Homework 2, MATH455: Due Friday, 02/02/2018

Alexander Van Roijen

February 25, 2018

Instructions: The homework assignment editing this LaTeX document. Download the LaTeX source from the class web page and study it to learn more about LaTeX. Replace the text with appropriate information. Run "pdflatex" on this document.

You will submit this assignment in two parts:

1. Print out the PDF file and bring it to class, and
2. Send an e-mail to:

   gang@math.binghamton.edu

   before class on the due date with two attachments:

   - The LaTeX source file, and
   - The generated PDF document.

Please complete the following:

Here are prior proofs for further development

$E[\widehat{\beta}_1] = \beta_1$

Proof.

$$E[\widehat{\beta}_1] = E[\frac{\sum_{i=1}^n (x_i - \overline{x})y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}] = \frac{\sum_{i=1}^n (x_i - \overline{x})E[y_i]}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})\beta_0 + \beta_1 x_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = 0 + \frac{\sum_{i=1}^n \beta_1 (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \beta_1$$

□

$E[\widehat{\beta}_0] = \beta_0$

Proof.

$$E[\widehat{\beta}_1] = E[\overline{y} - \widehat{\beta}_1\overline{x}] = \frac{1}{n}E[\sum_{i=1}^{n}[\beta_0 + \beta_1 x_i + \varepsilon_i]] - E[\widehat{\beta}_1\overline{x}] = \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} = \beta_0$$

□

$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

Proof.

$$\mathrm{Var}(\widehat{\beta}_1) = \mathrm{Var}[\frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}] = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \mathrm{Var}[y_i]}{\sum_{i=1}^{n}(x_i - \overline{x})^4} = \frac{\sigma^2}{S_{xx}}$$

□

1. (20pts) For simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$i = 1, \ldots n$. Suppose the random errors ε_i's are independent and with $\mathbb{E}\varepsilon = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. The least square estimator of β_0 and β_1 is

$$\begin{cases} \widehat{\beta}_0 & = \overline{y} - \widehat{\beta}_1\overline{x} \\ \widehat{\beta}_1 & = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \end{cases}$$

(a) (10pts) $\mathrm{Var}(vuiroheh) = \sigma^2\left(\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}\right)$.

Solution:

Proof.

$$Var(\widehat{\beta}_0) = Var(\overline{y}) + \overline{x}^2 Var(\widehat{\beta}_1) = \frac{n\sigma^2}{n^2} + \frac{(\overline{x}^2\sigma^2)}{S_{xx}} = \frac{\sigma^2 S_{xx}}{nS_{xx}} + \frac{(\overline{x}^2\sigma^2 n)}{nS_{xx}} = \frac{\sigma^2(\overline{x}^2 n + S_{xx})}{nS_{xx}}$$

$$\text{So what is left is to show that } \overline{x}^2 n + S_{xx} = \sum_{i=1}^{n} x_i^2$$

$$\overline{x}^2 n + S_{xx} = \overline{x}^2 n + \sum_{i=1}^{n}(x_i - \overline{x})^2 = \overline{x}^2 n + \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{2n\overline{x}\sum_{i=1}^{n} x_i}{n} + 2n\overline{x}^2 =$$

$$\sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + 2n\overline{x}^2 = \sum_{i=1}^{n} x_i^2$$

□

(b) (10pts) $Cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\sigma^2 \left(\frac{\overline{x}}{S_{xx}}\right)$.

Solution: Put your answer here.

Proof.

$$Cov(\widehat{\beta}_0, \widehat{\beta}_1) = E[\widehat{\beta}_0\widehat{\beta}_1] - E[\widehat{\beta}_0]E[\widehat{\beta}_1] = E[\widehat{\beta}_0\widehat{\beta}_1] - \beta_0\beta_1$$

$$E[\widehat{\beta}_0\widehat{\beta}_1] = E[\overline{y}\widehat{\beta}_1 - \widehat{\beta}_1^2\overline{x}] = E[\overline{y}\widehat{\beta}_1] - E[\widehat{\beta}_1^2\overline{x}] = E[\overline{y}]E[\widehat{\beta}_1] - E[\widehat{\beta}_1^2\overline{x}] = \beta_1(\beta_0 + \beta_1\overline{x}) - \overline{x}E[\widehat{\beta}_1^2]$$

We can use a convenient trick to determine $E[\widehat{\beta}_1^2]$

$$Var(\widehat{\beta}_1) := E[\widehat{\beta}_1^2] - E[\widehat{\beta}_1]^2 \rightarrow E[\widehat{\beta}_1^2] = Var(\widehat{\beta}_1) + E[\widehat{\beta}_1]^2 = Var(\widehat{\beta}_1) + \beta_1^2 = \frac{\sigma^2}{S_{xx}} + \beta_1^2$$

Subbing this back in we get

$$\beta_1(\beta_0 + \beta_1\overline{x}) - \overline{x}E[\widehat{\beta}_1^2] = \beta_1(\beta_0 + \beta_1\overline{x}) - \overline{x}(\frac{\sigma^2}{S_{xx}} + \beta_1^2)$$

We now bring this back to our original equation and get

$$\beta_1\beta_0 + \beta_1^2\overline{x} - \overline{x}\frac{\sigma^2}{S_{xx}} + \overline{x}\beta_1^2 - \beta_0\beta_1 = -\overline{x}\frac{\sigma^2}{S_{xx}} = -\sigma^2\frac{\overline{x}}{S_{xx}}$$

□

2. (10pts) The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Ronald Fisher in his 1936 paper. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.

