Math 583 Exam 2 is on Friday, Oct. 20 and emphasizes homeworks 4-6 and quizzes 4-6. You are allowed 10 sheets of notes and a calculator. Any needed tables will be provided. CHECK FORMULAS: YOU ARE RESPONSIBLE FOR ANY ERRORS ON THIS HANDOUT!

36) The MLR model is $Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. This model is also called the **full model**. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Note that $x_{i,1} \equiv 1$.

37) The ordinary least squares OLS full model estimator $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes $Q_{OLS}(\boldsymbol{\beta}) = \sum_{i=1}^n r_i^2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. In the estimating equations $Q_{OLS}(\boldsymbol{\beta})$, the vector $\boldsymbol{\beta}$ is a dummy variable. The minimizer $\hat{\boldsymbol{\beta}}_{OLS}$ estimates the parameter vector $\boldsymbol{\beta}$ for the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Note that $\hat{\boldsymbol{\beta}}_{OLS} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

38) Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b} = x_{i,1}b_1 + \cdots + x_{i,p}b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$. A *response plot* for MLR is a plot of $\hat{Y}_i$ versus Y_i. A *residual plot* is a plot of $\hat{Y}_i$ versus r_i. If the e_i are iid from a unimodal distribution that is not highly skewed, the plotted points should scatter about the identity line and the $r = 0$ line.

39) LS CLT: Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. Assume p is fixed and $n \to \infty$. Also assume that $\max h_i \xrightarrow{P} 0$ as $n \to \infty$ and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$. Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\, \boldsymbol{V})$. Equivalently, $(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\, \boldsymbol{I}_p)$.

40) Use $\boldsymbol{Z}_n \sim AN_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let a be a constant, let $\boldsymbol{A}$ be a $k \times r$ constant matrix, and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_r\boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_r$,

$$a\boldsymbol{Z}_n \sim AN_r\left(a\boldsymbol{\mu}_n, a^2\boldsymbol{\Sigma}_n\right), \quad \text{and} \quad \boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\mu}_n + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_n\boldsymbol{A}^T\right),$$

$$\hat{\boldsymbol{\theta}}_n \sim AN_r\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}}{n}\right), \quad \text{and} \quad \boldsymbol{A}\hat{\boldsymbol{\theta}}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{c}, \frac{\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T}{n}\right).$$

41) Problems with the OLS full model: i) If $n = p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$ regardless of how bad the predictors are. ii) If $n < p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$ or the program fails. iii) Need $n > Jp$ where $J \geq 5$, and preferably $J \geq 10$ for good estimation. If $n < 5p$, the OLS full model overfits.

42)

Label	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	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$[\hat{L}_2, \hat{U}_2]$
$\vdots$			
x_p	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	$[\hat{L}_p, \hat{U}_p]$

The classical OLS large sample 95% CI for β_i is $\hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$. Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If $0 \in$ CI for β_i, then fail to reject H_0, and conclude x_i is not needed in the MLR model given the other predictors are in the model. If $0 \notin$ CI for β_i, then reject H_0, and conclude x_i is needed in the MLR model.

43) 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 $\hat{\boldsymbol{\eta}}$.

44) 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.

45)

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	*	*	*	*	$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_4)^T = \hat{\boldsymbol{\beta}}_{OLS}$

Model I_{min} is the model, among p candidates, that minimizes C_p if $n \geq 10$, or EBIC if $n < 10p$. Model I_j contains j predictors, $x_1^*, x_2^*, ..., x_j^*$ where $x_1^* = x_1 \equiv 1$, the constant.

46) Variable selection is a search for a subset of predictors that can be deleted without important loss of information if $n \geq 10p$ and such that model I (containing the remaining predictors that were not deleted) is good for prediction if $n < 10p$. Note that the "100%" shorth CI for a β_i that is a component of $\boldsymbol{\beta}_O$ is [0,0].

47) Underfitting occurs if $S \not\subseteq I$ so that $\boldsymbol{x}_I$ is missing important predictors. Under-

fitting will occur if $\boldsymbol{x}_I$ is $k \times 1$ with $d = k < a_S$. Overfitting occurs if $S \subset I$ with $S \neq I$ or if $n < 5k$.

