Model: \( Y = X\beta + \epsilon \)
- \( Y \) random vector of responses (we use \( y \) for the observed value of \( Y \))
- \( \epsilon \) random “error” vector
- \( X \) is the model matrix, each column is an explanatory variable, this is an \( n \times (k + 1) \) matrix (we assume the columns of \( X \) are linearly independent)
- \( \beta \) the vector of (unknown) coefficients (parameters)

Fitted model: \( y = X\hat{\beta} + e \)
- \( y \) observed values of response variable
- \( \hat{\beta} \) fitted values of coefficients, (this could either be an estimator or an estimate, notation is ambiguous)
- \( e \) residual vector (defined by this equation)
- \( X\hat{\beta} \) also called \( \hat{y} \), the fitted values of the response variable

“Least squares estimate” and its properties:

1. \( e = y - \hat{y} = y - X\hat{\beta} \) is orthogonal to the column space of \( X \).
2. \( y - \bar{y} = (\hat{y} - \bar{y}) + e \) is a decomposition of \( y - \bar{y} \) into two orthogonal vectors.
3. \(|y - \bar{y}|^2 = |\hat{y} - \bar{y}|^2 + |e|^2 \) (Pythagorean Theorem of Statistics)

Some important linear algebra facts:

1. The length of a vector \( u \) satisfies \(|u|^2 = u \cdot u = u^T u \)
2. Two vectors \( u \) and \( v \) are orthogonal if and only if \( u \cdot v = 0 \) if and only if \( u^T v = 0 \).

The “Hat” Matrix

\[
H = X \left( X^T X \right)^{-1} X^T
\]

Properties:

1. \( \hat{y} = Hy \)
2. \( H \) is symmetric \((H^T = H)\) and idempotent \((H^2 = H)\).
3. \( e = (I - H)y \)

Definitions: If \( Z \) is a random vector (with components \( Z_i \)),

1. \( E(Z) = \mu_Z \) is the vector with components \( E(Z_i) \)
2. \( \text{Cov}(Z) \) is the matrix with the \( i, j^{th} \) entry equal to \( \text{Cov}(Z_i, Z_j) \). (Note that the \( i^{th} \) diagonal entry of \( \text{Cov}(Z) \) is the variance of \( Z_i \).)

Note: \( \text{Cov}(Z) = E[(Z - \mu_Z)(Z - \mu_Z)^T] \).
Estimating $\beta$.

1. Assume $E[\epsilon] = 0$. Then $\hat{\beta}$ is an unbiased estimator of $\beta$.

$$\hat{\beta} = \left( X^T X \right)^{-1} X^T Y = \left( X^T X \right)^{-1} X^T (X\beta + \epsilon) = \beta + \left( X^T X \right)^{-1} X^T \epsilon$$

So

$$E[\hat{\beta}] = \beta + E[\left( X^T X \right)^{-1} X^T \epsilon] = \beta$$

2. Assume that $\text{Cov}(\epsilon) = \sigma^2 I$. Then $\text{Cov}(\hat{\beta}) = \sigma^2 \left( X^T X \right)^{-1}$.

$$\text{Cov}(\beta) = E[(\beta - \hat{\beta})(\beta - \hat{\beta})^T] = E \left[ \left( X^T X \right)^{-1} X^T \epsilon \epsilon^T X \left( \left( X^T X \right)^{-1} \right)^T \right]$$

Estimating $\sigma^2$.

1. $E(\epsilon) = 0$.

2. $\text{Cov}(\epsilon) = \sigma^2 (I - H)$

3. $\text{Var}(\epsilon_i) = \sigma^2 (1 - h_{ii})$

4. $\hat{\sigma}^2 = S^2 = \frac{|\epsilon|^2}{n - (k + 1)} = \frac{\text{SSE}}{n - (k + 1)}$ is an unbiased estimator of $\sigma^2$.

The standard, distributional assumption: $\epsilon_i$ are independent, normal with variance $\sigma^2$. Then

1. $\frac{(n - (k + 1))S^2}{\sigma^2} \sim \text{Chisq}((n - (k + 1)))$

2. $\hat{\beta}, \hat{Y}$, and $\epsilon$ all have normal distributions.

3. Let $W = \left( X^T X \right)^{-1}$ and $w_i = W_{ii}$. Let $\text{SE}(\hat{\beta}) = Sw_i$. Then

$$\frac{\hat{\beta}_i - \beta_i}{\text{SE}(\hat{\beta}_i)} \sim t(n - (k + 1))$$