PSTAT 126 - Homework 2

Emily Lu

16 April, 2019

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.

```
library(alr4)
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

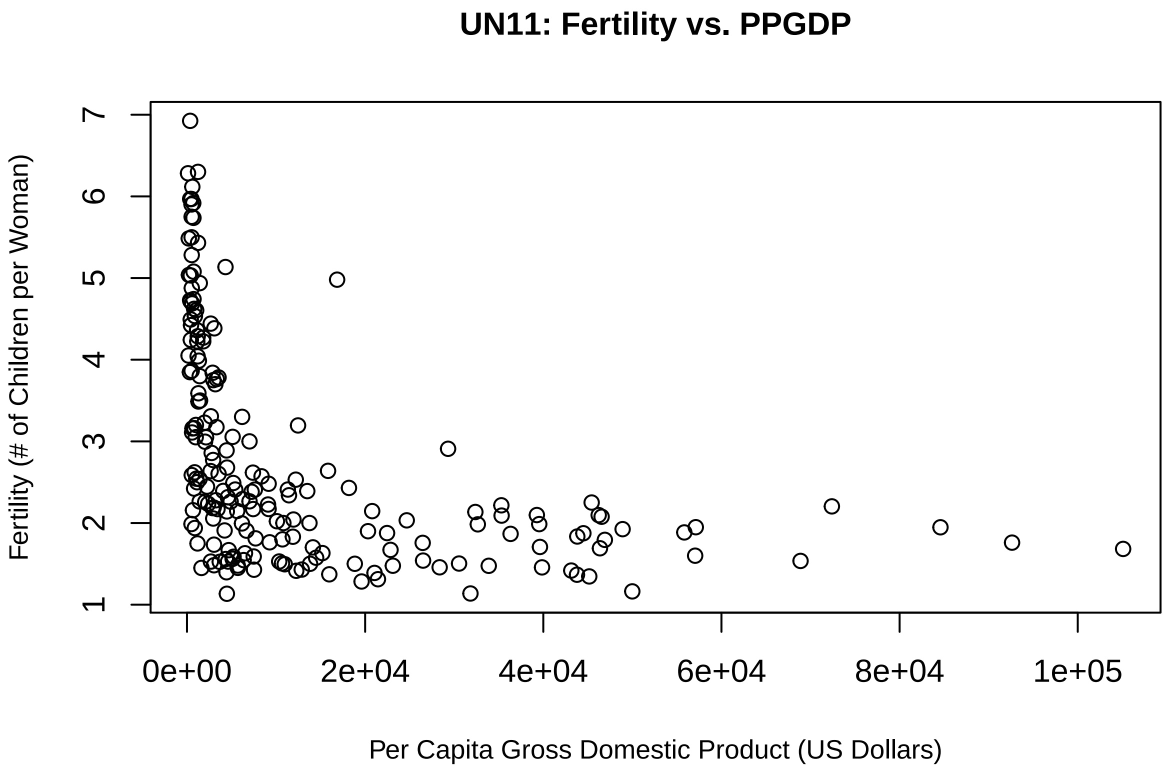
```
data("UN11")
fertility <- UN11$fertility
ppgdp <- UN11$ppgdp
```

(a) Identify the predicator and response.

The predictor is ppgdp and the response is fertility.

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

