Regression Analysis of Credit Accumulation at NJCU

Julian Garcia

FINC 665

4/27/2018

Executive Summary:

In my previous research predicting six year graduates here at NJCU I examined a large quantity of variables and their impact on graduation. From that research it was determined that the most important factor was credits accumulated by the end of a student's second year. For this linear regression analysis I wanted to look deeper into that data and analyze both pre-college variables and college performance variables impact on the accumulation of credits by the end of a student's second year. The pre-college indicators I use for this regression are High school percentile (hspctl) and SAT score (totnusat). The college performance indicators that will be used are cumulative GPA (CUMGPA), and accumulated credits (TermPassed) at the end of the first term.

Throughout this study I plan to explore and compare the variables and determine which combination impact credit accumulation in the most meaningful way. Six year graduation is one of the most widely reported statistics on student performance and lots of college rankings depend greatly on it. Learning more about the different impacts of certain variables that lead to graduation is imperative for Colleges and Universities to assist students in their careers and in turn yield results. Without informing ourselves on past information our strategies and plans are developed blindly.

The key components we assessed in this study are cumulative GPA, accumulated credits at the end of the first term in addition to high school percentile and SAT score. While some variables yielded higher levels of significance it was a combination that built the strongest model. This strongest model used all variables in addition to each variables interaction with the term credits passed. Accumulated credits assessed at these integral points in a student's career yielded the strongest impact and from past research it is evident that students who accumulate 60 credits by the end of their second year had put themselves in a very strong position to graduate within 6 years.

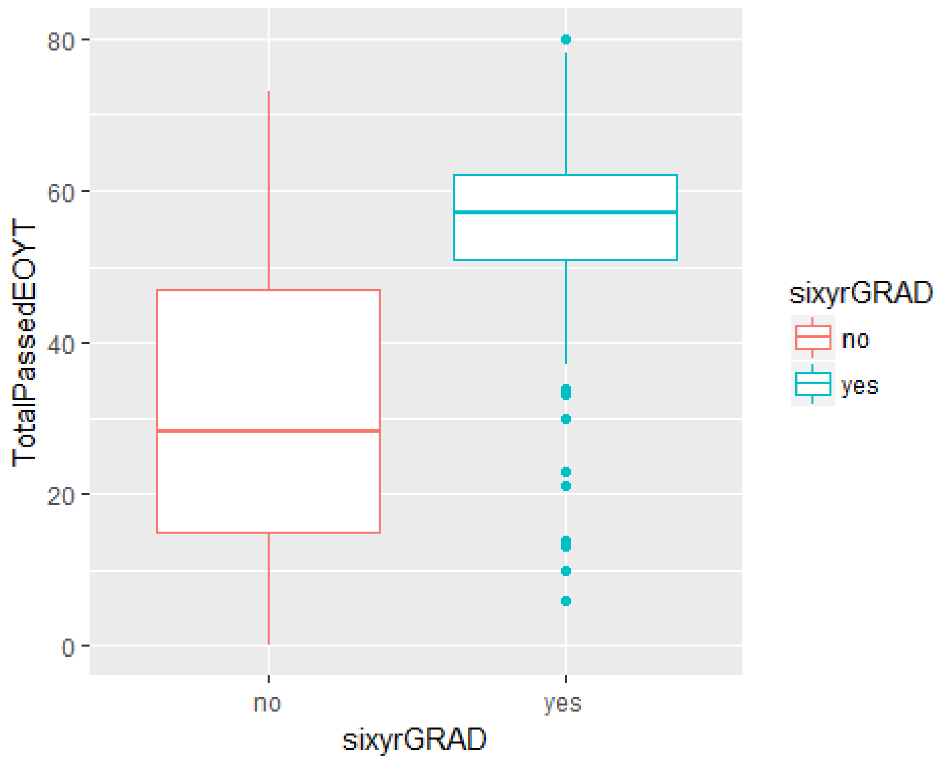
I believe from these findings, selecting students with strong pre-college characteristics is valuable but it is even important to closely monitor student college performance in both GPA and credit accumulation. Aiming to keep full time students on target to accumulate 15 degree credits per semester helps set a student's career on a path much more likely to graduate. Making it an integral part of the strategic plan and focusing on student persistence will not only improve the students experience and success but also bring greater recognition and prestige to the institution. Improving the overall number of full time first time students staying on target will in turn impact graduation rate in a positive manner.

