Applied linear statistical models

ch. 1 Some basic results in probability & statistics :-

→ Random variables & probability distributions ;

→ statistical estimation

Part I Basic regression analysis :

ch. 2 Linear regression with one independent variable :

Defn. Regression analysis, is a statistical tool that utilizes the relation betn. two or more quantitative variables, so that one variable can be predicted from another.

scope using a single predictor variable for predicting the variable of interest.

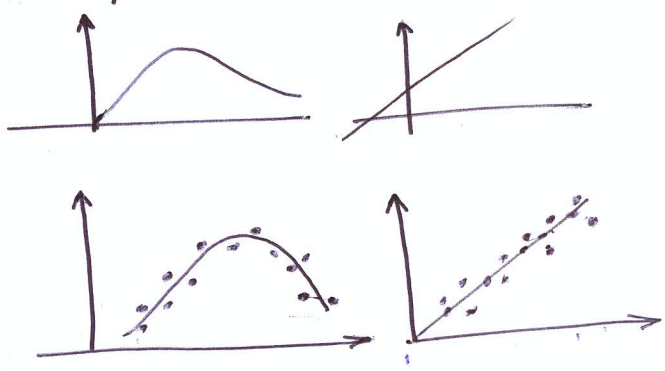
→ Basic ideas of regression analysis & estimation of the parameters of regression model.

2-1 Relation betn two variables :

→ functional relation $y = f(x)$

→ statistical relation

observations



2-2 Regression models & their uses :

Regression model concepts :

1- A tendency of the dependent variable Y to vary with the independent variable(s) X in a systematic fashion.

2- Scattering of observations arround the curve of statistical relationship.

Assuming that :

→ In population of the sampled procesn observations, there is a probability distribution of Y for each level of X. ⇒ (2)

→ Means of prob. distributions vary in some systematic manner with X. ⇒ (1)

relationship ≡ regression function (curve)

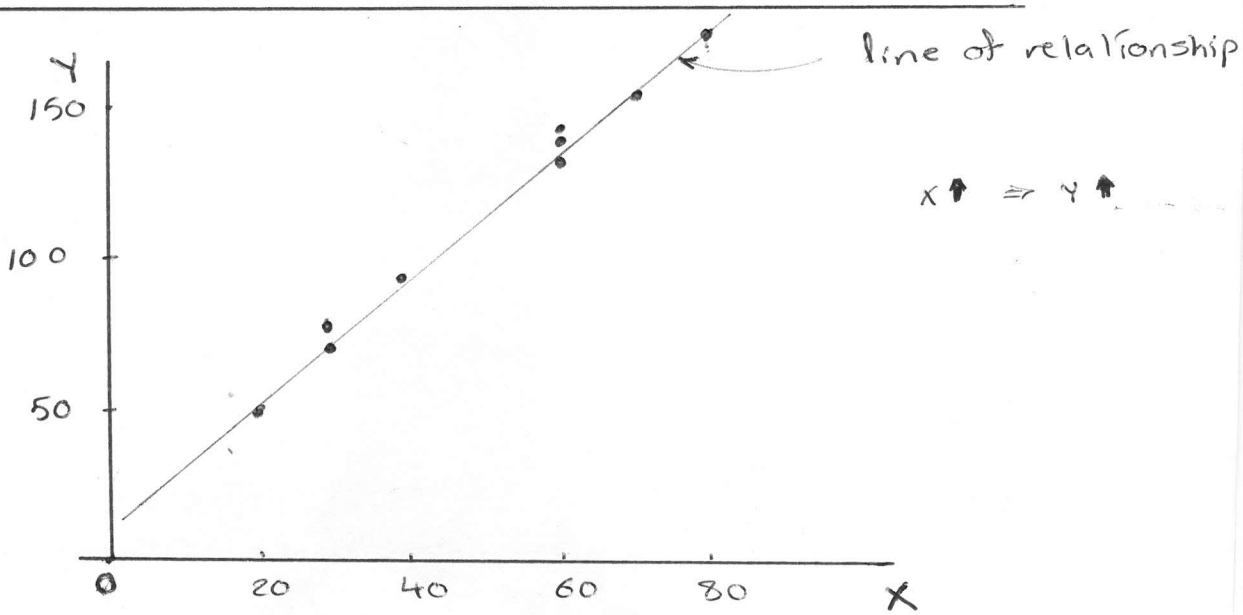
Example: Westwood Company:

Manufacturing of a certain spare part in lots which vary in size as demand fluctuates.

