Regression II

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- We have reviewed basic properties of the OLS estimator.
- We derived it using derivatives with regards to vectors.
- Suppose the regression model is correct, then it is unbiased and consistent for the regression coefficient.
- We can obtain predicted values using the projection matrix and regression residuals using the residual-making matrix.
- The omitted variable bias' magnitude is determined by the impact of the omitted variable on both the outcome and other regressors.

Inference in multivariate regression

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

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

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

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

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

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 $N * Var\left[\hat{\beta}\right]$ using its sample analogue:

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

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

- Therefore, $N * \widehat{Var}\left[\hat{\beta}\right] = \left(\frac{1}{N-P}\sum_{i=1}^{N}\hat{\varepsilon}_i^2\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$

Inference in multivariate regression: simulation

```
## The regression estimates are 3.963051 -2.920273 5.091297
## The regression standard error estimates are 0.2037381 0.
## The regression estimates are 3.963051 -2.920273 5.091297
## The regression standard error estimates are 0.2037381 0.
## The true standard errors are 0.2182289 0.07128418 0.1953
## The average regression standard error estimates are 0.20
```

Inference in multivariate regression

- 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.

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: simulation

```
## The regression estimates are 3.806672 -3.046103 5.89138
## The regression standard error estimates are 0.3994509 0.
## The robust standard error estimates are 0.3690726 0.1298
## The true standard errors are 0.3414864 0.1199145 0.4862
## The average regression standard error estimates are 0.39
## The average robust standard error estimates are 0.36907
```

Inference in multivariate regression (*)

- We do know that $\frac{\hat{\beta}_p - \beta_p}{\sqrt{Var[\hat{\beta}_p]}}$ 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 $\frac{\hat{\beta}_p - \beta_p}{\sqrt{\widehat{Var}[\hat{\beta}_p]}}$ **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.
- R is the number of hypotheses.
- For example, when $P = 3$ and the null hypotheses are $\beta_1 + \beta_2 = 0$ and $\beta_3 = 0$,

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

Hypothesis testing in multivariate regression

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

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

- Therefore, the Wald statistic

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

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

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

Hypothesis testing in multivariate regression

- A specific null hypothesis is

$$H_0 : \beta_p = 0.$$

- How do we write it in the linear form?

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

- In this case, the Wald statistic equals

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

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

Hypothesis testing in multivariate regression

- Another specific null hypothesis is

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_P = 0.$$

- 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

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

- 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
## -2.7287 -0.6187 -0.0555  0.5015  4.9659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.0534     0.2662  15.225  < 2e-16 ***
## XX1          -0.4312     0.0981  -4.395 4.88e-05 ***
## XX2           0.2465     0.2814   0.876    0.385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
## Residual standard error: 1.3 on 57 degrees of freedom
## Multiple R-squared:  0.2713, Adjusted R-squared:  0.2458
```

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.