Why do we run regressions?

Let's say we have two variables y and x. We may regress y on x, because

1. we want to see the linear association between y and x, OR
2. we want to predict y given values of x, OR
3. we want to measure the causal effect of x on y.

- What do we mean by the causal effect?
- We will use the definition of SW book:

A causal effect is defined to be the effect measured in an ideal randomized controlled experiment (IRCE).

IRCE

- Ideal: subjects all follow the treatment protocol, perfect compliance and no errors in reporting.
- Randomized: subjects from the population of interest are randomly assigned to a treatment or control group (no confounding factors).
- Controlled: having a control group permits measuring the differential effect of the treatment.
- Experiment: the treatment is assigned as part of the experiment (no *reverse causality* in which subjects choose the treatment they think will work best).

class size effect on test scores

- Recall the class size example: causal effect of reducing class size (STR) on test scores.
- What would an IRCE be for measuring the effect on Test Score of reducing STR?
  - Students would be randomly assigned to classes, which would have different sizes.
  - Because students are randomly assigned, all student characteristics (ε) would be distributed independently of STR.
  - The zero conditional mean assumption holds trivially in an IRCE, i.e., $\mathbb{E}(\varepsilon|STR) = 0$.
- Observational data are not collected from IRCEs. They often come from surveys.
- The zero conditional mean assumption almost certainly does not hold, resulting in omitted variables bias.

Setting

- The linear regression model is given by

$$y_i = \beta_0 \cdot 1 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \tag{1}$$

  for $i = 1, 2, \ldots, n$, where $(\beta_0, \ldots, \beta_k)'$ are the (unknown) parameters.

- Going back the example, STR can be one of the x's.
- What are the roles of other x's?
- They are the so-called control variables: they control for the omitted variables bias.
- Suppose x_{1i} is the STR variable. Then, the role of the control variables can be represented as:

$$\mathbb{E}(\varepsilon_i | STR_i, x_{2i}, \ldots, x_{ki}) = \mathbb{E}(\varepsilon | x_{2i}, \ldots, x_{ki}),$$

  for $i = 1, 2, \ldots, n$, i.e., the conditional mean independence assumption.

Setting

- Hence, controlling for $x_2, \ldots, x_k$, STR does not covary with all student characteristics (as if students were randomly assigned).
- The (almost) causal effect of STR on test scores then can be obtained.
- We will see that if any of the control variables is missing from the specification, the least squares estimator for the slope of STR will generally be biased.
- More specifically, we will see that for the omitted variable bias to occur, the omitted variable(s) will have to covary with STR.

Setting

- The model can be written more compactly

$$y_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + \varepsilon_i \tag{2}$$

for $i = 1, 2, \ldots, n$, where $\boldsymbol{x}_i = (1, x_{1i}, \ldots, x_{ki})^{'}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^{'}$ are the (unknown) parameters.

- Stacking over i's we can equivalently write

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^{'}$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{'}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{'}$

The least-squares methodology

- Let $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i^{'})^{'}$ be a random vector, i.e., each element is a random variable.
- A random variable is a mapping, $x(\omega) : \Omega \mapsto \mathbb{R}$.
- As ω varies over Ω, we obtain realizations $x(\omega)$ ranging over $\mathbb{R}$.
- The terminology random variable is a bit unfortunate, because a random variable is neither random nor a variable.
- x is not a variable, it is a real valued function.
- x is not random, it is fixed. But, $\omega \in \Omega$ is random.
- The realized sample is one of the possible realizations. There is an underlying uncertainty in $x(\omega)$ and $y(\omega)$, and anything derived from them will inherit that uncertainty.

The least-squares methodology

- From hereon, let's denote the unknown true population value of the parameters by $\{\boldsymbol{\beta}_0^{'}, \sigma_0^2\}$ and arbitrary values by $\{\boldsymbol{\beta}^{'}, \sigma^2\}$.
- The least-squares methodology tries to find the best value for the unknown $\boldsymbol{\beta}_0$ so that the difference between the realized values of the outcome variable and the guesses from the model is the smallest in some sense.
- Let $\widehat{\boldsymbol{\beta}}$ the value picked. With this, we can compute the difference between the realized value of the outcome variable and the guess from the model (fitted value) for each i. Call this difference the residuals.
- We cannot simply add the residuals over the i's to find the best value for $\boldsymbol{\beta}$, because negatives and positives will negate.
- Instead, the least-squares minimizes the sum of the squared residuals to find the best approximation to $\boldsymbol{\beta}_0$.
- The least-squares estimator solves

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta})^2 = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{'}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{4}$$

The least-squares methodology

- Using some matrix algebra, finding the solution is not difficult.
- Let $\mathcal{S}(\boldsymbol{\beta})$ denote $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.
- The first-order condition for a minimum is

$$\frac{\partial \mathcal{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \overset{set}{=} 0. \tag{5}$$

- Then, the solution from (5) follows given $\boldsymbol{X}$ is full column rank, i.e., $\boldsymbol{X}'\boldsymbol{X}$ is invertible,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{6}$$

- To see that this solution is indeed a minimum, it suffices to note that $\mathcal{S}(\boldsymbol{\beta})$ is convex.

