The Linear Regression Model

Paolo Zacchia

Econometric Theory

Lecture 7

Linear socio-economic relationships (1/2)

- This lecture introduces the **multivariate** linear regression model (the bivariate version is introduced in Lectures 3, 4 and 6; alongside some suitable estimators).

- In this model, a **dependent** variable Y_i, K independent or **explanatory** variables $\boldsymbol{x} = (X_{1i}, \ldots, X_{Ki})$, and in addition an unobserved "disturbance" or **error term** ε_i are related through a linear function.

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \varepsilon_i$$

- For the moment, no statistical assumption (like say a linear CEF) are imposed on this model.

- Statistical assumptions are introduced **after** the discussion of what is called here the "least squares solution" – and its algebraic properties.

Linear socio-economic relationships (2/2)

- Consider a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^{N}$ of size $N > K$, where y_i is the **realization** of Y_i for the i-th observation, while $\mathbf{x}_i$:

$$\mathbf{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ki} \end{bmatrix}$$

  is a vector of length K that collects all the **realizations** of the explanatory variables in $\boldsymbol{x}$ for the i-th observation.

- One can write the model *in terms of realizations* as follows.

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i$$

- The **parameter vector** of the model:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_K \end{bmatrix}^{\mathrm{T}}$$

  is the object of analysis and estimation.

The linear model: discussion

- How to think about these relationships? Traditionally, they received a *structural* interpretation.

- They were postulated as a representation of socio-economic phenomena, where an *endogenous* variable Y_i depends upon some K independent, *exogenous* variables $(X_{1i}, \ldots, X_{Ki})$ in a linear fashion.

- According to this interpretation the error term ε_i represents all other, generally unknown factors that also determine Y_i.

- Furthermore, ε_i removes determinism from the relationship (it can hold imperfectly in the data).

- This interpretation is partially outdated, but it can still be instructive from a pedagogical standpoint.

Constant terms

- Linear models usually include a **constant term** as follows.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{(K-1)} X_{(K-1)i} + \varepsilon_i$$

  Beside β_0, the model has $K-1$ independent variables X_{ki}, so that in total the model still has K parameters.

- Thus, the model can still be written as $y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i$, but here vector $\mathbf{x}_i$ is such that as $x_{i0} = 1$ for each observation.

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ x_{1i} \\ \vdots \\ x_{(K-1)i} \end{bmatrix}$$

- Constant terms allow to confidently assume that $\mathbb{E}[\varepsilon_i] = 0$: a nonzero mean can always be incorporated into β_0.

Introducing examples

- The next few slides introduce some very classical examples of linear models from econometrics.

- The first example, that concerns the *Keynesian consumption function*, deals with a **time-series** environment, one where variables are observed at different points in time.

- The second example, about the *Mincer equation*, is framed in a **cross-sectional** environment: that is, one where variables are observed across different socio-economic units.

- The second example may also extend to a **longitudinal** or **panel** dimension, one where variables are observed both at different points in time and across different units.

- For both examples, a "tentative" structural interpretation is offered: treat it with caution!

The Keynesian consumption function: history

- Macroeconomics emerged as a distinct discipline following the publication of Keynes' *General Theory*.

- This time period corresponded with the initial attempts at mathematically modeling macroeconomic phenomena.

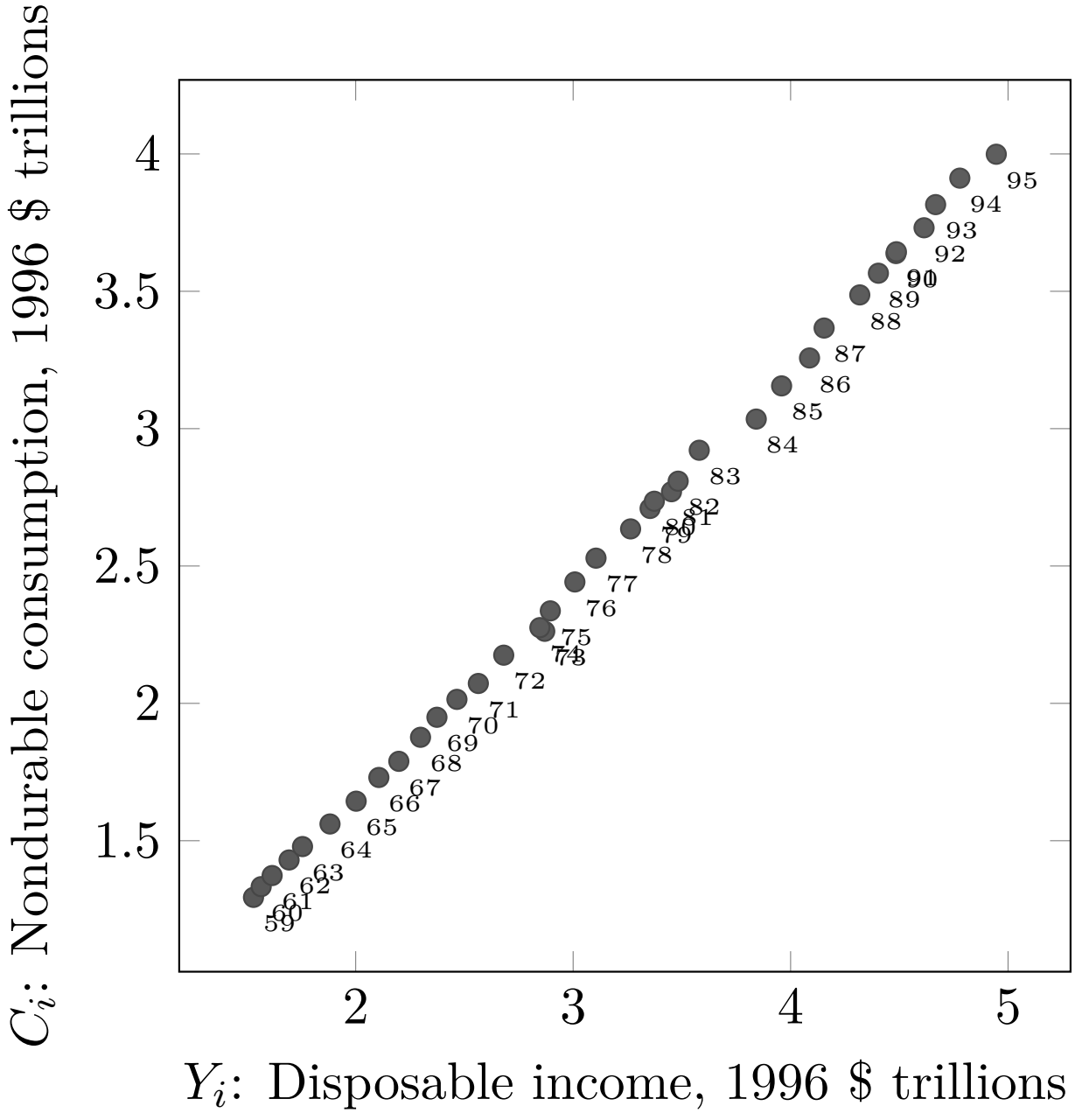
- One example is the "Keynesian" consumption function:

$$C_i = \mathrm{c}_0 + \mathrm{c}_1 Y_i + \varepsilon_i$$

  where C_i represents aggregate **consumption**, Y_i aggregate **disposable income**, ε_i is a disturbance term, and $(\mathrm{c}_0, \mathrm{c}_1)$ are the two parameters of interest.

- In particular, c_1 measures the dependence of C_i on Y_i: it is called **marginal propensity of consumption** and it was central in the economic policy of the time.

The Keynesian consumption function: data



The Keynesian consumption function: today

- The figure showcases the relationship between Y_i and C_i in historical time series data from the United States.
- The relationship displayed therein appears so linear that it seems "too good to be true..."
- Econometric research has shown that this relationship is, in fact, a statistical artifact.
- Specifically, it is believed to be a *spurious* relationship: one that is generated by the common dependence of both series upon other, external factors.
- This insight led economists to construct more sophisticated macroeconomic and macroeconometric models.

Human capital and wages: a theory

- Economic theories of the labor market postulate that wages depend upon workers' productivity.

- Productivity in turn depends upon all the skills that people develop in their life (with their education, experience, etc.). This goes under the name of **human capital**.

- In his seminal work, Mincer (1958) postulated the following structural relationship:

$$W_i = \exp\left(\beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i\right) \exp\left(\alpha_i + \epsilon_i\right)$$

  where, for every individual i: W_i is her/his **wage**, S_i is the acquired **education**, X_i is the **on-the-job experience**, α_i is **ability**, ϵ_i is the residual error term (e.g. **"luck"**).

- Typically α_i is hard to observe: it adds to the error term.

Human capital and wages: an empirical model

- Taking logs on both sides of such a structural relationship, one obtains the standard baseline **Mincer equation**:

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + \alpha_i + \epsilon_i$$

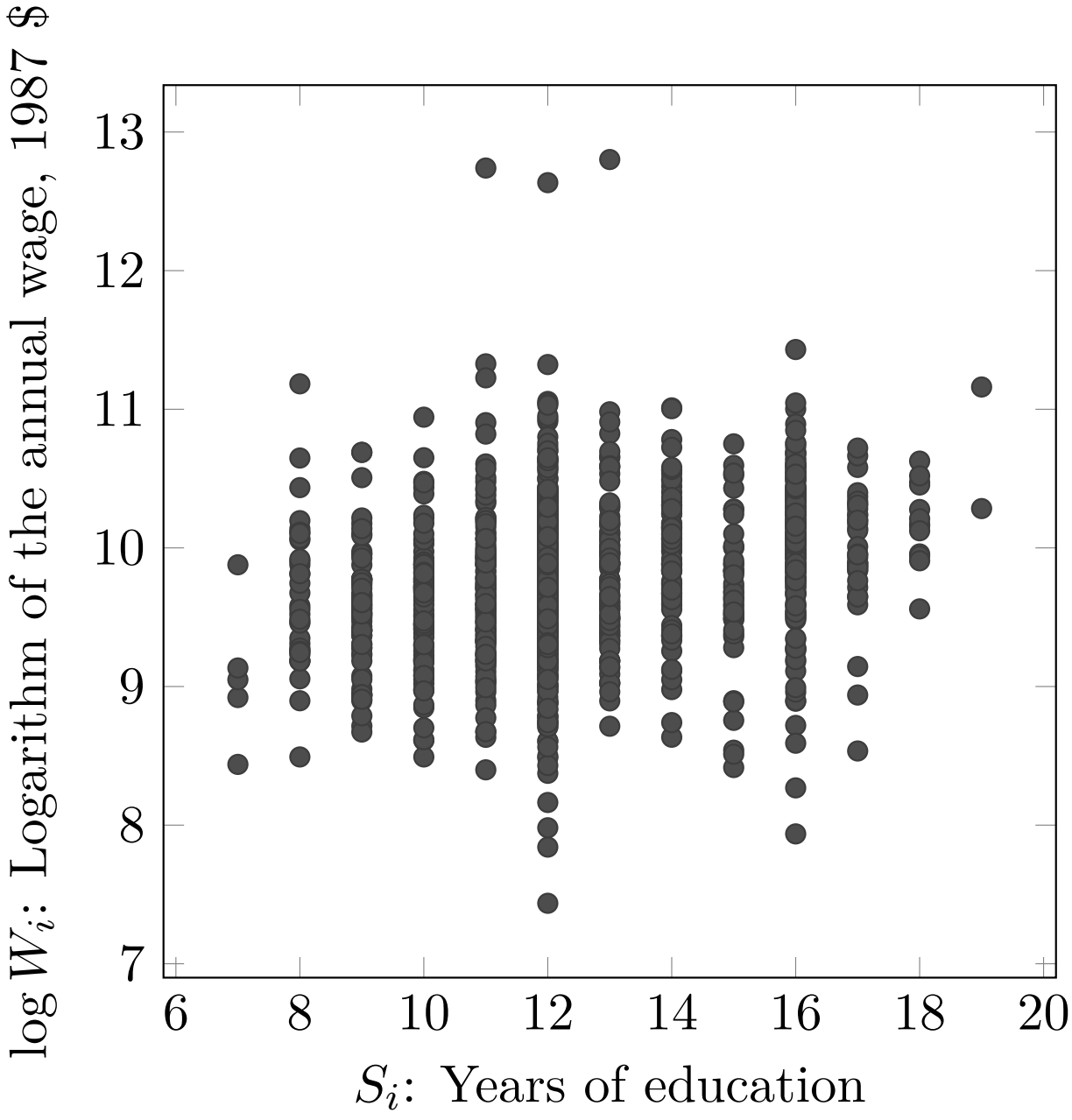
  while non-linear in its variables, this equation is **linear in the parameters** $\boldsymbol{\beta}_W = \{\beta_0, \beta_1, \beta_2, \beta_3\}$: thus, it is suited to analysis via the multivariate linear model.

