Name: Drashti Mehta

Lab Assignment 10

(40 points)

DIRECTIONS

- ***Be sure to show all of your work by providing your hypotheses, R commands, solutions/output, and interpretation.***
- ***Please circle or highlight your final answers.***

A. Answer these short answer questions.

1) **A student carefully computes the correlation coefficient by hand and gets the result of r = 1.36. What can you tell from this value (2 pts)?**

This value indicates that there was an error in the correlation measurement because r value must be between -1 and 1.

2) **A researcher found a positive correlation between temperature and grades. Lower temperatures corresponded with lower grades and higher temperature with higher grades. Is this enough to conclude that temperature causes grades to rise and fall (2 pts)?**

No because the correlation does not indicate causation. Temperature and grades may vary together but that doesn't mean one causes the other.

3) **Given a set of paired data (X and Y) give the expected correlation coefficient if**
   Y is independent of X:
   If Y is linearly dependent on X: (2 pts)?

If Y is independent of X, r ~ 0
If Y is linearly dependent on X, it should be approaching towards (or equal to) 1 or -1

4) **A scientist has a large number of data pairs (age, height) of humans from birth to death. He computes a correlation coefficient. Would you expect it to be positive or negative? Why? What would be the major problem with this approach (2 pts)?**

I would expect the correlation coefficient to be positive because your height increases as you grow older. One issue here is that most growth occurs early in life, after which height is fairly constant. We know that we don't expect the data to be linearly related when including points from whole lifespans.

5) **Explain SSR, SSE, and SST (2 pts).**

SSR is the sum square of the regression, SSE is the sum square of errors and SST is the total of sum squares. You calculate SSR by squaring all of the differences between y-bar and y-hat and summing them all up. For SST, you square all of the residuals and sum them up. SSE is the difference left when you subtract SSR from SST. You can also calculate SSE by subtracting all y-hats from y and squaring that result.

B. The following are the weights (kg) and blood glucose levels (mg/100ml) of 16 apparently healthy adult males.

Weight	Glucose
64	108
75.3	109
73	104
82.1	102
76.2	105
95.7	121
59.4	79
93.4	107
82.1	101
78.9	85
76.7	99
82.1	100
83.9	108
73	104

64.4	102
77.6	87

Find the correlation coefficient of the two variables by hand and in R. Test the correlation using R by finding the critical value and running the correlation in R (15 pts).

By hand:

H0: rho (ṕ) = 0
Ha: rho ≠ 0

	weight	glucose	x	y	xy	x^2	y^2
1.	64	108	-13.36	6.7	-89.51	178	44.89
2.	75.3	109	-2.06	7.7	-18.86	4.2436	58.29
3.	73	104	-4.36	2.7	-11.77	18.01	~~58~~7.29
4.	82.1	102	4.74	0.7	33.18	22.468	0.49
5.	76.2	105	-1.16	19.7	-42.92	1.3456	13.69
6.	95.7	121	18.84	-22.3	361.3	336.36	388.09
7.	59.4	79	-17.96	5.7	400.51	322.56	497.29
8.	93.4	107	16.04	-0.3	81.428	257.28	32.49
9.	82.1	101	4.74	-16.3	-1.422	28.468	0.9
10.	78.9	85	1.54	-2.3	-2.51	2.3716	265.69
11.	76.7	99	-0.66	-13	1.518	0.4356	5.29
12.	82.1	100	4.74	6.7	-6.162	22.468	1.69
13.	83.9	108	6.54	2.7	43.818	42.772	44.89
14.	73	104	-4.36	0.7	-11.77	19.01	7.29
15.	64.4	102	-12.96	-14.3	-9.072	167.96	0.49
16.	77.6	87	0.24	-	-3.432	0.576	204.49

