Practical 2 - Multiple Linear Regression

J. Chiquet, G. Rigaill and A.A Charantonis

September 2016

2000 US Presidential vote in Georgia

1. Preliminaires

We first load some usefull packages.

```
library(ggplot2); library(GGally)
library(car)
library(leaps)
```

- Load the `gavote` data set at [http://julien.cremeriefamily.info/reglin/gavote.txt]

```
## Read the data
gavote <- read.table(file = "gavote.txt", header=TRUE)
```

- Create the variable `undercount` (proportion of invalid ballots) and add it to the table. We'll look at the prediction of this variable using others. It is therefore the "response variable". Delete the variables `votes` and `ballots` (and explain why).

```
## Create the variable `undercount` (percentage of invalid ballots)
gavote$undercount <- (gavote$ballots - gavote$votes)/gavote$ballots
```

- Create variables `pergore`, `perbush` and `perother` (percentage of ballots for Gore, Bush and other candidates). Add them to the table. Delete the `gore` , `bush` and `other` variables of the table.

```
## Create variable `pergore`, `perbush` and ``perothers` (proportion of ballots for each candidate)
gavote$pergore <- gavote$gore/gavote$votes
gavote$perbush <- gavote$bush/gavote$votes
gavote$perother <- gavote$other/gavote$votes

## Suppress redondant variables
gavote$gore <- NULL
gavote$bush <- NULL
gavote$other <- NULL
```

We suppress the variable `votes` and `ballots` because predicting the proportion of invalid ballots (`undercount`) using those is not relevant (does not make sense), furthermore they will hide the effect of other possibly relevant variables.

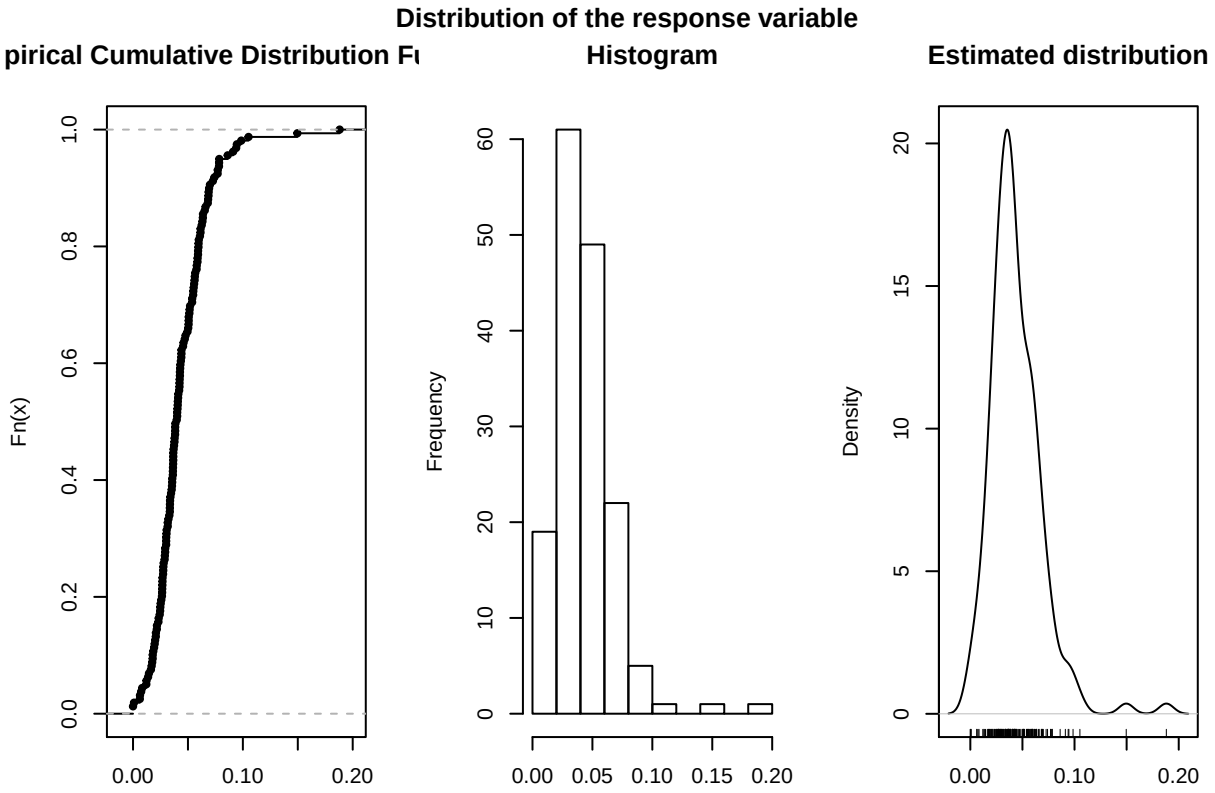
```
gavote$votes   <- NULL
gavote$ballots <- NULL
```

2. Descriptive Analysis

Make a descriptive analysis. This analysis will help you to choose a particular linear models and simplify your interpretation of the results.

Let's start with the `undercount` variable and have a look at its empirical distribution.

```
par(mfrow=c(1,3))
plot(ecdf(gavote$undercount), main="Empirical Cumulative Distribution Function", xlab="")
hist(gavote$undercount,main="Histogram",xlab="")
plot(density(gavote$undercount),main="Estimated distribution", xlab="")
rug(gavote$undercount)
title(outer=TRUE, main ="\nDistribution of the response variable", sub="Percentage")
```

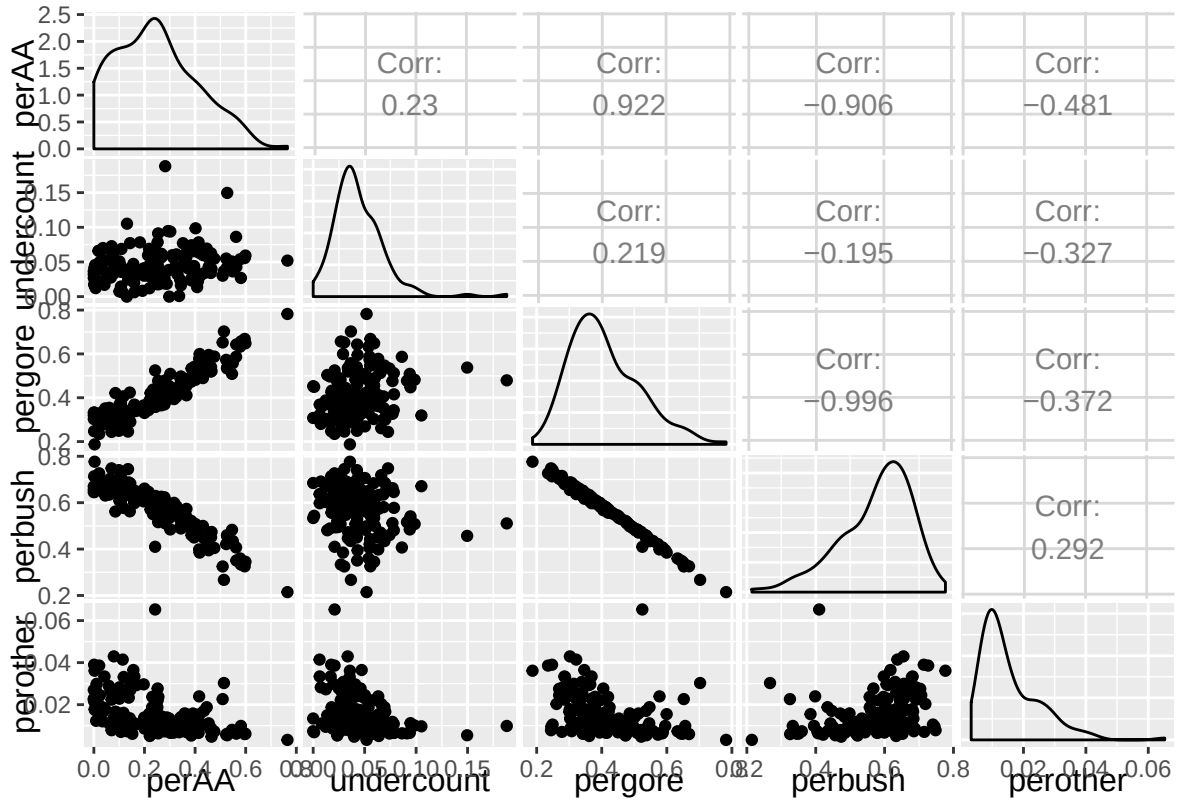


In the table there are both categorical and quantitative variables. The descriptive analysis must be made accordingly.

```
are.factor  <- sapply(gavote, is.factor)
```

