had $[0, D_{(U_B)}] = [0, 2.908]$ with $D_{\mathbf{0}} = 1.577$. So fail to reject H_0. Using the prediction region method with the I_{min} forward selection model had $[0, D_{(U_B)}] = [0, 3.258]$ while $D_{\mathbf{0}} = 1.245$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test was 0.392. (Use (4.35) with $\boldsymbol{S}_T^*$ and D from forward selection for the numerator, and from the full model for the denominator.) Hence the forward selection bootstrap test was more precise than the full model bootstrap test. Some R code used to produce the above output is shown below.

```
library(leaps)
y <- log(mussels[,5]); x <- mussels[,1:4]
x[,4] <- log(x[,4]); x[,2] <- log(x[,2])
out <- regboot(x,y,B=1000)
tem <- rowboot(x,y,B=1000)
outvs <- vselboot(x,y,B=1000) #get bootstrap CIs
outfs <- fselboot(x,y,B=1000) #get bootstrap CIs
apply(out$betas,2,shorth3);
apply(tem$betas,2,shorth3);
apply(outvs$betas,2,shorth3)  #for all subsets
apply(outfs$betas,2,shorth3) #for forward selection
ls.print(outvs$full)
ls.print(outvs$sub)
ls.print(outfs$sub)
#test if beta_2 = beta_3 = beta_4 = 0
Abeta <- out$betas[,2:4]  #full model
#prediction region method with residual bootstrap
out<-predreg(Abeta)
Abeta <- outvs$betas[,2:4]
#prediction region method with Imin all subsets
outvs <- predreg(Abeta)
Abeta <- outfs$betas[,2:4]
#prediction region method with Imin forward sel.
outfs<-predreg(Abeta)
#ratio of volumes for forward selection and full model
(sqrt(det(outfs$cov))*outfs$D0^3)/(sqrt(det(out$cov))*out$D0^3)
```

Example 4.8. Consider the Gladstone (1905) data set that has 12 variables on 267 persons after death. The response variable was *brain weight*. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and two categor-

ical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The eight predictor variables shown in the output were used.

Output is shown below for the full model and the bootstrapped minimum C_p forward selection estimator. Note that the shorth intervals for *length* and *sex* are quite long. These variables are often in and often deleted from the bootstrap forward selection. Model I_I is the model with the fewest predictors such that $C_P(I_I) \leq C_P(I_{min})+1$. For this data set, $I_I = I_{min}$. The bootstrap CIs differ due to different random seeds.

```
large sample full model inference for Ex. 4.8
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -3021.255 1701.070 -1.77 0.077 [-6549.8,322.79]
age      -1.656    0.314 -5.27 0.000 [ -2.304,-1.050]
breadth  -8.717   12.025 -0.72 0.469 [-34.229,14.458]
cephalic 21.876   22.029  0.99 0.322 [-20.911,67.705]
circum    0.852    0.529  1.61 0.109 [ -0.065, 1.879]
headht    7.385    1.225  6.03 0.000 [  5.138, 9.794]
height   -0.407    0.942 -0.43 0.666 [ -2.211, 1.565]
len      13.475    9.422  1.43 0.154 [ -5.519,32.605]
sex      25.130   10.015  2.51 0.013 [  6.717,44.19]
output and shorth intervals for the min Cp submodel
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6151.6,-415.4]
age      -1.708    0.285 -5.99 0.000 [ -2.299,-1.068]
breadth   0                           [-32.992, 8.148]
cephalic  5.958    2.089  2.85 0.005 [-10.859,62.679]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.817]
headht    7.424    1.161  6.39 0.000 [  5.028, 9.732]
height    0                           [ -2.859, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,30.508]
sex      25.313    9.920  2.55 0.011 [  0.000,42.144]
output and shorth for I_I model
       Estimate  Std.Err t-val Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6104.9,-778.2]
age      -1.708    0.285 -5.99 0.000 [ -2.259,-1.003]
breadth   0                           [-31.012, 6.567]
cephalic  5.958    2.089  2.85 0.005 [ -6.700,61.265]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.866]
headht    7.424    1.161  6.39 0.000 [  5.221,10.090]
height    0                           [ -2.173, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,28.819]
sex      25.313    9.920  2.55 0.011 [  0.000,42.847]
```

The *R* code used to produce the above output is shown below. The last four commands are useful for examining the variable selection output.

```
x<-cbrainx[,c(1,3,5,6,7,8,9,10)]
```

```
y<-cbrainy
library(leaps)
out <- regboot(x,y,B=1000)
outvs <- fselboot(x,cbrainy) #get bootstrap CIs,
apply(out$betas,2,shorth3)
apply(outvs$betas,2,shorth3)
ls.print(outvs$full)
ls.print(outvs$sub)
outvs <- modIboot(x,cbrainy) #get bootstrap CIs,
apply(outvs$betas,2,shorth3)
ls.print(outvs$sub)
tem<-regsubsets(x,y,method="forward")
tem2<-summary(tem)
tem2$which
tem2$cp
```

4.6.5 Simulations

For variable selection with the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The shorth estimator can be applied to a bootstrap sample $\hat{\beta}^*_{i1}, ..., \hat{\beta}^*_{iB}$ to get a confidence interval for β_i. Here $T_n = \hat{\beta}_i$ and $\theta = \beta_i$.

