Quiz 1

Adon Rosen

Date: 2020-02-16

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 correlation 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]

This confidence interval suggests that if this sampling procedure was repeated, we would expect the standard deviation of the slope estimate to be about 0.429.

f.

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

This confidence interval suggests that if this sampling procedure was repeated, we would expect the predicted mean to fall between the range stated above in 95% samples.

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 abs[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, thus the assumption that the linear model fitted here provides some utility is tenable.

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 that the additional variable yields a .7541 proportional reduction in error across this model and the null model.

e.

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

This confidence interval suggests that if this sampling procedure was repeated, we would expect the standard deviation of the slope estimate to be about 0.7565.

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 the null hypothesis that the relationship between year and graduation rate is 0.

g.

Conditional variance is equivalent to $\frac{1}{n}\Sigma_{n=1}^{i}(y - \hat{y})^2$ = MSE. Taking the root of 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 unique, 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 restrict the data to only positive increases across years in graduation rates. As long as the model fit displays homoscedastic errors 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 criminologist studying the relationship between population density and robbery rates in medium-sized US cities collected the following data for a random sample of 16 cities; X is the population density of the city (number of people per unit area) and Y is the robbery rate last year (number of robberies per 100,000 people). Assume that a simple linear regression model is appropriate.

I: 1 2 3 4 5 6 7 8 X: 59 49 75 54 78 56 60 82 Y; 209 180 195 192 215 197 208 189

I: 9 10 11 12 13 14 15 16 X: 69 83 88 94 47 65 89 70 Y: 213 201 214 212 205 186 200 204

a.

The fitted regression model is: $\hat{Y}_i = 182.9725 + 0.2616 * X_i$

b.

The estimated correlation value is: 0.3650112

c.

The fitted standardized regression model is: $\hat{Y}_i = 0 + 0.365 * X_i$

d.

Here the $\beta_{popdens}$ is equal to 0.2616 which suggests for every unit increase in population density, there is a 0.2616 unit increase in robbery rate.

e.

The proportion reduction of error here is 0.1332, which suggests a small-medium effect between these two variables. This value tells us that we reduce this proportion in the error compared to a model that only includes an intercept term.

f.

We want to calculate the confidence interval for the mean prediction. The associated formula is:

$\hat{Y} \pm t_{14} * \sqrt{MSE} * \sqrt{\frac{1}{n} \frac{(X_i - \overline{X})^2}{SS_x}}$

Values will be:

$\hat{Y}$ = 201.2826968

t_{14} = 2.1447867

MSE= 92.6898777

$n = 16$

$X_i = 70$

$\overline{X}$ = 69.875

SS_x = 3331.75

C.I.M = [195.76, 206.8]

g.

We want to calculate the confidence interval for an individual predicition. The associated formula is:

$\hat{Y} \pm t_{14} * \sqrt{MSE} * \sqrt{1 + \frac{1}{n} \frac{(X_i - \overline{X})^2}{SS_x}}$

Values will be:

$\hat{Y}$ = 207.5604787

t_{14} = 2.1447867

MSE= 92.6898777

$n = 16$

$X_i = 94$

$\overline{X}$ = 69.875

SS_x = 3331.75

C.I.I = [183.01, 232.11]

h.

Here we want to estimate the 95% confidence interval for the slope of the regression model.

In order to perform this we are going to use the following formula:

$\hat{\beta}_1 \pm t_{df=14} * se(\hat{\beta}_1)$

With the standard error of $\hat{\beta}_1$ calculated as:

$\text{se} = \sqrt{\frac{MSE}{\Sigma_{i=1}^{n}(x_i - \overline{x})^2}}$

Values will be:

$\beta_1 = 0.2616$

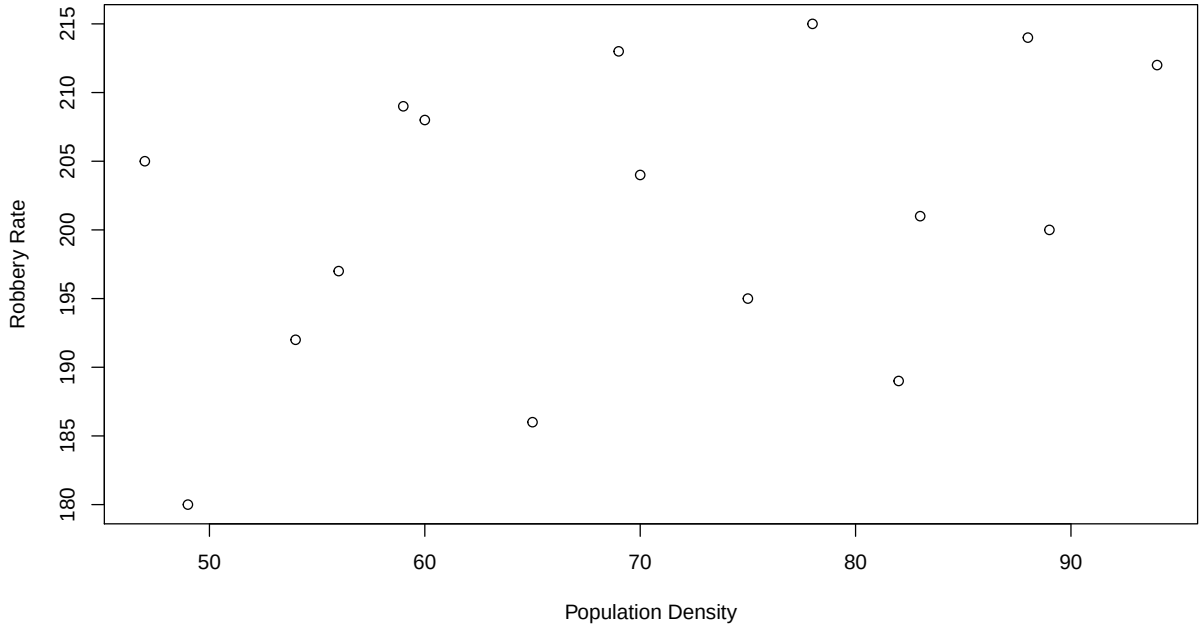
$t_{14} = 2.1447867$

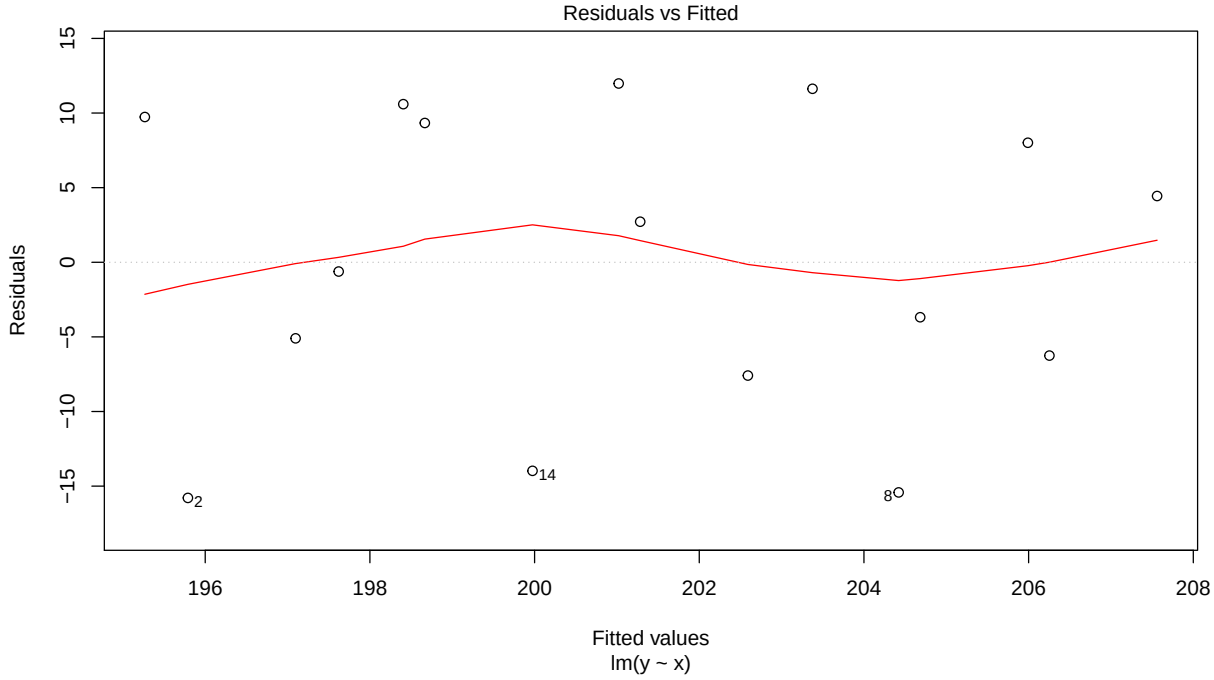
MSE = 92.6898777

