

Goodness of fit test

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

<http://www.scientificameriken.com/candy5.asp>

"To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (<http://us.mms.com/us/about/products/milkchocolate/>). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.

On average, our new mix of colors for M&M'S® Chocolate Candies is:

M&M'S® Milk Chocolate: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.

M&M'S® Peanut: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.

M&M'S® Kids MINIS®: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.

M&M'S® Crispy: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.

M&M'S® Peanut Butter and Almond: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA
A Division of Mars, Incorporated



How to accept or reject the null hypothesis that these probabilities are correct from a finite sample?

Pearson χ^2 Goodness of Fit Test

- Assume there is a **sample of size n** from a population
- There are **k classes** (e.g. six M&M colors) in the populations
- **Null hypothesis** H_0 : class i has frequency f_i in the population
- **Alternative hypothesis** H_1 : some population frequencies are inconsistent with f_i
- Let O_i be the **observed frequency** in the i th class.
- Let $E_i = n f_i$ be the **expected frequency** in the i th class.
- **Group together bins** with $E_i < 3$

The **test statistic** is

$$X_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (9-47)$$

P-value is calculated based on **chi-square distribution** with **$k-1$ degrees of freedom**:

$$\text{P-value} = \text{Prob}(H_0 \text{ is correct}) = 1 - \text{CDF_chi-squared}(X_0^2, k-1)$$

What is chi-squared distribution?

It is a PDF of the **sum of squares of k-1 standard normal (Gaussian) random variables**

k-1-degrees of freedom

$$f(x,n) \sim x^{(k-1)/2-1} \exp(-x/2)$$

It is just a Gamma PDF
with $r=(k-1)/2$, and $\lambda=1/2$

Mean value:

$$k-1$$

Standard deviation:

$$\sqrt{2(k-1)}$$

Play with Mathematica notebook

<http://demonstrations.wolfram.com/ChiSquaredDistributionAndTheCentralLimitTheorem/>

By Peter Falloon

M&M group exercise

- DO NOT EAT CANDY BEFORE COUNTING IS FINISHED! THEN, PLEASE, DO.
- We will be testing three null hypotheses:
 - M&M data: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown
 - Website data from <http://joshmadison.com/2007/12/02/mms-color-distribution-analysis>: 18.36% blue, 20.76% orange, 18.44% green, 14.08% yellow, 14.20% red, 14.16% brown
 - Uniform distribution: 1/6~16.67% of each candy color
- You will estimate P-values for each of these null hypothesis
- Hints: calculate E_i as (# candies in your sample)* f_i

Use **1-chi2cdf(X0squared, 5)** for P-value

$$X_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

M&M matlab exercise

- `p_mm=[0.24,0.2,0.16, 0.14, 0.13,0.13];`
- `p_u=1./6.*ones(1,6);`
- `p_website=[0.1836, 0.2076, 0.1844, 0.1408, 0.1420, 0.1416]`
- `sum1=sum(observed)`
- `predicted_u=sum1.*p_u; predicted_mm=sum1.*p_mm;`
`predicted_website=sum1.*p_website;`
- `gf_mm=0; gf_u=0; gf_website=0;`
- `for m=1:6;`
 `gf_mm=gf_mm+(observed(m)-predicted_mm(m)).^2./predicted_mm(m);`
 `gf_u=gf_u+(observed(m)-predicted_u(m)).^2./predicted_u(m);`
 `gf_website=gf_website+(observed(m)-`
 `predicted_website(m)).^2./predicted_website(m);`
`end;`
- `disp('goodness of fit of MM ='); disp(num2str(gf_mm));`
- `disp('p-value of MM ='); disp(num2str(1-chi2cdf(gf_mm,5))); disp(' ');`
- `disp('goodness of fit of website ='); disp(num2str(gf_website));`
- `disp('p-value of MM ='); disp(num2str(1-chi2cdf(gf_website,5))); disp(' ');`
- `disp('goodness of fit of uniform ='); disp(num2str(gf_u));`
- `disp('p-value of uniform='); disp(num2str(1-chi2cdf(gf_u,5)));`

Statistical tests of independence

- Did I mix M&M candy well?

	blue	orange	green	yellow	red	brown
group 1	56	62	36	36	37	35
group 2	59	67	29	39	32	25
group 3	58	63	29	28	33	24
group 4	58	60	36	22	37	36

How to test the hypothesis if multiple sample are drawn from the same population?

