Logistic Regression Model

Logistic Regression I

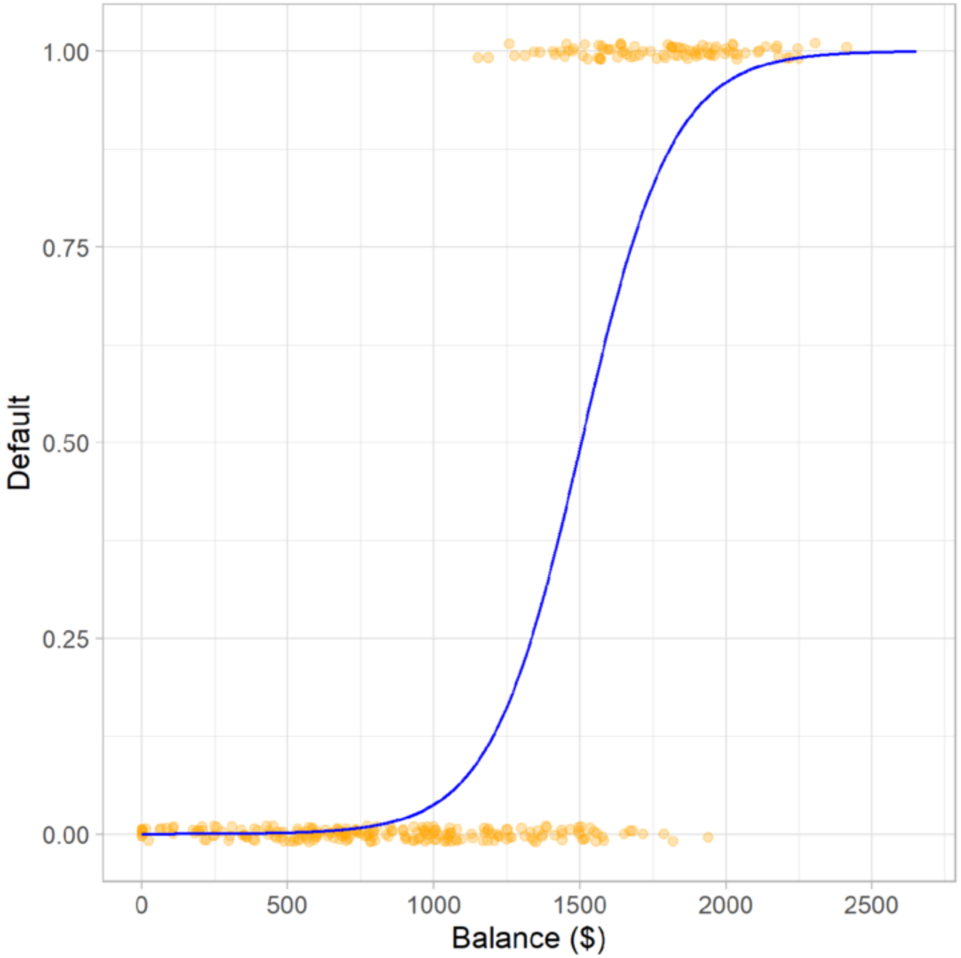
Learning Objectives

1. Understand Logistic Regression formulation.
2. Understand odds and log odds.
3. Understand Logistic Regression is predicting log odds.
4. Understand that the Maximum Likelihood Estimation (MLE) method is used to estimate the parameters, $\beta_0, \beta_1, ..., \beta_n$, of logistic regression model.

