## **First Week Revenue (F)**

The conventional wisdom in the industry is that the likely success of a film is gauged by how well it does in the first week after release. We now look at the likely causal relationships of different features to the revenue earned by a film in its first week after release. To avoid clutter we draw simpler DAGs that model the likely causal relationship between F and a feature and will include other features that were seen to have causal relationship, in the analysis earlier, with the feature under investigation.

The weekly revenue data available is inclusive of entertainment and service tax. To remove the impact of taxes we adjust the weekly revenue data by the ratio as Total-Nett-Gross/Total-Gross for the film.

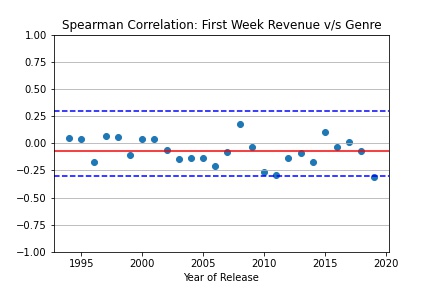
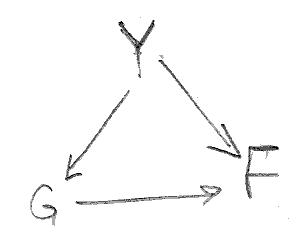
**Release Week**

The correlation between W and F can be easily ascertained using the Spearman Rank correlation.

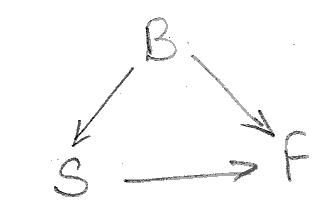
A Spearman Rank correlation value of 0.0642 indicates that there is likely **no direct causal path from W to F**.

**Genre**

Y is a confounder for the effect of G on F. We stratify the data by Y; this is effectively conditioning on Y and closes the backdoor path between F and G. We use the Spearman rank correlation to test the hypothesis that there is a direct causal path from G to F.

For all years the correlation coefficient stays within [-0.33, 0.33] leading to the conclusion that thre is likely **no direct causal path from G to F**.

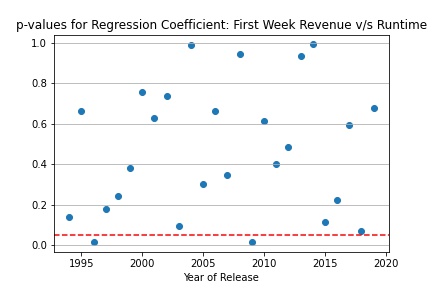
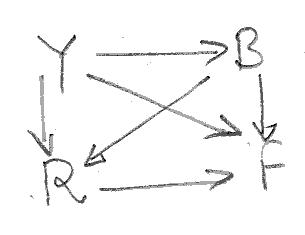
**Release Screens**

B is a confounder for the effect of S on F. To condition on B, we set up a linear regression model, regressing F on S and D. We ascertain the p-value for S to determine the likely causal path from S to F.

The p-value for the coefficient of S is less than 0.05 implying that S has a significant association with F and there is a **direct causal path from S to F**.

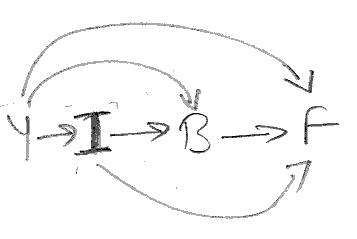
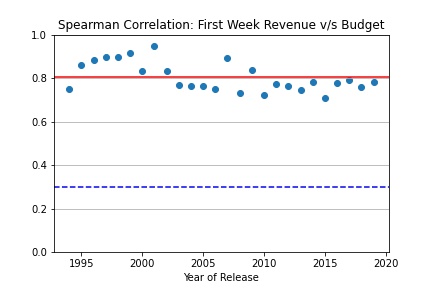
**Runtime**

Both B and Y are confounders to the effect of R on F. To condition on B and Y, we set up yearly regression models for each year, regressing F on R and B and then evaluate the correlation between F and R by examining the p-values of the regression coefficients of R.

