Two Variable Regression Analysis

Basic Ideas

Hypothetical Example

Concept of Population Regression Function

Meaning of Term “Linear”

Stochastic Specification of PRF

Significance of the Stochastic Disturbance Term

Sample Regression Function (SRF)

Hypothetical Example

A hypothetical Country:

Total population of 60 families.

Y = Weekly family consumption expenditure

X = Disposable family income

Divide 60 families into 10 groups of same income

Hypothetical Example

$Y \downarrow$ \ $X \rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y\|X)$	65	77	89	101	113	125	137	149	161	173

Conditional Probability: $\Pr(Y = 55 \mid X = 80) = \frac{1}{5} = 0.2$

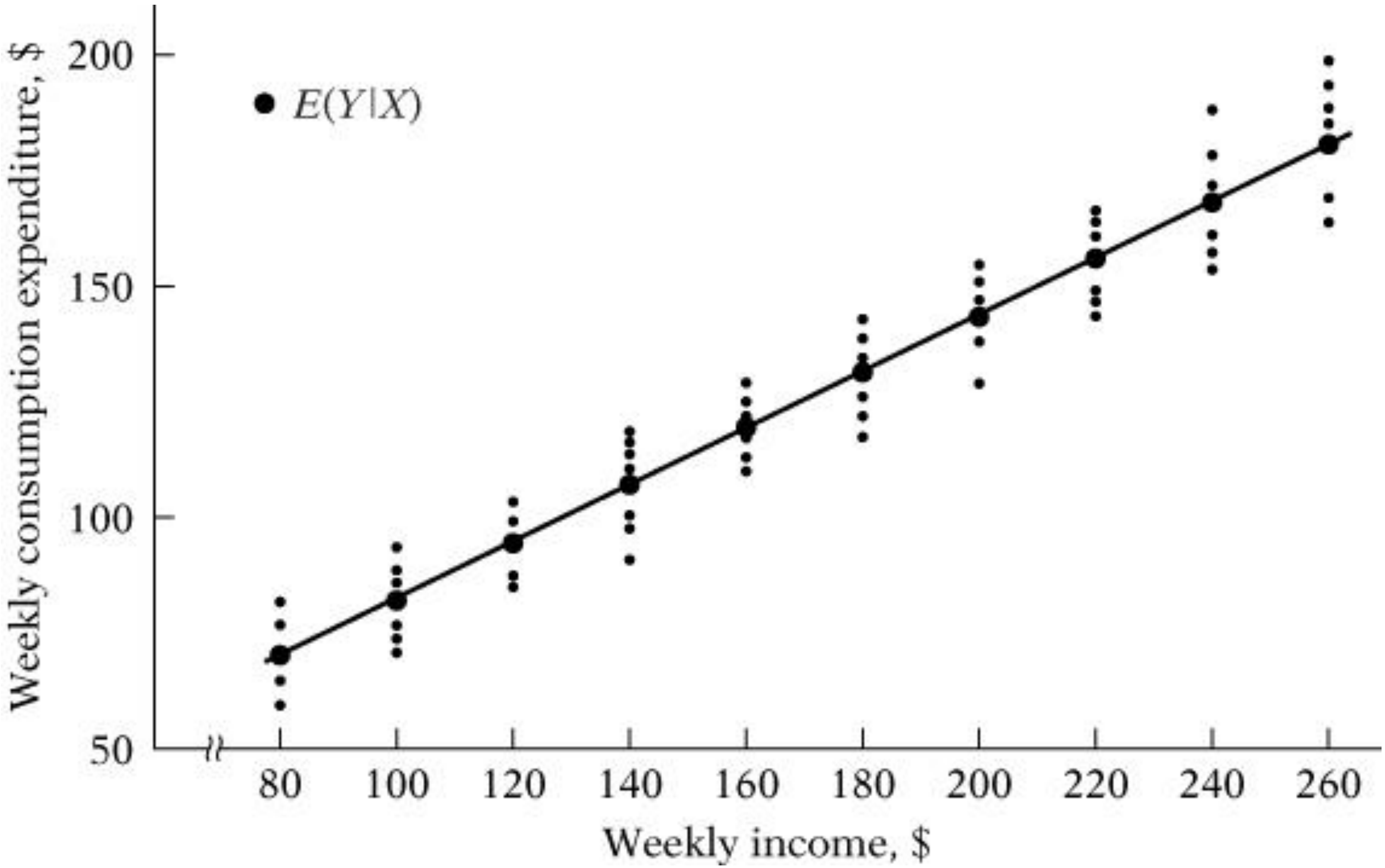
Conditional Mean:

$$E(Y \mid X = 80) = 55(0.2) + 60(0.2) + 65(0.2) + 70(0.2) + 75(0.2) = 65$$

Hypothetical Example

$X \rightarrow$ / $p(Y \mid X_i) \downarrow$	80	100	120	140	160	180	200	220	240	260
Conditional probabilities $p(Y \mid X_i)$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	—	$\frac{1}{6}$	—	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	—	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	—	—	—	$\frac{1}{7}$	—	—	—	$\frac{1}{7}$	—	$\frac{1}{7}$
Conditional means of Y	65	77	89	101	113	125	137	149	161	173

