UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

PhD Course
Advanced Biostatistics

Lecture 3
Linear Models:
ANOVA and Linear Regression

Peter Nazarov
petr.nazarov@lih.lu

30-05-2017

Outline

- **ANOVA (L3.1)**
  - 1-factor ANOVA
  - Multifactor ANOVA
  - Experimental design
- **Linear regression (L3.2)**
  - Simple linear regression
  - Multiple regression
  - Selecting variables

L3.1. ANOVA

Why ANOVA ?

Means for more than 2 populations

We have measurements for 5 conditions. Are the means for these conditions equal?

Validation of the effects

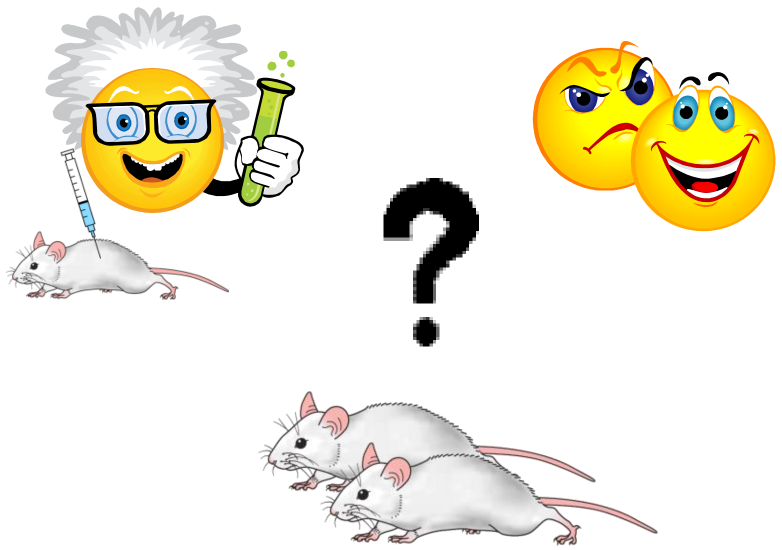
We assume that we have several factors affecting our data. Which factors are most significant? Which can be neglected?

ANOVA
example from Partek™

If we would use pairwise comparisons, what will be the probability of getting error?

Number of comparisons: $C_2^5 = \frac{5!}{2!3!} = 10$

Probability of an error: $1-(0.95)^{10} = 0.4$



http://easylink.playstream.com/affymetrix/ambsymposium/partek_08.wvx

L3.1. One-way ANOVA

Example

As part of a long-term study of individuals 65 years of age or older, sociologists and physicians at the Wentworth Medical Center in upstate New York investigated the relationship between geographic location and depression. A sample of 60 individuals, all in reasonably good health, was selected; 20 individuals were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. Each of the individuals sampled was given a standardized test to measure depression. The data collected follow; higher test scores indicate higher levels of depression.

Q: Is the depression level same in all 3 locations?

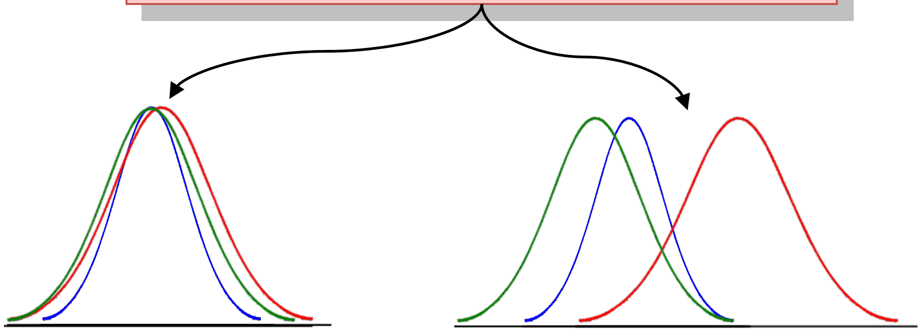
depression.txt

1. Good health respondents

Florida	New York	N. Carolina
3	8	10
7	11	7
7	9	3
3	7	5
8	8	11
8	7	8
...	...	...

