Simple Regression Model

Juergen Meinecke

Roadmap

The Population Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \dots, n$$

We have n observations, $(X_i, Y_i), i = 1, .., n$

- Y_i is the dependent variable
- X_i is the independent variable or explanatory variable or regressor
- u_i is the error term
- β_0 is the intercept
- β_1 is the slope

The error term u_i captures all factors that could explain Y_i *over and above the explanatory variable* X_i

Our main interest is to learn about the **expected effect** on Y of a unit change in X

This is often referred to as the **causal effect of X on Y**

Graphically, this causal effect is represented by the slope of the line

Technically, we need to study the question:
Given a scatterplot between two variables X and Y, how can we fit a line?

Fitting a line boils down to finding (estimating) the parameters β_0 (intercept) and β_1 (slope)

Statistical, or econometric, inference about β_0 and β_1 entails:

- Estimation

  How do we estimate β_0 and β_1?

  Answer: ordinary least squares (OLS)

- Hypothesis testing

  How to test if β_0 or β_1 are zero (or some other value)?

- Confidence intervals

  How to construct a confidence intervals?

Looking again at the classroom size example: the PRF is

$$TestScore = \beta_0 + \beta_1 STR$$

where

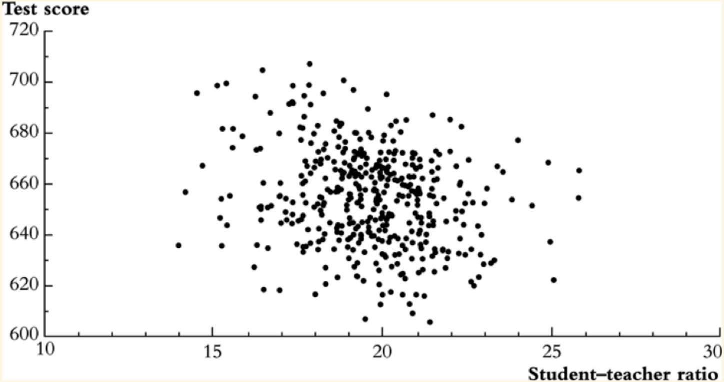
$$\begin{aligned} \beta_1 &= \text{slope of PRF} \\ &= \frac{\partial TestScore}{\partial STR} \\ &= \text{change in } TestScore \text{ for a unit change in } STR \end{aligned}$$

The parameters β_0 and β_1 are unobserved population parameters

Goal: statistical inference about β_0 and β_1

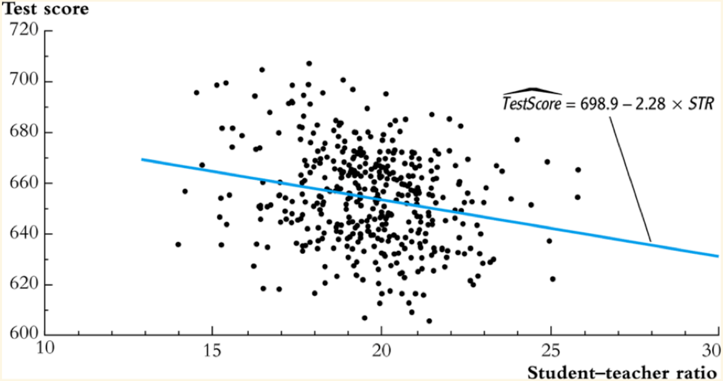
(Our priority is learning about β_1)

Given a scatterplot of the data …



…how does the estimated PRF look like?

Answer (after applying OLS estimation):



Estimated intercept $\hat{\beta}_0 = 698.9$

Estimated slope $\hat{\beta}_1 = -2.28$

Estimated PRF: $\widehat{TestScore} = 698.9 - 2.28 \cdot STR$

Interpretation of the estimated slope and intercept

$$\widehat{TestScore} = 698.9 - 2.28 \cdot STR$$

Districts with one more student per teacher on average have test scores that are 2.28 points lower

That is, $\frac{\partial \widehat{TestScore}}{\partial STR} = -2.28$

The intercept (taken literally) means that districts with zero students per teacher would have a (predicted) test score of 698.9

(This interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful)

