Kevin Skelly

# I Hate My Professor: An Examination into Underlying Biases in Student Evaluations

## Introduction:

When examining the performance of college professors, one of the most commonly employed strategies at the university level is the utilization of student evaluations of instruction (SEIs). While students taking the class seem to be a good representation of a professor's performance, have universities neglected the chance that students exhibit bias in their reviews? This report examines the possibility of bias related to attractiveness, sex, race, and age appearing in SEIs. This data comes from 463 observations related to course evaluations, course characteristics, and professor characteristics, collected by the University of Texas at Austin. In this report, I will utilize simple and multiple linear regression, log-linear regression, quadratic regression, and interaction terms to examine the effect of factors such as attractiveness, sex, race, and age on the course evaluation.

## Methods:

The dataset used in this report contains eight features, with each row representing a single class taught by a single instructor. Four binary variables–named *nnenglish*, *female*, *minority*, *onecredit*, and *intro*–represent whether the instructor is a non-native English speaker, whether the instructor is female, if the instructor is a minority, if the class is a one-credit class, and if the class is an intro level class, respectively. In all of these cases, a 1 represents the affirmative case, and a zero the negative case. Additionally, there are three ratio variables, *course_eval*, *beauty*, and *age*. The *course_eva*l column indicates the average evaluation score of the SEIs, ranging from 1 (very unsatisfactory) to 5 (excellent). *Beauty* denotes a rating of the instructor's physical appearance by a panel of six students, averaged across the six panelists, and shifted to have a mean of zero. *Age* is simply recorded as the age of the professor.

Analysis was performed by first using descriptive statistics to summarize the instructors and the courses in this dataset. Next, I employed descriptive statistics to gain an understanding of the course evaluations by finding the mean and standard deviation of the variable, since it is the target of the report. With this knowledge in mind, I conduct normality tests on course evaluation and estimate a simple linear regression on beauty. With this as a base case, I am then able to perform multiple linear regression, followed by quadratic and linear-log regression, and multiple regression with an interaction term, to assess an optimal model for assessing bias in student course evaluation. The results of the study are presented below.

**Results:**

Descriptive statistics proved successful in summarizing key features of the dataset. Regarding demographic characteristics of professors, the mean function concluded that 42% were female, 13% minority, and 6% were not native English speakers. The median age was 48 years with a standard deviation of roughly 10 years. Course analysis found that 33.9% of courses were introductory courses, and 5.8% were one-credit courses. Course evaluations had equal means and medians of 4.0, with a standard deviation of 0.55 points, the maximum score being 5.0, and the minimum of 2.1.

Simple linear regression indicated a moderate success in predicting course evaluations based on beauty. After passing normality tests and performing regression, the model demonstrated an $R^2$ of 0.036, with 95% confidence that beauty increases by at least 0.069 and at most 0.196 with a 1-point increase in course evaluation. Below is the graph of the model.

*Figure 1: Initial simple linear regression model*


Continuing to examine this relationship, the next step was checking for omitted variable bias with a multiple regression model, adding in independent variables intro, onecredit, female, minority, nnenglish, and age. After performing this regression, the wald test with simple regression returned a p-value of 3.268e-11, indicating that the added variables revealed omitted variable bias in the simple model. Fine-tuning this model, I tested models with a quadratic representation of age, and a logarithmic age variable. Wald-testing these models proved no significant difference between them and the quadratic model. The results of these models are presented below in table 1.

