

What's the Point of Stats?

Tom Glinnan

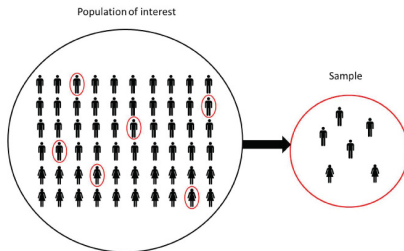
November 6, 2020

Why might we care about correlation?

- ▶ Correlation is not causation
- ▶ *But* it helps us look in the right place. It's where all science starts
- ▶ We want to understand how correlation works so we know exactly how useful it is \implies we should revise some ST102

Populations and Samples

We want to know how the world works, but we won't observe everyone → the problem of dumb luck



Names: Everyone in the world is the **Population**; our data is a **Sample**; the thing we are trying to estimate is called a **Parameter**

What is statistics *really*?

Statistics really has two main areas:

- ▶ **Estimation:** given our sample, what is our best guess for the population parameter?
→ e.g. what is the average height of 20-year olds in the UK?
- ▶ **Inference (Hypothesis Testing):** given our sample, how sure can we be that something is not true in the population?
→ e.g. is the average height of 20-year olds in the UK is above 170cm?

Both are fundamental to understanding correlation so we'll cover them here

Estimation I

‘An estimator is just a function from data to the parameter space’.
Ordinary Least Squares (OLS) is one type of estimator

Specifically:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

The Data: (x_i, y_i) for each person i

The Parameters: (α, β)

The Key Problem: Different Data \implies Different Estimates

We denote the estimates¹ by hats: e.g. $\hat{\alpha}$ is the estimate for α

¹An *estimator* is the function itself, the (point) *estimate* is the value of the function for a particular set of data

Estimation II

What properties might we like our estimator to have?

- ▶ **Unbiased:** on average it gives the true value
- ▶ **Consistent:** as our sample gets bigger, our estimate gets arbitrarily good eventually
- ▶ **Efficient:** our estimator has a small variance

In all statistics, we have to make some assumptions about the population to find properties of our estimator

Illustration

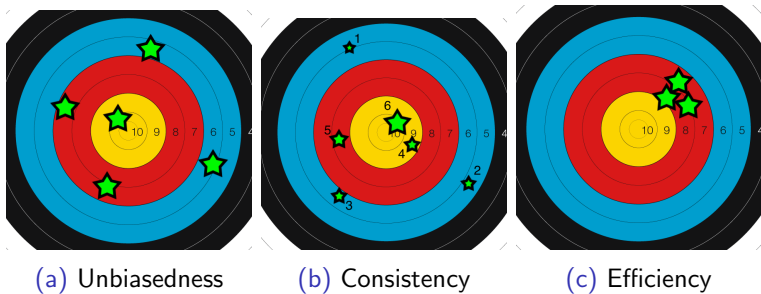


Figure: Three Key Properties

For consistency, labels 1, ..., 6 are the order of the shots

Why OLS?

Under some assumptions, OLS is unbiased, consistent, and has the smallest variance of **any** (linear) estimator

1. Our model ($y_i = \alpha + \beta x_i + \varepsilon_i$) actually is how the world works
2. Not all x_i are the same
3. No matter what x_i is, what we don't know is random noise ($E(\varepsilon_i|x) = 0$)
4. No matter what x_i is, what we don't know has the same magnitude on average ($Var(\varepsilon_i|x)$ is the same for all i)
5. What we don't know is for person i is correlated with what we don't know for person j ($Cov(\varepsilon_i, \varepsilon_j|x) = 0$)

We'll return to these issues when we talk about issues of using OLS in the real world

Multivariate Regression I

The thing written here is called a 'bivariate regression model':

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

If you imagine, we can extend this:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_n x_{ni} + \varepsilon_i$$

Where the y_i is called the **outcome** variable and the x_{ji} are called the **explanatory** variables. A **variable** is just a piece of data. A **coefficient** is the thing in front of a variable

For example: $\text{Wage}_i = \alpha + \beta_1 \text{Education Level}_i + \beta_2 \text{Gender}_i + \varepsilon_i$

Multivariate Regression II

What do the numbers actually mean?

When what we don't know is random, $E(\varepsilon_i|x) = 0$, so

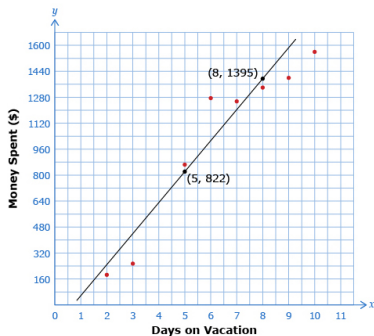
$$E(y|x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

and therefore:

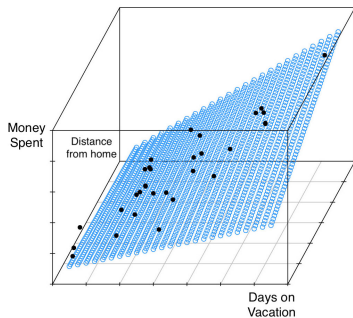
$$\beta_j = \frac{\partial E(y|x)}{\partial x_j}$$

Because of what partial derivatives mean: β_j is the change in the line of best fit in the direction x_j , holding others constant

Illustration



(a) Bivariate



(b) Multivariate

$$\text{Money Spent}_i = \alpha + \beta_1 \text{DOV}_i + \beta_2 \text{DFM}_i + \varepsilon_i$$

Multivariate Regression III

Because maths people like vectors, we can use the shorthand:

$$y_i = \beta' x_i + \varepsilon_i$$

where

$$\beta' = [\alpha \ \beta_1 \ \dots \ \beta_n] \quad \text{and} \quad x_i = \begin{bmatrix} 1 \\ x_{1i} \\ \vdots \\ x_{ni} \end{bmatrix}$$

and talk about estimating β with $\hat{\beta}$

Hypothesis Testing I

As $\hat{\beta}$ is a random variable, it has a distribution, so we can do testing on it.

