CEE 616: Probabilistic Machine Learning

M2 Linear Methods: L2C Linear Regression

Jimi Oke

UMass Amherst

College of Engineering

Thu, Oct 2, 2025

Outline

1. Introduction

2. OLS

3. Considerations

4. Irregularities

5. WLS

6. Outlook

Linear regression

Linear regression

Model of the form:

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2)$$

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{w}^\top \boldsymbol{x})^2\right) \tag{1}$$

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{w}^\top \boldsymbol{x})^2\right) \tag{1}$$

where:

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top\boldsymbol{x},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\boldsymbol{w}^\top\boldsymbol{x})^2\right) \tag{1}$$

where:

- y: response or dependent variable
- $\boldsymbol{\theta} = \boldsymbol{w}$: model parameters

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top\boldsymbol{x},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\boldsymbol{w}^\top\boldsymbol{x})^2\right) \tag{1}$$

where:

- y: response or dependent variable
- $\boldsymbol{\theta} = \boldsymbol{w}$: model parameters
- $\boldsymbol{w} = [w_0, w_1, \ldots, w_D]$

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top\boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{w}^\top\boldsymbol{x})^2\right) \qquad (1)$$

where:

- y: response or dependent variable
- $\boldsymbol{\theta} = \boldsymbol{w}$: model parameters
- $\boldsymbol{w} = [w_0, w_1, \ldots, w_D]$
- w_0: intercept, offset, bias, unconditional mean or baseline $\mathbb{E}[y]$

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{w}^\top \boldsymbol{x})^2\right) \tag{1}$$

where:

- y: response or dependent variable
- $\boldsymbol{\theta} = \boldsymbol{w}$: model parameters
- $\boldsymbol{w} = [w_0, w_1, \dots, w_D]$
- w_0: intercept, offset, bias, unconditional mean or baseline $\mathbb{E}[y]$
- $w_{1:D}$: weights or regression coefficients

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{w}^\top \boldsymbol{x})^2\right) \tag{1}$$

where:

- y: response or dependent variable
- $\boldsymbol{\theta} = \boldsymbol{w}$: model parameters
- $\boldsymbol{w} = [w_0, w_1, \ldots, w_D]$
- w_0: intercept, offset, bias, unconditional mean or baseline $\mathbb{E}[y]$
- $w_{1:D}$: weights or regression coefficients
- $\boldsymbol{x} = [1, x_1, \ldots, x_D]$: predictors or explanatory/independent variables

Linear regression

Model of the form:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \boldsymbol{w}^\top \boldsymbol{x})^2\right) \tag{1}$$

where:

- y: response or dependent variable
- $\boldsymbol{\theta} = \boldsymbol{w}$: model parameters
- $\boldsymbol{w} = [w_0, w_1, \ldots, w_D]$
- w_0: intercept, offset, bias, unconditional mean or baseline $\mathbb{E}[y]$
- $w_{1:D}$: weights or regression coefficients
- $\boldsymbol{x} = [1, x_1, \ldots, x_D]$: predictors or explanatory/independent variables
- D: number of features

Complexity

Complexity

- Simple linear regression: $D = 1$:

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

- Multivariate linear regression: $\boldsymbol{y} \in \mathbb{R}^J, J > 1$

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

- Multivariate linear regression: $\boldsymbol{y} \in \mathbb{R}^J, J > 1$

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{4}$$

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

- Multivariate linear regression: $\boldsymbol{y} \in \mathbb{R}^J, J > 1$

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{4}$$

- LR with nonlinear transformation of inputs:

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

- Multivariate linear regression: $\boldsymbol{y} \in \mathbb{R}^J, J > 1$

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{4}$$

- LR with nonlinear transformation of inputs:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \phi(\boldsymbol{x}), \sigma^2) \tag{5}$$

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

- Multivariate linear regression: $\boldsymbol{y} \in \mathbb{R}^J, J > 1$

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{4}$$

- LR with nonlinear transformation of inputs:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \phi(\boldsymbol{x}), \sigma^2) \tag{5}$$

where $\phi(\boldsymbol{x})$ is feature extractor or kernel

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

- Multivariate linear regression: $\boldsymbol{y} \in \mathbb{R}^J, J > 1$

$$p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{4}$$

- LR with nonlinear transformation of inputs:

$$p(y|\boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \phi(\boldsymbol{x}), \sigma^2) \tag{5}$$

  where $\phi(\boldsymbol{x})$ is feature extractor or kernel
- Polynomial regression (1D):

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

- Multivariate linear regression: $\boldsymbol{y} \in \mathbb{R}^J, J > 1$

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{4}$$

- LR with nonlinear transformation of inputs:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \phi(\boldsymbol{x}), \sigma^2) \tag{5}$$

  where $\phi(\boldsymbol{x})$ is feature extractor or kernel
- Polynomial regression (1D): $\phi(x) = [1, x, x^2, \ldots, x^d]$

Complexity

- Simple linear regression: $D = 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2) \tag{2}$$

- Multiple linear regression: $D > 1$:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{3}$$

- Multivariate linear regression: $\boldsymbol{y} \in \mathbb{R}^J, J > 1$

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2) \tag{4}$$

- LR with nonlinear transformation of inputs:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \phi(\boldsymbol{x}), \sigma^2) \tag{5}$$

  where $\phi(\boldsymbol{x})$ is feature extractor or kernel
- Polynomial regression (1D): $\phi(x) = [1, x, x^2, \ldots, x^d]$ (order/degree d)

Alternative view of simple linear regression model

Alternative view of simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

Alternative view of simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$y = w_0 + w_1 x + \epsilon \tag{6}$$

Alternative view of simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$\underbrace{y = w_0 + w_1 x}_{\text{true regression line}} + \underbrace{\epsilon}_{\text{random error term}} \tag{6}$$

Alternative view of simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$\underbrace{y = w_0 + w_1 x}_{\text{true regression line}} + \underbrace{\epsilon}_{\text{random error term}} \tag{6}$$

where ϵ is a normally distributed random variable:

Alternative view of simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$\underbrace{y = w_0 + w_1 x}_{\text{true regression line}} + \underbrace{\epsilon}_{\text{random error term}} \tag{6}$$

where ϵ is a normally distributed random variable:

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

Alternative view of simple linear regression model

For any fixed value of the independent variable x, the dependent variable y is related to x via the **model equation**:

$$y = w_0 + w_1 x + \epsilon \tag{6}$$

true regression line

random error term

where ϵ is a normally distributed random variable:

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

w_0 (intercept) and w_1 (slope) are the **regression coefficients**

Alternate view 1D LR (cont.)

Alternate view 1D LR (cont.)

Further considerations:

Alternate view 1D LR (cont.)

Further considerations:

- The linear regression model: y is treated as a **random variable** whose mean depends on x

$$y = w_0 + w_1 x + \epsilon$$
$$[\text{Response}] = [\text{mean (depending on } x)] + [\text{error}]$$

Alternate view 1D LR (cont.)

Further considerations:

- The linear regression model: y is treated as a **random variable** whose mean depends on x

$$y = w_0 + w_1 x + \epsilon$$
$$[\text{Response}] = [\text{mean (depending on } x)] + [\text{error}]$$

- $\mathbb{E}(y|x) = w_0 + w_1 x$

Alternate view 1D LR (cont.)

Further considerations:

- The linear regression model: y is treated as a **random variable** whose mean depends on x

$$y = w_0 + w_1 x + \epsilon$$
$$[\text{Response}] = [\text{mean (depending on } x)] + [\text{error}]$$

- $\mathbb{E}(y|x) = w_0 + w_1 x$
- Given data $x_n,\ n = 1, ..., N$, x is **treated as fixed**.

Alternate view 1D LR (cont.)

Further considerations:

- The linear regression model: y is treated as a **random variable** whose mean depends on x

$$y = w_0 + w_1 x + \epsilon$$
$$[\text{Response}] = [\text{mean (depending on } x)] + [\text{error}]$$

- $\mathbb{E}(y|x) = w_0 + w_1 x$
- Given data x_n, $n = 1, ..., N$, x is **treated as fixed**.
- ϵ - **error term** (**random variable**), accounts for:

Alternate view 1D LR (cont.)

Further considerations:

- The linear regression model: y is treated as a **random variable** whose mean depends on x

$$y = w_0 + w_1 x + \epsilon$$
$$[\text{Response}] = [\text{mean (depending on } x)] + [\text{error}]$$

- $\mathbb{E}(y|x) = w_0 + w_1 x$
- Given data x_n, $n = 1, ..., N$, x is **treated as fixed**.
- ϵ - **error term** (**random variable**), accounts for:
  - measurement error of y

Alternate view 1D LR (cont.)

Further considerations:

- The linear regression model: y is treated as a **random variable** whose mean depends on x

$$y = w_0 + w_1 x + \epsilon$$
$$[\text{Response}] = [\text{mean (depending on } x)] + [\text{error}]$$

- $\mathbb{E}(y|x) = w_0 + w_1 x$
- Given data x_n, $n = 1, ..., N$, x is **treated as fixed**.
- ϵ - **error term** (**random variable**), accounts for:
  - measurement error of y
  - the effects of other variables not in the model

Model in matrix notation

- With n independent observations on Y and the associated values of x_n, the complete model:

Model in matrix notation

- With n independent observations on Y and the associated values of x_n, the complete model:

$$\begin{aligned} y_1 &= w_0 + w_1 x_1 + \epsilon_1 \\ y_2 &= w_0 + w_1 x_2 + \epsilon_2 \\ &\vdots \\ y_N &= w_0 + w_1 x_N + \epsilon_N \end{aligned}$$

Model in matrix notation

- With n independent observations on Y and the associated values of x_n, the complete model:

$$
\begin{aligned}
y_1 &= w_0 + w_1 x_1 + \epsilon_1 \\
y_2 &= w_0 + w_1 x_2 + \epsilon_2 \\
&\vdots \\
y_N &= w_0 + w_1 x_N + \epsilon_N
\end{aligned}
$$

- In matrix notation:

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
$$

$$
\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}
$$

- We typically assume an intercept (represented by column of 1's in $\boldsymbol{X}$), except where explicitly noted otherwise.

Residual sum of squares (RSS)

Residual sum of squares (RSS)

Residual

Residual sum of squares (RSS)

Residual

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n,

Residual sum of squares (RSS)

Residual

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

Residual sum of squares (RSS)

Residual

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

$$e_n = y_n - \hat{y}_n \tag{8}$$

Residual sum of squares (RSS)

Residual

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

$$e_n = y_n - \hat{y}_n \tag{8}$$

The RSS is an overall measure of the fitness of a regression model:

Residual sum of squares (RSS)

Residual

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

$$e_n = y_n - \hat{y}_n \tag{8}$$

The RSS is an overall measure of the fitness of a regression model:

$$RSS =$$

Residual sum of squares (RSS)

Residual

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

$$e_n = y_n - \hat{y}_n \tag{8}$$

The RSS is an overall measure of the fitness of a regression model:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{N} e_n^2$$

Residual sum of squares (RSS)

Residual

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

$$e_n = y_n - \hat{y}_n \tag{8}$$

The RSS is an overall measure of the fitness of a regression model:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{N} e_n^2 = \sum_{i=1}^{N} (y_n - \hat{y}_n)^2 \tag{9}$$

Residual sum of squares (RSS)

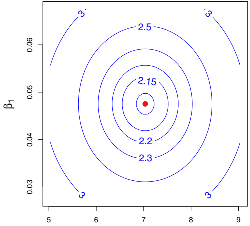
Residual

Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

$$e_n = y_n - \hat{y}_n \tag{8}$$

The RSS is an overall measure of the fitness of a regression model:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{N} e_n^2 = \sum_{i=1}^{N} (y_n - \hat{y}_n)^2 \tag{9}$$



Residual sum of squares (RSS)

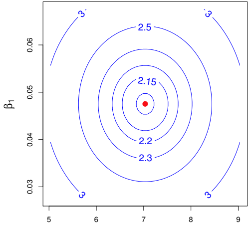
Residual

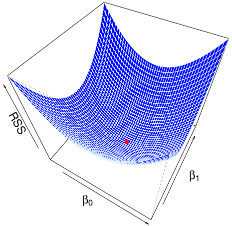
Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

$$e_n = y_n - \hat{y}_n \tag{8}$$

The RSS is an overall measure of the fitness of a regression model:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{N} e_n^2 = \sum_{i=1}^{N} (y_n - \hat{y}_n)^2 \tag{9}$$





Residual sum of squares (RSS)

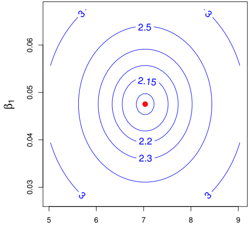
Residual

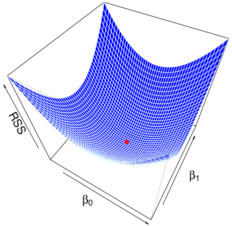
Given n independent observations $(x_1, y_1)...(x_n, y_n)$, with $\hat{y}_n$ as the predicted value for each y_n, the residual e_n is defined as:

$$e_n = y_n - \hat{y}_n \tag{8}$$

The RSS is an overall measure of the fitness of a regression model:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{N} e_n^2 = \sum_{i=1}^{N} (y_n - \hat{y}_n)^2 \tag{9}$$





RSS plotted against w_0 and w_1 for a given dataset using a contour plot (Left) and 3-D plot (right). The least squares estimate $(\hat{w}_0, \hat{w}_1)$ is indicated by the red dots.

Coefficient of determination

The **total sum of squares** is a measure of the total variance in the data.

