

# Logistic Regression: Part I

## Math 243

Michael Pearce

# Outline

Before fall break, we . . .

- Discussed penalization to improve MSE in regression models
- Learned two penalization methods: Ridge and Lasso.
- Started to think about **classification** with KNN.

In today's class, we will . . .

- Review **classification** and **KNN**
- Introduce **Logistic Regression** for Classification

## Classification Review

# Classification Problems

Suppose  $Y$  is a **categorical variable** with levels  $A_1, A_2, \dots, A_k$ .

# Classification Problems

Suppose  $Y$  is a **categorical variable** with levels  $A_1, A_2, \dots, A_k$ .

- Ex:  $Y$  indicates whether it is raining in Portland today

$$A_1 = \text{Raining} \quad A_2 = \text{Not Raining}$$

# Classification Problems

Suppose  $Y$  is a **categorical variable** with levels  $A_1, A_2, \dots, A_k$ .

- Ex:  $Y$  indicates whether it is raining in Portland today

$$A_1 = \text{Raining} \quad A_2 = \text{Not Raining}$$

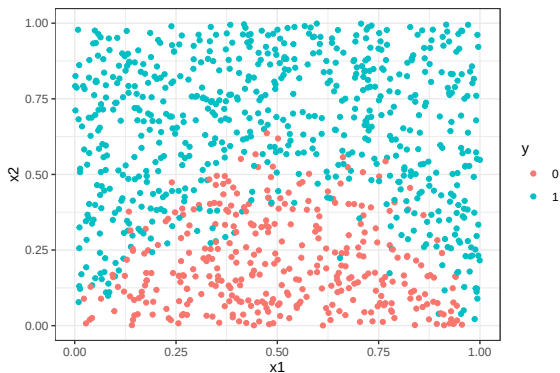
**Goal:** Build a model  $\hat{g}$  to classify an observation into levels  $(A_1, A_2, \dots, A_k)$  based on predictors  $(X_1, X_2, \dots, X_p)$

$$\hat{Y} = \hat{g}(X_1, X_2, \dots, X_p) \quad \text{where } \hat{g} \text{ take values in } \{A_1, \dots, A_k\}$$

# Classification Regions

Classification models divide the predictor space into different regions

- test points that fall into a region are classified in the same way

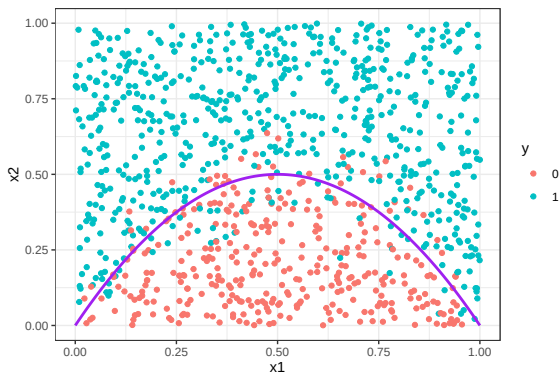


Different models will have different geometries for classification boundaries.

# Classification Regions

Classification models divide the predictor space into different regions

- test points that fall into a region are classified in the same way



Purple line indicates the optimal decision boundary, determined by the **Bayes Classifier**

# The Bayes Classifier

**Bayes Classifier:** For a test point  $x_0$ ,

- 1 Calculates true probability  $Y$  belongs to each class,  $A_1, \dots, A_k$

$$P(Y = A_j | X = x_0)$$

- 2 Classifies  $Y$  in class with highest probability.

# The Bayes Classifier

**Bayes Classifier:** For a test point  $x_0$ ,

- 1 Calculates true probability  $Y$  belongs to each class,  $A_1, \dots, A_k$

$$P(Y = A_j | X = x_0)$$

- 2 Classifies  $Y$  in class with highest probability.

- *Theoretically* minimizes error rate, but. . .
- **"True"** probabilities are unknown!

