Logistic Regression

Samantha Estrada PhD
ORSSP Research Design & Data Analysis Lab Consultant
Assistant Professor of Psychology
Dependent Variable

Independent Variable

X1 Horn length
X2 Mane color
X3 Coat Color
X4 Speed
Applications of Logistic Regression

• Retention studies
  – i.e., want to examine factors which predict whether college students will or will not stay in school

• Marriage/family studies
  – e.g., might look at variables which predict which couples will or will not divorce or factors which predict

• Medical research
  – Factors distinguishing between those who will and will not survive (e.g., surgery, a particular illness, etc.)
Logistic Regression

- Since logistic regression is nonparametric, you have more flexibility with variables because there are no normality assumptions.
- The outcome variable is categorical. The predictor variables can be a mix of categorical or continuous variables.
- Logistic regression is all about predicting the odds that a given outcome will occur.
  - Odds are different than probabilities.
  - Probabilities range from 0-1
  - Odds can range from negative infinity to positive infinity.
  - Positive odds means a thing is more likely to occur, and negative odds mean a thing is less likely to occur.
Brief Probability Review

• Probabilities are simply the likelihood that something will happen; a probability of .20 of rain means that there is a 20% chance of rain.

• If there is a 20% chance of rain, then there is an 80% chance of no rain; the odds, then, are:

\[ \text{Odds} = \frac{\text{prob(rain)}}{\text{prob(norain)}} = \frac{20}{80} = \frac{1}{4} = .25 \]

• Remember that probability can range from 0 to 1. But the odds can be greater than 1.
  – For instance, a 50% chance of rain has odds of 1.
Odds Ratio

• Odds ratio (OR) is the effect size for logistic regression
• Odds ratios greater than 1 = increase of the odds of that outcome
• Odds ratios less than 1 = decrease in the odds of that outcome.
• The comparison group is the group coded as 0.
  – So if your odds ratio is greater than 1, you have an increase in the odds of being in the 1 group.
  – Less than 1 decrease in odds of the 1 group (or increase in the 0 group).
Sample Size Requirements

• In terms of the adequacy of sample sizes, the literature has not offered specific rules applicable to logistic regression (Peng et al., 2002).

• Several authors on multivariate statistics (Tabachnick & Fidell, 2019) have recommended:
  – A minimum ratio of 10 (observations) to 1 (variable), with a minimum sample size of 100 or 50
Example: Logistic Regression

Data
The dataset for this example contains N = 275 observations and seven variables. In the following example we would like to predict heart attacks in males from the following data:

- **Nominal DV**: Heart Attack where 0=no heart attack and 1=heart attack.
- **Continuous IV**: AGE in years
- **Continuous IV**: Systolic blood pressure (SYSBP)
- **Continuous IV**: Diastolic blood pressure (DIABP)
- **Continuous IV**: Cholesterol (CHOLES)
- **Continuous IV**: Height (HT) height in inches
- **Continuous IV**: Weight (WT) weight in pounds

Research Question
Do body weight, height, blood pressure and age have an influence on the probability of having a heart attack (yes vs. no)?
Logistic Regression in R

# Installing the package for logistic regression
install.packages("caTools")
# Loading the packages
library(caTools)
library(haven)  # I use this package to import SPSS files

Next we import the file this can be done manually via the point and click option or via code.

# Loading the file
library(haven)
logistic.dat <-
as.data.frame(read_sav("~/Library/CloudStorage/OneDrive-TheUniversityofTexasatTyler/Teaching GD/PSYC 5340/PPT/8.5 Logistic Regression/logisitic.sav"))
If using R studio
## Descriptive info

```r
summary(logistic.dat)
```

