Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Model Diagnostics:

Overview

Assumptions of Normal Error Model

- **Linearity** of the regression relation
- **Normality** of the error terms
- **Constant variance** of the error terms
- **Independence** of the error terms

Consequences of Model Departures

- With regard to regression relation: serious
  - **Nonlinearity** of the regression relation
  - **Omission of important predictor variable(s)**
- With regard to error distribution: less serious
  - **Nonconstant variance (a.k.a. heteroscedasticity)** or **Nonindependence** $\Longrightarrow$ invalid variance estimation $\Longrightarrow$ invalid inference
  - **Nonnormality**: small departures – not serious; major departures – could be serious especially for small sample sizes
- **Outliers**: could be serious for small data sets

Residual Plots

- Examine regression relation and error variance:
  - residual vs. fitted value
  - residual vs. X variable(s)
  - residual vs. omitted X variable(s)
- Examine error distribution:
  - Normality: normal probability plot (Q-Q plot) of residuals
  - Independence: sequence plot of residuals
- Examine outliers or influential cases: studentized residuals, cook's distance

Remedial Measures

Mild departures often do not need to be fixed. For more serious departures:

- Fix regression relation: transformation of the response variable and/or transformation(s) of the X variable(s)
- Fix error distribution: transformation of the response variable
- Fix outliers: exclusion or robust regression

Model Diagnostics:

Nonlinearity Detection

Detection of Nonlinearity

residual vs. fitted value plot or residual vs. X variable plot:

- If these show a clear nonlinear pattern, then it is an indication of possible nonlinearity in the regression relation.
- This is because the nonlinearity unaccounted for by the model would be left in the residuals.

Simulation Experiment

- Data: 30 cases with $X \sim N(100, 16^2)$, $\varepsilon \sim N(0, 10^2)$,

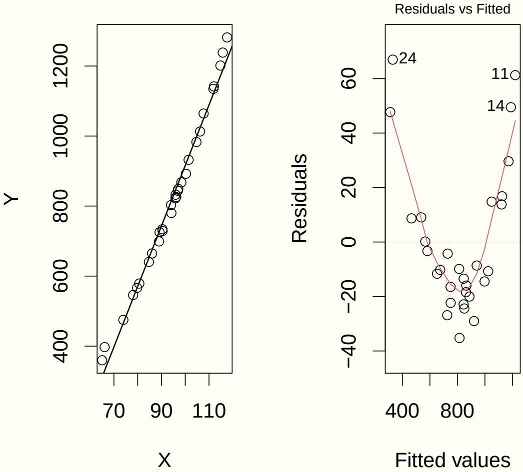
$$Y_i = 5 - X_i + 0.1X_i^2 + \varepsilon_i, \quad i = 1, \cdots, 30$$

- Fitted model: simple linear regression

Coefficients	Estimate	Std. Error	t value	$Pr(>\lvert t \rvert)$
Intercept	-811.8518	35.2767	-23.01	<2e-16 ***
X	17.2787	0.3695	46.76	<2e-16 ***

$\sqrt{MSE} = 27.6$, $R^2 = 0.9874$

Figure: Left: scatter plot; Right: residual vs. fitted value



Model Diagnostics: Unequal Variance Detection

Unequal Variance

- Sometimes variance increases (or decreases) with the value of the X variable. E.g., in financial data, the volume of transactions often has a role in the volatility of market.
- Data may come from different strata with different variability. E.g., measuring instruments with different precision may have been used to obtain the observations.

Detection of Nonconstancy in Variance

residual vs. fitted value plot:

- If it shows an unequal spread of the residuals along the horizontal axis, then this is an indication of unequal variance.

