Math 583 Exam 1 is on Wednesday, Sept. 27 and covers homeworks 1-4 and quizzes 1-4. 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!

1) For classical regression and multivariate analysis, we often want $n \geq Jp$, with $J \geq 5$. Often much larger $J$ is needed. High dimensional methods have $n \leq 5p$.

2) *Regression* investigates how the response variable $Y$ changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of predictors. The *estimated sufficient predictor* $ESP = \hat{\alpha} + \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$.

3) A *response plot* is a plot of ESP vs Y and a *residual plot* is a plot of ESP vs. $r$.

4) A plot of $w$ vs. $z$ puts $w$ on the horizontal axis and $z$ on the vertical axis.

5) 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 $I$.

6) For multiple linear regression (MLR), model MLR 1) 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 $i = 1, ..., n$. Here $n$ is the sample size and the random variable $e_i$ is the $i$th error. Assume that the $e_i$ are iid with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. In matrix notation, these $n$ equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ 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.

Model MLR 2) is

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

for $i = 1, ..., n$. For this model, we may use $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ with $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$.

7) For model MLR 1) $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the ordinary least squares (OLS) estimator is $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, which exists if $n > p$ and $\boldsymbol{X}$ has full rank $p$.

8) For model MLR 2) $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$, let $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$, let $\alpha$ be the intercept, and let the slopes vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. Let the population covariance matrices

$$\mathrm{Cov}(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{x}}, \text{ and}$$

$$\mathrm{Cov}(\boldsymbol{x}, Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}.$$

If the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population where $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exists and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is nonsingular, then the population coefficients from an OLS regression of $Y$ on $\boldsymbol{x}$ (even if a linear model does not hold) are

$$\alpha = \alpha_{OLS} = E(Y) - \boldsymbol{\beta}^T E(\boldsymbol{x}) \text{ and } \boldsymbol{\beta} = \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}.$$

9) The **sample mean** $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_i$ is the sample mean of the data in column $i$ corresponding to variable $X_i$. Let the sample covariance matrices be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \text{ and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}).$$

Then $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{S} = (S_{ij})$. That is, the $ij$ entry of $\boldsymbol{S}$ is the sample covariance $S_{ij}$.Let the method of moments estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$ and

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\,\overline{Y}.$$

10) For model MLR 2) $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$, a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ exists, then $\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\beta}}^T\overline{\boldsymbol{x}}$ and

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that $\sigma_Y^2$, $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then $\hat{\alpha} \xrightarrow{P} \alpha$ and

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta} \text{ as } \mathrm{n} \to \infty$$

where $\alpha$ and $\boldsymbol{\beta}$ are given by 8).

11) If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}),\ E(\boldsymbol{a} + \boldsymbol{Y}) = \boldsymbol{a} + E(\boldsymbol{Y}),\ \&\ E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}.$$

Also

$$\mathrm{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T.$$

Note that $E(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}E(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

12) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

13) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{X} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

14) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{a} + \boldsymbol{X} \sim N_p(\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Suppose

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix},\ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix},\ \textit{and}\ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

15) $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

16) Given a MVN distribution, be able to find the MVN distribution of subsets, pairs of independent random variables and the correlation
$\rho(X_i, X_j) = \frac{\sigma_{i,j}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = Cov(X_i, X_j)/\sqrt{V(X_i)V(X_j)}$.

17) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Often $X_1 = Y$ and $X_2 = X$. Then find $E(Y|X), V(Y|X)$ and $\rho(Y, X)$.

18) Know that if $Y_1, ..., Y_n$ are iid with $E(Y) = \mu$ and $\mathrm{V}(Y) = \sigma^2$, then $E(\overline{Y}) = \mu$ and $\mathrm{V}(\overline{Y}) = \sigma^2/n$. Know $E(S^2) = \sigma^2$.

19) Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

20) Multivariate Central Limit Theorem (MCLT): If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

21) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

22) Suppose $\boldsymbol{A}$ is a conformable constant matrix and $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$. Then $\boldsymbol{A}\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{X}$.

23) The behavior of convergence in distribution to a MVN distribution in B) is much like the behavior of the MVN distributions in A). The results in B) can be proven using the multivariate delta method. Let $\boldsymbol{A}$ be a $q \times k$ constant matrix, $b$ a constant, $\boldsymbol{a}$ a $k \times 1$ constant vector, and $\boldsymbol{d}$ a $q \times 1$ constant vector. Note that $\boldsymbol{a} + b\boldsymbol{X}_n = a + \boldsymbol{A}\boldsymbol{X}_n$ with $\boldsymbol{A} = b\boldsymbol{I}$. Thus i) and ii) follow from iii).

