REVIEW OF LOGISTIC REGRESSION

Dr. Aric LaBarr

Institute for Advanced Analytics

MATH REVIEW

Odds vs. Probability

- **Odds** is the ratio of events to non-events:

$$Odds = \frac{\#yes}{\#no}$$

- **Probability** is the ratio of event to the total number of outcomes:

$$p = \frac{\#yes}{\#yes + \ \#no}$$

- **Odds** and **Probability** are related:

$$Odds = \frac{p}{1 - p} \qquad p = \frac{Odds}{1 + Odds}$$

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	**84**
Checking	291	125	**416**
Total	**321**	**179**	**500**

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	84
Checking	291	125	416
Total	321	179	500

Probability of **NO BUY** in **Checking** account customers $= \frac{291}{416} = 0.70$

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	**84**
Checking	291	125	**416**
Total	**321**	**179**	**500**

Probability of **BUY** in **Checking** account customers $= \frac{125}{416} = 0.30$

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	84
Checking	291	125	416
Total	321	179	500

Odds of **BUY** in **Checking** account customers

$$= \frac{\text{Prob. of Buy}}{\text{Prob. of No Buy}} = \frac{0.30}{0.70} = 0.43$$

Odds Ratio

- **Odds Ratio** indicates how likely (in terms of odds) an event is for one group relative to another:

$$OR = \frac{Odds_A}{Odds_B}$$

- Since odds are always non-negative, so are odds ratios
  - OR > 1 → Event **more likely for A than for B**
  - OR < 1 → Event **more likely for B than for A**
  - OR = 1 → Event **equally likely in each group**

Probability versus Odds of an Outcome

	No Buy	Buy	Total
No Checking	30	54	84
Checking	291	125	416
Total	321	179	500

Odds of BUY in **No Checking** $= 1.77$

Odds of BUY in **Checking** $= 0.43$

Odds Ratio: No Checking to Checking $= \frac{1.77}{0.43} = 4.12$

Odds Ratio

Odds of **BUY** in **No Checking** $= 1.77$

Odds of **BUY** in **Checking** $= 0.43$

Odds Ratio: No Checking to Checking $= \dfrac{1.77}{0.43} = 4.12$

Non-Checking account customers have **4.12 times the odds** of buying the insurance product as compared to checking account customers.

Relative Risk

- **Relative Risk** indicates how likely (in terms of probability) an event is for one group relative to another:

$$RR = \frac{p_A}{p_B}$$

- Since probabilites are always non-negative, so are relative risks
  - RR > 1 → Event **more likely for A than for B**
  - RR < 1 → Event **more likely for B than for A**
  - RR = 1 → Event **equally likely in each group**

Math for Logistic Regression

- The following are rules involving the exponential function and natural logarithm:
  - $e^a > 0$ for any number *a*
  - $e^{a+b} = e^a e^b$, and $e^{a-b} = \frac{e^a}{e^b}$
  - $\log(a)$ can be any number, but $a > 0$
    - $\log(a) = -\infty$ if $a = 0$
    - $\log(a)$ **does not exist** if $a < 0$
  - $\log(a \times b) = \log(a) + \log(b)$, and $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$
  - $\log(e^a) = a$, and $e^{\log(a)} = a$
  - $a^{-1} = \frac{1}{a}$

BINARY LOGISTIC REGRESSION REVIEW

Assumptions for OLS Regression

- The random error term has a Normal distribution with a mean of zero.
- The random error term has constant variance.
- The error terms are independent.
- Linearity of the mean.
- No perfect collinearity.

Why Not Least Squares Regression?

$$y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i$$

- If the response variable is categorical, then how do you code the response numerically?
- If the response is coded (1=Yes and 0=No) and your regression equation predicts 0.5 or 1.1 or -0.4, what does that mean practically?
- If there are only two (or a few) possible response levels, is it reasonable to assume constant variance and normality?

