**Total Nett Gross Revenue (T)**

We want to build a prediction model that allows industry participants to predict the likely total nett gross revenue from a movie. To that end, we now look at the likely causal relationships of the different features to the total nett gross revenue earned by a film. To avoid clutter we draw simpler DAGs that model the likely causal relationship between T and a feature and will include other features that were seen to have causal relationship, in the analysis earlier, with the feature under investigation.

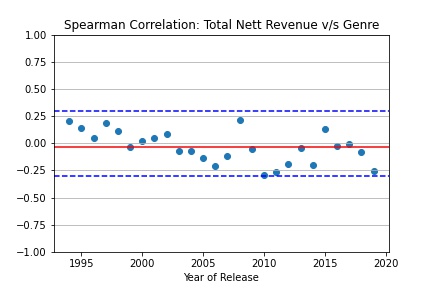
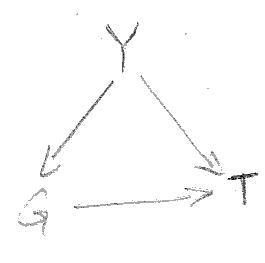
**Release Week**

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

A Spearman Rank correlation value of 0.0621 indicates that there is likely **no direct causal path from W to T**.

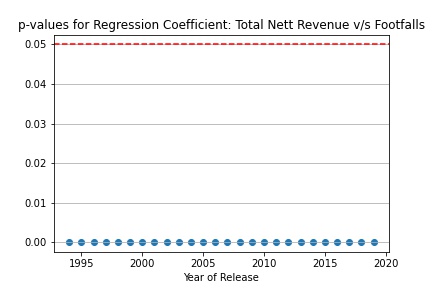
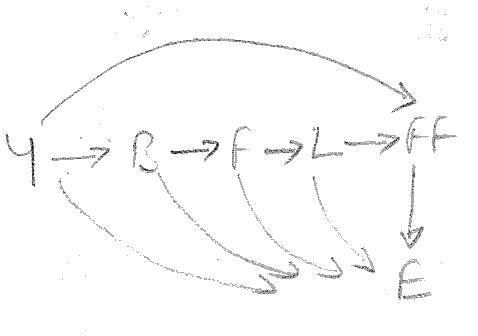
**Genre**

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

The average Spearman Rank correlation across years is -0.0327 and for all years the value stays within [-0.3, 0.3]. We can therefore conclude that there is likely **no direct causal path from G to T.**

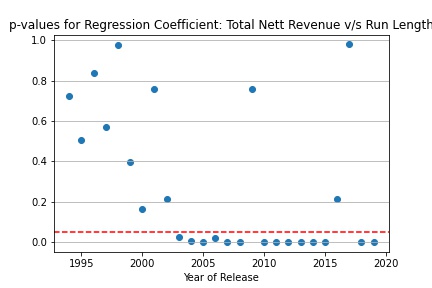
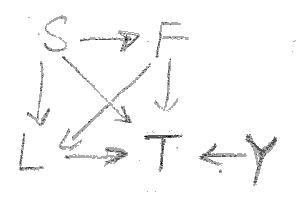
**Footfalls**

Y, B, F and L are likely confounders to the effect of FF on T. To condiction on the confounders we set up regresion model for each year, regressing T on B, F, L and FF and then evaluate the correlation between T and FF by examiming the p-values of the regression coefficients of FF.

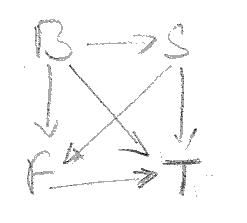
The p-values for the yearly regression models are all well below 0.05 and indicate that FF has a significant effect on T. We conclude that there is likely a **direct causal path from FF to T**.

**Run Length**

Y, S and F are likely confounders to the effect of L on T. To condition on the confounders we set up regression model for each year, regressing F on S, F and L and then evaluate the correlation between T and L by examining the p-values of the regression coefficients of L.

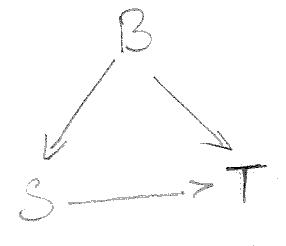
The p-values for the yearly regression model show a mixed picture. For 14 of the 24 years, the p-value is less than 0.05 but for the remaining 11 years it isn’t. However on balance, we assert that there is a **direct causal path from L to T**.

**First Week Revenue**

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

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

**Release Screens**

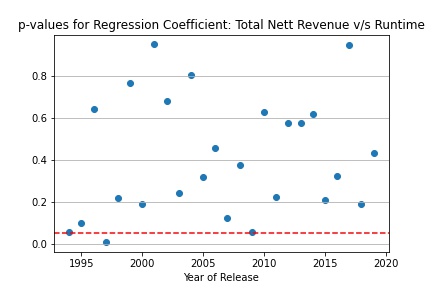
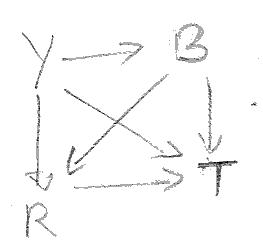
B is a confounder for the effect of S on T. Admittedly, S here is the number of screens the film opens to in its first week. We however make the assumption that the number of screens that a film gets through its run is likely a function of the number of screens that it opens to and hence we use S as 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 T on S and B. We ascertain the values of the coeffecient value and its p-value to determine the likely causal path from S to T.

