**Homework 01**

**PRINCIPAL COMPONENT ANALYSIS (PCA)**

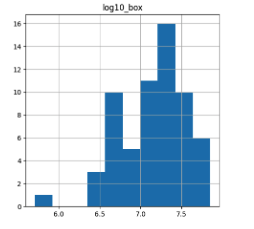
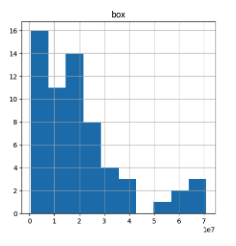
**BUAN 6383.001**

**Syam Menon**

Mihir Hirave (MDH230000)

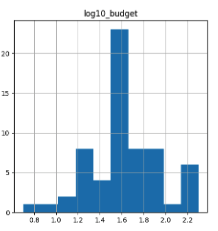
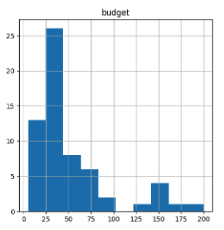
1. **Plot histograms of the continuous variables (box, budget, starpwr) to see if any transformations are needed. Are any of them skewed? Apply a log-transformation to all the skewed variables.**

**The histogram of box variable.**



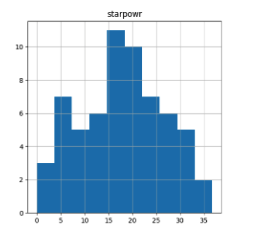
The histogram for the box variable looks right skewed (on the left). Therefore, we apply log transformation on the variable and the new histogram is on the right.

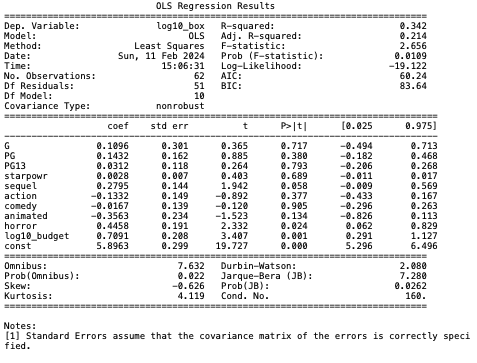
**The histogram of budget variable.**



The histogram for the budget variable looks right skewed (on the left). Therefore, we apply log transformation on the variable and the new histogram is on the right.

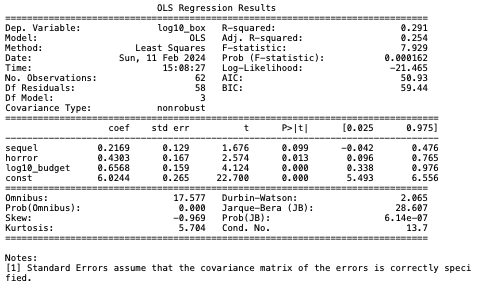
**The histogram of starpowr**



1. **Run a linear regression of box office revenues on the “traditional” variables (i.e., using all the independent variables (except the “buzz” variables). If any variables were transformed, be sure to use the transformed versions of those variables. What are the *R*2 and adjusted- *R*2 values? Which variables (if any) are significant at the 0.10 level, based on the t-statistics and associated probabilities (p > |t|)?**

R-squared is 0.342 and Adjusted R-squared is 0.214. Based on t-statistics and associated probabilities (p > |t|) with significance level at 0.1, budget, sequel, and horror are significant.

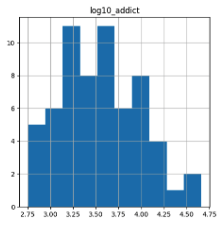
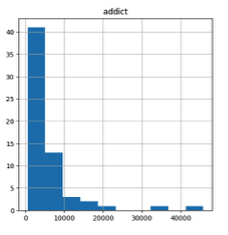
1. **Run another linear regression using only the variables that were significant (again, ignoring the “buzz” variables). What are the *R*2 and adjusted- *R*2 values? Are all the variables still significant at the 0.10 level?**



Running regression with budget, sequel, and horror variables. R-squared is 0.291 and Adjusted R-squared is 0.254. Based on t-statistics and associated probabilities (p > |t|) with significance level at 0.1, budget, sequel, and horror are all still significant. Here, the R-squared decreased but adjusted R-squared increased.