Key Problems:

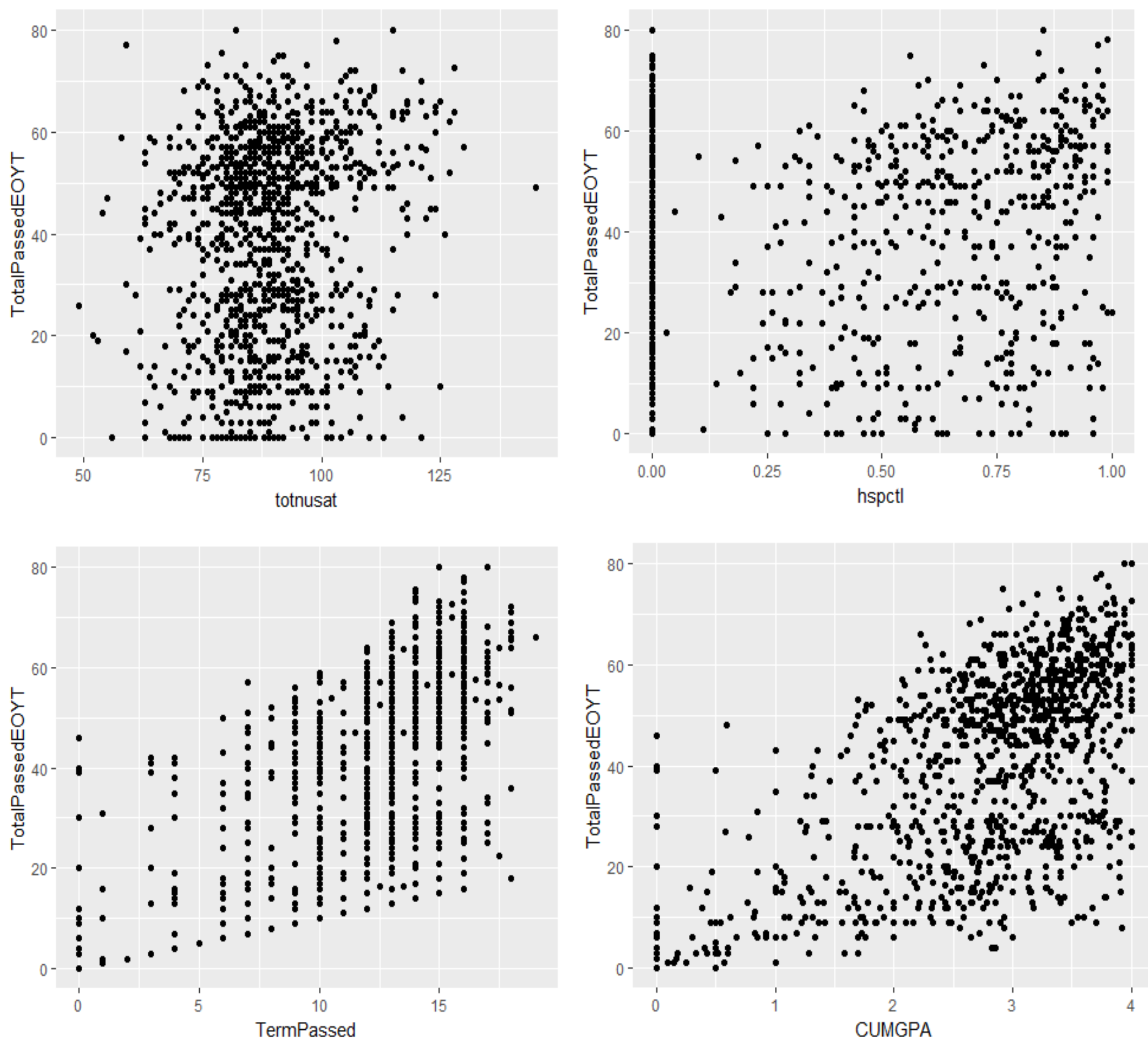
Six year graduation rates at NJCU are significantly lower in comparison to the national average. From previous analysis I have determined credit accumulation by the end of the second year to have the most important impact on graduation and plan to take a closer look at what impacts that variable. We want to determine if solely pre-college variables can predict 2nd year cumulated credits or if a combination of that and college performance variables are best.

Analysis and Evaluation:

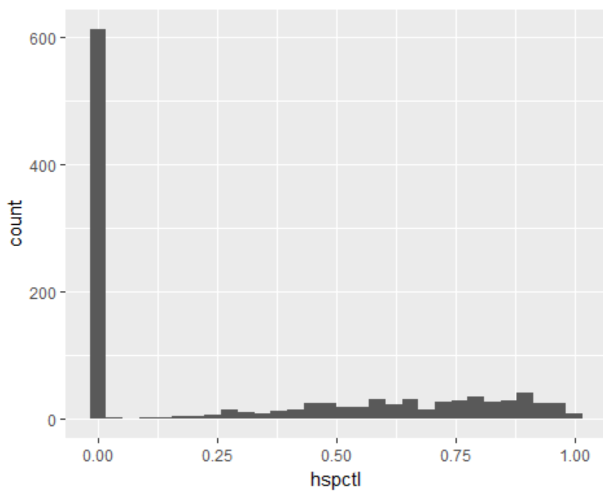
I wanted to use this boxplot to quickly illustrate the impact the dependent variable we are addressing in this analysis has on six year graduation. You can see the isolation in the two populations and how much tighter of a distribution of students that graduate exist near the 60 credit mark.

