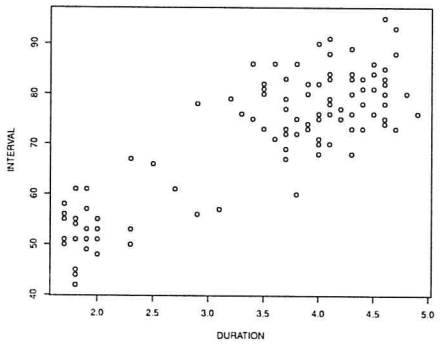
KEY

STAT 217: SLR Homework
Due Wednesday, April 1 in class

You may work in a group or individually. If you work in a group, turn in one paper for your group.

1. (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.




```
lm.out <- lm(INTERVAL~DURATION, data=faith.data)
summary(lm.out)
```

```
##
## 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
```

```
confint(lm.out)
```

```
##             2.5 % 97.5 %
## (Intercept)  29.3   38.3
## DURATION      9.5   12.0
```

(a) What is the estimated regression model?

A. $\widehat{duration}_i = 33.8 + 10.7 interval_i$

B. $duration_i = \beta_0 + \beta_1 interval_i + \epsilon_i$

C. $\widehat{interval}_i = 33.8 + 10.7 duration_i$

D. $interval_i = \beta_0 + \beta_1 duration_i + \epsilon_i$

(b) What is the true regression model?

A. $\widehat{duration}_i = 33.8 + 10.7 interval_i$

B. $duration_i = \beta_0 + \beta_1 interval_i + \epsilon_i$

C. $\widehat{interval}_i = 33.8 + 10.7 duration_i$

D. $interval_i = \beta_0 + \beta_1 duration_i + \epsilon_i$

(c) What is the estimated mean waiting time to the next eruption after an eruption lasting 4 minutes?

$33.8 + 10.7(4) = 76.6$

(d) Using the output above, interpret the slope coefficient estimate.

A. For a one hour increase in eruption duration, the true mean waiting time to the next eruption is estimated to increase by 10.7 minutes, with a 95% confidence interval from 9.49 to 11.98 minutes.

B. For a one minute increase in eruption duration, the true mean waiting time to the next eruption is estimated to increase by 10.7 minutes, with a 95% confidence interval from 4.31 to 14.98 minutes.

C. After an eruption that lasts 0 minutes, the true mean waiting time to the next eruption is estimated to be 33.8 minutes.

D. For a one minute increase in eruption duration, the true mean waiting time to the next eruption is estimated to increase by 10.7 minutes, with a 95% confidence interval from 9.49 to 11.98 minutes.

(e) Show how you would calculate confidence intervals for β_0 and β_1 if you weren't given the intervals in the output. Use $t^* = 1.98$.

for β_1 $10.741 \pm 1.98(0.626) = (9.502, 11.981)$

for β_0 $33.828 \pm 1.98(2.262) = (29.349, 38.307)$

(f) What proportion of the variability in the waiting time between eruptions is explained by the duration of the previous eruptions?

A. 0.737

B. 0.734

C. 10.74

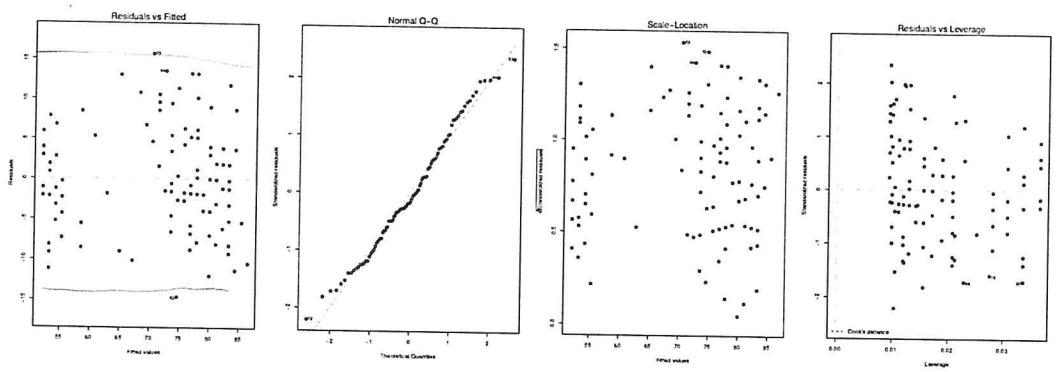
D. 6.68

(g) The estimates for the y-intercept and the slope coefficient shown in the model output above are known as least squares estimates because...

