Chapter 3

The Algebra of Least Squares

3.1 Introduction

In this chapter we introduce the popular least squares estimator. Most of the discussion will be algebraic, with questions of distribution and inference deferred to later chapters.

3.2 Samples

In Section 2.18 we derived and discussed the best linear predictor of Y given X for a pair of random variables $(Y, X) \in \mathbb{R} \times \mathbb{R}^k$ and called this the linear projection model. We are now interested in **estimating** the parameters of this model, in particular the projection coefficient

$$\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1} \mathbb{E}\left[XY\right]. \tag{3.1}$$

We can estimate β from samples which include joint measurements of (Y, X). For example, supposing we are interested in estimating a wage equation, we would use a dataset with observations on wages (or weekly earnings), education, experience (or age), and demographic characteristics (gender, race, location). One possible dataset is the Current Population Survey (CPS), a survey of U.S. households which includes questions on employment, income, education, and demographic characteristics.

Notationally we wish to distinguish observations (realizations) from the underlying random variables. The random variables are (Y, X). The observations are (Y_i, X_i). From the vantage of the researcher the latter are numbers. From the vantage of statistical theory we view them as realizations of random variables. For individual observations we append a subscript i which runs from 1 to n, thus the i^{th} observation is (Y_i, X_i). The number n is the sample size. The **dataset** or **sample** is $\{(Y_i, X_i) : i = 1, ..., n\}$.

From the viewpoint of empirical analysis a dataset is an array of numbers. It is typically organized as a table where each column is a variable and each row is an observation. For empirical analysis the dataset is fixed in the sense that they are numbers presented to the researcher. For statistical analysis we view the dataset as random, or more precisely as a realization of a random process.

The individual observations could be draws from a common (homogeneous) distribution or could be draws from heterogeneous distributions. The simplest approach is to assume homogeneity – that the observations are realizations from an identical underlying population F.

Assumption 3.1 The variables $\{(Y_1, X_1), ..., (Y_i, X_i), ..., (Y_n, X_n)\}$ are **identically distributed**; they are draws from a common distribution F.

This assumption does not need to be viewed as literally true. Rather it is a useful modeling device so that parameters such as β are well defined. This assumption should be interpreted as how we view an observation *a priori*, before we actually observe it. If I tell you that we have a sample with $n = 59$ observations set in no particular order, then it makes sense to view two observations, say 17 and 58, as draws from the same distribution. We have no reason to expect anything special about either observation.

In econometric theory we refer to the underlying common distribution F as the **population**. Some authors prefer the label the **data-generating-process** (DGP). You can think of it as a theoretical concept or an infinitely-large potential population. In contrast we refer to the observations available to us $\{(Y_i, X_i) : i = 1,...,n\}$ as the **sample** or **dataset**. In some contexts the dataset consists of all potential observations, for example administrative tax records may contain every single taxpayer in a political unit. Even in this case we view the observations as if they are random draws from an underlying infinitely-large population as this will allow us to apply the tools of statistical theory.

The linear projection model applies to the random variables (Y, X). This is the probability model as that described in Section 2.18. The model is

$$Y = X'\beta + e \tag{3.2}$$

where the linear projection coefficient β is defined as

$$\beta = \underset{b\in\mathbb{R}^k}{\operatorname{argmin}} S(b), \tag{3.3}$$

the minimizer of the expected squared error

$$S(\beta) = \mathbb{E}\left[\left(Y - X'\beta\right)^2\right]. \tag{3.4}$$

The coefficient has the explicit solution

$$\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right]. \tag{3.5}$$

3.3 Moment Estimators

We want to estimate the coefficient β defined in (3.5) from the sample of observations. Notice that β is written as a function of certain population expectations. In this context an appropriate estimator is the same function of the sample moments. Let's explain this in detail.

To start, suppose that we are interested in the population mean μ of a random variable Y with distribution function F

$$\mu = \mathbb{E}\left[Y\right] = \int_{-\infty}^{\infty} y dF(y). \tag{3.6}$$

The expectation μ is a function of the distribution F as written in (3.6). To estimate μ given n random variables Y_i from F a natural estimator is the sample mean

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Notice that we have written this using two pieces of notation. The notation $\overline{Y}$ with the bar on top is conventional for a sample mean. The notation $\widehat{\mu}$ with the hat "^" is conventional in econometrics to denote an estimator of the parameter μ. In this case $\overline{Y}$ is the estimator of μ, so $\widehat{\mu}$ and $\overline{Y}$ are the same. The sample mean $\overline{Y}$ can be viewed as the natural analog of the population mean (3.6) because $\overline{Y}$ equals the expectation (3.6) with respect to the empirical distribution – the discrete distribution which puts weight

$1/n$ on each observation Y_i. There are many other justifications for $\overline{Y}$ as an estimator for μ. We will defer these discussions for now. Suffice it to say that it is the conventional estimator in the lack of other information about μ or the distribution of Y.

Now suppose that we are interested in a set of population expectations of possibly nonlinear functions of a random vector Y, say $\mu = \mathbb{E}[h(Y)]$. For example, we may be interested in the first two moments of Y, $\mathbb{E}[Y]$ and $\mathbb{E}\left[Y^2\right]$. In this case the natural estimator is the vector of sample means,

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(Y_i).$$

We call $\widehat{\mu}$ the **moment estimator** for μ. For example, if $h(y) = (y, y^2)'$ then $\widehat{\mu}_1 = n^{-1}\sum_{i=1}^{n} Y_i$ and $\widehat{\mu}_2 = n^{-1}\sum_{i=1}^{n} Y_i^2$.

Now suppose that we are interested in a nonlinear function of a set of moments. For example, consider the variance of Y

$$\sigma^2 = \operatorname{var}[Y] = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2.$$

In general, many parameters of interest can be written as a function of moments of Y. Notationally, $\beta = g(\mu)$ and $\mu = \mathbb{E}[h(Y)]$. Here, Y are the random variables, $h(Y)$ are functions (transformations) of the random variables, and μ is the mean (expectation) of these functions. β is the parameter of interest, and is the (nonlinear) function $g(\cdot)$ of these means.

In this context a natural estimator of β is obtained by replacing μ with $\widehat{\mu}$. Thus $\widehat{\beta} = g\left(\widehat{\mu}\right)$. The estimator $\widehat{\beta}$ is often called a "plug-in" estimator. We typically call $\widehat{\beta}$ a moment, or moment-based, estimator of β since it is a natural extension of the moment estimator $\widehat{\mu}$.

Take the example of the variance $\sigma^2 = \operatorname{var}[Y]$. Its moment estimator is

$$\widehat{\sigma}^2 = \widehat{\mu}_2 - \widehat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2.$$

This is not the only possible estimator for σ^2 (there is also the well-known bias-corrected estimator) but $\widehat{\sigma}^2$ is a straightforward and simple choice.

3.4 Least Squares Estimator

The linear projection coefficient β is defined in (3.3) as the minimizer of the expected squared error $S(\beta)$ defined in (3.4). For given β, the expected squared error is the expectation of the squared error $\left(Y - X'\beta\right)^2$. The moment estimator of $S(\beta)$ is the sample average:

$$\widehat{S}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2 = \frac{1}{n}\mathrm{SSE}(\beta) \tag{3.7}$$

where

$$\mathrm{SSE}(\beta) = \sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2$$

is called the **sum of squared errors** function.

Since $\widehat{S}(\beta)$ is a sample average we can interpret it as an estimator of the expected squared error $S(\beta)$. Examining $\widehat{S}(\beta)$ as a function of β is informative about how $S(\beta)$ varies with β. Since the projection coefficient minimizes $S(\beta)$ an analog estimator minimizes (3.7).

We define the estimator $\widehat{\beta}$ as the minimizer of $\widehat{S}(\beta)$.

Definition 3.1 The **least squares estimator** is $\widehat{\beta} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \widehat{S}(\beta)$ where $\widehat{S}(\beta) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - X_i'\beta\right)^2$.

As $\widehat{S}(\beta)$ is a scale multiple of SSE(β) we may equivalently define $\widehat{\beta}$ as the minimizer of $SSE(\beta)$. Hence $\widehat{\beta}$ is commonly called the **least squares (LS)** estimator of β. The estimator is also commonly refered to as the **ordinary least squares (OLS)** estimator. For the origin of this label see the historical discussion on Adrien-Marie Legendre below. Here, as is common in econometrics, we put a hat "^" over the parameter β to indicate that $\widehat{\beta}$ is a sample estimate of β. This is a helpful convention. Just by seeing the symbol $\widehat{\beta}$ we can immediately interpret it as an estimator (because of the hat) of the parameter β. Sometimes when we want to be explicit about the estimation method, we will write $\widehat{\beta}_{\text{ols}}$ to signify that it is the OLS estimator. It is also common to see the notation $\widehat{\beta}_n$, where the subscript "n" indicates that the estimator depends on the sample size n.

It is important to understand the distinction between population parameters such as β and sample estimators such as $\widehat{\beta}$. The population parameter β is a non-random feature of the population while the sample estimator $\widehat{\beta}$ is a random feature of a random sample. β is fixed, while $\widehat{\beta}$ varies across samples.

3.5 Solving for Least Squares with One Regressor

For simplicity, we start by considering the case $k = 1$ so that there is a scalar regressor X and a scalar coefficient β. To illustrate, Figure 3.1(a) displays a scatter plot[1] of 20 pairs (Y_i, X_i).

The sum of squared errors SSE(β) is a function of β. Given β we calculate the "error" $Y_i - X_i\beta$ by taking the vertical distance between Y_i and $X_i\beta$. This can be seen in Figure 3.1(a) by the vertical lines which connect the observations to the straight line. These vertical lines are the errors $Y_i - X_i\beta$. The sum of squared errors is the sum of the 20 squared lengths.

The sum of squared errors is the function

$$\text{SSE}(\beta) = \sum_{i=1}^n \left(Y_i - X_i\beta\right)^2 = \left(\sum_{i=1}^n Y_i^2\right) - 2\beta\left(\sum_{i=1}^n X_iY_i\right) + \beta^2\left(\sum_{i=1}^n X_i^2\right).$$

This is a quadratic function of β. The sum of squared error function is displayed in Figure 3.1(b) over the range $[2,4]$. The coefficient β ranges along the x-axis. The sum of squared errors SSE(β) as a function of β is displayed on the y-axis.

The OLS estimator $\widehat{\beta}$ minimizes this function. From elementary algebra we know that the minimizer of the quadratic function $a - 2bx + cx^2$ is $x = b/c$. Thus the minimizer of SSE(β) is

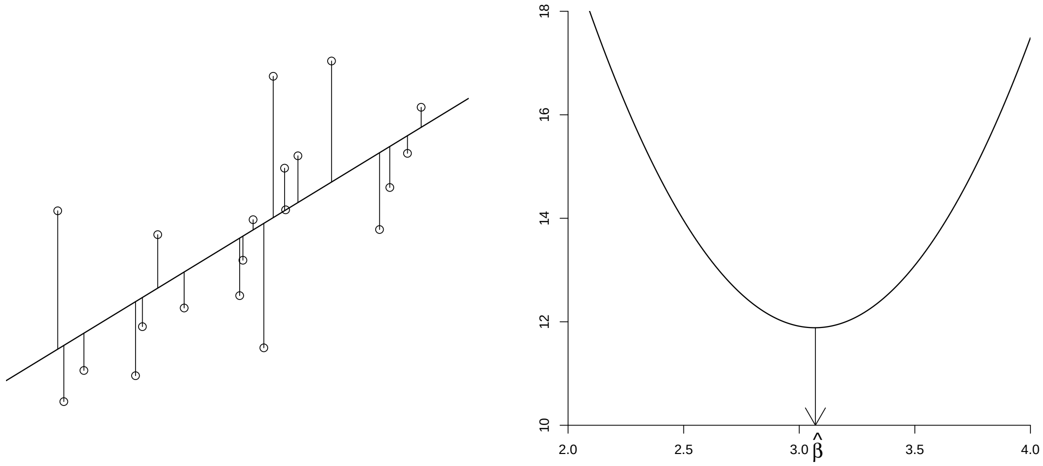
$$\widehat{\beta} = \frac{\sum_{i=1}^n X_iY_i}{\sum_{i=1}^n X_i^2}. \tag{3.8}$$

For example, the minimizer of the sum of squared error function displayed in Figure 3.1(b) is $\widehat{\beta} = 3.07$, and is marked on the x-axis.