48) In 45) 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

49) The out\$cp line gives $C_p(I_2), C_p(I_3), ..., C_p(I_p) = p$ and I_{min} is the I_j with the smallest C_p.

50) 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	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	$[\hat{L}_p, \hat{U}_p]$

51) The OLS SE is also accurate for forward selection with C_p if $\boldsymbol{X}^T\boldsymbol{X}/n \to \boldsymbol{V}^{-1} = diag(d_1, ..., d_p)$ where all $d_i > 0$. The diagonal limit matrix will occur if the predictors are orthogonal or if the nontrivial predictors are independent with 0 mean and finite variance.

52) Inference for OLS forward selection. Suppose p is fixed and I_{min} with C_p is used. The probability that I_{min} underfits goes to 0 as n goes to ∞, and $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under model 44). Hence the PI (3.33) is a large sample $100(1-\delta)\%$ PI for Y_f under mild conditions that is asymptotically optimal for a large class of iid zero mean unimodal distributions with finite variance. Consider using the residual bootstrap with OLS full model residuals and $\hat{\boldsymbol{\beta}}_{I_{min},0}$. In simulations, the shorth CIs for β_i and the prediction region method for testing $H_0 : \boldsymbol{\beta}_O = \boldsymbol{0}$ appear to be more precise than these methods for the OLS full model if $\beta_i = 0$ or H_0 is true if $n \geq 20p$ and $B \geq 50p$ are large enough. For 5000 runs and a 95% CI, if the CI coverage ≥ 0.94, then the shorter the

CI length, the more precise the CI inference. These simulated results have not yet been proven for forward selection since the forward selection estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is generally not asymptotically normal.

53) Suppose $n \geq 10p$ and $B \geq 50p$ are large enough, and that I_{min} is used with C_p. Assume the residual bootstrap is used with the OLS full model residuals. Consider testing $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ and assume $\hat{\boldsymbol{\beta}}^*_{O,i} = \mathbf{0}$ for $i = 1, ..., B$. Then the "100%" prediction region method confidence region is $\{\mathbf{0}\}$. Hence $pval = 1$ estimates the population pvalue. Note that the "100%" shorth CI for a β_i that is a component of $\boldsymbol{\beta}_O$ is [0,0]. **Conjecture:** If the above conditions hold, then fail to reject H_0 and the method can be used after looking at the data and bootstrap results.

54) In 44) suppose $\boldsymbol{\beta}_S$ is $a_S \times 1$. The population MLR model is *sparse* if a_S is small: few population coefficients β_i are nonzero. The population model is *dense* if $n/a_S < J$ where $J = 5$ or 10, say. The fitted model $\hat{\boldsymbol{\beta}}$ is *sparse* id the crude model degrees of freedom d is small. Often d is the number of nonzero estimated coefficients. The fitted model is *dense* if $n/d < J$.

55) Forward selection with OLS generates a sequence of M models. Let I_1 use $x^*_1 = x_1 \equiv 1$: the model has a constant but no nontrivial predictors. To form I_2, consider all models I with two predictors including x^*_1. Compute $Q_2(I) = SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_2 minimize $Q_2(I)$ for the $p-1$ models I that contain x^*_1 and one other predictor. Denote the predictors in I_2 by x^*_1, x^*_2. In general, to form I_j consider all models I with j predictors including variables $x^*_1, ..., x^*_{j-1}$. Compute $Q_j(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_j minimize $Q_j(I)$ for the $p-j+1$ models I that contain $x^*_1, ..., x^*_{j-1}$ and one other predictor not already selected. Denote the predictors in I_j by $x^*_1, ..., x^*_j$. Continue in this manner for $j = 2, ..., M$ to form $I_1, I_2, ..., I_M$. Often $M = \min(\lceil n/J \rceil, p)$.

56) Problems with forward selection: i) If n/p is large, the bootstrap inference for I_{min} with C_p has not yet been proven to work. ii) Forward selection can be slow. If $M = p$, then $\approx p(p-1)/2$ OLS models are fit. If $M = \lceil n/J \rceil$, then $\approx M(2p - M)/2$ models are fit. iii) If n/p is not large, there does not seem to be theory, although the PI (3.33) sometimes performed well with EBIC.

