

Statistics with Economics and Business Applications

Chapter 12 Simple Linear Regression

Simple Linear Regression Model, Least Squares Method, Coefficient of Determination, Test and Confidence Intervals, Estimation and Prediction

Introduction

So far we have done statistics on one variable at a time.

We now interested in **relationships** between two variables and how to use one variable to **predict** another variable.

- Does weight depend on height?
- Does blood pressure level predict life expectancy?
- Do SAT scores predict college performance?
- How do commercials affect sales?

Example: Age and Fatness

The following data was collected in a study of age and fatness in humans.

Age	23	23	27	27	39	41	45	49	50
% Fat	9.5	27.9	7.8	17.8	31.4	25.9	27.4	25.2	31.1
Age	53	53	54	56	57	58	58	60	61
% Fat	34.7	42	29.1	32.5	30.3	33	33.8	41.1	34.5

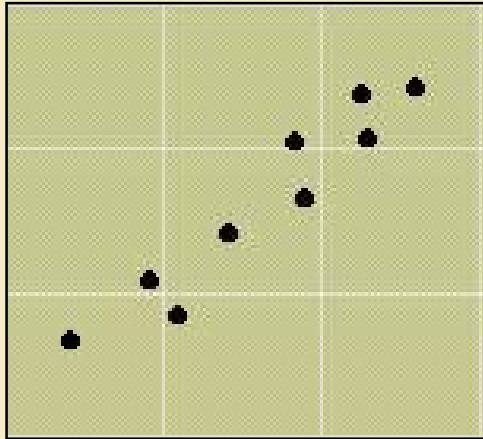
One of the questions was, “What is the relationship between age and fatness?”

Example: Age and Fatness

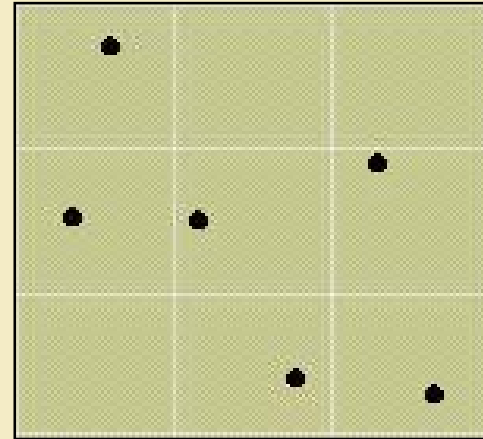
The following scatterplot shows that % fat in general tend to increase with age. The relationship is close, but not exactly, linear.



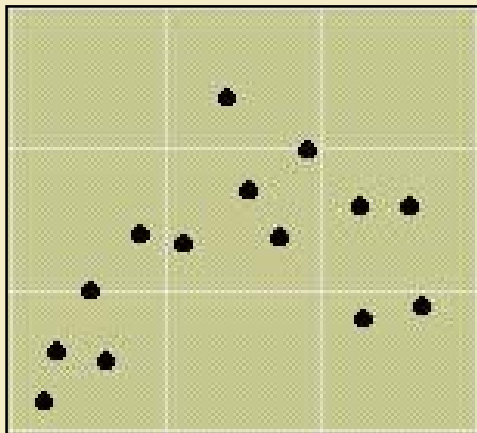
Shape and Trend



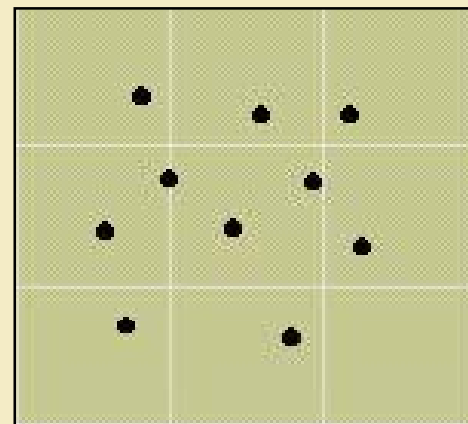
Positive linear - strong



Negative linear -weak



Curvilinear



No relationship

Investigation of Relationship

There are two approaches to investigate linear relationship

- Ch3: Correlation coefficient r_{xy} ---- a numerical measure of the **strength** and **direction** of the linear relationship between x and y .
- Ch12: Linear regression ---- a linear equation expresses the relationship between x and y . It provides a **form** of the relationship.

Review: Correlation Coefficient

The strength and direction of the relationship between x and y are measured using the **correlation coefficient r** .

$$r = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$s_{xx} = s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$$s_{yy} = s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1}$$

s_x = standard deviation of the X

s_y = standard deviation of the Y

s_{xy} = covariance of X and Y

Example

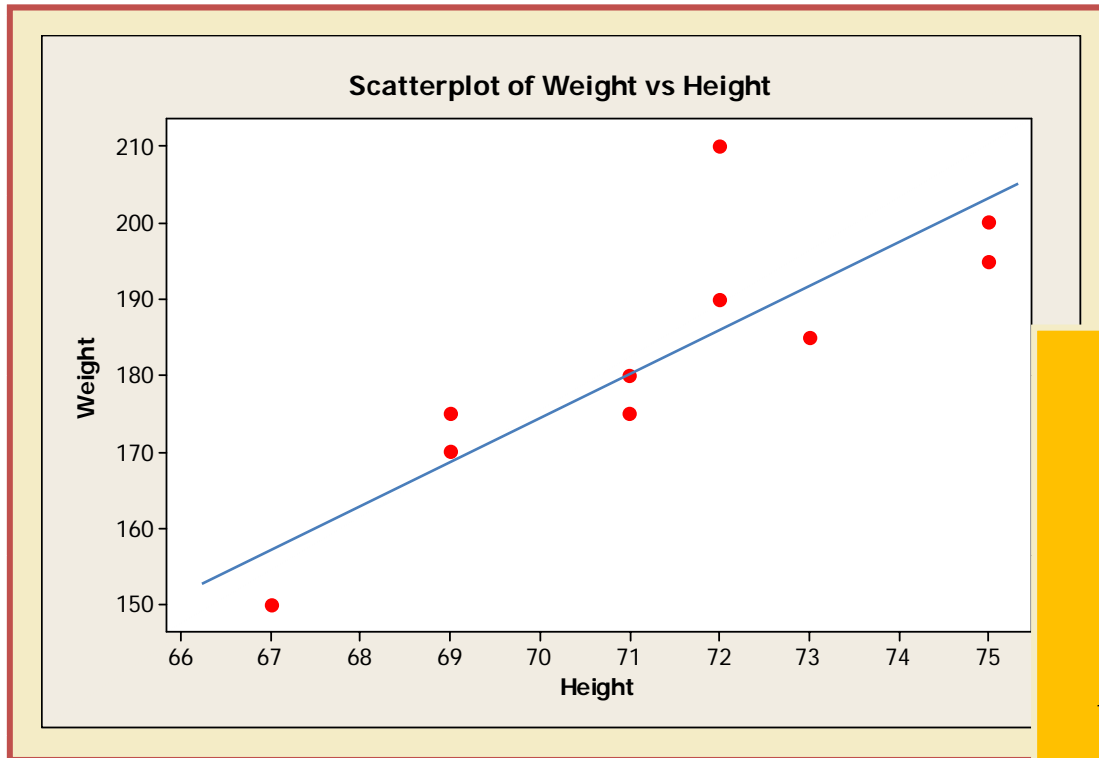
The table shows the heights and weights of $n = 10$ randomly selected college football players.

Player	1	2	3	4	5	6	7	8	9	10
Height, x	73	71	75	72	72	75	67	69	71	69
Weight, y	185	175	200	210	190	195	150	170	180	175

$$S_{xy} = 328 \quad S_{xx} = 60.4 \quad S_{yy} = 2610$$

$$r = \frac{328}{\sqrt{(60.4)(2610)}} = .8261$$

Football Players



$$r = .8261$$

Strong positive
correlation

As the player's height
increases, so does his
weight.



