Megan Faulkner
Educational Psychology Doctoral Student
meganfaulkner@unm.edu

EDPY 603: Applied Statistical Design & Analysis
December 2022

The Effect of First-Generation and Low-Income Status on Research Experience prior to Enrollment in Undergraduate STEM Coursework

Background Information

Involvement in research is a positive predictor for retention and success in undergraduate STEM degrees, with significant positive effects on four-year graduation rates, aspirations to continue to graduate school, GPA, satisfaction with the university experience, research self-efficacy, and plans to pursue a future career in research (Adedokun et al., 2013; Bowman & Holmes, 2018; Kilgo & Pascarella, 2016). While many students do not pursue research experiences until their junior or senior year, several studies indicate that earlier development of research skills produces stronger positive effects, particularly in terms of science identity development and higher order thinking skills (Adedokun et al., 2014; Thiry et al., 2012). Students that are traditionally underrepresented in higher education, particularly first-generation and low-income students, can benefit the most from research experiences, with participation in undergraduate research programs mitigating the low academic engagement, performance, and retention often associated with these populations (Harackiewicz et al., 2014; Soria & Stebleton, 2012).

Despite the evidence for the benefits of early exposure to research, very few studies investigate research experiences prior to enrollment in an undergraduate degree and the differences in involvement across underrepresented groups. Developing research experience prior to entering college may help students become involved with research earlier in their undergraduate degree and provide students with skills that can reduce the achievement gap experienced by first-generation and low-income students. At the University of New Mexico in particular, as an R1 institution serving a high percentage of first-generation and low-income students, understanding how these factors impact the research skills of incoming students is critical to supporting student success in STEM coursework.

The purpose of this study is to investigate the impact of first-generation and low-income student status on research experience prior to first-time enrollment in an undergraduate degree. Two research questions were developed to address this goal:

RQ1: Are there differences in research experience prior to college across first-generation and continuing-generation students, controlling for sex and race/ethnicity?

RQ2: What is the impact of family income on research experience prior to college, and does that impact differ between first-generation and continuing-generation students?

Methodology

The dataset used in this analysis is survey data collected by the University of New Mexico (UNM) Undergraduate Research, Arts, and Design Network (URAD) in the Fall 2020 semester. Incoming first-time freshmen enrolled in URAD-affiliated courses completed the pre-survey during the first two weeks of the semester, sharing information about their involvement in research experiences prior to college. UNM ID numbers from the survey were matched to students in UNM's Banner system to provide demographic and other background information. Of the 122 initial respondents, one was removed for incomplete data on the prior research experience questions, resulting in a dataset of 121 respondents.

<u>Predictor Variables</u>: Demographic information gathered from the survey and from UNM records are used as predictor variables in this study, namely sex, race/ethnicity, first-generation student status, and family income, as determined by Pell Grant eligibility. Distribution of respondents across these variables are reported in Table 1.

Outcome Variable: The outcome variable is a continuous score determined by answers to eight items on the URAD survey score measuring experience with research prior to entering college. Descriptive statistics across demographic categories are reported in Table 1.

Table 1. Prior research score statistics by demographic

	n	Mean	SD	SE
All respondents	121	20.07	20.92	1.90
Sex				
Female	57	22.74	21.10	2.79
Male	64	17.69	20.63	2.58
Race / ethnicity				
Hispanic	65	22.34	20.43	2.53
White	37	15.19	21.02	3.46
Other	19	21.79	21.86	5.01
Student Status				
First-Generation	60	21.65	20.79	2.68
Continuing Generation	61	18.51	21.10	2.70
Family Income				
Low Income	35	22.34	19.62	3.32
Not Low Income	86	19.14	21.47	2.31

Data Analysis

Three ANOVA models were developed to investigate the research questions. To address the first research question, a covariate only model was created with the variables sex (B_1, dummy coded with male as the reference group) and race/ethnicity (B_2 and B_3, effect coded with White as the reference group). First-generation student status was then added (B_4, dummy coded with continuing-generation students as the reference group). A 1df F-test was used to determine differences in prior research scores across first-generation and continuing-generation students.

To answer the second question, family income was added to the model, as measured by Pell Grant eligibility (B_5, dummy coded with non-Pell Grant eligibility as the reference group), as well as an interaction term (B_6) between Pell Grant eligibility and first-generation student status. A 1df F-test was used to determine the effect of family income on prior research experience and a second 1df F-test was run to determine if the effect differs between first-generation and continuing-generation students. All analyses were performed using R Studio.

$$\text{PriorRes} = B_0 + B_1 * Female + B_2 * Hispanic + B_3 * Other + B_4 * FirstGen + B_5 * Pell + B_6 * FirstGen * Pell$$

Results

Covariate only model. The covariate only model did not indicate significant effects of sex or race/ethnicity on prior research experience scores ($F(3, 117) = 1.81$, $p > 0.05$).

RQ1: Are there differences in research experience prior to college across first-generation and continuing-generation students, controlling for sex and race/ethnicity? This question was addressed by adding first-generation student status to the covariate-only model (Table 2). Analysis did not show significant differences in prior research experience between first-generation and continuing-generation students when controlling for sex and race/ethnicity ($F(1, 116) = 1.39$, $p > 0.05$). A comparison between this model and the covariate-only model did not indicate a significant difference between the two ($p > 0.05$).

RQ2: What is the impact of family income on research experience prior to college, and does that impact differ between first-generation and continuing-generation students? The second research question was answered by adding Pell Grant eligibility and the interaction between Pell Grant eligibility and first-generation student status to the previous model (Table 2). Results did not show a significant main effect of family income on prior research experience ($F(1, 115) = 1.12$, $p > 0.05$, nor was there a significant interaction between family income and first-generation student status, meaning that the effects of being low income did not differ across student status ($F(1, 114) = 1.01$, $p > 0.05$). Model comparisons did not reveal significant differences between these two models ($p > 0.05$), or between the family income main effect model and the RQ1 model ($p > 0.05$).

