Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Recap: Fitted Values and Residuals in Matrix Form

- Fitted values vector: $n \times 1$ column vector:

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

  where $\mathbf{H} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the **hat matrix**.

- Residuals vector: $n \times 1$ column vector:

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

- Note that, fitted values vector $\widehat{\mathbf{Y}}$ and residuals vector $\mathbf{e}$ are linear transformations of the observations vector $\mathbf{Y}$.

Recap: Hat Matrix

The hat matrix

$$\underset{n\times n}{\mathbf{H}} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and the matrix

$$\mathbf{I}_n - \mathbf{H} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

are $n \times n$ **projection matrices**. Meaning that they satisfy:

- **Symmetric**: $\mathbf{H}' = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})' = \mathbf{I}_n - \mathbf{H}$
- **Idempotent**: $\mathbf{H}^2 := \mathbf{H}\mathbf{H} = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n - \mathbf{H}$.

Moreover, $rank(\mathbf{H}) = 2$, $rank(\mathbf{I}_n - \mathbf{H}) = n - 2$ (provided that X_is are not all equal).

LS Estimation: Mean and Variance

Review: Linear Transformations of Random Vector

If $\mathbf{Z}$ is an $r \times 1$ random vector, and $\mathbf{A}$ is an $s \times r$ non-random matrix,

then

$$\underset{s\times 1}{\mathbf{W}} = \underset{s\times r}{\mathbf{A}}\ \underset{r\times 1}{\mathbf{Z}}$$

is an $s \times 1$ random vector with

$$\begin{aligned} \mathbf{E}\{\mathbf{W}\} &= \mathbf{E}\{\mathbf{AZ}\} = \mathbf{AE}\{\mathbf{Z}\} \\ \boldsymbol{\sigma}^2\{\mathbf{W}\} &= \boldsymbol{\sigma}^2\{\mathbf{AZ}\} = \mathbf{A}\boldsymbol{\sigma}^2\{\mathbf{Z}\}\mathbf{A}' \end{aligned}$$

If further $\mathbf{B}$ is a $t \times r$ non-random matrix, then

$$Cov(\mathbf{AZ}, \mathbf{BZ}) = \mathbf{A}\sigma^2\{\mathbf{Z}\}\mathbf{B}'$$

LS Estimation: Expectations

- LS estimator is unbiased:

$$\mathbf{E}\{\hat{\beta}\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}\{\mathbf{Y}\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta$$

- Expectation of the fitted values:

$$\mathbf{E}\{\widehat{\mathbf{Y}}\} = \mathbf{E}\{\mathbf{X}\hat{\beta}\} = \mathbf{X}\mathbf{E}\{\hat{\beta}\} = \mathbf{X}\beta = \mathbf{E}\{\mathbf{Y}\}$$

- Expectation of the residuals:

$$\mathbf{E}\{\mathbf{e}\} = \mathbf{E}\{\mathbf{Y} - \widehat{\mathbf{Y}}\} = \mathbf{E}\{\mathbf{Y}\} - \mathbf{E}\{\widehat{\mathbf{Y}}\} = \mathbf{0}_n$$

LS Estimation: Variance-Covariance Matrices

Variance-covariance of the LS estimator:

$$\begin{aligned}
\boldsymbol{\sigma}^2\{\hat{\boldsymbol{\beta}}\} &= \boldsymbol{\sigma}^2\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\} = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\boldsymbol{\sigma}^2\{\mathbf{Y}\}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} & -\frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ -\frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} & \frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2} \end{bmatrix}
\end{aligned}$$

What is the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$? What happens if $\overline{X} = 0$?

- Variance-covariance of the fitted values:

$$\boldsymbol{\sigma}^2\{\widehat{\mathbf{Y}}\} = \mathbf{H}\boldsymbol{\sigma}^2\{\mathbf{Y}\}\mathbf{H}' = \sigma^2\mathbf{H}$$

- Variance-covariance of the residuals:

$$\boldsymbol{\sigma}^2\{\mathbf{e}\} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\sigma}^2\{\mathbf{Y}\}(\mathbf{I}_n - \mathbf{H})' = \sigma^2(\mathbf{I}_n - \mathbf{H})$$

Are residuals uncorrelated? Do they have the same variance? How about the fitted values? What are the covariances between the residuals and fitted values?