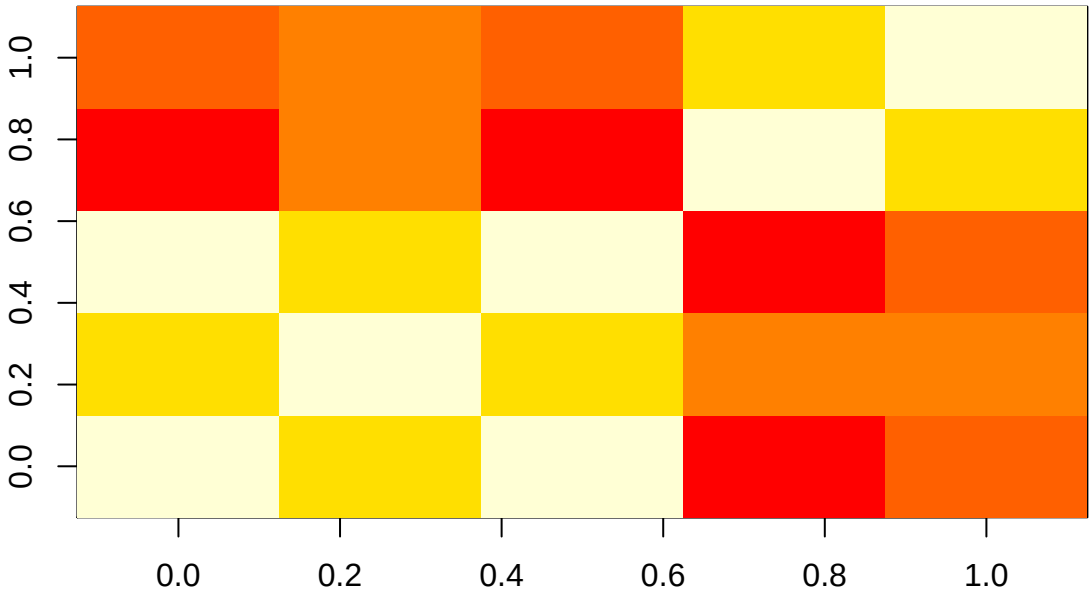
We first make a plot for all pairs of numerical variables.

```
ggpairs(gavote, columns = which(!are.factor))
```

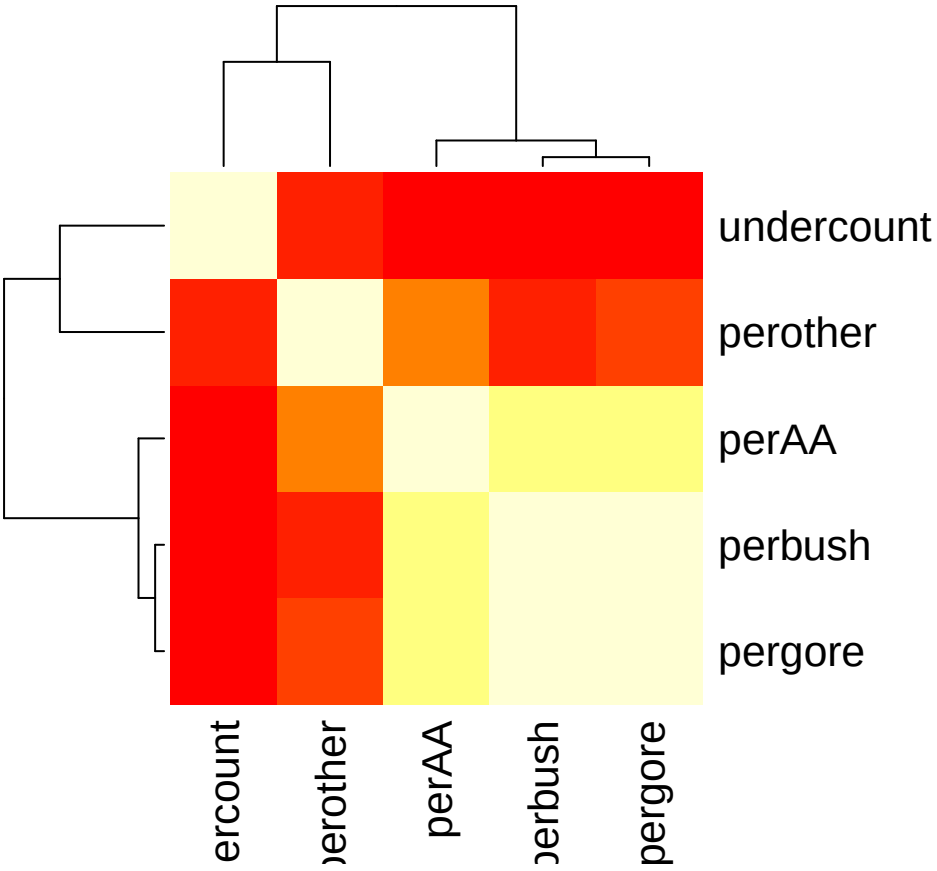


We then look at the correlation matrice between numerical variables.

```
image(cor(gavote[, !are.factor]))
```



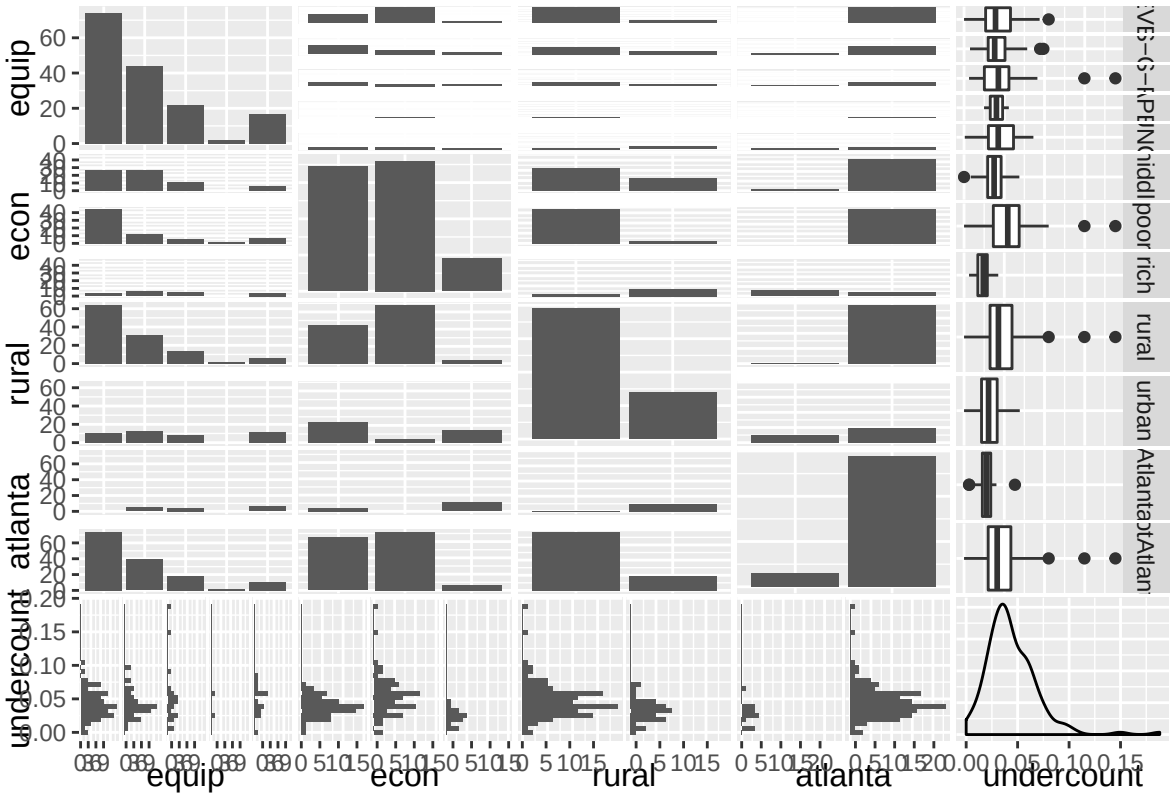
```
heatmap(abs(cor(gavote[, !are.factor])), symm=TRUE)
```



