Simple Linear Regression : Mortality Data

The data, taken from certain regions of Great Britain, Norway, and Sweden contains the mean annual temperature (in degrees F) and mortality index for neoplasms of the female breast.

response y	Mortality rate (M)	102.5	104.5	100.4	95.9	87.0	95.0	88.6	89.2
predictor x	Temperature (T)	51.3	49.9	50.0	49.2	48.5	47.8	47.3	45.1
	Mortality rate (M)	78.9	84.6	81.7	72.2	65.1	68.1	67.3	52.5
	Temperature (T)	46.3	42.1	44.2	43.5	42.3	40.2	31.8	34.0

規律

Objective : Obtaining the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women.

Website of my LM course
http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/

Initial Data Analysis (LM, Lab02) → **Getting Started**

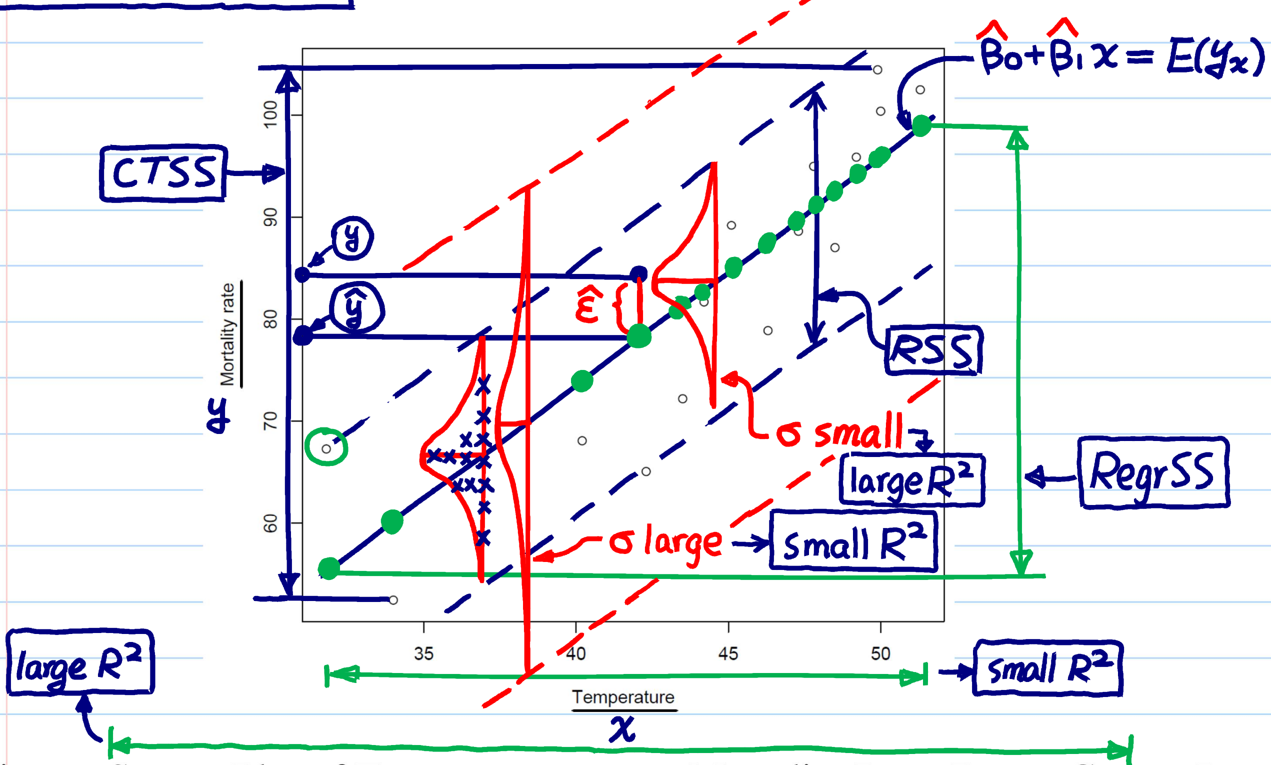


Figure: Scatter Plot of Temperature versus Mortality Rate, Breast Cancer Data.

