ENGG 319

Probability & Statistics for Engineers

Section #11

Simple Linear Regression & Correlation

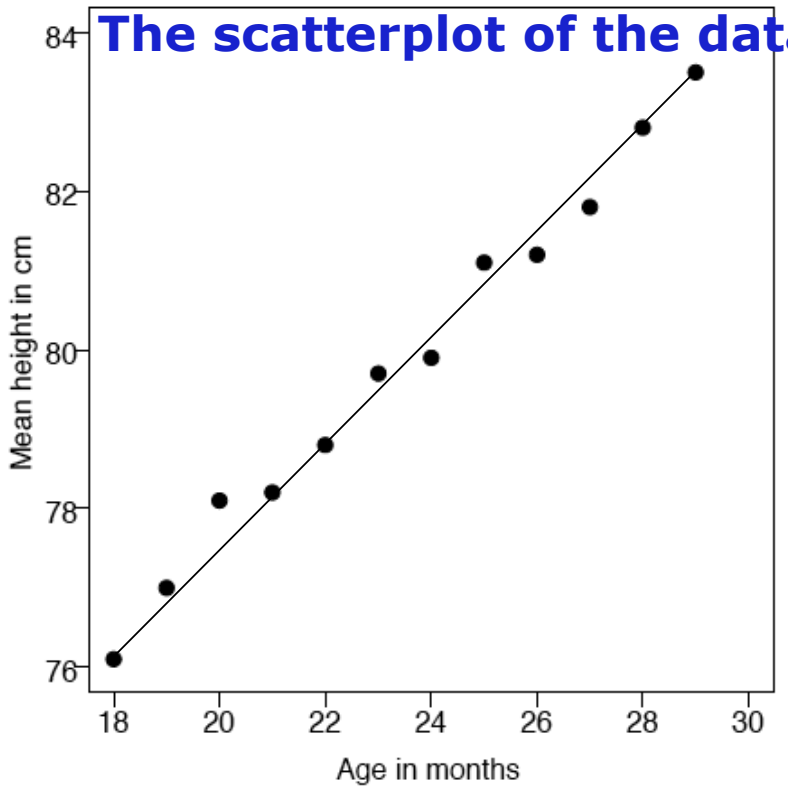
Dr. Sameh Nassar

Introduction

- **Regression models** are statistical models which describe the variation in one (or more) variable(s) when one or more other variable(s) vary. Inference based on such models is known as **regression analysis**.
- We use regression models for studying how changes in one (or more) variable(s) will change the value of another variable. Generalizing slightly, we can talk about a variable 'explaining' some of the variation in another variable.
  - Variables which are used to explain other variables are called **explanatory variables**, sometimes called an **independent variable**, or a **predictor**, or **regressor**.
  - The variable which is 'explained' is called the **response variable**, often called a **dependent variable.**

Example a: Age and Height of Children

Suppose you are interested in the general, overall, growth pattern of young children; one idea would be to follow a number of children over time and measure their heights at different ages. Such data would provide an indication of the overall growth pattern.



Each point in the **scatterplot** represents the average height of the children at the appropriate age.

Not surprisingly, the scatterplot shows that height and age are closely related.

A suitable model for this relationship might be a straight line describing the overall growth pattern, and an error term allowing for random variation away from the line.

Example b (1/6)
Measuring Mobility of Elderly People

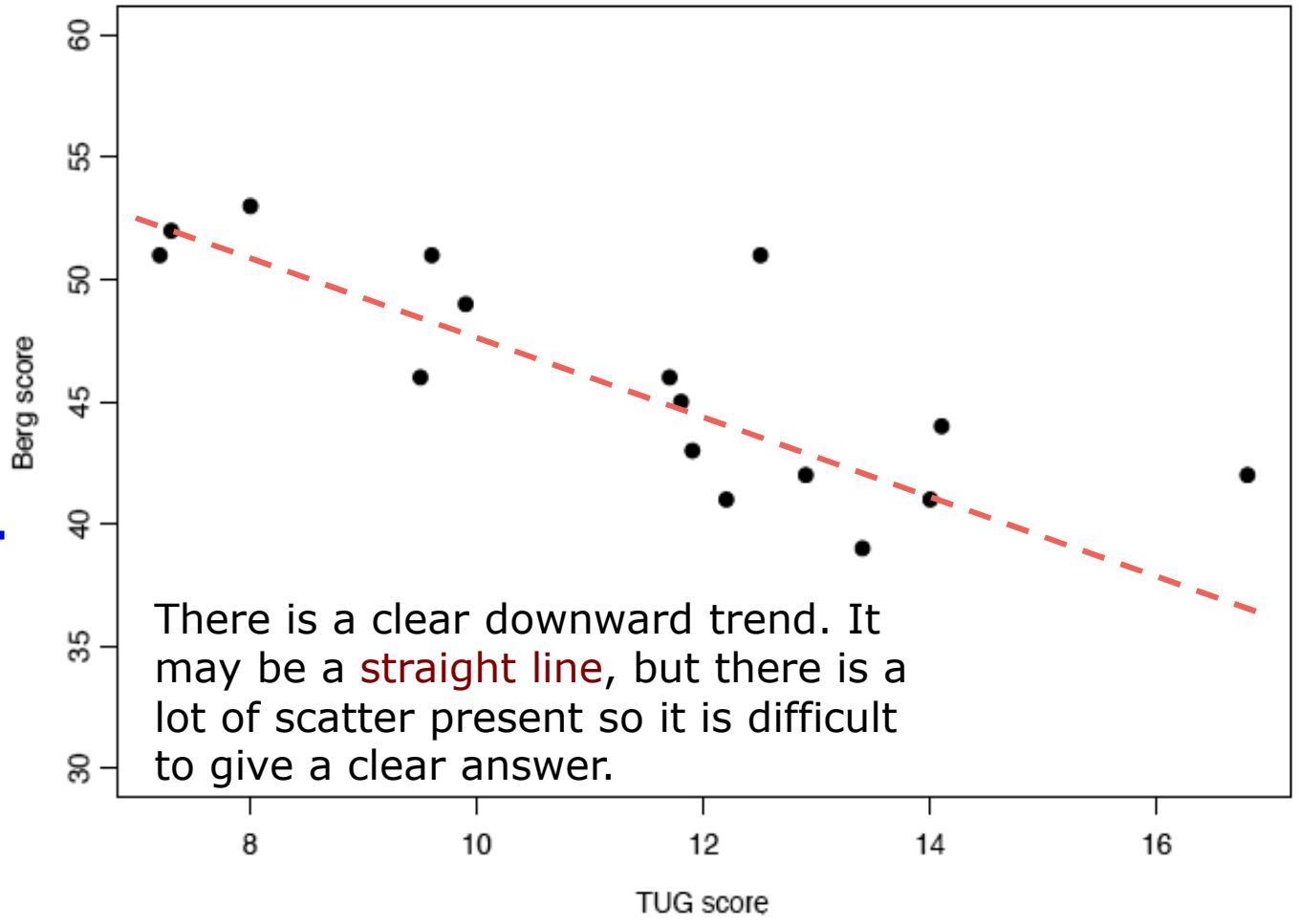
This example concerns two methods for measuring the mobility of elderly people. The two methods are the so-called **Berg score** and **Timed Up and Go (TUG)** score.

The **Berg score** is a measure based on how well the person performs in a number of different tasks. A low score corresponds to low mobility.

The **TUG score** is simply the time it takes a person to get up from a chair, walk three metres and return to the chair. Measuring the **Berg score** is much more demanding and time-consuming than measuring the **TUG score**.

The interest is whether the quick method (the TUG score) can be used to give a good prediction of the more thorough method (the Berg score).

Example b (2/6) : Reading the Scatterplot



the dependent variable

There is a clear downward trend. It may be a straight line, but there is a lot of scatter present so it is difficult to give a clear answer.

the independent variable

Which of the two variables is the dependent variable and which is the independent variable?

What can you say from the scatterplot about the nature of the relationship between the variables?

Independent & Dependent Variables

- In regression, **response variables** are always regarded as **random variables**, whereas **independent variables** are usually regarded as **non-random**.
  - As a consequence, all the scatter away from the main trend in a scatterplot is ascribed to the response variable, only.
  - This assumption makes good sense in the cases where data were collected from a study specifically designed to examine how the response variable depends on an independent variable.

