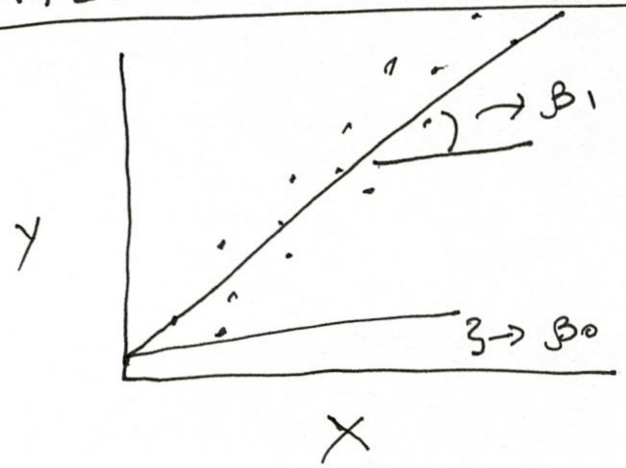
SIMPLE LINEAR REGRESSION



$$\boxed{Y = \beta_0 + \beta_1 x + \epsilon}$$

$\sim N(0, \sigma^2)$

Least squares.

$\hat{\beta}_1 , \hat{\beta}_0$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} , \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

E.g. $S_{xy} = 10$, $S_{xx} = 5$, $\bar{y} = 3$, $\bar{x} = 1$

$\hat{\beta}_1 , \hat{\beta}_0 \to ?$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10}{5} = 2 , \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3 - 2 \times 1 = 1$$

Residuals $e_i = y_i - \hat{y}_i$

$\hat{y}_i \to$ predicted

$y_i \to$ observation (data)

ANOVA

SS_T (Total sum of squares) $= S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$SS_T = SS_R + SS_E$

SS_R: Regression sum of squares

SS_E: Error sum of squares

$SS_R = \sum (\hat{y}_i - \bar{y})^2$

$SS_E = \sum (y_i - \hat{y}_i)^2$

$$R^2 = 1 - \frac{SS_E}{SS_T}$$

R^2: accuracy of the model

closer it is to 1, better the fit

↑ accuracy

Hypothesis Testing

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$$T_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \rightarrow \sqrt{\hat{\sigma}^2 / S_{xx}}$$

Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$

n → no. of observations

Ex.

$H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$

$\hat{\beta}_1 = 2$, $n = 15$, $S_{xx} = 5$, $\hat{\sigma}^2 = 1$

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{2}{\sqrt{1/5}} = 4.47$$

$t_{\alpha/2, n-2} = t_{0.025, 13} = 2.16$

$t_0 > t_{\alpha/2, n-2} \Rightarrow$ Reject H_0

ANOVA

$MS_R = \frac{SS_R}{1}$, $MS_E = \frac{SS_E}{n-p}$

p → (β_0, β_1)

$(\frac{1}{2})$

$$F_0 = \frac{MS_R}{MS_E}$$

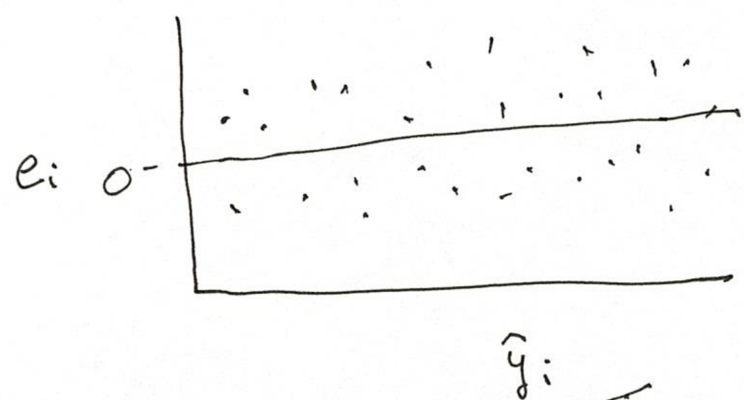
Reject H_0 if $f_0 > f_{\alpha, 1, n-2}$

