Statistical Modeling I
Practical in R – Output

Practical in R – Output

In this practical, we will work with the Swiss dataset provided with R. We will look at different subset models and their analysis.

Looking at the Swiss data in R, we find that it contains standardized fertility measure and socio-economic indicators for each of the 47 French-speaking provinces of Switzerland in about 1888. It is composed of 47 observations on 6 variables, each of which is a percentage (i.e. in [0,100]), including:

- Fertility
- Agriculture, % of men involved in agriculture as occupation;
- Examination, % draftees receiving highest mark on army examination;
- Education, % of education beyond primary school for draftees;
- Catholic, % of catholic (as opposed to protestant);
- Infant mortality, % of live births who live less than 1 year.

We are interested in predicting the infant mortality using multiple regression models.

1. First of all, we load the data by using the command data:

```
> data(swiss)
> head(swiss)
             Fertility Agriculture Examination Education Catholic Infant.Mortality
Courtelary        80.2        17.0          15        12     9.96             22.2
Delemont          83.1        45.1           6         9    84.84             22.2
Franches-Mnt      92.5        39.7           5         5    93.40             20.2
Moutier           85.8        36.5          12         7    33.77             20.3
Neuveville        76.9        43.5          17        15     5.16             20.6
Porrentruy        76.1        35.3           9         7    90.57             26.6
> ?swiss
```

   with the last command we have a look at the help, where all the informations related to the dataset are available. Then, we look at the plot of the data by using a command that creates different scatterplot across the variables of interest:

```
> pairs(swiss)
```

   Figure 1.1 shows the scatterplot between the six variables of interest and this plot gives the idea of the relations between the variables.
   This can be confirmed by using the correlation command across the variables:

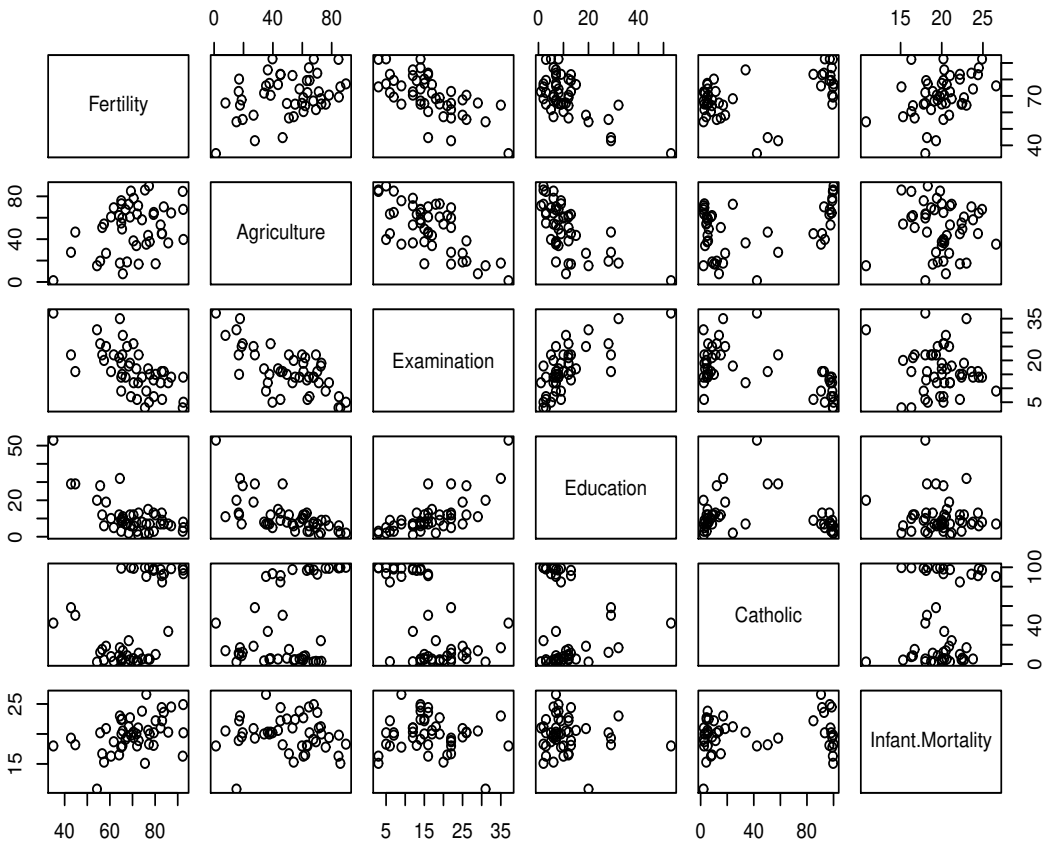

Figure 1.1: Scatterplot of the data

```
> 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
```

