Statistical Modeling I
Practical in R – Output

Practical in R – Output

In this practical, we will work with the dataset on presidential elections in US in year 2000 (on the https://electionlab.mit.edu/data is possible to found other data). We will look at how to select the best model by using the AIC and other measures.

In the file USElection.csv, we have different variables of interest, such as the fraction of the state's total counted vote for George W. Bush, which is the response variable. In the file, we find the following eleven columns for each of the US states:

- $Y = \%Bush$ which is the percentage of votes for G.W. Bush;
- $X_1 = UnEmpR$ which is the unemployment rate;
- $X_2 = Pop$ is the total population of the state;
- $X_3 = \%Male$ is the percentage of male;
- $X_4 = \%Pop > 65$ is the percentage of population older than 65;
- $X_5 = \%NonMetr$ is the percentage of rural (nonmetro) population;
- $X_6 = \%PopPov$ is the percentage of population below the poverty level;
- $X_7 = NuHouse$ is the total number of households;
- $X_8 = \%Inc > 50$ is the percentage of house income bigger than \$50000;
- $X_9 = \%Inc > 75$ is the percentage of house income bigger than \$75000;
- $X_{10} = \%Inc > 100$ is the percentage of house income bigger than \$100000.

Note to find the VIF values, we should use the command `vif(model)` and we should install and load the correct library:

```
> install.packages("car")
> library(car)
```

1. First of all we need to load the data in R:

```
> data <- read.csv("USElection.csv")
>
> Y<- data[,1]
> X1 <- data[,2]
> X2 <- data[,3]
> X3 <- data[,4]
> X4 <- data[,5]
> X5 <- data[,6]
> X6 <- data[,7]
> X7 <- data[,8]
> X8 <- data[,9]
> X9 <- data[,10]
> X10 <- data[,11]
```

After defining it, we fit the full model for the response variable by including all the explanatory variables

```
> mody <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10)
```

Further details are related to the computation of the VIF, which uses the following R command:

```
> vif(mody)
       X1        X2        X3        X4        X5        X6
 2.120386 22.175187  2.820395  1.775724  2.399138  4.690190
 X7        X8        X9       X10
21.735725 38.328787 69.432943 22.556284
```

Looking at the VIF values, we have different values larger than 10, f.e X_2, X_7, X_8, X_9 and X_{10}. Thus, there are problems with multicollinearity and it is likely to have inflated the standard errors and this may be why so many variables appear not significant.

2. In the first case, we define the full model with all the explanatory variables and then use the backwards elimination procedure:

```
  > reduced.model <- step(mody, direction="backward")
Start:  AIC=200.11
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq    RSS    AIC
- X2    1      0.29 1676.4 198.12
- X8    1      1.59 1677.7 198.16
- X6    1      4.57 1680.7 198.25
- X7    1      5.31 1681.4 198.27
```

```
- X9    1     16.24 1692.4 198.60
<none>              1676.1 200.11
- X3    1     78.30 1754.4 200.44
- X1    1    131.55 1807.7 201.97
- X10   1    136.59 1812.7 202.11
- X4    1    192.90 1869.0 203.67
- X5    1    424.86 2101.0 209.63

Step:  AIC=198.12
Y ~ X1 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq    RSS    AIC
- X8    1      1.46 1677.9 196.17
- X6    1      4.89 1681.3 196.27
- X9    1     15.96 1692.4 196.60
<none>              1676.4 198.12
- X3    1     83.52 1759.9 198.60
- X7    1    118.35 1794.8 199.60
- X1    1    131.64 1808.1 199.98
- X10   1    137.44 1813.9 200.14
- X4    1    192.72 1869.1 201.67
- X5    1    426.96 2103.4 207.69

Step:  AIC=196.17
Y ~ X1 + X3 + X4 + X5 + X6 + X7 + X9 + X10

       Df Sum of Sq    RSS    AIC
- X6    1     10.18 1688.0 194.47
- X9    1     25.92 1703.8 194.95
<none>              1677.9 196.17
- X3    1    105.99 1783.9 197.29
- X7    1    116.98 1794.9 197.60
- X1    1    134.21 1812.1 198.09
- X10   1    137.45 1815.3 198.18
- X4    1    191.44 1869.3 199.68
- X5    1    463.32 2141.2 206.60

Step:  AIC=194.47
Y ~ X1 + X3 + X4 + X5 + X7 + X9 + X10

       Df Sum of Sq    RSS    AIC
- X9    1     16.13 1704.2 192.96
<none>              1688.0 194.47
- X3    1     99.83 1787.9 195.41
- X7    1    128.20 1816.2 196.21
```

```
- X10    1    129.72 1817.8 196.25
- X1     1    159.35 1847.4 197.07
- X4     1    206.24 1894.3 198.35
- X5     1    487.54 2175.6 205.41

Step:  AIC=192.96
Y ~ X1 + X3 + X4 + X5 + X7 + X10

       Df Sum of Sq    RSS    AIC
<none>              1704.2 192.96
- X7    1    121.90 1826.1 194.48
- X3    1    123.55 1827.7 194.53
- X1    1    186.93 1891.1 196.27
- X4    1    205.60 1909.8 196.77
- X5    1    472.51 2176.7 203.44
- X10   1    658.58 2362.8 207.62
```

