PSTAT 126 - Assignment 4

Fall 2022

Due: Tuesday, October 25 at 11:59 pm on Canvas

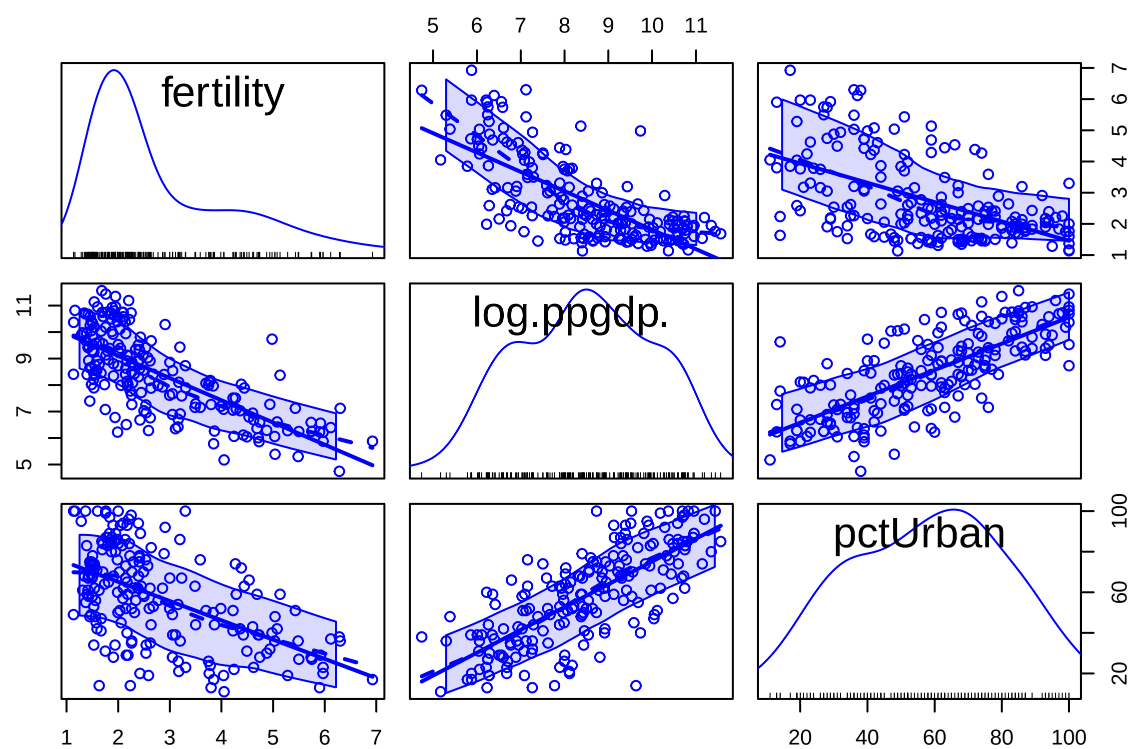
Note: ***Submit both your Rmd and generated pdf file to Canvas.*** *Use the same indentation level as* ***Solution*** *markers to write your solutions. Improper indentation will break your document.*

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

1. This problem uses the data set `UN11` from the `alr4` package.

   (a) Examine the figure generated by using `scatterplotMatrix` function for attributes (`fertility`, log(`ppgdp`), `pctUrban`), and comment on the marginal relationships.
   Solution:

```
scatterplotMatrix(~fertility+log(ppgdp)+pctUrban, data= UN11)
```



Fertility and log(ppgdp) have a negative correlation, Fertility and pctUrban have a negative correlation, log(ppgdp) and pctUrban have a positive correlation.

```
(b) Fit the two simple regressions ‘fertility‘ $\sim$ log(‘ppgdp‘) and ‘fertility‘ $\sim$ ‘pctUrban‘, ar

    **Solution**:
```

```
fit1 <- lm(fertility ~ log(ppgdp), data = UN11)
fit2 <- lm(fertility ~ pctUrban,data = UN11)
summary(fit1)
```

```
##
## 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(fit2)
```

