Standard linear regression model, model assumptions, OLS approximation

Empirical Methods, Lecture 2

Tunga Kantarcı, FEB, Groningen University, Spring 2025

SLM

Hereafter SLM stands for the standard linear regression model.

SLM, an empirical question

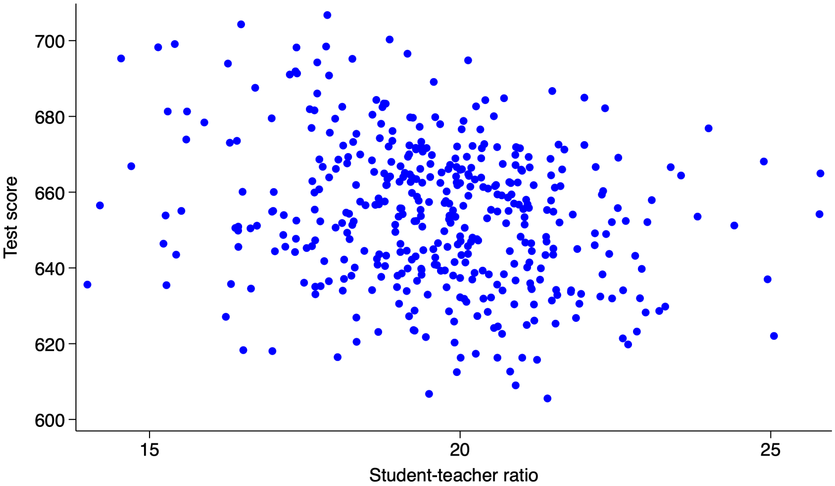
The California Standardized Testing and Reporting (STAR) dataset contains data on test performance and school characteristics. The data are from districts in California collected in 1999 by the California Department of Education. Test scores are the average of the reading and math scores on a standardized test administered to 5th grade students. The student-teacher ratio is the number of students divided by the number of teachers working full-time in a district. The data is analyzed in Kruger and Whitmore (Economic Journal, 2001).

```
. summarize tstscr str
```

Variable	Obs	Mean	Std. dev.	Min	Max
tstscr	420	654.1565	19.05335	605.55	706.75
str	420	19.64043	1.891812	14	25.8

SLM, an empirical question

Scatter plot of test scores vs student-teacher ratio:



SLM, an empirical question

We want to analyze how test scores vary with the student-teacher ratio.

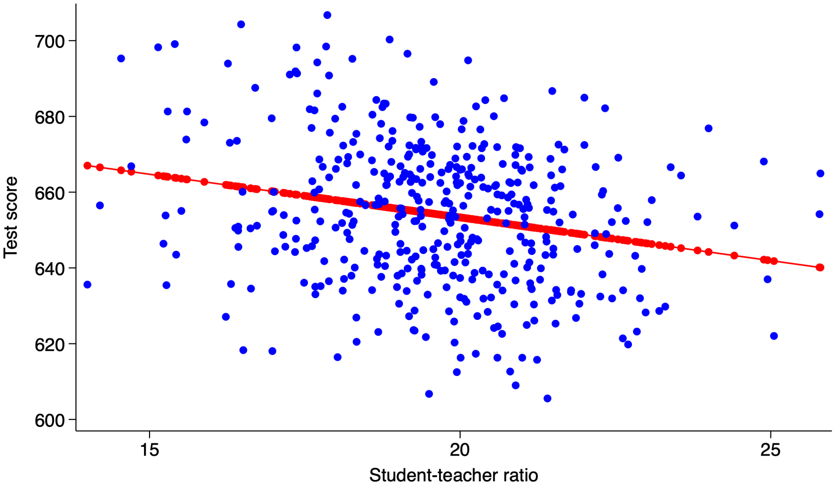
To do this we will use the standard linear regression model.

We will first discuss the model and the assumptions we make while using the model.

We will then discuss the method to use to fit this model to the data.