Interpreting r

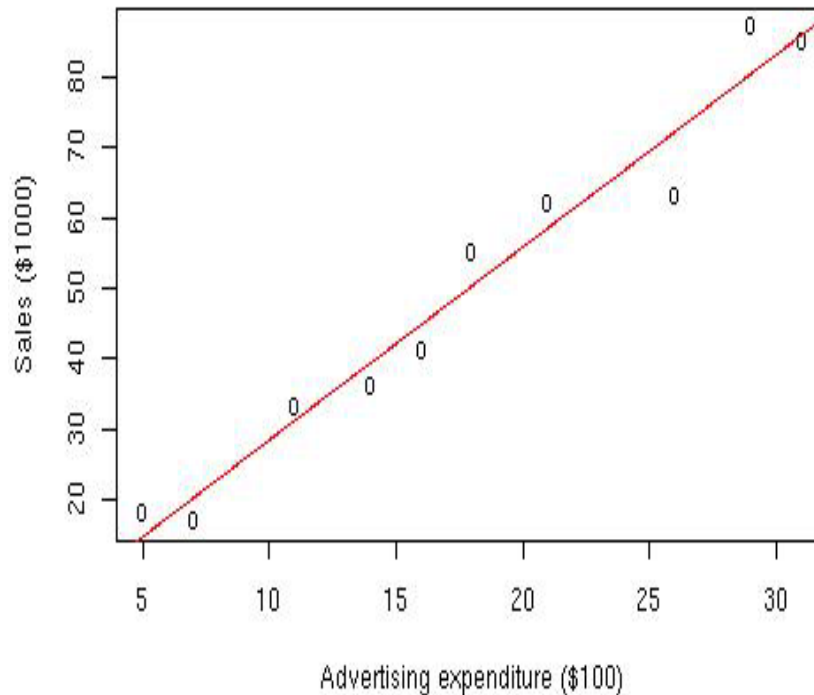
- $-1 \leq r \leq 1$ Sign of r indicates direction of the linear relationship.
- $r \approx 0$ No relationship; random scatter of points
- $r \approx 1$ or -1 Strong relationship; either positive or negative
- $r = 1$ or -1 All points fall exactly on a straight line.

Example: Advertising and Sale

The following table contains sales (y) and advertising expenditures (x) for 10 branches of a retail store.

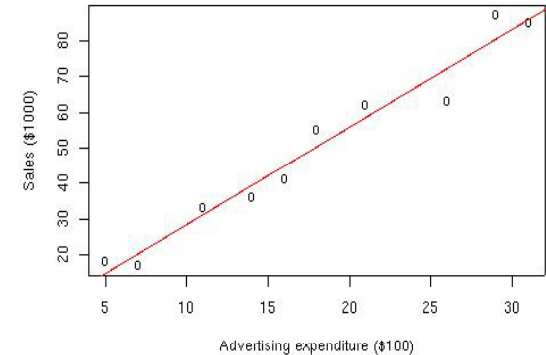
x (\$100)	5	7	11	14	16	18	21	26	29	31
y (\$1000)	17	18	33	36	41	55	62	63	85	87

**Scatter
Diagram:**



$$r_{xy} = 0.98$$

Example: Advertising expenditures and Sale

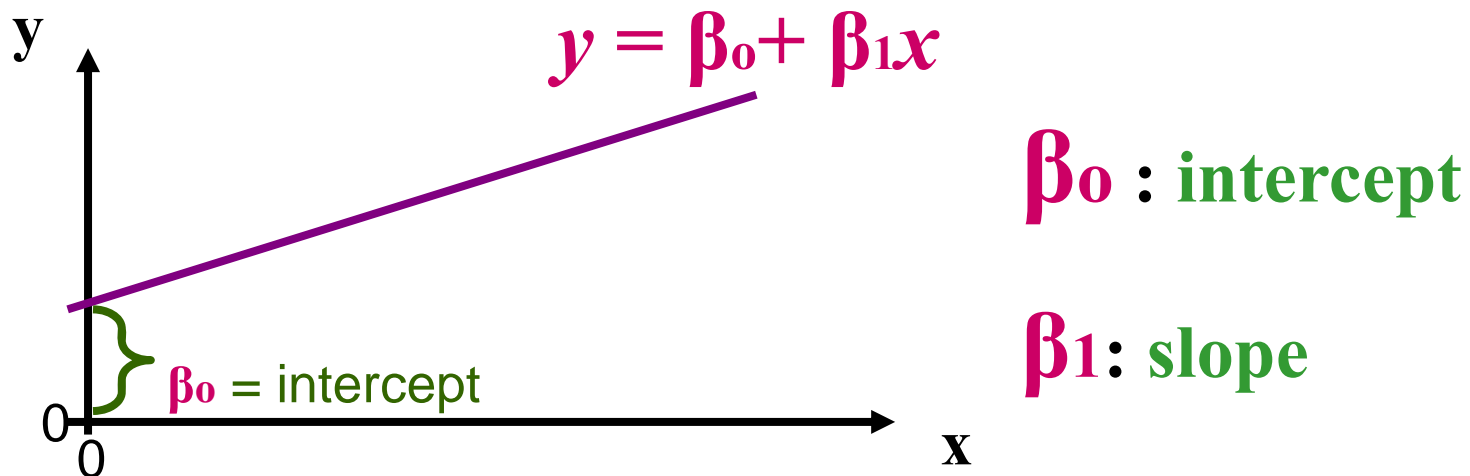


Often we want to investigate the **form** of the relationship for the purpose of **description**, **control** and **prediction**. For the advertising and sale example, sale increases as advertising expenditure increase. The relationship is almost, but not exact linear.

- **Description**: how sales depends on advertising expenditure
- **Control**: how much to spend on advertising to reach certain goals on sales
- **Prediction**: how much sales do we expect if we spend certain amount of money on advertising

Introduction to Linear Model

- x --- independent variables: variable being used to predict
 y --- dependent variable: variable being predicted
*For example, $x=age$ and $y=\%fat$;
 $x=advertising\ expenditure$, $y=sales$;*
- We want to find how y depends on x , or how to predict y by using x
- One of the simplest mathematical relationship between two variables x and y is a straight line



Reality

- The deterministic straight line is not adequate.
- Observations of (x, y) do not fall on a straight line.
 - Age influences fatness. But it is not the sole influence. There are other factors such as sex, body type and random variation (e.g. measurement error)
 - Other factors such as time of year, state of economy and size of inventory, besides the advertising expenditure, can influence the sale

We use an error term to represents random fluctuation in y from the straight line.

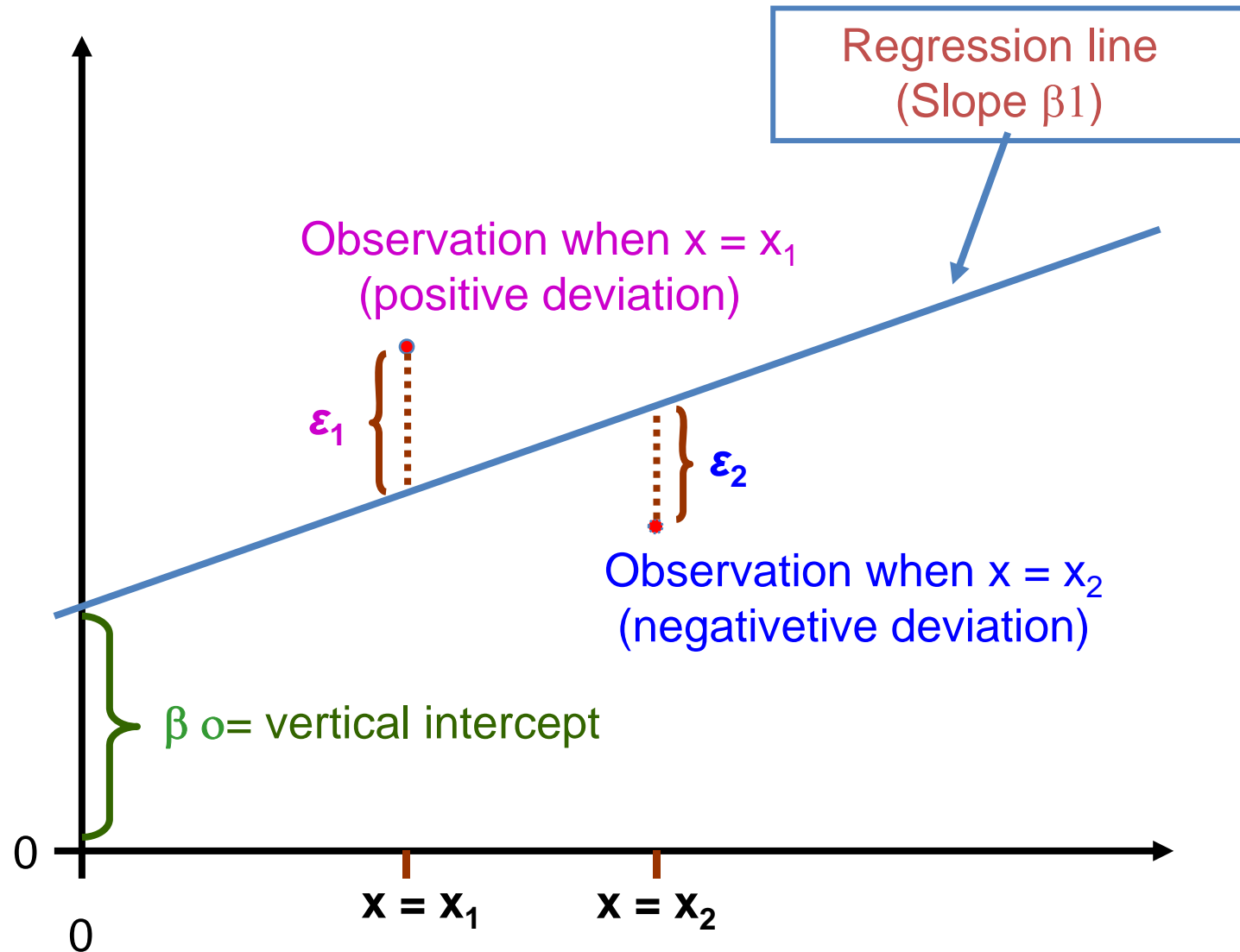
Simple linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- The error term ε accounts for the variability in y that cannot be explained by the linear relationship
- β_0 and β_1 are parameters of the model

Comments: Without the random deviation ε , all observed points (x, y) points would fall exactly on the deterministic line. The inclusion of ε in the model equation allows points to deviate from the line by random amounts.