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sysbp</th>
<th>diabp</th>
<th>choles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>23.00</td>
<td>90.00</td>
<td>55.00</td>
<td>135.0</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>36.00</td>
<td>110.0</td>
<td>75.50</td>
<td>254.0</td>
</tr>
<tr>
<td>Median</td>
<td>45.00</td>
<td>120.0</td>
<td>80.00</td>
<td>285.0</td>
</tr>
<tr>
<td>Mean</td>
<td>45.03</td>
<td>124.2</td>
<td>82.97</td>
<td>297.3</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>52.00</td>
<td>130.0</td>
<td>90.00</td>
<td>336.5</td>
</tr>
<tr>
<td>Max.</td>
<td>70.00</td>
<td>190.0</td>
<td>112.0</td>
<td>520.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ht</th>
<th>wt</th>
<th>coron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>62.00</td>
<td>108.0</td>
<td>0.0000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>67.00</td>
<td>150.0</td>
<td>0.0000</td>
</tr>
<tr>
<td>Median</td>
<td>68.00</td>
<td>166.0</td>
<td>0.0000</td>
</tr>
<tr>
<td>Mean</td>
<td>68.45</td>
<td>167.7</td>
<td>0.3636</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>70.00</td>
<td>181.0</td>
<td>1.0000</td>
</tr>
<tr>
<td>Max.</td>
<td>74.00</td>
<td>262.0</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
# Frequency of the Dependent Variable

```r
library(tidyverse)
library(formattable)

logistic.dat %>%
group_by(coron) %>%
summarize(Freq=n()) %>%
mutate(freq = percent(Freq / sum(Freq))) %>%
arrange(desc(Freq))
```

## A tibble: 2 x 3

<table>
<thead>
<tr>
<th>coron</th>
<th>Freq</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;dbl&gt;</td>
<td>&lt;int&gt;</td>
<td>&lt;formttbl&gt;</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>175 63.64%</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>100 36.36%</td>
</tr>
</tbody>
</table>

63.6% of the patients have not had a heart attack, and 36.4% of the patients have had one.
Collinearity

- We don’t want to have variables that explain the same thing in our regression, or that are too highly correlated.
- Logistic regression does not have to meet the assumptions of normality or heterogeneity of variance, but we do have to check for multicollinearity.
• We will do Simple Linear Regression to find the multicollinearity indicators

```
# Simple Linear Regression
model = lm(corn ~ age + sysbp + diabp + choles + ht + wt, data = logistic.dat)
```

\[ y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c \]
# Collinearity Diagnostics

```r
install.packages("olsrr")
library(olsrr)

ols_vif_tol(model)
```

```
## Variables Tolerance      VIF
## 1 age   0.6363933 1.571355
## 2 sysbp 0.2798345 3.573540
## 3 diabp 0.2694661 3.711041
## 4 choles 0.8096193 1.235148
## 5 ht     0.7425696 1.346675
## 6 wt     0.7023589 1.423774
```

Everything looks good according to our rules of thumb VIF < 10 and Tolerance > .01
Code: Logistic Regression

```r
logistic_model = glm(coron ~ age + sysbp + diabp + choles + ht + wt,
    data = logistic.dat,
    family = "binomial")

y = b_1x_1 + b_2x_2 + ... + b_nx_n + c

# Summary
summary(logistic_model)
```
## Call:
```
glm(formula = coron ~ age + sysbp + diabp + choles + ht + wt,
    family = "binomial", data = logistic.dat)
```

## Deviance Residuals:
```
    Min       1Q   Median       3Q      Max
-1.8538  -0.8391  -0.4360   0.8906   1.9273
```

## Coefficients:
```
                  Estimate Std. Error   z value  Pr(>|z|)
(Intercept)    -5.328605   5.076190  -1.0500  0.29384
age             0.072286   0.016487   4.3840 1.16e-05 ***
sysbp           0.012845   0.014852   0.8650  0.38708
diabp          -0.029113   0.026398  -1.1030  0.27009
choles          0.007676   0.002390   3.2119  0.00132 **
ht              -0.053164   0.070796  -0.7511  0.45269
wt              0.020838   0.006768   3.0789  0.00208 **
```

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 360.51  on 274  degrees of freedom
Residual deviance: 288.26  on 268  degrees of freedom
AIC: 302.26

Number of Fisher Scoring iterations: 4
Model Fit & Effect Size

Under the **Model Fit submenu** select Deviance, Overall model test, and all the pseudo $R^2$

1. **Deviance:** This stat shows the predictive success of the model. The smaller the number, the better the model (in SPSS this is called 2 Log Likelihood in case you ever need to know).