Simulation Experiment

- Data: 100 cases with $X_i = \frac{i}{10}$, $\varepsilon_i \sim N(0,1)$,

$$Y_i = 2 + 3X_i + \sigma(X_i)\varepsilon_i, \quad i = 1, \cdots, 100,$$

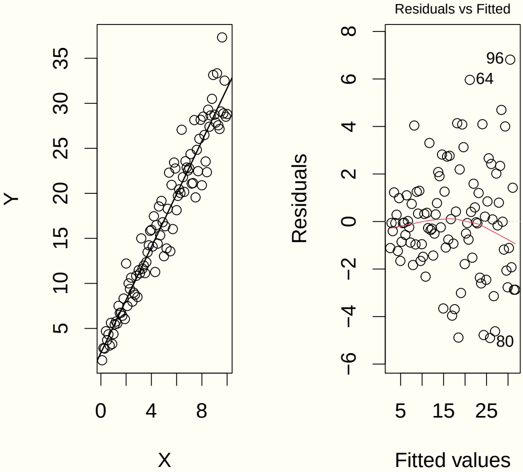
  where $\log \sigma^2(x) = 1 + 0.1x$.

- Fitted model: simple linear regression

Coefficients	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	2.29130	0.46689	4.908	3.67e-06 ***
X	2.93869	0.08027	36.612	$<$ 2e-16 ***

$\sqrt{MSE} = 2.317$, $R^2 = 0.9319$.

Figure: Left: scatter plot; Right: residual vs. fitted value



Model Diagnostics:

Non-normality Detection

Detection of Non-normality

Normal probability plot (a.k.a. Normal Q-Q plot) of residuals:

- If the residuals are normally distributed, then the points on the Q-Q plot should be (nearly) on a straight line.
- Departures from that could indicate **skewed** (non-symmetry) or **heavy-tailed** (more probability mass on tails than a Normal distribution) distributions.
- Other types of departures (e.g., nonlinearity) may affect the distribution of the residuals, thus it is better to examine these before checking normality.

Q-Q Plot

Q-Q stands for quantile-quantile. Q-Q plot is a graphical tool to compare the empirical distribution (of a sample) with a reference distribution.

- $e_{(k)}$'s – the *sample quantiles or empirical quantiles*: the kth smallest data in the sample
- $z_{(k)}$'s – the *theoretical quantiles* under the reference distribution
- Q-Q plot is simply the scatter plot of $e_{(k)}$'s vs. $z_{(k)}$'s
- A (nearly) straight line pattern indicates that the sample is likely from the reference distribution.

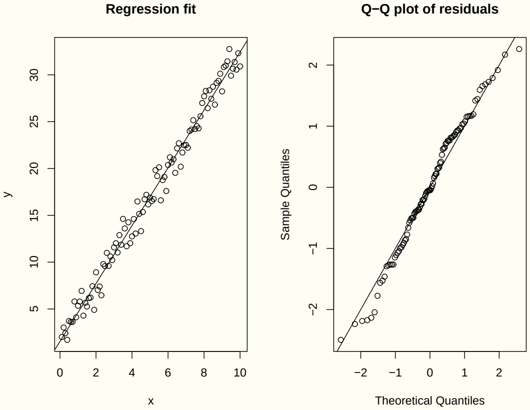
Case i	X_i	Y_i	$\widehat{Y}_i$	e_i
1	0.22	1.79	2.33	-0.54
2	3.55	5.66	5.90	-0.23
3	1.86	3.34	4.09	-0.75
4	3.29	5.83	5.62	0.22
5	1.25	4.74	3.43	1.31

$e_{(2)}$, the second smallest residual, is -0.54 and its corresponding theoretical quantile under Normality is:

$$z_{(2)} = \sqrt{MSE} \times Z\left((2 - 0.375)/(5 + 0.25)\right)$$

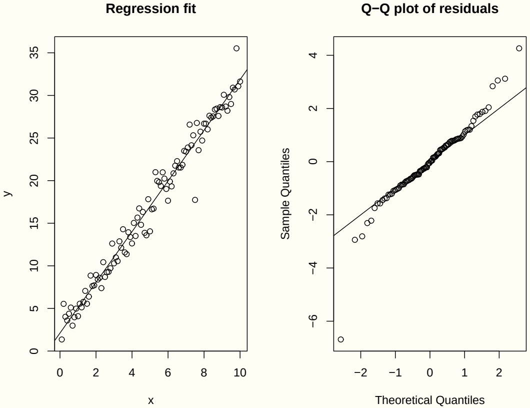
$$= \sqrt{0.8905} \times Z(0.31) = 0.944 \times (-0.497) = -0.469.$$