ADEQUACY OF MODEL

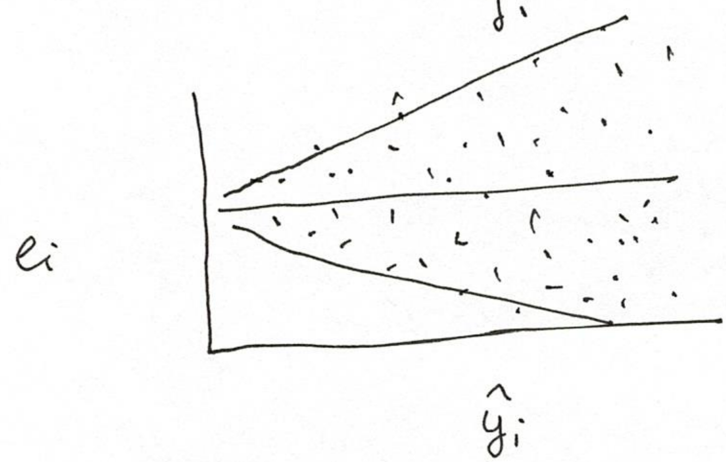
→ Validate assumptions of model

→ $\boxed{e_i \quad (y_i - \hat{y}_i) \longrightarrow x_i, \ \hat{y}_i}$

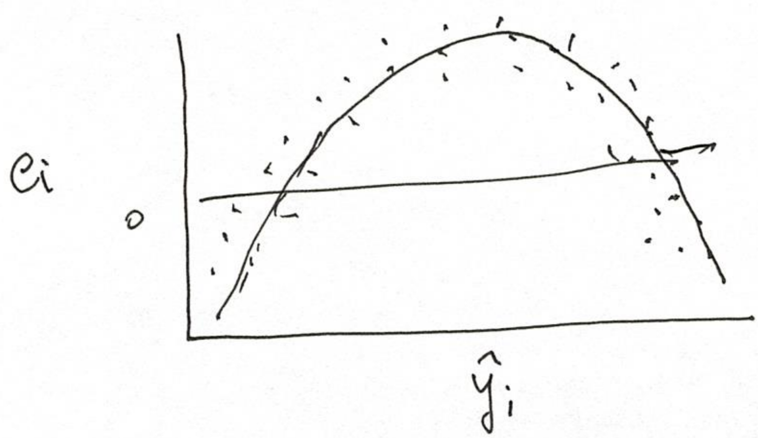
→ Normal probability plot of e_i



Acceptable



Transform variables
eg. transform y to $\sqrt{y}$, $\ln y$, $1/y$



Include higher order terms
x^2, x^3

MULTIPEE REGRESSION

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

$$Y = \beta_0 + \beta_1 \underbrace{x}_{x_1} + \beta_2 \underbrace{x^2}_{x_2} + \beta_3 \underbrace{x^3}_{x_3} + \epsilon$$

$$Y = \underline{\beta_0} + \underline{\beta_1 x_1} + \underline{\beta_2 x_2} + \underbrace{\beta_{12}}_{\beta_3} \underbrace{x_1 x_2}_{x_3} + \epsilon$$

$\downarrow$

Least squares

$R^2 =$

$$\text{Adjusted } R^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

will increase only if addition of a new variable produces a large enough reduction in SS_R to compensate for loss of one residual degree of freedom

INFERENCES

→ Test for significance of regression

$$\boxed{p = k+1}$$

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$H_1:$ At least one $\beta_j \neq 0$

ANOVA

$$MS_R = \frac{SS_R}{k}, \quad MS_E = \frac{SS_E}{n-p}$$

$k \to p-1$

$$f_0 = \frac{MS_R}{MS_E}$$

Reject H_0 if $f_0 > f_{\alpha, k, n-p}$

Individual Regression Coefficients

$H_0: \beta_j = 0 \quad , \quad H_1: \beta_j \neq 0$

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Reject H_0 if $|t_0| > t_{\alpha/2, n-p}$

E.g. $\hat{\beta}_j = 0.5 \quad , \quad se(\hat{\beta}_j) = 5 \quad , \quad n = 15, \quad p = 3$

$t_0 = \frac{0.5}{5} = 0.1 \quad , \quad t_{\alpha/2, n-p} = t_{0.025, 12} = 2.179$

Fail to reject H_0!

