Chapter 1

Linear Regression with One Predictor Variable

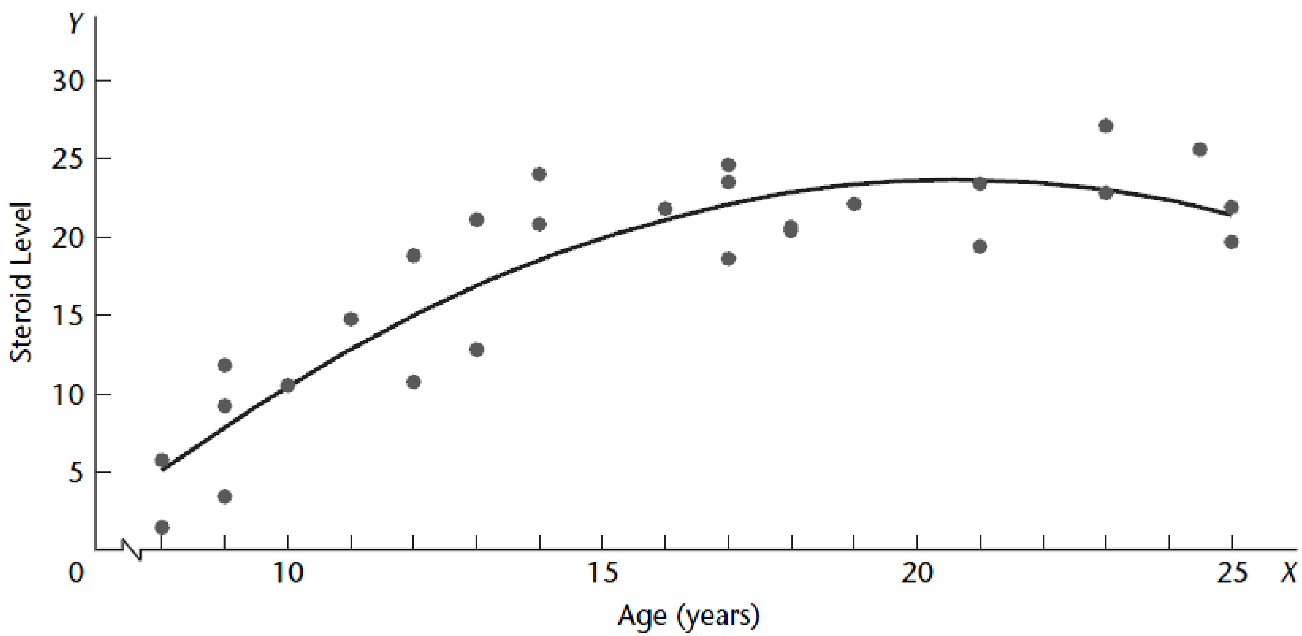
Outline

- Relations between Variables
- Concepts in Regression Models
  - random error, residuals, fitted value, ……
- **Simple Linear Regression Model with Distribution of Error Terms Unspecified**
  - Least square estimators (LSEs)
  - Properties of LSEs
- **Normal Error Regression Model**
  - Maximum likelihood estimators (MLEs)
  - Properties of MLEs

1.1 Relations between Variables

- **Functional Relation between Two Variables**
  - $Y = f(X)$
- **Statistical Relation between Two Variables**
  - $Y = f(X) + \varepsilon$

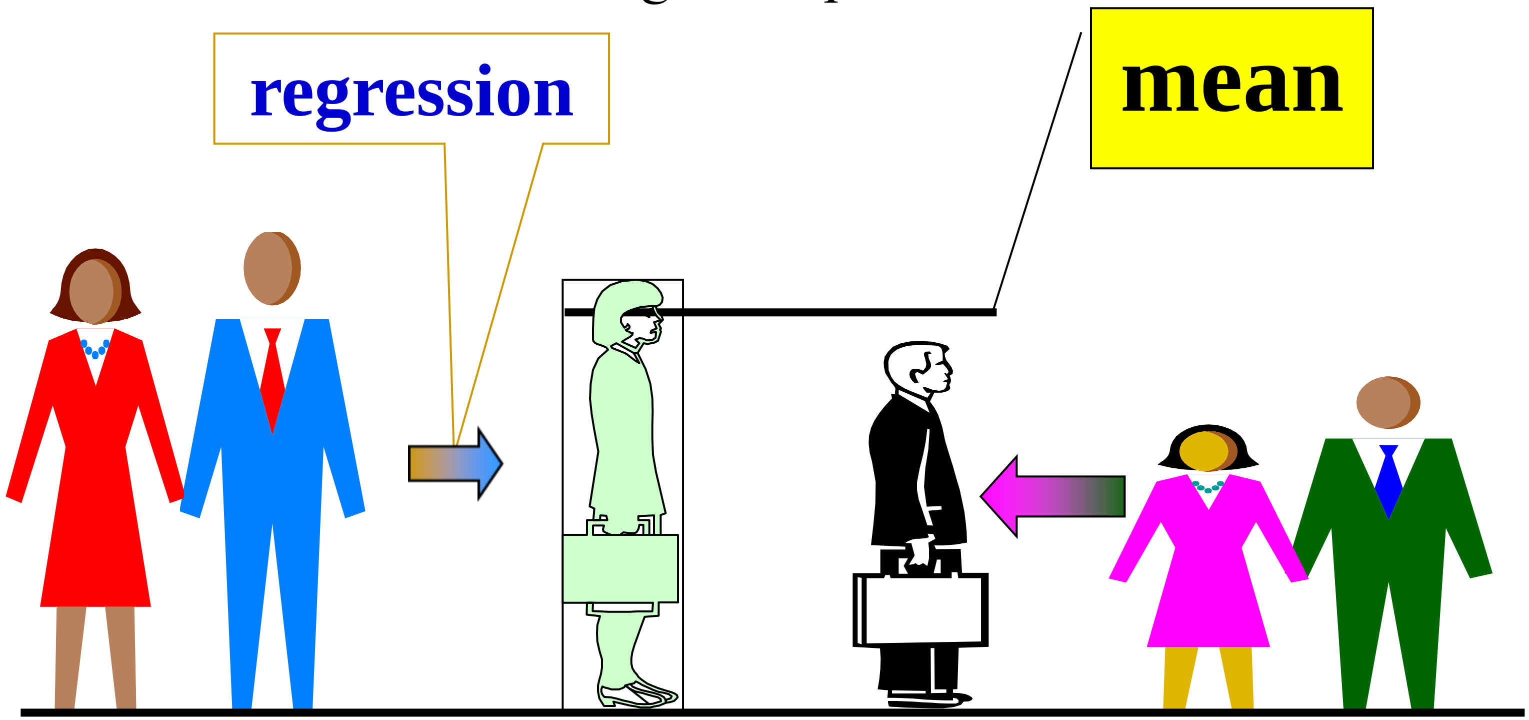
FIGURE 1.3 Curvilinear Statistical Relation between Age and Steroid Level in Healthy Females Aged 8 to 25.



