Math 583 Exam 2 is on Wednesday, Nov. 1 and covers homeworks 5-8 and quizzes 5-8. You are allowed 9 sheets of notes and a calculator. Any needed tables will be provided. CHECK FORMULAS: YOU ARE RESPONSIBLE FOR ANY ERRORS ON THIS HANDOUT!

60) A random vector $\boldsymbol{u}$ has a *mixture distribution* if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if 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 $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j.

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

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

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j],$$

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

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

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

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

62) 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 the same probabilities π_{kn} of the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the I_k are randomly selected.

63) **Theorem.** 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

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

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$.

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

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{1}$$

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 $\boldsymbol{\beta}$: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

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.

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

65) Use the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \leq t_1, ..., z_p \leq t_p)$ by $P(\boldsymbol{z} \leq \boldsymbol{t})$. Then

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \leq \boldsymbol{t}] = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

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

66) **Theorem.** 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

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

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$.

67) Data splitting divides the data into two sets: the training set (or modeling set) H that has n_H cases, and the validation set V that has $n - n_H = n_V$ cases. For regression, assume the cases are independent. Build a model I_H using a predictors using the cases in H. During the building process, you can examine the data (use the response to build the model) and you can use variable selection.
Then fit the model I_H using only the cases in V. Assuming $n \geq Ja$ (with $J \geq 5$ or 10 etc.), perform the usual model checks (such as a response plot) and the usual model inference.

Pros: with data splitting, you can look at the data for the cases in H, use data splitting for high dimensional data, and perform standard inference. (With all n cases, a) using the response to build a model invalidates inference, and b) variable selection inference is complicated.) Also standard theory applies to the n_V cases, hence often iid cases are not needed.

Drawbacks: a) Since n_V cases are used, there is a loss of efficiency compared to using all n cases if I_H could have be chosen without looking at the response or using variable selection. b) Models that are much better than I_H may exist, especially in high dimensions.

68) Let $Q(\boldsymbol{\eta})$ be a real valued function of the $k \times 1$ vector $\boldsymbol{\eta}$. The gradient of $Q(\boldsymbol{\eta})$ is the $k \times 1$ vector

$$\bigtriangledown Q = \bigtriangledown Q(\boldsymbol{\eta}) = \frac{\partial Q}{\partial \boldsymbol{\eta}} = \frac{\partial Q(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \frac{\partial}{\partial \eta_1} Q(\boldsymbol{\eta}) \\ \frac{\partial}{\partial \eta_2} Q(\boldsymbol{\eta}) \\ \vdots \\ \frac{\partial}{\partial \eta_k} Q(\boldsymbol{\eta}) \end{bmatrix}.$$

Suppose there is a model with unknown parameter vector $\boldsymbol{\eta}$. A set of *estimating equations* $f(\boldsymbol{\eta})$ is used to maximize or minimize $Q(\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ is a dummy variable vector.

As a mnemonic (memory aid) for the following theorem, note that the derivative $\frac{d}{dx} ax = \frac{d}{dx} xa = a$ and $\frac{d}{dx} ax^2 = \frac{d}{dx} xax = 2ax$.

Theorem. a) If $Q(\boldsymbol{\eta}) = \boldsymbol{a}^T \boldsymbol{\eta} = \boldsymbol{\eta}^T \boldsymbol{a}$ for some $k \times 1$ constant vector $\boldsymbol{a}$, then $\bigtriangledown Q = \boldsymbol{a}$.

b) If $Q(\boldsymbol{\eta}) = \boldsymbol{\eta}^T \boldsymbol{A} \boldsymbol{\eta}$ for some $k \times k$ constant symmetric matrix $\boldsymbol{A}$, then $\bigtriangledown Q = 2\boldsymbol{A}\boldsymbol{\eta}$.

c) If $Q(\boldsymbol{\eta}) = \sum_{i=1}^k |\eta_i| = \|\boldsymbol{\eta}\|_1$, then $\bigtriangledown Q = \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$ where $s_i = \text{sign}(\eta_i)$ where $\text{sign}(\eta_i) = 1$ if $\eta_i > 0$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 0$. This gradient is only defined for $\boldsymbol{\eta}$ where none of the k values of η_i are equal to 0.

69) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} =$
$\{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ where $0 < \delta < 1$. A large sample $100(1-\delta)\%$ confidence region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$. A prediction interval (PI) $[L_n, U_n]$ is a special case of a prediction region and a confidence interval (CI) $[L_n, U_n]$ is a special case of a confidence region. (We often want the probability to $\xrightarrow{P} 1 - \delta$.)

70) Let $D_i^2 = D^2_{\boldsymbol{x}_i} = D^2_{\boldsymbol{x}_i}(\overline{\boldsymbol{x}}, \boldsymbol{S}) = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})$ where $\boldsymbol{S} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ is the sample covariance matrix. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and $q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n)$, otherwise. If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i. The large sample $100(1-\delta)\%$ *nonparametric prediction region* is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(U_n)}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}$. This prediction region needs $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ to be nonsingular and the $\boldsymbol{x}_i, \boldsymbol{x}_f$ iid for $i = 1, ..., n$.

71) Data splitting divides the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into two sets: H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $(T_H, \boldsymbol{C}_H)$ is computed using the data set H. Then the squared validation distances $D_j^2 = D^2_{\boldsymbol{x}_{i_j}}(T_H, \boldsymbol{C}_H) = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1} (\boldsymbol{x}_{i_j} - T_H)$ are computed for the $j = 1, ..., n_V$ cases in the validation set V. Let $D^2_{(U_V)}$ be the U_Vth order statistic of the D_j^2 where

$$U_V = \min(n_V, \lceil (n_V + 1)(1 - \delta) \rceil).$$

The large sample $100(1-\delta)\%$ *data splitting prediction region* for $\boldsymbol{x}_f$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(T_H, \boldsymbol{C}_H) \leq D^2_{(U_V)}\}.$$

This prediction region can be used for high dimensional data if $\boldsymbol{C}_H$ is nonsingular, e.g. if $\boldsymbol{C}_H = \boldsymbol{I}_p$. The "coverage" is bounded below by $U_V/(n_V+1)$ (with exact coverage for $\boldsymbol{x}_i$ from a continuous multivariate distribution). The coverage is no larger than $n_V/(n_V+1)$ which can be low for small n_V. Prediction regions with much smaller volume may exist.

n_V	δ	U_V	$U_V/(n_V+1)$
1	1/2	1	1/2
2	1/3	2	2/3
19	0.05	19	0.95
20	0.05	20	0.9594
99	0.05	95	0.95
100	0.05	96	0.95
600	0.95	571	0.9501

72) In low dimensions with variable selection, possibly with data splitting, if model I is selected, often $\boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y) = \boldsymbol{\beta}_F(\boldsymbol{x}, Y)$ where F is the full model and $(\boldsymbol{x}_I, Y)$ means regress Y on $\boldsymbol{x}_I$. Hence $\hat{\boldsymbol{\beta}}_{I,0} \xrightarrow{P} \boldsymbol{\beta}_F$, i.e. $\hat{\boldsymbol{\beta}}_{I,0}$ is a consistent estimator of $\boldsymbol{\beta}_F$. Assume I is sparse with $\boldsymbol{\beta}_I$ an $a \times 1$ vector with $n \geq Ja$ and $J \geq 10$ or 20 etc. Check the model with the usual checks such as the response plot. Perform the usual inference on the model applied to the cases in the validation set V. The following table shows what $\hat{\boldsymbol{\beta}}_I$ or $\hat{\boldsymbol{\beta}}_{I,0}$ is estimating, along with errors that are common in the high dimensional literature. OPLS and MMLE need iid cases. For iid cases, $\boldsymbol{\beta}_{I,OLS}(\boldsymbol{x}_I, Y) = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}_I} \boldsymbol{\Sigma}_{\boldsymbol{x}_I, Y}$. Results that often hold under reasonable conditions in low dimensions, fail to hold under reasonable conditions in high dimensions. Let HD stand for high dimensions. I could be I_{min} or I_H.