Assume p is fixed, $n \geq 20p$, and that the error distribution is unimodal and not highly skewed. Then the plotted points in the response and residual plots should scatter in roughly even bands about the identity line (with unit slope and zero intercept) and the $r = 0$ line, respectively. See Figure 1.1. If the error distribution is skewed or multimodal, then much larger sample sizes may be needed.

Next, we describe a small simulation study that was done using $B = \max(1000, n/25, 50p)$ and 5000 runs. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$ and $50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$ where k and ψ are defined in the following paragraph. In the simulations, we use $\theta = \boldsymbol{A\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_S = \mathbf{1}$ and $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_E = \mathbf{0}$.

Let $\boldsymbol{x} = (1\ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\mathbf{0}, \boldsymbol{I})$ where the $m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{AA}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+(m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(m-2)\psi^2]$. Hence the correlations are $Cor(x_i, x_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$,

then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k+1$ ones and $p-k-1$ zeros. The zero mean errors e_i were iid from five distributions: i) N(0,1), ii) t_3, iii) EXP(1) - 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). Only distribution iii) is not symmetric.

When $\psi = 0$, the full model least squares confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when $n = 100$ and the iid zero mean errors have variance σ^2. The simulation computed the Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ (whether first $k+1$ $\beta_i = 1$) and $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ (whether the last $p-k-1$ $\beta_i = 0$). The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 suggests coverage is close to the nominal value.

The regression models used the residual bootstrap on the forward selection estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Table 4.2 gives results for when the iid errors $e_i \sim N(0, 1)$ with $n = 100$, $p = 4$, and $k = 1$. Table 4.2 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for forward selection. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (4.32), hybrid region (4.34), and Bickel and Ren region (4.33). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$. The length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B,T)}]$ where $D_{(U_B)}$ or $D_{(U_B,T)}$ is the cutoff for the confidence region. The cutoff will often be near $\sqrt{\chi^2_{g,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is close to 2.45 for the full model regression bootstrap tests.

Volume ratios of the three confidence regions can be compared using (4.35), but there is not enough information in Table 4.2 to compare the volume of the confidence region for the full model regression versus that for the forward selection regression since the two methods have different determinants $|\boldsymbol{S}^*_T|$.

The inference for forward selection was often as precise or more precise than the inference for the full model. The coverages were near 0.95 for the regression bootstrap on the full model, although there was slight undercoverage for the tests since $(n-p)/n = 0.96$ when $n = 25p$. Suppose $\psi = 0$. Then from Section 4.2, $\hat{\boldsymbol{\beta}}_S$ has the same limiting distribution for I_{min} and the full model. Note that the average lengths and coverages were similar for the full model and forward selection I_{min} for β_1, β_2, and $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$. Forward selection inference was more precise for $\boldsymbol{\beta}_E = (\beta_3, \beta_4)^T$. The Bickel and Ren (4.33) cutoffs and coverages were at least as high as those of the hybrid region (4.34).

For $\psi > 0$ and I_{min}, the coverages for the β_i corresponding to $\boldsymbol{\beta}_S$ were near 0.95, but the average length could be shorter since I_{min} tends to have

Table 4.2 Bootstrapping OLS Forward Selection with C_p, $e_i \sim N(0,1)$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.946	0.950	0.947	0.948	0.940	0.941	0.941	0.937	0.936	0.937
len	0.396	0.399	0.399	0.398	2.451	2.451	2.452	2.450	2.450	2.451
vs,0	0.948	0.950	0.997	0.996	0.991	0.979	0.991	0.938	0.939	0.940
len	0.395	0.398	0.323	0.323	2.699	2.699	3.002	2.450	2.450	2.457
reg,0.5	0.946	0.944	0.946	0.945	0.938	0.938	0.938	0.934	0.936	0.936
len	0.396	0.661	0.661	0.661	2.451	2.451	2.452	2.451	2.451	2.452
vs,0.5	0.947	0.968	0.997	0.998	0.993	0.984	0.993	0.955	0.955	0.963
len	0.395	0.658	0.537	0.539	2.703	2.703	2.994	2.461	2.461	2.577
reg,0.9	0.946	0.941	0.944	0.950	0.940	0.940	0.940	0.935	0.935	0.935
len	0.396	3.257	3.253	3.259	2.451	2.451	2.452	2.451	2.451	2.452
vs,0.9	0.947	0.968	0.994	0.996	0.992	0.981	0.992	0.962	0.959	0.970
len	0.395	2.751	2.725	2.735	2.716	2.716	2.971	2.497	2.497	2.599

less multicorrelation than the full model. For $\psi \geq 0$, the I_{min} coverages were higher than 0.95 for β_3 and β_4 and for testing $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ since zeros often occurred for $\hat{\beta}_j^*$ for $j = 3, 4$. The average CI lengths were shorter for I_{min} than for the OLS full model for β_3 and β_4. Note that for I_{min}, the coverage for testing $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ was higher than that for the OLS full model.

