Regression I

Ye Wang
University of North Carolina at Chapel Hill

Mathematics and Statistics For Political Research
POLI783

From CEF to regression

- In social science, we are interested in how multiple variables change together (e.g., democracy and prosperity).
- It can be described by their joint distribution.
- It is usually difficult to learn the entire distribution.
- We instead investigate how the expectation of one variable Y varies with other variables $\mathbf{X} \in \mathbb{R}^P$, which is Y's CEF given $\mathbf{X}$:

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

- $\mu(\mathbf{X})$ is the best predictor of Y in the MSE sense.
- The functional form of $\mu(\mathbf{X})$ is determined by $F(Y, \mathbf{X})$ and can be arbitrary.
- Nevertheless, when $F(Y, \mathbf{X})$ is jointly normal, $\mu(\mathbf{X})$ is linear:

$$\mu(\mathbf{X}) = \mathbf{X}' \left(\mathbb{E}[\mathbf{X}'\mathbf{X}]\right)^{-1} \mathbb{E}[\mathbf{X}'Y] = \mathbf{X}'\beta,$$

assuming that both Y and $\mathbf{X}$ are mean-zero.

From CEF to regression

- Remember that the error of the CEF is defined as $e_\mu = Y - \mu(\mathbf{X})$, with $\mathbb{E}\left[e_\mu \mid \mathbf{X}\right] = 0$.
- Therefore, for an i.i.d. sample drawn from this joint distribution

$$Y_i = \mathbf{X}_i' \beta + \varepsilon_i = \sum_{p=1}^{P} X_{ip}\beta_p + \varepsilon_i.$$

- Following the convention, we represent the error for unit i, $e_{\mu i}$, with ε_i, and $\mathbb{E}\left[\varepsilon_i \mid \mathbf{X}_i\right] = 0$.
- This is known as a linear regression model.
- When $F(Y, \mathbf{X})$ is not jointly normal, we can still use $\mathbf{X}'\beta$ as the first-order approximate of $\mu(\mathbf{X})$.

From CEF to regression

- We usually include the constant vector $\iota = (1, 1, \ldots, 1)'$ in $\mathbf{X}$ and an intercept β_0 in β:

$$\underbrace{\mathbf{X}}_{N\times(P+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1P} \\ 1 & x_{21} & x_{22} & \ldots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{N1} & x_{N2} & \ldots & x_{NP} \end{pmatrix}, \quad \underbrace{\beta}_{(P+1)\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{pmatrix}.$$

- Note that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)'$.
- We can write the linear regression model as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

with $\underbrace{\mathbf{Y}}_{N\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}$ and $\underbrace{\varepsilon}_{N\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$.

From CEF to regression

- There is no guarantee that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] = 0$.
- ε_i may include higher-order terms of $\mu(\mathbf{X})$'s Taylor expansion:

$$\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] = \mathbb{E}\left[Y_i - \mathbf{X}_i'\beta \mid \mathbf{X}_i\right] = \mu(\mathbf{X}_i) - \mathbf{X}_i'\beta.$$

- Nevertheless, we can always include higher-order terms into $\mathbf{X}_i$.
- Results obtained from the wrong model can still be informative.
- Therefore, the following linear regression model is still widely used in social science:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$
$$\mathbb{E}[\varepsilon_i|\mathbf{X}_i] = 0.$$

- The second part is sometimes known as "exogeneity."
- This is our DGP, and we want to estimate the coefficients β and study the estimator's sampling distribution.
- We may also test hypothesis regarding β.

Bivariate regression

- We start from the simple case with $P = 1$:

$$
\begin{aligned}
Y_i &= \mu + \tau D_i + \varepsilon_i, \\
\mathbb{E}[\varepsilon_i | D_i] &= 0.
\end{aligned}
$$

- Y_i: the outcome, the response, the dependent variable, the label.
- D_i: the treatment, the regressor/predictor, the independent variable, the feature.
- What have we assumed (and not assumed) in this model?
- A linear relationship between Y and D and a constant effect.
- No variable can be correlated with both Y and D.
- The variance is potentially heteroscedastic: $Var(\varepsilon_i | D_i) = \sigma_i^2$.
- No requirement on the error term's distribution.

