INFSCI 2725 Data Analytics

Assignment 9

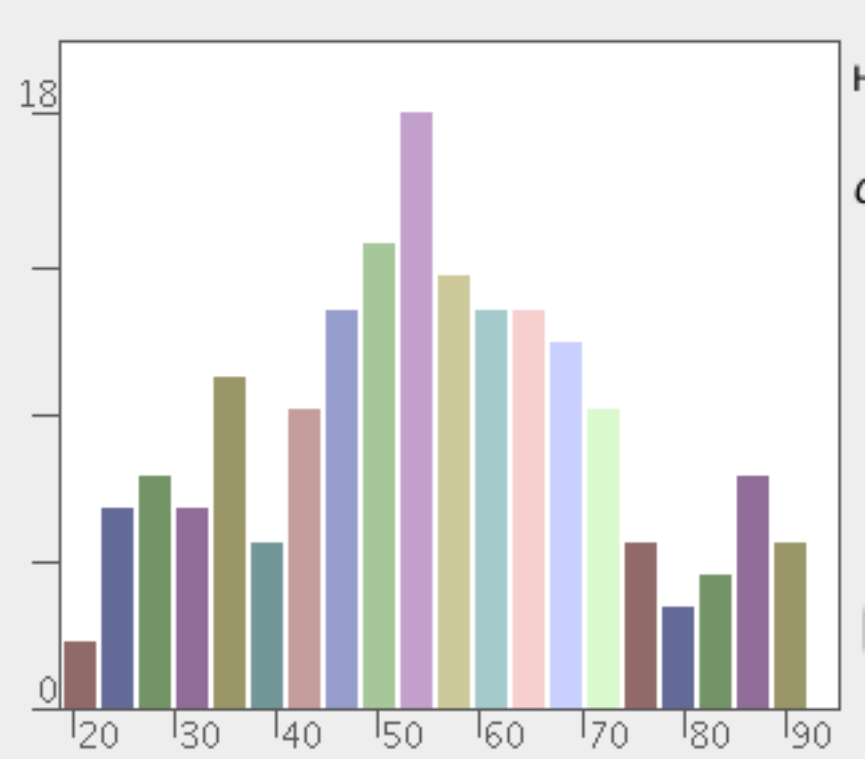
Retention Causal Analysis

Zheng Liu E-mail:zhengliu@pitt.edu
Luda Wang E-mail:luw20@pitt.edu
Kewei Li E-mail:kel137@pitt.edu

1. Normal Distribution Test

1.1. APRET

Histogram with 20 bins



Normality Tests for: apret (sample size:170)

Kolmogorov Smirnov:

K-S Statistic: 0.0418415

Significance Levels:	.20	.15	.10	.05	.01
K-S Critical Values:	0.0418	0.0821	0.0874	0.0936	0.1043
Test Result:	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT

H0 = apret is Normal.
(Normal if ACCEPT.)

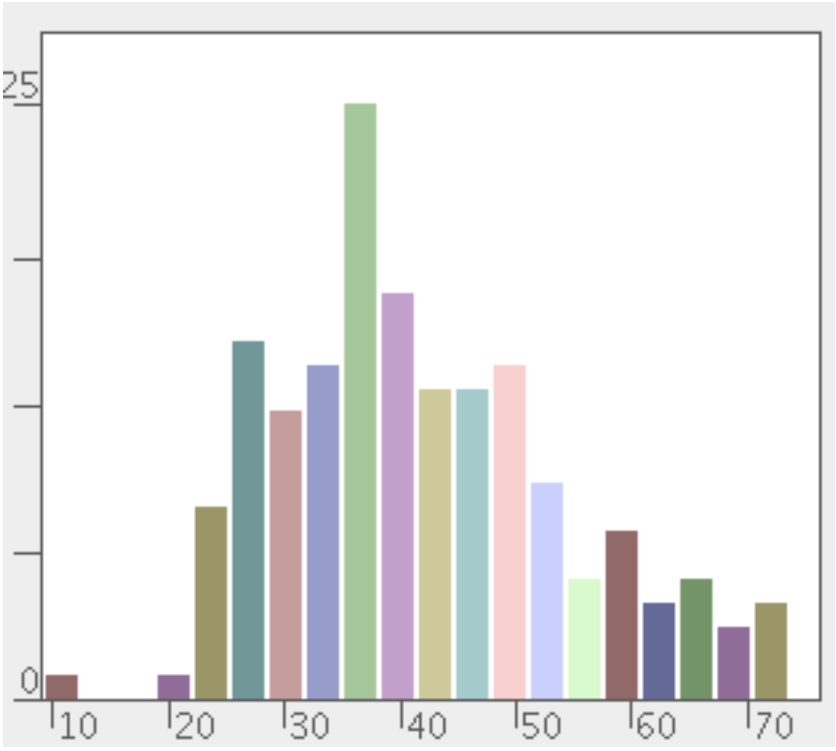
Anderson Darling Test:

A^2 = 0.3762
A^2* = 0.3779
p = 0.4080

H0 = apret is Non-normal.
(Normal if p > alpha.)

Based on the histogram and two statistical tests, the APRET is normal distribution.

1.2. PACC



Normality Tests for: pacc (sample size:170)

Kolmogorov Smirnov:

K-S Statistic: 0.0820562

Significance Levels:	.20	.15	.10	.05	.01
K-S Critical Values:	0.0821	0.0821	0.0874	0.0936	0.1043
Test Result:	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT

H0 = pacc is Normal.
(Normal if ACCEPT.)

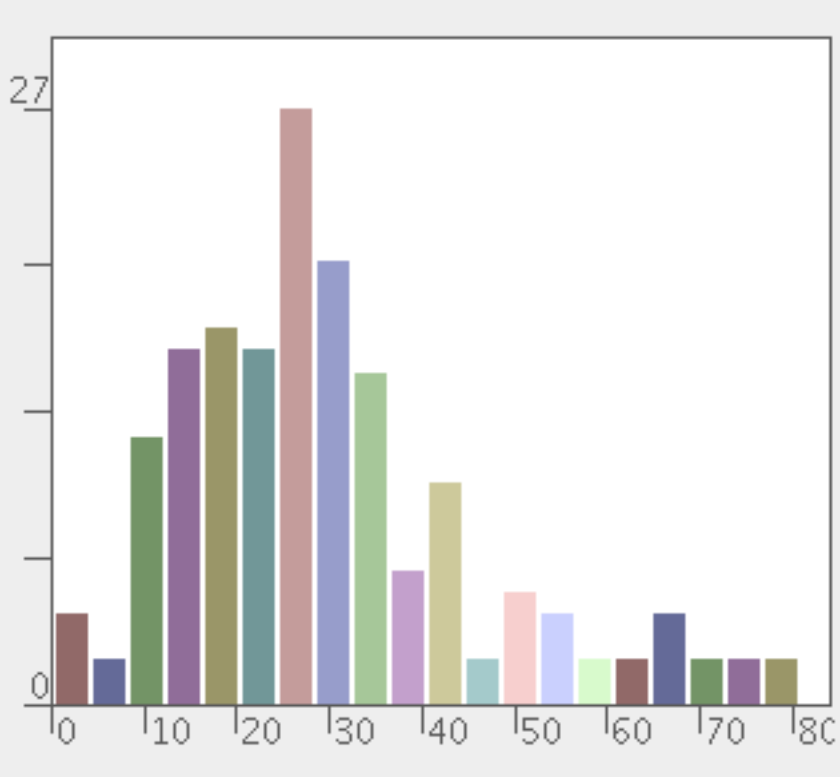
Anderson Darling Test:

A^2 = 1.3904
A^2* = 1.3966
p = 0.0013

H0 = pacc is Non-normal.
(Normal if p > alpha.)

Based on the histogram and two statistical tests, the PACC is normal distribution.

1.3. REJR



Normality Tests for: rejr (sample size:170)

Kolmogorov Smirnov:

K-S Statistic: 0.1192178

Significance Levels:	.20	.15	.10	.05	.01
K-S Critical Values:	0.1192	0.0821	0.0874	0.0936	0.1043
Test Result:	FAIL	FAIL	FAIL	FAIL	ACCEPT

H0 = rejr is Normal.
(Normal if ACCEPT.)

