**Assignment : Causal Discovery**

8th April 2018

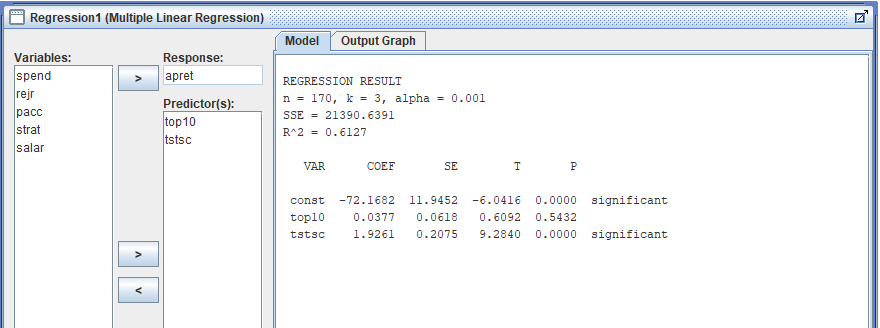
Team Members:

Prashant Wankhede (PRW19@pitt.edu)

Abhishek Shankarnarayanan (ABS100@pitt.edu)

**Highlights of the Results after analysis of retention.txt using Tetrad:**

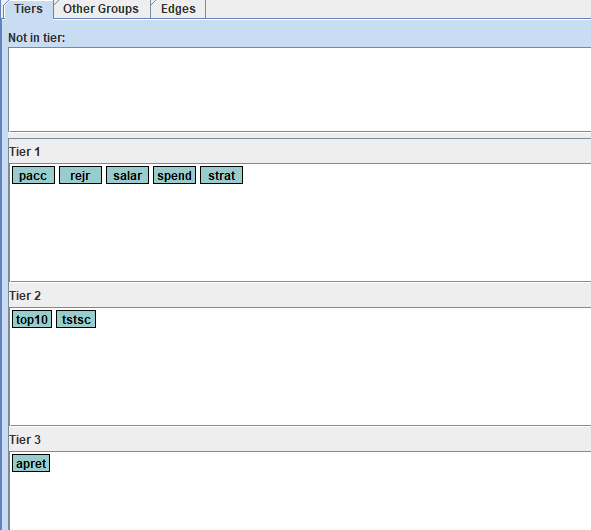
* We can concur that the 1993 data (retention.txt) does support the research findings in Druzdzel & Glymour 1994
* It can be deduced from the patterns that student retention is caused by average tests scores as inferred in the study.
* We devised alternate temporal orderings, as inferred through descriptive statistics (later discussed in the report), we still found that student retention (apret) was steadily and directly caused by average tests scores (tstsc)
* Changing the significance levels does not affect the primary result structure of the pattern, thus we can still conclude that apret (average student retention rate) is causally and directly dependent on (tstsc) average tests scores.
* By changing the significance levels, we get rid of other “latent common cause” connections, salar at p <= 0.51 and pacc at p < 0.248
* Data present in retention.txt can be almost said to be normally distributed and linearly dependent, because of a few skewed distributions (rejr, spend). The PC algorithm also assumes that the data is normally distributed, we verified the same by running descriptive statistics on it.



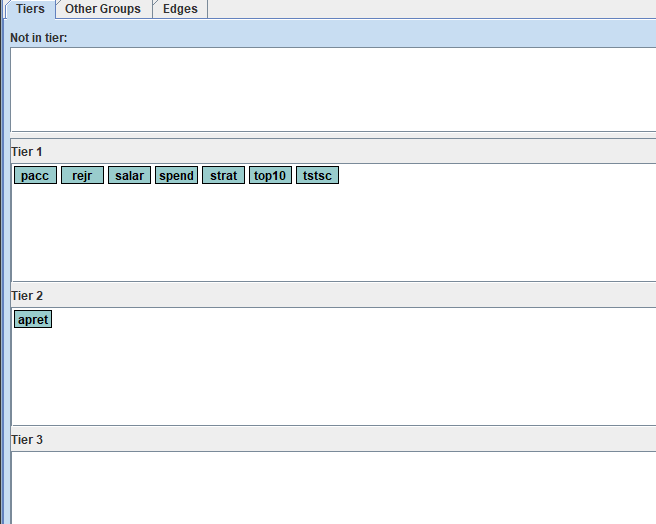
**Causal graphs/Patterns suggested by Tetrad**

Temporal Ordering:

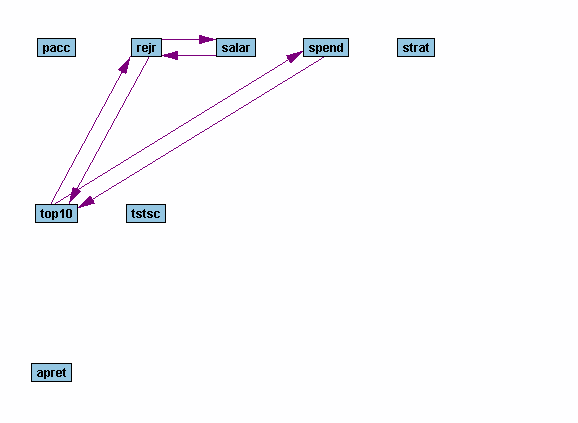
As per the linear dependencies we inferred through the scatterplots and regression lines we ran on the dataset, we set experimented with 3 different temporal orderings



(1)



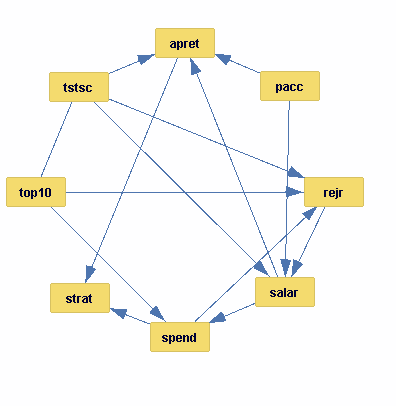
(2)



