M1 INTERMEDIATE ECONOMETRICS

ORDINARY LEAST SQUARES

Koen Jochmans

August 10, 2025

1. The least-squares problem

Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^k$ be random variables. The best linear predictor of Y given X (in the mean squared error sense) was the linear combination $X'\beta$,

$$\beta = \arg\min_b \mathbb{E}((Y - X'b)^2).$$

A random sample of size n on (Y, X) is obtained on independently drawing $(Y_1, X_1), \ldots, (Y_n, X_n)$ from the joint distribution of (Y, X). Moreover, for $1 \leq i < j \leq n$, the variables (Y_i, X_i) and (Y_j, X_j) are independent and identically distributed. A sample version of the best linear predictor then is $X'\hat{\beta}$,

$$\hat{\beta} = \arg\min_b \frac{1}{n} \sum_{i=1}^n (Y_i - X_i'b)^2.$$

Contrary to β, the coefficient vector $\hat{\beta}$ is a random variable because it is a function of the sample, which is random. Here we look at the statistical properties of $\hat{\beta}$ as an estimator of β.

2. The ordinary least-squares estimator

It is useful to proceed in matrix form. By random sampling we have, for each $1 \leq i \leq n$,

$$Y_i = X_i'\beta + e_i, \qquad \mathbb{E}(X_i e_i) = 0.$$

Collect the n outcome variables in the $n \times 1$ vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, the n vectors of k regressors in the $n \times k$ matrix $\boldsymbol{X} = (X_1', \ldots, X_n')'$, and the n prediction errors in the $n \times 1$ vector $\boldsymbol{e} = (e_1, \ldots, e_n)'$. Stacking the n equations gives

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}.$$

Hence,

$$\hat{\beta} = \arg\min_b (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) = \arg\min_b \|\boldsymbol{Y} - \boldsymbol{X}b\|^2,$$

for $\|\cdot\|$ the Euclidean norm. The normal equations for a minimum here are

$$\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\, b,$$

with unique solution

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$$

provided that the design matrix $\boldsymbol{X}'\boldsymbol{X}$ has rank k. This no-multicolinearity condition requires that the k columns of $\boldsymbol{X}$ are linearly independent. Indeed, $\boldsymbol{X}'\boldsymbol{X}$ is invertible if it is positive definite, that is, if $\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a} > 0$ for any vector $\boldsymbol{a}$ (different from the zero vector). But $\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a} = \|\boldsymbol{X}\boldsymbol{a}\|^2$ and indeed $\boldsymbol{X}\boldsymbol{a} \neq 0$ is the requirement that the $n \times k$ matrix $\boldsymbol{X}$ has maximal column rank.

3. Least-squares projection

We can define the residual vector $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$ as a sample version of the error $\boldsymbol{e}$. Then

$$\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = \|\hat{\boldsymbol{e}}\|^2 = \min_b \|\boldsymbol{Y} - \boldsymbol{X}b\|^2$$

By construction,

$$\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{Y}) - (\boldsymbol{X}'\boldsymbol{X})\,(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}) = 0$$

so the residual vector $\hat{\boldsymbol{e}}$ is orthogonal to the regressor matrix $\boldsymbol{X}$. This is a sample version of the orthogonality condition that $\mathbb{E}(Xe) = 0$ for the prediction error e.

It is useful to define the $n \times n$ matrices

$$\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}', \qquad \boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

These are projection matrices, as they are symmetric and satisfy $\boldsymbol{P}_{\boldsymbol{X}}^2 = \boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{M}_{\boldsymbol{X}}^2 = \boldsymbol{M}_{\boldsymbol{X}}$. The matrix $\boldsymbol{P}_{\boldsymbol{X}}$ projects onto the k-dimensional linear subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{X}$ (that is, the space that contains $\boldsymbol{X}b$ for any b). The matrix $\boldsymbol{M}_{\boldsymbol{X}}$ projects onto the orthogonal complement of this subspace (that is, the space that contains all vectors $\boldsymbol{u}$ that satisfy $\boldsymbol{X}'\boldsymbol{u} = 0$). Ordinary least squares looks for that linear combination of the columns of $\boldsymbol{X}$ that is closest to $\boldsymbol{Y}$, by making the residual vector $\hat{\boldsymbol{e}}$ as short as possible. We have

