# Multiple Regression

Kirk Bansak

February 16, 2023

# Bivariate Regression Review

# Using Bivariate Linear Regression for $\hat{f}$

For each election $i$, where $i$ is used to index different observations,

Let:

$Vote_i$ = Incumbent Vote Share in election $i$.
$Approval_i$ = Incumbent Net Approval in election $i$.

Employing linear function to relate $Approval_i$ to $Vote_i$:

$$\text{Vote}_i \quad = \quad \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i$$

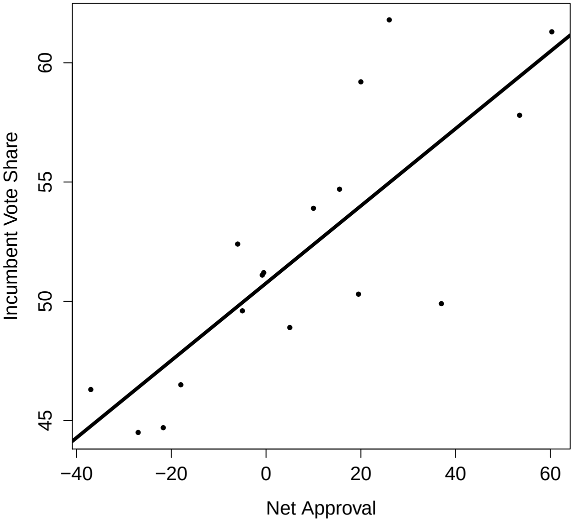
**After Estimation:**

$$\text{Vote}_i \quad = \quad \underbrace{\hat{\beta}_0 + \hat{\beta}_1 \text{Approval}_i}_{\widehat{\text{Vote}_i}} + \hat{\epsilon}_i$$

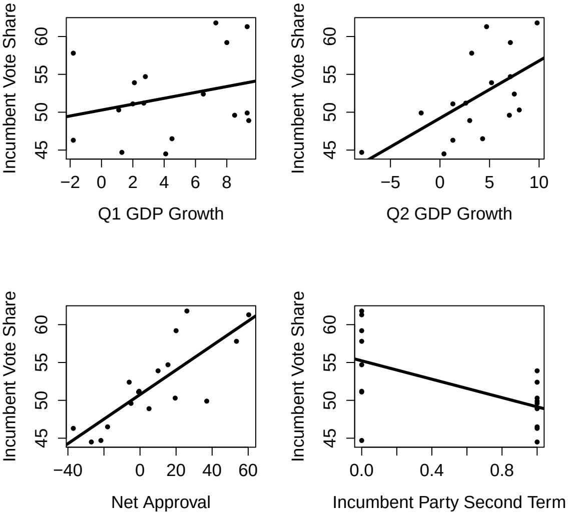
With $\hat{\beta}_0$ and $\hat{\beta}_1$ chosen via:

$$\underset{\hat{\beta}_0, \hat{\beta}_1}{\arg\min} \quad \sum_{i=1}^{N} \left( \text{Vote}_i - \hat{\beta}_0 - \hat{\beta}_1 \text{Approval}_i \right)^2$$

# Bivariate Regression: Geometric Perspective



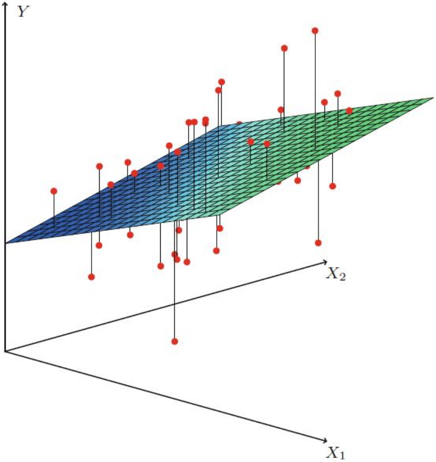
# Many Separate Bivariate Regressions



# Moving to Multiple Regression

# Multiple Regression: Geometric Perspective

**Example Predicting $Y$ with Two Variables At the Same Time**
**($X_1$ and $X_2$)**



# Multiple Regression: Function Perspective

Now consider our election case with **four** predictors.

