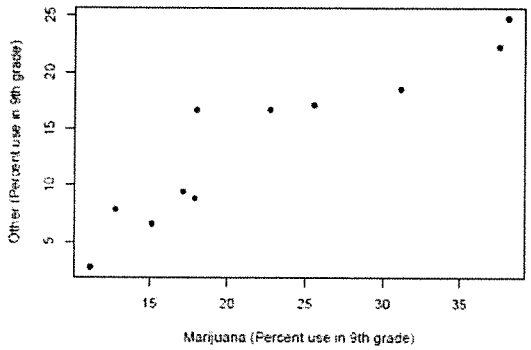
Key

Stat 217 Homework 7
Due Friday, October 30th

1. The *European School Study Project on Alcohol and Other Drugs* published an investigation of the usage of marijuana and other drugs. Data from 11 countries were collected, including the percentage of ninth graders who reported smoking marijuana and who have used other drugs (cocaine, amphetamines, etc.). We are interested in how the percentage of marijuana users effects the percentage of other drug users.

   a. A scatterplot of the percentage of other drug users versus the percentage of marijuana users is shown. What is the direction of the association?



   (A) Positive
   B) Negative
   C) No clear direction
   D) Not enough information

   b. What is the response variable?

   Other drug use (% use in 9th grade)

   c. What is the explanatory variable?

   Marijuana drug use (% use in 9th grade)

   d. The hypothesized regression model for examining this association is $Other_i = \beta_o + \beta_1 Marijuana_i + \epsilon_i$. How is β_o interpreted in the context of the problem?

   A) The proportion of the variability in the percentage of other drug usage that is explained by the regression model with marijuana usage as a predictor.
   B) The difference in the observed other drug usage and the predicted other drug usage.
   (C) The average other drug usage in countries with no marijuana usage.
   D) The average change in other drug usage associated with a 1% increase in marijuana usage.

   e. Below is the output for the simple linear regression model. Report the estimated regression equation.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.63435    2.23274  -0.732    0.483
maryJ        0.68606    0.09211   7.449  3.9e-05

Residual standard error: 2.764 on 9 degrees of freedom
Multiple R-squared:  0.8604,  Adjusted R-squared:  0.8449
F-statistic: 55.48 on 1 and 9 DF,  p-value: 3.898e-05
```

$\widehat{Other}_i = -1.635 + 0.686\, Marijuana_i$

What is the average percentage of other drug usage we would expect for a country that has 25% of ninth graders who have used marijuana?

$\widehat{Other} = -1.635 + 0.686(25)$

$= 15.515\%$

g. What is the proportion of the variability in other drug usage explained by the linear model with marijuana usage as a predictor?

0.8604

h. We decide to predict the percentage of other drug usage from a country with no marijuana usage. What is the potential problem with doing this?

Extrapolation! - Dangerous to use model to make predictions outside range of x's studied - country with lowest % marijuana use was 10%. - no countries saw 0% or fewer

i. Calculate a 95% confidence interval for the slope using $t^* = 2.27$.

$b_1 \pm t^* SE(b_1)$

$0.686 \pm 2.27(0.09211)$

$(0.477, 0.895)$

j. Interpret the interval you calculated above in the context of the problem.

For a 1% increase in the percentage of marijuana users, we are 95% confident that the true mean change in the percentage of other drug users lies between 0.477% and 0.895%.

k. Conduct a test for a linear relationship between the percent marijuana usage and percent other drug usage.

i. State the null and alternative hypotheses.

$H_0: \beta_1 = 0$ - There is no linear relationship between percent of marijuana usage and the percent of other drug usage in the pop.

$H_0: \beta_1 \neq 0$ - There is a linear relationship between the percent of marijuana usage and the percent of other drug usage in the pop.

ii. What is the test statistic?

$t = 7.449$

iii. What distribution does the test statistic follow under H_0?

t_9

iv. What is the p-value?

< 0.0001

v. State your conclusion in the context of the problem.

There is very strong evidence to suggest there is a linear relationship between the percent marijuana usage and percent of other drug usage in the pop.

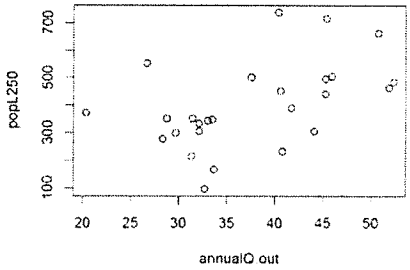
1. True or False. We can conclude that increases in marijuana usage causes an increase in other drug usage in these sampled countries since we had a small p-value for our slope.

2. This dataset is from a mark/recapture study done on the Kootenai River in Northwestern Montana. Each year scientists use mark/recapture methods to track the population of various trout species downstream of Libby Dam. We will explore the relationship between annual outflow from the dam (measured in hundreds of thousands of cubic feet of water) and rainbow trout population.

```
fish <- subset(recruitAbundLength.fwp, section == "FP")
fish2 <- fish[, c(6, 37)]
fish2$annualQ.out <- fish2$annualQ.out/100
cor(fish2)

##              popL250 annualQ.out
## popL250        1.000       0.488
## annualQ.out    0.488       1.000

plot(popL250 ~ annualQ.out, data = fish2)
```



a. Describe the relationship you see in the scatterplot.

moderately strong positive linear relationship
no obvious outliers or subgroups
relatively constant variance

b. What is the response variable?

rainbow trout population

c. What is the explanatory variable?

annual outflow from the Libby Dam

d. Use the output below to report the estimated regression equation.