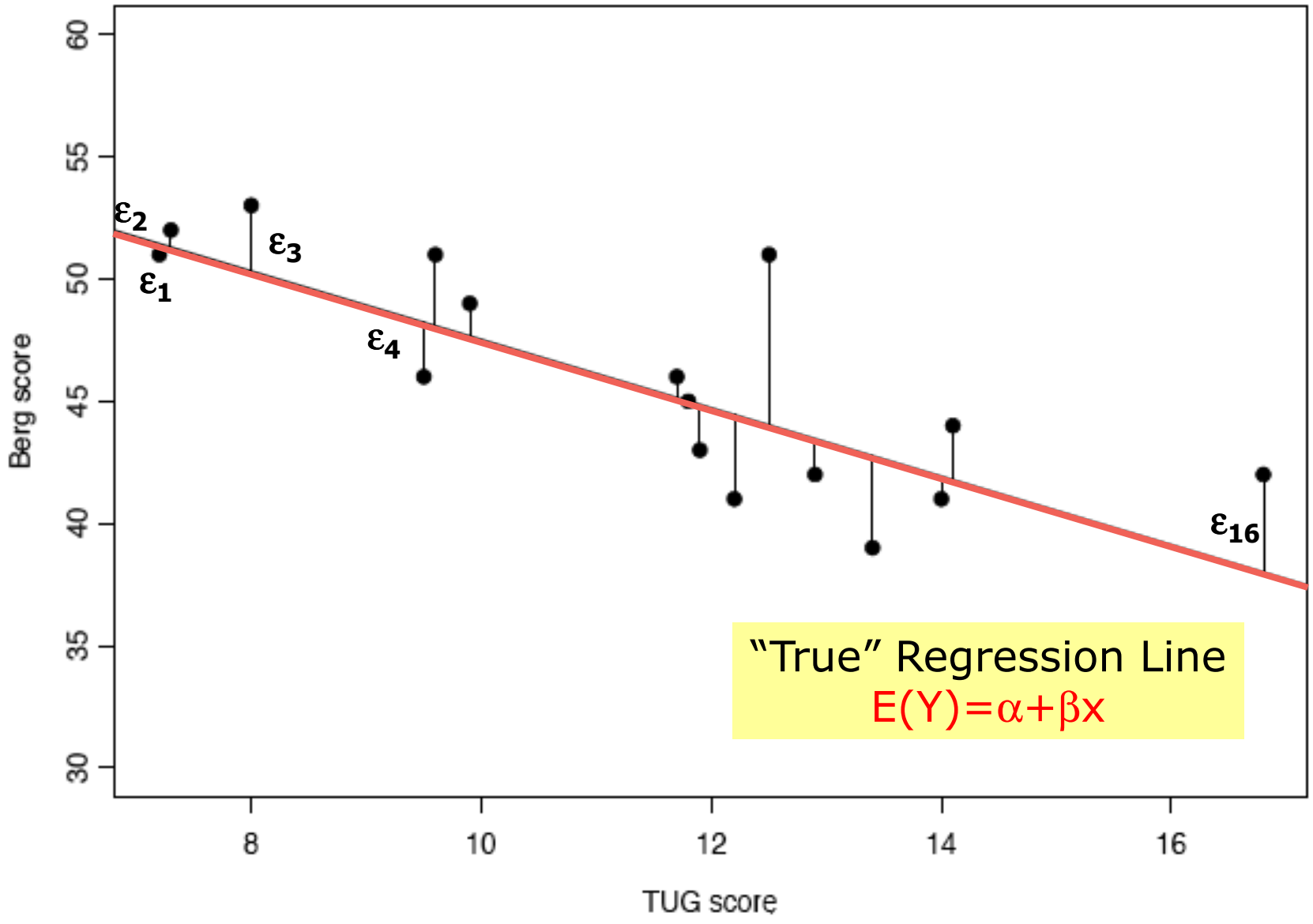
Definitions

- Since a response variable is random, we shall denote it by an upper-case letter, e.g. Y_i,
- While a non-random independent variable will be by a lower-case letter, e.g. x_i.
- The **general** regression model is given by:

$$Y = \alpha + \beta x + \varepsilon$$

where α and β are unknown intercept and slope parameters respectively, and ε is a random variable that is assumed to be distributed with **$E(\varepsilon)=0$** and **$Var(\varepsilon)=\sigma^2$**, often called the error variance.

Example b (3/6): Random Errors



Notes of Caution

- The first note concerns the validity of the model.
- A regression model is determined on the basis of the data we have observed
  - Any functional relationship between the response variable and the explanatory variable displayed in the scatterplot relates to our particular set of data.
  - For data outside the range of the observations, the relationship may, in some cases, be very different.

In general, one should not use a model outside the observed range

Notes of Caution

- The second note relates to the interpretation of conclusions drawn from a regression model. It is sometimes tempting to interpret a relationship in a regression model as if the explanatory variable is the reason for changes in the response variable.

- **Do changes in the explanatory variables *cause* changes in the response variable?**

- A regression model does not say anything about causation, it simply states that if the value of the explanatory is changed, the value of the response variable also changes.

Example b (4/6): Notes of Caution

- We found that a person who has a high **TUG score** will usually have a low **Berg score**. But that does not mean that a high **TUG score** *causes* a low **Berg score**.
- Presumably, the relationship is due to the fact that a person with poor mobility will score high on the TUG score, and low on the Berg score.
- Hence, it is a third variable 'mobility' that causes changes in both the TUG and the Berg scores.

Simple Linear Regression

- Suppose that we have a response variable **Y** and an explanatory variable **x**, then the **simple linear regression model** for **Y** on **x** is given by:

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad \forall i \in [1, n]$$

- The name simple linear regression model refers to the fact that the mean value of the response:

$$\mu_{Y|x} = E[Y_i] = \alpha + \beta x$$

is a linear function of the regression parameters α and β. It is called the **"true" regression line**.

- α and β are called the **regression parameters**.

Simple Linear Regression

- The random error ε_i is the term which accounts for the variation of the i^{th} response variable Y_i away from the linear predictor $\alpha+\beta x_i$ at the point x_i. That is:

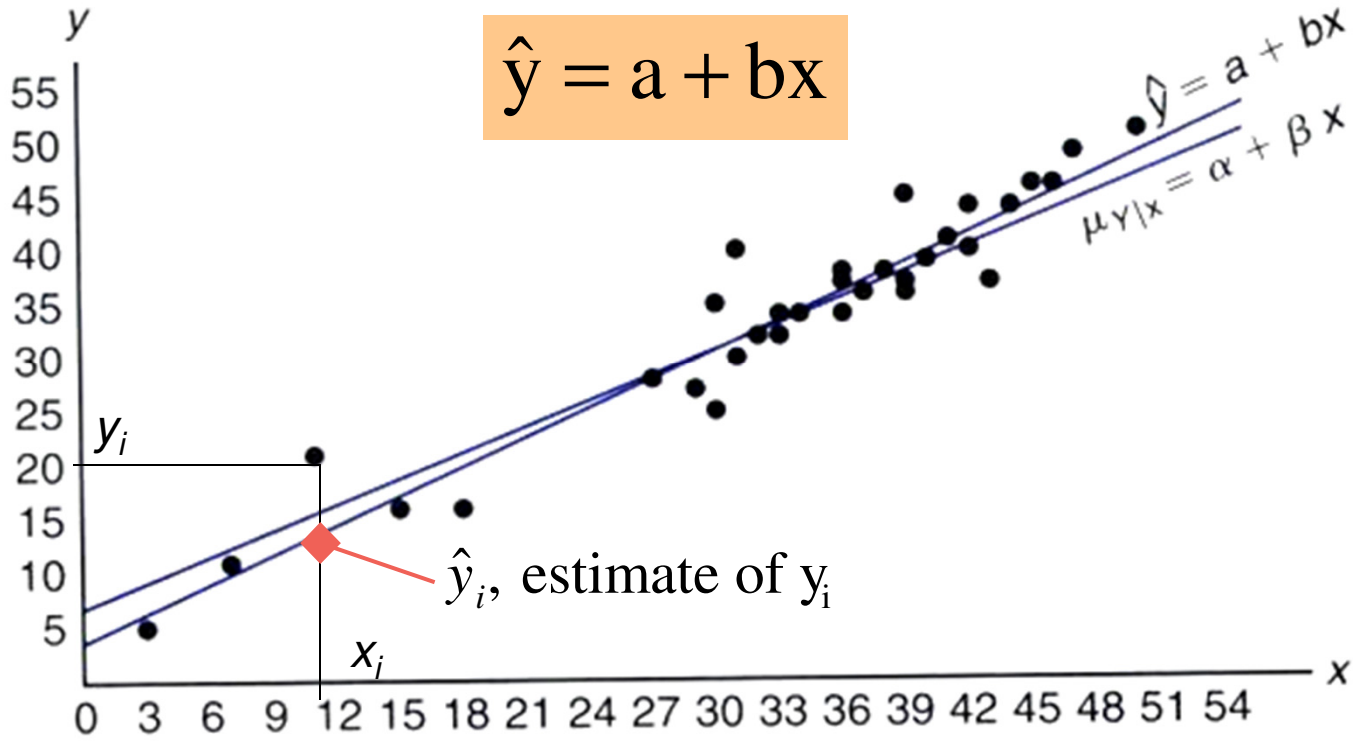
$$\varepsilon_i = Y_i - (\alpha + \beta x_i), \quad \forall i \in [1, n]$$

- The ε_i are independent random variables with the same variance and zero mean. Hence, the response variables Y_i are independent with means $(\alpha+\beta\, x_i)$, and constant variance equal to the variance of ε_i.

