Introduction to Interaction Models

2018-07-30

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_id`: Professor ID number
- `avg_eval`: Average course rating
- `num_courses`: Number of courses for which the professor has evaluations
- `num_students`: Number of students enrolled in the professor's courses
- `perc_evaluating`: Average percentage of enrolled students who completed an evaluation
- `beauty`: Measure of the professor's beauty composed of the average score on six standardized beauty ratings
- `tenured`: Is the professor tenured? (0 = non-tenured; 1 = tenured)
- `native_english`: Is the professor a native English speaker? (0 = non-native English speaker; 1 = native English speaker)
- `age`: Professor's age (in years)
- `female`: Is the professor female? (0 = male; 1 = female)

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.

```
# Load libraries
library(broom)
library(corrr)
library(dotwhisker)
library(dplyr)
library(ggplot2)
library(readr)
library(sm)
library(tidyr)

# Read in data
evals = read_csv(file = "~/Documents/github/epsy-8251/data/evaluations.csv")
head(evals)
```

```
# A tibble: 6 x 10
  prof_id avg_eval num_courses num_students perc_evaluating beauty tenured
    <int>    <dbl>       <int>        <int>           <dbl>  <dbl>   <int>
1       1     4              4          416            62.0  0.202       0
2       2     3.53           3          104            87.0 -0.826       1
3       3     3.45           2          250            78.5 -0.660       1
```

```
4       4     4.01           8          223            84.3 -0.766       1
5       5     4.35           6          331            81.8  1.42        0
6       6     4.44           7         1849            59.8  0.500       1
# ... with 3 more variables: native_english <int>, age <int>, female <int>
```

Main-Effects Models

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

```
lm.1 = lm(avg_eval ~ 1 + beauty + tenured, data = evals)

glance(lm.1)
```

```
   r.squared adj.r.squared     sigma statistic   p.value df    logLik
1 0.03869989    0.01757241 0.4570477  1.831733 0.1659908  3 -58.25682
       AIC      BIC deviance df.residual
1 124.5136 134.6868 19.00923          91
```

```
tidy(lm.1)
```

```
         term   estimate  std.error  statistic      p.value
1 (Intercept) 3.89901335 0.07382955 52.8110147 4.592495e-70
2      beauty 0.11061124 0.05813069  1.9028028 6.022802e-02
3     tenured 0.01183151 0.09711096  0.1218349 9.032984e-01
```

Here there is a statistically significant effect of gender ($p = .060$) controlling for differences in beauty. There is no effect of beauty after controlling for differences in gender ($p = .903$). Interpreting the magnitude of the 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 0.11-point difference in course evaluation score, controlling for differences in gender.
- Tenured professors have a slightly higher average course evaluations than non-tenured professors controlling for differences in beauty. This difference is 0.01-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.

```
profs = crossing(
  beauty = seq(from = -1.6, to = 1.9, by = 0.1),
  tenured = c(0, 1)
  )

profs %>%
  mutate(
    # Get y-hat values
    yhat = predict(lm.1, newdata = profs),
    # Make tenured a factor for better plotting
```

