Methods of Data Analysis I

Derek Li

Contents

1 Review

1.1 Expectation

- $\mathbb{E}[a] = a, a \in \mathbb{R}$.
- $\mathbb{E}[aY] = a\mathbb{E}[Y]$.
- $\mathbb{E}[X \pm Y] = \mathbb{E}[x] \pm \mathbb{E}[Y]$.
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are independent.
- Tower rule: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$.

1.2 Variance and Covariance

- $\text{Var}[a] = 0, a \in \mathbb{R}$.
- $\text{Var}[aY] = a^2\text{Var}[Y]$.
- $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
- $\text{Cov}(Y, Y) = \text{Var}[Y]$.
- $\text{Var}[Y] = \text{Var}[\mathbb{E}[Y|X]] + \mathbb{E}[\text{Var}[Y|X]]$.
- $\text{Var}[X \pm Y] = \text{Var}[X] + \text{Var}[Y] \pm 2\text{Cov}(X, Y)$.
- $\text{Cov}(X, Y) = 0$ if X and Y are independent.
- $\text{Cov}(aX + bY, cU + dW) = ac\text{Cov}(X, U) + ad\text{Cov}(X, W) + bc\text{Cov}(Y, U) + bd\text{Cov}(Y, W)$.

1.3 Correlation

If X and Y are random variables, a symmetric measure of the direction and strength of the linear dependence between them is their correlation

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}}.$$

1.4 Distributions

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
- Let $U = Z^2$, then $U \sim \chi^2_{(1)}$.
- If Z and $X \sim \chi^2_{(m)}$ are independent, then $\frac{Z}{\sqrt{X/m}} \sim t_{(m)}$.
- If $X \sim \chi^2_{(m)}, Y \sim \chi^2_{(n)}$ are independent, then $\frac{X/m}{Y/n} \sim F_{(m,n)}$.
- $t_{(m)} \xrightarrow{D} Z$, as $m \to \infty$.

1.4.1 Bivariate Normal Distribution

X and Y are jointly normally distributed is their joint density function is

$$f(x,y) = \frac{e^{-\frac{Q}{2}}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}},$$

where

$$Q = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right].$$

Two marginal distributions are

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \text{ and } Y \sim \mathcal{N}(\mu_y, \sigma_y^2).$$

The conditional distribution of Y given $X = x$ is

$$Y|X = x \sim \mathcal{N}\left(\mu_y + \rho\sigma_y\left(\frac{x-\mu_x}{\sigma_x}\right), (1-\rho^2)\sigma_y^2\right).$$

Theorem 1.1. If X and Y are jointly normally distributed, then a zero covariance between X and Y implies that they are statistically independent.

1.5 Matrix

1.5.1 Random Vectors and Matrices

Suppose $(X_1, \cdots, X_n)$ and $(Y_1, \cdots, Y_n)$ are sets of random variables, the random vectors

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

are $(n \times 1)$ vectors and $\mathbf{X}^T = (X_1, \cdots, X_n), \mathbf{Y}^T = (Y_1, \cdots, Y_n)$.

1.5.2 Expectation of Random Vectors

The mean of a random variable is a vector of means, i.e.,

$$\mathbb{E}(\mathbf{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}.$$

If $a \in \mathbb{R}$, then

$$a\mathbf{X} = \begin{pmatrix} aX_1 \\ \vdots \\ aX_n \end{pmatrix} \text{ and } \mathbb{E}[a\mathbf{X}] = a\mathbb{E}[\mathbf{X}].$$

If $\mathbf{a} \in \mathbb{R}^n$, then

$$\mathbf{a}^T\mathbf{X} = (a_1, \cdots, a_n)\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \sum_{i=1}^{n} a_i X_i$$

and therefore $\mathbb{E}[\mathbf{a}^T\mathbf{X}]$ is a scalar.

1.5.3 Expectation of Random Matrices

If $\mathbf{A} \in \mathbb{R}^{m\times n}$, then

$$\mathbf{AX} = \begin{pmatrix} \sum\limits_{i=1}^{n} a_{1i}X_i \\ \vdots \\ \sum\limits_{i=1}^{n} a_{mi}X_i \end{pmatrix} \text{ and } \mathbb{E}[\mathbf{AX}] = \mathbf{A}\mathbb{E}[\mathbf{X}].$$

If $a, b \in \mathbb{R}$, then $\mathbb{E}[a\mathbf{X} + b\mathbf{Y}] = a\mathbb{E}[\mathbf{X}] + b\mathbb{E}[\mathbf{Y}]$.

1.5.4 Special Vectors and Matrices

Definition 1.1 (Symmetric Matrix)**.** If $\mathbf{A} = \mathbf{A}^T$, then $\mathbf{A}$ is a symmetric matrix.

Example 1.1. If

$$\mathbf{X}_{n\times 2} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

then

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum X_n \\ \sum X_n & \sum X_n^2 \end{pmatrix} = n \begin{pmatrix} 1 & \overline{X} \\ \overline{X} & \frac{1}{n}\sum X_i^2 \end{pmatrix}$$

is symmetric. Besides, we call $\mathbf{X}$ as design matrix in SLR.

Property 1.1. A symmetric matrix is square.

Definition 1.2 (Diagonal Matrix)**.** A diagonal matrix is a square matrix whose off-diagonal elements are all zero.

Definition 1.3 (Identity Matrix)**.** The identity matrix of unit matrix, denoted by $\mathbf{I}$, is a diagonal matrix whose elements on the main diagonal are all one.

Property 1.2. For any $q \times q$ matrix $\mathbf{A}$, we have $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$.

Definition 1.4 (Unit Vector)**.** A unit vector, denoted by $\mathbf{1}$, is a vector whose elements are all one.

Definition 1.5. A square matrix with all elements 1 will be denoted by $\mathbf{J}$.

Property 1.3. $\mathbf{1}^T\mathbf{1} = n$ and $\mathbf{1}\mathbf{1}^T = \mathbf{J}$.

Definition 1.6 (Idempotent Matrix)**.** A square matrix is idempotent if and only if $\mathbf{A}^2 = \mathbf{A}$.

Property 1.4. If $\mathbf{A}$ is idempotent, then $\text{Trace}(\mathbf{A}) = \text{Rank}(\mathbf{A})$.

Property 1.5. $\mathbf{A}$ is idempotent if and only if $\text{Rank}(\mathbf{A}) + \text{Rank}(\mathbf{I} - \mathbf{A}) = n$, where n is the dimension of $\mathbf{A}$ and $\mathbf{I}$ is the $n \times n$ identity matrix.

Property 1.6. If $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2$, where $\mathbf{A}, \mathbf{A}_1, \mathbf{A}_2$ are all idempotent, then

$$\text{Rank}(\mathbf{A}) = \text{Rank}(\mathbf{A}_1) + \text{Rank}(\mathbf{A}_2).$$

1.5.5 Basic Operations for Matrices

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$.
- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$.
- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.
- $\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$.
- $k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B}$.
- $(\mathbf{A}^T)^T = \mathbf{A}$.
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
- $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$.
- $(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$.
- $(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$.
- $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$.

1.5.6 Variance-Covariance Matrix of Random Vectors

The variance-covariance matrix of a random vector $\mathbf{Y}$ is a symmetric matrix with the ith diagonal element is $\mathrm{Var}[Y_i]$ and the ijth element is $\mathrm{Cov}(Y_i, Y_j), i \neq j$, i.e.,

$$\begin{aligned}
\mathrm{Var}[\mathbf{Y}] &= \mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])^T] \\
&= \mathbb{E}\left[\begin{pmatrix} (Y_1 - \mathbb{E}[Y_1])^2 & (Y_1 - \mathbb{E}[Y_1])(Y_2 - \mathbb{E}[Y_2]) & \cdots \\ (Y_2 - \mathbb{E}[Y_2])(Y_1 - \mathbb{E}[Y_1]) & (Y_2 - \mathbb{E}[Y_2])^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}\right] \\
&= \begin{pmatrix} \mathrm{Var}[Y_1] & \mathrm{Cov}(Y_1, Y_2) & \cdots \\ \mathrm{Cov}(Y_2, Y_1) & \mathrm{Var}[Y_2] & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.
\end{aligned}$$

Note that $\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(Y_j, Y_i)$.

Property 1.7. If $\mathbf{A}$ is a square matrix of constants, then $\mathrm{Var}[\mathbf{AY}] = \mathbf{A}\mathrm{Var}[\mathbf{Y}]\mathbf{A}^T$.

Proof. We have

$$\begin{aligned}
\mathrm{Var}[\mathbf{AY}] &= \mathbb{E}[(\mathbf{AY} - \mathbb{E}[\mathbf{AY}])(\mathbf{AY} - \mathbb{E}[\mathbf{AY}])^T] \\
&= \mathbb{E}[\mathbf{A}(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T\mathbf{A}^T] \\
&= \mathbf{A}\mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T]\mathbf{A}^T \\
&= \mathbf{A}\mathrm{Var}[\mathbf{Y}]\mathbf{A}^T.
\end{aligned}$$

□

1.5.7 Rank of a Matrix

Definition 1.7 (Rank)**.** The rank of a matrix is the maximum number of linearly independent rows or columns in the matrix.

