Brief Notes #11
Linear Regression

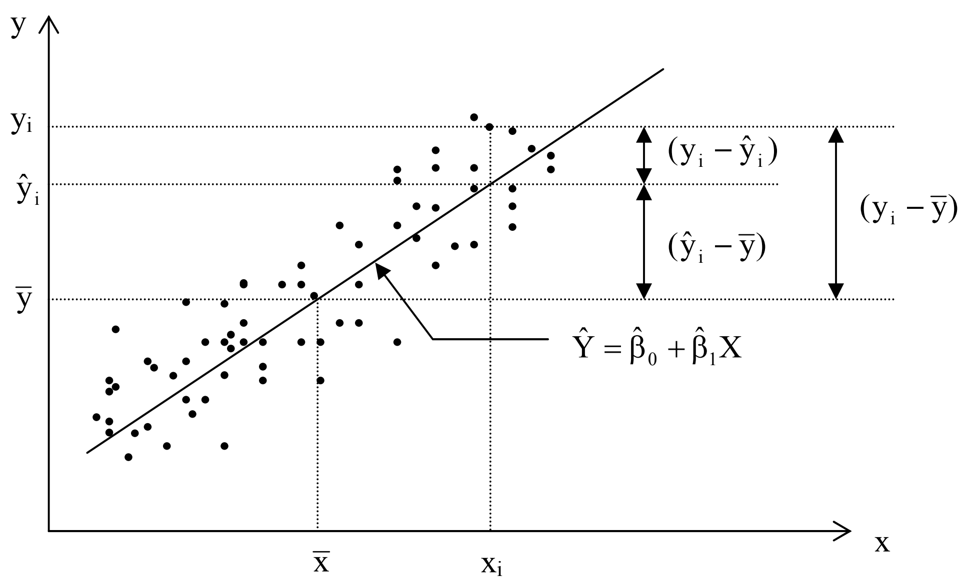
(a) Simple Linear Regression

- **Model:** $Y = \beta_0 + \beta_1\, g(X) + \varepsilon$, with $\varepsilon \sim (0,\sigma^2)$

  $\Rightarrow Y = \beta_0 + \beta_1 X + \varepsilon$

- **Data:** (X_i, Y_i), with $i = 1, \ldots, n$

  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



- **Least Squares Estimation of (β_0, β_1)**

  Find $(\hat{\beta}_0, \hat{\beta}_1)$ such that $\sum_i (Y_i - \hat{Y}_i)^2 = \min$, where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

- **Solution**

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = \overline{Y} - \frac{S_{XY}}{S_{XX}} \overline{X}, \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}},$$

where $\overline{X} = \frac{1}{n} \sum_i X_i$, $\quad \overline{Y} = \frac{1}{n} \sum_i Y_i$

$$S_{XX} = \sum_i (X_i - \overline{X})^2, \qquad S_{XY} = \sum_i (X_i - \overline{X})(Y_i - \overline{Y}).$$

- **Properties of $\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$ for $\varepsilon_i \sim$ iid $N(0, \sigma^2)$**

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \ \sigma^2 \begin{bmatrix} \left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}} \right) & \left(-\frac{\overline{X}}{S_{XX}} \right) \\ \left(-\frac{\overline{X}}{S_{XX}} \right) & \left(\frac{1}{S_{XX}} \right) \end{bmatrix} \right)$$

- **Properties of Residuals,** $e_i = Y_i - \hat{Y}_i$

- $\sum_i e_i = \sum_i e_i X_i = \sum_i e_i Y_i = 0$
- $SS_e = \sum_i (Y_i - \hat{Y}_i)^2$ = the residual sum of squares.

$$\frac{SS_e}{\sigma^2} \sim \chi^2_{n-2}, \qquad E[SS_e] = \sigma^2(n-2)$$

$$\Rightarrow \quad \hat{\sigma}^2 = SS_e/(n-2) = MS_e \text{ (mean square error)}$$

$$\hat{\sigma} = \sqrt{MS_e} = \text{"standard error of regression"}$$

- **Significance of Regression**

Let $S_{YY} = \sum_i (Y_i - \overline{Y})^2$ = <u>total sum of squares</u>.

<u>Property:</u> $S_{YY} = SS_e + SS_R$,

where $SS_e = \sum_i (Y_i - \hat{Y}_i)^2$ = residual sum of squares,

$SS_R = \sum_i (\hat{Y}_i - \overline{Y})^2$ = sum of squares explained by the regression.

Also SS_e and SS_R are statistically independent.