Table 4.3 Bootstrap CIs with C_p, $p = 10, k = 8, \psi = 0.9$, error type v)

n	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
250	0.945	0.824	0.822	0.827	0.827	0.824	0.826	0.817	0.827	0.999
shlen	0.825	6.490	6.490	6.482	6.485	6.479	6.512	6.496	6.493	6.445
250	0.946	0.979	0.980	0.985	0.981	0.983	0.983	0.977	0.983	0.998
prlen	0.807	7.836	7.850	7.842	7.830	7.830	7.851	7.840	7.839	7.802
250	0.947	0.976	0.978	0.984	0.978	0.978	0.979	0.973	0.980	0.996
brlen	0.811	8.723	8.760	8.765	8.736	8.764	8.745	8.747	8.753	8.756
2500	0.951	0.947	0.948	0.948	0.948	0.947	0.949	0.944	0.951	0.999
shlen	0.263	2.268	2.271	2.271	2.273	2.262	2.632	2.277	2.272	2.047
2500	0.945	0.961	0.959	0.955	0.960	0.960	0.961	0.958	0.961	0.998
prlen	0.258	2.630	2.639	2.640	2.632	2.632	2.641	2.638	2.642	2.517
2500	0.946	0.958	0.954	0.960	0.956	0.960	0.962	0.955	0.961	0.997
brlen	0.258	2.865	2.875	2.882	2.866	2.871	2.887	2.868	2.875	2.830
25000	0.952	0.940	0.939	0.935	0.940	0.942	0.938	0.937	0.942	1.000
shlen	0.083	0.809	0.808	0.806	0.805	0.807	0.808	0.808	0.809	0.224
25000	0.948	0.964	0.968	0.962	0.964	0.966	0.964	0.964	0.967	0.991
prlen	0.082	0.806	0.805	0.801	0.800	0.805	0.805	0.803	0.806	0.340
25000	0.949	0.969	0.972	0.968	0.967	0.971	0.969	0.969	0.973	0.999
brlen	0.082	0.810	0.810	0.805	0.804	0.809	0.810	0.808	0.810	0.317

Results for other values of n, p, k, and distributions of e_i were similar. For forward selection with $\psi = 0.9$ and C_p, the hybrid region (4.34) and shorth confidence intervals occasionally had coverage less than 0.93. It was also rare for the bootstrap to have one or more columns of zeroes so $\boldsymbol{S}_T^*$ was singular.

For error distributions i)-iv) and $\psi = 0.9$, sometimes the shorth CIs needed $n \geq 100p$ for all p CIs to have good coverage. For error distribution v) and $\psi = 0.9$, even larger values of n were needed. Confidence intervals based on (4.32) and (4.33) worked for much smaller n, but tended to be longer than the shorth CIs.

See Table 4.3 for one of the worst scenarios for the shorth, where shlen, prlen, and brlen are for the average CI lengths based on the shorth, (4.32), and (4.33), respectively. In Table 4.3, $k = 8$ and the two nonzero π_j correspond to the full model $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{S,0}$. Hence $\beta_i = 1$ for $i = 1, ..., 9$ and $\beta_{10} = 0$. Hence confidence intervals for β_{10} had the highest coverage and usually the shortest average length (for $i \neq 1$) due to zero padding. Theory in Section 4.2 showed that the CI lengths are proportional to $1/\sqrt{n}$. When $n = 25000$, the shorth CI uses the 95.16th percentile while CI (4.32) uses the 95.00th percentile, allowing the average CI length of (4.32) to be shorter than that of the shorth CI, but the distribution for $\hat{\beta}_i^*$ is likely approximately symmetric for $i \neq 10$ since the average lengths of the three confidence intervals were about the same for each $i \neq 10$.

When BIC was used, undercoverage was a bit more common and severe, and undercoverage occasionally occurred with regions (4.32) and (4.33). BIC also occasionally had 100% coverage since BIC produces more zeroes than C_p.

Some *R* code for the simulation is shown below.

```
record coverages and ``lengths" for
b1, b2, bp-1, bp, pm0, hyb0, br0, pm1, hyb1, br1

regbootsim3(n=100,p=4,k=1,nruns=5000,type=1,psi=0)
$cicov
[1] 0.9458 0.9500 0.9474 0.9484 0.9400 0.9408 0.9410
0.9368 0.9362 0.9370
$avelen
[1] 0.3955 0.3990 0.3987 0.3982 2.4508 2.4508 2.4521
[8] 2.4496 2.4496 2.4508
$beta
[1] 1 1 0 0
$k
[1] 1
library(leaps)
vsbootsim4(n=100,p=4,k=1,nruns=5000,type=1,psi=0)
$cicov
[1] 0.9480 0.9496 0.9972 0.9958 0.9910 0.9786 0.9914
0.9384 0.9394 0.9402
$avelen
[1] 0.3954 0.3987 0.3233 0.3231 2.6987 2.6987 3.0020
[8] 2.4497 2.4497 2.4570
```

```
$beta
[1] 1 1 0 0
$k
[1] 1
```

4.7 Data Splitting

Data splitting is used for inference after model selection. Use a training set to select a full model, and a validation set for inference with the selected full model. Here $p >> n$ is possible. See Chapter 6, Hurvich and Tsai (1990, p. 216) and Rinaldo et al. (2019). Typically when training and validation sets are used, the training set is bigger than the validation set or half sets are used, often causing large efficiency loss.

Let J be a positive integer and let $\lfloor x \rfloor$ be the integer part of x, e.g., $\lfloor 7.7 \rfloor = 7$. Initially divide the data into two sets H_1 with $n_1 = \lfloor n/(2J) \rfloor$ cases and V_1 with $n - n_1$ cases. If the fitted model from H_1 is not good enough, randomly select n_1 cases from V_1 to add to H_1 to form H_2. Let V_2 have the remaining cases from V_1. Continue in this manner, possibly forming sets $(H_1, V_1), (H_2, V_2), ..., (H_J, V_J)$ where H_i has $n_i = in_1$ cases. Stop when H_d gives a reasonable model I_d with a_d predictors if $d < J$. Use $d = J$, otherwise. Use the model I_d as the full model for inference with the data in V_d.

