Regression I

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- We can rely on either the asymptotic approach or resampling techniques for statistical inference.
- The latter includes Fisher's randomization test, bootstrap, and jackknife.
- The attraction is that we may avoid technical details such as calculating the variance or obtaining critical values.
- But the FRT only works under the sharp null.
- Bootstrap requires a smooth estimator.
- The Efron method works only when the true distribution is symmetric.
- The percentile-t method provides the best approximation as the t-statistic is pivotal.

Bivariate regression

- We have been familiar with the linear regression model with one predictor:
$$Y_i = \mu + \tau D_i + \varepsilon_i,$$
$$E[\varepsilon_i | D_i] = 0.$$
- 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 confounder and potentially heteroscedasticity: $Var(\varepsilon_i | D_i) = \sigma_i^2$.
- No requirement on the error term's distribution.

Bivariate regression

- The regression coefficients can be estimated via

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

- They are solutions to the minimization problem:

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

- This is known as the ordinary least squares (OLS) method.
- The estimator is independent to the model we use.

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).
- $R^2 = \frac{Var[Y_i] - SSR}{Var[Y_i]}$ measures the prediction power of the regressor(s).

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

Bivariate regression in practice

- Remember that the coefficient τ tells us the change in Y when D increases by 1 unit.
- It makes more sense when Y is continuous and D is either binary or continuous.
- When Y is binary, we call the regression model the "linear probability model."
- We interpret τ as the effect of D on the probability for Y to be 1.
- One concern is that the predicted outcome may be beyond the range of $[0, 1]$.
- We can fix this problem by using alternative models such as Probit or Logit.
- But the linear probability model is Ok if you don't care about prediction.

Bivariate regression in practice

- When Y is categorical or a count variable, a τ units increase in it is hard to interpret.
- We may respectively use multinomial logit and count models, such as the Poisson model or the negative binomial model.
- No model is more correct than the others, and you should choose the one that facilitates your interpretation.
- When D is categorical, it is better to include dummies standing for each of the category as regressors.
- It is also common to transform Y to $\log Y$, then

$$\tau = \frac{d \log Y}{dD} = \frac{1}{Y}\frac{dY}{dD} \approx \frac{\Delta Y}{Y}.$$

- The coefficient can be interpreted as the change of Y in percentages as X increases by 1 unit.
- This is known as elasticity in economics.

Bivariate regression in practice

- When Y may take the value of 0, we replace $\log Y$ with $\log(Y+1)$ or $\log(Y+\sqrt{Y^2+1})$ (the inverse hyperbolic sine transformation).
- They behave in very similar ways.
- But it is crucial to understand what 0 stands for.
- If your thermometer toward Trump is 0, maybe you just hate him.
- If your monthly income is 0, it may suggest you are not on the labor market.
- In the latter case, $\log(Y+1)$ is not appropriate if there are many 0s in data (Chen and Roth 2023).
- The change from 0 to 1 (the extensive margin) is very different from that from 1 to 2 (the intensive margin).
- We know that for any positive number c, $\log(cY+1) \approx \log c + \log Y$.
- The magnitude of the extensive margin effect can be driven by Y's unit.

Multivariate regression

- Now, let's consider the multivariate regression model

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

where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)'$, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)'$.
- Note that $\mathbf{X}_i$ is a $P \times 1$ vector, hence $\mathbf{X}$ is a $N \times P$ matrix.
- In bivariate regression, $\mathbf{X}_i = (1, D_i)'$ and $\beta = (\mu, \tau)'$.
- Similarly, we estimate β by solving the minimization problem

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

Multivariate regression

- We treat $\sum_{i=1}^{N}(Y_i - \mathbf{X}_i'\beta)^2$ as a function of β: $f(\beta) = \sum_{i=1}^{N}(Y_i - \mathbf{X}_i'\beta)^2$.
- The goal is to find $\hat{\beta}$ that minimizes $f(\beta)$, which can be done via taking the derivative of $f(\beta)$ with regards to β.
- We need some rules to compute the derivative with regards to a vector.
- For any function $f(\beta)$, where β is a column-vector, we require that $\frac{df(\beta)}{d\beta}$ is also a column-vector.

Multivariate regression

- With this rule in mind, 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 clearly a linear estimator.
- 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 $\mathbf{X}$, $\hat{\mathbf{Y}}$.
- 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

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

- 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 $\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_N)'$ is the vector of regression residuals.

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 $E[\hat{\beta}] = \beta$, and as $N \to \infty$,

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

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

Multivariate regression: omitted variables

- Suppose the true DGP is

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

- 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 3.951469 -3.008785 5.050837
## The regression estimates are 3.951469 -3.008785 5.050837
## [1]  4.005401 -3.003982  4.992981
```

References I

Chen, Jiafeng, and Jonathan Roth. 2023. "Logs with Zeros? Some Problems and Solutions." *The Quarterly Journal of Economics*, qjad054.