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

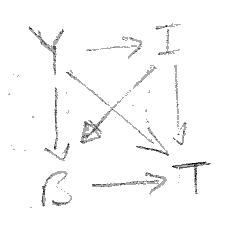
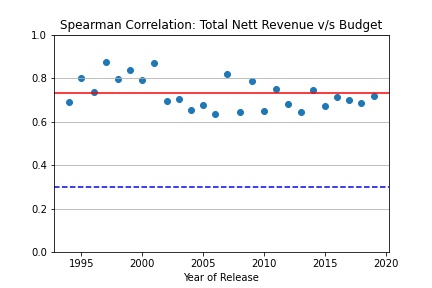
**Runtime**

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

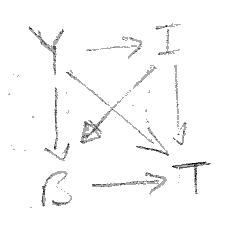
For all years except one, the p-value of the regression coefficient of R is greater than 0.05, leading to a conclusion that R is not significantly related to T and there is likely **no direct causal path from R to T**.

**Budget**

Release Year and Inflation are likely confounders to the effect of B on T. We stratify on Y; this is effectively conditioning on Y and closes one of the two the backdoor paths between T and B. However, we note that in a year I has the same value for all films and hence conditioning on Y is also effectively conditioning on I and closing the second backdoor path between T and B. We use the Spearman Rank correlation to test the hypothesis that there is a direct causal path from T to F.

The average Spearman rank correlation across years is 0.7290 and for all years the value stays above 0.6. We can therefore conclude that there is a **direct causal path from B to T.**

**Release Year**

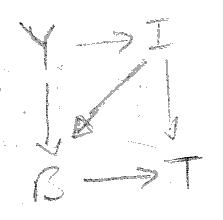
There are 4 possible causal paths from Y to T: Y -> T, Y -> B -> T, Y -> I -> T and Y -> I -> B -> T. Conditioning on I closes the latter 2 paths. We achieve this by adjusting the values of both B and T for inflation. If we calculate the total effect of Y on inflation adjusted T, then we get the combined effect of the first two paths: Y -> T and Y -> B -> T, where Y -> B represents the direct effect of Y on B (not mediated by I).

The total effect of Y on T, conditioned on T, is 0.2484. We know from earlier analysis that the direct effect of Y on B is 0.5111.

The average Spearman Rank correlation across years between B and T represents the direct effect of B on T and is 0.7290.

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

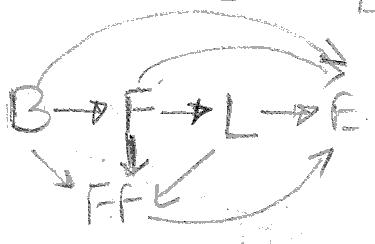
0.2484 = x + 0.5111\*0.7290, i.e. x = -0.1242. This value of x leads to the conclusion that there is likely **no** **direct causal path from Y to T**.

The earlier DAG now reduces to the DAG presented alongside. The Spearman Rank correlation between Y and T represents the total effect of Y on T and is 0.6234. From earlier analysis we know that Spearman Rank correlation between I and B is 0.2922.

We can therefore calculate the direct effect of I on F by manipulating the following equation:

0.6234 = 0.5111\*0.7290+ 1\* 0.2922\*0.7290 + x, i.e. x = -0.0378. This value of x leads to the conclusion that there is likely **no direct causal path from I to F**.

**Summary**

In summary, we conclude that the **Total Nett Gross Revenue of a film is influenced by:**

* **total footfalls,**
* **the run length,**
* **the number of screens the film releases to**
* **first week revenue, and**
* **budget of the film.**

The DAG for this model is presented alongside.

We update the summary table of causation:

|  |  |
| --- | --- |
| **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 |
|  |  |
| Footfalls | Release Year, Budget, First Week Revenue, Run Length |
| Total Nett Revenue | Footfalls, Run Length, First Week Revenue, Release Screens, Budget |

**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: 91.69
* Percentage of estimates for test set that are off by less than 35% from true value: 95.38
* Percentage of estimates for test set that are off by less than 45% from true value: 97.85
* Percentage of estimates for test set that are off by less than 55% from true value: 98.46

**Test Set Performance**

* Percentage of estimates for test set that are off by less than 25% from true value: 73.54
* Percentage of estimates for test set that are off by less than 35% from true value: 85.85
* Percentage of estimates for test set that are off by less than 45% from true value: 91.38
* Percentage of estimates for test set that are off by less than 55% from true value: 94.46