

Regression analysis

Two variables

(Montgomery and Runger: ch 11

Brani Vidakovic: ch 14)

Reminder

Covariance Defined

Covariance is a number quantifying average dependence between two random variables.

The covariance between the random variables X and Y , denoted as $\text{cov}(X, Y)$ or σ_{XY} is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y \quad (5-14)$$

The units of σ_{XY} are units of X times units of Y .

Unlike the range of variance, $-\infty < \sigma_{XY} < \infty$.

Correlation is “normalized covariance”

- Also called:
Pearson correlation coefficient

$\rho_{XY} = \sigma_{XY} / \sigma_X \sigma_Y$
is the covariance
normalized to
be $-1 \leq \rho_{XY} \leq 1$



Karl Pearson (1852– 1936)
English mathematician and biostatistician

Covariance and Scatter Patterns

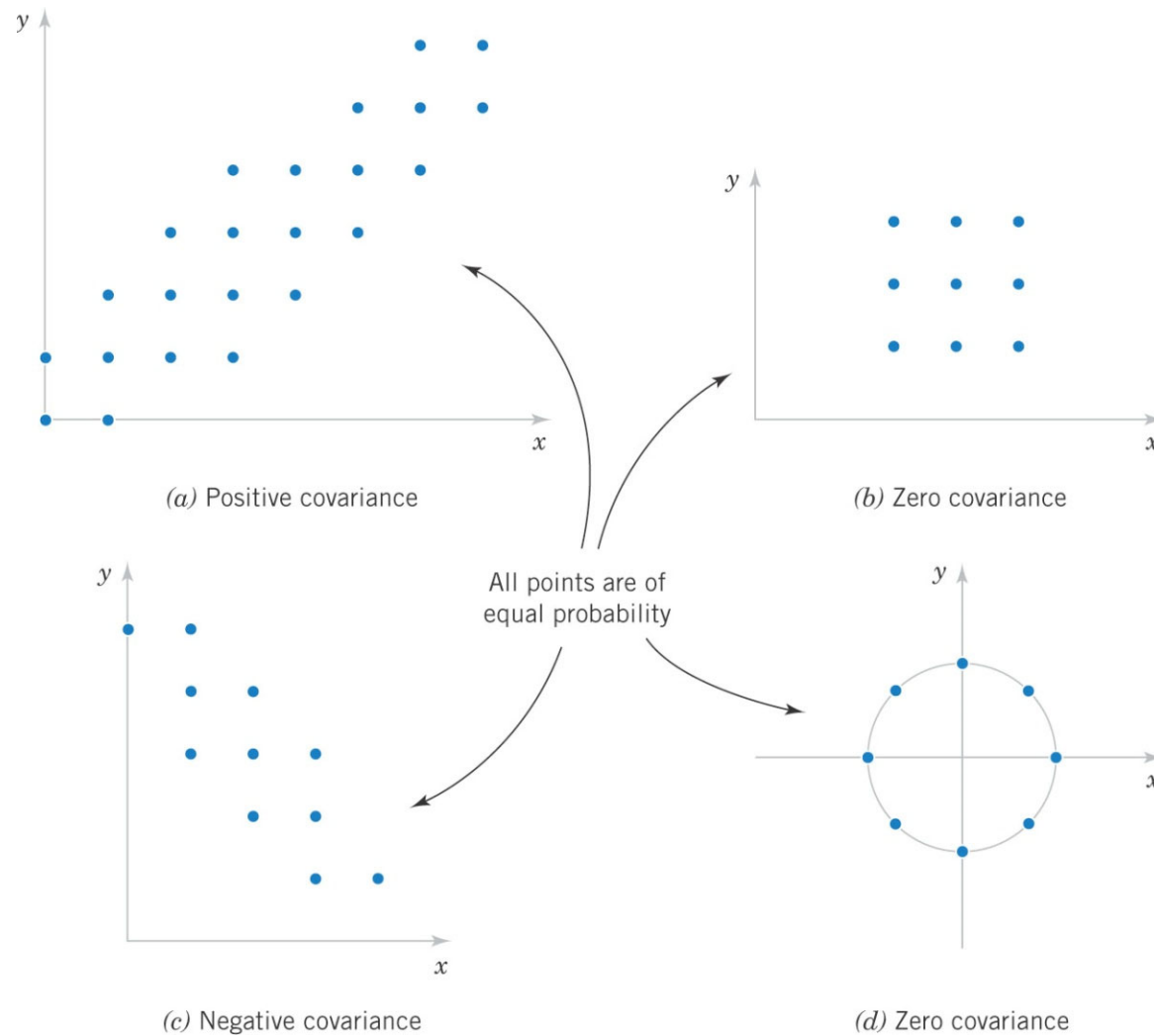
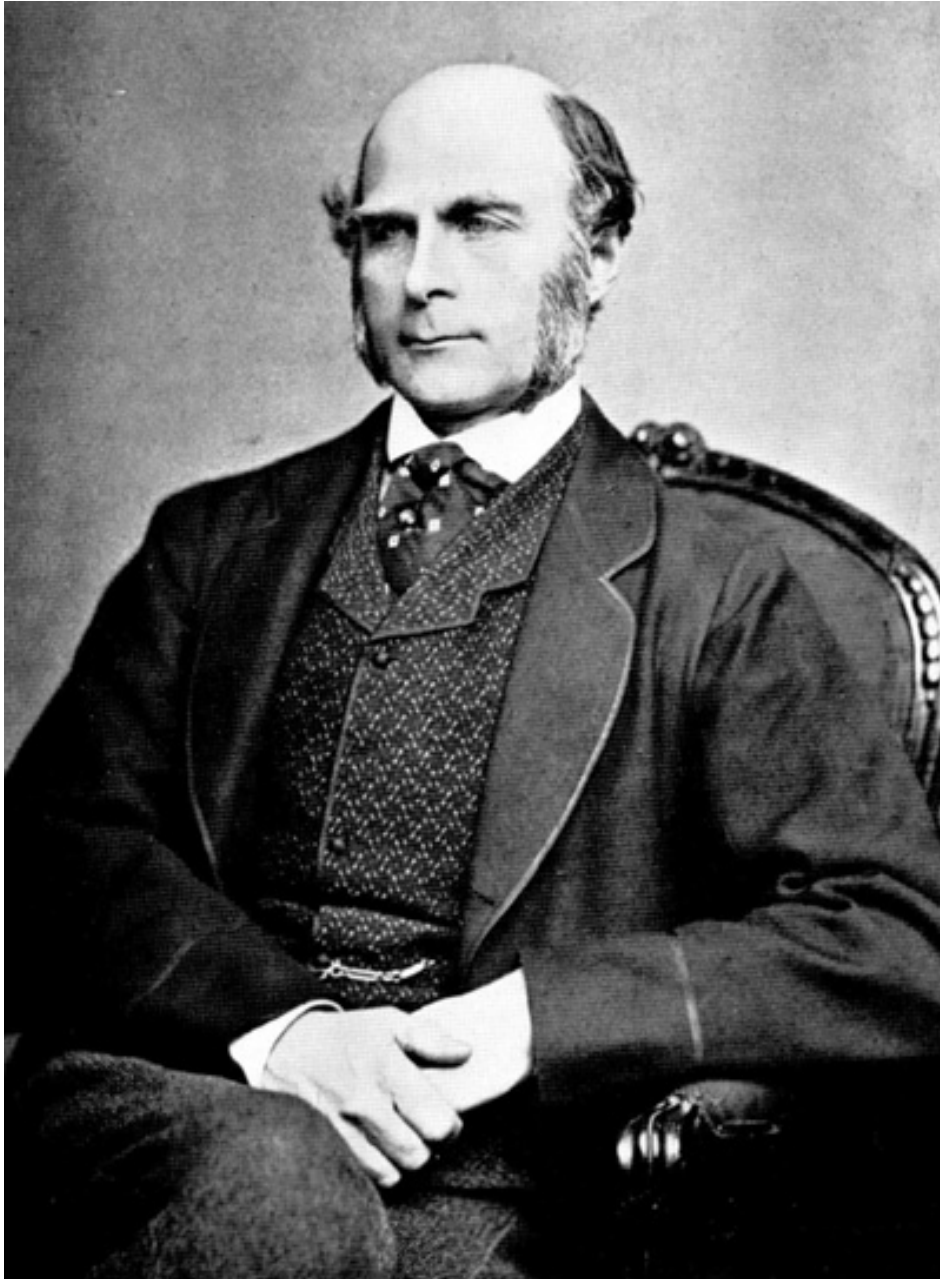


