Homework 4

Started: Feb 18 at 7:01pm

Quiz Instructions

Welcome to homework 4, which covers the introduction to the lm() command

Your instructions are here: **Homework-4-2020.html**

The brief R guide is here: **R-guide.html**

and data

- **GDP_vs_Satisfaction.csv**
- **playbill.csv**

The deadline for this homework is 23.59 TUESDAY 18-FEB

You can resubmit this quiz as many times as you like before the deadline, but no feedback will be given before then.

How to upload pictures. Click the little image, then choose the canvas tab, click on my files, then you can click upload files. you might want to make a new folder in my files in canvas so that you can keep them all in the same place.

Question 1 5 pts

What is the response variable for your dataset? There is also an optional comment section at the end of the quiz in case you want to clarify any answers.

- ○ Production
- ◉ Current Week
- ○ Last week
- ○

Question 2

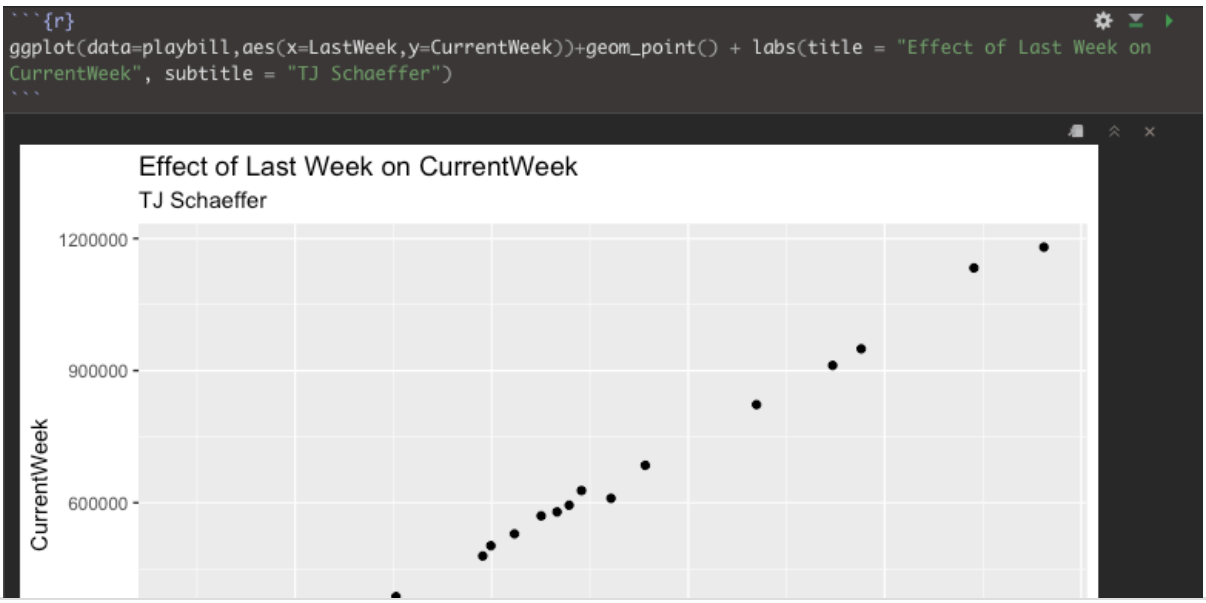
5 pts

Upload a screenshot of a scatterplot of your data and your code. For full marks this should be neat, with all axes labelled and a title.

Additional bonus point for a ggplot2 graph. (e.g. 6/5)

HTML Editor

```{r}
ggplot(data=playbill,aes(x=LastWeek,y=CurrentWeek))+geom_point() + labs(title = "Effect of Last Week on CurrentWeek", subtitle = "TJ Schaeffer")
```



p

0 words

Question 3

5 pts

Upload a screenshot of your linear model code and outputs.

