Data, Environment and Society: Lecture 9: Intro to regression

Instructor: Duncan Callaway
GSI: Salma Elmallah

September 26, 2019

Announcements

Today

- Review bias-variance tradeoff
- Regression
  - K-nearest neighbors
  - Linear least squares

Reading

- Today's lecture draws from DS100 Ch10, ISLR Ch 2, ISLR Ch 3.1
- For next week
  - Read Alstone *et al* for next Tuesday – in class discussion
  - Review ISLR Ch 3.1-3.2

Before moving on, a little linear algebra:

Here are two vectors:

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Then the “dot” product of the two vectors is

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2$$

Next, a little more linear algebra:

We can also multiply *matrices* and vectors. Matrices are like column vectors stacked side by side

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Then matrix multiplication gives us

$$\mathbf{Ab} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} a_{11}b_1 + a_{12}b_2 \\ a_{21}b_1 + a_{22}b_2 \end{bmatrix}$$

Next, a little more linear algebra:

We can also multiply *matrices* and vectors. Matrices are like column vectors stacked side by side

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Then matrix multiplication gives us

$$\mathbf{Ab} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} a_{11}b_1 + a_{12}b_2 \\ a_{21}b_1 + a_{22}b_2 \end{bmatrix}$$

Each element of the resulting matrix (or vector) is the dot product of a row of the first term (**A**) and a column of the second (**b**)

Therefore: the horizontal "dimension" of the first must be the same as the vertical "dimension" of the second.

Let's define matrices for our data:

Suppose we have n observations, (x_i, y_i). We'll arrange them all into a matrix form:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Note: when we start working with more than one independent variable, X will have a new column for each new variable.

And then a lot more linear algebra:

Let's define the 'transpose':

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \Rightarrow \quad X^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}$$

And then a lot more linear algebra:

Let's define the 'transpose':

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \Rightarrow \quad X^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}$$

Now a challenge question: what's the product of these two matrices:

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Product of a matrix and its transpose

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Product of a matrix and its transpose

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$= \begin{bmatrix} \text{1st row dot 1st col} & \text{1st row dot 2nd col} \\ \text{2nd row dot 1st col} & \text{2nd row dot 2nd col} \end{bmatrix}$$

Product of a matrix and its transpose

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$= \begin{bmatrix} \text{1st row dot 1st col} & \text{1st row dot 2nd col} \\ \text{2nd row dot 1st col} & \text{2nd row dot 2nd col} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} 1 \cdot 1 & \sum_{i=1}^{n} 1 \cdot x_i \\ \sum_{i=1}^{n} 1 \cdot x_i & \sum_{i=1}^{n} x_i \cdot x_i \end{bmatrix}$$

Product of a matrix and its transpose

$$X^T X = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$= \begin{bmatrix} \text{1st row dot 1st col} & \text{1st row dot 2nd col} \\ \text{2nd row dot 1st col} & \text{2nd row dot 2nd col} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} 1 \cdot 1 & \sum_{i=1}^{n} 1 \cdot x_i \\ \sum_{i=1}^{n} 1 \cdot x_i & \sum_{i=1}^{n} x_i \cdot x_i \end{bmatrix} \quad = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Doing linear algebra in numpy:

See the in-class workbook!

Finally, the “normal equations”

We showed a way to compute β coefficients individually a few slides ago.

Finally, the "normal equations"

We showed a way to compute β coefficients individually a few slides ago.

However that can get tedious if you're doing *multiple* linear regression – i.e. if you have more than one independent variable.

The so-called "normal equations" give a nice, compact form to get the parameters.

$$\Theta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} X^T Y$$

$$= \left(\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

A note for computing and linear algebra geeks

The normal equations are an efficient way to solve the least squares linear regression problem *when the number of independent variables is relatively small.*

But! Inverting a matrix (the $(\cdot)^{-1}$ part) is a heavy computational lift – especially as the size of the matrix gets big.

Later in the semester we'll talk about an alternative approach, called "gradient descent",

- It searches for the optimal point on the cost function in a more manual way.
- But it's actually faster than getting the solution using the normal equations.

Unbiased estimators

If certain conditions (to be covered thursday) are met, then the β values are *unbiased*.

What does that mean?

Unbiased estimators

If certain conditions (to be covered thursday) are met, then the β values are *unbiased*.

What does that mean?

It means that the β estimates you'd get from repeatedly sampling the population will equal, **on average**, the true β values.

