Social Statistics

Interactions

November 30, 2023

Warm Up: midd_survey.csv

- Everyone: regress gpa on number of siblings
- Group 1: Add control for gender to original model
  - → Predict gpa for men with 3 siblings and women with 4 siblings
- Group 2: Add control for class to original model
  - → Predict gpa for middle class student with 2 siblings and upper middle class student with 1 sibling
- Group 3: Add controls for gender and class to original model
  - → Predict gpa for lower class men with 0 siblings

Warm Up - Original Model

Regress gpa on number of siblings

```
gpa_sibs_model <- lm(gpa ~ siblings,
  data = midd_survey)

summary(gpa_sibs_model)
```

Warm Up - Original Model

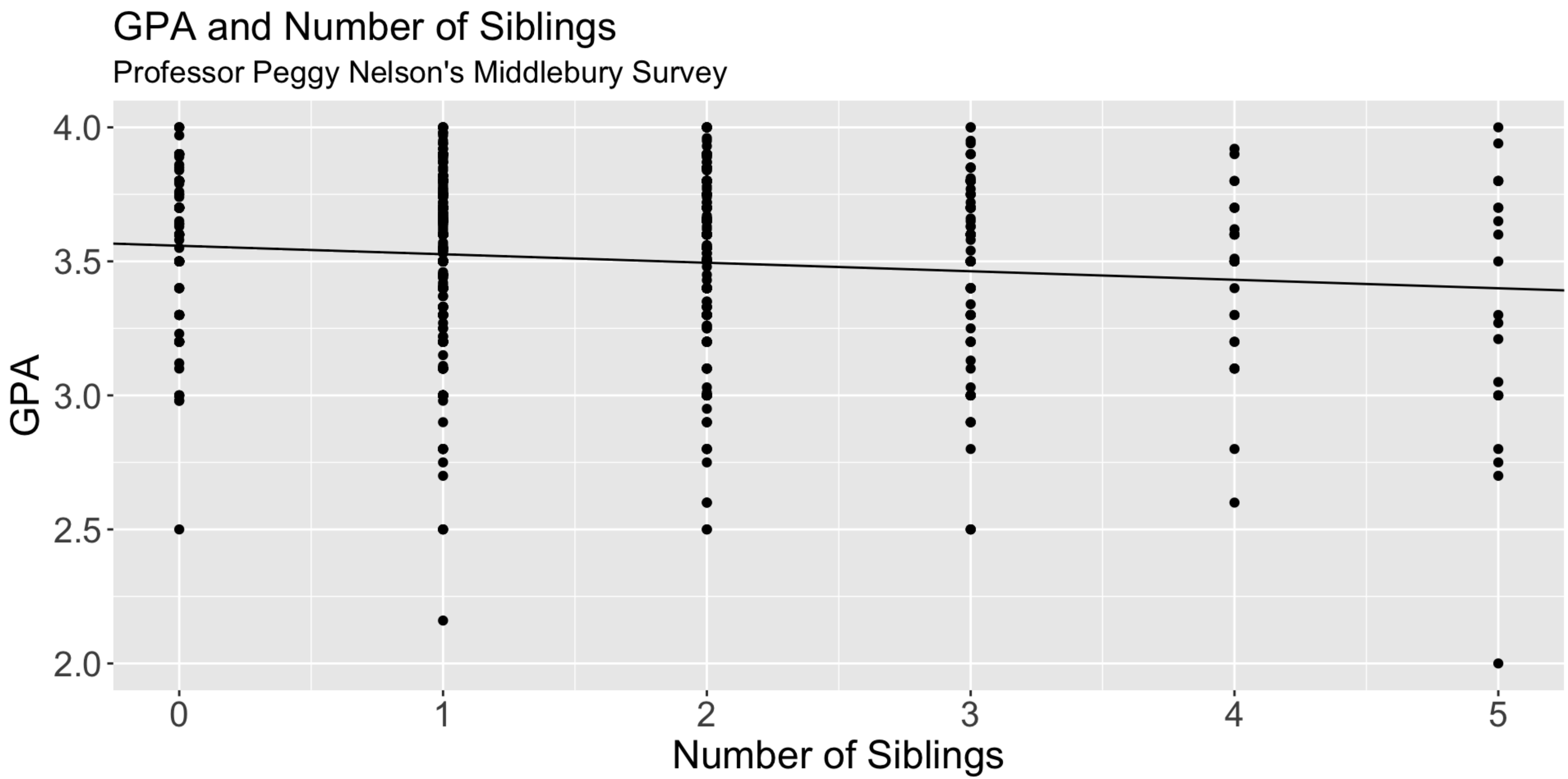
```
Call:
lm(formula = gpa ~ siblings, data = midd_survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.39967 -0.15695  0.04305  0.20487  0.60033 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.558767   0.017163 207.348  < 2e-16 ***
siblings    -0.031819   0.009198  -3.459 0.000564 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 983 degrees of freedom
Multiple R-squared:  0.01203,	Adjusted R-squared:  0.01102
```

