Career Choice and Academic Performance

Chris Cioffi, Kristina Frazier, Aidan Hennessy, Mike McHenry

Overview

1. Introduction and Question
2. Background and Related Literature
3. Data
4. Exploratory Analysis
5. Modeling
6. Diagnostic Tools
7. Remedial Measures
8. Conclusion

What Do You Want to Be When You Grow Up?

- During high school, young adults are often asked to make decisions regarding post-secondary education that can have a profound and lasting impact on their lives in the future.
- We investigate what factors in high school may be related to future academic performance.

Research Question

- Question: How is college GPA related to prospective career path in high school? How are other characteristics about a student’s background and high school environment related to their college GPA?
- This study aims to investigate whether students who have a desired future career path in the 9th grade perform better than students who do not, and if choice of career path matters.

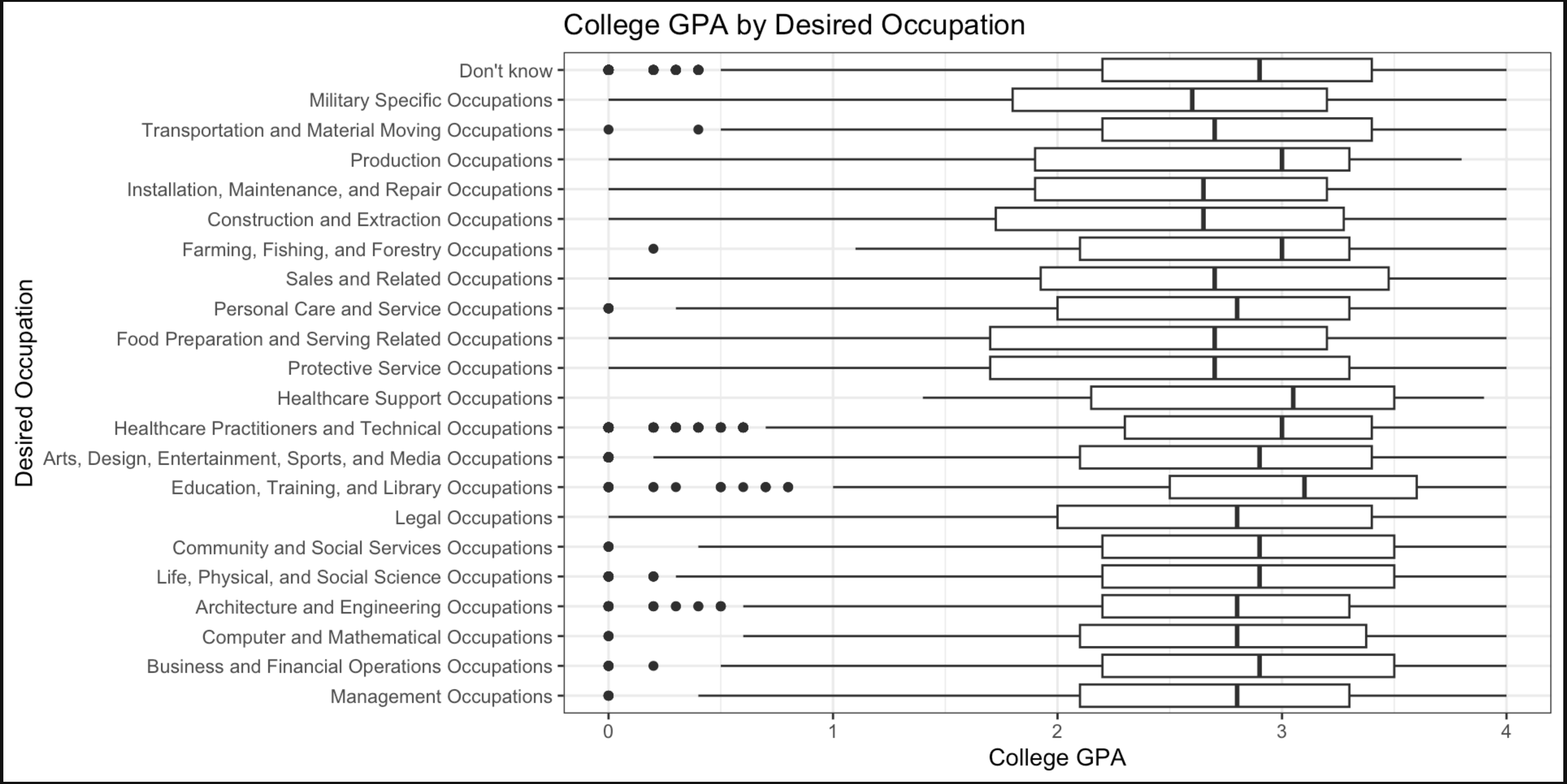
Data

- High School Longitudinal Study of 2009 (HSLS:09) from the National Center for Education Statistics.
    - Interviewed 9th graders across the United States in 2009.
    - Followed up with subjects in three subsequent interview rounds.
    - Offers a variety of information on students, parents, and school.