- In this model, most of the interest falls on parameter β_3: it represents the **returns of education** to individual **wages**: a parameter of interest for policy!

- The following figure: a cross-sectional excerpt from a longer panel survey (only data from 1987 were selected) displays a typical empirical relationship between $\log W_i$ and S_i.

- Note the "discrete" measure of S_i, typical of "micro" data.

Human capital and wages: data



Human capital, ability and wages

- The impossibility to proper measure α_i causes problems to the empirical analysis of this model.

- In short: are the empirical measures of β_3 due to education, or due to the empirical link between education and ability?

- In econometrics, this problem is called **endogeneity** and is elaborated in more depth in later Lectures.

- In more elaborate studies, the Mincer equation is implicitly or explicitly augmented with a **model for education**, like:

$$S_i = \gamma_0 + \gamma_1 Z_i + \phi_1 X_i + \phi_2 X_i^2 + \psi_0 \alpha_i + \eta_i$$

  where Z_i is some other factor affecting the individual choice to pursue more education, while η_i is another error term.

- This extension is analyzed later in Lectures 9 and 10.

Back to the model: compact notation

Let us return to the more general analysis of the linear model. It is useful to develop a **compact matrix notation**. Let:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}; \ \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_N^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{K1} \\ x_{12} & x_{22} & \dots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \dots & x_{KN} \end{bmatrix}; \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

that is, $\mathbf{y}$, $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are obtained by vertically stacking over **all observations**, respectively, the realization y_i of the dependent variable, the transpose of the vector $\mathbf{x}_i$, and the error term ε_i.

- Observe that if the model includes a constant term β_0, the first column of $\mathbf{X}$ is a vector of ones.

In econometrics, it is quite typical to describe models *in terms of realizations*: so the linear model can be expressed as follows.

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

Optimal prediction

- The objective of the analysis is to "evaluate" the parameter vector $\boldsymbol{\beta}$ in light of the data.

- There shall be no talk of "estimation" for now – estimation is discussed later alongside statistical assumptions.

- Hence, the "least squares solution" is initially motivated by a **population prediction problem**.

- Specifically, this is the problem of specifying a **prediction function** written as follows.

$$\widehat{y}_i = m_y\left(\mathbf{x}_i\right)$$

  This is meant to yield the "best guess" for some **unknown** value of $Y_i = y_i$, on the basis the observation of a vector of independent variables $\mathbf{x}_i$.

- How to choose the "optimal" prediction function $m_y\left(\mathbf{x}_i\right)$?

Loss functions

- It is obvious that the farther away the prediction $\widehat{y}_i$ is from the actual realization of Y_i, the worse it is for the analysts.

- This implies the existence of some **loss function**:

$$L\left(e_i\right) = L\left(y_i - \widehat{y}_i\right)$$

  with $e_i \equiv y_i - \widehat{y}_i$ and where $L\left(e_i\right)$ has the properties that it is increasing in $|y_i - \widehat{y}_i|$ and that $L\left(0\right) = 0$.

- If analysts aim at specifying a prediction function that is as consistent across different realizations of $\mathbf{x}_i$ as it is possible, a sensible criterion is to choose $m_y\left(\mathbf{x}_i\right)$ so that it minimizes the **expected loss**:

$$\mathbb{E}\left[L\left(Y_i - \widehat{Y}_i\right)\right] = \mathbb{E}\left[L\left(Y_i - m_y\left(\boldsymbol{x}_i\right)\right)\right]$$

  where the expectation is taken over the joint support of Y_i and $(X_{1i}, \ldots, X_{Ki})$.

Quadratic loss, mean squared error, and more

- Which actual loss function $L(e_i)$ to choose, however?
- Here the focus is on the **quadratic loss** $L(e_i) = e_i^2$: it is appealing because prediction errors are disproportionately more "harmful" the higher they are.
- The expected quadratic loss is called the **mean squared error** of prediction: a related result is discussed next.

$$\text{MSE} = \mathbb{E}\left[(Y_i - m_y(\boldsymbol{x}_i))^2\right]$$

- Alternatives include the **absolute loss** $L(e_i) = |e_i| \ldots$
- $\ldots$ or the more general **quantile loss**, for $p \in (0, 1)$.

$$L(e_i) = p\,|e_i| \cdot \mathbb{1}[e_i \geq 0] + (1-p)\,|e_i| \cdot \mathbb{1}[e_i < 0]$$

The quantile loss is asymmetric: for prediction errors of the same absolute size $|e_i|$, it punishes *underprediction* ($e_i > 0$) more than *overprediction* ($e_i < 0$) if $p > 0.5$, and vice versa.

Optimal predictor and the CEF (1/2)

Theorem 1

CEF as Optimal Predictor under Quadratic Loss. *Under the condition that* $\mathbb{V}\mathrm{ar}\left[Y_i|\,\boldsymbol{x}_i\right] < \infty$*, the predictor* $m_y\left(\boldsymbol{x}_i\right)$ *that minimizes the Mean Squared Error is the Conditional Expectation Function of* Y_i *given* $\boldsymbol{x}_i$ *(CEF):* $m_y\left(\boldsymbol{x}_i\right) = \mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right]$.

Proof.

The standard decomposition of the MSE from Lecture 5 applies also to this setting.

$$
\begin{aligned}
\mathbb{E}\left[\left(Y_i - m_y\left(\boldsymbol{x}_i\right)\right)^2\right] &= \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right] + \mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)^2\right] \\
&\quad + 2\,\mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right]\right)\left(\mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)\right] \\
&= \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)^2\right] \\
&= \mathbb{V}\mathrm{ar}\left[Y_i|\,\boldsymbol{x}_i\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y_i|\,\boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)^2\right]
\end{aligned}
$$

(Continues...)

Optimal predictor and the CEF (2/2)

Theorem 1

Proof.

(Continued.) Note that the last term in the third line vanishes since:

$$\mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right)\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right)\right] =$$

$$= \mathbb{E}\left[\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - m_y\left(\boldsymbol{x}_i\right)\right) \cdot \underbrace{\mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right) \middle| \boldsymbol{x}_i\right]}_{=0}\right] = 0$$

an observation that exploits the Law of Iterated Expectations. Notice how the result in the third line is minimized if $\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] = m_y\left(\boldsymbol{x}_i\right)$, so long as $\mathbb{V}\mathrm{ar}\left[Y_i \middle| \boldsymbol{x}_i\right] < \infty$. □

Note: with different loss functions, different results would apply.

- For, example, with the absolute loss the optimal predictor is the **conditional median**.
- With a quantile loss, the optimal predictor is the corresponding **conditional quantile** (for the same $p \in (0, 1)$).

Optimal linear predictors

- This result highlights the importance of studying the CEF.
- The exact functional form of the CEF, however, is unlikely to be known by analysts.
- What if analysts restricted their attention only to **optimal linear predictors**, written as $p_y(\boldsymbol{x}_i)$?
- That is, let $m_y(\boldsymbol{x}_i) = p_y(\boldsymbol{x}_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*$ where:

$$\boldsymbol{\beta}^* \in \underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\arg\min}\, \mathbb{E}\left[\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right]$$

  that is, $\boldsymbol{\beta}^*$ is one specific coefficient vector which, amongst all predictors that are **linear** in $\boldsymbol{x}_i$ (there might be many), minimizes the Mean Squared Error.

- The next result is associated with optimal linear predictors.

Optimal linear predictor and the CEF

Theorem 2

Optimal Linear Predictor as best approximation to the CEF.
Consider any optimal linear predictor $\boldsymbol{\beta}^*$. *If* $\mathbb{V}\mathrm{ar}\left[Y_i \middle| \boldsymbol{x}_i\right] < \infty$, *then:*

$$\boldsymbol{\beta}^* \in \underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\arg\min}\, \mathbb{E}\left[\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right].$$

In words, any optimal linear predictor of Y_i *is also an optimal linear predictor of the CEF,* $\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]$, *in the MSE sense.*

Proof.

The demonstration is analogous to that of Theorem 1:

$$\begin{aligned}
\mathbb{E}\left[\left(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] &= \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] + \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \\
&= \mathbb{E}\left[\left(Y_i - \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right] \\
&= \mathbb{V}\mathrm{ar}\left[Y_i \middle| \boldsymbol{x}_i\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2\right]
\end{aligned}$$

where the cross-term in the second line vanishes thanks to the Law of Iterated Expectations. The conclusion then follows easily. □

Implications of linear predictors

- By virtue of this result, if the CEF is unknown an optimal linear predictor is the best choice for approximating it!

- This has had a very deep impact at motivating the use of the linear model, even if just as an "exploratory" tool.

- The First Order Conditions associated with $\boldsymbol{\beta}^*$ are:

$$\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] \boldsymbol{\beta}^* = \mathbb{E}\left[\boldsymbol{x}_i Y_i\right]$$

  and a **unique solution** exists if $\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]$ is nonsingular.

$$\boldsymbol{\beta}^* = \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i Y_i\right]$$

- The **optimal linear predictor** – also (population) **linear projection** of Y_i given $\boldsymbol{x}_i$ – is thus as follows.

$$p_y\left(\boldsymbol{x}_i\right) = \boldsymbol{x}_i^{\mathrm{T}} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i Y_i\right]$$

Example: approximating a quadratic CEF (1/5)

- It is useful to develop an **example**. Suppose that the true CEF is **quadratic**.

$$\mathbb{E}\left[Y_i \middle| X_i\right] = X_i - \frac{1}{10} X_i^2$$

- How is an optimal linear predictor constructed? Since this is a bivariate environment, it would be:

$$p_y\left(X_i\right) = \beta_0^* + \beta_1^* X_i$$

where β_0^* and β_1^* are defined as follows.

$$\left(\beta_0^*, \beta_1^*\right) \in \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\arg\min}\ \mathbb{E}\left[\left(Y_i - \beta_0 - \beta_1 X_i\right)^2\right]$$

- In order to find the optimal linear predictor in this setting, it is necessary to solve for the above problem.