Logistic Regression Model

- Logistic model: S-shaped curve representing probability that `Y=1` for a given predictor variable `X`.
- This probability, $p(X) = Pr(Y = 1|X)$, is given by:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

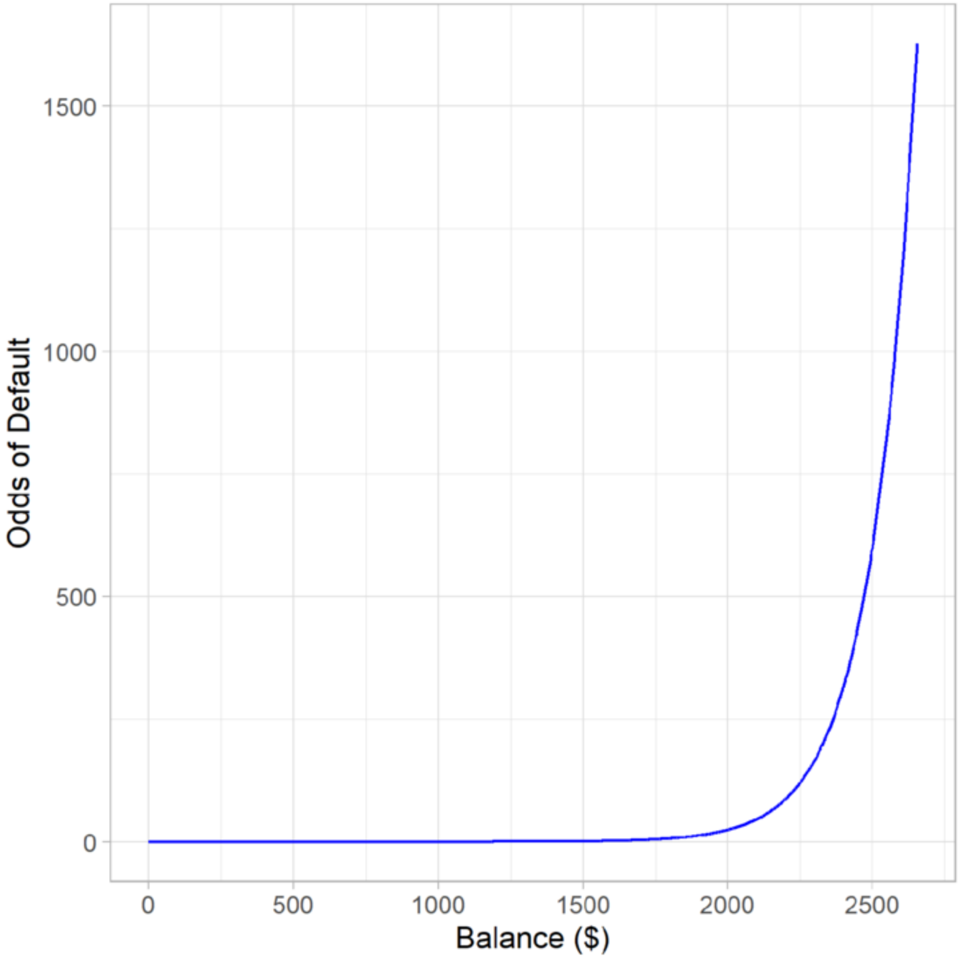


Odds

- Simplifying Eq.1,

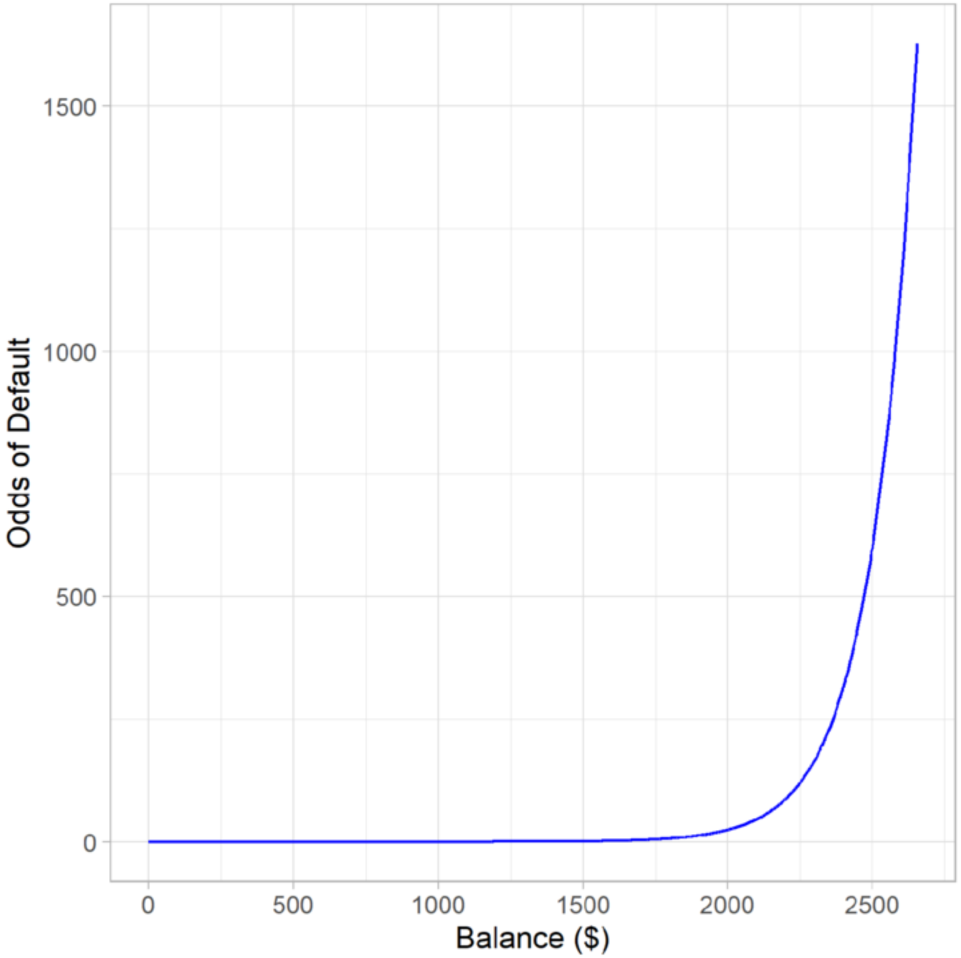
$$\frac{Pr(Y = 1|X)}{Pr(Y = 0|X)} = \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \qquad (2)$$

- In Eq.2, $p(X)/[1 - p(X)]$ represents the odds.



Odds

- Credit default: Odds of default, is defined as, the probability of default divided by the probability of no default.
    - Say, probability of default, $p(X) = 0.2 = 1/5$, i.e. 1 in 5 customers default.
    - Then the odds $= 0.2/(1 - 0.2) = 1/4$.
    - This means that default occurs once for every 4 customers who do not default, or, in 5 customers, we can expect 1 customer to default and 4 to not default.
- Odds can take values from 0 to ∞. Odds closer to 0 indicate very low probabilities of default. Odds closer to ∞ indicate very high probabilities of default.

