Our goal is to build a model to predict the first week’s revenue, run length, footfalls, total revenue and exhibitor’s reveneue of a film. To that end, we want to establish which variables have a causal effect on all these dependent variables.

The list of candidate variables that likely affect FWR are:

|  |  |  |
| --- | --- | --- |
| Release Year | Inflation | Genre |
| Budget | Runtime | Release Week |
| Release Screens |  |  |

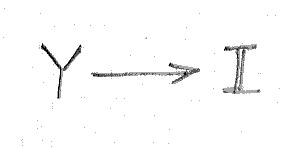
Of these I added inflation which is strictly not part of the variables that a film producer controls but maybe required to explain part of the effect of other variables like budget.

We use Directed Acyclic Graphs (DAGs) to aid building of an inferential model that identifies for each variable what other variables likely have a causal effect.

### Release Year (Y)

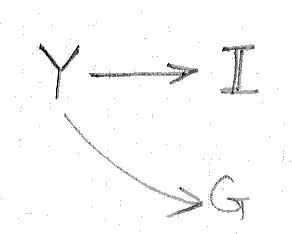
We begin with Y. We choose Y as it is not affected by any other variable in the list and is easily considered a random variable. A producer decides to release a film and the decision happens to be in a year. There is no other variable that impacts the choice of the Y.

### Inflation (I)

By definition I measures the price change year-on-year and we should expect Y to cause I.

A Spearman Rank correlation between Y and I of 1 is the result of the definition and hence there is a **direct** **causal path from Y to I**.

### Genre (G)

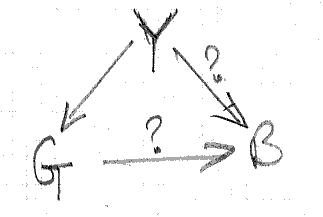
The decision to produce a film implicitly involves the decision of the genre of the film. We hypothesize that G is caused by Y. Moreover, given that the choice of Y is random, there are no other common causes to Y and Genre. We use the chi-square test to test this hypothesis.

A Chi-square statistic of 628.16 and an associated p-value < 10-6 clearly confirms that there is a **direct** **causal path from Y to G**.

There is **no causal path from I to G**. I can directly impact only such variables as may be quantified monetarily and G is clearly not such a variable.

### Budget (B)

Once a producer decides to make a film of a certain genre, the producer must decide the budget for the film.

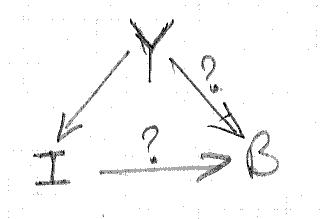
**Genre**

We hypothesize that G causes B – some genres are more expensive to make than others. We have already shown that Y is a cause for G and therefore Y is a likely common cause of B and G.

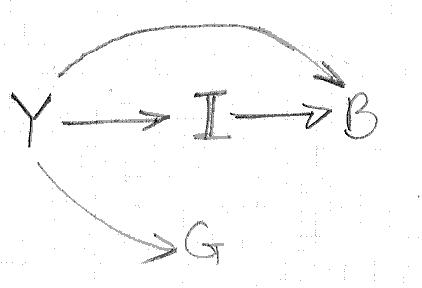
To test this hypothesis, we set up a regression model, regressing B over G and Y. Since G is a categorical variable we use dummy variables to one-hot encode G. All variables in the regression model are standardized. We ascertain the regression coefficient and the p-value for G to determine the likely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

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 B**.

**Release Year and Inflation**

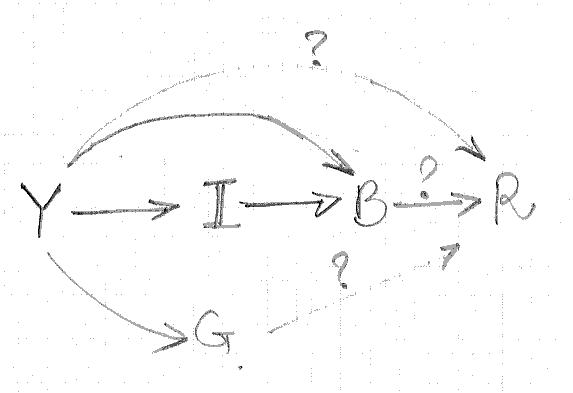
We hypothesize that Y has a causal effect on B. There are 2 likely causal paths: one that captures the direct effect of Y on B (in some years, finance is more easily available than in others) and the other that captures the indirect effect of Y on B, mediated by inflation (I).

To disaggregate the causal effect of Y and I over B we set up a regression model, regressing B over Y and I. All variables in the regression model are standardized. We ascertain the regression coefficient and the p-value for Y and I to determine the likely causal path from Y to B and I to B.

The p-values for coefficients for both Y and I are lower than 0.05 and the values of the regression coefficients are 1.0705 and 0.5474 respectively. We therefore conclude that there is **direct** **causal path from Y to B** and a **direct** **causal path from I to B**.

The overall DAG thus far is shown alongside.

### Runtime (R)

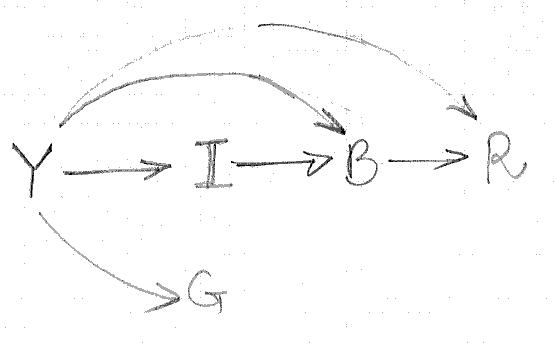
Once a film is made, the producer/director must decide the final cut and hence the runtime for the film. We hypothesize that there are causal paths from Y, B and G to R. The hypothesized DAG is shown alongside.

