1 Linear Regression and Alike

1.1 Best Linear Prediction and Ordinary Least Squares

Let Y denote the the outcome variable, $X \in \mathbb{R}^p$ denote predictive variables, and P denote the joint distribution of (Y, X). The *best linear prediction (BLP)* parameter β is defined as

$$\beta = \operatorname*{argmin}_{b \in \mathbb{R}^p} \mathbb{E}_P[(Y - X'b)^2]. \tag{1}$$

That is, the goal is to choose b to minimize the expected squared error when predicting the value of Y by a linear function $X'b$. By the first-oder conditions,

$$\mathbb{E}_P[X(Y - X'\beta)] = 0,$$

so provided that $(\mathbb{E}_P[XX'])^{-1}$ exists,

$$\beta = (\mathbb{E}_P[XX'])^{-1}\mathbb{E}_P[XY].$$

An alternative, and perhaps more familiar way of defining the same parameter is:

$$Y = X'\beta + \varepsilon, \quad \mathbb{E}_P[\varepsilon X] = 0. \tag{2}$$

The requirement $\mathbb{E}_P[\varepsilon X]$ is called an *orthogonality condition*. If X includes a constant, this condition requires: (i) $\mathbb{E}[\varepsilon] = 0$; and (ii) ε is uncorrelated with X.

Equation (1) has a projection interpretation: $X'\beta$ is an orthogonal projection of Y onto the space of linear functions of X. For this reason, we will write:

$$\text{Proj}(Y|\text{lin}(X)) = X'\beta,$$

$$\varepsilon = Y - \text{Proj}(Y|\text{lin}(X)).$$

The BLP parameter is a particular feature of the joint distribution of (Y, X). Given a random sample $\{(Y_i, X_i)\}_{i=1}^n$, the analogy principle suggests estimating the BLP parameter by:

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i Y_i$$

This estimator is also known as the Ordinary Least Squares estimator because it can be obtained by solving a sample analog of (1):

$$\hat{\beta}_n = \underset{b \in \mathbb{R}^p}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2$$

It can also be written using matrix notation as

$$\hat{\beta}_n = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}, \tag{3}$$

where $\mathbb{Y} = (Y_1, \ldots, Y_n)$ and $\mathbb{X}$ is a $n \times p$ matrix with rows X_i'.

Assuming that $\mathbb{E}_P[X_i X_i']$ is invertible, the LLN combined with CMT immediately implies that $\hat{\beta}_n$ is a consistent estimator for β. Furthermore, using a multivariate version of CLT and Slutsky's Theorem, one can show that

$$\sqrt{n}(\hat{\beta}_n - \beta) \quad \to_d \quad N\left(0,\ \underbrace{(\mathbb{E}[XX'])^{-1}\mathbb{E}[\varepsilon^2 XX'](\mathbb{E}[XX'])^{-1}}_{\Sigma}\right).$$

Typically, we are interested in a parameter of the form $\theta = c'\beta$, which is naturally estimated by $\hat{\theta}_n = c'\hat{\beta}_n$. By the CMT,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \quad \to_d \quad N(0,\ \underbrace{c'\Sigma c}_{\sigma^2}).$$

A consistent estimator of Σ, under regularity conditions, can be obtained as:[1]

$$\hat{\Sigma}_n = \left(\frac{1}{n}\sum_i X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 X_i X_i'\right) \left(\frac{1}{n}\sum_i X_i X_i'\right)^{-1},$$

so that a consistent estimator of $\hat{\sigma}^2$ is $\hat{\sigma}_n^2 = c'\hat{\Sigma}_n c$. Arguing as before,

$$Z_n = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\hat{\sigma}_n} \quad \to_d \quad N(0, 1),$$

which allows to easily construct confidence intervals and tests for θ.

[1] The proof of this is a bit convoluted and not very deep, so we will skip it.

1.2 Frisch-Waugh-Lowell Theorem

Suppose we split $X_i = (D_i, W_i)$, where D_i is a scalar covariate of interest (the target covariate) and W_i is a vector of additional covariates (controls). Then, we can write

$$Y = D\beta_1 + W'\beta_2 + \varepsilon; \qquad \mathbb{E}[\varepsilon(D, W')] = 0 \tag{4}$$

where β_1 is the parameter of interest. This parameter measures how much the linear prediction of Y changes if we change D by one unit. It is called the *linear predictive effect* and should not be confused with causal or treatment effect. Think of it as a feature of the joint distribution of (Y, D, W).

