Math 584 Exam 3 is on Tuesday, April 27. You are allowed 20 sheets of notes and a calculator. CHECK FORMULAS! The Final is Tuesday, May 4: 2:45-4:45 with 30 sheets of notes.

85) If the MLR (multiple linear regression) model contains a constant, then SSTO = SSE + SSR where the total sum of squares (corrected for the mean) $SSTO = \sum_{i=1}^n (Y_i - \overline{Y})^2$, the regression sum of squares $SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ and error (or residual) sum of squares $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2$.

86) If the MLR model contains a constant, then the coefficient of multiple determination $R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \dfrac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \dfrac{\mathrm{SSE}}{\mathrm{SSTO}}$.

87) MLR output for the Anova F test $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ has the form shown below where under "Source," "Model" often replaces "Regression" and "Error" often replaces "Residual." Here MS = SS/df.

Source	df	SS	MS	F	p-value
Regression	p-1	SSR	MSR	Fo=MSR/MSE	for Ho:
Residual	n-p	SSE	MSE		$\beta_1 = \cdots = \beta_{p-1} = 0$

88) If the MLR model has a constant β_0, then

$$F_o = \frac{MSR}{MSE} = \frac{R^2}{1 - R^2} \ \ \frac{n-p}{p-1}.$$

89) R^2 does not decrease and usually increases as predictors are added to the linear model. Want $n \geq 10p$.

90) If $\hat{\boldsymbol{\theta}} \sim AN_p(\boldsymbol{\theta}, \boldsymbol{\Sigma}_n)$, then $\boldsymbol{a}^T\hat{\boldsymbol{\theta}} \sim AN_p(\boldsymbol{a}^T\boldsymbol{\theta}, \boldsymbol{a}^T\boldsymbol{\Sigma}_n\boldsymbol{a})$, and a large sample $100(1-\delta)\%$ confidence interval (CI) for $\boldsymbol{a}^T\boldsymbol{\theta}$ is $\boldsymbol{a}^T\hat{\boldsymbol{\theta}} \pm t_{d_n,1-\delta} SE(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{a}^T\hat{\boldsymbol{\theta}} \pm t_{d_n,1-\delta}\sqrt{\boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}_n\boldsymbol{a}}$ where $d_n \to \infty$ as $n \to \infty$.

91) For the full rank OLS model, $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. So $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} \sim AN_1(\boldsymbol{a}^T\boldsymbol{\beta},\ MSE\ \ \boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a})$, and a large sample $100(1-\delta)\%$ confidence interval (CI) for $\boldsymbol{a}^T\boldsymbol{\beta}$ is $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} \pm t_{n-p,1-\delta}\sqrt{MSE\ \boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}}$ where $P(T \leq t_{n-p,1-\delta}) = 1 - \delta$ if $T \sim t_{n-p}$.

92) A large sample $100(1-\delta)\%$ CI for β_i uses $\boldsymbol{a}^T = (0, ..., 0, 1, 0, ..., 0)$ where the 1 is in the $(i+1)$th position for $i = 0, ..., p-1$. Then $\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}$ is the $(i+1)$th diagonal element d_{ii} of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. So the CI is $\hat{\beta}_i \pm t_{n-p,1-\delta}\sqrt{MSE\ \ d_{ii}} = \hat{\beta}_i \pm t_{n-p,1-\delta} SE(\hat{\beta}_i)$.

93) Suppose there are k $100(1-\delta_S)\%$ CIs where $1-\delta_S$ is the confidence for a single confidence interval. Suppose we want the overall familywise confidence $1-\delta_T$ (probability before gathering the data and making the k CIs) that all k CIs contain their θ_i. Hence δ_T = P(at least one of the k CIs does not contain its θ_i, before gathering data). Then δ_T is called the familywise error rate. Can get procedures where $1 - \delta_T \geq 1 - \delta$.

94) Assume $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. i) The Bonferroni t intervals use $\delta_S = \delta/k$. Then $1 - \delta_T \geq 1 - \delta$.

ii) Scheffe's CIs for $\boldsymbol{a}^T\boldsymbol{\beta}$ have the form $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} \pm \sqrt{pF_{p,n-p,1-\delta}}\sqrt{MSE}\sqrt{\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}}$, and have $1 - \delta_T \geq 1 - \delta$.

95) Scheffe's CIs are longer than the corresponding Bonferroni CIs. Scheffe's CIs allow data snooping: you can decide on the $\boldsymbol{a}^T\boldsymbol{\beta}$ to use after getting the data. Usually need to decide on the $\boldsymbol{a}^T\boldsymbol{\beta}$ to use before gathering data for valid inference.

96) If the normality in 94) does not hold, then the large sample familywise confidence $1 - \delta_{T,n} \rightarrow 1 - \delta_T \geq 1 - \delta$ as $n \rightarrow \infty$ for a large class of 0 mean iid error distributions.

97) A large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is a set C_n such that $P(\boldsymbol{\theta} \in C_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$. For the full rank OLS model, a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\beta}$ is $C_n = \{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq MSE\ p\ F_{p,n-p,1-\delta}\} = \{\boldsymbol{\beta} : D^2_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}) \leq MSE\ p\ F_{p,n-p,1-\delta}\}$, a hyperellipsoid for $\boldsymbol{\beta}$ centered at $\hat{\boldsymbol{\beta}}$.

98) For the full rank OLS model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$, a large sample $100(1-\delta)\%$ CI for $E(Y) = E(Y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ is $\boldsymbol{x}^T\hat{\boldsymbol{\beta}} \pm t_{n-p,1-\delta}\sqrt{MSE}\sqrt{\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}} =$
$\hat{Y} \pm t_{n-p,1-\delta}\sqrt{MSE}\sqrt{h_{\boldsymbol{x}}}$. Want $h_{\boldsymbol{x}} \leq \max(h_1, ..., h_n)$, where $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$, to avoid extrapolation.

99) Consider predicting future test value Y_f given $\boldsymbol{x}_f$ and training data $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$ where $Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$ and $Y_f = \boldsymbol{x}_f\boldsymbol{\beta} + \epsilon_f$. For the full rank OLS model, $\epsilon_1, ..., \epsilon_n, \epsilon_f$ are iid and $Y_f \perp\!\!\!\perp Y_1, ..., Y_n$. Hence $Y_f \perp\!\!\!\perp \hat{Y}_f = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$ since $\hat{\boldsymbol{\beta}}$ is computed using the training data. Then $E(\hat{Y}_f - Y_f) = 0$ and $V(\hat{Y}_f - Y_f) = \sigma^2(1 + h_f)$ where $h_f = \boldsymbol{x}_f(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$ is the leverage of $\boldsymbol{x}_f$. Want $h_f \leq \max(h_1, ..., h_n)$ to avoid extrapolation.

100) A large sample $100(1-\delta)\%$ *prediction interval* (PI) has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \xrightarrow{P} 1 - \delta$ as the sample size $n \rightarrow \infty$. If the highest density region is an interval, then a PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage.

101) The length of a large sample CI goes to 0 while the length of a good PI goes to $U - L$ as $n \rightarrow \infty$, where $P(Y_f \in [L, U] | \boldsymbol{x}_f) \geq 1 - \delta$.

102) **Know:** 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) $= [\mathrm{Z_{(d)}}, \mathrm{Z_{(d+c-1)}}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$ correspond to the interval with the smallest distance.

103) Let $k_n = \lceil n(1-\delta) \rceil$ where $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Let $a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1 + h_f)}$. Apply the shorth$(c = k_n)$ estimator to the residuals $e_1, ..., e_n$: shorth$(c) = [e_{(d)}, e_{(d+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then a large sample 100 $(1-\delta)\%$ PI for Y_f is

$$[\hat{Y}_f + a_n\tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n\tilde{\xi}_{1-\delta_2}].$$

For the full rank OLS model, this PI is asymptotically optimal if the $\boldsymbol{x}_i$ are bounded in probability and the iid ϵ_i come from a large class of zero mean unimodal distributions.

