Discrete Response Model
Lecture 3

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 logit(π), the log of the odds ratio, is dependent on the specific value of x_2.
- Implication: We no longer can only look at β_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 logit(π).
- 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

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

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 $\beta 1$, 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

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

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

Discrete Response Model
Lecture 3

datascience@berkeley

Variable Transformation, Part 1: Interactions Among Explanatory Variables–An Example

Example

A specification with distance, wind, and the distance × wind interaction:

```
# Load the data
setwd("/Users/jeffrey/Documents/JStuff/AdvStat/pgms/CatData/Chapter2")
#list.files("/Users/jeffrey/Documents/JStuff/AdvStat/pgms/CatData/Chapter2")

placekick<-read.table(file = "placekick.csv", header = TRUE, sep = ",")
str(placekick)
head(placekick)

# Estimate a GLM() model with distance and wind interaction
mod.fit.Ha<-glm(formula = good ~ distance + wind +
    distance:wind, family = binomial(link = logit), data =
    placekick)
summary(mod.fit.Ha)
```

Example

```
Call:
glm(formula = good ~ distance + wind + distance:wind, family = binomial(link = logit),
    data = placekick)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7291   0.2465   0.2465   0.3791   1.8647

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.684181   0.335962  16.919   <2e-16 ***
distance      -0.110253   0.008603 -12.816   <2e-16 ***
wind           2.469975   1.662144   1.486   0.1373
distance:wind -0.083735   0.043301  -1.934   0.0531 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1013.43  on 1424  degrees of freedom
Residual deviance:  767.42  on 1421  degrees of freedom
AIC: 775.42

Number of Fisher Scoring iterations: 6
```

Example (Log-Likehilood Ratio Test)

- To perform a LRT, we can fit the model under the null hypothesis and then use the anova() function:

```
# Likelihood-Ratio Test
mod.fit.Ho<-glm(formula = good ~ distance + wind, family
    = binomial(link = logit), data = placekick)
anova(mod.fit.Ho, mod.fit.Ha, test = "Chisq")
```

```
> mod.fit.Ho<-glm(formula = good ~ distance + wind, family
+     = binomial(link = logit), data = placekick)
> anova(mod.fit.Ho, mod.fit.Ha, test = "Chisq")
Analysis of Deviance Table

Model 1: good ~ distance + wind
Model 2: good ~ distance + wind + distance:wind
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1422     772.53
2      1421     767.42  1   5.1097  0.02379 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The test statistic is $-2\log(\Lambda) = 5.1097$, and the p-value is 0.0238, which provides an empirical evidence of a distance and wind interaction.

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Variable Transformation, Part 1: Interactions Among Explanatory Variables–An Example

Example (Log-Likehilood Ratio Test)

- Another way to obtain the LRT information would be to use the Anova() function from the car package:

```
> library(car)
> Anova(mod.fit.Ha, test = "LR")
Analysis of Deviance Table (Type II tests)

Response: good
              LR Chisq Df Pr(>Chisq)
distance       238.053  1    < 2e-16 ***
wind             3.212  1    0.07312 .
distance:wind    5.110  1    0.02379 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Berkeley SCHOOL OF INFORMATION