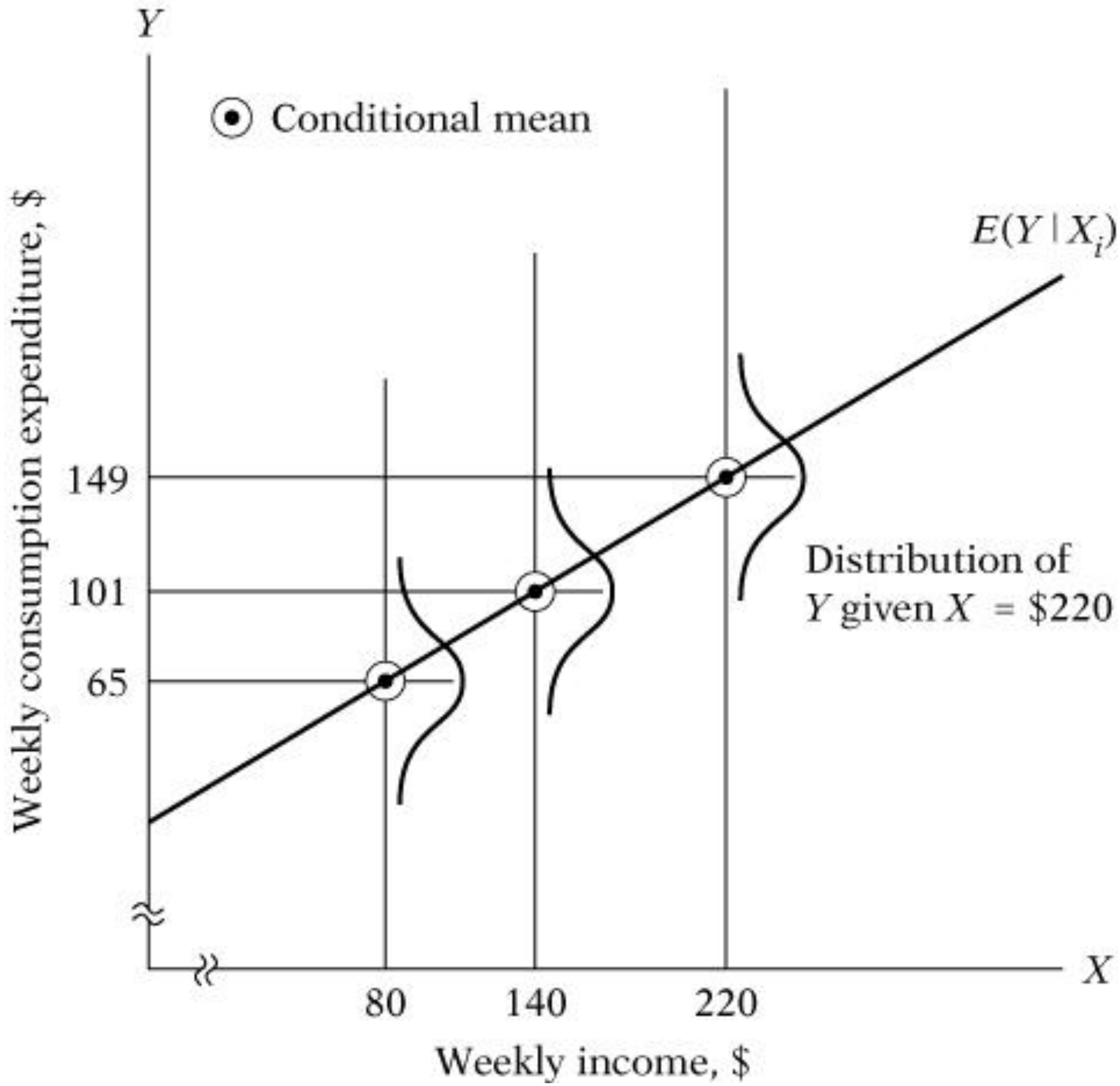
Hypothetical Example



Hypothetical Example

$p(Y \mid X_i)$ ↓ \ X →	80	100	120	140	160	180	200	220	240	260
Conditional probabilities $p(Y \mid X_i)$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	–	$\frac{1}{6}$	–	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	–	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	–	–	–	$\frac{1}{7}$	–	–	–	$\frac{1}{7}$	–	$\frac{1}{7}$
Conditional means of Y	65	77	89	101	113	125	137	149	161	173

Hypothetical Example



Concept of Population Regression Function

Conditional Mean $E(Y|X_i)$ is function of X_i

$$E(Y|X_i) = f(X_i)$$

Population Regression Function $E(Y|X_i)$ is linear function of X_i

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

where β_1 and β_2 are unknown but fixed parameters or regression coefficient.

Meaning of Term "Linear"

Linearity in Variables

Linear function: $E(Y \mid X_i) = \beta_1 + \beta_2 X_i$

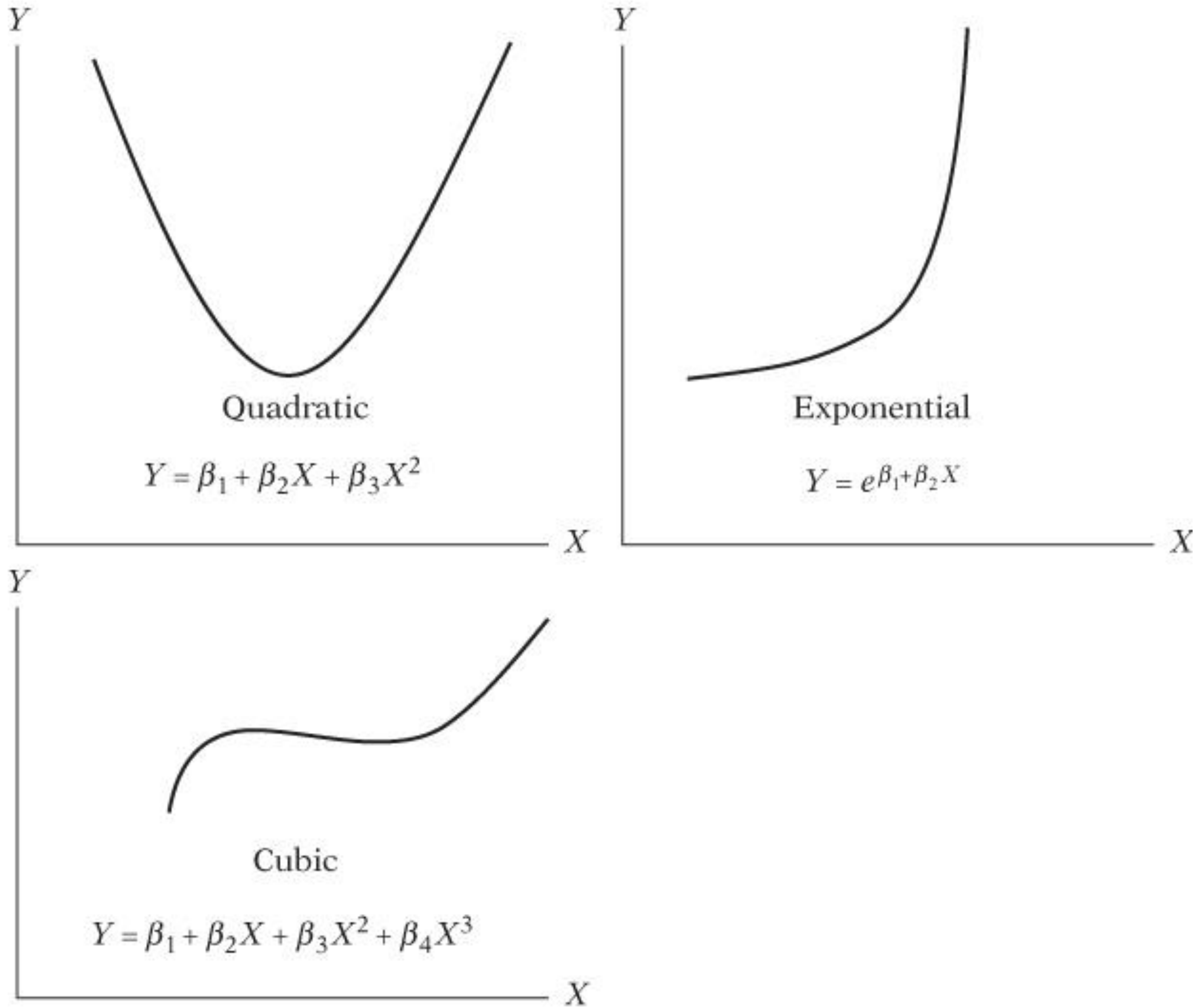
Nonlinear function: $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$

