Math 50 Fall 2017

Homework #1

(1) Verify that the least squares estimators are given by :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = S_{xy}/S_{xx}$$

Hint. Start with the equations

$$\begin{aligned} -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) &= 0 \\ -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i &= 0 \end{aligned}$$

and solve for the unknowns.

From the first equation we get

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) \\ 0 &= -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\hat{\beta}_0 + \sum_{i=1}^{n}\hat{\beta}_1 x_i \\ 0 &= -n\bar{y} + n\hat{\beta}_0 + n\hat{\beta}_1\bar{x} \\ 0 &= -\bar{y} + \hat{\beta}_0 + \hat{\beta}_1\bar{x} \end{aligned}$$

which gives $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.
This is the wanted formula for $\hat{\beta}_0$. Next we substitute this into the second eqaution

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}\left(y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i\right) x_i \\ 0 &= \sum_{i=1}^{n}\left(y_i - \bar{y}\right) x_i + \hat{\beta}_1 \sum_{i=1}^{n}\left(\bar{x} - x_i\right) x_i \end{aligned}$$

Then a rearrangement gives $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})x_i}{\sum_{i=1}^{n}(x_i-\bar{x})x_i}$. However the denominator is not exactly how we defined S_{xx}. Next we want to show denominator is indeed S_{xx} :

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right) \\ &= \sum_{i=1}^{n}\left(x_i - \bar{x}\right) x_i - \bar{x}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) \\ &= \sum_{i=1}^{n}\left(x_i - \bar{x}\right) x_i \end{aligned}$$

since $\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$. Therefore we obtained $\hat{\beta}_1 = S_{xy}/S_{xx}$.

(2) Note that in our definition

$$S_{xy} = \sum_{i=1}^{n} y_i\left(x_i - \bar{x}\right).$$

This is not the covariance since $covar(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$.
First show that

$$S_{xy} = n\ covar(x,y)$$

and using that conclude

$$\hat{\beta}_1 = \frac{covar(x,y)}{var(x)}.$$

Similar to the last step of the previous problem a calculation gives

$$\begin{aligned}
covar(x,y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,y_i - \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,\bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,y_i - \frac{1}{n}\bar{y}\sum_{i=1}^{n}(x_i-\bar{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,y_i
\end{aligned}$$

Then substituting this

$$\begin{aligned}
\frac{covar(x,y)}{var(x)} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,y_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= S_{xy}/S_{xx} \\
&= \hat{\beta}_1.
\end{aligned}$$

(3) Calculating the expectations show that
   (a) $\hat{\beta}_1$ is an unbiased estimator of β_1.
   (b) $\|\varepsilon\|^2$ is an unbiased estimator of $n\sigma^2$.

For part (a)

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_1\right) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \\
&= \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})\,\mathbb{E}(y_i)\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})\,\mathbb{E}(\beta_0+\beta_1 x_i+\varepsilon_i)\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(\beta_0+\beta_1 x_i)\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,\beta_0+\sum_{i=1}^{n}(x_i-\bar{x})\,\beta_1 x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \beta_1\frac{\sum_{i=1}^{n}(x_i-\bar{x})\,x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \beta_1\frac{\sum_{i=1}^{n}(x_i-\bar{x})\,x_i}{\sum_{i=1}^{n}(x_i-\bar{x})\,x_i} \\
&= \beta_1
\end{aligned}$$

For part (b), lets first recall from probability that

$$Var(v) = \mathbb{E}\left(v^2\right) - \left[\mathbb{E}(v)\right]^2$$

But in our case $\mathbb{E}(\varepsilon_i) = 0$, therefore $Var(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2)$. Then

$$\begin{aligned}\mathbb{E}\left(\|\varepsilon\|^2\right) &= \mathbb{E}\left(\sum_{i=1}^{n} \varepsilon_i^2\right) \\ &= \sum_{i=1}^{n} \mathbb{E}\left(\varepsilon_i^2\right) \\ &= \sum_{i=1}^{n} Var\left(\varepsilon_i\right) \\ &= n\sigma^2\end{aligned}$$

as desired.

(4) For a given sample, suppose that we calculated the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. In addition, let us assume we know that at $x_5 = 0$ the values ε_5 and e_5 are equal. Would that give you extra information about how good your least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Explain. (Hint: The goal of this question is to help understand the difference between ε_i and e_i. Write the equations that define them and use the given information.)

At $x_5 = 0$ we have the following two equations :

$$\begin{aligned}y_5 &= \beta_0 + \beta_1 0 + \varepsilon_5 \\ e_5 &= y_5 - \hat{y}_5\end{aligned}$$

Since ε_5 and e_5 are equal, we get

$$\begin{aligned}e_5 &= \varepsilon_5 \\ y_5 - \hat{y}_5 &= y_5 - \beta_0\end{aligned}$$

which gives

$$\begin{aligned}\beta_0 &= \hat{y}_5 \\ &= \hat{\beta}_0 + \hat{\beta}_1 0 \\ &= \hat{\beta}_0\end{aligned}$$

thus our estimator $\hat{\beta}_0$ is exactly equal to β_0 in this case. We can't say additional information about β_1.

