M&M group exercise

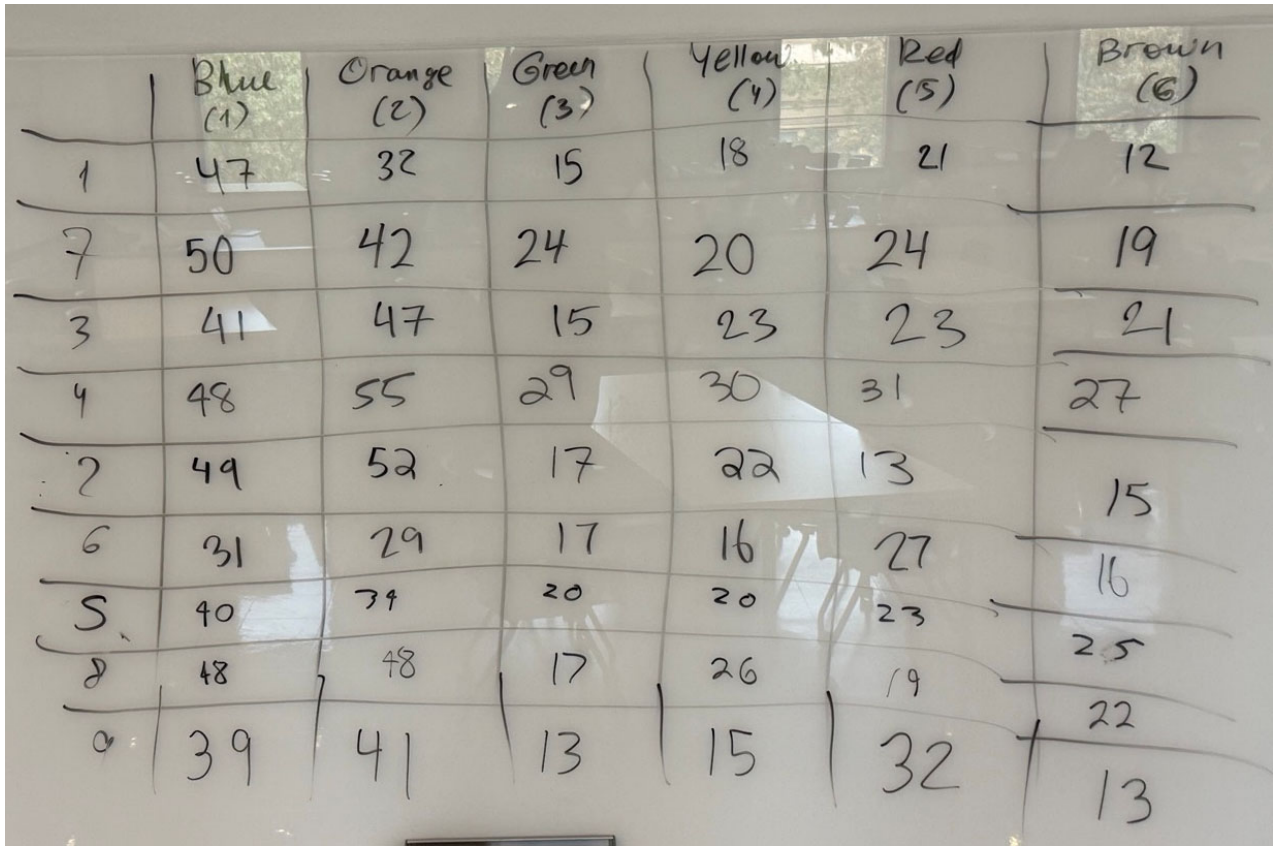
- DO NOT EAT CANDY BEFORE COUNTING IS FINISHED! THEN, **PLEASE, DO**.
- We will be testing three null hypotheses one after another:
  - M&M peanut official data: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown
  - Website (fan collected) data from http://joshmadison.com/2007/12/02/mms-color-distribution-analysis: 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 one of these null hypotheses
- Hints: O_i – is the observed # of candies of color i; calculate the expected # E_i=(# 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

```
% Groups 1 through 5 got samples from an unmixed box, Groups 6 through 9 got samples from a mixed box. The box was Milk M&Ms from Costco
% 1- blue,2-orange,3-green,4-yellow,5-red,6-brown
ngroups=9;
mm_table=zeros(ngroups,6);
% columns: Blue, Orange, Green, Yellow, Red, Brown
mm_table(1,:)=[47,32,15,18,21,12];
mm_table(2,:)=[49,52,17,22,13,15];
mm_table(3,:)=[41,47,15,23,23,21];
mm_table(4,:)=[48,55,29,30,31,27];
mm_table(5,:)=[40,34,20,20,23,25];
mm_table(6,:)=[31,29,17,16,27,16];
mm_table(7,:)=[50,42,24,20,24,19];
mm_table(8,:)=[48,48,17,26,19,22];
mm_table(9,:)=[39,41,13,15,32,13];
for group=1:ngroups;
observed=mm_table(group,:);
group % use when analyzing one group
f_mm=[0.24,0.2,0.16, 0.14, 0.13,0.13];
f_website=[18,21,18,14,14,14];
f_website=f_website./sum(f_website);
f_u=1./6.*ones(1,6);
n=sum(observed);
expected_u=n.*f_u;
expected_mm=n.*f_mm;
expected_website=n.*f_website;
gf_mm=0; gf_u=0; gf_website=0;
for m=1:6;
gf_mm=gf_mm+(observed(m)...
-expected_mm(m)).^2./expected_mm(m);
gf_u=gf_u+(observed(m)-expected_u(m)).^2./expected_u(m);
gf_website=gf_website+(observed(m)...
-expected_website(m)).^2./expected_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 website ='); 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)));
end;
```