Finally we look at the response variable as a function of categorical variables.

```
ggpairs(gavote, columns = which(colnames(gavote) == "undercount" | are.factor))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

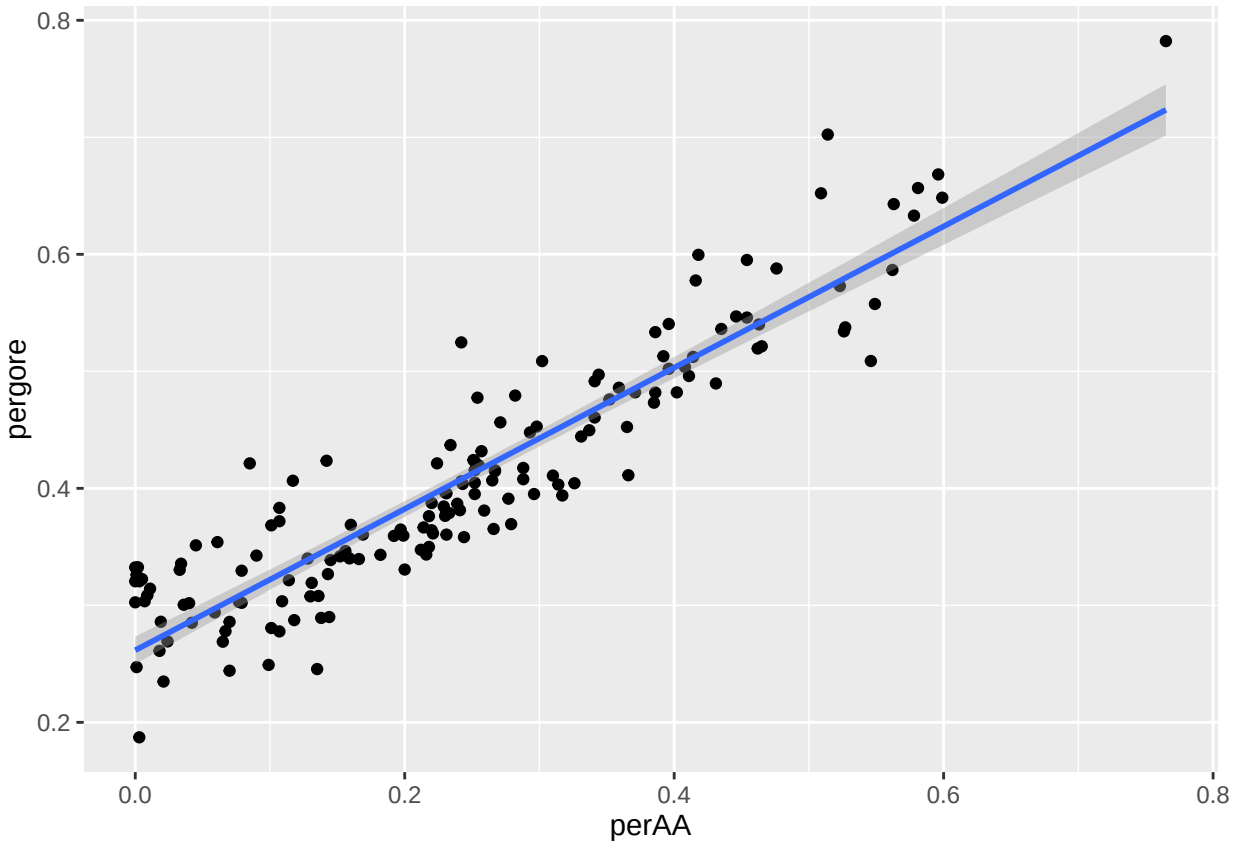


3. Some simple linear models

Votes among African Americans

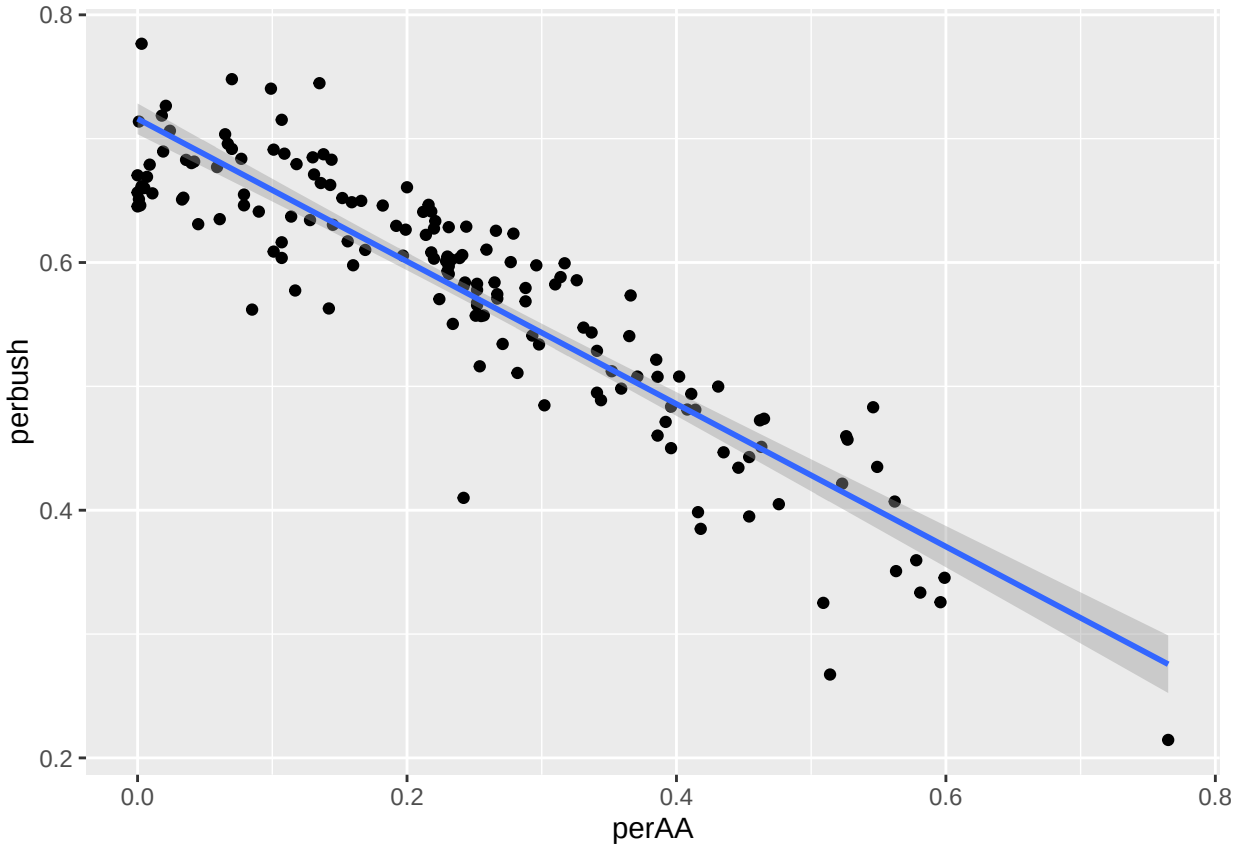
- Draw the scatterplot between `pergore` and `perAA`. Fit a simple regression model and draw the regression line between `pergore` and `perAA`.

```
ggplot(gavote, aes(x=perAA,y=pergore)) + geom_point() + stat_smooth(method="lm", formula=y~x)
```



We could draw the proportion of votes for Bush among afro-american and make similar conclusions.

```
ggplot(gavote, aes(x=perAA,y=perbush)) + geom_point() + stat_smooth(method="lm", formula=y~x)
```



- Regress `undercount` with `pergore`, then with `perAA` and finally with `pergore` and `perAA`. Compare these models and interpret the results.

```
M0 <- lm(undercount~1,gavote)
M11 <- lm(undercount~perAA,gavote)
M12 <- lm(undercount~pergore,gavote)
M2 <- lm(undercount~pergore+perAA,gavote)
anova(M0,M12,M2)
```

```
## Analysis of Variance Table
##
## Model 1: undercount ~ 1
## Model 2: undercount ~ pergore
## Model 3: undercount ~ pergore + perAA
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1    158 0.098477
## 2    157 0.093764  1 0.0047129 7.8845 0.005623 **
## 3    156 0.093249  1 0.0005151 0.8617 0.354701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(M0,M11,M2)
```

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

```
## Model 1: undercount ~ 1
## Model 2: undercount ~ perAA
## Model 3: undercount ~ pergore + perAA
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1    158 0.098477
## 2    157 0.093282  1 0.0051953 8.6914 0.003689 **
## 3    156 0.093249  1 0.0000327 0.0547 0.815309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`pergore` is not usefull to predict `perAA` and vice versa.

