Rhea Qianqi Rao

Regression and Multivariate Data Analysis

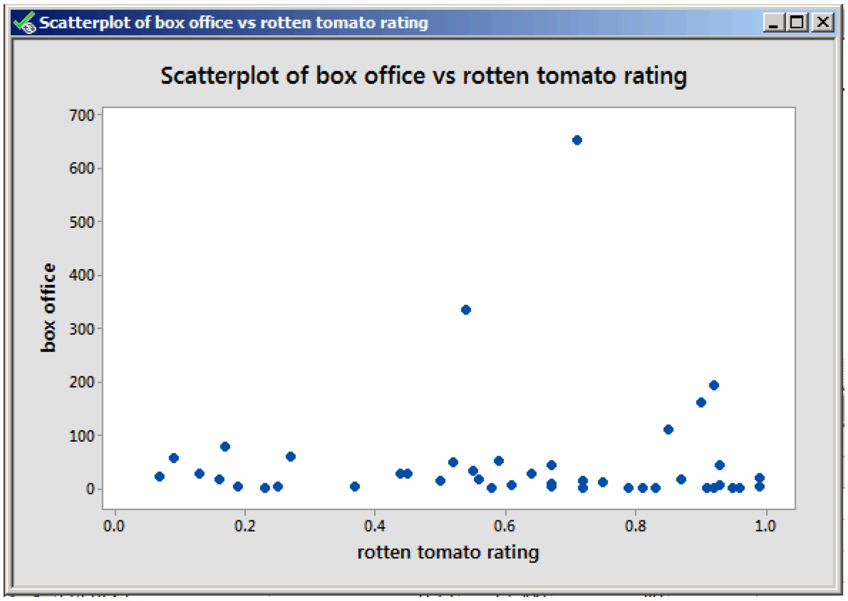
The Relationship Between The Rotten Tomato Score Of A Movie and Its Box Office Revenue

And Between Release Dates and Box Office Revenue

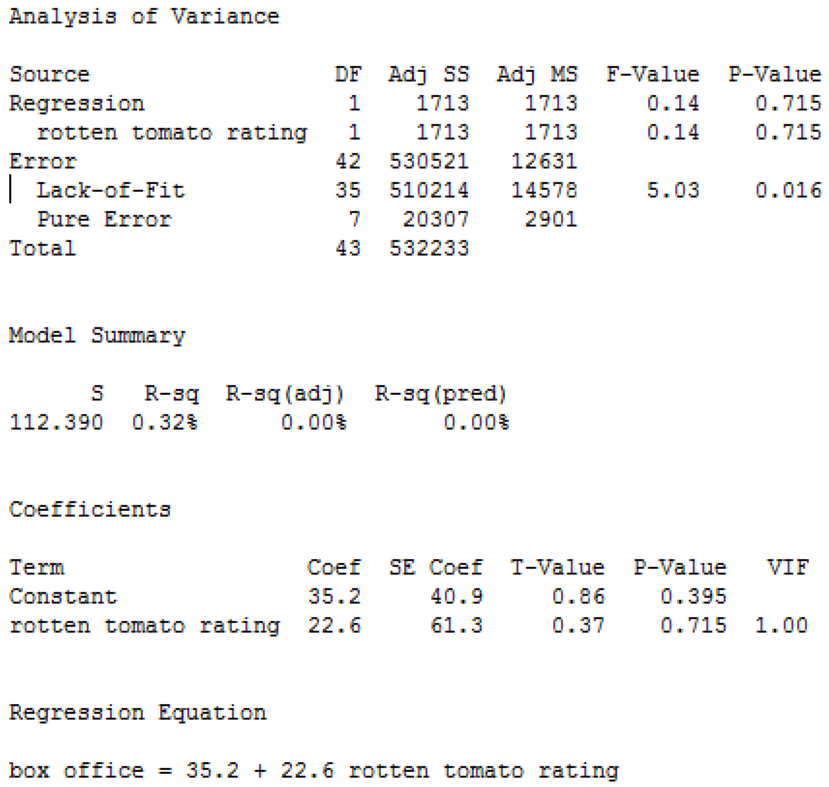
I'm interested in identifying the relationship between the rated score of a certain movie on Rotten Tomatoes, and the listed box office revenue of that movie. I obtained a sample of 45 movies from the website of Rotten Tomatoes (www.rottentomatoes.com). I focus on the recently released movies whose release dates range from July 2015 to September 2015. To eliminate as much noise in the data as I can, I random chose my data from a pool of English-speaking movies that are all released in United States. The box office data were all domestic revenue. While I think the place the movie is made (eg: in United States or International) might have an impact on my data as well, such information wasn't available on Rotten Tomatoes.

