

# REGRESSION AND QUALITATIVE DATA

Data Analysis for Journalism and Political Communication  
(Spring 2026)

Prof. Bell

# WHAT IS REGRESSION?

- So far, we've learned how to compare the mean of two groups using a **t-test**

## WHAT IS REGRESSION?

- So far, we've learned how to compare the mean of two groups using a **t-test**
  - But often, a t-test is too restrictive for the analysis we want to conduct:

# WHAT IS REGRESSION?

- So far, we've learned how to compare the mean of two groups using a **t-test**
- But often, a t-test is too restrictive for the analysis we want to conduct:
  - ▶ What if we have two continuous variables? Income, age, and years of education are common variables that we may not want to force into two discrete categories.

# WHAT IS REGRESSION?

- So far, we've learned how to compare the mean of two groups using a **t-test**
- But often, a t-test is too restrictive for the analysis we want to conduct:
  - ▶ What if we have two continuous variables? Income, age, and years of education are common variables that we may not want to force into two discrete categories.
  - ▶ What about **confounders**?

# WHAT IS REGRESSION?

- So far, we've learned how to compare the mean of two groups using a **t-test**
- But often, a t-test is too restrictive for the analysis we want to conduct:
  - ▶ What if we have two continuous variables? Income, age, and years of education are common variables that we may not want to force into two discrete categories.
  - ▶ What about **confounders**?

## Confounder

A variable that explains change in both the independent and dependent variable.

# WHAT IS REGRESSION?

- So far, we've learned how to compare the mean of two groups using a **t-test**
- But often, a t-test is too restrictive for the analysis we want to conduct:
  - ▶ What if we have two continuous variables? Income, age, and years of education are common variables that we may not want to force into two discrete categories.
  - ▶ What about **confounders**?

## Confounder

A variable that explains change in both the independent and dependent variable.

- ▶ When we fail to account for confounders, we face **omitted variable bias** and our estimates are not accurate.

## IN-CLASS EXAMPLE

**Linear regression is just the best fit line through a scatter plot of two or more variables.**

# ESTIMATING THE REGRESSION (BEST FIT) LINE

The linear regression equation is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$Y$  = dependent variable

$\beta_0$  = intercept

$\beta_1$  = slope, also called a coefficient

$X$  = independent variable

$\epsilon$  = error

# ESTIMATING THE REGRESSION (BEST FIT) LINE

The linear regression equation is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$Y$  = dependent variable

$\beta_0$  = intercept

$\beta_1$  = slope, also called a coefficient

$X$  = independent variable

$\epsilon$  = error

This looks very similar to a linear equation you might have learned before:

$$y = ax + b$$

# ESTIMATING THE REGRESSION (BEST FIT) LINE

The linear regression equation is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$Y$  = dependent variable

$\beta_0$  = intercept

$\beta_1$  = slope, also called a coefficient

$X$  = independent variable

$\epsilon$  = error

This looks very similar to a linear equation you might have learned before:

$$y = ax + b$$

$$y = b + ax$$

## IN-CLASS EXAMPLE

**The lm() function calculates the intercept and slope for the linear regression equation.**

## IN-CLASS EXAMPLE

You want to estimate the effect of income on the allocation to welfare applicants. The linear regression equation is:

$$\text{Allocation} = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}$$

The  $\hat{\phantom{x}}$  is the mathematical notation for “estimate of the mean”

## IN-CLASS EXAMPLE

You want to estimate the effect of income on the allocation to welfare applicants. The linear regression equation is:

$$\hat{\text{Allocation}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}$$

The  $\hat{\phantom{x}}$  is the mathematical notation for “estimate of the mean”

If “Income” in thousands has a value of 100, and the intercept is \$605, and the coefficient is -.5, what is our estimated mean of “Allocation”?