Significance of group of regressors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2} + \epsilon$$

$H_0: \beta_{12} = \beta_{11} = \beta_{22} = 0$

H_1: At least one of β_j's $\neq 0$

Full model (FM) : $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \underbrace{\beta_{r+1} x_{r+1} + \cdots + \cdots + \beta_k x_k}_{= 0} + \epsilon$

Reduced model (RM) : $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \epsilon$

$H_0: \beta_{r+1} = \beta_{r+2} = \cdots = \beta_k = 0$

H_1: At least one of β's $\neq 0$

$$F_0 = \frac{[SS_E(RM) - SS_E(FM)] / (k-r)}{SS_E(FM) / (n-p)}$$

$k - r \rightarrow p-1$

$n \rightarrow$ no. of observations

$p \rightarrow k+1$

MODEL ADEQUACY

Standardized residual : $d_i = \dfrac{e_i}{\sqrt{\hat{\sigma}^2}}$ $\quad e_i \rightarrow y_i - \hat{y}_i$

Studentized residual : $r_i = \dfrac{e_i}{\sqrt{\hat{\sigma}^2 (1-h_{ii})}}$ $\quad \rightarrow 0 \leq h_{ii} \leq 1$

$r_i > d_i$

$$
\begin{cases}
\hat{y}_1 = h_{11} y_1 + h_{12} y_2 + \cdots + h_{1n} y_n \\
\hat{y}_2 = h_{21} y_1 + h_{22} y_2 + \cdots + h_{2n} y_n \\
\vdots \qquad\qquad \ddots \\
\qquad\qquad\qquad\qquad + h_{nn} y_n \\
\hat{y}_n = h_{n1} y_1 + \cdots
\end{cases}
$$

$\rightarrow h_{ii} \geq \dfrac{2p}{n} \rightarrow$ leverage point $\leftarrow (x_i)$

Cook's distance measure : $D_i = \dfrac{r_i^2}{p} \dfrac{h_{ii}}{(1-h_{ii})}$

$D_i > 1 \Rightarrow$ influential point

Multicollinearity

Variance Inflation factor (VIF)

$\text{VIF}(\beta_j) = \dfrac{1}{1 - R_j^2}$ $\rightarrow R^2$ from regressing x_j other than x's

VIF > 10

$\rightarrow$ Remove x_j $\quad \rightarrow$ Try centering $\rightarrow x_i \rightarrow (x_i - \bar{x})$

$\rightarrow$ Use something other than least squares.

CATEGORICAL VARIABLES

$$x_i = \begin{cases} 0 \\ 1 \end{cases} \quad \text{(e.g. male/ female)}$$

Indicator variables

Region	x_1	x_2
East	0	0
Midwest	1	0
West	0	1

(e.g. does salary depend on region?)

VARIABLE SELECTION TECHNIQUES

$$C_p = \frac{SS_E\,(p)}{\hat{\sigma}^2\,(FM)} - n + 2p$$

→ Small value of C_p is desirable

[Go over table 6-10]

STEPWISE REGRESSION

→ Backward → start with all regressors and successively eliminate based on t-test

$t < t_{out}$ (cutoff) ⇒ remove regressor

[table 6-11]

→ Forward → start with no variables → add one ~~at~~ at a time

→ variable that results in largest t-stat. inserted as long as it is > threshold t_{in}

[table 6-12]