<u>Post-Hoc Power Analysis</u>: While a literature search did not find an appropriate meta-analysis for the effect of first-generation student status on research experience prior to college, a study by Harackiewicz et al. (2014) reported an effect size of 0.39 for first-generation student status on academic achievement for undergraduate freshmen. A post-hoc power analysis for first-generation student status using this effect size and parameter estimates from Model 4 resulted in a value of 0.41. This result indicates that the sample used in this study had only a 41% probability of finding an effect size of 0.39 or greater.

Table 2. Impact of first-generation student status on research experience prior to college

Predictor	B	SE	df	p < .05	ES
Model 1: Covariates					
F(3, 117) = 1.81					
Intercept	11.78	4.04	1	*	
Female	6.00	3.82	1		0.29
Hispanic	7.78	4.28	1		0.38
Other	8.11	5.92	1		0.40
Model 2: Covariates and first-generation student status					
F(1, 116) = 1.39					
FirstGen	1.52	3.85	1		0.07
Model 3: Effects of first-generation status across income level					
F(1, 115) = 1.12					
Pell	1.20	4.50	1		0.06
Model 4: Effects of first-generation status across income level					
F(1, 114) = 1.01					
FirstGen * Pell	-6.5	8.90	1		-0.32

*Notes: * indicates p < .05; ES reported using Cohen's d, using sd of model residuals as denominator.*

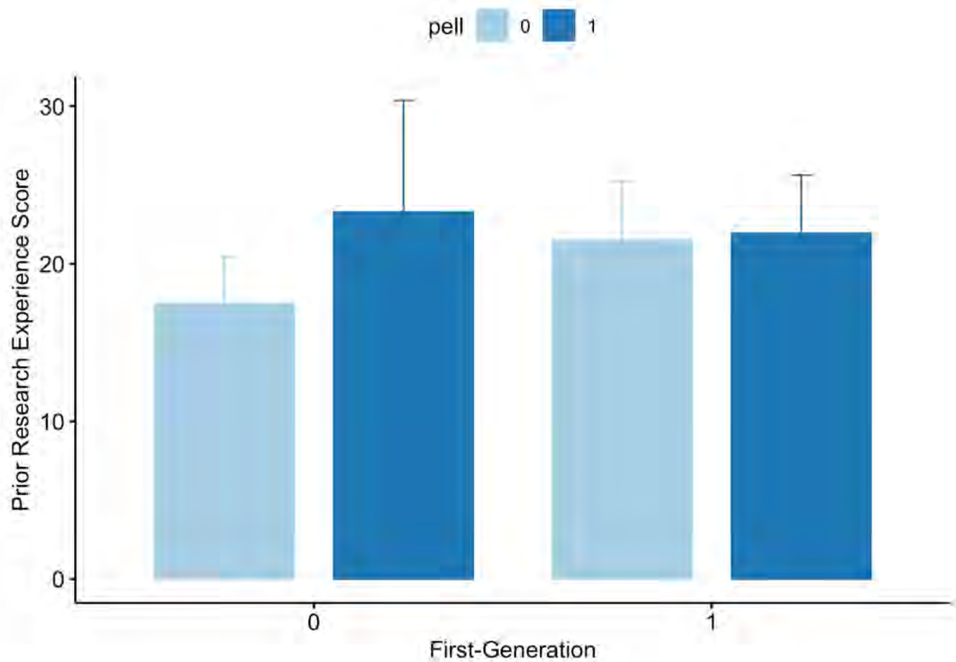

Figure 1: Bar graph of group means with 1 standard error bar

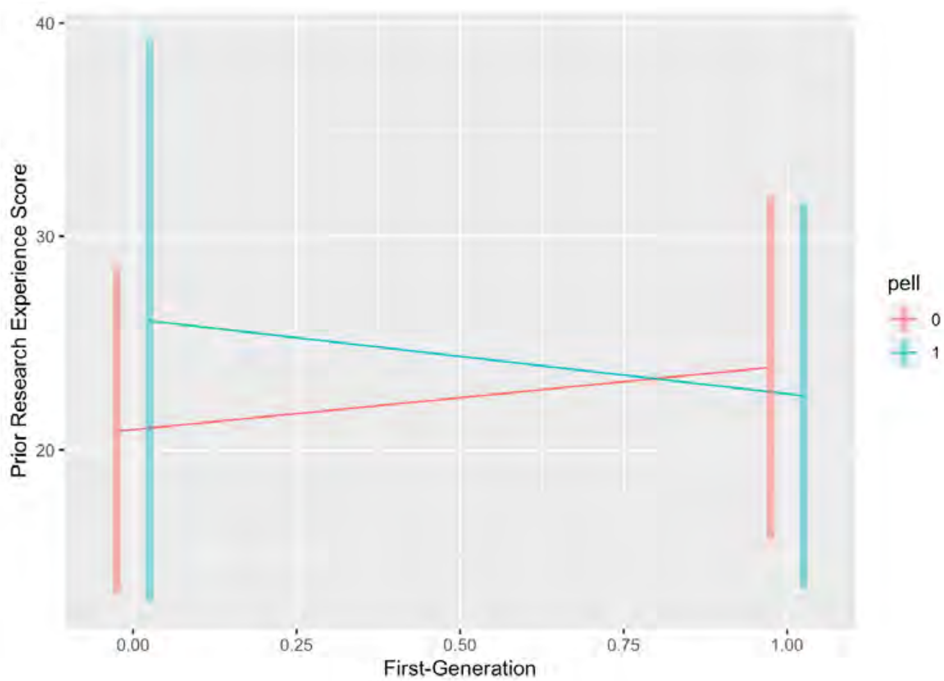

Figure 2: Line graph of adjusted group means with confidence intervals

Discussion and Conclusion

The analyses in this study did not find significant main effects of first-generation student status, family income, race/ethnicity, or sex on research experience prior to high school, nor was a significant interaction found between first-generation status and family income. Because 47% of students in the studied sample did not have prior research experience and because research experience is a predictor for success in undergraduate STEM degrees, understanding why students do not engage in research is a worthwhile goal. While not available in this dataset, existing research suggests that high school factors, such as the overall quality of the school and the academic rigor of the curriculum, parenting practices, cultural expectations and norms, and student beliefs, such as sense of belonging in a college environment, are significant contributions to differences in student performance between social classes, including low income and first-generation students (Harackiewicz et al., 2014). Though these factors are not explicitly related to research experience prior to high school, further investigation into these factors may reveal a relationship.