~~723~~

sum: xy: 723.4875 1573.438

1419.298

$$r = \frac{xy}{\sqrt{x^2 y^2}} = \frac{(723.4875)}{[(1419.298) \times (1573.438)]^{1/2}} = \underline{\underline{0.484}}$$

test statistics $(t) = r\sqrt{\frac{n-2}{1-r^2}}$

$= 0.48\sqrt{\frac{16-2}{1-0.2304}}$

$= 0.48\sqrt{\frac{14}{0.7696}} = 0.48 \times \sqrt{18.191}$

$= \cancel{2.1047}\ 2.047.$

- ~~Hence~~, the critical value for $\alpha = 0.05$ & $df = 14$ is ~~2.147~~. 2.144.
- In R:

> qt (0.025, 14, lower.tail = F)
> 2.144

- ∴ Test statistics does not exceed the critical value, hence we ~~do not~~ have enough evidence to reject the Ho.
∴ we accept Ha that ~~H~~ $\rho \neq 0$

In R:

> weight<-c(64,75.3,73,82.1,76.2,95.7,59.4,93.4,82.1,78.9,76.7,82.1,83.9,73,64.4,77.6)
> glucose<-c(108,109,104,102,105,121,79,107,101,85,99,100,108,104,102,87)
> cor.test(weight,glucose)

Pearson's product-moment correlation

data: weight and glucose
t = 2.0703, df = 14, p-value = 0.0574
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.01521939 0.79020296
sample estimates:
cor
0.48413

C. After the Chernobyl accident, radiation levels were calculated at 7 different distances from the Chernobyl power plant (15 points total). Radiation levels are measured in seivert.

Distance (meters)	Radiation Levels (Sievert)
10	155
20	121
50	110
100	84
200	45
500	10
1000	4

1. **Perform a linear regression in hand and in R using between Distance and Radiation Level.**

C:1

			XY	X^2	Y^2
	10	155	1550	100	24025
	20	121	2420	400	14641
	50	110	5500	2500	12100
	100	84	8400	10000	7056
	200	45	9000	40000	2025
	500	10	5000	250000	100
	1000	4	4000	1000000	16
sum:	1880	529	35870	1303000	59963
mean:	268.5714	75.571	5124.286	186142.9	8566.143

$$B_1 = \frac{\left[xy(\text{sum}) - ((x\text{sum}) * (y\text{sum})/n)\right]}{\left[(x^2\text{sum}) - (x\text{sum}^2)/n\right]}$$

$$= \frac{35870 - (1880 \times 529)/7}{1303000 - (1880)^2/7}$$

$$= -0.13307$$

$$B_0 = y\ \text{mean} - (B_1 \times X\ \text{mean})$$

$$= 75.571 - (-0.13307 \times 268.5714)$$

$$= 111.3102$$

line $= 111.31 - 0.13307x$

In R:

```
distance<-c(10,20,50,100,200,500,1000)
radiation<-c(155,121,110,84,45,10,4)
mydata<-data.frame(radiation,distance)
mydata
distanceModel<-lm(radiation~distance, mydata)
distanceModel
Model2<-lm(radiation~distance-1, mydata)
Model2
summary(distanceModel)
```

Results: Coefficients: (Intercept) =111.3112, distance= -0.1331

2. **Calculate the critical value of the F-statistic, and the p-value in R. At 0.05 alpha, make a conclusion.**

Critical value: qf(0.95,1,5) = 6.607891, F statistic = 12.07, P-value = 0.01776
Because our test (F) statistic is greater than the critical value and our P-value is less than 0.05, we reject the null hypothesis.

3. **Interpret the R^2 describing the data and the regression.**

R2 = 0.7072 and that means that almost 71% of variation in the response variable(radiation) can be explained by the regression model.

