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.
- It is an estimator that 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 $\log(Y+\sqrt{Y^2+1})$.
- 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

- The first-order condition is

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

- It leads to

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

- $\hat{\beta}$ 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 $\mathbf{X}$, $\hat{\mathbf{Y}}$.
- Each diagonal element, P_{ii}, is called the leverage of unit i.

Multivariate regression

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

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

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

$$\begin{aligned}Var\left[\hat{\beta}\right] &= Var\left[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}\varepsilon)\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}\varepsilon\varepsilon'\mathbf{X}')(\mathbf{X}'\mathbf{X})^{-1}\right] \\ &\to \mathbf{0}.\end{aligned}$$

- Note that $Var\left[\hat{\beta}\right]$ is a $P \times P$ matrix (the variance-covariance matrix).
- Hence, $\hat{\beta} \to \beta$ when $N \to \infty$.

Inference in multivariate regression

- Define the vector of regression residuals as $\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N)'$, where $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i'\hat{\beta}$.
- We can estimate the variance of $\hat{\beta}$ using

$$\widehat{Var}\left[\hat{\beta}\right] = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}\hat{\Sigma}\mathbf{X}')(\mathbf{X}'\mathbf{X})^{-1},$$

  where $\hat{\Sigma} = \hat{\varepsilon}\hat{\varepsilon}'$.
- This is known as the sandwich variance estimator.
- Since the units are independent to each other, we impose the constraint that $\hat{\Sigma}$ is diagonal, hence $\mathbf{X}\hat{\Sigma}\mathbf{X}' = \sum_{i=1}^{N} \hat{\varepsilon}_i^2 \mathbf{X}_i \mathbf{X}_i'$.
- This is the Eicker-Huber-White (EHW) robust variance estimator.
- Under homoscedasticity, $E[\varepsilon_i^2 | \mathbf{X}_i] = \sigma^2$ for any i, and $Var\left[\hat{\beta}\right] = \sigma^2 E\left[(\mathbf{X}'\mathbf{X})^{-1}\right]$.
- The sandwich variance estimator can then be simplified to $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, where $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N} \hat{\varepsilon}_i^2$.

Inference in multivariate regression

- It is easy to show that

$$\sqrt{N}(\hat{\beta} - \beta) \to \mathcal{N}\left(0, NVar\left[\hat{\beta}\right]\right).$$

- Hence, we can construct the 95% confidence interval of any element in β as

$$\left[\hat{\beta}_p - 1.96 * \sqrt{\widehat{Var}\left[\hat{\beta}_p\right]}, \hat{\beta}_p + 1.96 * \sqrt{\widehat{Var}\left[\hat{\beta}_p\right]}\right].$$

- In theory, the coverage rate should be 95%.
- But in practice, it is usually much lower than that (the Behrens–Fisher problem).

Inference in multivariate regression (*)

- We do know that $\frac{\hat{\beta}_p - \beta_p}{\sqrt{Var\left[\hat{\beta}_p\right]}}$ converges to normality at the root-N rate.
- But we replace the denominator with an estimate, which creates complex asymptotics in the statistic.
- When ε is normal, we know that $\frac{\hat{\beta}_p - \beta_p}{\sqrt{\widehat{Var}\left[\hat{\beta}_p\right]}}$ **obeys** the t-distribution.
- Using critical values from the normal distribution causes bias.
- After all, asymptotic distribution is an approximation!

Inference in multivariate regression (*)

- Multiple solutions have been proposed (but never welcomed).
- We can modify the variance estimate or the critical value.
- There are multiple variance estimators.
- HC1: multiply $\widehat{Var}\left[\hat{\beta}\right]$ by $\frac{N}{N-P+1}$.
- HC2: replace each $\hat{\varepsilon}_i$ with $\frac{\hat{\varepsilon}_i}{\sqrt{1-P_{ii}}}$, where P_{ii} is the (i,i)th entry of the projection matrix.
- HC3: replace each $\hat{\varepsilon}_i$ with $\frac{\hat{\varepsilon}_i}{1-P_{ii}}$.
- We can use the critical value from the t-distribution rather than the normal distribution.
- The t-distribution requires researchers to specify the degree of freedom of the model.
- See Imbens and Kolesar (2016) for technical details.

Hypothesis testing in multivariate regression

- The regression model enables us to test hypothesis regarding a linear combination of β.
- They usually take the form of $\mathbf{R}\beta = \mathbf{r}$, where $\mathbf{R}$ is a $R \times P$ matrix.
- For example, when $P = 3$ and the null hypothesis is $\beta_1 + \beta_2 = 0$ and $\beta_3 = 0$,

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

- How do we test the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$?

Hypothesis testing in multivariate regression

- Using the asymptotic normality of $\hat{\beta}$, we know that

$$\begin{aligned}\sqrt{N}(\mathbf{R}\hat{\beta} - \mathbf{R}\beta) &= \sqrt{N}(\mathbf{R}\hat{\beta} - \mathbf{r}) \\ &\rightarrow \mathcal{N}\left(0, N\mathbf{R} * Var\left[\hat{\beta}\right]\mathbf{R}'\right).\end{aligned}$$

- Therefore, the Wald statistic

$$W = (\mathbf{R}\hat{\beta} - \mathbf{r})'\left(\mathbf{R} * Var\left[\hat{\beta}\right]\mathbf{R}'\right)^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}) \rightarrow \chi^2(R).$$

- We reject the null hypothesis if W is sufficiently large.
- The Wald test is equivalent to the F-test under homoscedasticity, as

$$F = \frac{W}{R} \sim F(R, N - P).$$

References I

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

Imbens, Guido W, and Michal Kolesar. 2016. "Robust Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics* 98 (4): 701–12.