Regression

Data Analysis for Journalism and Political Communication (Fall 2025)

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**? For example, we found that Americans allocate less money to welfare applicants who are rated as "poor" workers compared to "excellent" workers. What are some possible confounders we need to consider?

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**? For example, we found that Americans allocate less money to welfare applicants who are rated as "poor" workers compared to "excellent" workers. What are some possible confounders we need to consider?
- **Regression** is a set of tools that allow us to efficiently evaluate the correlation between two variables while accounting for potential confounders

Introduction to Linear Regression

- We will focus on the simplest regression method called Ordinary Least Squares (OLS), or **linear regression**

Introduction to Linear Regression

- We will focus on the simplest regression method called Ordinary Least Squares (OLS), or **linear regression**
- Linear regression is used to estimate the effect of a change in the **independent** (explanatory) variable on the mean of the **dependent** (outcome) variable

INTRODUCTION TO LINEAR REGRESSION

- We will focus on the simplest regression method called Ordinary Least Squares (OLS), or **linear regression**
- Linear regression is used to estimate the effect of a change in the **independent** (explanatory) variable on the mean of the **dependent** (outcome) variable
- Why is it called "linear" regression? In practice, we are just running a best fit line through a scatter plot of two variables.

Estimating the Regression (Best Fit) Line

The linear regression equation is:

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

Y = dependent variable
β_0 = intercept
β_1 = slope, also called a coefficient
X = independent variable
ϵ = error

Estimating the Regression (Best Fit) Line

The linear regression equation is:

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

Y = dependent variable
β_0 = intercept
β_1 = slope, also called a coefficient
X = independent variable
ϵ = 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
β_0 = intercept
β_1 = slope, also called a coefficient
X = independent variable
ϵ = error

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

$$y = ax + b$$
$$y = b + ax$$

In-class exercise

Example

"Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America's Deserving Poor"
Christopher D. DeSante, *Am. Journal of Political Science* vol. 57 iss. 2 (2013)

WORK FIRST ASSISTANCE APPLICATION

Applicant Name: Latoya [redacted] Date of Application: [redacted]
Address: [redacted] Telephone: [redacted]
County: [redacted]
Case No.: [redacted] District No.: [redacted]

HOUSEHOLD: List all household members for whom Assistance is being requested:
(Non-applicant household members are not required to provide a social security number, immigrant or citizenship status)

Name	Date of Birth	Sex	Social Security No.	Citizen/ eligible immigrant	Relationship
[redacted]	08/16/1998	M	[redacted]	Y	Son
[redacted]	04/14/2001	F	[redacted]	Y	Daughter

Does the household include a child who meets the Work First age rule? ☒ Yes ☐ No
Is the child living with an adult who meets the Work First kinship rule? ☒ Yes ☐ No
Has anyone listed on the EA Application ever received EA? ☐ Yes When: ______ ☒ No
Does anyone live in the home that is not listed on the EA Application? ☐ Yes ☒ No
If yes, is the individual(s) a roomer/boarder? ☐ Yes ☐ No

[redacted]

Total assessed monthly need: $ 900.00

APPLICANT 1 Worker Quality Assessment (circle one): Poor Average Excellent

Applicant Statement: I understand that it is against the law for me to make false statements and that I am subject to prosecution if I do. I certify that the information I have provided is a true and complete statement of facts according to my best knowledge and belief. I certify, under penalty of perjury, that all persons for whom I am applying are U.S. citizens or qualified immigrants. I declare under penalty of perjury (and being subject to prosecution under 28 U. S. C. § 1746) that the foregoing is true and correct. I give the agency permission to verify any information necessary to determine my eligibility for Emergency Assistance.

Chris D. DeSante

Witness's Signature Applicant's/Representative's Signature Date

WORK FIRST ASSISTANCE APPLICATION

Applicant Name: Keisha [redacted] Date of Application: [redacted]
Address: [redacted] Telephone: [redacted]
County: [redacted]
Case No.: [redacted] District No.: [redacted]

HOUSEHOLD: List all household members for whom Assistance is being requested:
(Non-applicant household members are not required to provide a social security number, immigrant or citizenship status)

Name	Date of Birth	Sex	Social Security No.	Citizen/ eligible immigrant	Relationship
[redacted]	05/07/2005	M	[redacted]	Y	Son
[redacted]	03/20/2007	F	[redacted]	Y	Daughter

Does the household include a child who meets the Work First age rule? ☒ Yes ☐ No
Is the child living with an adult who meets the Work First kinship rule? ☒ Yes ☐ No
Has anyone listed on the EA Application ever received EA? ☐ Yes When: ______ ☒ No
Does anyone live in the home that is not listed on the EA Application? ☐ Yes ☒ No
If yes, is the individual(s) a roomer/boarder? ☐ Yes ☐ No

[redacted]

Total assessed monthly need: $ 900.00

APPLICANT 2 Worker Quality Assessment (circle one): Poor Average Excellent

Applicant Statement: I understand that it is against the law for me to make false statements and that I am subject to prosecution if I do. I certify that the information I have provided is a true and complete statement of facts according to my best knowledge and belief. I certify, under penalty of perjury, that all persons for whom I am applying are U.S. citizens or qualified immigrants. I declare under penalty of perjury (and being subject to prosecution under 28 U. S. C. § 1746) that the foregoing is true and correct. I give the agency permission to verify any information necessary to determine my eligibility for Emergency Assistance.