SLM, an empirical question

Scatter plot of test scores vs student-teacher ratio, fitted model:



SLM, notation

But first we need to be clear about notation.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

SLM, notation

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} 1 & x_{12} & \dots & x_{1K} \\ 1 & x_{22} & \dots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \dots & x_{nK} \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

SLM, notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{y}$: dependent variable. n$\times$1. The bold font is for observations. A vector is always a column vector.

y_i: an observation in a row of $\boldsymbol{y}$.

i: unit of study.

SLM, notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{X}$: matrix of variables. $n \times K$. The bold font indicates multiple observations. The big font indicates multiple variables.

$\boldsymbol{x}_k$: a column in $\boldsymbol{X}$. $n \times 1$. It contains n observations for variable k. k, l, m are used to indicate different columns. The bold font indicates multiple observations.

$\boldsymbol{x}_i'$: a row in $\boldsymbol{X}$. $1 \times K$. It contains observations for K variables for unit i. i, j, t, s are used to indicate different rows. The bold font indicates multiple variables.

$\boldsymbol{x}_i$: column vector formed by the transpose of a row in $\boldsymbol{X}$. K$\times$1.

x_{ik}: an observation in row i, column k of $\boldsymbol{X}$.

SLM, notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{\beta}$: true, or population, coefficient vector. K$\times$1. Unobserved.

β_k: a coefficient in a row of $\boldsymbol{\beta}$. These are slope parameters.

If you let $\boldsymbol{x}_0$ be a column of 1s, β_0 is the constant term, or the intercept.

SLM, notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{\varepsilon}$: error. n×1. Unobserved.

ε_i: an element in a row of $\boldsymbol{\varepsilon}$.

SLM, notation

So for i, we have

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

and if

$$\boldsymbol{x}' = \begin{bmatrix} 1 & x \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

we have

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

SLM, what is it?

What is a linear regression model?

Which assumptions make a liner regression model ‘standard’?

SLM, assumptions, linearity

A1. Linearity: the model is linear in the parameters.

SLM, assumptions, linearity

The model

$$y_i = \beta_1 + \beta_2 x_{i2}^2 + \varepsilon_i$$

is linear in the parameters, nonlinear in the regressor, linear in the squared regressor

SLM, assumptions, linearity

The model

$$y_i = x_{i2}^{\beta_2} + \varepsilon_i$$

is nonlinear in the parameter.

SLM, assumptions, full column rank

A2. Full column rank: $rank\,(\boldsymbol{X}) = K$. Remember that $\boldsymbol{X}$ is $n \times K$ matrix. It contains K columns. Hence, A2 means $\boldsymbol{X}$ has full column rank.

This means that the columns of X are linearly independent, meaning no column can be written as a linear combination of other columns.

SLM, assumptions, full column rank

A2 is not satisfied in two cases.

SLM, assumptions, full column rank

First, if $n < K$. Note that $rank\,(\boldsymbol{X}) \leq min\,(n, K)$. Hence, $rank\,(\boldsymbol{X})$ cannot be K if $n < K$. In practice this is not likely.

SLM, assumptions, full column rank

Second is the case where there is an exact relationship among any of the columns of $\boldsymbol{X}$.

SLM, assumptions, full column rank

E.g., consider the regression

$$wage_i = x_{0i}\beta_0 + d_i^{female}\beta_1 + d_i^{male}\beta_2 + \varepsilon_i$$

where $x_{0i} = 1$, and

$$d_i^{female} = \begin{cases} 1 & if \quad i = female \\ 0 & if \quad i = male \end{cases}$$

$$d_i^{male} = \begin{cases} 0 & if \quad i = female \\ 1 & if \quad i = male \end{cases}$$

Sum of the values in each row of $\boldsymbol{d}^{female}$ and $\boldsymbol{d}^{male}$ is equal to the value in that row of $\boldsymbol{x}_0$. Hence, one value can be perfectly predicted from other values. $rank\,(\boldsymbol{X}) \neq K$. This is perfect multicollinearity.

SLM, assumptions, full column rank

$$\begin{bmatrix} \boldsymbol{x}_0 & \boldsymbol{d}^{female} & \boldsymbol{d}^{male} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

SLM, assumptions, full column rank

We will learn later in this lecture about the OLS method to estimate $\boldsymbol{\beta}$.