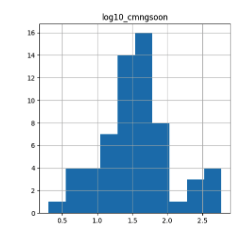
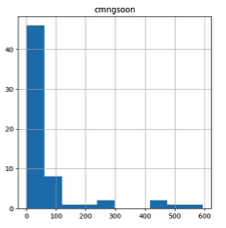
1. **Plot histograms of the four “buzz” variables. Are any of them skewed? Apply a log-transformation to all the skewed variables.**

**The histogram of addict variable.**



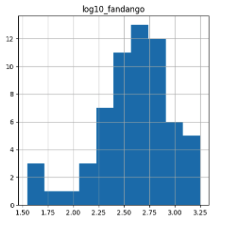
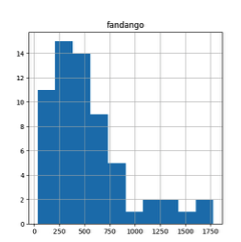
The histogram for the box variable looks right skewed (on the left). Therefore, we apply log transformation on the variable and the new histogram is on the right.

**The histogram of cmngsoon variable.**



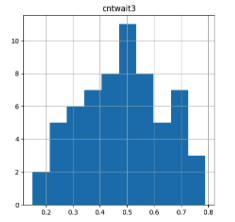
The histogram for the box variable looks right skewed (on the left). Therefore, we apply log transformation on the variable and the new histogram is on the right.

**The histogram of fandango variable.**

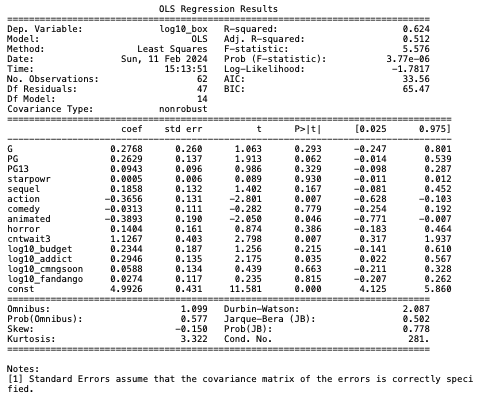


The histogram for the box variable looks right skewed (on the left). Therefore, we apply log transformation on the variable and the new histogram is on the right.

**The histogram of cntwait3 variable.**

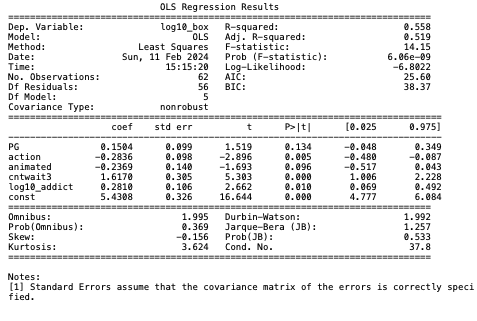


1. **Run a linear regression of box office revenues on *all* the independent variables, including the “buzz” variables (transformed as needed). What are the *R*2 and adjusted- *R*2 values? Which variables (if any) are significant at the 0.10 level, based on the t-statistics and associated probabilities (p > |t|)?**



R-squared is 0.624 and Adjusted R-squared is 0.512. Based on t-statistics and associated probabilities (p > |t|) with significance level at 0.1, PG, action, animated, addict, and cntwait3 are significant.