Example: approximating a quadratic CEF (2/5)

- The First Order Conditions are identical to those from the analysis of the bivariate model in Lecture 3!

$$\mathbb{E}\left[Y_i - \beta_0^* - \beta_1^* X_i\right] = 0$$
$$\mathbb{E}\left[X_i \left(Y_i - \beta_0^* - \beta_1^* X_i\right)\right] = 0$$

- Hence, the solutions are also alike.

$$\beta_0^* = \mathbb{E}\left[Y_i\right] - \frac{\mathbb{E}\left[X_i Y_i\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_i\right]}{\mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X_i\right]\right)^2} \cdot \mathbb{E}\left[X_i\right]$$
$$\beta_1^* = \frac{\mathbb{E}\left[X_i Y_i\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_i\right]}{\mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X_i\right]\right)^2}$$

- In general, the exact values for β_0^* and β_1^* depend upon the joint distribution of X_i and Y_i, but in this **particular case** some shortcuts can be taken.

Example: approximating a quadratic CEF (3/5)

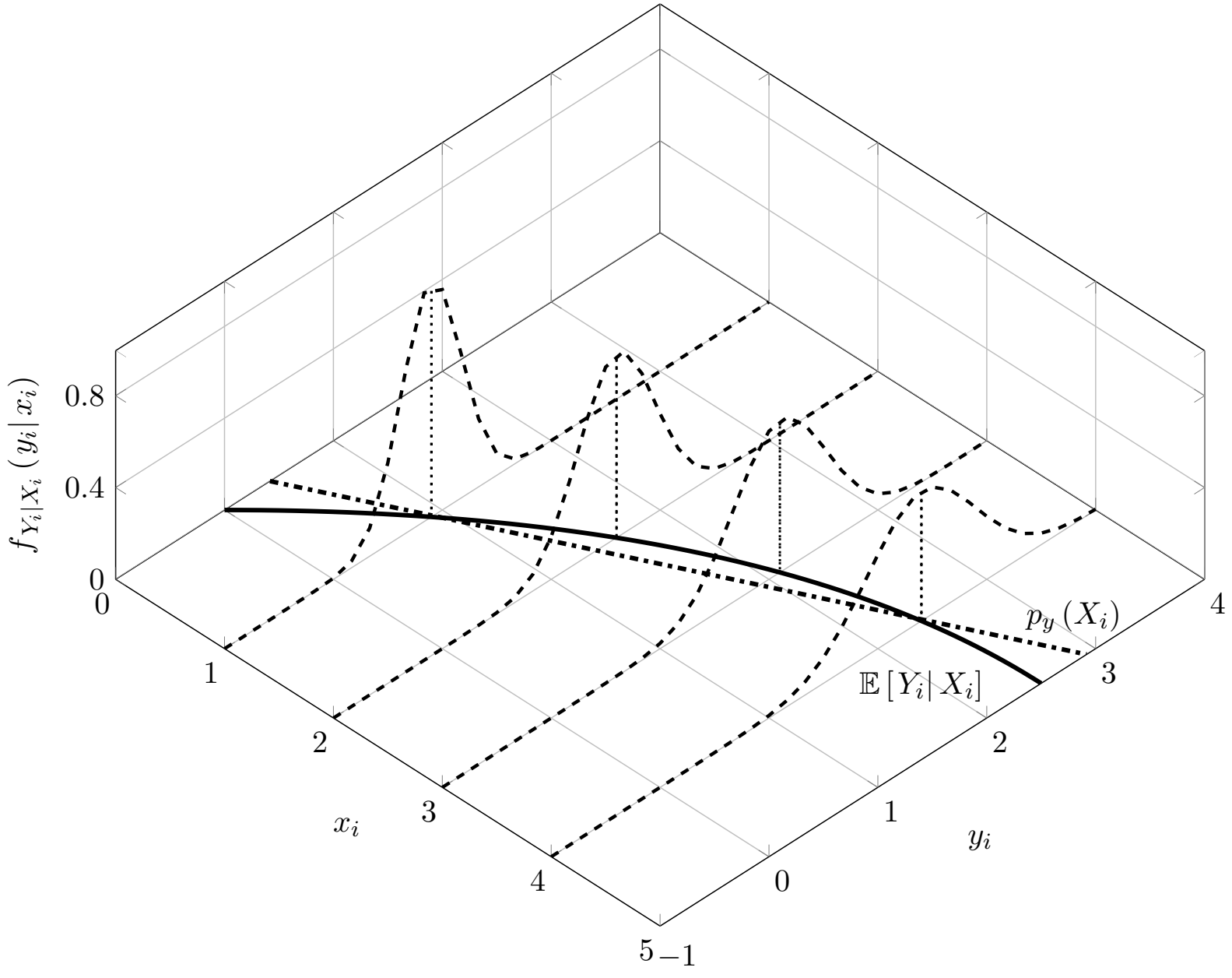
- Observe that under the hypothesis of a quadratic CEF, the moments of the form $\mathbb{E}\left[X_i^r Y_i\right]$ – for any nonnegative integer r – can be obtained easily. Here they are as follows.

$$\begin{aligned}
\mathbb{E}\left[X_i^r Y_i\right] &= \mathbb{E}\left[\mathbb{E}\left[X_i^r Y_i \middle| X_i\right]\right] \\
&= \mathbb{E}\left[X_i^r \cdot \mathbb{E}\left[Y_i \middle| X_i\right]\right] \\
&= \mathbb{E}\left[X_i^r \left(X_i - \frac{1}{10} X_i^2\right)\right] \\
&= \mathbb{E}\left[X_i^{r+1}\right] - \frac{1}{10}\,\mathbb{E}\left[X_i^{r+2}\right]
\end{aligned}$$

- **In this particular example**, β_0^* and β_1^* solely depend on the moments of X_i. If for example's sake it is $X_i \sim \mathcal{U}\left[0, 5\right]$, then $\mathbb{E}\left[X_i^r\right] = 5^r / \left(r + 1\right)$; thus the optimal linear predictor is as follows.

$$p_y\left(X_i\right) = \frac{5}{12} + \frac{1}{2} X_i$$

Example: approximating a quadratic CEF (4/5)



Example: approximating a quadratic CEF (5/5)

- In the figure, the solid line is the CEF, the dash-dotted line is the optimal linear predictor.
- To **help visualize** the **random nature** of the relationship linking Y_i with X_i, some **conditional** p.d.f.s of Y_i given X_i are displayed for $x_i = \{1, 2, 3, 4\}$.
- However, they need not be normal! This is an example.
- Observe how the optimal linear predictor approximates the CEF closely over the entire support of X_i!
- This finding, however, largely relies on the fact that in this working example the distribution of X_i is uniform.
- It is a good exercise to try to replicate this example and/or figure, but using different distributions.

The least squares problem

- In light of the properties of optimal linear predictors, it is natural to ask how to apply them to the data.

- While optimal linear predictors are a **population** objects, the **analogy principle** is useful in this regard.

- Given a sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, define the vector $\mathbf{b}$ as:

$$\mathbf{b} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2$$

  or equivalently, using compact matrix notation, as follows.

$$\mathbf{b} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \frac{1}{N} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right)^{\mathrm{T}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right)$$

- Thus, $\mathbf{b}$ is the vector of coefficients minimizing the average or sum of the squared analogues of prediction errors.

The least squares solution

- The First Order Conditions, also **normal equations**, are:

$$-\frac{2}{N}\sum_{i=1}^{N}\mathbf{x}_i\left(y_i-\mathbf{x}_i^{\mathrm{T}}\mathbf{b}\right)=\mathbf{0}$$

- ...and thus a unique solution $\mathbf{b}$ exists if the $K\times K$ matrix $\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ is **invertible**: it reads as follows.

$$\mathbf{b}=\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i y_i\right)$$

- It is more elegant to write this result via compact notation. The normal equations are:

$$-\frac{2}{N}\mathbf{X}^{\mathrm{T}}\left(\mathbf{y}-\mathbf{X}\mathbf{b}\right)=\mathbf{0}$$

- ...while, if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, the solution is as follows.

$$\mathbf{b}=\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

The least squares solution: discussion

- Obviously, the $1/N$ factor is irrelevant towards the solution.
- This solution is quite reminiscent of a Method of Moments approach – despite it features an optimization problem.
- Indeed, the statistical estimator based on the least squares solution, to be introduced later, can be framed via MM.
- However, the "longer route" – going via the minimization of the empirical, squared prediction errors – helps develop the link with optimal linear predictors, and their properties.
- It is helpful to familiarize with both the transparent vector-based notation, featuring $\mathbf{x}_i$ and y_i, and with the compact, quite convenient matrix notation, featuring $\mathbf{X}$ and $\mathbf{y}$.
- Observe that matrices $\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ are identical!

Fitted Values and residuals

- For every observation i, define its **fitted value** as:

$$\widehat{y}_i \equiv \mathbf{x}_i^{\mathrm{T}} \mathbf{b}$$

- . . . and its **residual** (*empirical prediction error*) as follows.

$$\begin{aligned} e_i &\equiv y_i - \widehat{y}_i \\ &= y_i - \mathbf{x}_i^{\mathrm{T}} \mathbf{b} \end{aligned}$$

- It is convenient to stack these objects vertically in order to handle them in compact notation.

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_N \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

Projection matrix

- The vector of fitted values can be expressed compactly as:

$$\begin{aligned}\widehat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\ &= \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &= \mathbf{P}_{\mathbf{X}}\mathbf{y}\end{aligned}$$

  where:

$$\mathbf{P}_{\mathbf{X}} \equiv \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$$

  is called the **projection matrix**.

- Pre-multiplying $\mathbf{y}$ by the projection matrix $\mathbf{P}_{\mathbf{X}}$ results in the vector of fitted values $\widehat{\mathbf{y}}$.

Residual-maker matrix

- Furthermore:

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \widehat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\mathbf{b} \\ &= (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\,\mathbf{y} \\ &= \mathbf{M}_{\mathbf{X}}\mathbf{y}\end{aligned}$$

  where:

$$\mathbf{M}_{\mathbf{X}} \equiv \mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$$

  is the so-called **residual maker** matrix.

- Pre-multiplying $\mathbf{y}$ by the residual maker matrix $\mathbf{M}_{\mathbf{X}}$ clearly results in the vector of residuals $\mathbf{e}$.

Properties of $\mathbf{P_X}$ and $\mathbf{M_X}$ (1/4)

- The projection and residual maker matrices have important properties.

- They are both **symmetric**:

$$\mathbf{P_X} = \mathbf{P_X^T}$$
$$\mathbf{M_X} = \mathbf{M_X^T}$$

- . . . they are both **idempotent**:

$$\mathbf{P_X P_X} = \mathbf{P_X}$$
$$\mathbf{M_X M_X} = \mathbf{M_X}$$

- . . . and they are **orthogonal** to one another.

$$\mathbf{P_X M_X} = \mathbf{M_X P_X} = \mathbf{0}$$

Properties of $\mathbf{P_X}$ and $\mathbf{M_X}$ (2/4)

- In addition, it is easy to see that:

$$\mathbf{P_X X} = \mathbf{X}$$
$$\mathbf{M_X X} = \mathbf{0}$$

  having a straightforward interpretation: if one projects the columns of $\mathbf{X}$ onto themselves, the projection is identical to $\mathbf{X}$ and the residuals, consequently, are zero.

- Finally, observe that:

$$\begin{aligned}\mathbf{y} &= (\mathbf{I} + \mathbf{P_X} - \mathbf{P_X})\,\mathbf{y} \\ &= \mathbf{P_X y} + \mathbf{M_X y} \\ &= \widehat{\mathbf{y}} + \mathbf{e}\end{aligned}$$

  (Continues. . .)