Additionally, a post-hoc power analysis suggests that the study was underpowered to find significant effects, although the analysis was based on one study, which is not best practice for determining a suitable effect size goal. Future research should pursue ways to increase the power for analysis, possibly by increasing the sample size and finding (or generating through meta-analysis) more accurate effect size estimates.

Model Assumptions

Normal distribution of error terms:

A histogram of the residuals for Model 4 was generated with two curves added for interpretation (Figure 3). The green line follows the model residuals, and the blue line follows a normal distribution. From this graph, it appears that the residuals for this model follow a non-normal distribution. To investigate the possibility that the high percentage of outcomes scores of zero (n = 59 or 49%) were responsible for the distribution, a second histogram was generated from a model (Model 5) using only prior research scores greater than zero (Figure 4). While this distribution is much closer to normal, omission of zero scores reduces the ability to answer the research questions for this study, so the original dataset was used for all analyses, unless otherwise noted. By using this dataset, it is possible that standard errors will be overestimated, resulting in larger confidence intervals and more conservative results.

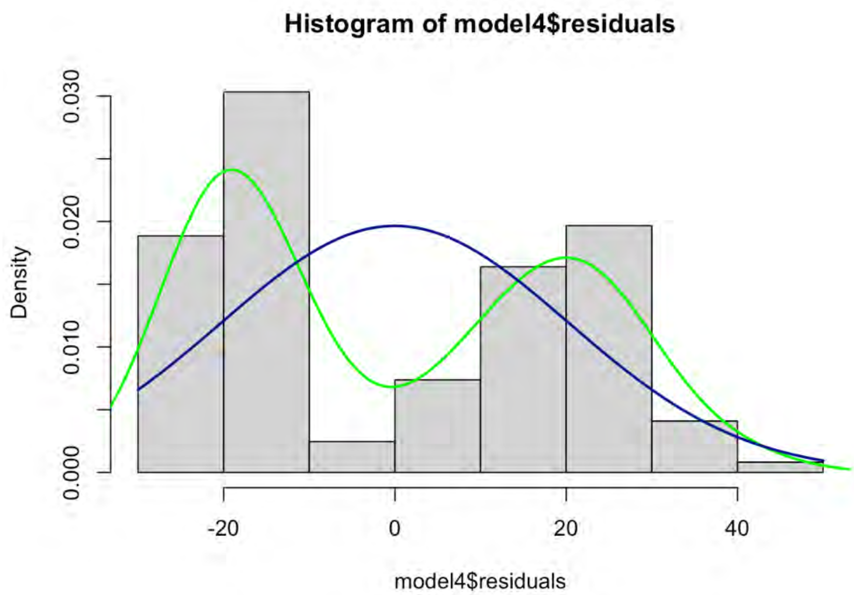

Figure 3: Histogram of Model 4 residuals

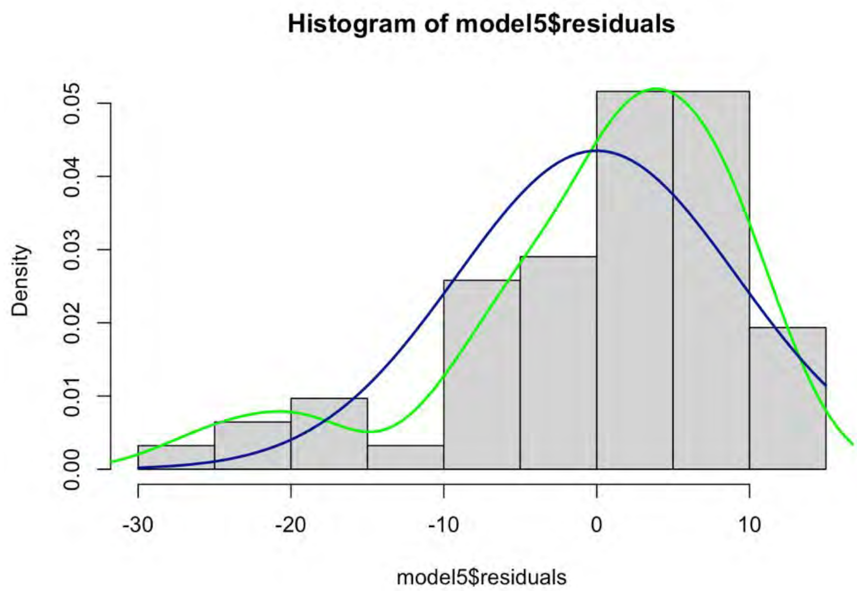

Figure 4: Histogram of Model 5 residuals

<u>Homoscedasticity</u>:

