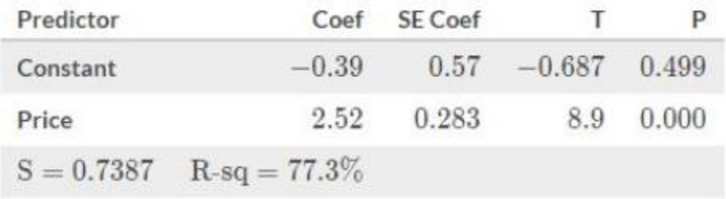
Homework 6

Statistical Inference, Spring 2021

1- Answer to the following questions:

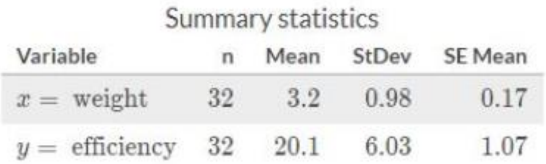
1-1 A panel of judges was asked to judge the quality of a random sample of 24 brands of potato chips. Here is computer output from a least-squares regression analysis on the relationship between the price of each brand (in dollars per pack) and its rating:



Predictor	Coef	SE Coef	T	P
Constant	−0.39	0.57	−0.687	0.499
Price	2.52	0.283	8.9	0.000
S = 0.7387 R-sq = 77.3%				

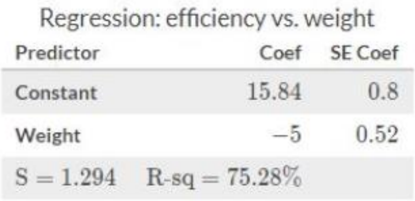
Assume that all conditions for inference have been met.
Calculate 90% confidence interval for the slope of the least-squares regression line?

1-2 Christine collected data about the weight (in thousand kilograms) and fuel efficiency (in kilometer per liter) of a random sample of 32 car models. Here is computer output on the sample data:



Summary statistics

Variable	n	Mean	StDev	SE Mean
$x =$ weight	32	3.2	0.98	0.17
$y =$ efficiency	32	20.1	6.03	1.07



Regression: efficiency vs. weight

Predictor	Coef	SE Coef
Constant	15.84	0.8
Weight	−5	0.52
S = 1.294 R-sq = 75.28%		

a. What conditions should be met for the above inference?

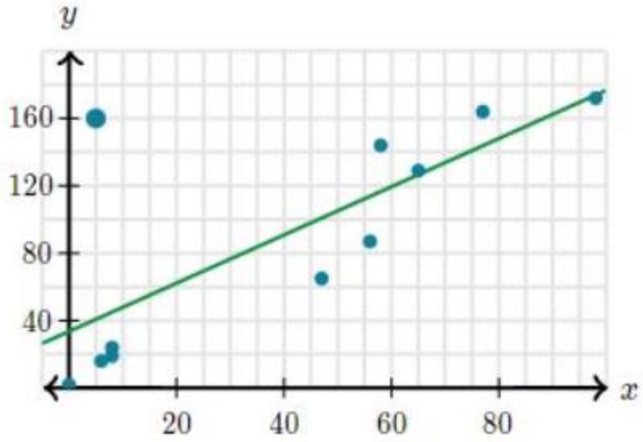
b. Write an appropriate test statistic for testing the null hypothesis that the population slope in this setting is 0?

1-3 Determine if the following statements are true or false. If false, explain why?

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

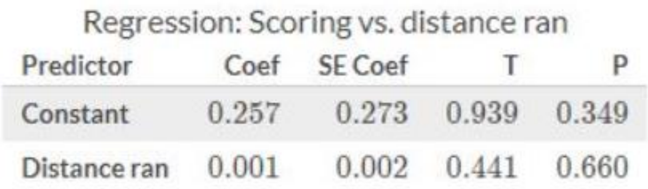
b. Correlation is a measure of the association between any two variables.

2- The scatter plot below displays a set of bivariate data along with its least-squares regression line. Consider removing the outlier and calculating a new least-squares regression line. Specify which point is an outlier? Why? What effect(s) removing this point would have?



3- Answer to the following questions:

3-1 Sara took a random sample of professional soccer players and found a positive linear relationship between how much they ran (in meters per minute) and how much they scored per game. Here is computer output from a least-squares regression analysis on her sample:



Regression: Scoring vs. distance ran

Predictor	Coef	SE Coef	T	P
Constant	0.257	0.273	0.939	0.349
Distance ran	0.001	0.002	0.441	0.660

Sara wants to test $H_0: \beta = 0 \ v.s. H_a: \beta \neq 0$. Assume that all conditions for inference have been met. At the $\alpha = 0.01$ level of significance, is there sufficient evidence to conclude a linear relationship between these variables for all players in this population? Why?

