# Chapter 3

# Linear regression

Let $X$ be a $K$ dimensional vector of random variables with a constant as its first element. Let $Y$ be a scalar random variable. Assume that the joint distribution of $X$ and $Y$ in the population of interest is known. Consider the following prediction problem: the analyst observes $X$ for a random draw from the population and wishes to use this information to form a 'good' prediction of $Y$ (which is not observed). As a concrete example consider a college admissions officer who wishes to use combined SAT score ($X$) to predict end-of-freshman-year GPA ($Y$).

As formulated thus far, our problem is identical to the one considered in an earlier chapter. We now depart from our earlier set-up by constraining ourselves to select a prediction function from a certain family. Let $X'b$ be a candidate prediction function. This function is *linear* in $X$; our analyst is restricting herself, *a priori*, to the class of **linear predictors**. This implies that choosing a prediction function is equivalent to choosing a coefficient vector, $b$.

The analyst will choose the prediction $x'b$ for $Y$ whenever she observes $X = x$. Prediction error, for a random draw, therefore equals $U(b) = Y - X'b$.

Making the notion of a 'good' prediction precise requires attaching a cost to prediction error. We will again assume that the analyst's **loss** is proportional to squared prediction error:

$$U\left(b\right)^2 = (Y - X'b)^2.$$

Since choosing $b$ to minimize loss is non-operational (doing so would require knowledge of $Y$), we instead choose this function to minimize *average* loss over a large number of replications of our prediction problem. Average loss or **risk** takes the mean squared error (MSE) form:

$$\mathbb{E}\left[U\left(b\right)^2\right] = \mathbb{E}\left[\left(Y - X'b\right)^2\right]. \tag{3.0.1}$$

We can evaluate risk (3.0.1) for different candidate values of $b \in \mathbb{R}^K$ since the joint distribution of $X$ and $Y$ is known. Choosing $b$ to minimize (3.0.1) corresponds to choosing a prediction rule that does well *on average*.

**Euclidean Norm**

As we develop basic results on linear regression we will make repeated use of various properties of matrix norms. A norm will we often use is the Euclidean Norm. The Euclidean Norm of an $m \times 1$ vector $a$ is

$$\|a\| = (a'a)^{1/2} = \left(\sum_{i=1}^{m} a_i^2\right)^{1/2}. \tag{3.0.2}$$

The Euclidean Norm of an $m \times n$ matrix $A$, also called the **Frobenius Norm**, is

$$\begin{aligned} \|A\| &= \|\mathrm{vec}\,(A)\| \\ &= \mathrm{tr}\,(A'A)^{1/2} \\ &= \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}. \end{aligned} \tag{3.0.3}$$

We can use the trace representation of the Frobenius Norm to derive a useful equality. Let $a$ and $b$ be $m \times 1$ vectors and consider the Frobenius Norm of the $m \times m$ matrix $ab'$:

$$\|ab'\| = \mathrm{tr}\,(ba'ab')^{1/2} = (b'ba'a)^{1/2} = \|a\|\,\|b\|. \tag{3.0.4}$$

Manipulating (3.0.1) we get an optimization problem of

$$\min_{b\in\mathbb{R}^K} \left\{\mathbb{E}\left[Y^2\right] - 2b'\mathbb{E}\left[XY\right] + b'\mathbb{E}\left[XX'\right]b\right\}. \tag{3.0.5}$$

Let $\beta_0$ denote the solution to (3.0.5). In order for this solution to be well-defined and unique we must impose some regularity conditions on the joint population distribution of $X$ and $Y$. Our results will apply to populations satisfying these conditions.