Chris D. DeSante

Witness's Signature Applicant's/Representative's Signature Date

Example

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

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

The $\hat{}$ is the mathematical notation for "estimate of the mean"

Example

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

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

The $\hat{}$ 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”?

Example

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

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

The $\hat{}$ 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"?

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

EXAMPLE

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

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

The $\hat{}$ 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"?

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

$$\widehat{Allocation} = 555$$

Example

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

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

The $\hat{}$ 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"?

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

$$\widehat{Allocation} = 555$$

So the coefficient (β_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:

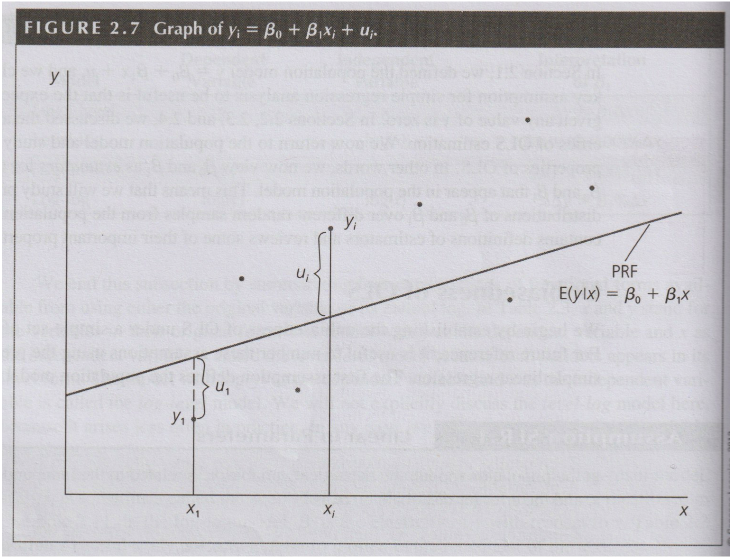
$$Y = \beta_0 + \beta_1 X + \epsilon$$

Y = dependent variable
β_0 = intercept
β_1 = slope, also called a coefficient
X = independent variable
ϵ = error

What Is Error?

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

WHAT IS ERROR?



What Is Error?

- Error (ϵ 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 (ϵ 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 ϵ = 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
β_0 = intercept
β_i = slope coefficient
X_i = independent variable
ϵ = error

Multiple Linear Regression

The multiple linear regression equation is:

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

Y = dependent variable
β_0 = intercept
β_i = slope coefficient
X_i = independent variable
ϵ = error

What is the interpretation of β_1?

β_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:

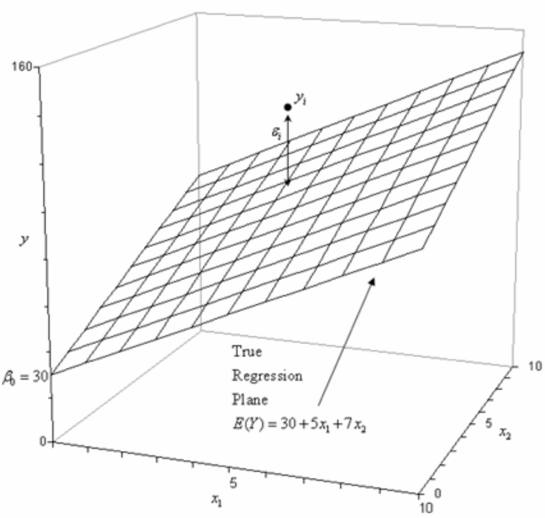
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Y = dependent variable
β_0 = intercept
β_i = slope coefficient
X_i = independent variable
ϵ = error

What is the interpretation of β_2?

β_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)

Multiple Linear Regression



Example

$$\hat{Allocation} = \hat{\beta}_0 + \hat{\beta}_1 Race + \hat{\beta}_2 WorkEthic + \hat{\beta}_3 Age + \hat{\beta}_4 Gender + \hat{\beta}_5 PoliticalParty + \hat{\beta}_6 Income$$

How Many Independent Variables?

- Our goal is to include all of the independent variables that could plausibly affect the dependent variable

How Many Independent Variables?

- Our goal is to include all of the independent variables that could plausibly affect the dependent variable
- But we do not include all possible variables (the kitchen sink method) because it makes our regression mathematically less efficient

How Many Independent Variables?

- Our goal is to include all of the independent variables that could plausibly affect the dependent variable
- But we do not include all possible variables (the kitchen sink method) because it makes our regression mathematically less efficient
- One way to evaluate how well you are explaining change in Y is through the R^2, technically called the "coefficient of determination"

How Many Independent Variables?

- Our goal is to include all of the independent variables that could plausibly affect the dependent variable
- But we do not include all possible variables (the kitchen sink method) because it makes our regression mathematically less efficient
- One way to evaluate how well you are explaining change in Y is through the R^2, technically called the “coefficient of determination”
- R^2 ranges from 0 to 1 and represents the proportion of change in Y explained by X_i

How Many Independent Variables?

- Our goal is to include all of the independent variables that could plausibly affect the dependent variable
- But we do not include all possible variables (the kitchen sink method) because it makes our regression mathematically less efficient
- One way to evaluate how well you are explaining change in Y is through the R^2, technically called the “coefficient of determination”
- R^2 ranges from 0 to 1 and represents the proportion of change in Y explained by X_i
- My pet peeve is over-relying on the R^2 - theory should drive your modeling decisions, not a summary statistic

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