3-2 Nader tracked how much toothpaste he used (in mg) and how long he brushed his teeth (in seconds) for a random sample of brushings. He saw a positive relationship between the amounts and times. A 95% confidence interval for the slope of the regression line was 0.38±0.56.

Nader 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 Nader’s brushing habits?

a) Fail to reject H0. Nader can't conclude a linear relationship between how much toothpaste he uses and how long he brushes.

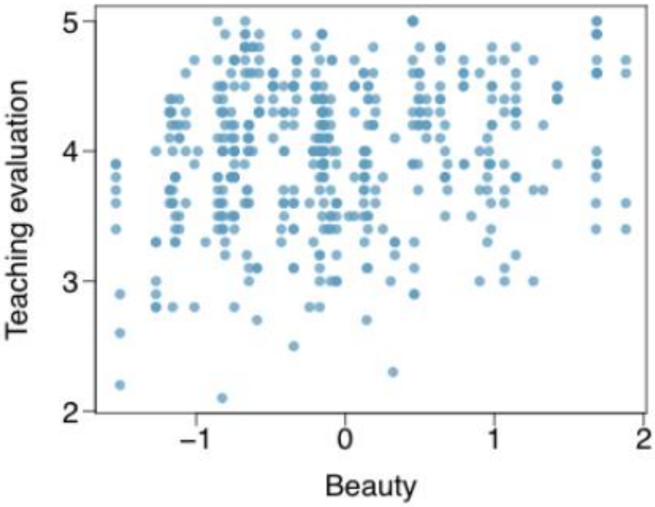
b) Reject H0. Nader can't conclude a linear relationship between how much toothpaste he uses and how long he brushes.

c) Fail to reject H0. This suggests a linear relationship between how much toothpaste he uses and how long he brushes.

d) Reject H0. This suggests a linear relationship between how much toothpaste he uses and how long he brushes.

4- Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at the University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

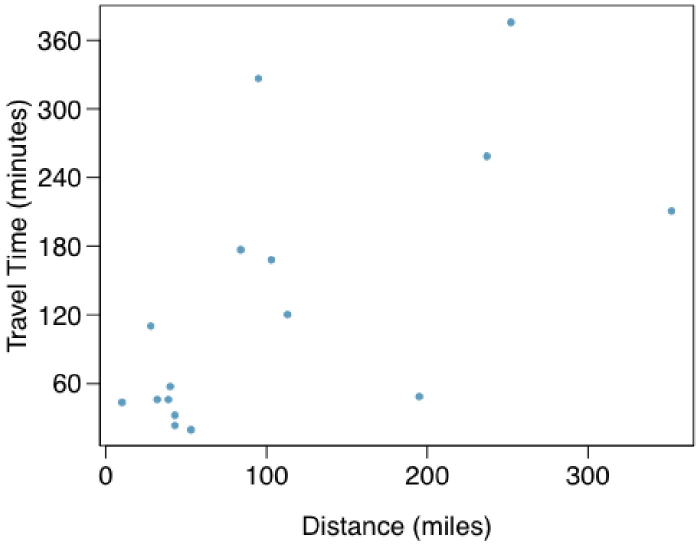
	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	_____	0.0322	4.13	0.0000



a) Given that the average standardized beauty score is −0.0883 and the average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.
c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.

5- The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).



The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

a) Write the equation of the regression line for predicting travel time.
b) Interpret the slope and the intercept in this context.
c) Calculate R^2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R^2 in the context of the application.
d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.
e) It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.
f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

6- (**R**) The dataset **starbucks** in the **openintro** package contains nutritional information on 77 Starbucks food items. Spend some time reading the help file of this dataset. For this problem, you will explore the relationship between the calories and carbohydrate grams in these items.

a) Create a scatterplot of this data with calories on the x-axis and carbohydrate grams on the y-axis, and describe the relationship you see.

b) In the scatterplot you made, what is the explanatory variable? What is the response variable? Why might you want to construct the problem in this way?

c) Fit a simple linear regression to this data, with carbohydrate grams as the dependent variable and the calories as the explanatory variable. Use the lm() function.

d) Write the fitted model out using mathematical notation. Interpret the slope and the intercept parameters.

e) Find and interpret the value of R^2 for this model.

f) Create a residual plot. The ggplot2 function fortify can help a lot with this. Describe what you see in the residual plot. Does the model look like a good fit?