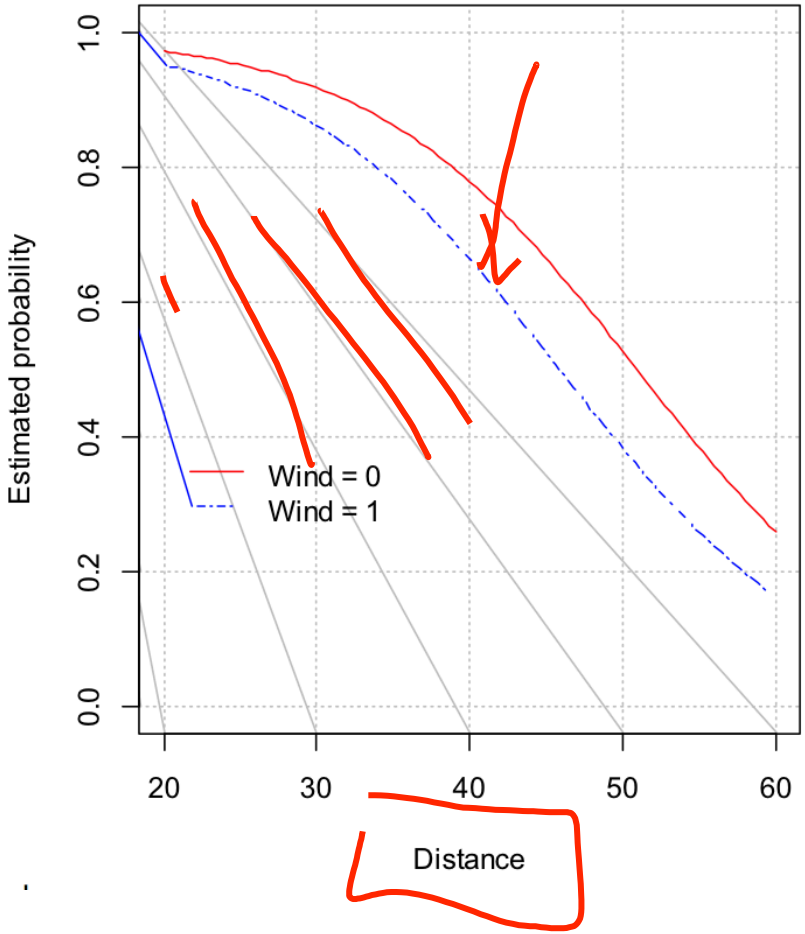
Discrete Response Model
Lecture 3

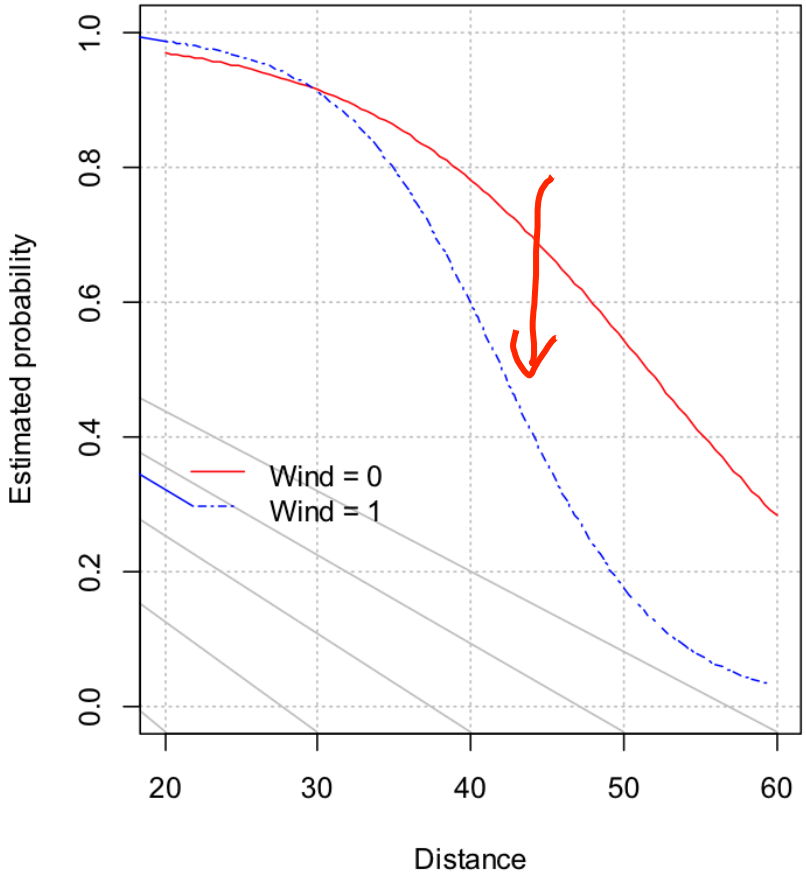
datascience@berkeley

Variable Transformation, Part 1: Interactions Among Explanatory Variables–An Example

Example: Visualization of the Interaction Effect

- The plot on the left does not include the interaction, whereas the plot on the right does.
- Observe how fast the estimated probability drops as the distance increases in the model with interaction effect.





Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Variable Transformation, Part 1: Interactions Among Explanatory Variables

Odds Ratio

- With interaction effect incorporated in the model, we need to find the odds ratio for wind comparing windy (1) vs. non-windy (0), holding distance constant.
- Recall the formula from our earlier discussion:

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

- For this equation, x_1 is distance and x_2 is wind. Of course, c = 1 for this setting due to wind being binary.
- Because distance could be anywhere from 18 to 66 yards, we use distances of 20, 30, 40, 50, and 60 yards to interpret the odds ratio for wind.

Interpretation of the Distance Odds Ratio

- For the distance odds ratio, we need to hold wind constant at 0 or 1.

- We also need to choose a value for **c** with respect to distance. The OR equation is

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

```
> round(data.frame(wind = wind, OR.hat = 1/OR.dist, OR.low
+     = 1/exp(ci.log.OR.up), OR.up = 1/exp(ci.log.OR.low)),2)
  wind OR.hat OR.low OR.up
1    0   3.01   2.54  3.56
2    1   6.96   3.03 15.98
```

Notice the odds ratios are inverted. Below are the interpretations:

- With 95% confidence, the odds of a success change by an amount between 2.54 to 3.56 times for every 10-yard decrease in distance under non-windy conditions.
- With 95% confidence, the odds of a success change by an amount between 3.03 to 15.98 times for every 10-yard decrease in distance under windy conditions.

Odds Ratio

```
# Odds Ratios
beta.hat<-mod.fit.Ha$coefficients[2:4]
c<-1
distance<-seq(from = 20, to = 60, by = 10)

OR.wind<-exp(c*(beta.hat[2] + beta.hat[3]*distance))
cov.mat<-vcov(mod.fit.Ha)[2:4,2:4]

#Var(beta^_2 + distance*beta^_3)
var.log.OR<-cov.mat[2,2] + distance^2*cov.mat[3,3] + 2*distance*cov.mat[2,3]

ci.log.OR.low<-c*(beta.hat[2] + beta.hat[3]*distance) - c*qnorm(p =
0.975)*sqrt(var.log.OR)
ci.log.OR.up<-c*(beta.hat[2] + beta.hat[3]*distance) + c*qnorm(p =
0.975)*sqrt(var.log.OR)

round(data.frame(distance = distance, OR.hat = 1/OR.wind, OR.low =
1/exp(ci.log.OR.up), OR.up = 1/exp(ci.log.OR.low)),2)
```

	distance	OR.hat	OR.low	OR.up
1	20	0.45	2.34	2.34
2	30	1.04	2.71	2.71
3	40	2.41	5.08	5.08
4	50	5.57	20.06	20.06
5	60	12.86	99.13	99.13

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Quadratic Term: An Introduction

Incorporate Quadratic Terms in R

- Quadratic and higher-order polynomials are one way to capture the nonlinear relationship between an explanatory variable and logit(π).
- To include this type of transformation in a formula argument, we can use the carat symbol ^ with the degree of the polynomial.
- However, as we saw earlier in this section, the carat symbol is used to denote the order of the interaction between explanatory variables. Thus,

  formula = y ~ x1 + x1^2

  would NOT be interpreted as $\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$.

- R interprets the x1^2 part as "all two-way interactions involving x1." Because only one explanatory variable is given in x1^2, R interprets this as simply x1. Also, because x1 was already given in the formula argument, the variable is not duplicated, so $\text{logit}(\pi) = \beta_0 + \beta_1 x_1$ is estimated instead!

Incorporate Quadratic Terms in R

- In order to obtain a x_1^2 terms, we need to use the I() function with x1^2.

- The I() function instructs R to interpret arguments as it normally would.
- Thus, formula = y ~ x1 + I(x1^2) would be interpreted as

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Quadratic Term: An Introduction

Odds ratios involving polynomial terms are dependent on the explanatory variable of interest. For example, to find OR for x_1 in

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

The corresponding odds ratio is

$$OR = \frac{\text{Odds}_{x_1+c}}{\text{Odds}_{x_1}} = \frac{e^{\beta_0+\beta_1(x_1+c)+\beta_2(x_1+c)^2}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_1^2}} = e^{c\beta_1+2cx_1\beta_2+c^2\beta_2} = e^{c\beta_1+c\beta_2(2x_1+c)}$$

The standard interpretation becomes

> The odds of a success change by $e^{c\beta_1+c\beta_2(2x_1+c)}$ times for a c-unit increase in x_1 when x_1 is at a value of ___.

