Midterm Solutions

1. Compare PG-13 and R movies based on amount of money made. Do R rated movies make less?

This problem was graded on four components, and each component was graded on a 0-3 point scale. Thus there is a total of 12 points.

1) Mechanics: did you make any mistakes using R? Any computational errors?
2) Summary Stats: Did you use appropriate summary statistics and provide a coherent description of the data?
3) Assumptions and checks: did you state the appropriate assumptions needed? did you check to see whether they were true? Did you evaluate what the effect would be on your conclusion if the assumptions turn out to be wrong?
4) Inference Procedure: did you choose appropriate statistical tests? Did you carry them out correctly? Did you interpret the results?

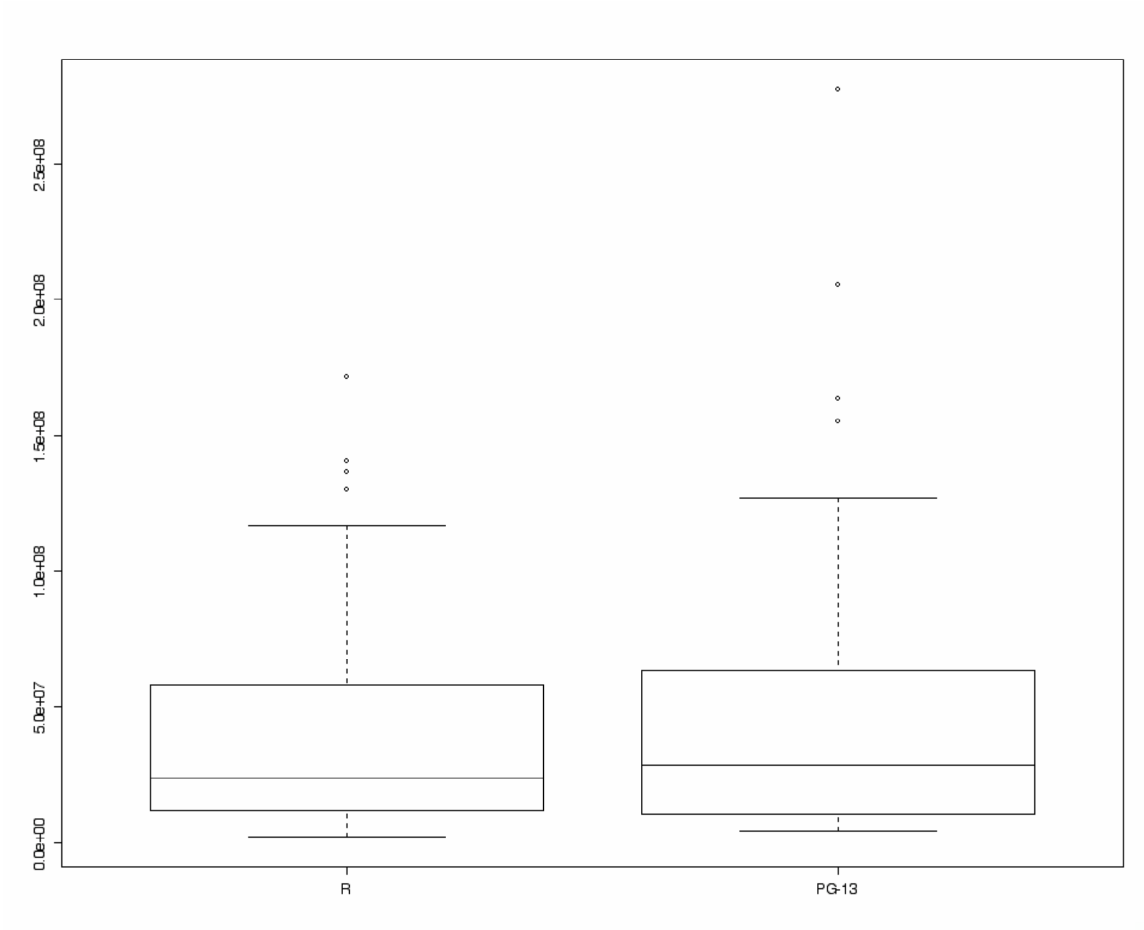
First you need to construct appropriate variables.

```
> r.rated <- UltimateGross[RRating==1]
> pg.rated <- UltimateGross[PG13Rating==1]
```

These two commands create new variables that have the ultimate gross for r-rated and pg-13 rated movies.