Perfect multicollinearity is a problem for estimating $\boldsymbol{\beta}$. The OLS estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Rank of $\boldsymbol{X}$ is not K. Hence rank of $\boldsymbol{X}'\boldsymbol{X}$ is not K. Square matrices are invertible if they have full rank. $\boldsymbol{X}'\boldsymbol{X}$ does not have full rank and hence it is not invertible. This implies that $\hat{\boldsymbol{\beta}}$ has multiple solutions.

This means that when there is perfect multicollinearity between variables, we cannot use the OLS method. This is why Stata will drop a variable to avoid perfect multicollinearity with that variable.

SLM, assumptions, exogeneity

A3. Strict exogeneity:

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{x}_k\right] = 0.$$

What does this moment condition say? Recall that $\boldsymbol{x}_k$ contains n observations for variable k. The stated condition says that the expected value of the error at observation i in the sample is independent of the explanatory variable k observed at any observation, including observation i. It says that the average of the error is the same across all observations of the independent variable, and that this average is 0. More on this later.

SLM, assumptions, exogeneity

Why is it strict? Take a look at the definition of weak exogeneity:

$$\mathrm{E}\left[\varepsilon_i \mid x_{ik}\right] = 0.$$

x_{ik} is observation i for variable k. That is, we do not consider all n observations of variable k, but just observation i. That is why

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{x}_k\right] = 0$$

is strict, since all n observations of variable k are considered.

SLM, assumptions, exogeneity

Note that strict exogeneity,

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{x}_k\right] = 0,$$

can be considered to apply to all K variables as

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{X}\right] = 0.$$

But this is beside the point. What makes it strict is about n not K.

SLM, assumptions, exogeneity

Why do we need A3? The model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Taking the expectation conditional on $\boldsymbol{X}$,

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{y} \mid \boldsymbol{X}\right] &= \mathrm{E}\left[\boldsymbol{X}\boldsymbol{\beta} \mid \boldsymbol{X}\right] + \mathrm{E}\left[\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] \\ &= \boldsymbol{X}\boldsymbol{\beta}. \end{aligned}$$

That is, A3 gives the conditional expectation function, or the population regression function.

Later we will study the other reasons for A3, and how we can relax it.

SLM, assumptions, exogeneity

A3 has two implications. First, by the LIE,

$$\mathrm{E}[\varepsilon_i] = \mathrm{E}_{\boldsymbol{X}}[\mathrm{E}[\varepsilon_i \mid \boldsymbol{X}]] = 0.$$

SLM, assumptions, exogeneity

Second, note that

$$\operatorname{Cov}[\varepsilon_i, \boldsymbol{X}] = \operatorname{E}[\varepsilon_i \boldsymbol{X}] - \operatorname{E}[\varepsilon_i] \operatorname{E}[\boldsymbol{X}]$$

and

$$\operatorname{E}[\varepsilon_i \boldsymbol{X}] = \operatorname{E}_{\boldsymbol{X}}[\operatorname{E}[\varepsilon_i \boldsymbol{X} \mid \boldsymbol{X}]] = \operatorname{E}_{\boldsymbol{X}}[\boldsymbol{X} \operatorname{E}[\varepsilon_i \mid \boldsymbol{X}]].$$

Hence, if

$$\operatorname{E}[\varepsilon_i \mid \boldsymbol{X}] = \mathbf{0},$$

then

$$\operatorname{Cov}[\varepsilon_i, \boldsymbol{X}] = \mathbf{0}.$$

It says that ε_i is not correlated with $\boldsymbol{X}$, or any function of $\boldsymbol{X}$.

