Stats 500 Homework 2

Online Submission to Canvas, Due date: 11:59pm, September 19, 2025

Problem 1 (Maximum 2 pages). The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988.

1. Fit a regression model with weekly wages as the response and years of education and experience as predictors. Present the output.
2. What percentage of variation in the response is explained by these predictors?
3. Which observation has the largest (positive) residual? Give the case number.
4. Compute the mean and median of the residuals. Explain what the difference between the mean and the median indicates.
5. For two people with the same education and one year difference in experience, what would be the difference in predicted weekly wages?
6. Compute the correlation of the residuals with the fitted values. Plot residuals against fitted values. Explain the value of this correlation using the geometric (projection) interpretation of least squares.

Hints: Useful R functions: data(), lm(), summary(), residuals(), fitted(), which.max(), mean(), median(), cor(), plot(). Note that the experience variable has some negative values which most likely indicate missing data. Those observations should be removed from the analysis.

Problem 2. Prove the unbiasedness of $\hat{\sigma}^2$, that is, $E[\hat{\sigma}^2] = \sigma^2$.

Hints: Refer to lecture slides (Chapter 2).

Problem 3. [Exercise 3, Chapter 2 of Faraway's book] In this question, we investigate the relative merits of methods for computing the coefficients. Generate some artificial data by:

```
> x <- 1:20
> y <- x+rnorm(20)
```

Fit a polynomial in x for predicting y. Compute $\hat{\beta}$ in two ways – by $lm()$ and by using the direct calculation described in the chapter (i.e., the OLS formula introduced in lecture). Perform such regressions with polynomials of increasing degrees. At what degree of polynomial does the direct calculation method fail? Explain why this fails. (Note the need for the $I()$ function in fitting the polynomial, for instance, $lm(y \sim x + I(x^2))$).

Note: This problem tells us that directly using the OLS formula to calculate $\hat{\beta}$ may not be a good option in practice. The $lm()$ function in R uses a computationally better way; see Chapter 2.7 of the textbook for details.