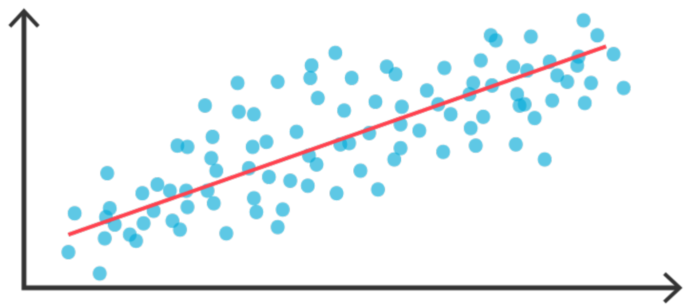
w203: Statistics for Data Science

w203 Instructors

2022-02-07

Contents

Cover



Chapter 1

Regression

We write a k-vector (of scalars) as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

The transpose of $\boldsymbol{x}$ as

$$\boldsymbol{x}' = \begin{bmatrix} x_1 & x_2 & \dots & x_k \end{bmatrix}.$$

We use uppercase letters $X, Y, Z, \dots$ to denote random variables. Random vectors are denoted by bold uppercase letters $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \dots$, and written as a column vector. For example,

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}_{k \times 1}$$

In order to distinguish random matrices from vectors, a random matrix is denoted by $\mathbb{X}$.

The expectation of $\boldsymbol{X}$ is defined as

$$\mathbb{E}[\boldsymbol{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_k] \end{bmatrix}$$

The $k \times k$ covariance matrix of $\boldsymbol{X}$ is

$$V[\boldsymbol{X}] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])'] \tag{1.1}$$
$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2}^2 & \dots & \sigma_k^2 \end{bmatrix}_{k \times k} \tag{1.2}$$

where $\sigma_j = V[X_j]$ and $\sigma_{ij} = Cov[X_i, X_j]$ for $i, j = 1, 2, \dots, k$ and $i \neq j$.

1.1 Conditional Expectation Function

Theorem 1.1. If $\mathbb{E}\left[Y^2\right] < \infty$ and $\boldsymbol{X}$ is a random vector such that $Y = m(\boldsymbol{X}) + e$, then the following statements are equivalent:
1. $m(X) = E[Y|\boldsymbol{X}]$, the CEF of Y given $\boldsymbol{X}$
2. $\mathbb{E}\left[e|\boldsymbol{X}\right] = 0$

1.2 Best Linear Predictor

Let Y be a random variable and $\boldsymbol{X}$ be a random vector. We denote the best linear predictor of Y given $\boldsymbol{X}$ by $\mathcal{P}[Y|\boldsymbol{X}]$. It's also called the linear projection of Y on $\boldsymbol{X}$.

Theorem 1.2 (Best Linear Predictor). Under the following assumptions

1. $\mathbb{E}\left[Y^2\right] < \infty$
2. $\mathbb{E}||\boldsymbol{X}||^2 < \infty$
3. $\mathbb{Q}_{\boldsymbol{XX}} := \mathbb{E}\left[\boldsymbol{XX}'\right]$ is positive-definite

the best linear predictor exists uniquely, and has the form

$$\mathcal{P}[Y|\boldsymbol{X}] = \boldsymbol{X}'\beta,$$

where $\beta = \left(\mathbb{E}[\boldsymbol{XX}']\right)^{-1}\mathbb{E}[\boldsymbol{X}Y]$.

Theorem 1.3 (Best Linear Predictor Error). If the BLP exists, the linear projection error $e = Y - \mathcal{P}[Y|\boldsymbol{X}]$ follows the following properties:

1. $\mathbb{E}[\boldsymbol{X}e] = \boldsymbol{0}$
2. $\mathbb{E}[e] = 0$ if $\boldsymbol{X}' = \begin{bmatrix} 1 & X_1 & \dots & X_k \end{bmatrix}$.

