Chapter 1

Linear methods in R

1.1 Least Squares (LS) regression in R

1.1.1 Parameter estimation

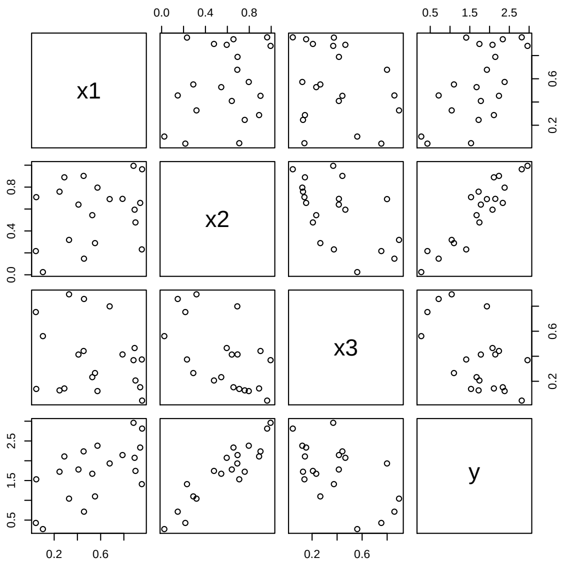

Figure 1.1: Multiple regression with simulated data: regression of y on three x-variables

- *Generation of the data*

```
>set.seed(123)
>x <- matrix(runif(60), ncol = 3)
>y <- x %*% c(1, 2, 0) + 0.1 * rnorm(20)
>colnames(x) = paste("x", 1:3, sep = "")
>d = data.frame(x, y = y)
>plot(d)
```

- *Model using only a constant term*

```
> lm0<-lm(y~1, data = d)
> lm0

Call:
lm(formula = y ~ 1, data = d)
```

```
Coefficients:
(Intercept)
       1.72
```

LS regression is computed by **lm()**. The estimated value of the intercept is $\beta_0 = 1.72$

- *Model with one explanatory variable*

```
> lm1<-lm(y~x1, data = d)
> lm1

Call:
lm(formula = y ~ x1, data = d)

Coefficients:
(Intercept)           x1
     0.9157       1.4600
```

- *Fit of a full model*

```
> lm3<-lm(y~x1+x2+x3, data = d)
> lm3

Call:
lm(formula = y ~ x1 + x2 + x3, data = d)

Coefficients:
(Intercept)           x1           x2           x3
    0.09585      0.91834      1.99804     -0.08761
```

1.1.2 Tests and confidence intervals

- *Testing the coefficients for significance*

```
> summary(lm3)

Call:
lm(formula = y ~ x1 + x2 + x3, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11566 -0.06133 -0.01260  0.06785  0.18004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09585    0.08200   1.169    0.260
x1           0.91834    0.06623  13.867 2.47e-10 ***
x2           1.99804    0.08453  23.637 7.18e-14 ***
x3          -0.08761    0.09060  -0.967    0.348
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08621 on 16 degrees of freedom
Multiple R-squared:  0.9882,       Adjusted R-squared:  0.986
F-statistic: 446.5 on 3 and 16 DF,  p-value: 1.251e-15
```

  - The t statistic of $x1$ and $x2$ is highly significant and the p-value of each variable is below 0.05. Therefore, both variables have a great impact on the explanation of the regressor and the null hypothesis can be rejected. The regressor $x3$ provides no significant additional contribution.

– The model provides a good fit (R squared), 98.82% of the variance of y can be explained by the model. The value 98.6% of the adjusted R squared is very high as well.

– *> qf(0.95, 3, 16)*

  [1] 3.238872

  The value of the "F statistic" of 446.5 is larger than the F quantile $F_{3,16;0.95} = 3.24$, therefore the null hypothesis $\beta_i = 0, \forall\ i = 1, \ldots, p$ can be rejected. This could also be concluded by the p-value that is close to 0.

