Siddharth Sharma

STAT 430

Project 2

1. Write up on the correlation and regression of variables of interest:
   1. The correlation coefficient is 0.46434. Since this value is greater than 0, it means that although the two variables seem to be uncorrelated (as depicted by the basic scatter plot, which shows no apparent regression line,) there is a positive correlation between them.
   2. ANOVA results from the PROC REG shows the p-value = 0.0097. This p value is statistically significant since it is less than the standard alpha value of 0.5. Further analysis shows that this p value is of the independent variable, which means that all of the predictors are significant and cannot be eliminated.
   3. Regression results and what they mean:
      1. The regression question is **tv = 1.74446 + (0.60818)\*movies**
      2. The p-value for the intercept is 0.0022 and for the independent variable (movies) is 0.0097
      3. Since both the p-values are less than 0.5 (alpha), we can conclude that the values are statistically significant. This means that the low p values are likely to be a meaningful addition to our model because changes in predictors’ values are related to changes in response variables.
      4. The residual analysis of the data shows that the points are spread not very close to the regression line. The residual values are between -2 and 2 mostly and only a few are close to 0, which means that the regression model is somewhat accurate, however it could be better.

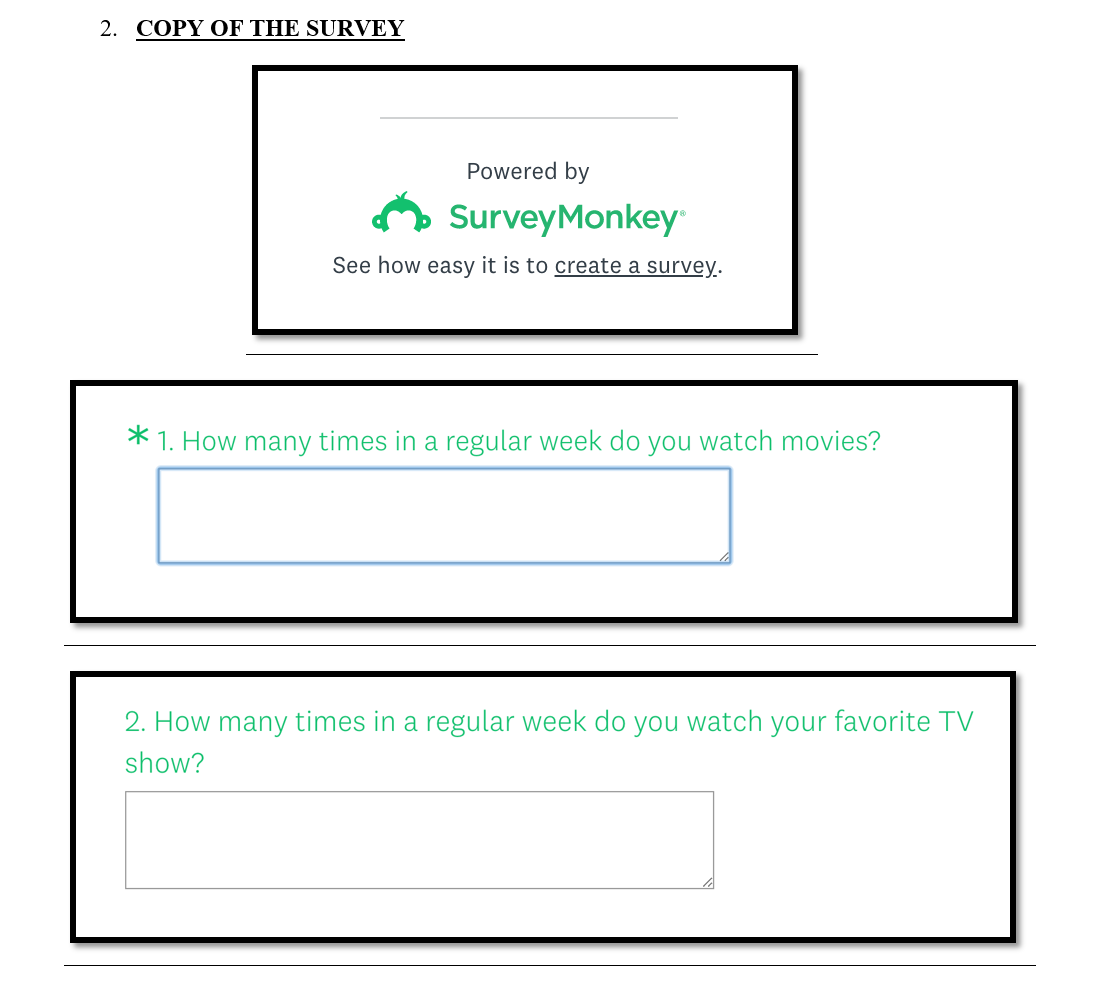
Since the data more or less fits the normal line, we can assume that the process data is normally distributed.

The residuals by regression for the dependent variable plot show the data is randomly distributed, meaning that the linear regression model is appropriate. Some values are very close to the horizontal line, meaning that the regression model is appropriate. The results for this are shown in the residual plots of the dependent variable.

1. The assumption for the regression model are met. The first assumption is that for this linear model, at least 20 independent samples are analyzed. Second assumption is that all variables are normal, which is shown in the reg procedure. Another assumption is that linear regression assumes that there is little or no multicollinearity in the data.

All these assumptions are necessary to check that the data can be analyzed using a linear regression model. For example, if we find from the normal probability plot that the data is not normal, we can conclude that we need a non-linear transformation for the data (log-transformation, for example.)

1. The sample data collected from the survey is not very good. The residuals are not all fairly close to 0 and the predicted values are not generally close to the regression line. By this we can conclude that even though the regression model describes the data, it is not very good. To make it better, more data could be collected to make the distribution more normal. This would also give us a sense of what values fit the regression line and what values are outliers.
2. The ultimate goal of this project is to find correlation between the two variables of interest (movies and tv). Data from Project 1 is used for this project.



Link to my survey: Here is a link to my survey: <https://www.surveymonkey.com/r/3T5W6K8>

1. ***SAS Code***

data survey;

input movies tv;

datalines;

2 5

2 1

1 0

0 3

0 0

4 4

1 1

1 3

2 0

1 4

1 2

2 3

2 2

3 2

3 1

3 5

0 3

4 7

2 4

1 3

0 2

1 3

2 2

5 7

2 3

5 4

3 5

0 5

1 1

3 2

;

/\*Statistical summary of the data using univariate procedure for detailed descriptions\*/

proc univariate data=survey plot;

var movies tv;

run;

/\*Basic scatter plot for initial impressions about the correlation of the two variables of interest\*/

proc sgplot data=survey;

scatter x = movies y = tv;

run;

/\*The correlation procedure to determine correlation between movies and tv\*/

proc corr data=survey;

var movies tv;

run;

/\*The reg procedure, Model Y(tv) = X(movies)\*/

proc reg data=survey;

model tv=movies;

plot tv\*movies

residual. \* tv;

run;

1. ***SAS Output (Including Graphs)***