Table 1: Regression Summary: Data Splitting and/or Variable Selection

low dimensions	HD but sparse I	high dimensional error
general: $\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_F = \boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
lasso VS: $\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_F = \boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
OLS: $\boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y) = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$ or $\boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}_I} \boldsymbol{\Sigma}_{\boldsymbol{x}_I, Y}$	$\boldsymbol{\beta}_E = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_F$
OPLS: $\boldsymbol{\beta}_{OPLS} = \lambda \boldsymbol{\Sigma}_{\boldsymbol{x}, Y}$	$\boldsymbol{\beta}_{I,OPLS} = \lambda_I \boldsymbol{\Sigma}_{\boldsymbol{x}_I, Y}$	$\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_F$
MMLE: $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}, Y}$	$\boldsymbol{\beta}_{I,MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}_I, Y}$	$\boldsymbol{\beta}_{MMLE} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_F$

73) Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Then OLS minimizes the OLS criterion $Q_{OLS}(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ where RSS is the residual sum of squares. One *ridge regression estimator* $\hat{\boldsymbol{\beta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\beta}) = \frac{1}{a} RSS(\boldsymbol{\beta}) + \frac{\lambda_{1n}}{a} \boldsymbol{\beta}^T \boldsymbol{\beta}.$$

Here $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ common. Then $\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, and the inverse matrix exists provided $\lambda_{1n} > 0$.

74) Another ridge regression estimator $\tilde{\boldsymbol{\beta}}_R$ minimizes

$$Q_{RR}(\boldsymbol{\beta}) = \frac{1}{a} RSS(\boldsymbol{\beta}) + \frac{\lambda_{1n}}{a}(\boldsymbol{\beta}^T\boldsymbol{\beta} - \beta_1^2).$$

This ridge estimator criterion is similar to the lasso criterion, but it is likely that $\tilde{\boldsymbol{\beta}}_{RR}$ does not have a simple formula unless 75) below is used.

75) It is common to center and scale the predictors and/or to center the response so that the constant term disappears. Then then two ridge regression estimators agree. Software often does this.

76) Refer to 73). It can be shown that

$$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda_{1n}\boldsymbol{I}_n)^{-1}\boldsymbol{Y}$$

where the inverse matrices exist for any $\lambda_{1,n} > 0$. The first formula is better if $n >> p$ while the 2nd formula is better if $n << p$. If $\lambda_{1,n} = 0$, then $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_{OLS}$. If $\hat{\lambda}_{1,n} \to \infty$, then $\hat{\boldsymbol{\beta}}_R \to \boldsymbol{0}$ and $\hat{\boldsymbol{Y}} \to \boldsymbol{0}$. Hence ridge regression is a shrinkage estimator and is regularized if $\lambda_{1,n} > 0$.

If $n > p$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists, then $\hat{\boldsymbol{\eta}}_R = \boldsymbol{A}_n\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{B}_n\hat{\boldsymbol{\beta}}_{OLS}$ where

$$\boldsymbol{A}_n = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X} \text{ and } \boldsymbol{B}_n = [\boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}].$$

77) **RR CLT.** Assume p is fixed and that the conditions of the OLS CLT (Theorem 3.1 and point 29 a)) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ then $\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V})$.

78) For k-fold cross validation (k-fold CV), randomly divide the training data into k groups (folds) of approximately equal size $n_j \approx n/k$ for $j = 1, ..., k$. Leave out the 1st fold, fit the method to the $k-1$ remaining folds, then compute some criterion for the 1st fold. Repeat for folds 2, ..., k.