As stated in Figure 1.1, the Infant mortality is positively correlated with Fertility (0.41) and with being Catholic (0.17), while is negatively correlated with Examination (-0.11); with Education (-0.099) and with Agriculture (-0.06). On the other hand, fertility is strongly positively correlated with being Catholic (0.46) and with Agriculture (0.35), while it is strongly negatively correlated with Education (-0.66) and Examination (-0.64)

2. Subsequently, we look at the multiple regression models, where the response variable is the infant mortality and we consider a maximum number of 5 predictors. Thus, the R command used is

```
> best.subset <- regsubsets(Infant.Mortality~., swiss, nvmax=5)
```

```
> best.subset.summary <- summary(best.subset)
> best.subset.summary
Subset selection object
Call: regsubsets.formula(Infant.Mortality ~ ., swiss, nvmax = 5)
5 Variables  (and intercept)
            Forced in Forced out
Fertility       FALSE      FALSE
Agriculture     FALSE      FALSE
Examination     FALSE      FALSE
Education       FALSE      FALSE
Catholic        FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         Fertility Agriculture Examination Education Catholic
1  ( 1 ) "*"       " "         " "         " "       " "
2  ( 1 ) "*"       " "         " "         "*"       " "
3  ( 1 ) "*"       "*"         " "         "*"       " "
4  ( 1 ) "*"       "*"         "*"         "*"       " "
5  ( 1 ) "*"       "*"         "*"         "*"       "*"
```

(for more details on this function, please refer to the help in R). If one want to look at the selection algorithm, it is possible to use the `outmat` selection in summary:

```
> best.subset.summary$outmat
         Fertility Agriculture Examination Education Catholic
1  ( 1 ) "*"       " "         " "         " "       " "
2  ( 1 ) "*"       " "         " "         "*"       " "
3  ( 1 ) "*"       "*"         " "         "*"       " "
4  ( 1 ) "*"       "*"         "*"         "*"       " "
5  ( 1 ) "*"       "*"         "*"         "*"       "*"
```

This function shows the best subset of predictors for 1 to 5 predictor models. For example, the best model with two variables includes Fertility and Education as predictors for the infant mortality (see second row from the top). Another detail is the inclusion of fertility in all the models from 1 to 5 predictors and a second detail is related to the variable Education, which is included in all the models with at least two variables. Thus it seems that two important variables are Fertility and Education for predicting the infant mortality, while being Catholic is not an important variable.

In order to see the best model, we need to check the adjusted R^2 and not the simple R^2 since we are comparing models with different number of exogenous variables.

```
> best.subset.summary$adjr2
[1] 0.1551527 0.1946250 0.1875454 0.1719056 0.1517086
```

In this case, we have the adjusted R^2 for all the models, where the first number refers to the model with 1 exogenous variable, the second with 2 exogenous variable and the last with 5 exogenous variables. In case we have a huge amount of numbers, we can use the `which.max` function that show the maximum value across the five adjusted R^2:

```
> best.subset.by.adjr2 <- which.max(best.subset.summary$adjr2)
> best.subset.by.adjr2
[1] 2
```

Thus we can see that the best model regarding the adjusted R^2 is the model with two explanatory variables followed by the model with three explanatory variables.

As another metric of interest, we can use the Mallow's C_p by using the following command:

```
 > best.subset.summary$cp
[1] 1.8173002 0.7739736 2.1834297 4.0000213 6.0000000
```

Contrary to the case of the adjsuted R^2, we need to find the minimum value and thus we use the following command:

```
> best.subset.by.cp <- which.min(best.subset.summary$cp)
> best.subset.by.cp
[1] 2
```

As for the $adj(R^2)$, we have that the best model is the model that include two explanatory variables. Note that the first model (the one with one regressors) has a value near 2, the number of parameters.

3. As a further step, we can also look at the plot of the adjusted R^2 and of the Mallow's statistic in case we have different models to compare. The command used for the plots are the following:

```
> plot(best.subset.summary$adjr2, xlab="Number of Variables",
+      ylab="Adjusted RSq", type="l")
> plot(best.subset.summary$cp, xlab="Number of Variables",
+      ylab="CP", type="l")
```

In Figure 1.2, we shows the adjusted R^2 across the models (left panel) and the Mallow's C_k statistics across models (right panel). As stated above, we can see that the maximum value is in the second model for the left panel, while the minimum value is in the second model for the right panel.

