STAT 131A

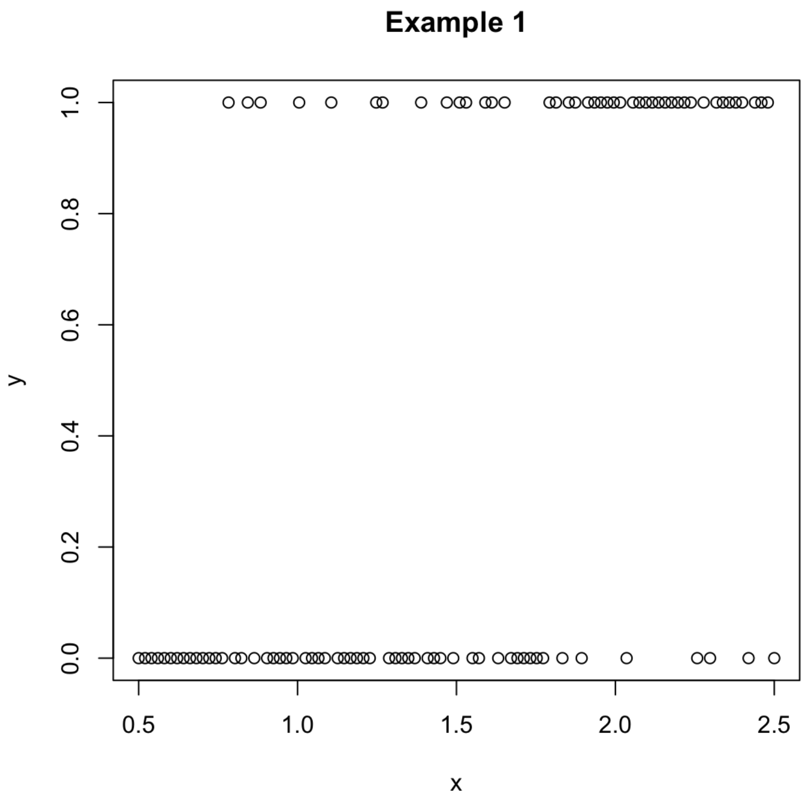
LECTURE 21: LOGISTIC REGRESSION I

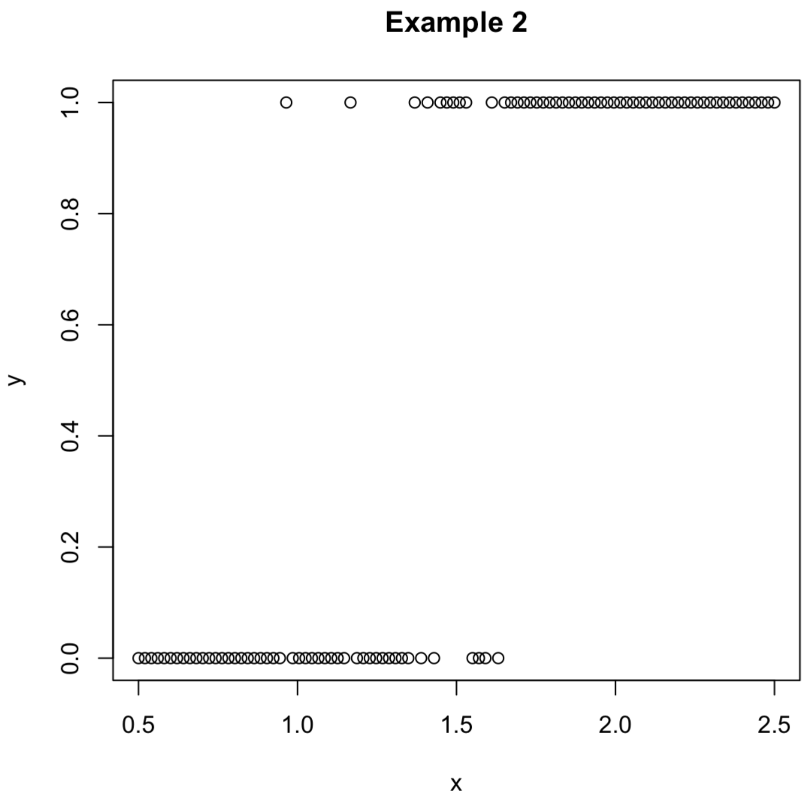
NOVEMBER 18, 2025

MOTIVATION:
WHEN THE **OUTCOME** IS **YES / NO**

- Many questions in data science have binary outcomes: yes/no, 0/1, success/failure.
- For instance, Did a patient develop a disease? Did a user click an ad? Did a loan default?
- We want to **relate the probability of "yes" to one or more predictors**
- **Logistic regression**: the standard tool for modeling **binary outcomes**

LOGISTIC REGRESSION SETUP


Example 1


Example 2

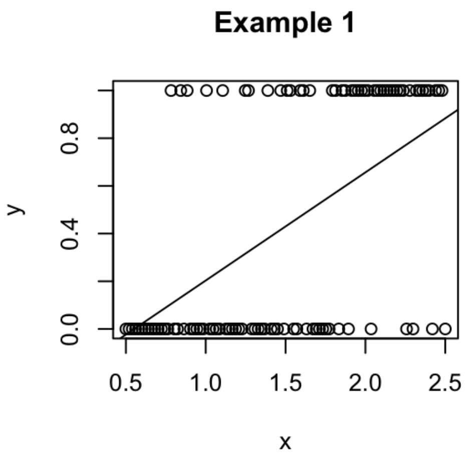
NOTE THAT …

- Logistic regression does **not directly try to predict values 0 or 1**
- Instead, tries to predict the *probability* that *y* is 1, as a function of its variables