Simple Regression Model

Juergen Meinecke

Roadmap

Let's tentatively assume we know how to estimate β_0 and β_1

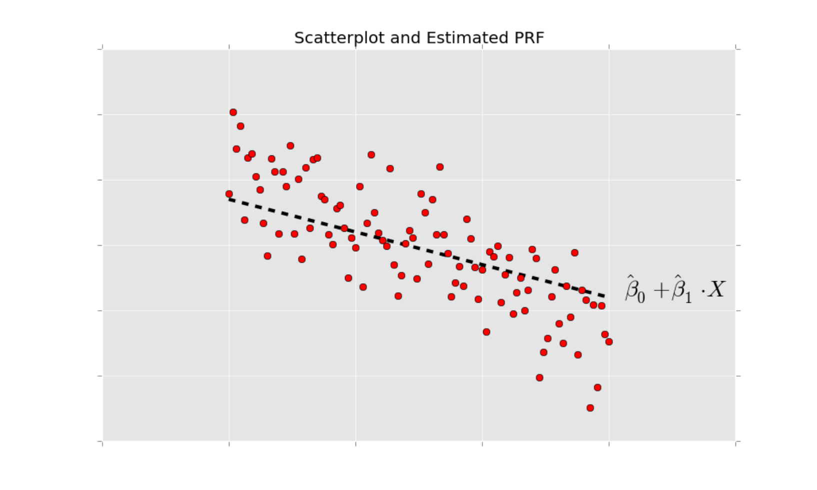
By convention, their estimators will be denoted $\hat{\beta}_0$ and $\hat{\beta}_1$

This results in the following estimated PRF

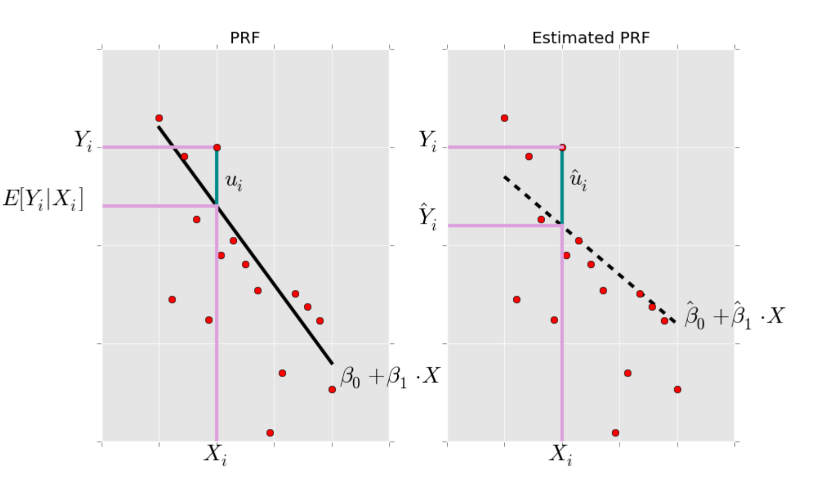
$$\widehat{\text{PRF}} := \hat{\beta}_0 + \hat{\beta}_1 X$$

It should be obvious that PRF $\neq \widehat{\text{PRF}}$

So at the end of the day, this is the picture we will actually see



Before we proceed to derive the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to clarify some terminology



(Again: we only get to see the picture on the right side)

Definition

The **predicted value** of Y_i is given by $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 X_i$.

The predicted value is the estimated PRF.

Difference between errors and residuals

Definition

The **error** is given by $u_i := Y_i - \beta_0 - \beta_1 X_i$.

It is the difference between Y_i and the PRF.

Definition

The **residual** is given by $\hat{u}_i := Y_i - \hat{Y}_i$.

It is the difference between Y_i and the predicted value.

Corollary

$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i = \hat{Y}_i + \hat{u}_i.$

Corollary

$\beta_0 + \beta_1 X_i + u_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i.$

By the way, it should be clear that in general

$$\beta_0 \neq \hat{\beta}_0 \qquad \beta_1 \neq \hat{\beta}_1 \qquad u_i \neq \hat{u}_i$$

Where do $\hat{\beta}_0$ and $\hat{\beta}_1$ come from?

Least squares criterion for estimators:

- minimize (in some sense) the difference between the estimated population regression function and the observations Y_i
- but some error terms will be above the line and some will be below, won't they cancel each other out?
- trick: look at the **squared residual** instead

  $$(Y_i - b_0 - b_1 X_i)^2$$
- Now, choose b_0 and b_1 such that the **sum of squared residuals** is minimized

  $$SSR(b_0, b_1) := \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

The set of solutions b_0 and b_1 that minimize SSR are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$

Nice fact:
turns out, there is only one unique minimizer to the least squares problem

Nice fact:
it is reasonably easy to compute that minimizer

We'll turn to the computation now…

Definition

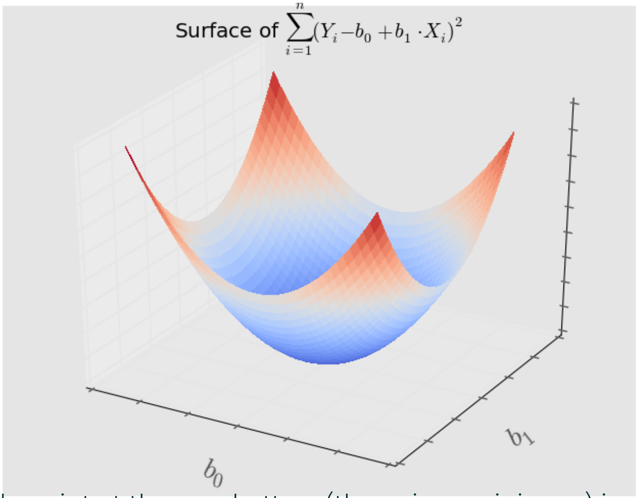
The **Ordinary Least Squares (OLS) estimators** are defined by

$$\hat{\beta}_0, \hat{\beta}_1 := \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

