

Universitat Autònoma de Barcelona
Facultat de Ciències



PRÀCTICA 2

Authors:

Gerard Lahuerta & Ona Sánchez
1601350 — 1601181

23 de Març del 2022

Índex

0	Introduction	3
1	Exercise 1	4
1.1	Statement	4
1.2	R-commands and Analysis	4
2	Exercise 2	7
2.1	Statement	7
2.2	R-commands and Analysis	7
3	Exercise 3	10
3.1	Statement	10
3.2	R-commands and Analysis	10

0 Introduction

The data set we have chosen to work with is the EuStockMarkets. The data set contains the daily closing prices of major European stock indices: Germany DAX, Switzerland SMI, France CAC, and UK FTSE. The data are sampled in business time, i.e., weekends and holidays are omitted.

The data set has four columns: time (from 1991 to 1998), DAX, SMI, CAC and FTSE. It looks like this:

	DAX	SMI	CAC	FTSE
1991.496	1628.75	1678.1	1772.8	2443.6
1991.500	1613.63	1688.5	1750.5	2460.2
1991.504	1606.51	1678.6	1718.0	2448.2
1991.508	1621.04	1684.1	1708.1	2470.4
1991.512	1618.16	1686.6	1723.1	2484.7
...

For this exercises, we will work with SMI, CAC and FTSE as the possible predictors for the best fitting model, and DAX will be the response variable.

1 Exercise 1

1.1 Statement

Provide details of the chosen dataset. Design models to be analysed for this dataset.

1.2 R-commands and Analysis

Dataset initialization and details:

```
1 library(datasets)
2 data("EuStockMarkets")
3 esm = data.frame(EuStockMarkets)
4 modelfull<-lm(DAX~., data = esm)
5 summary(modelfull)
```

Output of the code:

Residuals:

	Min	1Q	Median	3Q	Max
	-335.26	-80.75	9.98	82.99	328.04

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-175.94567	44.66573	-3.939	8.48e-05 ***
SMI	0.49277	0.01532	32.163	< 2e-16 ***
CAC	0.49565	0.01544	32.105	< 2e-16 ***
FTSE	-0.01720	0.02089	-0.823	0.41

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.4 on 1856 degrees of freedom

Multiple R-squared: 0.9898, Adjusted R-squared: 0.9898

F-statistic: 6.032e+04 on 3 and 1856 DF, p-value: < 2.2e-16

Analysis:

We observe that without any changes to the model, it has a very small p -value and a high R -Squared so we can deduce that the modifications that we will do to the model by changing the predictors will not be really significant, so there will not be a lot of differences between them (with the exception of the variable *FTSE* that has a very high p -value so we can't say anything about how it could affect future models).

Models to be analysed:

```

1 Model = update(modelfull, ~.-SMI)
2 summary(Model)
3 Model = update(modelfull, ~.-CAC)
4 summary(Model)
5 Model = update(modelfull, ~.-FTSE)
6 summary(Model)
7
8 Model = update(modelfull, ~.-SMI-CAC)
9 summary(Model)
10 Model = update(modelfull, ~.-SMI-FTSE)
11 summary(Model)
12 Model = update(modelfull, ~.-FTSE-CAC)
13 summary(Model)