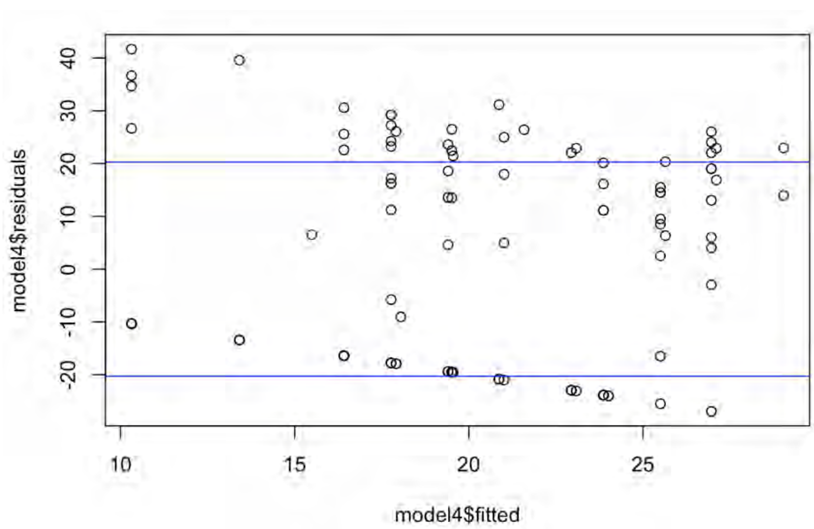

Figure 5: Scatterplot of Model 4 residuals with standard deviation lines

A scatterplot of residuals for Model 4 was generated, with a +1 and a -1 standard deviation line added around the mean (Figure 5). The residuals appeared to generally fall into a band with no discernible pattern indicating heteroscedasticity.

<u>Independence of residuals</u>:

A scatterplot of Model 4 residuals exhibits clustering of data points near the bottom of the graph (Figure 6). Similar to the assumption of normality graph above, a second scatterplot was generated from Model 5, in which outcome scores of zero were removed from the dataset (Figure 7). The clusters are no longer apparent, indicating that there is a clustering effect due to the zero scores. While the original dataset including the zero scores is used for all analyses in this study, further investigation into the characteristics of students with zero prior research scores would be useful in determining specific characteristics of this group that would suggest different effects for them than what was found in the main analysis.

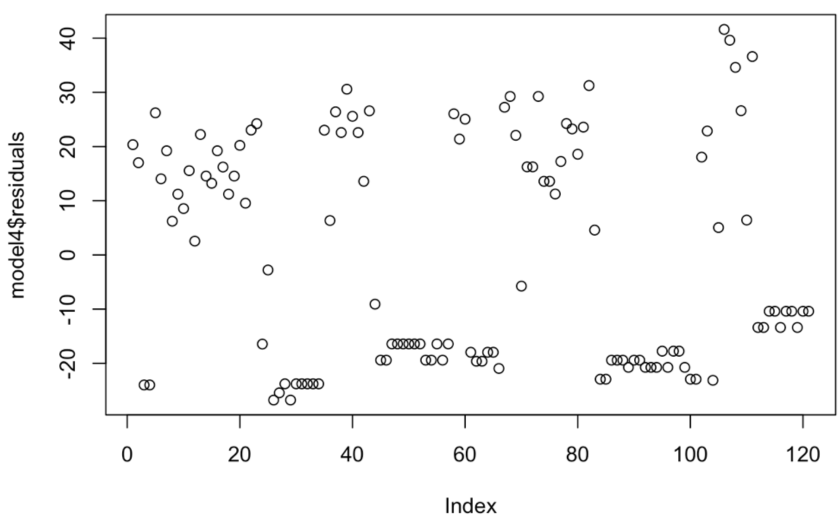

Figure 6: Scatterplot of Model 4 residuals

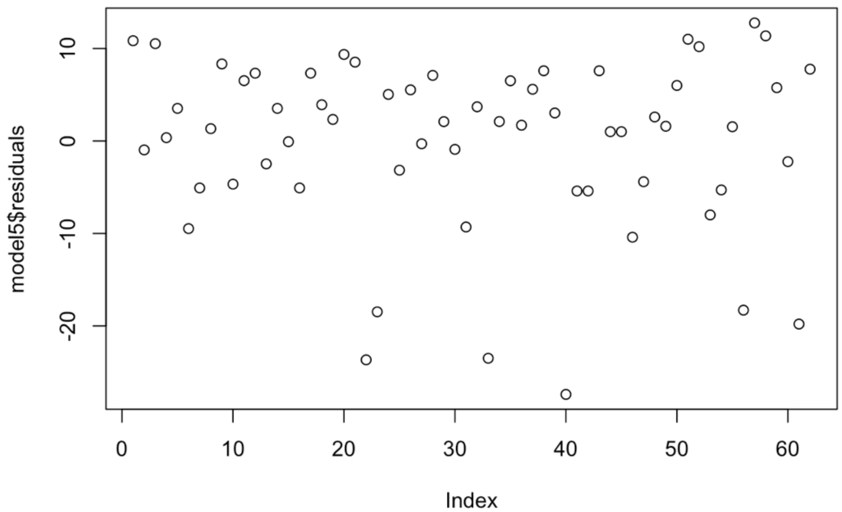

Figure 7: Scatterplot of Model 5 residuals

Linearity: All predictor variables are categorical; therefore, the model must be linear.

Accurate measurement of variables: Low frequencies of ethnicities/races other than White and Hispanic resulted in the use of an “Other” category to include identities such as Black/African-American, Asian, and Multi-racial. As these groups are not homogenous, results for the “Other” category should not be applied to members of the groups included in this variable. A larger sample size with a higher percentage of these races/ethnicities would allow for more accurate measurement of groups other than White and Hispanic. The “sex” variable may also be inaccurately measured, as the data source, the UNM admission application, only includes “male” and “female” as choices. Recent research reports that nearly 5% of adults under 30 in the United States identify as non-binary or transgender (Brown, 2022), indicating the likelihood of respondents who do not identify as “male” or “female” answering inaccurately.

Cronbach’s Alpha was calculated for the eight items in the prior research score scale. The calculated α was 0.86, which is considered a “good” score for internal reliability.

Relevant variables included: This assumption is addressed in the Discussion section above.