	Blue (1)	Orange (2)	Green (3)	Yellow (4)	Red (5)	Brown (6)
1	47	32	15	18	21	12
7	50	42	24	20	24	19
3	41	47	15	23	23	21
4	48	55	29	30	31	27
2	49	52	17	22	13	15
6	31	29	17	16	27	16
5	40	34	20	20	23	25
8	48	48	17	26	19	22
9	39	41	13	15	32	13

Box of M&M milk chocolate from Costco

Groups 1 through 5 – box was not mixed

Groups 6 through 9 – box was mixed

Statistical tests of independence

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(Group 1 and Color = green) = P(Group 1)*P(Color green)***
- Compute for all groups/colors 6*4=24 in our case

$E_{green}(group\ 1) = n_{tot} * (group1/n_{tot}) * (green/n_{tot})$

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

- Was the Milk M&M box mixed well?
  Let's compare the data from groups 1 through 5
- Using $\chi^2 =$
  $\sum_{groups\ \&\ colors} \frac{\left(O_{color}\,(group) - E_{color}(group)\right)^2}{E_{color}(group)}$

with # degrees of freedom =
(colors-1)*(groups-1)=(6-1)*(5-1)=20

Find P-value of null hypothesis H_0 that
samples are independent from each other

- For milk the P-value is 0.4431. I mixed this bag well!

Was the bag of milk M&Ms well mixed?

```
mm_table_all=mm_table;
%%
% Test of statistical independence.
% Did I mix the box well?
% Let's check the first 5 groups which got an unmixed sample
clear mm_table
mm_table=mm_table_all(1:5, :);
ngroups=5;
ncolors=6;
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, (ngroups-1)*(ncolors-1))
%gof = 20.2377
%P_value_gof = 0.4431
```

Meta-analysis

now that I know that me not mixing the box did not have any effect
I can compute the P-values of three hyposeses on meta-analysis of all 9 samples

```
observed=sum(mm_table_all, 1) f_mm=[0.24,0.2,0.16, 0.14, 0.13,0.13];
f_website=[18,21,18,14,14,14];
f_website=f_website./sum(f_website);
f_u=1./6.*ones(1,6);
n=sum(observed);
expected_u=n.*f_u;
expected_mm=n.*f_mm;
expected_website=n.*f_website;
gf_mm=0; gf_u=0; gf_website=0;
for m=1:6;
gf_mm=gf_mm+(observed(m)...
-expected_mm(m)).^2./expected_mm(m);
gf_u=gf_u+(observed(m)-expected_u(m)).^2./expected_u(m);
gf_website=gf_website+(observed(m)...
-expected_website(m)).^2./expected_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 website ='); 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)));
```

Meta-analysis results

- goodness of fit of MM = 52.764
  p-value of MM = 3.7603e-10

- goodness of fit of website = 115.5977
  p-value of website = 0

- goodness of fit of uniform =220.4052
  p-value of uniform = 0

- All three null-hypotheses are rejected!