Properties of $\mathbf{P_X}$ and $\mathbf{M_X}$ (3/4)

- **(Continued.)** ... and that:

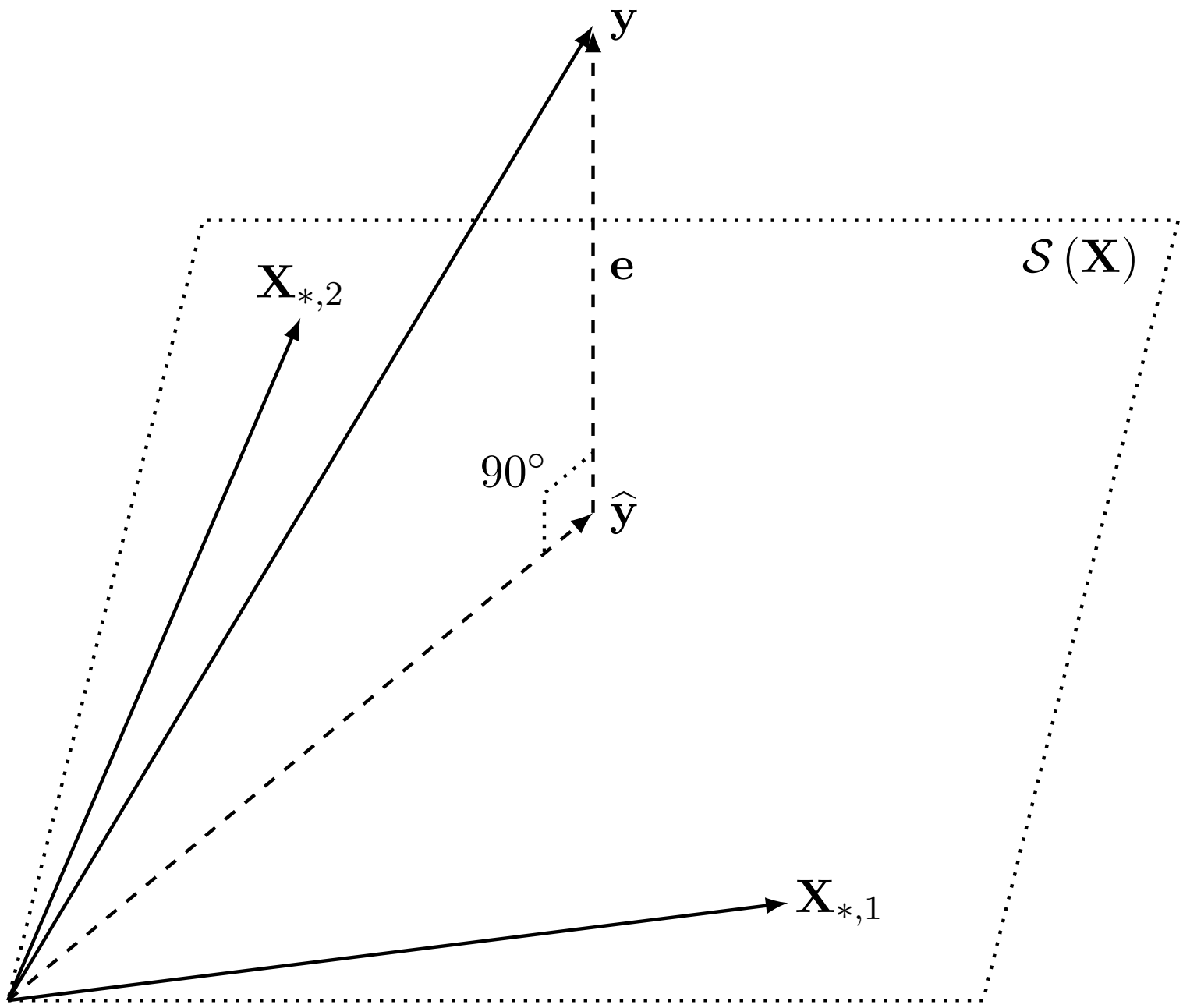
$$\begin{aligned}\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{e} &= \mathbf{y}^{\mathrm{T}}\mathbf{P_X}\mathbf{M_X}\mathbf{y} \\ = \mathbf{e}^{\mathrm{T}}\widehat{\mathbf{y}} &= \mathbf{y}^{\mathrm{T}}\mathbf{M_X}\mathbf{P_X}\mathbf{y} \\ &= 0\end{aligned}$$

  indicating that the decomposition of the vector $\mathbf{y}$ between the fitted values $\widehat{\mathbf{y}}$ and the residuals $\mathbf{e}$ is such that the two components at hand are **orthogonal** to one another.

- This relates to the important **geometric interpretation** of $\mathbf{b}$: via the linear combination $\widehat{\mathbf{y}} = \mathbf{P_X}\mathbf{y} = \mathbf{X}\mathbf{b}$, the least squares solution provides the **geometrical projection** of $\mathbf{y}$ onto the column space of $\mathbf{X}$ (call this $\mathcal{S}(\mathbf{X})$).

- This is represented in the next figure for $K = 2$.

Properties of $\mathbf{P_X}$ and $\mathbf{M_X}$ (4/4)



Partitioned regression

- The result to be discussed next, the **Frisch-Waugh-Lovell Theorem**, is central in the analysis of the linear model.

- This concerns **partition regression**: a linear model where two sets of independent variables are distinguished:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

  where $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$ and $\boldsymbol{\beta}^{\mathrm{T}} = \begin{bmatrix}\boldsymbol{\beta}_1^{\mathrm{T}} & \boldsymbol{\beta}_2^{\mathrm{T}}\end{bmatrix}$.

- The two sub-vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have respectively length K_1 and K_2 (with $K_1 + K_2 = K$).

- The K normal equations are rephrased here as:

$$\begin{bmatrix}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2 \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1 & \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{b}_1 \\ \mathbf{b}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} \\ \mathbf{X}_2^{\mathrm{T}}\mathbf{y}\end{bmatrix}$$

  where $\mathbf{b}^{\mathrm{T}} = \begin{bmatrix}\mathbf{b}_1^{\mathrm{T}} & \mathbf{b}_2^{\mathrm{T}}\end{bmatrix}$.

Frisch-Waugh-Lovell Theorem (1/2)

Theorem 3

Frisch-Waugh-Lovell Theorem. *The least squares solution for* $\mathbf{b}_2$ *in the partitioned model can be written as:*

$$\mathbf{b}_2 = \left(\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2^*\right)^{-1}\mathbf{X}_2^{*\mathrm{T}}\mathbf{y}$$

where:

$$\mathbf{X}_2^* \equiv \mathbf{M}_{\mathbf{X}1}\mathbf{X}_2$$

and $\mathbf{M}_{\mathbf{X}1}$ *is the residual maker matrix of* $\mathbf{X}_1$*.*

$$\mathbf{M}_{\mathbf{X}1} \equiv \mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}$$

Furthermore, a symmetrical result is obtained for $\mathbf{b}_1$*.*

Proof.

(Continues...)

Frisch-Waugh-Lovell Theorem (2/2)

Theorem 3

Proof.

(Continued.) By the algebra of partitioned matrices, one can write $\mathbf{b}_1$ as a function of $\mathbf{b}_2$ as:

$$\mathbf{b}_1 = \left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} - \left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\mathbf{b}_2$$

plugging the above in the lower block of K_2 normal equation gives:

$$\mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{y} - \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\mathbf{b}_2 + \mathbf{X}_2^{\mathrm{T}}\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2^{\mathrm{T}}\mathbf{y}$$

with solution:

$$\begin{aligned}\mathbf{b}_2 = &\left[\mathbf{X}_2^{\mathrm{T}}\left(\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\right)\mathbf{X}_2\right]^{-1} \times \\ &\times \left[\mathbf{X}_2\left(\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1\right)^{-1}\mathbf{X}_1^{\mathrm{T}}\right)\mathbf{y}\right] \\ = &\left(\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}\end{aligned}$$

as per the statement ($\mathbf{M}_{\mathbf{X}1}$ is symmetric and idempotent). The result for $\mathbf{b}_1$ is symmetrical. □

Frisch-Waugh-Lovell Theorem: discussion

The theorem says that any component ($\mathbf{b}_2$) of the least squares solution is *algebraically equivalent* to a least squares solution that is alternatively obtained by:

1. projecting the explanatory variables of interest ($\mathbf{X}_2$) to the other explanatory variables ($\mathbf{X}_1$);
2. calculating the corresponding residuals ($\mathbf{X}_2^* = \mathbf{M}_{\mathbf{X}1}\mathbf{X}_2$);
3. and, finally, projecting the dependent variable $\mathbf{y}$ onto these residuals $\mathbf{X}_2^*$.

Therefore a least squares coefficient b_k can be interpreted as the overall "contribution" of X_{ki} to Y_i, *after the contributions of the other* $K-1$ *explanatory variables to* Y_i *has been netted out* – or, using more technical terminology – *partialled out*.

This interpretation corresponds with the typical *ceteris paribus* thought experiments in science.

Partial correlation

- An illustrative case of the Frisch-Waugh-Lovell Theorem is the one where $K_1 = K - 1$ and $K_2 = 1$. Let here $\mathbf{X}_2 = \mathbf{s}$.

- The K-th Least Squares coefficient would be as follows.

$$b_K = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}}$$

- This quantity relates to a statistical object called **partial correlation coefficient** ρ^*_{YS} between Y_i and $X_{Ki} = S_i$.

$$\rho^*_{YS} = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}{\sqrt{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}}\sqrt{\mathbf{y}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}}$$

- This is the sample counterpart of the **population partial correlation** between Y_i and $X_{Ki} = S_i$, given $\boldsymbol{x}_i$.

$$\mathbb{C}\mathrm{orr}\left[Y_i, S_i \middle| \boldsymbol{x}_1\right] = \frac{\mathbb{C}\mathrm{ov}\left[Y_i, S_i \middle| \boldsymbol{x}_1\right]}{\sqrt{\mathbb{V}\mathrm{ar}\left[Y_i \middle| \boldsymbol{x}_1\right]}\sqrt{\mathbb{V}\mathrm{ar}\left[S_i \middle| \boldsymbol{x}_1\right]}}$$

Demeaned Models

- Another oft-invoked application of the Frisch-Waugh-Lovell Theorem is **demeaning**: the subtraction of sample **means** from both the explanatory and dependent variables.

- Suppose that $\mathbf{X}_1 = \boldsymbol{\iota}$ is the constant term of the model (an N-sized vector of ones). Then:

$$\mathbf{D} \equiv \mathbf{M}_{\mathbf{X}1} = \mathbf{I} - \boldsymbol{\iota}\left(\boldsymbol{\iota}^{\mathrm{T}}\boldsymbol{\iota}\right)^{-1}\boldsymbol{\iota}^{\mathrm{T}} = \mathbf{I} - \frac{1}{N}\boldsymbol{\iota}\boldsymbol{\iota}^{\mathrm{T}}$$

  and for any vector $\mathbf{a}$ of length N and given $\bar{a} \equiv \frac{1}{N}\sum_{i=1}^{N} a_i$, it is as follows.

$$\mathbf{D}\mathbf{a} = \mathbf{a} - \bar{a}\boldsymbol{\iota}$$

- Hence $\mathbf{b}_2$ can be obtained as the solution of a least squares problem applied to the **demeaned model**:

$$y_i - \bar{y} = \beta_1\left(x_{i1} - \bar{x}_1\right) + \cdots + \beta_{(K-1)}\left(x_{i(K-1)} - \bar{x}_{(K-1)}\right) + \varepsilon_i$$

  where $\bar{y} \equiv \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{x}_k \equiv \frac{1}{N}\sum_{i=1}^{N} x_{ik}$ for all k.

