Regression Analysis

Daniel Mao

Contents

Chapter 1

Simple Linear Regression

1.1 Simple Linear Regression

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
Or equivalently, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

- β_0, β_1, and σ are fixed, unknown variables.
- ε_i is unobserved random error term.
- y_i and x_i are the observed data.
- Treat x_i as fixed.

Regression Coefficients (β_0, β_1)
β_0 is an intercept.
$E[y_i|x_i = 0] = \beta_0 + \beta_1 0 = \beta_0$
β_1 is a slope.
$E[y_i|x_i = x^*] = \beta_0 + \beta_1 x^*$.
$E[y_i|x_i = x^* + 1] = \beta_0 + \beta_1(x^* + 1)$.
So $E[y_i|x_i = x^* + 1] - E[y_i|x_i = x^*] = \beta_1$.

Chapter 2

Multiple Linear Regression

2.1 The Model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Or,

$$y_i \mid \boldsymbol{x_i} \overset{indep}{\sim} N(\beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{x_i}, \sigma^2)$$

Or

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Or

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 I).$$

Or

$$\boldsymbol{y} \sim MVN(X\boldsymbol{\beta}, \sigma^2 I).$$

2.2 Estimating $\boldsymbol{\beta}$

2.2.1 Ordinary Least Squares Estimation

Definition (Ordinary Least Squares Estimation)**.** *We define the* ***ordinary least squares estimation*** $\hat{\boldsymbol{\beta}}_{OLS}$ *of* $\boldsymbol{\beta}$ *to be the vector given by*

$$\hat{\boldsymbol{\beta}}_{OLS} := \underset{\boldsymbol{\beta}\in\mathbb{R}^{1+p}}{\operatorname{argmin}}\big\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\big\}.$$

Proposition 2.2.1. *The OLS estimation* $\hat{\boldsymbol{\beta}}_{OLS}$ *of* $\boldsymbol{\beta}$ *is*

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

Proof. Let S denote the function $\boldsymbol{\beta} \mapsto (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$. Then

$$\begin{aligned}
S(\beta) &= (y - X\beta)^\top(y - X\beta) \\
&= y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta \\
&= y^\top y - 2y^\top X\beta + \beta^\top X^\top X\beta. \\
\frac{\partial S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= -2X^\top y + 2X^\top X\beta. \\
\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2} &= 2\boldsymbol{X}^\top\boldsymbol{X}.
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}_{OLS}}$ solves the equation $\frac{\partial S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 0$, we get

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

Since $\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2}$ is positive definite, $\hat{\boldsymbol{\beta}}_{OLS}$ is indeed a minimum point. ■

Proposition 2.2.2. *The mean and variance of* $\tilde{\boldsymbol{\beta}}_{OLS}$ *are:*

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\beta}}_{OLS}] &= \boldsymbol{\beta} \text{ and} \\
\operatorname{var}[\tilde{\boldsymbol{\beta}}_{OLS}] &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1},
\end{aligned}$$

assuming that $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ *exists.*

Proof.

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\beta}}_{OLS}] &= \mathbb{E}[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}] \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbb{E}[\boldsymbol{y}] \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta} \text{ and} \\
\operatorname{var}[\tilde{\boldsymbol{\beta}}_{OLS}] &= \operatorname{var}[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}] \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\operatorname{var}[\boldsymbol{y}]((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^\top \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\sigma^2 I((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^\top \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^\top \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.
\end{aligned}$$

That is,

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\beta}}_{OLS}] &= \boldsymbol{\beta} \text{ and} \\
\operatorname{var}[\tilde{\boldsymbol{\beta}}_{OLS}] &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.
\end{aligned}$$

■

Proposition 2.2.3. *The sampling distribution of $\tilde{\boldsymbol{\beta}}$ is*

$$\tilde{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}).$$

2.2.2 Maximum Likelihood Estimation

$$y \sim MVN(X\beta, \sigma^2 I)$$

The maximum likelihood function is:

$$\mathcal{L}(\beta, \sigma^2 \mid Y) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2 I|^{\frac{1}{2}}}\exp\big\{-\frac{1}{2}(y - X\beta)^\top(\sigma^2 I)^{-1}(y - X\beta)\big\}.$$

The log likelihood function is:

$$\ell(\beta, \sigma^2 \mid Y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^\top(y - X\beta).$$

So $\hat{\beta}_{MLE} = \hat{\beta}_{LS} = (X^\top X)^{-1}X^\top y$.

