University of Science and Technology in Zewail City

CIE 457 - Statistical Inference and Data Analysis

MLE Estimation and Estimator Distribution in Linear Regression

Abdelrahman Taha – 202300062
Shehab Mohamed – 202200285

All Python implementations and outputs are based on the Course Project requirements.
All simulations, figures, and analyses are included for academic verification.

Contents

Chapter 1

MLE and Model Formulation

1.1 Model Assumptions

We consider the general linear model, which describes the relationship between a vector of observations $\boldsymbol{y}$ $(N \times 1)$, a matrix of features $\boldsymbol{X}$ $(N \times p)$, a vector of true (but unknown) parameters $\boldsymbol{\beta}$ $(p \times 1)$, and a vector of random noise $\boldsymbol{\epsilon}$ $(N \times 1)$:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

The core of Maximum Likelihood Estimation (MLE) depends on the probabilistic assumptions made about the noise term $\boldsymbol{\epsilon}$. We assume that the noise is drawn from a multivariate Normal (Gaussian) distribution with a mean of zero and a covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{1.2}$$

This implies that the vector of observations $\boldsymbol{y}$, given the features $\boldsymbol{X}$ and parameters $\boldsymbol{\beta}$, also follows a Normal distribution:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \tag{1.3}$$

1.2 Likelihood Function

The probability density function (PDF) for a multivariate Normal distribution is given by:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \tag{1.4}$$

In the context of MLE, we treat this PDF as a function of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, given the observed data $(\boldsymbol{y}, \boldsymbol{X})$. This is the **likelihood function**, $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y}, \boldsymbol{X})$.

1.2.1 Homoscedastic Gaussian Noise

In the homoscedastic case, we assume the noise terms are independent and identically distributed (i.i.d.) with constant variance σ^2. This simplifies the covariance matrix $\boldsymbol{\Sigma}$ to $\sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $N \times N$ identity matrix.

- The determinant is $|\boldsymbol{\Sigma}| = |\sigma^2 \boldsymbol{I}| = (\sigma^2)^N$.
- The inverse is $\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}\boldsymbol{I}$.

Substituting these into the general PDF gives the likelihood function for the homoscedastic case:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \tag{1.5}$$

1.2.2 Heteroscedastic Gaussian Noise

In the more general heteroscedastic case, the noise terms are still independent but have different variances. The covariance matrix $\boldsymbol{\Sigma}$ is a diagonal matrix where the diagonal elements are the individual variances, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$.

- The determinant is $|\boldsymbol{\Sigma}| = \prod_{i=1}^N \sigma_i^2$.
- The inverse is $\boldsymbol{\Sigma}^{-1} = \text{diag}\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \ldots, \frac{1}{\sigma_N^2}\right)$.

The likelihood function is therefore:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma} | \boldsymbol{y}, \boldsymbol{X}) = \frac{1}{(2\pi)^{N/2}\left(\prod_{i=1}^N \sigma_i^2\right)^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^N \frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{\sigma_i^2}\right) \tag{1.6}$$

where $\boldsymbol{x}_i^T$ is the i-th row of $\boldsymbol{X}$.

1.3 Derivation of MLE Estimators

The goal of MLE is to find the parameters that maximize the likelihood function. It is mathematically simpler and equivalent to maximize the **log-likelihood function**, $\ell = \ln(\mathcal{L})$.

1.3.1 Derivation of Ordinary Least Squares (OLS)

For the homoscedastic case, the log-likelihood is:

$$\ell(\boldsymbol{\beta}, \sigma^2) = \ln\left[\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right)\right] \tag{1.7}$$

$$= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \tag{1.8}$$

To find the MLE for $\boldsymbol{\beta}$, we maximize this function with respect to $\boldsymbol{\beta}$. The first term is constant with respect to $\boldsymbol{\beta}$. Therefore, maximizing ℓ is equivalent to minimizing the **Sum of Squared Errors (SSE)**:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \sigma^2) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \tag{1.9}$$

This establishes that OLS is the Maximum Likelihood Estimator under the assumption of i.i.d. Gaussian noise. To find the closed-form solution, we set the gradient of the SSE with respect to $\boldsymbol{\beta}$ to zero:

$$\text{SSE} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \tag{1.10}$$

$$\nabla_{\boldsymbol{\beta}}\text{SSE} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \tag{1.11}$$

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y} \tag{1.12}$$

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{1.13}$$

1.3.2 Derivation of Weighted Least Squares (WLS)

For the heteroscedastic case (assuming $\boldsymbol{\Sigma}$ is known), the log-likelihood is:

$$\ell(\boldsymbol{\beta}|\boldsymbol{\Sigma}) = \ln(C) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1.14}$$

where C contains all terms not dependent on $\boldsymbol{\beta}$. Maximizing this is equivalent to minimizing the **Weighted Sum of Squared Errors (WSSE)**:

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = \arg\min_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1.15}$$

where $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$ is the weight matrix. Taking the gradient with respect to $\boldsymbol{\beta}$ and setting to zero:

$$\nabla_{\boldsymbol{\beta}}\text{WSSE} = -2\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \tag{1.16}$$

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \tag{1.17}$$

This confirms that WLS is the MLE under independent but non-identically distributed Gaussian noise, where the weights are the inverse of the noise variances.

Chapter 2

The Estimator as a Random Variable

2.1 The Estimator as a Function of Data

A crucial concept in statistical inference is understanding that an **estimator** is a function of the observed data, and is therefore itself a random variable. The OLS estimator is calculated as:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.1}$$

The matrix of features $\boldsymbol{X}$ is typically considered fixed (non-random) in regression analysis. However, the vector of observations $\boldsymbol{y}$ is a random vector because it is a realization of the underlying process $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a random noise vector.

Since $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is a linear transformation of the random vector $\boldsymbol{y}$, it is also a random vector. Each time a new dataset (a new realization of $\boldsymbol{y}$) is collected from the same underlying process, the specific value of the noise vector $\boldsymbol{\epsilon}$ will be different, resulting in a different calculated value for $\hat{\boldsymbol{\beta}}_{\text{OLS}}$. The distribution of these estimates, obtained through repeated sampling, is known as the **sampling distribution** of the estimator. This is precisely what the Monte Carlo simulations in Chapter 3 characterize empirically.