1. **Run another linear regression using only the variables that were significant. What are the *R*2 and adjusted- *R*2 values? Are all the variables still significant at the 0.10 level?**



R-squared is 0.558 and Adjusted R-squared is 0.519. Based on t-statistics and associated probabilities (p > |t|) with significance level at 0.1, action, animated, addict, and cntwait3 are significant, but PG is not significant anymore.

1. **Compare the models developed so far – which of these would you choose, and why?**

We would choose the last model built in Q6 because it has the highest Adj. R-squared of 0.519 suggesting that a better fit to the data. This model is also simpler as it has only 5 variables. Although the model in Q5 has a higher R-squared, indicating overfitting.

1. **Apply Principal Component Analysis to just the 4 “buzz” variables. If you transformed any of them, make sure you use the transformed versions. Also make sure that you standardize the variables first. What are the eigen values associated with each component? How many principal components are selected using (i) Kaiser's Rule, and using “explained variance" thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90%?**

Eigen values associated with

PC1: 2.41420026

PC2: 0.77519959

PC3: 0.45214886

PC4: 0.3584513

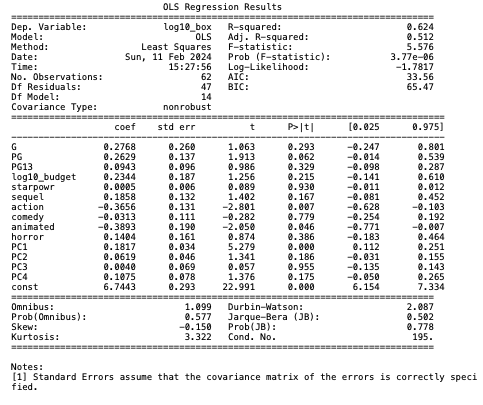


Number of principal components using

1. Kaiser’s Rule: 1 (PC1, as it is greater than 1)

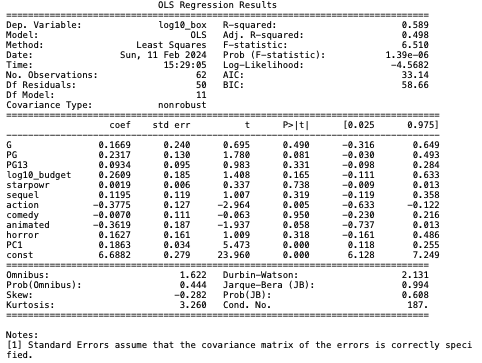
Explained variance thresholds of

1. 60%: 1 (PC1)
2. 70%: 2 (PC1, PC2)
3. 80%: 3 (PC1, PC2, PC3)
4. 90%: 3 (PC1, PC2, PC3)
5. **Run a linear regression using all the “traditional” independent variables (if transformed, use the transformed versions) *and* all 4 principal components (the only variables you should not use here are the four “buzz” variables). What are the *R*2 and adjusted- *R*2 values? Which variables (if any) are significant at the 0.10 level?** **In particular, are any of the principal components significant? What can you say about this model vis-à-vis the other models built so far?**

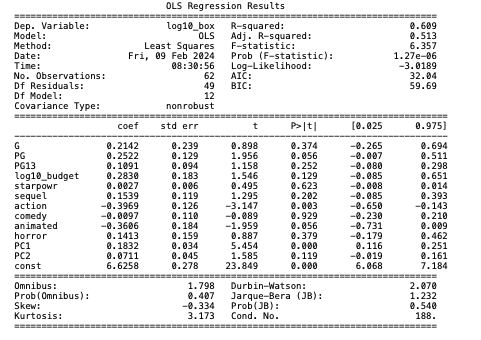


R-squared is 0.624 and Adjusted R-squared is 0.512. Based on t-statistics and associated probabilities (p > |t|) with significance level at 0.1, PG, action, animated, and PC1 are significant. *This model is* *similar to model built in Q5, regression on all variables including buzz variable.*

