Logistic Regression Mini-Project

Carl Larson

2/2/2018

To load the data of interest, the National Health Interview Survey:

```
NH11 <- readRDS("/Users/EagleFace/Documents/!logistic_regression/dataSets/NatHealth2011.rds")

labs <- attributes(NH11)$labels
```

First to use logistic regression to look at the relationship of hypertension as a function of age, sex, sleep, and BMI.

```
str(NH11$hypev) # checking structure
```

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

```
levels(NH11$hypev) # checking levels
```

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

```
# Swaps in NA in for any missing value
NH11$hypev <- factor(NH11$hypev, levels=c("2 No", "1 Yes"))

#Running the regression
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
```

These are coming out in log odds, so a transformation is needed to make these results more readily interprable.

```
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
```

This is showing high z values meaning we can reject the null hypothesis that these variables have nothing to do with the predicted variable, hypertension. It's showing that the strongest correlation of risk of hypertension is with BMI, and age is nearly as good a predictor. Sleep was also a good predictor, and sex was the worst predictor.

Next we are asking the question:

How much more likely is a 63 year old woman to have hypertension than a woman of age 33

```r
#Creating a bespoke dataset to use for this
predDat <- with(NH11,
                expand.grid(age_p = c(33, 63),
                            sex = "2 Female",
                            bmi = mean(sleep, na.rm = TRUE),
                            sleep = mean(sleep, na.rm = TRUE)))

#Predicting hypertension at these age 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 7.86221 7.86221 0.08950318 0.002858571              1
## 2    63 2 Female 7.86221 7.86221 0.37783748 0.006864154              1
```

They are saying it should show 13% probability for 33 yrs & 48% for 63 yrs. Noting the code doesn't really query that data, it seems to try to predict hypertension off sleep and bmi, which may be interesting, but the problem was literally asking something the difference in likelihood between getting hypertension between 33 year olds and 63 year olds.

Sleep and hypertension look like bad predictors of hypertension for 33 year olds, but decent predictors for 63 year olds. This makes sense too, and something that nuanced would need longitudinal backup to see if 33 year olds with worse sleep and bmi were more likely to develop hypertension by the time they were 63.

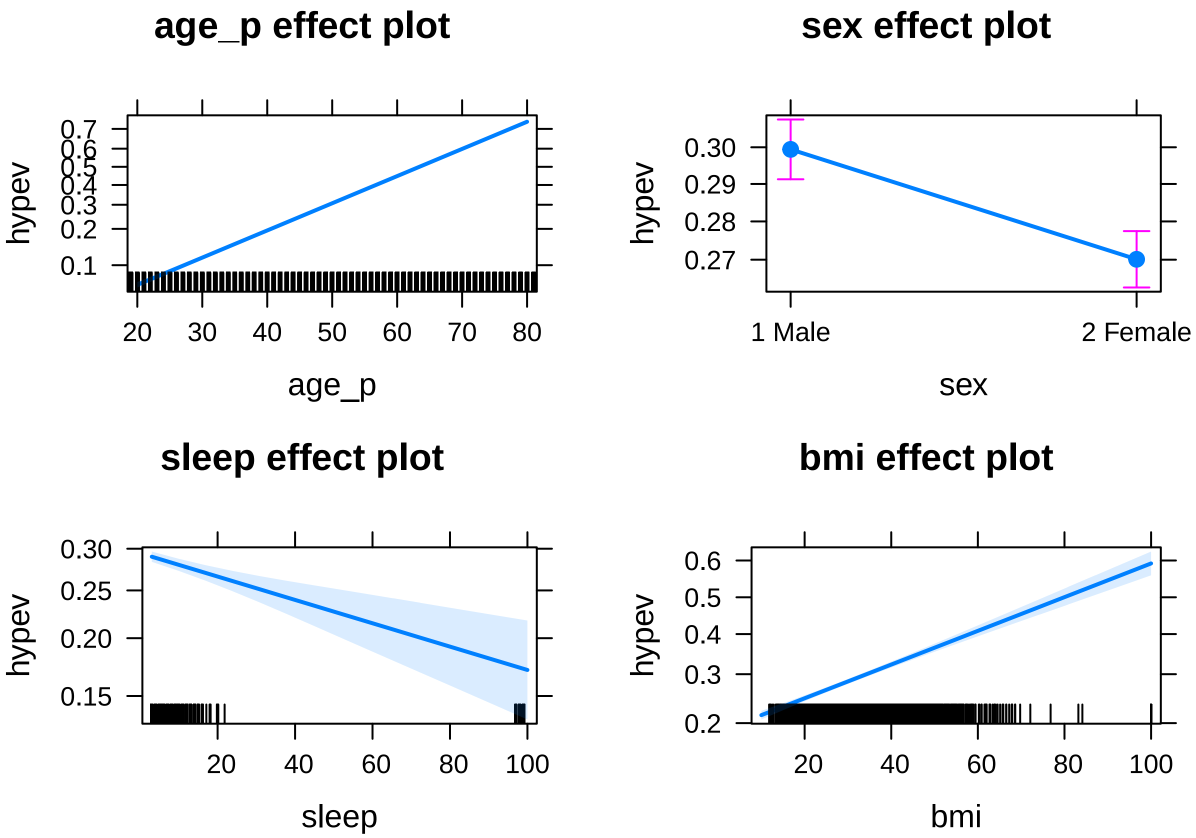
It seems this question could have just taken the proportion of each group that had hypertension and call that the expected value, but trying to "predict" anything off this snapshot would be premature.

```r
library(effects)
```

```
## Loading required package: carData
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
plot(allEffects(hyp.out))
```