Variance of the sample mean?

First, review:

- Population: all possible realizations of a data generating process.
- Sample: the subset of the population that you *observe*.

Define:

- μ = population mean
- $\hat{\mu}_i$ = sample mean.

i indexes the sample.

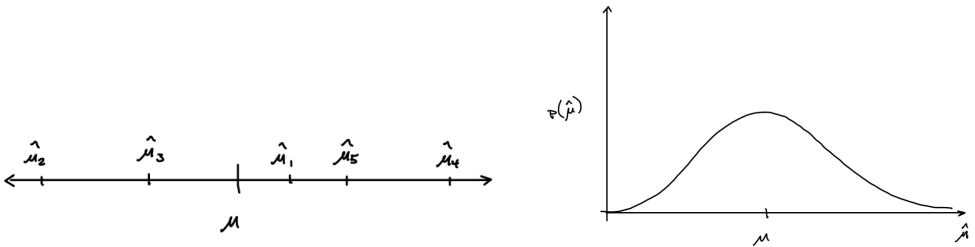
- Suppose your population is all countries in the world
- Randomly sample 20 of them.
  - First random sample of 20 $\rightarrow i = 1$
  - Second random sample of 20 $\rightarrow i = 2$
  - etc

Distribution of means

Suppose you're drawing many different samples from a population. What happens to the means?

Distribution of means

Suppose you're drawing many different samples from a population. What happens to the means?



You get many different values, and in general they will be normally distributed.

Standard error of the mean

If the sampling process is *unbiased*:

$$\text{avg}(\hat{\mu}) - \mu = 0$$

$$\text{var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

Standard error of the mean

If the sampling process is *unbiased*:

$$\text{avg}(\hat{\mu}) - \mu = 0$$

$$\text{var}(\hat{\mu}) = \frac{\sigma^2}{n} \equiv \text{SE}(\hat{\mu})^2$$

σ is the variance of ϵ, i.e. the changes in *y* that are not correlated with *x across the entire population*.

Population variance

Of course we rarely have the population variance.

- We don't usually know the true model
- We don't usually sample the whole population

Instead we use

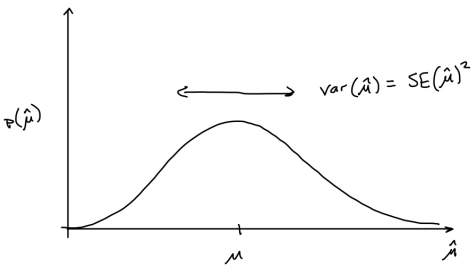
$$\hat{\text{SE}}(\hat{\mu})^2 = \hat{\sigma}^2 \frac{1}{n} = \frac{\text{RSS}}{(n-1)} \frac{1}{n}$$

How do we interpret the standard error of the mean?

In words: it is an estimate of the variance of the sample means, if we were to repeatedly sample.

How do we interpret the standard error of the mean?

In words: it is an estimate of the variance of the sample means, if we were to repeatedly sample.



This will be really useful in constructing “confidence intervals”, in just a few slides.

Ordinary least squares coefficients

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

We can think of the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ in the same conceptual terms as the sample means.

$$\text{avg}(\hat{\beta}_0) - \beta_0 = 0 \quad \text{(unbiased)}$$

$$\text{SE}(\hat{\beta}_0)^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

$$\text{avg}(\hat{\beta}_1) - \beta_1 = 0 \quad \text{(unbiased)}$$

$$\text{SE}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Confidence intervals

For a normal distribution:

$$\text{mean} \pm 2(\text{standard deviation}) = \mu \pm 2\sigma$$

is...

Confidence intervals

For a normal distribution:

$$\text{mean} \pm 2(\text{standard deviation}) = \mu \pm 2\sigma$$

is...the region containing 95% of the probability mass in the distribution.

Therefore the 95% "confidence intervals" are

$$\hat{\beta}_0 \pm 2\text{SE}(\hat{\beta}_0)$$
$$\hat{\beta}_1 \pm 2\text{SE}(\hat{\beta}_1)$$

If certain conditions are met (we'll cover Thursday) then

How to interpret the confidence interval?

How to interpret the confidence interval?

There is a 95% probability that the “true” model coefficient lies within the 95% confidence interval around the estimated coefficient.

Let’s explore this concept with an in-class Jupyter notebook.

See “lecture_09_supporting.ipynb” in the “supporting notebooks” directory for this lecture.