  - $p(x) = P(Y = 1 \mid x)$

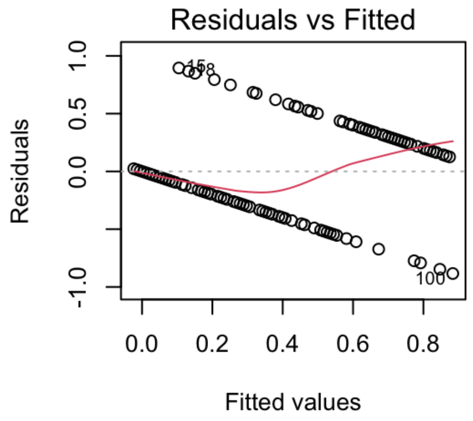
WHY NOT JUST USE LINEAR REGRESSION?

- If we regress a 0/1 outcome on x:
  - Predicted values can fall **below 0, or above 1** !
- Linear regression assumes constant variance; binary data has variance tied to the mean
  - $\operatorname{Var}(Y \mid X = x) = p(x)\left[1 - p(x)\right]$
- Errors are not normally distributed when the outcome is 0/1
- We need a model that
  - keeps predicted probabilities between 0 and 1
  - respects the structure of the data

WITH LINEAR FIT

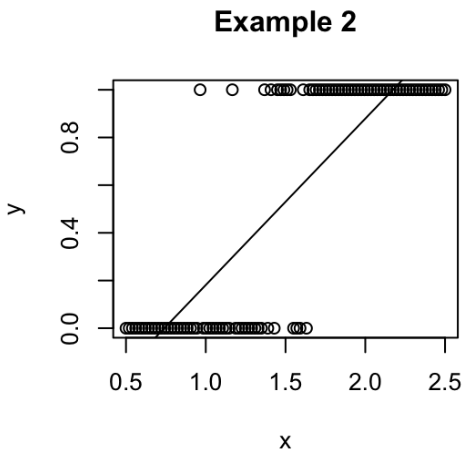


Example 1

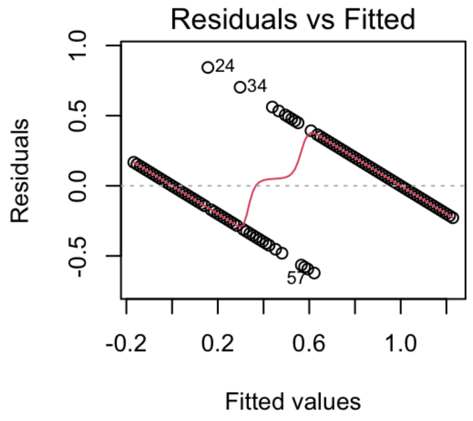


Residuals vs Fitted

- ✦This is problematic !
- ✦ Recall “heteroscedasticity”