# The Bayes Classifier

**Bayes Classifier:** For a test point  $x_0$ ,

- 1 Calculates true probability  $Y$  belongs to each class,  $A_1, \dots, A_k$

$$P(Y = A_j | X = x_0)$$

- 2 Classifies  $Y$  in class with highest probability.

- *Theoretically* minimizes error rate, but. . .
- **"True"** probabilities are unknown!

The classification probabilities above are just like  $f$  in regression:

$$Y = f(X) + \epsilon$$

- Regression:  $f$  captures relationship between  $Y$  and  $X$

# The Bayes Classifier

**Bayes Classifier:** For a test point  $x_0$ ,

- 1 Calculates true probability  $Y$  belongs to each class,  $A_1, \dots, A_k$

$$P(Y = A_j | X = x_0)$$

- 2 Classifies  $Y$  in class with highest probability.

- *Theoretically* minimizes error rate, but. . .
- **"True" probabilities are unknown!**

The classification probabilities above are just like  $f$  in regression:

$$Y = f(X) + \epsilon$$

- Regression:  $f$  captures relationship between  $Y$  and  $X$
- Classification:  $P(Y = A_j | X)$  for all  $j$  captures relationship between  $Y$  and  $X$
- In both settings, **these functions/probabilities are unknown!** We must estimate them using data.

## K-Nearest Neighbors

A while back, we approximated the probabilities  $P(Y = A_j | X = x_0)$  using **KNN**:

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

where  $N_0$  is the set of  $K$  training data points *closest* to  $x_0$

## K-Nearest Neighbors

A while back, we approximated the probabilities  $P(Y = A_j | X = x_0)$  using **KNN**:

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

where  $N_0$  is the set of  $K$  training data points *closest* to  $x_0$

- KNN classifies points based on modal/average  $y$  among the  $K$ -nearest neighbors.

## K-Nearest Neighbors

A while back, we approximated the probabilities  $P(Y = A_j | X = x_0)$  using **KNN**:

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

where  $N_0$  is the set of  $K$  training data points *closest* to  $x_0$

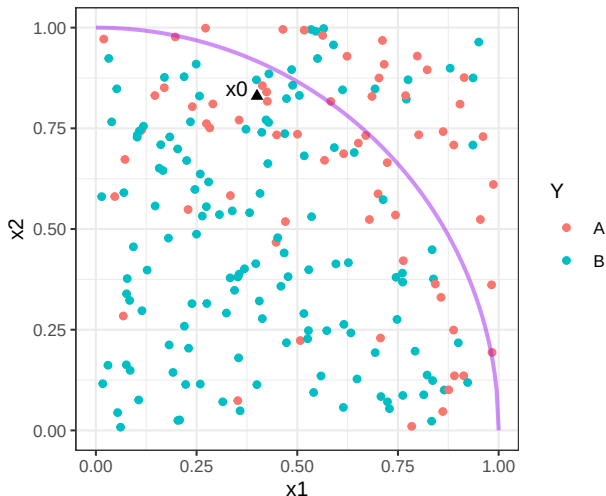
- KNN classifies points based on modal/average  $y$  among the  $K$ -nearest neighbors.
- **Recall:** The  $K$  in KNN determines model complexity:

Small  $K \approx$  \*more\* complex      Large  $K \approx$  \*less\* complex

# KNN Example

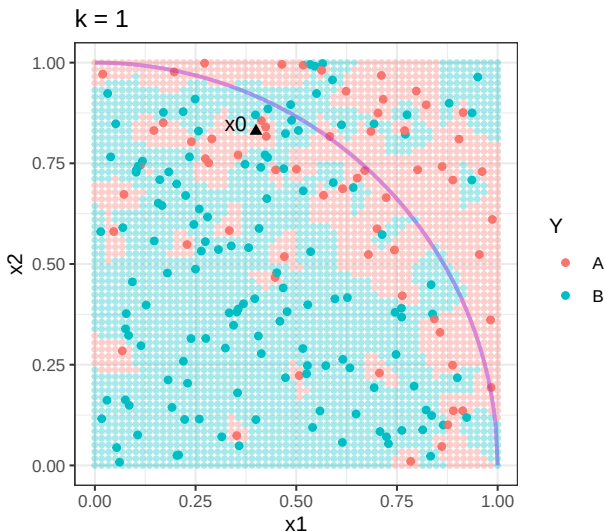
Here are the classification boundaries for a variety of values of  $K$ .

