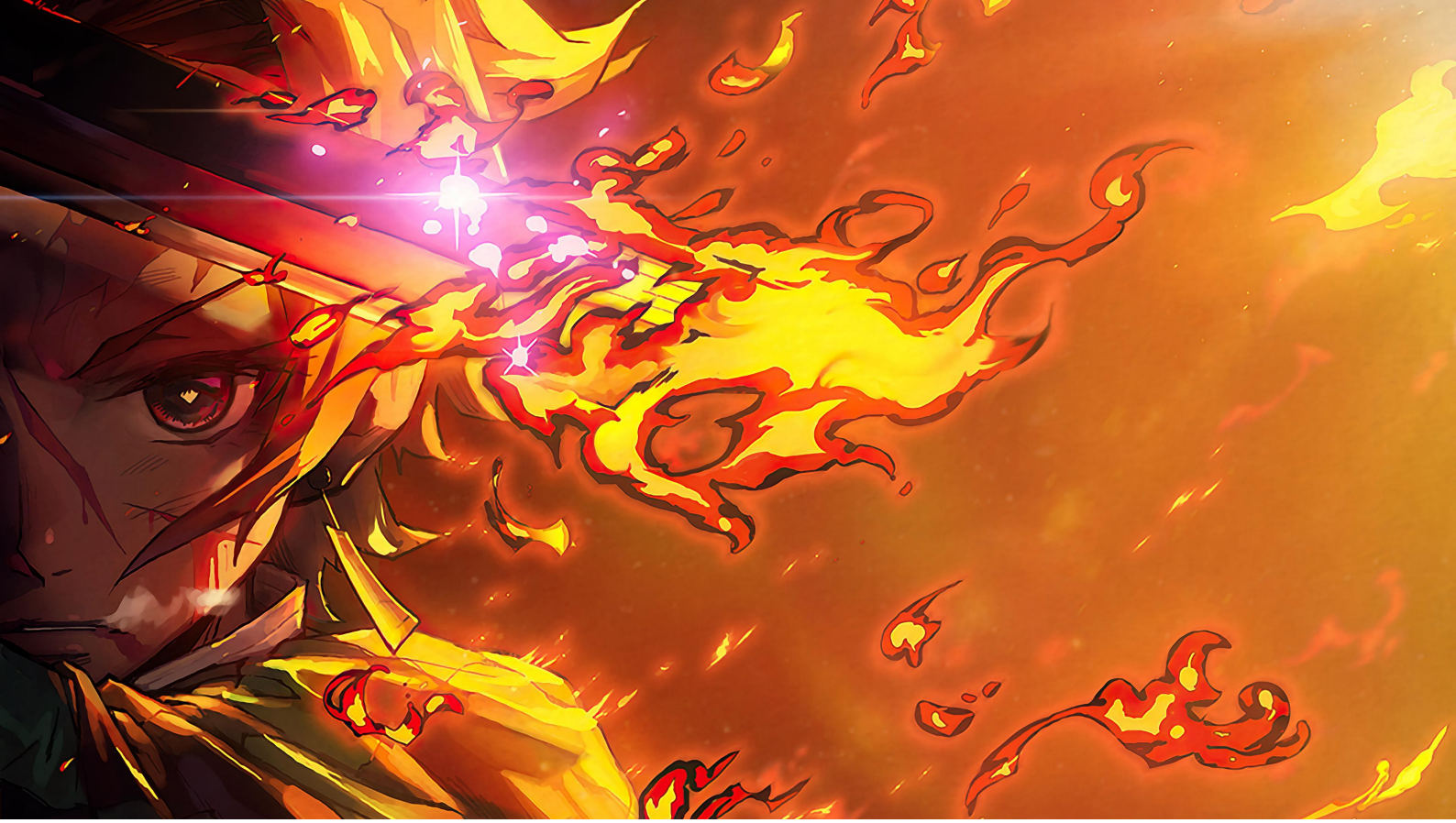

R.A. Note

Author: Yukina

Date: December 31, 2025

Contents

Part I

POO

Chapter 1

Linear Models & OLS

§1.1
LM and OLS

1.1.1 Matrix Form

- Let $(Y_i; X_{i1}, \cdots, X_{ip})$ $(1 \leqslant i \leqslant n)$ be n observations of Y and $X_1, X_2, \cdots, X_p$
- The linear regression model is given by:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

- The error terms satisfy the following Gauss-Markov conditions.

Definition 1.1. (Gauss-Markov Conditions) The error terms ε_i satisfy:

1. $\mathbb{E}(\varepsilon_i) = 0$ (no systematic bias in measurement errors)
2. $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ (homoscedasticity/constant variance)
3. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ (uncorrelated observations)

Definition 1.2. The linear regression model can be expressed in matrix form as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Here $\mathbf{X}$ is called the **design matrix** and is assumed to be of full column rank, i.e.

$$\mathrm{rank}(\mathbf{X}) = p + 1.$$

1.1.2 Estimate of β, σ^2

OLS of β

Theorem 1.3. (OLS Solution) The ordinary least squares estimate $\hat{\boldsymbol{\beta}}$ that minimizes RSS($\boldsymbol{\beta}$) is given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

provided that $\mathbf{X}^T\mathbf{X}$ is invertible.

Definition 1.4. (Hat matrix) The Hat matrix $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Remark 1.5.

$$\mathbf{H}^2 = \mathbf{H} \quad \text{and} \quad \mathbf{H}^\top = \mathbf{H}$$

The hat matrix is symmetric and idempotent.

Unbiased Estimator of Error Variance

Definition 1.6. (Residuals) $Y_i - \hat{Y}_i$ is called the residual, denoted by $\hat{\varepsilon}_i$:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \cdots, n.$$

Let the residual vector be $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n)^T$, then

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Theorem 1.7. Properties of Residuals

1. The sum of the least squares residuals is always zero:

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

2. $\mathbf{X}^T\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ (orthogonality condition)

3. $\hat{\mathbf{Y}}^T\hat{\boldsymbol{\varepsilon}} = 0$

Definition 1.8. The residual sum of squares (SSE) is defined as:

$$\text{SSE} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y},$$

Remark 1.9. Since $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$, we have

$$\text{SSE} = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y} = \hat{\boldsymbol{\varepsilon}}^T(\mathbf{I} - \mathbf{H})\hat{\boldsymbol{\varepsilon}}$$

Theorem 1.10. The error variance σ^2 can be unbiasedly estimated by:

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p-1} = \frac{1}{n-p-1}\mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

where $p+1$ is the number of unknown parameters in the linear regression model.

Proof. Taking expectations and using the cyclic property of trace:

$$\begin{aligned}
\mathbb{E}(\text{SSE}) &= \mathbb{E}\left[\text{tr}\left(\boldsymbol{\varepsilon}^T(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\right)\right] \\
&= \mathbb{E}\left[\text{tr}\left((\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)\right] \\
&= \text{tr}\left((\mathbf{I}-\mathbf{H})\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\right) \\
&= \sigma^2\text{tr}(\mathbf{I}-\mathbf{H}) \\
&= \sigma^2\left[n - \text{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\right] \\
&= \sigma^2\left[n - \text{tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\right)\right] \\
&= \sigma^2(n-p-1),
\end{aligned}$$

Therefore,

$$\mathbb{E}\left(\frac{\text{SSE}}{n-p-1}\right) = \sigma^2,$$

which proves that $\hat{\sigma}^2$ is an unbiased estimator of σ^2. □

1.1.3 Properties of OLS

Theorem 1.11. (Gauss-Markov Theorem) Under the Gauss-Markov assumptions:

1. $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ (unbiasedness)
2. $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
3. $\hat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator (BLUE)

Proof. Part (1): Unbiasedness. Taking expectations:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{\beta}$$

since $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ by Gauss-Markov assumption (i).