Linearity in Parameters

Linear Regression: $E(Y \mid Z_i = X_i^2) = \beta_1 + \beta_2 Z_i$

Nonlinear in parameter: $E(Y \mid X_i) = \beta_1 + \sqrt{\beta_2}\, X_i$

Examples: Linear-in-Parameters



Examples:

$Y \downarrow$ \ $X \rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y, \$	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y\mid X)$	65	77	89	101	113	125	137	149	161	173

$$Y_1 = 55 = \beta_1 + \beta_2(80) + u_1$$
$$Y_2 = 60 = \beta_1 + \beta_2(80) + u_2$$
$$Y_3 = 65 = \beta_1 + \beta_2(80) + u_3$$
$$Y_4 = 70 = \beta_1 + \beta_2(80) + u_4$$
$$Y_5 = 75 = \beta_1 + \beta_2(80) + u_5$$

Deviation of Y from its conditional mean

$$Y_i = 55, E(Y \mid X_i = 80) = 65$$

$$u_i = Y_i - E(Y \mid X_i)$$

Stochastic Specification of PRF

$$u_i = Y_i - E(Y|X_i)$$

or $$Y_i = E(Y|X_i) + u_i$$

where:

Deviation u_i is unobservable random variable or stochastic disturbance

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

where: $$E(u_i|X_i) = 0$$

Significance of the Stochastic Disturbance Term

1. Vagueness of theory
2. Unavailability of Data
3. Core variables vs. Control variables
4. Intrinsic randomness in human behavior
5. Poor proxy variables
6. Principle of Parsimony
7. Wrong functional form

Sample Regression Function (SRF)

Observing sample from population,we might get sample that does not represent population.

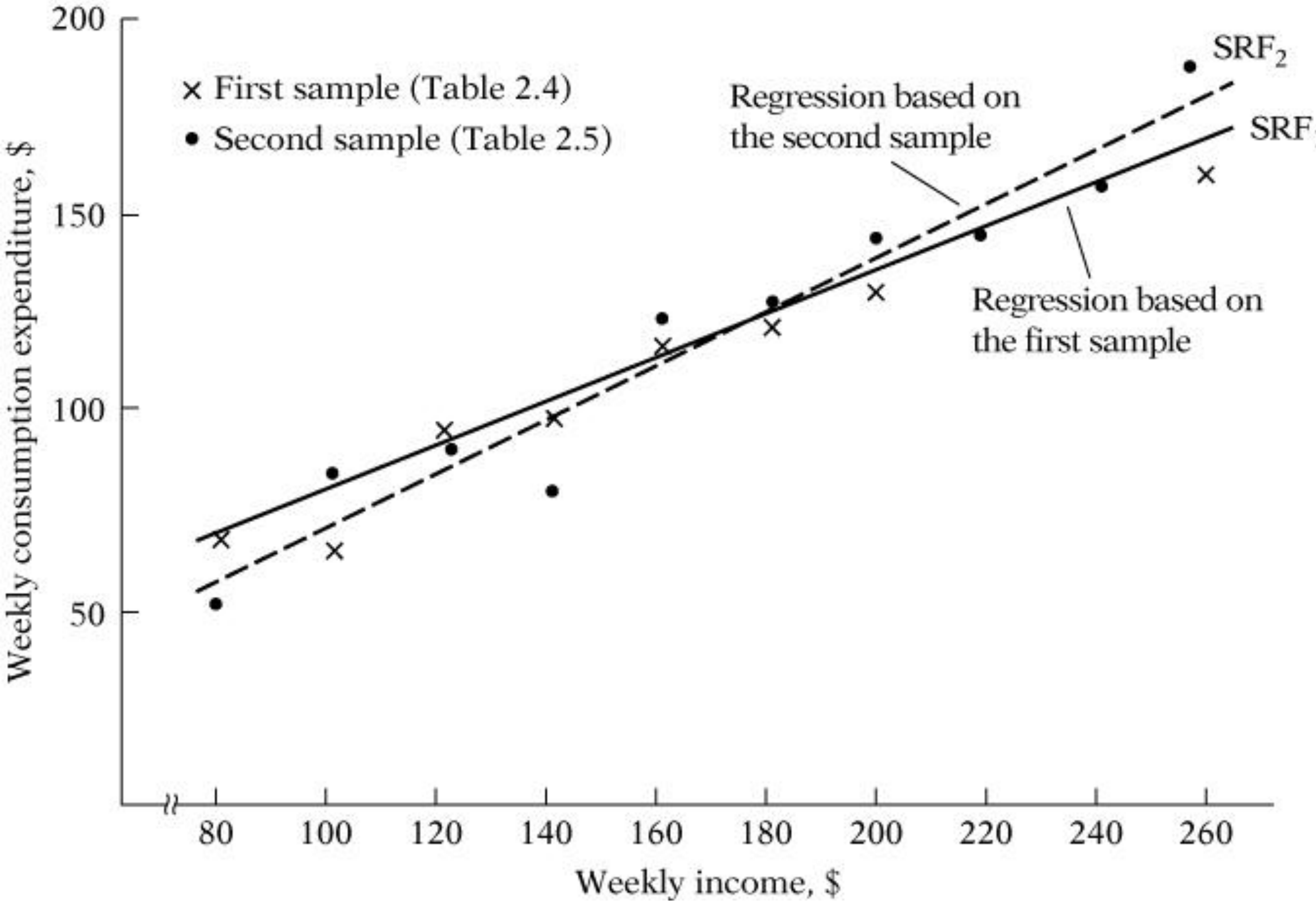
TABLE 2.4
A Random Sample from the Population of Table 2.1

