Analyzing Data from the MIDFIELD Project

Javier Alvarez, Leah Hunt, and Joshua Kwak

1 Introduction

1.1 Background

The Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) is a database that contains population data for 2.2 million undergraduates at 20 higher education institutions with engineering programs. MIDFIELD contains a substantial amount of information about students including their race, gender, high school GPA, degree major, and much more. By using the MIDFIELD database to identify patterns and trends as well as find evidence for statistical research questions, we gain a better understanding of student performances in school and encourage other researchers to conduct studies using the database. MIDFIELD continues to work towards improving data validation methods as well as expanding the number of institutions included in the database. [5]

1.2 Statistical Research Question

In this observational study, we are interested in looking at how the proportion of women in STEM has changed over the years by school. To do this, we will split the time period of 1994 – 2003 by year and observe the differences in the proportion of women pursuing STEM degrees by school. This makes ANOVA a perfect tool to use for this research question because we can compare the sample arithmetic means of each of those time periods and identify if there is a relation between the proportion of women in STEM and school attended. Put formally, the statistical research question is: How has the proportion of women pursuing STEM degrees changed over time?

Additionally, we want to observe how race and gender affects how well the course Calculus I acts as a "weed-out" class. Namely, how does race and gender affect the rate of students deciding to quit pursuing a STEM field and pursue something different after taking Calculus I? To do this, we will look at the number of students who have dropped out of a STEM degree shortly following Calc I as compared to the total number of students who have taken Calc I. By using a logistic regression model, we will model the probability of a student dropping out of a STEM degree using race and gender as predictors.

2 Exploring the Change in the Proportion of Women Graduates in STEM Majors over Time

Our first analysis will be exploring is whether the proportion of women pursuing STEM degrees has change throughout the years. In order to do this, we will begin by discussing our study design strategies.

2.1 Methods

2.1.1 Study Design

In our study, the response will be the proportion of women graduating with STEM degrees, and the measurement units will be 10 institutions given in the MIDFIELD database. Because the treatments are applied to the institutions, the 10 institutions are also the experimental units. The factor is each year between 1994 and 2003 which means there are 10 treatment levels. We chose the time period 1994 – 2003 because the availability of data was best during this period and would produce the most accurate information about the differences between school years. Because we are using an in-subject repeated measures design, the block will be the same as the subject, the institution.

The institutions that we got the data about the proportion of women in STEM from were randomly chosen for each year. Each institution randomly selected a subset of its population, and we measured the proportion of women in that population. It is relevant to note that the number of institutions measured differs per year so there is not an even distribution of data points. Although the institutions were randomized to account for this fact, it does provide a potential source of error/inaccuracy of the model in our ANOVA analysis.

2.2 Discussion of Models

The Hasse Diagram is given in Figure 1. We are using the school year as the treatment while blocking on institutions.

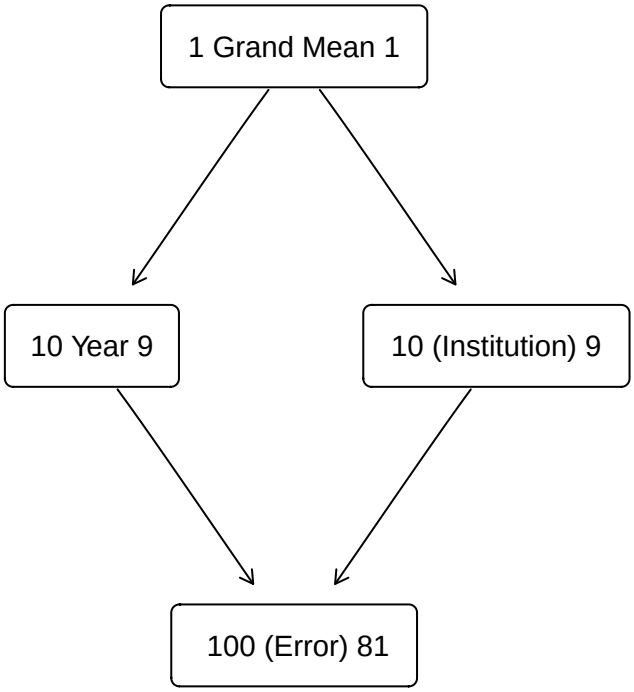


Figure 1: Hasse Diagram

As explained above, we will use ANOVA to conduct the analysis. After analyzing the Hasse diagram found in Figure 1, we decided the best model to use for our study is a One Way Repeated Measures (Within-Subjects) ANOVA Model. This method is effective because we can observe if there are statistically significant

differences in the sample arithmetic means between each school year. This will also allow to reduce the amount of subject-to-subject variation present within our model by applying each "year treatment" to each individual institution. If we were to find evidence to reject the null hypothesis that there is no statistically significant difference between the sample arithmetic means, we could then conduct post-hoc analysis to determine if later years caused the proportion of women in STEM to increase. The null and alternative hypothesis are formally expressed below.

INSERT HYPOTHESIS CORRECTLY

2.2.1 Type I Error Rate

We have decided that we will control Experimentwise Error Rate at 10% ($\mathcal{E}_I = 0.10$) for both components of our study (ANOVA and Logistic Regression portions). Given the low-stakes nature of our study, we believed it was reasonable to use a more liberal Type I Error control method. Having a Type I Error would not cause a significant impact nor incur overwhelming costs, so having a low evidentiary requirement is justified. The EER will be naturally controlled through the use of our ANOVA F Test of our model. We also plan to set our unusualness threshold at the same level as our Type I Error Rate ($UT = 0.10$).

2.3 Exploring the Data

2.3.1 Attributes

There are two attributes we are concerned about in this study: the time periods in which we are interested and the proportion of women pursuing STEM. Because the question we are interested in is the change in the proportion of women over time, graphing the data will provide some general useful insights.

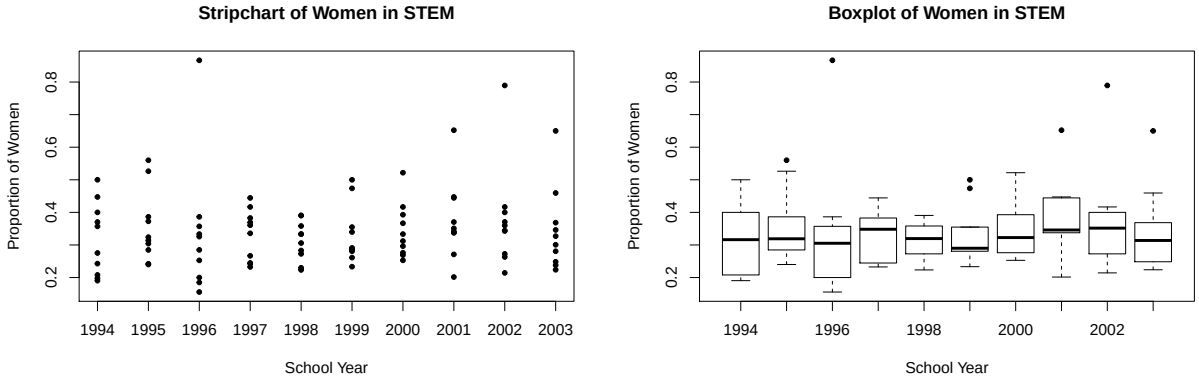


Figure 2: Stripchart and Boxplot of Proportion of Women over Time

Figure 2 gives plots of the data for all institutions over the entire time period. Both plots appear to indicate that he majority of institutions had a proportion of women between 0.2 – 0.5. Given that, it does not appear that there are any significant changes between years. While it does look like there are some differences between years, there is no clear upward/downward pattern that the data displays as the years increase. This would suggest that as time went on, the proportion of women in STEM had not increased significantly during the period of 1994 – 2003 and that there are not statistically significant differences in the sample arithmetic means between the years. ANOVA analysis will provide greater insight into these details.

Table 1: Statistics for MIDFIELD Data

Year	Sample.Max	Q3	Sample.Median	Q1	Sample.Min	SAM	SAV	Sample.Skewness
1994	0.500	0.393	0.316	0.217	0.190	0.319	0.012	0.218
1995	0.560	0.383	0.319	0.290	0.240	0.355	0.012	0.753
1996	0.867	0.351	0.305	0.213	0.156	0.335	0.041	1.674
1997	0.444	0.379	0.348	0.250	0.233	0.330	0.006	-0.003
1998	0.391	0.352	0.320	0.275	0.223	0.312	0.004	-0.119
1999	0.500	0.351	0.290	0.282	0.233	0.331	0.008	0.859
2000	0.522	0.386	0.323	0.281	0.253	0.344	0.007	0.799
2001	0.652	0.426	0.346	0.338	0.201	0.375	0.015	0.860
2002	0.789	0.393	0.352	0.290	0.214	0.377	0.025	1.593
2003	0.650	0.363	0.314	0.257	0.224	0.344	0.016	1.223

Table 1 gives the values of various descriptive statistics for the proportion of women pursuing STEM degrees by time periods. From the table, we can observe that the sample arithmetic means range from 0.31 – 0.38 which indicates that the means are very similar to each other. The year with the lowest proportion of women in STEM is 1998 and the year with the highest proportion is 2002. The sample arithmetic variance is also very small across all the years. The sample skewness values vary across the years which indicate that there may exist some outlier cases. The graphs shown above outline some of the years with outlier cases.

