SLR

Distribution of the data points:

$$y_i \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

LSE of $\hat{\beta}_0$ & $\hat{\beta}_1$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x}) y_i}{S_{xx}} = \frac{S_{xy}}{S_{xx}}$$

① Distn:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{n S_{xx}}\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\frac{\hat{\beta}_i - \beta_i}{se(\beta_i)} \sim t_{n-2}$$

$$se(\beta_i) \to Var(\beta_i)\big|_{\sigma^2 \leftrightarrow s^2}$$

② Covariance:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{S_{xx}}$$

Residuals:

$$r_i = y_i - \hat{y}_i$$

Properties:

① $\sum r_i = 0$

② $\sum r_i x_i = 0$

③ $\sum r_i \hat{y}_i = 0$

Variance estimator (unbiased):

$$s^2 = \frac{1}{n-2} \sum r_i^2$$

Mean response estimation

$$\hat{\mu} \sim N\left(\beta_0 + \beta_1 x_p, \sigma^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]\right)$$

$$\frac{\hat{\mu}-\mu}{se(\hat{\mu})} \sim t_{n-2}$$

Prediction

$$y_p - \hat{y}_p \sim N\left(0,\ \sigma^2\left[1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}\right]\right)$$

ANOVA

① Overall rel.

$$\underbrace{SST}_{\sum(y_i-\bar{y})^2} = \underbrace{SSE}_{\sum(y_i-\hat{y}_i)^2} + \underbrace{SSR}_{\sum(\hat{y}_i-\bar{y})^2}$$

② Table

SS	d.f.	MS	E
SSE	n-2	SSE/n-2	σ^2
SSR	1	SSR	$\sigma^2+\beta_1^2 S_{xx}$
SST	n-1	SST/n-1	

③ F distn.

$$\frac{MSR}{MSE} \sim F_{1,n-2}$$

④ R^2:

$$R^2 = \frac{SSR}{SST}$$

MLR

Laws of random vectors:

① Expectation

$$E[X] = \begin{bmatrix} \vdots \\ E(x_i) \\ \vdots \end{bmatrix} \qquad E[AX+B] = AE[X] + B$$

② Variance:

$$Var(X) = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_n) \\ & \ddots & & \\ & & \ddots & \\ & & & Var(X_n) \end{bmatrix} \qquad Var[AX+B] = A\,Var[X]A'$$

Linear model formulation

$$Y = X\beta + \varepsilon$$

$$\varepsilon \sim MVN(0, \sigma^2 I) \implies Y \sim MVN(X\beta, \sigma^2 I)$$

MVN Properties

1. $Y \sim MVN(\mu, \Sigma) \Rightarrow AY + B \sim MVN(A\mu + B, A'\Sigma A)$

2. $Y \sim MVN(\mu, \Sigma) \Rightarrow y_i \sim N(\mu_i, \Sigma_{ii})$

3. $Y \sim MVN(\mu, \Sigma) \Rightarrow$ partitions of $Y_1, Y_2 \sim MVN(\mu_i, \Sigma_i)$

4. $Var(Y) = \Sigma = diag\{\ldots\} \iff y_i$ is indep.

5. $Y \sim MVN(\mu, \Sigma)$, $W = AY$, $V = BY \Rightarrow V \perp W \iff A'\Sigma B = 0$

Derivative Properties

1. $\frac{\partial}{\partial \beta} A'\beta = A$

2. $\frac{\partial}{\partial \beta} A\beta' = A$

3. $\frac{\partial}{\partial \beta} \beta' A \beta = (A + A')\beta$

LSE of MLR

1. Formula for $\hat{\beta}$:

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

2. Distn:

$$\hat{\beta} \sim MVN(\beta, \sigma^2 (X'X)^{-1})$$

Fitted Values

① Formula:

$$\hat{Y} = X\breve{\beta}$$
$$= \underbrace{X(X'X)^{-1}X'}_{\underline{\underline{H}}}Y$$

② Hat matrix:

Symmetric: $H' = H$ Idempotent: $HH = H$

② Distn:

$$\hat{Y} \sim MVN(X\beta, \sigma^2 X(X'X)^{-1}X')$$

Residuals

$$r = Y - \hat{Y}$$
$$= (I - H)Y$$

① Distn:

$$r \sim MVN(0, \sigma^2(I-H))$$

② Properties:

a) $\sum r_i = 0$

b) $\sum r_i x_i = 0$

c) $\sum r_i \hat{y}_i = 0$

Variance estimate

$$MSE = \frac{1}{n-(p+1)} \cdot SSE = s^2$$

Inference

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[\sigma^2 (X'X)^{-1}_{jj}]}} \sim t_{n-(p+1)}$$

Expected value

$$\hat{\mu}_c \sim MVN(c'\beta, \sigma^2 c'(X'X)^{-1}c)$$

$$\frac{\hat{\mu}_c - \mu_c}{\sqrt{s^2 \, c'(X'X)^{-1}c}} \sim t_{n-(p+1)}$$

Prediction

$$y_p - \hat{y}_p \sim MVN\left(0, \sigma^2 \left[1 + c'(X'X)^{-1}c\right]\right)$$

$$\therefore \frac{y_p - \hat{y}_p}{\sqrt{s^2 \left[1 + c'(X'X)^{-1}c\right]}} \sim t_{n-(p+1)}$$

ANOVA

① Table

Source	SS	d.f.	M.S.
Error	$\sum (y_i - \hat{y}_i)^2$	$n-(p+1)$	$SSE / n-(p+1)$
Regression	$\sum (\hat{y}_i - \bar{y})^2$	p	SSR / p
Total	$\sum (y_i - \bar{y})^2$	$n-1$	

② F-test of overall significance

$H_0: \beta_0 = \beta_1 = \dots = \beta_p = 0$ vs. H_a: one is not 0

Test stat: $$\frac{SSR/p}{SSE/n-(p+1)} = \frac{MSR}{MSE} \sim F_{p,\, n-(p+1)}$$

③ R^2:

Normal (% of variation expl. by model) = $\frac{SSR}{SST}$

Adjusted:

$$R^2 = 1 - \frac{n-1}{n-(p+1)} \cdot (1 - R^2)$$

SPECIFICATION ISSUES

Categorical variables

① Model

$$y_i = \beta_0 + \beta_1 x_1 + \underbrace{\beta_2 x_2 + \beta_3 x_3 + \dots}_{\text{Indicator variables (\# of cat. - 1)}}$$

↳ Avg. diff in response b/w baseline & category

② Testing:

Obj: diff in response / category

A: $H_0: \beta_i = 0$ (if against baseline) $\Rightarrow$ t-test

B: $H_0: \beta_i = \beta_j \Rightarrow \beta_i - \beta_j = 0$ (if not against baseline)

$$\therefore \quad t = \frac{\hat{\beta}_i - \hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \left(Var(\hat{\beta}_i) + Var(\hat{\beta}_j) - 2Cov(\hat{\beta}_i, \hat{\beta}_j) \right)}}$$

Interaction effects

① Model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

② Test:

Obj: is relationship b/w y & x_1 diff. between diff. values of x_2?

$H_0: \beta_3 = 0$ vs. $H_a: \beta_3 \neq 0 \Rightarrow$ t-test

GENERAL LINEAR HYPOTHESES

$$\frac{(SSE_A - SSE)/l}{SSE/n-p-1} \sim F_{l, n-(p+1)}$$

SSE of reduced model if H_0 true

RESIDUAL ANALYSIS

Types of residuals

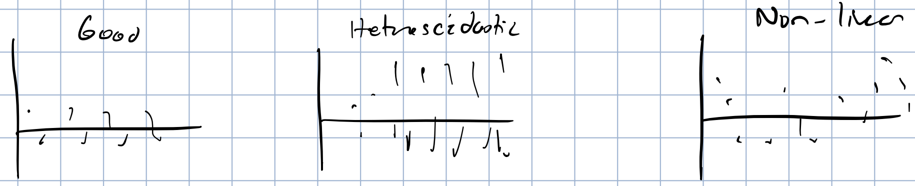
① Raw: $r_i = y_i - \hat{y}_i$

② Standardized: $d_i = \dfrac{r_i}{\hat{\sigma}}$

③ Studentized: $e_i = \dfrac{r_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}$

