Chapter 3
Simple Regression Analysis
(Part 1)

Terry Dielman

Applied Regression Analysis:
A Second Course in Business and Economic Statistics, fourth edition

3.1 Using Simple Regression to Describe a Relationship

- *Regression analysis* is a statistical technique used to describe relationships among variables.
- The simplest case is one where a *dependent variable* *y* may be related to an *independent* or *explanatory variable* *x*.
- The equation expressing this relationship is the line:

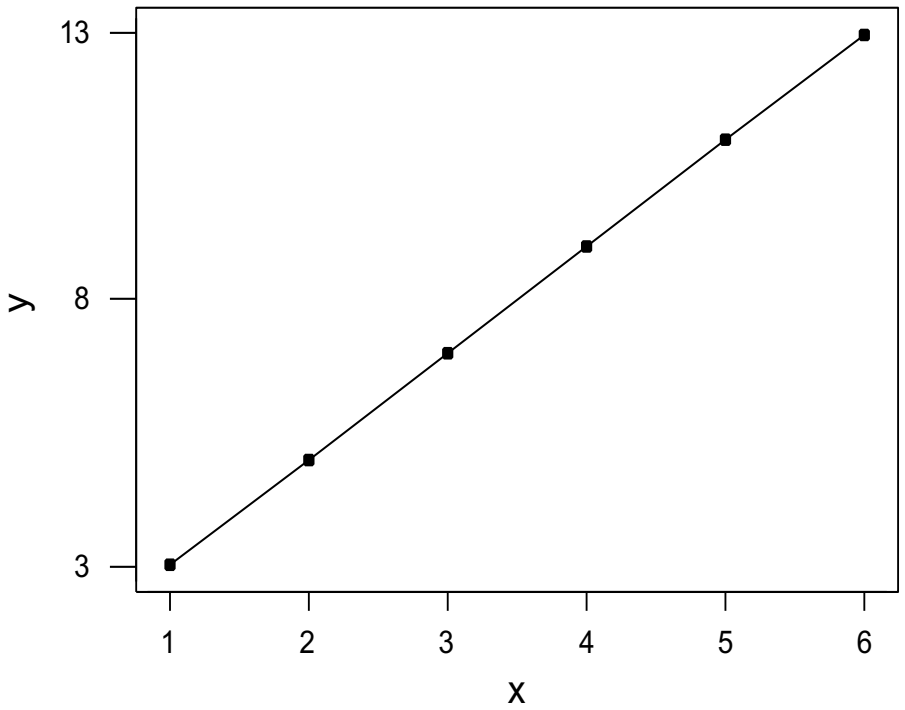
$$y = b_0 + b_1 x$$

Slope and Intercept

- For a given set of data, we need to calculate values for the slope b_1 and the intercept b_0.
- Figure 3.1 shows the graph of a set of six (x, y) pairs that have an exact relationship.
- Ordinary algebra is all you need to compute $y = 1 + 2x$

Figure 3.1 Graph of An Exact Relationship

x	*y*
1	3
2	5
3	7
4	9
5	11
6	13

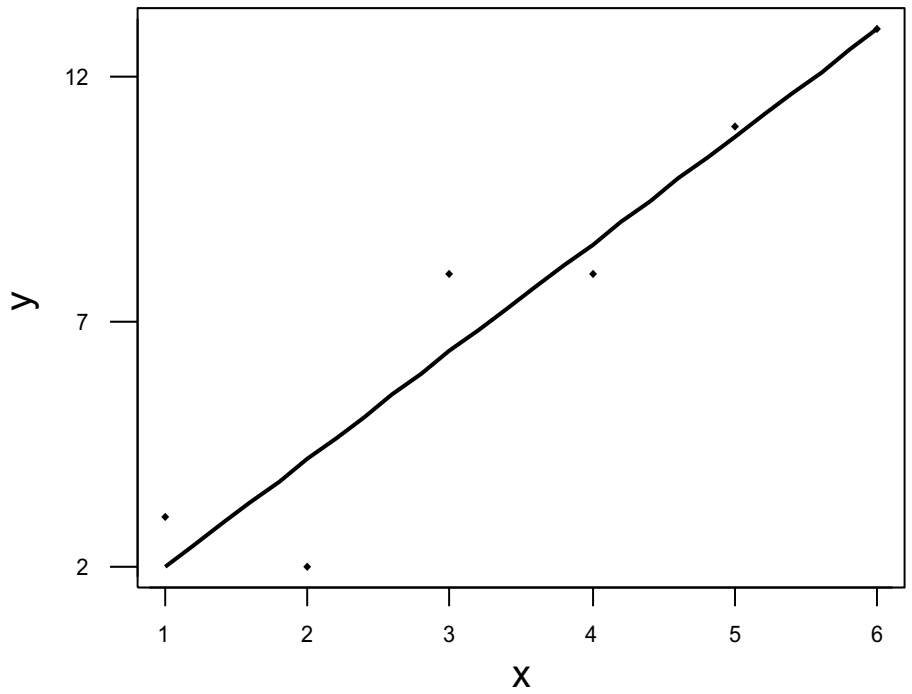


Error in the Relationship

- In real life, we usually do not have exact relationships.
- Figure 3.2 shows a situation where the *y* and *x* have a strong tendency to increase together but it is not perfect.
- You can use a ruler to put a line in approximately the "right place" and use algebra again.
- A good guess might be $\hat{y} = 1 + 2.5x$