The intercept-only model is the special case $X_i = 1$. In this case we find

$$\widehat{\beta} = \frac{\sum_{i=1}^n 1Y_i}{\sum_{i=1}^n 1^2} = \frac{1}{n}\sum_{i=1}^n Y_i = \overline{Y}, \tag{3.9}$$

[1] The observations were generated by simulation as $X \sim U[0,1]$ and $Y \sim \mathrm{N}[3X,1]$.



(a) Deviation from Fitted Line

(b) Sum of Squared Error Function

Figure 3.1: Regression With One Regressor

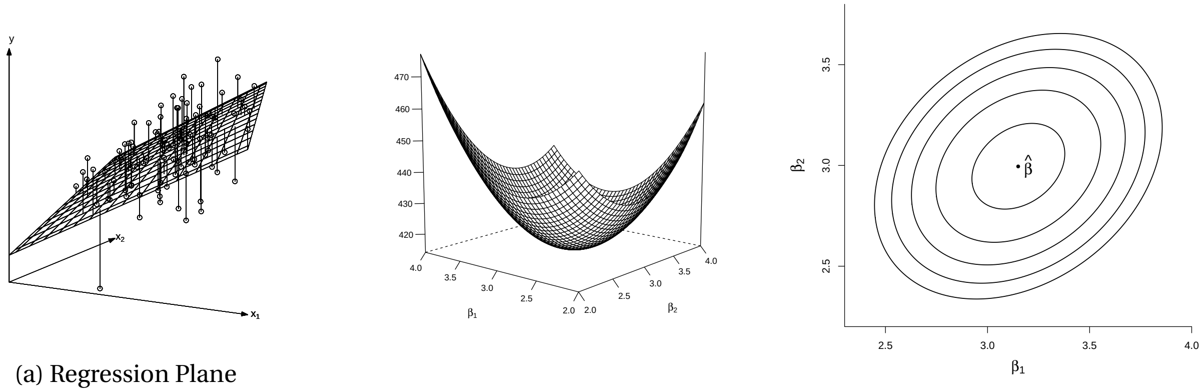
the sample mean of Y_i. Here, as is common, we put a bar "$^-$" over Y to indicate that the quantity is a sample mean. This shows that the OLS estimator in the intercept-only model is the sample mean.

Technically, the estimator $\widehat{\beta}$ in (3.8) only exists if the denominator is non-zero. Since it is a sum of squares it is necessarily non-negative. Thus $\widehat{\beta}$ exists if $\sum_{i=1}^{n} X_i^2 > 0$.

3.6 Solving for Least Squares with Multiple Regressors

We now consider the case with $k > 1$ so that the coefficient $\beta \in \mathbb{R}^k$ is a vector.

To illustrate, Figure 3.2(a) displays a scatter plot of 100 triples (Y_i, X_{1i}, X_{2i}). The regression function $x'\beta = x_1\beta_1 + x_2\beta_2$ is a 2-dimensional surface and is shown as the plane in Figure 3.2(a).



(a) Regression Plane

(b) Sum of Squared Error Function

(c) SSE Contour

Figure 3.2: Regression with Two Variables

The sum of squared errors SSE(β) is a function of the vector β. For any β the error $Y_i - X_i'\beta$ is the vertical distance between Y_i and $X_i'\beta$. This can be seen in Figure 3.2(a) by the vertical lines which connect the observations to the plane. As in the single regressor case these vertical lines are the errors $e_i = Y_i - X_i'\beta$. The sum of squared errors is the sum of the 100 squared lengths.

The sum of squared errors can be written as

$$\mathrm{SSE}(\beta) = \sum_{i=1}^{n} Y_i^2 - 2\beta' \sum_{i=1}^{n} X_i Y_i + \beta' \sum_{i=1}^{n} X_i X_i' \beta.$$

As in the single regressor case this is a quadratic function in β. The difference is that in the multiple regressor case this is a vector-valued quadratic function. To visualize the sum of squared errors function Figure 3.2(b) displays $\mathrm{SSE}(\beta)$. Another way to visualize a 3-dimensional surface is by a contour plot. A contour plot of the same $\mathrm{SSE}(\beta)$ function is shown in Figure 3.2(c). The contour lines are points in the (β_1, β_2) space where $\mathrm{SSE}(\beta)$ takes the same value. The contour lines are elliptical.

The least squares estimator $\widehat{\beta}$ minimizes $\mathrm{SSE}(\beta)$. A simple way to find the minimum is by solving the first-order conditions. The latter are

$$0 = \frac{\partial}{\partial \beta}\mathrm{SSE}(\widehat{\beta}) = -2\sum_{i=1}^{n} X_i Y_i + 2\sum_{i=1}^{n} X_i X_i' \widehat{\beta}. \tag{3.10}$$

We have written this using a single expression, but it is actually a system of k equations with k unknowns (the elements of $\widehat{\beta}$).

The solution for $\widehat{\beta}$ may be found by solving the system of k equations in (3.10). We can write this solution compactly using matrix algebra. Dividing (3.10) by 2 we obtain

$$\sum_{i=1}^{n} X_i X_i' \widehat{\beta} = \sum_{i=1}^{n} X_i Y_i. \tag{3.11}$$

This is a system of equations of the form $\boldsymbol{A}\boldsymbol{b} = \boldsymbol{c}$ where $\boldsymbol{A}$ is $k \times k$ and $\boldsymbol{b}$ and $\boldsymbol{c}$ are $k \times 1$. The solution is $\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{c}$, and can be obtained by pre-multiplying $\boldsymbol{A}\boldsymbol{b} = \boldsymbol{c}$ by $\boldsymbol{A}^{-1}$ and using the matrix inverse property $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k$. Applied to (3.11) we find an explicit formula for the least squares estimator

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right). \tag{3.12}$$

This is the natural estimator of the best linear projection coefficient β defined in (3.3), and can also be called the linear projection estimator.

Recall that we claimed that $\widehat{\beta}$ in (3.12) is the minimizer of $\mathrm{SSE}(\beta)$, and we found this by solving the first-order conditions. To be complete we should verify the second-order conditions. We calculate that

$$\frac{\partial^2}{\partial \beta \partial \beta'}\mathrm{SSE}(\beta) = 2\sum_{i=1}^{n} X_i X_i' > 0$$

which is a positive definite matrix. This shows that the second-order condition for minimization is satisfied so $\widehat{\beta}$ is indeed the unique minimizer of $\mathrm{SSE}(\beta)$.

Returning to the example sum of squared errors function $\mathrm{SSE}(\beta)$ displayed in Figure 3.2(b), the least squares estimator $\widehat{\beta}$ is the the pair $(\widehat{\beta}_1, \widehat{\beta}_2)$ which minimize this function; visually it is the low spot in the 3-dimensional graph, and is marked in Figure 3.2(c) as the center point of the contour plots.

Returning to equation (3.12) suppose that $k = 1$. In this case X_i is scalar so $X_i X_i' = X_i^2$. Then (3.12) simplifies to the expression (3.8) previously derived. The expression (3.12) is a notationally simple generalization but requires a careful attention to vector and matrix manipulations.

Alternatively, equation (3.5) writes the projection coefficient β as an explicit function of the population moments $\boldsymbol{Q}_{XY}$ and $\boldsymbol{Q}_{XX}$. Their moment estimators are the sample moments

$$\widehat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$
$$\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'.$$

The moment estimator of β replaces the population moments in (3.5) with the sample moments:

$$\begin{aligned}\widehat{\beta} &= \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{XY} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) \\ &= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right)\end{aligned}$$

which is identical with (3.12).

Technically, the estimator $\widehat{\beta}$ in (3.12) exists and is unique only if the inverted matrix is actually invertible, which holds if (and only if) this matrix is positive definite. This excludes the case that X_i contains redundant regressors. This will be discussed further in Section 3.24.

Theorem 3.1 If $\sum_{i=1}^{n} X_i X_i' > 0$, the least squares estimator equals

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right).$$

Adrien-Marie Legendre

The method of least squares was first published in 1805 by the French mathematician Adrien-Marie Legendre (1752-1833). Legendre proposed least-squares as a solution to the algebraic problem of solving a system of equations when the number of equations exceeded the number of unknowns. This was a vexing and common problem in astronomical measurement. As viewed by Legendre, (3.2) is a set of n equations with k unknowns. As the equations cannot be solved exactly, Legendre's goal was to select β to make the set of errors as small as possible. He proposed the sum of squared error criterion and derived the algebraic solution presented above. As he noted, the first-order conditions (3.10) is a system of k equations with k unknowns which can be solved by "ordinary" methods. Hence the method became known as **Ordinary Least Squares** and to this day we still use the abbreviation OLS to refer to Legendre's estimation method.

3.7 Illustration

We illustrate the least squares estimator in practice with the data set used to calculate the estimates reported in Chapter 2. This is the March 2009 Current Population Survey, which has extensive information on the U.S. population. This data set is described in more detail in Section 3.22. For this illustration we use the sub-sample of married (spouse present) Black female wage earners with 12 years potential work experience. This sub-sample has 20 observations.

In Table 3.1 we display the observations for reference. Each row is an individual observation which are the data for an individual person. The columns correspond to the variables (measurements) for the individuals. The second column is the reported wage (total annual earnings divided by hours worked). The third column is the natural logarithm of the wage. The fourth column is years of education. The fifth and six columns are further transformations, specifically the square of *education* and the product of *education* and log(*wage*). The bottom row are the sums of the elements in that column.

Table 3.1: Observations From CPS Data Set

Observation	Wage	log(Wage)	Education	Education^2	Education*log(Wage)
1	37.93	3.64	18	324	65.44
2	40.87	3.71	18	324	66.79
3	14.18	2.65	13	169	34.48
4	16.83	2.82	16	256	45.17
5	33.17	3.50	16	256	56.03
6	29.81	3.39	18	324	61.11
7	54.62	4.00	16	256	64.00
8	43.08	3.76	18	324	67.73
9	14.42	2.67	12	144	32.03
10	14.90	2.70	16	256	43.23
11	21.63	3.07	18	324	55.44
12	11.09	2.41	16	256	38.50
13	10.00	2.30	13	169	29.93
14	31.73	3.46	14	196	48.40
15	11.06	2.40	12	144	28.84
16	18.75	2.93	16	256	46.90
17	27.35	3.31	14	196	46.32
18	24.04	3.18	16	256	50.76
19	36.06	3.59	18	324	64.53
20	23.08	3.14	16	256	50.22
Sum	515	62.64	314	5010	995.86

Putting the variables into the standard regression notation, let Y_i be log wages and X_i be years of education and an intercept. Then from the column sums in Table 3.1 we have

$$\sum_{i=1}^{n} X_i Y_i = \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix}$$

and

$$\sum_{i=1}^{n} X_i X_i' = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}.$$

Taking the inverse we obtain

$$\left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} = \left(\begin{array}{cc} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{array}\right).$$

Thus by matrix multiplication

$$\widehat{\beta} = \left(\begin{array}{cc} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{array}\right)\left(\begin{array}{c} 995.86 \\ 62.64 \end{array}\right) = \left(\begin{array}{c} 0.155 \\ 0.698 \end{array}\right).$$

In practice the regression estimates $\widehat{\beta}$ are computed by computer software without the user taking the explict steps listed above. However, it is useful to understand that the least squares estimator can be calculated by simple algebraic operations. If your data is in a spreadsheet similar to Table 3.1, then the listed transformations (logarithm, squares and cross-products, column sums) can be computed by spreadsheet operations. $\widehat{\beta}$ could then be calculated by matrix inversion and multiplication. Once again, this is rarely done by applied economists since computer software is available to ease the process.

We often write the estimated equation using the format

$$\widehat{\log(wage)} = 0.155\ education + 0.698. \tag{3.13}$$

An interpretation of the estimated equation is that each year of education is associated with a 16% increase in mean wages.

Equation (3.13) is called a **bivariate regression** as there are two variables. It is also called a **simple regression** as there is a single regressor. A **multiple regression** has two or more regressors and allows a more detailed investigation. Let's take an example similar to (3.13) but include all levels of experience. This time we use the sub-sample of single (never married) Asian men which has 268 observations. Including as regressors years of potential work experience (*experience*) and its square (*experience*^{2}/100) (we divide by 100 to simplify reporting) we obtain the estimates

$$\widehat{\log(wage)} = 0.143\ education + 0.036\ experience - 0.071\ experience^2/100 + 0.575. \tag{3.14}$$

These estimates suggest a 14% increase in mean wages per year of education holding experience constant.