Warm Up - Original Model



Warm Up 1

Regress gpa on number of siblings, controlling for gender

```r
gpa_sibs_gender_model <- lm(gpa ~ siblings + gender,
      data = midd_survey)

summary(gpa_sibs_gender_model)
```

Warm Up 1

```
Call:
lm(formula = gpa ~ siblings + gender, data = midd_survey)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3859 -0.1459  0.0541  0.2141  0.6248 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.531540   0.021186 166.689  < 2e-16 ***
siblings    -0.031269   0.009187  -3.404 0.000691 ***
genderOther -0.058342   0.084524  -0.690 0.490210    
genderWoman  0.045629   0.019653   2.322 0.020455 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Warm Up 1

Predict gpa for men with 3 siblings and women with 4 siblings

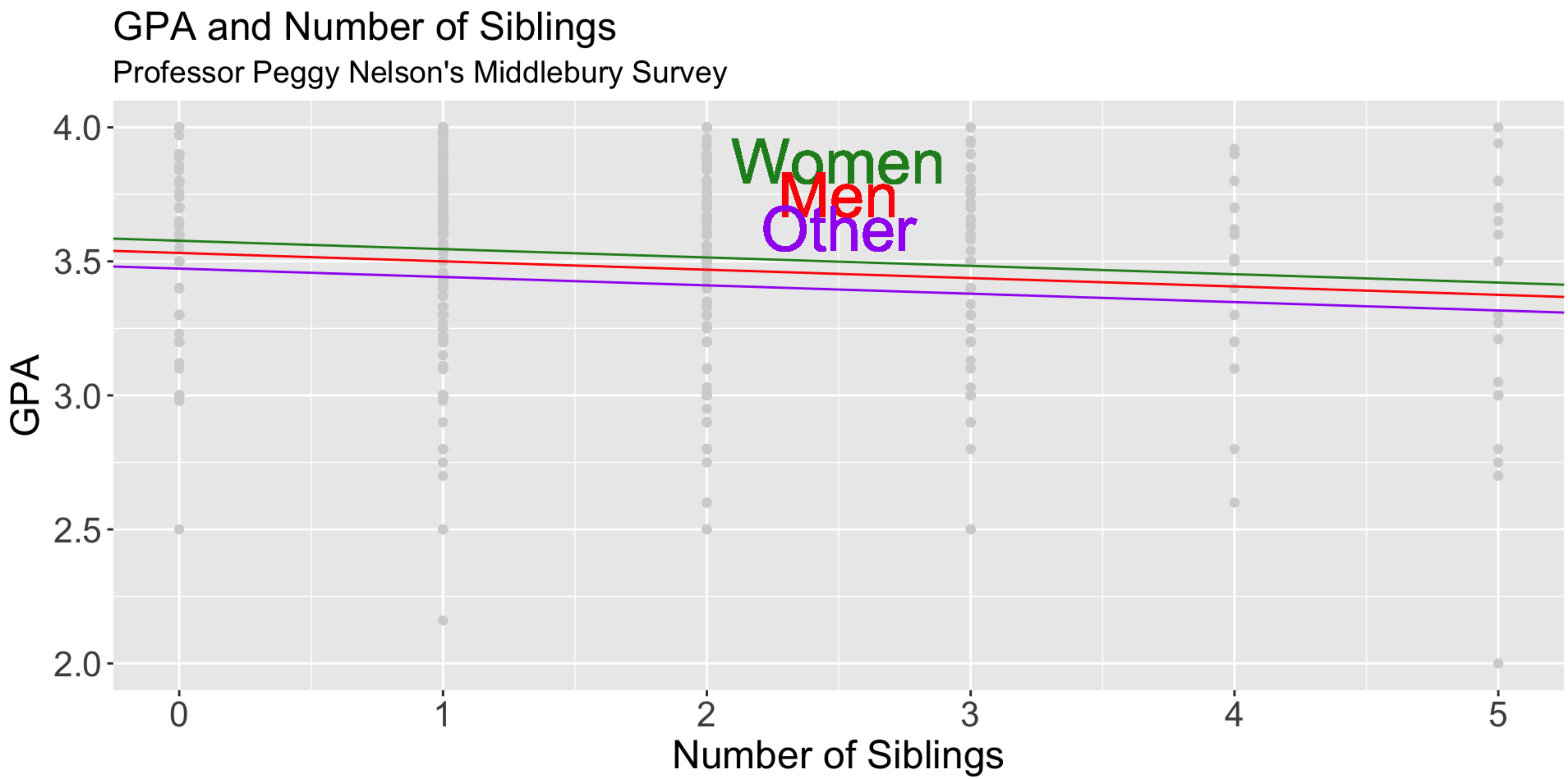
```
# For Men With 3 Siblings:
3.531540 - .031269*3
```

```
[1] 3.437733
```

```
# For Women With 4 Siblings:
3.531540 - .031269*4 + .045629
```

```
[1] 3.452093
```

Warm Up 1


GPA and Number of Siblings
Professor Peggy Nelson's Middlebury Survey
Women
Men
Other
4.0
3.5
3.0
2.5
2.0
GPA
0
1
2
3
4
5
Number of Siblings

Warm Up 2

Regress gpa on number of siblings, controlling for class.

