BB503/BB602 - R Training - Week XIII

Ege Ulgen

Logistic Regression

The data we'll use is `birthwt` from the `MASS` package. The `birthwt` data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

```r
# install.packages("MASS")
library(MASS)

data(birthwt)
?birthwt

dim(birthwt)
```

```
## [1] 189  10
```

```r
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

```r
# turn categorical variables into factor
birthwt$low <- as.factor(birthwt$low)
birthwt$race <- as.factor(birthwt$race)
birthwt$smoke <- as.factor(birthwt$smoke)
birthwt$ht <- as.factor(birthwt$ht)
birthwt$ui <- as.factor(birthwt$ui)

summary(birthwt)
```

```
##  low          age            lwt      race   smoke        ptl         ht
##  0:130   Min.   :14.0   Min.   : 80   1:96   0:115   Min.   :0.000   0:177
##  1: 59   1st Qu.:19.0   1st Qu.:110   2:26   1: 74   1st Qu.:0.000   1: 12
##          Median :23.0   Median :121   3:67           Median :0.000
##          Mean   :23.2   Mean   :130                  Mean   :0.196
##          3rd Qu.:26.0   3rd Qu.:140                  3rd Qu.:0.000
##          Max.   :45.0   Max.   :250                  Max.   :3.000
##  ui           ftv             bwt
##  0:161   Min.   :0.000   Min.   : 709
##  1: 28   1st Qu.:0.000   1st Qu.:2414
##          Median :0.000   Median :2977
##          Mean   :0.794   Mean   :2945
##          3rd Qu.:1.000   3rd Qu.:3487
##          Max.   :6.000   Max.   :4990
```

We'll be using logistic regression to identify risk factors associated with low infant birth weight (birth weight less than 2.5 kg).

```
fit0 <- glm(low~.-bwt, data = birthwt, family = binomial)
summary(fit0)
```

```
##
## Call:
## glm(formula = low ~ . - bwt, family = binomial, data = birthwt)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.895  -0.821  -0.532   0.982   2.212
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.48062    1.19689    0.40   0.6880
## age         -0.02955    0.03703   -0.80   0.4249
## lwt         -0.01542    0.00692   -2.23   0.0258 *
## race2        1.27226    0.52736    2.41   0.0158 *
## race3        0.88050    0.44078    2.00   0.0458 *
## smoke1       0.93885    0.40215    2.33   0.0196 *
## ptl          0.54334    0.34540    1.57   0.1157
## ht1          1.86330    0.69753    2.67   0.0076 **
## ui1          0.76765    0.45932    1.67   0.0947 .
## ftv          0.06530    0.17239    0.38   0.7048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 201.28  on 179  degrees of freedom
## AIC: 221.3
##
## Number of Fisher Scoring iterations: 4
```

We'll use only the significant variables:

```
fit1 <- glm(low~lwt + race + smoke + ht, data = birthwt, family = binomial)
summary(fit1)
```

```
##
## Call:
## glm(formula = low ~ lwt + race + smoke + ht, family = binomial,
##     data = birthwt)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.775  -0.875  -0.571   0.963   2.113
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3520     0.9244    0.38   0.7033   
## lwt          -0.0179     0.0068   -2.63   0.0084 **
## race2         1.2877     0.5216    2.47   0.0136 * 
## race3         0.9436     0.4234    2.23   0.0258 * 
## smoke1        1.0716     0.3875    2.77   0.0057 **
## ht1           1.7492     0.6908    2.53   0.0113 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 208.25  on 183  degrees of freedom
## AIC: 220.2
## 
## Number of Fisher Scoring iterations: 4
```

The final model:

```
fit_final <- glm(low~I(lwt - min(lwt)) + race + smoke + ht, data = birthwt, family = binomial)
summary(fit_final)
```

```
## 
## Call:
## glm(formula = low ~ I(lwt - min(lwt)) + race + smoke + ht, family = binomial, 
##     data = birthwt)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -1.775  -0.875  -0.571   0.963   2.113  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)   
## (Intercept)         -1.0805     0.4829   -2.24   0.0253 * 
## I(lwt - min(lwt))  -0.0179     0.0068   -2.63   0.0084 **
## race2               1.2877     0.5216    2.47   0.0136 * 
## race3               0.9436     0.4234    2.23   0.0258 * 
## smoke1              1.0716     0.3875    2.77   0.0057 **
## ht1                 1.7492     0.6908    2.53   0.0113 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 208.25  on 183  degrees of freedom
## AIC: 220.2
## 
## Number of Fisher Scoring iterations: 4
```