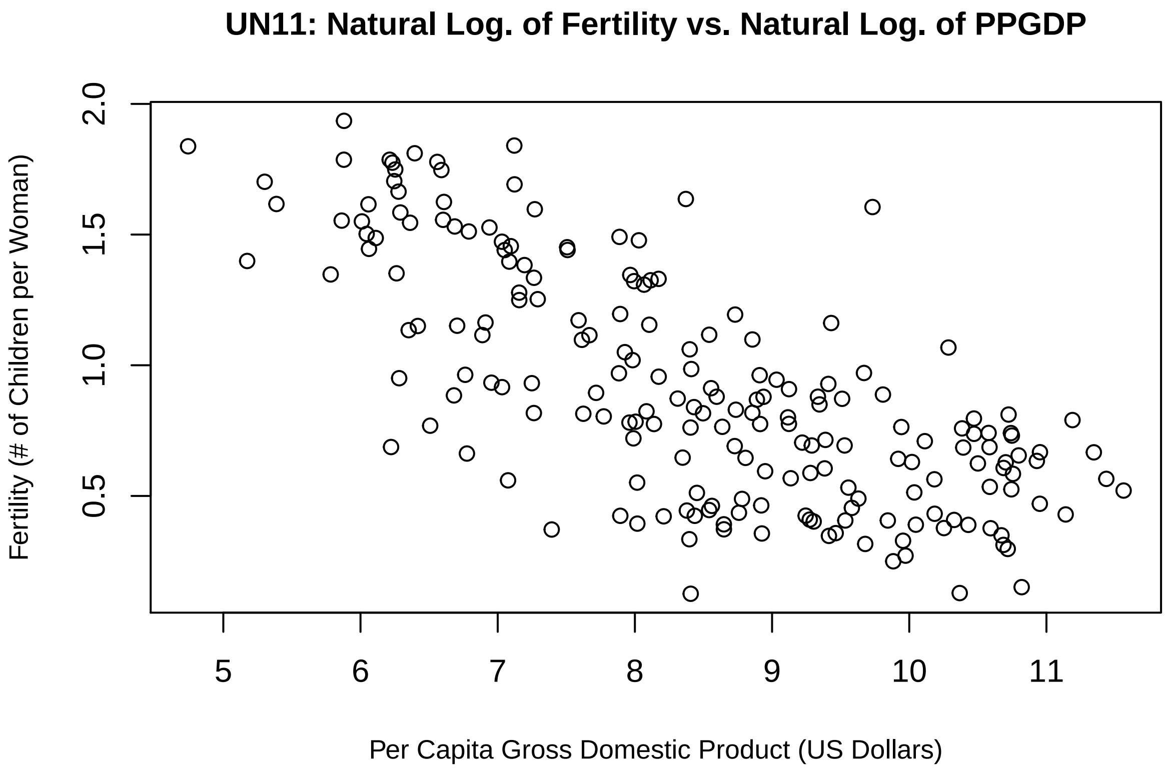
```
plot(ppgdp, fertility,
     xlab = 'Per Capita Gross Domestic Product (US Dollars)',
     ylab = 'Fertility (# of Children per Woman)',
     main = 'UN11: Fertility vs. PPGDP',
     cex.main = 1,
     cex.lab = 0.8)
```



The trend does not appear linear.

(c) Replace both variables by their natural logarithms and draw another scatterplot. Does the simple linear regression model seem plausible for a summary of this graph?

