Stat 157/260: Section 3 Notes - Fall 2019

Ordinary Least Square and its Friends

Chaoran Yu

We first set up some notations. Let $\mathbf{y}$ be an $(n \times 1)$ vector, collecting observations of the response (or dependent) variable. Let $\mathbf{X}$ be an $(n \times k)$ matrix, collecting observations of the explanatory (or independent) variables.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\intercal \\ \mathbf{x}_2^\intercal \\ \vdots \\ \mathbf{x}_n^\intercal \end{bmatrix}$$

Here, the rows of $\mathbf{X}$ store observations (individuals, countries, etc.) and the columns store variables. The row vector $\mathbf{x}_i^\intercal = (x_{i1}, x_{i2}, ..., x_{ik})$ contains the values of each of the k regressors for observation i. Unless otherwise stated, for this section, we assume that $x_{i1} = 1$ is a constant term.

1 Linear Projection

Given observed data $\left(y_i, \mathbf{x}_i\right)_{i=1}^n$, we are interested in describing the relationship between y_i and $\mathbf{x}_i$. An often used device is the conditional expectation function, $\mu(\mathbf{x}_i) \equiv \mathbb{E}\left[y_i \mid \mathbf{x}_i\right]$. Its popularity stems from an oft-touted property of the conditional expectation function is that it has an interpretation as an optimal predictor.

Proposition 1.1. *Assume Y_i is square integrable, then $\mu(\mathbf{x}_i) = \arg\min_{m \in \mathcal{M}} \mathbb{E}\left[\left(y_i - m(\mathbf{x}_i)\right)\right]$, where $\mathcal{M}$ is set of all square integrable functions.*

Proof. We decompose the square loss function as follows

$$\mathbb{E}\left[\left(y_i - m(\mathbf{x}_i)\right)^2\right] = \mathbb{E}\left[\left[\left(y_i - \mu(\mathbf{x}_i)\right) - \left(\mu(\mathbf{x}_i) - m(\mathbf{x}_i)\right)\right]^2\right] = \mathbb{E}\left[\left(y_i - \mu(\mathbf{x}_i)\right)^2\right] + \mathbb{E}\left[\left(\mu(\mathbf{x}_i) - m(\mathbf{x}_i)\right)^2\right].$$

For the optimization problem, the first term does not matter and the second is 0 when we set $m(\mathbf{x}_i) = \mu(\mathbf{x}_i)$. □

For this reason the conditional expectation is sometimes called the "best" or minimum mean square error predictor. It is worth pointing out that this optimality property is an artifact of the mean square loss function, to which we need not attribute any special importance. Were we to adopt an absolute loss function, the optimal predictor is the conditional median function.

The estimation of $\mu(\mathbf{x}_i)$ can be difficult, since without any restriction on the joint distribution, this object is an infinite dimensional object. Instead we can consider a linear approximation/predictor. Let $\text{Proj}\left(y_i \mid \mathbf{x}_i\right) \equiv \mathbf{x}_i^\intercal \beta$ with $\beta \equiv \mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\intercal\right]^{-1} \mathbb{E}\left[\mathbf{x}_i y_i\right]$ to be the population linear projection, where we assumed the invertibility of $\mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\intercal\right]$.

Proposition 1.2. *Assume Y_i is square integrable, then* $\text{Proj}\left(y_i \mid \mathbf{x}_i\right) = \arg\min_{b \in \mathbb{R}^k} \mathbb{E}\left[\left(y_i - \mathbf{x}_i^\intercal b\right)^2\right]$.

Proof. Same as above. □

$\text{Proj}\left(y_i \mid \mathbf{x}_i\right)$ is thus termed as the best linear predictor.

The formulation of β is obtained from the first order condition of the criterion function

$$\mathbb{E}\left[\mathbf{x}_i \left(y_i - \mathbf{x}_i^\intercal \beta\right)\right] = 0.$$

To facilitate interpretation, we assume that $\mathbf{x}_i$ include a constant (unless otherwise stated), so then the necessary condition for optimality is a covariance restriction which states that the prediction error $u_i \equiv y_i - \mathbf{x}_i^\intercal \beta$ such that

$\mathbb{E}[u_i] = 0$ must exhibit no (linear) covariation with the predictor variables $\mathbf{x}_i$. Assuming that $\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\intercal\right]$ is invertible, we obtain

$$\beta = \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\intercal\right]^{-1}\mathbb{E}\left[\mathbf{x}_i y_i\right],$$

whose sample analogue is the least square estimator. Before moving onto least square properties, we wish to state two properties of the linear projection operator as it inherits properties from the conditional expectation operator.

