Regression and prediction: OLS

Section 4

Sima Biondi

Spring 2025

Gov 51: Data Analysis and Politics

Overview

1. Regression: basics
2. Regression: estimation

Housekeeping

- Problem Set I: almost done! Due tonight @ 11:59pm

Housekeeping

- Problem Set I: almost done! Due tonight @ 11:59pm
- CA office hours: great for coding questions

Housekeeping

- Problem Set I: almost done! Due tonight @ 11:59pm
- CA office hours: great for coding questions
  - ↪ Pranav: walk-in hours

Housekeeping

- Problem Set I: almost done! Due tonight @ 11:59pm
- CA office hours: great for coding questions
  - ↪ Pranav: walk-in hours
  - ↪ Ben: `https://calendly.com/bheilbronn-college/ben-gov51-oh`

Roadmap

Last week: matching to improve our estimates

Roadmap

Last week: matching to improve our estimates

$\rightarrow$ Assumptions: probabilistic treatment, ignorability, SUTVA

Roadmap

Last week: matching to improve our estimates

- $\rightarrow$ Assumptions: probabilistic treatment, ignorability, SUTVA
- $\rightarrow$ When does matching work? If $X_i \approx X_i^M$ then we expect $Y_i \approx Y_i^M$

Roadmap

Last week: matching to improve our estimates

- → Assumptions: probabilistic treatment, ignorability, SUTVA
- → When does matching work? If $X_i \approx X_i^M$ then we expect $Y_i \approx Y_i^M$
- → Different ways to match: (a) various distance algorithms, (b) 1-to-1 or 1-to-many

Roadmap

Last week: matching to improve our estimates

- $\rightarrow$ Assumptions: probabilistic treatment, ignorability, SUTVA
- $\rightarrow$ When does matching work? If $X_i \approx X_i^M$ then we expect $Y_i \approx Y_i^M$
- $\rightarrow$ Different ways to match: (a) various distance algorithms, (b) 1-to-1 or 1-to-many

Roadmap

Last week: matching to improve our estimates

- $\rightarrow$ Assumptions: probabilistic treatment, ignorability, SUTVA
- $\rightarrow$ When does matching work? If $X_i \approx X_i^M$ then we expect $Y_i \approx Y_i^M$
- $\rightarrow$ Different ways to match: (a) various distance algorithms, (b) 1-to-1 or 1-to-many

In this section, we look at the nuts and bolts of actual estimation: our old friend **linear regression**

Overview

1. Regression: basics

2. Regression: estimation

What is linear regression?

Linear regression estimates:

1. size, and

of **the relationship** between an independent variable and a dependent variable.

What is linear regression?

Linear regression estimates:

1. size, and
2. direction

of **the relationship** between an independent variable and a dependent variable.

Terminology and notation

Lots of synonyms/aliases/pen names!

- Independent variable(s):

Terminology and notation

Lots of synonyms/aliases/pen names!

- Independent variable(s):
  - $\rightarrow$ *a.k.a.* explanatory variable(s)

Terminology and notation

Lots of synonyms/aliases/pen names!

- Independent variable(s):
  - $\rightarrow$ *a.k.a.* explanatory variable(s)
  - $\rightarrow$ *a.k.a.* covariate(s)

Terminology and notation

Lots of synonyms/aliases/pen names!

- Independent variable(s):
  - $\rightarrow$ *a.k.a.* explanatory variable(s)
  - $\rightarrow$ *a.k.a.* covariate(s)
  - $\Rightarrow$ *Notation*: X_i

Terminology and notation

Lots of synonyms/aliases/pen names!

- Independent variable(s):
  - $\rightarrow$ *a.k.a.* explanatory variable(s)
  - $\rightarrow$ *a.k.a.* covariate(s)
  - $\Rightarrow$ *Notation*: X_i
- Dependent variable:

Terminology and notation

Lots of synonyms/aliases/pen names!

- Independent variable(s):
  - $\rightarrow$ *a.k.a.* explanatory variable(s)
  - $\rightarrow$ *a.k.a.* covariate(s)
  - $\Rightarrow$ *Notation*: X_i
- Dependent variable:
  - $\rightarrow$ *a.k.a.* outcome

Terminology and notation

Lots of synonyms/aliases/pen names!