For this project, the predicting variable is the rated score of the movie on Rotten Tomatoes, while the response variable is the box office revenue of that certain movie. Before looking at the data, I expected to see a somewhat possible correlation between the two variables. Rotten Tomatoes is one of the most popular indicators of the likability of the movie among the audience, and a positive rating would likely attract more watcher into the cinema. On the Y-axis is the box office revenue in millions, while the X-axis is rotten tomato rating in decimal points (0.2 indicates a rating of 2 out of 10).

I first start by looking at the scatter plot of the variables:

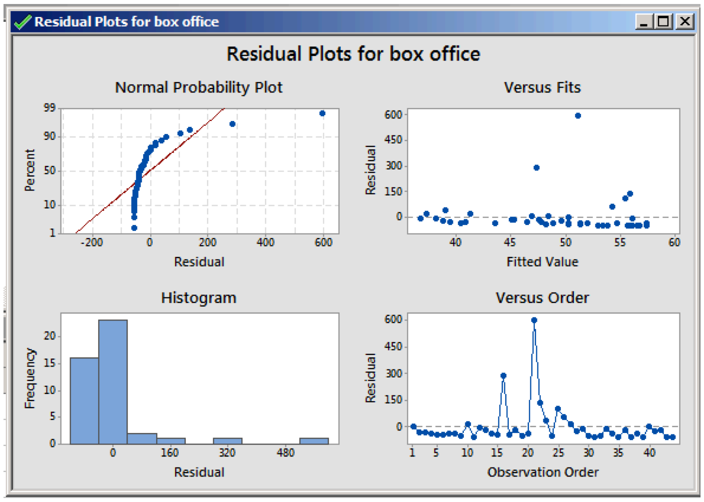


Judging from the scatterplot, there really doesn't seem to be any linear relationship between box office revenue and rotten tomato rating what so ever. The box office revenue data tend to stay below the 100 million. There are 2 very obvious outliers. And 3 abnormal observation points appearing between the 100 and 200 millions mark. A least square regression performed on the data is as below:



```
Analysis of Variance

Source               DF  Adj SS  Adj MS  F-Value  P-Value
Regression            1    1713    1713     0.14    0.715
  rotten tomato rating  1    1713    1713     0.14    0.715
Error                42  530521   12631
| Lack-of-Fit         35  510214   14578     5.03    0.016
  Pure Error          7   20307    2901
Total                43  532233


Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
112.390  0.32%      0.00%       0.00%


Coefficients

Term                  Coef  SE Coef  T-Value  P-Value   VIF
Constant              35.2     40.9     0.86    0.395
rotten tomato rating  22.6     61.3     0.37    0.715  1.00


Regression Equation

box office = 35.2 + 22.6 rotten tomato rating
```

The regression is very weak to say the least. R Square is as low as 0.32%, with R-sq(adj) even lower than R-sq and the difference is somewhat big compared to the number themselves. This not only indicates that there is no apparent linear relationship between rated score and box office revenue, but also that the sample size is too small to tell us anything significant about the data. The F-statistic, unsurprisingly, is very small. While the intercept here is positive, indicating a positive correlation between revenue and rated score, the number and the data itself is not at all meaningful. The standard error is even bigger than the coefficient itself. There is no point to further calculate the confidence interval or the prediction interval.