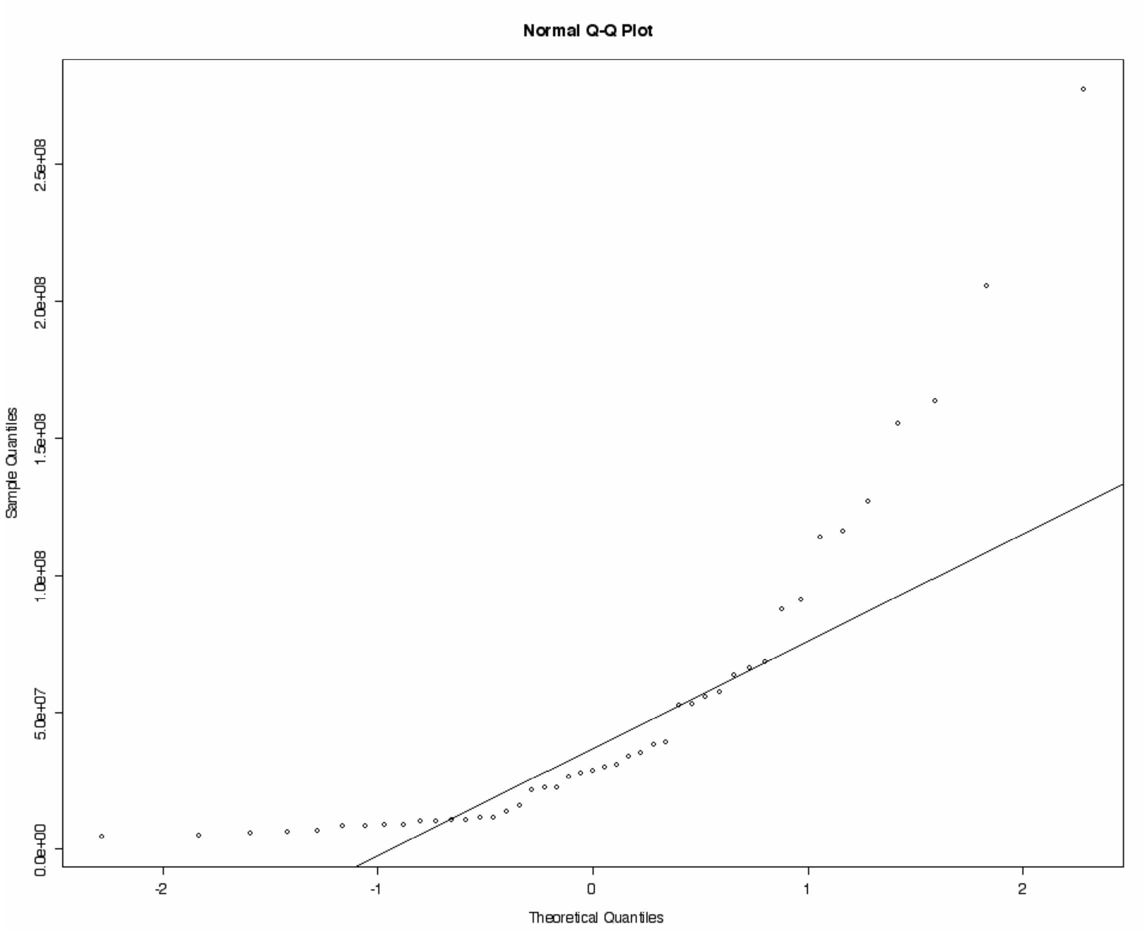
A summary command shows that the R-rated movies do seem to make quite a bit less on average (39 million as compared to 50 million). However, their medians are not all that different (24 as compared to 29), which suggests highly skewed distributions.

```
> boxplot(r.rated, pg.rated, names=c("R", "PG-13"))
```



Which shows that the PG-13 movies do seem to generally make more, but there also is quite a bit more skew.

One test to compare is the t-test, and this assumes each variable is normally distributed. A qq-norm plot (a pair, actually -- one for each variable) will tell us whether this is true. Neither of them is. Shown below is the qq norm plot for the pg-13 movies, and the r-rated movies look pretty similar:



Still, the t-test is supposed to be robust to violations of normality, and so it is probably okay to use it, given that the sample sizes are pretty large (45 for PG-13, 68 for R).

Our null hypothesis is that the mean gross of R-rated movies is the same as for PG-13. The alternative is that it is less, which means that the (Mean R-rated) minus (Mean PG-13) should be negative.

```
> t.test(r.rated, pg.rated, alternative="less")

     Welch Two Sample t-test

data:  r.rated and pg.rated
t = -1.1174, df = 69.841, p-value = 0.1338
alternative hypothesis: true difference in means is less
than 0
95 percent confidence interval:
    -Inf 5479726
sample estimates:
```

```
mean of x mean of y
 39251823  50392635
```

This test fails to reject the difference and we conclude there is no statistically significant difference between the mean gross of R movies and the mean gross of PG.

Note that this test does NOT assume that the variances of the two groups are equal. This assumption is rarely justified in practice, and should be avoided. Second, this is a one-sided test, since the alternative hypothesis is that the mean of the R movies is less than the PG-13 movies.

Discussion

There are at least two things that make us a little uneasy about this test. First, even though the t-test is robust if the variables are not from a normally distributed population, a sample size of 45 isn't all that large. Second, the t-test assumes that we have a random sample, and these clearly are not randomly sampled. They're not even a sample -- they're the entire population of r-rated and pg-13 rated movies in 1999.

One way around the first problem (but not the second problem) is to take log transforms of both variables, which does a surprisingly good job of turning them into normal distributions. (Not perfect, but pretty good). Performing the t-test on the log of these variables does not change our conclusions, which is nice.

The second problem is more problematic. We can make-up a new population: the population of all future R-rated movies and all future PG-rated movies. This makes sense because our purpose is to decide whether a *future* movie should be rated R or rated PG-13. However, this forces us to "pretend" that the movies were randomly selected from the future-- which is clearly impossible. Or we can "pretend" they represent a random sample from 1999 and that 1999 will look just like the future. Neither approach is terribly satisfactory.

