# Data Analytics

ECON 1008, Semester 1, 2019

Giulio Zanella
University of Adelaide
School of Economics

# Back to Chapter 4:
# Describing two numerical variables

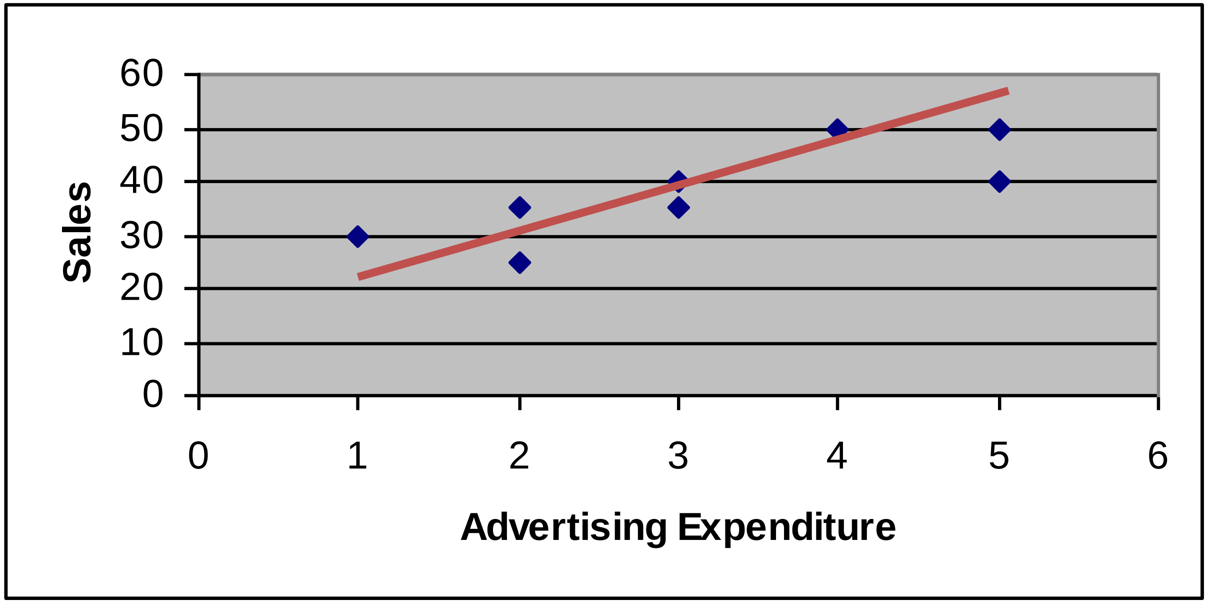
Remember: often we are interested in the **relationship between two numerical variables.**

# Scatter diagram

A scatter diagram can describe the relationship between advertising expenditure and sales.

| Advert | Sales |
|---|---|
| 1 | 30 |
| 3 | 40 |
| 5 | 40 |
| 4 | 50 |
| 2 | 35 |
| 5 | 50 |
| 3 | 35 |
| 2 | 25 |

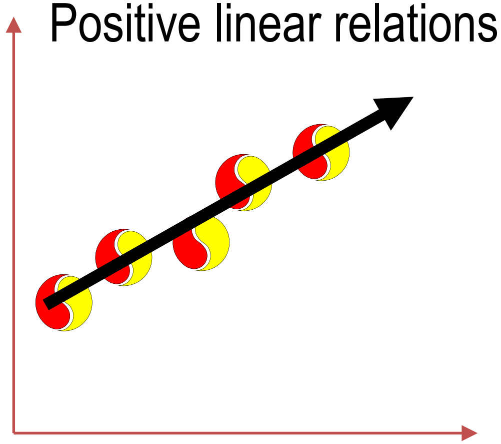
Excel scatter diagram



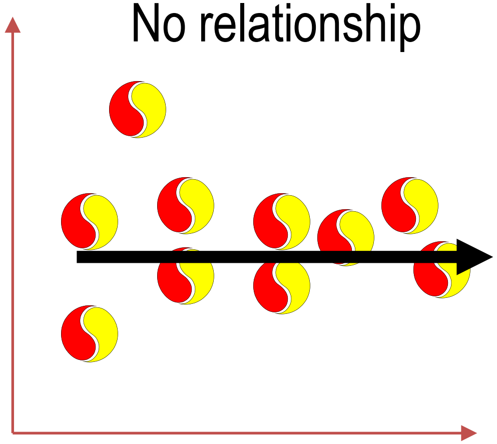
We see a **positive linear relationship**

# Typical patterns

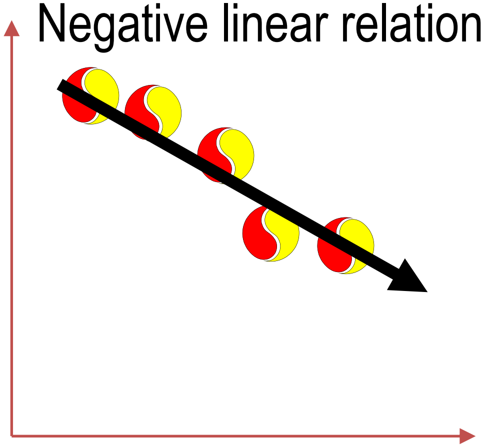
Positive linear relationship



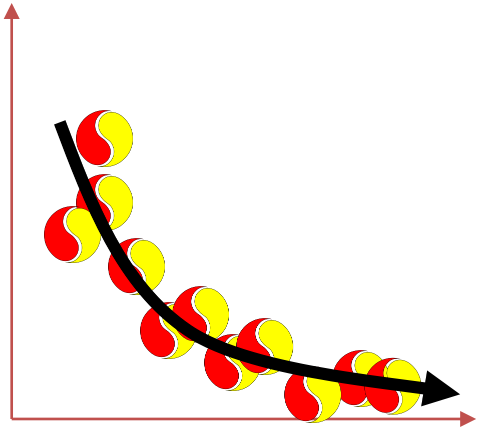
No relationship



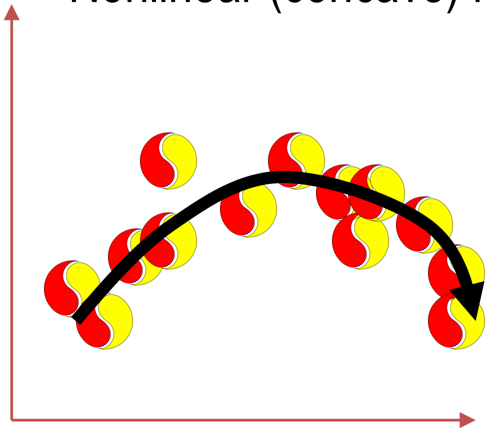
Negative linear relationship



Negative nonlinear relationship



Nonlinear (concave) relationship



# Back to Chapter 5: Measures of Association

Two numerical measures for the **description of a linear relationship** between two variables depicted in a scatter diagram.

1. **Covariance** (is there any pattern to the way two variables move together?)
2. **Correlation** coefficient (how strong is the linear relationship between two variables?)

# Covariance…

population mean of variable X, variable Y

$$\text{Population covariance} = \sigma_{xy} = \frac{\sum_{i=1}^{N} (\mathrm{x}_\mathrm{i} - \mu_x)(y_i - \mu_y)}{\mathrm{N}}$$

sample mean of variable X, variable Y

$$\text{Sample covariance} = \mathrm{s}_{xy} = \frac{\sum_{i=1}^{n} (\mathrm{x}_\mathrm{i} - \bar{x})(y_i - \bar{y})}{\mathrm{n - 1}}$$

Note: divisor is n-1, not n as you may expect.