Key Variables

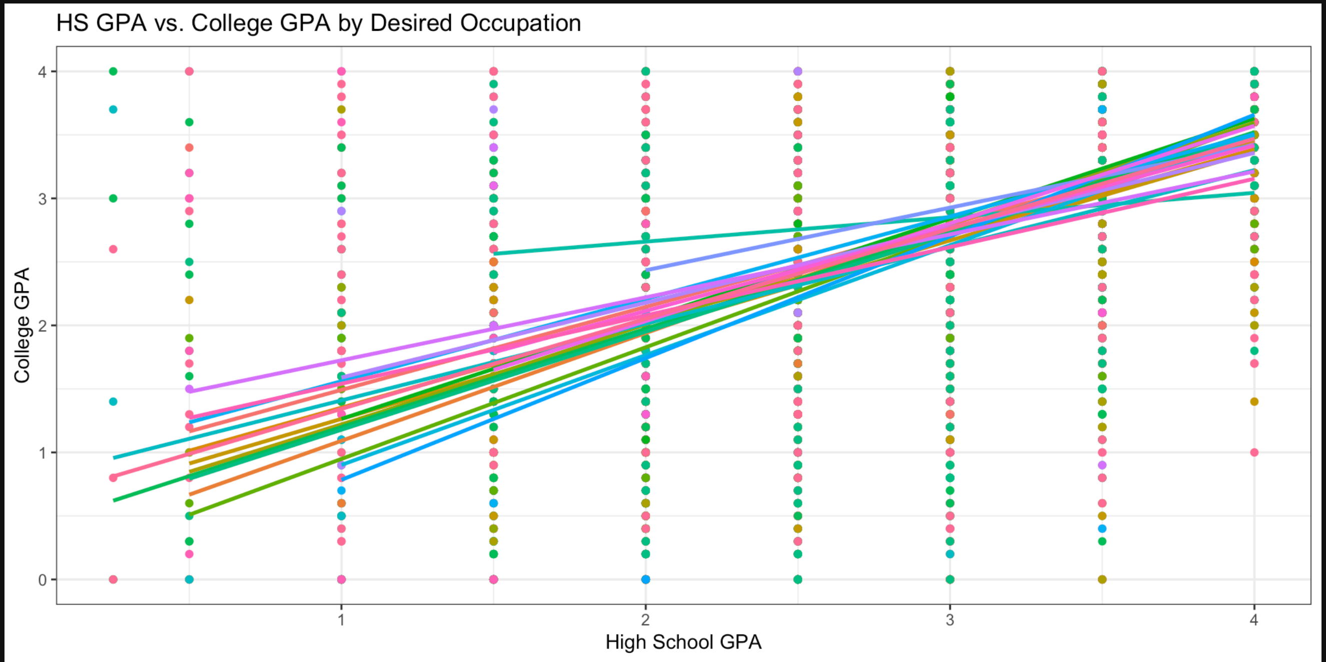
- Response Variable: College GPA
- Primary Predictor of Interest: Desired occupation at age 30.
    - A categorical variable with 22 occupation groups.
- Additional predictors:
    - Academic: High school GPA, credits earned for AP/IB courses, School engagement, Stem/non-stem desired occupation
    - Geographic and Socioeconomic Factors: Family Income, High School urbanicity, High School type

A Look at Desired Occupation



Desired Occupation and Academic Performance

Color-coded by Planned student occupation at age 30



```
            College_GPA    HS_GPA
College_GPA   1.0000000 0.5630064
HS_GPA        0.5630064 1.0000000
```

Model: Simple Linear Regression

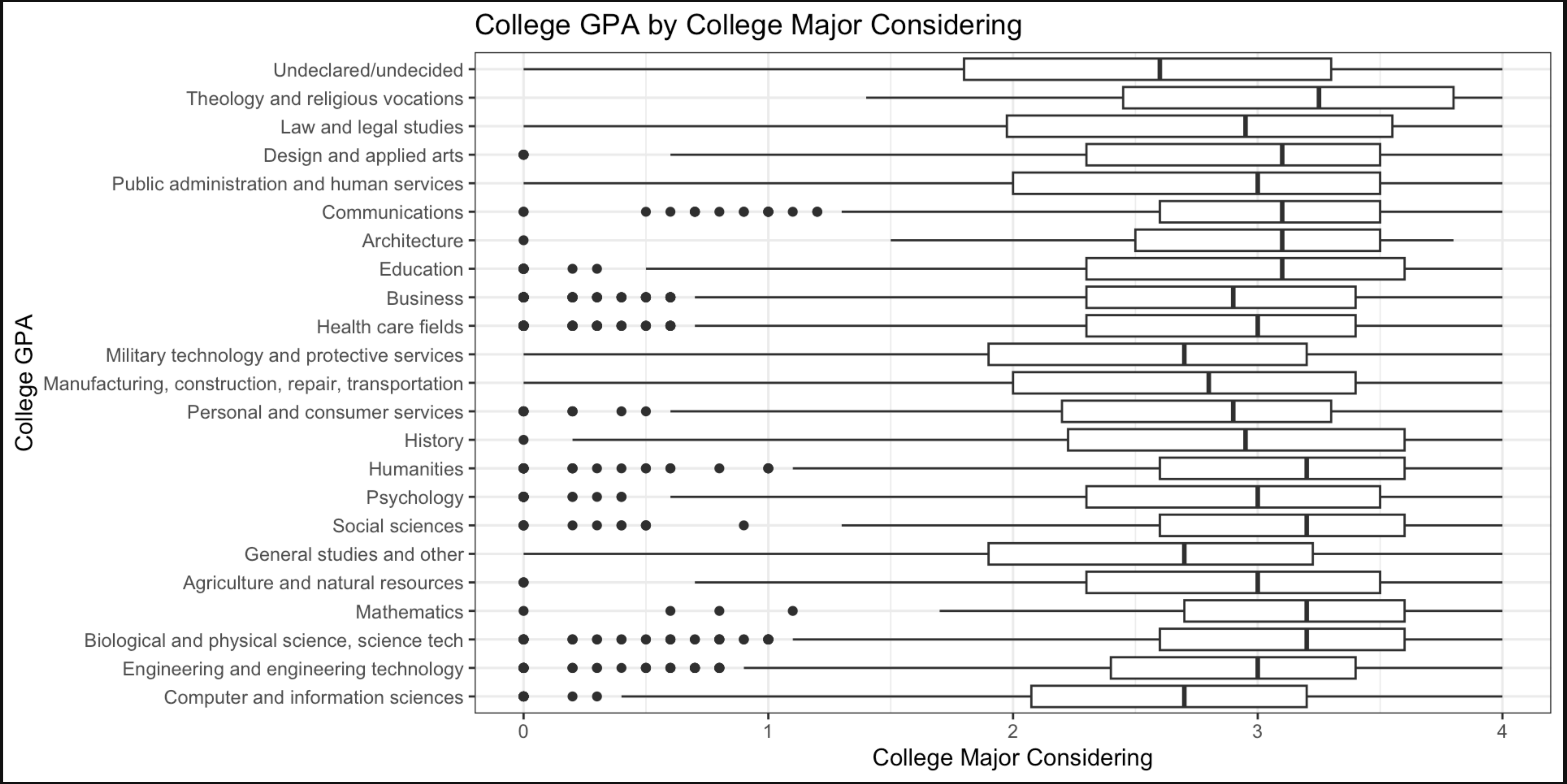
- Set reference group to those students who answered “Don’t Know”.
- Model takes the form of $\mathrm{College\_GPA} = \beta_0 + \beta_1 \mathrm{future\_job} + \epsilon$.