Linear Probability Model

$$p_i = \beta_0 + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i}$$

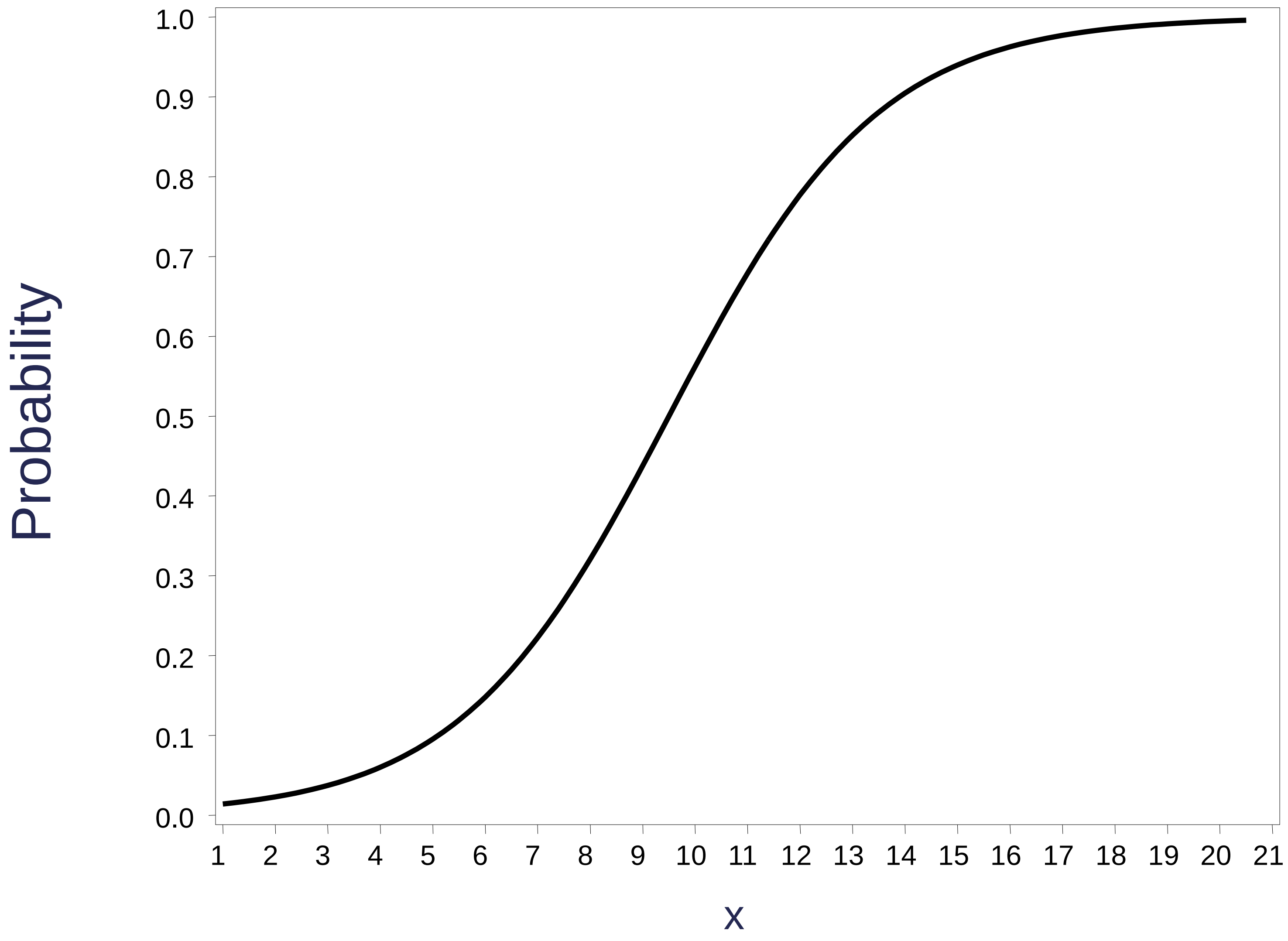
- Problems:
  - Probabilities are bounded, but linear functions can take on any value. (How do you interpret a predicted value of -0.4 or 1.1?)
  - The relationship between probabilities and X is usually nonlinear. Example, one unit change in X will have different effects when the probability is near 1 or 0.5.

Logistic Regression Model

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i})}}$$

- Has desired properties:
  - The predicted probability will always be between 0 and 1.
  - The parameter estimates do not enter the model equation linearly.
  - The rate of change of the probability varies as the X’s vary.