```
fish.lm <- lm(popL250 ~ annualQ.out, data = fish2)
summary(fish.lm)

##
## Call:
## lm(formula = popL250 ~ annualQ.out, data = fish2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -259.8  -49.7  -16.7   30.2  311.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    59.95     127.34    0.47    0.642
## annualQ.out     9.08       3.31    2.74    0.011
##
## Residual standard error: 140 on 24 degrees of freedom
## Multiple R-squared:  0.238,Adjusted R-squared:  0.206
## F-statistic:  7.5 on 1 and 24 DF,  p-value: 0.0114
```

$\hat{pop_i} = 59.95 + 9.08 outflow_i$

e. Interpret the coefficients (b_o and b_1) in the context of the problem.

$b_0 = 59.95$: For 4 years with an annual outflow from the dam of 0 hundred thousand cubic feet of water, we expect the average rainbow trout population to be 59.95 fish.

$b_1 = 9.08$: For a 1 hundred thousand cubic feet of water increase in annual outflow from the dam, we expect the average number of rainbow trout to increase by 9.08 fish.

OR: For each additional 100,000 ft^3 of water in annual outflow from the dam, we expect the average number of rainbow trout to increase by 9.08 fish

f. Give the predicted rainbow trout population for a year in which 3,000,000 cubic feet of water were released from the dam. (Careful: remember how the outflow is measured in the data)

$\widehat{Pop}_i = 59.95 + 9.08(30)$

$= \boxed{332.35}$ fish

g. Give the R^2 for the trout model and interpret it in the context of the problem.

$R^2 = 0.238$

~23.8% of the variability in the rainbow trout population is explained by the Dams annual outflow.

h. Calculate a 95% confidence interval for the slope using $t^* = 2.8$.

$b_1 \pm t^* SE(b_1)$

$9.08 \pm 2.8(3.31)$

$(-0.188, 18.348)$

i. Interpret the interval you calculated above in the context of the problem.

For a 100,000 ft^3 of water increase in annual outflow from the dam, we are 95% confident that the true mean change in rainbow trout population lies between -0.188 fish and 18.348 fish.

j. Conduct a test for a linear relationship between annual outflow from the dam and rainbow trout population.

i. State the null and alternative hypotheses.

$H_0: \beta_1 = 0$ - There is no linear relationship between annual outflow from the dam and the rainbow trout population in the pop.

$H_a: \beta_1 \neq 0$ - There is a linear relationship between annual outflow from the dam and the rainbow trout population in the pop.

ii. What is the test statistic?

2.74

iii. What distribution does the test statistic follow under H_0?

t_{24}

iv. What is the p-value?

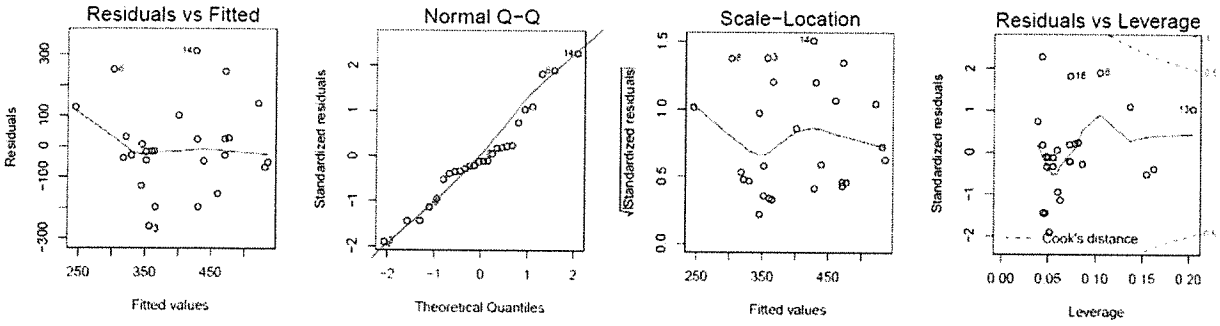
0.011

v. State your conclusion in the context of the problem. Include a scope of inference.

There is strong evidence to suggest there is a linear relationship between annual outflow from the dam and the rainbow trout population. This was not a random sample, so the results only apply to the fish in this study from the Kootenai River. There was not random assignment (can't assign annual outflow from the dam) so we cannot conclude the change in rainbow trout population is caused by the annual outflow from the dam.

k. Assess the assumptions providing justification:

```
par(mfrow = c(1, 4))
plot(fish.lm)
```



- Quantitative variables condition:

  annual outflow ✓ fish pop ✓

- Independent Observations:

  doesn't seem reasonable - one years outflow could definitely impact another years fish pop.

- Linearity of relationship:

  no curve in residuals vs fitted

  - assume met

- Equal (constant) variance:

  no obvious fanshape in residuals vs fitted

  - no increasing/decreasing variability

- Normality of the residuals:

  points lie on 1-1 line in Normal Q-Q plot

  - assume met

- No influential points:

  no influential points

  - all observations have Cooks D's < 0.5