WHEN TO USE CONTINUOUS VS. ORDINAL

Dr. Aric LaBarr
Institute for Advanced Analytics

WHAT'S THE PROBLEM?

What is the Problem?

- It is always a question of whether to treat a variable as an ordinal variable or continuous variable in a regression model.
- Typically the question is asked as:
  - How many levels before its continuous?

What is the Problem?

- It is always a question of whether to treat a variable as an ordinal variable or continuous variable in a regression model.
- Typically the question is asked as:
  - How many levels before its continuous?
    - Have rules of thumb used in industry…
      "Anything fewer than 10 is probably ordinal…"
      "Anything more than 20 is probably continuous…"
      "Anything in between… well… it depends… ☺"

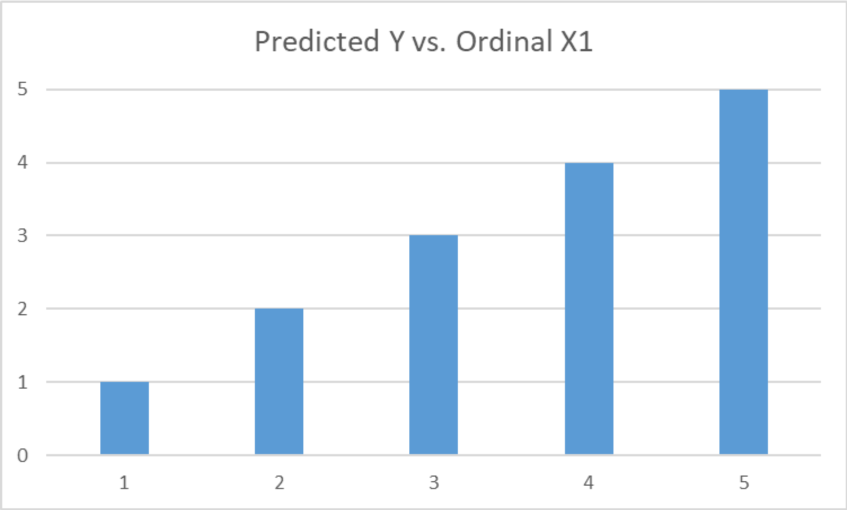
What is the Problem?

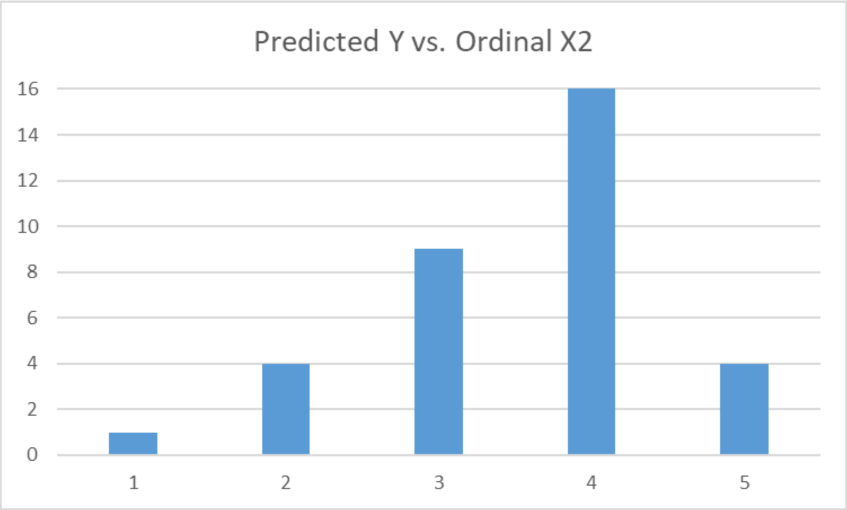
- It is always a question of whether to treat a variable as an ordinal variable or continuous variable in a regression model.
- Typically the question is asked as:
  - ~~How many levels before its continuous?~~

- This is actually the wrong question!