Because the odds ratio is dependent on the explanatory variable value, it is better to change the interpretation to

> The odds of a success are $e^{c\beta_1+c\beta_2(2x_1+c)}$ times as large for x_1 = __ + c than for x_1 = __,

where you need to put in the appropriate value of x_1. Also, this means multiple odds ratios may be needed to fully understand the effect of x_1 on the response.

Wald Confidence Interval

Wald confidence intervals are found in a similar manner as for interaction terms.

For the model of $\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$, the interval is

$$e^{c\hat{\beta}_1 + c\hat{\beta}_2(2x_1+c) \pm cZ_{1-\alpha/2}\sqrt{\text{Var}(\hat{\beta}_1 + \hat{\beta}_2(2x_1+c))}}$$

where

$$\text{Var}(\hat{\beta}_1 + \hat{\beta}_2(2x_1 + c)) = \text{Var}(\hat{\beta}_1) + (2x_1 + c)^2 \text{Var}(\hat{\beta}_2) + 2(2x_1 + c)\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$

Profile likelihood ratio intervals can be calculated as well, but they are subject to the same problems as before with the mcprofile package.

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Quadratic Term: An Example

Example

Suppose x represents the distance of the placekick. Below is how we can estimate the model.

$$\text{logit}(\pi) = \beta_0 + \beta_1 x + \beta_2 x^2$$

```
mod.fit.distsq<-glm(formula = good ~ distance + I(distance^2), family =
binomial(link = logit), data = placekick)
summary(mod.fit.distsq)
```

Example

```
Call:
glm(formula = good ~ distance + I(distance^2), family = binomial(link = logit),
    data = placekick)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8625   0.2175   0.2175   0.4011   1.2865

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.8446831  1.0009079   7.838 4.59e-15 ***
distance     -0.2407073  0.0579403  -4.154 3.26e-05 ***
I(distance^2) 0.0017536  0.0007927   2.212    0.027 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1013.43  on 1424  degrees of freedom
Residual deviance:  770.95  on 1422  degrees of freedom
AIC: 776.95

Number of Fisher Scoring iterations: 6
```

The estimated models is

$$\text{logit}(\hat{\pi}) = 7.8447 - 0.2407x + 0.001754x^2$$

The p-value for the Wald test of $H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$ is 0.027, suggesting there is marginal evidence of a quadratic relationship between distance and the response.

Likelihood Ratio Test

A LRT provides a similar p-value.

```
library(package = car)
Anova(mod.fit.distsq)
```

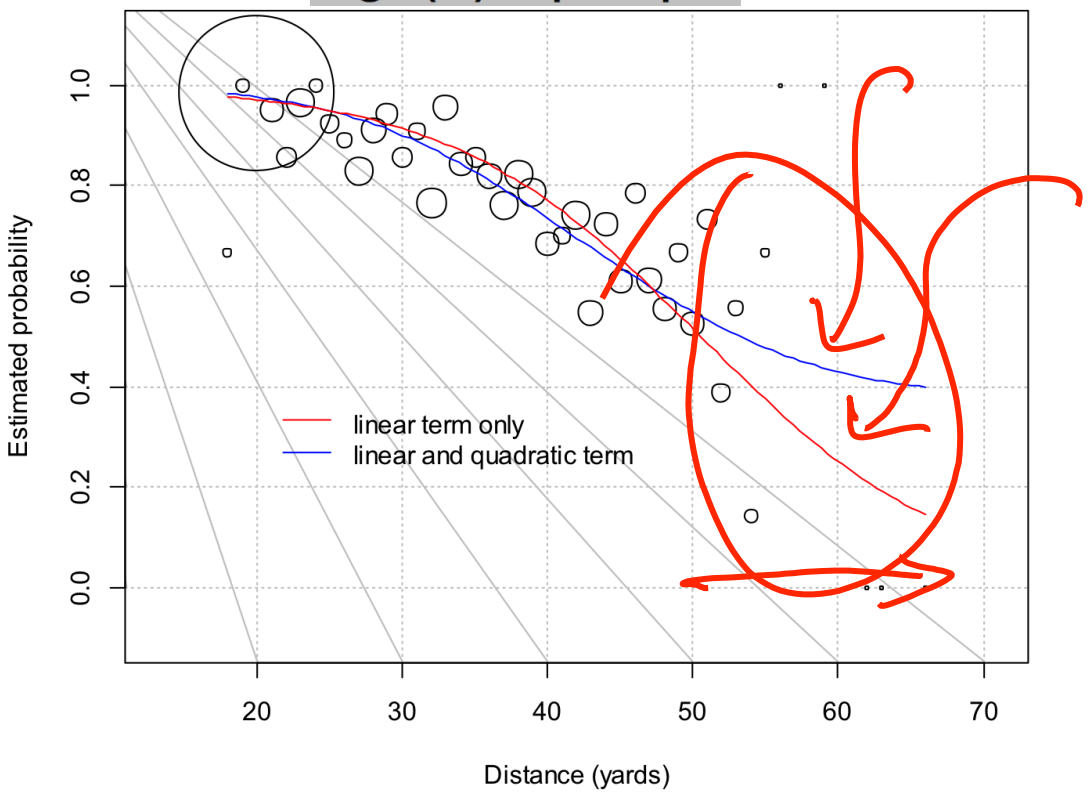
```
Analysis of Deviance Table (Type II tests)

Response: good
              LR Chisq Df Pr(>Chisq)
distance       16.9246  1  3.889e-05 ***
I(distance^2)   4.7904  1    0.02862 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Plot of the Estimated Model

Below is a plot of the estimated model, where the estimate of

$$\text{logit}(\pi) = \beta_0 + \beta_1 x$$



The main difference between the two models appears to be for the larger distances. Given the small number of observations at those distances, it may be difficult to justify the need for the quadratic term.

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Categorical Explanatory Variables

Introduction

- As we saw earlier with the change variable in the placekicking dataset, a categorical explanatory variable with two levels can be represented simply as a binary variable to reflect these two levels.
- When there are q levels (where **q > 2**), only **q – 1** binary (often referred to as "indicator") variables are needed to represent the variable, just like those in classical linear regression.

Formulation (via an Example)

Suppose an explanatory variable has levels of A, B, C, and D. Three indicator variables can be used to represent the explanatory variable in a model:

	Indicator variables		
Levels	x_1	x_2	x_3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

Notice how each level of the explanatory variable has a unique coding. The logistic regression model is

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Informally, we could also write out the full model as

$$\text{logit}(\pi) = \beta_0 + \beta_1 B + \beta_2 C + \beta_3 D$$

where it is assumed that B, C, or D are corresponding indicator variables in the model.

Interpretation

For example, A is the "base" level, and it is represented in the model with $x_1 = x_2 = x_3 = 0$ so that

$$\text{logit}(\pi) = \beta_0$$

For category B, the model becomes

$$\text{logit}(\pi) = \beta_0 + \beta_1$$

Thus, β_1 measures the effect of level B when compared to level A.

How R Treats Categorical Variables

- R treats categorical variables as a *factor* class type.
- By default, R orders the levels within a factor alphabetically, where numbers are ordered before letters and uppercase letters are before lowercase letters:

0, 1, 2, …, 9, …, a, A, b, B, …, z, Z

- To see the ordering of any factor, the levels() function can be used.
- This ordering of levels is important because R uses it to construct indicator variables with the "set first level to 0" method of construction.

Example: Control of the Tomato Spotted-Wilt Virus

- Plant viruses are often spread by insects.
- This occurs by insects feeding on plants already infected with a virus and subsequently becoming carriers of the virus.
- When they feed on other plants, insects may transmit this virus back to these new plants.
- To better understand the tomato spotted-wilt virus and how to control thrips that spread it, researchers at Kansas State University performed an experiment in a number of greenhouses.
- 100 uninfected tomato plants were put into each greenhouse, and they were introduced to the virus ("infested") in one of two ways (coded levels of the corresponding variable are given in parentheses):

1. Interspersing additional infected plants among the clean ones, and then releasing "uninfected" thrips to spread the virus (1).
2. Releasing thrips that carry the virus (2).

Example

To control the spread of the virus to the plants, the researchers used one of three methods:

1) Biologically through using predatory spider mites (B)
2) Chemically using a pesticide (C)
3) None (N)

The number of plants not displaying symptoms of infection were recorded for each greenhouse after eight weeks.