Votes according to the standard of living and one-way ANOVA

To study the effect of the `econ` factor on the variable `undercount`, we propose the following ANOVA model:

$$Y_i = \mu + \mathbf{1}_{\{i \in k\}} \mu_k + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where y_i describes the variable `undercount` associated with the i th individual, μ the intercept and μ_k an additive term associated with the group k (that is to say to the modality k of the variable `econ`).

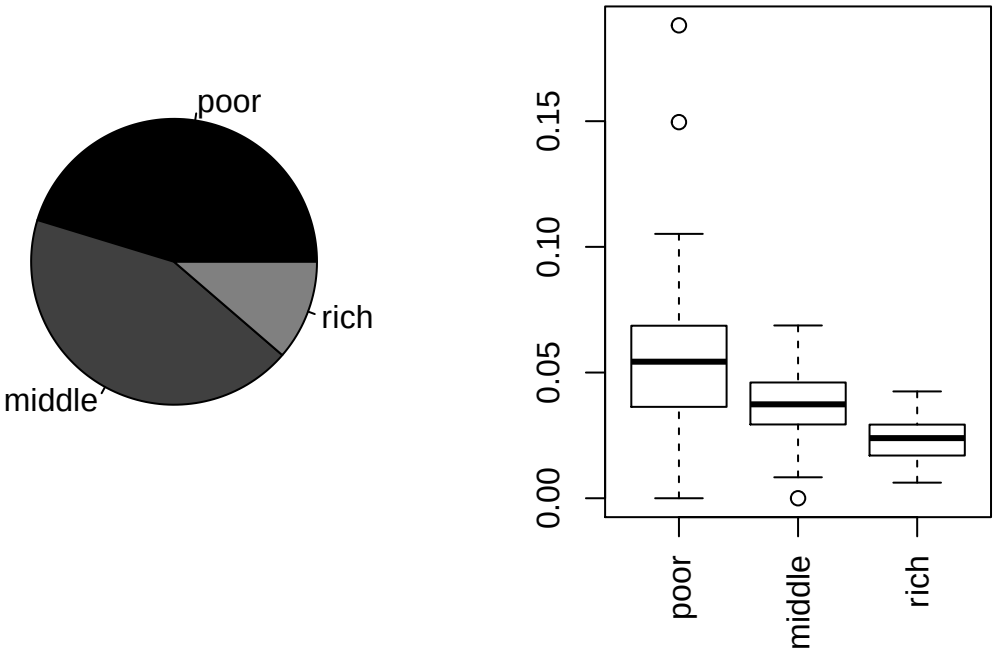
- Observe the levels of the categorical variables `econ`. Recode this variable by classifying these terms in an easily readable order.

```
gavote$econ <-factor(gavote$econ, levels=c("poor","middle","rich"))
```

- Represent the distribution of the variable `econ`.

To study categorical variable the `pie` and `barplot` command provide interpretable results.

```
par(mfrow=c(1,2))
pie(table(gavote$econ),col=gray(0:4/4))
boxplot(undercount~econ,gavote,las=3)
```



These boxplots clearly show that there is a trend. Rich towns have fewer invalid ballots and poor ones have more.

- Show that the above model can be written as a linear model of the form

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Take $\mathbf{X}$ the following $n \times (\text{n.group} + 1)$ matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & \vdots & \vdots & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & \dots & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Column 1 corresponds to the intercept μ ; column 2 indicates individual from group 1 and corresponds to parameter μ_1; column 3 indicates individual from group 2 and corresponds to parameter μ_2; etc. The vector $\boldsymbol{\beta} = (\mu, \mu_1, \dots, \mu_K)$.

- Study the effect of `econ` factor.

```
M.eco <- lm(log(1+undercount)~econ,gavote)
summary(M.eco)
```

```
##
## Call:
## lm(formula = log(1 + undercount) ~ econ, data = gavote)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.052992 -0.011944  0.000129  0.008945  0.119380
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.052992   0.002489  21.291  < 2e-16 ***
## econmiddle  -0.016048   0.003558  -4.510 1.27e-05 ***
## econrich    -0.030427   0.005566  -5.467 1.78e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02112 on 156 degrees of freedom
## Multiple R-squared:  0.1983, Adjusted R-squared:  0.188
## F-statistic: 19.29 on 2 and 156 DF,  p-value: 3.267e-08
```

The variable `econ` has a significant effect on the proportion of invalid ballots. Rich towns are associated to fewer invalid ballots and poor towns have more. It is also possible to quantify this effect using the value of the parameters.

```
## mean percentageof undercount relative to the reference : `poor`
100/coef(M.eco)[1] * coef(M.eco)[2]
```

```
## (Intercept)
##   -30.28385
```

```
100/coef(M.eco)[1] * coef(M.eco)[3]
```

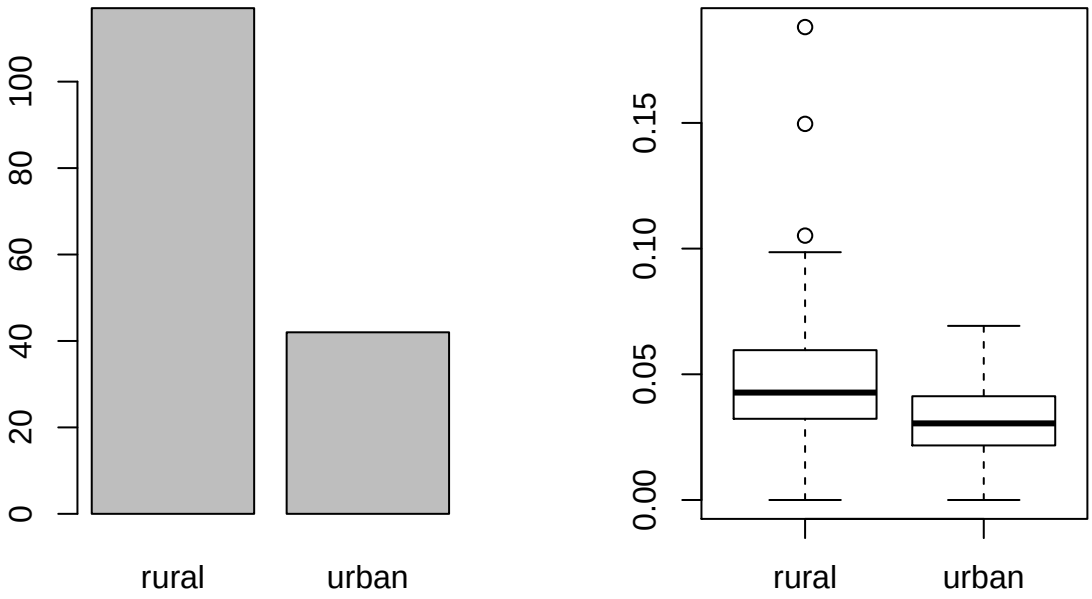
```
## (Intercept)
##   -57.41824
```

Votes according . . .

Illustrate a relevant behavior of your choice in the dataset by fitting a simple linear model different from those considered so far.

We propose to study the link between vote and the rural or urban status of the town. Visually it seems that there is link with the `undercount` variable.

```
par(mfrow=c(1,2))
barplot(sort(table(gavote$rural),decreasing=TRUE))
boxplot(undercount~rural,gavote)
```



This effect is confirmed by a test. There are less invalid ballots in urban towns.

```
M3 <- lm(undercount~rural,gavote)
anova(M3)
```

```
## Analysis of Variance Table
##
## Response: undercount
##            Df   Sum Sq   Mean Sq F value    Pr(>F)
```

```
## rural       1 0.007706 0.0077064  13.329 0.0003554 ***
## Residuals 157 0.090771 0.0005782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Construction of multiple linear models

Construction of a model using tests

- Fit a linear regression model that explains `undercount` based on the quantitative variables `perbush`, `pergore`, `perother` and `perAA`. How do you explain the results? Propose an alternative model by removing a variable (and justify). Remove this variable from the `gavote` data table.