```
coef(fit_final)
```

```
##       (Intercept) I(lwt - min(lwt))             race2             race3 
##         -1.080477         -0.017907          1.287662          0.943645 
##            smoke1               ht1 
##          1.071566          1.749163 
```

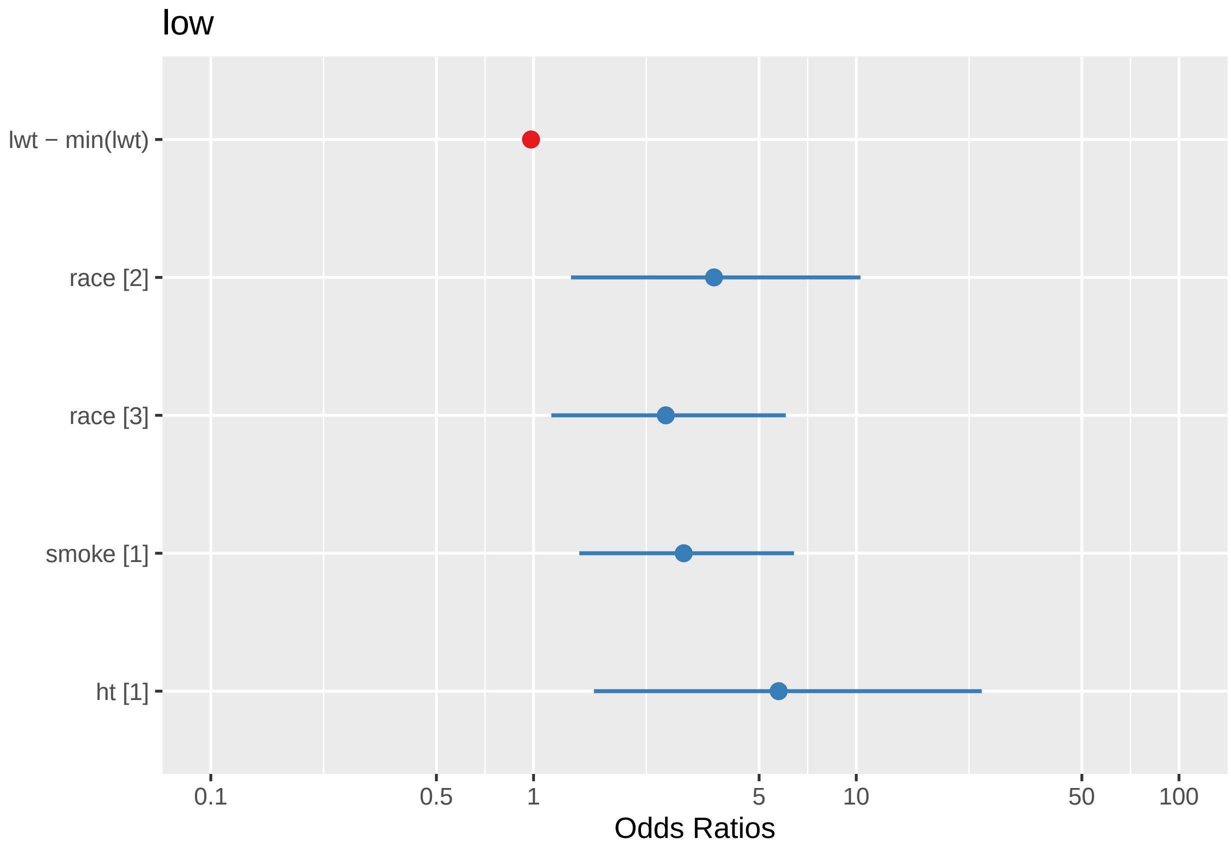
```
# OR
exp(coef(fit_final))[-1]
```

```
## I(lwt - min(lwt))             race2             race3            smoke1
##           0.98225           3.62430           2.56933           2.91995
##               ht1
##           5.74979
```

```
# % change in odds
(exp(coef(fit_final)) - 1)[-1] * 100
```

```
## I(lwt - min(lwt))             race2             race3            smoke1
##           -1.7747          262.4304          156.9329          191.9950
##               ht1
##          474.9786
```

```
# install.packages("sjPlot")
sjPlot::plot_model(fit_final)
```



Poisson Regression

The data we'll use is `epilepsy` from the `HSAUR` package. The dataset is for a randomized clinical trial investigating the effect of an anti-epileptic drug (Progabide).