HTML Editor

12pt Parag

```{r}
lm(CurrentWeek ~ LastWeek, data = playbill)
```

```
Call:
lm(formula = CurrentWeek ~ LastWeek, data = playbill)

Coefficients:
(Intercept)     LastWeek
  6804.8860       0.9821
```

p 0 words

Question 4 6 pts

CurrentWeek = 6804.8860 + 0.9821 LastWeek

Your answers should be to 4 decimal places

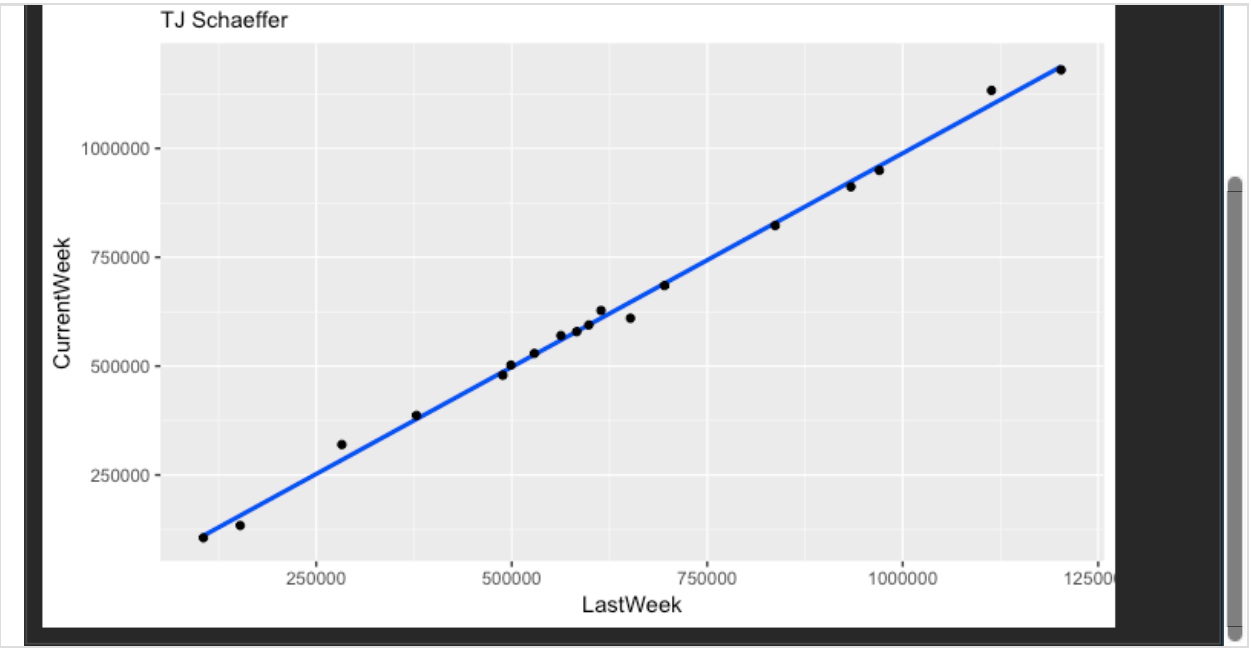
3 points for each of b0 and b1

Question 5 5 pts

Make another scatterplot of the data and overlay the fitted regression line onto the scatterplot using the abline() command. Use your favorite colour for the fitted line. Attach a screenshot of the R code line/lines and the plot. Here is a link to an R colour palette. (http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf (http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf)) or you can use hex codes for any color at all.