1.2 Regression Models and Their Uses

Historical Origins

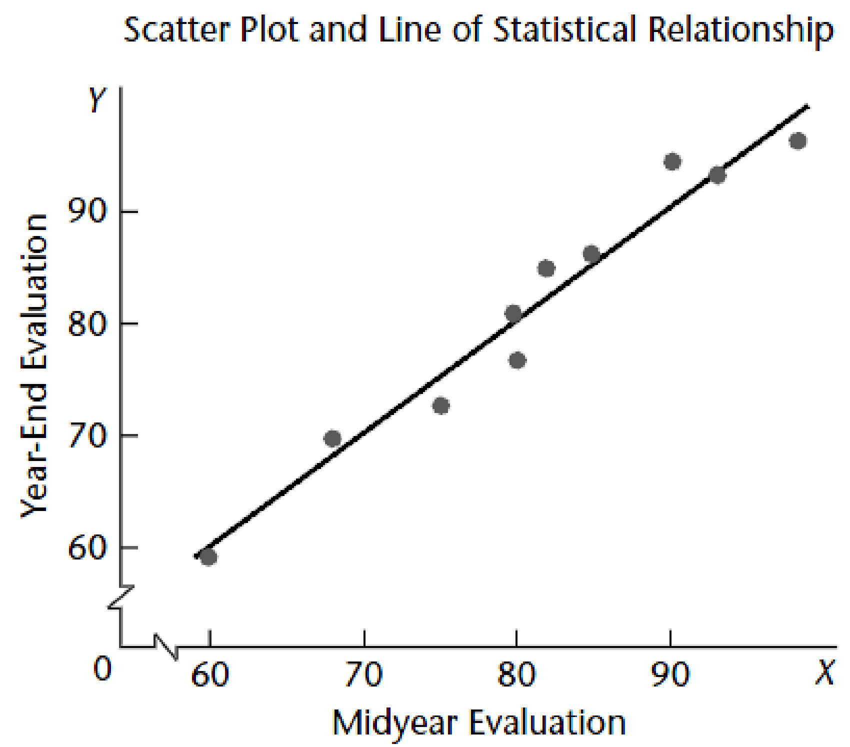
- First developed by Sir Francis Galton in the 19th century.
- The relation between heights of parents and children.

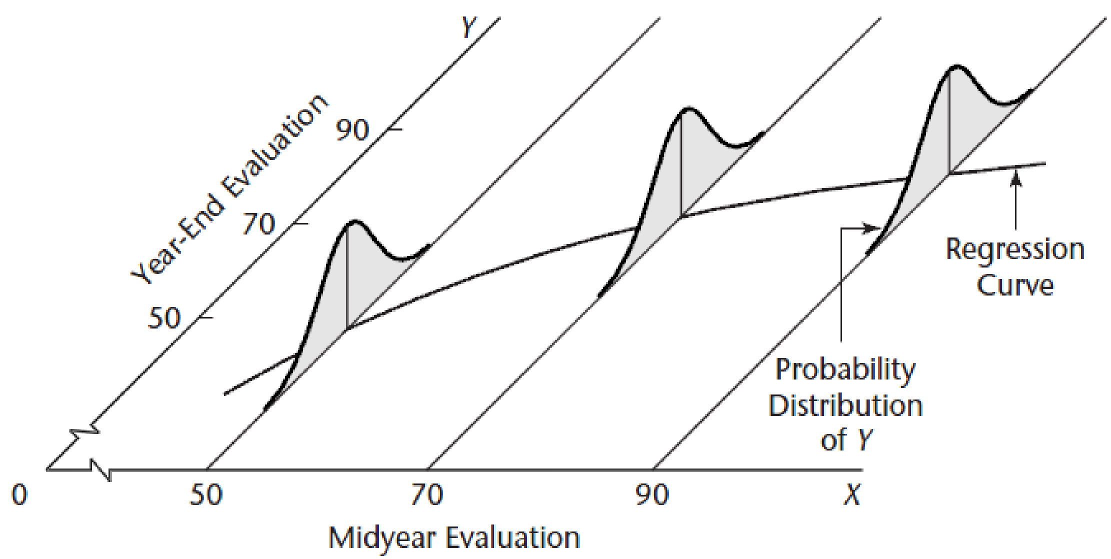


Sir Francis Galton's study in 1877

Basic concepts

- There is a probability distribution of Y for each level of X.
- The means of these probability distributions vary in some fashion with X.
- *e.g.* $Y \sim N(\alpha+\beta X, \sigma^2)$
  $\Leftrightarrow Y = \alpha+\beta X+\varepsilon,$
  $\varepsilon \sim N(0, \sigma^2)$





Goals of Regression Analysis

- Regression model describes an association between *X* and *Y*
  - model a statistical relationship between an "predictor variable" (input, independent variable, etc.) and a "response variable " (output, dependent variable, etc.)
- Two distinct goals
  - (Estimation) Understanding the relationship between predictor variables and response variables
  - (Prediction) Predicting the future response given the new observed predictors.

Use of regression analysis

- Description
- Control
- Prediction

- **Always** need to consider scope of the model.
- Statistical relationship generally does **not** imply **causality.**

1.3 Simple Linear Regression Model with Distribution of Error Terms Unspecified

Model – Error Distribution Unspecified

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i=1,2,\ldots n \qquad (1.1)$$

- Y_i : value of the response variable in the i-th trial
- X_i : a fixed known constant, the value of the predictor variable in the i-th trial
- ε_i : a random error term with $E(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i)=\sigma^2$, ε_i and ε_j are uncorrelated.
- β_0, β_1, and σ^2 are unknown parameters (constants).