For each election $i$, where $i$ is used to index different observations,

Let:

$Vote_i$ = Incumbent Vote Share in election $i$.
$Approval_i$ = Incumbent Net Approval in election $i$.
$Q1\ GDP_i$ = GDP Growth in Quarter immediately preceding election $i$.
$Q2\ GDP_i$ = GDP Growth Two Quarters immediately preceding election $i$.
$Inc\ 2nd\ Term_i$ = Indicator for whether or not incumbent is currently serving in second or greater term in election $i$.

Linear regression model is now:

$$\begin{aligned}\text{Vote}_i &= f(\text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i) + \epsilon_i \\ &= \beta_0 + \beta_1 \text{Approval}_i + \beta_2 \text{Q1 GDP}_i \\ &\quad + \beta_3 \text{Q2 GDP}_i + \beta_4 \text{Inc 2nd Term}_i + \epsilon_i\end{aligned}$$

# Multiple Regression: Function Perspective

Regression Model:

$$\text{Vote}_i = \beta_0 + \beta_1 \text{Approval}_i + \beta_2 \text{Q1 GDP}_i + \beta_3 \text{Q2 GDP}_i + \beta_4 \text{Inc 2nd Term}_i + \epsilon_i$$

After Estimation:

$$\text{Vote}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Approval}_i + \hat{\beta}_2 \text{Q1 GDP}_i + \hat{\beta}_3 \text{Q2 GDP}_i + \hat{\beta}_4 \text{Inc 2nd Term}_i + \hat{\epsilon}_i$$

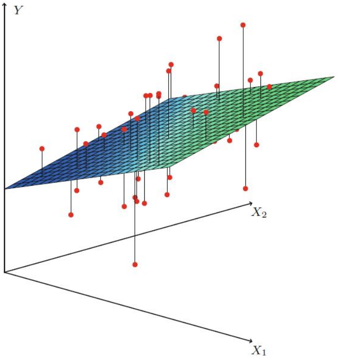
Prediction Function:

$$\widehat{\text{Vote}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Approval}_i + \hat{\beta}_2 \text{Q1 GDP}_i + \hat{\beta}_3 \text{Q2 GDP}_i + \hat{\beta}_4 \text{Inc 2nd Term}_i$$

# Test Your Knowledge

Consider again an example with 2 predictors, leading to the fitted regression:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\epsilon}_i$$




1) What does $\hat{\beta}_0$ correspond to?
2) What does $\hat{\beta}_1$ correspond to?
3) What does $\hat{\beta}_2$ correspond to?
4) What would be interpolation vs. extrapolation?

# Fitting a Multiple Regression

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_p x_{pi} + \hat{\epsilon}_i$$

$$\begin{aligned}\hat{\epsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - ... - \hat{\beta}_p x_{pi}\end{aligned}$$

# Fitting a Multiple Regression

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_p x_{pi} + \hat{\epsilon}_i$$

$$\begin{aligned}\hat{\epsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - ... - \hat{\beta}_p x_{pi}\end{aligned}$$

As in the bivariate case, in the multiple regression case, we (our software!) will choose $\hat{\beta}$ values to minimize the sum of the squared residuals:

$$\underset{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_j}{\arg\min} \quad \sum_{i=1}^{N} \hat{\epsilon}_i^2$$

# Fitting a Multiple Regression

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_p x_{pi} + \hat{\epsilon}_i$$

$$\begin{aligned} \hat{\epsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - ... - \hat{\beta}_p x_{pi} \end{aligned}$$

As in the bivariate case, in the multiple regression case, we (our software!) will choose $\hat{\beta}$ values to minimize the sum of the squared residuals:

$$\underset{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_j}{\arg\min} \sum_{i=1}^{N} \hat{\epsilon}_i^2$$

That is,

$$\underset{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p}{\arg\min} \sum_{i=1}^{N} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - ... - \hat{\beta}_p x_{pi} \right)^2$$

# Conditional Expectation Function Property

Conditional Expectation Function (CEF): $E[Y|X]$

# Conditional Expectation Function Property

Conditional Expectation Function (CEF): $E[Y|X]$

*Aside from its mathematical elegance, another desirable property of minimizing sum of squared errors:*

**CEF Prediction Property**:

Let $m(X)$ be any function of $X$. The CEF solves

$$\underset{m(X)}{\arg\min}\, E[(Y - m(X))^2] = E[Y|X]$$

# Conditional Expectation Function Property

Conditional Expectation Function (CEF): $E[Y|X]$

*Aside from its mathematical elegance, another desirable property of minimizing sum of* *squared* *errors:*

**CEF Prediction Property**:

Let $m(X)$ be any function of $X$. The CEF solves

$$\underset{m(X)}{\arg\min}\, E[(Y - m(X))^2] = E[Y|X]$$

In other words, minimizing the mean squared error at the population level results in the function $m(X)$ being the CEF.