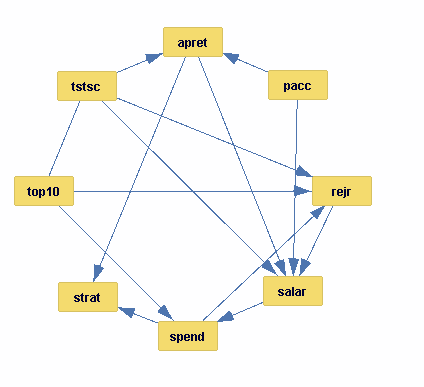
(3)

**Patterns suggested by Tetrad (For temporal ordering (1)):**

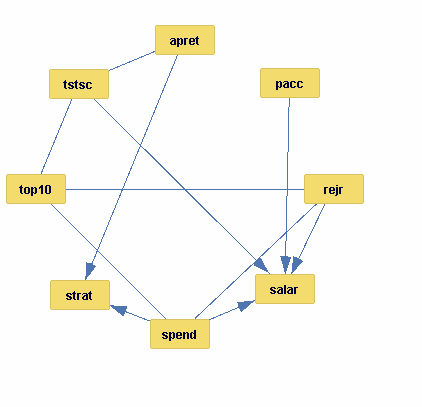
p = 0.6, We see 3 variables (salar,pacc,tstsc) acting as causes for apret



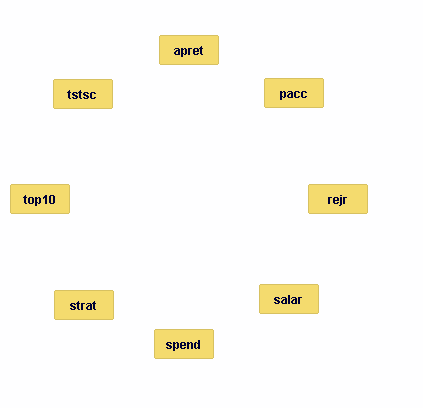
p = 0.248, We see 2 variables (pacc,tstsc) acting as causes for apret, we get rid of salar as a cause by decreasing p



p = 0.247, We see 1 variable (tstsc) acting as causes for apret, we get rid of pacc as a cause by decreasing p even further



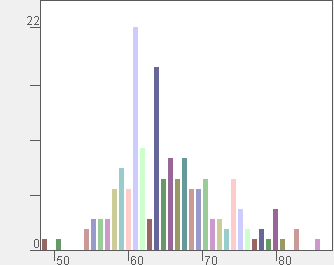
p = 6.0E-18, We see 0 variables acting as causes for apret, we get rid of tstsc as a cause by decreasing p even further



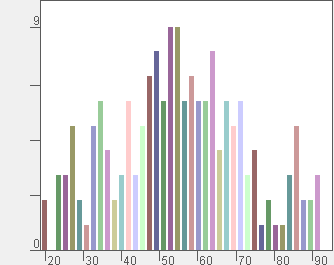
**Testing the assumptions about the data present in Retention.txt**

We check if the data is normally distributed and linearly dependent, for this we will create histograms and run regression. Results are as below (All results are not included to keep the report short)

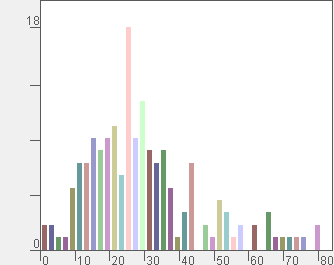
Observation: Except spend and rejr who have right skewed distribution, all other have normal distribution.



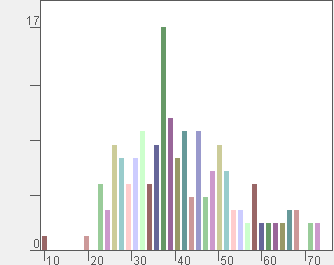
Tstsc



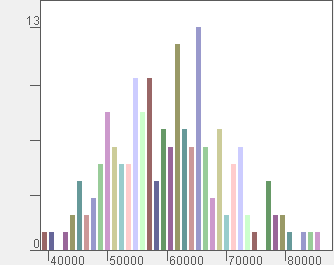
apret



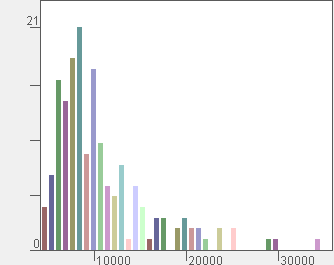
Rejr



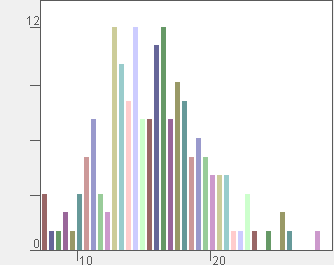
pacc



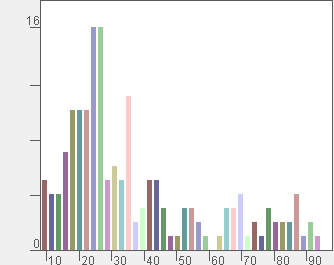
salar



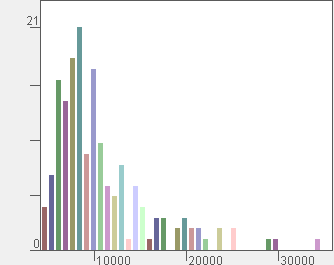
spend



strat



Top10



Spend

**Scatterplots (Regression lines):**