- Table: **samples (Student groups) – rows**, **classes (M&M colors) – columns**
- Test if color fractions are independent from group
- $P(\text{Group 1 and Color = green}) = P(\text{Group 1}) * P(\text{Color green})$
- Compute for all groups/colors $6*4=24$ in our case

$$E_{\text{green}}(\text{group 1}) = n_{\text{tot}} * (\text{group 1} / n_{\text{tot}}) * (\text{green} / n_{\text{tot}})$$

- $\chi^2 = \sum_{\text{groups \& colors}}^{n_{\text{tot}}} \frac{(O_{\text{color}}(\text{group}) - E_{\text{color}}(\text{group}))^2}{E_{\text{color}}(\text{group})}$
- # degrees of freedom = $(\text{colors}-1) * (\text{groups}-1)$

- Did I mix M&M candy well?

	blue	orange	green	yellow	red	brown
group 1	56	62	36	36	37	35
group 2	59	67	29	39	32	25
group 3	58	63	29	28	33	24
group 4	58	60	36	22	37	36

- Using $\chi^2 = \sum_{groups \& colors} \frac{(O_{color}(group) - E_{color}(group))^2}{E_{color}(group)}$
 with # degrees of freedom $(colors-1)*(groups-1)$
 Find P-value of null hypothesis H_0 that
samples are independent from each other

My solution

- `disp(mm_table)`
- `sumt=sum(sum(mm_table))`
- `sum_color=sum(mm_table, 1)`
- `sum_group=sum(mm_table, 2)`
- `mm_exp=kron(sum_group,sum_color)./sumt`
- `gof=sum(sum((mm_table-mm_exp).^2./mm_exp))`
- `P_value_gof=1-chi2cdf(gof, (4-1)*(6-1))`
- `%gof =10.8858; P_value_gof =0.7606`
- Null model that samples are independent is not rejected
- I mixed this bag well!

Batch effect

Does color composition vary from bag to bag?

- This year: 231 252 130 125 139 120
- Last year: 283 286 304 124 191 173
- Test if they are significantly different from each other:
- Same test expect `ngroups=2; ncolors=6;`
- Results:
Goodness of Fit = 39.7111
P_value = 1.7077e-07
- Batch effect is highly statistically significant!