Continuous vs. Ordinal

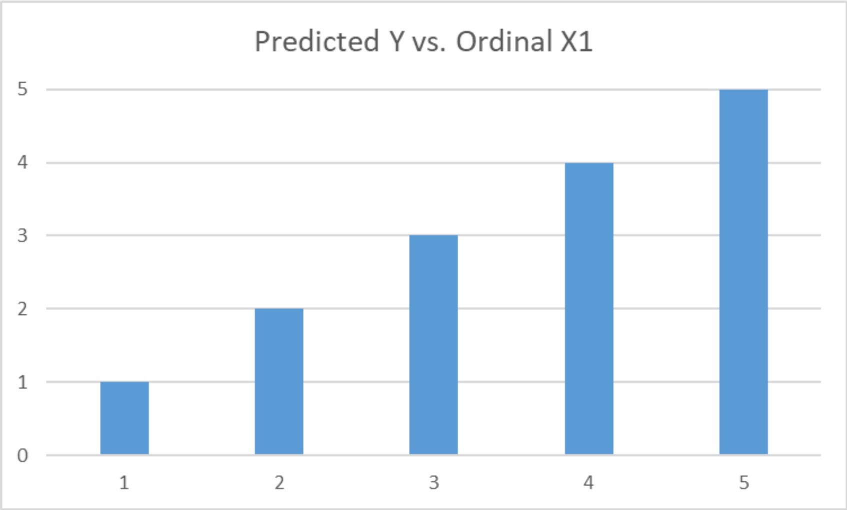
- Must understand the relationship between ordinal and continuous variables.
- Let's compare visually:

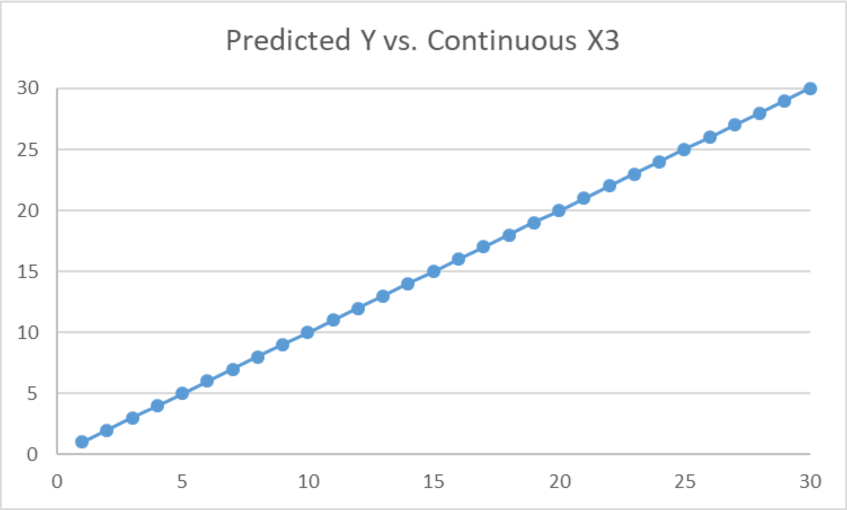




Continuous vs. Ordinal

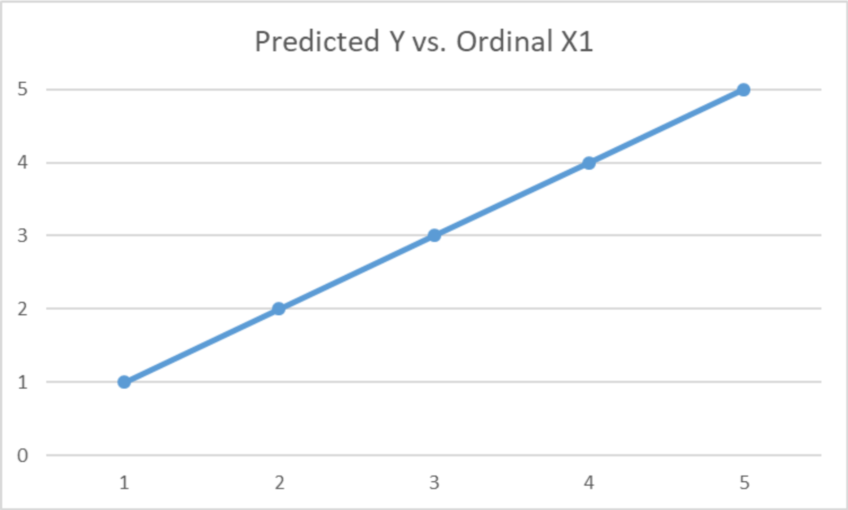
- Must understand the relationship between ordinal and continuous variables.
- Let's compare visually:

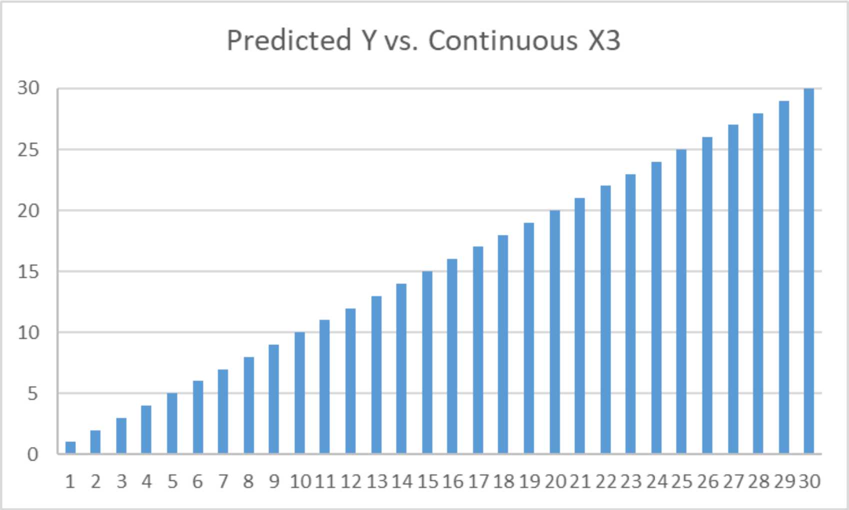




Continuous vs. Ordinal

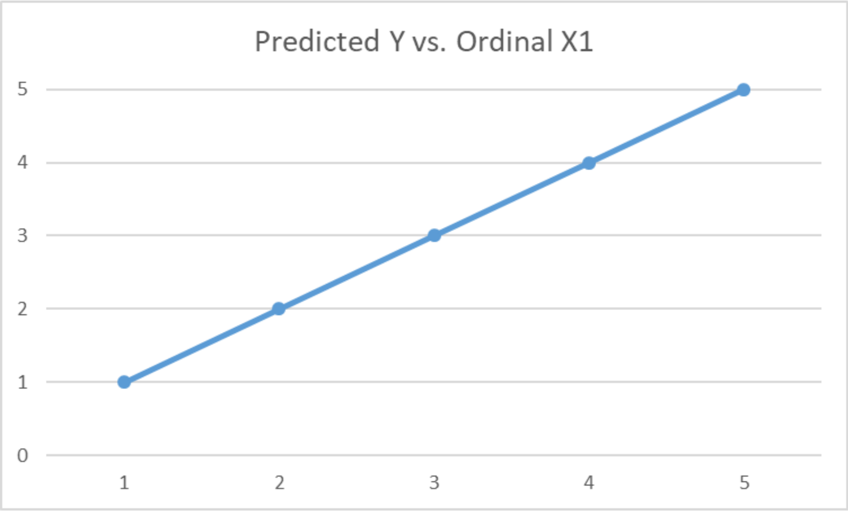
- Must understand the relationship between ordinal and continuous variables.
- Let's compare visually:

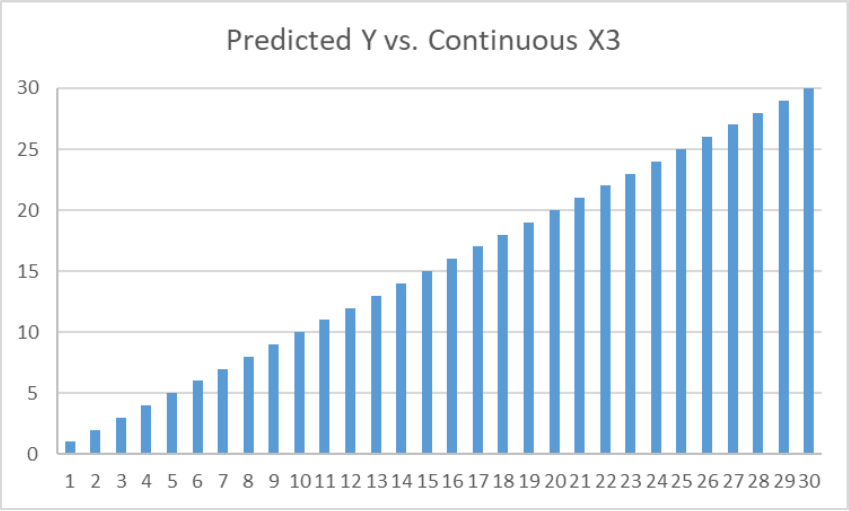




Continuous vs. Ordinal

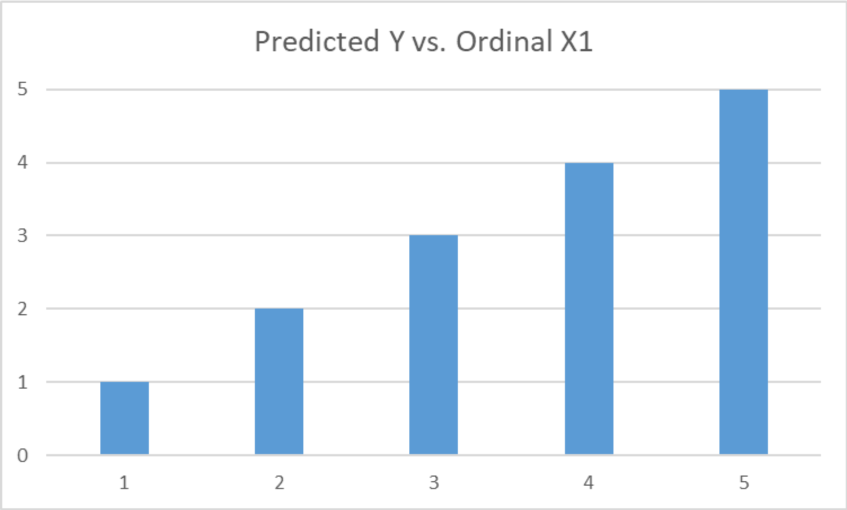
- What does the straight line imply?
- The jumps (in y) between points (values of x) are the same.

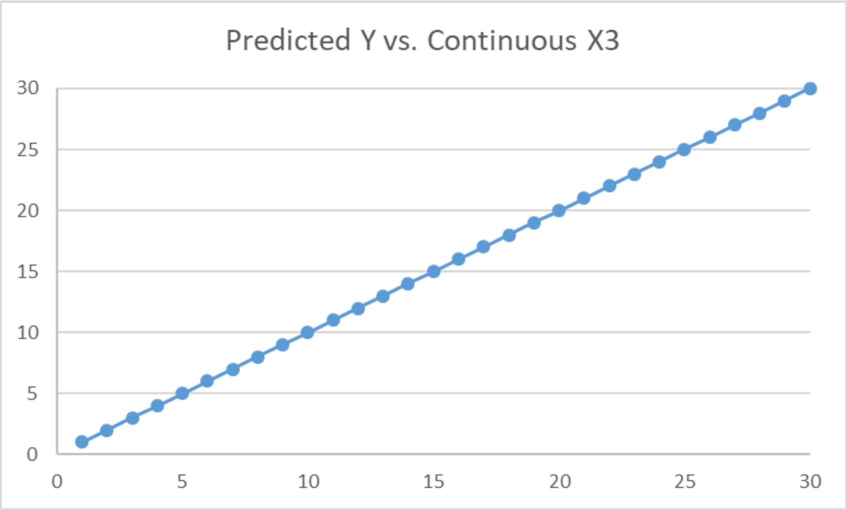




Continuous vs. Ordinal

- What does the straight line imply?
- The jumps (in y) between points (categories of x) are the same.





What is the Problem?

- It is always a question of whether to treat a variable as an ordinal variable or continuous variable in a regression model.
- Typically the question is asked as:
  - ~~How many levels before its continuous?~~
- Are you willing to assume the jumps (slope) between each value of x (category) are the same?

What is the Problem?

- Are you willing to assume the jumps (slope) between each value of x (category) are the same?
- More mathematically… are you willing to substitute equation 1 for equation 2?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \cdots + \hat{\beta}_{k-1} D_{k-1}$$

Design variables for categories of x_1

WHAT'S THE SOLUTION?

Comparison of Models

- Let the data decide if it should be modeled as continuous or categorical.
- Compare the following models:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \cdots + \hat{\beta}_{k-1} D_{k-1}$$

- The first model is a special case (nested within) of the second model!

How is it Nested?