One approach that solves both the problem of normality and the lack of a random sample is the permutation test. This was above and beyond the call of duty, but would be appropriate here.

PROBLEM 2

Each part of the problem was graded on a scale of 0-3, and so there were 24 points possible.

First some common problems/misconceptions.

1) To answer the question: "what is the distribution of variable y", many people did a plot of the variable on the y axis and the case-index on the x axis. You get this plot by typing

plot(y). While this plot is one way of using the "identify()" function to identify individual points, it is NOT a graph of the distribution. The distribution of a variable is a function that gives the frequency or density of values of a variable. There are three commonly used methods for picturing distributions: histograms, boxplots, and dotplots. Stemplots are another, slightly less common, method. Histograms are probably your best bet. Boxplots give you a less detailed picture, and tend to over-react to potential outliers (which is sometimes a good thing, sometimes not.)

2) While we sometimes put several plots on one page in order to save space (and sometimes to facilitate comparisons), you should be aware that your perception of any trends in a scatterplot will be affected by the shape of the plot. If the plot you make is too small, or is distorted to fit onto the page, then you might see things or not see things that you should or shouldn't see, if you get my point. When examining a plot, look at the plot by itself. Try resizing and changing the dimensions. (Although the default dimension is usually good enough.) You are doing detective work, looking for any indications of a failure of the model.

3) The r-squared does not evaluate the assumptions of the model. R-squared measures how tightly points are clustered about the line (or in multiple regression, about the plane), *assuming the model is correct*. It's possible to get a high r-squared with a horribly fitting model. R-squared is useful only if you're satisfied the model is a good one. So don't look at the R-squared to decide if your transformations are good or not. Look at residual plots and scatterplots and qqnorm plots to determine which transformation is best. Once you are assured the model is linear, then you can trust the r-squared statistic.

4) All assumptions are not created equal. There are three biggies: linear association, constant variance, and normally distributed errors. Of these, the normally distributed errors are probably the least important. The reason for this is the central limit theorem, which tells us that if the sample size is large enough, linear combinations of the data (such as the estimates of the slope and intercept) are approximately normally distributed. (some peopole mistakenly wrote that the central limit theorem says that the population is normally distributed if the sample size is large enough. The population either is or is not normally distributed, and no amount of data will change this fact. In fact, the more data you collect, the more the distribution of the data looks like the distribution of the population. So if your population's distribution is right-skewed, your sample will be right-skewed (when the sample size is large enough). One implication of this is that when you're trying out different transformations, ones that achieve linearity and/or constant variance might be more valuable than ones that only achieve normality. (but all of this just "depends" ...on the sample size, on how non-normal the residuals seem to be distributed, and on what you intend to do with the model.

5) Careful when interpreting slope. Many of you said something like "when the Friday gross increases by one dollar, the gross increases by blah dollars." This is wrong for two reasons: (a) in this data set, it is impossible for movies to change their Friday gross. So the "if" statement you are making is an impossible event, and so this interpretation doesn't mean anything. What it does mean is that movies that make 1$ more than other

movies make, on average, blah dollars more. This brings us to (b). The regression line tells us about mean values of the response value, not the response itself. So we are not saying that movies that differ by 1 dollar on Friday gross differ by exactly so many dollars on their total gross. We're saying that their *mean* values differ by this much.

6) The question " is the model useful for prediction" could be answered in different ways. First, you should check whether the slope is statistically significant (if this is a meaningful thing to look for). If it is not, this means that the model will not be useful because there is no association between the variables. Then, if the slope is significant, you might wonder whether or not your predictions will be good or bad. The r-squared helps you out here. Big r-squared means your observations will be closer to your predictions. But how close is close enough is a matter of opinion and will vary from person to person.

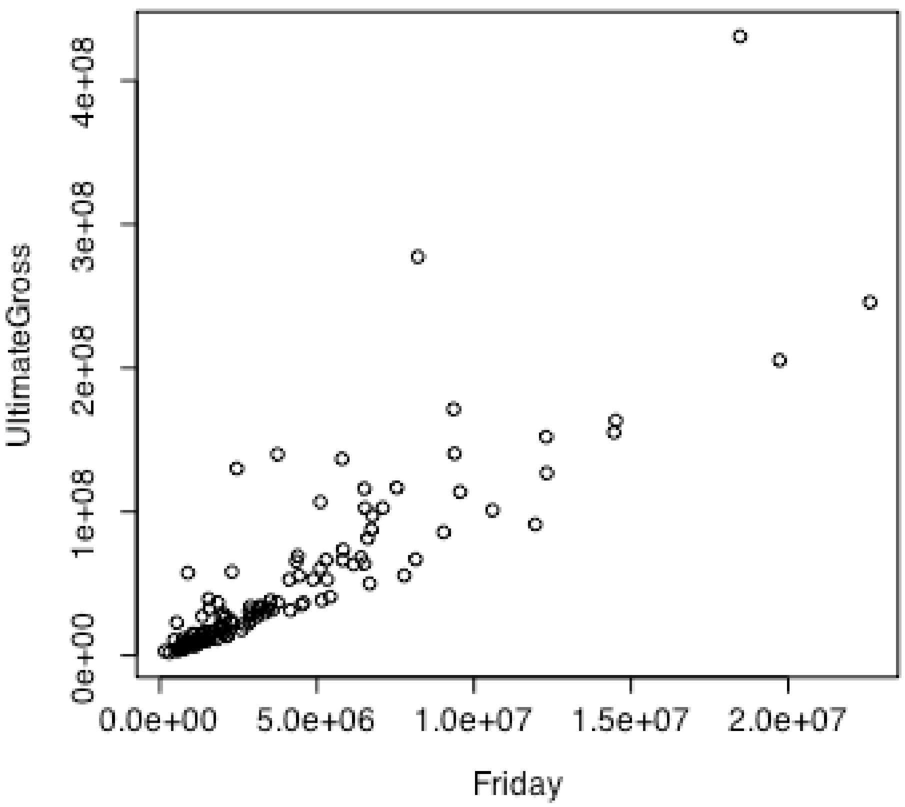
The variable "Friday" contains the amount of money the film made on the first Friday that it opened. (Most movies opened on Friday, and so this would be the very first day of release. But some opened on the Wednesday before.) Our ultimate goal will be to determine how well the first Friday predicts the ultimate gross.