R Studio Code:

```
library(readr)
urad <- read_csv("~/Desktop/PhD Program/EDPY 603:504/603 Final Project/603_finalproject.csv")
View(urad)

#----------Dummy Coding Ethnicity/Race-----------
Hispanic <- ifelse(urad$ethnicity_race == "Hispanic",1,0)
Other <- ifelse(urad$ethnicity_race == "Other",1,0)

#----------Descriptive Statistics----------------

library(psych)
describe(urad$prior_research_score)
describeBy(urad$prior_research_score, group = urad$female)
describeBy(urad$prior_research_score, group = urad$ethnicity_race)
describeBy(urad$prior_research_score, group = urad$first_gen)
describeBy(urad$prior_research_score, group = urad$pell)

#-----------ANOVA Model Building---------------

#---Model 1: Covariates---
model1 <- lm(prior_research_score ~ female + Hispanic + Other, data = urad)
summary(model1)
sd(model1$residuals)

#Standardized coefficients
library(lm.beta)
lm.beta(model1)

#---Model 2: Covariates and First-Generation Status; main effect of first-gen status---
model2 <- lm(prior_research_score ~ female + Hispanic + Other + first_gen, data = urad)
summary(model2)
sd(model2$residuals)

#Standardized coefficients
lm.beta(model2)

#Calculate differences between models 1 and 2
anova(model1, model2)

#---Model 3: Add Pell Grant eligibility; main effect of family income---
model3 <- lm(prior_research_score ~ female + Hispanic + Other + first_gen + pell, data = urad)
summary(model3)
sd(model3$residuals)

#Standardized coefficients
lm.beta(model3)

#Calculate differences between models 2 and 3
anova(model2, model3)
```

```
#---Model 4: Add interaction between Pell and First-Gen---
model4 <- lm(prior_research_score ~ female + Hispanic + Other + first_gen + pell + first_gen * pell, data = urad)
summary(model4)
sd(model4$residuals)

#Standardized coefficients
lm.beta(model4)

#Calculate differences between models 3 and 4
anova(model3, model4)
```

```
#---Model 5: Same as Model 4, but with zero scores removed from dataset---
#Creating urad_2 dataset
urad_2 <- subset(urad, prior_research_score!=0)
Hispanic <- ifelse(urad_2$ethnicity_race == "Hispanic",1,0)
Other <- ifelse(urad_2$ethnicity_race == "Other",1,0)

model5 <- lm(prior_research_score ~ female + Hispanic + Other + first_gen + pell + first_gen * pell, data = urad_2)
summary(model5)
```

```
#-----------Post-Hoc Power Analysis---------------
#"A" value: 0.39 effect size (Harackiewicz et al., 2014) * sd of Model 4 residuals (20.38)
#"s" value: standard error of first-gen parameter estimate from Model 4
#"df" value: degrees of freedom from Model 4
library(retrodesign)
retrodesign(A = 7.95, s = 4.55, df = 115)
```

```
#-----------Assumptions Testing--------------

#---Normality---
#Because half of sample size scored 0, this may affect the distribution of the error terms
#Histogram of residuals from Model 4
hist(model4$residuals, freq = F)
#Add kernel density curve
lines(density(model4$residuals), col = "green", lwd = 2)
#Add normal curve
curve(dnorm(x, mean = mean(model4$residuals), sd = sd(model4$residuals)), col = 'darkblue', lwd = 2, add = TRUE, yaxt = 'n')

#Histogram of residuals from Model 5 (omitting zero scores)
hist(model5$residuals, freq = F)
#Add kernel density curve
lines(density(model5$residuals), col = "green", lwd = 2)
#Add normal curve
curve(dnorm(x, mean = mean(model5$residuals), sd = sd(model5$residuals)), col = 'darkblue', lwd = 2, add = TRUE, yaxt = 'n')
```

```
#---Homogeneity of variance---
plot(model4$residuals~model4$fitted)

#Calculate standard deviation of model4 residuals
sd <- sd(model4$residuals)
#Add +1/-1 sd to residuals plot
abline(h = sd, col = 'blue')
abline(h = -sd, col = 'blue')
```

```
#---Independence---
#Model 4 plot
plot(model4$residuals)

#Model 5 plot, with zero scores removed
plot(model5$residuals)
```

```
#---Linearity---
#No test necessary as all predictor variables are categorical

#---No measurement error---
#Calculating Cronbach's Alpha for prior research experience scale
library(readr)
score <- read_csv("~/Desktop/PhD Program/EDPY 603:504/603 Final Project/score.csv")
#Recode missing data from 999 to NA
score[score == "999"] <- NA
#Listwise deletion of rows with missing data
complete_score <- score[complete.cases(score), ]
View(complete_score)
```

```
library(psych)
scale <- data.frame(complete_score$q5, complete_score$q8a, complete_score$q8b, complete_score$q6,
                    complete_score$q7a, complete_score$q7b, complete_score$q10a, complete_score$q10b)
alpha(scale)

#---All predictor variables included in model---
#Addressed in write-up
```

```
#-----------Graphs--------------

#Bar graph of group means with standard error bar

#Convert first_gen and pell variables to factors
urad$first_gen <- factor(urad$first_gen)
urad$pell <- factor(urad$pell)

library(ggpubr)
#ggbarplot(urad, x = "first_gen", y = "prior_research_score",add = "mean_se",
#   fill = "pell", color="pell",palette = "Paired",
#   ylab = "Prior Research Experience Score", xlab = "First-Generation",
#   position = position_dodge(.9))

#Line graph of adjusted means by group
library(emmeans)
emmip(model4, pell ~ first_gen, CIs = T, ylab = "Prior Research Experience Score",
      xlab = "First-Generation")
```

