PSTAT 126 Project 2

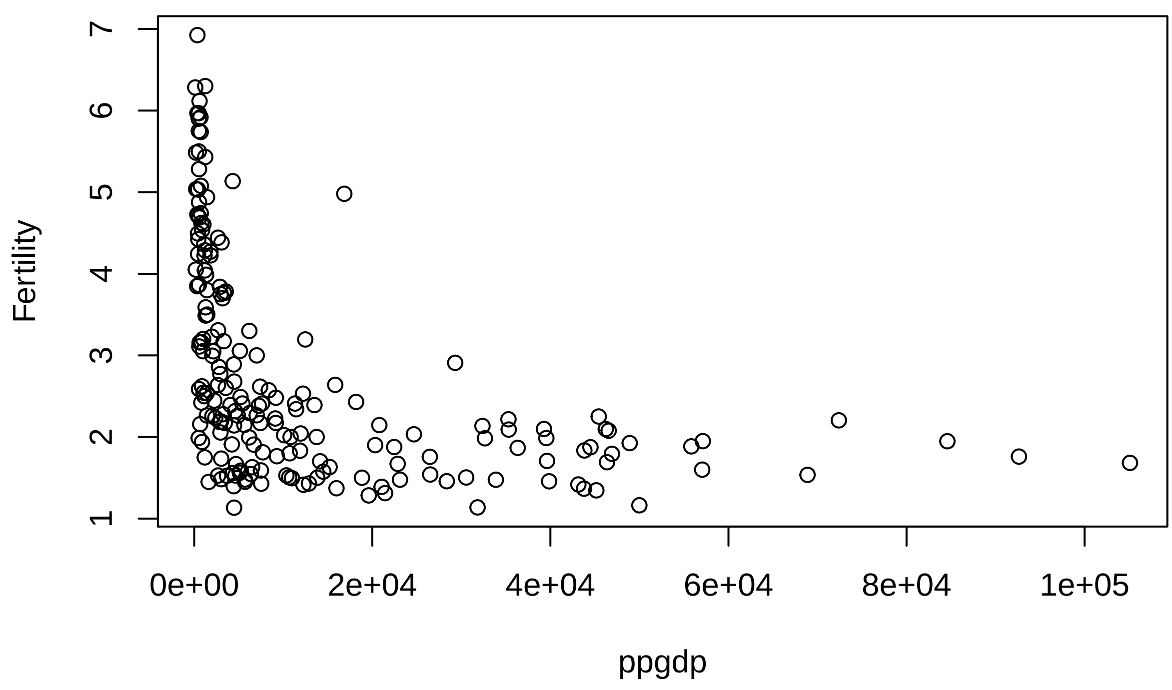
Problem 1

1. The R package `alr4` contains a dataset called `UN11` that includes the U.S national gross product per person (Predictor) and Fertility (Response). Answer for 1a.

```
library(alr4)
```

1b) Ploting the data

```
x=UN11$ppgdp
y=UN11$fertility
plot(x,y,xlab="ppgdp",ylab="Fertility")
```

