Logistic Regression

Using R

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

Dependent Variable

Y
Dichotomous
Yes/No
1/0

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:

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

- 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)?

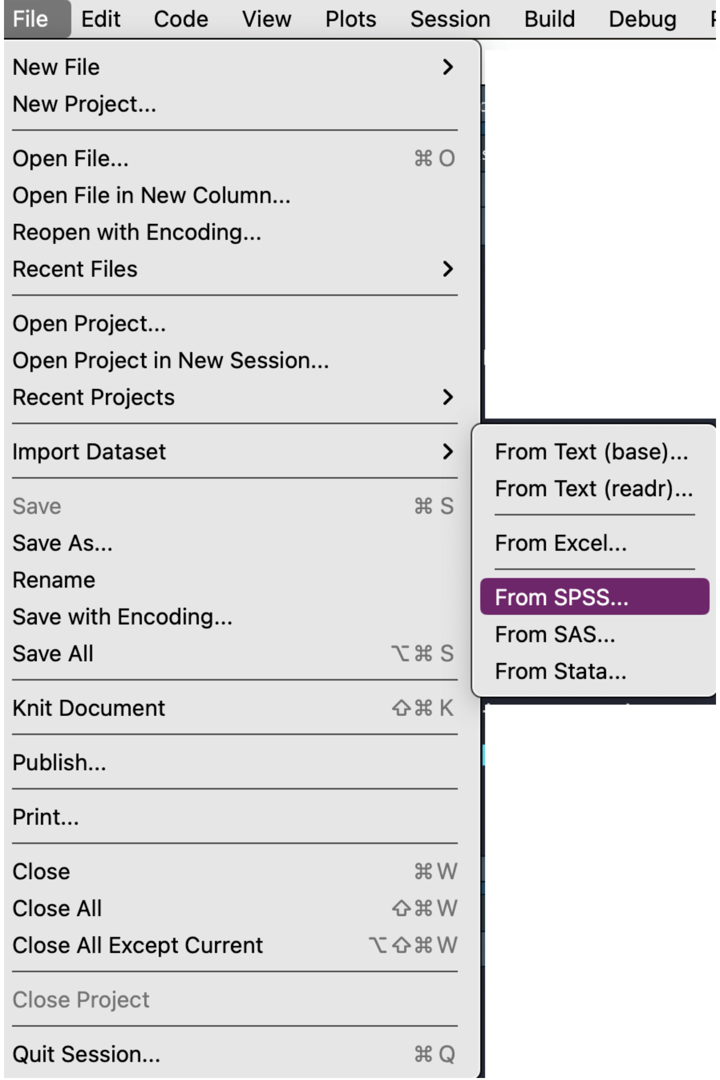
Logistic Regression in R

```
# 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



Descriptives

```
# Descriptive info
summary(logistic.dat)
##       age            sysbp           diabp            choles
##  Min.   :23.00   Min.   : 90.0   Min.   : 55.00   Min.   :135.0
##  1st Qu.:36.00   1st Qu.:110.0   1st Qu.: 75.50   1st Qu.:254.0
##  Median :45.00   Median :120.0   Median : 80.00   Median :285.0
##  Mean   :45.03   Mean   :124.2   Mean   : 82.97   Mean   :297.3
##  3rd Qu.:52.00   3rd Qu.:130.0   3rd Qu.: 90.00   3rd Qu.:336.5
##  Max.   :70.00   Max.   :190.0   Max.   :112.00   Max.   :520.0
##        ht              wt            coron
##  Min.   :62.00   Min.   :108.0   Min.   :0.0000
##  1st Qu.:67.00   1st Qu.:150.0   1st Qu.:0.0000
##  Median :68.00   Median :166.0   Median :0.0000
##  Mean   :68.45   Mean   :167.7   Mean   :0.3636
##  3rd Qu.:70.00   3rd Qu.:181.0   3rd Qu.:1.0000
##  Max.   :74.00   Max.   :262.0   Max.   :1.0000
```

Frequencies

```
# Frequency of the Dependent Variable
library(tidyverse)
library(formattable)
logistic.dat %>%
  group_by(coron) %>%
  summarize(Freq=n()) %>%
  mutate(freq = percent(Freq / sum(Freq))) %>%
  arrange(desc(Freq))
```

```
## # A tibble: 2 × 3
##   coron  Freq freq
##   <dbl> <int> <formttbl>
## 1     0   175 63.64%
## 2     1   100 36.36%
```

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 multicollinearty.

- We will do Simple Linear Regression to find the multicollinearity indicators

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

$$y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$$

```r
# Collinearity Diagnostics
# 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

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

$$y = b_1x_1 + b_2x_2 + \dots + b_nx_n + c$$

```
# 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  -1.050  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
```

REGRESSION EQUATION

B

P-VALUES

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
```

REGRESSION EQUATION

DEVIANCE

AKAIKE INFORMATION CRITERION

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.)**

Pseudo R^2

```
#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/
- Schedule a consultant appointment with me: https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/ors-research-design-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/

References

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3–14.

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., & Arppe, A. (2016). DescTools: Tools for descriptive statistics. R package version 0.99. 18. *R Found. Stat. Comput., Vienna, Austria*.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics*. Pearson.