The least-squares methodology

- Equivalently, we can compute the second-order derivatives. The second-order condition for a minimum is

$$\frac{\partial^2 \mathcal{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 2\mathbf{X}'\mathbf{X} \tag{7}$$

must be positive definite.

- To see, $\boldsymbol{X}'\boldsymbol{X}$ is positive definite, let $\boldsymbol{c}$ be an arbitrary $(k+1) \times 1$ nonzero vector and let $\boldsymbol{v} = \boldsymbol{X}\boldsymbol{c}$.
- Then,

$$\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c} = \boldsymbol{v}'\boldsymbol{v} = \sum_{i=1}^{n} v_i^2. \tag{8}$$

- (8) equals zero only if $v_i = 0$ for $i = 1, \ldots, n$. It can happen only if the columns of $\boldsymbol{X}$ linearly dependent.
- This is not allowed if $\boldsymbol{X}$ is full column rank.

Properties

- $\boldsymbol{X}$ is orthogonal to the residuals by construction, i.e., the inner product of every column of $\boldsymbol{X}$ and $\widehat{\boldsymbol{\varepsilon}}$ is zero (they form a right angle):

$$\begin{aligned}\boldsymbol{X}^{'}\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{X}^{'}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}^{'}\boldsymbol{y} - \boldsymbol{X}^{'}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{'}\boldsymbol{y} - \boldsymbol{X}^{'}\boldsymbol{X}(\mathbf{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y} \\ = \boldsymbol{X}^{'}\boldsymbol{y} - \boldsymbol{X}^{'}\boldsymbol{y} = \mathbf{0}.\end{aligned} \tag{9}$$

- Recall that the first column in $\mathbf{X}$ is a column of 1's. Let $\boldsymbol{l}$ denote it. The least-squares residuals sum to zero. From (9),

$$\boldsymbol{X}^{'}\widehat{\boldsymbol{\varepsilon}} = (\boldsymbol{l}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)^{'}\widehat{\boldsymbol{\varepsilon}} = \begin{pmatrix} \boldsymbol{l}^{'}\widehat{\boldsymbol{\varepsilon}} \\ \boldsymbol{x}_1^{'}\widehat{\boldsymbol{\varepsilon}} \\ \vdots \\ \boldsymbol{x}_k^{'}\widehat{\boldsymbol{\varepsilon}} \end{pmatrix} = \mathbf{0}.$$

- Notice that $\boldsymbol{l}^{'}\widehat{\boldsymbol{\varepsilon}}$ is the sum of least squares residuals and equals zero.

Properties

- The regression hyperplane passes through the data means: $\bar{y} = \bar{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}$.
- From (9), we have $\boldsymbol{X}^{'}\boldsymbol{y} = \boldsymbol{X}^{'}\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Then,

$$\begin{pmatrix} \boldsymbol{l}^{'} \\ \mathbf{x}_1^{'} \\ \vdots \\ \mathbf{x}_k^{'} \end{pmatrix} \mathbf{y} = \begin{pmatrix} \boldsymbol{l}^{'} \\ \mathbf{x}_1^{'} \\ \vdots \\ \mathbf{x}_k^{'} \end{pmatrix} (\boldsymbol{l}, \mathbf{x}_1, \ldots, \mathbf{x}_k)\widehat{\boldsymbol{\beta}}.$$

- From the first row, $\boldsymbol{l}^{'}\boldsymbol{y} = (\boldsymbol{l}^{'}\boldsymbol{l}, \boldsymbol{l}^{'}\boldsymbol{x}_1, \ldots, \boldsymbol{l}^{'}\boldsymbol{x}_k)\widehat{\boldsymbol{\beta}}$, hence

$$\bar{y} = \frac{1}{n}\boldsymbol{l}^{'}\boldsymbol{y} = (\frac{1}{n}\boldsymbol{l}^{'}\boldsymbol{l}, \frac{1}{n}\boldsymbol{l}^{'}\boldsymbol{x}_1, \ldots, \frac{1}{n}\boldsymbol{l}^{'}\boldsymbol{x}_k)\widehat{\boldsymbol{\beta}} = \bar{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}.$$