Part (2): Variance. The variance-covariance matrix is:

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T] \\
&= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

Last equation is based on Gauss-Markov assumptions (ii) and (iii), $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2\mathbf{I}$

Part (3): Best Linear Unbiased Estimator (BLUE). Let $\tilde{\boldsymbol{\beta}} = \mathbf{CY}$ be any other linear unbiased estimator of $\boldsymbol{\beta}$. Since it's unbiased:

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \mathbf{C}\mathbb{E}[\mathbf{Y}] = \mathbf{CX}\boldsymbol{\beta} = \boldsymbol{\beta} \quad \Rightarrow \quad \mathbf{CX} = \mathbf{I}$$

The variance of this alternative estimator is:

$$\text{Var}(\tilde{\boldsymbol{\beta}}) = \mathbf{C}\text{Var}(\mathbf{Y})\mathbf{C}^T = \sigma^2\mathbf{C}\mathbf{C}^T$$

Let $\mathbf{D} = \mathbf{C} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Then:

$$\mathbf{DX} = \mathbf{CX} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I} - \mathbf{I} = \mathbf{0}$$

Now compute:

$$\begin{aligned}\mathbf{C}\mathbf{C}^T &= [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}][(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}]^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^T + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^T + \mathbf{DX}(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

Since $\mathbf{DX} = \mathbf{0}$ and $\mathbf{X}^T\mathbf{D}^T = (\mathbf{DX})^T = \mathbf{0}$, we have:

$$\mathbf{C}\mathbf{C}^T = (\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^T$$

Therefore:

$$\text{Var}(\tilde{\boldsymbol{\beta}}) = \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^T] = \text{Var}(\hat{\boldsymbol{\beta}}) + \sigma^2\mathbf{D}\mathbf{D}^T$$

Since $\mathbf{D}\mathbf{D}^T$ is positive semi-definite, $\text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}})$ is positive semi-definite, meaning $\hat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators. □

Corollary 1.12. Under the Gauss-Markov assumptions:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Theorem 1.13. Under the Gauss-Markov assumptions:

1. $\dfrac{SSE}{\sigma^2} = \dfrac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$;
2. $\hat{\boldsymbol{\beta}}$ and SSE (or $\hat{\sigma}^2$) are mutually **independent**.

Proof. part(1). $SSE = \boldsymbol{\varepsilon}^T(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$. Since $(\mathbf{I} - \mathbf{H})$ is a symmetric idempotent matrix, we have

$$\text{rank}(\mathbf{I} - \mathbf{H}) = \text{tr}(\mathbf{I} - \mathbf{H}) = n - p - 1.$$

Therefore, there exists an $n \times n$ orthogonal matrix $\mathbf{P}$ such that:

$$\mathbf{P}^T(\mathbf{I} - \mathbf{H})\mathbf{P} = \begin{pmatrix} \mathbf{I}_{n-p-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Let:

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_n)^T = \mathbf{P}^T \boldsymbol{\varepsilon},$$

then $E(\boldsymbol{\eta}) = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{\eta}) = \sigma^2 \mathbf{P}^T \mathbf{P} = \sigma^2 \mathbf{I}$, so $\boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, and:

$$\frac{1}{\sigma^2} SSE = \frac{1}{\sigma^2} \boldsymbol{\eta}^T \mathbf{P}^T (\mathbf{I} - \mathbf{H}) \mathbf{P} \boldsymbol{\eta} = \frac{1}{\sigma^2} (\eta_1^2 + \eta_2^2 + \cdots + \eta_{n-p-1}^2).$$

Since $\eta_1, \eta_2, \cdots, \eta_{n-p-1}$ are independent $N(0, \sigma^2)$ random variables, we have:

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-p-1}.$$

part(2). Under the Gauss–Markov conditions, we have $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$. The least squares estimator $\hat{\boldsymbol{\beta}}$ and residual vector $\hat{\boldsymbol{\varepsilon}}$ are linear transformations of $\mathbf{Y}$:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\varepsilon}} \end{bmatrix} = \begin{bmatrix} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \\ \mathbf{I} - \mathbf{H} \end{bmatrix} \mathbf{Y}$$

Since $\mathbf{Y}$ is multivariate normal, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}})$ follows a joint normal distribution.

Now compute the covariance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$:

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}) &= \mathrm{Cov}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}, (\mathbf{I} - \mathbf{H}) \mathbf{Y} \right) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \cdot \mathrm{Cov}(\mathbf{Y}, \mathbf{Y}) \cdot (\mathbf{I} - \mathbf{H})^\top \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \cdot (\sigma^2 \mathbf{I}) \cdot (\mathbf{I} - \mathbf{H}) \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{I} - \mathbf{H}) = \mathbf{0} \end{aligned}$$

Since $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}})$ has a joint normal distribution and their covariance is zero, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are independent. □

Remark 1.14. partition the $n \times n$ orthogonal matrix $\mathbf{P}$ as $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2)$, where $\mathbf{P}_1$ consists of the first $n - p - 1$ columns of $\mathbf{P}$ and $\mathbf{P}_2$ consists of the last $p + 1$ columns. Partition $\boldsymbol{\eta}$ accordingly:

$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{P}_1^T \\ \mathbf{P}_2^T \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \mathbf{P}_1^T \boldsymbol{\varepsilon} \\ \mathbf{P}_2^T \boldsymbol{\varepsilon} \end{pmatrix}.$$

Since $\boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, we have:

$$\boldsymbol{\eta}_1 = \mathbf{P}_1^T \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n-p-1}), \quad \boldsymbol{\eta}_2 = \mathbf{P}_2^T \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{p+1}),$$

and $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are independent,so we have:

$$SSE = \boldsymbol{\varepsilon}^T (\mathbf{I} - \mathbf{H}) \boldsymbol{\varepsilon} = \boldsymbol{\eta}_1^T \boldsymbol{\eta}_1. \tag{1.1}$$

Now consider:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\varepsilon} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{P} \boldsymbol{\eta} = \mathbf{D} \boldsymbol{\eta},$$

where $\mathbf{D} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P} = (\mathbf{D}_1, \mathbf{D}_2)$, with $\mathbf{D}_1$ and $\mathbf{D}_2$ consisting of the first $n-p-1$ and last $p+1$ columns of $\mathbf{D}$, respectively. Since $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{H}) = \mathbf{0}$, we have:

$$(\mathbf{D}_1, \mathbf{D}_2)\begin{pmatrix}\mathbf{I}_{n-p-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} = \mathbf{D}\mathbf{P}^T(\mathbf{I}-\mathbf{H})\mathbf{P} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{H})\mathbf{P} = \mathbf{0},$$

which implies $\mathbf{D}_1 = \mathbf{0}$. Therefore:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{D}_1, \mathbf{D}_2)\begin{pmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{pmatrix} = \mathbf{D}_2\boldsymbol{\eta}_2. \tag{1.2}$$

Since $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are independent, equations (1.1) and (1.2) show that SSE and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ are independent, and therefore SSE and $\hat{\boldsymbol{\beta}}$ are independent.