– The test statistic from above can be used for the calculation of a confidence interval for $\hat{\beta}_j$. From the approximation of the 95% confidence interval, we obtain for $\hat{\beta}_1$ the interval

$$0.91834 \pm 2 * 0.06623 = [0.78, 1.06]$$

  and for $\hat{\beta}_3$

$$-0.08761 \pm 2 * 0.09060 = [-0.27, 0.09]$$

  The interval for $\hat{\beta}_1$ does not include zero, and thus the null hypothesis can be rejected at a 95% level. The interval for $\hat{\beta}_3$ includes zero, which confirms the acceptance of the null hypothesis due to a p-value of 0.348.

1.2 Variable selection in R

1.2.1 Model comparison with anova()

```
> anova(lm3)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 3.9799  3.9799 535.4639 9.991e-14 ***
x2         1 5.9693  5.9693 803.1073 4.199e-15 ***
x3         1 0.0070  0.0070   0.9351    0.3479
Residuals 16 0.1189  0.0074
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An F-test is computed for every additional explanatory variable, starting with the empty model and following the order of the formula. Regressor x3 does not improve the fit of the model and can be left out.

```
> lm2<-lm(y~x1+x2, data=d)
> anova(lm0, lm1, lm2, lm3)

Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1
Model 3: y ~ x1 + x2
Model 4: y ~ x1 + x2 + x3
  Res.Df     RSS Df Sum of Sq        F    Pr(>F)
```

```
1     19 10.0751
2     18  6.0951  1    3.9799 535.4639 9.991e-14 ***
3     17  0.1259  1    5.9693 803.1073 4.199e-15 ***
4     16  0.1189  1    0.0070   0.9351    0.3479
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here several nested models are compared in the specified order. This allows simultaneous testing of the significance of more than one parameter. Here, again, model lm3 does not improve the fit.

1.2.2 Body fat data

- *Scanning of the data and explanation of the variables*

```
> library("UsingR")
> data(fat)
> attach(fat)
> fat$body.fat[fat$body.fat == 0] <- NA
# exclude observations that are not used for the analysis
> fat <- fat[, -cbind(1, 3, 4, 9)]
# exclude a sample with wrong body height
> fat <- fat[-42, ]
# transform the body height in centimeter
> fat[, 4] <- fat[, 4] * 2.54
```

The data set “fat” consists of 15 physical measurements of 251 men. The data can be found in the `library(UsingR)`.

  - `body.fat`: percentage of body-fat calculated by Brozek's equation
  - `age`: age in years
  - `weight`: weight (in pounds)
  - `height`: height (in inches)
  - `BMI`: adiposity index
  - `neck`: neck circumference (cm)
  - `chest`: chest circumference (cm)
  - `abdomen`: abdomen circumference (cm)
  - `hip`: hip circumference (cm)
  - `thigh`: thigh circumference (cm)
  - `knee`: knee circumference (cm)
  - `ankle`: ankle circumference (cm)
  - `bicep`: extended biceps circumference (cm)
  - `forearm`: forearm circumference (cm)
  - `wrist`: wrist circumference (cm)

To measure the percentage of body-fat in the body, an extensive (and expensive) underwater technique has to be performed. The goal here is to establish a model which allows the prediction of the percentage of body-fat with easily measurable and collectible variables in order to avoid the underwater procedure. Nowadays, a new, very effortless method called bio-impedance analysis provides a reliable method to determine the body-fat percentage.

