Introduction to Interaction Models

2017-03-10

Preparation

In this set of notes, you will learn about interaction models. To do so, we will examine the question of whether there is a differential effect of beauty by gender on course evaluation scores. The data we will use in this set of notes is collected from student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations. The variables are:

- `prof`: Professor ID number
- `avgeval`: Average course rating
- `btystdave`: Measure of the professor's beauty composed of the average score on six standardized beauty ratings
- `tenured`: 0 = non-tenured; 1 = tenured
- `nonenglish`: 0 = native English speaker; 1 = non-native English speaker
- `age`: Professor's age (in years)
- `female`: 0 = male; 1 = female
- `students`: Number of students enrolled in the course
- `percentevaluating`: Percentage of enrolled students who completed an evaluation

These source of these data is: Hamermesh, D. S. & Parker, A. M. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369–376. The data were made available by: Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

```
# Read in data
beauty = read.csv(file = "~/Google Drive/Documents/epsy-8251/data/beauty.csv")
head(beauty)
```

```
  prof avgeval  btystdave tenured nonenglish age female students
1    1     4.3  0.2015666       0          0  36      1       43
2    2     4.5 -0.8260813       1          0  59      0       20
3    3     3.7 -0.6603327       1          0  51      0       55
4    4     4.3 -0.7663125       1          0  40      1       46
5    5     4.4  1.4214450       0          0  31      1       48
6    6     4.2  0.5002196       1          0  62      0      282
  percentevaluating
1          55.81395
2          85.00000
3         100.00000
4          86.95652
5          87.50000
6          64.53901
```

```
# Load libraries
library(dplyr)
library(ggplot2)
library(sm)
```

Main-Effects Models

We will explore the effects of beauty and gender on course evaluation scores. You might fit the regression model that includes both predictors.

```
lm.1 = lm(avgeval ~ 1 + btystdave + female, data = beauty)
summary(lm.1)
```

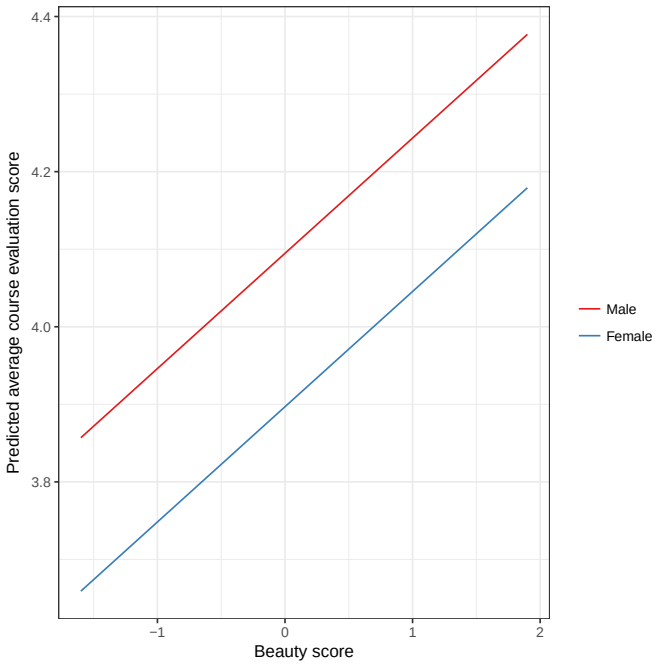
```
Call:
lm(formula = avgeval ~ 1 + btystdave + female, data = beauty)

Residuals:
     Min       1Q   Median       3Q      Max
-1.87196 -0.36913  0.03493  0.39919  1.03237

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  4.09471    0.03328  123.03    < 2e-16 ***
btystdave    0.14859    0.03195    4.65 0.00000434 ***
female      -0.19781    0.05098   -3.88    0.00012 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5373 on 460 degrees of freedom
Multiple R-squared:  0.0663,    Adjusted R-squared:  0.06224
F-statistic: 16.33 on 2 and 460 DF,  p-value: 0.0000001407
```