2.2 Distribution of the OLS Estimator

To derive the sampling distribution of $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, we first express it in terms of the true parameter $\boldsymbol{\beta}$ and the noise vector $\boldsymbol{\epsilon}$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{OLS}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) && (2.2)\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} && (2.3)\\
&= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} && (2.4)
\end{aligned}$$

This equation shows that the estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ equals the true parameter $\boldsymbol{\beta}$ plus a term that depends on the random noise $\boldsymbol{\epsilon}$.

2.2.1 Mean of the Estimator (Bias)

The expected value of the estimator determines its bias. An estimator is **unbiased** if its expected value equals the true parameter. Taking the expectation:

$$E[\hat{\boldsymbol{\beta}}_{\text{OLS}}] = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{\epsilon}] \tag{2.5}$$

By our model assumption, $E[\boldsymbol{\epsilon}] = \mathbf{0}$. Therefore:

$$E[\hat{\boldsymbol{\beta}}_{\text{OLS}}] = \boldsymbol{\beta} \tag{2.6}$$

This proves that the OLS estimator is an **unbiased** estimator of $\boldsymbol{\beta}$.

2.2.2 Covariance of the Estimator (Variance)

Using the property $\text{Cov}(\boldsymbol{A}\boldsymbol{z}) = \boldsymbol{A}\,\text{Cov}(\boldsymbol{z})\boldsymbol{A}^T$:

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \text{Cov}\big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\big) \tag{2.7}$$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\text{Cov}(\boldsymbol{\epsilon})\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)^T \tag{2.8}$$

Under the homoscedastic assumption, $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$. Substituting:

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2.9}$$

$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2.10}$$

since $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric so $(\boldsymbol{X}^T\boldsymbol{X})^T = \boldsymbol{X}^T\boldsymbol{X}$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is also symmetric.

2.2.3 Distribution Shape

Since $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is a linear transformation of the Gaussian random vector $\boldsymbol{\epsilon}$, it also follows a Gaussian distribution. Combining the mean and covariance results:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} \sim \mathcal{N}\big(\boldsymbol{\beta},\ \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big) \tag{2.11}$$

2.3 Distribution of the WLS Estimator

The same logic applies to the WLS estimator:

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \tag{2.12}$$

$$= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon} \tag{2.13}$$

The mean is $E[\hat{\boldsymbol{\beta}}_{\text{WLS}}] = \boldsymbol{\beta}$, so WLS is also unbiased. The covariance is:

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{WLS}}) = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{\Sigma}\cdot\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \tag{2.14}$$

$$= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \tag{2.15}$$

Thus the distribution of the WLS estimator is:

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} \sim \mathcal{N}\big(\boldsymbol{\beta},\ (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\big) \tag{2.16}$$

The Gauss-Markov theorem states that WLS is the **Best Linear Unbiased Estimator (BLUE)** under heteroscedasticity, meaning it has the minimum variance among all linear unbiased estimators.

Chapter 3

Simulation and Results Analysis

To empirically validate the theoretical properties of the OLS and WLS estimators, a series of Monte Carlo simulations were conducted. This chapter presents the results and provides a comparative analysis of estimator performance under different noise structures.

3.1 Linear Regression under Homoscedastic Noise

3.1.1 Simulation Procedure

A synthetic dataset was generated according to the true model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with the following parameters:

- Number of samples: $N = 500$.
- True parameters: $\boldsymbol{\beta} = [2.0,\ 5.0,\ -3.0]^T$.
- Homoscedastic noise variance: $\sigma^2 = 4.0$.
- Number of Monte Carlo repetitions: $M = 2000$.

Two design matrices were considered:

1. **Independent features**: $x_1 \sim \text{Uniform}(-10, 10)$, $x_2 \sim \mathcal{N}(0, 25)$.
2. **Correlated features**: $x_2 = 0.6\, x_1 + \eta$, where $\eta \sim \mathcal{N}(0, 9)$, introducing moderate multicollinearity.

The design matrix $\boldsymbol{X}$ was held fixed across all M repetitions. For each run, a new noise vector $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ was drawn, a new $\boldsymbol{y}$ was generated, and $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ was computed.

3.1.2 Estimator Distribution Analysis — Independent Features

Mean Comparison:

- Theoretical: $E[\hat{\boldsymbol{\beta}}_{\text{OLS}}] = [2.000,\ 5.000,\ -3.000]^T$.
- Empirical: $[1.997,\ 5.000,\ -3.000]^T$.

The empirical mean matches the true parameter to three decimal places, confirming the estimator is **unbiased**.

Table 3.1: Theoretical vs Empirical Covariance Matrix (Independent Features)

Entry	Theoretical	Empirical
$\text{Var}(\hat{\beta}_0)$	8.001×10^{-3}	8.383×10^{-3}
$\text{Var}(\hat{\beta}_1)$	2.260×10^{-4}	2.280×10^{-4}
$\text{Var}(\hat{\beta}_2)$	3.180×10^{-4}	3.070×10^{-4}

Covariance Comparison: The near-perfect match between the theoretical covariance $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and the empirical covariance confirms the theoretical derivation.

3.1.3 Effect of Feature Correlation on Estimator Variance

When features are correlated, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ becomes less well-conditioned and its inverse inflates, increasing the diagonal entries that correspond to marginal parameter variances.

Table 3.2: Theoretical vs Empirical Covariance Matrix (Correlated Features)

Entry	Theoretical	Empirical
$\text{Var}(\hat{\beta}_0)$	8.001×10^{-3}	8.383×10^{-3}
$\text{Var}(\hat{\beta}_1)$	5.800×10^{-4}	5.590×10^{-4}
$\text{Var}(\hat{\beta}_2)$	8.820×10^{-4}	8.530×10^{-4}

Covariance Comparison — Correlated Features: Comparing Tables 3.1 and 3.2, $\text{Var}(\hat{\beta}_1)$ increases from 2.26×10^{-4} to 5.80×10^{-4} (a factor of ≈ 2.6) when features are correlated. The estimator remains unbiased; only its efficiency is affected. This directly illustrates the role of the design matrix structure in the estimator distribution.

