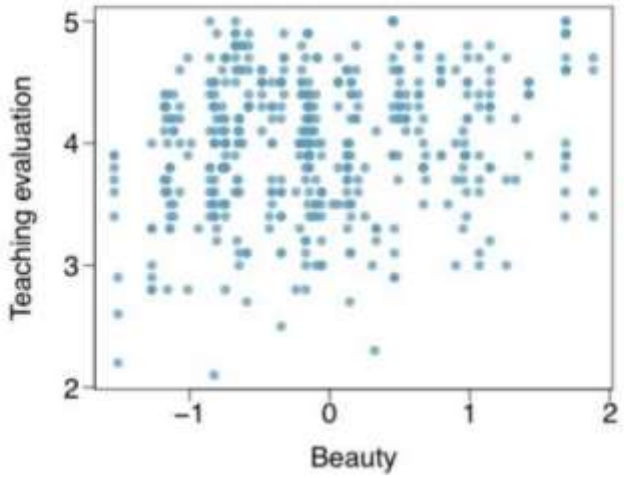
Homework 6

Statistical Inference

1. 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, a 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.

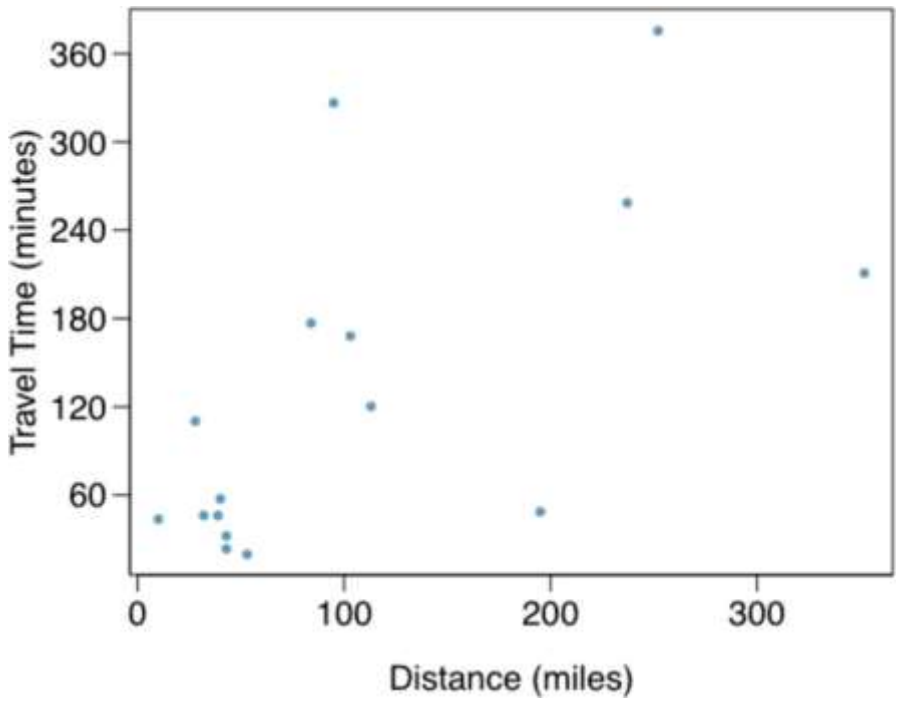
	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	_____	0.0322	4.13	0.0000



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.

2. 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.

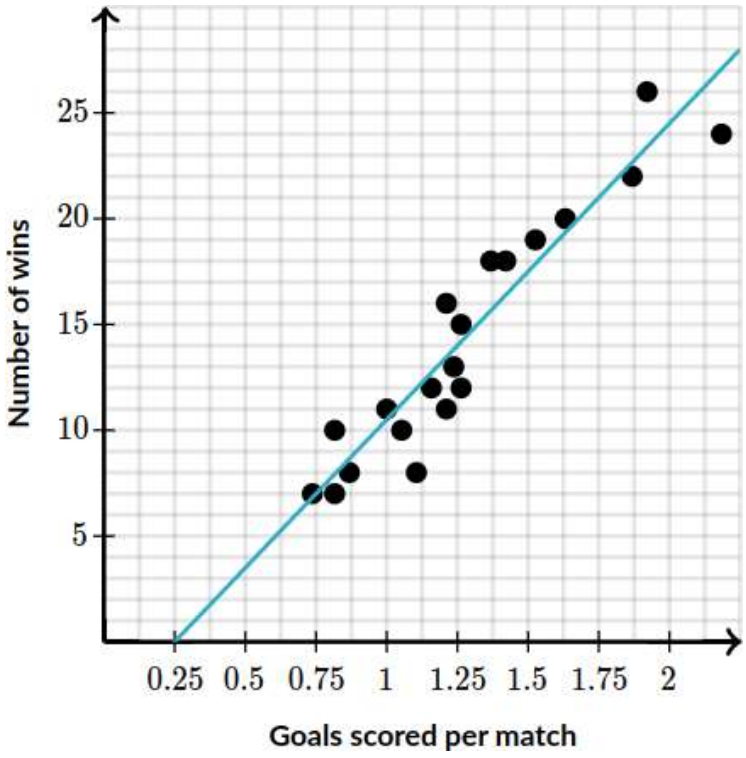
c. Calculate $R2$ of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret $R2$ in the context of the application.

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?

3. In one League, there were 20 soccer teams, and each team played a total of 38 matches. The scatter plot below shows the average number of goals each team scored per match, and how many total matches each team won. Each dot on the scatter plot represents a team. A line was fit to the data to model the relationship between scoring goals and winning games. Calculate linear equations best describes the given model.