A) Suppose $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

i) $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) $\boldsymbol{a} + b\boldsymbol{X} \sim N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.

iii) $\boldsymbol{A}\boldsymbol{X} + \boldsymbol{d} \sim N_q(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

(Find the mean and covariance matrix of the left hand side and plug in those values for the right hand side. **Be careful with the dimension** $k$ or $q$.)

B) Suppose $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

i) $\boldsymbol{A}\boldsymbol{X}_n \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) $\boldsymbol{a} + b\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.

iii) $\boldsymbol{A}\boldsymbol{X}_n + \boldsymbol{d} \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

24) $\boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}} = \mathrm{Cov}(\boldsymbol{x},\boldsymbol{y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T]$
$\mathrm{Cov}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{B}\boldsymbol{y}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{B}^T$
$\boldsymbol{\Sigma}_{\boldsymbol{A}\boldsymbol{x},Y} = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ (take $\boldsymbol{B} = \boldsymbol{I}_1 = 1$)
$\boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{B}\boldsymbol{y}} = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}\boldsymbol{B}^T$ (take $\boldsymbol{A} = \boldsymbol{I}$)
If $\boldsymbol{w}_i = \boldsymbol{A}\boldsymbol{x}_i$ for $i = 1, ..., n$, then
$\overline{\boldsymbol{w}} = \boldsymbol{A}\overline{\boldsymbol{x}}$,
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}^T$,
$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}^T$,
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$,
$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$,
$\mathrm{Cov}(\sum_{i=1}^n \boldsymbol{x}_i, \sum_{j=1}^m \boldsymbol{z}_j) = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{z}_j)$,

25) All subsets of a MVN distribution are MVN. Suppose

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and if $\boldsymbol{X}_1$ is a $q \times 1$ vector, then $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

26) Let $\boldsymbol{x}_n = (x_{1n}, ..., x_{pn})^T$ and $\boldsymbol{x} = (x_1, ..., x_p)^T$ be random vectors. Then $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$ implies $x_{in} \xrightarrow{D} x_i$ for $i = 1, ..., p$. Hence all subsets of $\boldsymbol{x}_n$ converge in distribution to the corresponding subsets of $\boldsymbol{x}$: if $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, then

$$\begin{pmatrix} x_{i_1,n} \\ \vdots \\ x_{i_k,n} \end{pmatrix} \xrightarrow{D} \begin{pmatrix} x_{i_1} \\ \vdots \\ x_{i_k} \end{pmatrix}.$$

Typically marginal convergence in distribution $x_{i_n} \xrightarrow{D} x_i$ for $i = 1, ..., p$ **does not imply** $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$.

27) The OLS regression of $Y$ on $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is a $k \times p$ constant matrix with full rank $k$, is $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = (\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$, provided the inverse matrices exist.

28) Under the conditions of 27), if the cases $(\boldsymbol{x}_i, Y_i)$ are iid, then the population OLS regression of $Y$ on $\boldsymbol{w}$ is $\boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{w}Y} = (\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, provided the inverse matrices exist.

29) **OLS CLTs.** Consider the MLR model and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\mathrm{VAR}(e_i) = \sigma^2$. If the $\boldsymbol{x}_i$ are random vectors, assume that the cases $(\boldsymbol{x}_i, Y_i)$ are independent, and that the $\boldsymbol{e}_i$ and $\boldsymbol{x}_i$ are independent. Also assume that $\max_i(h_1, ..., h_n) \to 0$ and

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

as $n \to \infty$ where the convergence is in probability if the $\boldsymbol{x}_i$ are random vectors (instead of nonstochastic constant vectors).

a) For MLR model 1) $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the OLS estimator $\hat{\boldsymbol{\beta}}$ satisfies

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

b) For MLR model 2) $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$, the OLS estimator $\hat{\boldsymbol{\phi}}$ satisfies

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

c) Suppose the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population and MLR model 2) $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ holds. Assume that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ exist. Then b) holds and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1})$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$.

30) If the $e_i$ are iid and $p$ is fixed, then under mild condition $\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{P} \boldsymbol{\beta}_{OLS}$. Need iid cases for $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$.