Here there is an effect of gender ($p = .001$) controlling for differences in beauty, and there is also an effect of beauty after controlling for differences in gender ($p < .001$). Interpreting these effects:

- Compared to professors who are rated as less beautiful, professors rated as more beautiful tend to have higher course evaluation scores, controlling for differences in gender. Each one-point difference in beauty is associated with a 0.15-point difference in course evaluation score, controlling for differences in gender,
- Female professors have a lower average course evaluation than male professors controlling for differences in beauty. This difference is 0.19-points, on average,

Visually, we can display these effects by showing the fitted regression line for female and male professors that uses beauty to predict course evaluation scores.

```
myData = expand.grid(
  btystdave = seq(from = -1.6, to = 1.9, by = 0.1),
  female = c(0, 1)
  )
myData = myData %>% mutate(yhat = predict(lm.1, newdata = myData))

myData$gender = factor(myData$female, levels = c(0, 1), labels = c("Male", "Female"))

ggplot(data = myData, aes(x = btystdave, y = yhat, color = gender)) +
  geom_line() +
  theme_bw() +
  xlab("Beauty score") +
  ylab("Predicted average course evaluation score") +
  scale_color_brewer(name = "", palette = "Set1")
```



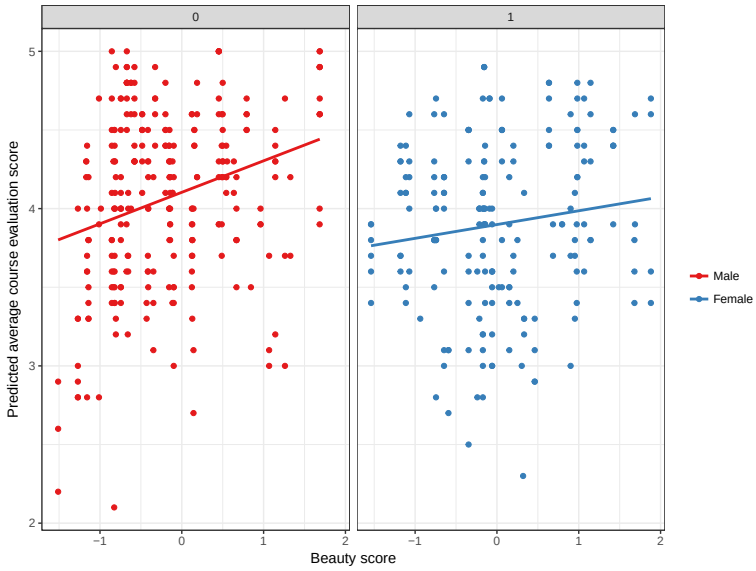
This display helps us see that the effect of beauty (slopes of the lines) is THE SAME for both males and females. We also see that the effect of gender (the vertical distance between the lines) is THE SAME for every level of beauty.

This type of model where the effect of a predictor is THE SAME for eahc level of another predictor is referred to as a main-effects model. All the models we have fitted thus far have been main-effects models.

Differential Effects Models: Interaction Models

Another question a researcher might have is whether the effect of beauty IS DIFFERENT for males and females. Examining the raw data suggests that this might be the case. In the scatterplots below, the samle data suggests that the effect of beauty on average course evaluation scores may be greater for male professors (steeper slope) than for female professors.

```
ggplot(data = beauty, aes(x = btystdave, y = avgeval, color = factor(female))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw() +
  xlab("Beauty score") +
  ylab("Predicted average course evaluation score") +
  scale_color_brewer(name = "", palette = "Set1", labels = c("Male", "Female")) +
  facet_wrap(~female)
```



Differential effects of beauty on course evaluation scores imply that the slopes of the regression lines for males and females are not equal (i.e., the lines are not parallel). This is in stark contrast to the main-effects model which implies parallel regression lines, or equal effects of beauty for both genders. In statistical terms we describe differential effects as *interaction effects*. We would say there is an interaction effect between beauty and gender on course evaluation scores.

