# A Comprehensive Study on Chem Grades

[chenziwe] [zhanghon]

# Problems

1. What are the best predictors for performance in CHM 129 and CHM 221?

2. Are there any interactions between explanatory variables? In particular, for what combinations of variables, is a student at risk of getting below B?

3. Are there any delayed risk factors, i.e. factors associated with the phenomenon that some students succeeded in the 100-level course but failed in the 200-level one?

# Thinking Process

1. Why did we choose chemistry grades?
    a. Only Chemistry, Biology and Math have sample size larger than 2000
    b. Only Biology and Chemistry have both 100 and 200-level course information
    c. Chemistry has a more balanced 100 v.s. 200-level sample size (1693 v.s. 1193)
2. How did we clean up the dataset?
    a. Combine tables
    b. Filter out grades not from Grinnell
    c. Filter out audit/satisfactory/withdraw-passing grades
    d. Convert SAT to ACT (and SAT math to ACT math)
    e. Process GSP data (Not invited, invited but not participated, participated)
    f. Divide Primary Academic Interest into Interested in Science and Not Interested in Science
    g. Filter out obviously wrong records
3. How did we look for the best model?

# How did we select these predictors?

- For numerical predictors, we looked at their t-score and p-values
- For categorical predictors, we used ANOVA to conduct model comparison tests between models with and without that particular predictor, and looked at their F-scores and p-values
- We used Y^2 transformation to improve our diagnostic plots
- We also used VIF and residual plots as diagnostic information and adjusted R-squared as model evaluation tool

# Model for CHM Grade Points

Response: Grade_Points$^2$

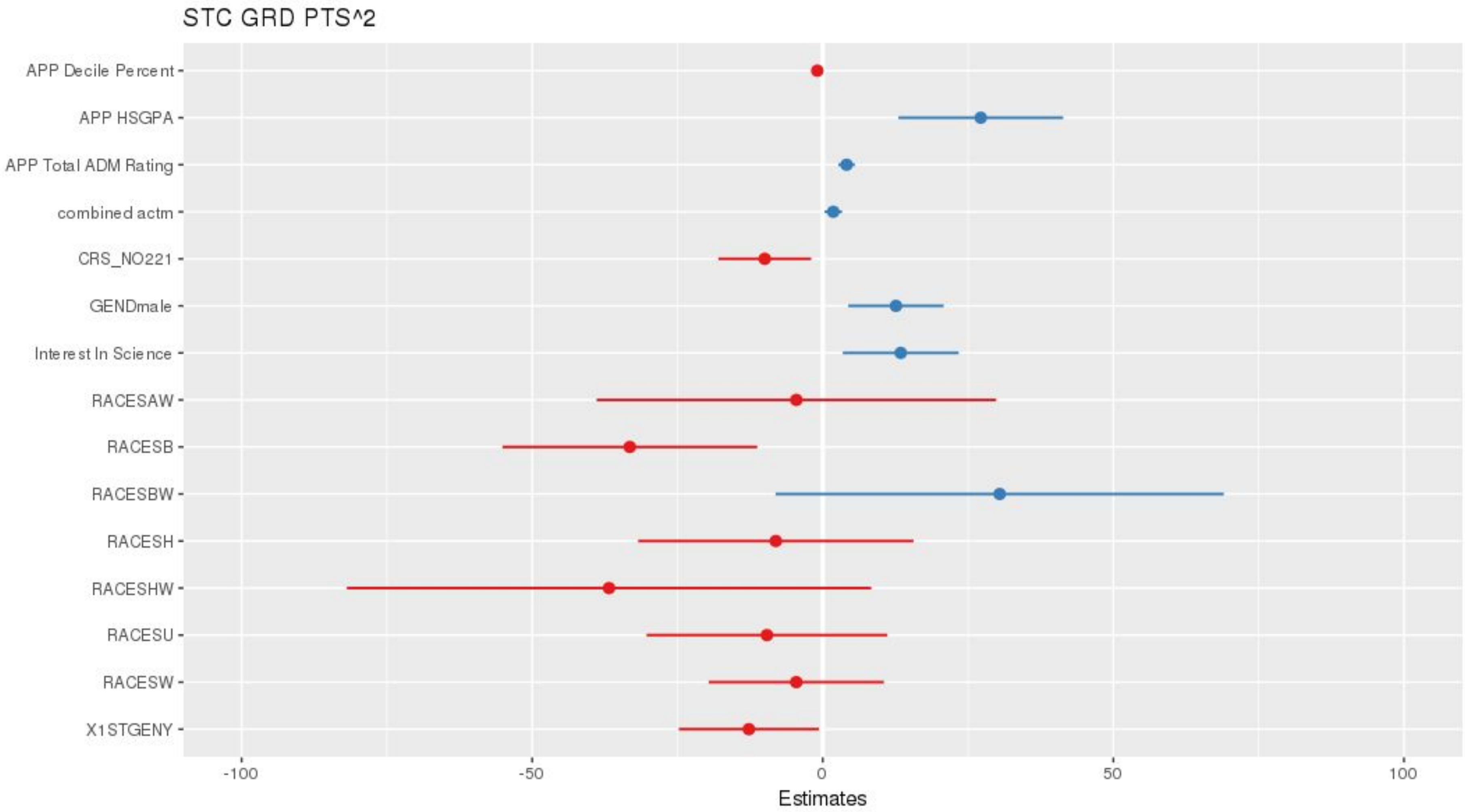
| Significant at $\alpha$=0.05 |
|---|
| First Generation |
| Standardized Test Math Score |
| CHM 221 |
| Race |