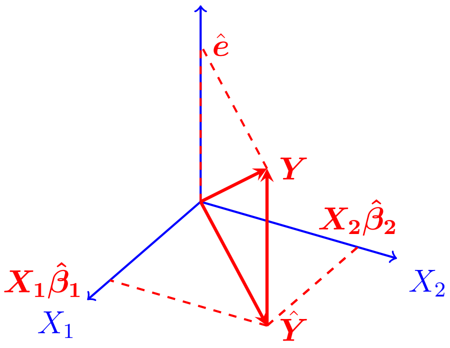
$$\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} = \boldsymbol{X}\hat{\beta} = \hat{\boldsymbol{Y}}, \qquad \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} = \hat{\boldsymbol{e}}.$$

Verify that $\hat{\boldsymbol{Y}}'\hat{\boldsymbol{e}} = 0$ as $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}} = 0$. Therefore,

$$\|\boldsymbol{Y}\|^2 = \|\hat{\boldsymbol{Y}} + \hat{\boldsymbol{e}}\|^2 = \|\hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{e}}\|^2.$$

The explanatory power of a regression is often measured by $R^2 = \|\hat{\boldsymbol{Y}}\|^2/\|\boldsymbol{Y}\|^2$. Figure 1 visually illustrates the least-squares projection in a three-dimensional space ($n = 3$) with two regressors ($k = 2$). In the plot the two regressors

Figure 1: Least-squares projection in a three-dimensional space



are orthogonal, i.e., the design matrix is diagonal, as can be seen by the 90° angle between the two regressor vectors.

For $\boldsymbol{X}$ to have maximal column rank we need to have $k \leq n$. The limit situation occurs when $k = n$, in which case we get no dimension reduction and so $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$; we fit the data perfectly. in the sequel we will be concerned with the sampling behavior of $\hat{\beta}$ when n grows, but k will be held fixed throughout. Such a sequence is not well suited for situations where k is not small relative to n.

4. Partitioned regression

Partition $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ so that

$$\boldsymbol{Y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{e}.$$

Then

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{Y} \\ \boldsymbol{X}_2'\boldsymbol{Y} \end{pmatrix}.$$

Some algebra using formulae for partitioned matrix inversion shows that the inverse matrix is equal to

$$\begin{pmatrix} (\boldsymbol{X}_1'\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1)^{-1} & -(\boldsymbol{X}_1'\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1} \\ -(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1)^{-1} & (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1} + (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2)(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1} \end{pmatrix},$$

from which we obtain

$$\hat{\beta}_1 = (\boldsymbol{X}_1'\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1'\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{Y}).$$

Because $\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1$ is the residual from a regression of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$, we can thus obtain $\hat{\beta}_1$ by including $\boldsymbol{X}_2$ as control variables in a multiple linear regression or by first filtering $\boldsymbol{X}_1$ from its linear dependence on $\boldsymbol{X}_2$ and next regressing $\boldsymbol{Y}$ on this filtered version of $\boldsymbol{X}_1$, which is $\boldsymbol{M}_{\boldsymbol{X}_2}\boldsymbol{X}_1$.

5. Small-sample behavior

The estimator $\hat{\beta}$ is a function of the sample. Because the sample is random so is $\hat{\beta}$. What can be said about its sampling distribution depends on several factors.

5.1. When $e|X = x \sim N(0, \sigma^2)$

In this case we know that

$$Y|X = x \sim N(x'\beta, \sigma^2).$$

This corresponds to the classical linear-regression setting. Because the (Y_i, X_i) are independent and identically distributed,

$$\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}_n), \qquad \text{or} \qquad \boldsymbol{e}|\boldsymbol{X} \sim N(0, \sigma^2\boldsymbol{I}_n).$$

Because $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}) = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{e})$ and β is constant, the estimator $\hat{\beta}$ is random because of the second term. Let $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Conditional on $\boldsymbol{X}$, $\boldsymbol{Ae}$ is a linear combination of independent normal random variables with variance σ^2. It is, therefore, also normally distributed. We have $\boldsymbol{Ae}|\boldsymbol{X} \sim N(0, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$. Here, the variance is obtained by noting that

$$\operatorname{var}(\boldsymbol{Ae}|\boldsymbol{X}) = \mathbb{E}(\boldsymbol{Aee}'\boldsymbol{A}'|\boldsymbol{X}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{ee}'|\boldsymbol{X})\boldsymbol{A}' = \sigma^2\boldsymbol{AA}' = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

As $\hat{\beta} = \beta + \boldsymbol{Ae}$ it immediately follows that

$$\hat{\beta}|\boldsymbol{X} \sim N(\beta, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

The conditioning here is important. The unconditional distribution of $\hat{\beta}$ is the mixture of this normal distribution with the distribution of $\boldsymbol{X}$ and will be non-normal, in general.