104) For the full rank OLS model, the 100 $(1-\delta)\%$ classical PI for Y_f is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}\sqrt{MSE\ (1 + h_f)}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n - p$ degrees of freedom. Asymptotically, this PI estimates $[E(Y_f|\boldsymbol{x}_f) - \sigma Z_{1-\delta/2}, E(Y_f|\boldsymbol{x}_f) + \sigma Z_{1-\delta/2}]$, the interval between two quantiles of a $N(E(Y_f|\boldsymbol{x}_f), \sigma^2)$ distribution where $P(Z \leq Z_\alpha) = \alpha$ if $Z \sim$

$N(0,1)$. This PI may not perform well if the ϵ_i are not iid $N(0,\sigma^2)$ since the normal quantiles are not the correct quantiles for other error distributions.

105) One of the best ways to check the linear model is to make a **residual plot** of $\hat{Y}$ versus e and a **response plot** of $\hat{Y}$ versus Y with the identity line that has unit slope and zero intercept added as a visual aid. For multiple linear regression (MLR), assume the zero mean errors are iid from a unimodal distribution that is not highly skewed. If the iid constant variance MLR model is useful, then i) the plotted points in the response plot should scatter about the identity line with no other pattern, and ii) the plotted points in the residual plot should scatter about the $e = 0$ line with no other pattern. If either i) or ii) is violated, then the iid constant variance MLR model *is not sustained.* In other words, if the plotted points in the residual plot show some type of dependency, eg increasing variance or a curved pattern, then the MLR model may be inadequate.

106) Omitting important predictors, known as underfitting, can be a serious problem. Then for the multiple linear regression model, $\hat{\boldsymbol{\beta}}$ tends to be a biased estimator of $\boldsymbol{\beta}$, and leaving out important predictors could destroy the linearity of the model and could result in a model that has a nonconstant variance function. Let

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ x_{p-1} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \beta_{p-1} \end{pmatrix},$$

assume that $Y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$ is a good OLS model. Hence $E(Y|\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{\beta}_1^T\boldsymbol{x}_1 + \beta_{p-1}x_{p-1}$ and $V(Y|\boldsymbol{x}) = \sigma^2$. If x_{p-1} is omitted from the model, then $E(Y|\boldsymbol{x}_1) = \boldsymbol{\beta}_1^T\boldsymbol{x}_1 + \beta_{p-1}E(x_{p-1}|\boldsymbol{x}_1)$ and $V(Y|\boldsymbol{x}_1) = \sigma^2 + \beta_{p-1}^2 V(Y|\boldsymbol{x}_1)$. Note that linearity is destroyed if $E(x_{p-1}|\boldsymbol{x}_1)$ is nonlinear and the model has a nonconstant variance function if $V(Y|\boldsymbol{x}_1)$ is not constant and so depends on $\boldsymbol{x}_1$. On the other hand, if $E(x_{p-1}|\boldsymbol{x}_1) = \boldsymbol{\theta}^T\boldsymbol{x}_1$ and $V(x_{p-1}|\boldsymbol{x}_1) = \tau^2$, then $E(Y|\boldsymbol{x}_1) = \boldsymbol{\eta}^T\boldsymbol{x}_1$ is linear and $V(Y|\boldsymbol{x}_1) = \sigma^2 + \beta_{p-1}^2\tau^2 = \gamma^2$ is constant, where $\boldsymbol{\eta} = \boldsymbol{\beta}_1 + \beta_{p-1}\boldsymbol{\theta}$.

107) Suppose $x_1 = 1$ and $(Y, x_2, ..., x_{p-1})^T = (Y, \boldsymbol{w}^T)^T \sim$

$$N_p\left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\mu_w} \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\Sigma}_{Y,\boldsymbol{w}} \\ \boldsymbol{\Sigma}_{\boldsymbol{w},Y} & \boldsymbol{\Sigma}_{\boldsymbol{w}} \end{pmatrix} \right).$$

Then $Y|x_{i1}, ..., x_{ik}$ follows a linear model with constant variance: $Y_i = \beta_{0k} + \beta_{1k}x_{i1} + \cdots + \beta_{kk}x_{ik} + \epsilon_{ik}$ where $V(\epsilon_{ik}) = \sigma_k^2$. Models with lower σ_k^2 are better.

108) Can also get linear models with underfitting if the columns of $\boldsymbol{X}$ are orthogonal: predictors can be omitted without changing the $\hat{\beta}_i$ of the predictors that are in the model.

109) Having too many predictors, known as overfitting, is much less serious than omitting important predictors. The $\hat{\beta}_i$ for unneeded x_i tend to have $\hat{\beta}_i \xrightarrow{P} 0$. Suppose $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$ is the appropriate OLS model where $\boldsymbol{X}_1$ is an $n \times k$ matrix, $\boldsymbol{X} = (\boldsymbol{X}_1\ \boldsymbol{X}_2)$ is $n \times p$, and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{0} \end{pmatrix}$$

since $\boldsymbol{\beta}_2 = \boldsymbol{0}$. Consider overfitting by fitting the OLS model using $\boldsymbol{X}$ instead of $\boldsymbol{X}_1$. Then large sample inference is correct using $\boldsymbol{X}$, but not as precise as the model that omits predictors with $\beta_i = 0$ (the model that uses $\boldsymbol{X}_1$). For the overfitted model, R^2 is

too high, and CIs for β_i are longer using $\boldsymbol{X}$ than using $\boldsymbol{X}_1$ for $i = 1, ..., k$. Also want $n \geq 10p$ for the overfitted model and $n \geq 10k$ for the model using $\boldsymbol{X}_1$.

110) If $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ but $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$ instead of $\sigma^2\boldsymbol{I}$, then under regularity conditions $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$, but typically i) $\text{Cov}(\hat{\boldsymbol{\beta}}) \neq \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and ii) $E(MSE) \neq \sigma^2$. GLS can be used if $\boldsymbol{V}$ is known. A sandwich estimator can also be used to get a consistent estimator of $\text{Cov}(\hat{\boldsymbol{\beta}})$.

111) Outliers can often be found using the response and residual plots. The OLS fitted values (so identity line) will often go right through a cluster of gross outliers. Look for a gap separating the outliers from the bulk of the data. Fit OLS to the bulk of the data producing OLS estimator $\boldsymbol{b}$. Then make the response and residual plots for all of the data using $\hat{Y} = \boldsymbol{x}^T\boldsymbol{b}$. If the identity line still goes through the far away cluster, then it may be a cluster of "good leverage cases," otherwise the cases are likely outliers. Robust estimators attempt to automatically fit the bulk of the data well.

112) For **variable selection**, the model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \epsilon$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*. The **full model is always a submodel**.

113) 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. 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. 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. Be able to find model I_I from computer output.

114) **Forward selection** Step 1) $k = 1$: Start with a constant $w_1 = x_1$. Step 2) $k = 2$: Compute C_p for all models with $k = 2$ containing a constant and a single predictor x_i. Keep the predictor $w_2 = x_j$, say, that minimizes C_p.
Step 3) $k = 3$: Fit all models with $k = 3$ that contain w_1 and w_2. Keep the predictor w_3 that minimizes C_p. ...
Step j) $k = j$: Fit all models with $k = j$ that contains $w_1, w_2, ..., w_{j-1}$. Keep the predictor w_j that minimizes C_p. ...
Step p): Fit the full model.

Backward elimination: All models contain a constant $= u_1$. Step 0) $k = p$: Start with the full model that contains $x_1, ..., x_p$. We will also say that the full model contains $u_1, ..., u_p$ where $u_1 = x_1$ but u_i need not equal x_i for $i > 1$.
Step 1) $k = p - 1$: Fit each model with $k = p - 1$ predictors including a constant. Delete the predictor u_p, say, that corresponds to the model with the smallest C_p. Keep $u_1, ..., u_{p-1}$.
Step 2) $k = p - 2$: Fit each model with $p - 2$ predictors including a constant. Delete the predictor u_{p-1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-2}$. ...
Step j) $k = p - j$: fit each model with $p - j$ predictors including a constant. Delete the predictor u_{p-j+1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-j}$. ...
Step $p - 2$) $k = 2$. The current model contains u_1, u_2 and u_3. Fit the model u_1, u_2 and the model u_1, u_3. Assume that model u_1, u_2 minimizes C_p. Then delete u_3 and keep u_1 and u_2.