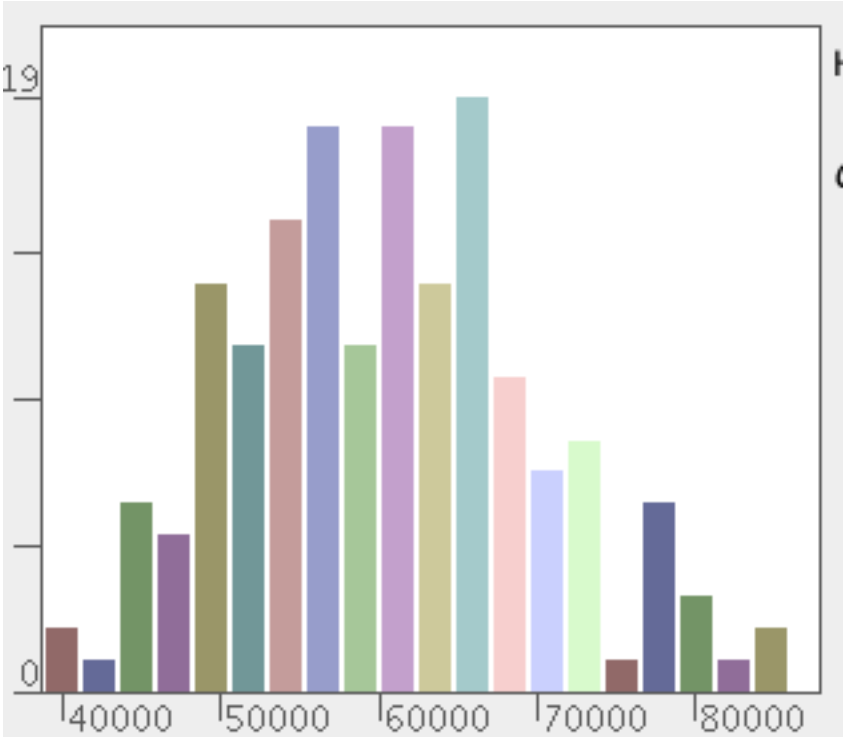
Anderson Darling Test:

A^2 = 3.5695
A^2* = 3.5855
p = 0.0000

H0 = rejr is Non-normal.
(Normal if p > alpha.)

Based on the histogram and two statistical tests, the REJR is not normal distribution. We can tell REJR is right skewed by histogram.

1.4. SALAR



Normality Tests for: salar (sample size:170)

Kolmogorov Smirnov:

K-S Statistic: 0.0439651

Significance Levels:	.20	.15	.10	.05	.01
K-S Critical Values:	0.0440	0.0821	0.0874	0.0936	0.1043
Test Result:	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT

H0 = salar is Normal.
(Normal if ACCEPT.)

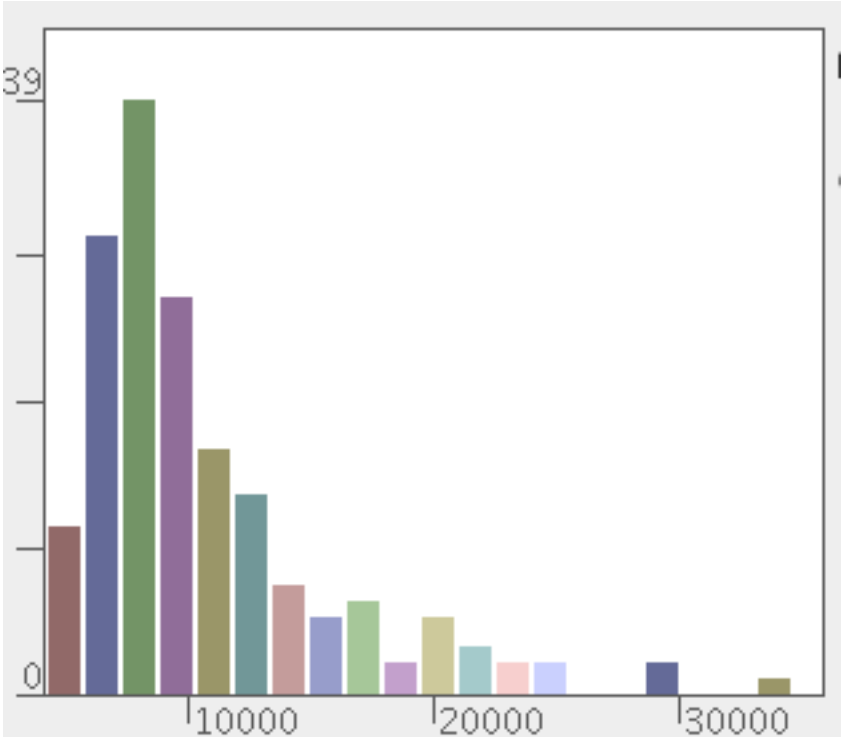
Anderson Darling Test:

A^2 = 0.3263
A^2* = 0.3278
p = 0.5181

H0 = salar is Non-normal.
(Normal if p > alpha.)

