**Questions**

**1 Interpretations**

**1.1 Descriptives with scatterplots (2 points total)**

For each of the following plots

1. Identify the strength of the relationship of the two variables in the data (eg: weak, moderate or strong, linear or not, any obvious trends).
2. Would it be reasonable to ﬁt a linear model? Why or why not?

\*\*\*In an order of up-to-down and left-to-right, I name the graphs from Graph 1 to Graph6\*\*\*

Graph1

In this graph, the relationship between two variables is moderate and linear, with an obvious trend. The points spreads out and cover a large range, but general trend is easy to tell as linear. So it is reasonable to fit a linear model with negative gratitude.

Graph2

The relationship between two variables in this graph is moderate and very possible to be linear. Applying a linear model is reasonable but with a lot of disturbance. The general trend is obvious that as x increase, y increase for sure, but there could exist some other possible model except linear for the graph, such as polynomial and exponential.

Graph3

The relationship between two variables is extremely weak, thus the specific relationship is hard to predict. No obvious trend could be observed in this graph. It could be a linear or anything else. It is worthy trying the linear model because any model would have a big variance.

Graph4

In this graph, the relationship is strong and obvious, and this is not a linear relationship. The trend seems more likely to be a secondary polynomial opening up.

Graph5

The relationship between two variables is strong. Dots in the graph with coordinates on x-axis larger than -1 is extremely concentrated, which almost form a line. Although there are numerous possibility for the graph since the points when x less than 1 spread out. Even though points with small x value covers a large range, but the trend is more likely to be exponential or polynomial, so it is not very reasonable to fit a linear model.

Graph6

Two variables have a strong relationship, which seems to be a parabola. Since there trend going up (where x increase, y increase) before the maximum y value where x is approximately -1, and also a trend going downwards after the maximum y value with more points. Hence, a linear line would not be able to represent the whole trend.

**1.2 Correlations (2 points total)**

Match the following calculated correlations to the corresponding six scatter-plots above:

-0.7, -0.7, 0.77, 0.3, -0.9, -0.01. Give a brief justification for your choices.

\*\*\*Use the same naming way as 1.1 \*\*\*

From large to small, place the calculated correlations in order. Then we have:

0.77, 0.3, -0.01, -0.7, -0.7, -0.9

Then list the graph in order: firstly find positive slope graph, and sort from graphs with concentrated points and close to a linear line to graphs with points spreading out. Secondly find negative slope and sort from graphs with dots spreading out to graphs with concentrated dots in linear shape. For graphs that is not reasonable to fit linear, classify them as ‘Close to no correlation’ for now.

Positive slop graphs are: Graph2,

Negative slope graphs are: Graph1,

Close to no correlation: Graph3, Graph4, Graph5, Graph6.

Since Graph3 really does not have a trend, so the correlation would be around zero. In Graph4, more points exist on the side with positive slope, so its correlation is likely to be a small positive number. Graph5 would have negative correlation because more than half of the dots form a line close to linear with negative slope. In Graph6, much more dots exist on the side with negative slope, so it is more likely to have a negative correlation.

Then most likely, the order of graph from large correlation to small would be:

Graph2, Graph4, Graph3, Graph5, Graph6, Graph1

Therefore match the correlations with the graphs:

Graph1: -0.9

Graph2: 0.77;

Graph3: -0.01

Graph4: 0.3

Graph5: -0.7

Graph6: -0.7

**1.3 Residuals (2 points per model)**

A statistician fitted several regression models to different variable pairs. For each model, she ran diagnostic plots for the residuals. For each of the following residual plots say which, if any, of the following assumptions are violated:

1. No mean trend
2. Normal distribution
3. Constant variance

Based on your answer, which models are good fits?

**Model 1:**

Here a) no mean trend is violated because the points are pretty concentrated around the line, not spreading out randomly, even though the line is almost horizontal, it shows that there is a trend.

From Normal Q-Q graph, it is a normal distribution. However, scale-Location have a little bit concern because the points could have a trend going up, but since the points are not very concentrated, I would say constant variance is not violated.

**Model 2:**

Both a) No mean trend and c) constant variance are violated. They have obvious trend, and all points are very concentrated along the trend.

Normal distribution assumption concerns me due to some outliers, but since only three points are far from the linear line, I would say it is not violated. The three points are also shown on the residuals vs Leverage graph as influential cases.

**Model 3:**

No obvious violation.