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

TABLE 2.5
Another Random Sample from the Population of Table 2.1

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

Sample Regression Function (SRF)



Sample Regression Function (SRF)

Sample regression line or Sample regression function (SRF): $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

where $\hat{Y}_i$ = estimator of $E(Y|X_i)$

$\hat{\beta}_1$ = estimator of β_1

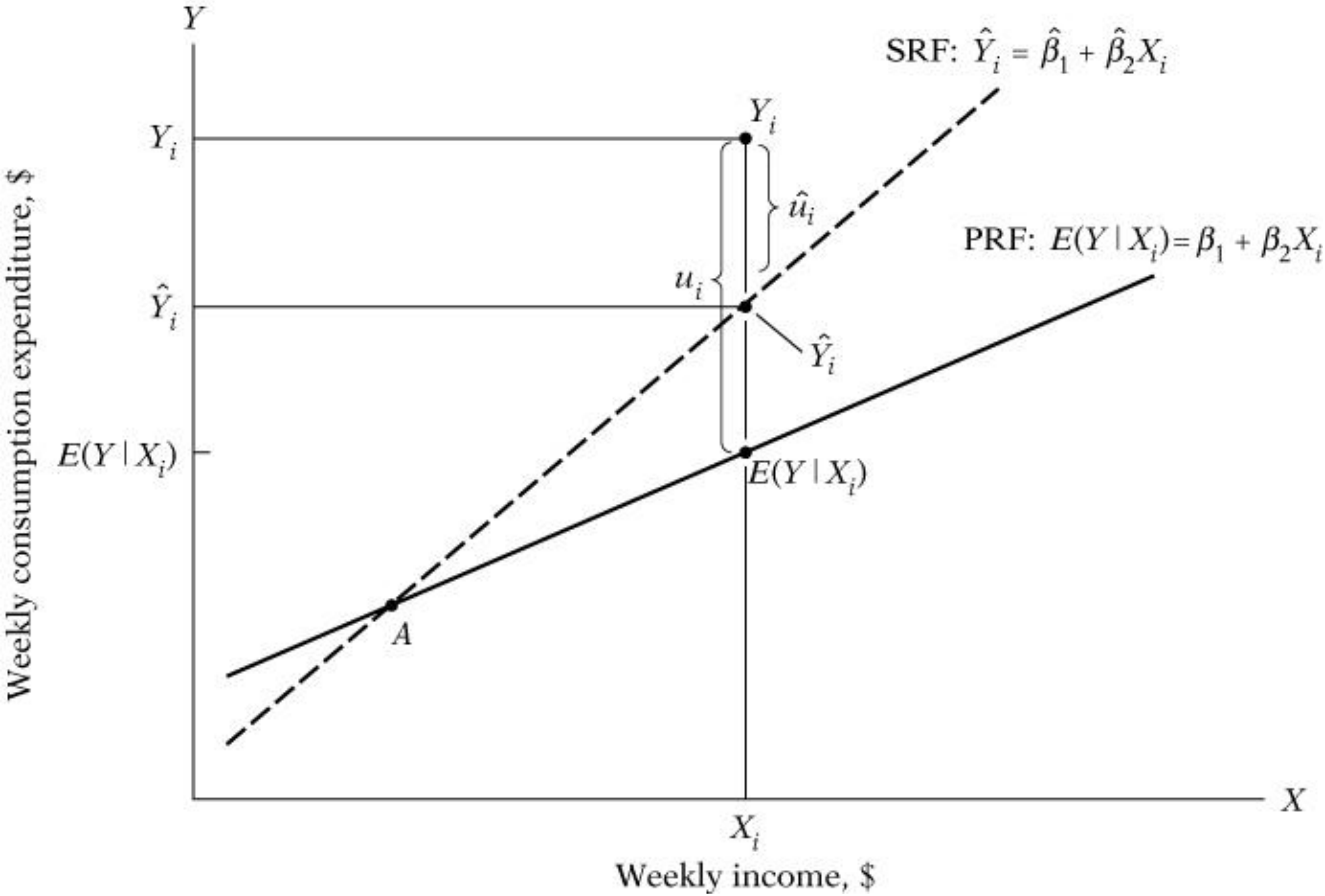
$\hat{\beta}_2$ = estimator of β_2

SRF: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

PRF: $E(Y|X_i) = \beta_1 + \beta_2 X_i$

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

Sample Regression Function (SRF)



Two Variable Regression Model

Problems of Estimation

1. Method of Ordinary Least Squares
2. Classical Linear Regression Model (CLRM):
   Assumptions Underlying the Method of Least Squares
3. Standard Errors of Least-Squares Estimates
4. Properties of Least-Squares Estimators:
   Gauss-Markov Theorem

