Multiple Linear Regression

GENOME 560, Spring 2026

Saori Sakaue (sakaue@uw.edu)

Section 1

Recap: Simple Linear Regression

A Linear Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- $\beta_0 =$ y-intercept
- $\beta_1 =$ slope
- $\epsilon =$ random error

Least Square Solutions for Linear Regression

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Deriving Least Square Solutions

Set partial derivatives to zero:

$$\frac{\partial f}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial f}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Deriving Least Square Solutions (continued)

From the first equation:

$$n\beta_0 = \sum y_i - \beta_1 \sum x_i \implies \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Substituting into the second equation and solving for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

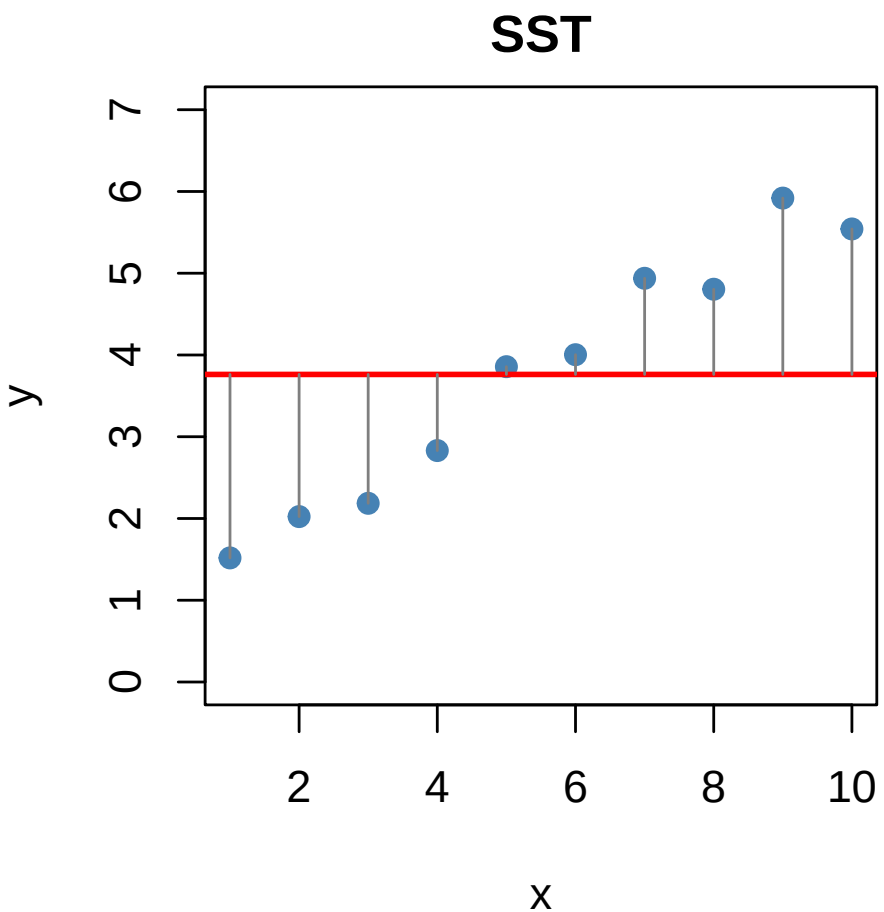
Sum of Squared Errors

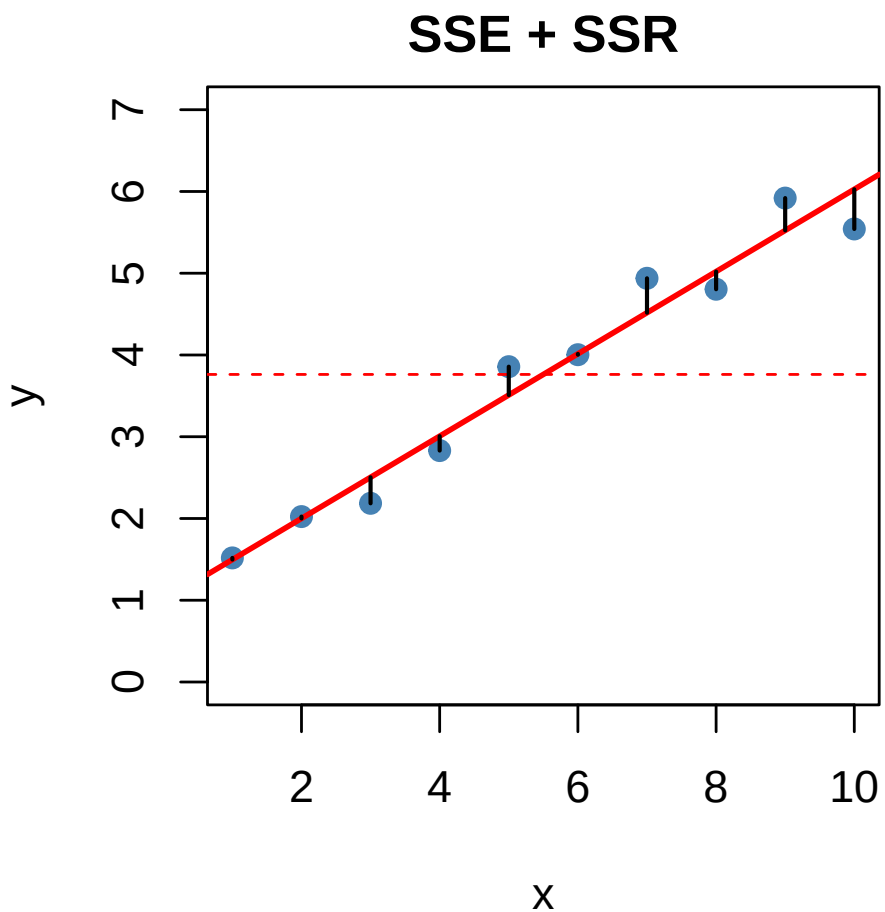
How different is the model prediction from the observed value?

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

Decomposing Variance

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{SSR (explained)}} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{SSE (not explained)}}$$





R^2 and Correlation

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \qquad r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

For simple linear regression: $R^2 = r^2$

Correlation Coefficient r

Used to describe the degree of **sample correlation**

- $-1 \leq r \leq 1$
- $r > 0$: positive correlation; $r < 0$: negative correlation
- $r = 0$: no linear correlation

Section 2