Figure 5-13 Joint probability distributions and the sign of $\text{cov}(X, Y)$. Note that covariance is a measure of linear relationship. Variables with non-zero covariance are **correlated**.

Regression analysis

- Many problems in engineering and science involve sample in which two or more variables were measured. They may not be independent from each other and one (or several) of them can be used to predict another
- Everyday example: in most samples height and weight of people are related to each other
- Biological example: in a cell sorting experiment the copy number of a protein may be measured alongside its volume
- **Regression analysis** uses a sample to build a model to predict protein copy number given a cell volume



Sir Francis Galton, (1822 -1911) was an English **statistician**, anthropologist, proto-geneticist, psychometrician, **eugenicist**, (“Nature vs Nurture”, inheritance of intelligence), tropical explorer, geographer, inventor (Galton Whistle to test hearing), meteorologist (weather map, anticyclone).

Invented both **correlation** and **regression analysis** when studied **heights of fathers and sons**

Found that fathers with height above average tend to have sons with height also above average but closer to the average.
Hence **“regression” to the mean**

Two variable samples

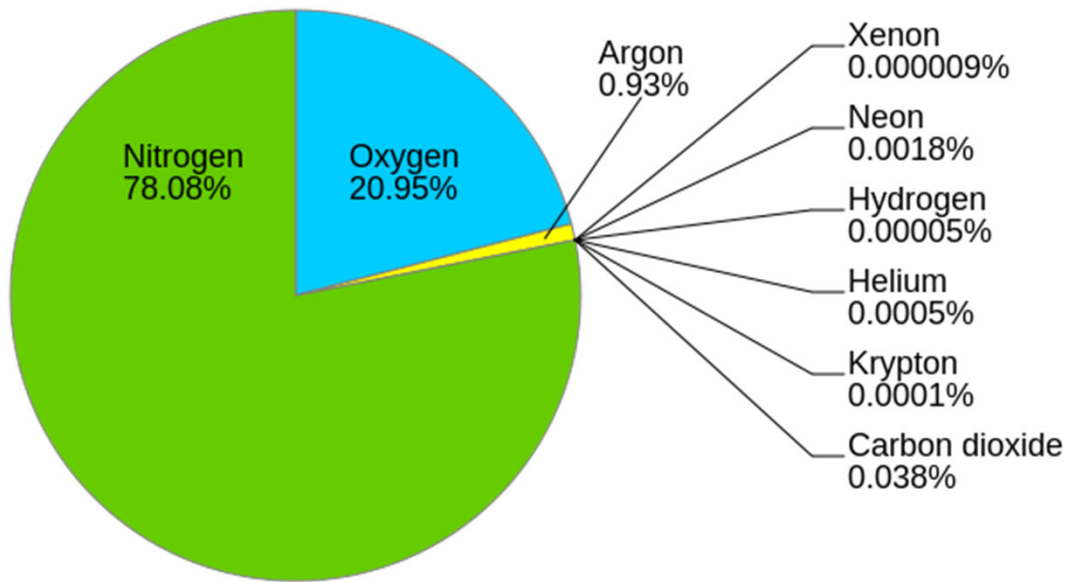


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go **kaboom!!!**
- When more hydrocarbons were removed, the remaining oxygen stays cleaner
- Except we don't know how dirty was the air to begin with