**Can you see an analogy with the variance?**

# Covariance…

There was a ‘shortcut’ for calculating sample variance without having to calculate the sample mean.

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

There is also a shortcut for calculating sample covariance without having to first calculate the means:

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

# Covariance…

When two variables move in the ***same direction*** (both increase or both decrease), the covariance will be a ***large positive number***.

When two variables move in ***opposite directions***, the covariance is a ***large negative number***.

When there is ***no particular pattern***, the covariance is a ***small number***.

However, it is often difficult to determine whether a particular covariance is large or small…

# Coefficient of Correlation…

The coefficient of correlation is defined as the covariance divided by the standard deviations of the variables:

$$\text{Population coefficient of correlation: } \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Greek letter ‘rho’

$$\text{Sample coefficient of correlation: } r = \frac{s_{xy}}{s_x s_y}$$

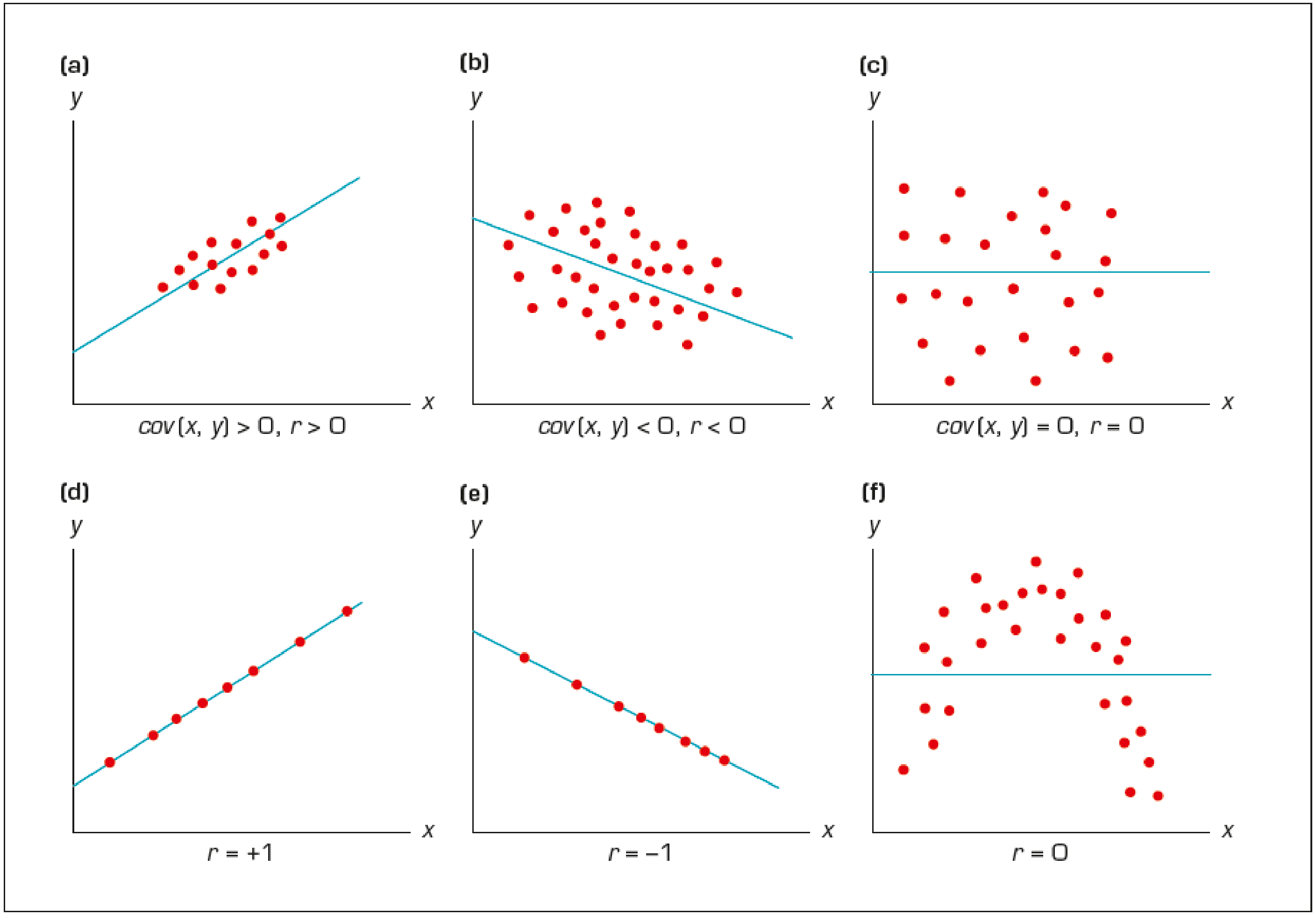
# Coefficient of Correlation…

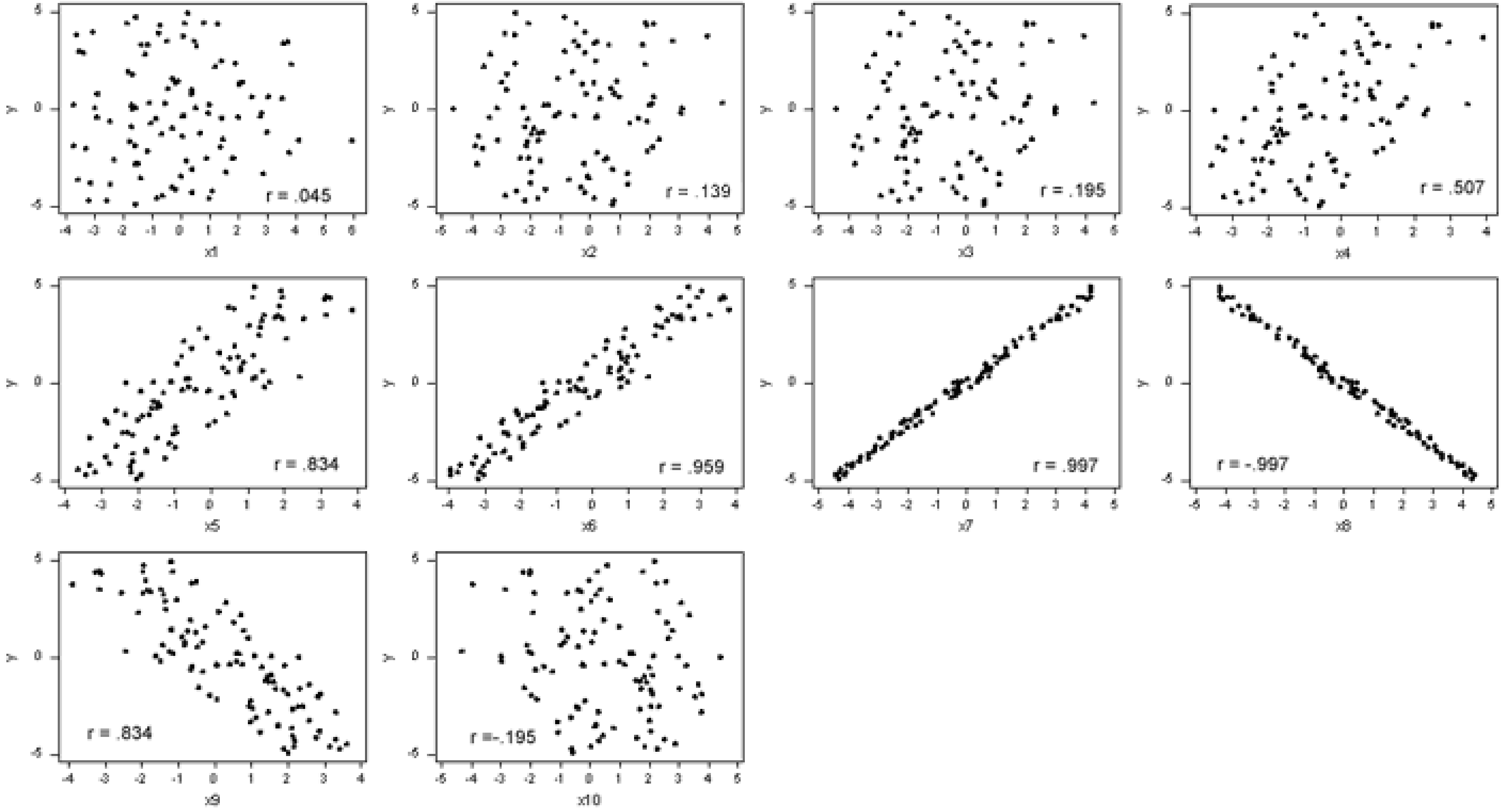
The coefficient of correlation

- can take positive or negative values.
- It can take only values between –1 and +1.
- answers the question: How **strong** is the association between X and Y?

# Coefficient of Correlation…

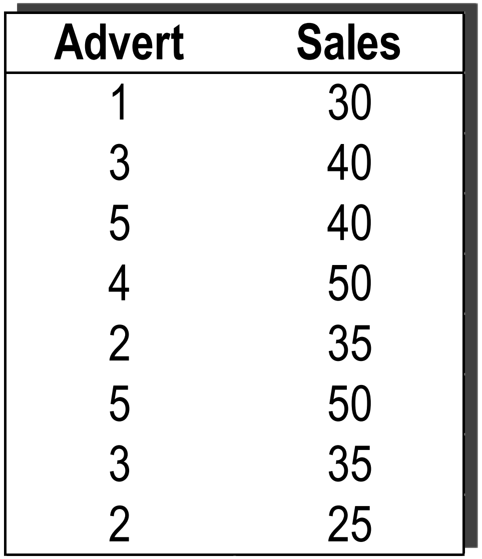
**Figure 5.6** Covariance and correlation for various scatter diagrams




y
5
0
-5
r = .045
-4 -3 -2 -1 0 1 2 3 4 5 6
x1
r = .139
-5 -4 -3 -2 -1 0 1 2 3 4 5
x2
r = .195
x3
r = .507
-4 -3 -2 -1 0 1 2 3 4
x4
r = .834
x5
r = .959
x6
r = .997
x7
r = -.997
x8
r = .834
x9
r =-.195
x10

