Math 583 HW 1 Fall 2017. Due Friday, Sept. 1.

Lab1 in Neckers 258 (back) on Monday August 28 will cover some of HW1. See the lab1 handout for some information on *R*. Quiz 1 on Friday, Sept. 1 is similar to HW 1. Use 2 sheets of notes. Place your solutions on a separate sheet of paper. DO NOT place solutions side by side. YOU ARE BEING GRADED FOR WORK NOT JUST THE FINAL ANSWER. As a rule of thumb, you should have some idea of what you were doing, even without the book or notes. You are encouraged to form groups to discuss ideas and HW problems, but do not copy.

Problem numbers are from the Olive text but usually are not yet in the handout text.

A 1.10. This problem uses some of the *R* commands at the end of Section 1.2.1. A problem with response and residual plots is that there can be a lot of black in the plot if the sample size n is large (more than a few thousand). A variant of the response plot for the additive error regression model $Y = m(\boldsymbol{x}) + e$ would plot the identity line, the two lines parallel to the identity line corresponding to the Olive (2017, Section 2.1) large sample $100(1-\delta)\%$ prediction intervals for Y_f that depends on $\hat{Y}_f$. Then plot points corresponding to training data cases that do not lie in their $100(1-\delta)\%$ PI. We will use $\delta = 0.01$, $n = 100000$, and $p = 8$.

a) Copy and paste the commands for this part into *R*. They make the usual response plot with a lot of black. Do not include the plot in *Word*.

b) Copy and paste the commands for this part into *R*. They make the response plot with the points within the pointwise 99% prediction interval bands omitted. Include this plot in *Word*. For example, left click on the plot and hit the *Ctrl* and *c* keys at the same time to make a copy. Then paste the plot into *Word*, e.g., get into *Word* and hit the *Ctrl* and *v* keys at the same time.

c) The additive error regression model is a 1D regression model. What is the sufficient predictor $= h(\boldsymbol{x})$?

B 4.7. The Rousseeuw and Leroy (1987, p. 26) Belgian telephone data has response $Y =$ *number of international phone calls* (in tens of millions) made per year in Belgium. The predictor variable $x =$ year (1950-1973). From 1964 to 1969 total number of minutes of calls was recorded instead, and years 1963 and 1970 were also partially effected. Hence there are 6 large outliers and 2 additional cases that have been corrupted.

a) The simple linear regression model is $Y = \alpha + \beta x + e = SP + e$. Copy and paste the *R commands* for this part to make a response plot of $ESP = \hat{Y} = \hat{\alpha} + \hat{\beta} x$ versus Y for this model. Include the plot in *Word*.

b) The additive model is $Y = \alpha + S(x) + e = AP + e$ where S is some unknown function of x. The *R commands* make a response plot of $EAP = \hat{\alpha} + \hat{S}(x)$ versus Y for this model. Include the plot in *Word*.

c) The simple linear regression model is a special case of the additive model with $S(x) = \beta x$. The additive model is a special case of the additive error regression model $Y = m(x) + e$ where $m(x) = \alpha + S(x)$. The response plots for these three models are used in the same way as the response plot for the multiple linear regression model: if the model is good, then the plotted points should cluster about the identity line with no other pattern. Which response plot is better for showing that something is wrong with the model? Explain briefly.

C) 1.11 The *slpack* function `tplot2` makes transformation plots for the multiple linear regression model $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$. Type = 1 for full model OLS and should not be used if $n < 5p$, type = 2 for elastic net, 3 for lasso, 4 for ridge regression, 5 for PLS, 6 for PCR, and 7 for forward selection with C_p if $n \geq 10p$ and EBIC if $n < 10p$. These methods are discussed in Chapter 3.

Copy and paste the three library commands near the top of *slrhw* into *R*.

For parts a) and b), $n = 100, p = 4$ and $Y = \log(Z) = 0x_1 + x_2 + 0x_3 + 0x_4 + e = x_2 + e$. ($Y$ and Z are swapped in the *R* code.)

a) Copy and paste the commands for this part into *R*. This makes the response plot for the elastic net using $Y = Z$ and $\boldsymbol{x}$ when the linear model needs $Y = \log(Z)$. Do not include the plot in *Word*, but explain why the plot suggests that something is wrong with the model $Z = \boldsymbol{x}^T\boldsymbol{\beta} + e$.

b) Copy and paste the command for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word.* Right click *Stop* 3 more times so that the cursor returns in the command window.

c) Is the response plot linear?

For the remaining parts, $n = p - 1 = 100$ and $Y = \log(Z) = 0x_1 + x_2 + 0x_3 + \cdots + 0x_{101} + e = x_2 + e$. Hence the model is sparse.

d) Copy and paste the commands for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word.* Right click *Stop* 3 more times so that the cursor returns in the command window.

e) Is the plot linear?

f) Copy and paste the commands for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word.* Right click *Stop* 3 more times so that the cursor returns in the command window. PLS is probably overfitting since the identity line nearly interpolates the fitted points.

D) 1.12 Get the *R* commands for this problem from (http://parker.ad.siu.edu/Olive/slrhw.txt). The data is such that $Y = 2 + x_2 + x_3 + x_4 + e$ where the zero mean errors are iid [exponential(2) - 2]. Hence the residual and response plots should show high skew. Note that $\boldsymbol{\beta} = (2, 1, 1, 1)^T$. The *R* code uses 3 nontrivial predictors and a constant, and the sample size $n = 1000$.

a) Copy and paste the commands for part a) of this problem into *R*. Include the response plot in *Word.* Is the lowess curve fairly close to the identity line?

b) Copy and paste the commands for part b) of this problem into *R*. Include the residual plot in *Word*: press the *Ctrl* and *c* keys as the same time. Then use the menu commands "Edit>Paste" in *Word.* Is the lowess curve fairly close to the $r = 0$ line? The lowess curve is a flexible scatterplot smoother.

c) The output out$coef gives $\hat{\boldsymbol{\beta}}$. Write down $\hat{\boldsymbol{\beta}}$ or copy and paste $\hat{\boldsymbol{\beta}}$ into *Word.* Is $\hat{\boldsymbol{\beta}}$ close to $\boldsymbol{\beta}$?