Thus in this case, the best model is the one that includes X_1; X_3; X_4; X_5; X_7 and X_{10} with an AIC equal to 192.96.

On the other hand, we define the null model, which is the model with only the intercept and then we apply the forward fit model:

```
> modyn <- lm(Y ~ 1)
> aic.forward.model <- step(modyn, scope=~X1 + X2 + X3 + X4 + X5 +
 X6 + X7 + X8 + X9 + X10, direction="forward")
Start:  AIC=239.89
Y ~ 1

       Df Sum of Sq    RSS    AIC
+ X10   1   2165.90 3245.5 215.81
+ X5    1   1919.41 3492.0 219.55
+ X9    1   1822.81 3588.6 220.94
+ X8    1   1555.76 3855.7 224.60
+ X3    1   1523.81 3887.6 225.02
+ X4    1    232.61 5178.8 239.65
<none>              5411.4 239.89
+ X2    1    107.39 5304.0 240.86
+ X7    1     66.31 5345.1 241.26
+ X1    1     58.66 5352.8 241.33
+ X6    1      0.36 5411.1 241.88

Step:  AIC=215.81
Y ~ X10

       Df Sum of Sq    RSS    AIC
```

```
+ X3    1    874.89 2370.6 201.79
+ X4    1    615.32 2630.2 207.09
+ X5    1    539.36 2706.2 208.54
+ X6    1    148.70 3096.8 215.42
<none>              3245.5 215.81
+ X9    1     84.27 3161.3 216.47
+ X8    1     71.54 3174.0 216.68
+ X1    1     30.97 3214.6 217.32
+ X7    1      8.49 3237.0 217.68
+ X2    1      5.26 3240.3 217.73

Step:  AIC=201.79
Y ~ X10 + X3

       Df Sum of Sq    RSS    AIC
+ X5    1   274.362 2096.3 197.52
+ X4    1    91.232 2279.4 201.79
<none>              2370.6 201.79
+ X1    1    44.884 2325.8 202.82
+ X8    1    20.492 2350.1 203.35
+ X9    1     6.968 2363.7 203.64
+ X6    1     0.515 2370.1 203.78
+ X2    1     0.426 2370.2 203.78
+ X7    1     0.087 2370.6 203.79

Step:  AIC=197.52
Y ~ X10 + X3 + X5

       Df Sum of Sq    RSS    AIC
+ X4    1   117.355 1978.9 196.58
+ X7    1    93.620 2002.7 197.19
+ X2    1    82.674 2013.6 197.47
<none>              2096.3 197.52
+ X1    1    68.807 2027.5 197.82
+ X9    1    23.099 2073.2 198.96
+ X8    1    17.487 2078.8 199.09
+ X6    1     9.085 2087.2 199.30

Step:  AIC=196.58
Y ~ X10 + X3 + X5 + X4

       Df Sum of Sq    RSS    AIC
+ X1    1   152.833 1826.1 194.48
+ X7    1    87.804 1891.1 196.27
+ X2    1    78.533 1900.4 196.52
```

```
<none>                1978.9 196.58
+ X6    1    42.094 1936.8 197.49
+ X9    1    31.527 1947.4 197.76
+ X8    1    30.767 1948.2 197.78

Step:  AIC=194.48
Y ~ X10 + X3 + X5 + X4 + X1

       Df Sum of Sq    RSS    AIC
+ X7    1   121.902 1704.2 192.96
+ X2    1   115.359 1710.7 193.16
<none>                1826.1 194.48
+ X9    1     9.835 1816.2 196.21
+ X6    1     5.217 1820.9 196.34
+ X8    1     2.076 1824.0 196.43

Step:  AIC=192.96
Y ~ X10 + X3 + X5 + X4 + X1 + X7

       Df Sum of Sq    RSS    AIC
<none>                1704.2 192.96
+ X9    1   16.1324 1688.0 194.47
+ X8    1    4.5332 1699.7 194.82
+ X6    1    0.3919 1703.8 194.95
+ X2    1    0.0756 1704.1 194.96
```