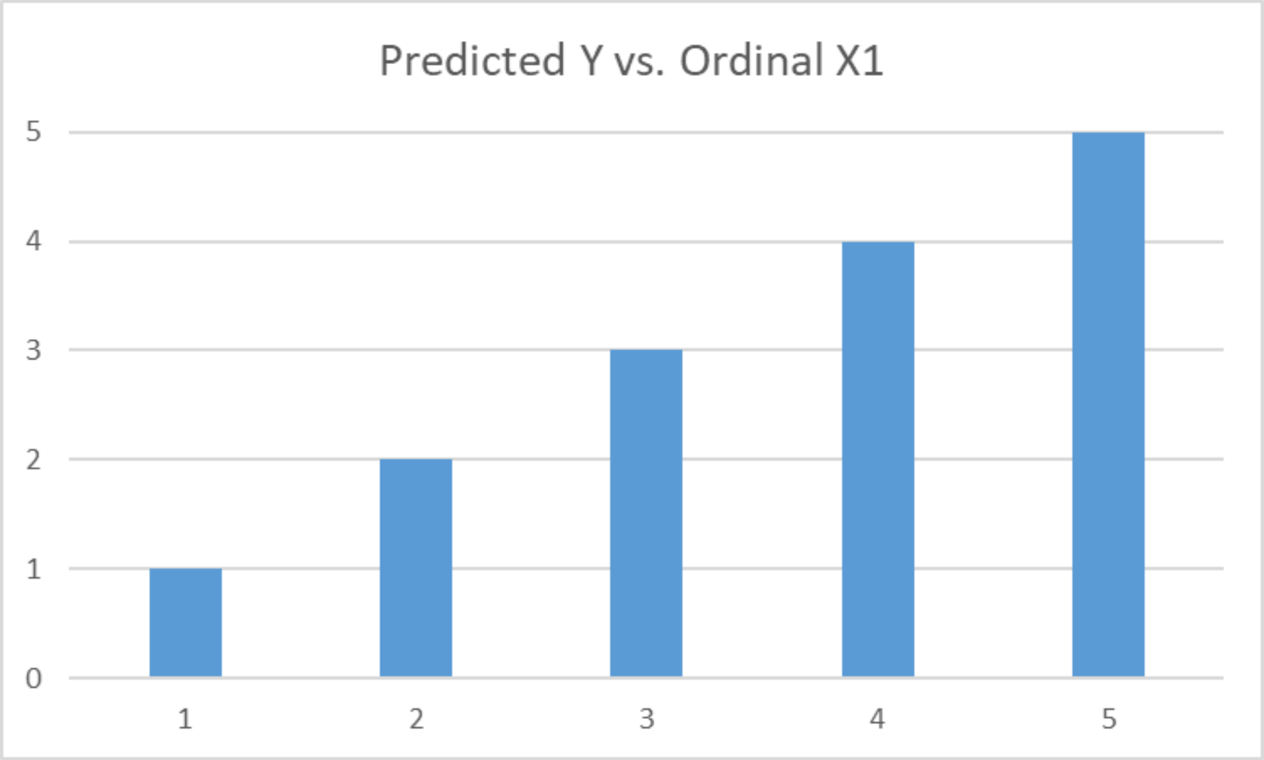
- Let's examine that second model a little closer to see how the first one could be a special case.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \cdots + \hat{\beta}_{k-1} D_{k-1}$$

How is it Nested?

- Let's examine that second model a little closer to see how the first one could be a special case.