# Example 1

Compute the covariance and the coefficient of correlation between advertising expenditure and sales level and discuss the strength and direction of the relationship between them. Base your calculation on the data (in millions) provided below.



| Advert | Sales |
|---|---|
| 1 | 30 |
| 3 | 40 |
| 5 | 40 |
| 4 | 50 |
| 2 | 35 |
| 5 | 50 |
| 3 | 35 |
| 2 | 25 |

# Example 1: Solution

Use the short-cut formulae below to obtain the required covariance and the coefficient of correlation.

$$\boxed{\operatorname{cov}(X,Y)=\frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(x_i-\overline{x})(y_i-\overline{y})}{n-1}=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}\right]}$$

$$\boxed{r=\frac{\operatorname{cov}(X,Y)}{s_xs_y}}$$

$$s_x^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right]$$

| Month | x | y | xy | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| **1** | 1 | 30 | 30 | 1 | 900 |
| **2** | 3 | 40 | 120 | 9 | 1600 |
| **3** | 5 | 40 | 200 | 25 | 1600 |
| **4** | 4 | 50 | 200 | 16 | 2500 |
| **5** | 2 | 35 | 70 | 4 | 1225 |
| **6** | 5 | 50 | 250 | 25 | 2500 |
| **7** | 3 | 35 | 105 | 9 | 1225 |
| **8** | 2 | 25 | 50 | 4 | 625 |
| **Sum** | **25** | **305** | **1025** | **93** | **12175** |

$$\operatorname{cov}(X,Y) = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

$$= \frac{1}{7}\left[1025 - \frac{25\times 305}{8}\right] = 10.268$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

$$= \frac{1}{7}\left[93 - \frac{25^2}{8}\right] = 2.125$$

$$s_x = \sqrt{2.125} = 1.458$$

Similarly, $s_y$ = 8.839

$$r = \frac{\operatorname{cov}(X,Y)}{s_x s_y} = \frac{10.268}{1.458 \times 8.839} = .797$$

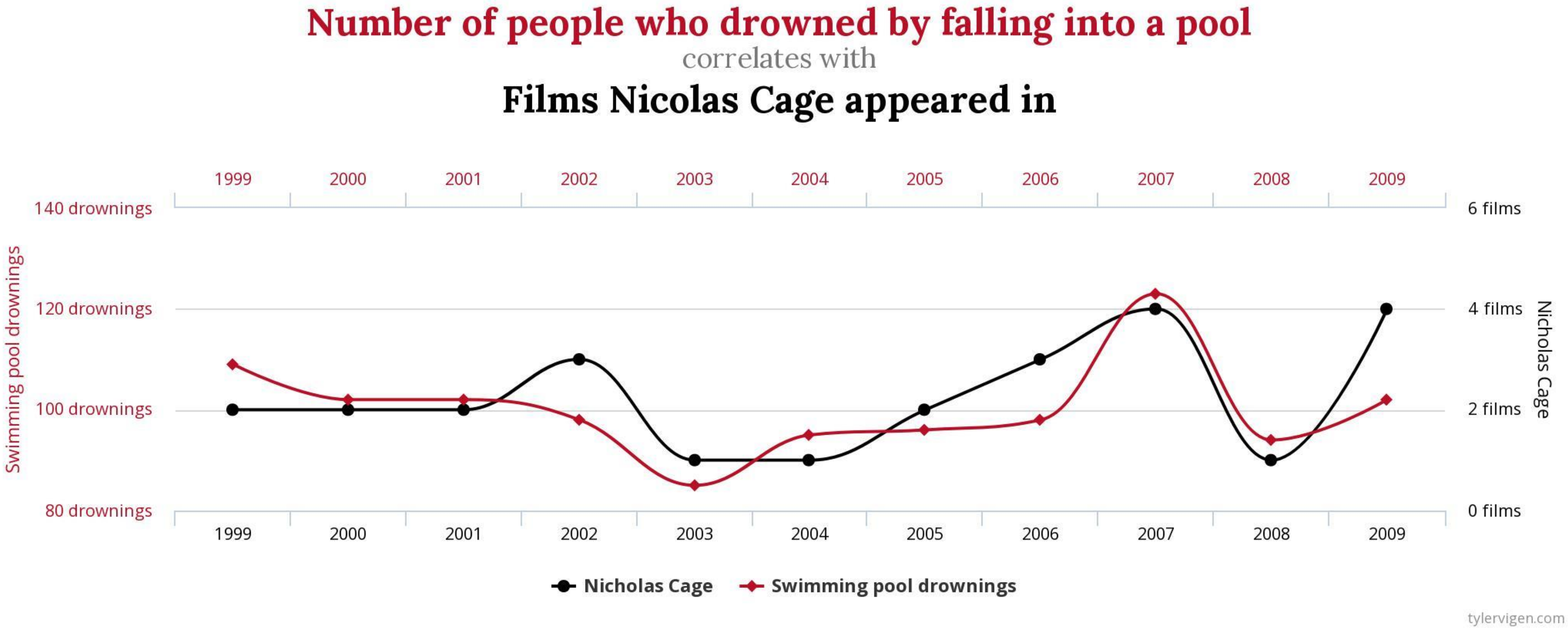
# Interpreting Correlation

if two variables X and Y are linearly related, it does not mean that X is causing Y.