```
gpa_sibs_class_model <- lm(gpa ~ siblings + class,
  data = midd_survey)

summary(gpa_sibs_class_model)
```

Warm Up 2

```
Call:
lm(formula = gpa ~ siblings + class, data = midd_survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.27129 -0.15165  0.02979  0.22156  0.62156 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              3.405229   0.029775 114.365  < 2e-16 ***
siblings                -0.026787   0.009059  -2.957  0.00318 ** 
classMiddle Class        0.130127   0.031589   4.119 4.12e-05 ***
classUpper Class         0.158991   0.036167   4.396 1.22e-05 ***
classUpper Middle Class  0.191764   0.028396   6.753 2.48e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Warm Up 2

Predict gpa for middle class student with 2 siblings and upper middle class student with 1 sibling

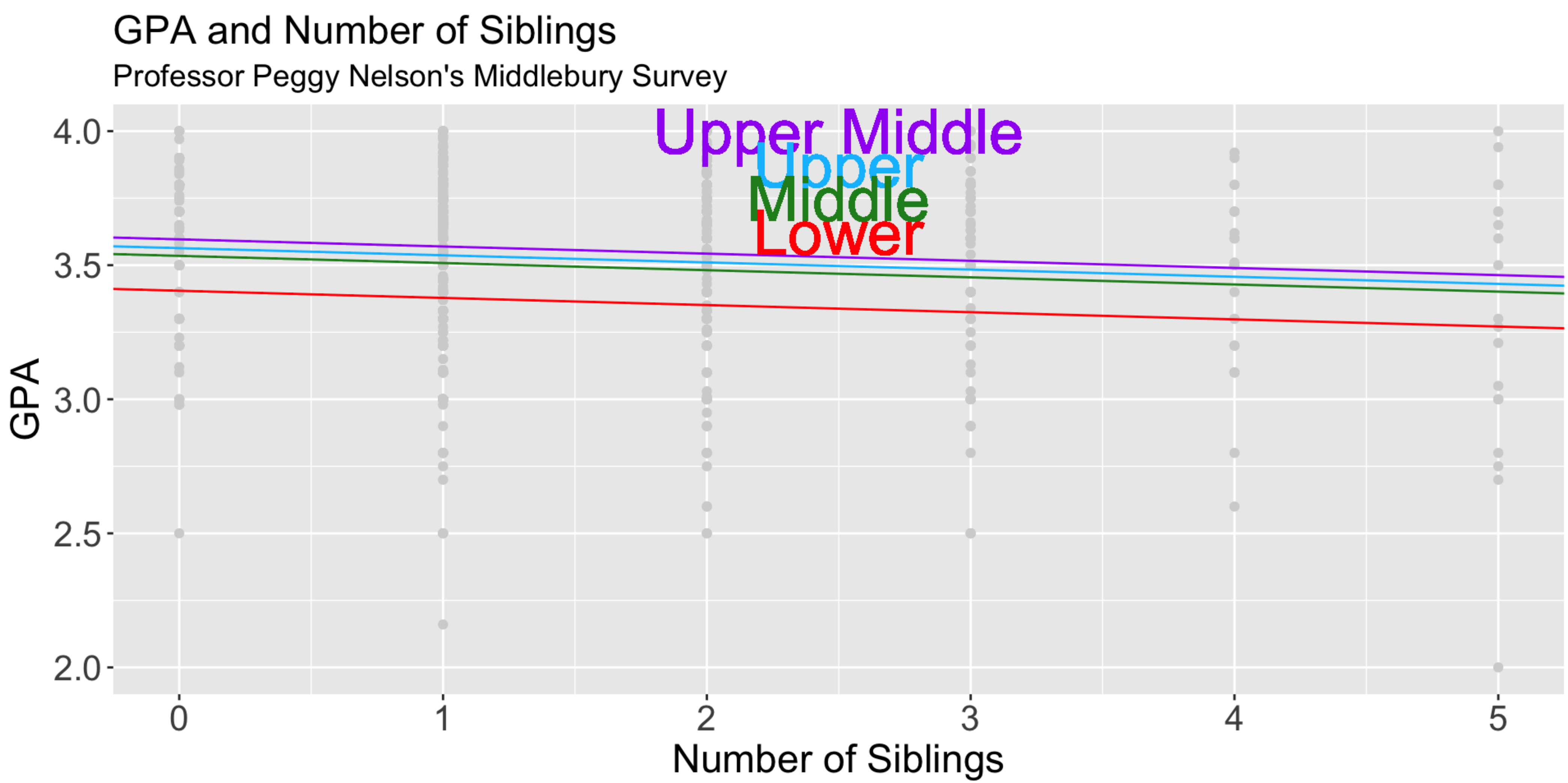
```
# For middle class student with 2 siblings:
3.405229 - .026787*2 + .130127
```

```
[1] 3.481782
```

```
# For upper middle class student with 1 sibling:
3.405229 -  .026787*1 + .191764
```

```
[1] 3.570206
```