**Some useful inequalities**

**Expectation Inequality:** For any random matrix $Y$ with $\mathbb{E}\left[\|Y\|\right] < \infty$

$$\|\mathbb{E}\left[Y\right]\| \leq \mathbb{E}\left[\|Y\|\right]. \tag{3.0.6}$$

**Cauchy-Schwarz:** For any random $m \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$

$$\mathbb{E}\left[\|\mathbf{X}'\mathbf{Y}\|\right] \leq \mathbb{E}\left[\|\mathbf{X}\|^2\right]^{1/2}\mathbb{E}\left[\|\mathbf{Y}\|^2\right]^{1/2}. \tag{3.0.7}$$

**Assumption 1.** *(i)* $\mathbb{E}[Y^2] < \infty$*, (ii)* $\mathbb{E}\left[\|X\|^2\right] < \infty$*, and (iii)* $\mathbb{E}\left[(\alpha' X)^2\right] > 0$ *for any non-zero* $\alpha \in \mathbb{R}^K$*.*

Parts (i) and (ii) of Assumption 1 imply that the elements of $\mathbb{E}[XX']$ and $\mathbb{E}[XY]$ are finite (i.e., that the means, variance and covariances of $X$ and $Y$ are finite). To see this we use the Expectation/Modulus Inequality (3.0.6), our result about Frobenius Norms (3.0.4) and part (ii) of Assumption 1 to get

$$\|\mathbb{E}[XX']\| \leq \mathbb{E}[\|XX'\|] = \mathbb{E}\left[\|X\|^2\right] < \infty,$$

and the Expectation and the Cauchy-Schwarz Inequalities and parts (i) and (ii) of Assumption 1 to get

$$\|\mathbb{E}[XY]\| \leq \mathbb{E}[\|XY\|] \leq \mathbb{E}\left[\|X\|^2\right]^{1/2} \mathbb{E}\left[Y^2\right]^{1/2} < \infty.$$

Finiteness of the elements of $\mathbb{E}[XX']$ and $\mathbb{E}[XY]$ then follows by noting that the Euclidian norm of a matrix/vector is larger than the absolute value of any of its components. Therefore $\|\mathbb{E}[XY]\| < \infty$ implies finiteness of each element of $\mathbb{E}[XY]$ and similarly for $\mathbb{E}[XX']$. Collectively parts (i) and (ii) ensure that the analyst's maximand (3.0.5) is finite for any finite $b$.

Part (iii) of Assumption 1 requires that no single predictor corresponds to a linear combination of the others (i.e., that the elements of $X$ be linearly independent). This condition ensures invertibility of the **design matrix** $\mathbb{E}[XX']$ since

$$\mathbb{E}\left[(\alpha' X)^2\right] = \alpha' \mathbb{E}[XX']\,\alpha,$$

condition (iii) implies positive-definiteness of $\mathbb{E}[XX']$. This, in turn, implies that the determinant of $\mathbb{E}[XX']$ is non-zero (non-singularity) and hence that $\mathbb{E}[XX']^{-1}$ is well-defined.

Differentiating (3.0.5) with respect to $b$, setting the result equal to zero and dividing by two we get

$$\mathbb{E}[XX']\,\beta_0 - \mathbb{E}[XY] = 0.$$

Since $\mathbb{E}[XX']$ is invertible we can directly solve for $\beta_0$:

$$\beta_0 = \mathbb{E}[XX']^{-1} \times \mathbb{E}[XY]. \tag{3.0.8}$$

The corresponding best (i.e., MSE-minimizing) linear predictor of $Y$ given $X = x$ is

$$\mathbb{E}^*[Y \,|\, X = x] = x'\beta_0. \tag{3.0.9}$$

Define $U = Y - X'\beta_0$ to be the prediction error associated with (3.0.9). From the first order conditions to (3.0.5) we get

$$\mathbb{E}[XU] = 0. \tag{3.0.10}$$

Equation (3.0.10) indicates that $\beta_0$ is chosen to ensure that the covariance between $X$ and $U$ is zero. Recall that the first element of $X$ is a constant so that (3.0.10) implies

$$\mathbb{E}[U] = 0$$

or zero average prediction error. Our analyst does not systematically under- or over-predict freshman GPA. This is an unsurprising property. If she were systematically making such a prediction error she could eliminate it by raising or lowering the intercept in (3.0.9).