Exercise 1.A

Let $\left(y_i, \mathbf{x}_i, \mathbf{z}_i\right)_{i=1}^n$ be observed variables, and assume invertibility of $\mathbb{E}\left[(\mathbf{x}_i, \mathbf{z}_i)^\intercal (\mathbf{x}_i, \mathbf{z}_i)\right]$. Prove that

1. if $\mu(\mathbf{x}_i)$ is linear, then $\text{Proj}\left(y_i \mid \mathbf{x}_i\right) = \mu(\mathbf{x}_i)$;
2. the tower property holds, $\text{Proj}\left(y_i \mid \mathbf{x}_i\right) = \text{Proj}\Big(\text{Proj}\left(y_i \mid \mathbf{x}_i, \mathbf{z}_i\right) \mid \mathbf{x}_i\Big)$.

From the previous discussion, we can posit that

$$y_i = \mathbf{x}_i^\intercal\beta + u_i,$$

which is simply a decomposition of the dependent variable into a linear predictor component $\mathbf{x}_i^\intercal\beta$ and a prediction error term u_i for $u_i \equiv y_i - \text{Proj}\left(y_i \mid \mathbf{x}_i\right)$. And u_i satisfies the following two properties $\mathbb{E}[u_i] = 0$ and $\mathbb{E}\left[u_i \mid \mathbf{x}_i\right] = 0$.

Example 1.3 (Freedman, 2008; Lin, 2013)**.** The estimation of the average treatment effect can also be cast in this regression framework. Let $\left(y_i(1), y_i(0)\right)_{i=1}^n$ be fixed potential outcomes. Let d_i be a treatment indicator such that $d_i \perp\!\!\!\perp \left(y_i(1), y_i(0)\right)$ for all i and that

$$y_i = d_i y_i(1) + (1 - d_i)y_i(0).$$

Rearranging the assignment equation, we have

$$y_i = d_i\left(y_i(1) - y_i(0)\right) + y_i(0) = d_i\tau_i + \varepsilon_i,$$

where $\tau_i = y_i(1) - y_i(0)$ and $\varepsilon_i = y_i(0)$. This equation still deviates from the regression framework in two fronts. First the coefficient on the explanatory variable d_i is usually fixed whilst $\tau_i \neq \tau_j$ in general. Secondly note that $n^{-1}\sum_{i=1}^n \varepsilon_i = n^{-1}\sum_{i=1}^n y_i(0) \neq 0$. To fit the regression framework, we rewrite the equation as

$$\begin{aligned}
y_i &= \beta_0 + \beta_1(d_i - \bar{d}) + u_i,\\
\beta_0 &= \bar{d}\,\bar{y}(1) + (1 - \bar{d})\bar{y}(0),\\
\beta_1 &= \bar{y}(1) - \bar{y}(0),\\
\beta_{1i} &= \left(y_i(1) - \bar{y}(1)\right) - \left(y_i(0) - \bar{y}(0)\right),\\
\beta_{0i} &= \bar{d}\left(y_i(1) - \bar{y}(1)\right) + (1 - \bar{d})\left(y_i(0) - \bar{y}(0)\right),\\
u_i &= \beta_{0i} + \beta_{1i}\left(d_i - \bar{d}\right),
\end{aligned}$$

for $\bar{d} = n^{-1}\sum_{i=1}^n d_i$, $\bar{y}(1) = n^{-1}\sum_{i=1}^n y_i(1)$ and $\bar{y}(0) = n^{-1}\sum_{i=1}^n y_i(0)$. The centering of d_i renders the coefficient β_0 more interpretable. Without the centering, β_0 becomes $(1 - \bar{d})\bar{y}(0)$. Observe now that $\sum_{i=1}^n \beta_{0i} = \sum_{i=1}^n \beta_{1i} = 0$ by construction, and $\mathbb{E}\left[u_i \mid d_1, \ldots, d_n\right] = \beta_{0i}$ and $\mathbb{E}\left[\sum_{i=1}^n u_i \mid d_1, \ldots, d_n\right] = 0$, so that this formulation satisfies certain "weak" form of regression conditions. We can now write the least square estimator of y_i on $1, (d_i - \bar{d})$ as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \left(d_i - \bar{d}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^n \left(d_i - \bar{d}\right)^2} = \beta_1 + \frac{\sum_{i=1}^n \left(d_i - \bar{d}\right)u_i}{\sum_{i=1}^n \left(d_i - \bar{d}\right)^2},$$

where the second term is interpreted as errors from not observing all $y_i(1), y_i(0)$.