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal

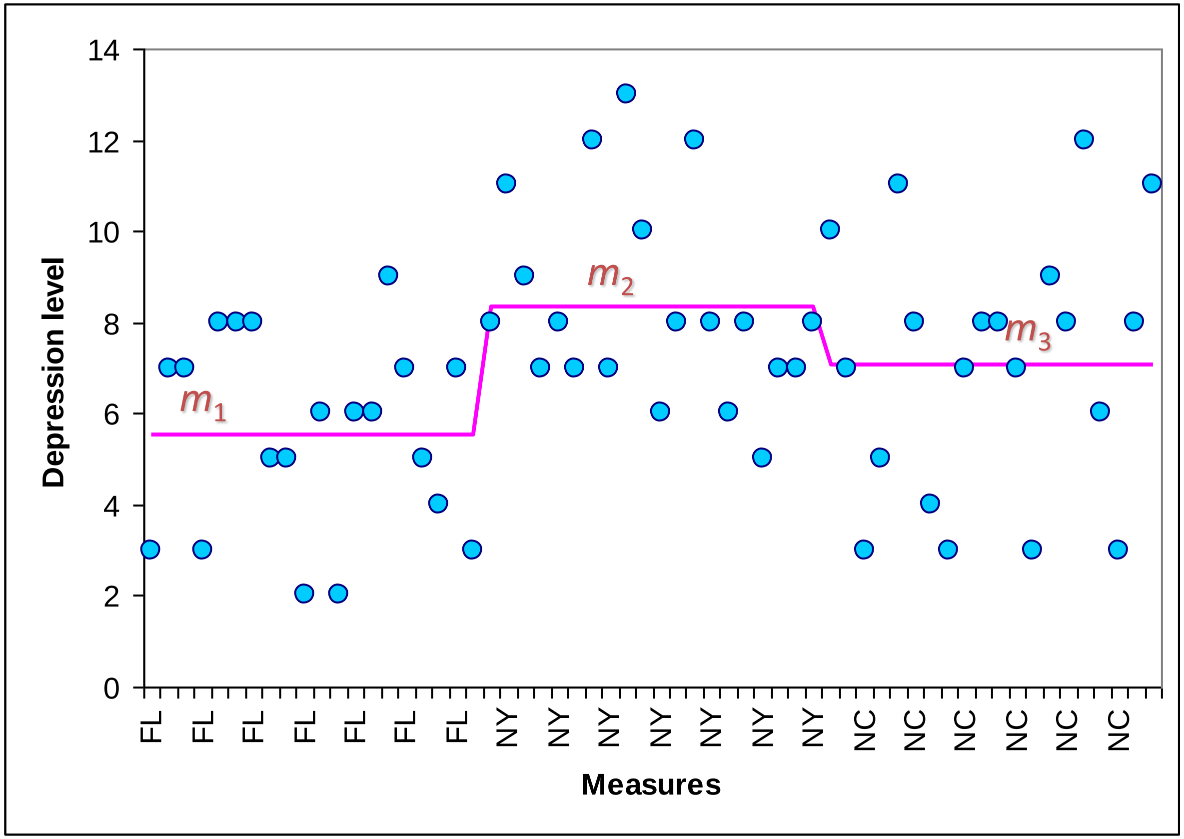


L3.1. One-way ANOVA

Meaning

H_0: $\mu_1 = \mu_2 = \mu_3$

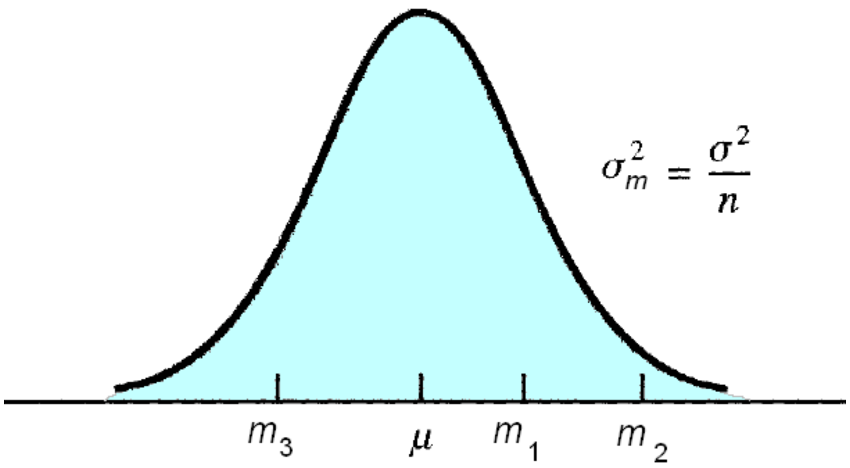
H_a: not all 3 means are equal



Assumption for ANOVA

Assumptions for Analysis of Variance

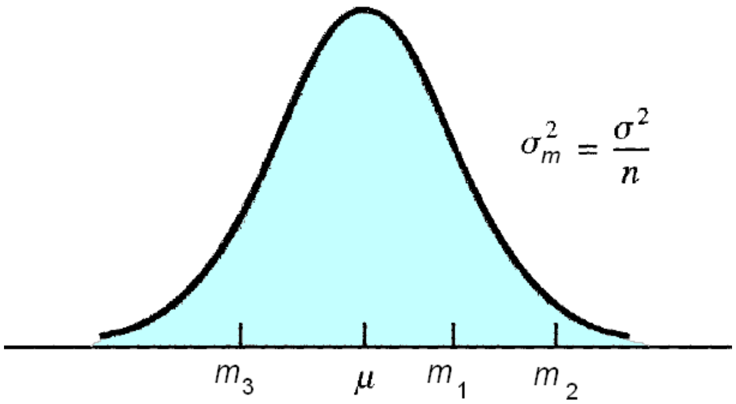
1. For each population, the response variable is normally distributed
2. The variance of the respond variable, denoted as σ^2 is the same for all of the populations.
3. The observations must be independent.



L3.1. One-way ANOVA

Some Calculations

Parameter	Florida	New York	N. Carolina
m=	5.55	8.35	7.05
overall mean=	6.98333		
var=	4.5763	4.7658	8.0500



Let's estimate the variance of sampling distribution. If H_0 is true, then all m_i belong to the same distribution

$$\sigma_m^2 = \frac{\sum_{i=1}^{k}(m_i - \overline{m})^2}{k-1} = \frac{(5.55-6.98)^2 + (8.35-6.98)^2 + (7.05-6.98)^2}{3-1} = 1.96$$

$$\sigma^2 = n\sigma_m^2 = 20 \times 1.96 = 39.27$$

– this is called between-treatment estimate, works only at H_0

At the same time, we can estimate the variance just by averaging out variances for each populations:

$$\sigma^2 = \frac{\sum_{i=1}^{k}\sigma_i}{k} = \frac{4.58+4.77+8.05}{3} = 5.8$$

– this is called within-treatment estimate

Does between-treatment estimate and within-treatment estimate give variances of the same "population"?

L3.1. One-way ANOVA

Theory

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_a: not all k means are equal

Means for treatments

$$m_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

Variances treatments

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - m_j)^2}{n_j - 1}$$

Total mean

$$\overline{m} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T}$$

$$n_T = n_1 + n_2 + \cdots + n_k$$

due to treatment

Sum squares

$$SSTR = \sum_{j=1}^{k} n_j (m_j - \overline{m})^2$$

Mean squares, $\sigma_{beetween}^2$

$$MSTR = \frac{SSTR}{k-1}$$

due to error

Sum squares

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2$$

Mean squares, σ_{within}^2

$$MSE = \frac{SSE}{n_r - k}$$

Test of variance equality

$$F = \frac{SSTR}{MSE}$$

p-value for the treatment effect

$$p - value$$

L3.1. One-way ANOVA

The Main Equation

Total sum squares

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{m}\right)^2$$

SS due to treatment

$$SSTR = \sum_{j=1}^{k} n_j \left(m_j - \overline{m}\right)^2$$

$$SST = SSTR + SSE$$

SS due to error

$$SSE = \sum_{j=1}^{k} \left(n_j - 1\right) s_j^2$$

Total variability of the data include variability due to treatment and variability due to error