If $\mathbb{E}[X_k U] > 0$ for some $k$ then large values of $U$ would be associated with large values of $X_k$ on average. In such a situation our analyst could reduce risk by decreasing the coefficient on $X_k$ (i.e., by lowering $\beta_{k0}$). This would result in her lowering her prediction when observing large realizations of $X_k$.

Equations (3.0.8), (3.0.9) and (3.0.10) define the MSE-minimizing linear predictor (LP) of $Y$ given $X$.

It is important to keep in mind that the CEF $\mathbb{E}[Y|X]$ is a feature of the *conditional* distribution of $Y$ given $X$, while the LP $\mathbb{E}^*[Y|X]$ is a feature of the *joint* distribution of $Y$ and $X$. While, keeping the conditional distribution of $Y$ given $X$ fixed, $\mathbb{E}[Y|X]$ is invariant to changes in the marginal distribution of $X$ this is not the case for $\mathbb{E}^*[Y|X]$.

## 3.1 Short and long linear regression

Consider a researcher interested in the conditional distribution of the logarithm of weekly wages ($Y$) given years of competed schooling ($X$) and vector of additional worker attributes. This vector could include variables such as age, childhood test scores, and race. Let $W$ be this $J \times 1$ vector of additional variables. For concreteness I will refer to $W$ as childhood IQ, however the notation that follows will allow for $W$ to be vector-valued.

The linear predictor which uses $W$ in addition to $X$ is the **long linear regression** function:

$$\mathbb{E}^*[Y|X, W] = X'\beta_0 + W'\gamma_0. \tag{3.1.1}$$

The coefficient $\beta_0$ measures how much our prediction of log-earnings changes with schooling, holding IQ fixed.

The **short linear regression** function, which uses $X$ alone, is

$$\mathbb{E}^*[Y|X] = X'\delta_0. \tag{3.1.2}$$

Here $\delta_0$ measures how much our prediction of log-earnings changes with schooling unconditionally; that is without holding IQ fixed.

In what follows we will explore the relationship between these two regression functions. Specifically, how are the coefficients on schooling in the two regression functions, respectively $\beta_0$ and $\delta_0$, related?

Let $\mathbb{E}^*[W|X] = \Pi_0 X$ be the multivariate linear predictor of the $J \times 1$ vector $W$ given the $K \times 1$ vector $X$ where $\Pi_0$ is a $J \times K$ matrix of linear predictor coefficients. This multivariate linear predictor may be constructed by stacking up $j = 1, \ldots, J$ univariate linear predictors for each element of $W$. Let $\mathbb{E}^*[W_j|X] = X'\pi_{j0}$, where $W_j$ is the $j^{th}$ element of $W$, be the $j^{th}$ such linear predictor. We form $\Pi_0$ by setting its $j^{th}$ row equal to the transpose of $\pi_{j0}$. It is a useful exercise to show that $\Pi_0 = \mathbb{E}[WX'] \times \mathbb{E}[XX']^{-1}$. In what follows we will call

$$\mathbb{E}^*[W|X] = \Pi_0 X \tag{3.1.3}$$

the **auxiliary linear regression**.

The coefficient on $X$ in this regression, $\Pi_0$, measures how much our (linear) prediction of a randomly sampled individual's IQ would change given a unit change in their years of completed schooling. We will see shortly that economic models of optimal schooling choice often imply that 'high ability' individuals complete more years of schooling. Therefore we might expect $\Pi_0$ to be positive or for schooling to be predictive of an individual's childhood IQ. It turns out that we can use the long and auxiliary regression functions to interpret the short regression function coefficient on schooling.

