Guinea pig 1

Introduction:

The movie industry is huge, in terms of economics. The potential for the amount of money to be earned can reach billion figures. For example, Avatar is the highest grossing film of all time, with a profit of $2.787, 965,087 (URL 3). Yet at the same time, the amount of money invested into it was said to be nearly $300,000,000. Not all the time can a movie guarantee the same success. Not only is the film business a lucrative trade, but it is also extremely risky. Is there any way for Hollywood producers to predict how much money they could make from a film? Is there perhaps a point when the budget is high enough to guarantee some significant profit is made? This is what I want to investigate. For this investigation, the budget of the movie is how much was spent making the film (film production), while the gross profit is how much the movie made in US theatres.

My question for this investigation is: I wonder if there is a relationship between the Budget of a film (amount in millions of US dollars spent in making the film), and the Gross Profit (amount of money in millions of US dollars that the film made in the US) that a film makes.

While the major focus of the film industry is to make money, there are still many directors who aim for artistic integrity, and make sure their movies are of a high quality. However, looking at the list of top grossing movies, compared to the top rated movies, there is quite a difference. As I mentioned. Avatar is the highest grossing film of all time (3). However, on imdb.com, the highest rated film of all time is Shawshank Redemption (URl 4), which only had a gross profit of $16 million, a meagre amount compared to Avatar's gross profit of $2,1816 million. This raises the question. Does quality in a film truly matter anymore? Can a film be of low quality, and still receive a high gross profit? So I would also like to investigate another matter. A rating will reflect the quality of the film, as a movie with a high rating was obviously considered an excellent film by the rest of the public.

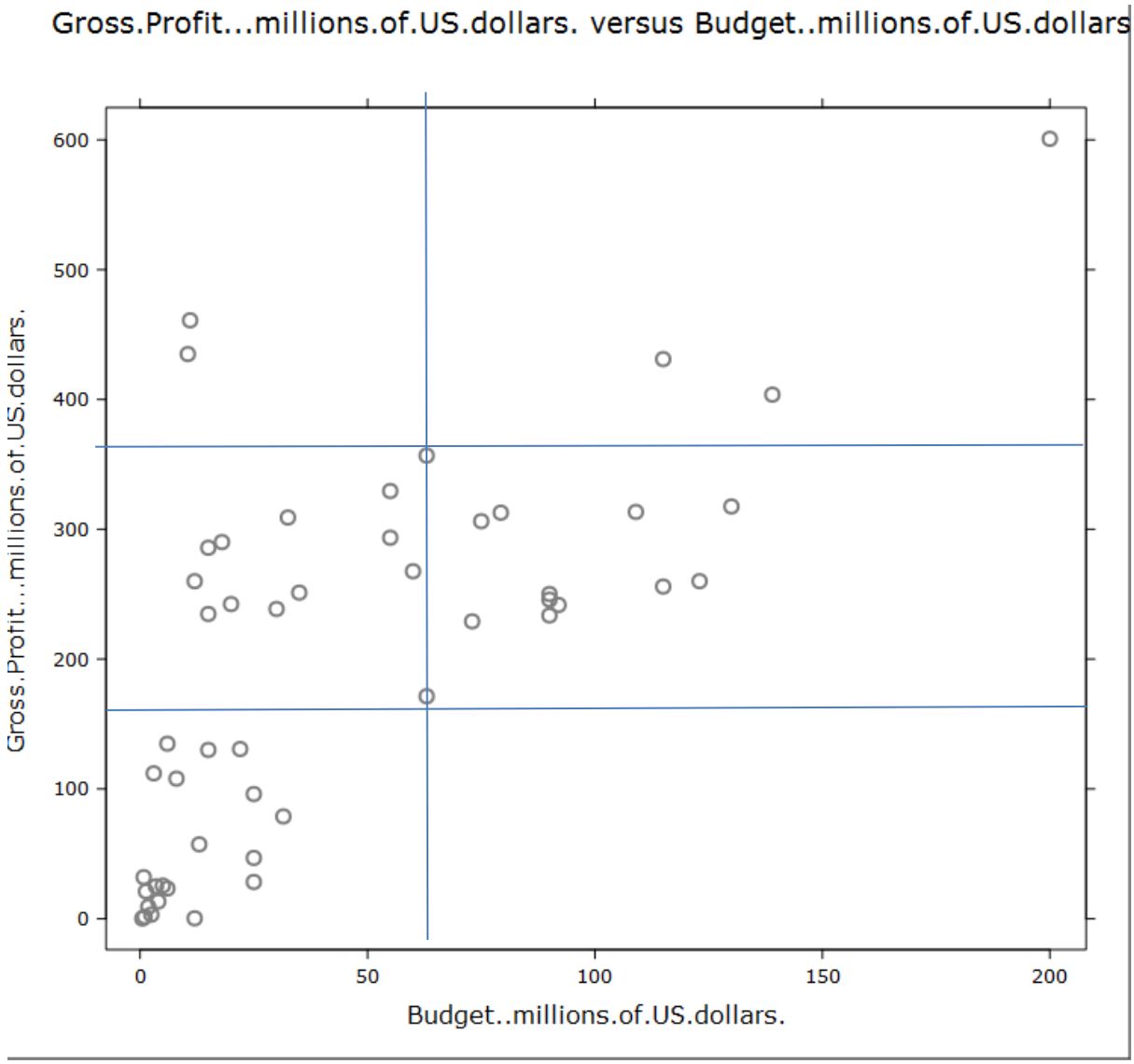
Another equation for this investigation is: I wonder if there is a relationship between the Rating (opinion of registered users of IMDB, which are then used to calculate an average score), and the gross profit that a film makes.

Finally, I will compare the two models to see which variable, budget or ratings, is a better model to predict the gross profit of a film. This will hopefully reflect whether it is the budget of the movie or the quality of a movie that really counts.

The data used is part of the data set collected by Mohammad Raza of Wolfson College, Cambridge, for his May 2003 Mathematical Tripos Part III essay 'Analysis of a large and complex data set'. It contains 50 famous and recent movies.

For this investigation, I chose the x variable, also known as the explanatory variable to be the budget, or the rating respectively. The response variable, or the y variable, is the Gross Profit. This is because the film makers invest money, or invest the budget to make a movie which produces a profit. In other words, the budget of a film leads to the profit. Similarly, the movie makers will make their film of a certain quality to receive a certain profit.

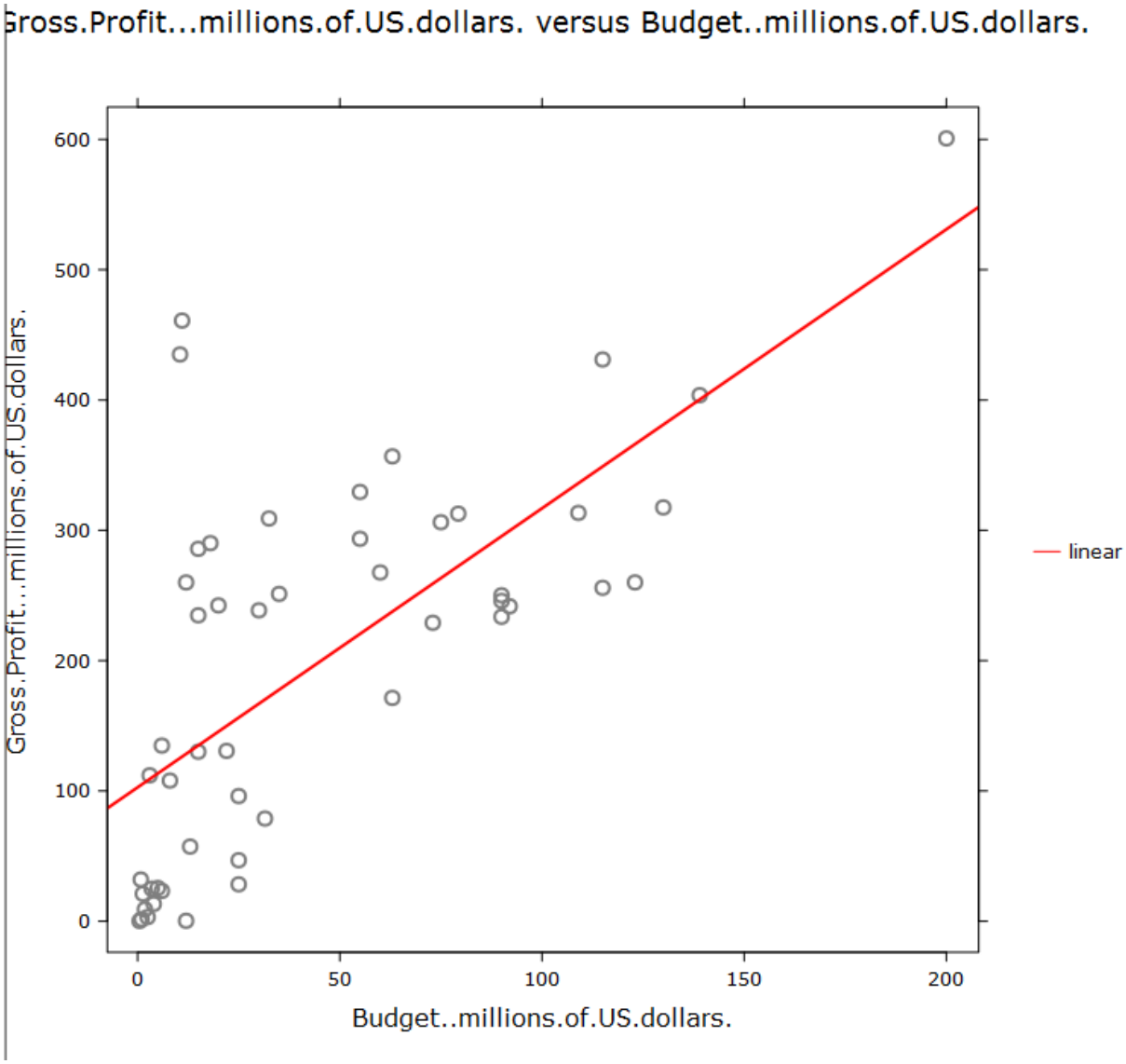
The units for this investigation is $US millions. So any number written is in millions.



Looking at the graph, there appears to be a weak positive association between the Budget ($US millions) and the Gross profit ($US millions) of a movie. This indicates that the more money that is put into making a movie (the higher the budget), the more money that movie will make (a higher gross profit). However, the relationship between the two appear to be very weak, as there is a lot of scatter in the graph. This means that for the budget of a certain movie, there is a large range of possible gross profits that movie could make. For example, if a movie had a budget of US $65, then the values of the gross profit could be between $160 and $370. This is quite a significant amount of scatter. Looking at the values, there do seem to be a few unusual values. For example there is one film (Titanic) that had a budget of $200, and had a gross profit of $600.743. This is both a high budget and a high gross profit, as the film with the next highest budget (Spider Man) had a budget of $139 million, and a gross profit of $403.706. I expected Titanic to have such a high gross profit, as it is the 2nd highest grossing movie of all time *(3).* There is also another group of films (Star Wars and ET) had a budget of $11 Million and 10.5 million respectively, both which are low budgets. But the Former had a gross profit of $460.9 million, and the latter had a gross profit of $434.9

million. These are extremely high gross profits for such a low budget film, and both these films can be considered y outliers. Later on I will investigate the effect of removing these outliers.

The relationship of the budget and gross profit appear to be approximately linear, so I am going to add a linear regression line.




```
Linear Trend

Gross.Profit...millions.of.US.dollars. = 2.1412 * Budget..millions.of.US.dollars. + 102.74
Correlation = 0.68369

Sample size: 50
```