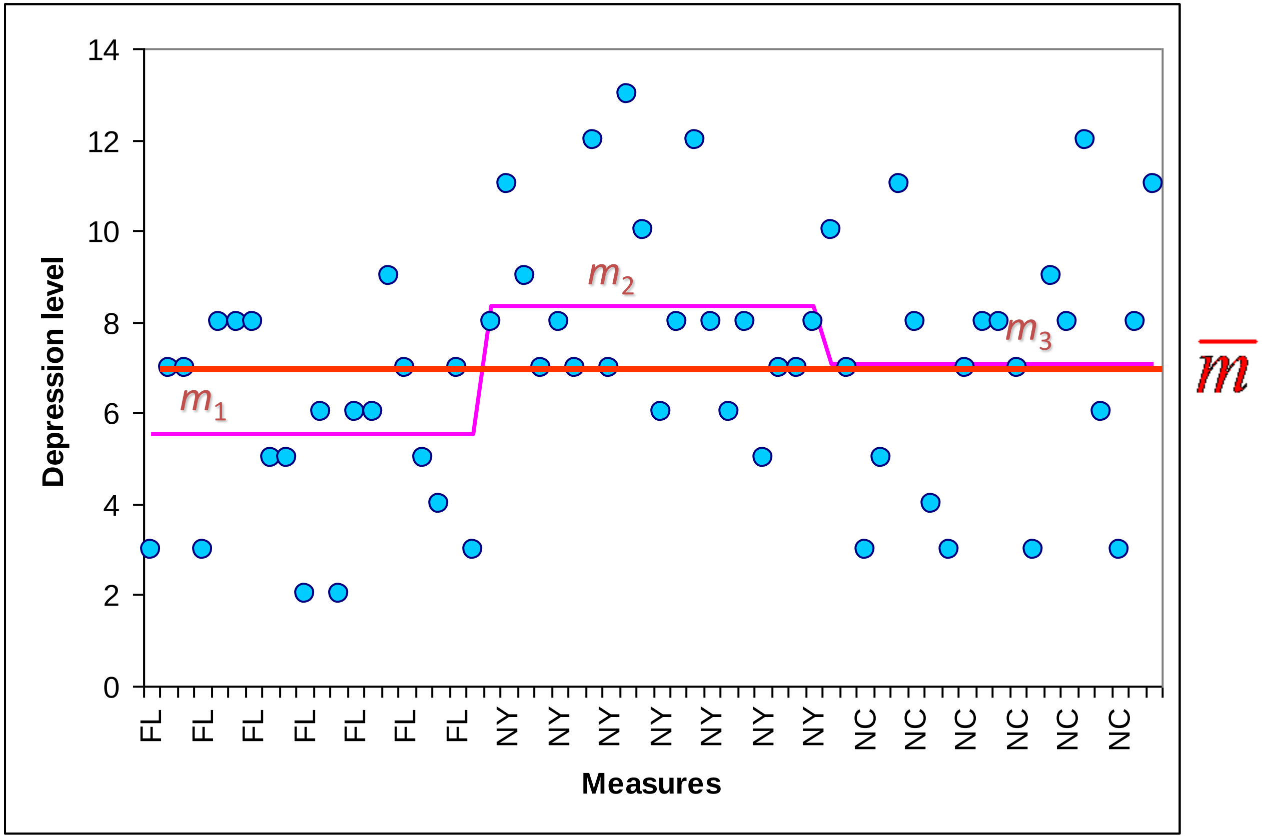
$$d.f.(SST) = d.f.(SSTR) + d.f.(SSE)$$
$$n_T - 1 = (k-1) + (n_T - k)$$

Partitioning

The process of allocating the total sum of squares and degrees of freedom to the various components.

L3.1. One-way ANOVA

Example



$$SST = SSTR + SSE$$

L3.1. ANOVA

Example

ANOVA table

A table used to summarize the analysis of variance computations and results. It contains columns showing the source of variation, the sum of squares, the degrees of freedom, the mean square, and the *F* value(s).

Let's perform for dataset 1: "good health"

`depression2.txt`

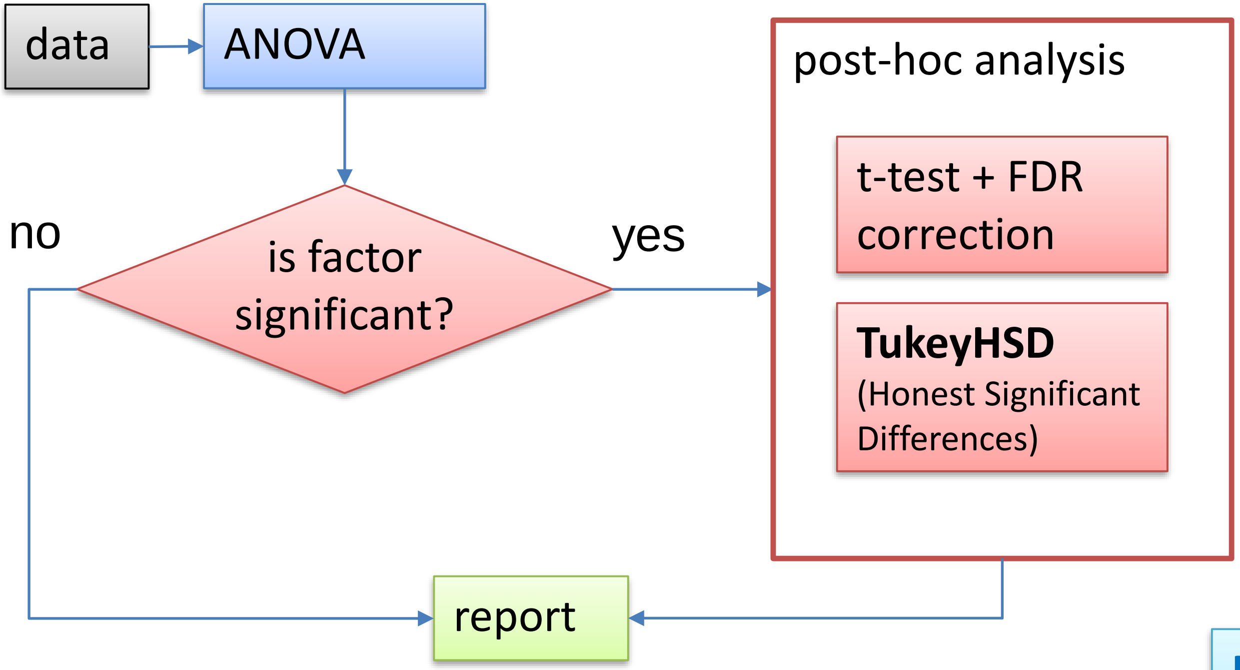
In R use:

- `aov(...)`

```
             Df Sum Sq Mean Sq F value Pr(>F)
Location      2   78.5   39.27   6.773 0.0023 **
Residuals    57  330.4    5.80
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-hoc Analysis

Post-hoc analysis
allows for additional exploration of significant differences in the data, when significant effect of the factor was already confirmed (for example, by ANOVA).



In R use:

- `TukeyHSD(...)`

L3.1. Multi-factor ANOVA

Some Definitions

Factor
Another word for the independent variable of interest.

Factorial experiment
An experimental design that allows statistical conclusions about two or more factors.

Treatments
Different levels of a factor.

`depression2.txt`

Factor 1: Health → good health, bad health

Factor 2: Location → Florida, New York, North Carolina

$$\text{Depression} = \mu + \text{Health} + \text{Location} + \text{Health}\times\text{Location} + \varepsilon$$

Interaction
The effect produced when the levels of one factor interact with the levels of another factor in influencing the response variable.

ANOVA Table

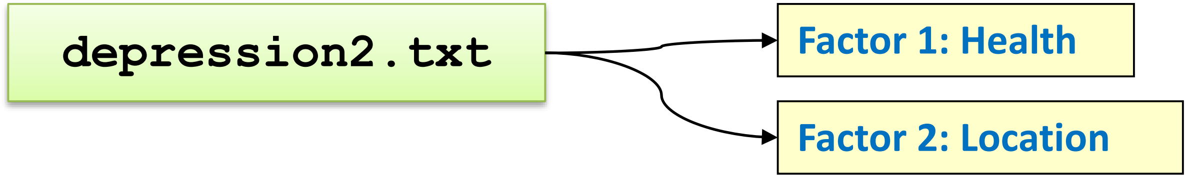
Replications
The number of times each experimental condition is repeated in an experiment.

a = number of levels of factor A
b = number of levels of factor B
r = number of replications
n_T = total number of observations taken in the experiment; $n_T = abr$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Factor A	SSA	$a - 1$	$\text{MSA} = \frac{\text{SSA}}{a-1}$	$\frac{\text{MSA}}{\text{MSE}}$
Factor B	SSB	$b - 1$	$\text{MSB} = \frac{\text{SSB}}{b-1}$	$\frac{\text{MSB}}{\text{MSE}}$
Interaction	SSAB	$(a - 1)(b - 1)$	$\text{MSAB} = \frac{\text{SSAB}}{(a-1)(b-1)}$	$\frac{\text{MSAB}}{\text{MSE}}$
Error	SSE	$ab(r - 1)$	$\text{MSE} = \frac{\text{SSE}}{ab(r-1)}$	
Total	SST	$n_T - 1$		

L3.1. Multi-factor ANOVA

Example

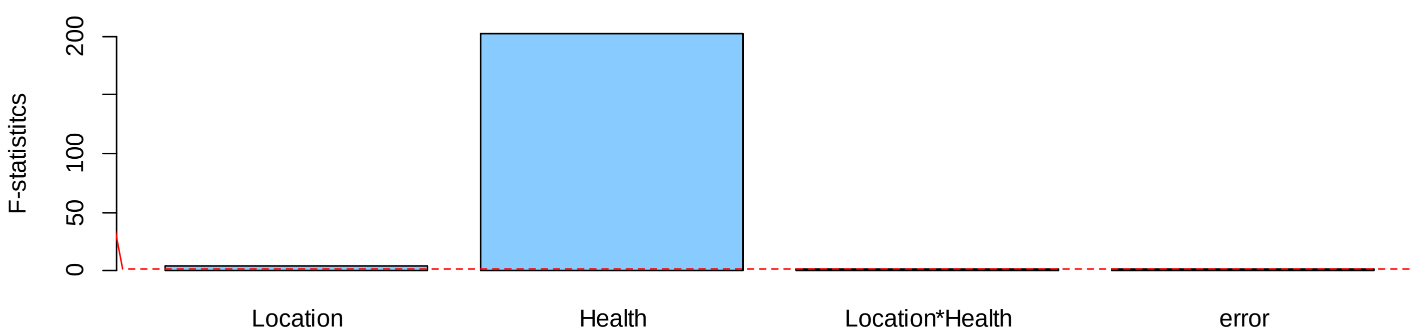