Based on the histogram and two statistical tests, the SALAR is normal distribution.

1.5. SPEND



Normality Tests for: spend (sample size:170)

Kolmogorov Smirnov:

K-S Statistic: 0.1686838

Significance Levels:	.20	.15	.10	.05	.01
K-S Critical Values:	0.1687	0.0821	0.0874	0.0936	0.1043
Test Result:	FAIL	FAIL	FAIL	FAIL	FAIL

H0 = spend is Normal.
(Normal if ACCEPT.)

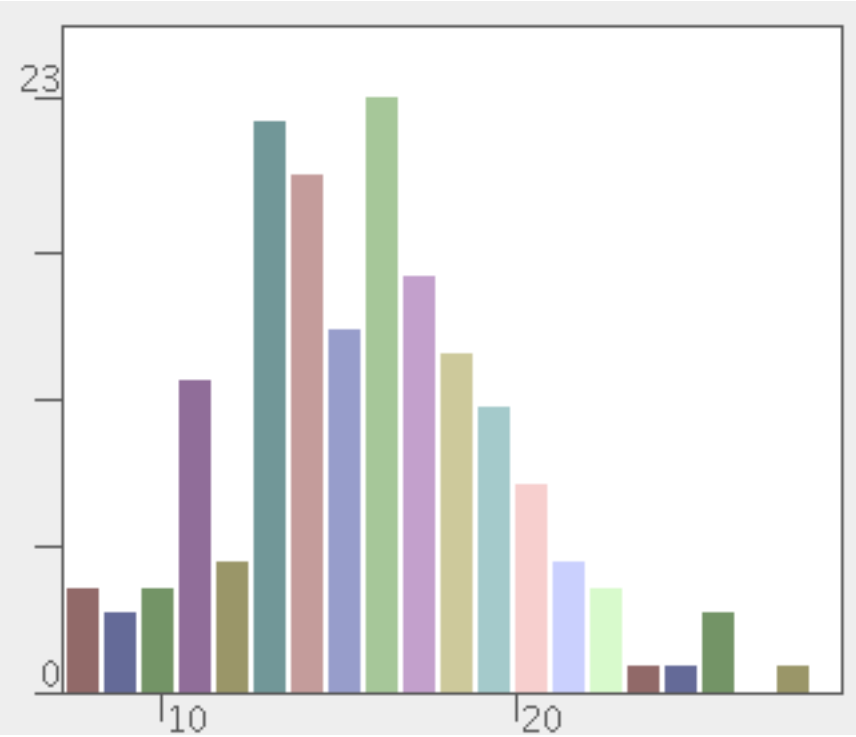
Anderson Darling Test:

A^2 = 8.4133
A^2* = 8.4510
p = 0.0000

H0 = spend is Non-normal.
(Normal if p > alpha.)

Based on the histogram and two statistical tests, the SPEND is not normal distribution. We can tell SPEND is right skewed by histogram.

1.6. STRAT



Normality Tests for: strat (sample size:170)

Kolmogorov Smirnov:

K-S Statistic: 0.0416381

Significance Levels:	.20	.15	.10	.05	.01
K-S Critical Values:	0.0416	0.0821	0.0874	0.0936	0.1043
Test Result:	ACCEPT	ACCEPT	ACCEPT	ACCEPT	ACCEPT

H0 = strat is Normal.
(Normal if ACCEPT.)

