PSTAT126 Homework3

Yanjie Qi

2019/8/26

1.

(q)

Loading the required data set and print some observations for the dataset prostate:

```r
library(faraway)
data(prostate)
attach(prostate)
head(prostate)
```

```
##       lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
## 1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
## 2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
## 3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
## 4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
## 5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
## 6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547
```

estimate the regression using lm():

```r
m.fit<-lm(lpsa~lcavol,data=prostate)
```

print summary:

```r
summary(m.fit)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50730    0.12194   12.36   <2e-16 ***
## lcavol       0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

produce an ANOVA table:

```r
anova(m.fit)
```

```
## Analysis of Variance Table
##
```

```
## Response: lpsa
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lcavol     1 69.003  69.003  111.27 < 2.2e-16 ***
## Residuals 95 58.915   0.620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b)

since The coeffient of determination–R-squared = 1-sse/ssto = 1-58.915/(69.003+58.915) = 0.5394. The coefficient of determination represents the variability in lpsa that is explained by the model.That is 53.94% of variability in lpsa that is explained by the model;the remaining is 100%-53.94%=46.06%. Therefore, 46.06% of variability in lpsa is left unexplained by the regression model.

2.

Loading the data set baeskel from the alr4 package:

```r
library(alr4)
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

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

## Loading required package: effects

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

##
## Attaching package: 'alr4'

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

```r
data(baeskel)
attach(baeskel)
```

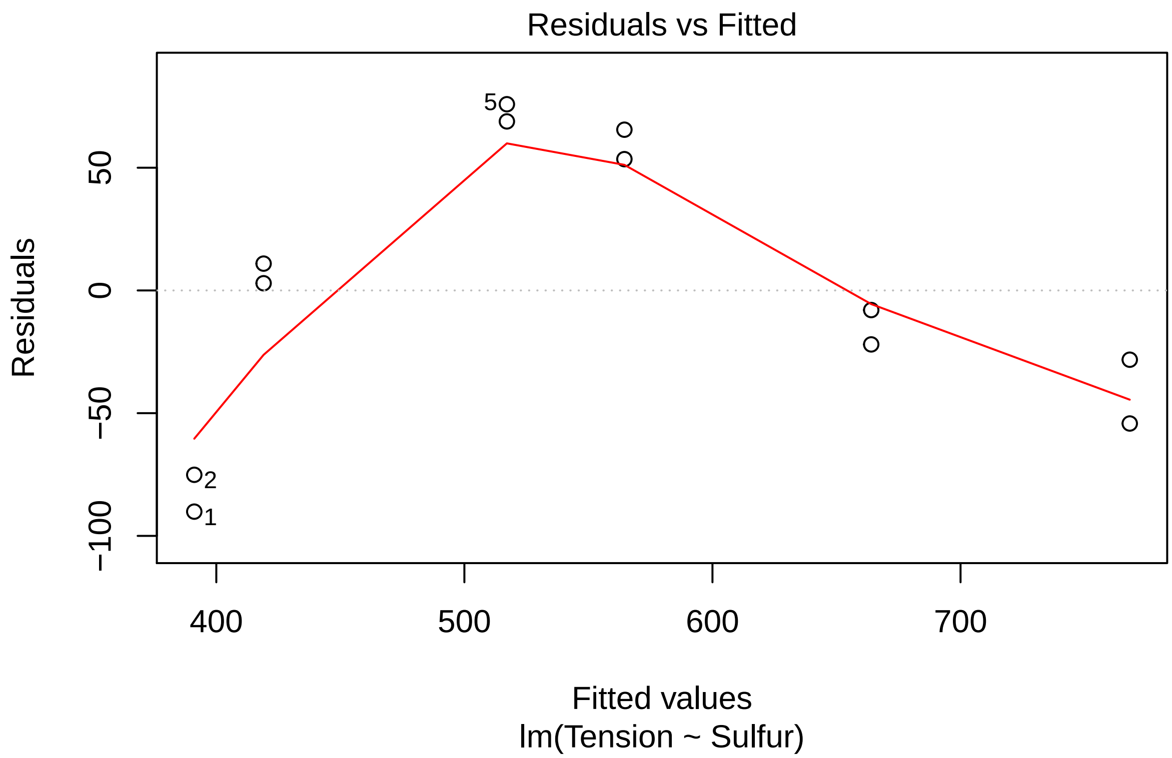
(q)

construct linear model:

```r
mod = lm(Tension~Sulfur, data = baeskel)
```

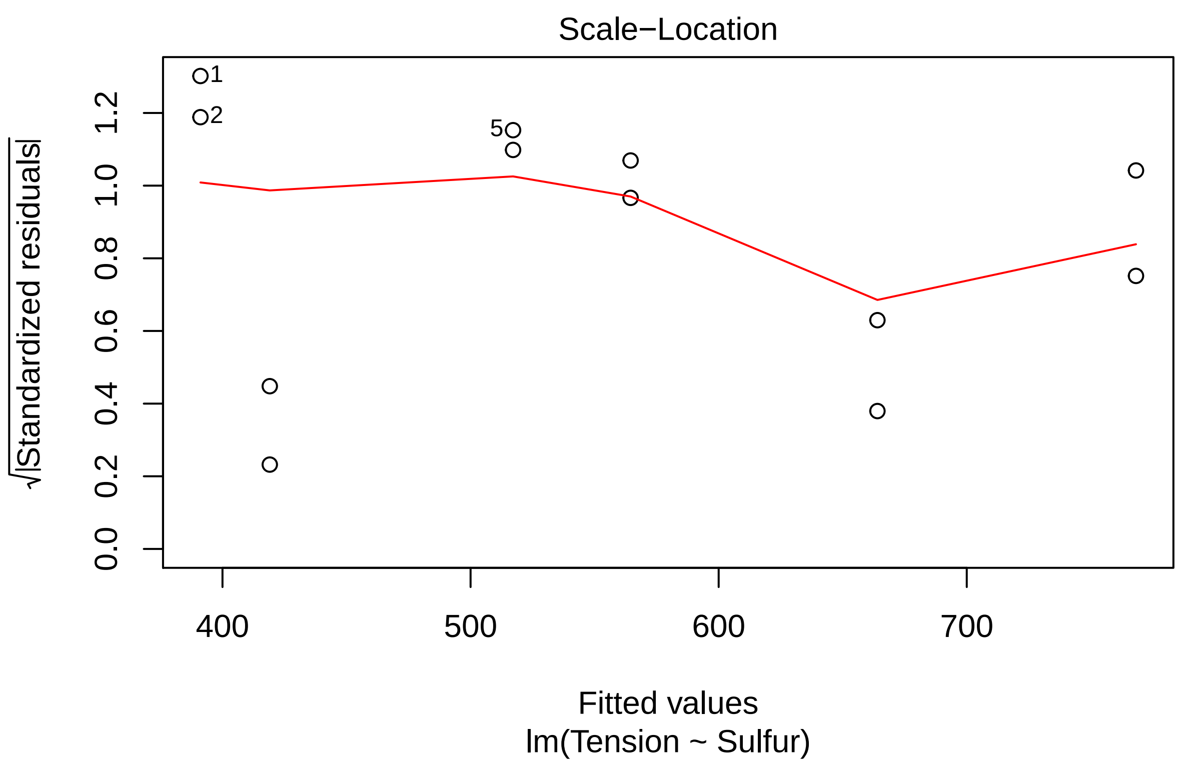
Residuals vs. Fitted Plot:

```r
plot(mod, which = 1)
```



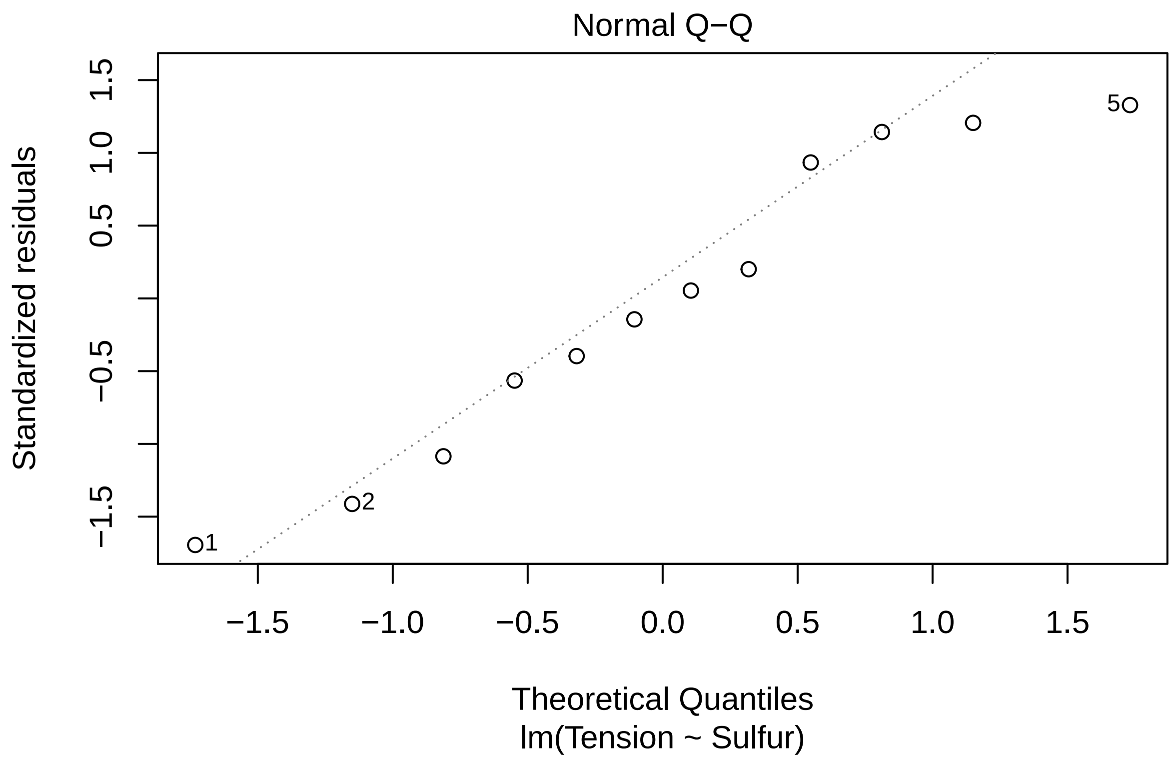
Scale-Location:

```r
plot(mod, which = 3)
```



QQ-Plot:

```r
plot(mod, which = 2)
```

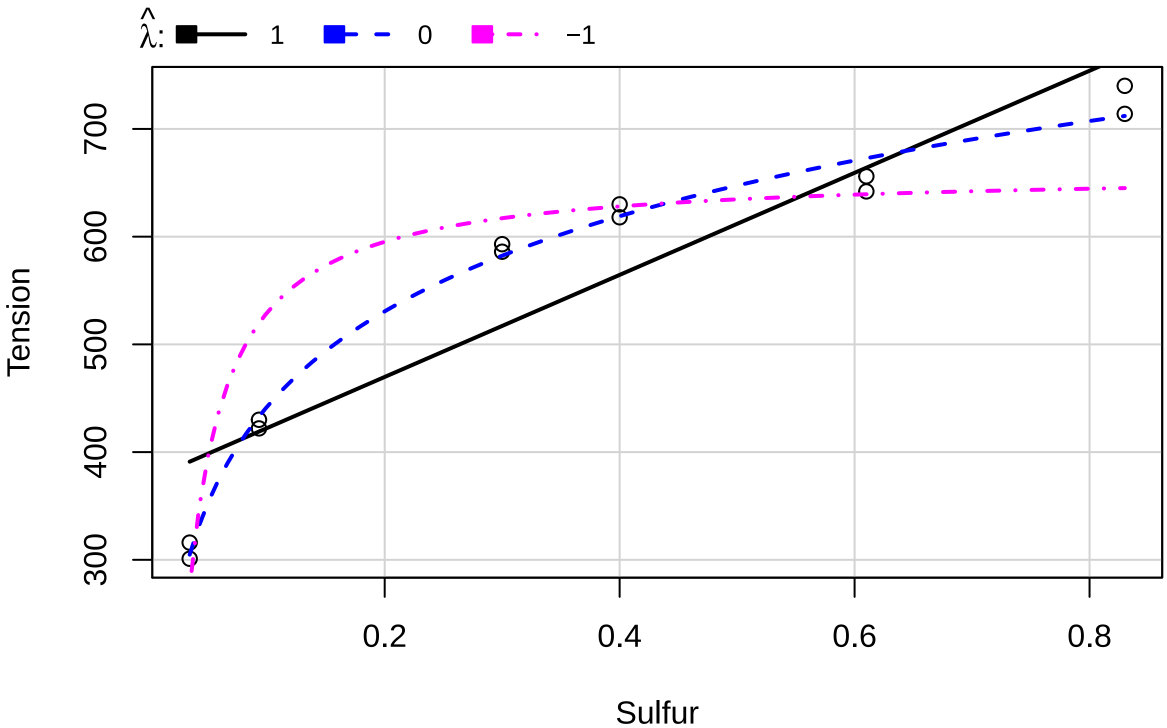


Conclusion: From Risidual vs. Fitted Plot, we could see a parabola, a distinctive pattern which means it is not a standard simple linear model; From QQ-plot, we could see its non-normality due to the graph which is a little bit right-skewed to the residuals.

(b)

fit these two transformations and plot the regression fits along with the part a)

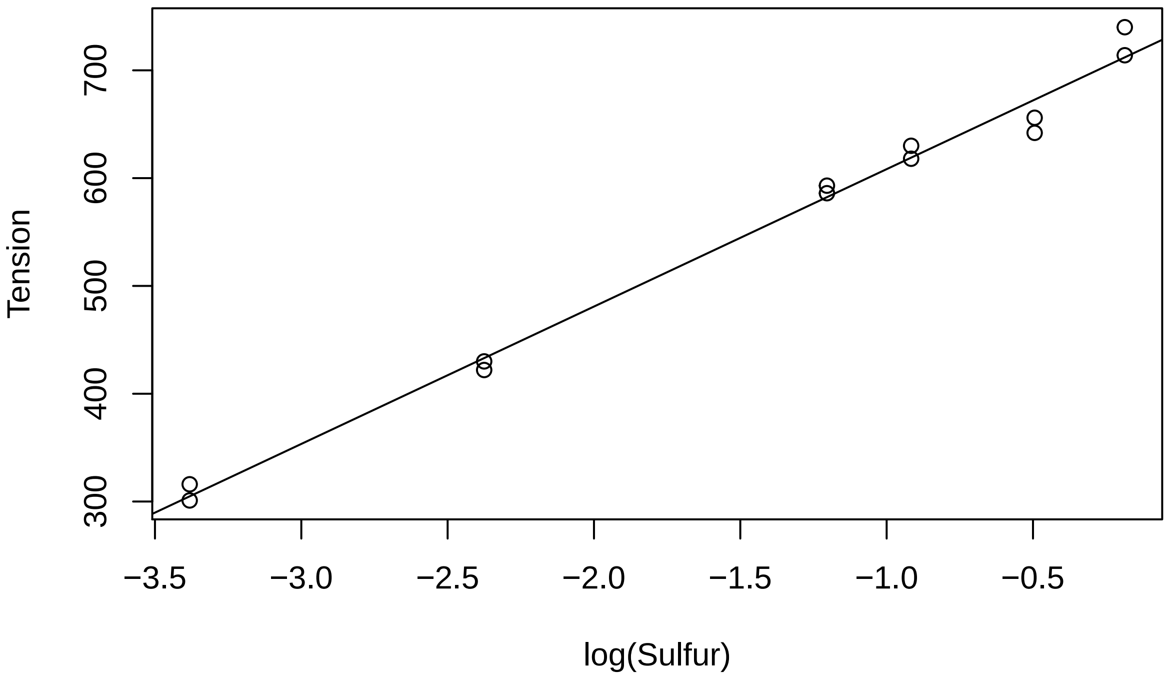
```r
invTranPlot(Tension~Sulfur, lambda = c(1,0,-1), optimal = F)
```



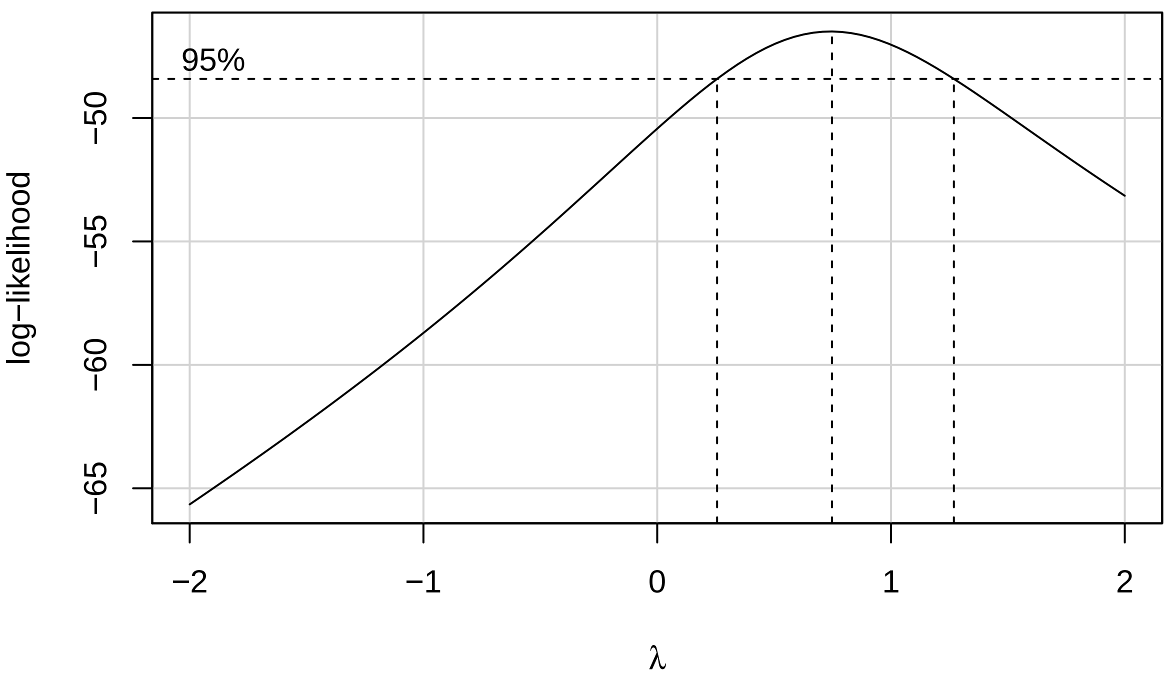