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

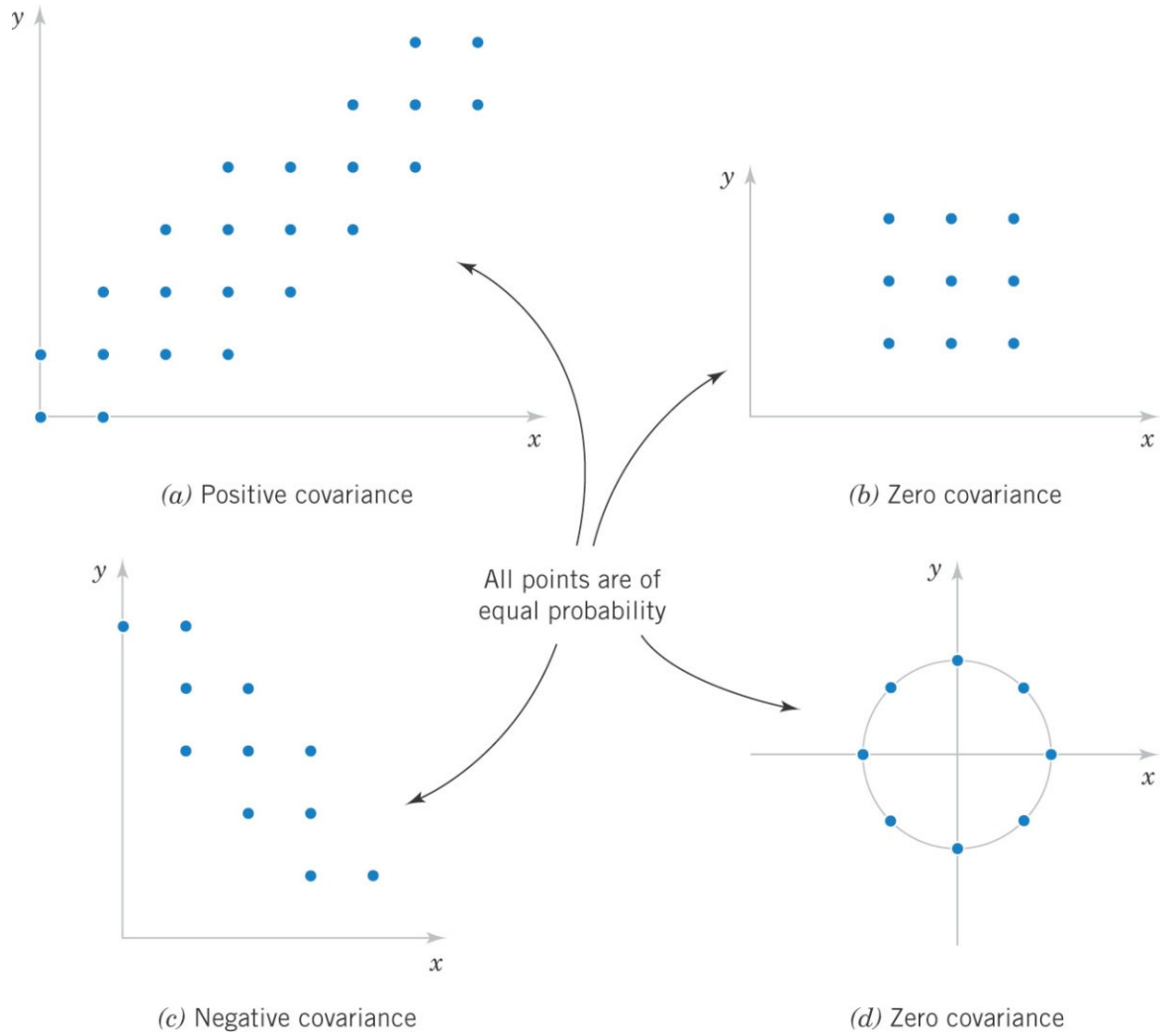


Figure 5-13 Joint probability distributions and the sign of cov(X, Y). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.