Coefficient of determination

The **total sum of squares** is a measure of the total variance in the data.

$$TSS = \sum_{n=1}^{N} (y_n - \overline{y}_n)^2 = \sum_{n=1}^{N} y_n^2 - \frac{1}{N} \left(\sum_{n=1}^{N} y_n \right)^2 \tag{10}$$

The coefficient of determination R^2 is a measure of the **proportion of variance explained** by the regression model:

Coefficient of determination

The **total sum of squares** is a measure of the total variance in the data.

$$TSS = \sum_{n=1}^{N} (y_n - \overline{y}_n)^2 = \sum_{n=1}^{N} y_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} y_n\right)^2 \tag{10}$$

The coefficient of determination R^2 is a measure of the **proportion of variance explained** by the regression model:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{11}$$

Coefficient of determination

The **total sum of squares** is a measure of the total variance in the data.

$$TSS = \sum_{n=1}^{N} (y_n - \overline{y}_n)^2 = \sum_{n=1}^{N} y_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} y_n\right)^2 \tag{10}$$

The coefficient of determination R^2 is a measure of the **proportion of variance explained** by the regression model:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{11}$$

The **adjusted** R^2 penalizes for model complexity:

Coefficient of determination

The **total sum of squares** is a measure of the total variance in the data.

$$TSS = \sum_{n=1}^{N} (y_n - \overline{y}_n)^2 = \sum_{n=1}^{N} y_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} y_n\right)^2 \tag{10}$$

The coefficient of determination R^2 is a measure of the **proportion of variance explained** by the regression model:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{11}$$

The **adjusted** R^2 penalizes for model complexity:

$$R_a^2 = 1 - \frac{RSS/df_{RSS}}{TSS/df_{TSS}}$$

Coefficient of determination

The **total sum of squares** is a measure of the total variance in the data.

$$TSS = \sum_{n=1}^{N} (y_n - \overline{y}_n)^2 = \sum_{n=1}^{N} y_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} y_n\right)^2 \tag{10}$$

The coefficient of determination R^2 is a measure of the **proportion of variance explained** by the regression model:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{11}$$

The **adjusted** R^2 penalizes for model complexity:

$$R_a^2 = 1 - \frac{RSS/df_{RSS}}{TSS/df_{TSS}} = 1 - \frac{(1 - R^2)(N-1)}{N - D - 1} \tag{12}$$

Principle of least squares

Principle of least squares

Given the regression model:

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

$$RSS(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$$

Principle of least squares

Given the regression model:

$$\hat{y}_n = \mathbf{w}^\top \mathbf{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

$$RSS(\mathbf{w}) = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \left(y_n - \mathbf{w}^\top \mathbf{x}_n\right) \tag{14}$$

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

$$RSS(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right) \tag{14}$$

In matrix form, we write:

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

$$RSS(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right) \tag{14}$$

In matrix form, we write:

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{15}$$

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

$$RSS(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right) \tag{14}$$

In matrix form, we write:

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{15}$$

To minimize the RSS, we take the derivative w.r.t. to $\boldsymbol{w}$ and set to 0:

$$RSS = \boldsymbol{y}^\top \boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{y} - \boldsymbol{y}^\top (\boldsymbol{X}\boldsymbol{w}) + (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{X}\boldsymbol{w}$$

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

$$RSS(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right) \tag{14}$$

In matrix form, we write:

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{15}$$

To minimize the RSS, we take the derivative w.r.t. to $\boldsymbol{w}$ and set to 0:

$$\begin{aligned} RSS &= \boldsymbol{y}^\top \boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{y} - \boldsymbol{y}^\top (\boldsymbol{X}\boldsymbol{w}) + (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2(\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{y} + \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} \end{aligned}$$

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

$$RSS(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right) \tag{14}$$

In matrix form, we write:

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{15}$$

To minimize the RSS, we take the derivative w.r.t. to $\boldsymbol{w}$ and set to 0:

$$\begin{aligned} RSS &= \boldsymbol{y}^\top \boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{y} - \boldsymbol{y}^\top (\boldsymbol{X}\boldsymbol{w}) + (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2(\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{y} + \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} \\ RSS' &= -2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} \end{aligned}$$

Principle of least squares

Given the regression model:

$$\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n \tag{13}$$

then the residual sum of squares (loss function) is given by

$$RSS(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right) \tag{14}$$

In matrix form, we write:

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{15}$$

To minimize the RSS, we take the derivative w.r.t. to $\boldsymbol{w}$ and set to 0:

$$\begin{aligned}
RSS &= \boldsymbol{y}^\top \boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{y} - \boldsymbol{y}^\top (\boldsymbol{X}\boldsymbol{w}) + (\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{X}\boldsymbol{w} \\
&= \boldsymbol{y}^\top \boldsymbol{y} - 2(\boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{y} + \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} \\
RSS' &= -2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = -2\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})
\end{aligned}$$

Principle of least squares (cont.)

Principle of least squares (cont.)

If $\boldsymbol{X}_{N\times D+1}$ has full column rank (i.e. all columns are linearly independent), then:

Principle of least squares (cont.)

If $\boldsymbol{X}_{N\times D+1}$ has full column rank (i.e. all columns are linearly independent), then:

- $\boldsymbol{X}^\top \boldsymbol{X}$ is positive definite
- $\boldsymbol{X}^\top \boldsymbol{X}$ is nonsingular and therefore invertible

This means there is a unique solution to the normal equations:

Principle of least squares (cont.)

If $\boldsymbol{X}_{N \times D+1}$ has full column rank (i.e. all columns are linearly independent), then:

- $\boldsymbol{X}^\top \boldsymbol{X}$ is positive definite
- $\boldsymbol{X}^\top \boldsymbol{X}$ is nonsingular and therefore invertible

This means there is a unique solution to the normal equations:

$$\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0 \tag{16}$$

Principle of least squares (cont.)

If $\boldsymbol{X}_{N\times D+1}$ has full column rank (i.e. all columns are linearly independent), then:

- $\boldsymbol{X}^\top\boldsymbol{X}$ is positive definite
- $\boldsymbol{X}^\top\boldsymbol{X}$ is nonsingular and therefore invertible

This means there is a unique solution to the normal equations:

$$\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}) = 0 \tag{16}$$

$$(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{y} \tag{17}$$

Principle of least squares (cont.)

If $\boldsymbol{X}_{N \times D+1}$ has full column rank (i.e. all columns are linearly independent), then:

- $\boldsymbol{X}^\top \boldsymbol{X}$ is positive definite
- $\boldsymbol{X}^\top \boldsymbol{X}$ is nonsingular and therefore invertible

This means there is a unique solution to the normal equations:

$$\begin{aligned} \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) &= 0 \quad (16) \\ (\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{w} &= \boldsymbol{X}^\top \boldsymbol{y} \quad (17) \\ \hat{\boldsymbol{w}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \quad (18) \end{aligned}$$

Principle of least squares (cont.)

If $\boldsymbol{X}_{N\times D+1}$ has full column rank (i.e. all columns are linearly independent), then:

- $\boldsymbol{X}^\top \boldsymbol{X}$ is positive definite
- $\boldsymbol{X}^\top \boldsymbol{X}$ is nonsingular and therefore invertible

This means there is a unique solution to the normal equations:

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0 \tag{16}$$

$$(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{y} \tag{17}$$

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \tag{18}$$

We therefore obtain the predictions as:

Principle of least squares (cont.)

If $\boldsymbol{X}_{N\times D+1}$ has full column rank (i.e. all columns are linearly independent), then:

- $\boldsymbol{X}^\top\boldsymbol{X}$ is positive definite
- $\boldsymbol{X}^\top\boldsymbol{X}$ is nonsingular and therefore invertible

This means there is a unique solution to the normal equations:

$$\begin{aligned}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}) &= 0 && (16)\\ (\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{w} &= \boldsymbol{X}^\top\boldsymbol{y} && (17)\\ \hat{\boldsymbol{w}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} && (18)\end{aligned}$$

We therefore obtain the predictions as:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \tag{19}$$

Ordinary least squares (OLS) model

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

$\hat{\boldsymbol{w}}$ is unbiased:

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

$\hat{\boldsymbol{w}}$ is unbiased: $\mathbb{E}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{w}}$.

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

$\hat{\boldsymbol{w}}$ is unbiased: $\mathbb{E}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{w}}$.
The error terms are assumed to satisfy:

Ordinary least squares (OLS) model

$$\underset{(N\times1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times1)}{\boldsymbol{w}} + \underset{(N\times1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

$\hat{\boldsymbol{w}}$ is unbiased: $\mathbb{E}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{w}}$.
The error terms are assumed to satisfy:

- **Zero mean:**

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

$\hat{\boldsymbol{w}}$ is unbiased: $\mathbb{E}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{w}}$.
The error terms are assumed to satisfy:

- **Zero mean:** $\mathbb{E}(\epsilon) = \mathbf{0}$

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

$\hat{\boldsymbol{w}}$ is unbiased: $\mathbb{E}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{w}}$.
The error terms are assumed to satisfy:

- **Zero mean:** $\mathbb{E}(\epsilon) = \mathbf{0}$
- **Constant variance:**

Ordinary least squares (OLS) model

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times 1)}{\boldsymbol{w}} + \underset{(N\times 1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

$\hat{\boldsymbol{w}}$ is unbiased: $\mathbb{E}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{w}}$.
The error terms are assumed to satisfy:

- **Zero mean:** $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$
- **Constant variance:** $Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$

Ordinary least squares (OLS) model

$$\underset{(N\times1)}{\boldsymbol{y}} = \underset{(N\times(D+1))}{\boldsymbol{X}} \cdot \underset{((D+1)\times1)}{\boldsymbol{w}} + \underset{(N\times1)}{\boldsymbol{\epsilon}} \tag{20}$$

where the least squares estimator can be written compactly in matrix form as:

$$\hat{\boldsymbol{w}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{21}$$

$\hat{\boldsymbol{w}}$ is unbiased: $\mathbb{E}(\hat{\boldsymbol{w}}) = \hat{\boldsymbol{w}}$.
The error terms are assumed to satisfy:

- **Zero mean:** $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$
- **Constant variance:** $Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$

The regression line passes through the centroid of the data points $(\overline{x}, \overline{y})$

Obtaining predictions

Obtaining predictions

Once we have the OLS estimate, $\hat{\boldsymbol{w}}_{OLS}$, we can predict a set of responses $\hat{\boldsymbol{y}}$ from observations $\boldsymbol{X}$ using:

Obtaining predictions

Once we have the OLS estimate, $\hat{\boldsymbol{w}}_{OLS}$, we can predict a set of responses $\hat{\boldsymbol{y}}$ from observations $\boldsymbol{X}$ using:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}_{OLS}$$

Obtaining predictions

Once we have the OLS estimate, $\hat{\boldsymbol{w}}_{OLS}$, we can predict a set of responses $\hat{\boldsymbol{y}}$ from observations $\boldsymbol{X}$ using:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}_{OLS} \tag{22}$$

Obtaining predictions

Once we have the OLS estimate, $\hat{\boldsymbol{w}}_{OLS}$, we can predict a set of responses $\hat{\boldsymbol{y}}$ from observations $\boldsymbol{X}$ using:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{w}}_{OLS} \quad (22) \\ &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \quad (23) \end{aligned}$$

Obtaining predictions

Once we have the OLS estimate, $\hat{\boldsymbol{w}}_{OLS}$, we can predict a set of responses $\hat{\boldsymbol{y}}$ from observations $\boldsymbol{X}$ using:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{w}}_{OLS} \tag{22}$$

$$= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \tag{23}$$

$$= \boldsymbol{X}^\dagger\boldsymbol{y} \tag{24}$$

Obtaining predictions

Once we have the OLS estimate, $\hat{\boldsymbol{w}}_{OLS}$, we can predict a set of responses $\hat{\boldsymbol{y}}$ from observations $\boldsymbol{X}$ using:

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{w}}_{OLS} && (22)\\
&= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} && (23)\\
&= \boldsymbol{X}^\dagger\boldsymbol{y} && (24)
\end{aligned}$$

- $\boldsymbol{X}^\dagger = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is known as the **hat matrix** or **projection matrix** or left pseudo-inverse of $\boldsymbol{X}$

Obtaining predictions

Once we have the OLS estimate, $\hat{\boldsymbol{w}}_{OLS}$, we can predict a set of responses $\hat{\boldsymbol{y}}$ from observations $\boldsymbol{X}$ using:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{w}}_{OLS} && (22) \\ &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} && (23) \\ &= \boldsymbol{X}^\dagger\boldsymbol{y} && (24) \end{aligned}$$

- $\boldsymbol{X}^\dagger = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is known as the **hat matrix** or **projection matrix** or left pseudo-inverse of $\boldsymbol{X}$

OLS assumptions

The ordinary least squares estimate for regression coefficients gives the *B*est *L*inear *U*nbiased *E*stimate (BLUE): **Gauss-Markov theorem**.

OLS assumptions

The ordinary least squares estimate for regression coefficients gives the *B*est *L*inear *U*nbiased *E*stimate (BLUE): **Gauss-Markov theorem**.

Assumes the following conditions are met:

OLS assumptions

The ordinary least squares estimate for regression coefficients gives the *B*est *L*inear *U*nbiased *E*stimate (BLUE): **Gauss-Markov theorem**.

Assumes the following conditions are met:

- Linearity: the parameters we are estimating using the OLS method must be themselves linear.
- Randomness: our data must have been randomly sampled from the population.
- NoN-Collinearity: the regressors being calculated are not perfectly correlated with each other.
- Exogeneity: the regressors are not correlated with the error term.
- Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

Qualitative predictors

Qualitative predictors

- Qualitative predictors arise when the *levels* are categories, e.g.