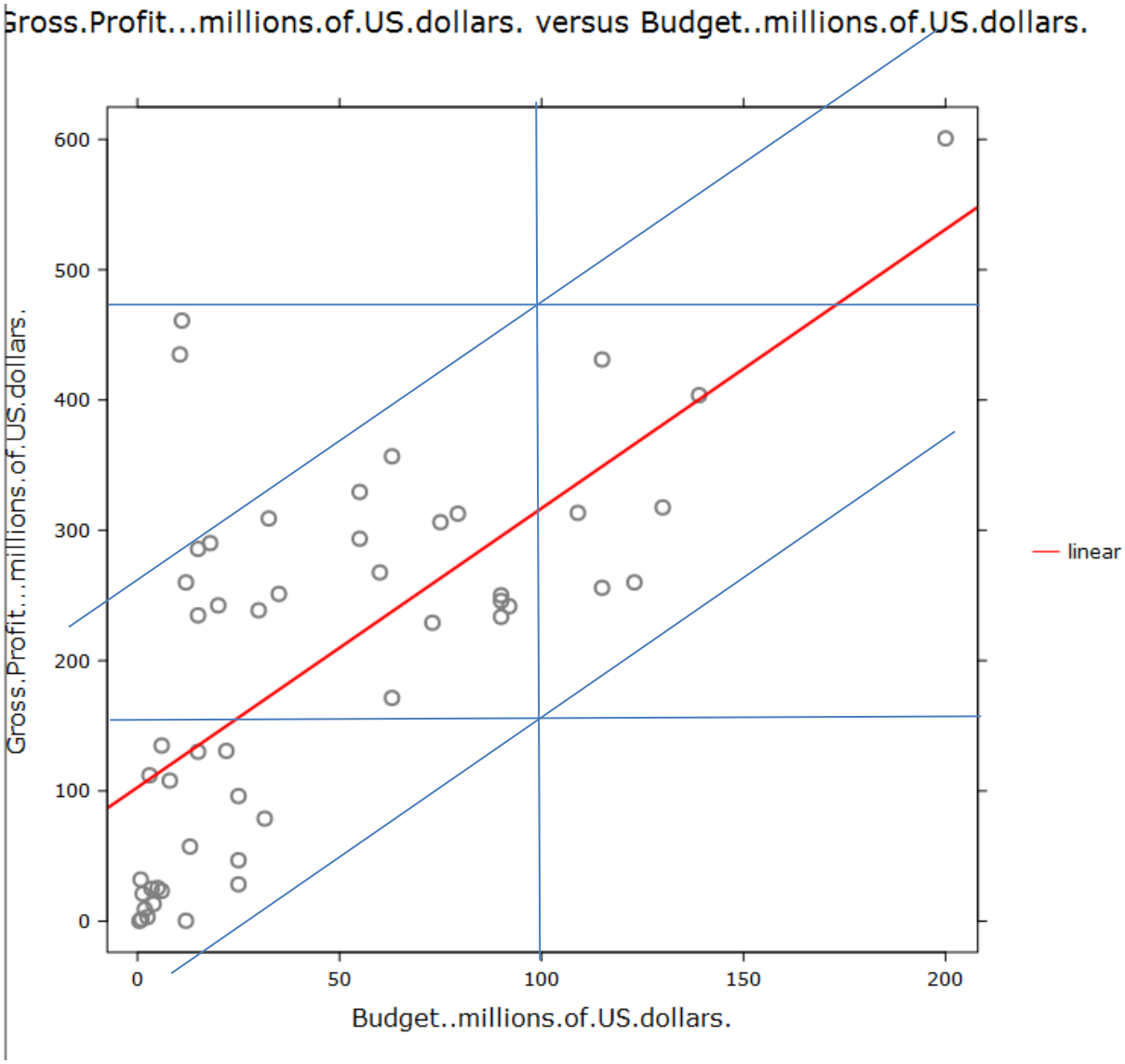
The equation for the Gross profit =2.1412*budget + 102.74. This means that for every dollar a movie maker invests, they can expect a profit of $2.14.

Just looking at the graph, there is a lot of scatter among the linear regression line. The correlation coefficient r for this graph is 0.68. This suggests that the linear relationship between the budget of a movie, and the gross profit it makes is weak.

I will use this equation to predict a value. For a movie with a budget of about $100 million dollars:

2.1412*100 + 102.74 = $316.86. So a movie with a $100 million budget can be expected to have a gross profit of $316.86 million.

There is quite a lot of scatter about the regression line, indicating a lot of variation. To take account of this variation, I will add prediction intervals about the linear regression line.

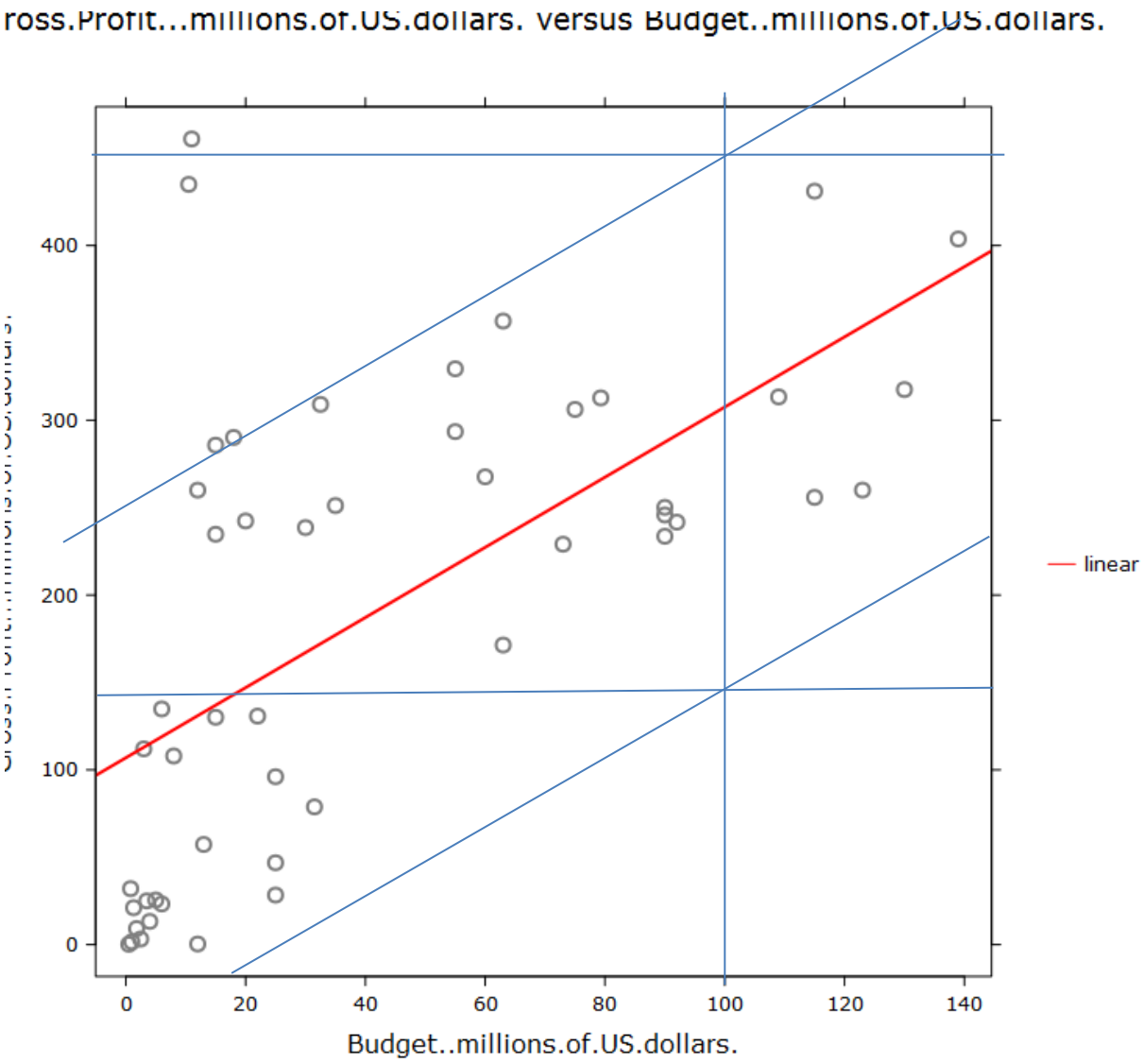


My prediction intervals indicate that for a film with a budget of $100, the grossing profit could be between $485 and $155.

So for a budget of $100 million, I can expect my gross profit to equal $316.86 +_ $165. As a percentage, a budget of $100 could have a gross profit of $316.86+_ 52%. This indicates that the prediction interval has a precision of 52%. The prediction interval is over a half of the actual graph this further shows that the relationship between these two variables is weak.

Removing Outliers:

The first outlier I will remove is the Titanic outlier, due to its high budget and high gross profit.



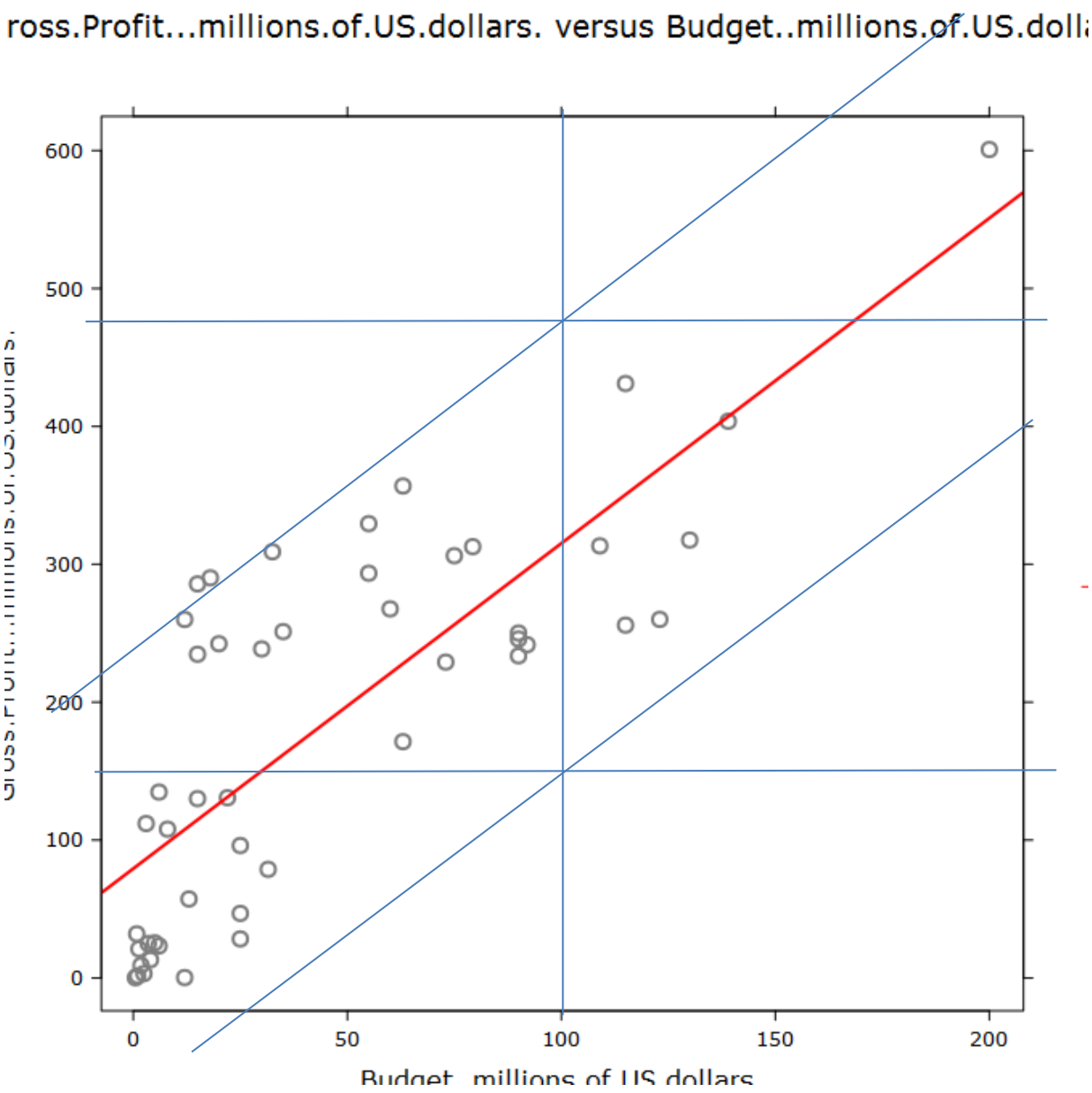

```
Linear Trend

Gross.Profit...millions.of.US.dollars. = 2.0067 * Budget..millions.of.US.dollars. + 106.88
Correlation = 0.61236

Sample size: 49
```

The new equation without the Titanic movie is the gross profit =2.0067*budget + 106.88. The Gradient has decreased from 2.141 down to 2.067, while the y intercept has increased from 102.74 to 106.88. The correlation coefficient r is now .61. The relationship is now even weaker than it was with the titanic outlier. The titanic movie acted as a point of high leverage, increasing the gradient.

The equation for a movie with a budget of $100 is 2.0067*100 + 106.88 = $307.55 gross profit. The prediction interval for a movie with a budget of $100 is between $140 and $475. This gives an equation of a $100 budget having a gross profit of $307.55 +_ 167 (approx.). 167/307.55 = 54.4%. The precision of the

prediction interval has increased even further, and the correlation coefficient has only gotten weaker. As this outlier has actually strengthened the relationship, I will keep it in the graph.

I will now remove the 2nd group of outliers, ET and Star Wars. Both these films have similar values, so I will remove them both at the same time.




```
Linear Trend

Gross.Profit...millions.of.US.dollars. = 2.3582 * Budget..millions.of.US.dollars. + 79.34
Correlation = 0.79565

Sample size: 48
```

Just looking at the graph, the scatter appears to have been reduced. While there still is quite a bit of scatter, it isn't to the same extent as the last graph. The equation for this new graph is
the gross profit = 2.3583*Budget +79.34. The gradient has increased from 2.141 to 2.358 and the y

intercept has also decreased from 102. 74 to 79.34. The correlation coefficient r has increased from .68 to .795. This indicates that the relationship has increased with the outliers remove.

Predicting a gross profit for a movie with a budget of $100 = 2.3582*100 +79.34 = 315.16.
The prediction interval is between $140 and $490. So for a movie with a budget of $100, the gross profit would equal 315.16 +- 157. As a percentage, the equation would be 315.16 +-49.8%. The prediction interval has a precision of 49.8%. This is a slightly more accurate precision compared to the previous graph, which had a precision of over 52%. This, and the fact that the correlation coefficient r has increased suggests that I should remove these two outliers (Star Wars and ET). But I should still keep the titanic outlier, as removing it has actually weakened the relationship when I removed it, and the precision increased from 52% to 54.4%.