Property 1.8. $\text{Rank}(\mathbf{AB}) \leqslant \min(\text{Rank}(\mathbf{A}), \text{Rank}(\mathbf{B}))$.

1.5.8 Invertibility

Property 1.9. We say a matrix $\mathbf{A}$ is invertible if

(1) $\mathbf{A}$ is non-singular, i.e., $\det(\mathbf{A}) \neq 0$;

(2) or rows or columns of $\mathbf{A}$ are linearly independent;

(3) or $\mathbf{A}$ is of full rank.

1.5.9 Matrix Differentiation

Property 1.10. If $\theta^T = (\theta_1, \cdots, \theta_k), \mathbf{c}^T = (c_1, \cdots, c_k)$ s.t.

$$f(\theta) = \mathbf{c}^T\theta = \theta^T\mathbf{c} = \sum_i c_i\theta_i$$

is a scalar, then

$$\frac{\partial f(\theta)}{\partial \theta} = \mathbf{c}.$$

Property 1.11. Let $\mathbf{A}$ be a $k \times k$ symmetric matrix, and suppose $f(\theta) = \theta^T\mathbf{A}\theta$, then

$$\frac{\partial f(\theta)}{\partial \theta} = 2\mathbf{A}\theta.$$

Example 1.2. Suppose $f(\beta) = \mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta$, where $\mathbf{X}$ is $n \times 2, \mathbf{Y}$ is $n \times 1$ and β is 2×1. we have

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

2 Sample Linear Regression

2.1 Statistical Model

$$Y = \beta_0 + \beta_1 X + e,$$

where Y is dependent or response variable, X is independent or explanatory variable, β_0 is intercept parameter, β_1 is slope parameter, and e is random error or noise (variation in measures that we cannot account for).

Given a specific value of $X = x$, we want to find the expected value of Y

$$\mathbb{E}[Y|X = x].$$

2.2 Estimating β_0, β_1

Given n pairs bivariate data $(x_1, y_1), \cdots, (x_n, y_n)$, we want to use $\widehat{\beta}_0$ and $\widehat{\beta}$ to estimate β_0 and β_1.

Consider the residual sum of squares

$$RSS = \sum_{i=1}^{n} \widehat{e}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i \right) \right]^2,$$

we can use least squares method that minimizes the criterion RSS to find the estimators.

2.2.1 Least Squares Method

Least squares method makes no statistical assumptions. We have

$$\frac{\partial RSS}{\partial \widehat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) \text{ and } \frac{\partial RSS}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) x_i.$$

Let $\frac{\partial RSS}{\partial \widehat{\beta}_0}$ and $\frac{\partial RSS}{\partial \widehat{\beta}_1}$ be 0, we get the normal equations

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) = 0 \text{ and } \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) x_i = 0.$$

Therefore, we have

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{\beta}_0 - \sum_{i=1}^{n} \widehat{\beta}_1 x_i = n\overline{y} - n\widehat{\beta}_0 - n\widehat{\beta}_1 \overline{x} = 0 \Rightarrow \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

Besides,

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \widehat{\beta}_0 x_i - \sum_{i=1}^{n} \widehat{\beta}_1 x_i^2 &= \sum_{i=1}^{n} x_i y_i - \left(\overline{y} - \widehat{\beta}_1 \overline{x} \right) \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{xy} + n\widehat{\beta}_1 \overline{x}^2 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0, \end{aligned}$$

i.e.,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} := \frac{SXY}{SXX}.$$

2.2.2 Interpretation

$\widehat{\beta}_0$: The expected value of y when $x = 0$. No practical interpretation unless 0 is within the range of the predictor values.

$\widehat{\beta}_1$: When x changes by 1 unit, the corresponding average change in y is the slope.

2.2.3 Estimation in R

```
model = lm(y ~ x)
summary(model)
```

2.3 Properties of Fitted Regression Line

Property 2.1.

$$\sum_{i=1}^{n} \widehat{e}_i = 0.$$

Proof. By definition,

$$\begin{aligned}\sum_{i=1}^{n} \widehat{e}_i &= \sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = \sum_{i=1}^{n} \left(y_i - \overline{y} + \widehat{\beta}_1 \overline{x} - \widehat{\beta}_1 x_i\right) \\ &= n\overline{y} - n\overline{y} + n\widehat{\beta}_1\overline{x} - n\widehat{\beta}_1\overline{x} = 0.\end{aligned}$$

□

Property 2.2. The sum of squares of residuals is not 0 unless the fit to the data is perfect.

Property 2.3.

$$\sum_{i=1}^{n} \widehat{e}_i x_i = 0.$$

Proof. By definition,

$$\begin{aligned}\sum_{i=1}^{n} \widehat{e}_i x_i &= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) x_i = \sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{xy} - \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = 0.\end{aligned}$$

□

Property 2.4.

$$\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i = 0.$$

Proof. By definition,

$$\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i = \sum_{i=1}^{n} \widehat{e}_i (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = \widehat{\beta}_0 \sum_{i=1}^{n} \widehat{e}_i + \widehat{\beta}_1 \sum_{i=1}^{n} \widehat{e}_i x_i = 0 + 0 = 0.$$

□

Property 2.5.

$$\sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i.$$

Proof. We have

$$\sum_{i=1}^{n} \widehat{e}_i = 0 = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{y}_i \Rightarrow \sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i.$$

□

2.4 Assumptions

The Gauss-Markov conditions are:

1. $\mathbb{E}[e_i] = 0$.

2. $\mathrm{Var}[e_i] = \sigma^2$, i.e., homoscedastic.

3. The errors are uncorrelated or $\mathrm{Cov}(e_i, e_j) = \rho(e_i, e_j) = 0$.

Theorem 2.1 (Gauss-Markov Theorem)**.** Under the conditions or the simple linear regression model, the least-squares parameter estimators are best linear unbiased estimators.

We assume that Y is relate to x by the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \cdots, n.$$

Under the conditions we have

$$\mathbb{E}[Y|X = x_i] = \beta_0 + \beta_1 x_i$$

and

$$\mathrm{Var}[Y|X = x_i] = \mathrm{Var}[\beta_0 + \beta_1 x_i + e_i | X = x_i] = \mathrm{Var}[e_i] = \sigma^2.$$

2.5 Estimating the Variance of the Random Error Term

The variance σ^2 is another parameter of the SLR model and we want to estimate σ^2 to measure the variability of our estimates of Y, and carry out inference on the model.

An unbiased estimate of σ^2 is

$$S^2 = \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{n-2} = \frac{RSS}{n-2}.$$

2.6 Properties of Least Squares Estimators

Since $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$,

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i - \overline{y}\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i.$$

Let $c_i = \frac{x_i - \overline{x}}{SXX}$, we can rewrite $\widehat{\beta}_1$ as

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i,$$

which is a linear combination of y_i.

We have

$$\begin{aligned}\mathbb{E}\left[\widehat{\beta}_1|X\right] &= \mathbb{E}\left[\sum_{i=1}^{n} c_i y_i|X = x_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}[y_i|X = x_i]\\ &= \sum_{i=1}^{n} c_i \mathbb{E}[\beta_0 + \beta_1 x_i] = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i\\ &= \frac{\beta_0}{SXX}\sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1 \sum_{i=1}^{n} \frac{(x_i - \overline{x})x_i}{SXX}\\ &= \beta_1 \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{SXX} = \beta_1.\end{aligned}$$

Therefore, $\widehat{\beta}_1$ is unbiased for β_1. Besides,

$$\begin{aligned}\mathrm{Var}\left[\widehat{\beta}_1|X\right] &= \mathrm{Var}\left[\sum_{i=1}^{n} c_i y_i|X\right] = \sum_{i=1}^{n} c_i^2 \mathrm{Var}[y_i|X = x_i]\\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{SXX^2} = \frac{\sigma^2}{SXX}.\end{aligned}$$

We have

$$\begin{aligned}\mathbb{E}\left[\widehat{\beta}_0|X\right] &= \mathbb{E}\left[\overline{y} - \widehat{\beta}_1\overline{x}|X = x_i\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}_1\overline{x}|X = x_i\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\beta_0 + \beta_1 x_i + e_i|X = x_i] - \overline{x}\mathbb{E}\left[\widehat{\beta}_1|X = x_i\right]\\ &= \frac{1}{n}n\beta_0 + \frac{1}{n}n\beta_1\overline{x} - \overline{x}\beta_1 = \beta_0.\end{aligned}$$

Therefore, $\widehat{\beta}_0$ is unbiased for β_0. Besides,

$$\begin{aligned}\mathrm{Var}\left[\widehat{\beta}_0|X\right] &= \mathrm{Var}\left[\overline{y} - \widehat{\beta}_1\overline{x}|X = x_i\right]\\ &= \mathrm{Var}\left[\overline{y}|X = x_i\right] + \mathrm{Var}\left[\widehat{\beta}_1\overline{x}|X = x_i\right] - 2\mathrm{Cov}\left(\overline{y}, \widehat{\beta}_1\overline{x}|X = x_i\right)\\ &= \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{SXX} - 0 = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right).\end{aligned}$$

Note that $\text{Cov}\left(\overline{y}, \widehat{\beta}_1\overline{x}|X = x_u\right) = \frac{\overline{x}\sigma^2}{n}\sum_{i=1}^{n} c_i = 0$.