Implies that if the CEF is linear, then the OLS solution (minimizing sum of squared errors) provides the CEF at the population level.

# Multiple Regression: Our Estimated Prediction Function

$$\widehat{\text{Vote}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Approval}_i + \hat{\beta}_2 \text{Q1 GDP}_i + \hat{\beta}_3 \text{Q2 GDP}_i + \hat{\beta}_4 \text{Inc 2nd Term}_i$$

$$\begin{aligned}\widehat{\text{Vote}}_i &= 51.01 + 0.10 \times \text{Approval}_i + 0.57 \times \text{Q1 GDP}_i \\ &\quad + 0.10 \times \text{Q2 GDP}_i - 4.35 \times \text{Inc 2nd Term}_i\end{aligned}$$

How do we interpret the $\hat{\beta}$'s?

# Regression in R

- Key Functions:
  - `lm`
  - `summary`
  - `predict`

To R!

Some Linear Algebra Basics

Vector: ordered n-tuple of numbers

- $1$
- $\pi$
- $(1, 2)$
- $(0, 0)$
- $(\pi, e)$
- $(3.1, 4.5, 6.11132)$
- $(\beta_0, \beta_1, \beta_2, \beta_3)$
- $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$

We will write vectors with bold ($\boldsymbol{\beta}$)

# Inner Product

Consider two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ and they are the same length. The define their inner product, $\boldsymbol{u} \cdot \boldsymbol{v}$, as

$$\begin{aligned} \boldsymbol{u} \cdot \boldsymbol{v} &= u_1 v_1 + u_2 v_2 + \ldots + u_p v_p \\ &= \sum_{j=1}^{p} u_j v_j \end{aligned}$$

# Rewriting our Prediction

Define:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4) \\
&= (51.01, 0.10, 0.57, 0.10, -4.35) \\
\boldsymbol{x}_i &= (1, \text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i)
\end{aligned}
$$

# Rewriting our Prediction

Define:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4) \\
&= (51.01, 0.10, 0.57, 0.10, -4.35) \\
\boldsymbol{x}_i &= (1, \text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i)
\end{aligned}
$$

Note that the first element of $\boldsymbol{x}_i$ is 1, corresponding to the intercept!

# Rewriting our Prediction

Define:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4) \\
&= (51.01, 0.10, 0.57, 0.10, -4.35) \\
\boldsymbol{x}_i &= (1, \text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i)
\end{aligned}$$

Note that the first element of $\boldsymbol{x}_i$ is 1, corresponding to the intercept!

Then, we can write our prediction as:

$$\widehat{\text{Vote}_i} = \widehat{\boldsymbol{\beta}} \cdot \boldsymbol{x}_i$$

# Rewriting our Prediction

Define:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4) \\
&= (51.01, 0.10, 0.57, 0.10, -4.35) \\
\boldsymbol{x}_i &= (1, \text{Approval}_i, \text{Q1 GDP}_i, \text{Q2 GDP}_i, \text{Inc 2nd Term}_i)
\end{aligned}$$

Note that the first element of $\boldsymbol{x}_i$ is 1, corresponding to the intercept!

Then, we can write our prediction as:

$$\widehat{\text{Vote}_i} = \widehat{\boldsymbol{\beta}} \cdot \boldsymbol{x}_i$$

To R!

# Useful Properties to Know About

Standard statistical software (e.g. `R`) will output standard errors for each $\hat{\beta}_j$, which are measures of uncertainty.

# Useful Properties to Know About

Standard statistical software (e.g. `R`) will output standard errors for each $\hat{\beta}_j$, which are measures of uncertainty.

Typically, 95% confidence intervals can be constructed for each $\hat{\beta}_j$ as such:

$$\left[\hat{\beta}_j - 1.96 \cdot SE(\hat{\beta}_j), \hat{\beta}_j + 1.96 \cdot SE(\hat{\beta}_j)\right]$$

where $SE(\hat{\beta}_j)$ denotes the standard error for $\hat{\beta}_j$.

# Useful Properties to Know About

Standard statistical software (e.g. R) will output standard errors for each $\hat{\beta}_j$, which are measures of uncertainty.