SLM, assumptions, exogeneity

$$\mathrm{E}\left[\varepsilon_i \mid \boldsymbol{X}\right] = \mathbf{0}$$

can be easily violated. E.g., suppose

$$\varepsilon_i^* = \varepsilon_i + \boldsymbol{x}_k \beta_k,$$

where $\beta_k \neq 0$, and $\boldsymbol{x}_k$ is correlated with $\boldsymbol{X}$. Then, ε_i^* is correlated with $\boldsymbol{X}$ because

$$\mathrm{E}\left[\varepsilon_i^* \mid \boldsymbol{X}\right] \neq 0.$$

This is restrictive in practice. We would want to include $\boldsymbol{x}_k$ in the model so that

$$\mathrm{E}\left[\varepsilon_i^* \mid \boldsymbol{X}\right] = 0.$$

But what if $\boldsymbol{x}_k$ is unobserved? We cannot include it.

SLM, assumptions, spherical errors

A4. Errors are homoskedastic and non-autocorrelated.

Homoskedasticity: each ε_i has the same variance σ^2 conditional on $\boldsymbol{X}$:

$$\text{Var}\left[\varepsilon_i \mid \boldsymbol{X}\right] = \sigma^2, \; \forall\, i.$$

Nonautocorrelation: each ε_i is uncorrelated with every other disturbance ε_j conditional on $\boldsymbol{X}$:

$$\text{Cov}\left[\varepsilon_i, \varepsilon_j \mid \boldsymbol{X}\right] = 0, \; \forall\, i \neq j.$$

Later we will study how we can relax this assumption.

SLM, assumptions, spherical errors

If $\mathrm{E}[\varepsilon_i \mid \boldsymbol{X}] = 0$,

$$\mathrm{Var}[\varepsilon_i \mid \boldsymbol{X}] = \mathrm{E}\left[\varepsilon_i^2 \mid \boldsymbol{X}\right] - (\mathrm{E}[\varepsilon_i \mid \boldsymbol{X}])^2 = \mathrm{E}[\varepsilon_i \varepsilon_i \mid \boldsymbol{X}] = \sigma^2,$$

and

$$\mathrm{Cov}[\varepsilon_i, \varepsilon_j \mid \boldsymbol{X}] = \mathrm{E}[\varepsilon_i \varepsilon_j \mid \boldsymbol{X}] - \mathrm{E}[\varepsilon_i \mid \boldsymbol{X}]\,\mathrm{E}[\varepsilon_j \mid \boldsymbol{X}] = \mathrm{E}[\varepsilon_i \varepsilon_j \mid \boldsymbol{X}] = 0.$$

The variance-covariance matrix for n errors is

$$\mathrm{Var}[\boldsymbol{\varepsilon} \mid \boldsymbol{X}] = \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] - \mathrm{E}[\boldsymbol{\varepsilon} \mid \boldsymbol{X}]\,\mathrm{E}\left[\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] = \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right].$$

Note that $\boldsymbol{\varepsilon}$ is $n \times 1$, and hence $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$ is $n \times n$. This implies that

$$\mathrm{Var}[\boldsymbol{\varepsilon} \mid \boldsymbol{X}] = \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] = \sigma^2 I_n = \sigma^2 \boldsymbol{I}$$

which is a $n \times n$ matrix.

SLM, assumptions, spherical errors

$$
\begin{aligned}
\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] &= \begin{bmatrix} \mathrm{E}\left[\varepsilon_1\varepsilon_1 \mid \boldsymbol{X}\right] & \mathrm{E}\left[\varepsilon_1\varepsilon_2 \mid \boldsymbol{X}\right] & \dots & \mathrm{E}\left[\varepsilon_1\varepsilon_n \mid \boldsymbol{X}\right] \\ \mathrm{E}\left[\varepsilon_2\varepsilon_1 \mid \boldsymbol{X}\right] & \mathrm{E}\left[\varepsilon_2\varepsilon_2 \mid \boldsymbol{X}\right] & \dots & \mathrm{E}\left[\varepsilon_2\varepsilon_n \mid \boldsymbol{X}\right] \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{E}\left[\varepsilon_n\varepsilon_1 \mid \boldsymbol{X}\right] & \mathrm{E}\left[\varepsilon_n\varepsilon_2 \mid \boldsymbol{X}\right] & \dots & \mathrm{E}\left[\varepsilon_n\varepsilon_n \mid \boldsymbol{X}\right] \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & 1 \end{bmatrix}}_{I_n} \sigma^2
\end{aligned}
$$

SLM, assumptions, random sampling

A5. Random sampling: the data $\{(\boldsymbol{x}_i, y_i) : i = 1, 2, \ldots, n\}$ is a random sample following the population model $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$. It says that all elements of the data have the same probability of being selected from the population. That is, the observations are i.i.d. This implies that the data have been chosen to be representative of the population.