4. Ali intends to investigate the relationship between study hours and caffeine consumption among the students of his school. He randomly selects 20 students from his school and records their caffeine intake (in milligrams) and time spent studying in a given week. Here is the computer output from the 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
$S = 1.532$	$R\text{-sq} = 60.0\%$			

Assume that all conditions for inference have been met. Calculate the 95% confidence interval for the slope of the least squares regression line.

5. Consider a research study on the behavior of customers of a web shop and the relationship between sales and appreciation of the website. A number of visitors to the website were asked to mark their appreciation of the website on a 5 point scale, ranging from 1 (very bad) to 5 (very good). For these visitors was also recorded whether they actually bought something on the website. The results of the study are summarized in the following contingency table:

Count			
Website Appreciation	Buy (0=no, 1=yes)		
	0	1	Total
1	6	2	8
2	5	2	7
3	7	6	13
4	3	7	10
5	1	8	9
Total	22	25	47

We introduce the variable Y to denote whether a customer has bought something (Y = 1) or not (Y = 0). The numerical variable X is introduced to denote the appreciation value for a customer X = 1,...,5. In a logistic regression analysis the probability of a buy (Y = 1), as a function of the appreciation value X.

a. Consider an arbitrary visitor of the website and assume that you do not have any information about how much this customer appreciates the website. How large is the probability that this customer has actually bought something?

b. If you would have to make a guess whether the customer has bought something, what would the best guess be?

c. If we decide not to use the appreciation value X we set $b_1 = 0$. In this case R glm function reports that the optimal value for b_0 then is 0.128. Explain why this is indeed a good value.

If we decide to use the appreciation value we can construct a better approximation for P(Y = 1), R calculates that the optimal values for b_0 and b_1 are $b_0 = -2.3190$ and $b_1 = 0.795$.

d. Calculate for each value of X = 1,...,5, the approximation for P(Y = 1), when the logistic model is $b_0 = -2.3190$ and $b_1 = 0.795$

e. Assuming a threshold of $p = 0.5$, what are the sensitivity and specificity of the model in part (c)? What are the sensitivity and specificity of the model in part (d)?

6. Suppose we wanted to know if people's ability to report words accurately was affected by which ear they heard them in. To investigate this, we performed a dichotic listening task. Each participant heard a series of words, presented randomly to either their left or right ear, and reported the words if they could. Each participant thus provided two scores: the number of words that they reported correctly from their left ear, and the number reported correctly from their right ear. Do participants report more words from one ear than the other? Use a Mann-Whitney-Wilcoxon Rank Sum Test to answer this question.

Number of words reported:

Participant	Left ear	Right ear
1	25	32
2	29	30
3	10	7
4	31	36
5	27	20
6	24	32
7	27	26
8	29	33
9	30	32
10	32	32
11	20	30
12	5	32

7. Suppose the sales manager of a company wishes to investigate how sales performance, y, depends on five independent variables:
x_1 = number of months the sales representative has been employed by the company
x_2 = sales of the company's product and competing products in the sales territory
x_3= dollar advertising expenditure in the territory
x_4= weighted average of the company's market share in the territory for the previous four years
x_5 = change in the company's market share in the territory over the previous four years

A random sample of 26 observations shows the following results:

	Coefficients	Standard Error
Intercept	- 1113.00	420.00
x_1	3.60	1.20
x_2	0.04	0.01
x_3	0.13	0.04
x_4	256.95	175.20
x_5	324.50	200.00

Partial ANOVA Table

Source	SS (in 1000)
Regression	39,500
Residual	3,500
Total	43,000

a. Test the overall significance (i.e., validity) of the multiple regression model using a 5% significance level.
b. Perform statistical tests to identify the independent variables that are significant in predicting sales performance.
c. Construct a 98% confidence interval for the intercept.
d. Compute adjusted R squared.
e. If a simple regression analysis is performed with x_1 as the only independent variable, the *Sum of Squares for Error* (*SSE*) is three times as large as the *SSE* in the above multiple regression model with 5 independent variables. Test if this simple regression model is significant or not.

8. **(R)** The dataset "uswages[1]" is drawn as a sample from the Current Population Survey in 1988. Predict the wage from the years of education.

    a. Make a plot of the two variables of interest that makes some effort to avoid the problems of overplotting. Repeat the plot but use a log scale for the response.

    b. Compute the default smoothing spline fit and display it on top of the data. Comment on the quality of the fit.

    c. Compute the default lowess fit and display it on the fit. Does this method work better than smoothing splines in this instance?

    d. For each number of years of education, compute both the mean and the median wage. Construct a plot showing how these means and medians change with education. Which summary works better?

    e. Instead of means and medians, compute the two quartiles and the median and display them on top of the data. (This is a form of quantile regression).

    f. Display the lowess fit on the log-transformed data. Do you think it is better to work on the log scale for this data?

9. **(R)** The dataset prostate[2] is from a study of 97 men with prostate cancer who were due to receive a radical prostatectomy. Predict the lweight using the age.

    a. Plot the data and comment on the relationship.

    b. Fit a curve using kernel methods, plotting the fit on top of the data. What is the effect of the outlier?

---

[1] https://rpubs.com/sadjei65320/754802

[2] https://rpubs.com/RobbyS/614744

c. Compute the smoothing spline fit with the default amount of smoothing. What type of curve has been fit to the data?

d. Fit a loess curve with a 95% confidence band. Do you think a linear fit is plausible for this data?

e. Display all three previous fits on top of the same display and compare.

f. Introduce lpsa as a second predictor and show the bivariate fit to the data using smoothing splines.

g. Plot the residuals and interpret them.