Chapter 3

Least Squares

Notation: y_i is a scalar, and $x_i = (x_{i1}, \ldots, x_{iK})'$ is a $K \times 1$ vector. $Y = (y_1, \ldots, y_n)'$ is an $n \times 1$ vector, and

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{22} & \cdots & x_{nK} \end{bmatrix}$$

is an $n \times K$ matrix.

The ordinary least squares (OLS) is the most basic estimation technique in econometrics. It is simple and transparent. Understanding it thoroughly paves the way to study more sophisticated linear estimators. Many nonlinear estimators resemble the behavior of OLS in a neighborhood of the true value.

In this lecture, we study the finite sample properties of OLS. We first learn a series of facts from the linear algebra. Next, we derive the finite-sample exact distribution under the normality of the error term. Finally, the Gauss-Markov theorem justifies the optimality of OLS under the classical assumptions.

3.1 Algebra of Least Squares

3.1.1 OLS

As we have learned from the linear project model, the projection coefficient β in the regression

$$y = x'\beta + e$$

can be written as

$$\beta = \left(E\left[xx'\right]\right)^{-1} E\left[xy\right]. \tag{3.1}$$

We draw a pair of (y, x) from the joint distribution, and we mark it as (y_i, x_i) for $i = 1, \ldots, n$ repeated experiments. We possess a *sample* $(y_i, x_i)_{i=1}^n$.

Remark 3.1. Is (y_i, x_i) random or deterministic? Before we make the observation, they are treated as random variables whose realized values are uncertain. After we make the observation, they become deterministic values which cannot vary anymore. (y_i, x_i) are treated as random when we talk about statistical properties — statistical properties of a fixed number

is meaningless.

Remark 3.2. All probability descriptions are thought experiments before observation. In reality, we have at hand fixed numbers (more recently, words, photos, audio clips, video clips, etc., which can all be represented in digital formats with 0 and 1) to feed into a computational operation, and the operation will return one or some numbers. All statistical interpretation about these numbers are drawn from the probabilistic thought experiments. A *thought experiment* is an academic jargon for a *story* in plain language. Under the axiomatic approach of probability theory, such stories are mathematical consistent and coherent. But mathematics is a tautological system, not science. The scientific value of a probability model depends on how close it is to the *truth* or implications of the truth. In this course, we suppose that the data are generated from some mechanism, which is taken as the truth. In the linear regression model for example, the joint distribution of (y, x) is the truth, while we are interested in the linear projection coefficient β, which is an implication of the truth as in (3.1).

The sample mean is a natural estimator of the population mean. Replace the population mean $E[\cdot]$ in (3.1) by the sample mean $\frac{1}{n}\sum_{i=1}^{n}\cdot$, and the resulting estimator is

$$\widehat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \left(X'X\right)^{-1} X'y$$

if $X'X$ is invertible. This is one way to motivate the OLS estimator.

Alternatively, we can derive the OLS estimator from minimizing the sum of squared residuals $\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2$, or equivalently

$$Q\left(\beta\right) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2 = \frac{1}{2n}\left(Y - X\beta\right)'\left(Y - X\beta\right) = \frac{1}{2n}\left\|Y - X\beta\right\|^2,$$

where the factor $\frac{1}{2n}$ is nonrandom and does not change the minimizer, and $\|\cdot\|$ is the Euclidean norm of a vector. Solve the first-order condition

$$\frac{\partial}{\partial\beta}Q\left(\beta\right) = -\frac{1}{n}X'\left(Y - X\beta\right) = 0.$$

This necessary condition for optimality gives exactly the same $\widehat{\beta} = \left(X'X\right)^{-1}X'y$. Moreover, the second-order condition

$$\frac{\partial^2}{\partial\beta\partial\beta'}Q\left(\beta\right) = \frac{1}{n}X'X$$

shows that $Q\left(\beta\right)$ is convex in β due to the positive semi-definite matrix $X'X$. (The function $Q\left(\beta\right)$ is strictly convex in β if $X'X$ is positive definite.)