Also in this case, we arrive at the same best model as before, thus the model that includes X_{10}; X_3; X_5; X_4; X_1 and X_7 with an AIC equal to 192.96.

3. As stated in the previous point, we run the linear regression with the following explanatory variables: X_{10}; X_3; X_5; X_4; X_1 and X_7. Thus in this case, the percentage of house income bigger than $100000; of male; of rural population; of population older than 65 and the unemployment rate and the total number of households are the important variables in the linear regression.

   We run the usual linear regression model and the summary states:

```
> modfinal <- lm(Y ~ X10 + X3 + X5 + X4 + X1 + X7)
> summary(modfinal)

Call:
lm(formula = Y ~ X10 + X3 + X5 + X4 + X1 + X7)

Residuals:
    Min      1Q  Median      3Q     Max
-16.369  -4.045   1.078   3.446  10.017
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.929e+01  7.535e+01  -0.654 0.516406
X10         -2.495e+00  6.050e-01  -4.124 0.000163 ***
X3           2.532e+00  1.418e+00   1.786 0.080983 .
X5           2.071e-01  5.931e-02   3.493 0.001102 **
X4          -1.416e+00  6.145e-01  -2.304 0.026006 *
X1          -2.122e+00  9.658e-01  -2.197 0.033340 *
X7           8.879e-07  5.005e-07   1.774 0.082970 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.223 on 44 degrees of freedom
Multiple R-squared:  0.6851,Adjusted R-squared:  0.6421
F-statistic: 15.95 on 6 and 44 DF,  p-value: 1.231e-09
```

The variables related to X_3 and X_7 are statistically significant, but only at the 10% significant level. Looking at the overall regression, thus at statistic F, we have that the overall regression is highly significant, with value equal to 15.95 and p-value really small. Moving to the adjusted R^2 it has values of 64.21%, which is bigger than the adjusted R^2 of the model with all the explanatory variables and thus the new model is better than the full model.

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

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X10        1 2165.90 2165.90 55.9208 2.292e-09 ***
X3         1  874.89  874.89 22.5886 2.172e-05 ***
X5         1  274.36  274.36  7.0837   0.01082 *
X4         1  117.36  117.36  3.0300   0.08873 .
X1         1  152.83  152.83  3.9460   0.05323 .
X7         1  121.90  121.90  3.1474   0.08297 .
Residuals 44 1704.18   38.73
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the ANOVA table, we have that variables 4, 1 and 7 are not statistically significant at 5% level once variable 10, 3 and 5 are included in the model. However, this variables are all included in the model by the AIC.