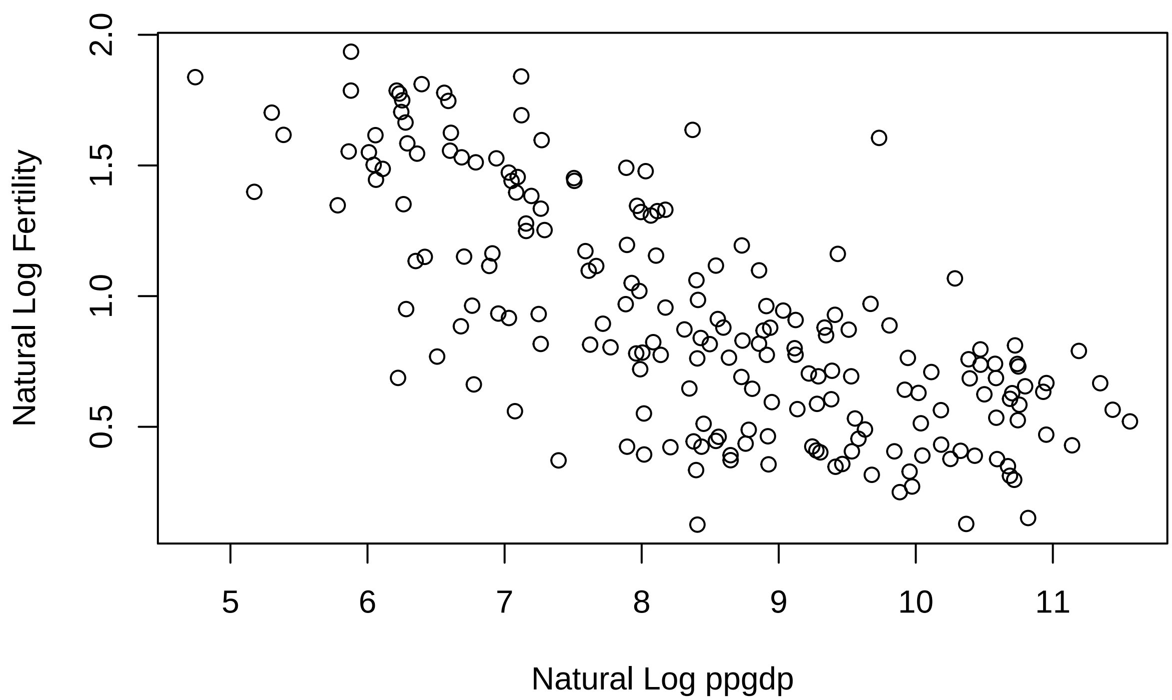


```
## The trend is not linear
```

The trend in the plot appears to be not linear.

1c) Replacing Variables

```
x1 = log(UN11$ppgdp)
y1 = log(UN11$fertility)
plot(x1,y1,xlab="Natural Log ppgdp",ylab="Natural Log Fertility")
```



The Simple Linear Regression Model is plausible for a summary of this graph.

