Statistical learning

Linear regression

V. Lefieux

1 Libraries

```
library(car) #For VIF
library(leaps) #For variables selection
```

2 Data

```
setwd("/Users/vincentlefieux/Documents/Data/")
ozone=read.table("data_ozone.txt",header=TRUE,sep=";",dec=",")
attach(ozone)
```

3 Simple linear regression

3.1 Estimation

```
reg1=lm(maxO3~T12)
reg1
```

```
##
## Call:
## lm(formula = maxO3 ~ T12)
##
## Coefficients:
## (Intercept)          T12
##     -27.420        5.469
```

```
names(reg1)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
summary(reg1)
```

```
##
## Call:
## lm(formula = maxO3 ~ T12)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.079 -12.735   0.257  11.003  44.671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.4196     9.0335  -3.035    0.003 **
## T12           5.4687     0.4125  13.258   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.57 on 110 degrees of freedom
## Multiple R-squared:  0.6151, Adjusted R-squared:  0.6116
## F-statistic: 175.8 on 1 and 110 DF,  p-value: < 2.2e-16
```

```r
plot(T12,maxO3,xlab="T12",ylab="maxO3")
abline(reg1,col="red")
```

```r
par(mfrow=c(2,2))
plot(reg1)
```

```
par(mfrow=c(1,1))
yhat1=reg1$fit
plot(maxO3,yhat1,xlab="maxO3",ylab="Fitted maxO3")
abline(a=0,b=1,col="red")
```

It's also possible to use `yhat1=reg1$fitted.values` or `yhat1=predict(reg1)`.

```r
round(mean(reg1$residuals),digits=2)
```

```
## [1] 0
```

```r
round(var(reg1$residuals),digits=2)
```

```
## [1] 305.84
```

```r
boxplot(reg1$residuals,main="Residuals of reg1")
```

```r
hist(reg1$residuals,main="Residuals of reg1")
```

We choose 5% as level of tests:

```
alpha=0.05
```

Dimensions of the data set are:

```
n=dim(ozone)[1]
p=2
```

We compute studentized residuals:

```
rstudent1=rstudent(reg1)
threshold_stud1=qt(1-alpha/2,n-p-1)
plot(rstudent1,xlab="Observation",ylab="Studentized residuals")
abline(h=threshold_stud1,col="red")
abline(h=-threshold_stud1,col="red")
```

```
ids1=abs(rstudent1)>threshold_stud1
rstudent1[ids1]
```

```
##        17        25        34        79
##  2.210261 -2.216162  2.395229  2.646881
```

We compute Cook distances:

```
infl1=influence.measures(reg1)
names(infl1)
```

```
## [1] "infmat" "is.inf" "call"
```

```
colnames(infl1$infmat)
```

```
## [1] "dfb.1_"  "dfb.T12" "dffit"   "cov.r"   "cook.d"  "hat"
```

```
cook1=infl1$infmat[,"cook.d"]
threshold_cook1=4/(n-p)
plot(cook1,type="h",xlab="Observation",ylab="Cook distance")
abline(h=threshold_cook1,col="red")
```

```
idc1 <- cook1>threshold_cook1
cook1[idc1]
```

```
##         34         52         53         54         56         58
## 0.04196648 0.03637132 0.04316194 0.05956720 0.05199605 0.03818800
##         79
## 0.09102743
```

We remove outliers:

```
to_remove1=(1:n)[ids1*idc1==1]
ozone1=ozone[-to_remove1,]
reg2=lm(maxO3~T12,data=ozone1)
summary(reg2)
```

```
##
## Call:
## lm(formula = maxO3 ~ T12, data = ozone1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.241 -12.170   0.543  11.423  38.876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.7493     8.6949  -3.076  0.00265 **
## T12           5.4013     0.3975  13.590  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.72 on 108 degrees of freedom
## Multiple R-squared:  0.631,  Adjusted R-squared:  0.6276
## F-statistic: 184.7 on 1 and 108 DF,  p-value: < 2.2e-16
```

