Practice 1 & 2 Solutions Draft

This practice follows your learning of Chapter 1 which is a review chapter. It is worthwhile gaining some practice using `R` on some real data. The real data to be used here is the `swiss` data which can be accessed in `R`. The aim is to build some good linear models to analyze the `swiss` data where `Fertility` is used as the response.

Detailed R commands and numerical and graphical outputs used are given in the Appendix.

1. *An initial data analysis that explores the numerical and graphical characteristics of the data.*

   The `swiss` data contain observations of standardized `Fertility` measure and 5 social-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. The response variable `Fertility` and the 5 social-economic indicators, which will be used as the predictors, are described in the following table:

`Fertility`	Ig, 'common standardized fertility measure'
`Agriculture`:	% of males involved in agriculture as occupation
`Examination`:	% draftees receiving highest mark on army examination
`Education`:	% education beyond primary school for draftees
`Catholic`:	% 'catholic' (as opposed to 'protestant')
`Infant.Mortality`:	live births who live less than 1 year.

   Numerical summary of the data shows that all the 6 variables are numerical with weak to moderate linear correlations among them. A matrix of scatter-plots for the 6 variables indicates `Fertility` has positive correlation with `Agriculture` and `Infant.Mortality`; negative correlation with `Examination` and `Education`; and a curvature correlation with `Catholic`. In addition, is seems the distribution of `Fertility` is not too different from the normal except for small values of `Fertility`.

2. *Variable selection to choose the best model.*

   We start by fitting a linear regression model

   ```
   lmod <- lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, swiss).
   ```

   By both a t-test and an ANOVA F test we find `Examination` does not have significant effect on `Fertility`. We then treat

   `Fertility ~ (Agriculture + Education + Catholic + Infant.Mortality)^2`

   as the full model, and use `step()` with BIC for selecting the best model. The fitted best model is

   $$\widehat{\text{Fertility}} = 53.75 - 0.134\text{Agriculture} - 0.515\text{Education} + 0.207\text{Catholic} + 1.24\text{Infant.Mortality} - 0.011\text{Education:Catholic}$$

   with $R^2 = 0.7318$ and $R^2_a = 0.699$. The terms in this model cannot be further reduced by the `drop1()` command.

3. *An exploration of transformations to improve the fit of the model.*

   It does not seem to need a transformation on the response variable because the empirical distribution of `Fertility` is not far from the normal. On the other hand, the relationship between `Fertility` and `Catholoc` seems to be curvature. Thus, we replace `Catholic` by `poly(Catholic, 2)` or `bs(Catholic, 3)` in the model `smallm` to see whether any improvement of fit can be made. It does not seem to achieve any significant improvement by doing this. Here the order 2 in `poly()` and df 3 in `bs()` are selected using a try-and-error approach.

4. `Diagnostics to check the assumptions of your model.`

   The 4 diagnostics plots given by `plot(smallm)` show that the model provides a good fit to the data in general. The residuals vs. leverage plot identifies 4 provinces that have the largest Cook distance values and are influential to model fitting. These 4 provinces are *Porrentruy*, *Sierre*, *Sion*, and *Rive Gauche*, which have the most extreme residuals in regard to model `smallm`, but do not have large leverage values. The predictors values of these 4 provinces are mostly unusual in comparison with those of other provinces.

5. *Some predictions of future observations for interesting values of the predictors.*

   While there should be many interesting values of the predictors, we chose to predict the `Fertility` value at the mean values of the predictors. The predicted value of `Fertility` equals 69.46289 with standard error 1.045219.

6. *An interpretation of the meaning of the model with respect to the particular area of application.*

   The selected model `smallm` suggests that all predictors except `Examination` are significantly related to `Fertility` with the directions of the relations been given in the `summary(smallm)` output. In addition, `Education` and `Catholic` have significant interaction effect on `Fertility`.

