Analyses Involving Categorical Dependent Variables

Gemma Wallace & Neil Yetz
PSY 653 Module 10 Lab
Apr 15, 2020

Logistic regression

- Logistic regression is used when you have a categorical outcome
- Uses a logit link to link the categorical outcome with the predictor variables
- You can derive interpretable odds ratios from logistic regression

Odds Ratios

Probability and odds ratios are fundamental for logistic regression

- A proportion can serve as an estimate for the probability (aka *risk*) of an event
  - E.g., If 75% of students who studied received a passing grade, the probability, or risk, of a randomly sampled student who studied getting a passing grade is 0.75
- Risk Ratios (aka relative risk) allow us to compare probability or risk *across multiple groups*
  - RR = (Probability of group 1/probability of group 2) → risk of outcome for one group compared to another group
- Odds Ratios = another way to describe the likelihood of an event
  - OR = (chances event will occur)/(chances event will not occur) → odds that the outcome will occur

Load Libraries

```
14 ## Load Libraries
15 ```{r}
16 library(tidyverse)
17 library(psych)
18 library(olsrr)
19 ```
```

Read in data

```
## Read in data
```{r}
lr <- read_csv("Logistic2.csv")
```
```

```
Parsed with column specification:
cols(
  Y = col_double(),
  X1 = col_double(),
  X2 = col_double(),
  X3 = col_double(),
  X4 = col_double()
)
```

This is a simulated dataset with N=164 and 5 variables:

Y: *A binary categorical variable* (Coded as 0 or 1).

X1: A binary variable (Coded as 0 or 1)

X2: A continuous variable ranging from 0 to 10

X3: A continuous variable ranging from 0 to 5

X4: A continuous variable ranging from 0 to 4

Goals for this activity:

1) Regress Y on all of the predictor (X) variables using OLS multiple regression and interpret the results.
2) Regress Y on all of the predictor (X) variables using logistic regression and interpret the results
3) Compare the results from the two analytic approaches

Though not shown in the demo, remember that it's good practice to examine descriptives and visualize your data before conducting analyses :)

Part 1: Conduct analyses with OLS Regression

OLS Regression Model 1

```
### Model 1
```{r}
ols_mod1 <- lm(Y ~ X1, data = lr)
ols_regress(ols_mod1)
```
```

```
                        Model Summary
---------------------------------------------------------------
R                       0.439       RMSE                 0.418
R-Squared               0.193       Coef. Var           60.719
Adj. R-Squared          0.188       MSE                  0.175
Pred R-Squared          0.174       MAE                  0.346
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
-----------------------------------------------------------------------
              Sum of
             Squares         DF    Mean Square      F          Sig.
-----------------------------------------------------------------------
Regression     6.785          1          6.785    38.764    0.0000
Residual      28.355        162          0.175
Total         35.140        163
-----------------------------------------------------------------------

                                 Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
-----------------------------------------------------------------------------------------
(Intercept)    0.500         0.045                 11.211    0.000    0.412    0.588
         X1    0.408         0.066        0.439     6.226    0.000    0.279    0.537
-----------------------------------------------------------------------------------------
```

```
51
```

Intercept: When X1 is zero, the expected Y is .500.

X1: For every one-unit increase X1, there is an expected .408 increase in Y.

This model explains 19.31% of the variance in Y.

We will use a hierarchical regression approach to build up to the full model that evaluates all of the predictors simultaneously. We'll compare model R^2 at each stage. No need to test interactions in this activity.

OLS Regression Model 2

```
### Model 2
```{r}
ols_mod2 <- lm(Y ~ X1 + X2, data = lr)
ols_regress(ols_mod2)
```
```

```
                         Model Summary
----------------------------------------------------------------
R                       0.442       RMSE                 0.419
R-Squared               0.196       Coef. Var           60.809
Adj. R-Squared          0.186       MSE                  0.176
Pred R-Squared          0.166       MAE                  0.345
----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square      F          Sig.
--------------------------------------------------------------------
Regression      6.876          2          3.438    19.585     0.0000
Residual       28.264        161          0.176
Total          35.140        163
--------------------------------------------------------------------

                                 Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     0.584         0.125                4.676     0.000     0.337     0.831
         X1     0.405         0.066        0.436   6.157     0.000     0.275     0.535
         X2    -0.013         0.019       -0.051  -0.721     0.472    -0.050     0.023
-----------------------------------------------------------------------------------------
```

```
66
```

Intercept: When all predictors are zero, the expected Y is .584.