1) To start, describe the distribution of the first Friday box-office. Identify (by name of film) any outliers. Also describe the distribution of UltimateGross and identify (by name of film) any outliers. The names are in the variable picture.name.

The distribution of both movies is right skewed with a few movies making considerably more than the rest. This histogram of the Friday returns doesn't indicate any outliers, although a boxplot does show a few potential outliers: The Mummy, Star Wars Episode 1, Austin Power 2, Toy Story 2. The histogram of the Ultimate Gross does show an outlier: Star Wars. A boxplot suggests even more: The six Sense, Austin Powers 2, The Matrix. It's mildly interesting that two of these movies, the Six Sense and the Matrix were not outliers for the Friday gross. This suggests that perhaps some movies did not do so well on opening weekend, but then surged ahead, perhaps based on word of mouth. (Sometimes these are called "sleepers".) This might be particularly true of the Six Sense, which is the second largest grossing film of the year.

Six Sense didn't do horribly it's first Friday. It earned $8215195, which means it was 1.07 standard deviations above average for that year. You should know that from the regression to the mean phenomenon, we might expect it to be closer to average in it's ultimate gross, and the fact that this is not the case indicates in some ways that this is a remarkable outcome.

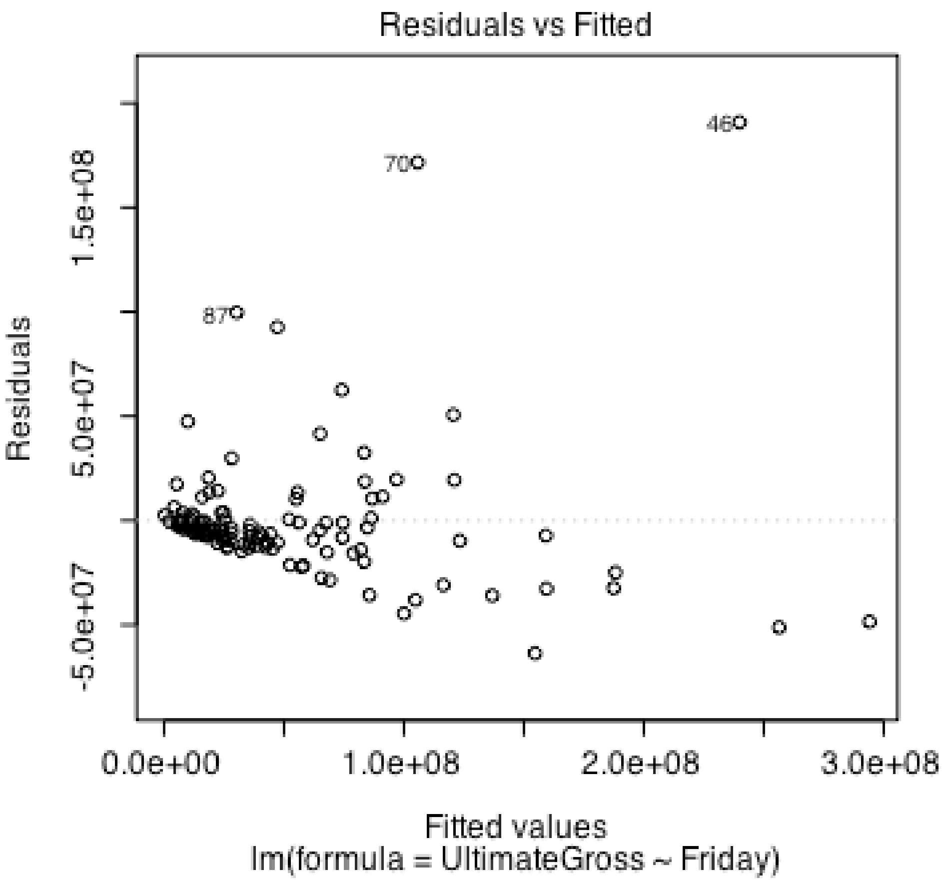
2) Make a scatterplot of UltimateGross against Friday. Describe the relationship between Friday profits and ultimate gross.