31) The one component partial least squares (OPLS) estimator is easy to compute in high and low dimensions. Compute $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\eta}}_{OPLS}$, then compute $W_i = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}_i$ for $i = 1, ..., n$. Then do the OLS regression of $Y$ on the $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}_i$ using the working model $Y = \alpha + \lambda W + \epsilon$ to get $\hat{\alpha}$ and $\hat{\lambda}$. Then $\hat{\alpha}_{OPLS} = \hat{\alpha}$ and

$$\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \ \text{ where } \ \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}}.$$

Under iid cases,

$$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \ \text{ where } \ \lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}}{\boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$, then $\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$. The OPLS MLR model is $Y = Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$.

32) CLT for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$: Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Assume $E(x_{ij}^k\ Y_i^m)$ exist for $j = 1, ..., p$ and $k, m = 0, 1, 2$. Let $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \boldsymbol{\mu}_Y)$ with sample mean $\overline{\boldsymbol{w}}_n$. Let $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. Then a)

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}),\ \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}),$$

$$\text{and } \ \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

b) Let $\boldsymbol{z}_i = \boldsymbol{x}_i(Y_i - \overline{Y}_n)$ and $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(Y_i - \overline{Y}_n)$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$. Hence $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$.

c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$, assume $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ is true, and assume $\hat{\lambda} \xrightarrow{P} \lambda \neq 0$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\boldsymbol{0}, \lambda^2\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{A}^T).$$

33) **Slutsky's Theorem**: Let $\boldsymbol{x}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{x} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix. If $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$ and $\boldsymbol{y}_n \xrightarrow{P} \boldsymbol{d}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{x}_n + \boldsymbol{y}_n \xrightarrow{D} \boldsymbol{x} + \boldsymbol{d}$ and

ii) $\boldsymbol{y}_n^T\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{d}^T\boldsymbol{x}$.

c) If $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$ and $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then $\boldsymbol{W}_n\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{x}$, $\boldsymbol{x}_n^T\boldsymbol{W}_n \xrightarrow{D} \boldsymbol{x}^T\boldsymbol{C}$, $\boldsymbol{W}_n^{-1}\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{C}^{-1}\boldsymbol{x}$, and $\boldsymbol{x}_n^T\boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{x}^T\boldsymbol{C}^{-1}$.

34) Let $\boldsymbol{u} = (x_{i_1}, ..., x_{i_k})^T$ where $n \geq Jk$ with $J \geq 5$. Sometimes much larger $J$ will be needed. Then apply 32ab) with $\boldsymbol{u}$ in place of $\boldsymbol{x}$ to do a hypothesis test of the form $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ or $H_0 : \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{0}$ where $\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{B}\boldsymbol{\Sigma}_{\boldsymbol{u},Y}$. In particular, if $\boldsymbol{\beta}_{OPLS} = (\beta_1, ..., \beta_p)^T$, use $\boldsymbol{A} = (0, ..., 0, 1, 0, ..., 0)$ with a 1 in the $i$th position to test $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If $i_1 < i_2 < \cdots < i_k$, then the $j$th row of $\boldsymbol{A}$ has a 1 in the $i_j$ position and all other row entries equal to 0 to test $H_0 : (\beta_{i_1}, ..., \beta_{i_k})^T = \boldsymbol{0}$. Note that $\boldsymbol{A} = \boldsymbol{I}_p$ can be used to test $H_0 : \boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ in low dimensions ($n \geq Jp$), but not in high dimensions. Similarly, $\boldsymbol{A} = [\boldsymbol{0} \;\; \boldsymbol{I}_k \;]$ can be used to test whether the last $k$ elements $(\beta_{p-k-1}, ..., \beta_p)^T = \boldsymbol{0}$. To test $H_0 : \beta_3 - \beta_4 = 0$, use $\boldsymbol{A} = (0, 0, 1, -1, 0, ..., 0)$. Using a test based on 32c) may not work well in high dimensions because $\hat{\lambda}$ may not be close to $\lambda$.

35) If the data $Y_1, ..., Y_n$ are arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the $i$th order statistic and the $Y_{(i)}$'s are called the *order statistics*. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \;\; \text{if n is odd,}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \;\; \text{if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. The *sample median absolute deviation* is $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \; i = 1, \ldots, n)$.

36) Suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}.$$

The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column $i$ corresponding to variable $X_i$.

37) Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ be an estimator of multivariate location and dispersion. The $i$th *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the $i$th *squared Mahalanobis distance* is $D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W}))$.

38) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median is $D_i^2 = D_i^2(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median $\text{MED}_j$: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

39) Let the *covmb2 set* $B$ of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set $B$. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \;\; \text{and} \;\; \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \; - \; 1}.$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

Table 1: 40) OPLS MLR Results

| General | $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\beta}_{OPLS}$ |
|---|---|
| $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{1}{\lambda}[Cov(\boldsymbol{x})]^{-1}\boldsymbol{\beta}_{OPLS}$ | $\boldsymbol{\beta}_{OLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ |
| $\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda Cov(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$ | $\boldsymbol{\beta}_{OPLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ |
| $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = Cov(\boldsymbol{x})\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$ | $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ |

41) **Multitude of MLR models:** Suppose

$$\begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} \sim N_{p+1}\left( \begin{pmatrix} \mu_Y \\ \boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{Y\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \end{pmatrix} \right).$$

Let $\boldsymbol{A}$ be a constant $k \times p$ matrix with full rank $k$. Let $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{x}$. Assume $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is nonsingular. Then

$$\begin{pmatrix} Y \\ \boldsymbol{w} \end{pmatrix} \sim N_{k+1}\left( \begin{pmatrix} \mu_Y \\ \boldsymbol{\mu}_{\boldsymbol{w}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{Y\boldsymbol{w}} \\ \boldsymbol{\Sigma}_{\boldsymbol{w}Y} & \boldsymbol{\Sigma}_{\boldsymbol{w}} \end{pmatrix} \right) \sim N_{k+1}\left( \begin{pmatrix} \mu_Y \\ \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{Y\boldsymbol{x}}\boldsymbol{A}^T \\ \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T \end{pmatrix} \right).$$

If $\boldsymbol{A} = \boldsymbol{\eta}^T$ and $w = \boldsymbol{w} = \boldsymbol{\eta}^T\boldsymbol{x}$ is a random variable, then

$$\begin{pmatrix} Y \\ \boldsymbol{\eta}^T\boldsymbol{x} \end{pmatrix} \sim N_2\left( \begin{pmatrix} \mu_Y \\ \boldsymbol{\eta}^T\boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\eta} \\ \boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta} \end{pmatrix} \right).$$

Using 17), $Y|\boldsymbol{\eta}^T\boldsymbol{x} \sim N(\alpha_{\boldsymbol{\eta}} + \boldsymbol{\beta}_{\boldsymbol{\eta}}^T\boldsymbol{x}, \sigma_{\boldsymbol{\eta}}^2)$ where $\alpha_{\boldsymbol{\eta}} = \mu_Y - \boldsymbol{\beta}_{\boldsymbol{\eta}}^T\boldsymbol{\mu}_{\boldsymbol{x}}$, $\boldsymbol{\beta}_{\boldsymbol{\eta}} = \lambda\boldsymbol{\eta}$,

$$\sigma_{\boldsymbol{\eta}}^2 = \Sigma_Y - \boldsymbol{\beta}_{\boldsymbol{\eta}}^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \Sigma_Y - \lambda\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \Sigma_Y - \frac{(\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\eta})^2}{\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta}},$$

and

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\eta}}{\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta}}.$$

Hence $Y|\boldsymbol{\eta}^T\boldsymbol{x}$ follows a population OLS regression model for the regression of $Y$ on $\boldsymbol{\eta}^T\boldsymbol{x}$: $Y|\boldsymbol{\eta}^T\boldsymbol{x} = \alpha_{\boldsymbol{\eta}} + \boldsymbol{x}^T(\lambda\boldsymbol{\eta}) + e$ where $e \sim N(0, \sigma_{\boldsymbol{\eta}}^2)$. Using $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ corresponds to OPLS. Using $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OLS}$ gives $Y|\boldsymbol{\beta}_{OLS}^T\boldsymbol{x} \sim Y|\boldsymbol{x} \sim N(E(Y|\boldsymbol{x}), V(Y|\boldsymbol{x}))$ or $Y|\boldsymbol{x} = \alpha_{OLS} + \boldsymbol{x}^T\boldsymbol{\beta}_{OLS} + e$ where $e \sim N(0, V(Y|\boldsymbol{x}))$ and $V(Y|\boldsymbol{x}) = \Sigma_Y - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$.