Recalling the projection interpretation from Equation (1), we can define the partialling-out operator:

$$\tilde{V} \equiv V - \text{Proj}(V|\text{lin}(W)),$$

which can be applied to any random variable V. Applying it to both sides of equation (4), we obtain:

$$\tilde{Y} = \tilde{D}\beta_1 + \varepsilon; \qquad \mathbb{E}[\tilde{D}\varepsilon] = 0 \tag{5}$$

because $W \in \text{lin}(W)$, and ε is orthogonal to $\text{lin}(W, D)$ by construction. Comparing Equations (4) and (5), we have partialled out W entirely and obtained the target coefficient β_1 as:

$$\beta_1 = \frac{\mathbb{E}[\tilde{Y}\tilde{D}]}{\mathbb{E}[\tilde{D}^2]}.$$

This result is known as the Frisch-Waugh-Lowell Theorem.

One can also apply the same idea in-sample with an appropriate notion of orthogonality. Specifically, suppose that we observe $V_1, \ldots, V_n$[2]. Let us denote the sample average operator as

$$\mathbb{E}_n[V] = \frac{1}{n}\sum_{i=1}^{n} V_i.$$

Letting $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the OLS coefficients, we can write the in-sample version of Equation (4):

$$\mathbb{Y} = \mathbb{D}\hat{\beta}_1 + \mathbb{W}\hat{\beta}_2 + \hat{e} \tag{6}$$

where $\mathbb{D} = (D_1, \ldots, D_n)$ and $\mathbb{W}$ is a $n \times d_W$ matrix with rows W_i'. Here, the vector

[2] Here V_i is just a placeholder for Y_i, or D_i, or W_iD_i, etc.

of OLS residuals $\hat{e}$ is orthogonal to both $\mathbb{D}$ and $\mathbb{W}$ in a sense that

$$\mathbb{E}_n[D\hat{e}] = 0; \qquad \mathbb{E}_n[W\hat{e}] = 0$$

The matrix $P_W = \mathbb{W}(\mathbb{W}'\mathbb{W})^{-1}\mathbb{W}'$ projects $n \times 1$ vectors (e.g. $\mathbb{Y}$ or $\mathbb{D}$) onto the column space of $\mathbb{W}$. Therefore, the in-sample analog of the partialling-out operator is:

$$\tilde{\mathbb{V}} \equiv \mathbb{V} - P_W \mathbb{V}.$$

Applying this to both sides of Equation (6), we obtain:

$$\tilde{\mathbb{Y}} = \tilde{\mathbb{D}}\hat{\beta}_1 + \hat{e},$$

because $P_W \mathbb{W} = \mathbb{W}$ and $P_W \hat{e} = 0$. Therefore, the OLS estimator $\hat{\beta}_1$ can be computed as:

$$\hat{\beta}_1 = \frac{\mathbb{E}_n[\tilde{Y}\tilde{D}]}{\mathbb{E}_n[\tilde{D}^2]}.$$

Note that this estimator is exactly equal to the first component of (3), and therefore has the same asymptotic properties. The new representation from the Frish-Waugh-Lowell Theorem will be useful later when we deal with high-dimensional W_i.

1.3 Omitted Variable Bias

Suppose we're interested in the BLP coefficient β_1 defined as:

$$Y = D\beta_1 + W'\beta_2 + \varepsilon; \qquad \mathbb{E}[\varepsilon(D, W')] = 0,$$

but the controls W are not observed. Since we do not have access to W, we may consider estimating β from:

$$Y = D\beta + \nu; \qquad \mathbb{E}[\nu D] = 0.$$

Then,

$$\beta = \frac{\mathbb{E}[DY]}{\mathbb{E}[D^2]} = \beta_1 + \frac{\mathbb{E}[DW'\beta_2]}{\mathbb{E}[D^2]}.$$

The last summand on the right is called the *omitted variable bias*, because it arises if we omit W from the estimating equation.

A classic example of such setting is a wage regression where Y is (log) wage, D is an indicator of college education, and W (positive scalar) is the level of ability. In this setting, we would intuitively expect the omitted variable bias to be positive. So, by omitting W, we would overestimate of the effect of college education on wage.