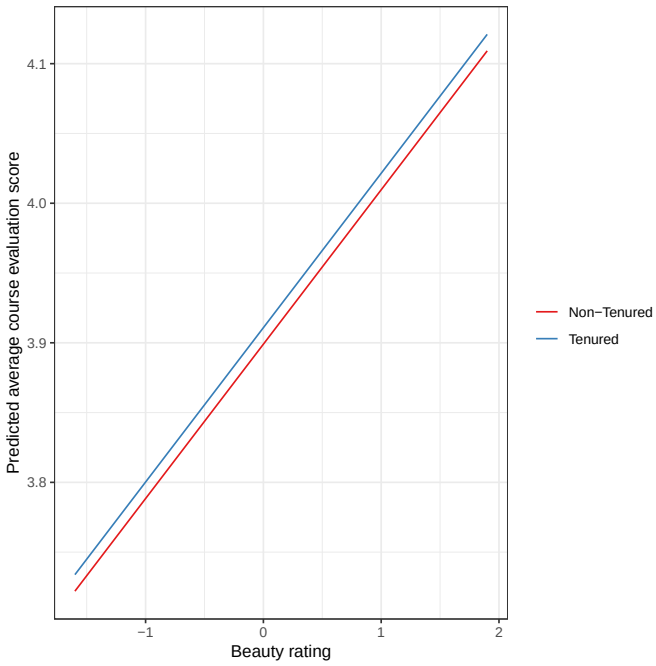


Figure 1: Fitted regression lines showing the predicted course evaluation ratings as a function of professor's beauty rating for tenured and non-tenured professors.

```
    tenured = factor(tenured, levels = c(0, 1), labels = c("Non-Tenured", "Tenured"))
    ) %>%
  ggplot(aes(x = beauty, y = yhat, color = tenured)) +
    geom_line() +
    theme_bw() +
    xlab("Beauty rating") +
    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 tenure (the vertical distance between the lines) is THE SAME for every level of beauty, and is essentially 0.

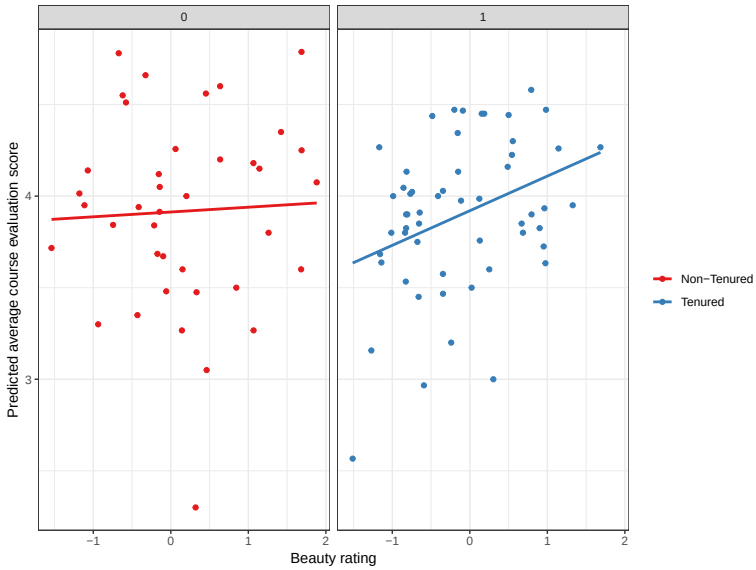
This type of model where the effect of a predictor is THE SAME for eahc level of another predictor is referred to as a **main-effects model**. All the models we have fitted thus far have been main-effects models.

Differential Effects Models: Interaction Models

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

```
ggplot(data = evals, aes(x = beauty, y = avg_eval, color = factor(tenured))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
```

```
  theme_bw() +
  xlab("Beauty rating") +
  ylab("Predicted average course evaluation score") +
  scale_color_brewer(name = "", palette = "Set1", labels = c("Non-Tenured", "Tenured")) +
  facet_wrap(~tenured)
```



Differential effects of beauty on course evaluation scores imply that the slopes of the regression lines for tenured and non-tenured professors 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 tenure status 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 sampling 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 tenure predictor by the beauty predictor to create the interaction term. Then we fit a model that includes the original tenure 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
evals = evals %>%
  mutate(
    bty_tenured = beauty * tenured
    )

head(evals)
```

```
# A tibble: 6 x 11
```

```
  prof_id avg_eval num_courses num_students perc_evaluating beauty tenured
    <int>    <dbl>       <int>        <int>           <dbl>  <dbl>   <int>
1       1     4              4          416            62.0  0.202       0
2       2     3.53           3          104            87.0 -0.826       1
3       3     3.45           2          250            78.5 -0.660       1
4       4     4.01           8          223            84.3 -0.766       1
5       5     4.35           6          331            81.8  1.42        0
6       6     4.44           7         1849            59.8  0.500       1
# ... with 4 more variables: native_english <int>, age <int>,
#   female <int>, bty_tenured <dbl>
```

```
# Fit interaction model
lm.2 = lm(avg_eval ~ 1 + beauty + tenured + bty_tenured, data = evals)
tidy(lm.2)
```

```
         term    estimate  std.error   statistic      p.value
1 (Intercept) 3.913137445 0.07411831 52.79582739 1.645383e-69
2      beauty 0.026187556 0.08332829  0.31426970 7.540436e-01
3     tenured 0.007007145 0.09665330  0.07249774 9.423667e-01
4 bty_tenured 0.162815860 0.11572006  1.40698042 1.628772e-01
```

