Logistic Regression

Nate Wells

Math 243: Stat Learning

October 25th, 2021

Outline

In today's class, we will. . .

- Review classification problems
- Discuss Logistic Regression for Classification

Section 1

Logistic Regression

Classificaiton Problems

- Suppose Y is a categorical variable with levels $A_1, A_2, \ldots, A_k$.

Classificaiton Problems

- Suppose Y is a categorical variable with levels $A_1, A_2, \ldots, A_k$.
  - Example: Let Y indicate whether it is raining in Portland at noon on 10/25/21.
  - Levels: $A_1 = \text{Raining}$, $A_2 = \text{Not Raining}$.

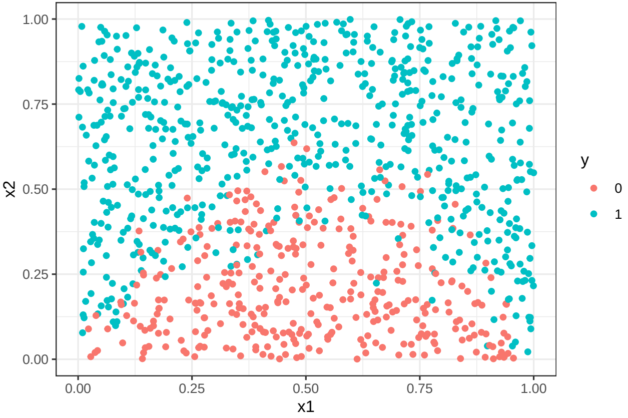
Classificaiton Problems

- Suppose Y is a categorical variable with levels $A_1, A_2, \ldots, A_k$.
    - Example: Let Y indicate whether it is raining in Portland at noon on 10/25/21.
    - Levels: $A_1 = \text{Raining}$, $A_2 = \text{Not Raining}$.
- Goal: Build a model f to classify an observation into levels $A_1, A_2, \ldots, A_k$ based on the values of several predictors $X_1, X_2, \ldots, X_p$ (quantitative or categorical)

$$\hat{Y} = f(X_1, X_2, \ldots, X_p) \qquad \text{where } f \text{ take values in } \{A_1, \ldots, A_k\}$$

Classification Regions

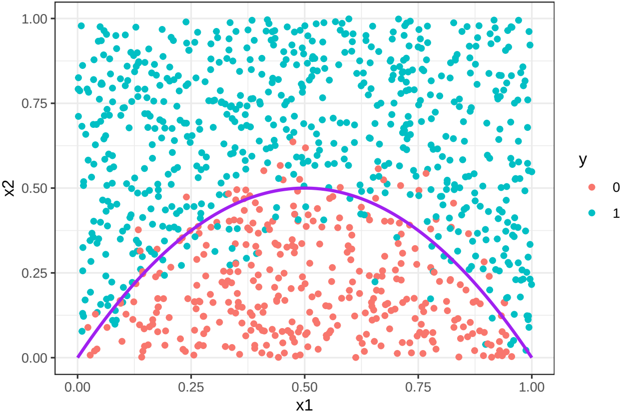
Any classification model will divide predictor space into unions of regions, where each point in a region will be classified in the same way.



Different models will have different geometries for classification boundaries.

Classification Regions

Any classification model will divide predictor space into unions of regions, where each point in a region will be classified in the same way.



The purple line indicates the optimal decision boundary.

The Bayes Classifier and KNN

- The Bayes classifier theoretically minimizes error rate

$$f(x_0) = \mathrm{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

The Bayes Classifier and KNN

- The Bayes classifier theoretically minimizes error rate

$$f(x_0) = \mathrm{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

  - In practice, these conditional probabilities are not known.

The Bayes Classifier and KNN

- The Bayes classifier theoretically minimizes error rate

$$f(x_0) = \mathrm{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

  - In practice, these conditional probabilities are not known.

- But we can approximate them using *KNN*:

$$P(Y = A_j \mid X = x_0) \approx \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

The Bayes Classifier and KNN

- The Bayes classifier theoretically minimizes error rate

$$f(x_0) = \mathrm{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

  - In practice, these conditional probabilities are not known.