## IN-CLASS EXAMPLE

You want to estimate the effect of income on the allocation to welfare applicants. The linear regression equation is:

$$\hat{\text{Allocation}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}$$

The  $\hat{\phantom{x}}$  is the mathematical notation for “estimate of the mean”

If “Income” in thousands has a value of 100, and the intercept is \$605, and the coefficient is -.5, what is our estimated mean of “Allocation”?

$$\hat{\text{Allocation}} = 605 + (-.5) * 100$$

## IN-CLASS EXAMPLE

You want to estimate the effect of income on the allocation to welfare applicants. The linear regression equation is:

$$\hat{\text{Allocation}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}$$

The  $\hat{\phantom{x}}$  is the mathematical notation for “estimate of the mean”

If “Income” in thousands has a value of 100, and the intercept is \$605, and the coefficient is -.5, what is our estimated mean of “Allocation”?

$$\hat{\text{Allocation}} = 605 + (-.5) * 100$$

$$\hat{\text{Allocation}} = 555$$

## IN-CLASS EXAMPLE

You want to estimate the effect of income on the allocation to welfare applicants. The linear regression equation is:

$$\hat{\text{Allocation}} = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}$$

The  $\hat{\phantom{x}}$  is the mathematical notation for “estimate of the mean”

If “Income” in thousands has a value of 100, and the intercept is \$605, and the coefficient is -.5, what is our estimated mean of “Allocation”?

$$\hat{\text{Allocation}} = 605 + (-.5) * 100$$

$$\hat{\text{Allocation}} = 555$$

So the coefficient ( $\beta_1$ ) is the effect of a **one-unit** change in income (thousands) on the mean allocation.

# WHAT IS ERROR?

Recall that the linear regression equation is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$Y$  = dependent variable

$\beta_0$  = intercept

$\beta_1$  = slope, also called a coefficient

$X$  = independent variable

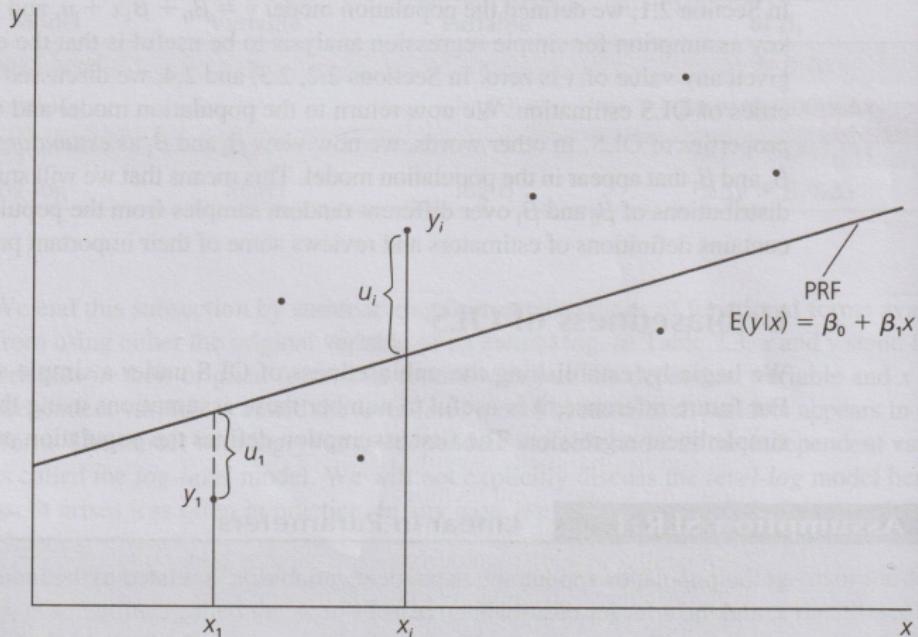
$\epsilon$  = error

# WHAT IS ERROR?

- Error ( $\epsilon$  or  $u$ ) is also called the residual (left over from  $Y = \beta_0 + \beta_1 X$ , our best fit line)

# WHAT IS ERROR?

FIGURE 2.7 Graph of  $y_i = \beta_0 + \beta_1 x_i + u_i$ .