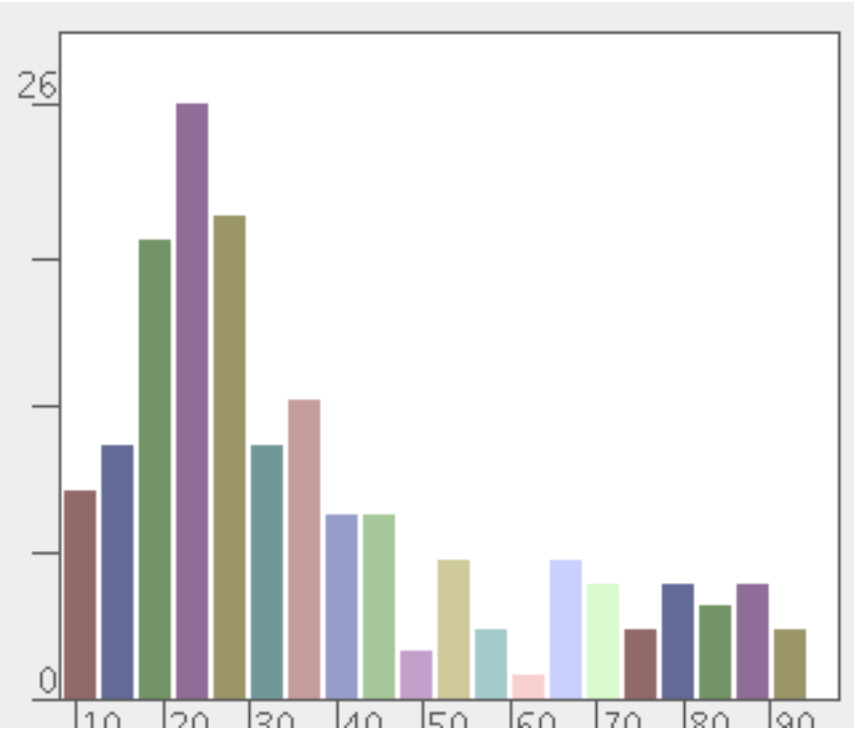
Anderson Darling Test:

A^2 = 0.3679
A^2* = 0.3695
p = 0.4266

H0 = strat is Non-normal.
(Normal if p > alpha.)

Based on the histogram and two statistical tests, the STRAT is normal distribution.

1.7. TOP10



Normality Tests for: top10 (sample size:170)

Kolmogorov Smirnov:

K-S Statistic: 0.1660248

Significance Levels:	.20	.15	.10	.05	.01
K-S Critical Values:	0.1660	0.0821	0.0874	0.0936	0.1043
Test Result:	FAIL	FAIL	FAIL	FAIL	FAIL

H0 = top10 is Normal.
(Normal if ACCEPT.)

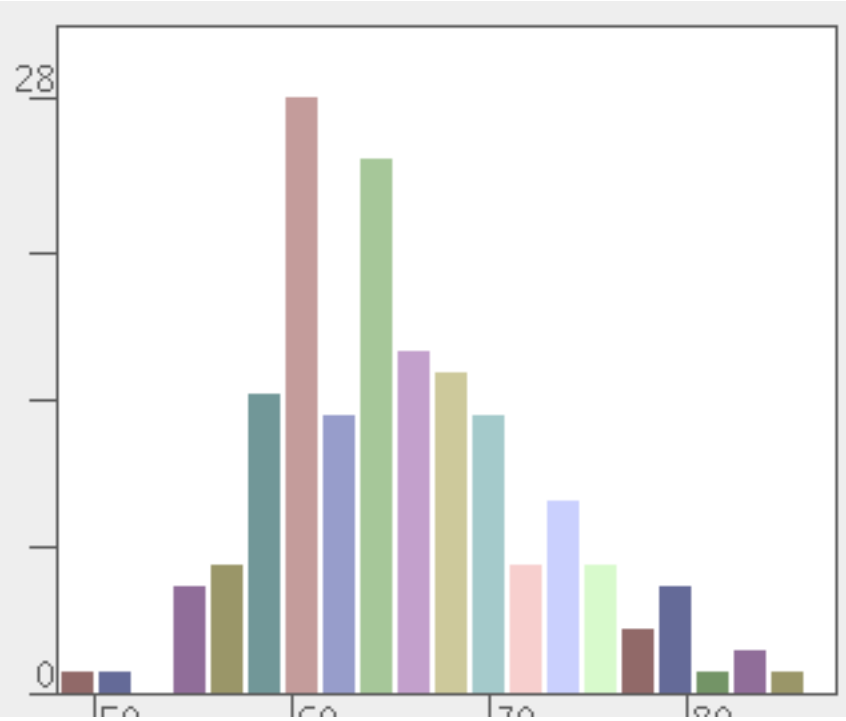
Anderson Darling Test:

A^2 = 7.9097
A^2* = 7.9452
p = 0.0000

H0 = top10 is Non-normal.
(Normal if p > alpha.)