57) Forward selection generates M models $I_1, ..., I_M$. Let $\boldsymbol{x}_I$ and $\boldsymbol{\beta}_I$ be $a \times 1$. For a given data set, p, n, and $\hat{\sigma}^2$ act as constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset I_{min} that minimizes the criterion. Let criteria $C_S(I)$ have the form

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

The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. Typically $\hat{\sigma}^2$ is the OLS full model

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

when n/p is large. $AIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2a$ and $BIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + a \log(n)$ need n/p large. $EBIC(I) = BIC(I) + 2 \log\left[\binom{p}{a}\right]$ may work when n/p is not large.

58) Another variable selection model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T\boldsymbol{\beta}_{S_i}$ for $i = 1, ..., J$ where there are $J \geq 2$ nonnested "true" submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. When this model holds, omitting all predictors x_j with a $\hat{\beta}_{ij}^* = 0$ in the bootstrap sample may result in underfitting.

59) In simulations, we used $\boldsymbol{\beta} = (1, 1, ..., 1, 0, ..., 0)^T$ where the first $k+1$ coefficients $\beta_i = 1$. Hence the population model has k active nontrivial predictors with $a_S = k+1$. The nontrivial predictors were such that $\rho = cor(x_i, x_j) = \dfrac{2\psi + (m-2)\psi^2}{1 + (m-1)\psi^2}$ where $m = p - 1$, and $x_i \neq x_j$ are nontrivial predictors. If $\psi = 0$, then $\rho = 0$. If ψ is close to 1 or $\psi > 0$ and p is large, then ρ gets close to 1. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ for $c > 0$. We expect ρ close to 1 to be favorable for PCR and PLS. The simulation used $\psi = 0, 1/\sqrt{p}$, and 0.9. Hence ρ gets close to 0, 0.5, and 1. The shorth 95% CIs for β_i were obtained. The prediction region method for testing $H_0 : (\beta_{k+2}, ..., \beta_p)^T = \mathbf{0}$ was also tested where H_0 was true. The output gives the proportion of times the prediction region method bootstrap test fails to reject H_0. The nominal proportion is 0.95. The average length of the interval $[0, D_{(U_B)}] = D_{(U_B)}$ is given where the test rejects H_0 if $D_{\mathbf{0}} > D_{(U_B)}$. If the statistic T is $r \times 1$ and asymptotically normal, and if n and B are large enough, we expect the average length to be near $\sqrt{\chi^2_{r,0.95}}$ where $r = p - k - 1$ for the simulation. This result occurred for the OLS full model. For the CIs, the lengths of the CIs give a measure of precision provided the coverage is not much less than the nominal coverage. For the test, $D_{(U_B)}$ does not indicate the volume of the confidence region, so the test "lengths" are not useful for measuring precision.

60) Principal components regression (PCR) uses $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. PCR uses orthogonal predictor variables that are projections on the axes of a hyperellipsoid determined by the eigenvalues and eigenvector of the correlation matrix $\boldsymbol{R_u}$. The first principal component V_1 is the projection on the longest axis, and the ith principal component is the projection on the ith longest axis. Model J_i does the OLS regression of $\boldsymbol{Z}$ on $V_1, ..., V_i$ for $i = 1, ..., M$ where $M \leq \min(p-1, n)$. If $n > p$ then using all $p-1$ principal components is the same as using the OLS full model to get $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\beta}}$. PCR is sometimes useful if there are just a few dominant principal components: e.g. $\sum_{i=1}^{d} \hat{\lambda}_i / \sum_{i=1}^{p-1} \hat{\lambda}_i \geq 0.9$ where d is small and the $\hat{\lambda}_i$ are the eigenvalues of $\boldsymbol{R_u}$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{p-1} \geq 0$.

61) Problems with PCR: i) In general $\hat{\boldsymbol{\beta}}_{PCR}$ is an inconsistent estimator of $\boldsymbol{\beta}$ unless $P(\hat{\boldsymbol{\beta}}_{PCR} = \hat{\boldsymbol{\beta}}_{OLS}) \to 1$ as $n \to \infty$. ii) There is no reason why $V_1, V_2, ..., V_M$ should decrease in importance for predicting Z or Y.