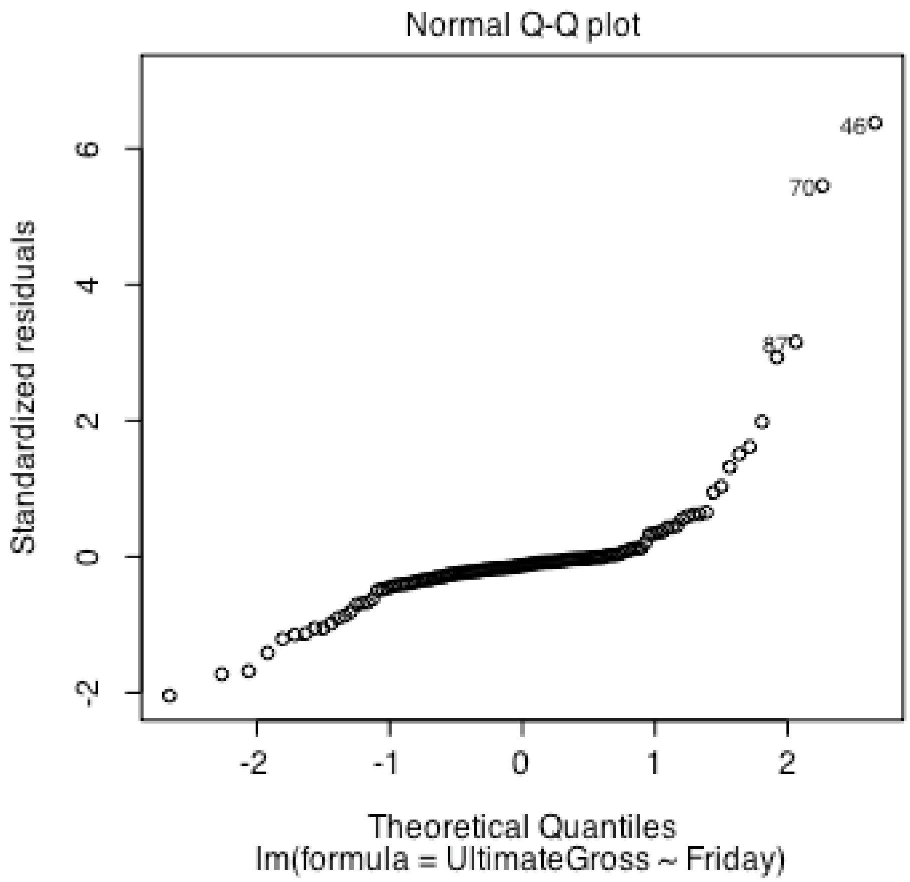
We saw a positive, nearly linear (maybe even linear -- it's hard to tell) relation between the amount of money earned by a film on the first Friday and the amount earned ultimately. This means that movies that more than average the first weekend, tend to make more in the end, too. Not terribly surprising.

3) We'll do something naive. Fit a linear model using Friday profits to predict the ultimate gross. Evaluate the fit. Does the assumption of a linear relationship between the mean ultimate gross and the Friday profits seem justified? Are there influential points? If so, what are they? Do the residuals seem to be normally distributed?

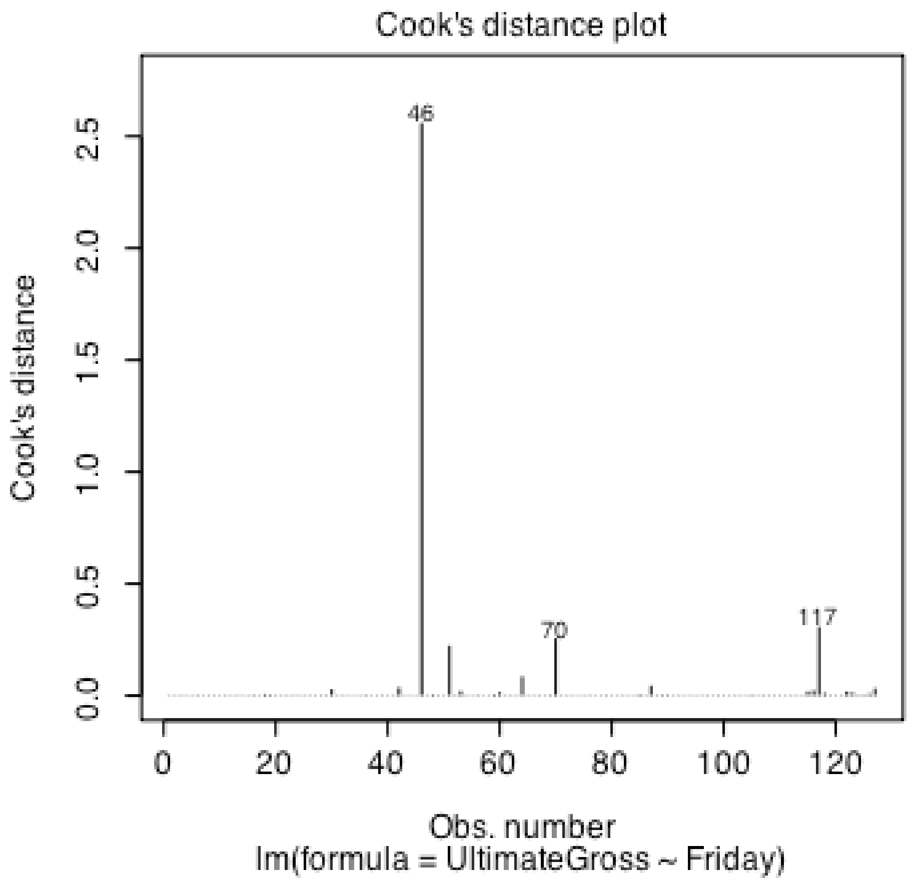
The fit isn't terrible, but nor is it great. A plot of the residuals against the fitted values shows a very prominent "fan" shape: For movies for which we're predicting a large gross, we are seeing lots of variability. For movies for which we predict smaller gross, we see little variability. This is a sign of non-constant variance, and one side-effect is that our p-values will be wrong and our confidence intervals wrong.



