BE7023 Homework 5

Mike Lape

Octobe 12, 2018

```r
#setwd("C:/Users/lapt3u/Box/UC/Fall_2018/BE7023_Adv_Biostats/adv_biostats/hw_5")
library(ggplot2)
# Build DF
no.yes <- c("No","Yes")
Smoking <- gl(2,1,8,no.yes)
Obesity <- gl(2,2,8,no.yes)
Snoring <- gl(2,4,8,no.yes)
Total <- c(60,17,8,2,187,85,51,23)
Hypertension <- c(5,2,1,0,35,13,15,8)
dat <- data.frame(Smoking,Obesity,Snoring,Total,Hypertension)
```

1. Postulate the logisitc regression model.

Z = β_0 + β_1 * Smoking(Yes) + β_2 * Snoring(Yes) + β_3 * Obesity(Yes)

Pr(Hypertension = Yes) = $\frac{e^Z}{(1+e^Z)}$

2. Fit the model to the data.
   Exhibit the output.
   Write the prediction model.

```r
# Fit the model
mod <- glm(cbind(Hypertension, Total-Hypertension) ~ Smoking + Snoring + Obesity, data = dat, family =
summary(mod)
```

```
##
## Call:
## glm(formula = cbind(Hypertension, Total - Hypertension) ~ Smoking +
##     Snoring + Obesity, family = binomial, data = dat)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
## -0.04344   0.54145  -0.25476  -0.80051   0.19759  -0.46602  -0.21262
##        8
##  0.56231
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.37766    0.38018  -6.254    4e-10 ***
## SmokingYes  -0.06777    0.27812  -0.244   0.8075
## SnoringYes   0.87194    0.39757   2.193   0.0283 *
## ObesityYes   0.69531    0.28509   2.439   0.0147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14.1259  on 7  degrees of freedom
## Residual deviance:  1.6184  on 4  degrees of freedom
## AIC: 34.537
##
## Number of Fisher Scoring iterations: 4
```

Using the fit model, the prediction equation is below.

Z = -2.378 - (0.068 * Smoking(Yes)) + (0.872 * Snoring(Yes)) + (0.695 * Obesity(Yes))

Pr(Hypertension = Yes) = $\frac{e^Z}{(1+e^Z)}$

3. Check the adequacy of the model.

```
# We can check the adequacy of the model by doing a simple hypothesis test.
# Our null hypothesis is that the response probability (Hypertension) follows the
# logistic model laid out above in question 2.  The alternate hypothesis is simply
# that the response probability does not follow the logistic model laid out.
# Test this: If null hypothesis is true then the residual deviance has a chi-squared
# distribution with the stated degrees of freedom.
# Let's pull out residual deviance and degrees of freedom

# Residual deviance:
dev <- mod$deviance
paste("Residual Deviance: ",round(dev, 3))
```

```
## [1] "Residual Deviance:  1.618"
```

```
# Degrees of freedom
dof <- mod$df.residual
paste("Degrees of Freedom: ", dof)
```

```
## [1] "Degrees of Freedom:  4"
```

```
# Now calculate p-value under null hypothesis
p <- pchisq(dev, dof, lower.tail = F)
paste("P-value: ", round(p, 3))
```

```
## [1] "P-value:  0.805"
```

```
# Because this p-value is much larger than 0.05 we accept the null hypothesis
# that the reposonse probability (hypertension) follows the logistic model
# laid out in question 2, and thus the logistic model can adequatly descibe
# our data.
```

4. Check the significance of the covariates.

```
summary(mod)
```

```
##
## Call:
## glm(formula = cbind(Hypertension, Total - Hypertension) ~ Smoking +
##     Snoring + Obesity, family = binomial, data = dat)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7
## -0.04344   0.54145  -0.25476  -0.80051   0.19759  -0.46602  -0.21262
##        8
##  0.56231
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.37766    0.38018  -6.254    4e-10 ***
## SmokingYes  -0.06777    0.27812  -0.244   0.8075
## SnoringYes   0.87194    0.39757   2.193   0.0283 *
## ObesityYes   0.69531    0.28509   2.439   0.0147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14.1259  on 7  degrees of freedom
## Residual deviance:  1.6184  on 4  degrees of freedom
## AIC: 34.537
##
## Number of Fisher Scoring iterations: 4
```