Production Run i	Lot size X_i	Man-hours Y_i
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

independent (predictor) variable

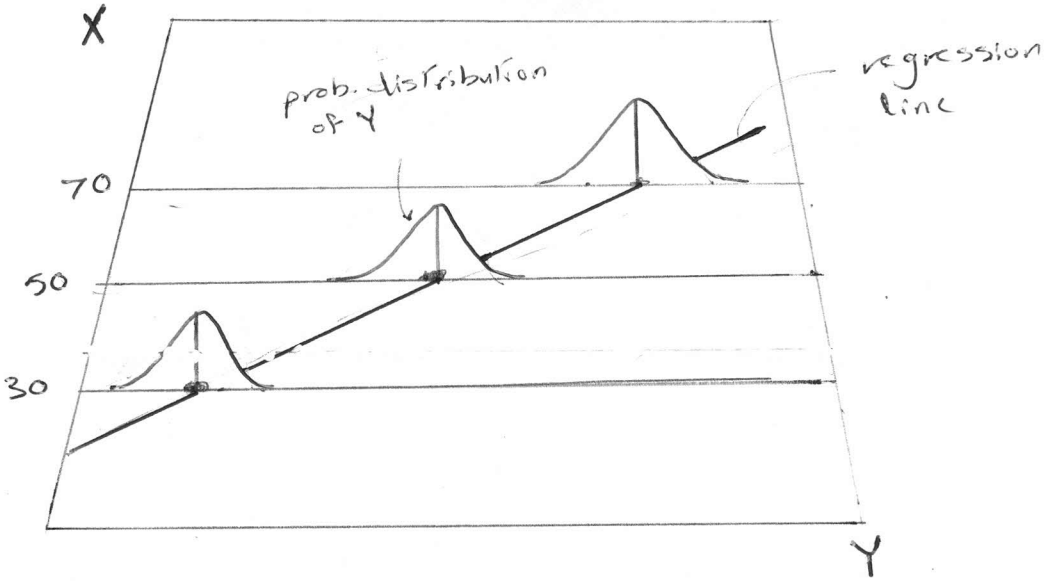
dependent (response) variable



scatter plot

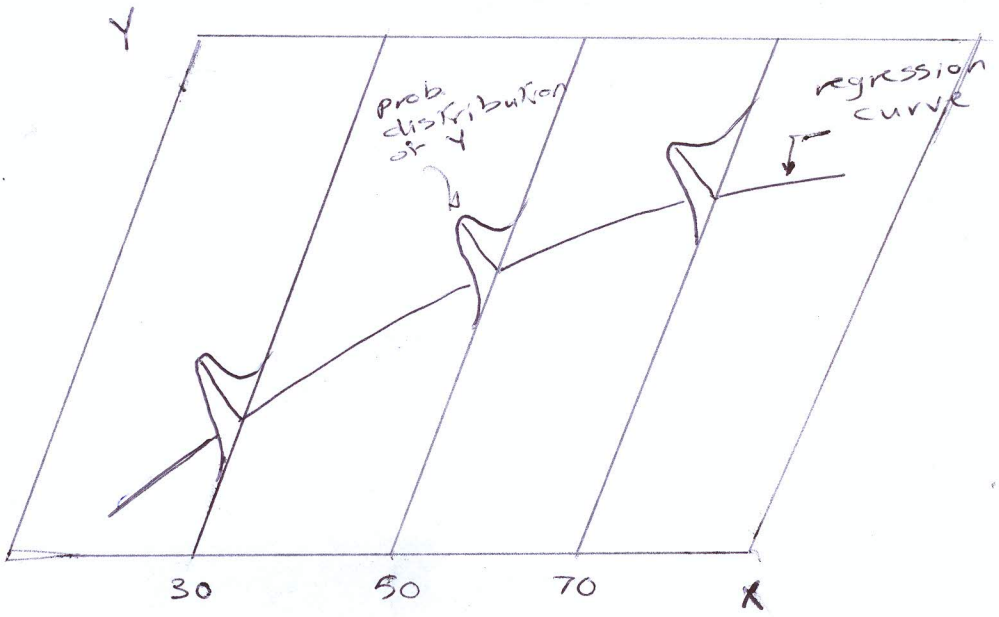
* Probability distributions of Y for lot sizes X = 30, 50, 70

Linear regression model



Curvilinear regression model

reflects economies of scale of larger for sizes.