2.7 Normal Error Regression Model

Given distributional assumption:

$$e_i \sim \mathcal{N}(0, \sigma^2),$$

we know:

(1) the errors are independent since $\rho = 0$;

(2) since $y_i = \beta_0 + \beta_1 x_i + e_i$, then $Y_i|X \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$;

(3) the least squares estimates of β_0, β_1 are equivalent to their maximum likelihood estimators.

(4) since $\widehat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ is a linear combination of the y_i's, $\widehat{\beta}_1|X \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SXX}\right)$; since $\overline{y}$ is normally distributed, $\widehat{\beta}_0|X \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right)\right)$.

Property 2.6. Under the normal error SLR model, where

$$e_i \sim \mathcal{N}(0, \sigma^2) \text{ and } S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{e}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2,$$

we have

$$\frac{(n-2)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i - \widehat{Y}_i}{\sigma^2}\right)^2 \sim \chi^2_{(n-2)}.$$

Property 2.7. Under the normal error SLR model,

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{SXX}}} \sim t_{(n-2)}.$$

Proof. We have $\widehat{\beta}_1|X = x_i \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SXX}\right)$, and thus

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{SXX}} \sim \mathcal{N}(0, 1).$$

Wherefore

$$\frac{\frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{SXX}}}{\sqrt{(n-2)S^2/\sigma^2/(n-2)}} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{SXX}}} \sim t_{(n-2)}.$$

□

2.8 Inference for the Parameter

2.8.1 Significance Test

- Step 1: $H_0 : \beta_1 = \beta_1^0$ against $H_a : \beta_1 \neq \beta_1^0$.
- Step 2: Test statistic $t = \frac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{S^2/SXX}}$, and under $H_0, t \sim t_{(n-2)}$.
- Step 3: $p-$value $= 2P(t_{(n-2)} \geqslant |t|)$.
- Step 4: The smaller the p-value, the greater the evidence against H_0 and the larger p-value indicate that the data is consistent with H_0.

p-value	Evidence against H_0
< 0.001	Very strong
(0.001, 0.01)	Strong
(0.01, 0.05)	Moderate
(0.05, 0.1)	Weak
> 0.1	None

Note that the test statistic for $\widehat{\beta}_0$ is $t = \frac{\widehat{\beta}_0 - \beta_0^0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right)}}$

2.8.2 Confidence Interval

The CI is

$$\text{Estimate} \pm 100\left(1 - \frac{\alpha}{2}\right) \text{th quantile} \times \text{Standard Error (Estimate)},$$

where α is the critical value.

For β_1, the CI is

$$\left[\widehat{\beta}_1 \pm t_{\frac{\alpha}{2}(n-2)}\sqrt{\frac{S^2}{SXX}}\right].$$

For β_0, the CI is

$$\left[\widehat{\beta}_0 \pm t_{\frac{\alpha}{2}(n-2)}\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right)}\right].$$

Note that a $100(1-\alpha)\%$ CI for θ consists of all those values of θ_0 for which $H_0 : \theta = \theta_0$ will not be rejected at level α. In other words, we do not reject H_0 is θ_0 lies within the CI, and we reject H_0 is the CI does not include θ_0.

2.9 The Pooled Two-Sample t-Procedure

We want to test $H_0 : \mu_x = \mu_y$, where

$$X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_x, \sigma_x^2) \text{ and } Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_y, \sigma_y^2).$$

Suppose two samples are independent and $\sigma_x^2 = \sigma_y^2 = \sigma^2$, then we have

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{(n_x+n_y-2)},$$

where $s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$.

2.10 Regression Analysis of Variance

Notice that $y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})$. We have

$$\begin{aligned} TSS &= \sum_i^n (y_i - \overline{y})^2, \\ RSS &= \sum_i^n (y_i - \widehat{y}_i)^2 = \sum_i^n \widehat{e}_i^2, \\ RegSS &= \sum_i^n (\widehat{y}_i - \overline{y})^2. \end{aligned}$$

RSS, residual SS, is the least square criterion, representing the unexplained variation in y's. $RegSS$, regression SS, is the amount of variation in y's explained by regression line.

Property 2.8. $RegSS = \widehat{\beta}_1^2 SXX$.

Proof. We have

$$\begin{aligned} RegSS &= \sum_i^n (\widehat{y}_i - \overline{y})^2 = \sum_i^n (\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \overline{y})^2 \\ &= \sum_i^n (\overline{y} - \widehat{\beta}_1\overline{x} + \widehat{\beta}_1 x_i - \overline{y})^2 = \widehat{\beta}_1^2 \sum_i^n (x_i - \overline{x})^2 = \widehat{\beta}_1^2 SXX. \end{aligned}$$

□

Property 2.9. $TSS = RSS + RegSS$.

Proof. We have

$$\begin{aligned} \sum_i^n (y_i - \overline{y})^2 &= \sum_i^n ((y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}))^2 \\ &= \sum_i^n (y_i - \widehat{y})^2 + \sum_i^n (\widehat{y}_i - \overline{y})^2 + 2\sum_i^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) \\ &= RSS + RegSS + 2\sum_i^n \widehat{e}_i(\widehat{y}_i - \overline{y}) \\ &= RSS + RegSS + 2\sum_i^n \widehat{e}_i\widehat{y}_i - 2\overline{y}\sum_i^n \widehat{e}_i \\ &= RSS + RegSS. \end{aligned}$$

□

2.10.1 Regression ANOVA Table

Source	SS	df	Mean SS
Regression Line	$RegSS = \widehat{\beta}_1^2 SXX$	1	$\widehat{\beta}_1^2 SXX$
Error	$RSS = \sum_{i=1}^{n} \widehat{e}_i^2$	$n-2$	$\frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n-2} = S^2$
Total	$TSS = \sum_{i}^{n} (y_i - \overline{y})^2$		

Property 2.10. Let

$$F = \frac{MRegSS}{MRSS} = \frac{RegSS/1}{RSS/(n-2)}.$$

If $\beta_1 = 0$, then

$$F \sim F_{(1,n-2)}.$$

Proof. If $\beta_1 = 0$, then $\widehat{\beta}_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{SXX}\right)$, i.e.,

$$\frac{\widehat{\beta}_1}{\sqrt{\sigma^2/SXX}} \sim \mathcal{N}(0,1) \Rightarrow \frac{\widehat{\beta}_1^2}{\sigma^2/SXX} \sim \chi^2_{(1)}.$$

Besides, $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{(n-2)}$, and we have

$$\frac{\frac{\widehat{\beta}_1^2}{\sigma^2/SXX}}{\frac{(n-2)S^2}{\sigma^2}/(n-2)} = \frac{\widehat{\beta}_1^2 SXX}{S^2} = F \sim F_{(1,n-2)}.$$

□

Note that F is another test of $H_0 : \beta_1 = 0$, and in R, we have:

```
anova(model)
```

2.10.2 Coefficient of Determination

Let

$$R^2 = \frac{RegSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Here are some comments about R^2:

- $R^2 \in [0,1]$.
- R^2 gives percentage of variation in y's explained by regression line.
- R^2 is not resistant to outliers.
- A high R^2 does not indicate that the estimated regression line is a good fit since:
  - ∗ we do not have absolute rules about how large it should be;
  - ∗ R^2 can get very high by overfitting.

- It is not meaningful for models without intercept.
- To compare 2 models, R^2 is only useful:
  - $*$ same observations, y's in original units (not transformed);
  - $*$ one set of predictor variables is a subset of the other.

2.10.3 Sample Correlation Coefficient

The estimate of the population correlation is Pearson's Product-Moment Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{SXY}{\sqrt{SXX \cdot SYY}},$$

which is the MLE of ρ. r is distribution free and is always a number between -1 and 1.

Theorem 2.2. $R^2 = r^2$.

Proof. We have

$$R^2 = \frac{RegSS}{TSS} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\widehat{\beta}_1^2 SXX}{SYY} = \frac{\frac{SXY^2}{SXX^2} \cdot SXX}{SYY} = \frac{SXY^2}{SXX \cdot SYY} = r^2.$$

□

Property 2.11. If $\rho = 0$,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\widehat{\beta}_1}{\sqrt{S^2/SXX}} \sim t_{(n-2)},$$

where $\widehat{\beta}_1$ is the slope estimate for the normal error SLR model.