5.2. When $\mathbb{E}(e|X = x) = 0$ only

When we drop the distributional assumption on $\boldsymbol{e}|\boldsymbol{X}$ and instead only require that this distribution has mean zero we have that the CEF is the linear function

$$\mathbb{E}(Y|X = x) = x'\beta.$$

Other moments of the conditional distribution are left fully unrestricted, however. For example, the conditional variance $\mathbb{E}(e^2|X = x)$ is allowed to change with x in an unrestricted manner.

Consequently we can only make statements about the conditional mean of $\hat{\beta}|\boldsymbol{X}$. Moreover,

$$\mathbb{E}(\hat{\beta}|\boldsymbol{X}) = \mathbb{E}(\beta + \boldsymbol{A}\boldsymbol{e}|\boldsymbol{X}) = \beta + \boldsymbol{A}\,\mathbb{E}(\boldsymbol{e}|\boldsymbol{X}) = \beta$$

because $\mathbb{E}(\boldsymbol{e}|\boldsymbol{X}) = 0$. Hence, $\hat{\beta}$ is conditionally unbiased for β. It is also unconditionally unbiased, as $\mathbb{E}(\hat{\beta}) = \mathbb{E}(\mathbb{E}(\hat{\beta}|\boldsymbol{X})) = \beta$.

5.3. When $\mathbb{E}(Xe) = 0$ only

In this final case $X\beta$ is only the best linear predictor of Y and it no longer corresponds to the CEF. The sampling distribution of $\hat{\beta}$ in this case cannot be characterized with any generality. In particular, it is not possible to claim (even unconditional) unbiasedness of $\hat{\beta}$ for β, as this would require that $\mathbb{E}(\boldsymbol{A}\boldsymbol{e}) = \mathbb{E}(\boldsymbol{(X'X)^{-1}X'e})$ equals zero, which is not implied by $\mathbb{E}(Xe) = 0$.

6. Large-sample behavior

6.1. Consistency

We say that $\hat{\beta}$ is consistent for β if, for any $\varepsilon > 0$, it holds that

$$\mathbb{P}(\|\hat{\beta} - \beta\| > \varepsilon) \to 0$$

as $n \to \infty$. That is, the probability that $\hat{\beta}$ lies beyond a distance ε from β goes to zero as the sample size increases. We write $\hat{\beta} \underset{p}{\to} \beta$ as $n \to \infty$ or $\text{plim}_{n\to\infty}\hat{\beta} = \beta$.

To show consistency of $\hat{\beta}$ for β we will use the following four conditions:

C1. Random sampling.

C2. $\mathbb{E}(Xe) = 0$.

C3. $\mathbb{E}(\|X\|^2) < +\infty$.

C4. $\operatorname{rank} \mathbb{E}(XX') = k$.

C1 and C2 are both maintained assumptions. Indeed, C1 is a restatement of our sampling assumption while C2 merely formalizes that β is the coefficient of the best linear predictor. C3 has previously been made implicitly. It demands that the regressors have finite second moments. This ensures that the matrix $\mathbb{E}(XX')$ is well defined. C4, finally, demands this matrix to be full rank, so that its inverse is well defined. This ensures that β is uniquely defined.

To see consistency we begin by noting that

$$\hat{\beta} - \beta = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1} \left(\frac{\boldsymbol{X}'\boldsymbol{e}}{n}\right)$$

is the ratio of two sample averages. Consider the numerator first. Because of C1 and C2 we have

$$\frac{\boldsymbol{X}'\boldsymbol{e}}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i e_i \underset{p}{\rightarrow} \mathbb{E}(Xe) = 0$$

by a direct application of the law of large numbers. For the denominator, we first see that, because C1 and C3,

$$\frac{\boldsymbol{X}'\boldsymbol{X}}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \underset{p}{\rightarrow} \mathbb{E}(XX')$$

again follows by the law of large numbers. Then, because C4 ensures that the inverse of this matrix is well defined, the continuous-mapping theorem implies that

$$\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1} \underset{p}{\rightarrow} \mathbb{E}(XX')^{-1}.$$

Finally, by Slutsky's theorem, the product of the two terms converges to the product of their respective probability limits, yielding $\hat{\beta} - \beta \underset{p}{\rightarrow} \mathbb{E}(XX')^{-1}\, 0$, and so $\operatorname{plim}_{n\rightarrow\infty}\hat{\beta} = \beta$.