Properties

- The mean of the fitted values from the regression equals the mean of the actual values: $\bar{\widehat{y}} = \bar{y}$. This is easily verified because $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

$$\bar{\widehat{y}} = \frac{1}{n}\boldsymbol{l}^{'}\mathbf{X}\widehat{\boldsymbol{\beta}} = \frac{1}{n}\boldsymbol{l}^{'}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \frac{1}{n}\boldsymbol{l}^{'}\widehat{\boldsymbol{\varepsilon}}$$
$$= \frac{1}{n}\boldsymbol{l}^{'}\left(\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\varepsilon}}\right) = \frac{1}{n}\boldsymbol{l}^{'}\boldsymbol{y} = \bar{y}.$$

Projection Matrices

- We will define two useful matrices. From the definition of the least squares residual

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\mathbf{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y} = \left(\mathbf{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\right)\boldsymbol{y} = \mathbb{M}_{\boldsymbol{x}}\boldsymbol{y}$$

  where $\mathbb{M}_{\boldsymbol{x}} = \left(\mathbf{I}_n - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\right)$.

- Also, notice that

$$\widehat{\boldsymbol{y}} = \boldsymbol{y} - \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \mathbb{M}_{\mathbf{x}}\boldsymbol{y} = \left(\mathbf{I}_n - \mathbf{I}_n + \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\right)\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y} = \mathbb{P}_{\boldsymbol{x}}\boldsymbol{y}$$

  where $\mathbb{P}_{\boldsymbol{x}} = \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$.

- We call $\mathbb{M}_{\boldsymbol{x}}$ and $\mathbb{P}_{\boldsymbol{x}}$ respectively as the orthogonal projection matrix and the projection matrix.

- Both matrices are symmetric and idempotent.

Projection Matrices

- Using $\mathbb{M}_{\boldsymbol{x}}$ and $\mathbb{P}_{\boldsymbol{x}}$, we have
  - $\boldsymbol{y} = \boldsymbol{X}^{'}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\varepsilon}} = \mathbb{P}_{\boldsymbol{x}}\boldsymbol{y} + \mathbb{M}_{\boldsymbol{x}}\boldsymbol{y}$,
  - $\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^{'}\mathbb{M}_{\boldsymbol{x}}^{'}\mathbb{M}_{\boldsymbol{x}}\boldsymbol{y} = \boldsymbol{y}^{'}\mathbb{M}_{\boldsymbol{x}}^{'}\boldsymbol{y} = \widehat{\boldsymbol{\varepsilon}}^{'}\boldsymbol{y}$,
  - $\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}} = (\boldsymbol{y} - \mathbb{P}_{\boldsymbol{x}}\boldsymbol{y})^{'}(\boldsymbol{y} - \mathbb{P}_{\boldsymbol{x}}\boldsymbol{y}) = \boldsymbol{y}^{'}\boldsymbol{y} - \boldsymbol{y}^{'}\mathbb{P}_{\boldsymbol{x}}^{'}\boldsymbol{y} - \boldsymbol{y}^{'}\mathbb{P}_{\boldsymbol{x}}\boldsymbol{y} + \boldsymbol{y}^{'}\mathbb{P}_{\boldsymbol{x}}^{'}\mathbb{P}_{\boldsymbol{x}}\boldsymbol{y} = \boldsymbol{y}^{'}\boldsymbol{y} - \boldsymbol{y}^{'}\mathbb{P}_{\boldsymbol{x}}^{'}\mathbb{P}_{\boldsymbol{x}}\boldsymbol{y} = \boldsymbol{y}^{'}\boldsymbol{y} - \widehat{\boldsymbol{\beta}}^{'}\boldsymbol{X}^{'}\boldsymbol{X}\widehat{\boldsymbol{\beta}}$,
  - $\mathbb{P}_{\boldsymbol{x}}\boldsymbol{X} = \boldsymbol{X}$ and $\mathbb{P}_{\boldsymbol{x}}\mathbb{M}_{\boldsymbol{x}} = \mathbf{0}$.

Goodness of fit

- A special symmetric idempotent matrix is the following

$$\mathbb{M}_0 = \mathbf{I}_n - \boldsymbol{l}_n(\boldsymbol{l}_n'\boldsymbol{l}_n)^{-1}\boldsymbol{l}_n' = \mathbf{I}_n - \frac{1}{n}\boldsymbol{l}_n\boldsymbol{l}_n'$$
$$= \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix}$$

  where $\boldsymbol{l}_n$ is $n \times 1$ vector of 1's.

- $\mathbb{M}_0$ is often called the deviation-from-the-mean matrix.