We can see that only 2 of the 3 predictors are signficant, snoring and obesity.
Snoring(Yes) [p-val: 0.0283]
Obesity(Yes) [p-val: 0.0147]

5. Obtain the odds ratios for: Hypertension vs Smoking Hypertension vs Obesity Hypertension vs Snoring.

```
# To calculate the odds ratio for each covariate we just take the
# calculated coefficient and reverse the natural log built into
# the model by raising e to the power of the coefficient for
# that covariate.
or <- exp(mod$coefficients)
paste("Odds ratio of Hypertension when...")
```

```
## [1] "Odds ratio of Hypertension when..."
```

```
paste("Smoking = Yes: ", round(or["SmokingYes"],3))
```

```
## [1] "Smoking = Yes:  0.934"
```

```
paste("Obesity = Yes: ", round(or["ObesityYes"],3))
```

```
## [1] "Obesity = Yes:  2.004"
```

```
paste("Snoring = Yes: ", round(or["SnoringYes"],3))
```

```
## [1] "Snoring = Yes:  2.392"
```

6. Obtain predicted probabilities as per the fitted model.

```
# We will calculate predicted probabilities for the data we used to fit the
# model, which is less than ideal, but its the only data we have.
pred <- predict(mod, newdata = dat, type = "response")
# Predicted probabilities for training data:
round(pred,3)
```

```
##     1     2     3     4     5     6     7     8
## 0.085 0.080 0.157 0.148 0.182 0.172 0.308 0.294
```

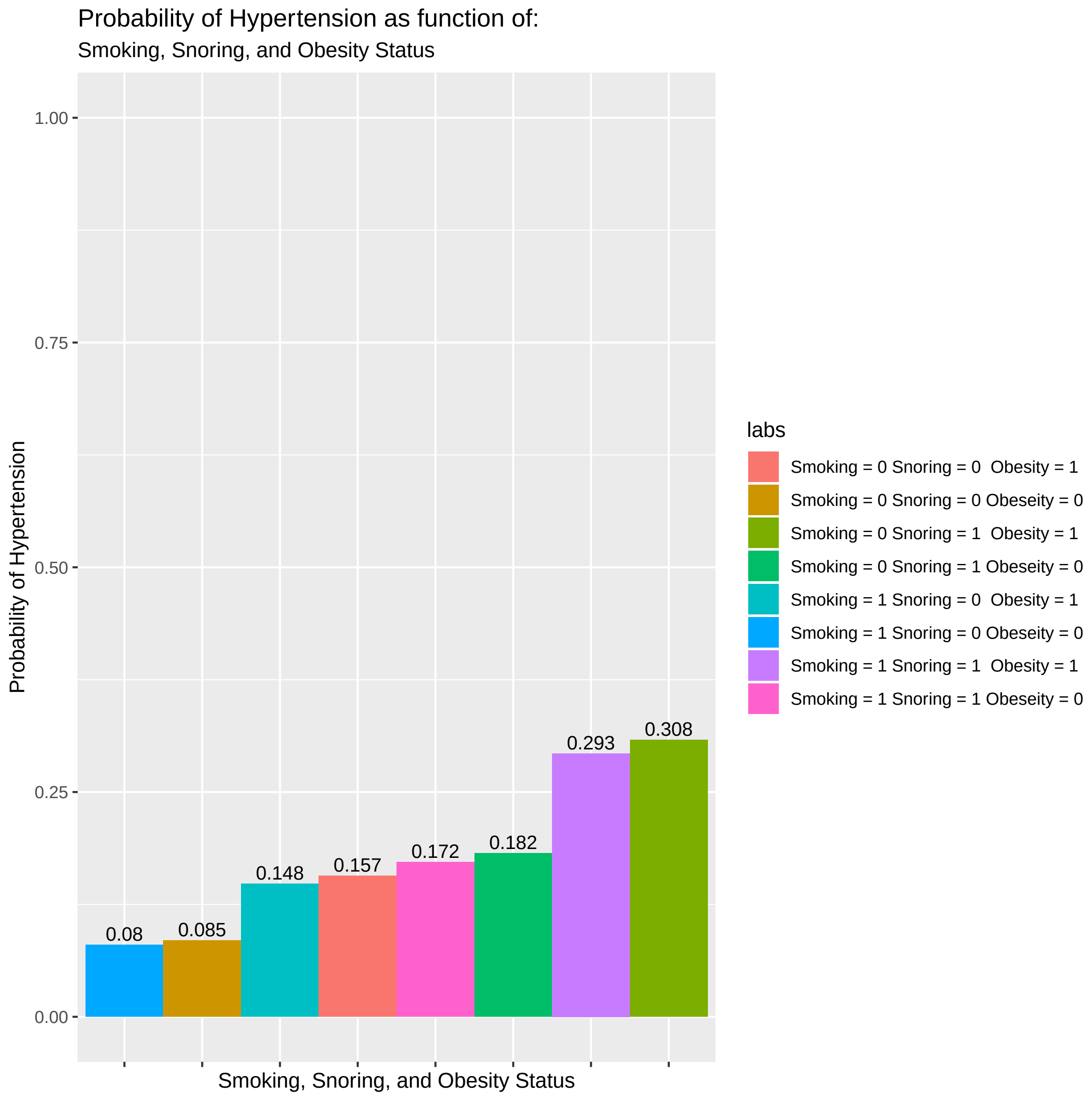
```
# We can also classify each by using 0.5 as a cutoff, where hypertension = Yes
# if the predicted probability is greater than or equal to 0.5, otherwise
# hypertension = No
pred_class <- ifelse(pred >= 0.50, 1, 0)
pred_class
```

```
## 1 2 3 4 5 6 7 8
## 0 0 0 0 0 0 0 0
```

```
# We can see that using the 0.5 cutoff none of patients are predicted to have
# hypertension, but we know in almost all groups some of the patients have it.
# We would probably need to move the cutoff point to fix this, which I won't
# do here.
```