6.2. Asymptotic normality

Consistency implies that, in large samples, $\hat{\beta}$ can be expected to be 'close' to β. We can say much more, however. To do so we first need to strengthen C3 as

C3'. $\mathbb{E}(\|X\|^4) < +\infty$,

and additionally assume

C5. $\mathbb{E}(e^4) < +\infty$.

Under the conditions in C1, C2, C3', and C5, the random variables $X_i e_i$ are independent and identically distributed, with a finite mean (of zero, in fact), and variance

$$\operatorname{var}(Xe) = \mathbb{E}(e^2 XX') = \mathit{\Omega}.$$

This variance exists because $\|Xe\|^2 = e^2\|X\|^2$ and so, by the Cauchy-Schwarz inequality, $\mathbb{E}(\|Xe\|^2) = \mathbb{E}(e^2\|X\|^2) \leq \mathbb{E}(e^4)^{1/2}\,\mathbb{E}(\|X\|^4)^{1/2} < +\infty$. Thus, an application of the central limit theorem gives

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i \underset{d}{\rightarrow} N(0, \mathit{\Omega}).$$

Also, from before, $(n^{-1}\sum_{i=1}^{n} X_i X_i')^{-1} \underset{p}{\rightarrow} \mathbb{E}(XX')^{-1} = Q^{-1}$, which was well defined. Therefore,

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}'\boldsymbol{e}}{\sqrt{n}}\right) \underset{d}{\rightarrow} N(0, V_\beta)$$

for $V_\beta = Q^{-1} \Omega\, Q^{-1}$, by an application of Slutsky's theorem. This results means that, for large n, the distribution of $\hat{\beta}$ is well-approximated by the normal distribution with mean β and variance V_β/n; we write this as $\overset{a}{\sim}$, so that

$$\hat{\beta} \overset{a}{\sim} N(\beta, V_\beta/n).$$

When $\mathbb{E}(e^2|X = x) = \sigma^2$, the regression error is homoskedastic, and we have

$$\Omega = \mathbb{E}(e^2 XX') = \mathbb{E}(E(e^2|X)\, XX') = \sigma^2 Q$$

so that $V_\beta = \sigma^2 Q^{-1}$. When the conditional variance of the regression error varies with x we say that the errors are heteroskedastic. A look at the scatter plot in Figure 2, for example, reveals that, for the different education levels, the spread of wages around their mean changes with the level of education.

6.3. Variance estimation

The variance V_β is unknown but can be estimated as

$$\hat{V}_\beta = \hat{Q}^{-1} \hat{\Omega}\, \hat{Q}^{-1},$$

where $\hat{Q} = n^{-1} \sum_{i=1}^n X_i X_i'$ and $\hat{\Omega} = n^{-1} \sum_{i=1}^n \hat{e}_i^2 X_i X_i'$. This estimator is consistent, in that

$$\hat{V}_\beta \underset{p}{\rightarrow} V_\beta$$

as $n \to \infty$. This is so because each of its components is consistent. For $\hat{Q}$ this has already been shown. To show the same for $\hat{\Omega}$, we first observe that

$$\frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 X_i X_i' = \frac{1}{n} \sum_{i=1}^n ((\hat{e}_i - e_i) + e_i)^2 X_i X_i'$$

and then work out the square to get

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 X_i X_i' + \frac{1}{n}\sum_{i=1}^{n}(\hat{e}_i - e_i)^2 X_i X_i' - \frac{2}{n}\sum_{i=1}^{n}(\hat{e}_i - e_i)e_i X_i X_i'.$$

Here,

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 X_i X_i' \underset{p}{\rightarrow} \Omega$$

by the law of large numbers. Next, as $(\hat{e}_i - e_i) = -X_i'(\hat{\beta} - \beta)$ and $\hat{\beta} \underset{p}{\rightarrow} \beta$,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{e}_i - e_i)e_i\, X_i X_i' = -\left(\frac{1}{n}\sum_{i=1}^{n} e_i\, X_i X_i' X_i'\right)(\hat{\beta} - \beta) \underset{p}{\rightarrow} 0$$

because the term in brackets converges in probability to a well defined finite limit because of C3' and C5. In the same way,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{e}_i - e_i)^2 X_i X_i' \underset{p}{\rightarrow} 0,$$

so that, indeed, $\hat{\Omega} \underset{p}{\rightarrow} \Omega$. By Slutsky's theorem, $\hat{V}_\beta \underset{p}{\rightarrow} V_\beta$ then follows as claimed.