```
> tomato<-read.table(file = "TomatoVirus.csv", header = TRUE, sep = ",")
> head(tomato)
  Infest Control Plants Virus8
1      1       C    100     21
2      2       C    100     10
3      1       B    100     19
4      1       N    100     40
5      2       C    100     30
6      2       B    100     30
```

Both the Control and Infest explanatory variables are categorical in nature.

```
> class(tomato$Control)
[1] "factor"
> levels(tomato$Control)
[1] "B" "C" "N"
```

Example

For demonstration purposes, we will change the Infest variable to be a factor in the data frame:

```
> tomato$Infest<-factor(tomato$Infest)
> class(tomato$Infest)
[1] "factor"
```

We estimate the model incorporating both the Infest and Control variables:

Example

```
> mod.fit<-glm(formula = Virus8/Plants ~ Infest +
+     Control, family = binomial(link = logit), data =
+     tomato, weight = Plants)
> summary(mod.fit)

Call:
glm(formula = Virus8/Plants ~ Infest + Control, family = binomial(link = logit),
    data = tomato, weights = Plants)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.288  -2.425  -1.467   1.828   8.379

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6652     0.1018  -6.533 6.45e-11 ***
Infest2       0.2196     0.1091   2.013   0.0441 *
ControlC     -0.7933     0.1319  -6.014 1.81e-09 ***
ControlN      0.5152     0.1313   3.923 8.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 278.69  on 15  degrees of freedom
Residual deviance: 183.27  on 12  degrees of freedom
AIC: 266.77

Number of Fisher Scoring iterations: 4
```

```
> levels(tomato$Control)
[1] "B" "C" "N"
> levels(tomato$Infest)
[1] "1" "2"
```

Example

Because the response variable is given in a binomial form, we used the **weight argument** along with the success/trials formulation in the formula argument. Estimated model:

$$\text{logit}(\hat{\pi}) = -0.6652 + 0.2196\text{Infest2} - 0.7933\text{C} + 0.5152\text{N}$$

- Based on the positive estimated parameter for Infest2, **the probability of showing symptoms** is estimated to be larger in greenhouses where infestation method 2 is used.
- Based on the estimated parameters for C and N, the **estimated probability of showing symptoms** is lowest for the chemical control method and highest for when no control method is used.
- Note that these interpretations rely on their not being an interaction between the explanatory variables in the model.

Hypothesis Testing

- As with classical linear regression, all indicator variables must be included in a hypothesis test to evaluate the importance of a categorical explanatory variable.

Consider again the example with the categorical explanatory variable having four levels:

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

To evaluate the importance of this explanatory variable, we need to test

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

H_a: At least one β not equal to 0

Note: Three separate Wald tests of $H_0: \beta_i = 0$ vs. $H_a: \beta_i \neq 0$ are not appropriate.

One could do one overall Wald test, but we will focus instead on using a LRT because it performs better.

Interactions including categorical explanatory variables

Multiply each indicator variable by the model terms representing the other explanatory variable(s).

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Odds Ratio in the Context of Categorical Explanatory Variables

Odds ratios are useful for interpreting a categorical explanatory variable in a model; however, they are easily misinterpreted.

For example, suppose we want to interpret $OR = e^{\beta_1}$ from the model

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

in the example with the categorical explanatory variable having 4 levels. A common **mistake** is to interpret this odds ratio as

The odds of a success are e^{β_1} times as large as for level B than *all of the other levels*

	Indicator variables		
Levels	x_1	x_2	x_3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

To see why this is wrong, note that the odds of a success at level B are

$$e^{\beta_0+\beta_1 1+\beta_2 0+\beta_3 0} = e^{\beta_0+\beta_1}$$

	Indicator variables		
Levels	x_1	x_2	x_3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

In order to have a resulting OR = e^{β_1}, we need the denominator of the odds ratio to be e^{β_0}. Thus, $x_1 = x_2 = x_3 = 0$ for this second odds, which corresponds to level A of the categorical variable. The correct interpretation of the odds ratio is then

> The odds of a success are e^{β_1} times as large as for level B than for level A

Similar interpretations are found for e^{β_2} (compare level C to A) and for e^{β_3} (compare level D to A).

What if you would like to compare level B to level C? You need to find the ratio of two odds:

$$OR = \frac{Odds_{x_1=1,x_2=0,x_3=0}}{Odds_{x_1=0,x_2=1,x_3=0}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0+\beta_2}} = e^{\beta_1-\beta_2}$$

Similarly, the odds ratios can be found for comparing level B to level D as $e^{\beta_1 - \beta_3}$ and level C to level D as $e^{\beta_2 - \beta_3}$.

- Again, please remember that an odds ratio is just the ratio of two odds.
- Whenever you have difficulty understanding an odds ratio, go back to the basics and form the ratio.

Example

- For illustration, let's start with a model without the interaction, though the interaction between infestation and control is significant

The estimate model is

$$\text{logit}(\hat{\pi}) = -0.6652 + 0.2196\text{Infest2} - 0.7933\text{C} + 0.5152\text{N}$$

The estimated odds ratios for the control methods are:

```
> exp(mod.fit$coefficients[3:4])
ControlC ControlN
0.452342 1.674025
```

```
> exp(mod.fit$coefficients[4] - mod.fit$coefficients[3])
ControlN
3.700795
```

For example, the estimated odds ratio comparing level N to level B is $e^{0.5152} = 1.67$.

The estimated odds of plants showing symptoms are 1.67 times as large for using no control methods than using a biological control, where the infestation method is held constant.

Because we would prefer to REDUCE the proportion of plants showing symptoms, it may be of more interest to invert the odds ratio:

> The estimated odds of plants showing symptoms are 1/1.67 = **0.5973 times** as large for using a biological control method than using no control methods, where the infestation method is held constant.
>
> **Thus, using the spider mites (biological control) is estimated to reduce the odds of a plant showing symptoms by approximately 40%.**

In order to compare the no control to the chemical control, the odds ratio is

$$OR = \frac{Odds_{C=0,N=1}}{Odds_{C=1,N=0}} = \frac{e^{\beta_0+\beta_3}}{e^{\beta_0+\beta_2}} = e^{\beta_3-\beta_2}$$

The estimated odds ratio is $e^{0.5152-(-0.7933)} = 3.70$.

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Odds Ratio in the Context of Categorical Explanatory Variables

Confidence Interval

In order to compare the no control to the chemical control, the odds ratio is

$$OR = \frac{Odds_{C=0,N=1}}{Odds_{C=1,N=0}} = \frac{e^{\beta_0+\beta_3}}{e^{\beta_0+\beta_2}} = e^{\beta_3-\beta_2}$$

The estimated odds ratio is $e^{0.5152-(-0.7933)} = 3.70$

```
> K<-matrix(data = c(0, 0,  1,  0,
+                    0, 0,  0,  1),  nrow = 2, ncol = 4,
+       byrow = TRUE)
> linear.combo<-mcprofile(object = mod.fit, CM = K)
> ci.log.OR<-confint(object = linear.combo, level = 0.95, adjust = "none")
> ci.log.OR

  mcprofile - Confidence Intervals

level:            0.95
adjustment:       none

   Estimate  lower  upper
C1   -0.793 -1.054 -0.536
C2    0.515  0.258  0.773
```

```
> comparison<-c("C vs. B", "N vs. B")
> data.frame(comparison, OR = exp(ci.log.OR$confint))
  comparison  OR.lower  OR.upper
1    C vs. B 0.3486325 0.5848772
2    N vs. B 1.2945688 2.1665987
```

For example, the 95% profile LR confidence interval comparing level N to level B is 1.29 to 2.17. Thus,

> With 95% confidence, the odds of plants showing symptoms are **between 1.29 and 2.17 times as large** when using no control methods rather than using a biological control (holding the infestation method constant)

Alternatively, we could also say

With 95% confidence, the odds of plants showing symptoms are between 0.46 and 0.77 times as large when using a biological control method rather than using no control methods (holding the infestation method constant). **Thus, using the spider mites (biological control) is estimated to reduce the odds of a plant showing symptoms by approximately 23% to 54%.**

Model with Interactions

The estimated model is

$$\text{logit}(\hat{\pi}) = -1.0460 + 0.9258\text{Infest2} - 0.1623\text{C} + 1.1260\text{N} - 1.2114\text{Infest2} \times \text{C} - 1.1662\text{Infest2} \times \text{N}$$

To understand the effect of Control on the response, we will need to calculate odds ratios where the level of Infest2 is fixed at either 0 or 1. The odds ratio comparing level N to level B with Infest2 = 0 is

$$OR = \frac{\text{Odds}_{C=0,N=1,\text{infest2}=0}}{\text{Odds}_{C=0,N=0,\text{infest2}=0}} = \frac{e^{\beta_0+\beta_3}}{e^{\beta_0}} = e^{\beta_3}$$

The odds ratio comparing level N to level B with Infest2 = 1 is

$$OR = \frac{\text{Odds}_{C=0,N=1,\text{infest2}=1}}{\text{Odds}_{C=0,N=0,\text{infest2}=1}} = \frac{e^{\beta_0+\beta_1+\beta_3+\beta_5}}{e^{\beta_0+\beta_1}} = e^{\beta_3+\beta_5}$$

Model with Interactions

Other odds ratios can be calculated in a similar manner. Below are all of the estimated odds ratios and corresponding confidence intervals for Control holding Infest2 constant:

```
mcprofile - Confidence Intervals

level:              0.95
adjustment:         none

   Estimate  lower  upper
C1   1.1260  0.750  1.508
C2  -0.0402 -0.400  0.318
C3  -0.1623 -0.536  0.210
C4  -1.3738 -1.750 -1.009
C5   1.2884  0.905  1.678
C6   1.3336  0.934  1.742
```

Model with Interactions

