WELCOME TO PSY 653 LAB!

MODULE 05:

ANALYSES INVOLVING CATEGORICAL DEPENDENT VARIABLES (LOGISTIC REGRESSION)

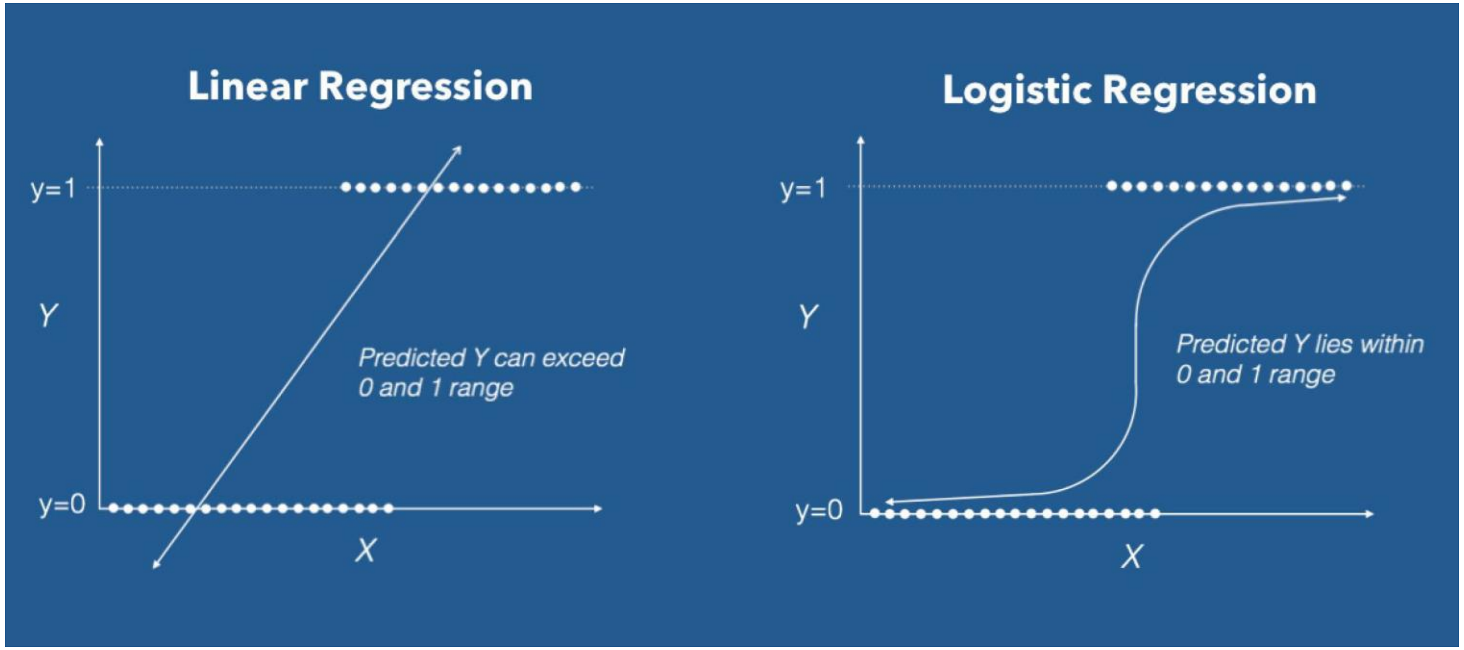
*Thanks to Gemma Wallace for her help with these slides