Error distribution: Normal$(0, 1)$



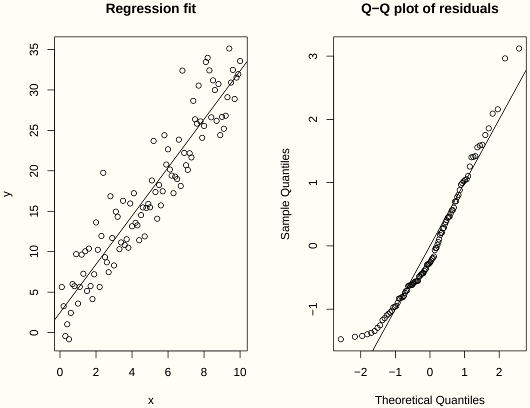
Normal Q-Q plot shows a straight line pattern.

Error distribution: $t_{(5)}$ – symmetrical but heavy-tailed



Normal Q-Q plot shows more probability mass on both tails compared to a Normal distribution.

Error distribution: centered $\chi^2_{(5)}$ – right-skewed



Normal Q-Q plot shows more probability mass on the right tail and less probability mass on the left tail compared to a Normal distribution.

Remedial Measures:

Transformations

Transformation of X

Linearize a nonlinear relationship:

- Increasing and concave downward: $X' = \log X$ or $X' = \sqrt{X}$
- Increasing and concave upward: $X' = X^2$ or $X' = \exp(X)$
- Decreasing and concave upward: $X' = 1/X$ or $X' = \exp(-X)$.
- Sometimes, add a constant to the transformation, e.g.

  $X' = 1/(c + X)$, to avoid negative or nearly zero values.

Transformation of Y

Fix error distribution such as unequal variance or non-normality.

- Unequal variance and non-normality often appear together.
- Commonly used transformations:
  - $Y' = \sqrt{Y}$
  - $Y' = \log Y$
  - $Y' = 1/Y$
  - Sometimes, add a constant to the transformation, e.g., $Y' = \log(c + Y)$, to avoid negative or nearly zero values.
- A simultaneous transformation of X might be needed to maintain a linear relationship.

Box-Cox Procedure

Choose a power transformation:

- For each $\lambda \in R$, define the transformed observations as

$$Y_i^* = \begin{cases} K_1 \frac{Y_i^\lambda - 1}{\lambda}, & if, \quad \lambda \neq 0 \\ K_2 \log(Y_i), & if, \quad \lambda = 0 \end{cases}, \; K_2 = (\prod_{j=1}^{n} Y_j)^{1/n}, \; K_1 = 1/K_2^{\lambda-1}$$

- For each λ, fit a regression model on the transformed data Y^* and derive $SSE(\lambda)$ (or maximum loglikelihood).
- Find the λ that minimizes SSE (or maximizes maximum loglikelihood) and apply the corresponding power transformation ($\lambda = 0$: logarithm transformation).

Simple Regression: Matrix Form

Simple Linear Regression in Matrix Form

The regression equations:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \cdots n$$

can be expressed in a compact matrix form:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}} \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

- **Response vector Y** and **error vector** : $n \times 1$ column vectors

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- **Design matrix**: $n \times 2$ matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_i \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

- **Coefficient vector**: 2×1 column vector:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

The model assumptions:

$$E(\epsilon_i) = 0, \quad Var(\epsilon_i) = \sigma^2, \quad \text{for all } i = 1, \cdots, n$$

$$Cov(\epsilon_i, \epsilon_j) = 0, \quad \text{for all } i \neq j$$

can be expressed in matrix form:

$$\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}_n, \quad \boldsymbol{\sigma}^{\mathbf{2}}\{\boldsymbol{\epsilon}\} = \sigma^2 \mathbf{I}_n.$$

Mean of the error vector:

$$\mathbf{E}\{\boldsymbol{\epsilon}\} := \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}_n,$$

where $\mathbf{0}_n$ is the $n \times 1$ zero vector.