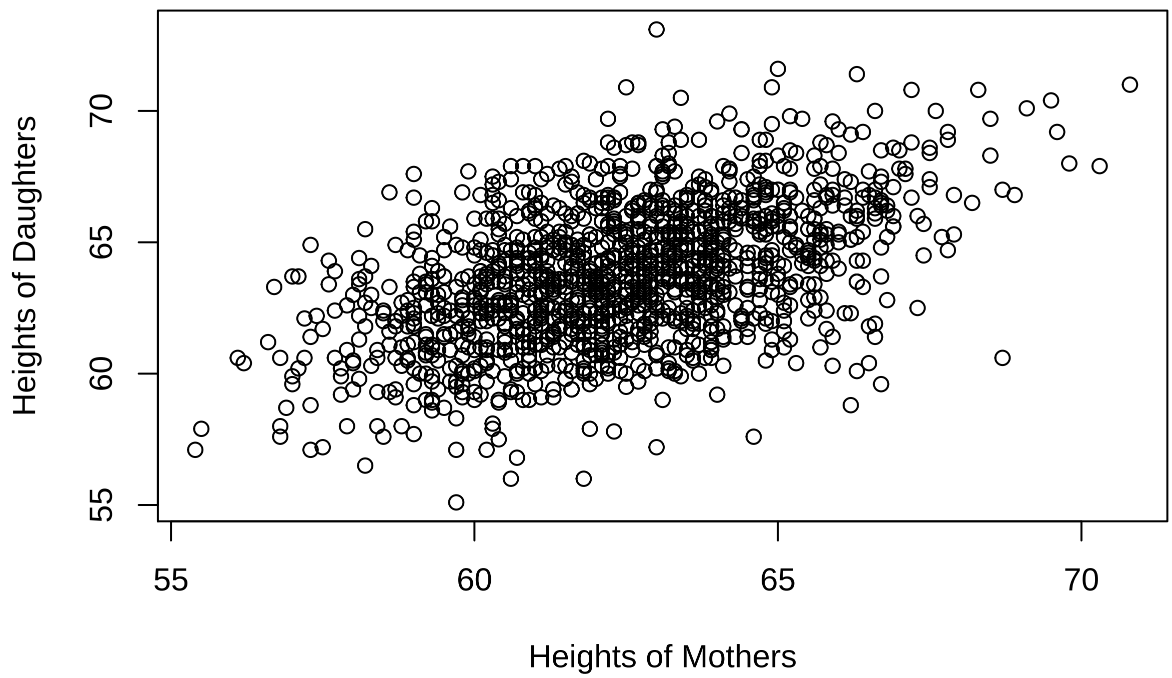
Problem 2

2. The R package `alr4` contains a dataset called `Heights` that includes the heights of families in England. The data set includes 1375 pairs of heights of mothers (mheight) and their daughters (dheights) in inches.

```
library(alr4)
```

2a) Drawing the scatterplot

```r
##Predictor = mheight response = dheight
x = Heights$mheight
y = Heights$dheight
plot(x,y,xlab="Heights of Mothers",ylab="Heights of Daughters")
```



2b) Computations

```r
##Computing x bar
xbar = mean(x)
xbar
```

```
## [1] 62.4528
```

```r
## Computing y bar
ybar = mean(y)
ybar
```

```
## [1] 63.75105
```

```r
## Computing Sxx
Sxx = sum((x - xbar)^2)
Sxx
```

```
## [1] 7620.907
```

```r
## Computing Syy
Syy = sum((y - ybar)^2)
Syy
```

```
## [1] 9288.616
```

```r
## Computing Sxy
Sxy = sum((x - xbar)*(y - ybar))
Sxy
```

```
## [1] 4128.603
```

```r
## Compute the Least Squares Estimate of Intercept
r = Sxy/(sqrt(Sxx*Syy))
r
```

```
## [1] 0.4907094
```

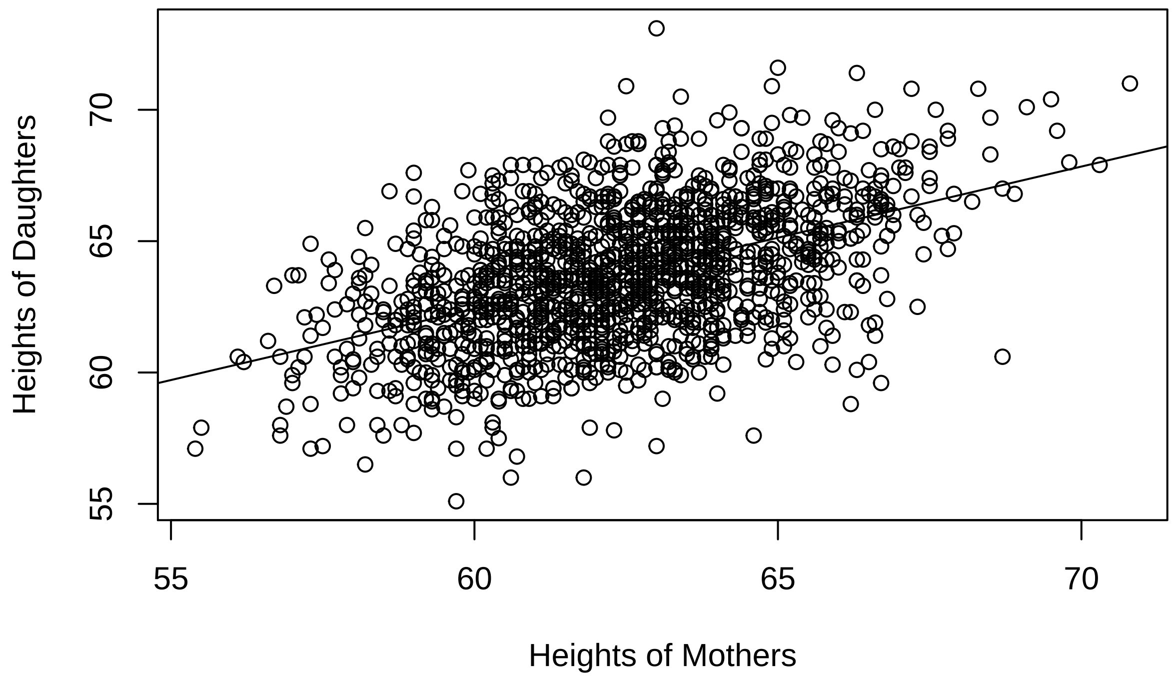
```r
## Compute the Slope for Simple Linear Regression Model
b1 = r*(sqrt(Syy/Sxx))
b1
```

```
## [1] 0.541747
```

```r
b0 = ybar - (b1*xbar)
b0
```

```
## [1] 29.91744
```

```r
## Drawing the Fitted Line
plot(x,y,xlab="Heights of Mothers",ylab="Heights of Daughters")
abline(b0,b1)
```