Proof. Since $r^2 = R^2$, then

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\frac{\widehat{\beta}_1\sqrt{SXX}}{\sqrt{SXY}}\sqrt{n-2}}{\sqrt{(n-2)S^2/SXY}} = \frac{\widehat{\beta}_1}{\sqrt{S^2/SXX}}.$$

If $\rho = 0$, then $\beta_1 = 0$, i.e.,

$$\frac{\widehat{\beta}_1}{\sqrt{S^2/SXX}} \sim t_{(n-2)}.$$

□

2.11 Confidence Interval for the Population Regression Line

We want to find a CI for the unknown population regression line at a given value of X, denoted by x^*, i.e.,

$$\mathbb{E}[Y|X=x^*] = \beta_0 + \beta_1 x^*.$$

The point estimate for $\mathbb{E}[Y|X=x^*]$ is

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

We have

$$\mathbb{E}\left[\widehat{y}^*\right] = \mathbb{E}\left[\widehat{y}|X=x^*\right] = \beta_0 + \beta_1 x^*,$$

i.e., $\widehat{y}^*$ is unbiased for $\mathbb{E}[Y|X=x^*]$.

Recall that $\text{Var}\left[\widehat{\beta}_0|X\right] = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right), \text{Var}\left[\widehat{\beta}_1|X\right] = \frac{\sigma^2}{SXX}$, then

$$\text{Cov}\left[\widehat{\beta}_0, \widehat{\beta}_1|X\right] = \text{Cov}\left[\overline{y} - \widehat{\beta}_1\overline{x}, \widehat{\beta}_1|X\right] = -\overline{x}\text{Var}\left[\widehat{\beta}_1|X\right] = -\frac{\overline{x}\sigma^2}{SXX}.$$

Wherefore

$$\begin{aligned}\text{Var}\left[\widehat{y}^*\right] &= \text{Var}\left[\widehat{y}|X=x^*\right] = \text{Var}\left[\widehat{\beta}_0 + \widehat{\beta}_1 x|X=x^*\right] \\ &= \text{Var}\left[\widehat{\beta}_0|X=x^*\right] + (x^*)^2\text{Var}\left[\widehat{\beta}_1|X=x^*\right] + 2x^*\text{Cov}\left[\widehat{\beta}_0, \widehat{\beta}_1|X=x^*\right] \\ &= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right) + (x^*)^2\frac{\sigma^2}{SXX} - \frac{2x^*\overline{x}\sigma^2}{SXX} = \sigma^2\left(\frac{1}{n} + \frac{(x^*-\overline{x})^2}{SXX}\right).\end{aligned}$$

Hence, as $n\uparrow, \text{Var}\left[\widehat{y}^*\right]\downarrow$; as x^* closer to $\overline{x}, \text{Var}\left[\widehat{y}^*\right]\downarrow$.

Using $S^2 = MRSS$, we get the standard error of the estimate of $\mathbb{E}[Y|X=x^*]$,

$$\sqrt{S^2\left(\frac{1}{n} + \frac{(x^*-\overline{x})^2}{SXX}\right)}.$$

Hence, a $100(1-\alpha)\%$ CI for $\mathbb{E}[Y|X=x^*]$, the mean response for all the elements in the population with $X=x^*$ is

$$\left[\widehat{y}^* \pm t_{\frac{\alpha}{2}(n-2)}\sqrt{S^2\left(\frac{1}{n} + \frac{(x^*-\overline{x})^2}{SXX}\right)}\right].$$

Notice that it is only valid for x^* in the range of the original data values of X but not for extrapolation.

2.12 Prediction Interval for Actual Value of Y

A confidence interval is always reported for a parameter while a prediction interval is reported for the value of a random variable. We want to find a PI for the actual value of Y at $X=x^*$, i.e., $Y^* = Y|X=x^*$.

The point estimate for Y^* is

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

The error in our prediction is

$$\varepsilon^* = Y^* - \widehat{y}^*.$$

The predicted value $\widehat{y}^*$ has two sources of variability:

- Since the regression line is estimated at $\widehat{\beta}_0 + \widehat{\beta}_1 X$;
- due to ε^*, some points do not fall exactly on the line.

We have

$$\begin{aligned}\operatorname{Var}\left[Y^* - \widehat{y}^*\right] &= \operatorname{Var}\left[Y - \widehat{y}|X = x^*\right] \\ &= \operatorname{Var}[Y|X = x^*] + \operatorname{Var}\left[\widehat{y}|X = x^*\right] - 2\operatorname{Cov}(Y, \widehat{y}|X = x^*) \\ &= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right) - 0 = \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right).\end{aligned}$$

Notice that $\operatorname{Cov}(Y, \widehat{y}|X = x^*) = 0$ since Y^* is a new observation.

Hence, a $100(1-\alpha)\%$ PI for $Y|X = x^*$ is

$$\left[\widehat{y}^* \pm t_{\frac{\alpha}{2}(n-2)}\sqrt{S^2\left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right)}\right].$$

PIs for Y^* have the same center but are wider than CIs for $\mathbb{E}[Y|X = x^*]$.

3 Diagnostics and Transformations for Simple Linear Regression

3.1 Phenomena and Fallacies in Regression

- The regression effect: Regression to the mean - more values near the average than away from it; unusually large or small measurements tend to be followed by measurements that are closer to the mean.
- The regression fallacy: when the regression effect is mistaken for a real effect.
- Ecological fallacy/Correlation: inference is made about an individual based on aggregate data for a group.

3.2 Validity of SLR Model: Model Linearity

3.2.1 Residuals

Recall that the residuals is $\widehat{e}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}x_i$.

Property 3.1. $\mathbb{E}\left[\widehat{e}_i\right] = 0$.

Proof. We have

$$\mathbb{E}\left[\widehat{e}_i\right] = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i = 0.$$

□

Property 3.2. $\operatorname{Var}\left[\widehat{e}_i\right] = (1-h_{ii})\sigma^2$.

Proof. We have

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}, \widehat{\beta}_1 = \frac{\sum\limits_{j=1}^{n}(x_j-\overline{x})y_j}{SXX}, \overline{y} = \frac{1}{n}\sum_{j=1}^{n}y_j.$$

Thus

$$\begin{aligned}\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i &= \frac{1}{n}\sum_{j=1}^{n}y_j + \frac{1}{SXX}\sum_{j=1}^{n}(x_j-\overline{x})(x_i-\overline{x})y_j \\ &= \sum_{j=1}^{n}\left(\frac{1}{n} + \frac{(x_i-\overline{x})(x_j-\overline{x})}{SXX}\right)y_j := \sum_{j=1}^{n}h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i}h_{ij}y_j.\end{aligned}$$

Hence,

$$\widehat{e}_i = y_i - \widehat{y}_i = (1-h_{ii})y_i + \sum_{j\neq i}h_{ij}y_j,$$

and thus

$$\begin{aligned}\operatorname{Var}\left[\widehat{e}_i\right] = \operatorname{Var}\left[(1-h_{ii})y_i + \sum_{j\neq i}h_{ij}y_j\right] &= (1-h_{ii})^2\sigma^2 + \sum_{j\neq i}h_{ij}^2\sigma^2 + 0 \\ &= \left(1 - 2h_{ii} + \sum_{j=1}^{n}h_{ij}^2\right)\sigma^2.\end{aligned}$$

Since

$$\sum_{j=1}^{n} h_{ij}^2 = \sum_{j=1}^{n}\left(\frac{1}{n} + \frac{(x_i-\overline{x})(x_j-\overline{x})}{SXX}\right)^2 = \frac{1}{n} + 0 + \sum_{j=1}^{n}\frac{(x_i-\overline{x})^2(x_j-\overline{x})^2}{SXX^2}$$
$$= \frac{1}{n} + \frac{(x_i-\overline{x})^2}{SXX} = h_{ii},$$

then $\mathrm{Var}\left[\widehat{e}_i\right] = (1-h_{ii})\sigma^2$, where $h_{ii} = \frac{1}{n} + \frac{(x_i-\overline{x})^2}{SXX}$. □

Property 3.3. $\sum\limits_{j=1}^{n} h_{ij} = 1$.

Proof. We have

$$\sum_{j=1}^{n} h_{ij} = \sum_{j=1}^{n}\left(\frac{1}{n} + \frac{(x_i-\overline{x})(x_j-\overline{j})}{SXX}\right) = \frac{n}{n} + \frac{x_i-\overline{x}}{SXX}\sum_{j=1}^{n}(x_j-\overline{j}) = 1+0 = 1.$$

□

Property 3.4. $\sum\limits_{j=1}^{n} h_{ij}^2 = h_{ii}$.

Property 3.5. $\sum\limits_{j=1}^{n} h_{ij}x_j = x_i$.

Proof. We have

$$\sum_{j=1}^{n} h_{ij}x_j = \sum_{j=1}^{n}\left(\frac{x_j}{n} + \frac{(x_i-\overline{x})(x_j-\overline{x})x_j}{SXX}\right) = \overline{x} + \frac{(x_i-\overline{x})SXX}{SXX} = x_i.$$

□

Property 3.6. $\mathrm{Var}\left[\widehat{y}_i\right] = h_{ii}\sigma^2$.

Proof. We have

$$\mathrm{Var}\left[\widehat{y}_i\right] = \mathrm{Var}\left[\sum_{j=1}^{n} h_{ij}y_j\right] = \sum_{j=1}^{n} h_{ij}^2 \mathrm{Var}[y_j] = h_{ii}\sigma^2.$$

□

Property 3.7. $\mathrm{Cov}\left(\widehat{e}_i, \widehat{e}_j\right) = -h_{ij}\sigma^2$, for $i \neq j$.

