Regression analysis
Two variables

(Montgomery and Runger: ch 11
Brani Vidakovic: ch 14)

Reminder

Covariance Defined

Covariance is a number qunatifying 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

$$\sigma_{XY} = E\left[(X - \mu_X)(Y - \mu_Y)\right] = E(XY) - \mu_X \mu_Y \qquad (5\text{-}14)$$

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

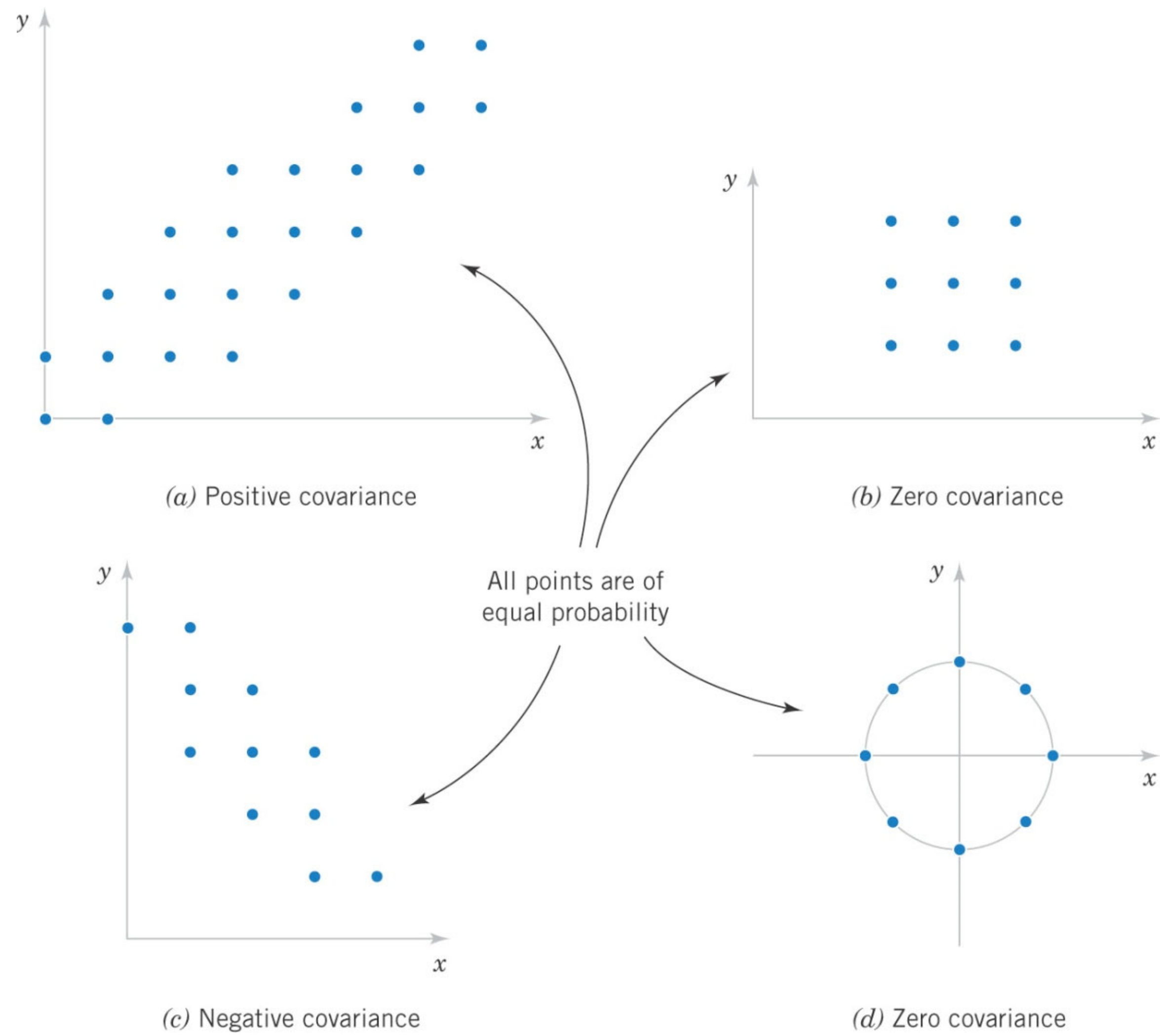


Figure 5-13 Joint probability distributions and the sign of cov(*X*, *Y*). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.

