STA 207 HW-3 Solution

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:

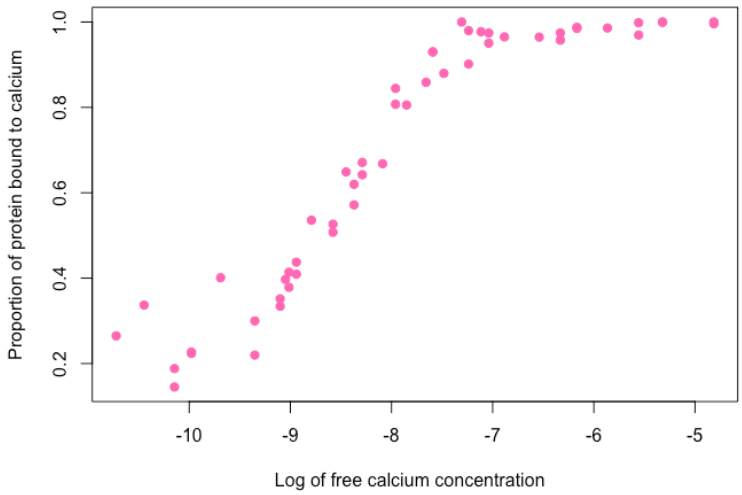
```
library(Stat2Data)
data("Fluorescence")
names(Fluorescence) #check variable names
```

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.

The correlation between log of the free calcium concentration and the proportion of protein bound to calcium is 0.914, indicating a strong and positive linear relation.

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

Figure 1



The relationship is strong and positive, however the pattern appears to be non-linear especially after the proportion arrives at the value of 1.

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.

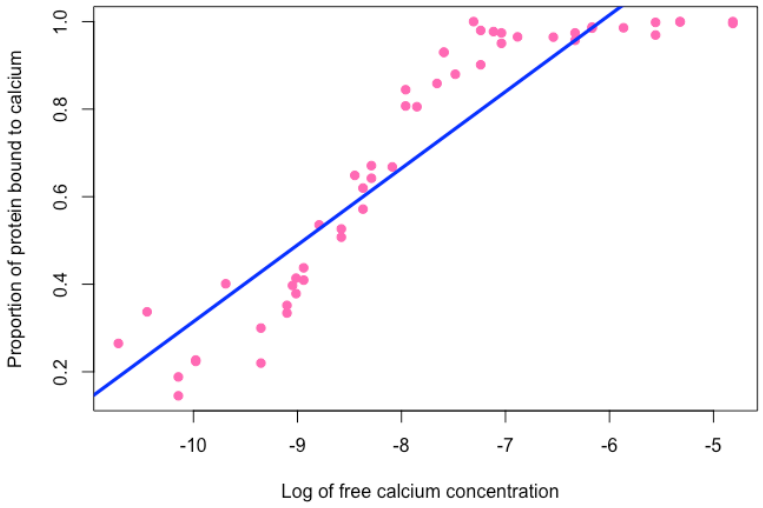
$$\hat{Y} = 2.066 + 0.175X,$$

X is the free calcium concentration, and
Y is the 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?

At first glance, the model seems to capture the general trend, but we do see a curve or changing slope shown in the dots. Therefore, a linear model (straight line) is unable to explain some parts of the relationship.

Figure 2

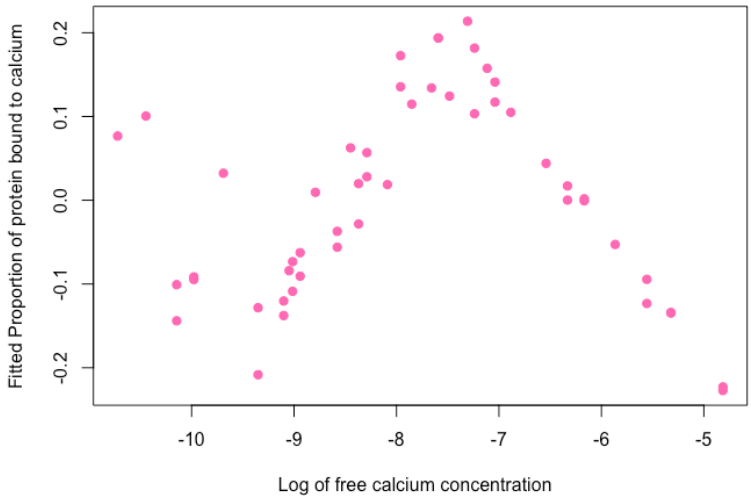


e. [5 points] Interpret the slope estimate.
For a one unit increase in log of the free calcium concentration, we expect the proportion of protein bound to calcium to increase by 0.175.