Example 2



Residuals vs Fitted

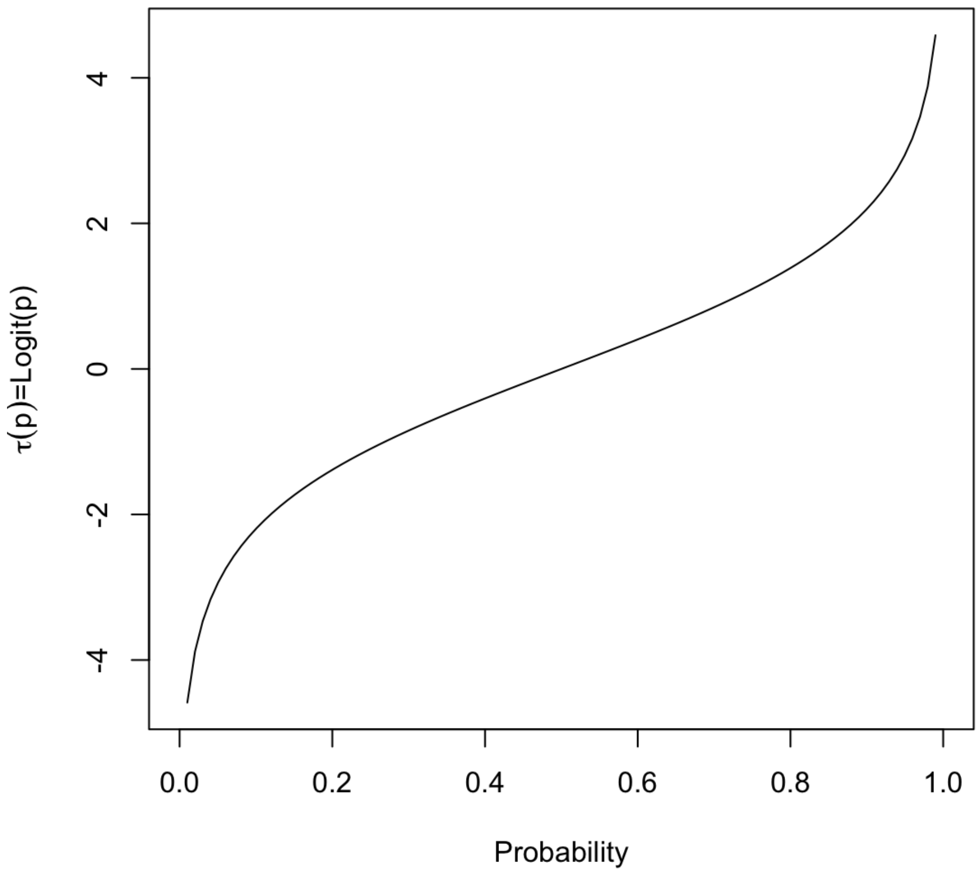
THE ODDS

- $odds(E) := \dfrac{P(E \text{ happens})}{P(E \text{ does not happen})} = \dfrac{P(E)}{1 - P(E)} = \dfrac{p}{1 - p}$
- $\text{odds} = \dfrac{p}{1 - p}$
- $p = \dfrac{\text{odds}}{1 + \text{odds}}$