OBJECTIVES

- Quick review of logistic regression
- Odds ratios
- Coding tutorial

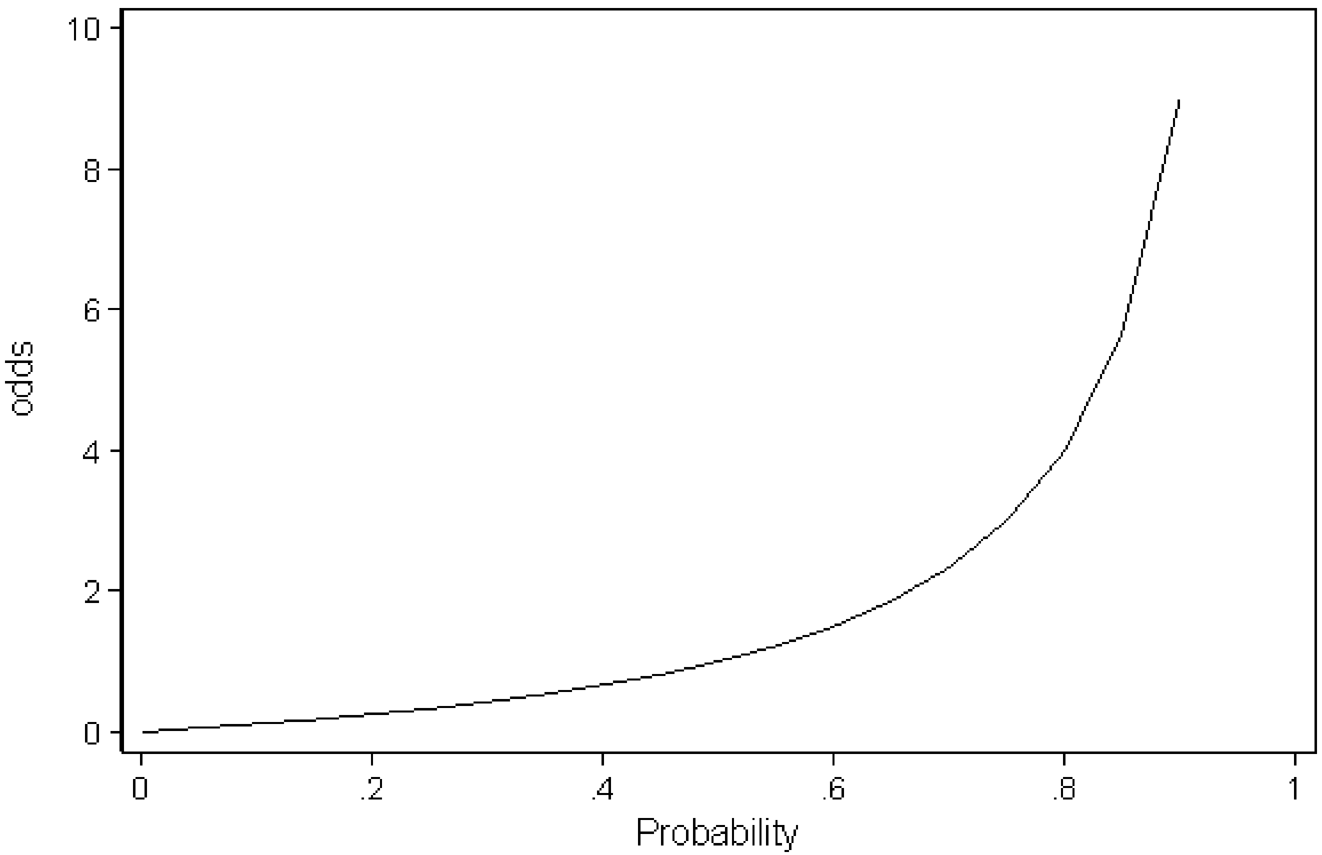
LOGISTIC REGRESSION

- Logistic regression is used when you have a binomial outcome
- Uses a logit link to link the categorical outcome with the predictor variables
- You can derive interpretable odds ratios from logistic regression
  - Which is super neat! Logistic makes it extremely easy to obtain these Odds ratios.



ODDS RATIOS (OR)

- Odds ratios range from zero to infinity!
- An OR **LOWER** than 1 indicates that an event is less likely to occur
- An OR **ABOVE** 1 indicates that the even is more likely to occur
- An OR of **EXACTLY** 1 indicates there is no relationship between the predictor and binary outcome
- In logistic regression, we'll be exponentiating our coefficients to obtain the odds ratios



CREATE A NEW R-PROJECT AND R-NOTEBOOK!