Proposition 2.2.4. *The mean and variance of the maximum likelihood estimator $\tilde{\boldsymbol{\beta}}^{MLE}$ are*

$$\begin{aligned}\mathbb{E}[\tilde{\boldsymbol{\beta}}] &= \boldsymbol{\beta} \text{ and} \\ \operatorname{var}[\tilde{\boldsymbol{\beta}}] &= \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1},\end{aligned}$$

Proof.

$$\begin{aligned}E[\hat{\boldsymbol{\beta}}] &= E[(X^\top X)^{-1} X^\top Y] \\ &= (X^\top X)^{-1} X^\top E[Y] \\ &= (X^\top X)^{-1} X^\top (X\boldsymbol{\beta}) \\ &= (X^\top X)^{-1} (X^\top X)\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \\ \operatorname{var}[\hat{\boldsymbol{\beta}}] &= \operatorname{var}[(X^\top X)^{-1} X^\top Y] \\ &= (X^\top X)^{-1} X^\top \operatorname{var}[Y]((X^\top X)^{-1} X^\top)^\top \\ &= (X^\top X)^{-1} X^\top \operatorname{var}[Y] X (X^\top X)^{-1} \\ &= (X^\top X)^{-1} X^\top (\sigma^2 I) X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} (X^\top X)(X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1}.\end{aligned}$$

■

Distribution

$$\hat{\boldsymbol{\beta}} \sim N\big(\boldsymbol{\beta}, \sigma^2 (X^\top X)^{-1}\big).$$

That is,

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})$$

where $V = (X^\top X)^{-1}$.

2.3 Fitted Values

$$\begin{aligned}&\hat{\boldsymbol{y}} \\ &= X\hat{\boldsymbol{\beta}} \\ &= X(X^\top X)^{-1} X^\top \boldsymbol{y} \\ &= H\boldsymbol{y}.\end{aligned}$$

2.4 Estimate the Standard Deviation

2.5 Estimate the Residuals

$$\begin{aligned}
&\boldsymbol{e}\\
&= \boldsymbol{y} - \hat{\boldsymbol{y}}\\
&= \boldsymbol{y} - X\hat{\boldsymbol{y}}\\
&= \boldsymbol{y} - X(X^\top X)^{-1}X^\top \boldsymbol{y}\\
&= \boldsymbol{y} - H\boldsymbol{y}\\
&= (I-H)\boldsymbol{y}.
\end{aligned}$$

Mean

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{e}]\\
&= \mathbb{E}[(I-H)\boldsymbol{y}]\\
&= (I-H)\mathbb{E}[\boldsymbol{y}]\\
&= (I-H)X\boldsymbol{\beta}\\
&= X\boldsymbol{\beta} - X(X^\top X)^{-1}X^\top X\boldsymbol{\beta}\\
&= X\boldsymbol{\beta} - X\boldsymbol{\beta}\\
&= \mathbf{0}.
\end{aligned}$$

Variance

$$\begin{aligned}
&\operatorname{var}[\boldsymbol{e}]\\
&= \operatorname{var}[(I-H)\boldsymbol{y}]\\
&= (I-H)\operatorname{var}[\boldsymbol{y}](I-H)^\top\\
&= (I-H)\sigma^2(I-H)^\top\\
&= \sigma^2(I-H).
\end{aligned}$$

Distribution

$$\boldsymbol{e} \sim N\big(\mathbf{0}, \sigma^2(I-H)\big).$$

2.6 Properties of the Hat Matrix

Proposition 2.6.1. *H is symmetric. i.e., $H^\top = H$.*

Proposition 2.6.2. *H is idempotent. i.e., $HH = H$.*

Proposition 2.6.3. *$I - H$ is symmetric. i.e., $(I-H)^\top = I - H$.*

Proposition 2.6.4. *$I - H$ is idempotent. i.e., $(I-H)(I-H) = I - H$.*

Proposition 2.6.5. *The trace $\mathrm{tr}(H)$ of the hat matrix is*

2.7 Weighted Least Squares Estimation

Definition (Weight Matrix)**.** *We define the **weight matrix** W to be the matrix given by*

$$W = \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{bmatrix}_{n\times n}$$

where w_i is the weight assigned to sample i, for $i \in \{1, ..., n\}$.