Training Data and Bayes Classifier



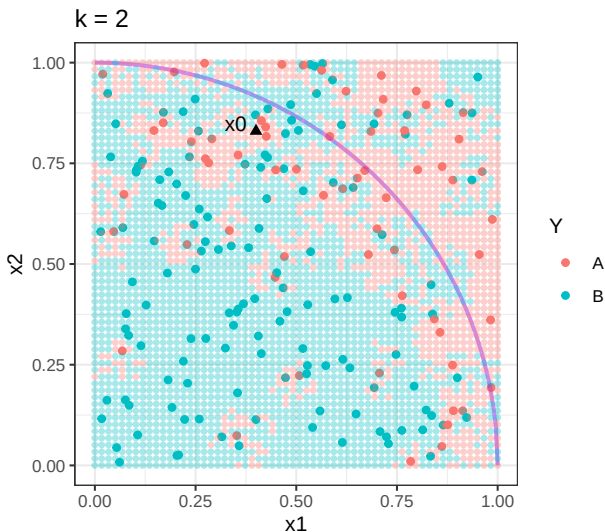
$K=1$ 

Here are the classification boundaries for a variety of values of  $K$ .



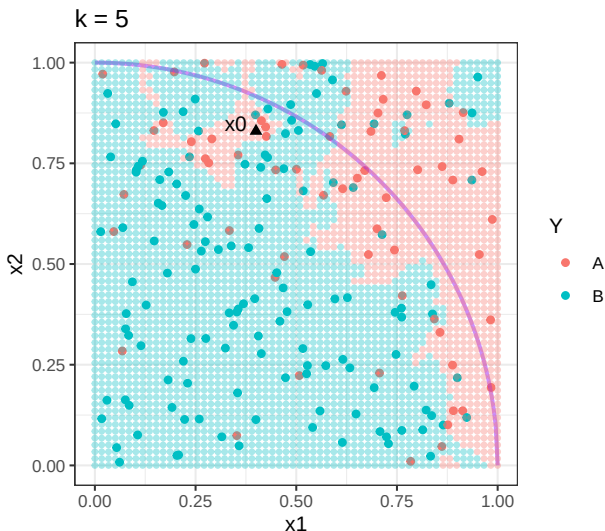
$K=2$

Here are the classification boundaries for a variety of values of  $K$ .



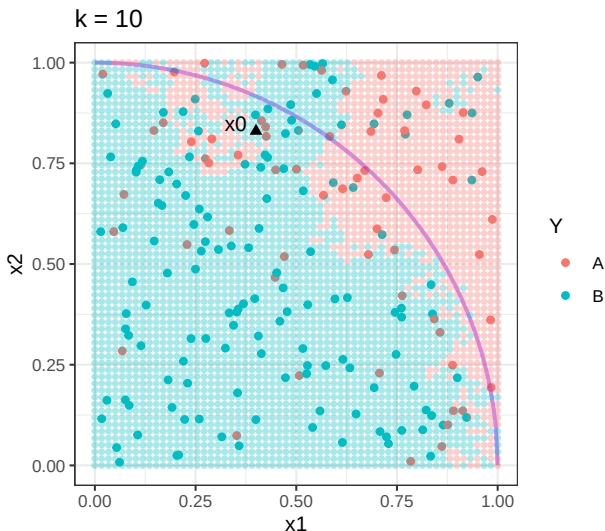
$K=5$

Here are the classification boundaries for a variety of values of  $K$ .



