```
Math 583    HW 6 Fall 2020        Due Friday, Oct. 9.
Quiz 6 on Wednesday will have problems on large sample theory and. Note that
Exam 2 is now Friday Oct. 23. Final: Monday, Dec. 7, 8-10 AM.
```

Problem numbers are from Olive (2020). Do the source commands from homework 4.

5.7. By the OLS CLT, under mild regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$. If $\boldsymbol{A}$ is a constant $k \times p$ matrix with rank k, what is the limiting distribution of $\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{A}\boldsymbol{\beta})$?

A) Sketch a DD plot if outliers are present. Make sure you include the identity line.

B) Make a rough sketch (drawing) of the two plots that should be made with any multiple linear regression analysis. Place the name of each plot below each sketch.

R problems

4.3. The function `predsim2` computes the data splitting prediction region. The output gives cvr = observed coverage, up ≈ actual coverage, and mnhsq = mean cutoff $D^2_{(U_V)}$. With 5000 runs, expect observed coverage $\in [0.94, 0.96]$ if the actual coverage is close to 0.95.

a) When xtype=3 and dtype=1, $(T, C) = (\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$ where $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. If $n \geq \max(20p, 200)$ and $n_V = 100$, then $D^2_{(U_V)}$ should estimate the population percentile $\chi^2_{p,0.95}$. Copy and paste the commands for this problem into *R*. Include the output in *Word.*

i) Was the observed coverage near the actual coverage?

ii) Was the mnchsq near 18.3?

b) When xtype = 1, $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, diag(1, ..., p))$ and the χ^2 approximation no longer holds. Copy and paste the commands for this problem into *R*. Include the output in *Word.*

i) Was the observed coverage near the actual coverage?

ii) Was the mnchsq a lot larger than 18.3? (If so, then the volume of the prediction region is much larger than that in a).)

c) Copy and paste the commands for this problem into *R*. Include the output in *Word.* Now $p > n$. Were the observed and actual coverages close?

5.12. This problem fits OLS to n inliers and k outliers. The inliers follow the model $Y = x + e$ (the mean function is the identity line) while the outliers are a near point mass with $(x, y) \approx (20, -20)$. Copy and paste the commands for this problem into *R*. Then copy and paste the four plots into *Word.*

The first three plots a), b), and c) use 1 outlier and $n = 10, 100$, and 1000. The OLS line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x$ is added to each plot. When $n = 10$, the OLS line is tilted away from the identity line. There is still some tilt for $n = 100$ but little tilt for $n = 1000$. Plot d) uses 40 outliers but 10000 inliers, and the OLS line is close to the identity line. (The outlier resistance occurs since OLS minimizes $\sum r_i^2$. If the OLS line goes through the outliers, then the inliers are fit badly. If there are enough inliers, then fitting the inliers well and the outliers poorly leads to a lower OLS criterion than fitting the outliers well. One outlier can tilt OLS arbitrarily badly, but the one outlier needs to be very far from the bulk of the data if the number of inliers is large. A small percentage of outliers, e.g. 1%, can tilt OLS even if the outliers are not very fall from the bulk of the data.)