Figure 3.2 Graph of a Relationship That is NOT Exact

x	*y*
1	3
2	2
3	8
4	8
5	11
6	13



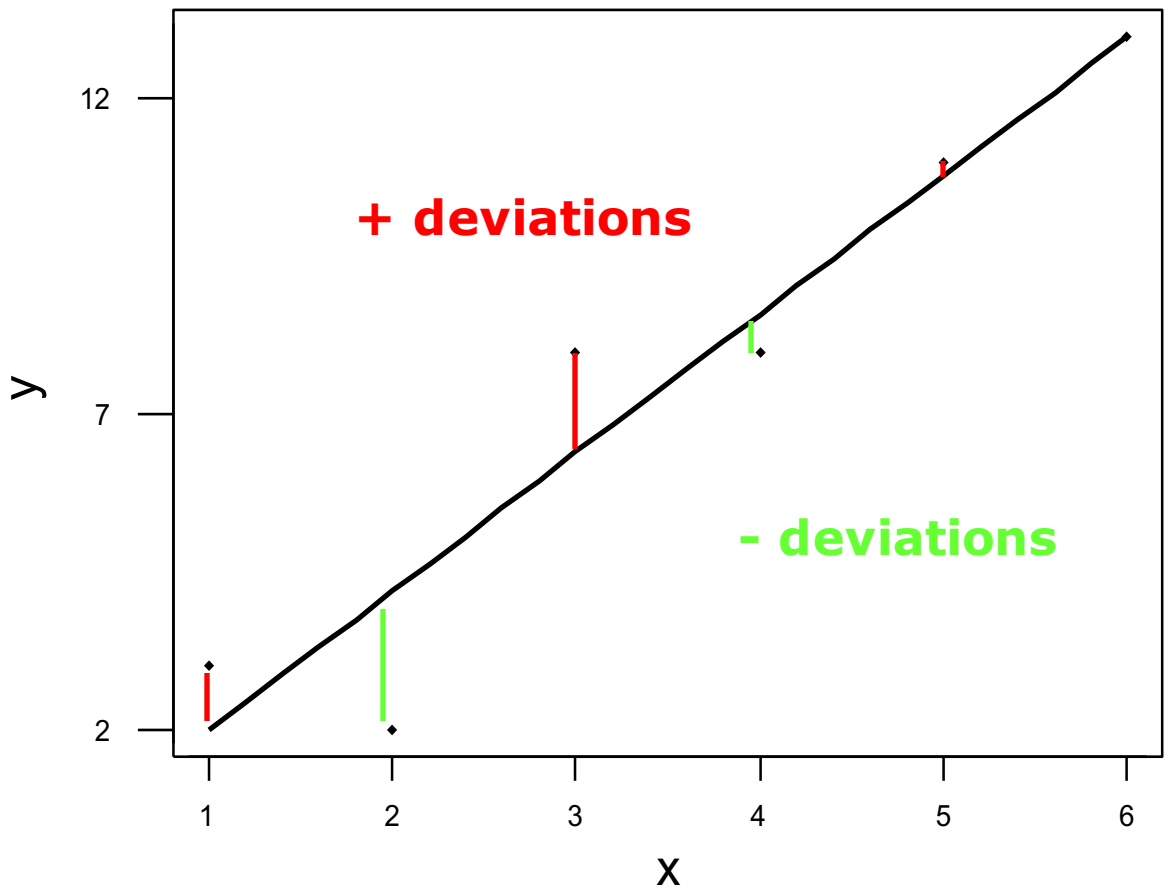
Everybody Is Different

- The drawback to this technique is that everybody will have their own opinion about where the line goes.
- There would be ever greater differences if there were more data with a wider scatter.
- We need a precise mathematical technique to use for this task.

Residuals

- Figure 3.3 shows the previous graph where the "fit error" of each point is indicated.
- These *residuals* are positive if the point is above the line and negative if the line is above the point.
- We want a technique that will make the + and – even out.

Figure 3.3 Deviations From the Line



Computation Ideas (1)

We can search for a line that minimizes the sum of the residuals:

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

While this is a good idea, it can be shown that any line passing through the point $(\bar{x}, \bar{y})$ will have this sum = 0.

Computation Ideas (2)

We can work with absolute values and search for a line that minimizes:

$$\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Such a procedure—called LAV or *least absolute value* regression—does exist but usually is found only in specialized software.

Computation Ideas (3)

By far the most popular approach is to square the residuals and minimize:

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

This procedure is called *least squares* and is widely available in software. It uses calculus to solve for the b_0 and b_1 terms and gives a unique solution.

Least Squares Estimators

- There are several formula for the b_1 term. If doing it by hand, we might want to use:

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

- The intercept is $b_0 = \bar{y} - b_1 \bar{x}$

Figure 3.5 Computations Required for b_1 and b_0

x_i	y_i	x_i^2	$x_i y_i$
1	3	1	3
2	2	4	4
3	8	9	24
4	8	16	32
5	11	25	55
6	13	36	78
21	**45**	**91**	**196**

Totals

Calculations

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

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

The Unique Minimum

- The line we obtained was:

$$\hat{y} = -0.2 + 2.2x$$

- The sum of squared errors (SSE) is:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 8.80$$

- No other linear equation will yield a smaller SSE. For the line *1 + 2.5x* we guessed earlier, the SSE is 10.75