2.4 ANOVA Assumption Testing

In order to properly make use of our One Way Repeated Measures (Within-Subjects) ANOVA Model, we must first check that our data passes the proper assumption for such a model. These assumptions are the normality of our residuals as well as our random effects, homoscedasticity of our residuals, as well as a lack of interaction between our subjects. Since we are also expecting a lack of independence between our observations, we will be using Mauchly's test of Sphericity to compensate for such violations. Upon checking the assumptions of our original data, we noticed many issues regarding the normality and homoscedasticity within the data, and decided to use a log-transformation of our response variable in order to adjust to our model [1].

[1] An inclusion of the assumption visualizations for the original data along with a detailed explanation has been included within Appendix B

2.4.1 Normality of Residuals and Random Effects

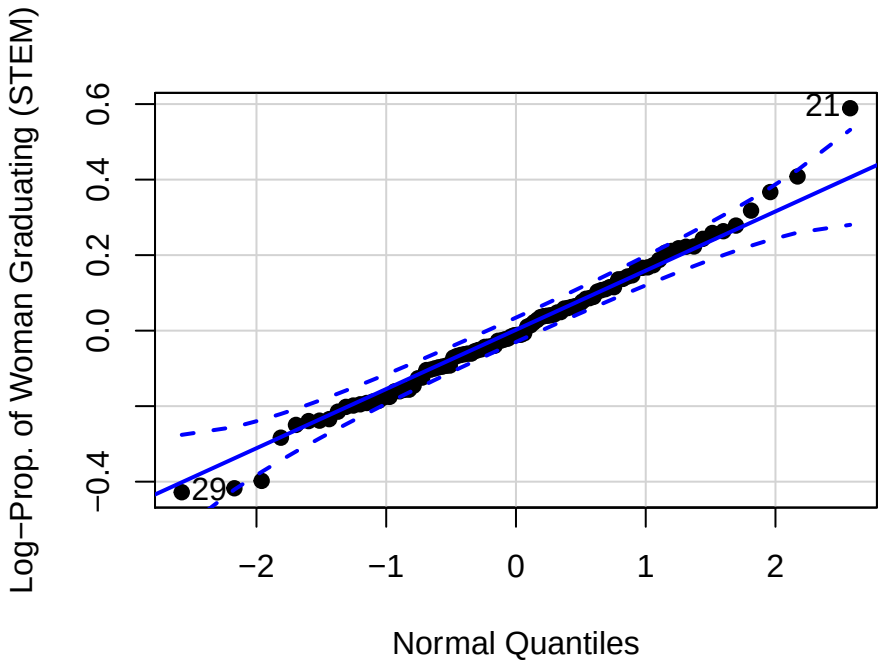


Figure 3: Normal Quantile Plot (Residuals) for Log-Transformed MIDFIELD Data

Looking at the Normal Quantile Plot of the residuals for our log-transormed data seen in Figure 3, we see that our residuals exist mostly within our 90% confidence envelope. There do appear to be a couple of value either on the edge or just outside of the confidence envelope, which could represent slight potential outliers within the data. However, these values appear to deviate very slightly from a normal distribution, thus we will assume that our residuals appear to be normally distributed.

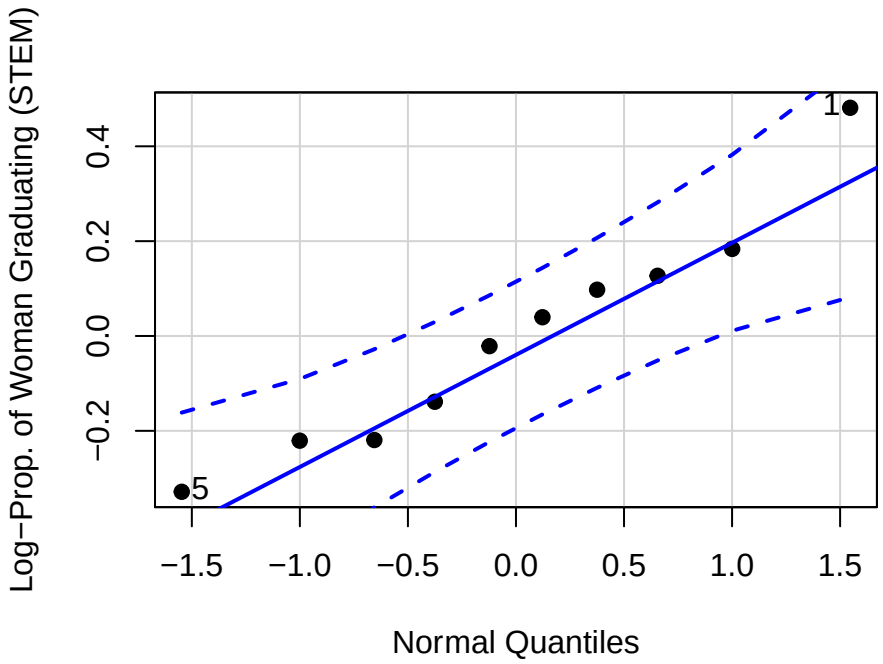


Figure 4: Normal Quantile Plot (Random Effects) for Log-Transformed MIDFIELD Data

Within Figure 4 above, we are given an visualization of the Normal Quantile plot for our random effects (our institutions). Within the plot, all values appear to be within the 90% confidence envelope. There appears to be no presence in potential outliers within the plot. Thus, our random effects appear to be normally distributed, and we can assume that normality of our residuals and random effects have been satisfied.

2.4.2 Homoscedasticity of Residuals

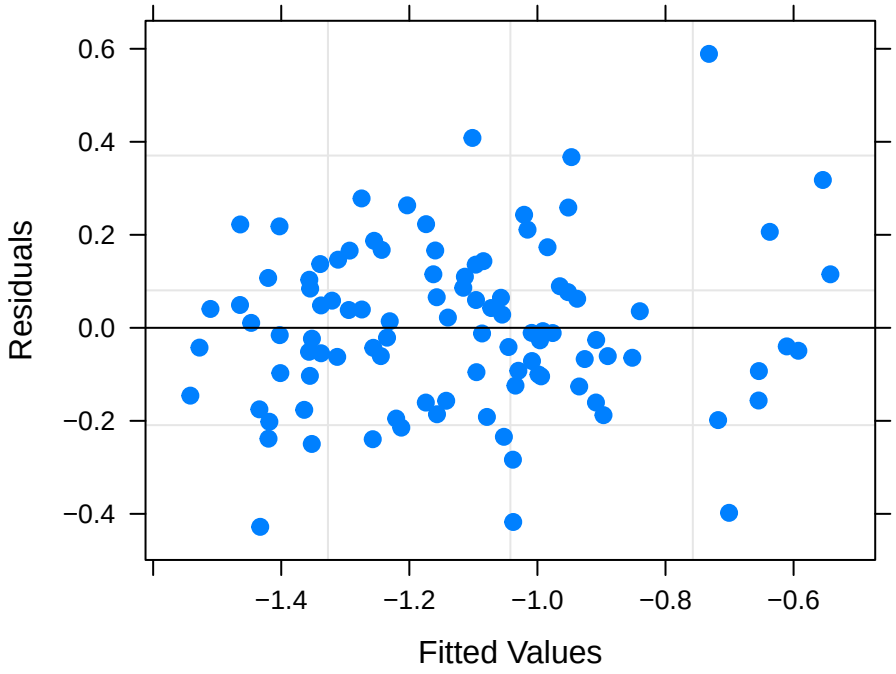
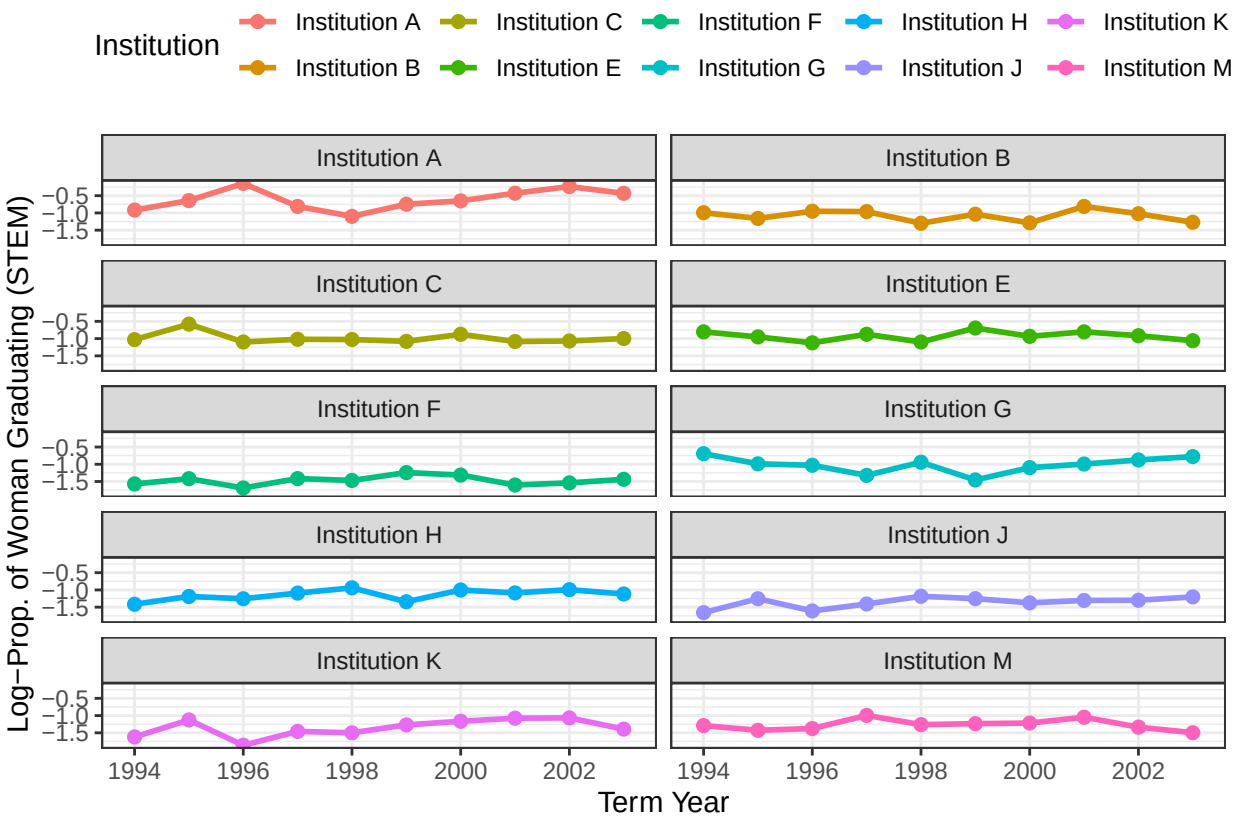