3.8 Least Squares Residuals

As a by-product of estimation we define the **fitted value** $\widehat{Y}_i = X_i'\widehat{\beta}$ and the **residual**

$$\widehat{e}_i = Y_i - \widehat{Y}_i = Y_i - X_i'\widehat{\beta}. \tag{3.15}$$

Sometimes $\widehat{Y}_i$ is called the predicted value but this is a misleading label. The fitted value $\widehat{Y}_i$ is a function of the entire sample, including Y_i, and thus cannot be interpreted as a valid prediction of Y_i. It is thus more accurate to describe $\widehat{Y}_i$ as a *fitted* rather than a *predicted* value.

Note that $Y_i = \widehat{Y}_i + \widehat{e}_i$ and

$$Y_i = X_i'\widehat{\beta} + \widehat{e}_i. \tag{3.16}$$

We make a distinction between the **error** e_i and the **residual** $\widehat{e}_i$. The error e_i is unobservable while the residual $\widehat{e}_i$ is an estimator. These two variables are frequently mislabeled which can cause confusion.

Equation (3.10) implies that

$$\sum_{i=1}^{n} X_i \widehat{e}_i = 0. \tag{3.17}$$

To see this by a direct calculation, using (3.15) and (3.12),

$$\begin{aligned}\sum_{i=1}^{n} X_i \widehat{e}_i &= \sum_{i=1}^{n} X_i \left(Y_i - X_i'\widehat{\beta}\right) \\ &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i X_i'\widehat{\beta} \\ &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i X_i' \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right) \\ &= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i Y_i = 0.\end{aligned}$$

When X_i contains a constant an implication of (3.17) is

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i = 0. \tag{3.18}$$

Thus the residuals have a sample mean of zero and the sample correlation between the regressors and the residual is zero. These are algebraic results, and hold true for all linear regression estimates.

3.9 Demeaned Regressors

Sometimes it is useful to separate the constant from the other regressors, and write the linear projection equation in the format

$$Y_i = X_i'\beta + \alpha + e_i$$

where α is the intercept and X_i does not contain a constant. The least squares estimates and residuals can be written as $Y_i = X_i'\widehat{\beta} + \widehat{\alpha} + \widehat{e}_i$.

In this case (3.17) can be written as the equation system

$$\begin{aligned}\sum_{i=1}^{n} \left(Y_i - X_i'\widehat{\beta} - \widehat{\alpha}\right) &= 0 \\ \sum_{i=1}^{n} X_i \left(Y_i - X_i'\widehat{\beta} - \widehat{\alpha}\right) &= 0.\end{aligned}$$

The first equation implies

$$\widehat{\alpha} = \overline{Y} - \overline{X}'\widehat{\beta}.$$

Subtracting from the second we obtain

$$\sum_{i=1}^{n} X_i \left(\left(Y_i - \overline{Y}\right) - \left(X_i - \overline{X}\right)'\widehat{\beta}\right) = 0.$$

Solving for $\widehat{\beta}$ we find

$$\begin{aligned}\widehat{\beta} &= \left(\sum_{i=1}^{n} X_i \left(X_i - \overline{X}\right)'\right)^{-1} \left(\sum_{i=1}^{n} X_i \left(Y_i - \overline{Y}\right)\right) \\ &= \left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)'\right)^{-1} \left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right).\end{aligned} \tag{3.19}$$

Thus the OLS estimator for the slope coefficients is a regression with demeaned data.

The representation (3.19) is known as the demeaned formula for the least squares estimator.

3.10 Model in Matrix Notation

For many purposes, including computation, it is convenient to write the model and statistics in matrix notation. The linear equations $Y_i = X_i'\beta + e_i$ make a system of n equations. We can stack these n equations together as

$$\begin{aligned} Y_1 &= X_1'\beta + e_1 \\ Y_2 &= X_2'\beta + e_2 \\ &\vdots \\ Y_n &= X_n'\beta + e_n. \end{aligned}$$

Define

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Observe that $\boldsymbol{Y}$ and $\boldsymbol{e}$ are $n \times 1$ vectors and $\boldsymbol{X}$ is an $n \times k$ matrix. The system of n equations can be compactly written in the single equation

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}. \tag{3.20}$$

Sample sums can be written in matrix notation. For example

$$\begin{aligned} \sum_{i=1}^{n} X_i X_i' &= \boldsymbol{X}'\boldsymbol{X} \\ \sum_{i=1}^{n} X_i Y_i &= \boldsymbol{X}'\boldsymbol{Y}. \end{aligned}$$

Therefore the least squares estimator can be written as

$$\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right).$$

The matrix version of (3.16) and estimated version of (3.20) is

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

Equivalently the residual vector is

$$\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}.$$

Using the residual vector we can write (3.17) as

$$\boldsymbol{X}'\widehat{\boldsymbol{e}} = 0.$$

It can also be useful to write the sum of squared error criterion as

$$SSE\left(\beta\right) = \left(\boldsymbol{Y} - \boldsymbol{X}\beta\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\beta\right).$$

Using matrix notation we have simple expressions for most estimators. This is particularly convenient for computer programming as most languages allow matrix notation and manipulation.

Theorem 3.2 Important Matrix Expressions

$$\begin{aligned}\widehat{\beta} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)\\ \widehat{\boldsymbol{e}} &= \boldsymbol{Y}-\boldsymbol{X}\widehat{\beta}\\ \boldsymbol{X}'\widehat{\boldsymbol{e}} &= 0.\end{aligned}$$

Early Use of Matrices

The earliest known treatment of the use of matrix methods to solve simultaneous systems is found in Chapter 8 of the Chinese text *The Nine Chapters on the Mathematical Art*, written by several generations of scholars from the 10^{th} to 2^{nd} century BCE.

3.11 Projection Matrix

Define the matrix

$$\boldsymbol{P}=\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'.$$

Observe that

$$\boldsymbol{P}\boldsymbol{X}=\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{X}.$$

This is a property of a **projection matrix**. More generally, for any matrix $\boldsymbol{Z}$ which can be written as $\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{\Gamma}$ for some matrix $\boldsymbol{\Gamma}$ (we say that $\boldsymbol{Z}$ lies in the **range space** of $\boldsymbol{X}$), then

$$\boldsymbol{P}\boldsymbol{Z}=\boldsymbol{P}\boldsymbol{X}\boldsymbol{\Gamma}=\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Gamma}=\boldsymbol{X}\boldsymbol{\Gamma}=\boldsymbol{Z}.$$

As an important example, if we partition the matrix $\boldsymbol{X}$ into two matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ so that $\boldsymbol{X}=[\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$ then $\boldsymbol{P}\boldsymbol{X}_1=\boldsymbol{X}_1$. (See Exercise 3.7.)

The projection matrix $\boldsymbol{P}$ has the algebraic property that it is **idempotent**: $\boldsymbol{P}\boldsymbol{P}=\boldsymbol{P}$. See Theorem 3.3.2 below. For the general properties of projection matrices see Section A.11.

The matrix $\boldsymbol{P}$ creates the fitted values in a least squares regression:

$$\boldsymbol{P}\boldsymbol{Y}=\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}=\boldsymbol{X}\widehat{\beta}=\widehat{\boldsymbol{Y}}.$$

Because of this property $\boldsymbol{P}$ is also known as the **hat matrix**.

A special example of a projection matrix occurs when $\boldsymbol{X}=\boldsymbol{1}_n$ is an n-vector of ones. Then

$$\boldsymbol{P}=\boldsymbol{1}_n\left(\boldsymbol{1}_n'\boldsymbol{1}_n\right)^{-1}\boldsymbol{1}_n'=\frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n'.$$

Note that in this case

$$\boldsymbol{P}\boldsymbol{y}=\boldsymbol{1}_n\left(\boldsymbol{1}_n'\boldsymbol{1}_n\right)^{-1}\boldsymbol{1}_n'\boldsymbol{Y}=\boldsymbol{1}_n\overline{Y}$$

creates an n-vector whose elements are the sample mean $\overline{Y}$.

The projection matrix $\boldsymbol{P}$ appears frequently in algebraic manipulations in least squares regression. The matrix has the following important properties.

Theorem 3.3 The projection matrix $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ for any $n \times k$ $\boldsymbol{X}$ with $n \geq k$ has the following algebraic properties

1. $\boldsymbol{P}$ is symmetric ($\boldsymbol{P}' = \boldsymbol{P}$).
2. $\boldsymbol{P}$ is idempotent ($\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$).
3. $\operatorname{tr}\boldsymbol{P} = k$.
4. The eigenvalues of $\boldsymbol{P}$ are 1 and 0. There are k eigenvalues equalling 1 and $n-k$ equalling 0.
5. $\operatorname{rank}(\boldsymbol{P}) = k$.

We close this section by proving the claims in Theorem 3.3. Part 1 holds since

$$\begin{aligned}
\boldsymbol{P}' &= \left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)' \\
&= \left(\boldsymbol{X}'\right)'\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)'(\boldsymbol{X})' \\
&= \boldsymbol{X}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)'\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{X}\left((\boldsymbol{X})'\left(\boldsymbol{X}'\right)'\right)^{-1}\boldsymbol{X}' = \boldsymbol{P}.
\end{aligned}$$

To establish part 2, the fact that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ implies that

$$\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' = \boldsymbol{P}.$$

as claimed. For part 3,

$$\operatorname{tr}\boldsymbol{P} = \operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right) = \operatorname{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\right) = \operatorname{tr}(\boldsymbol{I}_k) = k.$$

See Appendix A.5 for definition and properties of the trace operator.

For part 4, it is shown in Appendix A.11 that the eigenvalues λ_i of an idempotent matrix are all 1 and 0. Since $\operatorname{tr}\boldsymbol{P}$ equals the sum of the n eigenvalues and $\operatorname{tr}\boldsymbol{P} = k$ by part 3, it follows that there are k eigenvalues equalling 1 and the remainder $(n-k)$ equalling 0.

For part 5, observe that $\boldsymbol{P}$ is positive semi-definite since its eigenvalues are all non-negative. By Theorem A.4.5 its rank equals the number of positive eigenvalues, which is k as claimed.

3.12 Annihilator Matrix

Define

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Note that

$$\boldsymbol{M}\boldsymbol{X} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{P}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = 0. \tag{3.22}$$

Thus $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal. We call $\boldsymbol{M}$ the **annihilator matrix**, due to the property that for any matrix $\boldsymbol{Z}$ in the range space of $\boldsymbol{X}$ then

$$\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{Z} - \boldsymbol{P}\boldsymbol{Z} = 0.$$

For example, $\boldsymbol{M}\boldsymbol{X}_1 = 0$ for any subcomponent $\boldsymbol{X}_1$ of $\boldsymbol{X}$, and $\boldsymbol{M}\boldsymbol{P} = 0$ (see Exercise 3.7).

The annihilator matrix $\boldsymbol{M}$ has similar properties with $\boldsymbol{P}$, including that $\boldsymbol{M}$ is symmetric ($\boldsymbol{M}' = \boldsymbol{M}$) and idempotent ($\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$). It is thus a projection matrix. Similarly to Theorem 3.3.3 we can calculate

$$\operatorname{tr} \boldsymbol{M} = n - k. \tag{3.23}$$

(See Exercise 3.9.) One implication is that the rank of $\boldsymbol{M}$ is $n - k$.

While $\boldsymbol{P}$ creates fitted values, $\boldsymbol{M}$ creates least squares residuals:

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta} = \widehat{\boldsymbol{e}}. \tag{3.24}$$

As discussed in the previous section, a special example of a projection matrix occurs when $\boldsymbol{X} = \mathbf{1}_n$ is an n-vector of ones, so that $\boldsymbol{P} = \mathbf{1}_n \left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n'$. In this case the annihilator matrix is

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P} = \boldsymbol{I}_n - \mathbf{1}_n \left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n'.$$

While $\boldsymbol{P}$ creates a vector of sample means, $\boldsymbol{M}$ creates demeaned values:

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y} - \mathbf{1}_n\overline{Y}.$$

For simplicity we will often write the right-hand-side as $\boldsymbol{Y} - \overline{Y}$. The i^{th} element is $Y_i - \overline{Y}$, the **demeaned** value of Y_i.