f. [5 points] Report the standard errors for regression parameter estimates and interpret them.
The sample to sample variability in the intercept estimate is 0.089 and the variability in the slope is 0.011.

g. [15 points] Is this model a good fit. Justify?
   a.) The coefficient of determination is 0.8363, so 83.63% variation in proportion of protein bound to calcium is explained by log of the free calcium concentration.
   b.) The residual standard error is 0.1199, thus showing that the average variation in fitted model is approximately 0.1199, which seems moderate.
   c.) The residual plot in Figure 3 shows a clear pattern so the model is not able to extract the entire information.

Figure 3



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] With Y as MAT SAT scores and X as the # of hours spent studying for the test, compute the slope estimate. Interpret the slope.

With Y as MAT SAT scores and X as the # of hours spent studying for the test,

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

Thus, $\hat{\beta}_1 = 11.05 * 2.291 = 25.32$

For every additional hour someone spend studying, the average test score is expected to increase by 25.32 points.

(b) How much variability is there in the above slope estimate? The sample-to-sample variability in slope estimate is 2.3 points.

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

Yes, the slope is positive, indicating an increase in average math SAT score as number of hours spent studying increased.

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:

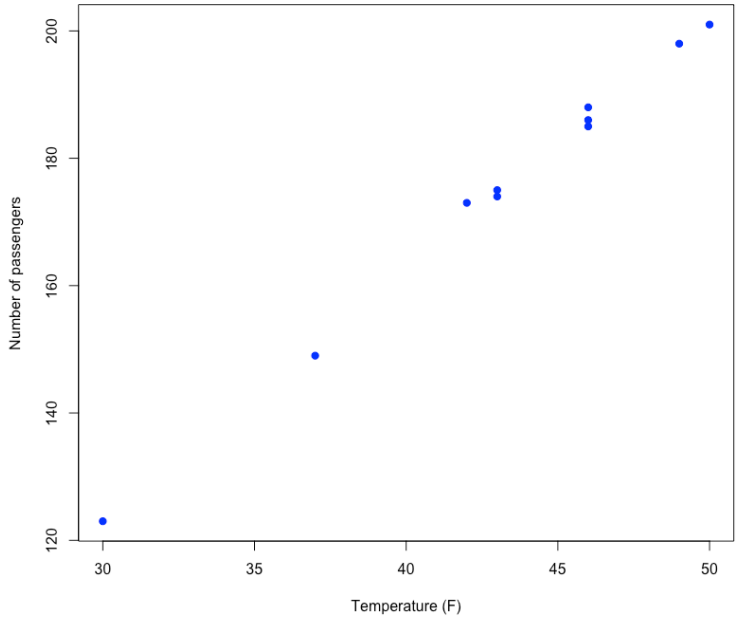
Temperature	**Passengers**
42	173
37	149
46	185
30	123
50	201
43	174
43	175
46	188
46	186
49	198

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 between the # of passengers and temperature is 0.998, which is almost a perfect, positive linear relationship. The scatterplot in Figure 1 below shows that the points are aligned in a perfect positive linear way. The small sample size is important to note.

Figure 1



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

The least squares regression equation is:

$$\hat{Y} = 4.413 + 3.953X,$$

X is the temperature in Fahrenheit, and

Y is the number of passengers

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

We know that for any value of X, we have an observed Y value and a predicted value. The residual is the difference of these two Y values, that is, $e = Y - \hat{Y}$. For $X = 42, Y = 173$ (data value) and the predicted value is

$$\hat{Y} = 4.413 + 3.953(42)$$
$$= 170.439,$$

Thus, the residual is

$$e = Y - \hat{Y}$$
$$= 173 - 170.439 = 2.561$$

Using R, we also see that the residual corresponding to X=42 is 2.54408060. Since we are using slope and intercept rounded to three decimals and R's calculation is not, the $\hat{Y} = 170.439$ in this

calculation and R has a $\hat{Y} = 170.4559$, therefore there is a small numerical difference due to rounding.