Figure 5: Tukey-Anscombe Plot for Log-Transformed MIDFIELD Data

In order to interpret the homoscedasticity of our residuals, we turn to the Tukey-Anscombe plot within Figure 5. Upon our initial analysis, there appears to be no fanning present within our plot, which is a good indicator that homoscedasticity is present. We can also roughly see that their is no large difference in the vertical spread of our residuals. Thus, we will decide that our assumption of normality has been satisfied.

2.4.3 Interaction



In order to use our ANOVA model, we must check for a lack of interaction between our subjects. Using the interaction plots above, we are given the individual plots for the log-proportion of women graduating in STEM by term year. Within the plot, we see potential interactions present such as between the years 1995 and 1996 for institution's A and G. However there appear to be few visual cues of interactions present, so we will proceed with caution and assume that there is no interaction present.

2.4.4 Sphericity

Table 2: Mauchly's Test for Sphericity

	Test Statistic	p-value
Term_Year	0	0.019

Note: Computer rounding has made test statistic appear to be zero.

Lastly, we will be using the results from Mauchly's test of Sphericity in order to determine if our repeated measures are compoundly symmetrical. Upon analysis of Table 2 above, we see that we acheived a p-value of 0.019. Comparing our p-value to our Type I Error Rate of 0.1, we determine that our p-value is much less than our error rate. THus, we will decide that our data violates the assumption of sphericity. We will proceed with our omnibus testing, however we will include adjustments methods to combat our violation of sphericity.

2.5 Omnibus Results

Table 3: ANOVA Table Within-Subjects: Log of Proportion of Women Graduating in STEM

	Numerator DF	Denominator DF	Term SS	Error SS	F Ratio	p-value
(Intercept)	1	9	125.639	5.820	194.279	0.000
Term_Year	9	81	0.386	3.011	1.155	0.335

Note: Computer rounding has made the p-values look like zero.

Looking at the results or our Repeated Measures ANOVA Model above in Table 3, we see that our term year achieved an F Ratio of 1.155. In other words, our term year factor accounts for around 1.15 times the variation as our error term. We would most likely expect to achieve an F Ratio such as this 33% (p-value = 0.335) of time under our model's null hypothesis, which is a considerable amount of time. Comparing our p-value with our unusualness threshold ($UT = 0.10$), we can see that our p-value is not less than our threshold, which in turn would result in the decision of failing to reject our model's null hypothesis. However, due to the issues with sphericity that we noted earlier, we must adjust our model to account for our violations.

Table 4: Adjustments for Sphericity Violations

	Greenhouse-Geiser	p-value	Huynh-Feldt	p-value
Term_Year	0.498	0.347	1.052	0.335

Note: Huynh-Feldt eps is being treated as 1.

In order to handle our violations of sphericity, we look upon the Greenhouse-Geiser and Huynh-Feldt adjustments of our p-value. Table 4 provides us with these adjustments. Our p-values using Greenhouse-Geiser and Huynh-Feldt adjustments are 0.347 and 0.335, respectively. Even when we use these adjusted p-values, we still see that our p-values are much greater than our unusualness threshold ($UT = 0.10$). Thus, we will fail to reject our null hypothesis and continue to act as if there is no significant difference in the log proportion of women graduating with STEM degrees in our collection of institutions due to the term year. In other words, we interpret that there is not a significant increase or decrease in the proportion of women graduating in STEM fields throughout our selected time period.

3 Exploring the Effect of Race and Gender on Drop Out Rates after Introductory Calculus

The second portion to our analysis considers the role of the introductory calculus course as a weed out course for STEM majors. Introductory calculus is a requirement for nearly all STEM majors and is one of the most commonly cited reasons for students to leave the STEM major. [3] While works such as that done by Rasmussen and Ellis have considered factors impacting the probability of continuing the calculus curriculum after the first calculus course, we wish to consider factors that impact students' probability of leaving the STEM major after introductory calculus, in particular to see how the probability varies by race and gender.

To consider this problem, we use the MIDFIELD data to identify STEM students who took Calculus I and label them as either having dropped out of the STEM field in association with Calculus I or not having dropped out of the STEM field in association with introductory calculus.[2] We then develop a logistic

[2]A more detailed description of the data wrangling techniques and the methods by which the drop out variable was created can be found in Appendix A.

regression model to model each students' probability of dropping out after introductory calculus based on their race and gender in order to test our hypothesis that at least one of race and gender affect a student's probability of dropping out of the STEM program after introductory calculus against the null hypothesis that neither is a significant factor.

3.1 Exploratory Analysis

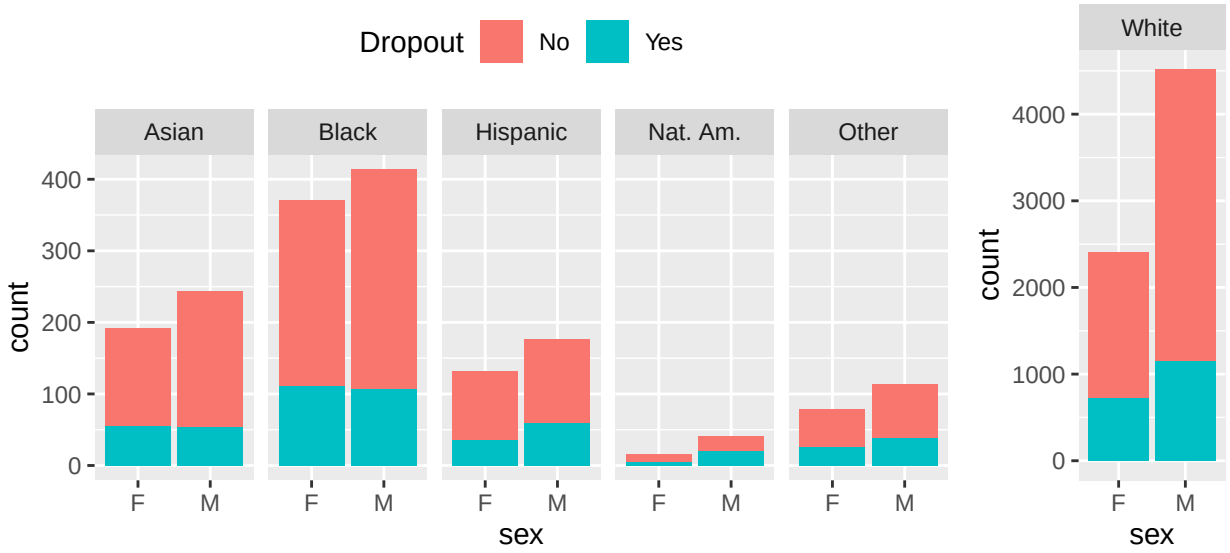


Figure 6: Dropouts by race and gender

We begin by visualizing the collected dropout data as seen in Figure 6. We see that we are treating gender as a factor, for which we will assume only two levels. Race is also a factor, which we will treat as having six levels: White, Asian, Black, Hispanic, Native American, and Other Minorities. Note that the other minority group also contains the category international, which was included in the MIDFIELD dataset but regrouped to other minorities for this analysis. It is also worth considering the wide variety in sample size between the different categories, with White having an overwhelming majority and Native American having only barely enough to be considered as an individual category.