Using an alpha-value 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., $H_0 : \beta_{\text{bty\_tenured}} = 0$, $p = .163$). 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.

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

Recall that the predictor X_{Tenured} is a dummy coded predictor that is 1 for tenured professors and 0 for non-tenured professors. We can use that to write individual regression equations, based on the interaction model, for each tenure status. For example, the regression model for non-tenured professors is,

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

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

The regression model for tenured professors is,

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

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 (tenured professors) and the reference group (non-tenured professors).

Consider if the interaction slope (β_3) were zero. Then the beauty effect for tenured professors which is $[\beta_1 + \beta_3]$ would be $[\beta_1 + 0] = \beta_1$. This would imply that the beauty effect for tenured and non-tenured professors 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,

$$\hat{\text{Avg. Course Eval}} = 3.91 + 0.03(\text{Beauty}) + 0.01(\text{Tenured}) + 0.16(\text{Beauty})(\text{Tenured}).$$

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

Non-Tenured Professors:

$$\begin{aligned} \hat{\text{Avg. Course Eval}} &= 3.91 + 0.03(\text{Beauty}) + 0.01(0) + 0.16(\text{Beauty})(0) \\ &= 3.91 + 0.03(\text{Beauty}) \end{aligned}$$

The intercept from the interaction model ($\hat{\beta}_0 = 3.91$) is the estimated average course evaluation score for non-tenured professors who have an average beauty rating of zero. The beauty effect from the interaction model ($\hat{\beta}_1 = 0.03$) suggests that for non-tenured professors, a one-unit difference in beauty rating is generally associated with a 0.03-point difference in average course evaluation scores.

Tenured Professors:

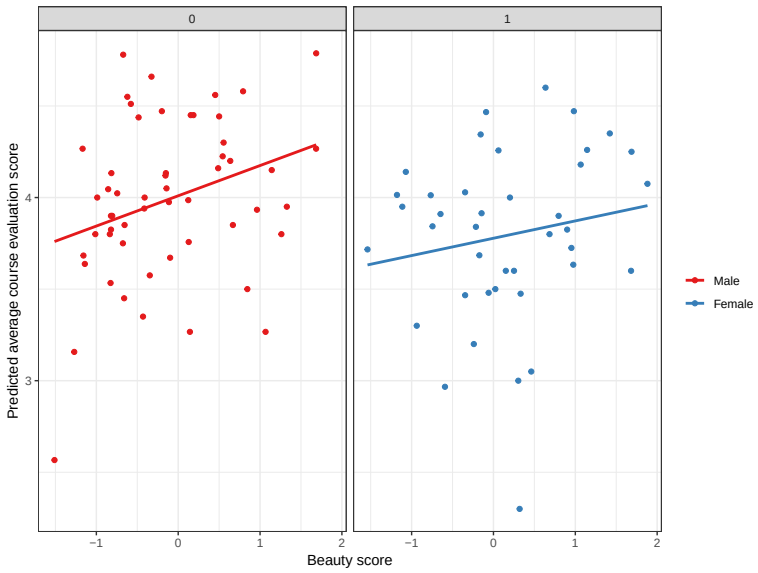
$$\begin{aligned} \hat{\text{Avg. Course Eval}} &= 3.91 + 0.03(\text{Beauty}) + 0.01(1) - 0.16(\text{Beauty})(1) \\ &= 3.91 + 0.03(\text{Beauty}) + 0.01 + 0.16(\text{Beauty}) \\ &= [3.91 + 0.01] + [0.03 + 0.16]\,(\text{Beauty}) \end{aligned}$$