It may mean that another variable is causing both X and Y or that Y is causing X. Remember

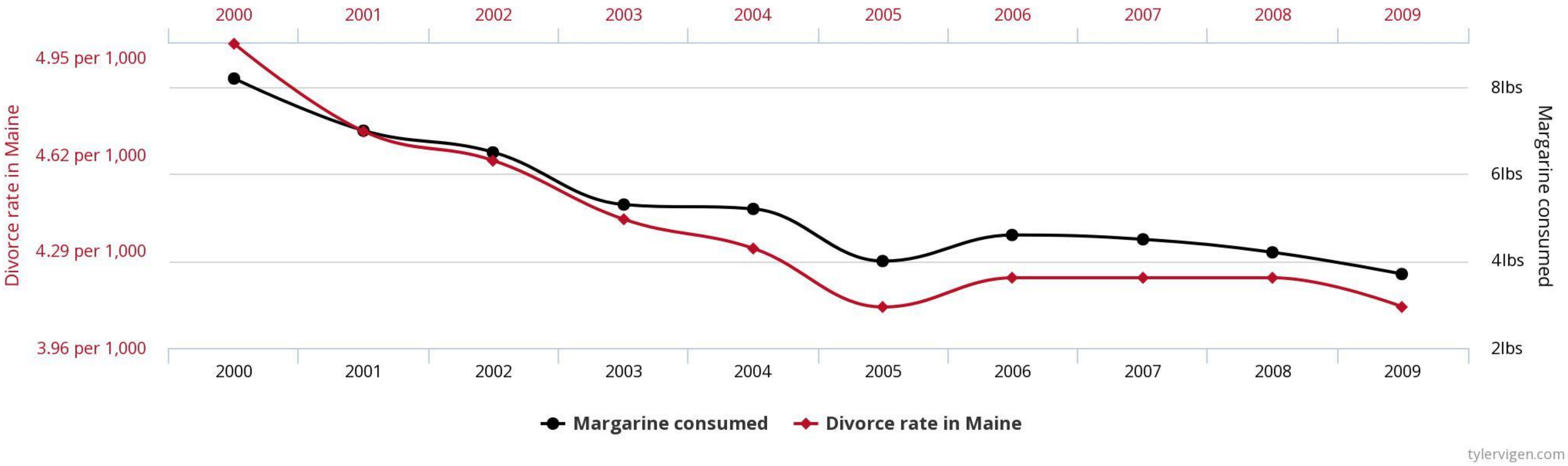
**Correlation is not Causation**

Correlations are everywhere, causal relations are much harder to establish! Examples…

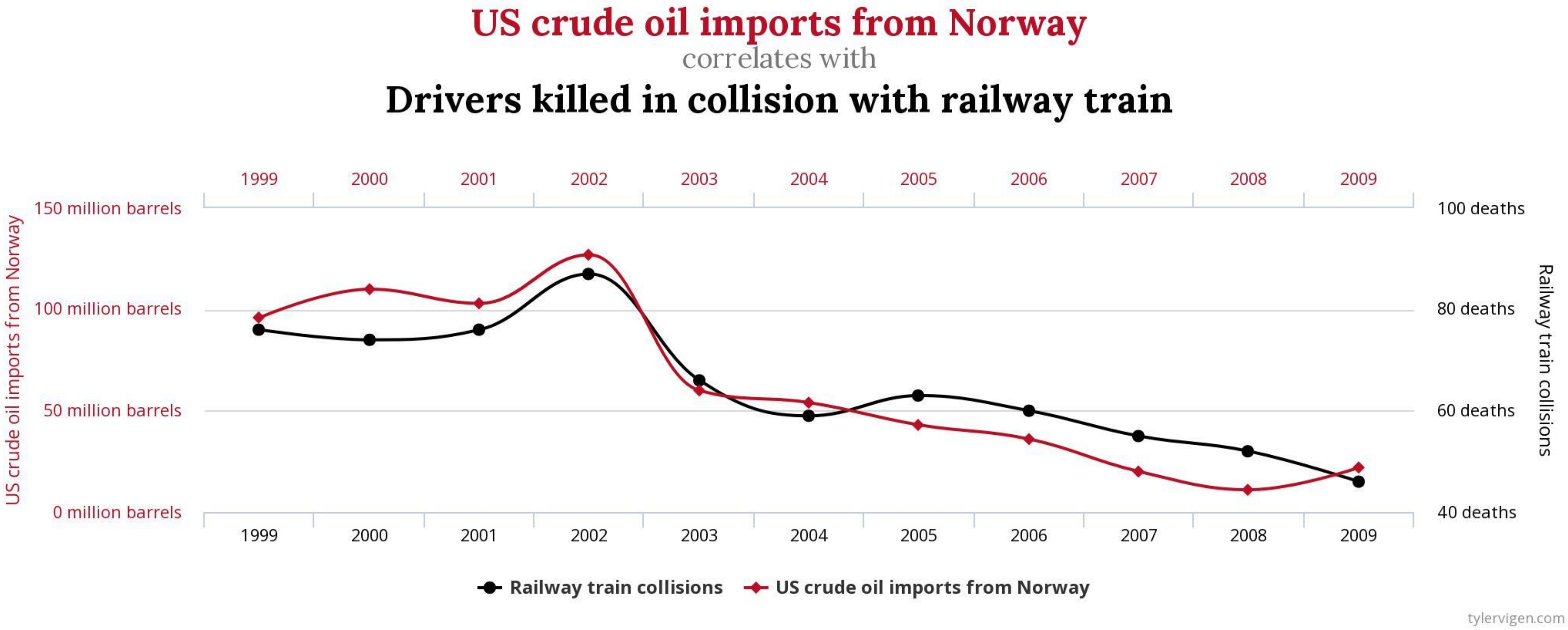


Correlation coefficient = 0.67

**Divorce rate in Maine**
correlates with
**Per capita consumption of margarine**



Correlation coefficient = 0.993



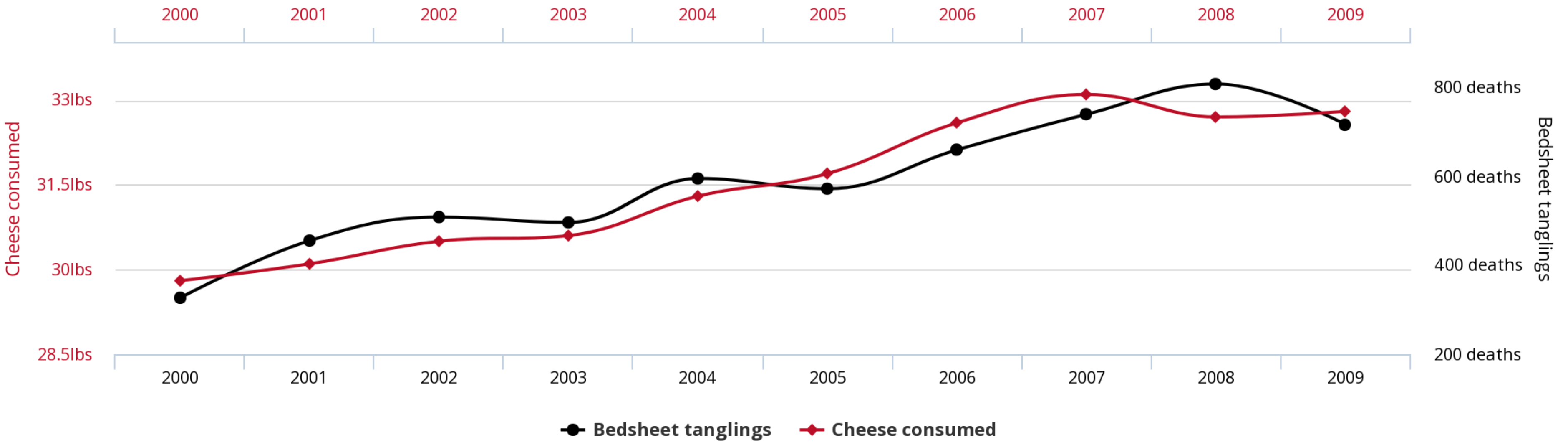
US crude oil imports from Norway
correlates with
Drivers killed in collision with railway train