```
res = aov(Depression ~  Location + Health + Location*Health, Data)
summary(res)
source("http://sablab.net/scripts/drawANOVA.r")
drawANOVA(res)
```

ANOVA

ANOVA model: Depression = m + Location + Health + Location * Health + e

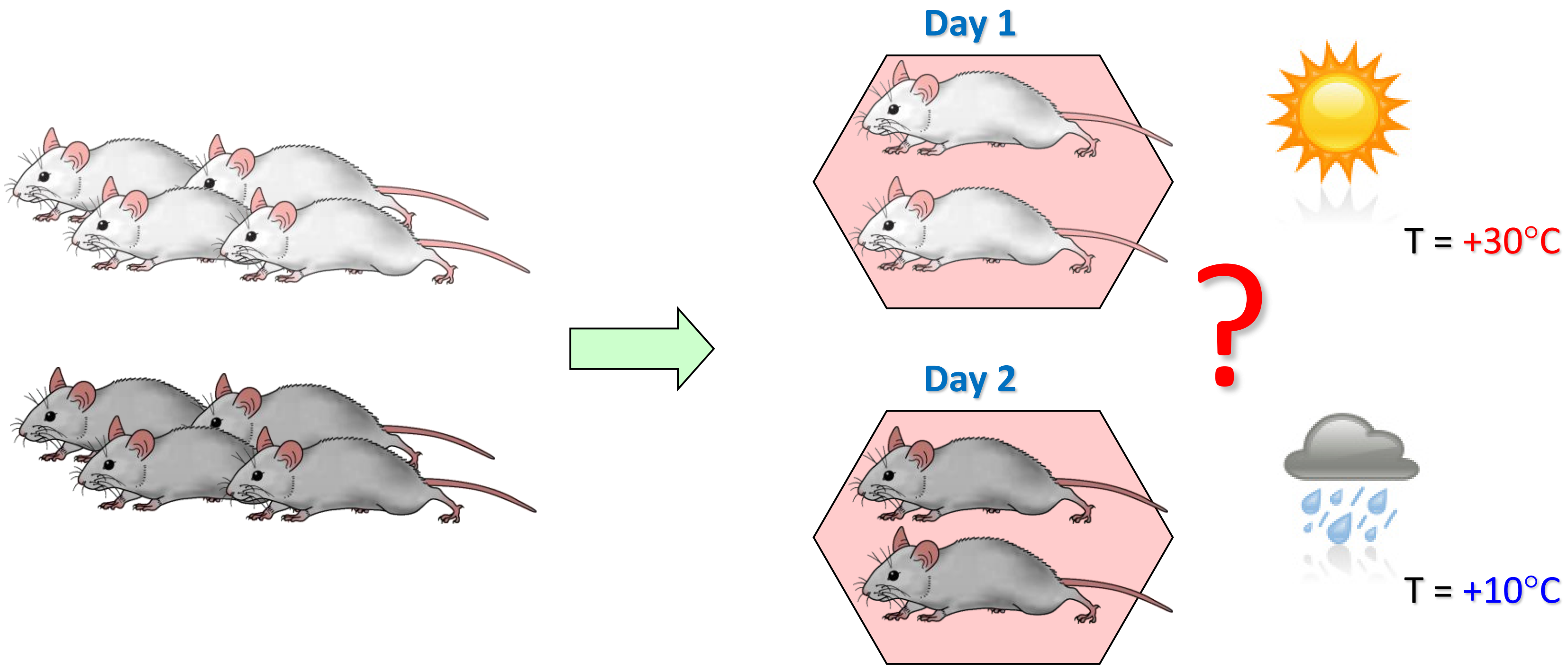
Factor	Df	Sum Sq	Mean Sq	F value	p-value	
Location	2	73.85	36.93	4.29	1.5981e-02	*
Health	1	1748.03	1748.03	203.094	4.3961e-27	***
Location:Health	2	26.12	13.06	1.517	2.2373e-01	



Experimental Design

Aware of Batch Effect !

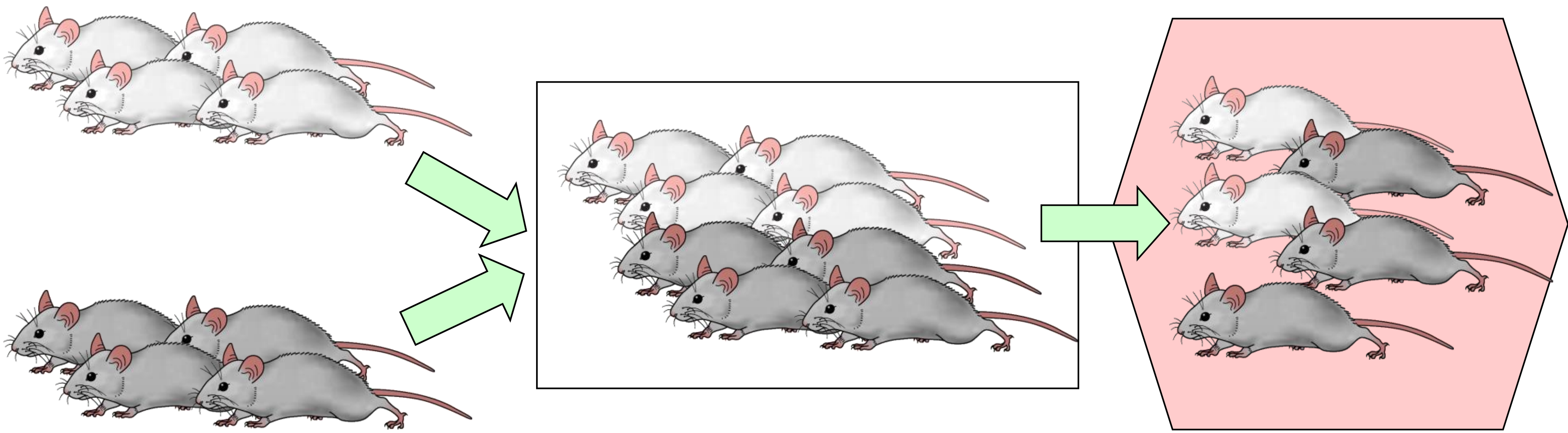
When designing your experiment always remember about various factors which can effect your data: batch effect, personal effect, lab effect...



L3.1. Multi-factor ANOVA

Experimental Design

Completely randomized design
An experimental design in which the treatments are randomly assigned to the experimental units.



We can nicely randomize:

Day effect

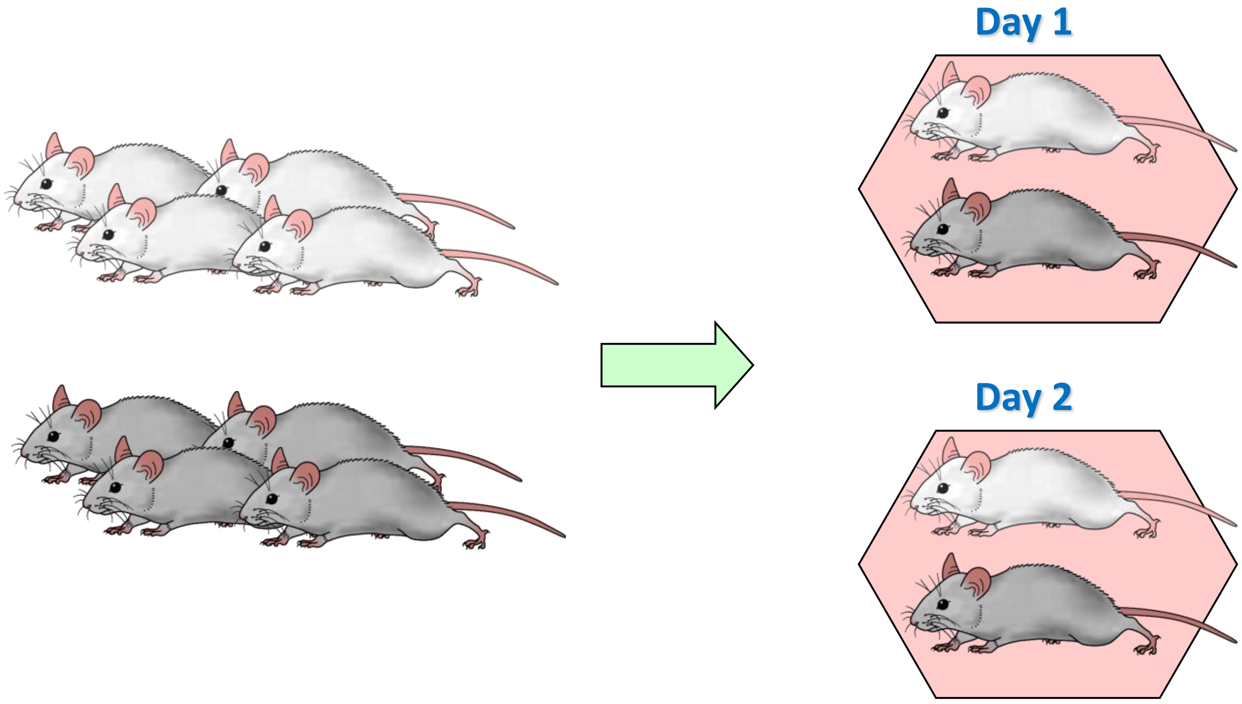
Batch effect

L3.1. Multi-factor ANOVA

Experimental Design

Blocking

The process of using the same or similar experimental units for all treatments. The purpose of blocking is to remove a source of variation from the error term and hence provide a more powerful test for a difference in population or treatment means.



L3.1. Multi-factor ANOVA

A good suggestion… ☺

Block what you can block, randomize what you cannot, and try to avoid unnecessary factors