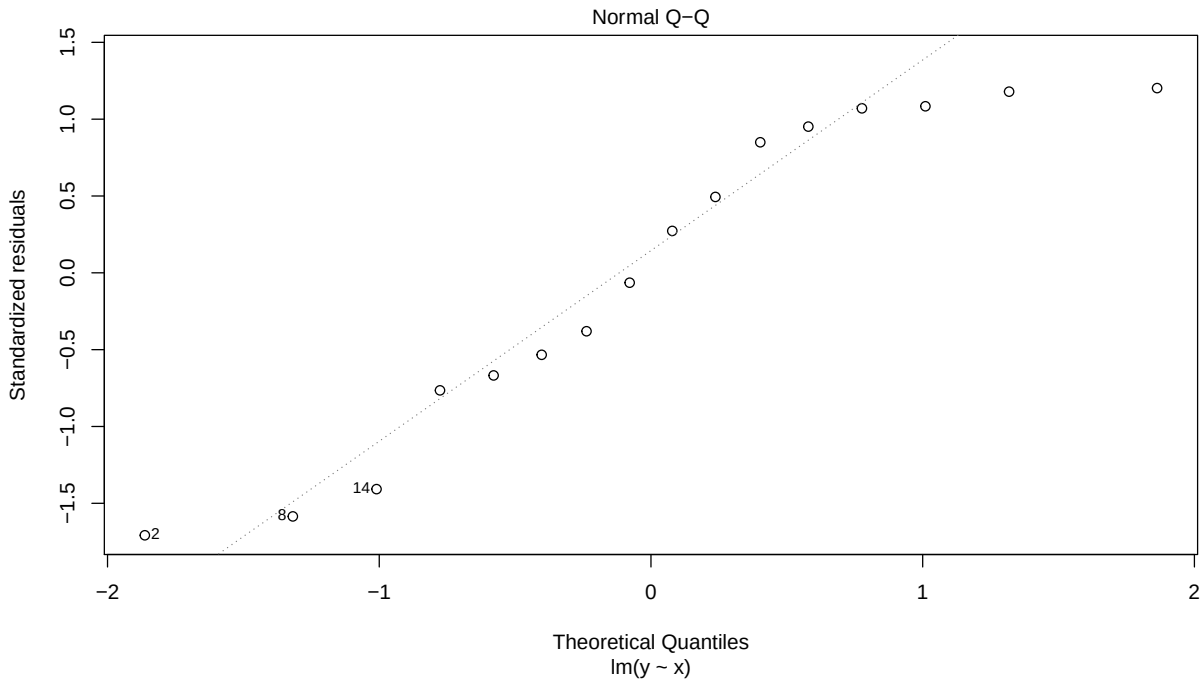
$SS_x = 3331.75$

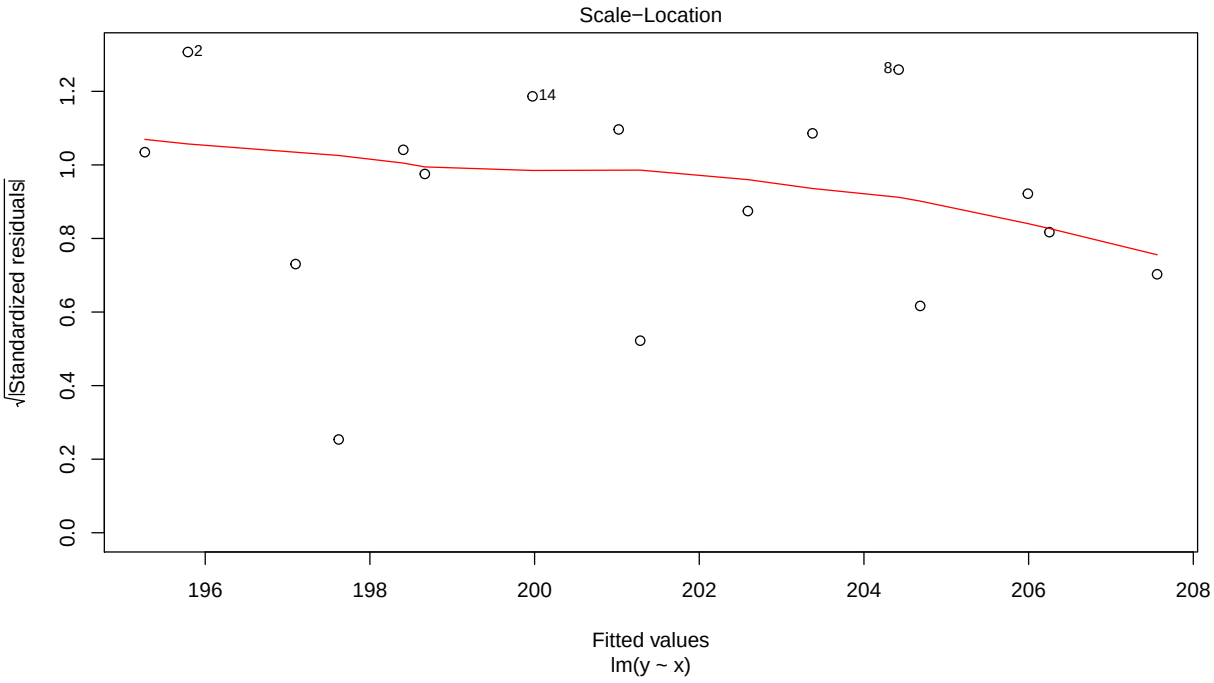
Accordingly, the 95% CI for β_1 is equal to [-0.12, 0.64]