Bivariate regression

- Our estimands are the two parameters, μ and τ.
- We estimate the unknown parameters through solving the minimization problem:

$$(\hat{\mu}, \hat{\tau})' = \arg\min_{\mu,\tau} \sum_{i=1}^{N} (Y_i - \mu - \tau D_i)^2.$$

- Here, the objective function is a sample analogue of the MSE.
- It results in the following estimator:

$$\hat{\tau} = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})(D_i - \bar{D})}{\sum_{i=1}^{N} (D_i - \bar{D})^2}$$
$$\hat{\mu} = \bar{Y} - \hat{\tau}\bar{D}.$$

- This is known as the ordinary least squares (OLS) method.
- The estimator is independent to the model we use.
- We can derive the same estimator through MLE.

Bivariate regression

- Define $f(\mu, \tau) = \sum_{i=1}^{N}(Y_i - \mu - \tau D_i)^2$, we can see that

$$\frac{\partial f(\mu, \tau)}{\partial \mu} = -2\sum_{i=1}^{N}(Y_i - \mu - \tau D_i),$$
$$\frac{\partial f(\mu, \tau)}{\partial \tau} = -2\sum_{i=1}^{N} D_i(Y_i - \mu - \tau D_i).$$

- The first order conditions lead to the estimators.
- Then, we predict the outcome with $\hat{Y}_i = \hat{\mu} + \hat{\tau} D_i$.
- The regression residual is $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ and $\sum_{i=1}^{N} \hat{\varepsilon}_i^2$ is called the sum of squared residuals (SSR).
- We define the total sum of squares (SST) as $\sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2$, which is proportional to Y_i's sample variance.
- $R^2 = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST}$ measures the prediction power of the regressor(s).
- It is bounded between 0 and 1 and captures the proportion of Y's variance that can be explained by the regressors.

Properties of the OLS estimator

- We focus on the properties of $\hat{\tau}$:

$$\begin{aligned}
\hat{\tau} &= \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(D_i - \bar{D})}{\sum_{i=1}^{N}(D_i - \bar{D})^2} \\
&= \frac{\sum_{i=1}^{N}(\tau(D_i - \bar{D}) + \varepsilon_i - \bar{\varepsilon})(D_i - \bar{D})}{\sum_{i=1}^{N}(D_i - \bar{D})^2} \\
&= \tau + \frac{\sum_{i=1}^{N}(\varepsilon_i - \bar{\varepsilon})(D_i - \bar{D})}{\sum_{i=1}^{N}(D_i - \bar{D})^2}.
\end{aligned}$$

- We can see that $\mathbb{E}[\hat{\tau}] = \tau$.
- $\lim_{N\to\infty} \hat{\tau} = \tau$ when conditions for the law of large numbers are satisfied.

Multivariate regression

- Now, let's consider the multivariate regression model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$
$$\mathbb{E}[\varepsilon_i | \mathbf{X}_i] = 0.$$

- In bivariate regression, $\mathbf{X}_i = (1, D_i)'$ and $\beta = (\mu, \tau)'$.
- Similarly, we estimate β by solving the minimization problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - \mathbf{X}_i'\beta)^2.$$

- We treat $\sum_{i=1}^{N} (Y_i - \mathbf{X}_i'\beta)^2$ as a function of β: $f(\beta) = \sum_{i=1}^{N} (Y_i - \mathbf{X}_i'\beta)^2$.
- We can find $\hat{\beta}$ that minimizes $f(\beta)$ with matrix calculus.

Multivariate regression

- Taking the derivative of $f(\beta)$ with regards to β, we have

$$\begin{aligned}
\frac{df(\beta)}{d\beta} &= \sum_{i=1}^{N} \frac{d(Y_i - \mathbf{X}_i'\beta)^2}{d\beta} \\
&= \sum_{i=1}^{N} 2(Y_i - \mathbf{X}_i'\beta)\frac{d(Y_i - \mathbf{X}_i'\beta)}{d\beta} \\
&= \sum_{i=1}^{N} 2(Y_i - \mathbf{X}_i'\beta)\mathbf{X}_i
\end{aligned}$$

- The first-order condition is

$$2\sum_{i=1}^{N} \mathbf{X}_i(Y_i - \mathbf{X}_i'\hat{\beta}) = 0.$$

- It leads to

$$\sum_{i=1}^{N} \mathbf{X}_i Y_i = \mathbf{X}'\mathbf{Y} = \sum_{i=1}^{N} \mathbf{X}_i\mathbf{X}_i'\hat{\beta} = \mathbf{X}'\mathbf{X}\hat{\beta}.$$