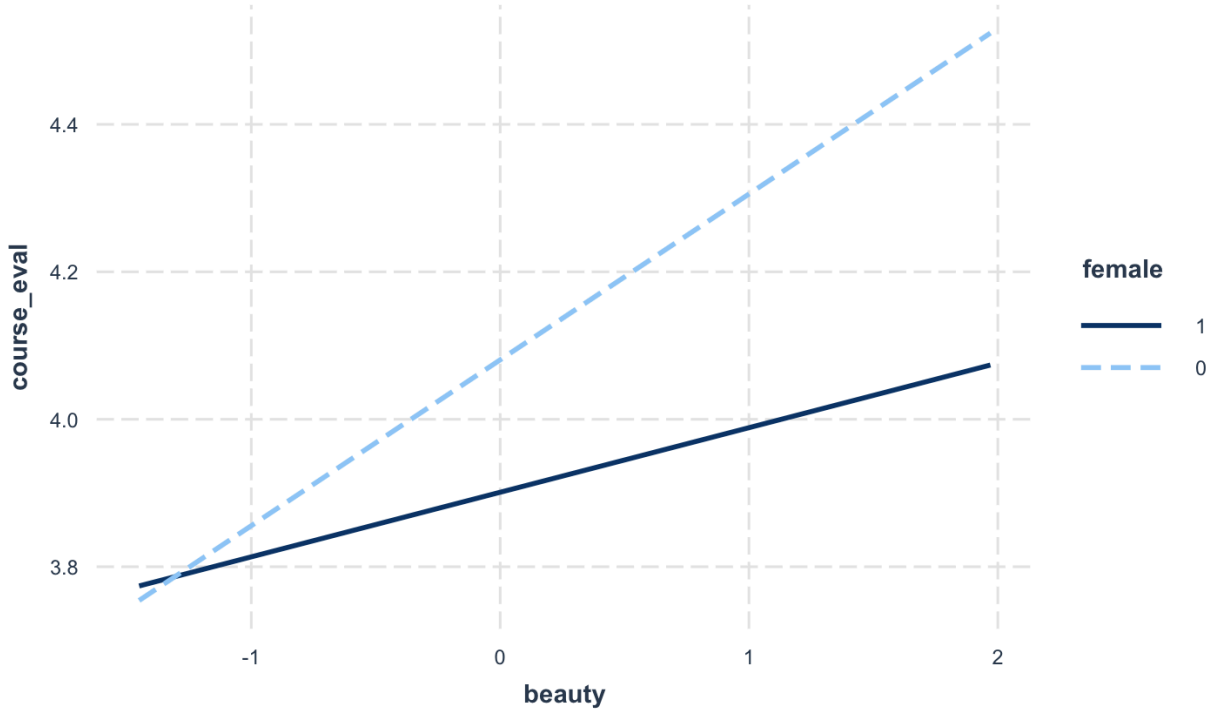
*Table 1: Results of various regressions on course evaluation using the TeachingRatings dataset*

| ***Dependent Variable:*** *[LD1] Course Evaluation Score* | | | | | |
|---|---|---|---|---|---|
| ***Independent Variable*** | ***Simple Linear*** | ***Multiple Linear*** | ***Quadratic Model*** | ***Linear Log Model*** | ***Beauty & Age Interaction*** |

| | | | | | |
|---|---|---|---|---|---|
| *Intercept* | *4.00*** (0.025)* | *4.17*** (0.139)* | *3.68*** (0.550)* | *4.37*** (0.474)* | *4.14*** (0.140)* |
| *Beauty* | *0.13*** (0.032)* | *0.16*** (0.031)* | *0.16*** (0.031)* | *0.16** (0.031)* | *0.22*** (0.047)* |
| *Female* | - | *-0.18*** (0.052)* | *-0.19*** (0.052)* | *-0.18*** (0.052)* | *-0.18*** (0.052)* |
| *Minority* | - | *-0.17** (0.068)* | *-0.17** (0.068)* | *-0.17** (0.068)* | *0.00 (0.56)* |
| *Non-native English Speaker* | - | *-0.24** (0.096)* | *-0.24** (0.096)* | *-0.24** (0.096* | *-0.27** (0.095)* |
| *Intro* | - | *0.01 (0.057)* | *0.00 (0.056)* | *0.01 (0.057)* | *0.00 (0.095)* |
| *One Credit* | - | *0.63*** (0.108)* | *0.62*** (0.109)* | *0.63*** (0.108)* | *0.66*** (0.108)* |
| *Age* | - | *0.00 (0.003)* | *0.02* (0.023)* | *-0.02 (0.023)* | *0.00 (0.003)* |
| *$Age^2$* | - | - | *0.00 (0.000)* | - | - |
| *Log(Age)* | - | - | - | *-0.08 (0.121)* | - |
| *Beauty: Female* | - | - | - | - | *-0.14* (0.063)* |
| ***Summary Statistics*** | | | | | |
| *Adjusted $R^2$* | *0.035* | *0.143* | *0.142* | *0.142* | *0.150* |
| *n* | *463* | *463* | *463* | *463* | *463* |

