Additional Exercise 12

Jonas Moss

October 1, 2018

Loading

We begin by loading the data.

```
url = "http://www.uio.no/studier/emner/matnat/math/STK3100/data/lowbirthweight.txt"
lowbirthweight = read.table(file = url,
                            header = TRUE)
rm(url)
```

And take a look at the first few lines of the data frame.

```
head(lowbirthweight)
```

```
##   low age lwt  race smoke ht ui ftv ptl  bwt
## 1   0  19 182 black     0  0  1   0   0 2523
## 2   0  33 155 other     0  0  0   1   0 2551
## 3   0  20 105 white     1  0  0   1   0 2557
## 4   0  21 108 white     1  0  1   1   0 2594
## 5   0  18 107 white     1  0  1   0   0 2600
## 6   0  21 124 other     0  0  0   0   0 2622
```

The data frame includes a covariate `bwt` that we should ignore.

The full model

We'll take a look at the full model first.

```
full_model = glm(formula = low ~ . - bwt,
                 data = lowbirthweight,
                 family = binomial(link = "logit"))
```

Using `~ .` in a formula tells `R` to use all the covariates in a data frame (except the response). The minus sign removes the covariate `bwt`.

Backwards elimination

I have made a function to do backwards elimination with the `drop1` function.

First we will need the following, which returns the identity of the worst covariate.

```
#' Find the worst covariate of a model in terms of significance
#'
#' @param model A model object compatible with \code{drop1}.
#' @param ... Additional arguments to be passed to \code{drop1}.
#' @return A list containing the name of the worst variable and its significance
#'      level.

get_worst = function(model, ...) {

  anova_object = drop1(model, ...)
```

```r
  significances = tail(anova_object$Pr, -1)
  variables = tail(rownames(anova_object), -1)

  list(variable = variables[which.max(significances)],
       significance = max(significances))
}
```

Messing around with Rs formula object, we can write

```r
#' Backwards elimination.
#'
#' @param model A model object compatible with \code{drop1}.
#' @param level The desired significance level when elimination is stopped.
#' @param ... Additional arguments to be passed to \code{drop1}.
#' @return A list containing the final model, the ordered vector of removed
#'    variables, and an ordered vector of their significacne levels.

backwards_elimination = function(model, level, ...) {

  top_model = drop1(model, ...)
  variables = c()
  significances = c()

  for(i in 1:(nrow(top_model) - 1)) {

    worst = get_worst(model, test = "LRT")
    variables[i] = worst$variable
    significances[i] = worst$significance

    if(significances[i] <= level) break;

    call = model$call
    call[[2]] = update(old = formula(deparse(formula(model))),
                       new = paste0("~ . -",  variables[i]))
    model = eval(call)

  }

  return(list(model = model,
              variables = variables,
              significances = significances))

}
```

Applying backwards elimination

Using `backwards_elimination` we can do exercise b:

```r
backwards_elimination(full_model, level = 0.05, test = "LRT")
```

```
## $model
##
## Call:  glm(formula = low ~ lwt + race + smoke + ht + ptl, family = binomial(link = "logit"),
##     data = lowbirthweight)
##
```

```
## Coefficients:
## (Intercept)           lwt    raceother     racewhite          smoke
##      1.3217       -0.0165      -0.3955       -1.2584         0.8801
##          ht           ptl
##      1.7620        1.2312
##
## Degrees of Freedom: 188 Total (i.e. Null);  182 Residual
## Null Deviance:         234.7
## Residual Deviance: 200.6      AIC: 214.6
##
## $variables
## [1] "ftv"    "age"    "ui"     "smoke"
##
## $significances
## [1] 0.74004706 0.30202831 0.10285867 0.02584185
```

And we can try a different level:

```
backwards_elimination(full_model, level = 0.01, test = "LRT")
```

```
## $model
##
## Call:  glm(formula = low ~ lwt + ht + ptl, family = binomial(link = "logit"),
##     data = lowbirthweight)
##
## Coefficients:
## (Intercept)           lwt           ht          ptl
##     0.98799      -0.01706      1.88841      1.40712
##
## Degrees of Freedom: 188 Total (i.e. Null);  185 Residual
## Null Deviance:         234.7
## Residual Deviance: 210.3      AIC: 218.3
##
## $variables
## [1] "ftv"    "age"    "ui"     "smoke" "race"  "ht"
##
## $significances
## [1] 0.740047062 0.302028310 0.102858671 0.025841853 0.096744169 0.006725508
```

3