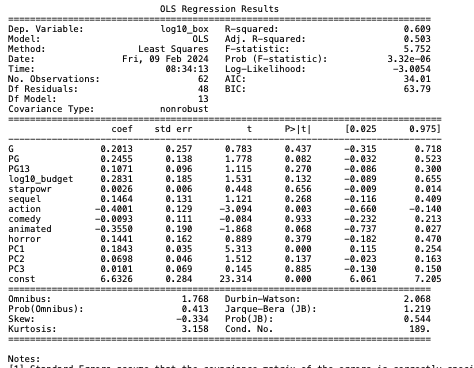
1. **Now run regressions using the number of principal components based on (i) Kaiser's Rule and “explained variance” thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90% (if any of the models are identical, point this out and run it only once). Compare all the regression models involving the principal components (including the one involving all four components). Which of these would you recommend, and why?**
2. Based on Kaiser’s Rule, we only take the principal components whose eigen value is greater than 1, therefore we run regression using PC1 and dropping the other Principal components. The result of regression summary is below.



1. The above model built in (i) satisfies the “explained variance” threshold of 60%, as it only selects PC1 as well.
2. Using the “explained variance” threshold of 70%, we select PC1 and PC2, dropping PC3 and PC4. The result of regression summary is below.



1. Using the “explained variance” threshold of 80%, we select PC1, PC2, and PC3, dropping PC4. The result of regression summary is below.



1. The above model built in (iv) satisfies the “explained variance” threshold of 90%, as it also selects PC1, PC2, and PC3.

After comparing all the models, we would recommend the model built in (iii) with “explained variance” threshold of 70%, because it has the highest Adjusted R-squared out of all models.

1. **Now apply Principal Component Analysis to the 4 “buzz” variables *and* the other continuous variables (budget and starpowr). Again, use transformed versions of the variables if any were transformed and standardize the variables first. What are the eigen values associated with each component? How many principal components are selected using (i) Kaiser's Rule, and using “explained variance" thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90%?**

