STA 321 Professor Gabrosek Fall 2020 Homework #1

Directions: This homework is the culminating experience for Learning Module 1: Simple Linear Regression. This homework is to be completed individually. While you may discuss problems with each other, the work you turn in must be your own. You will turn in this cover page with your assignment. Note: Any problem that begins CH Problem is a problem from the textbook.

By signing your name below, you verify that the work done on this homework is your own.

Print Name _____Monica Klosin ☺_____

1. Skip
2. CHP Problem 1.3, pages 23-24

   In each of the following sets of variables, identify which of the variables can be regarded as a response and which can be used as predictors (explain)?

   A response variable is the variable that is what the question is focused on. A predictor variable is what explains the changes in the response variable.

   a) Number of cylinders and gasoline consumption of cars
      response: gasoline consumption of cars
      predictors: cylinders
      why: Cylinders are where gasoline is burned and turned into power. The larger cylinder a car has, will then determine how much gas a car can hold and therefor consume.

   b) SAT scores, GPAs, and college admission
      response: college admission
      predictors: SAT scores and GPAs
      why: College admission is based on the scores of the SAT and the GPA a student has. The higher an SAT and GPA, the higher chance a student has in being admitted to college.

   c) Supply and Demand of certain goods
      response: Supply
      predictors: Demand
      why: Based on how much demand of an item there is, the supply will either increase or decrease.

   d) Company's assets, return on a stock, and net sales
      response: return on stock
      predictors: company's assets and net sales
      why: based on a company's assets and net sales, the stock of the company will either rise or fall. That is why the stock is the response variable, and the predictors is the company's assets and net sales.

e) The distance of a race, the time to run race, and weather conditions
response: time to run race
predictors: distance of race and weather conditions
why: The time to run a race depends on how long the race is, as well as if it is sunny dry weather, or rainy slippery weather.

f) The weight of a person, whether or not the person is a smoker, and whether or not the person has a lung cancer
response: whether the person has lung cancer
predictors : whether or not the person is a smoker or the weight of the person
why: The chance of a person having lung cancer is determined on if they smoke or if they are obese.

g) The height and weight of a child, their parent's height and weight, and the gender and age of the child
response: the height and weight of child
predictors: the gender and age of child or their parent's height and weight
why: The height and weight of a child can be determined by their age and gender (due to genetics) or their parent's height and weight (also genetic for height, and weight based on living conditions and the diet they are on, since their diet is most likely the diet their parents are on).

3. CH Problem 1.4, page 24 – Do only part (a) and number the answers as parts (a) to (g). For example, I've done the first one for you. Note that qualitative and categorical are the same thing.

Classify each variable as either quantitative or qualitative.

(a) Quantitative = gas consumption; number of cylinders
(b) Quantitative = SAT scores; GPAs; admission rate (I am viewing this problem as if this is a dashboard on college board for a college, and the college stats list the average SAT score being 1290, average GPA being 3.8, and the admission rate being 83% acceptance)
(c) Quantitative = Supply of goods; demand of goods
(d) Quantitative = net sales, return on stock, company's assets
(e) Quantitative = distance of race; time to run race
Categorical = weather conditions
(f) Quantitative = weight of person
Categorical = person is a smoker; person has lung cancer
(g) Quantitative = height/weight of child; height/weight of parents; age of child
Categorical = gender of child

Software Investigation

In this software investigation, we will use a data set from the textbook that investigates the following research question:

Do short people tend to marry short people and do tall people tend to marry tall people?

There are 96 observations in the data set. The two variables are:

- Wife – height of wife in cm
- Husband – height of husband in cm

4. Find numerical summaries for the variables Husband and Wife. **See LM 1 – Basic Numerical Summaries on 321 R How-to.**

   (a) Use the snipping tool to copy the two tables produced and paste them here.

```
favstats(~Husband, data= Heights, na.rm=TRUE)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
152	166.75	175.5	182.25	192	174.3229	9.960443	96	0

1 row

```
favstats(~Wife, data= Heights, na.rm=TRUE)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
141	158	164.5	170.25	181	163.8958	9.128877	96	0

1 row

   (b) What is the height of the shortest wife in the data set? Include the units.

   Shortest wife: 141cm

   (c) What is the height **in inches** of the tallest husband in the data set? Include the units.

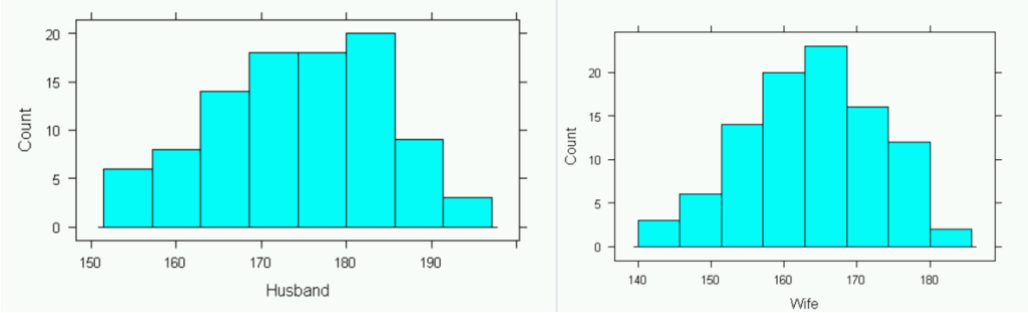
   Tallest husband: 75.5906 inches

   (d) What is it about the mean, median, and standard deviation that suggest the distributions of the heights for each sex are roughly symmetric?

