Homework 3

Animesh Sengupta

9/27/2022

```
setwd("/Users/animeshsengupta/Work Directory/DACSS/STAT625/Homeworks")
library(alr4)  # loads the installed package into the workspace so you can use it
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

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

```
library(summarytools)
library(ggplot2)
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

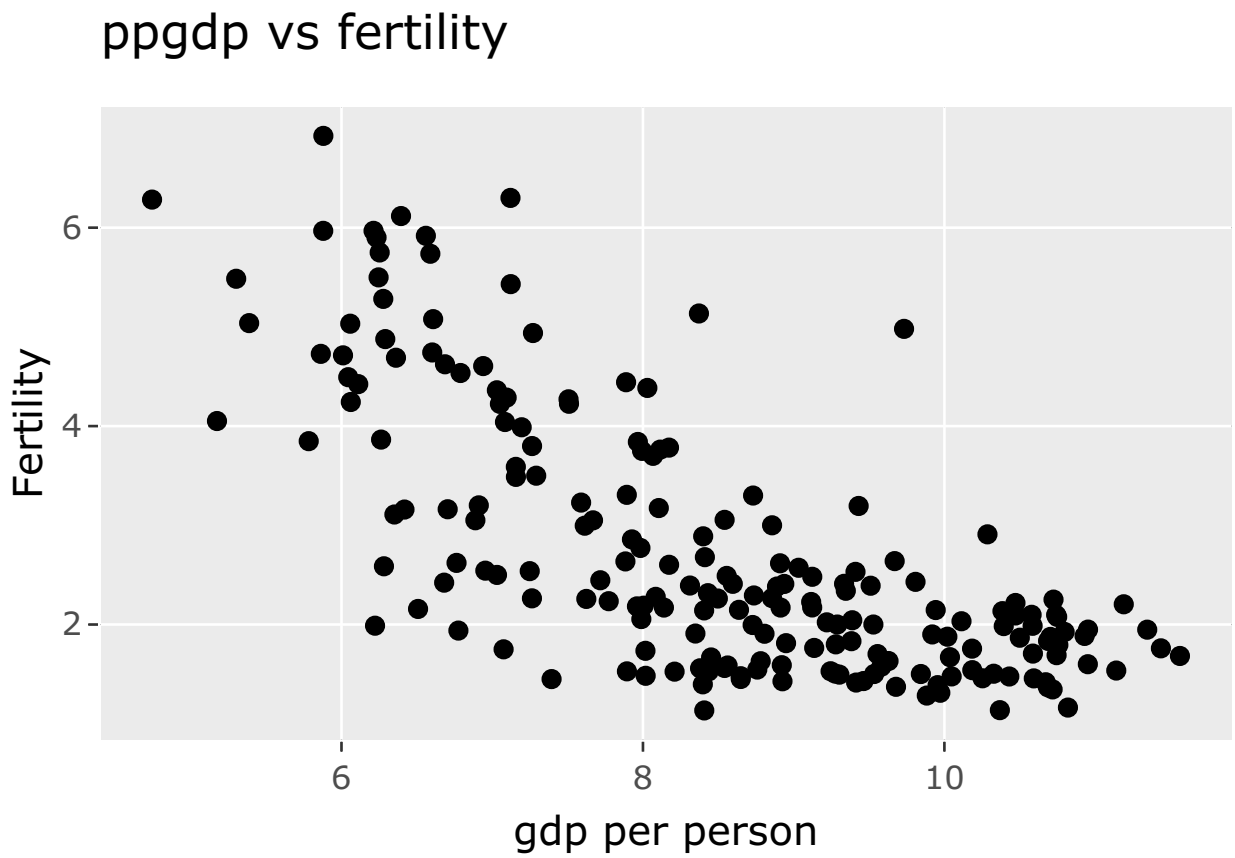
```
## The following object is masked from 'package:graphics':
##
##     layout
```

Answer 3.1

```
colnames(UN11)
```

```
## [1] "region"    "group"     "fertility" "ppgdp"     "lifeExpF"  "pctUrban"
```

```
b1<- ggplot(UN11,aes(x=log(ppgdp), y=fertility)) +
    geom_point()+
    xlab("gdp per person")+
    ylab("Fertility")+
    ggtitle("ppgdp vs fertility")
ggplotly(b1)
```




```
UN11$logppgdp=round(log(UN11$ppgdp),2)
ans1<-UN11%>%filter(logppgdp==9.73|logppgdp==8.37)
#ans2<-UN11%>%filter(logppgdp==8.37)
ans1
```