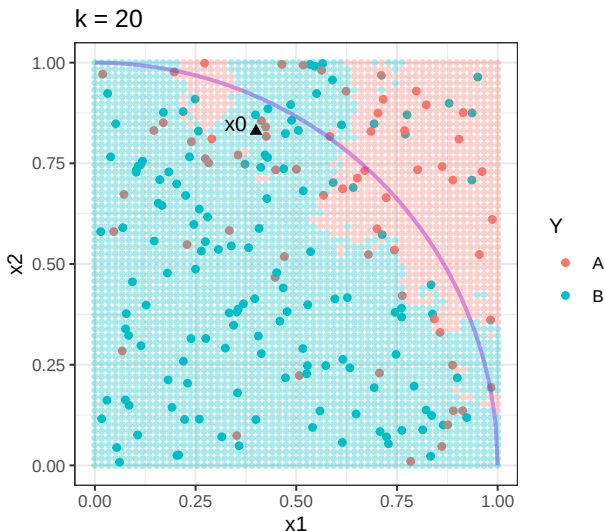
$K=10$

Here are the classification boundaries for a variety of values of  $K$ .



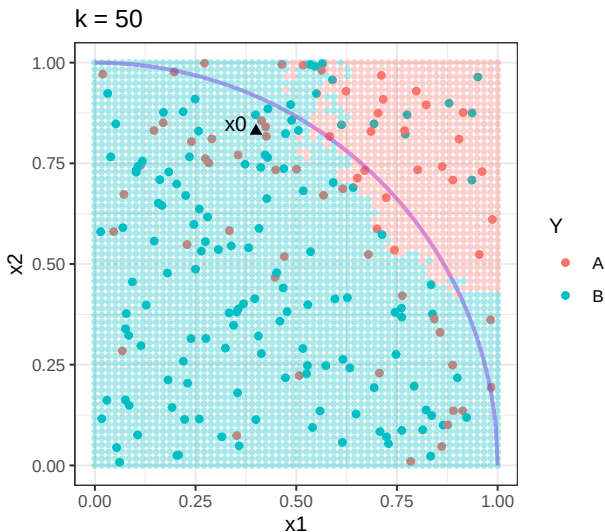
$K=20$

Here are the classification boundaries for a variety of values of  $K$ .



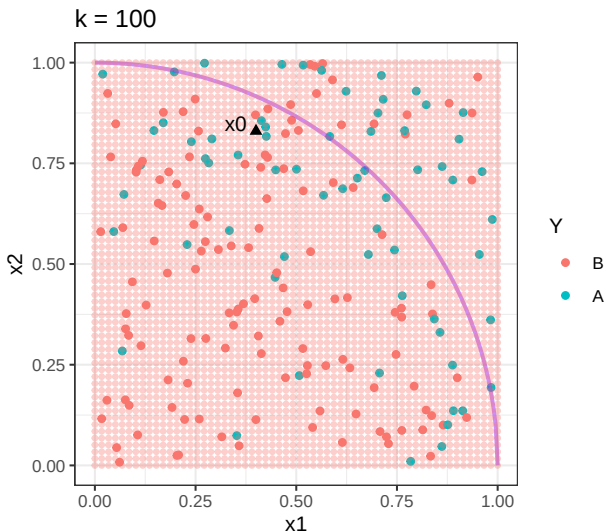
$K=50$

Here are the classification boundaries for a variety of values of  $K$ .



# K=100

Here are the classification boundaries for a variety of values of  $K$ .



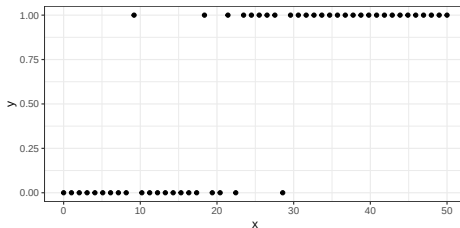
# KNN Cons

- ① Fast to “train” models, but slow to “test” (especially for large  $K$ )
- ② Hard to interpret (can’t do inference)
- ③ If data truly has a structure/parametric form, KNN will be too complex.
- ④ Suffers from the “curse of dimensionality”.
- ⑤ Requires large sample sizes to be accurate!