Results: Simple Linear Regression

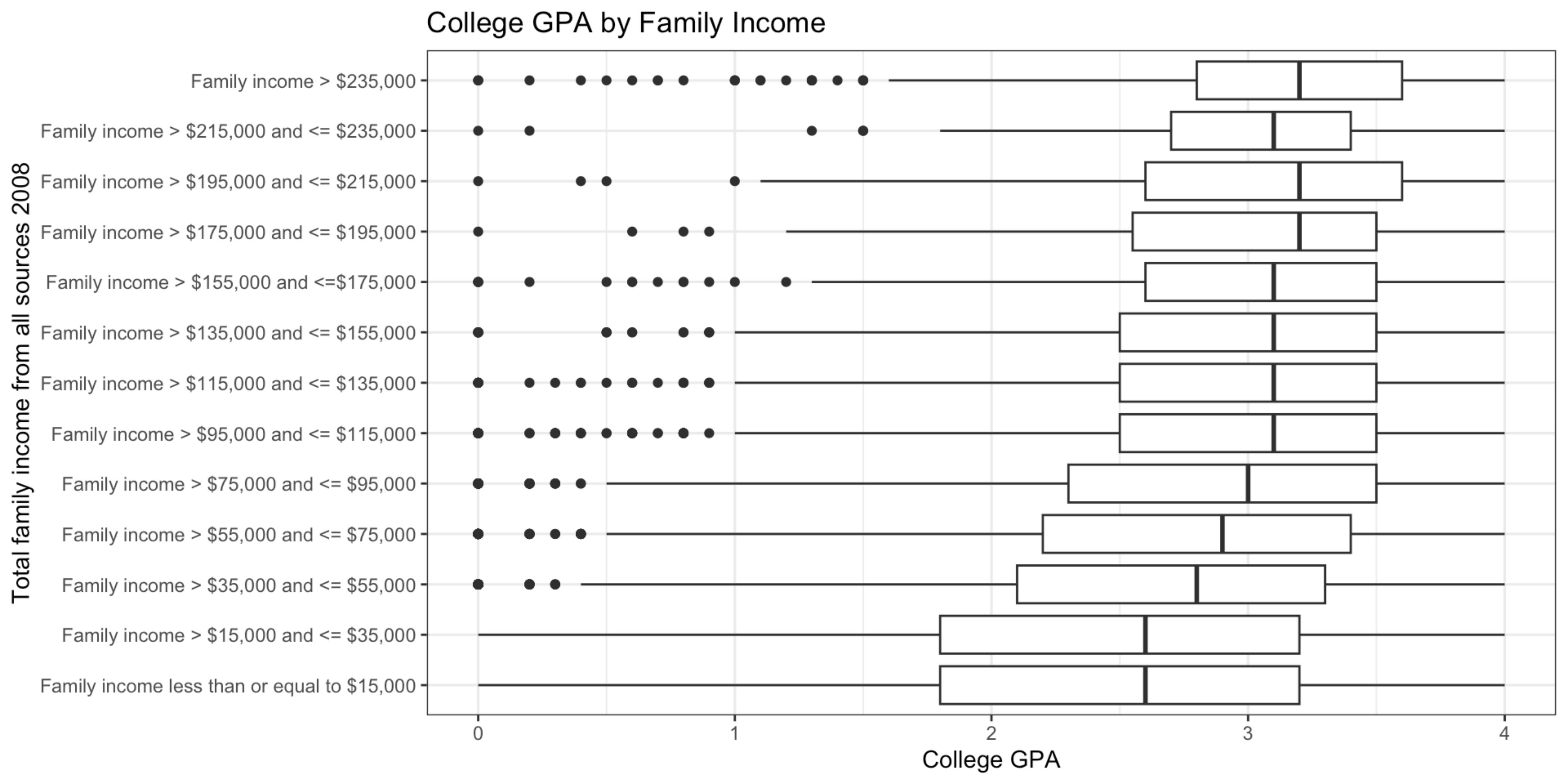
- Showing only results with a p-value < 0.10.

```
# A tibble: 7 × 5
  term                                      estimate std.error statistic
p.value
  <chr>                                        <dbl>     <dbl>     <dbl>
<dbl>
1 (Intercept)                                  2.72     0.0170    160.   0
2 Education, Training, and Library Occupat…    0.192    0.0454      4.24
2.28e-5
3 Arts, Design, Entertainment, Sports, and…   -0.109    0.0303     -3.60
3.16e-4
4 Protective Service Occupations              -0.317    0.0595     -5.33
1.00e-7
5 Food Preparation and Serving Related Occ…   -0.329    0.0878     -3.75
1.79e-4
6 Installation, Maintenance, and Repair Oc…   -0.271    0.104      -2.62
8.84e-3
```