This procedure is simple for a fixed data set, but it would be good to automate the procedure. Forward selection with the Chen and Chen (2008) EBIC criterion and lasso are useful for finding a reasonable fitted model. BIC and the Hurvich and Tsai (1989) AIC_C criterion can be useful if $n \geq \max(2p, 10a_d)$. For example, if $n = 500000$ and $p = 90$, using $n_1 = 900$ would result in a much smaller loss of efficiency than $n_1 = 250000$.

4.8 Summary

1) A *model for variable selection* can be described by $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$. Assume p is fixed while $n \to \infty$.

2) If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. For the OLS model with $S \subseteq I$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$ where $(\boldsymbol{X}_I^T\boldsymbol{X}_I)/(n\sigma^2) \xrightarrow{P} \boldsymbol{V}_I^{-1}$.

3) **Theorem 4.4, Variable Selection CLT.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $T_n = \hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \hat{\boldsymbol{\beta}}_{I_j,0}$. Let $T_n = T_{kn} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the π_k with $S \subseteq I_k$ by π_j. The other $\pi_k = 0$ since $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ and $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v}$$

where $\boldsymbol{A}\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

4) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$ where $0 < \delta < 1$. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

5) Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and $q_n = \min(1-\delta/2, 1-\delta+10\delta p/n)$, otherwise. If $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq h\}$ is a large sample $100(1-\delta)\%$ prediction regions if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i. The large sample $100(1-\delta)\%$ nonparametric prediction region $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}$ uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. We want $n \geq 10p$ for good coverage and $n \geq 50p$ for good volume.

6) Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Make a confidence region and reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region. Let q_B and U_B be as in 5) with n replaced by B and p replaced by g. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$. a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}^*_T) \leq D^2_{(U_B)}\}$ where $D^2_{(U_B)}$ is computed from $D^2_i = (T^*_i - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (T^*_i - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}^*_T]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. This procedure applies the nonparametric prediction region to the bootstrap sample. b) The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \leq D^2_{(U_B,T)}\}$ where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample

quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. c) The hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}$.

If $g = 1$, confidence intervals can be computed without $\boldsymbol{S}_T^*$ or D^2 for a), b), and c).

For some data sets, $\boldsymbol{S}_T^*$ may be singular due to one or more columns of zeroes in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model if n and B are large enough. Let $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$, and consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_O = \boldsymbol{0}$. If $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{O,i} = \boldsymbol{0}$ for greater than $B\delta$ of the bootstrap samples $i = 1, ..., B$, then fail to reject H_0. (If $\boldsymbol{S}_T^*$ is nonsingular, the $100(1-\delta)\%$ prediction region method confidence region contains $\boldsymbol{0}$.)

7) **Theorem 4.7: Geometric Argument.** Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$.

8) Applying the nonparametric prediction region (4.24) to the iid data $T_1, ..., T_B$ results in the $100(1-\delta)\%$ confidence region $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T \boldsymbol{S}_T^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}(T_n, \boldsymbol{S}_T)\}$ where $D^2_{(U_B)}(T_n, \boldsymbol{S}_T)$ is computed from the $(T_i - T_n)^T \boldsymbol{S}_T^{-1} (T_i - T_n)$ provided the $\boldsymbol{S}_T = \boldsymbol{S}_{T_n}$ are "not too ill conditioned." For OLS variable selection, assume there are two or more component clouds. The bootstrap component data clouds have the same asymptotic covariance matrix as the iid component data clouds, which are centered at $\boldsymbol{\theta}$. The jth bootstrap component data cloud is centered at $E(T_{ij}^*)$ and often $E(T_{jn}^*) = T_{jn}$. Confidence region (4.32) is the prediction region (4.24) applied to the bootstrap sample, and (4.32) is slightly larger in volume than (4.24) applied to the iid sample, asymptotically. The hybrid region (4.34) shifts (4.32) to be centered at T_n. Shifting the component clouds slightly and computing (4.24) does not change the axes of the prediction region (4.24) much compared to not shifting the component clouds. Hence by the geometric argument, we expect (4.34) to have coverage at least as high as the nominal, asymptotically, provided the $\boldsymbol{S}_T^*$ are "not too ill conditioned." The Bickel and Ren confidence region (4.33) tends to have higher coverage and volume than (4.34). Since $\overline{T}^*$ tends to be closer to $\boldsymbol{\theta}$ than T_n, (4.32) tends to have good coverage.

9) Suppose m independent large sample $100(1-\delta)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the m runs. Let Y count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1 - \delta_n)$ where $1 - \delta_n$ is the true coverage. Simulation can be used to see if the true or actual coverage $1 - \delta_n$ is close to the nominal coverage $1 - \delta$. A prediction region with $1 - \delta_n < 1 - \delta$ is liberal and a region with $1 - \delta_n > 1 - \delta$ is conservative. It is better to be conservative by 3% than

liberal by 3%. Parametric prediction regions tend to have large undercoverage and so are too liberal. Similar definitions are used for confidence regions.

10) For the bootstrap, perform variable selection on $\boldsymbol{Y}_i^*$ and $\boldsymbol{X}$ (or $\boldsymbol{X}^*$ for the nonparametric bootstrap), fit the model that minimizes the criterion, and add 0s corresponding to the omitted variables, resulting in estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$ where $\hat{\boldsymbol{\beta}}_i^* = \hat{\boldsymbol{\beta}}_{I_{min},0,i}^*$.