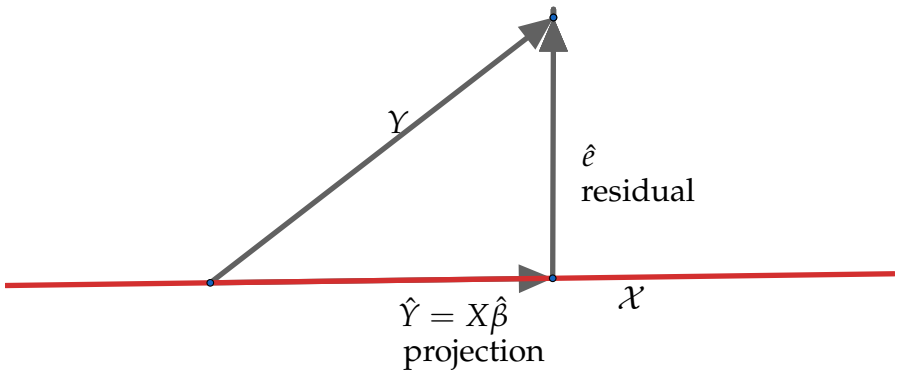
Here are some definitions and properties of the OLS estimator.

- Fitted value: $\widehat{Y} = X\widehat{\beta}$.
- Projection matrix: $P_X = X\left(X'X\right)^{-1}X$; Residual maker matrix: $M_X = I_n - P_X$.
- $P_X X = X$; $X'P_X = X'$.
- $M_X X = 0$; $X'M_X = 0$.

- $P_X M_X = M_X P_X = 0$.
- If $AA = A$, we call it an *idempotent* matrix. Both P_X and M_X are idempotent.
- Residual: $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - X(X'X)^{-1}X'Y = (I - P_X)Y = M_X Y = M_X(X\beta + e) = M_X e$. Notice $\widehat{e}$ and e are two different objects.
- $X'\widehat{e} = X'M_X e = 0$.
- $\sum_{i=1}^{n} \widehat{e_i} = 0$ if x_i contains a constant.

  (Because $X'\widehat{e} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ * & * & \cdots & * \\ \cdots & \cdots & \ddots & \vdots \\ * & * & \cdots & * \end{bmatrix} \begin{bmatrix} \widehat{e_1} \\ \widehat{e_2} \\ \vdots \\ \widehat{e_n} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ and the the first row implies $\sum_{i=1}^{n} \widehat{e_i} = 0$.)

The operation of OLS bears a natural geometric interpretation. Notice $\mathcal{X} = \{W : W = X\beta,\ \beta \in \mathbb{R}^K\}$ is the linear space spanned by the K columns of $X = [X_{\cdot 1}, \ldots, X_{\cdot K}]$, which is of K-dimension if the columns are linearly independent. The OLS estimator is the minimizer of $\min_{\beta \in \mathbb{R}^K} \|Y - X\beta\|$ (Square the Euclidean norm or not does not change the minimizer because a^2 is a monotonic transformation for $a \geq 0$). In other words, $X\widehat{\beta}$ is the point in $\mathcal{X}$ such that it is the closest to the vector Y in terms of the Euclidean norm.



The relationship $Y = X\widehat{\beta} + \widehat{e}$ decomposes Y into two orthogonal vectors $X\widehat{\beta}$ and $\widehat{e}$ as $\left\langle X\widehat{\beta}, \widehat{e} \right\rangle = \widehat{\beta}X'\widehat{e} = 0$, where $\langle \cdot, \cdot \rangle$ is the *inner product* of two vectors. Therefore $X\widehat{\beta}$ is the *projection* of Y onto $\mathcal{X}$, and $\widehat{e}$ is the corresponding *projection residuals.* The Pythagorean theorem implies

$$\|Y\|^2 = \|X\widehat{\beta}\|^2 + \|\widehat{e}\|^2 .$$

3.1.2 Goodness of Fit

R-squared is a popular measure of goodness-of-fit in the linear regression. R-squared is well defined only when a constant is included in the regressors. Let $M_\iota = I_n - \frac{1}{n}\iota\iota'$, where ι is an $n \times 1$ vector of 1's. M_ι is the *demeaner*, in the sense that $M_\iota (z_1, \ldots, z_n)' = (z_1 - \bar{z}, \ldots, z_n - \bar{z})'$, where $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$.

