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 |  |  |

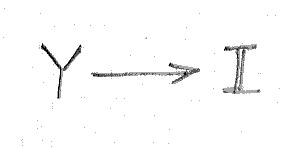
Of these I added inflation which is strictly not part of the variables that a film producer controls but maybe required to explain part of the effect of other variables like budget.

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)

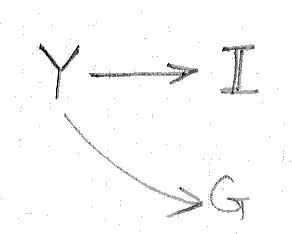
We begin with Y. We choose Y as it is not affected by any other variable in the list and is easily considered a random variable. A producer decides to release a film and the decision happens to be in a year. There is no other variable that impacts the choice of the Y.

### 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**.

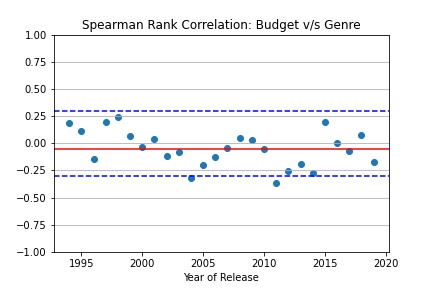
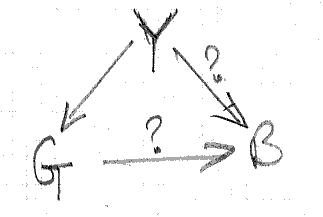
### Genre (G)

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.

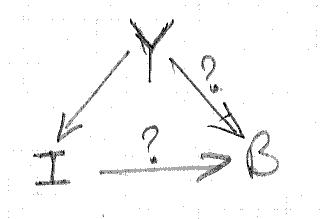
**Genre**

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.

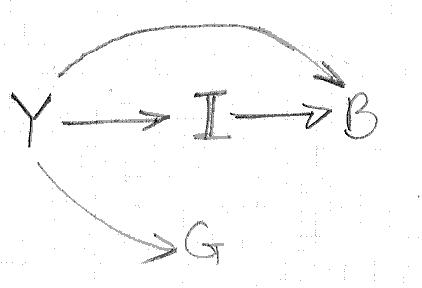
To test this hypothesis, we stratify the data by Y; this is effectively conditioning on Y and closes the backdoor path between G and B. We use the Spearman rank correlation to test the hypothesis.

The average Spearman rank correlation across years is -0.0483 and for all but two years and the value stays within [-0.3, 0.3]; even when it falls outside the interval its not by much. We can therefore conclude that there is likely **no direct causal path from G to B.**

**Release Year**

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).

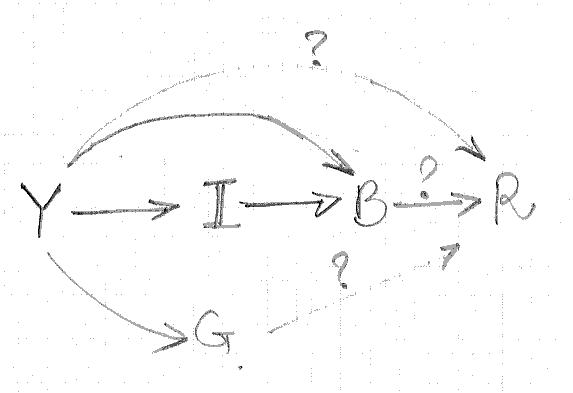
The total effect of Y on B is ascertained by the Spearman Rank correlation between the two. The direct effect of Y on B is ascertained by the Spearman Rank correlation beween Y and B adjusted to 2019 prices, i.e. conditioned on I. Finally, the indirect effect of Y on B is the difference between the total and direct effect of Y on B.

We confirm that Spearman Rank correlation for total effect of Y on B is 0.8033 and the Spearman Rank correlation for direct effect of Y on B is 0.5111; both correlation coefficient being significant implies that there is a **direct** **causal path from Y to B** and a **direct** **causal path from I to B**.

Moreover, since the Y and I are perfectly correlated, the Spearman Rank correlation between I and B is (0.8033 – 0.5111)/1 = 0.2922.