We can also use (3.24) to write an alternative expression for the residual vector. Substituting $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$ into $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{Y}$ and using $\boldsymbol{M}\boldsymbol{X} = \mathbf{0}$ we find

$$\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\left(\boldsymbol{X}\beta + \boldsymbol{e}\right) = \boldsymbol{M}\boldsymbol{e} \tag{3.25}$$

which is free of dependence on the regression coefficient β.

3.13 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}\left[e_i^2\right]$ is a moment, so a natural estimator is a moment estimator. If e_i were observed we would estimate σ^2 by

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{3.26}$$

However, this is infeasible as e_i is not observed. In this case it is common to take a two-step approach to estimation. The residuals $\widehat{e}_i$ are calculated in the first step, and then we substitute $\widehat{e}_i$ for e_i in expression (3.26) to obtain the feasible estimator

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i^2. \tag{3.27}$$

In matrix notation, we can write (3.26) and (3.27) as $\widetilde{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e}$ and

$$\widehat{\sigma}^2 = n^{-1}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}. \tag{3.28}$$

Recall the expressions $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{e}$ from (3.24) and (3.25). Applied to (3.28) we find

$$\widehat{\sigma}^2 = n^{-1}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}} = n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{e} = n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} \tag{3.29}$$

the third equality since $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

An interesting implication is that

$$\widetilde{\sigma}^2 - \widehat{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e} - n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} = n^{-1}\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} \geq 0.$$

The final inequality holds because $\boldsymbol{P}$ is positive semi-definite and $\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}$ is a quadratic form. This shows that the feasible estimator $\widehat{\sigma}^2$ is numerically smaller than the idealized estimator (3.26).

3.14 Analysis of Variance

Another way of writing (3.24) is

$$\boldsymbol{Y}=\boldsymbol{P}\boldsymbol{Y}+\boldsymbol{M}\boldsymbol{Y}=\widehat{\boldsymbol{Y}}+\widehat{\boldsymbol{e}}. \tag{3.30}$$

This decomposition is **orthogonal**, that is

$$\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{e}}=(\boldsymbol{P}\boldsymbol{Y})'(\boldsymbol{M}\boldsymbol{Y})=\boldsymbol{Y}'\boldsymbol{P}\boldsymbol{M}\boldsymbol{Y}=0. \tag{3.31}$$

It follows that

$$\boldsymbol{Y}'\boldsymbol{Y}=\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}+2\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{e}}+\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}=\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}+\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n}Y_i^2=\sum_{i=1}^{n}\widehat{Y}_i^2+\sum_{i=1}^{n}\widehat{e}_i^2.$$

Subtracting $\overline{Y}$ from both sides of (3.30) we obtain

$$\boldsymbol{Y}-\mathbf{1}_n\overline{Y}=\widehat{\boldsymbol{Y}}-\mathbf{1}_n\overline{Y}+\widehat{\boldsymbol{e}}.$$

This decomposition is also orthogonal when $\boldsymbol{X}$ contains a constant, as

$$\left(\widehat{\boldsymbol{Y}}-\mathbf{1}_n\overline{Y}\right)'\widehat{\boldsymbol{e}}=\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{e}}-\overline{Y}\mathbf{1}_n'\widehat{\boldsymbol{e}}=0$$

under (3.18). It follows that

$$\left(\boldsymbol{Y}-\mathbf{1}_n\overline{Y}\right)'\left(\boldsymbol{Y}-\mathbf{1}_n\overline{Y}\right)=\left(\widehat{\boldsymbol{Y}}-\mathbf{1}_n\overline{Y}\right)'\left(\widehat{\boldsymbol{Y}}-\mathbf{1}_n\overline{Y}\right)+\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\widehat{e}_i^2.$$

This is commonly called the **analysis-of-variance** formula for least squares regression.

A commonly reported statistic is the **coefficient of determination** or **R-squared**:

$$R^2=\frac{\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}=1-\frac{\sum_{i=1}^{n}\widehat{e}_i^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}.$$

It is often described as the fraction of the sample variance of Y_i which is explained by the least squares fit. R^2 is a crude measure of regression fit. We have better measures of fit, but these require a statistical (not just algebraic) analysis and we will return to these issues later. One deficiency with R^2 is that it increases when regressors are added to a regression (see Exercise 3.16) so the “fit” can be always increased by increasing the number of regressors.

The coefficient of determination was introduced by Wright (1921).

3.15 Projections

One way to visualize least squares fitting is as a projection operation.

Write the regressor matrix as $\boldsymbol{X}=[\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots\ \boldsymbol{X}_k]$ where $\boldsymbol{X}_j$ is the j^{th} column of $\boldsymbol{X}$. The range space $\mathscr{R}(\boldsymbol{X})$ of $\boldsymbol{X}$ is the space consisting of all linear combinations of the columns $\boldsymbol{X}_1$,$\boldsymbol{X}_2$,...,$\boldsymbol{X}_k$. $\mathscr{R}(\boldsymbol{X})$ is a k

dimensional surface contained in $\mathbb{R}^n$. If $k = 2$ then $\mathscr{R}(\boldsymbol{X})$ is a plane. The operator $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ projects vectors onto $\mathscr{R}(\boldsymbol{X})$. The fitted values $\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$ are the projection of $\boldsymbol{Y}$ onto $\mathscr{R}(\boldsymbol{X})$.

To visualize examine Figure 3.3. This displays the case $n = 3$ and $k = 2$. Displayed are three vectors $\boldsymbol{Y}$, $\boldsymbol{X}_1$, and $\boldsymbol{X}_2$, which are each elements of $\mathbb{R}^3$. The plane created by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is the range space $\mathscr{R}(\boldsymbol{X})$. Regression fitted values are linear combinations of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and so lie on this plane. The fitted value $\widehat{\boldsymbol{Y}}$ is the vector on this plane closest to $\boldsymbol{Y}$. The residual $\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ is the difference between the two. The angle between the vectors $\widehat{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{e}}$ is 90°, and therefore they orthogonal as shown.

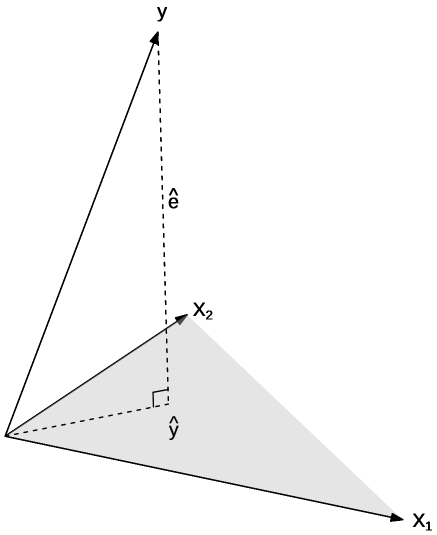


Figure 3.3: Projection of $\boldsymbol{Y}$ onto $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$

3.16 Regression Components

Partition $\boldsymbol{X} = [\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$ and $\beta = (\beta_1, \beta_2)$. The regression model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{e}. \tag{3.32}$$

The OLS estimator of $\beta = (\beta_1', \beta_2')'$ is obtained by regression of $\boldsymbol{Y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ and can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\widehat{\beta} + \widehat{\boldsymbol{e}} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{e}}. \tag{3.33}$$

We are interested in algebraic expressions for $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

Let's focus on finding an algebraic expression for $\widehat{\beta}_1$. The least squares estimator by definition is found by the joint minimization

$$\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = \underset{\beta_1, \beta_2}{\operatorname{argmin}} \operatorname{SSE}\left(\beta_1, \beta_2\right) \tag{3.34}$$

where

$$\operatorname{SSE}\left(\beta_1, \beta_2\right) = \left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1 - \boldsymbol{X}_2\beta_2\right)'\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1 - \boldsymbol{X}_2\beta_2\right).$$

An equivalent expression for $\widehat{\beta}_1$ can be obtained by concentration (nested minimization). The solution (3.34) can be written as

$$\widehat{\beta}_1 = \underset{\beta_1}{\operatorname{argmin}}\left(\min_{\beta_2} \operatorname{SSE}\left(\beta_1, \beta_2\right)\right). \tag{3.35}$$

The inner expression $\min_{\beta_2} \mathrm{SSE}\left(\beta_1, \beta_2\right)$ minimizes over β_2 while holding β_1 fixed. It is the lowest possible sum of squared errors given β_1. The outer minimization $\operatorname{argmin}_{\beta_1}$ finds the coefficient β_1 which minimizes the "lowest possible sum of squared errors given β_1". This means that $\widehat{\beta}_1$ as defined in (3.34) and (3.35) are algebraically identical.

Examine the inner minimization problem in (3.35). This is simply the least squares regression of $\boldsymbol{Y}-\boldsymbol{X}_1\beta_1$ on $\boldsymbol{X}_2$. This has solution

$$\underset{\beta_2}{\operatorname{argmin}} \mathrm{SSE}\left(\beta_1, \beta_2\right)=\left(\boldsymbol{X}_2^{\prime} \boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2^{\prime}\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right)\right)$$

with residuals

$$\begin{aligned}
\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1-\boldsymbol{X}_2\left(\boldsymbol{X}_2^{\prime} \boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2^{\prime}\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right)\right) & =\left(\boldsymbol{M}_2 \boldsymbol{Y}-\boldsymbol{M}_2 \boldsymbol{X}_1 \beta_1\right) \\
& =\boldsymbol{M}_2\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right)
\end{aligned}$$

where

$$\boldsymbol{M}_2=\boldsymbol{I}_n-\boldsymbol{X}_2\left(\boldsymbol{X}_2^{\prime} \boldsymbol{X}_2\right)^{-1} \boldsymbol{X}_2^{\prime} \tag{3.36}$$

is the orthogonal projection matrix for $\boldsymbol{X}_2$. This means that the inner minimization problem (3.35) has minimized value

$$\begin{aligned}
\min_{\beta_2} \mathrm{SSE}\left(\beta_1, \beta_2\right) & =\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right)^{\prime} \boldsymbol{M}_2 \boldsymbol{M}_2\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right) \\
& =\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right)^{\prime} \boldsymbol{M}_2\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right)
\end{aligned}$$

where the second equality holds since $\boldsymbol{M}_2$ is idempotent. Substituting this into (3.35) we find

$$\begin{aligned}
\widehat{\beta}_1 & =\underset{\beta_1}{\operatorname{argmin}}\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right)^{\prime} \boldsymbol{M}_2\left(\boldsymbol{Y}-\boldsymbol{X}_1 \beta_1\right) \\
& =\left(\boldsymbol{X}_1^{\prime} \boldsymbol{M}_2 \boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1^{\prime} \boldsymbol{M}_2 \boldsymbol{Y}\right) .
\end{aligned}$$

By a similar argument we find

$$\widehat{\beta}_2=\left(\boldsymbol{X}_2^{\prime} \boldsymbol{M}_1 \boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2^{\prime} \boldsymbol{M}_1 \boldsymbol{Y}\right)$$

where

$$\boldsymbol{M}_1=\boldsymbol{I}_n-\boldsymbol{X}_1\left(\boldsymbol{X}_1^{\prime} \boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1^{\prime} \tag{3.37}$$

is the orthogonal projection matrix for $\boldsymbol{X}_1$.

Theorem 3.4 The least squares estimator $\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ for (3.33) has the algebraic solution

$$\widehat{\beta}_1=\left(\boldsymbol{X}_1^{\prime} \boldsymbol{M}_2 \boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1^{\prime} \boldsymbol{M}_2 \boldsymbol{Y}\right) \tag{3.38}$$

$$\widehat{\beta}_2=\left(\boldsymbol{X}_2^{\prime} \boldsymbol{M}_1 \boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2^{\prime} \boldsymbol{M}_1 \boldsymbol{Y}\right) \tag{3.39}$$

where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are defined in (3.37) and (3.36), respectively.

3.17 Regression Components (Alternative Derivation)*

An alternative proof of Theorem 3.4 uses an algebraic argument based on the population argument presented in Section 2.22. Since this is a classic derivation we present it here for completeness.