11) Let $Z_1, ..., Z_n$ be random variables, let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics, and let c be a positive integer. Compute $Z_{(c)} - Z_{(1)}, Z_{(c+1)} - Z_{(2)}, ..., Z_{(n)} - Z_{(n-c+1)}$. Let shorth(c) $= [\mathrm{Z_{(d)}}, \mathrm{Z_{(d+c-1)}}]$ correspond to the interval with the shortest length.

The large sample $100(1-\delta)\%$ *shorth*(c) *CI* uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length. Here $c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\,\rceil\,\rceil)$. The shorth CI is computed by applying the shorth PI to the bootstrap sample.

4.9 Complements

This chapter followed Olive (2017b, ch. 5) and Pelawa Watagoda and Olive (2019ab) closely. Also see Olive (2013a, 2018), Pelawa Watagoda (2017), and Rathnayake and Olive (2019). For MLR, Olive (2017a: p. 123, 2017b: p. 176) showed that $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a consistent estimator. Olive (2014: p. 283, 2017ab, 2018) recommended using the shorth(c) estimator for the percentile method. Olive (2017a: p. 128, 2017b: p. 181, 2018) showed that the prediction region method can simulate well for the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Hastie et al. (2009, p. 57) noted that variable selection is a shrinkage estimator: the coefficients are shrunk to 0 for the omitted variables.

Good references for the bootstrap include Efron (1979, 1982), Efron and Hastie (2016, ch. 10–11), and Efron and Tibshirani (1993). Also see Chen (2016) and Hesterberg (2014). One of the sufficient conditions for the bootstrap confidence region is that T has a well behaved Hadamard derivative. Fréchet differentiability implies Hadamard differentiability, and many statistics are shown to be Hadamard differentiable in Bickel and Ren (2001), Clarke (1986, 2000), Fernholtz (1983), Gill (1989), Ren (1991), and Ren and Sen (1995). Bickel and Ren (2001) showed that their method can work when Hadamard differentiability fails.

There is a massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Leeb and Pötscher (2005, 2006, 2008), Leeb et al. (2015), Tibshirani et al. (2016), and Tibshirani et al. (2018). Knight and Fu (2000) have some results on the residual bootstrap that uses residuals from one estimator, such as full model OLS, but fit another estimator, such as lasso.

Inference techniques for the variable selection model, other than data splitting, have not had much success. For multiple linear regression, the methods are often inferior to data splitting, often assume normality, or are asymptotically equivalent to using the full model, or find a quantity to test that is not $\boldsymbol{A\beta}$. See Ewald and Schneider (2018). Berk et al. (2013) assumes normality, needs p no more than about 30, assumes σ^2 can be estimated independently of the data, and Leeb et al. (2015) say the method does not work. The bootstrap confidence region (4.32) is centered at $\overline{T}^* \approx \sum_j \rho_{jn} T_{jn}$, which is closely related to a model averaging estimator. Wang and Zhou (2013) show that the Hjort and Claeskens (2003) confidence intervals based on frequentist model averaging are asymptotically equivalent to those obtained from the full model. See Buckland et al. (1997) and Schomaker and Heumann (2014) for standard errors when using the bootstrap or model averaging for linear model confidence intervals.

Efron (2014) used the confidence interval $\overline{T}^* \pm z_{1-\delta} SE(\overline{T}^*)$ assuming $\overline{T}^*$ is asymptotically normal and using delta method techniques, which require nonsingular covariance matrices. There is not yet rigorous theory for this method. Section 4.2 proved that $\overline{T}^*$ is asymptotically normal: under regularity conditions: if $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$ and$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$, then under regularity conditions $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$. If $g = 1$, then the prediction region method large sample $100(1-\delta)\%$ CI for θ has $P(\theta \in [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]) \to 1 - \delta$ as $n \to \infty$. If the Frey CI also has coverage converging to $1-\delta$, than the two methods have the same asymptotic length (scaled by multiplying by $\sqrt{n}$), since otherwise the shorter interval will have lower asymptotic coverage.

For the mixture distribution with two or more component groups, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by Theorem 4.4 b). If $\sqrt{n}(T_i^* - c_n) \xrightarrow{D} \boldsymbol{u}$ then c_n must be a value such as $c_n = \overline{T}^*$, $c_n = \sum_j \rho_{jn} T_{jn}$, or $c_n = \sum_j \pi_j T_{jn}$. Next we will examine $\overline{T}^*$. If $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$, and for the parametric and nonparametric bootstrap, $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j,0} - \hat{\boldsymbol{\beta}}_{I_j,0}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$. Let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0} = \boldsymbol{A}\boldsymbol{D}_{j0}\boldsymbol{Y}$ using notation from Section 4.6. Let $\boldsymbol{\theta} = \boldsymbol{A\beta}$. Hence from Section 4.5.3, $\sqrt{n}(\overline{T}^*_j - T_{jn}) \xrightarrow{P} \mathbf{0}$. Assume $\hat{\rho}_{in} \xrightarrow{P} \rho_i$ as $n \to \infty$. Then $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) =$

$$\sum_i \hat{\rho}_{in}\sqrt{n}(\overline{T}^*_i - \boldsymbol{\theta}) = \sum_j \hat{\rho}_{jn}\sqrt{n}(\overline{T}^*_j - \boldsymbol{\theta}) + \sum_k \hat{\rho}_{kn}\sqrt{n}(\overline{T}^*_k - \boldsymbol{\theta})$$

$= d_n + a_n$ where $a_n \xrightarrow{P} \mathbf{0}$ since $\rho_k = 0$. Now

$$d_n = \sum_j \hat{\rho}_{jn}\sqrt{n}(\overline{T}^*_j - T_{jn} + T_{jn} - \boldsymbol{\theta}) = \sum_j \hat{\rho}_{jn}\sqrt{n}(T_{jn} - \boldsymbol{\theta}) + c_n$$

where $c_n = o_P(1)$ since $\sqrt{n}(\overline{T}_j^* - T_{jn}) = o_P(1)$. Hence under regularity conditions, if $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$ then $\sum_j \rho_j \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$.