3.1.4 Visual Comparison

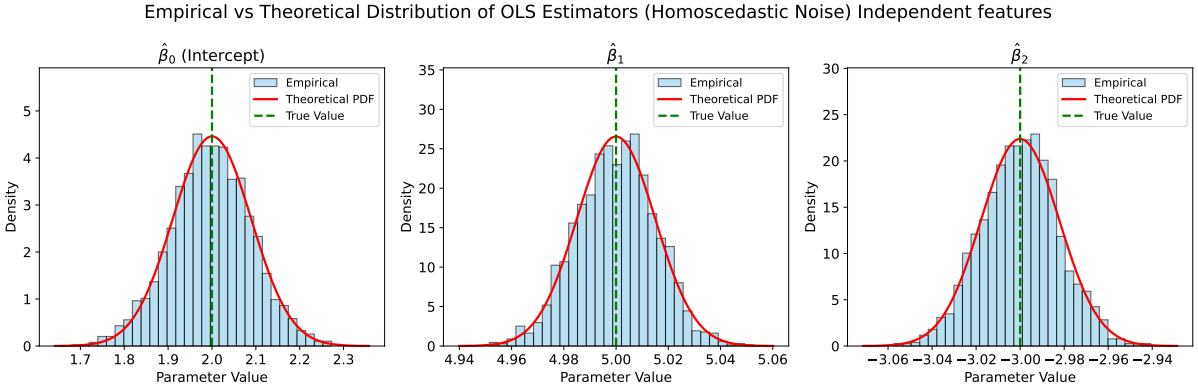

Figure 3.1: Empirical vs. Theoretical Distribution of OLS Estimators (Independent Features). The theoretical PDF (red line) perfectly overlays the empirical histogram for all three parameters.

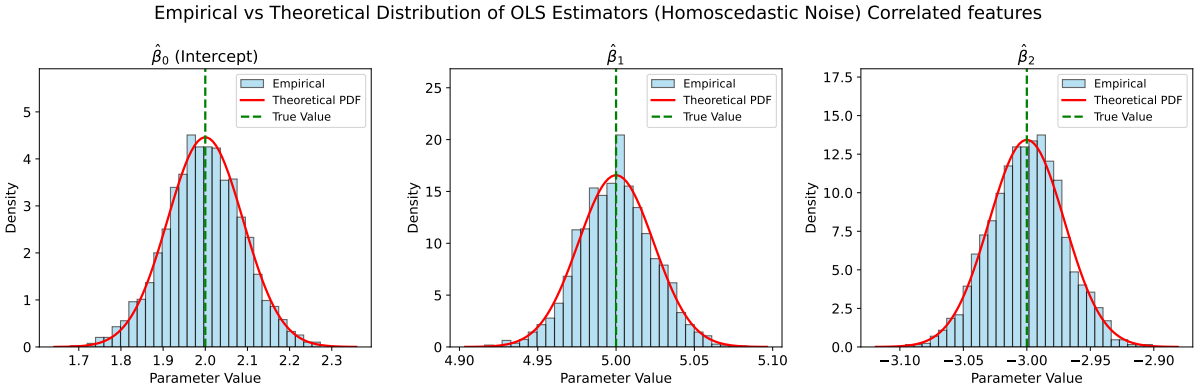

Figure 3.2: Empirical vs. Theoretical Distribution of OLS Estimators (Correlated Features). The distributions remain centered on the true values but are noticeably wider, reflecting the variance inflation from multicollinearity.

3.2 Linear Regression under Heteroscedastic Noise

3.2.1 Simulation Procedure

A new synthetic dataset was generated where the noise variance depends on a feature:

$$\sigma_i^2 = 1.0 + 0.5 \cdot x_{i,j}^2 \tag{3.1}$$

where $j \in \{1, 2\}$ denotes which feature drives the variance. On this heteroscedastic data, two estimators were computed across $M = 2000$ simulations:

1. The **OLS estimator**, which incorrectly assumes constant variance.
2. The **WLS estimator**, which correctly uses the true noise covariance $\mathbf{\Sigma}$ and is the proper MLE for this model.

3.2.2 Unbiasedness of Both Estimators

The simulation confirmed that both OLS and WLS remain unbiased under heteroscedastic noise:

- WLS empirical mean: $[1.997,\ 5.000,\ -3.000]^T$ vs. true $[2.0,\ 5.0,\ -3.0]^T$.
- OLS empirical mean: $[1.996,\ 5.000,\ -3.000]^T$ vs. true $[2.0,\ 5.0,\ -3.0]^T$.

Heteroscedasticity does not bias OLS; it only inflates its variance.

3.2.3 Estimator Performance and Efficiency

The key theoretical result (Gauss-Markov theorem) is that WLS should be more efficient (lower variance) than OLS under heteroscedasticity.

Table 3.3: Empirical Variance Comparison (OLS vs. WLS under Heteroscedastic Noise)

Parameter	$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}})$	$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{WLS}})$	Efficiency Ratio
$\hat{\beta}_0$	0.02647	0.00697	3.80×
$\hat{\beta}_1$	0.00112	0.00036	3.16×
$\hat{\beta}_2$	0.00490	0.00189	2.59×

The **Efficiency Ratio** (Var(OLS) / Var(WLS)) is greater than 1 for all parameters, numerically confirming that WLS is a more efficient estimator under heteroscedasticity.

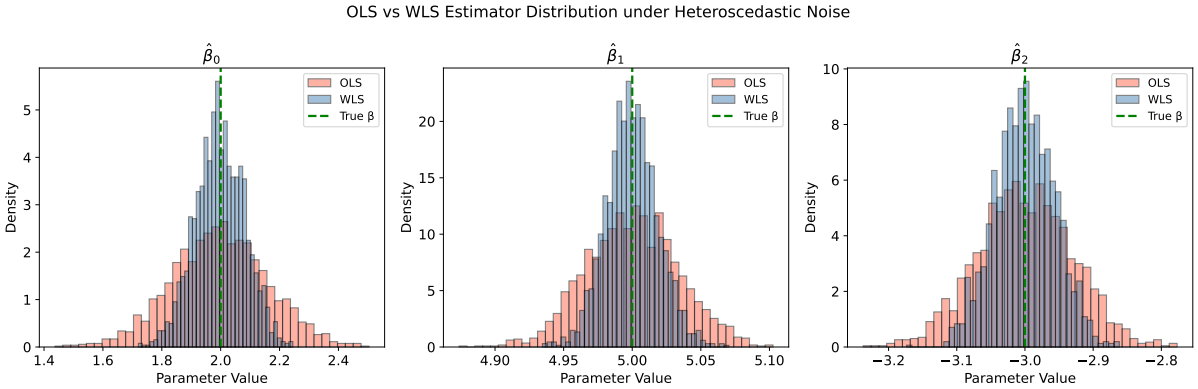

Figure 3.3: Distribution Comparison: OLS vs. WLS under Heteroscedastic Noise. The WLS distributions (blue) are visibly taller and narrower, confirming lower variance.