Proposition 1.15. If the error vector $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then the residual vector

$$\hat{\boldsymbol{\varepsilon}} \sim N(\mathbf{0}, \sigma^2(\mathbf{I}-\mathbf{H}))$$

Proof. The residual vector can be expressed as:

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}-\mathbf{H})\mathbf{Y} = (\mathbf{I}-\mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon},$$

since $(\mathbf{I}-\mathbf{H})\mathbf{X} = \mathbf{0}$.

$$\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0} \implies \mathbb{E}[\hat{\boldsymbol{\varepsilon}}] = \mathbb{E}[(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}] = (\mathbf{I}-\mathbf{H})\mathbb{E}[\boldsymbol{\varepsilon}] = (\mathbf{I}-\mathbf{H})\mathbf{0} = \mathbf{0}. \quad \text{Expectation}$$

The variance-covariance matrix of the residual vector is:

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) &= \mathbb{E}[\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^T] - \mathbb{E}[\hat{\boldsymbol{\varepsilon}}]\mathbb{E}[\hat{\boldsymbol{\varepsilon}}]^T\\ &= \mathbb{E}[(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T(\mathbf{I}-\mathbf{H})^T] - \mathbf{0}\\ &= (\mathbf{I}-\mathbf{H})\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T](\mathbf{I}-\mathbf{H})^T\end{aligned}$$

Since $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2\mathbf{I}$ and $(\mathbf{I}-\mathbf{H})$ is symmetric and idempotent:

$$\mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = (\mathbf{I}-\mathbf{H})(\sigma^2\mathbf{I})(\mathbf{I}-\mathbf{H})^T = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \sigma^2(\mathbf{I}-\mathbf{H}).$$

□

1.1.4 Constrained Linear Regression

Definition 1.16. (Constrained LM) Consider the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with the linear constraint:

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{b},$$

where:

- $\mathbf{A}$ is an $m \times (p+1)$ known matrix with full row rank ($\mathrm{rank}(\mathbf{A}) = m$)

- $\mathbf{b}$ is an m-dimensional known vector
- $m < p + 1$ (number of constraints less than number of parameters)

Theorem 1.17. The constrained least squares estimator of $\boldsymbol{\beta}$ under the constraint $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$ is given by:

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b}),$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the unconstrained least squares estimator.

This estimator satisfies:

1. $\mathbf{A}\hat{\boldsymbol{\beta}}_c = \mathbf{b}$ (constraint satisfaction)
2. For all $\boldsymbol{\beta}$ satisfying $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$, we have $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \geqslant \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_c\|^2$ (optimality)

Proof. The Lagrangian is:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}^T(\mathbf{A}\boldsymbol{\beta} - \mathbf{b}),$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the vector of Lagrange multipliers.Take partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + 2\mathbf{A}^T\boldsymbol{\lambda} = \mathbf{0},$$
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = 2(\mathbf{A}\boldsymbol{\beta} - \mathbf{b}) = \mathbf{0}.$$

This gives the system:

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{A}^T\boldsymbol{\lambda} = \mathbf{X}^T\mathbf{Y}, \tag{1.3}$$
$$\mathbf{A}\boldsymbol{\beta} = \mathbf{b}. \tag{1.4}$$

From equation (1.3):

$$\hat{\boldsymbol{\beta}}_c = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\hat{\boldsymbol{\lambda}}.$$

Substitute into equation (1.4):

$$\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\hat{\boldsymbol{\lambda}} = \mathbf{b}.$$

Solve for $\hat{\boldsymbol{\lambda}}$:

$$\hat{\boldsymbol{\lambda}} = (\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b}).$$

Substitute back to obtain the constrained estimator:

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b}).$$

(i) Direct computation shows $\mathbf{A}\hat{\boldsymbol{\beta}}_c = \mathbf{b}$.

(ii) For any $\boldsymbol{\beta}$ satisfying $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$, consider the decomposition:

$$\begin{aligned}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 &= \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_c + \mathbf{X}(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta})\|^2 \\ &= \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_c\|^2 + \|\mathbf{X}(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta})\|^2 + 2(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta})^T\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_c).\end{aligned}$$

The cross term vanishes because:

$$\begin{aligned}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_c) &= \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_c \\ &= \mathbf{A}^T\hat{\boldsymbol{\lambda}} \quad \text{(from equation 1.3)},\end{aligned}$$

and

$$(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta})^T\mathbf{A}^T\hat{\boldsymbol{\lambda}} = (\mathbf{A}\hat{\boldsymbol{\beta}}_c - \mathbf{A}\boldsymbol{\beta})^T\hat{\boldsymbol{\lambda}} = (\mathbf{b} - \mathbf{b})^T\hat{\boldsymbol{\lambda}} = 0.$$

Therefore, $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \geqslant \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_c\|^2$ for all feasible $\boldsymbol{\beta}$. □

Remark 1.18. (Geometric Interpretation) The constrained estimator can be viewed as the projection of the unconstrained estimator onto the constraint subspace $\{\boldsymbol{\beta} : \mathbf{A}\boldsymbol{\beta} = \mathbf{b}\}$.

Theorem 1.19. Under the normality assumption $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$:

1. $\hat{\boldsymbol{\beta}}_c \sim N(\boldsymbol{\beta}, \sigma^2\mathbf{V}_c)$, where

$$\mathbf{V}_c = (\mathbf{X}^T\mathbf{X})^{-1} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}$$

2. The constrained residual sum of squares is:

$$SSE_c = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_c\|^2 = SSE + (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b})$$

3. $\dfrac{SSE_c - SSE}{\sigma^2} \sim \chi^2(m)$ and is independent of SSE

Proof. Part (1): Since $\hat{\boldsymbol{\beta}}_c$ is a linear function of $\mathbf{Y}$, it follows a normal distribution. The mean and variance can be computed directly from the expression for $\hat{\boldsymbol{\beta}}_c$.

Part (2): Using the expression for $\hat{\boldsymbol{\beta}}_c$ and some algebra:

$$\begin{aligned}SSE_c &= \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_c\|^2 \\ &= \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_c)\|^2 \\ &= SSE + \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_c)\|^2 \quad \text{(cross term vanishes)}.\end{aligned}$$

The result follows by substituting the expression for $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_c$.

Part (3): This follows from the independence of $\mathbf{A}\hat{\boldsymbol{\beta}}$ and SSE, and the fact that $\mathbf{A}\hat{\boldsymbol{\beta}} \sim N(\mathbf{A}\boldsymbol{\beta}, \sigma^2\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)$. □

§1.2

Hypothesis Testing in Linear Models

1.2.1 Foundational Concepts

Definition 1.20. (Errors) **1. Error Sum of Squares (SSE)**

- *Component form:*

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- *Matrix form:*

$$\text{SSE} = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

  where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix.

- *Degrees of freedom:* $n - p - 1$

2. Regression Sum of Squares (SSR)

- *Component form:*

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- *Matrix form:*

$$\text{SSR} = \mathbf{Y}^T(\mathbf{H} - \frac{1}{n}\mathbf{J})\mathbf{Y}$$

  where $\mathbf{J} = \mathbf{1}\mathbf{1}^T$ is the matrix of ones.

- *Degrees of freedom:* p

3. Total Sum of Squares (SST)

- *Component form:*

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- *Matrix form:*

$$\text{SST} = \mathbf{Y}^T(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y}$$

- *Degrees of freedom:* $n - 1$

Remark 1.21.

$$\begin{aligned}\sum_{i=1}^{n} (Y_i - \bar{Y})^2 &= \sum_{i-1}^{n} Y_i^2 - n\left(\bar{Y}\right)^2 \\ &= \mathbf{Y}^T\mathbf{Y} - \frac{1}{n}\left(\mathbf{Y}^T\mathbf{1}\mathbf{1}^T\mathbf{Y}\right)\end{aligned}$$

Theorem 1.22. The total sum of squares decomposes as:

$$\text{SST} = \text{SSR} + \text{SSE}$$

Proof. Using the matrix forms:

$$\begin{aligned}
\text{SSR} + \text{SSE} &= \mathbf{Y}^T(\mathbf{H} - \frac{1}{n}\mathbf{J})\mathbf{Y} + \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y} \\
&= \mathbf{Y}^T\left[(\mathbf{H} - \frac{1}{n}\mathbf{J}) + (\mathbf{I} - \mathbf{H})\right]\mathbf{Y} \\
&= \mathbf{Y}^T(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y} \\
&= \text{SST}
\end{aligned}$$

□

1.2.2 Linear Hypothesis Testing

Definition 1.23. (Linear Hypothesis) A general linear hypothesis in the linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ takes the form:

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{b} \quad \text{vs} \quad H_1 : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{b},$$

Example 1.24. Common Hypothesis Testing Scenarios:

- **Overall Model Significance**: Test whether all slope coefficients are zero

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

- **Individual Coefficient Test**: Test whether a specific predictor is significant

$$H_0 : \beta_j = 0$$

- **Equality of Effects**: Test whether two predictors have equal effects

$$H_0 : \beta_i = \beta_j$$

- **Linear Combination Test**: Test specific linear relationships among parameters

$$H_0 : c_1\beta_1 + c_2\beta_2 + \cdots + c_p\beta_p = d$$

- **Subset Significance**: Test whether a group of predictors are jointly significant

$$H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$$

1.2.3 F-Test Theory

Theorem 1.25. (General F-Test) For the normal linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2\mathbf{I})$, consider testing

$$H_0 : \mathbf{A}\beta = \mathbf{b} \quad \text{vs} \quad H_1 : \mathbf{A}\beta \neq \mathbf{b}$$

The F-test statistic is:

$$F = \frac{(\text{SSE}(H_0) - \text{SSE})/m}{\text{SSE}/(n-p-1)} \sim F(m, n-p-1)$$

where:

$$\text{SSE} = \mathbf{Y}^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y} \quad \text{(Full model)}$$
$$\text{SSE}(H_0) = \min_{\mathbf{A}\beta=\mathbf{b}} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \quad \text{(Constrained model)}$$

Equivalently, the numerator can be expressed as:

$$\text{SSE}(H_0) - \text{SSE} = (\mathbf{A}\hat{\beta} - \mathbf{b})^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\beta} - \mathbf{b})$$

Global Significance Testing

Definition 1.26. (Global Null Hypothesis) Tests whether the regression model has any explanatory power:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs} \quad H_1 : \text{At least one } \beta_j \neq 0$$

Theorem 1.27. (Global F-Test Statistic)

$$F = \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)} \sim F(p, n-p-1)$$

Proof. For the global null hypothesis:

- Constraint matrix: $\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_p \end{pmatrix}$, $\mathbf{b} = \mathbf{0}$
- $\text{SSE}(H_0) = \text{SST}$ (only intercept model)
- $\text{SSE}(H_0) - \text{SSE} = \text{SST} - \text{SSE} = \text{SSR}$

The result follows directly from the general F-test. □

Example 1.28. (ANOVA Table for Global Test)

Source	Sum of Squares	df	Mean Square	F
Regression	SSR	p	MSR = SSR/p	MSR/MSE
Error	SSE	$n-p-1$	MSE = SSE/$(n-p-1)$	
Total	SST	$n-1$		

Testing Individual Coefficients

Definition 1.29. (Single Coefficient Test)

$$H_0 : \beta_j = c \quad \text{vs} \quad H_1 : \beta_j \neq c$$

Theorem 1.30. (t-Test for Single Coefficient)

$$t = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t(n-p-1)$$

where $(\mathbf{X}^T\mathbf{X})^{-1}_{jj}$ is the j-th diagonal element.

Derivation from F-Test Framework. For $H_0 : \beta_j = c$, the constraint is $\mathbf{A}\boldsymbol{\beta} = c$ where $\mathbf{A}$ has 1 in position j and 0 elsewhere.

From the general F-test:

$$F = \frac{(\text{SSE}(H_0) - \text{SSE})/1}{\text{SSE}/(n-p-1)} \sim F(1, n-p-1)$$

It can be shown that:

$$\text{SSE}(H_0) - \text{SSE} = \frac{(\hat{\beta}_j - c)^2}{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}$$

Therefore:

$$F = \frac{(\hat{\beta}_j - c)^2}{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}} = t^2$$

Since $F(1, \nu)$ is the square of $t(\nu)$, we have:

$$t = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t(n-p-1)$$

□

Direct Distributional Proof. From the distributional properties of OLS:

- $\hat{\beta}_j \sim N(\beta_j, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj})$
- $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$
- $\hat{\beta}_j$ and $\hat{\sigma}^2$ are independent

Under $H_0 : \beta_j = c$:

$$Z = \frac{\hat{\beta}_j - c}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim N(0,1)$$

By definition of the t-distribution:

$$t = \frac{Z}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2(n-p-1)}}} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t(n-p-1)$$

□

Testing Subsets of Coefficients

Definition 1.31. (Subset Hypothesis) Test whether a subset of k coefficients are simultaneously zero:

$$H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_k} = 0$$

Theorem 1.32. (Partial F-Test) Let SSE_f be the error sum of squares from the full model and SSE_r from the reduced model (excluding the k coefficients). Then:

$$F = \frac{(\text{SSE}_r - \text{SSE}_f)/k}{\text{SSE}_f/(n-p-1)} \sim F(k, n-p-1)$$

Proof. This is a special case of the general F-test where $\mathbf{A}$ selects the k coefficients to test and $\mathbf{b} = \mathbf{0}$. □

Testing Linear Combinations

Definition 1.33. (Linear Combination Test)

$$H_0 : \mathbf{c}^T\boldsymbol{\beta} = d \quad \text{vs} \quad H_1 : \mathbf{c}^T\boldsymbol{\beta} \neq d$$

where $\mathbf{c}$ is a $(p+1) \times 1$ vector of constants.

Theorem 1.34. (t-Test for Linear Combination)

$$t = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - d}{\hat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \sim t(n-p-1)$$

1.2.4 Comprehensive Summary

Test Type	Hypothesis	Statistic	Distribution
Global Regression	$\beta_1 = \cdots = \beta_p = 0$	$F = \dfrac{\text{SSR}/p}{\text{SSE}/(n-p-1)}$	$F(p, n-p-1)$
Single Coefficient	$\beta_j = c$	$t = \dfrac{\hat{\beta}_j - c}{\text{SE}(\hat{\beta}_j)}$	$t(n-p-1)$
Subset of Coefficients	$\beta_{j_1} = \cdots = \beta_{j_k} = 0$	$F = \dfrac{(\text{SSE}_r - \text{SSE}_f)/k}{\text{SSE}_f/(n-p-1)}$	$F(k, n-p-1)$
Linear Combination	$\mathbf{c}^T\boldsymbol{\beta} = d$	$t = \dfrac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - d}{\text{SE}(\mathbf{c}^T\hat{\boldsymbol{\beta}})}$	$t(n-p-1)$
General Constraints	$\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$	$F = \dfrac{(\text{SSE}(H_0) - \text{SSE})/m}{\text{SSE}/(n-p-1)}$	$F(m, n-p-1)$

Table 1.1: Hypothesis Test Decision Rules

Test	Statistic	Critical Value	P-value
Chi-square	χ^2	$\chi^2 > \chi^2_{\alpha,df}$	$p < \alpha$
F-test	F	$F > F_{\alpha,df_1,df_2}$	$p < \alpha$
t-test	t	$\lvert t\rvert > t_{\alpha/2,df}$ (two-tailed)	$p < \alpha$

§1.3

Multicollinearity and Ridge Estimation

1.3.1 Standardized Linear Regression Model

Definition 1.35. (Standardized Variables) Let $(Y_i; X_{i1}, \cdots, X_{ip})$ for $1 \leqslant i \leqslant n$ be observations from the random vector $(Y; X_1, X_2, \cdots, X_p)$. Define the sample statistics:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad j = 1, \ldots, p$$

$$s_Y^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad s_j^2 = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2, \quad j = 1, \ldots, p$$

The standardized variables are:

$$\tilde{Y} = \frac{Y - \bar{Y}}{s_Y}, \quad \tilde{X}_j = \frac{X_j - \bar{X}_j}{s_j}, \quad j = 1, \ldots, p$$

with corresponding standardized observations:

$$\tilde{Y}_i = \frac{Y_i - \bar{Y}}{s_Y}, \quad \tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, p$$

Definition 1.36. (Matrix Formulation) Define the standardized response vector and design matrix:

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \vdots \\ \tilde{Y}_n \end{pmatrix}, \quad \tilde{\mathbf{X}} = \begin{pmatrix} \tilde{X}_{11} & \tilde{X}_{12} & \cdots & \tilde{X}_{1p} \\ \tilde{X}_{21} & \tilde{X}_{22} & \cdots & \tilde{X}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{X}_{n1} & \tilde{X}_{n2} & \cdots & \tilde{X}_{np} \end{pmatrix}$$

The standardized linear regression model is:

$$\tilde{\mathbf{Y}} = \alpha\mathbf{1} + \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

where $\mathbf{1}$ is an n-dimensional vector of ones, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are the standardized coefficients.

