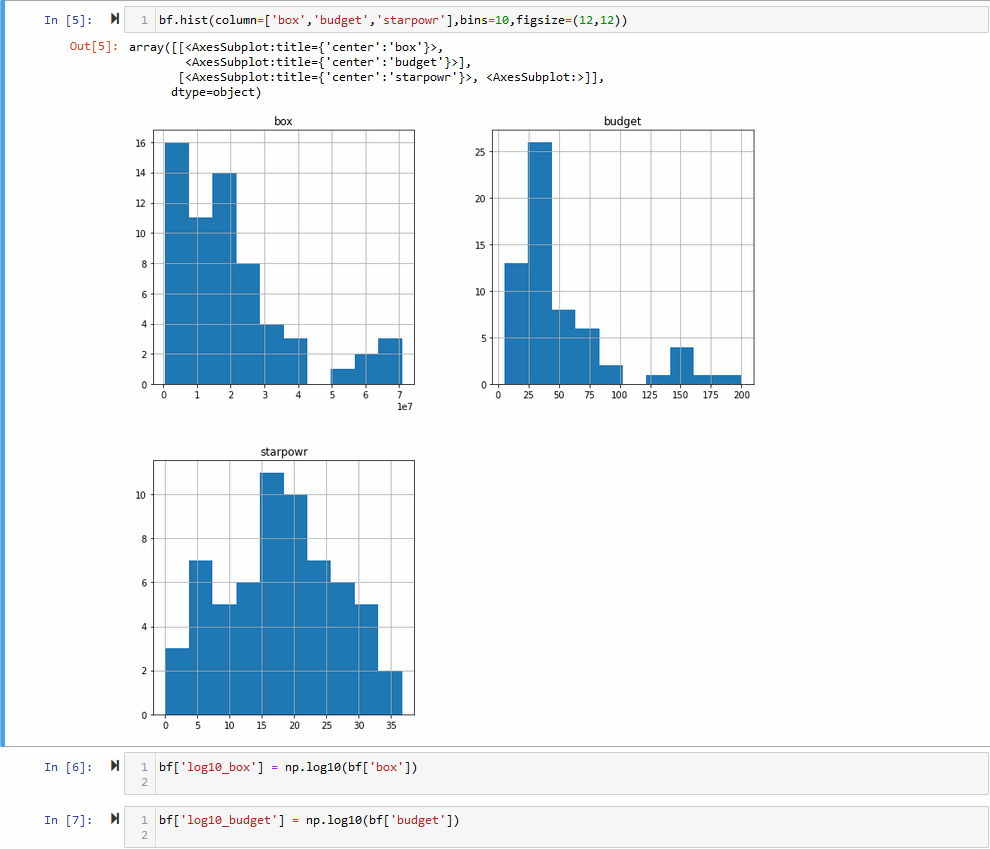
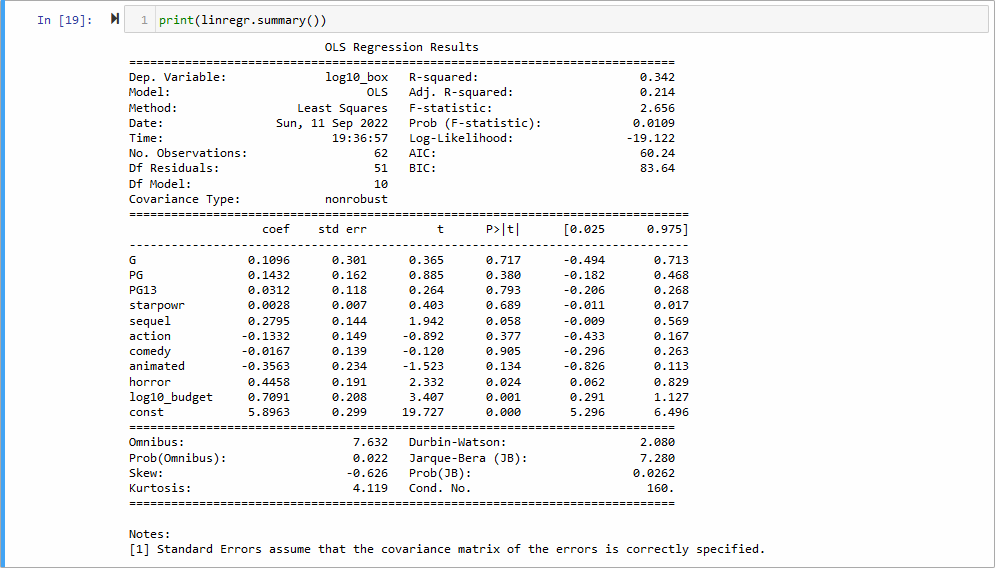
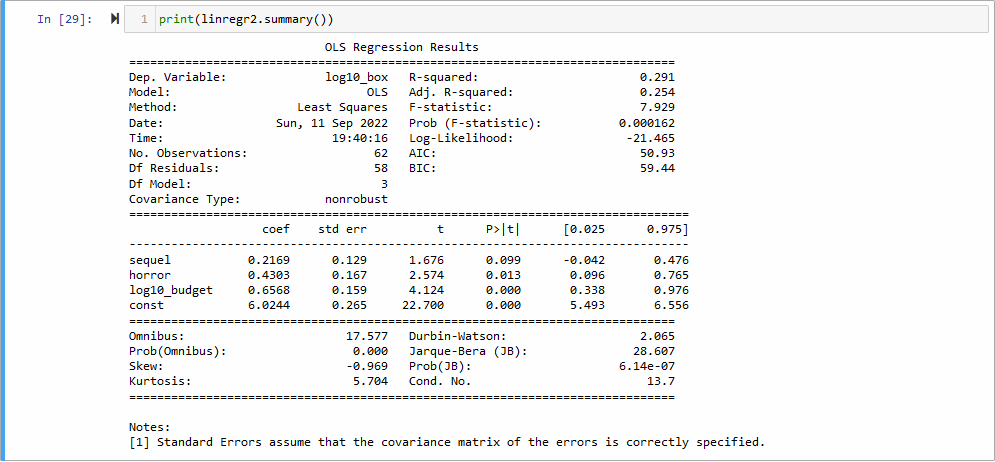
1: Plotted histograms. Box and Budget appear skewed, so applied log-transformation for those variables.