Implications of constant terms

If a constant term β_0 is included in the model, the least squares solution has the following notable properties.

1. All the residuals sum up to zero: $\sum_{i=1}^{N} e_i = 0$.

2. Hence, the mean of the fitted values $\widehat{y}_i$ coincides with that of dependent variable y_i.

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^{\mathrm{T}}\mathbf{b} = \frac{1}{N}\sum_{i=1}^{N} \widehat{y}_i = \bar{y}$$

3. From this, it also follows that the point $(\bar{y}, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)$ lies on the $p(\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\mathbf{b}$ hyperplane.

Together, these properties allow to express a measure about the **goodness of fit**: how well the least squares solution "explains" Y_i via $(X_{i1}, \ldots, X_{iK})$ through the linear combination $\widehat{y}_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{b}$.

Coefficient of determination

This measure is called **coefficient of determination** R^2 and is defined as:

$$\mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum_{i=1}^{N} \left(\widehat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2} \in [0, 1]$$

where:

1. the numerator, called **Explained Sum of Squares** (ESS), relates to the variance of the fitted values:

$$\mathrm{ESS} \equiv \sum_{i=1}^{N} \left(\widehat{y}_i - \bar{y}\right)^2 = \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{b}$$

2. whereas the denominator, named **Total Sum of Squares** (TSS), corresponds with the grand *empirical* variance of Y_i.

$$\mathrm{TSS} \equiv \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2 = \mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}$$

Residual sum of squares

The difference between the TSS and ESS equals the sum of the squared residuals and, as a consequence, is called the **Residual Sum of Squares** (RSS).

$$\text{RSS} \equiv \text{TSS} - \text{ESS} = \sum_{i=1}^{N} e_i^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e}$$

To see why, note that if the model features a constant term it is $\mathbf{De} = \mathbf{e}$, and recall that $\widehat{\mathbf{y}} = \mathbf{Xb}$ and $\mathbf{e}$ are orthogonal. So:

$$\underbrace{\mathbf{y}^{\mathrm{T}}\mathbf{Dy}}_{=\text{TSS}} = (\mathbf{Xb}+\mathbf{e})^{\mathrm{T}}\,\mathbf{D}\,(\mathbf{Xb}+\mathbf{e}) = \underbrace{\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{DXb}}_{=\text{ESS}} + \underbrace{\mathbf{e}^{\mathrm{T}}\mathbf{e}}_{=\text{RSS}}$$

and thus the coefficient of determination R^2 can also be written as follows.

$$\mathrm{R}^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \in [0,1]$$

Coefficient of determination and correlation

Intuitively, the R^2 coefficient is close to 1 if the model 'explains' most of the variation in Y_i; it is close to 0 if only a small portion of it is 'explained.' To better appreciate this, observe that:

$$\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{b} = \widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\widehat{\mathbf{y}} = \widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\left(\mathbf{y}+\mathbf{e}\right) = \widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\mathbf{y}$$

and therefore:

$$\mathrm{R}^2 = \frac{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\widehat{\mathbf{y}}}{\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}} = \frac{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\mathbf{y}}{\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}} \cdot \underbrace{\frac{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\mathbf{y}}{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\widehat{\mathbf{y}}}}_{=1} = \frac{\widehat{\mathbf{y}}^{\mathrm{T}}\mathbf{D}\mathbf{y}}{\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}} =$$

$$= \frac{\left[\sum_{i=1}^{N}\left(y_i-\bar{y}\right)\left(\widehat{y}_i-\bar{y}\right)\right]^2}{\left[\sum_{i=1}^{N}\left(\widehat{y}_i-\bar{y}\right)^2\right]\left[\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2\right]}$$

thus, the R^2 coefficient is equal to the square of the correlation coefficient between y_i and the fitted values $\widehat{y}_i$ (hence its name).

The "adjusted" R^2 coefficient

- One should not mistake the R^2 coefficient for a measure of the overall "quality" of the projection.

- Every empirical setting is different (with different variances of Y_i, values of $\mathbf{b}$), and leads to different measures of R^2!

- In addition, the coefficient increases "mechanically" simply by adding more explanatory variables.

- Because of this, the so-called **"adjusted"** R^2 coefficient is sometimes used: it accounts for varying K.

$$\begin{aligned}\overline{\mathrm{R}}^2 &= 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \frac{N-1}{N-K} \\ &= 1 - \left(1 - \mathrm{R}^2\right) \frac{N-1}{N-K}\end{aligned}$$

- Notice: $\overline{\mathrm{R}}^2$ may even turn negative for low values of R^2!

Fitting a linear consumption function (1/2)

- To showcase some properties of least squares, as well as the R^2 coefficient, let us return to the earlier examples.

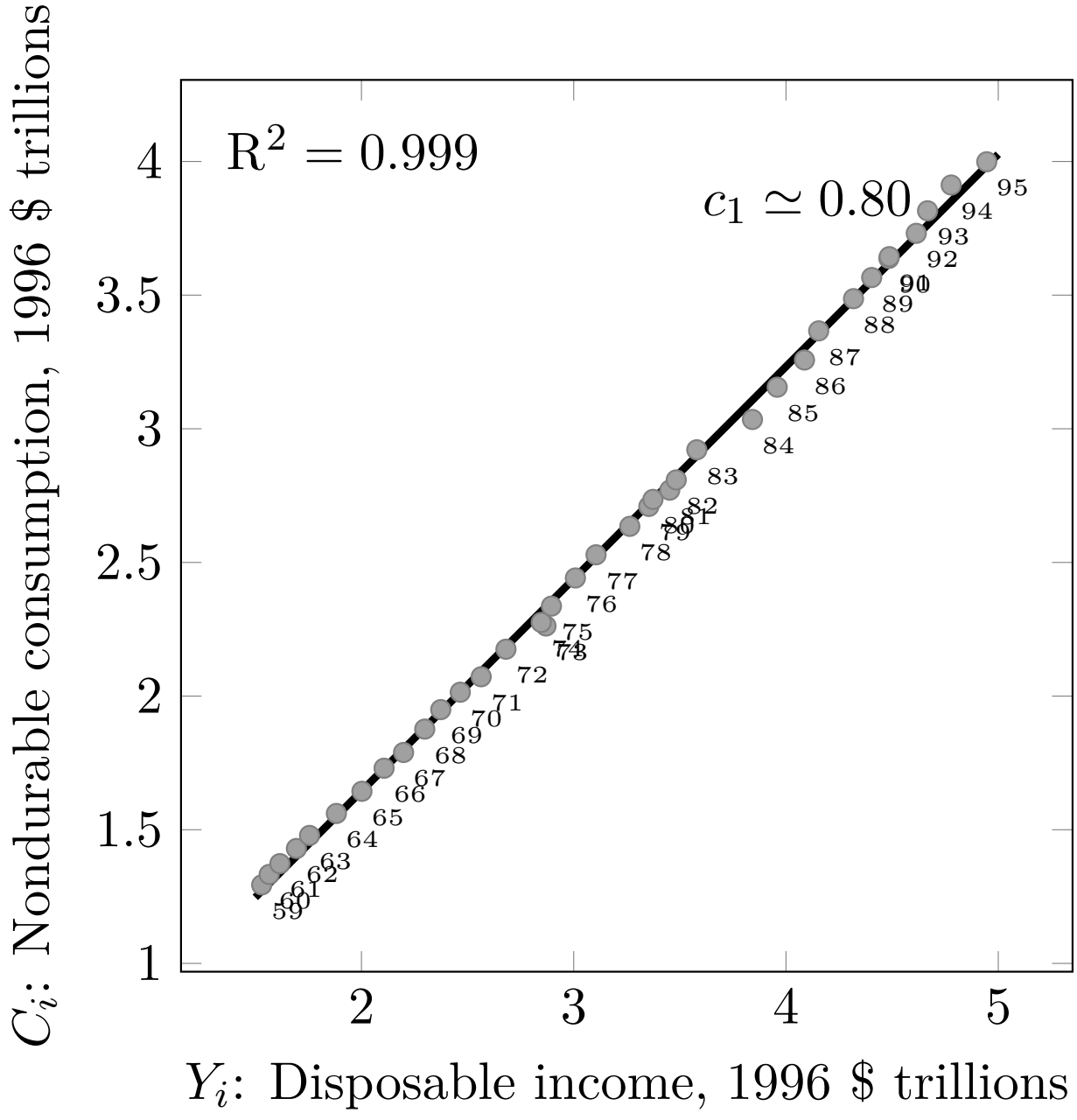
- Consider the initial example on the Keynesian consumption function: there, a linear fit returns a value for the marginal propensity of consumption equal to $c_1 \simeq 0.80$.

- In this specific, case R^2 is virtually close to 1!

- See the next figure for graphical evidence.

- Yet, as discussed earlier macroeconomists nowadays discard structural interpretations of such a model; as a consequence this particular R^2 is not especially informative.

Fitting a linear consumption function (2/2)



Fitting the Mincer equation (1/3)

- In trying to fit a **simplified** version of the Mincer equation **without** the squared term for experience X_i, one obtains a value $b \simeq 0.064$ for the coefficent of education, as well as an R^2 coefficient equal to 0.076. See the next figure.

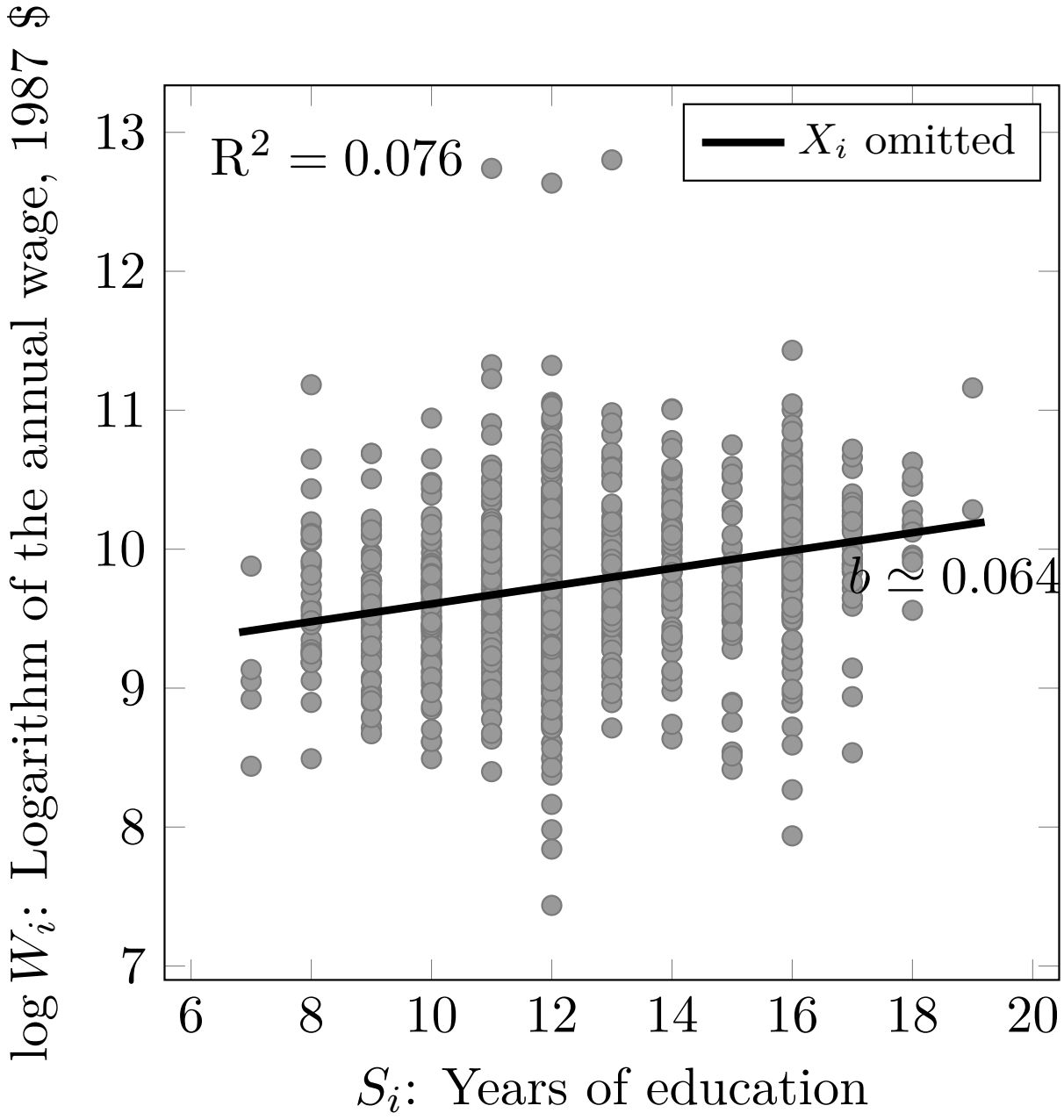
- Education explains quite a substantial part of the log-wage variance: about 7.6%. This is not bad!