| Significant at $\alpha$=0.01 |
|---|
| Gender |
| $\frac{HighSchoolRank}{HighSchoolSize}$ |
| Interest In Science |

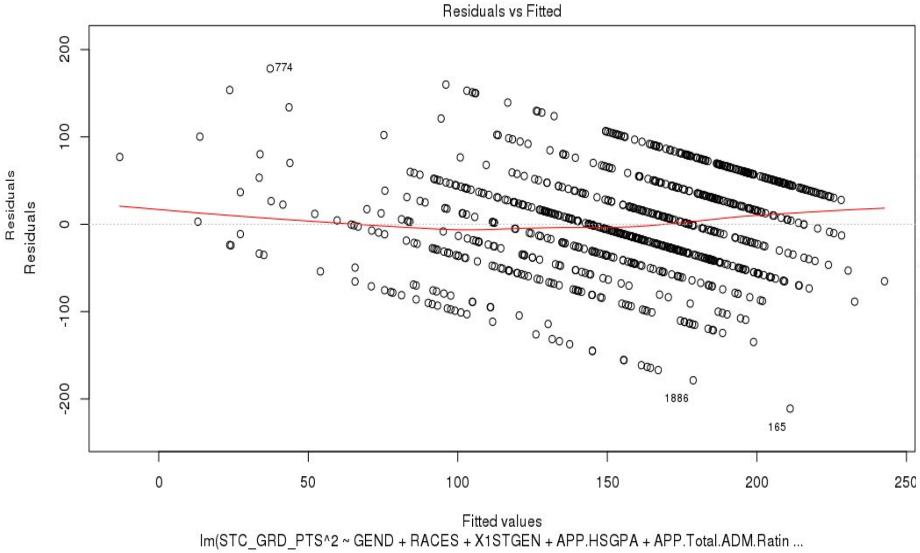
| Significant at $\alpha$=0.001 |
|---|
| High School GPA (Not CIRP) |
| Admissions Rating |