Partition $\widehat{\boldsymbol{Q}}_{XX}$ as

$$\widehat{\boldsymbol{Q}}_{XX} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2 \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}$$

and similarly $\widehat{\boldsymbol{Q}}_{XY}$ as

$$\widehat{\boldsymbol{Q}}_{XY} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1Y} \\ \widehat{\boldsymbol{Q}}_{2Y} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{Y} \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{Y} \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.3)

$$\widehat{\boldsymbol{Q}}_{XX}^{-1} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix}^{-1} \overset{\text{def}}{=} \begin{bmatrix} \widehat{\boldsymbol{Q}}^{11} & \widehat{\boldsymbol{Q}}^{12} \\ \widehat{\boldsymbol{Q}}^{21} & \widehat{\boldsymbol{Q}}^{22} \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1} \end{bmatrix} \tag{3.40}$$

where $\widehat{\boldsymbol{Q}}_{11\cdot 2} = \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}$ and $\widehat{\boldsymbol{Q}}_{22\cdot 1} = \widehat{\boldsymbol{Q}}_{22} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{12}$. Thus

$$\begin{aligned} \widehat{\beta} &= \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} \\ &= \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1Y} \\ \widehat{\boldsymbol{Q}}_{2Y} \end{bmatrix} \\ &= \begin{pmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{1Y\cdot 2} \\ \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{2Y\cdot 1} \end{pmatrix}. \end{aligned}$$

Now

$$\begin{aligned} \widehat{\boldsymbol{Q}}_{11\cdot 2} &= \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1 \end{aligned}$$

and

$$\begin{aligned} \widehat{\boldsymbol{Q}}_{1y\cdot 2} &= \widehat{\boldsymbol{Q}}_{1Y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{2Y} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{Y} - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{Y} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{Y}. \end{aligned}$$

Equation (3.39) follows.

Similarly to the calculation for $\widehat{\boldsymbol{Q}}_{11\cdot 2}$ and $\widehat{\boldsymbol{Q}}_{1Y\cdot 2}$ you can show that $\widehat{\boldsymbol{Q}}_{2Y\cdot 1} = \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{Y}$ and $\widehat{\boldsymbol{Q}}_{22\cdot 1} = \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2$. This establishes (3.38). Together, this is Theorem 3.4.

3.18 Residual Regression

As first recognized by Frisch and Waugh (1933) and extended by Lovell (1963), expressions (3.38) and (3.39) can be used to show that the least squares estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ can be found by a two-step regression procedure.

Take (3.39). Since $\boldsymbol{M}_1$ is idempotent, $\boldsymbol{M}_1 = \boldsymbol{M}_1\boldsymbol{M}_1$ and thus

$$\begin{aligned}\widehat{\beta}_2 &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{Y}\right)\\ &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{Y}\right)\\ &= \left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{e}}_1\right)\end{aligned}$$

where $\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$ and $\widetilde{\boldsymbol{e}}_1 = \boldsymbol{M}_1\boldsymbol{Y}$.

Thus the coefficient estimator $\widehat{\beta}_2$ is algebraically equal to the least squares regression of $\widetilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$. Notice that these two are $\boldsymbol{Y}$ and $\boldsymbol{X}_2$, respectively, premultiplied by $\boldsymbol{M}_1$. But we know that multiplication by $\boldsymbol{M}_1$ is equivalent to creating least squares residuals. Therefore $\widetilde{\boldsymbol{e}}_1$ is simply the least squares residual from a regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1$, and the columns of $\widetilde{\boldsymbol{X}}_2$ are the least squares residuals from the regressions of the columns of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.

We have proven the following theorem.

Theorem 3.5 Frisch-Waugh-Lovell (FWL)
In the model (3.32), the OLS estimator of β_2 and the OLS residuals $\widehat{\boldsymbol{e}}$ may be computed by either the OLS regression (3.33) or via the following algorithm:

1. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{e}}_1$;
2. Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{X}}_2$;
3. Regress $\widetilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$, obtain OLS estimates $\widehat{\beta}_2$ and residuals $\widehat{\boldsymbol{e}}$.

In some contexts (such as panel data models, to be introduced in Chapter 17), the FWL theorem can be used to greatly speed computation.

The FWL theorem is a direct analog of the coefficient representation obtained in Section 2.23. The result obtained in that section concerned the population projection coefficients; the result obtained here concern the least squares estimators. The key message is the same. In the least squares regression (3.33) the estimated coefficient $\widehat{\beta}_2$ algebraically equals the regression of $\boldsymbol{Y}$ on the regressors $\boldsymbol{X}_2$ after the regressors $\boldsymbol{X}_1$ have been linearly projected out. Similarly, the coefficient estimate $\widehat{\beta}_1$ algebraically equals the regression of $\boldsymbol{Y}$ on the regressors $\boldsymbol{X}_1$ after the regressors $\boldsymbol{X}_2$ have been linearly projected out. This result can be insightful when interpreting regression coefficients.

A common application of the FWL theorem is the demeaning formula for regression obtained in (3.19). Partition $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ where $\boldsymbol{X}_1 = \mathbf{1}_n$ is a vector of ones and $\boldsymbol{X}_2$ is a matrix of observed regressors. In this case $\boldsymbol{M}_1 = \boldsymbol{I}_n - \mathbf{1}_n\left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n'$. Observe that $\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2 = \boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2$ and $\boldsymbol{M}_1\boldsymbol{Y} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ are the "demeaned" variables. The FWL theorem says that $\widehat{\beta}_2$ is the OLS estimate from a regression of $Y_i - \overline{Y}$ on $X_{2i} - \overline{X}_2$:

$$\widehat{\beta}_2 = \left(\sum_{i=1}^n \left(X_{2i} - \overline{X}_2\right)\left(X_{2i} - \overline{X}_2\right)'\right)^{-1}\left(\sum_{i=1}^n \left(X_{2i} - \overline{X}_2\right)\left(Y_i - \overline{Y}\right)\right).$$

This is (3.19).

Ragnar Frisch

Ragnar Frisch (1895-1973) was co-winner with Jan Tinbergen of the first Nobel Memorial Prize in Economic Sciences in 1969 for their work in developing and applying dynamic models for the analysis of economic problems. Frisch made a number of foundational contributions to modern economics beyond the Frisch-Waugh-Lovell Theorem, including formalizing consumer theory, production theory, and business cycle theory.

3.19 Leverage Values

The **leverage** values for the regressor matrix $\boldsymbol{X}$ are the diagonal elements of the projection matrix $\boldsymbol{P}=\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. There are n leverage values, and are typically written as h_{ii} for $i=1,...,n$. Since

$$\boldsymbol{P}=\left(\begin{array}{c} X_1' \\ X_2' \\ \vdots \\ X_n' \end{array}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\begin{array}{cccc} X_1 & X_2 & \cdots & X_n \end{array}\right)$$

they are

$$h_{ii}=X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}X_i. \tag{3.41}$$

The leverage value h_{ii} is a normalized length of the observed regressor vector X_i. They appear frequently in the algebraic and statistical analysis of least squares regression, including leave-one-out regression, influential observations, robust covariance matrix estimation, and cross-validation.

A few properties of the leverage values are now listed.

Theorem 3.6

1. $0\le h_{ii}\le 1$.
2. $h_{ii}\ge 1/n$ if X includes an intercept.
3. $\sum_{i=1}^n h_{ii}=k$.

We prove Theorem 3.6 below.

The leverage value h_{ii} measures how unusual the i^{th} observation X_i is relative to the other observations in the sample. A large h_{ii} occurs when X_i is quite different from the other sample values. A measure of overall unusualness is the maximum leverage value

$$\overline{h}=\max_{1\le i\le n} h_{ii}. \tag{3.42}$$

It is common to say that a regression design is **balanced** when the leverage values are all roughly equal to one another. From Theorem 3.6.3 we can deduce that complete balance implies $h_{ii}=\overline{h}=k/n$. An example where complete balance occurs is when the regressors are all orthogonal dummy variables, each of which have equal occurrance of 0's and 1's.

A regression design is **unbalanced** if some leverage values are highly unequal from the others. The most extreme case is $\overline{h}=1$. An example where this occurs is when there is a dummy regressor which takes the value 1 for only one observation in the sample.

The maximal leverage value (3.42) will change depending on the choice of regressors. For example, consider equation (3.14), the wage regression for single Asian men which has $n=268$ observations. This regression has $\overline{h}=0.33$. If the squared experience regressor is omitted, the leverage drops to $\overline{h}=0.10$. If a cubic in experience is added, it increases to $\overline{h}=0.76$. And if a fourth and fifth power are added, it increases to $\overline{h}=0.99$.

Some inference procedures (such as robust covariance matrix estimation and cross-validation) are sensitive to high leverage values. We will return to these issues later.

We now prove Theorem 3.6. For part 1 let s_i be an $n\times 1$ unit vector with a 1 in the i^{th} place and zeros elsewhere so that $h_{ii}=s_i'\boldsymbol{P}s_i$. Then applying the Quadratic Inequality (B.18) and Theorem 3.3.4,

$$h_{ii}=s_i'\boldsymbol{P}s_i\leq s_i's_i\lambda_{\max}(\boldsymbol{P})=1$$

as claimed.

For part 2 partition $X_i=(1,Z_i')'$. Without loss of generality we can replace Z_i with the demeaned values $Z_i^*=Z_i-\overline{Z}$. Then since Z_i^* and the intercept are orthgonal

$$\begin{aligned} h_{ii}&=(1,Z_i^{*\prime})\left[\begin{array}{cc} n & 0\\ 0 & \boldsymbol{Z}^{*\prime}\boldsymbol{Z}^* \end{array}\right]^{-1}\left(\begin{array}{c} 1\\ Z_i^* \end{array}\right)\\ &=\frac{1}{n}+Z_i^{*\prime}\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1}Z_i^*\geq\frac{1}{n}. \end{aligned}$$

For part 3, $\sum_{i=1}^n h_{ii}=\operatorname{tr}\boldsymbol{P}=k$ where the second equality is Theorem 3.3.3.

3.20 Leave-One-Out Regression

There are a number of statistical procedures – residual analysis, jackknife variance estimation, cross-validation, two-step estimation, hold-out sample evaluation – which make use of estimators constructed on sub-samples. Of particular importance is the case where we exclude a single observation and then repeat this for all observations. This is called **leave-one-out** (LOO) regression.

Specifically, the leave-one-out estimator of the regression coefficient β is the least squares estimator constructed using the full sample excluding a single observation i. This can be written as

$$\begin{aligned} \widehat{\beta}_{(-i)}&=\left(\sum_{j\neq i}X_jX_j'\right)^{-1}\left(\sum_{j\neq i}X_jY_j\right)\\ &=\left(\boldsymbol{X}'\boldsymbol{X}-X_iX_i'\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}-X_iY_i\right)\\ &=\left(\boldsymbol{X}_{(-i)}'\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}'\boldsymbol{Y}_{(-i)}. \end{aligned} \tag{3.43}$$

Here, $\boldsymbol{X}_{(-i)}$ and $\boldsymbol{Y}_{(-i)}$ are the data matrices omitting the i^{th} row. The notation $\widehat{\beta}_{(-i)}$ or $\widehat{\beta}_{-i}$ is commonly used to denote an estimator with the i^{th} observation omitted. There is a leave-one-out estimator for each observation, $i=1,\ldots,n$, so we have n such estimators.

The leave-one-out predicted value for Y_i is $\widetilde{Y}_i=X_i'\widehat{\beta}_{(-i)}$. This is the predicted value obtained by estimating β on the sample without observation i and then using the covariate vector X_i to predict Y_i. Notice that $\widetilde{Y}_i$ is an authentic prediction as Y_i is not used to construct $\widetilde{Y}_i$. This is in contrast to the fitted values $\widehat{Y}_i$ which are functions of Y_i.

The **leave-one-out residual**, **prediction error**, or **prediction residual** is $\widetilde{e}_i = Y_i - \widetilde{Y}_i$. The prediction errors may be used as estimates of the errors instead of the residuals. The prediction errors are better estimates than the residuals since the former are based on authentic predictions.

The leave-one-out formula (3.43) gives the unfortunate impression that the leave-one-out coefficients and errors are computationally cumbersome, requiring n separate regressions. In the context of linear regression this is fortunately not the case. There are simple linear expressions for $\widehat{\beta}_{(-i)}$ and $\widetilde{e}_i$.