The sample selection model deals with situations where this assumption fails. This course does not study this model.

SLM, assumptions, random sampling

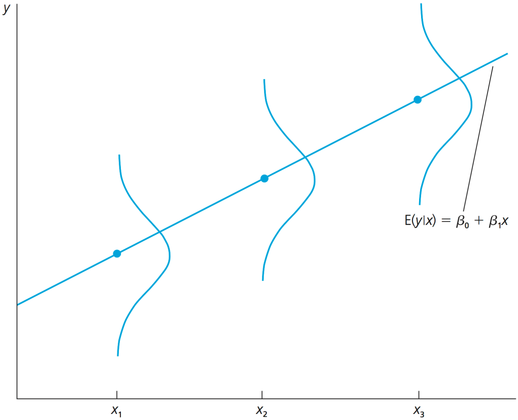
A6. ε_i is normal. That is, ε_i has the mean and variance given by A3 and A4, and has a normal distribution. That is,

$$\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N\left[\mathbf{0}, \sigma^2 \boldsymbol{I}\right].$$

We will use this assumption if n is small. We will drop this assumption if n is large.

SLM, summary of assumptions

A1: regression line is linear in $\boldsymbol{\beta}$. A3: the conditional expectation function. A4: errors have a constant variance conditional on $\boldsymbol{X}$, and hence so do $\boldsymbol{y}$. The latter because $\text{Var}[\varepsilon_i \mid \boldsymbol{X}] = \text{Var}[y_i \mid \boldsymbol{X}]$. A6: errors are normal, and hence so do $\boldsymbol{y}$. The following figure demonstrates all of these assumptions:



OLS approximation

Consider the SLM

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$\boldsymbol{\beta}$ is unknown and we want to estimate it. The best estimate is the one that makes $\boldsymbol{y}$ as close to $\boldsymbol{X}\boldsymbol{\beta}$ as possible since our aim is to explain $\boldsymbol{y}$ with $\boldsymbol{X}\boldsymbol{\beta}$ as much as possible. Let $\hat{\boldsymbol{\beta}}$ be a candidate for $\boldsymbol{\beta}$ that intends to minimise the sum of squared residuals

$$S(\hat{\boldsymbol{\beta}}) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

The necessary condition for a minimum is

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{0}.$$

$\partial S(\hat{\boldsymbol{\beta}})/\partial \hat{\boldsymbol{\beta}}$ is calculated using matrix differentiation. If A2 holds, $S(\hat{\boldsymbol{\beta}})$ attains a minimum at $\hat{\boldsymbol{\beta}}_{OLS}$ which takes the form

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

OLS approximation

The derivation above used matrix differentiation and notation. In this course we will hardly use matrix algebra and notation.

That is, the model of interest in standard form is

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

The sum of squared residuals in standard form is then

$$\sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

OLS approximation

Taking the derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and setting them equal to 0, and using some algebra tricks, the OLS estimator of β_0 is

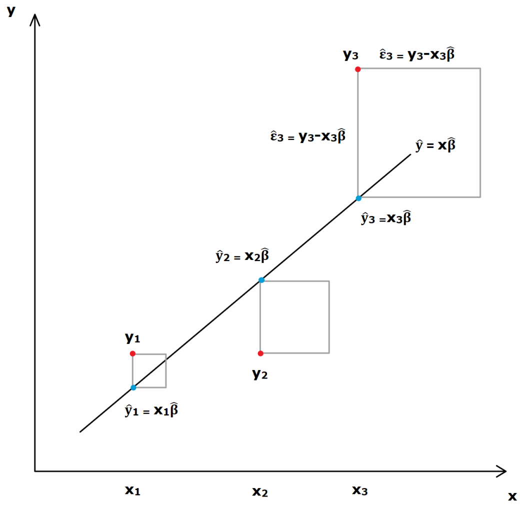
$$\hat{\beta}_{0,OLS} = \bar{y} - \hat{\beta}_{1,OLS}\bar{x}$$