Method of Ordinary Least Squares

From $$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

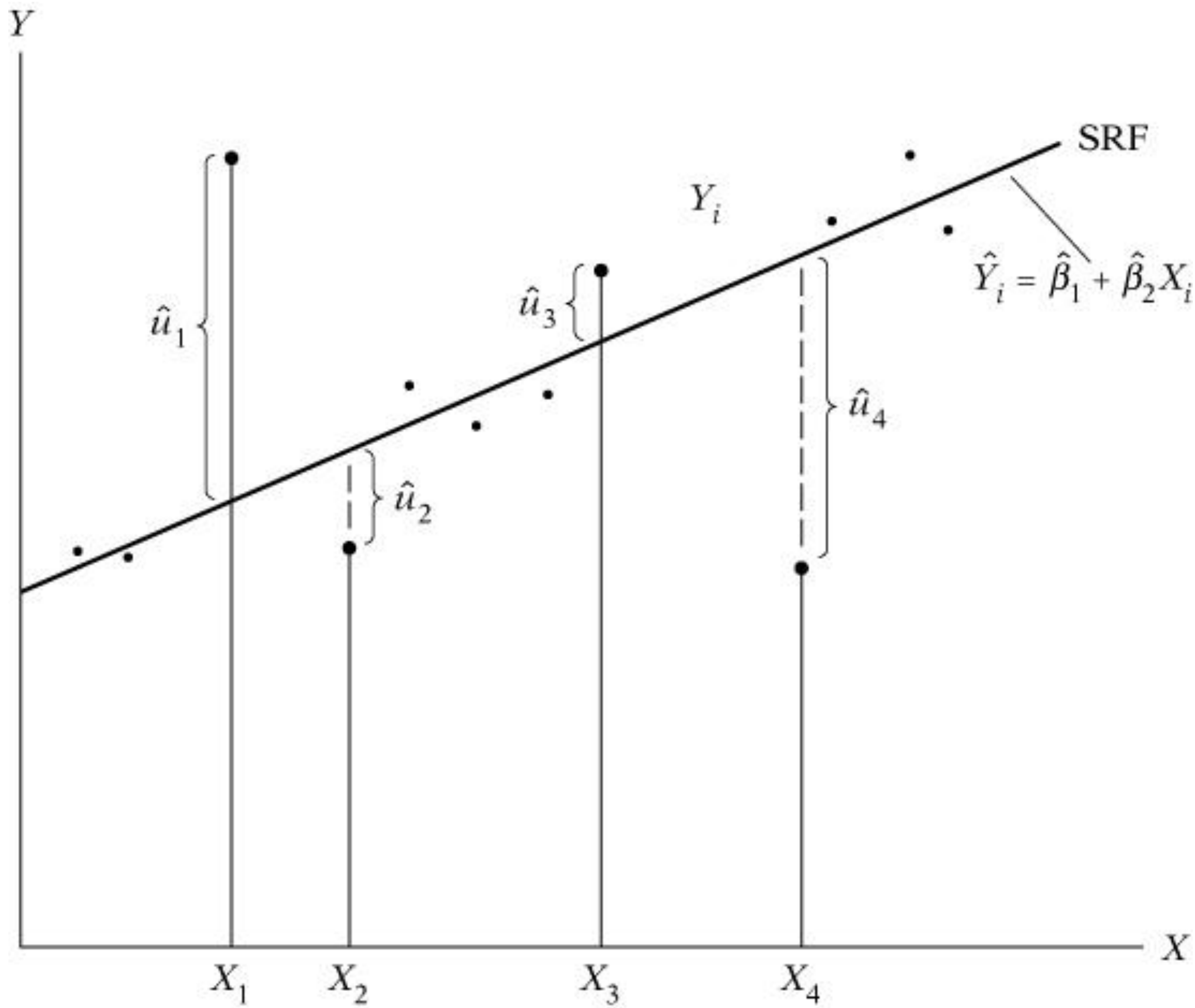
Then $$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

OLS try to minimize sum of squares residuals

$$\sum \hat{u}_i^2 = \sum \left(Y_i - \hat{Y}_i \right)^2 \to 0$$

$$If \quad \sum \hat{u}_i^2 = 0 \quad then \quad Y_i = \hat{Y}_i$$

Method of Ordinary Least Squares



Method of Ordinary Least Squares

$$Set \quad \frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = 0 \quad then \quad \sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$Set \quad \frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = 0 \quad then \quad \sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

Solve these normal equations:

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

Method of Ordinary Least Squares

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

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

$$= \frac{\sum x_i y_i}{\sum x_i^2}$$

Method of Ordinary Least Squares

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

$$= \bar{Y} - \hat{\beta}_2 \bar{X}$$

Method of Ordinary Least Squares

Statistical Properties of OLS Estimators

1. Easily to compute
2. Point Estimators

Properties of Regression Line

1. Pass through the sample mean of Y and X
2. Mean of estimated Y_i equals to mean of actual Y_i
3. Mean value of residual is zero
4. Residuals are uncorrelated with predicted Y_i
5. Residuals are uncorrelated with X_i

