

# Discrete Response Model

## Lecture 3

---

**[datascience@berkeley](mailto:datascience@berkeley)**

# Variable Transformation, Part 1: Interactions Among Explanatory Variables

# Introduction

Similar to linear regression models, we can include various transformations of explanatory variables in a logistic regression model.

In this section, we will study two of the most common transformations

- 1) two-way interactions
- 2) quadratic terms

Interactions between explanatory variables are needed when the effect of one explanatory variable on the probability of success depends on the value for a second explanatory variable.

# Formulation in R: Example 1

Consider the model of

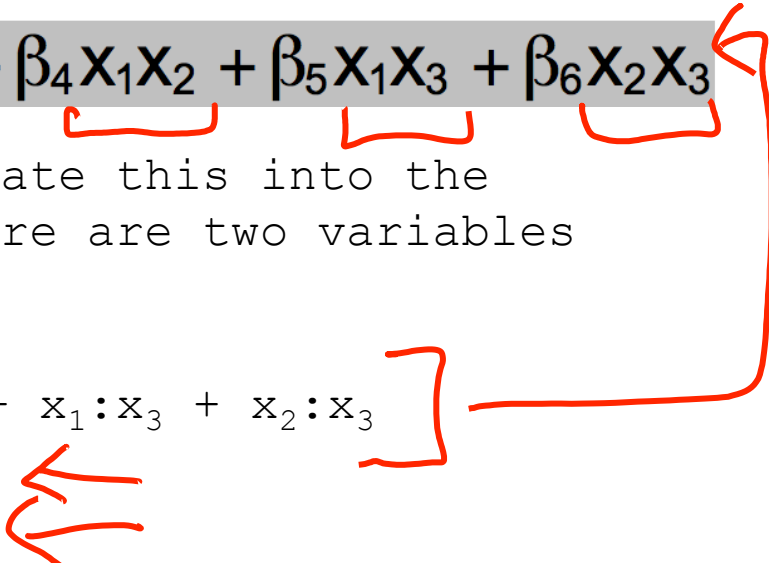
$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

There are several ways to incorporate this into the formula argument of **glm()** when there are two variables called  $x_1$  and  $x_2$  in a data frame:

```
formula = y ~ x1 + x2 + x1:x2  
formula = y ~ x1*x2  
formula = y ~ (x1 + x2)^2
```

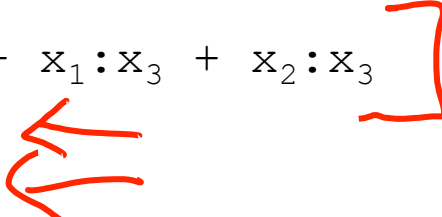
# Formulation in R: Example 2

Next, consider the model of

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3$$


There are several ways to incorporate this into the formula argument of `glm()` when there are two variables called  $x_1$  and  $x_2$  in a data frame:

```
formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3
formula = y ~ x1*x2 + x1*x3 + x2*x3
formula = y ~ (x1 + x2 + x3)^2
```



- I personally prefer the first setup because it mimics the actually underlying formula.
- However, data scientists have different preferences, and I have seen a lot of the second setup, too.
- In practice, you have to decide what's best for you (and the team in which you work). In many cases, you may just need to follow the convention/standards used in your company.

# Interpretation and Understanding Interaction

Consider again the model of

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Note that with the interaction, the effect that  $x_1$  has on  $\text{logit}(\pi)$ , the log of the odds ratio, is dependent on the specific value of  $x_2$ .
- Implication: We no longer can only look at  $\beta_1$  when trying to understand the effect  $x_1$  has on the response.
- In this specific setup, this also applies to the effect that  $x_2$  has on  $\text{logit}(\pi)$ .
- While we can still use odds ratios to interpret these effects, it is now a little more difficult.
- However, remember that the interpretation still is based on the ratio of two odds.

# Interpretation and Understanding Interaction

For the above model, the odds ratio for  $x_2$  holding  $x_1$  constant is

$$OR = \frac{\text{Odds}_{x_2+c}}{\text{Odds}_{x_2}} = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 (x_2+c) + \beta_3 x_1 (x_2+c)}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}} = e^{c\beta_2 + c\beta_3 x_1} = e^{c(\beta_2 + \beta_3 x_1)}$$

Notice how the  $x_1$  is influencing the effect of  $x_2$  on OR, which depends on both  $\beta_2$ , which measures the effect of  $x_2$  on OR without the interaction with  $x_1$ , and both  $\beta_3$ , which measures the impact of  $x_1$  on OR without the interaction with  $x_2$ , and the  $x_1$  itself

In other words, it needs to include  $x_1$  when interpreting  $x_2$ 's corresponding odds ratio, and the effect is a function and not a constant.

The odds of a success change by  $e^{c(\beta_2 + \beta_3 x_1)}$  times for every  $c$ -unit increase in  $x_2$  when  $x_1$  is fixed at a value of \_\_\_\_.

# Confidence Intervals

- Wald and profile likelihood ratio intervals again can be found for OR.
- With respect to the Wald interval, we use the same basic form as before, but now with a more complicated variance expression.
- For example, the interval for the  $x_2$  odds ratio in the

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

model is

$$e^{c(\hat{\beta}_2 + \hat{\beta}_3 x_1) \pm cZ_{1-\alpha/2} \sqrt{\text{Var}(\hat{\beta}_2 + \hat{\beta}_3 x_1)}}$$

where

$$\text{Var}(\hat{\beta}_2 + \hat{\beta}_3 x_1) = \text{Var}(\hat{\beta}_2) + x_1^2 \text{Var}(\hat{\beta}_3) + 2x_1 \text{Cov}(\hat{\beta}_2, \hat{\beta}_3)$$

- Variances and covariances can be found from the estimated covariance matrix.



# Remarks

- Profile likelihood ratio intervals are generally preferred, but they are more difficult to calculate due to the additional parameters included in the odds ratio, similar to what we saw in previous lecture.
- As in the previous lecture, I recommend always calculating both Wald and profile likelihood ratio intervals. If there appears to be problems with using the **mcprofile package**, use the Wald interval. (Refer to the details in the last lecture.)

# Example

- Continue with the placekick example in the last lecture.
- Suppose a 50-yard placekick will have a longer time period that the wind can affect it than a 20-yard placekick.
- In other words, a distance and wind interaction would be of interest to test.
- The wind explanatory variable in the dataset is a binary variable for placekicks attempted in windy conditions (1) vs. non-windy conditions (0), where windy conditions are defined as a wind stronger than 15 miles per hour at kickoff in an outdoor stadium.

# Berkeley

SCHOOL OF  
INFORMATION