**Run Length (L)**

The conventional wisdom in the industry is that the fate of a film is decided by its first week’s performace at the box-office. However, over years, we have all heard of films that took time to *catch fire*. Sometimes theater owners may be making a hasty decision in writing off a film just based on its first week performance. We define *run length* as the number of weeks a film showed in theater i.e. recorded revenues. We want to investigate the features that affect a film’s run length. We now look at the likely causal relationships of the different features to the film’s run length. To avoid clutter we draw simpler DAGs that model the likely causal relationship between L 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 used to establish the run length of a film.

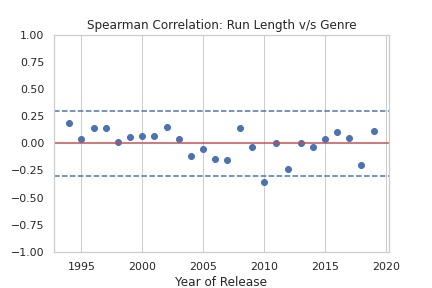
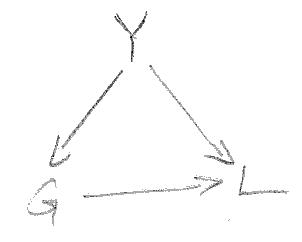
**Release Week**

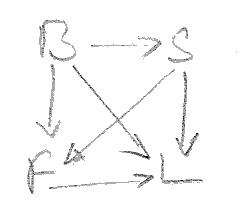
The correlation between W and L can be easily ascertained using the Spearman Rank correlation.

A Spearman Rank correlation value of -0.0241 indicates that there is likely **no direct causal path from W to L**.

**Genre**

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

The average Spearman rank correlation across years is 0.0005 and for all years but one the value stays within [-0.3, 0.3]. We can therefore conclude that there is likely **no direct causal path from G to L.**

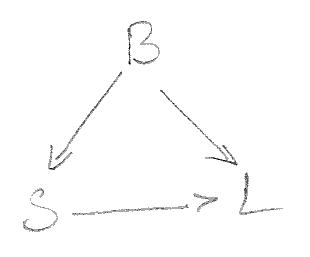
**First Week Revenue**

B and S are likely confounders to the effect of F on L. To condition on all the confounders we set up a regression model, regressing L on B, S, and F. We ascertain the p-value of the coefficeint of F to determine the existense of a causal path from F to L.

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

However, we also note that the R-squared value for the regression model is pretty low indicating that a linear model may not be adequate to explain the effect of F on L.

**Release Screens**

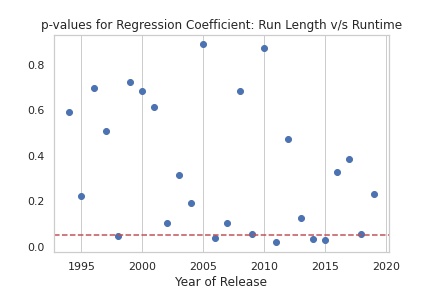
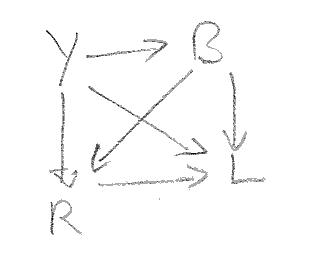
B is a confounder for the effect of S on L. Admittedly, S here is not the number of screens the film is playing to through the run. We however make the assumption that the number of screens that a film gets in its run is likely a function of the number of screens that it opens to and hence we use S as a proxy variable for the number of screens that the film plays to during its run.

To condition on B, we set up a linear regression model, regressing L on S and B. We ascertain the p-value of the coefficient of S to determine the likely causal path from S to L.

The p-value for the coefficient of S is less than 0.05 implying that S has a significant association with F. However the R-squared value for the model is pretty low and we conclude and there is a **direct causal path from S to L** but a linear model may not be appropriate to explain the causal effect.

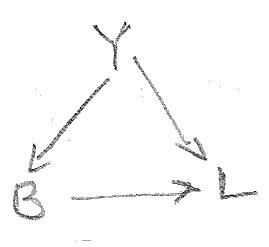
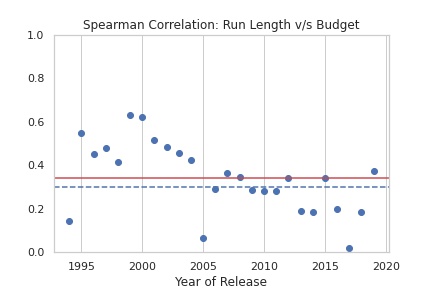
**Runtime**

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

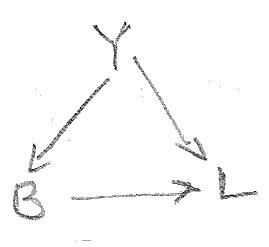
For all years except 5, the p-value of the regression coefficient of R is greater than 0.05, leading to a conclusion that R is not significantly related to L and there is likely **no direct causal path from R to L**.

**Budget**

Release Year is a confounder to the effect of B on L. We stratify on Y; this is effectively conditioning on Y and closes the backdoor paths between L and B. We use the Spearman rank correlation to test the hypothesis that there is a direct causal path from L to F.

The average Spearman rank correlation across years is 0.3422 indicating that there is likely **no** **direct causal path from B to L.** It is interesting to note that after around 2005, the correlation coefficient is reducing even below 0.33, perhaps pointing to reducing ability of a film to get an extended run just on its producers’ ability to spend.

**Release Year**

There are 2 possible causal paths from Y to L: Y -> L and Y -> B -> L. If we calculate the total effect of Y on L, then we get the combined effect of the two paths.

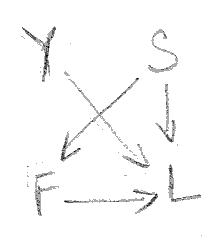
The total effect of Y on L is -0.6032. We know from earlier analysis that the direct effect of Y on B is 0.5111.

The Spearman Rank correlation between B and L represents the total effect of B on L and is 0.3422.

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

-0.6032 = x + 0.5111\*0.3422, i.e. x = -0.7781. This value of x leads to the conclusion that there is a **direct causal path from Y to L**.

**Summary**

In summary, we conclude that the **run length of a film is influenced by**

* **the first week revenue,**
* **the number of screens that it opens to, 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 |  |
| Release Screens | Budget |
| First Week Revenue | Budget, Release Screens |
| Run Length | Release Year, Release Screens, First Week Revenue |

**Prediction Model Summary**

We fitted a Gradient Boosted Ensemble 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: 74.77
* Percentage of estimates for test set that are off by less than 35% from true value: 91.38
* Percentage of estimates for test set that are off by less than 45% from true value: 95.08
* Percentage of estimates for test set that are off by less than 55% from true value: 98.77

**Test Set Performance**

* Percentage of estimates for test set that are off by less than 25% from true value: 75.38
* Percentage of estimates for test set that are off by less than 35% from true value: 90.77
* Percentage of estimates for test set that are off by less than 45% from true value: 94.15
* Percentage of estimates for test set that are off by less than 55% from true value: 97.54