Fitting the Model

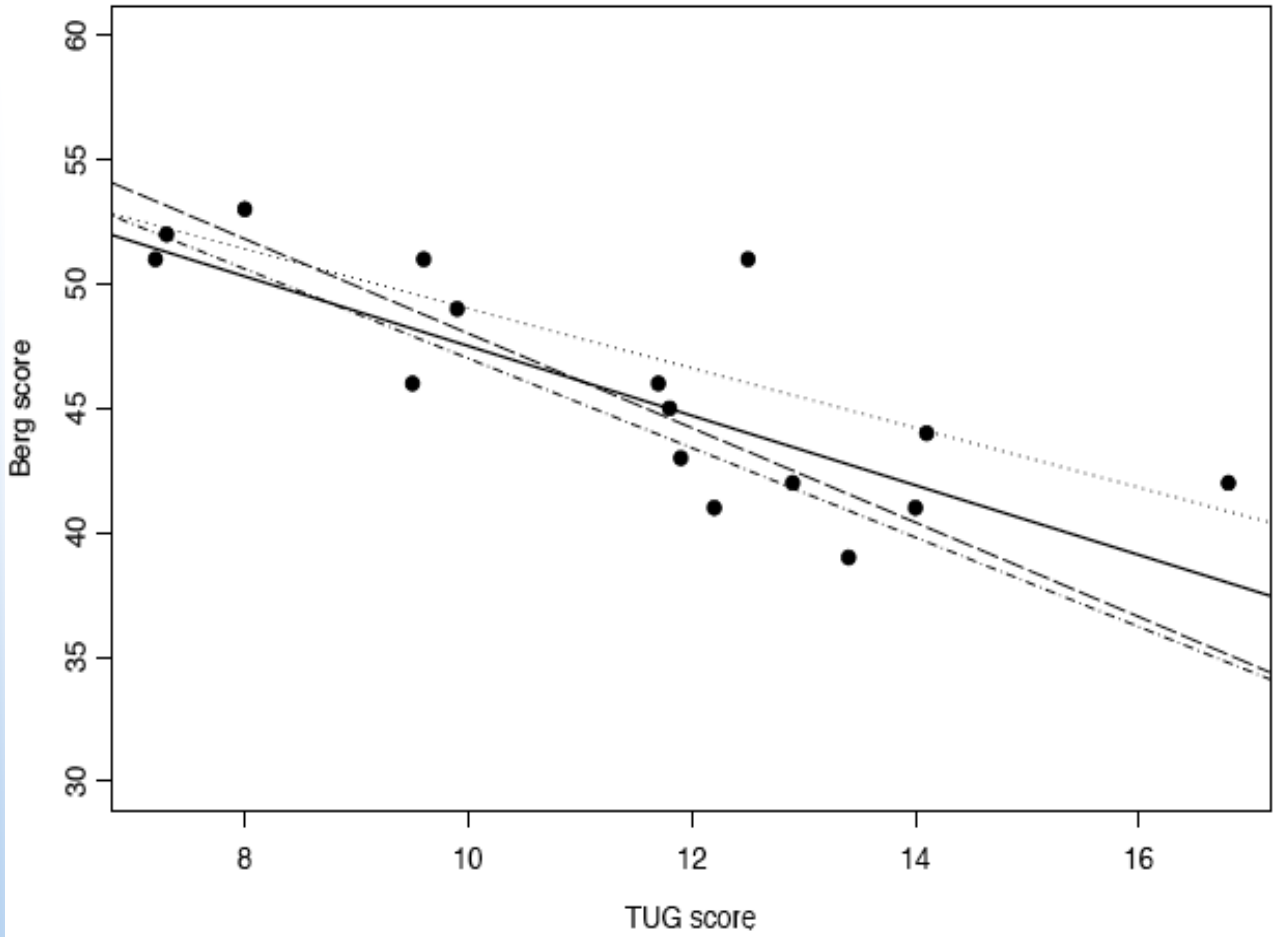
- We need to estimate the regression parameters α and β, in order to estimate the "true" regression line.
- Suppose we denote the estimates a for α and b for β. The estimated or **fitted regression** line is given by:

$$\hat{y} = a + bx$$



Example b (5/6): Fitting the Model

We have decided that a straight line might describe the relationship in the data well.



4 different lines are added to the scatterplot for the data on mobility of elderly people.

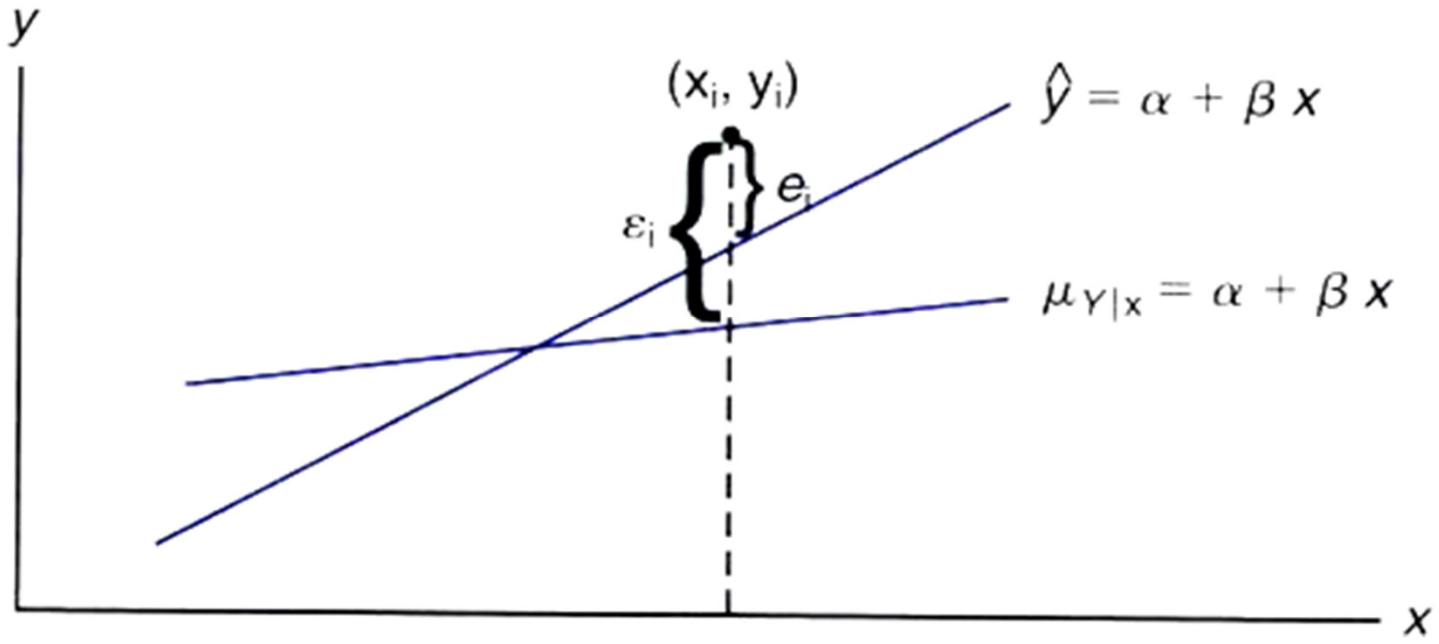
1 or 2 of the lines may look a little better than others, but it is difficult to decide: **which line is the best?**.

Least-Squares and the Fitted Model (1/4)

- The observed value of ε_i is the difference between the i^{th} observation y_i and the linear predictor $(a+b\,x_i)$ at the point x_i. That is:

$$e_i = y_i - \hat{y}_i \quad , \quad \forall i \in [1, n]$$

- e_i, the observed values of ε_i, are called **residuals**.



Least Squares and the Fitted Model (2/4)

- The sum of squared residuals, or, as it is usually called, the residual sum of squares, is denoted by the Sum of Squares of the Errors (**SSE**):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \Leftrightarrow \quad SSE = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- The better the line fits the data, the smaller the residuals will be => To estimate α and β, we need to minimize **SSE** with respect to a and b.

We differentiate **SSE** with respect to a and b and get:

$$\frac{\partial SSE}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) \qquad \frac{\partial SSE}{\partial b} = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i$$

Least Squares and the Fitted Model (3/4)

- Putting the derivatives equal to zero and re-arranging the terms, yields the following equations:

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{cases}$$

- Solving the equations will provide the least squares estimates a and b of the regression parameters α and β, respectively.

Least Squares and the Fitted Model (4/4)

- The least-squares estimates are given by:

$$\begin{cases} b = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ a = \bar{y} - b\bar{x} \end{cases}$$

Shorthand Notation

It is convenient to use the following shorthand notation for the sums involved in the expressions for the parameter estimates

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{N} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

→ **corrected sums of squares**

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

→ **corrected sums of cross product**

Least-Squares Estimators

In the shorthand notation, the least squares estimates of the regression parameters a and b of the slope and intercept of the regression line are given by:

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{S_{xy}}{S_{xx}} \end{cases}$$

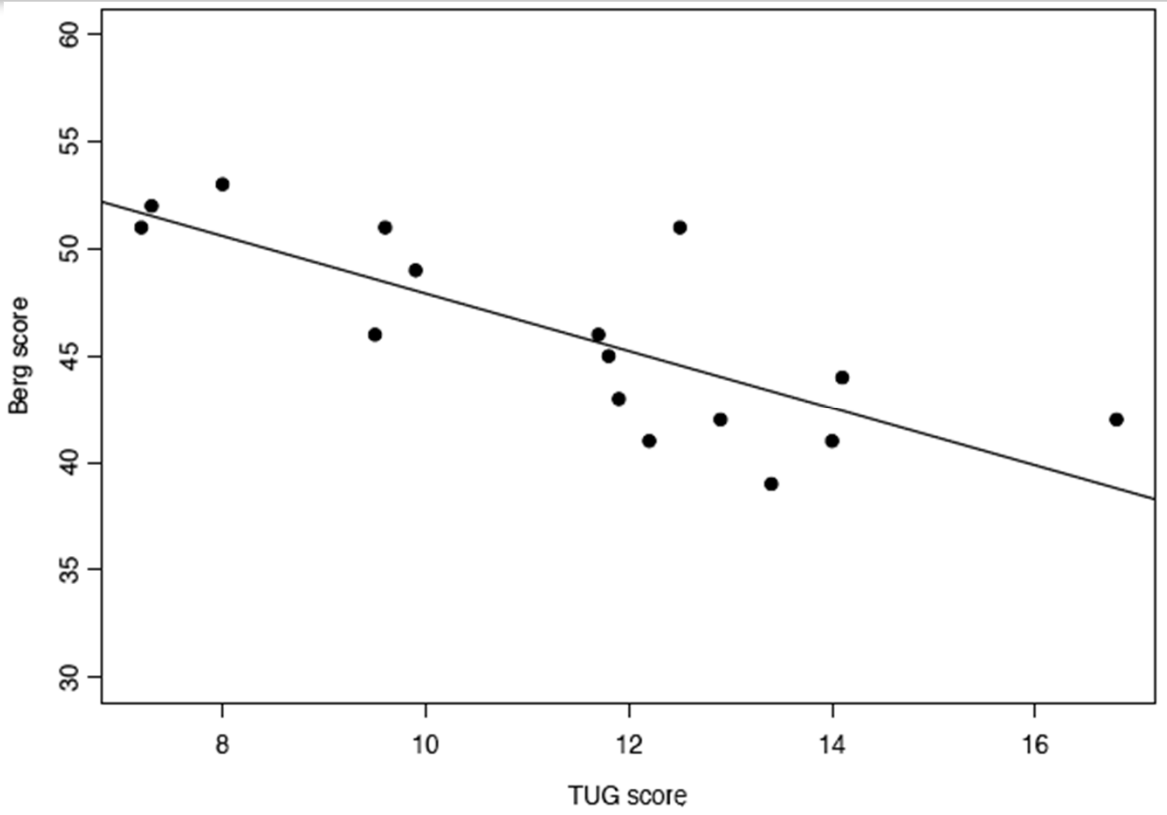
In the shorthand notation, **SSE** becomes:

$$SSE = \sum_{i=1}^{n} e_i^2 = S_{yy} - bS_{xy}$$

Example b (6/6): The Fitted Model

- For the data on mobility of elderly people, the least-squares estimates of the regression parameters are given by:
  - a = 61.314
  - b = -1.340
- So, the fitted least squares line has equation:

$$\hat{y} = 61.314 - 1.340x$$



Example #1

a) Calculate the least-squares estimates of a and b for the data in the table relating the mean insulation compression (dependent variable y in units of 0.10 inches) with the pressure (independent variable x in units of 10 pounds per square inch). Then compute the SSE.

Specimen	Pressure x	Compression y
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

b) Give a practical interpretation of the results

Example #1 (Sol.)

	xi	yi	power(xi,2)	xi.yi	power(yi,2)
	1	1	1	1	1
	2	1	4	2	1
	3	2	9	6	4
	4	2	16	8	4
	5	4	25	20	16
Totals	15	10	55	37	26

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 55 - \frac{1}{5}(15)^2 = 10$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 37 - \frac{(15)(10)}{5} = 7$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{7}{10} = 0.7 \quad , \quad a = \bar{y} - b\bar{x} = \frac{10}{5} - (0.7)\frac{15}{5} = -0.1$$

Example #1 (Sol.)

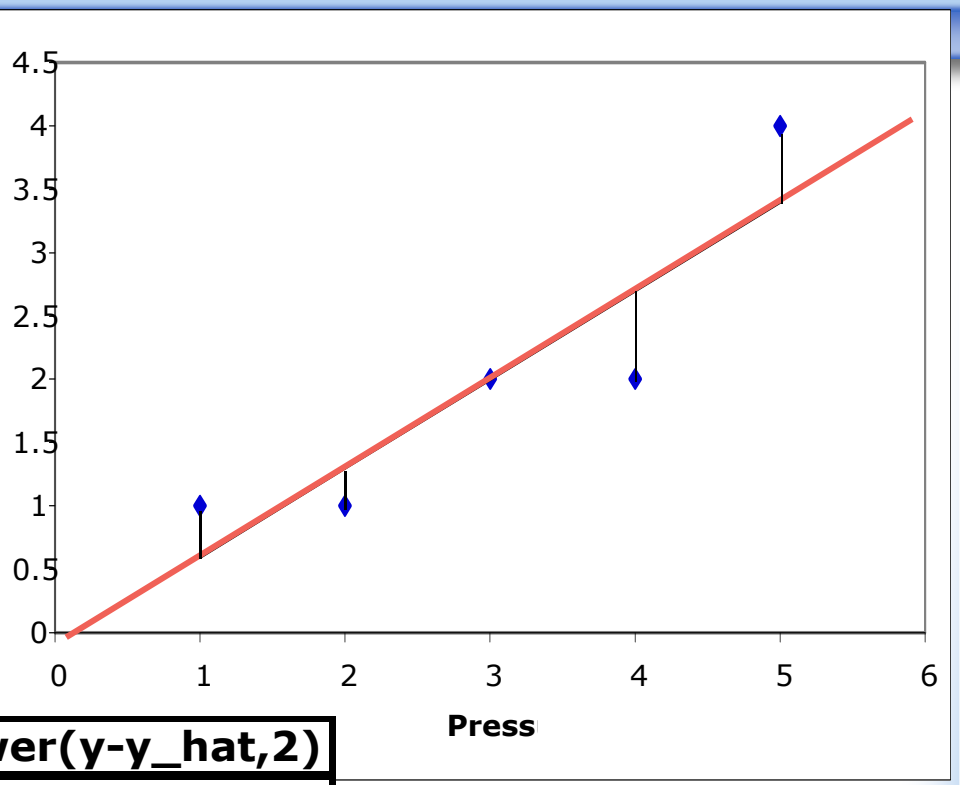
The least - square line is thus :

$$\hat{y} = -0.1 + 0.7x$$

$$SSE = S_{yy} - bS_{xy}$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 26 - \frac{1}{5}(10)^2 = 6$$

$$\text{Thus } SSE = 6 - (0.7)(7) = 1.10$$



xi	yi	y_hat=-.1+.7x	y-y_hat	power(y-y_hat,2)
1	1	0.60	0.40	0.16
2	1	1.30	-0.30	0.09
3	2	2.00	0.00	0.00
4	2	2.70	-0.70	0.49
5	4	3.40	0.60	0.36
			0.00	1.10
			Sum of errors	SSE

Example #1 (Sol.)

b) Interpretation of the least-squares slope b=0.7:

The compression y will increase 0.7 unit for every 1-unit increase in pressure x.

Interpretation of the least-squares intercept a=-0.1:

It is our estimate of compression when pressure is set at 0. Can compression be negative?

We are attempting to use least-squares model to predict y for a value of x that is outside the range of the sample data and this is impractical. => **a will not always have a practical interpretation!**

The Coefficient of Correlation

- A statistical concept closely related to linear regression is that of correlation.

- **Correlation** refers to the situation where we have two random variables **X** and **Y**, and wish to measure the strength of the linear association between the two: the association is **strong** if knowing the value of one variable can give us a (reasonably) precise idea of the value of the other variable; the association is **weak** if we can only get a very rough estimate.

- **Note:** that there is an important difference between this situation and the linear regression situation: here, both X and Y are random, whereas in regression we always regard the explanatory variable x as a non-random variable.

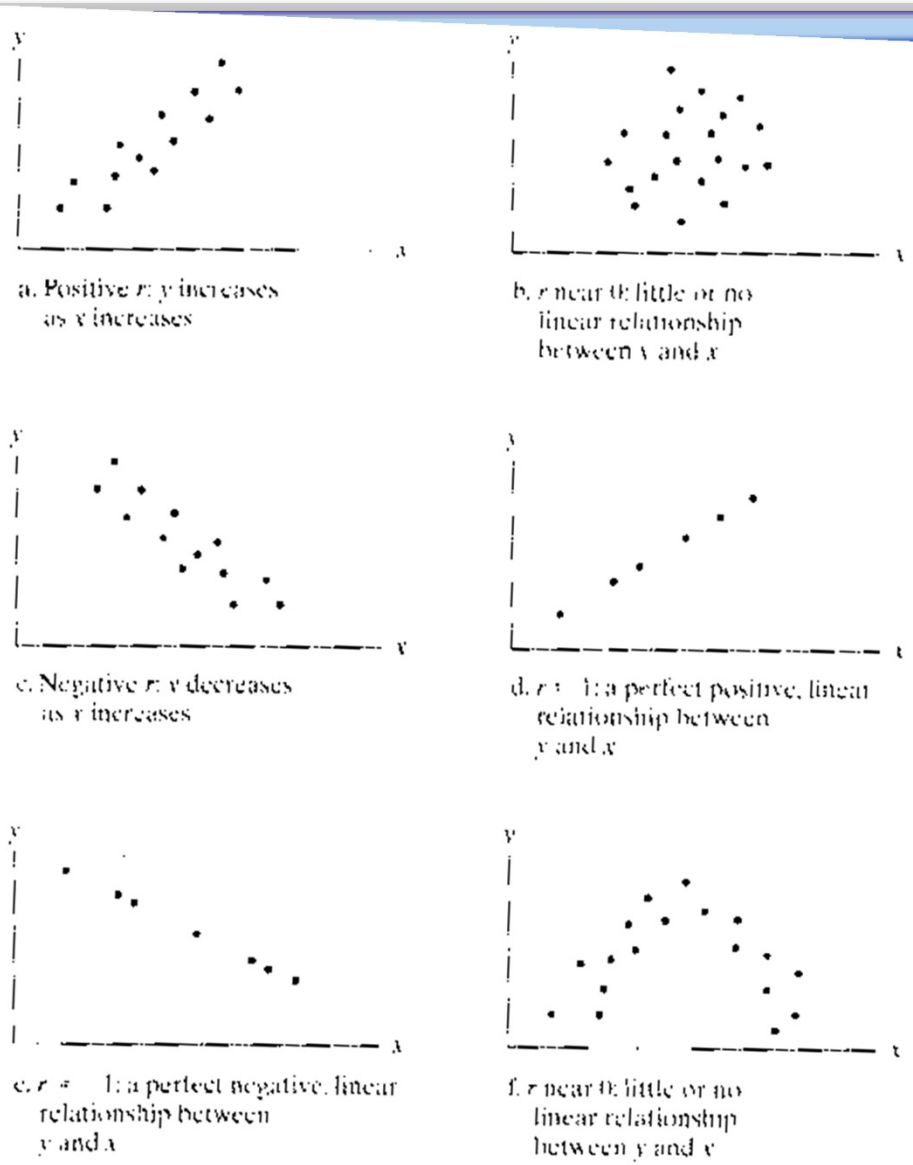
The Coefficient of Correlation

- The **coefficient of correlation r** is a measure of the strength of the linear relationship between 2 variables x and y **in the sample**. It is computed by:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

r is *scaleless* (unitless) and is always between -1 and +1.