For a vector r the linear combination $r'\hat{\beta}$ is, in large samples, has variance $(r'\, V_\beta\, r)/n$, an estimator of which is $(r'\, \hat{V}_\beta\, r)/n$. The standard error of $r'\hat{\beta}$ is

$$\sqrt{\frac{(r'\, \hat{V}_\beta\, r)}{n}}.$$

It is an estimator of the variability of the estimator $r'\hat{\beta}$.

Under homoskedasticity an other estimator of V_β would be $\hat{\sigma}^2 \hat{Q}^{-1}$, where

$$\hat{\sigma}^2 = (\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}})/(n-k).$$

This is consistent for $\sigma^2 Q^{-1}$ because $\hat{\sigma}^2 \underset{p}{\rightarrow} \mathbb{E}(e^2)$ (the degrees-of-freedom correction is standard but not needed for this). However, if such an estimator is used when the errors are actually heteroskedastic, we are not estimating the correct variance. The resulting standard errors are not reliable and so should not be used.

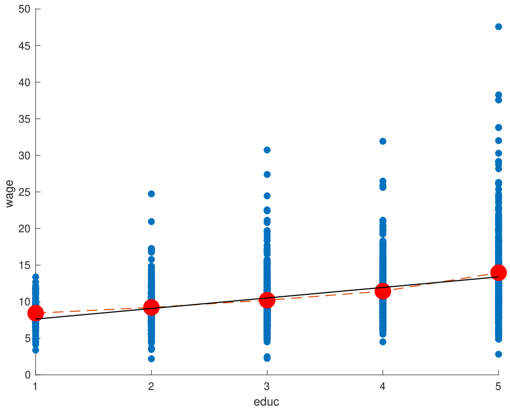
6.4. Illustrations

Wage data As a first example we consider a simple regression of hourly wage (in euro), `wage`, on an indicator of the education level (ranging from 1 to 5, with 1 being the lowest level and 5 the highest), `educ`. The data set covers 1472 individuals and comes from the European Community Household Panel.

In Figure 2 we provide a scatter plot of `wage` against `educ`. Because `educ` takes on only few values we can consider a saturated specification of the CEF that is linear in dummy variables for each of the levels of education. In this case the matrix $\boldsymbol{X}'\boldsymbol{X}$ is diagonal, with the diagonal entries equal to the number of observations for each of the levels of education in our data. The estimated regression coefficients then equal the average wage for each of these subsamples. These averages are marked with a red dot in Figure 2. The progression appears fairly linear. We verify this by fitting a simple linear regression of `wage` on (a constant and) `educ`, given by the solid black line in the plot. The coefficient estimate on `educ` is 1.44, implying that we predict an increase in the hourly wage rate of 1.44 euro for each additional level of education.

The data also includes a measure of labor market experience (the years of experience), `exper`. The correlation between `wage` and `exper` is 0.31. The correlation between `educ` and `exper` is -0.29. So more experienced workers

Figure 2: A simple wage regression



tend to earn more but also tend to have lower education. A multiple least squares regression of `wage` on (a constant and) `educ` and `exper` should thus gives us an upward revision of the coefficient estimate on `educ`. Indeed, we now find 1.93. Thus, we now predict an increase in the hourly wage rate of 1.93 euro for each additional level of education while keeping the experience level fixed. The first two columns in Table 1 give the detailed regression results. The numbers between brackets are the standard errors obtained as explained above.

We can go further in trying to net out the correlation of `educ` with other factors. In our data the only remaining variable we have is a binary indicator for gender, `male`, which is one for males and zero otherwise. The third column in Table 1 gives regression results for the specification where we additionally include `male` as a regressor. The coefficient of 1.346 shows that we predict a wage differential of 1.35 euro/hour for males relative to females with the same levels of education and experience. This estimate is of the same magnitude as the returns to education (which here is estimated at 1.99 euro/hour). In our final specification we refine this specification by allowing for the returns to education and experience to be different for males and females. We do this by

