Zach Yek

STAT 351

Final Project

Understanding How Students Use Fredonia’s Tutoring Services

**Introduction**:

Throughout this experiment, we wish to understand which students (hereafter: tutees) make the most use of Fredonia’s tutoring services, measured as the total time spent per tutoring session. The variables of interest include:

1. GPA: tutee’s cumulative GPA at the time of recording; graded on a 4.0 scale.
2. Student-athlete status: whether the tutee is a student-athlete (binary variable; Yes, or No).
3. Class standing: tutee’s class standing at the time of recording (Freshman, Sophomore, Junior, Senior, or Grad Student).

Since the results of our experiment could potentially help Fredonia’s Learning Center offer more generally productive services, plus target their messaging towards the correct demographics, we will attempt to examine both the relationship between individual predictors and the response, as well as the collective relationship between predictors and response.

**Design**:

Beginning with the data collection phase, we worked with Fredonia’s coordinator of tutoring services – Adam Hino, to collect user demographic and usage data by having tutees sign in/out when entering/leaving the learning center. We initially had tutees fill out physical forms but switched to digital channels roughly 8 weeks into the semester to streamline turnover rates. Since we had not yet decided on what types of relationships we wanted to investigate at this stage, we decided to err on the side of caution and collect more data than was ultimately necessary (see Appendix B for a full list of variables recorded). Some care was reserved to ensure the data we collected were consistent with the Learning Center’s existing archives, such that we could incorporate legacy data into our analyses to capture temporal relationships as well. The names of tutees were recorded as per Fredonia’s Learning Center policies but were omitted from our analyses due to privacy concerns.

After acquiring the raw data, which included sessions as early as 08/17/2015 and as late as 04/30/2023, we began the data cleaning phase by importing the files into Python using Jupyter Notebooks and Pandas. We cleaned the dataset by:

* addressing any data entry errors (e.g. missing or duplicate values),
* engineering additional features (e.g. defining mins\_spent as time\_out - time\_in),
* performing ordinal encoding on the categorical variables relevant to our study (e.g. encoding student-athlete status as 0: not student-athlete, 1: student-athlete; class standing as 0: freshman, 1: sophomore, 2: junior, 3: senior, 4: grad student).

Once the data were in good shape, we exported everything into a CSV file for further study. Appendix C contains screenshots of the full notebook.

Finally, we initialized the data analysis phase by importing the cleaned data into RStudio. Then, with the parsimony principle for statistical models in mind, we picked the variables of interest:

* response : total time spent per tutoring session (in minutes)
* predictors : cumulative GPA, student-athlete status, class standing

All our variables are quantitative after the cleaning phase, though it’s worth noting that both student-athlete status and class standing are categorical in nature. Also of note, one could of course pick more variables to include in the study, at the expense of interpretability of results (or lack thereof). Subsequently, we performed 2 types of statistical tests:

1. t-test: for determining whether individual predictors are statistically significant with respect to the response
2. F-test (ANOVA): for determining whether the collection of variables are statistically significant with respect to the response

Based on our results, we constructed a regression model for predicting minutes spent in tutoring. Appendix A has all relevant RStudio outputs.

Only post-2020 datapoints were used to account for the effects of COVID-19.

**Results**:

Starting with GPA, our hypotheses are

* ; there is no relationship between GPA and minutes spent in tutoring.
* ; there is a relationship between GPA and minutes spent in tutoring.

Using the t-test, we find that

At a standard significance , since , we conclude that there is sufficient evidence to reject the null in favor of the alternative, implying that there is indeed a significant relationship between a tutee’s GPA and their minutes spent in tutoring.

Next, with student-athlete status, our hypotheses are

* ; there is no relationship between whether a tutee is a student-athlete and minutes spent in tutoring.
* ; there is a relationship between whether a tutee is a student-athlete and minutes spent in tutoring.

Using the t-test, we find that

At a standard significance , since , we conclude that there is not enough evidence to reject the null. This is indeed surprising, especially with the context that student-athletes are typically required to log a set number of study hall hours per week through the Learning Center. Further investigation is beyond the scope of this project, though this could serve as an enlightening follow-up study.

Moving on, with class standing, our hypotheses are

* ; there is no relationship between a tutee’s class standing and minutes spent in tutoring.
* ; there is a relationship between a tutee’s class standing and minutes spent in tutoring.