Both Residuals vs Fitted graph and Scale-Location graph have dots randomly spread out with fitted line very close to be horizontal. Only Normal Q-Q concerns me a little bit because the extreme values are a little bit off the linear trend, but they are not far from the trend, so it should not violate the assumptions.

**Model 4:**

All three assumptions are violated. Residuals vs Fitted has a parabola trend and dots are very concentrated. Normal Q-Q graphs have a curve, and almost half of the points are off the linear fitted line. Scale-Location graph also have an obvious trend.

**Model 5:**

No obvious violation

All three graphs are close to what it should looks like if all assumptions are not violated. Residuals vs fitted and scale-location graphs have dots spread out randomly with fitted line almost horizontal. Although scale-location graph’s fitted line is a little bit tilted, but no trend could be observed correspondingly. As for the Normal Q-Q graph, even though there exist some small distortion, generally, all points are lying along the linear line. No influential outlier cases exist in this case.

**Model 6:**

No obvious violation:

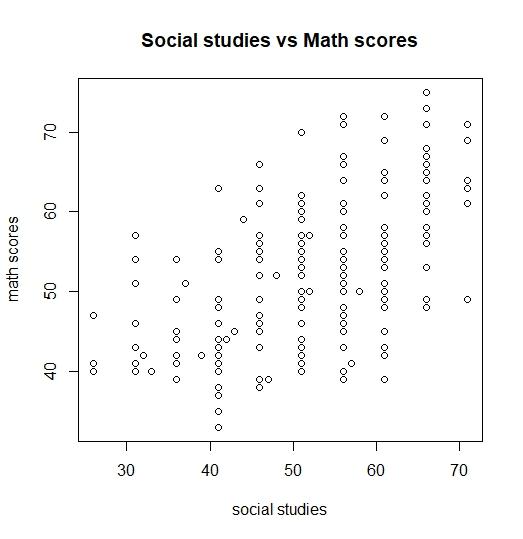
Both residuals vs fitted and scale-location graphs have horizontal fitted line and scatter points randomly. Even though in normal Q-Q graph, some extreme values, especially between -3 and -2 standardized residuals, are off the linear fitted line, but they are not far from the whole trend, and the rest of the points are nicely lying along the trend.

**Overall, model 3, model 5, and model 6 are good fits.**

**2 Linear regression in R**

Consider the hsb2 dataset.

1. **Do a scatter plot of social studies (socst) scores vs math scores (math). Report the correlation. Describe the strength in relationship between these two variables**

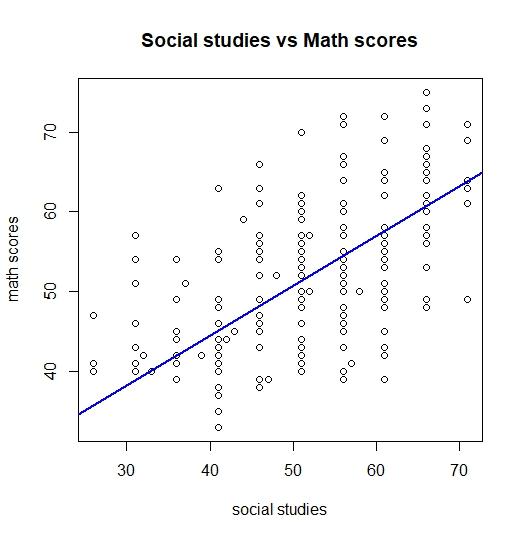
****

The correlation is 0.544803 ~= 0.545.

The relationship between two variables are moderate. We can observe that as students’ social studies score increase, their math score also increase. However, the dots are not concentrated but spread widely.

1. **Fit a linear regression model that can be used to predict social studies scores based on math scores. Write down the model equation that R gives you. Overlay this regression line on top of the scatterplot.**

According to the intercept and coefficient given by R, the line equation is:



Interpret the intercept and the slope. Is the intercept meaningful? Why or why not?