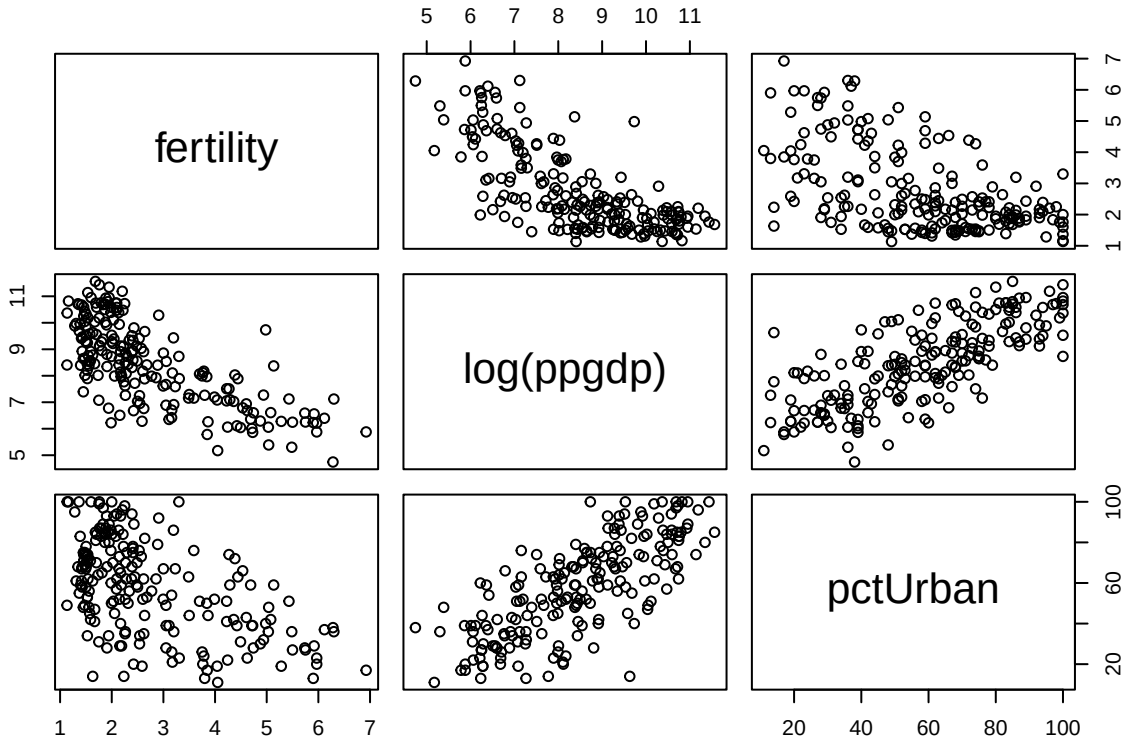
```
##                   region  group fertility   ppgdp lifeExpF pctUrban logppgdp
## Angola            Africa africa     5.135  4321.9    53.17       59     8.37
## Equatorial Guinea Africa africa     4.980 16852.4    52.91       40     9.73
```

There are some regions and localities in Africa(as per above dataframe) which has a high log(ppgdp) and even higher fertility. they all have lower life expectency, around the range of 52% in common.

Answer 3.2

3.2.1

```
ans321<-pairs(~ fertility+log(ppgdp)+pctUrban, data=UN11)
```




```
ans321
```

```
## NULL
```

The most clear marginal relationship is between log(ppgdp) and pcturban, As evident from the scatterplot a linear model with a positive slope will fit it perfectly well.