2c) Computing Estimates and Standard Errors

```
yhat = b0 + b1*x
e = y - yhat
n = length(x)
sigma2hat = (sum(e^2))/(n - 2)

## Estimated Standard Error
sigmahat = sqrt(sigma2hat)
sigmahat
```

```
## [1] 2.266311
```

T testing

```
## Standard Error for b0
se_b0 = sigmahat*sqrt(1/n + mean(x)^2/Sxx)
se_b0
```

```
## [1] 1.622469
```

```r
## Standard Error for b1
se_b1 = sigmahat/sqrt(Sxx)
se_b1
```

```
## [1] 0.02596069
```

```r
##T test for null Hypothesis where b0=0,Test Stat, P value
t_stat_b0 = b0/se_b0
t_stat_b0
```

```
## [1] 18.43945
```

```r
p_val_b0 = pt(t_stat_b0,df = n - 2, lower.tail = FALSE)
p_val_b0
```

```
## [1] 2.60594e-68
```

```r
##T test for null hypothesis where b1=0, Test stat, P value
t_stat_b1 = b1/se_b1
t_stat_b1
```

```
## [1] 20.86797
```

```r
p_val_b1 = pt(t_stat_b1, df = n - 2, lower.tail = FALSE)
p_val_b1
```

```
## [1] 1.608457e-84
```

2d) 99% Confidence interval for $b1$

```r
t_pct = qt(p=.995, df = n - 2)
ci_b1_99 = b1 + c(-1,1) * se_b1 * qt(p=.995, df = (length(x)))
ci_b1_99
```