## Could we classify with a Linear Model?

Let  $Y$  be a *binary* variable:  $Y = 0$  or  $Y = 1$

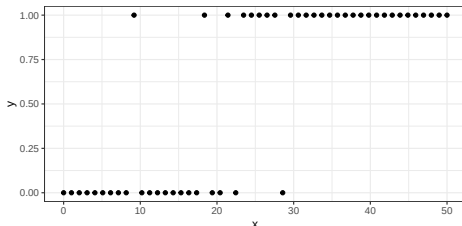
- We have a single quantitative predictor  $X$ .
- Want to estimate  $p(X) = P(Y = 1|X)$



## Could we classify with a Linear Model?

Let  $Y$  be a *binary* variable:  $Y = 0$  or  $Y = 1$

- We have a single quantitative predictor  $X$ .
- Want to estimate  $p(X) = P(Y = 1|X)$



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

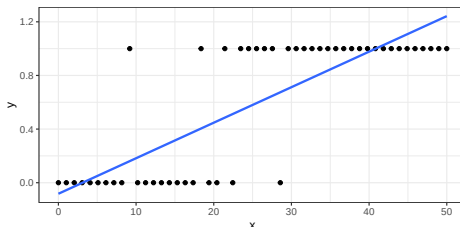
Predictions:

- Predict 1 if  $\hat{p}(x_0) \geq 0.5$
- Predict 0 if  $\hat{p}(x_0) < 0.5$

## Could we classify with a Linear Model?

Let  $Y$  be a *binary* variable:  $Y = 0$  or  $Y = 1$

- We have a single quantitative predictor  $X$ .
- Want to estimate  $p(X) = P(Y = 1|X)$

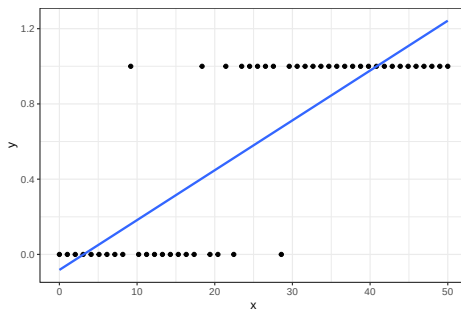


Fit by Least Squares:  $\hat{p}(X) = -0.08 + 0.03X$

Predictions:

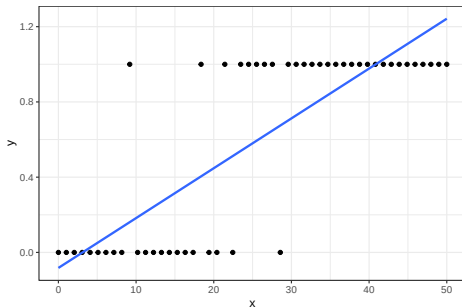
- Predict 1 if  $X \geq 22.0$
- Predict 0 if  $X < 22.0$

# Problems with the linear model idea



**TPS:** What are some issues with this linear model idea?

## Problems with the linear model idea



**TPS:** What are some issues with this linear model idea?

- ① Our prediction  $\hat{p}(X)$  may take values outside 0 and 1.
- ② Conceptually, a linear model for  $p(X)$  is wrong! → **Let's focus on this**

## Problems with linear model idea

Ex: Let's say  $Y$  is an indicator of finishing a marathon

- $Y = 1$  if you *finish* the marathon
- $Y = 0$  if you *don't* finish the marathon

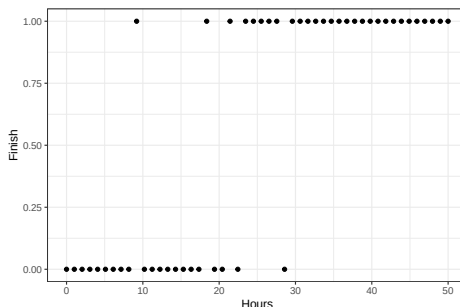
Predictor  $X$  is the number of hours you run each week