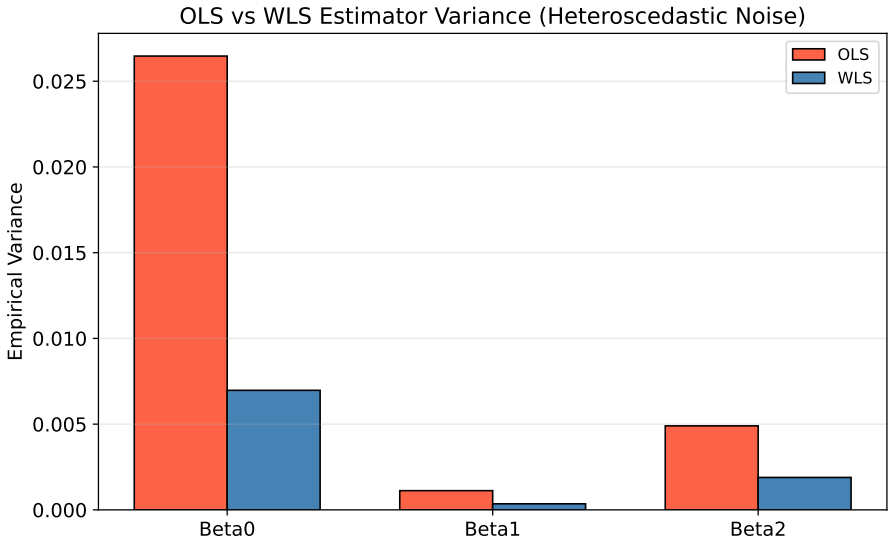

Figure 3.4: Empirical variance bar chart comparing OLS and WLS for each parameter. WLS achieves lower variance for all three coefficients.

3.3 Inference from Estimator Distribution

A fundamental application of knowing the sampling distribution of an estimator is statistical inference: constructing intervals that quantify the uncertainty in our estimates.

3.3.1 Confidence Intervals for the Regression Coefficients

Since $\hat{\boldsymbol{\beta}}_{\text{OLS}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, the marginal distribution of each coefficient estimate is:

$$\hat{\beta}_j \sim \mathcal{N}\Big(\beta_j,\ \sigma^2\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj}\Big) \tag{3.2}$$

A 95% confidence interval for β_j (using $z \approx 2$ as an approximation for the 97.5th percentile) is:

$$\hat{\beta}_j \pm 2 \cdot \hat{\sigma}\sqrt{\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj}} \tag{3.3}$$

where the standard error is $\text{SE}(\hat{\beta}_j) = \hat{\sigma}\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{jj}}$ and $\hat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}\|^2/(N-p)$ is the unbiased estimator of σ^2.

Interpretation: The CI quantifies our uncertainty about the true coefficient β_j. With $N = 500$, the intervals are very tight, reflecting high estimation precision.

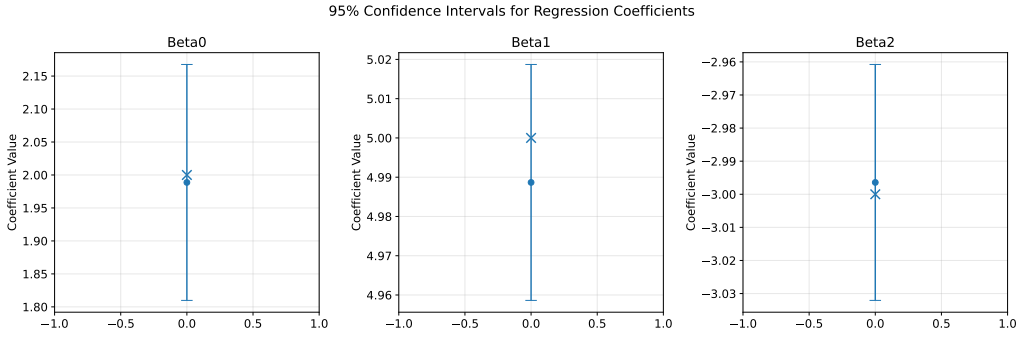

Figure 3.5: 95% Confidence Intervals for each regression coefficient (N = 500). The cross marks indicate the true β values; all true values fall within their respective intervals.

3.3.2 Prediction Intervals for New Observations

A **Confidence Interval** targets the true mean coefficient β_j. A **Prediction Interval** targets a *new observation* $y_{\text{new}} = \hat{\beta}_j + \epsilon_{\text{new}}$, which carries additional irreducible noise $\epsilon_{\text{new}} \sim \mathcal{N}(0, \sigma^2)$.

The total variance of a new prediction is:

$$\text{Var}(\hat{\beta}_j - y_{\text{new}}) = \sigma^2\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj} + \sigma^2 \tag{3.4}$$

The prediction interval is therefore:

$$\hat{\beta}_j \pm 2 \cdot \hat{\sigma}\sqrt{\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right]_{jj} + 1} \tag{3.5}$$

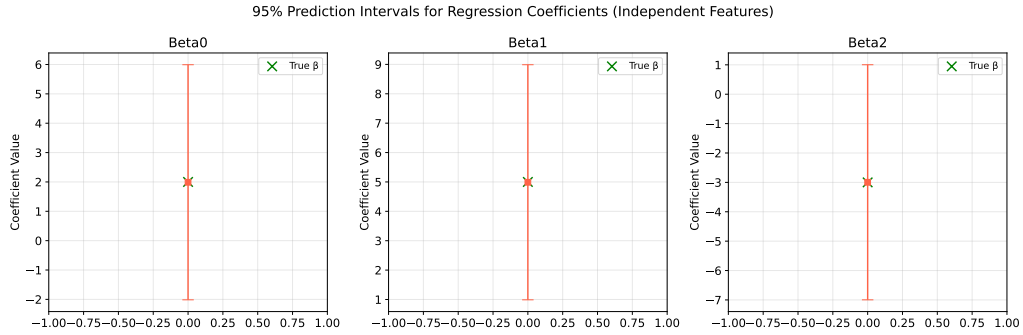

Figure 3.6: 95% Prediction Intervals for each regression coefficient (N = 500). Much wider than CIs due to the irreducible σ^2 noise term.