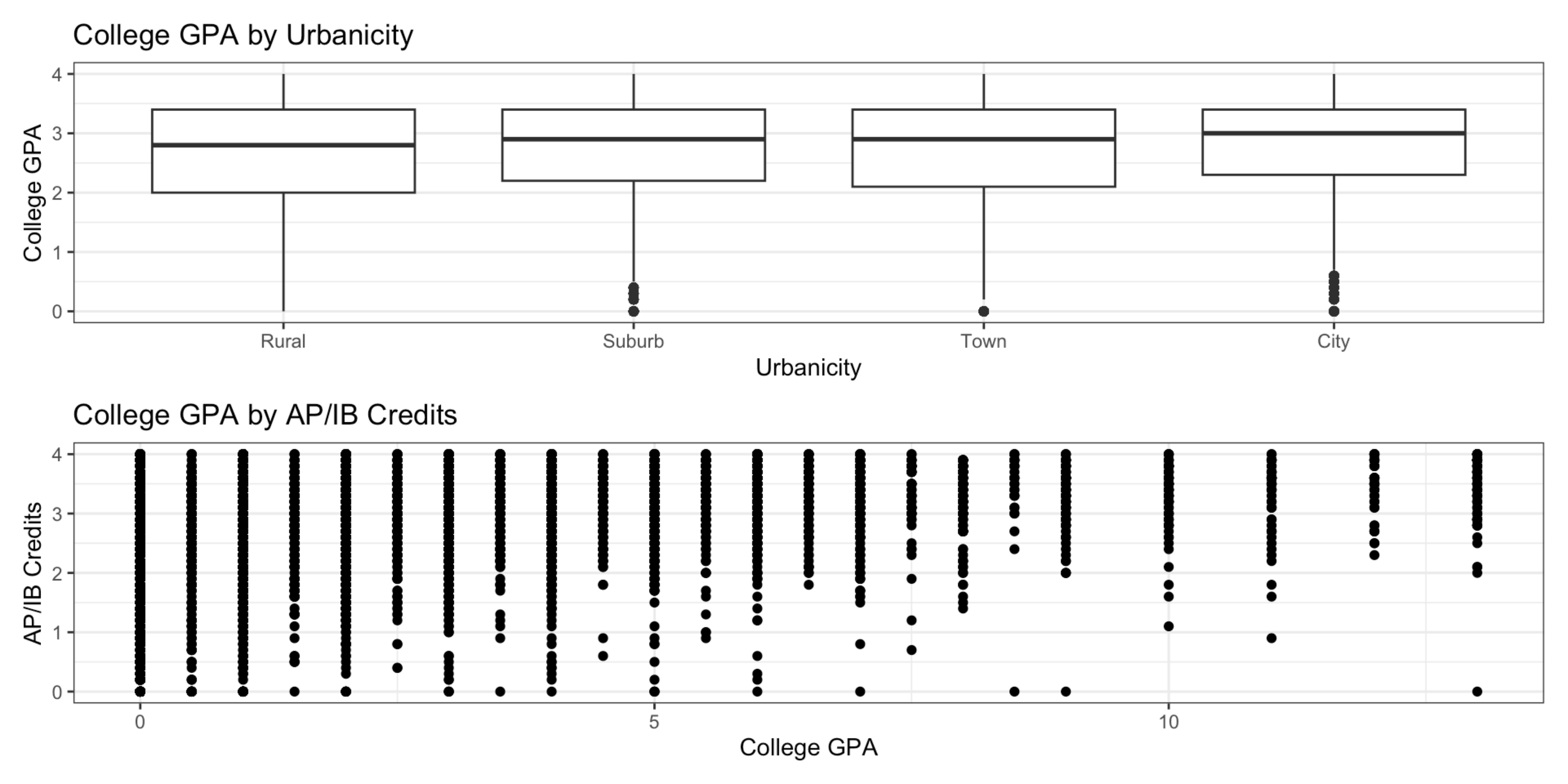
Additional Variables



Additional Variables



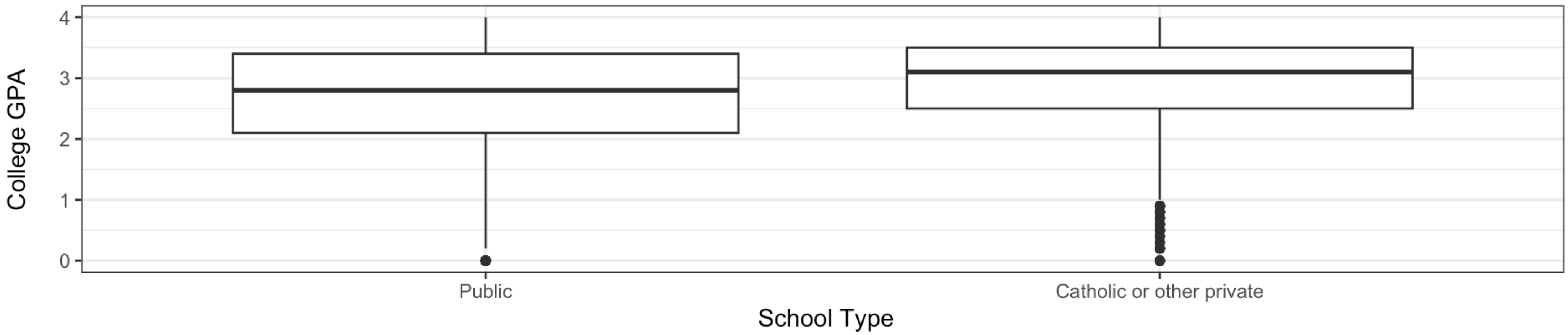
Additional Variables

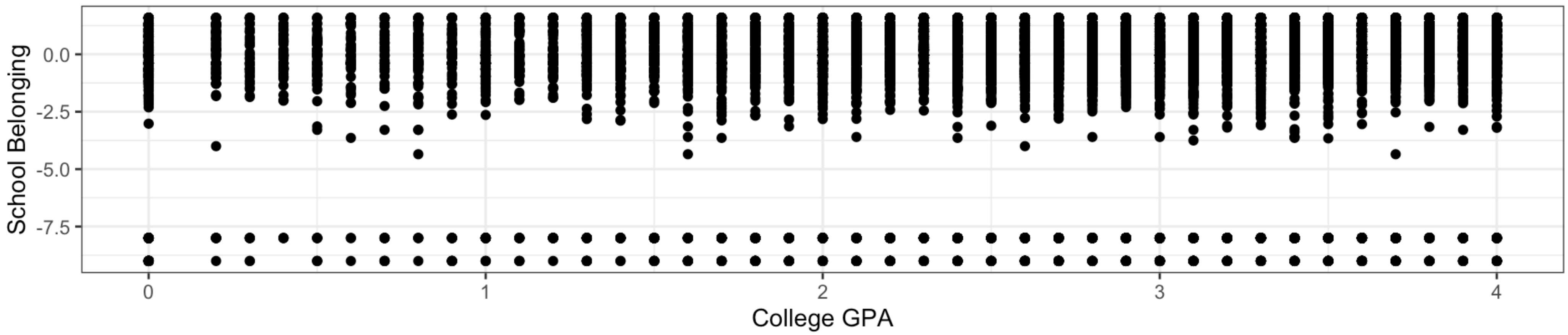


College GPA by Urbanicity

College GPA by AP/IB Credits

Additional Variables


College GPA by School Type


College GPA by Student’s Sense of School Belonging

Multiple Linear Regression

Initial MLR Model:

$$\text{College\_GPA} = \beta_0 + \beta_1\, \text{future\_job} + \beta_2\, \text{college\_gpa} + \beta_3\, \text{major\_considering} + \beta_4\, \text{family\_income} + \beta_5\, \text{credits} + \beta_6\, \text{school\_type} + \beta_7\, \text{urbanicity} + \beta_8\, \text{school\_belonging} + \epsilon$$

- Adjusted R^2: 0.3563
- F-statistic: 66.59 on 62 and 7286 DF, p-value: < 2.2e-16

Remove Urbanicity & School Belonging?

- Urbanicity & School Belonging: All Betas are insignificant at alpha = 0.10

Lack of Fit Test

- Null Hypothesis: $\beta_7 \text{urbanicity} = \beta_8 \text{school\_belonging} = 0$
- Alternative hypothesis: either of the betas for these variables is a non-zero value
- Use alpha = 0.10

Remove Urbanicity & School Belonging?

```
Analysis of Variance Table

Model 1: X5GPAALL ~ X1STU30OCC2 + X3TGPAACAD + X4ENTRYMAJ23 + X1FAMINCOME + 
    X3TCREDAPIB + X1CONTROL
Model 2: X5GPAALL ~ X1STU30OCC2 + X3TGPAACAD + X4ENTRYMAJ23 + X1LOCALE + 
    X1FAMINCOME + X3TCREDAPIB + X1CONTROL + X1SCHOOLBEL
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1   7290 3608.0                         
2   7286 3605.7  4    2.2764 1.15  0.331
```

- With a P-value of 0.331, there is insufficient evidence to reject the null hypothesis that the values for the betas of these two predictors are not zero.
- The lack of significant relationship between Urbanicity & School Belonging was seen in earlier plots.

Variable Selection

New MLR Model:

$$\text{College\_GPA} = \\ \beta_0 + \beta_1 \, future\_job + \beta_2 \, college\_gpa + \beta_3 \, major\_considering + \\ \beta_4 \, family\_income + \beta_5 \, credits + \beta_6 \, school\_type + \epsilon$$

- Stepwise selection did not remove additional variables

Diagnostics

Linearity

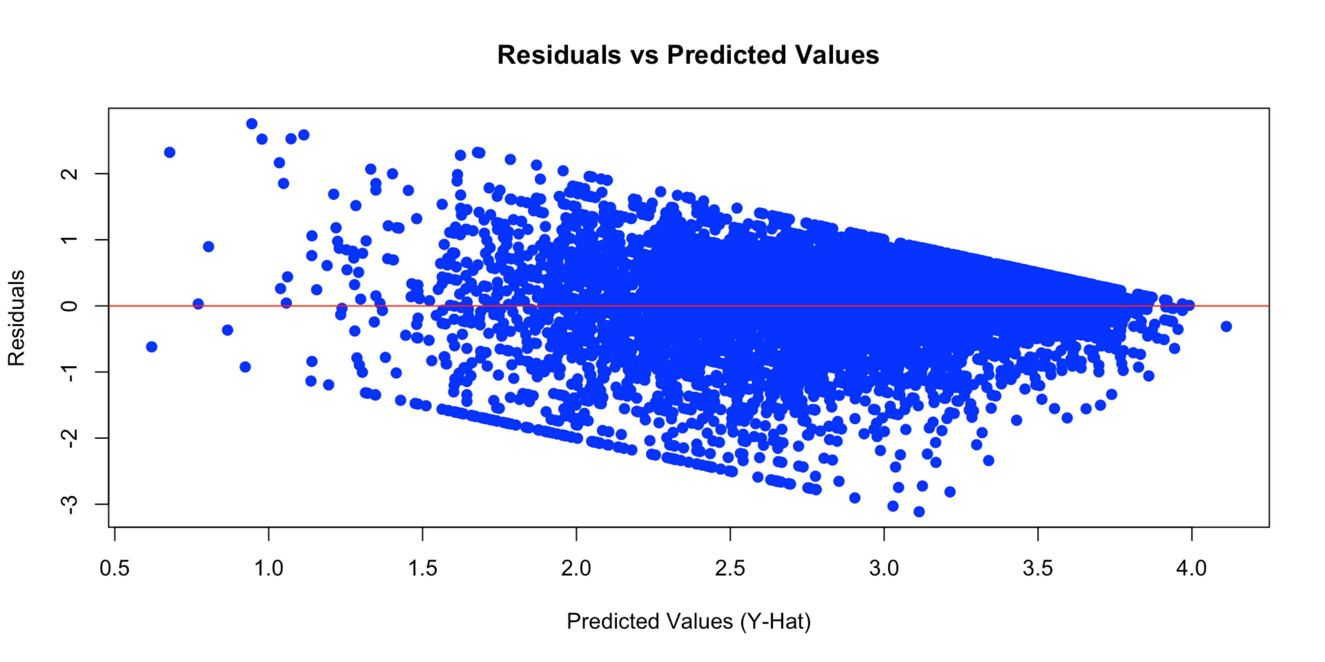
- F-statistic: 71.1 on 58 and 7290 DF, p-value: < 2.2e-16

```
Call:
lm(formula = X5GPAALL ~ X1STU30OCC2 + X3TGPAACAD + X4ENTRYMAJ23 +
    X1FAMINCOME + X3TCREDAPIB + X1CONTROL, data = MLR_all)

Residuals:
     Min       1Q   Median       3Q      Max
-3.11365 -0.34477  0.09234  0.43827  2.75491

Coefficients:

Estimate
(Intercept)
0.4541565
X1STU30OCC2Management Occupations
0.0153304
```