Model – Error Distribution Unspecified

- The response Y_i = deterministic term + random term
  - deterministic term $\beta_0 + \beta_1 X_i$;
  - random term ε_i with E(ε_i) = 0, Var(ε_i) = σ^2 , ε_i and ε_j are uncorrelated

$\Rightarrow$ Implies Y_i is a random variable

$$E\{Y_i\} = E\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \beta_0 + \beta_1 X_i + E\{\varepsilon_i\} = \beta_0 + \beta_1 X_i + 0 = \beta_0 + \beta_1 X_i$$

$$\text{var}\{Y_i\} = \text{var}\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \text{var}\{\varepsilon_i\} = \sigma^2$$

$$\text{cov}\{Y_i, Y_j\} = \text{cov}\{\beta_0 + \beta_1 X_i + \varepsilon_i, \beta_0 + \beta_1 X_j + \varepsilon_j\} = \text{cov}\{\varepsilon_i, \varepsilon_j\} = 0 \ \forall\, i \neq j$$

Alternative Form:

$$Y_i = \beta_0 + \beta_1\left(X_i - \overline{X}\right) + \beta_1 \overline{X} + \varepsilon_i = \beta_0^* + \beta_1\left(X_i - \overline{X}\right) + \varepsilon_i \qquad \beta_0^* = \beta_0 + \beta_1 \overline{X}$$

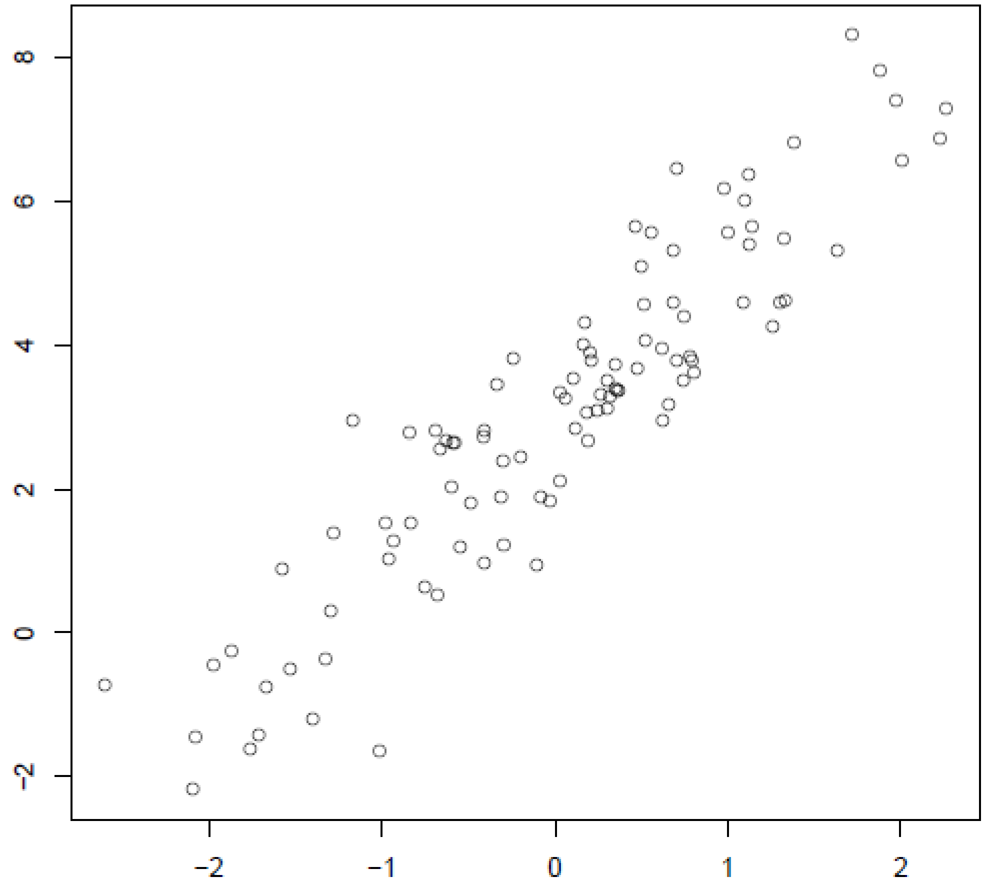
1.4 Data for Regression Analysis

- Observational Data
  - Example: relation between age of employee (X) and number of days of illness last year (Y)
  - Cannot be controlled!
- Experimental Data
  - Example: an insurance company wishes to study the relation between productivity of its analysts in processing claims (Y) and length of training X.
  - Treatment: the length of training
  - Experimental Units: the analysts included in the study.
- Completely Randomized Design: Most basic type of statistical design

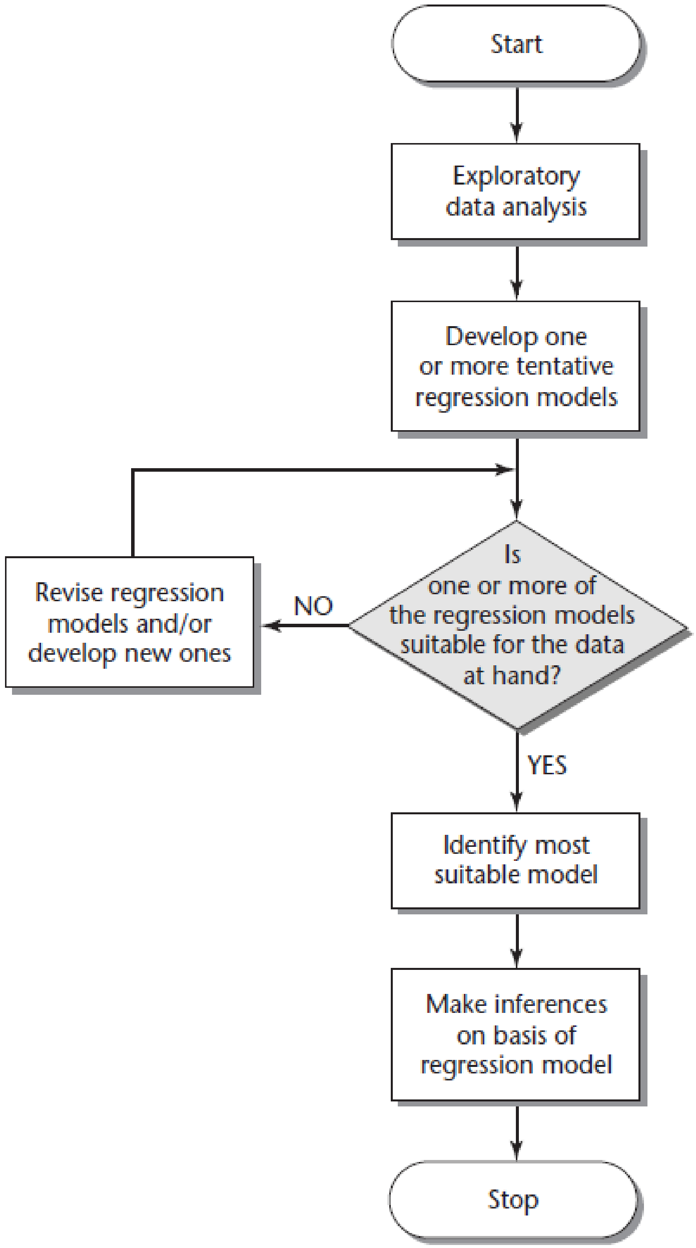
Simple Linear Regression

- Dataset: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

Why is it called *SLR*?
Simple: only one predictor X
Linear: regression function is linear