Warm Up 2



GPA and Number of Siblings
Professor Peggy Nelson's Middlebury Survey
Upper Middle
Upper
Middle
Lower
4.0
3.5
3.0
2.5
2.0
GPA
0
1
2
3
4
5
Number of Siblings

Warm Up 3

Group 3: Add controls for gender and class to original model

```
gpa_sibs_class_gender_model <-
lm(gpa ~ siblings + class + gender, data = midd_survey)

summary(gpa_sibs_class_gender_model)
```

Warm Up 3

```
Call:
lm(formula = gpa ~ siblings + class + gender, data = midd_survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.28880 -0.14141  0.03547  0.20969  0.65715 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              3.368622   0.033079 101.837  < 2e-16 ***
siblings                -0.025774   0.009051  -2.848  0.00450 ** 
classMiddle Class        0.132890   0.031883   4.168 3.34e-05 ***
classUpper Class         0.166871   0.036480   4.574 5.39e-06 ***
classUpper Middle Class  0.195771   0.028842   6.788 1.97e-11 ***
genderOther              0.049050   0.084249   0.582  0.56056    
genderWoman              0.051690   0.019277   2.681  0.00745 ** 
```

Warm Up 3

Predict gpa for lower class men with 0 siblings

```
3.368622
```

```
[1] 3.368622
```

