STAT 210P

Homework 2

Round to the nearest 1st decimal place (nearest tenth) when possible.

1. Say a regression equation if fit to data, and the correlation coefficient estimate ,R ,between X and Y is 0.5.
State if true or false.

a. The slope of the regression line is 0.5.

b. The regression model with X explains 50% of the variation in Y,

c. 25% of the variation in Y is explained by X.

d. 50% of the variation in Y is explained by X.

e. X is positively associated with Y.

f. Even if X and Y have a non linear relationship, the value of 0.5 can be used to measure the association between X and Y.

2. Suppose a simple linear model is fit to predict $Y=$ weight in kilograms using $X=$ height in centimeters of an adult. But say a new simple linear model is fit using $Y=$ height in centimeters and $X=$ weight in kilograms (that is to say Y and X have reversed).
State whether each of the following would be the same for this new model as it was for the original model, or it would be different and explain in a sentence or two.

a. The value of R.

b. The value of R^2.

c. The estimate of β_1.

d. The estimate of β_0.

e. The test statistic to test if the correlation between explanatory and response is equal to 0.

3. This question will use the MidwestSales data set that is on the class website.
Run the following lines once you import the dataset to give names to the variables. Say you named the dataset "Midwest":

```
names(Midwest)=c("id","price","sqft","bed","bath","ac","garage","pool","year","quality",
"style","lot","hwy")
```

a. Fit a linear model where the square footage of the house is used to predict the sale price. (X=sqft and Y=price). Write the estimated regression equation.

b. Interpret the estimate of the slope.

c. Test whether square footage has a significant linear relationship with price. Use $\alpha = 0.05$ level of significance. Write the null and alternative hypothesis, state the test statistic and p-value, and make a conclusion.

d. Now test whether square footage has a significant positive linear relationship with price. Use $\alpha = 0.05$ level of significance. Write the null and alternative hypothesis, state the test statistic and p-value, and make a conclusion.

e. Find and interpret a 95% confidence interval for the mean price when sqft=2000.

f. Find and interpret a 95% prediction interval for the price when sqft=2000.

g. What would happen to the interval from part f. if the confidence level is decreased to 90%? Explain in a sentence or two.

h. Would it make sense to predict the sale price of a house that is 8500 square feet?

4. This question will use the skin cancer data set that is on the class website. Say it is of interest if latitude is predictive of mortality rate due to skin cancer. Fit a simple linear model where X=Lat and Y=Mort.

a. Using a $\alpha = 0.05$ significance level, conduct a formal statistical test of whether latitude has a linear association with the mortality.

b. Find and interpret a 99% confidence interval for the mean mortality rate when Lat=40.

c. Find and interpret a 99% prediction interval for the mortality rate when Lat=40.

d. What can you say about the center of the confidence interval and prediction interval. Is it the same? Explain in a sentence or two why or why not it would be the same.

e. How does the width of the confidence interval compare to the prediction interval. Explain in a sentence or two.

5. 1 point each. Remember the form of the prediction interval. The standard error of the prediction at a value of $X = X_p$ was $SE(\hat{Y}_p) = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right)}$

For each part, state what will happen to the prediction interval (stay the same, be wider, or be narrower).

a. The sample size is increased.

b. If X_p gets closer to $\bar{X}$, the average of the X covariate.

c. If the variability of the response variable decreases.

d. The average of the response is increased.