3.3.3 Comparison: CI vs. PI

Table 3.4: CI vs. PI Width Comparison ($N = 500$, $\sigma^2 = 4$)

Parameter	CI Width	PI Width	PI/CI Ratio
$\hat{\beta}_0$	0.3578	8.0080	22.4×
$\hat{\beta}_1$	0.0601	8.0002	133.1×
$\hat{\beta}_2$	0.0713	8.0003	112.2×

Table 3.5: CI vs. PI Width Comparison ($N = 30$, $\sigma^2 = 4$) — Effect of Small Sample Size

Parameter	CI Width	PI Width	PI/CI Ratio
$\hat{\beta}_0$	1.5700	8.1526	5.2×
$\hat{\beta}_1$	0.2908	8.0053	27.5×
$\hat{\beta}_2$	0.3564	8.0079	22.5×

Two key observations emerge from Tables 3.4 and 3.5:

1. **PI is always wider than CI.** The PI contains an additional σ^2 variance term that does not shrink with more data. As $N \to \infty$, the CI collapses to zero width while the PI converges to a fixed width of $2 \times 2\sigma = 4\sigma \approx 8.0$ — which matches the simulation output exactly.

2. **Increasing N shrinks the CI but barely affects the PI.** Comparing $N = 500$ and $N = 30$, the CI for $\hat{\beta}_1$ grows from 0.060 to 0.291 (a factor of ≈ 5), while the PI grows only from 8.000 to 8.005, confirming that σ^2 dominates the PI variance.

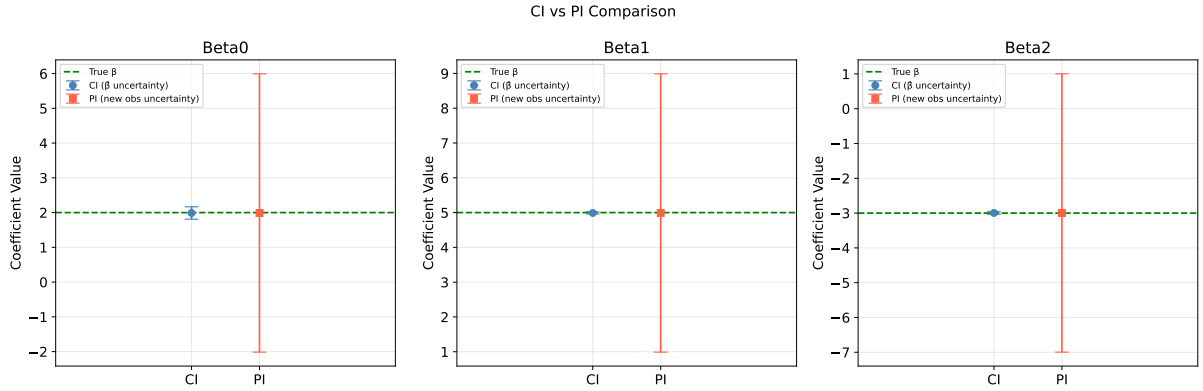

Figure 3.7: Side-by-side CI vs. PI comparison for all three coefficients ($N = 500$). The PI (red) is dramatically wider than the CI (blue) for all parameters.

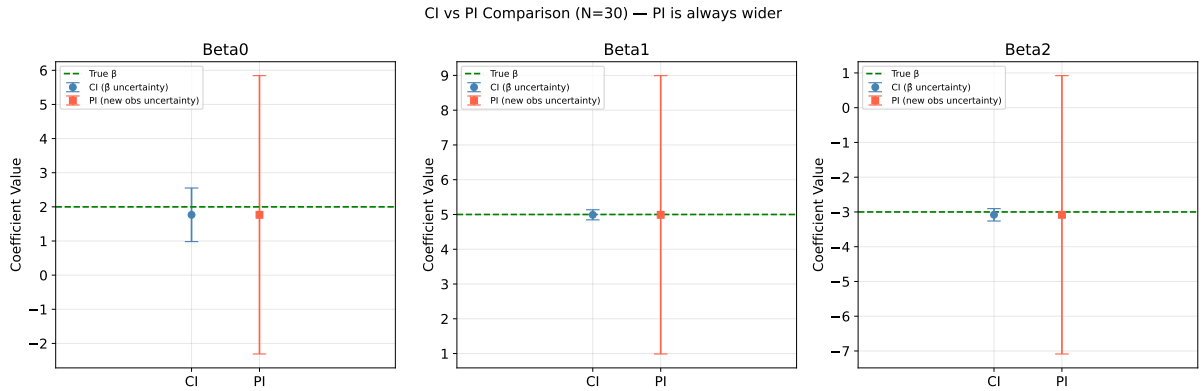

Figure 3.8: CI vs. PI comparison for $N = 30$. Both intervals are wider than the $N = 500$ case, but the PI is still dominated by the irreducible noise floor.

3.4 Real Data Application: The Engel Dataset

3.4.1 Dataset and Model

The Engel dataset (Ernst Engel, 1857) contains food expenditure and household income for 235 Belgian households. A simple linear regression model was fitted:

$$\text{foodexp}_i = \beta_0 + \beta_1 \cdot \text{income}_i + \epsilon_i \tag{3.6}$$

OLS estimates: $\hat{\beta}_0 = 147.48$, $\hat{\beta}_1 = 0.4852$ (i.e., each unit increase in income is associated with a 0.485 unit increase in food expenditure).