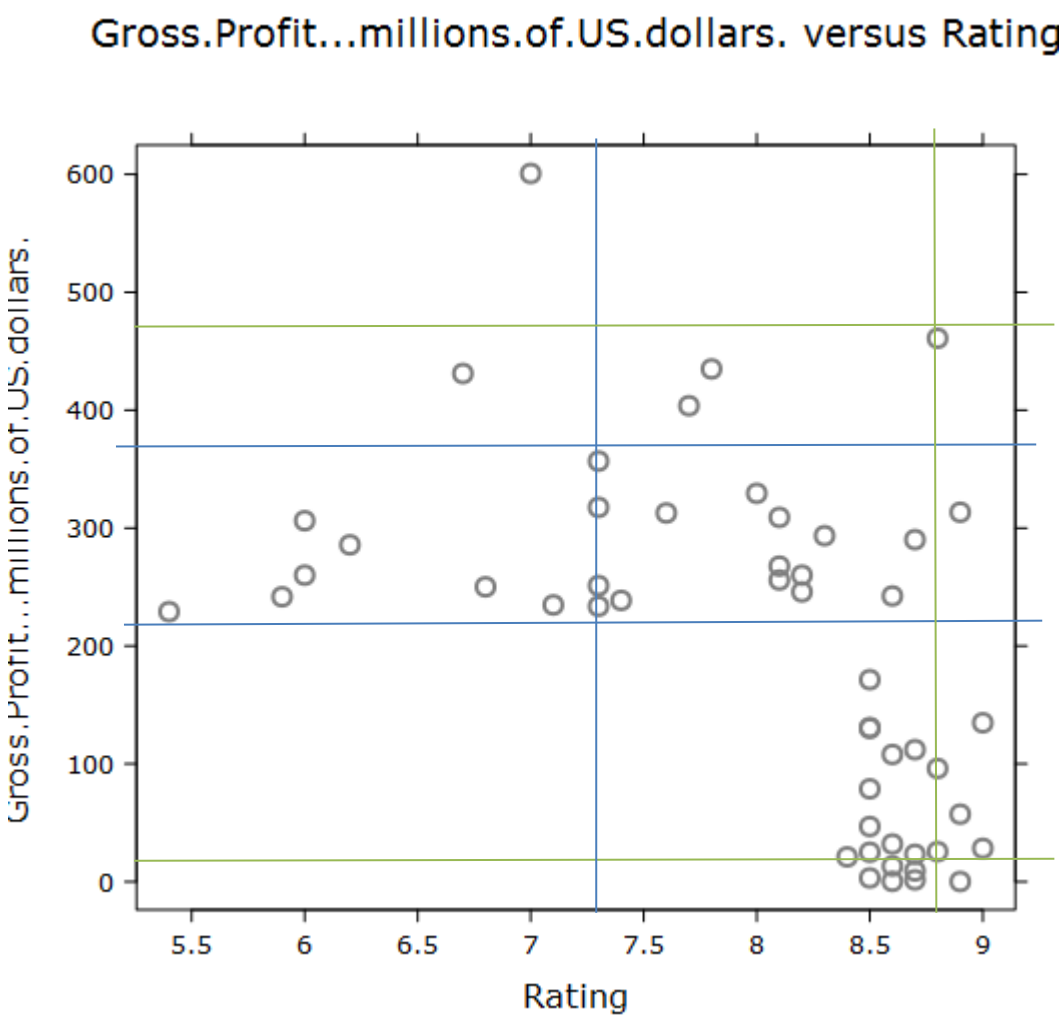
Saying that, this prediction model, even without the two outliers, has a precision of just under 50%, so it is not extremely reliable. It does appear that no matter how much money you invest into a movie, you can never be sure of its return. For example, a movie like Shawshank Redemption only grossed $16, which barely covered its budget (2). Yet a movie such as an E.T, with a gross profit of around $10, managed to bring in over $400.

For both the outliers Star Wars and E.T

Looking for research, Star Wars had a cheaper budget due to its use of green/blue screens, which is a cheaper alternative compared to building full-scale sets. Yet the special effects were still very advanced and drew in crowds (8). This technological advancement explains the unusually high gross profit and low budget that Star wars had. Meanwhile, ET was described as the 'the movie that touched the world'. Both adults and children immensely enjoyed it, and in the 80's, it was re-released twice in theatres twice (9). This immense popularity explains why ET had such a high gross profit, even though its budget was extremely low.

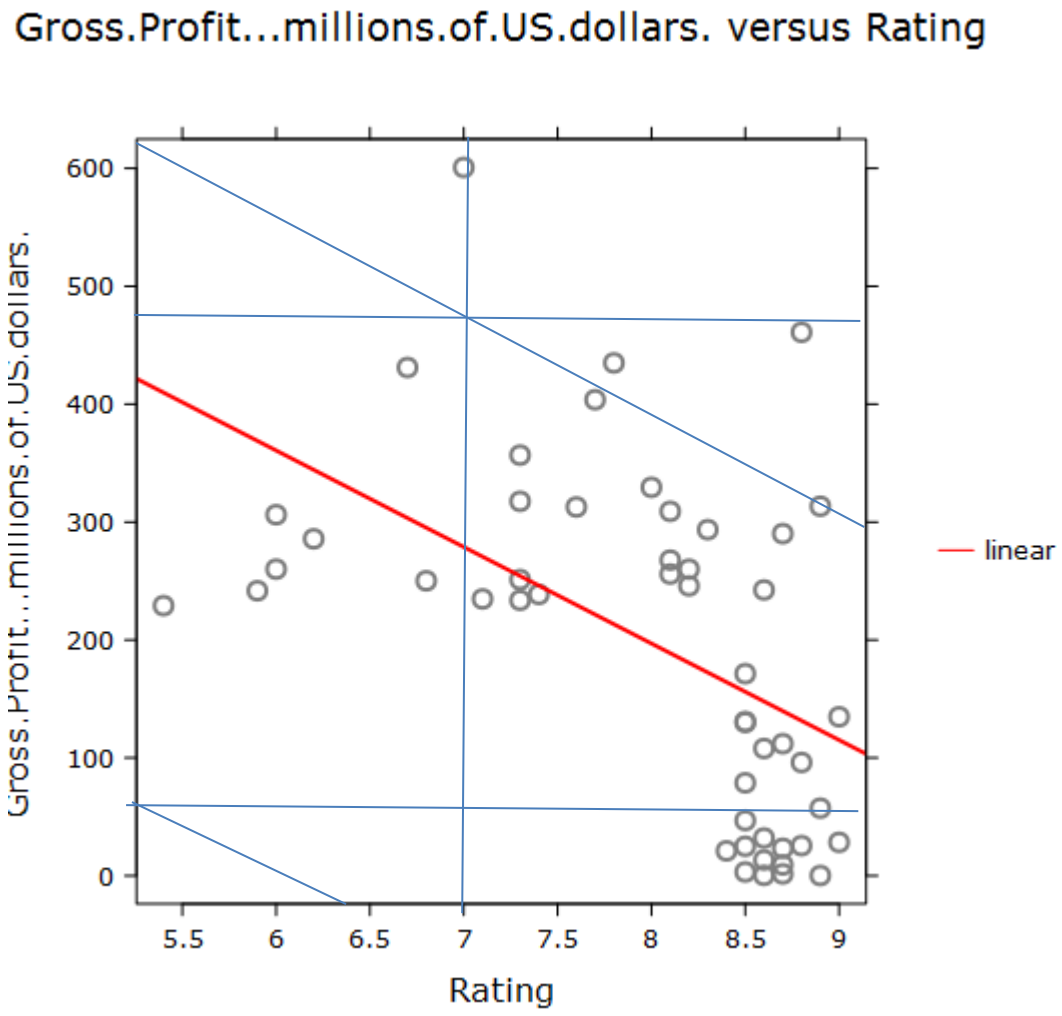
Investigating the 2nd Variable: Movie ratings.

I will now investigate the relationship between the ratings (average score from ratings given by users of the website IMDB), and the gross profit of a movie. This is to see if the ratings of a movie will be a better model to predict the gross profit.



Looking at the relationship, there appears to be a very weak, negative association between the rating of a film, and the Gross profit that a film makes. This suggests that as the rating of a movie increases, the gross profit it is expected to make will decrease. There is a lot of scatter, suggesting that this relationship is very weak. For any rating a movie receives, there is a large range of possible gross profits that movie could've made. For example, for a movie that earned a rating of 7.4 (approximately), the range of gross profits could be between $210 and $385. If you were to find to look at another movie with a rating of about 8.8, the gross profit could be anywhere between $1 and$470. This huge difference in values for gross profits, and the variation is nearly double the amount among certain ratings, leads me to suspect that there is some lurking variable which could explain the differences in scatter. There does appear to be one unusual value, Titanic, which despite its average score of 7, received a huge gross profit of $600.

Though it is a very weak relationship, the relationship still is approximately linear. And, because I want to compare this model specifically with the model used for the relationship between budget and gross profit, I will use add a linear regression line.




```
Linear Trend

Gross.Profit...millions.of.US.dollars. = -81.855 * Rating + 851.81
Correlation = -0.52211

Sample size: 50

'
```

The equation for this line is -81.55*rating + 851.81, which suggests that for every point of rating, the gross profit will decreased by $81.
The r correlation coefficient for this graph is -0.5. This is extremely weak, compared to the new revised graph (with ET and Star wars removed) for budget and gross profit, which had a correlation coefficient of .795. In fact it is questionable whether or not this graph can be modelled by a linear regression line. However, for the purpose of comparing it my original model for budget and gross profit of a movie, I will add a prediction interval around the linear regression line to take into account the scatter.

Before I do that, I will use the equation to predict the gross profit of a movie with a rating of 7.

Gross profit = -81.855*7 +851.81 =$278.825.

The prediction interval suggests that for a movie that had a rating of 7, the gross profit will be somewhere between $70 and $490. So the gross profit of a movie with a rating of 75 could equal $278.85 +- $210. As a

percentage, this is 278.85 +- 75%. This means that our prediction interval has a precision of 43%. As this prediction interval is nearly 75%. This is an unreasonable model – there is no point trying to use ratings to predict the gross profit of a movie, as the prediction interval is simply too high. Another model, perhaps a square, or a cubic model might fit the graph better, though that would be another investigation to consider. Based on this investigation, it is clearly better to use the budget of a movie to predict its gross profit.

Based on the graph above, one could come to the conclusion that the higher rating a movie earns, the less profit it would make; higher quality movies appreciated by the audience will earn less. Obviously, this is an incorrect conclusion. Correlation does not indicate causation. There must be some other lurking variable which explains such a relationship.
One reason could be due to genre. Traditionally, the most popular films were war films, musicals and historical dramas (3). If a movie is part of a popular genre, more people will want to watch it, regardless of its quality. So a high quality comedy movie might still not do as a well as a low quality musical. Also, there are release times. A movie that is released during the summer period, or Christmas period, will receive more views as it is during the holiday periods of USA. Thus there will be more people who have the time to go watch movies. So a movie released in summer could have a higher gross profit compared to a movie released in autumn, regardless of quality. (5). There is also inflation. The data we received did not take into account inflation. Money had a lower value in the past compared to now. So films released earlier would make less money due to the money value back then having a lower value. For example, Gone with the Wind does not make the top fifty highest grossing movies. But if inflation is accounted for, it would be the highest grossing film (3).

Conclusion

In conclusion, the model for the budget of a movie and its gross profit had a correlation coefficient of .795 and a prediction interval percentage of 49.8%. The model for the rating of a movie and its gross profit had a correlation coefficient r of -.5 and a prediction interval of 75%. Clearly rating does not fit a linear model, but for the sake of comparison, I placed a linear model on it in order to compare it to the budget of a movie. It is clear from this investigation that the budget of a film is a better model to predict the gross profit of a movie. The quality of a movie is no indication of whether or not it will make a high gross profit. Even so, the budget itself is still an unreliable prediction model for the gross profit of a movie. My research from a report (URL 5) says that a high budget means a high risk. No matter how much money you invest into a movie, you will never be able to predict its gross profit. It ultimately comes down to the audience.

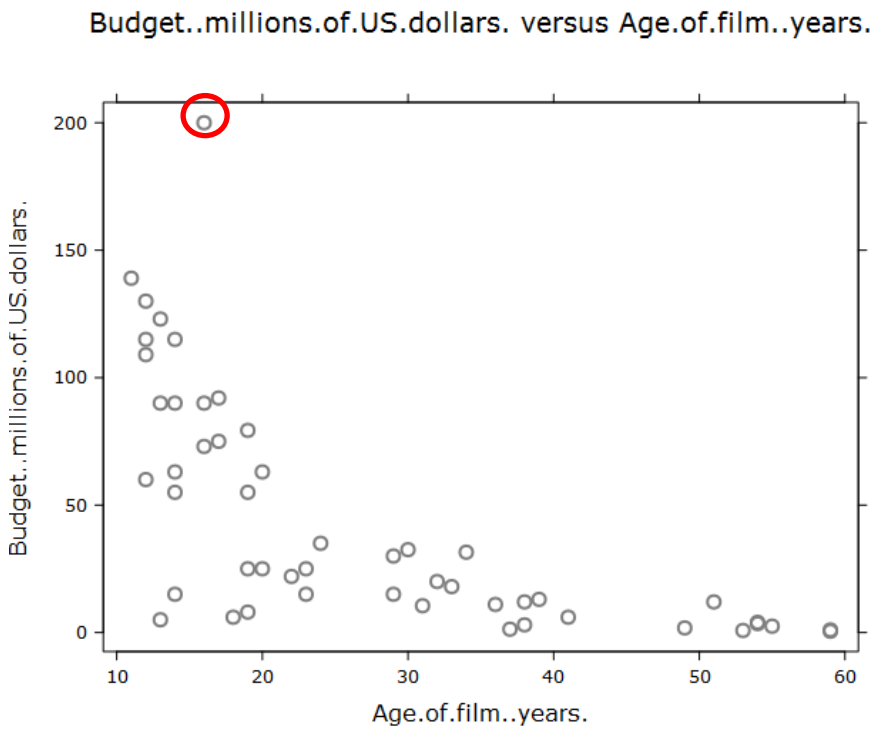
Awarded M only due inappropriate use of linear model and r in second question which was then repeated in her conclusion.