Appendix

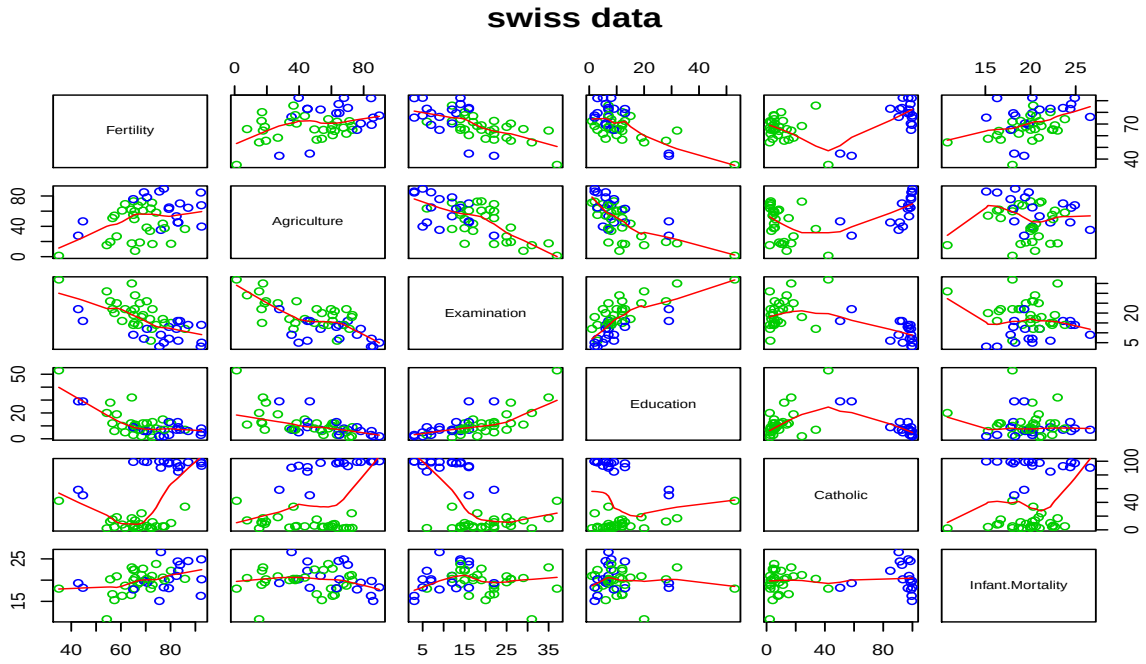
```
library(faraway); require(graphics); data(swiss); help(swiss); dim(swiss); head(swiss)

####1.###
##numerical summary
summary(swiss)
```

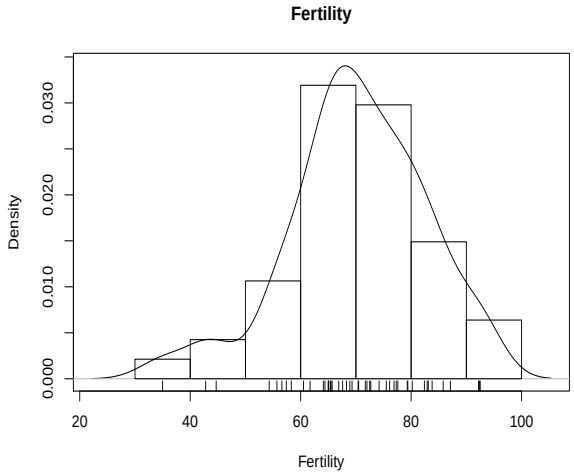
```
   Fertility      Agriculture     Examination      Education        Catholic       Infant.Mortality
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00   Min.   :  2.150   Min.   :10.80
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00   1st Qu.:  5.195   1st Qu.:18.15
 Median :70.40   Median :54.10   Median :16.00   Median : 8.00   Median : 15.140   Median :20.00
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98   Mean   : 41.144   Mean   :19.94
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00   3rd Qu.: 93.125   3rd Qu.:21.70
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00   Max.   :100.000   Max.   :26.60
```