- Write the linear regression model in deviation from the mean form:

$$\mathbb{M}_0\boldsymbol{y} = \mathbb{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \mathbb{M}_0\widehat{\boldsymbol{\varepsilon}} = \mathbb{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\varepsilon}}.$$

- The second equality follows because $\mathbb{M}_0\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\varepsilon}} - \frac{1}{n}\boldsymbol{l}_n\boldsymbol{l}_n'\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\varepsilon}}$ as $\boldsymbol{l}_n'\widehat{\boldsymbol{\varepsilon}} = 0$.

Goodness of fit

- We can write the sum of the squared deviations of the elements of $\boldsymbol{y}$ from their mean

$$\begin{aligned}(\mathrm{M}_0\boldsymbol{y})^{'}\mathrm{M}_0\boldsymbol{y} = \boldsymbol{y}^{'}\mathrm{M}_0\boldsymbol{y} &= (\mathrm{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\widehat{\boldsymbol{\varepsilon}})^{'}(\mathrm{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\widehat{\boldsymbol{\varepsilon}}) \\ &= \widehat{\boldsymbol{\beta}}^{'}\boldsymbol{X}^{'}\mathrm{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}^{'}\boldsymbol{X}^{'}\widehat{\boldsymbol{\varepsilon}}+\widehat{\boldsymbol{\varepsilon}}^{'}\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}} \\ &= \widehat{\boldsymbol{\beta}}^{'}\boldsymbol{X}^{'}\mathrm{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}} \\ &= \widehat{\boldsymbol{y}}^{'}\mathrm{M}_0\widehat{\boldsymbol{y}}+\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}}.\end{aligned}$$

- Recall that $\bar{\widehat{y}} = \bar{y}$ and $\widehat{\boldsymbol{y}} - \boldsymbol{l}_n\bar{\widehat{y}} = \widehat{\boldsymbol{y}} - \boldsymbol{l}_n\bar{y}$ when an intercept is included in $\boldsymbol{X}$.

- Let $\boldsymbol{y}^{'}\mathrm{M}_0\boldsymbol{y}$, $\widehat{\boldsymbol{y}}^{'}\mathrm{M}_0\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}}$ be denoted respectively by total sum of squares (TSS), explained sum of squares (ESS) and residual sum of squares (RSS).

- Clearly, $TSS = ESS + RSS$, where ESS is the portion of the variation in $\boldsymbol{y}$ explained by the model, and the RSS is the portion of the variation in $\boldsymbol{y}$ unexplained by the model.

Goodness of fit

- A natural measure of goodness-of-fit is

$$R^2 = \frac{ESS}{TSS} = \frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\mathbb{M}_0\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{\boldsymbol{y}'\mathbb{M}_0\boldsymbol{y}} = 1 - \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{\boldsymbol{y}'\mathbb{M}_0\boldsymbol{y}} = 1 - \frac{RSS}{TSS}$$

  which is the square of the sample correlation between $\boldsymbol{y}$ and $\mathbb{P}_{\boldsymbol{x}}\boldsymbol{y}$.

- We will refer to this measure as the coefficient of determination. It lies between 0 and 1.

- One serious drawback of R^2 is that in never decreases as we include more explanatory variables in $\boldsymbol{X}$, even when they are irrelevant (superfluous) in explaining $\boldsymbol{y}$.

- Hence, we modify R^2 so that it penalizes for superfluous regressors.

- The modified measure is called the adjusted R^2 and is given by

$$\bar{R}^2 = 1 - \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}/(n-k-1)}{\boldsymbol{y}'\mathbb{M}_0\boldsymbol{y}/(n-1)}.$$

Goodness of fit

- As k increases relative to the number of observations the last term increases and $\bar{R}^2$ falls.
- One can show that
$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\left(1 - R^2\right).$$
- In fact, $\bar{R}^2$ increases if and only if the t-statistic of a newly added regressor is greater than one in absolute value and $\bar{R}^2$ may even get negative values.
- For linear regression models without an intercept, the interpretation of R^2 is not easily available and R^2 should not be used for model comparison.
- Instead, one can use the following measures for model comparison (including nonlinear models):
  - Akaike Information Criterion:
$$AIC = \log\left(\frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{n}\right) + \frac{2(k+1)}{n},$$
  - Bayesian (or Schwartz) Information Criterion:
$$BIC = \log\left(\frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{n}\right) + \frac{k\log(n)}{n}.$$

Assumptions

- Characterizing the finite sample distribution of the least-squares estimator is feasible, which is often not the case for many other estimators.
- We make the following assumptions:

1. $\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbf{0}$ a.s.,
2. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{'}|\boldsymbol{X}) = \sigma_0^2\mathbf{I}_n$ a.s.,
3. rank$(\boldsymbol{X}) = k+1$ a.s.