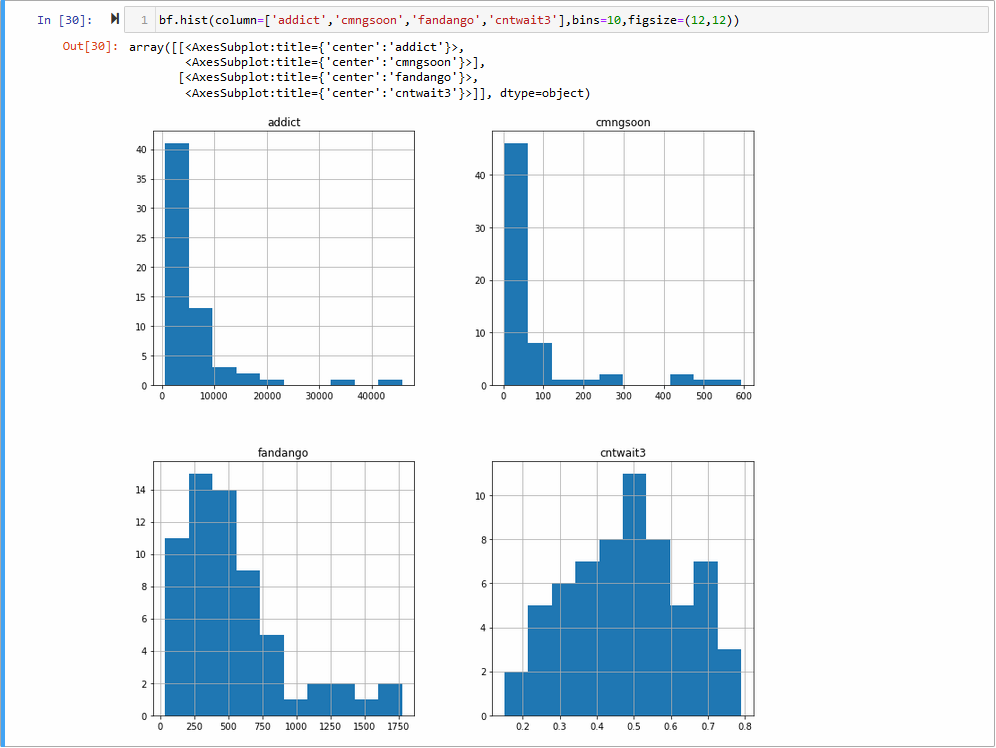


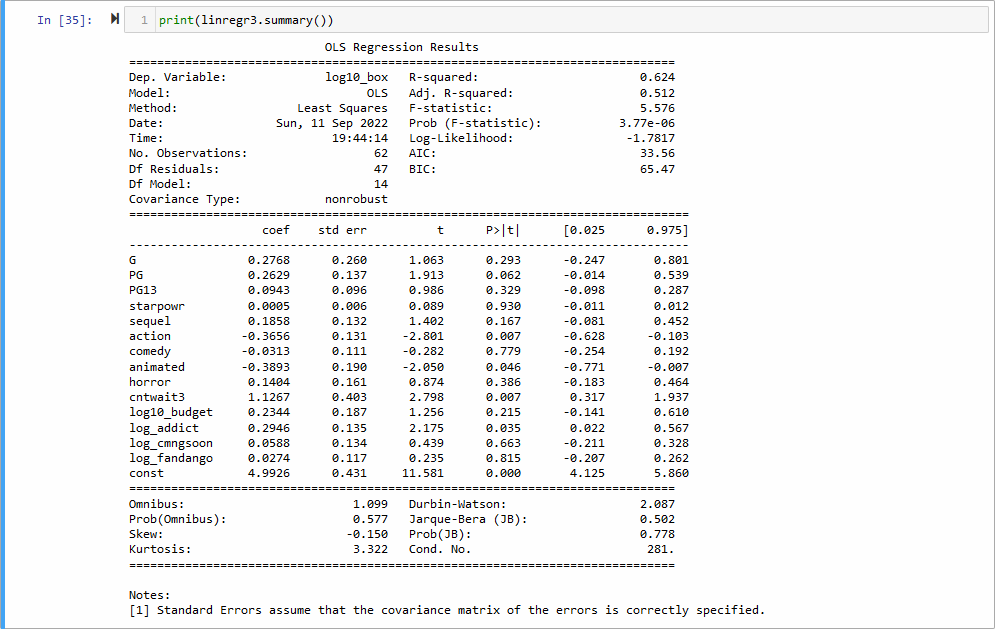
2: Ran linear regression. R-squared of .342, Adjusted R-Squared of .214. Model is both a poor predictor of variation in the (log of) box office, and is inefficient given the amount of information provided. Only the sequel indicator, horror indicator, (log of) budget, and constant were significant.

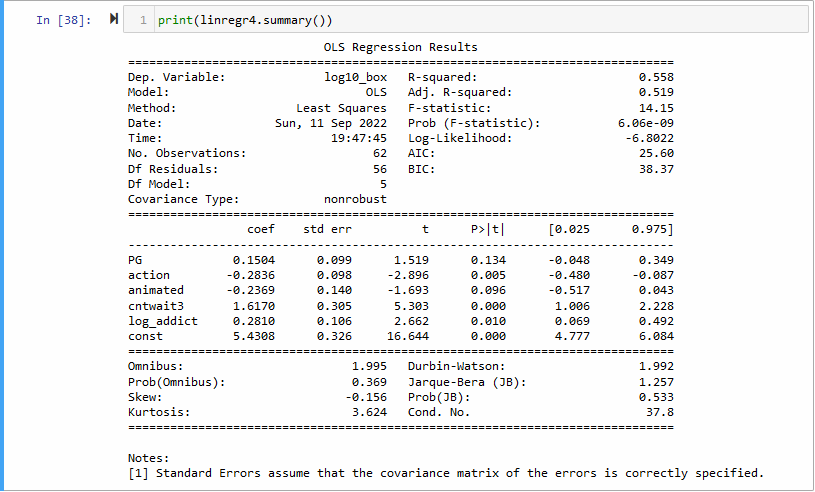


3: Ran linear regression. R-squared is .291, adjusted is .254. Model becomes worse at predicting box office returns, but is more efficient given the amount of variables provided. All variables remain constant at the 10% level.

4: Only cntwait3 was un-skewed. The scores for addict, coming soon, and fandango all displayed heavy right skew. Preformed logistic transformation on all buzz variables except the can’t wait score.



5: Ran Regression. R squared of .624, adjusted of .512. Model is far better than its predecessors. PG indicator, action indicator, animated indicator, as well as the buzz variables of cntwait3 and (log of) addict are all significant.