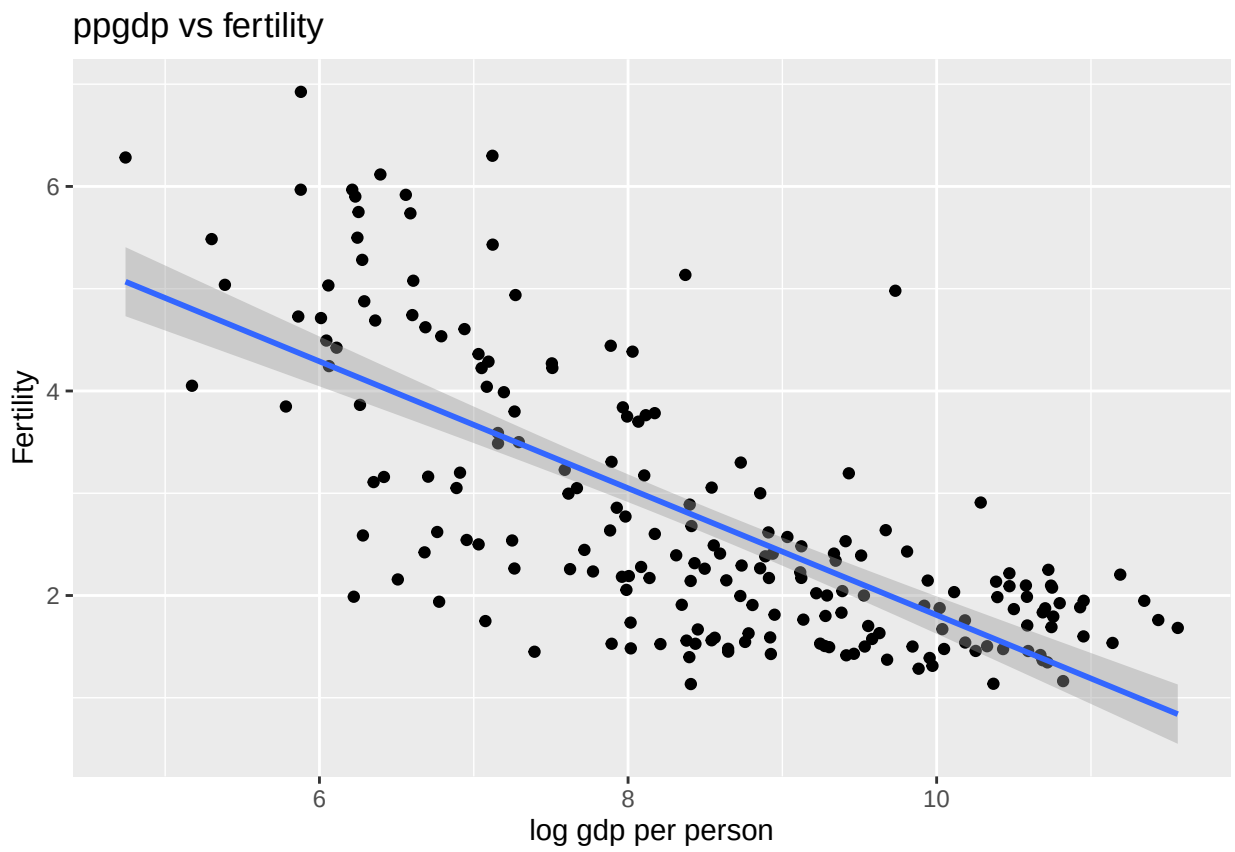
However the relationship between fertility and log(ppgdp) as per the scatterplot seems a bit complex. A lot of data points are clustered in one region , which shows skewed distribution over log(ppgdp). a single fit of line wouldnt be able to do justice for modelling. The same can be said about the marginal relationship between fertility and pcturban.

3.2.2

```
unml1<-lm(fertility~log(ppgdp),UN11)
unml2<-lm(fertility~pctUrban,UN11)
b1<- ggplot(UN11,aes(x=log(ppgdp), y=fertility)) +
    geom_point()+
    stat_smooth(method="lm")+
    xlab("log gdp per person")+
```