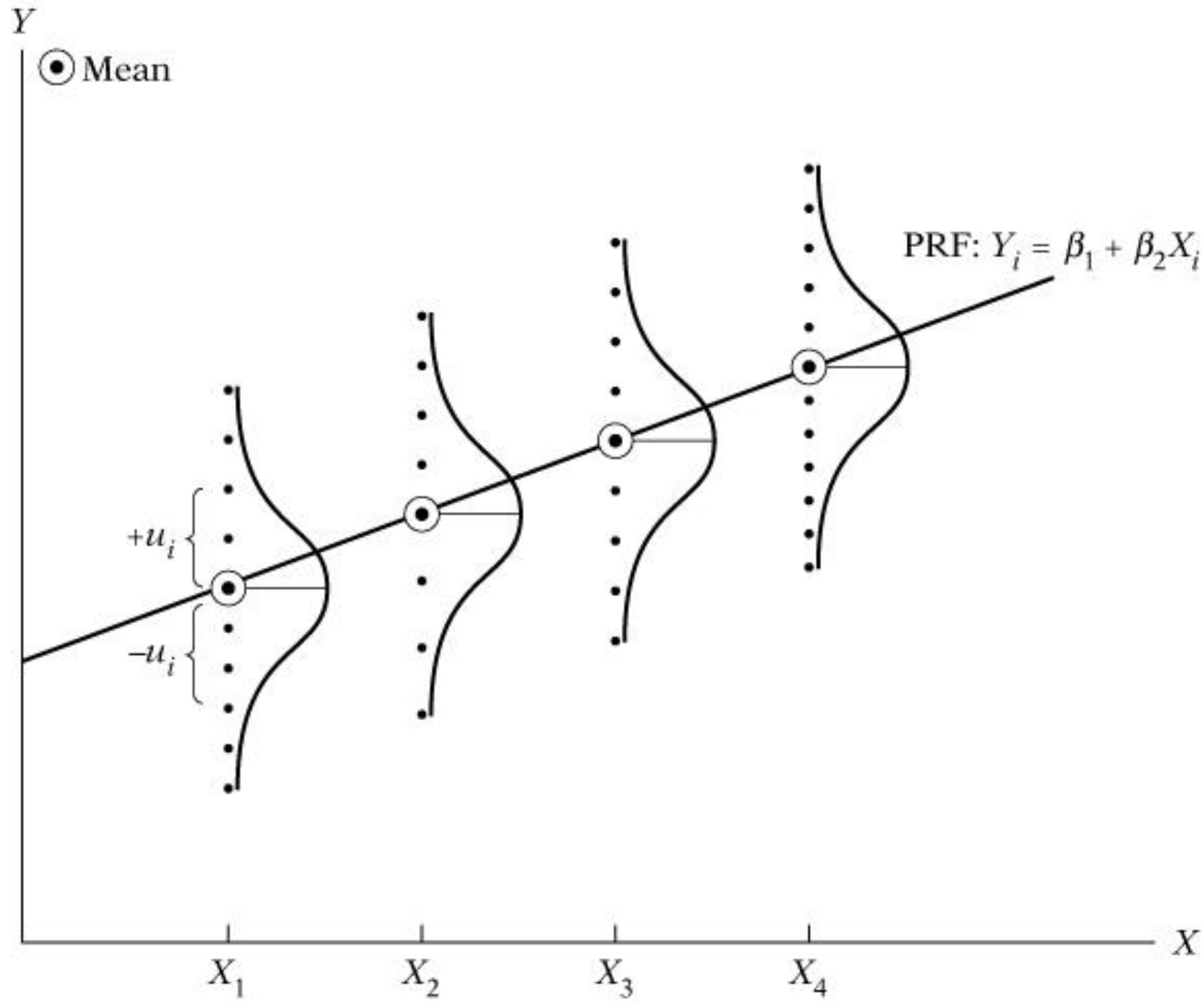
Assumptions of Least Squares

1. Linear Regression Model
2. X_i values are fixed in repeated sampling
3. Zero mean value of disturbance u_i
4. Homoscedasticity or equal variance of u_i
5. No autocorrelation between the disturbance
6. Zero covariance between u_i and X_i
7. Number of observations must be greater than number of parameters to be estimated
8. Variability in X_i values
9. Regression model is correctly specified
10. No perfect multicollinearity

Assumptions of Least Squares

1. Linear Regression Model
   - Linear in parameter
2. X values are fixed in repeated sampling
   - X_i is assumed to be *nonstochastic*
3. Zero mean value of disturbance u_i
   - $E\left(u_i \middle| X_i\right) = 0$