## WHAT IS ERROR?

- Error ( $\epsilon$  or  $u$ ) is also called the residual (left over from  $Y = \beta_0 + \beta_1 X$ , our best fit line)
- Our goal in regression is to fit the best line that minimizes the error

## WHAT IS ERROR?

- Error ( $\epsilon$  or  $u$ ) is also called the residual (left over from  $Y = \beta_0 + \beta_1 X$ , our best fit line)
- Our goal in regression is to fit the best line that minimizes the error
- However, we can never get the  $\epsilon = 0$ , and often we don't even get close. We just do the best we can to make the "best" best fit line.

# MULTIPLE LINEAR REGRESSION

The multiple linear regression equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$Y$  = dependent variable

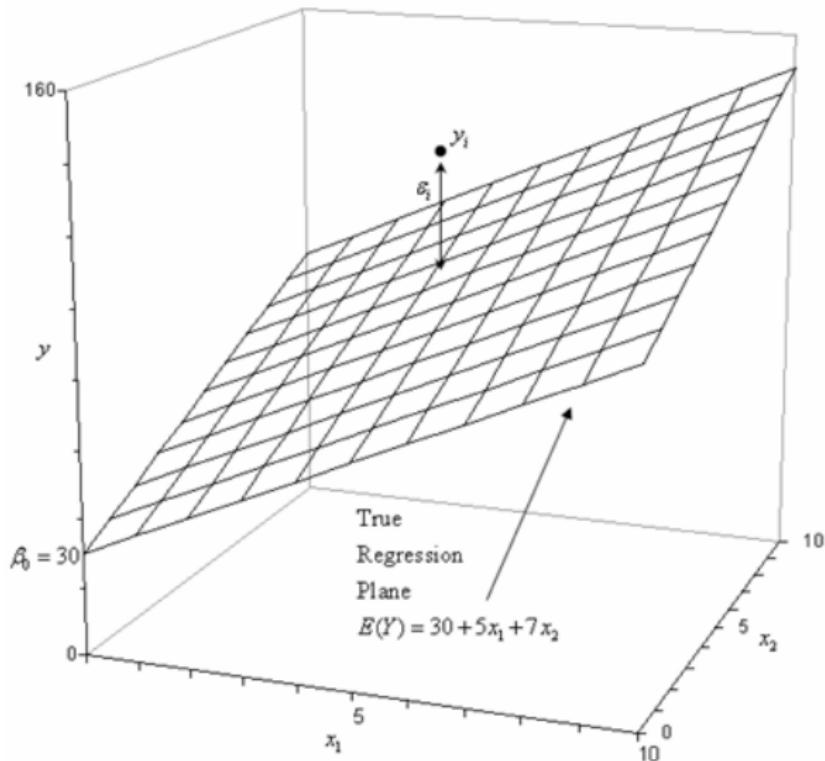
$\beta_0$  = intercept

$\beta_i$  = slope coefficient

$X_i$  = independent variable

$\epsilon$  = error

# MULTIPLE LINEAR REGRESSION



# MULTIPLE LINEAR REGRESSION

The multiple linear regression equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$Y$  = dependent variable

$\beta_0$  = intercept

$\beta_i$  = slope coefficient

$X_i$  = independent variable

$\epsilon$  = error

What is the interpretation of  $\beta_1$ ?

$\beta_1$  is the effect of a one-unit change in  $X_1$  on the mean of  $Y$ , holding  $X_2$  constant (the independent effect of  $X_1$  on  $Y$ )

# MULTIPLE LINEAR REGRESSION

The multiple linear regression equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$Y$  = dependent variable

$\beta_0$  = intercept

$\beta_i$  = slope coefficient

$X_i$  = independent variable

$\epsilon$  = error

What is the interpretation of  $\beta_2$ ?