R Studio Output:

```
> library(readr)
> urad <- read_csv("~/Desktop/PhD Program/EDPY 603:504/603 Final Project/603_finalproject.csv")
Rows: 121 Columns: 6
── Column specification ──────────────────────────────────────────────────────
Delimiter: ","
chr (1): ethnicity_race
dbl (5): ecure_id, pell, first_gen, female, prior_research_score

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.
> View(urad)
>
> #----------Dummy Coding Ethnicity/Race-----------
> Hispanic <- ifelse(urad$ethnicity_race == "Hispanic",1,0)
> Other <- ifelse(urad$ethnicity_race == "Other",1,0)
>


> #----------Descriptive Statistics----------------
>
> library(psych)
> describe(urad$prior_research_score)
   vars   n  mean    sd median trimmed   mad min max range skew kurtosis  se
X1    1 121 20.07 20.92      9   18.84 13.34   0  53    53 0.21    -1.77 1.9
> describeBy(urad$prior_research_score, group = urad$female)

 Descriptive statistics by group
group: 0
   vars  n  mean    sd median trimmed mad min max range skew kurtosis   se
X1    1 64 17.69 20.63      0   16.04   0   0  53    53 0.42    -1.65 2.58
------------------------------------------------------------------------------------------------
group: 1
   vars  n  mean   sd median trimmed   mad min max range  skew kurtosis   se
X1    1 57 22.74 21.1     32   22.15 28.17   0  53    53 -0.02    -1.84 2.79
```

```
> describeBy(urad$prior_research_score, group = urad$ethnicity_race)

 Descriptive statistics by group
group: Hispanic
   vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 65 22.34 20.43     29   21.66 29.65   0  53    53 -0.02    -1.77 2.53
------------------------------------------------------------------------------------------------
group: Other
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 19 21.79 21.86     26   21.41 35.58   0  50    50 0.04    -1.98 5.01
------------------------------------------------------------------------------------------------
group: White
   vars  n  mean    sd median trimmed mad min max range skew kurtosis   se
X1    1 37 15.19 21.02      0   13.19   0   0  53    53 0.72    -1.38 3.46

> describeBy(urad$prior_research_score, group = urad$first_gen)

 Descriptive statistics by group
group: 0
   vars  n  mean   sd median trimmed mad min max range skew kurtosis  se
X1    1 61 18.51 21.1      0   17.06   0   0  52    52 0.34    -1.78 2.7
------------------------------------------------------------------------------------------------
group: 1
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 60 21.65 20.79     25   20.65 37.06   0  53    53 0.09    -1.77 2.68
> describeBy(urad$prior_research_score, group = urad$pell)

 Descriptive statistics by group
group: 0
   vars  n  mean    sd median trimmed mad min max range skew kurtosis   se
X1    1 86 19.14 21.47      0    17.7   0   0  53    53 0.32    -1.76 2.31
------------------------------------------------------------------------------------------------
group: 1
   vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 35 22.34 19.62     28   21.93 26.69   0  52    52 -0.06    -1.78 3.32
>
```

```
> #-----------ANOVA Model Building---------------
>
> #---Model 1: Covariates---
> model1 <- lm(prior_research_score ~ female + Hispanic + Other, data = urad)
> summary(model1)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other,
    data = urad)

Residuals:
    Min      1Q  Median      3Q     Max
-25.897 -19.568  -8.785  20.103  41.218

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.782      4.037   2.918  0.00422 **
female         6.003      3.823   1.570  0.11908
Hispanic       7.786      4.284   1.817  0.07173 .
Other          8.112      5.924   1.369  0.17351
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.71 on 117 degrees of freedom
Multiple R-squared:  0.04436,	Adjusted R-squared:  0.01986
F-statistic:  1.81 on 3 and 117 DF,  p-value: 0.1491


> sd(model1$residuals)
[1] 20.44912
>
> #Standardized coefficients
> library(lm.beta)
> lm.beta(model1)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other,
    data = urad)

Standardized Coefficients::
(Intercept)      female    Hispanic       Other
         NA   0.1438474   0.1863554   0.1416710
```

```
> #---Model 2: Covariates and First-Generation Status; main effect of first-gen status---
> model2 <- lm(prior_research_score ~ female + Hispanic + Other + first_gen, data = urad)
> summary(model2)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other + 
    first_gen, data = urad)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.177 -18.831  -9.654  19.584  40.693 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   11.307      4.227   2.675  0.00855 **
female         5.826      3.863   1.508  0.13431   
Hispanic       7.523      4.351   1.729  0.08643 . 
Other          7.762      6.011   1.291  0.19917   
first_gen      1.521      3.854   0.395  0.69375   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.78 on 116 degrees of freedom
Multiple R-squared:  0.04564,	Adjusted R-squared:  0.01274 
F-statistic: 1.387 on 4 and 116 DF,  p-value: 0.2428
```

```
> sd(model2$residuals)
[1] 20.4354
> 
> #Standardized coefficients
> lm.beta(model2)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other + 
    first_gen, data = urad)

Standardized Coefficients::
(Intercept)      female    Hispanic       Other   first_gen 
         NA  0.13959028  0.18007552  0.13556090  0.03651205 

> 
> #Calculate differences between models 1 and 2
> anova(model1, model2)
Analysis of Variance Table

Model 1: prior_research_score ~ female + Hispanic + Other
Model 2: prior_research_score ~ female + Hispanic + Other + first_gen
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    117 50180                           
2    116 50113  1    67.317 0.1558 0.6938
```

```
> #---Model 3: Add Pell Grant eligibility; main effect of family income---
> model3 <- lm(prior_research_score ~ female + Hispanic + Other + first_gen + pell, data = urad)
> summary(model3)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other + 
    first_gen + pell, data = urad)

Residuals:
   Min     1Q Median     3Q    Max 
-26.86 -18.73 -10.65  19.64  40.75 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   11.247      4.250   2.647  0.00927 **
female         5.912      3.892   1.519  0.13157   
Hispanic       7.204      4.529   1.591  0.11444   
Other          7.484      6.124   1.222  0.22421   
first_gen      1.293      3.963   0.326  0.74474   
pell           1.201      4.507   0.267  0.79026   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.87 on 115 degrees of freedom
Multiple R-squared:  0.04623,	Adjusted R-squared:  0.004766 
F-statistic: 1.115 on 5 and 115 DF,  p-value: 0.3564
```

```
> sd(model3$residuals)
[1] 20.42909
> 
> #Standardized coefficients
> lm.beta(model3)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other + 
    first_gen + pell, data = urad)

Standardized Coefficients::
(Intercept)      female    Hispanic       Other   first_gen        pell 
         NA  0.14165265  0.17243915  0.13070896  0.03103985  0.02614931 

>
```

```
> #Calculate differences between models 2 and 3
> anova(model2, model3)
Analysis of Variance Table

Model 1: prior_research_score ~ female + Hispanic + Other + first_gen
Model 2: prior_research_score ~ female + Hispanic + Other + first_gen +
    pell
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    116 50113
2    115 50082  1     30.95 0.0711 0.7903
>
```

```
> #---Model 4: Add interaction between Pell and First-Gen---
> model4 <- lm(prior_research_score ~ female + Hispanic + Other + first_gen + pell + first_gen * pell, data = urad)
> summary(model4)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other +