Testing for an Interaction Effect

The inferential question is whether the interaction effect that we are seeing in the sample data is real, or whether it is an artifact of samplng error. To examine this we need a way to test whether the slopes of the two regression lines are equal.

To do this, we create another predictor that is the product of the two predictors we believe interact and include that product term in the regression model along with the original predictors we used to create it (i.e., also include the constituent main-effects). In our example, we multiply the gender predictor by the beauty predictor to create the interaction term. Then we fit a model that includes the original gender predictor, the original beauty predictor, and the newly created interaction term. We then pay attention to the coefficient and p-value for the interaction term.

```
# Create interaction term
beauty = beauty %>% mutate(bty_female = btystdave * female)
head(beauty)
```

```
  prof avgeval  btystdave tenured nonenglish age female students
1    1     4.3  0.2015666       0          0  36      1       43
2    2     4.5 -0.8260813       1          0  59      0       20
3    3     3.7 -0.6603327       1          0  51      0       55
4    4     4.3 -0.7663125       1          0  40      1       46
5    5     4.4  1.4214450       0          0  31      1       48
6    6     4.2  0.5002196       1          0  62      0      282
  percentevaluating bty_female
1          55.81395  0.2015666
2          85.00000  0.0000000
3         100.00000  0.0000000
4          86.95652 -0.7663125
5          87.50000  1.4214450
```

```
6          64.53901  0.0000000
```

```r
# Fit interaction model
lm.2 = lm(avgeval ~ 1 + btystdave + female + bty_female, data = beauty)
summary(lm.2)
```

```
Call:
lm(formula = avgeval ~ 1 + btystdave + female + bty_female, data = beauty)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83820 -0.37387  0.04551  0.39876  1.06764

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  4.10364    0.03359 122.158    < 2e-16 ***
btystdave    0.20027    0.04333   4.622 0.00000495 ***
female      -0.20505    0.05103  -4.018 0.00006851 ***
bty_female  -0.11266    0.06398  -1.761     0.0789 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5361 on 459 degrees of freedom
Multiple R-squared:  0.07256,   Adjusted R-squared:  0.0665
F-statistic: 11.97 on 3 and 459 DF,  p-value: 0.0000001471
```

If we are using a strict cutoff of $\alpha = .05$ to evaluate the predictors, we would fail to reject the null hypothesis that the partial slope for the interaction term is zero (i.e., $H_0 : \beta_{\text{bty\_female}} = 0$). This suggests that the differential effects we saw in the raw data are likely just an artifact of sampling error.

Mathematical Expression of the Interaction Model

In general, the interaction model (with two predictors) can be written as,

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \beta_3\,(X_1X_2) + \epsilon.$$

First notice that if β_3, the coefficient on the interaction term, is zero, this equation reduces to the equation for the main-effects model, namely

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \epsilon.$$

In practice, if we fail to reject the null hypothesis that the coefficient for the interaction term is zero, we would drop the interaction term from the model, and instead adopt the main-effects model.

To understand how testing whether the slope associated with the interaction term is equivalent to testing whether the regression lines are parallel, we will write out the interaction model for our example.

$$Y = \beta_0 + \beta_1(X_{\text{beauty}}) + \beta_2(X_{\text{Female}}) + \beta_3\,(X_{\text{Beauty}}X_{\text{Female}}) + \epsilon.$$

Recall that the predictor X_{Female} is a dummy coded predictor that is 1 for females and 0 for males. We can use that to write individual regression equations, based on the interaction model, for both genders. For example, the regression model for males is,

$$
\begin{aligned}
Y &= \beta_0 + \beta_1(X_{\text{beauty}}) + \beta_2(X_{\text{Female}}) + \beta_3\left(X_{\text{Beauty}}X_{\text{Female}}\right) + \epsilon \\
&= \beta_0 + \beta_1(X_{\text{beauty}}) + \beta_2(0) + \beta_3\left(X_{\text{Beauty}} \times 0\right) + \epsilon \\
&= \beta_0 + \beta_1(X_{\text{beauty}}) + \epsilon.
\end{aligned}
$$

The intercept from the interaction model (β_0) turns out to be the intercept term for the reference group (males). The slope associated with beauty from the interaction model (β_1) turns out to be the beauty effect for the reference group (males).

