STAT 540, Fall 2015

Due on **12:00 A.M., Friday, Oct 23rd**

Homework 7

Homework format for all STAT 540 homework this term: Please label all problems clearly and turn in an organized homework assignment. You don't need to spend hours producing beautifully typeset homework, but you won't get credit if we can't find or read your answer. Unless noted otherwise, turn in the following (as appropriate for the problem).

- Theoretical derivation (when asked for).
- Numerical results **with an explanation of your solution**, written in complete sentences. If computer code is absolutely necessary to provide context here, then include it–nicely formatted–within the solution (otherwise, see below).
- Appropriate graphics. Use informative labels, including titles and axis labels. Try to put multiple plots on the page by using, for example, the R command `par(mfrow=c(2,2))`.
- **Only as necessary**: Final clean computer code used to answer the problem **attached to the end of your homework**. Only include the rare code excerpts without which we wouldn't be able to figure out what you did. Annotate your code. Number and order the code in order of the problems. When in doubt, leave it out; consider that we will probably never read it.
- Some problems will be relatively open-ended, such as "Here are some data. Analyze them and write a report." I will provide further instructions about reports later. They should be self-contained, with suitable EDA, graphs, numerical results, and **scientific interpretation**. No computer code should be included. The report should be concise: "no longer than necessary".

(1) The weight of a bear is an important measure of how well it is doing. Weighing a bear in the wild is difficult. It is a lot easier to measure the length of various parts of the bear's body. The following data were collected in an attempt to find simpler measures that could adequately predict bear weight. 54 bears were located in the wild (assume this is a random sample of bears from this location). Each was anesthetized, weighed, and measured. The data are in `bear.txt` on the class web site.

We will consider the data from three X variables: chest, headlen, and neck. These are the girth of the chest, the length of the head and the length of the neck. All are measured in inches. The goal is to predict the weight (Y) of the bear, measured in pounds.

For the purpose of this assignment, use only these four variables (weight, chest, headlen, and neck). Do not worry about the rest of the variables in the data set. Also, do not worry about assumptions. We'll assume that the model is linear and that the errors have equal variances.

(a) Plot weight vs head length (headlen) for these bears. Describe the relationship. Does this relationship make sense (biologically)?

(b) Estimate β_H in the simple linear regression model: $Y = \beta_0 + \beta_H X_{\text{headlength}} + \epsilon$. What are the units of β_H? Does the value "make sense"?

(c) Estimate β_H in the multiple linear regression model: $Y = \beta_0 + \beta_H X_{\text{headlength}} + \beta_C X_{\text{chest}} + \beta_N X_{\text{neck}} + \epsilon$. What are the units of β_H? Does this value make sense?

(d) Should the values in parts (b) and (c) be the same? Explain why or why not.

Use the multiple linear regression in 4c for the next four parts of this question.

(e) Test $H_0 : \beta_H = 0$, using a T statistic. What is the p-value? Write a one-sentence conclusion.

(f) Test $H_0 : \beta_H = 0$ using the model comparison approach.

(g) Construct a test of $H_0 : \beta_H = 0$ and $\beta_N = 0$. What is the p-value (at least approximately)? Write a one-sentence conclusion. Hint: thinking about the models that are being compared may help you construct this test and write a conclusion.

(h) Construct a test of $H_0 : \beta_H = 0$ and $\beta_N = 0$ and $\beta_C = 0$. What is the p-value? Write a one-sentence conclusion.

The final parts of this problem will use the two variable model:

$$Y = \beta_0 + \beta_1 X_{\text{chest}} + \beta_2 X_{\text{neck}} + \epsilon$$

(i) Estimate β_0, β_1 and β_2; then predict the weight for the following 3 bears

bear	chest	neck
A	35in	20in
B	55in	30in
C	50in	15in

(j) Calculate the s.e. of the predicted average (i.e. the s.e. of the line) for each of the three bears in the previous part.

(k) For these 54 bears, the average chest size is 35.6 inches; the average neck size is 20.5 inches. Explain why the s.e. for bear C is higher than that for bear B, even though both the chest and neck measurements for bear C are closer to the average values.

(2) The problem is about linerboard production. The file `mill.txt` contains production information about linerboard, a paper product. The amount of linerboard produced for each of 25 months is recorded (PRODUCT), along with the cost of raw materials (RAWMAT), energy usage in btus (ENERGY), mill depreciation (DEPREC) and labor costs (LABOR).

In economics, a Cobb-Douglas production function has the form

$$\text{PRODUCT} = \beta_0(\text{RAWMAT})^{\beta_1}(\text{ENERGY})^{\beta_2}(\text{DEPREC})^{\beta_3}(\text{LABOR})^{\beta_4} \times (\text{Error})$$

One hypothesis of particular interest is $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$. This is the hypothesis of constant returns to scale–if the hypothesis is true, then multiplying all of the predictors (inputs) by a constant would lead to the response (output) being multiplied by the same constant.

Fit the Cobb-Douglas function to these data. Ask if you don't know or don't remember how to convert functions like this to a multiple linear regression. Then, answer the following questions.

(a) Test the hypothesis that $\beta_1 = 0, \beta_2 = 0, \beta_3 = 0$, and $\beta_4 = 0$. Report the test statistic and p-value; write an appropriate conclusion.

(b) Test the hypothesis $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ in the Cobb-Douglas model. Again, report the test statistic, a p-value and write an appropriate conclusion.

(c) Estimate $\beta_1 + \beta_2 + \beta_3 + \beta_4$ and its s.e. Calculate a 95% confidence interval for the sum.