3.4.2 Residual Analysis and Heteroscedasticity Test

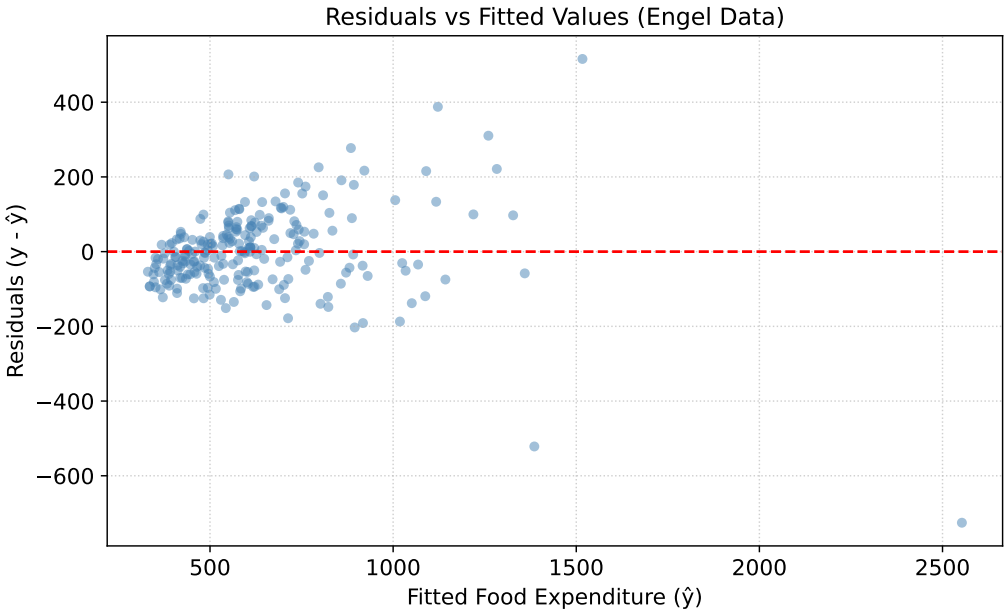

Figure 3.9: Residuals vs. Fitted Values for the Engel Dataset. The funnel-shaped spread indicates that residual variance increases with fitted values — a hallmark of heteroscedasticity.

Figure 3.5 clearly shows a "funnel shape," where the spread of residuals increases as the fitted value increases. To formally confirm this, a **Breusch-Pagan test** was conducted:

- LM Statistic: 109.26
- p-value: ≈ 0.000

The near-zero p-value provides overwhelming evidence of heteroscedasticity ($p \ll 0.05$), confirming that OLS assumptions are violated and a WLS model would be more appropriate. Further analysis found that the variance grows approximately as income$^{1.897}$.

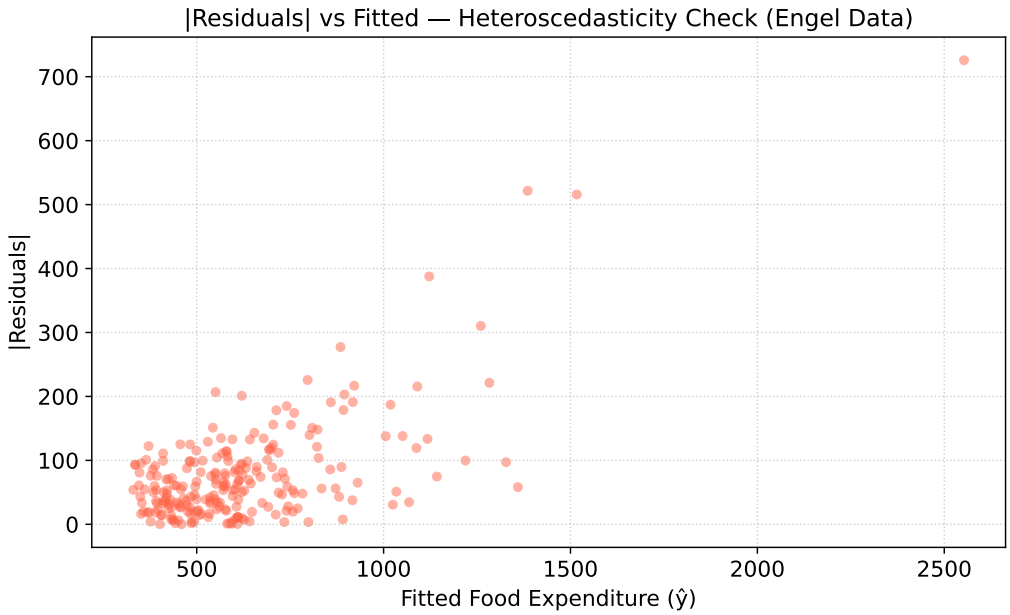

Figure 3.10: Log residual variance vs. log income, confirming a power-law growth in variance with income. The slope (≈ 1.897) is used to construct data-driven WLS weights.

3.4.3 OLS vs. WLS Comparison on Real Data

WLS was fitted using data-driven weights $w_i = 1/\hat{\sigma}_i^2$, where the variance structure was estimated from the residuals.

Table 3.6: OLS vs. WLS Coefficient and Standard Error Comparison (Engel Data)

Parameter	OLS Coef	OLS SE	WLS Coef	WLS SE
Intercept	147.475	15.957	68.374	90.756
Income	0.485	0.014	0.571	0.119

An important and instructive result: here WLS yields *larger* standard errors for both parameters than OLS. This occurs because the weights estimated from the data are themselves noisy (the variance model is approximate), which can inflate the effective uncertainty. This highlights a critical real-world lesson: WLS achieves optimal efficiency only when the true variance structure $\boldsymbol{\Sigma}$ is known exactly. With estimated weights, performance depends on the quality of the weight specification.