79) For the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, compute $\hat{Y}_i(j)$ for each Y_i in the fold j left out. Then $MSE_j = \frac{1}{n_j}\sum_{i=1}^{n_j}(Y_i - \hat{Y}_i(j))^2$, and the overall criterion is $CV_{(k)} = \frac{1}{k}\sum_{j=1}^{k} MSE_j$. Note that if each $n_j = n/k$, then $CV_{(k)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i(j))^2$. Then $CV_{(k)} \equiv CV_{(k)}(I_i)$ is computed for $i = 1, ..., J$, and the model I_c with the smallest $CV_{(k)}(I_i)$ is selected.

80) Output like that below means cases 7, 12, 14, 18, 21, and 23 are in fold 1 while cases 1, 16, 22, 24, and 25 are in fold 4.

```
folds: 4 2 3 5 3 3 1 5 2 2 5 1 2 1 3 4 2 1 5 5 1 4 1 4 4 3
```

81) For MLR, k-fold CV picks a model I_c that is good for prediction. Models $I_1, ..., I_J$ are considered. For example, I_j corresponds to λ_j when a grid of J values of λ is used: $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_J$ where $\lambda_j = \lambda_{1,n,j}$. Ridge regression, lasso, and elastic net use such a grid.

82) A fitted or population regression model is sparse if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is nonsparse. A high dimensional population regression model is abundant or dense if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model. High dimensional methods have $n \leq 5p$.

83) In low dimensions, want to estimate $\boldsymbol{\beta} = \boldsymbol{\beta}_F$ that uses all predictors. In high dimensions, it may not be possible to get a consistent estimator of $\boldsymbol{\beta}_F$, but wit iid cases, we want to greatly outperform the null model that uses iid $Y_1, ..., Y_n$ with no predictors.

84) The *lasso estimator* $\hat{\boldsymbol{\beta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=2}^{p} |\beta_i| \tag{2}$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$.

85) Lasso CLT. Assume p is fixed and that the conditions of the OLS CLT (Theorem 3.1 and point 29 a))hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{V}).$$

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

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

c) If $\hat{\lambda}_{1,n}/n \xrightarrow{P} 0$, then $\hat{\boldsymbol{\beta}}_L \xrightarrow{P} \boldsymbol{\beta} = \boldsymbol{\beta}_{F,OLS}$.

86) Like ridge regression, lasso uses a grid of λ values: $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_J$ where $\lambda_j = \lambda_{1,n,j}$. The value of λ_J is the smallest value of λ such that $\hat{\boldsymbol{\beta}}_2 = \cdots = \hat{\boldsymbol{\beta}}_p = 0$ (no nontrivial predictors are active, so none are used). The value of λ_1 tends to be proportional to $n^{3/4}$, which makes the lasso estimator at most $n^{1/4}$ consistent instead of the much better $\sqrt{n}$ consistent. In low dimensions, λ_1 is often selected, and $J = 100$ is fine. In high dimensions, increasing J gives more models that could be good for prediction (multitude of models).

87) In low dimensions, since lasso is a consistent estimator of $\boldsymbol{\beta} = \boldsymbol{\beta}_F$, $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ where I_{min} corresponds to the $\hat{\lambda} = \lambda_j$ chosen by k-fold CV. Hence the lasso variable selection estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. In high dimensions with $p > n$, either all p of the $\hat{\boldsymbol{\beta}}_i$ are nonzero or at most $n + 1$ of the $\hat{\boldsymbol{\beta}}_i$ are nonzero (including a constant). Hence if $\hat{\boldsymbol{\beta}}_S$ is $(a + 1) \times 1$ with $a > n$, then the lasso variable selection estimator can not be a consistent estimator of $\boldsymbol{\beta} = \boldsymbol{\beta}_F = \boldsymbol{\beta}_{OLS}$. Data splitting can be used for inference after checking that the model selected by lasso variable selection is good. Since lasso can select $n_H + 1$ predictors including a constant, may need $n_V > n_H + 1$ to compute OLS on the validation set (e.g. for simulations).