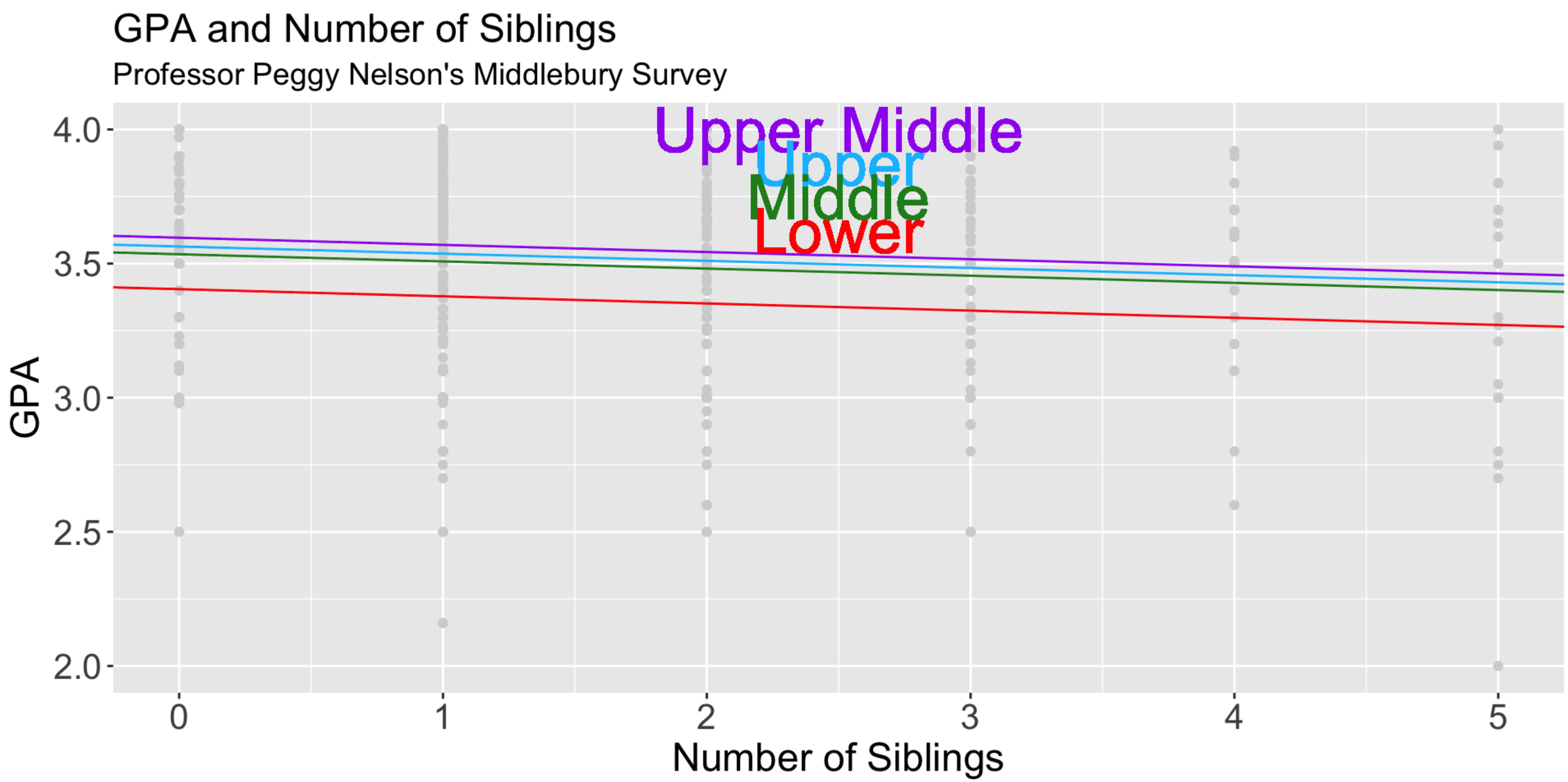
Where We Are Now

- We know lines can have different starting points. That means there can be different alphas, or intercepts.
- We know the predicted values on the lines can be different from the observed values. Those differences are the residuals.

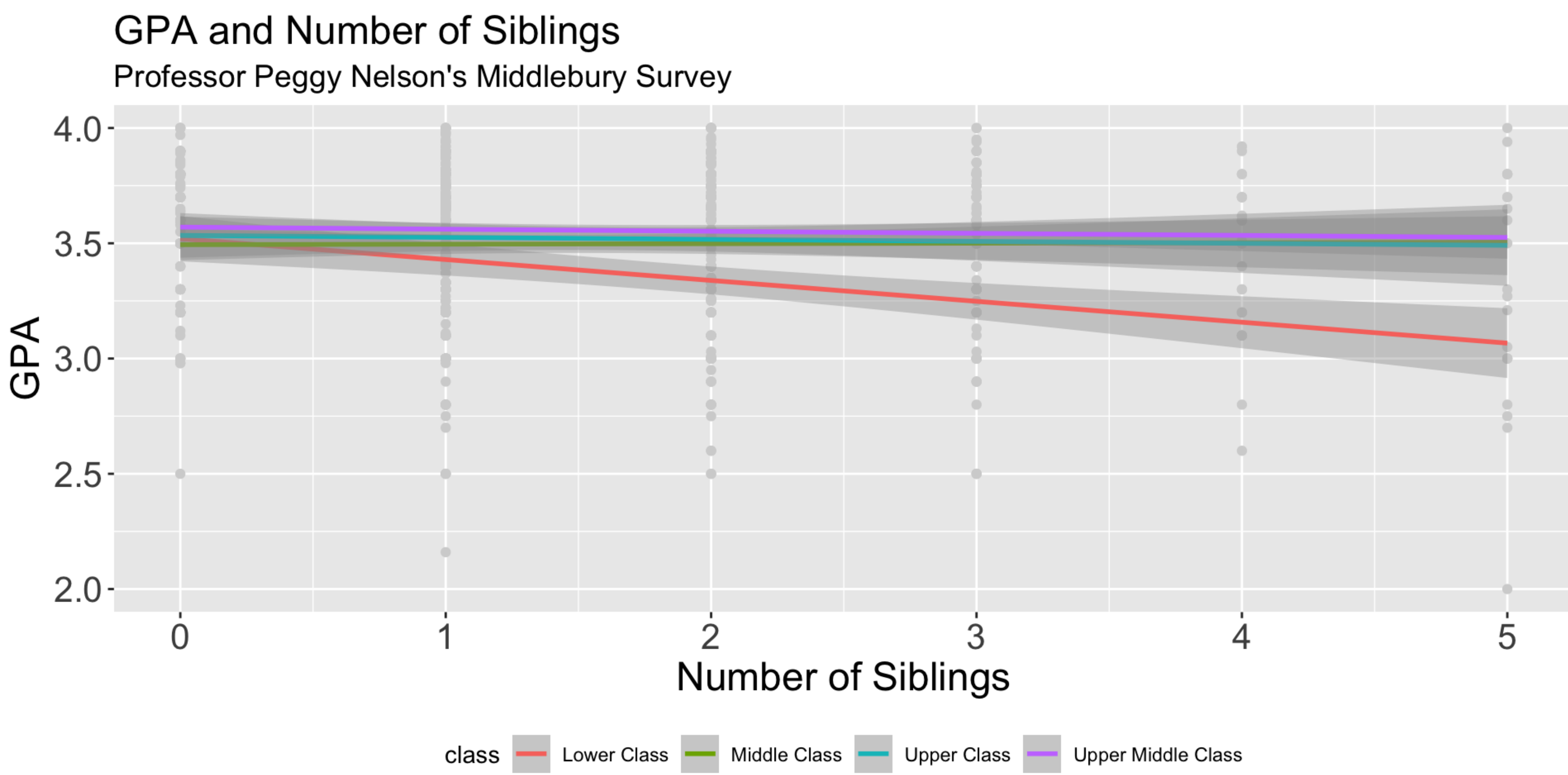
Where We Are Now

- One reason we might have big residuals is because we assume that the change for each increase in our X variable is the same for all values of our control variable(s).
- If that is not true, we need a way to let the slopes of our lines vary too.
- More formally, we want to know if the average change in Y for a change in X changes as the value of our control variable changes