Fitting the Regression Line

not regarded as r.v., no measurement error

- Underlying Model :

r.v. — r.v. : random component

model matrix $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean structure of } y \text{ : systematic component}} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

- Coefficients are estimated by minimizing

$$\begin{bmatrix} \frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = 0 \\ \frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = 0 \end{bmatrix} \leftarrow \sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2. \equiv S(\beta_0, \beta_1)$$

- **Least Squares Estimates**

$\boxed{\hat{\rho}_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}}$

Estimated Coefficients :

function of x_i's only, irrelevant to $\underset{\sim}{\beta}, \sigma^2$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad var(\hat{\beta}_1) = \frac{\sigma^2 \cdot 1}{\sum (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad , \quad var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right),$$

$$\bar{x} = \frac{1}{N} \sum x_i \quad , \quad \bar{y} = \frac{1}{N} \sum y_i.$$

$\boxed{Cov(\hat{\underset{\sim}{\beta}}) = \sigma^2 (X^T X)^{-1}}$

規律：？%，隨機：？%

Explanatory Power of the Model

Q: How well the model explain the data? "goodness of fit" measure

- The total variation in y can be measured by corrected total sum of squares

$CTSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$.

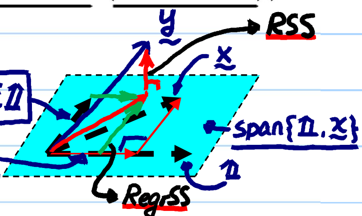
variation in y_i's

- This can be decomposed into two parts (Analysis of Variance (ANOVA)):

source of variation in y_i's

$-\hat{y}_i + \hat{y}_i$

$$CTSS = \underset{\text{規律}}{RegrSS} + \underset{\text{隨機}}{RSS},$$



where

source of variation : x_i's → $RegrSS$ = Regression sum of squares $= \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2, = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2$

$\bar{y}$: mean of $\hat{y}_i$'s

$\hat{\sigma}^2 = \frac{RSS}{n - p}$ ← # of $\underset{\sim}{\beta}$

source of variation : ε_i's → RSS = Residual sum of squares $= \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$.

$\hat{\varepsilon}_i$: residuals

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the predicted value of y_i at x_i.

- $R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS}$ measures the proportion of variation in y explained by the fitted model. ← check the graph in LNp.2-2

ANOVA Table for Simple Linear Regression

↳ decomposition of sum of squares (sources of variation in $\underset{\sim}{y}$)

ANOVA Table for Simple Linear Regression

Source	Degrees of Freedom	Sum of Squares	Mean Squares
source of variation: x_i's → regression	1 (β_1)	$\hat\beta_1^2\sum(x_i-\bar x)^2$ ($RegrSS$)	$\hat\beta_1^2\sum(x_i-\bar x)^2$
source of variation: ε_i's → residual	$N-2$ (β_0,β_1)	$\sum_{i=1}^N(y_i-\hat y_i)^2$ (RSS ← RSS under H_A)	$\frac{\sum_{i=1}^N(y_i-\hat y_i)^2}{(N-2)}$
total (corrected)	$N-1$ (β_0)	$\sum_{i=1}^N(y_i-\bar y)^2$ ($CTSS$ ← RSS under H_0)	

overall F-test
$H_0: y=\beta_0+\varepsilon$
$H_A: y=\beta_0+\beta_1 x+\varepsilon$

R^2

$F=\frac{MS_{RegrSS}}{MS_{RSS}} \sim F_{df_{RegrSS},\, df_{RSS}}$

ANOVA Table for Breast Cancer Example

Source	Degrees of Freedom	Sum of Squares	Mean Squares
regression	1	2599.53	2599.53
residual	14	796.91	56.92
total (corrected)	15 (16 obs.)	3396.44	

$\div \to \sim F_{1,14}$

equivalent to the overall F-test in LN p. 2-5 for simple regression → ***t*-Statistic** → null model = ? alternative model = ?

why interested in it ?

- To test the null hypothesis $H_0:\beta_j=0$ against the alternative hypothesis $H_A:\beta_j\neq 0$ under the full model, use the test statistic

can be changed to other constant a for testing $H_0: \beta_j=a$

pdf of $\hat\beta_j$ pdf of $\hat\beta_j$

$t_j^2=F$

$$t_j=\frac{\hat\beta_j-0}{s.d.(\hat\beta_j)}.$$

$s.d.(\hat\beta_j) \leftarrow \hat\sigma\sqrt{(X^TX)^{-1}_{jj}}$

- The higher the value of $|t_j|$, the more significant is the coefficient. → i.e., more confident to reject $H_0: \beta_j=0$