- But we can approximate them using *KNN*:

$$P(Y = A_j \mid X = x_0) \approx \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

- Our model for P is therefore $\hat{P}_j(x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$.

The Bayes Classifier and KNN

- The Bayes classifier theoretically minimizes error rate

$$f(x_0) = \mathrm{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

  - In practice, these conditional probabilities are not known.
- But we can approximate them using *KNN*:

$$P(Y = A_j \mid X = x_0) \approx \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

- Our model for P is therefore $\hat{P}_j(x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$.
- And our classifier model is $\hat{g}(x_0) = \mathrm{argmax}_{A_j} \hat{P}_j(x_0)$

Why not always just use KNN?

1. KNN has very low training time (basically none), but often large test time (especially for large K)

Why not always just use KNN?

1. KNN has very low training time (basically none), but often large test time (especially for large K)
2. KNN models are hard to interpret, so often not ideal for inference questions.

Why not always just use KNN?

1. KNN has very low training time (basically none), but often large test time (especially for large K)
2. KNN models are hard to interpret, so often not ideal for inference questions.
3. If a linear or more structured model is more appropriate (i.e. accurately captures the true form of f), then KNN will be less stable.

Why not always just use KNN?

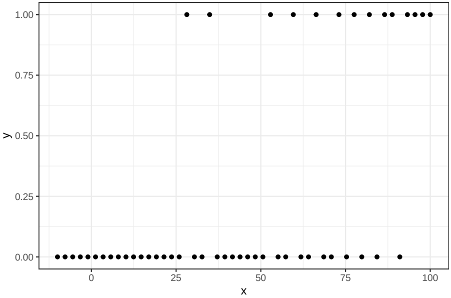
1. KNN has very low training time (basically none), but often large test time (especially for large K)
2. KNN models are hard to interpret, so often not ideal for inference questions.
3. If a linear or more structured model is more appropriate (i.e. accurately captures the true form of f), then KNN will be less stable.
4. KNN suffers from the "curse of dimensionality". For fixed K and large p, adding more predictors increases bias and variance.

Why not always just use KNN?

1. KNN has very low training time (basically none), but often large test time (especially for large K)
2. KNN models are hard to interpret, so often not ideal for inference questions.
3. If a linear or more structured model is more appropriate (i.e. accurately captures the true form of f), then KNN will be less stable.
4. KNN suffers from the "curse of dimensionality". For fixed K and large p, adding more predictors increases bias and variance.
5. KNN requires large sample sizes (compared to alternatives)

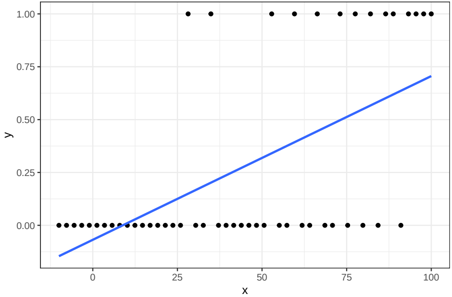
Alternatives

- Suppose Y is a binary categorical variable with a single quantitative predictor X. We want to model $p(X) = P(Y = 1|X)$



Alternatives

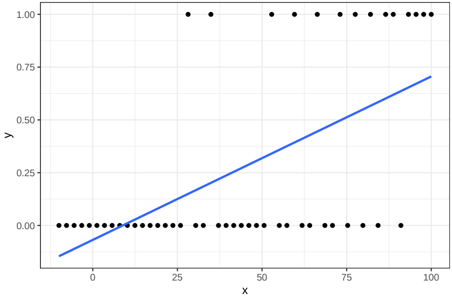
- Suppose Y is a binary categorical variable with a single quantitative predictor X. We want to model $p(X) = P(Y = 1|X)$



- Linear model: $p(X) = \beta_0 + \beta_1 X = -0.07 + 0.008X$

Alternatives

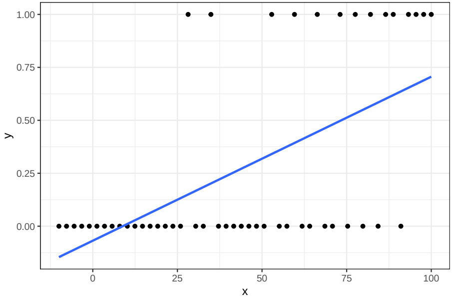
- Suppose Y is a binary categorical variable with a single quantitative predictor X. We want to model $p(X) = P(Y = 1|X)$



- Linear model: $p(X) = \beta_0 + \beta_1 X = -0.07 + 0.008X$
- Predict 1 if $\hat{P}(x) \geq 0.5$, and 0 otherwise.