1.4 Conditional Expectation

Consider a population with two observed traits Y and X. We can study the marginal distribution of Y, or conditional distributions of Y for the subpopulations with a particular value of $X = x$. For example, take one person at random from the US population, let Y denote her wage and X denote her number of years of education. Then, $\mathbb{E}[Y]$ measures the average wage in the US population, and $\mathbb{E}[Y|X = x]$ measures the average wage among people with x years of education. The function $g(x) = \mathbb{E}[Y|X = x]$ is called the *conditional distribution function.*

Often, we will need to work with a random variable $g(X)$, denoted simply $\mathbb{E}[Y|X]$ and refer to it as the conditional expectation of Y given X. Conditional expectations have a number of useful properties:

1. *Linearity:* one can split the sum and pull the constants out

$$\mathbb{E}[\alpha_1 Y_1 + \alpha_2 Y_2|X] = \alpha_1 \mathbb{E}[Y_1|X] + \alpha_2 \mathbb{E}[Y_2|X].$$

2. The *Law of Iterated Expectations:*

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y].$$

3. Conditional Expectation and Independence: if Y and X are independent, then

$$\mathbb{E}[Y|X] = \mathbb{E}[Y].$$

4. Conditioning is “knowing”: since X is “known”, $g(X)$ behaves like a constant

$$\mathbb{E}[g(X)Y|X] = g(X)\mathbb{E}[Y|X].$$

The properties of conditional expectation unlock another interpretation for the

BLP coefficient from Equation (1). Note that

$$\begin{aligned}\mathbb{E}[(Y - X'b)^2] \quad &= \mathbb{E}[((Y - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X] - X'b))^2] \\ &= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + 2 \cdot \mathbb{E}[\,(Y - \mathbb{E}[Y|X])((\mathbb{E}[Y|X] - X'b))\,] \\ &+ \mathbb{E}[(\mathbb{E}[Y|X] - X'b)^2]\end{aligned}$$

Here $\mathbb{E}[(Y - \mathbb{E}[Y|X])^2]$ does not depend on b and, by the Law of Iterated Expectations,

$$\mathbb{E}[\,(Y - \mathbb{E}[Y|X])((\mathbb{E}[Y|X] - X'b))\,] = 0.$$

Therefore:

$$\operatorname*{argmin}_{b \in \mathbb{R}^d} \mathbb{E}[(Y - X'b)^2] = \operatorname*{argmin}_{b \in \mathbb{R}^d} \mathbb{E}[(\mathbb{E}[Y|X] - X'\beta)^2].$$

Therefore, the BLP coefficient β is the coefficient in the *best linear approximation* $x'\beta$ of the conditional mean function $\mathbb{E}[Y|X = x]$.

1.5 Linear Regression Model

The model:

$$Y = X'\beta + \varepsilon; \qquad \mathbb{E}[\varepsilon|X] = 0$$

is known as the *linear regression model.* It is substantially different from the BLP setup because the assumption $\mathbb{E}[\varepsilon|X] = 0$ restricts the distribution of the data: it implies that $\mathbb{E}[Y|X = x] = x'\beta$, i.e., the conditional expectation function is linear. In turn, the orthogonality restriction $\mathbb{E}[\varepsilon X] = 0$ merely defines the parameter β as the BLP coefficient. We can also see the implication formally:

$$\mathbb{E}[\varepsilon X] = \mathbb{E}[\mathbb{E}[\varepsilon X|X]] = \mathbb{E}[X\mathbb{E}[\varepsilon|X]] = 0,$$

using properties 2 and 4 of the conditional expectation.

The vector of parameters β can be estimated by OLS in exactly the same fashion as above, and has the same asymptotic properties. However, it has a different interpretation because of the assumption $\mathbb{E}[Y|X = x] = x'\beta$. The requirement $\mathbb{E}[\varepsilon|X] = 0$ is known as *exogeneity* or *conditional mean independence* assumption.

Let $X = (D, W)$ and $D \in \{0, 1\}$ be the target covariate. Suppose we want to estimate the *predictive effect*:

$$\mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0].$$

In the linear regression model

$$Y=\beta_0+\beta_1 D+\varepsilon; \qquad \mathbb{E}[\varepsilon|D]=0,$$

we have

$$\mathbb{E}[Y|D=1]=\beta_0+\beta_1,$$
$$\mathbb{E}[Y|D=0]=\beta_0$$

so that $\beta_1=\mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0]$ can be estimated by OLS. Alternatively, consider the *conditional predictive effect:*

$$\mathbb{E}[Y|D=1,W=w]-\mathbb{E}[Y|D=0,W=w].$$

In the linear regression model

$$Y=D\beta_1+W'\beta_2+\varepsilon; \qquad \mathbb{E}[\varepsilon|D,W]=0,$$

we have

$$\mathbb{E}[Y|D=1,W=w]=\beta_1+w'\beta_2$$
$$\mathbb{E}[Y|D=0,W=w]=w'\beta_2$$

so that $\beta_1=\mathbb{E}[Y|D=1,W=w]-\mathbb{E}[Y|D=0,W=w]$ can be estimated by OLS.