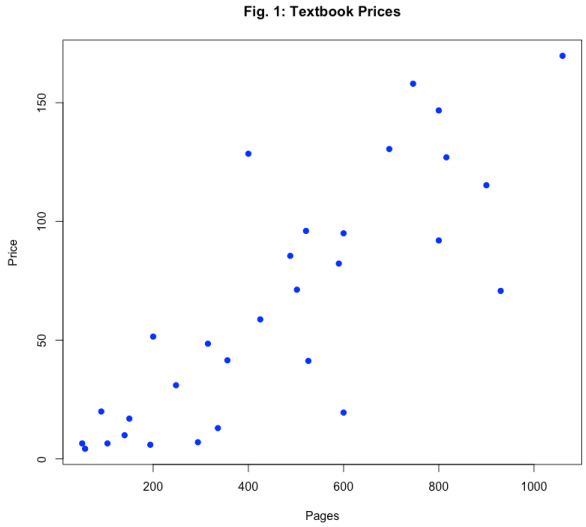
Problem 4 [25 points]

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 scatterplot below shows a positive, strong, linear relationship between the number of pages and price of textbooks. I used correlation between the variables which was 0.82 to determine the strength of the relationship.



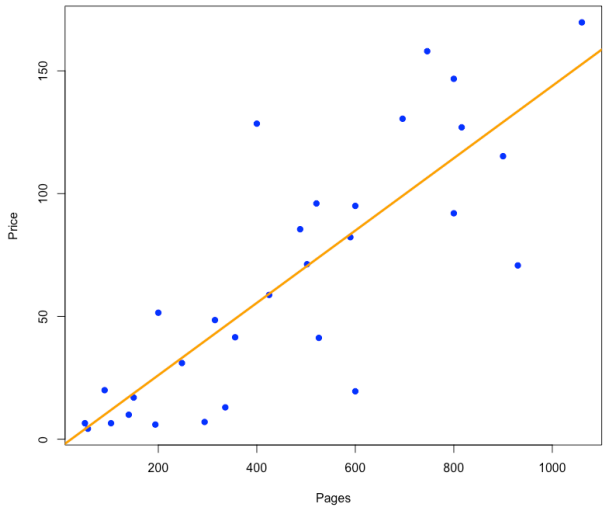
b. [7.5 points] What is the fitted model equation for predicting price from number of pages. Make a scatterplot with overlayed fitted line and comment on the fit.

$\hat{Y} = -3.42 + 0.15X$,

Where Y: Textbook price in $ and X is the number of pages.

The fitted model seems to be fitting the data well explaining the overall trend as represented in Fig. 2 below.

Fig. 2: Textbook Prices



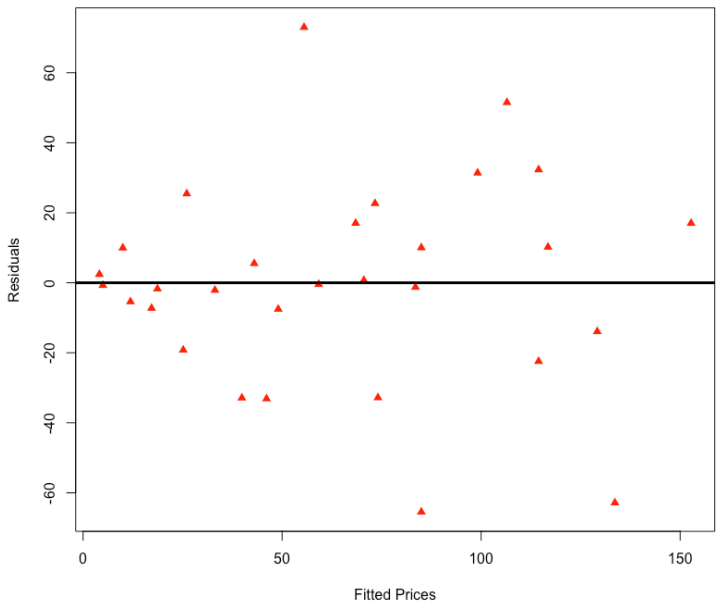
c. [15 points] Is this a good model? Justify.

a.) The coefficient of determination is 0.68, so 68% variation in textbook prices is explained by the number of pages.

b.) The residual standard error is 29.76$, thus showing that the average variation in fitted model is approximately 30$, which seems large in this context.

c.) The residual plot is not completely random as there is some pattern left in Fig. 3 below.

Fig. 3: Residual Plot



Overall, the fit is moderate.