Diagnostics

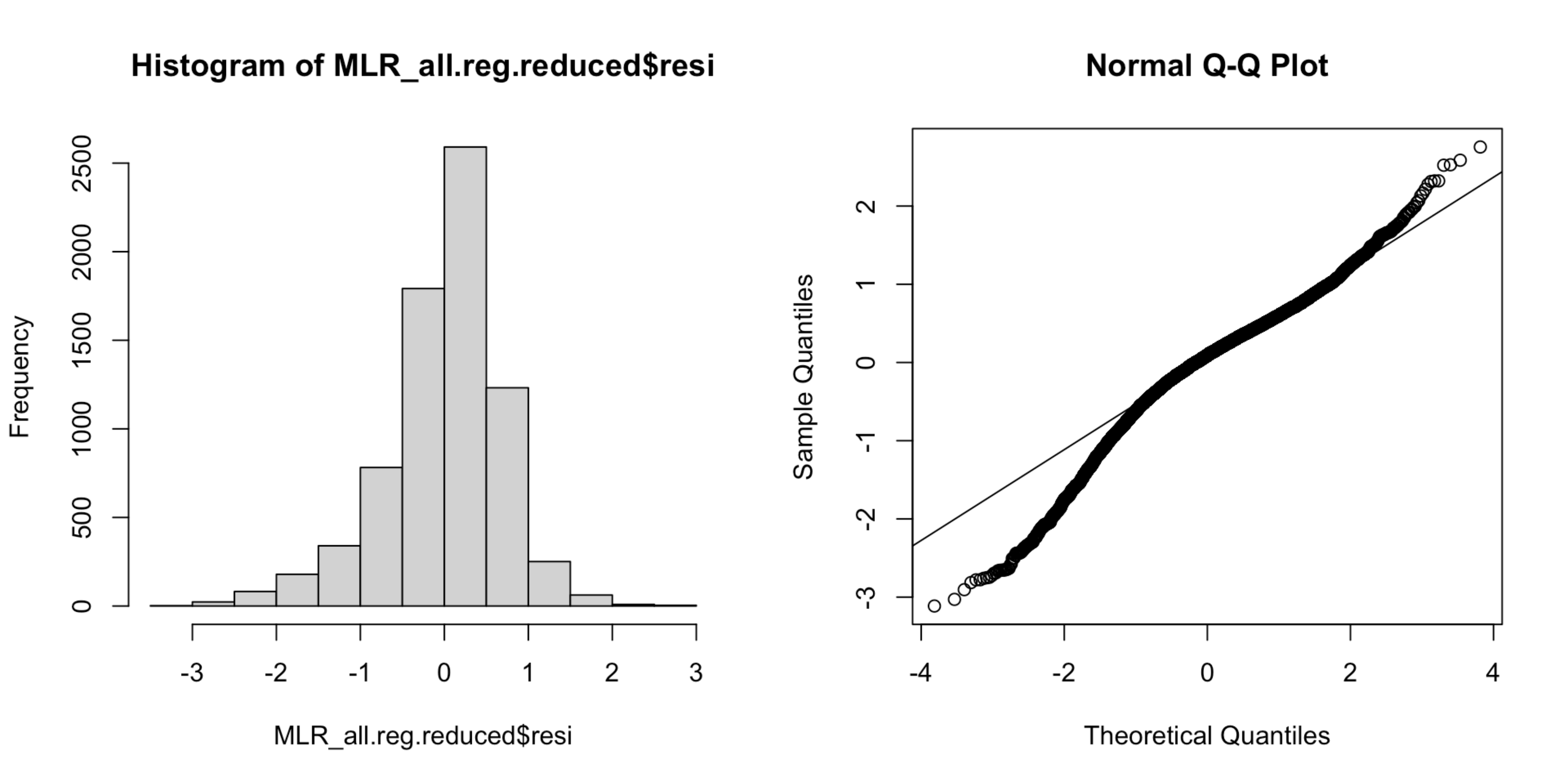
Constant Variance

- Breusch-Pagan test yields a p-value < 0.0001.
- Because the response variable is bound.



Diagnostics

Normality



Remedial Measures

- Need to address non-constant variance first, and then recheck normality assumption
- Try Box-Cox transformation
- Weighted Least Squares
- Recheck model diagnostics.