The qqnorm plot shows that the residuals are not normal. This is a minor sin, but a sin none-the-less. It is minor, because regression is moderately robust from departures from normality. But still, if we can improve it, we will have more accurate p-values and confidence intervals.



We identified outliers in the first part because outliers can have an influence on our regression. The Cook's distance plot shows that observation 46 has a large influence.



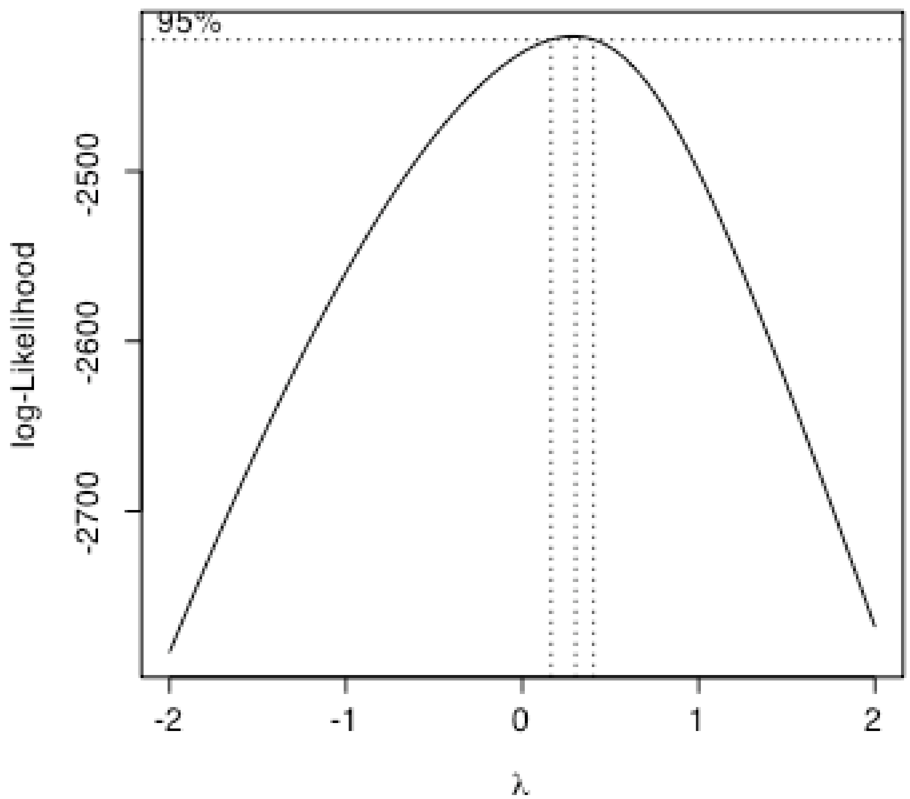
This is Star Wars.
> picture_name[46]
[1] STAR_WARS_EP1_PHANTOM_MENACE

Not so surprising, since this was such a large outlier.