Using 17), $Y|\boldsymbol{w}$ also follows a population OLS regression model for the regression of $Y$ on $\boldsymbol{w}$: $Y|\boldsymbol{w} = \alpha_{\boldsymbol{w},OLS} + \boldsymbol{w}^T\boldsymbol{\beta}_{\boldsymbol{w},OLS} + e$ where $e \sim N(0, \sigma_{Y|\boldsymbol{w}}^2)$. Here

$$\boldsymbol{\beta}_{\boldsymbol{w},OLS} = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{w}Y} = (\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y},$$

and $\alpha_{\boldsymbol{w},OLS} = E(Y) - \boldsymbol{\beta}_{\boldsymbol{w},OLS}^T E(\boldsymbol{w})$.

42) Referring to 40) and 41), be able to compute $\lambda_{OPLS}$ from 31), $\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$ if the cases are iid, the MLR model holds, and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is a diagonal matrix. Be able to recognize $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$ from the MLR model.

43) Let $\|\boldsymbol{a}\|_2 = \sqrt{\boldsymbol{a}^T\boldsymbol{a}}$ be the Euclidean norm.

44) The marginal maximum likelihood estimator (MMLE or marginal least squares estimator) computes the marginal regression of $Y$ on $x_i$ resulting in the estimator $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M})$ for $i = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, ..., \hat{\beta}_{p,M})^T$.

45) For multiple linear regression, the marginal estimators are the simple linear regression (SLR) estimators, and $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M}) = (\hat{\alpha}_{i,SLR}, \hat{\beta}_{i,SLR})$. Hence

$$\hat{\boldsymbol{\beta}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}.$$

If the $\boldsymbol{w}_i$ are the predictors standardized to have unit sample variances, then

$$\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} = \boldsymbol{I}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{w}, Y)$$

where $(\boldsymbol{w}, Y)$ denotes that $Y$ was regressed on $\boldsymbol{w}$, and $\boldsymbol{I}$ is the $p \times p$ identity matrix.

46) The MMLE is interesting since if each predictor satisfies a marginal model, then the marginal model theory can be used to find a confidence interval for $\beta_i$ for $i = 1, ..., p$ where $\beta_i$ is the $i$th component of $\boldsymbol{\beta}_{MMLE}$. For MLR, let $\boldsymbol{V} = diag(\boldsymbol{\Sigma}_{\boldsymbol{x}}) = diag(\sigma_1^2, ..., \sigma_p^2)$. For iid cases, $\boldsymbol{\beta}_{MMLE} = \boldsymbol{V}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{V}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\beta}_{OLS}$.

For standardized predictors, let $s_j$ and $\sigma_j$ be the sample and population standard deviations of $x_j$. Let $\boldsymbol{w}_i = \hat{\boldsymbol{D}} \boldsymbol{x}_i = diag(1/s_1, ..., 1/s_p) \boldsymbol{x}_i$ and $\boldsymbol{u}_i = \boldsymbol{D} \boldsymbol{x}_i = diag(1/\sigma_1, ..., 1/\sigma_p) \boldsymbol{x}_i$. Note that $\sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) = \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y}) + \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) = O_P(1) + \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y})$ under mild regularity conditions for iid cases. Hence $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{u},Y}$. Note that $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is the correlation matrix of $\boldsymbol{x}$.

47) Theorem: Consider the MMLE for MLR. Suppose the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some distribution. Let $\boldsymbol{w}_i$ be the standardized predictors and assume $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u},Y}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u}}$ where the $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}$ are nonsingular for large enough $n$ and $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is nonsingular.

$$a) \hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{w}, Y) \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u},Y} =$$

$$\boldsymbol{\eta}_{OPLS}(\boldsymbol{u}, Y) = \boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}} [\boldsymbol{\Sigma}_{\boldsymbol{u}}]^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u},Y} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \boldsymbol{\beta}_{OLS}(\boldsymbol{u}, Y).$$

b) Let $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}(\boldsymbol{u}, Y)$. Then $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}} \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}$ if $\boldsymbol{\beta}_{OLS} = \boldsymbol{0}$ or if $\boldsymbol{\beta}_{OLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ with eigenvalue = 1.