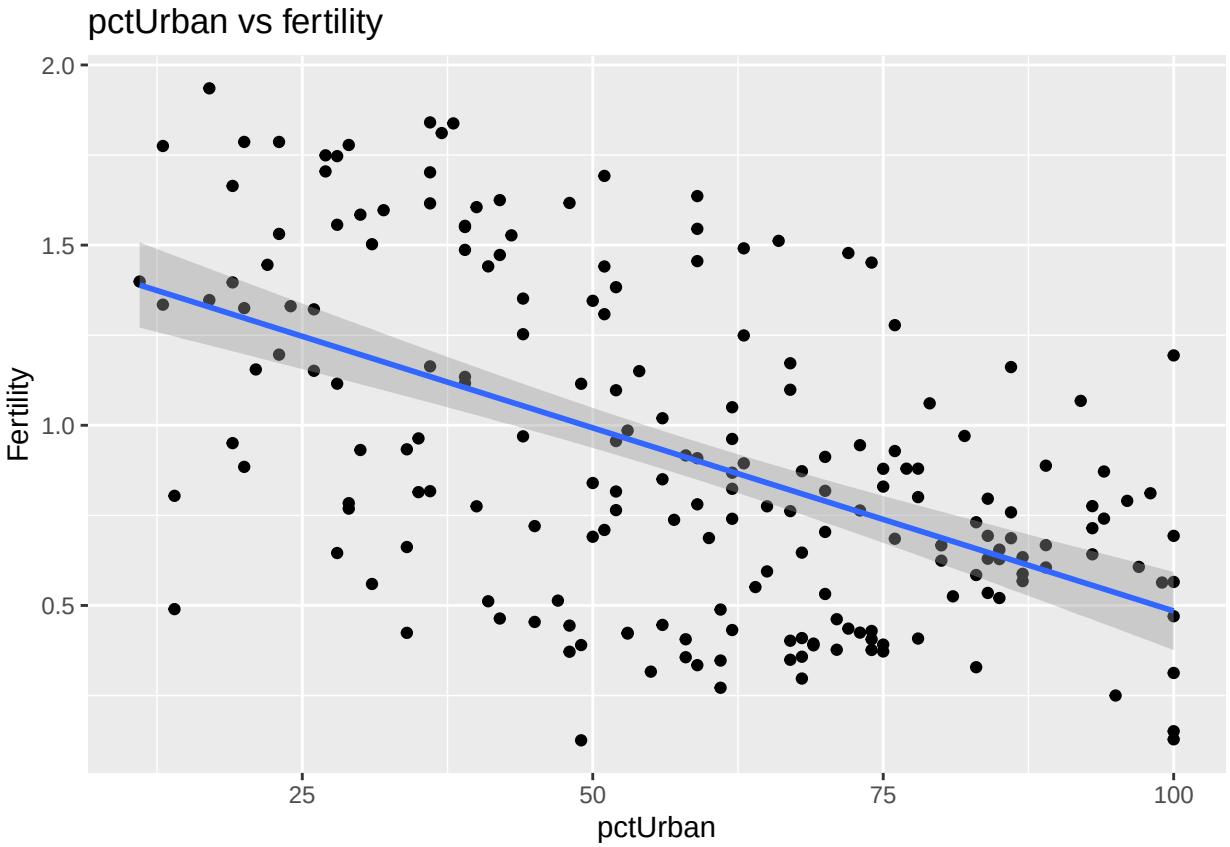
```
    ylab("Fertility")+
    ggtitle("ppgdp vs fertility")
b2<-ggplot(UN11,aes(x=pctUrban, y=log(fertility))) +
    geom_point()+
    stat_smooth(method="lm")+
    xlab("pctUrban")+
    ylab("Fertility")+
    ggtitle("pctUrban vs fertility")
b1
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
b2
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
summary(unml1)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16313 -0.64507 -0.06586  0.62479  3.00517
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.00967    0.36529   21.93   <2e-16 ***
## log(ppgdp)  -0.62009    0.04245  -14.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9305 on 197 degrees of freedom
## Multiple R-squared:   0.52,  Adjusted R-squared:  0.5175
## F-statistic: 213.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
summary(unml2)
```