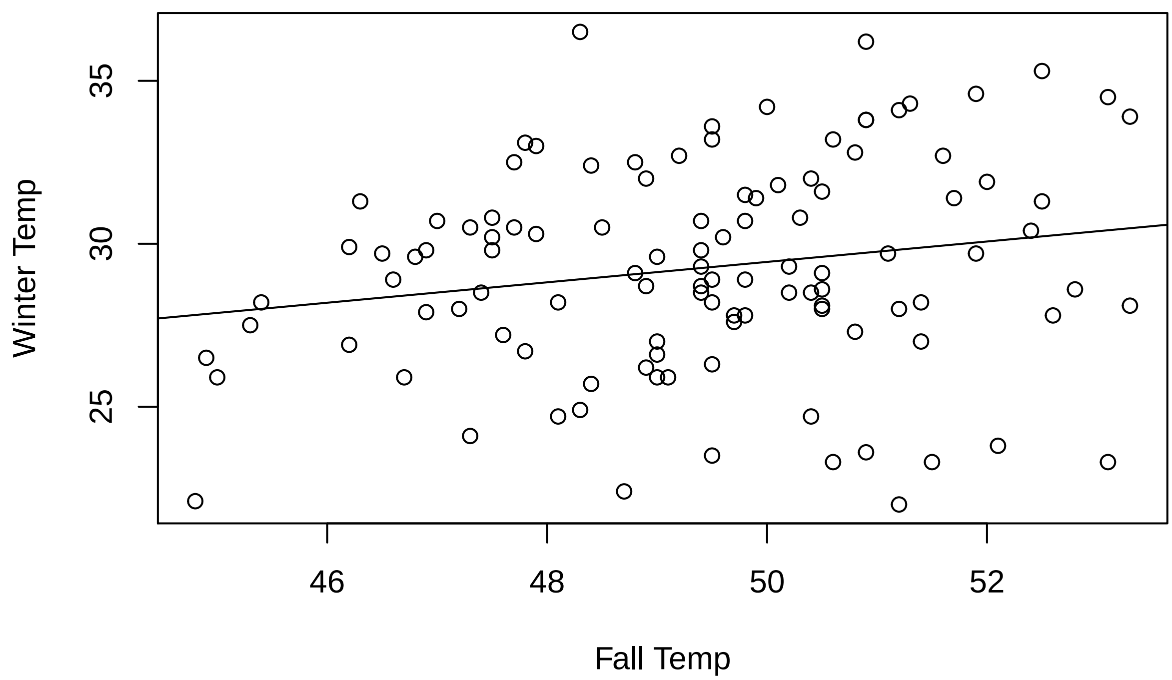
```
## [1] 0.4747838 0.6087103
```

Problem 3

3. The R package `alr4` contains a dataset called `ftcollinstemp` that includes the temperatures of Fall and Winter. The dataset includes the temperatures in degree Farenheit of the temperatures of the years 1900 to 2010.

3a) Drawing the Regression line

```r
library(alr4)
attach(ftcollinstemp)
x = fall
y = winter
fit1 = lm(y ~ x)
plot(x,y,xlab= "Fall Temp", ylab = "Winter Temp")
abline(fit1$coef[1], fit1$coef[2])
```



3b) Testing the Null Hypothesis where the slope = 0

```r
## test H0: slope = 0 vs Ha: slope != 0
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8186 -1.7837 -0.0873  2.1300  7.5896
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7843     7.5549   1.825   0.0708 .
## x             0.3132     0.1528   2.049   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.179 on 109 degrees of freedom
## Multiple R-squared:  0.0371, Adjusted R-squared:  0.02826
## F-statistic:   4.2 on 1 and 109 DF,  p-value: 0.04284
```

```
## Because the p value is less than your alpha = .05
## this is significant evidence that the slope = 0.
## therefore fall weather can predict winter weather, we accept the null hypothesis
```

3c) Computing the 99% Confidence Interval

```
## Compute the T percentile
t_pct = qt(p=.975, df = length(x) - 2)
sxx = sum((x-mean(x))^2)
fit1 = lm(y ~ x)
yhat = fit1$coef[1] + fit1$coef[2] * x
e = y - yhat
sigma2hat = sum(e^2)/(length(x) - 2)
sigmahat = sqrt(sigma2hat)
se_b1 = sigmahat/sqrt(sxx)
## Confidence interval 99%
ci_b1_99 = fit1$coef[2] + c(-1,1) * t_pct * se_b1
ci_b1_99
```

```
## [1] 0.01028623 0.61605204
```

3d) Conclusion

We are 99% confident that the true variation in winter is exmplained by the variation of fall lies within the interval (.2278,.3985).