Qualitative predictors

- Qualitative predictors arise when the *levels* are categories, e.g.
  - gender

Qualitative predictors

- Qualitative predictors arise when the *levels* are categories, e.g.
  - gender
  - income levels

Qualitative predictors

- Qualitative predictors arise when the *levels* are categories, e.g.
  - gender
  - income levels
  - commuting mode

Qualitative predictors

- Qualitative predictors arise when the *levels* are categories, e.g.
    - gender
    - income levels
    - commuting mode
    - type of soil

Qualitative predictors

- Qualitative predictors arise when the *levels* are categories, e.g.
  - gender
  - income levels
  - commuting mode
  - type of soil
  - climate zone

Qualitative predictors

- Qualitative predictors arise when the *levels* are categories, e.g.
  - gender
  - income levels
  - commuting mode
  - type of soil
  - climate zone
- These predictors are referred to as *factors*

Representing qualitative predictors: two-level case

Representing qualitative predictors: two-level case

A factor with two possible values is represented by an **indicator** (or **dummy**) variable.

Representing qualitative predictors: two-level case

A factor with two possible values is represented by an **indicator** (or **dummy**) variable.

For instance, if X is a factor denoting married status, then:

Representing qualitative predictors: two-level case

A factor with two possible values is represented by an **indicator** (or **dummy**) variable.

For instance, if X is a factor denoting married status, then:

$$x_n = \begin{cases} 0 & \text{if } i\text{th person is unmarried} \\ 1 & \text{if } i\text{th person is married} \end{cases}$$

Representing qualitative predictors: two-level case

A factor with two possible values is represented by an **indicator** (or **dummy**) variable.

For instance, if X is a factor denoting married status, then:

$$x_n = \begin{cases} 0 & \text{if } i\text{th person is unmarried} \\ 1 & \text{if } i\text{th person is married} \end{cases}$$

Then in the simple case, the linear model would be:

Representing qualitative predictors: two-level case

A factor with two possible values is represented by an **indicator** (or **dummy**) variable.

For instance, if X is a factor denoting married status, then:

$$x_n = \begin{cases} 0 & \text{if } i\text{th person is unmarried} \\ 1 & \text{if } i\text{th person is married} \end{cases}$$

Then in the simple case, the linear model would be:

$$y_n = w_0 + w_1 x_n + \epsilon_n = \begin{cases} w_0 + \epsilon_n & \text{if } i\text{th person is unmarried} \\ w_0 + w_1 + \epsilon_n & \text{if } i\text{th person is married} \end{cases}$$

Representing qualitative predictors: two-level case

A factor with two possible values is represented by an **indicator** (or **dummy**) variable.

For instance, if X is a factor denoting married status, then:

$$x_n = \begin{cases} 0 & \text{if } i\text{th person is unmarried} \\ 1 & \text{if } i\text{th person is married} \end{cases}$$

Then in the simple case, the linear model would be:

$$y_n = w_0 + w_1 x_n + \epsilon_n = \begin{cases} w_0 + \epsilon_n & \text{if } i\text{th person is unmarried} \\ w_0 + w_1 + \epsilon_n & \text{if } i\text{th person is married} \end{cases}$$

An alternative to the $\{0/1\}$ scheme is $\{-1/1\}$.

Representing qualitative predictors: multi-level case

Representing qualitative predictors: multi-level case

When more than two levels exist, additional dummy variables may be used.

Representing qualitative predictors: multi-level case

When more than two levels exist, additional dummy variables may be used.

Generally, $k-1$ dummy variables are used to represent k levels.

Representing qualitative predictors: multi-level case

When more than two levels exist, additional dummy variables may be used.

Generally, $k-1$ dummy variables are used to represent k levels.

$$x_{nB} = \begin{cases} 0 & \notin B \\ 1 & \in B \end{cases}$$

$$x_{nC} = \begin{cases} 0 & \notin C \\ 1 & \in C \end{cases}$$

$$y_n = w_0 + w_B x_{nB} + w_C x_{nC} = \begin{cases} w_0 + \epsilon_n & \notin B \cup C \\ w_0 + w_B + \epsilon_n & \in B \\ w_0 + w_C + \epsilon_n & \in C \end{cases}$$

The level with no dummy variable is termed the **baseline**. If $A = \overline{B \cup C}$, then A is the baseline.

Representing qualitative predictors: multi-level case

When more than two levels exist, additional dummy variables may be used.

Generally, $k-1$ dummy variables are used to represent k levels.

$$x_{nB} = \begin{cases} 0 & \notin B \\ 1 & \in B \end{cases}$$

$$y_n = w_0 + w_B x_{nB} + w_C x_{nC} = \begin{cases} w_0 + \epsilon_n & \notin B \cup C \\ w_0 + w_B + \epsilon_n & \in B \\ w_0 + w_C + \epsilon_n & \in C \end{cases}$$

The level with no dummy variable is termed the **baseline**. If $A = \overline{B \cup C}$, then A is the baseline.

Representing qualitative predictors: multi-level case

When more than two levels exist, additional dummy variables may be used.

Generally, $k-1$ dummy variables are used to represent k levels.

$$x_{nB} = \begin{cases} 0 & \notin B \\ 1 & \in B \end{cases}$$

$$x_{nC} = \begin{cases} 0 & \notin C \\ 1 & \in C \end{cases}$$

$$y_n = w_0 + w_B x_{nB} + w_C x_{nC} = \begin{cases} w_0 + \epsilon_n & \notin B \cup C \\ w_0 + w_B + \epsilon_n & \in B \\ w_0 + w_C + \epsilon_n & \in C \end{cases}$$

Representing qualitative predictors: multi-level case

When more than two levels exist, additional dummy variables may be used.

Generally, $k-1$ dummy variables are used to represent k levels.

$$x_{nB} = \begin{cases} 0 & \notin B \\ 1 & \in B \end{cases}$$

$$x_{nC} = \begin{cases} 0 & \notin C \\ 1 & \in C \end{cases}$$

$$y_n = w_0 + w_B x_{nB} + w_C x_{nC} = \begin{cases} w_0 + \epsilon_n & \notin B \cup C \\ w_0 + w_B + \epsilon_n & \in B \\ w_0 + w_C + \epsilon_n & \in C \end{cases}$$

The level with no dummy variable is termed the **baseline**.

Representing qualitative predictors: multi-level case

When more than two levels exist, additional dummy variables may be used.

Generally, $k - 1$ dummy variables are used to represent k levels.

$$x_{nB} = \begin{cases} 0 & \notin B \\ 1 & \in B \end{cases}$$

$$x_{nC} = \begin{cases} 0 & \notin C \\ 1 & \in C \end{cases}$$

$$y_n = w_0 + w_B x_{nB} + w_C x_{nC} = \begin{cases} w_0 + \epsilon_n & \notin B \cup C \\ w_0 + w_B + \epsilon_n & \in B \\ w_0 + w_C + \epsilon_n & \in C \end{cases}$$

The level with no dummy variable is termed the **baseline**. If $A = \overline{B \cup C}$, then A is the baseline.

Interaction terms

Interaction terms

We introduce interaction terms into a linear model if *synergistic* effects are observed between two variables, i.e. the effect of one variable depends on the value of the other.

Interaction terms

We introduce interaction terms into a linear model if *synergistic* effects are observed between two variables, i.e. the effect of one variable depends on the value of the other.

For instance, if x_1 and x_2 interact in the standard model:

Interaction terms

We introduce interaction terms into a linear model if *synergistic* effects are observed between two variables, i.e. the effect of one variable depends on the value of the other.

For instance, if x_1 and x_2 interact in the standard model:

$$Y = w_0 + w_1 x_1 + w_2 x_2 + \epsilon \tag{25}$$

Interaction terms

We introduce interaction terms into a linear model if *synergistic* effects are observed between two variables, i.e. the effect of one variable depends on the value of the other.

For instance, if x_1 and x_2 interact in the standard model:

$$Y = w_0 + w_1 x_1 + w_2 x_2 + \epsilon \tag{25}$$

then, we include the **interaction term** $x_1 x_2$:

Interaction terms

We introduce interaction terms into a linear model if *synergistic* effects are observed between two variables, i.e. the effect of one variable depends on the value of the other.

For instance, if x_1 and x_2 interact in the standard model:

$$Y = w_0 + w_1 x_1 + w_2 x_2 + \epsilon \tag{25}$$

then, we include the **interaction term** $x_1 x_2$:

$$Y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + \epsilon \tag{26}$$

In models with interaction terms, the **main effects** must be included by keeping the single variables.

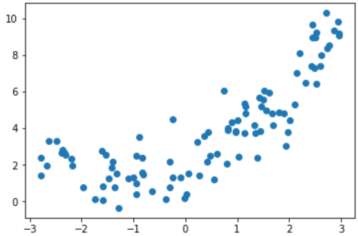
Polynomial regression

Polynomial regression

In some cases, a polynomial function provides a good approximation to the true regression function for a given dataset.

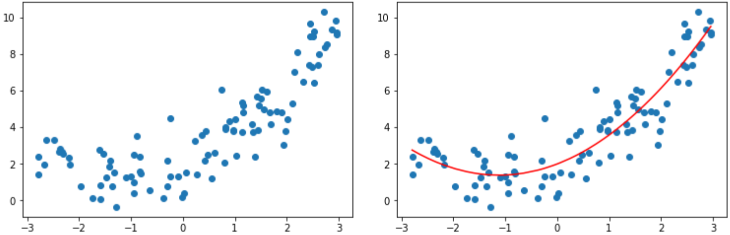
Polynomial regression

In some cases, a polynomial function provides a good approximation to the true regression function for a given dataset.



Polynomial regression

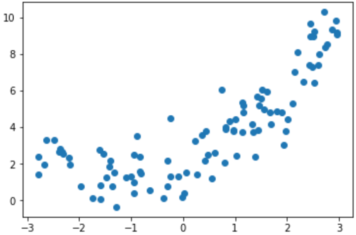
In some cases, a polynomial function provides a good approximation to the true regression function for a given dataset.

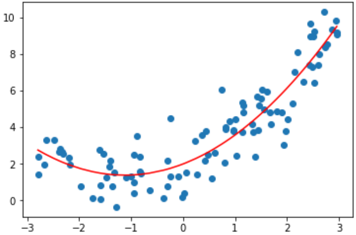


d-th degree polynomial regression model equation

Polynomial regression

In some cases, a polynomial function provides a good approximation to the true regression function for a given dataset.



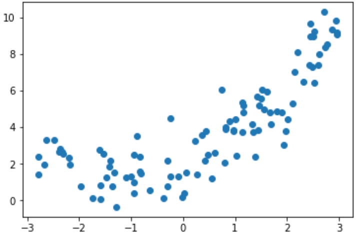


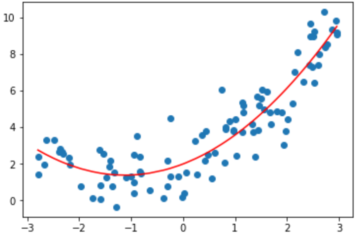
d-th degree polynomial regression model equation

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d + \epsilon \tag{27}$$

Polynomial regression

In some cases, a polynomial function provides a good approximation to the true regression function for a given dataset.





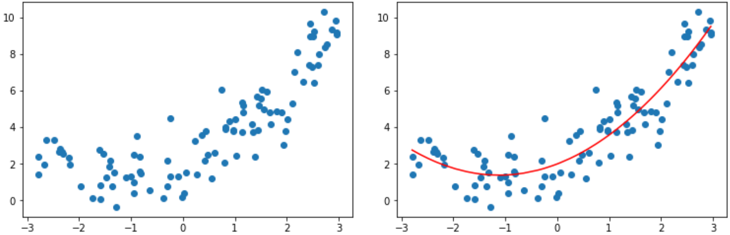
d-th degree polynomial regression model equation

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d + \epsilon \tag{27}$$

where ϵ is normally distributed:

Polynomial regression

In some cases, a polynomial function provides a good approximation to the true regression function for a given dataset.



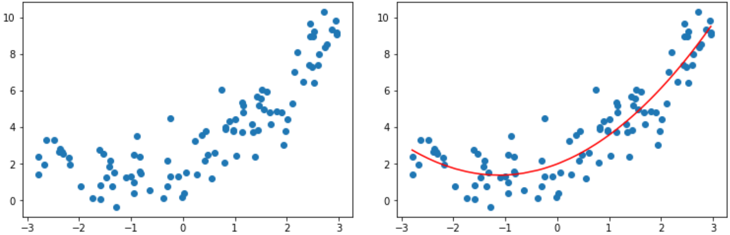
d-th degree polynomial regression model equation

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d + \epsilon \tag{27}$$

where ϵ is normally distributed: $\mathcal{N}(0, \sigma^2)$

Polynomial regression

In some cases, a polynomial function provides a good approximation to the true regression function for a given dataset.



d-th degree polynomial regression model equation

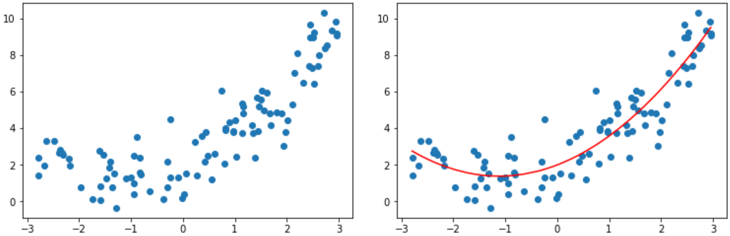
$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d + \epsilon \tag{27}$$

where ϵ is normally distributed: $\mathcal{N}(0, \sigma^2)$

This means that the expected value of y given x can be written as:

Polynomial regression

In some cases, a polynomial function provides a good approximation to the true regression function for a given dataset.



d-th degree polynomial regression model equation

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d + \epsilon \tag{27}$$

where ϵ is normally distributed: $\mathcal{N}(0, \sigma^2)$

This means that the expected value of y given x can be written as:

$$\mathbb{E}(y|x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d \tag{28}$$

Polynomial regression and multiple regression

Polynomial regression can be considered a special case of multiple linear regression.