```
data.frame(Infest2 = c(0, 1, 0, 1, 0, 1), comparison, OR
  = round(exp(ci.log.OR$estimate),2), OR.CI =
  round(exp(ci.log.OR$confint),2))
```

	Infest2	comparison	Estimate	OR.CI.lower	OR.CI.upper
C1	0	N vs. B	3.08	2.12	4.52
C2	1	N vs. B	0.96	0.67	1.37
C3	0	C vs. B	0.85	0.58	1.23
C4	1	C vs. B	0.25	0.17	0.36
C5	0	N vs. C	3.63	2.47	5.36
C6	1	N vs. C	3.79	2.54	5.71

Model with Interactions

```
ci.logit.wald<-confint(object = save.wald, level = 0.95,
  adjust = "none")
data.frame(Infest2 = c(0, 1, 0, 1, 0, 1), comparison, OR
  = round(exp(ci.log.OR$estimate),2), lower =
  round(exp(ci.logit.wald$confint[,1]),2), upper =
  round(exp(ci.logit.wald$confint[,2]),2))
```

	Infest2	comparison	Estimate	lower	upper
C1	0	N vs. B	3.08	2.11	4.50
C2	1	N vs. B	0.96	0.67	1.38
C3	0	C vs. B	0.85	0.59	1.23
C4	1	C vs. B	0.25	0.17	0.37
C5	0	N vs. C	3.63	2.46	5.34
C6	1	N vs. C	3.79	2.53	5.68

Model with Interactions

The columns of K are ordered corresponding to the 6 parameters estimated by the model. For example, row 2 corresponds to estimating

$$OR = \frac{Odds_{C=0,N=1,infest2=1}}{Odds_{C=0,N=0,infest2=1}} = \frac{e^{\beta_0+\beta_1+\beta_3+\beta_5}}{e^{\beta_0+\beta_1}} = e^{\beta_3+\beta_5}$$

where the 4^{th} and 6^{th} columns of K have 1's for the 4^{th} and 6^{th} parameters. Remember the first parameter in the model is β_0 so this is why the column numbers are 1 higher than the indices for the β's.

The estimated odds ratio comparing level N to level B with Infest2 = 1 is $e^{1.1260 - 1.1662} = 0.96$. Thus,

> The estimated odds of plants showing symptoms are 0.96 times as large for using no control than using a biological control when infected thrips are released into the greenhouse.

We can also see why the interaction between Infest and Control was significant. The N vs. B and C vs. B odds ratio differ by a large amount over the two levels of Infest. However, there is not much of a difference for N vs. C over the levels of Infest.

Model with Interactions

The 95% profile likelihood ratio interval comparing level N to level B with Infest2 = 1 is 0.67 < OR < 1.38. Thus,

> With 95% confidence, the odds of plants showing symptoms are between 0.67 and 1.38 times as large for using no control methods than using a biological control when infected thrips are released in the greenhouse. Because 1 is within the interval, there is not sufficient evidence to conclude a biological control is effective in this setting.

Notice the interval for comparing level N to level B with Infest2 = 0 is 2.11 < OR < 4.50. Because the interval is above 1, there is sufficient evidence to conclude the biological control reduces the odds of plants showing symptoms when interspersing infected plants with uninfected thrips.

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Convergence Criteria and the Case of Complete Separation

glm() Convergence Criteria

The glm() function uses IRLS until convergence is obtained or until the maximum number of iterations are reached. To determine convergence, glm() does not look at the successive estimates of the parameters directly; rather it examines the residual deviance. If we let $G^{(k)}$ denote the residual deviance at iteration k, then convergence occurs when

$$\frac{\left|G^{(k)} - G^{(k-1)}\right|}{0.1 + \left|G^{(k)}\right|} < \acute{\iota}$$

where ε is some specified small number greater than 0. The numerator provides a measure to determine if the $\hat{\pi}_i$ for $i = 1, \ldots, n$ are changing much from one iteration to the next. The denominator helps to take into account the relative size of the residual deviance.

glm() Convergence Criteria

The glm() function provides a few ways to control how convergence is decided:

- The epsilon argument sets ε above. The default is $\varepsilon = 10^{(-8)}$.
- The maxit argument states the maximum number of iterations allowed for the numerical procedure. The default is maxit = 25.
- The trace argument value can be used to see the actual $G^{(k)}$ values for each iteration. The default is trace = FALSE (do not show these values).

glm() Convergence Criteria

Consider the model with only distance as the explanatory variable:

$$\text{logit}(\hat{\pi}) = 5.8121 - 0.1150\text{distance}$$

Below are the results from using the glm() to estimate the model and from including the three arguments controlling convergence. Note that these three argument values were chosen for illustrative purposes only:

```
> mod.fit<-glm(formula = good ~ distance, family =
+   binomial(link = logit), data = placekick, trace = TRUE,
+   epsilon = 0.0001, maxit = 50)
Deviance = 836.7715 Iterations - 1
Deviance = 781.1072 Iterations - 2
Deviance = 775.8357 Iterations - 3
Deviance = 775.7451 Iterations - 4
Deviance = 775.745 Iterations - 5
> mod.fit$control
$epsilon
[1] 1e-04

$maxit
[1] 50

$trace
[1] TRUE
```

glm( ) Convergence Criteria

The convergence criteria value for iteration k = 5 is

$$\frac{\left|G^{(5)} - G^{(4)}\right|}{0.1 + \left|G^{(4)}\right|} = \frac{\left|775.745 - 775.7451\right|}{0.1 + \left|775.745\right|} = 1.3 \times 10^{-7}$$

which is less than the stated ε = 0.0001, so the iterative numerical procedure stopped. For iteration k = 4, the convergence criteria value is 0.00012, which is greater than 0.0001, so this is why the procedure continued.

If the value for maxit was changed to 3, the message

```
Warning message:
glm.fit: algorithm did not converge
```

would be printed to warn you that convergence was not obtained. Of course, you would NOT use the parameter estimates in this situation!

Complete Separation

Consider a simple dataset with one explanatory variable x_1 that is less than 6 when y = 0 and greater than or equal to 6 when y = 1. Because x_1 perfectly separates out the two possible values of y, complete separation occurs. Below is the corresponding R code and output:

```
> set1<-data.frame(x1 = c(1,2,3,4,5,6,7,8,9,10), y =
+     c(0,0,0,0,0,1,1,1,1,1))
> set1
   x1 y
1   1 0
2   2 0
3   3 0
4   4 0
5   5 0
6   6 1
7   7 1
8   8 1
9   9 1
10 10 1
```

Complete Separation

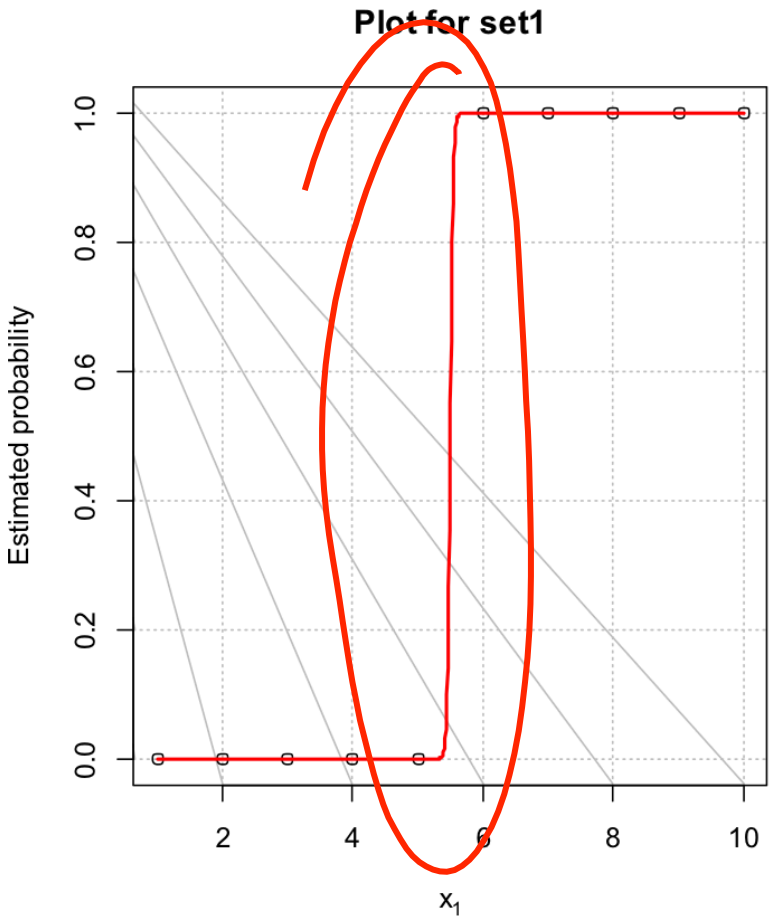
```
> mod.fit1<-glm(formula = y ~ x1, data = set1, family = binomial(link = logit), trace =
TRUE)
Deviance = 4.270292 Iterations - 1
Deviance = 2.574098 Iterations - 2
```

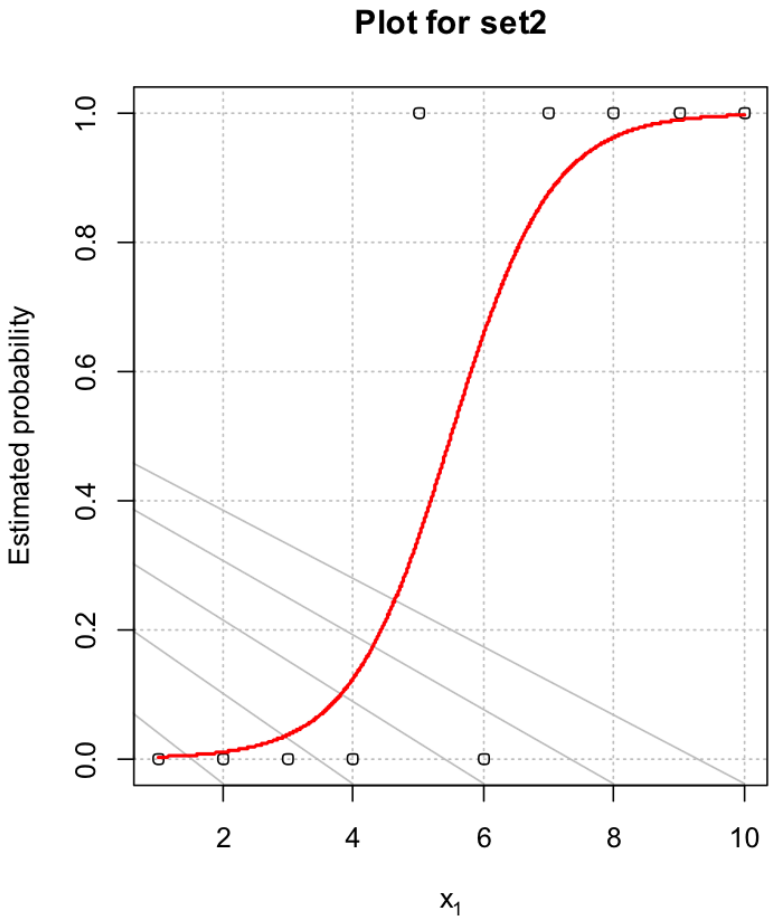
```
Deviance = 2.137736e-09 Iterations - 24
Deviance = 7.864775e-10 Iterations - 25
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
> mod.fit1$coefficients
(Intercept)          x1
 -245.84732    44.69951
```

R indicates that both convergence did not occur and at least some estimates of π are 0 or 1. Next is a plot (left side) of the data and the model at iteration #25:

Complete Separation





Complete Separation

Because there is a separation between the y = 0 and 1 values, the slope of the line between x = 5 and 6 will continue to get larger as the iterations continue. Essentially, the β_1 estimate is going to infinity with continued iterations. Notice this means the estimate is VERY biased.

Interestingly, R indicates "convergence" after 26 iterations if you increase maxit! However, the same

glm.fit: fitted probabilities numerically 0 or 1 occurred

message will occur. You should not use logistic regression here because the parameter estimates will continue to change for a larger number of iterations. Try this yourself with a larger maxit and smaller epsilon.

By reversing the y values at x_1 = 5 and 6, we obtain model convergence in 6 iterations (not shown here). The right plot above shows the data and the final model. The slope of the model is now not as great as was before.

Complete Separation: Remarks

- Complete separation is not necessarily bad if you want to distinguish between the response levels of y. The problem is that the model estimated by maximum likelihood does not provide a good way to interpret the relationship between y and the explanatory variables.
- It can be difficult to see complete separation graphically if there is more than one explanatory variable. There may be times even when the glm() function does not provide a warning. When parameter estimates are very large or very small with large estimated standard deviations, this is a sign that complete separation may exist. These types of parameter estimates can then lead to many observations with estimated probabilities of success close to 0 or 1.

Berkeley SCHOOL OF INFORMATION

Discrete Response Model
Lecture 3

datascience@berkeley

Generalized Linear Model

Generalized Linear Model

Logistic regression models fall within a family of models called *generalized linear models*. Each generalized linear model has three different components:

1. **Random:** This specifies the distribution for Y. For the logistic regression model, Y has a Bernoulli distribution.
2. **Systematic:** This specifies a linear combination of the β parameters with the explanatory variables, and it is often referred to as the linear predictor. For the logistic regression model, we have $\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$.