Using the t-test, we find that

At a standard significance , since , we conclude that there is sufficient evidence to reject the null in favor of the alternative, implying that there is indeed a significant relationship between a tutee’s class standing and their minutes spent in tutoring.

Subsequently, we fit the following multiple regression model, accounting for interaction terms by taking the product of all possible permutations of .

Our goal is to use ANOVA to determine if this collection of predictors is statistically significant. Thus, our hypotheses are

* ; there are no significant predictors for modeling minutes spent in tutoring
* ; there is at least one significant predictor for modeling minutes spent in tutoring.

Using the ANOVA table and model summary, we find that

* Adjusted-

At a standard significance , since , we conclude that there is sufficient evidence to reject the null in favor of the alternative, implying that the given model is statistically significant in predicting minutes spent in tutoring.

However, notice from the previous test’s outputs the individual F-statistics and corresponding p-values for each of the predictors.

Text

Description automatically generated

Specifically, we notice all terms involving student-athlete status are not statistically significant according to the F-test. This observation, along with our previous findings that student-athlete status in isolation is not statistically significant for predicting minutes spent in tutoring, leads us to believe an improved model would be one that includes only GPA and class standing as predictors. Thus, we fit the following improved multiple regression model,

where the same indices for are retained for ease of comparison, and perform the same ANOVA tests to verify whether there are any improvements. Our hypotheses are

* ; there are no significant predictors for modeling minutes spent in tutoring
* ; there is at least one significant predictor for modeling minutes spent in tutoring.

Using the ANOVA table and model summary, we find that

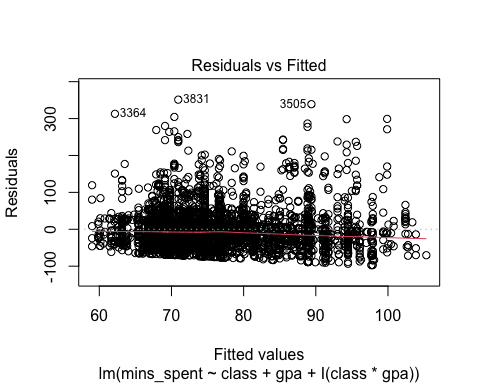
* Adjusted-

At a standard significance , since , we conclude that there is sufficient evidence to reject the null in favor of the alternative, implying that the given model is statistically significant in predicting minutes spent in tutoring. Moreover, at roughly the same adjusted-R-squared value, but almost double the previous F-statistic, we can also determine that the improved model is better at predicting minutes spent in tutoring than the previous one!

The final, improved model is thus

where GPA, and class standing.

It remains to verify the model assumptions via the appropriate residual plots.



From the residuals vs fitted values scatterplot, we notice the data are well distributed about the zero-line, with a few outliers above but not below. This, along with the fact that there are no clear signs of grouping/clustering, nor fanning/tailing, means the conditions of zero-mean and constant variance are satisfied.

Chart, line chart

Description automatically generated

From the normal Q-Q plot, we observe the model residuals are fairly well behaved, closely following the linear line with some tailing at the end, likely due to the outliers we observed from the previous plot. However, all things considered, the residuals look pretty normal, thus satisfying the condition of normality.

We also checked for multicollinearity via a correlation matrix.

Text

Description automatically generated

Aside from a slight positive correlation between GPA and class standing, there doesn’t seem to be any issues of note in our final model.

**Conclusion**:

Throughout this study, we’ve identified statistically significant predictors (out of an initial 3) for modeling minutes spent in tutoring. We then used those findings to construct a statistically significant regression model for predicting minutes spent in tutoring. While the tests involving GPA and class standing were consistent with expectations:

* The lower a tutee’s GPA, the more time they spent in tutoring
* The higher a tutee’s class standing, the more time they spent in tutoring

It came as a bit of a surprise that a tutee’s student-athlete status is not a significant predictor, not least for the fact that Fredonia’s student-athletes are required to log a set number of hours per week through the Learning Center. Some possible explanations are:

* the athletic departments stopped enforcing this policy post-COVID19
* student-athletes tend to check-in but *not* check-out of their sessions, causing their recorded minutes per session to appear nonsensically large (which would’ve been thrown out during the data cleaning phase)

Further work includes:

1. Investigating why student-athlete status is not a significant predictor of minutes spent in tutoring.
2. Investigating the causal relationships between GPA/class standing and minutes spent in tutoring.
3. Incorporating more predictors to see if we can improve the model’s predictive power. Alternatively, we could repeat the same analysis of statistical significance but for a different set of predictors.