Simple Linear Regression Model



Basic Assumptions of ε

The error term ε accounts for the variability in y that cannot be explained by the linear relationship between x and y . *ε is a random variable !*

1. The distribution of ε at any particular x value has mean value 0. i.e. $E(\varepsilon)=0$
2. The standard deviation of ε is the same for any particular value of x . This standard deviation is denoted by σ . i.e. $\text{Std. Dev}(\varepsilon)=\sigma$
3. The distribution of ε at any particular x value is normal.
4. The random errors are independent of one another.

$$\overset{\text{indep}}{\varepsilon} \sim \mathbf{N}(0, \sigma^2)$$

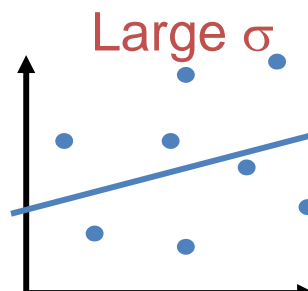
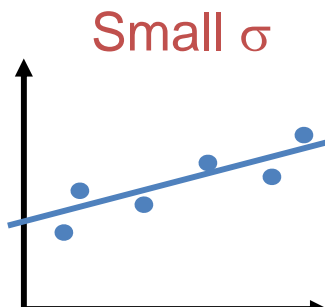
Interpretation of Terms

1. The straight line $\beta_0 + \beta_1 x$ describes **expected value** of y for any fixed value of x . (*Based on the assumption $E(\varepsilon) = 0$*)

Simple Linear Regression Equation:

$$E(y) = \beta_0 + \beta_1 x$$

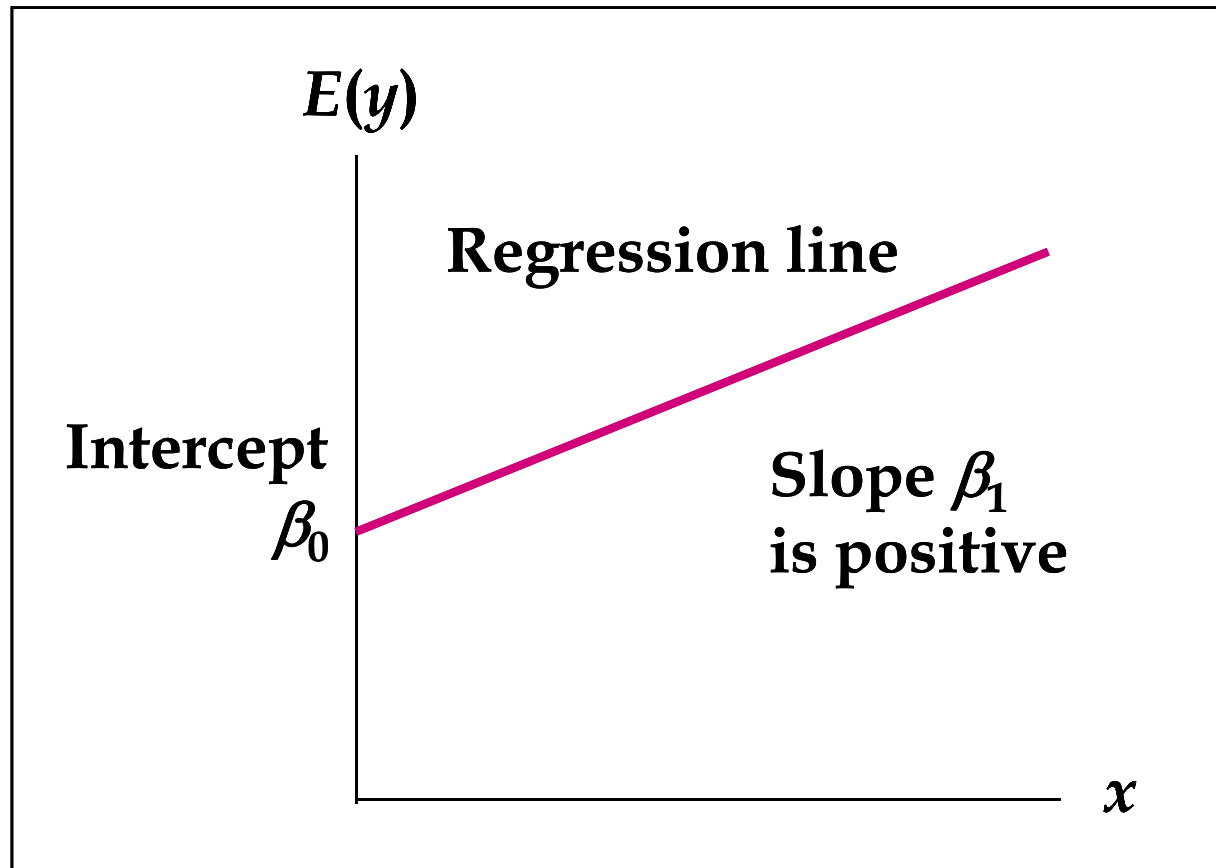
2. The **slope** β_1 of the regression line is the **average** change in y associated with a **1**-unit increase in x . The **intercept** β_0 is the height of the line when $x = 0$.
3. The size of σ determines the extent to which (x, y) observations deviate from the population line.



σ = Std. dev of error ε

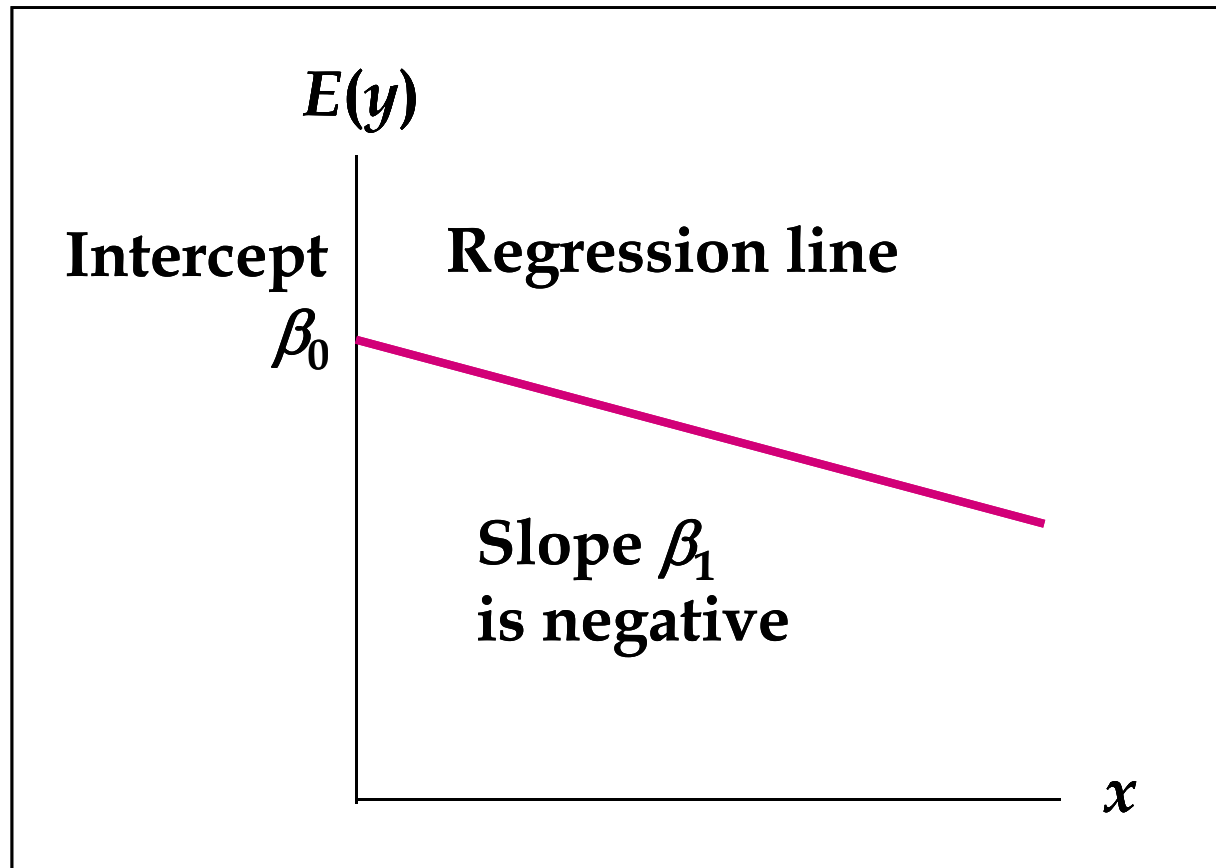
Simple Linear Regression Equation: $E(y) = \beta_0 + \beta_1 x$