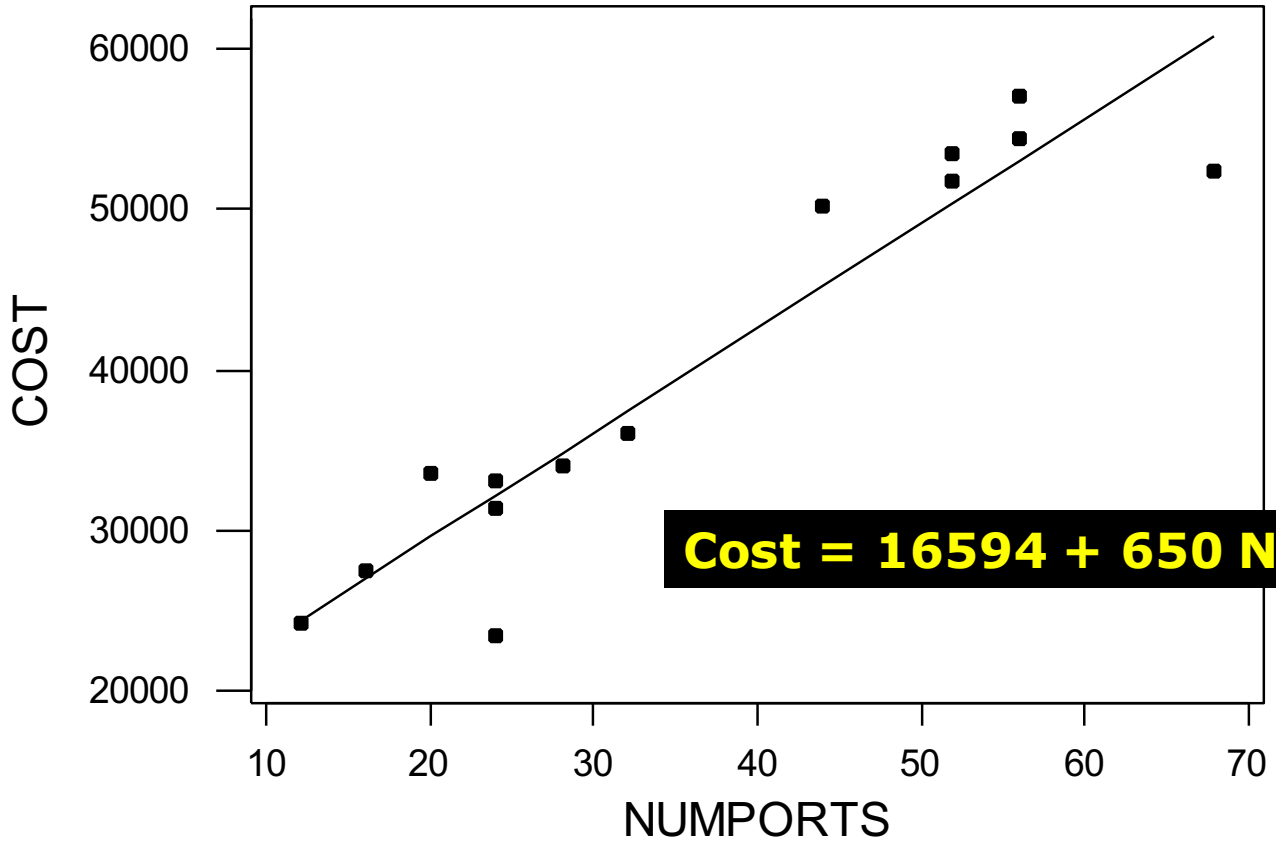
3.2 Examples of Regression as a Descriptive Technique

Example 3.2 Pricing Communications Nodes

A Ft. Worth manufacturing company was concerned about the cost of adding nodes to a communications network. They obtained data on 14 existing nodes.

They did a regression of cost (the *y*) on number of ports (*x*).