Appendix A (RStudio code – data analysis):

library(readr)

## Warning: package 'readr' was built under R version 4.1.2

data <- read\_csv("data.csv")

## New names:  
## Rows: 18740 Columns: 12  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," chr  
## (6): visit\_location, visit\_purpose, college, major, ethnicity, professor dbl  
## (6): ...1, visit\_year, class, student\_athlete, gpa, mins\_spent  
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## • `` -> `...1`

head(data)

## # A tibble: 6 × 12  
## ...1 visit\_year visit\_l…¹ visit…² college major class ethni…³ stude…⁴ profe…⁵  
## <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl> <chr>   
## 1 0 2018 Reed Fou… Writing Libera… Engl… 2 White 1 Ploetz…  
## 2 1 2018 Reed Fou… Writing Libera… Engl… 2 White 1 FitzPa…  
## 3 2 2018 Reed Fou… Writing Libera… Engl… 2 White 1 FitzPa…  
## 4 3 2018 Reed Fou… Writing Libera… Engl… 2 White 1 FitzPa…  
## 5 4 2018 Reed Fou… Writing Libera… Engl… 2 White 1 FitzPa…  
## 6 5 2018 Reed Fou… Writing Libera… Engl… 2 White 1 FitzPa…  
## # … with 2 more variables: gpa <dbl>, mins\_spent <dbl>, and abbreviated  
## # variable names ¹​visit\_location, ²​visit\_purpose, ³​ethnicity,  
## # ⁴​student\_athlete, ⁵​professor

data\_subset <- subset(data, visit\_year >= 2020)  
attach(data\_subset)  
head(data\_subset)

## # A tibble: 6 × 12  
## ...1 visit\_year visit\_l…¹ visit…² college major class ethni…³ stude…⁴ profe…⁵  
## <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl> <chr>   
## 1 3101 2022 Reed Fou… Writing Libera… Comm… 2 Asian 0 Quatro…  
## 2 3102 2022 Reed Fou… Genera… Libera… Comm… 3 White 0 Cranda…  
## 3 3103 2022 Reed Fou… Genera… Colleg… Chil… 2 White 0 Conti,…  
## 4 3104 2022 Reed Fou… Writing Visual… Visu… 1 White 0 Noel, …  
## 5 3105 2022 Reed Fou… Writing Libera… Engl… 2 White 0 Johnst…  
## 6 3106 2022 Reed Fou… Genera… Libera… Appl… 3 White 0 Cheng,…  
## # … with 2 more variables: gpa <dbl>, mins\_spent <dbl>, and abbreviated  
## # variable names ¹​visit\_location, ²​visit\_purpose, ³​ethnicity,  
## # ⁴​student\_athlete, ⁵​professor

gpa\_model = lm(mins\_spent~gpa)  
summary(gpa\_model)

##   
## Call:  
## lm(formula = mins\_spent ~ gpa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -89.30 -32.66 -10.37 17.37 350.77   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.788 3.751 24.471 < 2e-16 \*\*\*  
## gpa -5.122 1.174 -4.363 1.32e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52.47 on 3891 degrees of freedom  
## Multiple R-squared: 0.004868, Adjusted R-squared: 0.004613   
## F-statistic: 19.04 on 1 and 3891 DF, p-value: 1.316e-05

athlete\_model = lm(mins\_spent~student\_athlete)  
summary(athlete\_model)

##   
## Call:  
## lm(formula = mins\_spent ~ student\_athlete)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -76.43 -32.87 -11.81 14.21 352.63   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 75.7884 0.8708 87.030 <2e-16 \*\*\*  
## student\_athlete 0.8101 3.4713 0.233 0.815   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52.6 on 3891 degrees of freedom  
## Multiple R-squared: 1.4e-05, Adjusted R-squared: -0.000243   
## F-statistic: 0.05446 on 1 and 3891 DF, p-value: 0.8155

class\_model = lm(mins\_spent~class)  
summary(class\_model)