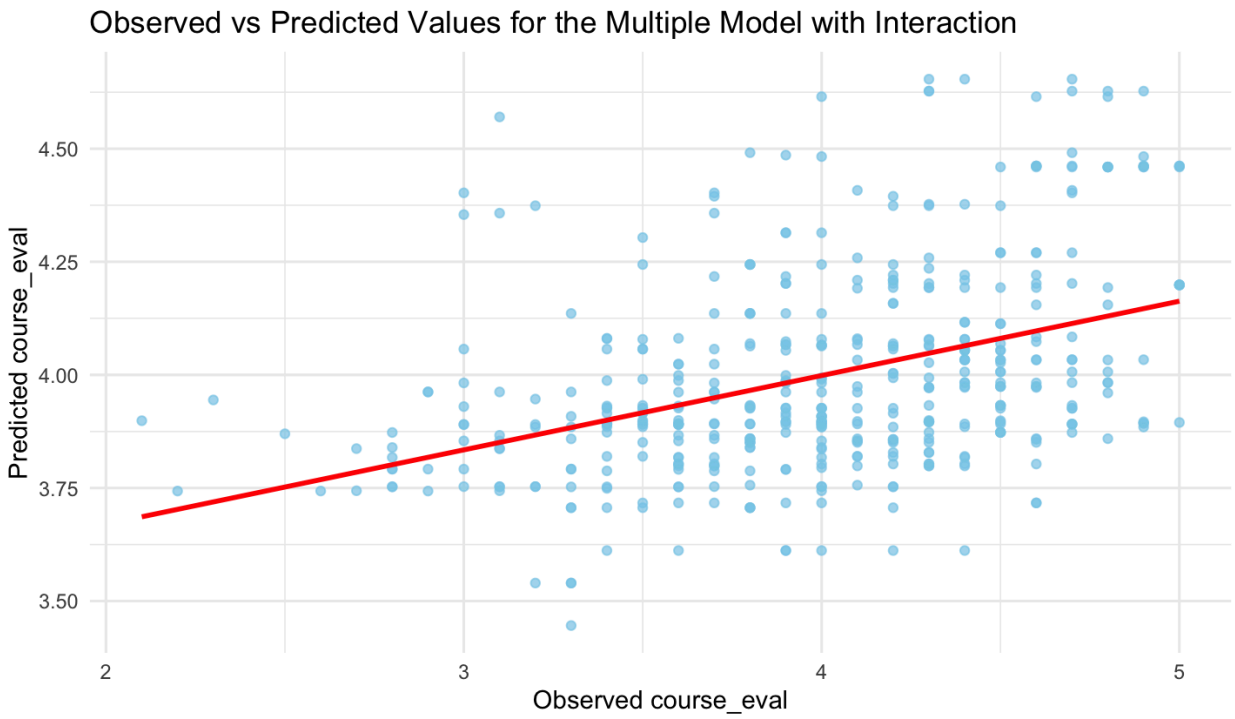
Finally, as seen in the table, I tested the multiple linear regression with an interaction term, female*beauty. This added term demonstrated a slight significance, indicating an improved model.

*Figure 2: Interaction plot of beauty on course evaluation given sex*




To further evidence this, the model had an adjusted $R^2$ of 0.15, the highest of all the models. Because of this, I have added below the comparison of observed and predicted variables for this regression, which performed the best of our results.

*Figure 3: Comparison of observed vs. predicted Values for multiple regression with interaction*




**Conclusion**

The most conclusive result of the regressions performed below can be gleaned from the regression table. Notable factors that had an effect on course evaluation scores (denoted by asterisks in table 1) include: one credit, non-native English speaking, beauty, sex, and beauty given sex. These terms are listed in order of their coefficients in the regression. What conclusions can we draw from this? First, we can conclude that having a one-credit course has a high positive effect on course evaluation, and non-native teachers tend to perform worse. These can be explained quite simply, as students will tend to perform well in one-credit classes, and thus will be likely to leave a high review. Similarly, they might struggle if there is a language barrier between them and their teacher, prompting a lower review.

Now remains beauty, sex, and their interaction. As beauty increases, students are shown to give higher evaluations, evidenced by the coefficient of 0.22. Similarly, as sex increases (i.e. the binary variable switches from 0 for male to 1 for female), students give lower reviews

(coefficient -0.18). Finally, the interaction plot demonstrates that the positive relationship exists for beauty and course evaluation, but it is much more of a positive relationship for males than females.

From these findings, it is clear that students are harsher on female teachers than males in their course evaluations. Additionally, they tend to give more attractive teachers higher ratings, but this relationship is way less notable for female teachers. It can certainly be argued that this attractiveness bias has a hidden relationship, as there are studies that humans find those they like (i.e. professors they give high ratings) more attractive than people that they dislike. However, the relationship between sex and course evaluation indicates that students have a clear bias towards male teachers over females. This evidences an underlying gender bias in academic studies, particularly demonstrated among students. Because of this, students should not continue administering course evaluations at this university.