The regression model for females is,

$$
\begin{aligned}
Y &= \beta_0 + \beta_1(X_{\text{beauty}}) + \beta_2(X_{\text{Female}}) + \beta_3\left(X_{\text{Beauty}}X_{\text{Female}}\right) + \epsilon \\
&= \beta_0 + \beta_1(X_{\text{beauty}}) + \beta_2(1) + \beta_3\left(X_{\text{Beauty}} \times 1\right) + \epsilon \\
&= \beta_0 + \beta_1(X_{\text{beauty}}) + \beta_2 + \beta_3(X_{\text{beauty}}) + \epsilon \\
&= [\beta_0 + \beta_2] + \beta_1(X_{\text{beauty}}) + \beta_3(X_{\text{beauty}}) + \epsilon \\
&= [\beta_0 + \beta_2] + [\beta_1 + \beta_3]\left(X_{\text{beauty}}\right) + \epsilon
\end{aligned}
$$

Now we can see that the other two terms used in the interaction model, β_2 and β_3, describe the differences in intercept and slope, respectively, between the non-reference group (females) and the reference group (males).

Consider if the interaction slope (β_3) were zero. Then the beauty effect for females which is $[\beta_1 + \beta_3]$ would be $[\beta_1 + 0] = \beta_1$. This would imply that the beauty effect for males and females would be exactly the same (i.e., they would have the same slope).

Interpreting the Fitted Model's Coefficients

Here we will use the interaction model we fitted earlier to understand how to interpret the different coefficientsin the model. This is purely for pedagogical purposes. In practice, since we failed to reject the null hypothesis that the interaction effect was zero, we would drop the interaction term and interpret the main-effects model's coefficients.

Based on the fitted interaction model, we can write the equation for the fitted model as,

$$
\widehat{\text{Avg. Course Eval}} = 4.1 + 0.20(\text{Beauty}) - 0.21(\text{Female}) - 0.11(\text{Beauty})(\text{Female}).
$$

The easiest way to determine how to interpret the coefficients is to actually compute the regression equations for males and females from the fitted interaction model.

Males:

$$
\begin{aligned}
\widehat{\text{Avg. Course Eval}} &= 4.1 + 0.20(\text{Beauty}) - 0.21(0) - 0.11(\text{Beauty})(0) \\
&= 4.1 + 0.20(\text{Beauty})
\end{aligned}
$$

The intercept from the interaction model ($\hat{\beta}_0 = 4.1$) is the estimated average course evaluation score for male professors who have an average beauty rating of zero. The beauty effect from the interaction model ($\hat{\beta}_1 = 0.20$) suggests that for male professors, a one-unit difference in beauty rating is generally associated with a 0.20-point difference in average course evaluation scores.

Females:

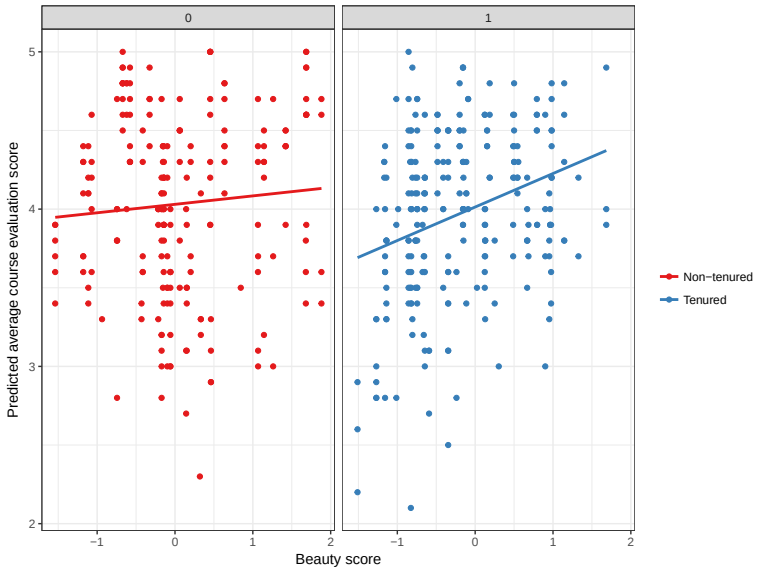
$$\begin{aligned}\widehat{\text{Avg. Course Eval}} &= 4.1 + 0.20(\text{Beauty}) - 0.21(1) - 0.11(\text{Beauty})(1) \\ &= 4.1 + 0.20(\text{Beauty}) - 0.21 - 0.11(\text{Beauty}) \\ &= [4.1 - 0.21] + [0.20 - 0.11]\,(\text{Beauty})\end{aligned}$$

The female effect from the interaction model ($\hat{\beta}_2 = -0.21$) indicates that female professors with a beauty rating of zero have average course evaluation scores that are 0.21-points lower than male professors with beauty ratings of zero, on average. The interaction effect ($\hat{\beta}_3 = -0.11$) indicates that for female professors, a one-unit difference in beauty rating is generally associated with a 0.11-point lower difference in average course evaluation scores than male professors for the same change in beauty rating. Put differently, a one-unit difference in beauty rating for male professors is associated with a 0.20-point difference in average course evaluation scores; but for female professors, a one-unit difference in beauty rating is only associated with a 0.09-point difference in average course evaluation scores.

Interaction Effect of Tenure and Beauty

Let's examine whether there is a differential effect of beauty on course evaluation scores for professors with tenure status compared to those with non-tenured status.

```
ggplot(data = beauty, aes(x = btystdave, y = avgeval, color = factor(tenured))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw() +
  xlab("Beauty score") +
  ylab("Predicted average course evaluation score") +
  scale_color_brewer(name = "", palette = "Set1", labels = c("Non-tenured", "Tenured")) +
  facet_wrap(~tenured)
```



Judging by the sample data, it appears that there is a larger effect of beauty on course evaluation scores for tenured professors than for non-tenured professors. To determine whether this difference in effects is just due to sampling error, we will fit the interaction model and evaluate the interaction term.

```r
# Create interaction predictor
beauty = beauty %>% mutate(bty_tenured = btystdave * tenured)
head(beauty)
```

```
  prof avgeval  btystdave tenured nonenglish age female students
1    1     4.3  0.2015666       0          0  36      1       43
2    2     4.5 -0.8260813       1          0  59      0       20
3    3     3.7 -0.6603327       1          0  51      0       55
4    4     4.3 -0.7663125       1          0  40      1       46
5    5     4.4  1.4214450       0          0  31      1       48
6    6     4.2  0.5002196       1          0  62      0      282
  percentevaluating bty_female bty_tenured
1          55.81395  0.2015666   0.0000000
2          85.00000  0.0000000  -0.8260813
3         100.00000  0.0000000  -0.6603327
4          86.95652 -0.7663125  -0.7663125
5          87.50000  1.4214450   0.0000000
6          64.53901  0.0000000   0.5002196
```

```r
# Fit interaction model
lm.3 = lm(avgeval ~ 1 + btystdave + tenured + bty_tenured, data = beauty)
summary(lm.3)
```

```
Call:
lm(formula = avgeval ~ 1 + btystdave + tenured + bty_tenured,
    data = beauty)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74756 -0.36467  0.04294  0.40445  1.16889

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.03039    0.03754 107.350   <2e-16 ***
btystdave    0.05369    0.04495   1.195   0.2329
tenured     -0.01709    0.05166  -0.331   0.7410
bty_tenured  0.15934    0.06501   2.451   0.0146 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5429 on 459 degrees of freedom
Multiple R-squared:  0.0489,    Adjusted R-squared:  0.04268
F-statistic: 7.866 on 3 and 459 DF,  p-value: 0.00003971
```