1.5 Overview of Steps in Regression Analysis



1.6 Estimation of Regression Function

Example

- An experimenter gave three subjects a very difficult task. Data on the age of the subject (X) and on the number of attempts to accomplish the task before giving up (Y) follow:

Subject i	1	2	3
Age X_i	20	55	30
Number of Attempts Y_i	5	12	10

- Want to find parameters for a function of the form

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

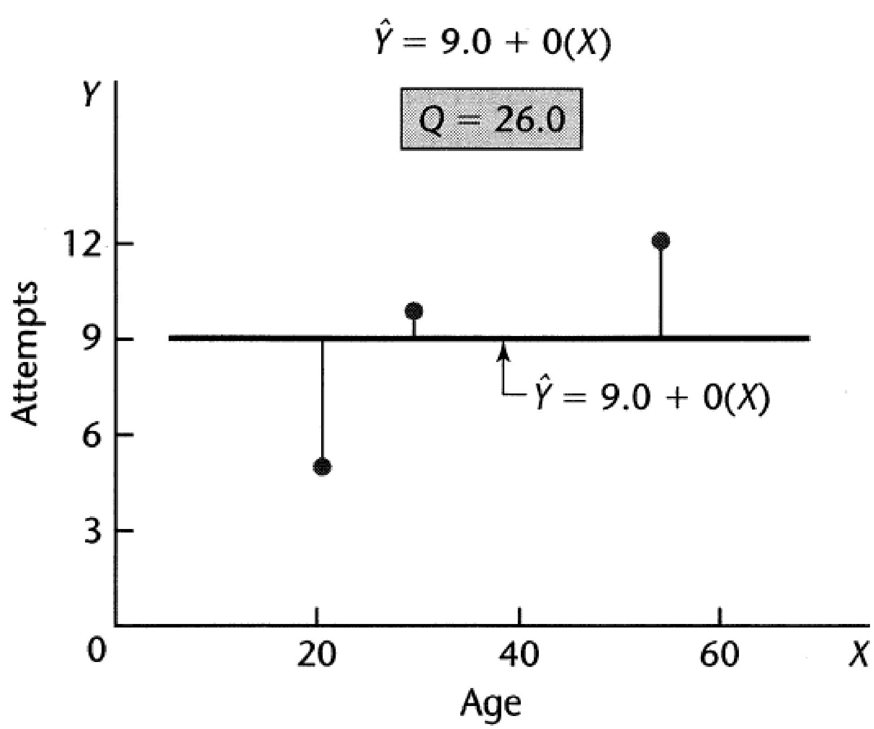
Least Squares Estimation

- Goal: make Y_i and $\beta_0 + \beta_1 X_i$ close for all i .
- Proposal 1: minimize $Q = \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i \right)$
- Proposal 2: minimize $Q = \sum_{i=1}^{n} | \varepsilon_i | = \sum_{i=1}^{n} | Y_i - \beta_0 - \beta_1 X_i |$
- Proposal 3 (Final Proposal): minimize

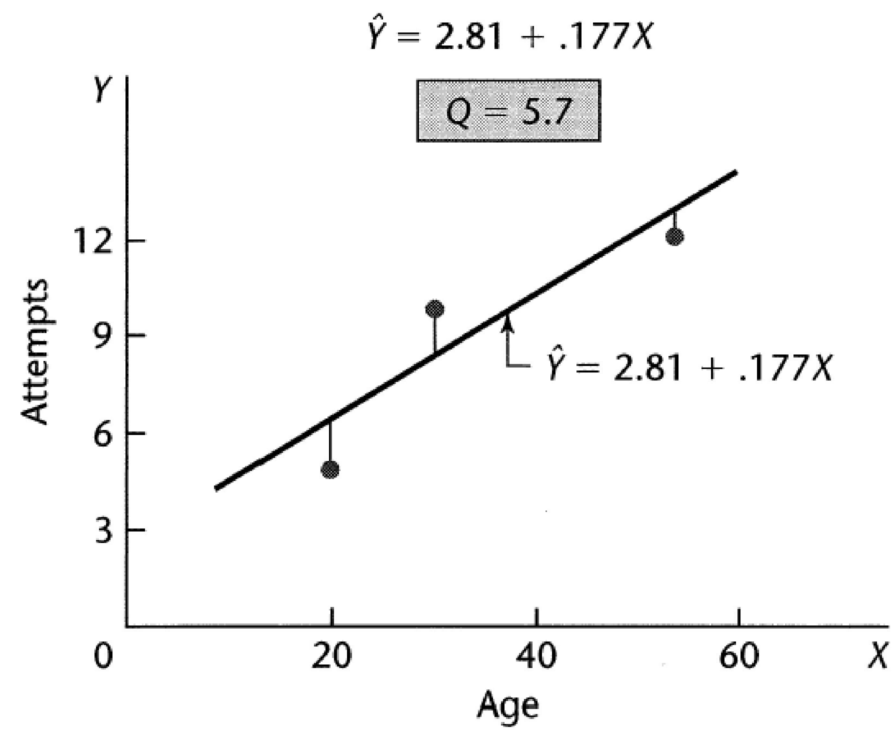
$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i \right)^2$$

  - Choose b_0 and b_1 as estimators for β_0 and β_1.
  - b_0 and b_1 will minimize the criterion Q for the given sample observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

Comparison



$\hat{Y} = 9.0 + 0(X)$
Q = 26.0
Y
Attempts
12
9
6
3
0
20
40
60
X
Age
$\hat{Y} = 9.0 + 0(X)$



$\hat{Y} = 2.81 + .177X$
Q = 5.7
Y
Attempts
12
9
6
3
0
20
40
60
X
Age
$\hat{Y} = 2.81 + .177X$

Repetition- The Summation Operator

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = \sum_{i=1}^{n} X_i - n\overline{X} = 0$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_0 + \beta_1 X_i + \varepsilon_i\right) = \beta_0 + \beta_1\overline{X} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$$

$$SS_{XX} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

$$SS_{YY} = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$$

$$SS_{XY} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}$$

Question: The expectations of random variables $\overline{Y}, SS_{YY}, SS_{XY}$?