Recap: Matrix Operations

Sum of Matrices

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix}$$

Matrices must have the same dimensions.

Product of a Scalar and a Matrix

$$c\mathbf{A} = \begin{pmatrix} ca_{11} & ca_{12} \\ ca_{21} & ca_{22} \end{pmatrix}$$

Product of Two Matrices

$$\mathbf{A}_{m\times n} \cdot \mathbf{B}_{n\times p} = \mathbf{C}_{m\times p}$$

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

The number of columns of $\mathbf{A}$ must equal the number of rows of $\mathbf{B}$.

Product of Two Matrices — Example

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \cdot 5 + 2 \cdot 7 & 1 \cdot 6 + 2 \cdot 8 \\ 3 \cdot 5 + 4 \cdot 7 & 3 \cdot 6 + 4 \cdot 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$

Transpose and Product

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

Inverse

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Not all matrices have inverses. A matrix is **invertible** (non-singular) if $\det(\mathbf{A}) \neq 0$.

Section 3

Multiple Linear Regression

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- p predictors (independent variables)
- β_j = effect of X_j on Y, holding all other predictors constant

Key Assumptions of Multiple Regression

1. **Linearity:** $E[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
2. **Independence:** observations are independent
3. **Homoscedasticity:** $\text{Var}(\epsilon_i) = \sigma^2$ for all i
4. **Normality:** $\epsilon_i \sim N(0, \sigma^2)$

Least Square Solution in Matrix Form

In matrix notation:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}$ is $n \times (p+1)$, β is $(p+1) \times 1$.