Polynomial regression and multiple regression

Polynomial regression can be considered a special case of multiple linear regression.
For example, consider the model:

$$Y = w_0 + w_1 x + w_2 x^2 + \epsilon$$

Polynomial regression and multiple regression

Polynomial regression can be considered a special case of multiple linear regression.
For example, consider the model:

$$Y = w_0 + w_1 x + w_2 x^2 + \epsilon$$

If we set $x \to z_1$ and $x^2 \to z_2$, then the model becomes:

Polynomial regression and multiple regression

Polynomial regression can be considered a special case of multiple linear regression.
For example, consider the model:

$$Y = w_0 + w_1 x + w_2 x^2 + \epsilon$$

If we set $x \to z_1$ and $x^2 \to z_2$, then the model becomes:

$$Y = w_0 + w_1 z_1 + w_2 z_2 + \epsilon$$

Polynomial regression and multiple regression

Polynomial regression can be considered a special case of multiple linear regression.
For example, consider the model:

$$Y = w_0 + w_1 x + w_2 x^2 + \epsilon$$

If we set $x \to z_1$ and $x^2 \to z_2$, then the model becomes:

$$Y = w_0 + w_1 z_1 + w_2 z_2 + \epsilon$$

which can be taken as a multiple linear regression model with two predictor variables.

Improving model quality via higher-order terms

Improving model quality via higher-order terms

The presence of nonlinearity in the "residuals vs. fitted" plot usually indicates the form of the required polynomial.

Improving model quality via higher-order terms

The presence of nonlinearity in the “residuals vs. fitted” plot usually indicates the form of the required polynomial.

Example 2.1: `mpg` $\sim$ `weight`

In the `Auto` dataset, we regress `mpg` on `weight`.

Improving model quality via higher-order terms

The presence of nonlinearity in the "residuals vs. fitted" plot usually indicates the form of the required polynomial.

Example 2.1: `mpg` ∼ `weight`

In the `Auto` dataset, we regress `mpg` on `weight`.

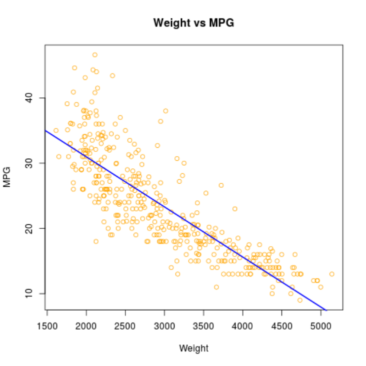


Figure: Simple linear model: $y = w_0 + w_1 x$. $R^2 = 0.69$. Is this a good fit?

Improving model quality via higher-order terms (cont.)

Improving model quality via higher-order terms (cont.)

Example 2.1: `mpg` $\sim$ `weight` (cont.)

We observe a pattern in the residual plot against the fitted values.

Improving model quality via higher-order terms (cont.)

Example 2.1: `mpg` ∼ `weight` (cont.)

We observe a pattern in the residual plot against the fitted values.

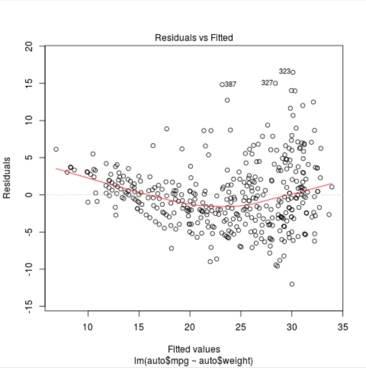


Figure: "Residuals vs. fitted values" plot for the simple linear model. What do you observe?

Improving model quality via higher-order terms (cont.)

Example 2.1: mpg $\sim$ weight (cont.)

We observe a pattern in the residual plot against the fitted values.

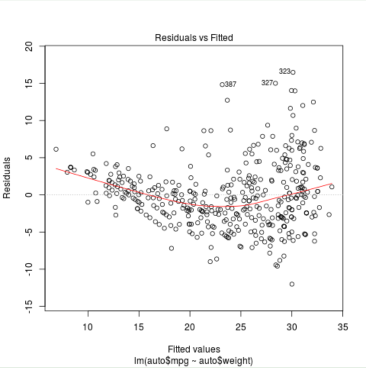


Figure: "Residuals vs. fitted values" plot for the simple linear model. What do you observe?

This further confirms that a linear fit is not the best approximation for this data.

Improving model quality via higher-order terms (cont.)

Improving model quality via higher-order terms (cont.)

Example 2.1: mpg ∼ weight (cont.)

Improving model quality via higher-order terms (cont.)

Example 2.1: `mpg` $\sim$ `weight` (cont.)

Now, we try the model:

Improving model quality via higher-order terms (cont.)

Example 2.1: `mpg` $\sim$ `weight` (cont.)

Now, we try the model: $Y = w_0 + w_1 X + w_2 X^2$.

Improving model quality via higher-order terms (cont.)

Example 2.1: mpg ∼ weight (cont.)

Now, we try the model: $Y = w_0 + w_1 X + w_2 X^2$.

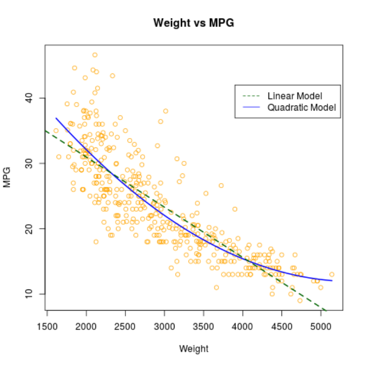


Figure: Linear and quadratic models for regressing mpg on weight.

Improving model quality via higher-order terms (cont.)

Example 2.1: `mpg` $\sim$ `weight` (cont.)

Now, we try the model: $Y = w_0 + w_1 X + w_2 X^2$.

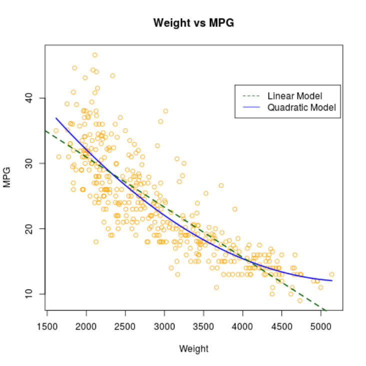


Figure: Linear and quadratic models for regressing `mpg` on `weight`.

The adjusted R^2 is now 0.71.

Improving model quality via higher-order terms (cont.)

Improving model quality via higher-order terms (cont.)

Example 2.1: `mpg` $\sim$ `weight` (cont.)

Improving model quality via higher-order terms (cont.)

Example 2.1: mpg ∼ weight (cont.)

We compare the residual plots.

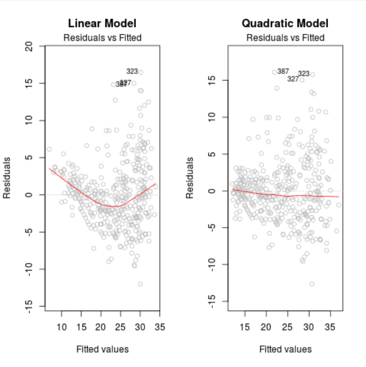


Figure: Residual plots for the linear and quadratic models for regressing `mpg` on `weight`.

Improving model quality via higher-order terms (cont.)

Example 2.1: `mpg` ~ `weight` (cont.)

We compare the residual plots.

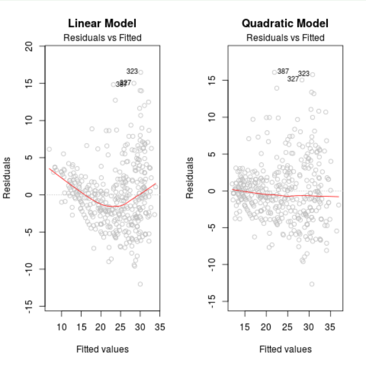


Figure: Residual plots for the linear and quadratic models for regressing `mpg` on `weight`.

While we have explained some of the variance with the second-order term, the statistics reveal that other variables may also be important predictors.

When OLS assumptions fail

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

Data issues encountered in linear regression

1. Nonlinearity

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

Data issues encountered in linear regression

1. Nonlinearity
2. Correlation of error terms

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

Data issues encountered in linear regression

1. Nonlinearity
2. Correlation of error terms
3. Non-constant variance of error (heteroscedasticity)

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

Data issues encountered in linear regression

1. Nonlinearity
2. Correlation of error terms
3. Non-constant variance of error (heteroscedasticity)
4. Outliers

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

Data issues encountered in linear regression

1. Nonlinearity
2. Correlation of error terms
3. Non-constant variance of error (heteroscedasticity)
4. Outliers
5. High-leverage points

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

Data issues encountered in linear regression

1. Nonlinearity
2. Correlation of error terms
3. Non-constant variance of error (heteroscedasticity)
4. Outliers
5. High-leverage points
6. Collinearity

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

Data issues encountered in linear regression

1. Nonlinearity
2. Correlation of error terms
3. Non-constant variance of error (heteroscedasticity)
4. Outliers
5. High-leverage points
6. Collinearity

When OLS assumptions fail

In some cases, assumptions for the ordinary least squares may not hold due a combination of the following.

Data issues encountered in linear regression

1. Nonlinearity
2. Correlation of error terms
3. Non-constant variance of error (heteroscedasticity)
4. Outliers
5. High-leverage points
6. Collinearity

We discuss a few approaches to handle these.

Nonlinearity

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w} x}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w} x}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \quad \xrightarrow{y' = \ln(y)}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \quad \xrightarrow{y'=\ln(y)} \quad y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \quad \xrightarrow{y'=\ln(y)} \quad y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \xrightarrow{y'=\ln(y)} y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}} \xrightarrow{y'=\ln(y),\, x'=\ln(x)}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \xrightarrow{y'=\ln(y)} y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}} \xrightarrow{y'=\ln(y),\, x'=\ln(x)} y' = \ln(\alpha) + \boldsymbol{w}x'$
3. $y = \alpha + \boldsymbol{w} \cdot \log(x)$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \xrightarrow{y'=\ln(y)} y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}} \xrightarrow{y'=\ln(y),\, x'=\ln(x)} y' = \ln(\alpha) + \boldsymbol{w}x'$
3. $y = \alpha + \boldsymbol{w} \cdot \log(x)$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \xrightarrow{y'=\ln(y)} y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}} \xrightarrow{y'=\ln(y),\, x'=\ln(x)} y' = \ln(\alpha) + \boldsymbol{w}x'$
3. $y = \alpha + \boldsymbol{w} \cdot \log(x) \xrightarrow{x'=\log(x)}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \xrightarrow{y'=\ln(y)} y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}} \xrightarrow{y'=\ln(y),\, x'=\ln(x)} y' = \ln(\alpha) + \boldsymbol{w}x'$
3. $y = \alpha + \boldsymbol{w} \cdot \log(x) \xrightarrow{x'=\log(x)} y = \alpha + \boldsymbol{w}x'$
4. $y = \alpha + \boldsymbol{w} \cdot \frac{1}{x}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \xrightarrow{y'=\ln(y)} y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}} \xrightarrow{y'=\ln(y),\, x'=\ln(x)} y' = \ln(\alpha) + \boldsymbol{w}x'$
3. $y = \alpha + \boldsymbol{w} \cdot \log(x) \xrightarrow{x'=\log(x)} y = \alpha + \boldsymbol{w}x'$
4. $y = \alpha + \boldsymbol{w} \cdot \frac{1}{x}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \xrightarrow{y'=\ln(y)} y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}} \xrightarrow{y'=\ln(y),\, x'=\ln(x)} y' = \ln(\alpha) + \boldsymbol{w}x'$
3. $y = \alpha + \boldsymbol{w} \cdot \log(x) \xrightarrow{x'=\log(x)} y = \alpha + \boldsymbol{w}x'$
4. $y = \alpha + \boldsymbol{w} \cdot \frac{1}{x} \xrightarrow{x'=\frac{1}{x}}$

Nonlinearity

There are situations in which we can tell from investigations that the predictors do not exhibit a linear relationship with the response variable.

To handle these, we consider **linearly transforming** the data (either the inputs or outputs or both.)