$$E(\overline{Y}) = E\left(\beta_0 + \beta_1\overline{X} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right) = \beta_0 + \beta_1\overline{X}, \quad \operatorname{var}(\overline{Y}) = \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n}\varepsilon_i\right) = \frac{\sigma^2}{n}$$

Least Squares Estimation

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

Find least square estimators b_0, b_1 that minimize Q

$$Q(b_0, b_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$$

$$\frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)\left(-1\right) \overset{\text{set}}{=} 0 \Rightarrow \sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i \qquad (1)$$

$$\frac{\partial Q}{\partial \beta_1} = 2\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)\left(-X_i\right) \overset{\text{set}}{=} 0 \Rightarrow \sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \qquad (2)$$

Normal equations

Least Squares Estimation

$$(1):\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i; \qquad (2):\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

Solving by multiplying (1) by $\frac{1}{n}\sum_{i=1}^{n} X_i$ and taking $(2)-(1)$:

$$\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right) = b_1\left(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right)$$

$$\Rightarrow SS_{XY} = b_1 SS_{XX}$$

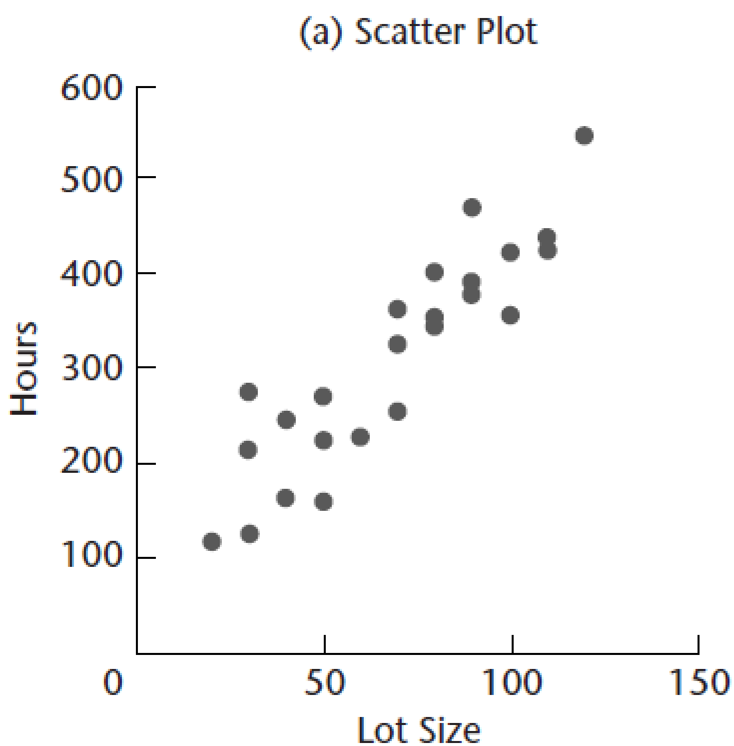
$$\Rightarrow b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

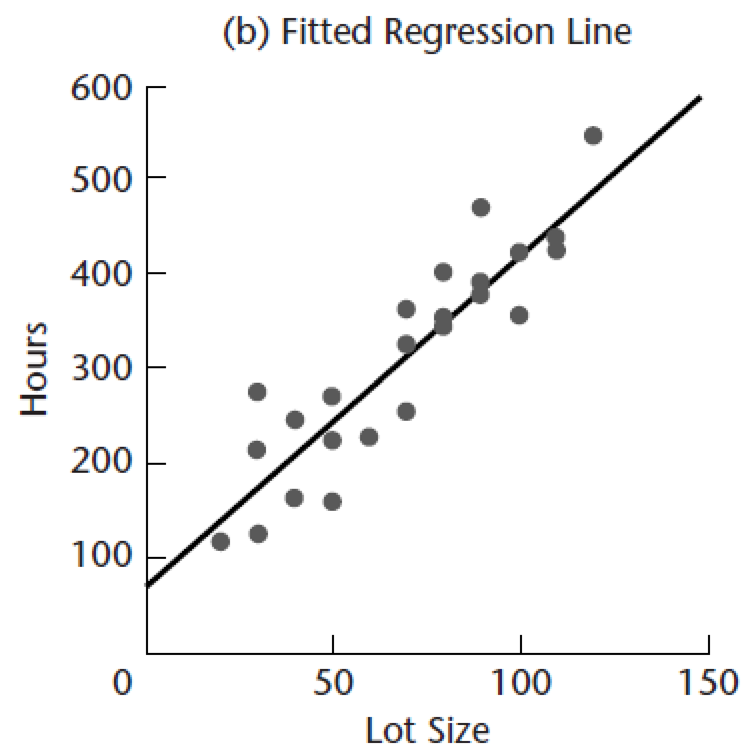
From (1): $\quad b_0 = \overline{Y} - b_1 \overline{X}$

$\boxed{\text{Fitted line goes through } (\overline{X}, \overline{Y})}$

Toluca Company Example

- The Toluca Company manufactures refrigeration equipment as well as many replacement parts.
- Company officials wished to determine the relationship between lot size and labor hours required to produce the lot.





LS Estimation for the example

Run i	(1) Lot Size X_i	(2) Work Hours Y_i	(3) $X_i - \bar{X}$	(4) $Y_i - \bar{Y}$	(5) $(X_i - \bar{X})(Y_i - \bar{Y})$	(6) $(X_i - \bar{X})^2$	(7) $(Y_i - \bar{Y})^2$
1	80	399	10	86.72	867.2	100	7,520.4
2	30	121	−40	−191.28	7,651.2	1,600	36,588.0
3	50	221	−20	−91.28	1,825.6	400	8,332.0
···	···	···	···	···	···	···	···
23	40	244	−30	−68.28	2,048.4	900	4,662.2
24	80	342	10	29.72	297.2	100	883.3
25	70	323	0	10.72	0.0	0	114.9
Total	1,750	7,807	0	0	70,690	19,800	307,203
Mean	70.0	312.28					

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{70{,}690}{19{,}800} = 3.5702$$

$$b_0 = \bar{Y} - b_1\bar{X} = 312.28 - 3.5702(70.0) = 62.37$$

$$\hat{Y} = 62.37 + 3.5702X$$

Fitted Values and Residuals

- True regression line $E(Y) = \beta_0 + \beta_1 X$.
- Using the estimated parameters, the fitted regression line is

$$\hat{Y} = b_0 + b_1 X \qquad \widehat{\mathrm{E}(Y)} = b_0 + b_1 X$$

- ***Residual:*** the difference between the observed and fitted predicted value. $e = Y - \hat{Y}$
- The *fitted value* for the *i*th case $\hat{Y}_i = b_0 + b_1 X_i \quad i = 1, \ldots, n$
- The *i*th *residual*