1.2.3 Full model

```
> model.lm<-lm(body.fat~., data = fat)
> summary(model.lm)

Call:
lm(formula = body.fat ~ ., data = fat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1062  -2.6605  -0.2011   2.8920   9.2619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -44.91075   36.67739  -1.224  0.22200
age           0.05740    0.03004   1.911  0.05725 .
weight       -0.16239    0.10076  -1.612  0.10838
height        0.17192    0.20001   0.860  0.39089
BMI           0.75340    0.73339   1.027  0.30534
neck         -0.42594    0.21857  -1.949  0.05251 .
chest        -0.05969    0.09907  -0.603  0.54740
abdomen       0.87126    0.08569  10.168  < 2e-16 ***
hip          -0.22543    0.13796  -1.634  0.10359
thigh         0.21780    0.13660   1.594  0.11220
knee         -0.01257    0.22965  -0.055  0.95639
ankle         0.12398    0.20837   0.595  0.55243
bicep         0.16357    0.16000   1.022  0.30769
forearm       0.39166    0.18627   2.103  0.03656 *
wrist        -1.49585    0.49586  -3.017  0.00284 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.988 on 235 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7432,        Adjusted R-squared:  0.7279
F-statistic: 48.58 on 14 and 235 DF,  p-value: < 2.2e-16
```

The coefficients `age`, `neck`, `abdomen`, `forearm` and `wrist` have very large t-values and very small p-values, therefore the null hypothesis $\beta_i = 0$ should be rejected. Due to the very small p-value of the F-statistic, the null hypothesis $\beta_i = 0, \ \forall \ i = 1, \ldots, p$ should be rejected as well. With an R squared = 0.7432 we can assume that the model provides a good fit.

1.2.4 Best subset regression with Leaps and Bound algorithm

```
> library(leaps)
> lm.regsubset<-regsubsets(body.fat~., data=fat, nbest = 1, nvmax = 8)
> summary(lm.regsubset)

Subset selection object
Call: regsubsets.formula(body.fat ~ ., data = fat, nbest = 1, nvmax = 8)
14 Variables  (and intercept)
        Forced in Forced out
age         FALSE      FALSE
weight      FALSE      FALSE
height      FALSE      FALSE
BMI         FALSE      FALSE
neck        FALSE      FALSE
chest       FALSE      FALSE
abdomen     FALSE      FALSE
hip         FALSE      FALSE
thigh       FALSE      FALSE
knee        FALSE      FALSE
ankle       FALSE      FALSE
bicep       FALSE      FALSE
forearm     FALSE      FALSE
```

```
wrist      FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         age weight height BMI neck chest abdomen hip thigh knee ankle bicep
1  ( 1 ) " " " "    " "    " " " "  " "   "*"     " " " "   " "  " "   " "
2  ( 1 ) " " "*"    " "    " " " "  " "   "*"     " " " "   " "  " "   " "
3  ( 1 ) " " "*"    " "    " " " "  " "   "*"     " " " "   " "  " "   " "
4  ( 1 ) " " "*"    " "    " " " "  " "   "*"     " " " "   " "  " "   " "
5  ( 1 ) " " "*"    " "    " " "*"  " "   "*"     " " " "   " "  " "   " "
6  ( 1 ) " " "*"    " "    " " "*"  " "   "*"     " " " "   " "  " "   "*"
7  ( 1 ) "*" "*"    " "    " " "*"  " "   "*"     " " "*"   " "  " "   " "
8  ( 1 ) "*" "*"    " "    " " "*"  " "   "*"     "*" "*"   " "  " "   " "
         forearm wrist
1  ( 1 ) " "     " "
2  ( 1 ) " "     " "
3  ( 1 ) " "     "*"
4  ( 1 ) "*"     "*"
5  ( 1 ) "*"     "*"
6  ( 1 ) "*"     "*"
7  ( 1 ) "*"     "*"
8  ( 1 ) "*"     "*"
```

regsubsets() in `library(leaps)` provides the "best" model for different sizes of subsets. Here only one "best" model per subset size was considered. The ranking of the models is done using the BIC measure.

```
> lm.regsubset2<-regsubsets(body.fat~., data=fat, nbest = 2, nvmax = 8)
> plot(lm.regsubset2)
```