Guinea pig 2

I wonder if there is a relationship between the age of a film and the budget spent making the film ($US)?

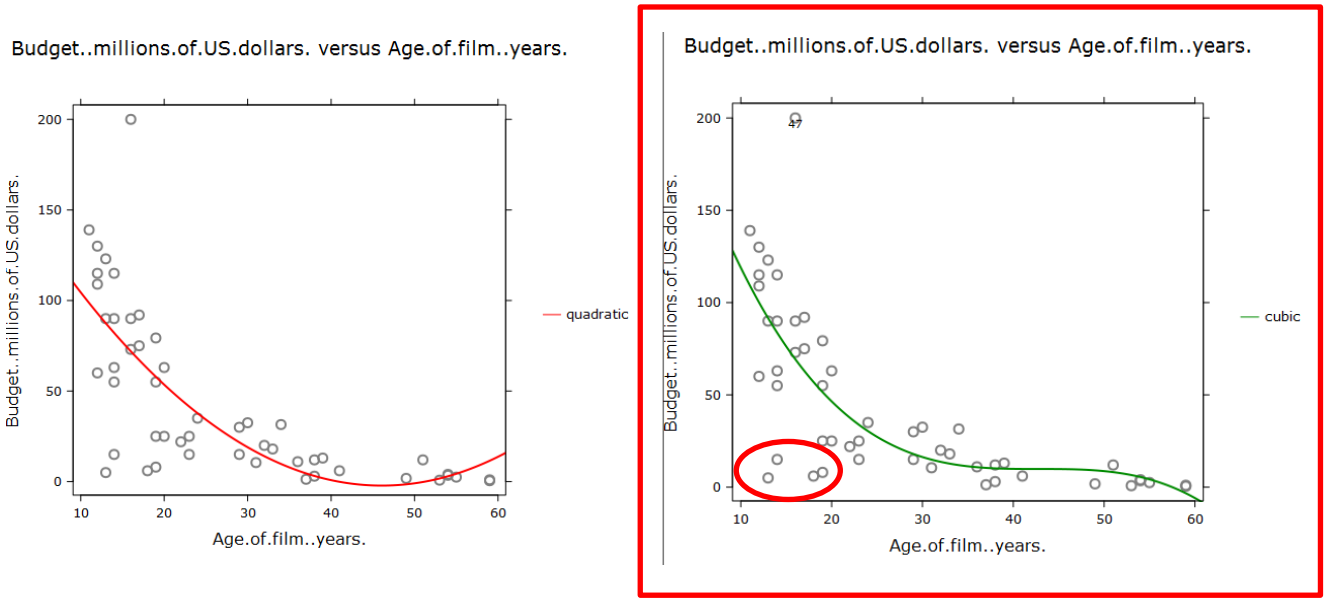
It is widely believed that the excessive budgets for films are always on the rise and 'will continue to balloon' (2). So, is this true? With the highest 20 film budgets being from the last decade, with the earliest release being in 2006 (3), I expect to agree. A large and continuously expanding portion of a film's budget is from its marketing. Only a few years ago the marketing costs would make up around 70% of a films budget, while today that percentage may be up to 80% (1). Studios can afford to these massive budgets as generally 'they get what they pay for, in that if they give a higher budget to a film bound to be a success, they end up with a larger return' (2). Clearly the strategy for film production companies is spend big, earn big. This appears to be successful as the highest 20 film budgets have led to profits for their production companies, with the worldwide gross profit being significantly more than the budget (3).

To undertake this investigation whether there is a relationship between the age and budget of a film, I shall place the data of 50 famous recent movies collected by Mohammad Raza of Wolfson College, Cambridge for his May 2003 Mathematical Tripos Part III essay 'Analysis of a large and complex data set' on a scatterplot and then add a regression line to further analyse the relationship. Following this, I shall separate the data into two groups; films that won awards and those that did not win awards, to see if there is a connection between the amount spent on a film and whether or not it won an award over time.



From this scatterplot, I observe a moderate, non-linear, negative association. This indicates that as the age of the film increases, the amount that it was made for decreases. There appears to be an outlier which has an unusually high budget for the time it was made. This value is from the 1997 film, '*Titanic*', which had a large budget of US$200 million compared to other films of the time whose budgets range between US$75 million and US$100 million. This extravagant budget was caused by the elaborate sets and the equipment

needed to recreate the ill-fated ship. As the data in the scatterplot is clearly nonlinear, I shall add either a cubic or a quadratic regression line to further analyse the relationship between the age and budget of the 50 famous recent films.




```
Cubic Trend

Budget..millions.of.US.dollars. = -16.978 * Age.of.film..years. + 0.39667 * Age.of.film..years.^2 + -0.0030891 * Age.of.film..years.^3 + 252.07

Sample size: 50
```

From placing both a quadratic and a cubic regression line on the scatterplot, I observe that the cubic regression line fits the data best. This also confirms my initial observation that the data was not linear as the regression line that fits best is curved. By adding the regression line I can see that the relationship between the age and budget of a film is fairly strong up until more recent films which are less than 25 years old. As there is a fair amount of data below the trend line for films made within the last 25 years, it is likely that predictions for this time using this regression line are likely to be overestimated. Is also appears as though the budget for films remained fairly consistent around US$1 million and US$20 million until about 35 years ago when the budget for films clearly increases.

Also from adding the regression line, I have created an equation to predict the budget of a film from when it was made.

For example: I can find an estimated budget for a film made 20 years ago by placing it in the equation formed by the cubic regression line.

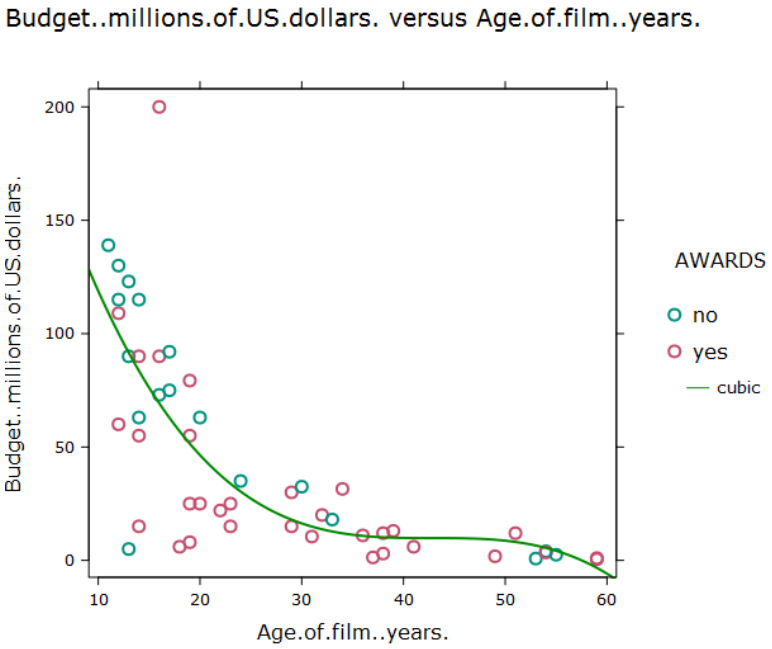
(-16.978*Age of Film+0.39667*Age of Film^2+-0.0030891*Age of Film^3+252.07)

-16.978*20+0.39667*20+-0.0030891*20+252.07=46.5

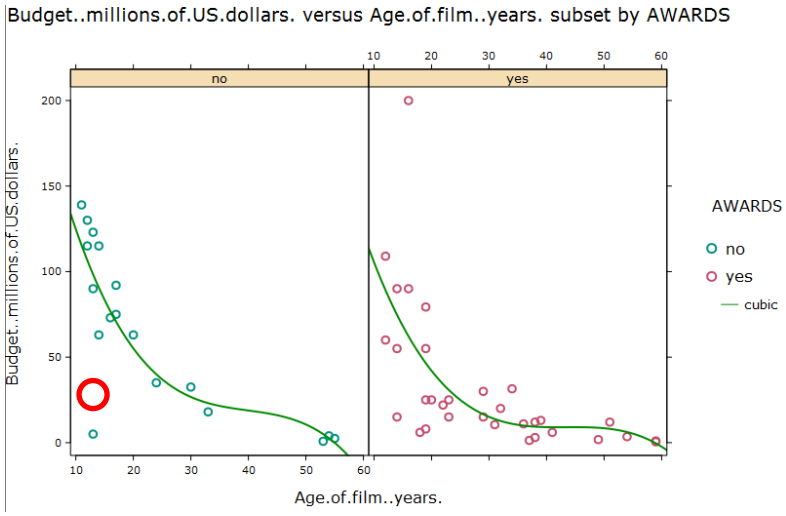
Therefore, I can estimate that a film which is 20 years old was made for a budget of US$46.5 million.

As this regression line is nonlinear, I cannot use the correlation coefficient to confirm the strength of the relationship.

I wonder if there is a difference in the age and budget of films which did win awards and those that did not win awards



To see whether the budget influenced a film winning an award, such as an Academy Award, Golden Globe, or BAFTA, over time, I have separated the data into two groups; films that won an award and those that did not. Initially, there is no clear observation that there is much of a difference between films with awards and films without awards as they appear to be fairly evenly scattered throughout the data. I will have to separate the two groups into individual scatterplots to further investigate whether the budget of a film has influenced whether it has won an award or not over time.



After separating the two groups, films which have and have not won awards, into individual scatterplots, I still cannot observe any obvious indications that the budget of a film influences whether it wins an award. However, I have noted that both groups data still follow a cubic regression line and the gradient of the films without an award is steeper than the gradient of the films with an award. This indicates a more dramatic increase of budget over time for films without an award. Also the data for films without an award is closer to the trend line than the data for films with an award, except for an unusual value which had a much lower budget than other films at the time. This indicates that the relationship of the budgets of films without awards over time is much stronger than the relationship of the budgets of films with an award over time. From splitting the data set into two groups, I have found that the budget of a film does not influence whether it will win an award or not, and therefore the budget does not indicate the quality of a film.

From this investigation I have found enough evidence to agree with my initial hypothesis that there is indeed a relationship between the age of a film and its budget. There is however, no indication that over

time there has been a suggestion of budget influencing whether a film wins awards.
This increase in film budgets has been caused by production companies spending all that they have to in order to create a film that will appeal to the biggest target audience possible so they can gain the biggest profit possible from the production. However, a factor that has had large influence on the film budgets is the inflation of the US economy. I have observed that film budgets began to increase dramatically from films that were 25 years old when this data was collected in 2003, therefore from the mid-1970s, film budgets have been rising at an increasing rate. This fits with the high inflation rates in the US during the 1970s. By the end of the decade, inflation had increased by 103% (4). Therefore, it may not actually be that films are costing more, just that film budgets change according to inflation. Also, I can only confirm my hypothesis based on the data from within the time frame in which the 50 recent famous movies were made and that data from outside this time frame may not necessarily fit with the trend line of the scatterplot.

Graded Excellence

Lacked the depth of other scripts but good research, appropriate, non-linear analysis with a consideration of award/no award groupings and no obvious misunderstandings.

Guinea pig 3

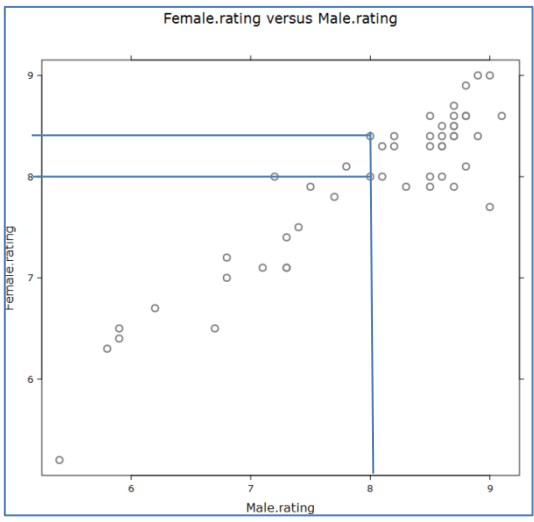
I wonder if there is a relationship between the male rating and female rating of films

Introduction

Gender stereotypes in regards to film genre preferences have largely expected 'a gender bias such that males would be partial to "men's films", and females would be partial to "women's films"' [1], as is consistent with the Social Identity Theory, which believes people seek out messages supporting their own personal identity. [2] According to Fischoff's studies, "males and females showed the widest differences in genre preference in action- adventure and romance."[1] In another research paper, "it was confirmed that females prefer romantic films, whereas males favour action films." [3]An understanding of gender preferences towards certain film genres, and any level of agreeability is especially important when marketing a film. Some bug films have an advertising expense of over $100 million, so aiming certain films at the gender most likely to respond positively, is essential for success. Similarly, couples often watch movies together, and if the marketing effectively captures the interest of both genders, they will be more likely to compromise on a genre they might not have picked out by themselves.

