Section 14.6: Building a Regression Model

Building a Regression Model Using Forward Selection

To build a regression model in R using one of the stepwise selection methods, use the following steps. Following along Example 2 in Section 14.6, we want to find the best regression model to describe miles per gallon.

Step 1: First, type or import you dataset into R. Table 10 from Example 2 in Section 14.6 has been imported.

Step 2: Establish your base models. In R, we must establish an empty model and a full model where every variable is used:

model1 <- lm(Miles.Per.Gallon ~ 1, data = Table10)

model2 <- lm(Miles.Per.Gallon ~ Weight + Engine + Cylinders + Drive + Clearance, data = Table10)

Note: *model1* is the empty model that will the be the start of a forward selection process. *model2* is the full model that will start the backward elimination process.

```
model1 <- lm(Miles.Per.Gallon ~ 1, data = Table10)
model2 <- lm(Miles.Per.Gallon ~ Weight + Engine + Cylinders + Drive + Clearan
ce, data = Table10)
```

Step 3: Use the following command to select your model using forward selection:

step(model1, direction = "forward", scope = list(lower=model1, upper=model2), test = "F")

Note: You begin the command with *model1* because that is the empty model that starts forward selection. If you do not want to see all of the results in R, you can add *trace=0* to the command.

```
step(model1, direction = "forward", scope = list(lower=model1, upper=model2),
test = "F")

## Start:  AIC=21.47
## Miles.Per.Gallon ~ 1
##
##             Df Sum of Sq    RSS     AIC F value    Pr(>F)
## + Engine     1    34.774 20.159  8.4341 22.4249 0.0003895 ***
## + Weight     1    30.798 24.135 11.1342 16.5891 0.0013183 **
## + Cylinders  1    22.222 32.711 15.6950  8.8315 0.0108096 *
## + Drive      1    14.631 40.303 18.8255  4.7193 0.0489128 *
## <none>                   54.933 21.4711
## + Clearance  1     6.686 48.247 21.5243  1.8015 0.2025008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Step:  AIC=8.43
## Miles.Per.Gallon ~ Engine
##
##             Df Sum of Sq    RSS     AIC F value Pr(>F)
## + Drive      1    7.7718 12.387  3.1293  7.5288 0.0178 *
## <none>                   20.159  8.4341
## + Weight     1    0.7280 19.431  9.8823  0.4496 0.5152
## + Cylinders  1    0.3313 19.828 10.1855  0.2005 0.6623
## + Clearance  1    0.0055 20.154 10.4300  0.0033 0.9553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=3.13
## Miles.Per.Gallon ~ Engine + Drive
##
##             Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                   12.387 3.1293
## + Cylinders  1  0.221780 12.165 4.8583  0.2005 0.6630
## + Clearance  1  0.012335 12.375 5.1143  0.0110 0.9185
## + Weight     1  0.000170 12.387 5.1291  0.0002 0.9904

##
## Call:
## lm(formula = Miles.Per.Gallon ~ Engine + Drive, data = Table10)
##
## Coefficients:
## (Intercept)       Engine        Drive
##      28.683       -1.516       -1.647
```

This command starts with the empty model and adds variables one by one. The selection process adds *Engine* first, because it has the highest F-statistic. It then adds *Drive* and stops.

$$MPG = 28.68 - 1.516 * Engine - 1.647 * Drive$$

Building a Regression Model Using Backward Elimination

Use the following command to select the best model using backward elimination:

step(model2, direction = "backward", scope = list(lower=model1, upper=model2), test = "F")