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

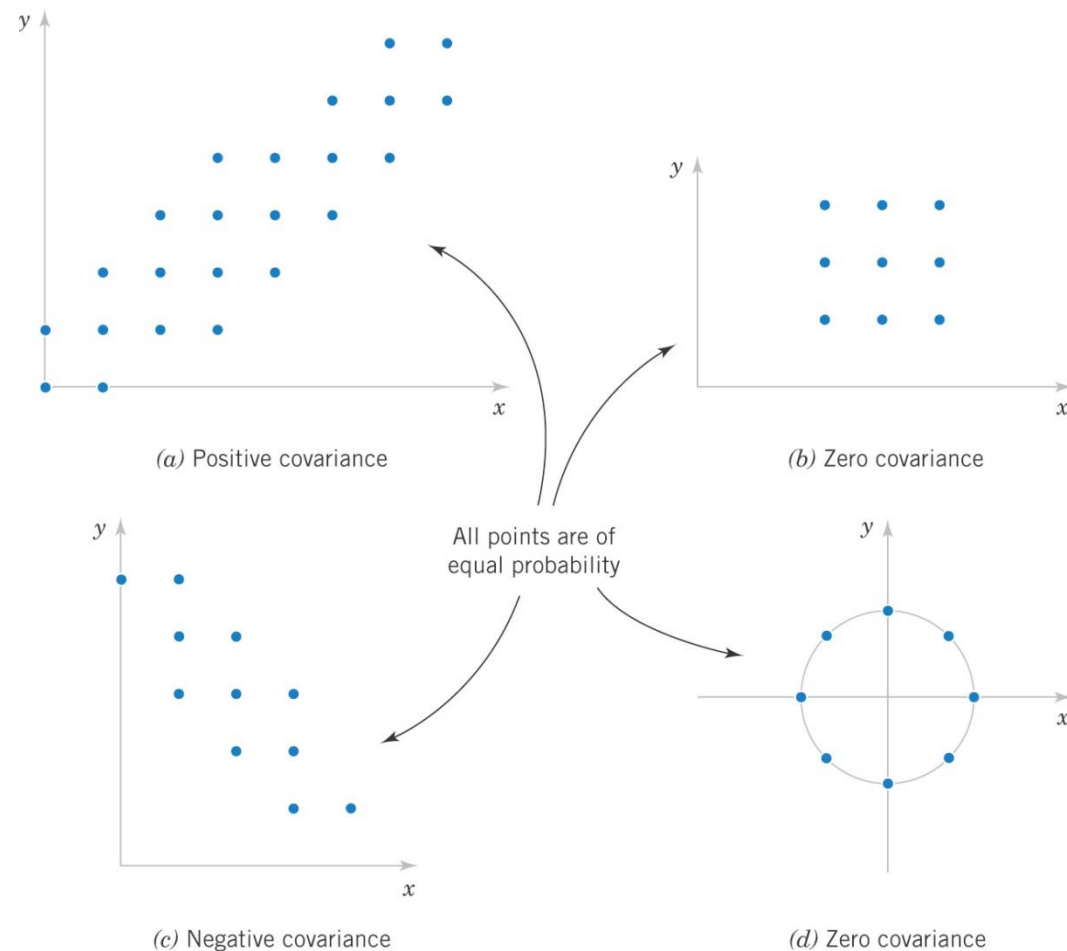


Figure 5-13 Joint probability distributions and the sign of $\text{cov}(X, Y)$. Note that covariance is a measure of linear relationship. Variables with non-zero covariance are **correlated**.

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

Credit: XKCD
comics

QUESTIONS FOUND IN GOOGLE AUTOCOMPLETE



WHY ARE THERE WEEKS IN MAY DO I FEEL DIZZY

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARIOUS PRIETIES
WHY ARE OLD KLINGONS DIFFERENT



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

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 AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE ANTS IN MY LAPTOP



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 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



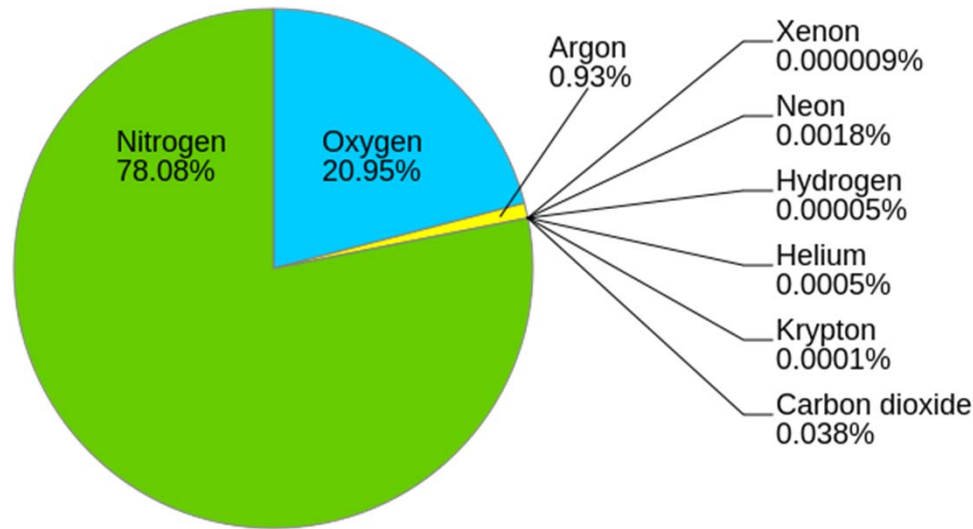
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

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

Regression analysis

- Many problems in engineering and science involve sample in which two or more variables were measured
- **Regression analysis** is a statistical technique that is very useful for these types of problems
- 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

Two variable samples



- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go kaboom!!!
- The more hydrocarbons were removed the cleaner is the remaining oxygen
- 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