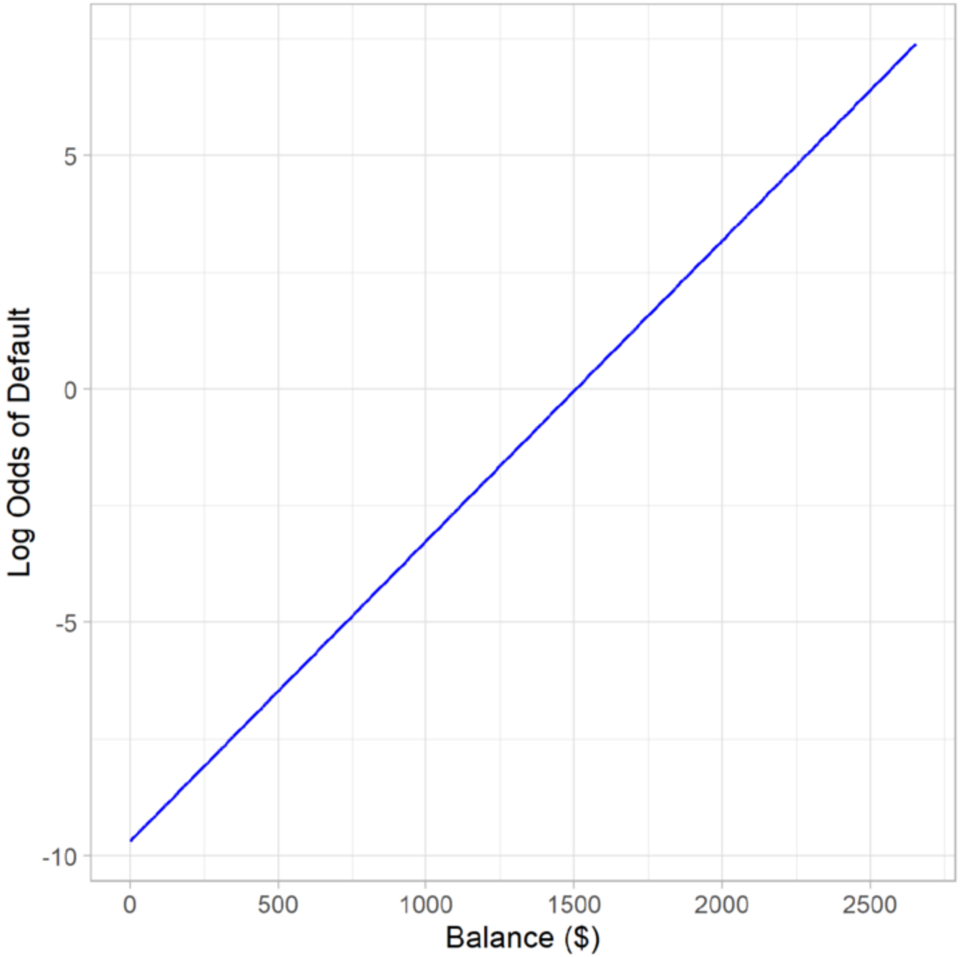


Log Odds

- Taking logarithms on the odds in Eq.2, we get the log odds or Logits,

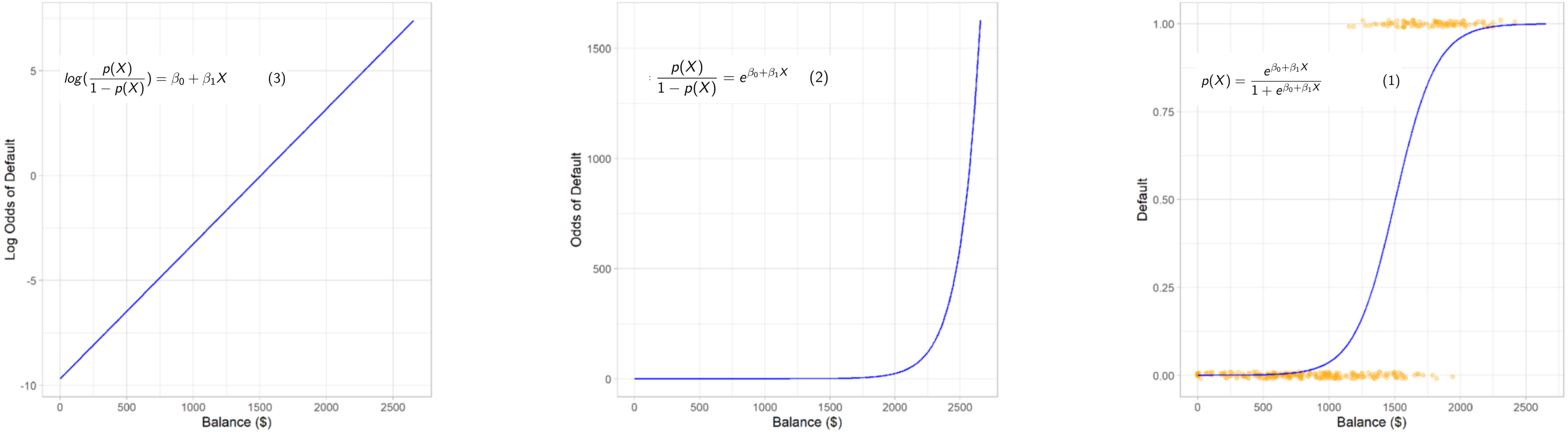
$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X \qquad (3)$$

- Logistic model is *linear* in `X` for the Logit.