115) Can do all subsets variable selection for up to about 30 predictors. Criterion other than C_p, such as MSE(I) or $\overline{R}^2(I) = 1 - [1 - R^2(I)]\frac{n}{n-k}$ where model I contains k predictors, including a constant (adjusted R^2), can be used. C_p needs a good full model with $n \geq 10p$ or $5p$.

116) Collinearity occurs when at least one column of $\boldsymbol{X} = [\boldsymbol{v}_0, \boldsymbol{v}_1, ..., \boldsymbol{v}_{p-1}]$ is highly correlated with a linear combination of the other columns. Regress $\boldsymbol{v}_j$ on $\boldsymbol{v}_0, \boldsymbol{v}_1, ..., \boldsymbol{v}_{j-1}$, $\boldsymbol{v}_{j+1}, ..., \boldsymbol{v}_{p-1}$. Let R_j^2 be the coefficient of determination (squared multiple correlation coefficient) from the regression. The *variance inflation factor* $VIF_j = \dfrac{1}{1 - R_j^2}$. Let d_{jj} be given in 92). Then

$$SE(\hat{\beta}_j) = \sqrt{MSE} \;\; d_{jj} = \frac{\sqrt{MSE}\sqrt{VIF_j}}{SD(\boldsymbol{v}_j)\sqrt{n-1}}$$

where $SD(\boldsymbol{v}_j)$ is the sample standard deviation of the n elements of $\boldsymbol{v}_j$. Collinearity does not affect prediction much, provided that the software does not fail because $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular.

117) The ith residual $e_i = \epsilon_i + N_i$ where $N_i = \boldsymbol{x}_i^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \sim AN_1(0, MSE\ h_i) \xrightarrow{P} 0$ if $\max(h_1, ..., h_n) \to 0$ as $n \to \infty$.

118) In experimental design models or design of experiments (DOE), the entries of $\boldsymbol{X}$ are coded, often as $-1, 0$ or 1. Often $\boldsymbol{X}$ is not a full rank matrix.

119) Some DOE models have one Y_i per $\boldsymbol{x}_i$ and lots of $\boldsymbol{x}_i$'s. Then the response and residual plots are used like those for MLR.

120) Some DOE models have n_i Y_i's per $\boldsymbol{x}_i$, and only a few distinct values of $\boldsymbol{x}_i$. Then the response and residual plots no longer look like those for MLR.

121) A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i.

122) Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(y)$. A one way fixed effects ANOVA model has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2. The one way fixed effects normal ANOVA model is the special case where $Y|(W = a_i) \sim N(\mu_i, \sigma^2)$.

123) The *response plot* is a plot of $\hat{Y}$ versus Y. For the one way Anova model, the response plot is a plot of $\hat{Y}_{ij} = \hat{\mu}_i$ versus Y_{ij}. Often the identity line with unit slope and zero intercept is added as a visual aid. Vertical deviations from the identity line are the residuals $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i$. The plot will consist of p dot plots that scatter about the identity line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $Y_{i,1}, ..., Y_{i,n_i}$. Assume that each

$n_i \geq 10$. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

The *residual plot* is a plot of $\hat{Y}$ versus e where the residual $e = Y - \hat{Y}$. The plot will consist of p dot plots that scatter about the $e = 0$ line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $e_{i,1}, ..., e_{i,n_i}$. Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot than with the response plot.

124) Rule of thumb: Let R_i be the range of the ith dot plot $= \max(Y_{i1}, ..., Y_{i,n_i}) - \min(Y_{i1}, ..., Y_{i,n_i})$. If the $n_i \approx n/p$ and if $\max(R_1, ..., R_p) \leq 2\min(R_1, ..., R_p)$, then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable.

125) Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}. Be able to find $\hat{\mu}_i$ from data.

126) The **cell means model** for the fixed effects one way Anova is $Y_{ij} = \mu_i + \epsilon_{ij}$ where Y_{ij} is the value of the response variable for the jth trial of the ith factor level for $i = 1, ..., p$ and $j = 1, ..., n_i$. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The ϵ_{ij} are iid from the location family with pdf $f_Z(z)$, zero mean and unknown variance $\sigma^2 = V(Y_{ij}) = V(\epsilon_{ij})$. For the normal cell means model, the ϵ_{ij} are iid $N(0, \sigma^2)$. The estimator $\hat{\mu}_i = \overline{Y}_{i0} = \sum_{j=1}^{n_i} Y_{ij}/n_i = \hat{Y}_{ij}$. The ith residual is $e_{ij} = Y_{ij} - \overline{Y}_{i0}$, and $\overline{Y}_{00}$ is the sample mean of all of the Y_{ij} and $n = \sum_{i=1}^{p} n_i$. The total sum of squares SSTO $= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{00})^2$, the treatment sum of squares SSTR $= \sum_{i=1}^{p} n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2$, and the error sum of squares SSE = RSS $= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2$. The MSE is an estimator of σ^2. The Anova table is the same as that for multiple linear regression, except that SSTR replaces the regression sum of squares and that SSTO, SSTR and SSE have $n-1$, $p-1$ and $n-p$ degrees of freedom.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

127) Shown is a one way ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F." SSE is often replaced by RSS = residual sum of squares.

128) In matrix form, the cell means model is the linear model without an intercept (although $\mathbf{1} \in C(\boldsymbol{X})$), where $\boldsymbol{\mu} = \boldsymbol{\beta} = (\mu_1, ..., \mu_p)^T$, and $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{\epsilon} =$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{p,1} \\ \vdots \\ \epsilon_{p,n_p} \end{bmatrix}.$$

129) For the cell means model, $\boldsymbol{X}^T\boldsymbol{X} = diag(n_1, ..., n_p)$, $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = diag(1/n_1, ..., 1/n_p)$, and $\boldsymbol{X}^T\boldsymbol{Y} = (Y_{10}, ..., Y_{p0})^T$. So $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\mu}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T$. Then $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\mu}}$, and $\hat{Y}_{ij} = \overline{Y}_{i0}$. Hence the ijth residual $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \overline{Y}_{i0}$ for $i = 1, ..., p$ and $j = 1, ..., n_i$.

130) In the response plot, the dot plot for the jth treatment crosses the identity line at $\overline{Y}_{j0}$.

131) For the one way anova F test has hypotheses $H_0 : \mu_1 = \cdots = \mu_p$ and $H_A : not H_0$ (not all of the p population means are equal). The one way Anova table for this test is given above 127). Let RSS = SSE. The test statistic

$$F = \frac{MSTR}{MSE} = \frac{[RSS(H) - RSS]/(p-1)}{MSE} \sim F_{p-1,n-p}$$

if the ϵ_{ij} are iid $N(0, \sigma^2)$. If H_0 is true, then $Y_{ij} = \mu + \epsilon_{ij}$ and $\hat{\mu} = \overline{Y}_{00}$. Hence $RSS(H) = SSTO = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{00})^2$. Since $SSTO = SSE + SSTR$, the quantity $SSTR = RSS(H) - RSS$, and $MSTR = SSTR/(p-1)$.

132) The one way Anova F test is a large sample test if the ϵ_{ij} are iid with mean 0 and variance σ^2. Then the Y_{ij} come from the same location family with the same variance $\sigma_i^2 = \sigma^2$ and different mean μ_i for $i = 1, ..., p$. Thus the p treatments (groups, populations) have the same variance $\sigma_i^2 = \sigma^2$. The $V(\epsilon_{ij}) \equiv \sigma^2$ assumption (which implies that $\sigma_i^2 = \sigma^2$ for $i = 1, ..., p$) is a much stronger assumption for the one way Anova model than for MLR, but the test has some resistance to the assumption that $\sigma_i^2 = \sigma^2$ by 124).