$$R^2 = \frac{\widehat{Y}' M_\iota \widehat{Y}}{Y' M_\iota Y} = \frac{\frac{1}{n}\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\frac{1}{n}\sum_{i=1}^n (y - \overline{y})^2}$$

is the ratio between the predicted variance and the total variance.

Exercise 3.1. Show $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i = \sum_{i=1}^n \widehat{y}$.

We can decompose $Y = P_X Y + M_X Y = \widehat{Y} + \widehat{e}$. The total variation is

$$Y' M_\iota Y = \left(\widehat{Y} + \widehat{e}\right)' M_\iota \left(\widehat{Y} + \widehat{e}\right) = \widehat{Y}' M_\iota \widehat{Y} + 2\widehat{Y}' M_\iota \widehat{e} + \widehat{e}' M_\iota \widehat{e} = \widehat{Y}' M_\iota \widehat{Y} + \widehat{e}'\widehat{e}$$

where the last equality follows by $M_\iota \widehat{e} = \widehat{e}$ as $\frac{1}{n}\sum_{i=1}^n \widehat{e}_i = 0$ (This property fails if the regression has no constant), and $\widehat{Y}'\widehat{e} = Y' P_X M_X e = 0$. .

The magnitude of R-squared varies in different contexts. In macro models with the lagged dependent variables are presents, it is not unusually to observe R-squared larger than 90%. In cross sectional regressions it is often below 20%. In financial market shorter-term predictive models, a person may become a billion if he can systematically achieve 2% out-of-sample R-squared.

3.1.3 Frish-Waugh-Lovell Theorem

The Frish-Waugh-Lovell (FWL) theorem is an algebraic fact about the formula of a subvector of the OLS estimator. To derive the FWL theorem we need to use the inverse of partitioned matrix. For a positive definite symmetric matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{pmatrix}$, the inverse can be written as

$$A^{-1} = \begin{pmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}'\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & \left(A_{22} - A_{12}'A_{11}^{-1}A_{12}\right)^{-1} \end{pmatrix}.$$

In our context of OLS estimator, let $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$

$$\begin{aligned} \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} &= (X'X)^{-1}X'Y \\ &= \left(\begin{pmatrix} X_1' \\ X_2' \end{pmatrix}\begin{pmatrix} X_1 & X_2 \end{pmatrix}\right)^{-1}\begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix} \\ &= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1}\begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix} \\ &= \begin{pmatrix} \left(X_1'M_{X_2}'X_1\right)^{-1} & -\left(X_1'M_{X_2}'X_1\right)^{-1}X_1'X_2\left(X_2'X_2\right)^{-1} \\ * & * \end{pmatrix}\begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix}. \end{aligned}$$

The subvector

$$\begin{aligned}\widehat{\beta}_1 &= \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' Y - \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' X_2 \left(X_2' X_2\right)^{-1} X_2' Y \\ &= \left(X_1' M_{X_2}' X_1\right)^{-1} \left(X_1' Y - X_1' P_{X_2} Y\right) \\ &= \left(X_1' M_{X_2}' X_1\right)^{-1} X_1' M_{X_2} Y.\end{aligned}$$

Notice that $\widehat{\beta}_1$ can be obtained by the following:

1. Regress y on X_2, obtain residuals $\tilde{e}_2$;
2. Regress X_1 on X_2, obtain residuals $\tilde{X}_2$;
3. Regress $\tilde{e}_2$ on $\tilde{X}_2$, obtain OLS estimates $\widehat{\beta}_1$.

Similar derivation can also be carried out in the population linear projection. See Hansen (2020) [E] Chapter 2.22-23.