Simple Regression: Geometric Interpretation

Some Notations

- Let $\mathbf{1}_n$ denote the n-dimensional column vector of ones;
- Let $\mathbf{x} = (X_1, \cdots, X_n)^T$ denote the n-dimensional column vector of X values;
- The design matrix $\mathbf{X}$ for simple regression is formed by these two column vectors:

$$\mathbf{X} = (\mathbf{1}_n, \mathbf{x})$$

Column Space of the Design Matrix

- Let $\mathrm{col}(\mathrm{X})$ denote the set of linear combinations of the two column vectors of **X**:

  $$\mathrm{col}(\mathrm{X}) = \{\mathbf{v} \in \mathbb{R}^{\mathrm{n}} : \text{ there exists } \mathrm{c}_0, \mathrm{c}_1 \in \mathrm{R}, \mathrm{s.t.}, \mathbf{v} = \mathrm{c}_0 \mathbf{1}_{\mathrm{n}} + \mathrm{c}_1 \mathbf{x}\}.$$

- $\mathrm{col}(\mathrm{X})$ is referred to as the *column space of the design matrix* **X**.
- Note that $\mathrm{col}(\mathrm{X})$ forms a **linear subspace** of $\mathbb{R}^n$.

Projection to $\mathrm{col}(\mathrm{X})$ by Hat Matrix

The hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ "orthogonally projects" vectors in $\mathbb{R}^n$ to $\mathrm{col}(\mathrm{X})$ in the following sense:

For (any) $\mathbf{w} \in \mathbb{R}^n$:

- $\mathbf{Hw} \in \mathrm{col}(\mathrm{X})$, i.e., we can find two constants $c_0, c_1 \in \mathbf{R}$ such that $\mathbf{Hw} = c_0\mathbf{1}_n + c_1\mathbf{x}$.
- $\mathbf{w} - \mathbf{Hw} \perp \mathrm{col}(\mathrm{X})$, i.e., for any $\mathbf{v} \in \mathrm{col}(\mathrm{X})$, $(\mathbf{w} - \mathbf{Hw})'\mathbf{v} = 0$.

Notes: "$\in$" is read as "belongs to"; "$\perp$" is read as "is orthogonal to".

What is $\mathbf{HX}$? What is $\mathbf{H1}_n, \mathbf{Hx}$? What is $\mathbf{Hv}$ for $\mathbf{v} \in \mathrm{col}(\mathrm{X})$?

Optional Reading: Proof of the Projection Properties of **H**

- By definition of $\mathbf{H}$, $\mathbf{Hw} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}$. Now, let $\mathbf{c} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}$ (which is a 2×1 vector). We can see $\mathbf{Hw} = \mathbf{Xc} = c_0\mathbf{1}_n + c_1\mathbf{x}$, where c_0, c_1 are the 1st and 2nd elements of the vector $\mathbf{c}$.
- By definition of $\mathrm{col}(\mathrm{X})$, for (any) $\mathbf{v} \in \mathrm{col}(\mathrm{X})$, we can find a 2×1 vector $\mathbf{c}$ such that $\mathbf{v} = \mathbf{Xc}$. So $(\mathbf{w} - \mathbf{Hw})'\mathbf{v} = (\mathbf{w} - \mathbf{Hw})'\mathbf{Xc}$=$\mathbf{w}'(\mathbf{I}_n - \mathbf{H})\mathbf{Xc}$. Note that, $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{HX} = \mathbf{X} - \mathbf{X} = \mathbf{0}_{n\times 2}$. So $(\mathbf{w} - \mathbf{Hw})^T\mathbf{v} = 0$.

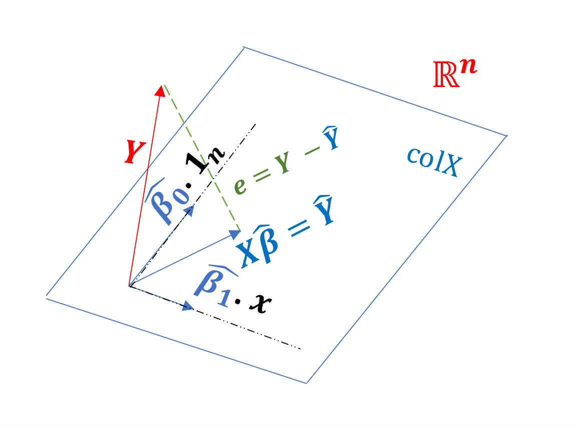
Fitted Values and Residuals

- The fitted values vector $\widehat{\mathbf{Y}} = \mathbf{HY} = \hat{\beta}_0 \mathbf{1}_n + \hat{\beta}_1 \mathbf{x}$ belongs to $\mathrm{col}(\mathrm{X})$.
- The residuals vector $\mathbf{e} = \mathbf{Y} - \mathbf{HY}$ is orthogonal to $\mathrm{col}(\mathrm{X})$.
- Also note that, $\mathbf{1}_n, \mathbf{x} \in \mathrm{col}(\mathrm{X})$.
- Therefore:

$$
\begin{aligned}
< \mathbf{e}, \mathbf{1}_n > &= \sum_{i=1}^{n} e_i = 0 \\
< \mathbf{e}, \mathbf{x} > &= \sum_{i=1}^{n} X_i e_i = 0 \\
< \mathbf{e}, \widehat{\mathbf{Y}} > &= \sum_{i=1}^{n} \hat{Y}_i e_i = 0
\end{aligned}
$$

Geometric Interpretation of Regression

Figure: To regress Y onto X is to "orthogonally project" the response vector $\mathbf{Y}$ onto the column space of the design matrix $\mathbf{X}$



Sums of Squares: Matrix Form

Error Sum of Squares

$$SSE = \sum_{i=1}^{n} e_i^2$$

can be expressed in matrix form as:

$$SSE = \mathbf{e}'\mathbf{e} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})'(\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

It can be shown that:

- $\mathbf{I}_n - \mathbf{H}$ is a projection matrix.
- $rank(\mathbf{I}_n - \mathbf{H}) = n - 2 = df(SSE)$.

Total Sum of Squares

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2$$

can be expressed in matrix form as:

$$SSTO = \mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}_n\mathbf{Y} = \mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y},$$

where $\mathbf{J}_n$ is the $n \times n$ matrix of ones. It can be shown that:

- $\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$ is a projection matrix.
- $rank(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n) = n - 1 = df(SSTO)$.

Regression Sum of Squares

$$SSR = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$$

can be expressed in matrix form as:

$$\begin{aligned} SSR &= \left(\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\right)' \left(\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\right), \qquad \overline{\mathbf{Y}} := \frac{1}{n}\mathbf{J}_n\mathbf{Y} \\ &= \mathbf{Y}' \left(\mathbf{H} - \frac{1}{n}\mathbf{J}_n\right)' \left(\mathbf{H} - \frac{1}{n}\mathbf{J}_n\right) \mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y} \end{aligned}$$

It can be shown that:

- $\mathbf{H} - \frac{1}{n}\mathbf{J}_n$ is a projection matrix.
- $rank(\mathbf{H} - \frac{1}{n}\mathbf{J}_n) = 1 = df(SSR)$.

Review: Matrix Trace Operation and Properties

- $\mathbf{M} = (m_{ij})$ is an $s \times s$ square matrix, its trace is the summation of its diagonal elements: $\mathrm{Tr}(\mathbf{M}) = \sum_{\mathrm{i=1}}^{\mathrm{s}} \mathrm{m_{ii}}$. Specifically, a scalar c may be viewed as a 1×1 square matrix, and $\mathrm{Tr(c) = c}$.

- Trace is a linear operator: For two square matrices **A** and **B** (of the same dimension) and a scalar c: $\mathrm{Tr}(\mathbf{A} + \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B})$, $\mathrm{Tr}(\mathrm{c} \cdot \mathbf{A}) = \mathrm{c} \cdot \mathrm{Tr}(\mathbf{A})$.

- Consequently, for a random (square) matrix **A**: $E(\mathrm{Tr}(\mathbf{A})) = \mathrm{Tr}(\mathrm{E}(\mathbf{A}))$.

- If **A** is an $s \times t$ matrix and **B** is a $t \times s$ matrix, then $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$.

Deriving Expectation of SSE

Note that: $SSE = \mathrm{Tr}(\mathrm{SSE}) = \mathrm{Tr}(\mathbf{Y}'(\mathbf{I}_\mathrm{n} - \mathbf{H})\mathbf{Y})$, so

$$\begin{aligned}
E(SSE) &= E\left(\mathrm{Tr}(\mathbf{Y}'(\mathbf{I}_\mathrm{n} - \mathbf{H})\mathbf{Y})\right) = E\left(Tr((\mathbf{I}_n - \mathbf{H})\mathbf{Y}\mathbf{Y}')\right) \\
&= Tr\left(E((\mathbf{I}_n - \mathbf{H})\mathbf{Y}\mathbf{Y}')\right) = Tr\left((\mathbf{I}_n - \mathbf{H})E(\mathbf{Y}\mathbf{Y}')\right) \\
&= Tr\Big((\mathbf{I}_n - \mathbf{H})(\sigma^2\mathbf{I}_n + \mathbf{X}\beta\beta'\mathbf{X}')\Big) \\
&= \sigma^2 Tr\left(\mathbf{I}_n - \mathbf{H}\right) + Tr\left((\mathbf{I}_n - \mathbf{H})\mathbf{X}\beta\beta'\mathbf{X}'\right) \\
&= (n-2)\sigma^2
\end{aligned}$$