As a further step, we have a look at the standardized residuals for this final model. Figure 1.1 shows the standardized residuals versus the fitted values (left panel) and the QQ plot (right panel)

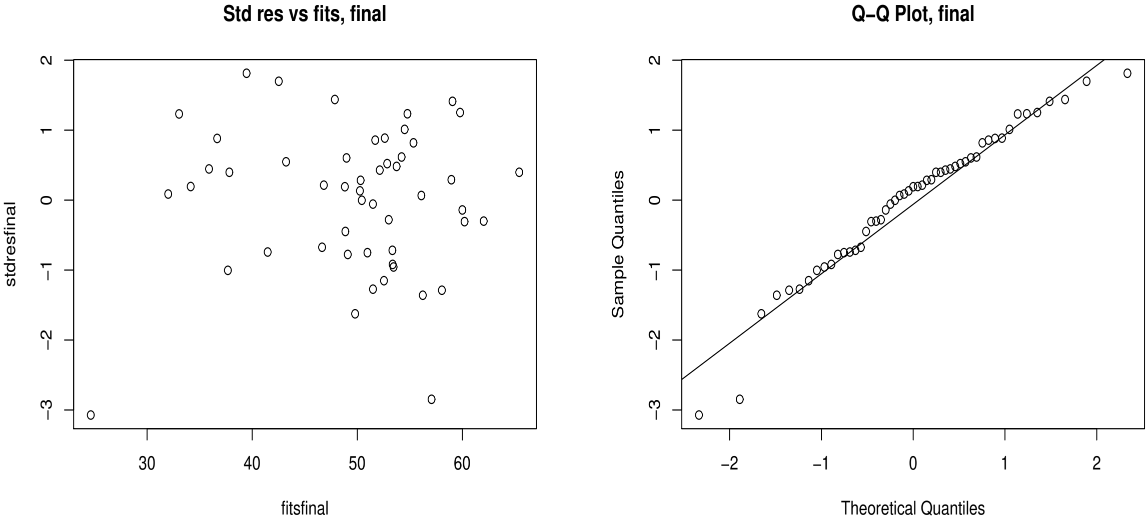

Figure 1.1: Plot of standardized residuals versus fitted values (left) and QQ plot (right) for the model with six explanatory variables.

Left panel of Figure 1.1 shows the constant variance assumption is okay, while the normality assumption (right panel) shows some problems in the left tails. We have a look at the usual Shapiro-Wilk test in R

```
> shapiro.test(stdresfinal)

Shapiro-Wilk normality test

data:  stdresfinal
W = 0.95802, p-value = 0.06866
```

with p-value equal to 0.069, which is nearly to significance at 5% level, so we might have to think to transforming the dependent variable.

In conclusion, looking at the VIF, we have

```
> vif(modfinal)
     X10       X3       X5       X4       X1       X7
1.550514 1.720006 2.033735 1.734214 1.144043 1.418790
```

which are all less than 10, thus there is no problems with multicollinearity.

Then we can compute the leverage values

```
> hatvalues(modfinal)
         1          2          3          4          5          6          7
0.07645769 0.57344144 0.13774677 0.06409605 0.35113013 0.16988339 0.15938588
         8          9         10         11         12         13         14
0.09167765 0.32603150 0.30887106 0.10180379 0.24333131 0.10182275 0.05688161
        15         16         17         18         19         20         21
```

```
0.05995066 0.10258757 0.03424660 0.05838242 0.10231233 0.09764033 0.09994505
        22         23         24         25         26         27         28
0.18610877 0.04213988 0.03497601 0.21376281 0.04095702 0.14538842 0.05339417
        29         30         31         32         33         34         35
0.26688781 0.11821401 0.10750926 0.06957300 0.20610314 0.03962528 0.11862422
        36         37         38         39         40         41         42
0.07374407 0.05221588 0.09467141 0.11750209 0.14397171 0.07612399 0.14077983
        43         44         45         46         47         48         49
0.04607636 0.35687017 0.21888827 0.14660385 0.13422723 0.12592447 0.17361545
        50         51
0.03572137 0.10217406
```