However, linearity of the conditional mean $\mathbb{E}[Y|W,D]$ is a very restrictive assumption. For example, it implies that β_1 is constant for all w. We can obtain better estimators of the conditional predictive effect by considering non-linear and non-parametric regression models, e.g.,

$$\mathbb{E}[Y|D,W]=\beta_1(W)D+\beta_2(W).$$

We will discuss such models later in the course.

1.6 Potential Outcomes Framework and Causal Parameters

Consider an experiment in which some units are treated and some are not (e.g. a job training program, a vaccine trial, etc.). The treated units form the *treatment group*, the non-treated units form the *control group*. Let $D \in \{0, 1\}$ indicate the treatment status, and $(Y(1), Y(0))$ denote the *potential outcomes* of a unit if it's treated and not treated correspondingly. The observed outcome is $Y = DY(1) + (1 - D)Y(0)$.

One causal parameter of interest is the *average treatment effect*:

$$ATE = \mathbb{E}[Y(1) - Y(0)].$$

If the experiment is randomized (i.e. each unit is assigned to the treatment or control group by "flipping a coin"), then D is independent from the potential outcomes $(Y(0), Y(1))$. Such experiments are often called Randomized Controlled Trials or A/B tests, and the independence assumption is called *strict exogeneity*. Then,

$$\begin{aligned} \mathbb{E}[Y|D] &= \mathbb{E}[Y(1)|D] \cdot D + \mathbb{E}[Y(0)|D] \cdot (1 - D) \\ &= \underbrace{\mathbb{E}[Y(0)]}_{\beta_0} + \underbrace{\mathbb{E}[Y(1) - Y(0)]}_{\beta_1} \cdot D \end{aligned}$$

Therefore, the ATE equals β_1 in a regression model $Y = \beta_0 + \beta_1 D + \varepsilon$ with $\mathbb{E}[\varepsilon|D] = 0$ and can simply be estimated by OLS. In this way, the coefficient in a linear regression model is given a causal interpretation. Without strict exogeneity, the slope coefficient in the linear regression model does not estimate the ATE.[3]

The treatment effects may be heterogeneous, i.e., may vary with observed characteristics W of the units. For this reason, another parameter of interest is *conditional average treatment effect:*

$$CATE(w) = \mathbb{E}[Y(1) - Y(0)|W = w].$$

If the experiment is randomized conditional on W (i.e. the treatment status is determined by flipping a coin within each subpopulation with fixed $W = w$), the treatment status D is independent from the potential outcomes $(Y(0), Y(1))$ conditional on W. Such experiments are called *stratified randomized experiments* and the assumption is

[3] As and optional exercise, you can show that in this case $\beta_1 = ATE + \text{bias}$.

called *conditional exogeneity* assumption. In this case,

$$\begin{aligned}\mathbb{E}[Y|D,W] &= \mathbb{E}[Y(1)|D,W]\cdot D + \mathbb{E}[Y(0)|D,W]\cdot(1-D)\\ &= \underbrace{\mathbb{E}[Y(0)|W]}_{\beta_0(W)} + \underbrace{\mathbb{E}[Y(1)-Y(0)|W]}_{\beta_1(W)}\cdot D.\end{aligned}$$

So, $\beta_1(W)$ corresponds to the CATE parameter. Both $\beta_0(W)$ and $\beta_1(W)$ can be estimated either non-parametrically or imposing additional assumptions. For example, assuming that the conditional expectations are linear in W:

$$\mathbb{E}[Y(0)|W] = \beta_0 + \beta_2' W;$$

$$\mathbb{E}[Y(1)-Y(0)|W] = \beta_1 + \beta_3' W,$$

one can write down a linear regression model:

$$Y = \beta_0 + \beta_1 D + \beta_2' W + \beta_3'(DW) + \varepsilon; \qquad \mathbb{E}[\varepsilon|D,W] = 0.$$

This model can be estimated using OLS to obtain $\widehat{CATE}(W) = \hat{\beta}_1 + \hat{\beta}_3' W$. Of course, causal interpretation of this parameter hinges on the linearity assumption, which is rather strong. We will return to non-parametric estimation in this context later in the course.