Based on the histogram and two statistical tests, the TOP10 is not normal distribution. We can tell TOP10 is right skewed by histogram.

1.8. TSTSC



Normality Tests for: tstsc (sample size:170)

Kolmogorov Smirnov:

K-S Statistic: 0.0921306

Significance Levels:	.20	.15	.10	.05	.01
K-S Critical Values:	0.0921	0.0821	0.0874	0.0936	0.1043
Test Result:	FAIL	FAIL	ACCEPT	ACCEPT	ACCEPT

H0 = tstsc is Normal.
(Normal if ACCEPT.)

Anderson Darling Test:

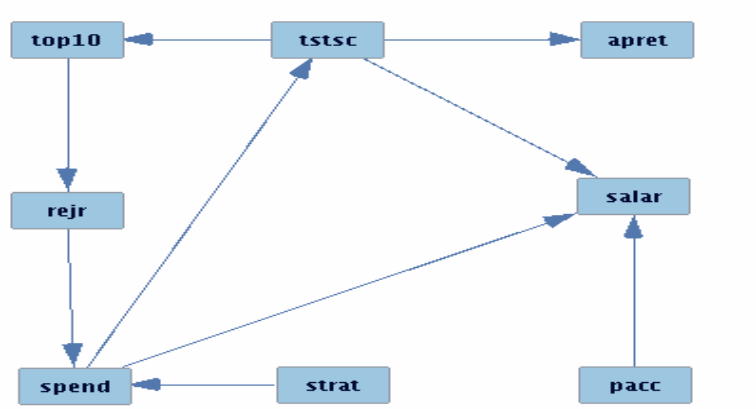
A^2 = 1.9165
A^2* = 1.9251
p = 0.0001

H0 = tstsc is Non-normal.
(Normal if p > alpha.)

Based on the histogram and two statistical tests, the TSTSC is not normal distribution when apply high significance levels. We can tell TSTC is a little bit right skewed by histogram

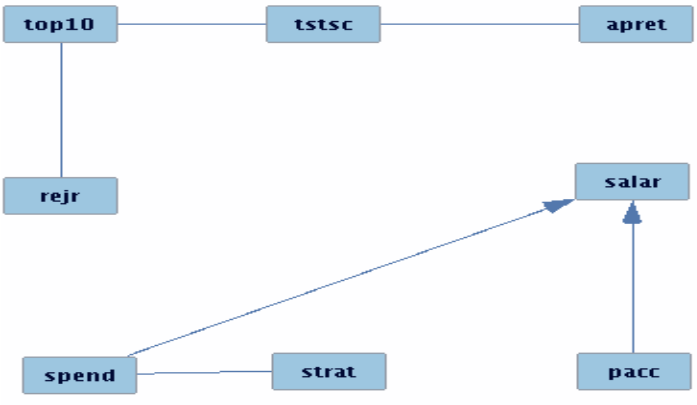
2. Causal Analysis

2.1. 0.02 cut off without knowledge



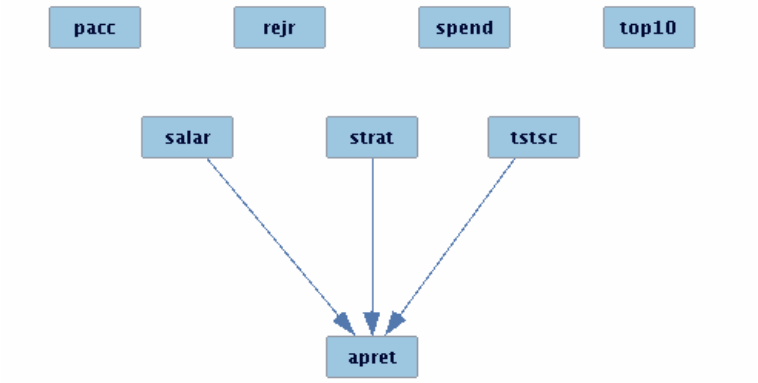
We can see that, the only direct causal of APRET is TSTSC, which is same as Professor Druzdzel's paper.