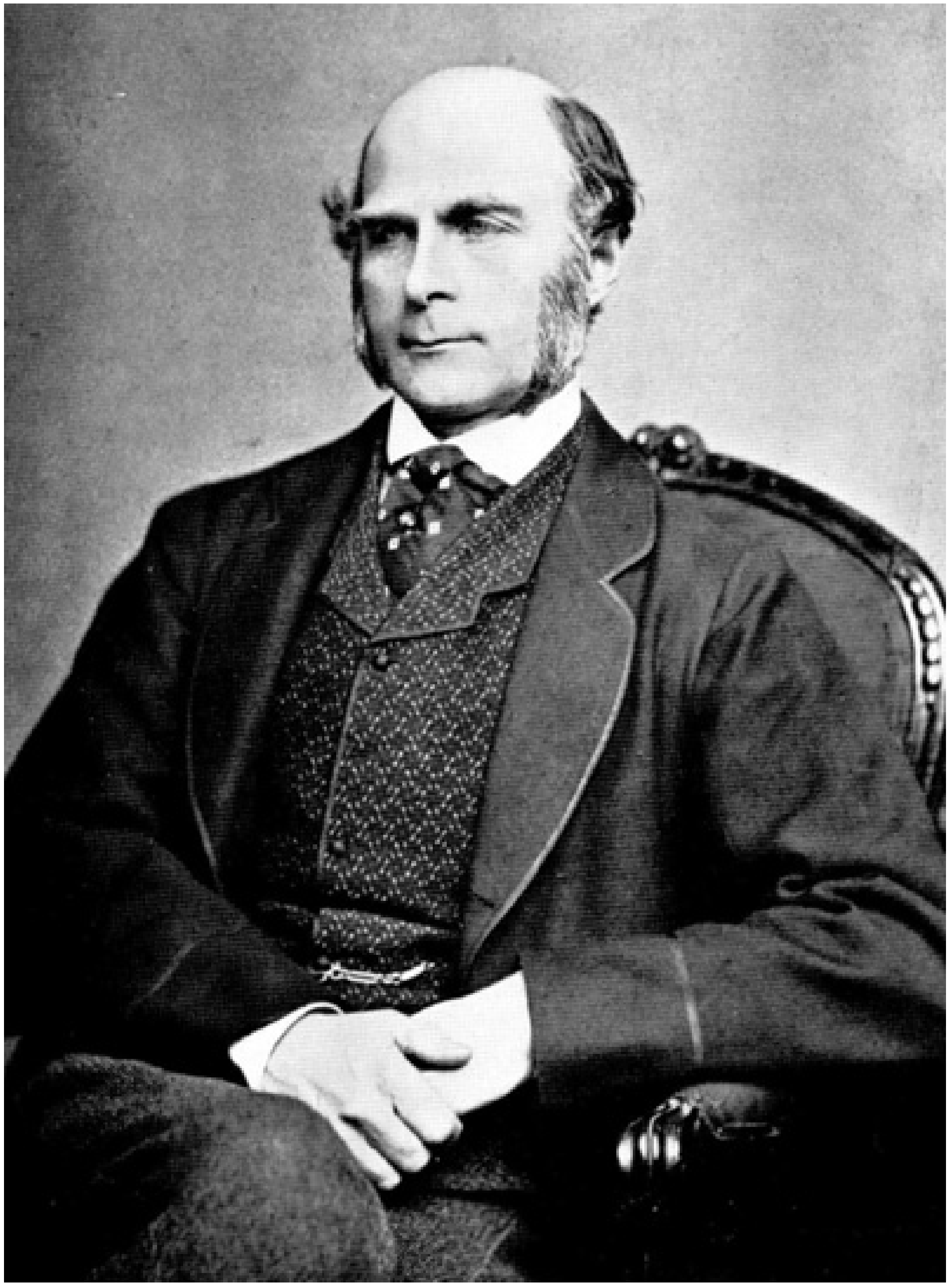
Regression analysis

• 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

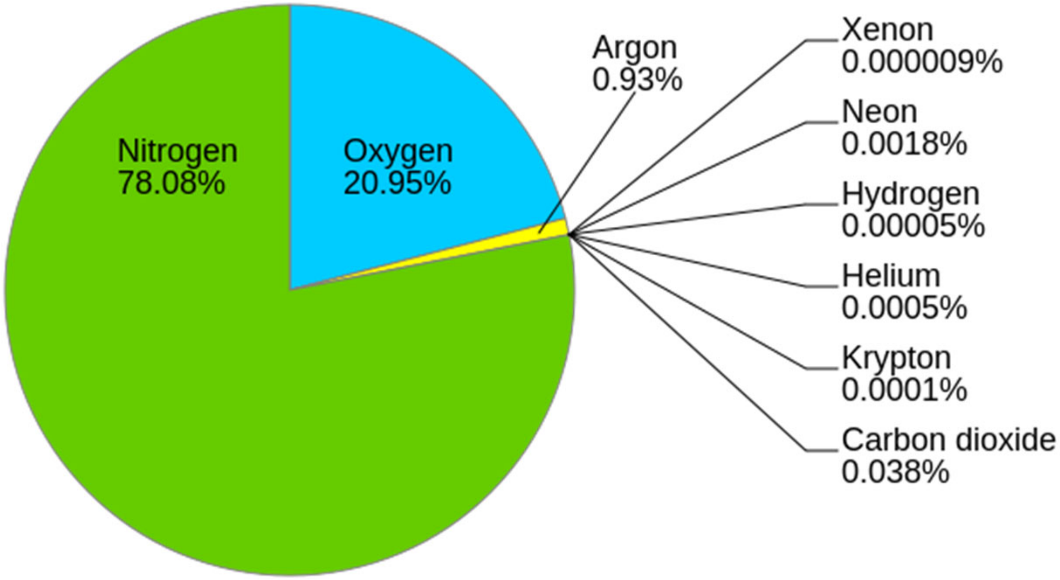


Sir Francis Galton,
(1822 -1911) was an English statistician, anthropologist, proto-geneticist, psychometrician, eugenicist, ("Nature vs Nurture", inheritance of intelligence), tropical explorer, geographer, inventor (Galton Whistle to test hearing), meteorologist (weather map, anticyclone).