```
##   lambda       RSS
## 1      1 35824.332
## 2      0  2535.896
## 3     -1 35691.735
```

(c)

```
mod2 = lm(Tension~log(Sulfur))
plot(Tension~log(Sulfur))
abline(mod2)
```




```r
bc = boxCox(mod2)
```




```r
lambda.opt = bc$x[which.max(bc$y)]
lambda.opt
```

```
## [1] 0.7474747
```

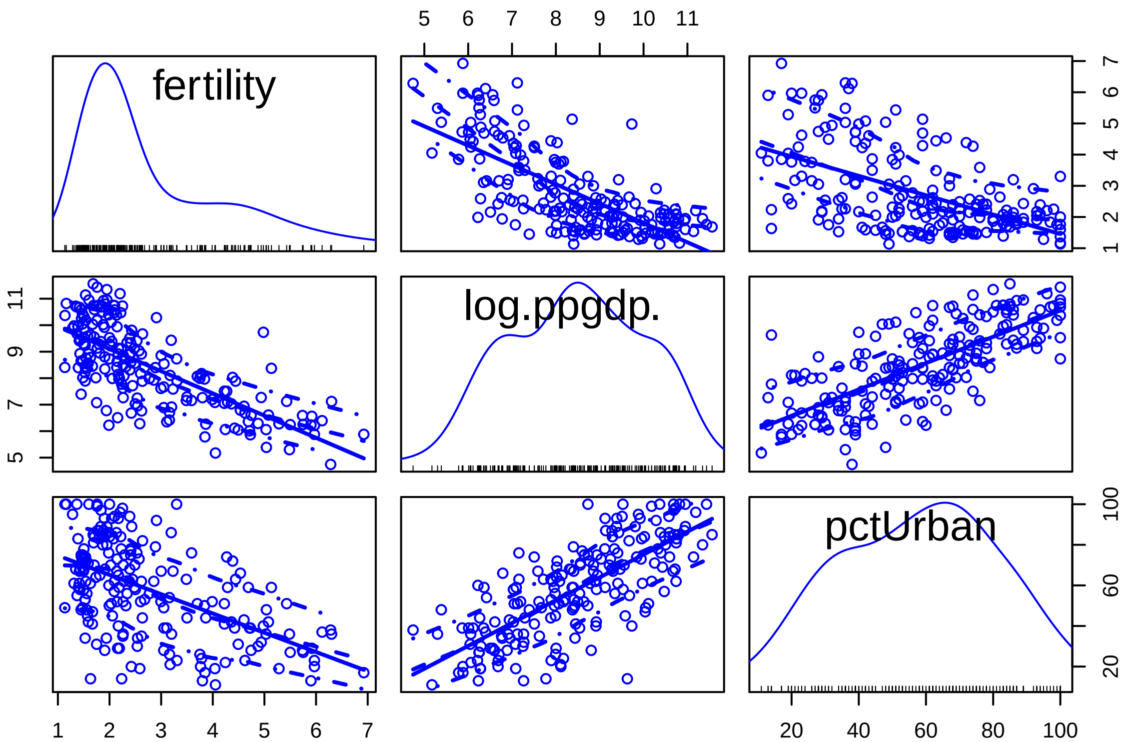
From part b and c, we can see that lambda = 1 and lambda = 0.7474747 is within the confidence interval, so we shouldn't transform the variable.

3.

(a)

We have already had the alr4 package loaded therefore, plot the scatterplot matrix:

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



We could moment from above that: Fertility and log(ppgdp) has a negative correlation; Fertility and pctUrban has a negative correlation; log(ppgdp) and pctUrban has a positive correlation.

(b)

construct OLS regression:

```
ols.fit=lm(fertility~log(ppgdp)+pctUrban,data=UN11)
ols.fit1=lm(fertility~log(ppgdp),data=UN11)
ols.fit2=lm(fertility~pctUrban,data=UN11)
```

Obtain the coefficients and p-values:

```
summary(ols.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
```

```r
summary(ols.fit2)
```

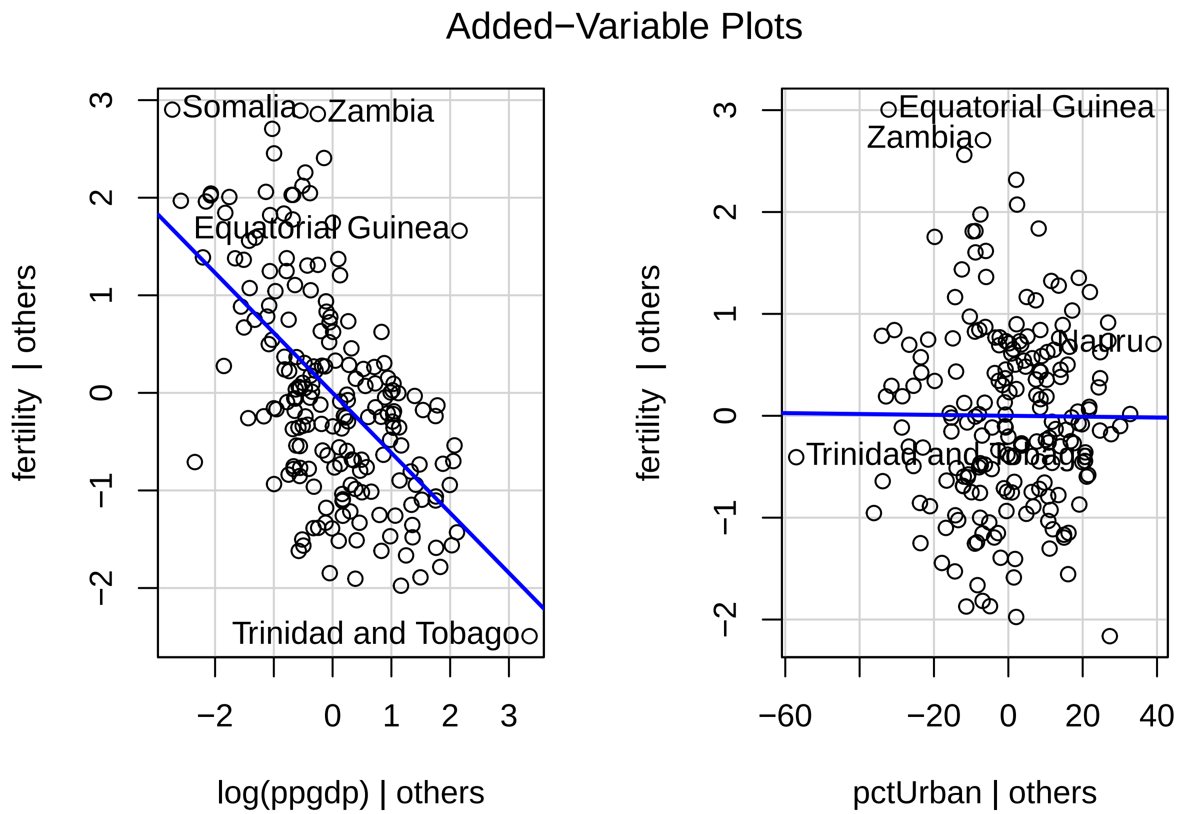
```
##
## Call:
## lm(formula = fertility ~ pctUrban, data = UN11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4932 -0.7795 -0.1475  0.6517  2.9029
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.559823   0.213681  21.339   <2e-16 ***
##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

From the summaries, we can see that the coefficients are both significantly different from zero at any conventional level of significance.

(c)

Obtain the added-variable plots:

```r
avPlots(ols.fit)
```



From above, we can say that log(ppgdp) is useful as it shows a steep slope while pctUrban, which is not useful, is neutral to fertility.