1.7 Endogeneity and Instrumental Variables

In many situations, the exogeneity assumption in the linear regression model may fail: the latent variable ε is likely to be correlated with some of the regressors X. The top three reasons are systematic measurement errors, omitted variables, and simultaneity.

1.7.1 Measurement Errors

Consider the BLP setting (or a regression model):

$$Y = X'\beta + \varepsilon; \quad \mathbb{E}[\varepsilon X] = 0,$$

but suppose the researcher observes $\tilde{X} = X + \eta$, where η denotes the measurement error. Then:

$$Y = \tilde{X}'\beta + \underbrace{(\varepsilon - \eta'\beta)}_{\tilde{\varepsilon}}.$$

Therefore, we have $\mathbb{E}[\tilde{X}\eta'] \neq 0$ by construction, which means that, the OLS estimator $\hat{\beta}_n$ based on the above equation will be inconsistent.

On the other hand, suppose we have available a random vector Z such that $\mathbb{E}[\varepsilon Z] = 0$ and $\mathbb{E}[Z\eta'] = 0$. Then,

$$\mathbb{E}[(Y - \tilde{X}'\beta)Z] = 0 \quad \Longrightarrow \quad \beta = (\mathbb{E}[Z\tilde{X}'])^{-1}\mathbb{E}[ZY],$$

provided that $\mathbb{E}[Z\tilde{X}']$ is invertible. The elements of Z are called *instrumental variables.*

When dealing with measurement errors, a common empirical approach is to obtain another noisy measurement of X, say $Z = X + \nu$ for which the measurement error is uncorrelated with the original one, i.e. $\mathbb{E}[\nu\varepsilon] = 0$ and $\mathbb{E}[\nu\eta'] = 0$. Such Z satistifes the stated assumption and allows to construct a consistent estimator for the parameter β. Therefore, having two noisy measurements of X allows to recover β.

For a recent empirical application of this idea, see Chalfin and McCrary (2013), who employ multiple measurements of police force growth rates to address measurement error in a regression of crime growth on police force growth.

1.7.2 Omitted Variables

Suppose we're interested in a parameter β_1 from a regression model

$$Y = X_1'\beta_1 + X_2'\beta_2 + \varepsilon,$$

but the variables X_2 are unobservable. Here, even if X_1 is uncorrelated with ε, it mights still be correlated with X_2. As a result, for $\tilde{\varepsilon} = X_2'\beta_2 + \varepsilon$, we have $\mathbb{E}[X_1\tilde{\varepsilon}] \neq 0$, so the OLS estimator will be inconsistent.

Suppose that we have a random vector Z such that $\mathbb{E}[ZX_2'] = 0$ and $\mathbb{E}[Z\varepsilon] = 0$. Then:

$$\mathbb{E}[(Y - \beta'X_1)Z] = 0 \quad \Longrightarrow \quad \beta = (\mathbb{E}[ZX_1])^{-1}\mathbb{E}[ZY],$$

provided that $\mathbb{E}[ZX_1]$ is invertible. As before, the plug-in principle immediately suggests a consistent estimator.

Omitted variables are a common concern in the study of returns to education; see e.g. Card (2001).

1.7.3 Simultaneity

Consider the following structural model from Angrist and Krueger (1991):

$$\begin{aligned} Y &= D\beta_1 + W'\beta_2 + U \\ D &= Z\alpha_1 + W'\alpha_2 + V \end{aligned}$$

where Y is the log of wage, D is years of education, W includes geographic indicators, year of birth, race, marital status, and a constant, and Z is a dummy for the fourth quarter of birth. Here, (Y, D) are called *endogenous* variables, (W, Z) — *exogenous* variables, and (U, V) — the error terms. We will assume that $(U, V) \perp (W, Z)$, but U and V are potentially correlated.

The causal parameter of interest is β_1, measuring returns to schooling. Estimating this parameter is difficult because it is likely that $\mathbb{E}[DU] \neq 0$ due to $\mathbb{E}[UV] \neq 0$ due to a common unobserved factor that influences both wage and the choice of education, such as innate ability. This is precisely the issue of simultaneity. As a result, running OLS for the first equation will produce inconsistent estimates.