As stated above, the best model is the model with 2 variables, thus we need to look at the estimated coefficients of the best model:

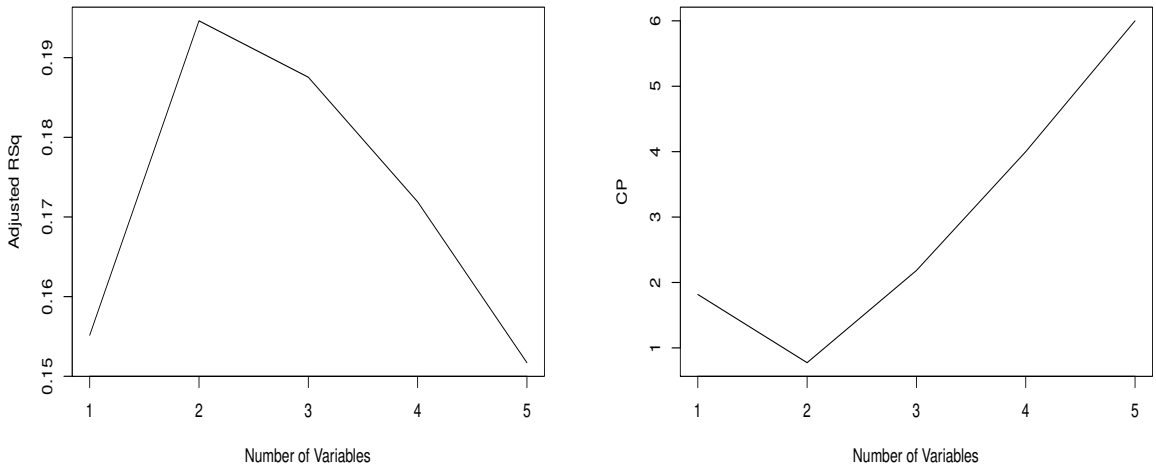

Figure 1.2: Plot of adjusted R^2 (left) and of the Mallow's C_k statistic (right) across the models.

```
> coef(best.subset,2)
(Intercept)   Fertility   Education
 8.63757624  0.14615350  0.09594897
```

4. In order to check the two models, first of all we run the linear regression model with two explanatory variables:

```
> mod <- lm(swiss$Infant.Mortality ~ swiss$Fertility
+ swiss$Education )
> summary(mod)

Call:
lm(formula = swiss$Infant.Mortality ~ swiss$Fertility
+ swiss$Education)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6927 -1.4049  0.2218  1.7751  6.1685

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.63758    3.33524   2.590 0.012973 *
swiss$Fertility  0.14615    0.04125   3.543 0.000951 ***
swiss$Education  0.09595    0.05359   1.790 0.080273 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.614 on 44 degrees of freedom
Multiple R-squared:  0.2296,Adjusted R-squared:  0.1946
F-statistic: 6.558 on 2 and 44 DF,  p-value: 0.003215
```

Then we run the anova table for this model by using the following commands:

```
 > anova(mod)
Analysis of Variance Table

Response: swiss$Infant.Mortality
                 Df  Sum Sq Mean Sq F value   Pr(>F)
swiss$Fertility   1  67.717  67.717  9.9108 0.002949 **
swiss$Education   1  21.902  21.902  3.2055 0.080273 .
Residuals        44 300.636   6.833
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Looking at the linear regression, the fertility variable is statistically significant ($p - value = 0.00095$), while the Education is weakly statistically significant only at 1% (p-value equal to 0.08). Regarding the intercept, it is statistically significant but only at 5% and all the coefficients are positive related. Then we look at the standardized residuals, Figure 1.3 shows the standardized residuals versus the fitted values (left panel) and the QQ plot (right panel). The plot of standardized residuals versus fitted values seems

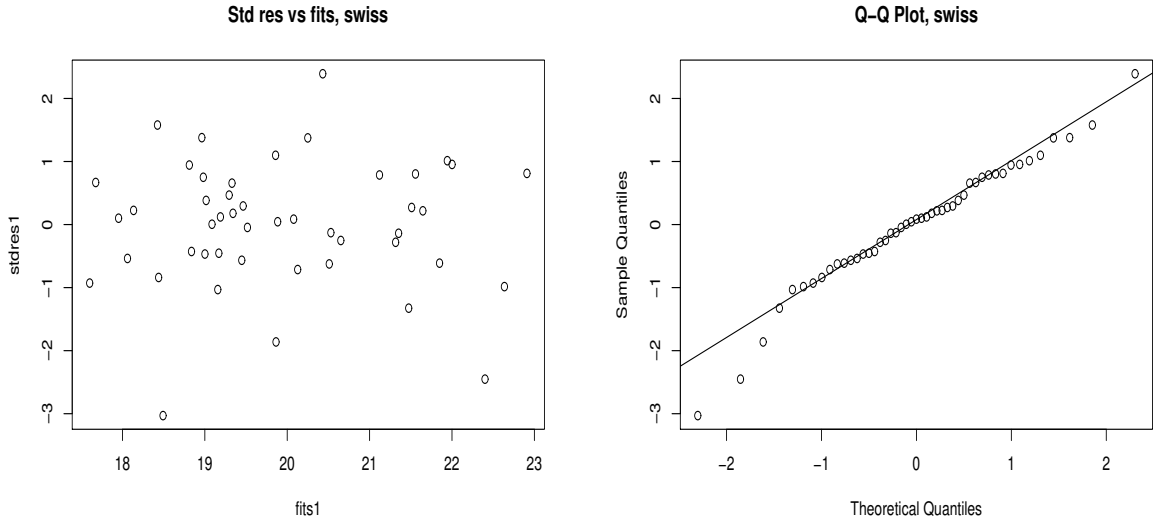

Figure 1.3: Plot of standardized residuals versus fitted values (left) and QQ plot (right) for the model with two explanatory variables.

random with a possible outlier, while the QQ plot shows some deviation from the line in

the tails but the Shapiro-Wilk test is not significant, thus we cannot reject the normality assumption.

Moving to the second model, we include only one explanatory variable

