Math 158 Project: Assignment 2

Sophia Hui and Allison Kirkegaard

February 19, 2018

Introduction

Our data comes from an evaluation of Teach for America conducted by Mathematica Policy Research in 2004, which was designed as a year-long randomized controlled trial including 17 elementary schools from six regions across the United States. It comprises student pre- and posttest scores in math and reading, student survey responses, and teacher survey responses. We chose to have students be our observational units, and merged their teachers' data with their own so that our model of student test performance can include both student and teacher characteristics.

Hypotheses

We are introducing a new variable: the difference between pretest and posttest scores in math; the variable will be called 'diff'. We will be testing the linear relationship between 'diff' and the teacher's years of experience. Our explanatory variable will be the student's teacher's total years of teaching experience, and our response variable will be Y = posttest score - pretest score in math. The linear regression model we will be using is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

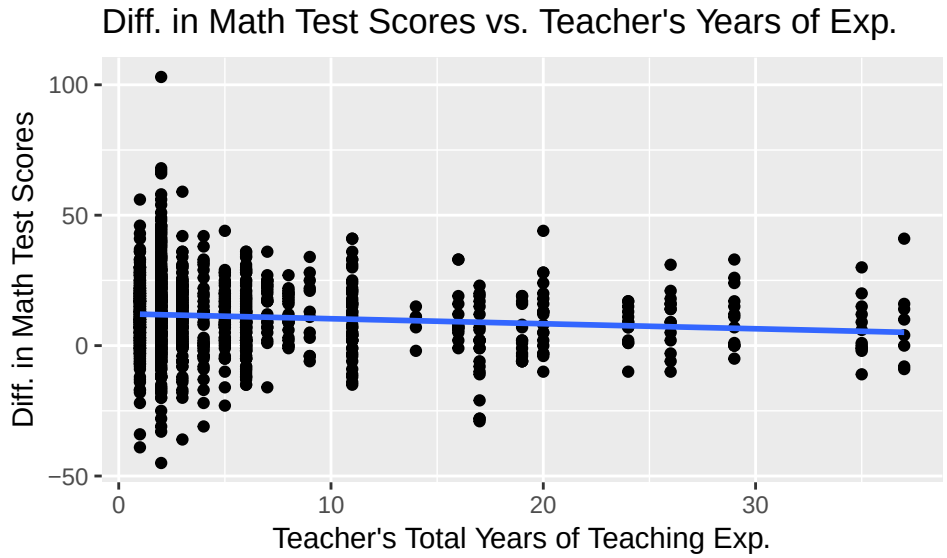
Our hypothesis is:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

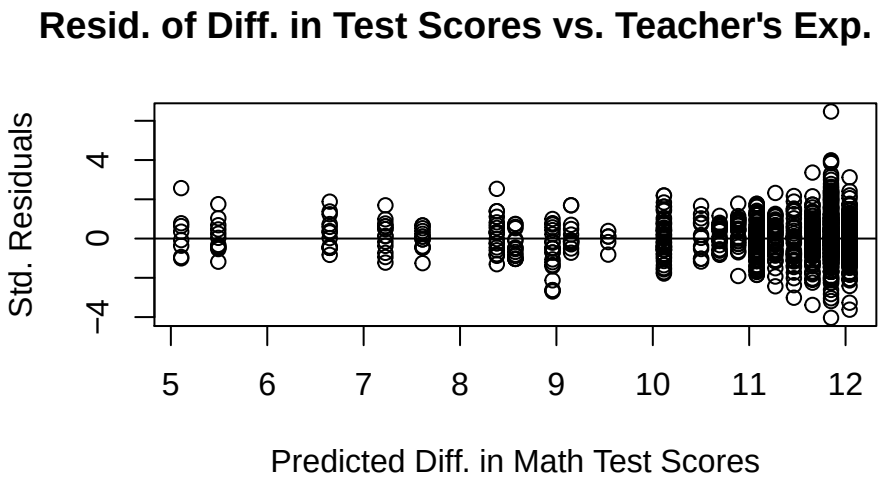
Note that 100 observations were removed because those teachers did not provide how many years of experience they had in teaching. It is possible that those teachers were primarily first-year teachers who thought they had no experience, but since the survey instructed teachers to include the current school year in their total years of teaching experience and thus all teachers should have had at least one year, we chose to omit these missing values rather than recode them as zeros.

