Stat 217 Homework 4
Due: Friday, October 24th, beginning of class

1. Write the letter of the scatterplot that corresponds to each correlation coefficient below.

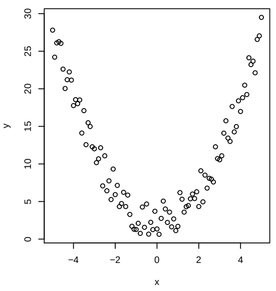
   r = 0.85 ______________________

   r=-0.85 ______________________

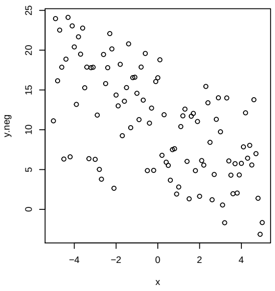
   r=0 ______________________

   r=-0.2 ______________________

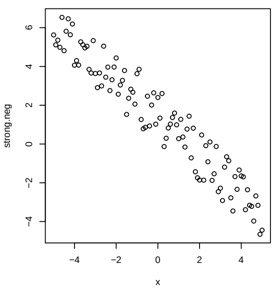
Plot A



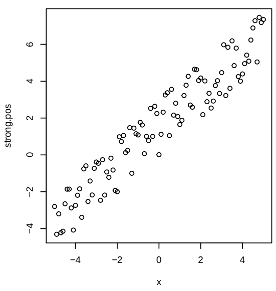
Plot B



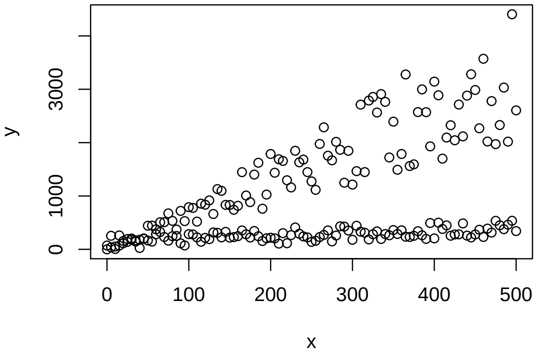
Plot C



Plot D



2. Page 167 of the textbook provides a list of some general things to look for in scatterplots. What items from the list do you see in the following scatterplot?



3. Suppose you toss a ball in the air, and you make a plot showing the height of the ball from time $t = 0$ to the time when the ball hits the ground. What is wrong with the following statement: “There is a strong correlation between time and height of the ball”.

   A. Correlation does not imply causation

   B. We generally use the word correlation to describe a linear relationship between two variables, and the relationship between time and height is curved

   C. You should not use correlation to describe a relationship between two quantitative variables

   D. There is nothing wrong with this statement

4. Which of the following statements about correlation are true?

   I Correlation is a value between 0 and 1.
   II Correlation measures the strength and direction of a linear relationship.
   III Correlation is not resistant to outliers.

   A. I only
   B. II only
   C. III only
   D. I and III only
   E. I and II only
   F. II and III only
   G. All of the above are true

5. Which of the following pairs of variables would you expect to have a correlation near 1?

   A. Size of car engine (liters) and top speed of a car.
   B. Age of adult women and likelihood of getting pregnant
   C. Type of pet and size of yard
   D. Ounces of water drank in a day and score on a statistics exam that evening

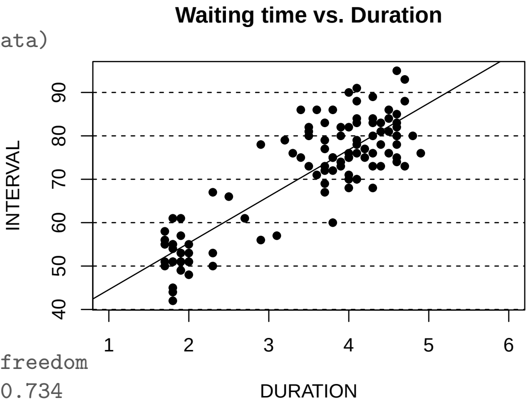
6. (Old Faithful): Old Faithful Geyser in Yellowstone National Park derives its names and fame from the regularity (and beauty) of its eruptions. Rangers usually post the predicted times of eruptions for visitors. R. A. Hutchinson, a park geologist, collected measurements of the eruption durations (in minutes) and the subsequent time intervals before the next eruption (in minutes) over an 8-day period. Help rangers use the data to explain the relationship between duration and subsequent time to the next eruption.

```
Call:
lm(formula = INTERVAL ~ DURATION, data = faith.data)

Residuals:
   Min     1Q Median     3Q    Max
-14.64  -4.44  -1.09   4.47  15.65

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.828      2.262    15.0   <2e-16
DURATION      10.741      0.626    17.1   <2e-16

Residual standard error: 6.68 on 105 degrees of freedom
Multiple R-squared:  0.737,Adjusted R-squared:  0.734
F-statistic:  294 on 1 and 105 DF,  p-value: <2e-16
```



(a) A simple linear regression model was fit, and the ouput is shown above. What is the estimated regression equation?