Useful intrinsically linear functions

1. $y = \alpha e^{\boldsymbol{w}x} \xrightarrow{y'=\ln(y)} y' = \ln(\alpha) + \boldsymbol{w}x$
2. $y = \alpha x^{\boldsymbol{w}} \xrightarrow{y'=\ln(y),\, x'=\ln(x)} y' = \ln(\alpha) + \boldsymbol{w}x'$
3. $y = \alpha + \boldsymbol{w} \cdot \log(x) \xrightarrow{x'=\log(x)} y = \alpha + \boldsymbol{w}x'$
4. $y = \alpha + \boldsymbol{w} \cdot \frac{1}{x} \xrightarrow{x'=\frac{1}{x}} y = \alpha + \boldsymbol{w}x'$

Autocorrelation

Correlation of error terms can lead to underestimation of coefficient standard errors:

Autocorrelation

Correlation of error terms can lead to underestimation of coefficient standard errors: $Cov(\epsilon_n, \epsilon_k) \neq 0 \quad \forall \quad n \neq k, k = 1, ..., n$

Autocorrelation

Correlation of error terms can lead to underestimation of coefficient standard errors: $Cov(\epsilon_n, \epsilon_k) \neq 0 \quad \forall \quad n \neq k, k = 1, ..., n$

- Identified by patterns in residual trends

Autocorrelation

Correlation of error terms can lead to underestimation of coefficient standard errors: $Cov(\epsilon_n, \epsilon_k) \neq 0 \quad \forall \quad n \neq k, k = 1, ..., n$

- Identified by patterns in residual trends
- This usually occurs in time series or geographical data

Autocorrelation

Correlation of error terms can lead to underestimation of coefficient standard errors: $Cov(\epsilon_n, \epsilon_k) \neq 0 \quad \forall \quad n \neq k, k = 1, ..., n$

- Identified by patterns in residual trends
- This usually occurs in time series or geographical data

Autocorrelation

Correlation of error terms can lead to underestimation of coefficient standard errors: $Cov(\epsilon_n, \epsilon_k) \neq 0 \quad \forall \quad n \neq k, k = 1, ..., n$

- Identified by patterns in residual trends
- This usually occurs in time series or geographical data

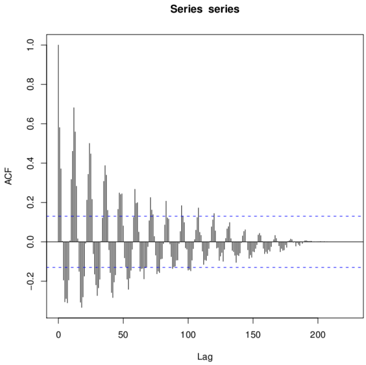


Figure: Example of error correlations for a time series dataset. The blue dashed lines represent the 95% confidence interval. Correlations outside of this band are statistically

Non-constant variance of error terms

Non-constant variance of error terms

Recall a fundamental assumption of linear regression is **homoscedasticity**, i.e.

Non-constant variance of error terms

Recall a fundamental assumption of linear regression is **homoscedasticity**, i.e.

$$Var(\epsilon_n) = \sigma^2 \tag{29}$$

Non-constant variance of error terms

Recall a fundamental assumption of linear regression is **homoscedasticity**, i.e.

$$Var(\epsilon_n) = \sigma^2 \quad (29)$$

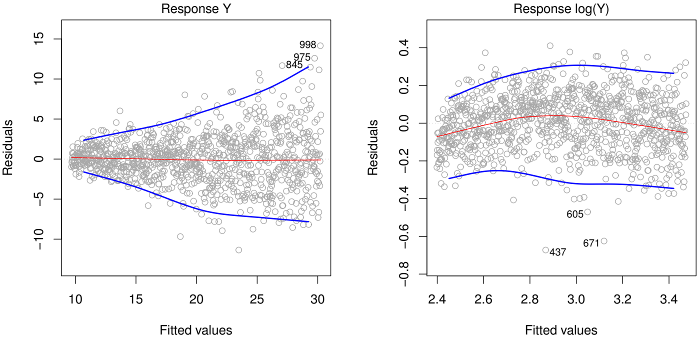


Figure: Heteroscedasticity (left) and homoscedasticity (right); fixed by log-transformation of the response.

Outliers

Outliers

- Outliers occur when some observations produce residuals that are significantly larger than average

Outliers

- Outliers occur when some observations produce residuals that are significantly larger than average
- They can be efficiently identified using **studentized residuals**:

$$t_n = \frac{\hat{e}_n}{RSE\sqrt{1 - h_n}} \tag{30}$$

Outliers

- Outliers occur when some observations produce residuals that are significantly larger than average
- They can be efficiently identified using **studentized residuals**:

$$t_n = \frac{\hat{e}_n}{RSE\sqrt{1 - h_n}} \tag{30}$$

- To address this issue, outliers can be **carefully** removed from the training data

Outliers

- Outliers occur when some observations produce residuals that are significantly larger than average
- They can be efficiently identified using **studentized residuals**:

$$t_n = \frac{\hat{e}_n}{RSE\sqrt{1 - h_n}} \tag{30}$$

- To address this issue, outliers can be **carefully** removed from the training data

Outliers

- Outliers occur when some observations produce residuals that are significantly larger than average
- They can be efficiently identified using **studentized residuals**:

$$t_n = \frac{\hat{e}_n}{RSE\sqrt{1 - h_n}} \tag{30}$$

- To address this issue, outliers can be **carefully** removed from the training data

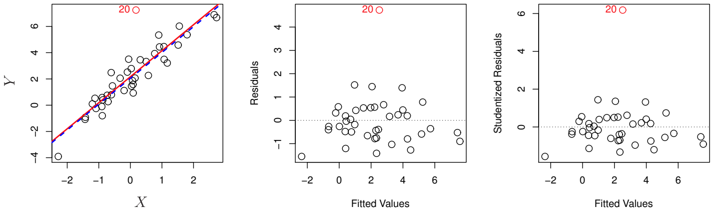


Figure: Outlier in a regression model. The studentized residual confirms the outlier cannot be ignored.

High leverage points

High leverage points

Leverage

The leverage of an observation is a measure of its influence on the regression model:

High leverage points

Leverage

The leverage of an observation is a measure of its influence on the regression model:

$$h_n = \frac{1}{n} + \frac{(x_n - \overline{x})^2}{\sum_{k=1}^{N}(x_k - \overline{x})^2} \tag{31}$$

High leverage points

Leverage

The leverage of an observation is a measure of its influence on the regression model:

$$h_n = \frac{1}{n} + \frac{(x_n - \overline{x})^2}{\sum_{k=1}^{N}(x_k - \overline{x})^2} \tag{31}$$

- Leverage increases with the deviation from the mean

High leverage points

Leverage

The leverage of an observation is a measure of its influence on the regression model:

$$h_n = \frac{1}{n} + \frac{(x_n - \overline{x})^2}{\sum_{k=1}^{N}(x_k - \overline{x})^2} \tag{31}$$

- Leverage increases with the deviation from the mean
- Leverage is often plotted against the standardized residuals to identify problematic observations

High leverage points

Leverage

The leverage of an observation is a measure of its influence on the regression model:

$$h_n = \frac{1}{n} + \frac{(x_n - \overline{x})^2}{\sum_{k=1}^{N}(x_k - \overline{x})^2} \tag{31}$$

- Leverage increases with the deviation from the mean
- Leverage is often plotted against the standardized residuals to identify problematic observations
- Another quantity, the Cook's distance, explicitly computes the scaled average of the changes to regression model when the observation of interest is removed

High leverage points

Leverage

The leverage of an observation is a measure of its influence on the regression model:

$$h_n = \frac{1}{n} + \frac{(x_n - \overline{x})^2}{\sum_{k=1}^{N}(x_k - \overline{x})^2} \tag{31}$$

- Leverage increases with the deviation from the mean
- Leverage is often plotted against the standardized residuals to identify problematic observations
- Another quantity, the Cook's distance, explicitly computes the scaled average of the changes to regression model when the observation of interest is removed
- $\frac{1}{n} \leq h_n \leq 1$; $\mathbb{E}(h_n) = \frac{p+1}{n}$

Collinearity

Collinearity

Collinearity arises when two or more predictors are correlated.

Collinearity

Collinearity arises when two or more predictors are correlated. This can usually be identified via correlation plots:

Collinearity

Collinearity arises when two or more predictors are correlated. This can usually be identified via correlation plots:

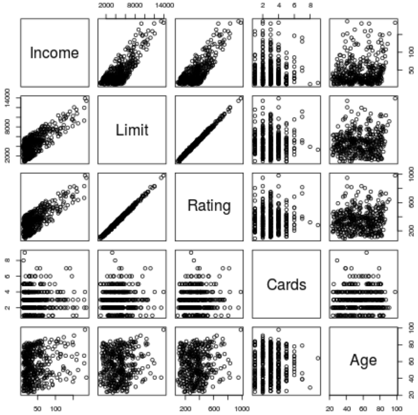


Figure: Correlation plot for a few of the predictors in the `Credit` dataset. Which of the predictors in this plot are highly correlated?

Identifying collinearity

Identifying collinearity

Consequences

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

How do we **identify** collinearity?

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

How do we **identify** collinearity?

- Identify using correlation plots and values

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

How do we **identify** collinearity?

- Identify using correlation plots and values
- Use the variance inflation factor (VIF), which is robust to *multicollinearity*:

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

How do we **identify** collinearity?

- Identify using correlation plots and values
- Use the variance inflation factor (VIF), which is robust to *multicollinearity*:

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

How do we **identify** collinearity?

- Identify using correlation plots and values
- Use the variance inflation factor (VIF), which is robust to *multicollinearity*:

$$VIF(\hat{w}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{32}$$

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

How do we **identify** collinearity?

- Identify using correlation plots and values
- Use the variance inflation factor (VIF), which is robust to *multicollinearity*:

$$VIF(\hat{w}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{32}$$

  where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

- Variables with similarly high VIFs are multicollinear.

Identifying collinearity

Consequences

- Increases uncertainty of model estimates (standard errors)
- Reduces the power of the hypothesis test (probability of correctly rejecting the null hypothesis)

How do we **identify** collinearity?

- Identify using correlation plots and values
- Use the variance inflation factor (VIF), which is robust to *multicollinearity*:

$$VIF(\hat{w}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{32}$$

  where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.
- Variables with similarly high VIFs are multicollinear.
- The smallest value of the VIF is 1.

Correcting for collinearity

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

- Remove the problematic variables (e.g. if 3 variables are multicollinear, drop 2 of them from the model)

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

- Remove the problematic variables (e.g. if 3 variables are multicollinear, drop 2 of them from the model)
- Create a composite variable from the collinear variables

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

- Remove the problematic variables (e.g. if 3 variables are multicollinear, drop 2 of them from the model)
- Create a composite variable from the collinear variables

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

- Remove the problematic variables (e.g. if 3 variables are multicollinear, drop 2 of them from the model)
- Create a composite variable from the collinear variables

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset)

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

- Remove the problematic variables (e.g. if 3 variables are multicollinear, drop 2 of them from the model)
- Create a composite variable from the collinear variables

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset)

For example, if we estimate the model:

$$Y = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3 + w_4 X_4$$

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

- Remove the problematic variables (e.g. if 3 variables are multicollinear, drop 2 of them from the model)
- Create a composite variable from the collinear variables

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset)

For example, if we estimate the model:

$$Y = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3 + w_4 X_4$$

where Y, X_1, X_2, X_3 are `Balance`, `Income`, `Limit`, `Rating` and `Age`, respectively.

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

- Remove the problematic variables (e.g. if 3 variables are multicollinear, drop 2 of them from the model)
- Create a composite variable from the collinear variables

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset)

For example, if we estimate the model:

$$Y = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3 + w_4 X_4$$

where Y, X_1, X_2, X_3 are `Balance`, `Income`, `Limit`, `Rating` and `Age`, respectively. The adjusted R^2 is 0.876 and the VIF values are:

Correcting for collinearity

To correct for [multi]collinearity, there are two approaches:

- Remove the problematic variables (e.g. if 3 variables are multicollinear, drop 2 of them from the model)
- Create a composite variable from the collinear variables

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset)

For example, if we estimate the model:

$$Y = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3 + w_4 X_4$$

where Y, X_1, X_2, X_3 are `Balance`, `Income`, `Limit`, `Rating` and `Age`, respectively. The adjusted R^2 is 0.876 and the VIF values are:

```
Income:    2.52
Limit:     149
Rating:    149
Age:       1.03
```

Correcting for collinearity (cont.)

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset); cont.

Correcting for collinearity (cont.)

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset); cont.

We see from the VIF that Limit and Age are highly correlated.

Correcting for collinearity (cont.)

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset); cont.

We see from the VIF that Limit and Age are highly correlated. So we remove `Limit` (X_2) and estimate the new model:

$$Y = w_0 + w_1 X_1 + w_3 X_3 + w_4 X_4$$

Correcting for collinearity (cont.)

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset); cont.

We see from the VIF that Limit and Age are highly correlated. So we remove `Limit` (X_2) and estimate the new model:

$$Y = w_0 + w_1 X_1 + w_3 X_3 + w_4 X_4$$

The adjusted R^2 is only slightly lower at 0.875, and the VIFs are:

Correcting for collinearity (cont.)

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset); cont.

We see from the VIF that Limit and Age are highly correlated. So we remove `Limit` (X_2) and estimate the new model:

$$Y = w_0 + w_1 X_1 + w_3 X_3 + w_4 X_4$$

The adjusted R^2 is only slightly lower at 0.875, and the VIFs are:

```
Income:   2.52
Rating:   2.48
Age:      1.02
```

Correcting for collinearity (cont.)

Example: Regressing Balance on Income, Limit, Rating and Age (`Credit` dataset); cont.

We see from the VIF that Limit and Age are highly correlated. So we remove `Limit` (X_2) and estimate the new model:

$$Y = w_0 + w_1 X_1 + w_3 X_3 + w_4 X_4$$

The adjusted R^2 is only slightly lower at 0.875, and the VIFs are:

```
Income:    2.52
Rating:    2.48
Age:       1.02
```

Thus, we have corrected for collinearity without decreasing the quality of the fit.