Scatter plot to visually detect trends

Red line is called **linear regression**

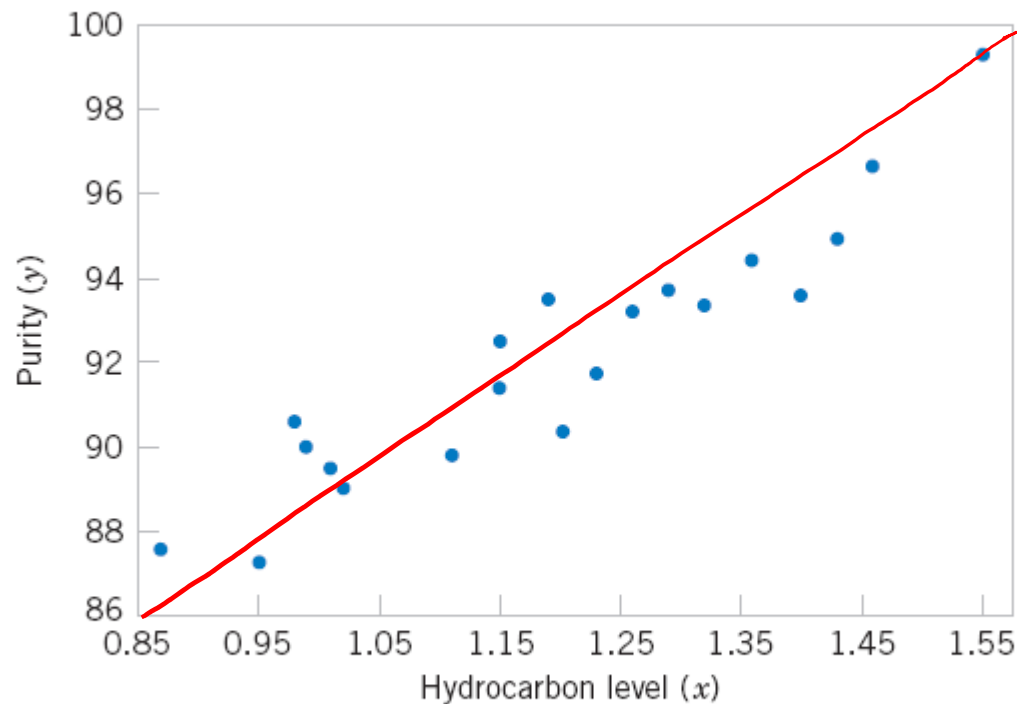
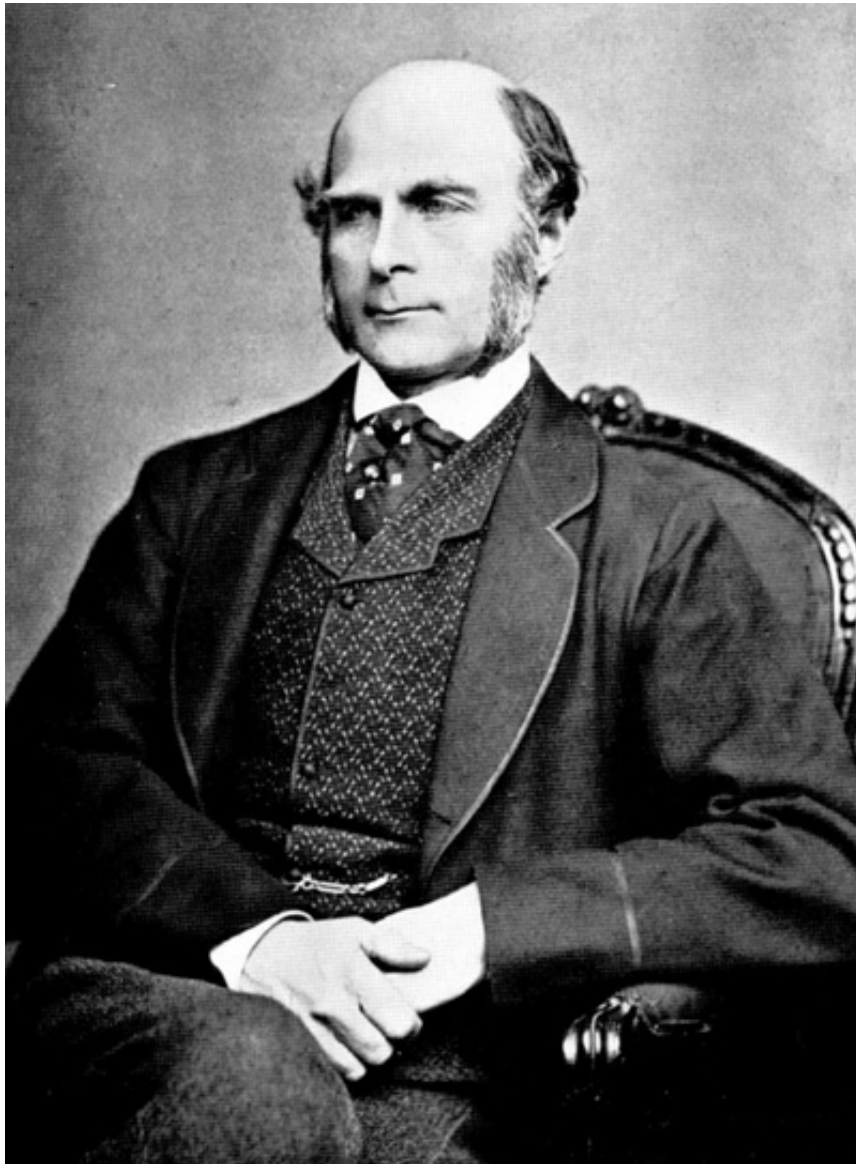