L3.1. ANOVA

Please go through the code at:

http://edu.sablab.net/abs2017/scripts3.html

Section 3.1

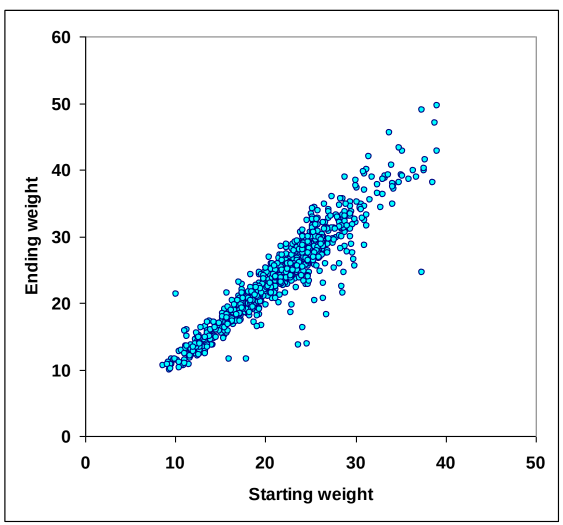
Do Exercises 3.1

L3.2. Linear Regression

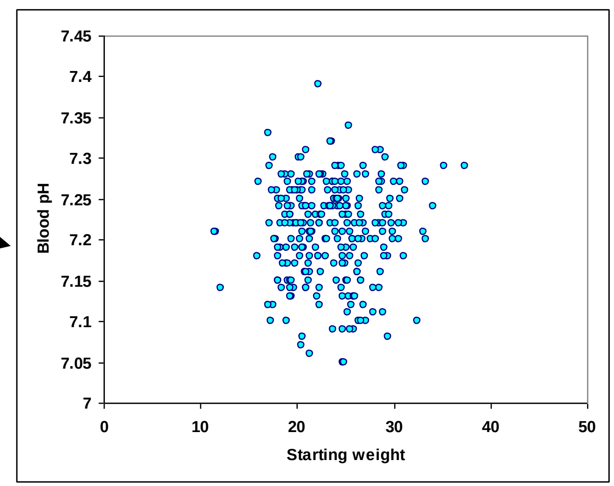
Dependent and Independent Variables

mice.xls

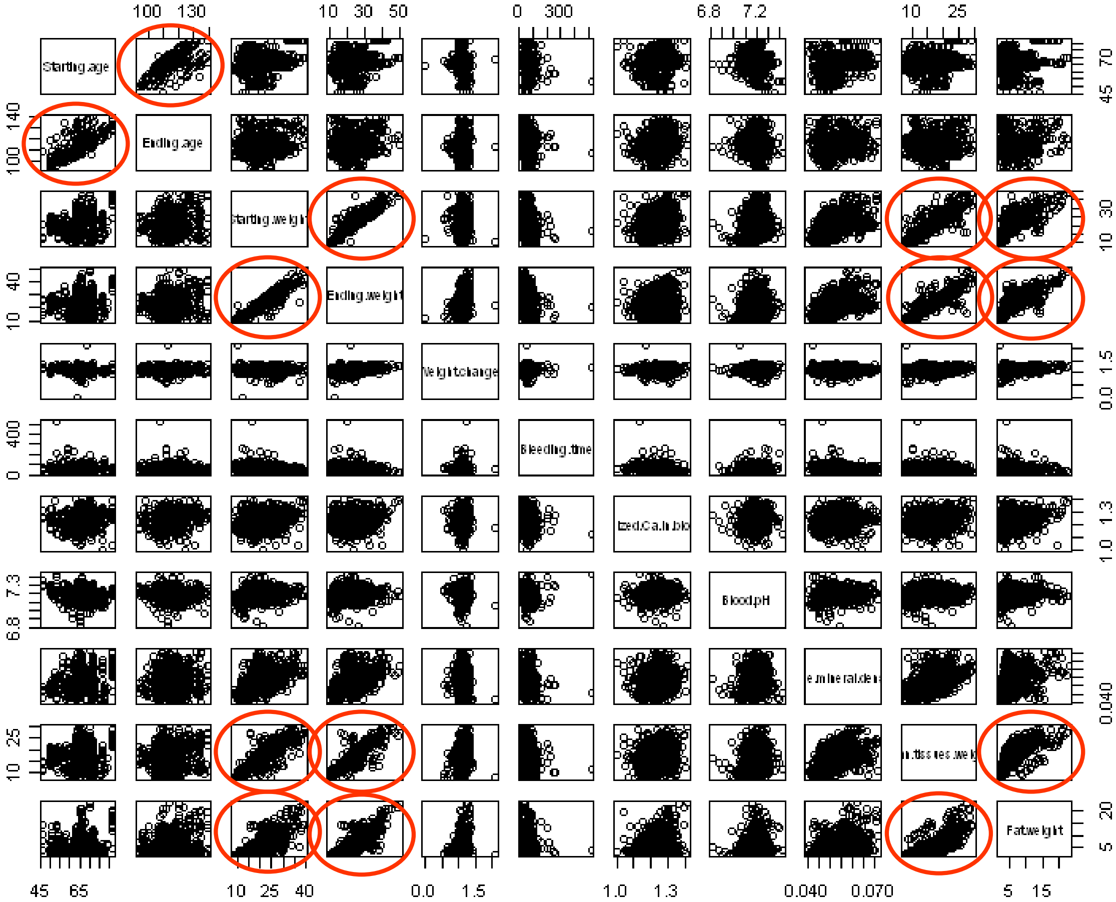
Ending weight vs. Starting weight



Blood pH vs. Starting weight



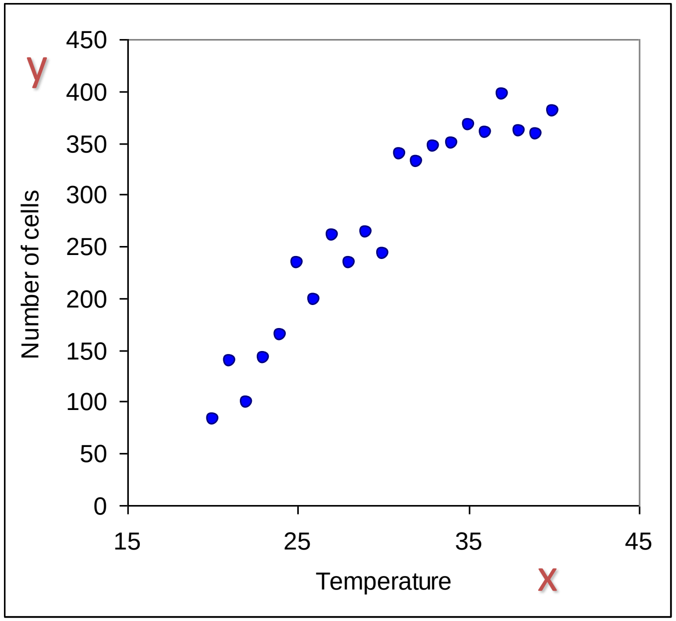
Dependent and Independent Variables



L3.2. Linear Regression

Example

Temperature	Cell Number
20	83
21	139
22	99
23	143
24	164
25	233
26	198
27	261
28	235
29	264
30	243
31	339
32	331
33	346
34	350
35	368
36	360
37	397
38	361
39	358
40	381



Cells are grown under different temperature conditions from 20° to 40°. A researched would like to find a dependency between T and cell number.

`cells.txt`

Dependent variable

The variable that is being predicted or explained. It is denoted by *y*.

Independent variable

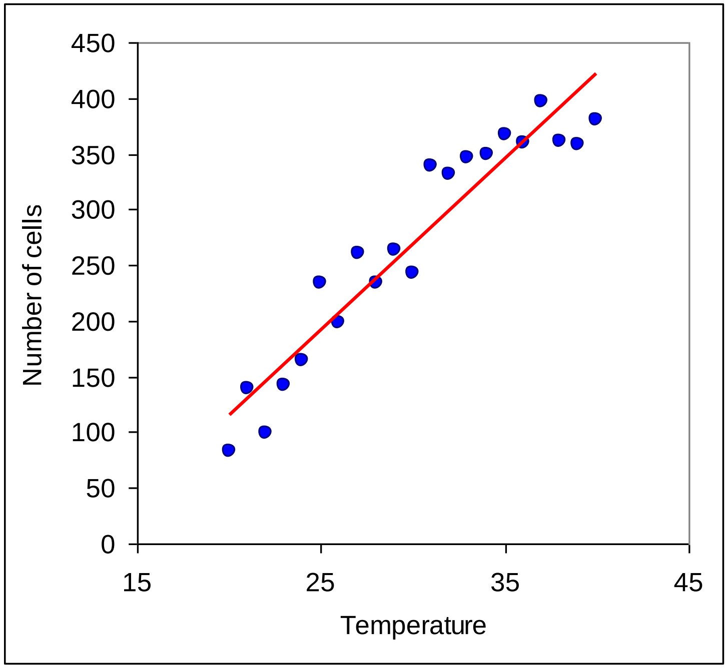
The variable that is doing the predicting or explaining. It is denoted by *x*.

Regression Model and Regression Line

Simple linear regression

Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line.

- Building a *regression* means finding and tuning the model to explain the behaviour of the data

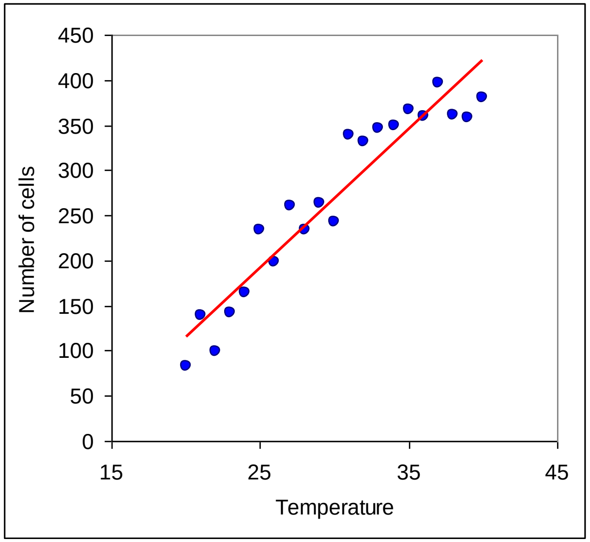