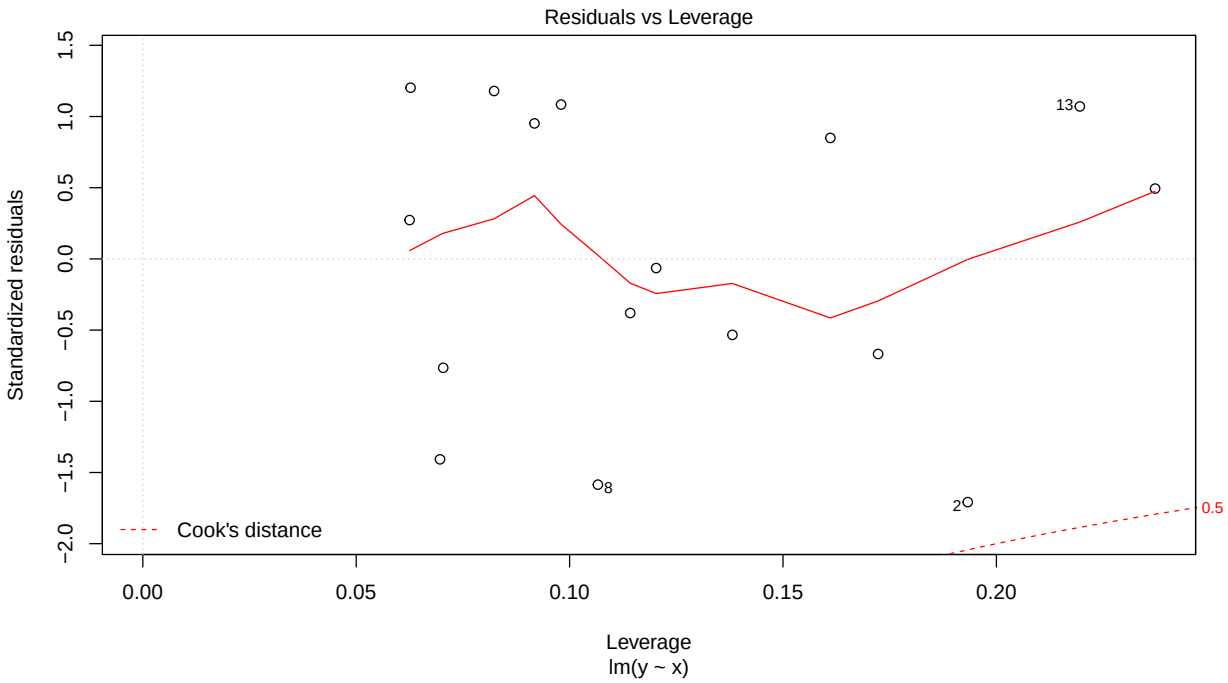
i.


Robbery Rate
180
185
190
195
200
205
210
215
50
60
70
80
90
Population Density


Residuals vs Fitted
Residuals
15
10
5
0
−5
−10
−15
2
14
8
196
198
200
202
204
206
208
Fitted values
lm(y ~ x)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
2
8
14


Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ x)
2
8
14



As we observe from the preceding plots, these data appear to display homoscedastic errors, they are independent samples, and they have also been randomly sampled. These characteristics taken together means the linear regression model fitted here is appropriate for the data.

Question 4

Analysis of Variance

Model DF: 1

Model SS: 9.874148

Model MSE: 9.874148

Model F: 53.27865

Error DF: 48

Error SS: 8.895852

Error MSE: 0.1853303

Corrected Total DF: 49

R-square: 0.5260601

Parameter Estimates

intercept standard error: 0.1793638

intercept standardized estimate: 0

size parameter estimate: 0.414932

Pr $>$ |t|: 1.2769322×10^{-9}

size standardized estimate: 0.7253