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



- 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

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

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

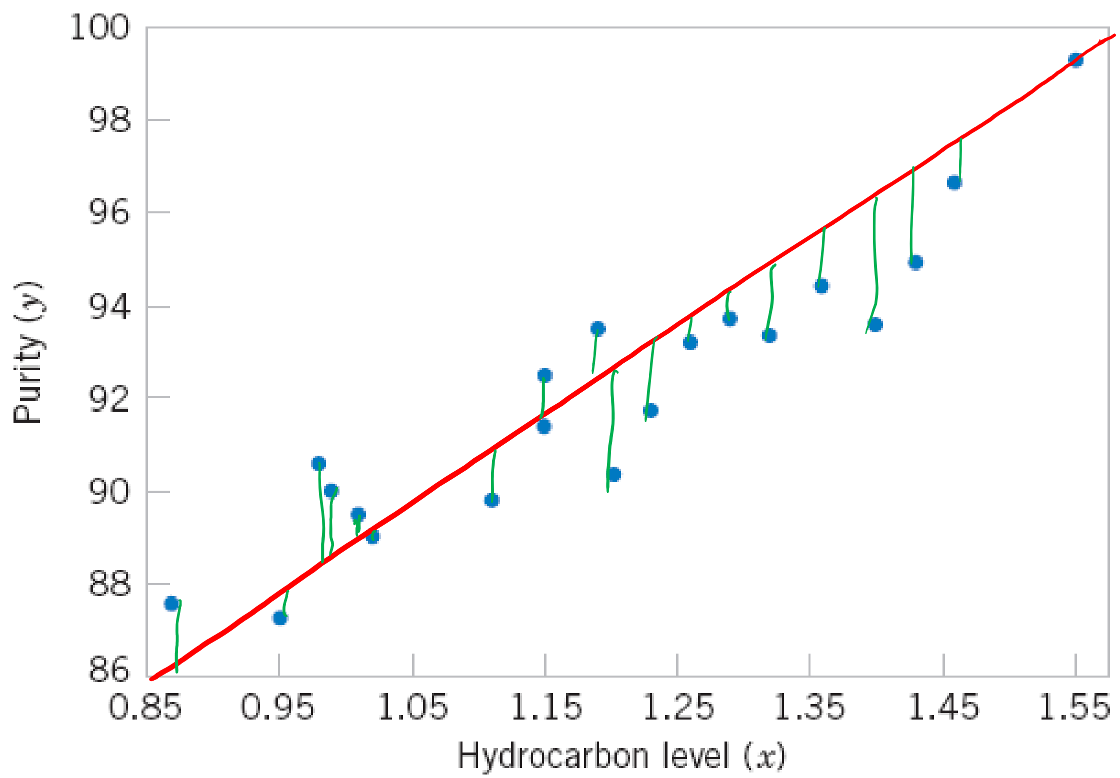


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

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

Linear regression

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

$$Y = \beta_0 + \beta_1 X + \varepsilon = \hat{Y} + \varepsilon$$

ε is the **random error term**

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

Note: Y, $\hat{Y}$, X and ε are random variables

The minimal assumption: $E(\varepsilon \mid x)=0$ →

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

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

How does one find β_0 & β_1?

$$\mathrm{Cov}(Y; X) = \mathrm{Cov}\big((\beta_0 + \beta_1 X + \epsilon); X\big) =$$

$$= \mathrm{Cov}(\beta_0; X) + \beta_1 \mathrm{Cov}(X, X) + \mathrm{Cov}(\epsilon; X)$$

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

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

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

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

Thus

$$\boxed{\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}} \; ; \quad \boxed{\beta_0 = E(Y) - \beta_1 E(X)}$$

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.

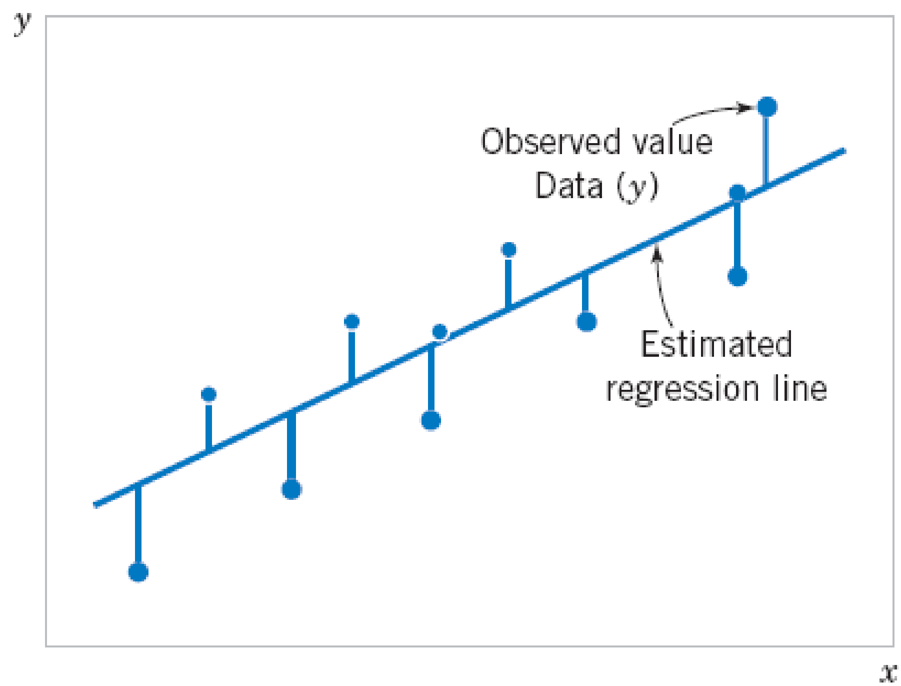


Figure 11-3 Deviations of the data from the estimated regression model.

