PSTAT 126 - Assignment 2

Fall 2022

Due: Tuesday, October 11 at 11:59 pm on Canvas

Note: ***Submit both your `Rmd` and generated pdf file to Canvas.*** *Use the same indentation level as* ***Solution*** *markers to write your solutions. Improper indentation will break your document.*

```
library(alr4)
library(ggplot2)
data(UN11)
```

1. The data set `UN11` in the `alr4` package contains several variables, including `ppgdp`, the gross national product per person in U.S. dollars, and `fertility`, the birth rate per 1000 females, from the year 2009. The data are for 199 localities, and we will study the regression of `fertility` on `ppgdp`.

   (a) Identify the predictor and response.

   Solution:

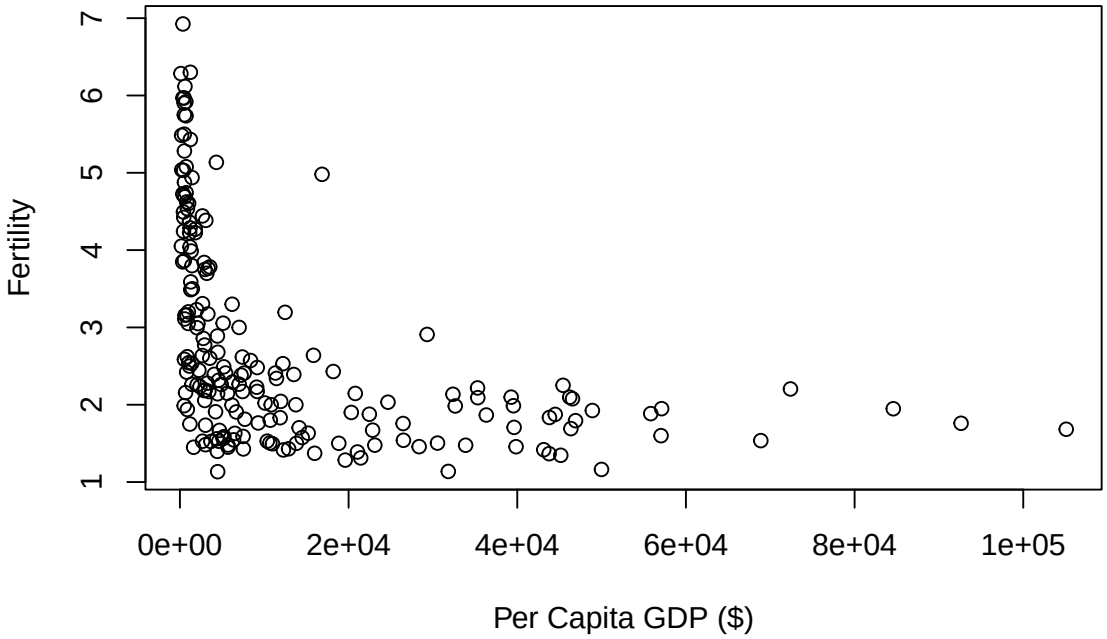
   ```
   fertility <- UN11$fertility
   ppgdp <- UN11$ppgdp
   ```

   The response is fertility, the predictor is ppgdp.

   (b) Draw the scatterplot of fertility against ppgdp and describe the relationship between these two variables. Is the trend linear?

   Solution:

   ```
   plot(ppgdp,fertility,
   xlab = 'Per Capita GDP ($)',
   ylab = 'Fertility')
   ```