Least Square Solution

Minimize:

$$f(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

Solving by Partial Differentiation

$$\frac{\partial f}{\partial \beta} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta = 0$$

Matrix Differentiation May Be Easier

Setting the derivative to zero:

$$\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{Y}$$

Normal equation — solution:

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

Matrix Differentiation (continued)

Expand $f(\beta)$:

$$f = \mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

Using matrix calculus rules:

$$\frac{\partial}{\partial\beta}(\beta^T\mathbf{a}) = \mathbf{a}, \quad \frac{\partial}{\partial\beta}(\beta^T\mathbf{A}\beta) = 2\mathbf{A}\beta$$

Simple Linear Regression in Matrix Form

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Simple Linear Regression Solution

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

This gives the same $\hat{\beta}_0$ and $\hat{\beta}_1$ as before.

Properties of $\hat{\beta}$ for Hypothesis Testing

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

For individual coefficients:

$$t_j = \frac{\hat{\beta}_j}{\mathsf{SE}(\hat{\beta}_j)} \sim t_{n-p-1}$$

under $H_0 : \beta_j = 0$.

Multiple Correlation Coefficient

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Adjusted R^2 penalizes for number of predictors:

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

Section 4

Potential Issues in Regression

In the Real World...

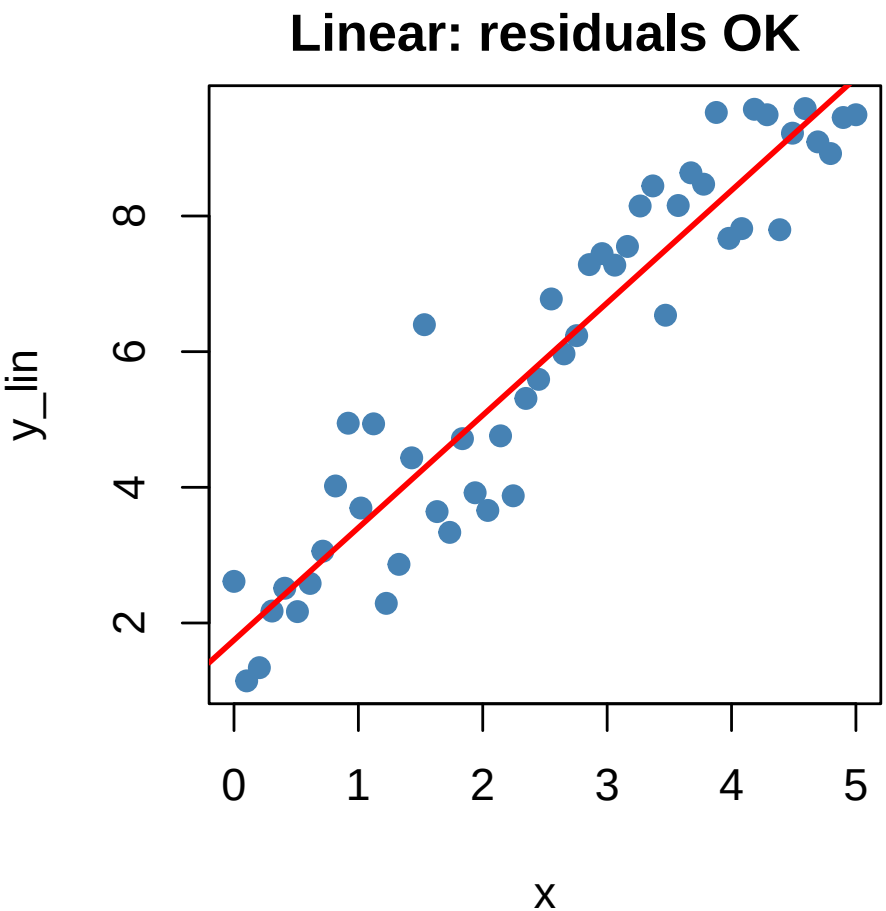
Real data often violates assumptions. We need to check for:

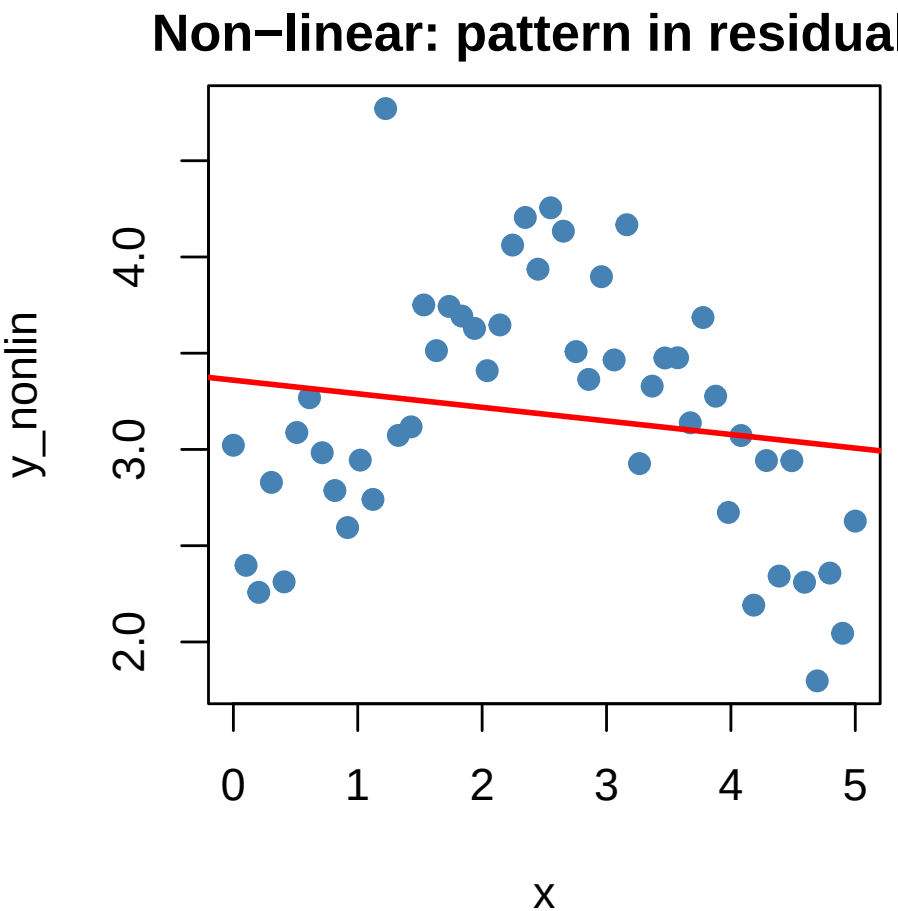
- Non-linearity
- Correlation of error terms
- Non-constant variance (heteroscedasticity)
- Collinearity