The p-value associated with the interaction term ($p = .015$) is statistically significant. This indicates the differential beauty effect we saw in the raw data is larger than we would expect because of sampling error. It is important to note that even though the main-effects of tenure and beauty are not statistically significant, we need to keep them in the model. This is because the interaction effect was composed of these two effects. So, to appropriately interpret the interaction effect we need to keep all constitutent main-effects in the model.

To aid interpretation of the model, we will write the fitted regression equations for tenured and non-tenured professors based on the interaction model.

$$\textbf{Non} - \textbf{tenured}: \quad \widehat{\text{Avg. Course Eval}} = 4.03 + 0.05(\text{Beauty})$$
$$\textbf{Tenured}: \quad \widehat{\text{Avg. Course Eval}} = 4.01 + 0.21(\text{Beauty})$$

The average course evaluation scores for non-tenured professors (the reference group) with a beauty rating of zero is predicted to be 4.03. For tenured professors with the same beauty rating of zero, the average course evaluation score is 0.02-points lower ($p = .741$). For non-tenured professors, there is likely no effect of beauty on average course evaluation scores ($\hat{\beta} = 0.05$, $p = .233$). For tenured professors, however, there is a slight effect of beauty on course evaluation scores. Each one-unit difference in beauty rating, is associated with a 0.21-point difference in average course evaluation scores for tenured professors. The difference in the effect of beauty between tenured and non-tenured professors is statistically significant ($p = .015$).