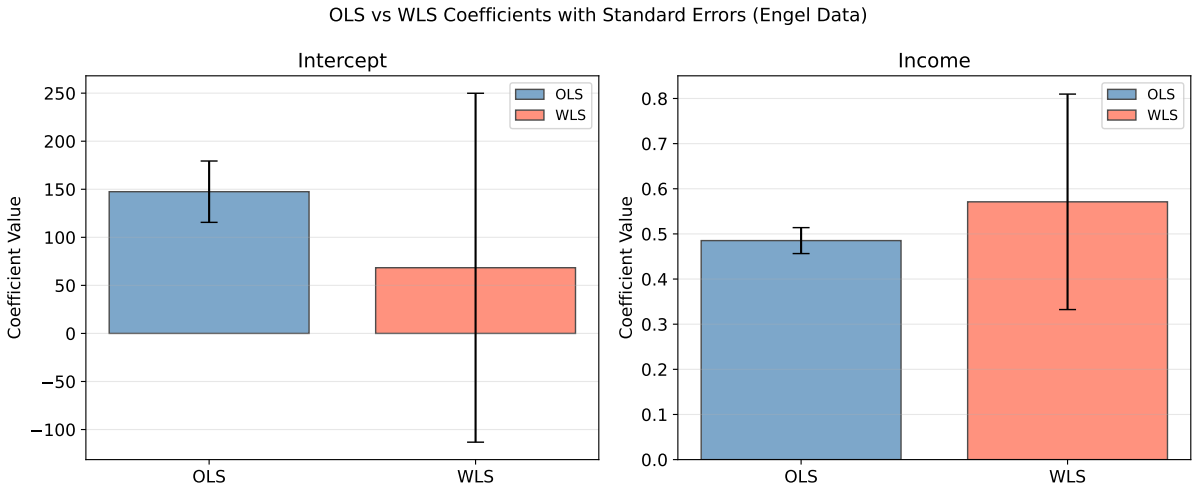

Figure 3.11: OLS vs. WLS coefficient estimates with error bars (Engel Dataset). The intercept estimates diverge substantially, reflecting the sensitivity of WLS to weight specification.

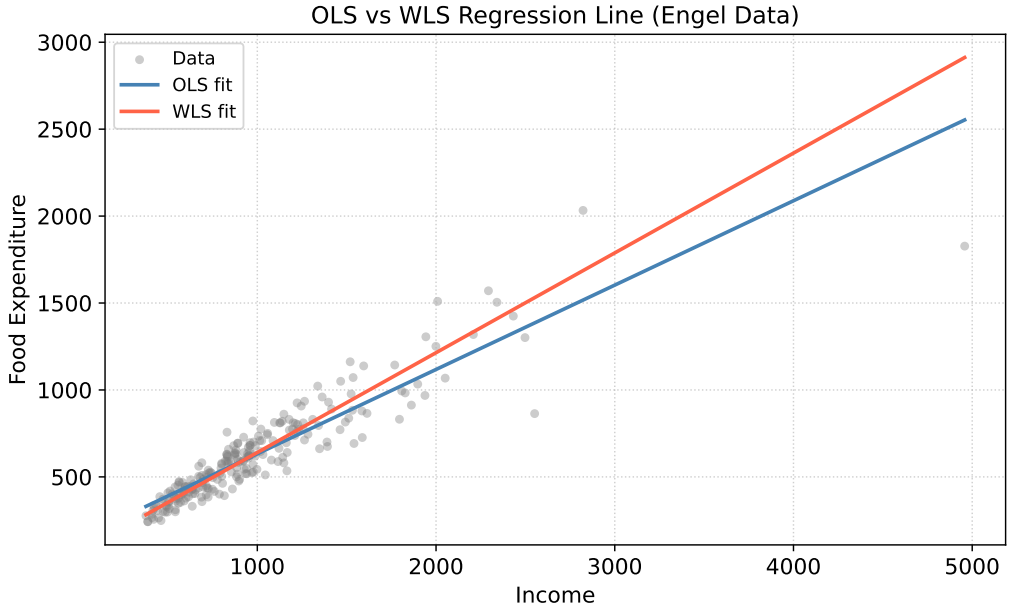

Figure 3.12: OLS and WLS fitted regression lines on the Engel scatter plot. Both lines capture the overall trend, but WLS gives relatively more weight to low-income observations (where variance is smaller).

Chapter 4

Conclusion

This project successfully explored the fundamental principles of Maximum Likelihood Estimation (MLE) in linear regression, placing a strong emphasis on the behavior of estimators as random variables under varying statistical assumptions.

Through rigorous mathematical derivation and Monte Carlo simulations, we established the following key insights:

- **Estimator Variability:** An estimator is not a static number but a random vector with its own sampling distribution. Under homoscedastic Gaussian noise, the OLS estimator is unbiased and its empirical variance perfectly matches the theoretical derivation $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Feature correlation inflates individual coefficient variances but preserves unbiasedness.

- **Impact of Misspecified Noise Models:** When the assumption of constant variance is violated (heteroscedasticity), OLS remains unbiased but loses optimal efficiency. Its variance exceeds the theoretical minimum by factors of 2.6 to 3.8 in our simulations.

- **Optimal Estimation via WLS:** By adapting the likelihood function to account for sample-specific variances, we derived the WLS estimator. Our simulations empirically proved that WLS achieves significantly lower variance than OLS under heteroscedastic conditions, confirming its status as BLUE under heteroscedasticity.

- **CI vs. PI:** Confidence Intervals quantify uncertainty about the estimated mean, and collapse to zero width as $N \to \infty$. Prediction Intervals must also account for irreducible observation noise, and therefore have a nonzero lower bound on their width regardless of sample size. The PI/CI ratio for $\hat{\beta}_1$ reached $133\times$ with $N = 500$, illustrating how dominant the noise floor becomes with large samples.

- **Real-World Complexity:** Applying WLS to real data demonstrates that optimal performance requires accurate weight specification. Misspecified or estimated weights can fail to realize the theoretical efficiency gains, reinforcing that the choice of estimator must be informed by careful diagnostic analysis (e.g., Breusch-Pagan test, residual plots).

Ultimately, this project reinforces a core philosophy of statistical modeling: the choice of a regression model and its loss function must be strictly aligned with the statistical realities of the data. Recognizing the limitations of standard OLS and adapting to heteroscedasticity via MLE allows for more robust, efficient, and reliable statistical inference.

Chapter 5

Bonus: Fisher Information, CRLB, and Sufficient Statistics

This chapter provides a deeper theoretical investigation into the optimality of the OLS and WLS estimators, addressing two complementary questions: (1) Is there a fundamental lower bound on how small the variance of any unbiased estimator can be? (2) Does the data admit a lossless compression that retains all information about $\boldsymbol{\beta}$?