Implications of the Coefficient of Correlation



a. Positive r: y increases as x increases

b. r near 0: little or no linear relationship between y and x

c. Negative r: y decreases as x increases

d. $r = 1$: a perfect positive, linear relationship between y and x

e. $r = -1$: a perfect negative, linear relationship between y and x

f. r near 0: little or no linear relationship between y and x

Warning:

High correlation does not imply causality. If a large positive or negative value of the sample correlation coefficient **r** is observed, it is incorrect to conclude that a change in x causes a change in y. The only valid conclusion is that a linear trend **may** exist between x and y.

Coefficient of Determination

- We used the principle of least-squares to fit the 'best' straight line to data. But how well does the least-squares line explain the variation in the data?

- We are going to describe here a measure for roughly assessing how well a fitted line describes the variation in data: the **coefficient of determination.**

- The coefficient of determination compares the amount of variation in the data away from the fitted line with the total amount of variation in the data.

Coefficient of Determination

- If we did not have the linear model we would have to use the 'naive' model instead, i.e. $\hat{y} = \bar{y}$

  The variation away from the naive model is:

$$S_{yy} = \sum_{i=1}^{n} (Y_i - \bar{y})^2 \quad = \quad SST$$

  This is the **total** amount of variation in the data and is known as the total corrected sum of squares (**SST**).

- However, if we use the least squares line as model, the variation away from model is only:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

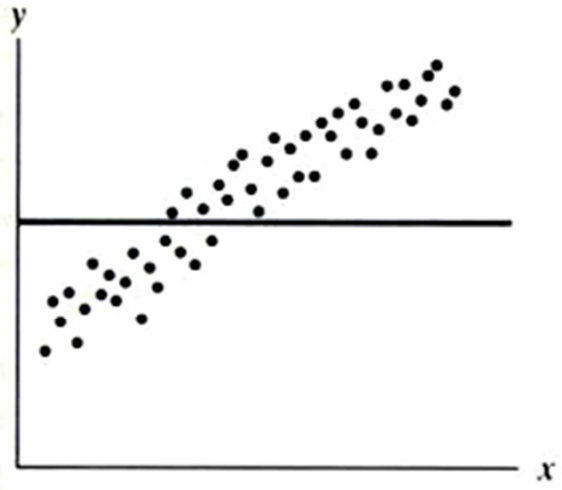
Coefficient of Determination

- A measure of the strength of the linear relationship between Y and x is the **coefficient of determination r^2.**

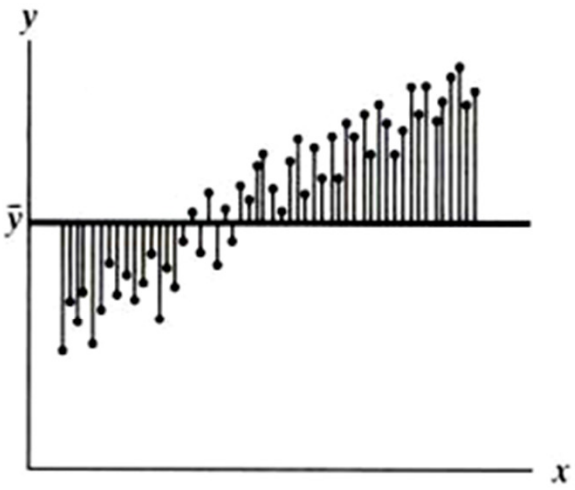
- It is the proportional reduction in variation obtained by using the least squares line instead of the naive model. That is, the reduction in variation away from the model (S_{yy}-SSE) as a proportion of the total variation S_{yy}:

$$r^2 = \frac{S_{yy} - SSE}{S_{yy}} = \frac{SST - SSE}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Coefficient of Determination



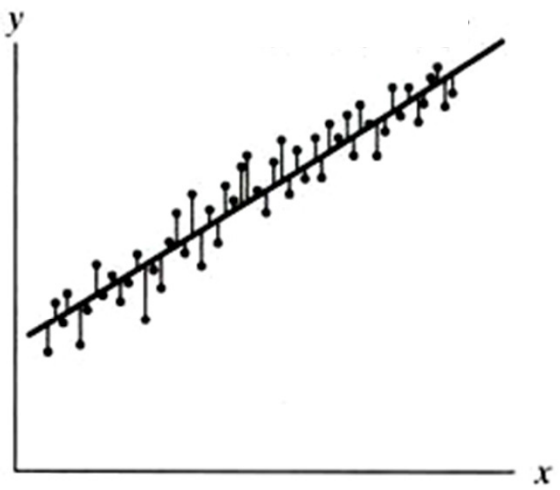
a. Scattergram of data



b. Assumption: x contributes no information for predicting y;

$$\hat{y} = \bar{y}$$

SSE ≈ S_{yy}



c. Assumption: x contributes information for predicting y;

$$\hat{y} = a + bx$$

SSE minimal

Coefficient of Determination

- The value of r^2 will always lie between 0 and 1 (or, in percentage, between 0% and 100%).

- $r^2 = 1$ if $b \neq 1$ and SSE = 0,
  if all the data points lie precisely on the fitted straight line.

- If r^2 is close to 1, it is an indication that the data points lie close to the least-squares line.

- $r^2 = 0$ if SSE = S_{yy},
  if the fitted straight-line model offers no more information about the value of Y than the naive model does.

Example #2

Based on the data given in **Example #1** and the corresponding obtained results, compute the coefficient of determination and comment on its value.

Example #2 (Sol.)

Known from/Computed in Example #1:

$S_{xx} = 10$ $S_{xy} = 7$ $S_{yy} = 6$ SSE = 1.10

$$\hat{y} = -0.1 + 0.7x$$

What to Use $r^2 = \dfrac{S_{yy} - SSE}{S_{yy}}$

$$r^2 = \frac{S_{yy} - SSE}{S_{yy}} = \frac{6 - 1.10}{6} = 0.8167$$

82% of the sample variation in compression values can be explained by the least-squares line

Estimating the Variance

- We found that the principle of least squares can provide estimates of the regression parameters α and β in a simple linear regression model.
- If you repeat the experiment over and over again, each time using the same fixed variables x, the resulting estimates will most likely differ from experiment to experiment. The distributional assumptions imply that the Y_i's are independently distributed with mean $\mu_{Y|xi}=\alpha+\beta x$ and equal variances $\sigma^2_{Y|xi}=\sigma^2$ for i=1, …, n.
- In order to fit the model we also need an estimate for the common variance σ^2.
- Such an estimate is required for making statistical inferences about the true straight-line relationship between x and y.

Estimating the Variance

- Since σ^2 is the common variance of the residuals e_i $i \in [1,n]$, it would be natural to estimate it by the sample variance of the fitted residuals.

- An **unbiased estimate of the common variance**, σ^2, is given by:

$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

- $\mathbf{s^2}$ is a point estimate for $\sigma^2_{Y|xi} = \sigma^2$.

Example #3

Based on the data given in **Example #1** and the corresponding obtained results, estimate the variance.

Example #3 (Sol.)

Known from/Computed in Example #1:

$S_{xx} = 10$ $S_{xy} = 7$ $S_{yy} = 6$ SSE = 1.10

$$\hat{y} = -0.1 + 0.7x$$

What to Use $s^2 = \frac{SSE}{n-2}$

$$s^2 = \frac{SSE}{n-2} = \frac{1.10}{5-2} = 0.367$$

Confidence Interval

- A 100(1-α)% confidence interval for the mean response $\mu_{Y|x0}$ is:

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

where $t_{\alpha/2}$ is a value of the t-distribution with df =n - 2.

Example #4

Based on the data given in **Example #1** and the corresponding obtained results in **Example #1** and **Example #3** , find a 95% Confidence Interval for the mean insulation compression when the pressure is 4 units.

Example #4 (Sol.)

Known from/Computed in Example #1:

$S_{xx} = 10$ $n = 5$ $\bar{x} = 15/5 = 3$ $\hat{y} = -0.1 + 0.7x$

Known from/Computed in Example #3:

$s^2 = 0.367$

Given:

$x_0 = 4$ $\alpha = 0.05$

Required: 95% CI of the mean $\mu_{Y|x_0}$

What to Use:

$$C.I \text{ for } \mu_{Y|x_0} = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Example #4 (Sol.)

$$\hat{y}_0 = -0.1 + 0.7x_0 = -0.1 + 0.7 * 4 = 2.7$$

$\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \rightarrow t_{\alpha/2} = 3.182 \quad (df = 5 - 2 = 3)$

$$C.I \text{ for } \mu_{Y|x_0} = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$C.I \text{ for } \mu_{Y|4} = 2.7 \pm 3.182 * \sqrt{0.367} \sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$C.I \text{ for } \mu_{Y|4} = \mathbf{2.7 \pm 3.182 * 0.606 * 0.548}$$

$$= 2.7 \pm 1.057$$

$$= 1.643 \ , \ 3.757$$

Prediction Interval

- A 100(1-α)% prediction interval for a single response y_0 is given by:

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

where $t_{\alpha/2}$ is a value of the t-distribution with df=n-2.

Example #5

Based on the data given in **Example #1** and the corresponding obtained results in **Example #1** and **Example #3** , predict the amount of compression for an individual piece of insulation subjected to a pressure of 4 units. Use a 95% Prediction Interval (PI).

Example #5 (Sol.)