4 Multiple linear regression

4.1 Model selection

The complete regression is:

```
reg_compl=lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+maxO3v,data=ozone)
summary(reg_compl)
```

```
##
## Call:
## lm(formula = maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + maxO3v,
##     data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.768  -7.845  -1.359   8.134  38.984
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.70548   13.10860   0.969  0.33467
## T9          -0.63596    1.03462  -0.615  0.54011
## T12          2.50600    1.39946   1.791  0.07625 .
## T15          0.71381    1.13674   0.628  0.53142
## Ne9         -2.76057    0.89157  -3.096  0.00252 **
## Ne12        -0.37193    1.34590  -0.276  0.78283
## Ne15         0.09028    0.99934   0.090  0.92819
## maxO3v       0.37774    0.06121   6.171 1.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.43 on 104 degrees of freedom
## Multiple R-squared:  0.7546, Adjusted R-squared:  0.738
## F-statistic: 45.68 on 7 and 104 DF,  p-value: < 2.2e-16
```

We remove ne15:

```
reg_compl=lm(maxO3~T9+T12+T15+Ne9+Ne12+maxO3v,data=ozone)
summary(reg_compl)
```

```
##
## Call:
## lm(formula = maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + maxO3v, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.689  -7.796  -1.447   8.147  38.929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.84916   12.95015   0.992  0.32338
## T9          -0.62976    1.02745  -0.613  0.54124
## T12          2.56018    1.25841   2.034  0.04443 *
## T15          0.65787    0.94878   0.693  0.48960
## Ne9         -2.76526    0.88585  -3.122  0.00232 **
```

```
## Ne12        -0.30796    1.13912  -0.270  0.78742
## maxO3v       0.37752    0.06087   6.202 1.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.36 on 105 degrees of freedom
## Multiple R-squared:  0.7545, Adjusted R-squared:  0.7405
## F-statistic:  53.8 on 6 and 105 DF,  p-value: < 2.2e-16
```

We remove ne12:

```r
reg_compl=lm(maxO3~T9+T12+T15+Ne9+maxO3v,data=ozone)
summary(reg_compl)
```

```
##
## Call:
## lm(formula = maxO3 ~ T9 + T12 + T15 + Ne9 + maxO3v, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.246  -7.607  -1.295   8.285  38.477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.28440   11.53397   0.978   0.3301
## T9          -0.73127    0.95218  -0.768   0.4442
## T12          2.66487    1.19211   2.235   0.0275 *
## T15          0.66817    0.94386   0.708   0.4806
## Ne9         -2.92578    0.65451  -4.470 1.97e-05 ***
## maxO3v       0.37960    0.06012   6.314 6.48e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.3 on 106 degrees of freedom
## Multiple R-squared:  0.7544, Adjusted R-squared:  0.7428
## F-statistic: 65.11 on 5 and 106 DF,  p-value: < 2.2e-16
```

We remove T9:

```r
reg_compl=lm(maxO3~T12+T15+Ne9+maxO3v,data=ozone)
summary(reg_compl)
```

```
##
## Call:
## lm(formula = maxO3 ~ T12 + T15 + Ne9 + maxO3v, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.068  -7.767  -1.605   8.446  40.187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.13680   11.16838   0.818   0.4151
## T12          2.23175    1.04826   2.129   0.0355 *
## T15          0.62772    0.94058   0.667   0.5060
## Ne9         -2.96393    0.65137  -4.550 1.42e-05 ***
```

```
## maxO3v       0.37019    0.05875   6.301 6.71e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.27 on 107 degrees of freedom
## Multiple R-squared:  0.753,  Adjusted R-squared:  0.7438
## F-statistic: 81.55 on 4 and 107 DF,  p-value: < 2.2e-16
```

We remove T15:

```r
reg_compl=lm(maxO3~T12+Ne9+maxO3v,data=ozone)
summary(reg_compl)
```

```
##
## Call:
## lm(formula = maxO3 ~ T12 + Ne9 + maxO3v, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.385  -7.872  -1.941   7.899  41.513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.76225   11.10038   0.879    0.381
## T12          2.85308    0.48052   5.937 3.57e-08 ***
## Ne9         -3.02423    0.64342  -4.700 7.71e-06 ***
## maxO3v       0.37571    0.05801   6.477 2.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.23 on 108 degrees of freedom
## Multiple R-squared:  0.752,  Adjusted R-squared:  0.7451
## F-statistic: 109.1 on 3 and 108 DF,  p-value: < 2.2e-16
```

