STAT 700
Homework 3 Problems
due Wed. Sept. 21

2 Problems. Show all work.
Please follow the Lab report directions off the homework web page for R Problems.
Please work in Groups 2 (or 3)!

1. Suppose we fit the model,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{X}$ involves $r = 3$ independent variables. Assume that there are $n = 10$ observations and ε_i are independent $N(0, \sigma^2)$ random variables.

(a) Write model (1) in full matrix notation, indicating all individual elements and dimensions of matrices. You may assume there is an intercept.

(b) Give the form of the least squares estimator of $\boldsymbol{\beta}$, for the fit to model (1). Call it $\hat{\boldsymbol{\beta}}$. You may use the notation of model (1).

(c) Find $E(\hat{\boldsymbol{\beta}})$ for model (1). Is $\hat{\boldsymbol{\beta}}$ a biased or unbiased estimator of $\boldsymbol{\beta}$? Explain. Find the distribution of $\hat{\boldsymbol{\beta}}$.

Now assume that the true model involves an additional $s = 2$ independent variables contained in $\boldsymbol{W}$, so the true model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{\gamma}$ is the vector of regression coefficients for the independent variables contained in $\boldsymbol{W}$.

(d) Write model (2) in full matrix notation, indicating all individual elements and dimensions of matrices. To save us from re-writing part (a), you can just write down the matrix $\boldsymbol{W}$ and the vector $\boldsymbol{\gamma}$. Note: matrix $\boldsymbol{W}$ should not have a column of 1's.

(e) Under the true model (2), find $E(\hat{\boldsymbol{\beta}})$. In general, is $\hat{\boldsymbol{\beta}}$ an biased or unbiased estimator of $\boldsymbol{\beta}$? Explain. Find the distribution of $\hat{\boldsymbol{\beta}}$.

Next page.

2. **GPA data**. (Ref: Graybill and Iyer (1994)) Consider the population of high school graduates who were admitted to a particular university during the a ten year time period and who completed at least the first year of course work after being admitted. We are interested in investigating how well the first year grade point average (GPA) can be predicted by using the following quantities with 20 students:

X_1 = the score on the mathematics part of the SAT (SATmath)
X_2 = the score on the verbal part of the SAT (SATverb)
X_3 = the grade point average of all high school mathematics courses (HSmath)
X_4 = the grade point average of all the high school English courses (HSenglish)

We will use data available off the class web page:
`http://www.rohan.sdsu.edu/~babailey/stat700/gpa.dat`

(a) Plot the data using the `pairs` function.

(b) Fit a linear model using all the predictor variables. Include summary and diagnostic plots from `lm`. How well does the linear model fit the data?

(c) Test whether the regression is significant at the 0.05 significance level. Be sure to state the null and alternative hypotheses.

(d) Use the R function `step` (or `drop1`) and the AIC model selection criteria to determine the "best" model. You should call the function with your lm object from part (b).

Examine the AIC values. If you want to drop just 1 variable from the full model (so you would have 3 variables included), which one would you drop? What is the AIC for this model?

Note: If you have trouble understanding the output, you can always fit each model and use the `extractAIC` function with each model as the argument and it will return the AIC value.

Is there an even a "better" model than the previous one, based on AIC. If so, what is that model and what is the AIC value?