Positive Linear Relationship



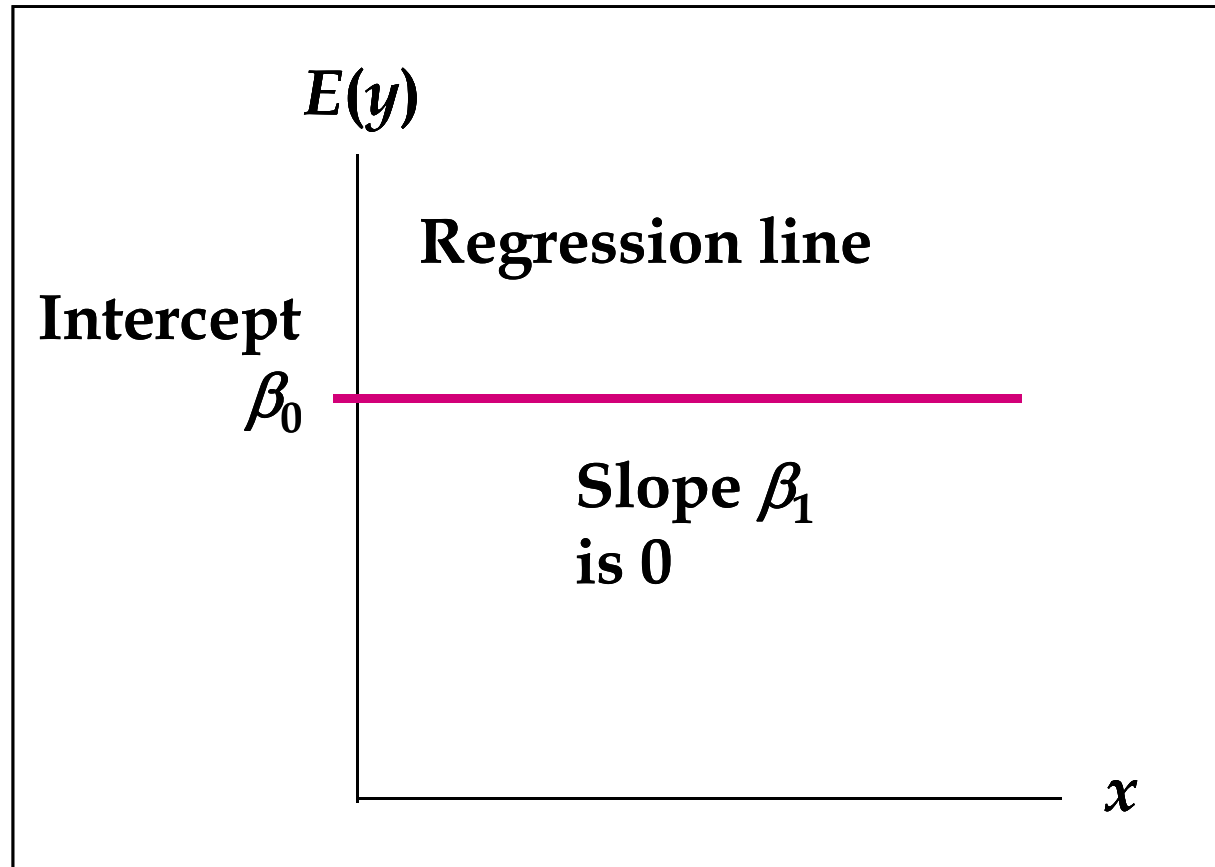
Simple Linear Regression Equation: $E(y) = \beta_0 + \beta_1 x$

Negative Linear Relationship



Simple Linear Regression Equation: $E(y) = \beta_0 + \beta_1 x$

No Relationship



Steps in Regression Analysis

To perform simple regression analysis, we usually use a step-by-step approach as below:

1. Fit the model to data – estimate parameters β_0 and β_1 . (12.2)
2. Use the analysis of variance t test (or F test) and r^2 to determine how well the model fits the data. (12.3, 12.5)
3. Proceed to estimate or predict the quantity of interest (12.7)
4. Use diagnostic plots to check for violation of the regression assumptions about ε . (12.8)

Data

1. Data: n pairs of observations of independent and dependent variables

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

2. Simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1, \dots, n$$

ε_i are independent normal with mean 0 and standard deviation σ .

3. Usually three parameters, β_0 , β_1 and σ , are unknown. We need to estimate them from data.

$$b_0 \longrightarrow \beta_0$$

$$b_1 \longrightarrow \beta_1$$

Simple Linear Regression

Example: Reed Auto Sales

Reed Auto periodically has a special week-long sale. As part of the advertising campaign, Reed runs one or more television commercials during the weekend preceding the sale.

Data from a sample of 5 previous sales are shown on the right.

<u>Number of TV Ads</u>	<u>Number of Cars Sold</u>
1	14
3	24
2	18
1	17
3	27

Least Squares Method

- The equation of the best-fitting line is calculated using n pairs of data (x_i, y_i) .
- We choose our estimates b_0 and b_1 to estimate β_0 and β_1 so that the vertical distances of the points from the line are minimized.