Known from/Computed in Example #1:

$S_{xx} = 10$ $n = 5$ $\bar{x} = 15/5 = 3$ $\hat{y} = -0.1 + 0.7x$

Known from/Computed in Example #3:

$s^2 = 0.367$

Given:

$x_0 = 4$ $\alpha = 0.05$

Required: 95% PI of the response y_0

What to Use:

$$\text{P.I for } y_0 = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Example #5 (Sol.)

$$y_0 = -0.1 + 0.7x_0 = -0.1 + 0.7 * 4 = 2.7$$

$\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \rightarrow t_{\alpha/2} = 3.182 \quad (df = 5 - 2 = 3)$

$$\text{P.I for } y_0 = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$\text{P.I for } y_0 \mid_4 = 2.7 \pm 3.182 * \sqrt{0.367} \sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$\begin{aligned} \text{P.I for } y_0 \mid_4 &= \mathbf{2.7 \pm 3.182 * 0.606 * 1.140} \\ &= 2.7 \pm 2.199 \\ &= 0.501 \ , \ 4.899 \end{aligned}$$

Example #6 (1/2)

A contract engineer at DuPont Corp. studied the rate at which a spilled volatile liquid spread across a surface (*Chemical Eng. Progress, Jan. 2005*). Assume 50 gallons of methanol spills onto a level surface outdoors. The engineer used derived empirical formulas to calculate the mass (in pounds) of the spill after a period of time ranging from 0 to 60 minutes. The calculated mass values are given in the table.

a) Construct a scatterplot for the data with y: calculated mass, x: time.

b) Find the least-squares line relating mass (y) to time (x). How well the regression line fit the sample data?

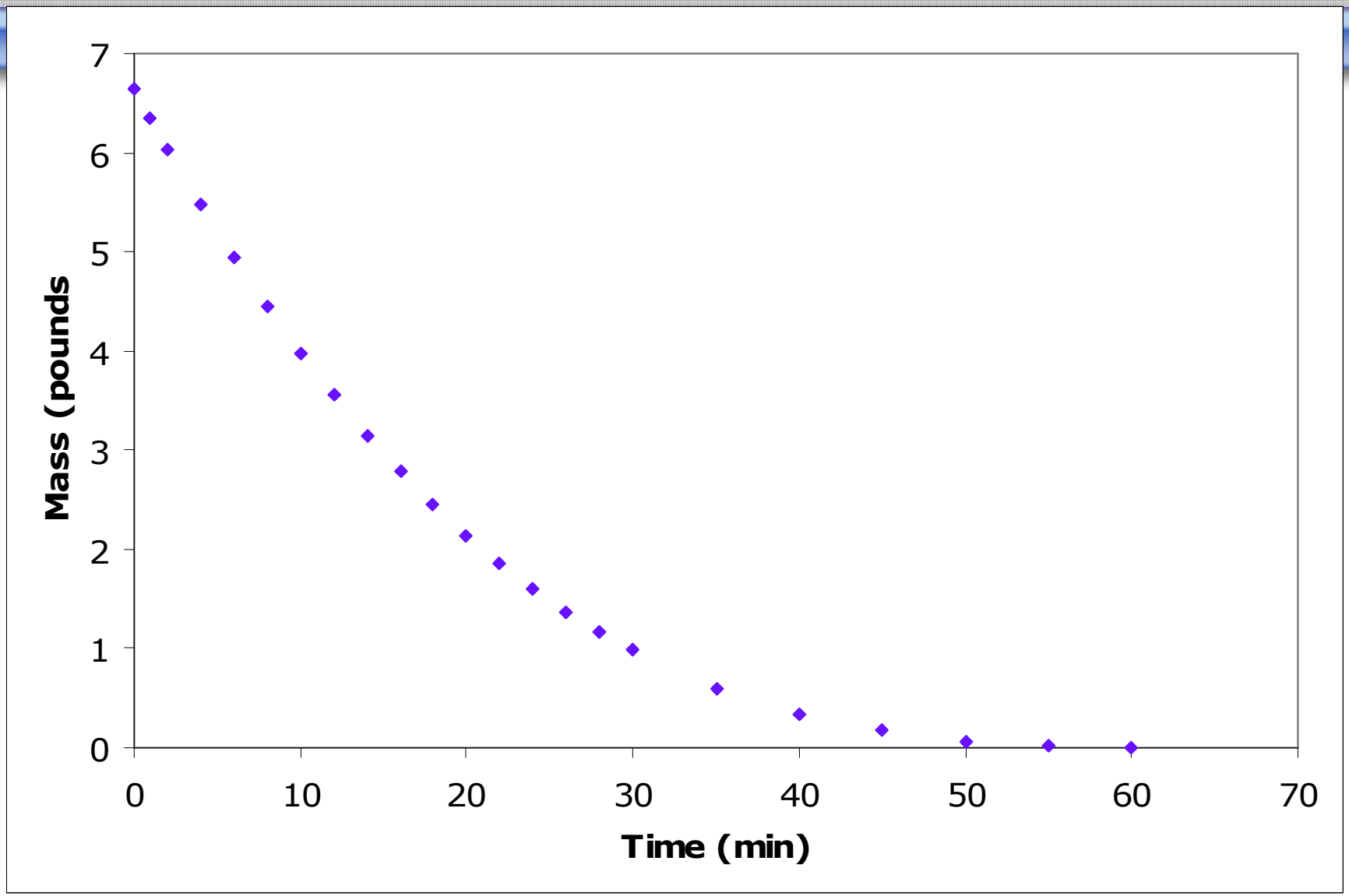
c) Find a 95% confidence interval for the mean mass of all spills with an elapsed time of 8 min.

d) Find a 90% prediction interval for the mass of a spill when the elapsed time is 8 min.

Example #6 (2/2)

Time (min)	Mass (pounds)	Time (min)	Mass (pounds)
0	6.64	20	2.14
1	6.34	22	1.86
2	6.04	24	1.6
4	5.47	26	1.37
6	4.94	28	1.17
8	4.44	30	0.98
10	3.98	35	0.6
12	3.55	40	0.34
14	3.15	45	0.17
16	2.79	50	0.06
18	2.45	55	0.02
		60	0

Example #6 (Sol.) "a"

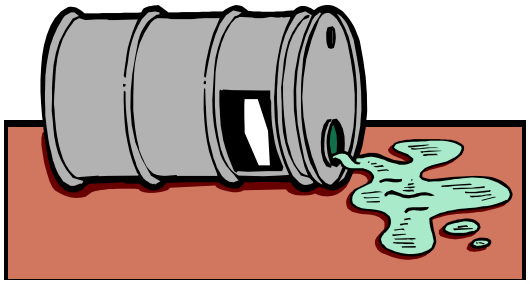


Example #6 (Sol.) "b"

Given: $n = 23$

Computed: $\bar{x} = 22.8696$ $\bar{y} = 2.613$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 6906.6087$$



$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = -787.5109$$

$$\text{slope } b = \frac{S_{xy}}{S_{xx}} = -0.114 \quad , \quad \text{intercept } a = \bar{y} - b\bar{x} = 5.221$$

Example #6 (Sol.) "b"

How well the regression line fit the sample data?

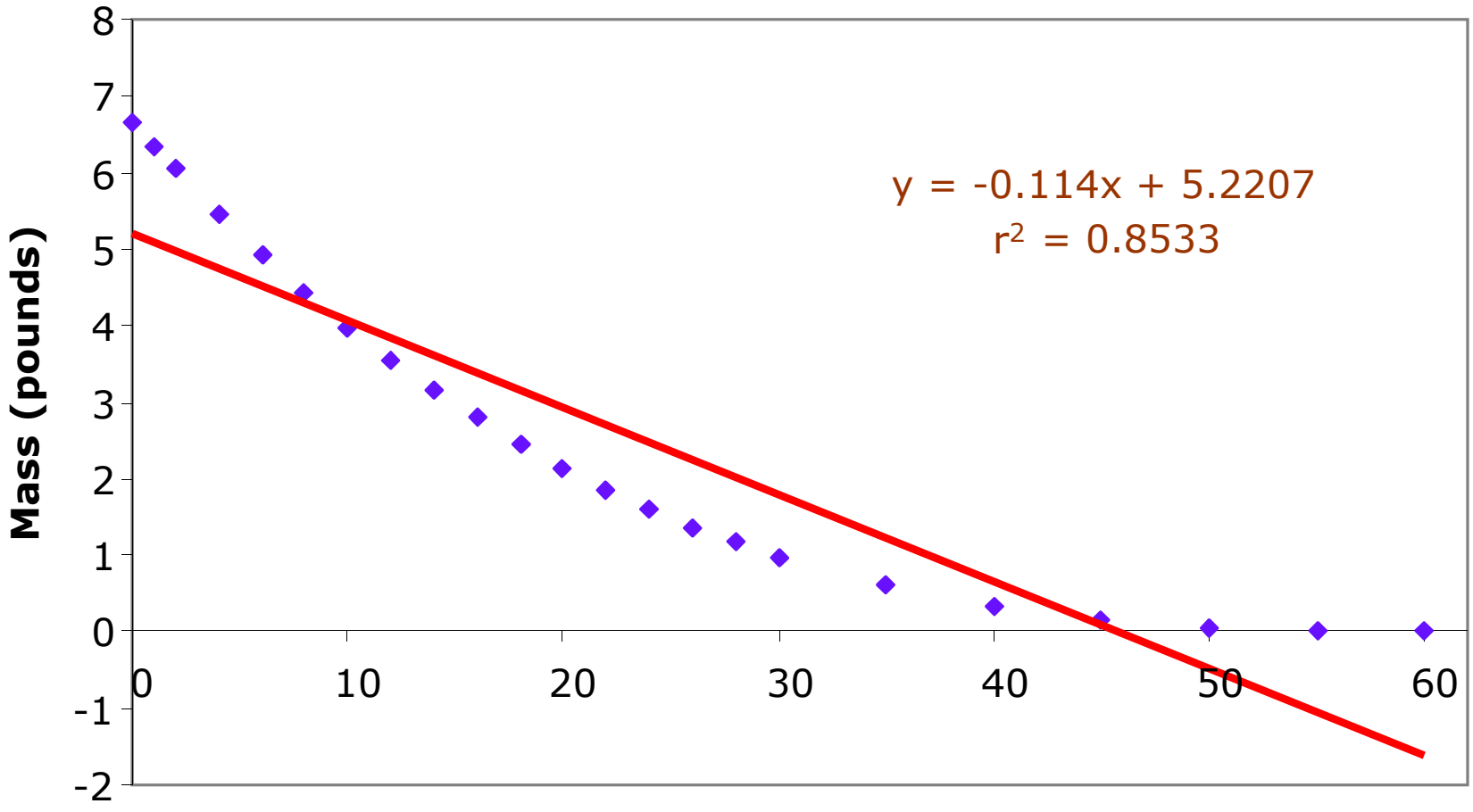
coefficient of determination r^2

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 105.2269$$

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = 0.853$$

Example #6 (Sol.) "b"