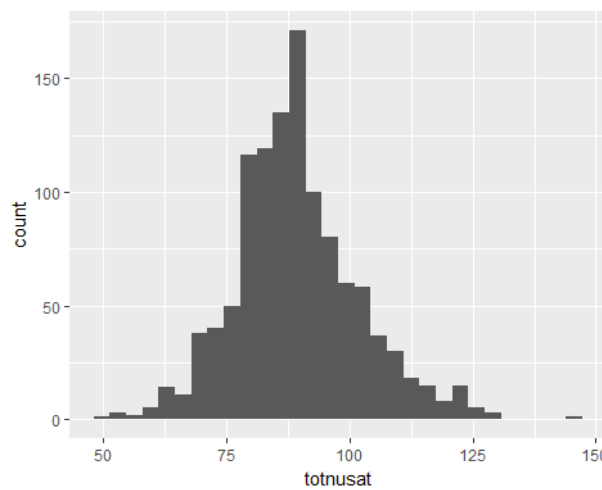


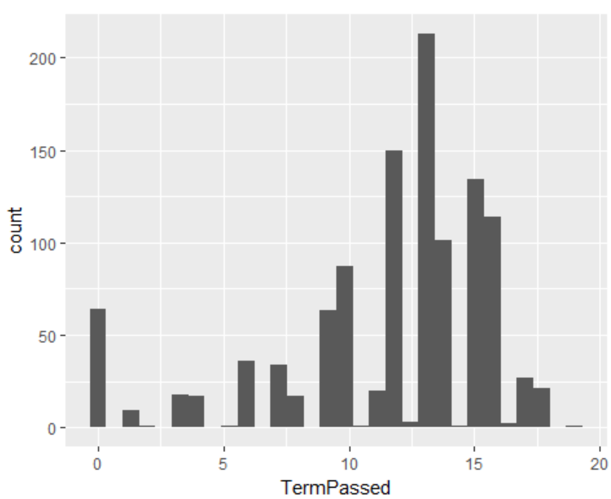
These scatter plots illustrate a general idea of the relationships each independent variable has with the dependent variable. The first 2 variables are pre-college indicators and show very little in terms of relation to the end of 2nd year credit accumulation. As we move into the performance indicators we can see movement start to take shape in the way the plot forms with dots clustering towards the upper right hand corner.

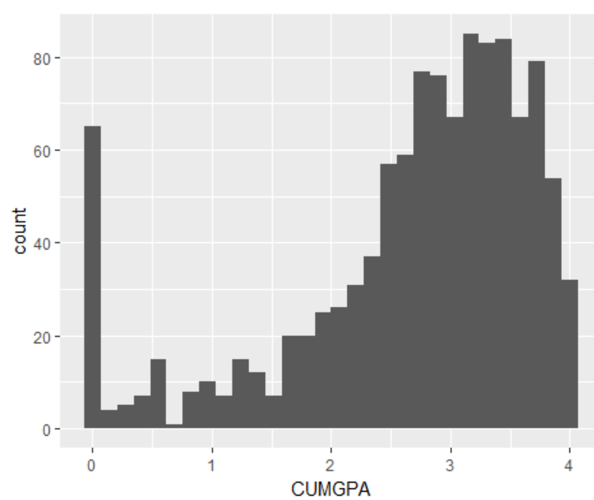


The histograms illustrated below give a general idea of the distribution of each variable. The first chart illustrates high school percentile which is relatively flat with an ever so slightly fatter upper tail. The second chart is SAT scores and has the most normal distribution of the charts. The final two charts dealing with term credits passed and cumulative GPA have much fatter upper tails.