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.



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**

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

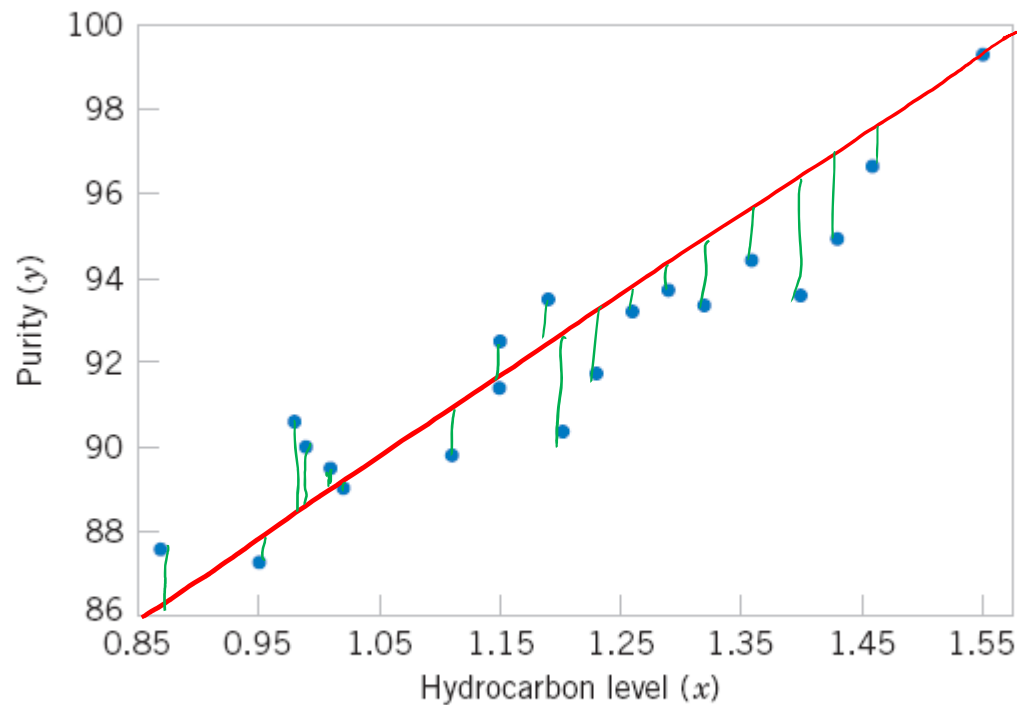


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

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

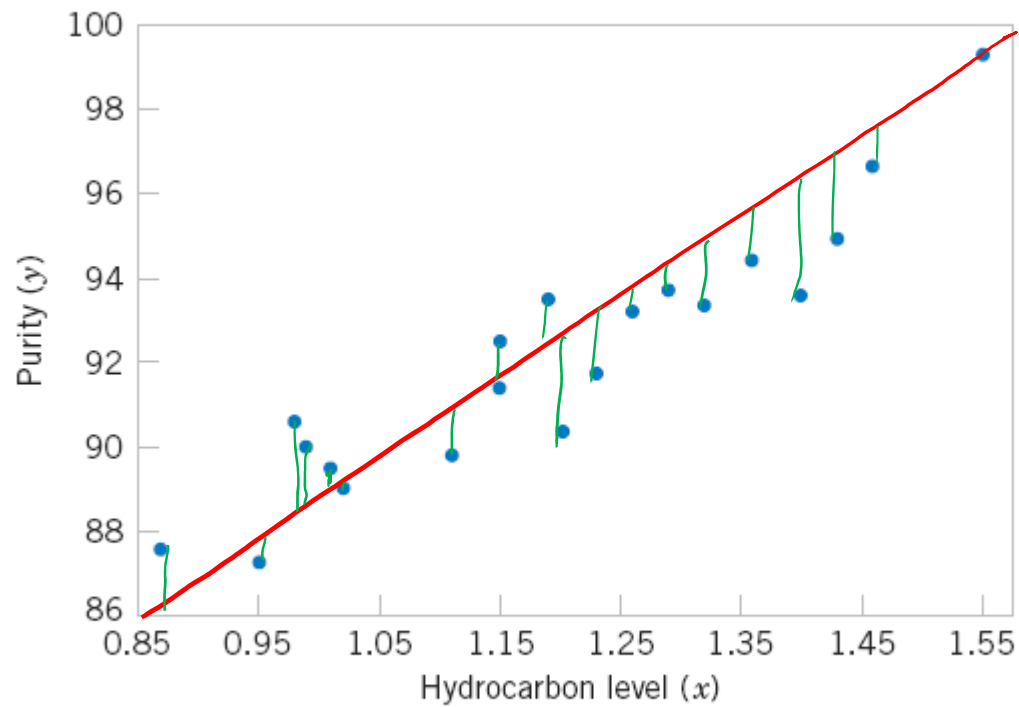


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

