Logistic Regression

Uohna Thiessen

February 16, 2017

Regression with binary outcomes: Logistic regression

Load dataset into R and generate list of variables

```
NH11 <- readRDS("dataSets/NatHealth2011.rds")
labs <- attributes(NH11)$labels
```

Logistic regression example

Look at the structure of the dataset. How many levels to the outcome variable?

```
str(NH11$hypev) # check stucture of hypev
```

```
##  Factor w/ 5 levels "1 Yes","2 No",..: 2 2 1 2 2 1 2 2 1 2 ...
```

```
levels(NH11$hypev) # check levels of hypev
```

```
## [1] "1 Yes"             "2 No"              "7 Refused"
## [4] "8 Not ascertained" "9 Don't know"
```

Collapse all outcome to binary level variable

```
NH11$hypev <- factor(NH11$hypev, levels=c("2 No", "1 Yes"))
```

Let’s predict the probability of being diagnosed with hypertension based on age, sex, sleep, and bmi.Generate model and print summary

```
hyp.out <- glm(hypev~age_p+sex+sleep+bmi,
              data=NH11, family="binomial")
coef(summary(hyp.out))
```

```
##                 Estimate   Std. Error    z value     Pr(>|z|)
## (Intercept) -4.269466028 0.0564947294 -75.572820 0.000000e+00
## age_p        0.060699303 0.0008227207  73.778743 0.000000e+00
## sex2 Female -0.144025092 0.0267976605  -5.374540 7.677854e-08
## sleep       -0.007035776 0.0016397197  -4.290841 1.779981e-05
## bmi          0.018571704 0.0009510828  19.526906 6.485172e-85
```

Logistic regression coefficients

One solution is to transform the coefficients to make them easier to interpret

```
hyp.out.tab <- coef(summary(hyp.out))
hyp.out.tab[, "Estimate"] <- exp(coef(hyp.out))
hyp.out.tab
```

```
##                Estimate   Std. Error    z value     Pr(>|z|)
## (Intercept) 0.01398925 0.0564947294 -75.572820 0.000000e+00
## age_p       1.06257935 0.0008227207  73.778743 0.000000e+00
## sex2 Female 0.86586602 0.0267976605  -5.374540 7.677854e-08
## sleep       0.99298892 0.0016397197  -4.290841 1.779981e-05
## bmi         1.01874523 0.0009510828  19.526906 6.485172e-85
```

Generating predicted values

"How much more likely is a 63 year old female to have hypertension compared to a 33 year old female?".

Create a dataset with predictors set at desired levels

```
predDat <- with(NH11,
                expand.grid(age_p = c(33, 63), sex = "2 Female",
                bmi = mean(bmi, na.rm = TRUE), sleep = mean(sleep, na.rm = TRUE)))
```

Predict hypertension at those levels

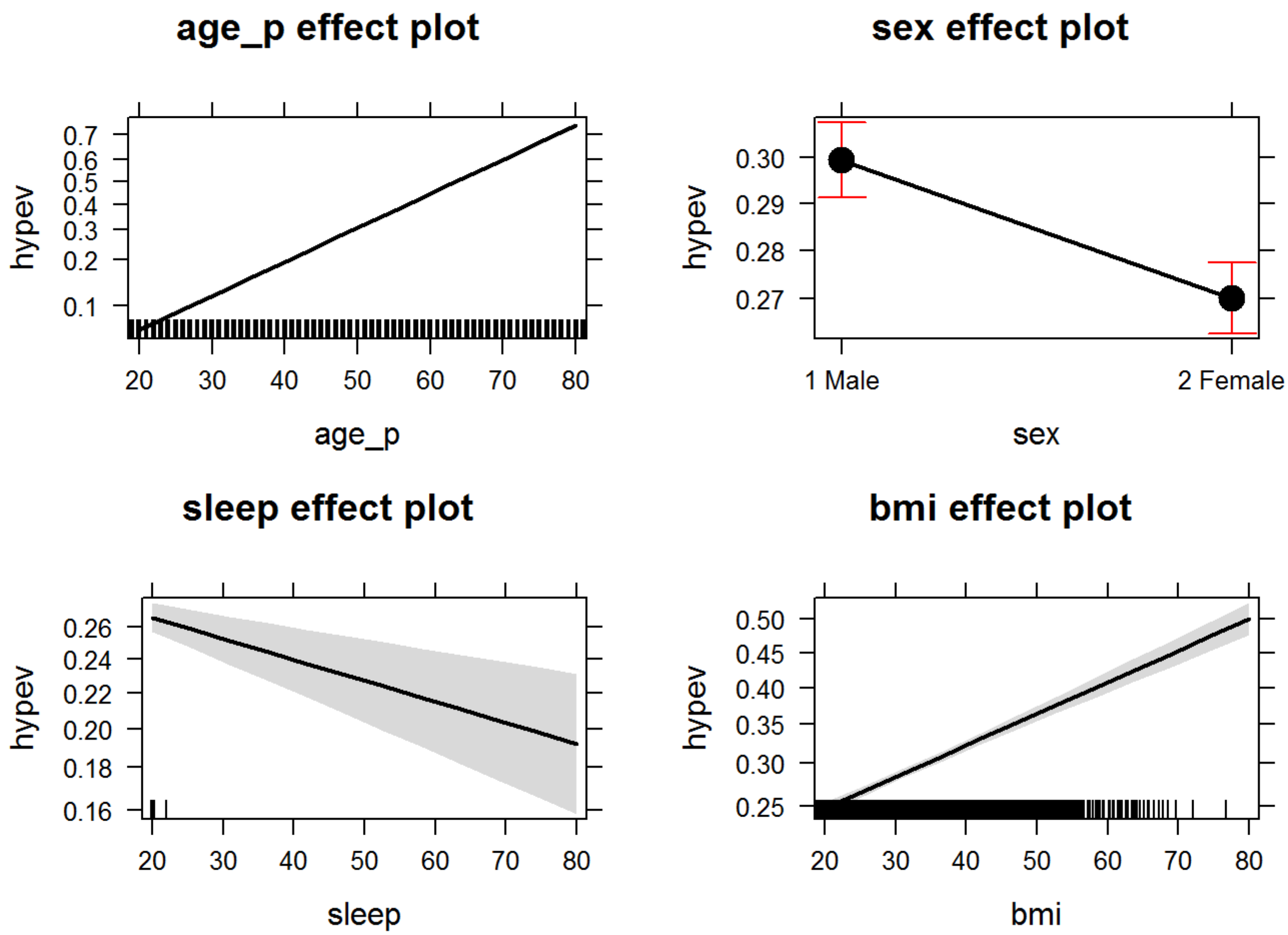
```
cbind(predDat, predict(hyp.out, type = "response",
                       se.fit = TRUE, interval="confidence",
                       newdata = predDat))
```

```
##   age_p      sex      bmi   sleep       fit      se.fit residual.scale
## 1    33 2 Female 29.89565 7.86221 0.1289227 0.002849622              1
## 2    63 2 Female 29.89565 7.86221 0.4776303 0.004816059              1
```