- The Frisch-Waugh-Lovell Theorem helps visualize the effect of including experience: the ensuing figure shows a linear fit of **log-wages** on the **residuals** obtained from **projecting education** on a **squared term** for **experience**.

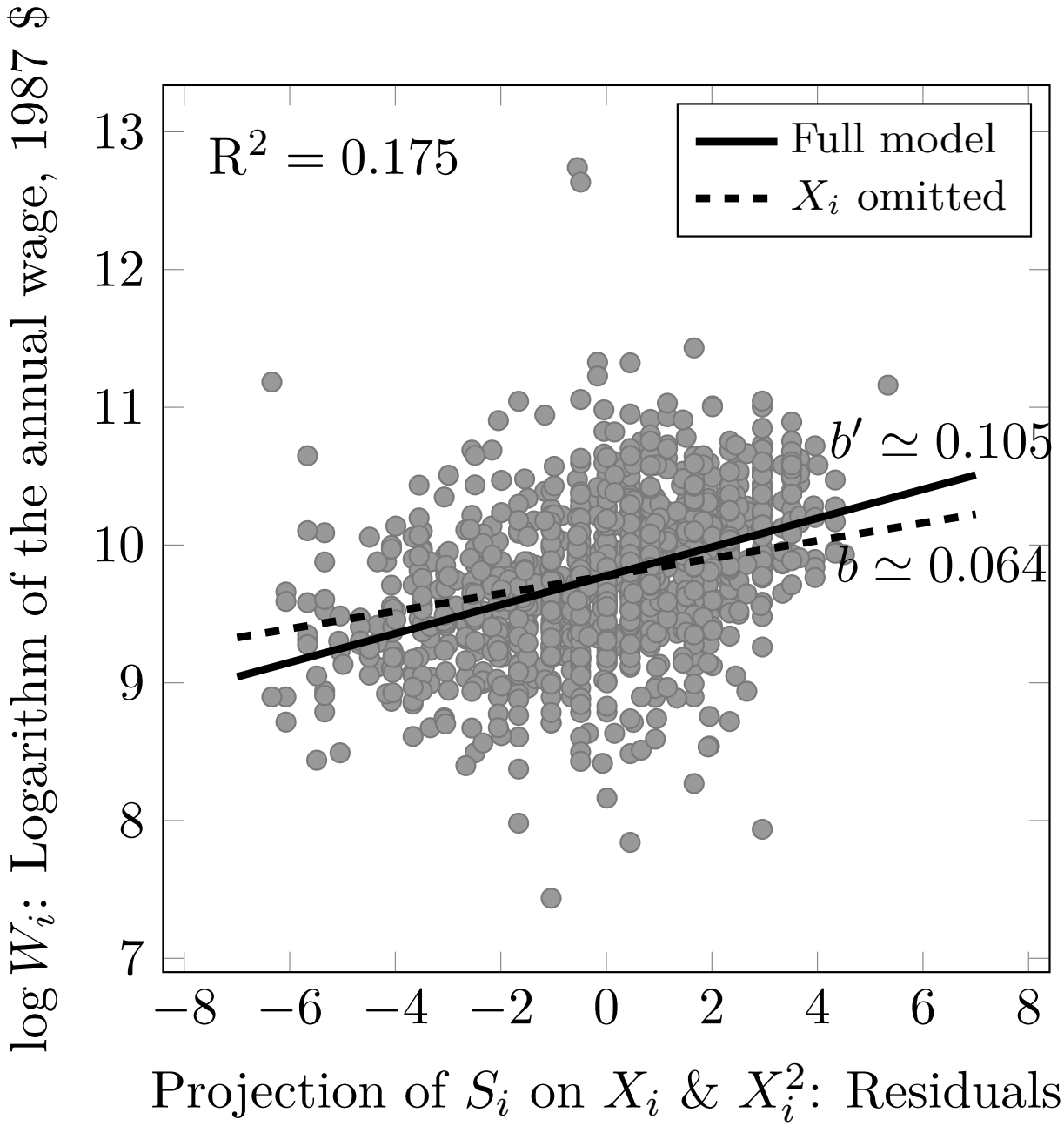
- As one would expect, R^2 increases, to about 0.175. But also the education coefficient does! This is now up to $b' \simeq 0.105$.

- The education coefficient was dragged down by the negative (mechanical) correlation between education and experience.

Fitting the Mincer equation (2/3)



Fitting the Mincer equation (3/3)



Linear regression

- It is finally time to endow the workhorse linear model with statistical assumptions, and an associated **estimator**.

- The resulting **Ordinary Least Squares** (OLS) estimator inherits all the properties of the least squares solution.

- The leading **distributional assumption** discussed here is **linearity of the CEF** of Y_i given $\boldsymbol{x}_i$:

$$\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0$$

  where $\boldsymbol{\beta}_0$ denotes the "true" vector of parameters.

- A linear model enriched with such a hypothesis on the joint distribution of $(\boldsymbol{x}_i, Y_i)$ is called a **linear regression** model.

- In this environment, $\boldsymbol{x}_i$ are called the **regressors**, and Y_i is called the **regressand**.

Implications of CEF linearity

An implication of CEF linearity is that, so long as $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, the expectation of the linear model's error term ε_i, **conditional** on the explanatory variables $\boldsymbol{x}_i$, is zero:

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] &= \mathbb{E}\left[Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \middle| \boldsymbol{x}_i\right] \\
&= \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \\
&= 0
\end{aligned}$$

which by the Law of Iterated Expectations implies the following.

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{x}_i\varepsilon_i\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}\left[\boldsymbol{x}_i\varepsilon_i \middle| \boldsymbol{x}_i\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{x}_i \cdot \mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right]\right] \\
&= \mathbf{0}
\end{aligned}$$

The opposite however is not true: specifically, $\mathbb{E}\left[\boldsymbol{x}_i\varepsilon_i\right] = \mathbf{0}$ does **not** imply $\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] = 0$.

Least squares and linear predictors

By the standard properties of probability limits it can be shown that (note the random sequence notation with the subscript N):

$$\mathbf{b}_N = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i \xrightarrow{p} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i Y_i \right] = \boldsymbol{\beta}^*$$

so long as matrices $\frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$ and $\mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \right]$ have full rank.

This means that:

$$\mathbf{x}_i^{\mathrm{T}} \mathbf{b}_N \xrightarrow{p} \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* = p_y \left(\boldsymbol{x}_i = \mathbf{x}_i \right)$$

that is, the Least Squares projection converges in probability to the **optimal linear predictor**.

This should not be too surprising: such a result generally holds for sample analogs of population moments.

Least squares, linear CEF and linear predictors

What is relevant here is that under hypothesis of a linear CEF, the optimal linear predictor **coincides with the CEF** for any given realization $\boldsymbol{x}_i = \mathbf{x}_i$:

$$
\begin{aligned}
p_y\left(\boldsymbol{x}_i = \mathbf{x}_i\right) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* &= \mathbf{x}_i^{\mathrm{T}} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i Y_i\right] \\
&= \mathbf{x}_i^{\mathrm{T}} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i \, \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right]\right] \\
&= \mathbf{x}_i^{\mathrm{T}} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right] \boldsymbol{\beta}_0 \\
&= \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \\
&= \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i = \mathbf{x}_i\right]
\end{aligned}
$$

where the second line exploits the Law of Iterated Expectations.

In light of Theorems 1 and 2 this result should not really be too surprising.

Ordinary Least Squares

Combining all these observations, it is clear that when the CEF is linear the least squares solution $\mathbf{b}$ converges in probability to the "true" CEF parameters.

$$\mathbf{b}_N \xrightarrow{p} \boldsymbol{\beta}_0$$

If seen as an estimator, the least squares solution $\mathbf{b}$ is called the **Ordinary Least Squares** (OLS) estimator – where "ordinary" sets it apart from its various extensions.

Note that if the CEF is linear, but only if some 'other' variables not originally included in $\boldsymbol{x}_i$ are added to the set of explanatory variables, the estimator is still inconsistent.

Omitted variable bias (1/2)

To analyze the last issue in more detail, suppose that the 'true' CEF is:

$$\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i, \boldsymbol{s}_i\right] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \boldsymbol{s}_i^{\mathrm{T}} \boldsymbol{\delta}_0$$

where $\boldsymbol{s}_i$ is another set of explanatory variables not included in the analysis alongside $\boldsymbol{x}_i$. The associated parameter set is $\boldsymbol{\delta}_0$.

Thus, the probability limit of least squares here is as follows.

$$\begin{aligned}
\mathbf{b}_N &\xrightarrow{p} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i Y_i\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i \, \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i, \boldsymbol{s}_i\right]\right] \\
&= \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \boldsymbol{x}_i \boldsymbol{s}_i^{\mathrm{T}} \boldsymbol{\delta}_0\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{s}_i^{\mathrm{T}}\right] \boldsymbol{\delta}_0
\end{aligned}$$

Note the application of the Law of Iterated Expectations.

Omitted variable bias (2/2)

The intuition is developed best when $\boldsymbol{s}_i = S_i$ is a single variable with associated parameter δ_0. Then:

$$\mathbf{b}_N \xrightarrow{p} \boldsymbol{\beta}_0 + \delta_0 \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i S_i\right]$$

is the (in)famous formula of the **omitted variable bias**.

The formula suggests that if the variable S_i is **omitted**, the bias is asymptotically nonzero unless **one** of two conditions realizes:

1. the coefficient of S_i in the CEF is zero: $\delta_0 = 0$;
2. S_i bears no correlation with $\boldsymbol{x}_i$: that is, $\mathbb{E}\left[\boldsymbol{x}_i S_i\right] = 0$.

The Mincer equation with omitted experience X_i is an excellent example of a **negative** omitted variable bias.

Why using Ordinary Least Squares?

This Lecture has already developed two relevant motivations for using the OLS estimator in empirical analysis.