## Problems with linear model idea

Ex: Let's say  $Y$  is an indicator of finishing a marathon

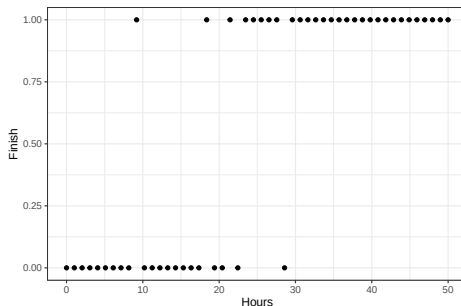
- $Y = 1$  if you *finish* the marathon
- $Y = 0$  if you *don't* finish the marathon

Predictor  $X$  is the number of hours you run each week



- As hours ran increases, it seems like the prob of finishing the marathon increases

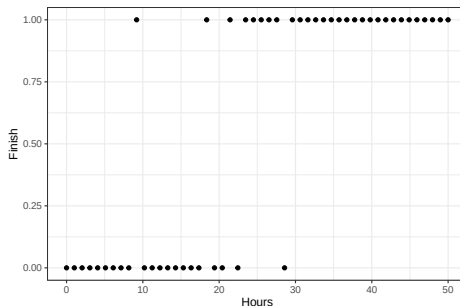
## Problems with linear model idea



**Q:** Why might we think that the probability of finishing the marathon is *not* linear function of  $X$ ?

$$P(Y = 1|X) = \beta_0 + \beta_1 X$$

# Problems with linear model idea



Q: Why might we think that the probability of finishing the marathon is *not* linear function of  $X$ ?

$$P(Y = 1|X) = \beta_0 + \beta_1 X$$

- e.g., this model assumes the change in probability from 15 to 25 hours ran would be the same as from 40 to 50 hours ran!

## Section 2

# Logistic Regression

# Logistic Regression

**Logistic Regression:** For a binary (0 or 1) response  $Y$ , logistic regression models:

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

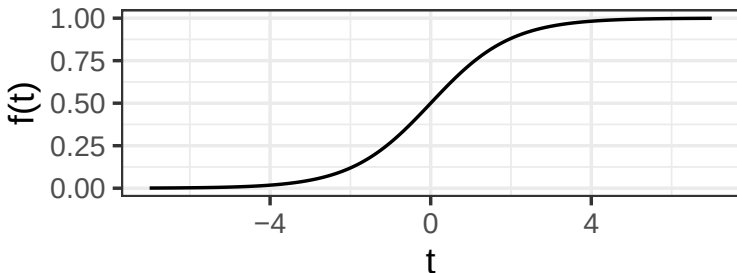
# Logistic Regression

**Logistic Regression:** For a binary (0 or 1) response  $Y$ , logistic regression models:

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

This uses the **Logistic Function**:

$$f(t) = \frac{e^t}{1 + e^t}$$



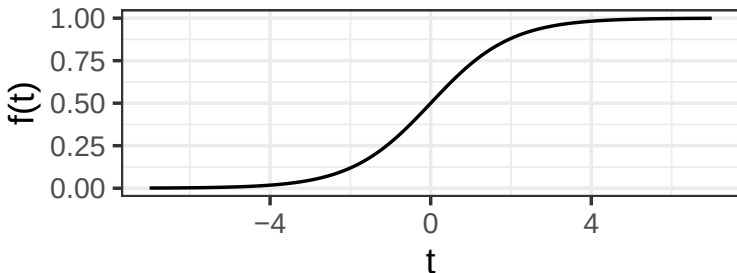
# Logistic Regression

**Logistic Regression:** For a binary (0 or 1) response  $Y$ , logistic regression models:

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

This uses the **Logistic Function**:

$$f(t) = \frac{e^t}{1 + e^t}$$



- The logistic function is always between 0 and 1!
- $f(t)$  does not change much on the extremes of  $t$

# Logistic Regression

Logistic Regression:

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

# Logistic Regression

Logistic Regression:

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

Q: If  $\beta_1 > 0$ , what happens to  $p(X)$  as  $X$  increases? Decreases?