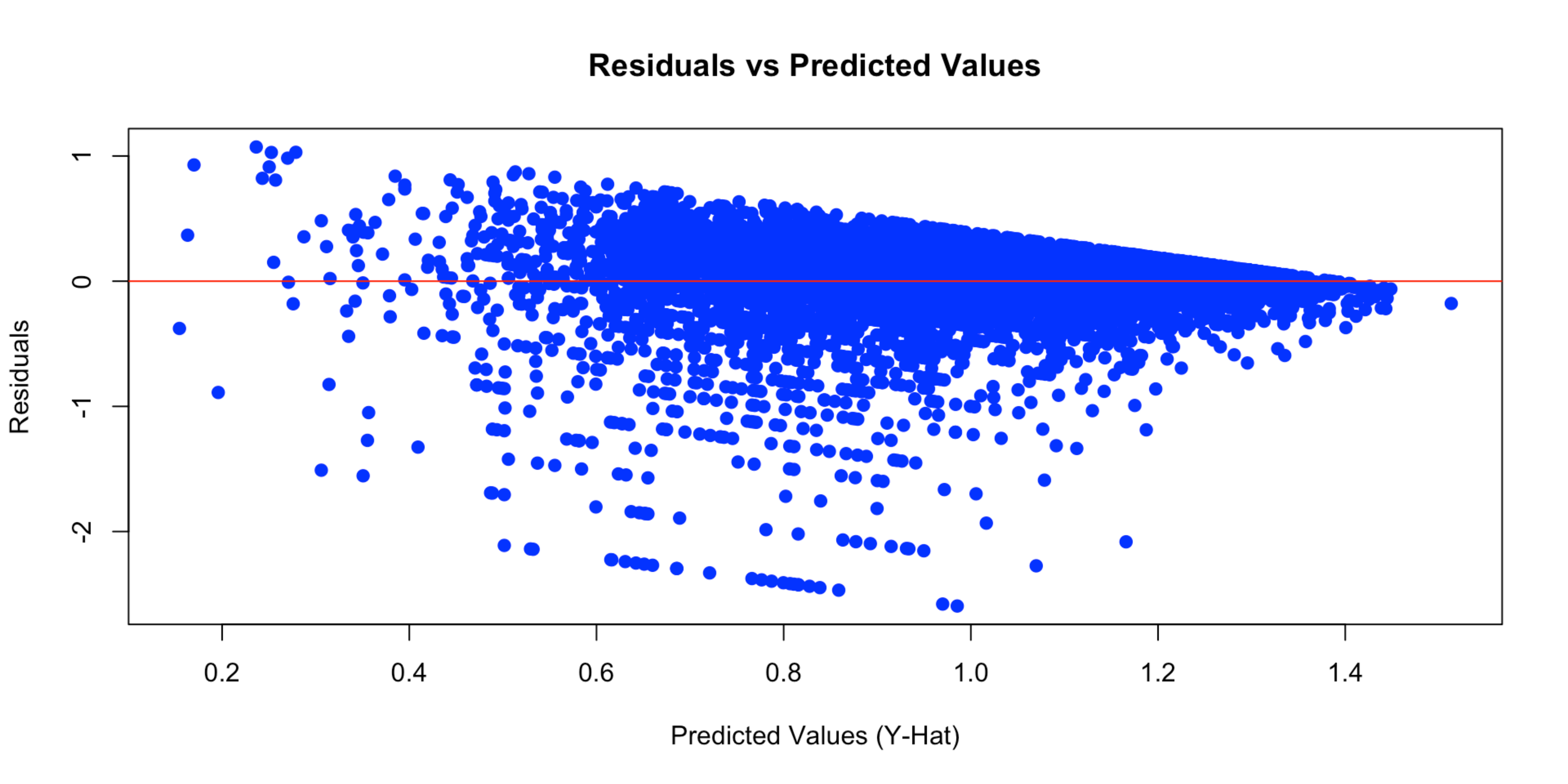
Try Box-Cox Transformation

Need to get rid of all 0.0 GPAs

```
Call:
lm(formula = log(X5GPAALL) ~ X1STU30OCC2 + X3TGPAACAD + X4ENTRYMAJ23 +
    X1FAMINCOME + X3TCREDAPIB + X1CONTROL, data = MLR_all_no_0)

Coefficients:
                                                          (Intercept)
                                                            0.0109720
                                    X1STU30OCC2Management Occupations
                                                            0.0093697
             X1STU30OCC2Business and Financial Operations Occupations
                                                            0.0510748
                     X1STU30OCC2Computer and Mathematical Occupations
                                                           -0.0217388
                  X1STU30OCC2Architecture and Engineering Occupations
                                                            0.0244767
```