```
summary(lm(undercount~perother+perbush+pergore+perAA, gavote))
```

```
##
## Call:
## lm(formula = undercount ~ perother + perbush + pergore + perAA,
##     data = gavote)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.050222 -0.014154 -0.002319  0.011856  0.137301
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08515    0.03586   2.375 0.018783 *
## perother    -0.82761    0.24389  -3.393 0.000876 ***
## perbush     -0.04385    0.04642  -0.945 0.346305
## pergore           NA         NA      NA       NA
## perAA       -0.01335    0.03218  -0.415 0.678791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02366 on 155 degrees of freedom
## Multiple R-squared:  0.1191, Adjusted R-squared:  0.102
## F-statistic: 6.985 on 3 and 155 DF,  p-value: 0.0001942
```

The design matrix is singular because of the high correlation between its variables. The $\mathbf{X}$ matrix is not full-ranked and thus is not invertible.

```
XtX <- cov(gavote[, colnames(gavote) %in% c("perbush","pergore", "perAA", "perother")])
rcond(XtX)
```

```
## [1] 7.386712e-18
```

We discard one variable, for example `perbush` that has a 0.99 correlation with `pergore`.

```
cov2cor(XtX)
```

```
##               perAA    pergore    perbush   perother
## perAA     1.0000000  0.9216525 -0.9056253 -0.4806871
## pergore   0.9216525  1.0000000 -0.9963448 -0.3723278
## perbush  -0.9056253 -0.9963448  1.0000000  0.2916860
## perother -0.4806871 -0.3723278  0.2916860  1.0000000
```

```
gavote <- gavote[, colnames(gavote) != "perbush"]
```

- Fit a model that explain `undercount` based on all the (remaining) explanatory variables. Build a model `Model.test` that integrates only the significant variables.

A preliminary log transform slightly improves the results, but this is marginal and one could keep the original data.

```
summary(lm(undercount~.,gavote))$adj.r.squared
```

```
## [1] 0.2392041
```

```
summary(lm(log(1+undercount)~.,gavote))$adj.r.squared
```

```
## [1] 0.2415461
```

Only the intercept and the `equip` and `econ` variables are significant (t-test). We thus build the following model:

```
model.test <- lm(undercount~equip+econ,gavote)
summary(model.test)
```

```
##
## Call:
## lm(formula = undercount ~ equip + econ, data = gavote)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.062688 -0.012753 -0.002154  0.010154  0.117329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.050732   0.002911  17.429  < 2e-16 ***
## equipOS-CC   0.008791   0.004315   2.037  0.04336 *
## equipOS-PC   0.020060   0.005460   3.674  0.00033 ***
## equipPAPER  -0.010485   0.015616  -0.671  0.50294
## equipPUNCH   0.012872   0.005952   2.162  0.03215 *
## econmiddle  -0.020629   0.003833  -5.382 2.73e-07 ***
## econrich    -0.039223   0.006013  -6.523 9.68e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0217 on 152 degrees of freedom
## Multiple R-squared:  0.2734, Adjusted R-squared:  0.2447
## F-statistic: 9.531 on 6 and 152 DF,  p-value: 6.926e-09
```

Keeping the `rural` and/or `perother` variables we get a slightly better R^2. But an analysis of variance show that these models are not significantly better. In case our goal is prediction (rather than explanation) we could keep them. Here we only keep significant variables.

```
model.test2 <- lm(undercount~equip+econ+rural,gavote)
summary(model.test2)$adj.r.squared
```

```
## [1] 0.2516078
```

```
anova(model.test, model.test2)
```

```
## Analysis of Variance Table
##
## Model 1: undercount ~ equip + econ
## Model 2: undercount ~ equip + econ + rural
##   Res.Df      RSS Df Sum of Sq     F Pr(>F)
## 1    152 0.071556
## 2    151 0.070434  1 0.0011213 2.404 0.1231
```

```
model.test3 <- lm(undercount~equip+econ+perother,gavote)
summary(model.test3)$adj.r.squared
```

```
## [1] 0.2474172
```

```
anova(model.test, model.test3)
```

```
## Analysis of Variance Table
##
## Model 1: undercount ~ equip + econ
## Model 2: undercount ~ equip + econ + perother
##   Res.Df      RSS Df  Sum of Sq      F Pr(>F)
## 1    152 0.071556
## 2    151 0.070829  1 0.00072694 1.5498 0.2151
```

```
model.test4 <- lm(undercount~equip+econ+rural+perother,gavote)
summary(model.test4)$adj.r.squared
```

```
## [1] 0.2518474
```

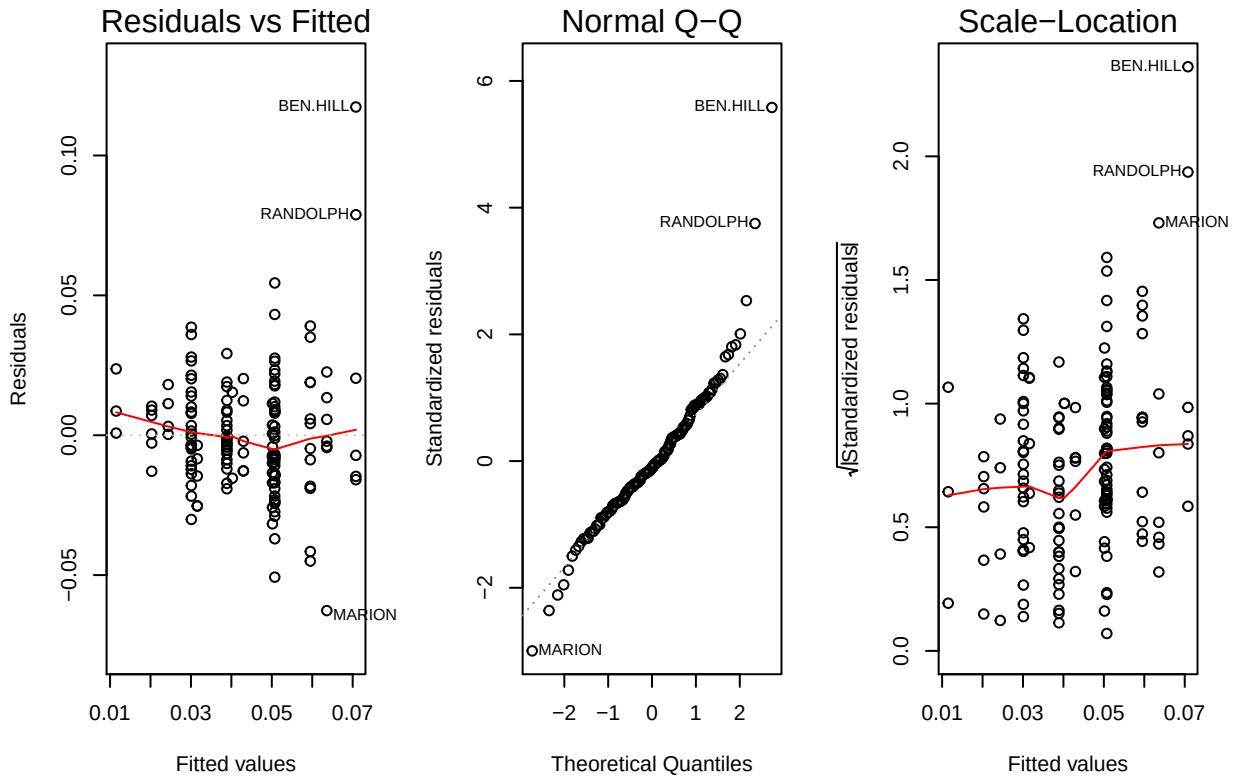
```
anova(model.test, model.test4)
```

```
## Analysis of Variance Table
##
## Model 1: undercount ~ equip + econ
## Model 2: undercount ~ equip + econ + rural + perother
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    152 0.071556
## 2    150 0.069946  2 0.0016102 1.7265 0.1814
```

- Study the residue of `model.test`.

Three data-points disturb normality and homosedasticity.

```
par(mfrow=c(1,3))
plot(model.test, which=1:3)
```



Construction by stepwise regression

- Using the procedure `step`, propose a model using the forward / backward procedureusing the AIC and BIC criteria. We will call these model `model.BIC` and `model.AIC`.