```r
NLppgdp <- log(ppgdp)
NLfertility <- log(fertility)

plot(NLppgdp, NLfertility, xlab = 'Per Capita Gross Domestic Product (US Dollars)',
     ylab = 'Fertility (# of Children per Woman)',
     main = 'UN11: Natural Log. of Fertility vs. Natural Log. of PPGDP',
     cex.main = 1,
     cex.lab = 0.8)
```



After taking the natural logarithms of both variables, the simple linear regression model does seem plausible for a 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).

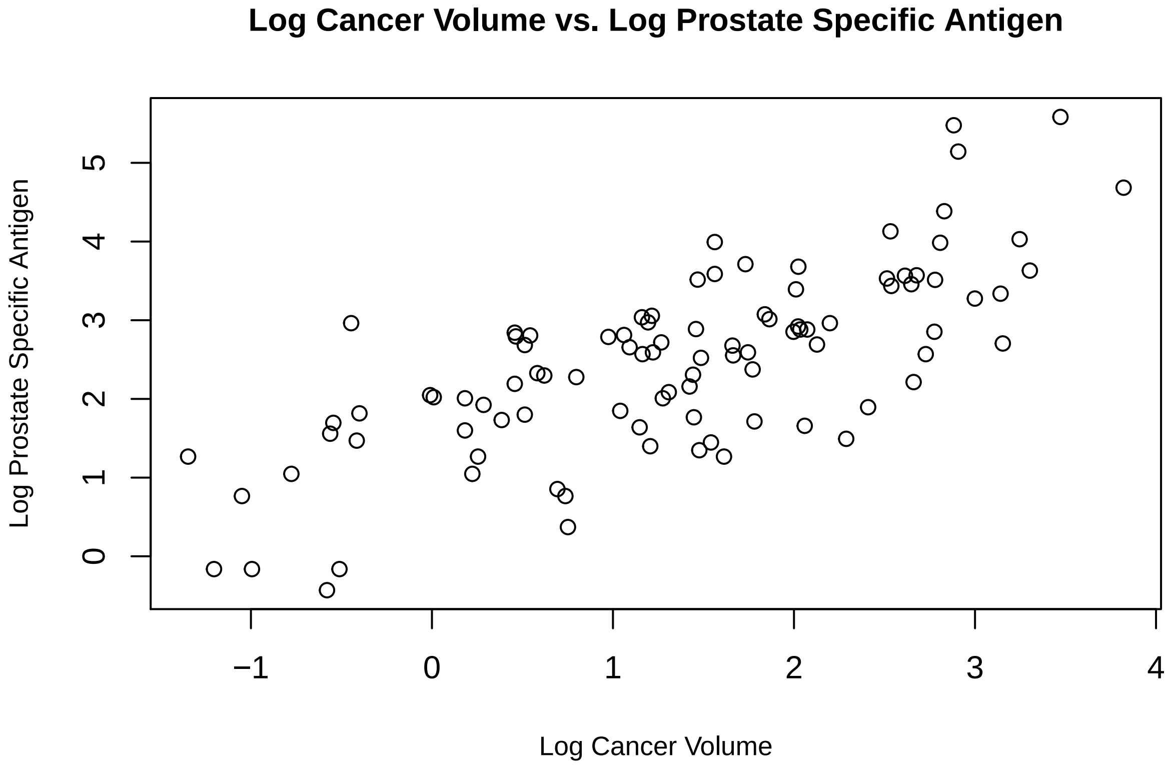
```r
library(faraway)
```

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

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

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

```r
data("prostate")
x <- prostate$lcavol
y <- prostate$lpsa
```

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

```r
plot(x, y,
     xlab = "Log Cancer Volume",
     ylab = "Log Prostate Specific Antigen",
     main = "Log Cancer Volume vs. Log Prostate Specific Antigen",
     cex.main = 1,
     cex.lab = 0.8)
```



This scatterplot shows that these data set can be a reasonable simple linear regression model since the trend appears to have a linearity, normality and constant variance.

(b) Without using the function lm, 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).

From the computations below, we have that $\bar{x} = 1.35001$, $\bar{Y} = 2.478387$, $S_{XX} = 133.359$, $S_{YY} = 127.9176$, and $S_{XY} = 95.92784$. For the simple linear regression model, the ordinary least squares estimates of the intercept is 1.5072979 and the ordinary least squares estimates of the slope is 0.7193201.