Weighted least squares (WLS)

Weighted least squares (WLS)

When variances are not heteroskedastic (i.e. they are not constant), then we use the WLS model:

Weighted least squares (WLS)

When variances are not heteroskedastic (i.e. they are not constant), then we use the WLS model:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2(\boldsymbol{x})$$

Weighted least squares (WLS)

When variances are not heteroskedastic (i.e. they are not constant), then we use the WLS model:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top\boldsymbol{x}, \sigma^2(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left(-\frac{1}{2\sigma^2(\boldsymbol{x})}(y - \boldsymbol{w}^\top\boldsymbol{x})^2\right) \tag{33}$$

Weighted least squares (WLS)

When variances are not heteroskedastic (i.e. they are not constant), then we use the WLS model:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2(\boldsymbol{x})) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left(-\frac{1}{2\sigma^2(\boldsymbol{x})}(y - \boldsymbol{w}^\top \boldsymbol{x})^2\right) \tag{33}$$

Thus, we can write the WLS model for the entire dataset as:

Weighted least squares (WLS)

When variances are not heteroskedastic (i.e. they are not constant), then we use the WLS model:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top\boldsymbol{x}, \sigma^2(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left(-\frac{1}{2\sigma^2(\boldsymbol{x})}(y - \boldsymbol{w}^\top\boldsymbol{x})^2\right) \tag{33}$$

Thus, we can write the WLS model for the entire dataset as:

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \boldsymbol{\Lambda}^{-1}) \tag{34}$$

Weighted least squares (WLS)

When variances are not heteroskedastic (i.e. they are not constant), then we use the WLS model:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top\boldsymbol{x}, \sigma^2(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left(-\frac{1}{2\sigma^2(\boldsymbol{x})}(y - \boldsymbol{w}^\top\boldsymbol{x})^2\right) \quad (33)$$

Thus, we can write the WLS model for the entire dataset as:

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \boldsymbol{\Lambda}^{-1}) \quad (34)$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(1/\sigma^2(\boldsymbol{x}_n))$ is a diagonal weight matrix.

Weighted least squares (WLS)

When variances are not heteroskedastic (i.e. they are not constant), then we use the WLS model:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left(-\frac{1}{2\sigma^2(\boldsymbol{x})}(y - \boldsymbol{w}^\top \boldsymbol{x})^2\right) \quad (33)$$

Thus, we can write the WLS model for the entire dataset as:

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \boldsymbol{\Lambda}^{-1}) \quad (34)$$

where $\boldsymbol{\Lambda} = \text{diag}(1/\sigma^2(\boldsymbol{x}_n))$ is a diagonal weight matrix.

- In OLS, we assume $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ [constant variance]

Weighted least squares (WLS)

When variances are not heteroskedastic (i.e. they are not constant), then we use the WLS model:

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{x})}} \exp\left(-\frac{1}{2\sigma^2(\boldsymbol{x})}(y - \boldsymbol{w}^\top \boldsymbol{x})^2\right) \quad (33)$$

Thus, we can write the WLS model for the entire dataset as:

$$p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \boldsymbol{\Lambda}^{-1}) \quad (34)$$

where $\boldsymbol{\Lambda} = \text{diag}(1/\sigma^2(\boldsymbol{x}_n))$ is a diagonal weight matrix.

- In OLS, we assume $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ [constant variance]
- In WLS, we assume $\epsilon_n \sim \mathcal{N}(0, \sigma^2(x_n)^2)$ [non-constant variance; different for each observation]

WLS (cont.)

WLS (cont.)

The WLS estimator (coefficients) is given by:

WLS (cont.)

The WLS estimator (coefficients) is given by:

$$\hat{\boldsymbol{w}}_{WLS} = (\boldsymbol{X}^\top \boldsymbol{\Lambda} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{y} \tag{35}$$

WLS (cont.)

The WLS estimator (coefficients) is given by:

$$\hat{\boldsymbol{w}}_{WLS} = (\boldsymbol{X}^\top \boldsymbol{\Lambda} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{y} \tag{35}$$

- Since $\boldsymbol{\Lambda} = \mathrm{diag}(1/\sigma^2(\boldsymbol{x}_n)) = \mathrm{diag}(w_1, \ldots, w_N)$ is a diagonal weight matrix, we can write $\boldsymbol{\Lambda} = \boldsymbol{W}$.

WLS (cont.)

The WLS estimator (coefficients) is given by:

$$\hat{\boldsymbol{w}}_{WLS} = (\boldsymbol{X}^\top \boldsymbol{\Lambda} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{y} \tag{35}$$

- Since $\boldsymbol{\Lambda} = \mathrm{diag}(1/\sigma^2(\boldsymbol{x}_n)) = \mathrm{diag}(w_1, \ldots, w_N)$ is a diagonal weight matrix, we can write $\boldsymbol{\Lambda} = \boldsymbol{W}$.
- Thus, the WLS estimate can be written as:

WLS (cont.)

The WLS estimator (coefficients) is given by:

$$\hat{\boldsymbol{w}}_{WLS} = (\boldsymbol{X}^\top \boldsymbol{\Lambda} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{y} \tag{35}$$

- Since $\boldsymbol{\Lambda} = \mathrm{diag}(1/\sigma^2(\boldsymbol{x}_n)) = \mathrm{diag}(w_1, \ldots, w_N)$ is a diagonal weight matrix, we can write $\boldsymbol{\Lambda} = \boldsymbol{W}$.
- Thus, the WLS estimate can be written as:

$$\hat{\boldsymbol{w}}_{WLS} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} \tag{36}$$

WLS (cont.)

The WLS estimator (coefficients) is given by:

$$\hat{\boldsymbol{w}}_{WLS} = (\boldsymbol{X}^\top \boldsymbol{\Lambda} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{y} \tag{35}$$

- Since $\boldsymbol{\Lambda} = \text{diag}(1/\sigma^2(\boldsymbol{x}_n)) = \text{diag}(w_1, \ldots, w_N)$ is a diagonal weight matrix, we can write $\boldsymbol{\Lambda} = \boldsymbol{W}$.
- Thus, the WLS estimate can be written as:

$$\hat{\boldsymbol{w}}_{WLS} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} \tag{36}$$

- When $\boldsymbol{W}$ is not known, some common estimators are: $w_n = \frac{1}{x_n}$ or $w_n = n$
- Weights can also be iteratively estimated

Reading assignments

- **PMLI** 11.1-2
- **ESL** 3.2
- **PMLCE** 8.1

Note: Appendices follow in the next 2 dozen slides.

Hypothesis testing

Hypothesis testing

Hypothesis testing provides a framework for evaluating parameter(s) of a population with respect to a desired or known outcome.

Given that in most cases, we can only estimate these parameters, hypothesis testing allows us to determine if the estimate supports a **research hypothesis**.

The results of this testing is useful for **decision-making**.

Formulating a hypothesis test

Formulating a hypothesis test

A hypothesis is a statement regarding a parameter.
In a test, there are usually two competing hypotheses:

Formulating a hypothesis test

A hypothesis is a statement regarding a parameter.
In a test, there are usually two competing hypotheses:

- H_0:

Formulating a hypothesis test

A hypothesis is a statement regarding a parameter.
In a test, there are usually two competing hypotheses:

- H_0: the *null* hypothesis
- H_1:

Formulating a hypothesis test

A hypothesis is a statement regarding a parameter.
In a test, there are usually two competing hypotheses:

- H_0: the *null* hypothesis
- H_1: the *alternative* hypothesis (H_a is also used to denote this)

Formulating a hypothesis test

A hypothesis is a statement regarding a parameter.
In a test, there are usually two competing hypotheses:

- H_0: the *null* hypothesis
- H_1: the *alternative* hypothesis (H_a is also used to denote this)

The null hypothesis is usually framed as an equality, i.e.:

$$H_0 : \mu = \mu_0 \tag{37}$$

where μ_0 is the specified standard.
The alternative is given by

Formulating a hypothesis test

A hypothesis is a statement regarding a parameter.
In a test, there are usually two competing hypotheses:

- H_0: the *null* hypothesis
- H_1: the *alternative* hypothesis (H_a is also used to denote this)

The null hypothesis is usually framed as an equality, i.e.:

$$H_0 : \mu = \mu_0 \tag{37}$$

where μ_0 is the specified standard.
The alternative is given by

$$H_1 : \mu \neq \mu_0 \tag{38}$$

Outcomes of a hypothesis test

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it.

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**:

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**:

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**:

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject $\boldsymbol{H_0}$**: because of sufficient sample evidence in support of H_1
- **Fail to reject $\boldsymbol{H_0}$**:

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject $\boldsymbol{H_0}$**: because of sufficient sample evidence in support of H_1
- **Fail to reject $\boldsymbol{H_0}$**: because of insufficient evidence in support of H_0

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**: because of insufficient evidence in support of H_0

No truth test for the null hypothesis

The failure to reject H_0 does not mean that H_0 is true.

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**: because of insufficient evidence in support of H_0

No truth test for the null hypothesis

The failure to reject H_0 does not mean that H_0 is true.

	H_0 is true	**H_1 is true**

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**: because of insufficient evidence in support of H_0

No truth test for the null hypothesis

The failure to reject H_0 does not mean that H_0 is true.

	H_0 is true	**H_1 is true**
Fail to reject H_0		

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**: because of insufficient evidence in support of H_0

No truth test for the null hypothesis

The failure to reject H_0 does not mean that H_0 is true.

	H_0 is true	**H_1 is true**
Fail to reject H_0	Correct decision	

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**: because of insufficient evidence in support of H_0

No truth test for the null hypothesis

The failure to reject H_0 does not mean that H_0 is true.

	H_0 is true	**H_1 is true**
Fail to reject H_0	Correct decision	Type II error

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**: because of insufficient evidence in support of H_0

No truth test for the null hypothesis

The failure to reject H_0 does not mean that H_0 is true.

	H_0 is true	**H_1 is true**
Fail to reject H_0	Correct decision	Type II error
Reject H_0		

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**: because of insufficient evidence in support of H_0

No truth test for the null hypothesis

The failure to reject H_0 does not mean that H_0 is true.

	H_0 is true	**H_1 is true**
Fail to reject H_0	Correct decision	Type II error
Reject H_0	Type I error	

Outcomes of a hypothesis test

The null hypothesis is presumed unless there is sufficient evidence to discard it. The alternative hypothesis, however, is what we hope to support.

Thus there are **two outcomes** of a hypothesis test:

- **Reject H_0**: because of sufficient sample evidence in support of H_1
- **Fail to reject H_0**: because of insufficient evidence in support of H_0

No truth test for the null hypothesis

The failure to reject H_0 does not mean that H_0 is true.

	H_0 is true	**H_1 is true**
Fail to reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Type I and Type II errors

Type I and Type II errors

Type I error (false positive)

- The incorrect rejection of H_0 is a Type I error.

Type I and Type II errors

Type I error (false positive)

- The incorrect rejection of H_0 is a Type I error.
- The probability of a Type I error is the **level of significance**, α

Type I and Type II errors

Type I error (false positive)

- The incorrect rejection of H_0 is a Type I error.
- The probability of a Type I error is the **level of significance**, α

Type I and Type II errors

Type I error (false positive)

- The incorrect rejection of H_0 is a Type I error.
- The probability of a Type I error is the **level of significance**, α

Examples of Type I error

Type I and Type II errors

Type I error (false positive)

- The incorrect rejection of H_0 is a Type I error.
- The probability of a Type I error is the **level of significance**, α

Examples of Type I error

- Convicting a defendant of a crime when they are innocent
- Diagnosing a patient with a disease when in fact they do not have it

Type I and Type II errors

Type I error (false positive)

- The incorrect rejection of H_0 is a Type I error.
- The probability of a Type I error is the **level of significance**, α

Examples of Type I error

- Convicting a defendant of a crime when they are innocent
- Diagnosing a patient with a disease when in fact they do not have it (i.e. the null hypothesis is that the disease is NOT present)

Type I and Type II errors

Type I error (false positive)

- The incorrect rejection of H_0 is a Type I error.
- The probability of a Type I error is the **level of significance**, α

Examples of Type I error

- Convicting a defendant of a crime when they are innocent
- Diagnosing a patient with a disease when in fact they do not have it (i.e. the null hypothesis is that the disease is NOT present)

Type II error (false negative)

- Failure to reject H_0 when in fact H_1 is true is a Type II error.

Note: We cannot compute $\boldsymbol{w}$ except the alternative hypothesis H_1 is specified.

Type I and Type II errors

Type I error (false positive)

- The incorrect rejection of H_0 is a Type I error.
- The probability of a Type I error is the **level of significance**, α

Examples of Type I error

- Convicting a defendant of a crime when they are innocent
- Diagnosing a patient with a disease when in fact they do not have it (i.e. the null hypothesis is that the disease is NOT present)

Type II error (false negative)

- Failure to reject H_0 when in fact H_1 is true is a Type II error.
- The probability of a Type II error is denoted $\boldsymbol{w}$

Note: We cannot compute $\boldsymbol{w}$ except the alternative hypothesis H_1 is specified.

Summary of hypothesis testing approach

Summary of hypothesis testing approach

1. *Define* the **null** (H_0) and **alternative** (H_1) hypotheses

Summary of hypothesis testing approach

1. *Define* the **null** (H_0) and **alternative** (H_1) hypotheses
2. *Determine* the appropriate **test statistic** (and distribution)

Summary of hypothesis testing approach

1. *Define* the **null** (H_0) and **alternative** (H_1) hypotheses
2. *Determine* the appropriate **test statistic** (and distribution)
3. *Estimate* the test statistic from the sample data

Summary of hypothesis testing approach

1. *Define* the **null** (H_0) and **alternative** (H_1) hypotheses
2. *Determine* the appropriate **test statistic** (and distribution)
3. *Estimate* the test statistic from the sample data
4. *Specify* or *identify* the **level of significance** (α)

Summary of hypothesis testing approach

1. *Define* the **null** (H_0) and **alternative** (H_1) hypotheses
2. *Determine* the appropriate **test statistic** (and distribution)
3. *Estimate* the test statistic from the sample data
4. *Specify* or *identify* the **level of significance** (α)
5. *Define* the **region of rejection/critical region** of the null hypothesis by choosing the **critical value**.