Chapter 2

Ordinary Least Squares

Let Y be our outcome random variable and

$$\boldsymbol{X} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}_{(k+1)\times 1}$$

be our predictor vector containing k predictors and a constant. We denote the joint distribution of $(Y, \boldsymbol{X})$ by $F(y, \boldsymbol{x})$, i.e.,

$$F(y, \boldsymbol{x}) = \mathbb{P}(Y \leq y, \boldsymbol{X} \leq \boldsymbol{x}) = \mathbb{P}(Y \leq y, X_1 \leq x_1, \dots, X_k \leq x_k).$$

The dataset or sample is a collection of observations $\{(Y_i, \boldsymbol{X}_i) : i = 1, 2, \dots, n\}$. We assume that each observation $(Y_i, \boldsymbol{X}_i)$ is a random vector drawn from the common distribution or population F.

2.1 Matrix Formulation

For a given vector of (unknown) coefficients $\boldsymbol{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \dots & \beta_k\end{bmatrix}' \in \mathbb{R}^{k+1}$, we define the following cost function:

$$\widehat{S}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{X_i}'\boldsymbol{\beta})^2.$$

The cost function $\widehat{S}(\boldsymbol{\beta})$ can also be thought of as the average sum of residuals. In fact, $\widehat{S}(\boldsymbol{\beta})$ is the moment (plug-in) estimator of the mean squared error,

$$S(\boldsymbol{\beta}) = \mathbb{E}\left[(Y - \boldsymbol{X}'\boldsymbol{\beta})^2\right].$$

We now minimize $\widehat{S}(\boldsymbol{\beta})$ over all possible choices of $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$. When the minimizer exists and is unique, we call it the least squares estimator, denoted $\hat{\boldsymbol{\beta}}$.

Definition 2.1 ((Ordinary) Least Squares Estimator). The least square estimator is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{k+1}}{\arg\min}\ \widehat{S}(\boldsymbol{\beta}),$$

provided it exists uniquely.

2.2 Solution of OLS

We rewrite the cost function as

$$\widehat{S}(\boldsymbol{\beta}) = \frac{1}{n} SSE(\boldsymbol{\beta}),$$

where $SSE(\boldsymbol{\beta}) := \sum_{i=1}^{n} (Y_i - \boldsymbol{X_i}'\boldsymbol{\beta})^2$.

We now express $SSE(\boldsymbol{\beta})$ as a quadratic function of $\boldsymbol{\beta}'$.

$$\begin{aligned}
SSE &= \sum_{i=1}^{n} (Y_i - \boldsymbol{X_i}'\boldsymbol{\beta})^2 && (2.1)\\
&= \sum_{i=1}^{n} Y_i^2 - 2\sum_{i=1}^{n} Y_i(\boldsymbol{X_i}'\boldsymbol{\beta}) + \sum_{i=1}^{n} (\boldsymbol{X_i}'\boldsymbol{\beta})^2 && (2.2)\\
&= \sum_{i=1}^{n} Y_i^2 - 2\sum_{i=1}^{n} Y_i(\boldsymbol{\beta}'\boldsymbol{X_i}) + \sum_{i=1}^{n} (\boldsymbol{X_i}'\boldsymbol{\beta})(\boldsymbol{X_i}'\boldsymbol{\beta}) && (2.3)\\
&= \sum_{i=1}^{n} Y_i^2 - 2\sum_{i=1}^{n} \boldsymbol{\beta}'(Y_i\boldsymbol{X_i}) + \sum_{i=1}^{n} (\boldsymbol{\beta}'\boldsymbol{X_i})(\boldsymbol{X_i}'\boldsymbol{\beta}) && (2.4)\\
&= \left(\sum_{i=1}^{n} Y_i^2\right) - 2\boldsymbol{\beta}'\left(\sum_{i=1}^{n} \boldsymbol{X_i} Y_i\right) + \boldsymbol{\beta}'\left(\sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}'\right)\boldsymbol{\beta} && (2.5)
\end{aligned}$$

Taking partial derivative w.r.t. β_j, we get

$$\frac{\partial}{\partial \beta_j} SSE(\boldsymbol{\beta}) = -2\left[\sum_{i=1}^{n} \boldsymbol{X_i} Y_i\right]_j + 2\left[\left(\sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}'\right)\boldsymbol{\beta}\right]_j.$$