To examine the last term and $\boldsymbol{w}$, let the $n \times 1$ vector $\boldsymbol{Y}$ have characteristic function $\phi_{\boldsymbol{Y}}$, $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, and $\mathrm{Cov}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}$. Let $\boldsymbol{Z} = (\boldsymbol{Y}^T, ..., \boldsymbol{Y}^T)^T$ be a $Jn \times 1$ vector with J copies of $\boldsymbol{Y}$ stacked into a vector. Let $\boldsymbol{t} = (\boldsymbol{t}_1^T, ..., \boldsymbol{t}_J^T)^T$. Then $\boldsymbol{Z}$ has characteristic function $\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}}(\sum_{j=1}^J \boldsymbol{t}_i) = \phi_{\boldsymbol{Y}}(\boldsymbol{s})$. Now assume $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$. Then $\boldsymbol{t}^T \boldsymbol{Z} = \boldsymbol{s}^T \boldsymbol{Y} \sim N(\boldsymbol{s}^T \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{s}^T \boldsymbol{s})$. Hence $\boldsymbol{Z}$ has a multivariate normal distribution by Definition 1.23 with $E(\boldsymbol{Z}) = (\boldsymbol{X}\boldsymbol{\beta}^T, ..., \boldsymbol{X}\boldsymbol{\beta}^T)^T$, and $\mathrm{Cov}(\boldsymbol{Z})$ a block matrix with $J \times J$ blocks each equal to $\sigma^2 \boldsymbol{I}$. Then

$$\sum_j \rho_j T_{jn} = \sum_j \rho_j \boldsymbol{A}\boldsymbol{D}_{j0}\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{Y} \sim N_g(\boldsymbol{\theta}, \sigma^2 \boldsymbol{B}\boldsymbol{B}^T) =$$

$$N_g(\boldsymbol{\theta}, \sigma^2 \sum_j \sum_k \rho_j \rho_k \boldsymbol{A}\boldsymbol{D}_{j0}\boldsymbol{D}_{k0}^T \boldsymbol{A})$$

since $E(T_{jn}) = E(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}) = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\theta}$ if $S \subseteq I_j$. Since $(T_{1n}^T, ..., T_{jn}^T)^T = diag(\boldsymbol{A}D_{10}, ..., \boldsymbol{A}\boldsymbol{D}_{J0})\boldsymbol{Z}$, then $(T_{1n}^T, ..., T_{jn}^T)^T$ is multivariate normal and

$$\sum_j \rho_j T_{jn} \sim N_g[\boldsymbol{\theta}, \sum_j \sum_k \pi_j \pi_k \mathrm{Cov}(T_{jn}, T_{kn})].$$

Now assume $n\boldsymbol{D}_{j0}\boldsymbol{D}_{k0}^T \xrightarrow{P} \boldsymbol{W}_{jk}$ as $n \to \infty$. Then

$$\sum_j \rho_j \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w} \sim N_g(\boldsymbol{0}, \sigma^2 \sum_j \sum_k \rho_j \rho_k \boldsymbol{A}\boldsymbol{W}_{jk}\boldsymbol{A}).$$

We conjecture that this result may hold under milder conditions than $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, but even the above results are not yet rigorous. If $\sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}_j \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$, then a possibly poor approximation is $\overline{T}^* \approx \sum_j \rho_j T_{jn} \approx N_g[\boldsymbol{\theta}, \sum_j \sum_k \rho_j \rho_k Cov(T_{jn}, T_{kn})]$, and estimating $\sum_j \sum_k \rho_j \rho_k Cov(T_{jn}, T_{kn})$ with delta method techniques may not be possible.

The double bootstrap technique may be useful. See Hall (1986) and Chang and Hall (2015) for references. The double bootstrap for $\overline{T}^* = \overline{T}_B^*$ says that $T_n = \overline{T}^*$ is a statistic that can be bootstrapped. Let $B_d \geq 50 g_{max}$ where $1 \leq g_{max} \leq p$ is the largest dimension of $\boldsymbol{\theta}$ to be tested with the double bootstrap. Draw a bootstrap sample of size B and compute $\overline{T}^* = T_1^*$. Repeat for a total of B_d times. Apply the confidence region (4.32), (4.33), or (4.34) to the double bootstrap sample $T_1^*, ..., T_{B_d}^*$. If $D_{(U_{B_d})} \approx D_{(U_{B_d},T)} \approx \sqrt{\chi^2_{g,1-\delta}}$, then $\overline{T}^*$ may be approximately multivariate normal. The CI (4.32) applied to the double bootstrap sample could be regarded as a modified Frey CI

without delta method techniques. Of course the double bootstrap tends to be too computationally expensive to simulate.