6: Ran regression. R-squared of .557, adjusted of .519. PG is no longer significant, but action, cntwait3, and log of addict all increase in significance.

7: Of the models so far, we would choose the one with all of the variables and buzz variables (regression 3, in question 5). While the last model has a slightly higher adjusted R-squared, the model with all of the variables has a much higher raw R-squared. In addition, we believe variables such as budget or starpower are too valuable to ignore, and their lack of significance may hint at a collinearity or other issue in the model. Excluding them may risk more missed information, despite lack of statistical power.

8: The eigen values are 2.4142, .7752, .1452, and .3584. This results in an explained variance of 60.36% for the first component, 19.38% for the second, 11.30% for the third, and 8.96% for the fourth. Via Kaiser’s rule, we would only use the first principal component. The explained variance thresholds would result in the following:

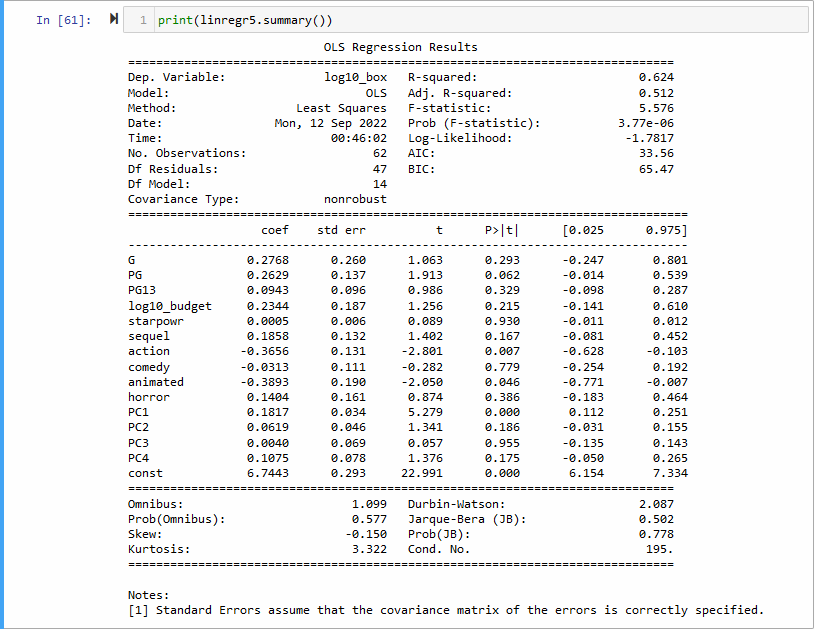
60% - use only the first component

70% - use only the first two components

80% - use the first three components

90% - use the first three components

9: This model has an R-squared of .624 and adjusted of .512. The significant variables are the PG, action, and animated indicator, the constant, and the first principal component. This model performs identically to the model we chose above – with all of the variables and buzz variables compared to the log of box office returns. This makes sense, as we have excluded none of the principal components/variance.

