Regression II

Ye Wang
University of North Carolina at Chapel Hill

Mathematics and Statistics For Political Research
POLI783

Inference in multivariate regression

- Now, let's examine the variance of $\hat{\beta}$:

$$
\begin{aligned}
&\text{Var}\left[\hat{\beta}\right] = \text{Var}\left[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\varepsilon)\right] \\
=&\mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\varepsilon\varepsilon'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\right] \\
=&\frac{1}{N}\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\varepsilon_i^2\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\right].
\end{aligned}
$$

- $\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\varepsilon_i^2 \to \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\varepsilon_i^2\right] = \sigma^2\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right]$ if $\text{Var}\left[\varepsilon_i\right] = \sigma^2$.
- Then,

$$
\begin{aligned}
N * \text{Var}\left[\hat{\beta}\right] \to &\left(\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right]\right)^{-1}\sigma^2\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right]\left(\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right]\right)^{-1} \\
=&\sigma^2\left(\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right]\right)^{-1}.
\end{aligned}
$$

- $N * \text{Var}\left[\hat{\beta}\right]$ converges to a $(P+1)\times(P+1)$ matrix (the variance-covariance matrix).
- $\hat{\beta} \to \beta$ when $N \to \infty$.

Best linear unbiased estimator

- When $\text{Var}[\varepsilon_i] = \sigma^2$, we can show that the OLS estimator is the best linear unbiased estimator (BLUE).
- Consider another linear estimator $\tilde{\beta} = \mathbf{AY}$, where $\mathbf{A}$ is a $(P+1) \times N$ matrix.
- For unbiasedness, we need $\mathbb{E}[\mathbf{AY}] = \mathbb{E}[\mathbf{AX}\beta] = \beta$.
- Therefore, $\mathbb{E}[\mathbf{AX}] = \mathbf{I}$, and we can define $\check{\mathbf{A}} = \mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$.
- Then, $\mathbb{E}\left[\check{\mathbf{A}}\mathbf{X}\right] = \mathbf{I} - \mathbf{I} = \mathbf{0}$.
- Following the same logic, we can show that

$$\begin{aligned} N * \text{Var}\left[\tilde{\beta}\right] &= \mathbb{E}\left[\mathbf{A}\varepsilon\varepsilon'\mathbf{A}'\right] \\ \to N * \text{Var}\left[\hat{\beta}\right] &+ \sigma^2\mathbb{E}\left[\check{\mathbf{A}}_i\check{\mathbf{A}}_i'\right]. \end{aligned}$$

- $N * \text{Var}\left[\tilde{\beta}\right] - N * \text{Var}\left[\hat{\beta}\right]$ converges to a semi positive-definite matrix, thus $\hat{\beta}$ is more efficient.

Inference in multivariate regression

- Remember that the vector of regression residuals is $\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N)'$, where $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i'\hat{\beta}$.
- We can estimate $N * \text{Var}\left[\hat{\beta}\right]$ using its sample analogue:

$$\hat{\sigma}^2 = \frac{1}{N-P-1}\sum_{i=1}^{N}\hat{\varepsilon}_i^2,$$

$$\left(\widehat{\mathbb{E}}\left[\mathbf{X}_i\mathbf{X}_i'\right]\right)^{-1} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\right)^{-1} = \left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1}.$$

- $N - P - 1$ is known as the degree of freedom (dof) of the model, and it equals the trace of the residual-making matrix $\mathbf{Q}$.
- Therefore, $\widehat{\text{Var}}\left[\hat{\beta}\right] = \left(\frac{1}{N-P-1}\sum_{i=1}^{N}\hat{\varepsilon}_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}.$

Inference in multivariate regression: simulation

```
## The regression estimates are 4.185089 -2.975853 5.005425
## The regression standard error estimates are 0.2011642 0.
## The regression estimates are 4.185089 -2.975853 5.005425
## The regression standard error estimates are 0.2011642 0.
## The true standard errors are 0.2248701 0.07598621 0.1962
## The average regression standard error estimates are 0.20
```

Inference in multivariate regression

- In practice, heteroscedasticity is more common, and the previous estimator no longer works.
- In this case, we can estimate the variance of $\hat{\beta}$ using

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

  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.

Inference in multivariate regression

- It is easy to show that

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

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

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

- 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: simulation

```
## The regression estimates are 3.96365 -2.912993 5.643425
## The regression standard error estimates are 0.3586602 0.
## The robust standard error estimates are 0.2632776 0.1208
## The true standard errors are 0.2868046 0.1200374 0.5611
## The average regression standard error estimates are 0.35
## The average robust standard error estimates are 0.26327
```

Inference in multivariate regression (*)

- We do know that $\frac{\hat{\beta}_p - \beta_p}{\sqrt{\mathsf{Var}\left[\hat{\beta}_p\right]}}$ 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 $\frac{\hat{\beta}_p - \beta_p}{\sqrt{\widehat{\mathsf{Var}}\left[\hat{\beta}_p\right]}}$ **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 $\widehat{\text{Var}}\left[\hat{\beta}\right]$ 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.