Considering this visualization relative to our subject, we see that females and several minority groups appear to be overall underrepresented in this data. Females also appear to generally have higher numbers of dropouts relative to population size for most races.

3.2 Assumption Testing

Logistic regression requires five assumptions to be satisfied: appropriate outcome structure, independence of observations, absence of multicollinearity, linearity of independent vars and log odds, and sufficiently large sample size. [4] The first and last of these are clearly met as we designed the dependent variable to be binary and the size of the sample considered is over 8,000 students.[3] We do conceed, however,that some individual groups, such as female native americans, do have a somewhat small sample size within this large group, so this conclusions drawn for such groups must be closely ridiculed.

[3]The full MIDFIELD dataset contains over 97,000 students. The smaller number used in the analysis is a result of filtering the students to STEM students who were identified as having taken an introductory calculus course.

Considering the linearity of independent variables relative to log odds, we compare the dropout rate of each individual group in the below plot.

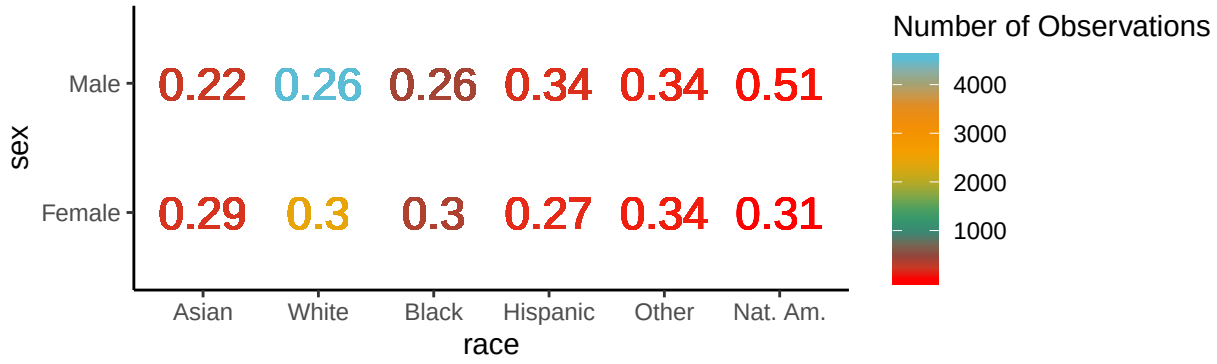


Figure 7: Dropout rates compared by race and gender

We see that for the three largest racial groups females have higher dropout rates than males with other minorities as equal and the last two categories having the reverse trend. Likewise, for most pairs of races, similarities and differences in proportion dropping increase or decrease consistently for males and females with the Hispanic and Native American groups again providing the only mild exceptions, for example comparing either to the Other Minorities category. This implies that there could be some mild issues with the assumption of the independent variables' linearity with log odds. We will, however, still proceed with the analysis noting that these mild issues primarily occur among the least common groups in the population and that any results of the analysis pertaining to these smaller groups will need to be done cautiously.[4]

To consider multicollinearity, we can use the variance inflation factor (vif). The calculated values for vif were 1.009 for each predictor, which is less than 4, the commonly used indicator of collinearity problems. Therefore, we may proceed under the assumption that this assumption holds.

The last assumption, and most tricky to consider, is the independence of observations. While each student's measurements are taken individually without regard to any other student in the study, our data does not give any way for us to know that students, particularly in the same institution at the same time period, interact with or influence each other in a meaningful way. However, it is reasonable to assume that if these influences exist, they would be negligible, particularly considering that we are considering students' choices in educational field which would rarely be impacted by infuences from the choices of the other student.

Therefore, aside from a small linearity issue that we will continue to acknowledge, we decide that all of our assumptions have been sufficiently met, so we may proceed with our analysis.

3.3 Developing the Model

In developing the logistic regression model, we will use a baseline of a white male. The resulting model is shown in the below table.

[4]Particularly considering the smaller sample size in the Native American and to a lesser extent Hispanic groups, it may seem reasonable to merge one or both of these groups into the Other Minorities category. A condensed summary of the analysis making both of these choices is shown in Appendix C.

Table 5: Odds Ratios Relative to white men

	Odds Ratio	P-value
Female	1.217	0.000
Asian	0.888	0.298
Black	1.012	0.883
Hispanic	1.196	0.157
Native American	2.283	0.002
Other Minorities	1.356	0.049

Based on our test and the p values it produced, the only significant terms are gender, the Native American race, and Other Minority races. The p value for the impact of race rounds to 0 in the table, implying that assuming the null hypothesis we would expect to see results at least as extreme as we did nearly none of the time. Similarly, for the Native American group, we would expect to see results this extreme only .2% of the time and for the other minority groups we would expect to see a value this extreme only 4.9% of the time. Their odds ratios, which are set using the white man as a baseline, then tell us how much more likely these students are to drop out. For example, a randomly selected female is predicted to be 1.217 times more likely to drop out than randomly selected male from the population and a randomly selected Native American is predicted to be over twice as likely to drop out after introductory calculus compared to a randomly selected White student. This tells us that our data suggests that gender does influence the impact of introductory calculus as a weed out course with significance. Race, however, only statistically significantly affects the dropout rate after calculus 1 for students of Native American heritage and those from races that did not fall into any of the given categories. We should point out, however, that the Native American group was the smallest group, containing only approximately 60 total students, and contained some assumption issues, so the result should be taken with caution. Similarly the group of Other Minorities, most likely contains a rather diverse group of students and therefore is difficult to interpret entirely on its own. Similarly, recall that the groups Hispanic and Native American followed a slightly different pattern in our assumptions check, so we must be cautious on assertions we make about these groups, both in the effects of race and gender. Note, however, that the results remain consistent even when these groups are merged into the Other category, the results of which are shown in Appendix C.

We can also view the predicted probability of dropping out by race and gender in the below table.

Table 6: Prediction of Drop Out Probability by Race and Gender

	White	Asian	Black	Hispanic	Native Americans	Other Minorities
Male	0.258	0.236	0.26	0.294	0.442	0.320
Female	0.297	0.273	0.30	0.336	0.491	0.365

These predictions show a similar story where the females are predicted to be more likely to drop out than the men and Native Americans and Other Minorities have the largest predicted dropout rates among the different races.

We should also consider the effect size of the model. The below table gives the values for McFadden's, Coxsnell's, and Nagelkerke's pseudo-R^2 statistics.

Table 7: Effect Sizes for the Effects of Race and Gender on Dropout Rate

McFadden	CoxSnell	Nagelkerke
0.003	0.003	0.005

McFadden's and Coxsnell's statistics suggest that the model explains .3% of the variation while Nagelkerke's suggests that the model explains .5% of the variation. While this may seem like a very small amount, keep in mind that we are attempting to predict dropout rates based solely on race and gender. We would not expect, and certainly not hope, that these factors would explain a large amount of variation in dropout rates nor was that the intended purpose of our analysis. It is also worth noting that using a likelihood ratio test and chi^2 statistic, we get a p value of less than .0001, which implies that our model should still be considered statistically significant relative to the null model.

4 Discussion and Further Study

From the MIDFIELD data, we have explored two aspects of minority participation in the STEM field. The first area we explored was whether there has been a growth of the proportion of women graduating with STEM degrees. Through the omnibus results from our study, we found that there was not a sufficient decrease or increase in the proportion of women graduating with STEM degrees with our collection of data.

The second area we explored was how the probability of dropping out of the STEM field after taking an introductory calculus course varied among different races and genders. In this case, we found that females were more likely to drop out after introductory calculus than males, and some minority groups also had a significantly higher probability of dropping out.