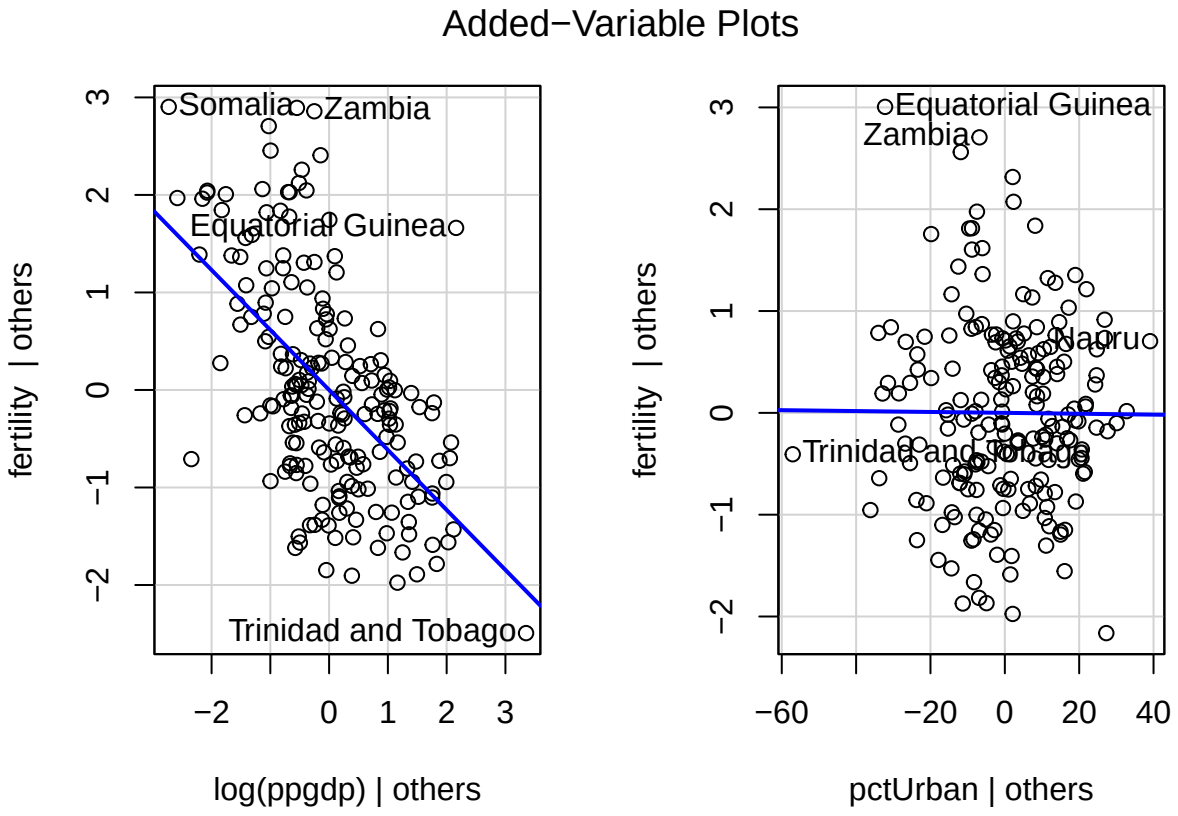
```
##
```

```
## Call:
## lm(formula = fertility ~ pctUrban, data = UN11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4932 -0.7795 -0.1475  0.6517  2.9029
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.559823   0.213681  21.339   <2e-16 ***
## pctUrban    -0.031045   0.003421  -9.076   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.128 on 197 degrees of freedom
## Multiple R-squared:  0.2948, Adjusted R-squared:  0.2913
## F-statistic: 82.37 on 1 and 197 DF,  p-value: < 2.2e-16
```

Since the p value of both the models are very less in the range of 2e-16 , we can conclude that they are significantly different from 0 at any conventional level of significance like =0.05 or 0.001

3.2.3

```
unml3<-lm(fertility~log(ppgdp)+pctUrban,UN11)
avPlots(unml3)
```




```
summary(unml3)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp) + pctUrban, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.9932699  0.3993367  20.016   <2e-16 ***
## log(ppgdp)  -0.6151425  0.0641565  -9.588   <2e-16 ***
## pctUrban    -0.0004393  0.0042656  -0.103    0.918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9328 on 196 degrees of freedom
## Multiple R-squared:   0.52,  Adjusted R-squared:  0.5151
## F-statistic: 106.2 on 2 and 196 DF,  p-value: < 2.2e-16
```

Acc to the added variable plots, the log(ppgdp) is useful after adjusting for pcturban because the model somewhat fits to the data and there is less variability across the fitted line. it models negatively with y=-x model type. the same cant be said for the pcturban after adjusting log(ppgdp) because it poorly fits the data and there is a lot of deviation from fitted line.

3.2.4

As seen from the summary of both the models, the estimated coefficient as -0.62 from single regression of log(ppgdp) and multiple regression with pcturban.