Variance-covariance matrix of the error vector:

$$
\boldsymbol{\sigma^2}\{\boldsymbol{\epsilon}\} := \begin{bmatrix} Var(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_2, \epsilon_1) & Var(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\epsilon_n, \epsilon_1) & Cov(\epsilon_n, \epsilon_2) & \cdots & Var(\epsilon_n) \end{bmatrix}
$$

$$
= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n,
$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Mean response vector: $n \times 1$ column vector:

$$\mathbf{E}\{\mathbf{Y}\} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_i) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_i \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_i \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}.$$

Summary

simple regression in matrix form:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}}\ \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

- $\boldsymbol{\epsilon}$ is a random vector with $\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}_n,\ \boldsymbol{\sigma}^{\mathbf{2}}\{\boldsymbol{\epsilon}\} = \sigma^2\mathbf{I}_n$.
- Normal error model: $\boldsymbol{\epsilon} \sim \mathrm{Normal}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$.
- In terms of the response vector:

$$\mathbf{E}\{\mathbf{Y}\} = \mathbf{X}\boldsymbol{\beta}, \quad \boldsymbol{\sigma}^{\mathbf{2}}\{\mathbf{Y}\} = \sigma^2\mathbf{I}_n.$$

Least Squares Estimation:

Matrix Form

Least Squares Estimation in Matrix Form

Least squares criterion:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

can be expressed in matrix form : $\mathbf{b} = (b_0, b_1)^T$

$$Q(\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\mathbf{b})' (\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}.$$

LS estimators:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \overline{Y} - \hat{\beta}_1 \overline{X} \\ \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{bmatrix},$$

provided that X_is are not all equal.

- $\hat{\boldsymbol{\beta}}$ is linear in the observations $\mathbf{Y}$.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}.$$

When

$$D := n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2 = n\sum_{i=1}^n (X_i - \overline{X})^2 \neq 0$$

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \begin{bmatrix} \frac{\sum_{i=1}^n X_i^2}{n\sum_{i=1}^n (X_i - \overline{X})^2} & -\frac{\sum_{i=1}^n X_i}{n\sum_{i=1}^n (X_i - \overline{X})^2} \\ -\frac{\sum_{i=1}^n X_i}{n\sum_{i=1}^n (X_i - \overline{X})^2} & \frac{n}{n\sum_{i=1}^n (X_i - \overline{X})^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} & -\frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ -\frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} & \frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2} \end{bmatrix}. \end{aligned}$$

Deriving LS Estimator

- Differentiate $Q(\cdot)$ with respect to $\mathbf{b}$: $\frac{\partial}{\partial \mathbf{b}} Q = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}$.
- Set the gradient to zero $\Longrightarrow$ *normal equation*:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

- Multiply both sides by $(\mathbf{X}'\mathbf{X})^{-1}$:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

- The left hand side becomes $\mathbf{I}_2\mathbf{b} = \mathbf{b}$, and the right hand side is the solution.

Fitted Value and Residual:

Matrix Form

Fitted Values and Residuals

- 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:

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

- Fitted values $\widehat{\mathbf{Y}}$ and residuals $\mathbf{e}$ are linear in the observations $\mathbf{Y}$.

Hat Matrix

$\mathbf{H}$ plays an important role in model diagnostics.

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

are $n \times n$ **projection matrices**:

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

Error Sum of Squares

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

can be expressed in matrix form:

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

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

LS Estimation: Mean and Variance

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

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

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:

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

- Variance-covariance of the fitted values:

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

- Variance-covariance of the residuals:

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

Expectation of SSE

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

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