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

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

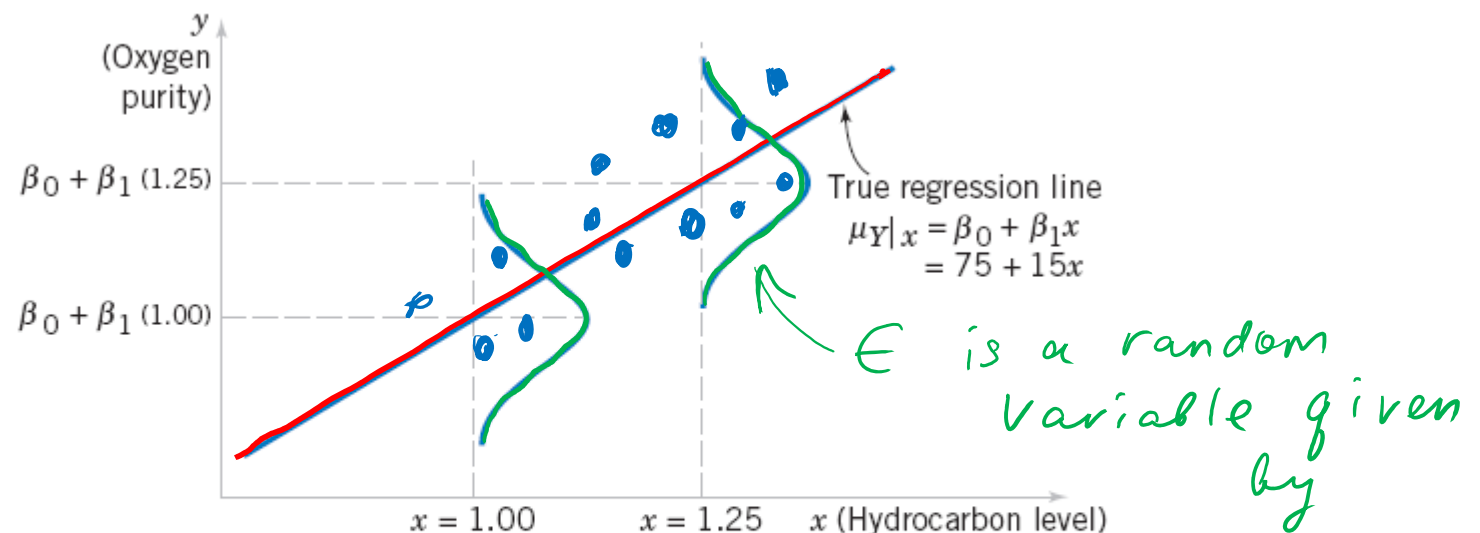


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity–hydrocarbon data.

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

$$E(\epsilon | x) = 0 \text{ for all } x$$

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

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.