Theorem 1.37. For the standardized linear regression model:

1. The least squares estimates are:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}, \quad \hat{\alpha} = 0$$

2. The model can be written through the origin:

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

3. The variance of the estimator is:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$$

4. $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ is the sample correlation matrix of the original explanatory variables

Proof. Note that $\mathbf{1}^T\tilde{\mathbf{X}} = 0$ and $\mathbf{1}^T\tilde{\mathbf{Y}} = 0$. The parameter estimates are:

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}\end{pmatrix} = \left((\mathbf{1},\tilde{\mathbf{X}})^T(\mathbf{1},\tilde{\mathbf{X}})\right)^{-1}(\mathbf{1},\tilde{\mathbf{X}})^T\tilde{\mathbf{Y}} = \begin{pmatrix} n & 0\\ 0 & \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{1}^T\tilde{\mathbf{Y}}\\ \tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}\end{pmatrix}$$
$$= \begin{pmatrix} 1/n & 0\\ 0 & (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\end{pmatrix}\begin{pmatrix}0\\ \tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}\end{pmatrix} = \begin{pmatrix}0\\ (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}\end{pmatrix}$$

The correlation matrix property follows from:

$$(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})_{jk} = \sum_{i=1}^{n}\tilde{X}_{ij}\tilde{X}_{ik} = \frac{\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(X_{ik}-\bar{X}_k)}{s_j s_k}$$

□

1.3.2 Multicollinearity and Its Effects

Definition 1.38. (Multicollinearity) The explanatory variables $X_1, X_2, \cdots, X_p$ exhibit multicollinearity if there exists a non-zero vector $\mathbf{c} = (c_1, \ldots, c_p)^T$ such that:

$$c_1X_1 + c_2X_2 + \cdots + c_pX_p \approx \text{constant}$$

In standardized form, this is equivalent to:

$$c_1\tilde{X}_1 + c_2\tilde{X}_2 + \cdots + c_p\tilde{X}_p \approx 0$$

Theorem 1.39. Let $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$ be the eigenvalues of $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$. The following are equivalent:

1. Multicollinearity exists among the explanatory variables

2. $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ is ill-conditioned (some $\lambda_j \approx 0$)

3. The condition number $k = \lambda_1/\lambda_p$ is large

4. $\det(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}) = \prod_{j=1}^{p} \lambda_j \approx 0$

Proof. The equivalence follows from the spectral decomposition $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathbf{\Phi\Lambda\Phi}^T$. If $\lambda_j \approx 0$ for some j, then the corresponding eigenvector $\boldsymbol{\varphi}_j$ satisfies:

$$\|\tilde{\mathbf{X}}\boldsymbol{\varphi}_j\|^2 = \boldsymbol{\varphi}_j^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\boldsymbol{\varphi}_j = \lambda_j \approx 0$$

Thus $\tilde{\mathbf{X}}\boldsymbol{\varphi}_j \approx \mathbf{0}$, indicating an approximate linear dependency among the columns of $\tilde{\mathbf{X}}$. □

Theorem 1.40. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$. Then:

1. Total variance: $\sum_{j=1}^{p} \mathrm{Var}(\hat{\beta}_j) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$

2. Expected squared norm: $E(\|\hat{\boldsymbol{\beta}}\|^2) = \|\boldsymbol{\beta}\|^2 + \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$

3. Individual variance: $\mathrm{Var}(\hat{\beta}_j) = \sigma^2[(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}]_{jj}$

Proof. (1) Since $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$, we have:

$$\sum_{j=1}^{p} \mathrm{Var}(\hat{\beta}_j) = \mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) = \sigma^2\mathrm{tr}((\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

(2) Note that:

$$E(\|\hat{\boldsymbol{\beta}}\|^2) = \sum_{j=1}^{p} E(\hat{\beta}_j^2) = \sum_{j=1}^{p} [\mathrm{Var}(\hat{\beta}_j) + (E(\hat{\beta}_j))^2] = \mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) + \|\boldsymbol{\beta}\|^2$$

(3) For individual variance, $\mathrm{Var}(\hat{\beta}_j)$ is the jth diagonal element of $\sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$. □

1.3.3 Ridge Estimation

Definition 1.41. (Ridge Estimator) For the standardized linear regression model $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the ridge estimator is:

$$\hat{\boldsymbol{\beta}}(c) = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + c\mathbf{I})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}, \quad c > 0$$

where c is the ridge (shrinkage) parameter.

Theorem 1.42. (Geometric Interpretation) The ridge estimator solves the constrained optimization problem:

$$\min_{\boldsymbol{\beta}} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|^2 \leqslant t$$

where c is the Lagrange multiplier corresponding to the constraint $\|\boldsymbol{\beta}\|^2 \leqslant t$.

Proof. The Lagrangian is:

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) = \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2 + \lambda(\|\boldsymbol{\beta}\|^2 - t)$$

Taking derivatives:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -2\tilde{\mathbf{X}}^T(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = \mathbf{0}$$

Solving gives $\hat{\boldsymbol{\beta}}(\lambda) = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{I})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}$. □

Theorem 1.43. (Bias-Variance Properties) The ridge estimator has:

1. **Bias:** $E[\hat{\boldsymbol{\beta}}(c)] = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + c\mathbf{I})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\boldsymbol{\beta}$
2. **Variance:** $\text{Var}[\hat{\boldsymbol{\beta}}(c)] = \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + c\mathbf{I})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + c\mathbf{I})^{-1}$
3. **Shrinkage:** $\|\hat{\boldsymbol{\beta}}(c)\| < \|\hat{\boldsymbol{\beta}}\|$ for all $c > 0$

Proof. (3) Using spectral decomposition $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T$ and $\mathbf{Z} = \tilde{\mathbf{X}}\boldsymbol{\Phi}$, $\boldsymbol{\alpha} = \boldsymbol{\Phi}^T\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\alpha}}(c) = (\boldsymbol{\Lambda} + c\mathbf{I})^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\alpha}}$$

Since $\|\hat{\boldsymbol{\alpha}}(c)\|^2 = \sum_{j=1}^{p}\left(\frac{\lambda_j}{\lambda_j + c}\right)^2 \hat{\alpha}_j^2 < \sum_{j=1}^{p}\hat{\alpha}_j^2 = \|\hat{\boldsymbol{\alpha}}\|^2$, and $\boldsymbol{\Phi}$ is orthogonal, we have $\|\hat{\boldsymbol{\beta}}(c)\| < \|\hat{\boldsymbol{\beta}}\|$. □

Theorem 1.44. In canonical form, the MSE of the ridge estimator is:

$$\text{MSE}[\hat{\boldsymbol{\alpha}}(c)] = \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j + c)^2} + c^2\sum_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j + c)^2}$$

There exists $c > 0$ such that $\text{MSE}[\hat{\boldsymbol{\beta}}(c)] < \text{MSE}[\hat{\boldsymbol{\beta}}]$.