133) Other design matrices $\boldsymbol{X}$ can be used for the full model. One design matrix adds a column of ones to the cell means design matrix. This model is no longer a full rank model.

134) A full rank one way Anova model with an intercept adds a constant but deletes the last column of the $\boldsymbol{X}$ for the cell means model. Then $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are as in the cell means model. Then $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^T = (\mu_p, \mu_1 - \mu_p, \mu_2 - \mu_p, ..., \mu_{p-1} - \mu_p)^T$. So $\beta_0 = \mu_p$ and $\beta_i = \mu_i - \mu_p$ for $i = 1, ..., p-1$.

It can be shown that the OLS estimators are $\hat{\beta}_0 = \overline{Y}_{p0} = \hat{\mu}_p$, and $\hat{\beta}_i = \overline{Y}_{i0} - \overline{Y}_{p0} = \hat{\mu}_i - \hat{\mu}_p$ for $i = 1, ..., p-1$. (The cell means model has $\hat{\beta}_i = \hat{\mu}_i = \overline{Y}_{i0}$.) In matrix form the model is

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{p,1} \\ \vdots \\ \epsilon_{p,n_p} \end{bmatrix}.$$

This model is interesting since the one way Anova F test of $H_0 : \mu_1 = \cdots = \mu_p$ versus H_A : not H_0 corresponds to the MLR Anova F test of $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ versus H_A : not H_0.

135) A contrast $\theta = \sum_{i=1}^{p} c_i \mu_i$ where $\sum_{i=1}^{p} c_i = 0$. The estimated contrast is $\hat{\theta} = \sum_{i=1}^{p} c_i \overline{Y}_{i0}$. Then $SE(\hat{\theta}) = \sqrt{MSE}\sqrt{\sum_{i=1}^{p} \frac{c_i^2}{n_i}}$ and a $100(1-\delta)\%$ CI for θ is $\hat{\theta} \pm t_{n-1,1-\delta/2} SE(\hat{\theta})$. CIs for one way Anova are less robust to the assumption that $\sigma_i^2 \equiv \sigma^2$ than the one way Anova F test.

136) Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $\theta_{ij} = \mu_i - \mu_j$ where $i \neq j$. The Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

Inference After Variable Selection

137) Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. A *model for variable selection* is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. 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). If $S \subseteq I$, 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, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. Note that $\boldsymbol{\beta}_E = \boldsymbol{0}$. Let $k_S = a_S - 1 =$ the number of population active nontrivial predictors. Then $k = a - 1$ is the number of active predictors in the candidate submodel I.

138) A simple method for inference after variable selection is **data splitting:** let the *training set* have $n_T \leq n/2$ cases and the *validation set* have $n_V = n - n_T \geq n/2$ cases. Select the n_T cases without replacement from the n cases. Assume the cases are

independent and follow a statistical model, e.g. MLR. i) Build model I with the training set, possibly using variable selection and the response to select predictors, predictor transformations, and the response transformation. ii) Act as if model I with k predictors is the full model for the validation set.

Want $n \geq 5k$ and preferably $n \geq 10k$. We need I to be a good model for the data. The efficiency is $\approx n_V/n = 1 - n_T/n$. Inefficient inference is much better than the invalid inference that results when I is built using all n cases and then treated as the full model on the same data set of n cases.

139) Let $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ be iid random vectors from a distribution with cdf F, mean $\boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{Z}) = \boldsymbol{\Sigma}$. Let $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ be the observed values of the $\boldsymbol{Z}_i$. The distribution of the random vector $\boldsymbol{w}$ is the *empirical distribution* if $\boldsymbol{w}$ is a discrete random vector with the following pmf. Then the sample mean and sample covariance matrix where $E(\boldsymbol{w}) = \overline{\boldsymbol{z}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i$ and $\text{Cov}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{z}_i - \overline{\boldsymbol{z}})(\boldsymbol{z}_i - \overline{\boldsymbol{z}})^T$.

$\boldsymbol{w}$	$\boldsymbol{z}_1$	$\boldsymbol{z}_2$	$\cdots$	$\boldsymbol{z}_n$
$P(\boldsymbol{w} = \boldsymbol{z})$	$1/n$	$1/n$	$\cdots$	$1/n$

140) The nonparametric **bootstrap** uses B bootstrap samples where a bootstrap sample is a sample of size n drawn with replacement from $x_1, ..., x_n$ (iid wrt the empirical distribution). Let $x_{i1}^*, ..., x_{in}^*$ denote the ith bootstrap sample. Let $T_i^* = g(x_{i1}^*, ..., x_{in}^*)$ be the statistic computed from $x_{i1}^*, ..., x_{in}^*$ for $i = 1, ..., B$. Be able to compute T_i^* for simple statistics such as the sample mean and sample median.

141) Let T_n be a $g \times 1$ statistic, e.g, $T_n = \hat{\boldsymbol{\beta}}$. If you had an iid sample $T_{1n}, ..., T_{Bn}$ you could figure out how the statistic behaves, but you only have $T_n = T_{1n}$. Under regularity conditions, if $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma})$, then $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma})$. So $\sqrt{n}(T_1^* - T_n), ..., \sqrt{n}(T_B^* - T_n)$ is pseudodata for $\sqrt{n}(T_{1n} - \boldsymbol{\theta}), ..., \sqrt{n}(T_{Bn} - \boldsymbol{\theta})$.

142)

I_j	model	x_2	x_3	x_4	x_5	$\hat{\boldsymbol{\beta}}_{I_j,0}$ if $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_j}$
I_2	1		*			$(\hat{\beta}_1, 0, \hat{\beta}_3, 0, 0)^T$
I_3	2		*	*		$(\hat{\beta}_1, 0, \hat{\beta}_3, \hat{\beta}_4, 0)^T$
I_4	3	*	*	*		$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, 0)^T$
I_5	4	*	*	*	*	$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_4)^T = \hat{\boldsymbol{\beta}}_{OLS}$

143) In 142) sometimes TRUE = * and FALSE = blank. The x_i may be replaced by the variable name or letters like a b c d.

I_j	model	x_2	x_3	x_4	x_5
I_2	1	FALSE	TRUE	FALSE	FALSE
I_3	2	FALSE	TRUE	TRUE	FALSE
I_4	3	TRUE	TRUE	TRUE	FALSE
I_5	4	TRUE	TRUE	TRUE	TRUE

144) Typical bootstrap output for forward selection, lasso, and elastic net is shown below. The SE column is usually omitted except possibly for forward selection. The term "coef" might be replaced by "Estimate." This column gives $\hat{\boldsymbol{\beta}}_{I,0}$ where $I = I_{min}$ for forward selection, $I = L$ for lasso, and $I = EN$ for elastic net. Note that the SE entry is omitted if $\hat{\beta}_i = 0$ so variable x_i was omitted by the variable selection method. In the

output below, $\hat{\beta}_2 = \hat{\beta}_3 = 0$. The SE column corresponds to the OLS SE obtained by acting as if the OLS full model contains a constant and the variables not omitted by the variable selection method. The OLS SE is incorrect unless the variables were selected before looking at the data for forward selection.