88) Consider intervals that contain c cases $[Y_{(1)}, Y_{(c)}], [Y_{(2)}, Y_{(c+1)}], ..., [Y_{(n-c+1)}, Y_{(n)}]$. Compute $Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, ..., Y_{(n)} - Y_{(n-c+1)}$. Then the estimator shorth$(c) = [Y_{(s)}, Y_{(s+c-1)}]$ is the interval with the shortest length.

89) 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

$$\boldsymbol{R_u} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n}.$$

Then regression through the origin is used for the model $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$ 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 $\hat{\boldsymbol{\eta}}$.

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

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

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression, $j = 1$ corresponds to lasso, and $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS_W(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Usually a grid of M values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$ is used where $\lambda_i = \lambda_{1,n,i}$. 10-fold CV is often used to select $\lambda_S = \hat{\lambda}_{1,n}$.

91) Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, and let $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$ be used to fit elastic net. Then $\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\eta}}_{EN}$, and $\overline{\boldsymbol{Y}}$ are used to find $\hat{\boldsymbol{\beta}}_{EN}$ and $\hat{\boldsymbol{Y}}$. The elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimizes the criterion $Q_{EN}(\boldsymbol{\eta}) = RSS_W(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1$ where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$. Let the $(n+p-1) \times (p-1)$ augmented matrix $\boldsymbol{W}_A$ and the $(n+p-1) \times 1$ augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_1}\,\boldsymbol{I}_{p-1} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{Z}_{\mathrm{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0}$ is the $(p-1) \times 1$ zero vector. Let $RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2$. Then $\hat{\boldsymbol{\eta}}_{EN}$ can be obtained from the lasso of $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$: that is, $\hat{\boldsymbol{\eta}}_{EN}$ minimizes
$Q_L(\boldsymbol{\eta}) = RSS_A(\boldsymbol{\eta}) + \lambda_2\|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta})$.

92) The k-component estimator

$$\hat{\boldsymbol{\beta}}_{kE} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{A}}_{k,n}^T(\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T)^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y} = \hat{\boldsymbol{\Lambda}}_k\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$$

$$= \hat{\boldsymbol{A}}_{k,n}^T(\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T)^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}, Y) = \hat{\boldsymbol{\Lambda}}_k\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}, Y).$$

Suppose $k = p$ and $\hat{\boldsymbol{A}}_{p,n}^{-1}$ exists. Then $\hat{\boldsymbol{\beta}}_{pE} = \hat{\boldsymbol{\beta}}_{OLS}$.

93) The matrix $\boldsymbol{A}$ has eigenvalue λ with eigenvector $\boldsymbol{x} \neq \boldsymbol{0}$ if $\boldsymbol{Ax} = \lambda\boldsymbol{x}$. Let $\boldsymbol{e}$ be an eigenvector of $\boldsymbol{A}$ with unit length: $\|\boldsymbol{e}\|_2 = 1$. If the corresponding eigenvalue is unique, then $\boldsymbol{e}$ and $-\boldsymbol{e}$ are the only such eigenvectors. Suppose $\boldsymbol{A}$ is $p \times p$ and symmetric. Then the eigenvalues of $\boldsymbol{A}$ are real. Then $\boldsymbol{A}$ is positive definite, $\boldsymbol{A} > 0$, if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, and $\boldsymbol{A}$ is positive semidefinite, $\boldsymbol{A} \geq 0$, then $\lambda_p \geq 0$. A positive definite matrix is nonsingular: $\boldsymbol{A}^{-1}$ exists.

94)

$$\hat{\boldsymbol{A}}_{k,n} = \begin{pmatrix} \hat{\boldsymbol{\eta}}_1 \\ \hat{\boldsymbol{\eta}}_2 \\ \vdots \\ \hat{\boldsymbol{\eta}}_k \end{pmatrix}.$$

95) For principle components regression (PCR), let $\hat{\boldsymbol{D}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ or $\hat{\boldsymbol{D}} = \hat{\boldsymbol{R}}_{\boldsymbol{x}}$, the sample correlation matrix of the $\boldsymbol{x}$ in the MLR model $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Let $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ be the eigenvalue eigenvector pairs of $\hat{D}$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \cdots \geq \hat{\lambda}_p$ and with the eigenvectors chosen to be orthonormal. Then $\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{e}}_i$ in 92) and 94). There are $p+1$ PCR estimators: $\hat{\boldsymbol{\beta}}_{1PCR}, ..., \hat{\boldsymbol{\beta}}_{pPCR}$ and the model selection PCR estimator $\hat{\boldsymbol{\beta}}_{MS,PCR} = \hat{\boldsymbol{\beta}}_{k^*PCR}$ where k^* is chosen by model selection such as 10-fold CV.