Pricing Communications Nodes

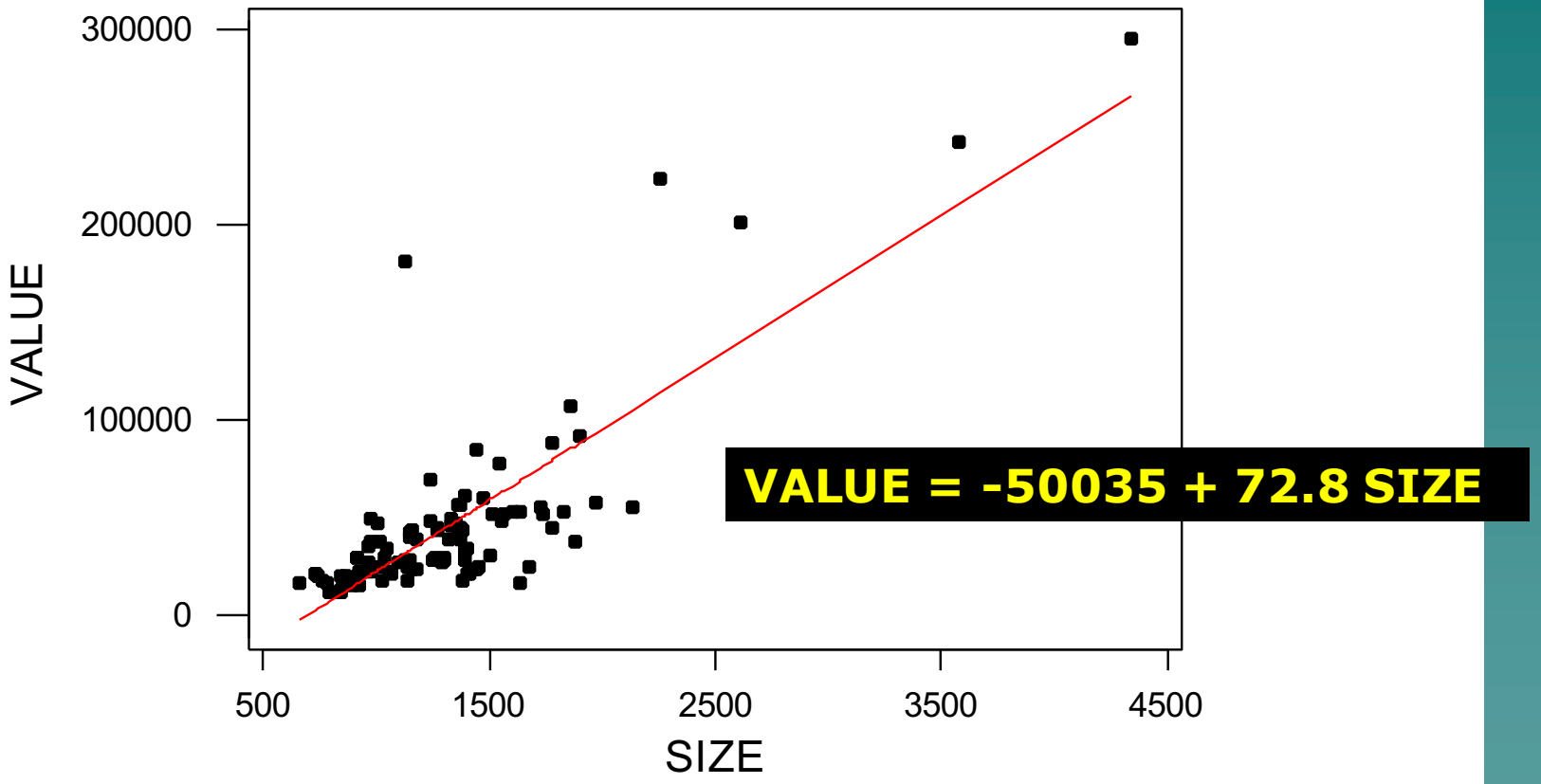


Example 3.3 Estimating Residential Real Estate Values

The Tarrant County Appraisal District uses data such as house size, location and depreciation to help appraise property.

Regression can be used to establish a weight for each factor. Here we look at how price depends on size for a set of 100 homes. The data are from 1990.

Tarrant County Real Estate



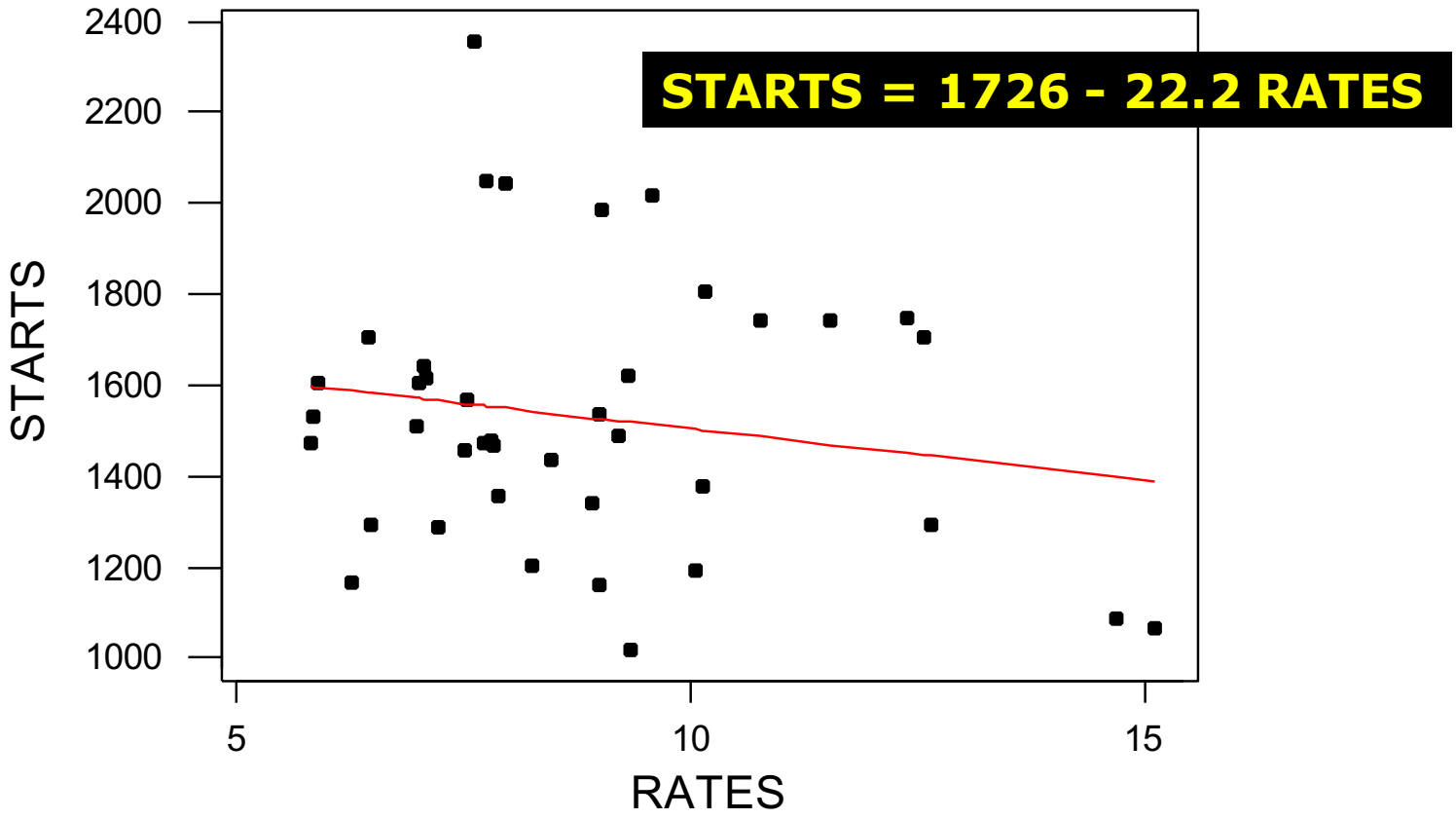
Example 3.4 Forecasting Housing Starts

Forecasts of various economic measures is important to the government and various industries.

Here we analyze the relationship between US housing starts and mortgage rates. The rate used is the US average for new home purchases.

Annual data from 1963 to 2002 is used.

US Housing Starts



3.3 Inferences From a Simple Regression Analysis

- So far regression has been used as a way to describe the relationship between the two variables.
- Here we will use our sample data to make inferences about what is going on in the underlying population.
- To do that, we first need some assumptions about how things are.

3.3.1 Assumptions Concerning the Population Regression Line

- Lets use the communications nodes example to illustrate. Costs ranged from roughly $23000 to $57000 and number of ports from 12 to 68.
- Three times we had projects with 24 ports, but the three costs were all different. The same thing occurred at repeated observations at 52 and 56 ports.
- This illustrates how we view things: at each value of *x* there is a *distribution* of potential *y* values that can occur.