**Ch. 2**

48) Refer to point 5) for a model for variable selection: $\boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S + \boldsymbol{x}_E^T \boldsymbol{\beta}_E = \boldsymbol{x}_S^T \boldsymbol{\beta}_S$. If $S \subseteq I$, then $\boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + \boldsymbol{x}_O^T \boldsymbol{0} = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$. To clarify notation, suppose $p = 3$, a constant $\alpha$ is always in the model, and $\boldsymbol{\beta} = (\beta_1, 0, 0)^T$. Then the $J = 2^p = 8$ possible subsets of $\{1, 2, ..., p\}$ are $I_1 = \emptyset$, $S = I_2 = \{1\}$, $I_3 = \{2\}$, $I_4 = \{3\}$, $I_5 = \{1, 2\}$, $I_6 = \{1, 3\}$, $I_7 = \{2, 3\}$, and $I_8 = \{1, 2, 3\}$. There are $2^{p - a_S} = 4$ subsets $I_2, I_5, I_6$, and $I_8$ such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_2, \hat{\beta}_3)^T$ and $\boldsymbol{x}_{I_7} = (x_2, x_3)^T$.

Let $I_{min}$ correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, use zero padding to form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are $J$ subsets, e.g. $J = 2^p$.

49)

| $I_J$ | model | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_j}$ |
|---|---|---|---|---|---|---|
| $I_1$ | 1 | | * | | | $(0, \hat{\beta}_2, 0, 0)^T$ |
| $I_2$ | 2 | | * | * | | $(0, \hat{\beta}_2, \hat{\beta}_3, 0)^T$ |
| $I_3$ | 3 | * | * | * | | $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, 0)^T$ |
| $I_4$ | 4 | * | * | * | * | $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^T = \hat{\boldsymbol{\beta}}_{OLS}$ |

Model $I_{min}$ is the model, among $p$ candidates, that minimizes $C_p$, AIC, BIC if $n \geq 10$, or EBIC if $n < 10p$. Model $I_j$ contains $j$ predictors, $x_1^*, x_2^*, ..., x_j^*$ if forward selection is used.

50) 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 $\beta_i$ that is a component of $\boldsymbol{\beta}_O$ is [0,0].

51) Underfitting occurs if $S \not\subseteq I$ so that $\boldsymbol{x}_I$ is missing important predictors. Underfitting 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 $n < 5k$.

52) In 49) 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 |

53) 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 $C_p$.

54) 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 $\beta_i$ |
|---|---|---|---|
| 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 | | $[\hat{L}_4, \hat{U}_4]$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_p$ | $\hat{\beta}_p$ | $SE(\hat{\beta}_p)$ | $[\hat{L}_p, \hat{U}_p]$ |

55) Forward selection generates a sequence of $J$ models. Assume all models with a constant $\alpha$ contain a constant. To form $I_1$, consider all models $I$ with one predictor. Compute criterion $C(x_i)$ and let $x_1^*$ minimize the criterion. Then $I_1$ corresponds to $x_1^*$ and $p$ models are fit. To form $I_2$, compute $C(x_1^*, x_j)$ for the $p-1$ models where $x_j \neq x_1^*$. Continue in this manner. The last model fit is $I_J$ using $C(x_1^*, ..., x_{J-1}^*, x_j)$ for the $p-J+1$ models where $x_j$ is not one of the $J-1$ variables already selected. $I_{min}$ is the model corresponding to the smallest criterion. Often $J = min(n-r, p)$ for some integer $r \geq 0$.

The same regression method, eg OLS or a GLM, is used to compute each fitted model $\hat{\boldsymbol{\beta}}_I$. For MLR, typically $C(I) = RSS(I) + p(I)$ where $SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$.

56) Forward selection can be slow if $n$ and $p$ are large with the number of fitted models close to $n(p - n/2)$ if $p >> n$.

57) Lasso variable selection uses a grid of $J$ models depending on a parameter $\lambda_i$ for $i = 1, ..., J$ to form $I_1, ..., I_J$. The `glmnet` default appears to be 100, but researchers often use much larger $J$ for high dimensional data.

58) MMLE variable selection: Let the $\boldsymbol{x}_i$ be the predictors. Let the $\boldsymbol{w}_i$ be the standardized predictors such that the sample variance of each predictor is 1. Find the $J$ variables $x_1^*, ..., x_J^*$ corresponding to the largest $|\hat{\beta}_i|$. Note that these variables are not standardized.

59) A two stage variable selection method is to use 58) to find $r$ variables, then use forward selection or lasso variable selection on the $r$ variables to get model $I_{min}$ with variables $x_1^*, ..., x_a^*$.

**Math 583 Exam 1 is on Wednesday, Sept. 27 and covers homeworks 1-4 and quizzes 1-4. You are allowed 9 sheets of notes and a calculator.**