```
# sample means
xbar <- mean(x)
ybar <- mean(y)
c(xbar, ybar)
```

```
## [1] 1.350010 2.478387
```

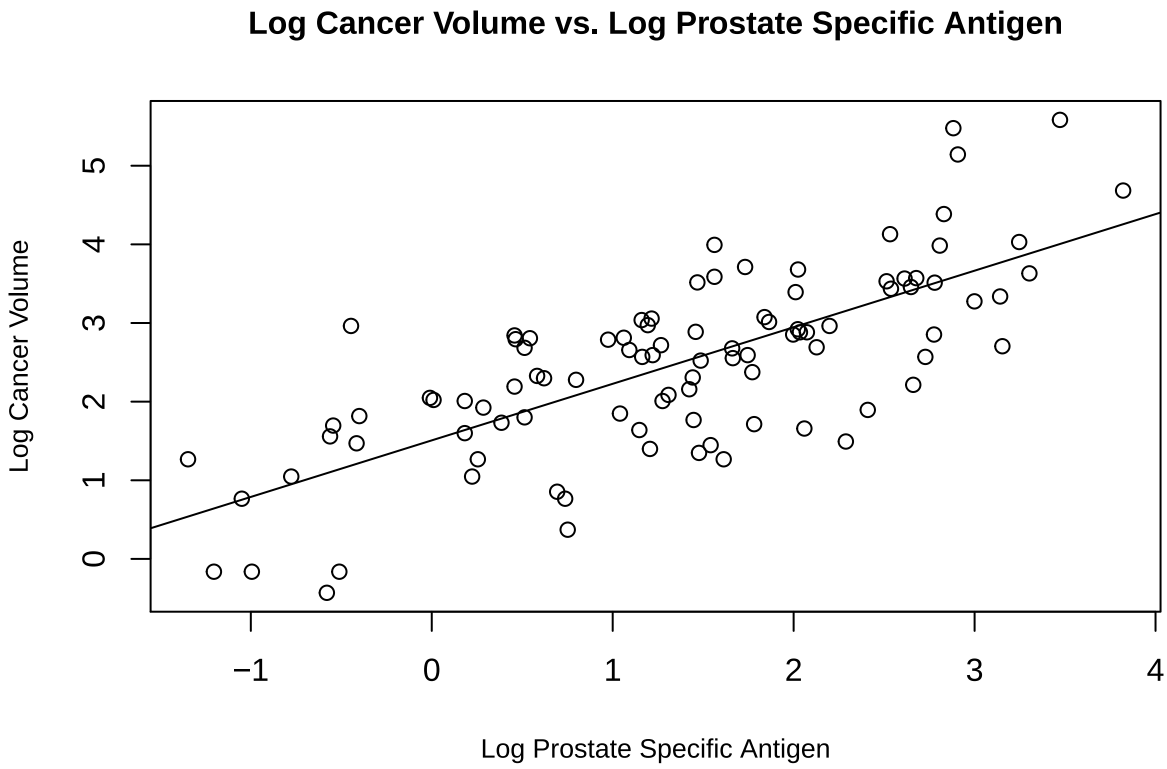
```
# sum of squares
Sxx <- sum((x - xbar)^2)
Syy <- sum((y - ybar)^2)
Sxy <- sum((x - xbar)*(y - ybar))
c(Sxx, Syy, Sxy)
```

```
## [1] 133.35903 127.91758  95.92784
```

```
# intercept and slope
b1 <- Sxy/Sxx
b0 <- ybar - b1*xbar
c(b0, b1)
```

```
## [1] 1.5072979 0.7193201
```

```
plot(x, y,
     xlab = "Log Prostate Specific Antigen",
     ylab = "Log Cancer Volume",
     main = "Log Cancer Volume vs. Log Prostate Specific Antigen",
     cex.main = 1,
     cex.lab = 0.8)
abline(b0, b1)
```



(c) Obtain the estimate of σ^2 and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Also find the estimated covariance between these two estimates. 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.

```
# estimate of sigma squared (MSE)
yhat <- b0 + b1*x
n <- length(y)
mse <- (1/(n-2))*sum((y-yhat)^2)
mse
```

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

```
# estimated errors of B0 hat
SEb0hat <- sqrt(mse*(1/n + (xbar^2)/Sxx))
```

```
SEb0hat
```

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

```
# estimated errors of B1 hat
SEb1hat <- sqrt(mse/Sxx)
SEb1hat
```

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

```
# estimated covariance between these two estimates
covb0b1 <- -xbar*mse/Sxx
covb0b1
```

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

```
# test statistic and p-value for B0 = 0
testStat_b0 = b0/SEb0hat
pValue_b0 = 2*pt(abs(testStat_b0), df = n-2, lower.tail = F)
c(testStat_b0, pValue_b0)
```