Label	Estimate or coef	SE	shorth 95% CI for β_i
Constant=intercept= x_1	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$[\hat{L}_1, \hat{U}_1]$
x_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$[\hat{L}_2, \hat{U}_2]$
x_3	0		$[\hat{L}_3, \hat{U}_3]$
x_4	0		$[\hat{L}_4, \hat{U}_4]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	$[\hat{L}_p, \hat{U}_p]$

145) We will consider the nonparametric bootstrap, parametric bootstrap, and residual bootstrap. Let T_n be a statistic, e.g. $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\overline{T}^* = \frac{1}{B}\sum_{i=1}^{n} \boldsymbol{T}_i^*$ and $\boldsymbol{S}_T^* = \frac{1}{B-1}\sum_{i=1}^{B}(\boldsymbol{T}_i^* - \overline{\boldsymbol{T}}^*)(\boldsymbol{T}_i^* - \overline{\boldsymbol{T}}^*)^T$. be the sample mean and sample covariance matrix of the bootstrap sample. For OLS, assume $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

146) Suppose the data set has n cases $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$, e.g, $\boldsymbol{z}_i = (Y_i, \boldsymbol{x}_i^T)^T$ and a statistic $T_n = T_n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$. The **nonparametric bootstrap** (naive, empirical, rowwise, pairwise bootstrap) draws a sample of size n with replacement from the n cases (from the empirical distribution of the cases). Let the ith bootstrap sample be $\boldsymbol{z}_{i1}^*, ..., \boldsymbol{z}_{in}^*$. Then $T_i^* = T(\boldsymbol{z}_{i1}^*, ..., \boldsymbol{z}_{in}^*)$ for $i = 1, ..., B$. The nonparametric bootstrap often works well if the cases are iid from some population. This assumption is very strong for MLR. For MLR, let $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T$. The nonparametric bootstrap has $\sqrt{n}(\hat{\boldsymbol{\beta}}_i^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V})$ if the $(Y_i, \boldsymbol{u}_i^T)^T$ are iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We can write $\boldsymbol{Y}_j^* = \boldsymbol{X}_j^*\boldsymbol{\beta} + \boldsymbol{\epsilon}_j^*$ for $j = 1, ..., B$ where $Y_{ij}^* = \boldsymbol{x}_{ij}^*\boldsymbol{\beta} + \epsilon_{ij}^*$. Hence $\boldsymbol{\epsilon}_j^*$ consists of the unknown ϵ_i sampled with replacement from $\epsilon_1, ..., \epsilon_n$ corresponding to the indices of the cases $(Y_i, \boldsymbol{x}_i^T)^T$ sampled with replacement for the jth bootstrap sample.

147) The **residual bootstrap** samples with replacement from the full model OLS residuals. $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ 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

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

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

Then $\hat{\boldsymbol{\beta}}_{I_j}^* = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $\text{Cov}(\hat{\boldsymbol{\beta}}_{I_j}^*) = \sigma_n^2(\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}_{I_j}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{P}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ since $\boldsymbol{P}\boldsymbol{X}_{I_j} = \boldsymbol{X}_{I_j}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant. It can be shown that $\sqrt{n}(\hat{\boldsymbol{\beta}}_i^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

148) The **parametric bootstrap** for MLR has $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2 \boldsymbol{I}) \sim N_n(\boldsymbol{PY}, \hat{\sigma}_n^2 \boldsymbol{I})$ where **we are not assuming** that the $\epsilon_i \sim N(0, \sigma^2)$, and $\hat{\sigma}_n^2 = MSE = \dfrac{1}{n-p}\sum_{i=1}^{n} r_i^2$ where the residuals are from the full OLS model. Thus $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^* \sim N_{a_I}(\hat{\boldsymbol{\beta}}_I, \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1})$ since $E(\hat{\boldsymbol{\beta}}_I^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{PY} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{PX}_I = \boldsymbol{X}_I$, and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I^*) = \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$. Hence

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

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

149) Note that for the residual bootstrap, $\hat{\sigma}_n^2 = (n-p)MSE/n$ while for the parametric bootstrap, $\hat{\sigma}_n^2 = MSE$.

150) Refer to 145). Consider $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\} \tag{1}$$

where $D^2_{(U_B)}$ is the $100q_B$th sample quantile (where $q_B \downarrow 1-\delta$) computed from $D_i^2 = (T_i^* - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. The corresponding test rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B,T)}\} \tag{2}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - T_n)$. The corresponding test rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$.

The hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\}. \tag{3}$$

The corresponding test rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

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) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Suppose $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned." Then

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

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

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

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

are well behaved. If $(n\boldsymbol{S}_T^*)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T\boldsymbol{\Sigma}_A^{-1}\boldsymbol{u}$. 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 (1), (2),

and (3) will have coverage near $1-\delta$. The regularity conditions for the prediction region method are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

151) A random vector $\boldsymbol{u}$ has a **mixture distribution** of random vectors $\boldsymbol{u}_j$ with probabilities π_j if $\boldsymbol{u}$ equals random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. Let $\boldsymbol{u}$ and $\boldsymbol{u}_j$ be $p \times 1$ random vectors. Then the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$$

where the probabilities π_j satisfy $0 \le \pi_j \le 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \ge 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of $\boldsymbol{u}_j$.

Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)] \text{ and } \mathrm{E}(\boldsymbol{u}) = \sum_{\mathrm{j}=1}^{\mathrm{J}} \pi_{\mathrm{j}} \mathrm{E}[\boldsymbol{u}_{\mathrm{j}}].$$

Hence $\mathrm{Cov}(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$
$\sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

$$\sum_{j=1}^{J} \pi_j \mathrm{Cov}(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T.$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$\mathrm{Cov}(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j \mathrm{Cov}(\boldsymbol{u}_j).$$

152) 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. 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.

153) 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. Then

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

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.

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

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{4}$$

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

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{5}$$

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.

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

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{6}$$

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.

154) **Geometric argument:** 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 $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$, and R_C is a $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$as $n, B \to \infty$ if $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. We also need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved (not too ill conditioned) for each $n \geq 20g$, say. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T^*_i - T_n)$ both converge in distribution to $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, say, 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. Then the hybrid region (3) is a confidence region by the geometric argument, and (1) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$.

155) By 153), the Geometric argument holds for iid $T_1, ..., T_B$ where $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$. 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 $\rho_{jn} \to \rho_j$ 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

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

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 $\boldsymbol{S}^*_{T,j}$. 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 $\boldsymbol{S}_{T,j}$.

156) For the residual, parametric and nonparametric bootstrap with C_p, the jth component of an iid sample $T_1, ..., T_B$ and the jth component of the bootstrap sample $T^*_1, ..., T^*_B$ have the same variability asymptotically. Since $E(T_{jn}) = \boldsymbol{\theta}$, each component of

the iid sample is centered at $\boldsymbol{\theta}$. Since $E(T_{jn}^*) = T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, the bootstrap components are centered at T_{jn}. Geometrically, separating the component clouds so that they are no longer centered at one value makes the overall data cloud larger. Thus the variability of T_n^* is larger than that of T_n for variable selection, asymptotically. Hence the prediction region applied to the bootstrap sample is slightly larger than the prediction region applied to the iid sample, asymptotically (we want $n \geq 20p$). Hence cutoff $\hat{D}^2_{1,1-\delta} = D^2_{(U_B)}$ gives coverage close to or higher than the nominal coverage for confidence regions (1) and (3), using the geometric argument. The deviation $T_i^* - T_n$ tends to be larger in magnitude than the deviations $\overline{T}^* - \boldsymbol{\theta}$, $T_n - \boldsymbol{\theta}$, and $T_i^* - \overline{T}^*$. Hence the cutoff $\hat{D}^2_{2,1-\delta} = D^2_{(U_B,T)}$ tends to be larger than $D^2_{(U_B)}$, and region (2) tends to have higher coverage than region (3) for a mixture distribution.

Lasso, Lasso Variable Selection, Ridge Regression, Elastic Net

157) Let $\boldsymbol{x}_i^T = (1 \;\; \boldsymbol{u}_i^T)$. It is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W} = (W_{ij})$. For $j = 1, ..., p-1$, let W_{ij} denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Then the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n}.$$

Then regression through the origin is used for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$ where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. Thus the centered response $Z_i = Y_i - \overline{Y}$ and $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Then $\hat{\boldsymbol{\eta}}$ does not depend on the units of measurement of the predictors. Linear combinations of the $\boldsymbol{u}_i$ can be written as linear combinations of the $\boldsymbol{x}_i$, hence $\hat{\boldsymbol{\beta}}$ can be found from $\hat{\boldsymbol{\eta}}$.

158) Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \lambda_{1,n}\sum_{i=1}^{p-1}|\eta_i|^j \tag{7}$$

where $\lambda_{1,n} \geq 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression, and $j = 1$ corresponds to lasso. In the literature, $Q(\boldsymbol{\eta})/c$ is often used, where $c = 2, n$, or $2n$ are common. The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$.

Lasso and ridge regression use a maximum value λ_M of λ and a grid of M λ values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{M-1} < \lambda_M$. For lasso, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ for $i < M$.

The elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{8}$$

where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ with $0 \leq \alpha \leq 1$.

159) Assume that the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}$ where $\boldsymbol{V}^{-1} = \boldsymbol{\rho}_{\boldsymbol{u}}$, the population correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$, if the $\boldsymbol{u}_i$ are a random sample from a population. If $\lambda_{1,n}/n \to 0$ then

$$\frac{\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}, \quad \text{and} \quad \mathrm{n}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \lambda_{1,\mathrm{n}}\boldsymbol{I}_{\mathrm{p-1}})^{-1} \xrightarrow{\mathrm{P}} \boldsymbol{V}.$$

Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$. By the OLS CLT, $\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

160) For ridge regression, $\hat{\boldsymbol{\eta}}_R =$

$$(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z} = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

$$[\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{B}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

$$\hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1n}}{n}n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

For the lasso estimator $\hat{\boldsymbol{\eta}}_L$:

$$\frac{-1}{n}\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_L) + \frac{\lambda_{1,n}}{2n}\boldsymbol{s}_n = \boldsymbol{0} \quad \text{or} \quad -\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_{\mathrm{L}}) + \frac{\lambda_{1,\mathrm{n}}}{2}\boldsymbol{s}_{\mathrm{n}} = \boldsymbol{0}$$

where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_{i,L})$ if $\hat{\eta}_{i,L} \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 1$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 1$. Note that $\boldsymbol{s}_n = \boldsymbol{s}_{n,\hat{\boldsymbol{\eta}}_L}$ depends on $\hat{\boldsymbol{\eta}}_L$. Thus $\hat{\boldsymbol{\eta}}_L$

$$= (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n.$$

By standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for the elastic net, $\hat{\boldsymbol{\eta}}_{EN}$ is optimal if

$$2\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{EN} - 2\boldsymbol{W}^T\boldsymbol{Z} + 2\lambda_1\hat{\boldsymbol{\eta}}_{EN} + \lambda_2\boldsymbol{s}_n = 0, \quad \text{or}$$

$$(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})\hat{\boldsymbol{\eta}}_{EN} = \boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \quad \text{or}$$

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_R - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{9}$$

Hence

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_1}{n}\, n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_2}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, \boldsymbol{s}_n$$

$$= \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, [\frac{\lambda_1}{n}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n].$$

Note that if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then $\hat{\lambda}_1/\sqrt{n} \xrightarrow{P} (1-\psi)\tau$ and $\hat{\lambda}_2/\sqrt{n} \xrightarrow{P} 2\psi\tau$. Under these conditions,

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - n(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_1\boldsymbol{I}_{p-1})^{-1}\, [\frac{\hat{\lambda}_1}{\sqrt{n}}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\hat{\lambda}_2}{2\sqrt{n}}\boldsymbol{s}_n].$$

161) The following theorem shows the elastic net, lasso, and ridge regression are asymptotically equivalent to the OLS full model if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$. Let $\hat{\boldsymbol{\eta}}_A$ be $\hat{\boldsymbol{\eta}}_{EN}$, $\hat{\boldsymbol{\eta}}_L$, or $\hat{\boldsymbol{\eta}}_R$. Note that c) follows from b) if $\psi = 0$, and d) follows from b) (using $2\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 2\tau$) if $\psi = 1$. Recall that we are assuming that p is fixed.

RR CLT, Lasso CLT, EN CLT: *Assume that the conditions of the OLS CLT hold for the model* $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2 \boldsymbol{V}).$$

b) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$, *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

c) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}).$$

d) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

162) Usually $\hat{\lambda}_{1,n}$ is selected using a criterion such as k–fold CV or GCV. It is not clear whether $\hat{\lambda}_{1,n} = o(n)$. For the elastic net and lasso, λ_M/n does not go to zero as $n \to \infty$ since $\hat{\boldsymbol{\eta}} = \boldsymbol{0}$ is not a consistent estimator. Hence λ_M is likely proportional to n, and using $\lambda_i = i\lambda_M/M$ for $i = 1, ..., M$ will not produce a consistent estimator.

163) Lasso and elastic net can be regarded as methods for variable selection: often some of the $\hat{\beta}_i = 0$. Let the **active set** be the set of x_i that have nonzero $\hat{\beta}_i$. The relaxed lasso estimator is OLS applied to the lasso active set while the relaxed elastic net estimator is OLS applied to the elastic net active set. If $\hat{\lambda}_{1n}/\sqrt{n} \to \tau > 0$, then lasso tends to have at least one $\hat{\beta}_j = 0$ for large n. Lasso may not be $\sqrt{n}$ consistent if lasso selects S with high probability, but then relaxed lasso tends to be $\sqrt{n}$ consistent.

Let I_{min} be the lasso or elastic net active set. Expect relaxed lasso and relaxed elastic net perform better than lasso and elastic net unless $(\boldsymbol{X}_{I_{min}}^T \boldsymbol{X}_{I_{min}})^{-1}$ is ill conditioned.

If $P(S \subseteq I) \to 1$ as $n \to \infty$, then relaxed lasso and relaxed elastic net have a CLT given by 153).

MREG

164) The **multivariate linear regression model**

$$\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$ where $x_1 \equiv 1$ is the trivial predictor. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (1, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$ where the 1 could be omitted. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{e}_i$ is defined below. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are to be estimated,

and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid. The data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Y}]$ except usually the first column $\mathbf{1}$ of $\boldsymbol{X}$ is omitted. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ design matrix of predictor variables is

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where $\boldsymbol{v}_1 = \mathbf{1}$. The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \dots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Considering the ith row of $\boldsymbol{Z}, \boldsymbol{X}$ and $\boldsymbol{E}$ shows that $\boldsymbol{y}_i^T = \boldsymbol{x}_i^T \boldsymbol{B} + \boldsymbol{\epsilon}_i^T$.

165) We have changed notation for multiple linear regression, using $\boldsymbol{e}$ and $\boldsymbol{\epsilon}$ for errors and $\hat{\boldsymbol{\epsilon}}$, $\boldsymbol{r}$, and $\hat{\epsilon}_{ij}$ for residuals. For the *multiple linear regression model*, $m = 1$ and

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{10}$$

for $i = 1, \dots, n$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

166) Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix $\boldsymbol{X}$** of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$ and error vector $\boldsymbol{e}_j$ change and thus depend on j.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)$ which corresponds to the ith row of $\boldsymbol{Z}$ and the ith row of $\boldsymbol{X}$. Then

$$\left[\begin{array}{c} Y_{i1} = \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} + \epsilon_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \epsilon_{i1} \\ Y_{i2} = \beta_{12}x_{i1} + \cdots + \beta_{p2}x_{ip} + \epsilon_{i2} = \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \epsilon_{i2} \\ \vdots \\ Y_{im} = \beta_{1m}x_{i1} + \cdots + \beta_{pm}x_{ip} + \epsilon_{im} = \boldsymbol{x}_i^T\boldsymbol{\beta}_m + \epsilon_{im} \end{array} \right]$$

or, suppressing the condition $\boldsymbol{y}_i|\boldsymbol{x}_i$, we have $\boldsymbol{y}_i = \boldsymbol{\mu}_{\boldsymbol{x}_i} + \boldsymbol{\epsilon}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{\mu}_{\boldsymbol{x}_i} = \boldsymbol{B}^T\boldsymbol{x}_i = \left[\begin{array}{c} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_m \end{array} \right].$$

167) The **least squares estimators** are

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \left[\begin{array}{cccc} \hat{\boldsymbol{\beta}}_1 & \hat{\boldsymbol{\beta}}_2 & \ldots & \hat{\boldsymbol{\beta}}_m \end{array} \right].$$