1. If the CEF of Y_i given $\boldsymbol{x}_i$ is linear, then the OLS estimator consistently estimates its parameters.

2. If the CEF is not linear, the OLS estimator asymptotically converges to the optimal linear predictor – which is shown to be the best approximation to the CEF (which in turn is the optimal predictor under quadratic loss).

In what follows, three additional motivations for OLS are given.

1. It can handle selected **non-linear** model.

2. It nicely handles **categorical** (discrete) data.

3. It also approximates the **derivative** of (non-linear) CEFs.

Least squares for linearized models

- Least squares and OLS can often be adapted to structural **non-linear** models.
- The trick is to identify some appropriate **transformation** of the model that makes it **linear in the parameters**.
- Leading examples are models that follow from **logarithmic** transformations: the **log-lin** and **log-log** models.
- In such cases, care should be taken as to how to **interpret** the parameters!
- Suitable transformations might not exist. In such cases, the best feasible option may be a **Non-Linear Least Squares** (NLLS) estimation approach.
- Some examples follow suit.

Example: a log-lin model

- Consider a Mincer equation with $\varepsilon_i \equiv \alpha_i + \epsilon_i$.

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + \varepsilon_i$$

- It follows from taking the **logarithm** on **both sides** of the motivating structural equation, yielding a **log-lin** model.

- Here, the coefficient for education: β_3 is a **semi-elasticity**.

- That is, it indicates the **relative** increase of the dependent variable (here: wages W_i) that on average follows a **unitary** increase in the independent variable (here: education S_i).

- The previous estimate $b' = 0.105$ indicates that, on average, wages increase by 10.5% for every extra year of education.

Example: a log-log model

- A **Cobb-Douglas production function** is a very typical component of many economic models:

$$Y_i = A_i K_i^{\beta_K} L_i^{\beta_L}$$

  it relates **output** Y_i, **capital** K_i, **labor** L_i and total factor **productivity** A_i.

- Taking **logarithms** on both sides delivers a **log-log** model:

$$\log Y_i = \alpha + \beta_K \log K_i + \beta_L \log L_i + \omega_i$$

  for $\log A_i = \alpha + \omega_i$; here ω_i is an unobserved **productivity shock** with $\mathbb{E}[\omega_i] = 0$ (a hard-to-defend assumption).

- Here, the two coefficients β_K and β_L are **elasticities**.

- They denote the **relative** change of the dependent variable (here: output Y_i) that on average follows a **relative** change in the independent variables (here: capital K_i or labor L_i).

Example: a non-linearizable model

- Consider the following 'augmented' production function of the Cobb-Douglas kind:

$$\log Y_i = \alpha + \beta_K \log K_i + \beta_L \log L_i + \delta \exp(-\lambda D_i) U_i + \omega_i$$

  where U_i is the size of some **local university**, while D_i is its **distance** from firm i.

- In this model, parameter δ represents the semi-elasticity of firm productivity with respect to the university's size...

- ...but weighted by the **distance decay** factor $\exp(-\lambda D_i)$, which is parametrized by λ.

- This model cannot be estimated via OLS; one shall proceed via Non-Linear Least Squares.

A simple dummy variable model (1/4)

A main advantage of OLS is that it is well suited at estimating differences between **groups** or **categories** that it is possible to separately distinguish in the data. This is due to the connection between OLS and linear CEFs.

The technical devices used for this sake are **dummy variables**: binary variables that can only take – for example – value 0 or 1 and that identify different groups or categories.

To illustrate how dummy variables operate, it is useful to study the following simple bivariate model:

$$Y_i = \pi_0 + \pi_1 D_i + \eta_i$$

where Y_i is some outcome of interest, D_i is a **dummy variable** that identifies a **group** of interest (for example: females, blacks, foreigners, young people) and η_i is an error term.

A simple dummy variable model (2/4)

In such a model, the vector of Least Squares coefficients (p_0, p_1) which is obtained through a sample $\{y_i, d_i\}_{i=1}^{N}$ is calculated as:

$$\begin{aligned}
\begin{bmatrix} p_0 \\ p_1 \end{bmatrix} &= \begin{bmatrix} N & N_D \\ N_D & N_D \end{bmatrix}^{-1} \begin{bmatrix} N\bar{y} \\ N_D\bar{y}_D \end{bmatrix} \\
&= \frac{1}{N - N_D} \begin{bmatrix} N\bar{y} - N_D\bar{y}_D \\ -N\bar{y} + N\bar{y}_D \end{bmatrix} \\
&= \frac{1}{N - N_D} \begin{bmatrix} N\bar{y} - N_D\bar{y}_D \\ (N - N_D)\,\bar{y}_D - N\bar{y} + N_D\bar{y}_D \end{bmatrix} \\
&= \begin{bmatrix} \bar{y}_{\setminus D} \\ \bar{y}_D - \bar{y}_{\setminus D} \end{bmatrix}
\end{aligned}$$

where N_D is the total number of observations with $d_i = 1$. The analysis of the terms $\bar{y}$, $\bar{y}_D$ and $\bar{y}_{\setminus D}$ follows. **(Continues...)**

A simple dummy variable model (3/4)

(Continued.) In the expression:

$$\begin{bmatrix} p_0 \\ p_1 \end{bmatrix} = \begin{bmatrix} \bar{y}_{\setminus D} \\ \bar{y}_D - \bar{y}_{\setminus D} \end{bmatrix}$$

the terms $\bar{y}$, $\bar{y}_D$ and $\bar{y}_{\setminus D}$ are as follows:

- $\bar{y}$ is the grand average of y_i in the sample;
- $\bar{y}_D \equiv N_D^{-1} \sum_{i=1}^{N} y_i d_i$ is the average of y_i in the "dummy" group with $d_i = 1$;
- while the term

$$\bar{y}_{\setminus D} \equiv \frac{1}{N - N_D} \sum_{i=1}^{N} y_i \left(1 - d_i\right) = \frac{N\bar{y} - N_D \bar{y}_D}{N - N_D}$$

  is instead the average of y_i in the complementary group with $d_i = 0$.

A simple dummy variable model (4/4)

By the Laws of Large Numbers, and by the properties of linear projections, it follows that:

$$\begin{bmatrix} p_0 \\ p_1 \end{bmatrix} = \begin{bmatrix} \bar{y}_{\setminus D} \\ \bar{y}_D - \bar{y}_{\setminus D} \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \mathbb{E}\left[Y_i \middle| D_i = 0\right] \\ \mathbb{E}\left[Y_i \middle| D_i = 1\right] - \mathbb{E}\left[Y_i \middle| D_i = 0\right] \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}$$

endowing the two regression parameters π_0 and π_1 with a clear interpretation in terms of group-specific population averages.

A similar result can be obtained through the following, simpler alternative model, with two dummy variables and no constant:

$$Y_i = \pi_1' D_i + \pi_2' \left(1 - D_i\right) + \eta_i'$$

and it is even easier to show that:

$$\begin{bmatrix} p_1' \\ p_2' \end{bmatrix} = \begin{bmatrix} \bar{y}_{\setminus D} \\ \bar{y}_D \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \mathbb{E}\left[Y_i \middle| D_i = 1\right] \\ \mathbb{E}\left[Y_i \middle| D_i = 0\right] \end{bmatrix} = \begin{bmatrix} \pi_1' \\ \pi_2' \end{bmatrix}$$

with an even more straightforward interpretation.

The dummy variable trap

Observe that it is **impossible** to run a model like the previous one with the addition of a constant term:

$$Y_i = \pi_0'' + \pi_1'' D_i + \pi_2'' (1 - D_i) + \eta_i'' \quad (?)$$

because no unique vector of Least Squares coefficient is possibly computed.

The reason is that here, the columns of the regressors matrix $\mathbf{X}$ are by construction **linearly dependent**, implying that matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ – here, of size 3×3 – is singular.

This problem, which can be generalized to higher dimensions, is popularly known as the **dummy variable trap**.

Care should be taken in designing a model that features dummy variables!

Analysis of variance (1/2)

- In these simple examples, the use of OLS for the analysis of group differences requires **no statistical assumptions**!

- This fact can be **generalized**: suppose that a population is partitioned between K non-overlapping **groups**, which may also represent the intersections of more aggregate divisions.

- For example, a partition intersecting **two** binary groups like "gender" and "age" is the set with size $K = 4$ and elements $\{male\&young, female\&young, male\&old, female\&old\}$.

- If a *unique* dummy variable X_{ki} is associated to each group for $k = 1, \ldots, K$ (the dummy variable trap is thus avoided) for every dependent variable Y_i the following holds.

$$\mathbb{E}\left[Y_i \middle| X_{1i}, \ldots, X_{Ki}\right] = \mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i\right] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\pi}_0$$

Analysis of variance (2/2)

- The K parameters $\boldsymbol{\pi}_0$ from the expression $\mathbb{E}[Y_i|\,\boldsymbol{x}_i] = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\pi}_0$ are interpreted as the K **group-specific means** of Y_i.

- Such a model is called a **fully saturated regression** (that is, saturated with dummy variables – one per group).

- The typical use of models like this is in statistical exercises going under the name **"Analysis of Variance"** (ANOVA) – their objective is to examine group differences in selected populations.

- A fully saturated regression model can be useful to provide selected least squares parameters with an interpretation as the "average CEF derivative" – more on this soon.

- Moreover, it is used as a building block for regression-based **causal inference** (see Lecture 9).

Dummies in econometrics

- In econometrics, dummy variables are also typically used to understand average group differences in the **response** of Y_i to **explanatory variables** X_{ik}.

- This is done by **interacting** the explanatory variables with the dummy: by adding variables *à la* $X_{ik}D_i$ to the model.

- In practice, this can effectively amount to running separate group-specific regression models.

- It may be useful to perform this exercise if the focus of the analyst lies in a variable common to all groups.

- In this case it can be advisable to keep the model "flexible" and introduce multiple dummy variables and group-specific responses via interaction terms, thus fully leveraging on the *ceteris paribus* logic of the Frisch-Waugh-Lovell theorem.

Dummies in econometrics: an example (1/2)

- To substantiate, consider a Mincer equation that omits the squared term for experience X_i. Suppose that interest falls on differences between races (e.g. black and whites).

- The ensuing figure graphically displays the working dataset with observations colored by race (blacks are about 33%).