3.2 Statistical Properties of Least Squares

In this section we return to the classical statistical framework under restrictive distributional assumption $y_i | x_i \sim N\left(x_i'\beta, \gamma\right)$, where $\gamma = \sigma^2$ to ease the differentiation. This assumption is equivalent to $e_i | x_i = \left(y_i - x_i'\beta\right) | x_i \sim N\left(0, \gamma\right)$. Because the distribution of e_i is invariant to x_i, the error term $e_i \sim N\left(0, \gamma\right)$ and is statistically independent of x_i. This is a very strong assumption.

3.2.1 Maximum Likelihood Estimation

The *conditional* likelihood of observing a *random sample* $(y_i, x_i)_{i=1}^n$ is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right),$$

and the (conditional) log-likelihood function is

$$L\left(\beta, \gamma\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\gamma - \frac{1}{2\gamma}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2.$$

The maximum likelihood estimator $\widehat{\beta}_{MLE}$ can be found using the FOC:

$$\frac{\partial}{\partial\beta}L\left(\beta, \gamma\right) = \frac{1}{2\gamma}\sum_{i=1}^{n} 2x_i\left(y_i - x_i'\beta\right)^2 = 0.$$

Rearranging the above equation in matrix form $X'Y = X'X\widehat{\beta}_{MLE}$, we explicitly solve

$$\widehat{\beta}_{MLE} = (X'X)^{-1}X'Y.$$

The maximum likelihood estimator (MLE) coincides with the OLS estimator. Similarly, another FOC gives $\widehat{\gamma}_{\text{MLE}} = \widehat{e}'\widehat{e}/n$.

3.2.2 Classical Finite Sample Distribution

We can show the finite-sample exact distribution of $\widehat{\beta}$ assuming the error term follows a Gaussian distribution. *Finite sample distribution* means that the distribution holds for any n; it is in contrast to *asymptotic distribution*,

which is a large sample approximation to the finite sample distribution. We first review some properties of a generic jointly normal random variables.

Fact 3.1. *Let $z \sim N(\mu, \Omega)$ be an $l \times 1$ random vector with a positive definite variance-covariance matrix Ω. Let A be an $m \times l$ non-random matrix where $m \leq l$. Then $Az \sim N(A\mu, A\Omega A')$.*

Fact 3.2. *If $z \sim N(0,1)$, $w \sim \chi^2(d)$ and z and w are independent. Then $\frac{z}{\sqrt{w/d}} \sim t(d)$.*

The OLS estimator

$$\widehat{\beta} = (X'X)^{-1} X'Y = (X'X)^{-1} X' (X'\beta + e) = \beta + (X'X)^{-1} X'e,$$

and its conditional distribution can be written as

$$\begin{aligned}
\widehat{\beta}|X &= \beta + (X'X)^{-1} X'e|X \\
&\sim \beta + (X'X)^{-1} X' \cdot N\left(0_n, \sigma^2 \cdot I_n\right) \\
&\sim N\left(\beta, \sigma^2 (X'X)^{-1} X'X (X'X)^{-1}\right) \sim N\left(\beta, \sigma^2 (X'X)^{-1}\right)
\end{aligned}$$

by Fact 3.1. The k-th element of the vector coefficient

$$\widehat{\beta}_k|X = \eta_k'\widehat{\beta}|X \sim N\left(\beta_k, \sigma^2 \eta_k' (X'X)^{-1} \eta_k\right) \sim N\left(\beta_k, \sigma^2 (X'X)^{-1}_{kk}\right),$$

where $\eta_k = (1\{l = k\})_{l=1,\ldots,K}$ is the selector of the k-th element.