(5) Consider the Rocket Propellant example. The data is at
https://math.dartmouth.edu/~m50f17/propellant.csv
Write an R script to do the following in the given order.
  (a) Plot the scatterer diagram. (x-axis is Age, y-axis is Shear Strength)
  (b) Fit a simple linear regression model to the data. Plot the fitted line onto the scatterer diagram.
  (c) Suppose that you realized there is an error in the table and in the last entry the age should be 11.5 instead of 21.5. After correcting that data, plot the new fitted line onto your plot using a different color. (Your plot should have scatterer data and two line plots).
  (d) Compute the mean and variance of residuals of the last regression model.

HW1 Question 5

Code

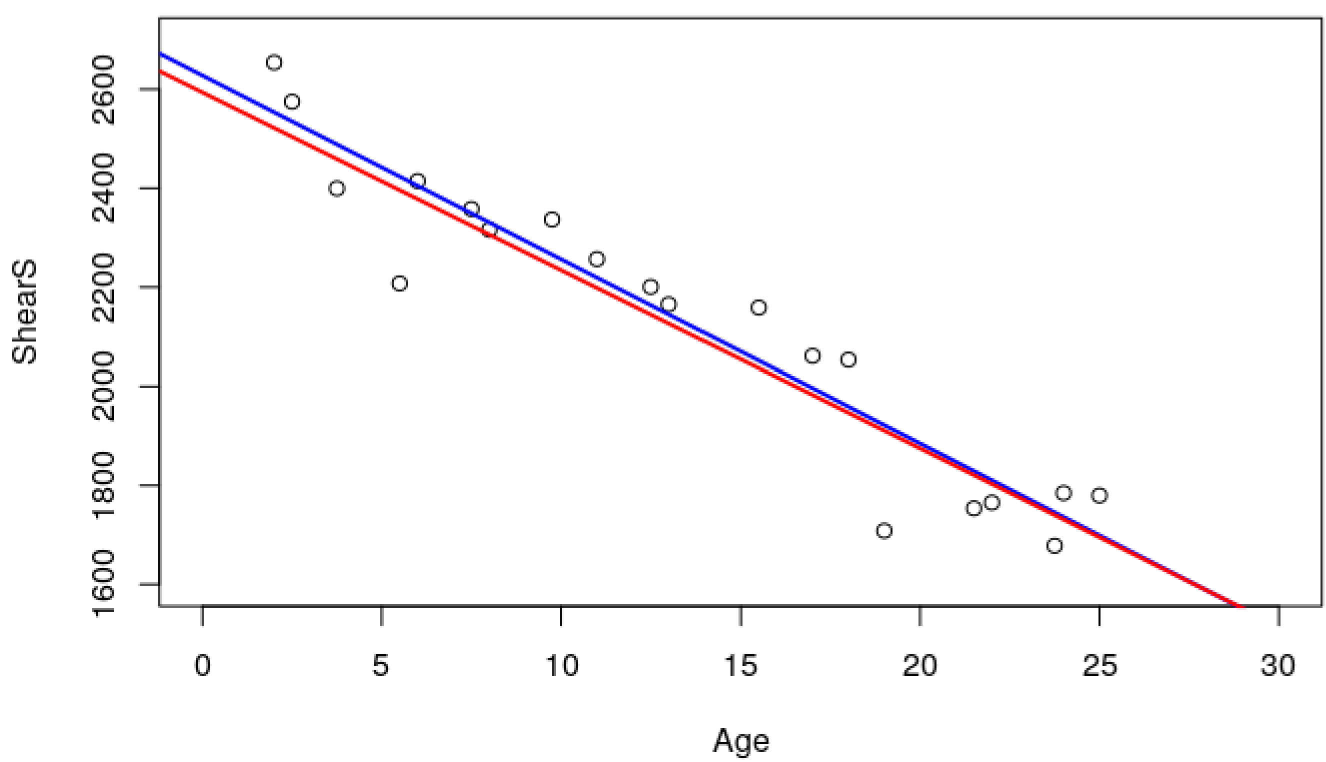
Hide

```
prop = read.table("https://math.dartmouth.edu/~m50f17/propellant.csv", header=T,
sep=",")
Age <- prop$Age
ShearS <- prop$ShearS

# Part (a)
plot(Age, ShearS, xlim=c(0,30), ylim=c(1600,2700))
# Part (b)
fitted = lm(ShearS ~ Age)
abline(fitted$coef, lwd = 2, col = "blue")
```

Hide

```
# Part (c)
AgeCor <- Age
AgeCor[20] <- 11.5
fittedCor = lm(ShearS ~ AgeCor)
abline(fitted2$coef, lwd = 2, col = "red")
```



Hide

```
# Part (d)
res = resid(fitted2)
cat ( ' Mean of residuals : ', mean(res) , '\n' )
```

```
 Mean of residuals :  3.730349e-15
```

Hide

```
cat ( ' Variance of residuals : ', var(res) , '\n' )
```

```
 Variance of residuals :  18517.54
```