So what is a test?

Whenever you hear 'hypothesis', just think 'initial guess'. We aim to see if our estimate $\hat{\beta} \approx \beta_0$, our initial guess

- ▶ Make a *null hypothesis* H_0
- ▶ Assume that the null is true
- ▶ Work out the probability we saw what we just did, or more extreme
- ▶ If under our assumption it was really unlikely to see what we just saw, then our assumption was probably wrong!

Hypothesis Testing II (non-essential)

- ▶ For mathy/philosophy-types: this is like a generalised version of a proof by contradiction
- ▶ We never 'accept' a hypothesis.
 - To show why, say we have two possible null hypotheses:
 $H' : \beta_1 = 0$ and $H'' : \beta_1 = 0.000000001$
 - any test would probably reject both in very similar circumstances.
 - Accepting both would mean you think that β_1 takes two different values, which is philosophically a bit dodgy
- ▶ We can do a test using its test statistic T or its p value
- ▶ Interpretation of a p value: assuming that the null is true, what is the probability that we saw what we just did, or something more 'extreme'

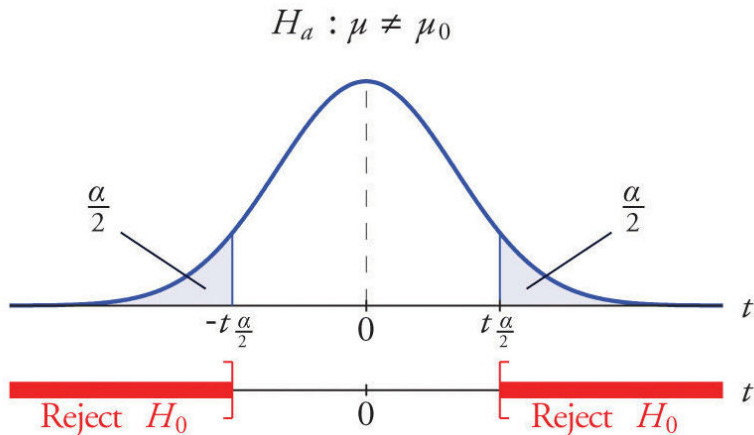
Hypothesis Testing III



Figure: A bottle-filling machine

The inventor of the t-stat worked in a beer factory. The aim was to check if bottles were over-filled or under-filled

Illustration



Technical Note: we need to assume the distribution of $\hat{\beta}$. When small amounts of data assume t . When large amounts of data, Normal becomes an arbitrarily good approximation by the central limit theorem

Errors

I promise errors are a less confusing concept than you think!

Type I: $\hat{\beta}_j$ failed the test but it should have passed

Type II: $\hat{\beta}_j$ passed the test but it should have failed

Note that:

- ▶ $P(\text{Type I}) = \alpha$ (the significance level)
→ You get to pick this
- ▶ $\text{Power} = 1 - P(\text{Type II})$

More power comes from either having more data, or using a more efficient estimator

Understanding Errors

A court case is effectively a hypothesis test:

- ▶ The jury have some beliefs about whether or not the defendant is guilty
- ▶ Someone is 'innocent until proven guilty', i.e. H_0 : Innocent and H_1 : Guilty
- ▶ The jury sees evidence and decides using the same principle as hypothesis testing: assuming that they are innocent, how likely is it that this evidence exists? Level of reasonable doubt $= \alpha$
- ▶ What are Type I and Type II errors here?
- ▶ If we reduce Type I (fewer innocents go to jail) then more guilty people go free (increase Type II), and vice versa
- ▶ How to solve? More evidence, or better inference from the evidence we have!

Testing in practice

What might we want to test?

- ▶ Whether a given $\beta_j = 0$ - This would mean that x_j wouldn't help to explain y_i (t test)
→ if we conclude $\beta_j \neq 0$ then we say that β_j is 'statistically significant'
- ▶ Whether *all* of the β are 0 This would mean that the model is useless (f test)
→ This is related to R^2 : the percentage of variation in y_i that the model explains
→ This is Regression ANOVA from ST102
- ▶ Anything else depending on context!

Remember to realise the Type I / Type II tradeoffs! Usually pick $\alpha = 10\%^*, 5\%^{**}, 1\%^{***}$

Looking at stata

. reg price weight

Source	SS	df	MS	★ Number of obs	=	74
				F(1, 72)	=	29.42
Model	184233937	1	184233937	★ Prob > F	=	0.0000
Residual	450831459	72	6261548.04	★ R-squared	=	0.2901
				Adj R-squared	=	0.2802
Total	635065396	73	8699525.97	Root MSE	=	2502.3

★ price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
weight	2.044063	.3768341	5.42	0.000	1.292857	2.795268
_cons	-6.707353	1174.43	-0.01	0.995	-2347.89	2334.475

Questions

- ▶ How do we rule out that a correlation is due to dumb luck?
- ▶ What is the most important property an estimator can have: unbiasedness, consistency, efficiency, something else?
- ▶ Why might OLS not be the best way to estimate a regression slope?
- ▶ Under which circumstances would you want to test a hypothesis using a high α ? A low one?