→ Regression models with more than one independent variable; تحليل الانحدار المتعدد

→ Construction of regression models:

1- selection of independent variables

2- functional form of regression equation

3- scope of the model (coverage of some region of values of indep. variables)

4- use of regression analysis:

→ 1- description الوصف

2- control التحكم

3- prediction والتنبؤ

2-3 Regression model with distribution of error terms unspecified:

Simple linear regression model (one indep. variable)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , i = 1, 2, \ldots, n$$

(Y_i observed)

Y_i ... value of response variable in ith trial

β_0, β_1 ... parameters

X_i ... (const.) value of independent variable in ith trial

ε_i ... random error term; mean $E(\varepsilon_i) = 0$

variance $\sigma^2_{\varepsilon_i} = \sigma^2$

$\text{Cov}\,\sigma(\varepsilon_i, \varepsilon_j) = 0 \; ; \; \forall\, i \neq j$

Features of the model

- $Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{const. term}} + \underbrace{\varepsilon_i}_{\text{random term}}$

- $E(Y_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i$

  i.e. regression fun.: $E(Y) = \beta_0 + \beta_1 X$

- variation of Y_i from the value of regression fun. is ε_i which has a const. variance σ^2

  then, $\sigma^2(Y_i) = \sigma^2$

- error terms are uncorrelated, i.e. outcome in one trial has no effect on error term for any other trial.

- response variable observations Y_i has a probability distribution with mean $\beta_0 + \beta_1 X_i$ & variance σ^2 and any two observations Y_i, Y_j are uncorrelated.

* Meaning of regression parameters:

β_0, β_1 ... regression coefficients

→ slope of regression line
(change in mean of the prob. distribution of Y per unit increase in X)

→ Y-intercept of the regression line
(if the scope of the model includes $X=0$, then β_0 is the mean of prob. distribution of Y at $X=0$
(Cost of setting up production process, no matter lot size)

Example: for Westwood company lot size application

Assume regression function

$$Y_i = 9.5 + 2.1 X_i + \varepsilon_i$$

then $E(Y) = 9.5 + 2.1X$

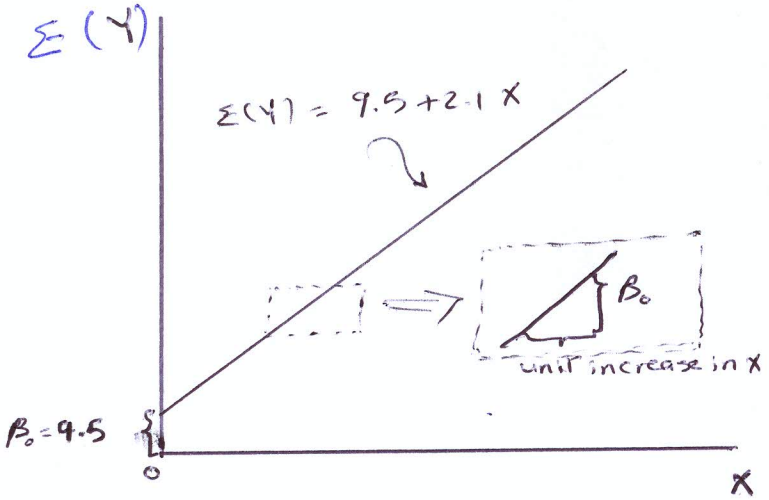
suppose in ith trial, $X_i = 45$ & $Y_i = 108$