    first_gen + pell + first_gen * pell, data = urad)

Residuals:
    Min      1Q  Median      3Q     Max
-26.778 -19.414  -9.083  19.222  41.610

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      10.390      4.417   2.352   0.0204 *
female            6.031      3.904   1.545   0.1251
Hispanic          7.364      4.544   1.621   0.1078
Other             7.568      6.138   1.233   0.2201
first_gen         2.993      4.601   0.650   0.5167
pell              5.176      7.066   0.732   0.4654
first_gen:pell   -6.506      8.898  -0.731   0.4662
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.91 on 114 degrees of freedom
Multiple R-squared:  0.05069,	Adjusted R-squared:  0.0007223
F-statistic: 1.014 on 6 and 114 DF,  p-value: 0.4195
```

```
> sd(model4$residuals)
[1] 20.38136
>
> #Standardized coefficients
> lm.beta(model4)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other +
    first_gen + pell + first_gen * pell, data = urad)

Standardized Coefficients::
   (Intercept)         female       Hispanic          Other      first_gen           pell first_gen:pell
            NA     0.14450821     0.17627000     0.13216586     0.07182973     0.11264831    -0.12453043

>
```

```
> #Calculate differences between models 3 and 4
> anova(model3, model4)
Analysis of Variance Table

Model 1: prior_research_score ~ female + Hispanic + Other + first_gen +
    pell
Model 2: prior_research_score ~ female + Hispanic + Other + first_gen +
    pell + first_gen * pell
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    115 50082
2    114 49848  1    233.76 0.5346 0.4662
>
```

```
> #---Model 5: Same as Model 4, but with zero scores removed from dataset---
> #Creating urad_2 dataset
> urad_2 <- subset(urad, prior_research_score!=0)
> Hispanic <- ifelse(urad_2$ethnicity_race == "Hispanic",1,0)
> Other <- ifelse(urad_2$ethnicity_race == "Other",1,0)
>
> model5 <- lm(prior_research_score ~ female + Hispanic + Other + first_gen + pell + first_gen * pell, data = urad_2)
> summary(model5)

Call:
lm(formula = prior_research_score ~ female + Hispanic + Other +
    first_gen + pell + first_gen * pell, data = urad_2)

Residuals:
    Min      1Q  Median      3Q     Max
-27.410  -4.605   1.897   6.519  12.762
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     39.2376     3.1042  12.640   <2e-16 ***
female           0.6715     2.6876   0.250   0.8036
Hispanic         0.1721     3.1111   0.055   0.9561
Other            2.6625     4.2673   0.624   0.5353
first_gen        2.3962     3.4212   0.700   0.4866
pell             2.5635     4.5060   0.569   0.5717
first_gen:pell -12.3703     5.7079  -2.167   0.0346 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.656 on 55 degrees of freedom
Multiple R-squared:  0.1657,    Adjusted R-squared:  0.07466
F-statistic:  1.82 on 6 and 55 DF,  p-value: 0.112
```

```
>
> #-----------Post-Hoc Power Analysis---------------
> #"A" value: 0.39 effect size (Harackiewicz et al., 2014) * sd of Model 4 residuals (20.38)
> #"s" value: standard error of first-gen parameter estimate from Model 4
> #"df" value: degrees of freedom from Model 4
> library(retrodesign)
> retrodesign(A = 7.95, s = 4.55, df = 115)
$power
[1] 0.4080234

$typeS
[1] 0.0003688294

$exaggeration
[1] 1.541408
```

```
>
> #-----------Assumptions Testing---------------
>
> #---Normality---
> #Because half of sample size scored 0, this may affect the distribution of the error terms
> #Histogram of residuals from Model 4
> hist(model4$residuals, freq = F)
> #Add kernel density curve
> lines(density(model4$residuals), col = "green", lwd = 2)
> #Add normal curve
> curve(dnorm(x, mean = mean(model4$residuals), sd = sd(model4$residuals)), col = 'darkblue', lwd = 2, add = TRUE, yaxt = 'n')
>
> #Histogram of residuals from Model 5 (omitting zero scores)
> hist(model5$residuals, freq = F)
> #Add kernel density curve
> lines(density(model5$residuals), col = "green", lwd = 2)
> #Add normal curve
> curve(dnorm(x, mean = mean(model5$residuals), sd = sd(model5$residuals)), col = 'darkblue', lwd = 2, add = TRUE, yaxt = 'n')
>
```

```
> #---Homogeneity of variance---
> plot(model4$residuals~model4$fitted)
>
> #Calculate standard deviation of model4 residuals
> sd <- sd(model4$residuals)
> #Add +1/-1 sd to residuals plot
> abline(h = sd, col = 'blue')
> abline(h = -sd, col = 'blue')
>
> #---Independence---
> #Model 4 plot
> plot(model4$residuals)
>
> #Model 5 plot, with zero scores removed
> plot(model5$residuals)
>
```

```
> #---Linearity---
> #No test necessary as all predictor variables are categorical
>
> #---No measurement error---
> #Calculating Cronbach's Alpha for prior research experience scale
> library(readr)
> score <- read_csv("~/Desktop/PhD Program/EDPY 603:504/603 Final Project/score.csv")
Rows: 63 Columns: 10
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (10): ecure_id, research_exp_dummy, q5, q8a, q8b, q6, q7a, q7b, q10a, q10b


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.
> #Recode missing data from 999 to NA
> score[score == "999"] <- NA
> #Listwise deletion of rows with missing data
> complete_score <- score[complete.cases(score), ]
> View(complete_score)
>
> library(psych)
> scale <- data.frame(complete_score$q5, complete_score$q8a, complete_score$q8b, complete_score$q6,
+                     complete_score$q7a, complete_score$q7b, complete_score$q10a, complete_score$q10b)
> alpha(scale)
```

```
 Item statistics
                     n raw.r std.r r.cor r.drop mean  sd
complete_score.q5   59  0.64  0.61  0.52   0.49  4.7 2.1
complete_score.q8a  59  0.81  0.83  0.83   0.75  5.5 1.6
complete_score.q8b  59  0.79  0.81  0.81   0.72  5.5 1.6
complete_score.q6   59  0.60  0.58  0.50   0.46  4.5 1.8
complete_score.q7a  59  0.57  0.56  0.46   0.42  4.1 1.9
complete_score.q7b  59  0.65  0.65  0.57   0.53  4.7 1.8
complete_score.q10a 59  0.87  0.88  0.90   0.82  5.3 1.6
complete_score.q10b 59  0.81  0.82  0.83   0.74  5.3 1.7

Non missing response frequency for each item
                       1    2    3    4    5    6    7 miss
complete_score.q5   0.10 0.10 0.08 0.17 0.05 0.22 0.27    0
complete_score.q8a  0.07 0.00 0.02 0.07 0.20 0.37 0.27    0
complete_score.q8b  0.05 0.03 0.02 0.10 0.12 0.39 0.29    0
complete_score.q6   0.08 0.07 0.12 0.17 0.22 0.19 0.15    0
complete_score.q7a  0.10 0.19 0.10 0.14 0.20 0.19 0.08    0
complete_score.q7b  0.07 0.08 0.07 0.14 0.29 0.17 0.19    0
complete_score.q10a 0.07 0.03 0.00 0.10 0.19 0.42 0.19    0
complete_score.q10b 0.07 0.03 0.02 0.12 0.17 0.39 0.20    0
>
> #---All predictor variables included in model---
> #Addressed in write-up
>
```

```
Reliability analysis
Call: alpha(x = scale)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
      0.86      0.87    0.89      0.45 6.4 0.029    5 1.2     0.39

 lower alpha upper     95% confidence boundaries
0.8 0.86 0.91

 Reliability if an item is dropped:
                    raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
complete_score.q5        0.86      0.86    0.89      0.47 6.3    0.029 0.038  0.40
complete_score.q8a       0.82      0.83    0.85      0.42 5.0    0.036 0.029  0.39
complete_score.q8b       0.83      0.84    0.85      0.42 5.1    0.035 0.028  0.39
complete_score.q6        0.86      0.87    0.88      0.48 6.5    0.029 0.035  0.40
complete_score.q7a       0.86      0.87    0.89      0.49 6.7    0.028 0.033  0.43
complete_score.q7b       0.85      0.86    0.88      0.47 6.1    0.031 0.042  0.40
complete_score.q10a      0.81      0.82    0.84      0.40 4.7    0.038 0.025  0.38
complete_score.q10b      0.82      0.83    0.85      0.42 5.0    0.036 0.029  0.39
```