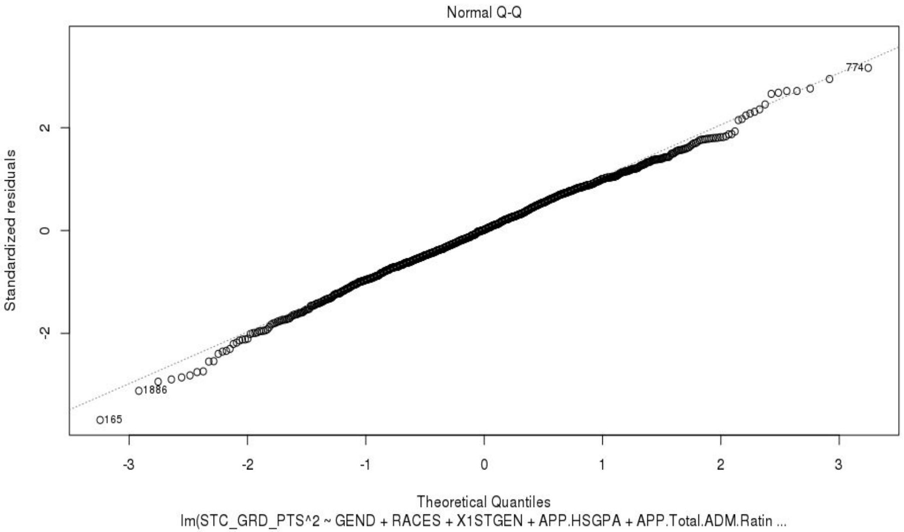
*All predictors in our model are significant!

| Total Degree of Freedom | Model Utility Test P-value | Adjusted R$^2$ |
|---|---|---|
| 850 | <2.2*10$^{-16}$ | 33.03% |


STC GRD PTS^2
APP Decile Percent
APP HSGPA
APP Total ADM Rating
combined actm
CRS_NO221
GENDmale
Interest In Science
RACESAW
RACESB
RACESBW
RACESH
RACESHW
RACESU
RACESW
X1STGENY
-100
-50
0
50
100
Estimates

# Check Model Assumption





*All VIF < 3 except for Admissions Ratings and Race (both around 3)

# Model with Interactions

- Factors examined (all categorical predictors in the previous model)
    a. Gender
    b. Race
    c. 1stGen
    d. Interest in Science
    e. Course Number
- Removed groups that have size smaller than 10
- 3 significant combinations of factors found!

