Unit 5 - Logistic Regression


Exercises 1-3 utilize a data set provided by Afifi, Clark and May (2004). The data are a study of depression and was a longitudinal study. The purpose of the study was to obtain estimates of the prevalence and incidence of depression and to explore its risk factors. The study variables were of several types - demographics, life events, stressors, physical health, health services utilization, medication use, lifestyle, and social support.

Depressed Data set  Please download the depress_small.dta data set from online and set it in your working directory. If you recall, there are two ways to upload this data set into R. You can point and click by finding the Import Dataset command or you can do it via the command line, which we will do here.

```r
library(foreign)

# I saved the original stata data into stata 12 version.

# head(dat)
```

Consider the following three variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Codings</th>
<th>Label in R</th>
</tr>
</thead>
<tbody>
<tr>
<td>drink</td>
<td>1 = yes 2 = no</td>
<td>Regular Drinker</td>
</tr>
<tr>
<td>sex</td>
<td>1 = male 2 = female</td>
<td></td>
</tr>
<tr>
<td>cases</td>
<td>0 = normal 1 = Depressed</td>
<td>Depressed is cesd &gt;= 16</td>
</tr>
</tbody>
</table>


Using R, load the depression data set and execute the commands needed to create a 2X2 table featuring the variables drink and sex. [Hint: see section 2a from R for Categorical Data Analysis, the descr library].

```r
library(descr)
crosstab(dat$drink, dat$sex, plot = FALSE)
```

## Cell Contents
## |-------------------------|
## | Count |
What are the odds that a woman is a regular drinker? \( \frac{139}{44} = 3.2 \)

What are the odds that a man is a regular drinker? \( \frac{95}{16} = 5.9 \)

What is the odds ratio? That is, compared to a man, what is the relative odds (odds ratio) that a woman is a regular drinker? \( \text{OR} = \frac{\text{odds for woman}}{\text{odds for man}} = \frac{3.2}{5.9} = 0.54 \)

2. Repeat the tabulation that produced for problem #1 two times, one for persons who are depressed and the other for persons who are not depressed.

To visualize a table with factored levels, such drink by sex BY depressed, we do the following. There are other ways, but this is the most simple way.

```r
with(dat, table(drink, sex, depressed))
```

```r
# , , depressed = 0
#
#   sex
#   # drink 1. male 2. female
#   # 1. yes 87 106
#   # 2. no 14 37
# #
# # , , depressed = 1
# #
#   sex
#   # drink 1. male 2. female
#   # 1. yes 8 33
#   # 2. no 2 7
```

From here, you may compute the following odds ratios:

Among persons who are NOT depressed:

- OR (Relative odds, compared to a man, that a woman is a regular drinker): \( \frac{106*14}{37*87} = 0.46 \)

Among persons who are depressed:

- OR (Relative odds, compared to a man, that a woman is a regular drinker): \( \frac{33*2}{7*8} = 1.18 \)
3. Fit a logistic regression model using these variables. Use drink as the dependent variables and cases and sex as independent variables. Also, include as an independent variable the appropriate interaction term.

Before we dive into the model, let’s rearrange the factored levels in our variables. It is important to note that the baseline is 0 while the...non-baseline is 1. It is also important to note that continuous variables are continuous and categorical variables are factored into R. Please follow along below and if you get lost, look at it again or simply email me.

```r
class(dat$drink)  # our dep. variable is a factor, which is what we want

## [1] "factor"

levels(dat$drink)  # but the baseline is YES when it should be no. Let's fix that.

## [1] "1. yes" "2. no"

dat$drink1 <- ifelse(dat$drink == "2. no", as.numeric(dat$drink) - 2, dat$drink)
levels(dat$drink1) <- c("0. no", "1. yes")  # fixed

class(dat$female)  # we see that it is numeric and we must change it to a factor

## [1] "numeric"

dat$female1 <- factor(dat$female)

class(dat$cases)  # factor - good

## [1] "factor"

levels(dat$cases)  # the levels are set

## [1] "0. normal" "1. depressed"

That was a bit of an esoteric exercise in examining the structure of our data. My apologies as I would not want to include that in this course but it is good to get some experience with it. Moving on, let’s create our model with our newly formatted variables.

```r
logmodel <- glm(drink1 ~ cases + female1 + cases*female1, family = binomial(link = "logit"), data = dat)
summary(logmodel)
```

## Call:
## glm(formula = drink1 ~ cases + female1 + cases * female1, family = binomial(link = "logit"), data = dat)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max

## Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 1.8269 | 0.2880 | 6.344 | 2.24e-10 *** |
| cases1. depressed | -0.4406 | 0.8414 | -0.524 | 0.601 |
| female11 | -0.7743 | 0.3455 | -2.241 | 0.025 * |
| cases1. depressed:female11 | 0.9386 | 0.9579 | 0.980 | 0.327 |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 297.53 on 293 degrees of freedom
Residual deviance: 291.92 on 290 degrees of freedom
AIC: 299.92

Number of Fisher Scoring iterations: 4

Please write out the full model:

\[ \text{logit}[pr(\text{drinker} = \text{yes})] = \beta_0 + \beta_1[\text{cases}] + \beta_2[\text{female}] + \beta_3[\text{cases} \times \text{female}] \]

where:

- \text{cases} = 1 if depressed; 0 otherwise
- \text{female} = 1 if female; 0 otherwise
- \text{cases} \times \text{female} = as is

and

- \beta_0 = 1.8269 (also known as the intercept)
- \beta_1 = -0.4406
- \beta_2 = -0.7743
- \beta_3 = 0.9386

Is the interaction term in your model significant? No. We see that the p-value is \( \approx 0.33 \).

How does your answer to problem #3 compare to your answer in problem #2? Please comment.

The answers match.

Among depressed: \textbf{OR = 1.18} Among non-depressed: \textbf{OR = 0.46}

\[ \text{logit}[pr(\text{drinker} = \text{yes})] = \beta_0 + \beta_1[\text{cases}] + \beta_2[\text{female}] + \beta_3[\text{cases} \times \text{female}] \]

Furthermore, let’s interpret the model amongst those who were depressed and those who were not depressed.

\text{logit[female]} = 1.8269 - 0.4406 - 0.7743 + 0.9386 = \textbf{1.5506}

\text{logit[male]} = 1.8269 - 0.4406 = \textbf{1.3863}

\text{logit[female]} - \text{logit[male]} = 1.5506 - 1.3863 = \textbf{0.1643}

\text{OR[women compared to men]} = \exp(\text{logit[p1]} - \text{logit[p0]}) = \exp(0.1643) = \textbf{1.1786}
logit[female] = 1.8269 - 0.7743 = \textbf{1.0526}

logit[male] = \textbf{1.8269}

logit[female] - logit[male] = 1.0526 - 1.8269 = \textbf{-0.7743}

OR[women compared to men] = \exp\{\text{logit}[p1] - \text{logit}[p0]\} = \exp(-0.7743) = \textbf{0.4610}

Q4: Please refer to the other version of HW solution.