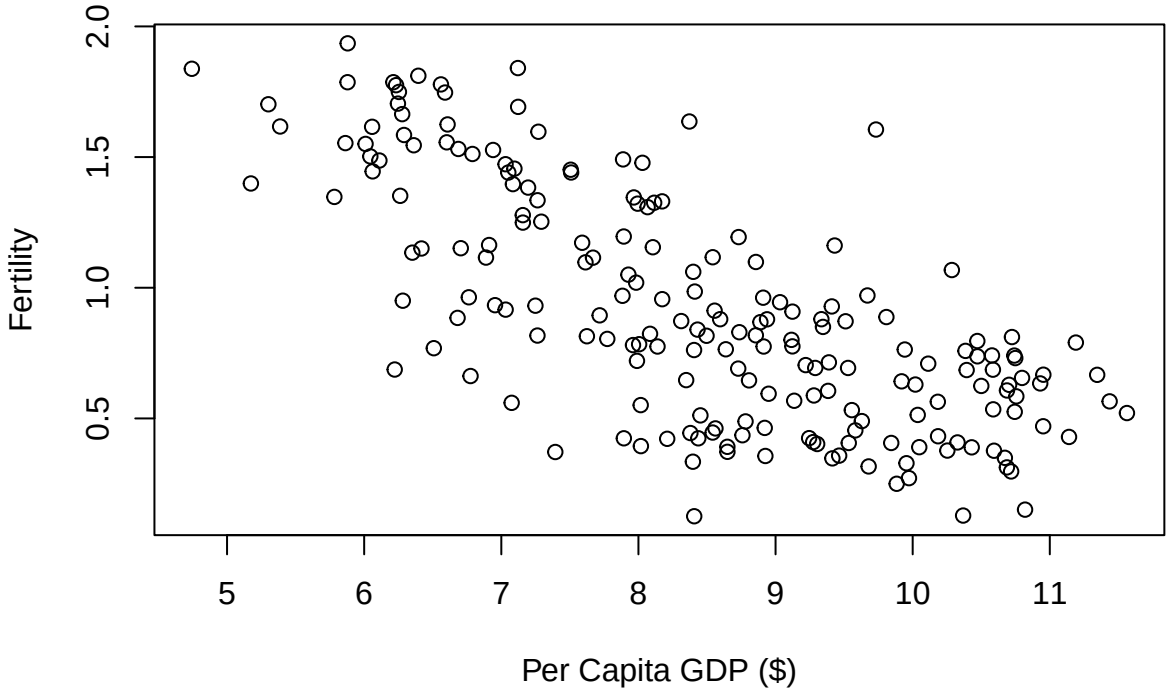


The trend is NOT linear, we can see that there is a decreasing relationship between ppgdp and fertility.

(c) Replace both variables by their natural logarithms and draw another scatterplot. Does the simple li

Solution:

```
plot(log(ppgdp),log(fertility),
xlab = 'Per Capita GDP ($)',
ylab = 'Fertility')
```



Yes, after taking the log of both variables, an SLR model seems plausible for the summary of the graph.

2. The data set `prostate` in the `faraway` package is from a study of 97 men with prostate cancer. Interest is in predicting `lpsa` (log prostate specific antigen) with `lcavol` (log cancer volume). You may not use the function `lm` for this question.

   (a) Draw a scatterplot - does a simple linear regression model seem reasonable?
   Solution:

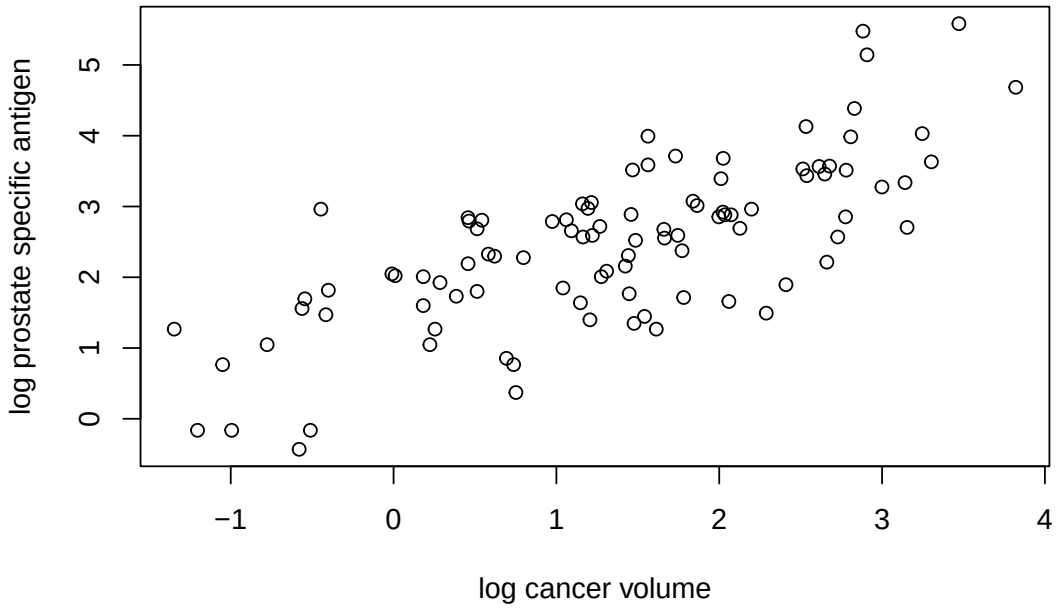
```
library(faraway)
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:alr4':
##
##     cathedral, pipeline, twins
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
data("prostate")
lcavol <- prostate$lcavol
lpsa <- prostate$lpsa
plot(lcavol,lpsa,
     xlab= 'log cancer volume',
     ylab= 'log prostate specific antigen')
```



Yes, a SLR model seems reasonable.