Notice: if $\beta_1 = 0$, then $\dfrac{SS_R}{\sigma^2} \sim \chi_1^2$

Definition:

$$R^2 = \frac{SS_R}{S_{YY}} = 1 - \frac{SS_e}{S_{YY}}, \quad \text{coefficient of determination of the regression.}$$

- **Hypothesis Testing for the Slope β_1**

1. H_0: $\beta_1 = \beta_{1_0}$ against H_1: $\beta_1 \neq \beta_{1_0}$ **(t-test)**

Property: $t(\beta_1) = \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_e / S_{XX}}} \sim t_{n-2}$

$\Rightarrow$ Accept H_0 at confidence level α if:

$$|\, t(\beta_{1_0})\,| < t_{n-2,\alpha/2}$$

2. H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$ **(F-test)**

From distributional properties and independence of SS_R and SS_e, and under H_0,

$$F = \frac{SS_R / 1}{SS_e / (n-2)} \sim F_{1,n-2}$$

$\Rightarrow$ Accept H_0 if $F < F_{1,\, n-2,\, \alpha}$

Notice that for H_0: $\beta_1 = 0$, the t-test and the F-test are equivalent.

(b) Multiple Linear Regression

- **Model:** $Y = \beta_0 + \sum_{j=1}^{k} \beta_j g_j(\underline{X}) + \varepsilon$, with $\varepsilon \sim (0, \sigma^2)$

  $$\Rightarrow\; Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon$$

- **Data:** $(Y_i, \underline{X}_i)$, with i = 1, … , n

  $Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i$, with i = 1, … , n

  Let $\underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$, $\quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$, $\quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, $\quad \underline{H} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}$

  $$\Rightarrow \underline{Y} = \underline{H}\,\underline{\beta} + \underline{\varepsilon}$$

- **Least Squares Estimation**

  $$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j X_{ij}$$

  $$\underline{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \underline{Y} - \underline{H}\underline{\hat{\beta}}$$

  $$SS_e = \sum_i e_i^2 = \sum_i \underline{e}^T \underline{e} = (\underline{Y} - \underline{H}\underline{\hat{\beta}})^T (\underline{Y} - \underline{H}\underline{\hat{\beta}}) = \underline{Y}\underline{Y}^T - 2\underline{Y}^T \underline{H}\underline{\hat{\beta}} + \underline{\hat{\beta}}^T \underline{H}^T \underline{H}\underline{\hat{\beta}}$$

  $$\frac{dSS_e(\underline{\beta})}{d\underline{\beta}} = \underline{0} \qquad \Rightarrow \qquad \underline{\hat{\beta}} = (\underline{H}^T \underline{H})^{-1} \underline{H}^T \underline{Y}$$

- **Properties of $\hat{\underline{\beta}}$** (if $\varepsilon_i \sim$ iid $N(0, \sigma^2)$)

$$\hat{\underline{\beta}} \sim N(\underline{\beta}, \sigma^2 (\underline{H}^T \underline{H})^{-1})$$

- **Properties of Residuals**

$$\frac{SS_e}{\sigma^2} \sim \chi^2_{n-k-1}$$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{SS_e}{n-k-1} = MS_e$$

$$S_{YY} = SS_e + SS_R$$

$$R^2 = 1 - \frac{SS_e}{S_{YY}}$$

- **Hypothesis Testing**

Let $\underline{\beta} = \begin{bmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{bmatrix}$, where $\underline{\beta}_1$ has τ_1 components and $\underline{\beta}_2$ has $\tau_2 = k - \tau_1$ components. We want to test H_0: $\underline{\beta}_2 = \underline{0}$ against H_1: $\underline{\beta}_2 \neq \underline{0}$ (at least one component of $\underline{\beta}_2$ is non-zero).

The procedure is as follows:

- Fit the complete regression model and calculate SS_R and SS_e;

- Fit the reduced model with $\underline{\beta}_2 = \underline{0}$, and calculate SS_{R_1};

- Let $SS_{2|1} = SS_R - SS_{R_1}$ = extra sum of squares due to $\underline{\beta}_2$ when $\underline{\beta}_1$ is in the regression.

- Distributional property of $SS_{2|1}$. Under H_0,

$$\frac{SS_{2|1}}{\sigma^2} \sim \chi^2_{\tau_2}$$

Also, $SS_{2|1}$ and SS_e are independent. Therefore,

$$F = \frac{SS_{2|1} / \tau_2}{SS_e / (n-k-1)} \sim F_{\tau_2, n-k-1}$$

$$\Rightarrow \text{Accept } H_0 \text{ if } F < F_{\tau_2, n-k-1, \alpha}$$