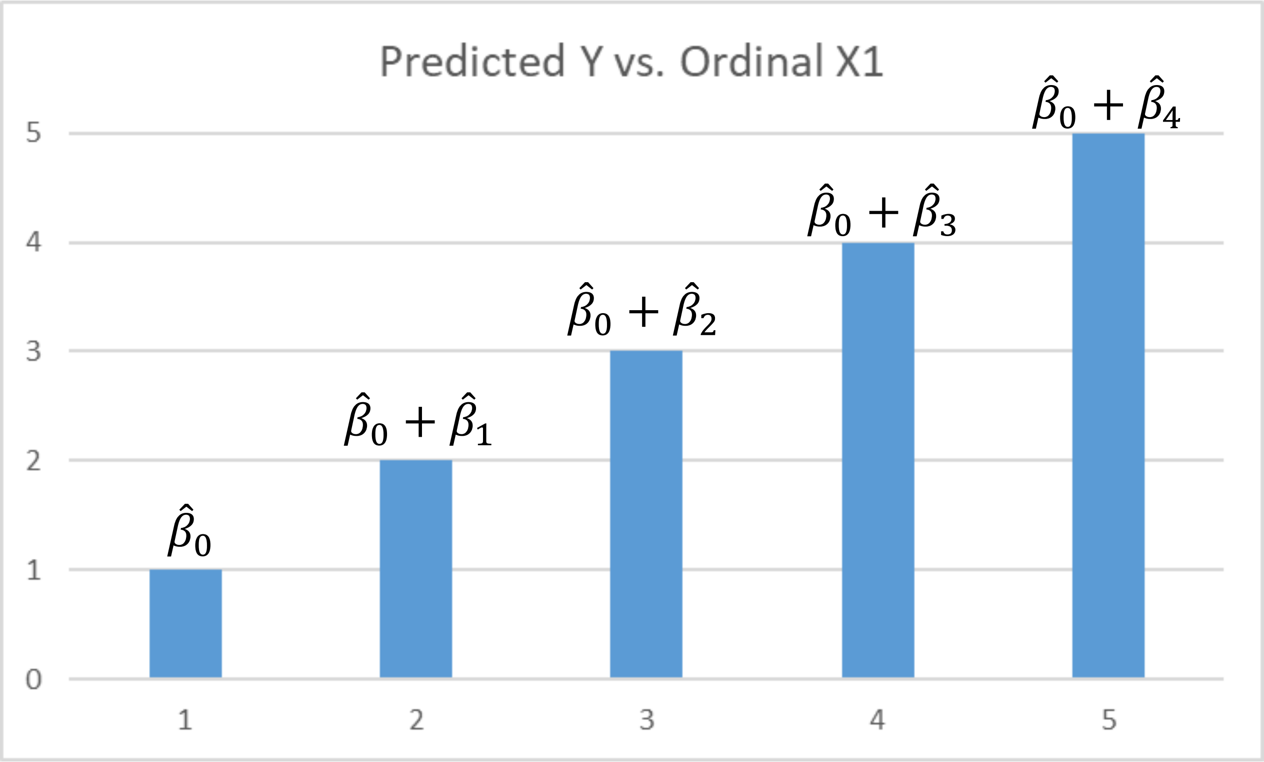
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 D_4$$



How is it Nested?

- Let's examine that second model a little closer to see how the first one could be a special case.

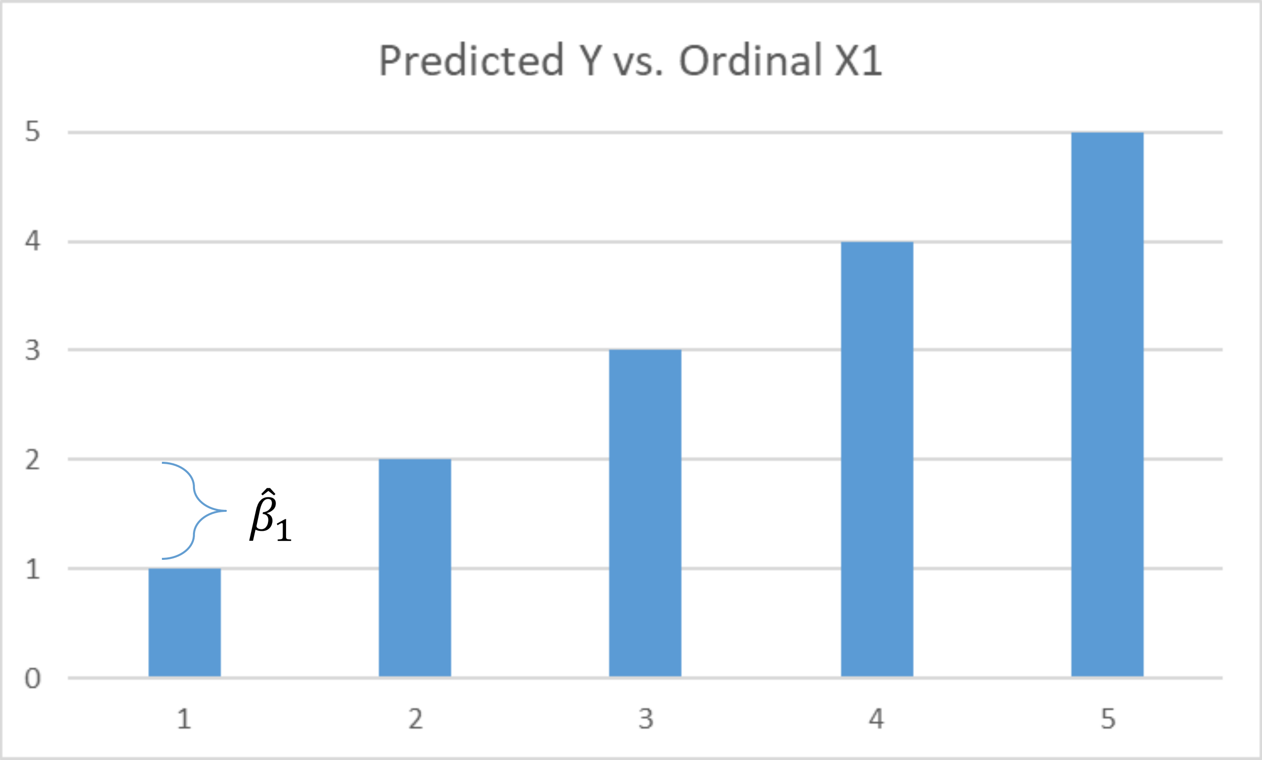
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 D_4$$