- These assumptions correspond to the statistical model of random regressors with independent sampling.
- The first two assumptions mostly rule out models of time-series data.
- By their nature, time series data almost always exhibit dependence of data points.

Remarks

- σ_0^2 is a constant not depending on $\boldsymbol{X}$. $\sigma_0^2 \boldsymbol{I}_n$ is the conditional covariance of $\boldsymbol{\varepsilon}$, but since it does not depend on $\boldsymbol{X}$, it is also the unconditional covariance of $\boldsymbol{\varepsilon}$.
- From the first two assumptions, we have $\mathbb{E}(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma_0^2$ for $i = 1, \ldots, n$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$ (unconditionally).
- The independent random sampling provides the uncorrelatedness.
- However, uniformity of the variances is simply of convenience, but otherwise , it is necessary to know what the variances are.
- The last assumption is needed for the existence of the least squares estimator. Failure of this assumption is known as perfect collinearity.
- Near failure of this assumption results in $\boldsymbol{X}^{'} \boldsymbol{X}$ being poorly conditioned for inversion. It is known as multicollinearity.

Unbiasedness

Claim

Assumptions 1–3 imply that $\widehat{\boldsymbol{\beta}}$ is unbiased.

Proof.

We can write

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}) = \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}.$$

Hence, the expectation of $\widehat{\boldsymbol{\beta}}$ conditional on $\boldsymbol{X}$ is,

$$\begin{aligned}\mathbb{E}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \boldsymbol{\beta}_0 + \mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}|\boldsymbol{X}\right] \\ &= \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) \\ &= \boldsymbol{\beta}_0,\end{aligned}$$

where we used Assumption 1. Furthermore, by the LIE, the unconditional expectation of $\widehat{\boldsymbol{\beta}}$ is

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbb{E}\left[\mathbb{E}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})\right] = \mathbb{E}(\boldsymbol{\beta}_0) = \boldsymbol{\beta}_0.$$

□

Unbiasedness

- In repeated samples (i.e., sample infinitely many times), the distribution of the estimator is centered on the true value $\boldsymbol{\beta}_0$.
- In repeated samples, the estimator captures the truth on average.
- This is a desirable property for an estimator, but by no means a sufficient condition for the estimator to be useful.
- Next, we look at the bias property of the least-squares estimator of σ_0^2.
- Recall that the least squares estimator simply uses the sample variance of the residuals, $\widehat{\sigma}^2 = \widehat{\boldsymbol{\varepsilon}}' \widehat{\boldsymbol{\varepsilon}}/n$.
- Recall that

$$\widehat{\boldsymbol{\varepsilon}} = \mathbb{M}_{\boldsymbol{x}} \boldsymbol{y} = \mathbb{M}_{\boldsymbol{x}} (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbb{M}_{\boldsymbol{x}} \boldsymbol{\varepsilon}$$

since $\mathbb{M}_{\boldsymbol{x}} \boldsymbol{X} = \mathbf{0}$.

Unbiasedness

- Then, $\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}} = (\mathbb{M}_{\boldsymbol{x}}\boldsymbol{\varepsilon})^{'}(\mathbb{M}_{\boldsymbol{x}}\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^{'}\mathbb{M}_{\boldsymbol{x}}\boldsymbol{\varepsilon}$.
- Using the fact that trace of scalar equals itself and by Assumption 2, we have

$$\begin{aligned}\mathbb{E}(\boldsymbol{\varepsilon}^{'}\mathbb{M}_{\boldsymbol{x}}\boldsymbol{\varepsilon}|\boldsymbol{X}) &= \mathbb{E}\left[\text{tr}(\boldsymbol{\varepsilon}^{'}\mathbb{M}_{\boldsymbol{x}}\boldsymbol{\varepsilon})|\boldsymbol{X}\right] = \mathbb{E}\left[\text{tr}(\mathbb{M}_{\boldsymbol{x}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{'})|\boldsymbol{X}\right] \\ &= \text{tr}\left[\mathbb{M}_{\boldsymbol{x}}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{'}|\boldsymbol{X})\right] = \sigma_0^2\text{tr}(\mathbb{M}_{\boldsymbol{x}}).\end{aligned}$$

- Furthermore, $\text{tr}(\mathbb{M}_{\boldsymbol{x}}) = \text{tr}(\mathbf{I}_n) - \text{tr}\left[\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\right] = \text{tr}(\mathbf{I}_n) - \text{tr}\left[(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{X}\right] = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{I}_{k+1}) = n - k - 1$.
- Hence,

$$\mathbb{E}(\widehat{\sigma}^2|\boldsymbol{X}) = \frac{1}{n}\mathbb{E}(\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = \frac{1}{n}\mathbb{E}(\boldsymbol{\varepsilon}^{'}\mathbb{M}_{\boldsymbol{x}}\boldsymbol{\varepsilon}|\boldsymbol{X}) = \left(\frac{n-k-1}{n}\right)\sigma_0^2.$$

- Also, by the LIE, this is true unconditionally.