Correlation coefficient = 0.955

## Per capita cheese consumption

correlates with

## Number of people who died by becoming tangled in their bedsheets



tylervigen.com

Correlation coefficient = 0.947

# Causality

These are examples of spurious correlations: two variables move together but neither of the two is “causing” the other

Uncovering causality requires an experimental design, which falls in the domain of **causal inference**

# Experimental design

CLINICAL TRIAL DESIGN
ELIGIBLE PATIENTS
RANDOMIZATION
TRIAL DRUG 1
TRIAL DRUG 6
PLACEBO
TRIAL DRUG 2
TRIAL DRUG 5
UG 3
C. Lynn

# Introduction to regression analysis

When the problem objective is to ***analyse the relationship*** between numerical variables, ***correlation*** and ***regression analysis*** is the first tool we employ.

Regression analysis is used to predict the value of one variable (the ***dependent variable***) on the basis of other variables (the ***independent variables***).

Dependent variable: denoted Y

Independent variables: denoted $X_1$, $X_2$, …, $X_k$

# Correlation analysis

If we are interested ***only*** in determining whether a relationship ***exists***, we employ ***correlation analysis***.

Chapter 17 examines the relationship between ***two variables***, sometimes called ***simple linear regression***.

We learn how to estimate such a relationship, measure the strength and make inferences on the relationship.

Mathematical equations describing these relationships are also called ***models***, and they fall into two types: deterministic or probabilistic.

# Model types

**Deterministic Model:** an equation or set of equations that allow us to ***fully determine*** the value of the dependent variable from the values of the independent variables.

Contrast this with…

**Probabilistic Model:** a method used to capture the ***randomness*** that is part of a real-life process.

E.g. do all houses of the same size (measured in square metre) sell for exactly the same price?

# A model

To create a probabilistic model, we start with a deterministic model that ***approximates the relationship*** we want to model and add a ***random term*** that measures the ***error*** of the deterministic component.

## Deterministic Model:

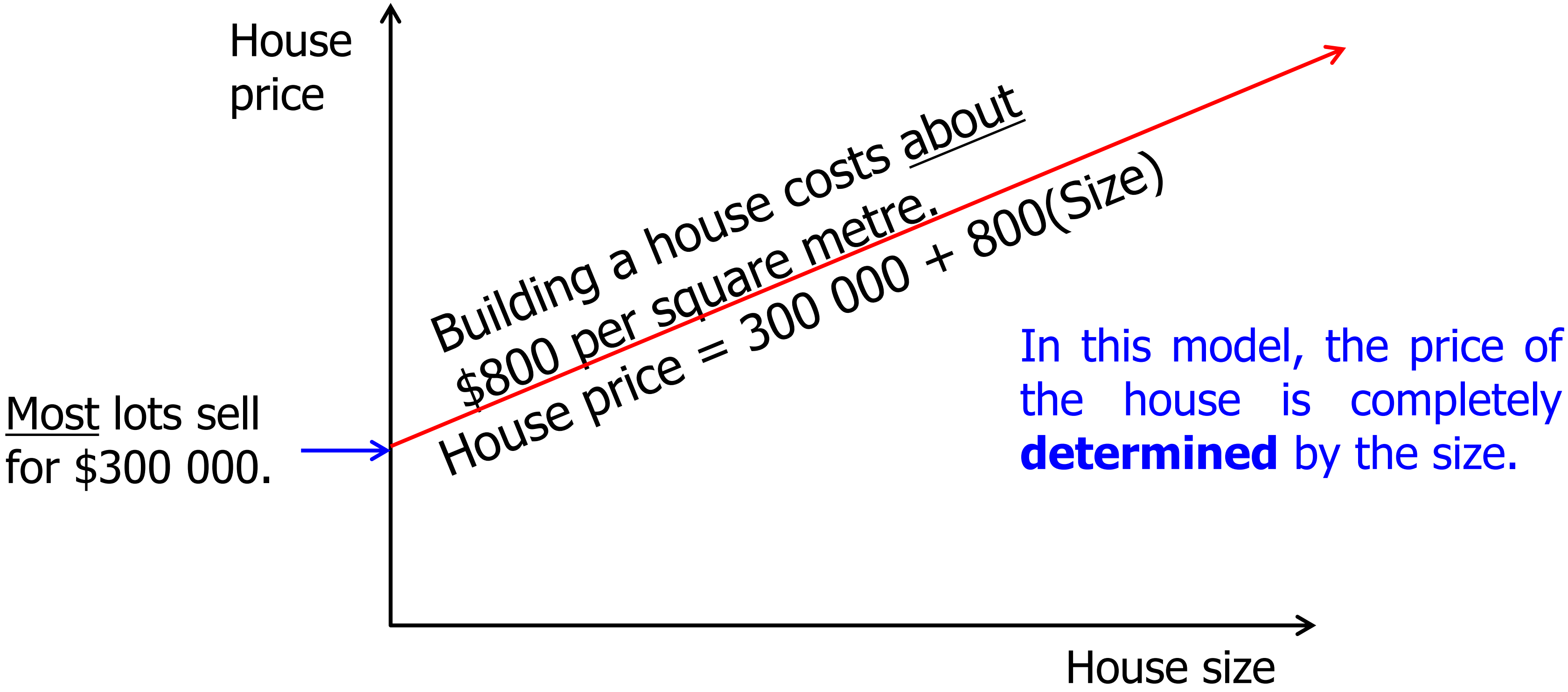
The cost of building a new house is about \$800 per square metre and most lots sell for about \$300 000. Hence the approximate selling price (**y**) would be:

$$y = \$300\,000 + \$800(x)$$

(where **x** is the size of the house in square metres)

# A model…

A model of the relationship between house size (independent variable) and house price (dependent variable) would be:



Most lots sell for $300 000.