##   
## Call:  
## lm(formula = mins\_spent ~ class)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -82.74 -32.32 -11.74 17.53 348.81   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68.902 1.672 41.21 < 2e-16 \*\*\*  
## class 3.571 0.744 4.80 1.65e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52.44 on 3891 degrees of freedom  
## Multiple R-squared: 0.005886, Adjusted R-squared: 0.00563   
## F-statistic: 23.04 on 1 and 3891 DF, p-value: 1.649e-06

model = lm(mins\_spent~class+student\_athlete+gpa+I(class\*student\_athlete)+I(class\*gpa)+I(student\_athlete\*gpa)+I(class\*student\_athlete\*gpa))  
summary(model)

##   
## Call:  
## lm(formula = mins\_spent ~ class + student\_athlete + gpa + I(class \*   
## student\_athlete) + I(class \* gpa) + I(student\_athlete \* gpa) +   
## I(class \* student\_athlete \* gpa))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -98.10 -31.62 -9.24 20.00 351.42   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 55.166 5.928 9.305 < 2e-16 \*\*\*  
## class 28.852 3.337 8.646 < 2e-16 \*\*\*  
## student\_athlete 71.268 26.655 2.674 0.00753 \*\*   
## gpa 3.786 1.947 1.945 0.05189 .   
## I(class \* student\_athlete) -42.785 16.013 -2.672 0.00757 \*\*   
## I(class \* gpa) -7.515 1.025 -7.332 2.74e-13 \*\*\*  
## I(student\_athlete \* gpa) -21.198 8.572 -2.473 0.01344 \*   
## I(class \* student\_athlete \* gpa) 12.407 4.985 2.489 0.01286 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 51.86 on 3885 degrees of freedom  
## Multiple R-squared: 0.02918, Adjusted R-squared: 0.02743   
## F-statistic: 16.68 on 7 and 3885 DF, p-value: < 2.2e-16

anova(model)

## Analysis of Variance Table  
##   
## Response: mins\_spent  
## Df Sum Sq Mean Sq F value Pr(>F)   
## class 1 63356 63356 23.5532 1.263e-06 \*\*\*  
## student\_athlete 1 1705 1705 0.6338 0.42602   
## gpa 1 97079 97079 36.0902 2.057e-09 \*\*\*  
## I(class \* student\_athlete) 1 3336 3336 1.2402 0.26550   
## I(class \* gpa) 1 130197 130197 48.4020 4.057e-12 \*\*\*  
## I(student\_athlete \* gpa) 1 1736 1736 0.6453 0.42183   
## I(class \* student\_athlete \* gpa) 1 16661 16661 6.1938 0.01286 \*   
## Residuals 3885 10450279 2690   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

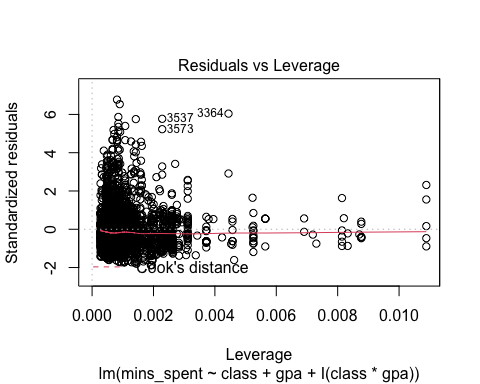
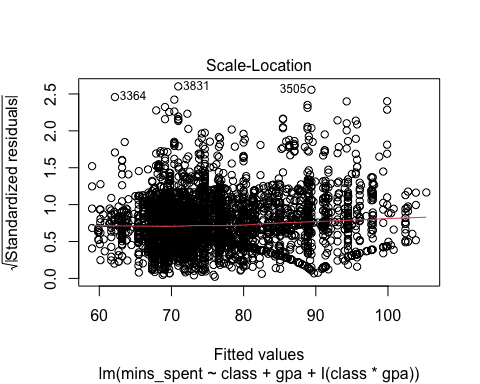
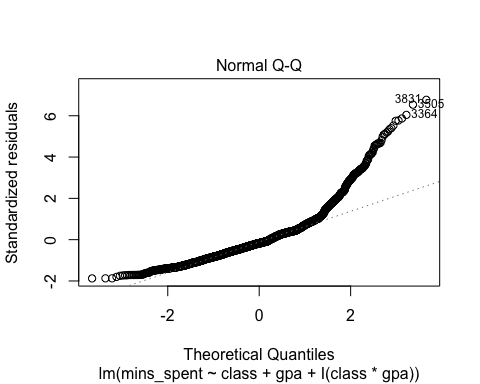
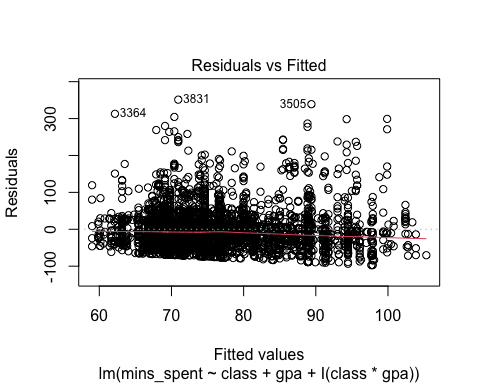
model\_improved = lm(mins\_spent~class+gpa+I(class\*gpa))  
summary(model\_improved)