4.2 Automatic variables selection

With the *backward* method:

```r
regselect1=regsubsets(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+maxO3v,data=ozone,method="backward")
summary(regselect1)
```

```
## Subset selection object
## Call: regsubsets.formula(maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 +
##     maxO3v, data = ozone, method = "backward")
## 7 Variables  (and intercept)
##        Forced in Forced out
## T9         FALSE      FALSE
## T12        FALSE      FALSE
## T15        FALSE      FALSE
## Ne9        FALSE      FALSE
## Ne12       FALSE      FALSE
## Ne15       FALSE      FALSE
## maxO3v     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: backward
```

```
##          T9  T12 T15 Ne9 Ne12 Ne15 maxO3v
## 1  ( 1 ) " " "*" " " " " " "  " "  " "
## 2  ( 1 ) " " "*" " " " " " "  " "  "*"
## 3  ( 1 ) " " "*" " " "*" " "  " "  "*"
## 4  ( 1 ) "*" "*" " " "*" " "  " "  "*"
## 5  ( 1 ) "*" "*" "*" "*" " "  " "  "*"
## 6  ( 1 ) "*" "*" "*" "*" "*"  " "  "*"
## 7  ( 1 ) "*" "*" "*" "*" "*"  "*"  "*"
```

```
plot(regselect1)
plot(regselect1,scale="bic")
```

```
plot(regselect1,scale="Cp")
```

```
plot(regselect1,scale="adjr2")
```

With the *forward* method:

```
regselect2=regsubsets(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+maxO3v,data=ozone,method="forward")
summary(regselect2)
```

```
## Subset selection object
## Call: regsubsets.formula(maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 +
##     maxO3v, data = ozone, method = "forward")
## 7 Variables  (and intercept)
##        Forced in Forced out
## T9         FALSE      FALSE
## T12        FALSE      FALSE
## T15        FALSE      FALSE
## Ne9        FALSE      FALSE
## Ne12       FALSE      FALSE
## Ne15       FALSE      FALSE
## maxO3v     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: forward
##          T9  T12 T15 Ne9 Ne12 Ne15 maxO3v
## 1  ( 1 ) " " "*" " " " " " "  " "  " "
## 2  ( 1 ) " " "*" " " " " " "  " "  "*"
## 3  ( 1 ) " " "*" " " "*" " "  " "  "*"
## 4  ( 1 ) "*" "*" " " "*" " "  " "  "*"
## 5  ( 1 ) "*" "*" "*" "*" " "  " "  "*"
## 6  ( 1 ) "*" "*" "*" "*" "*"  " "  "*"
## 7  ( 1 ) "*" "*" "*" "*" "*"  "*"  "*"
```

```
plot(regselect2)
plot(regselect2,scale="bic")
```

```
plot(regselect2,scale="Cp")
```

```
plot(regselect2,scale="adjr2")
```

4.3 Predictive power of `reg_compl`

We partition the sample, estimate the model on 80% of the data set and calculate the forecast error on the 20% left.

```
n=dim(ozone)[1]
sel=sample(1:n,round(n*0.8))
ozone_estim=ozone[sel,]
ozone_pred=ozone[-sel,]
reg_compl_tronc=lm(maxO3~T12+Ne9+maxO3v-1,data=ozone_estim)
summary(reg_compl_tronc)
```

```
##
## Call:
## lm(formula = maxO3 ~ T12 + Ne9 + maxO3v - 1, data = ozone_estim)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.401  -7.782  -1.426   9.085  39.740
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## T12     3.08829    0.30615  10.087 2.68e-16 ***
## Ne9    -2.57916    0.53320  -4.837 5.64e-06 ***
## maxO3v  0.40871    0.06821   5.992 4.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.87 on 87 degrees of freedom
## Multiple R-squared:  0.9761, Adjusted R-squared:  0.9753
## F-statistic:  1186 on 3 and 87 DF,  p-value: < 2.2e-16
```

```
pred=predict(reg_compl_tronc,ozone_pred)
res_pred=ozone_pred$maxO3-pred
```

We can use for example the RMSE (Root Mean Square Error):

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2}.$$

or the MAPE (Mean Average Percentage Error):

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{y_t - \widehat{y}_t}{y_t}\right|.$$

```
rmse=sqrt(mean((res_pred)^2))
round(rmse,digits=2)
```

```
## [1] 11.48
```

```
mape=mean(abs(1-pred/ozone_pred$maxO3))*100
round(mape,digits=2)
```

```
## [1] 11.96
```