HW 1

Adon Rosen

Date: 2020-02-14

Question 1

A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 4 women from each 10-year age group, beginning with age 40 and ending with age 79, resulting in a total sample size of n=16. Some results follow, where X is the age, and Y is a measure of the muscle mass. Assume a simple linear regression model is appropriate.
For these data:

Mean(X) = 61.69 SD(X) = 14.67
Mean(Y) = 52.21 SD(Y) = 24.16

y-hat = 142.68 – 1.47 *X

SE(Slope) = 0.200 Root MSE = 11.38

a.

Parameters:

1. β_0 = 142.68 = The predicted muscle mass in a women when her X value, age is equal to 0
2. β_1 = -1.47 = The change in muscle mass for every one unit change in age

b.

$\hat{Y}_{63} = 142.68 - 1.47 * 63 = 50.07$

c.

In order to calculate the corellation from the slope first multiply the slope of the regression line by the standard deviation of X and then divide by the standard deviation of Y.

Slope: -1.47

SDx: 14.67

SDy: 24.16

r = (Slope * SDx) / SDy = -0.8925869

d.

State our hypotheses:

$H_0 : r = 0$

$H_a : r \neq 0$

$\alpha = 0.05$

We would like to calculate the t value of the correlation. The t value's formula is: $t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$

t = -7.4072668

p = 3.3156464×10^{-6}

Given this p value, we reject the null hypothesis and conclude there is a non-zero relationship between age and muscle mass.

e.

The 95% C.I. for β_1 = -1.47 +/- 0.429; [-1.899,-1.041]

I will interpret this later

f.

The 95% C.I. for a participant with an age of 63 is equal to [23.043,77.097].

I will interpret this later

g.

The residual is equal to $Y - \hat{Y}$ = 45 - 54.48 = -9.48

h.

PRE = 0.7967114

$\beta_{standardized} = \beta\frac{SD_x}{SD_y}$ = -0.8925869

Here we see that we explain roughly 0.7967 percent of the total variance in the outcome. Furthermore, it appears that the that the standardized coefficient between age and muscle mass is: -0.8925869 which suggests a very strong relationship between these two variables.

i.

If two women differ in age by 10 years the predicted difference in muscle mass will be -14.7 units.

j.

It is difficult to tell if the linear assumption holds with these data. However the provided model fits the data very well.

Question 2

The following SAS output contains an analysis in which 6-year graduation rates were collected over a 6-year period and analyzed for trends. The response variable Y is graduation rate (measured as a percentage). The explanatory variable X is year (beginning with year 1)

a.

$\hat{Y} = 61.52667 + 0.95429 * year$

b.

$\hat{Y}_6 = 61.52667 + 0.95429 * 6 = 67.25241$

c.

The residual is equal to $Y - \hat{Y}$ = 67.8 - 66.29812 = 1.50188

Here the residual is 1.5, meaning the value was underpredicted by 1.5%.

d.

PRE = 0.7541 This PRE suggests an excellent model fit.

e.

The 95% C.I. for β_1 = 0.95429 +/- 0.7565; [0.19779,1.71079]

I will interpret this later

f.

State our hypotheses:

$H_0 : \beta_{year} = 0$

$H_a : \beta_{year} \neq 0$

$\alpha = 0.05$

We conclude that year is a significant predictor (t=3.50, p<0.05) and reject that null hypothesis that the relationship between year and graduation rate is not 0.

g.

Conditional standard deviation is equivalent to $\frac{1}{n}\Sigma_{n=1}^{i}(y - \hat{y})^2$ = MSE. Taking the root os the MSE gives you the conditional SD. Here the root mean squared error is equal to 1.13982, therefore the conditional SD is also 1.13982.

h.

This study had 6 observations, assuming all of the observations were unqiue, this study had 6 years worth of data.

i.

The value 15.93 is the sum of squares for the model. Another phrase for this is the explained sum of squares (ESS). ESS reflects how much of the variance the model explains from the original total sum of squares.

j.

While the model reported here suggests a positive relationship between year and graduation rate, this does not restirct the data to only allowing positive increases across years in graduation rates. As long as the model fit displays homoskedacity this is not a concern for the model that was fitted.

k.

$\beta_{year} = 0.95429$

$SD_x = 1.8708287$

$SD_y = 2.0558857$

$\beta_{standardized} = \beta\frac{SD_x}{SD_y} = 0.8643246$

Question 3

a.

```
Call:
lm(formula = y ~ x, data = q.three.dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.790  -6.587   1.048   9.433  11.979 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 182.9725    12.7223  14.382 8.87e-10 ***
x             0.2616     0.1783   1.467    0.164    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.29 on 14 degrees of freedom
Multiple R-squared:  0.1332,	Adjusted R-squared:  0.07132 
F-statistic: 2.152 on 1 and 14 DF,  p-value: 0.1645
```

b.

```
[1] 0.3650112
```

c.

```
Call:
lm(formula = yStand ~ xStand, data = q.three.dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.47840 -0.61676  0.09816  0.88324  1.12160 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.776e-17  2.409e-01   0.000    1.000
xStand      3.650e-01  2.488e-01   1.467    0.164

Residual standard error: 0.9637 on 14 degrees of freedom
Multiple R-squared:  0.1332,	Adjusted R-squared:  0.07132 
F-statistic: 2.152 on 1 and 14 DF,  p-value: 0.1645
```

d.

Question 4