Alternatives

- Suppose Y is a binary categorical variable with a single quantitative predictor X. We want to model $p(X) = P(Y = 1|X)$



- Linear model: $p(X) = \beta_0 + \beta_1 X = -0.07 + 0.008X$
- Predict 1 if $\hat{P}(x) \geq 0.5$, and 0 otherwise.
  - Solving the linear equation, predict 1 if $X \geq 73.4$

Problems with linear model

1. Our prediction $p(X)$ may take values outside 0 and 1.

Problems with linear model

1. Our prediction $p(X)$ may take values outside 0 and 1.
2. Too inflexible (enormous bias).

Problems with linear model

1. Our prediction $p(X)$ may take values outside 0 and 1.
2. Too inflexible (enormous bias).
3. In practice, $p(X)$ is rarely close to linear.

Odds

- Suppose a certain event occurs with probability p. The odds of the event occurring are

$$\text{odds} = \frac{p}{1-p}$$

Odds

- Suppose a certain event occurs with probability p. The odds of the event occurring are

$$\mathrm{odds} = \frac{p}{1-p}$$

  - If $p = .75$, then $\mathrm{odds} = 3$ (or 3 to 1).
  - If $p = .5$, then $\mathrm{odds} = 1$ (or even odds).

Odds

- Suppose a certain event occurs with probability p. The odds of the event occurring are

$$\text{odds} = \frac{p}{1-p}$$

  - If $p = .75$, then $\text{odds} = 3$ (or 3 to 1).
  - If $p = .5$, then $\text{odds} = 1$ (or even odds).
- But odds compress unlikely events towards 0, while stretching likely events towards infinity.

Odds

- Suppose a certain event occurs with probability p. The odds of the event occurring are

$$\text{odds} = \frac{p}{1-p}$$

  - If $p = .75$, then $\text{odds} = 3$ (or 3 to 1).
  - If $p = .5$, then $\text{odds} = 1$ (or even odds).

- But odds compress unlikely events towards 0, while stretching likely events towards infinity.
  - Events that are less likely to happen than not have odds between 0 and 1, while events that are more likely to happen than not have odds between 1 and infinity.

Odds

- Suppose a certain event occurs with probability p. The odds of the event occurring are

$$\text{odds} = \frac{p}{1-p}$$

  - If $p = .75$, then $\text{odds} = 3$ (or 3 to 1).
  - If $p = .5$, then $\text{odds} = 1$ (or even odds).
- But odds compress unlikely events towards 0, while stretching likely events towards infinity.
  - Events that are less likely to happen than not have odds between 0 and 1, while events that are more likely to happen than not have odds between 1 and infinity.
- So instead, we consider log odds:

$$\text{log odds} = \ln \frac{p}{1-p} = \ln p - \ln(1-p)$$

Logistic Regression

- Suppose Y is binary categorical, and that the log odds of the event "$Y = 1$" is linear in X. That is,

Logistic Regression

- Suppose Y is binary categorical, and that the log odds of the event "$Y = 1$" is linear in X. That is,

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

Logistic Regression

- Suppose Y is binary categorical, and that the log odds of the event "$Y = 1$" is linear in X. That is,

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- Increasing X by 1 increases the log odds of $Y = 1$ by a constant amount.