Q: If  $\beta_1 < 0$ , what happens to  $p(X)$  as  $X$  increases? Decreases?

# Logistic Regression

Logistic Regression:

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

Q: If  $\beta_1 > 0$ , what happens to  $p(X)$  as  $X$  increases? Decreases?

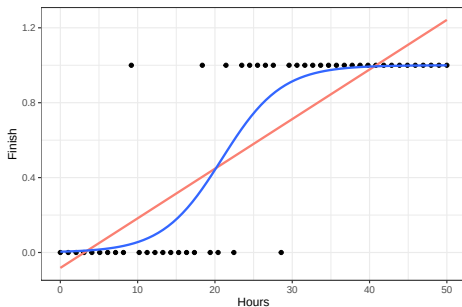
Q: If  $\beta_1 < 0$ , what happens to  $p(X)$  as  $X$  increases? Decreases?

Predictions: for test point  $x_0$

- Predict 1 if  $\hat{p}(x_0) \geq 0.5$
- Predict 0 if  $\hat{p}(x_0) < 0.5$

## Plotting the curve

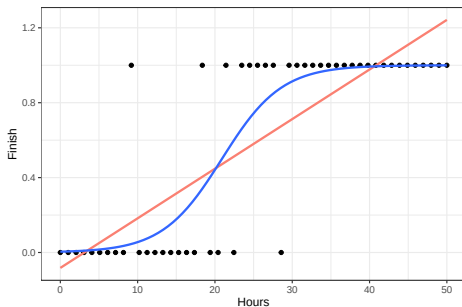
Consider previous data where  $Y$  is an indicator of finishing a marathon, and  $X$  is hours ran/week.



- Linear Model (Red):  $\hat{p}(X) = -0.08 + 0.03X$
- Logistic Model (Blue):  $\hat{p}(X) = \frac{e^{-5.4+0.26X}}{1+e^{-5.4+0.26X}}$

## Plotting the curve

Consider previous data where  $Y$  is an indicator of finishing a marathon, and  $X$  is hours ran/week.



- Linear Model (Red):  $\hat{p}(X) = -0.08 + 0.03X$
- Logistic Model (Blue):  $\hat{p}(X) = \frac{e^{-5.4+0.26X}}{1+e^{-5.4+0.26X}}$

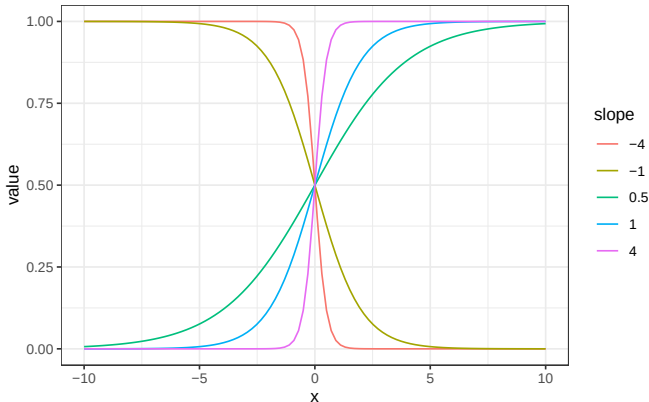
Q: How does the logistic regression model improve upon the linear model?