**Genre**

We hypothesize that G causes R – some genres need more time to tell the story. We have already shown that Y is a cause for G and therefore Y is a likely common cause of B and R. To test this hypothesis, we set up a regression model, regressing R over G and Y. 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 R.

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 R**.

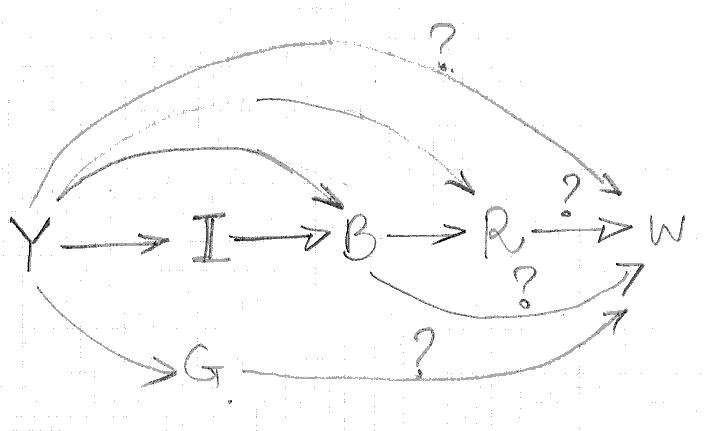
**Budget and Release Year**

To disaggregate the causal effect of B and Y on R we set up a regression model, regressing R over B, and Y - I can not have a causal relationship with R (a non-monteray measure). All variables in the regression model are standardized. We ascertain the regression coefficient and the p-values for B and Y to determine the likely causal path from B to R and Y to R.

The p-value of the regression coefficients for B and Y are significantly lower than 0.05 and the values of the regression coeffcients are 0.3907 and -0.5705 respectively. We conclude that there is a  **direct causal path from B to R** and a **direct causal path from Y to R**.

The updated DAG is shown above.

### Release Week (W)

A film is now ready and the producer must decide when to release the film. We look for correlation of W with other features we have discussed earlier, one feature at a time. The hypothesized DAG is shown alongside.

**Genre**

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

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 W**.

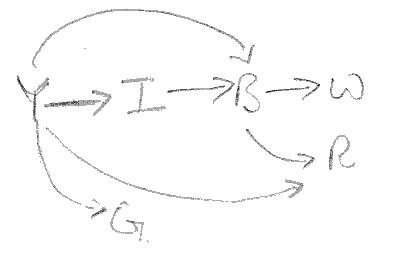
**Runtime**

Y and B are confounders to the effect of R on W. To ascertain the causal effect of R on W we set up a regression model, regressing W over R, B and Y. All variables in the regression model are standardized. We ascertain the regression coefficient and the p-value for R to determine the likely causal path from R to W.

The p-value of the regression coefficient for R is significantly greater than 0.05 and we conclude that there is likely **no direct causal path from R to W**.

**Budget and Release Year**

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

The p-value of the regression coefficient for Y is significantly lower than 0.05 and the regression coefficient is -0.0998 and we conclude that there islkely **no direct causal path from Y to W**.

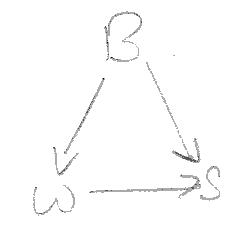
The p-value of the regression coefficient for Y is significantly lower than 0.05 and the regression coefficient is 0.1833. There is likely a weak causal effect of B on W and we conclude that there is **a direct causal path from B to W**.

The updated DAG is now shown alongside.

### Screens (S)

It seems almost axiomatic that a film that gets more number of screens is likely to get a wider audience. However, screens are a scarce commodity and controlled by not the producers but the exhibitors. We look for correlation of S with other features we have discussed earlier, one feature at a time.

**Release Week**

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

The p-value of the regression coefficient for W is 0.059 but the regression coefficient is -0.0299 and we conclude that there is likely **no direct causal path from W to S**.

**Genre**

Y is a confounder for the effect of G on S. To ascertain the causal effect of G on W, we set up a regression model, regressing W over G and Y. We ascertain the regression coefficient and the p-value for G to determine the likely causal path from G to W.

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 W**.

**Runtime**

Y and B are confounders to the effect of R on S. To ascertain the causal effect of R on S, we set up a regression model, regressing S over R, B and Y. We ascertain the regression coefficient and the p-value for R to determine the likely causal path from R to S.

The p-value of the regression coefficient for R is lower than 0.05 but the regression coefficient is 0.0647 and we conclude that there islkely **no direct causal path from R to S**.

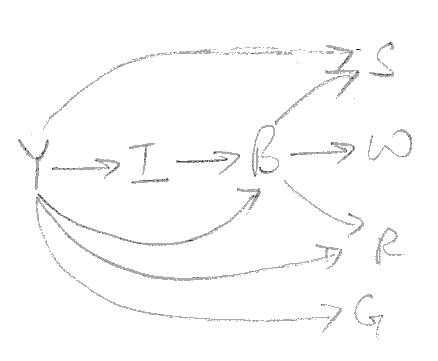
**Budget and Release Year**

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

The p-value of the regression coefficient for Y is significantly lower than 0.05 and the regression coefficient is 0.4324 and we conclude that there is a **direct causal path from Y to S**.

The p-value of the regression coefficient for B is significantly lower than 0.05 and the regression coefficient is 0.5922 and we conclude that there is **a direct causal path from B to S**.

## Summary

The relationships between the different features discovered by the causal inquiry process is shown in the DAG alongside and summarised in the table below:

|  |  |
| --- | --- |
| **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 |