Traditional notation

Definition

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

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

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

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

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

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} y_i x_i - \frac{\left(\displaystyle\sum_{i=1}^{n} y_i\right)\left(\displaystyle\sum_{i=1}^{n} x_i\right)}{n}}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad (11\text{-}8)$$

Handwritten note: $\frac{\sum y_i x_i}{n} - \frac{(\sum y_i)(\sum x_i)}{n^2}$ over $\frac{\sum x_i^2}{n} - \frac{(\sum x_i)^2}{n^2}$ $= \frac{Cov(X, Y)}{Var(X)}$

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

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

Symbolically,

$$SS_T = SS_R + SS_E \qquad (11\text{-}25)$$

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,

$$R^2 = SS_R/SS_T$$
$$= 152.13/173.38$$
$$= 0.877$$

- 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

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

where SS_E can be easily computed using

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

11-3: Properties of the Least Squares Estimators

- Slope Properties

$$E(\hat{\beta}_1) = \beta_1 \qquad V(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{S_{xx}} = \frac{\hat{\sigma}_\epsilon^2}{n\,\sigma_x^2}$$

Large n → small variance of β_1

- Intercept Properties

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

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

11-4: Hypothesis Tests in Simple Linear Regression

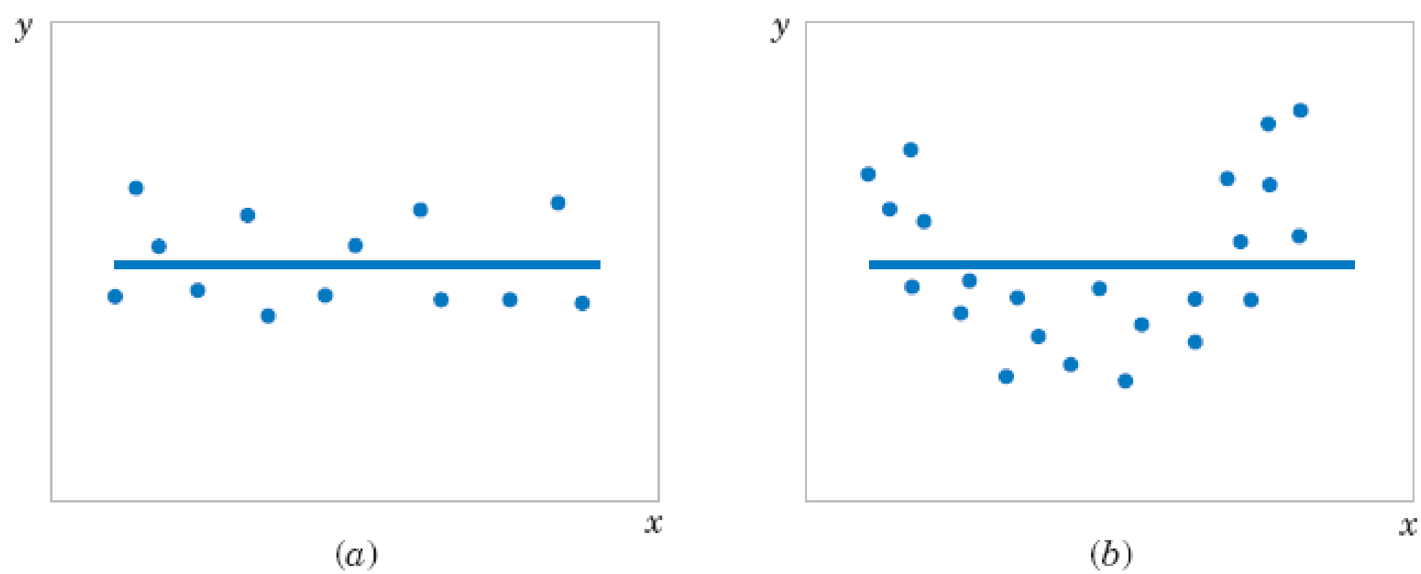


Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

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

11-4: Hypothesis Tests in Simple Linear Regression

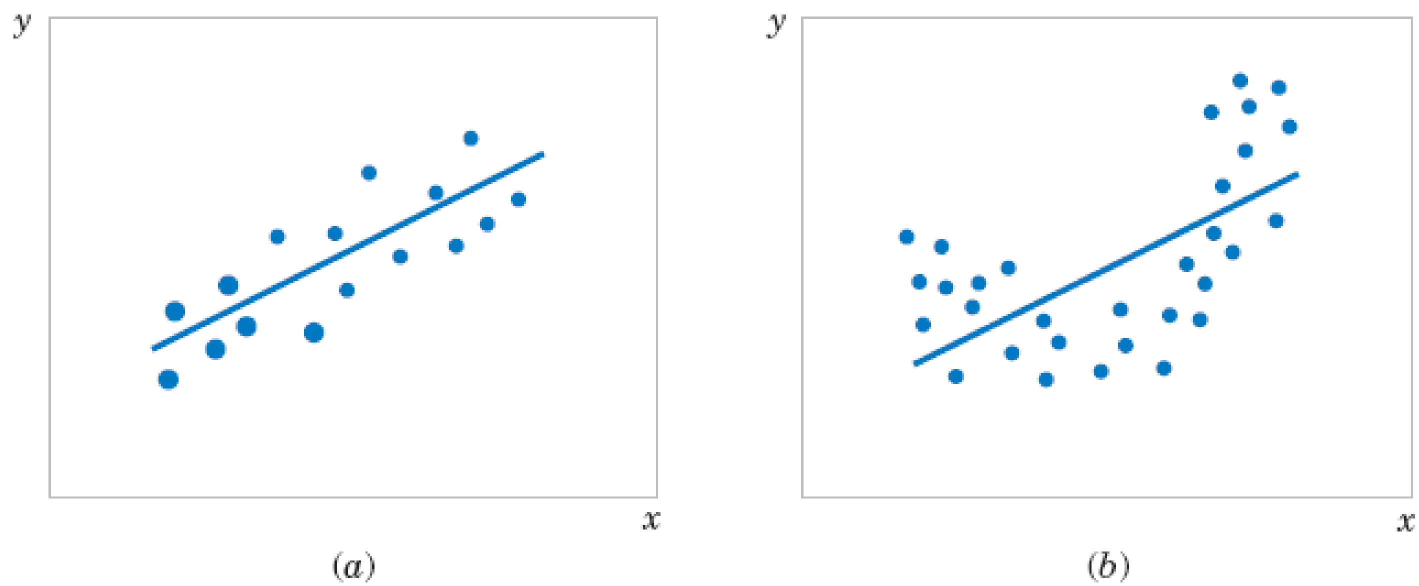


Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

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)