This tells us that a 33 year old female has a 13% probability of having been diagnosed with hypertension, while and 63 year old female has a 48% probability of having been diagnosed.

Packages for computing and graphing predicted values

```
library(effects)
plot(allEffects(hyp.out))
```



Exercise: logistic regression ———————————————————————————

Use the NH11 data set that we loaded earlier.

1. Use glm to conduct a logistic regression to predict ever worked (everwrk) using age (age_p) and marital status (r_maritl).

```
everwrk_mod <- glm(everwrk~age_p+r_maritl,
                   data=NH11, family="binomial")
```

Generate a summary of the model

```
summary(everwrk_mod)
```

```
##
## Call:
## glm(formula = everwrk ~ age_p + r_maritl, family = "binomial",
##     data = NH11)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0429  -0.5675  -0.4452  -0.3384   2.7223
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                 -0.454159   0.093080  -4.879
## age_p                                       -0.029346   0.001633 -17.966
## r_maritl2 Married - spouse not in household  0.081460   0.213836   0.381
## r_maritl4 Widowed                            0.686882   0.083623   8.214
## r_maritl5 Divorced                          -0.732113   0.111145  -6.587
## r_maritl6 Separated                         -0.116447   0.150190  -0.775
## r_maritl7 Never married                      0.355230   0.068865   5.158
## r_maritl8 Living with partner               -0.446226   0.137654  -3.242
## r_maritl9 Unknown marital status             0.541038   0.457838   1.182
##                                             Pr(>|z|)
## (Intercept)                                 1.07e-06 ***
## age_p                                        < 2e-16 ***
## r_maritl2 Married - spouse not in household  0.70324
## r_maritl4 Widowed                            < 2e-16 ***
## r_maritl5 Divorced                          4.49e-11 ***
## r_maritl6 Separated                          0.43814
## r_maritl7 Never married                     2.49e-07 ***
## r_maritl8 Living with partner                0.00119 **
## r_maritl9 Unknown marital status             0.23731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11182  on 14064  degrees of freedom
## Residual deviance: 10411  on 14056  degrees of freedom
##   (18949 observations deleted due to missingness)
## AIC: 10429
##
## Number of Fisher Scoring iterations: 5
```

2. Predict the probability of working for each level of marital status. Transform the coefficients from log odds to odds

```
everwrk_mod.tab <- coef(summary(everwrk_mod))
everwrk_mod.tab[, "Estimate"] <- exp(coef(everwrk_mod))
everwrk_mod.tab
```

```
##                                             Estimate  Std. Error
## (Intercept)                               0.6349819 0.093080415
## age_p                                     0.9710807 0.001633363
## r_maritl2 Married - spouse not in household 1.0848694 0.213835768
## r_maritl4 Widowed                         1.9875095 0.083623142
## r_maritl5 Divorced                        0.4808920 0.111144918
## r_maritl6 Separated                       0.8900773 0.150189947
## r_maritl7 Never married                   1.4265083 0.068864919
## r_maritl8 Living with partner             0.6400391 0.137653720
## r_maritl9 Unknown marital status          1.7177898 0.457837543
##                                               z value     Pr(>|z|)
## (Intercept)                                -4.8792091 1.065121e-06
## age_p                                     -17.9664355 3.569242e-72
## r_maritl2 Married - spouse not in household  0.3809446 7.032444e-01
## r_maritl4 Widowed                           8.2140223 2.138998e-16
## r_maritl5 Divorced                         -6.5870087 4.487759e-11
## r_maritl6 Separated                        -0.7753316 4.381438e-01
## r_maritl7 Never married                     5.1583553 2.491285e-07
## r_maritl8 Living with partner              -3.2416562 1.188373e-03
## r_maritl9 Unknown marital status            1.1817259 2.373145e-01
```