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

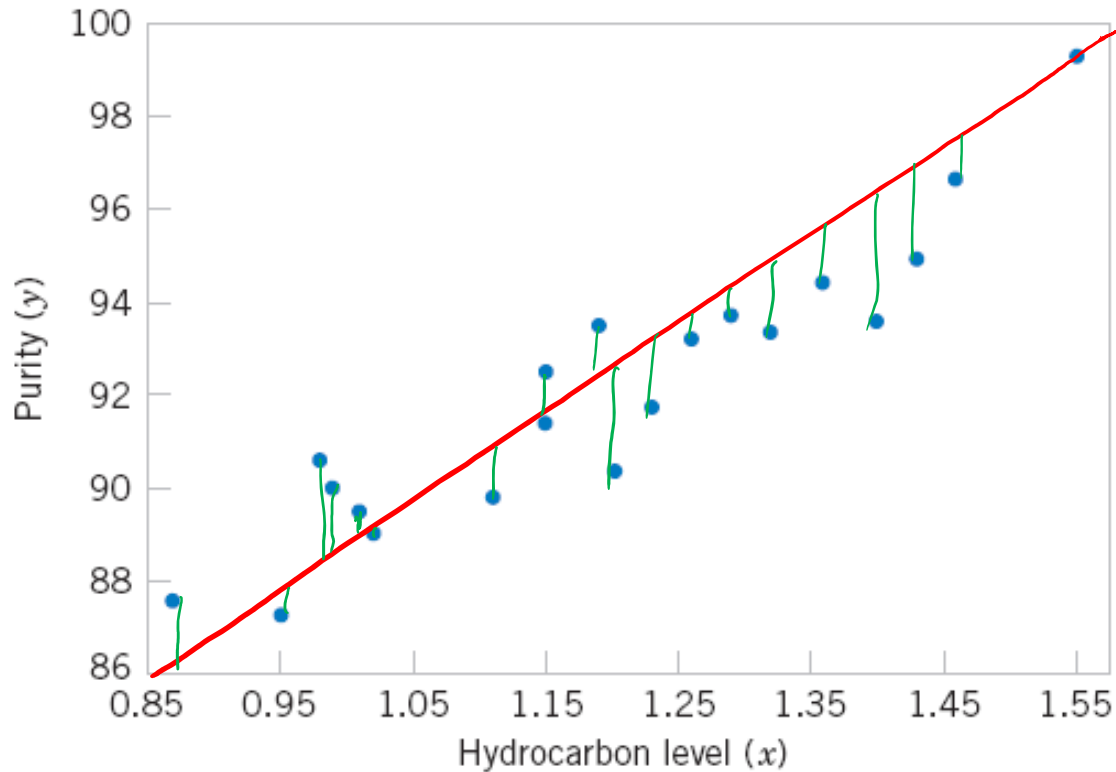


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

$$Y = 75 + 15 \cdot X + \epsilon$$

Linear regression

The **simple linear regression model** is given by

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

ε is the **random error term**

slope β_1 and intercept β_0 of the line are called **regression coefficients**

Note: Y , X and ε are random variables

The minimal assumption: $E(\varepsilon | x) = 0 \rightarrow$

$$E(Y | x) = \beta_0 + \beta_1 x + E(\varepsilon | x) = \beta_0 + \beta_1 x$$

$$Y = \beta_0 + \beta_1 X + \epsilon ; \quad E(\epsilon | x) = 0 \quad \forall x$$

How does one find β_0 & β_1 ?

$$\begin{aligned} \text{Cov}(Y, X) &= \text{Cov}(\beta_0 + \beta_1 X + \epsilon, X) = \\ &= \cancel{\text{Cov}(\beta_0, X)} + \beta_1 \text{Cov}(X, X) + \cancel{\text{Cov}(\epsilon, X)} \end{aligned}$$

$\text{Cov}(\beta_0, X) = 0$ since β_0 is constant

$$\text{Cov}(X, X) = E(X^2) - E(X)^2 = \text{Var}(X)$$

$$\text{Cov}(\epsilon, X) = E(\epsilon \cdot X) - \cancel{E(\epsilon)} \cdot \cancel{E(X)} =$$

$$= E(\epsilon \cdot X) = \sum_{\text{all } x} x \cdot \cancel{E(\epsilon | x)} = 0$$

Thus

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$\beta_0 = E(Y) - \beta_1 E(X)$$

Method of least squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

Figure 11-3 Deviations of the data from the estimated regression model.