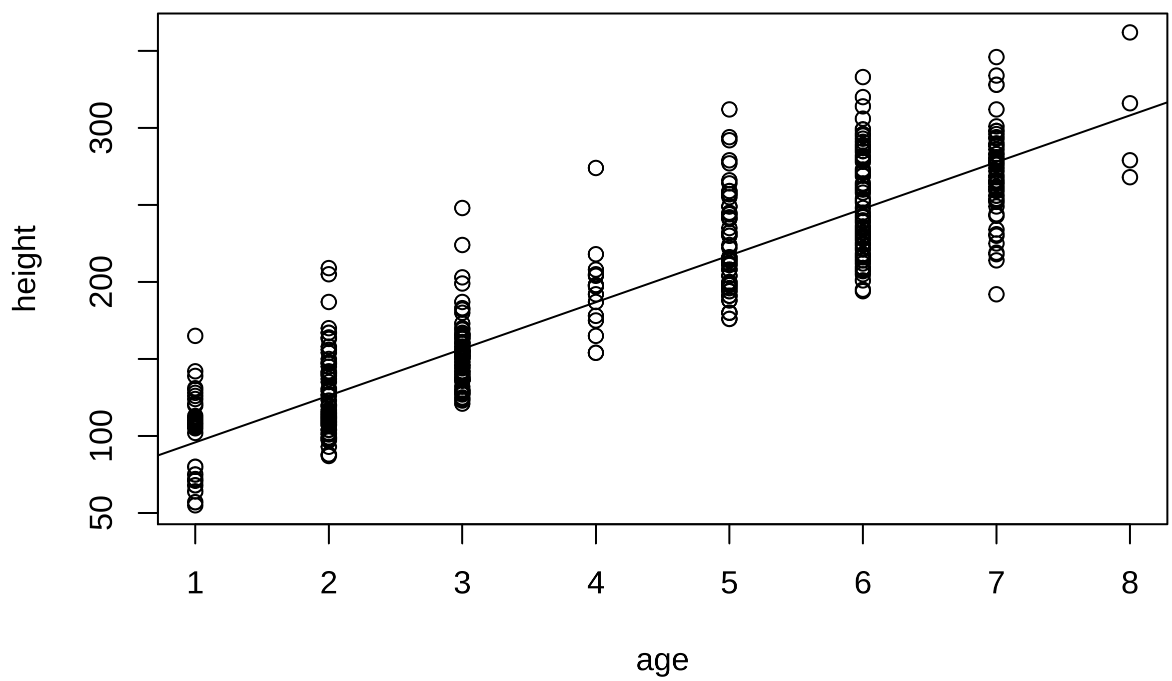
Problem 4

4. The R package `alr4` contains a dataset called `wblake` that includes the samples of small mouth bass collected in Minnesota. The data set includes the length and age of these bass.

4a) Ploting the Regression Line

```
library(alr4)
attach(wblake)
x = Age
y = Length
```

```r
fit2 = lm(y ~ x)
plot(x,y,xlab="age",ylab="height")
abline(fit2$coef[1],fit2$coef[2])
```



4b) Testing the Null Hypothesis where slope = 0

```r
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.794 -19.499  -4.499  16.177  94.853
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.5272     3.1974   20.49   <2e-16 ***
## x            30.3239     0.6877   44.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 28.65 on 437 degrees of freedom
## Multiple R-squared:  0.8165, Adjusted R-squared:  0.8161
## F-statistic:  1944 on 1 and 437 DF,  p-value: < 2.2e-16
```

```r
##Since the p value is less than our alpha value of .05,
##we can reject our Null Hypothesis.
##Therefore the slope does not equal to 0.
```

4c) Finding the 95% Confidence Interval for mean length at age 4 years

```r
age = data.frame(x = 4)
predict(fit2, newdata = age, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 186.8227 184.1217 189.5237
```

4d) Finding the 95% Confidence Interval for mean length at age 9 years

```r
age9 = data.frame(x = 9)
pre = predict(fit2, newdata = age9, interval = "confidence", level = 0.95)
pre
```

```
##        fit      lwr      upr
## 1 338.4422 331.4231 345.4612
```

```r
## This interval is not trustworthy because at the age of 9 years the
## bass is not fully reached its full length.
```