Property 3.8. $\widehat{e}_i \sim \mathcal{N}\left(0, (1-h_{ii})\sigma^2\right)$.

We can plot the residuals in three way:

- $\widehat{e}_i$ against observation order/time i;
- $\widehat{e}_i$ against predictor value x_i;
- $\widehat{e}_i$ against fitted value $\widehat{y}_i$.

Residuals plot can be used to assess whether an appropriate model has been fit to the data: if no pattern is found, then the model provides an adequate summary of the data; or the shape of the pattern provides information on the function of x that is missing from the model.

3.2.2 Diagnostics

We can use ***scatter plot***, ***residual plot*** and ***added-variable plot*** to check for SLR model.

- Residual plot: Using residuals against x_i or $\widehat{y}_i$ will yield the same information. If the model is linear, there should be no pattern.
- Added-Variable plot: Using residuals against other potential predictors. Any pattern indicates that the other predictor should be included in the model.

If the assumption is violated, we can add additional predictors or transform X and/or Y.

3.3 Validity of SLR Model: Uncorrelated Errors

It is based on the design of the study and we can use randomization, where possible, to satisfy the assumption and widen the scope of inferences. Since Cov $(\widehat{e}_i, \widehat{e}_j) = -h_{ij}\sigma^2, i \neq j$, residuals are not uncorrelated even if errors are independent. However the covariance is usually so small and can be ignored in practice.

We can use ***residual plot*** to check, using residuals against observation order. If the errors are uncorrelated, there should be no pattern.

If the predictor is time-dependent, auto-correlation may exist. If the assumption is violated, we can fit a time series model or do longitudinal data analysis.

3.4 Validity of SLR Model: Homoscedasticity

3.4.1 Standardized Residuals

Definition 3.1 (Standardized Residuals)**.** The ith standardized residual is

$$r_i = \frac{\widehat{e}_i}{\sqrt{S^2(1-h_{ii})}},$$

where $\widehat{\sigma}^2 = S^2 = \frac{1}{n-2}\sum_{j=1}^{n} \widehat{e}_j^2$.

3.4.2 Diagnostics

We can use ***residual plot*** to check, using $\sqrt{|\text{Residuals}|}$ or $\sqrt{|\text{Standardized Residuals}|}$ against x_i. If the variance is constant, there should be no pattern.

If the assumption is violated, we can transform X and/or Y, do weighted least squares or fit a generalized linear model (models variance as a function of the mean).

3.5 Normality

3.5.1 Q-Q Plots

Probability plots are useful tools to graphically assess goodness-of-fit: we plot the observed order statistics against the expected theoretical quantiles, and if the data follows the particular distribution, the plot should look roughly linear. The most common probability plot is the normal Q-Q plot.

3.5.2 The Shapiro-Wilk Test

We test H_0 : the data follow a normal distribution, and the test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{2}(x_i - \overline{x})^2},$$

where $a_i = \frac{\mathbf{m}^T V^{-1}}{(\mathbf{m}^T V^{-1} V^{-1}\mathbf{m})^{1/2}}$, $\mathbf{m}^T = (m_1, \cdots, m_n)$ are expected values of standard normal order statistics and V is the covariance matrix of those normal order statistics.

If the p-value is less than the significance level, there is evidence that the data is not normal.

In R, we can use

```
shapiro.test(data)
```

3.5.3 Diagnostics

The assumption of normal errors is needed in small samples for the validity of t distribution based hypothesis tests and CI, and for all sample sizes for PI.

We can use ***normal Q-Q plot*** of residuals or standardized residuals. If the errors are normally distributed, then there should be a linear relationship.

Recall that $\sum_{j=1}^{n} h_{ij} = 1$ and $\sum_{j=1}^{n} h_{ij} x_j = x_i$ and thus

$$\begin{aligned}\widehat{e}_i = y_i - \sum_{j=1}^{n} h_{ij} y_j &= \beta_0 + \beta_1 x_i + e_i - \sum_{j=1}^{n}(\beta_0 + \beta_1 x_j + e_j) h_{ij} \\ &= \beta_0 + \beta_1 x_i + e_i - \beta_0 - \beta_1 x_i - \sum_{j=1}^{n} h_{ij} e_j = e_i - \sum_{j=1}^{n} h_{ij} e_j.\end{aligned}$$

In small to moderate samples, second term can dominate e_i and the residuals can look like they come from any distributions including normal even if the errors do not. As n increases, second term has a much smaller variance than that of e_i and thus for large samples, the residuals can be used to assess normality of the errors.

Note that do not bother to check normality in the presence of other issues.

3.6 Outliers, Leverage Points, and Influential Points

Outliers are the points that do not follow the pattern of the data. Outliers with respect to the explanatory variable (in the x direction) are called leverage points. If we remove the leverage points from the data, the fitted model is substantially different, then we call it influential points.

A good leverage point has no adverse effect on the estimated regression coefficients, decreases the standard errors, and increases R^2.

3.6.1 Quantifying Leverage

Note that h_{ii} is the leverage of the ith data point and it varies only by the squared distance of x_i from its mean but not the values of the y's. Recall that

$$\widehat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j,$$

and h_{ii} shows how y_i affects $\widehat{y}_i$.

For SLR,

$$\overline{h}_{ii} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(x_i-\overline{x})^2}{SXX}\right) = \frac{1}{n}\left(\frac{n}{n} + \frac{SXX}{SXX}\right) = \frac{2}{n}$$

and we define a point is a leverage point if

$$h_{ii} > 2\overline{h}_{ii} = \frac{4}{n}.$$

If $h_{ii} \approx 1$, then for $i \neq j, h_{ij} \approx 0$ (since $\sum\limits_{j=1}^{n} h_{ij} = 1$) and thus $\widehat{y}_i \approx y_i$, so the ith point is a leverage point, i.e., the fitted line is attracted by the point. $\widehat{e}_i$ has small variance and

$$\text{Var}\left[\widehat{y}_i\right] \approx \text{Var}[y_i].$$

When leverage points do not exist, there is little or no difference in the plots of residuals when compared to plots of standardized residuals.

In small to moderate size data sets, an outlier point is one if $|r_i| > 2$. In very large data sets, an outlier point is one if $|r_i| > 4$.

An influential point is a leverage point which is also an outlier.

In R, we have:

```r
# Calculate h_{ii}.
lm.influence(model)$hat
hatvalues(model)


# Calculate r_i.
```

```
rstandard(model)

# Calculate hat{e_i}.
model$residuals
```

3.6.2 Cook's Distance

Definition 3.2 (Cook's Distance)**.** Cook's distance for the ith point is given by

$$D_i = \frac{\sum\limits_{j=1}^{n} \left(\widehat{y}_{j(i)} - \widehat{y}_j\right)^2}{2S^2},$$

where $\widehat{y}_{j(i)}$ is the j fitted value based on the fit obtained when the ith case has been removed from the fit.

Cook's distance measures influence of the ith observation.

Property 3.9. $D_i = \frac{r_i^2}{2}\left(\frac{h_{ii}}{1-h_{ii}}\right)$, where r_i is the ith standardized residual and h_{ii} is the leverage of the ith point.

Hence a large D_i may due to large r_i, large h_{ii} or both.

If the largest D_i is much less than 1, deletion of a case will not change the estimate of $\widehat{\beta}$ by much.

Definition 3.3. A point is noteworthy if

$$D_i > \frac{4}{n-2}.$$

In practice, we look for gaps in the values of Cook's distance and not just whether values exceed the suggested cut-off.

In R, we have:

```
cooks.distance(model)
```

3.6.3 Recommendations

- Base estimates and confidence intervals only on valid model.
- Unusual points should be thoroughly investigated and should not be routinely deleted from an analysis.
- Outliers often point to important features of the problem not considered before.

3.7 Diagnostic Plots from R

In R, we have:

```
plot(model)
```

We will have 4 plots:

- $\widehat{e}_i$ against $\widehat{y}_i$.
- Normal Q-Q plot of r_i.
- $\sqrt{r_i}$ against $\widehat{y}_i$.
- r_i against h_{ii} with Cook's distance.

3.8 Transformations

With transformations, we can overcome problems due to non-constant variance, estimate percentage effects, overcome problems due to nonlinearity, and remedy non-normality.

3.8.1 Common Transformations

Common monotonic transformations are $X^2, \ln X, \sqrt{X}$.

- If relationship is non-linear but variance of Y is nearly constant - transform X.
- If relationship is non-linear and variance is non-constant - transform Y.
- If relationship is linear but variance is non-constant - transform both X and Y.

Note that transforming changes the relative spacing of the observations.