In this example, Z is an instrumental variable, which satisfies $\mathbb{E}[U(Z, W')] = 0$, or equivalently

$$\mathbb{E}\left[(Y - D\beta_1 - W'\beta_2)\begin{pmatrix} Z \\ W \end{pmatrix}\right] = 0$$

This is a system of $\dim(W) + 1$ equations with the same number of unknowns, which can be solved for (β_1, β_2) in the same fashion as in the previous examples.

Another common example of simultaneity comes from demand estimation. The following discussion is based on Berry (1994). Suppose there are $j = 1, \dots, J$ products available, and each customer decides which product to purchase (not choosing any product is also a possibility, called an *outside option* and recorded as $j = 0$). Under certain assumptios on how individuals make choices, Berry (1994) shows that the market shares S_j satisfy

$$\log S_j/S_0 = P_j\beta_1 + W_j'\beta_2 + \varepsilon_j,$$

where P_j is the price of product j, W_j are observable characteristics of product j, and ε_j are product characteristics observed by the consumers but unobserved by the econometrician.

To estimate (β_1, β_2) by OLS would require $\mathbb{E}[\varepsilon_j P_j] = 0$ and $\mathbb{E}[\varepsilon_j W_j] = 0$. The former assumption is particularly problematic: under oligopolistic competition, the price charged by the firm should reflect any unobservable characteristic that makes the product more appealing to consumers. In other words, $\mathbb{E}[\varepsilon_j P_j] > 0$.

To resolve this problem, we need to find a variable Z_j satisfying $\mathbb{E}[Z_j \varepsilon_j] = 0$. In this context, a common choice is the price of an input used in the production of a good j (i.e. a cost shifter). Then,

$$\mathbb{E}\left[(\log S_j/S_0 - \beta_1 P_j - W_j'\beta_2)\begin{pmatrix} Z \\ W \end{pmatrix}\right] = 0,$$

which allows to construct a consistent estimator using the plug-in principle as discussed before.

1.7.4 Estimation

All of the above examples share the following structure:

$$Y = X'\beta + \varepsilon; \quad \mathbb{E}[\varepsilon X] \neq 0, \text{ but } \mathbb{E}[\varepsilon Z] = 0,$$

for a vector of *instrumental variables* (or simply *instruments*) Z with $\dim(Z) \geqslant \dim(X)$. Equivalently,

$$\mathbb{E}[(Y - X'\beta)Z] = 0. \tag{7}$$

Instrumental variables are required to satisfy two conditions: (i) $\mathbb{E}[\varepsilon Z] = 0$, known as the *validity* condition, and (ii) $Z \not\perp X$, known as the *relevance* condition. In words, Z should affect X but be uncorrelated with ε. Finding a good instrument is art! We will discuss some examples later.

To estimate β, we could follow the plug-in principle and find $\hat{\beta}_n$ solving:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n)Z_i = 0. \tag{8}$$

This works just fine if $\dim(Z) = \dim(X)$, but when $\dim(Z) > \dim(X)$, this system of equations is unlikely to have solutions due to the sampling uncertainty. Instead, we can pick a symmetric weighting matrix Ω_n and solve:

$$\hat{\beta}_n^{IV} = \underset{b}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'b) Z_i \right)' \Omega_n \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'b) Z_i \right).$$

This is a convex minimization problem that has a simple closed-form solution:

$$\hat{\beta}_n^{IV} = (\mathbb{X}'\mathbb{Z}\Omega_n\mathbb{Z}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Z}\Omega_n\mathbb{Z}'\mathbb{Y},$$

provided $(\mathbb{X}'\mathbb{Z}\Omega_n\mathbb{Z}'\mathbb{X})^{-1}$ exists. Here, $\mathbb{Z}$ is a $n \times d_Z$ matrix with rows Z_i' and $\mathbb{X}$ and $\mathbb{Y}$ are as before.

One can show that such $\hat{\beta}_n^{IV}$ is consistent, i.e. $\hat{\beta}_n^{IV} \to_p \beta$, for any weighting matrix Ω_n which converges in probability to a non-degenerate matrix Ω. The question is then which Ω_n to choose. Here are three options:

1. $\Omega_n = I_d$, which is the same as solving (8) if $\dim(X) = \dim(Z)$.