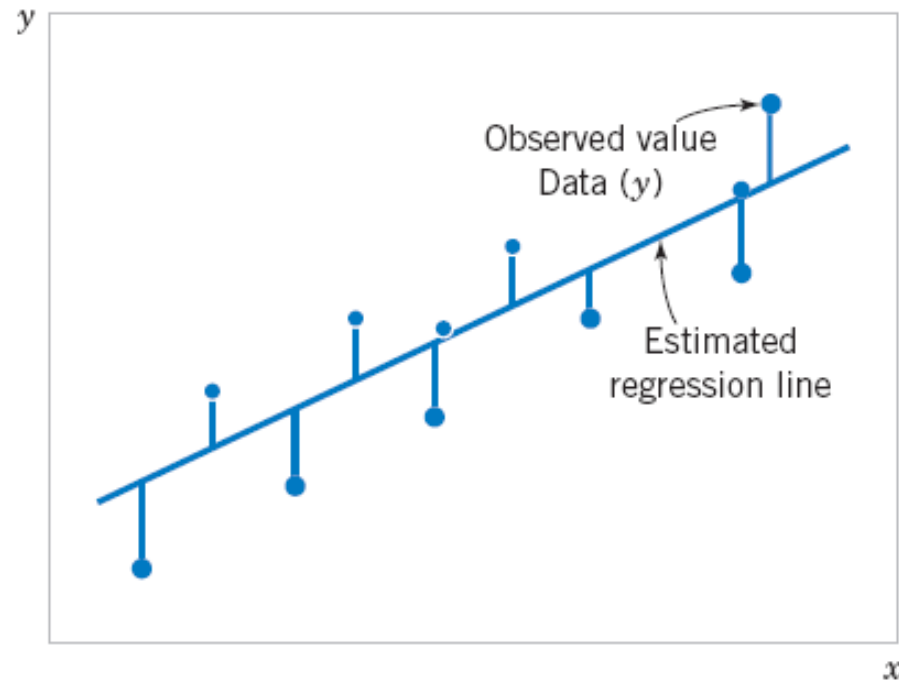


Figure 11-3 Deviations of the data from the estimated regression model.

Traditional notation

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

11-2: Simple Linear Regression

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The **analysis of variance** identity is

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11-24)$$

Symbolically,

$$SS_T = SS_R + SS_E \quad (11-25)$$

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2) VERY COMMONLY USED

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \leq R^2 \leq 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

- For the oxygen purity regression model,

$$\begin{aligned}R^2 &= SS_R/SS_T \\ &= 152.13/173.38 \\ &= 0.877\end{aligned}$$

- Thus, the model accounts for 87.7% of the variability in the data.

11-2: Simple Linear Regression

Estimating σ_ε^2

An **unbiased estimator** of σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{SS_E}{n - 2} \quad (11-13)$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \quad (11-14)$$

11-3: Properties of the Least Squares Estimators

- Slope Properties

$$E(\hat{\beta}_1) = \beta_1$$

$$V(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{S_{xx}} = \frac{\hat{\sigma}_\varepsilon^2}{n \hat{\sigma}_x^2}$$

Large $n \rightarrow$ small variance of β_1

- Intercept Properties

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \hat{\sigma}_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] =$$

$$= \hat{\sigma}_\varepsilon^2 \left[1 + \frac{\mu_x^2}{\hat{\sigma}_x^2} \right] \times \frac{1}{n}$$