In words

- we look at the rhs as a function in b_0 and b_1
- that function happens to be quadratic
- we find the values of b_0 and b_1 that minimize that function
- the values that minimize that function are called solution
- we give the solution a specific name: $\hat{\beta}_0$ and $\hat{\beta}_1$

Geometry of the minimization problem



The single point at the very bottom (the unique minimum) is denoted $(\hat{\beta}_0, \hat{\beta}_1)$

Simple Regression Model

Juergen Meinecke

Roadmap

- Ordinary Least Squares Estimation
  - Derivation of OLS Estimator

The mathematics of finding the solution

The basic approach is *multivariate calculus* which you know from high school or EMET1001 or both

First step: differentiate $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$

$$\frac{\partial SSR}{\partial b_0} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$

$$\frac{\partial SSR}{\partial b_1} = -2\sum_{i=1}^{n} X_i \cdot (Y_i - b_0 - b_1 X_i)$$

This is a set in two linear equations and two unknowns (b_0 and b_1)

Second step: set derivative to zero
(at this step, $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$)

$$0 = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

$$0 = -2 \sum_{i=1}^{n} X_i \cdot (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

These two equations are the *first order necessary condtions (foc)* for a minimum

Third step: using the first foc, solve for $\hat{\beta}_0$ as function of $\hat{\beta}_1$

$$
\begin{aligned}
0 &= -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\
&= \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\
&= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 X_i \\
&= \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i
\end{aligned}
$$

which is equivalent to

$$
n\hat{\beta}_0 = \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i
$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

and

$$\begin{aligned}\hat{\beta}_0 &= \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} X_i}{n} \\ &= \bar{Y} - \hat{\beta}_1 \bar{X}\end{aligned}$$

This is an elegant result:
$\hat{\beta}_0$ is the sample average of Y minus $\hat{\beta}_1$ times the sample average of X

Fourth step: substitute $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ into the second first order condition from the second step and solve for $\hat{\beta}_1$

$$
\begin{aligned}
0 &= -2 \sum_{i=1}^{n} X_i \cdot (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\
&= \sum_{i=1}^{n} X_i \cdot (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\
&= \sum_{i=1}^{n} X_i \cdot (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i) \\
&= \sum_{i=1}^{n} X_i \cdot (Y_i - \bar{Y} - \hat{\beta}_1 (X_i - \bar{X})) \\
&= \sum_{i=1}^{n} X_i Y_i - \bar{Y} X_i - \hat{\beta}_1 (X_i^2 - \bar{X} X_i) \\
&= \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} (X_i^2 - \bar{X} X_i)
\end{aligned}
$$

Continuing...

$$\begin{aligned} 0 &= \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} (X_i^2 - \bar{X} X_i) \\ &= \left(\sum_{i=1}^{n} X_i Y_i \right) - n \bar{X} \bar{Y} - \hat{\beta}_1 \sum_{i=1}^{n} (X_i^2 - \bar{X} X_i), \end{aligned}$$

where we use $\sum_{i=1}^{n} X_i = n\bar{X}$ (we'll prove this in the workshop)
Rearranging

$$\hat{\beta}_1 \sum_{i=1}^{n} (X_i^2 - \bar{X} X_i) = \left(\sum_{i=1}^{n} X_i Y_i \right) - n \bar{X} \bar{Y}$$

where the lhs can be simplified

$$\sum_{i=1}^{n} (X_i^2 - \bar{X} X_i) = \sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i = \left(\sum_{i=1}^{n} X_i^2 \right) - n \bar{X}^2$$

Isolating $\hat{\beta}_1$ on the left results in...

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} X_i Y_i\right) - n\bar{X}\bar{Y}}{\left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2}$$

Now we exploit a property of the summation operator:
(we'll prove this in the workshop)

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \left(\sum_{i=1}^{n} X_i Y_i\right) - n\bar{X}\bar{Y}$$

Now we use this property to simplify the result for $\hat{\beta}_1$:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\left(\sum_{i=1}^{n} X_i Y_i\right) - n\bar{X}\bar{Y}}{\left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2} \\ &= \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\end{aligned}$$

What is the rhs equal to?

With a tiny modification we see that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- denominator: sample variance of X_i
- numerator: sample covariance between X_i and Y_i

The OLS estimator of the slope is equal to the ratio of sample covariance and sample variance!

In summary, we have now derived the OLS estimators of β_0 and β_1, they are

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The OLS estimators are functions of the sample data only

Given the sample data (X_i, Y_i) we can first compute the rhs for $\hat{\beta}_1$ and then we can compute the rhs for $\hat{\beta}_0$

Computer programs such as `Python` easily calculate the rhs for you