3. **Link:** This specifies how the expected value of the random component E(Y) is linked to the systematic component. In logistic regression model, logit is the link function.

Note that "linear" in generalized linear models comes from the β parameters simply being coefficients for the explanatory variables in the model. Nonlinear models involve more complex functional forms such as x^{β}.

Probit Regression Model

While the logit-link function is the most prevalently used for binary regression, there are two other functions that are common:

Inverse CDF of a standard normal distribution: This produces what is known as a probit regression model.

Suppose $\Phi(\cdot)$ denotes the CDF of a standard normal distribution. The model is written as

$$\pi = \Phi(\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p)$$

or equivalently as

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

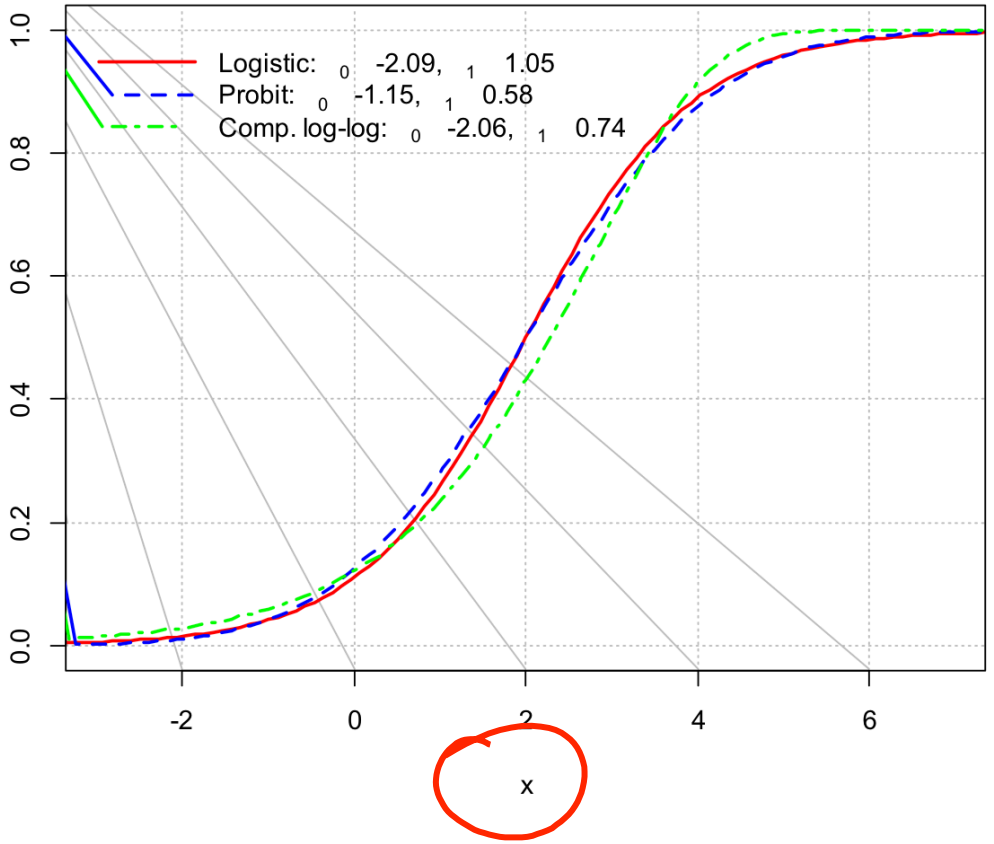
A very common way to express the model is

$$\text{probit}(\pi) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

where probit is used to denote the inverse CDF transformation in a similar manner as the logit transformation is for logistic regression.

Comparison of Three Types of Models

Suppose the linear predictor has only one explanatory variable of the form $\beta_0 + \beta_1 x$. By choosing values of β_0 and β_1 so that the mean is 2 and the variance is 3 for the corresponding CDFs, we obtain the plots of the models displayed below (see the corresponding program for code):



Comparison of Three Types of Models

Estimation: Probit and complementary log-log models are estimated in the same way as the logistic regression model. The difference now is that π is represented in the log-likelihood function by the corresponding probit or complementary log-log model specification.

Inference: Once the parameter estimates are found, the same inference procedures as used for the logistic regression model are available for the probit and complementary log-log models.

Odds ratios: Odds ratios are not as easy to interpret with probit and complementary log-log regression models as they were for logistic regression models. The same simplification does not hold true for probit and complementary log-log models. This is one of the main reasons why logistic regression models are the most used binary regression models.

Comparison of Three Types of Models

Example: Odds ratios used with probit models

Consider the model probit(π) = $\beta_0 + \beta_1 x$. The odds of a success are

$$\text{Odds}_x = \Phi(\beta_0 + \beta_1 x) / \left[1 - \Phi(\beta_0 + \beta_1 x)\right]$$

at a particular value of x. The odds of a success with a c-unit increase in x are

$$\text{Odds}_{x+c} = \Phi(\beta_0 + \beta_1 x + \beta_1 c) / \left[1 - \Phi(\beta_0 + \beta_1 x + \beta_1 c)\right]$$

When the odds ratio is formed from these two odds, x will remain in the final expression! Therefore, the odds ratio for a probit regression model depends on the value of the explanatory variable.

Berkeley SCHOOL OF INFORMATION