STA 207 HW-3

Due Date: Sept. 19, by 10:20AM in Moodle

DERIN GEZGIN | CAMEL ID: 00468038

Problem 1: Fluorescence Experiment (45 points)

Suzanne Rohrback used a novel approach in a series of experiments to examine calcium-binding proteins. The variable Calcium is the log of the free calcium concentration and ProteinProp is the proportion of protein bound to calcium.

You may access the data by running the following R code in RStudio:

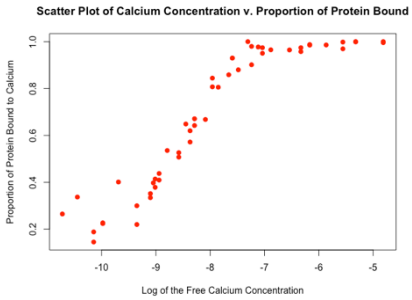
```
library(Stat2Data)
data("Fluorescence")
?Fluorescence  #this line gives you the details of the dataset
```

a. **[5 points] Report the correlations between the two variables, Calcium and ProteinProp through correlation coefficient and comment on the strength and direction of the relation.**

Correlation Coefficient between the variables → ***0.914***

*Correlation between the variables is **very strong** in the **positive** direction considering it's very close to 1.*

b. **[5 points] Make a scatter plot with Calcium as X and ProteinProp as Y. Comment on the relationship.**



There is a strong linear relationship from -9 to -7 in the positive direction. From -7 and beyond the relationship curves and becomes more horizontal. In general, we can say that it's not a linear relationship as it curves towards the end of the plot.

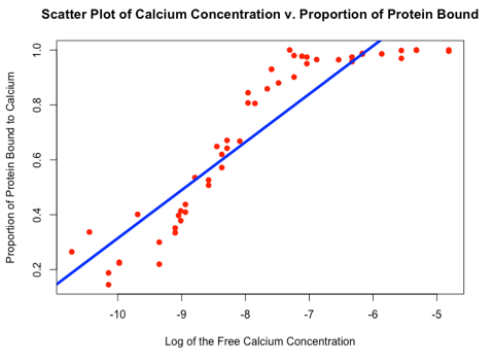
c. **[5 points] Fit an SLR for predicting the proportion of protein bound to calcium using the log of the free calcium concentration. Report the fitted model.**

The fitted model is $\hat{Y} = 2.066 + 0.175x$

x = Log of the free calcium concentration

$\hat{Y}$ = *Estimated proportion of Protein Bound to Calcium*

d. **[5 points] Plot the regression line and all the points on a scatterplot. Does it seem to be a good fit?**



*I think that it **seems** like a good fit as most of the points are around the line. But the part at the end can cause a problem. I'd say it's kind of fitted but there's a lot of points that'd have a high residual. I wouldn't classify this as a good fit as there are a lot of points that'd be classified with a high residual.*

e. [5 points] Interpret the slope estimate.

*The slope estimate is **0.175**. Which is the expected **difference** in the estimated "proportion of Protein bound to calcium" for each 1 unit increase in the "log of the Free Calcium Concentration."*

f. [5 points] Report the standard errors for regression parameter estimates and interpret them.

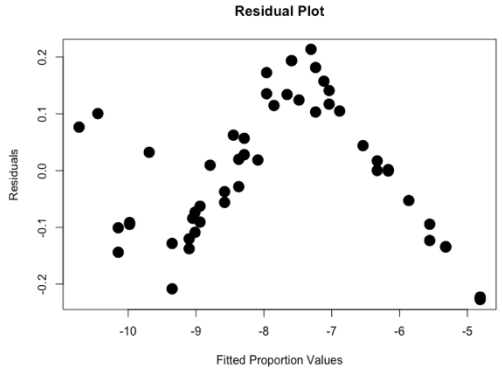
*The standard error for the Y-Intercept estimate ($\hat{Y}$) is 0.0889. This means that, on an **average**, we can expect ± **0.0889** sample to sample variability in the intercept estimate of 2.066.*

*The standard error for the slope estimate is ± **0.011**. This means that, on an **average**, we can expect ± **0.011** variability in the slope estimate of 0.175.*

g. [15 points] Is this model a good fit. Justify?

There are 3 different things we can check to assess if a model is good fit or not.

1. ***Coefficient Determination***
   For this factor, we can see that the Multiple R-squared values is ***0.836*** which means that the 84% of the proportion of protein bound to calcium can be explained by log of the free calcium concentration. It's the variability in Y explained by X.
2. ***Variability in Errors***
   The average distance between the observed values of "proportion of Protein bound to calcium" and fitted values of it is ***±0.1199***. It's very close to 0.
3. ***Residual Plot***



We can see that there's a somewhat clear step pattern in the residual plot.

Considering all these factors, I don't think that we can say there's an extremely strong relationship between these variables. While we have around 84% accuracy in how X explains Y, our residual plot has somewhat clear pattern which means that there's still a better way to explain the relationship between the variables.

Problem 2 (15 points)

To determine whether there is a relationship between Math SAT scores and the number of hours spent studying for the test? A study was conducted involving 20 students as they prepared for and took the Math section of the SAT Examination. The regression output for the study is shown below:

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	353.165	24.337	14.51	2.24e-11 ***
x	-	2.291	11.05	1.87e-09 ***

Answer the following:

(a) [5 points] Compute the slope estimate. Interpret the slope.