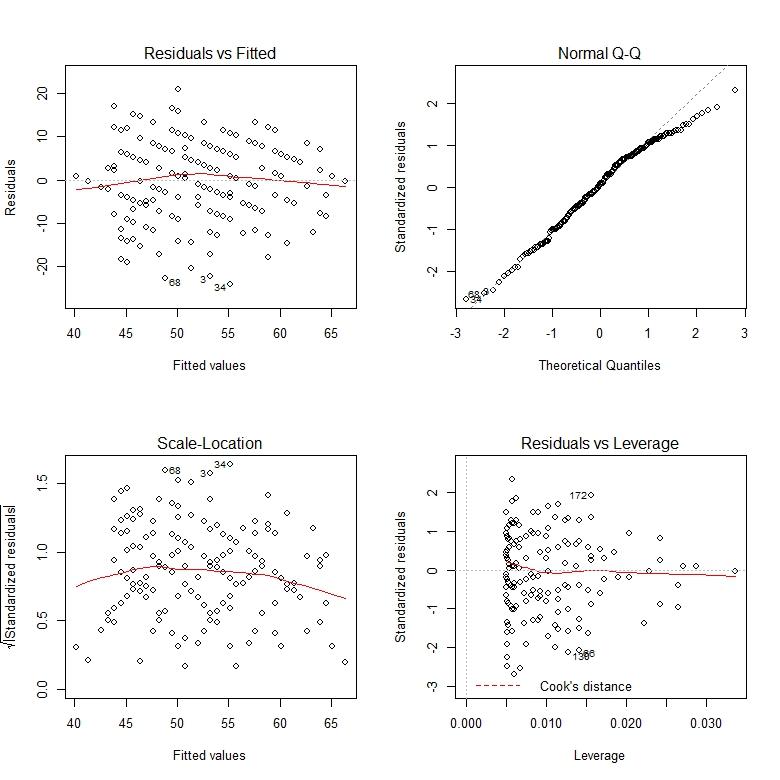
The intercept means when the student have zero math score, they averagely still get around 20 points. This is a meaningful number to predicting how well a student will do in social studies if knowing their math level. Nonetheless, we need to consider if it comes to a student who can do zero on a test, then he or she could be in a pretty different situation from the students that are included in this dataset, where people at least have thirties.

1. **R reports the p-value of the slope and the intercept. What is the meaning of these p-values? What are the null and alternative hypotheses in these tests? For the p-values in your model, what can you conclude?**

P-values for slope and intercept means if it is significant to have such value for the whole trend. The null hypothesis is that it is not significant to have this variable to illustrate the trend of data. The alternative hypothesis is that it is significant to include the variable with such coefficient or intercept. Basically, we compare the given intercept value or coefficient to zero.

R gives us the p-values 2.36-07 for intercept and <2e-16 for coefficient of variable math. Since both of them are smaller than 0.01, it is significant to include both value to illustrate the trend of data.

1. **Do a diagnostic plot for your model. Say which, if any, of the (a) no mean trend, (b) normal distribution and (c) constant variance assumptions are violated. Based on your answer, is the model a good fit? That is, is there a linear relationship between social studies scores and math scores?**



Residual vs Fitted graph has a horizontal fitted line with dots pretty spread out randomly, so the assumption of no mean trend holds. For Normal Q-Q plot, only a little bit of high extreme points are off the linear fit line, but generally lies on the linear trend. Thus, the normal distribution assumption also holds. In scale-location graph, the fitted line is almost horizontal, and no obvious trend exist in the points. Hence, the third assumption of constant variance holds.

Since the model applies all three assumptions, the model should be a good model. The relationship between social studies and math is linear.

1. **Report the R2 value of your model. What is the meaning of this value?**

R gives the following value for r square:

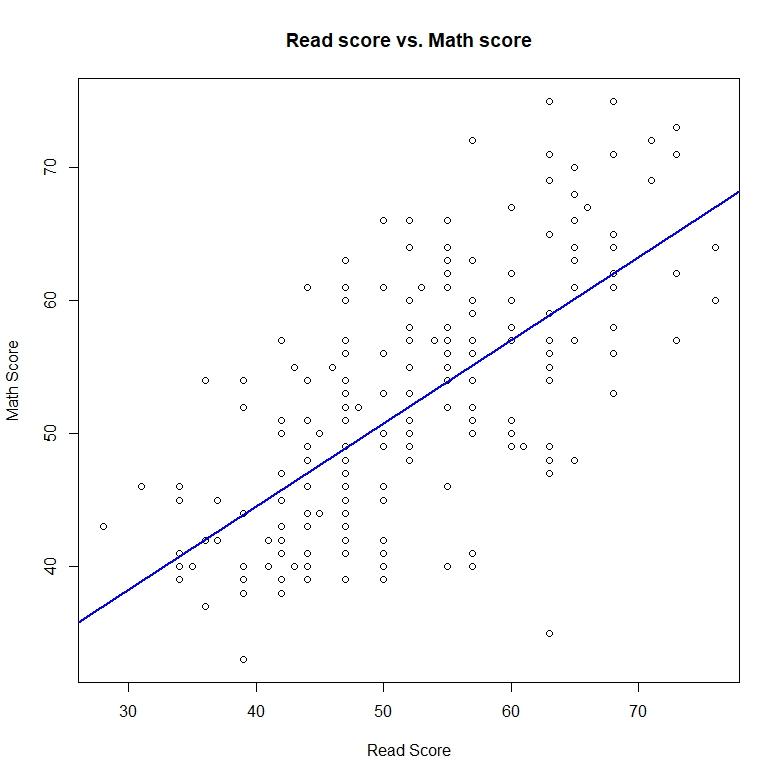
**Multiple R-squared: 0.2965, Adjusted R-squared: 0.2929**