Least Squares Criterion

Estimated value of y for x_i : $\hat{y}_i = b_0 + b_1 x_i$

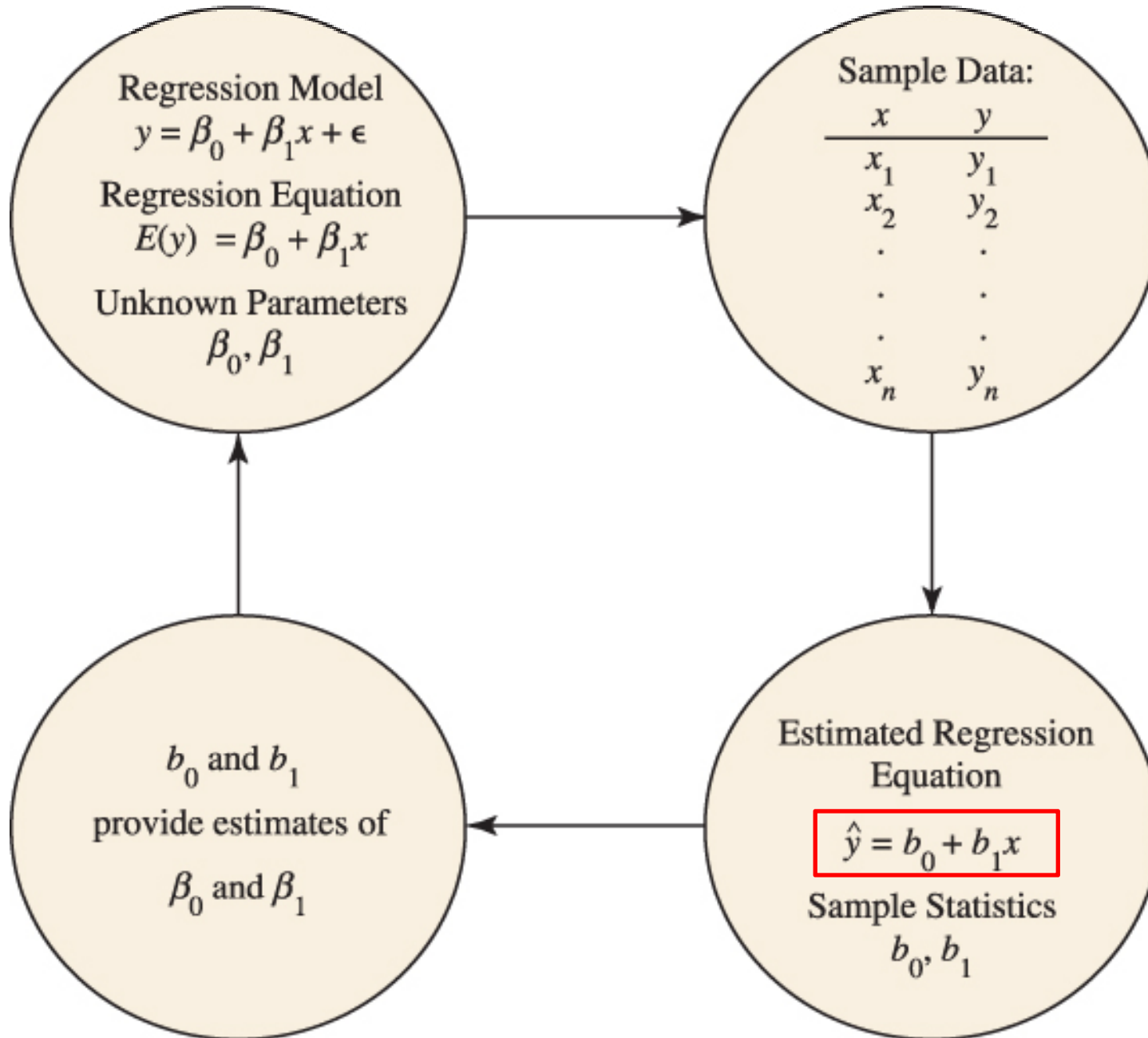
Choose b_0 and b_1 to minimize

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

**SSE --- sum of squares
due to error**

$$= \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

Estimation Process in Simple Linear Regression



Least Square Estimator for β_1 and β_0

- **Slope** for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- **y-Intercept** for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1 \bar{x}$$

where:

x_i = value of independent variable for i th observation

y_i = value of dependent variable for i th observation

\bar{x} = mean value for independent variable

\bar{y} = mean value for dependent variable

n = total number of observations

LS Estimator: : $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ is minimized for such b_0 and b_1

Simple Linear Regression

Example: Reed Auto Sales

Reed Auto periodically has a special week-long sale. As part of the advertising campaign, Reed runs one or more television commercials during the weekend preceding the sale.

Data from a sample of 5 previous sales are shown on the right.

<u>Number of TV Ads</u>	<u>Number of Cars Sold</u>
1	14
3	24
2	18
1	17
3	27

Estimated Regression Equation

Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{20}{4} = 5$$

y-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1\bar{x} = 20 - 5(2) = 10$$

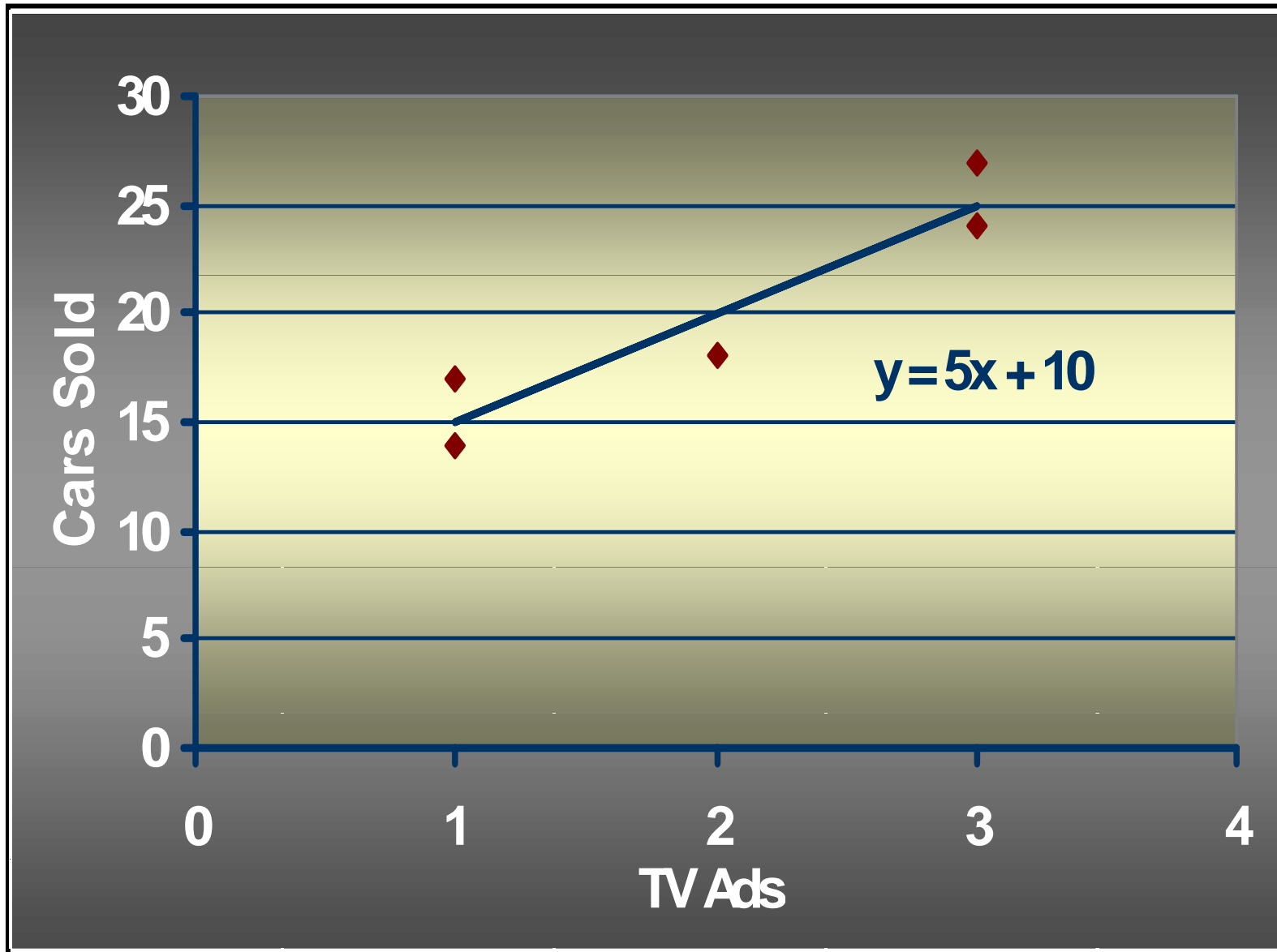
Estimated Regression Equation

$$\hat{y} = 10 + 5x$$

Comments:

- 1. Positive slope (b1=5) implies as number of commercials increases, sales increase.**
- 2. Can you predict the expected number of car sold if Reed run 3 TV commercials?**

Scatter Diagram and Trend Line



Least Squares Estimators

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i) / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

The second form is often recommended when using a calculator to compute b_1

Example: Ads and Sales

$$n=5, \sum x_i=10, \bar{x}=2, \sum y_i=100, \bar{y}=20$$

$$\sum x_i^2 = 1+9+4+1+9 = 24$$

$$\sum x_i y_i = 14+72+36+17+81=220$$

$$\text{numerator} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 220 - \frac{10*100}{5} = 20.$$

$$\text{denominator} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 24 - \frac{100}{5} = 4$$

Then

$$b_1 = \frac{20}{4} = 5$$

$$b_0 = \bar{y} - b_1 \bar{x} = 20 - 5*2 = 10$$

Steps in Regression Analysis

To perform simple regression analysis, we usually use a step-by-step approach as below:

1. Fit the model to data – estimate parameters β_0 and β_1 . (12.2)
2. Use the analysis of variance t test (or F test) and r^2 to determine how well the model fits the data. (12.3, 12.5)
3. Proceed to estimate or predict the quantity of interest (12.7)
4. Use diagnostic plots to check for violation of the regression assumptions about ε . (12.8)

12.3 Coefficient of Determination

- The i th residual (error) = $y_i - \hat{y}_i$

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

- **SSE** (sum of squares for error): the measure of the error in using the estimated regression equation to estimate the values of y in the sample.
- SSE is the quantity that is minimized by the least square method.

# of <u>Ads</u> (Xi)	# of cars sold (Yi)	$\hat{y}_i = 10 + 5x_i$	Residuals $y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	14	15	-1	1
3	24	25	-1	1
2	18	20	-2	4
1	17	15	2	4
3	27	25	2	4

$$SSE = \sum(\hat{y}_i - y_i)^2 = 1+1+4+4+4=14$$

SST (**total sum of squares**): measures the error involved in using \bar{y} to estimate y_i , (the total variation in y):

$$SST = \sum (y_i - \bar{y})^2$$

- **SST** can be divided into two parts:

$$SST = SSR + SSE$$

✓ **SSR** (**sum of squares for regression**): measures how much \hat{y} values on the estimated regression line deviate from \bar{y}

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

✓ **SSE** (**sum of squares for error**): measures the error in using \hat{y} to estimate y .