In reality, σ^2 is an unknown parameter, and

$$s^2 = \widehat{e}'\widehat{e}/\left(n-K\right) = e'M_X e/\left(n-K\right)$$

is an unbiased estimator of σ^2. Consider the t-statistic

$$\begin{aligned} T_k &= \frac{\widehat{\beta}_k - \beta_k}{\sqrt{s^2\left[\left(X'X\right)^{-1}\right]_{kk}}} = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\sigma^2\left[\left(X'X\right)^{-1}\right]_{kk}}} \cdot \frac{\sqrt{\sigma^2}}{\sqrt{s^2}} \\ &= \frac{\left(\widehat{\beta}_k - \beta_k\right)/\sqrt{\sigma^2\left[\left(X'X\right)^{-1}\right]_{kk}}}{\sqrt{\frac{e'}{\sigma}M_X\frac{e}{\sigma}/\left(n-K\right)}}. \end{aligned}$$

The numerator follows a standard normal, and the denominator follows $\frac{1}{n-K}\chi^2\left(n-K\right)$. Moreover, the numerator and the denominator are statistically independent (Basu's theorem). As a result, we conclude $T_k \sim t\left(n-K\right)$ by Fact 3.2. This finite sample distribution allows us to conduct statistical inference.

3.2.3 Mean and Variance

Now we relax the normality assumption and statistical independence. Instead, we represent the regression model as $Y = X\beta + e$ and

$$\begin{aligned} E[e|X] &= 0_n \\ \text{var}\left[e|X\right] &= E\left[ee'|X\right] = \sigma^2 I_n. \end{aligned}$$

where the first condition is the *mean independence* assumption, and the second condition is the *homoskedasticity* assumption. These assumptions are about the first and second *moments* of e_i conditional on x_i. Unlike the normality assumption, they do not restrict the distribution of e_i.

- Unbiasedness:

$$\begin{aligned} E\left[\widehat{\beta}|X\right] &= E\left[\left(X'X\right)^{-1}XY|X\right] = E\left[\left(X'X\right)^{-1}X\left(X'\beta+e\right)|X\right] \\ &= \beta + \left(X'X\right)^{-1}XE\left[e|X\right] = \beta. \end{aligned}$$

  Unbiasedness does not rely on homoskedasticity.

- Variance:

$$\begin{aligned} \text{var}\left[\widehat{\beta}|X\right] &= E\left[\left(\widehat{\beta}-E\widehat{\beta}\right)\left(\widehat{\beta}-E\widehat{\beta}\right)'|X\right] \\ &= E\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'|X\right] \\ &= E\left[\left(X'X\right)^{-1}X'ee'X\left(X'X\right)^{-1}|X\right] \\ &= \left(X'X\right)^{-1}X'E\left[ee'|X\right]X\left(X'X\right)^{-1} \end{aligned}$$

  where the second equality holds as $E\left[\widehat{\beta}\right] = E\left[E\left[\widehat{\beta}|X\right]\right] = \beta$. Under the assumption of homoskedasticity, it can be simplified as

$$\text{var}\left[\widehat{\beta}|X\right] = \left(X'X\right)^{-1}X'\left(\sigma^2 I_n\right)X\left(X'X\right)^{-1} = \sigma^2\left(X'X\right)^{-1}.$$

Example 3.1. (Heteroskedasticity) If $e_i = x_i u_i$, where x_i is a scalar random variable, u_i is statistically independent of x_i, $E[u_i] = 0$ and $E\left[u_i^2\right] = \sigma^2$. Then $E[e_i|x_i] = 0$ but $E\left[e_i^2|x_i\right] = \sigma^2 x_i^2$ is a function of x_i. We say e_i^2 is a heteroskedastic error.

It is important to notice that independently and identically distributed sample (iid) (y_i, x_i) does not imply homoskedasticity. Homoskedasticity or heterskdasticity is about the relationship between $(x_i, e_i = y_i - \beta x)$, whereas iid is about the relationship between (y_i, x_i) and $\left(y_j, x_j\right)$ for $i \neq j$.

3.2.4 Gauss-Markov Theorem

Gauss-Markov theorem is concerned about the optimality of OLS. It justifies OLS as the efficient estimator among all linear unbiased ones. *Efficient* here means that it enjoys the smallest variance in a family of estimators.