Proof. We have:

$$\text{MSE}[\hat{\boldsymbol{\alpha}}(c)] = \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j + c)^2} + c^2\sum_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j + c)^2}$$

The derivative at $c = 0$ is negative, guaranteeing improvement for small $c > 0$. □

Chapter 2

Generalized Linear Models (GLM)

§ 2.1

The definition of Generalized Linear Models

2.1.1 exponential family

Definition 2.1. (Exponential Family) A probability distribution belongs to the exponential family if its probability density or mass function can be written in the canonical form:

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

where the components are:

- θ: natural parameter (canonical parameter)
- ϕ: dispersion parameter (scale parameter)
- $b(\theta)$: cumulant function (log-normalization constant)
- $a(\phi)$: dispersion function, typically $a(\phi) = \frac{\phi}{\omega}$ where ω are known weights
- $c(y,\phi)$: normalization constant ensuring $\int f(y|\theta,\phi)dy = 1$

Theorem 2.2. Let

$$Y \sim \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

- **Mean**: $\mathbb{E}[Y] = \mu = b'(\theta)$
- **Variance**: $\mathrm{Var}[Y] = b''(\theta)a(\phi) = V(\mu)a(\phi)$

where $V(\mu)$ is called the variance function.

Proof. We have

$$\int f(y|\theta,\phi)dy = 1 \implies \int \frac{\partial}{\partial\theta} f(y|\theta,\phi)dy = 0$$

Computing the derivative:

$$\begin{aligned}\frac{\partial}{\partial\theta} f(y|\theta,\phi) &= f(y|\theta,\phi)\cdot\frac{\partial}{\partial\theta}\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \\ &= f(y|\theta,\phi)\cdot\frac{y - b'(\theta)}{a(\phi)}\end{aligned}$$

Therefore:

$$\int \frac{y - b'(\theta)}{a(\phi)} f(y|\theta,\phi)dy = 0$$

Since $a(\phi) \neq 0$, we have:

$$\int y f(y|\theta,\phi)dy - b'(\theta)\int f(y|\theta,\phi)dy = 0$$

Using $\int f(y|\theta,\phi)dy = 1$ and $\mathbb{E}[Y] = \int y f(y|\theta,\phi)dy$:

$$\mathbb{E}[Y] - b'(\theta) = 0 \quad \implies \quad \mathbb{E}[Y] = b'(\theta) = \mu$$

Now, differentiate the mean result with respect to θ:

$$\frac{\partial}{\partial\theta}\mathbb{E}[Y] = b''(\theta)$$

Using the definition of expectation and differentiating under the integral sign:

$$\frac{\partial}{\partial\theta}\mathbb{E}[Y] = \int y\cdot\frac{y - b'(\theta)}{a(\phi)} f(y|\theta,\phi)dy \quad \text{(from earlier calculation)}$$

So we have:

$$b''(\theta) = \frac{1}{a(\phi)}\int y(y - b'(\theta)) f(y|\theta,\phi)dy$$

Expanding and using linearity of expectation:

$$\begin{aligned}b''(\theta) &= \frac{1}{a(\phi)}\left[\int y^2 f(y|\theta,\phi)dy - b'(\theta)\int y f(y|\theta,\phi)dy\right] \\ &= \frac{1}{a(\phi)}\left[\mathbb{E}[Y^2] - \mu\cdot\mathbb{E}[Y]\right] \\ &= \frac{1}{a(\phi)}\left[\mathbb{E}[Y^2] - \mu^2\right] \\ &= \frac{1}{a(\phi)}\mathrm{Var}[Y]\end{aligned}$$

Therefore:

$$\mathrm{Var}[Y] = b''(\theta)a(\phi)$$

□

2.1.2 Link Functions

Definition 2.3. (Link Function) A link function $g(\cdot)$ is a monotonic differentiable function that satisfies:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \tag{2.2}$$

Its inverse is called the response function:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

Morever if the link function satisfies:

$$g(\mu) = \theta$$

meaning the link function maps the mean to the natural parameter, it is called the canonical link function.

2.1.3 Generalized Linear Models (GLM)

Definition 2.4. (Generalized Linear Model) A Generalized Linear Model consists of three components:

1. **Random Component**: Response variables $Y_1, Y_2, \ldots, Y_n$ are independent and follow an exponential family distribution:

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

2. **Systematic Component**: Linear predictor combining covariates:

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

3. **Link Function**: Monotonic differentiable function connecting mean to linear predictor:

$$g(\mu_i) = \eta_i \quad \text{with} \quad \mu_i = \mathbb{E}[Y_i]$$

The complete model specification is:

$$\begin{cases} Y_i \sim \text{ExpFamily}(\theta_i, \phi) \\ \mathbb{E}[Y_i] = \mu_i = b'(\theta_i) \\ g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \\ \text{Var}[Y_i] = b''(\theta_i)a(\phi) = V(\mu_i)a(\phi) \end{cases} \tag{2.3}$$

Table 2.1: Generalized Linear Models (GLM): Exponential Family Distributions Summary

Distribution	Density Function	θ	ϕ	$b(\theta)$	Mean & Variance	Canonical Link
Normal	$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	μ	σ^2	$\frac{\theta^2}{2}$	$\mathbb{E}[Y] = \mu$ $\text{Var}(Y) = \sigma^2$	μ
Poisson	$P(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}$	$\log \lambda$	1	e^θ	$\mathbb{E}[Y] = \lambda$ $\text{Var}(Y) = \lambda$	$\log \mu$
Binomial	$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$	$\log\left(\frac{p}{1-p}\right)$	1	$n \log(1 + \exp(\theta))$	$\mathbb{E}[Y] = np$ $\text{Var}(Y) = np(1-p)$	$\log\left(\frac{\mu}{1-\mu}\right)$
Gamma	$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$	$-\beta$	$\frac{1}{\alpha}$	$-\log(-\theta)$	$\mathbb{E}[Y] = \frac{\alpha}{\beta}$ $\text{Var}(Y) = \frac{\alpha}{\beta^2}$	μ^{-1}
Inverse Gaussian	$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right)$	$-\frac{1}{2\mu^2}$	$\frac{1}{\lambda}$	$-\sqrt{-2\theta}$	$\mathbb{E}[Y] = \mu$ $\text{Var}(Y) = \frac{\mu^3}{\lambda}$	μ^{-2}
Negative Binomial	$P(Y = y) = \binom{y+r-1}{y} p^r (1-p)^y$	$\log\left(\frac{1-p}{p}\right)$	1	$-r \ln(1 - e^\theta)$	$\mathbb{E}[Y] = \frac{r(1-p)}{p}$ $\text{Var}(Y) = \frac{r(1-p)}{p^2}$	$\log\left(\frac{\mu}{\mu+r}\right)$
Geometric	$P(Y = y) = p(1-p)^y$	$\log(1-p)$	1	$-\log(1 - e^\theta)$	$\mathbb{E}[Y] = \frac{1-p}{p}$ $\text{Var}(Y) = \frac{1-p}{p^2}$	$\log(1+\mu)$
Exponential	$f(y) = \lambda \exp(-\lambda y)$	$-\lambda$	1	$-\log f(-\theta)$	$\mathbb{E}[Y] = \frac{1}{\lambda}$ $\text{Var}(Y) = \frac{1}{\lambda^2}$	μ^{-1}

Note: Exponential family form: $f(y|\theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$

Additional functions: $a(\phi) = \phi$ for Normal, Gamma, Inverse Gaussian; $a(\phi) = 1$ for others.