3.2.5 and 3.2.3 mean function computation

```
unml4<-lm(pctUrban~log(ppgdp), UN11)
unml5<-lm(log(ppgdp)~pctUrban, UN11)
unml6<-lm(residuals(unml1)~residuals(unml4), UN11)
unml7<-lm(residuals(unml2)~residuals(unml5), UN11)
unml8<-lm(residuals(unml3)~residuals(unml6))
unml9<-lm(residuals(unml3)~residuals(unml7))

summary(unml8)
```

```
## Warning in summary.lm(unml8): essentially perfect fit: summary may be unreliable

##
## Call:
## lm(formula = residuals(unml3) ~ residuals(unml6))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -9.484e-16 -7.192e-17  2.882e-17  1.005e-16  1.153e-15
##
## Coefficients:
##                     Estimate Std. Error    t value Pr(>|t|)
## (Intercept)      -1.574e-17  1.579e-17 -9.970e-01     0.32
## residuals(unml6)  1.000e+00  1.705e-17  5.864e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.227e-16 on 197 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.439e+33 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
summary(unml9)
```

```
##
## Call:
## lm(formula = residuals(unml3) ~ residuals(unml7))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.588e-14 -1.337e-16  1.082e-16  2.834e-16  6.010e-15
##
## Coefficients:
##                     Estimate Std. Error   t value Pr(>|t|)
## (Intercept)      1.417e-16  1.367e-16 1.037e+00    0.301
## residuals(unml7) 1.000e+00  1.476e-16 6.774e+15   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.928e-15 on 197 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.589e+31 on 1 and 197 DF,  p-value: < 2.2e-16
```

The slope from avgplots and from the mean function are 1 with very small p value hence we can conclude that they are identical. Thus residuals from avgplot and mean function are same.

3.2.6

```
summary(unml1)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16313 -0.64507 -0.06586  0.62479  3.00517
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.00967    0.36529   21.93   <2e-16 ***
## log(ppgdp)  -0.62009    0.04245  -14.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9305 on 197 degrees of freedom
## Multiple R-squared:   0.52,  Adjusted R-squared:  0.5175
## F-statistic: 213.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
summary(unml7)
```

```
##
## Call:
## lm(formula = residuals(unml2) ~ residuals(unml5), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.986e-16  6.596e-02   0.000        1
## residuals(unml5) -6.151e-01  6.399e-02  -9.613   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9305 on 197 degrees of freedom
```

```
## Multiple R-squared:  0.3193, Adjusted R-squared:  0.3158
## F-statistic:  92.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
summary(unml3)
```

```
##
## Call:
## lm(formula = fertility ~ log(ppgdp) + pctUrban, data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15114 -0.64929 -0.06604  0.63253  2.99102
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.9932699  0.3993367  20.016   <2e-16 ***
## log(ppgdp)  -0.6151425  0.0641565  -9.588   <2e-16 ***
## pctUrban    -0.0004393  0.0042656  -0.103    0.918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9328 on 196 degrees of freedom
## Multiple R-squared:   0.52,  Adjusted R-squared:  0.5151
## F-statistic: 106.2 on 2 and 196 DF,  p-value: < 2.2e-16
```

As evident from the summary , all the 3 values of t test varies. Since another term is added i.e pcturban, the variation of data will be explained by that added term and hence the variance will reduce and so will tvalue. hence we see a reduction from -14.61 to -9.63

Answer 3.4

3.4.1

The line will be parallel to y axis since x1 is fixed , thus x2 will be fixed to becoz of linear relationship

3.4.2

it will be something similar to null plot since there is a relation between y and x with no error hence the residuals wil be in a straight line.

3.4.3

under the condition of coefficient of determination (R squared) for Y~X1 is near 0 , which means x1 addds no variability to Y, then it will be same as Y~X2

3.4.4

True

Answer 4

if Pr(> |t|) = .08 then the p-value will be 0.08.

Answer 5

One applied situation where mulitple regeression will be helpful in the case of total sales of a particular grocery stores. The predictors could be the population density near store location and the range of items. hence the response would be sales and the predictors would be population density and range of items sold. A step by step multiple regression would help us determining where it would be prudent to open a store. it can show a relation between opening a store in a populous location versus having a wide range of items.