```
## [1] 1.236130e+01 1.722234e-21
```

```
# test statistic and p-value for B1 = 0
testStat_b1 = b1/SEb1hat
pValue_b1 = 2*pt(q = testStat_b1, df = n-2, lower.tail = F)
c(testStat_b1, pValue_b1)
```

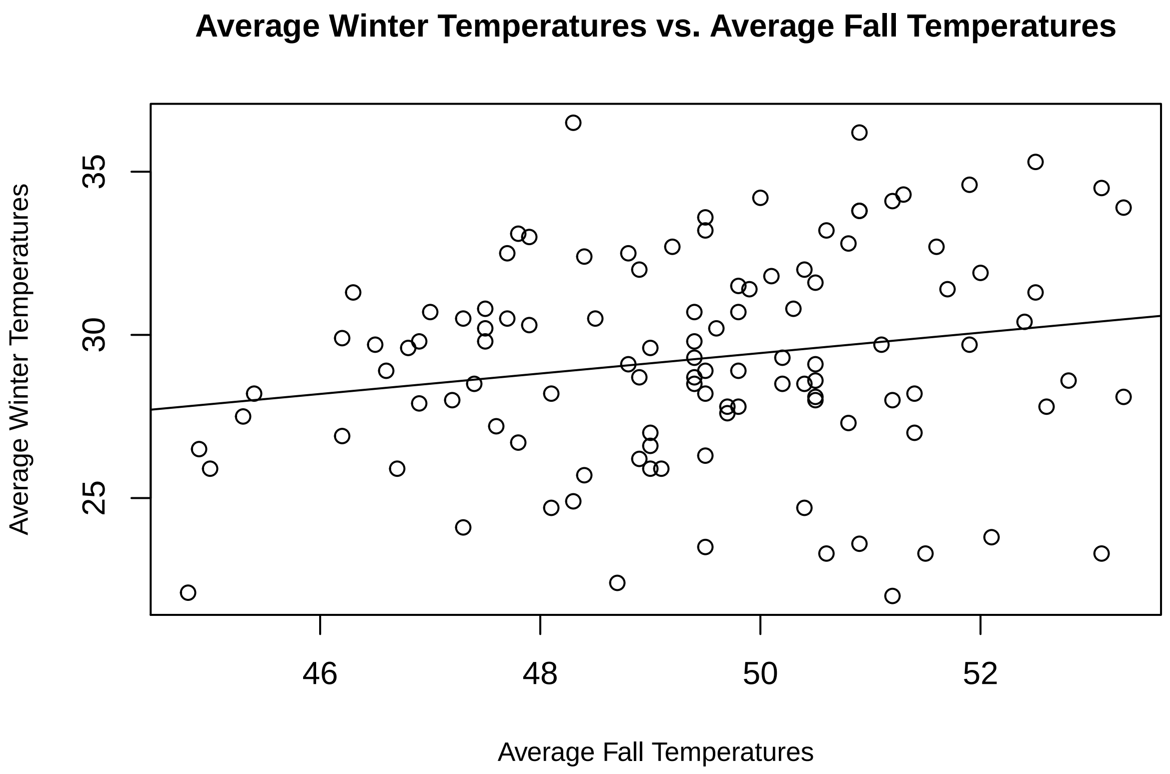
```
## [1] 1.054832e+01 1.118616e-17
```

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")
x <- ftcollinstemp$fall
y <- ftcollinstemp$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.

```
fit1 <- lm(y~x)
plot(x, y,
     xlab = "Average Fall Temperatures",
     ylab = "Average Winter Temperatures",
     main = "Average Winter Temperatures vs. Average Fall Temperatures",
     cex.main = 1,
     cex.lab = 0.8)
abline(fit1$coefficients[1], fit1$coefficients[2])
```



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

```
# null hypothesis is "the slope of the regression line is equal to 0"
summary(fit1)$coefficients
```

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

```
pValue <- 0.04284
pValue
```

```
## [1] 0.04284
```

Since the p-value, 0.04284, is greater than the $\alpha = 0.01$, this means we could accept the null hypothesis. In other words, the slope of regression line is equal to 0.

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

```
summary(fit1)$r.squared
```

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

3.71% of the variability in winter is explained by fall.