X1: Holding all other variables constant; For every one-unit increase X1, there is an expected .405 increase in Y.

X2: Holding all other variables constant; For every one-unit increase X2, there is an expected .013 decrease in Y.

This model explains 19.58% of the variance in Y. This is not much higher than Model 1 (19.3%)

OLS Regression Model 3

```
### Model 3
```{r}
ols_mod3 <- lm(Y ~ X1 + X2 + X3, data = lr)
ols_regress(ols_mod3)
```
```

```
                        Model Summary
-----------------------------------------------------------------
R                       0.465       RMSE                 0.415
R-Squared               0.216       Coef. Var           60.210
Adj. R-Squared          0.202       MSE                  0.172
Pred R-Squared          0.176       MAE                  0.336
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
--------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square      F          Sig.
--------------------------------------------------------------------
Regression      7.603          3          2.534    14.724     0.0000
Residual       27.538        160          0.172
Total          35.140        163
--------------------------------------------------------------------

                                 Parameter Estimates
-----------------------------------------------------------------------------------------------
      model       Beta     Std. Error     Std. Beta       t        Sig      lower      upper
-----------------------------------------------------------------------------------------------
(Intercept)      0.680          0.132                   5.143     0.000      0.419      0.941
         X1      0.345          0.071         0.372     4.838     0.000      0.204      0.486
         X2     -0.012          0.018        -0.046    -0.661     0.509     -0.049      0.024
         X3     -0.053          0.026        -0.158    -2.054     0.042     -0.105     -0.002
-----------------------------------------------------------------------------------------------
```

Intercept: When all predictors are zero, the expected Y is .680.

X1: Holding all other variables constant; For every one-unit increase X1, there is an expected .345 increase in Y.

X2: Holding all other variables constant; For every one-unit increase X2, there is an expected -.0.12 increase in Y.

X3: Holding all other variables constant; For every one-unit increase X3, there is an expected -.053 increase in Y.

This model explains 21.64% of the variance in Y. This is not much higher than Model 2 (19.58%).

OLS Regression Model 4

```
### Model 4
```{r}
ols_mod4 <- lm(Y ~ X1 + X2 + X3 + X4, data = lr)
ols_regress(ols_mod4)
```
```

Model Summary

R	0.518	RMSE	0.402
R-Squared	0.268	Coef. Var	58.360
Adj. R-Squared	0.250	MSE	0.162
Pred R-Squared	0.222	MAE	0.314

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	9.431	4	2.358	14.581	0.0000
Residual	25.710	159	0.162		
Total	35.140	163			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.308	0.169		1.823	0.070	-0.026	0.643
X1	0.286	0.071	0.308	4.012	0.000	0.145	0.427
X2	-0.013	0.018	-0.050	-0.734	0.464	-0.049	0.022
X3	-0.029	0.026	-0.086	-1.114	0.267	-0.081	0.023
X4	0.117	0.035	0.255	3.362	0.001	0.048	0.185

Intercept: When all predictors are zero, the expected Y is .316.

X1: Holding all other variables constant; For every one-unit increase X1, there is an expected .287 increase in Y.

X2: Holding all other variables constant; For every one-unit increase X2, there is an expected -.013 increase in Y.

X3: Holding all other variables constant; For every one-unit increase X3, there is an expected -.028 increase in Y.

X4: Holding all other variables constant; For every one-unit increase X4, there is an expected .115 increase in Y.

This model explains 26.87% of the variance in Y. This is a bit higher than Model 3 (21.64%).

Summary of OLS model comparisons

Percent variance explained in Y:

Model 1 (Y~X1) = 19.31
Model 2 (Y~X1+X2) = 19.58
Model 3 (Y~X1+X2+X3) = 21.64
Model 4(Y~X1 + X2 + X3 + X4) = 26.87

While the effect size (i.e., R^2 difference) is most important, you can also test for significant improvements to model fit using anova().

```
### Hierarchical comparison
```{r}
anova(ols_mod1,
      ols_mod2,
      ols_mod3,
      ols_mod4)
```
```

```
Analysis of Variance Table

Model 1: Y ~ X1
Model 2: Y ~ X1 + X2
Model 3: Y ~ X1 + X2 + X3
Model 4: Y ~ X1 + X2 + X3 + X4
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1    162 28.355
2    161 28.264  1   0.09126  0.5644 0.4536063
3    160 27.538  1   0.72643  4.4926 0.0355952 *
4    159 25.709   1.82804 11.3055 0.0009681 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Part 2: Conduct analyses with Logistic Regression

The glm() function