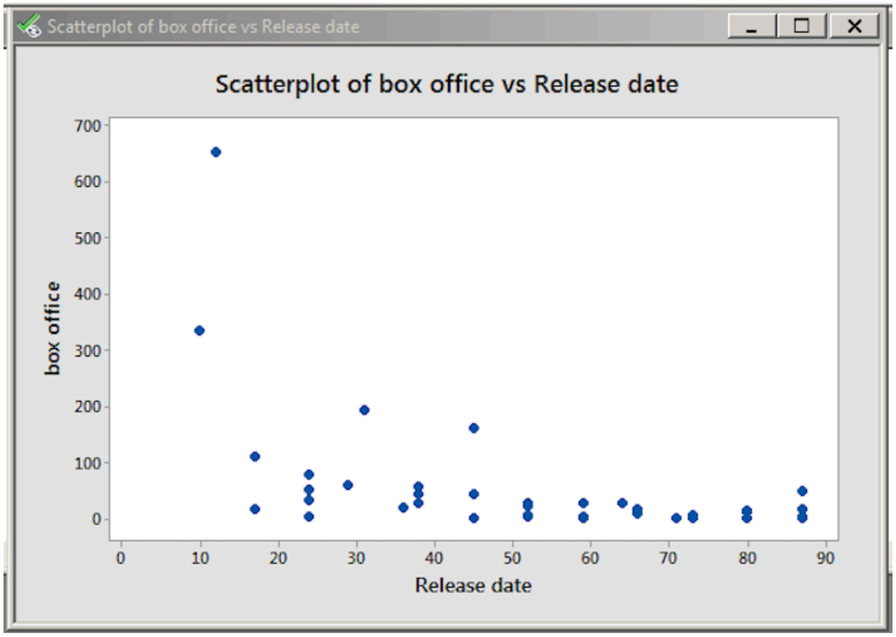
Looking again at these data, we see that they don't follow the normal distribution at all. However, we see two very obvious outliers that have extremely high revenue, with other observation points all under 100 million. The two observation points are Minions and Jurassic Park. Both of which are continuations of famous movie series that achieved great precedent success. So regardless of these two movies' rated scores, they have publicity and a large fan base, which subsequently contributed to the large box office sale.

The sample of data indicated a lack of linear relationship between the Rotten Tomato Score of a movie and its Box Office Revenue. I suspected that other factors might have affected the revenue. For example, among these 45 samples, some are released in July, some in August, while the others in September. The movies released in July is likely to have a much higher box office revenue than the ones released in August, let alone the ones in September.

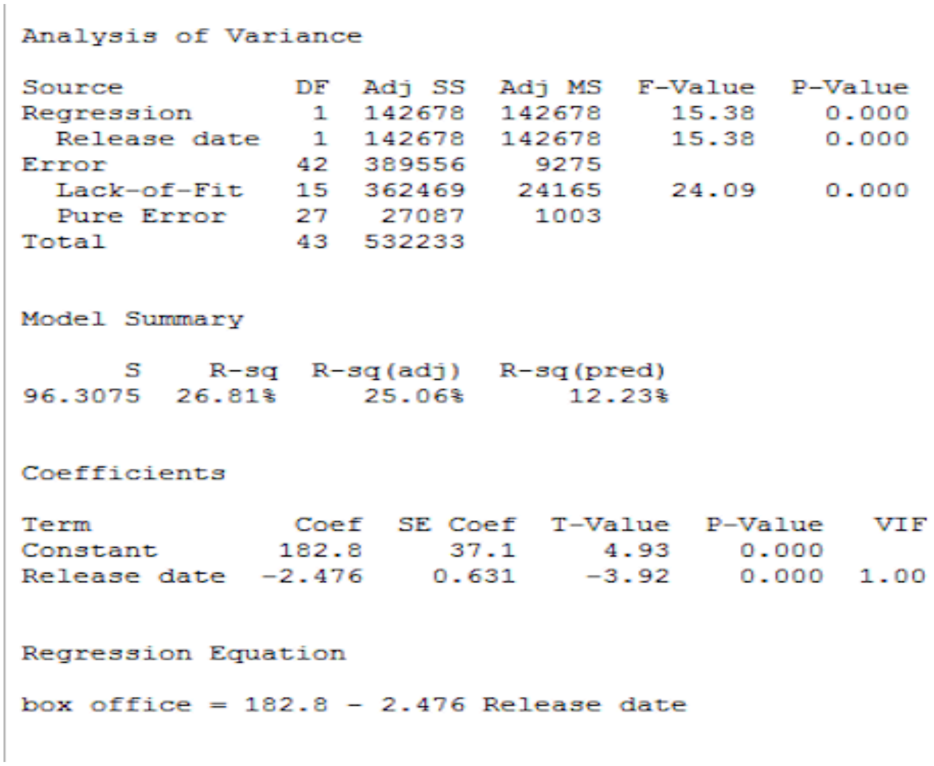
This fact violated the assumptions that $E(\varepsilon i)$ = 0 for all i. The movies released in July is likely to have y value systematically above the ones released in September. Also, the movie that has high revenue in July is likely to have high revenue in August as well, as it has created enough momentum and publicity that attract more audiences in the coming August and September. So it violated the assumption that εi and εj are not correlated with each other for i that is not equal to j. All the violations of assumptions meant that least squares regression doesn't apply to these two variables.