Let $U = Y - \mathbb{E}^*[Y|X, W]$ be the prediction error associated with the long regression (3.1.1). By construction this prediction error is uncorrelated with $X$ and $W$. This implies that

$$\mathbb{E}^*[U|X] = 0. \tag{3.1.4}$$

We can now write

$$Y = X'\beta_0 + W'\gamma_0 + U,$$

which gives, after applying the linear predictor operator to both sides,

$$\begin{aligned}\mathbb{E}^*[Y|X] &= X'\beta_0 + \mathbb{E}^*[W'|X]\gamma_0 + \mathbb{E}^*[U|X] \\ &= X'\beta_0 + X'\Pi_0'\gamma_0 \\ &= X'\left(\beta_0 + \Pi_0'\gamma_0\right).\end{aligned}$$

The first equality above follows from the fact that the linear predictor operator is a linear operator (i.e., the best LP of $Z = X + Y$ is equal to the sum of the best LPs of $X$ and $Y$ alone). The second equality follows from the fact that the best LP of $X$ given $X$ is $X$ itself and equation (3.1.4) above.

To recap we have shown:

$$\delta_0 = \beta_0 + \Pi_0'\gamma_0. \tag{3.1.5}$$

Or, equivalently, the **Law of Iterated Linear Predictors (LILP):**

$$\mathbb{E}^*[Y|X] = \mathbb{E}^*[\mathbb{E}^*[Y|X,W]|X]. \tag{3.1.6}$$

The magnitude of the short regression coefficient, $\delta_0$, reflects a combination of two effects. First, it varies with how much our prediction of log-earnings would change given a unit change in schooling holding IQ fixed (i.e., the long regression coefficient, $\beta_0$). The second effect arises from the fact that (i) IQ and schooling covary in the population ($\Pi_0 \neq 0$) and (ii) conditional on schooling IQ is predictive of wages ($\gamma_0 \neq 0$). Relationship (3.1.5) is sometimes called the 'omitted variable bias formula'. I dislike this terminology as it implicitly privileges the long regression function.

An algebraic derivation of (3.1.5), that also facilitates the development of some additional results, follows. Define

$$\begin{array}{lll}\Sigma_{XX} = \mathbb{E}[XX'] & \Sigma_{WW} = \mathbb{E}[WW'] & \Sigma_{XW} = \mathbb{E}[XW'] \\ \Sigma_{XY} = \mathbb{E}[XY] & & \Sigma_{WY} = \mathbb{E}[WY]\end{array}$$

and $C = \Sigma_{WW} - \Pi_0\Sigma_{XW}$. Recalling that $\delta_0 = \Sigma_{XX}^{-1}\Sigma_{XY}$ and $\Pi_0 = \Sigma_{XW}'\Sigma_{XX}^{-1}$ we can partition

the vector of long regression coefficients as follows

$$\begin{aligned}
\begin{pmatrix} \beta_0 \\ \gamma_0 \end{pmatrix} &= \mathbb{E}\left[\begin{pmatrix} X \\ W \end{pmatrix}\begin{pmatrix} X \\ W \end{pmatrix}'\right]^{-1} \times \mathbb{E}\left[\begin{pmatrix} X \\ W \end{pmatrix} Y\right] \\
&= \begin{pmatrix} \Sigma_{XX} & \Sigma_{XW} \\ \Sigma_{XW}' & \Sigma_{WW} \end{pmatrix}^{-1} \times \begin{pmatrix} \Sigma_{XY} \\ \Sigma_{WY} \end{pmatrix} \\
&= \begin{pmatrix} \Sigma_{XX}^{-1} + \Pi_0' C^{-1} \Pi_0 & -\Pi_0' C^{-1} \\ -C^{-1}\Pi_0 & C^{-1} \end{pmatrix} \times \begin{pmatrix} \Sigma_{XY} \\ \Sigma_{WY} \end{pmatrix} \\
&= \begin{pmatrix} \delta_0 - \Pi_0' C^{-1} \left(\Sigma_{WY} - \Pi_0 \Sigma_{XY}\right) \\ C^{-1}\left(\Sigma_{WY} - \Pi_0 \Sigma_{XY}\right) \end{pmatrix} \\
&= \begin{pmatrix} \delta_0 - \Pi_0' \gamma_0 \\ \gamma_0 \end{pmatrix}.
\end{aligned}$$

Rearranging the first block of the last line above we have (3.1.5) above.

