Homework One

Statistics 151a (Linear Models)

Due by 11:59 pm on September 6, 2018

August 28, 2018

1. Consider simple linear regression where there is one response variable y and an explanatory variable x and there are n subjects with values $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$.

   a) Write down (no need to calculate) the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 for the regression $\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$ of y on x. (**0.3 points**)

   b) Write down (again no need to calculate) the estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ of α_0 and α_1 for the regression $\mathbb{E}(x_i|y_i) = \alpha_0 + \alpha_1 y_i$ of x on y. (**0.3 points**)

   c) Intuition might suggest that $\hat{\alpha}_1 = 1/\hat{\beta}_1$. Is this true? (**0.2 point**)

   d) Consider the Pearson height data used in class where y is Son's height and x is Father's height. Plot the data and the two lines $y = \hat{\beta}_0 + \hat{\beta}_1 x$ and $x = \hat{\alpha}_0 + \hat{\alpha}_1 y$. (**0.5 points**)

2. Download the dataset *meap93.Rdata* from bcourses. Take y to be the variable *math10* which denotes the percentage of tenth graders at a high school receiveing a passing score on a standardized mathematics exam. Take x to be the variable *lnchprg* which denotes the percentage of students who are eligible for a federally funded school lunch program.

   a) Fit a simple linear regression model for y on x. Report the estimates of β_0 and β_1 together with their standard errors. (**0.6 points**)

   b) We would expect the lunch program to have a positive effect on student performance. Does your model support such a positive relationship? If yes, explain why. If no, what went wrong? (**0.5 points**)

3. Consider again the data *meap93.Rdata*. Now we want to explore the relationship kbetween the math pass rate (*math10*) and spending per student (*expend*).

a) Do you think each additional dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate? Explain. (**0.4 points**)

b) In the model $math10 = \beta_0 + \beta_1 \log(expend) + e$, argue that $\beta_1/10$ is the percentage point change in *math10* given a 10 percent increase in *expend*. (**0.4 points**)

c) Use the data to estimate the parameters β_0 and β_1 in the above model. Report the estimates and standard errors. (**0.5 points**)

d) How big is the estimated spending effect i.e., if spending increases by 10 percent, what is the estimated percentage point increase in *math10*? (**0.2 points**)

e) One might worry that regression analysis can produce fitted values for *math10* that are greater than 100. Why is this not much of a worry in this dataset? (**0.4 points**)

4. Consider the setting of simple linear regression from class. The simple regression line gives the prediction $\hat{\beta}_0 + \hat{\beta}_1 a$ when the explanatory variable value equals a.

a) Using the formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$, show that (**0.5 points**)

$$\hat{\beta}_0 + \hat{\beta}_1 a = \sum_{i=1}^{n} y_i \left\{ \frac{1}{n} + \frac{(a-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}.$$

b) Under the assumption that $(x_1, y_1), \ldots, (x_n, y_n)$ are independent with $var(y_i|x_i) = \sigma^2$, show that (**0.5 points**)

$$var(\hat{\beta}_0 + \hat{\beta}_1 a | x_1, \ldots, x_n) = \frac{\sigma^2}{n} + \frac{\sigma^2 (a-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

c) As a varies, argue that $var(\hat{\beta}_0 + \hat{\beta}_1 a | x_1, \ldots, x_n)$ is smallest when $a = \bar{x}$. (**0.5 points**)

5. Consider the Anscombe dataset available in R which can be accessed (and plotted) via

```
library(datasets)
a <- anscombe
par(mfrow=c(2,2))
plot(a$x1,a$y1, main=paste("Dataset One"))
plot(a$x2,a$y2, main=paste("Dataset Two"))
plot(a$x3,a$y3, main=paste("Dataset Three"))
plot(a$x4,a$y4, main=paste("Dataset Four"))
```

a) For each of these four datasets, fit a linear model for the response variable on the explanatory variable (including the intercept term). Plot these four datasets (in the same graphics window as above) along with the fitted regression lines. Does the linear model make sense for these datasets? (**0.6 points**)

b) In each of the four datasets, predict the response variable when the explanatory variable equals 10. Do these predictions make sense? (**0.6 points**)

6. Consider simple linear regression where there is one response variable y and one explanatory variable x and there are n subjects with values $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$. Suppose I believe that the average value of y should equal zero when $x = 0$. In that case, I would fit the model

$$\mathbb{E}(y|x) = \beta_1 x$$

for modeling the relationship between y and x.

a) Compute the least squares estimate $\hat{\beta}_1$ of β_1. (**0.4 points**)

b) Show that $\hat{\beta}_1$ is unbiased. (**0.4 points**)

c) Compute the variance of $\hat{\beta}_1$ under the assumption $var(y|x) = \sigma^2$. (**0.4 points**)

7. The goal of this problem is to investigate the unbiasedness property of the least squares estimator in simulations and also to check if the standard error formulas for the least squares estimators are valid under heteroskedasticity. We work in the setting of simple linear regression.

Consider independent observations $(x_1, y_1), \ldots, (x_n, y_n)$ drawn according to the following regression model:

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, 25) \qquad \text{if } x_i \leq 65,$$

$$y_i|x_i \sim \beta_0 + \beta_1 x_i + 10T_i \qquad \text{if } 65 < x_i \leq 70$$

where T_i is a t distributed random variable with 3 degrees of freedom, and

$$y_i|x_i \sim \beta_0 + \beta_1 x_i + U_i \qquad \text{if } x_i > 70$$

where U_i is a random variable that is uniformly distributed in the interval $[-8, 8]$.

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the least squares estimates of β_0 and β_1 based on the data $(x_1, y_1), \ldots, (x_n, y_n)$.

a) Are $\hat{\beta}_0$ and $\hat{\beta}_1$ conditionally (on $x_1, \ldots, x_n$) unbiased for β_0 and β_1 respectively? (**0.35 points**)

b) Check the unbiasedness of $\hat{\beta}_0$ and $\hat{\beta}_1$ via the following simulation. Set $n = 100, \beta_0 = 32, \beta_1 = 0.5$ and the values $x_1, \ldots, x_n$ via

```
x = seq(59, 76, length.out = 100).
```

Also set $M = 10000$. For every $j = 1, \ldots, M$, generate data $y_1^{(j)}, \ldots, y_n^{(j)}$ from the given model. Compute the least squares estimators $\hat{\beta}_0^{(j)}$ and $\hat{\beta}_1^{(j)}$ from the data $(x_1, y_1^{(j)}), \ldots, (x_n, y_n^{(j)})$. Estimate the bias of $\hat{\beta}_i, i = 0, 1$ by

$$\frac{1}{M} \sum_{j=1}^{M} \hat{\beta}_i^{(j)} - \beta_i.$$

Are the estimates of the bias close to zero? (**0.55 points**)

c) Show that the assumption of homoskedasticity is not valid here. (**0.35 points**)

d) In the simulation, compute the standard deviation of the values $\hat{\beta}_i^{(1)}, \ldots, \beta_i^{(\hat{M})}$ for $i = 0, 1$. Also compute, for each $j = 1, \ldots, M$, the standard errors reported by the *lm* function in R based on the data $(x_1, y_1^{(j)}), \ldots, (x_n, y_n^{(j)})$ and plot a histogram of these standard errors. Are the standard errors reported by R reliable when homoskedasticity is violated? (**0.55 points**)

References