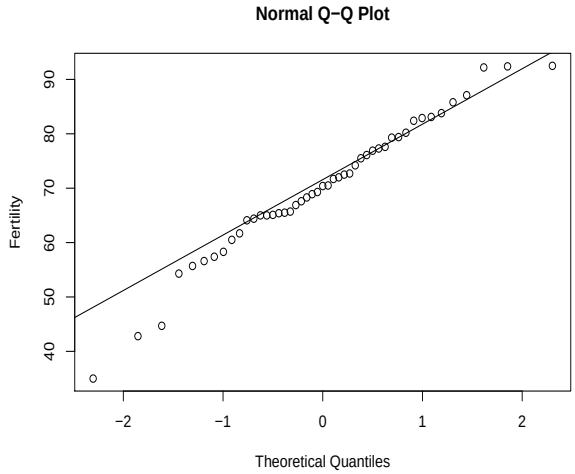
```
cor(swiss)
```

```
                 Fertility Agriculture Examination   Education   Catholic Infant.Mortality
Fertility        1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847       0.41655603
Agriculture      0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951      -0.06085861
Examination     -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418      -0.11402160
Education       -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589      -0.09932185
Catholic         0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000       0.17549591
Infant.Mortality 0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959       1.00000000
```

```
#graphical summary
pairs(swiss, panel = panel.smooth, main = "swiss data", col = 3 + (swiss$Catholic > 50))

plot(density(swiss$Fertility),main="Fertility",xlab="Fertility")
rug(swiss$Fertility)
hist(swiss$Fertility,freq=F,add=T)

qqnorm(swiss$Fertility, ylab="Fertility")
qqline(swiss$Fertility)
```



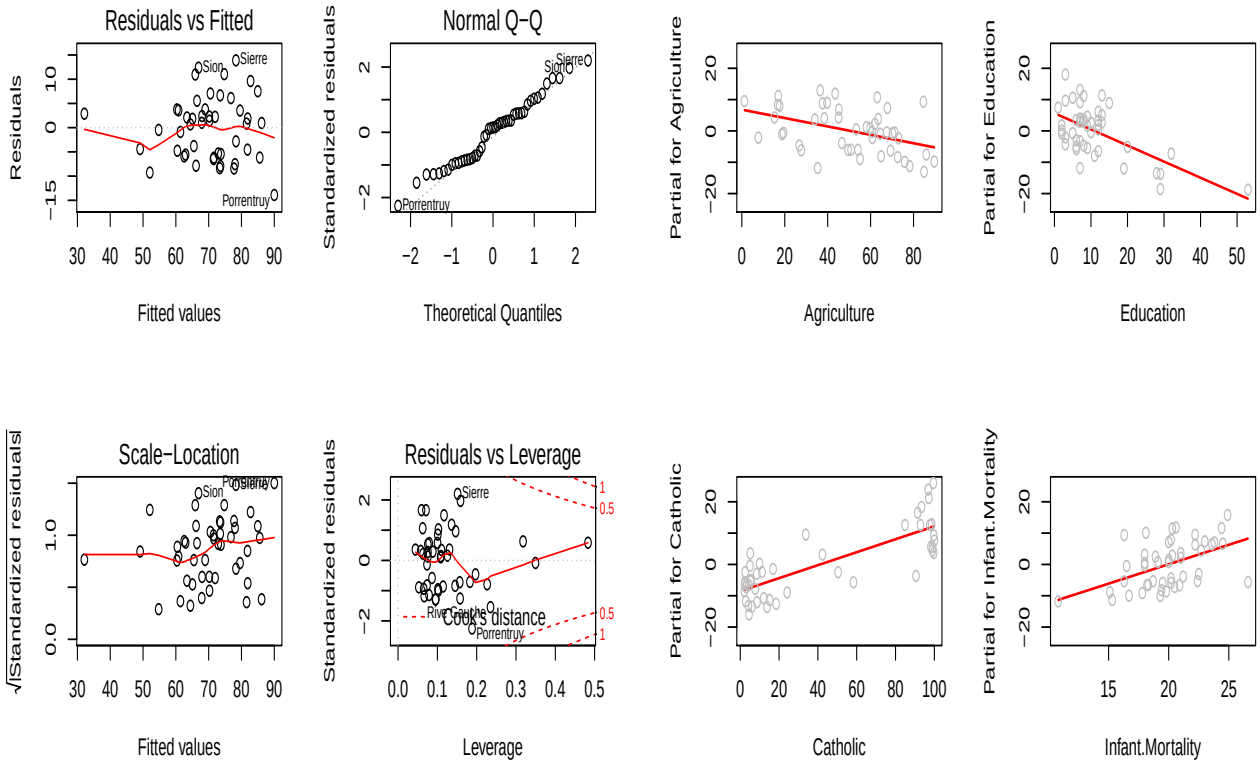





```
####2.
lmod<-lm(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality, swiss)
summary(lmod); drop1(lmod, test="F")
lmod1<-lm(Fertility~Agriculture+Education+Catholic+Infant.Mortality, swiss)
summary(lmod1); anova(lmod1, lmod)

lmodi<-lm(Fertility~(Agriculture+Education+Catholic+Infant.Mortality)^2, swiss)
smallm <- step(lmodi,trace=FALSE, k=log(47)) #BIC. AIC if k=2
summary(smallm)

Call: lm(formula = Fertility ~ Agriculture + Education + Catholic +
    Infant.Mortality + Education:Catholic, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-13.9060  -5.4997   0.9556   3.6698  13.8934

Coefficients:
```