The r-squared value is very low, which means the many of the dots are far away from the fitted line.

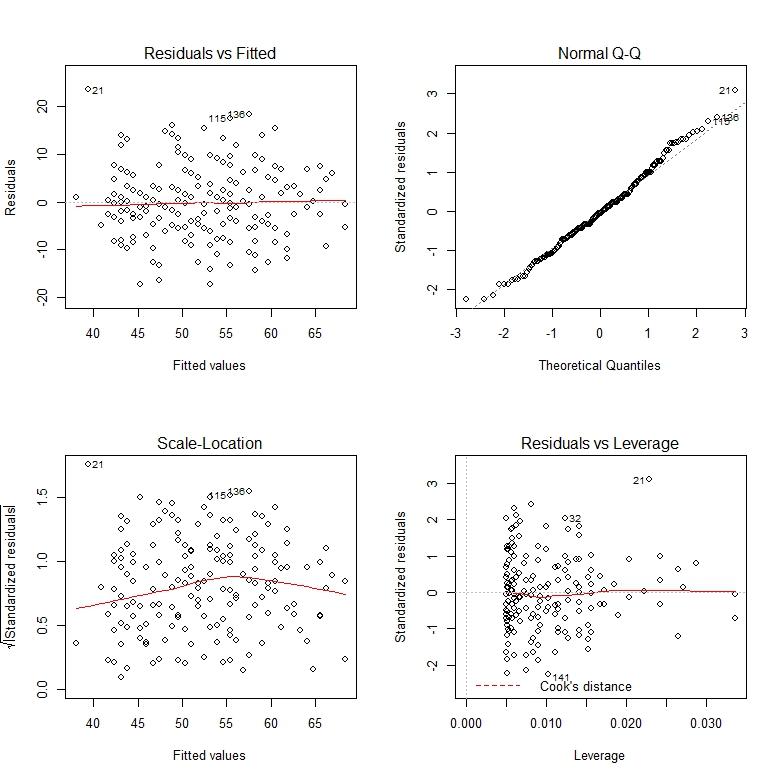
1. **Summarize the relationship between social studies scores and math scores in a paragraph that utilizes all of the numbers in the previous questions.**

The general relationship between social studies scores and math scores is linear, which means as math score increase, the social studies scores also increase. However, this is the average trend, so each individuals have very different cases, which mean two student have same math scores, could have very different score in social studies.

**Bonus question: perform the same analysis for read scores vs math scores. Provide a scatterplot overlaid with the regression line, and a paragraph that summarizes of the relationship between these two variables like question 2.7.**



The correlation value between two variables is 0.6622801 ~= 0.66. The relationship between read score and math score is moderate, which means that we can easily see that as math score increase, the read score also increase. But there is a lot of variance. By the summary of R, the adjusted r-square value is 0.4358. The intercept value is 14.07254 and the coefficient value of variable math is 0.72481. Hence the regression model equation is . The slope refers to how math score can determine the read score, and the intercept here means the average read score if students made a zero on math test. The p-value for intercept is 1.08e-05 and <2.2e-16 for math variable coefficient. Hence, both values are statistically significant. The diagnostics graphs could be seen in the following:



From the diagnostics, we could see that all three assumptions are not violated. The residuals vs Fitted graph has a horizontal fitted line with randomly scattered points. The normal Q-Q graph have all point lies along the linear line only except few extreme point. The scale-location plot also give as an approximately horizontal fitted line. Although the line is not perfectly straight, but the points spreads out pretty evenly. And no outliers in residuals vs Leverage graph.

Overall, this model is a good model, and the relationship between read score and math score is linear.