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**

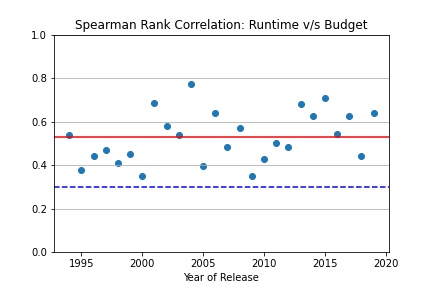
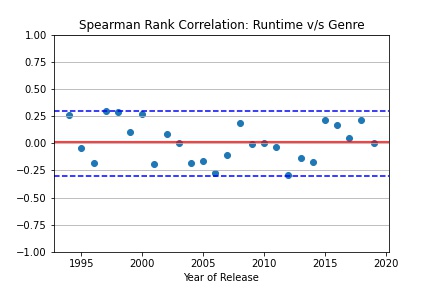
We stratify the data by Y; this is effectively conditioning on Y and closes the backdoor path between R and G. We use the Spearman rank correlation to test the hypothesis that there is a causal path from G to R.

The average Spearman rank correlation across years is 0.0143 and for all years the value stays within [-0.3, 0.3]. We can therefore conclude that there is **no direct causal path from G to R.**

**Budget**

Stratifying data by Y also closes all the backdoor paths betwen R and B. We use the Spearman correlation to test the hypothesis that there is a causal path between B and R.

The average Spearman Rank correlation across years is 0.5279 and for all years the value stays stays above 0.3. We can therefore conclude that there is a **direct** **causal path from B to R**.

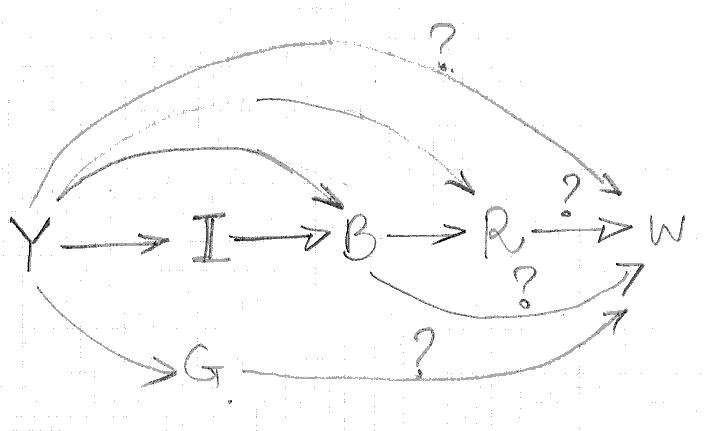
**Release Year**

There are now three causal paths from Y to R, two which pass through B. To find the direct effect of Y on R, we calculate the total effect of Y on R and then deduct from it the indirect effect of Y on R mediated by B. The indirect effect of Y on R mediated by B is the total effect of Y on B multipled by the direct effect of B on R.

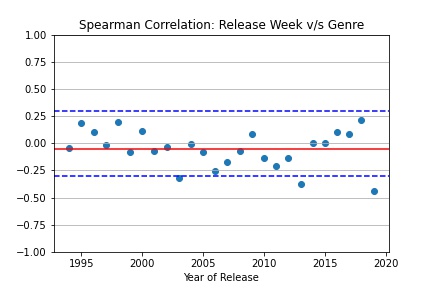
The Spearman Rank correlation between Y and R is -0.3872. We earlier established that the total effect of Y on B (using Spearman Rank correlation) is 0.8033 and the direct effect of B on R (using Spearman Rank correlation ) is 0.5279. Manipulating the following equation: -0.3872 = 0.8033\*0.5279 + x, yields that the indirect effect of Y on R x = -0.8113. We therefore conclude that there is a direct **direct** **causal path from Y to R**.

The updated DAG is shown below.

### 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 above.

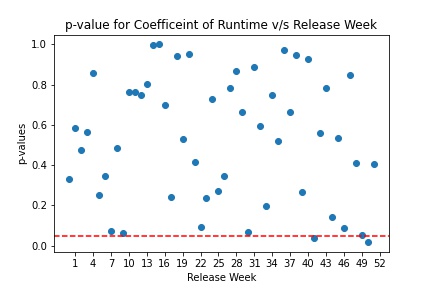
**Genre**

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

The average Spearman rank correlation across years is -0.0516 and for all years, but three, the value stays within [-0.3, 0.3]. Moreover, in years that the value does exceed -0.3, it does so by a small margin. We can therefore conclude that there is **no direct causal path from G to W.**

**Runtime**

To close all the backdoor paths between W and R, we need to condition on B and Y. We do this by setting up a multiclass logistic regression, regressing W over R, Y and B.