# ANOVA: Main Effects & Interactions

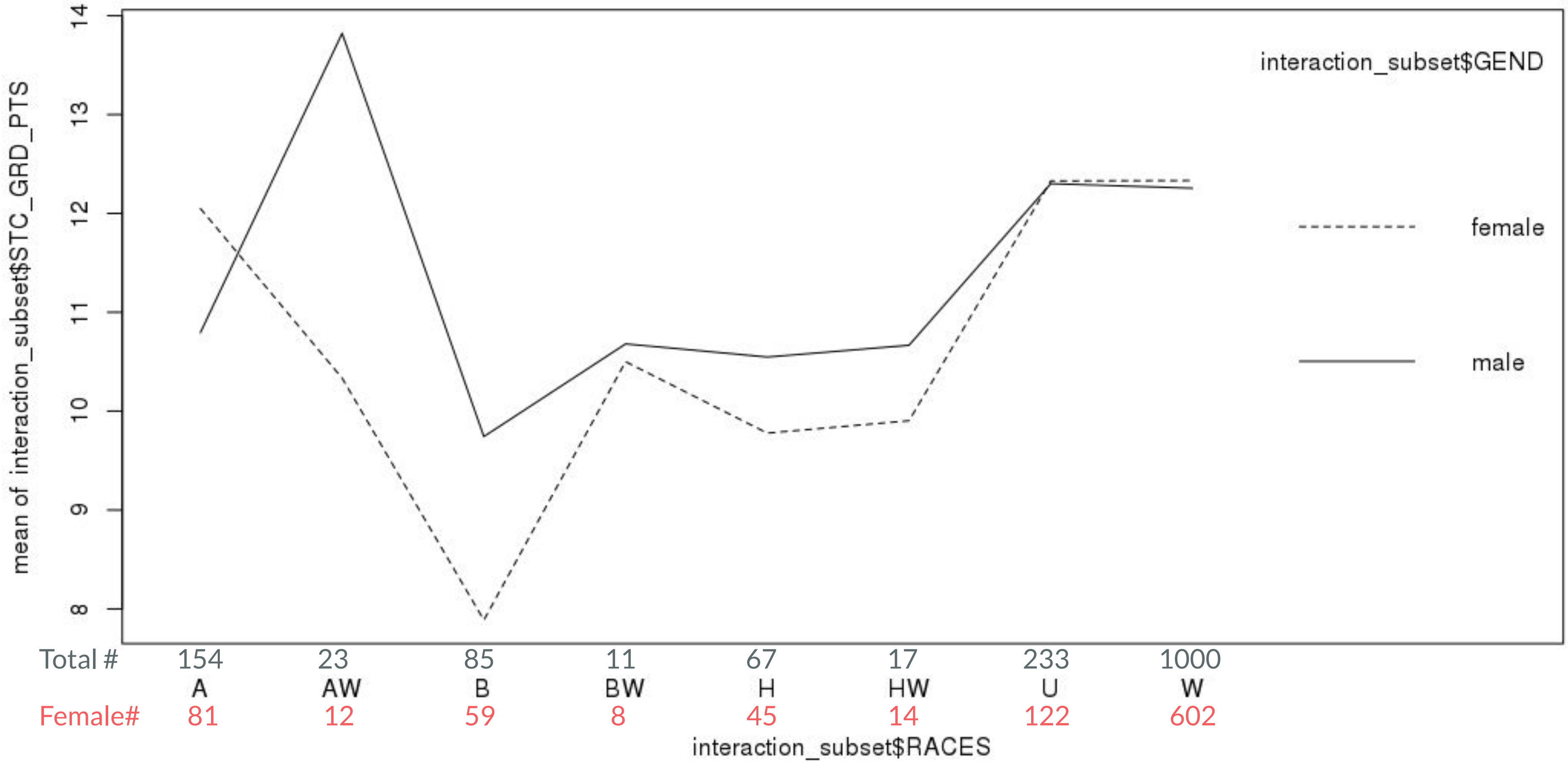
Response: Grade_Points$^2$

| Significant at $\alpha$=0.05 |
|---|
| Gender:Race |
