Math 583 HW 2 2023 Due Wednesday, Sept. 6.
Two pages, problems A) - E).

Do not forget to copy and paste the two source commands into *R*.

Do not forget to hit Enter after pasting the commands into *R*.

A) The trees data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. The girth is the diameter of the tree (in inches) measured at 4.5 feet above the ground. The response variable is volume with predictor variables girth and height.

a) The function `OPLSplot` make the response plot and residual plot for multiple linear regression based on one component partial least squares. Copy and paste the *R* commands for this problem into *R*. Include the two plots in *Word*.

b) The function `OPLSEEplot` plots the OPLS ESP versus the OLS ESP. For MLR, the ESP is the fitted value: ESP $= \hat{Y}$. Copy and paste the *R* commands for this problem into *R*. i) Include the plot in *Word*.

ii) Do the OLS and OPLS estimators seem to be nearly the same? (Is the correlation in the plot nearly one so that the plotted points nearly fall on the identity line?)

c) The function `rcovxy` makes the classical and three robust estimators of $\boldsymbol{\eta}$, and makes a scatterplot matrix of the 4 ESPs and Y. Only two robust estimators are made if $n \leq 2.5p$. The top plots are similar in appearance to the response plots based on the estimators (the scaling is different so the plotted points scatter about a line that is usually not the identity line).

Copy and paste the *R* commands for this problem into *R*. i) Include the plot in *Word*.

ii) Are the four ESPs highly correlated?

B) The function `covxycis` obtains the large sample $100(1-\alpha)\%$ confidence intervals $\hat{\eta}_j \pm t_{n-1,1-\alpha} SE(\hat{\eta}_j)$ for $\eta_j = \mathrm{Cov}(x_j, Y)$ for $j = 1, ..., p$. Let $n = p = 100$. When 100 independent large sample 95% CIs are made, the nominal coverage is 95%. Let the simulated coverage be the number W of 95% CIs that contain the population parameter.

Suppose the simulation uses K runs and $W_i = 1$ if μ is in the ith CI, and $W_i = 0$ otherwise, for $i = 1, ..., K$. Then the W_i are iid binomial(1,$1 - \delta_n$) where $\rho_n = 1 - \delta_n$ is the true coverage of the CI when the sample size is n. Let $\hat{\rho}_n = \overline{W}$. Since $\sum_{i=1}^{K} W_i \sim$ binomial(K, ρ_n), the standard error $SE(\overline{W}) = \sqrt{\rho_n(1-\rho_n)/K}$. For $K = 5000$ and ρ_n near 0.9, we have $3SE(\overline{W}) \approx 0.01$. Hence an observed coverage of $\hat{\rho}_n$ within 0.01 of the nominal coverage $1 - \delta$ suggests that there is no reason to doubt that the nominal CI coverage is different from the observed coverage. So for a large sample 95% CI, we want the observed coverage to be between 0.94 and 0.96 if $K = 5000$. Also a difference of 0.01 is not large. Coverage slightly higher than the nominal coverage is better than coverage slightly lower than the nominal coverage. For a large sample 95% CI, we want the observed coverage to be between 0.88 and 1.0 if $K = 100$.

a) Here $\boldsymbol{y} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$ with $Y_i \sim N(0,1) \perp\!\!\!\perp \boldsymbol{x}_i \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Hence $\mathrm{Cov}(x_j, Y) = 0$ for $j = 1, ..., p = 100$. Hence $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$, the 100×1 vector of zeroes. Copy and paste the *R* commands for this problem into *R*. i) Include the plot in *Word*. The plot shows the 100 CIs. The left most o and + correspond to the 95% CI for $\mathrm{Cov}(x_1, Y)$, ..., and the

rightmost o and + correspond to the 95% CI for $\mathrm{Cov}(x_p, Y)$ with $p = 100$. Most of the 95% CIs contain 0.