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

# of <u>Ads</u> (Xi)	# of cars sold (Yi)	$\hat{y}_i = 10 + 5x_i$
1	14	15
3	24	25
2	18	20
1	17	15
3	27	25

$$SSE = 14$$

$$\bar{y} = 20$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = 100$$

$$SST = \sum (y_i - \bar{y})^2 = ?$$

Any two of these sum of squares are known, the third can be computed easily!

In the Ads and Sales example, we developed the estimated regression equation $\hat{y}_i = 10 + 5x_i$ to approximate the relationship between the number of Ads (X) and the number of cars sold (Y) .

How well does the estimated regression equation fit the data?

SST, SSR and SSE can be used to provide a measure of the goodness of fit for the estimated regression equation-

-----**Coefficient of Determination.**

A perfect fit happens when SSR=SST. Why?

$$\hat{y}_i - y_i = 0 \Rightarrow SSE = 0$$

SSR = the explained portion of SST using the regression line

SSE = the unexplained portion of SST.

Coefficient of Determination

The coefficient of determination is defined as

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The ratio r^2 takes values between 0 and 1 and is used to evaluate *the goodness of fit* of the estimated regression equation.

- r^2 can be interpreted as the percentage of SST that can be explained by using the estimated regression equation.

For the Ads and Sales example: $r^2=?$

SSR=100, SST=114,

$$r^2 = 100/114 = 0.8772 = 87.72\%$$

- The regression relationship is very strong
- 87.72% of SST can be explained by the regression equation $\hat{y}_i = 10 + 5x_i$.
- 87.72% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

Correlation Coefficient

Recall:
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}}$$

$$r_{xy} = (\text{sign of } b_1) \sqrt{\text{Coefficient of Determination}}$$

$$r_{xy} = (\text{sign of } b_1) \sqrt{r^2}$$

where:

b_1 = the slope of the estimated regression
equation $\hat{y} = b_0 + b_1x$

Correlation Coefficient

For the previous example: Coefficient of Determination $r^2 = 0.8772$. **What is r_{xy} ?**

$$r_{xy} = (\text{sign of } b_1) \sqrt{r^2}$$

The sign of b_1 in the equation $\hat{y} = 10 + 5x$ is “+”.

$$r_{xy} = +\sqrt{.8772}$$

$$r_{xy} = \mathbf{+.9366}$$

+0.9366 indicates a strong positive linear relationship exists between x and y

Some Comments about r^2 :

1. Correlation Coefficient r_{xy} is restricted to a linear relationship between two variables x and y .
2. Coefficient of Determination r^2 can be used for *nonlinear relationship* and for relationship that have two or more independent variables. Thus it provides a wider range of applicability.
3. A high value of r^2 can arise even though the relationship between the two variables is *non-linear*. The fit of a linear model should never simply be judged from the r^2 value alone.
4. In the social science, values of r^2 as low as 0.25 are often considered useful. In the physical or life science, $r^2 > 0.6$. In business applications, r^2 values vary largely.

12.5 Testing for Significance

To see if the regression relationship is significant, we must conduct a hypothesis test to determine whether the value of β_1 is zero. (Why?)

Two tests are commonly used:

***t* Test**

and

***F* Test**

Both the *t* test and *F* test require an estimate of σ^2 , the variance of ε in the regression model.

12.5 Testing for Significance

- **An Estimate of σ**

The mean square error (MSE) provides a good estimate of σ^2 :

$$S_E^2 = \text{MSE} = \text{SSE}/(n - 2)$$

where:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

n - 2 is the degree of freedom for SSE.

Note: Every sum of square has a degree of freedom.

Statisticians have shown that SSE has n-2 degree of freedom. n is number of observations and 2 is the number of parameters in the model.

Testing for Significance

- An Estimate of σ
To estimate σ we take the square root of S_E^2 .

The resulting S_E is called the standard error of the estimate.

$$s_E = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

$$s_E^2 = MSE \rightarrow \sigma^2$$

$$s_E \rightarrow \sigma$$

- Previously, we have $\hat{y}_i = 10 + 5x_i$, $b_0=10$, $b_1=5$ are the least square estimates of β_0 and β_1 based on one sample.

- What would happen if we used a different sample for the same regression ?

b_0 and b_1 should be different for difference samples.

- b_0 and b_1 are sample statistics with their own sampling distribution.

Sampling distribution of b_1

When the basic assumptions of the simple linear regression model are satisfied, the following are true:

- Expected value

$$E(b_1) = \beta_1$$

- Standard Deviation

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- Distribution Form

Normal

Since σ is unknown,

$$s_E = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

$$s_E \rightarrow \sigma$$

$$s_{b_1} \triangleq \frac{s_E}{\sqrt{\sum (x_i - \bar{x})^2}} \rightarrow \sigma_{b_1}$$

Testing for Significance: t Test

$$E(y) = \beta_0 + \beta_1 x \quad \beta_1 = 0 \Leftrightarrow \text{No linear relation between X and Y}$$

Hypothesis: $H_0: \beta_1 = 0$

$$H_a: \beta_1 \neq 0$$

Test Statistic:

$$t^* = \frac{b_1}{s_{b_1}} \quad \text{where} \quad s_{b_1} = \frac{s_E}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Use t distribution with $df = n - 2$

Reject H_0 if $p\text{-value} \leq \alpha$

For the previous example:

$$H_0: \beta_1 = 0 \quad SSE = \sum (\hat{y}_i - y_i)^2 = 14$$

$$H_a: \beta_1 \neq 0 \quad s_E^2 = MSE = \frac{SSE}{n-2} = \frac{14}{5-2} \approx 4.67$$

$$\alpha = 0.05$$

$$s_E = \sqrt{4.67} \approx 2.16$$

$$\bar{x} = 2, \sum (x_i - \bar{x})^2 = 1 + 1 + 0 + 1 + 1 = 4$$

$$s_{b_1} = \frac{s_E}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{2.16}{\sqrt{4}} \approx 1.08$$

$$t^* = \frac{b_1}{s_{b_1}} = \frac{5}{1.08} = 4.62$$

$$df = n - 2 = 5 - 2 = 3$$

$$p\text{-value} = 2P(t > 4.62)$$

- For $df = 3$, $t = 4.541$ provides an area of 0.01 in the upper tail
- P-value $< 2 * 0.01 = 0.02 < 0.05$

Hence, We can reject H_0 and conclude that β_1 is not equal to 0.

This evidence is sufficient to conclude that a significant relationship exists between the number of TV Ads and the number of cars sold.

Confidence Interval for β_1

- The form of a confidence interval for β_1 is:

$$b_1 \pm t_{\alpha/2} S_{b_1}$$

b_1 is the
point
estimator

margin
of error

where $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with $n - 2$ degrees of freedom

Confidence Interval for β_1

$$t_{3,.025} = 3.182$$

$$95\%CI = b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.182 * 1.08 = (1.5634, 8.4366)$$

Note:

We can also use a 95% confidence interval for β_1 to test the hypotheses at 5% level of significance.

Rejection Rule:

$H_0 : \beta_1 = 0$ is rejected if 0 is not included in the confidence interval for β_1 .

For our example, the H_0 is rejected b/c the CI doesn't include 0.

Testing for Significance: F Test

Hypotheses $H_0: \beta_1 = 0$

$$H_a: \beta_1 \neq 0$$

Test Statistic

$$F = MSR/MSE$$

where $MSE = SSE / (n - 2)$

$$\begin{aligned} MSR &= SSR / (\text{Number of Independent Variables}) \\ &= SSR / 1 = SSR \end{aligned}$$

F follows an F distribution with $F_{1, n-2}$

1 degree of freedom in the numerator and
 $n - 2$ degrees of freedom in the denominator

Reject H_0 if $p\text{-value} \leq \alpha$

F Test for our example: Recall $SSR=100$, $MSE=14/3=4.67$

1. Determine the hypotheses. $H_0: \beta_1 = 0$

$$H_a: \beta_1 \neq 0$$

2. Specify the level of significance. $\alpha = .05$

3. Compute the test statistic. $F^* = MSR/MSE=100/4.67$
 $= 21.43$

4. Determine whether to reject H_0

F Table (P655)

$F_{1,3} = 17.44$ provides an area of .025 in the upper tail.