| First Generation:Interest in Science |

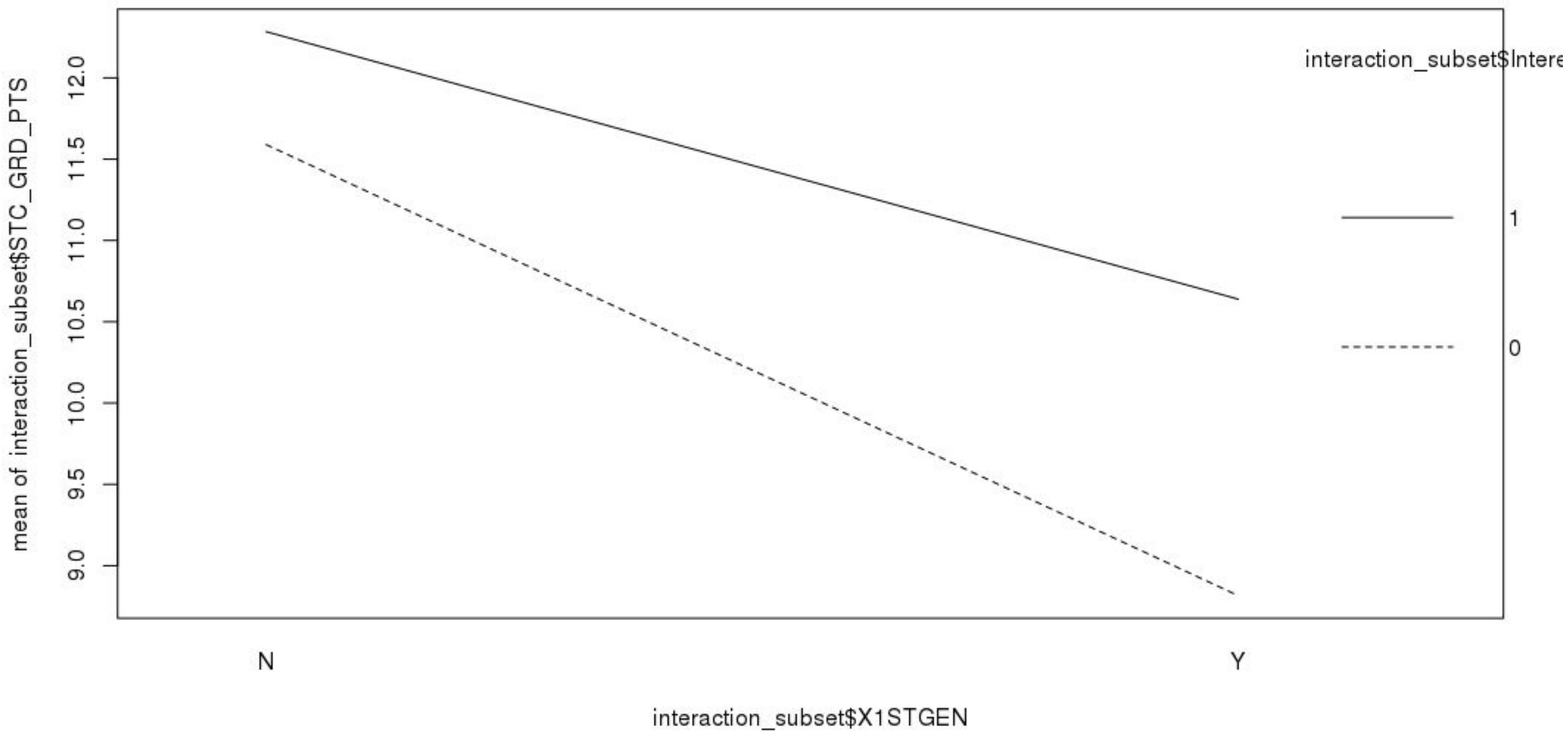
| Significant at $\alpha$=0.01 |
|---|
| Race:First Generation |