The formulation in this section implies that the least square procedure reduces the potentially very complicated structure of the data into a lower dimensional parametric model capable of being directly interpreted. Even if this low

dimensional model comes at a cost of misspecification, it is still guaranteed to provide certain types of approximations which we can interpret mechanically without knowing the exact misspecification.

2 Regression Algebra

Given a sample of $\left(y_i, \mathbf{x}_i\right)_{i=1}^n$, we could estimate the predictor coefficient β by

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\intercal\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i y_i\right) = \left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\left(\mathbf{X}^\intercal\mathbf{y}\right),$$

which is solution to the following optimization problem

$$\hat{\beta} = \underset{b\in\mathbb{R}^k}{\arg\min}\left(\mathbf{y}-\mathbf{X}b\right)^\intercal\left(\mathbf{y}-\mathbf{X}b\right) = \underset{b\in\mathbb{R}^k}{\arg\min}\sum_{i=1}^n\left(y_i - \mathbf{x}_i^\intercal b\right)^2.$$

We now define two matrices for understanding matrix algebra, the projection matrix $\mathbf{P}_\mathbf{X} \equiv \mathbf{X}\left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal$, and the annihilator matrix $\mathbf{M}_\mathbf{X} = \mathbf{I}_n - \mathbf{P}_\mathbf{X}$.[1] These two matrices have special properties.

Exercise 2.A

Prove that $\mathbf{P}_\mathbf{X}$ and $\mathbf{M}_\mathbf{X}$ are idempotent and symmetric, and $\mathbf{P}_\mathbf{X}\mathbf{M}_\mathbf{X} = 0$.

It is immediate to see that $\mathbf{P}_\mathbf{X}\mathbf{X} = \mathbf{X}$ and $\mathbf{M}_\mathbf{X}\mathbf{X} = 0$. More importantly, we have

$$\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{X}\hat{\beta} = \hat{\mathbf{y}} \text{ and } \mathbf{M}_\mathbf{x}\mathbf{y} = \mathbf{y} - \hat{\mathbf{y}} = \hat{\mathbf{u}},$$

which gives respectively the fitted values and the residuals from the regression.

2.1 Frisch-Waugh-Lovell Theorem

Let us partition the regressor matrix $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$ into two parts $\mathbf{X}_1$, $n\times k_1$ matrix, and $\mathbf{X}_2$, $n\times k_2$ matrix, such that $k = k_1 + k_2$. The regression model can now be written as

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u}. \tag{2.1}$$

Define $\mathbf{P}_1 \equiv \mathbf{X}_1\left(\mathbf{X}_1^\intercal\mathbf{X}_1\right)^{-1}\mathbf{X}_1^\intercal$ and $\mathbf{P}_2 \equiv \mathbf{X}_2\left(\mathbf{X}_2^\intercal\mathbf{X}_2\right)^{-1}\mathbf{X}_2^\intercal$, and the associated annihilator matrix $\mathbf{M}_1$ and $\mathbf{M}_2$. If we pre-multiply the regression equation by $\mathbf{M}_1$, we have

$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\beta_2 + \mathbf{M}_1\mathbf{u}. \tag{2.2}$$

We can now state the Frisch-Waugh-Lovell theorem.