Multivariate regression

- Multiplying $(\mathbf{X}'\mathbf{X})^{-1}$ to both sides, we can see that

$$\hat{\beta} = \left(\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i Y_i\right) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}).$$

- $\hat{\beta}$ is a linear transformation of $\mathbf{Y}$.
- The predicted outcome equals $\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$.
- $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is known as the projection matrix.
- It transforms $\mathbf{Y}$ to an element in the space spanned by $\mathbf{X}$, $\hat{\mathbf{Y}}$.
- Note that $\mathbf{P}$ is symmetric and $\mathbf{P}^2 = \mathbf{P}'\mathbf{P} = \mathbf{P}$.
- $\mathbf{P}$ is known as an idempotent matrix.
- What is the value of $\mathbf{PX}$?
- Each diagonal element, P_{ii}, is called the leverage of unit i.

Multivariate regression

- $\mathbf{Q} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is known as the residual-making matrix, where $\mathbf{I}$ is the identity matrix.
- $\mathbf{Q}$ is also an idempotent matrix, and $\mathbf{QP} = \mathbf{PQ} = \mathbf{P} - \mathbf{P} = \mathbf{0}$.
- We can see that

$$\begin{aligned}\mathbf{QY} =& \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ =& \mathbf{Y} - \mathbf{X}\beta \\ =& \mathbf{Y} - \hat{\mathbf{Y}} = \hat{\varepsilon},\end{aligned}$$

  where $\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_N)'$ is the vector of regression residuals.
- Note that $\hat{\mathbf{Y}}'\hat{\varepsilon} = \mathbf{Y}'\mathbf{PQY} = 0$.
- The SSR equals $\hat{\varepsilon}'\hat{\varepsilon} = \mathbf{Y}'\mathbf{QY}$ and $R^2 = 1 - \frac{SSR}{STT}$ remains bounded between 0 and 1, since

$$SST = \mathbf{Y}'\mathbf{Y} = \left(\hat{\mathbf{Y}} + \hat{\varepsilon}\right)'\left(\hat{\mathbf{Y}} + \hat{\varepsilon}\right) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\varepsilon}'\hat{\varepsilon} \geq SSR.$$

Multivariate regression: properties

- As before, we plug in the regression equation, and obtain

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\varepsilon) \\ &= \beta + \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\varepsilon_i\right)\end{aligned}$$

- It is straightforward to see that $\mathbb{E}[\hat{\beta}] = \beta$, and as $N \to \infty$,

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i' \to \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right], \\ &\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\varepsilon_i \to \mathbb{E}\left[\mathbf{X}_i\varepsilon_i\right] = \mathbb{E}\left[\mathbb{E}\left[\varepsilon_i \mid \mathbf{X}_i\right]\mathbf{X}_i\right] = 0.\end{aligned}$$

- $\hat{\beta}$ is an unbiased and consistent estimator for β.

Multivariate regression: omitted variables

- Suppose the true DGP is

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \delta\mathbf{U} + \varepsilon, \\ \mathbb{E}[\varepsilon_i|\mathbf{X}_i, U_i] &= 0. \end{aligned}$$

- But U_i is not controlled by the researcher when fitting the regression model.
- Now, we can see that

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \\ &= \beta + \delta(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'U) + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\varepsilon_i) \\ &\to \beta + \delta\gamma, \end{aligned}$$

  where γ is the limit of the OLS estimate when regressing $\mathbf{U}$ on $\mathbf{X}$.
- U_i is often referred to as the "omitted variable."

Multivariate regression: omitted variables

- The asymptotic bias of the OLS estimator $\hat{\beta}$ equals

$$\lim_{N\to\infty}(\hat{\beta}-\beta)=\delta\gamma,$$

  which is known as the "omitted variable bias (OVB)."
- The bias equals zero when either δ or γ equals zero.
- No OVB when U_i is uncorrelated with either Y_i or $\mathbf{X}_i$.
- We will see that this logic generalizes to cases where linear models fail.

Multivariate regression: simulation

```
## The regression estimates are 3.915022 -3.049459 5.342146
## The regression estimates are 3.915022 -3.049459 5.342146
## [1]  4.002375 -2.996764  4.997096
```

References I