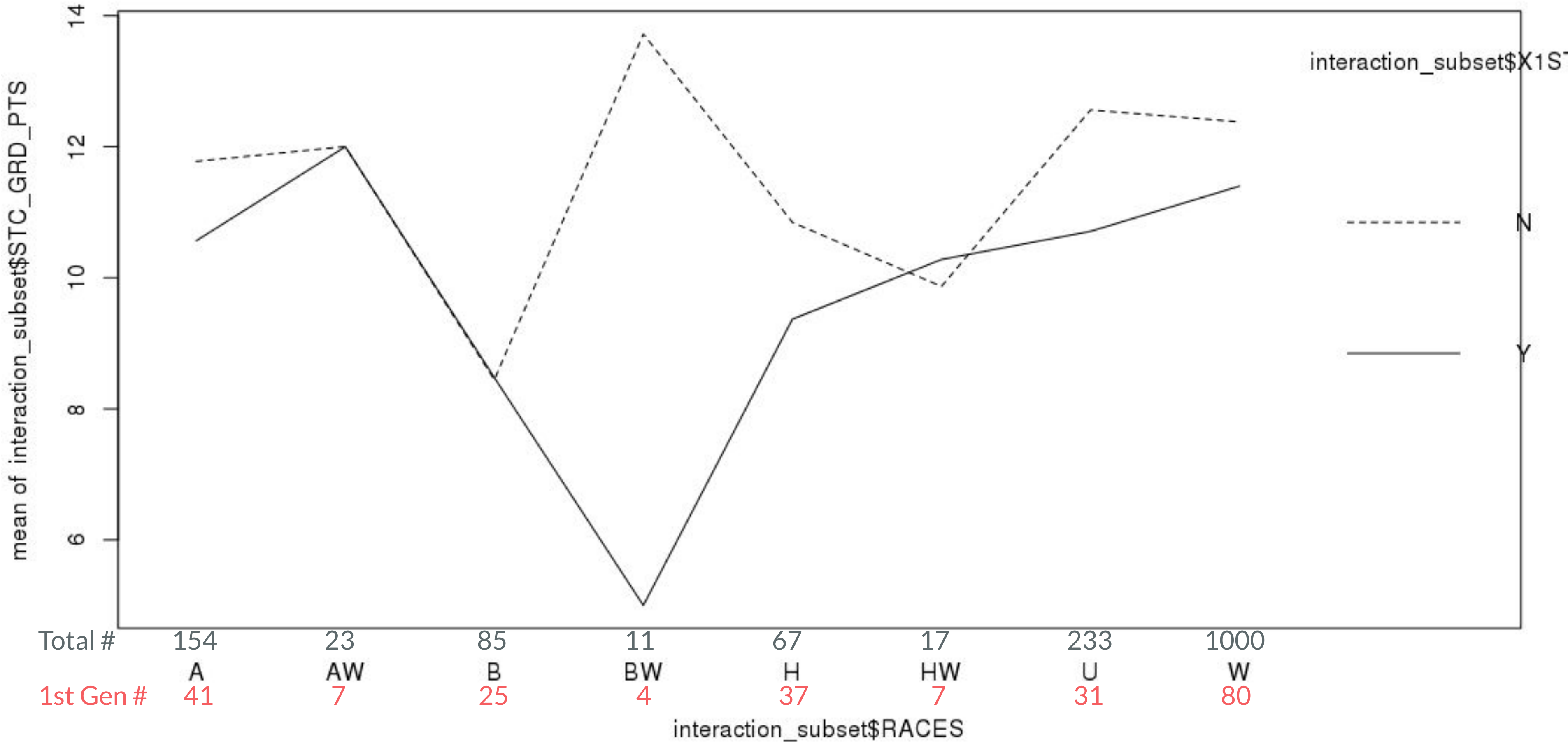
| Significant at $\alpha$=0.001 |
|---|
| Race |
| First Generation |
| Interest in Science |