Theorem 3.7 The leave-one-out estimator and prediction error equal

$$\widehat{\beta}_{(-i)} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i \tag{3.44}$$

and

$$\widetilde{e}_i = (1 - h_{ii})^{-1} \widehat{e}_i \tag{3.45}$$

where h_{ii} are the leverage values as defined in (3.41).

We prove Theorem 3.7 at the end of the section.

Equation (3.44) shows that the leave-one-out coefficients can be calculated by a simple linear operation and do not need to be calculated using n separate regressions. Equation (3.45) for the prediction error is particularly convenient. It shows that the leave-one-out residuals are a simple scaling of the least squares residuals.

Another interesting feature of equation (3.45) is that the prediction errors $\widetilde{e}_i$ are a simple scaling of the residuals $\widehat{e}_i$ with the scaling dependent on the leverage values h_{ii}. If h_{ii} is small then $\widetilde{e}_i \simeq \widehat{e}_i$. However if h_{ii} is large then $\widetilde{e}_i$ can be quite different from $\widehat{e}_i$. Thus the difference between the residuals and predicted values depends on the leverage values, that is, how unusual is X_i.

To write (3.45) in vector notation, define

$$\begin{aligned}\boldsymbol{M}^* &= \left(\boldsymbol{I}_n - \operatorname{diag}\{h_{11}, .., h_{nn}\}\right)^{-1} \\ &= \operatorname{diag}\{(1-h_{11})^{-1}, .., (1-h_{nn})^{-1}\}.\end{aligned}$$

Then (3.45) is equivalent to

$$\widetilde{\boldsymbol{e}} = \boldsymbol{M}^* \widehat{\boldsymbol{e}}. \tag{3.46}$$

One use of the prediction errors is to estimate the out-of-sample mean squared error:

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widetilde{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n} (1-h_{ii})^{-2} \widehat{e}_i^2. \tag{3.47}$$

This is known as the **sample mean squared prediction error**. Its square root $\widetilde{\sigma} = \sqrt{\widetilde{\sigma}^2}$ is the **prediction standard error**.

We complete the section with a proof of Theorem 3.7. The leave-one-out estimator (3.43) can be written as

$$\widehat{\beta}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X} - X_i X_i'\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{Y} - X_i Y_i\right). \tag{3.48}$$

Multiply (3.48) by $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X} - X_i X_i'\right)$. We obtain

$$\widehat{\beta}_{(-i)} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i X_i' \widehat{\beta}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y} - X_i Y_i\right) = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i Y_i.$$

Rewriting

$$\widehat{\beta}_{(-i)} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \left(Y_i - X_i'\widehat{\beta}_{(-i)}\right) = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i$$

which is (3.44). Premultiplying this expression by X_i' and using definition (3.41) we obtain

$$X_i'\widehat{\beta}_{(-i)} = X_i'\widehat{\beta} - X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i\widetilde{e}_i = X_i'\widehat{\beta} - h_{ii}\widetilde{e}_i.$$

Using the definitions for $\widehat{e}_i$ and $\widetilde{e}_i$ we obtain $\widetilde{e}_i = \widehat{e}_i + h_{ii}\widetilde{e}_i$. Re-writing we obtain (3.45).

3.21 Influential Observations

Another use of the leave-one-out estimator is to investigate the impact of **influential observations**, sometimes called **outliers**. We say that observation i is influential if its omission from the sample induces a substantial change in a parameter estimate of interest.

For illustration consider Figure 3.4 which shows a scatter plot of realizations (Y_i, X_i). The 25 observations shown with the open circles are generated by $X_i \sim U[1,10]$ and $Y_i \sim N(X_i, 4)$. The 26^{th} observation shown with the filled circle is $X_{26} = 9$, $Y_{26} = 0$. (Imagine that $Y_{26} = 0$ was incorrectly recorded due to a mistaken key entry.) The figure shows both the least squares fitted line from the full sample and that obtained after deletion of the 26^{th} observation from the sample. In this example we can see how the 26^{th} observation (the "outlier") greatly tilts the least squares fitted line towards the 26^{th} observation. In fact, the slope coefficient decreases from 0.97 (which is close to the true value of 1.00) to 0.56, which is substantially reduced. Neither Y_{26} nor X_{26} are unusual values relative to their marginal distributions so this outlier would not have been detected from examination of the marginal distributions of the data. The change in the slope coefficient of -0.41 is meaningful and should raise concern to an applied economist.

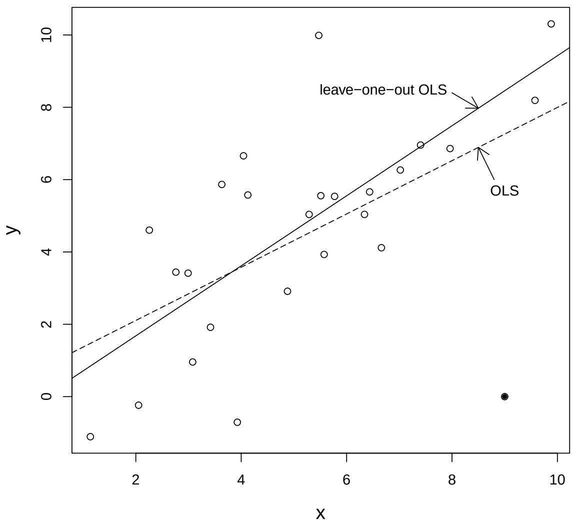


Figure 3.4: Impact of an Influential Observation on the Least-Squares Estimator

From (3.44) we know that

$$\widehat{\beta} - \widehat{\beta}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i\widetilde{e}_i. \tag{3.49}$$

By direct calculation of this quantity for each observation i, we can directly discover if a specific observation i is influential for a coefficient estimate of interest.

For a general assessment, we can focus on the predicted values. The difference between the full-sample and leave-one-out predicted values is

$$\widehat{Y}_i - \widetilde{Y}_i = X_i'\widehat{\beta} - X_i'\widehat{\beta}_{(-i)} = X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i\widetilde{e}_i = h_{ii}\widetilde{e}_i$$

which is a simple function of the leverage values h_{ii} and prediction errors $\widetilde{e}_i$. Observation i is influential for the predicted value if $|h_{ii}\widetilde{e}_i|$ is large, which requires that both h_{ii} and $|\widetilde{e}_i|$ are large.

One way to think about this is that a large leverage value h_{ii} gives the potential for observation i to be influential. A large h_{ii} means that observation i is unusual in the sense that the regressor X_i is far from its sample mean. We call an observation with large h_{ii} a **leverage point**. A leverage point is not necessarily influential as the latter also requires that the prediction error $\widetilde{e}_i$ is large.

To determine if any individual observations are influential in this sense several diagnostics have been proposed (some names include DFITS, Cook's Distance, and Welsch Distance). Unfortunately, from a statistical perspective it is difficult to recommend these diagnostics for applications as they are not based on statistical theory. Probably the most relevant measure is the change in the coefficient estimates given in (3.49). The ratio of these changes to the coefficient's standard error is called its DFBETA, and is a postestimation diagnostic available in Stata. While there is no magic threshold, the concern is whether or not an individual observation meaningfully changes an estimated coefficient of interest. A simple diagnostic for influential observations is to calculate

$$\text{Influence} = \max_{1\le i\le n}\left|\widehat{Y}_i - \widetilde{Y}_i\right| = \max_{1\le i\le n}|h_{ii}\widetilde{e}_i|.$$

This is the largest (absolute) change in the predicted value due to a single observation. If this diagnostic is large relative to the distribution of Y_i it may indicate that that observation is influential.

If an observation is determined to be influential what should be done? As a common cause of influential observations is data entry error, the influential observations should be examined for evidence that the observation was mis-recorded. Perhaps the observation falls outside of permitted ranges, or some observables are inconsistent (for example, a person is listed as having a job but receives earnings of $0). If it is determined that an observation is incorrectly recorded, then the observation is typically deleted from the sample. This process is often called "cleaning the data". The decisions made in this process involve a fair amount of individual judgment. [When this is done the proper practice is to retain the source data in its original form and create a program file which executes all cleaning operations (for example deletion of individual observations). The cleaned data file can be saved at this point, and then used for the subsequent statistical analysis. The point of retaining the source data and a specific program file which cleans the data is twofold: so that all decisions are documented, and so that modifications can be made in revisions and future research.] It is also possible that an observation is correctly measured, but unusual and influential. In this case it is unclear how to proceed. Some researchers will try to alter the specification to properly model the influential observation. Other researchers will delete the observation from the sample. The motivation for this choice is to prevent the results from being skewed or determined by individual observations. This latter practice is viewed skeptically by many researchers who believe it reduces the integrity of reported empirical results.

For an empirical illustration consider the log wage regression (3.14) for single Asian men. This regression, which has 268 observations, has Influence = 0.29. This means that the most influential observation, when deleted, changes the predicted (fitted) value of the dependent variable log(*wage*) by 0.29, or equivalently the average wage by 29%. This is a meaningful change and suggests further investigation. We examine the influential observation, and find that its leverage h_{ii} is 0.33, which is the maximum in the sample as described in Section 3.19. It is a rather large leverage value, meaning that the regressor X_i is unusual. Examining further, we find that this individual is 65 years old with 8 years education, so that his potential work experience is 51 years. This is the highest experience in the subsample – the next highest is 41 years. The large leverage is due to his unusual characteristics (very low education and very high experience) within this sample. Essentially, regression (3.14) is attempting to estimate the conditional mean at *experience*= 51 with only one observation. It is not surprising that this observation determines the fit and is thus influential. A reasonable conclusion is the regression function can only be estimated over a smaller range of *experience*. We restrict the sample to individuals with less than 45 years experience, re-estimate, and obtain the following estimates.

$$\widehat{\log(wage)} = 0.144\ education + 0.043\ experience - 0.095\ experience^2/100 + 0.531. \tag{3.50}$$

For this regression, we calculate that Influence = 0.11, which is greatly reduced relative to the regression (3.14). Comparing (3.50) with (3.14), the slope coefficient for education is essentially unchanged, but the coefficients in experience and its square have slightly increased.

By eliminating the influential observation equation (3.50) can be viewed as a more robust estimate of the conditional mean for most levels of *experience*. Whether to report (3.14) or (3.50) in an application is largely a matter of judgment.

3.22 CPS Data Set

In this section we describe the data set used in the empirical illustrations.

The Current Population Survey (CPS) is a monthly survey of about 57,000 U.S. households conducted by the Bureau of the Census of the Bureau of Labor Statistics. The CPS is the primary source of information on the labor force characteristics of the U.S. population. The survey covers employment, earnings, educational attainment, income, poverty, health insurance coverage, job experience, voting and registration, computer usage, veteran status, and other variables. Details can be found at www.census.gov/cps and dataferrett.census.gov.

From the March 2009 survey we extracted the individuals with non-allocated variables who were full-time employed (defined as those who had worked at least 36 hours per week for at least 48 weeks the past year), and excluded those in the military. This sample has 50,742 individuals. We extracted 14 variables from the CPS on these individuals and created the data set `cps09mar`. This data set, and all others used in this textbook, are available at http://www.ssc.wisc.edu/~bhansen/econometrics/

3.23 Numerical Computation

Modern econometric estimation involves large samples and many covariates. Consequently calculation of even simple statistics such as the least squares estimator requires a large number (millions) of arithmetic operations. In practice most economists don't need to think much about this as it is done swiftly and effortlessly on our personal computers. Nevertheless it is useful to understand the underlying calculation methods as occasionally choices can make substantive differences.

While today nearly all statistical computations are made using statistical software running on personal computers, this was not always the case. In the nineteenth and early twentieth centures, "computer" was a job label for workers who made computations by hand. Computers were employed by astronomers and statistical laboratories to execute numerical calculations. This fascinating job (and the fact that most computers employed in laboratories were women) has entered popular culture. For example the lives of several computers who worked for the early U.S. space program is described in the book and popular movie *Hidden Figures*, a fictional computer/astronaut is the protagonist of the novel *The Calculating Stars*, and the life of computer/astronomer Henrietta Swan Leavitt is dramatized in the play *Silent Sky*.

Until programmable electronic computers became available in the 1960s economics graduate students were routinely employed as computers. Sample sizes were considerably smaller than those seen today, but still the effort required to calculate by hand (for example) a regression with $n = 100$ observations and $k = 5$ variables is considerable! If you are a current graduate student you should feel fortunate that the profession has moved on from the era of human computers! (Now research assistants do more elevated tasks such as writing Stata and Matlab code.)