Plots:

① Residuals vs. fitted values

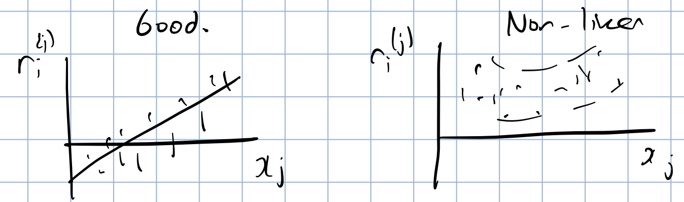


② Residuals vs. expl. variable

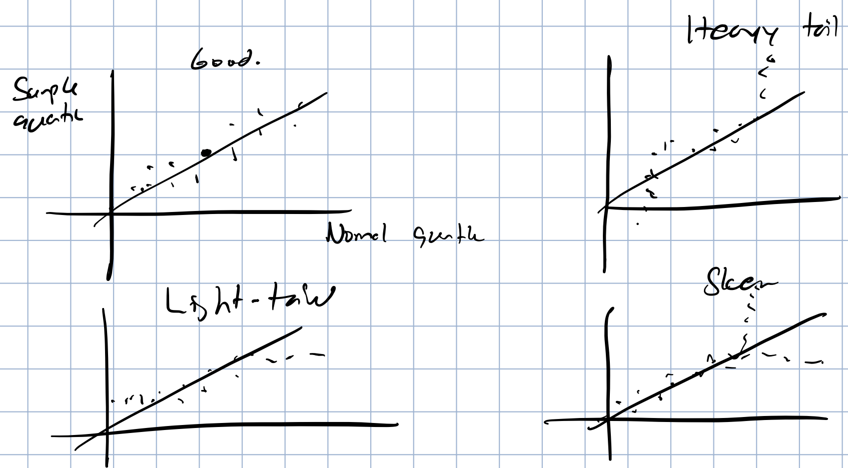
SAME

③ Partial residual

$$r_i^{j} = r_i - \hat{\beta}_j x_j$$



④ Q-Q



Correction

① Raw correction:

$$Var(y_i) = h(\mu_i)\cdot\sigma^2 \Rightarrow (g'(\mu_i))^2 \propto \frac{1}{h(\mu_i)}$$

transform

② Box-Cox:

$$g(y_i) = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log y_i, & \lambda = 0 \end{cases}$$

$\lambda = 1/2$, $\lambda = 1$, $\lambda = 0$, $\lambda = -1$

EFFECTS OF INDIVIDUAL OBSERVATIONS

Outliers:

(1) In response:

$$|e_i| > 3 \rightarrow \text{obs. } i \text{ is outlier}$$

(2) In explanatory

$$h_{ii} > 2\bar{h} = \frac{2(p+1)}{n} \rightarrow \text{obs. } i \text{ is outlier}$$

Influential Datapoints

(1) Cook's Distance:

$$D_i = e_i \cdot \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{p+1}$$

$$D_i = \begin{cases} > 0.5 \rightarrow \text{might be inf.} \\ > 1 \rightarrow \text{def. inf.} \end{cases}$$

MODEL SELECTION

Selection Criteria

(1) R^2_{adj}:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

$$= 1 - \frac{n-1}{n-k-1}(1-R^2)$$

(2) Akaike Info Criteria:

$$AIC = 2q - 2\ln(L(\hat{\theta}))$$ → lower AIC, better model

(3) Bayesian Info Criteria:

$$BIC = q\ln(n) - 2\ln(L(\hat{\theta}))$$