Thus, the p -value corresponding to $F = 21.43$ is less than .025. Hence, we reject H_0 .

Note: The T test and F test provide identical results for simple linear regression. . Recall: $t^*=4.62$, $F^*=(t^*)^2$

Some Cautions about the Interpretation of Significance Tests

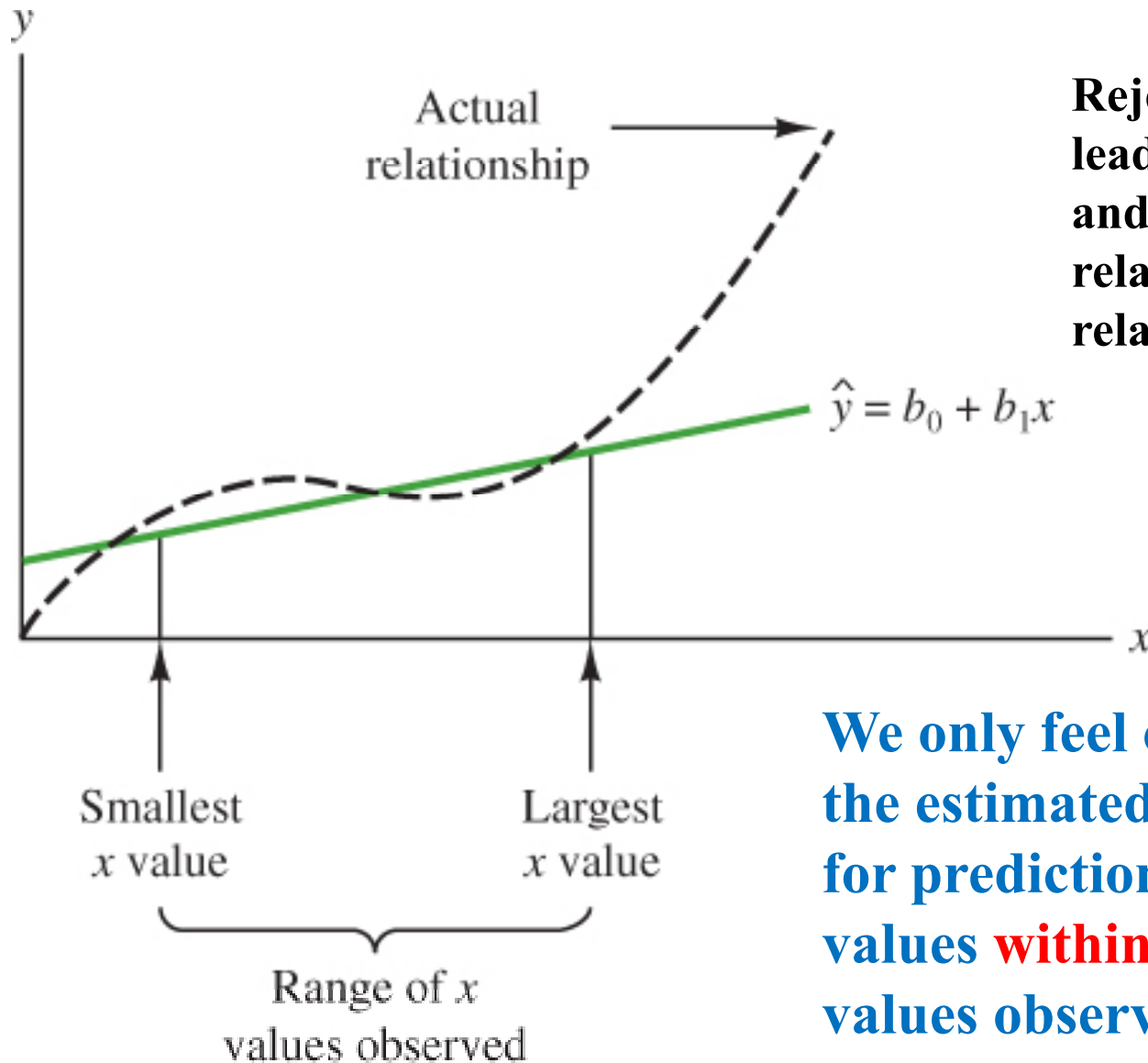
Rejecting $H_0: \beta_1 = 0$ and concluding that the relationship between x and y is significant.

However,

(1) It does not enable us to conclude that a cause-and-effect relationship is present between x and y . There may be some other unknown variables at work! (**time of year, state of economy, etc.**)

(2) It does not enable us to conclude that there is a linear relationship between x and y . We can state only that x and y are related and a linear relationship explains a significant portion of the variability in y over the range of values for x observed in the sample.

Example of a linear approximation of a nonlinear relationship



Rejection of $H_0: \beta_1=0$ leads to the conclusion **X** and **Y** are significantly related but the actual relationship is not linear.

We only feel confident in using the estimated regression equation for prediction corresponding to x values **within** the range of the x values observed in the sample.

Question:

Is it reasonable to use the regression line $\hat{y}_i = 10 + 5x_i$ to estimate the number of car sold when Reed runs 5 TV Ads?

No, b/c 5 is not within the range of the observed number of Ads, which is (1,2,3).

Steps in Regression Analysis

To perform simple regression analysis, we usually use a step-by-step approach as below:

1. Fit the model to data – estimate parameters β_0 and β_1 . (12.2)
2. Use the analysis of variance t test (or F test) and r^2 to determine how well the model fits the data. (12.3, 12.5)
3. Use estimated regression equation to estimate or predict (12.7)
4. Use diagnostic plots to check for violation of the regression assumptions about ε . (12.8)

12.7 Estimation and Prediction

If the significant relationship exists between x and y , and the r^2 shows the fit is good, the estimated regression line should be useful for **estimation** and **prediction**.

Point estimation:

Recall: $\hat{y}_i = 10 + 5x_i$

(1) If 3 TV ads are run prior to a sale, we expect the **mean number** of cars sold to be:

$$10 + 5(3) = 25 \text{ cars}$$

—————> **E(y) when x=3**

(2) In **one particular week** 3 TV ads are run, we expect the number of cars sold to be:

$$10 + 5(3) = 25 \text{ cars}$$

—————> **an individual value of y when x=3**

Interval Estimation

Point estimates do not provide any information about the precision associated with an estimate.

Confidence Intervals and Prediction intervals show the precision of the regression results.

Narrower intervals provide a higher degree of precision.

In developing Interval estimate, we define:

x_p : the particular or given value of x

$\hat{y}_p = b_0 + b_1 x_p$: the estimated value of y when $x = x_p$

$y_p = \beta_0 + \beta_1 x_p + \epsilon_p$ ——— An individual value of y
involves a random error!

$E(y_p) = \beta_0 + \beta_1 x_p$

The prediction of an individual value of y , y_p , is more difficult.

Using the Estimated Regression Equation for Estimation and Prediction

To estimate the mean value of y ($E(y_p)$) when $x = x_p$:

$$\hat{y}_p \pm t_{\alpha/2} S_E \sqrt{\left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

To predict an individual value of y (y_p) when $x = x_p$:

$$\hat{y}_p \pm t_{\alpha/2} S_E \sqrt{\left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

where $S_E = \sqrt{MSE}$, $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom

Note:

1. The confidence interval and prediction interval are the most precise (narrowest interval) when $x_p = \bar{x}$. The further x_p is from \bar{x} the larger $x_p - \bar{x}$ becomes. As a result, the interval is wider.

2. The prediction interval is always wider than confidence interval of the mean value of y .

If 3 TV ads are run prior to a sale, we expect the mean number of cars sold to be:

$$y = 10 + 5(3) = 25 \text{ cars}$$

The 95% confidence interval estimate of **the mean number** of cars sold when 3 TV ads are run is:

$$25 \pm 4.61 = 20.39 \text{ to } 29.61 \text{ cars}$$

The 95% prediction interval estimate of the number of cars sold **in one particular week** when 3 TV ads are run is:

$$25 \pm 8.28 = 16.72 \text{ to } 33.28 \text{ cars}$$

The prediction interval is wider than confidence interval of the mean value of y.

Steps in Regression Analysis

To perform simple regression analysis, we usually use a step-by-step approach as below:

1. Fit the model to data – estimate parameters β_0 and β_1 . (12.2)
2. Use the analysis of variance t test (or F test) and r^2 to determine how well the model fits the data. (12.3, 12.5)
3. Use estimated regression equation to estimate or predict (12.7)
4. Use diagnostic plots to check for violation of the regression assumptions. (12.8)

12.8 Residual Analysis

Validating Model Assumptions

The results of a regression analysis are only valid when the necessary assumptions have been satisfied.