To obtain the least squares estimator $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)$ we need to either invert $\boldsymbol{X}'\boldsymbol{X}$ or solve a system of equations. To be specific, let $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{c} = \boldsymbol{X}'\boldsymbol{Y}$ so that the least squares estimator can be written as either the solution to

$$\boldsymbol{A}\widehat{\beta} = \boldsymbol{c} \tag{3.51}$$

or as

$$\widehat{\beta} = \boldsymbol{A}^{-1}\boldsymbol{c}. \tag{3.52}$$

The equations (3.51) and (3.52) are algebraically identical but they suggest two distinct numerical approaches to obtain $\widehat{\beta}$. (3.51) suggests solving a system of k equations. (3.52) suggests finding $\boldsymbol{A}^{-1}$ and then multiplying by $\boldsymbol{c}$. While the two expressions are algebraically identical the implied numerical approaches are different.

In a nutshell, solving the system of equations (3.51) is numerically preferred to the matrix inversion problem (3.52). Directly solving (3.51) is faster and produces a solution with a higher degree of numerical accuracy. Thus (3.51) is generally recommended over (3.52). However, in most practical applications the choice will not make any practical difference. Contexts where the choice may make a difference is when the matrix $\boldsymbol{A}$ is ill-conditioned (to be discussed in Section 3.24) or of extremely high dimension.

Numerical methods to solve the system of equations (3.51) and calculate $\boldsymbol{A}^{-1}$ are discussed in Sections A.18 and A.19, respectively.

Statistical packages use a variety of matrix methods to solve (3.51). Stata uses the sweep algorithm which is a variant of the Gauss-Jordan algorithm discussed in Section A.18. (For the sweep algorithm see Goodnight (1979).) In R, `solve(A,b)` uses the QR decomposition. In Matlab, `A\b` uses the Cholesky decomposition when $\boldsymbol{A}$ is positive definite and the QR decomposition otherwise.

3.24 Collinearity Errors

For the least squares estimator to be uniquely defined the regressors cannot be linearly dependent. However, it is quite easy to *attempt* to calculate a regression with linearly dependent regressors. This can occur for many reasons, including the following.

1. Including the same regressor twice.

2. Including regressors which are a linear combination of one another, such as *education*, *experience* and *age* in the CPS data set example (recall, *experience* is defined as *age-education-6*).

3. Including a dummy variable and its square.

4. Estimating a regression on a sub-sample for which a dummy variable is either all zeros or all ones.

5. Including a dummy variable interaction which yields all zeros.

6. Including more regressors than observations.

In any of the above cases the regressors are linearly dependent so $\boldsymbol{X}'\boldsymbol{X}$ is singular and the least squares estimator is not uniquely defined. If you attempt to estimate the regression, you are likely to encounter an error message. (A possible exception is Matlab using "`A\b`", as discussed below.) The message may be that "system is exactly singular", "system is computationally singular", a variable is "omitted because of collinearity", or a coefficient is listed as "NA". In some cases (such as estimation in R using explicit matrix computation or Matlab using the `regress` command) the program will stop execution. In other cases the program will continue to run. In Stata (and in the `lm` package in R), a regression will be reported but one or more variables will be omitted to achieve non-singularity.

If any of these warnings or error messages appear, the correct response is to stop and examine the regression coding and data. Did you make an unintended mistake? Have you included a linearly dependent regressor? Are you estimating on a subsample for which the variables (in particular dummy variables) have no variation? If you can determine that one of these scenarios caused the error, the solution is immediately apparent. You need to respecify your model (either sample or regressors) so that the redundancy is eliminated. All empirical researchers encounter this error in the course of empirical work. You should not, however, simply accept output if the package has selected variables for omission. It is the researcher's job to understand the underlying cause and enact a suitable remedy.

There is also a possibility that the statistical package will not detect and report the matrix singularity. If you compute in Matlab using explicit matrix operations and use the recommended `A\b` command to compute the least squares estimator Matlab may return a numerical solution without an error message even when the regressors are algebraically dependent. It is therefore recommended that you perform a numerical check for matrix singularity when using explicit matrix operations in Matlab.

How can we numerically check if a matrix $\boldsymbol{A}$ is singular? A standard diagnostic is the **reciprocal condition number**

$$C = \frac{\lambda_{\min}(\boldsymbol{A})}{\lambda_{\max}(\boldsymbol{A})}.$$

If $C = 0$ then $\boldsymbol{A}$ is singular. If $C = 1$ then $\boldsymbol{A}$ is perfectly balanced. If C is extremely small we say that $\boldsymbol{A}$ is **ill-conditioned**. The reciprocal condition number can be calculated in Matlab or R by the `rcond` command. Unfortunately, there is no accepted tolerance for how small C should be before regarding $\boldsymbol{A}$ as numerically singular, in part since `rcond(A)` can return a positive (but small) result even if $\boldsymbol{A}$ is algebraically singular. However, in double precision (which is typically used for computation) numerical accuracy is bounded by $2^{-52} \simeq$ 2e-16, suggesting the minimum bound $C \geq$ 2e-16.

Checking for numerical singularity is complicated by the fact that low values of C can also be caused by unbalanced or highly correlated regressors.

To illustrate, consider a wage regression using the sample from (3.14) on powers of experience X from 1 through k (e.g. $X, X^2, X^3, ..., X^k$). We calculated the reciprocal condition number C for each k, and found that C is decreasing as k increases, indicating increasing ill-conditioning. Indeed, for $k = 5$, we find $C =$ 6e-17, which is lower than double precision accuracy. This means that a regression on (X, X^2, X^3, X^4, X^5) is ill-conditioned. The regressor matrix, however, is not singular. The low value of C is not due to algebraic singularity but rather is due to a lack of balance and high collinearity.

Ill-conditioned regressors have the potential problem that the numerical results (the reported coefficient estimates) will be inaccurate. It is not a major concern as this only occurs in extreme cases and because high numerical accuracy is not typically an important goal in econometric estimation. Nevertheless, we should try and avoid ill-conditioned regressions whenever possible.

There are strategies which can reduce or even eliminate ill-conditioning. Often it is sufficient to rescale the regressors. A simple rescaling which often works for non-negative regressors is to divide each by its sample mean, thus replace X_{ji} with $X_{ji}/\overline{X}_j$. In the above example with the powers of experience, this means replacing X_i^2 with $X_i^2/\left(n^{-1}\sum_{i=1}^n X_i^2\right)$, etc. Doing so dramatically reduces the ill-conditioning. With this scaling regressions for $k \leq 11$ satisfy $C \geq$ 1e-15. Another rescaling specific to a regression with powers is to first rescale the regressor to lie in $[-1,1]$ before taking powers. With this scaling, regressions for $k \leq 16$ satisfy $C \geq$ 1e-15. A simpler scaling option is to rescale the regressor to lie in $[0,1]$ before taking powers. With this scaling, regressions for $k \leq 9$ satisfy $C \geq$ 1e-15. This is often sufficient for applications.

Ill-conditioning can often be completely eliminated by orthogonalization of the regressors. This is achieved by sequentially regressing each variable (each column in $\boldsymbol{X}$) on the preceeding variables (each preceeding column), taking the residual, and then rescaling to have a unit variance. This will produce regressors which algebraically satisfy $\boldsymbol{X}'\boldsymbol{X} = n\boldsymbol{I}_n$ and have a condition number of $C = 1$. If we apply this method to the above example, we obtain a condition number close to 1 for $k \leq 20$.

What this shows is that when a regression has a small condition number it is important to examine the specification carefully. It is possible that the regressors are linearly dependent in which case one or more regressors will need to be omitted. It is also possible that the regressors are badly scaled in which case it may be useful to rescale some of the regressors. It is also possible that the variables are highly collinear in which case a possible solution is orthogonalization. These choices should be made by the researcher not by an automated software program.

3.25 Programming

Most packages allow both interactive programming (where you enter commands one-by-one) and batch programming (where you run a pre-written sequence of commands from a file). Interactive programming can be useful for exploratory analysis but eventually all work should be executed in batch mode. This is the best way to control and document your work.

Batch programs are text files where each line executes a single command. For Stata, this file needs to have the filename extension ".do", and for MATLAB ".m". For R there is no specific naming requirements, though it is typical to use the extension ".r". When writing batch files it is useful to include comments for documentation and readability. To execute a program file you type a command within the program.

Stata: `do chapter3` executes the file *chapter3.do*

MATLAB: `run chapter3` executes the file *chapter3.m*

R: `source(“chapter3.r”)` or `source(‘chapter3.r’)` executes the file *chapter3.r*

There are similarities and differences between the commands used in these packages. For example:

1. Different symbols are used to create comments. `*` in Stata, `#` in R, and `%` in Matlab.

2. Matlab uses the symbol `;` to separate lines. Stata and R use a hard return.

3. Stata uses `ln()` to compute natural logarithms. R and Matlab use `log()`.

4. The symbol = is used to define a variable. R prefers `<-`. Double equality == is used to test equality.

We now illustrate programming files for Stata, R, and MATLAB, which execute a portion of the empirical illustrations from Sections 3.7 and 3.21. For the R and Matlab code we illustrate using explicit matrix operations. Alternatively, R and Matlab have built-in functions which implement least squares regression without the need for explicit matrix operations. In R the standard function is `lm`. In Matlab the standard function is `regress`. The advantage of using explicit matrix operations as shown below is that you know exactly what computations are done and it is easier to go "out of the box" to execute new procedures. The advantage of using built-in functions is that coding is simplified and you are much less likely to make a coding error.

Stata do File

```
*       Clear memory and load the data
clear
use cps09mar.dta
*       Generate transformations
gen wage = ln(earnings/(hours*week))
gen experience = age - education - 6
gen exp2 = (experience^2)/100
*       Create indicator for subsamples
gen mbf = (race == 2) & (marital <= 2) & (female == 1)
gen mbf12 = (mbf == 1) & (experience == 12)
gen sam = (race == 4) & (marital == 7) & (female == 0)
*        Regressions
reg wage education if mbf12 == 1
reg wage education experience exp2 if sam == 1
*       Leverage and influence
predict leverage, hat
predict e, residual
gen d=e*leverage/(1-leverage)
summarize d if sam ==1
```

R Program File

```
#       Load the data and create subsamples
dat <- read.table("cps09mar.txt")
experience <- dat[,1]-dat[,4]-6
mbf <- (dat[,11]==2)&(dat[,12]<=2)&(dat[,2]==1)&(experience==12)
sam <- (dat[,11]==4)&(dat[,12]==7)&(dat[,2]==0)
dat1 <- dat[mbf,]
dat2 <- dat[sam,]
#       First regression
y <- as.matrix(log(dat1[,5]/(dat1[,6]*dat1[,7])))
x <- cbind(dat1[,4],matrix(1,nrow(dat1),1))
xx <- t(x)%*%x
xy <- t(x)%*%y
beta <- solve(xx,xy)
print(beta)
#       Second regression
y <- as.matrix(log(dat2[,5]/(dat2[,6]*dat2[,7])))
experience <- dat2[,1]-dat2[,4]-6
exp2 <- (experience^2)/100
x <- cbind(dat2[,4],experience,exp2,matrix(1,nrow(dat2),1))
xx <- t(x)%*%x
xy <- t(x)%*%y
beta <- solve(xx,xy)
print(beta)
#       Create leverage and influence
e <- y-x%*%beta
xxi <- solve(xx)
leverage <- rowSums(x*(x%*%xxi))
r <- e/(1-leverage)
d <- leverage*e/(1-leverage)
print(max(abs(d)))
```

MATLAB Program File

```
% Load the data and create subsamples
dat = load cps09mar.txt;
# An alternative to load the data from an excel file is
# dat = xlsread('cps09mar.xlsx');
experience = dat(:,1)-dat(:,4)-6;
mbf = (dat(:,11)==2)&(dat(:,12)<=2)&(dat(:,2)==1)&(experience==12);
sam = (dat(:,11)==4)&(dat(:,12)==7)&(dat(:,2)==0);
dat1 = dat(mbf,:);
dat2 = dat(sam,:);
%       First regression
y = log(dat1(:,5)./(dat1(:,6).*dat1(:,7)));
x = [dat1(:,4),ones(length(dat1),1)];
xx = x'*x
xy = x'*y
beta = xx\xy;
display(beta);
%       Second regression
y = log(dat2(:,5)./(dat2(:,6).*dat2(:,7)));
experience = dat2(:,1)-dat2(:,4)-6;
exp2 = (experience.^2)/100;
x = [dat2(:,4),experience,exp2,ones(length(dat2),1)];
xx = x'*x
xy = x'*y
beta = xx\xy;display(beta);
%       Create leverage and influence
e = y-x*beta;
xxi = inv(xx)
leverage = sum((x.*(x*xxi))')';
d = leverage.*e./(1-leverage);
influence = max(abs(d));
display(influence);
```

3.26 Exercises

Exercise 3.1 Let Y be a random variable with $\mu = \mathbb{E}[Y]$ and $\sigma^2 = \operatorname{var}[Y]$. Define

$$g\left(y,\mu,\sigma^2\right) = \left(\begin{array}{c} y-\mu \\ (y-\mu)^2-\sigma^2 \end{array}\right).$$

Let $(\widehat{\mu},\widehat{\sigma}^2)$ be the values such that $\overline{g}_n(\widehat{\mu},\widehat{\sigma}^2)=0$ where $\overline{g}_n(m,s) = n^{-1}\sum_{i=1}^n g\left(y_i,m,s\right)$. Show that $\widehat{\mu}$ and $\widehat{\sigma}^2$ are the sample mean and variance.