- In the following model, parameter π_0 expresses the average racial differences, *holding returns to education constant.*

$$\log W_i = \beta_0 + \pi_0 D_i + \beta_1 S_i + \varepsilon_i$$

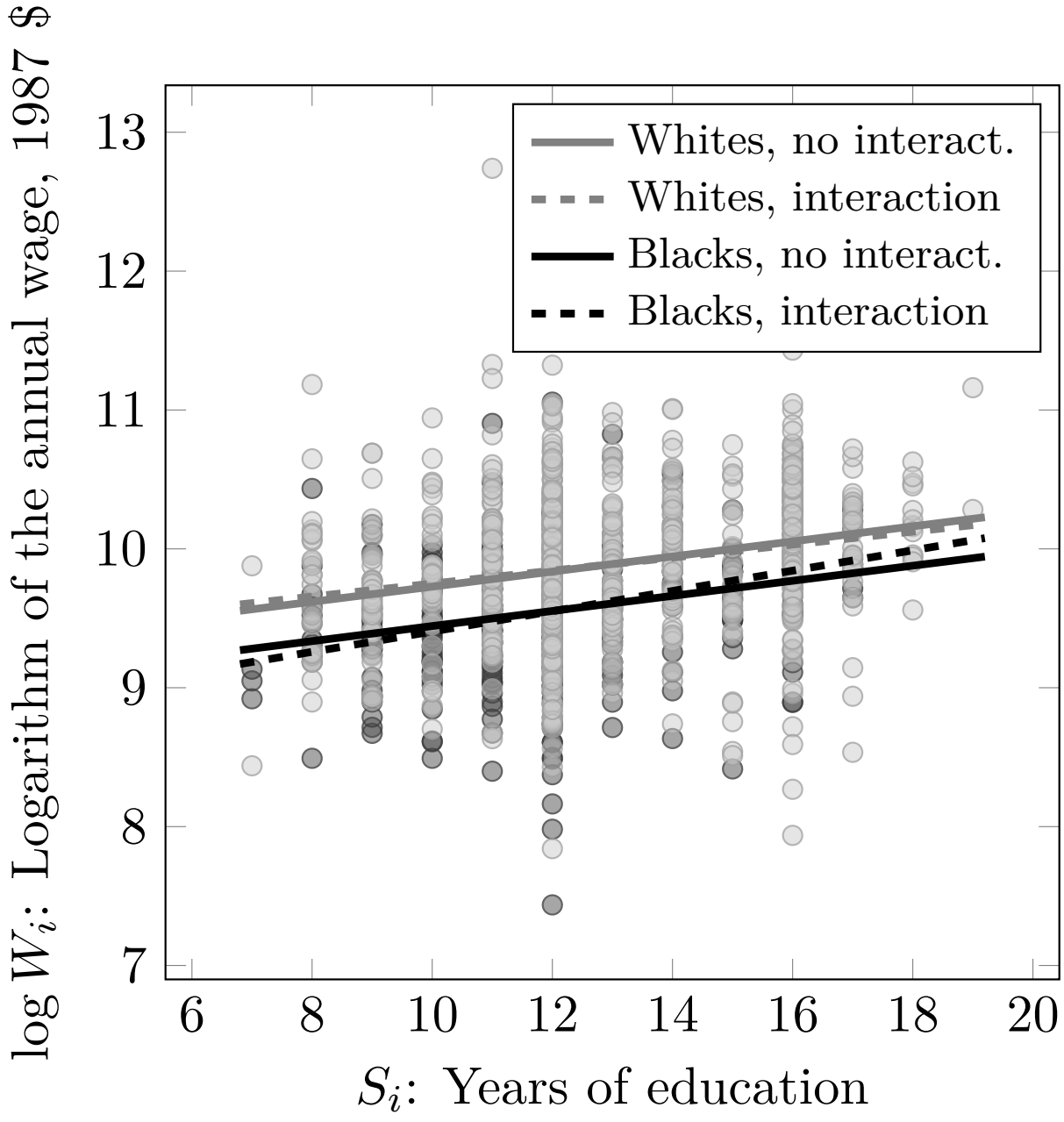
  Fitting this model results in the figure's solid lines.

- Adding an interaction term introduces *race-specific returns to education.*

$$\log W_i = \beta_0 + \pi_0 D_i + \beta_1 S_i + \pi_1 D_i S_i + \varepsilon_i$$

  Fitting this model results in the figure's dashed lines.

Dummies in econometrics: an example (2/2)



Regression and the CEF derivative (1/4)

- There is yet another property of the least squares solution that helps motivate the use of OLS in empirical analyses.

- In short: even if the true CEF is **unknown**, least squares (OLS) estimation can provide a useful **approximation** to the **average derivative** of the CEF.

- Suppose that the analyst's interest falls on the effect of an independent variable S_i upon a dependent variable Y_i. Their statistical relationship and CEF are unknown.

- However, in this setting the analyst can identify a vector of independent variables $\boldsymbol{x}_i$ such that:

$$\mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\pi}_0$$

  which occurs, for example, if $\boldsymbol{x}_i$ consists of a fully saturated set of dummy variables.

Regression and the CEF derivative (2/4)

- Define the following **derivative** of the CEF $\mu'_{Y|S,\boldsymbol{x}}(s_i;\boldsymbol{x}_i)$:

$$\mu'_{Y|S,\boldsymbol{x}}(s_i;\boldsymbol{x}_i) \equiv \frac{\partial}{\partial S_i}\,\mathbb{E}\left[Y_i|\,S_i;\boldsymbol{x}_i\right]\Big|_{S_i=s_i}$$

  and observe that this is a function of $\boldsymbol{x}_i$, evaluated at every point s_i on the support of S_i.

- Consider a least squares fit of Y_i on $\boldsymbol{x}_i$ and S_i.

$$Y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + \delta_0 S_i + \varepsilon_i$$

- The coefficient associated with S_i is the following variation of a partial correlation coefficient.

$$\widehat{\delta}_{OLS} = \frac{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\mathbf{y}}{\mathbf{s}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\mathbf{s}}$$

Regression and the CEF derivative (3/4)

- If the CEF of S_i conditional on $\boldsymbol{x}_i$ is linear, it is possible to show that, asymptotically:

$$\widehat{\delta}_{OLS} \xrightarrow{p} \delta^* = \frac{\mathbb{C}\text{ov}\left[Y_i, S_i \middle| \boldsymbol{x}_i\right]}{\mathbb{V}\text{ar}\left[S_i \middle| \boldsymbol{x}_i\right]}$$

- . . . and since $\mathbb{E}\left[S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right] = 0$, the above simplifies as:

$$\delta^* = \frac{\mathbb{E}\left[Y_i \left(S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)\right]}{\mathbb{E}\left[S_i \left(S_i - \mathbb{E}\left[S_i \middle| \boldsymbol{x}_i\right]\right)\right]}$$

- . . . and furthermore, Yitzhaki (1996) and then Angrist and Krueger (1999), incrementally demonstrated the following.

$$\delta^* = \frac{\mathbb{E}_{\boldsymbol{x}}\left[\int_{\mathbb{X}_S} \mu'_{Y|S,\boldsymbol{x}}\left(s_i; \boldsymbol{x}_i\right) \phi\left(s_i; \boldsymbol{x}_i\right) ds_i\right]}{\mathbb{E}_{\boldsymbol{x}}\left[\int_{\mathbb{X}_S} \phi\left(s_i; \boldsymbol{x}_i\right) ds_i\right]}$$

Regression and the CEF derivative (4/4)

- The Yitzhaki-Angrist-Krueger expression for δ^* is obtained directly from the manipulation of the probability limit.

- It allows for an interpretation of δ^* as the CEF **derivative** $\mu'_{Y|S,\boldsymbol{x}}(s_i;\boldsymbol{x}_i)$, **averaged** over the entire support of $\boldsymbol{x}_i$, and incorporating **weights** expressed by the term $\phi(s_i;\boldsymbol{x}_i)$.

- The weights are defined as:

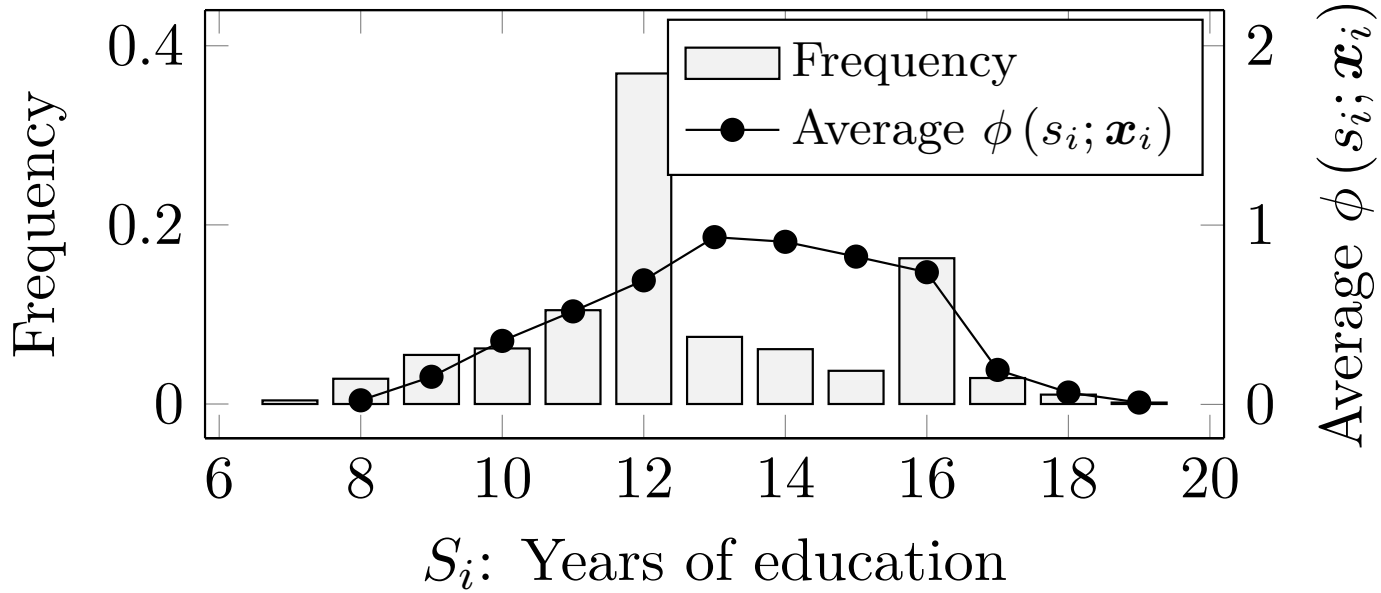
$$\phi(s_i;\boldsymbol{x}_i) \equiv \{\mathbb{E}[S_i|\, S_i \geq s_i, \boldsymbol{x}_i] - \mathbb{E}[S_i|\, S_i < s_i, \mathbf{x}_i]\} \times \\ \times \{\mathbb{P}(S_i \geq s_i|\, \boldsymbol{x}_i)[1 - \mathbb{P}(S_i \geq s_i|\, \boldsymbol{x}_i)]\}$$

  an expression which is not very easy to interpret. However, they clearly take higher values around the **median** of S_i.

- In selected contexts, like the one developed in the following example, this interpretation can prove quite useful.

Example: average returns to schooling (1/3)

- Consider the analysis of returns to schooling; let S_i here be education (measured as discrete years) whereas $\boldsymbol{x}_i$ is a fully saturated set of demographic indicators.
- The figure below shows the empirical frequency of S_i in the working data. Most observations report 12 (i.e. high school degree) or 16 (i.e. college degree) years of education.
- The **weights** $\phi(s_i; \boldsymbol{x}_i)$ are higher around these key values.



Example: average returns to schooling (2/3)

- The returns to schooling can be also expressed here as:

$$\mu'_{\log W|S,\boldsymbol{x}}(s_i;\boldsymbol{x}_i) \equiv$$
$$\equiv \mathbb{E}[\log W_i|\, S_i = s_i;\boldsymbol{x}_i] - \mathbb{E}[\log W_i|\, S_i = s_i - 1;\boldsymbol{x}_i]$$

  this is the **discrete** variation in the CEF of log-wages W_i following one extra year of education S_i.

- If the CEF is non-linear, this can be any function of s_i.

- Weighting $\mu'_{\log W|S,\boldsymbol{x}}(s_i;\boldsymbol{x}_i)$ by $\phi(s_i;\boldsymbol{x}_i)$ makes sense here!

- In fact, more weight is given to those values of S_i that are of major interest for **education policy**: $s_i \in \{12,\ldots,16\}$.

- In the working data, such weighted average is about 0.064: just like the least squares coefficient for education from the Mincer equation that omits experience!

Example: average returns to schooling (3/3)

- The figure below summarizes these findings. It displays the regression line from the model that omits experience...
- ...alongside the empirical CEF of log-wages – calculated as yearly averages – as well as its discrete variations.