$\beta_2$  is the effect of a one-unit change in  $X_2$  on the mean of  $Y$ , holding  $X_1$  constant (the independent effect of  $X_2$  on  $Y$ )

# NON-CONTINUOUS DEPENDENT VARIABLES

- One of the assumptions of linear regression is that the relationship between  $X$  and  $Y$  is linear (you can draw a best fit line)

## NON-CONTINUOUS DEPENDENT VARIABLES

- One of the assumptions of linear regression is that the relationship between  $X$  and  $Y$  is linear (you can draw a best fit line)
- This assumption is usually fine when we are working with continuous dependent variables like income

## NON-CONTINUOUS DEPENDENT VARIABLES

- One of the assumptions of linear regression is that the relationship between  $X$  and  $Y$  is linear (you can draw a best fit line)
- This assumption is usually fine when we are working with continuous dependent variables like income
- What about categorical dependent variables, like party ID? The linear regression model is not well suited for these.

## NON-CONTINUOUS DEPENDENT VARIABLES

- One of the assumptions of linear regression is that the relationship between  $X$  and  $Y$  is linear (you can draw a best fit line)
- This assumption is usually fine when we are working with continuous dependent variables like income
- What about categorical dependent variables, like party ID? The linear regression model is not well suited for these.
- What about binary dependent variables, like support for a policy?

We call this a linear probability model because it estimates the effect of a one-unit change in  $X$  on the *probability* (percent chance) of a value of “1” for the dependent variable

# IN-CLASS EXAMPLE

# QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count

# QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count
  - But what about data that isn't numbers? This is **qualitative data**

# QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count
  - But what about data that isn't numbers? This is **qualitative data**
  - There are at least twenty different types of qualitative data analysis, including several found in the social sciences:

# QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count
  - But what about data that isn't numbers? This is **qualitative data**
  - There are at least twenty different types of qualitative data analysis, including several found in the social sciences:
    - ▶ Case studies

# QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count
- But what about data that isn't numbers? This is **qualitative data**
- There are at least twenty different types of qualitative data analysis, including several found in the social sciences:
  - ▶ Case studies
  - ▶ Ethnography

# QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count
- But what about data that isn't numbers? This is **qualitative data**
- There are at least twenty different types of qualitative data analysis, including several found in the social sciences:
  - ▶ Case studies
  - ▶ Ethnography
  - ▶ Content analysis

## QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count
  - But what about data that isn't numbers? This is **qualitative data**
  - There are at least twenty different types of qualitative data analysis, including several found in the social sciences:
    - ▶ Case studies
    - ▶ Ethnography
    - ▶ Content analysis
    - ▶ Mixed-methods

# QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count
- But what about data that isn't numbers? This is **qualitative data**
- There are at least twenty different types of qualitative data analysis, including several found in the social sciences:
  - ▶ Case studies
  - ▶ Ethnography
  - ▶ Content analysis
  - ▶ Mixed-methods
- Our goal isn't to find a "best fit line," but to find patterns, themes, and meaning from "thick descriptions" (Geertz, 1973)

# QUALITATIVE DATA ANALYSIS

- So far, we've focused on quantitative data — things we can count
- But what about data that isn't numbers? This is **qualitative data**
- There are at least twenty different types of qualitative data analysis, including several found in the social sciences:
  - ▶ Case studies
  - ▶ Ethnography
  - ▶ Content analysis
  - ▶ Mixed-methods
- Our goal isn't to find a "best fit line," but to find patterns, themes, and meaning from "thick descriptions" (Geertz, 1973)
- We will focus on one type of qualitative analysis: text analysis

# TEXT ANALYSIS

- Think of:
  - ▶ Interview transcripts

# TEXT ANALYSIS

- Think of:
  - ▶ Interview transcripts
  - ▶ Open-ended survey responses

# TEXT ANALYSIS

- Think of:
  - ▶ Interview transcripts
  - ▶ Open-ended survey responses
  - ▶ News articles or social media posts

# TEXT ANALYSIS

- Think of:

- ▶ Interview transcripts
- ▶ Open-ended survey responses
- ▶ News articles or social media posts
- ▶ Focus group notes

# TEXT ANALYSIS

- Think of:
  - ▶ Interview transcripts
  - ▶ Open-ended survey responses
  - ▶ News articles or social media posts
  - ▶ Focus group notes
- The most common way to analyze qualitative data is **thematic coding**