Logistic Regression Curve

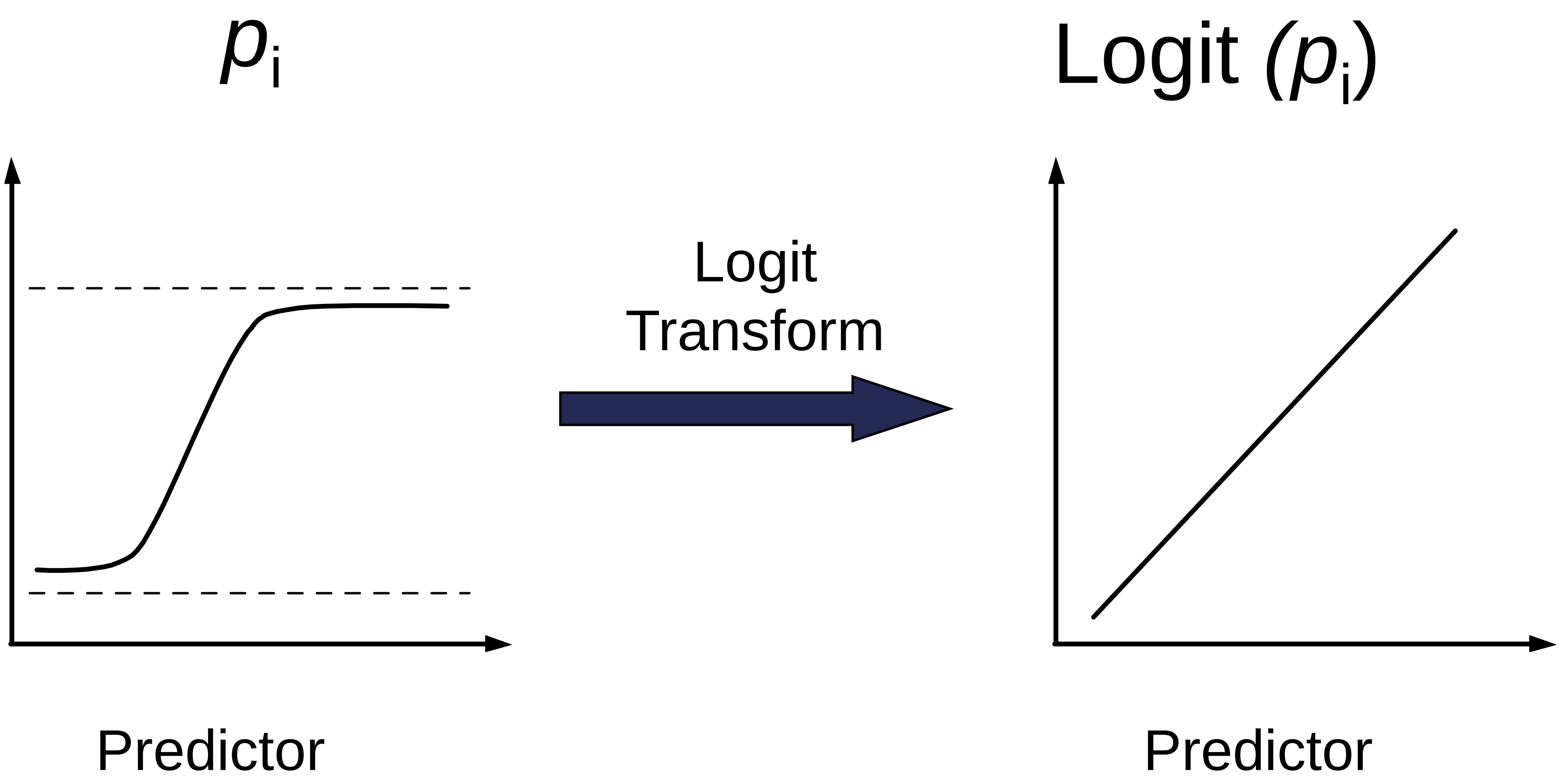


The Logit Link Transformation

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots \beta_k x_{k,i}$$

- To create a linear model, a link function (logit) is applied to the probabilities.
- The relationship between the parameters and the logits are linear.
- Logits unbounded.

The Logit Link Transformation



CATEGORICAL INPUTS

Reference Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Reference coding** is a common way to code categorical variables.
- 2 Category Example (A, B):

$$x = \begin{cases} 1 & \text{if A} \\ 0 & \text{if B} \end{cases}$$

- 3 Category Example (A, B, C):

	x_1	x_2
A	1	0
B	0	1
C	0	0

Reference Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Reference** coding is a common way to code categorical variables.
- 3 Category Example (A, B, C):

	x_1	x_2
A	1	0
B	0	1
C	0	0

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category A and C.

Reference Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Reference coding** is a common way to code categorical variables.
- 3 Category Example (A, B, C):

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$
A	1	0
B	0	1
C	0	0

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category B and C.

Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- 2 Category Example (A, B):

$$x = \begin{cases} 1 & \text{if A} \\ -1 & \text{if B} \end{cases}$$

- 3 Category Example (A, B, C):

	x_1	x_2
A	1	0
B	0	1
C	**-1**	**-1**

Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category A and the overall average of categories **A, B, & C**.

	x_1	x_2
A	1	0
B	0	1
C	**-1**	**-1**

Effects Coding

- Categorical variables need to be coded differently because they are not numerical in nature.
- **Effects coding** is another common way to code categorical variables.
- 3 Category Example (A, B, C):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Average difference between category B and the overall average of categories **A, B, & C**.

	x_1	x_2
A	1	0
B	0	1
C	**-1**	**-1**