There are other factors other than Rotten Tomato Scores and release dates that might affect the box office revenue as well, such as the advertisement expenses, the country in which the movie was made. These all have violated the basic assumptions of least square regression.

Just out of curiosity, I ran a least square regressions on the box office revenue and the release dates of the movies. The response variable is box office revenue in the millions, while the predicting variable is the difference of days between the movie release date and July 1st. I expected to see a negative correlations between the two variables, the revenue should be getting smaller as the release move further away from July 1st. the scatter plot is as the following:



Looking at the graph, the two variables look like might be negatively correlated, but the points look rather scattered and we are still uncertain of the relationship between the two variables. So I ran least square regression on the data, the result comes back like this:



```
Analysis of Variance

Source            DF  Adj SS  Adj MS  F-Value  P-Value
Regression         1  142678  142678    15.38    0.000
  Release date     1  142678  142678    15.38    0.000
Error             42  389556    9275
  Lack-of-Fit     15  362469   24165    24.09    0.000
  Pure Error      27   27087    1003
Total             43  532233


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
96.3075  26.81%     25.06%      12.23%


Coefficients

Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant       182.8     37.1     4.93    0.000
Release date  -2.476    0.631    -3.92    0.000  1.00


Regression Equation

box office = 182.8 - 2.476 Release date
```

With an R-sq of 26.81% and R-sq(adj) of 25.06%, the regression isn't particularly strong. The F-value is larger than the one obtained for rotten tomatoes, yet still a weak one. The regression indicates that one unit change in release day, in other words, one day later for the movie to be released, is associated with an estimated expected change of -2.476 in unit for the box office revenue, which corresponds to 2.4

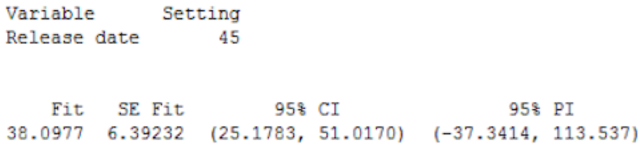
million decrease in box office sales. The standard error of the estimates of coefficient tells us that this model could be used to predict changes in release dates to within ± 1.26 %, 95% of the time.

The intercept of the slope is 182.8, this indicates that the estimated expected box office revenue of a movie is 182 million when it is released on July 1st.

To further test to validity of the model, we look at the t-statistic for null hypothesis of $\beta 1=0$:

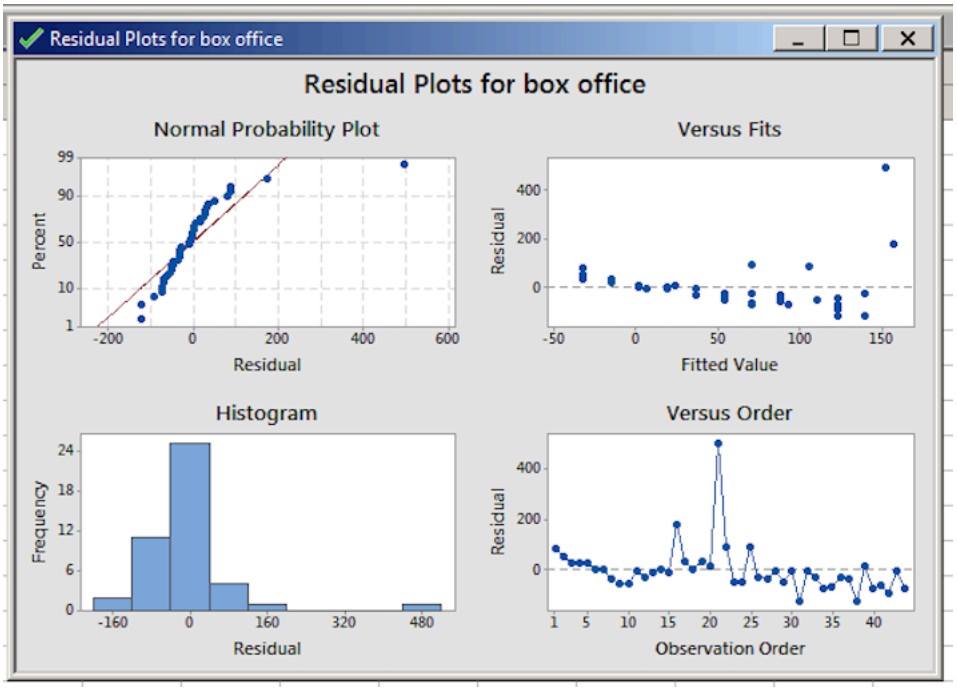
$t = \frac{-2.476-0}{0.631}$ = 3.924. So the tail probability for this test is approximately 0.005, and the data is statistically significant at the usual Type 1 error level, we reject the hypothesis. So there is a certain level of association between the two variables, although not a strong one.

Now that we have a model, we can use it to obtain certain data about the box office revenue when it is given the release date (In this situation, the release date has to be within 2 months after July 2015.)



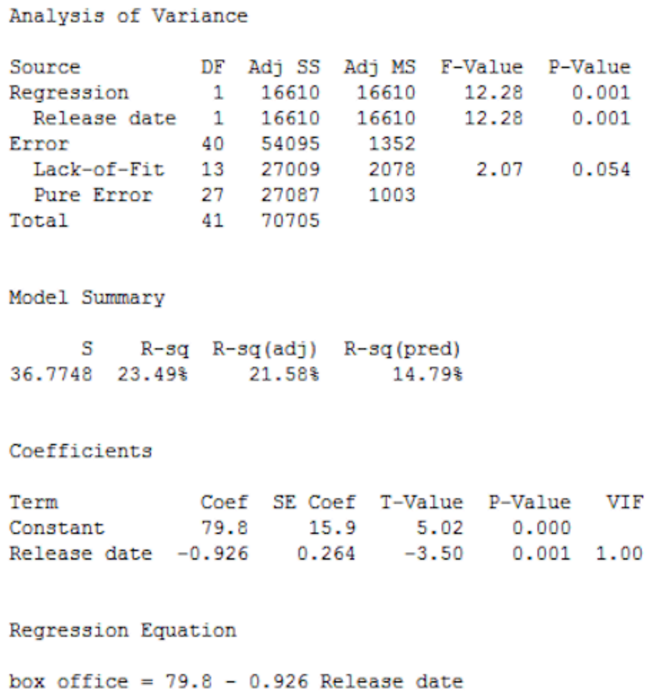
```
Variable      Setting
Release date       45


    Fit   SE Fit         95% CI              95% PI
38.0977  6.39232  (25.1783, 51.0170)  (-37.3414, 113.537)
```

The average of the differences between movie release dates and July 1st is 45. As shown above, for any movies estimated to release 45 days after July 4th, our guess for the average box office revenue change would be from 25.18 millions to 51.02 millions. Our guess for the revenue for a specific movie released 45 days after July 1st, namely August 15th, would be (-37.34, 113.54) . In real life, the revenue for a certain movie could be negative, so the movie revenue is estimated to be within 0 and 113.54 millions.