and the OLS estimator of β_1 is

$$\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})}.$$

OLS approximation

By minimizing the sum of "squared" residuals, OLS fits as good as possible a regression line to the data points.



OLS approximation

OLS is an approximation, or estimation, method. It is not a model.

The model is the standard liner regression model.

We estimate the population, or slope, parameters of the model using the OLS method.

OLS approximation

The solution to the least squares problem is

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta}_{OLS} - \boldsymbol{X}'\boldsymbol{y} = -\boldsymbol{X}'(\boldsymbol{y} - \underbrace{\boldsymbol{X}\hat{\beta}_{OLS}}_{\hat{\boldsymbol{y}}}) = -\boldsymbol{X}'\hat{\varepsilon} = 0.$$

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta}_{OLS} = \boldsymbol{X}'\boldsymbol{y}$$

are also called the normal equations.

OLS approximation

Recall the population regression function given by

$$\mathrm{E}\left[\boldsymbol{y} \mid \boldsymbol{X}\right] = \boldsymbol{X}\boldsymbol{\beta}.$$

The solution to the least squares problem gives

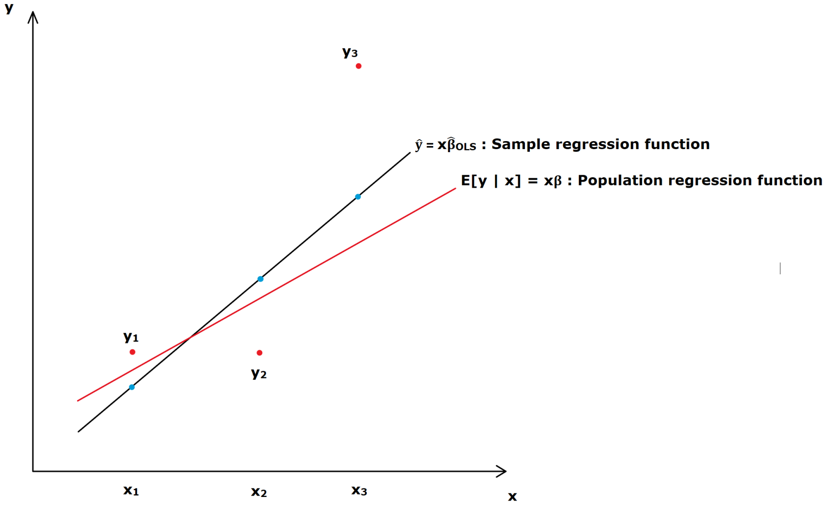
$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}$$

which represent the predictions of the regression model. This is the sample regression function. It is the estimate of the population regression function:

$$\mathrm{E}\widehat{\left[\boldsymbol{y} \mid \boldsymbol{X}\right]} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}.$$

OLS approximation

Compare the unknown population regression function to the sample regression function.



OLS approximation

Consider the sample regression function

$$\widehat{\mathrm{E}\left[\boldsymbol{y} \mid \boldsymbol{X}\right]} = \boldsymbol{X}\hat{\beta}_{OLS}.$$

Considering discrete changes in the predicted dependent variable and the independent variable, we have:

$$\frac{\Delta \widehat{\mathrm{E}\left[\boldsymbol{y} \mid \boldsymbol{X}\right]}}{\Delta \boldsymbol{X}} = \hat{\beta}_{OLS}.$$

This tells that when the independent variable changes by some unit, the dependent variable changes on average by the OLS estimate.