Example #6 (Sol.) "c"

95% C.I. for the mean mass with $x_0 = 8$:

$$\hat{y}_0 = 5.221 - 0.114 * 8 \quad = 4.3085$$

$$\alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \rightarrow t_{\alpha/2} = 2.080 \quad (df = 23 - 2 = 21)$$

$$\text{C.I for } \mu_{Y|x_0} = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

$$s^2 = \frac{105.2269 - (-0.114) * -787.5109}{23 - 2} = 0.7357$$

Example #6 (Sol.) "c"

$$\text{C.I for } \mu_{Y|x_0} = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$\text{C.I for } \mu_{Y|8} = 4.3085 \pm 2.080 * \sqrt{0.7357} \sqrt{\frac{1}{23} + \frac{(8 - 22.8696)^2}{6906.6087}}$$

$$\begin{aligned} \text{C.I for } \mu_{Y|8} &= 4.3085 \pm 2.080 * 0.8577 * 0.2747 \\ &= 4.3085 \pm 0.4901 \\ &= 3.8184 \ , \ 4.7986 \end{aligned}$$

Example #6 (Sol.) "d"

90% P.I. for the mass with $x_0 = 8$:

$\alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \rightarrow t_{\alpha/2} = 1.721 \quad (df = 23 - 2 = 21)$

$$P.I \text{ for } y_0 = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$P.I \text{ for } y_{0|8} = 4.3085 \pm 1.721 * \sqrt{0.7357} \sqrt{1 + \frac{1}{23} + \frac{(8 - 22.8696)^2}{6906.6087}}$$

$$P.I \text{ for } y_{0|8} = 4.3085 \pm 1.721 * 0.8577 * 1.0371$$
$$= 4.3085 \pm 1.5308$$
$$= 2.7777 \ , \ 5.8393$$

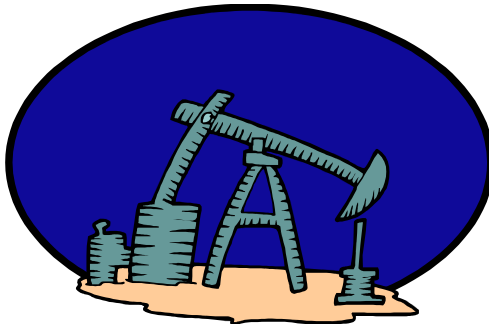
Example #7 (1/2)

Two processes for hydraulic drilling of rock are dry drilling and wet drilling. In a dry hole, compressed air is forced down the drill rods to flush the cuttings and drive the hammer; in a wet hole, water is forced down. An experiment was conducted to determine whether the time y it takes to dry drill a distance of 5 feet in rock increases with depth x. The results for one portion of the experiment are shown in the table:

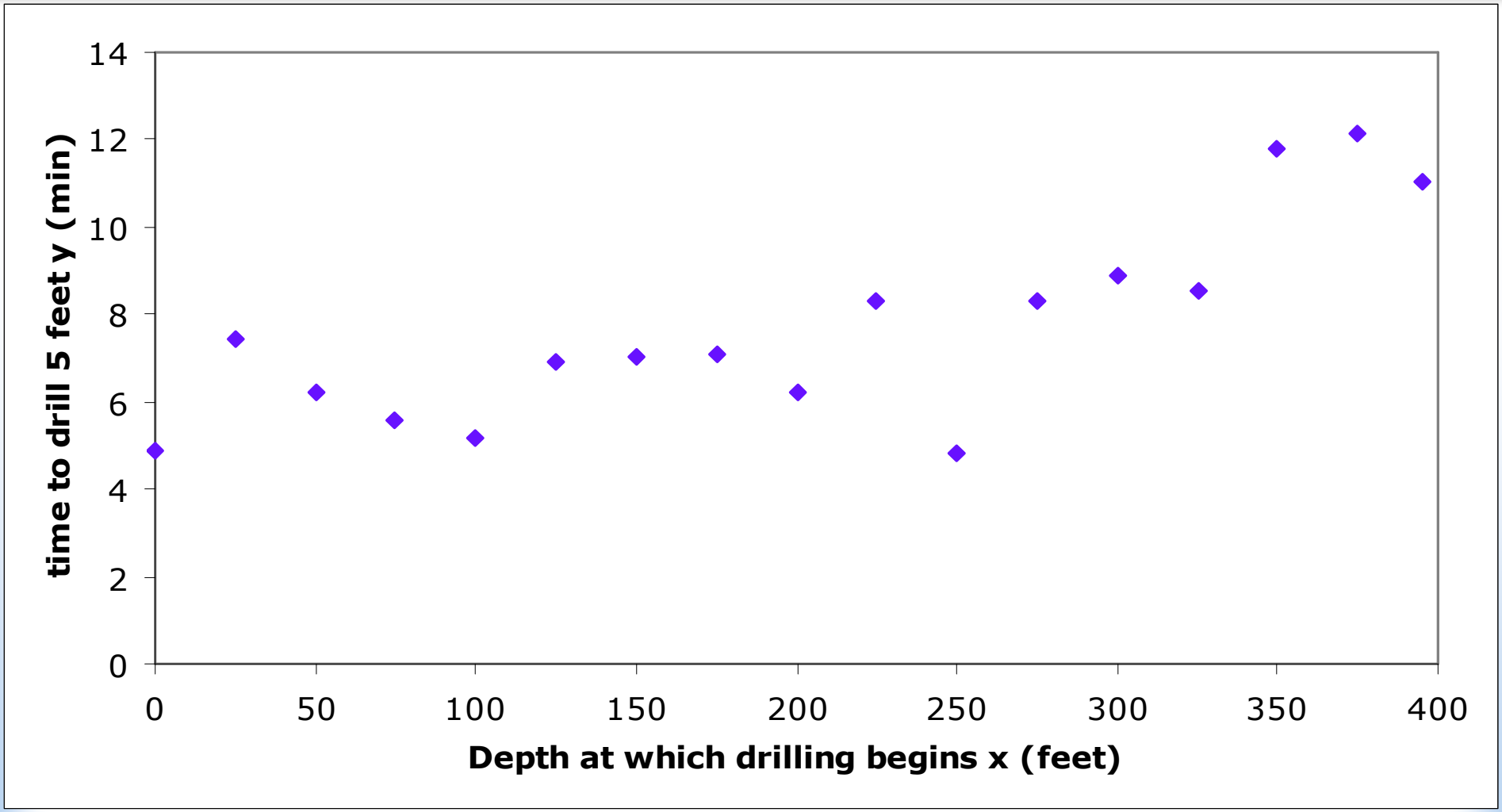
a) Construct a scatterplot for the data.
b) Find the least-squares line relating time (y) to depth (x). How well the regression line fit the sample data?
c) Find a 90% confidence interval for the mean time to dry a drill of 5 feet when x = 178.
d) Find a 90% prediction interval for the the time to dry a drill of 5 feet when x = 178.

Example #7 (2/2)

D at which drilling begins (feet)	Time to drill 5 feet (min)
0	4.9
25	7.41
50	6.19
75	5.57
100	5.17
125	6.89
150	7.05
175	7.11
200	6.19
225	8.28
250	4.84
275	8.29
300	8.91
325	8.54
350	11.79
375	12.12
395	11.02



Example #7 (Sol.) "a"



Example #7 (Sol.) "b"

Given: $n = 17$

Computed: $\bar{x} = 199.7059$ $\bar{y} = 7.6629$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 253023.529$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 3640.465$$

$$\text{slope } b = \frac{S_{xy}}{S_{xx}} = 0.0144 \quad , \quad \text{intercept } a = \bar{y} - b\bar{x} = 4.7896$$

Example #7 (Sol.) "b"

How well the regression line fit the sample data?