Linear Regression

Because our explanatory variable is discrete, it is difficult to see a linear relationship between our explanatory and response variables. However, there is no curvilinear relationship evident either, so in the interest of parsimony we will assume that the relationship is linear.



The residuals from our linear model appear to be normally distributed. Upon first glance, their variance appears to be nonconstant (with larger residuals associated with higher predicted differences in test scores), but this actually seems to be a result of having many observations (students) with low values for the total years of their teacher's experience, which is associated with higher predicted differences in test scores in our model. Since we have so many observations at these values, there are more values at the extremes, but the density of points in the middle is also much greater. Thus we have normally distributed residuals with approximately constant variance, which satisfies the technical conditions for linear regression.



Having determined that simple linear regression is appropriate, we can now examine our model. Our linear model is

$$\hat{Y} = 12.235 + -0.193X,$$

$p < 0.001$. With such a low p-value, we can reject H_0 and conclude that there exists a linear relationship between students' differences in pre- and posttest scores and their teachers' total years of teaching experience.

```
tidy(test_lm)
```

```
##          term estimate std.error statistic   p.value
```

```
## 1 (Intercept)   12.235    0.5129     23.85 2.47e-103
## 2        a1_a   -0.193    0.0574     -3.35  8.20e-04
```

We can verify that this relationship exists by conducting an F-test of $\beta_1 = 0$ versus $\beta_1 \neq 0$. Finding an F statistic of 11.3 and an associated p-value less than 0.001, we can indeed reject $H_0 : \beta_1 = 0$.

```
tidy(anova(test_lm))
```

```
##        term   df  sumsq meansq statistic p.value
## 1      a1_a    1   2235   2235      11.3 0.00082
## 2 Residuals 1207 239781    199        NA      NA
```

Based on the residual plot above, our linear model appears to fit our data well. However, our R^2 value is very low. Only 0.924% of the variability of the response data is explained by our model:

```
glance(test_lm)$r.squared
```

```
## [1] 0.00924
```

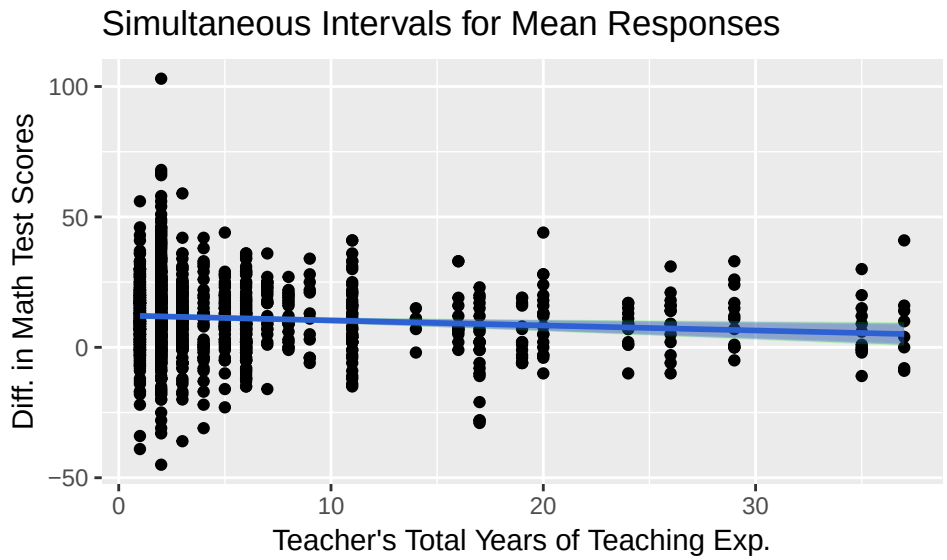
This means that while there is a linear relationship between students' differences in test scores and their teachers' years of teaching experience, there are other explanatory variables in our dataset that explain much more of the variability in the students' differences in test scores.