We Are About To Move From This…



To This…


GPA and Number of Siblings
Professor Peggy Nelson's Middlebury Survey
GPA
4.0
3.5
3.0
2.5
2.0
0
1
2
3
4
5
Number of Siblings
class
Lower Class
Middle Class
Upper Class
Upper Middle Class

Introducing Interactions

- An interaction is the product of two (or more) variables.
- When we wanted to add another control variable, we used a plus sign:

```
gpa_sibs_class_model <- lm(gpa ~ siblings + class,
        data = midd_survey)
```

- When we want to include the product of two variables, we use a star:

```
gpa_sibsXclass_model <-
lm(gpa ~ siblings * class,
        data = midd_survey)

summary(gpa_sibsXclass_model)
```

Introducing Interactions

```
Call:
lm(formula = gpa ~ siblings * class, data = midd_survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.26944 -0.15217  0.03884  0.20683  0.73316 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       3.52009    0.04076  86.351  < 2e-16 ***
siblings                         -0.09065    0.01801  -5.033 5.75e-07 ***
classMiddle Class                -0.02612    0.05244  -0.498 0.618514    
classUpper Class                  0.01418    0.06503   0.218 0.827416    
classUpper Middle Class           0.05007    0.04787   1.046 0.295767    
siblings:classMiddle Class        0.09276    0.02606   3.559 0.000390 ***
siblings:classUpper Class         0.08201    0.03191   2.570 0.010328 *  
```

Introducing Interactions

- This model has *main effects*: Siblings, Middle Class, Upper Class, Upper Middle Class
- And it has *interaction effects*: Siblings X Middle Class, Siblings X Upper Class, and Siblings X Upper Middle Class
- The interaction term tells us how the slope varies for each value of the other variable.
- The slope for our reference group (Lower Class) is the coefficient for siblings: -0.09065

Introducing Interactions

```
Call:
lm(formula = gpa ~ siblings * class, data = midd_survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.26944 -0.15217  0.03884  0.20683  0.73316 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       3.52009    0.04076  86.351  < 2e-16 ***
siblings                         -0.09065    0.01801  -5.033 5.75e-07 ***
classMiddle Class                -0.02612    0.05244  -0.498 0.618514    
classUpper Class                  0.01418    0.06503   0.218 0.827416    
classUpper Middle Class           0.05007    0.04787   1.046 0.295767    
siblings:classMiddle Class        0.09276    0.02606   3.559 0.000390 ***
siblings:classUpper Class         0.08201    0.03191   2.570 0.010328 *  
```

Introducing Interactions

- The slope for our other groups is the coefficient for siblings plus the respective interaction term
- For Middle Class:
  - → -0.09065 + 0.09276 = 0.00211
- For Upper Class:
  - → -0.09065 + 0.08201 = -0.00864
- For Upper Middle Class:
  - → -0.09065 + 0.08165 = -0.009

Interactions and Predictions

- For predictions, use the full equation

```
3.52009 - 0.09065*(siblings) - 0.02612*(middle class) +
0.01418*(upper class) +0.05007*(upper middle class) +
0.09276*(siblings*middle class) +
0.08201*(siblings*upper class) +
0.08165*(siblings*upper middle class)
```

- This still makes the intercept the predicted gpa for a lower class student with zero siblings:

```
[1] 3.52009
```

Interactions and Predictions

- Without interactions, we estimated the predicted gpa for a middle class student with 2 siblings to be 3.481782.
- What is the prediction with interactions?

```
3.52009 - 0.09065*(2) - 0.02612*(1) + 0.01418*(0) +
0.05007*(0) + 0.09276*(2*1) + 0.08201*(2*0) + 0.08165*(2*0)
```

```
[1] 3.49819
```