The Conditional Mean

- Our first assumption is that the means of these distributions all lie on a straight line:

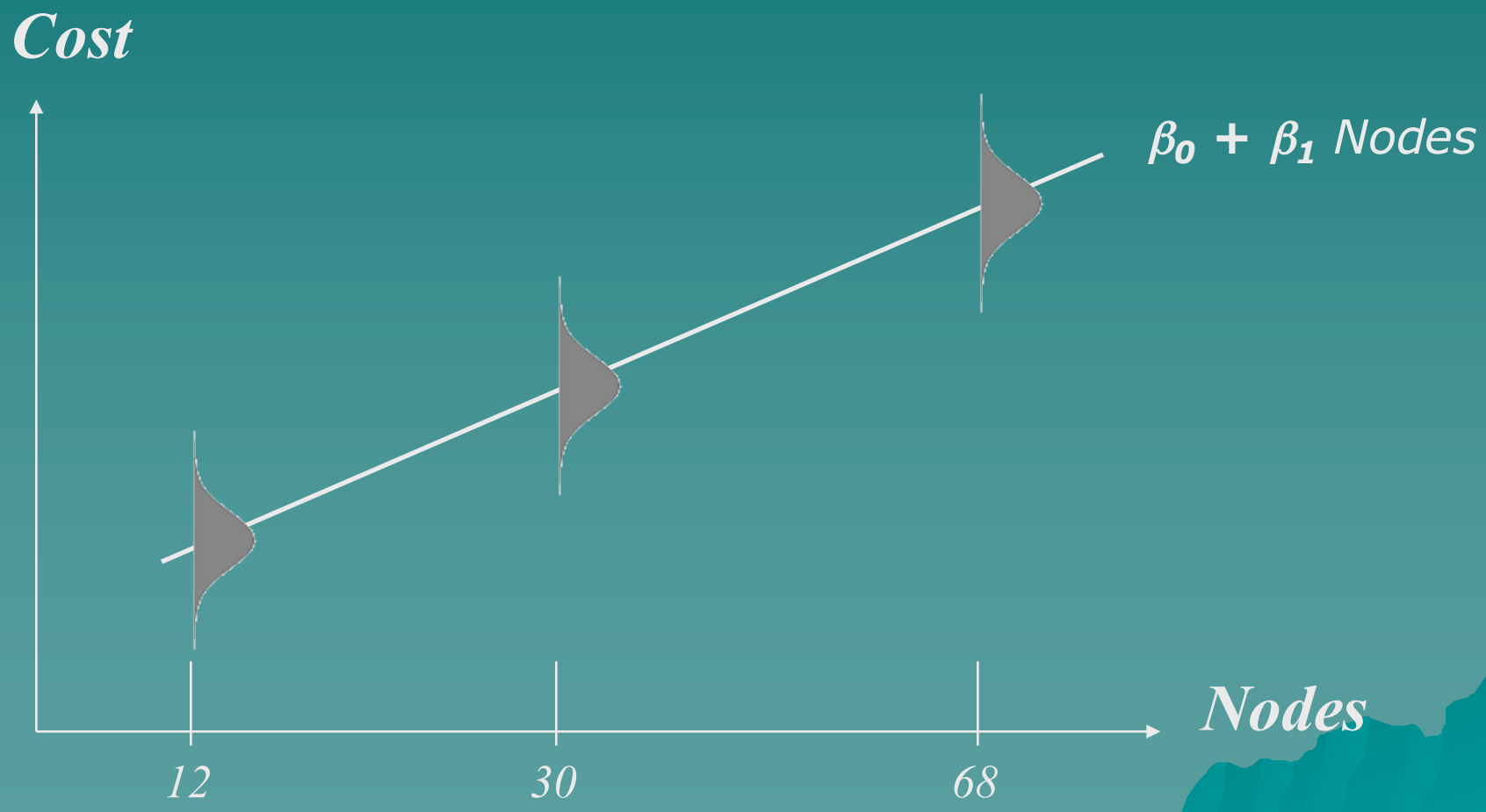
$$\mu_{y|x} = \beta_0 + \beta_1 x$$

- For example, at projects with 30 ports, we have:

$$\mu_{y|x=30} = \beta_0 + 30\beta_1$$

- The actual cost of projects with 30 ports are going to be distributed about the mean. This also happens at other sizes of projects, so you might see something like the next slide.

Figure 3.12 Distribution of Costs around the Regression Line



The Disturbance Terms

- Because of the variation around the regression line, it is convenient to view the individual costs as:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- The e_i are called the *disturbances* and represent how y_i differs from its conditional mean. If y_i is above the mean, its disturbance has a + value.

Assumptions

1. We expect the average disturbance e_i to be zero so the regression line passes through the conditional mean of y.
2. The e_i have constant variance σ_e^2.
3. The e_i are normally distributed.
4. The e_i are independent.

Notation for the true line (*population line*):

$$E(Y|X) = \beta_0 + \beta_1 X$$

expected value of Y given X

true intercept

true slope

Think of E(Y | X) as the average price of houses with size X.

Some houses (with size X) will have price bigger than the expected value, some smaller. The true line tells us what to expect on average.

true line	fitted line
$E(Y\|X) = \beta_0 + \beta_1 X$	$\hat{Y} = b_0 + b_1 X$

b_0 is an estimate of β_0

b_1 is an estimate of β_1

For a given X,

$\hat{Y}$ is an estimate of $E(Y|X)$

Even if you knew the true line (i.e., β_0 and β_1 are known), there would still be uncertainty about your prediction of Y given X.