How is it Nested?

- Let's examine that second model a little closer to see how the first one could be a special case.

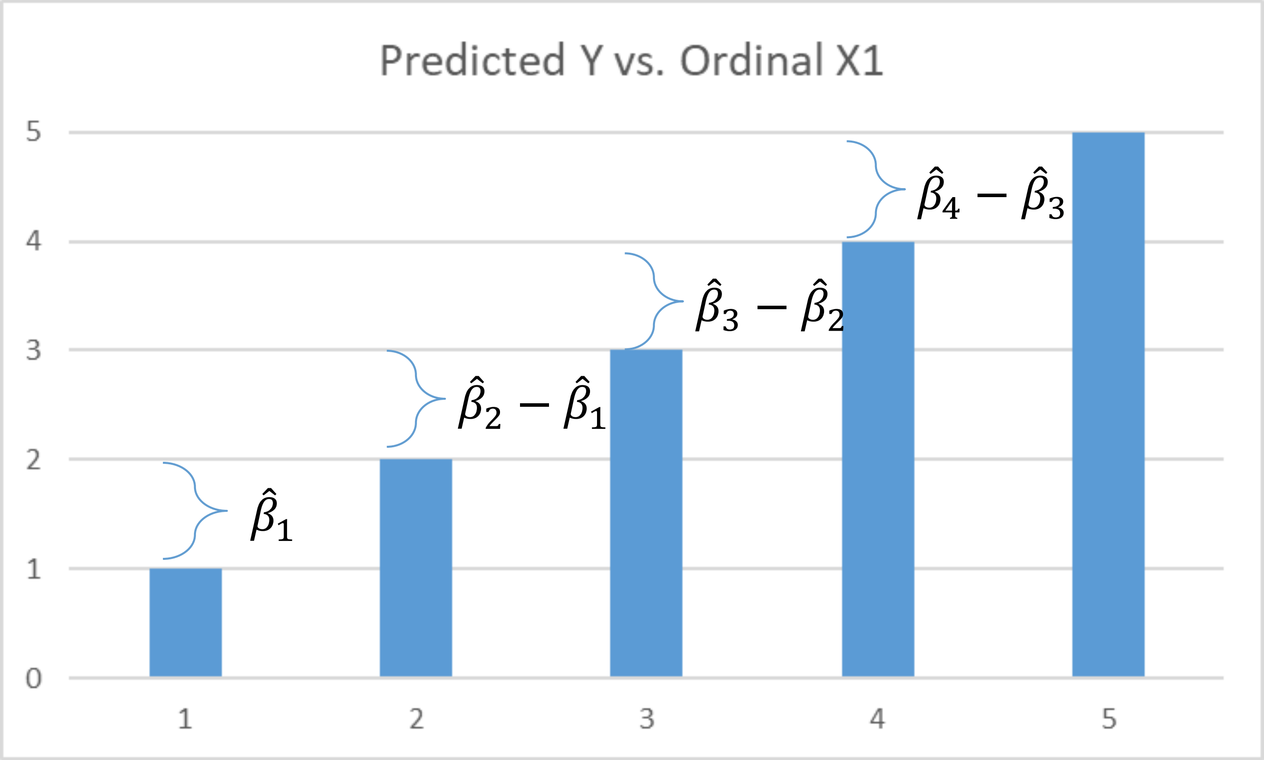
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 D_4$$



How is it Nested?

