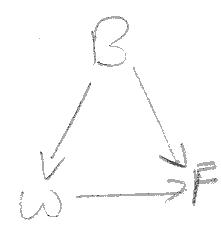
## **First Week Revenue (F)**

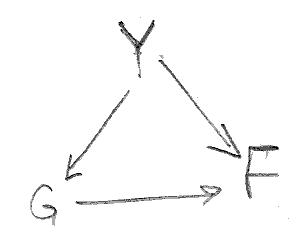
The conventional wisdom in the industry is that the likely success of a film is gauged by how well it does in the first week after release. We now look at the likely causal relationships of different features to the revenue earned by a film in its first week after release. To avoid clutter we draw simpler 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**

B is a confounder for the effect of W on F. To ascertain the causal effect of W on F we set up a regression model, regressing F over B and W.

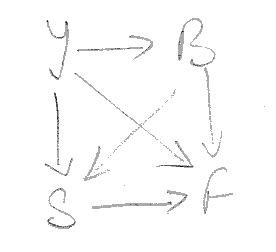
The p-value of the regression coefficient for W is 0.806 and we conclude that there is likely **no direct causal path from W to F**.

**Genre**

Y is a confounder for the effect of G on F. To condition on Y, we set up a regression model, regressing F over G and F. Since G is a categorical variable we use dummy variables to one-hot encode G. All variables in the regression model are standardized. We ascertain the regression coefficient and the p-value for G to determine the likely causal path from G to F.

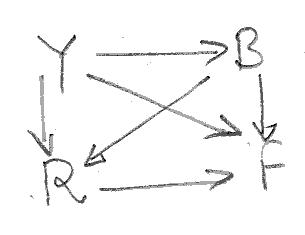
For all genre categories, the p-value of the regression coefficient is significantly greater than 0.05 and we conclude that there is likely **no direct causal path from G to F**.

**Release Screens**

Y and B are confounder for the effect of S on F. To condition on Y ad B, we set up a linear regression model, regressing F on S, B and Y. All variables in the regression model are standardized. We ascertain the regression coefficient and the p-value for S to determine the likely causal path from S to F.

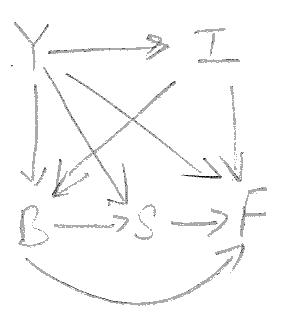
The regression coefficeint for S is 0.7567 and its p-value is less than 0.05 implying that S has a significant association with F and there is a **direct causal path from S to F**.

**Runtime**

Both B and Y are confounders to the effect of R on F. To condition on B and Y, we set up a regression models, regressing F on Y, R and B and then evaluate the correlation between F and R by examining the p-value and the regression coefficients of R. All variables in the regression model are standardized.

The p-value of the regression coefficient for R is less than 0.05 but the value of the coefficient is 0.0396 indicating a very poor causal correlation and there is likely **no direct causal path from R to F**.

**Budget**

Y and I are confounder to the effect of B on F and S is a mediator for the effect of B on F. To disaggregate the causal effect of B on F, we set up a regression model, regressing F over I, B and S. (Conditioning on I effectively also conditions on Y). We evaluate the causal correlation between B and F by examining the p-value and the regression coefficients of B. All variables in the regression model are standardized.

The p-value of the regression coefficient for B is less than 0.05 and the value of the coefficient is 0.2916 indicating a weak but relevant causal correlation and we conclude that there is a **direct causal path from B to F**.

**Inflation**

To close all the backdoor paths between I and F we condition on B and Y. We set up a regression model, regressing F over I, Y and B and evaluate the causal correlation between F and I by examining the p-value and the regression coefficient of I.

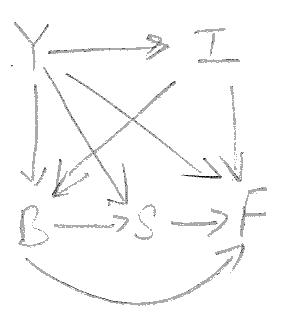
The p-value of the regression coefficient for I is less than 0.05 and the value of the coefficient is 0.2681 indicating a weak but relevant causal effect and we conclude that there is a **direct causal path from I to F**.

**Release Year**

To ascertain the direct causal effect of Y on F, we set up a regression model, regressing F over Y, I, B and S. This closes all but the direct path from Y to S. We evaluate the casual correlation between B and F by examining the p-value and the regression coefficient of Y.

The p-value of the regression coefficient for Y is less than 0.05 and the value of the coefficient is -0.2492 indicating a sweak but relevant causal correlation there is a a **direct causal path from Y to F**.

**Summary**

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

* **the budget of the film,**
* **the number of screens that is opens to,**
* **the inflation factor of the year of release, and**
* **the year of release.**

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 | Budget |
| Release Screens | Release Year, Budget |
| First Week Revenue | Budget, Release Screens, Inflation and Release Year |

**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: 45.23
* Percentage of estimates for test set that are off by less than 35% from true value: 61.23
* Percentage of estimates for test set that are off by less than 45% from true value: 72.62
* Percentage of estimates for test set that are off by less than 55% from true value: 81.85