```
# install.packages("HSAUR")
library(HSAUR)
```

```
## Loading required package: tools
```

```
?epilepsy

data("epilepsy")

dim(epilepsy)
```

```
## [1] 236   6
```

```
head(epilepsy)
```

```
##     treatment base age seizure.rate period subject
## 1     placebo   11  31            5      1       1
## 110   placebo   11  31            3      2       1
## 112   placebo   11  31            3      3       1
## 114   placebo   11  31            3      4       1
## 2     placebo   11  30            3      1       2
## 210   placebo   11  30            5      2       2
```

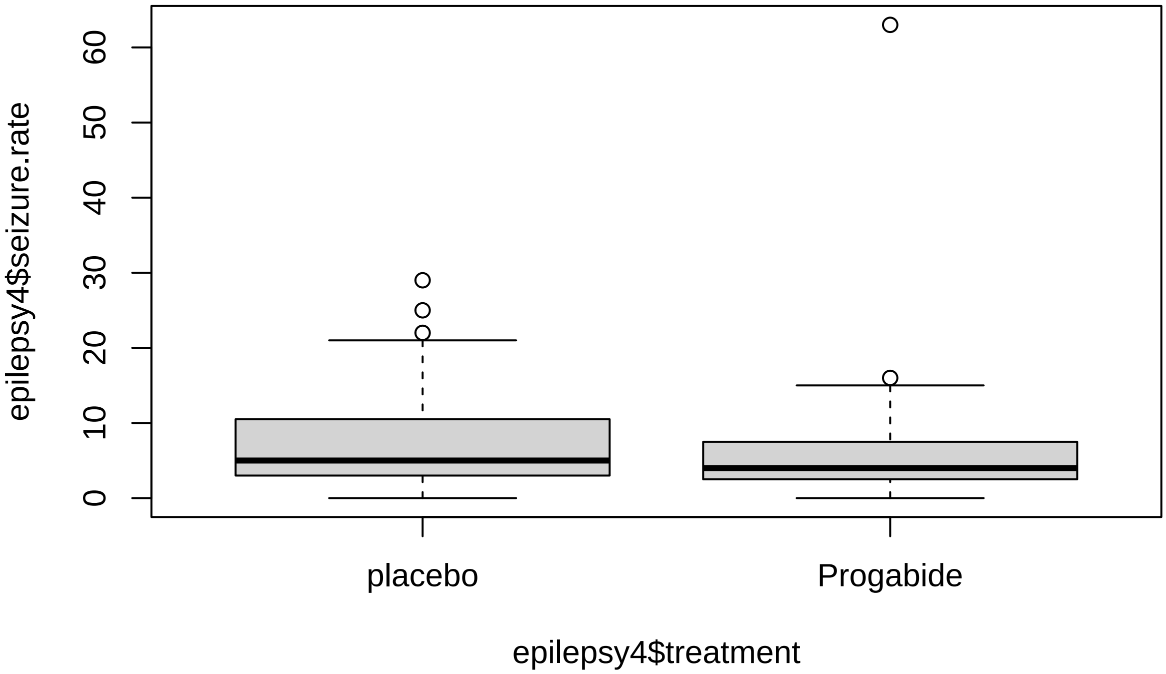
```
summary(epilepsy)
```

```
##      treatment        base             age         seizure.rate     period
##  placebo  :112   Min.   :  6.0   Min.   :18.0   Min.   :  0.00   1:59
##  Progabide:124   1st Qu.: 12.0   1st Qu.:23.0   1st Qu.:  2.75   2:59
##                  Median : 22.0   Median :28.0   Median :  4.00   3:59
##                  Mean   : 31.2   Mean   :28.3   Mean   :  8.26   4:59
##                  3rd Qu.: 41.0   3rd Qu.:32.0   3rd Qu.:  9.00
##                  Max.   :151.0   Max.   :42.0   Max.   :102.00
##
##     subject
##  1      :  4
##  2      :  4
##  3      :  4
##  4      :  4
##  5      :  4
##  6      :  4
##  (Other):212
```

We'll only inspect period 4:

```
epilepsy4 <- epilepsy[epilepsy$period == 4, ]

boxplot(epilepsy4$seizure.rate~epilepsy4$treatment)
```



Let's inspect the effect of treatment adjusting for base and age:

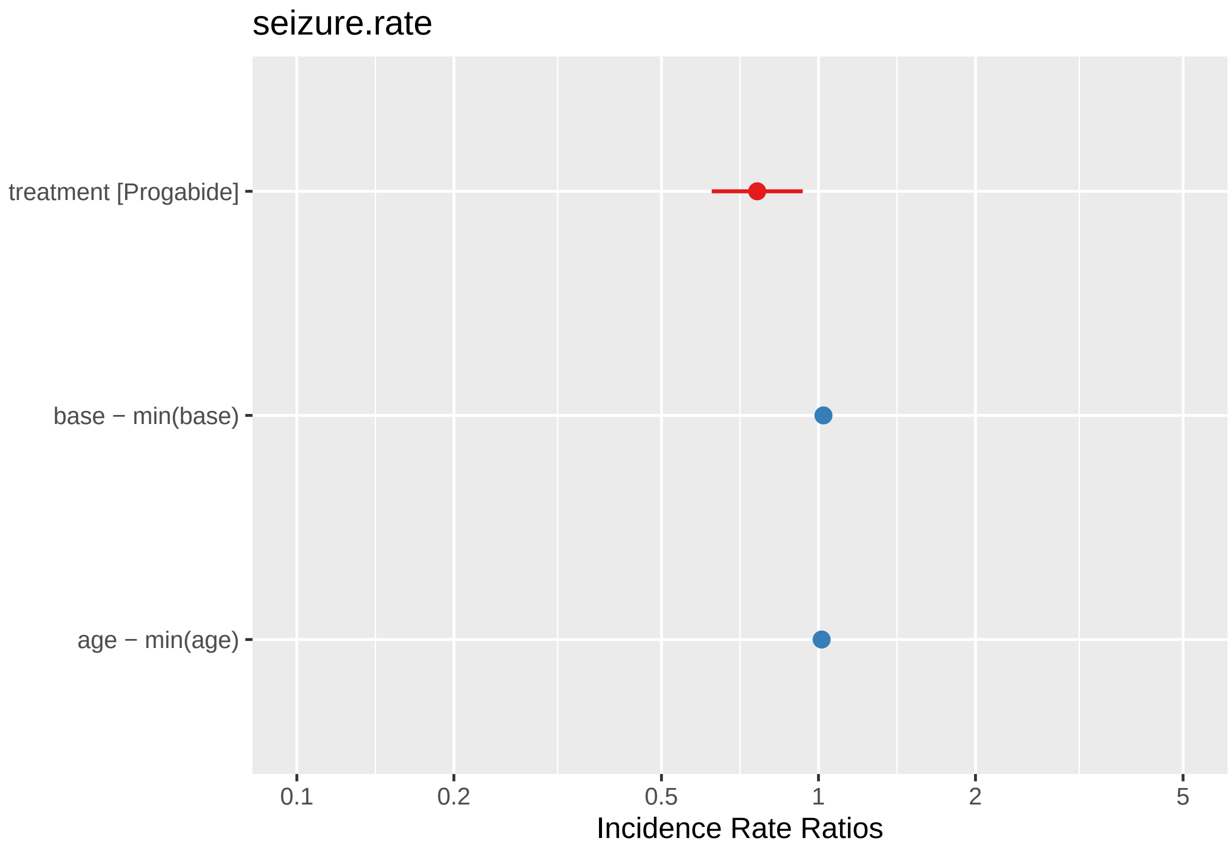
```
fit_pois <- glm(seizure.rate ~ treatment + I(base - min(base)) + I(age - min(age)), data = epilepsy4, f
summary(fit_pois)
```

```
##
## Call:
## glm(formula = seizure.rate ~ treatment + I(base - min(base)) +
##     I(age - min(age)), family = poisson, data = epilepsy4)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.164  -1.025  -0.144   0.487   3.899
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.16070    0.14121    8.22   <2e-16 ***
## treatmentProgabide  -0.27048    0.10187   -2.66   0.0079 **
## I(base - min(base))  0.02206    0.00109   20.27   <2e-16 ***
## I(age - min(age))    0.01404    0.00858    1.64   0.1017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 476.25  on 58  degrees of freedom
## Residual deviance: 147.02  on 55  degrees of freedom
## AIC: 342.8
##
## Number of Fisher Scoring iterations: 5
```

```
(exp(coef(fit_pois)[-1]) - 1) * 100
```

```
##  treatmentProgabide I(base - min(base))   I(age - min(age))
##            -23.6989              2.2302              1.4143
```

```
sjPlot::plot_model(fit_pois)
```


seizure.rate
treatment [Progabide]
base − min(base)
age − min(age)
Incidence Rate Ratios