ii) If you hit "Enter" after you pasted the commands into R, you get the number of the 100 large sample 95% CIs that contained the population covariance $\mathrm{Cov}(X_j, Y) = 0$. Write down that number. (A number between 88 and 100 gives no reason to doubt that the actual coverage is near the nominal coverage of 0.95.)

b) Now $Y = 0 + \boldsymbol{x}^T\boldsymbol{\beta} + e = x_1 + x_2 + e$ with $\boldsymbol{\beta} = (1, 1, 0, ..., 0)^T$ a 100×1 vector and $\boldsymbol{\Sigma_x} = diag(1, 2, 3, ..., p)$. Then $\boldsymbol{\Sigma_{xY}} = (1, 2, 0, ..., 0)^T$. Copy and paste the R commands for this problem into R. The output gives the number of the 100 large sample 95% CIs that contained the population covariance $\mathrm{Cov}(X_j, Y)$. Write down that number.

c) Now $Y = 0 + \boldsymbol{x}^T\boldsymbol{\beta} + e = \sum_{i=1}^{p} x_i + e$ with $\boldsymbol{\beta} = \mathbf{1}$, the a 100×1 vector of ones, and $\boldsymbol{\Sigma_x} = diag(1, 2, 3, ..., p)$. Then $\boldsymbol{\Sigma_{xY}} = (1, 2, ..., p)^T$. Copy and paste the R commands for this problem into R. The output gives the number of the 100 large sample 95% CIs that contained the population covariance $\mathrm{Cov}(X_j, Y)$. Write down that number.

C) Suppose the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Suppose $Y|\boldsymbol{x} = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $\boldsymbol{x}_i \perp\!\!\!\perp e_i$. Then under mild conditions, $(\alpha, \boldsymbol{\beta}) = (\alpha_{OLS}, \boldsymbol{\beta}_{OLS})$. Hence if a MLR model generated the data, $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma_{xY}}$. Solve this equation for $\boldsymbol{\Sigma_{xY}}$.

D) Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid where the $\boldsymbol{x}_i$ are $p \times 1$ random vectors. Let $\boldsymbol{\mu_x} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let the iid $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu_x})(Y_i - \boldsymbol{\mu}_Y)$ with sample mean $\overline{\boldsymbol{w}}_n$. Then $E(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ and $\mathrm{Cov}(\boldsymbol{w}_i) = \boldsymbol{\Sigma_w}$. Use the MCLT to find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta})$ for some vector $\boldsymbol{\eta}$. (State what $\boldsymbol{\eta}$ is.)

E) Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$ be iid with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ where $\boldsymbol{x}$ is $p \times 1$ with k much smaller than p. Consider testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_A : \boldsymbol{\mu} \neq \mathbf{0}$. Now $\boldsymbol{\mu} = \mathbf{0}$ iff $\|\boldsymbol{\mu}\| = 0$ iff $\boldsymbol{\mu}^T\boldsymbol{\mu} = 0$. Let $n = \mathrm{floor}(\mathrm{k}/2) = \lfloor \mathrm{k}/2 \rfloor$ be the integer part of $k/2$. So $\mathrm{floor}(100/2) = \mathrm{floor}(101/2) = 50$. Let the iid random variables $W_i = \boldsymbol{x}_{2i-1}^T\boldsymbol{x}_{2i}$ for $i = 1, ..., n$. Hence $W_1, W_2, ..., W_n = \boldsymbol{x}_1^T\boldsymbol{x}_2, \boldsymbol{x}_3^T\boldsymbol{x}_4, ..., \boldsymbol{x}_{2n-1}^T\boldsymbol{x}_{2n}$. Then $E(W_i) = \boldsymbol{\mu}^T\boldsymbol{\mu}$ and $V(W_i) = \sigma_W^2$. Use the CLT to find the limiting distribution of $\sqrt{n}(\overline{W} - \boldsymbol{\mu}^T\boldsymbol{\mu})$. The usual t-test and t-CI can be used on the W_i, but $\sigma_W/\sqrt{n}$ can be quite large if p is large. Better tests will be covered in Chapter 9.