Definition (Weighted Least Squares Estimation)**.** *We define the **weighted least squares estimation** $\hat{\boldsymbol{\beta}}_{WLS}$ of $\boldsymbol{\beta}$ to be the vector given by*

$$\hat{\boldsymbol{\beta}}_{WLS} := \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{1+p}} \big\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top W (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \big\}.$$

Proposition 2.7.1. *The WLS estimation $\hat{\boldsymbol{\beta}}_{WLS}$ of $\boldsymbol{\beta}$ is*

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^\top W \boldsymbol{X})^{-1} \boldsymbol{X}^\top W \boldsymbol{y}.$$

Proof. Let $S(\boldsymbol{\beta})$ denote the function $\boldsymbol{\beta} \mapsto (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top W (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Then

$$\begin{aligned}
S(\boldsymbol{\beta}) &= (\boldsymbol{y}^\top - \boldsymbol{\beta}^\top \boldsymbol{X}^\top) W (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{y}^\top W - \boldsymbol{\beta}^\top \boldsymbol{X}^\top W)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= (\boldsymbol{y}^\top W \boldsymbol{y} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top W \boldsymbol{y}) - (\boldsymbol{y}^\top W \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top W \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{\beta}^\top \boldsymbol{X}^\top W \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}^\top W \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top W \boldsymbol{y} + \boldsymbol{y}^\top W \boldsymbol{y}. \\
\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \boldsymbol{\beta}^\top (\boldsymbol{X}^\top W \boldsymbol{X} + \boldsymbol{X}^\top W^\top \boldsymbol{X}) - \boldsymbol{y}^\top W \boldsymbol{X} - \boldsymbol{y}^\top W^\top \boldsymbol{X} + 0 \\
&= 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top W \boldsymbol{X} - 2\boldsymbol{y}^\top W \boldsymbol{X}. \\
\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} &= 2\boldsymbol{X}^\top W^\top \boldsymbol{X} \\
&= 2\boldsymbol{X}^\top W \boldsymbol{X}.
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}_{WLS}$ solves the equation $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, we get

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^\top W^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top W^\top \boldsymbol{y} = (\boldsymbol{X}^\top W \boldsymbol{X})^{-1} \boldsymbol{X}^\top W \boldsymbol{y}.$$

Since $\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}$ is a positive definite matrix, $\hat{\boldsymbol{\beta}}_{WLS}$ is indeed a minimum point. ■

Proposition 2.7.2. *The mean and variance of $\tilde{\boldsymbol{\beta}}_{WLS}$ are*

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}] =$$

Chapter 3

Robust Regression

3.1 Sensitivity Curve

Definition (Sensitivity Curve)**.**

3.2 Breakdown Point

Definition (Breakdown Point)**.**

Example 3.2.1 (Mean)**.** *The breakdown point of mean is* $\frac{1}{n}$ *where* n *is the sample size.*

Example 3.2.2 (Median)**.** *The breakdown point of median is* $\frac{1}{2}$*.*

Definition (Least Median Squares Estimator)**.** *We define the* ***least median squares estimator*** $\hat{\boldsymbol{\beta}}_{LMS}$ *of* $\boldsymbol{\beta}$ *to be the vector given by*

$$\hat{\boldsymbol{\beta}}_{LMS} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \big\{ \underset{i=1..n}{\operatorname{median}} (y_i - x_i^\top \boldsymbol{\beta})^2 \big\}.$$

Chapter 4

Smoothing Spline

4.1 Definitions

Definition (Smoothing Spline)**.** *We define a **smoothing spline**, denoted by $\hat{\mu}$, to be a second-differentiable function estimate of μ given by*

$$\hat{\mu} := \operatorname{argmin} \left\{ \sum_{i=1}^{n} (y_i - \mu(x_i))^2 + \lambda \int_{\mathbb{R}} (\mu''(x))^2 dx \right\}$$

for some non-negative smoothing parameter λ.

Proposition 4.1.1.

- *When $\lambda = 0$, we get perfect fit.*
- *When $\lambda = \infty$, we get OLS fit for simple linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.*
- *When λ is a Lagrange multiplier, we get a linear fit.*

Theorem 1. *Suppose f is a real function whose value is known only at a set of n distinct points $x_1, ..., x_n$. The points $(x_1, f(x_1)), ..., (x_n, f(x_n))$ can be used to determine the natural cubic splines s such that $s(x_i) = f(x_i)$ for $i = 1, ..., n$. Then*

$$\int_{\mathbb{R}} (s''(x))^2 dx \leq \int_{\mathbb{R}} (f''(x))^2 dx.$$

i.e., the natural cubic splines s is smoother than the actual function f.

Theorem 2. *For any fixed λ, the solution to the problem*

$$\operatorname{argmin} \left\{ \sum_{i=1}^{n} (y_i - \mu(x_i))^2 + \lambda \int_{\mathbb{R}} (\mu''(x))^2 dx \right\}$$

is a natural cubic spline with knots at $x_1, ..., x_n$.