The tenure effect from the interaction model ($\hat{\beta}_2 = 0.01$) indicates that tenured professors with a beauty rating of zero have average course evaluation scores that are 0.01-points higher than non-tenured professors with beauty ratings of zero, on average. The interaction effect ($\hat{\beta}_3 = 0.16$) indicates that for tenured professors, a one-unit difference in beauty rating is generally associated with a 0.16-point higher difference in average course evaluation scores than non-tenured professors for the same change in beauty rating. Put differently, a one-unit difference in beauty rating for non-tenured professors is associated with a 0.03-point difference in average course evaluation scores; but for tenured professors, athe same one-unit difference in beauty rating is associated with a 0.19-point difference in average course evaluation scores.

Interaction Effect of Gender and Beauty

Let's examine whether there is a differential effect of beauty on course evaluation scores for female and male professors.

```
ggplot(data = evals, aes(x = beauty, y = avg_eval, 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)
```



Judging by the sample data, it appears that the effect of beauty on course evaluation scores for female and male professors is roughly the same. To further evaluate this, we will fit the interaction model and evaluate the interaction term.

```
# Create interaction predictor
evals = evals %>%
  mutate(
    bty_female = beauty * female
    )

head(evals)
```

```
# A tibble: 6 x 12
  prof_id avg_eval num_courses num_students perc_evaluating beauty tenured
    <int>    <dbl>       <int>        <int>           <dbl>  <dbl>   <int>
1       1     4              4          416            62.0  0.202       0
2       2     3.53           3          104            87.0 -0.826       1
3       3     3.45           2          250            78.5 -0.660       1
4       4     4.01           8          223            84.3 -0.766       1
5       5     4.35           6          331            81.8  1.42        0
6       6     4.44           7         1849            59.8  0.500       1
# ... with 5 more variables: native_english <int>, age <int>,
#   female <int>, bty_tenured <dbl>, bty_female <dbl>
```

```r
# Fit interaction model
lm.3 = lm(avg_eval ~ 1 + beauty + female + bty_female, data = evals)
tidy(lm.3)
```

```
         term    estimate  std.error  statistic      p.value
1 (Intercept)  4.00946754 0.06107170 65.6518116 8.181477e-78
2      beauty  0.16519709 0.07602264  2.1729987 3.240702e-02
3      female -0.23173564 0.09402349 -2.4646569 1.561315e-02
4  bty_female -0.07052814 0.11344212 -0.6217104 5.357041e-01
```

The p-value associated with the interaction term ($p = .535$) is statistically significant. This indicates that it is likely the beauty effect is the same for male and female professors in the population.