§ 2.2

GLM Parameter Estimation Methods

2.2.1 Likelihood Equations

For independent observations from an exponential family distribution, the log-likelihood function is:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right)$$

Theorem 2.5. (Score Equations) The maximum likelihood estimates satisfy the system of equations:

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i) X_{ir}}{a(\phi) V(\mu_i) g'(\mu_i)} = 0, \quad r = 1, 2, \cdots, p$$

Proof. Using the chain rule:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_r} &= \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_r} \\ &= \sum_{i=1}^{n} \frac{Y_i - \mu_i}{a(\phi)} \cdot \frac{1}{V(\mu_i)} \cdot \frac{1}{g'(\mu_i)} \cdot X_{ir} \end{aligned}$$

Setting derivatives to zero gives the score equations. □

2.2.2 Iteratively Reweighted Least Squares (IRLS)

Definition 2.6. (Working Variables) Define the **working response** and **weights**:

$$\begin{aligned} Z_i &= \eta_i + (Y_i - \mu_i) g'(\mu_i) \\ w_i &= \frac{1}{a(\phi) V(\mu_i) (g'(\mu_i))^2} \end{aligned}$$

Theorem 2.7. (IRLS) The score equations are equivalent to the weighted least squares normal equations:

$$\mathbf{X}^T \mathbf{W} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

which yields the updating formula:

$$\boldsymbol{\beta}^{(new)} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Z}$$

Algorithm 1 Iteratively Reweighted Least Squares (IRLS)

1: Initialize $\boldsymbol{\beta}^{(0)}$
2: $k \leftarrow 0$
3: **repeat**
4: Compute $\boldsymbol{\eta}^{(k)} = \mathbf{X}\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\mu}^{(k)} = g^{-1}(\boldsymbol{\eta}^{(k)})$
5: Compute working response: $Z_i^{(k)} = \eta_i^{(k)} + (Y_i - \mu_i^{(k)})g'(\mu_i^{(k)})$
6: Compute weights: $w_i^{(k)} = \dfrac{1}{a(\phi)V(\mu_i^{(k)})(g'(\mu_i^{(k)}))^2}$
7: Update: $\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{Z}^{(k)}$
8: $k \leftarrow k + 1$
9: **until** convergence

2.2.3 Fisher Scoring Method

Definition 2.8. (Newton-Raphson Method) The Newton-Raphson algorithm uses the observed information matrix:

$$\boldsymbol{\beta}^{(new)} = \boldsymbol{\beta}^{(old)} - \left[\nabla^2 \ell(\boldsymbol{\beta}^{(old)})\right]^{-1} \nabla \ell(\boldsymbol{\beta}^{(old)})$$

where:

- $\nabla \ell(\boldsymbol{\beta})$ is the score function
- $\nabla^2 \ell(\boldsymbol{\beta})$ is the observed information matrix

Definition 2.9. (Fisher Scoring) Fisher scoring replaces the observed information with the expected information (Fisher information):

$$\boldsymbol{\beta}^{(new)} = \boldsymbol{\beta}^{(old)} + \left[\mathcal{I}(\boldsymbol{\beta}^{(old)})\right]^{-1} \nabla \ell(\boldsymbol{\beta}^{(old)})$$

where $\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}[-\nabla^2 \ell(\boldsymbol{\beta})]$ is the Fisher information matrix.

Theorem 2.10. (Fisher Information for GLM) For GLMs, the Fisher information matrix has the form:

$$\mathcal{I}(\boldsymbol{\beta}) = \frac{1}{\phi}\mathbf{X}^T\mathbf{W}\mathbf{X}$$

where $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$ with $w_i = \dfrac{1}{a(\phi)V(\mu_i)(g'(\mu_i))^2}$.

Algorithm 2 Fisher Scoring for GLM

Initialize $\boldsymbol{\beta}^{(0)}$
$k \leftarrow 0$
repeat
  Compute $\boldsymbol{\eta}^{(k)} = \mathbf{X}\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\mu}^{(k)} = g^{-1}(\boldsymbol{\eta}^{(k)})$
  Compute weights: $w_i^{(k)} = \dfrac{1}{a(\phi)V(\mu_i^{(k)})(g'(\mu_i^{(k)}))^2}$
  Compute score: $\mathbf{S}^{(k)} = \mathbf{X}^T\mathbf{W}^{(k)}(\mathbf{Y} - \boldsymbol{\mu}^{(k)})$
  Compute Fisher information: $\mathcal{I}^{(k)} = \mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X}$
  Update: $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (\mathcal{I}^{(k)})^{-1}\mathbf{S}^{(k)}$
  $k \leftarrow k + 1$
until convergence

2.2.4 Algorithm Equivalence

Theorem 2.11. For GLMs, the Fisher scoring algorithm is equivalent to the IRLS algorithm.

Proof. The Fisher scoring update is:

$$\boldsymbol{\beta}^{(new)} = \boldsymbol{\beta}^{(old)} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{Y} - \boldsymbol{\mu})$$

Substituting the working response $\mathbf{Z} = \boldsymbol{\eta} + \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\mu})$ where $\mathbf{D} = \text{diag}(g'(\mu_i))$, we get:

$$\begin{aligned}\boldsymbol{\beta}^{(new)} &= \boldsymbol{\beta}^{(old)} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{D}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}^{(old)}) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Z}\end{aligned}$$

which is exactly the IRLS update. □

Remark 2.12. For canonical links, $g(\mu) = \theta$, so $g'(\mu) = 1/V(\mu)$ and the observed information equals the expected information:

$$-\nabla^2\ell(\boldsymbol{\beta}) = \mathcal{I}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$$

Therefore, Newton-Raphson and Fisher scoring coincide for canonical links.

§ 2.3 Hypothesis Testing in Generalized Linear Models

2.3.1 The Three Classical Tests

Consider testing the hypothesis:

$$H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20} \quad \text{vs} \quad H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_{20}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ is partitioned into parameters of interest ($\boldsymbol{\beta}_2$, dimension q) and nuisance parameters ($\boldsymbol{\beta}_1$, dimension $p - q$).

2.3.2 Likelihood Ratio Test (LRT)

Definition 2.13. (Likelihood Ratio Test) The likelihood ratio test statistic is defined as:

$$\lambda_{LR} = 2[\ell(\hat{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})]$$

where:

- $\ell(\hat{\boldsymbol{\beta}})$: log-likelihood at the unrestricted MLE
- $\ell(\tilde{\boldsymbol{\beta}})$: log-likelihood at the restricted MLE (under H_0)

Theorem 2.14. Under H_0 and regularity conditions:

$$\lambda_{LR} \xrightarrow{d} \chi^2(q) \quad \text{as } n \to \infty$$

where q is the number of restrictions imposed by H_0.

Proof. Using Taylor expansion around $\hat{\boldsymbol{\beta}}$:

$$\ell(\tilde{\boldsymbol{\beta}}) \approx \ell(\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \mathcal{I}(\hat{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})$$

Thus:

$$\lambda_{LR} \approx (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^T \mathcal{I}(\hat{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \sim \chi^2(q)$$

□

2.3.3 Wald Test

Definition 2.15. (Wald Test) The Wald test statistic for $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$ is:

$$W = (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})^T [\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_2)]^{-1} (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})$$

where $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_2)$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_2$.

Remark 2.16. For GLMs, the Wald test statistic can be computed as:

$$W = (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})^T [\mathbf{I}^{22}(\hat{\boldsymbol{\beta}})]^{-1} (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})$$

where $\mathbf{I}^{22}(\hat{\boldsymbol{\beta}})$ is the appropriate submatrix of the Fisher information inverse.

Theorem 2.17. Under H_0:

$$W \xrightarrow{d} \chi^2(q)$$

2.3.4 Score Test (Lagrange Multiplier Test)

Definition 2.18. (Score Test) The score test statistic is based on the score function evaluated at the restricted MLE:

$$S = \mathbf{U}(\tilde{\boldsymbol{\beta}})^T \mathcal{I}(\tilde{\boldsymbol{\beta}})^{-1} \mathbf{U}(\tilde{\boldsymbol{\beta}})$$

where:

- $\mathbf{U}(\tilde{\boldsymbol{\beta}}) = \nabla \ell(\tilde{\boldsymbol{\beta}})$: score function at restricted MLE
- $\mathcal{I}(\tilde{\boldsymbol{\beta}})$: Fisher information at restricted MLE

Remark 2.19. For GLMs with canonical links, the score test simplifies to:

$$S = (\mathbf{Y} - \tilde{\boldsymbol{\mu}})^T \mathbf{X}[\mathbf{X}^T \mathbf{W} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{Y} - \tilde{\boldsymbol{\mu}})$$

where $\tilde{\boldsymbol{\mu}}$ are the fitted values under H_0.