Therefore,

$$\frac{\partial}{\partial \boldsymbol{\beta}} SSE(\boldsymbol{\beta}) = -2\left(\sum_{i=1}^{n} \boldsymbol{X_i} Y_i\right) + 2\left(\sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}'\right)\boldsymbol{\beta}.$$

In order to miniminize $SSE(\boldsymbol{\beta})$, a necessary condition for $\widehat{\boldsymbol{\beta}}$ is

$$\frac{\partial}{\partial \boldsymbol{\beta}} SSE(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \mathbf{0},$$

i.e.,

$$-2\left(\sum_{i=1}^{n} \boldsymbol{X_i} Y_i\right) + 2\left(\sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}'\right)\widehat{\boldsymbol{\beta}} = \mathbf{0}$$

So,

$$\left(\sum_{i=1}^{n} \boldsymbol{X_i} Y_i\right) = \left(\sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}'\right)\widehat{\boldsymbol{\beta}} \tag{2.6}$$

Both the left and right hand side of the above equation are $k+1$ vectors. So, we have a system of $(k+1)$ linear equations with $(k+1)$ unknowns—the elements of $\boldsymbol{\beta}$.

Let us define

$$\widehat{\mathbb{Q}}_{\boldsymbol{XX}} = \frac{1}{n}\left(\sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}'\right) \text{ and } \widehat{\mathbb{Q}}_{\boldsymbol{X}Y} = \frac{1}{n}\left(\sum_{i=1}^{n} \boldsymbol{X_i} Y_i\right).$$

Rewriting (2.6), we get

$$\widehat{\mathbb{Q}}_{\boldsymbol{X}Y} = \widehat{\mathbb{Q}}_{\boldsymbol{XX}}\hat{\boldsymbol{\beta}}. \tag{2.7}$$

Equation (2.7) is sometimes referred to as the first-order moment condition. For the uniqueness of solution, we require that $\widehat{\mathbb{Q}}_{\boldsymbol{XX}}$ is non-singular. In that case, we can solve for $\widehat{\boldsymbol{\beta}}$ to get,

$$\hat{\boldsymbol{\beta}} = \left[\widehat{\mathbb{Q}}_{\boldsymbol{XX}}\right]^{-1}\widehat{\mathbb{Q}}_{\boldsymbol{X}Y}.$$

To verify that the above choice minimizes $SSE(\boldsymbol{\beta})$, one can consider the second-order moment conditions.

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} SSE(\boldsymbol{\beta}) = 2\widehat{\mathbb{Q}}_{\boldsymbol{XX}}.$$

If $\widehat{\mathbb{Q}}_{\boldsymbol{XX}}$ is non-singular, it is also positive-definite. So, we have actually proved the following theorem.

Theorem 2.1. If $\widehat{\mathbb{Q}}_{\boldsymbol{XX}}$ is non-singular, then the least squares estimator is unique, and is given by

$$\hat{\boldsymbol{\beta}} = \left[\widehat{\mathbb{Q}}_{\boldsymbol{XX}}\right]^{-1}\widehat{\mathbb{Q}}_{\boldsymbol{X}Y}.$$

2.3 Errors and Residuals

We first define the fitted value as

$$\widehat{Y}_i = \boldsymbol{X}_i'\hat{\boldsymbol{\beta}} \text{ for } i = 1, 2, \dots, n.$$

For the least squares estimators, we define the errors and residuals in the following way:

$$e_i = Y_i - \boldsymbol{X}'\boldsymbol{\beta}, \text{ and } \hat{e}_i = Y_i - \widehat{Y}_i.$$

Theorem 2.2 (Least Squares Error). If $\widehat{\mathbb{Q}}_{\boldsymbol{XX}}$ is non-singular, then