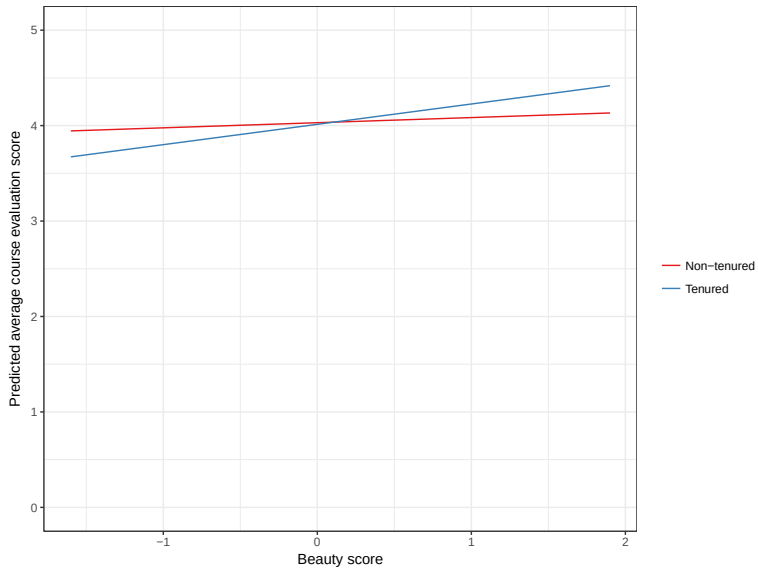
Visually Displaying the Model

It is useful to visually display the fitted interaction model to aid model interpretation. To do this we need to create a dataset that includes the predictors `btystdave`, `tenured`, and `bty_tenured`. The last predictor, recall, was the product of the two main effects. So, when creating our data set, we use `expand.grid()` to create the data for the main effects and then we add the product term afterward.

```
# Create new data set with main effects
myData = expand.grid(
  btystdave = seq(from = -1.6, to = 1.9, by = 0.1),
  tenured = c(0, 1)
)

# Compute interaction effect
myData = myData %>% mutate( bty_tenured = btystdave * tenured )

# Note this can be done in one computation as well
myData = expand.grid(
    btystdave = seq(from = -1.6, to = 1.9, by = 0.1),
    tenured = c(0, 1)
  ) %>%
  mutate( bty_tenured = btystdave * tenured )

# Use fitted model to compute fitted values for the data
myData = myData %>% mutate( yhat = predict(lm.3, newdata = myData) )
head(myData)
```

```
  btystdave tenured bty_tenured     yhat
1      -1.6       0           0 3.944484
2      -1.5       0           0 3.949853
3      -1.4       0           0 3.955223
4      -1.3       0           0 3.960592
5      -1.2       0           0 3.965961
6      -1.1       0           0 3.971330
```

```r
# Plot the fitted model
ggplot(data = myData, aes(x = btystdave, y = yhat, color = factor(tenured))) +
  geom_line() +
  theme_bw() +
  xlab("Beauty score") +
  ylab("Predicted average course evaluation score") +
  scale_color_brewer(name = "", palette = "Set1", labels = c("Non-tenured", "Tenured")) +
  ylim(0, 5)
```



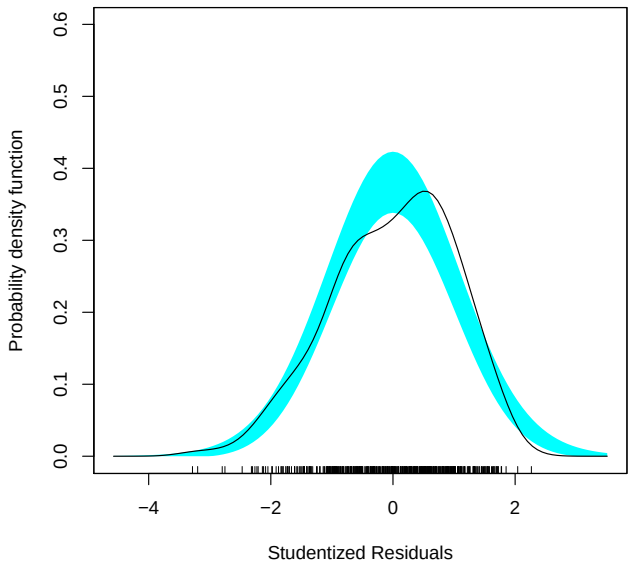
Model Assumptions

Just like main-effects models, we need to examine the assumptions for any fitted interaction model. We do this in the exact same way we did for main effects models.

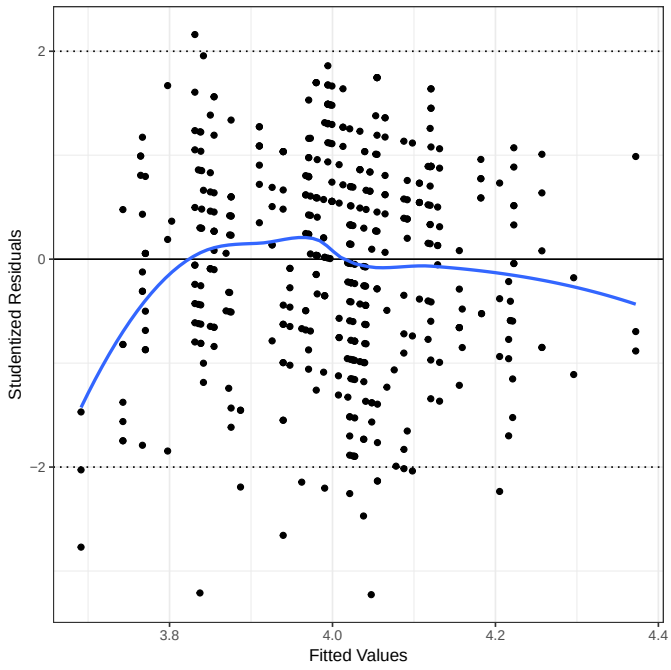
```r
# Create fortified data
fort_lm3 = fortify(lm.3)
head(fort_lm3)
```

```
  avgeval  btystdave tenured bty_tenured        .hat    .sigma
1     4.3  0.2015666       0   0.0000000 0.004909552 0.5433530
2     4.5 -0.8260813       1  -0.8260813 0.006820278 0.5425994
3     3.7 -0.6603327       1  -0.6603327 0.005489975 0.5434281
4     4.3 -0.7663125       1  -0.7663125 0.006293151 0.5430789
5     4.4  1.4214450       0   0.0000000 0.017563551 0.5433123
6     4.2  0.5002196       1   0.5002196 0.007698717 0.5434752
       .cooksd  .fitted      .resid  .stdresid
1 0.0002816478 4.041214  0.25878633  0.4778521
2 0.0025754687 3.837324  0.66267640  1.2248165
3 0.0001403153 3.872633 -0.17263271 -0.3188614
4 0.0010943981 3.850056  0.44994397  0.8314052
5 0.0013277035 4.106711  0.29328875  0.5450378
6 0.0000425897 4.119863  0.08013703  0.1481818
```

```r
# Examine normality assumption
sm.density(fort_lm3$.stdresid, model = "normal", xlab = "Studentized Residuals")
```




```r
# Examine other assumptions
ggplot(data = fort_lm3, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_hline(yintercept = c(-2, 2), linetype = "dotted") +
  geom_smooth(se = FALSE) +
  theme_bw() +
  xlab("Fitted Values") +
  ylab("Studentized Residuals")
```