L3.2. Linear Regression

Regression Model and Regression Line

Regression model
The equation describing how y is related to x and an error term; in simple linear regression, the regression model is $y = \beta_0 + \beta_1 x + \varepsilon$

Regression equation
The equation that describes how the mean or expected value of the dependent variable is related to the independent variable; in simple linear regression,
$E(y) = \beta_0 + \beta_1 x$

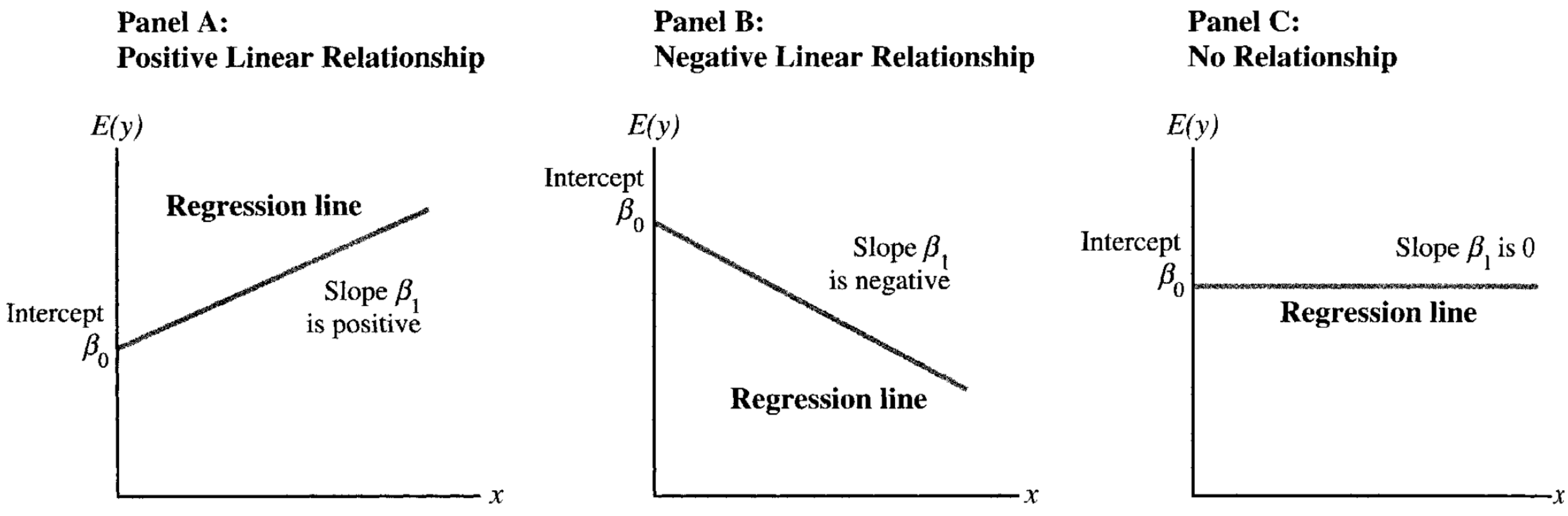


- Model for a simple linear regression:

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

Regression Model and Regression Line

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



L3.2. Linear Regression

Estimation

Estimated regression equation

The estimate of the regression equation developed from sample data by using the least squares method. For simple linear regression, the estimated regression equation is $y = b_0 + b_1 x$

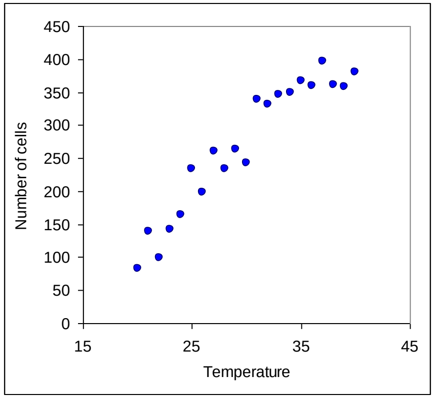
$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

$$\hat{y}(x) = b_1 x + b_0$$

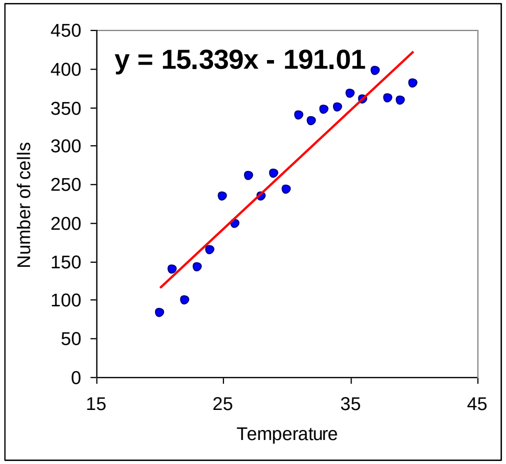
$$E[y(x)] = b_1 x + b_0$$

`cells.xls`

1. Make a scatter plot for the data.



2. Right click to "Add Trendline". Show equation.



Slope and Intercept

Least squares method

A procedure used to develop the estimated regression equation.

The objective is to minimize $\sum (y_i - \hat{y}_i)^2$

