# Homework 6

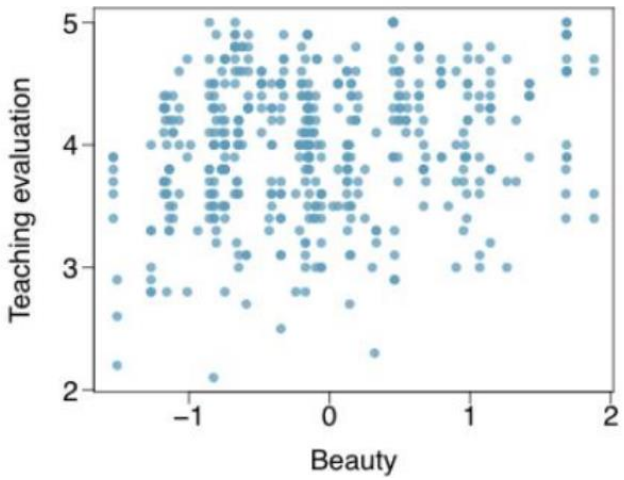
Statistical Inference, Fall 1401

**1.** Show that the likelihood ratio statistic for the multiple regression model is a monotonic function of the model F-statistic for the multiple regression model.

**2.** What is the AIC statistic (smaller is better) for the unrestricted model regression analysis? For the restricted model? When does the restricted model have a lower AIC statistic?

**3.** 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.

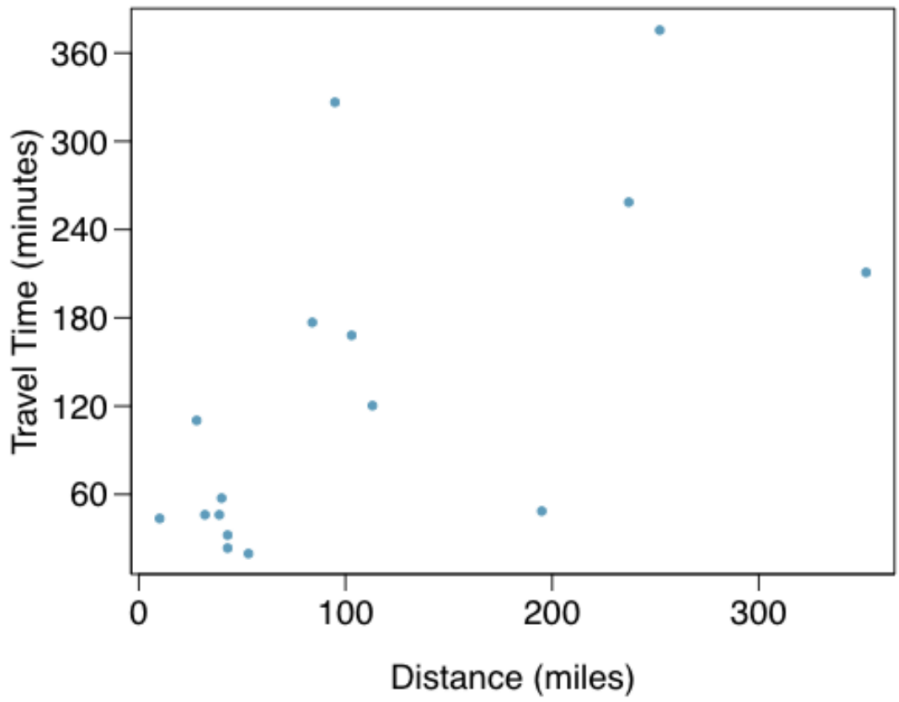
| | Estimate | Std. Error | T value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.010 | 0.0255 | 157.21 | 0.0000 |
| beauty | _____ | 0.0322 | 4.13 | 0.0000 |




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.

**4.** 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.

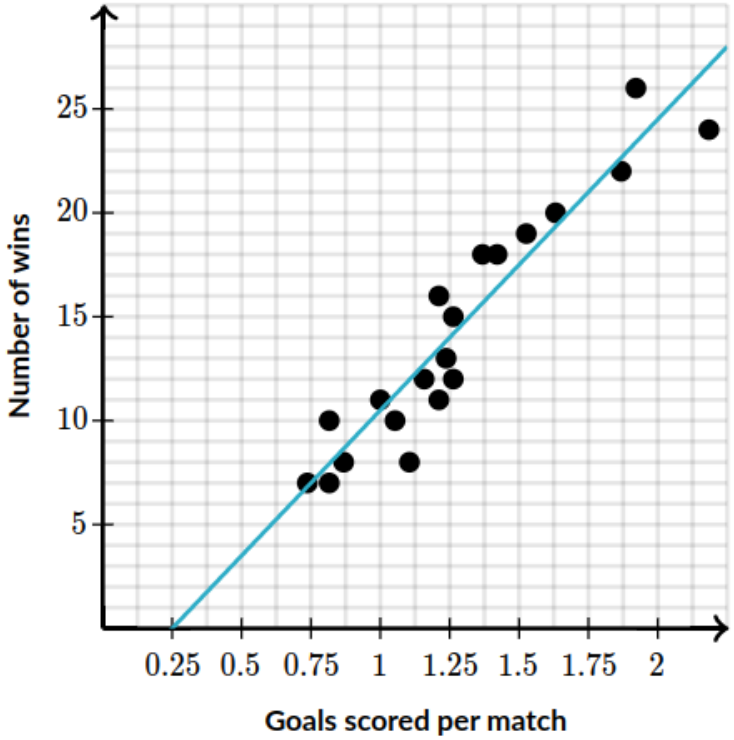
e. It actually takes the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

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?

**5.** 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.

# Homework 6

Statistical Inference, Fall 1401




**6.** 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.

# Homework 6

Statistical Inference, Fall 1401

**7.** **(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?

**8.** **(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.