The summary statistics displayed below show residual standard errors decrease from model 1 through model 4. While the error decreases accros models the r squared and adjusted r squared increases until we arrive at our final model using all variables in addition to their interactions with term credits passed. We will further see these implications as we move through individual model comparison through the use of anovas.

```
summary(model1)

##
## Call:
## lm(formula = TotalPassedEOYT ~ hspctl + totnusat, data = totalpassed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.667 -15.899   5.432  16.188  42.420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.10892    4.24753   4.028  6.0e-05 ***
## hspctl       4.15004    1.64574   2.522   0.0118 *
## totnusat     0.22788    0.04712   4.837  1.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.93 on 1132 degrees of freedom
## Multiple R-squared:  0.02692,    Adjusted R-squared:  0.0252
## F-statistic: 15.66 on 2 and 1132 DF,  p-value: 1.964e-07

summary(model2)

##
## Call:
## lm(formula = TotalPassedEOYT ~ hspctl + totnusat + CUMGPA, data = totalpas
sed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.341 -10.100   1.713  11.084  41.270
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.43222    3.28968  -1.347  0.17815
## hspctl       1.90421    1.24464   1.530  0.12632
## totnusat     0.09429    0.03586   2.630  0.00866 **
## CUMGPA      12.71746    0.43475  29.253  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.05 on 1131 degrees of freedom
## Multiple R-squared:  0.446,  Adjusted R-squared:  0.4446
## F-statistic: 303.6 on 3 and 1131 DF,  p-value: < 2.2e-16
```

```r
summary(model3)
```

```
##
## Call:
## lm(formula = TotalPassedEOYT ~ hspctl + totnusat + CUMGPA + TermPassed,
##     data = totalpassed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.668 -10.664   2.572  10.521  46.519
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.16264    3.10480  -1.663   0.0966 .
## hspctl       2.09716    1.17457   1.785   0.0745 .
## totnusat     0.03576    0.03419   1.046   0.2958
## CUMGPA       7.39293    0.60868  12.146   <2e-16 ***
## TermPassed   1.74329    0.14723  11.841   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.2 on 1130 degrees of freedom
## Multiple R-squared:  0.5072, Adjusted R-squared:  0.5054
## F-statistic: 290.7 on 4 and 1130 DF,  p-value: < 2.2e-16
```

```r
summary(model4)
```

```
##
## Call:
## lm(formula = TotalPassedEOYT ~ (hspctl + totnusat + CUMGPA) *
##     TermPassed, data = totalpassed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.966  -9.328   2.039  10.638  39.807
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.181025   8.125450   0.022 0.982230
## hspctl               4.199134   3.395603   1.237 0.216479
## totnusat             0.030533   0.091243   0.335 0.737962
## CUMGPA               3.324496   0.911585   3.647 0.000278 ***
## TermPassed           1.075744   0.664212   1.620 0.105602
## hspctl:TermPassed   -0.231226   0.269178  -0.859 0.390519
## totnusat:TermPassed -0.002857   0.007397  -0.386 0.699434
## CUMGPA:TermPassed    0.492185   0.082814   5.943 3.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14 on 1127 degrees of freedom
## Multiple R-squared:  0.5224, Adjusted R-squared:  0.5195
## F-statistic: 176.1 on 7 and 1127 DF,  p-value: < 2.2e-16
```

Model Comparison:

	RSE	R^2	Adjusted R^2
Model 1	19.93	0.0269	0.0252
Model 2	15.05	0.4460	0.4446
Model 3	14.20	0.5072	0.5054
Model 4	14.00	0.5224	0.5195

Model 2 differs from model 1 in that we introduce CUMGPA (one of the college performance indicators) in addition to solely using pre-college variables. We can reject model 1 for model 2 at the 5% and 1% level showing a large improvement.

```
anova(fit1,fit2)

## Analysis of Variance Table
##
## Model 1: TotalPassedEOYT ~ hspctl + totnusat
## Model 2: TotalPassedEOYT ~ hspctl + totnusat + CUMGPA
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1132 449811
## 2   1131 256069  1    193742 855.72 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 3 differs from model 2 in that we introduce another one of the college performance indicators (term credits passed). F = 140.21 and a p-value of 2.2e-16 we can reject model 2 for model 3 at the 5% and 1% level showing another large improvement and continuing to highlight the importance of college performance indicators in these models.

```
anova(fit2,fit3)

## Analysis of Variance Table
##
## Model 1: TotalPassedEOYT ~ hspctl + totnusat + CUMGPA
## Model 2: TotalPassedEOYT ~ hspctl + totnusat + CUMGPA + TermPassed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1131 256069
## 2   1130 227803  1     28265 140.21 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 4 differs from model 3 with the addition of all variables interaction with term credits passed. F = 11.986 has a p-value of 9.965e-08 we can reject model 3 for model 4 at the 5% and 1% level showing another large improvement and continuing to highlight the importance of college performance indicators in these models and in particular the term credits passed.

```
anova(fit3,fit4)
```

```
## Analysis of Variance Table
##
## Model 1: TotalPassedEOYT ~ hspctl + totnusat + CUMGPA + TermPassed
## Model 2: TotalPassedEOYT ~ (hspctl + totnusat + CUMGPA) * TermPassed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1130 227803
## 2   1127 220760  3    7043.8 11.986 9.965e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Citations

Mendenhall, William, and Terry Sincich. *A Second Course in Statistics: Regression Analysis*. 7th ed., Prentice Hall, 2012.

Wooldridge, Jeffrey M. *Introductory Econometrics: a Modern Approach*. 6th ed., Cengage Learning, 2016.