PC1: 2.83823382

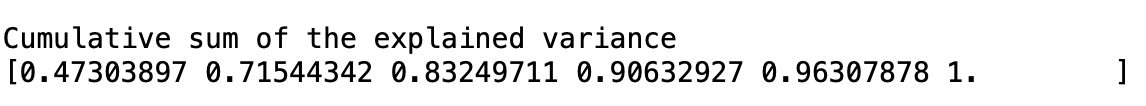
PC2: 1.45442671

PC3: 0.70232212

PC4: 0.44299297

PC5: 0.34049709

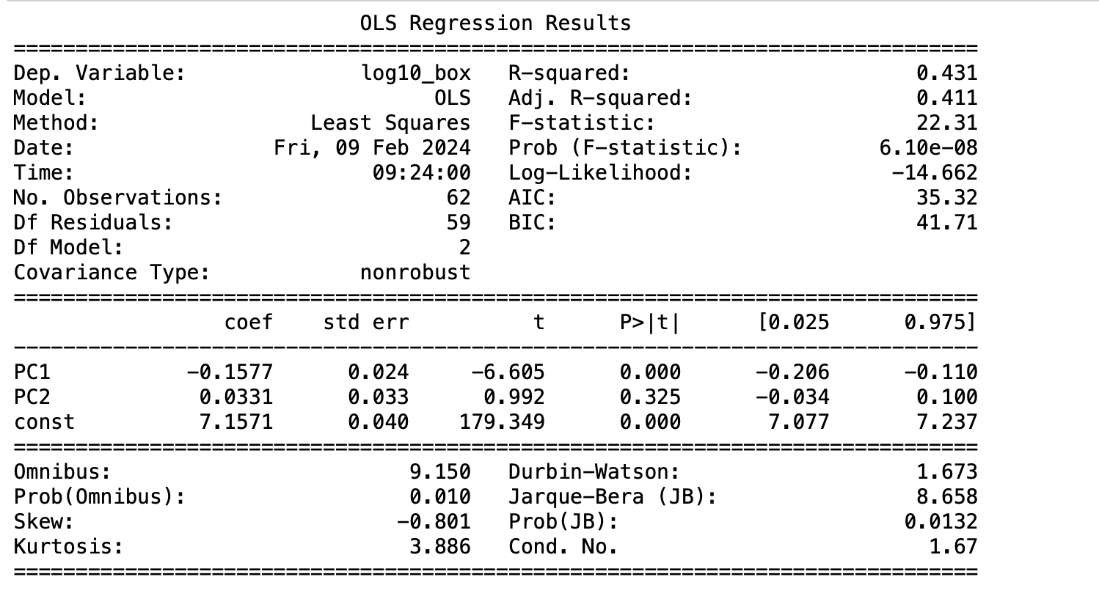
PC6: 0.2215273

Number of principal components using

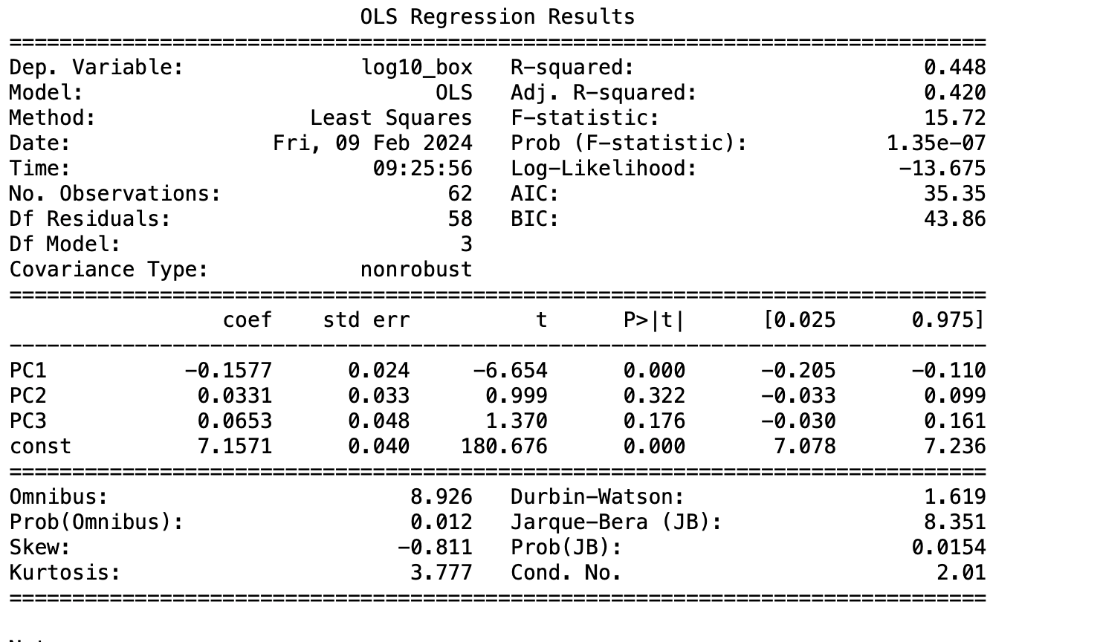
1. Kaiser’s Rule: 2 (PC1 and PC2, as it is greater than 1)