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(b_0 + b_1 X_i\right) \quad i = 1, \ldots, n$$

  - Distinguish between the model error term value

$$\varepsilon_i = Y_i - E(Y_i) = Y_i - \left(\beta_0 + \beta_1 X_i\right) \quad i = 1, \ldots, n$$

- Sum of the squared residuals

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

Fitted Values, Residuals, and Squared Residuals—Toluca Company Example

$$\hat{Y}_1 = 62.37 + 3.5702(80) = 347.98$$

Run i	(1) Lot Size X_i	(2) Work Hours Y_i	(3) Estimated Mean Response $\hat{Y}_i$	(4) Residual $Y_i - \hat{Y}_i = e_i$	(5) Squared Residual $(Y_i - \hat{Y}_i)^2 = e_i^2$
1	80	399	347.98	51.02	2,603.0
2	30	121	169.47	−48.47	2,349.3
3	50	221	240.88	−19.88	395.2
...	...	...	...	...	...
23	40	244	205.17	38.83	1,507.8
24	80	342	347.98	−5.98	35.8
25	70	323	312.28	10.72	114.9
Total	1,750	7,807	7,807	0	54,825

$$\hat{Y}_i = b_0 + b_1 X_i = \left(\overline{Y} - \frac{SS_{XY}}{SS_{XX}}\overline{X}\right) + \frac{SS_{XY}}{SS_{XX}} X_i = \overline{Y} + \frac{SS_{XY}}{SS_{XX}}\left(X_i - \overline{X}\right)$$

Alternative Model

- Using the alternative format of linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i = \beta_0^* + \beta_1\left(X_i - \overline{X}\right) + \varepsilon_i, \quad \beta_0^* = \beta_0 + \beta_1 \overline{X}$$

- The least squares estimators

$$b_1 = \frac{SS_{XY}}{SS_{XX}}, \qquad b_0 = \overline{Y}$$

  - b_1 for β_1 remains the same as before, and

$$b_0^* = \overline{Y} = \left(\overline{Y} - b_1\overline{X}\right) + b_1\overline{X} = b_0 + b_1\overline{X}$$

- Hence the estimated regression function is

$$\boxed{\hat{Y}_i = b_0^* + b_1\left(X_i - \overline{X}\right) = \overline{Y} + \frac{SS_{XY}}{SS_{XX}}\left(X_i - \overline{X}\right)}$$

- In the Toluca Company example, $\bar{Y} = 312.28$ and $\bar{X} = 70.0$

$$\hat{Y} = 312.28 + 3.5702(X - 70.0)$$

Properties of Fitted regression line

(1) $\sum e_i = 0$

(2) $\sum e_i^2$ is minimized

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

(4) $\sum X_i e_i = 0$

(5) $\sum \hat{Y}_i e_i = 0$

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

- These properties follow directly from the least squares criterion and normal equations (pg 23-24)

Proof:

$$(1)\ \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} [Y_i - \bar{Y} - b_1 (X_i - \bar{X})] = 0$$

$$\Rightarrow (3)\ \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

$$(4) \sum_i X_i e_i = \sum_i (X_i - \bar{X}) e_i$$

$$= \sum_i (X_i - \bar{X})[Y_i - \bar{Y} - b_1 (X_i - \bar{X})] = SS_{xy} - b_1 SS_{xx} = 0$$

$$(5)\ \sum_i \hat{Y}_i e_i = \sum_i e_i [\bar{Y} + b_1 (X_i - \bar{X})]$$

$$= \bar{Y} \sum_i e_i + b_1 \sum_i e_i (X_i - \bar{X}) = 0$$

1.7 Estimation of Error Terms Variance σ^2

$$\sigma^2 = \text{var}\{\varepsilon\} = E\left\{\left(\varepsilon - E(\varepsilon)\right)^2\right\} = E\left\{\left(\varepsilon - 0\right)^2\right\} = E\left\{\varepsilon^2\right\}$$

ε unobservable since $\varepsilon = Y - \left(\beta_0 + \beta_1 X\right)$

We use residual e to "estimate" ε

$$e = Y - \hat{Y} = Y - \left(b_0 + b_1 X\right)$$

Obtain the "average" squared residual to estimate σ^2 :

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{SSE}{n-2} = MSE$$

- Toluca Company example, we obtain: $SSE = 54825$,

$$s^2 = MSE = \frac{54{,}825}{23} = 2{,}384$$

Properties of Estimators

Under linear regression model (1.1) in which the errors have expectation zero and are uncorrelated and have equal variances σ^2.

(1) Least squares estimators b_0 and b_1 are linear combinations of $\{Y_i\}$

(2) (*Gauss-Markov theorem*) Least squares estimators b_0 and b_1 are BLUE (best linear unbiased estimators) of β_0 and β_1 respectively.

- Best: have minimum variance among all unbiased linear estimators

(3) MSE is an unbiased estimator of σ^2, i.e. $E(MSE) = \sigma^2$.

Properties of Estimators

(1) Proof:

$$b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \sum_{i=1}^{n}\frac{\left(X_i - \overline{X}\right)}{SS_{XX}}Y_i = \sum_{i=1}^{n}k_i Y_i$$

$$b_0 = \overline{Y} - b_1\overline{X} = \sum_{i=1}^{n}\left(\frac{1}{n} - k_i\overline{X}\right)Y_i = \sum_{i=1}^{n}l_i Y_i$$

(2) Proof:

$$k_i = \frac{X_i - \overline{X}}{SS_{XX}}$$

Note $\sum_{i=1}^{n}k_i = 0, \quad \sum_{i=1}^{n}k_i X_i = 1, \quad \sum_{i=1}^{n}k_i^2 = \frac{1}{SS_{XX}}$

$$E(b_1) = \sum_{i=1}^{n}k_i E(Y_i) = \sum_{i=1}^{n}k_i\left(\beta_0 + \beta_1 X_i\right) = \beta_0\sum_{i=1}^{n}k_i + \beta_1\sum_{i=1}^{n}k_i X_i = \beta_1$$

$$E\left\{b_0\right\} = E\left\{\overline{Y} - b_1\overline{X}\right\} = \left(\beta_0 + \beta_1\overline{X}\right) - \beta_1\overline{X} = \beta_0$$

So b_0 and b_1 unbiased estimators of β_0 and β_1 respectively.

