Data, Environment and Society: Lecture 10: Linear Regression, continued

Instructor: Duncan Callaway
GSI: Seigi Karasaki

September 25, 2018

Announcements

Today

- Standard errors
- Confidence intervals
- We'll compare these to what we get by 'simulating' the confidence interval.

Reading

- For thursday: Alstone et al. See github README for instructions on what to read.

Question:

- Seigi out of town through rest of week. How to support your week?
  - Jupyter notebook with further review
  - Skype-in office hours

Review

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

$$\frac{\partial \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\partial \hat{\beta}_0} = 0 \qquad \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\partial \hat{\beta}_0} = 0 \qquad \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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.

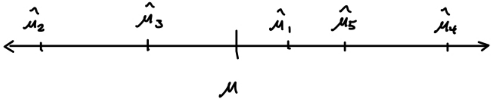
- Suppose your population is all countries in the world
- Randomly sample 20 of them.
  - First random sample of 20 $\rightarrow i = 1$
  - Second random sample of 20 $\rightarrow i = 2$
  - etc

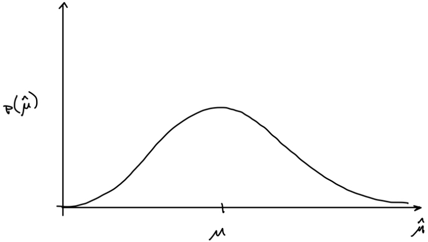
Distribution of means

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

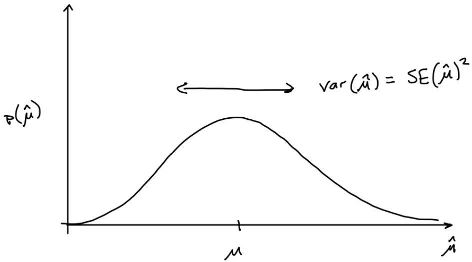
$$\hat{\text{SE}}(\hat{\mu})^2 = \hat{\sigma}^2 \frac{1}{n} = \frac{\text{RSS}}{(n-1)} \frac{1}{n}$$

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.

$$\begin{aligned}
\text{avg}(\hat{\beta}_0) - \beta_0 &= 0 \quad \text{(unbiased)} \\
\text{SE}(\hat{\beta}_0)^2 &= \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \\
\text{avg}(\hat{\beta}_1) - \beta_1 &= 0 \quad \text{(unbiased)} \\
\text{SE}(\hat{\beta}_1)^2 &= \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\end{aligned}$$

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_10_supporting.ipynb” in the “supporting notebooks” directory for this lecture.

What if the confidence interval contains zero?

For example, if

$$-10.3 < \beta_1 < 24.8?$$

This implies there is more than a remote chance that there is no significant relationship between the dependent and independent variables.

p-values

p-values measure the probability that the estimated coefficients arose by chance from a data generating process that actually has *no* relationship between the inputs and outputs.

p = 0.05 implies a 5% chance that the true parameter value is *zero*.

A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.

p-hacking?

What's wrong with these practices:

- Stop collecting data once $p < 0.05$
- Analyze many independent variables, but only report those for which $p < 0.05$
- Collect and analyze many data samples, but only report those with $p < 0.05$
- Exclude participants to get $p < 0.05$.
- Transform the data to get $p < 0.05$.

(credit to Leif Nelson, UCB Haas)

The trouble with p-hacking...

...is that by looking for the data set and the models that give low p-values, you could just be looking for those 5% “chances” where the real relationship is non-existent.

The trouble with p-hacking...

...is that by looking for the data set and the models that give low p-values, you could just be looking for those 5% “chances” where the real relationship is non-existent.

Some estimates suggest that this practice leads to false positive rates of 61%!

Model accuracy: R^2

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Model accuracy: R^2

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 measures the fraction of variation in the dependent variable that is captured by the model.