- Independent variable(s):
  - $\rightarrow$ *a.k.a.* explanatory variable(s)
  - $\rightarrow$ *a.k.a.* covariate(s)
  - $\Rightarrow$ *Notation*: X_i
- Dependent variable:
  - $\rightarrow$ *a.k.a.* outcome
  - $\Rightarrow$ *Notation*: Y_i

Notation

- Gov 50 notation: $Y_i = \alpha + \beta X_i + \epsilon_i$

Notation

- Gov 50 notation: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Gov 51 notation:

Notation

- Gov 50 notation: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Gov 51 notation:
  - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow$

Notation

- Gov 50 notation: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Gov 51 notation:
    - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow$

Notation

- Gov 50 notation: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Gov 51 notation:
  - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow$ "assumed model"
  - $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \rightarrow$

Notation

- Gov 50 notation: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Gov 51 notation:
    - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow$ "assumed model"
    - $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \rightarrow$

Notation

- Gov 50 notation: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Gov 51 notation:
  - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow$ "assumed model"
  - $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \rightarrow$ "fitted model"

Notation

- Gov 50 notation: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Gov 51 notation:
  - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow$ "assumed model"
  - $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \rightarrow$ "fitted model"

What part of the notation describes the relationship between our variables?

Notation

- Gov 50 notation: $Y_i = \alpha + \beta\, X_i + \epsilon_i$
- Gov 51 notation:
    - $Y_i = \beta_0 + \beta_1\, X_i + \epsilon_i \rightarrow$ "assumed model"
    - $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1\, X_i \rightarrow$ "fitted model"

What part of the notation describes the relationship between our variables?

β_0 and $\beta_1 \Rightarrow \quad Y_i = \beta_0 + \beta_1\, X_i + \epsilon_i$

Overview

1 Regression: basics

2 Regression: estimation

Estimation

If β_0 and β_1 help us describe the relationship between Y_i and X_i, how do we choose β_0 and β_1?

Estimation

If β_0 and β_1 help us describe the relationship between Y_i and X_i, how do we choose β_0 and β_1?

Loss functions! Picking the right loss function matters because it's the metric by which we decide what line fits our data best.

Estimation

Loss functions options:

- Ordinary Least Squares (OLS)

Estimation

Loss functions options:

- Ordinary Least Squares (OLS)
- Absolute deviation

Estimation

Loss functions options:

- Ordinary Least Squares (OLS)
- Absolute deviation
- Penalized Least Squares

Estimation: OLS

OLS is **BLUE** conditional on assumptions:

- ⋆ Best

Estimation: OLS

OLS is **BLUE** conditional on assumptions:

- ⋆ Best
- ⋆ Linear

Estimation: OLS

OLS is **BLUE** conditional on assumptions:

- ⋆ Best
- ⋆ Linear
- ⋆ Unbiased

Estimation: OLS

OLS is **BLUE** conditional on assumptions:

- ⋆ Best
- ⋆ Linear
- ⋆ Unbiased
- ⋆ Estimator

Estimation: OLS

OLS is **BLUE** conditional on assumptions:

- ⋆ Best
- ⋆ Linear
- ⋆ Unbiased
- ⋆ Estimator

Estimation: OLS

OLS is **BLUE** conditional on assumptions:

- ⋆ Best
- ⋆ Linear
- ⋆ Unbiased
- ⋆ Estimator

Assumptions: (a) linearity, (b) independence, (c) homoscedasticity, (d) no (perfect) multicollinearity, (e) zero conditional mean of errors

Limits of OLS

- Using statistical packages, we can estimate linear regressions on any two variables

Limits of OLS

- Using statistical packages, we can estimate linear regressions on any two variables
  - BUT! not all regression are useful regressions

Limits of OLS

- Using statistical packages, we can estimate linear regressions on any two variables
    - BUT! not all regression are useful regressions
- Significant coefficients are NOT causal estimates without identification

Multi-variate regression

- Why would we want to add more variables?