Proposition 4.1.2.

$$\hat{\boldsymbol{\beta}}_\lambda = (N^\top N + \lambda\Omega)^{-1} N^\top \boldsymbol{y}.$$

$\hat{\boldsymbol{\beta}}_\lambda$ *is a linear function on the response* $\boldsymbol{y}$.

Proof. Say

$$\hat{\mu} = \sum_{j=1}^{n} \beta_j N_j.$$

Define

$$N := \begin{bmatrix} N_1(x_1) & \dots & N_n(x_1) \\ \vdots & \ddots & \vdots \\ N_1(x_n) & \dots & N_n(x_n) \end{bmatrix}.$$

Define

$$\Omega := \begin{bmatrix} \int_{\mathbb{R}} N_1''(x) N_1''(x) dx & \dots & \int_{\mathbb{R}} N_1''(x) N_n''(x) dx \\ \vdots & \ddots & \vdots \\ \int_{\mathbb{R}} N_n''(x) N_1''(x) dx & \dots & \int_{\mathbb{R}} N_n''(x) N_n''(x) dx \end{bmatrix}.$$

Define

$$S(\boldsymbol{\beta}) := \sum_{i=1}^{n} (y_i - \mu(x_i))^2 + \lambda \int_{\mathbb{R}} (\mu''(x))^2 dx.$$

Then

$$
\begin{aligned}
S(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - \mu(x_i))^2 + \lambda \int_{\mathbb{R}} (\mu''(x))^2 dx \\
&= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \beta_j N_j(x_i) \right)^2 + \lambda \int_{\mathbb{R}} \left(\sum_{j=1}^{n} \beta_j N_j''(x) \right)^2 dx \\
&= (\boldsymbol{y} - N\boldsymbol{\beta})^\top (\boldsymbol{y} - N\boldsymbol{\beta}) + \lambda \int_{\mathbb{R}} \left(\sum_{j=1}^{n} \beta_j N_j''(x) \right)^2 dx \\
&= (\boldsymbol{y} - N\boldsymbol{\beta})^\top (\boldsymbol{y} - N\boldsymbol{\beta}) + \lambda \int_{\mathbb{R}} \sum_{j=1}^{n} \sum_{k=1}^{n} \beta_j \beta_k N_j''(x) N_k''(x) dx \\
&= (\boldsymbol{y} - N\boldsymbol{\beta})^\top (\boldsymbol{y} - N\boldsymbol{\beta}) + \lambda \sum_{j=1}^{n} \sum_{k=1}^{n} \beta_j \beta_k \int_{\mathbb{R}} N_j''(x) N_k''(x) dx \\
&= (\boldsymbol{y} - N\boldsymbol{\beta})^\top (\boldsymbol{y} - N\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \Omega \boldsymbol{\beta} \\
&= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{\beta}^\top N \boldsymbol{y} - \boldsymbol{y}^\top N \boldsymbol{\beta} + \boldsymbol{\beta}^\top N^\top N \boldsymbol{\beta} + \lambda \boldsymbol{\beta}^\top \Omega \boldsymbol{\beta} \\
&= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top N \boldsymbol{\beta} + \boldsymbol{\beta}^\top (N^\top N + \lambda \Omega) \boldsymbol{\beta} \\
\frac{\partial S(\boldsymbol{\beta})}{\partial \beta} &= -2\boldsymbol{y}^\top N + 2\boldsymbol{\beta}^\top (N^\top N + \lambda \Omega)^\top \\
&\overset{set}{=} 0.
\end{aligned}
$$

Since the estimate $\hat{\boldsymbol{\beta}}_\lambda$ solves $\frac{\partial S(\boldsymbol{\beta})}{\partial \beta} = 0$, we get

$$
-2\boldsymbol{y}^\top N + 2\hat{\boldsymbol{\beta}}_\lambda^\top (N^\top N + \lambda \Omega)^\top = 0.
$$

Solving for $\hat{\boldsymbol{\beta}}_\lambda$ gives

$$
\hat{\boldsymbol{\beta}}_\lambda = (N^\top N + \lambda \Omega)^{-1} N^\top \boldsymbol{y}.
$$

■

4.2 Effective Degrees of Freedom

Definition. *We define the* ***effective degrees of freedom****, denoted by df_λ, to be a number given by*

$$
df_\lambda := (S_\lambda)
$$

where $S_\lambda := N(N^\top N + \lambda \Omega)^{-1} N^\top \boldsymbol{y}$.