References

Adedokun, O. A., Bessenbacher, A. B., Parker, L. C., Kirkham, L. L., & Burgess, W. D. (2013). Research skills and STEM undergraduate research students' aspirations for research careers: Mediating effects of research self-efficacy: RESEARCH SKILLS AND STEM UNDERGRADUATE RESEARCH. *Journal of Research in Science Teaching*, *50*(8), 940–951. https://doi.org/10.1002/tea.21102

Adedokun, O. A., Parker, L. C., Childress, A., Burgess, W., Adams, R., Agnew, C. R., Leary, J., Knapp, D., Shields, C., Lelievre, S., & Teegarden, D. (2014). Effect of Time on Perceived Gains from an Undergraduate Research Program. *CBE—Life Sciences Education*, *13*(1), 139–148. https://doi.org/10.1187/cbe.13-03-0045

Bowman, N. A., & Holmes, J. M. (2018). Getting off to a good start? First-year undergraduate research experiences and student outcomes. *High Educ*.

Brown, A. (2022, June 7). *About 5% of young adults in the U.S. say their gender is different from their sex assigned at birth*. Pew Research Center. Retrieved December 7, 2022, from https://www.pewresearch.org/fact-tank/2022/06/07/about-5-of-young-adults-in-the-u-s-say-their-gender-is-different-from-their-sex-assigned-at-birth/

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Giffen, C. J., Blair, S. S., Rouse, D. I., & Hyde, J. S. (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. *Journal of Educational Psychology*, *106*(2), 375–389. https://doi.org/10.1037/a0034679

Kilgo, C. A., & Pascarella, E. T. (2016). Does independent research with a faculty member enhance four-year graduation and graduate/professional degree plans? Convergent results with different analytical methods. *Higher Education*, *71*(4), 575–592. https://doi.org/10.1007/s10734-015-9925-3

Soria, K. M., & Stebleton, M. J. (2012). First-generation students' academic engagement and retention. *Teaching in Higher Education*, *17*(6), 673–685. https://doi.org/10.1080/13562517.2012.666735

Thiry, H., Weston, T. J., Laursen, S. L., & Hunter, A.-B. (2012). The Benefits of Multi-Year Research Experiences: Differences in Novice and Experienced Students' Reported Gains from Undergraduate Research. *CBE—Life Sciences Education*, *11*(3), 260–272. https://doi.org/10.1187/cbe.11-11-0098