2. $\Omega_n = (\frac{1}{n} \sum_{i=1}^{n} Z_i Z_i')^{-1}$, which corresponds to *Two-Stage Least Squares*, or *2SLS*:

   (1) Run OLS of each component X_j on Z and obtain predicted values $\hat{X}_{ji}$.

   (2) Run OLS of Y on $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_d)$, and obtain $\hat{\beta}_n^{2SLS}$.

3. $\Omega_n = (\frac{1}{n} \sum_{i=1}^{n} Z_i Z_i' \hat{e}_i^2)^{-1}$, which leads to the lowest possible asymptotic variance, and therefore is a preferred choice. It is sometimes called *Three-Stage Least Squares* and computed as follows:

   (1) Obtain a consistent estimator $\tilde{\beta}$ (e.g. using options 1 or 2).

   (2) Construct the residuals $\hat{e}_i = Y - X_i'\tilde{\beta}$ and compute Ω_n.

   (3) Compute $\hat{\beta}_n^{IV}$.

All of the resulting estimators satisfy:

$$\hat{\beta}_n^{IV} \quad \to_p \quad \beta$$

$$\sqrt{n}(\hat{\beta}_n^{IV} - \beta) \quad \to_d \quad N(0, V),$$

where V can be consistently estimated from the data. Like with OLS, this result allows to construct tests and confidence intervals for the components of β.

1.8 Challanges with Many Covariates

Researchers in economics often have to deal with many covariates or instruments. Examples include cross-country growth regressions (Sala-i-Martin 1997), demand estimation (Bajari, Nekipelov, Ryan, and Yang 2015), macro/finance forecasting (papers by Stock and Watson), or using "judge fixed-effects" as instruments for judges decisions (Aizer and Doyle 2015). Moreover, researchers often use technical regressors (polynomials and interaction terms constructed from the original regressors) to model a non-linear relationship or capture heterogeneous effects, or use many instruments (again, interaction terms) in the first stage of the 2SLS procedure to increase its R^2 and obtain "more precise" IV estimates. Another example is series estimation, a non-parametric estimation method that we will discuss later.

Using many covariates poses both practical and conceptual challenges. The main practical challenge is *multicollinearity* or approximate linear dependence between the regressors. Recall the OLS formula:

$$\hat{\beta}_n = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y},$$

where $\mathbb{X}$ is the $n \times p$ matrix with rows X_i'. If the columns of $\mathbb{X}$ are approximately linearly dependent, the matrix $\mathbb{X}'\mathbb{X}$ will be approximately degenerate and its inverse will not be numerically stable. Intuitively, taking an inverse of an approximately degenerate matrix is like dividing by a number approximately equal to zero. This may lead to very large standard errors and wide confidence intervals. Indeed, in the simplest case of homoskedastic error terms, $\mathbb{V}ar(\hat{\beta}_n|\mathbb{X}) = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$. These issues become more pronounced as the number of regressors approaches the sample size.

When using too many instruments relative to the sample size, the 2SLS estimator tends to be biased towards the OLS estimator. This is a problem of overfitting: Z predicts X too well in the first stage (think of $Z \approx X$, in which case $\hat{\beta}_n^{OLS} \approx \hat{\beta}_n^{2SLS}$). The classic references are Angrist and Krueger (1991) and Bound, Baker and Jaeger (1995). We will talk about overfitting a lot in this course.

Using many covariates also raises a conceptual problem. When doing inference, we rely on the asymptotic approximation in which p is fixed and n goes to infinity. We do not actually think that the researcher will keep on collecting infinite amounts of data. Instead, we say that an estimator is approximately Normal if n is large and construct a confidence interval. But is this a good approximation for a given n? If

$p = 5$ and $n = 1000$, it probably is, if $p = 90$ and $n = 100$, perhaps not, and if $p > n$ we cannot even compute the OLS estimator.

As a way out, one can employ a different approximation in which $p = p(n)$ is allowed to increase with n, explicitly acknowledging the fact that p is large. In a classic paper, Huber (1973) shows that consistency and asymptotic normality of the vector $\hat{\beta}_n$ require $p/n \to 0$ as $n \to \infty$. Formally, one can show that:

$$\mathbb{E}\left[\left|\left|\hat{\beta}_n - \beta\right|\right|_2^2\right] \sim \frac{\sigma^2 p}{n}, \tag{9}$$

meaning that for large p/n, $\hat{\beta}_n$ will be far from β, on average.[4] Recently, Cattaneo, Jansson, and Newey (2018) focus on a single coefficient $\hat{\beta}_k$ and show that consistency and asymptotic normality can be achieved even if $p/n \to c > 0$.