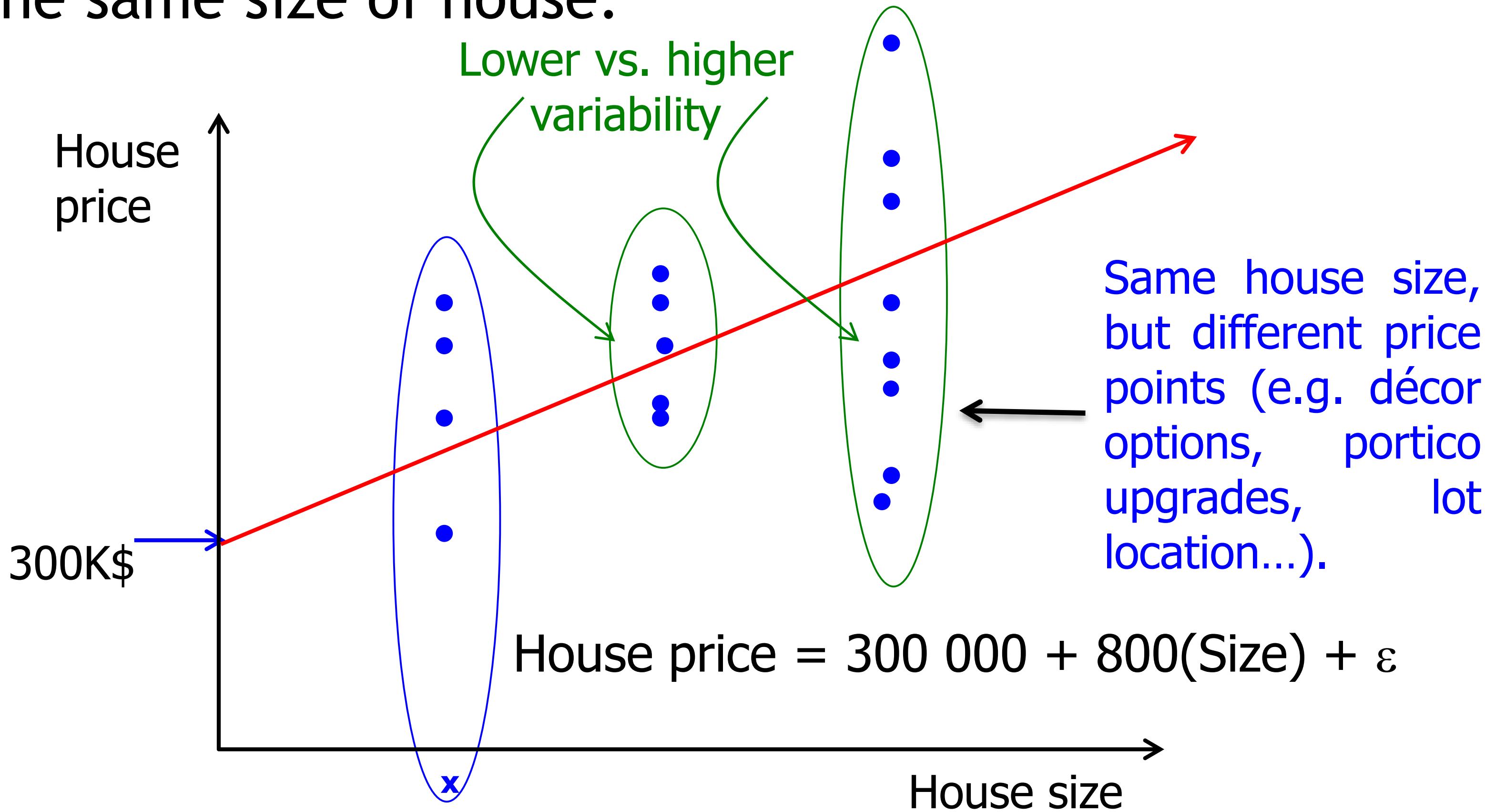
Building a house costs about $800 per square metre.

House price = 300 000 + 800(Size)

In this model, the price of the house is completely **determined** by the size.

# A model…

In real life, however, the house cost will vary even among the same size of house:



Same house size, but different price points (e.g. décor options, portico upgrades, lot location…).

House price = 300 000 + 800(Size) + $\varepsilon$

# Random term

We now represent the price of a house as a function of its size in this probabilistic model:

$$y = 300\,000 + 800x + \varepsilon$$

where $\varepsilon$ (Greek letter epsilon) is the ***random term*** (also known as ***error variable***). It is the difference between the ***actual*** selling price and the ***estimated*** price based on the size of the house. Its value will vary from house sale to house sale, even if the area of the house (i.e. **x**) remains the same due to other factors such as the location, age, décor etc of the house.

# 17.1 Model

A straight line model with one independent variable is called a ***first order linear model*** or a ***simple linear regression model***. It is written as:

dependent variable

independent variable

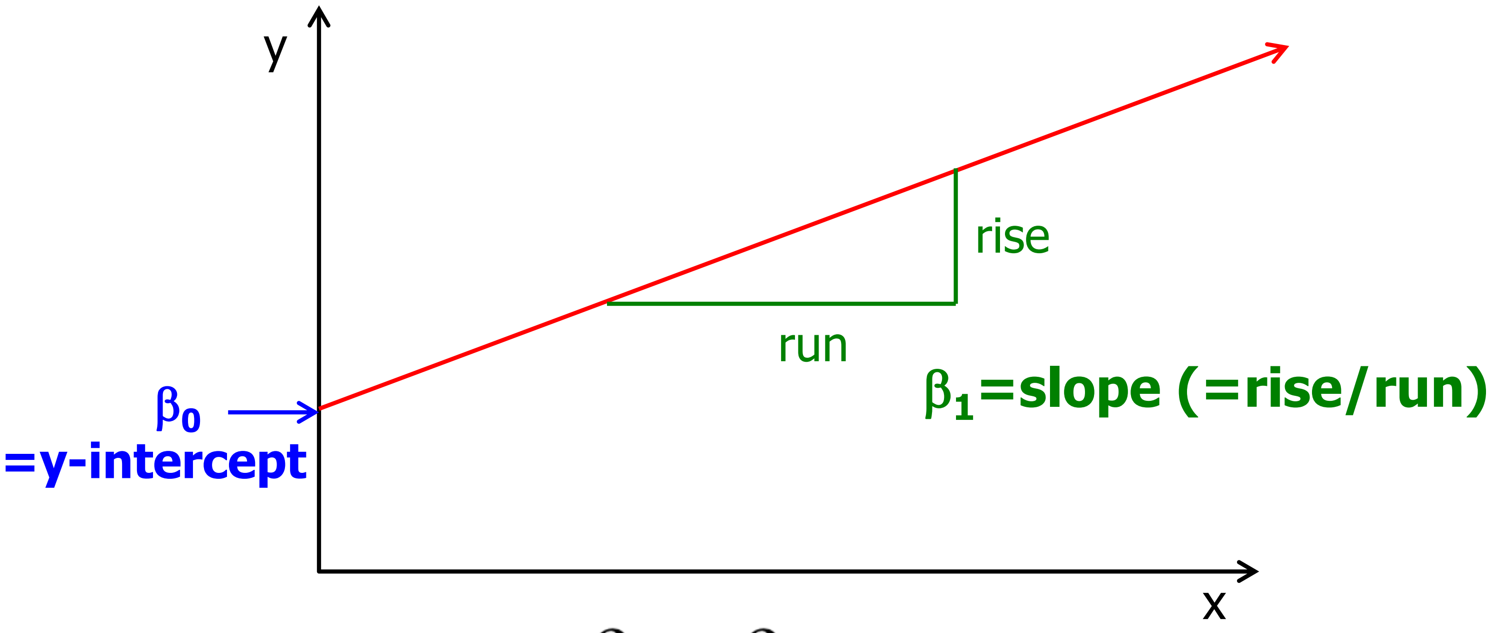
$$\boxed{y} = \boxed{\beta_0} + \boxed{\beta_1}\boxed{x} + \boxed{\varepsilon}$$

y-intercept

slope of the line

error variable

# Simple linear regression model…



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

Note that both $\beta_0$ and $\beta_1$ are **population parameters** which are usually unknown, and hence *estimated* from the data.