Unbiasedness

- This result shows that the least-squares estimator $\widehat{\sigma}^2$ underestimates σ_0^2 (biased towards zero), but the bias is small for large samples.
- The unbiased least-squares estimator of σ_0^2 is

$$\widetilde{\sigma}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^{'}\widehat{\boldsymbol{\varepsilon}}}{n-k-1}$$

where $\widetilde{\sigma}$ is called the standard error of the regression.

- The sampling variance of $\widehat{\boldsymbol{\beta}}$ can be driven as follows:

$$\begin{aligned}
\mathrm{Var}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)^{'}|\boldsymbol{X}\right] \\
&= (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{'}|\boldsymbol{X}\right)\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1} \\
&= \sigma_0^2(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}.
\end{aligned}$$

- Unlike the conditional mean of $\widehat{\boldsymbol{\beta}}$, conditional variance of $\widehat{\boldsymbol{\beta}}$ depends on $\boldsymbol{X}$.
- Hence, $\sigma_0^2(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}$ is not the unconditional variance of $\widehat{\boldsymbol{\beta}}$, unless $\boldsymbol{X}$ is deterministic.

Sampling variance of the LSE

- Applying the LIE, we get $\text{Var}(\widehat{\boldsymbol{\beta}}) = \mathbb{E}(\text{Var}(\widehat{\boldsymbol{\beta}})|\boldsymbol{X}) = \sigma_0^2 \mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]$.
- Also, note that $\mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \neq \left[\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})\right]^{-1}$.
- We must assume that $\mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]$ exists, otherwise the variance, conditional or unconditional, is not well-defined.[1]
- For randomly drawn $\boldsymbol{X}$ in repeated samples from an experiment, the dispersion of the distribution of $\widehat{\boldsymbol{\beta}}$ is given by $\sigma_0^2 \mathbb{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]$, i.e., the average case.
- But for practical purposes, the conditional variance can be the relevant measure of dispersion.
- Since $\boldsymbol{X}$ is observed, in repeated samples one can consider confining attention to those samples with this $\boldsymbol{X}$ and the relevant measure of dispersion becomes the conditional variance.
- But, note that variance (of the distribution of an estimator) is used to compare the efficiency of estimators, hence the average case is the natural choice in that case.

[1] For any random variables y and x, $\mathbb{E}(y|x)$ is well-defined only if $\mathbb{E}|y| < \infty$.

Efficiency of the LSE

- Let $\boldsymbol{L} \in \mathbb{R}^{(k+1)\times n}$ some matrix such that $\boldsymbol{\beta}^{\dagger} = \boldsymbol{L}\boldsymbol{y}$.
- Here, $\boldsymbol{\beta}^{\dagger}$ denotes a class of linear estimators of $\boldsymbol{\beta}_0$.
- The Gauss-Markov Theorem claims that the least-squares estimator $\widehat{\boldsymbol{\beta}}$ is the best in the sense that within the class of linear estimators of $\boldsymbol{\beta}_0$ the least squares estimator has the smallest variance.
- In other words, it makes the most efficient use of the information in a given sample.
- Before, we proceed to state the theorem and its proof, we make the following assumptions:

4. $\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}, \boldsymbol{L}) = \mathbf{0}$ a.s.
5. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{'}|\boldsymbol{X}, \boldsymbol{L}) = \sigma_0^2 \boldsymbol{I}_n$ a.s.
6. $\boldsymbol{L}\boldsymbol{X} = \mathbf{I}_{k+1}$ a.s.

Efficiency of the LSE

- Given these assumptions we have

$$\mathbb{E}(\boldsymbol{\beta}^{\dagger}|\boldsymbol{X},\boldsymbol{L}) = \boldsymbol{L}\boldsymbol{X}\boldsymbol{\beta}_0 + \mathbb{E}(\boldsymbol{L}\boldsymbol{\varepsilon}|\boldsymbol{X},\boldsymbol{L}) = \boldsymbol{\beta}_0,$$

  and using the LIE we obtain

$$\mathbb{E}(\boldsymbol{\beta}^{\dagger}) = \mathbb{E}(\mathbb{E}(\boldsymbol{\beta}^{\dagger}|\boldsymbol{X},\boldsymbol{L})) = \mathbb{E}(\boldsymbol{L}\boldsymbol{X})\boldsymbol{\beta}_0 + \mathbb{E}(\boldsymbol{L}\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X},\boldsymbol{L})) = \boldsymbol{\beta}_0.$$