3.8.2 Variance Stabilizing Transformations

Theorem 3.1 (Delta Method)**.** *Suppose Y has a distribution with mean μ and variance σ_Y^2, and $Z = f(Y)$. We have*

$$\mathbb{E}[Z] \approx f(\mu), \mathrm{Var}[Z] \approx \sigma_Y^2 \left[f'(\mu)\right]^2 .$$

Proof. Recall the first order Taylor series expansion of Z, we have

$$Z = f(Y) = f(\mu) + (Y - \mu)f'(\mu) + o(Y - \mu)$$

and thus

$$\mathbb{E}[Z] = \mathbb{E}[f(\mu)] + \mathbb{E}[(Y - \mu)f'(\mu)] + \mathbb{E}[o(Y - \mu)] \approx f(\mu).$$

Besides, $Z - f(\mu) = (Y - \mu)f'(\mu) + o(Y - \mu)$ and thus

$$\mathrm{Var}[Z] = \mathbb{E}[(Z - f(\mu))^2] \approx \mathbb{E}[((Y - \mu)f'(\mu))^2] = \sigma_Y^2[f'(\mu)]^2.$$

□

Example 3.1. Suppose $Y \sim \text{Poisson}(\mu)$, then $\mathbb{E}[Y] = \mu = \text{Var}[Y]$. Hence variance is linearly related to expectation and we want a $f(Y)$ s.t. $\text{Var}[f(Y)]$ will be constant (independent of μ), i.e., we want

$$\text{Var}[f(Y)] \approx \mu[f'(\mu)]^2 = C \Rightarrow [f'(\mu)]^2 \propto \frac{1}{\mu} \Rightarrow f'(\mu) \propto \frac{1}{\sqrt{\mu}}.$$

Thus,

$$f(\mu) \propto \sqrt{\mu}.$$

4 Weighted Least Squares Regression

To overcome non-constant variance of the error term and transformation may create an inappropriate regression relationship, we can use WLS.

4.1 Statistical Model

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

where e_i has mean 0 but variance $\frac{\sigma^2}{w_i}$.

We can consider three cases:

- Large w_i : $\frac{\sigma^2}{w_i}$ is close to 0 and estimates of β_0, β_1 are s.t. fitted line at x_i is close to y_i.
- Small w_i : $\frac{\sigma^2}{w_i}$ is large and estimates of β_0, β_1 take little account of (x_i, y_i).
- Zero w_i : $\frac{\sigma^2}{w_i}$ is infinite and the ith case should be ignored in fitting the line.

4.2 Weighted Least Square Criterion

Consider

$$WRSS = \sum_{i=1}^{n} w_i \left(y_i - \widehat{y}_{w_i}\right)^2 = \sum_{i=1}^{n} w_i \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2.$$

The larger the value of w_i, the more the i th case is taken into account. $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are obtained by minimizing WRSS, and we have

$$\widehat{\beta}_{0w} = \frac{\sum\limits_{i=1}^{n} w_i y_i}{\sum\limits_{i=1}^{n} w_i} - \widehat{\beta}_{1w} \frac{\sum\limits_{i=1}^{n} w_i x_i}{\sum\limits_{i=1}^{n} w_i} = \overline{y}_w - \widehat{\beta}_{1w} \overline{x}_w,$$

and

$$\widehat{\beta}_{1w} = \frac{\sum\limits_{i=1}^{n} w_i \left(x_i - \overline{x}_w\right)\left(y_i - \overline{y}_w\right)}{\sum\limits_{i=1}^{n} w_i \left(x_i - \overline{x}_w\right)^2}.$$

5 Multiple Linear Regression

5.1 SLR in Matrix Form

5.1.1 Statistic Model

Consider the simple linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e},$$

where $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ and $\mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$. Note that $\mathbf{X}$ is called the design matrix.

5.1.2 Normal Error Regression Model

The Gauss-Markov conditions are:

1. $\mathbb{E}[\mathbf{e}] = \mathbf{0}$.

2. $\text{Var}[\mathbf{e}] = \sigma^2\mathbf{I}$.

In addition, we assume the error terms follow a multivariate normal distribution:

$$\mathbf{e} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}).$$

Therefore, we have $\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$ and $\text{Var}[\mathbf{Y}|\mathbf{X}] = \sigma^2\mathbf{I}$.

5.1.3 Least Squares Method and Properties

We find estimates of β_0 and β_1 by minimizing RSS. We have

$$\begin{aligned} RSS\left(\widehat{\beta}\right) &= \sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2 \\ &= \left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right)^T\left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right) \\ &= \left(\mathbf{Y}^T - \widehat{\beta}^T\mathbf{X}^T\right)\left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right) \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\widehat{\beta} - \widehat{\beta}^T\mathbf{X}^T\mathbf{Y} + \widehat{\beta}^T\mathbf{X}^T\mathbf{X}\widehat{\beta} \\ &= \mathbf{Y}^T - 2\widehat{\beta}^T\mathbf{X}^T\mathbf{Y} + \widehat{\beta}^T\mathbf{X}^T\mathbf{X}\widehat{\beta}. \end{aligned}$$

Let

$$\frac{\partial RSS\left(\widehat{\beta}\right)}{\partial\widehat{\beta}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\widehat{\beta} = 0,$$

we have

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

provided $\mathbf{X}^T\mathbf{X}$ is invertible.

Besides,

$$\begin{aligned}\widehat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \frac{1}{SXX}\begin{pmatrix}\frac{1}{n}\sum x_i^2 & -\overline{x} \\ -\overline{x} & 1\end{pmatrix}\begin{pmatrix} n\overline{y} \\ \sum x_i y_i\end{pmatrix} = \begin{pmatrix}\frac{\sum x_i^2 \overline{y} - \sum x_i y_i \overline{x}}{SXX} \\ \frac{\sum x_i y_i - n\overline{xy}}{SXX}\end{pmatrix}.\end{aligned}$$

We also have

$$\mathbb{E}\left[\widehat{\beta}|\mathbf{X}\right] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{Y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \mathbf{I}\beta = \beta,$$

and

$$\begin{aligned}\mathrm{Var}\left[\widehat{\beta}|\mathbf{X}\right] &= \mathrm{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}|\mathbf{X}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Var}[\mathbf{Y}|\mathbf{X}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\ &= \sigma^2\mathbf{X}^{-1}(\mathbf{X}^T)^{-1}\mathbf{X}^T(\mathbf{X}^{-1}(\mathbf{X}^T)^{-1}\mathbf{X}^T)^T = \sigma^2\mathbf{X}^{-1}(\mathbf{X}^T)^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \begin{pmatrix}\frac{\sigma^2\sum x_i^2}{nSXX} & -\frac{\overline{x}\sigma^2}{SXX} \\ -\frac{\overline{x}\sigma^2}{SXX} & \frac{\sigma^2}{SXX}\end{pmatrix}.\end{aligned}$$

5.1.4 Properties of Residuals

We have

$$\widehat{\mathbf{e}} = \begin{pmatrix}\widehat{e}_1 \\ \vdots \\ \widehat{e}_n\end{pmatrix} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\beta} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} := \mathbf{Y} - \mathbf{H}\mathbf{Y},$$

where $\mathbf{H}$ is the hat matrix and the elements of $\mathbf{H}$ are h_{ij}. Therefore,

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

Property 5.1. $\mathbf{H}$ is symmetric and idempotent.

Proof. We have

$$\mathbf{H}^T = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}.$$

Also

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}.$$

□

Hence,

$$\begin{aligned}\widehat{\mathbf{e}} &= (\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \mathbf{e}) = (\mathbf{I} - \mathbf{H})\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{e} \\ &= \mathbf{X}\beta - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{e} = \mathbf{X}\beta - \mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{e}.\end{aligned}$$

Thus,

$$\mathbb{E}\left[\widehat{\mathbf{e}}|X\right] = \mathbb{E}[(\mathbf{I} - \mathbf{H})\mathbf{e}|X] = (\mathbf{I} - \mathbf{H})\mathbb{E}[\mathbf{e}|X] = \mathbf{0}$$

and

$$\mathrm{Var}\left[\widehat{\mathbf{e}}|X\right] = \mathrm{Var}[(\mathbf{I} - \mathbf{H})\mathbf{e}|X] = (\mathbf{I} - \mathbf{H})\mathrm{Var}[\mathbf{e}|X](\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{I} - \mathbf{H}).$$

5.1.5 Analysis of Variance

We can write $TSS, RSS, RegSS$ in quadratic form (i.e., for a vector $\mathbf{Y}$ and a symmetric matrix $\mathbf{A}, \mathbf{Y}^T\mathbf{A}\mathbf{Y}$ is a quadratic form):

$$\begin{aligned}
TSS &= \sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n y_i^2 - n\overline{y}^2 = \mathbf{Y}^T\mathbf{Y} - \frac{1}{n}\mathbf{Y}^T\mathbf{J}\mathbf{Y} = \mathbf{Y}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}, \\
RSS &= \sum_{i=1}^n \widehat{e}_i = \widehat{\mathbf{e}}^T\widehat{\mathbf{e}} = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})\mathbf{Y} = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}, \\
RegSS &= TSS - RSS = \mathbf{Y}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} - \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y} = \mathbf{Y}^T\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}.
\end{aligned}$$

Also, we have

$$\begin{aligned}
&\text{Rank}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = \text{Rank}(\mathbf{I}) - \text{Rank}\left(\frac{1}{n}\mathbf{J}\right) = n-1, \\
&\text{Rank}(\mathbf{I}-\mathbf{H}) = \text{Rank}(\mathbf{I}) - \text{Rank}(\mathbf{H}) = n-2, \\
&\text{Rank}\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) = \text{Rank}(\mathbf{H}) - \text{Rank}\left(\frac{1}{n}\mathbf{J}\right) = 1.
\end{aligned}$$

5.2 Multiple Linear Regression

5.2.1 Statistic Model

In a multiple linear regression model,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, i = 1, \cdots, n.$$

Note that there are $p(p > 1)$ predictor variables, the model is linear in the β's, the predictors may be non-linear (e.g., $\ln X_1, X_1 \cdot X_2$), parameters include $(p+1)$ β's and σ^2, and thus there are $(p+2)$ parameters in this model (so we need at least that many observations).