##   
## Call:  
## lm(formula = mins\_spent ~ class + gpa + I(class \* gpa))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -97.24 -31.76 -9.28 20.19 351.29   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 58.436 5.776 10.117 < 2e-16 \*\*\*  
## class 26.963 3.263 8.262 < 2e-16 \*\*\*  
## gpa 2.827 1.894 1.492 0.136   
## I(class \* gpa) -6.973 1.002 -6.960 3.99e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 51.89 on 3889 degrees of freedom  
## Multiple R-squared: 0.02712, Adjusted R-squared: 0.02637   
## F-statistic: 36.14 on 3 and 3889 DF, p-value: < 2.2e-16

anova(model\_improved)

## Analysis of Variance Table  
##   
## Response: mins\_spent  
## Df Sum Sq Mean Sq F value Pr(>F)   
## class 1 63356 63356 23.528 1.280e-06 \*\*\*  
## gpa 1 98179 98179 36.459 1.705e-09 \*\*\*  
## I(class \* gpa) 1 130431 130431 48.437 3.986e-12 \*\*\*  
## Residuals 3889 10472382 2693   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(model\_improved)



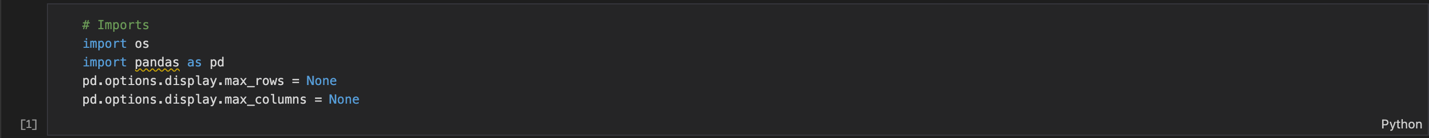
data\_subset\_selected <- data\_subset[, c("mins\_spent", "class", "gpa")]  
cor(data\_subset\_selected)

## mins\_spent class gpa  
## mins\_spent 1.00000000 0.07671838 -0.06977339  
## class 0.07671838 1.00000000 0.28408433  
## gpa -0.06977339 0.28408433 1.00000000

Appendix B (Full list of variables recorded):

* GPA: tutee’s cumulative GPA at the time of recording; out of 4.0 scale.
* Student-athlete status: whether the tutee is a student-athlete; binary variable (Yes, or No).
* Class standing: tutee’s class standing at the time of recording (Freshman, Sophomore, Junior, Senior, or Grad Student). Undergraduates in their fifth year or more were labeled as Seniors.
* College: tutee’s primary college (Liberal Arts and Sciences, School of Business, etc.)
* Major: tutee’s primary major
* Ethnicity: tutee’s identified ethnicity (White, Black, Asian, Hispanic, etc.)
* Professor: name of professor associated with tutee’s coursework; recorded in the format “LastName, First Name”
* Visit purpose: purpose of visit (General Tutoring, Writing, Language Support, etc.)
* Visit location: location of visit (4th floor Reed Library, Mason Hall)
* Time in: timestamp when tutee checked in
* Time out: timestamp when tutee checked out
* Date: date of session

Appendix C (Python code – data cleaning):

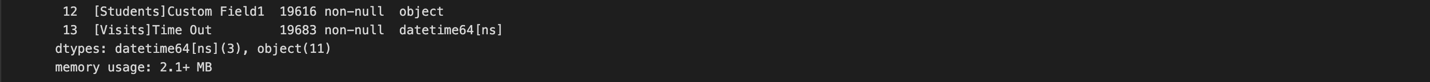


Text

Description automatically generated

Text

Description automatically generated



Text

Description automatically generated



Text

Description automatically generated