We have shown that OLS is unbiased in that $E\left[\widehat{\beta}\right] = \beta$. There are numerous linearly unbiased estimators. For example, $\left(Z'X\right)^{-1} Z'y$ for $z_i = x_i^2$ is unbiased because $E\left[\left(Z'X\right)^{-1} Z'y\right] = E\left[\left(Z'X\right)^{-1} Z'\left(X\beta + e\right)\right] = \beta$. We cannot say OLS is better than those other unbiased estimators because they are equally good in this aspect. Thus, we move to the second order property of variance: an estimator is better if its variance is smaller.

Fact 3.3. *For two generic random vectors X and Y of the same size, we say X's variance is smaller or equal to Y's variance if $(\Omega_Y - \Omega_X)$ is a positive semi-definite matrix. The comparison is defined this way because for any non-zero*

constant vector c, the variance of the linear combination of X

$$\text{var}\left(c'X\right) = c'\Omega_X c \leq c'\Omega_Y c = \text{var}\left(c'Y\right)$$

is no bigger than the same linear combination of Y.

Let $\tilde{\beta} = A'y$ be a generic linear estimator, where A is any $n \times K$ functions of X. As

$$E\left[A'y|X\right] = E\left[A'\left(X\beta + e\right)|X\right] = A'X\beta.$$

So the linearity and unbiasedness of $\tilde{\beta}$ implies $A'X = I_n$. Moreover, the variance

$$\text{var}\left(A'y|X\right) = E\left[\left(A'y - \beta\right)\left(A'y - \beta\right)'|X\right] = E\left[A'ee'A|X\right] = \sigma^2 A'A.$$

Let $C = A - X\left(X'X\right)^{-1}$.

$$\begin{aligned} A'A - \left(X'X\right)^{-1} &= \left(C + X\left(X'X\right)^{-1}\right)'\left(C + X\left(X'X\right)^{-1}\right) - \left(X'X\right)^{-1} \\ &= C'C + \left(X'X\right)^{-1}X'C + C'X\left(X'X\right)^{-1} \\ &= C'C, \end{aligned}$$

where the last equality follows as

$$\left(X'X\right)^{-1}X'C = \left(X'X\right)^{-1}X'\left(A - X\left(X'X\right)^{-1}\right) = \left(X'X\right)^{-1} - \left(X'X\right)^{-1} = 0.$$

Therefore $A'A - (X'X)^{-1}$ is a positive semi-definite matrix. The variance of any $\tilde{\beta}$ is no smaller than the OLS estimator $\widehat{\beta}$. The above derivation shows OLS achieves the smallest variance among all linear unbiased estimators.

Homoskedasticity is a restrictive assumption. Under homoskedasticity, $\text{var}\left[\widehat{\beta}\right] = \sigma^2 (X'X)^{-1}$. Popular estimator of σ^2 is the sample mean of the residuals $\widehat{\sigma}^2 = \frac{1}{n}\widehat{e}'\widehat{e}$ or the unbiased one $s^2 = \frac{1}{n-K}\widehat{e}'\widehat{e}$. Under heteroskedasticity, Gauss-Markov theorem does not apply.

3.3 Summary

The linear algebraic properties holds in finite sample no matter the data are taken as fixed numbers or random variables. The exact distribution under the normality assumption of the error term is the classical statistical results. The Gauss Markov theorem holds under two crucial assumptions: linear CEF and homoskedasticity.

Historical notes: Carl Friedrich Gauss (1777–1855) claimed he had come up with the operation of OLS in 1795. With only three data points at hand, Gauss successfully applied his method to predict the location of the drawf planet Ceres in 1801. While Gauss did not publish the work on OLS until 1809, Adrien-Marie Legendre (1752–1833) presented this method in 1805. Today people tend to attribute OLS to Gauss, assuming that a giant like Gauss had no need to tell a lie to steal Legendre's discovery.

MLE was promulgated and popularized by Ronald Fisher (1890–1962). He was a major contributor of the frequentist approach which dominates mathematical statistics today, and he sharply criticized the Bayesian approach. Fisher collected the Iris flower dataset of 150 observations in his biological study in 1936, which can be displayed in R by typing `iris`.

`Zhentao Shi.  September 15, 2020`