The last equality is because $Tr\left(\mathbf{I}_n - \mathbf{H}\right) = Tr\left(\mathbf{I}_n\right) - Tr\left(\mathbf{H}\right) = n - 2$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{H}\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$. *Why?*

Can you derive $E(SSR)$ and $E(SSTO)$ in a similar fashion?

Optional Reading: Eigen-decomposition of Projection Matrices

- A projection matrix $\mathbf{P}$ can be decomposed as $\mathbf{P} = Q\Lambda Q^T$, where Q is an orthogonal matrix formed by $\mathbf{P}$'s eigenvectors and Λ is a diagonal matrix consisting of its eigenvalues.
- Moreover, $\mathbf{P}$'s eigenvalues are either 1 or 0.
- Consequently, the number of its nonzero eigenvalues equals its trace equals its rank.
- In simple linear regression:

$$rank(\mathbf{H}) = \mathbf{tr}(\mathbf{H}) = 2, \quad rank(\mathbf{I}_n - \mathbf{H}) = \mathbf{tr}(\mathbf{I}_n - \mathbf{H}) = n - 2$$

Optional Reading: Sampling Distribution of *SSE* under Normal Error Model

- $\mathbf{I}_n - \mathbf{H}$ is a projection matrix with rank $n-2 \Longrightarrow$ its spectral decomposition looks like:

$$\mathbf{I}_n - \mathbf{H} = \mathbf{Q}^T \mathbf{\Lambda} \mathbf{Q},$$

  where $\mathbf{\Lambda} = \mathrm{diag}\{1, \cdots, 1, 0, 0\}$ and $\mathbf{Q}$ is an orthogonal matrix.

- $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0} \Longrightarrow$

$$\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon}$$

- $SSE = \mathbf{e}^T\mathbf{e} = \boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon} = (\mathbf{Q}\boldsymbol{\epsilon})^T\boldsymbol{\Lambda}(\mathbf{Q}\boldsymbol{\epsilon})$
- Let $\mathbf{z} = \mathbf{Q}\boldsymbol{\epsilon}$, then

$$SSE = \sum_{i=1}^{n} \lambda_i z_i^2 = \sum_{i=1}^{n-2} z_i^2.$$

where λ_i is the ith diagonal element of $\boldsymbol{\Lambda}$ and we have $\lambda_i = 1$ for $i = 1, \cdots, n-2$ and $\lambda_i = 0$ for $i = n-1, n$.

- Moreover

$$\mathbf{E}(\mathbf{z}) = \mathbf{Q}\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}, \quad \boldsymbol{\sigma}^2\{\mathbf{z}\} = \mathbf{Q}\boldsymbol{\sigma}^2\{\boldsymbol{\epsilon}\}\mathbf{Q}^T = \sigma^2\mathbf{Q}\mathbf{Q}^T = \sigma^2\mathbf{I}_n$$

So under Normal error model, z_is are i.i.d. $N(0, \sigma^2)$.

- Thus, by the definition of chi-squares distributions,

$SSE \sim \sigma^2\chi^2_{(n-2)}$.

Confidence Interval for σ^2

Under the Normal error model, a $(1-\alpha)100\%$-confidence interval for σ^2 is :

$$\left[\frac{SSE}{\chi^2(1-\alpha/2; n-2)}, \frac{SSE}{\chi^2(\alpha/2; n-2)}\right]$$

- $\chi^2(1-\alpha/2; n-2)$ and $\chi^2(\alpha/2; n-2)$ are the $(1-\alpha/2)100\%$ percentile and $(\alpha/2)100\%$ percentile of the chi-squares distribution with $n-2$ degrees of freedom.

How to derive a $(1-\alpha)100\%$-confidence interval for the error standard deviation σ? What is a 95% confidence interval for the error variance in the "Heights example"?

Optional Reading: Deriving C.I. for σ^2

This is through the pivotal quantity:

$$\frac{SSE}{\sigma^2}$$

and the fact that it follows $\chi^2_{(n-2)}$ distribution since $SSE \sim \sigma^2\chi^2_{(n-2)}$.

Probability Density Curves of χ^2 Distributions