A pseudo r-squared value of 0.0202 indicates that the fitted multiclass logistic regression model likely does a poor job explaining W using R, Y and B.

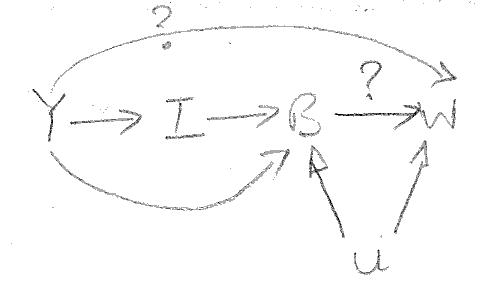
Specifically, if we consider the p-values for the coefficients of R across the per-week logistic regression model, we see that for all weeks, but two, the p-values are in excess of 0.05. This leads to the conclusion that R is not significant in predicting the value of W.

We can therefore conclude that there is likely **no direct causal path from R to W**.

**Budget**

To close all the backdoor paths between W and B, we need to condition on Y. We stratify the data by Y; this is effectively conditioning on Y and closes the backdoor path between W and B. We use the Spearman rank correlation to test the hypothesis that there is a direct causal path from B to W.

The average Spearman rank correlation across years is 0.1718 indicating that there is no causal path from B to W. However there might be a slight twist in this relationship.

In analyzing the causal relationship between B and W, we want to incorporate the idea that certain high profile producers/production houses are able to *choose* release weeks for their large budget films. This changes the DAG somewhat and relevant portion of it is presented alongside.

U here represents the profile of the production house. *It is worth noting that introduction of U does not in any way affect any of the discussions / conclusions drawn so far.* We name this variable U because it is difficult to measure the profile of a production house and is therefore *unknowable*.

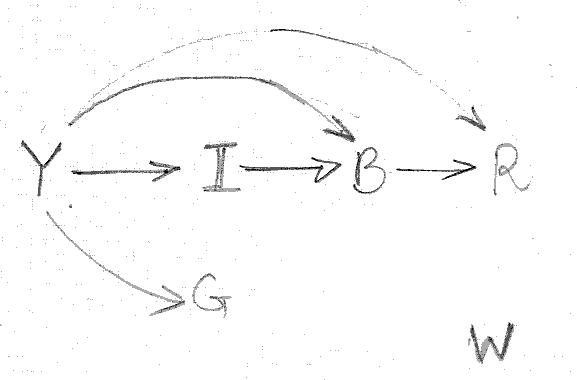
There are now three backdoor paths from W to B, W <- Y -> I -> B, W <- Y -> B and W <- U -> B. Conditioning on Y closes the first two backdoor paths but not the third and we use this fact to calculate the total effect of B on W of the remaining two paths: B -> W and W <- U -> B.

The average Spearman Rank correlation we estimated earlier ( 0.1718) now represents the total effect of B on W of the remaining two paths.

To disaggregate the effects of the two paths, we use I as an instrumental variable. We can do so because I satisfies the three conditions:

1. no causal path between I and U; inflation can’t affect the profile of the production house,
2. causal path between I and B (proven earlier), and
3. no direct causal path between I and W; inflation can only affect monetarily quantifiable variables

Using the Spearman Rank correlation we determine that the total effect of I and W is -0.0004, the effect of I on B is 0.2922 (from earlier analysis) and therefore the direct effect of B on W is -0.0004 / 0.2922 = -0.0013, leading to the conclusion that this is likely **no direct causal path from B to W** even accounting for the profile of the production house releasing the film.

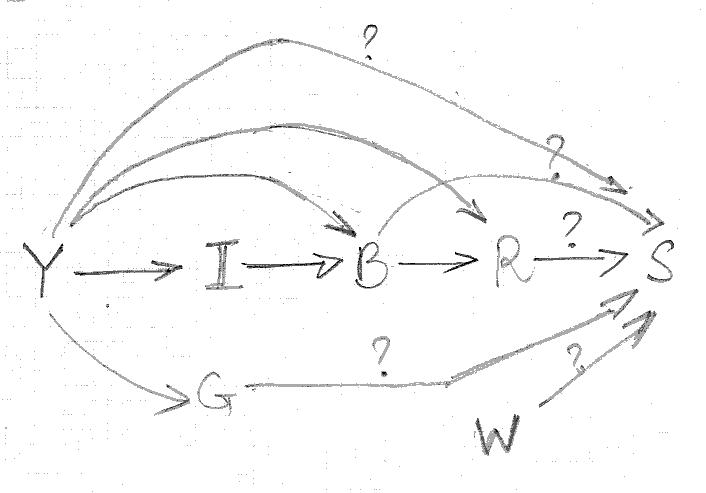
Finally, we investigate the causal path from Y to W. Given the earlier discussion in this section, the only causal path that possibly exists between Y and W is the direct one. We investigate the hypothesis using the Spearman Rank correlation.