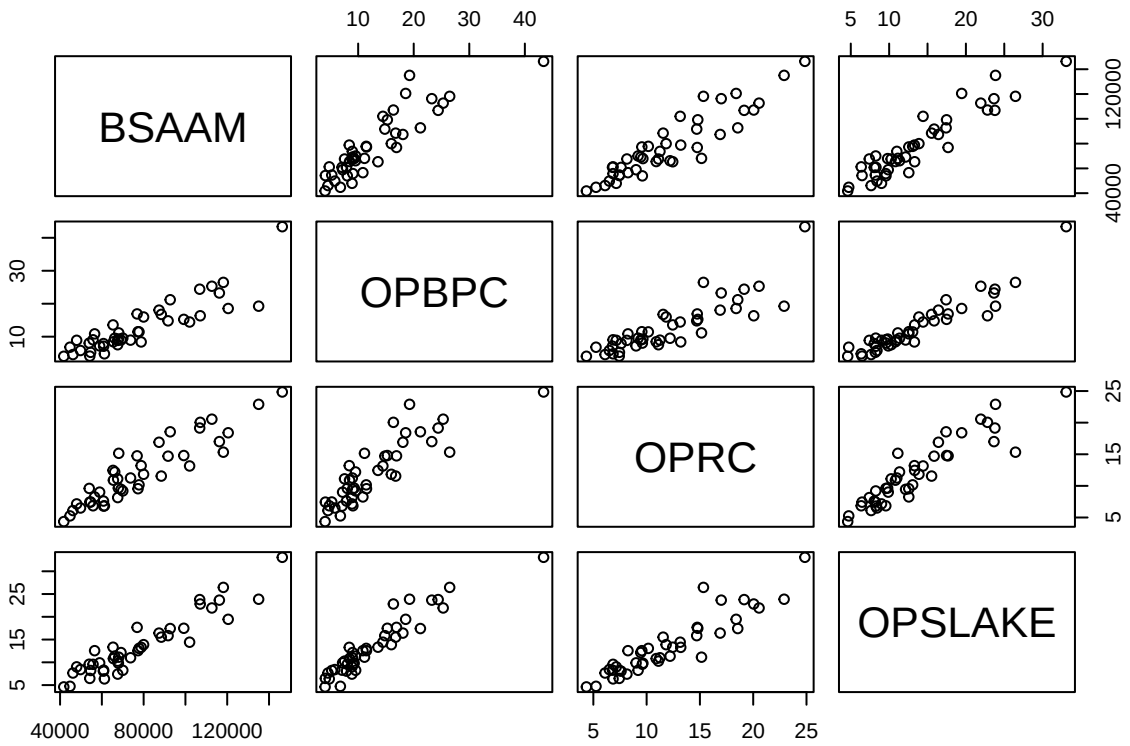
Answer 3.6

3.6.1

```
colnames(water)
```

```
## [1] "Year"    "APMAM"   "APSAB"   "APSLAKE" "OPBPC"   "OPRC"    "OPSLAKE"
## [8] "BSAAM"
```

```
ans361<-pairs(~ BSAAM+OPBPC+OPRC+OPSLAKE, data=water)
```



```
cormatrix<-water%>%select(BSAAM,OPBPC,OPRC,OPSLAKE)%>%cor()
cormatrix
```

```
##             BSAAM     OPBPC      OPRC   OPSLAKE
## BSAAM   1.0000000 0.8857478 0.9196270 0.9384360
## OPBPC   0.8857478 1.0000000 0.8647073 0.9433474
## OPRC    0.9196270 0.8647073 1.0000000 0.9191447
## OPSLAKE 0.9384360 0.9433474 0.9191447 1.0000000
```

As per the scatterplot matrix, opbpc and opslake fits each other fantastically along the line y=x with good variation across the line. In general a lot of the predictors fit each other very well. OPBPC and OPRC also fits very well but not as well as the previous example. OPSlake fits all the other predictors very well , with a model type y=x i.e a positive slope. These ideas are reinforces with the correlation matrix as wel. The correlation matrix row of opslake has value equal to 0.9ish which shows a strong correlation.

3.6.2

```
unml12<-lm(BSAAM~OPBPC+OPRC+OPSLAKE,water)

summary(unml12)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15964.1  -6491.8   -404.4   4741.9  19921.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
## OPBPC          40.61     502.40   0.081  0.93599
## OPRC         1867.46     647.04   2.886  0.00633 **
## OPSLAKE      2353.96     771.71   3.050  0.00410 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8304 on 39 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8941
## F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

The t value columnn shows the associated t test with the significance of estimated parameters. That means a t-test was performed on the estimators (B0,B1)etc which can be used to accept or reject the hypothesis.