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        53.752308   9.919330   5.419 2.89e-06 ***
Agriculture        -0.134055   0.065843  -2.036  0.04825 *
Education          -0.515105   0.252478  -2.040  0.04781 *
Catholic            0.207038   0.046184   4.483 5.81e-05 ***
Infant.Mortality    1.239697   0.372195   3.331  0.00184 **
Education:Catholic -0.011255   0.005058  -2.225  0.03161 *
---
Residual standard error: 6.853 on 41 degrees of freedom
Multiple R-squared:  0.7318,    Adjusted R-squared:  0.699
F-statistic: 22.37 on 5 and 41 DF,  p-value: 9.443e-11
```

```
par(mfrow=c(2,2)); termplot(smallm,partial=T,terms=NULL); plot(smallm)
```




```
####3.
library(MASS)
Wlmodp<-lm(Fertility~Agriculture+Education+poly(Catholic,2)+Infant.Mortality + Education:poly(Catholic,2), swiss)
summary(Wlmodp)

Wlmodp1<-lm(Fertility~Agriculture + Education + poly(Catholic,2) + Infant.Mortality + Education:Catholic, swiss)
summary(Wlmodp1)

Call:
lm(formula = Fertility ~ Agriculture + Education + poly(Catholic,
    2) + Infant.Mortality + Education:Catholic, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8316  -5.2273   0.2632   4.0651  14.2838

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)        61.826469   9.825002   6.293 1.83e-07 ***
Agriculture        -0.152171   0.068575  -2.219 0.032221 *
Education          -0.517682   0.252752  -2.048 0.047145 *
poly(Catholic, 2)1 55.884416  13.372902   4.179 0.000155 ***
poly(Catholic, 2)2  9.820777  10.261947   0.957 0.344311
Infant.Mortality    1.269834   0.373906   3.396 0.001556 **
Education:Catholic -0.009239   0.005484  -1.685 0.099837 .
---
Residual standard error: 6.86 on 40 degrees of freedom
Multiple R-squared:  0.7378,    Adjusted R-squared:  0.6984
F-statistic: 18.75 on 6 and 40 DF,  p-value: 3.078e-10
```

```
plot(Wlmodp1)
termplot(Wlmodp,partial=T,terms=NULL)

library(splines)
Wlmods<-lm(Fertility~Agriculture+Education+bs(Catholic,3)+Infant.Mortality + Education:bs(Catholic,3), swiss)
summary(Wlmods)

Wlmods1<-lm(Fertility~Agriculture+Education+bs(Catholic,3)+Infant.Mortality + Education:Catholic, swiss)
summary(Wlmods1)
```

```
Call:
lm(formula = Fertility ~ Agriculture + Education + bs(Catholic,
    3) + Infant.Mortality + Education:Catholic, data = swiss)

Residuals:
    Min      1Q  Median      3Q     Max
-14.377  -5.072   0.321   4.014  14.446

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        55.455117  10.117571   5.481 2.72e-06 ***
Agriculture        -0.148741   0.070264  -2.117 0.040702 *
Education          -0.479725   0.284130  -1.688 0.099316 .
bs(Catholic, 3)1   -3.217588  11.717565  -0.275 0.785077
bs(Catholic, 3)2   11.056766  19.100029   0.579 0.565994
bs(Catholic, 3)3   19.202744   4.706975   4.080 0.000216 ***
Infant.Mortality    1.259959   0.379587   3.319 0.001964 **
Education:Catholic -0.010382   0.006687  -1.553 0.128614
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.939 on 39 degrees of freedom
Multiple R-squared:  0.7384,    Adjusted R-squared:  0.6914
F-statistic: 15.72 on 7 and 39 DF,  p-value: 1.35e-09
```