1. $\sum_{i=1}^{n} \boldsymbol{X}_i \hat{e}_i = \boldsymbol{0}$
2. $\sum_{i=1}^{n} \hat{e}_i = 0$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} \boldsymbol{X}_i \hat{e}_i &= \sum_{i=1}^{n} \boldsymbol{X}_i (Y_i - \widehat{Y}_i) && (2.8)\\
&= \sum_{i=1}^{n} \boldsymbol{X}_i Y_i - \sum_{i=1}^{n} \boldsymbol{X}_i \widehat{Y}_i && (2.9)\\
&= \sum_{i=1}^{n} \boldsymbol{X}_i Y_i - \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\hat{\boldsymbol{\beta}} && (2.10)\\
&= \widehat{Q}_{\boldsymbol{X}Y} - \widehat{Q}_{\boldsymbol{XX}}\hat{\boldsymbol{\beta}} && (2.11)\\
&= \widehat{Q}_{\boldsymbol{X}Y} - \widehat{Q}_{\boldsymbol{XX}} \left(\widehat{Q}_{\boldsymbol{XX}}^{-1}\widehat{Q}_{\boldsymbol{X}Y}\right) && (2.12)\\
&= \boldsymbol{0} && (2.13)
\end{aligned}$$

From the first row of (1) we get

$$\sum_{i=1}^{n} X_{i1}\hat{e}_i = 0.$$

Since $X_{i1} = 1$ for all i, we have that

$$\sum_{i=1}^{n} \hat{e}_i = 0.$$

Hence the result. □

2.4 Model in Matrix Notation

Taking the definition of errors from the last section, we can write down a system of n linear equations:

$$Y_1 = \boldsymbol{X_1}'\boldsymbol{\beta} + e_1 \tag{2.14}$$
$$Y_2 = \boldsymbol{X_2}'\boldsymbol{\beta} + e_2 \tag{2.15}$$
$$\vdots \tag{2.16}$$
$$Y_n = \boldsymbol{X_1}'\boldsymbol{\beta} + e_n \tag{2.17}$$

Define

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}, \ \mathbb{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix}_{n\times(k+1)}, \text{ and } \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}.$$

We can now rewrite the system as the following:

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

Note that

$$\mathbb{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}$$

We also note that

$$\widehat{Q}_{\boldsymbol{XX}} = \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{X}_i = \mathbb{X}'\mathbb{X},$$

and

$$\widehat{Q}_{\boldsymbol{X}Y} = \sum_{i=1}^{n} \boldsymbol{X}_i Y_i = \mathbb{X}'\boldsymbol{Y}.$$

So, we have write the least squares estimator as

$$\hat{\boldsymbol{\beta}} = \left[\mathbb{X}'\mathbb{X}\right]^{-1} \mathbb{X}\boldsymbol{Y}.$$

Similarly, the residual vector is

$$\hat{\boldsymbol{e}} = \boldsymbol{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}.$$

As a consequence, we can write

$$\mathbb{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}.$$

Chapter 3

Linear Conditional Expectation Function

3.1 Variance of Error

We first compute the (unconditional) variance of the error vector $\boldsymbol{e}$. The covariance matrix

$$\mathbb{V}[\boldsymbol{e}] = \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}'\right] - \mathbb{E}\left[\boldsymbol{e}\right]\mathbb{E}\left[\boldsymbol{e}'\right] = \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}'\right] \overset{\text{def}}{=} \mathbb{D}.$$

For $i \neq j$, the errors e_i,e_j are independent. As a result, $\mathbb{E}\left[e_i e_j\right] = \mathbb{E}\left[e_i\right]\mathbb{E}\left[e_j\right] = 0$. So, $\mathbb{D}$ is a diagonal matrix with the i-th diagonal element σ_i^2:

$$\mathbb{D} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}.$$

3.2 Variance of OLS Estimators

Chapter 4

Large-Sample Regression

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

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + e.$$

we have from Theorem 1.3

$$\mathbb{E}\left[e\right] = 0, \text{ and } \mathbb{E}\left[\boldsymbol{X}e\right] = \mathbf{0}.$$

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

$$Y_i = \boldsymbol{X_i}'\boldsymbol{\beta} + e_i.$$

In matrix notation, we can write

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

Then

$$\mathbb{E}\left[\boldsymbol{e}\right] = \mathbf{0}$$

.