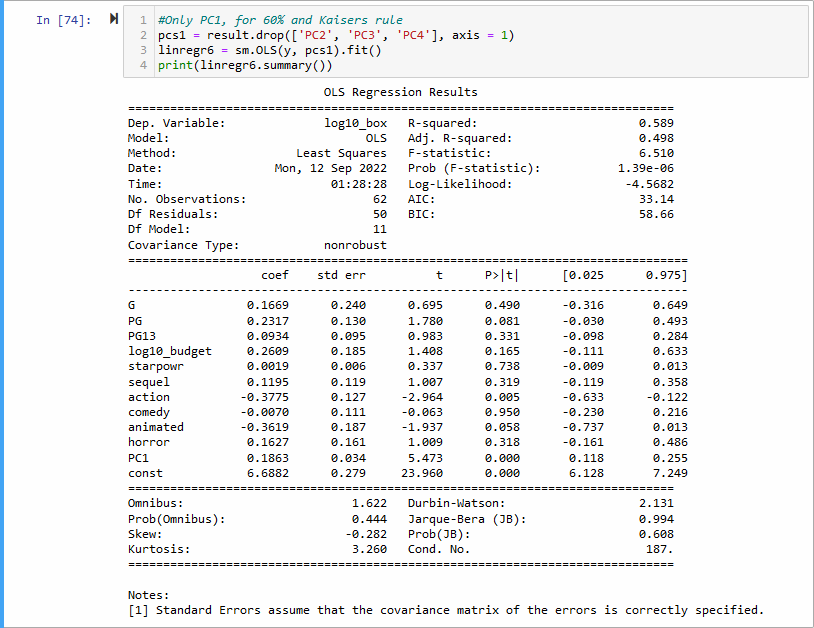


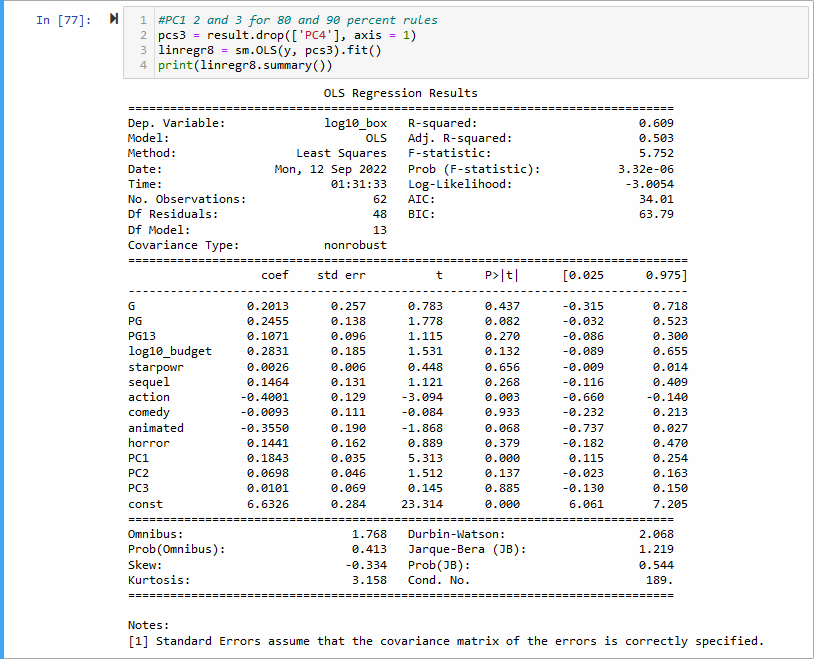
10: There will be three regressions run. The first will only have the first principal component, as per Kaiser’s rule and the explained variance threshold of 60%. The second will have the first two components, to explain 70% of the variance, and the last one will have the first three, to fulfill the 80% and 90% rule.

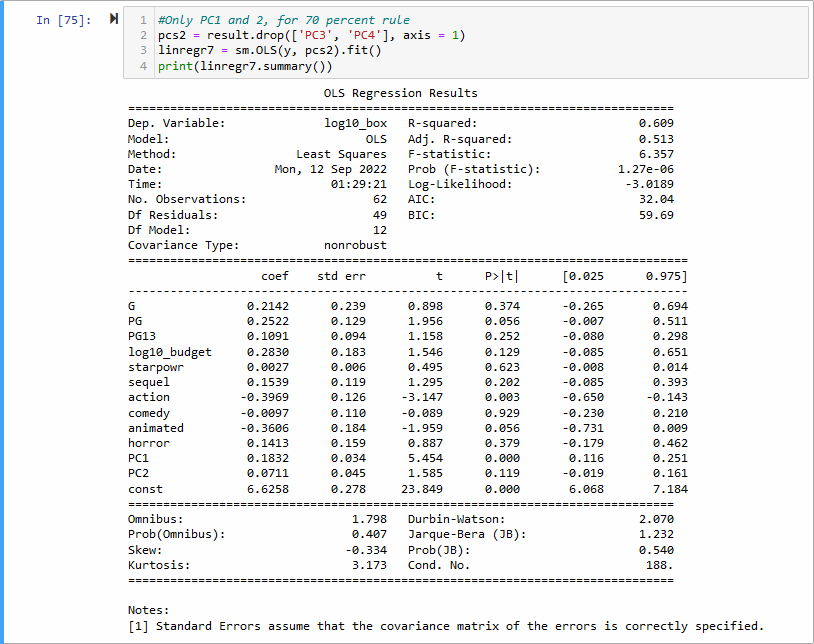
The regression outputs are below. We would likely use the one with the first two principal components. The first regression of this problem, with only PC one, was less efficient with a lower R-squared and adjusted R-squared than the full model. However, the R-squared dropped by 3.5%, while the adjusted only dropped by 1.4%, implying some improvement in relative efficiency.