Today I will be investigating whether, like in previous studies, the difference in genre preference is evident between females and males, and if so, to what extent? Although similar studies have been previously conducted, they were completed at least a decade ago. I am interested to see whether the results of my own investigation are similar, or whether gender agreeability has increased during the past few years. To investigate this, I will be comparing the male rating and female rating of movies within the dataset by Mohammad Raza, and investigate whether the relationship between these ratings differ with different genres i.e. are females and males more likely to agree on their perceptions of a movie from one genre compared to another. For the interests of this study, I have combined the different genres into two subset genres: Action, Fantasy, and Thriller as the more action-based subset "A", and Comedy and Drama as the more relaxed, easy watching subset "C". I predict that there will be an obvious correlation between male and female ratings, but that it will be only of moderate strength. I also predict that there will be different levels of agreeability within each subset, compared with when they are all analysed together. Because gender differences have minimised over the past decade, and increasingly, movies are aimed at wider audiences of both genders, I believe that the results from the older studies (that females and males part ways when it comes to Action/ Drama) are less applicable and less accurate to today.

Initial Analysis

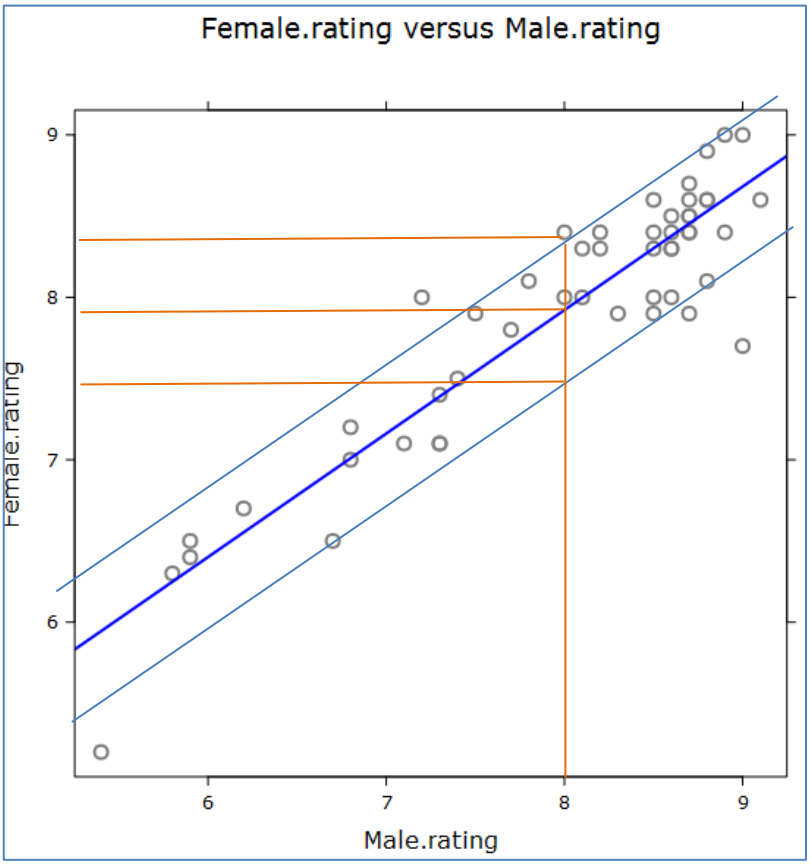


Upon initial analysis of this graph, there appears to be a moderately strong, positive association between the male rating and female rating of the films in this dataset. Being a positive relationship, a high male rating is expected to have a high female rating and vice versa. Although there is an evident correlation between the two variables, this is not due to a causational relationship, as female and male ratings are determined independently of each other. This is a moderately strong relationship as there is a clear linear relationship with only minimal scatter

around the trend. For example, a typical male rating of around 8, has female ratings ranging between 8 and 8.5, which shows scatter, both only on a minimal level.

Linear Regression Line

Because there is a noticeable linear trend, it is sensible to add a linear regression line that can calculate the strength of this relationship and give an indication of its usefulness as a predictor, which can then be compared to the individual analysis of the separate sub-sets.




```
Linear Trend

Female.rating = 0.76041 * Male.rating + 1.84
Correlation = 0.91981

Sample size: 50
```

Retrieving the summary statistics as below…

… gives us an equation for the relationship between male and female ratings in the form of y=mx+c, where:

Female rating = 0.76041 x Male rating + 1.84

Because the scales of both variables are the same, the gradient 0.76 already indicates a degree of agreeability. The correlation coefficient 'r', however, gives us the greatest indication of the strength of the association between female and male ratings. The correlation coefficient r= 0.920, which is close to one, concurs with the initial analysis that the relationship is moderately strong.

The equation derived from the linear regression line can also be used to predict values. For instance, using a typical male rating of 8, we can predict what the female rating will be:

Female rating = 0.76041 x Male rating + 1.84

= 0.76041 x 8 + 1.84 = 7.923

By this working, a typical male rating of 8 is expected to have a female rating between 7.53 and 8.32 (7.923± 0.4/5%). This range has been calculated from a prediction uncertainty set up on the graph to capture around 95% of the points around the trend line, thus, factoring in the amount of scatter and variability. As previously observed, this scatter is quite minimal, hence its strong relationship and low uncertainty of only 5%. The predicted rating of a female based off of a male rating shows a high degree of

agreeability between opinions regarding films of all genres collectively. Such a strong relationship was not what was seen in previous studies, already giving some indication that gender differences in regards to film popularity have changed over the past decade, becoming increasingly similar.

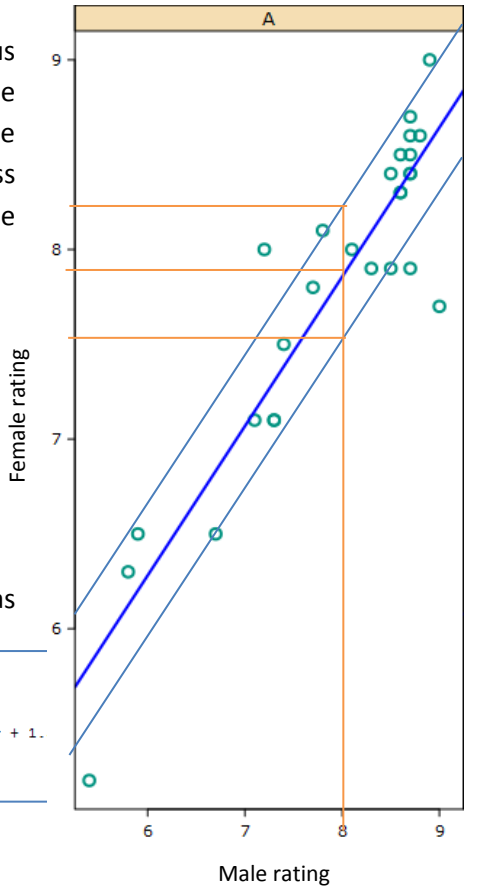
Groups

Previous studies mentioned in the introduction, have noted that males tend towards action-adventure films, while women prefer romance, and this preference demonstrates a wide difference between the two genders. Given that my initial analysis, which did not segregate genres, shows a very strong relationship contrary to what the other research papers concluded, I wonder if the same nature of agreeability between genders is apparent within the subset genres as well, and whether there is one genre that both genders are more likely to agree upon, and finally, how the following graphs compare to the original analysis. Coding variables by the subset genres has provided a graph which shows Genre A (Action/ Fantasy/ Thriller) in green and Genre C (Drama/ Comedy) in pink. Looking at this, there is no obviously significant difference between the two subsets, and there is a large amount of cross over within the same region. It does appear however that genre A has slightly more scatter (which would indicate less agreeability between genders) than genre C. Disregarding the minimum point for genre A, there is no noticeable shift or difference in the range of either genre either, which is expected, as both genres are expected to have a range of negatively and positively rated films. Upon initial analysis, it appears as if there is minimal difference between the two subsets, but this will be further analysed in the following graphs.

Group A: (Action/ Fantasy/ Thriller)

Based on the results of the previous studies, "males tend to focus more on action"[1], which are included within Genre A. By these conclusions, although the correlation coefficient might not change, the male rating should be higher than the female rating, and therefore, a less steep linear regression line would be expected. This is not seen on the graph to the right, providing some support for my hypothesis that the difference in gender preference has decreased over the years, at least, for the action/ fantasy/ thriller genre.

Female rating vs. Male Rating by subset Genre



Like the very first graph analysed, there is a steep, moderately strong positive association between the male rating and female rating, again, with only minimal scatter.

These observations are supported by the summary statistics, which show a correlation coefficient of 0.92, which is the exact same as the first r value calculated, indicating there is the same degree of strength between

```
Summary for Genre.1 = A

Linear Trend

Female.rating = 0.7876 * Male.rating + 1.
Correlation = 0.91781

Sample size: 27
```

the two relationships. The gradient of 0.78 is also very similar to the first gradient of 0.76, showing yet another aspect of this new graph which has the tiniest of difference to the original.

Again, a y=mx+c equation can be derived and used to predict a female rating from a typical male rating of 8:

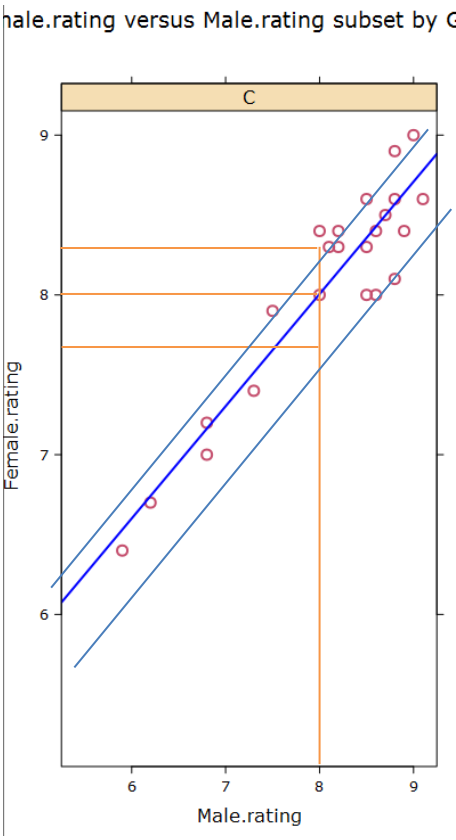
Female rating = 0.7876 x Male rating + 1.56

= 0.7876 x 8 + 1.56 = 7.8608

By this working, a typical male rating of 8 for a film in genre A is expected to have a female rating between 7.56 and 8.16 (7.86± 0.3/3.8%). The similarity in male ratings and expected female ratings is, again, very close, indicating a high level of agreeability. The prediction uncertainty range has decreased to 3.8% showing how removing genre B has decreased the amount of scatter, and increased the precision of the predicted female rating, but only by the tiniest of amounts. The strong level of agreeability contrasts what was concluded in past research papers, showing how males and females are now more likely to agree on their views or perceptions of a film in the action, thriller or fantasy genre and that this level of agreeability is not that different from the level seen when genres were not distinguished.

Group C: (Comedy/ Drama)

"Female viewers reported a greater preference for low-arousal films compared to male viewers,"[4] according to one study, which was backed up by another that "confirmed that females prefer films without a focus on action."[3] Again, by these conclusions it is expected that, even if the correlation coefficient remains similar, the female rating for this genre (which includes comedy and drama; "low-arousal" films "without a focus on action") should be higher than the male rating, and therefore, a steeper linear regression line would be expected. As with the previous subset investigation, this expectation is not seen in the graph to the right, and reinforces my belief that gender agreeability has increased across all genres within subsets A and C over the past decade.



Like the very first graph analysed, there is a steep, moderately strong positive association between the male rating and female rating, again, with only minimal scatter.

```
Summary for Genre.1 = C

Linear Trend

Female.rating = 0.7026 * Male.rating + 2.39
Correlation = 0.93484

Sample size: 23
```

These observations are supported by the summary statistics, which show a correlation coefficient of 0.93, which is only 0.01 higher that the first r value calculated, indicating that the degree of strength between the two relationships is very, very similar. The gradient of 0.70 is slightly lower than the gradient of the subset genre A of 0.78, meaning that there might be a little less agreeability within genre C than in genre A, though this may be affected by the also slightly higher y intercept. Again, however, this is only of minimal difference. For more in depth comparison, we can therefore use the prediction values.

A y=mx+c equation can be derived and used to predict a female rating from a typical male rating of 8:

Female rating = 0.7026 x Male rating + 2.39

= 0.7026 x 8 + 2.39=8.0108

By this working, a typical male rating of 8 for a film in genre C is expected to have a female rating between 7.71 and 8.31 (8.01± 0.3/3.7%). The similarity in male ratings and expected female ratings is, again, very close, indicating a high level of agreeability in both genre A and C. The prediction uncertainty range has decreased to 3.7% showing, again, how the subsets have less scatter when separated from each other. Again, this prediction shows a great degree of agreeability, and the high correlation coefficient indicates this relationship is quite strong.