Next, consider variances of b_0 and b_1.

$$\text{var}(b_1) = \text{var}\left(\sum_{i=1}^{n} k_i Y_i\right) = \sum_{i=1}^{n} k_i^2 \,\text{var}(Y_i) = \sigma^2 \sum_{i=1}^{n} k_i^2 = \frac{\sigma^2}{SS_{XX}}$$

$$\text{cov}\left\{b_1, Y_i\right\} = \text{cov}\left\{\sum_{i=1}^{n} k_i Y_i, Y_i\right\} = \sum_{j=1}^{n} \text{cov}\left\{k_j Y_j, Y_i\right\} = \text{cov}\left\{k_i Y_i, Y_i\right\}_i = k_i \sigma^2$$

$$\text{cov}\left\{b_1, \overline{Y}\right\} = \text{cov}\left\{b_1,\ \sum_{i=1}^{n} \frac{1}{n} Y_i\right\} = \frac{1}{n}\sum_{i=1}^{n} k_i \sigma^2 = 0$$

$$\text{var}\left\{b_0\right\} = \text{var}\left\{\overline{Y} - b_1\overline{X}\right\} = \text{var}\left\{\overline{Y}\right\} + \overline{X}^2 \,\text{var}\left\{b_1\right\} - 2\overline{X}\,\text{cov}\left\{\overline{Y}, b_1\right\}$$

$$= \text{var}\left\{\overline{Y}\right\} + \overline{X}^2 \,\text{var}\left\{b_1\right\} = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{SS_{XX}}\right) = \frac{\sum X_i^2}{nSS_{XX}}\sigma^2$$

Properties of Estimators

$$\operatorname{cov}(b_0, b_1) = \operatorname{cov}(\bar{Y} - b_1\bar{X}, b_1) = -\bar{X}\operatorname{var}(b_1) = -\frac{\bar{X}}{SS_{XX}}\sigma^2$$

Variance matrix of (b_0, b_1)

$$\frac{\sigma^2}{SS_{XX}}\begin{pmatrix} \frac{1}{n}\sum X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}$$

Properties of Estimators

- Among all unbiased linear estimators of the form

$$\hat{\beta}_1 = \sum c_i Y_i$$

- As this estimator must be unbiased we have

$$\mathbb{E}(\hat{\beta}_1) = \sum c_i \,\mathbb{E}(Y_i) = \sum c_i(\beta_0 + \beta_1 X_i)$$
$$= \beta_0 \sum c_i + \beta_1 \sum c_i X_i \; = \beta_1$$

  - Clearly it must be the case that $\sum c_i = 0$ and $\sum c_i X_i = 1$

- Now define

$$d_i = c_i - k_i \quad \text{where } \; k_i = \frac{X_i - \overline{X}}{SS_{XX}}$$

Properties of Estimators

- The variance of this estimator

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \sum c_i^2 \,\mathrm{Var}(Y_i) = \sigma^2 \sum (k_i + d_i)^2 \\
&= \sigma^2 \left(\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \right)
\end{aligned}
$$

- Note we just demonstrated that $\sigma^2 \sum k_i^2 = \mathrm{Var}(b_1)$
- Recall $\sum c_i = 0$ and $\sum c_i X_i = 1$
- Now by showing that

$$
\sum k_i d_i \;=\; \sum k_i (c_i - k_i) \;=\; \sum k_i c_i - \sum k_i^2
$$

$$
= \sum c_i \left(\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right) - \frac{1}{\sum (X_i - \bar{X})^2} = \frac{\sum c_i X_i - \bar{X} \sum c_i}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} = 0
$$

Properties of Estimators

- So we are left with

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(b_1) + \sigma^2 \left(\sum d_i^2\right)$$

- It is minimized when all the $d_i = 0$. This means that the least squares estimator b_1 is BLUE of β_1.

- Similarly, we can show b_0 is BLUE of β_0.

Properties of Estimators

(3) Proof:

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i = Y_i - \left(\overline{Y} - b_1 \overline{X}\right) - b_1 X_i = (Y_i - \overline{Y}) - b_1 (X_i - \overline{X})$$

$$E(e_i) = E(Y_i - b_0 - b_1 X_i) = EY_i - Eb_0 - E(b_1) X_i = \beta_0 + \beta_1 X_i - \beta_0 - \beta_1 X_i = 0$$

$$\text{var}(e_i) = \text{var}[Y_i - \overline{Y} - b_1 (X_i - \overline{X})]$$

$$= \text{var}(Y_i) + \text{var}(\overline{Y}) + \text{var}(b_1)(X_i - \overline{X})^2 - 2\,\text{cov}(Y_i, \overline{Y}) - 2(X_i - \overline{X})\left[\text{cov}(Y_i, b_1) - \text{cov}(\overline{Y}, b_1)\right]$$

$$= \sigma^2 + \frac{\sigma^2}{n} + \frac{(X_i - \overline{X})^2 \sigma^2}{SS_{XX}} - \frac{2\sigma^2}{n} - \frac{2(X_i - \overline{X})^2 \sigma^2}{SS_{XX}} + 0$$

$$= \frac{(n-1)\sigma^2}{n} - \frac{(X_i - \overline{X})^2 \sigma^2}{SS_{XX}}$$

$$E(SSE) = E\left(\sum_{i=1}^{n} e_i^2\right) = \sum_{i=1}^{n} E(e_i^2) = \sum_{i=1}^{n} \mathrm{var}(e_i)$$

$$= \sum_{i=1}^{n}\left[\frac{(n-1)\sigma^2}{n} - \frac{(X_i - \overline{X})^2\sigma^2}{SS_{XX}}\right] = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2$$

$$E(MSE) = \frac{E(SSE)}{n-2} = \sigma^2$$

- Question: For any $i \neq j$, ε_i and ε_j are uncorrelated.

  Are e_i and e_j uncorrelated?

$$\boxed{0 = \operatorname{var}\left(\sum_{i=1}^{n} e_i\right) \neq \sum_{i=1}^{n} \operatorname{var}(e_i) = (n-2)\sigma^2, \quad \text{for } n > 2}$$

$$0 = \operatorname{var}\left(\sum_{i=1}^{n} e_i\right) = \sum_{i=1}^{n} \operatorname{var}(e_i) + \sum_{\substack{i,j=1 \\ j \neq i}}^{n} \operatorname{cov}(e_i, e_j)$$

$$\Rightarrow \sum_{\substack{i,j=1 \\ j \neq i}}^{n} \operatorname{cov}(e_i, e_j) = -\sum_{i=1}^{n} \operatorname{var}(e_i) = -(n-2)\sigma^2$$

In fact, we can get $\operatorname{cov}(e_i, e_j) = -\dfrac{\sigma^2}{n} - \dfrac{(X_i - \overline{X})(X_j - \overline{X})\sigma^2}{SS_{XX}}$

for $i \neq j$. Then $\sum_{\substack{i,j=1 \\ j \neq i}}^{n} \operatorname{cov}(e_i, e_j) = -(n-1)\sigma^2 + \sigma^2 = -(n-2)\sigma^2$,

$$\text{since } 0 = \left[\sum_{i=1}^{n} (X_i - \overline{X})\right]^2 = SS_{XX} + \sum_{\substack{i,j=1 \\ j \neq i}}^{n} (X_i - \overline{X})(X_j - \overline{X})$$

1.8 Normal Error Regression Model

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

with ε_i are i.i.d and $\varepsilon_i \sim N(0, \sigma^2)$.