These analyses combine to suggest that these aspects of women and minorities in STEM should be further studied and indicates continuing underlying biases against them. That being said, both parts of this analysis did suffer from some issues that arose from incomplete data, and it should be kept in mind that even the most recent data in the MIDFIELD study only includes up to 2017. We also acknowledge that the original intent of the MIDFIELD study was to target engineering, which could have impacted our results. We would suggest that more study on these matters would be useful, especially to consider more recent data that targets all of STEM instead of primarily targetting engineering. We would also suggest further study into the patterns we found, in particular to explore what factors may be impeding the movement to get women more involved in the STEM field and the causation behind the more adverse effect that calculus has on women in comparision to men as well as on only select minority groups.

References

[1] Richard Layton, Russell Long and Matthew Ohland (2019). midfielddata: Student Record Data for 98,000 Undergraduates. R package version 0.1.0. https://github.com/MIDFIELDR/midfielddata

[2] National Center for Education Statistics. (n.d.). Browse CIP Codes. Retrieved from https://nces.ed.gov/ipeds/cipcode/browse.aspx?y=55

[3] Rasmussen, C., & Ellis, J. (2013). Students who switch out of calculus and the reasons why they leave. In Martinez, M. & Castro Superfine, A (Eds.). *Proceedings of the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 457-464). Chicago, IL: University of Illinois at Chicago.

[4] Schreiber-Gregory, D. (2018). PDF.

[5] Ohland, Matthew W., and Russell A. Long. The Multiple-Institution Database for Investigating Engineering Longitudinal Development: an Experiential Case Study of Data Sharing and Reuse. Advances in Engineering Education, 2016, advances.asee.org/wp-content/uploads/vol05/issue02/Papers/AEE-18-Ohland.pdf.

5 Appendix A: Data Wrangling Methods

Due to the extensiveness of the data wrangling necessary for this project, we leave the description of the process to this appendix and exclude the code generating the model ready data for either component of the project from the Code Appendix. For scripts to generate the model ready data, contact the authors of this work. The data released by the MIDFIELD project was originally separated into four data tables, whose names and descriptions as well as the pertinent data to this analysis pulled from them can be found in the table below.

Table 8: MIDFIELD Data Tables

Table	Purpose	UsefulData
midfieldterms	Shows the students' status by term	Program of study by term
midfieldcourses	Shows each instance of a student taking a class	Semester in which Calculus 1 was taken
midfieldstudents	Shows personal information about each student	Race and Gender of each student
midfielddegrees	Shows the end degree, or lack thereof, earned by each student	Students that earned STEM degrees

In order to properly combine this data for our analysis, there were several major obsticals to overcome. The first consideration was identifying the STEM programs in the data. This was done by translating the six digit cip codes into a binary variable that identified the code as either within STEM or not within STEM. [2] This method was used for both portions of the analysis.

5.1 Analysis Specific to Women in STEM

While finding the proportion of women STEM graduates by institution by year was rather straightforward, there were a few modelling decisions regarding the data. The first decision was to exclude all data that came from a year and institution that graduated fewer than 10 STEM students in the given year. This was done with the intention of helping to improve the reliability of the data.

A challenge that was ran into was fitting the data into the One Way Repeated Measures (Within-Subjects) ANOVA Model. Due to the result of the data exclusion mentioned above. Our subjects for our study (the individual institutions) no longer had a response value for every treatment, making our blocks incomplete. To combat this issue, two decisions were made. First, two institutions, institutions l and d, were removed from the data due to the lack of response values from the treatments. The second decision was to subset the Midfield data to include response values from the time period of 1994-2003, rather than the original time period of 1989-2016. This decision was reached to include the largest set of consecutive years in which each institution provided a proper response value, thus making the blocks complete.

5.2 Analysis Specific to Dropout Rates after Calculus 1

The most complicated challenge that we faced was to identify which semester the student took their first calculus course. This issue primarily involves the midfieldcourses data table. There are two identifiers for course in this table: course title and course code. While the course title can rather clearly identify courses that could be considered Calculus 1, only three institutions, institutions c, d, and l, provided course titles in the data.

This leaves the course code as an identifier of course. These codes, however, vary by institution and are not standardized to match to particular courses. In order to approach this issue and allow more of the data to

be used, we began cross-referencing subsets of course codes from particular institutions in order to attempt to match the anonymized data to a particular institution or a particular coding key in order to figure out which course code matches to Calculus 1. Note that in the attempt to create this matching, we excluded math courses from the matching set then ensured the predicted Calculus 1 course appeared in the data set. The institutions that were identified using this method were institutions a, b, e, h, j, and l. We were unable to identify the remaining institutions, so they were excluded from the second part of our analysis.

Another major considertation to make was how to define when a student drops out of the STEM field in association with Calculus 1. Noting that the data considers each year to contain six semesters, we defined a student to have dropped out of the STEM field in association with Calculus 1 if they do not graduate with a STEM degree and they do not appear in a STEM program in the fourth, fifth, or sixth semesters following the most recent taking of the course. Simply put, we sample the time period between six months and a year after the course and see whether or not the student is still in STEM. The consideration of whether or not the graduate in STEM prevents the accidental elimination of any student that gradutates immediately or shortly after Calculus 1 while the slight delay between the measurement era and the actual taking of the course gives time for the student to choose to drop out or change majors and for this change to be recorded.

6 Appendix B: Assumptions on Non-transformed Data

Upon performing our assumption testing for a One Way Repeated Measures (Within-Subjects) ANOVA Model, we had noticed issues with the normality and homoscedasticity of the original data. We have included the graphs used to interpret the normality and homoscedasticity below.

6.1 Normality

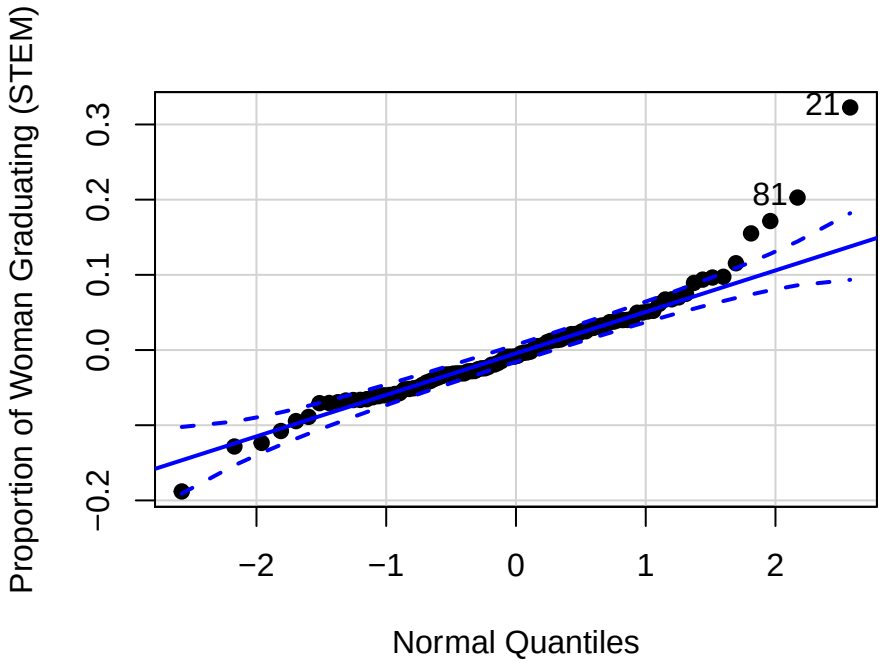


Figure 8: Normal Quantile Plot (Residuals) for Original MIDFIELD Data

Looking at the Normal Quantile Plot of the original data's residuals above in Figure 7, We quickly notice a large positive skew in our residuals. Due to the large amount of values existing outside of our 90% confidence envelope, we can interpret that our residuals do not appear to normally distributed. This lack of normality is a sign that a potential transformation of the data would be sufficient.

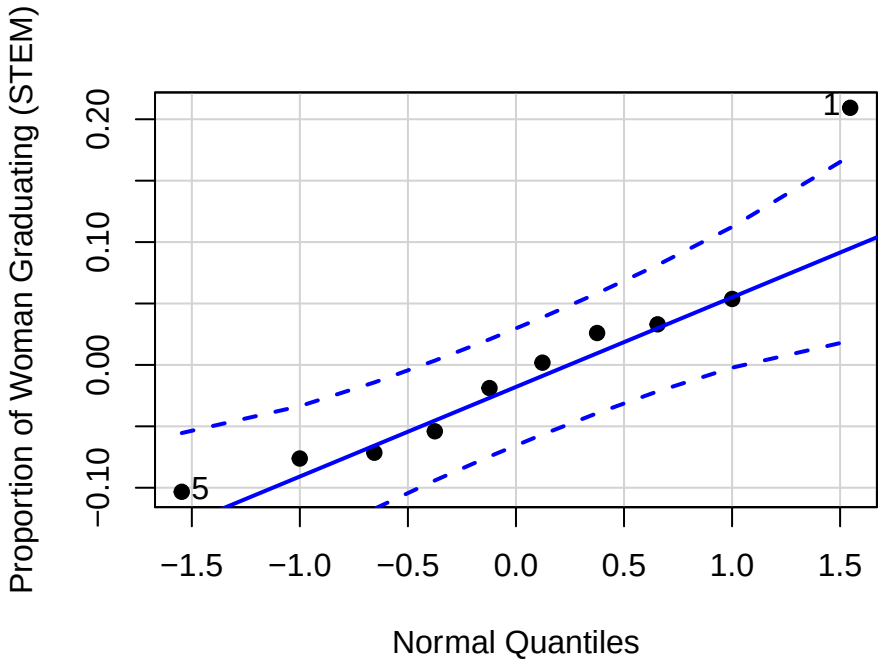


Figure 9: Normal Quantile Plot (Random Effects) for Original MIDFIELD Data

Upon analysis of the Normal Quantile plot for our original data's random effects (see Figure 8 above), we see a potentially troublesome data point within the plot. More specifically, there seems to a potential outlier represented the the top right of our plot. Since this value is outside of 90% confidence envelope, we assume that the normality of our random effects is not satisfied. Due to the lack of robustness within a random effects model, even slight deviances to normality such as the one we faced above can not be tolerated by the ANOVA model.