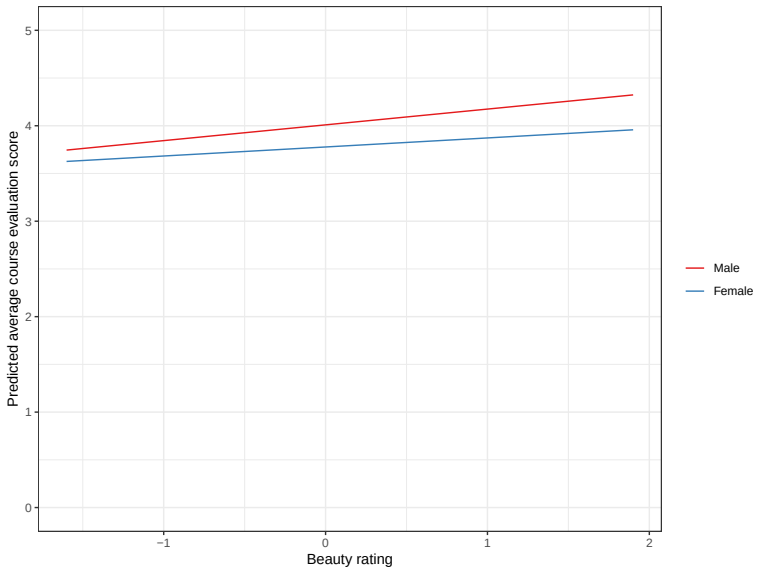
Visually Displaying the Model

It is often useful to visually display the fitted interaction model to aid model interpretation. To do this for the last model we fitted, we need to create a dataset that includes the predictors `beauty`, `female`, and `bty_female`. The last predictor, recall, was the product of the two main effects. So, when creating our data set, we use `crossing()` we include the data for the main effects and then `mutate()` the product term afterward.

```r
# Create new data set with main effects
profs = crossing(
  beauty = seq(from = -1.6, to = 1.9, by = 0.1),
  female = c(0, 1)
  ) %>%
  # Mutate on product term
  mutate(
    bty_female = beauty * female
    )

profs %>%
  mutate(
    # Compute fitted values for the data
    yhat = predict(lm.3, newdata = profs),
    # Make female a factor for better plotting
    female = factor(female, levels = c(0, 1), labels = c("Male", "Female"))
    ) %>%
  ggplot(aes(x = beauty, y = yhat, color = female)) +
    geom_line() +
```

```
    theme_bw() +
    xlab("Beauty rating") +
    ylab("Predicted average course evaluation score") +
    scale_color_brewer(name = "", palette = "Set1") +
    ylim(0, 5)
```



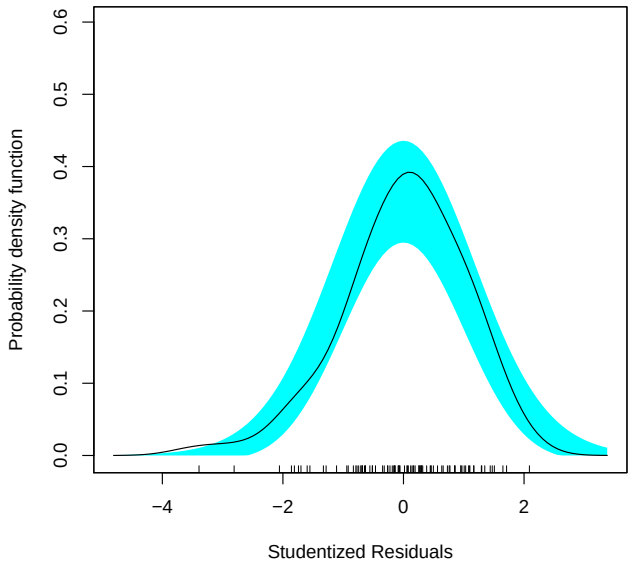
Model Assumptions

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.