Exercise 3.2 Consider the OLS regression of the $n\times 1$ vector $\boldsymbol{y}$ on the $n\times k$ matrix $\boldsymbol{X}$. Consider an alternative set of regressors $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{C}$, where $\boldsymbol{C}$ is a $k\times k$ non-singular matrix. Thus, each column of $\boldsymbol{Z}$ is a mixture of some of the columns of $\boldsymbol{X}$. Compare the OLS estimates and residuals from the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ to the OLS estimates from the regression of $\boldsymbol{Y}$ on $\boldsymbol{Z}$.

Exercise 3.3 Using matrix algebra, show $\boldsymbol{X}'\widehat{\boldsymbol{e}} = 0$.

Exercise 3.4 Let $\widehat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{Y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$. Find $\boldsymbol{X}_2'\widehat{\boldsymbol{e}}$.

Exercise 3.5 Let $\widehat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Find the OLS coefficient from a regression of $\widehat{\boldsymbol{e}}$ on $\boldsymbol{X}$.

Exercise 3.6 Let $\widehat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. Find the OLS coefficient from a regression of $\widehat{\boldsymbol{Y}}$ on $\boldsymbol{X}$.

Exercise 3.7 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$ and $\boldsymbol{M}\boldsymbol{X}_1 = 0$.

Exercise 3.8 Show that $\boldsymbol{M}$ is idempotent: $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

Exercise 3.9 Show that $\operatorname{tr}\boldsymbol{M} = n - k$.

Exercise 3.10 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and $\boldsymbol{X}_1'\boldsymbol{X}_2 = 0$ then $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$.

Exercise 3.11 Show that when $\boldsymbol{X}$ contains a constant, $\frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i = \overline{Y}$.

Exercise 3.12 A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be vectors of 1's and 0's, with the i^{th} element of $\boldsymbol{D}_1$ equaling 1 and that of $\boldsymbol{D}_2$ equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are n_1 men and n_2 women in the sample. Consider fitting the following three equations by OLS

$$\boldsymbol{Y} = \mu + \boldsymbol{D}_1\alpha_1 + \boldsymbol{D}_2\alpha_2 + \boldsymbol{e} \tag{3.53}$$

$$\boldsymbol{Y} = \boldsymbol{D}_1\alpha_1 + \boldsymbol{D}_2\alpha_2 + \boldsymbol{e} \tag{3.54}$$

$$\boldsymbol{Y} = \mu + \boldsymbol{D}_1\phi + \boldsymbol{e} \tag{3.55}$$

Can all three equations (3.53), (3.54), and (3.55) be estimated by OLS? Explain if not.

(a) Compare regressions (3.54) and (3.55). Is one more general than the other? Explain the relationship between the parameters in (3.54) and (3.55).

(b) Compute $\boldsymbol{1}_n'\boldsymbol{D}_1$ and $\boldsymbol{1}_n'\boldsymbol{D}_2$, where $\boldsymbol{1}_n$ is an $n \times 1$ vector of ones.

Exercise 3.13 Let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be defined as in the previous exercise.

(a) In the OLS regression

$$\boldsymbol{Y} = \boldsymbol{D}_1\widehat{\gamma}_1 + \boldsymbol{D}_2\widehat{\gamma}_2 + \widehat{\boldsymbol{u}},$$

show that $\widehat{\gamma}_1$ is the sample mean of the dependent variable among the men of the sample ($\overline{Y}_1$), and that $\widehat{\gamma}_2$ is the sample mean among the women ($\overline{Y}_2$).

(b) Let $\boldsymbol{X}$ ($n \times k$) be an additional matrix of regressors. Describe in words the transformations

$$\boldsymbol{Y}^* = \boldsymbol{Y} - \boldsymbol{D}_1\overline{Y}_1 - \boldsymbol{D}_2\overline{Y}_2$$

$$\boldsymbol{X}^* = \boldsymbol{X} - \boldsymbol{D}_1\overline{X}_1' - \boldsymbol{D}_2\overline{X}_2'$$

where $\overline{X}_1$ and $\overline{X}_2$ are the $k \times 1$ means of the regressors for men and women, respectively.

(c) Compare $\widetilde{\beta}$ from the OLS regression

$$\boldsymbol{Y}^* = \boldsymbol{X}^* \widetilde{\boldsymbol{\beta}} + \widetilde{\boldsymbol{e}}$$

with $\widehat{\beta}$ from the OLS regression

$$\boldsymbol{Y} = \boldsymbol{D}_1 \widehat{\alpha}_1 + \boldsymbol{D}_2 \widehat{\alpha}_2 + \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}}.$$

Exercise 3.14 Let $\widehat{\beta}_n = \left(\boldsymbol{X}_n' \boldsymbol{X}_n\right)^{-1} \boldsymbol{X}_n' \boldsymbol{Y}_n$ denote the OLS estimate when $\boldsymbol{Y}_n$ is $n \times 1$ and $\boldsymbol{X}_n$ is $n \times k$. A new observation (Y_{n+1}, X_{n+1}) becomes available. Prove that the OLS estimate computed using this additional observation is

$$\widehat{\beta}_{n+1} = \widehat{\beta}_n + \frac{1}{1 + X_{n+1}' \left(\boldsymbol{X}_n' \boldsymbol{X}_n\right)^{-1} X_{n+1}} \left(\boldsymbol{X}_n' \boldsymbol{X}_n\right)^{-1} X_{n+1} \left(Y_{n+1} - X_{n+1}' \widehat{\beta}_n\right).$$

Exercise 3.15 Prove that R^2 is the square of the sample correlation between $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$.

Exercise 3.16 Consider two least squares regressions

$$\boldsymbol{Y} = \boldsymbol{X}_1 \widetilde{\beta}_1 + \widetilde{\boldsymbol{e}}$$

and

$$\boldsymbol{Y} = \boldsymbol{X}_1 \widehat{\beta}_1 + \boldsymbol{X}_2 \widehat{\beta}_2 + \widehat{\boldsymbol{e}}.$$

Let R_1^2 and R_2^2 be the R-squared from the two regressions. Show that $R_2^2 \geq R_1^2$. Is there a case (explain) when there is equality $R_2^2 = R_1^2$?

Exercise 3.17 For $\widetilde{\sigma}^2$ defined in (3.47), show that $\widetilde{\sigma}^2 \geq \widehat{\sigma}^2$. Is equality possible?

Exercise 3.18 For which observations will $\widehat{\beta}_{(-i)} = \widehat{\beta}$?

Exercise 3.19 For the intercept-only model $Y_i = \beta + e_i$, show that the leave-one-out prediction error is

$$\widetilde{e}_i = \left(\frac{n}{n-1}\right)\left(Y_i - \overline{Y}\right).$$

Exercise 3.20 Define the leave-one-out estimator of σ^2,

$$\widehat{\sigma}_{(-i)}^2 = \frac{1}{n-1} \sum_{j \neq i} \left(Y_j - X_j' \widehat{\beta}_{(-i)}\right)^2.$$

This is the estimator obtained from the sample with observation i omitted. Show that

$$\widehat{\sigma}_{(-i)}^2 = \frac{n}{n-1} \widehat{\sigma}^2 - \frac{\widehat{e}_i^2}{(n-1)(1-h_{ii})}.$$

Exercise 3.21 Consider the least squares regression estimators

$$Y_i = X_{1i} \widehat{\beta}_1 + X_{2i} \widehat{\beta}_2 + \widehat{e}_i$$

and the "one regressor at a time" regression estimators

$$Y_i = X_{1i} \widetilde{\beta}_1 + \widetilde{e}_{1i}, \qquad Y_i = X_{2i} \widetilde{\beta}_2 + \widetilde{e}_{2i}$$

Under what condition does $\widetilde{\beta}_1 = \widehat{\beta}_1$ and $\widetilde{\beta}_2 = \widehat{\beta}_2$?

Exercise 3.22 You estimate a least squares regression

$$Y_i = X_{1i}'\widetilde{\beta}_1 + \widetilde{u}_i$$

and then regress the residuals on another set of regressors

$$\widetilde{u}_i = X_{2i}'\widetilde{\beta}_2 + \widetilde{e}_i$$

Does this second regression give you the same estimated coefficients as from estimation of a least squares regression on both set of regressors?

$$Y_i = X_{1i}'\widehat{\beta}_1 + X_{2i}'\widehat{\beta}_2 + \widehat{e}_i$$

In other words, is it true that $\widetilde{\beta}_2 = \widehat{\beta}_2$? Explain your reasoning.

Exercise 3.23 The data matrix is $(\boldsymbol{Y}, \boldsymbol{X})$ with $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$, and consider the transformed regressor matrix $\boldsymbol{Z} = [\boldsymbol{X}_1, \boldsymbol{X}_2 - \boldsymbol{X}_1]$. Suppose you do a least squares regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$, and a least squares regression of $\boldsymbol{Y}$ on $\boldsymbol{Z}$. Let $\widehat{\sigma}^2$ and $\widetilde{\sigma}^2$ denote the residual variance estimates from the two regressions. Give a formula relating $\widehat{\sigma}^2$ and $\widetilde{\sigma}^2$? (Explain your reasoning.)

Exercise 3.24 Use the data set from Section 3.22 and the sub-sample used for equation (3.50) (see Section 3.25) for data construction)

(a) Estimate equation (3.50) and compute the equation R^2 and sum of squared errors.

(b) Re-estimate the slope on education using the residual regression approach. Regress log(wage) on experience and its square, regress education on experience and its square, and the residuals on the residuals. Report the estimates from this final regression, along with the equation R^2 and sum of squared errors. Does the slope coefficient equal the value in (3.50)? Explain.

(c) Are the R^2 and sum-of-squared errors from parts (a) and (b) equal? Explain.

Exercise 3.25 Estimate equation (3.50) as in part (a) of the previous question. Let $\widehat{e}_i$ be the OLS residual, $\widehat{Y}_i$ the predicted value from the regression, X_{1i} be education and X_{2i} be experience. Numerically calculate the following:

(a) $\sum_{i=1}^n \widehat{e}_i$

(b) $\sum_{i=1}^n X_{1i}\widehat{e}_i$

(c) $\sum_{i=1}^n X_{2i}\widehat{e}_i$

(d) $\sum_{i=1}^n X_{1i}^2\widehat{e}_i$

(e) $\sum_{i=1}^n X_{2i}^2\widehat{e}_i$

(f) $\sum_{i=1}^n \widehat{Y}_i\widehat{e}_i$

(g) $\sum_{i=1}^n \widehat{e}_i^2$

Are these calculations consistent with the theoretical properties of OLS? Explain.

Exercise 3.26 Use the data set from Section 3.22.

(a) Estimate a log wage regression for the subsample of white male Hispanics. In addition to education, experience, and its square, include a set of binary variables for regions and marital status. For regions, create dummy variables for Northeast, South and West so that Midwest is the excluded group. For marital status, create variables for married, widowed or divorced, and separated, so that single (never married) is the excluded group.

(b) Repeat using a different econometric package. Compare your results. Do they agree?