We can get a prediction region by randomly dividing the data into two half sets H and V where H has $n_H = \lceil n/2 \rceil$ of the cases and V has the remaining $m = n_V = n - n_H$ cases. Compute $(\overline{\boldsymbol{x}}_H, \boldsymbol{S}_H)$ from the cases in H. Then compute the distances $D_i^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_H)^T \boldsymbol{S}_H^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_H)$ for the m vectors $\boldsymbol{x}_i$ in V. Then a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{x}_F$ is $\{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\overline{\boldsymbol{x}}_H, \boldsymbol{S}_H) \leq D_{(k_m)}^2\}$ where $k_m = \lceil m(1-\delta) \rceil$. This prediction region may give better coverage than the nonparametric prediction region (4.24) if $5p \leq n \leq 20p$.

The iid sample $T_1, ..., T_B$ has sample mean $\overline{T}$. Let $T_{in} = T_{ijn}$ if T_{jn} is chosen D_{jn} times where the random variables $D_{jn}/B \xrightarrow{P} \pi_{jn}$. The D_{jn} follow a multinomial distribution. Then the iid sample can be written as

$$T_{1,1}, ..., T_{D_{1n},1}, ..., T_{1,J}, ..., T_{D_{Jn},J},$$

where the T_{ij} are not iid. Denote $T_{1j}, ..., T_{D_{jn},j}$ as the jth component of the iid sample with sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$. Thus

$$\overline{T} = \frac{1}{B} \sum_{i=1}^{B} T_{ijn} = \sum_j \frac{D_{jn}}{B} \frac{1}{D_{jn}} \sum_{i=1}^{D_{jn}} T_{ij} = \sum_j \hat{\pi}_{jn} \overline{T}_j.$$

Hence $\overline{T}$ is a random linear combination of the $\overline{T}_j$. Conditionally on the D_{jn}, the T_{ij} are independent, and $\overline{T}$ is a linear combination of the $\overline{T}_j$. Note that $\mathrm{Cov}(\overline{T}) = \mathrm{Cov}(T_n)/B$.

Software. The simulations were done in *R*. See R Core Team (2016). We used several *R* functions including forward selection as computed with the `regsubsets` function from the `leaps` library. Several *linmodpack* functions were used. The function `predrgn` makes the nonparametric prediction region and determines whether $\boldsymbol{x}_f$ is in the region. The function `predreg` also makes the nonparametric prediction region, and determines if $\mathbf{0}$ is in the region. For multiple linear regression, the function `regboot` does the residual bootstrap for multiple linear regression, `regbootsim` simulates the residual bootstrap for regression, and the function `rowboot` does the empirical nonparametric bootstrap. The function `vsbootsim` simulates the bootstrap for all subsets variable selection, so needs p small, while `vsbootsim2` simulates the prediction region method for forward selection. The functions `fselboot` and `vselboot` bootstrap the forward selection and all subsets variable selection estimators that minimize C_p. See Examples 4.7 and 4.8. The `shorth3` function computes the shorth(c) intervals with the Frey (2013) correction used when $g = 1$. Table 4.2 was made using `regbootsim3` for the OLS full model and `vsbootsim4` for forward selection. The functions `bicboot` and `bicbootsim` are useful if BIC is used instead of C_p. For forward selection

with C_p, the function `vscisim` was used to make Table 4.3, and can be used to compare the shorth, prediction region method, and Bickel and Ren CIs for β_i.

4.10 Problems

4.1. Consider the Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) listed below. Find shorth(7). Show work.

```
0.0  0.8  1.0  1.2  1.3  1.3  1.4  1.8  2.4  4.6
```

4.2. Find shorth(5) for the following data set. Show work.

```
6   76   90   90   94   94   95   97   97   1008
```

4.3. Find shorth(5) for the following data set. Show work.

```
66  76   90   90   94   94   95   95   97   98
```

4.4. Suppose you are estimating the mean θ of losses with the maximum likelihood estimator (MLE) $\overline{X}$ assuming an exponential (θ) distribution. Compute the sample mean of the fourth bootstrap sample.

actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$

bootstrap samples:

2, 10, 1, 2, 2: $\overline{X} = 3.4$

50, 10, 50, 2, 2: $\overline{X} = 22.8$

10, 50, 2, 1, 1: $\overline{X} = 12.8$

5, 2, 5, 1, 50: $\overline{X} =$?

4.5. The data below are a sorted residuals from a least squares regression where $n = 100$ and $p = 4$. Find shorth(97) of the residuals.

```
number     1     2     3     4   ...  97   98   99   100
residual -2.39 -2.34 -2.03 -1.77 ... 1.76 1.81 1.83  2.16
```

4.6. To find the sample median of a list of n numbers where n is odd, order the numbers from smallest to largest and the median is the middle ordered number. The sample median estimates the population median. Suppose the sample is $\{14, 3, 5, 12, 20, 10, 9\}$. Find the sample median for each of the three bootstrap samples listed below.

Sample 1: 9, 10, 9, 12, 5, 14, 3

Sample 2: 3, 9, 20, 10, 9, 5, 14

Sample 3: 14, 12, 10, 20, 3, 3, 5

4.7. Suppose you are estimating the mean μ of losses with $T = \overline{X}$.

actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$,

a) Compute $T_1^*, ..., T_4^*$, where T_i^* is the sample mean of the ith bootstrap sample. bootstrap samples:

2, 10, 1, 2, 2:

50, 10, 50, 2, 2:

10, 50, 2, 1, 1:

5, 2, 5, 1, 50:

b) Now compute the bagging estimator which is the sample mean of the T_i^*: the bagging estimator $\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ where $B = 4$ is the number of bootstrap samples.