```
plot(Wlmods1)
termplot(Wlmods,partial=T,terms=NULL)


########4.
(1:47)[rownames(swiss)=="Sion"]  #38
(1:47)[rownames(swiss)=="Sierre"] #37
(1:47)[rownames(swiss)=="Porrentruy"] #6
(1:47)[rownames(swiss)=="Rive Gauche"] #47

swiss[c(6,37,38,47),]
rownames(swiss)[c(6,37,38,47)]
hatvalues(smallm)
influence.measures(smallm)
cooks.distance(smallm)
```

```
  Courtelary     Delemont Franches-Mnt      Moutier   Neuveville   Porrentruy
0.0201154852 0.0039410826 0.0366691513 0.0292755138 0.0351584235 0.1960420472
       Broye        Glane      Gruyere       Sarine      Veveyse        Aigle
0.0013622033 0.0492469430 0.0002214636 0.0265934794 0.0004467372 0.0047153887
     Aubonne     Avenches     Cossonay    Echallens     Grandson     Lausanne
```

```
0.0010608615 0.0005897404 0.0052804731 0.0090110500 0.0005400547 0.0305009295
   La Vallee       Lavaux       Morges       Moudon        Nyone         Orbe
0.0006271319 0.0002206554 0.0010221776 0.0292633997 0.0073849102 0.0160432221
        Oron      Payerne Paysd'enhaut        Rolle        Vevey      Yverdon
0.0032624985 0.0024832884 0.0136542028 0.0002399045 0.0167444949 0.0101748505
     Conthey    Entremont       Herens     Martigwy      Monthey   St Maurice
0.0086585000 0.0292802881 0.0161075451 0.0187071209 0.0169381979 0.0190298484
      Sierre         Sion       Boudry La Chauxdfnd     Le Locle    Neuchatel
0.1441475824 0.1222319428 0.0009958496 0.0494736931 0.0062623898 0.0309042646
  Val de Ruz ValdeTravers V. De Geneve  Rive Droite  Rive Gauche
0.0126224991 0.0199766071 0.0534481565 0.0189033468 0.1224375602
```

```
sort(cooks.distance(smallm))
```

```
      Lavaux      Gruyere        Rolle      Veveyse     Grandson     Avenches
0.0002206554 0.0002214636 0.0002399045 0.0004467372 0.0005400547 0.0005897404
   La Vallee       Boudry       Morges      Aubonne        Broye      Payerne
0.0006271319 0.0009958496 0.0010221776 0.0010608615 0.0013622033 0.0024832884
        Oron     Delemont        Aigle     Cossonay     Le Locle        Nyone
0.0032624985 0.0039410826 0.0047153887 0.0052804731 0.0062623898 0.0073849102
     Conthey    Echallens      Yverdon   Val de Ruz Paysd'enhaut         Orbe
0.0086585000 0.0090110500 0.0101748505 0.0126224991 0.0136542028 0.0160432221
      Herens        Vevey      Monthey     Martigwy  Rive Droite   St Maurice
0.0161075451 0.0167444949 0.0169381979 0.0187071209 0.0189033468 0.0190298484
ValdeTravers   Courtelary       Sarine       Moudon      Moutier    Entremont
0.0199766071 0.0201154852 0.0265934794 0.0292633997 0.0292755138 0.0292802881
    Lausanne    Neuchatel   Neuveville Franches-Mnt        Glane La Chauxdfnd
0.0305009295 0.0309042646 0.0351584235 0.0366691513 0.0492469430 0.0494736931
V. De Geneve         Sion  Rive Gauche       Sierre   Porrentruy
0.0534481565 0.1222319428 0.1224375602 0.1441475824 0.1960420472
```

```
#########5
pdf <- data.frame(Agriculture=mean(swiss$Agriculture),  Examination=mean(swiss$Examination),
        Education=mean(swiss$Education), Catholic=mean(swiss$Catholic),  Infant.Mortality=mean(swiss$Infant.Mortality))
pp <- predict(smallm,new=pdf, se.fit=T); pp
```

```
$`fit`   69.46289
$se.fit  1.045219
$df 41
$residual.scale  6.852949
```