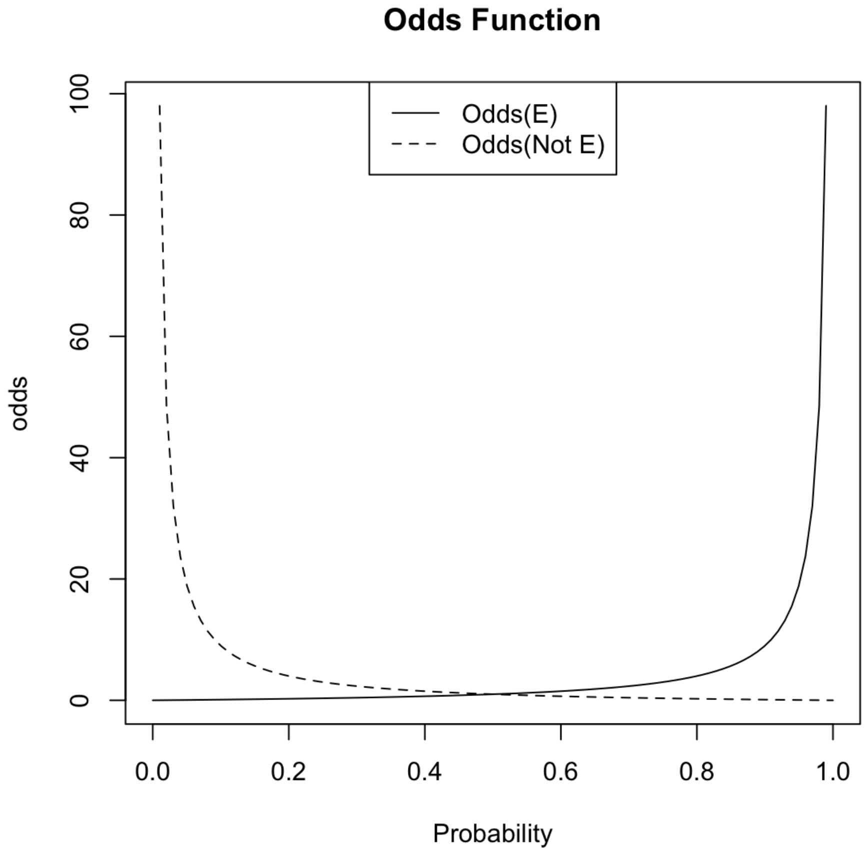
LOGIT FUNCTION & LOG-ODDS

- $z(x) = \tau(p(x))$
- $\hat{p}(x) = \tau^{-1}(\hat{z}(x))$.
- $\tau(p) = logit(p) = log(\frac{p}{1-p})$.