- So, $\boldsymbol{\beta}^{\dagger}$ is unbiased.
- Furthermore, the sampling variance of $\boldsymbol{\beta}^{\dagger}$ is given by

$$\mathrm{Var}(\boldsymbol{\beta}^{\dagger}|\boldsymbol{X},\boldsymbol{L}) = \mathbb{E}\left[\left(\boldsymbol{\beta}^{\dagger}-\boldsymbol{\beta}\right)\left(\boldsymbol{\beta}^{\dagger}-\boldsymbol{\beta}\right)^{'}|\boldsymbol{X},\boldsymbol{L}\right] = \mathbb{E}\left(\boldsymbol{L}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{'}\boldsymbol{L}^{'}|\boldsymbol{X},\boldsymbol{L}\right) = \sigma_0^2\boldsymbol{L}\boldsymbol{L}^{'}$$

  and by the LIE we have

$$\mathrm{Var}(\boldsymbol{\beta}^{\dagger}) = \sigma_0^2\mathbb{E}(\boldsymbol{L}\boldsymbol{L}^{'}).$$

Efficiency of the LSE

Theorem (Gauss-Markov)

The difference between Var($\boldsymbol{\beta}^{\dagger}$) and Var($\widehat{\boldsymbol{\beta}}$) is a positive semi-definite matrix for every $\boldsymbol{L}$ satisfying Assumptions 4–6.

- Note that we have not used the terminology BLUE here.
- If indeed we had deterministic regressors, we would simply state the theorem as the least-squares estimator is BLUE.
- However, this is not necessarily true for a regression model with stochastic regressors.

Efficiency of the LSE

Proof.

Let $\boldsymbol{D} = \boldsymbol{L} - (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$, so that $\boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}$ by assumption. Then, note that

$$\begin{aligned}\boldsymbol{L}\boldsymbol{L}^{'} &= (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1} + (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{D}^{'} + \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{D}^{'} \\ &= (\boldsymbol{X}^{'}\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{D}^{'}.\end{aligned}$$

It follows that

$$\text{Var}(\boldsymbol{\beta}^{\dagger}|\boldsymbol{X},\boldsymbol{L}) = \text{Var}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{L}) + \sigma_0^2\boldsymbol{D}\boldsymbol{D}^{'},$$

where $\boldsymbol{D}\boldsymbol{D}^{'}$ positive semi-definite (well-known result). Then, taking unconditional expectations

$$\text{Var}(\boldsymbol{\beta}^{\dagger}) = \text{Var}(\widehat{\boldsymbol{\beta}}) + \sigma_0^2\mathbb{E}(\boldsymbol{D}\boldsymbol{D}^{'}),$$

and noting that for arbitrary fixed $(k+1)$-vector $\boldsymbol{c}$,

$$\boldsymbol{c}^{'}\mathbb{E}(\boldsymbol{D}\boldsymbol{D}^{'})\boldsymbol{c} = \mathbb{E}(\boldsymbol{c}^{'}\boldsymbol{D}\boldsymbol{D}^{'}\boldsymbol{c}) \geq 0$$

since $\boldsymbol{c}^{'}\boldsymbol{D}\boldsymbol{D}^{'}\boldsymbol{c}$ is non-negative (a sum of squares) for any $\boldsymbol{D}$. □

Efficiency of the LSE

- Note again that in the non-stochastic $\boldsymbol{X}$ and $\boldsymbol{L}$ case, Assumption 6 holds for every member of the class $\boldsymbol{\beta}^{\dagger}$.
- Hence $\widehat{\boldsymbol{\beta}}$ is said to be BLUE, best (minimum variance) in the class of linear unbiased estimators.
- But, in the stochastic case, note that

$$\mathbb{E}(\boldsymbol{\beta}^{\dagger}) = \mathbb{E}(\boldsymbol{L}\boldsymbol{X})\boldsymbol{\beta}_0 + \mathbb{E}(\boldsymbol{L}\boldsymbol{\varepsilon})$$

  and given Assumption 4, it is sufficient for unbiasedness if $\mathbb{E}(\boldsymbol{L}\boldsymbol{X}) = \mathbf{I}_{k+1}$.
- This is weaker than Assumption 6.
- If $\boldsymbol{L}\boldsymbol{X} \neq \mathbf{I}_{k+1}$ with positive probability, the estimator is conditionally biased.
- However, under $\mathbb{E}(\boldsymbol{L}\boldsymbol{X}) = \mathbf{I}_{k+1}$, the bias terms average out over the distribution of $\boldsymbol{L}\boldsymbol{X}$.