```

Output of the code:

```

Call:
lm(formula = DAX ~ CAC + FTSE, data = esm)

Residuals:
    Min       1Q   Median       3Q      Max
-427.33  -79.59    7.40   89.52  389.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.575e+03  1.260e+01 -125.00 <2e-16 ***
CAC          8.479e-01  1.358e-02   62.46 <2e-16 ***
FTSE        6.215e-01  8.956e-03   77.09 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136.5 on 1857 degrees of freedom
Multiple R-squared:  0.9842, Adjusted R-squared:  0.9842
F-statistic: 5.78e+04 on 2 and 1857 DF,  p-value: < 2.2e-16
#####

Call:
lm(formula = DAX ~ SMI + FTSE, data = esm)

Residuals:
    Min       1Q   Median       3Q      Max
-319.20 -101.46    -9.06   109.06   566.57

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.37022    37.42814    2.307 <2e-16 ***
SMI          0.84173    0.01346   62.52 <2e-16 ***
FTSE        -0.33572    0.02292  -14.65 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136.4 on 1857 degrees of freedom
Multiple R-squared:  0.9842, Adjusted R-squared:  0.9842
F-statistic: 5.787e+04 on 2 and 1857 DF,  p-value: < 2.2e-16
#####

Call:
lm(formula = DAX ~ SMI + CAC, data = esm)

Residuals:
    Min       1Q   Median       3Q      Max
-336.83  -79.21   10.15   82.37   326.60

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.102e+02  1.617e+01  -13.01 <2e-16 ***
SMI          4.809e-01  4.741e-03   101.42 <2e-16 ***
CAC          5.017e-01  1.358e-02   36.93 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.4 on 1857 degrees of freedom
Multiple R-squared:  0.9898, Adjusted R-squared:  0.9898
F-statistic: 9.05e+04 on 2 and 1857 DF,  p-value: < 2.2e-16
#####

Call:
lm(formula = DAX ~ FTSE, data = esm)

Residuals:
    Min       1Q   Median       3Q      Max
-408.43 -172.53  -45.71   137.68   989.96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.331e+03  2.109e+01  -63.12 <2e-16 ***
FTSE        1.083e+00  5.705e-03   189.84 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 240.3 on 1858 degrees of freedom
Multiple R-squared:  0.951, Adjusted R-squared:  0.9509
F-statistic: 3.604e+04 on 1 and 1858 DF,  p-value: < 2.2e-16
#####

Call:
lm(formula = DAX ~ CAC, data = esm)

Residuals:
    Min       1Q   Median       3Q      Max
-576.32 -250.24    2.49   245.09   548.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.493e+03  2.573e+01  -58.04 <2e-16 ***
CAC          1.806e+00  1.118e-02   161.62 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 279.6 on 1858 degrees of freedom
Multiple R-squared:  0.9336, Adjusted R-squared:  0.9336
F-statistic: 2.612e+04 on 1 and 1858 DF,  p-value: < 2.2e-16
#####

Call:
lm(formula = DAX ~ SMI, data = esm)

Residuals:
    Min       1Q   Median       3Q      Max
-285.76 -106.88   -20.15   104.20   603.45

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.478e+02  7.558e+00   46.02 <2e-16 ***

```

```

SMI          6.465e-01  2.008e-03  321.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 144 on 1858 degrees of freedom
Multiple R-squared:  0.9824, Adjusted R-squared:  0.9824
F-statistic: 1.036e+05 on 1 and 1858 DF,  p-value: < 2.2e-16

```

Analysis:

As we can see in the outputs, the value of the $p - value$ does not change depending on which predictors we use, it always stays in $< 2.2e - 16$. Instead, the *R-squared error* varies between 0.9336 (when the only predictor is CAC) and 0.9898 (when we use all the predictors).

We also have to consider the full model, analysed in section [1.2](#) as a model to be analysed.

2 Exercise 2

2.1 Statement

Apply backward selection to find the best fit model using p-value and AIC criteria. Compare the results found by both methods. Do the same for forward selection. Comment on the results.

2.2 R-commands and Analysis

Backward selection using p-value:

```
1 library(MASS)
2
3 Model=update(modelfull, .~-SMI)
4 summary(Model)
5 Model=update(modelfull, .~-CAC)
6 summary(Model)
7 Model=update(modelfull, .~-FTSE)
8 summary(Model)
9 Model=modelfull
10 summary(Model)
```

The output of the first three models have already been explained in section 1.2.

The output of the last model has already been explained in section 1.2.

Backward selection using AIC criteria:

```
1 modelbackward = stepAIC(modelfull, trace = TRUE, direction = "backward")
```

Output of the code:

```
Start:  AIC=17469.1
DAX ~ SMI + CAC + FTSE
```

	Df	Sum of Sq	RSS	AIC
- FTSE	1	8113	22217334	17468
<none>			22209221	17469
- CAC	1	12333661	34542881	18289
- SMI	1	12378364	34587584	18291

```
Step:  AIC=17467.78
DAX ~ SMI + CAC
```

	Df	Sum of Sq	RSS	AIC
<none>			22217334	17468
- CAC	1	16315505	38532840	18490
- SMI	1	123050941	145268275	20958

Forward selection using p-value:

```
1 colnames(esm)
2 modelnull = lm(DAX~1, data = EuStockMarkets)
3 summary(modelnull)
4 Model = update(modelnull, .~.+SMI)
5 summary(Model)
6 Model = update(modelnull, .~.+CAC)
7 summary(Model)
8 Model = update(modelnull, .~.+FTSE)
9 summary(Model)
10 Model = update(modelnull, .~.+SMI+CAC)
11 summary(Model)
12 Model = update(modelnull, .~.+SMI+FTSE)
13 summary(Model)
14 Model = update(modelnull, .~.+SMI+CAC+FTSE)
15 summary(Model)
```

The outputs of these models have already been explained in [1.2](#).