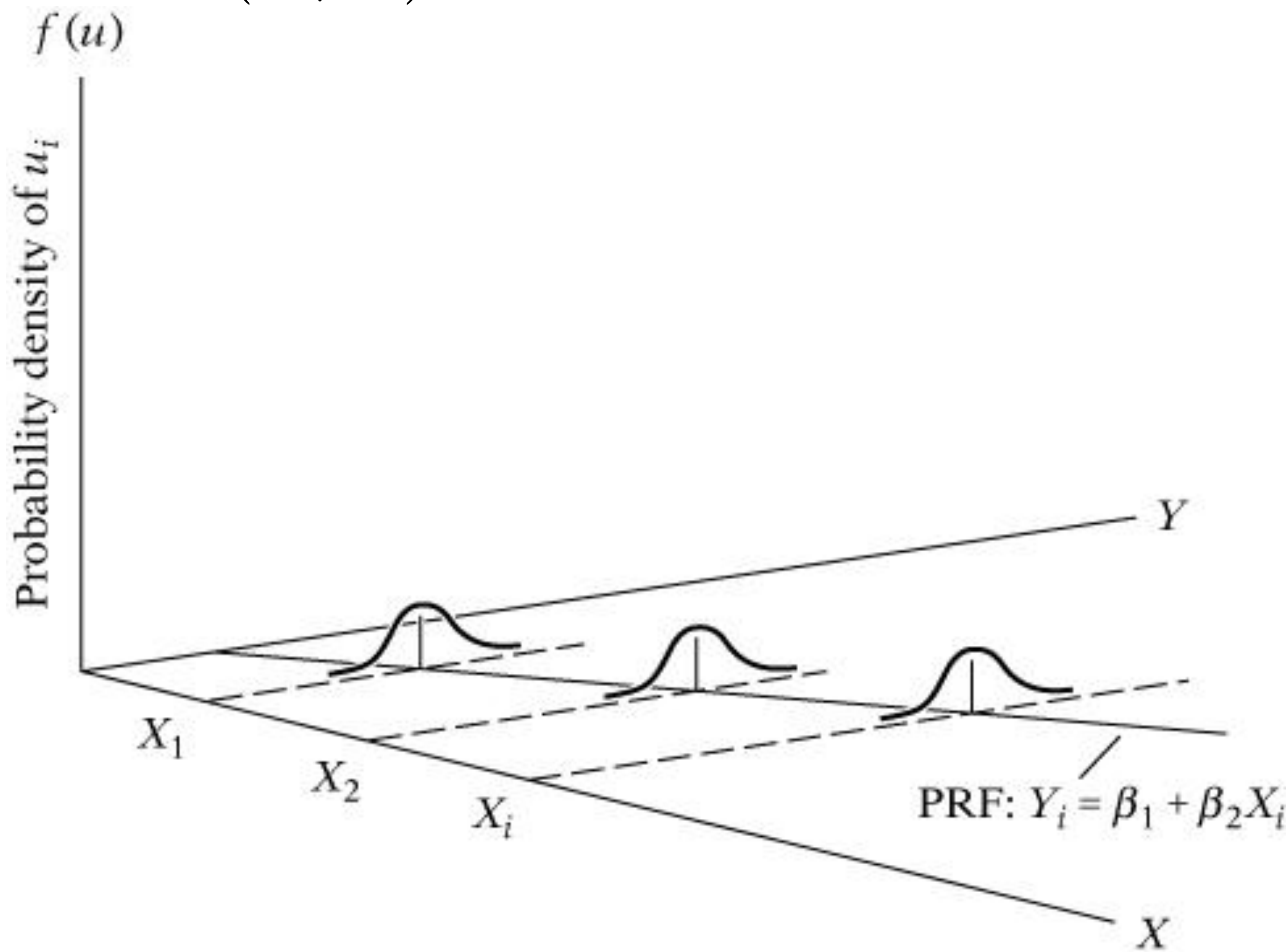
Assumptions of Least Squares



Assumptions of Least Squares

4. Homoscedasticity or equal variance of u_i

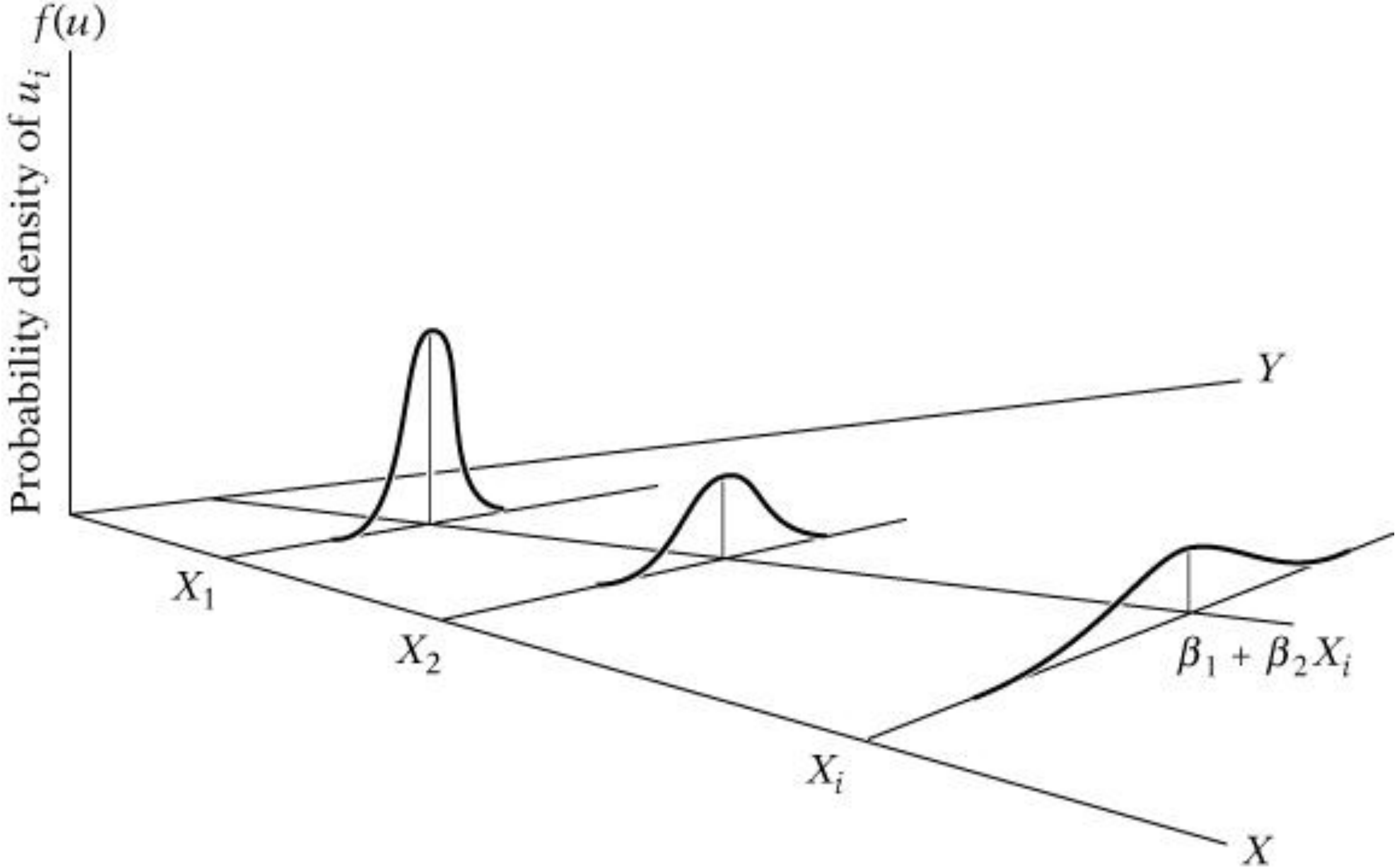
- $\text{var}\left(u_i \mid X_i\right) = \sigma^2$ or Constant variance