(d) A friend claim there is a simpler way to compute the s.e. of $\beta_1 + \beta_2 + \beta_3 + \beta_4$. Since the `R` output or `SAS` output includes s.e. for each β_j, you could then compute $\sqrt{\text{se}^2_{\beta_1} + \text{se}^2_{\beta_2} + \text{se}^2_{\beta_3} + \text{se}^2_{\beta_4}}$. This value is 0.182 for these data. Explain why this is not the s.e. of $\beta_1 + \beta_2 + \beta_3 + \beta_4$.

(3) Consider a competition among 5 table tennis players labeled 1 through 5. For $1 \leq i < j \leq 5$, define y_{ij} to be the score for player i minus the score for player j when player i plays a game against player j. Suppose for $1 \leq i < j \leq 5$,

$$y_{ij} = \beta_i - \beta_j + \epsilon_{ij}$$

where $\beta_1, \ldots, \beta_5$ are unknown parameters and the ϵ_{ij} terms are random errors with mean 0. Suppose four games will be played that will allow us to observe y_{12}, y_{34}, y_{25}, and y_{15}. Let

$$\mathbf{y} = \begin{bmatrix} y_{12} \\ y_{34} \\ y_{25} \\ y_{15} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}, \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_{12} \\ \epsilon_{34} \\ \epsilon_{25} \\ \epsilon_{15} \end{bmatrix}.$$

(a) Define a design matrix $\mathbf{X}$ so that the model above may be written as the standard linear model.

(b) is $\beta_1 - \beta_2$ estimable? Prove that your answer is correct.

(c) is $\beta_1 - \beta_3$ estimable? Prove that your answer is correct.

(d) Write down a general expression for the normal equation.

(e) Find a solution to the normal equation in this particular problem involving table tennis players. Hint: find a matrix $\mathbf{G}$ such that $(\mathbf{X}'\mathbf{X})\mathbf{G}(\mathbf{X}'\mathbf{X}) = (\mathbf{X}'\mathbf{X})$, which is a generalized inverse of $\mathbf{X}'\mathbf{X}$, that is $(\mathbf{X}'\mathbf{X})^-$ (you can use the function `ginv` in `R`).

(f) Find the ordinary least square (OLS) estimator for $\beta_1 - \beta_5$.

(g) For a general model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, what must we assume about $\boldsymbol{\varepsilon}$ in order for the OLS estimator of an estimable function $\mathbf{c}'\boldsymbol{\beta}$ to be unbiased?

(h) For a general model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, what must we assume about $\boldsymbol{\varepsilon}$ in order for the OLS estimator of an estimable function $\mathbf{c}'\boldsymbol{\beta}$ to have the smallest variance among all linear unbiased estimators?

(i) Give a linear unbiased estimator of $\beta_1 - \beta_5$ that is not the OLS estimator.

(4) Suppose that $\mathbf{X}$ is an $n \times p$ design matrix, show that $\mathcal{C}(\mathbf{X}) = \mathcal{C}(P_{\mathbf{X}})$.

(5) Consider a multiple regression model that does not contain an intercept term, i.e.,

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

where $\epsilon_i \sim$ i.i.d $N(0, \sigma^2)$ for $i = 1, \ldots, n$.

(a) Present the objective function that is minimized to obtain the least squares estimators of β_1 and β_2.

(b) Present the normal equations that must be solved to obtain the least squares estimators for β_1 and β_2.

(c) What is the matrix formula for the least squares estimator $\hat{\boldsymbol{\beta}}$? Give formulae for $(\mathbf{X}'\mathbf{X}), (\mathbf{X}'\mathbf{X})^{-1}, \mathbf{X}'\mathbf{y}, \hat{\boldsymbol{\beta}}$ based on this particular problem.

(d) Give the degrees of freedom associated with the sum of squared residuals.

(6) Suppose $\mathbf{X}$ is an $n \times p$ matrix and $\mathbf{y}$ is an $n \times 1$ vector. Suppose that $\mathbf{z} \in \mathcal{C}(\mathbf{X})$ and $\mathbf{z} \neq P_{\mathbf{X}}\mathbf{y}$. Prove that $||\mathbf{y} - \mathbf{z}|| > ||\mathbf{y} - P_{\mathbf{X}}\mathbf{y}||$. Hint: note that for any vector $\mathbf{a}$ and any vector $\mathbf{b} \neq \mathbf{0}$ such that $\mathbf{a}'\mathbf{b} = 0$

$$\begin{aligned} ||\mathbf{a} + \mathbf{b}||^2 &= (\mathbf{a} + \mathbf{b})'(\mathbf{a} + \mathbf{b}) = (\mathbf{a}' + \mathbf{b}')(\mathbf{a} + \mathbf{b}) \\ &= \mathbf{a}'\mathbf{a} + \mathbf{a}'\mathbf{b} + \mathbf{b}'\mathbf{a} + \mathbf{b}'\mathbf{b} \\ &= ||\mathbf{a}||^2 + ||\mathbf{b}||^2 + 2\mathbf{a}'\mathbf{b} \\ &= ||\mathbf{a}||^2 + ||\mathbf{b}||^2 \\ &> ||\mathbf{a}||^2, \end{aligned}$$

and note that

$$||\mathbf{y} - \mathbf{z}||^2 = ||\mathbf{y} - P_{\mathbf{X}}\mathbf{y} + P_{\mathbf{X}}\mathbf{y} - \mathbf{z}||^2 = \ldots$$

.

(7) Textbook problems: Problems 7.4 Grocery retailer; 7.32 Reduced model.