In the same way also the Cook's distance:

```
> cooks.distance(modfinal)
           1            2            3            4            5
2.446345e-02 1.964268e-01 3.999502e-04 7.687996e-04 2.041358e-01
           6            7            8            9           10
4.852367e-02 5.381035e-03 1.121319e-02 6.521981e-01 2.099686e-01
          11           12           13           14           15
2.974862e-03 2.529163e-02 2.538802e-02 3.913898e-03 2.115124e-03
          16           17           18           19           20
3.020160e-02 3.977916e-03 3.864297e-05 4.699225e-02 2.505069e-02
          21           22           23           24           25
2.511123e-03 2.474720e-04 1.262582e-03 6.869031e-03 1.475145e-02
          26           27           28           29           30
3.297901e-08 2.298521e-03 1.227827e-02 2.678554e-02 8.751435e-04
          31           32           33           34           35
2.611278e-02 9.025633e-03 1.401261e-03 1.604397e-03 1.634187e-03
          36           37           38           39           40
4.172669e-04 5.275827e-03 8.429698e-03 5.706726e-03 2.420961e-02
          41           42           43           44           45
8.655097e-03 2.110421e-03 2.213652e-05 1.565047e-03 7.981325e-02
          46           47           48           49           50
1.989402e-01 8.052362e-03 1.241146e-02 2.402969e-03 4.838958e-03
          51
2.574073e-03
```

We can also include the graphical representation of them:

```
> i = (1:51)
> plot(i,hatvalues(modfinal), main = "Leverage values, Election")
> plot(i,cooks.distance(modfinal), main = "Cook's distance, Election
> qf(p=0.5,df1 = 7, df2 = 44)
[1] 0.920551
```

Figure 1.2 shows the leverage values and Cook's distance for the model with 6 explanatory variables plus the intercept. In this case, we have $n = 51$ and $p = 7$ (6 regressors + 1 intercept), thus a leverage value is large if it is $\frac{14}{51} = 0.274$ and very large if $\frac{21}{51} = 0.412$. Looking at thep plot we see that there is one very large value and a few other large ones (The very large one is actually Alaska which has a number of unusual regressor values).

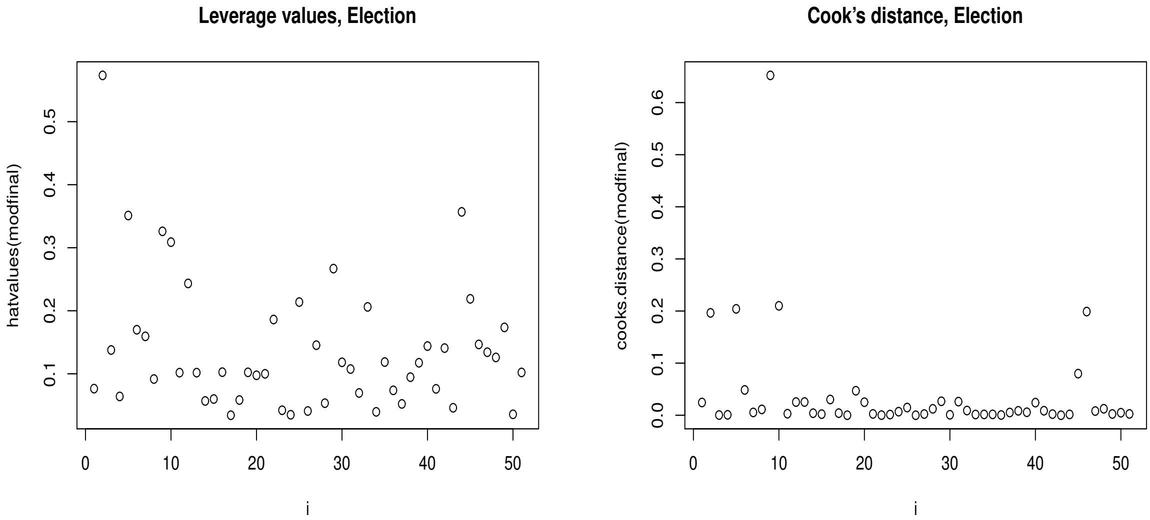

Figure 1.2: Plot of leverage values (left) and Cook's distance (right) for the model with six explanatory variables.

Moving to the Cook's distance, the critical value for Cook's distance is 0.92 from a F distribution with 7 and 44 degrees of freedom and the highest Cook's distance for observation 9 is smaller than that. Thus it is nevertheless more influential than any other state (It is actually District of Columbia, the area surrounding Washington).