- Let's examine that second model a little closer to see how the first one could be a special case.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 D_4$$



How is it Nested?

- Let's examine that second model a little closer to see how the first one could be a special case.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 D_4$$

- What if the following were true?

$$\hat{\beta}_1 = \hat{\beta}_2 - \hat{\beta}_1 = \hat{\beta}_3 - \hat{\beta}_2 = \hat{\beta}_4 - \hat{\beta}_3$$

How is it Nested?

- Let's examine that second model a little closer to see how the first one could be a special case.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 D_4$$

- What if the following were true?

$$\hat{\beta}_1 = \hat{\beta}_2 - \hat{\beta}_1 = \hat{\beta}_3 - \hat{\beta}_2 = \hat{\beta}_4 - \hat{\beta}_3$$

SAME AS

$$\hat{\beta}_2 = 2\hat{\beta}_1 \qquad \hat{\beta}_3 = 3\hat{\beta}_1 \qquad \hat{\beta}_4 = 4\hat{\beta}_1$$

How is it Nested?

- Let's examine that second model a little closer to see how the first one could be a special case.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 D_4$$

- What if the following were true?

$$\hat{\beta}_1 = \hat{\beta}_2 - \hat{\beta}_1 = \hat{\beta}_3 - \hat{\beta}_2 = \hat{\beta}_4 - \hat{\beta}_3$$

SAME AS

$$\hat{\beta}_2 = 2\hat{\beta}_1 \quad \hat{\beta}_3 = 3\hat{\beta}_1 \quad \hat{\beta}_4 = 4\hat{\beta}_1$$

Only need to really estimate $\hat{\beta}_1$!!!

Comparison of Models

- Let the data decide if it should be modeled as continuous or categorical.
- Compare the following models:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \cdots + \hat{\beta}_{k-1} D_{k-1}$$

- The first model is a special case (nested within) of the second model!

Statistical Tests for Nested Models

- Many different statistical tests can compare two models if they are nested.
- Linear regression primarily uses the Nested F Test.
- Logistic regression primarily uses the Likelihood Ratio Test (LRT).

HOW TO IMPLEMENT SOLUTION?

Statistical Tests for Nested Models

- Many different statistical tests can compare two models if they are nested.
- Linear regression primarily uses the Nested F Test.
- Logistic regression primarily uses the Likelihood Ratio Test (LRT).

Statistical Tests for Nested Models

- Many different statistical tests can compare two models if they are nested.
- Linear regression primarily uses the Nested F Test.
- Logistic regression primarily uses the Likelihood Ratio Test (LRT).

$$H_0\text{: Models are the same}$$

OR

$$H_0: \hat{\beta}_1 = \hat{\beta}_2 - \hat{\beta}_1 = \hat{\beta}_3 - \hat{\beta}_2 = \hat{\beta}_4 - \hat{\beta}_3$$

Statistical Tests for Nested Models

- Many different statistical tests can compare two models if they are nested.
- Linear regression primarily uses the Nested F Test.
- Logistic regression primarily uses the Likelihood Ratio Test (LRT).

$$H_0\text{: Models are the same}$$

- If models are the same USE THE SIMPLER MODEL (aka treat the variable as continuous)!

Likelihood Ratio Test Example – R

```
logit.model.c <- glm(INS ~ CCPURC,
                     data = ins_t,
                     family = binomial(link = "logit"))
logit.model.o <- glm(INS ~ factor(CCPURC),
                     data = ins_t,
                     family = binomial(link = "logit"))

anova(logit.model.o, logit.model.c, test = 'LRT')
```

Likelihood Ratio Test Example – R

```
## Analysis of Deviance Table
##
## Model 1: INS ~ factor(CCPURC)
## Model 2: INS ~ CCPURC
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7415     9611.7
## 2      7418     9628.2 -3  -16.514 0.0008896 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What is the Problem? What is the Solution? **How to Implement Solution?**