```
model.AIC <- step(lm(undercount~.,gavote), k=2)
```

```
## Start:  AIC=-1205.45
## undercount ~ equip + econ + perAA + rural + atlanta + pergore +
##     perother
##
##            Df Sum of Sq      RSS     AIC
## - atlanta   1 0.0000034 0.069708 -1207.4
## - pergore   1 0.0000868 0.069792 -1207.2
## - perAA     1 0.0001935 0.069899 -1207.0
## - perother  1 0.0006455 0.070351 -1206.0
## - rural     1 0.0007085 0.070414 -1205.8
## <none>                  0.069705 -1205.5
## - equip     4 0.0084239 0.078129 -1195.3
## - econ      2 0.0078679 0.077573 -1192.5
##
## Step:  AIC=-1207.44
```

```
## undercount ~ equip + econ + perAA + rural + pergore + perother
##
##            Df Sum of Sq      RSS     AIC
## - pergore   1 0.0000878 0.069796 -1209.2
## - perAA     1 0.0001910 0.069899 -1209.0
## - perother  1 0.0006611 0.070369 -1207.9
## - rural     1 0.0007126 0.070421 -1207.8
## <none>                  0.069708 -1207.4
## - equip     4 0.0086883 0.078397 -1196.8
## - econ      2 0.0082120 0.077920 -1193.7
##
## Step:  AIC=-1209.24
## undercount ~ equip + econ + perAA + rural + perother
##
##            Df Sum of Sq      RSS     AIC
## - perAA     1 0.0001493 0.069946 -1210.9
## - perother  1 0.0005815 0.070378 -1209.9
## - rural     1 0.0007368 0.070533 -1209.6
## <none>                  0.069796 -1209.2
## - equip     4 0.0087722 0.078568 -1198.4
## - econ      2 0.0083559 0.078152 -1195.3
##
## Step:  AIC=-1210.9
## undercount ~ equip + econ + rural + perother
##
##            Df Sum of Sq      RSS     AIC
## - perother  1 0.0004889 0.070434 -1211.8
## - rural     1 0.0008833 0.070829 -1210.9
## <none>                  0.069946 -1210.9
## - equip     4 0.0086956 0.078641 -1200.3
## - econ      2 0.0090961 0.079042 -1195.5
##
## Step:  AIC=-1211.79
## undercount ~ equip + econ + rural
##
##         Df Sum of Sq      RSS     AIC
## <none>               0.070434 -1211.8
## - rural  1 0.0011213 0.071556 -1211.3
## - equip  4 0.0084709 0.078905 -1201.7
## - econ   2 0.0150415 0.085476 -1185.0
```

```
model.BIC <- step(lm(undercount~.,gavote), k=log(nrow(gavote)))
```

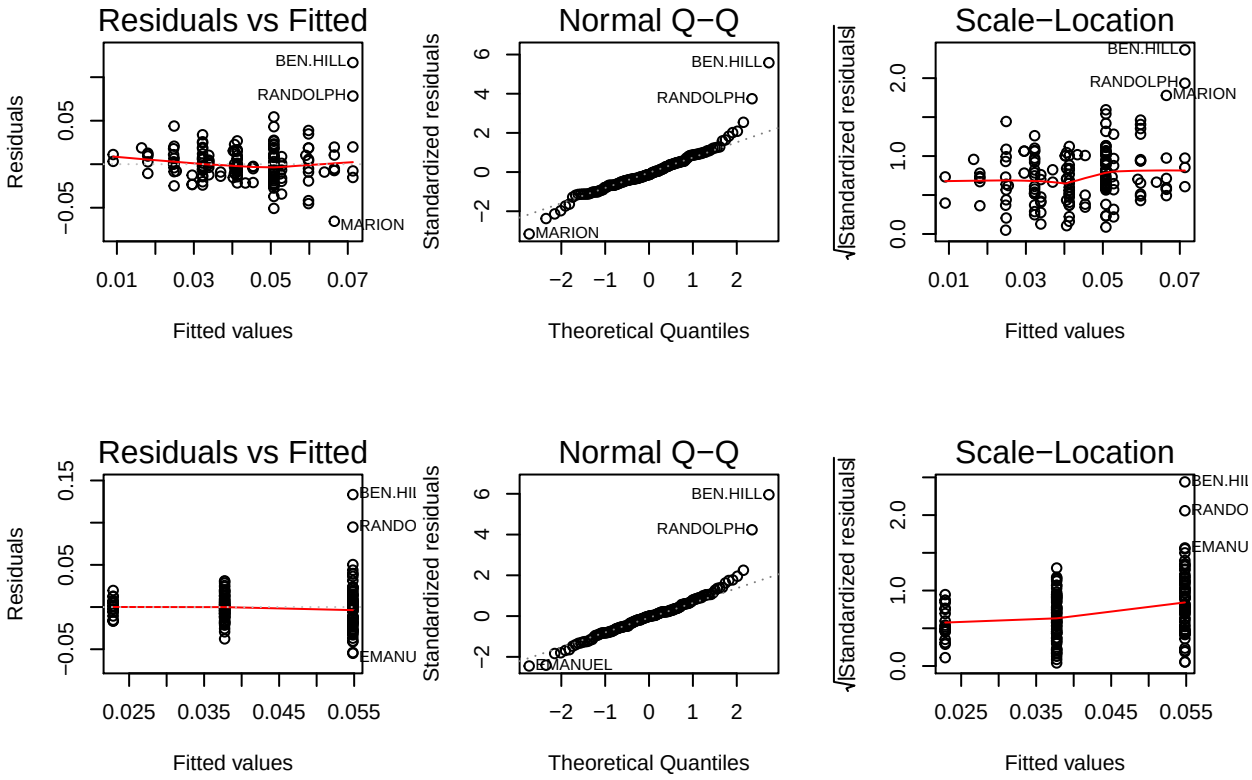
```
## Start:  AIC=-1168.62
## undercount ~ equip + econ + perAA + rural + atlanta + pergore +
##     perother
##
##            Df Sum of Sq      RSS     AIC
## - atlanta   1 0.0000034 0.069708 -1173.7
## - pergore   1 0.0000868 0.069792 -1173.5
## - perAA     1 0.0001935 0.069899 -1173.2
## - perother  1 0.0006455 0.070351 -1172.2
## - rural     1 0.0007085 0.070414 -1172.1
## - equip     4 0.0084239 0.078129 -1170.8
```

```
## <none>                  0.069705 -1168.6
## - econ      2 0.0078679 0.077573 -1161.8
##
## Step:  AIC=-1173.68
## undercount ~ equip + econ + perAA + rural + pergore + perother
##
##            Df Sum of Sq      RSS     AIC
## - pergore   1 0.0000878 0.069796 -1178.5
## - perAA     1 0.0001910 0.069899 -1178.3
## - perother  1 0.0006611 0.070369 -1177.2
## - rural     1 0.0007126 0.070421 -1177.1
## - equip     4 0.0086883 0.078397 -1175.3
## <none>                  0.069708 -1173.7
## - econ      2 0.0082120 0.077920 -1166.1
##
## Step:  AIC=-1178.55
## undercount ~ equip + econ + perAA + rural + perother
##
##            Df Sum of Sq      RSS     AIC
## - perAA     1 0.0001493 0.069946 -1183.3
## - perother  1 0.0005815 0.070378 -1182.3
## - rural     1 0.0007368 0.070533 -1182.0
## - equip     4 0.0087722 0.078568 -1180.0
## <none>                  0.069796 -1178.5
## - econ      2 0.0083559 0.078152 -1170.7
##
## Step:  AIC=-1183.28
## undercount ~ equip + econ + rural + perother
##
##            Df Sum of Sq      RSS     AIC
## - perother  1 0.0004889 0.070434 -1187.2
## - rural     1 0.0008833 0.070829 -1186.4
## - equip     4 0.0086956 0.078641 -1184.9
## <none>                  0.069946 -1183.3
## - econ      2 0.0090961 0.079042 -1174.0
##
## Step:  AIC=-1187.24
## undercount ~ equip + econ + rural
##
##         Df Sum of Sq      RSS     AIC
## - rural  1 0.0011213 0.071556 -1189.8
## - equip  4 0.0084709 0.078905 -1189.5
## <none>               0.070434 -1187.2
## - econ   2 0.0150415 0.085476 -1166.6
##
## Step:  AIC=-1189.8
## undercount ~ equip + econ
##
##         Df Sum of Sq      RSS     AIC
## - equip  4 0.0077405 0.079296 -1193.7
## <none>               0.071556 -1189.8
## - econ   2 0.0249718 0.096528 -1152.3
##
## Step:  AIC=-1193.74
```

```
## undercount ~ econ
##
##        Df Sum of Sq      RSS     AIC
## <none>              0.079296 -1193.7
## - econ  2  0.019181 0.098477 -1169.4
```

- Study the residues of these models.