A. $\widehat{duration} = 10.74 + 33.83 interval_i$

B. $\widehat{duration} = 33.83 + 10.74 interval_i$

C. $\widehat{interval} = 33.83 + 10.74 duration_i$

D. $\widehat{interval} = 10.74 + 33.83 duration_i$

E. $\widehat{duration} = \beta_0 + \beta_1 interval_i$

F. $\widehat{interval} = \beta_0 + \beta_1 duration_i$

(b) Interpret the estimate of the y-intercept.

A. After an eruption that lasts 0 minutes, the true mean waiting time to the next eruption is estimated to be 33.83 minutes.

B. The estimated change in the interval for a one minute increase in duration.

C. After an eruption that lasts 1 minute, the true mean waiting time to the next eruption is estimated to be 45 minutes.

D. For a duration of 0, the interval is estimated to be 2.26 minutes.

(c) Interpret the estimate of the slope.

A. The estimated interval at a duration of 0.

B. For a one second increase in eruption duration, the true mean waiting time to the next eruption is estimated to increase by 10.74 seconds.

C. For a one minute increase in eruption duration, the true mean waiting time to the next eruption is estimated to increase by 10.74 minutes.

D. For a one minute increase in duration, the change in the interval is estimated to be 0.626 minutes.

(d) What symbol do we use to describe the slope in the population? What symbol do we use to describe the y-intercept in the population?

A. slope=β_0; y-intercept = β_1

B. slope=β_1; y-intercept = β_0

C. slope=b_0; y-intercept = b_1

D. slope=b_1; y-intercept = b_0

E. slope=μ_1; y-intercept = μ_2

F. slope=ρ; y-intercept = r

(e) Why are b_0 and b_1 called *least squares estimates*?

A. b_0 and b_1 are found by minimizing the sum of the squared residuals.

B. b_0 and b_1 are found by minimizing the type I error rates of the t-tests in the coefficient table.

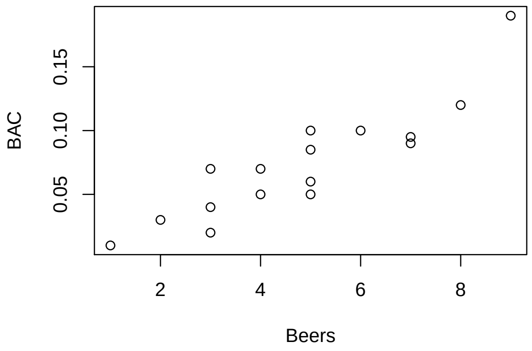
C. b_0 and b_1 are the estimates for the slope and y-intercept, and they are square numbers.

D. King Charles II made up this name.

7. Below, the Beers vs. BAC data are shown. The scatterplot and the output of the regression model are also shown. Find the sum of the squared residuals for this example. Recall: each residual can be found with the equation $e_i = y_i - \hat{y}$, and $\hat{y}$ can be found from the estimated regression equation. The sum of the squared residuals is then $\sum_{i=1}^{16} e_i^2$.

```
   Beers   BAC
1      5 0.100
2      2 0.030
3      9 0.190
4      8 0.120
5      3 0.040
6      7 0.095
7      3 0.070
8      5 0.060
9      3 0.020
10     5 0.050
11     4 0.070
12     6 0.100
13     5 0.085
14     7 0.090
15     1 0.010
16     4 0.050
```

```
with(BB, plot(Beers, BAC))
```



```
lm.beer <- lm(BAC~Beers, data=BB)
summary(lm.beer)



Call:
lm(formula = BAC ~ Beers, data = BB)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.02712 -0.01735  0.00177  0.00862  0.04103 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0127     0.0126   -1.00     0.33
Beers         0.0180     0.0024    7.48    3e-06

Residual standard error: 0.0204 on 14 degrees of freedom
Multiple R-squared:   0.8,Adjusted R-squared:  0.786 
F-statistic: 55.9 on 1 and 14 DF,  p-value: 2.97e-06
```