Logistic Regression

- Suppose Y is binary categorical, and that the log odds of the event "$Y = 1$" is linear in X. That is,

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- Increasing X by 1 increases the log odds of $Y = 1$ by a constant amount.
- Increasing X by 1 increases the odds of $Y = 1$ by a constant *relative rate*

Logistic Regression

- Suppose Y is binary categorical, and that the log odds of the event "$Y = 1$" is linear in X. That is,

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- Increasing X by 1 increases the log odds of $Y = 1$ by a constant amount.
- Increasing X by 1 increases the odds of $Y = 1$ by a constant *relative rate*
- Solving for odds:

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

Logistic Regression

- Suppose Y is binary categorical, and that the log odds of the event "$Y = 1$" is linear in X. That is,

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X$$

- Increasing X by 1 increases the log odds of $Y = 1$ by a constant amount.
- Increasing X by 1 increases the odds of $Y = 1$ by a constant *relative rate*
- Solving for odds:

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

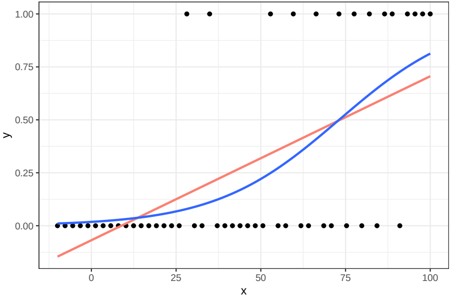
- Solving for $p(X)$:

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

The Logistic Curve

- The conditional probability $p(X)$ takes the form of a logistic curve:

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

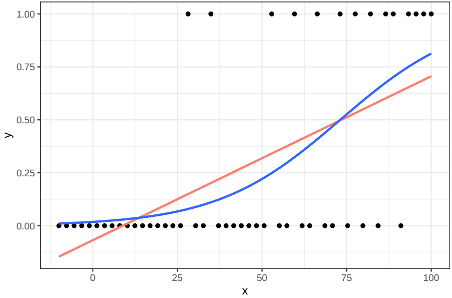


- Logistic model: $p(X) = \frac{e^{-4+0.05X}}{1+e^{-4+0.05X}}$

The Logistic Curve

- The conditional probability $p(X)$ takes the form of a logistic curve:

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

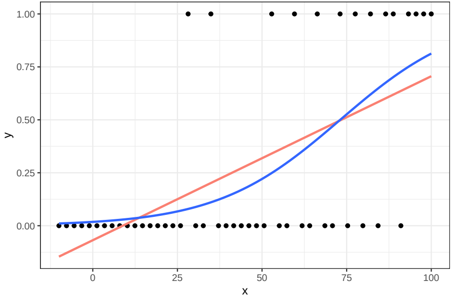


- Logistic model: $p(X) = \frac{e^{-4+0.05X}}{1+e^{-4+0.05X}}$
- Predict 1 if $\hat{P}(x) \geq 0.5$ (or if $\log \mathrm{odds} \geq 0$)

The Logistic Curve

- The conditional probability $p(X)$ takes the form of a logistic curve:

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



- Logistic model: $p(X) = \frac{e^{-4+0.05X}}{1+e^{-4+0.05X}}$
- Predict 1 if $\hat{P}(x) \geq 0.5$ (or if $\log \mathrm{odds} \geq 0$)
  - Solving the linear equation, predict 1 if $X \geq 73.1$

Multiple Logistic Regression

- Nothing stops us from modeling Y based on more than 1 predictor.

Multiple Logistic Regression

- Nothing stops us from modeling Y based on more than 1 predictor.

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Multiple Logistic Regression

- Nothing stops us from modeling Y based on more than 1 predictor.

$$\ln \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Solving for $p(X)$:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

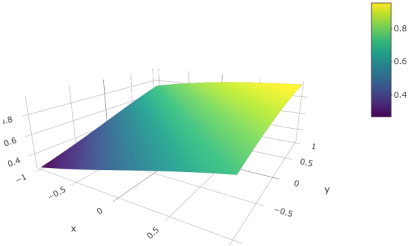
Multiple Logistic Regression

- Nothing stops us from modeling Y based on more than 1 predictor.

$$\ln \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Solving for $p(X)$:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$



Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method. . .

1. For historical reasons

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method. . .

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation
4. Because it often gives reasonable predictions

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation
4. Because it often gives reasonable predictions

Logistic regression has been used to...

1. Create spam filters

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation
4. Because it often gives reasonable predictions

Logistic regression has been used to...

1. Create spam filters
2. Forecast election results

Applications of Logistic Regression

Logistic Regression is the most commonly used binary classification method...