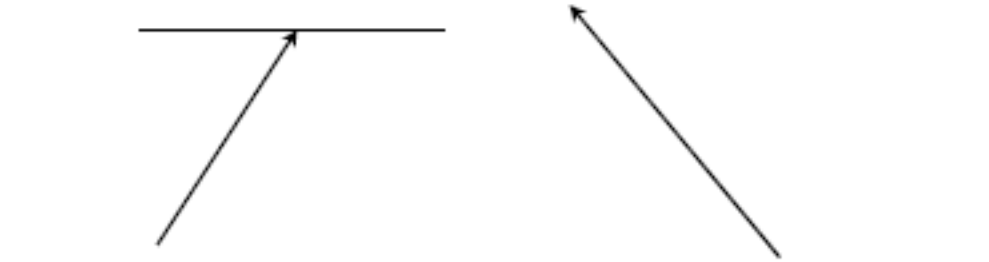
The true line only gives the expected value of Y given X: $E(Y|X) = \beta_0 + \beta_1 X$

Re-write as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where ε is the difference between Y and E(Y|X).

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



$E(\varepsilon \mid X) = 0$

Positive ε ==> Y larger than $E(Y \mid X)$

Negative ε ==> Y smaller than $E(Y \mid X)$

We assume that ε has a normal distribution:

$$\varepsilon \sim N(0, \sigma^2)$$

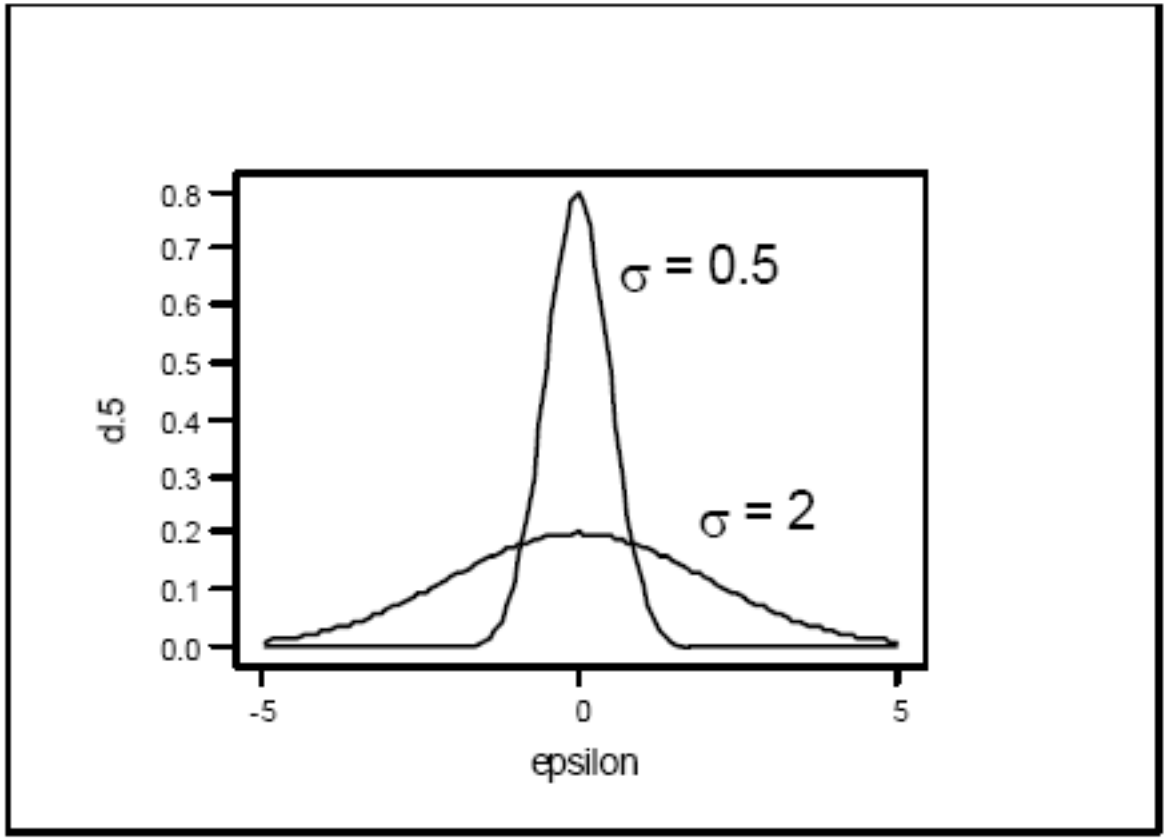
mean of ε is zero (sometimes Y above line, sometimes Y below line)

variance of ε is σ^2

If σ^2 is small, ε tends to be small (close to zero).
If σ^2 is large, ε tends to be large (far from zero).

Quick Review of the Normal Distribution

Graphs of two normal distributions, both having mean zero:



Recall:

$$X \sim N(\mu, \sigma^2)$$

implies

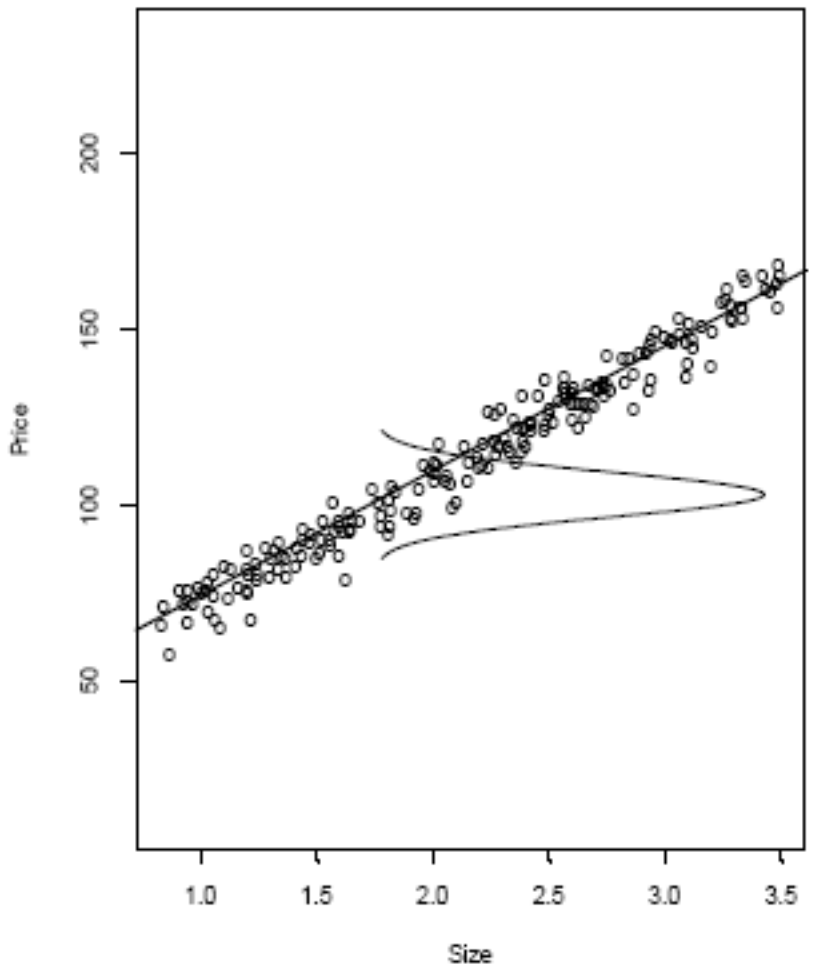
$$E(X) = \mu \qquad Var(X) = \sigma^2$$

$$\Pr(\mu - \sigma \le X \le \mu + \sigma) = 0.68$$

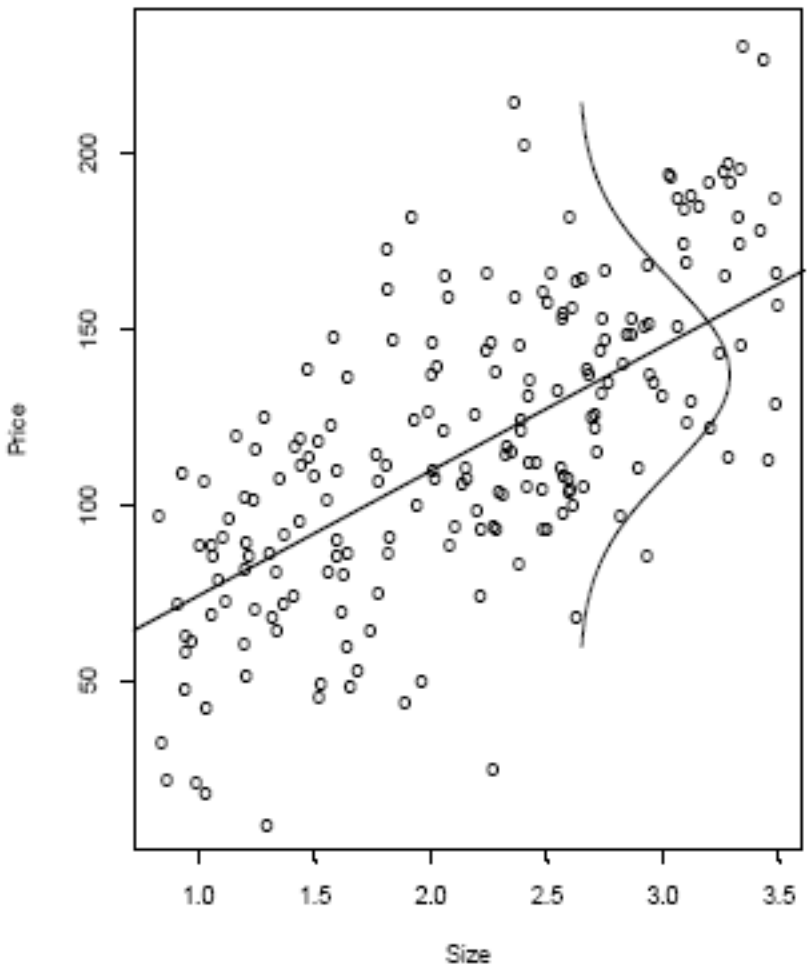
$$\Pr(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.9544$$

$$\Pr(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.997$$

ε tends to be small:



ε tends to be large:



Conditional distribution of Y given X:

$$\mu_{Y|X} = E(Y|X) = E(\beta_0 + \beta_1 X + \varepsilon|X) = \beta_0 + \beta_1 X$$

$$\sigma^2_{Y|X} = Var(Y|X) = Var(\beta_0 + \beta_1 X + \varepsilon|X) = \sigma^2$$

Under linearity and normality assumptions, the conditional distribution of Y given X is normally distributed:

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

3.3.2 Inferences About β_0 and β_1

- We use our sample data to estimate β_0 by b_0 and β_1 by b_1. If we had a different sample, we would not be surprised to get different estimates.
- Understanding how much they would vary from sample to sample is an important part of the inference process.
- We use the assumptions, together with our data, to construct the sampling distributions for b_0 and b_1.

The Sampling Distributions

- The estimators have many good statistical properties. They are unbiased, consistent and minimum variance.
- They have normal distributions with standard errors that are functions of the x values and σ_e^2.
- Full details are in Section 3.3.2

Estimate of σ_e^2

- This is an unknown quantity that needs to be estimated from data.
- We estimate it by the formula:

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

- The term *MSE* stands for mean squared error and is more or less the average squared residual.

Standard Error of the Regression

- The divisor *n-2* used in the previous calculation follows our general rule that degrees of freedom are sample size – the number of estimates we make (b_0 and b_1) before estimating the variance.
- The square root of *MSE* is S_e which we call the *standard error of the regression*.
- S_e can be roughly interpreted as the "typical" amount we miss in estimating each *y* value.