Assumptions of Least Squares

In case of violation or Heteroscedasticity

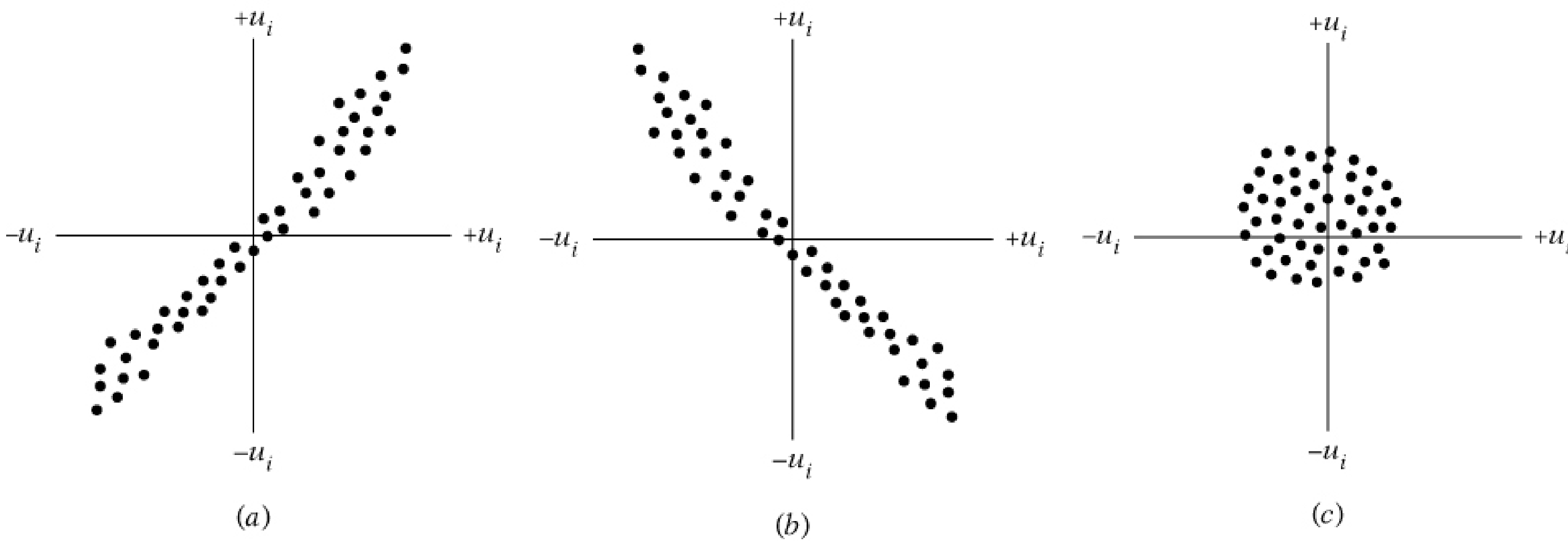
- $\mathrm{var}\left(u_i \middle| X_i\right) = \sigma_i^2$



Assumptions of Least Squares

5. No autocorrelation between the disturbances

- $\text{cov}\left(u_i u_j \mid X_i, X_j\right) = 0$



Positive Autocorrelation Negative Autocorrelation No Autocorrelation

Assumptions of Least Squares

6. Zero covariance between u_i and X_i

   - $E(u_i X_i) = 0$ then $\operatorname{cov}(u_i X_i) = 0$

7. Number of observations must be greater than number of parameters to be estimated

   Number of observations must be greater than number of explanatory variables

8. Variability in X_i values

   X_i values in given sample must not all be the same

9. Regression model is correctly specified

   No specification bias or error

Assumptions of Least Squares

10. No perfect multicollinearity

    There are no perfect linear relationships among the explanatory variables.

Standard Errors of Least-Squares

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$$

$$\operatorname{var}\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum x_i^2}$$

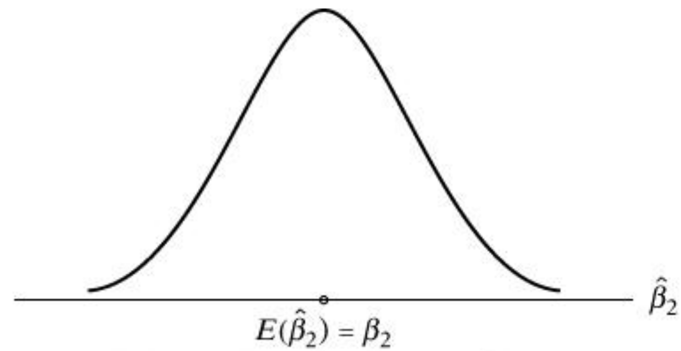
$$\operatorname{var}\left(\hat{\beta}_1\right) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$$

Properties of Least-Squares Estimator

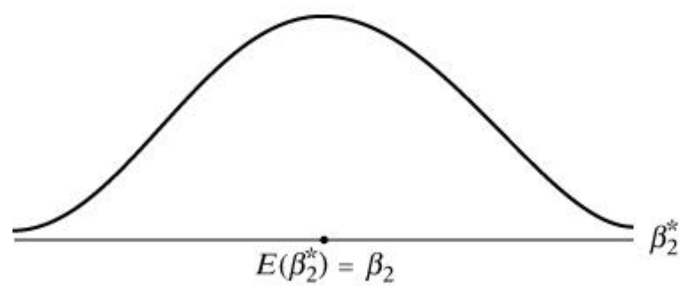
1. Linear
2. Unbiased
3. Efficient estimator

Gauss-Markov Theorem:

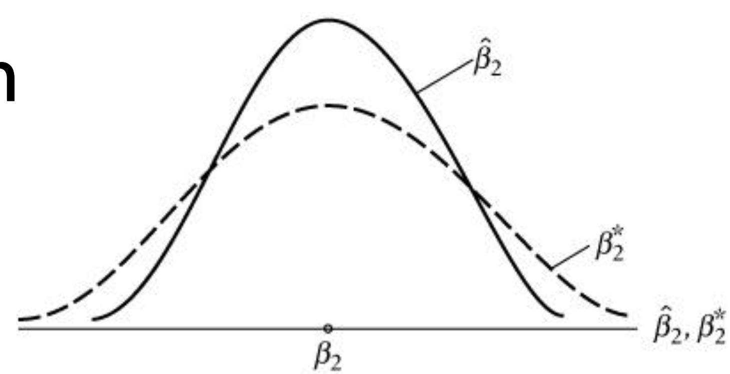
Given assumptions of CLRM, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance -- they are Best Linear Unbiased Estimators (BLUE).



(a) Sampling distribution of β_2



(b) Sampling distribution of β_2^*



(c) Sampling distributions of β_2 and β_2^*