2. Cox & Snell $R^2$ and Nagelkerke $R^2$ :*These two numbers in the model summary box are similar to $R^2$ in multiple regression (a proportion of the variance in the DV accounted for by the variables in model). We will report both of them as “% of variance accounted for”.
   
   – **Effect size notes:** Cox and Snell $R^2$ based on likelihoods and sample size BUT never can reach 1, even if you achieve perfect fit.
   
   – Use Nagelkerke $R^2$ which adjusts Cox and Snell so that the upper limit is 1 (most people report this type of effect size.)
## Call:

```
glm(formula = coron ~ age + sysbp + diabp + choles + ht + wt,
    family = "binomial", data = logistic.dat)
```

## Deviance Residuals:

```
     Min       1Q   Median       3Q      Max
-1.8538 -0.8391 -0.4360   0.8906   1.9273
```

## Coefficients:

```
                         Estimate Std. Error   z value  Pr(>|z|)
(Intercept)            -5.328605   5.076190  -1.050  0.29384
age                   -0.072286   0.016487   4.384 1.16e-05 ***
sysbp                  -0.012845   0.014852   0.865  0.38708
diabp                 -0.029113   0.026398  -1.103  0.27009
choles                 0.007676   0.002390   3.212  0.00132 **
ht                     -0.053164   0.070796  -0.751  0.45269
wt                     -0.020838   0.006768   3.079  0.00208 **
```

---

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

## (Dispersion parameter for binomial family taken to be 1)

Null deviance: 360.51 on 274 degrees of freedom
Residual deviance: 288.26 on 268 degrees of freedom
AIC: 302.26

## Number of Fisher Scoring iterations: 4
Model Fit & Effect Size

Under the Model Fit submenu select Deviance, Overall model test, and all the pseudo $R^2$.

1. **Deviance**: This stat shows the predictive success of the model. The smaller the number, the better the model (in SPSS this is called 2 Log Likelihood in case you ever need to know).

2. **Cox & Snell $R^2$ and Nagelkerke $R^2$**: These two numbers in the model summary box are similar to $R^2$ in multiple regression (a proportion of the variance in the DV accounted for by the variables in model). We will report both of them as “% of variance accounted for”.
   - **Effect size notes**: Cox and Snell $R^2$ based on likelihoods and sample size BUT never can reach 1, even if you achieve perfect fit.
   - Use Nagelkerke $R^2$ which adjusts Cox and Snell so that the upper limit is 1 (most people report this type of effect size.)
#install and load DescTools package
# install.packages('DescTools')
library(DescTools)

#calculate pseudo R-squared for model
PseudoR2(logistic_model, c("McFadden", "Nagel", "CoxSnell"))
##   McFadden Nagelkerke CoxSnell
## 0.2004085  0.3163152  0.2310492
Code: Odds Ratio

#Odds Ratio
exp(coef(logistic_model))

## (Intercept)         age         sysbp         diabp         choles         ht
##  0.004850832  1.074962215  1.012928265  0.971306984  1.007705705  0.948224528
##           wt
##  1.021056162
Interpreting Odds Ratio

What if...?

• **Scenario 1** Imagine height was significant and the odds ratio (OR) was .94. Then we would interpret the odds ratio like this:

  *The odds ratio indicates that for every unit increase in height the odds of the outcome decrease by a factor of .94.*

Odds Ratio for Categorical Variables

• **Scenario 2** Imagine that Weight is a categorical variable coded as in Weight = 0 means “not overweight” and Weight = 1 is “overweight.” Then we would interpret the odds ratio like this:

  *The odds that a person will experience the outcome are 1.02 times higher for those who are overweight than for those who are not.*
Resources

• Research Design & Data Analysis Lab: [https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/](https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/)

• Schedule a consultant appointment with me: [https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/ors-research-design-data-analysis-data-analysis-lab-consultants/](https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/ors-research-design-data-analysis-data-analysis-lab-consultants/)

• Check out Lab Resources (including recording of this webinar): [https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/resources/](https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/resources/)
References