Explained variance thresholds of

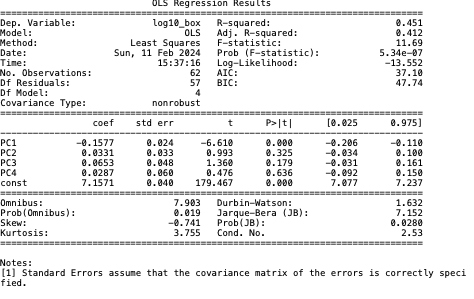
1. 60%: 2 (PC1 and PC2)
2. 70%: 2 (PC1, and PC2)
3. 80%: 3 (PC1, PC2, and PC3)
4. 90%: 4 (PC1, PC2, PC3, and PC4)
5. **Next, run regressions using the number of principal components based on (i) Kaiser's Rule and “explained variance” thresholds of (ii) 60%, (iii) 70%, (iv) 80% and (v) 90%. Compare these regression models and explain which one you would recommend, and why?**
6. Based on Kaiser’s Rule, we only take the principal components whose eigen value is greater than 1, therefore we run regression using PC1 and PC2 and dropping the other Principal components. The result of regression summary is below.



1. The above model built in (i) satisfies the “explained variance” threshold of 60%, as it also selects PC1 and PC2.
2. The above model built in (i) satisfies the “explained variance” threshold of 70%, as it also selects PC1 and PC2.
3. Using the “explained variance” threshold of 80%, we select PC1, PC2, and PC3, dropping PC4, PC5, and PC6. The result of regression summary is below.



1. Using the “explained variance” threshold of 90%, we select PC1, PC2, PC3, and PC4, dropping PC5, and PC6. The result of regression summary is below.



We would use the model built in (iv), using the “explained variance” threshold of 80%, because it has the highest Adjusted R-square among other regression models.

1. **Are the “buzz” variables helping build a better model? How about PCA?**

By comparing the R-squared and the adjusted R-squared of the models with or without buzz variables, buzz variables help explain a great amount of variability. Thus, buzz variables help develop a better model.

The primary use of PCA is to reduce the dimensionality of our data. When we compare the R-squared and the adjusted R-squared of models without PCA and the best PCA models, we find that the values of the two models do not differ much. However, PCA models are simpler. Thus, PCA helps in building a better model.

1. **Did you learn anything surprising while doing these analyses? Can you provide some managerial takeaways?**
   1. In Question 5, we see that the independent variable PG was deemed significant when it was compared against all other independent variables. However, when we compared only to other significant variables, in Question 6, PG was no longer deemed significant. This indicates that other independent variables play some role in the relationship between one another. This could be because independent variables can capture some of the same information as one another and cause other variables to become insignificant.
   2. It is surprising that the budget in and of itself was not statistically significant (in models involving buzz variables), while we thought it would play a significant role in how a movie performed in terms of opening weekend box office collection. Achieving commercial success does not necessarily require a large budget. Rather, we need to consider other factors such as story, content quality and audience engagement.
   3. We observed that budget and sequel were statistically significant in models which did not include the buzz variables. As soon as buzz variables were introduced, we observed that the statistical significance of those two variables reduced. Thus, our key takeaway is that while production budget and sequel do impact opening weekend box office collection, their significance is shallow when compared to the significance of buzz surrounding a movie.
   4. It is surprising that ‘starpowr’ is not statistically significant as one might expect. The reason could be while huge movies stars help with generating buzz, the number of high-profile actors in a film does not contribute to the extent of domestic opening weekend collection. Thus, the key takeaway would be that casting more high-profile actors does not also lead to proportional gains in collection. In fact, casting more well-known actors will end up increasing the cost of the movie. So, hire fewer high-profile actors. This is also seen in the 'starpowr' histogram, where the 'starpowr' is mostly between 15-20. This would mean that movies tend to cast around 1-2 well-known actors.
   5. One of the reasons “buzz variables” are contributing to box could be because box is domestic opening weekend box office revenues ($) of a movie. Opening weekends are dictated by online chatter about a movie. Therefore, creating a thought-out marketing strategy to create buzz could help to increase the box office collection. Utilizing creative campaigns to engage with audiences can contribute to high box office success, as demonstrated by the commercial success of the movie ‘Barbie’.
   6. As a manager, I would suggest the production companies to invest in movie genres such as action and animation and try to aim for a PG rating as this combination is more likely to receive higher returns.