Efficiency of the LSE

- Therefore, the estimator is unbiased in repeated sampling under the statistical model with stochastic regressors.
- But, we cannot conclude from $\mathbb{E}(\boldsymbol{L}\boldsymbol{X}) = \mathbf{I}_{k+1}$ that $\mathbb{E}(\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}) = \mathbf{0}$, therefore the proof above breaks down.
- Hence, in such a case, although the estimator is unbiased, we cannot show that the least squares estimator is at least as efficient as this linear estimator.
- Therefore, for a statistical model with random regressors, it is not correct to claim that $\widehat{\boldsymbol{\beta}}$ is BLUE (except in the context of the conditional distribution).
- In practice, in a repeated sampling experiment where $\boldsymbol{X}$ is held fixed and a new $\boldsymbol{\varepsilon}$ is sampled in each replication of the experiment, the least squares has the variance of $\sigma_0^2(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}$ and therefore the most efficient in the class of linear unbiased estimators. However, if both $\boldsymbol{\varepsilon}$ and $\boldsymbol{X}$ are sampled in each replication of the experiment, the least squares has the variance of $\sigma_0^2\mathbb{E}\big[(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\big]$, and we cannot claim that the least squares estimator is the best in the linear unbiased class. But, there is no obvious candidate for a more efficient estimator, and this result is more formal than practical significance.

Finite sample properties of the LSE

- To sum, Assumptions 1–3 are sufficient to characterize the mean and variance of the finite sample distribution of the least-squares estimator.
- Note that we have not specified any distribution for $\boldsymbol{y}|\boldsymbol{X}$, such as the normal distribution or some other distribution.
- Indeed, if we further make the normality assumption, we can characterize the exact distribution of the least-squares estimator in finite samples.
- In other words, if we additionally assume that $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}_0,\, \sigma_0^2 \mathbf{I}_n)$, then

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim N(\boldsymbol{\beta}_0, \sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

- Note again that this is the exact distribution of $\widehat{\boldsymbol{\beta}}|\boldsymbol{X}$. We are not utilizing any approximation methods (large sample methods).

Finite sample properties of the LSE

- You can already see that this result can be used for inference.
- In significance testing exercises, we are interested in whether a given regressor belongs to the model.
- Say, we'd like to test if the jth regressor belongs to the model, i.e., $H_0 : \beta_{j0} = 0$.
- We can compute
$$\frac{\widehat{\beta}_j - 0}{\sqrt{(\sigma_0^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{jj}}} \sim N(0,1).$$
where jj subscript denotes the jth row jth column of the matrix.
- This is the t-test statistic you are familiar with. The only problem is that it is not feasible because we do not know σ_0^2.
- We replace the unknown σ_0^2 with its unbiased estimator $\widetilde{\sigma}^2$ to make it feasible:
$$\frac{\widehat{\beta}_j - 0}{\sqrt{(\widetilde{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1})_{jj}}} \sim t(n-k-1).$$

Remarks

- For Gauss-Markov theorem to hold the normality assumption is not necessary.
- Note that we have not characterized the large sample distribution of the least-squares estimator.
- The behavior of the least-squares estimator can be studied in large samples (as n grows without a bound) using the so-called large sample tools such as LLNs and CLTs.
- Indeed, it is straightforward to show that the LSE in large samples is approximately distributed as normal, and the larger n is the better the approximation is.
- Also, there is a large sample counterpart of the Gauss-Markov theorem.
- The t-test statistic is also approximately normal in large samples, and the larger n is the better the approximation is.

Hypothesis testing

- The term hypotheses testing refers to the process of trying to decide the truth or falsity of hypotheses on the basis of experimental/observational evidence.
- Experimental/observational measurements are subject to random error.
- Therefore, any decision about the truth or falsity of the hypothesis, based on experimental/observational evidence, also is subject to error.
- It will not be possible to avoid an occasional decision error, but it will be possible to construct tests so that such errors occur infrequently at some prescribed rate.
  1. Type I error: Reject a true H_0.
  2. Type II error: Fail to reject a false H_0.
- For a simple null hypothesis, the probability of rejecting a true H_0, is referred to as the significance level of the test.

Hypothesis testing

Hypothesis testing involve the following steps:

- Decide on your tolerance level (or level of significance) for making a Type I error, i.e., out of 100 times how many times you're willing to reject a true null hypothesis?
- Calculate a test statistic $\mathcal{T}_n$.
- Obtain the critical value c and
  - reject H_0 if $\mathcal{T}_n > c$,
  - fail to reject H_0 if $\mathcal{T}_n \leq c$.