A. they are calculated by minimizing the sum of squared residuals
B. they are unbiased
C. they are estimates of variance parameters
D. they are calculated by R

(h) What assumptions should you check before trusting the results shown in the simple linear regression model output? Circle all that apply.

A. Independence
B. Constant Variance
C. Quantitative Variables
D. Categorical Variables
E. No Influential Points
F. Normality
G. All Expected Cell Counts are greater than 5
H. Linearity
I. Homogeneity Assumption



(i) In one sentence, describe how you check the following assumptions using the above plots. Then state whether the assumption is met or not (yes or no).

- Constant Variance

Check for approximately equal spread of residuals across fitted values. in Residuals vs Fitted Values plot

Yes-Met

- No Influential Points

Check for points with large Cook's Distance in the Residuals vs Leverage plot

Yes-Met

- Linearity

Make sure no curvature is present in the resids vs. fitted values plot

Yes-Met

(j) What hypotheses are being tested in the DURATION line of the model output?

A. $H_0 : \beta_0 = 0$ vs. $H_A : \beta_0 \neq 0$

B. $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

C. $H_0 : b_0 = 0$ vs. $H_A : b_0 \neq 0$

D. $H_0 : b_1 = 0$ vs. $H_A : b_1 \neq 0$

(k) What is your conclusion for the hypotheses being tested in the DURATION line of the model output?

A. There is strong evidence of a linear relationship between eruption duration and waiting time to the next eruption in Old Faithful eruptions (p-value< 0.0001 from t-stat=15.0 on 105 df).

B. There is strong evidence of a linear relationship between eruption duration and waiting time to the next eruption in Old Faithful eruptions (p-value< 0.0001 from t-stat=17.1 on 105 df).

C. There is strong evidence of a linear relationship between eruption duration and waiting time to the next eruption in Old Faithful eruptions (p-value< 0.0001 from F-stat=294 on 1 and 105 df).

D. There is very weak evidence of a linear relationship between eruption duration and waiting time to the next eruption in Old Faithful eruptions (p-value< 0.0001 from t-stat=17.1 on 105 df).

(l) What hypotheses are being tested in the (Intercept) line of the model output?

A. $H_0 : \beta_0 = 0$ vs. $H_A : \beta_0 \neq 0$

B. $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$

C. $H_0 : b_0 = 0$ vs. $H_A : b_0 \neq 0$

D. $H_0 : b_1 = 0$ vs. $H_A : b_1 \neq 0$

(m) What is your conclusion for the hypotheses being tested in the (Intercept) line of the model output?

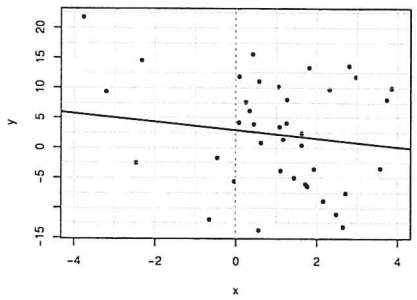
A. There is strong evidence of a linear relationship between eruption duration and waiting time to the next eruption in Old Faithful eruptions (p-value< 0.0001 from t-stat=15.0 on 105 df).

B. There is strong evidence of a linear relationship between eruption duration and waiting time to the next eruption in Old Faithful eruptions (p-value< 0.0001 from t-stat=17.1 on 105 df).

C. There is strong evidence that the true mean waiting time to the next eruption after an eruption lasting 0 minutes is nonzero (p-value< 0.0001 from t-stat=15.0 on 105 df).

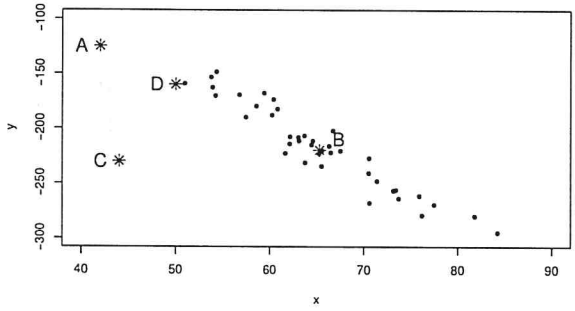
D. There is strong evidence that the true mean waiting time to the next eruption after an eruption lasting 0 minutes is nonzero (p-value< 0.0001 from t-stat=17.1 on 105 df).

2. Estimate the correlation coefficient, r, in the scatterplot.



A. -1
B. -0.32
C. 0.25
D. -0.9

3. In the scatterplot:



(a) Which point has the highest leverage?