The second model performs slightly worse than the original with all Principal Components – with the R-squared dropping 1.5%. However, the adjusted R-squared does increase by .1%, which is a small but noticeable improvement, and is more impressive relative to the non-adjusted R-squared. In addition, the model with the first two PC’s has the highest significance for the budget indicator – while it has a P-value of .129, which is not statistically significant, this improvement indicates that this model may suffer from less collinearity or other issues relative to the other models.

The final model with the first three principal components keeps the same R-squared as the second model (.609), but the adjusted R-squared drops a full percentage point, implying there is effectively no benefit to adding the third principal component to the model.







11: The eigen values of these 6 components are 2.838, 1.454, .7023, .4430, .3405, and .2215, respectively. The first two would be selected with Kaisers rule. With the explained variance thresholds, we would select:

60% and 70% - the first two PC’s (.473 and .242, total .7154) (same as kaisers)

80% - the first three (additional .117, for .8324)

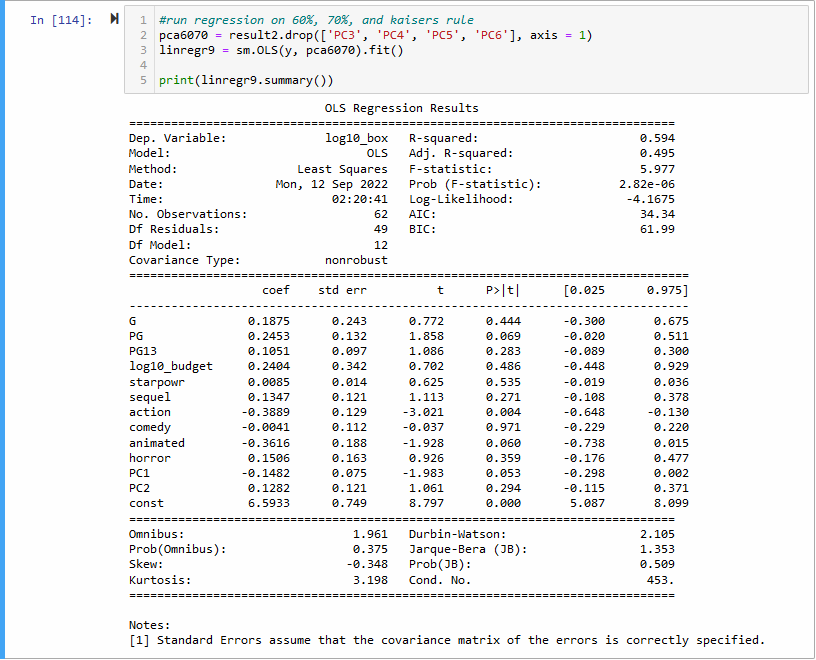
90% - the first four (additional .0738, for .9063)

Of these regressions we would likely take, again, the second one, following the 80% of variance rule. The first one (60, 70, Kaisers), produces a model with reasonable results, but lackluster compared to some of the other models. With an R-squared below .5, we would rather take any of the 5 models before this PCA that would produce better results.

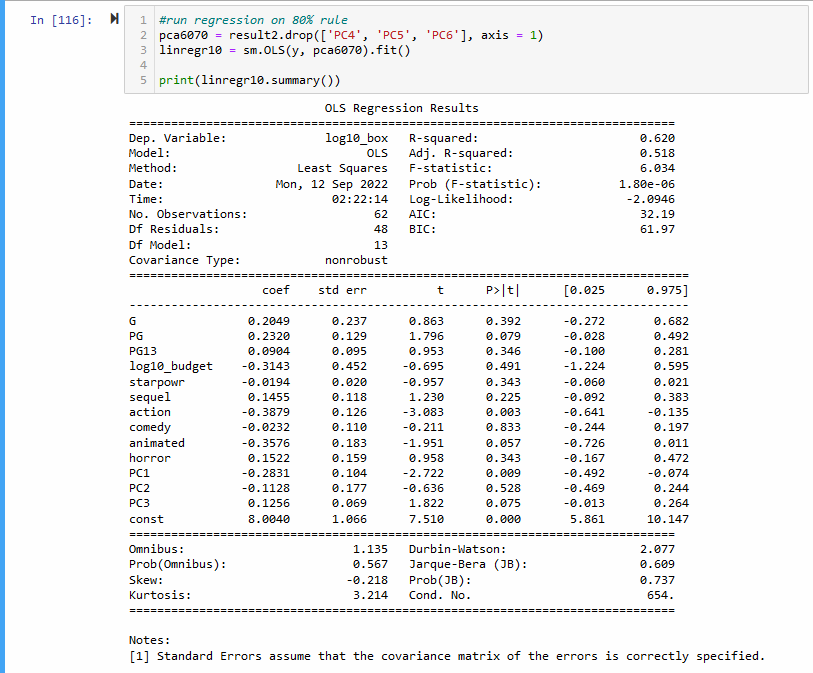
The second model (80%) is far more powerful. It is the closest model we have gotten so far to our “complete” model from question 5, in R-squared (.620, compared to .624 from earlier), and is higher in adjusted R-squared (.518, compared to .512). Due to having a higher adjusted R-squared and condensed data, this model should be comparably reliable while being easier to scale.