Note: You begin the command with *model2* because that is the full model that starts backward elimination.

```
step(model2, direction = "backward", scope = list(lower=model1, upper=model2)
, test = "F")
```

```
## Start:  AIC=8.83
## Miles.Per.Gallon ~ Weight + Engine + Cylinders + Drive + Clearance
##
##             Df Sum of Sq    RSS     AIC F value  Pr(>F)
## - Weight     1    0.0011 12.146  6.8346  0.0008 0.97829
```

```
## - Clearance  1    0.0191 12.164  6.8569  0.0142 0.90791
## - Cylinders  1    0.2295 12.375  7.1142  0.1701 0.68969
## <none>                  12.145  8.8333
## - Engine     1    2.6350 14.780  9.7786  1.9526 0.19580
## - Drive      1    7.0180 19.163 13.6742  5.2005 0.04853 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=6.83
## Miles.Per.Gallon ~ Engine + Cylinders + Drive + Clearance
##
##             Df Sum of Sq    RSS     AIC F value  Pr(>F)
## - Clearance  1    0.0192 12.165  4.8583  0.0158 0.90252
## - Cylinders  1    0.2286 12.375  5.1143  0.1882 0.67362
## <none>                  12.146  6.8346
## - Engine     1    4.1869 16.333  9.2773  3.4471 0.09303 .
## - Drive      1    7.6732 19.820 12.1792  6.3173 0.03073 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=4.86
## Miles.Per.Gallon ~ Engine + Cylinders + Drive
##
##             Df Sum of Sq    RSS     AIC F value  Pr(>F)
## - Cylinders  1    0.2218 12.387  3.1293  0.2005 0.66298
## <none>                  12.165  4.8583
## - Engine     1    4.7421 16.908  7.7957  4.2878 0.06270 .
## - Drive      1    7.6623 19.828 10.1855  6.9282 0.02332 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=3.13
## Miles.Per.Gallon ~ Engine + Drive
##
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                12.387  3.1293
## - Drive   1    7.7718 20.159  8.4341  7.5288 0.0178032 *
## - Engine  1   27.9154 40.303 18.8255 27.0427 0.0002218 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = Miles.Per.Gallon ~ Engine + Drive, data = Table10)
##
## Coefficients:
## (Intercept)       Engine        Drive
##      28.683       -1.516       -1.647
```

This command starts with the full model and subtracts variables one by one. The selection process subtracts *Weight* first, because it has the lowest F-statistic. It then subtracts *Clearance* and *Cylinders*, which leaves *Engine* and *Drive* as the variables in the final model.

$$MPG = 28.68 - 1.516 * Engine - 1.647 * Drive$$

Building a Regression Model Using Stepwise Regression

Use the following command to select your model using stepwise regression:

step(model1, direction = "both", scope = list(lower=model1, upper=model2), test = "F")

Note: In Stepwise Regression, you begin the command with either *model1* or *model2* because Stepwise Regression can add and drop variables, unlike Forward Selection and Backward Elimination.

```
step(model1, direction = "both", scope = list(lower=model1, upper=model2), te
st = "F")
```

```
## Start:  AIC=21.47
## Miles.Per.Gallon ~ 1
##
##             Df Sum of Sq    RSS     AIC F value    Pr(>F)
## + Engine     1    34.774 20.159  8.4341 22.4249 0.0003895 ***
## + Weight     1    30.798 24.135 11.1342 16.5891 0.0013183 **
## + Cylinders  1    22.222 32.711 15.6950  8.8315 0.0108096 *
## + Drive      1    14.631 40.303 18.8255  4.7193 0.0489128 *
## <none>                  54.933 21.4711
## + Clearance  1     6.686 48.247 21.5243  1.8015 0.2025008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=8.43
## Miles.Per.Gallon ~ Engine
##
##             Df Sum of Sq    RSS     AIC F value    Pr(>F)
## + Drive      1     7.772 12.387  3.1293  7.5288 0.0178032 *
## <none>                  20.159  8.4341
## + Weight     1     0.728 19.431  9.8823  0.4496 0.5152221
## + Cylinders  1     0.331 19.828 10.1855  0.2005 0.6622993
## + Clearance  1     0.005 20.154 10.4300  0.0033 0.9553130
## - Engine     1    34.774 54.933 21.4711 22.4249 0.0003895 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=3.13
## Miles.Per.Gallon ~ Engine + Drive
##
##             Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                  12.387  3.1293
```

```
## + Cylinders  1    0.2218 12.165  4.8583  0.2005 0.6629816
## + Clearance  1    0.0123 12.375  5.1143  0.0110 0.9184887
## + Weight     1    0.0002 12.387  5.1291  0.0002 0.9904081
## - Drive      1    7.7718 20.159  8.4341  7.5288 0.0178032 *
## - Engine     1   27.9154 40.303 18.8255 27.0427 0.0002218 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = Miles.Per.Gallon ~ Engine + Drive, data = Table10)
##
## Coefficients:
## (Intercept)       Engine        Drive
##      28.683       -1.516       -1.647
```

Similar to the Forward Selection example, this command adds *Engine* and then *Drive* the the empty model.

$$MPG = 28.68 - 1.516 * Engine - 1.647 * Drive$$