In matrix notation, the multiple regression model is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e},$$

where $\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}.$

5.2.2 Model Assumptions

The Gauss-Markov conditions are:

(1) $\mathbb{E}[\mathbf{e}] = \mathbf{0}$.

(2) $\text{Var}[\mathbf{e}] = \sigma^2\mathbf{I}$.

Therefore, $\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$ and $\text{Var}[\mathbf{Y}|\mathbf{X}] = \sigma^2\mathbf{I}$. Besides, for inference, normal error assumption is needed.

5.2.3 Least Squares Method

Least squares estimates of β are

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

5.2.4 Analysis of Variance

Property 5.2. Rank$(\mathbf{I}-\mathbf{H}) = n-p-1$.

Proof. We know $\dim(\mathbf{H}) = n \times n, \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Since

$$\text{Trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \text{Trace}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) = \text{Trace}(\mathbf{I}_{p+1}) = p+1,$$

then Rank$(\mathbf{H}) = $ Trace$(\mathbf{H}) = p+1,$ Rank$(\mathbf{I}-\mathbf{H}) = n-p-1$. □

Hence, we have:

Source	SS	df	Mean SS	F
Regression Line	$RegSS = \mathbf{Y}^T\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y}$	p	$MRegSS = \frac{RegSS}{p}$	$F = \frac{MRegSS}{MSE}$
Error	$RSS = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}$	$n-p-1$	$MSE = \frac{RSS}{n-p-1}$	
Total	$TSS = \mathbf{Y}^T\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)$	$n-1$		

5.2.5 Estimating σ^2 in MLR

Define

$$S^2 = MSE = \frac{RSS}{n-p-1},$$

where $RSS = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y} = \hat{\mathbf{e}}^T\hat{\mathbf{e}}$.

Theorem 5.1. S^2 is an unbiased estimate of σ^2.

Proof. We have

$$\mathbb{E}[S^2] = \frac{1}{n-p-1}\mathbb{E}[RSS] = \frac{1}{n-p-1}\mathbb{E}[\hat{\mathbf{e}}^T\hat{\mathbf{e}}|\mathbf{x}].$$

Since $\hat{\mathbf{e}} = (\mathbf{I}-\mathbf{H})\mathbf{e}$,

$$\mathbb{E}[S^2] = \frac{1}{n-p-1}\mathbb{E}[\mathbf{e}^T(\mathbf{I}-\mathbf{H})\mathbf{e}|\mathbf{x}] = \frac{1}{n-p-1}(\mathbb{E}[\mathbf{e}^T\mathbf{e}|\mathbf{x}] - \mathbb{E}[\mathbf{e}^T\mathbf{H}\mathbf{e}|\mathbf{x}]).$$

We know $\mathbb{E}[\mathbf{e}^T\mathbf{e}|\mathbf{x}] = \sum_{i=1}^{n}\mathbb{E}[e_i^2|\mathbf{x}] = n\sigma^2$ and

$$\mathbb{E}[\mathbf{e}^T\mathbf{H}\mathbf{e}|\mathbf{x}] = \mathbb{E}\left[\sum_{i=1}^{n} e_i^2 h_{ii} + \sum_{i\neq j} e_i e_j h_{ij}\middle|\mathbf{x}\right] = \mathbb{E}\left[\sum_{i=1}^{n} e_i^2 h_{ii}\middle|\mathbf{x}\right] = (p+1)\sigma^2.$$

Therefore, $\mathbb{E}[S^2] = \frac{1}{n-p-1}(n\sigma^2 - (p+1)\sigma^2) = \sigma^2$. □

5.2.6 Coefficient of Multiple Determination

Let

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\mathbf{Y}^T(\mathbf{H} - \frac{1}{n}\mathbf{J})\mathbf{Y}}{\mathbf{Y}^T(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y}}.$$

R^2 gives the percentage of variation in Y explained by the model with all the p predictors and is not the square of a correlation.

R^2 increases as p increases since TSS remains the same, $RegSS$ stays the same or increases, and RSS stays the same or decreases. Hence R^2 is not helpful in telling whether additional predictors are useful for explaining the response.

Therefore, we define adjusted R^2 :

$$R^2_{\text{adj}} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} = 1 - (n-1)\frac{MSE}{TSS},$$

where $MSE = S^2 = \frac{RSS}{n-p-1}$ is an unbiased estimate of σ^2. R^2_{adj} is adjusted for the number of predictors in the model and is always less than R^2.

5.2.7 Global F-Test

We test

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \text{ against } H_a : \text{At least one } \beta_j \text{ is not } 0,$$

i.e., to test whether there is a linear association between Y and all predictors. The test statistic is

$$F = \frac{MRegSS}{MSE} = \frac{RegSS/p}{RSS/(n-p-1)} \overset{H_0}{\sim} F_{(p,n-p-1)}.$$

5.2.8 Partial F-Test

We test

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \text{ against } H_a : H_0 \text{ is not true},$$

where $k < p$, i.e., to test the added predictive value of $X_1, \cdots, X_k$ over and above the other predictors or to test whether the unexplained variation reduced by a significant amount when the predictors are added to the model. The test statistic is

$$F = \frac{(RSS_{\text{reduced}} - RSS_{\text{full}})/(df_{\text{reduced}} - df_{\text{full}})}{MSE_{\text{full}}} = \frac{(RSS_{\text{reduced}} - RSS_{\text{full}})/k}{RSS_{\text{full}}/(n-p-1)} \overset{H_0}{\sim} F_{(k,n-p-1)}.$$

In general, $TSS_{\text{full}} = TSS_{\text{reduced}}, RegSS_{\text{full}} \geqslant RegSS_{\text{reduced}}, RSS_{\text{full}} \leqslant RSS_{\text{reduced}}$. Notice that

$$RSS_{\text{full}} = \min_{\beta \in \mathbb{R}^{p+1}} \left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right)^T \left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right)$$

and

$$RSS_{\text{reduced}} = \min_{\beta \in \mathbb{R}^{k+1}} \left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right)^T \left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right).$$

The minimum in a larger dimensional space is always the same or less.

The order in which individual predictors are added to the model is important.

5.2.9 Analysis of Covariance

Consider the situation where we want to model a response variable based on both quantitative and qualitative variables.

Coincident regression line: the simplest model in the given situation is one in which the dummy variable has no effect on Y :

$$Y = \beta_0 + \beta_1 x + e.$$

Parallel regression lines: the dummy variable d produces only an additive change on Y :

$$Y = \beta_0 + \beta_1 x + \beta_2 d + e.$$

Unrelated regression lines: the dummy variable d changes the size of the effect of x on Y :

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (d \times x) + e.$$

Note that if significant interaction appears, then it is difficult to talk about effects of an individual predictor variable sin it depends on another. If no significant interaction appears, remove the interaction term(s), and use the additive model to talk about individual effects of predictors.

5.3 Inference for a Single Regression Coefficient

We want to test

$$H_0 : \beta_j = 0 \text{ against } H_a : \beta_j \neq 0.$$

The test statistic is

$$t = \frac{\widehat{\beta}_j}{\mathrm{SE}\left(\widehat{\beta}_j\right)} \overset{H_0}{\sim} t_{(n-p-1)}.$$

This test gives an indication of whether or not X_j contributes to the prediction of the response variable over and above all the other predictors and is a special case of the partial F-test with $k = 1$.

The confidence interval for β_j is

$$\left[\widehat{\beta}_j \pm t_{\frac{\alpha}{2}(n-p-1)\mathrm{SE}\left(\widehat{\beta}_j\right)}\right],$$

where $\widehat{\beta}_j$ is the unbiased estimator of β_j.

5.4 Global F-Test and Individual t-Test

- If the global F-test is significant:

  (1) All or some of the t-tests are significant, then there are some useful explanatory variables for predicting Y.

  (2) All the t-tests are not significant, then this is an indication of multicollinearity, which implies that individual X's do not contribute to the prediction of Y over and above other X's.

- If the global F-test is not significant:

  (1) All the t-tests are not significant, then none of the X's contribute to the prediction of Y.

  (2) Some of the t-tests are significant, then

  (i) The model has no predictive ability. Likely, if there are many predictors, there are type 1 errors in the t tests.