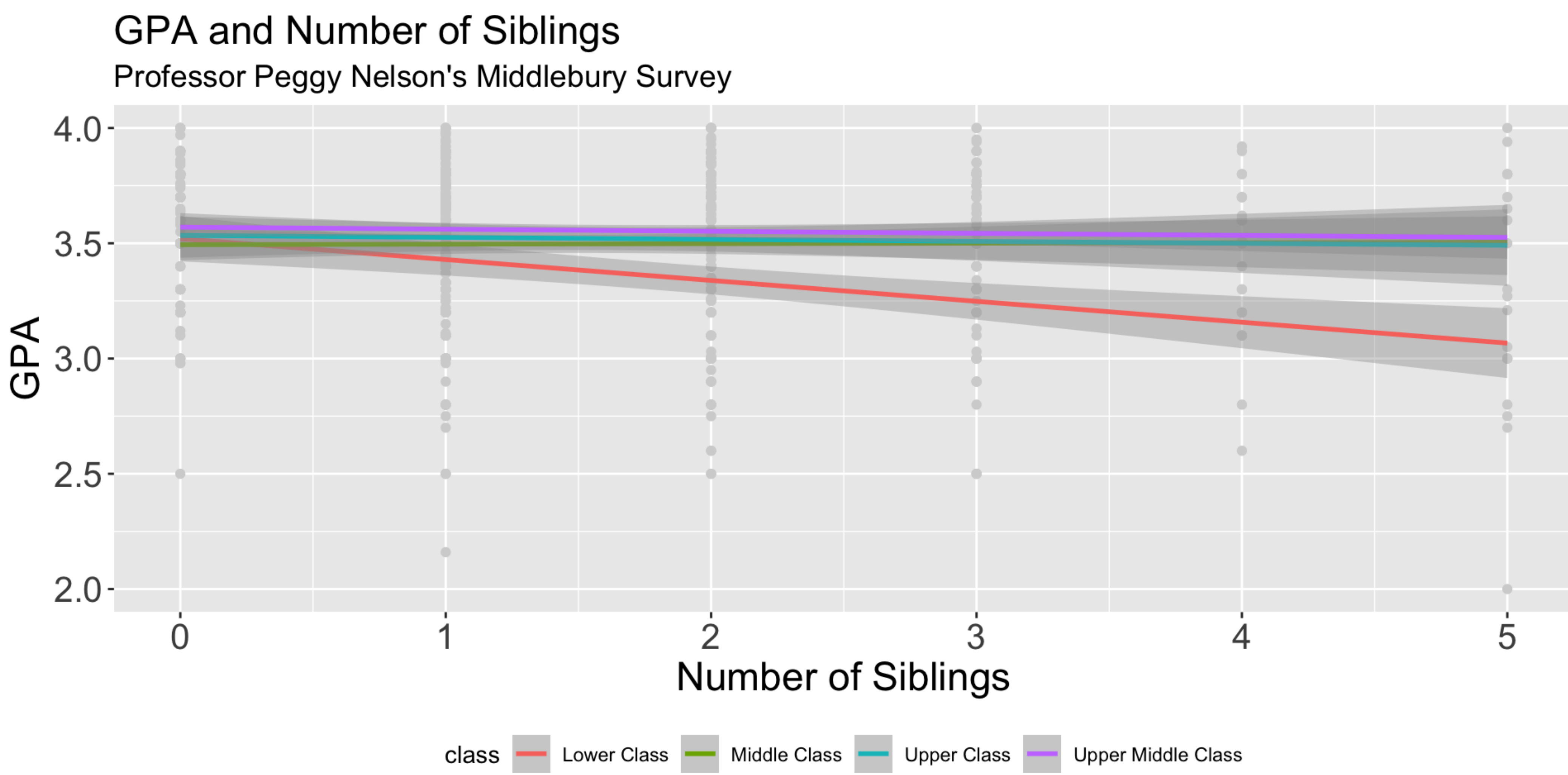
Plotting Interactions

- Add your control variable to the aesthetics map as the color. The regular `geom_smooth(method = lm)` function includes interactions by default.

```
gpa_sibs_class_plot <- ggplot(midd_survey,
      aes(x = siblings, y = gpa, color = class))

gpa_sibs_class_plot + geom_point(color = "Light Gray") +
     geom_smooth(method = lm) +
     labs(x = "Number of Siblings", y = "GPA",
      title = "GPA and Number of Siblings",
      subtitle = "Professor Peggy Nelson's Middlebury Survey") +
     theme(legend.position = "bottom")
```

Plotting Interactions



Understanding Interactions

- Interactions are not always significant! If *none* are significant, do not use that model (usually).
- Have to include the *main effects* when you have an interaction. R does this automatically; other programs do not.
- If the main effects are not significant when you add the interactions but the interactions are significant, that's okay.
- With lots of interactions, can be hard to imagine a plot…much easier to calculate predictions when you have interactions.

Understanding Interactions

- Key takeaway is that a significant interaction effect tells you that the change for a one unit change in X is different for at least one value of another variable. That means the slope varies.

Interactions - Example 2

- Our questions before interactions: On a scale from 1-5, would you expect students to disagree or agree that they are actively looking to start a relationship at Middlebury (`midd_lookingfor_relationship`)?
  - → Will the average responses vary across genders? Would school year explain that variation?
- Our question with interactions: Would you expect any differences across genders to vary by school year?