To illustrate the above ideas consider the following empirical analogs of the short, long and auxiliary regressions described above. Our fitted regression functions are based on $N = 1,896$ White male respondents from the cross-sectional sample of National Longitudinal Survey of Youth 1979 (NLSY79) cohort. The outcome of interest is the logarithm of average real wages from 1990 to 1993 (measured in 2010 prices). Here $X$ corresponds to years of completed schooling at age 28 and $W$ a respondent's percentile score on the Armed Forces Qualification Test (AFQT). All respondents were aged 14 to 22 in 1979.

The fitted short regression is

$$\widehat{\text{LogWage}} = \underset{(0.1226)}{8.6148} + \underset{(0.0091)}{0.1386}\ \text{YrsSch}.$$

The fitted long regression is

$$\widehat{\text{LogWage}} = \underset{(0.1272)}{8.7971} + \underset{(0.0118)}{0.1069}\ \text{YrsSch} + \underset{(0.0011)}{0.0043}\ \text{AFQT}.$$

The fitted auxiliary regression is

$$\widehat{\text{AFQT}} = -\underset{(2.35)}{42.79} + \underset{(0.17)}{7.43}\ \text{YrsSch}.$$

Note that $0.1386 \approx 0.1069 + 0.0043 \times 7.43$ as suggested by our results above (the small difference is due to rounding error). In the NLSY79 AFQT is predictive of both wages and years of schooling; consequently the estimated short and long regression coefficients on schooling differ considerably.

## 3.2 Residual regression

The algebraic derivation of the Law of Iterated Linear Predictors given in the previous section may be used to derive another useful property of LPs. Consider the following **residual linear regression**. First compute the multivariate linear predictor of $W$ given $X$ and the associated the $J \times 1$ vector of prediction errors $V = W - \Pi_0 X$. Second compute the linear regression of $Y$ given $V$ or the 'residual regression'

$$\mathbb{E}^*[Y|V] = V'\eta_0, \qquad \eta_0 = \mathbb{E}[VV']^{-1} \times \mathbb{E}[VY]. \tag{3.2.1}$$

Note that

$$\begin{aligned}
\mathbb{E}[VV'] &= \mathbb{E}\left[(W - \Pi_0 X)(W - \Pi_0 X)'\right] \\
&= \Sigma_{WW} - \Sigma_{XW}'\Pi_0' - \Pi_0\Sigma_{XW} + \Pi_0\Sigma_{XX}\Pi_0' \\
&= C - \Sigma_{XW}'\Sigma_{XX}^{-1}\Sigma_{XW} + \Sigma_{XW}'\Sigma_{XX}^{-1}\Sigma_{XX}\Sigma_{XX}^{-1}\Sigma_{XW} \\
&= C,
\end{aligned}$$

with $E[VY] = (\Sigma_{WY} - \Pi_0\Sigma_{XY})$. Therefore $\eta_0 = C^{-1}(\Sigma_{WY} - \Pi_0\Sigma_{XY}) = \gamma_0$. Using the equalities $\delta_0 = \beta_0 + \Pi_0'\gamma_0$ and $\eta_0 = \gamma_0$ we may write:

$$\begin{aligned}
\mathbb{E}^*[Y|X,W] &= X'\beta_0 + W'\gamma_0 \\
&= X'\left(\delta_0 - \Pi_0'\gamma_0\right) + W'\gamma_0 \\
&= X'\left(\delta_0 - \Pi_0'\gamma_0\right) + (W - \Pi_0 X)'\gamma_0 + (\Pi_0 X)'\gamma_0 \\
&= X'\delta_0 + V'\gamma_0 \\
&= X'\delta_0 + V'\eta_0 \\
&= \mathbb{E}^*[Y|X] + \mathbb{E}^*[Y|V].
\end{aligned}$$

Thus the long regression is the sum of the short and residual regression functions, a result know as the **Frisch-Waugh-Lovell Theorem**.