```
par(mfrow=c(2,3))
plot(model.AIC, which=1:3)
plot(model.BIC, which=1:3)
```



Exhaustive search

When the number of predictors remains reasonable (<30), it is possible to test all linear models and choose the best in the sense of a criterion.

- Use the package `regsubsets` and its procedure `leaps`, to identify the best model with respect to the adjusted R^2. Comment.

```
model.all <- regsubsets(undercount~.,gavote)
summary(model.all)$outmat[which.max(summary(model.all)$adjr2), ]
```

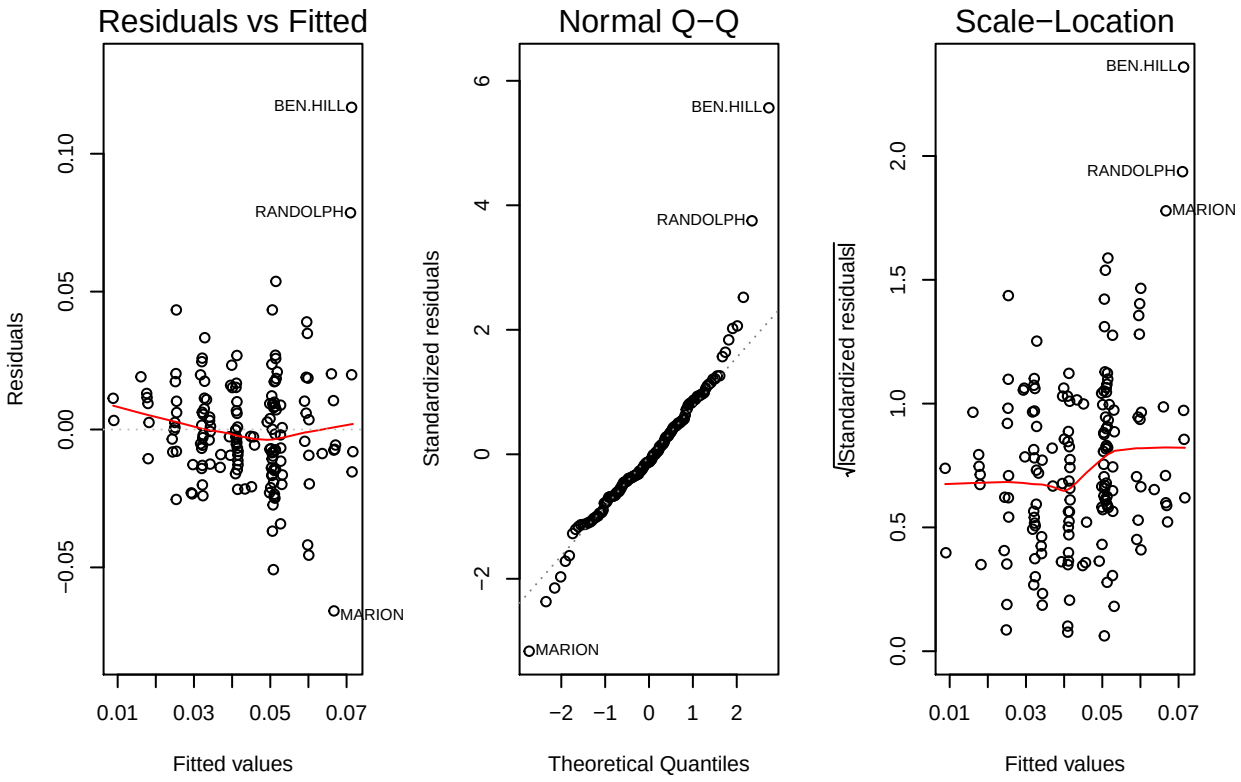
```
##        equipOS-CC        equipOS-PC        equipPAPER        equipPUNCH
##               "*"               "*"               " "               "*"
##        econmiddle          econrich             perAA        ruralurban
```

```
##                "*"              "*"               " "               "*"
## atlantanotAtlanta           pergore          perother
##                " "              " "               "*"
```

```
model.r2 <- lm(undercount~equip+econ+rural+pergore, gavote)
```

- Study the residue of `model.R2`.

```
par(mfrow=c(1,3))
plot(model.r2, which=1:3)
```



5. Study and refine of the final model

- Choose a model between `model.test`,`model.AIC`, `model.BIC` and `model.R2`. Make a complete diagnosis (residue, test hypothesis related to the linear model, Cook's distance, levers). You can use the functions `plot.lm`, `cook.distance`, `rstudent` ...

The 3 models are imbedded! They use the same variables that we already identified.

```
anova(model.BIC,model.AIC,model.r2)
```

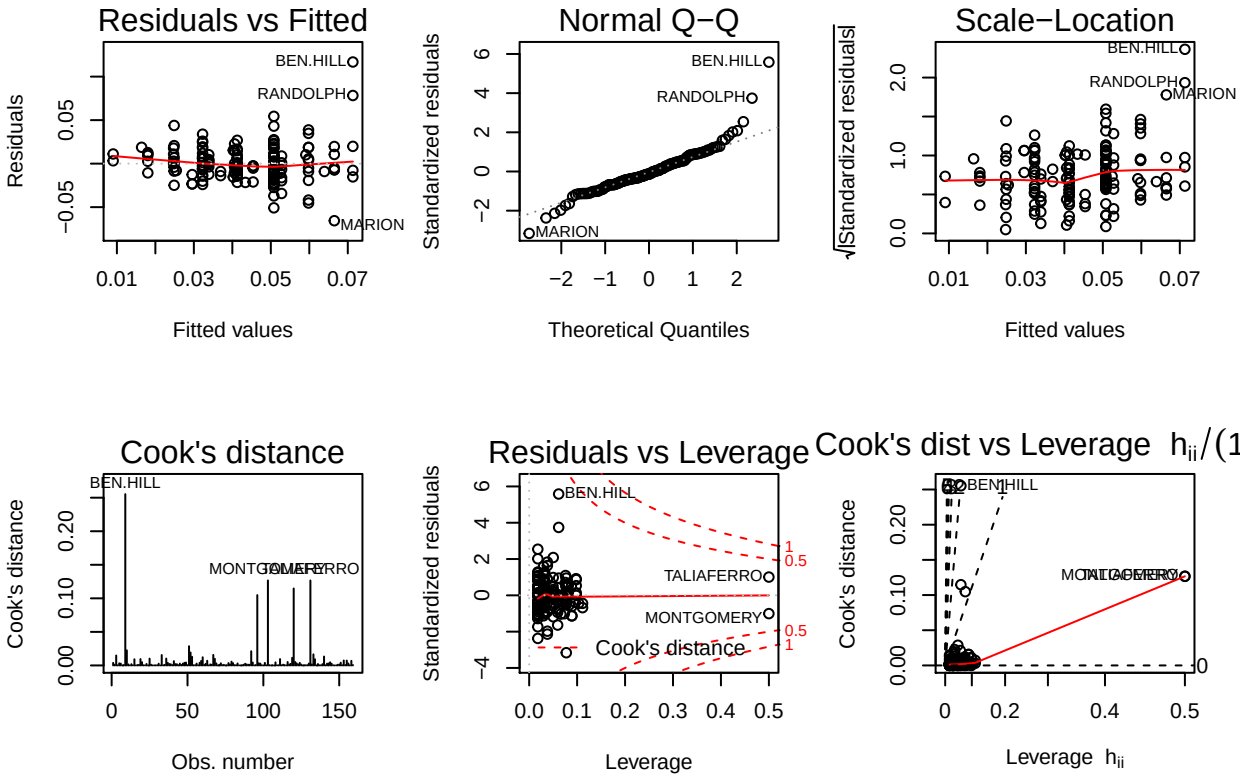
```
## Analysis of Variance Table
##
## Model 1: undercount ~ econ
## Model 2: undercount ~ equip + econ + rural
## Model 3: undercount ~ equip + econ + rural + pergore
```

```
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1    156 0.079296
## 2    151 0.070434  5 0.0088618 3.7760 0.002999 **
## 3    150 0.070405  1 0.0000289 0.0616 0.804291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Try to improve the fit to the data by excluding outliers, transforming the answer or some predictors. You can also merge the levels of categorical variable if they have similar effects.