The last model (90%) is practically a mirror of the model from question 5, with the same R and R-squared. However, due to PCA, it is harder to interpret than it’s predecessor, and with no advantages visa vis significance of variables, there is no reason to choose it.

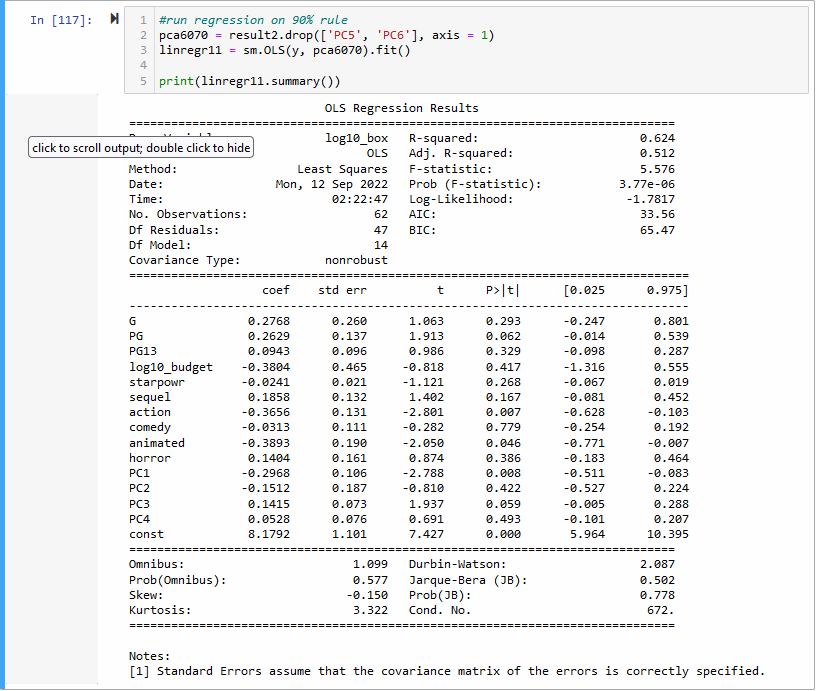
60, 70, Kaisers:



80% rule



90% Rule



13: The “buzz” variables proved to be incredibly helpful in building a better model. They allowed us to actually reach a measure of reliability visa vis our R and R-squared variables. However, PCA was only moderately helpful. While it did produce comparable models – specifically at the 80% of variance level – we were working with a minimal amount of data – only ever compressing ~6 columns. This showcased the power of PCA when working with much larger amounts of variables, but when it comes to editing regressions, applying econometric theory to define interaction variables seems like it could be more valuable.

14: I’m surprised at how little “budget” was considered statistically significant, and would argue that it likely is a problem with data, rather than real-world significance. It was also interesting to note how the dummy variables interacted – for example, PG and action movies remaining significant, while many of the other classifiers were not.

As a managerial takeaway, it appears as if a producer should strive to increase their “can’t wait” and “addict” scores. They can do this by developing good/better trailers, which should drive up trailer views (addict) and can’t wait votes (cntwait3). It does not seem to be worth drumming up general discussion about the movie – the fandango and cmngsoon scores are less impactful – but rather garnering passionate fans. I would also recommend creating a PG-rated, non-animated, non-action movie. PG rated movies typically receive higher returns, while animated and action movies are statistically lower-return genres/styles across most of the models.