96) For partial least squares (PLS), the k-component PLS estimator regresses Y on $\hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x}$. There are several equivalent ways to get the $\hat{\boldsymbol{\eta}}_i$. One way is to use

$$\hat{\boldsymbol{\eta}}_1 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}, \hat{\boldsymbol{\eta}}_2 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}, ..., \hat{\boldsymbol{\eta}}_k = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{k-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}.$$

Then $\hat{\boldsymbol{\beta}}_{1PLS} = \hat{\boldsymbol{\beta}}_{OPLS}$. There are $p+1$ PLS estimators: $\hat{\boldsymbol{\beta}}_{1PLS}, ..., \hat{\boldsymbol{\beta}}_{pPLS}$ and the model selection PLS estimator $\hat{\boldsymbol{\beta}}_{MS,PLS} = \hat{\boldsymbol{\beta}}_{k^*PLS}$ where k^* is chosen by model selection such as 10-fold CV.

97) Since OLS is used, need $k < n - 1$ in high dimensions.

ch. 8

98) A multiple linear regression model with heterogeneity is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i$$

for $i = 1, ..., n$ where the e_i are independent with $E(e_i) = 0$ and $V(e_i) = \sigma_i^2$. In matrix form, this model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent 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. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma}_{\boldsymbol{e}} = diag(\sigma_i^2) = diag(\sigma_1^2, ..., \sigma_n^2)$ is an $n \times n$ positive definite matrix. In chapters 2 and 3, the constant variance assumption was used: $\sigma_i^2 = \sigma^2$ for all i. Hence heterogeneity means that the constant variance assumption does not hold.

99) For 98), under iid cases and additional regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V})$$

where $\boldsymbol{V}^{-1} = E[\boldsymbol{x}_i\boldsymbol{x}_i^T]$, $\boldsymbol{\Omega} = E[e_i^2\boldsymbol{x}_i\boldsymbol{x}_i^T]$, and

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T \xrightarrow{P} \boldsymbol{V}^{-1}.$$

100) Under iid cases and 98) but $Y = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$, then $Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x}$ is often an MLR model with heterogeneity. Then $\sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ as in 32).

101) Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and $\boldsymbol{x}_i^T = (1 \;\; \boldsymbol{u}_i^T)$. 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 $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ 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 $\hat{\boldsymbol{\eta}}$.

102) A single index model is $Y = m(SP) + e$ where $E(Y|SP) = m(SP)$ and $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ or $SP = \boldsymbol{\beta}^T\boldsymbol{x}$. If the cases are iid, OPLS theory still holds. Hence $\sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ as in 32).

103) Poisson regression: $Y|SP \sim \text{Poisson}(\exp(SP))$. For Poisson regression, estimate $E(Y|\boldsymbol{x})$ with e^{ESP}. The response plot is a plot of ESP versus Y with the estimated mean function e^{ESP} and lowess added as visual aids. The lowess curve should track the exponential curve fairly closely except possible for the largest values of ESP.

104) Let $Z_i = Y_i$ if $Y_i > 0$, and let $Z_i = 0.5$ if $Y_i = 0$. The **minimum chi–square estimator** of the parameters $(\alpha, \boldsymbol{\beta})$ in a Poisson regression model is $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$, and is found from the weighted least squares regression of $\log(Z_i)$ on $\boldsymbol{x}_i$ with weights $w_i = Z_i$. Equivalently, use the ordinary least squares (OLS) regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{x}_i^T)^T$. If the cases are iid, OPLS theory still holds with Z replacing Y. Hence $\sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z} - \boldsymbol{\Sigma}_{\boldsymbol{x}Z}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ as in 32).