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

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>=3

11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

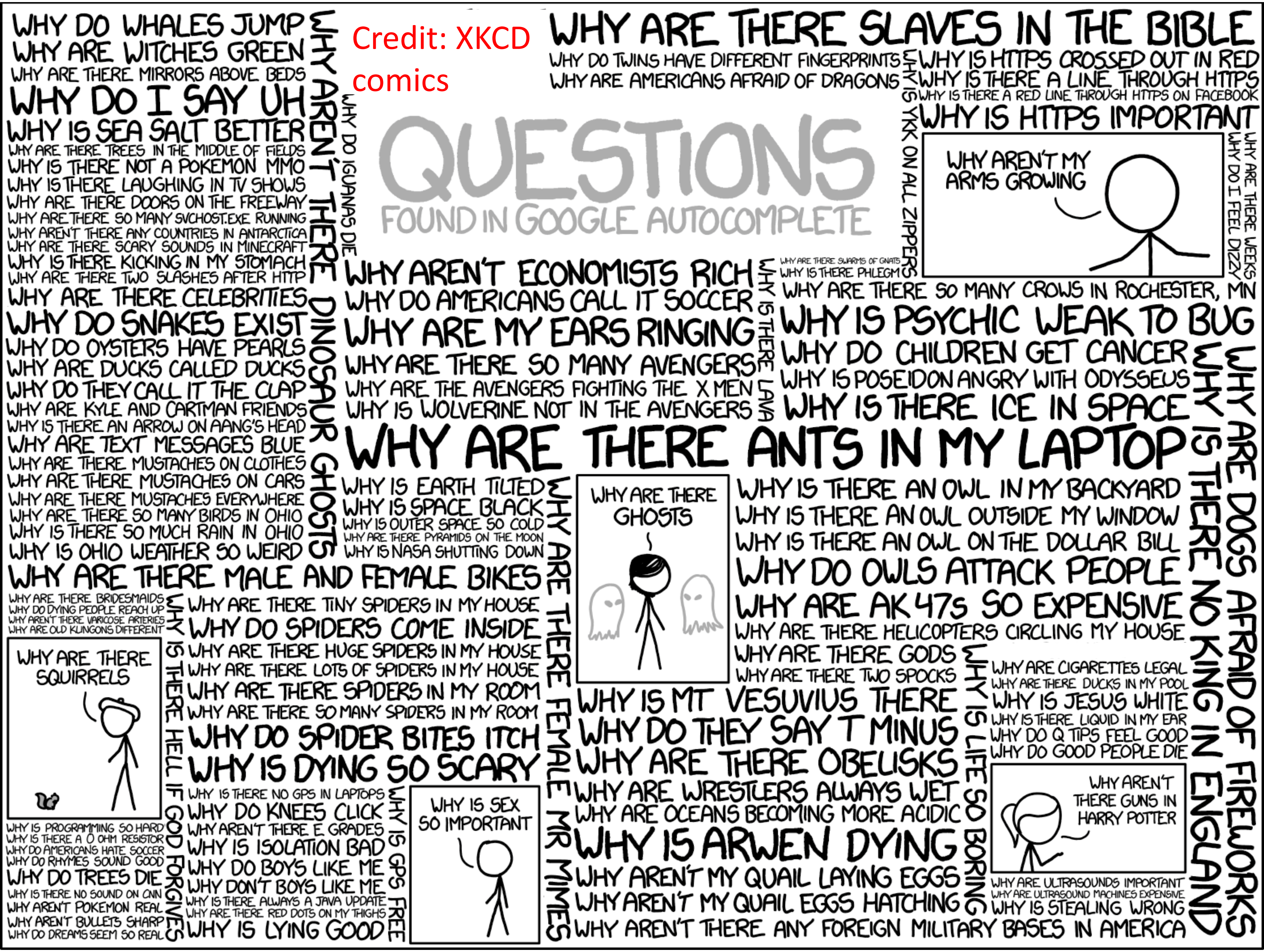
Choose α (e.g. $\alpha = 5\%$ for 95% confidence in rejecting H_0)

$$T = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}_\epsilon}{\hat{\sigma}_x} \cdot \frac{1}{\sqrt{n}}}$$

$t_{\alpha/2, n-2}$ is such

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

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

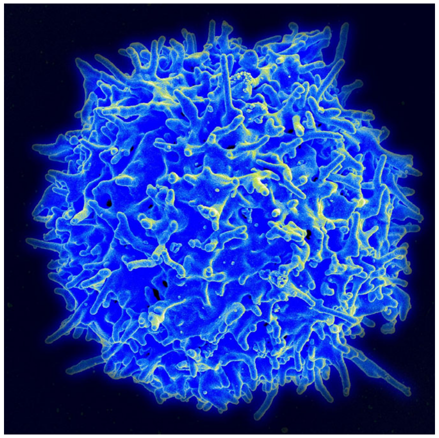


Credit: XKCD comics

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