Since dealing with a large number of covariates is hard, there is a natural question: do all of them really matter? If not, can we simply choose a small subset and proceed as before? Should we rely on economic intuition/models or the data? The answers ultimately depend on the application at hands and our beliefs about the world. In the following sections, we will consider a variety of penalization and dimensionality reduction methods that help address the problems with many covariates.

1.9 Homework

1. In this problem, you will revisit the unemployment insurance experiment performed by the US department of labor in the 1980-s. Unemployment insurance reduces the risks and the costs associated with being unemployed. On the other hand, it incentives people to put less effort into getting a job, which negatively affects the economy. One possible solution is to offer a monetary bonus for finding a new job quickly.

   To assess the effectiveness of this potential solution, the department of labor randomly split the unemployed into control and treatment groups and offered a cash bonus for finding a job within a certain period of time to the latter. The data is contained in `penn.csv` and described in the codebook posted on the class website.

[4] Here $||a - b||_2^2 = \sum_{j=1}^{p}(a_j - b_j)^2$.

(1) Select a subsample that only includes the control group and one of the treated groups, `tg = 0` and, say, `tg = 4`.

(2) Define the outcome variable to be the logarithm of the number of weeks $Y = \log(\texttt{inuidur1})$, create the treatment indicator $D = \mathbf{1}(\texttt{tg} = 4)$, and a vector of controls $W =$ (sex, race, and age dummies).

(3) Assume that the conditional exogeneity assumption is satisfied. Estimate the model:

$$Y = \beta_0 + \beta_1 D + \varepsilon.$$

How do you interpret the coefficients? Under what assumptions do the coefficients have a *causal* interpretation? Do the results suggest that the program was successful in reducing the length of unemployment?

(4) Estimate the model:

$$Y = \beta_0 + \beta_1 D + \beta_2' W + \varepsilon.$$

How do you interpret the coefficients? Under what assumptions do the coefficients have a *causal* interpretation? What is the difference with part (3)? Do the results suggest that the program was successful in reducing the length of unemployment?

(5) Estimate the model:

$$Y = \beta_0 + \beta_1 D + \beta_2' W + \beta_3' WD + \varepsilon$$

How do you interpret the coefficients? Do the results suggest that the model from part (3) was reasonable?

(6) Do your conclusions change for different treatment groups? (do not include all of the estimation results, just discuss them).

2. Perform Monte-Carlo simulations to investigate performance of different weighting matrices Ω_n used in the IV estimator. Below is a suggested roadmap. Use $M = 1000$ simulations and $n = 200$ observations.

(0) Pick some true value of $\beta = (\beta_1, \beta_2) \in \mathbb{R}^2$.

(1) For each simulation, create $(Y_i, X_i, Z_i)_{i=1}^n$ where $d_Z > d_X$, X is endogenous, and Z is a valid instrument. To obtain interesting results, introduce some heteroskedasticity.

(2) Calculate $\hat{\beta}_{n,j}^{IV}$ for $j = 1, 2, 3$ using the three choices of Ω_n.

(3) Plot the histogram of $\hat{\beta}_{n,j}^{IV}$ across simulations along with the true value. Compare the results.

How do your results change for $n = 30$ or $n = 500$?

3.* This problem is optional. A researcher estimated the model:

$$Y = \beta_0 + \beta_1 X + \varepsilon; \qquad \mathbb{E}[\varepsilon|X] = 0$$

using OLS and was unhappy with the results: the confidence interval for β_1 turned out too wide. So, she came up with an alternative estimator constructed as follows: (i) For each pair of points (x_i, y_i) and (x_j, y_j) from the dataset, fit a straight line through them and let $\beta_1^{(ij)}$ denote the slope of that line; (ii) Set $\hat{\beta}_1 = \frac{1}{\binom{n}{2}} \sum_{i \neq j} \beta_1^{(ij)}$. Answer the following questions:

(1) Is $\hat{\beta}_1$ unbiased, i.e., is $\mathbb{E}[\hat{\beta}_1] = \beta_1$?

(2) Is $\hat{\beta}_1$ consistent? (Hint: use the Law of Iterated Expectations and the Law of Large Numbers)

(3) Is this $\hat{\beta}_1$ likely to have shorter confidence interval? (Hint: look up and apply the Gauss-Markov Theorem)