Recall Assumptions:

1. The relationship between x and y is linear, given by $y = \beta_0 + \beta_1x + \varepsilon$.
2. The random error terms ε are independent and, for any value of x , have a normal distribution with mean 0 and variance σ^2 .

Residual Analysis: Validating Model Assumptions

If the assumptions about the error term ε appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.

Recall: Residual for Observation i

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

The residuals provide the best information about ε thus we can use them to check violations in the assumptions about random errors.

Much of the residual analysis is based on an examination of graphical plots.

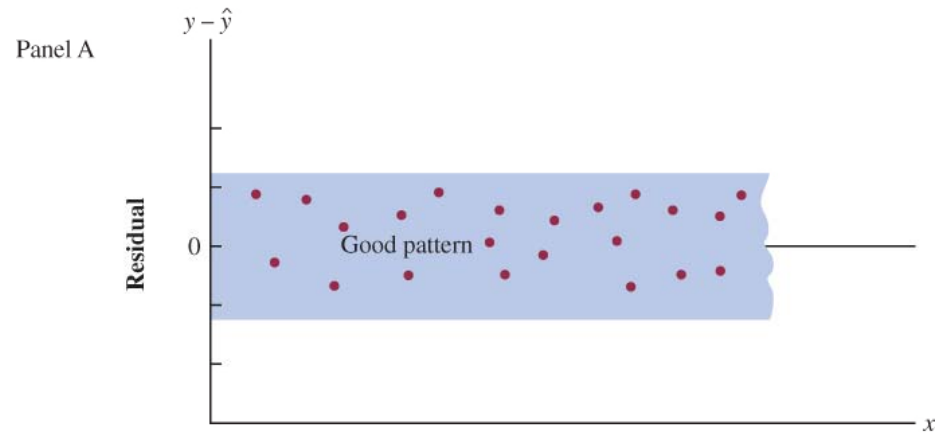
Residual Plot Against x

If the assumption that the variance of ε is the same for all values of x is valid, and the assumed linear regression model is an adequate representation of the relationship between the variables x and y , then

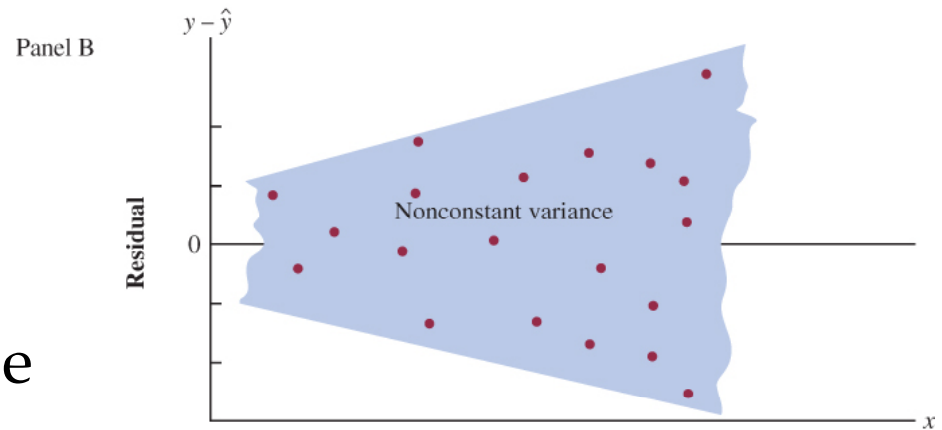
the residual plot should give an overall impression of a horizontal band of points.

Residual Pattern: Three general Pattern

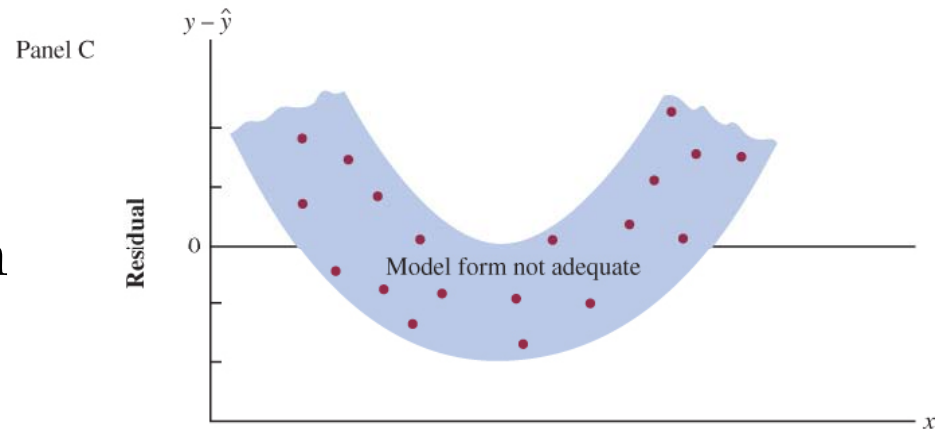
Good Pattern



Non-constant Variance



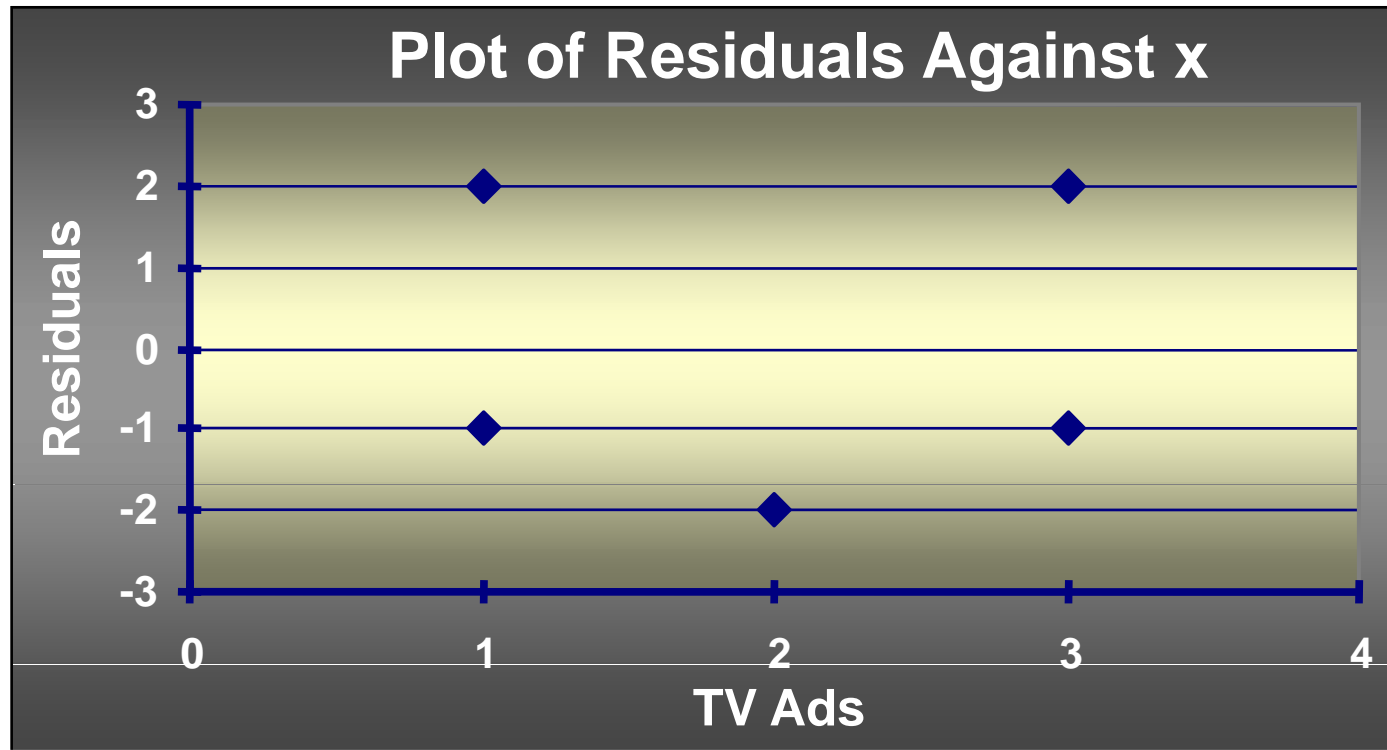
Non-linear Pattern



Example:

# of <u>Ads</u> (X_i)	# of cars sold (Y_i)	$\hat{y}_i = 10 + 5x_i$	Residuals $y_i - \hat{y}_i$
1	14	15	-1
3	24	25	-1
2	18	20	-2
1	17	15	2
3	27	25	2

Residual Plot Against x (*TV Ads*)



The residuals appear to approximate the horizontal pattern but here the sample size is too small.