```
## Logistic regression
```{r}

log_mod <- glm(Y ~ X1 + X2 + X3 + X4, family = binomial, data = lr)
summary(log_mod)

```
```

family = binomial tells the model that the outcome variable is binary (zeros and ones)

Logistic regression model

```
## Logistic regression
```{r}

log_mod <- glm(Y ~ X1 + X2 + X3 + X4, family = binomial, data = lr)
summary(log_mod)

```
```

```
Call:
glm(formula = Y ~ X1 + X2 + X3 + X4, family = binomial, data = lr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3956  -0.7618   0.3744   0.7864   1.6046

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.86309    0.98886  -0.873 0.382766
X1           1.77209    0.48405   3.661 0.000251 ***
X2          -0.08569    0.11189  -0.766 0.443785
X3          -0.15597    0.15370  -1.015 0.310210
X4           0.59549    0.20668   2.881 0.003962 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 203.32  on 163  degrees of freedom
Residual deviance: 155.59  on 159  degrees of freedom
AIC: 165.59

Number of Fisher Scoring iterations: 5
```

```
```

The model displays the **log odds** of each predictor variable (While controlling for all other predictors in the model) on the outcome of Y.

We can see that X1 and X4 are statistically significant. However, to have a better interpretation of each with odds ratios, we need to exponentiate the coefficients.

Note: we don’t need to build up the model step-by-step here because we can back-calculate the variance explained by each added variable (shown later!)

Exponentiate the coefficients and confidence intervals to obtain interpretable Odds ratios.

```
## Get ORs & 95% confidence intervals
```{r}
exp(coefficients(log_mod))
exp(confint(log_mod))
```
```

```
(Intercept)          X1          X2          X3          X4
  0.4218572   5.8831439   0.9178798   0.8555857   1.8139261
Waiting for profiling to be done...
                  2.5 %    97.5 %
(Intercept) 0.05830423  2.898238
X1          2.37624313 16.216290
X2          0.73114151  1.138133
X3          0.63116883  1.157472
X4          1.22321689  2.762664
```

Exponentiate the coefficients and confidence intervals to obtain interpretable odds ratios

```
## Get ORs & 95% confidence intervals
```{r}
exp(coefficients(log_mod))
exp(confint(log_mod))
```
```

```
(Intercept)          X1          X2          X3          X4
  0.4218572   5.8831439   0.9178798   0.8555857   1.8139261
Waiting for profiling to be done...
                   2.5 %    97.5 %
(Intercept) 0.05830423  2.898238
X1          2.37624313 16.216290
X2          0.73114151  1.138133
X3          0.63116883  1.157472
X4          1.22321689  2.762664
```

ODDS RATIOS

CONFIDENCE INTERVALS

Logistic regression model interpretations

```
## Get ORs & 95% confidence intervals
```{r}
exp(coefficients(log_mod))
exp(confint(log_mod))
```
```

```
(Intercept)          X1          X2          X3          X4
  0.4218572   5.8831439   0.9178798   0.8555857   1.8139261
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.05830423  2.898238
X1          2.37624313 16.216290
X2          0.73114151  1.138133
X3          0.63116883  1.157472
X4          1.22321689  2.762664
```

In logistic regression, an effect is significant if the confidence interval **does not contain 1** (not zero, as in ols analyses; odds of 1 represent equal odds)

Intercept: When all of the X variables are zero, the odds are .421 times as likely of developing the outcome of Y (Or we can take the inverse and state the they are 2.38 times as likely NOT to develop the outcome of Y). This is not statistically significant.

X1 (Binary): After controlling for all variables in the model, Those coded as 1 are 5.88 times as likely to develop the outcome of Y as compared to those coded 0. This is statistically significant.

X2 (Continuous): After controlling for all variables in the model, For every one unit increase in X2, there is an expected increase of 0.918 times of developing Y (Or we can take the inverse and state that for every one unit increase in X2, there is a 1.09 increase in the odds of NOT developing the outcome of Y). This is not statistically significant.

X3 (Continuous): After controlling for all variables in the model, For every one unit increase in X3, there is an expected increase of 0.856 times in the odds of developing Y (Or we can take the inverse and state that for every one unit increase in X2, there is a 1.17 increase in the odds of NOT developing the outcome of Y). This not is statistically significant.

X4 (Continuous): After controlling for all variables in the model, For every one unit increase in X4, there is an expected increase of 1.81 times in the odds of developing Y. This is statistically significant.

Logistic regression: Examine deviance between models