   Both medians are slightly greater than their means, which indicates a slight left skew. Since the sd is around 9, this indicates a similar spread of data.

5. Make histograms of the variables Husband and Wife. **See LM 1 – Histogram on 321 R How-to.**

(a) Use the snipping tool to copy the two histograms and paste them here.



(b) Describe the distribution of heights of wives. Mention peaks, shape, and outliers.

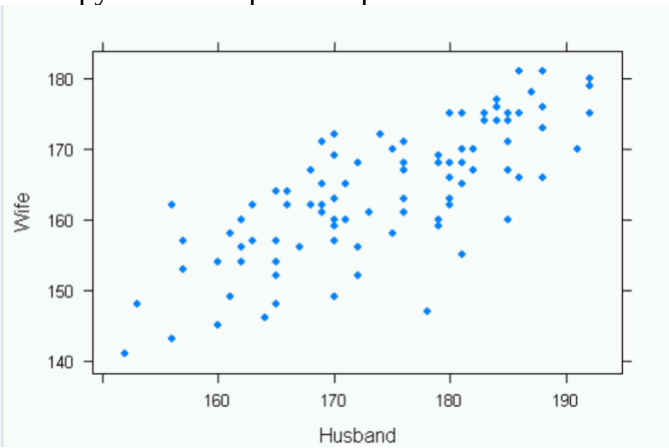
Looks unimodal, got a bell shape, doesn't seem that there are any outliers.

(c) Describe the distribution of heights of husbands. Mention peaks, shape, and outliers.

Looks unimodal, very slightly left skewed due to the high concentration of husbands with a height of ~170, but not enough to call this a left skewed graph. Graph has a bell shape, doesn't seem that there are any outliers.

6. Make a scatterplot with X = Husband and Y = Wife (response). **See LM 1 – Scatterplot on 321 R How-to.**

(a) Use the snipping tool to copy the scatterplot and paste it here.



(b) Is there a linear relationship between wife height and husband height? If yes, describe the strength and direction of the linear relationship.

Yes, just by looking at the graph, you see that there is a positive linear relationship between wife height and husband height. After typing in R:

```
cor(Heights$Wife, Heights$Husband)
```

The output is 0.7633864, confirming that there is a moderate positive linear relationship between wife height and husband height.

7. Find the linear correlation between X = Husband and Y = Wife. **See LM 1 – Linear Correlation on 321 R How-to.**

   (a) What is the value of r?

```
```{r, Q7}
# find correlation coefficient
cat("r = ", cor(Wife~Husband, data = Heights))
```
```

```
r =  0.7633864
```

   (b) Describe what the linear correlation tells you about the linear relationship between the variables.

The linear correlation of r = 0.7633864 indicates that there is a moderate positive relationship between wife height and husband height.

8. Based on what you have done in Questions 6 and 7, answer the research question posed in this homework.

The research question posed in this homework is *Do short people tend to marry short people and do tall people tend to marry tall people?*

Based on the visuals of the scatterplot and the moderate positive linear correlation, it seems that the question's answer is yes. Short people tend to marry short people and tall people tend to marry tall people.

9. Since there is a linear relationship between X = Husband and Y = Wife, it is appropriate to fit the simple linear regression model. What is the model? Note: I want the model, not the regression equation. Be sure to tell me what the X and Y variables are.

The Simple Linear Regression (SLR) model is: $Y = \beta_0 + \beta_1 X + \varepsilon$.

The response variable is denoted by Y, which is Wife Height. The predictor variable denoted by X is Husband Height. β_0 indicates the y-intercept of the line that predicts the Height of Wife's. β_1 is the slope of the line that predicts the Height of Husbands. ε is the actual – predicated value for Husband Height.

10. Fit the regression model and save the results to an R Object named Model. **See LM 1 – Linear Regression on 321 R How-to.**

(a) Use the summary( ) function to get the output from Model. Use the snipping tool to copy and paste that output below. Start at the coefficients table.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.93015   10.66162   3.933 0.000161 ***
Husband      0.69965    0.06106  11.458  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.928 on 94 degrees of freedom
Multiple R-squared:  0.5828,    Adjusted R-squared:  0.5783
F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16


> anova(RObject)
Analysis of Variance Table

Response: Wife
          Df Sum Sq Mean Sq F value    Pr(>F)
Husband    1 4613.7  4613.7  131.29 < 2.2e-16 ***
Residuals 94 3303.3    35.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Write out the least-squares regression equation.