Summary of hypothesis testing approach

1. *Define* the **null** (H_0) and **alternative** (H_1) hypotheses
2. *Determine* the appropriate **test statistic** (and distribution)
3. *Estimate* the test statistic from the sample data
4. *Specify* or *identify* the **level of significance** (α)
5. *Define* the **region of rejection/critical region** of the null hypothesis by choosing the **critical value**.
6. *Decide.* If the test statistic is in the critical region, reject H_0. If not, do not reject H_0 (fail to reject it)

One-sided tests

One-sided tests

Case A: upper tail

One-sided tests

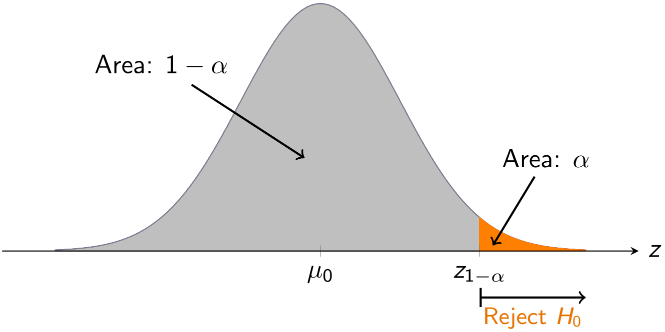
Case A: upper tail

- $H_0 : \mu = \mu_0$
- $H_1 : \mu > \mu_0$

One-sided tests

Case A: upper tail

- $H_0 : \mu = \mu_0$
- $H_1 : \mu > \mu_0$



One-sided tests (cont.)

One-sided tests (cont.)

Case B: lower tail

One-sided tests (cont.)

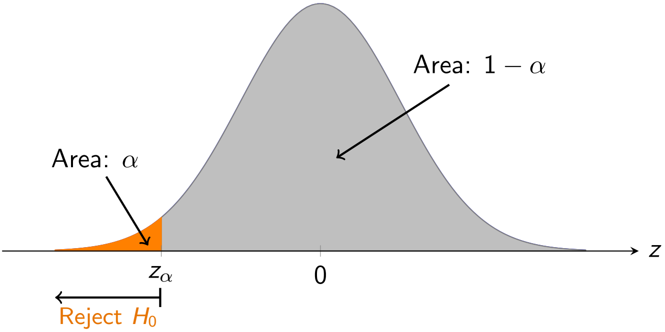
Case B: lower tail

- $H_0 : \mu = \mu_0$
- $H_1 : \mu < \mu_0$

One-sided tests (cont.)

Case B: lower tail

- $H_0 : \mu = \mu_0$
- $H_1 : \mu < \mu_0$



Two-sided tests

Two-sided tests

Case C: both tails

Two-sided tests

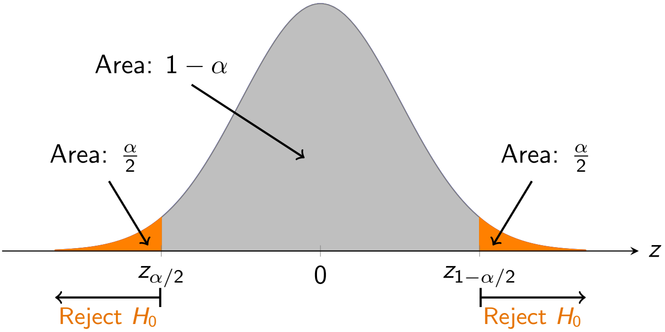
Case C: both tails

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$

Two-sided tests

Case C: both tails

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$



Hypothesis testing for known and unknown variance

Hypothesis testing for known and unknown variance

In the case where the **sample mean** as the test statistic:

1. If the variance is known, then we use the normal distribution to find the probability of the standardized Z-statistic $\frac{\overline{X}-\mu}{\boldsymbol{\sigma}/\sqrt{n}}$ and compare it to the appropriate critical value to test our hypotheses

Hypothesis testing for known and unknown variance

In the case where the **sample mean** as the test statistic:

1. If the variance is known, then we use the normal distribution to find the probability of the standardized Z-statistic $\frac{\overline{X}-\mu}{\boldsymbol{\sigma}/\sqrt{n}}$ and compare it to the appropriate critical value to test our hypotheses
2. If the variance is unknown, we use the t-distribution ($N-1$ degrees of freedom) to find the probability of the standardized T-statistic $\frac{\overline{X}-\mu}{\boldsymbol{s}/\sqrt{n}}$ and compare it to the appropriate critical value to test our hypotheses

Two-tailed tests: unknown variance

Two-tailed tests: unknown variance

Example 2: Golf ball production

A premium golf ball production line must produce all of its balls to 1.615 ounces in order to get the top rating (and therefore the top dollar). Samples are drawn hourly and checked. If the production line gets out of sync with a statistical significance of more than 1%, it must be shut down andrepaired. This hour's sample of 18 balls has a mean of 1.611 oz and a standard deviation of 0.065 oz. Do you shut down the line?

Two-tailed tests: unknown variance

Example 2: Golf ball production

A premium golf ball production line must produce all of its balls to 1.615 ounces in order to get the top rating (and therefore the top dollar). Samples are drawn hourly and checked. If the production line gets out of sync with a statistical significance of more than 1%, it must be shut down andrepaired. This hour's sample of 18 balls has a mean of 1.611 oz and a standard deviation of 0.065 oz. Do you shut down the line?

Step 1 Formulate hypotheses:

Two-tailed tests: unknown variance

Example 2: Golf ball production

A premium golf ball production line must produce all of its balls to 1.615 ounces in order to get the top rating (and therefore the top dollar). Samples are drawn hourly and checked. If the production line gets out of sync with a statistical significance of more than 1%, it must be shut down andrepaired. This hour's sample of 18 balls has a mean of 1.611 oz and a standard deviation of 0.065 oz. Do you shut down the line?

Step 1 Formulate hypotheses:

$$H_0 : \mu = 1.615$$
$$H_1 : \mu \neq 1.615$$

Two-tailed tests: unknown variance

Example 2: Golf ball production

A premium golf ball production line must produce all of its balls to 1.615 ounces in order to get the top rating (and therefore the top dollar). Samples are drawn hourly and checked. If the production line gets out of sync with a statistical significance of more than 1%, it must be shut down andrepaired. This hour's sample of 18 balls has a mean of 1.611 oz and a standard deviation of 0.065 oz. Do you shut down the line?

Step 1 Formulate hypotheses:

$$H_0 : \mu = 1.615$$
$$H_1 : \mu \neq 1.615$$

Two-tailed tests: unknown variance

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 2 Compute T-statistic:

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 2 Compute T-statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{1.611 - 1.615}{0.065/\sqrt{18}} = -0.261$$

Step 3 $\alpha = 1\% = 0.01$.
Given that this is a two-tailed test, we have two critical regions with areas: $\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$.
The lower tail is bounded by $t_{0.005}$ and the upper tail by $t_{1-0.005} = t_{0.995}$.

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 2 Compute T-statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

$$-0.261$$

Step 3 $\alpha = 1\% = 0.01$.
Given that this is a two-tailed test, we have two critical regions with areas: $\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$.
The lower tail is bounded by $t_{0.005}$ and the upper tail by $t_{1-0.005} = t_{0.995}$.

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 2 Compute T-statistic:

$$
\begin{aligned}
t &= \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \\
&= \frac{1.611 - 1.615}{0.065/\sqrt{18}} = -0.261
\end{aligned}
$$

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 2 Compute T-statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$
$$= \frac{1.611 - 1.615}{0.065/\sqrt{18}} = -0.261$$

Step 3 $\alpha = 1\% = 0.01.$

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 2 Compute T-statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{1.611 - 1.615}{0.065/\sqrt{18}} = -0.261$$

Step 3 $\alpha = 1\% = 0.01$.
Given that this is a two-tailed test, we have two critical regions with areas: $\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$.

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 2 Compute T-statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{1.611 - 1.615}{0.065/\sqrt{18}} = -0.261$$

Step 3 $\alpha = 1\% = 0.01$.
Given that this is a two-tailed test, we have two critical regions with areas: $\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$.
The lower tail is bounded by $t_{0.005}$ and the upper tail by $t_{1-0.005} = t_{0.995}$.

Two-tailed tests: unknown variance

Two-tailed tests: unknown variance

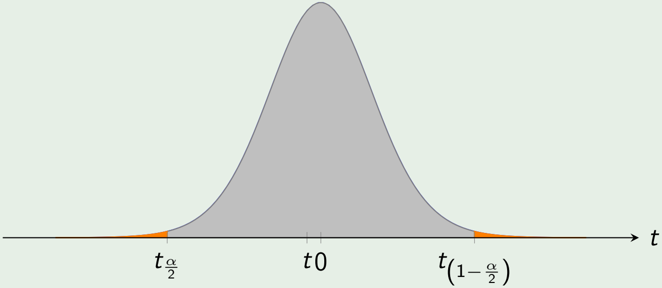
Example 2: Golf ball production

Step 4 The critical values are $t_{0.005} = -2.8982$ and $t_{0.95} = 2.8982$ (d.o.f. $= 17$).

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 4 The critical values are $t_{0.005} = -2.8982$ and $t_{0.95} = 2.8982$ (d.o.f. $= 17$).

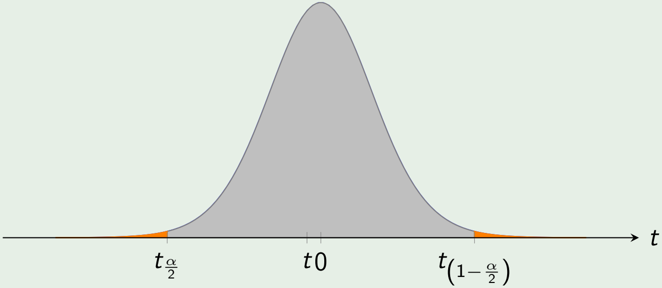


Step 5 We that the test statistic is within the region of nonrejection:

Two-tailed tests: unknown variance

Example 2: Golf ball production

Step 4 The critical values are $t_{0.005} = -2.8982$ and $t_{0.95} = 2.8982$ (d.o.f. $= 17$).



Step 5 We that the test statistic is within the region of nonrejection:

$$t_{\frac{\alpha}{2}} = -2.8982 < t = -0.261 < t_{\left(1-\frac{\alpha}{2}\right)} = 2.8982$$

Example 2: Golf ball production

Step 6 Thus, we **fail to reject** the null hypothesis.

Example 2: Golf ball production

Step 6 Thus, we **fail to reject** the null hypothesis.
In real terms, this means that the sample was within the bounds of what would be acceptable if the population mean were 1.615 oz. Therefore, we would not stop the production line.

Definition and usefulness of p-value

Definition and usefulness of *p*-value

Definition

The *p*-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.

Definition and usefulness of p-value

Definition

The p-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.
Equivalently,

Definition and usefulness of p-value

Definition

The p-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.
Equivalently, this is the minimum probability of a Type I error.

- Provides more information about the strength of a test

Definition and usefulness of p-value

Definition

The p-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.
Equivalently, this is the minimum probability of a Type I error.

- Provides more information about the strength of a test
- Indicates the smallest level at which the data is significant

Definition and usefulness of p-value

Definition

The p-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.
Equivalently, this is the minimum probability of a Type I error.

- Provides more information about the strength of a test
- Indicates the smallest level at which the data is significant
- Can be compared with α irrespective of which type of test was used

Definition and usefulness of p-value

Definition

The p-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.
Equivalently, this is the minimum probability of a Type I error.

- Provides more information about the strength of a test
- Indicates the smallest level at which the data is significant
- Can be compared with α irrespective of which type of test was used

Definition and usefulness of p-value

Definition

The p-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.
Equivalently, this is the minimum probability of a Type I error.

- Provides more information about the strength of a test
- Indicates the smallest level at which the data is significant
- Can be compared with α irrespective of which type of test was used

Alternative definition

Definition and usefulness of p-value

Definition

The p-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.
Equivalently, this is the minimum probability of a Type I error.

- Provides more information about the strength of a test
- Indicates the smallest level at which the data is significant
- Can be compared with α irrespective of which type of test was used

Alternative definition

The p-value is the probability of obtaining a test statistic value at least as contradictory to H_0 as the value that actually resulted.

Definition and usefulness of p-value

Definition

The p-value is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given dataset.
Equivalently, this is the minimum probability of a Type I error.

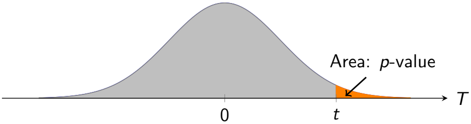
- Provides more information about the strength of a test
- Indicates the smallest level at which the data is significant
- Can be compared with α irrespective of which type of test was used

Alternative definition

The p-value is the probability of obtaining a test statistic value at least as contradictory to H_0 as the value that actually resulted.
The smaller the p-value, the more contradictory is the data to H_0.

