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.

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)
> summary(mody)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
    X10)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7014  -3.1110   0.9113   3.4952  11.0512

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.334e+01  1.025e+02  -0.520  0.60579
X1          -2.423e+00  1.368e+00  -1.772  0.08404 .
X2           5.796e-08  6.994e-07   0.083  0.93437
X3           2.581e+00  1.889e+00   1.367  0.17928
X4          -1.388e+00  6.468e-01  -2.146  0.03803 *
X5           2.133e-01  6.700e-02   3.184  0.00281 **
X6           1.982e-01  6.003e-01   0.330  0.74305
X7           7.250e-07  2.037e-06   0.356  0.72384
X8          -1.529e-01  7.852e-01  -0.195  0.84662
X9           1.227e+00  1.971e+00   0.623  0.53707
X10         -4.333e+00  2.400e+00  -1.805  0.07854 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.473 on 40 degrees of freedom
Multiple R-squared:  0.6903,Adjusted R-squared:  0.6128
F-statistic: 8.914 on 10 and 40 DF,  p-value: 1.738e-07

> anova(mody)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1   58.66   58.66  1.3999  0.243720
X2         1   95.92   95.92  2.2891  0.138145
```

```
X3         1 1483.08 1483.08 35.3930 5.571e-07 ***
X4         1    4.80    4.80  0.1146  0.736780
X5         1 1339.52 1339.52 31.9668 1.448e-06 ***
X6         1  137.02  137.02  3.2699  0.078087 .
X7         1   88.12   88.12  2.1030  0.154808
X8         1  350.41  350.41  8.3624  0.006167 **
X9         1   41.17   41.17  0.9825  0.327547
X10        1  136.59  136.59  3.2596  0.078536 .
Residuals 40 1676.13   41.90
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus once defined the full model, we can look further at the plots of the standardized residuals. We run two plots: the standardized residuals versus fitted values (left panel) and the QQ plot (right panel).

```
> stdresfull <-rstandard(mody)
> fitsfull<-fitted(mody)
>
> plot(fitsfull,stdresfull, main="Std res vs fits, full")
> qqnorm(stdresfull, main="Q-Q Plot, full")
> qqline(stdresfull)
```

Left panel of Figure 1.1 shows no reason to doubt that the variance is constant, while there are three values that show negative standardized residuals. Moving to the right panel, we have heavy left tails, thus we cast some doubts on the normality assumption. For looking at the normality assumption, we have a look at the Shapiro-Wilk test, which gives:

```
> shapiro.test(stdresfull)

Shapiro-Wilk normality test

data:  stdresfull
W = 0.95133, p-value = 0.03581
```

The p-value is smaller than the significance level, thus we reject the null hypothesis of normality assumption of the residuals.

2. From the summary statistics of the linear regression, we see that few variables are statistically significant, like X_4 (percentage of population older than 65) and X_5 (percentage of rural population), while X_1 (unemployment rate) and X_{10} (percentage of house income bigger than \$100000) are statistically significant but only at 10%. Moving to the Anova table, we have that few of the variables are significant in the presence of the other variables: X_3 (% percentage of male), X_5 (percentage of rural population) and X_8

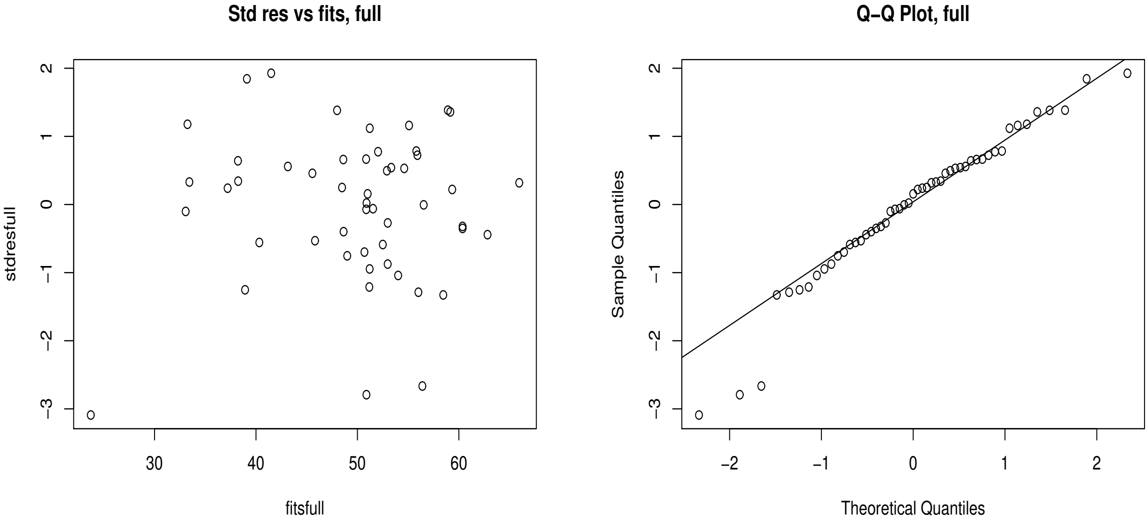

Figure 1.1: Plot of standardized residuals versus fitted values (left) and QQ plot (right) for the model with all the explanatory variables.

(percentage of house income bigger than $50000), while X_6 (percentage of population below poverty level) and X_{10} (percentage of house income bigger than $100000) are statistically significant but only at 10%.

Moving to the overall regression, the F statistic and relatively p-value indicate that the overall regression is highly significant (F = 8.91 and p-value = 1.73×10^{-7}). For the adjusted R^2, we have a value of 61.28%, which shows a lot of variation in the data not explained by all these variables.

3. 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.