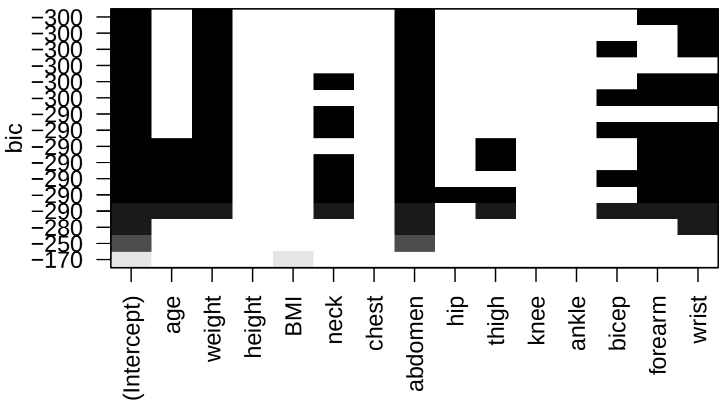

Figure 1.2: Model selection with **leaps()**

This plot shows the two best, by **regsubsets()** computed models with 1-8 regressors each. The BIC, coded in grey scale, does not improve after the fifth stage (starting from the bottom, see Figure 1.2). The optimal model can then be chosen from the models with "saturated" grey, and preferable that model is taken with the smallest number of variables.

1.2.5 Stepwise selection - automatic model search

- *Stepwise selection with* **drop1()**

```
> drop1(model.lm, test="F")
Single term deletions

Model:
body.fat ~ age + weight + height + BMI + neck + chest + abdomen +
    hip + thigh + knee + ankle + bicep + forearm + wrist
        Df Sum of Sq    RSS    AIC  F value    Pr(>F)
<none>                3738.3 706.23
age      1     58.08 3796.4 708.09   3.6511  0.057249 .
weight   1     41.32 3779.6 706.98   2.5974  0.108379
height   1     11.75 3750.0 705.02   0.7389  0.390892
BMI      1     16.79 3755.1 705.35   1.0553  0.305339
neck     1     60.41 3798.7 708.24   3.7978  0.052509 .
chest    1      5.78 3744.1 704.62   0.3630  0.547401
abdomen  1   1644.60 5382.9 795.38 103.3844 < 2.2e-16 ***
hip      1     42.47 3780.8 707.06   2.6700  0.103595
thigh    1     40.44 3778.7 706.92   2.5419  0.112202
knee     1      0.05 3738.3 704.23   0.0030  0.956395
ankle    1      5.63 3743.9 704.61   0.3540  0.552429
bicep    1     16.62 3754.9 705.34   1.0451  0.307694
forearm  1     70.33 3808.6 708.89   4.4213  0.036558 *
wrist    1    144.76 3883.1 713.73   9.1002  0.002837 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(update(model.lm,.~.-knee))
Call:
lm(formula = body.fat ~ age + weight + height + BMI + neck +
    chest + abdomen + hip + thigh + ankle + bicep + forearm +
    wrist, data = fat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0922  -2.6545  -0.1914   2.9011   9.2520

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.01721   36.54833  -1.232  0.21928
age           0.05699    0.02907   1.961  0.05107 .
weight       -0.16288    0.10014  -1.627  0.10516
height        0.17148    0.19942   0.860  0.39072
BMI           0.75481    0.73139   1.032  0.30312
neck         -0.42464    0.21682  -1.959  0.05135 .
chest        -0.05961    0.09885  -0.603  0.54704
abdomen       0.87123    0.08551  10.189  < 2e-16 ***
hip          -0.22594    0.13735  -1.645  0.10132
thigh         0.21554    0.12999   1.658  0.09862 .
ankle         0.12186    0.20432   0.596  0.55147
bicep         0.16398    0.15949   1.028  0.30491
forearm       0.39080    0.18520   2.110  0.03590 *
wrist        -1.49797    0.49329  -3.037  0.00266 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.98 on 236 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7432,       Adjusted R-squared:  0.7291
F-statistic: 52.54 on 13 and 236 DF,  p-value: < 2.2e-16
```

Elimination of the least significant variable, in this case `knee` is excluded from the model. The R squared (and adjusted R squared) do not change, the fit remains the same.