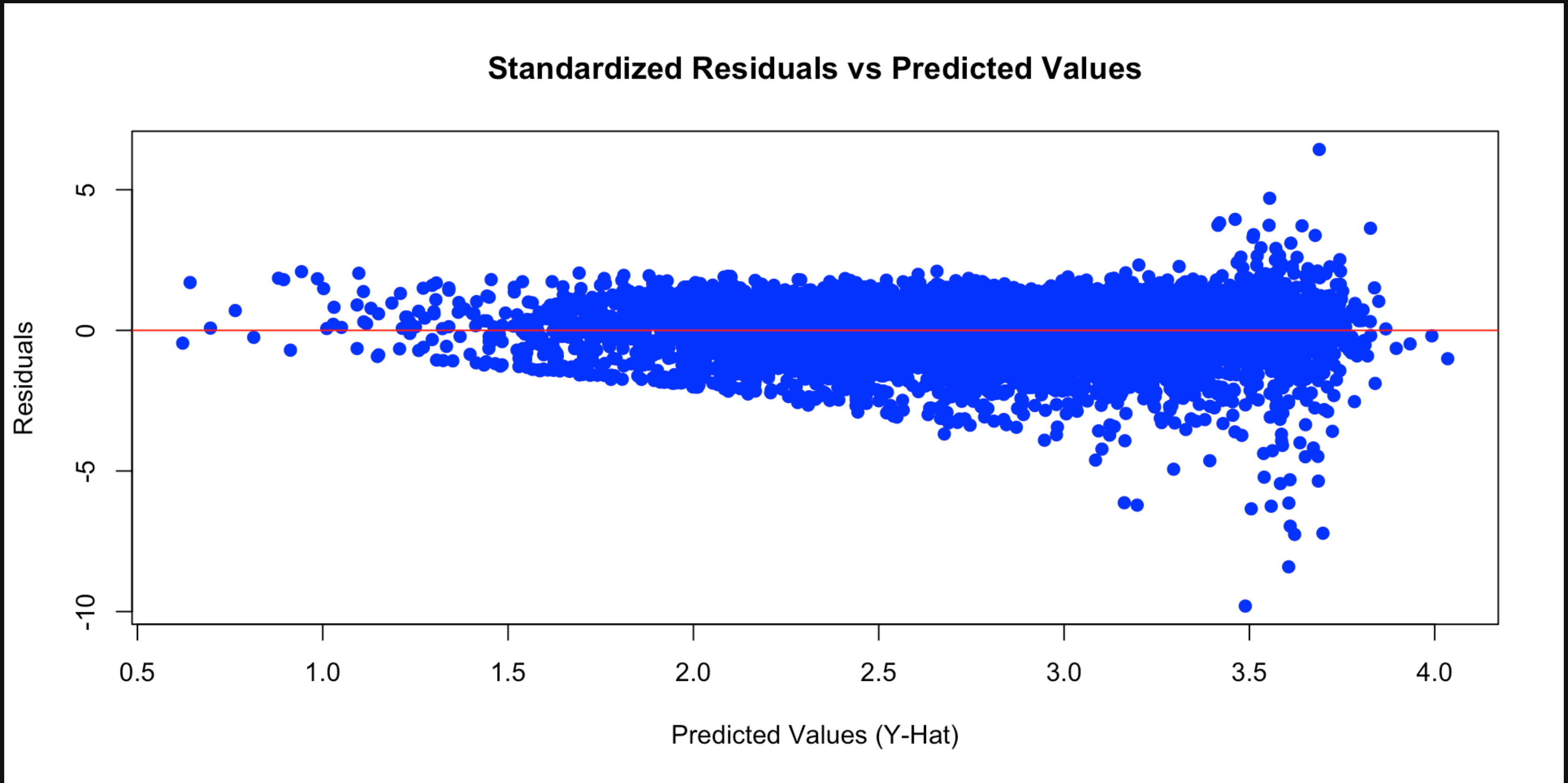
Box-Cox Residual Variance

Still a pattern



Weighted Least Squares on Full Model

Can’t reach coefficient convergence.

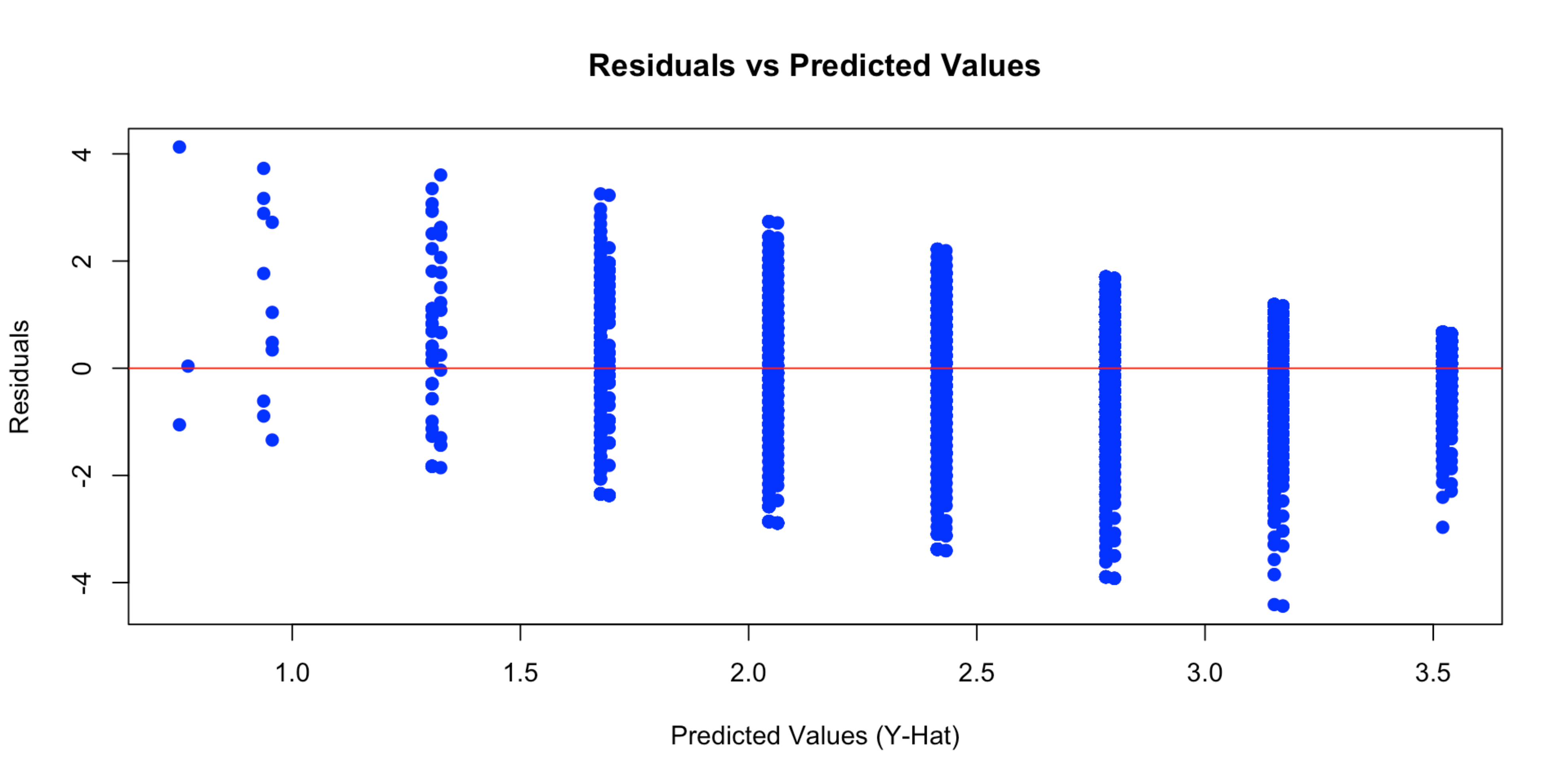


Make Predictor “Don’t Know” vs. “Know” Future Occupation

```
Call:
lm(formula = X5GPAALL ~ career + X3TGPAACAD, data = MLR_all)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1704 -0.3517  0.1174  0.4483  2.9471 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.58715    0.04001   14.67   <2e-16 ***
career      -0.01878    0.01859   -1.01    0.313    
X3TGPAACAD   0.73808    0.01212   60.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

0-1 Career Residual Variance

Still a pattern.



Implications & Limitations

- Non-constant variance could not be addressed through transformation and other remedies given this set of predictors
- Perhaps other predictors not captured by the High School Longitudinal Study are more reliably related to College GPA and choice of future career
- Additionally, other regression approaches, such as Quantile regression, may be useful for exploring the significance of this relationship