4 points for the line and 1 for the colour

HTML Editor



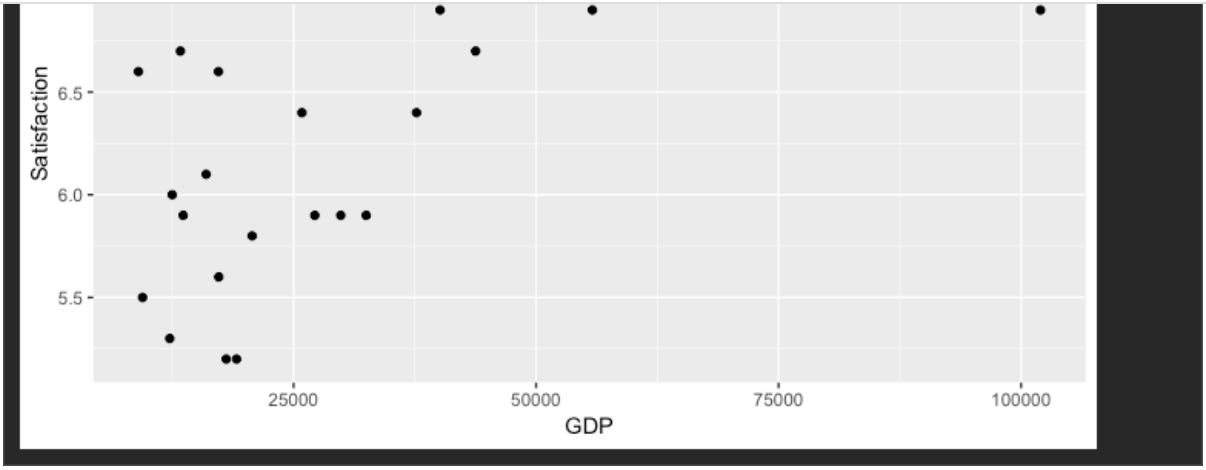
p 0 words

Question 6 6 pts

Comment on the GDP vs Satisfaction data. 3 marks given for showing you have explored the data (feel free to use screenshots). 3 marks for commenting on the

suitability of the data for regression

HTML Editor



The response variable is Life Satisfaction, GDP is the explanatory variable, and the relationship seems to be a strong, positive linear relationship despite some outliers.

p 25 words

Question 7 12 pts

If H0: r=0 and H1: r != 0, then there is evidence at a significance level of 5% to reject the null hypothesis.

Given the amount of data we have, there is a 95% chance that our r value falls between 0.4593531 and 0.8304061

3 marks for each

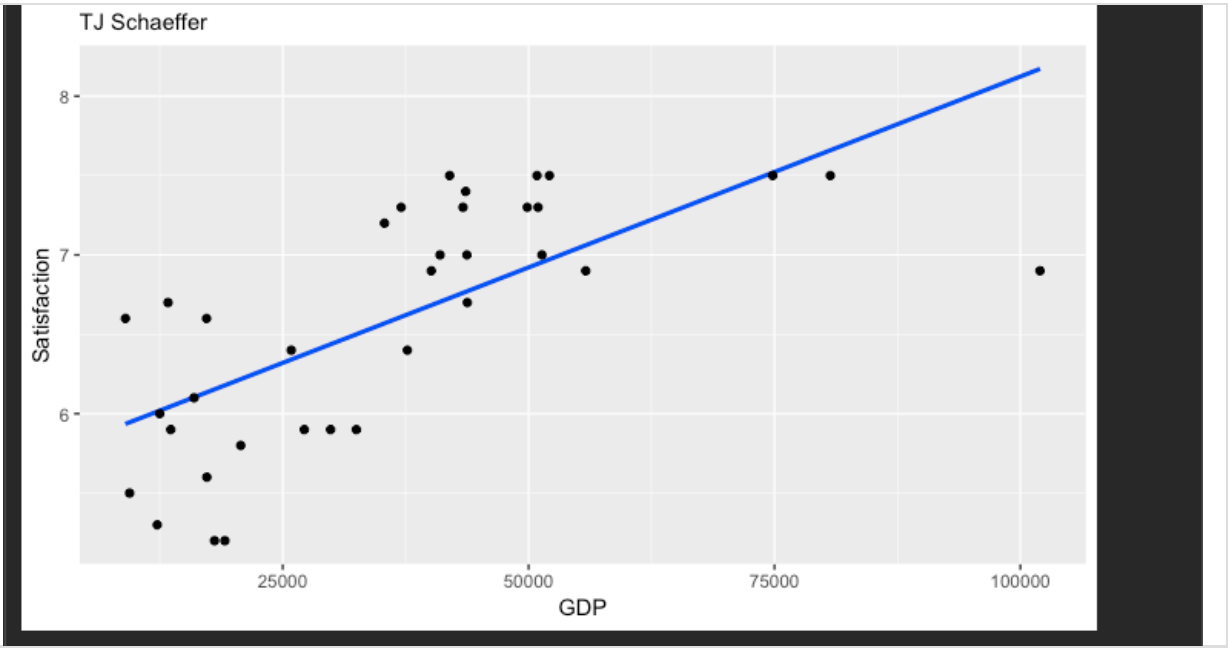
Question 8

6 pts

Regardless your opinion in part (b) fit a simple linear regression (SLR) model to this data. Attach a screenshots of your code, the R output of the SLR fit and the scatterplot with fitted line overlaid.