Potential Issues in Regression

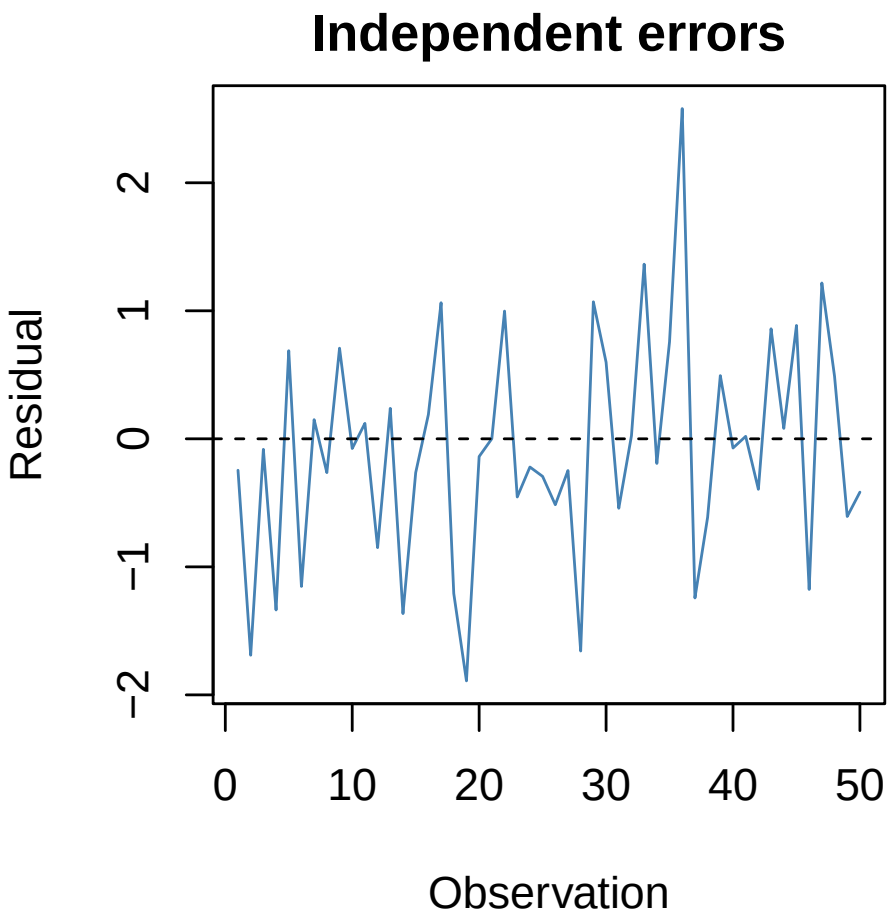
1. Non-linearity
2. Correlation of error terms
3. Non-constant variance
4. Collinearity