Typically, 95% confidence intervals can be constructed for each $\hat{\beta}_j$ as such:

$$\left[\hat{\beta}_j - 1.96 \cdot SE(\hat{\beta}_j), \hat{\beta}_j + 1.96 \cdot SE(\hat{\beta}_j)\right]$$

where $SE(\hat{\beta}_j)$ denotes the standard error for $\hat{\beta}_j$.

**Note: The bootstrap can also be used to construct this confidence interval as well as confidence intervals for other quantities of interest.**

# Useful Properties to Know About

Recall that the **Residual Sum of Squares (RSS)**, which measures the unexplained variation in the outcome, is the following:

$$RSS = \sum_{i=1}^{N} \hat{\epsilon}_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

# Useful Properties to Know About

Recall that the **Residual Sum of Squares (RSS)**, which measures the unexplained variation in the outcome, is the following:

$$RSS = \sum_{i=1}^{N} \hat{\epsilon}_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

In addition, the **Total Sum of Squares (TSS)**, which measures the total variation in the outcome, is the following:

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

# Useful Properties to Know About

Recall that the **Residual Sum of Squares (RSS)**, which measures the unexplained variation in the outcome, is the following:

$$RSS = \sum_{i=1}^{N} \hat{\epsilon}_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

In addition, the **Total Sum of Squares (TSS)**, which measures the total variation in the outcome, is the following:

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

The following is the **$\mathbf{R^2}$ Statistic**, which measures the proportion of variability in the outcome that is explained using the predictor(s):

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

# Important Considerations for Predictor Variables

- Qualitative Predictors

  e.g. $U.S.\ Citizen = \{\text{Yes, No}\}$

  e.g. $Nationality = \{\text{United States, Canada, Mexico, Germany, ...}\}$

# Important Considerations for Predictor Variables

- Qualitative Predictors

  e.g. $U.S.\ Citizen = \{\text{Yes, No}\}$

  e.g. $Nationality = \{\text{United States, Canada, Mexico, Germany, ...}\}$

- Interaction Terms

  e.g. $\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 Female_i + \hat{\beta}_2 Age_i + \hat{\beta}_3 Female_i \cdot Age_i$

# Important Considerations for Predictor Variables

- Qualitative Predictors

  e.g. $U.S.\ Citizen = \{\text{Yes, No}\}$

  e.g. $Nationality = \{\text{United States, Canada, Mexico, Germany, ...}\}$

- Interaction Terms

  e.g. $\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 Female_i + \hat{\beta}_2 Age_i + \hat{\beta}_3 Female_i \cdot Age_i$

- Polynomial Terms

  e.g. $\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 Age_i + \hat{\beta}_2 Age_i^2$

# Important Considerations for Predictor Variables

- Qualitative Predictors

  e.g. $U.S.\ Citizen = \{\text{Yes, No}\}$

  e.g. $Nationality = \{\text{United States, Canada, Mexico, Germany, ...}\}$

- Interaction Terms

  e.g. $\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 Female_i + \hat{\beta}_2 Age_i + \hat{\beta}_3 Female_i \cdot Age_i$

- Polynomial Terms

  e.g. $\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 Age_i + \hat{\beta}_2 Age_i^2$

- Collinearity
  - Highly correlated predictors
  - Perfectly correlated predictors

# Important Considerations for Predictor Variables

- Qualitative Predictors

  e.g. $U.S.\ Citizen = \{\text{Yes, No}\}$

  e.g. $Nationality = \{\text{United States, Canada, Mexico, Germany, ...}\}$

- Interaction Terms

  e.g. $\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 Female_i + \hat{\beta}_2 Age_i + \hat{\beta}_3 Female_i \cdot Age_i$

- Polynomial Terms

  e.g. $\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 Age_i + \hat{\beta}_2 Age_i^2$

- Collinearity
  - Highly correlated predictors
  - Perfectly correlated predictors

**To R**

**(Also refer to textbook, Chapter 3, Section 3 for more details!)**

# Intro to Classification

Classification refers to the process of predicting response variables that are qualitative (also often called categorical or discrete).

We will study approaches for classification in the case of binary response variables (response variables that have two possible values).

# Two Estimation Goals

Imagine we are trying to predict how a Senator will vote on a particular issue.