105) The cases $(\boldsymbol{x}_i, Y_i)$ follow a Weibull proportional hazards (PH) regression model if $\log(Y_i) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i + \epsilon_i$ follows an MLR model = Weibull accelerated failure time model.

106) Binary regression: $Y|SP \sim \text{bin}(1, \rho(\text{SP}))$. Here $\rho(SP) = E(Y|SP) = P(Y = 1|SP)$. If $(Y, \boldsymbol{x}^T)^T$ has a joint distribution, then $Y|\boldsymbol{\eta}^T\boldsymbol{x} \sim bin(1, \rho(\boldsymbol{\eta}^T\boldsymbol{x}))$ follows a binary regression model for every $\boldsymbol{\eta} \in \mathbb{R}^p$. However, the model and ρ could be poor. Visualize ρ with a response plot of $\hat{\boldsymbol{\eta}}^T\boldsymbol{x}$ versus Y with a scatterplot smoother added as a visual aid.

107) For binary logistic regression $Y = 0$ or $Y = 1$ and $\rho(SP) = \dfrac{e^{SP}}{1 + e^{SP}}$ Estimate $E(Y|\boldsymbol{x}) = P(Y = 1|\boldsymbol{x})$ with $\hat{\rho}(\boldsymbol{x}) = \hat{\rho}(ESP) = \dfrac{\exp(ESP)}{1 + \exp(ESP)}$. The response plot is a plot of ESP versus Y with the (estimated mean function) logistic curve $\hat{\rho}(ESP)$ and a step function added as visual aids. The step function heights are the sample proportion of cases with $Y = 1$ in each slice, and the step function should track the logistic curve fairly closely.

108) If $Y_i \sim D(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i, \boldsymbol{\gamma})$ and the regression method gives $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\gamma}$, then the *parametric bootstrap* generates $Y_i^* \sim D(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_f, \hat{\boldsymbol{\gamma}})$ for $i = 1, ..., B$. Then a $100(1-\delta)\%$ PI for Y_f is roughly the $100(1-\delta)\%$ shorth PI applied to $Y_1^*, ..., Y_B^*$. Here D is some parametric distribution. (Poisson regression, logistic regression, Weibull regression, etc.)

109) Lasso variable selection can be used for several models, such as MLR, LR, PR, and Weibull PH regression. Fit the lasso estimator $\hat{\boldsymbol{\beta}}_L = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. Then fit the

corresponding regression model (such as MLR, LR, PR, or Weibull PH regression) to the variables x_j corresponding to the $\hat{\boldsymbol{\beta}}_j \neq 0$.

110) The MMLE for regression methods is to fit the marginal regression model: regress Y on x_j to get $(\hat{\alpha}_j, \hat{\beta}_j)$, for $j = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{beta}_1, ..., \hat{\beta}_p)^T$. Let $W_i = \hat{\boldsymbol{\beta}}_{MMLE}\boldsymbol{x}_i$ for $i = 1, ..., n$. Then do the marginal regression of Y on W_i for the regression method (e.g. MLR, LR, PR, Weibull regression). Given the marginal regression output for $j = 1, ..., k$ be able to find $(\hat{\beta}_1, ..., \hat{\beta}_k)^T$ where $k = p$ is possible in low dimensions.

111) MMLE variable selection: get the standardized predictors $\boldsymbol{w}_i = (w_1, ..., w_p)^T$. Get the MMLE estimator $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{beta}_1, ..., \hat{\beta}_p)^T$ from the marginal regressions of Y on the w_i. Then take the J variables x_j corresponding to the largest $|\hat{\beta}_j|$. Regress Y on these J variables.

112) If J is small enough, data splitting can be used for inference after variable selection (lasso VS, MMLE VS, forward selection, etc.).