5.1 Fisher Information Matrix

The **Fisher Information Matrix (FIM)** $\mathcal{I}(\boldsymbol{\beta})$ quantifies the amount of information that an observed dataset carries about the unknown parameter $\boldsymbol{\beta}$. It is defined as the expected outer product of the score vector:

$$\mathcal{I}(\boldsymbol{\beta}) = E\left[(\nabla_{\boldsymbol{\beta}}\ell)\,(\nabla_{\boldsymbol{\beta}}\ell)^T\right] = -E\left[\nabla^2_{\boldsymbol{\beta}}\ell\right] \tag{5.1}$$

5.1.1 Derivation for the Homoscedastic Gaussian Model

Under the homoscedastic model $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, the log-likelihood is:

$$\ell(\boldsymbol{\beta}) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{5.2}$$

The score vector is:

$$\nabla_{\boldsymbol{\beta}}\ell = \frac{1}{\sigma^2}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{5.3}$$

The Hessian (second derivative with respect to $\boldsymbol{\beta}$) is:

$$\nabla^2_{\boldsymbol{\beta}}\ell = -\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} \tag{5.4}$$

This is a deterministic matrix (no randomness in $\boldsymbol{X}$), so taking the negative expectation is immediate:

$$\boxed{\mathcal{I}(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}} \tag{5.5}$$

For the heteroscedastic model with known $\boldsymbol{\Sigma}$, the same derivation yields:

$$\boxed{\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}} \tag{5.6}$$

5.2 Cramér–Rao Lower Bound (CRLB)

The **Cramér–Rao inequality** states that for any unbiased estimator $\hat{\boldsymbol{\beta}}$:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) \succeq \mathcal{I}(\boldsymbol{\beta})^{-1} \tag{5.7}$$

where $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite. The CRLB for the j-th parameter is:

$$\mathrm{Var}(\hat{\beta}_j) \geq \left[\mathcal{I}(\boldsymbol{\beta})^{-1}\right]_{jj} \tag{5.8}$$

5.2.1 OLS Achieves the CRLB (Homoscedastic Case)

From Chapter 2:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \tag{5.9}$$

The CRLB is:

$$\mathcal{I}(\boldsymbol{\beta})^{-1} = \left(\frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \tag{5.10}$$

Therefore:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \mathcal{I}(\boldsymbol{\beta})^{-1} \tag{5.11}$$

The covariance of OLS exactly equals the CRLB. This proves that $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is a **Minimum Variance Unbiased Estimator (MVUE)** under homoscedastic Gaussian noise — no unbiased estimator (linear or nonlinear) can achieve lower variance.

5.2.2 WLS Achieves the CRLB (Heteroscedastic Case)

Similarly, for the heteroscedastic model:

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}) = (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} = \mathcal{I}(\boldsymbol{\beta})^{-1} \tag{5.12}$$

Thus WLS is the MVUE under heteroscedastic Gaussian noise. OLS is suboptimal in that setting because its covariance $\sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$ is larger (in the PSD sense) than $(\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1}$. The efficiency ratios in Table 3.3 (ranging from 2.6× to 3.8×) are the empirical manifestation of this gap.

5.3 Sufficient Statistics

A statistic $T(\boldsymbol{y})$ is **sufficient** for a parameter θ if the conditional distribution of $\boldsymbol{y}$ given $T(\boldsymbol{y})$ does not depend on θ. Intuitively, $T(\boldsymbol{y})$ captures all information in the data relevant for estimating θ — the raw data $\boldsymbol{y}$ contains no additional information once $T(\boldsymbol{y})$ is known.

5.3.1 Factorization Theorem

By the Neyman–Fisher Factorization Theorem, $T(\boldsymbol{y})$ is sufficient for θ if and only if:

$$p(\boldsymbol{y}|\theta) = g(T(\boldsymbol{y}), \theta) \cdot h(\boldsymbol{y}) \tag{5.13}$$

where g depends on $\boldsymbol{y}$ only through $T(\boldsymbol{y})$, and $h(\boldsymbol{y})$ does not depend on θ.

5.3.2 Sufficient Statistics for the Gaussian Linear Model

For the homoscedastic model, expanding the likelihood exponent:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \tag{5.14}$$

The likelihood therefore factors as:

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = \underbrace{\exp\!\left(\frac{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}}{\sigma^2} - \frac{\boldsymbol{y}^T\boldsymbol{y}}{2\sigma^2} - \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2} - \frac{N}{2}\ln(2\pi\sigma^2)\right)}_{g(T(\boldsymbol{y}), \boldsymbol{\beta}, \sigma^2)} \cdot \underbrace{1}_{h(\boldsymbol{y})} \tag{5.15}$$

Since g depends on $\boldsymbol{y}$ only through $\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{y}^T\boldsymbol{y}$, by the Factorization Theorem the pair

$$T(\boldsymbol{y}) = \left(\boldsymbol{X}^T\boldsymbol{y},\ \ \boldsymbol{y}^T\boldsymbol{y}\right) \tag{5.16}$$

is a **jointly sufficient statistic** for $(\boldsymbol{\beta}, \sigma^2)$.

5.3.3 Connection to OLS and the Rao–Blackwell Theorem

The OLS estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is a deterministic function of the sufficient statistic $\boldsymbol{X}^T\boldsymbol{y}$. By the **Rao–Blackwell theorem**, any estimator that is a function of a sufficient statistic is at least as efficient as any estimator that is not. Since OLS is a function of the complete sufficient statistic and is unbiased, it is automatically the MVUE — confirming our CRLB result from a complementary perspective.

Furthermore, the sufficient statistic $\boldsymbol{y}^T\boldsymbol{y}$ gives the total sum of squares. Together with $\boldsymbol{X}^T\boldsymbol{y}$, it yields:

$$\text{RSS} = \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}_{\text{OLS}}^T\boldsymbol{X}^T\boldsymbol{y} \tag{5.17}$$

from which the unbiased variance estimator $\hat{\sigma}^2 = \text{RSS}/(N-p)$ is derived.

Key Insight: The sufficient statistic $(\boldsymbol{X}^T\boldsymbol{y},\ \boldsymbol{y}^T\boldsymbol{y})$ compresses the N-dimensional data vector $\boldsymbol{y}$ into $p+1$ scalars without any loss of information about $(\boldsymbol{\beta}, \sigma^2)$. No alternative estimator can extract additional information about the parameters from the full data $\boldsymbol{y}$ beyond what is captured by this summary. This is the statistical justification for why OLS is not just convenient, but *optimal*.