Next we are asked to conduct a logistical regression on “ever worked” predicted by age and marital status.

1. Use glm to conduct a logistic regression to predict ever worked using age and marital status

Just intuitively, it seems there should be a loose correlation for marital status with married people being more likely to have worked ever than non-married people, just because those groups also skew by age, which would be a hidden controlling variable.

```
#First to look at the structure of our variables.

str(NH11$r_maritl)
```

```
##  Factor w/ 10 levels "0 Under 14 years",..: 6 8 5 7 2 2 8 8 8 2 ...
```

```
levels(NH11$r_maritl)
```

```
##  [1] "0 Under 14 years"
##  [2] "1 Married - spouse in household"
##  [3] "2 Married - spouse not in household"
##  [4] "3 Married - spouse in household unknown"
##  [5] "4 Widowed"
##  [6] "5 Divorced"
##  [7] "6 Separated"
##  [8] "7 Never married"
##  [9] "8 Living with partner"
## [10] "9 Unknown marital status"
```

```r
length(NH11$r_maritl)
```

```
## [1] 33014
```

```r
str(NH11$age_p)
```

```
##  num [1:33014] 47 18 79 51 43 41 21 20 33 56 ...
```

```r
length(NH11$age_p)
```

```
## [1] 33014
```

```r
str(NH11$everwrk)
```

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

```r
levels(NH11$everwrk)
```

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

```r
length(NH11$everwrk)
```

```
## [1] 33014
```

We can see there may be a significant amount of "NA" values for the output variable, ever worked. There are 5 different possible factor outputs for that variable, and it seems likely that they will all correlate heavily with age, and each other. I am intuitively thinking that the older people who have worked in higher proportions will have answered "1 Yes" the most. I still want to combine the non-yes responses to the everworked variable.

Now to try to make the model.

```r
unworked.pred <- glm(everwrk ~ age_p + r_maritl,
                     data=NH11, family="binomial")
coef(summary(unworked.pred))
```

```
##                                                Estimate  Std. Error
## (Intercept)                                 -0.45415880 0.093080415
## age_p                                       -0.02934571 0.001633363
## r_maritl2 Married - spouse not in household  0.08145957 0.213835768
## r_maritl4 Widowed                            0.68688235 0.083623142
## r_maritl5 Divorced                          -0.73211254 0.111144918
## r_maritl6 Separated                         -0.11644701 0.150189947
## r_maritl7 Never married                      0.35522972 0.068864919
## r_maritl8 Living with partner               -0.44622604 0.137653720
## r_maritl9 Unknown marital status             0.54103849 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
```

2. Predict the probability of working for each level of marital status

Based on these results it seems, sadly, that widows are the most likely marital staus group to have worked already, and never married are the most likely to never have worked.

This does make some morbid sense, as widows are usually older and people who have never worked also are more likely to never have married and just be a young person.

Both of those values had low Pr(>|z|) values, meaning a low likelihood of the null hypothesis being true, meaning we are very sure that those variables essentially aren't wrong on predicting "ever worked" status.