A. A
B. B
C. C
D. D

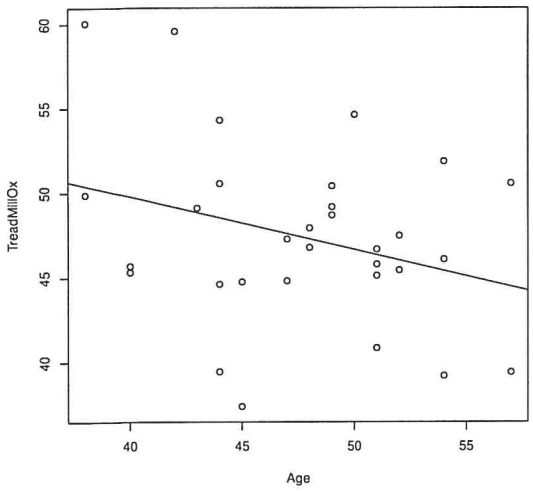
(b) What point has the highest influence?

A. A
B. B
C. C
D. D

(c) What point has the LOWEST influence?

A. A
B. B
C. C
D. D

4. In this problem, we revisit the treadmill dataset. The `TreadMillOx` variable is a measure of oxygen consumption in ml/kg/min on 32 subjects. We are interested in the relationship between oxygen consumption and age.



(a) Write out the true linear regression model for this example. Use greek letters and define parameters.

$Ox_i = \beta_0 + \beta_1 age_i + \varepsilon_i$

Ox_i: the oxygen value of the i^{th} subject
age_i: the age of the i^{th} subject
β_0: population y-intercept
β_1: population slope coefficient
ε_i: Residual error for the i^{th} observation

(b) We conduct a permutation test to test for a linear relationship between age and oxygen consumption. Write the null and alternative hypotheses for this test.

$H_0: \beta_1 = 0$
$H_A: \beta_1 \neq 0$

(c) The code below produces a permutation distribution of t-test statistics. In the code below, what does the `shuffle` function do?

A. shuffles a deck of cards

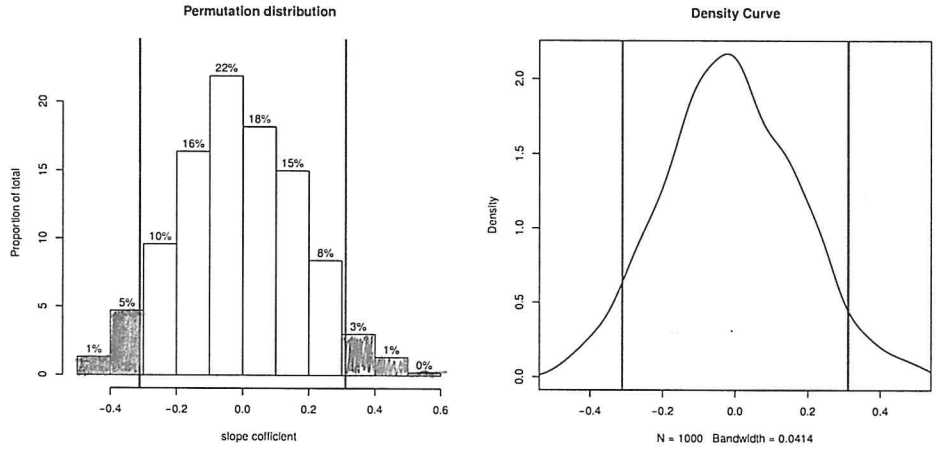
B. resamples the data with replacement

C. shuffles the order of the rows in the dataset

D. randomly reorders the subjects' ages in order to simulate no relationship between age and treadmill oxygen consumption

```
require(mosaic)
b1obs <- lm.tread$coef[2]

B <- 1000
slope <- matrix(NA, nrow=B)
for(b in (1:B)){
  slope[b]<-lm(TreadMill0x~shuffle(Age), data=treadmill)$coef[2]
}
```



p-value ≈ 0.10

(d) Above is the permutation distribution of slope coefficients simulated under the null hypothesis. In the original sample, the estimated slope coefficient is -0.3112. Shade in the area on the histogram and the density curve that represents the p-value and write an estimate for the p-value next to the plot.

(e) What is your decision at a significance level of 0.05?

FTR H_0

(f) Write your conclusion in the context of the problem.

There is weak evidence of a linear relationship between age and oxygen consumption in the population (p-value ≈ 0.10 from permutation test).