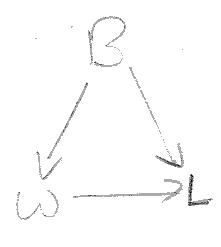
**Run Length (L)**

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.

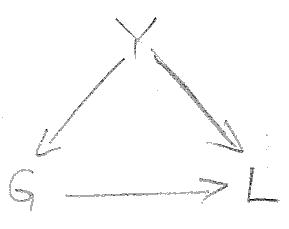
The weekly revenue data available is used to establish the run length of a film.

**Release Week**

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

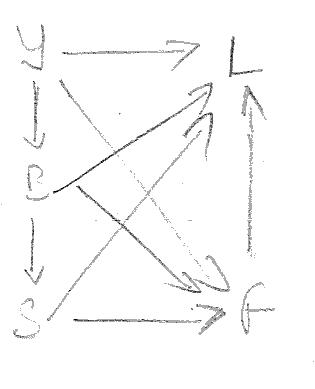
The p-value of the regression coefficient for W is 0.943 and we conclude that there is likely **no direct causal path from W to F**.

**Genre**

Y is a confounder for the effect of G on L. To condition on Y, we set up a regression model, regressing L 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 L.

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 L**.

**First Week Revenue**

Y, B and S are likely confounders to the effect of F on L. (We ignore I because I can causally affect variables that are monetary in nature. To condition on all the confounders we set up a regression model, regressing L on Y, B, S, and F. We ascertain regression coefficient and the p-value 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 and the regression coefficient of F is 0.553, implying that F has a significant association with L and there is a **direct causal path from F to L**.

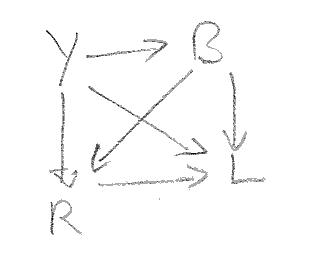
**Release Screens**

Y and B are confounders tand F is a mediator o 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 Y, B, and F we set up a linear regression model, regressing L on S, Y, B, F. We ascertain the p-value of the coefficient of S to determine the likely causal path from S to L. The causal diagram is the same as in the previous section and is not shown here again.

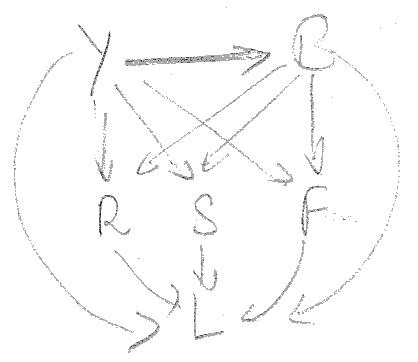
The p-value for the coefficient of S is less than 0.05 and the regression coefficient of S is -0.6864, implying S has a significant association with L and there is a **direct causal path from S to L**.

**Runtime**

Both B and Y are confounders to the effect of R on L. To condition on B and Y, we set up a regression model regressing L on R and B and then evaluate the causal correlation between L and R by examining the regression coefficients of R and its p-value.

The p-value for the coefficient of R is less than 0.05 and the regression coefficient of R is 0.2360, implying R has an association with L and there is a **direct causal path from R to L**.

**Budget**

Y is a confounder and S, R and F are mediators to the effect of B on L. To ascertain the causal relationship betweek B and L, we set up a regression model, regressing L over B, Y, S, R and F. We then evaluate the causal correlation between B and L by ascertaining the regression coefficient and its p-value for B.

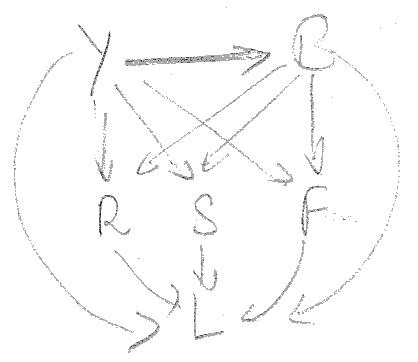
The p-value for the coefficient of B is less than 0.05 and the regression coefficient of R is -0.1094, implying at a weak causal relationship between B and L and we conclude that there is a **direct causal path from R to L**.

**Release Year**

B, R, S and F are mediators to the effect of Y on L. The causal diagram in the earlier section is applicable. To ascertain the causal relationship between Y and L, we use the results of the regression model in the previous section and ascertain the regression coefficient and its p-value for Y.

The p-value for the coefficient of B is less than 0.05 and the regression coefficient of R is -0.5251, implying at a strong causal relationship between Y and L and we conclude 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,**
* **the runtime of the film,**
* **the budget of the film, 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 | Release Year, Inflation, Budget, Release Screens |
| Run Length | Release Year, Budget, Release Screens, Runtime, 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: 71.69
* Percentage of estimates for test set that are off by less than 35% from true value: 85.54
* Percentage of estimates for test set that are off by less than 45% from true value: 88.62
* Percentage of estimates for test set that are off by less than 55% from true value: 92.00

**Test Set Performance**

* Percentage of estimates for test set that are off by less than 25% from true value: 76.00
* Percentage of estimates for test set that are off by less than 35% from true value: 84.31
* Percentage of estimates for test set that are off by less than 45% from true value: 93.54
* Percentage of estimates for test set that are off by less than 55% from true value: 96.00