Conclusion

The purpose of this investigation was to see whether there is a relationship between male rating and female rating, and if so, to what degree of agreeability do they share? Previous research indicated that there was a wide distinction between the gender preferences of certain genres, but this research came from 1994. Ten years on (based on the 2003 dataset), if this division is still apparent, the male rating and female rating relationship would be different when split into subset genres. If there was a distinct difference between the subsets then, for marketing purposes, it would make sense to look at different genres separately for information, rather than when together. The initial graph with its undistinguished genres showed a moderately strong relationship with a correlation coefficient close to one, and a prediction interval for the female rating which was very close to the typical male rating of 8. This shows a high level of agreeability between genders for films in general. The previous research papers implied that this initial strong relationship may be deceiving, because when split into subsets, they found a bigger difference in gender preference. Subsets genre A and genre C in my research, however, showed very little difference from the original analysis. Both subsets had a correlation coefficient that was equal to or very, very close to the original, indicating a very similar strength for the relationship between male rating and female rating. Similarly, the predicted range of a female rating for a typical male rating of 8 was very similar between subsets and in comparison to the original prediction. Despite minimal fluctuation between them, all indicated a high level of agreeability, as the prediction ranges all encompassed the male rating, 8. Because of this minimal difference between all graphs analysed, it does not make sense to investigate the different subsets separately, as the holistic graph gave us a very accurate portrayal to begin with. This backs up my hypothesis that less difference between genders is seen now. I did also predict there would be a positive relationship between female and male rating, but it turned out to be even stronger than the moderate strength I expected. Similarly, there is also less difference between genres than I expected. The only noticeable difference when the groups were split up, was the decreased prediction uncertainty range, which showed how separating them decreased the amount of scatter, but this was only to minimal effect

It can therefore be seen that, since the initial papers conducted twenty years ago which indicated wide gender difference for genre preference, there is now a much higher degree of agreeability. Stereotypical preferences have decreased, and movies such as those portrayed in this dataset (which were collated 10 years after Fischoff's study) now market themselves for both genders, or often aim to include aspects of multiple genres to appeal to a wider audience. This would help to explain why both males and females are more likely to agree on their ratings of certain films despite what genres they may have traditionally been aimed at just one gender.

Limitations

Although this research has shown sufficient evidence to back up my conclusion, there are some limitations of the process I went through, which may have impacted results. As an observational study, my data was limited and less recent. Derived from Mohammad Raza's dataset, there is no indication that the films investigated are a random, representative sample of all films in the industry. For his essay, Mohammad may have selected specific films to back up his findings, or selected specific films. If this dataset is not representative of the film industry and all its films, then my conclusion has less weight as a generalised statement for all genres and films. For instance, because these films are better known, it may have influenced the level of agreeability between genders, while less well-known films may have had a wider difference. Similarly, there may have been lurking variables which influenced this dataset that are not easy to trace, as it is an observational study. Age of the person voting may have had some influence, as well as the ratio of the genders voting. Fischoff too noted the possibility of an age difference, noting that "the gender-genre differences are most dramatically expressed in the age group below 26 yrs"[1] Because my correlation is so strong, however, and the conclusions are so very clear, these limitations may have had little effect, though, to be even more accurate, further tests might be conducted. To relate this study to 2014, a similar investigation could be conducted, which includes even more modern films, to see if there has developed an even stronger level of agreeability like it did from Fischoff's research to the one in this paper.

References

1. S. Fischoff, Ph. D et al. *Favourite Films and Film Genres As A Function of Race, Age, and Gender.* http://web.calstatela.edu/faculty/sfischo/media3.html
2. Abrams & Hogg, 1990
3. O. Chausson, *Assessing the Impact Of Gender And Personality On Film Preferences,* http://mypersonality.org/wiki/lib/exe/fetch.php?media=assessing_the_impact_of_gender_and_personality_on_film_preferences.pdf
4. http://comminfo.rutgers.edu/~kgreene/research/pdf/BanderjeeGreeneKrcmarBadRug2008JMPtry3.pdf
5. http://warmovies.about.com/od/FilmCriticism/fl/Removing-the-Mystery-of-Box-Office-and-film-Budgets/htm

Graded a strong Excellence script. Insight clearly shown in a thorough comparison with a previous study, limitations discussed and no inappropriate analysis.

Guinea pig 4

Introduction:

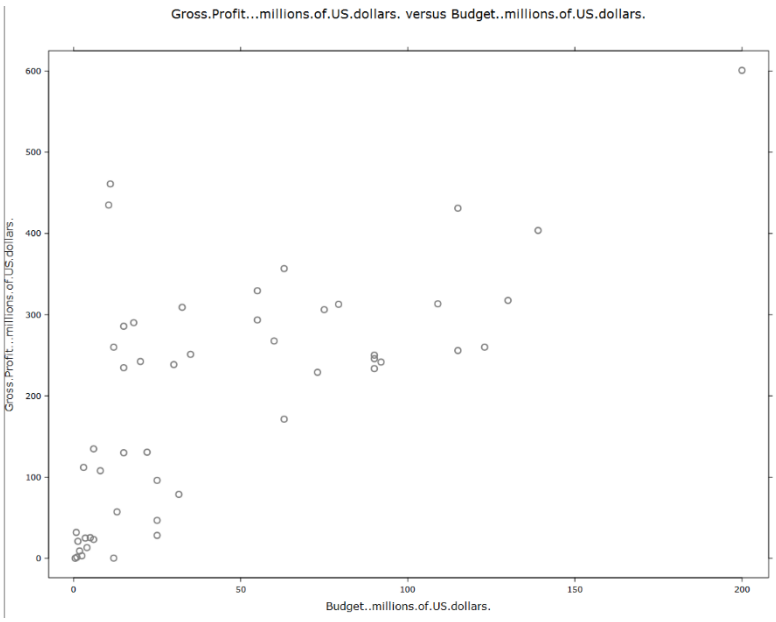
The movie making business has come to be a very profitable area, thus I wonder if the amount of budget a movie has is a good indicator of the gross profit it will receive. Research has suggested this "number of big budget films with potentially high revenue "(research number 3 .) Also general knowledge tells us that when directors have a higher budget they can hire "stars" to be in the film which could lead to more people seeing the movie due to good acting and as the celebrity is in the movie and thus a higher profit. Although , the profit of a movie is also dependent on lurking variables as research has suggested " three timing variables, holidays summer or Christmas release… in all three cases the third quartile of gross is noticeably higher for the seasonal release then non seasonal releases" (research number 2) . Thus this suggests that timing of the movie's release is also a factor to consider. Research also suggests that when movies have a higher rating they will thus have a higher profit " Shawshank redemption gross profit of £ 2,344,349 and a raiting of 9.3 /10 " (research number 7) compared to " chariman of the board gross profit being $306,795 and the rating being 2.3/ 10 " (research number 8) . This research shows that for a movie with a higher rating the profit they receive is much larger therefore I wonder if rating is also a variable that affects profit, I will group the variables to test this.

I am using a dataset of 50 most famous recent movies with the budgets and profits being taken of the IMDb website.

I am having the x variable as budget and the y (response variable) as profit.

Question:

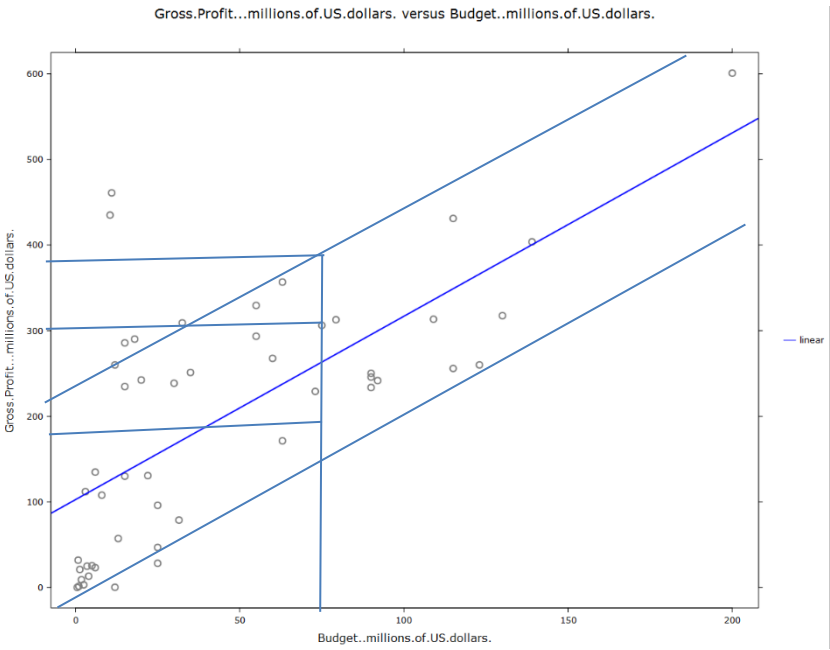
I wonder if there is a relationship between the budget of a movie and the gross profit they make.



Analyse:

There is a tendency for the gross profit to increase when the budget of a film is increased / is larger. The graph shows a positive linear relationship meaning that when the budget of a film increases, the profit made also increases. However, this relationship is moderately strong as there is a lot of scatter about the data , as when looking at a budget of 75 million US dollars, the profit made can vary between approximately 225 million dollars to 380 million dollars (measured both in US dollars). There are three outliers in the data set, these being; row 47 the movie; Titanic with it having a very large budget of 200 million dollars and a profit of 600, 743 million dollars. Also row 7 being the movie ET, with it having a small budget of 10.5 million dollars and generating a profit (as of 2013) of 434 949 million dollars. And also the outlier, row 33 being the movie Star Wars, having again a small budget of 11 million dollars but generating a large profit of 460.936 million dollars. I also notice that the budgets in the data set are majority below 150 million dollars and they generate a profit generally below 480 million dollars.

Since the relationship between the amount of budget the movie had and the profit generated is linear I will add a regression line to test how strong the relationship is, I expect it the correlation coefficient to be below 7 as there is not a strong relationship seen.



Linear Trend

Gross.Profit...millions.of.US.dollars. = 2.1412 * Budget..millions.of.US.dollars. + 102.74

Correlation = 0.68369

Sample size: 50.

Adding the regression line has given evidence that the relationship between the budget of a film and the profit it generates is not strong, it is fairly weak as the correlation coefficient , r is 0.68 meaning that it shows a moderately strong relationship. If the correlation coefficient r was 1 then the relationship would be called perfectly linear and there would be no scatter about the regression line. The regression line has a large amount of scatter about it, this gives evidence that there is a large amount of variability about the data set and the regression line. I also notice that the outliers are all highly above the regression line showing that they are outliers and do have large variability then the data set.

Adding the regression line also tells me that for every one million US dollar increase in the budget, the profit will increase by 2.14 million US dollars.

Prediction:

Guinea pig scripts for Bi-Variate presentation Workshop PN Auckland University Statistics day 2014

I will use the budget of a film to predict the profit it has made using the equation:
Gross.Profit...millions.of.US.dollars. = 2.1412 * Budget..millions.of.US.dollars. + 102.74

I will use a budget of 75 million dollars US = 2.14 x 75 + 102.74 = 263.24 million dollars.

However as there is a large amount of scatter about the regression line there is a large prediction interval of ± 120 million dollars. Thus the profit of a film for a budget of 75 million dollars could be 263.24 million dollars but it could be up to 383 million dollars and as low as 143 million dollars due to this large amount of scatter.

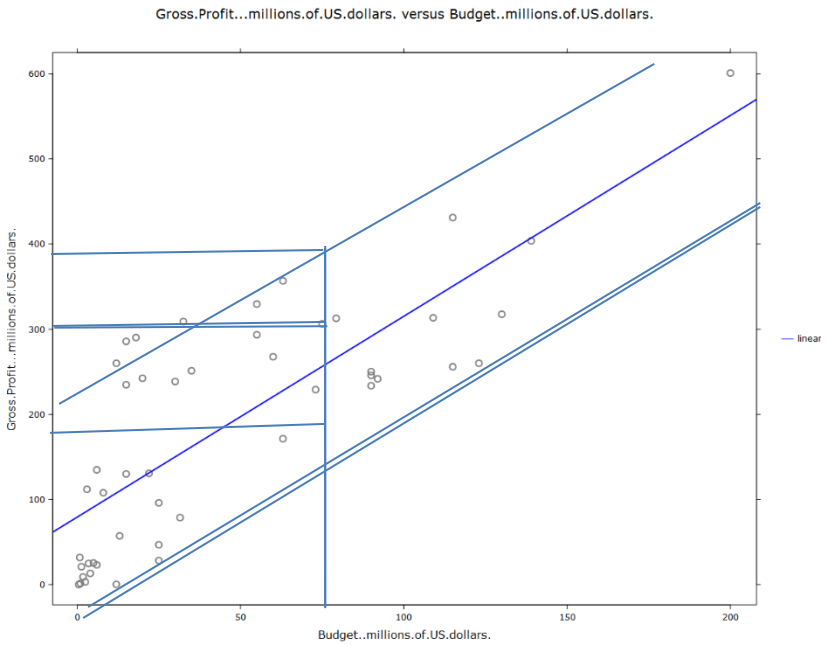
Percentage error:

I will test the amount of error in the data set to see how reliable and precise the prediction is:

383divided by 263 divided by 263 x 100 = 0.55 %. This means that there is a high amount of error in the data set thus it is not precise.

Outliers:

Since there are three outliers in the data set, I will see if removing them, makes r, the correlation coefficient to become stronger. (However I will not remove row 47 as it is in line with the rest of the data set and thus removing it will not affect r). As I see that the y outliers being row 7 and row 33, also make the regression line pull towards it. I will thus remove each outlier, taking the y outliers out together and to redo the predictions and see if r is made stronger. (I will not take out outlier's row 7 and 33 separately as they are extremely close together and taking one without the other will not change r as one will still be present to make the regression line still be pulled towards it.)