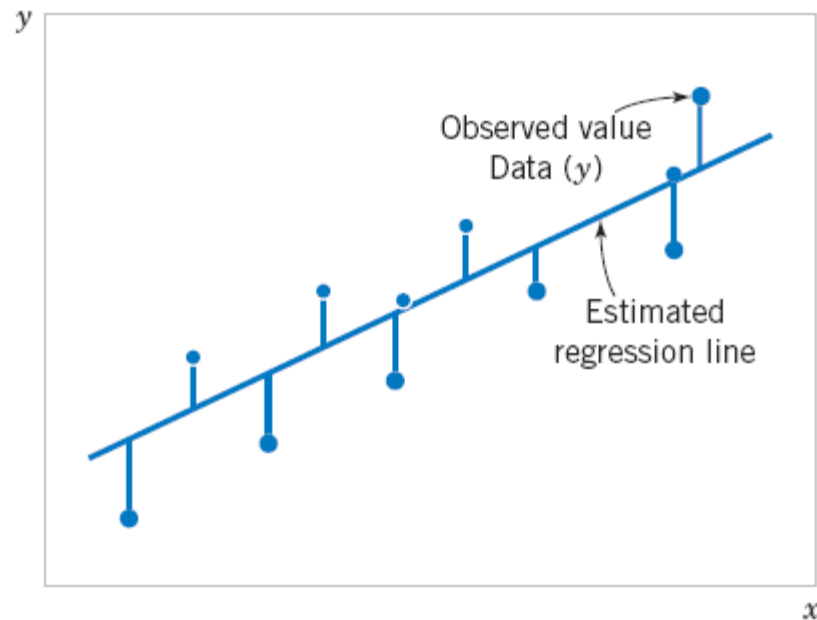


Figure 11-3 Deviations of the data from the estimated regression model.

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

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

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

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 USEFUL

- 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_e^2

The error sum of squares is

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

11-3: Properties of the Least Squares Estimators

- Slope Properties

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

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

Large $n \rightarrow$ small
variance
of β_1

- Intercept Properties

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

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

11-4: Hypothesis Tests in Simple Linear Regression

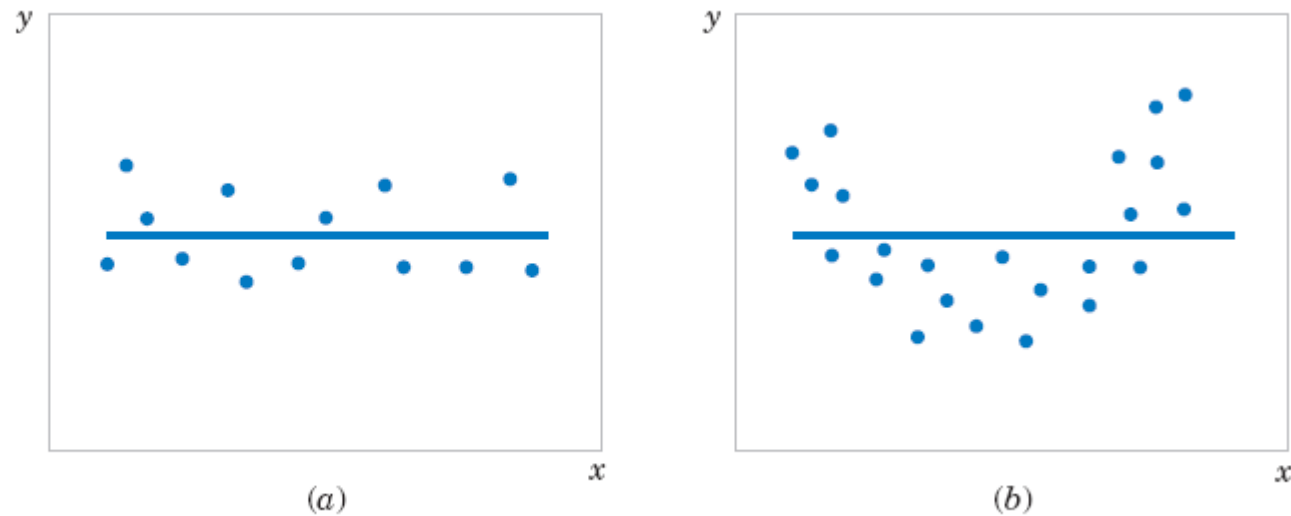


Figure 11-5 The hypothesis $H_0: \beta_1 = 0$ is not rejected.

