Stat 401 Homework 9: Due Tuesday, Dec. 5 by 1:40pm

General Instructions: Most homeworks will consist of a set of problems from the textbook, plus additional problems or data analyses. Turn in a physical copy of your homework at the beginning of class (or earlier) on the due date. For problems that require the use of R, include relevant and well-labeled R code and output with the problem. Additional R code and output may be included in an appendix.

Open Intro Exercises

• 7.2

• 7.8

• 7.24 (a), (b) and (d). For part (a), your description should address all four features of a scatterplot: form, direction, strength, and outliers. For part (d), explain your answer (don’t just write “yes” or “no”).

• 7.30. In part (a), write the equation of the fitted linear model equation using appropriate notation (not the “true” model).

• 7.44

• 8.4. Note: ”variance of the residuals” is $SSE \times (n - 1)$ and ”variance of the number of absent days for all students in the data set” is $SSR \times (n - 1)$. Additionally, (e) interpret the value of $R^2$ in context of this problem.

Additional Problems

1. For each of the following plots, explain (1) which of the four simple linear regression assumptions seem to be met and why, (2) which of the four simple linear regression assumptions seem to be violated and why, and (3) which of the four regression assumptions cannot be assessed from the given plot.

(a) Scatterplot of $y$ vs. $x$: 

![Scatterplot of Y vs. X](image)
2. For this problem, we are going to use salary data on faculty in a small Midwestern college in the early 1980s. The data were originally used for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data hold tenured or tenure track positions; temporary faculty are not included. Load the data set into R and few the variable descriptions using the following commands (you may need to install the alr4 package):

library(alr4)
data(salary)
?salary

(Note that it’s unfortunate the name of the data set is the same as the name of the response variable in the data set. Be careful while coding (or re-name the data set).)
(a) Fit the model

\[
\text{mod} <- \text{lm(salary} \sim \text{year*degree, data=salary)}
\]

i. Write the equation of the fitted regression model using appropriate statistical notation. Clearly define any variables used in the equation.

ii. Write a sentence interpreting each fitted coefficient in the model.

iii. Create a plot of salary versus year with the following properties:
- descriptive x-axis and y-axis labels, and a main title,
- different point characters and/or colors for the different levels of degree,
- fitted regression lines for each degree group using the fitted model above.

Write a few sentences describing what the plot tells you about the relationship between salary, year, and degree.

iv. Perform an F-test to compare the model with just year as a predictor to the model used above. What model is \( H_0 \)? \( H_a \)? Report the F test statistic, the p-value, and your conclusion for the test.

(b) Use the best subsets model selection procedure (ignoring interactions and transformations of variables) with adjusted \( R^2 \) as the model selection criterion to select a “best” model. Fit this model in R and report a summary of the model. Choose one of the coefficients in the model (not the intercept): Calculate a 95% confidence interval for this coefficient and write a sentence interpreting the interval.