OLS Regression Uncertainty

UC Berkeley, MIDS w203

Statistics for Data Science

February 11, 2022

Large-Sample Regression

We assume that the best linear predictor, $\mathscr{P}[Y|\boldsymbol{X}]$, of Y given $\boldsymbol{X}$ is $\boldsymbol{X}\beta$.

$$Y = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon.$$

We have from Theorem ref(thm:blperror)

$$\mathbb{E} = 0, \text{ and } \mathbb{E} = \mathbf{0}.$$

We also assume that the dataset $\{(Y_i, \boldsymbol{X}_i)\}$ is taken **i.i.d.** from the joint distribution of $(Y, \boldsymbol{X})$. For each i, we can write

$$Y_i = \boldsymbol{X_i}\beta + \varepsilon_i.$$

In matrix notation, we can write

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Then

$$\mathbb{E} = \mathbf{0}, \text{ and } \mathbb{E} = \mathbf{0}$$

We start by revealing an alternative expression for the OLS estimators $\widehat{\boldsymbol{\beta}}$ using matrix notation.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\boldsymbol{Y} \\
&= \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T(\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \\
&= \left[\mathbb{X}^T\mathbb{X}\right]^{-1}(\mathbb{X}^T\mathbb{X})\boldsymbol{\beta}+\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}+\left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\boldsymbol{\varepsilon}
\end{aligned}$$

So,

$$\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta} = \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\boldsymbol{\varepsilon} \tag{1}$$

We can then multiply by $\sqrt{n}$ both sides of Equation ref(eq:beta) to get

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i^T\varepsilon_i\right) \\
&= \widehat{\mathbb{Q}}_{\boldsymbol{XX}}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i^T\varepsilon_i\right)
\end{aligned}$$

From the consistency of OLS estimators, we already have

$$\widehat{\mathbb{Q}}_{\boldsymbol{XX}} \xrightarrow[p]{} \mathbb{Q}_{\boldsymbol{XX}}$$

Our aim now is to understand the distribution of the stochastic term (the second term) in the above expression.

We first note (from i.i.d. and Theorem ref(thm:blperror)) that

$$\mathbb{E} = \mathbb{E} = \mathbf{0}.$$

Let us compute the covariance matrix of $\boldsymbol{X}_i\varepsilon_i$. Since the expectation vector is zero, we have

$$\mathbb{V}[\boldsymbol{X}_i^T\varepsilon_i] = \mathbb{E} = \mathbb{E} \overset{\text{def}}{=} \mathbb{A}.$$

As any function of $\{(Y_i, \boldsymbol{X}_i)\}$'s are independent, $\{\boldsymbol{X}_i\varepsilon_i\}$'s are independent. By the (multivariate) **Central Limit Theorem**, as $n \to \infty$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i^T\varepsilon_i \xrightarrow[d]{} \mathcal{N}(\mathbf{0}, \mathbb{A}).$$

There is a small technicality here, we must have $\mathbb{A} < \infty$. This can be imposed by a stronger regularity condition on the moments, e.g., $\mathbb{E}, \mathbb{E} < \infty$. Putting everything together, we conclude

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow[d]{} \mathbb{Q}_{\boldsymbol{XX}}^{-1} \mathcal{N}(\mathbf{0}, \mathbb{A}) = \mathcal{N}\left(\mathbf{0}, \left[\mathbb{Q}_{\boldsymbol{XX}}^{-1}\right]^T \mathbb{A} \mathbb{Q}_{\boldsymbol{XX}}^{-1}\right) = \mathcal{N}\left(\mathbf{0}, \mathbb{Q}_{\boldsymbol{X}}^{-}\right.$$

We assume the following

1. The observations $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d from the joint distribution of $(Y, \boldsymbol{X})$
2. $\mathbb{E} < \infty$
3. $\mathbb{E} < \infty$
4. $\mathbb{Q}_{\boldsymbol{XX}} = \mathbb{E}$ is positive-definite.

Under these assumptions, as $n \to \infty$

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow[d]{} \mathcal{N}\left(\mathbf{0}, \mathbb{V}_{\beta}\right),$$

where

$$\mathbb{V}_{\boldsymbol{\beta}} \overset{\text{def}}{=} \mathbb{Q}_{\boldsymbol{XX}}^{-1} \mathbb{A} \mathbb{Q}_{\boldsymbol{XX}}^{-1}$$

and $\mathbb{Q}_{\boldsymbol{XX}} = \mathbb{E}$, $\mathbb{A} = \mathbb{E}$.

The covariance matrix $\mathbb{V}_{\boldsymbol{\beta}}$ is called the **asymptotic variance matrix** of $\widehat{\beta}$. The matrix is sometimes referred to as the **sandwich** form.

We now turn our attention to the estimation of the sandwich matrix using a finite sample.

Theorem ref(thm:asympvar) presented the asymptotic covariance matrix of $\sqrt{n}(\widehat{\beta} - \beta)$ is

$$\mathbb{V}_{\boldsymbol{\beta}} = \mathbb{Q}_{\boldsymbol{XX}}^{-1} \mathbb{A} \mathbb{Q}_{\boldsymbol{XX}}^{-1}.$$

Without imposing any homoskedasticity condition, we estimate $\mathbb{V}_\beta$ using a plug-in estimator.

We have already seen that $\widehat{\mathbb{Q}}_{\boldsymbol{XX}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{X}_i$ is a natural estimator for $\mathbb{Q}_{\boldsymbol{XX}}$. For $\mathbb{A}$, we use the moment estimator

$$\widehat{\mathbb{A}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{X}_i e_i^2,$$

where $e_i = (Y_i - \boldsymbol{X}_i\widehat{\beta})$ is the i-th residual. As it turns out, $\widehat{\mathbb{A}}$ is a consistent estimator for $\mathbb{A}$.

As a result, we get the following plug-in estimator for $\mathbb{V}_\beta$:

$$\widehat{\mathbb{V}}_\beta = \widehat{\mathbb{Q}}_{\boldsymbol{XX}}^{-1} \widehat{\mathbb{A}} \widehat{\mathbb{Q}}_{\boldsymbol{XX}}^{-1}$$

The estimator is also consistent. For a proof, see Hensen 2013.

As a consequence, we can get the following estimator for the variance, $\mathbb{V}_{\widehat{\beta}}$, of $\widehat{\beta}$ in the heteroskedastic case.

$$
\begin{aligned}
\widehat{\mathbb{V}}\left[\widehat{\boldsymbol{\beta}}\right] &= \frac{1}{n}\widehat{\mathbb{V}}^{\mathrm{HCo}}_{\boldsymbol{\beta}} \\
&= \frac{1}{n}\widehat{\mathbb{Q}}^{-1}_{\boldsymbol{XX}}\widehat{\mathbb{A}}\widehat{\mathbb{Q}}^{-1}_{\boldsymbol{XX}} \\
&= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}e_i^2\boldsymbol{X}_i^T\boldsymbol{X}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i\right)^{-1} \\
&= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{D}\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}
\end{aligned}
$$

where $\mathbb{D}$ is an $n \times n$ diagonal matrix with diagonal entries $e_1^2, e_2^2, \ldots, e_n^2$. The estimator, $\widehat{\mathbb{V}}\left[\widehat{\beta}\right]$, is referred to as the **robust error variance estimator** for the OLS coefficients $\widehat{\beta}$.