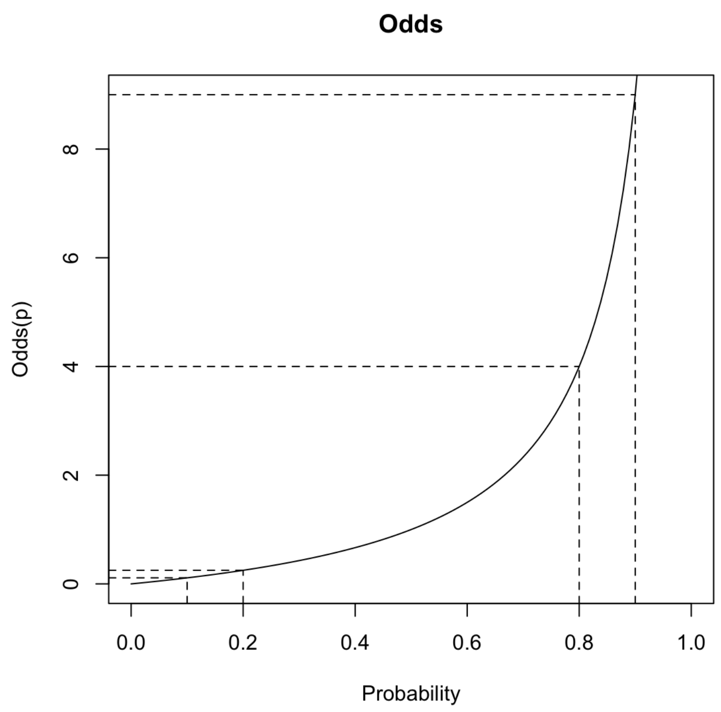
Logit Function

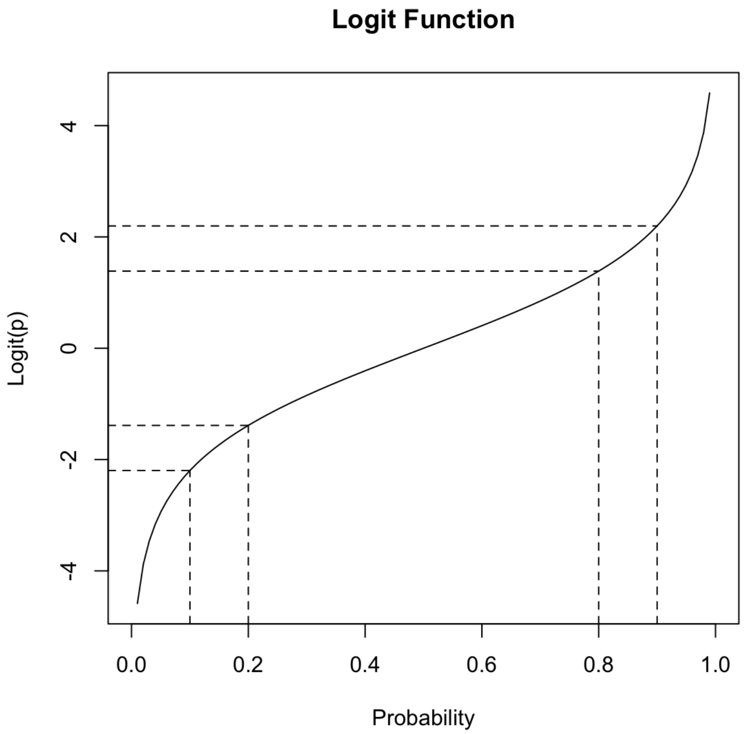


THE ODDS FUNCTION



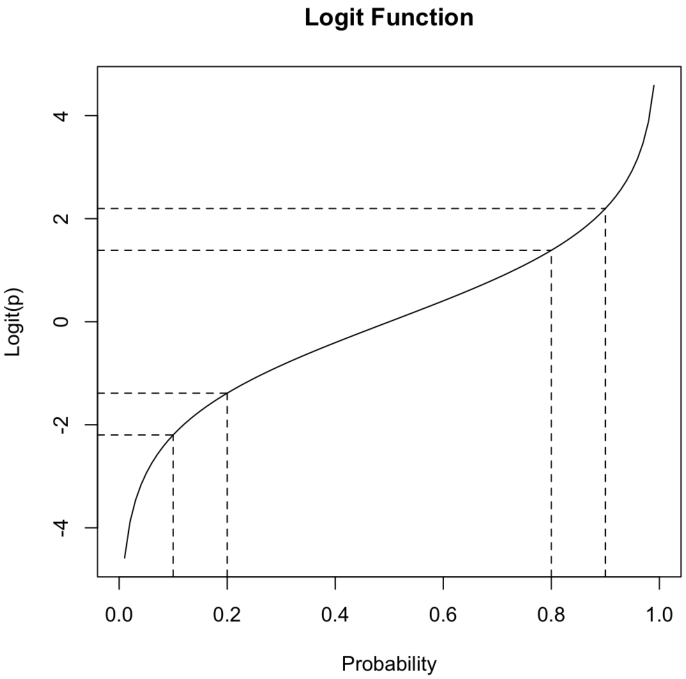
LOG ODDS

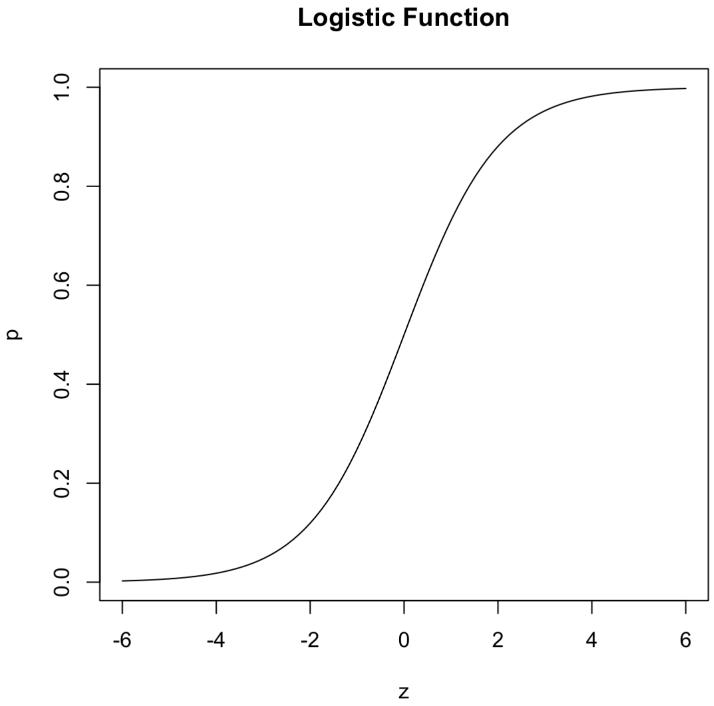




- **Log of odds**(E) : $log\left(\frac{p}{1-p}\right)$

THE LOGIT FUNCTION





- $z = logit(p)$
- $p = P(E) = \tau^{-1}(z) = \dfrac{e^{z}}{1+e^{z}} = \dfrac{1}{1+e^{-z}}\,.$