- For 2-sided alternatives, p-value $=\text{Prob}\left(|t_{df}|>|t_{obs}|\right)$, df = degrees of freedom for the t-statistic, t_{obs} = observed value of the t-statistic. If p-value is very small, then either we have observed something which rarely happens, or H_0 is not true. In practice, if p-value is less then $\alpha=0.05$ or 0.01, H_0 is rejected at level α.

r.v. df_{RSS} under H_0

tests $\xleftrightarrow{\text{duality}}$ Confidence Interval: collection of plausible β's

100(1 − α)% confidence interval for β_j is given by

$$\hat{\beta}_j \pm \underbrace{t_{N-2,\frac{\alpha}{2}}}_{\text{critical value}} \times s.d.(\hat{\beta}_j),$$

where $t_{N-2,\frac{\alpha}{2}}$ is the upper $\alpha/2$ point of the t distribution with $N-2$ degrees of freedom.

∵ 2-sided df_{RSS}

If the confidence interval for β_j does not contain 0, then H_0 is rejected.

$\boxed{\beta_j = 0}$

Test $H_0: \beta_j = \beta_j^*$ vs. $H_A: \beta_j \neq \beta_j^*$

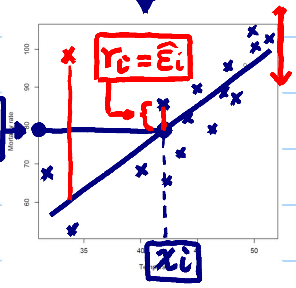
Acceptance region: $|t_j| = \left|\dfrac{\hat{\beta}_j - \beta_j^*}{s.d.(\hat{\beta}_j)}\right| \leq t_{df_{RSS},\frac{\alpha}{2}}$

$$\hat{\beta}_j - t \times s.d.(\hat{\beta}_j) \leq \beta_j^* \leq \hat{\beta}_j + t \times s.d.(\hat{\beta}_j)$$

estimate ± (critical value) × s.d.(estimate)

Predicted Values and Residuals

$y = \hat{y} + \hat{\varepsilon}$



- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value of y_i at x_i.
- $r_i = y_i - \hat{y}_i$ is the corresponding residual.
  or denoted by $\hat{\varepsilon}$

  used in $\hat{\sigma}^2 = \dfrac{RSS}{n-p}$
- [Studentized] Standardized residuals are defined as $\dfrac{r_i}{s.d.(r_i)}$.
- Plots of residuals are extremely useful to judge the "goodness" of fitted model.

  under ② & ③, $\hat{\varepsilon}$ contains more information than error distribution

  $\mathrm{cov}(\hat{\varepsilon}) = (I - H)\sigma^2$
  hat matrix
  H_{ii}: leverage

  future lecture LN p.3-22 ~ 3-25

  - under ① Normal probability plot (will be explained in Unit 3).
  - Residuals versus predicted values.
  - Residuals versus covariate x.
    under ② & ③

$$y = f(x_1, \cdots, x_m) + \varepsilon$$
$$= \hat{y} + \hat{\varepsilon}$$

③ ②

① If model correct,
$\hat{\varepsilon}$: carry information about 隨機 ε (error dist., error variance)
$\hat{y}$: carry information about 規律 f
② overfitting: $\varepsilon \to \hat{y}$
③ lack of fit: $f \to \hat{\varepsilon}$

Analysis of Breast Cancer Data

The regression equation is
M = - 21.79 + 2.36 T

t-test

Predictor	Coef	SE Coef	T	P
Constant	$\hat{\beta}_0 \to$ -21.79	15.67	-1.39	0.186
T	$\hat{\beta}_1 \to$ 2.3577	0.3489	6.76	0.000

s.d.($\hat{\beta}_0$) $H_0: \beta_0 = 0$
s.d.($\hat{\beta}_1$) $H_0: \beta_1 = 0$

null model = ?
alternative model = ?

S = 7.54466 ($\hat{\sigma}$)   R-Sq = 76.5%   R-Sq(adj) = 74.9%

$6.76^2 = 45.67$
(only for simple regression)

Analysis of Variance ← overall F-test

Source	DF	SS	MS	F	P
Regression	1	2599.5	2599.5	45.67	0.000
Residual Error	14	796.9	56.9 ($\hat{\sigma}^2$)		
Total	15	3396.4			

LNp.2-5

Unusual Observations

Obs	T	M	Fit	SE Fit	Residual	St Resid
15	31.8	67.30	53.18	4.85	14.12	2.44RX

check LNp.2-2

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.