The formula of the $t\ value$ is $\frac{\widehat{B_i}-B_{i,H_0}}{se(\widehat{B_i})}$. In our case as we say that for H_0, B_i is also 0 our formula would be t $=\frac{\widehat{B_i}}{se(\widehat{B_i})}$. From the given output we already know that t value is 11.05. At the same time, we know that $se(\widehat{B_i})$ is 2.291. If we do the multiplication B_i would be **25.316**.

This means that there's expected 25.316 change on average in the student's SAT math scores per 1 hour increase in the study time of the students.

(b) [5 points] How much variability is there in the above estimate?

The standard error for the slope estimate is **± 2.291**. *This means that, on an* ***average****, we can expect* ***± 2.291*** *variability in the slope estimate of* ***25.316***.

(c) [5 points] Based on this estimate, is there a positive relation between math SAT score and number of hours of study. Justify.

There's a positive relationship as the slope is positive. If there's an hour increase in the study time of the students, there's expected 25.316 change in the student's SAT math scores. We can also say this because the P-Value is less than 0.05 (It has ***) and there's enough evidence to reject the null hypothesis and conclude that there's a relationship between two variables.

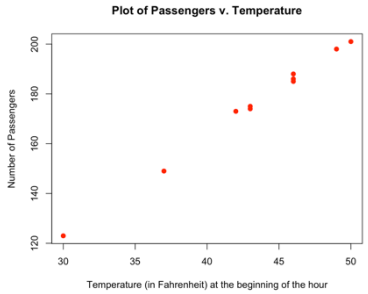
Problem 3 (15 points)

The city's transportation department is interested in studying the relationship between the temperature and number of passengers that ride the main bus line in order to better serve their customers. The manager recorded the temperature (in Fahrenheit) at the beginning of the hour, and then had a bus driver record the number of passengers that boarded the bus throughout the hour. Their findings are listed below:

Temperature	**Passengers**
42	173
37	149
46	185
30	123
50	201
43	174
43	175
46	188
46	186
49	198

The manager wishes to predict the number of passengers using the temperature.

(a) [5 points] Describe the relation between temperature and the number of passengers? Use scatterplot and correlation coefficient.



The correlation coefficient is **0.998** which makes it an extremely strong relationship. If we look at the scatter plot, we can also see that there's an obvious relationship between these variables.

(b) [5 points] Obtain the least squares regression equation and report it.

As the estimated *Y-Intercept* is 4.413 and estimated slope is 3.953 the least squares equation would be $\hat{Y} = 4.413 + 3.953x$

In this equation $\hat{Y}$ is Number of passengers and **x** is temperature in Fahrenheit.

(c) [5 points] Find residual for temperature=42 degrees Fahrenheit and passengers=173.

The predicted value would be $\hat{Y}$ = 4.413 + 3.952 * 42 = 170.397 passangers.

The residual is 173 – 170.397 = 2.603.

When I did *fit$residuals* the residual value I got is *2.544*. This difference is probably because I round the values in the least squares regression equation.

Problem 4 [25 points]

Two undergraduate students took a random sample of 30 textbooks from the campus bookstore. They recorded the price and number of pages in each book, in order to investigate the question of whether the number of pages can be used to predict price. The data can be accessed by running:

```
install.packages(Stat2Data)
library(Stat2Data)
data("TextPrices")
?TextPrices
```

a. [2.5 points] Produce a scatterplot to investigate the students' question. What does the plot reveal?

The plot shows that there's an upward relationship between the number of pages and the price of a book. But the points are very distributed so that it's hard to see a strong relationship.

b. [7.5 points] What is the fitted model equation for predicting price from number of pages. Make a scatterplot with overlayed fitted line and comment on the fit.

The fitted model has an estimated Y intercept of -3.422 and an estimated slope of 0.147. In this case, the fitted model would be

$$\hat{Y} = -3.422 + 0.147X.$$

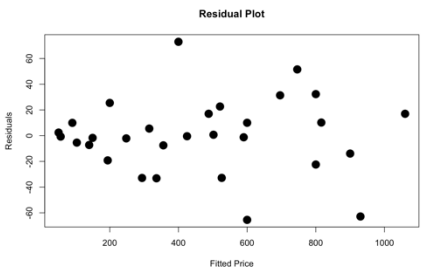
Where $\hat{Y}$ is the price and **x** is number of pages.

While the fitted line gives a general estimate for the direction of the relationship, I think that it's a bit far from being good at explaining the relationship between the variables as there's significant distance between the sample points and the fit line especially in 500-1000 range.

c. [15 points] Is this a good model? Justify.

There are 3 different things we can check to assess if a model is good fit or not.

1. ***Coefficient Determination***
   For this factor, we can see that the Multiple R-squared value is ***0.677*** which means that 68% of the book prices can be explained by the number of pages in a book. It's the variability in Y explained by X.
2. ***Variability in Errors***
   The average distance between the observed values of book prices differ from the fitted values of book prices is ***29.76***.
3. ***Residual Plot***



In this graph, we can see that there's somewhat random distribution in the residuals.

In general, our model can be classified as a somewhat good model as our multiple R-squared value is ***0.677***. It's certainly better than %50 but at the same time it's not as good as %90. At the same time, we can see that the variability in the errors is too big while our residual plot is somewhat random. I won't classify this as a very good model but also, I won't tell it a bad model.