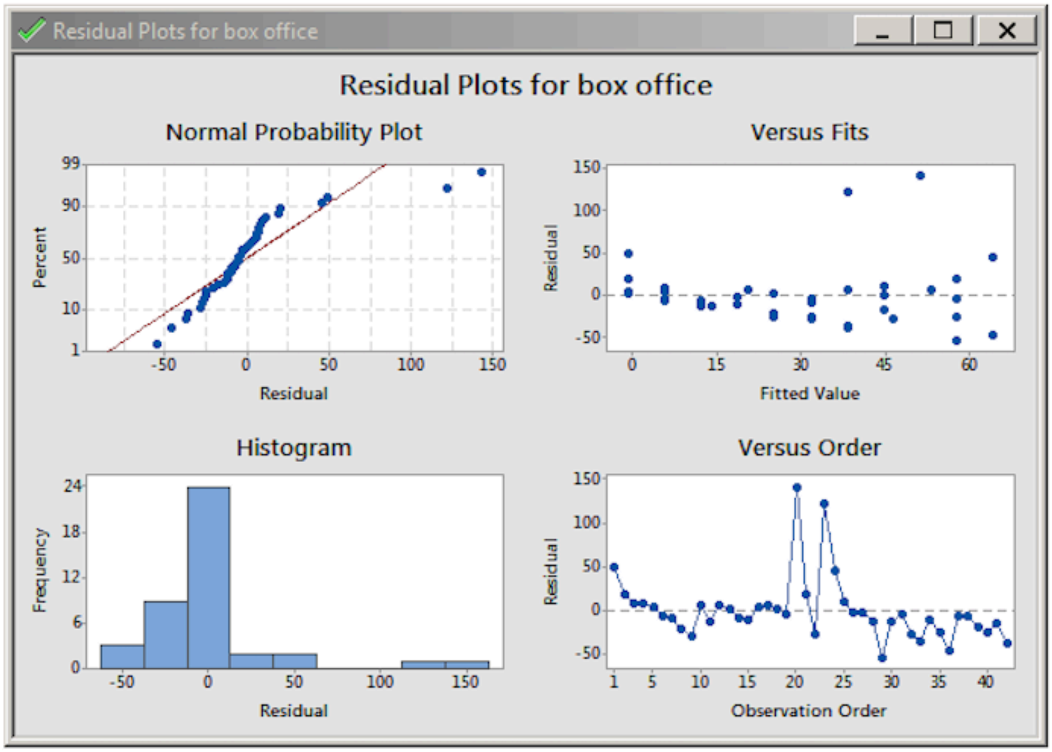
Looking at the graphs, the data is somewhat normally distributed. There are again two obviously unusual observations in these plots, including the movies Minions and Jurassic Park, which I have mentioned before. If we look at the outliers that stand on the far left at the normal probability plot, we could see that the points represent movie PHOENIX and MR. HOLMES. Both movies stand out in the sense that they have relatively low box office revenue despite its early release date. Surprisingly also, that Phoenix is a highly acclaimed movie with a rotten tomatoes score of 99%, the highest among all samples. Mr. Holmes also had a relatively high score of 87% I suspected that the abnormally low box office data occurred due to the genre and production of the movie. Both movies are made by small productions, not block bluster material.

After taking them out, we run regression with the rest of the data, the result is as follows:



```
Analysis of Variance

Source          DF  Adj SS  Adj MS  F-Value  P-Value
Regression       1   16610   16610    12.28    0.001
  Release date   1   16610   16610    12.28    0.001
Error           40   54095    1352
  Lack-of-Fit   13   27009    2078     2.07    0.054
  Pure Error    27   27087    1003
Total           41   70705


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
36.7748  23.49%     21.58%      14.79%


Coefficients

Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant        79.8     15.9     5.02    0.000
Release date  -0.926    0.264    -3.50    0.001  1.00


Regression Equation

box office = 79.8 - 0.926 Release date
```

The regression has gotten even weaker with R-sq equal to 23.49%, lower than the previous 25%. And calculating the t statistic, we realize that it has become smaller, indicating that the data is less significant than before we took out the outliers. It indicates that with outliers, namely the two movies with the highest revenues in the sample, the linear relationship is stronger. The reason behind this is probably that the two movies that naturally would have gotten highest revenues happen to be released earlier in time. In fact, they are both released in July. The rest of the data indicates a less strong relationship between release date and box office revenue.



As we remove the previous outliers, there are, not surprisingly, more outliers coming into the picture. The linear relationship between revenue and release date, nevertheless, is stronger than the linear relationship between revenue and rated score on rotten tomatoes. The relationship between revenue and release date is still not strong, possibly due to some of the reasons that could lead to violations of assumptions for linear regression as mentioned in one of the above paragraph.