  (ii) The predictors are poorly chosen. The contribution of one useful predictor among many poor ones may note be enough for the model to be significant.

5.5 General Comments

- The regression equation gives the mean response for each combination of explanatory variables.
- The regression equation will not be useful if it is very complicated or a function of a large number of explanatory variables.
- We generally want a model that is as simple as possible to adequately describe the response variable.

6 Diagnostics and Transformations for Multiple Linear Regression

6.1 Multicollinearity

Multicollinearity occurs when explanatory variables are highly correlated. In this case, it is difficult to measure the individual influence of one of the predictors on the response and the fitted equation is unstable. The estimated regression coefficients vary widely from data set to data set, depending on which predictor is included in the model, and may even have opposite sign than what is expected.

When some X's are perfectly correlated ,we cannot estimate β because $\mathbf{X}^T\mathbf{X}$ is singular. Even if $\mathbf{X}^T\mathbf{X}$ is close to singular, its determinant will be close to zero and the standard errors of estimated coefficients will be large.

6.1.1 Variance Inflation Factors

For the general multiple regression model

$$Y = \beta_0 + \beta 1 X_1 + \cdots + \beta_p X_p + e,$$

we have

$$\text{Var}\left[\widehat{\beta}_j\right] = \frac{\sigma^2}{(n-1)S^2_{X_j}(1-R^2_j)}, j = 1, \cdots, p,$$

where R^2_j is the value of R^2 from the regression of X_j on the other X's.

The jth variance inflation factor is

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

A commonly used cut-off is 5 and the larger, the worse.

6.2 Leverage Points

We classify the ith point as a leverage point in MLR model with p predictors if

$$h_{ii} > 2 \times \text{average}(h_{ii}) = \frac{2(p+1)}{n}.$$

6.3 Validity

When a valid model is fit, a plot of standardized residuals against any predictor or any linear combination of the predictors (such as the fitted values) will have the following features: (1) a random scatter of points around the horizontal axis, since the mean function of e_i is zero when a correct model is fit; (2) constant variability as we look along the horizontal axis.

Hence, any pattern in plot of standardized residuals in indicative that an invalid model is fit to the data. Note that any nonrandom pattern itself does not provide direct information on how the model is misspecified.

6.4 Box-Cox Transformation

The Box-Cox transformation is a general method for transforming a strictly positive (response or predictor) variable and aims to find transformation that makes the transformed variable close to normally distributed. It considers a family of power transformations and is based on maximizing a likelihood function. For example,

Power	Transformation
1	No transformation
0	Natural log
0.5	Square root
-1	Inverse

6.5 Added Variable Plot

Suppose our current model is

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

and we are considering the introduction of an additional predictor variable $\mathbf{Z}$, that is, our new model is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\alpha + \mathbf{e}.$$

The added-variable plot is obtained by plotting the residuals from $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ against the residuals from the model

$$\mathbf{Z} = \mathbf{X}\delta + \mathbf{e}.$$

With added variable plot, we can visually assess the effect of each predictor ($\mathbf{Z}$), having adjusted for the effects of the other predictors, visually estimate α, and identify points which have undue influence on the least squares estimate of α.

7 Variable Selection

Variable selection methods aim to choose the subset of explanatory variables that is best in a given sense. Overfitting occurs when too many predictors are in the final regression model and the opposite of overfitting is underfitting.

In general, there is a bias-variance trade-off: when we add more predictors to a valid model, the bias of the predictions gets smaller, but the variance of the estimated coefficients gets larger.

7.1 Information Criteria

7.1.1 Likelihood-Based Criteria

Suppose that $y_i, x_{1i}, \cdots, x_{pi}, i = 1, \cdots, n$ are the observed values of normal random variables and

$$y_i | x_{1i}, \cdots, x_{pi} \sim \mathcal{N}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \sigma^2).$$

Thus the conditional density of y_i given $x_{1i}, \cdots, x_{pi}$ is given by

$$f(y_i | x_{1i}, \cdots, x_{pi}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}))^2}{2\sigma^2} \right).$$

Assuming that the n observations are independent, then the likelihood function of the unknown parameters $\beta_0, \beta_1, \cdots, \beta_p, \sigma^2$ given Y is given by

$$\begin{aligned} L(\beta_0, \beta_1, \cdots, \beta_p, \sigma^2 | Y) &= \prod_{i=1}^{n} f(y_i | x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}))^2}{2\sigma^2} \right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}))^2 \right). \end{aligned}$$

The log-likelihood function is given by

$$l(\beta_0, \beta_1, \cdots, \beta_p, \sigma^2 | Y) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}))^2.$$

The MLEs of $\beta_0, \beta_1, \cdots, \beta_p$ can be obtained by minimizing the third term only, which is equivalent to minimizing the residual sum of squares RSS. Hence, MLEs of $\beta_0, \beta_1, \cdots, \beta_p$ are equal to the least squares estimates.

Therefore,

$$l\left(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p, \sigma^2 | Y \right) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} RSS.$$

Solving for the MLE of σ^2, we have

$$\sigma^2_{\text{MLE}} = \frac{RSS}{n},$$

which differs slightly from unbiased estimate of σ^2. Hence,

$$l\left(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p, \widehat{\sigma}^2 | Y\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\frac{RSS}{n} - \frac{n}{2}.$$

7.1.2 Akaike's Information Criterion

We define

$$AIC = 2\left[-l\left(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p, \widehat{\sigma}^2 | Y\right) + K\right],$$

where $K = p + 2$ is the number of parameters in the fitted model. Note that $(p + 1) + 1$ means $p + 1$ β's and σ^2. We write it as

$$AIC = 2\left(\frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln\frac{RSS}{n} + \frac{n}{2} + p + 2\right)$$

and in R, we only calculate the term related to p, i.e.,

$$AIC = n\ln\frac{RSS}{n} + 2p.$$

The smaller the value of AIC, the better the model. When the sample sample size is small, or when the number of parameters estimated is a moderate to large fraction of the sample size, AIC tends to overfit since the penalty for model complexity is not strong enough.

7.1.3 Bayesian Information Criterion

We define

$$BIC = -2l\left(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p, \widehat{\sigma}^2 | Y\right) + K\ln n,$$

where $K = p + 2$. We write it as

$$BIC = n\ln(2\pi) + n\ln\frac{RSS}{n} + p\ln n + 2\ln n$$

and in R, we only calculate the term related to p, i.e.,

$$BIC = n\ln\frac{RSS}{n} + p\ln n.$$

The smaller the value of BIC, the better the model. When $n \geqslant 8, \ln n \geqslant 2$, and the penalty term in BIC is greater than that in AIC - BIC penalizes complex model more heavily than AIC and thus favors simpler models than AIC.

7.1.4 Data Strategy

A popular data strategy is to calculate R^2_{adj}, AIC, corrected AIC and BIC, and compare the models which minimize AIC, corrected AIC and BIC with the model that maximizes R^2_{adj}.

7.2 Stepwise Regression

If there are k terms that can be added to the mean function apart from the intercept, then there are 2^k possible regression equations.

Backward elimination starts with all the potential terms in the model, then removes the term with the largest p-value each time to give a smaller information criterion.

Forward selection starts with no term in the model, then adds one term at a time with the smallest p-value until no further terms can be added to produce a smaller information criterion. Backward elimination and forward selection considers at most $\frac{k(k+1)}{2}$ of the 2^k possible predictor subsets.

Stepwise regression alternates forward steps with backward steps. It can consider more subsets than the backward or forward methods. The idea is to end up with a model where no variables are redundant given the other variables in the model. Selection overstates significance: estimates of regression coefficients are biased; p-values from F and t tests are generally smaller than their true values.

7.3 Penalized Linear Regression

Penalized linear regression performs variable selection and regression coefficient estimation simultaneously. We solve a constrained least squares optimization problem

$$\min_{\widehat{\beta}_p} \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}^T \mathbf{x}_i \right)^2 + \sum_{j=1}^{p} p_\lambda(\cdot),$$

where $p(\cdot)$ is the penalty function and $\lambda \geqslant 0$ is the penalty parameter. When $\lambda = 0$, the solution is the least squares estimates.

When $p_\lambda = \lambda \widehat{\beta}_j^2$, we call it ridge regression. When $p_\lambda = \lambda \left| \widehat{\beta}_j \right|$, we call it the lasso.

7.4 Cross Validation

K-Fold cross validation is a standard approach to assess the predictive ability of models by evaluating their performance on a new data set.

- Step 1: Divide the data randomly into k roughly equal sets.
- Step 2: Establish the model by using all but one of the k folds, and this set is called the training data set.
- Step 3: Use the remaining data set, called the test data to evaluate the model.
- Step 4: Repeat step 3 and 4 k times by changing the kth fold.
- Step 5: Calculate cross validation error by finding the average of the squared differences between the response and fitted values for the test set. Good candidate subsets will have small cross validation errors.

Note that if $k = n$, we call it leave-one-out cross validation (LOOCV).

Influential points will dramatically affect the results. Splitting the data randomly into training set and test set does not work well in small data sets.