# 17.2 Estimating the coefficients

In much the same way we base estimates of $\mu$ on $\bar{X}$, we estimate $\beta_0$ using $\hat{\beta}_0$ and $\beta_1$ using $\hat{\beta}_1$, the y-intercept and slope (respectively) of the ***least squares*** or ***regression line*** given by:

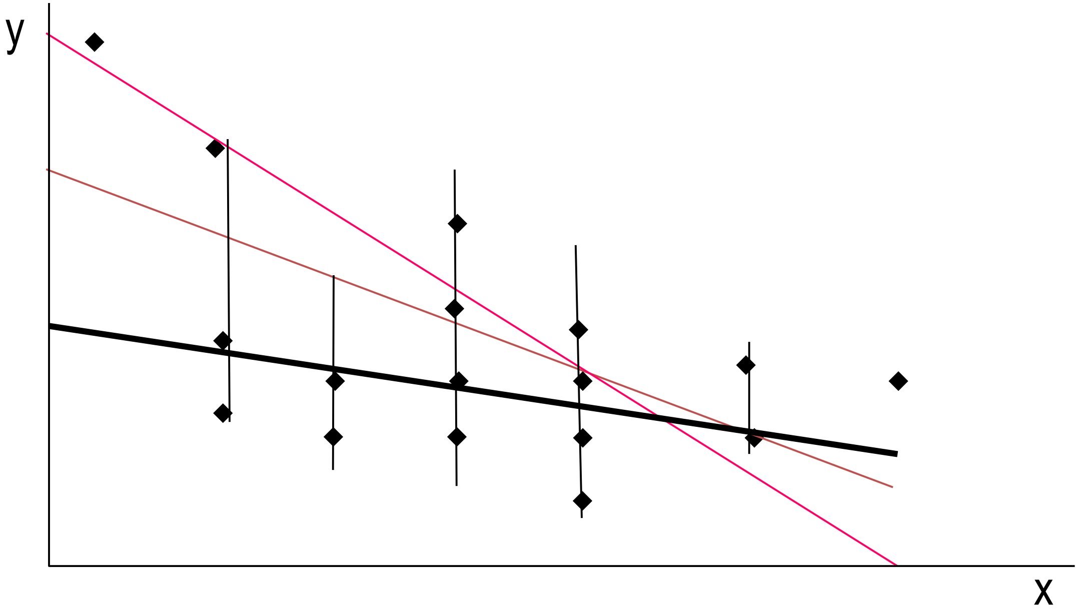
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

(Recall: this is an application of the least squares method and it produces a straight line that ***minimises*** the sum of the squared differences between the points and the line)

# Least squares method

The question is:

- Which straight line fits best?
- The least squares line minimises the sum of squared differences between the points and the line.

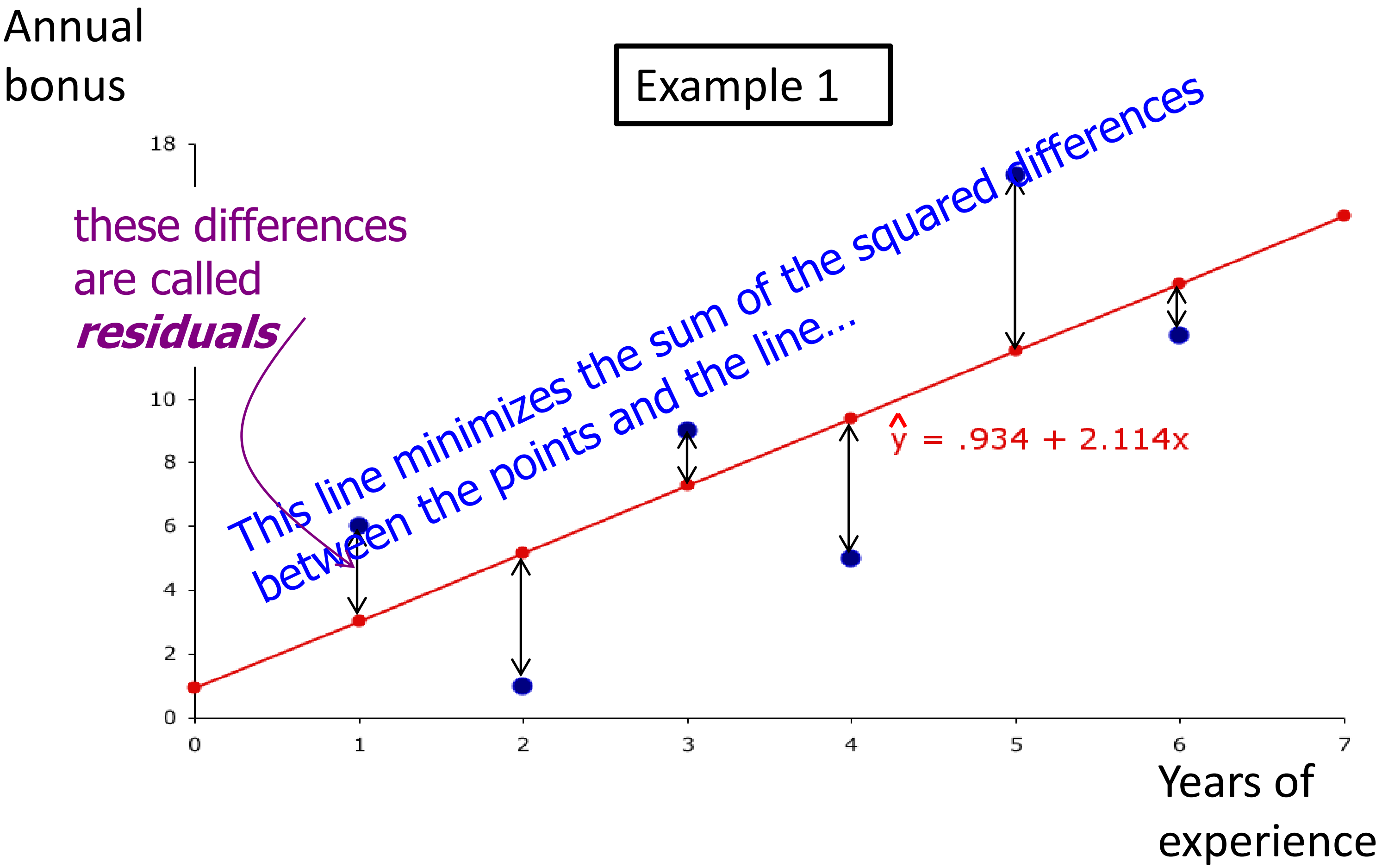


# Example 1

The annual bonuses ($1,000s) of six employees with different years of experience were recorded as follows. We wish to determine the straight line relationship between annual bonus and years of experience.

| Years of experience | x | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Annual bonus | y | 6 | 1 | 9 | 5 | 17 | 12 |