Theorem 2.20. Under H_0:

$$S \xrightarrow{d} \chi^2(q)$$

Theorem 2.21. All three tests are asymptotically equivalent:

$$\lambda_{LR} = W + o_p(1) = S + o_p(1)$$

For finite samples, the ordering of power is typically:

$$\text{LRT} \geq \text{Wald} \geq \text{Score}$$

where $\geq$ denotes "more powerful than".

§ 2.4

Logistic and Poisson Regression

2.4.1 Logistic Regression for Binary Data

Model Specification and GLM Verification

Definition 2.22. For binary response data $Y_i \in \{0, 1\}$, the logistic regression model assumes:

- **Random Component**: $Y_i \sim \text{Bernoulli}(p_i)$
- **Systematic Component**: $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- **Link Function**: Logit link $g(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$

Theorem 2.23. (Logistic Regression as GLM) The Bernoulli distribution belongs to the exponential family with:

$$\begin{aligned} \theta_i &= \log\left(\frac{p_i}{1 - p_i}\right) \\ b(\theta_i) &= \log(1 + e^{\theta_i}) = -\log(1 - p_i) \\ a(\phi) &= 1 \\ c(y_i, \phi) &= 0 \end{aligned}$$

The logit link is the canonical link function.

Proof. The Bernoulli PMF can be written as:

$$\begin{aligned} P(Y_i = y_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \exp\left\{ y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right\} \\ &= \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \end{aligned}$$

where $\theta_i = \log\left(\frac{p_i}{1-p_i}\right)$, $b(\theta_i) = \log(1+e^{\theta_i})$, $a(\phi) = 1$, and $c(y_i, \phi) = 0$. □

Parameter Estimation via IRLS

Theorem 2.24. (IRLS Components for Logistic Regression) For logistic regression:

- **Variance Function**: $V(\mu_i) = \mu_i(1-\mu_i)$ where $\mu_i = p_i$
- **Link Derivative**: $g'(\mu_i) = \dfrac{1}{\mu_i(1-\mu_i)}$
- **Weights**: $w_i = \mu_i(1-\mu_i)$
- **Working Response**: $Z_i = \eta_i + \dfrac{y_i - \mu_i}{\mu_i(1-\mu_i)}$

Algorithm 3 IRLS for Logistic Regression

1: Initialize $\boldsymbol{\beta}^{(0)}$, set $k = 0$
2: **repeat**
3: Compute linear predictor: $\eta_i^{(k)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(k)}$
4: Compute probabilities: $p_i^{(k)} = \dfrac{\exp(\eta_i^{(k)})}{1+\exp(\eta_i^{(k)})}$
5: Compute weights: $w_i^{(k)} = p_i^{(k)}(1-p_i^{(k)})$
6: Compute working response: $Z_i^{(k)} = \eta_i^{(k)} + \dfrac{y_i - p_i^{(k)}}{w_i^{(k)}}$
7: Update: $\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{Z}^{(k)}$
8: $k \leftarrow k+1$
9: **until** $\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\| < \epsilon$

Hypothesis Testing for Logistic Regression

Example 2.25. (Overall Model Significance) Test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ using deviance:

Null Model: $\log\left(\frac{p}{1-p}\right) = \beta_0$ **Full Model**: $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}$

Test statistic: $\lambda_{LR} = D_{\text{null}} - D_{\text{full}} \sim \chi^2(p)$

Example 2.26. (Individual Coefficients) For $H_0 : \beta_j = 0$:

Wald Test: $z = \dfrac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)} \sim N(0,1)$

Likelihood Ratio Test: Compare deviance of full model vs model without x_j

Example 2.27. (Goodness-of-Fit Test) **Hosmer-Lemeshow Test**:

- Group observations by predicted probabilities
- Test statistic: $C = \sum_{g=1}^{G} \dfrac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)} \sim \chi^2(G-2)$

2.4.2 Poisson Regression for Count Data

Model Specification and GLM Verification

Definition 2.28. For count data $Y_i \in \{0, 1, 2, \ldots\}$, the Poisson regression model assumes:

- **Random Component**: $Y_i \sim \text{Poisson}(\lambda_i)$
- **Systematic Component**: $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- **Link Function**: Log link $g(\lambda_i) = \log(\lambda_i)$

Theorem 2.29. (Poisson Regression as GLM) The Poisson distribution belongs to the exponential family with:

$$\begin{aligned} \theta_i &= \log(\lambda_i) \\ b(\theta_i) &= e^{\theta_i} = \lambda_i \\ a(\phi) &= 1 \\ c(y_i, \phi) &= -\log(y_i!) \end{aligned}$$

The log link is the canonical link function.

Proof. The Poisson PMF can be written as:

$$\begin{aligned} P(Y_i = y_i) &= \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \\ &= \exp\left\{ y_i \log(\lambda_i) - \lambda_i - \log(y_i!) \right\} \\ &= \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \end{aligned}$$

where $\theta_i = \log(\lambda_i)$, $b(\theta_i) = e^{\theta_i}$, $a(\phi) = 1$, and $c(y_i, \phi) = -\log(y_i!)$. □

Parameter Estimation via IRLS

Theorem 2.30. (IRLS Components for Poisson Regression) For Poisson regression:

- **Variance Function**: $V(\mu_i) = \mu_i$ where $\mu_i = \lambda_i$
- **Link Derivative**: $g'(\mu_i) = \dfrac{1}{\mu_i}$
- **Weights**: $w_i = \mu_i$
- **Working Response**: $Z_i = \eta_i + \dfrac{y_i - \mu_i}{\mu_i}$

Algorithm 4 IRLS for Poisson Regression

Initialize $\boldsymbol{\beta}^{(0)}$, set $k = 0$
repeat
  Compute linear predictor: $\eta_i^{(k)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(k)}$
  Compute rates: $\lambda_i^{(k)} = \exp(\eta_i^{(k)})$
  Compute weights: $w_i^{(k)} = \lambda_i^{(k)}$
  Compute working response: $Z_i^{(k)} = \eta_i^{(k)} + \dfrac{y_i - \lambda_i^{(k)}}{w_i^{(k)}}$
  Update: $\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{Z}^{(k)}$
  $k \leftarrow k + 1$
until $\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\| < \epsilon$

Hypothesis Testing for Poisson Regression

Example 2.31. (Overall Model Significance) Test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$:
Null Model: $\log(\lambda) = \beta_0$ **Full Model**: $\log(\lambda_i) = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}$
Test statistic: $\lambda_{LR} = 2[\ell(\text{full}) - \ell(\text{null})] \sim \chi^2(p)$

Example 2.32. Overdispersion] Test H_0: Poisson model vs H_1: Overdispersed model:
Score Test for Overdispersion:

$$S = \frac{\left[\sum_{i=1}^n (y_i - \hat{\lambda}_i)^2 - y_i\right]^2}{2\sum_{i=1}^n \hat{\lambda}_i^2} \sim \chi^2(1)$$

Alternative: Use negative binomial regression if overdispersion is present

Example 2.33. (Testing Rate Ratios) For $H_0 : \beta_j = 0$, the rate ratio is $e^{\beta_j} = 1$:
Wald Test: $z = \dfrac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)} \sim N(0,1)$
Confidence Interval for Rate Ratio: $\exp\left(\hat{\beta}_j \pm z_{1-\alpha/2} \cdot \text{SE}(\hat{\beta}_j)\right)$