1. For historical reasons
2. Due to its relative simplicity
3. For ease of interpretation
4. Because it often gives reasonable predictions

Logistic regression has been used to...

1. Create spam filters
2. Forecast election results
3. Investigate health outcomes based on patient risk factors

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ based on training data.

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ based on training data.

- We could use the Method of Least Squares, as we did with Linear Regression.

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \dots, X_p$, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - We need to estimate the parameters $\beta_0, \beta_1, \dots, \beta_p$ based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
  - But there isn't a closed-from solution as in Linear Regression

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \dots, X_p$, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - We need to estimate the parameters $\beta_0, \beta_1, \dots, \beta_p$ based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
  - But there isn't a closed-from solution as in Linear Regression
  - And in practice, residuals tend not to be approximately Normally distributed

Regression Coefficient Estimates

- Assume that the log-odds of $Y = 1$ is indeed linear in $X_1, \ldots, X_p$, so that

$$\ln \frac{p(X)}{1 - p(X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$ based on training data.
- We could use the Method of Least Squares, as we did with Linear Regression.
  - But there isn't a closed-from solution as in Linear Regression
  - And in practice, residuals tend not to be approximately Normally distributed
- Instead, we use the method of **Maximum Likelihood** (ML)

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- Then the probability of the observed data is

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{k} p(x_i) \prod_{j=k+1}^{n} (1 - p(x_j))$$

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- Then the probability of the observed data is

$$\ell(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^{k} p(x_i) \prod_{j=k+1}^{n} (1 - p(x_j))$$

  - View ℓ as a function of parameters $\beta_0, \dots, \beta_p$ for **fixed** observations $x_1, \dots, x_n$.

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
    - Assume we've relabeled indices so the first k observations have $y = 1$
    - As before, we assume
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$
- Then the probability of the observed data is
$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{k} p(x_i) \prod_{j=k+1}^{n} (1 - p(x_j))$$
    - View ℓ as a function of parameters $\beta_0, \ldots, \beta_p$ for **fixed** observations $x_1, \ldots, x_n$.
- The goal is to choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ so as to maximize ℓ

The Method of Maximum Likelihood

- Under ML, we compare all possible models and select the one for which the observed data had highest probability of occurring
- Suppose we have k observations with $y = 1$ and $n - k$ with $y = 0$.
  - Assume we've relabeled indices so the first k observations have $y = 1$
  - As before, we assume

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- Then the probability of the observed data is

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{k} p(x_i) \prod_{j=k+1}^{n} (1 - p(x_j))$$

  - View ℓ as a function of parameters $\beta_0, \ldots, \beta_p$ for **fixed** observations $x_1, \ldots, x_n$.
- The goal is to choose $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ so as to maximize ℓ
  - How? (Calculus or numeric methods, or R!)

Summary

- In a classification problem, we are interested a categorical response variable Y.

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.
- Ideally, we would like to estimate the conditional probability of Y given X

$$P(Y = A_j | X)$$

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.
- Ideally, we would like to estimate the conditional probability of Y given X

$$P(Y = A_j | X)$$

- For binary response Y, we can use logistic regression, which assumes the log-odds of $Y = 1$ is linear:

$$\ln \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.
- Ideally, we would like to estimate the conditional probability of Y given X

$$P(Y = A_j | X)$$

- For binary response Y, we can use logistic regression, which assumes the log-odds of $Y = 1$ is linear:

$$\ln \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- This implies the conditional probability is logistic:

$$P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

Summary

- In a classification problem, we are interested a categorical response variable Y.
- We might be interested in **predicting** the class for Y based on observations, or we might be interested in **inferring** the relationships between Y and predictors.
- Ideally, we would like to estimate the conditional probability of Y given X

$$P(Y = A_j|X)$$

- For binary response Y, we can use logistic regression, which assumes the log-odds of $Y = 1$ is linear:

$$\ln \frac{P(Y=1|X)}{1 - P(Y=1|X)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- This implies the conditional probability is logistic:

$$P(Y=1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

- To classify, we assign a test observation the value 1 if

$$P(Y=1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}} \geq 0.5$$