A Spearman Rank correlation of 0.004 leads to the conclusion that there is **no direct causal path from Y to W**.

The updated DAG is now shown alongside.

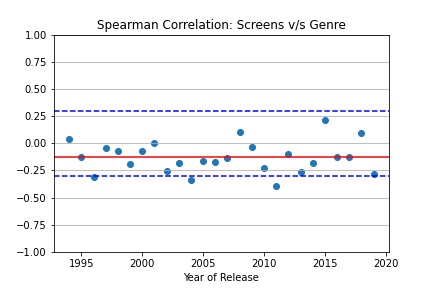
W is quite unique of all the features discussed so far in that it has no corelation with any of the earlier features.

### 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. The hypothesized DAG is alongside.

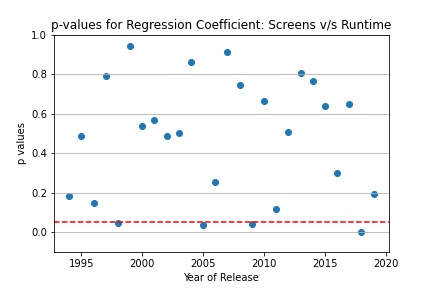
**Release Week**

The correlation between W and S can be easily ascertained using the Spearman Rank correlation. A Spearman Rank correlation value of 0.0809 helps us conclude that there is likely **no direct causal path from W to S**.

**Genre**

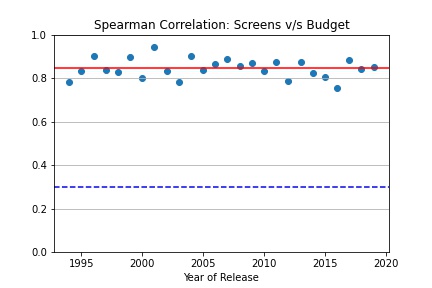
Y is a confounder for the effect of G on S. We stratify the data by Y; this is effectively conditioning on Y and closes the backdoor path between S and G. We use the Spearman rank correlation to test the hypothesis that there is a causal path from S to G.

The average Spearman rank correlation across years is -0.1281 and for all years, but two, the value stays within [-0.3, 0.3]. Moreover, in years that the value does exceed -0.3, it does so by a small margin. We can therefore conclude that there is likely **no direct causal path from G to S.**

**Runtime**

To close all the backdoor paths from S to R, we need to condition on Y and B. To achieve this, we set up for each year linear regression models, regressing S on B and R, and then evaluate the correlation between S and R by evaluating the p-values of the regression coefficients of R.

For all years, but four, the p-value for the regression coefficient of R is greater than 0.05 and three years the p-value is less than 0.05 its not significantly so. This leads to the conclusion that there is likely **no direct causal path from R to S**.

**Budget**

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

The average Spearman Rank correlation across years is 0.8449 and for all years the value stays above 0.75. We can therefore conclude that there is a **direct** **causal path from B to S.**

**Release Year**

There now remain 3 causal paths from Y to S: Y -> S, Y -> B -> S and Y -> I -> B -> S, two of which are mediated by B. To find the direct effect of Y on S, we calculate the total effect of Y on S and then deduct the indirect effect mediated by B. We can ascertain the indirect effect mediated by B because we know the total effect of Y on B and the direct effect of B on S.

The Spearman Rank correlation between Y and S is 0.8771 and represents the total effect of Y on S. The direct effect of Y on S is therefore: 0.8771 – 0.8033 \* 0.8449 = 0.1984. We therefore conclude that there is likely **no direct causal path from Y to S**[[1]](#footnote-2). The updated DAG is shown below.

## Summary

The relationships between the different features discovered by the causal inquiry process is summarised below:

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

1. This is a great example of *correlation does not imply causation*. Number of screens have increased with each passing year but it is not the passing year that is causing the increase in number of screens. [↑](#footnote-ref-2)