   (a) Download the data set "iris.txt" from blackboard under the folder "Data sets", read the data into R;

   (b) Create a scatter plot of "Sepals length" (y-axis) v.s. "Sepals width" (x-axis) such that (a) all three species are displayed in the same plot; (b) different colors and symbols are used for different species;

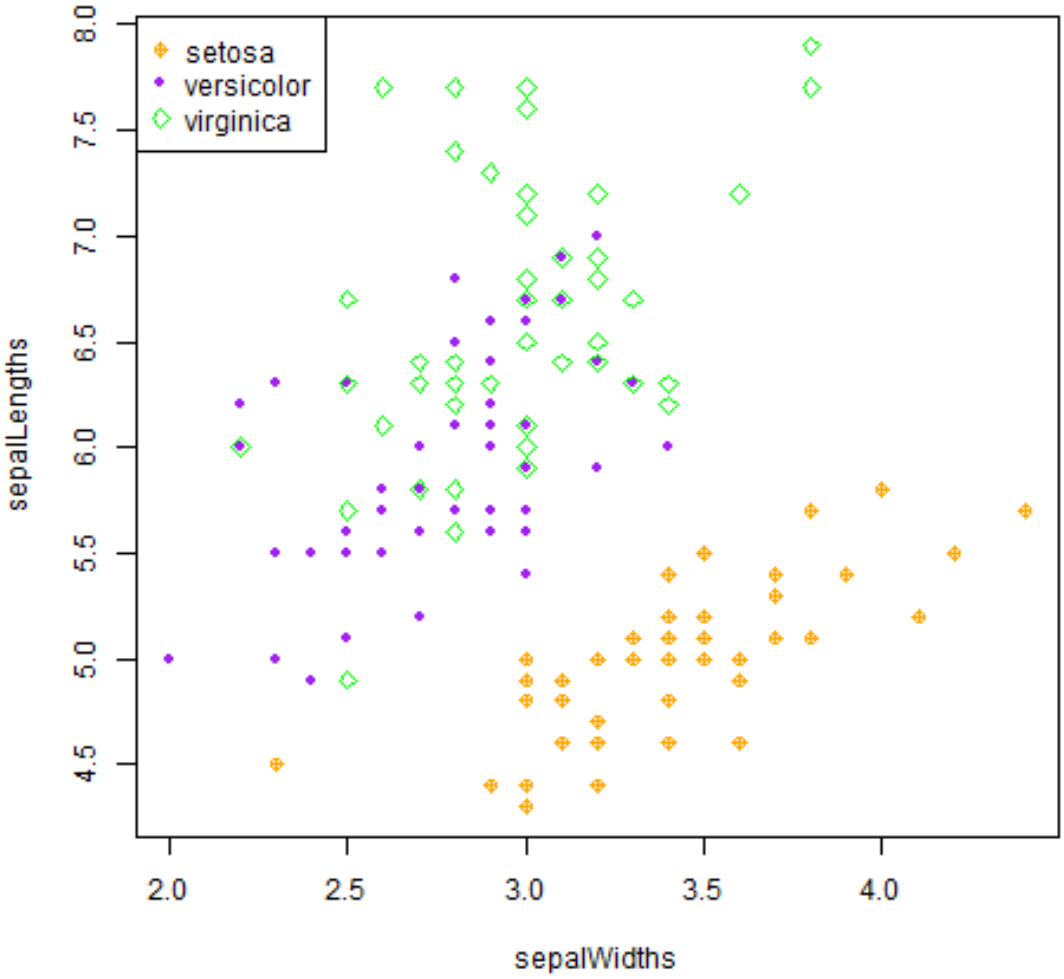

Figure 1: petal length vs petal width

(c) Can you distinguish one species from another by just looking at the plot you created? Observation: The orange points are clearer than the green and purple points since they do not overlap as much

3. (20pts) A TV game show called *let's Make a Deal*,was popular in the 1960s and 1970s. A contestant in the show as given a choice of three doors. Behind one door was a valuable prizes. After a contestant chose a door, say Door 1, the host opened one of the other two doors, say Door 3, showing a less valuable prize. He then gave the contestant the

opportunity to switch from Door 1 to Door 2. Would switching from Door 1 to Door 2 increase the contestant's chances of winning the car?

(a) Find the theoretical probability of winning if the contestant switch the door.

Solution:

$$P(D1) = \frac{1}{3} \rightarrow P(D2 \text{ or } D3) = \frac{2}{3}$$

Two things can happen here, either I chose the right door on my first guess,

meaning door 2 or door 3 will be opened with equal probablity, or I chose incorrectly,

and the other bogus door is revealed.

In the case I chose incorrectly, switching causes me to win every time

In the case I chose correctly, switching doors causes me to lose every time.

I will choose incorrectly to begin with $\frac{2}{3}$ of the time.

If I switch all the time, my chances of winning are thus $\frac{2}{3}$, which is an improvement!

(b) Use software R to conduct a simulation to confirm your solution.

Solution: Describe how did you design the simulation clearly.

```
#this is where i never switch
oDoor=sample(c(1,2,3),100,1)
results=table(oDoor)
myPick=sample(c(1,2,3),100,1)
correct = 0
for(i in 1:100){
if(oDoor[i]==myPick[i]){
correct = correct+1
#this shows the chance I got it right the first time
}
}
print(correct/100.0)
#this is when I always switch
oDoor=sample(c(1,2,3),100,1)
results=table(oDoor)
myPick=sample(c(1,2,3),100,1)
```

```
correct = 0
for(i in 1:100){
reveal=1:3
if(oDoor[i]==myPick[i]){
#i initially picked the correct door, but I am going to switch, so I dont get any co
}
else
{
#now I pick the correct door! he revealed the other bogus door and I switch to choos
correct=correct + 1
}
}
print(correct/100.0)
```

Hint: check out the use of R function "sample()";