Table 1: Wage regressions

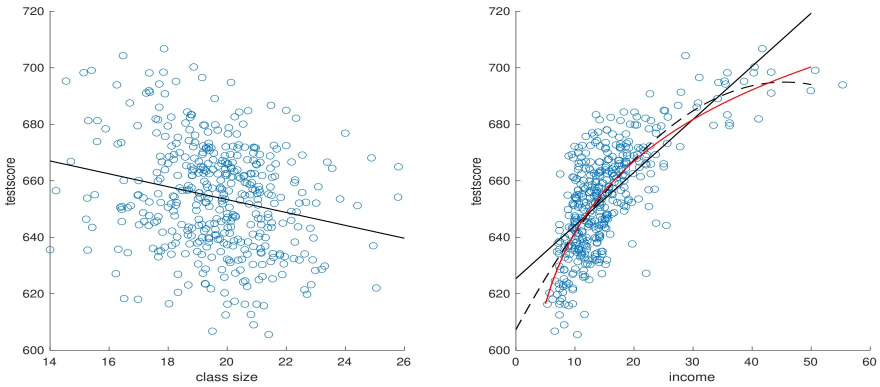
`wage`	(1)	(2)	(3)	(4)
`educ`	1.440	1.930	1.986	1.873
	(0.095)	(0.097)	(0.097)	(0.130)
`exper`		0.201	0.192	0.164
		(0.011)	(0.011)	(0.015)
`male`			1.346	0.032
			(0.188)	(0.760)
`educ*male`				0.168
				(0.183)
`exper*male`				0.045
				(0.021)
`constant`	6.185	1.074	0.214	1.053
	(0.280)	(0.414)	(0.448)	(0.533)
R-squared	0.152	0.344	0.366	0.368

including the interaction terms `educ*male` and `exper*male`. This amounts to running the previous regression separately for mean and women. Gender, by itself, now becomes virtually irrelevant for our prediction. We do find some heterogeneity in the different returns by gender. For females the return to education (conditional on experience) is estimated at 1.87 euro/hour while it is 1.873+0.168 = 2.04 euro/hour for men. The adjustment of 0.168 is not large relative to its standard error of 0.183, however. The returns to experience (holding the level of education fixed) are 0.16 euro/hour and 0.21 euro/hour, respectively.

Test score data Our second example is inspired by discussions in the textbook of Stock and Watson (Introduction to Econometrics, 2003) and uses data from 420 Californian school districts. In the left plot of Figure 3 we plot the average test-score for each district against the average class size in that district. The regression line for the simple regression of `testscore` on `class size` (and, as always, a constant) is also provided in the same plot.

The negative slope shows a negative correlation between test performance and class size. The right plot of the same figure scatters `testscore` against the average district income. Here the cloud of points suggests a nonlinear relationship. We provide three different regression lines. The first (in full black) is the simple linear regression of `testscore` on `income`. The second (in dashed black) includes the quadratic term `income`^{2} and is able to capture the decreasing returns to income. It is clear, however, that this line will bend back down as income increases, which is undesirable. A simple alternative is to regress `testscore` on `log(income)` instead; this is the third regression line, in full red. The regressions just discussed are given in Columns (1) to (4) of Table 2.

Figure 3: Simple test-score regressions



The results so far, taken independently, find a substantial negative impact of 2.28 points on test score for each additional one-person increase in class size and a predicted increase of 0.36 points for a 1% increase in income. In Column (5) we give the results of a regression of `testscore` on both these independent variables. While the coefficient on `log(income)` changes little, controlling for income drastically reduces the estimated negative impact of `class size` on test score performance. These changes, in part, reflect the heterogeneity

Table 2: Test score regressions

testscore	(1)	(2)	(3)	(4)	(5)	(6)
class size	-2.280				-0.879	-0.385
	(0.519)				(0.340)	(0.290)
income		1.879	3.851			
		(0.114)	(0.268)			
income^2			-0.042			
			(0.005)			
log(income)				36.420	35.620	28.360
				(1.397)	(1.400)	(1.329)
learner						-0.429
						(0.032)
constant	698.9	625.4	607.3	557.8	577.2	593.5
	(10.4)	(1.9)	(2.9)	(3.8)	(7.9)	(6.9)
R-squared	0.051	0.508	0.556	0.563	0.570	0.712

in school districts and the importance of this heterogeneity in explaining test scores. Also note the drastic difference in R^2 between the simple regressions on `class size` and `income` (or `log(income)`), respectively. To go further, Column (6) provides results where we include the variable `learner`, which is the percentage of children in the district that do not have English as their native language. Districts where there are more such children tend to have lower test scores and smaller class sizes (this can be seen in a regression not reported here), and so controlling for this confounding variable would seem important. Indeed, doing so further reduces the magnitude of the estimated coefficient on class size. The magnitude is now of a comparable order as its standard error.