11-4: Hypothesis Tests in Simple Linear Regression

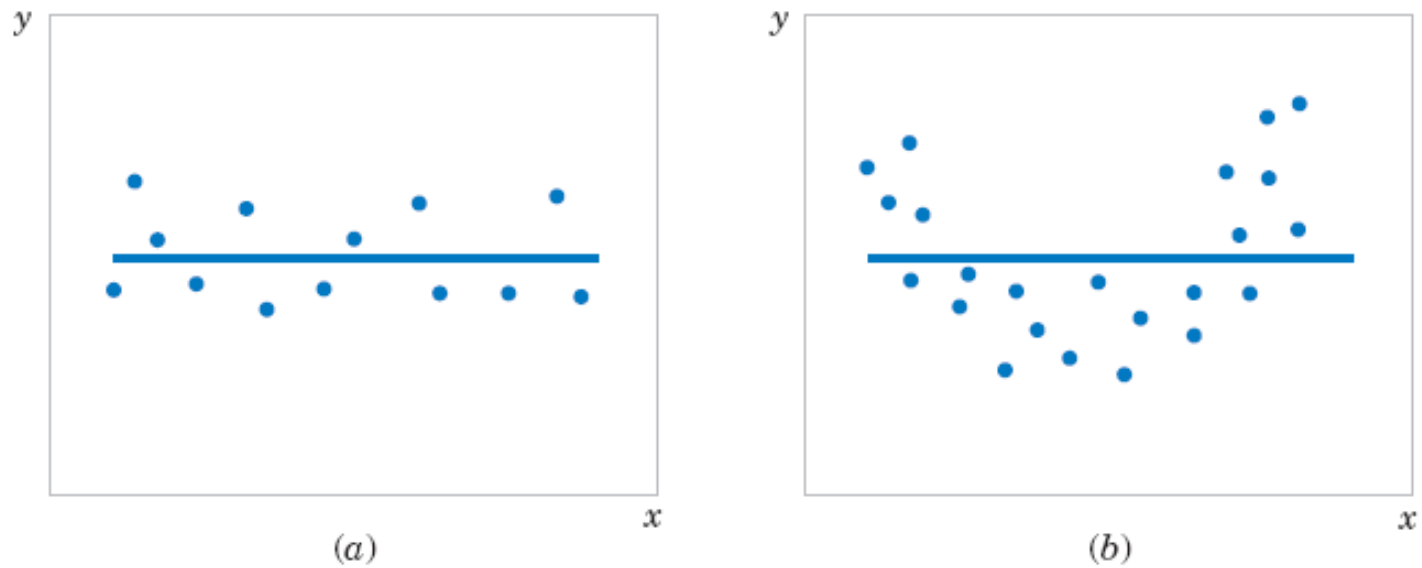


Figure 11-5 The hypothesis $H_0: \beta_1 = 0$ is not rejected.

Figure 11-5 The null hypothesis $H_0: \beta_1 = 0$ is accepted.