7. Plot the Probability(HypertensionYes) for different scenarios of Smoking, Obesity, and Snoring. Comment on the graph

```
apply_mod <- function(smoke, snore, obese)
{
  ret <- (exp(-2.378 - (0.068 * smoke) + (0.872 * snore) + (0.695 * obese)) /
        (1 + exp(-2.378 - (0.068 * smoke) + (0.872 * snore) + (0.695 * obese))))
  return(ret)
}

bar_1 <- round(apply_mod(0,0,0),3)
bar_2 <- round(apply_mod(1,0,0),3)
bar_3 <- round(apply_mod(0,1,0),3)
bar_4 <- round(apply_mod(1,1,0),3)
bar_5 <- round(apply_mod(0,0,1),3)
bar_6 <- round(apply_mod(1,0,1),3)
bar_7 <- round(apply_mod(0,1,1),3)
bar_8 <- round(apply_mod(1,1,1),3)

vals <- c(bar_1, bar_2,  bar_3,  bar_4,  bar_5,  bar_6, bar_7, bar_8 )

labs <- c(
        "Smoking = 0 Snoring = 0 Obeseity = 0",
        "Smoking = 1 Snoring = 0 Obeseity = 0",
        "Smoking = 0 Snoring = 1 Obeseity = 0",
        "Smoking = 1 Snoring = 1 Obeseity = 0",
        "Smoking = 0 Snoring = 0  Obesity = 1",
        "Smoking = 1 Snoring = 0  Obesity = 1",
        "Smoking = 0 Snoring = 1  Obesity = 1",
        "Smoking = 1 Snoring = 1  Obesity = 1"
        )

df <- data.frame(labs = labs, prob = vals)
df <- df[order(df$prob),]
pl <- ggplot(data = df, aes(x = reorder(labs,prob), y = prob, fill = labs)) +
          geom_bar(stat="identity", width = 1) +
          geom_text(aes(label=prob), vjust=-0.3, size=3.5) +
          ylim(0,1)+ labs(y = "Probability of Hypertension",
          x = "Smoking, Snoring, and Obesity Status", title =
            "Probability of Hypertension as function of:",
            subtitle = "Smoking, Snoring, and Obesity Status") +
          theme(axis.text.x=element_blank())

pl
```



We can easily see from the plot that when someone both snores and is obese they have the greatest probability of developing hypertension. Oddly, the probability of developing hypertension when you are obese, snore, and smoke is lower, indicating that smoking possibly is protective if you are already obese and snore. We can also see that the probability drops off a bit after these two situations, indicating that these 2 are at the highest proability of having hypertension. Indeed, it appears and this should have been obvious from the negative coefficient for smoking but smoking is protective for hypertension.