Interpreting Logistic Regression Coefficients

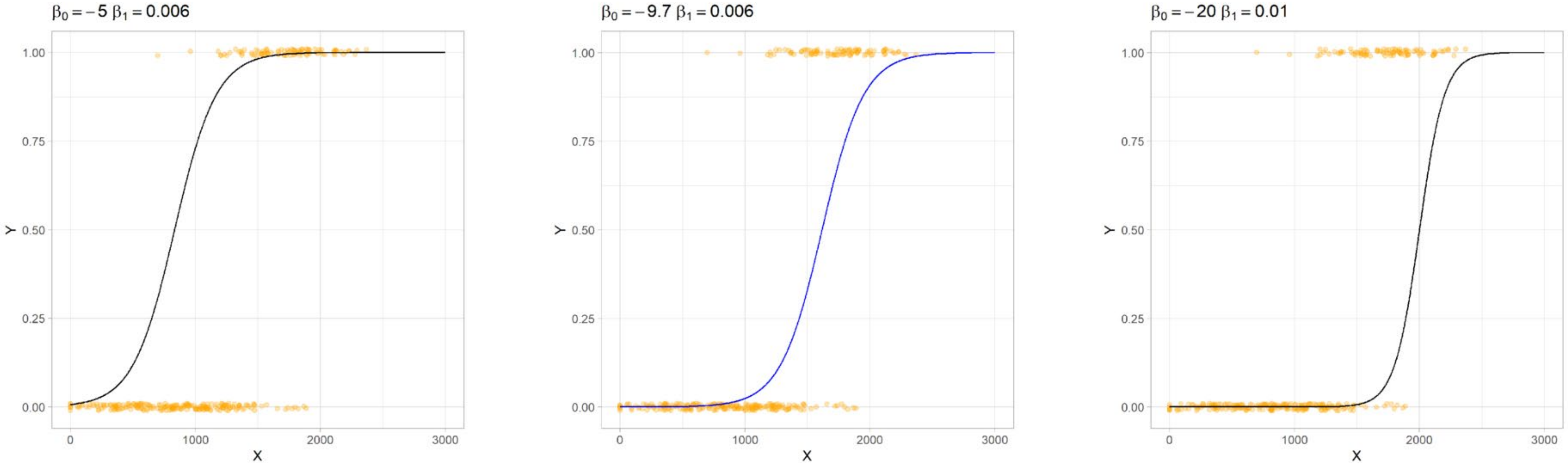
What does β_1 represent?



- In this model, $\beta_1 = 0.006$. Increasing X by 1 unit, increases the log odds by β_1, or 0.006.
- Increasing X by 1 unit, multiplies the odds by e^{β_1}, or $e^{0.006} = 1.006$. This is a 0.6% change in odds. So, change in odds depends on the value of X.

Estimating Logistic Regression Coefficients

What is the best fit S-shaped curve for this data?



Estimating Logistic Regression Coefficients

How is the best fit S-shaped curve estimated?

- Try to find β_0 and β_1 so that:
  - For those who default, $p(X)$ is close to 1, and
  - For those who do not default, $p(X)$ is close to 0
- Use likelihood function:

$$L(\beta_0, \beta_1) = \prod_{default=yes} p(X) \prod_{default=no} (1 - p(X)) \tag{4}$$

- Find β_0 and β_1 that maximises the likelihood function. The `glm()` function uses this maximum likelihood estimation method.

Multiple Logistic Regression

How can multiple predictors be included in the model?

- Adding more predictors, `balance`, `income` and `student` in the Logistic Regression model for the response variable, `default`.
- We can extend the Logistic Regression model for multiple predictor variables, X_1, X_2, ..., X_p:

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \quad (5)$$

- Increasing X_1 by 1 unit, holding other predictors fixed, on average:
  - changes the log odds by β_1, and
  - multiplies the odds by e^{β_1}
- Then, find β_0, β_1, ..., β_p that maximises the likelihood function.