11-4: Hypothesis Tests in Simple Linear Regression

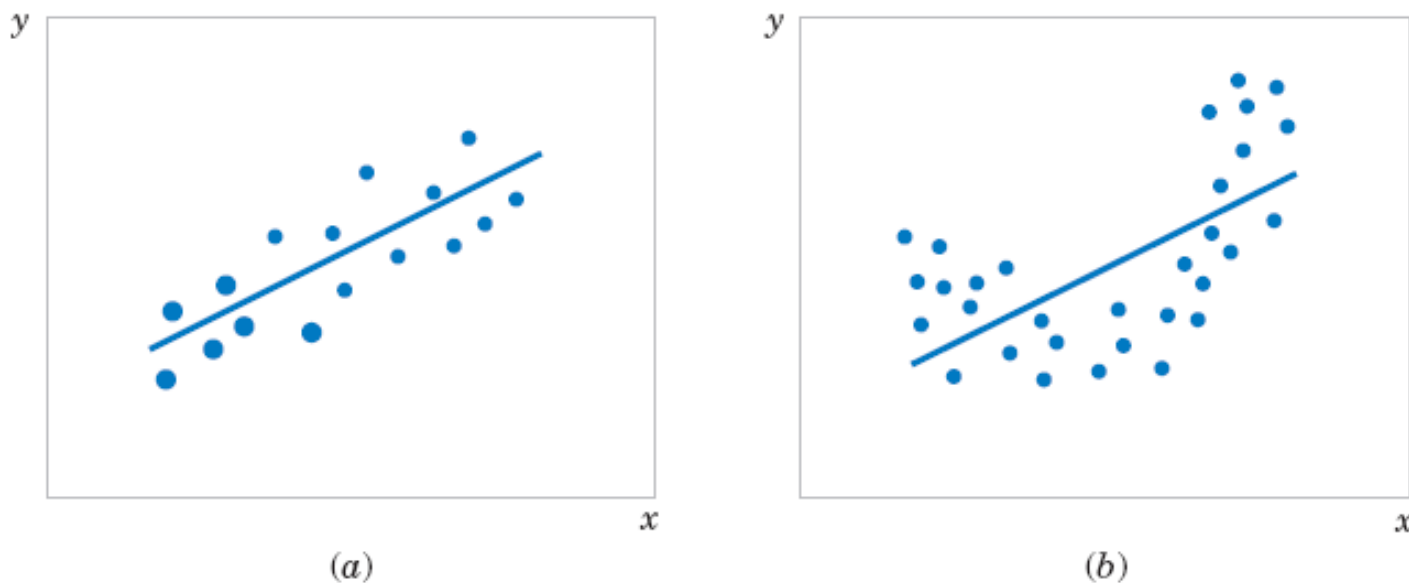


Figure 11-6 The hypothesis $H_0: \beta_1 = 0$ is rejected.

Figure 11-6 The **null hypothesis $H_0: \beta_1 = 0$ is rejected.**

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of Z-tests for large n

An important special case of the hypotheses of Equation 11-18 is

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**. *Failure to reject* H_0 is equivalent to **concluding that there is no linear relationship between X and Y** .

11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Choose α

(e.g. $\alpha = 5\%$
for 95%
confidence

in rejecting
 H_0)

$$Z = \frac{\hat{\beta}_1 - 0}{\frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_x} \cdot \frac{1}{\sqrt{n}}}$$

$$\sqrt{\text{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_x \sqrt{n}}$$

for $\alpha = 5\%$

Reject H_0 if $|Z| > Z_{\alpha/2} = 1.96$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of t -tests for smaller n .

The number of degrees of freedom in $n-2$

One can always fit a straight line through two points so one needs $n \geq 3$



11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$T = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}_e}{\sigma_x} \cdot \frac{1}{\sqrt{n}}}$$

Reject H_0 if $|T| > t_{\alpha/2, n-2}$

Choose α
(e.g. $\alpha = 5\%$
for 95%
confidence
in rejecting
 H_0)

$t_{\alpha/2, n-2}$ is such
 $1 - \frac{\alpha}{2} = \text{cdf}(t_{\alpha/2, n-2}, n-2)$

Credit: XKCD
comics

WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS

WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

QUESTIONS

FOUND IN GOOGLE AUTOCOMLETE



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

WHY AREN'T ECONOMISTS RICH

WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS THERE PHLEGM

WHY DO AMERICANS CALL IT SOCCER

WHY IS PSYCHIC WEAK TO BUG

WHY ARE MY EARS RINGING

WHY DO CHILDREN GET CANCER

WHY ARE THERE SO MANY AVENGERS

WHY IS POSEIDON ANGRY WITH ODYSSEUS

WHY ARE THE AVENGERS FIGHTING THE X MEN

WHY IS THERE ICE IN SPACE

WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL IN MY BACKYARD

WHY IS SPACE BLACK

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL OUTSIDE MY WINDOW

WHY IS OUTER SPACE SO COLD

WHY ARE THERE GHOSTS

WHY IS THERE AN OWL ON THE DOLLAR BILL

WHY ARE THERE PYRAMIDS ON THE MOON

WHY ARE THERE GHOSTS

WHY DO OWLS ATTACK PEOPLE

WHY IS NASA SHUTTING DOWN

WHY ARE THERE GHOSTS

WHY ARE AK 47s SO EXPENSIVE

WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE GHOSTS

WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE

WHY ARE THERE TINY SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY ARE THERE GODS

WHY DO SPIDERS COME INSIDE

WHY ARE THERE GHOSTS

WHY ARE THERE TWO SPOCKS

WHY ARE THERE HUGE SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY IS LIFE SO BORING

WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE

WHY ARE THERE GHOSTS

WHY ARE CIGARETTES LEGAL

WHY ARE THERE SPIDERS IN MY ROOM

WHY ARE THERE GHOSTS

WHY ARE THERE DUCKS IN MY POOL

WHY ARE THERE SO MANY SPIDERS IN MY ROOM

WHY ARE THERE GHOSTS

WHY IS JESUS WHITE

WHY DO SPIDER BITES ITCH

WHY ARE THERE GHOSTS

WHY IS THERE LIQUID IN MY EAR

WHY IS DYING SO SCARY

WHY ARE THERE GHOSTS

WHY DO Q TIPS FEEL GOOD

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS

WHY AREN'T THERE DINOSAUR GHOSTS

WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS

WHY DO IGUANAS DIE

WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY

WHY AREN'T THERE DINOSAUR GHOSTS

WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT

WHY AREN'T THERE DINOSAUR GHOSTS

WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES

WHY AREN'T THERE DINOSAUR GHOSTS

WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS

WHY AREN'T THERE DINOSAUR GHOSTS

WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD

WHY AREN'T THERE DINOSAUR GHOSTS

WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS

WHY AREN'T THERE DINOSAUR GHOSTS

WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO

WHY AREN'T THERE DINOSAUR GHOSTS

WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS

WHY AREN'T THERE DINOSAUR GHOSTS

WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KUNGONS DIFFERENT

WHY AREN'T THERE DINOSAUR GHOSTS

WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR

WHY AREN'T THERE DINOSAUR GHOSTS

WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE

WHY AREN'T THERE DINOSAUR GHOSTS



WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS

WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE



WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK

WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD



WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

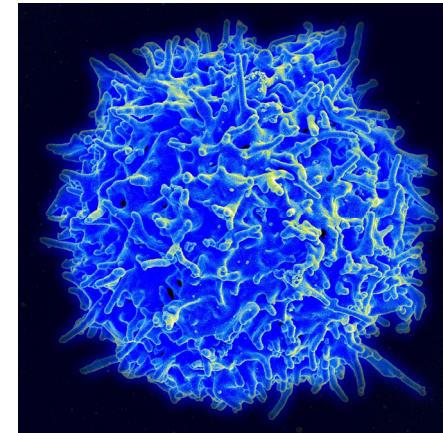
WHY IS LIFE SO BORING
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

Human T cell expression data

- The matrix contains **47 expression samples** from Lukk et al, Nature Biotechnology 2010
- All samples are **from T cells in different individuals**
- Only the **top 3000 genes** with the largest variability **were used**
- The value is **log2 of gene's expression level** in a given sample as measured by the microarray technology

a T cell



A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations | Corresponding author

Nature Biotechnology 28, 322–324 (2010) | doi:10.1038/nbt0410-322

Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (<http://www.ebi.ac.uk/gxa/array/U133A>) that allows the user to search for a gene of interest and

“Let’s Make a Deal” show with Monty Hall aired on NBC/ABC 1963-1986





**WHEEL OF
FORTUNE**

Gene Expression “Wheel of Fortune”

- Each group gets a pair of genes that are known to be correlated.
- Each group also gets a random pair of genes selected by the “Wheel of Fortune”. They may or may not be correlated
- Download (log-transformed) `expression_table.mat`
- Run command `fitlm(x,y)` on assigned and random pairs
- Record β_0 , β_1 , R^2 , P-value of the slope β_1 and write them on the blackboard
- Validate Matlab result for R^2 using your own calculations
- Look up gene names (see `gene_description` in your workspace) and write down a brief description of biological functions of genes. Does their correlation make biological sense?

Correlated pairs

plausible biological connection based
on short description

g1=1994; g2=188; group 1

g1=2872; g2=1269; group 2

g1=1321; g2=10; group 3

g1= 886; g2=819; group 4

g1=2138; g2=1364; group 5

no obvious biological common function

```
g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);  
disp([g1, g2])
```