```
> mod1 <- lm(swiss$Infant.Mortality ~ swiss$Fertility )
> summary(mod1)

Call:
lm(formula = swiss$Infant.Mortality ~ swiss$Fertility)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6038 -1.5673 -0.0607  1.8367  6.0788

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     13.12970    2.25063   5.834 5.51e-07 ***
swiss$Fertility  0.09713    0.03160   3.074  0.00359 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.677 on 45 degrees of freedom
Multiple R-squared:  0.1735,Adjusted R-squared:  0.1552
F-statistic: 9.448 on 1 and 45 DF,  p-value: 0.003585

> anova(mod1)
Analysis of Variance Table

Response: swiss$Infant.Mortality
                Df Sum Sq Mean Sq F value   Pr(>F)
swiss$Fertility  1  67.72  67.717  9.4477 0.003585 **
Residuals       45 322.54   7.168
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the variable Fertility is statistically significant with p-value equal to 0.00359 and also the intercept is statistically significant. In Figure 1.4, the residual plot again shows evidence of one outlier but no other problems with the assumption of constant variance. Regarding to normality assumption, it seems to have some movements in the tails, but looking at the Shapiro-Wilk test, we do not reject the null hypothesis with p-value equal to 0.3353.

In conclusion, the best model is the one with one explanatory variable, although it has a small adjusted R^2, but in the model with two explanatory variables, the new introduced variables is weakly statistically significant and the differences in term of adjusted R^2 are small.

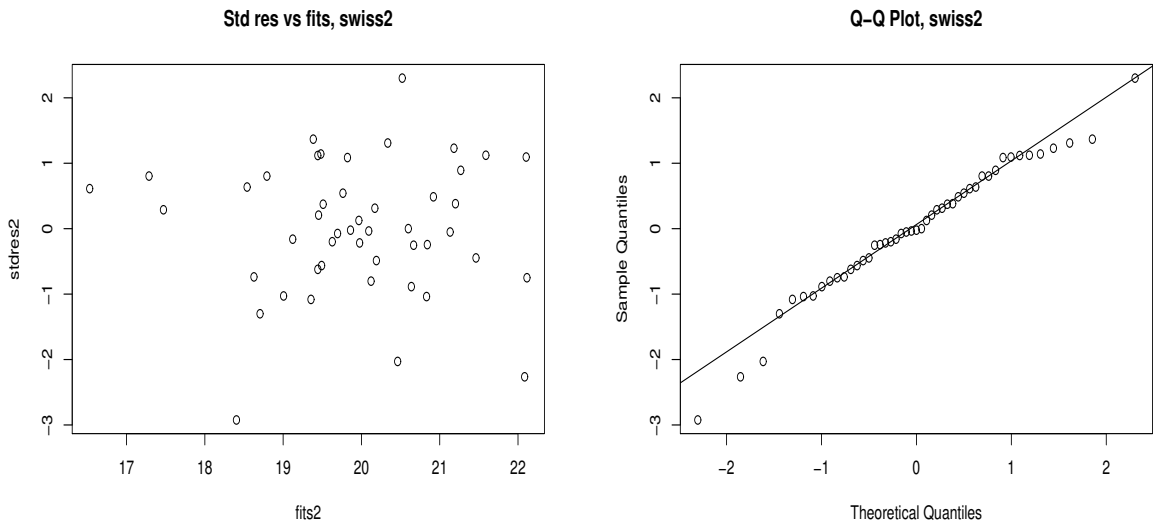

Figure 1.4: Plot of standardized residuals versus fitted values (left) and QQ plot (right) for the model with one explanatory variable.