Multi-variate regression

- Why would we want to add more variables?
    1. Better prediction

Multi-variate regression

- Why would we want to add more variables?
    1. Better prediction
    2. Easier interpretation: we intuitively understand what an effect size is if we hold other variables constant

Multi-variate regression

- Why would we want to add more variables?
    1. Better prediction
    2. Easier interpretation: we intuitively understand what an effect size is if we hold other variables constant
- Example: What effect does incumbency have on reelection, conditional on scandals?

Multi-variate regression

- Why would we want to add more variables?
    1. Better prediction
    2. Easier interpretation: we intuitively understand what an effect size is if we hold other variables constant
- **Example**: What effect does incumbency have on reelection, conditional on scandals?
    - We might want to know the incumbency advantage holding scandals constant

Multi-variate regression implementation

Implementation is similar to bivariate regression:

- Bivariate:

$$\beta_1 = \arg\min_{\beta_1} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 =$$

$$\arg\min_{\beta_1} \sum_{i=1}^{N} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Multi-variate regression implementation

Implementation is similar to bivariate regression:

- Bivariate:

$$\beta_1 = \arg\min_{\beta_1} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 =$$

$$\arg\min_{\beta_1} \sum_{i=1}^{N} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

- Multivariate:

$$\beta_1 = \arg\min_{\beta_1} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 =$$

$$\arg\min_{\beta_1} \sum_{i=1}^{N} (Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2))^2$$

Keep in mind

Coding:

- Package to produce tables in R: modelsummary

Conceptually:

Keep in mind

Coding:

- Package to produce tables in R: modelsummary

Conceptually:

- Hypothesis testing and β's as random variables

Keep in mind

Coding:

- Package to produce tables in R: modelsummary

Conceptually:

- Hypothesis testing and β's as random variables
  - Statistical significance

Keep in mind

Coding:

- Package to produce tables in R: modelsummary

Conceptually:

- Hypothesis testing and β's as random variables
  - Statistical significance
  - Interpreting p-values

Penalized regression

Address the problem of **over-fitting**

- Over-fitting: a model that begins to describe the error in the data rather than relationships between variables

Penalized regression

Address the problem of **over-fitting**

- Over-fitting: a model that begins to describe the error in the data rather than relationships between variables
  - In other words, our model may have high internal validity, but it has extremely low external validity

Penalized regression

Address the problem of **over-fitting**

- Over-fitting: a model that begins to describe the error in the data rather than relationships between variables
    - In other words, our model may have high internal validity, but it has extremely low external validity
    - We capture the relationships that are specific to our dataset and only your dataset,

Penalized regression

Address the problem of **over-fitting**

- Over-fitting: a model that begins to describe the error in the data rather than relationships between variables
    - In other words, our model may have high internal validity, but it has extremely low external validity
    - We capture the relationships that are specific to our dataset and only your dataset,

Penalized regression

Address the problem of **over-fitting**

- Over-fitting: a model that begins to describe the error in the data rather than relationships between variables
  - In other words, our model may have high internal validity, but it has extremely low external validity
  - We capture the relationships that are specific to our dataset and only your dataset,which (usually) isn't the goal!
- Can be particularly problematic if our variables are collinear

Penalized regression

Address the problem of **over-fitting**

- Over-fitting: a model that begins to describe the error in the data rather than relationships between variables
  - In other words, our model may have high internal validity, but it has extremely low external validity
  - We capture the relationships that are specific to our dataset and only your dataset,which (usually) isn't the goal!
- Can be particularly problematic if our variables are collinear
  - Collinearity refers to high correlation between covariates

Penalized regression

Address the problem of **over-fitting**

- Over-fitting: a model that begins to describe the error in the data rather than relationships between variables
  - In other words, our model may have high internal validity, but it has extremely low external validity
  - We capture the relationships that are specific to our dataset and only your dataset,which (usually) isn't the goal!
- Can be particularly problematic if our variables are collinear
  - Collinearity refers to high correlation between covariates
  - Makes it difficult to interpret our results!

Next week

- Upcoming: problem set due tonight!

Next week

- Upcoming: problem set due tonight!
- Questions about uncertainty and inference? Come to office hours!