Inference About β_1

- Interval estimates and hypothesis tests are constructed using the sampling distribution of b_1.
- The standard error of b_1 is:

$$S_{b_1} = S_e \sqrt{\frac{1}{(n-1)S_x^2}}$$

- Computer programs routinely compute this and report its value.

Interval Estimate

- The distribution we use is a *t* with n-2 degrees of freedom.

- The interval is:

$$b_1 \pm t_{n-2} s_{b_1}$$

- The value of *t*, of course, depends on the selected confidence level.

Tests About β_1

The most common test is that a change in the *x* variable does not induce a change in *y*, which can be stated:

$$H_0: \ \beta_1 = 0$$
$$H_a: \ \beta_1 \neq 0$$

If H_1 is true it implies the population regression equation is a flat line; that is, regardless of the value of *x*, *y* has the same distribution.

Test Statistic

The test would be performed by using the standardized test statistic:

$$t = \frac{b_1 - 0}{S_{b_1}}$$

Most computer programs compute this, and its associated p-value. and include them on the output.

The p-value is for the two-sided version of the test.

Inference About β_0

- We can also compute confidence intervals and perform hypothesis tests about the intercept in the population equation.
- Details about the tests and intervals are in Section 3.3.2, but in most problems we are not interested in this.
- The intercept is the value of *y* at *x=0* and in many problems this is not relevant; for example, we never see houses with zero square feet of floor space.
- Sometimes it is relevant, anyway. If we are estimating costs, we could interpret the intercept as the fixed cost. Even though we never see communication nodes with zero ports, there is likely to be a fixed cost associated with setting up each project.

Example 3.6 Pricing Communications Nodes (continued)

Inference questions:

1. What is the equation relating NUMPORTS to COST?
2. Is the relationship significant?
3. What is an interval estimate of β_1?
4. Is the relationship positive?
5. Can we claim each port costs at least $1000?
6. What is our estimate of fixed cost?
7. Is the intercept 0?

Minitab Regression Output

Regression Analysis: COST versus NUMPORTS

The regression equation is
COST = 16594 + 650 NUMPORTS

Predictor	Coef	SE Coef	T	P
Constant	16594	2687	6.18	0.000
NUMPORTS	650.17	66.91	9.72	0.000

S = 4307 R-Sq = 88.7% R-Sq(adj) = 87.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1751268376	1751268376	94.41	0.000
Residual Error	12	222594146	18549512		
Total	13	1973862521			

SAS Regression Output

The REG Procedure
Model: MODEL1
Dependent Variable: COST

Number of Observations Read	14
Number of Observations Used	14

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1751268376	1751268376	94.41	<.0001
Error	12	222594146	18549512		
Corrected Total	13	1973862522			

Root MSE	4306.91446	R-Square	0.8872
Dependent Mean	40186	Adj R-Sq	0.8778
Coeff Var	10.71758		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	16594	2687.05000	6.18	<.0001
NUMPORTS	1	650.16917	66.91389	9.72	<.0001

Is the relationship significant?

H_0: $\beta_1 = 0$ *(Cost does not change when number of ports increase)*

H_a: $\beta_1 \neq 0$ *(Cost does change)*

We will use a 5% level of significance and the *t* distribution with $(n-2) = 12$ degrees of freedom.

Decision rule: Reject H_0 if $t > 2.179$
or if $t < -2.179$

from Minitab output $t = 9.72$ (p-value =.000)

We conclude that there is a significant relationship between project size and cost.

What is an interval estimate of β_1?

Interval is: $b_1 \pm t_{n-2} s_{b_1}$

For a 95% interval use $t = 2.179$

$650.17 \pm 2.179(66.91) = 650.17 \pm 145.80$

We are 95% sure that the average cost for each additional node is between \$504 and \$796.

Can we claim a positive relationship?

H_0: $\beta_1 = 0$ *(Cost does not change when size increases)*
H_a: $\beta_1 > 0$ *(Cost increases when size increases)*

We will use a 5% level of significance and the *t* distribution with *(n-2)* = 12 degrees of freedom.

Decision rule: Reject H_0 if $t > 1.782$

From Minitab output $t = 9.72$ (p-value is half of the listed value of .000, which is still .000)

We conclude that the project cost does increase with project size.

Is the cost per port at least $1000?

H_0: $\beta_1 \geq 1000$ *(Cost per port at least $1000)*
H_a: $\beta_1 < 1000$ *(Cost is less than $1000)*

Again we will use a 5% level of significance and 12 degrees of freedom.

Decision rule: Reject H_0 if $t < -1.782$

$$\text{Here use} \quad t = \frac{b_1 - 1000}{S_{b_1}} = \frac{650.17 - 1000}{66.91} = -5.23$$

We conclude that the cost per node is (much) less than $1000.

Is the cost per port at least $1000?

H_0: $\beta_1 <= 1000$ *(Cost per port at least $1000)*
H_a: $\beta_1 > 1000$ *(Cost is less than $1000)*

Again we will use a 5% level of significance and 12 degrees of freedom.

Decision rule: Reject H_0 if $t > 1.782$

$$\text{Here use} \quad t = \frac{b_1 - 1000}{S_{b_1}} = \frac{650.17 - 1000}{66.91} = -5.23$$

We conclude that the cost per node is (much) less than $1000.

What is our estimate of fixed cost?

We can interpret the intercept of the equation as fixed cost, and the slope as variable cost. For the intercept, an interval is:

$$b_0 \pm t_{n-2} s_{b_0}$$

16594 ± 2.179(2687) = 16954 ± 5855

We are 95% sure the fixed cost is between $11,099 and $22,809.

Is the intercept 0?

H_0: $\beta_0 = 0$ *(Fixed cost is 0)*
H_a: $\beta_0 \neq 0$ *(Fixed cost is not 0)*

Again, use a 5% level of significance and 12 d.f.

Decision rule: Reject H_0 if $t > 2.179$
or if $t < -2.179$

from Minitab output $t = 6.18$ (p-value =.000)

We conclude that the fixed cost is not zero.