4.8. Consider the output for Example 4.7 for the minimum C_p forward selection model.

a) What is $\hat{\boldsymbol{\beta}}_{I_{min}}$?

b) What is $\hat{\boldsymbol{\beta}}_{I_{min},0}$?

c) The large sample 95% shorth CI for H is [0,0.016]. Is H needed is the minimum C_p model given that the other predictors are in the model?

d) The large sample 95% shorth CI for $\log(S)$ is [0.324,0.913] for all subsets. Is $\log(S)$ needed is the minimum C_p model given that the other predictors are in the model?

e) Suppose $x_1 = 1$, $x_4 = H = 130$, and $x_5 = \log(S) = 5.075$. Find $\hat{Y} = (x_1\ x_4\ x_5)\hat{\boldsymbol{\beta}}_{I_{min}}$. Note that $Y = \log(M)$.

4.9^Q. Suppose $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$ where where $E(\boldsymbol{r}^W) = \boldsymbol{0}$ and $Cov(\boldsymbol{r}^W) = Cov(\boldsymbol{Y}^*) = MSE\ \boldsymbol{I}_n$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$. Recall that $\boldsymbol{X}$ is an $n \times p$ constant matrix. Simplify quantities when possible.

a) What is $E(\hat{\boldsymbol{\beta}}^*)$?

b) What is $Cov(\hat{\boldsymbol{\beta}}^*)$?

c) Recall that $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}$. What is $E(\hat{\boldsymbol{\beta}}_I^*) = E[(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^*]$?

d) What is $Cov(\hat{\boldsymbol{\beta}}_I^*)$?

4.10^Q. Suppose $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \sigma_n^2\boldsymbol{I}_n)$. Hence $Y_i^* = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \epsilon_i^P$ where $E(\epsilon_i^P) = 0$ and $V(\epsilon_i^P) = \sigma_n^2$. Hence $\boldsymbol{A}\boldsymbol{Y}^* \sim N_g(\boldsymbol{A}\boldsymbol{X}\hat{\boldsymbol{\beta}}, \sigma_n^2\boldsymbol{A}\boldsymbol{A}^T)$ if $\boldsymbol{A}$ is a $g \times n$ constant matrix. Recall that $\boldsymbol{X}$ is an $n \times p$ constant matrix. Simplify quantities when possible.

a) What is the distribution of $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$?

b) Using a), what is $E(\hat{\boldsymbol{\beta}}^*)$?

c) Recall that $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}$. What is the distribution of $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^*$ if $\hat{\boldsymbol{\beta}}_I^*$ is $k \times 1$?

4.11^Q. Suppose $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$ where $E(\boldsymbol{r}^W) = \boldsymbol{0}$ and $Cov(\boldsymbol{r}^W) = Cov(\boldsymbol{Y}^*) = diag(r_i^2) = diag(r_1^2, ..., r_n^2)$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ is the least squares estimator from regressing $\boldsymbol{Y}^*$ on $\boldsymbol{X}$, an $n \times p$ constant matrix. This model is used for the wild bootstrap. Simplify quantities when possible. (Can simplify a) and c), but can't simplify b) and d) much.)

a) What is $E(\hat{\boldsymbol{\beta}}^*)$?

b) What is $Cov(\hat{\boldsymbol{\beta}}^*)$?

c) Recall that $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{PY}$. What is $E(\hat{\boldsymbol{\beta}}_I^*) = E[(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^*]$?

d) What is $Cov(\hat{\boldsymbol{\beta}}_I^*)$?

4.12.

4.13.

4.14.

4.15.

4.16.

4.17.

4.18.

4.19.

4.20.

R Problems

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `linmodpack` function, e.g. *regbootsim2*, will display the code for the function. Use the `args` command, e.g. *args(regbootsim2)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into *R*.

4.21. a) Type the *R* command `predsim()` and paste the output into *Word*.

This program computes $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, diag(1,2,3,4))$ for $i = 1, ..., 100$ and $\boldsymbol{x}_f = \boldsymbol{x}_{101}$. One hundred such data sets are made, and ncvr, scvr, and mcvr count the number of times $\boldsymbol{x}_f$ was in the nonparametric, semiparametric, and parametric MVN 90% prediction regions. The volumes of the prediction regions are computed and voln, vols, and volm are the average ratio of the volume of the ith prediction region over that of the semiparametric region. Hence vols is always equal to 1. For multivariate normal data, these ratios should converge to 1 as $n \to \infty$.

b) Were the three coverages near 90%?

4.22. Consider the multiple linear regression model $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + e_i$ where $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The function `regbootsim2` bootstraps the regression model, finds bootstrap confidence intervals for β_i and a bootstrap confidence region for $(\beta_3, \beta_4)^T$ corresponding to the test $H_0 : \beta_3 = \beta_4 = 0$ versus H_A: not H_0. See the *R* code near Table 4.3. The lengths of the CIs along with the proportion of times the CI for β_i contained β_i are given. The fifth interval gives the length of the interval $[0, D_{(c)}]$ where H_0 is rejected if $D_0 > D_{(c)}$ and the fifth "coverage" is the proportion of times the test fails to reject H_0. Since nominal 95% CIs were used and the nominal

level of the test is 0.05 when H_0 is true, we want the coverages near 0.95. The CI lengths for the first 4 intervals should be near 0.392. The residual bootstrap is used.

Copy and paste the commands for this problem into *R*, and include the output in *Word*.