Let $\text{Yes}_i$ denote the *ith* Senator's vote, where:
$\text{Yes}_i = 1$ if Senator $i$ votes Yes
$\text{Yes}_i = 0$ if Senator $i$ votes No (or Abstains)

Let $\boldsymbol{x}_i$ denote a vector of predictor values for Senator $i$.

# Two Estimation Goals

Imagine we are trying to predict how a Senator will vote on a particular issue.

Let $\text{Yes}_i$ denote the *ith* Senator's vote, where:
$\text{Yes}_i = 1$ if Senator $i$ votes Yes
$\text{Yes}_i = 0$ if Senator $i$ votes No (or Abstains)

Let $\boldsymbol{x}_i$ denote a vector of predictor values for Senator $i$.

Two Quantities to Estimate:

# Two Estimation Goals

Imagine we are trying to predict how a Senator will vote on a particular issue.

Let $\text{Yes}_i$ denote the *ith* Senator's vote, where:
$\text{Yes}_i = 1$ if Senator $i$ votes Yes
$\text{Yes}_i = 0$ if Senator $i$ votes No (or Abstains)

Let $\boldsymbol{x}_i$ denote a vector of predictor values for Senator $i$.

Two Quantities to Estimate:

- Probability of voting yes: $\widehat{\Pr}(\text{Yes}_i = 1 | \boldsymbol{x}_i)$

# Two Estimation Goals

Imagine we are trying to predict how a Senator will vote on a particular issue.

Let $\text{Yes}_i$ denote the *ith* Senator's vote, where:
$\text{Yes}_i = 1$ if Senator $i$ votes Yes
$\text{Yes}_i = 0$ if Senator $i$ votes No (or Abstains)

Let $\boldsymbol{x}_i$ denote a vector of predictor values for Senator $i$.

Two Quantities to Estimate:

- Probability of voting yes: $\widehat{\Pr}(\text{Yes}_i = 1|\boldsymbol{x}_i)$
- Classification of vote: $\widehat{\text{Yes}_i} = I(\widehat{\Pr}(\text{Vote}_i = 1|\boldsymbol{x}_i) > t)$ , where $t$ is a threshold

  That is, if $\widehat{\Pr}(\text{Vote}_i = 1|\boldsymbol{x}_i) > t$, then $I(\widehat{\Pr}(\text{Vote}_i = 1|\boldsymbol{x}_i) > t) = 1$, otherwise 0.

# Linear Probability Model

$$\text{Yes}_i = \boldsymbol{\beta} \cdot \boldsymbol{x}_i + \epsilon_i$$

# Linear Probability Model

$$\text{Yes}_i = \boldsymbol{\beta} \cdot \boldsymbol{x}_i + \epsilon_i$$

The $\beta$'s can be estimated using the exact same process as before (OLS Regression), ignoring the fact that the outcome is binary.

# Linear Probability Model

$$\text{Yes}_i = \boldsymbol{\beta} \cdot \boldsymbol{x}_i + \epsilon_i$$

The $\beta$'s can be estimated using the exact same process as before (OLS Regression), ignoring the fact that the outcome is binary.

But now, the predicted/fitted values should be interpreted differently, as predicted probabilities:

$$\widehat{\text{Pr}}(\text{Yes}_i = 1 | \boldsymbol{X}_i) = \widehat{\boldsymbol{\beta}} \cdot \boldsymbol{x}_i$$

# Linear Probability Model

$$\text{Yes}_i = \boldsymbol{\beta} \cdot \boldsymbol{x}_i + \epsilon_i$$

The $\beta$'s can be estimated using the exact same process as before (OLS Regression), ignoring the fact that the outcome is binary.

But now, the predicted/fitted values should be interpreted differently, as predicted probabilities:

$$\widehat{\Pr}(\text{Yes}_i = 1 | \boldsymbol{X}_i) = \widehat{\boldsymbol{\beta}} \cdot \boldsymbol{x}_i$$

And classifications can be made as follows:

$$\widehat{\text{Yes}}_i = 1 \text{ if } \widehat{\boldsymbol{\beta}} \cdot \boldsymbol{x}_i > t$$
$$\widehat{\text{Yes}}_i = 0 \text{ if } \widehat{\boldsymbol{\beta}} \cdot \boldsymbol{x}_i \leq t$$

# (Potential) Problems with Linear Probabilty Model

- Probabilities greater than 1, less than 0
- Potentially implausible relationship between covariates and response