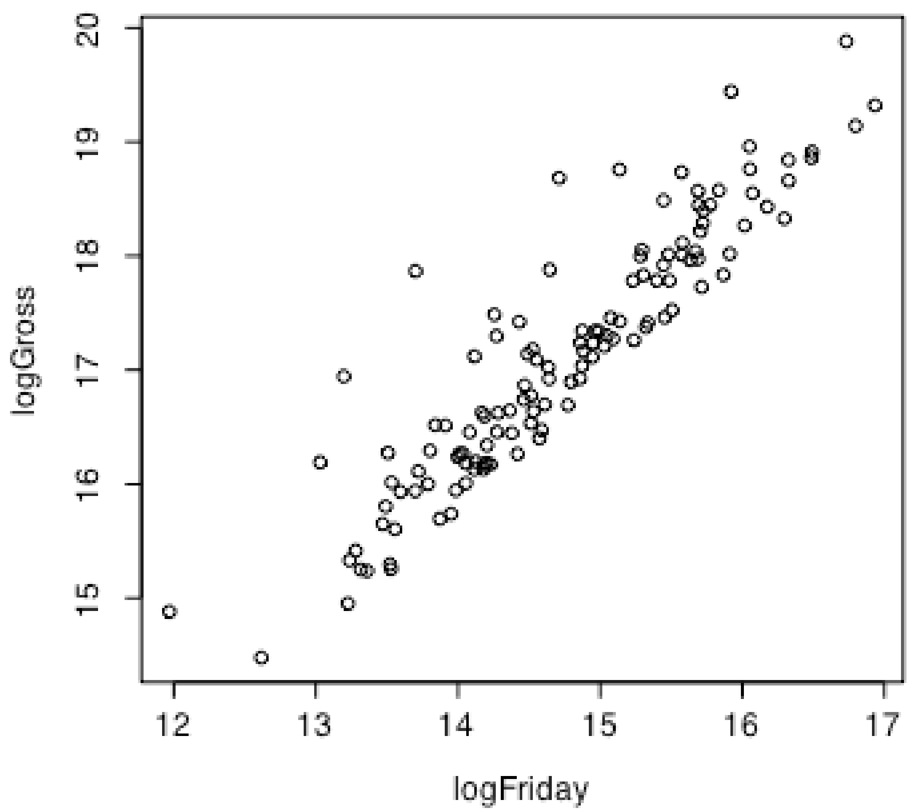
4)Find transforms of either variable, or both variables, that best satisfy the assumptions of the linear model. Justify your choice with the appropriate plots and/or summary statistics. (Warning: it's pretty hard, and may not be possible, to find transforms that satisfy all assumptions.)

One approach to this is to try various transformations until you get more normally distributed histograms. The boxcox function can help by suggesting some transformations to start.

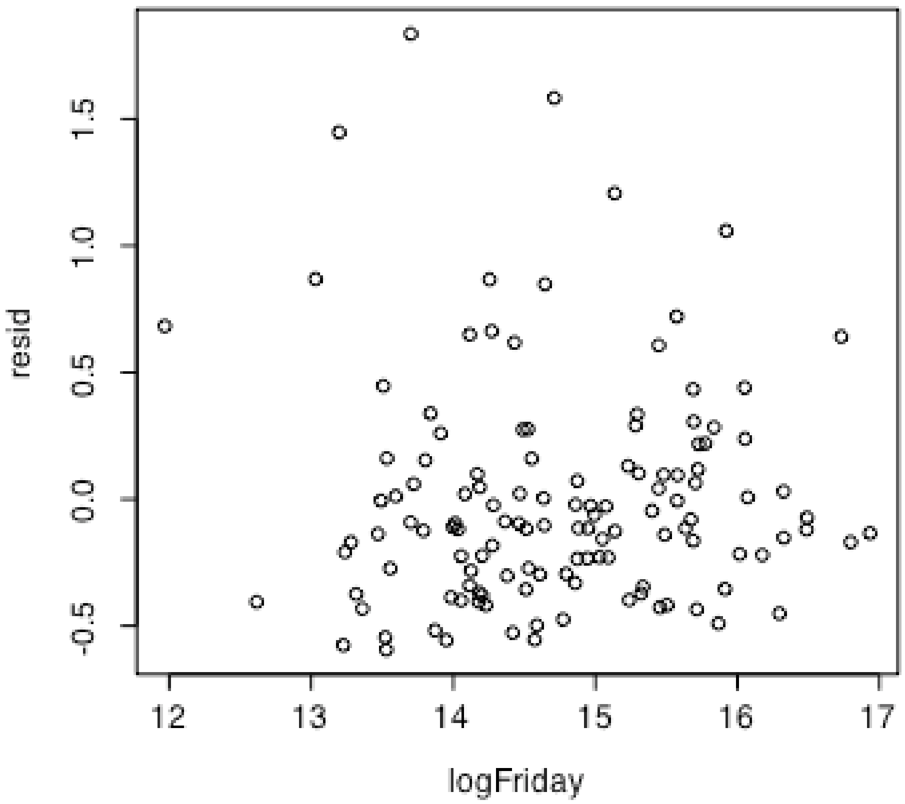
The result of a call to boxcox(UltimateGross~Friday) is this graph