| Total Degree of Freedom |
|---|
| 1589 |



| | A | AW | B | BW | H | HW | U | W |
|---|---|---|---|---|---|---|---|---|
| Total # | 154 | 23 | 85 | 11 | 67 | 17 | 233 | 1000 |
| Female# | 81 | 12 | 59 | 8 | 45 | 14 | 122 | 602 |


mean of interaction_subset$STC_GRD_PTS
12.0
11.5
11.0
10.5
10.0
9.5
9.0
N
Y
interaction_subset$X1STGEN
interaction_subset$Intere
1
0



| | A | AW | B | BW | H | HW | U | W |
|---|---|---|---|---|---|---|---|---|
| Total # | 154 | 23 | 85 | 11 | 67 | 17 | 233 | 1000 |
| 1st Gen # | 41 | 7 | 25 | 4 | 37 | 7 | 31 | 80 |

# Delayed Risk Factors

1. Cleaning data
   a. Find students who had good 129 grade (>=B)
   b. Find students who also took 221 within those students
   c. Add a column indicating good 221 grade (True if >=B)
2. Finding best model

# Model for Delayed Risk (logistic regression)

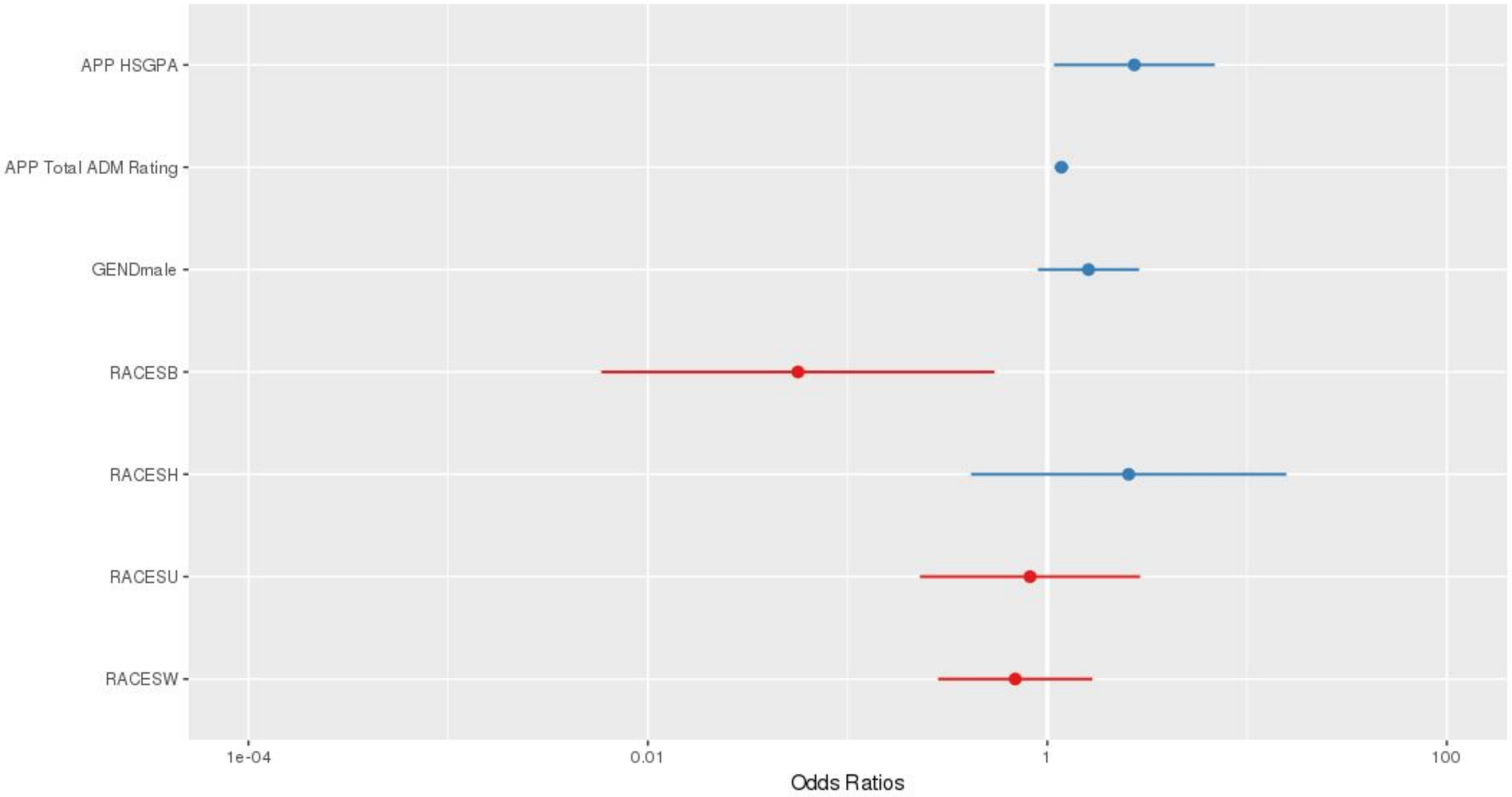
Response: whether the student achieved B or above in CHM 211

| Significant at $\alpha$=0.05 |
|---|
| High School GPA |

| Significant at $\alpha$=0.01 |
|---|
| Race |

| Significant at $\alpha$=0.001 |
|---|
| Admissions Rating |

| Total Degree of Freedom | Hosmer and Lemeshow GOF test |
|---|---|
| 353 | 0.8895 |



GRD GOOD

APP HSGPA
APP Total ADM Rating
GENDmale
RACESB
RACESH
RACESU
RACESW
1e-04
0.01
1
100
Odds Ratios

# Limitations

1. In the process of combining SAT scores with ACT scores, we lost the information of ACT subscores except math
2. Lurking variables, arbitrary success standard, missing values
3. 25% Failure, 75% Success in Model for Delayed Risk Factors
4. In Model for Chem Grade Points, the response is actually a categorical variable
5. Inaccurate records in the dataset: e.g. Grade D can be 4 or 0 grade points; High School Rank/Size/Quotient = 0 (and there are around 100 of them in our data set of only 1,951 observations); High School GPA = 0, etc.
6. Do not have full information about what each variable means