The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\begin{array}{cccc} \hat{\boldsymbol{Y}}_1 & \hat{\boldsymbol{Y}}_2 & \ldots & \hat{\boldsymbol{Y}}_m \end{array} \right].$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\left[\begin{array}{c} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{array} \right] = \left[\begin{array}{cccc} \boldsymbol{r}_1 & \boldsymbol{r}_2 & \ldots & \boldsymbol{r}_m \end{array} \right] = \left[\begin{array}{cccc} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{array} \right].$$

These quantities can be found from the m multiple linear regressions of Y_j on the predictors: $\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\widehat{\boldsymbol{Y}}_j = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j$ and $\boldsymbol{r}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} =$

$$\frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$, since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\boldsymbol{0}$. Let $\hat{\boldsymbol{\Sigma}}_\epsilon = \hat{\boldsymbol{\Sigma}}_{\epsilon,p}$ be the unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Also,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = (n-d)^{-1}\boldsymbol{Z}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z},$$

and

$$\hat{\boldsymbol{E}} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z}.$$

168)**Theorem**: Suppose $\boldsymbol{X}$ has full rank $p < n$ and the covariance structure of 164) holds. Then $E(\hat{\boldsymbol{B}}) = \boldsymbol{B}$ so $E(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\beta}_j$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k) = \sigma_{jk}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ for $j, k = 1, ..., p$. Also $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}}$ are uncorrelated, $E(\hat{\boldsymbol{E}}) = \boldsymbol{0}$ and

$$E(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = E\left(\frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p}\right) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}.$$

Also, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is a $\sqrt{n}$ consistent estimator of $\Sigma_{\boldsymbol{\epsilon}}$ under mild regularity conditions.

169) A **response plot** for the jth response variable is a plot of the fitted values $\hat{Y}_{ij}$ versus the response Y_{ij}. The identity line with slope one and zero intercept is added to the plot as a visual aid. A **residual plot** corresponding to the jth response variable is a plot of $\hat{Y}_{ij}$ versus r_{ij} where $i = 1, ..., n$. Make the m response and residual plots for any multivariate linear regression. For each response variable Y_{ij}, the response and residual plots behave just as they do for MLR.

170) Let the observed multivariate data $\boldsymbol{w}_i$ for $i = 1, ..., n$ be collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{w}_1^T, ..., \boldsymbol{w}_n^T$. 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 such as the sample covariance matrix. The ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D_{\boldsymbol{w}_i}^2(T, \boldsymbol{C}) = (\boldsymbol{w}_i - T)^T\boldsymbol{C}^{-1}(\boldsymbol{w}_i - T)$$

for each point $\boldsymbol{w}_i$.

171) The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

$$T = \overline{\boldsymbol{W}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}_i, \quad \text{and} \quad \boldsymbol{C} = \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{W}})(\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{W}})^{\mathrm{T}}$$

and will be denoted by MD_i. When T and $\boldsymbol{C}$ are estimators other than the sample mean and covariance, $D_i = \sqrt{D_i^2}$ will sometimes be denoted by RD_i. Then the **DD plot** is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i. The identity line is added as a visual aid. If n is large and the $\boldsymbol{w}_i$ are iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the data will cluster tightly about the identity line. Tight clustering about a line through the origin that is not the identity line suggests that the $\boldsymbol{w}_i$ are iid from an elliptically contoured distribution that is not multivariate normal.

172) A *large sample* $(1-\delta)100\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \to 1 - \delta$ as $n \to \infty$, and is *asymptotically optimal* if the volume of the region converges in probability to the volume of the population minimum volume covering region.

173) For multivariate linear regression, the classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and past data $(\boldsymbol{x}_1, \boldsymbol{y}_i), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ is $\{\boldsymbol{y} : D_{\boldsymbol{y}}^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \chi_{m,1-\delta}^2\}$ and does not work well unless the $\boldsymbol{\epsilon}_i$ are iid $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

174) Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

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

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. The large sample *nonparametric prediction region* that works for a large class of error vector distributions is $\{\boldsymbol{y} : D^2_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D^2_{(U_n)}\}$ where $D_{(U_n)}$ is the q_nth sample quantile of the $D_i = D_{\hat{\boldsymbol{z}}_i}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = D_{\hat{\boldsymbol{\epsilon}}_i}(\boldsymbol{0}, \boldsymbol{S}_r)$. In the DD plot, the cases to the left of the vertical line $MD = D_{(U_n)}$ correspond to $\boldsymbol{y}_i$ that are in their nonparametric prediction region when $\boldsymbol{x}_f = \boldsymbol{x}_i$. Hence $100q_n\%$ of the training data are in their prediction region, and $100q_n\% \rightarrow 100(1-\delta)\%$ as $n \rightarrow \infty$.

175) Consider testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. The *error or residual sum of squares and cross products matrix* is

$$\boldsymbol{W}_e = (\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{Z}^T[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Z}.$$

Note that $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$.

176) Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The *Roy's maximum root statistic* is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

The *Wilks' Λ statistic* is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m}(1+\lambda_i)^{-1}$.

The *Pillai's trace statistic* is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$.

The *Hotelling-Lawley trace statistic* is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\lambda_i$.

177) Let matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\; \ldots \;\; \boldsymbol{a}_p]$. Then the vec operator stacks the columns of $\boldsymbol{A}$ on top of one another so

$$vec(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix}.$$

Let $\boldsymbol{A} = (a_{ij})$ be an $m \times n$ matrix and $\boldsymbol{B}$ a $p \times q$ matrix. Then the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the $mp \times nq$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$. The following assumption is important.

178) **Theorem**: *The Hotelling-Lawley trace statistic*

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

179) **Assumption D1**: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max(h_1, ..., h_n) \to 0$ as $n \to \infty$, assume that the zero mean iid error vectors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

180) **Multivariate Least Squares Central Limit Theorem (MLS CLT):** For the least squares estimator, if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is $\sqrt{n}$ consistent and $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

181) **Theorem**: *If assumption D1 holds and if H_0 is true, then* $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

182) Consider testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. Assume the error distribution is multivariate normal $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Then under H_0,

$$[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \sim \chi^2_{rm},$$

and

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \tag{11}$$

The above equation also holds if the $\boldsymbol{\epsilon}_i$ are iid for a large class of distributions. A large sample level α test will reject H_0 if $pval < \alpha$ where

$$pval = P(\frac{T}{rm} < F_{rm,n-mp}). \tag{12}$$

183) Theorems 178) and 181) are useful for relating multivariate tests with the partial F_R test for multiple linear regression that tests whether a reduced model that omits some of the predictors can be used instead of the full model that uses all p predictors.

$$F_R = \frac{[\boldsymbol{L}\hat{\boldsymbol{\beta}}]^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}[\boldsymbol{L}\hat{\boldsymbol{\beta}}]}{r\hat{\sigma}^2}$$

is distributed as $F_{r,n-p}$ if H_0 is true and the errors are iid $N(0, \sigma^2)$. Note that for multiple linear regression with $m = 1$, $F_R = (n-p)U(\boldsymbol{L})/r$ since $\hat{\boldsymbol{\Sigma}}_{\epsilon}^{-1} = 1/\hat{\sigma}^2$. Hence the scaled Hotelling Lawley test statistic is the partial F test statistic extended to $m > 1$ predictor variables by Theorem 178).

184) By Theorem 181), for example, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of nonnormal error distribution. If $Z_n \sim F_{k,d_n}$, then $Z_n \xrightarrow{D} \chi^2_k/k$ as $d_n \to \infty$. Hence using the $F_{r,n-p}$ approximation gives a large sample test with correct asymptotic level, and the partial F test is robust to nonnormality.