Theorem 2.1. *The regression coefficients $\hat{\beta}_2$ from Equation (2.1) are identical to the regression coefficients $\tilde{\beta}_2$ from Equation (2.2). The residuals obtained from the respective regressions are identical.*

Proof. We first write out the equation for $\tilde{\beta}_2$ as

$$\tilde{\beta}_2 = \left(\mathbf{X}_2^\intercal\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2^\intercal\mathbf{M}_1\mathbf{y}.$$

By regression algebra we can express $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \hat{\mathbf{u}} = \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \mathbf{M}_\mathbf{X}\mathbf{y}.$$

[1] The motivation of the name of $\mathbf{P}_\mathbf{X}$ can be seen as the result of orthogonal projection. Alternatively we can use singular value decomposition to under $\mathbf{P}_\mathbf{X}$. By SVD, we write $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\intercal$, where $\mathbf{U}$ is $n\times k$ matrix, Σ is $k\times k$ diagonal matrix of singular values of $\mathbf{X}$ and $\mathbf{V}$ is $k\times k$ matrix, and that $\mathbf{U}^\intercal\mathbf{U} = \mathbf{V}^\intercal\mathbf{V} = \mathbf{I}_k$. Using this formulation, we see that $\mathbf{P}_\mathbf{X} = \mathbf{U}\mathbf{U}^\intercal$. The projection matrix takes in any vector $y\in\mathbb{R}^n$ and apply $\mathbf{U}^\intercal y$ to find coefficients that could express y using the basis of the column space of $\mathbf{X}$ and then return it as an element in the column space of $\mathbf{X}$.

Premultiply both sides by $\mathbf{X}_2^\intercal \mathbf{M}_1$ to get

$$\mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{y} = \mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2 + \mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{X}_2 \mathbf{y} = \mathbf{X}_2^\intercal \mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2,$$

where the second line follows from that fact that $\mathbf{M}_2$ annihilates $\mathbf{X}_2$. The first claim that $\hat{\beta}_2 = \tilde{\beta}_2$ now follows.

For the second claim, observe that

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \tilde{\beta}_2 + \mathbf{M}_1 \mathbf{M}_\mathbf{X} \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \tilde{\beta}_2 + \mathbf{M}_\mathbf{X} \mathbf{y},$$

so the residuals are $\mathbf{M}_\mathbf{X}\mathbf{y}$ which is identical to those residuals from the regression of Equation (2.1). □

The Frisch-Waugh-Lovell theorem is an elegant description of the anatomy of multiple regression. Loosely speaking, the regression coefficient $\hat{\beta}_2$ is interpreted as "the effect" of $\mathbf{X}_2$ on $\mathbf{y}$, holding $\mathbf{X}_1$ constant (or controlling for $\mathbf{X}_1$). This theorem makes clear of this statement through a two step procedure:

- **Step 1**: Regress the columns of $\mathbf{X}_2$ onto the columns of $\mathbf{X}_1$:

  $$\mathbf{X}_2 = \mathbf{X}_1 \gamma + \eta$$

  From the usual OLS formulas, we know that $\widehat{\gamma} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2$ and that the residuals from this regression are given by:

  $$\begin{aligned} \hat{\mathbf{u}}_2 &= \mathbf{X}_2 - \mathbf{X}_1 \hat{\gamma} \\ &= (\mathbf{I} - \mathbf{P_1}) \mathbf{X}_2. \end{aligned}$$

- **Step 2**: Regress $\mathbf{y}$ onto $\hat{\mathbf{u}}_2$:

  $$\mathbf{y} = \hat{\mathbf{u}}_2 \delta + \nu$$

  From this second regression, we get the following formula for $\widehat{\delta}$:

  $$\begin{aligned} \widehat{\delta} &= (\hat{\mathbf{u}}_2^\intercal \hat{\mathbf{u}}_2)^{-1} \hat{\mathbf{u}}_2^\intercal \mathbf{y} \\ \widehat{\delta} &= (\mathbf{X}_2^\intercal (\mathbf{I} - \mathbf{P_1}) \mathbf{X}_2)^{-1} \mathbf{X}_2^\intercal (\mathbf{I} - \mathbf{P_1}) \mathbf{y} \end{aligned}$$

  The Frisch-Waugh-Lovell theorem states that $\hat{\beta}_2$ is equal to $\hat{\delta}$.

So, what this says is that if we want to estimate the effect of $\mathbf{X}_2$ on $\mathbf{y}$, controlling for $\mathbf{X}_1$, we first partial out the effect of $\mathbf{X_1}$ on $\mathbf{X}_2$, and we regress $\mathbf{y}$ on only the component of $\mathbf{X_2}$ that is uncorrelated with $\mathbf{X_1}$.

References

Freedman, David A. 2008. "On regression adjustments in experiments with several treatments." *Ann. Appl. Stat.*, 2(1): 176–196.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *Ann. Appl. Stat.*, 7(1): 295–318.