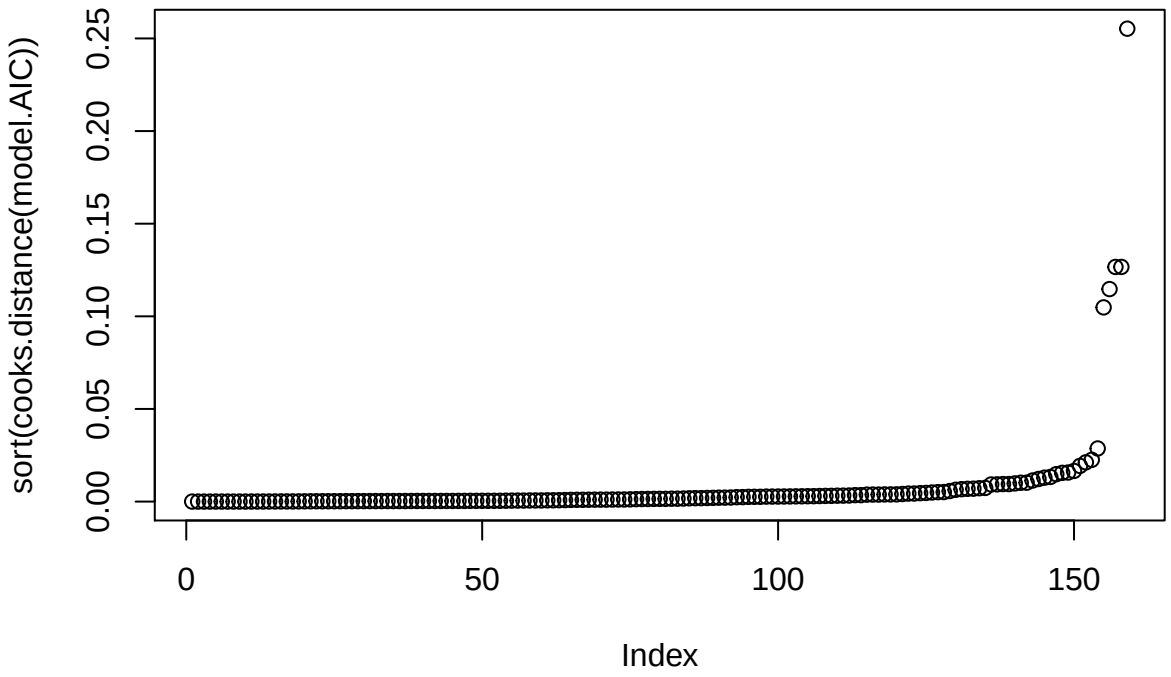
For example using the AIC model.

```
par(mfrow=c(2,3))
plot(model.AIC, which=1:6)
```



Some data-points are clearly outliers. They have high values for the Cook distance.

```
plot(sort(cooks.distance(model.AIC)))
```




```
remove <- order(cooks.distance(model.AIC), decreasing = TRUE)[1:5]
```

If we remove those data-points we get a better R^2 and the `rural` variable becomes significant at 5%

```
gavote2 <- gavote[-remove, ]
summary(lm(undercount~equip+econ+rural,gavote2))
```

```
##
## Call:
## lm(formula = undercount ~ equip + econ + rural, data = gavote2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.049471 -0.009938 -0.001223  0.009715  0.055697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.049471   0.002315  21.369  < 2e-16 ***
## equipOS-CC   0.007834   0.003423   2.289 0.023513 *
## equipOS-PC   0.009288   0.004546   2.043 0.042828 *
## equipPUNCH   0.019072   0.005092   3.746 0.000258 ***
## econmiddle  -0.015076   0.003260  -4.625 8.15e-06 ***
## econrich    -0.029872   0.005410  -5.522 1.48e-07 ***
## ruralurban  -0.007916   0.003807  -2.079 0.039319 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01716 on 147 degrees of freedom
## Multiple R-squared:  0.3218, Adjusted R-squared:  0.2941
## F-statistic: 11.63 on 6 and 147 DF,  p-value: 1.24e-10
```

In fact we remove one category of the `equip` variable!

```
table(gavote2$equip)
```

```
##
## LEVER OS-CC OS-PC PAPER PUNCH
##    74    44    20     0    16
```

We can try to recode this variable and fuse some categories. OS-CC and OS-PC have similar coefficients.

```
equip.new <- as.character(gavote2$equip)
equip.new[equip.new %in% c("OS-CC","OS-PC")] <- "OS"
gavote2$equip.new <- as.factor(equip.new )
```

We get a sligthly better R^2 and our model is easier to interpret (less categories).

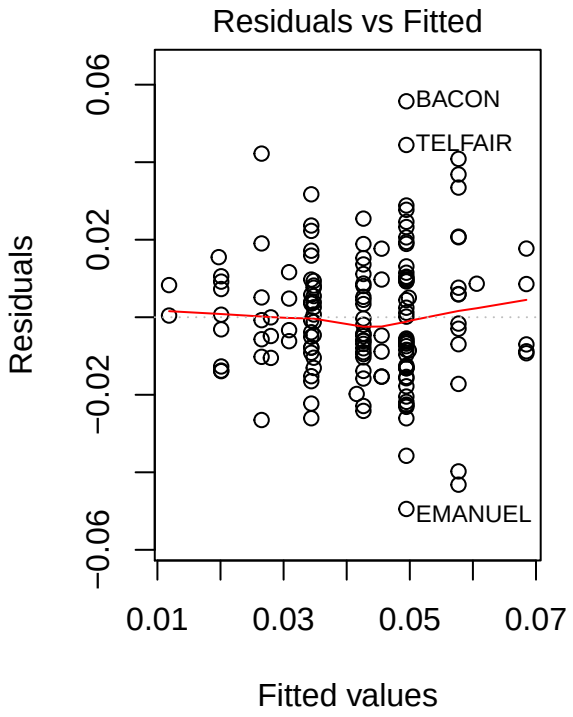
```
model.final <- lm(undercount~equip.new+econ+rural,gavote2)
summary(model.final)
```

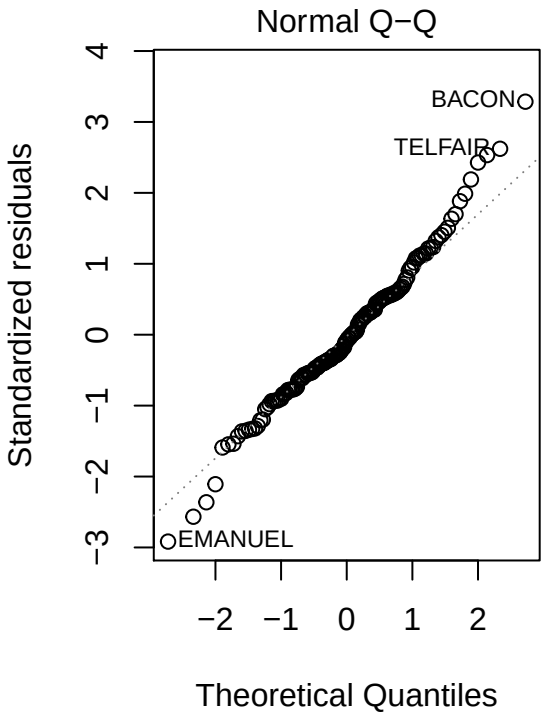
```
##
## Call:
## lm(formula = undercount ~ equip.new + econ + rural, data = gavote2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.049458 -0.010067 -0.001435  0.009371  0.055711
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.049458   0.002308  21.432  < 2e-16 ***
## equip.newOS     0.008261   0.003127   2.642 0.009133 **
## equip.newPUNCH  0.019021   0.005074   3.749 0.000254 ***
## econmiddle     -0.015071   0.003250  -4.637 7.69e-06 ***
## econrich       -0.029737   0.005376  -5.532 1.40e-07 ***
## ruralurban     -0.007874   0.003793  -2.076 0.039617 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01711 on 148 degrees of freedom
## Multiple R-squared:  0.3214, Adjusted R-squared:  0.2984
## F-statistic: 14.02 on 5 and 148 DF,  p-value: 3.246e-11
```

- Interpret the final model that you have.

The graph of residuals looks ok and the distribution is more or less symetric.

```
par(mfrow=c(1,2))
plot(model.final, which=1:2)
```





Residuals do not correlate with predictions.

```
durbinWatsonTest(model.final)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.07511671      1.829459   0.294
##  Alternative hypothesis: rho != 0
```

The distribution is not exactly Gaussian but that's okayish. (we could standardise the results).

```
shapiro.test(rstandard(model.final))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(model.final)
## W = 0.98517, p-value = 0.09868
```