- *Automatic model search with* **step()**

```
>model.lmstep<-step(model.lm)
Start:  AIC=706.23
body.fat ~ age + weight + height + BMI + neck + chest + abdomen +
```

```
    hip + thigh + knee + ankle + bicep + forearm + wrist

          Df Sum of Sq    RSS    AIC
- knee     1   0.04766 3738.3  704.2
- ankle    1       5.6 3743.9  704.6
- chest    1       5.8 3744.1  704.6
- height   1      11.8 3750.0  705.0
- bicep    1      16.6 3754.9  705.3
- BMI      1      16.8 3755.1  705.4
<none>                 3738.3  706.2
- thigh    1      40.4 3778.7  706.9
- weight   1      41.3 3779.6  707.0
- hip      1      42.5 3780.8  707.1
- age      1      58.1 3796.4  708.1
- neck     1      60.4 3798.7  708.2
- forearm  1      70.3 3808.6  708.9
- wrist    1     144.8 3883.1  713.7
- abdomen  1    1644.6 5382.9  795.4

Step:  AIC=704.23
body.fat ~ age + weight + height + BMI + neck + chest + abdomen +
    hip + thigh + ankle + bicep + forearm + wrist

          Df Sum of Sq    RSS    AIC
- ankle    1       5.6 3744.0  702.6
- chest    1       5.8 3744.1  702.6
- height   1      11.7 3750.1  703.0
- bicep    1      16.7 3755.1  703.4
- BMI      1      16.9 3755.2  703.4
<none>                 3738.3  704.2
- weight   1      41.9 3780.3  705.0
- hip      1      42.9 3781.2  705.1
- thigh    1      43.6 3781.9  705.1
- neck     1      60.8 3799.1  706.3
- age      1      60.9 3799.3  706.3
- forearm  1      70.5 3808.9  706.9
- wrist    1     146.1 3884.4  711.8
- abdomen  1    1644.6 5382.9  793.4
                             ⋮
Step:  AIC=697.41
body.fat ~ age + weight + neck + abdomen + hip + thigh + forearm +
    wrist

          Df Sum of Sq    RSS    AIC
<none>                 3786.2  697.4
- hip      1      37.4 3823.6  697.9
- age      1      59.3 3845.5  699.3
- neck     1      61.2 3847.4  699.4
- weight   1      74.7 3860.9  700.3
- thigh    1      77.5 3863.7  700.5
- forearm  1     114.0 3900.2  702.8
- wrist    1     135.8 3922.1  704.2
- abdomen  1    2712.5 6498.7  830.5

Call:
lm(formula = body.fat ~ age + weight + neck + abdomen + hip +
 thigh + forearm + wrist, data = fat)

Coefficients:
(Intercept)          age       weight         neck      abdomen          hip
  -18.46826      0.05577     -0.08081     -0.41183      0.87775     -0.20063
      thigh      forearm        wrist
    0.26719      0.46567     -1.39341
```

step() calls **add1()** and **drop1()** as long as the AIC cannot be reduced further.

- *Comparison of the models with* **anova()**

```
> anova(model.lm, model.lm1,model.lmstep)
Analysis of Variance Table

Model 1: body.fat ~ age + weight + height + BMI + neck + chest + abdomen +
    hip + thigh + knee + ankle + bicep + forearm + wrist
Model 2: body.fat ~ age + weight + height + BMI + neck + chest + abdomen +
    hip + thigh + ankle + bicep + forearm + wrist
Model 3: body.fat ~ age + weight + neck + abdomen + hip + thigh + forearm +
    wrist
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    235 3738.3
2    236 3738.3 -1    -0.048 0.0030 0.9564
3    241 3786.2 -5   -47.861 0.6017 0.6987
```

By using the smaller model model.lmstep no essential information is lost, therefore it can be used for the prediction instead of model.lm.