(b) Compute the values $\bar{x}, \bar{y}, S_{xx}, S_{yy}$ and S_{xy}. Compute the ordinary least squares estimates of the intercept and slope for the simple linear regression model, and draw the fitted line on your plot from part a).

Solution:

```
xbar <- mean(lcavol)
xbar
```

```
## [1] 1.35001
```

```
ybar <- mean(lpsa)
ybar
```

```
## [1] 2.478387
```

```
sxx <- sum((lcavol-xbar)^2)
sxx
```

```
## [1] 133.359
```

```
syy <- sum((lpsa-ybar)^2)
syy
```

```
## [1] 127.9176
```

```
sxy <- sum((lcavol-xbar)*(lpsa-ybar))
sxy
```

```
## [1] 95.92784
```

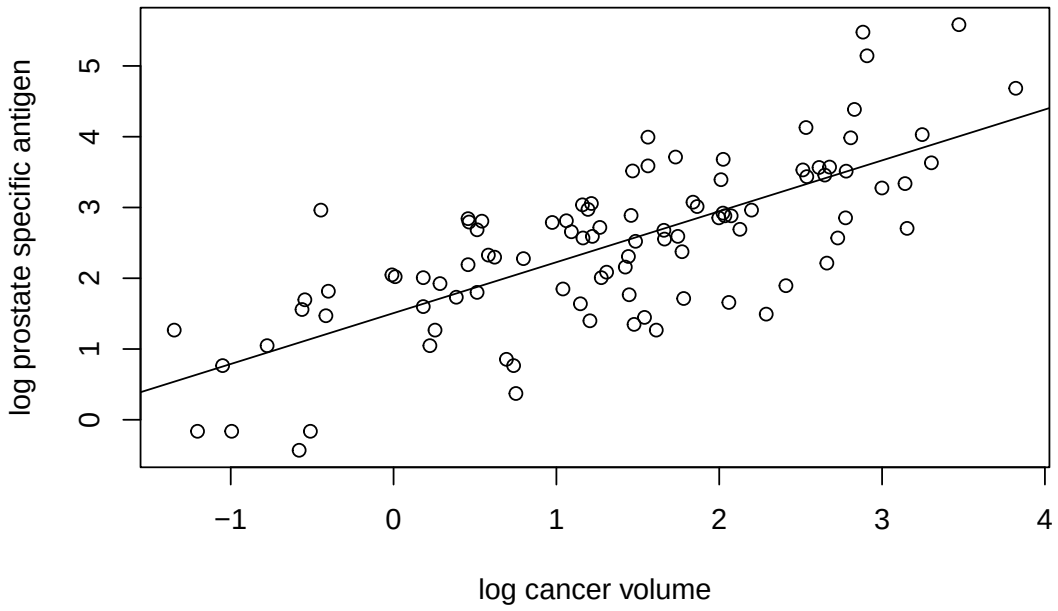
```
b1 <- sxy/sxx
b0 <- ybar-(b1*xbar)
b0
```

```
## [1] 1.507298
```

```
b1
```

```
## [1] 0.7193201
```

```
plot(lcavol,lpsa,
     xlab= 'log cancer volume',
     ylab='log prostate specific antigen')
abline(b0,b1)
```



(c) Compute $\hat{\sigma}^2$ and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Also find the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$.

Solution:

```
yhat <- b0+b1*lcavol
n <- length(lpsa)
mse <- (1/(n-2))*sum((lpsa-yhat)^2)
mse
```

```
## [1] 0.6201553
```

```
b0hatSE <- sqrt(mse*(1/n+(xbar^2)/sxx))
b0hatSE
```

```
## [1] 0.1219368
```

```
b1hatSE <- sqrt(mse/sxx)
b1hatSE
```

```
## [1] 0.06819288
```

```
covariance <- -xbar*mse/sxx
covariance
```

```
## [1] -0.006277907
```

(d) Carry out t-tests for the two null hypotheses $\beta_0 = 0$ and $\beta_1 = 0$, reporting the value of the test statistic and a p-value in each case.