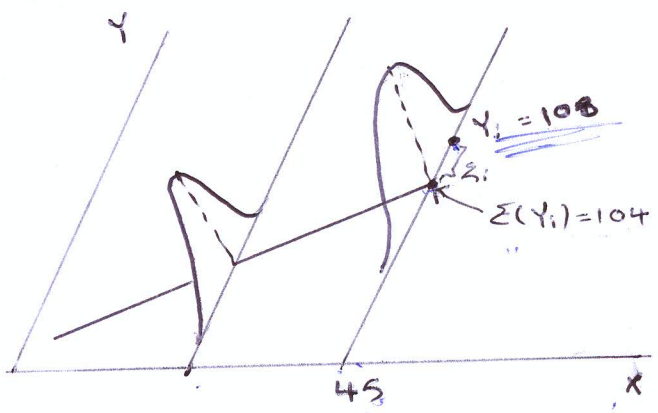
then $E(Y_i) = 9.5 + 2.1(45) = 104$

$Y_i - E(Y_i)$

i.e. $Y_i = 108 = 104 + 4 \Rightarrow \varepsilon_i = 4$

note: ε_i ... deviation of Y_i from its mean value $E(Y_i)$





Alternative versions of model:

① $Y_i = \beta_0 X_0 + \beta_1 X_i + \varepsilon_i$ where $X_0 \equiv 1$

② using $X_i - \bar{X}$ rather than X_i

$$Y_i = \beta_0 + \beta_1 (X_i - \bar{X}) + \beta_1 \bar{X} + \varepsilon_i$$

$$= \beta_0 + \beta_1 \bar{X} + \beta_1 (X_i - \bar{X}) + \varepsilon_i$$

$$= \beta_0^* + \beta_1 (X_i - \bar{X}) + \varepsilon_i$$

2.4 Estimation of regression function:

Obtaining needed sample data:

→ Experimental data – controlled

→ Nonexperimental data (observational) – not controlled

Method of least squares:

$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ for a given observation (X_i, Y_i)

sum of squared n deviations:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

values of β_0, β_1 that minimize Q, (b_0, b_1) point estimators of β_0, β_1

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\Rightarrow \sum (Y_i - b_0 - b_1 X_i) = 0 \quad \Rightarrow \sum Y_i - n b_0 - b_1 \sum X_i = 0$$

$$\sum X_i (Y_i - b_0 - b_1 X_i) = 0 \qquad \sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

$$\sum Y_i = n b_0 + b_1 \sum X_i \quad (1)$$
$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2 \quad (2)$$

normal equations

$\sum X_i$, $\sum X_i^2$, $\sum Y_i$, $\sum X_i Y_i$ are calculated then solve (1), (2)

multiply (1) $* \sum X_i \Rightarrow \sum X_i \sum Y_i = n b_0 \sum X_i + b_1 (\sum X_i)^2$

(2) $* n \Rightarrow n \sum X_i Y_i = n b_0 \sum X_i + n b_1 \sum X_i^2$

$$n \sum X_i Y_i - \sum X_i \sum Y_i = b_1 [n \sum X_i^2 - (\sum X_i)^2]$$

$$b_1 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}[\sum Y_i - b_1 \sum X_i] = \bar{Y} - b_1 \bar{X}$$

Properties of least-square estimators:

(Gauss-Markov theorem)

① unbiased $\quad E(b_0) = \beta_0$, $E(b_1) = \beta_1$

② have min. variance among all unbiased linear estimators. (b_0, b_1 are linear combinations of Y_i's, since X_i are const.'s).

Example: Westwood Company

basic calculations to obtain b_0, b_1

Y_i	X_i	$X_i Y_i$	X_i^2	Y_i^2
73	30	2,190	900	5,329
50	20	1,000	400	2,500
128	60	7,680	3,600	16,384
170	80	13,600	6,400	28,900
87	40	3,480	1,600	7,569
108	50	5,400	2,500	11,664
135	60	8,100	3,600	18,225
69	30	2,070	900	4,761
148	70	10,360	4,900	21,904
132	60	7,920	3,600	17,424
Σ 1,100	500	61,800	28,400	134,660

$n = 10$

use values to solve eqn.'s ①, ②

$$\text{or } \hat{\beta}_1 = b_1 = \frac{\Sigma X_i Y_i - \frac{\Sigma X_i \Sigma Y_i}{n}}{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}} = \frac{61800 - \frac{1}{10}(500)(1100)}{28400 - \frac{1}{10}(500)^2} = 2$$

$$\hat{\beta}_0 = b_0 = \frac{1}{n}\left[\Sigma Y_i - b_1 \Sigma X_i\right] = \frac{1}{10}\left[1100 - 2(500)\right] = 10$$