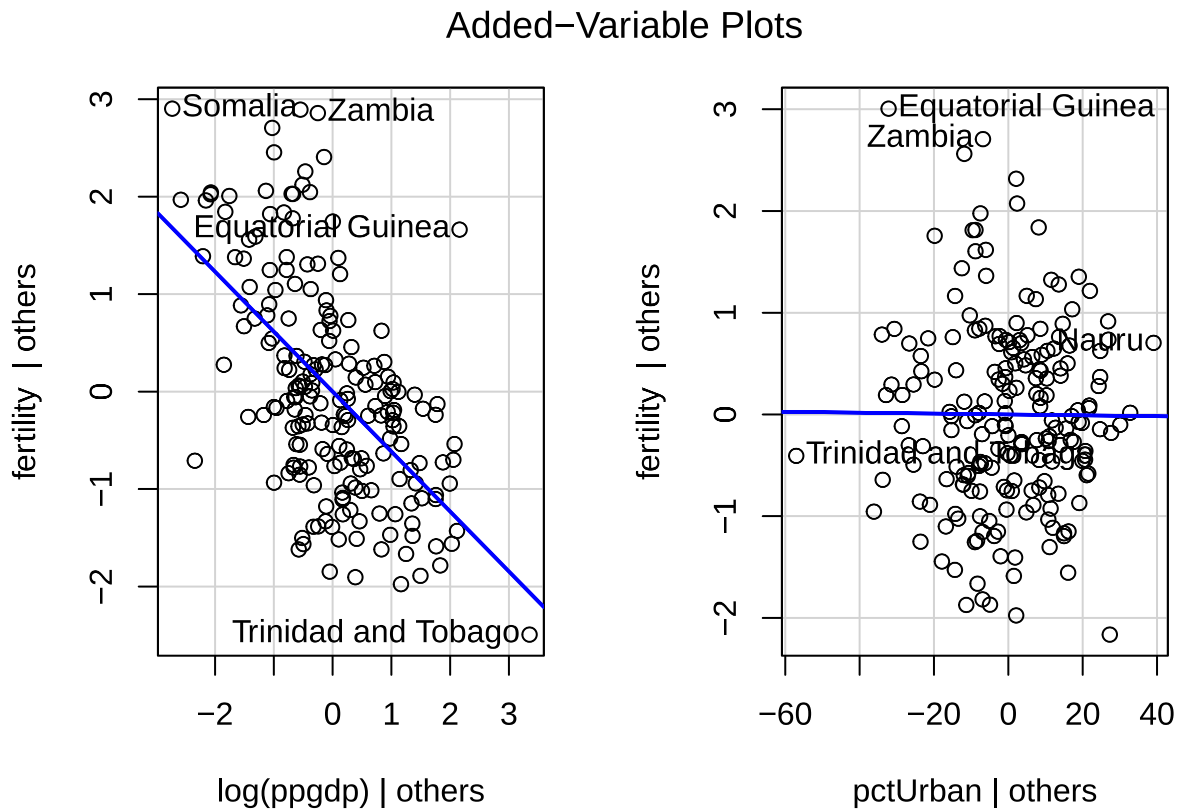
```
##
## Call:
## lm(formula = fertility ~ pctUrban, data = UN11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4932 -0.7795 -0.1475  0.6517  2.9029
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.559823   0.213681  21.339   <2e-16 ***
## 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
```

Observing the above summaries, Yes, the slopes are significantly different than zero at any conventional level.

```
(c) Obtain the added-variable plots for both predictors. Based on the added-variable plots, does log(‘p

    **Solution**:
```

```r
fit <- lm(fertility ~ log(ppgdp)+pctUrban,data = UN11)
avPlots(fit)
```



The av plot for log(ppgdp) after adjusting for pctUrban is useful as it maintains a steep slope, while pctUrban after adjusting for log(ppgdp) has a neutral slope and is not useful.

2. Consider a multiple linear regression model with two continuous predictors:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

   (a) Suppose that x_{i1} and x_{i2} are exactly related in that $x_{i1} = 2.2x_{i2}$ for all i. For example, x_{i2} could be weight in kilograms and x_{i1} weight in pounds for the i-th individual. Describe the appearance of the added variable plot for x_{i2} after adjusting for x_{i1}.
   Solution:

Since xi2 is a linear function of xi1 and are exactly related, the residuals are zero and therefor the avPlot will look like a vertical line.

```
b)  Suppose that $x_{i1}$ and $x_{i2}$ are not perfectly correlated, but that $Y_i = 3x_{i1}$, i.e. $Y_

    **Solution**:
```

Since xi1 and xi2 are not perfectly correlated, the avPlot for xi2 would look like a straight line (no-slope or little-to-no-slope)

```
c)  (**Bonus**): Simulate some data for each of the situations in parts a) and b) and create an added-va

    **Solution**:
```