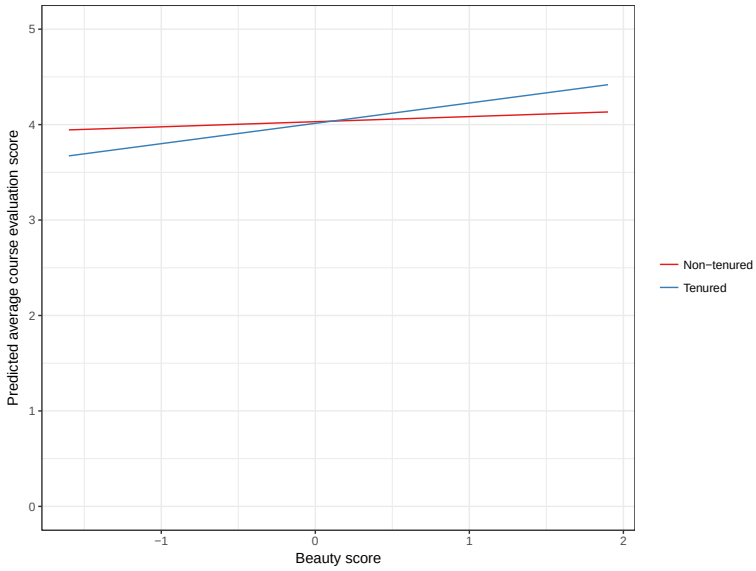
Based on the density plot of the studentized residuals, there is some question about whether the normality assumption is satisfied. The scatterplot of the model's studentized residuals versus its fitted values suggests that the assumption of homoskedasticity is reasonablly satisfied. The loess line (indicating the mean pattern of the conditional residuals) suggests that the average residual is close to zero for each fitted value. The exceptions seem to be at the extreme fitted values where there are too few residuals to suggest a linearity problem.

Two Interpretations of an Interaction Effect

There are always two interpretations of an interaction effect.

1. The effect of X_1 on Y differs depending on the level of X_2.
2. The effect of X_2 on Y differs depending on the level of X_1.

For example, in our tenure and beauty example, we interpreted the interaction as the effect of beauty on course evaluation scores is different for tenured and non-tenured faculty. In the visual display, this interpretation focuses on the difference in slopes.



We could also interpret the interaction as: the effect of tenure on course evaluation scores is different depending on professor's beauty rating. In the visual display, this interpretation focuses on the vertical distance between the lines.

Which one you use is up to you. Try them both. Although they both describe the same interaction, trying the different interpretations can sometimes lead to more information about or better ways of describing the effect.