

# Statistics 108

## Homework Assignments

Note, all problems listed here are to be written up and handed in on the due date provided. Homework is posted concurrently with the class material. Sections may be assigned at different times but have the same due date as previously assigned problems.

### Chapter 2 and chapter 3 problems have a due date of Oct 16

#### Chapter 2

- (P1) The data set **manager.xlsx** contains data on employee *rating*, *years of experience*, *salary* and the variable *origin*. We would like to investigate whether rating or experience are predictive of salary. Obtain scatterplots for both possible predictor variables and discuss whether a linear regression is plausible. Provide labels and titles for the scatterplots.
- (P2) For the data set **manager.xlsx** fit a linear regression model with *salary* the outcome variable and *years of experience* the predictor variable. Provide estimates of slope and intercept. Interpret the slope in context. Is there a meaningful interpretation of the slope? Why or why not?
- (P3) For the regression in problem 2, what is your estimate of the average salary for employees with 5 years experience? What would you estimate is the difference in average salary for employees with 15 years experience and with 5 years experience?
- (P4) For the regression in problem 2, what is your estimate of the difference in salary between 2 randomly chosen employees, one who has 7 years of experience and one, who has 17 years of experience. State all conclusions in context. Show the calculations and justify your answer.
- (P5) Analyze the relationship between rating and salary using the data set **manager.xlsx**. Obtain the least squares line and interpret the slope.
- (P6) Compare the average salary of an employee whose rating is 3 to that of an employee whose rating is 6. What do you conclude based on the difference? State all conclusions in context. Refer to the data set **manager.xlsx**.
- (P7) Using the regression obtained in problem 5, calculate the residual sum of squares. Recalculate the residuals for this data when the intercept is set to 40 and the slope is set to 5.5. Compare the result to the residual sum of squares for the least squares estimates.

- (P8) For the regression in problem 5, what would your response be to someone who asked you to predict the salary of an employee with a rating of 10?

### Chapter 3

- (P1) Returning to the data set **manager.xlsx**, obtain the t-test for the hypothesis  $H_0 : \beta_1 = 0$  vs  $H_a : \beta_1 \neq 0$  from the regression of salary on years of experience. Provide the value of the test-statistic, the p-value and a conclusion in context. Does the result help to explain the seemingly strange result of a negative slope? Justify your answer.
- (P2) For the data set **manager.xlsx**, obtain the t-test of the hypothesis  $H_0 : \beta_1 = 0$  vs  $H_a : \beta_1 > 0$ , where  $\beta_1$  is the slope for the regression of salary on manager rating. Provide the test-statistic, the p-value and a conclusion in context. Be careful to only use computer output from a regression program that is correct for the question asked.
- (P3) For the regression of salary on manager rating (**manager.xlsx**), find the average salary for subjects who have a rating of 3.5 and compare to the average salary for those with a rating of 8.5. For both ratings find the 95% mean interval. Do these intervals overlap?
- (P4) For the regression of salary on manager rating (**manager.xlsx**), find the predicted salary for a subject who has a rating of 2.5 and compare to the predicted salary of someone with a rating of 8.4. For both ratings find the 95% prediction interval. Carry out the calculation using the formulas provided in class. You can use R to obtain summary statistics for the inputs needed for the formulas but you cannot use R to obtain the prediction interval directly.
- (P5) The mean rating is approximately 5.9. Without using the regression to calculate the mean salary of those whose rating is average, can you find out what the average salary is based on the usual sample statistics of mean and standard deviations of your data. Recall that you have two variables, rating and salary for which you have or can get summary statistics from your sample. Justify your answer using facts about the regression line and points on that line.
- (P6) The data file **restauranttips.csv** contains information on the amount of a bill (bill), tips (the tip), whether the bill was paid by cash or credit card, a variable that indicates which server, percent tip and day of the week. Carry out a regression analysis for the amount of the tip (Y) on the amount of the bill (X). Obtain a relevant graph, find the regression line, carry out a t-test to assess whether there is a linear relationship between the amount of the bill and the amount of the tip. Find a confidence interval for the slope parameter. Predict the average tip for a bill that was \$40. Find the mean interval for a bill that was \$40.