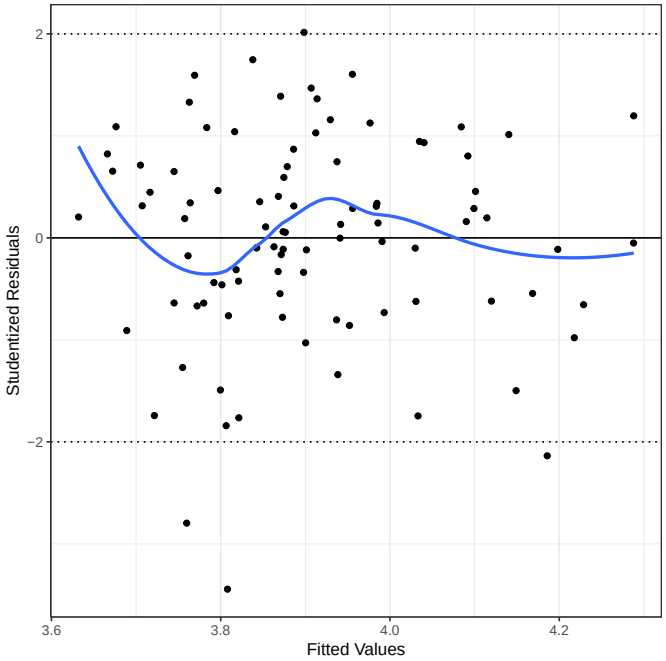
```
# Create fortified data
out.3 = augment(lm.3)
head(out.3)
```

```
  avg_eval     beauty female bty_female  .fitted    .se.fit     .resid
1 4.000000  0.2015666      1  0.2015666 3.796814 0.07021478  0.2031860
2 3.533333 -0.8260813      0  0.0000000 3.873001 0.08070967 -0.3396680
3 3.450000 -0.6603327      0  0.0000000 3.900382 0.07295781 -0.4503825
4 4.012500 -0.7663125      1 -0.7663125 3.705186 0.10514497  0.3073141
5 4.350000  1.4214450      1  1.4214450 3.912299 0.12700735  0.4377014
6 4.442857  0.5002196      0  0.0000000 4.092102 0.07666870  0.3507548
        .hat    .sigma     .cooksd .std.resid
1 0.02505096 0.4455771 0.001382151  0.4638595
2 0.03309927 0.4446054 0.005188851 -0.7786590
3 0.02704649 0.4434774 0.007362020 -1.0292458
4 0.05617516 0.4448488 0.007565432  0.7130507
5 0.08196435 0.4434748 0.023668410  1.0297500
6 0.02986782 0.4445108 0.004959709  0.8027343
```

```r
# Examine normality assumption
sm.density(out.3$.std.resid, model = "normal", xlab = "Studentized Residuals")
```




```r
# Examine other assumptions
ggplot(data = out.3, aes(x = .fitted, y = .std.resid)) +
  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")
```



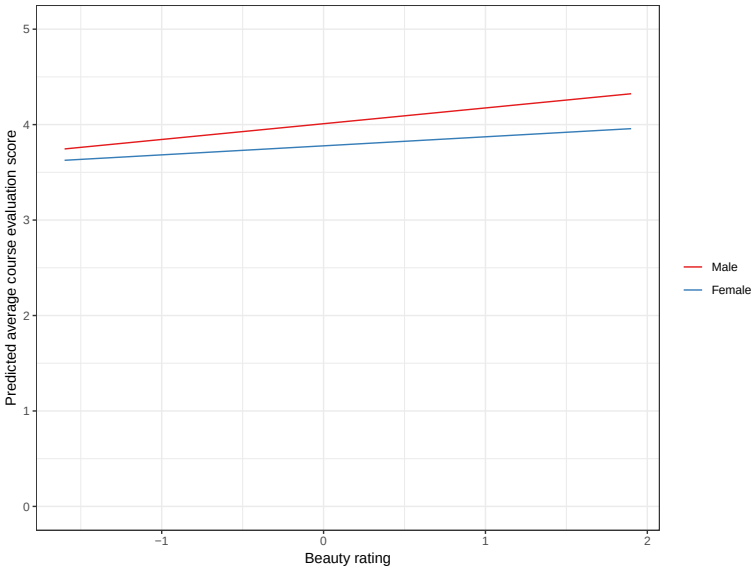
Based on the density plot of the studentized residuals, there is some question about whether the normality assumption is satisfied. The scatterplot of the model's studentized residuals versus its fitted values suggests that the assumption of homoskedasticity is reasonablly satisfied. The loess line (indicating 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 gender 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.