6.2 Homoscedasticity

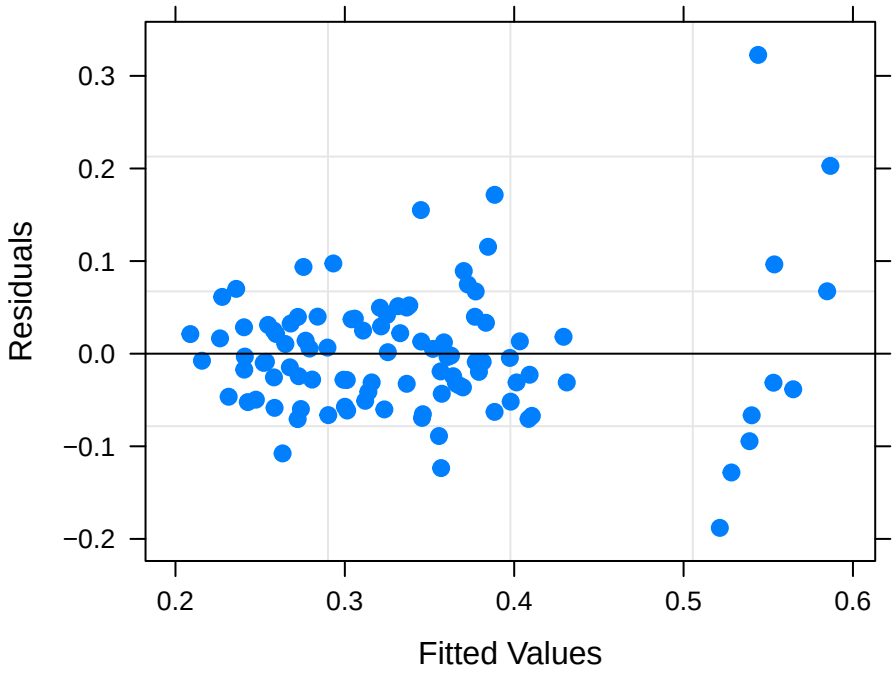


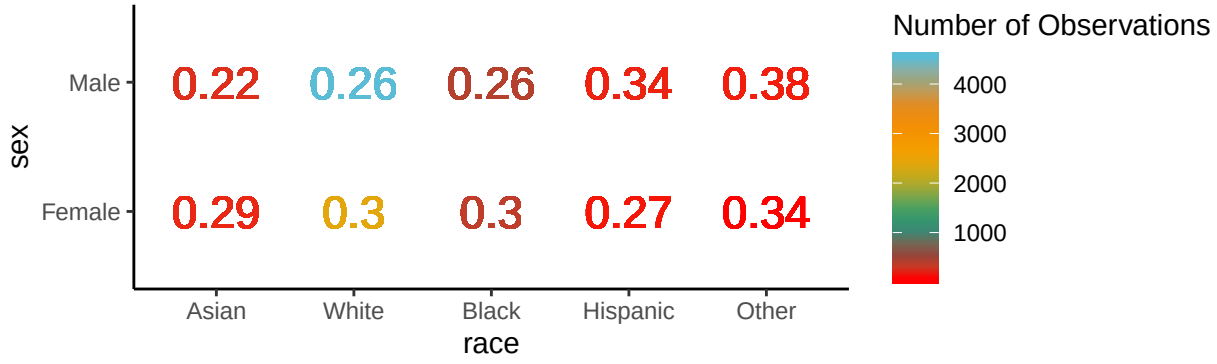
Figure 10: Tukey-Anscombe Plot of Original MIDFIELD Data

Finally, we take a look at the Tukey-Anscombe plot in Figure 9 in order to interpret our assumption of homoscedasticity. Looking at the plot, we notice a fanning pattern present within our residuals. Due to the lack of robustness within our model, we have decided to conclude that our residuals appear to not be homoscedastic.

Due to all the issues present when checking the assumptions for our model, we felt it neccessary to transform the original data. Upon several trials with different transformations, we decided that a log-transformation appeared to fit the most appropriately with our data.

7 Appendix C: Modified Logistic Regression Models

In the assumption testing of our data, we recognized some issues with the linearity of independent variables and log odds, the largest issue considering the Native American group, which was also distinctly the smallest group. It then makes sense to consider what happens when we merge the Native American groups into the Other Minorities category.



As we could have predicted, the result still has some issues with the Other Minorities and Hispanic categories in that they each have a higher male proportion of female dropouts. The remainder of the analysis remains similar.

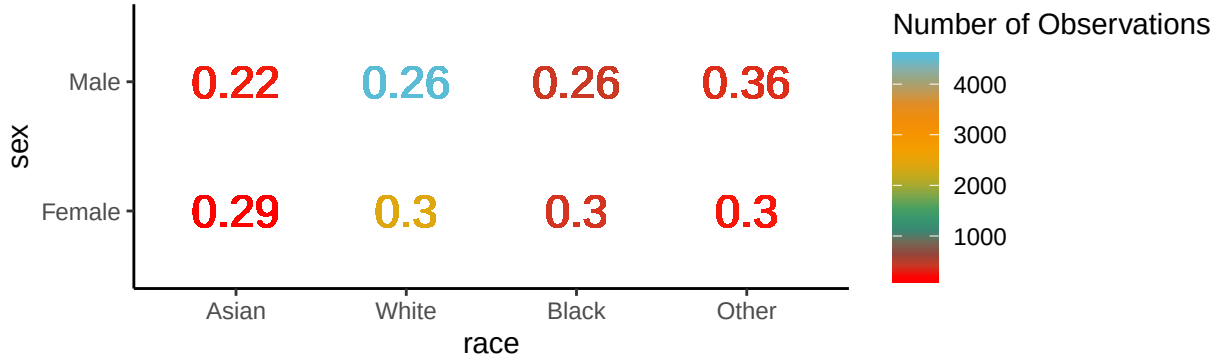
The results are shown in the below table.

Table 9: Odds Ratios Relative to white men

	Odds Ratio	P-value
Female	0.907	0.000
Asian	0.904	0.045
Black	0.804	0.024
Hispanic	0.916	0.252
Other Minorities	1.082	0.458

Other than the Other Minorities category, which clearly changed in composition, the odds ratios and p values are very similar to the original analysis, which shows us that this decision is not overly influential to our analysis and conclusions.

In an effort to better meet the linearity of independent observations and log odds, we can also merge the Hispanic group into the Other Minorities group.



Similar to the last case, the result is similar, with now only Other Minorities presenting mild issues to our assumption. The rest of analysis proceeds similarly with the test results shown below.

Table 10: Odds Ratios Relative to white men

	Odds Ratio	P-value
Female	0.908	0.000
Asian	0.954	0.304
Black	0.848	0.062
Other Minorities	0.966	0.625

Other than the Other Minorities category, which clearly changed in composition, the odds ratios and p values are still very similar to the original analysis, so the decision is still relatively unimportant. That being said, none of these models entirely meet assumptions, but the last does come relatively close and produces a similar result to our original model.