```
## Deviancy test
```{r}
anova(log_mod,test="Chisq")
```
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                  163     203.32
X1    1   34.604       162     168.72 4.04e-09 ***
X2    1    0.484       161     168.23 0.486443
X3    1    3.694       160     164.54 0.054620 .
X4    1    8.946       159     155.59 0.002781 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This compares deviance, an estimate of model fit, between each model and the null model. The values represent:

X1 = model with just X1 vs. NULL model
X2 = model with X1 + X2 vs. NULL model
X3 = model with X1 + X2 + X3 vs. NULL model
X4 = model with X1 + X2 + X3 + X4 vs. NULL model

These comparisons tell us whether adding information to the null model leads to better prediction. In this case, the X2 and X3 models do not significantly improve model fit.

Logistic regression: McFadden's R^2

McFadden R^2 = 1-(Deviance model/Deviance Null)

```
## Deviancy test
```{r}
anova(log_mod,test="Chisq")
```
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                   163     203.32
X1    1   34.604       162     168.72 4.04e-09 ***
X2    1    0.484       161     168.23 0.486443
X3    1    3.694       160     164.54 0.054620 .
X4    1    8.946       159     155.59 0.002781 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On the previous slide, we showed how deviance comparisons give information about how each subsequent model compares to the null model.

McFadden's R^2 allows you to estimate the *percent variance* explained by each model, which can serve as an effect size.

Logistic regression: McFadden's R^2

You can use the McFadden's R^2 values to compare changes in the percent of variance in Y for the addition of each variable, like we do in OLS hierarchical regression comparisons:

````
## Calculate Mcfadden R^2
```{r}
m1_mcfadden <- 1 - (168.72/203.32)
m2_mcfadden <- 1 - (168.23/203.32)
m3_mcfadden <- 1 - (164.54/203.32)
m4_mcfadden <- 1 - (155.59/203.32)

m1_mcfadden
m2_mcfadden
m3_mcfadden
m4_mcfadden
```
````

```
[1] 0.1701751
[1] 0.1725851
[1] 0.1907338
[1] 0.2347531
```

Percent variance explained in Y:

Model 1 (Y~X1) = 17.02
Model 2 (Y~X1+X2) = 17.26
Model 3 (Y~X1+X2+X3) = 19.07
Model 4(Y~X1 + X2 + X3 + X4) = 23.47

Part 3: Compare results from OLS vs. logistic regression

OLS regression

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig
(Intercept)	0.308	0.169		1.823	0.070
X1	0.286	0.071	0.308	4.012	0.000
X2	-0.013	0.018	-0.050	-0.734	0.464
X3	-0.029	0.026	-0.086	-1.114	0.267
X4	0.117	0.035	0.255	3.362	0.001

Percent variance explained in Y:

Model 1 (Y~X1) = 19.31
Model 2 (Y~X1+X2) = 19.58
Model 3 (Y~X1+X2+X3) = 21.64
Model 4(Y~X1 + X2 + X3 + X4) = 26.87

Logistic regression

```
Call:
glm(formula = Y ~ X1 + X2 + X3 + X4, family = binomial, data = lr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3956  -0.7618   0.3744   0.7864   1.6046

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.86309    0.98886  -0.873 0.382766
X1           1.77209    0.48405   3.661 0.000251 ***
X2          -0.08569    0.11189  -0.766 0.443785
X3          -0.15597    0.15370  -1.015 0.310210
X4           0.59549    0.20668   2.881 0.003962 **
```

Percent variance explained in Y:

Model 1 (Y~X1) = 17.02
Model 2 (Y~X1+X2) = 17.26
Model 3 (Y~X1+X2+X3) = 19.07
Model 4(Y~X1 + X2 + X3 + X4) = 23.47

While the results are similar, logistic regression is preferred because it provides concrete, meaningful interpretations.