Example 2 - The Basic Model

Start with the bivariate relationship

```
lookrel_gender_model <-
     lm(midd_lookingfor_relationship ~ gender,
      data = midd_survey)

summary(lookrel_gender_model)
```

Example 2 - Basic Model

```
Call:
lm(formula = midd_lookingfor_relationship ~ gender, data = midd_survey)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9819 -0.9819  0.2201  1.0181  2.3846 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.98187    0.06678  44.650   <2e-16 ***
genderOther -0.36648    0.36998  -0.991    0.322    
genderWoman -0.20200    0.08601  -2.349    0.019 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.312 on 982 degrees of freedom
```

Example 2 - Basic Model

- On average, women's responses tend to be lower than men's responses, meaning women are less likely than men to say they are looking to start a relationship at Middlebury. This difference is significant.
- Students in the other gender category also tend to have lower responses than men, on average. But this difference is not significant.

Example 2 - Control Variable

Control for year

```
lookrel_gender_year_model <-
lm(midd_lookingfor_relationship ~ gender + year,
      data = midd_survey)

summary(lookrel_gender_year_model)
```

Example 2 - Control Variable

```
Call:
lm(formula = midd_lookingfor_relationship ~ gender + year, data = midd_survey)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0879 -1.0279  0.1027  1.1027  2.4178 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    3.02792    0.09840  30.771   <2e-16 ***
genderOther   -0.30582    0.36930  -0.828   0.4078    
genderWoman   -0.19060    0.08585  -2.220   0.0266 *  
yearSophomore  0.06002    0.11928   0.503   0.6149    
yearJunior     0.02784    0.12383   0.225   0.8221    
yearSenior    -0.25514    0.11535  -2.212   0.0272 *  
---
```

Example 2 - Control Variable

- Controlling for school year, average scores for women are still significantly lower than average scores for men.
- Holding gender constant, average scores for seniors are significantly lower than average sores for first year students.

Interactions - Example 2 - Full Model

Add interaction between gender and year

```r
lookrel_genderXyear_model <-
lm(midd_lookingfor_relationship ~ gender * year,
data = midd_survey)

summary(lookrel_genderXyear_model)
```

Interactions - Example 2 - Full Model

```
Call:
lm(formula = midd_lookingfor_relationship ~ gender * year, data = midd_survey)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.02041 -0.97059  0.07333  1.04950  2.66667 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                2.95050    0.12975  22.740   <2e-16 ***
genderOther               -1.28383    0.76393  -1.681   0.0932 .  
genderWoman               -0.02801    0.17325  -0.162   0.8716    
yearSophomore              0.06991    0.18489   0.378   0.7054    
yearJunior                 0.03774    0.19193   0.197   0.8442    
yearSenior                 0.02009    0.18304   0.110   0.9126    
genderOther:yearSophomore  3.26342    1.51696   2.151   0.0317 *  
```

Interpreting - Example 2 - Full Model

- The difference between seniors and first years is .48 points lower for women than it is for men, on average. This difference is significant.
- Among sophomores, average scores for students in the other gender category are 3.26 points higher than the average scores for men. This difference is significant.

Summarizing Interactions

- With lots of variables, interpreting and plotting interactions can get messy.
- Easier to predict values from your full model and describe them.
- Remember the `fitted.values` function can calculate predictions and save them as a new variable:

```
midd_survey$pred_lookrel <-
  lookrel_genderXyear_model$fitted.values
```

Summarizing Interactions

- Then use `group_by()` and `summarize()` to describe the predictions for each combination of the variables you are interacting
- We'll also assert the order of the class years

```
lookrel_predictions <- midd_survey |>
    mutate(year = factor(year,
                         levels = c("First Year",
                                    "Sophomore",
                                    "Junior",
                                    "Senior"))) |>
     group_by(gender, year) |>
     summarize(agree_look_rel = round(mean(pred_lookrel),3))
```

Summarizing Predictions

```r
kbl(lookrel_predictions,
    booktabs = TRUE,
    align = rep("c", 2))
```

gender	year	agree_look_rel
Man	First Year	2.950
Man	Sophomore	3.020
Man	Junior	2.988
Man	Senior	2.971
Other	First Year	1.667
Other	Sophomore	5.000
Other	Junior	2.333
Other	Senior	2.833

gender	year	agree_look_rel
Woman	First Year	2.922
Woman	Sophomore	2.927
Woman	Junior	2.921
Woman	Senior	2.459

Summarizing Predictions

```
lookrel_predictions |>
  pivot_wider(names_from = "year", values_from = "agree_look_rel") |>
  kbl(booktabs = TRUE,
    align = rep("c", 4)) |>
  kable_paper()
```

gender	First Year	Sophomore	Junior	Senior
Man	2.950	3.020	2.988	2.971
Other	1.667	5.000	2.333	2.833
Woman	2.922	2.927	2.921	2.459