- $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, and $\{Y_i, i=1,2,\ldots n\}$ are independent

$$f(y_i) = f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left(y_i - (\beta_0 + \beta_1 X_i)\right)^2}{2\sigma^2}\right\} \quad i = 1,\ldots,n$$

- Likelihood:

$$L\left(\beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^{n} f(y_i) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 X_i)\right)^2\right\}$$

Maximum Likelihood estimators (MLEs)

Goal： select β_0 , β_1 , σ^2 to maximize L(or equivalenrly ln$L$)

$$l = \ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 X_i)]^2$$

We must select β_0 , β_1 to minimize

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2$$

Method of least square

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0, \beta_1}(l) = \arg\min_{\beta_0, \beta_1}\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 X_i)]^2 = (b_0, b_1)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \beta_1 X_i\right)\right)^2 \overset{set}{=} 0 \quad \Rightarrow$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{n-2}{n} MSE$$

- **MLEs**

$$\hat{\beta}_1 = b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

$$\hat{\beta}_0 = b_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{SS_E}{n} = \frac{n-2}{n} MSE$$

Properties of MLEs

In normal error regression model,

(1) MLEs of β_0 and β_1 are same with LSE estimators b_0 and b_1. They are linear combinations of $\{Y_i\}$.

(2) MLEs of β_0 and β_1 are BLUEs and normal distributed

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{SS_{XX}} \begin{pmatrix} \frac{1}{n}\sum X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix} \right)$$

(3) MSE of σ^2 is a biased estimator with

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2(n-2) \quad \text{and} \quad E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2 \to \sigma^2$$

(4) $(\hat{\beta}_0, \hat{\beta}_1, \bar{Y})$ and $\hat{\sigma}^2$ (or SSE) are independent.

Fisher's Theorem

(Fisher's Theorem) Let $X_1, X_2, \ldots, X_n$ be independent $N(\mu_i, \sigma^2)$ distributed random variables, and $Q=Q_1+Q_2+\ldots+Q_k$, where $Q, Q_1, Q_2, \ldots, Q_k$ are quadratic forms in $X_1, X_2, \ldots, X_n$, i.e., $Q=\mathbf{X}'\mathbf{A}\mathbf{X}$, and $Q_i=\mathbf{X}'\mathbf{A}_i\mathbf{X}$, $i=1,2,\ldots,k$. If

$$Q/\sigma^2 \sim \chi^2(r),\ Q_1/\sigma^2 \sim \chi^2(r_1), \cdots, Q_{k-1}/\sigma^2 \sim \chi^2(r_{k-1}),$$

then

(1) $Q_1, Q_2, \ldots, Q_k$ are independent.

(2) $Q_k/\sigma^2 \sim \chi^2(r_k)$, where $r_k = r-(r_1+\cdots+r_{k-1})$.

Fisher's Theorem is valid even if the quadratic forms are noncentral chi-square distributed.

Properties (3-4) of MLEs can be derived by Fisher's theorem.

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 X_i = \beta_0^* + \beta_1(X_i - \bar{X}), \qquad \beta_0^* = \beta_0 + \beta_1 \bar{X}$$

$$\hat{\beta}_0^* = \bar{Y} \sim N(\beta_0^*, \sigma^2 / n), \;\; \hat{\beta}_1 = SS_{XY} / SS_{XX} \sim N(\beta_1, \sigma^2 / SS_{XX}),$$

$$\sum (Y_i - \mu_i)^2 = \sum [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \mu_i)]^2$$

$$= \sum (\hat{Y}_i - \mu_i)^2 + \sum (Y_i - \hat{Y}_i)^2$$

$$= \sum [\hat{\beta}_0^* + \hat{\beta}_1(X_i - \bar{X}) - \beta_0^* - \beta_1(X_i - \bar{X})]^2 + SS_E$$

$$= n(\hat{\beta}_0^* - \beta_0^*)^2 + (\hat{\beta}_1 - \beta_1)^2 SS_{XX} + n\hat{\sigma}^2$$

$\boxed{Q_1}$ $\boxed{Q_2}$ $\boxed{Q_3}$

$$\begin{array}{llll} Q/\sigma^2 = & Q_1/\sigma^2 + & Q_2/\sigma^2 + & Q_3/\sigma^2 \\ \chi^2(n) & \chi^2(1) & \chi^2(1) & \boxed{\chi^2(n-2)} \end{array}$$

then Q_3 is chi-square distributed and Q_1, Q_2 , Q_3 are independent.

- $\bar{Y}(=\hat{\beta}_0^*), \hat{\beta}_1, \hat{\sigma}$ are independent each other.
- $(\hat{\beta}_0, \hat{\beta}_1)$ is independent with $\hat{\sigma}$.
- $\hat{\sigma}^2$ is biased estimator with

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2), \qquad E(\hat{\sigma}^2) = \frac{\sigma^2}{n} E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = \frac{n-2}{n}\sigma^2.$$

R code

```
toluca = read.table('D:\\Reg_licx\\Data_4e\\CH01TA01.txt',header=F)
names(toluca)<-c("Size", "Hours")   ##Change the column names
plot(toluca,xlim=c(0,150),ylim=c(0,600))  ##Scatter Plot

####Doing linear regression using R function lm()
fit = lm(Hours~Size, data=toluca);   summary(fit)
resi = fit$residuals   ##Residuals
yfit = predict(fit)   ##fitted values

####Verify the property of residuals
x = toluca[,1]
sum(resi);    sum(x*resi);    sum(yfit*resi)
```

Homework

- Under the linear regression model (1.1) with error distribution unspecified (in which the errors have expectation zero and are uncorrelated and have equal variances σ^2), calculate

  (1) the expectations of random variables SS_{YY} and SS_{XY}

  (2) $\operatorname{cov}(e_i, e_j), i \neq j.$
- pg 35~39: 1.21, 1.33, 1.34, 1.39, 1.41
- Optional: Show least square estimator b_0 is BLUE of β_0 in model (1.1) with error distribution unspecified.