OLS approximation

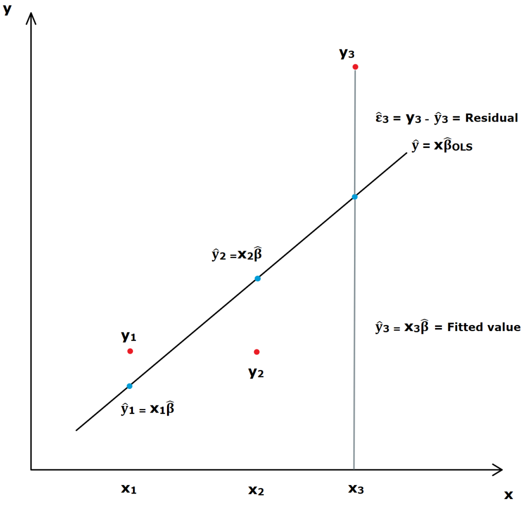
The solution to the least squares problem gives

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$$

which shows that we explain the dependent variable by the prediction of our model and our error.

OLS approximation

An observation of the dependent variable is explained by the prediction of our model and our error: $y_3 = \hat{y}_3 + \hat{\epsilon}_3$.



OLS approximation, implications

The solution has three implications:

1. If the first column of $\boldsymbol{X}$, $\boldsymbol{x}_0$, is a column of 1s, i.e. the regression includes a constant, the residuals, or deviations from the regression line, sum to zero:

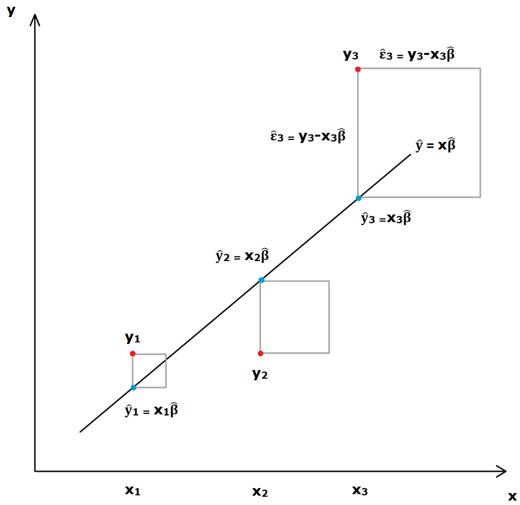
$$\boldsymbol{x}_0'\hat{\boldsymbol{\varepsilon}} = \sum_i^n \hat{\varepsilon}_i = 0.$$

2. $\bar{y} = \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_{OLS} + \bar{\hat{\varepsilon}}$, and since $\bar{\hat{\varepsilon}} = 0$ by the first implication, $\bar{y} = \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}_{OLS}$. This says that the regression hyperplane passes through the point of means of the data.

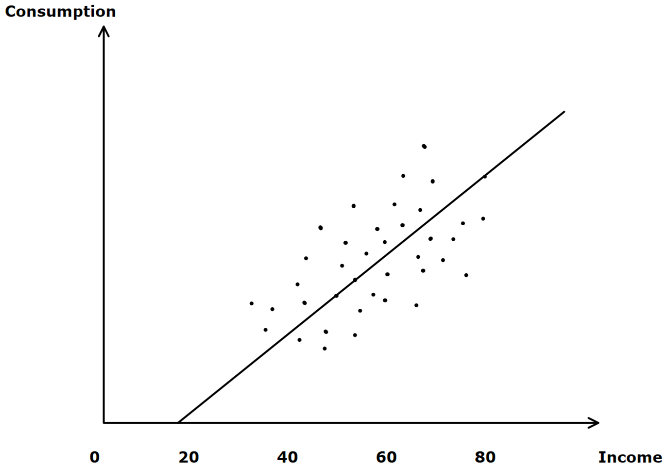
3. $\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}$ from the solution. Taking the means, we obtain $\bar{y} = \bar{\hat{y}} + \bar{\hat{\varepsilon}}$. Since $\bar{\hat{\varepsilon}} = 0$ by the first implication, we obtain $\bar{y} = \bar{\hat{y}}$.