Forward selection using AIC criteria:

```
1 modelfull = formula(DAX~SMI+CAC+FTSE)
2 modelnull = lm(DAX~1, data=esm)
3 modelforward = stepAIC(modelnull, trace = TRUE, direction = "forward",
                        scope = modelfull)
```

Output of the code:

Start: AIC=26000.62

DAX ~ 1

	Df	Sum of Sq	RSS	AIC
+ SMI	1	2149092423	38532840	18490
+ FTSE	1	2080369991	107255272	20394
+ CAC	1	2042356988	145268275	20958
<none>			2187625263	26001

Step: AIC=18489.96

DAX ~ SMI

	Df	Sum of Sq	RSS	AIC
+ CAC	1	16315505	22217334	17468
+ FTSE	1	3989958	34542881	18289
<none>			38532840	18490

Step: AIC=17467.78

DAX ~ SMI + CAC

	Df	Sum of Sq	RSS	AIC
<none>			22217334	17468
+ FTSE	1	8113.4	22209221	17469

Analysis:

For the backward selection using $p - value$, we start with the full model, and we must remove the element with the highest $p - value$ each time. As we can see in the outputs, all the models have the same $p - value$ ($2.2 \cdot 10^{-16}$) so we can not remove any element, staying with the full model.

Moreover, for the forward selection (also using $p - value$), we start with no variables and we add the one with the lowest $p - value$ each time. As in the backward selection, we finally add all the variables because the $p - value$ in all the models are the same ($2.2 \cdot 10^{-16}$).

When we use the AIC criteria, for both cases (backward and forward) the result is the same: we must stay with the three predictors, using the full model.

3 Exercise 3

3.1 Statement

Find the best possible subset of variables to select the best fit model. Compare the results with the final models obtained in the previous point.

3.2 R-commands and Analysis

```
1 library(olsrr)
2 Model = lm(DAX~SMI+CAC+FTSE, data = EuStockMarkets)
3 olsstepallpossible(Model)
```

Output of the code:

Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
1	1	1	SMI	0.9823860	0.9823765
3	2	1	FTSE	0.9509718	0.9509454
2	3	1	CAC	0.9335954	0.9335597
4	4	2	SMI CAC	0.9898441	0.9898331
5	5	2	SMI FTSE	0.9842099	0.9841929
6	6	2	CAC FTSE	0.9841894	0.9841724
7	7	3	SMI CAC FTSE	0.9898478	0.9898314

```
1 olsstepbestsubset(Model)
```

Output of the code:

```
Best Subsets Regression
-----
Model Index Predictors
-----
1 SMI
2 SMI CAC
3 SMI CAC FTSE
-----
```

Subsets Regression Summary											
Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.9824	0.9824	0.9823	1364.1468	23770.4141	18489.9164	23786.9991	38574317.3445	20761.1802	11.1679	0.0177
2	0.9898	0.9898	0.9898	2.6780	22748.2273	17469.7867	22770.3406	22253232.8989	11983.3972	6.4462	0.0102
3	0.9898	0.9898	0.9898	4.0000	22749.5479	17471.1138	22777.1896	22257098.3979	11991.9087	6.4508	0.0102

AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria

```
1 library(leaps)
2 modelsubsets = regsubsets(DAX~SMI+CAC+FTSE, data=EuStockMarkets, nbest=2)
3 summary(modelsubsets)$which
```

Output of the code:

	(Intercept)	SMI	CAC	FTSE
1	TRUE	TRUE	FALSE	FALSE
1	TRUE	FALSE	FALSE	TRUE
2	TRUE	TRUE	TRUE	FALSE
2	TRUE	TRUE	FALSE	TRUE
3	TRUE	TRUE	TRUE	TRUE

Analysis:

We conclude (with the the outputs above) that our best variable sample are all the variables of the dataset (*SMI*, *CAC*, *FTSE*), the same conclusion that we obtained in exercise 2.

This is because of many things (for example): it has one of the lowest *R-Squared error*, it has the most accurate Mallow's CP (number of predictors plus the intercept; indicates that the model produces relatively precise and unbiased estimates) and its Final Prediction Error is quite good. The last output shows all the possible variable subset selection done by *R* (the last one is the best).