# TEXT ANALYSIS

- Think of:
  - ▶ Interview transcripts
  - ▶ Open-ended survey responses
  - ▶ News articles or social media posts
  - ▶ Focus group notes
- The most common way to analyze qualitative data is **thematic coding**
- A **code** is just a descriptive label applied to a piece of text

# TEXT ANALYSIS

Well, that's one problem, that [my school is] pretty small, so 1 if you say one thing to one person, and then they decide to tell two people, then those two people tell two people, and in one period everybody else knows. 2 Everybody in the entire school knows that you said whatever it was. So. . . .

1 SPREADING RUMORS

2 KNOWING WHAT YOU SAID

# TEXT ANALYSIS

- Think of:
  - ▶ Interview transcripts
  - ▶ Open-ended survey responses
  - ▶ News articles or social media posts
  - ▶ Focus group notes
- The most common way to analyze qualitative data is **thematic coding**
- A **code** is just a descriptive label applied to a piece of text
- We can then quantify our codes: how frequently each code appears, in which parts of the text, from which speakers, etc.

# WHERE Do CODES COME FROM?

- **Deductive Coding:** You start with a list of codes *before* you read
  - ▶ This is most often used when the goal of the study is to count the appearances of a discrete set of themes, e.g., "How often does Donald Trump talk about trade in his speeches?"

# WHERE Do CODES COME FROM?

- **Deductive Coding:** You start with a list of codes *before* you read
  - ▶ This is most often used when the goal of the study is to count the appearances of a discrete set of themes, e.g., "How often does Donald Trump talk about trade in his speeches?"
- **Inductive Inductive:** You start *without* a list of codes. You read the data and let the themes emerge from the text
  - ▶ Many qualitative research studies do not begin with well-defined hypotheses, and researchers generate new ideas through the process of coding

# WHERE Do CODES COME FROM?

- **Deductive Coding:** You start with a list of codes *before* you read
  - ▶ This is most often used when the goal of the study is to count the appearances of a discrete set of themes, e.g., "How often does Donald Trump talk about trade in his speeches?"
- **Inductive Inductive:** You start *without* a list of codes. You read the data and let the themes emerge from the text
  - ▶ Many qualitative research studies do not begin with well-defined hypotheses, and researchers generate new ideas through the process of coding
- Most studies begin with a short list of deductive codes, then refine those codes or add new ones as you explore the data

## IN-CLASS EXERCISE

### ***From StoryCorps:***

In February of 2012, Jamal Faison was a 20-year-old college sophomore home on school break in New York City when he, along with two others, were arrested for attempting to steal mobile devices from a subway rider. Transit police arrested Jamal and he spent the next eight months on Rikers Island — New York City's massive main jail complex.

In September 2012, Jamal pleaded guilty to grand larceny and attempted robbery charges and a month later was released from custody. Jamal came to StoryCorps to remember the night he was released from Rikers, and to discuss his relationship with his uncle, Born.

# IN-CLASS EXERCISE

## WHAT IF WE HAVE TOO MUCH TEXT?

- Thematic coding by hand is great for 20 interviews, but what if we have 20,000 news articles? 2 million tweets?

## WHAT IF WE HAVE TOO MUCH TEXT?

- Thematic coding by hand is great for 20 interviews, but what if we have 20,000 news articles? 2 million tweets?
- This is where computational text analysis comes in

## WHAT IF WE HAVE TOO MUCH TEXT?

- Thematic coding by hand is great for 20 interviews, but what if we have 20,000 news articles? 2 million tweets?
- This is where computational text analysis comes in
- When working with computational text analysis, we call the full set of text and documents the **corpus**

## WHAT IF WE HAVE TOO MUCH TEXT?

- Thematic coding by hand is great for 20 interviews, but what if we have 20,000 news articles? 2 million tweets?
- This is where computational text analysis comes in
- When working with computational text analysis, we call the full set of text and documents the **corpus**
- The techniques we will learn in this class are called “bag of words” techniques, because unlike an LLM, they are agnostic about the order of the words

