Homework 6

Statistical Inference, Fall 2021

1- Answer to the following questions:

a. Ali is studying the correlation between hours spent studying and caffeine consumption among students at his school. During a given week, he randomly selects 20 students at his school and records their caffeine intake (mg) and studying time. Here is computer output from a least-squares regression analysis on his sample:

Predictor	**Coef**	**SE Coef**	**T**	**P**
constant	2.544	0.134	18.955	0.000
caffeine	0.164	0.057	2.862	0.010
S = 1.532 R-sq = 60%				

Assume that all conditions for inference have been met.

Which of these is a 95% confidence interval for the slope of the least-squares regression line?

b. Elvin compared the scores of a random sample of 41 students on a first exam in a certain class with their subsequent scores on the second exam. Here is computer output on the sample data:

Summary statistics

Variable	**n**	**Mean**	**stDev**	**SE Mean**
x = first exam score	41	59.5	19.7	3.1
y = second exam score	41	59.4	21.7	3.4

Regression: second exam score vs. first exam score

Predictor	**Coef**	**SE Coef**
constant	6.985	6.65
caffeine	0.881	0.11
S = 13.2 R-sq = 62.89%		

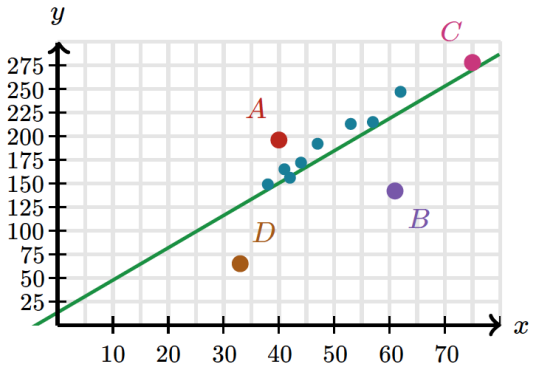
i. What conditions should be met for above inferences?

ii. Write an appropriate test statistic for testing the null hypothesis that the population slope in this setting is 0?

Homework 6

Statistical Inference, Fall 2021

c. Sara was curious which points in the scatterplot below were most influential in terms of the coefficient of determination, r^2.



Which point, if removed, would cause r^2 to decrease the most? why?

2- In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades.

Student	x_i	y_i
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

Find the Regression Equation.

3- Answer the following questions:

a. Each morning, Arianna runs. For a random sample of runs, she tracked the temperature (in degrees Celsius) and the distance run (in kilometers). The temperatures were negatively correlated with the distances. A 95% confidence interval for the slope of the regression line was (-0.02, 0.12).

Arianna wants to use this interval to test $H_0: \beta = 0$ vs $H_a: \beta \neq 0$ at the $\alpha = 0.05$ level of significance. Assume that all conditions for inference have been met.

Which of these is the most appropriate conclusion about the relationship between temperature and distance for Arianna's runs?

i. Reject H_0. Ariana can't conclude a linear relationship between temperature and distance.

ii. Fail to reject H_0. Ariana can't conclude a linear relationship between temperature and distance.

iii. Reject H_0. This suggests a linear relationship between temperature and distance.

iv. Fail to reject H_0. This suggests a linear relationship between temperature and distance.

b. Biologists observed a curved relationship between the average heart rate and life expectancy of several mammal species in a large sample. The biologist took the natural logarithm for both variables and observed a linear relationship in the transformed data.

The least-squares regression equation for the transformed data is shown here, where life expectancy (LE) is in years and heart rate (HR) is in beats per minute.

$$\widehat{ln(LE)} = 6.33 - 0.78\, ln(HR)$$

What is the predicted life expectancy of a mammal species whose average heart rate is 60 beats per minute according to this model? (Round your answer to the nearest year.)

4- **(R)** The *openintro* package contains a dataset called *absenteeism* that consists of data on 146 schoolchildren in a rural area of Australia. Spend some time reading the help file of this dataset. We are interested in seeing if the ethnicity (aboriginal or not), sex (male or female), and learning ability (average or slow) of the children affects the number of days they are absent from school.

a. Convert the *Eth*, *Sex*, and *Lrn* variables to binary variables. One way to do this is with the function *ifelse()*. You should construct them so that:

i. *Eth* = 1 if the student is not aboriginal and *Eth* = 0 if the student is aboriginal;
ii. *Sex* = 1 if the student is male and *Sex* = 0 if the student is female;
iii. *Lrn* = 1 if the student is a slow learner and *Lrn* = 0 is the student is an average learner.

b. Fit a linear model to the data with Days as the dependent variable and the three variables mentioned in (1) as explanatory variables.
c. Write the fitted model out using mathematical notation. Make sure you define the variables you use. Interpret all of the coefficients (including the intercept) in context.
d. Find and interpret the adjusted R^2 value for this model.

e. Create a residual plot. Describe what you see in the residual plot. Does the model look like a good fit?

Homework 6

Statistical Inference, Fall 2021

5- Determine if the following statements are true or false. If false, explain why.

a. By adding an explanatory variable to an existing MLR model if the variable is not a meaningful predictor, R2 will decrease and adjusted R2 will stay about the same.

b. A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.

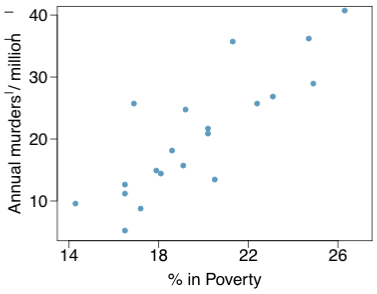
c. Correlation is a measure of any association between any two variables.

d. If predictors are collinear, then removing one variable will have no influence on the point estimate of another variable's coefficient.

6- The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−29.901	7.789	−3.839	0.001
poverty %	2.559	0.390	6.562	0.000

$s = 5.512$; $R^2 = 70.52\%$; $R^2_{adj} = 68.89\%$



a. Write out the linear model.

b. Interpret the intercept, slope, R^2.

c. Calculate the correlation coefficient.

d. What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

e. State the conclusion of the hypothesis test from part d in context of the data.

f. Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

g. Do your results from the hypothesis test and the confidence interval agree? Explain.

Homework 6

Statistical Inference, Fall 2021

7- A traveler thinks that the time it takes them to commute is different based on which day of the week it is. For a few weeks, they record their commute time each day, Monday through Friday. Test their claim at the 5% level.

a. What kind of test should we conduct?

b. What are the hypotheses?

c. Complete the missing spaces in partially filled table.

Source	DF	Sum of Squares	Mean Square	F Value	Prob.
Day(Groups)		14.28			
Error(Residuals)					
Total	19	30.2			

d. What is the correct decision?

e. What is the appropriate conclusion/interpretation?

8- **(R)** The built-in "state.77" dataset in R includes information about 50 states in US. Take life expectancy as the response and the remaining variables as predictors. Using this dataset, answer the following questions:

a. Using backward selection and p-value as the criterion, find a multiple (or simple) linear regression model which is the best for predicting life expectancy using others. Show each step in your report.
b. Find the simple linear model of data. (Assume X = Murder and Y = Life expectancy).
c. Plot histogram of residuals and find mean and sd for part (b).
d. Plot QQ-plot of residuals versus a zero-mean normal distribution with sd of part (b). Do you observe that residuals are nearly normal?