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 X$$
$$\hat{Y} = 41.9015 + 0.69965X$$

(c) Write a one sentence interpretation of the slope in the context of the problem.

The slope is 0.69965, which tells us that as the height of husbands goes up by 1cm, the height of wife goes up by 0.69965 cm.

(d) What proportion of the variation in wife height is explained by the least-squares line?

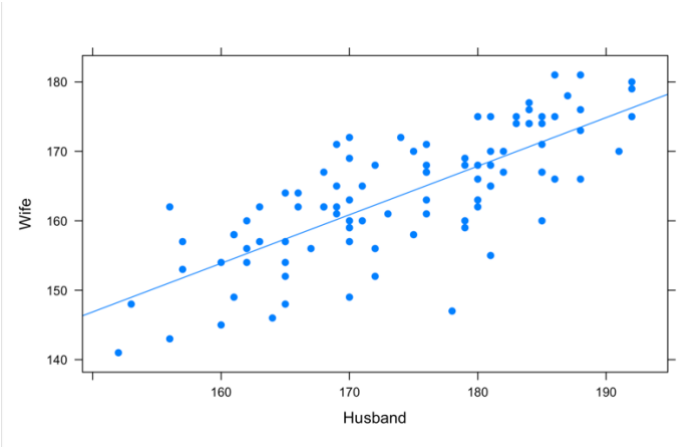
The proportion is 0.5828, so 58% of the variation in wife height is explained by the least-squares line.

(e) Determine whether there is a statistically significant positive linear relationship between wife height and husband height. Do the following: (i) write out Ho and Ha in notation, (ii) state the value of the test statistic, (iii) state the p-value, (iv) decide whether to reject Ho, and (v) write a couple sentences interpreting your decision referencing the research question.

i. $H_0 : \hat{\beta}_o = 0$
$H_1 : \hat{\beta}_1 \neq 0$

ii. test statistic = 11.458
iii. p-value = < 2.2e-16
iv. We reject H_o because $p < 0.05$.
v. The population slope is not equal to 0, indicating that as X changes Y changes. There is some non-0 linear relationship between Wife Height and Husband Height, which we have observed in earlier questions. This also makes sense in terms of the question, since it would be very hard to find data where all Wife's have all different heights and all their Husbands are the same height.

11. Make a scatterplot with X = Husband and Y = Wife that includes the regression line. Use the snipping tool to copy and paste that output below. **See LM 1 – Scatterplot with Regression Line on 321 R How-to.**



12. Find a 99% confidence interval for the slope. **See LM 1 – Confidence Interval on Slope on 321 R How-to.**

(a) What are the lower limit and the upper limit of the C.I.?

```
> confint(RObject, "Husband", level=0.99)
              0.5 %    99.5 %
Husband 0.5391137 0.8601938
```

(b) Write a one sentence interpretation of the C.I.

We are 99% confident that for the increase in 1cm of Height for Husband, the height of a Wife will go up between 0.53 and 0.86 cm.

13. Do the following intervals: **See LM 1 – Intervals on Y Given X on 321 R How-to.**
    - Find a 97% confidence interval for the mean wife height for husbands that are six feet (183 cm) tall.
    - Find a 97% prediction interval for the height of a wife whose husband is six feet tall.
    - Create a data frame with the results.
    - Subset Observations to only include where husband height is 183. **See Utilities – Subset Observations on 321 R How-to.**
    - Print Subset. Note: You will get a separate row of output for each row in the dataset where the value Husband = 183. That's okay.

(a) the snipping tool to copy the output and paste it here.

	Heights.Husband <dbl>	fit <dbl>	lwr <dbl>	upr <dbl>	fit.1 <dbl>	lwr.1 <dbl>	upr.1 <dbl>
18	183	169.9668	168.1946	171.7389	169.9668	156.7846	183.149
48	183	169.9668	168.1946	171.7389	169.9668	156.7846	183.149

2 rows

(b) What is the point estimate for the two intervals?

169.9668 is the point estimate for the two intervals.

(c) By hand, show how the point estimate was calculated.

$$\hat{Y} = 41.9015 + 0.69965X$$
$$\hat{Y} = 41.9015 + 0.69965(183)$$
$$\hat{Y} = 41.9015 + 128.03595$$
$$\hat{Y} = 169.9668$$

(d) Write a sentence that interprets the C.I. for the mean response.

We are 95% confident that the mean height of all wife's with husbands with a height of 183 cm is between 168.1946 cm and 171.7389 cm.

(e) Write a sentence that interprets the P.I. for an individual response.

We are 95% confident that the height of a wife with a husband whos height is 183 cm is between 156.7846 cm and 183.149 cm.

(f) Using the context of the data set, explain why the C.I. in part (d) is so much narrower than the P.I. in part (e).

A C.I. tells you about the estimation of a population, while a P.I. tells you about one specific case. It is harder to estimate the value for one specific case vs for a whole population, which is why the P.I is much wider than the C.I. .