4.1 Consistency of OLS Estimators

4.2 Asymptotic Normality

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

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= [\mathbb{X}'\mathbb{X}]^{-1}\,\mathbb{X}'\boldsymbol{Y} && (4.1)\\
&= [\mathbb{X}'\mathbb{X}]^{-1}\,\mathbb{X}'(\mathbb{X}\boldsymbol{\beta}+\boldsymbol{e}) && (4.2)\\
&= [\mathbb{X}'\mathbb{X}]^{-1}\,(\mathbb{X}'\mathbb{X})\boldsymbol{\beta}+[\mathbb{X}'\mathbb{X}]^{-1}\,\mathbb{X}'\boldsymbol{e} && (4.3)\\
&= \boldsymbol{\beta}+[\mathbb{X}'\mathbb{X}]^{-1}\,\mathbb{X}'\boldsymbol{e} && (4.4)
\end{aligned}$$

So,

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=[\mathbb{X}'\mathbb{X}]^{-1}\,\mathbb{X}'\boldsymbol{e} \tag{4.5}$$

We can then multiply by $\sqrt{n}$ both sides of Equation (4.5) to get

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i e_i\right) && (4.6)\\
&= \widehat{\mathbb{Q}}_{\boldsymbol{XX}}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i e_i\right) && (4.7)
\end{aligned}$$

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 1.3) that

$$\mathbb{E}\left[\boldsymbol{X}_i e_i\right]=\mathbb{E}\left[\boldsymbol{X}e\right]=\mathbf{0}.$$

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

$$\mathbb{V}[\boldsymbol{X}_i e_i]=\mathbb{E}\left[\boldsymbol{X}_i e_i(\boldsymbol{X}_i e_i)'\right]=\mathbb{E}\left[\boldsymbol{X}e(\boldsymbol{X}e)'\right]=\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}'e^2\right]\overset{\text{def}}{=}\mathbb{A}.$$

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

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

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., $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[||\boldsymbol{X}||^4\right] < \infty$. Putting everything together, we conclude

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

Theorem 4.1 (Asymptotic Distribution of OLS Estimators). 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. $\mathbb{E}\left[Y^4\right] < \infty$
3. $\mathbb{E}\left[||\boldsymbol{X}||^4\right] < \infty$
4. $\mathbb{Q}_{\boldsymbol{XX}} = \mathbb{E}\left[\boldsymbol{XX}'\right]$ is positive-definite. Under these assumptions, as $n \to \infty$

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

where

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

and $\mathbb{Q}_{\boldsymbol{XX}} = \mathbb{E}\left[\boldsymbol{XX}'\right]$, $\mathbb{A} = \mathbb{E}\left[\boldsymbol{XX}'e^2\right]$.

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

4.3 Covariance Matrix Estimation

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

4.3.1 Heteroskedastic Variance

Theorem 4.1 presented the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

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

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

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

$$\hat{\mathbb{A}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\hat{e}_i^2,$$

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

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

$$\widehat{\mathbb{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}} = \widehat{\mathbb{Q}}_{\boldsymbol{XX}}^{-1} \widehat{\mathbb{A}} \widehat{\mathbb{Q}}_{\boldsymbol{XX}}^{-1}$$

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

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

$$\begin{aligned}
\widehat{\mathbb{V}}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{HC0}} &= \frac{1}{n} \widehat{\mathbb{V}}_{\boldsymbol{\beta}}^{\mathrm{HC0}} && (4.8)\\
&= \frac{1}{n} \widehat{\mathbb{Q}}_{\boldsymbol{XX}}^{-1} \widehat{\mathbb{A}} \widehat{\mathbb{Q}}_{\boldsymbol{XX}}^{-1} && (4.9)\\
&= \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \hat{e}_i^2 \right) \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} && (4.10)\\
&= \left(\mathbb{X}\mathbb{X}' \right)^{-1} \mathbb{X}\mathbb{D}\mathbb{X}' \left(\mathbb{X}\mathbb{X}' \right)^{-1} && (4.11)
\end{aligned}$$

where $\mathbb{D}$ is an $n \times n$ diagonal matrix with diagonal entries $\hat{e}_1^2, \hat{e}_2^2, \dots, \hat{e}_n^2$. The estimator $\widehat{\mathbb{V}}_{\widehat{\boldsymbol{\beta}}}^{\mathrm{HC0}}$ is referred to as the robust error variance estimator for the OLS coefficients $\widehat{\boldsymbol{\beta}}$.

4.3.2 Homeskedastic Variance

Appendix A

Proofs