2.2. 0.001 cut off without knowledge



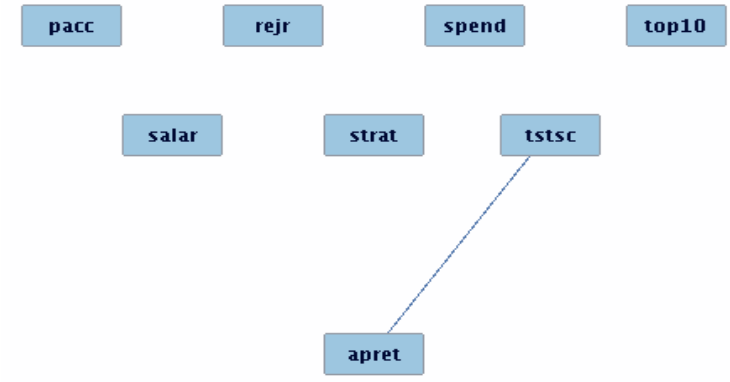
After decrease the significance level to 0.001, the number of causal relation decrease, too. However, there is still relation between TSTSC and APRET, which suggest that the relation between those two is strong.

2.3. 0.03 cut off with tier separated



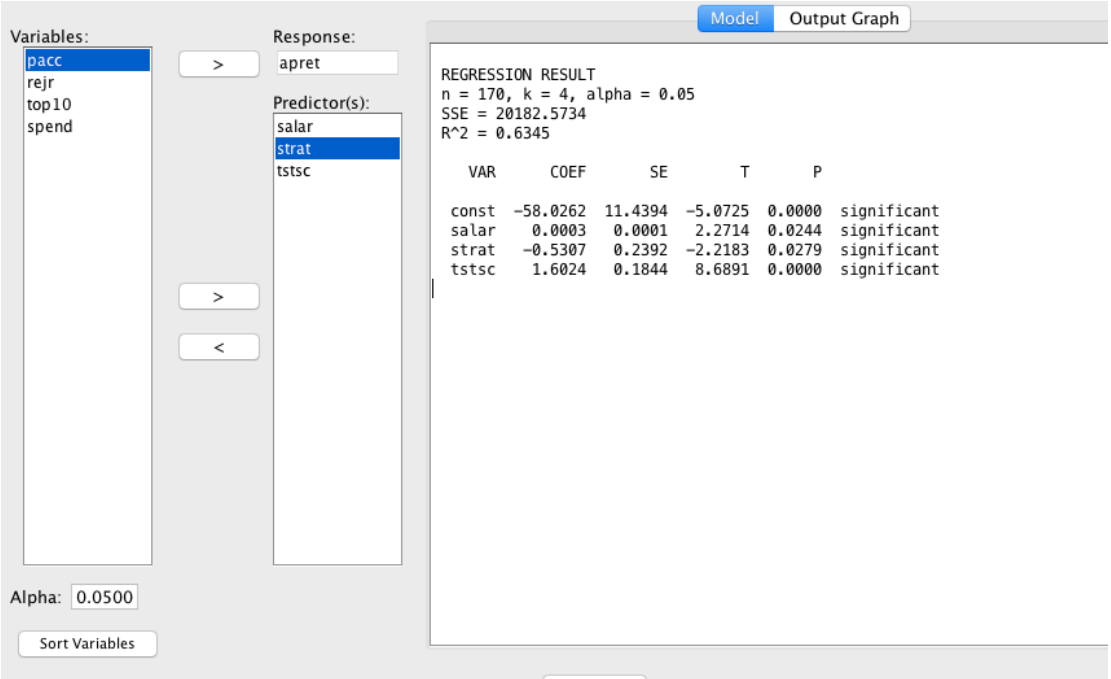
After putting the APRET in tier 2 and forbidding the relation within tier, we get this graph. It shows us that with 0.03 cut off, there are 3 features explain the APRET, SALAR, STRAT and TSTSC. The conclusion is different from Druzdzel.

2.4. 0.01 cut off with tier separated



When significance level decrease, TSTSC becomes the only feature again.

2.5. Multi-linear Regression



By using the 3 features selected by causal analysis, we get the Multi-linear regression model. The R-square is 0.6345 and all predictors are significant at 0.05 significance level.