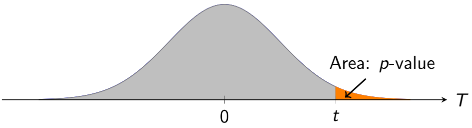
p-value for z tests

p-value for z tests



p-value: area in upper tail

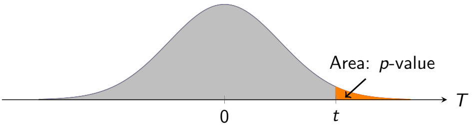
p-value for z tests

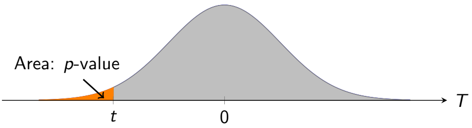


p-value: area in upper tail

$$p = 1 - F_T(t_\nu) \qquad (39)$$

p-value for z tests



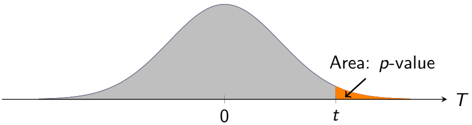


p-value: area in upper tail

$$p = 1 - F_T(t_\nu) \tag{39}$$

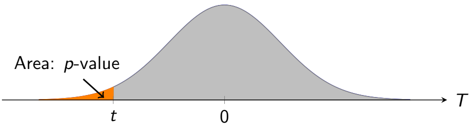
p-value: area in lower tail

p-value for z tests



p-value: area in upper tail

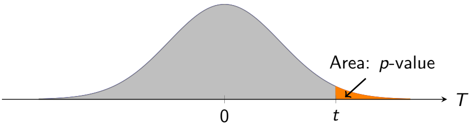
$$p = 1 - F_T(t_\nu) \tag{39}$$



p-value: area in lower tail

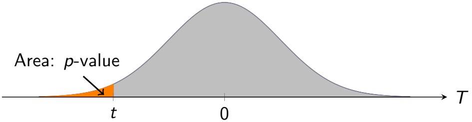
$$p = F_T(t_\nu) \tag{40}$$

p-value for z tests



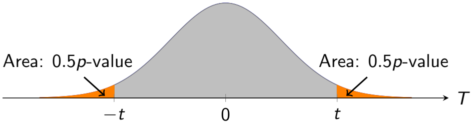
p-value: area in upper tail

$$p = 1 - F_T(t_\nu) \qquad (39)$$



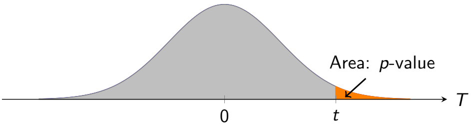
p-value: area in lower tail

$$p = F_T(t_\nu) \qquad (40)$$



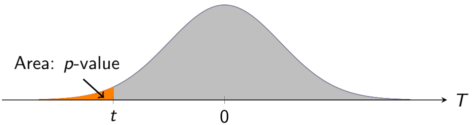
p-value: sum of area in both tails

p-value for z tests



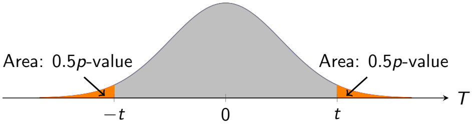
p-value: area in upper tail

$$p = 1 - F_T(t_\nu) \quad (39)$$



p-value: area in lower tail

$$p = F_T(t_\nu) \quad (40)$$



p-value: sum of area in both tails

$$p = 2(1 - F_T|t_\nu|)) \quad (41)$$

Estimates and standard error

From the model equation $Y = w_0 + w_1 X + \epsilon$, we see that Y is a random variable with true variance σ^2.

Estimates and standard error

From the model equation $Y = w_0 + w_1 X + \epsilon$, we see that Y is a random variable with true variance σ^2.
The point estimate of the true mean of Y is unbiased:

Estimates and standard error

From the model equation $Y = w_0 + w_1 X + \epsilon$, we see that Y is a random variable with true variance σ^2.
The point estimate of the true mean of Y is unbiased:

$$\hat{\mu} = \overline{y} = \frac{1}{n} \sum y_n \tag{42}$$

$$\mathbb{E}(\hat{\mu}) = \mu \tag{43}$$

Based on the Gauss-Markov theorem, the least squares estimates are also unbiased:

Estimates and standard error

From the model equation $Y = w_0 + w_1 X + \epsilon$, we see that Y is a random variable with true variance σ^2.
The point estimate of the true mean of Y is unbiased:

$$\hat{\mu} = \overline{y} = \frac{1}{n} \sum y_n \tag{42}$$

$$\mathbb{E}(\hat{\mu}) = \mu \tag{43}$$

Based on the Gauss-Markov theorem, the least squares estimates are also unbiased:

$$E\left(\hat{w}_0\right) = w_0 \tag{44}$$

$$E\left(\hat{w}_1\right) = w_1 \tag{45}$$

Standard errors

The population of Y is unknown, but we can estimate its parameters from a sample of n observations.

Standard errors

The population of Y is unknown, but we can estimate its parameters from a sample of n observations.

By the Central Limit Thereom (CLT), $\overline{Y}$ is normally distributed with mean μ and variance σ^2/n.

Standard errors

The population of Y is unknown, but we can estimate its parameters from a sample of n observations.

By the Central Limit Thereom (CLT), $\overline{Y}$ is normally distributed with mean μ and variance σ^2/n.

Standard error

The standard error of mean estimate is given by:

$$S\mathbb{E}(\hat{\mu}) = Var(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \tag{46}$$

Where σ is not known, it can be estimated from the data:

Standard errors

The population of Y is unknown, but we can estimate its parameters from a sample of n observations.

By the Central Limit Thereom (CLT), $\overline{Y}$ is normally distributed with mean μ and variance σ^2/n.

Standard error

The standard error of mean estimate is given by:

$$S\mathbb{E}(\hat{\mu}) = Var(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \tag{46}$$

Where σ is not known, it can be estimated from the data:

$$\hat{\sigma}^2 =$$

Standard errors

The population of Y is unknown, but we can estimate its parameters from a sample of n observations.

By the Central Limit Thereom (CLT), $\overline{Y}$ is normally distributed with mean μ and variance σ^2/n.

Standard error

The standard error of mean estimate is given by:

$$S\mathbb{E}(\hat{\mu}) = Var(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \tag{46}$$

Where σ is not known, it can be estimated from the data:

$$\hat{\sigma}^2 = \frac{\sum (y - \hat{y})^2}{N - 2} =$$

Standard errors

The population of Y is unknown, but we can estimate its parameters from a sample of n observations.

By the Central Limit Thereom (CLT), $\overline{Y}$ is normally distributed with mean μ and variance σ^2/n.

Standard error

The standard error of mean estimate is given by:

$$S\mathbb{E}(\hat{\mu}) = Var(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \tag{46}$$

Where σ is not known, it can be estimated from the data:

$$\hat{\sigma}^2 = \frac{\sum (y - \hat{y})^2}{N-2} = \frac{RSS}{N-2} = RSE^2 \tag{47}$$

Model accuracy

Model accuracy

We use the standard errors to evaluate the accuracy of the coefficient estimates.

Model accuracy

We use the standard errors to evaluate the accuracy of the coefficient estimates.

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{48}$$

$$SE(\hat{w}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{N} (x_n - \overline{x})^2} \right] \tag{49}$$

$$SE(\hat{w}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N} (x_n - \overline{x})^2} \tag{50}$$

Model accuracy

We use the standard errors to evaluate the accuracy of the coefficient estimates.

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{48}$$

$$SE(\hat{w}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_n - \overline{x})^2} \tag{50}$$

Model accuracy

We use the standard errors to evaluate the accuracy of the coefficient estimates.

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{48}$$

$$SE(\hat{w}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{N} (x_n - \overline{x})^2} \right] \tag{49}$$

$$SE(\hat{w}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N} (x_n - \overline{x})^2} \tag{50}$$

Model accuracy

We use the standard errors to evaluate the accuracy of the coefficient estimates.

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{48}$$

$$SE(\hat{w}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{N}(x_n - \overline{x})^2} \right] \tag{49}$$

$$SE(\hat{w}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_n - \overline{x})^2} \tag{50}$$

Hypothesis testing on the slope parameter

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Null hypothesis:

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Null hypothesis:

$$H_0 : w_1 = 0 \tag{51}$$

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Null hypothesis:

$$H_0 : w_1 = 0 \tag{51}$$

Alternative hypothesis:

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Null hypothesis:

$$H_0 : w_1 = 0 \tag{51}$$

Alternative hypothesis:

$$H_1 : w_1 \neq 0 \tag{52}$$

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Null hypothesis:

$$H_0 : w_1 = 0 \tag{51}$$

Alternative hypothesis:

$$H_1 : w_1 \neq 0 \tag{52}$$

Test statistic:

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Null hypothesis:

$$H_0 : w_1 = 0 \tag{51}$$

Alternative hypothesis:

$$H_1 : w_1 \neq 0 \tag{52}$$

Test statistic:

$$t = \frac{\hat{w}_1}{SE\left(\hat{w}_1\right)} \tag{53}$$

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Null hypothesis:

$$H_0 : w_1 = 0 \tag{51}$$

Alternative hypothesis:

$$H_1 : w_1 \neq 0 \tag{52}$$

Test statistic:

$$t = \frac{\hat{w}_1}{SE\left(\hat{w}_1\right)} \tag{53}$$

Decision

Hypothesis testing on the slope parameter

The so-called **model utility test** is used to determine if any relationship exists between X and Y.

Null hypothesis:

$$H_0 : w_1 = 0 \tag{51}$$

Alternative hypothesis:

$$H_1 : w_1 \neq 0 \tag{52}$$

Test statistic:

$$t = \frac{\hat{w}_1}{SE\left(\hat{w}_1\right)} \tag{53}$$

Decision

Reject null hypothesis if $t \leq t_{\alpha/2,N-2}$ or $t \geq t_{1-\alpha/2,N-2}$.

Regression and the F test

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

- $t^2 = f$

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

- $t^2 = f$
- $t^2_{1-\alpha/2,N-2} = F_{1-\alpha,1,N-2}$ (critical value)

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

- $t^2 = f$
- $t^2_{1-\alpha/2,N-2} = F_{1-\alpha,1,N-2}$ (critical value)

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

- $t^2 = f$
- $t^2_{1-\alpha/2,N-2} = F_{1-\alpha,1,N-2}$ (critical value)

Analysis of variance (ANOVA) table for simple linear regression:

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

- $t^2 = f$
- $t^2_{1-\alpha/2,N-2} = F_{1-\alpha,1,N-2}$ (critical value)

Analysis of variance (ANOVA) table for simple linear regression:

Source of variation	**d.o.f.**	**Sum of Squares**	**Mean Square**	$\boldsymbol{F}$

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

- $t^2 = f$
- $t^2_{1-\alpha/2,N-2} = F_{1-\alpha,1,N-2}$ (critical value)

Analysis of variance (ANOVA) table for simple linear regression:

Source of variation	**d.o.f.**	**Sum of Squares**	**Mean Square**	F
Regression	1	$TSS - RSS$	$TSS - RSS$	$\frac{(TSS-RSS)}{RSS/(N-2)}$

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

- $t^2 = f$
- $t^2_{1-\alpha/2,N-2} = F_{1-\alpha,1,N-2}$ (critical value)

Analysis of variance (ANOVA) table for simple linear regression:

Source of variation	**d.o.f.**	**Sum of Squares**	**Mean Square**	F
Regression	1	$TSS - RSS$	$TSS - RSS$	$\frac{(TSS-RSS)}{RSS/(N-2)}$
Error	$N-2$	RSS	$RSE^2 = \frac{RSS}{N-2}$	

Regression and the F test

The right-tail F test gives the exact same result as the model utility t test because:

- $t^2 = f$
- $t^2_{1-\alpha/2,N-2} = F_{1-\alpha,1,N-2}$ (critical value)

Analysis of variance (ANOVA) table for simple linear regression:

Source of variation	**d.o.f.**	**Sum of Squares**	**Mean Square**	F
Regression	1	$TSS - RSS$	$TSS - RSS$	$\frac{(TSS-RSS)}{RSS/(N-2)}$
Error	$N-2$	RSS	$RSE^2 = \frac{RSS}{N-2}$	
Total	$N-1$	TSS		

Confidence intervals

The $100(1-\alpha)\%$ confidence intervals of the coefficient estimates are given by:

Confidence intervals

The $100(1-\alpha)\%$ confidence intervals of the coefficient estimates are given by:

$$\langle w_0 \rangle_{1-\alpha} = \hat{w}_0 \pm t_{\left(1-\frac{\alpha}{2}, N-2\right)} SE\left(\hat{w}_0\right) \tag{54}$$

$$\langle w_1 \rangle_{1-\alpha} = \hat{w}_1 \pm t_{\left(1-\frac{\alpha}{2}, N-2\right)} SE\left(\hat{w}_1\right) \tag{55}$$

R^2 and correlation coefficient

R^2 and correlation coefficient

Recall the **sample correlation coefficient**:

R^2 and correlation coefficient

Recall the **sample correlation coefficient**:

$$r = \frac{\sum(x_n - \overline{x})(y_n - \overline{y})}{\sqrt{\sum(x_n - \overline{x})^2 \sum(y_n - \overline{y})^2}} \tag{56}$$

R^2 and correlation coefficient

Recall the **sample correlation coefficient**:

$$r = \frac{\sum(x_n - \overline{x})(y_n - \overline{y})}{\sqrt{\sum(x_n - \overline{x})^2 \sum(y_n - \overline{y})^2}} \tag{56}$$

In the univariate case, we can show that $R^2 = r^2$.

R^2 and correlation coefficient

Recall the **sample correlation coefficient**:

$$r = \frac{\sum(x_n - \overline{x})(y_n - \overline{y})}{\sqrt{\sum(x_n - \overline{x})^2 \sum(y_n - \overline{y})^2}} \tag{56}$$

In the univariate case, we can show that $R^2 = r^2$.

Thus, R^2 is also a measure of the linear relationship between X and Y.