Download the “Logistic2.csv” file from Canvas and save it into your R-project file

LOAD LIBRARIES

```
## Load Libraries
```{r}
library(tidyverse)
library(psych)

```
```

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

	Y	X1	X2	X3	X4
1	1	1	2.2	3	3
2	0	0	5.6	1	0
3	1	0	4.4	3	3
4	0	0	3.3	0	2
5	1	1	4.4	0	4
6	0	1	5.6	0	4
7	0	0	6.7	3	1
8	1	1	7.8	0	4
9	1	1	6.7	2	3
10	1	0	5.6	2	2
11	1	0	7.8	0	4
12	1	0	6.7	0	2

A LOOK AT OUR DATASET

- Notice how Y is just a series of 0's and 1's.

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)

RUN A SIMPLE LOGISTIC REGRESSION MODEL ONE BINARY PREDICTOR

```
## Start simple, include one binary variable
```{r}
log_mod1 <- glm(Y ~ X1, family = binomial, data = lr)
summary(log_mod1)
```
```

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

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.1839  -1.1774   0.4396   1.1774   1.1774

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.251e-15  2.132e-01   0.000        1
X1          2.288e+00  4.503e-01   5.082 3.74e-07 ***
---
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: 168.72  on 162  degrees of freedom
AIC: 172.72

Number of Fisher Scoring iterations: 4
```

The model displays the **log odds** of the predictor variable on the outcome of Y.

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

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

````
### Get OR and 95% confidence intervals
```{r}
exp(coefficients(log_mod1))
exp(confint(log_mod1))
```
````

```
(Intercept)          X1
   1.000000    9.857142
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.6574447  1.521041
X1          4.3043085 25.718327
```

ODDS RATIOS

CONFIDENCE INTERVALS

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

```
### Get OR and 95% confidence intervals
```{r}
exp(coefficients(log_mod1))
exp(confint(log_mod1))
```
```

```
(Intercept)          X1
   1.000000    9.857142
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.6574447  1.521041
X1          4.3043085 25.718327
```

(Intercept): When all of the X variables are zero, the odds are 1.00 times as likely of developing the outcome of Y (Meaning the odds are completely even of developing Y). This is **not** statistically significant.

X1 (Binary): Those coded as 1 are 9.86 times as likely to develop the outcome of Y as compared to those coded 0. **This is statistically significant**. But it is a very wide confidence interval!

RUN A ANOTHER LOGISTIC REGRESSION MODEL ONE BINARY PREDICTOR & ONE CONTINUOUS PREDICTOR

```
## Make it a little harder, add one binary and one continuous variable
```{r}
log_mod2 <- glm(Y ~ X1 + X2, family = binomial, data = lr)
summary(log_mod2)
```
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2341  -1.1308   0.4323   1.0331   1.2588

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.44788    0.68087   0.658    0.511
X1           2.27664    0.45079   5.050 4.41e-07 ***
X2          -0.07159    0.10326  -0.693    0.488
---
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: 168.23  on 161  degrees of freedom
AIC: 174.23

Number of Fisher Scoring iterations: 4
```

The model displays the **log odds** of the predictor variable on the outcome of Y.

We can see that X1 is statistically significant and X2 is not. **However, to have a better interpretation of each with odds ratios, we need to exponentiate the coefficients.**

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

```
### Get OR and 95% confidence intervals
```{r}
exp(coefficients(log_mod2))
exp(confint(log_mod2))
```
```

```
(Intercept)          X1          X2
  1.5649851   9.7438486   0.9309089
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.4145744  6.158430
X1          4.2494457 25.443254
X2          0.7568849  1.138565
```

- **Intercept:** When all of the X variables are zero, the odds are 1.56 times as likely of developing the outcome of Y. This is **not** statistically significant.
- **X1** (Binary): After controlling for all variables in the model, Those coded as 1 are 9.74 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 0.931 times of developing Y (Or we can take the inverse and state that for every one unit increase in X2, there is a 1.07 increase in the odds of NOT developing the outcome of Y). This is **not** statistically significant.

LOGISTIC REGRESSION MODEL

```r
## 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.

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