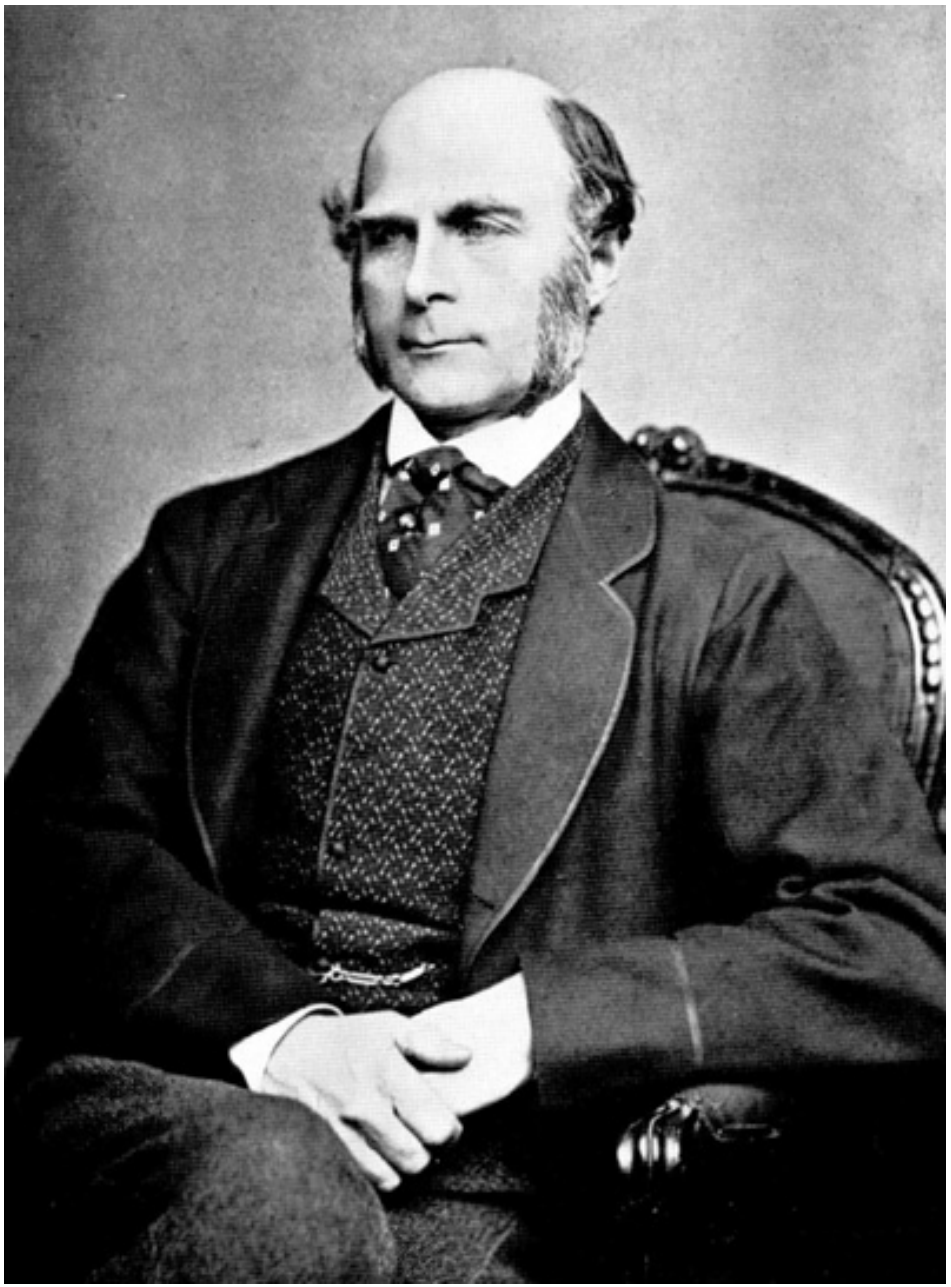
Regression analysis

• 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

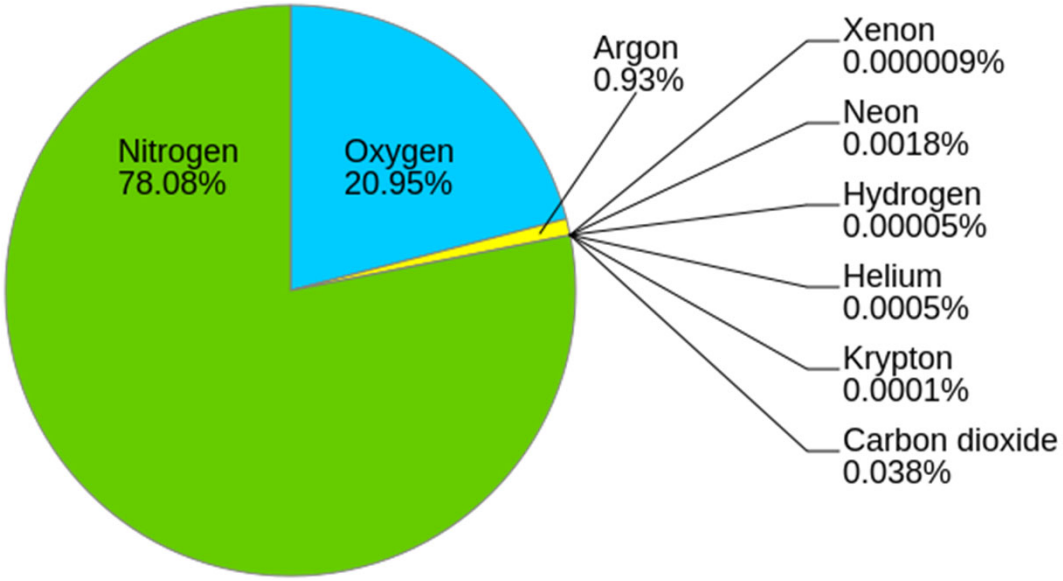


Sir Francis Galton,
(1822 -1911) was an English statistician, anthropologist, proto-geneticist, psychometrician, eugenicist, ("Nature vs Nurture", inheritance of intelligence), tropical explorer, geographer, inventor (Galton Whistle to test hearing), meteorologist (weather map, anticyclone).