## WHAT IF WE HAVE TOO MUCH TEXT?

- Thematic coding by hand is great for 20 interviews, but what if we have 20,000 news articles? 2 million tweets?
- This is where computational text analysis comes in
- When working with computational text analysis, we call the full set of text and documents the **corpus**
- The techniques we will learn in this class are called “bag of words” techniques, because unlike an LLM, they are agnostic about the order of the words
  - ▶ As if you are putting all the words into a bag, drawing them out at random, and seeing how frequently they occur

# LDA TOPIC MODELING

- **LDA** (Latent Dirichlet Allocation) is a popular method

# LDA TOPIC MODELING

- **LDA** (Latent Dirichlet Allocation) is a popular method
- It is an unsupervised method: you don't give it a list of codes. It finds the themes (called "topics") made up of words that tend to appear together frequently

# LDA TOPIC MODELING

- **LDA** (Latent Dirichlet Allocation) is a popular method
- It is an unsupervised method: you don't give it a list of codes. It finds the themes (called "topics") made up of words that tend to appear together frequently
- The researcher must decide what the semantic meaning of these topics (correlated words) mean

# RESEARCHER CHOICES IN LDA

- 1 How many topics?

# RESEARCHER CHOICES IN LDA

## ① How many topics?

- ▶ You must pre-define how many topics to create

# RESEARCHER CHOICES IN LDA

## ① How many topics?

- ▶ You must pre-define how many topics to create
- ▶ Too few: The topics might be too broad and meaningless

# RESEARCHER CHOICES IN LDA

## ① How many topics?

- ▶ You must pre-define how many topics to create
- ▶ Too few: The topics might be too broad and meaningless
- ▶ Too many: The topics might be too specific, overlap, or be junk

# RESEARCHER CHOICES IN LDA

## ① How many topics?

- ▶ You must pre-define how many topics to create
- ▶ Too few: The topics might be too broad and meaningless
- ▶ Too many: The topics might be too specific, overlap, or be junk
- ▶ This is a critical choice with no single right answer

# RESEARCHER CHOICES IN LDA

## ① How many topics?

- ▶ You must pre-define how many topics to create
- ▶ Too few: The topics might be too broad and meaningless
- ▶ Too many: The topics might be too specific, overlap, or be junk
- ▶ This is a critical choice with no single right answer

## ② What do the topics mean?

# RESEARCHER CHOICES IN LDA

## ① How many topics?

- ▶ You must pre-define how many topics to create
- ▶ Too few: The topics might be too broad and meaningless
- ▶ Too many: The topics might be too specific, overlap, or be junk
- ▶ This is a critical choice with no single right answer

## ② What do the topics mean?

- ▶ LDA just gives you a list of words: "economy," "jobs," "growth..."

# RESEARCHER CHOICES IN LDA

## ① How many topics?

- ▶ You must pre-define how many topics to create
- ▶ Too few: The topics might be too broad and meaningless
- ▶ Too many: The topics might be too specific, overlap, or be junk
- ▶ This is a critical choice with no single right answer

## ② What do the topics mean?

- ▶ LDA just gives you a list of words: “economy,” “jobs,” “growth...”
- ▶ Is that topic “The Economy”? Or “Labor Market Reports”? Or “Political Talking Points”?

# RESEARCHER CHOICES IN LDA

## ① How many topics?

- ▶ You must pre-define how many topics to create
- ▶ Too few: The topics might be too broad and meaningless
- ▶ Too many: The topics might be too specific, overlap, or be junk
- ▶ This is a critical choice with no single right answer

## ② What do the topics mean?

- ▶ LDA just gives you a list of words: “economy,” “jobs,” “growth...”
- ▶ Is that topic “The Economy”? Or “Labor Market Reports”? Or “Political Talking Points”?
- ▶ The researcher must interpret and label the topic. This is a subjective, human step!

# IN-CLASS EXERCISE

- Data at: Kaggle.com
- Tool: Voyant