62) Partial least squares (PLS) uses PLS components $V_1, ..., V_M$ that are linear combinations of the nontrivial predictors. Unlike PCR, the PLS components V_i are chosen using the response variable Y: want components highly correlated with Y. Let model J_i use $V_1, ..., V_i$ for $i = 1, ..., M$. If $n > p$ then using all $p-1$ PLS components is the same as using the OLS full model to get $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\beta}}$.

63) Problem: $\hat{\boldsymbol{\beta}}_{PLS}$ is not a consistent estimator of $\boldsymbol{\beta}$ unless $p/n \to 0$ as $n \to \infty$.

64) The matrix $\boldsymbol{A}$ has eigenvalue λ with eigenvector $\boldsymbol{x} \neq \mathbf{0}$ if $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. 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.

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

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

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 $RSS_W(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. 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 $\lambda_S = \hat{\lambda}_{1,n}$.

66) Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and let $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ be used to fit ridge regression. Then $\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\eta}}_R$, and $\overline{\boldsymbol{Y}}$ are used to find $\hat{\boldsymbol{\beta}}_R$ and $\hat{\boldsymbol{Y}}$.

67) The ridge regression estimator $\hat{\boldsymbol{\eta}}_R$ minimizes the criterion in 65) with $j = 2$, and the criterion can be written as

$$Q_R(\boldsymbol{\eta}) = \frac{1}{a}RSS_W(\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\|\boldsymbol{\eta}\|_2^2.$$

If $\lambda_{1,n} = 0$, then $\hat{\boldsymbol{\eta}}_R = \hat{\boldsymbol{\eta}}_{OLS}$. If $\hat{\lambda}_{1,n} \to \infty$, then $\hat{\boldsymbol{\eta}}_R \to \boldsymbol{0}$ and $\hat{\boldsymbol{Y}} \to \overline{\boldsymbol{Y}}$. Hence ridge regression is a shrinkage estimator and is regularized if $\lambda_{1,n} > 0$. Also,

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

where the inverse matrices exist for any $\lambda_{1,n} > 0$. If $n > p$ and $(\boldsymbol{W}^T\boldsymbol{W})^{-1}$ exists, then $\hat{\boldsymbol{\eta}}_R = \boldsymbol{A}_n\hat{\boldsymbol{\eta}}_R = \boldsymbol{B}_n\hat{\boldsymbol{\eta}}_R$ where

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

68) **RR CLT.** Assume p is fixed and that the conditions of the LS CLT Theorem 3.1 hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

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

69) Let the augmented matrix $\boldsymbol{W}_A$ and the augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_A = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0}$ is the $(p-1) \times 1$ zero vector. For $\lambda_{1,n} > 0$, the OLS estimator from regressing $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$ is

$$\hat{\boldsymbol{\eta}}_A = (\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z}_A = \hat{\boldsymbol{\eta}}_R.$$

70) Ridge regression can beat OLS if n/p is small or if $\boldsymbol{X}^T\boldsymbol{X}$ is ill conditioned (nearly singular). Ridge regression can beat lasso if $a_S > n$.

71) Ridge regression with 10-fold CV tends to underfit if both $a_s > 18$ and the predictors are highly correlated.

72) Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and let $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ be used to fit lasso. Then $\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\eta}}_L$, and $\overline{\boldsymbol{Y}}$ are used to find $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{Y}}$.

73) The lasso estimator $\hat{\boldsymbol{\eta}}_L$ minimizes the criterion in 65) with $j = 1$, and the criterion can be written as

$$Q_R(\boldsymbol{\eta}) = \frac{1}{a} RSS_W(\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a} \|\boldsymbol{\eta}\|_1.$$

If $\lambda_{1,n} = 0$, then $\hat{\boldsymbol{\eta}}_L = \hat{\boldsymbol{\eta}}_{OLS}$. If $\hat{\lambda}_{1,n} \to \infty$, then $\hat{\boldsymbol{\eta}}_L \to \boldsymbol{0}$ and $\hat{\boldsymbol{Y}} \to \overline{\boldsymbol{Y}}$. Hence lasso is a shrinkage estimator and is regularized if $\lambda_{1,n} > 0$. Usually a grid of M values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$ is used where $\lambda_i = \lambda_{1,n,i}$ and λ_M is the smallest value of λ such that $\hat{\boldsymbol{\eta}}_\lambda = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ for $i < M$. 10-fold CV is often used to select $\lambda_S = \hat{\lambda}_{1,n}$.