# Meaning of Slope Coefficients

## Logistic Regression Model:

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

Effect of Slope, with constant intercept of 0

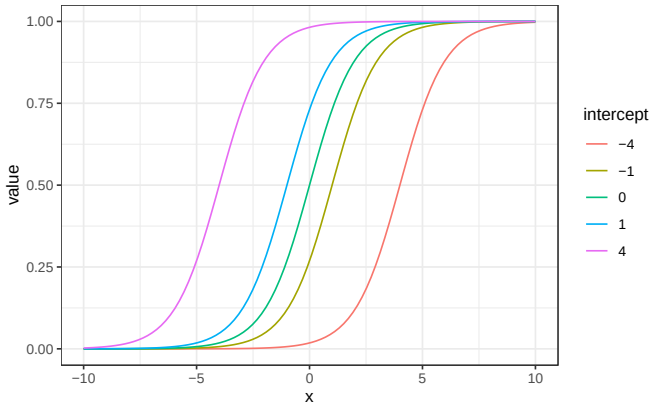


# Meaning of Intercept Coefficients

## Logistic Regression Model:

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

Effect of Intercept, with constant slope of 1

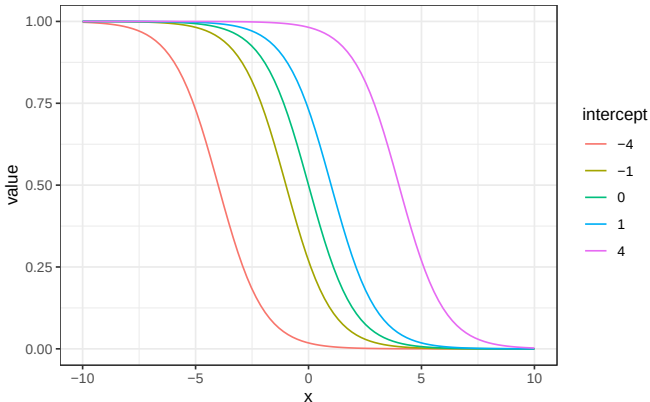


# Meaning of Intercept Coefficients

## Logistic Regression Model:

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

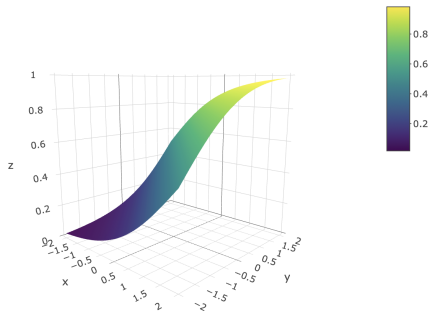
Effect of Intercept, with constant slope of -1



# Multiple Logistic Regression

Easily extends to more than one predictor:

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



- The logistic curve becomes a 3D curved surface.

# 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

## Summary

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

- Might be interested in **predicting** the class for  $Y$
- Might be interested in **inferring** the relationships between  $Y$  and predictors  $X$ .

## Summary

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

- Might be interested in **predicting** the class for  $Y$
- Might be interested in **inferring** the relationships between  $Y$  and predictors  $X$ .

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  $Y$ .

- Might be interested in **predicting** the class for  $Y$
- Might be interested in **inferring** the relationships between  $Y$  and predictors  $X$ .

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 conditional probability is a logistic function of  $X$ :

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

## Summary

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

- Might be interested in **predicting** the class for  $Y$
- Might be interested in **inferring** the relationships between  $Y$  and predictors  $X$ .

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 conditional probability is a logistic function of  $X$ :

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

### Prediction:

- Assign a test observation the value 1 if  $P(Y = 1 | X) \geq 0.50$
- Assign the value 0 if  $P(Y = 1 | X) < 0.50$

## Summary

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

- Might be interested in **predicting** the class for  $Y$
- Might be interested in **inferring** the relationships between  $Y$  and predictors  $X$ .

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 conditional probability is a logistic function of  $X$ :

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

### Prediction:

- Assign a test observation the value 1 if  $P(Y = 1|X) \geq 0.50$
- Assign the value 0 if  $P(Y = 1|X) < 0.50$

### Inference:

- If  $\beta_k > 0$ , then the probability of success *increases* as  $X_k$  increases

## Section 3

### Recap

# Recap

Today, we discussed. . .

- Classification via KNN (nonparametric)
- Classification via Logistic Regression (parametric)

Next time, we will. . .

- Continue discussing logistic regression