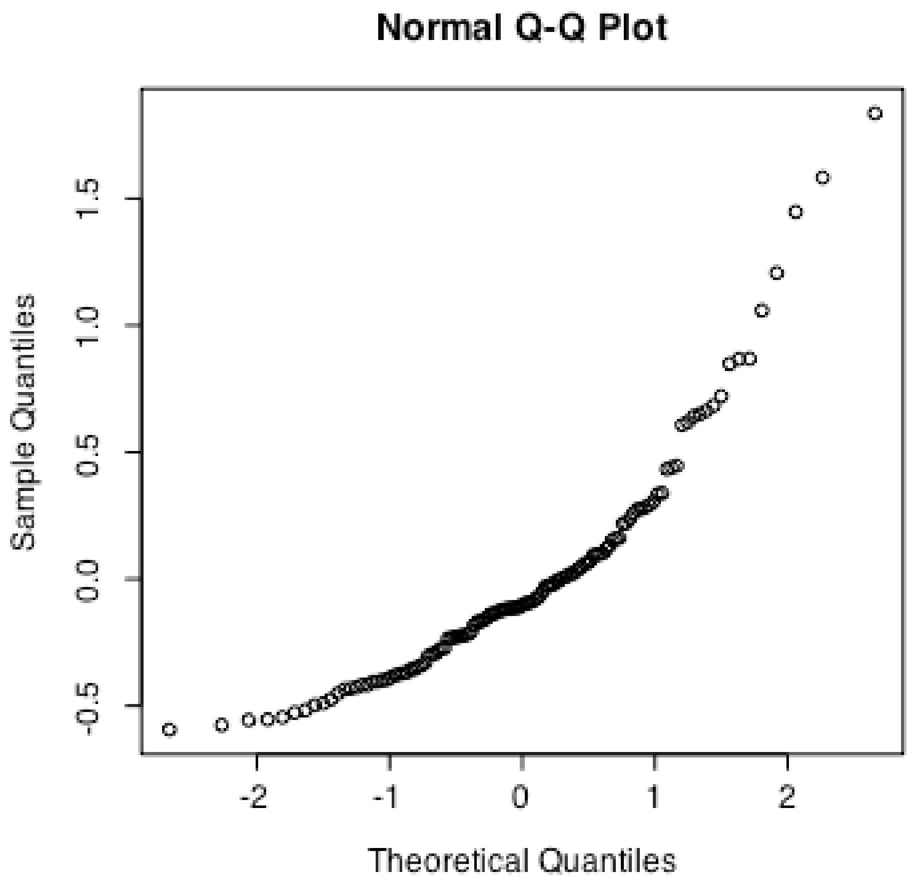
Which tells us that a good transform is to raise the x variable to maybe the 1/2 power. In fact, you can do even better by taking the log. And if you do this, and look at the scatterplot, and try a new linear model and examine the residuals, you'll see that the best fit is to look at logs. (This isn't too surprising: monetary variables are often improved by log transforms. Another rule of thumb is that if the maximum value is about ten times the min (or more), then a log transform might be useful.



This isn't perfect, but the non-constant variance problem has been (almost) eliminated, and while the residuals are still not normal, it appears that they are closer to being normally distributed.



Note the "fan" shape is gone. As before, the lack of a trend shows that the assumption of linearity is justified, or at least can't be easily dismissed.



Fit a linear model to your transformed data. According to the model, does knowledge of Friday's receipts help predict the ultimate gross? Explain. Interpret the slope and intercept.

The linear model is
logGross = 1.525 + 1.06*log(Friday)

The r-squared is .84, so we've explained 84% of the variation in the ultimate gross. The intercept and slopes are both statistically significant. The fact that the slope is not zero means that there is a relationship between the Friday returns and the gross, and the fairly high r-squared tells us that we should be able to make relatively precise predictions. Although whether they are precise enough to be useful for business decisions is something that can be answered only by those who have to make the decisions.

The slope here tells us that movies for which the log of Friday returns is one log-dollar higher make 1.06 log-dollars more, on average. But what is a log-dollar? Easier to "unwrap" this equation by taking exponents of both sides:

Ultimate Gross = exp(1.525)*Friday^1.06

Ultimate Gross = 4.572 (Friday)^1.06

One way of interpreting this is to look at ratios. For example, suppose we want to compare movies that made x amount on Friday with movies that made 10x. The Model says that the Ultimate Gross of the 10x movies compares to the x movies as follows: (UltimateGross(x))/UltimateGross(10x) = 10^(1.06) = 11.48.

This means that movies that tend to make ten times as much money on Friday tend to make about 11 times as much money in the ultimate gross, on average.

Predict how much a movie will ultimately make if the Friday receipts are \$1,200,000. Include a 95\% confidence interval.

The question calls for a prediction for an individual movie, and so we need a prediction interval.

Predict how much a movie will ultimately make if the Friday receipts are \$1,200,000. Include a 95\% confidence interval.

```
> newdata <- data.frame(logFriday=log(1200000))
> predict(logfit,newdata,interval="prediction")
       fit      lwr      upr
[1,] 16.34556 15.46096 17.23016
```

But we need to translate these into dollars:

We predict it will make $12554127, but we are 95% confident the amount will be between $5,183,339 and $30,406,288. Note that this is a very large range.

On average, how much money will movies that made \$1200000 on the first Friday make? Provide a 95\% confidence interval.

This calls for a confidence interval. We still predict that the mean gross of all movies that make $1.2 million on the opening Friday will make 12.5 million overall. but now the confidence interval says that we are 95% confident this mean amount will be tween $11.4 million and 13.9 million.

Evaluate this model. Do you think it would be useful for predicting gross income? What are the weak points?

One weak point is that we don't have normally distributed residuals, and so our confidence/prediction intervals, as well as our p-values are approximations and it is difficult to tell just how good these approximations are.

We also know there are three very influential points: Being John Malkovich, Cider House Rules, and Cradle Will Rock. We might try refitting the model without these three movies to see if we get radically different results. This doesn't mean we would trust the model any more once the points were removed---it would simply be a check on how sensitive our predictions were to certain points that we fit.

The prediction intervals are rather large, and possibly too large to be useful. This suggests that despite the large R-squared, we might want to look for a better method for making predictions. We might try to include more variables, since it seems reasonable that other variables (such as the time of year of the release) could be helpful and could explain more variation.