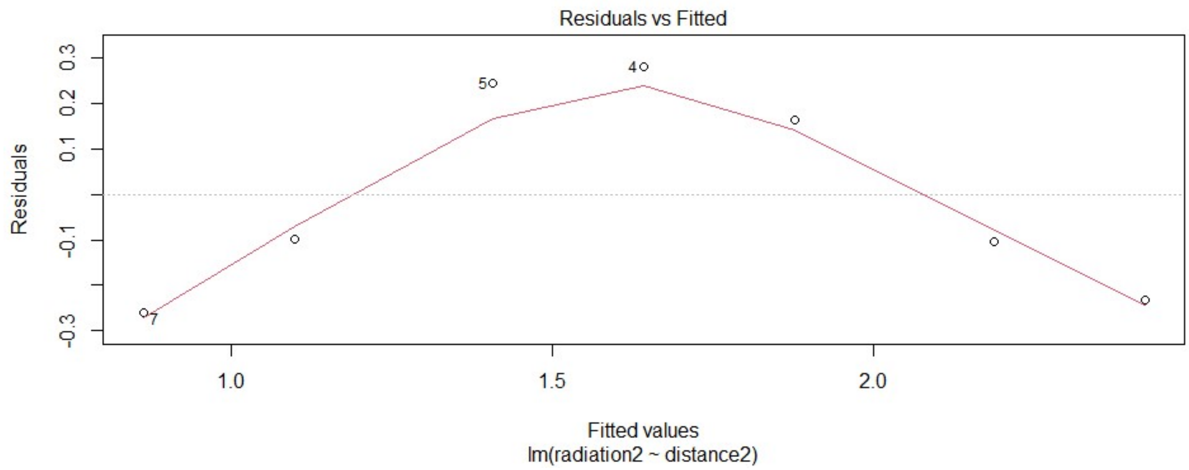
4. **Interpret the standard error describing radiation to the regression.**

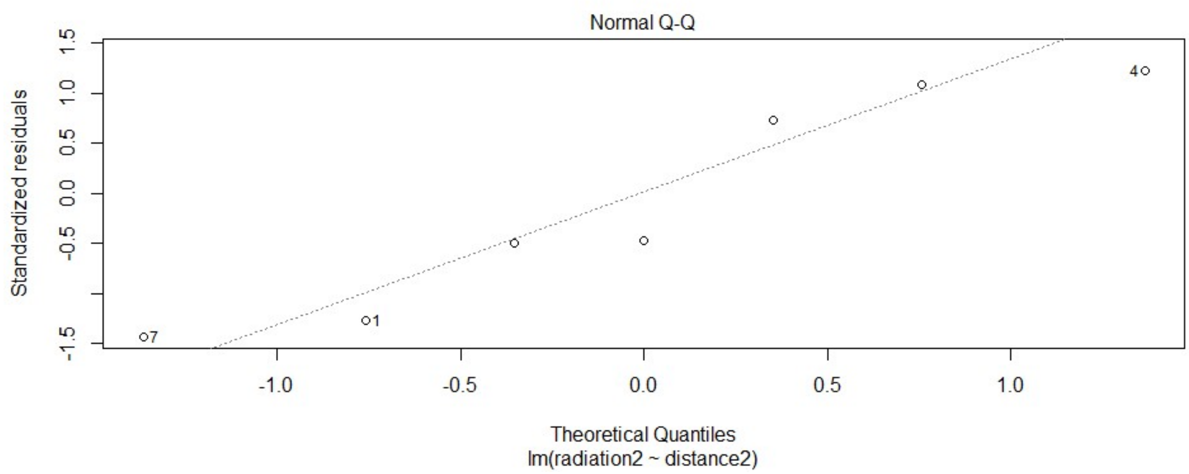
SSE = 16.5231 which can be interpreted as the total squared distance from the radiation levels to the regression line

5. **Create a data frame of the data that is log transformed. Perform single linear regression, interpret the results.**

```
distance2<-log10(c(10,20,50,100,200,500,1000))
radiation2<-log10(c(155,121,110,84,45,10,4))
mydata2<-data.frame(radiation2,distance2)
mydata2
distanceModel2<-lm(radiation2~distance2, mydata2)
distanceModel2
Model3<-lm(radiation2~distance2-1, mydata2)
```

Model3
summary(distanceModel2)
plot(distanceModel2)





line = 3.2010 - 0.7795x

Multiple R-squared: 0.8614, Adjusted R-squared: 0.8337
F-statistic: 31.07 on 1 and 5 DF, p-value: 0.002559 SSE = 0.295

The sum squared error was reduced significantly by the log transformation and there was an increase in R^2, so more of the variation in the response variable (radiation) could be explained by the regression model. We still reject the null hypothesis, due to a low p-value and F-statistic higher than the critical value.