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

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

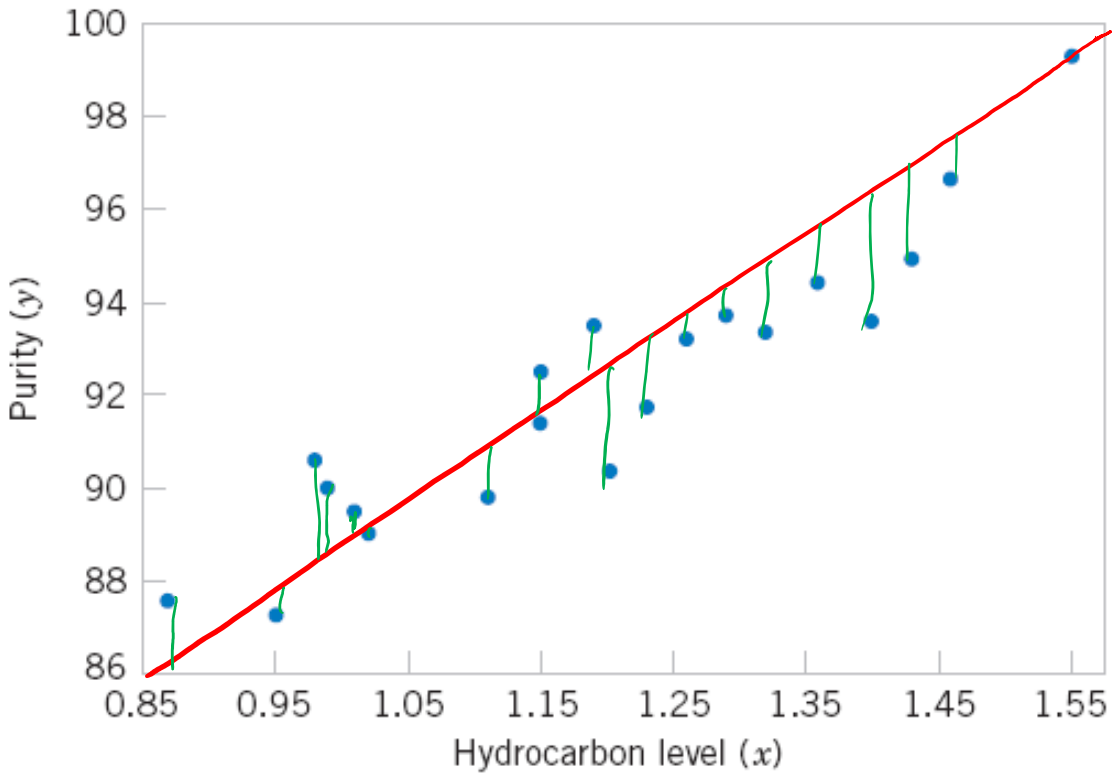


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

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

Linear regression

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

$$Y = \beta_0 + \beta_1 X + \mathrm{E} = \hat{Y} + \mathrm{E}$$

E is the **random error**

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

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

Let's assume that $\mathrm{E}(\mathrm{E}|\, x)=0$ →
$\mathrm{E}(Y|\, x) = \beta_0 + \beta_1 x + \mathrm{E}(\mathrm{E}|\, x) = \beta_0 + \beta_1 x$

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

How does one find β_0 & β_1?

$$Cov(Y, X) = Cov((\beta_0 + \beta_1 X + \epsilon), X) =$$

$$= Cov(\beta_0, X) + \beta_1 Cov(X, X) + Cov(\epsilon, X)$$

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

$Cov(X, X) = E(X^2) - E(X)^2 = Var(X)$

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

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

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

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.

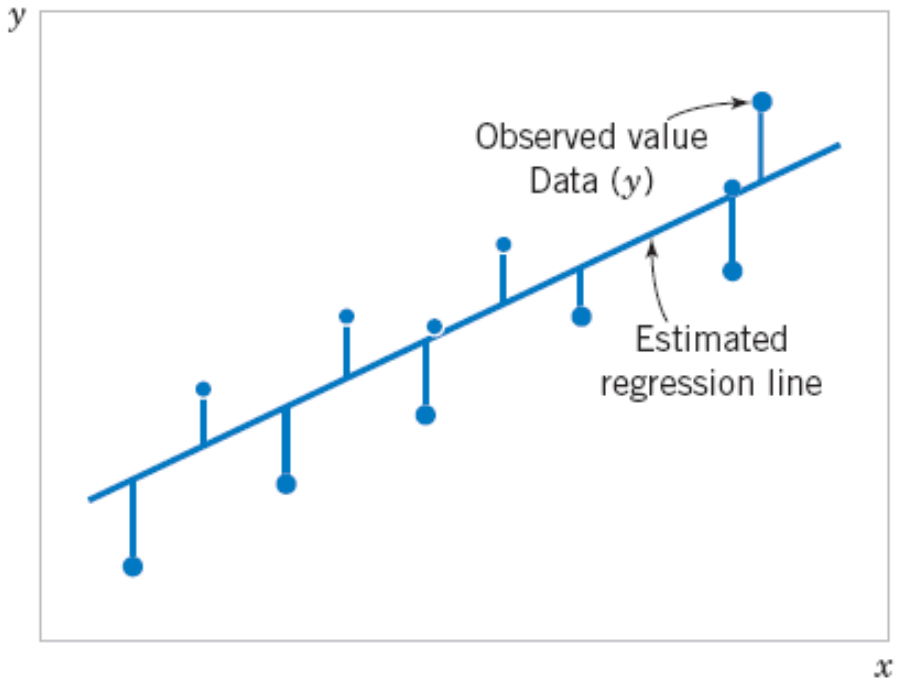


Figure 11-3 Deviations of the data from the estimated regression model.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{11-3}$$

and the sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{11-4}$$

The least squares estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \tag{11-5}$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \tag{11-6}$$

$$\hat{\beta}_0 \sum x_i = -\hat{\beta}_1 \frac{\left(\sum x_i\right)^2}{n} + \frac{\sum y_i \sum x_i}{n}$$

Traditional notation

Definition

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

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

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

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

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