8 Appendix D: Code Appendix

```
library(tidyr)
library(dplyr)
library(data.table)
library(midfielddata)
library(Metrics)
library(caret)
library(rcompanion)
library(ggplot2)
library(hasseDiagram)
library(kableExtra)
library(cowplot)
library(wesanderson)
library(magrittr)
library(psych)
suppressMessages(library(car))
ALabs <- c("1 Grand Mean 1", "10 Year 9",
             "10 (Institution) 9", " 100 (Error) 81")
AMat <- matrix(data = F, nrow = 4, ncol = 4)
AMat[1, c(2:4)] = AMat[c(2:3), 4] = T
hasseDiagram::hasse(AMat, ALabs)
m_f<-read.csv("~/master_anova.csv")
m_f$term_degree<-as.factor(m_f$term_degree)
m_f$institution<-as.factor(m_f$institution)
tpx <- m_f[,1]
tpx
instx <- m_f[,2]
px <- m_f[,3]
dfx <- data.frame(tpx,instx, px)
dfx

df <- dfx[-c((seq(1,31, by=1)), seq(132,163, by=1)),]
df
row.names(df) <- 1:nrow(df)

tp2 <- df[,1]
tp <- droplevels(tp2, exclude = if(anyNA(levels(tp2))) NULL else NA)
inst2 <- df[,2]
inst <- droplevels(inst2, exclude = if(anyNA(levels(inst2))) NULL else NA)
p <- df[,3]
df2 <- data.frame(tp,inst,p)
df2$tp <- as.factor(df2$tp)
df2$inst <- as.factor(df2$inst)
stripchart(p~tp, vertical = TRUE,
           pch = 20, ylab = "Proportion of Women",
           xlab = "School Year",
           main = "Stripchart of Women in STEM")

boxplot(p~tp,
        pch = 20, ylab =
        "Proportion of Women",
        xlab = "School Year",
```

```
        main = "Boxplot of Women in STEM")
#getStats <- function(df2)

A <- psych::describeBy(
  x = as.numeric(df2$p), group = df2$tp,
  na.rm = TRUE, interp = TRUE,
  quant= c(0.25,0.75), skew = TRUE,
  digits = 3, mat = TRUE)
B <- data.frame(
  Year = A$group1,
  Sample.Max = A$max,
  Q3= A$Q0.75,
  Sample.Median = A$median,
  Q1 = A$Q0.25,
  Sample.Min = A$min,
  SAM = A$mean,
  SAV = round((A$sd)^2,3),
  Sample.Skewness = A$skew)

knitr::kable(B,
             caption = "Statistics for MIDFIELD Data",
             align = c('c', rep('r', 9))) %>%
kableExtra::kable_styling(bootstrap_options = c("striped", "condensed"),
                          font_size = 9, latex_options = "HOLD_position")%>%
  column_spec(1, width = '0.3in') %>%
  column_spec(2, width = "0.7in")%>%
  column_spec(3, width = "0.25in")%>%
  column_spec(4, width = "0.8in")%>%
  column_spec(5, width = "0.25in")%>%
  column_spec(6, width = "0.7in")%>%
  column_spec(7, width = "0.4in")%>%
  column_spec(8, width = "0.4in")%>%
  column_spec(9, width = "0.8in")
df2$logp <- log(df2$p)
midfieldLogM1 <- lme4::lmer(logp ~ tp + (1| inst), data = df2)
df2Wide <- tidyr::pivot_wider(df2,id_cols = inst, names_from = tp, values_from = logp)
responseWide <- as.matrix(df2Wide[ , 2:11])
midfieldMV <- lm(responseWide ~ 1)
Term_Year <- factor(levels(df2$tp))

### Create the table object
outputMV <- car::Anova(midfieldMV, idata = data.frame(Term_Year),
                   idesign = ~Term_Year, type = "III")
midfieldMultivariate <- summary(outputMV, multivariate=FALSE)
anovaFE <- car::qqPlot(
  x = residuals(midfieldLogM1),
  distribution = "norm",
  envelope = 0.90,
  xlab = "Normal Quantiles",
  ylab = "Log-Prop. of Woman Graduating (STEM)",
  pch = 19
)
anovaRE <- car::qqPlot(
```

```
  x = lme4::ranef(midfieldLogM1)$inst[, "(Intercept)"],
  distribution = "norm",
  envelope = 0.90,
  xlab = "Normal Quantiles",
  ylab = "Log-Prop. of Woman Graduating (STEM)",
  pch = 19
)
plot(midfieldLogM1, which = 1, pch = 19, xlab = "Fitted Values", ylab = "Residuals")
ggplot(data = df2,
  mapping = aes(x = tp, y = logp, color = inst, group = inst)) +
  geom_point(size=2) +
  geom_line(size=1) +
  facet_wrap(~ inst, ncol = 2) +
  theme_bw() +
  xlab("Term Year") +
  ylab("Log-Prop. of Woman Graduating (STEM)") +
  labs(color = "Institution") +
  theme(legend.position="top") +
  scale_x_discrete(breaks=seq(1994, 2002, 2))

options(knitr.kable.NA= "")
knitr::kable(
  data.frame(unclass(midfieldMultivariate$sphericity.tests)),
  digits = 3,
  col.names = c("Test Statistic", "p-value"),
  caption = "Mauchly's Test for Sphericity",
  align = c('c',rep('c',2))
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12, latex_options = "HOLD_position") %>%
kableExtra::footnote(
    general = "Computer rounding has made test statistic appear to be zero.",
    footnote_as_chunk = T)

midfieldFrame <- data.frame(unclass(midfieldMultivariate$univariate.tests))
midfieldFrame <- midfieldFrame[,c(2,4,1,3,5,6)]
knitr::kable(
  midfieldFrame,
  digits = 3,
  col.names = c("Numerator DF", "Denominator DF", "Term SS", "Error SS", "F Ratio", "p-value"),
  caption = "ANOVA Table Within-Subjects: Log of Proportion of Women Graduating in STEM",
  align = c('c',rep('c',6))
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12, latex_options = "HOLD_position") %>%
  kableExtra::footnote(
    general = "Computer rounding has made the p-values look like zero.",
    footnote_as_chunk = T)
knitr::kable(
  data.frame(unclass(midfieldMultivariate$pval.adjustments)),
  digits = 3,
```

```
  col.names = c("Greenhouse-Geiser", "p-value", "Huynh-Feldt", "p-value"),
  caption = "Adjustments for Sphericity Violations",
  align = c('c',rep('c',4))
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12, latex_options = "HOLD_position") %>%
  kableExtra::footnote(
    general = "Huynh-Feldt eps is being treated as 1.",
    footnote_as_chunk = T)
#Load data for logistic regression analysis
train<-fread("~/master_glm.csv") #Assumed location of data file
train<-train[race!="Unknown" & sex!="Unknown"]
train<-train[,race:=ifelse(race=="International", "Other", race)]

# Create prep data for EDA plot
trainp<-train[,race:=ifelse(race=="Native American", "Nat. Am.", race)]
trainp<-trainp[race!="White"]
trainp<-trainp[,sex:=ifelse(sex=="Male", "M", "F")]
trainp<-trainp[,dropout:=ifelse(dropout==1, "Yes", "No")]

# Component of plot for minority groups
gg1<-ggplot(data=trainp, aes(x=sex, fill = as.factor(dropout))) +
  geom_bar() +
  facet_grid (~race) +
  theme(legend.position="top") +
  labs(fill="Dropout")

# Component of plot for white group
trainp<-train[race=="White"]
trainp<-trainp[,sex:=ifelse(sex=="Male", "M", "F")]
gg2<-ggplot(data=trainp, aes(x=sex, fill = as.factor(dropout))) +
  geom_bar() +
  facet_grid (~race) +
  theme(legend.position = "none")
# Combine plots
plot_grid(gg1, gg2, rel_widths = c(3.5,1))
# Prep data for proportion grid plot
trainp2<-train
trainp2<-trainp2[,drop_rate:=mean(dropout), keyby=c("race", "sex")]
trainp3<-train
trainp3<-trainp2[,n:=.N, keyby=c("race", "sex")]
trainp3$race<-factor(trainp2$race,
                     levels=c("Asian", "White", "Black", "Hispanic", "Other", "Nat. Am."))

# Make dropout proportion grid plot
ggplot(aes(x=race, y=sex, color=n), data=trainp3) +
  geom_text(aes(label=round(drop_rate, digits=2)), size=6) +
  scale_color_gradientn(colours = wes_palette("Darjeeling1", 10, type = "continuous")) +
  labs(color="Number of Observations") +
  theme_bw()+
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```

```r
# Correct training data to be ready for the rest of the analysis
train[,race:=ifelse(race=="Nat. Am.","Native American", race)]
train$sex<-as.factor(train$sex)
train$race<-as.factor(train$race)

# Set order for race and gender so that white men will be the baseline
Gender <- factor(train$sex, levels=c('Male', 'Female'))
Race<-factor(train$race,
             levels = c("White", "Asian", "Black", "Hispanic",
                        "International", "Native American", "Other"))
glm_model<-glm(dropout ~ Gender + Race, family = binomial, data = train)

# Check Multicollinearity
car::vif(glm_model)

# Calculate Odds Ratios
oddsrat <- exp(parameters::model_parameters(glm_model)[2])
p <- parameters::model_parameters(glm_model)["p"]
dat <- data.frame(Odds_Ratio=oddsrat, p=p)[-1,]
rownames(dat)<-c("Female", "Asian", "Black", "Hispanic", "Native American", "Other Minorities")
setnames(dat, c("Coefficient", "p"), c("Odds Ratio", "P-value"))