74) By the KKT conditions for convex optimality, $\hat{\boldsymbol{\eta}}_L = \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\hat{\lambda}_{1,n}\boldsymbol{s}_n/n$ where $s_{i,n} \in [-1, 1]$.

75) **Lasso CLT.** Assume p is fixed and that the conditions of the LS CLT Theorem 3.1 hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, then $\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right)$.

76) Lasso can beat OLS if n/p is small or if $\boldsymbol{X}^T\boldsymbol{X}$ is ill conditioned (nearly singular). At most n of the $\hat{\eta}_{iL} \neq 0$ even if $p > n$. This property can be useful if the population model is sparse. Lasso and ridge regression can be much faster than forward selection if both n and p are large.

77) Lasso with 10-fold CV tends to underfit if both $a_s > 18$ and the predictors are highly correlated. Ridge regression can beat lasso if $a_S > n$.

78) The relaxed lasso estimator $\hat{\boldsymbol{\beta}}_{RL}$ is OLS fit to the j variables, including a constant, that have $\hat{\eta}_{iL} \neq 0$. Hence relaxed lasso is a variable selection method that is a competitor of forward selection.

79) If n/p is large, the program should include $\lambda_1 = 0$ and a value like $\lambda_2 \approx \sqrt{n}/\log(n)$.

80) Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and let $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ 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

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

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
$Q_L(\boldsymbol{\eta}) = RSS_A(\boldsymbol{\eta}) + \lambda_2\|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta})$.

81) By the KKT conditions for convex optimality, $\hat{\boldsymbol{\eta}}_L =$
$\hat{\boldsymbol{\eta}}_R - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\hat{\lambda}_{1,n}\boldsymbol{s}_n/n$ where $s_{i,n} \in [-1, 1]$. 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 $\hat{\lambda}_2/\sqrt{n} \xrightarrow{P} 2\psi\tau$.

82) **EN CLT.** Assume p is fixed and that the conditions of the LS CLT Theorem 3.1 hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then $\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\mathbf{0}, \sigma^2 \boldsymbol{V})$.

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

83) The function `enet` does elastic net using 10-fold CV and a grid of α values $\{0, 1/am, 2/am, ..., am/am = 1\}$. The default uses $am = 10$.

84) The large sample $100(1-\delta)\%$ PI for Y_f given $\boldsymbol{x}_f$ and the training data $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$ is the PI in 34) with $\hat{m}(\boldsymbol{x}_f) = \hat{Y}_f$:

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

where

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \text{ if } \mathrm{d} \leq 8\mathrm{n}/9, \text{ and } \mathrm{b_n} = 5\left(1 + \frac{15}{\mathrm{n}}\right),$$

otherwise. Here d is a crude estimate of the model degrees of freedom (df). Also, $\hat{Y}_f = \boldsymbol{x}_{f,I_{min}}^T \hat{\boldsymbol{\beta}}_{I_{min}}$.

85) If $n \geq 10p$, using $d = p$ works ok for OLS, FS, PCR, PLS, RR, L, RL, and EN. OLS, FS, PCR, PLS, L, and RL use variables $x_1^*, ..., x_d^*$, and a better value for d is $d =$ number of variables used (including a constant) = number of $\hat{\beta}_i \neq 0$. FS, L, and RL have $x_i^* = x_j$ for some j while PCR and PLS have $x_i^* = v_i = \boldsymbol{\gamma}_i^T \boldsymbol{x}$, some linear combination of the predictors.

86) If n/p is large, in the simulations the PI had coverage near the nominal coverage. The length was near the asymptotically optimal length for $n \geq 100p$. If n/p is small and the population model is sparse, the PIs can work well under strong regularity conditions for forward selection with EBIC and for lasso and relaxed lasso.

87) The program for forward selection used C_p if $n \geq 10p$ and EBIC if $n < 10p$.

88) 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.

89) 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 $i = 1, ..., M$, and the model I_c with the smallest $CV_{(k)}(I_i)$ is selected.

90) Could modify the k-fold CV criterion by making PIs.

91) 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
```