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

11-4: Hypothesis Tests in Simple Linear Regression

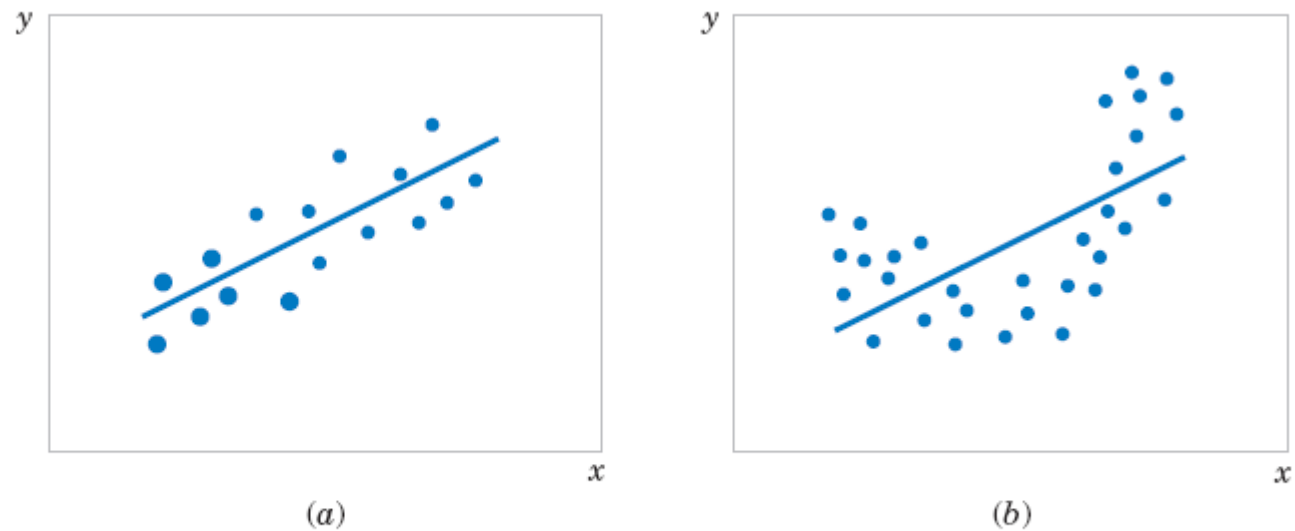


Figure 11-6 The hypothesis $H_0: \beta_1 = 0$ is rejected.

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 *t*-Tests

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)

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

for $\alpha = 5\%$

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

T-cell expression data

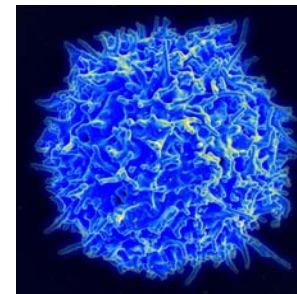
- The matrix contains **47 expression samples** from Lukk et al, Nature Biotechnology 2011
- All samples are **from T cells in different individuals**
- Only the **top 3000 genes** with the largest variability **were used**
- The value is **log2 of gene's expression level** in a given sample as measured by the microarray technology

A 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

Gene Expression “Wheel of Fortune”

- Each group gets a pair of genes that are known to be correlated.
- Each group will also get 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) and write down a brief description of biological functions of genes. Does their correlation make biological sense?



Correlated pairs

2907	2881
1994	188
2274	1597
2982	1353
2872	1269
1321	10
2935	1654
2745	1695
886	819
2138	1364

Random pairs

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

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

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

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

Matlab code

- load expression_table.mat
- g1=2907; g2=288;
- x=exp(g1,:)' ; y=exp(g2,:)' ;
- figure; plot(x,y,'ko');
- lm=fitlm(x,y)
- y_fit=lm.Fitted;
- hold on; plot(x,lm.Fitted,'r-');
- SST=sum((y-mean(y)).^2);
- SSR=sum((y_fit-mean(y)).^2);
- SSE=sum((y-y_fit).^2);
- R2=SSR./SST
- disp([gene_names(g1), gene_names(g2)]);
- disp(gene_description(g1)); disp(gene_description (g2));

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

Credit: XKCD
comics

QUESTIONS FOUND IN GOOGLE AUTOCOMPLETE



WHY ARE THERE WEEKS IN MAY DO I FEEL DIZZY

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARIOUS PRIETIES
WHY ARE OLD KLINGONS DIFFERENT



WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN



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 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 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 ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

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