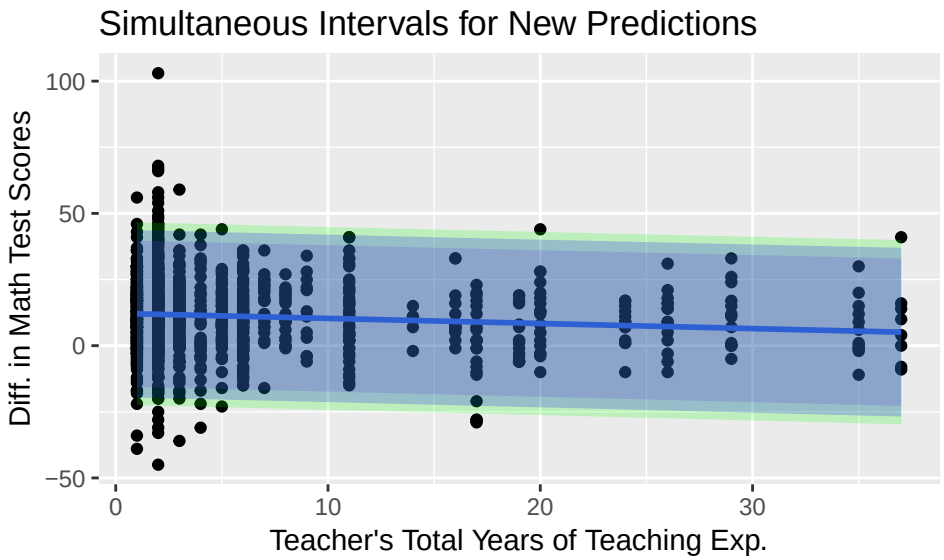
In this data, we might be particularly interested in the case where a student's teacher is in their first year of teaching. We can construct 95% prediction and confidence intervals at this value:

```
##   a1_a .fitted .se.fit lower_PI upper_PI lower_CI upper_CI
## 1    1      12    0.48    -15.6     39.7     11.1       13
```

By constructing a 95% prediction interval, we are 95% confident that the difference in pre- and post-test math scores will be in the interval (-15.6, 39.7) for a student who has a teacher with no prior experience ($X_h = 1$). We also constructed a 95% confidence interval, which tells us that if we were to repeatedly samples, then 95% of our confidence intervals will capture the true mean score difference for students with teachers with 1 year of experience (including the current school year).

We may also be interested in simultaneous inference. Here we construct confidence and prediction intervals for all of the observations in our dataset (purple = no adjustment, green = Working-Hotelling and Scheffe, blue = Bonferroni):





It is important to adjust for multiple comparisons because with over 1000 observations in our dataset, if we created intervals for each of them we could hardly expect all of them to cover the true values with probability $(1-\alpha)$. We need to adjust the intervals so that $(1-\alpha)$ is the probability that the total range of observations are contained in the appropriate confidence or prediction intervals. Otherwise, random chance might lead us to find a b_0 and b_1 such that the mean responses were only correct for a particular range of x values.

Conclusion

Initially, we were concerned by the scatterplot, because our explanatory variable only had discrete values, which yielded in a plot that did not look linear. However, we were able to reject our null hypothesis and conclude that there was a negative relationship between the total number of years of teaching experience and the difference in math pre- and post-test scores. This surprised us, because typically, more experienced teachers have better success in the classroom. We think this unexpected negative correlation may be a result of teachers' TFA status acting as a confounding variable, since almost all TFA teachers have one or two years of teaching experience and most non-TFA teachers have much more teaching experience. It could be that TFA teachers are more successful in raising student test scores for reasons other than their years of teaching experience, such as their educational backgrounds or the TFA training program. Thus in future assignments, we will want to add other explanatory variables to our model, and potentially interactions between TFA status and other characteristics.