Estimated regression function

$$\hat{Y} = b_0 + b_1 X$$

, $\hat{Y}$ unbiased estimator of $E(Y)$

fitted regression line

$$\hat{Y} = 10 + 2X$$

$$\hat{Y}_i = b_0 + b_1 X_i$$

point estimate for mean number of man-hours when lot size is $X = 55$, would be:

$$\hat{Y} = 10 + 2(55) = 120$$

Residuals:

(observed vertical deviation of Y_i from fitted regression line)

The ith residual is the difference between the observed value Y_i and the fitted value $\hat{Y}_i$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i$$

Westwood Company example: observed values, Fitted values, residuals, and squared residuals:

i	X_i	Y_i	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$	$e_i^2 = (Y_i - \hat{Y}_i)^2$
1	30	73	70	+3	9
2	20	50	50	0	0
3	60	128	130	-2	4
4	80	170	170	0	0
5	40	87	90	-3	9
6	50	108	110	-2	4
7	60	135	130	+5	25
8	30	69	70	-1	1
9	70	148	150	-2	4
10	60	132	130	+2	4
	500	1100	1100	0	60

* Properties of fitted regression line:

1- sum of residuals is zero $\sum_{i=1}^{n} e_i = 0$ (3)

$\sum Y_i - b_0 - b_1 \sum X_i = 0$ normal eqn: (1)

2- sum of squared residuals $\sum_{i=1}^{n} e_i^2$ is a minimum

$\sum Y_i = n b_0 + b_1 \sum X_i$

3- $\sum Y_i = \sum \hat{Y}_i$ (4)

4- $\sum_{i=1}^{n} X_i e_i = 0$ (5) sum of weighted residuals is zero

$\sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$ normal eqn: (2)

5- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ (6) from (3), (5)

6. The regression line always goes through the point $(\bar{X}, \bar{Y})$.

from the alternative regression model

$$Y_i = \beta_0^* + \beta_1 (X_i - \bar{X}) + \xi_i \quad , \quad \beta_0^* = \beta_0 + \beta_1 \bar{X}$$

then $b_0^* = b_0 + b_1 \bar{X} = \bar{Y} - b_1 \bar{X} + b_1 \bar{X} = \bar{Y}$

i.e. $\hat{Y} = \bar{Y} + b_1 (X - \bar{X})$ $\quad X = \bar{X} \Rightarrow \hat{Y} = \bar{Y}$

* Estimation of error terms (ξ_i) variance:
(point estimate of σ^2)

Regression model:

→ recall for a single population: estimate for σ^2

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$$

→ recall that $\sigma^2(Y_i) = \sigma^2(\xi_i) = \sigma^2$

→ deviation of an observation Y_i is calculated around its estimated mean $\hat{Y}_i$, then:

sum of squares:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \sum_{i=1}^{n} e_i^2$$

mean square:

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

(has $n-2$ degrees of freedom)

MSE unbiased estimator of σ^2

$E(MSE) = \sigma^2$

Alternative computational formulas:

$$SSE = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum X_i Y_i$$

if b_0, b_1 are calculated first (?)

or

$$SSE = \sum (Y_i - \bar{Y})^2 - \frac{\left[\sum (X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\sum (X_i - \bar{X})^2}$$

$$= \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} - \frac{\left[\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}\right]^2}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

Returning to Westwood Company example:

$$SSE = \sum e_i^2 = 60$$

$$MSE = \frac{SSE}{n-2} = \frac{60}{8} = 7.5$$

or $$SSE = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum X_i Y_i$$

$$= 134,660 - 10(1,100) - 2(61,800) = 60$$

Normal error regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \quad i = 1, \ldots, n$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{(independent)}$$

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Estimation of parameters using method of maximum likelihood: $(\beta_0, \beta_1, \sigma^2)$

The likelihood function for the normal error model [joint prob. distribution of parameters]

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum (Y_i - \beta_0 - \beta_1 X_i)^2}$$

Maximum likelihood estimators:

$$\hat{\beta}_0 = b_0 = \frac{1}{n}\left[\sum Y_i - b_1 \sum X_i\right]$$

$$\hat{\beta}_1 = b_1 = \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2}$$

$$\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n} \quad \text{(biased estimator)}$$

$$\Rightarrow \quad MSE = \frac{n\hat{\sigma}^2}{n-2} \quad \text{(unbiased)} \quad \left(\frac{SSE}{n-2}\right)$$