Similarly, using an $F_{rm,n-pm}$ approximation for the following test statistics gives large sample tests with correct asymptotic level and similar power for large n. The large sample test will have correct asymptotic level as long as the denominator degrees of freedom $d_n \to \infty$ as $n \to \infty$, and $d_n = n - pm$ reduces to the partial F test if $m = 1$ and $U(\boldsymbol{L})$ is used. Then the three test statistics are

$$\frac{-[n-p-0.5(m-r+3)]}{rm}\ \log(\Lambda(\boldsymbol{L})),\ \ \frac{n-p}{rm}\ V(\boldsymbol{L}),\ \text{and}\ \ \frac{\mathrm{n-p}}{\mathrm{rm}}\ \mathrm{U}(\boldsymbol{L}).$$

it can be shown that

$$V(\boldsymbol{L}) \leq -\log(\Lambda(\boldsymbol{L})) \leq U(\boldsymbol{L}).$$

Hence the Hotelling Lawley test will have the most power and Pillai's test will have the least power.

185) Under regularity conditions, $-[n-p+1-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

These statistics are robust against nonnormality.

For the Wilks' Lambda test,

$$pval = P\left(\frac{-[n-p+1-0.5(m-r+3)]}{rm} \ \log(\Lambda(\boldsymbol{L})) < F_{rm,n-rm}\right).$$

For the Pillai's trace test, $pval = P\left(\frac{n-p}{rm} \ V(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

For the Hotelling Lawley trace test, $pval = P\left(\frac{n-p}{rm} \ U(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

The above three tests are large sample tests, P(reject $H_0|H_0$ is true) $\rightarrow \alpha$ as $n \rightarrow \infty$, under regularity conditions.

186) The 4 step MANOVA F test of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0} \ \ \boldsymbol{I}_{p-1}]$:
i) State the hypotheses H_0: the nontrivial predictors are not needed in the mreg model
H_1: at least one of the nontrivial predictors is needed.
ii) Find the test statistic F_o from output.
iii) Find the pval from output.
iv) If pval $< \alpha$, reject H_0. If pval $\geq \alpha$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, ..., Y_m$ and the predictors X_2, ..., X_p. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, ..., Y_m$ and the predictors X_2, ..., X_p. (Get the variable names from the story problem.)

187) The 4 step F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ where the 1 is in the jth position. Let $\boldsymbol{b}_j^T$ be the jth row of $\boldsymbol{B}$. The hypotheses are equivalent to $H_0: \ \boldsymbol{b}_j^T = \boldsymbol{0}$
$H_1: \ \ \boldsymbol{b}_j^T \neq \boldsymbol{0}$. This test is a test for whether x_j is needed in the model.
i) State the hypotheses
H_0: x_j is not needed in the model H_1: x_j is needed in the model.
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $< \alpha$, reject H_0. If pval $\geq \alpha$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that X_j is needed in the mreg model for $Y_1, ..., Y_m$. If you fail to reject H_0, then conclude that X_j is not needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model.

The statistic

$$F_j = \frac{1}{d_j}\hat{\boldsymbol{B}}_j^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{B}}_j = \frac{1}{d_j}(\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jm})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \begin{pmatrix} \hat{\beta}_{j1} \\ \hat{\beta}_{j2} \\ \vdots \\ \hat{\beta}_{jm} \end{pmatrix}$$

where $\hat{\boldsymbol{B}}_j^T$ is the jth row of $\hat{\boldsymbol{B}}$ and $d_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The statistic F_j could be used for forward selection and backward elimination in variable selection.

188) The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. The ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $X_2, ..., X_p$ corresponds to the MANOVA F test.
i) State the hypotheses H_0: the reduced model is good
H_1: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $< \alpha$, reject H_0 and conclude that the full model should be used.
If pval $\geq \alpha$, fail to reject H_0 and conclude that the reduced model is good.

189) The 4 step MANOVA F test should reject H_0 if the response and residual plots look good, n is large enough and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small.

190) Recall the population OLS coefficients and second way to compute $\hat{\boldsymbol{\beta}}$ from 74) and 75). Similar results will hold for multivariate linear regression. Let $\boldsymbol{y} = (Y_1, ..., Y_m)^T$, let $\boldsymbol{w} = (x_2, ..., x_p)^T$, let $\hat{\boldsymbol{\beta}}_j = (\hat{\alpha}_j, \hat{\boldsymbol{\eta}}_j^T)^T$ where $\hat{\alpha}_j = \overline{Y}_j - \hat{\boldsymbol{\eta}}_j^T\overline{\boldsymbol{w}}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{wy}} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T$ which has jth column $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$ for $j = 1, ..., m$. Let

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{w} \end{pmatrix}, \quad E(\boldsymbol{u}) = \boldsymbol{\mu}_{\boldsymbol{u}} = \begin{pmatrix} E(\boldsymbol{y}) \\ E(\boldsymbol{w}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{y}} \\ \boldsymbol{\mu}_{\boldsymbol{w}} \end{pmatrix}, \quad \text{and} \quad \text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{yy}} & \boldsymbol{\Sigma}_{\boldsymbol{yw}} \\ \boldsymbol{\Sigma}_{\boldsymbol{wy}} & \boldsymbol{\Sigma}_{\boldsymbol{ww}} \end{pmatrix}.$$

Let the vector of constants be $\boldsymbol{\alpha}^T = (\alpha_1, ..., \alpha_m)$ and the matrix of slope vectors $\boldsymbol{B}_S = \begin{bmatrix} \boldsymbol{\eta}_1 & \boldsymbol{\eta}_2 & \cdots & \boldsymbol{\eta}_m \end{bmatrix}$. Then the population least squares coefficient matrix is

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{B}_S \end{pmatrix}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{B}_S^T\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\boldsymbol{B}_S = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{wy}}$ where $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{\Sigma}_{\boldsymbol{ww}}$.

If the $\boldsymbol{u}_i$ are iid with nonsingular covariance matrix Cov($\boldsymbol{u}$), the least squares estimator

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}^T \\ \hat{\boldsymbol{B}}_S \end{pmatrix}$$

where $\hat{\boldsymbol{\alpha}} = \overline{\boldsymbol{y}} - \hat{\boldsymbol{B}}_S^T\overline{\boldsymbol{w}}$ and $\hat{\boldsymbol{B}}_S = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{wy}}$. The least squares multivariate linear regression estimator can be calculated by computing the classical estimator $(\overline{\boldsymbol{u}}, \boldsymbol{S}_{\boldsymbol{u}}) = (\overline{\boldsymbol{u}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}})$ of multivariate location and dispersion on the $\boldsymbol{u}_i$, and then plug in the results into the formulas for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}_S$.

191)	Multiple Linear Regression	Multivariate Linear Regression
	$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$	$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$
1)	$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$	$E[\boldsymbol{Z}] = \boldsymbol{X}\boldsymbol{B}$
2)	$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$	$\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$
3)	$E(\boldsymbol{e}) = \boldsymbol{0}$	$E[\boldsymbol{E}] = \boldsymbol{0}$
4)	$\boldsymbol{H} = \boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$	$\boldsymbol{H} = \boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
5)	$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$	$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}$
6)	$\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$	$\widehat{\boldsymbol{Z}} = \boldsymbol{P}\boldsymbol{Z}$
7)	$\boldsymbol{r} = \hat{\boldsymbol{e}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	$\widehat{\boldsymbol{E}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Z}$
8)	$E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$	$E[\widehat{\boldsymbol{B}}] = \boldsymbol{B}$
9)	$E(\widehat{\boldsymbol{Y}}) = E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$	$E[\widehat{\boldsymbol{Z}}] = \boldsymbol{X}\boldsymbol{B}$
10)	$\hat{\sigma}^2 = \frac{\boldsymbol{r}^T\boldsymbol{r}}{n-p}$	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p}$
11)	$V(e_i) = \sigma^2$	$\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$
12)	$E(Y_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i$	$E[\boldsymbol{y}_i] = \boldsymbol{B}^T\boldsymbol{x}_i$
13)	$H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ $rF_R \xrightarrow{D} \chi^2_r$	$H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{0}$ $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$
14)	LS CLT $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{W})$	MLS CLT $\sqrt{n}\ \ vec(\widehat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.