Linear Trend

Gross.Profit...millions.of.US.dollars. = 2.3582 * Budget..millions.of.US.dollars. + 79.34

Correlation = 0.79565

Sample size: 48

With taking out the two y outliers being row 7 and 33, it has made r, the correlation coefficient become much stronger. The original correlation coefficient was 0.68 and now it has increased to 0.79. This is a large increase in R and has now made the relationship between budgets being a good indicator of profit be a moderately strong relationship as it is very close to being 0.80. This means that the outliers were making the regression line be pulled towards the outliers instead of it being representative of the data set.

I will redo the prediction to see whether the prediction interval has been made smaller with taking out the outliers.

I will use the same budget as before being a budget of 75 million US dollars.

2.3582 x 75 + 79.34 = 255.59 million US dollars. This is slightly smaller than the original prediction with the original amount being predicated to be 263 million US dollars. Thus there is a difference of 8 million dollars.

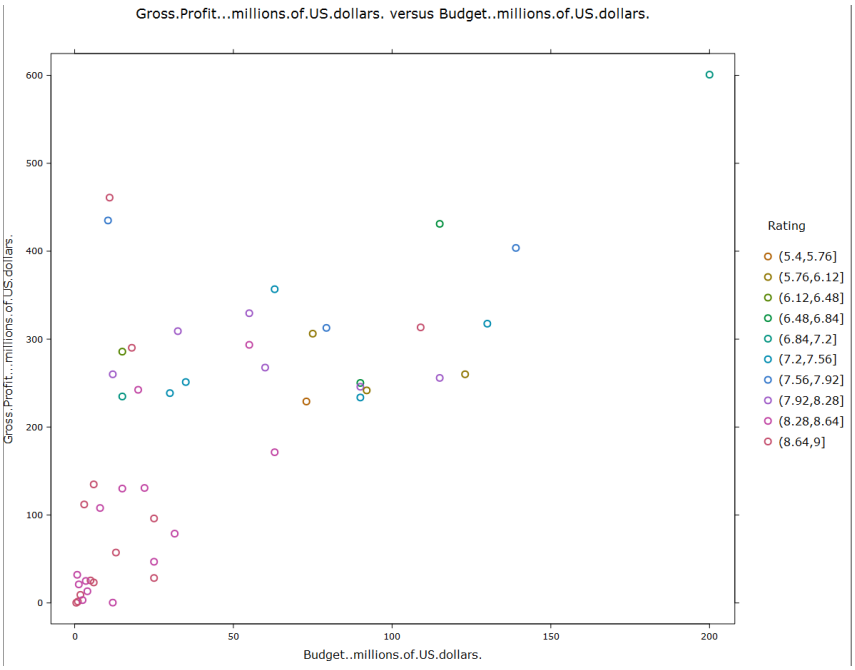
As there is still moderate scatter about the regression line the prediction interval is going to be ± 125 million dollars. This means that the profit made for a budget of 75 million US dollars could be 255 million dollars but it could be as high as 380 million dollars and as low as 130 million US dollars. This approximately the same prediction interval as with the original data set, meaning that taking the outliers out does not make a small prediction interval.

<u>Percentage error:</u>
380 divided by 255 divided by 255 x 100 = 0.58 , thus with removing the outliers, there is a slightly more precise prediction but it is by a very small amount as the original was 0.55 % and now it is 0.58.

<u>Groupings:</u>

I am going to group the variables to see whether movies with a bigger budget obtain more profit and thus have a higher rating then movies with smaller budgets and thus smaller profits.

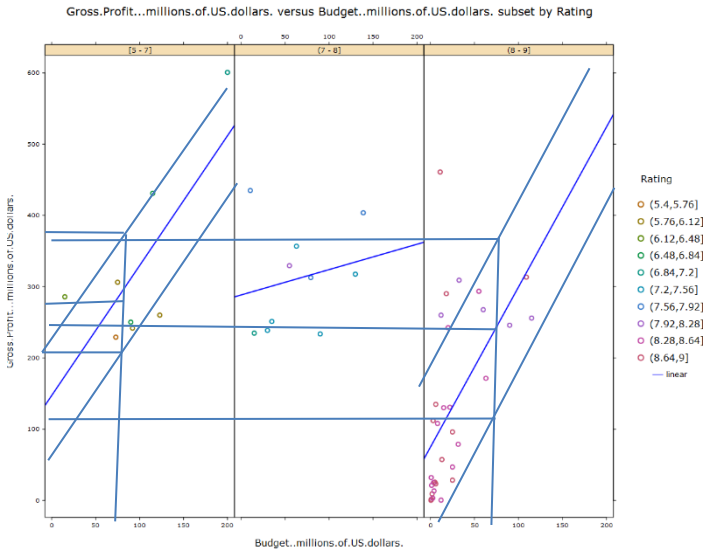
As I have found the research " In other words , movie viewers talk more about successful movies which affects revenues and ratings than unsuccessful movies " (research number 6). This means that then people see a movie that they enjoy they will talk about it more and rate it higher which thus will lead to higher revenue and thus higher gross profit. Thus I expect higher rated movies to have a higher profit.



From grouping the data into the ratings each movie was given it can now be seen that (from the key with dark pink and purples being the higher ratings and mustard and green colours being the lower ratings) that there is a strong relationship between the budget of a movie , and its profit and the ratings it gains . We can see that movies with a lower budget and thus a lower profit tend to be more highly rated on a scale of 1-10 with one being awful and 10 being excellent (source booklet). This is as there is a clear large group of pink and purple colours where there is low budget and profit movies. We also see that there is more highly rated movies having a rating bigger than 7.92 with a budget

lower than 60 million apart from 3 movies these being row; 18 being the movie LOTR FELLOWSHIP which had a rating of 8.9 , a budget of 109 million dollars and a profit of 313.364 million dollars. Also the row number 21 being the movie MONSTERS INC, having a rating of 8.1, a budget of 115 million US dollars and a profit of 255.87 million US dollars. Also the row number 48 which is the movie TOY STORY 2 which had a rating of 8.2, a budget of 90 million dollars and a gross profit of 245.823 million US dollars. (This is opposing to research as research states " high revenues come from animated films " (research number 1) , thus as monsters INC and toy story are both animated this is unusual. The group has a profit lower than 330 million and thus there is movies rated lower having a rating below 6.84 having a budget bigger than 70 million dollars and a profit higher than 280 million US dollars.

I will split the groups into rating to see whether the higher rated films with lower budgets and thus lower profits have a strong relationship, thus having a bigger correlation coefficient and a smaller prediction interval.



I will not be doing the rating of 7-8 as I am interested in the lowest and the highest ratings. Also the pink / purple which are the highest ratings are the ones I am looking at similarly the green and mustard coloured – the lowest ratings are the ratings I am interested in.

Summary for Rating = [5 - 7]

Linear Trend

Gross.Profit...millions.of.US.dollars. = 1.8237 * Budget..millions.of.US.dollars. + 147.09

Correlation = 0.74977

Sample size: 8

I notice that for the lowest ratings of 5-7 the correlation coefficient, r Is 0.75 , this is showing a moderately strong positive relationship between the budget of a movie , the profit and now the rating being between 5 and 7 . This correlation coefficient is larger than the original correlation coefficient as it was originally 0.68 showing that for movies with a budget from 0-200 and their profit being from 200 million dollars to 600 million dollars there is a strong relationship with them all having a rating between 5 and 7.

I will redo the prediction to see whether the prediction interval becomes smaller with there being lower rated films.

I will use the original figure of 75 million dollars budget to predict the profit of a movie;

= 1.8237 x 75 + 147.09 = 283.59 million US dollars. This is a larger amount of profit predicted then for the original data as the original prediction was 263 million US dollars. This shows that the original dataset was under predicting the profit for a movie rated between 5 and 7 with a budget of 75 million US dollars. This means that using the original dataset is not as reliable as grouping them and using their ratings as well.

However as there is scatter about the regression line , the profit could be 283.59 million US dollars but it could be in a prediction interval of ± 100 meaning it could be as high as 383 million US dollars and as low as

183 million US dollars. This prediction interval is slightly lower than the original as the original prediction interval was ±120, this shows that grouping and using a rating between 5 and 7 shows a more reliable prediction as the interval is smaller.

Summary for Rating = (8 – 9)

Linear Trend

Gross.Profit...millions.of.US.dollars. = 2.2442 * Budget..millions.of.US.dollars. + 74.94

Correlation = 0.55977

Sample size: 32

I notice that for the higher rated movies they have a correlation coefficient, r being 0.56 thus there is a moderately strong relationship between low budget and low profit movies having a higher rating. The correlation coefficient has now decreased from the original as the original was 0.68 and now it has decreased to 0.56 thus now there is a slightly weaker relationship. Thus now it is a strong relationship.

I will redo the prediction to see whether for using the same budget as the original, this being a budget of 75 million US dollars will predict a profit similar then before. I am testing also to see whether the prediction interval is made smaller.

2.24 X 75 + 74.94 = 243 million US dollars.

This prediction is higher than the original prediction as the original was 263 million US dollars, thus the original data set was over predicting the amount of profit made for a low budget movie, thus grouping them makes it a more reliable prediction.

This means that the amount of budget of 75 million dollars the profit will be 243 however as there is scatter about the regression line, the prediction interval is going to be ± 100 million US dollars. Thus the profit could be as high as 343 million US dollars and as low as 143 million US dollars. This prediction interval is lower than the original as the original was ± 120 , now it has decreased to ± 100 thus with having the low budget thus low profit and high rated movies, it has strengthened the relationship of budget being a good indicator for profit as the prediction interval has deceased. Thus grouping them has made a more reliable prediction.

From this I have found that higher rated movies have a lower budget and a lower profit, this means that profit is not affected by rating as the research suggested.

<u>Conclusion:</u>

I asked the question, I wonder if budget is a good predictor of the gross profit a movie will make. I have found out that there is a strong relationship between the two and generally when a movie has a higher budget the movie will have a higher gross profit. I wondered if grouping the variables into rating affected the amount of profit they received , I came to find that lower budget movies tend to have a lower profit however a higher rating , thus there does not seem to be a relationship between the rating of a movie and

the profit it receives. However, it has been seen that there are many lurking variables in this data set meaning there is reliability issues and limitations. This is as research states “release dates are one of the most important factors in determining the success of a movie” (research number 3) this means that as I have not accessed the release date of a movie , there is more factors that affect the profit then just the budget. Also I have found that the lurking variable that we do not know if any of the movies were released at the same time which as research says “ two movies of the same type opening at the same time can result in result in reduced revenues for one or the other.” (research number 5) thus if two movies were released at the same time – which I do not know , then budget could not have been the variable that was affecting profit thus this is a limitation. This investigation would have been useful for potential directors/ producers as it can tell them that the budget of a movie affects the profit they will make, also it can tell them that knowing when the other movies are being released and making sure there are no clashes will also affect the overall gross profit and revenue they will make.

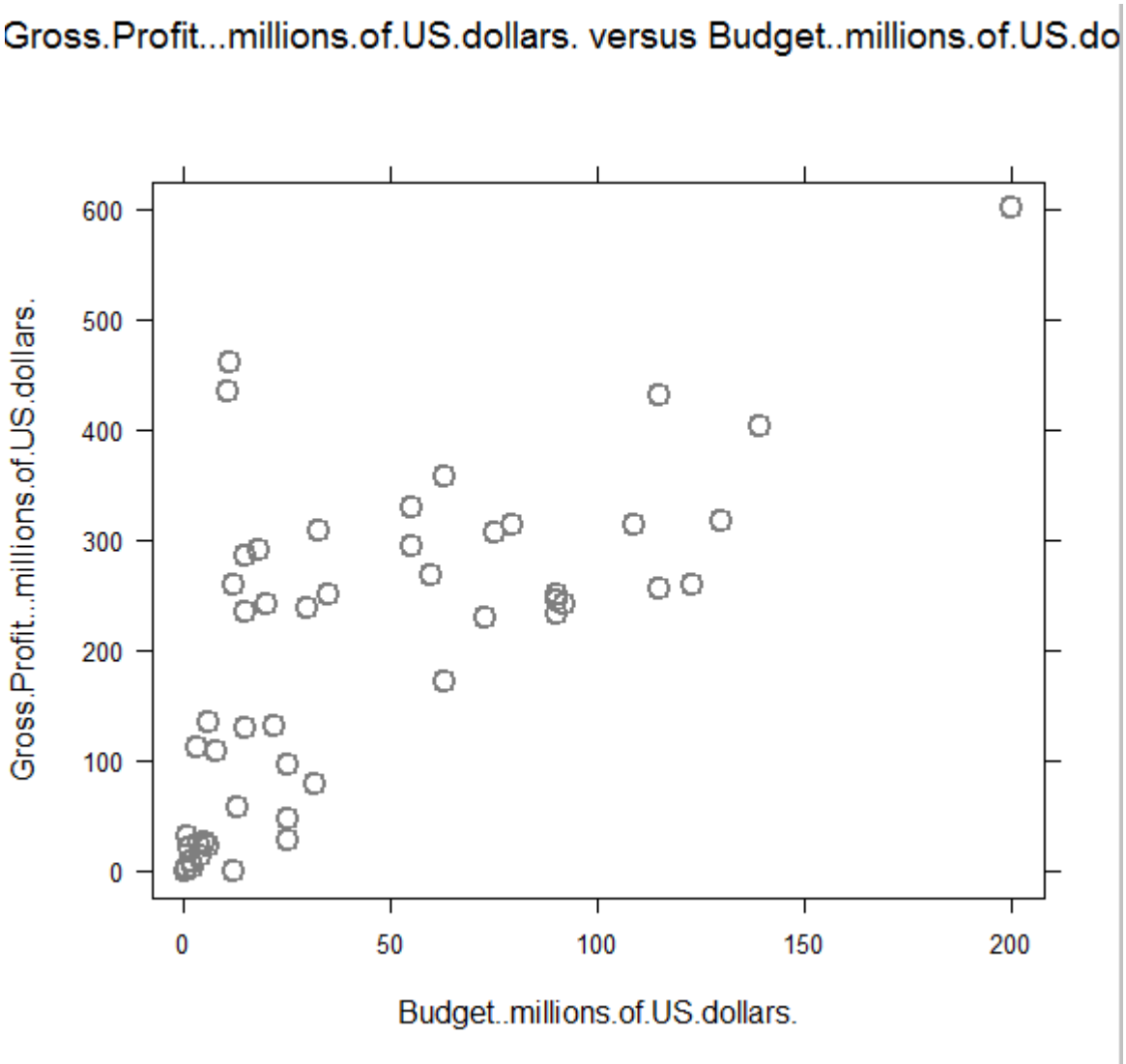
Grade:

Restricted to M as rating and grouping discussion was very hard to follow and putting linear models through small clusters (e.g sample size 8) and then commenting on the increase in r was inappropriate.