# Least Squares Line



Example 1

these differences are called ***residuals***

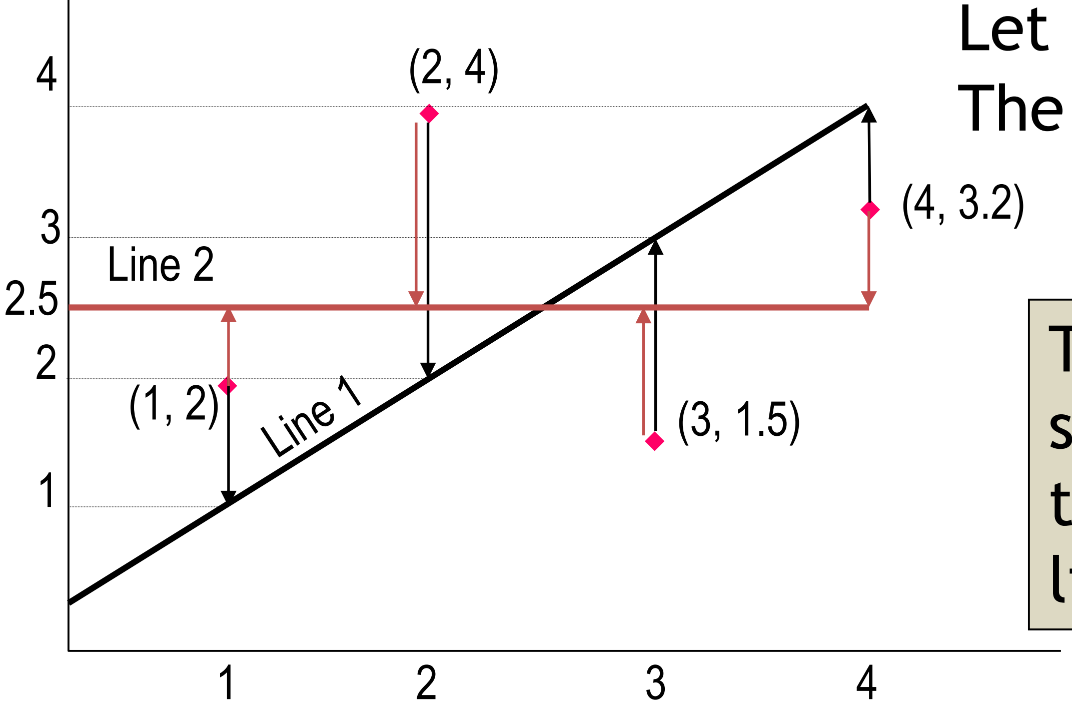
This line minimizes the sum of the squared differences between the points and the line…

# Example 2: Which line fits best?

The best line is the one that minimises the sum of squared vertical differences between the points and the line.

Line 1: Sum of squared differences $= (2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Line 2: Sum of squared differences $= (2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 -- 2.5)^2 = 3.99$



Let us compare two lines.
The second line is horizontal.

The smaller the sum of squared differences, the better the fit of the line to the data.

# Least squares estimates…

Then

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The estimated simple linear regression equation that estimates the equation of the first-order linear model is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

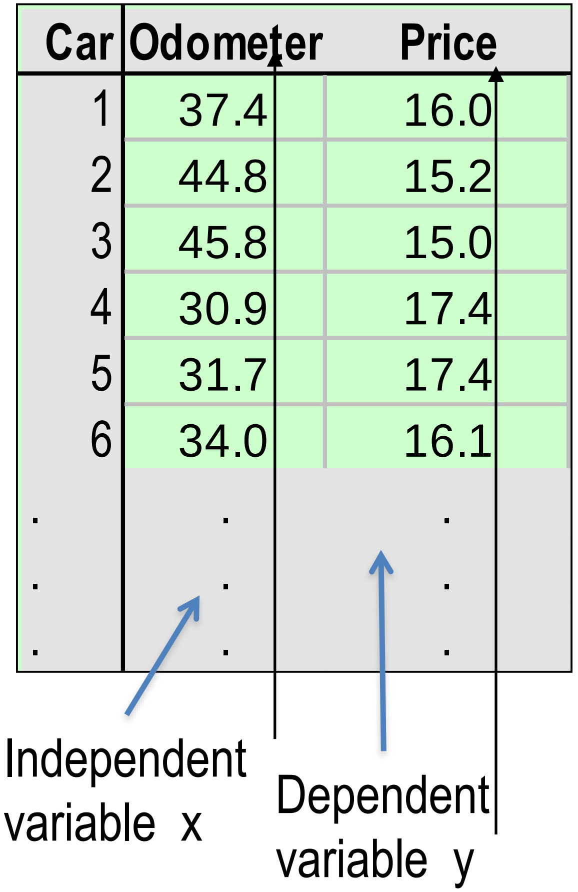
# Example 3 – Odometer readings and prices of used cars

*(Example 17.3, p717)*

A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.

A random sample of 100 cars is selected and the data are recorded.

Estimate a linear relationship between price and odometer reading..



| Car | Odometer | Price |
|---|---|---|
| 1 | 37.4 | 16.0 |
| 2 | 44.8 | 15.2 |
| 3 | 45.8 | 15.0 |
| 4 | 30.9 | 17.4 |
| 5 | 31.7 | 17.4 |
| 6 | 34.0 | 16.1 |
| . | . | . |
| . | . | . |
| . | . | . |

Independent variable x

Dependent variable y

# Example 3 - Solution

To calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to calculate several statistics first:

$$n = 100; \quad \sum x = 3601.1; \quad \sum y = 1623.7; \quad \bar{x} = 36.01; \quad \bar{y} = 16.24;$$

$$\sum x^2 = 133986.6; \quad \sum y^2 = 26421.9; \quad \sum xy = 58067.4$$

$$s_X^2 = \frac{1}{(n-1)}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) = \frac{4\,307.378}{99} = 43.509$$

$$s_{xy} = \frac{1}{(n-1)}\left(\sum x_i y_i - \frac{(\sum x_i \sum y_i)}{n}\right) = \frac{\text{-}403.6207}{99} = -4.077$$

# Example 3 – Solution…

Therefore,

$$\hat{\beta}_1 = \frac{s_{xy}}{s_X^2} = \frac{\text{-}4.077}{4\ 3.509} = \text{-}0.0937$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 16.24 - (-0.0937)(36.01) = 19.611$$

The estimated least squares regression line is

$$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 19.611 - 0.094x}$$

# Testing the slope

If no linear relationship exists between the two variables, we would expect the regression line to be ***horizontal***, that is, to have a ***slope of zero***.

We want to see if there is a linear relationship, i.e. we want to see if the slope ($\beta_1$) is something other than zero. Our research hypothesis becomes:

$H_A$: $\beta_1 \neq 0$ (linear relationship exists)

Thus the null hypothesis becomes:

$H_0$: $\beta_1 = 0$ (no linear relationship exists)

# Testing the slope…

We can use the following test statistic to test our hypotheses:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

where $s_{\hat{\beta}_1}$ is the standard deviation of $\hat{\beta}_1$, defined as:

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

- If the error var [illegible]rmally distributed, the test statistic has a Student *t*-distribution with *n*–2 degrees of freedom.
- The rejection region depends on whether or not we're doing a one- or two- tail test (two-tail test is most typical).

# Testing the slope…

If we wish to test for ***positive*** or ***negative*** linear relationships we conduct one-tail tests, i.e. our research alternate hypotheses become:

$H_A$: $\beta_1 < 0$ (testing for a negative slope)

or

$H_A$: $\beta_1 > 0$ (testing for a positive slope)

Of course, the null hypothesis remains: $H_0$: $\beta_1 = 0$.

# Example 3…

Test to determine whether there is enough evidence to infer that a linear relationship exists between the price and the odometer reading at the 5% significance level.

# Example 3 – Solution…

We want to test the hypothesis

$H_0$: $\beta_1 = 0$ (no linear relationship)

$H_A$: $\beta_1 \neq 0$ (a linear relationship exists)

If the null hypothesis is rejected, we conclude that there is a significant linear relationship between price and odometer reading.

Test statistic:

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

has a t-distribution with 98 (=100–2) degrees of freedom.

Level of significance $\alpha = 0.05$.

# Example 3 – Solution…

Decision rule: Reject $H_0$ if $|t| > t_{0.025,98} = 1.984$,
or reject $H_0$ if *p*-value < $\alpha$.

Value of the test statistic:

To compute *t* we need the values of $\hat{\beta}_1$ and $s_{\hat{\beta}_1}$.

$$\hat{\beta}_1 = -0.0937$$

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x{}^2}} = \frac{0.4526}{\sqrt{99(43.509)}} = 0.0069$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{-0.0937 - 0}{.0069} = -13.59$$

Conclusion: Comparing the decision rule with the calculated t-value (=-13.59), we reject Ho and conclude that the odometer readings do affect the sale price.