Solution:

```
testb0 = b0/b0hatSE
pvalb0 = 2*pt(abs(testb0), df= n-2, lower.tail = F)
testb0
```

```
## [1] 12.3613
```

```
pvalb0
```

```
## [1] 1.722234e-21
```

```
testb1 = b1/b1hatSE
pvalb1 = 2*pt(abs(testb1), df= n-2, lower.tail = F)
testb1
```

```
## [1] 10.54832
```

```
pvalb1
```

```
## [1] 1.118616e-17
```

3. The data set `ftcollinstemp` in the `alr4` package gives the mean temperature in the fall of each year, defined as September 1 to November 30 , and the mean temperature in the following winter, defined as December 1 to the end of February in the following calendar year, in degrees Fahrenheit, for Ft. Collins, CO (Colorado Climate Center, 2012). These data cover the time period from 1900 to 2010 . The question of interest is: Does the average fall temperature predict the average winter temperature?

```
library(alr4)
```

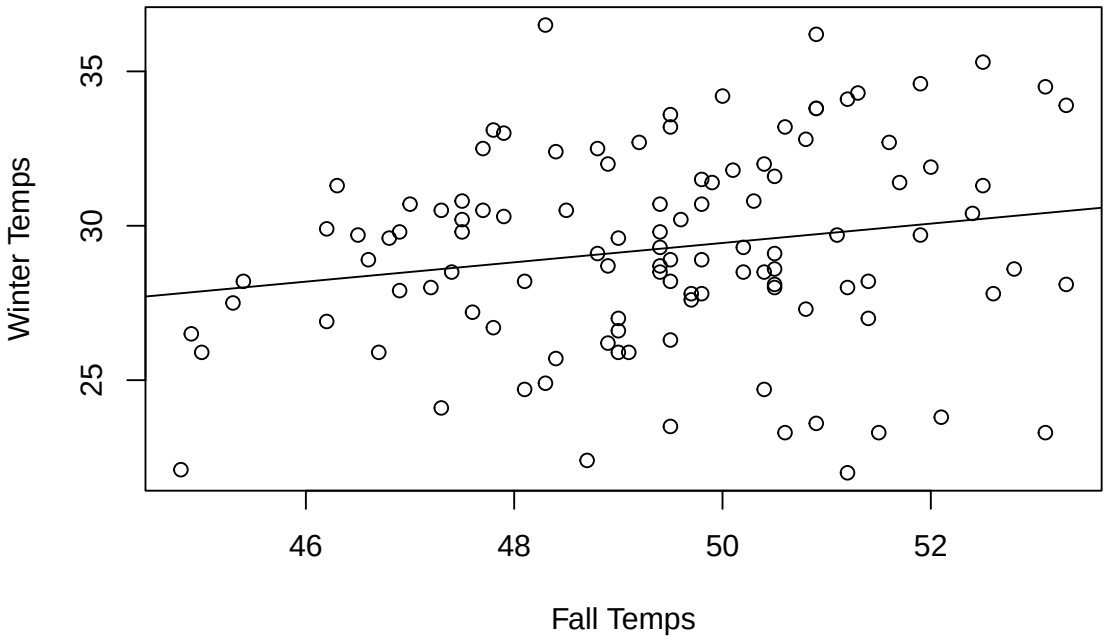
```
data("ftcollinstemp")
fall <- ftcollinstemp$fall
winter <- ftcollinstemp$winter
```

(a) Use the `lm` function in R to fit the regression of the response on the predictor. Draw a scatterplot of the data and add your fitted regression line.

Solution:

```
regression <- lm(winter~fall)
```

```r
plot(fall,winter,
xlab= 'Fall Temps',
ylab = 'Winter Temps',)
abline(coef(regression))
```



(b) Test the null hypothesis that the slope is 0 against a two-sided alternative at $\alpha = 0.01$, and interpret your findings.

Solution:

```r
summary(regression)$coefficients
```

```
##               Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 13.7843452  7.5548896 1.824559 0.07080657
## fall         0.3131691  0.1528193 2.049277 0.04283611
```

```r
pval <- .04283611
pval
```

```
## [1] 0.04283611
```

Given alpha = 0.01, we have that our p-val>alpha and thus the slope of the regression line is 0.

(c) What percentage of the variability in winter is explained by fall?

Solution:

```
summary(regression)$r.squared
```

```
## [1] 0.03709854
```

The percentage of the variability in winter explained by fall is equal to 3.71%