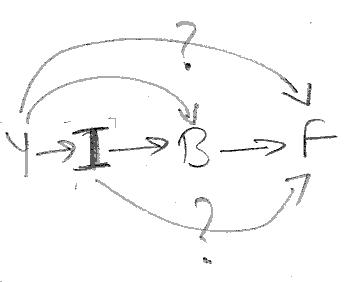
For all years except two, the p-value of the regression coefficient of R is greater than 0.05, leading to a conclusion that R is not significantly related to F and there is likely **no direct causal path from R to F**.

**Budget**

Release Year and Inflation are confounder to the effect of B on F. We stratify on Y; this is effectively conditioning on Y and closes one of the backdoor paths between F and B. However it is important to realize that within a year I has the same value and therefore to close the second backdoor needs no further conditioning. So in effect stratifying on Y conditiones on both Y and I and closes all the backdoor paths from F to B. We use the Spearman rank correlation to test the hypothesis that there is a direct causal path from B to F.

The average Spearman rank correlation across years is 0.8049 and for all years the value stays well above 0.6. We can therefore conclude that there is **direct causal path from B to F.**

**Release Year**

There are 4 possible causal paths from Y to F: Y -> F, Y -> B -> F, Y -> I -> F and Y -> I -> B -> F.

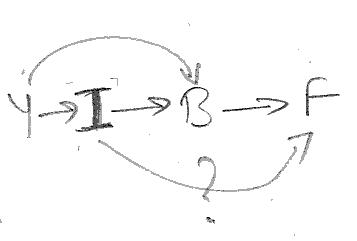
Conditioning on I closes the latter 2 paths. We achieve this by adjusting the values of both B and F for inflation. If we calculate the total effect of Y on inflation adjusted F, then we get the combined effect of the first two paths: Y -> F and Y -> B -> F, where Y -> B represents the direct effect of Y on B (not mediated by I).

The total effect of Y on inflation adjusted F is 0.5864. We know from earlier analysis that the direct effect of Y on B is 0.5111.

The Spearman Rank correlation between B and F represents the total effect of B on F and is 0.9229.

We can therefore calculate the direct effect of Y on F by manipulating the following equation:

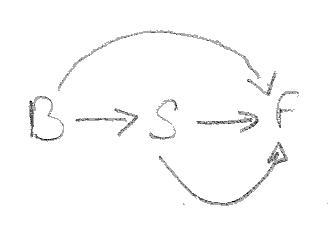
0.5864 = x + 0.5111\*0.9229, i.e. x = 0.1147. This value of x leads to the conclusion that there is likely **no direct causal path from Y to F**.

The earlier DAG now reduces to the DAG presented alongside. The Spearman Rank correlation between Y and F represents the total effect of Y on F and is 0.8165. From earlier analysis we know that Spearman Rank correlation between I and B is 0.2922.

We can therefore calculate the direct effect of I on F by manipulating the following equation:

0.8165 = 0.5111\*0.9229 + 1\* 0.2922\*0.9229 + x, i.e. x = 0.0751. This value of x leads to the conclusion that there is **no direct causal path from I to F**.

**Summary**



In summary, we conclude that the **first week revenue of a film is influenced by:**

* **the budget of the film, and**
* **the number of screens that is opens to**

The DAG for this model is presented alongside.

We update the summary table to reflect these causal relationships

|  |  |
| --- | --- |
| **Feature** | **Influenced/Affected By** |
| Release Year |  |
| Inflation | Release Year |
| Genre | Release Year |
| Budget | Release Year, Inflation |
| Runtime | Release Year, Budget |
| Release Week |  |
| Release Screens | Budget |
| First Week Revenue | Budget, Release Screens |

**Prediction Model Summary**

We fitted a RandomForest Regression Model to a training set and evaluated its performance against a test set, both sets drawn from the data available. The performance of the model:

* Percentage of estimates for test set that are off by less than 25% from true value: 44.92
* Percentage of estimates for test set that are off by less than 35% from true value: 60.00
* Percentage of estimates for test set that are off by less than 45% from true value: 69.85
* Percentage of estimates for test set that are off by less than 55% from true value: 75.69