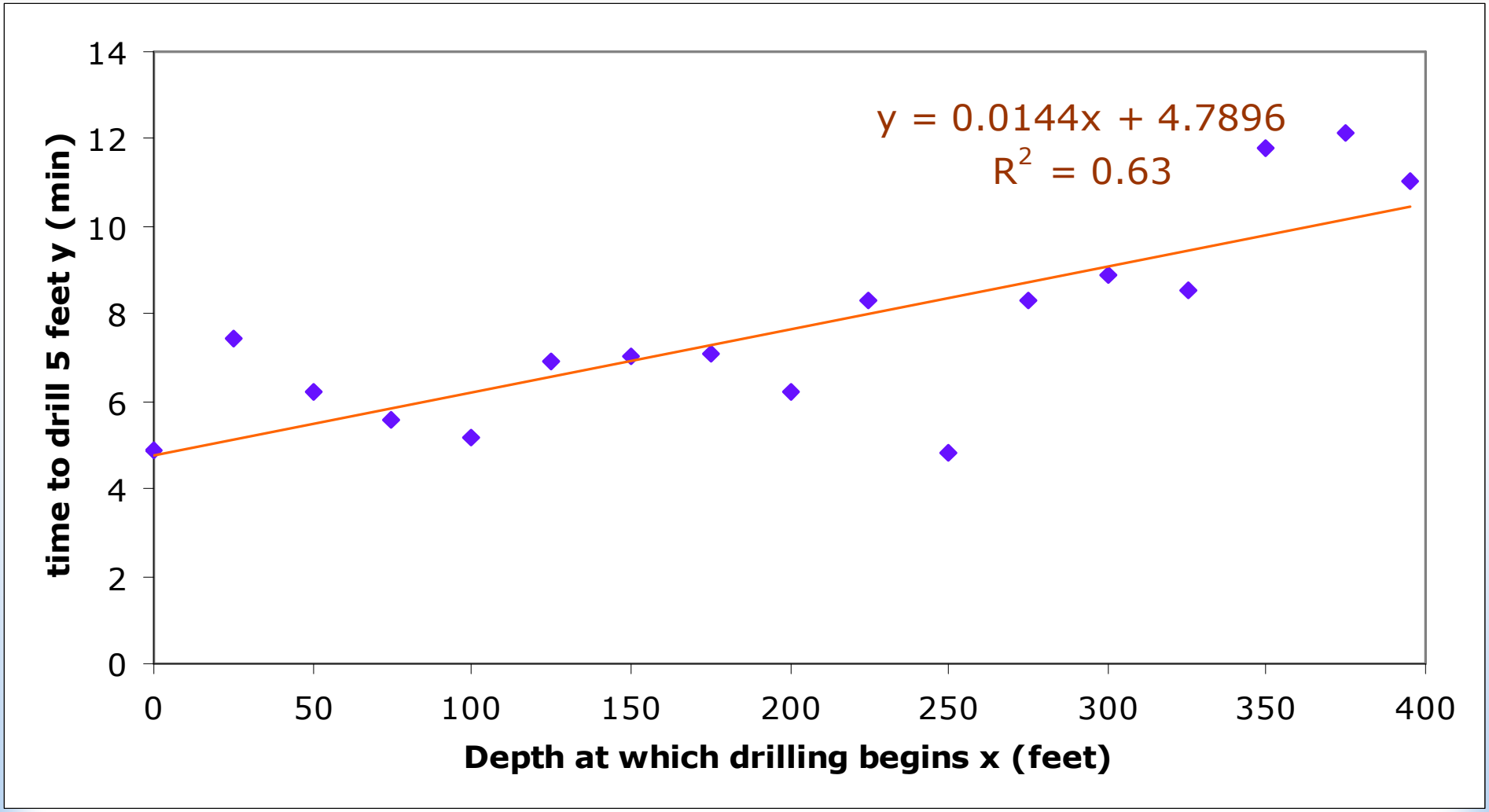
coefficient of determination r^2

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 83.146$$

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = 0.63$$

Example #7 (Sol.) "b"



Example #7 (Sol.) "c"

90% C.I. for the mean time to drill with $x_0 = 178$:

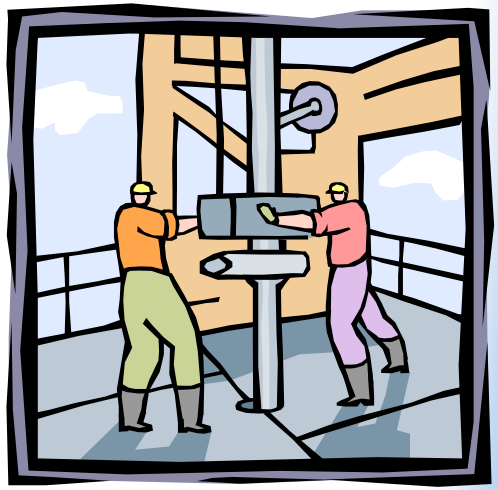
$$\hat{y}_0 = 4.7896 + 0.0144 * 178 \quad = 7.3528$$

$$\alpha = 0.10 \Rightarrow \alpha/2 = 0.05 \rightarrow t_{\alpha/2} = 1.753 \quad (df = 17 - 2 = 15)$$

$$\text{C.I for } \mu_{Y|x_0} = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

$$s^2 = \frac{83.146 - (0.0144) * 3640.465}{17 - 2} = 2.0482$$



Example #7 (Sol.) "c"

$$\text{C.I for } \mu_{Y|x_0} = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$\text{C.I for } \mu_{Y|178} = 7.3528 \pm 1.753 * \sqrt{2.0482} \sqrt{\frac{1}{17} + \frac{(178 - 199.7059)^2}{253023.529}}$$

$$\begin{aligned} \text{C.I for } \mu_{Y|178} &= 7.3528 \pm 1.753 * 1.4312 * 0.2463 \\ &= 7.3528 \pm 0.6181 \\ &= 6.7344 \ , \ 7.9709 \end{aligned}$$

Example #7 (Sol.) "d"

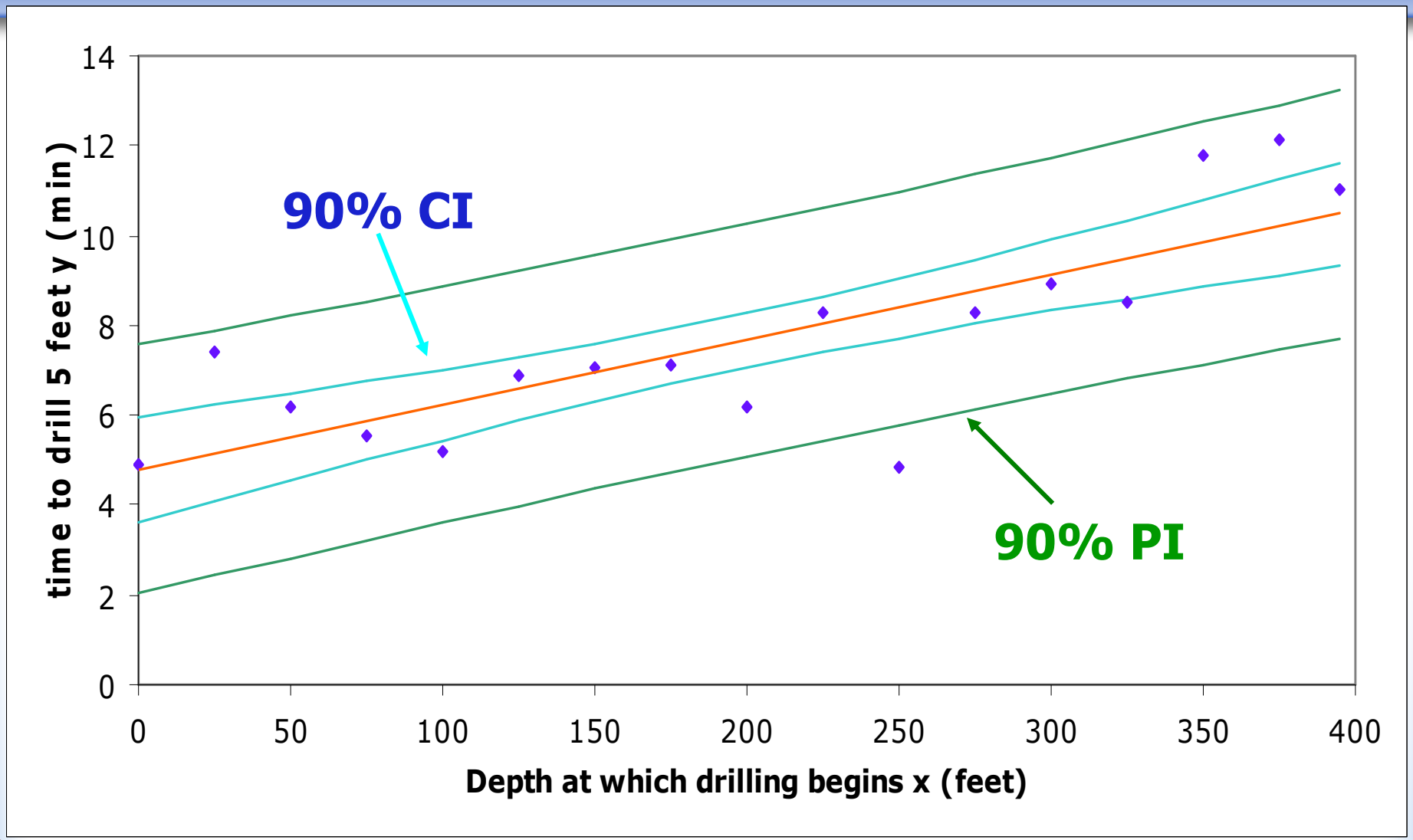
90% P.I. for the time to drill with $x_0 = 178$:

$$\text{P.I for } y_0 = \hat{y}_0 \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$\text{P.I for } y_{0|178} = 7.3528 \pm 1.753 * \sqrt{2.0482} \sqrt{1 + \frac{1}{17} + \frac{(178 - 199.7059)^2}{253023.529}}$$

$$\begin{aligned} \text{P.I for } y_{0|178} &= 7.3528 \pm 1.753 * 1.4312 * 1.0299 \\ &= 7.3528 \pm 2.5839 \\ &= 4.7689 \ , \ 9.9367 \end{aligned}$$

Example #7 (Sol.)



Textbook Sections

- 11.1
- 11.2
- 11.3
- 11.4
- 11.5
- 11.11
- 11.12