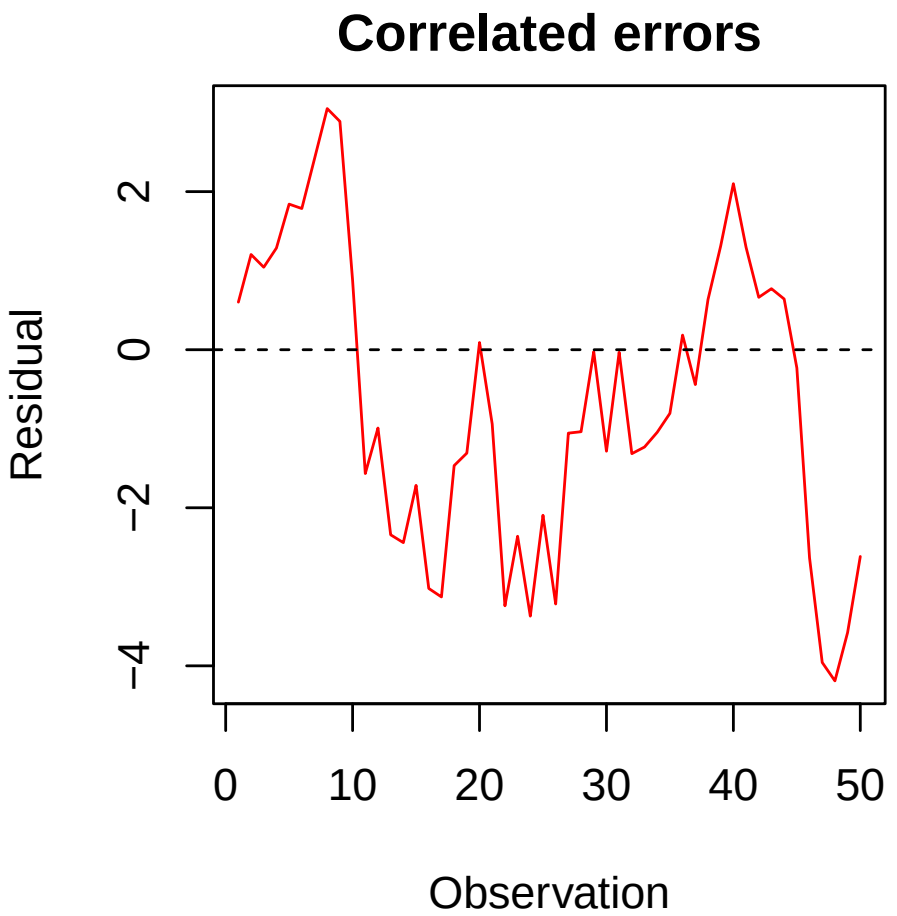
Non-linearity



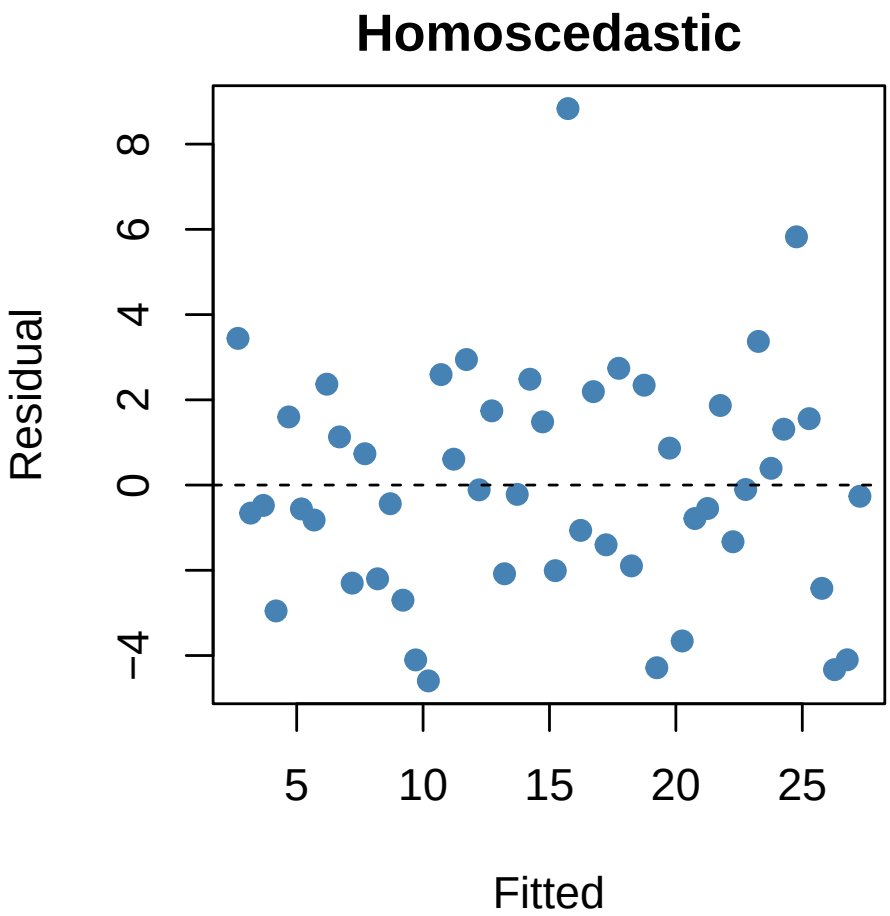