Guinea pig 5

For every movie made there is a budget of how much money will be spent on making the film, however the budget that is set for each film is different. Some movies will have much larger budgets than others http://en.wikipedia.org/wiki/budget. Once a movie is made and released it will start to make a gross profit. A gross product is defined as "a company's revenue minus its costs of goods sold. Gross profit is a company's residual profit after selling a product or service and deducting the cost associated with its product sale". The gross profit can also vary depending on how successful or not it is http://www.investopedia.com/terms/g/grossprofit.asp. From looking at research from http://www.the-numbers.com/movie /budgets/ we can see a list of films that have a large budget and make much more gross profit compared to films that have a smaller set budget. For example when we look at the movie Avatar it had a set production budget of $425,000,000 and when we look at the gross profit that this movie made we see that it is $2,783,918,982. When we look at a movie with a lower set production budget such as The last house on the left which had a budget of $87,000 and made a gross profit of $3,100,000. We see that there is a huge different of gross profit made from movies that have different budgets. Therefore I would like to know **"I wonder if there is a relationship between the budget of a movie and the gross profit that the movie makes."**

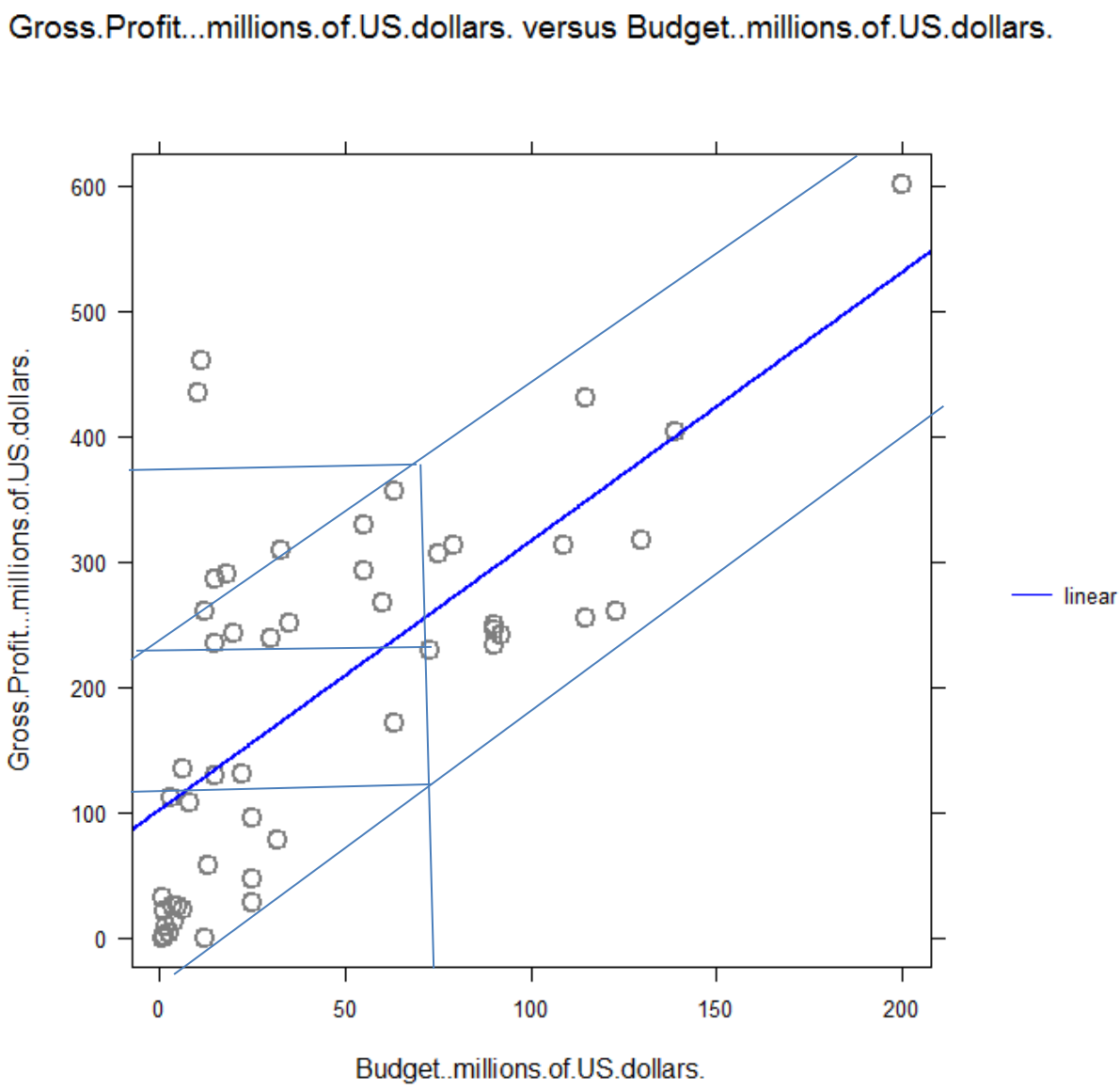
I will place the budget of the film on the x axis as this is the explanatory variable and the gross profit made on the Y axis as this is the response variable.



This is approximately a linear trend between the budget of a film and the gross profit made. There is a positive association, this is seen as

when the set budget of a film increase the gross profit that is made also increases. The relationship between the budget of a film and the gross profit made is moderate, this is seen with the moderate amount of scatter. We can see that the variation is fairly consistent. For example when we look at a film that has a budget of $25 million it may have a gross profit between approximately $20million and $230million. When we look at a film with a larger budget of $125million it may have a gross profit between approximately $250million and $450million. We can also see 2 outliers. The first is the movie ET (7) this has a set budget of $10.5 million which is considerably low compared to the gross profit it made which was $434,949 million. The second unusual point is the film Star Wars (33) which also had a considerably low budget of $11 million and also made a large gross profit of $460,936 million.

Since there is approximately a linear trend I will add a regression line.



Linear Trend

Gross.Profit...millions.of.US.dollars. = 2.1412 * Budget..millions.of.US.dollars. + 102.74

Correlation = 0.68369

Guinea pig scripts for Bi-Variate presentation Workshop PN Auckland University Statistics day 2014

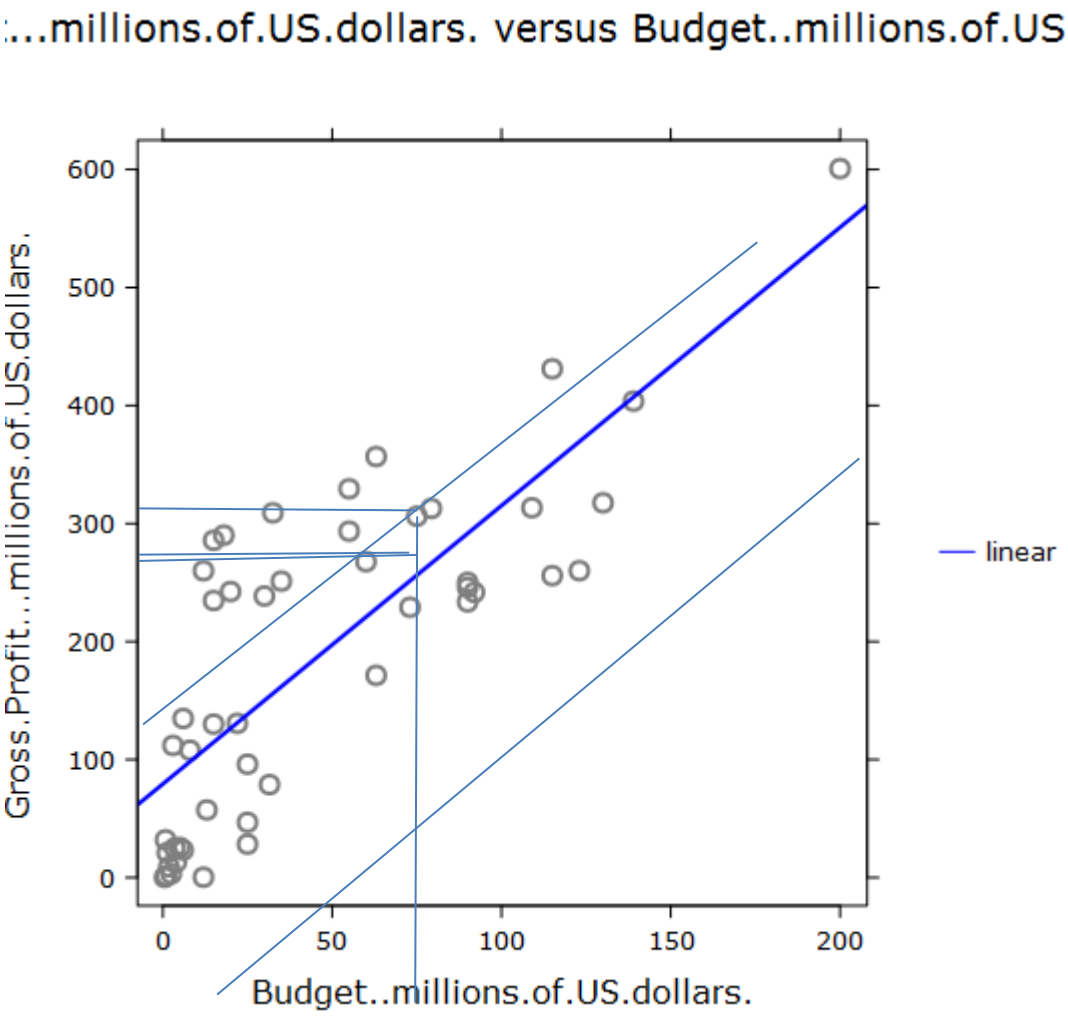
Sample size: 50.

The correlation co-efficient is approximately r=0.68, this is not close to r=1 which backs up my statement that this relationship is moderate as there is a reasonable amount of scatter against the regression line.

2.14*75+102.7= $236.7 million [$120 million, $380million]

A movie that has a budget of $75 million has a predicted gross profit of $236.7 million dollars However as there is a large amount of scatter about the regression line there is a large prediction interval. A movie which has a budget of $25 million will make a gross profit of somewhere between $120 million and $380 million

<u>Removing outliers</u>:



Linear Trend

Gross.Profit...millions.of.US.dollars. = 2.3582 * Budget..millions.of.US.dollars. + 79.34

Correlation = 0.79565

Sample size: 48

I chose to remove two outliers. These were (row 7 and row 33) Star wars and ET both these films had a low budget but made a large gross profit. This could be because of several reasons such as the time both movies were made.

When we remove the outliers Star wars and ET (row 33 and 7) we see that the correlation co-efficient has changed, this is now 0.79 this is closer to r=1 than the old correlation co-efficient before we removed the outliers which was 0.68. This shows us that those two outliers were affecting the place of the regression line. Since 0.79 is closer to r=0.1 it shows us that when these outliers are removed the relationship between the budget of a film and the gross profit becomes stronger.

2.35*75+79.34= $255.6 million

For a film with a budget of $75 million we predict that it will make a gross profit of $255.6 million.

Since there is a reasonable amount of scatter about the regression line the prediction interval will be large. For a movie with a budget of $75 million it is likely to make a gross profit of somewhere between $110 million and $390 million.

Conclusion

To answer my question "I wonder if there is a relationship between the budget of a movie and the gross profit that the movie makes" we see that movies with a smaller budget do tend to make a smaller profit than movies with a larger budget. However there was a moderate amount of scatter about the regression line, which showed a moderate relationship between the budget of a movie and the gross profit. Since the relationship was only moderate it is hard to say that it is true that movies with a larger budget will make a larger gross profit than those with a smaller budget. Therefore I think that there are several factors that can influence the success and gross profit a movie gains other than the budget that was set for the film. For example the actors who are in the film play a big part in the success as more people are likely to watch it if the actors are well known. The second factor that I think will influence the gross profit and success of a movie is the release date. Finally I also think that the amount of advertising which is done for the film also contributes to how successful and how much gross profit is made. Therefore the relationship between the budget of a film and the gross profit made is moderate.

Key omission in conclusion initially: no comparison of models with and without the outliers-no discussion as to which prediction was the more appropriate in this situation.

(initially given A but gained M after resubmission)