Measurement error

- Our outcome and regressors may be measured with error:

$$Y_i = Y_i^* + e_{Yi},$$
$$\mathbf{X}_i = \mathbf{X}_i^* + \mathbf{e}_{\mathbf{X}i}.$$

- If e_{Yi} ($\mathbf{e}_{\mathbf{X}i}$) is independent to Y_i^* ($\mathbf{X}_i^*$), $\mathbb{E}[e_{Yi}] = 0$ ($\mathbb{E}[\mathbf{e}_{\mathbf{X}i}] = \mathbf{0}$), and $\mathrm{Var}[e_{Yi}] < \infty$ ($\mathrm{Var}[\mathbf{e}_{\mathbf{X}i}] < \infty$), it is known as the classical measurement error.
- In this case, our true regression model is $\mathbf{Y}^* = \mathbf{X}\beta + \varepsilon$ (or $\mathbf{Y} = \mathbf{X}^*\beta + \varepsilon$).
- With the classical measurement error only in Y, we have

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

- The OLS estimator remains consistent, but its variance becomes larger.

Measurement error

- With the classical measurement error only in $\mathbf{X}$, we have

$$\begin{aligned}\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right)\\ &= \left((\mathbf{X}^* + \mathbf{e}_{\mathbf{X}})'(\mathbf{X}^* + \mathbf{e}_{\mathbf{X}})\right)^{-1}\left((\mathbf{X}^* + \mathbf{e}_{\mathbf{X}})'(\mathbf{X}^*\beta + \varepsilon)\right)\\ &\to \left(\mathbb{E}\left[\mathbf{X}_i^*(\mathbf{X}_i^*)'\right] + \mathrm{Var}\left[\mathbf{e}_{\mathbf{X}i}\right]\right)^{-1}\mathbb{E}\left[\mathbf{X}_i^*(\mathbf{X}_i^*)'\right]\beta \neq \beta.\end{aligned}$$

- In the bivariate case,

$$\hat{\beta} \to \frac{\mathbb{E}\left[(D_i^*)^2\right]\beta}{\mathbb{E}\left[(D_i^*)^2\right] + \mathrm{Var}\left[\mathbf{e}_{Di}\right]} < \beta.$$

- This is known as the attenuation bias.

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 $R \times (P+1)$ matrix.
- R is the number of hypotheses.
- For example, when $P = 2$ and the null hypotheses are $\beta_0 + \beta_1 = 0$ and $\beta_2 = 0$,

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

Hypothesis testing in multivariate regression

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

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

- Therefore, the Wald statistic

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

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

$$F = \frac{W}{R} \sim F(R, N - P - 1).$$

Hypothesis testing in multivariate regression

- A specific null hypothesis is

$$H_0 : \beta_p = 0.$$

- How do we write it in the linear form?

$$\mathbf{R} = (0, \ldots, 1, \ldots, 0)' \text{ and } \mathbf{r} = 0$$

- In this case, the Wald statistic equals

$$W = \hat{\beta}_p \left(\mathrm{Var}\left[\hat{\beta}_p\right] \right)^{-1} \hat{\beta}_p = \frac{\hat{\beta}_p^2}{\mathrm{Var}\left[\hat{\beta}_p\right]}.$$

- $W \to \chi^2(R)$ and $\sqrt{W} \to \mathcal{N}(0, 1)$.

Hypothesis testing in multivariate regression

- Another specific null hypothesis is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_P = 0.$$

- How do we write it in the linear form?

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

- In this case, the Wald statistic equals

$$W = \hat{\beta}'_{-0} \left(\text{Var} \left[\hat{\beta}_{-0} \right] \right)^{-1} \hat{\beta}_{-0}.$$

- We reject the null hypothesis if $\frac{W}{R}$'s value is larger than the 95% threshold under the F distribution.

Hypothesis testing in multivariate regression: simulation

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1969 -1.0197  0.0980  0.8774  4.6045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7664     0.3868   9.738 9.85e-14 ***
## XX1          -0.3003     0.1394  -2.154   0.0355 *
## XX2           0.8222     0.3917   2.099   0.0402 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 1.725 on 57 degrees of freedom
## Multiple R-squared:  0.1303, Adjusted R-squared:  0.0997
```

References I

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.