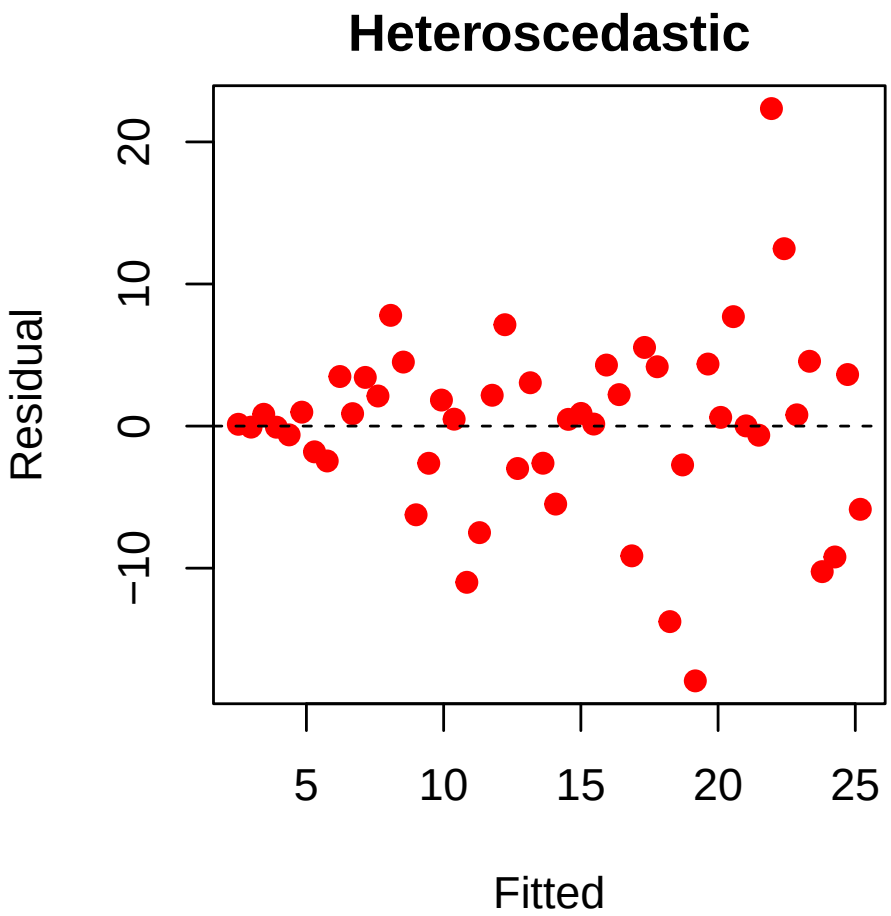
Correlation of Error Terms





Non-constant Variance (Heteroscedasticity)