OLS approximation, insights

OLS is not robust to outliers. y_3 contributes too much to the minimization problem of OLS: it pulls the regression towards itself.



OLS approximation, insights

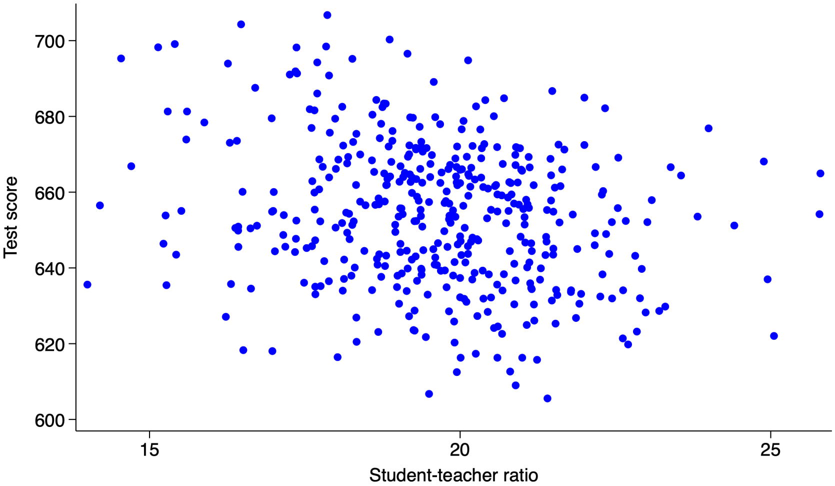


OLS approximation, insights

For incomes between 40 and 80, the consumption function can be approximated by the line (model). Does the line describe the consumption-income relationship for all incomes, or only for the those in the center? Only in the center! What is the predicted consumption when income is 10? A negative value! Models are approximations. Approximations do not work well if we move too far away from the point of approximation. OLS is a good approximator around the average value of x.

OLS approximation, example

Scatter plot of test scores vs student-teacher ratio:



OLS approximation, example

Our population regression model is:

$$\textit{testscr} = \beta_0 + \beta_1 \textit{str} + u.$$

We want to estimate the population parameters of the model.

OLS approximation, example

We use the OLS method

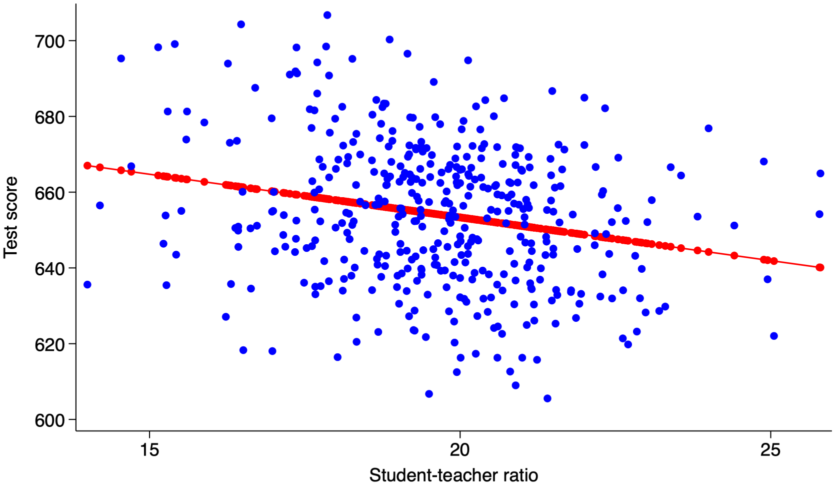
```
. regress testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11012         1  7794.11012   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

which gives the sample regression model, or the fitted model:

$$\widehat{testscr} = 698.93 - 2.28\ str$$

OLS approximation, example

For each i, we have a prediction on the regression line:



OLS approximation, example

Interpretation of the estimated coefficient of the independent variable:

$$\frac{\Delta \widehat{\mathrm{E}\left[testscr \mid str\right]}}{\Delta str} = -2.28.$$

On average, a unit increase in student-teacher ratio is associated with a -2.28 points decrease in test scores.