# Table for Odds Ratios
knitr::kable(
  dat,
  digits = 3,
  caption = "Odds Ratios Relative to white men",
  align = c('l',rep('c',4))
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12, latex_options = "HOLD_position")

# Get predictions for each combination of race and gender
test<-data.table(Race = c("White", "Asian", "Black", "Hispanic", "Native American", "Other",
                          "White", "Asian", "Black", "Hispanic",  "Native American", "Other"),
                 Gender=c("Male","Male","Male","Male","Male","Male",
                          "Female","Female","Female","Female","Female","Female"))
test$Prediction<-predict(glm_model,newdata = test,type="response")
preds<-data.frame(Sex = c("Male", "Female"),
                  White = c(test$Prediction[1], test$Prediction[7]),
                  Asian = c(test$Prediction[2], test$Prediction[8]),
                  Black = c(test$Prediction[3], test$Prediction[9]),
                  Hispanic = c(test$Prediction[4], test$Prediction[10]),
                  Nat = c(test$Prediction[5], test$Prediction[11]),
                  Other = c(test$Prediction[6], test$Prediction[12]))
setnames(preds, c("Nat", "Other", "Sex"), c("Native Americans", "Other Minorities", " "))

#Make table of predictions
knitr::kable(
  preds,
  digits = 3,
  caption = "Prediction of Drop Out Probability by Race and Gender",
```

```
  align = c('l',rep('c',6))
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12, latex_options = "HOLD_position")

# Calculate effect sizes
rpt<- rcompanion::nagelkerke(glm_model)
rpt

# Data for effect size table
effsize<-data.frame(McFadden=c(.003), CoxSnell = c(.003), Nagelkerke = c(.005))

# Table for Effect Sizes
knitr::kable(effsize,
             caption = "Effect Sizes for the Effects of Race and Gender on Dropout Rate") %>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "condensed"),
                            font_size = 12, latex_options = "HOLD_position")

# Create data for table on midfield data explanation
midtab<-data.frame(Table = c("midfieldterms",
                             "midfieldcourses",
                             "midfieldstudents",
                             "midfielddegrees"),
                   Purpose = c("Shows the students' status by term",
                               "Shows each instance of a student taking a class",
                               "Shows personal information about each student",
                               "Shows the end degree, or lack thereof, earned by each student"),
                   UsefulData = c("Program of study by term",
                                  "Semester in which Calculus 1 was taken",
                                  "Race and Gender of each student",
                                  "Students that earned STEM degrees"))

# Create table for midfield data explanation
knitr::kable(
  midtab,
  digits = 3,
  align = c('l',rep('c',2)), caption = "MIDFIELD Data Tables"
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12, latex_options = "HOLD_position")%>%
     column_spec(1, width = '1.1in') %>%
      column_spec(2, width = "2.5in")%>%
  column_spec(3, width = "2.5in")

options("contrasts" = c("contr.sum","contr.poly"))
midfieldM1 <- lme4::lmer(p ~ tp + (1| inst), data = df2)
anovaFE <- car::qqPlot(
  x = residuals(midfieldM1),
  distribution = "norm",
  envelope = 0.90,
  xlab = "Normal Quantiles",
```

```
  ylab = "Proportion of Woman Graduating (STEM)",
  pch = 19
)
anovaRE <- car::qqPlot(
  x = lme4::ranef(midfieldM1)$inst[, "(Intercept)"],
  distribution = "norm",
  envelope = 0.90,
  xlab = "Normal Quantiles",
  ylab = "Proportion of Woman Graduating (STEM)",
  pch = 19
)
plot(midfieldM1, which = 1, pch = 19, xlab = "Fitted Values", ylab = "Residuals")
# Reread data for Appendix C model 1
train<-fread("~/master_glm.csv")
train<-train[,race:=ifelse(race=="International", "Other", race)]
train<-train[,race:=ifelse(race=="Native American", "Other", race)]
train$sex<-as.factor(train$sex)
train$race<-as.factor(train$race)
train<-train[race!="Unknown" & sex!="Unknown"]
Gender <- factor(train$sex, levels=c('Male', 'Female'))
Race<-factor(train$race, levels=c("White", "Asian", "Black", "Hispanic", "Other"))

# Create prep data for plot
trainp=train
trainp<-trainp[race!="White"]
trainp<-trainp[,sex:=ifelse(sex=="Male", "M", "F")]
trainp<-trainp[,dropout:=ifelse(dropout==1, "Yes", "No")]

trainp2<-train
trainp2<-trainp2[,drop_rate:=mean(dropout), keyby=c("race", "sex")]

trainp3<-train
trainp3<-trainp2[,n:=.N, keyby=c("race", "sex")]
trainp3$race<-factor(trainp2$race, levels=c("Asian", "White", "Black", "Hispanic", "Other"))

#Make dropout by sex and race plot
ggplot(aes(x=race, y=sex, color=n), data=trainp3) +
  geom_text(aes(label=round(drop_rate, digits=2)), size=6) +
  scale_color_gradientn(colours = wes_palette("Darjeeling1", 10, type = "continuous")) +
  labs(color="Number of Observations") +
  theme_bw()+
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))

# Run logistic regression model
train$sex<-as.factor(train$sex)
train$race<-as.factor(train$race)
Gender <- factor(train$sex, levels=c('Male', 'Female'))
Race<-factor(train$race,
             levels = c("White", "Asian", "Black", "Hispanic", "Other"))
glm_model<-glm(dropout ~ Gender + Race, family = binomial, data = train)

# Check Multicollinearity
```

```r
car::vif(glm_model)

# Calculate Odds Ratios
oddsrat <- exp(parameters::model_parameters(glm_model)[2])
p <- parameters::model_parameters(glm_model)["p"]
dat <- data.frame(Odds_Ratio=oddsrat, p=p)[-1,]
rownames(dat)<-c("Female", "Asian", "Black", "Hispanic",  "Other Minorities")
setnames(dat, c("Coefficient", "p"), c("Odds Ratio", "P-value"))

# Table for Odds Ratios
knitr::kable(
  dat,
  digits = 3,
  caption = "Odds Ratios Relative to white men",
  align = c('l',rep('c',4))
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12, latex_options = "HOLD_position")

# Reread data for Appendix C model 2
train<-fread("~/master_glm.csv")
train<-train[,race:=ifelse(race=="International", "Other", race)]
train<-train[,race:=ifelse(race=="Native American", "Other", race)]
train<-train[,race:=ifelse(race=="Hispanic", "Other", race)]
train$sex<-as.factor(train$sex)
train$race<-as.factor(train$race)
train<-train[race!="Unknown" & sex!="Unknown"]
Gender <- factor(train$sex, levels=c('Male', 'Female'))
Race<-factor(train$race, levels=c("White", "Asian", "Black", "Other"))

# Create prep data for plot
trainp=train
trainp<-trainp[race!="White"]
trainp<-trainp[,sex:=ifelse(sex=="Male", "M", "F")]
trainp<-trainp[,dropout:=ifelse(dropout==1, "Yes", "No")]

trainp2<-train
trainp2<-trainp2[,drop_rate:=mean(dropout), keyby=c("race", "sex")]

trainp3<-train
trainp3<-trainp2[,n:=.N, keyby=c("race", "sex")]
trainp3$race<-factor(trainp2$race, levels=c("Asian", "White", "Black", "Other"))

#Create dropout rate plot
ggplot(aes(x=race, y=sex, color=n), data=trainp3) +
  geom_text(aes(label=round(drop_rate, digits=2)), size=6) +
  scale_color_gradientn(colours = wes_palette("Darjeeling1", 10, type = "continuous")) +
  labs(color="Number of Observations") +
  theme_bw()+
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```

```r
# Fit logistic regression model
train$sex<-as.factor(train$sex)
train$race<-as.factor(train$race)
Gender <- factor(train$sex, levels=c('Male', 'Female'))
Race<-factor(train$race,
             levels = c("White", "Asian", "Black", "Other"))
glm_model<-glm(dropout ~ Gender + Race, family = binomial, data = train)

# Check Multicollinearity
car::vif(glm_model)

# Calculate Odds Ratios
oddsrat <- exp(parameters::model_parameters(glm_model)[2])
p <- parameters::model_parameters(glm_model)["p"]
dat <- data.frame(Odds_Ratio=oddsrat, p=p)[-1,]
rownames(dat)<-c("Female", "Asian", "Black",  "Other Minorities")
setnames(dat, c("Coefficient", "p"), c("Odds Ratio", "P-value"))

# Table for Odds Ratios
knitr::kable(
  dat,
  digits = 3,
  caption = "Odds Ratios Relative to white men",
  align = c('l',rep('c',4))
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12, latex_options = "HOLD_position")
```