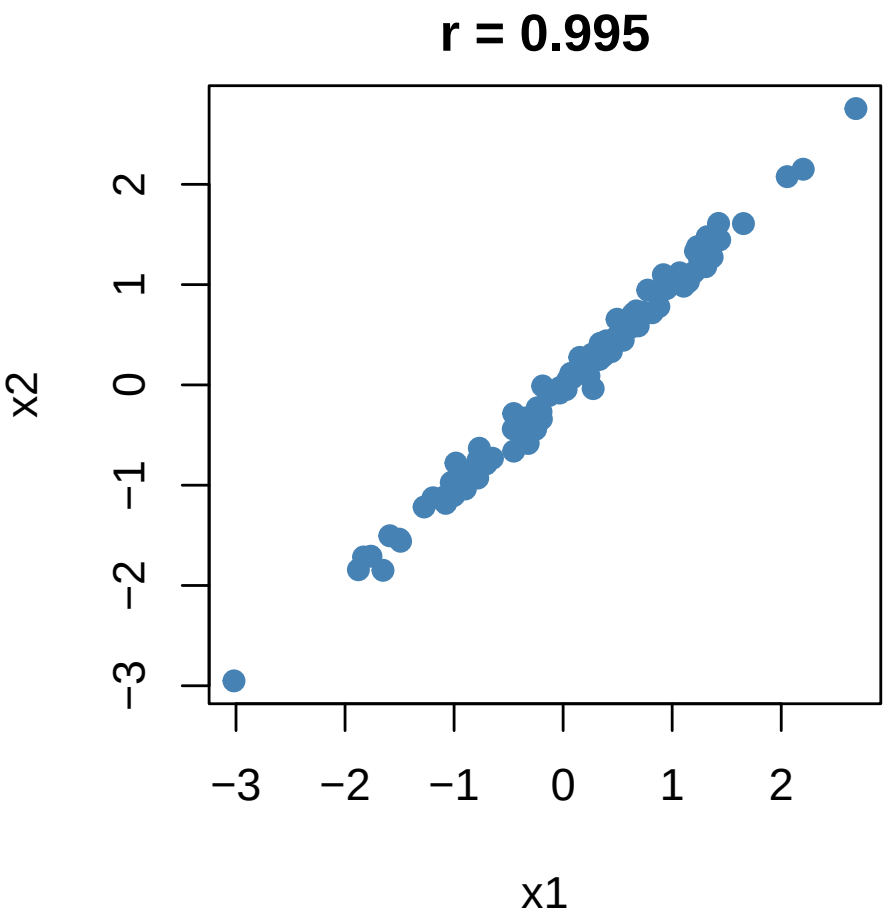


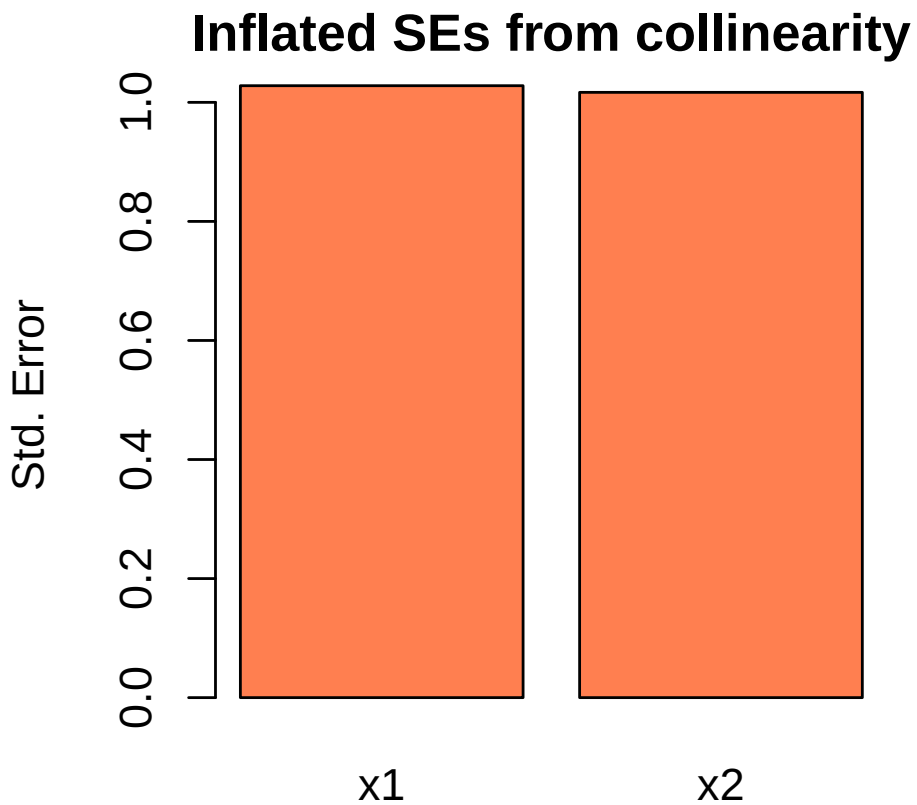
Collinearity

Two or more predictor variables are closely related to each other.

- Difficult to separate out individual effects of collinear variables on the response
- Increases variance of coefficient estimates $\rightarrow$ reduced power
- Can cause unstable estimates

Collinearity — Example





Collinearity — Detection

Variance Inflation Factor (VIF):

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 from regressing X_j on all other predictors.

- VIF $= 1$: no collinearity
- VIF > 5 or 10: problematic collinearity

Categorical Independent Variables

Use **dummy variables** (indicator variables):

Example: ABO blood type {A, B, O, AB}

	D_A	D_B	D_{AB}
Type A	1	0	0
Type B	0	1	0
Type AB	0	0	1
Type O	0	0	0

k categories $\rightarrow k-1$ dummy variables (reference category omitted to avoid collinearity)