[3 marks for the fit and 3 for the plot]

HTML Editor



p

0 words

Question 9

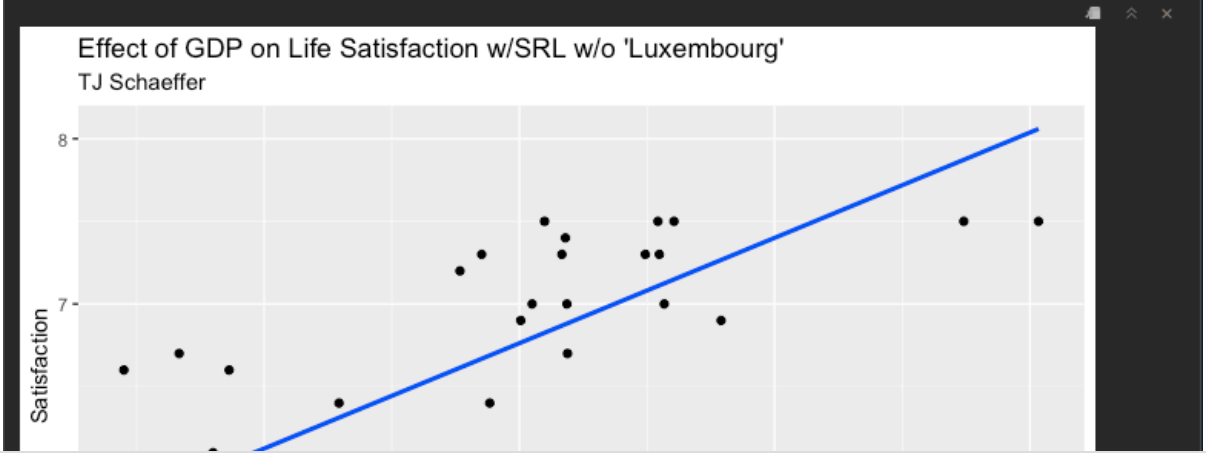
6 pts

In the original data, observation number 21 (“Luxembourg”) seems to be an outlier. Use the below code line to remove that observation from the dataset. Then re-fit the regression model to the reduced dataset.Attach the R output and the scatterplot of the reduced data with fitted model overlaid.

HTML Editor

12pt Parag

```{r}
ggplot(data=GDPSnew,aes(x=GDP,y=Satisfaction))+geom_point() + geom_smooth(method = "lm", se = FALSE, color = "blue4") + labs(title = "Effect of GDP on Life Satisfaction w/SRL w/o 'Luxembourg'", subtitle = "TJ Schaeffer")
```



p 0 words

Question 10 5 pts

Does it seems like model fit has improved once the “Luxembourg” was removed from the dataset? Explain

HTML Editor

12pt Parag

After viewing the plot, it does seem like the model fit has improved once "Luxembourg" was removed from the dataset because the regression line doesn't have to account for the outlier, which is evident in the increase from a .687 correlation with the "Luxembourg" point to .771 correlation without the "Luxembourg" point.

p 52 words

Question 11 5 pts

Write the equation of the least squares regression line of salary on years of experience.

HTML Editor

Salary = 25792.2 + 9450(YearsExperience)

p 6 words

Question 12 5 pts

IN YOUR OWN WORDS, Interpret the slope and the intercept in the context of the problem.

HTML Editor

Based on the equation before, we can determine that the average starting salary of all jobs is $25,792.20 and as a worker increases with years of experience, the salary increases by $9450 per year.

p 36 words

Question 13 5 pts

What is the coefficient of determination? We will cover this in Friday's lecture, or you can find out here: https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/a/r-squared-intuition (https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/a/r-squared-intuition) or any other good stats book

HTML Editor

12pt Parag

R^2 = SSReg / SST

SSReg = $2.0857 \cdot 10^{10}$

SST = $2.0857 \cdot 10^{10} + 9.3813 \cdot 10^{8} = 2.1795 \times 10^{10}$

R^2 = SSReg / SST = $2.0857 \cdot 10^{10} / 2.1795 \times 10^{10} = .957$

Based on the coefficient of determination, 95.7% of the the data in the salarydata is accounted for by the fitted regression line.

p 57 words

Question 14

4 pts

Q14. Add your **Rmd** CODE file here

Upload STAT462_HW4.Rmd

Your file has been successfully uploaded.

Question 15

0 pts

If you have any comments or clarifications for any question, there is space to write them here.

HTML Editor

12pt Parag

None

p 1 words

Quiz saved at 10:43pm Submit Quiz