Linear Regression Review

Example: Sex discrimination in bank salaries. In the 1970’s, Harris Trust was sued for sex discrimination in the salaries it paid its employees. One approach to addressing this issue was to examine the starting salaries of all skilled, entry-level clerical workers between 1965 and 1975. The data is saved under **banksalary.csv**, and relevant R code can be found in Class > Code under **banksalary.Rmd**.

*First, we will speculate on what we expect to find and then we will perform an analysis using the data.*

Speculating… **Open the data file** so that you can see the data, but **do not do anything else with the data** on the computer until question 6.

1. Identify the observational units, the response variable, and explanatory variables.

2. Given the mean starting salary of male workers ($5957) was 16% higher than the mean starting salary of female workers ($5139): Is this enough evidence to conclude sex discrimination exists? If not, what further evidence would you need?

3. How would you expect age, experience, and education to each be related to starting salary?

4. Why might it be important to control for seniority (number of years with the bank) if we are only concerned with the salary when the worker started?

5. Do you expect any explanatory variables (including sex) to be closely related to each other? What implications would this have for modeling?

Using the data…

*One approach is to construct a good model for beginning salaries while requiring sex as a predictor, to determine the significance of sex after controlling for the other covariates. Then we can explore interactions with sex to see if its effect is consistent across levels of other predictors.*

6. Use the data to address the primary question of interest here using only the beginning salary and sex variables. Be sure to discuss plots and summary statistics first, and then look at test(s) of significance.

7. Construct plots to investigate how each of the potential confounders (age, experience, education) is related to beginning salaries. Describe your findings.

8. Does seniority play a role in the variation of starting salaries? In what way?

9. Examine how the explanatory variables (including sex) are related to each other, if at all. What implications would this have for modeling?

10. Fit a simple linear regression model with starting salary as the response and education as the sole explanatory variable. Interpret the intercept and slope of this model; also interpret the R-squared value. Is there a significant relationship between education and starting salary?

Intercept:

Slope:

R2

Significance:

11. Does model1 from question 10 meet all linear regression assumptions? List each assumption and how you decided if it was met or not.

12. Is a model with all 4 confounding variables better than a model with just education? Justify with an appropriate significance test in addition to summary statistics of model performance.

13. You should have noticed that the term for age was not significant in the model3. What does this imply about age and about future modeling steps?

14. The relationship between experience and beginning salary exhibits some curvature. How might it be interpreted in this context? Determine whether a quadratic term in experience improves a model without curvature.

15. Based on model6, what conclusions can be drawn about sex discrimination at Harris Trust? Do these conclusions have to be qualified at all, or are they pretty clear cut? Interpret a 95% confidence interval for the male indicator variable in context to help with your response.

16. Do any explanatory variables exhibit an interaction with sex. If so, what are the implications for your answer in (15)?

17. Often salary data is logged before analysis. Would you recommend logging starting salary in this study? Support your decision analytically.

18. Regardless of your answer to (17), provide an interpretation for the coefficient for the male coefficient in model6a after logging starting salary.