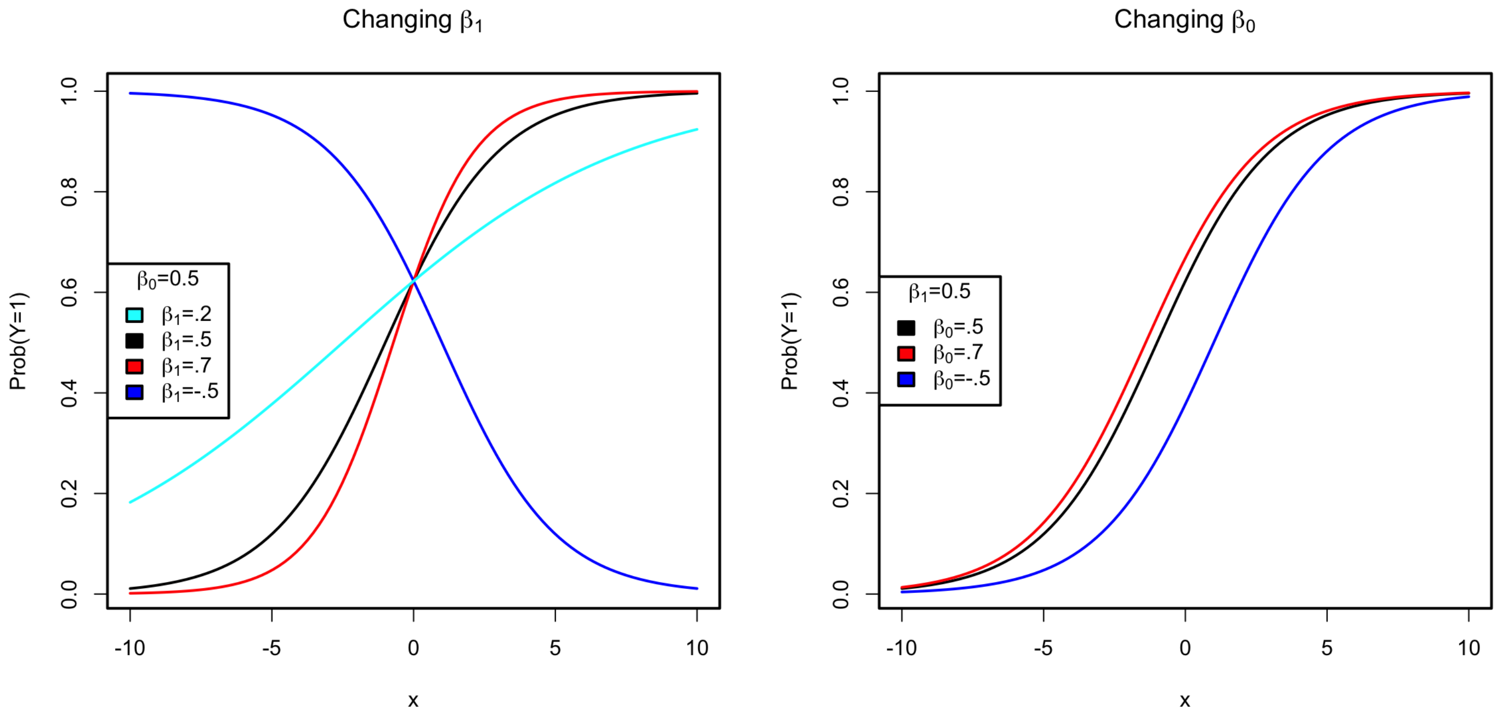
THE LOGISTIC REGRESSION MODEL

- $\log(\frac{p}{1-p}) = \log\big(odds(y=1)\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\,.$
- $p(x) = \dfrac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$

- Let Y be a binary outcome (0/1) and X a predictor
- We model p(x) = P(Y = 1 I X = x) : the S-shaped curve
- Logistic model: **guarantees 0 < p(x) < 1 for all x**

VISUALIZING THE LOGISTIC REGRESSION MODEL (FOR ONE VARIABLE)

- $log(\frac{p_i}{1-p_i}) = \log(odds(y_i = 1)) = \beta_0 + \beta_1 x_i$
- $p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$
-



INTERPRETING COEFFICIENTS VIA ODDS RATIOS

- In $\operatorname{logit}(p) = \beta_0 + \beta_1 x$, a one-unit increase in x changes log-odds by β_1
- Exponentiating: $e^{(\beta 1)}$ is the multiplicative change in odds for a 1-unit increase in x.
- If exp(β1) = 1.2, odds increase by 20% for each unit increase in x (holding other variables fixed).
- If exp(β1) = 0.7, odds decrease by 30% per unit increase in x.

INTERPRETING
THE COEFFICIENTS

MULTIPLE LOGISTIC REGRESSION

- We often have several predictors: age, sugar intake, ethinicity, SES, … etc.
- Binary targets/outcomes:
  - At risk for heart-disease, or hypertension or hospital-readmission or ..
  - All above are Y/N
- Model: $\mathrm{logit}\left(P\left(Y=1\,\middle|\,x\right)\right)=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p$
- Each coefficient describes the association
  - between its predictor and the log-odds of Y = 1
  - holding all other predictors constant

INTERPRETING THE INTERCEPT

- The intercept β_0 corresponds to log-odds of Y = 1 , when all predictors are at their **reference values**
- Exponentiating: $e^{(\beta 0)}$ gives the **baseline odds** of the event
- Often less interesting than slope coefficients but important conceptually
- In a logistic model with categorical predictors, choose baseline categories

INTERPRETING SLOPES

- Suppose logit(p) = $\beta_0 + \beta_1$ age + ...
- Then for a 1-year increase in age, log-odds of the outcome change by β_1
  - log-odds of the outcome = (say) risk-of-diabetes
- Essentially, the odds **are multiplied by** $e^{(\beta 1)}$

FOR A **BINARY** PREDICTOR

- Suppose smoker is coded 1 = smoker, 0 = non-smoker.
- Then β_{smoker} is the difference in log-odds of the outcome (risk of diabetes etc.) between smokers and non-smokers
- We often describe this as "multiplicative change in odds" comparing groups.

FITTING LOGISTIC REGRESSION

- Parameters β **are estimated by maximum likelihood**
  - Not by minimizing least squared errors !
- We choose β to make the observed 0/1 outcomes most probable under the model.
- We will leave it to software (R) for this optimization
  - Our focus is on the interpretation