y_i = observed value of the dependent variable for the ith observation

$\hat{y}_i$ = estimated value of the dependent variable for the ith observation

Slope:

$$b_1 = \frac{\sum (x_i - m_x)(y_i - m_y)}{(x_1 - m_x)^2}$$

Intercept:

$$b_0 = m_y - b_1 m_x$$

Coefficient of Determination

Sum squares due to error

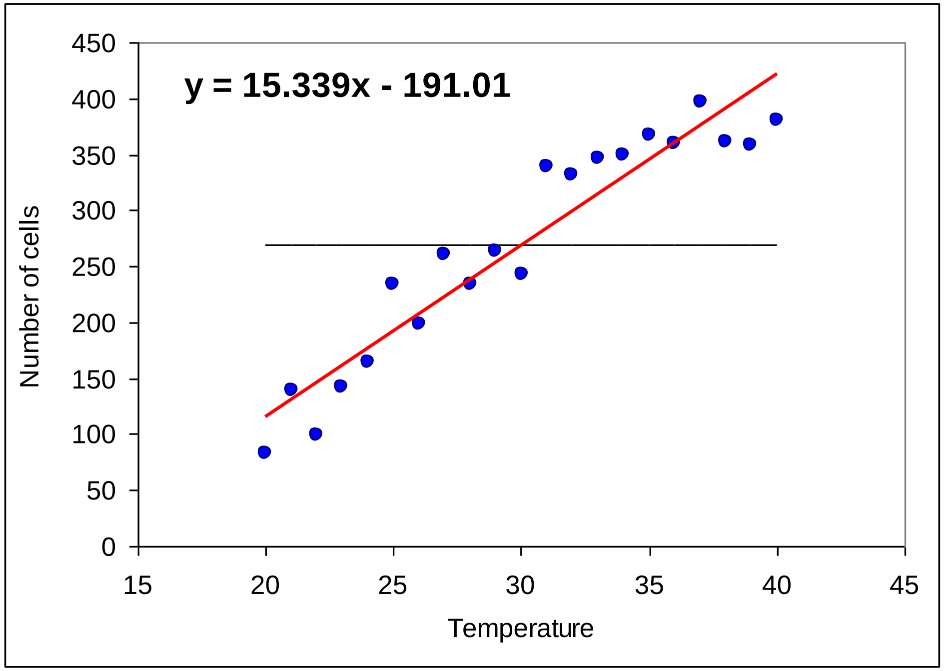
$$SSE = \sum (y_i - \hat{y}_i)^2$$

Sum squares total

$$SST = \sum (y_i - m_y)^2$$

Sum squares due to regression

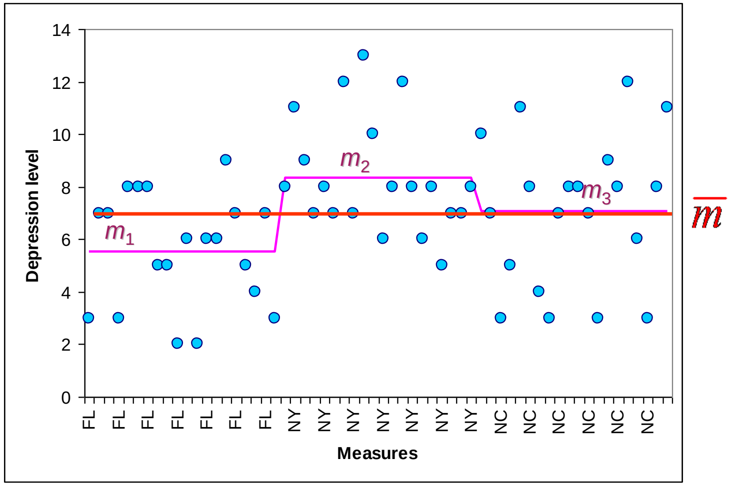
$$SSR = \sum (\hat{y}_i - m_y)^2$$



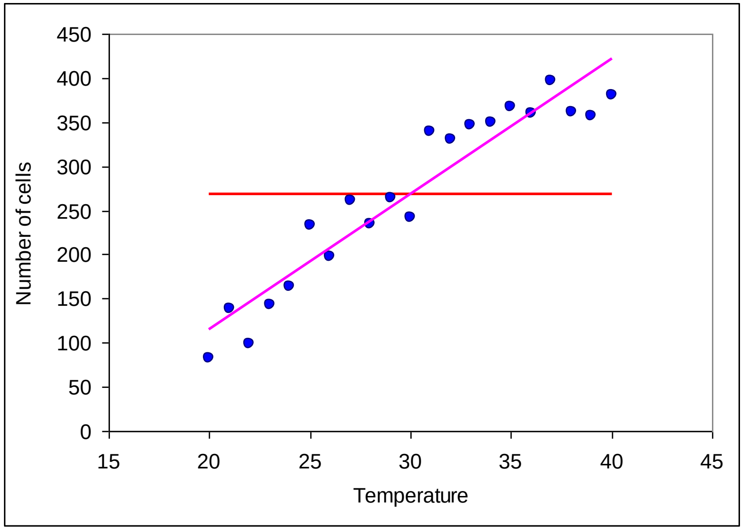
The Main Equation

$$SST = SSR + SSE$$

ANOVA and Regression



$$SST = SSTR + SSE$$



$$SST = SSR + SSE$$

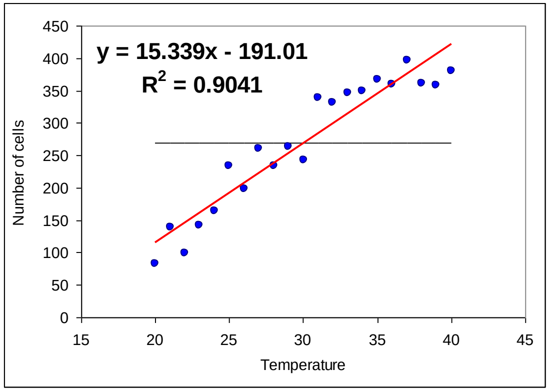
Coefficient of Determination

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SST = \sum (y_i - m_y)^2$$

$$SSR = \sum (\hat{y}_i - m_y)^2$$

$$SST = SSR + SSE$$



Coefficient of determination

A measure of the goodness of fit of the estimated regression equation. It can be interpreted as the proportion of the variability in the dependent variable *y* that is explained by the estimated regression equation.

$$R^2 = \frac{SSR}{SST}$$

Correlation coefficient

A measure of the strength of the linear relationship between two variables (previously discussed in Lecture 1).

$$r = \text{sign}(b_1)\sqrt{R^2}$$

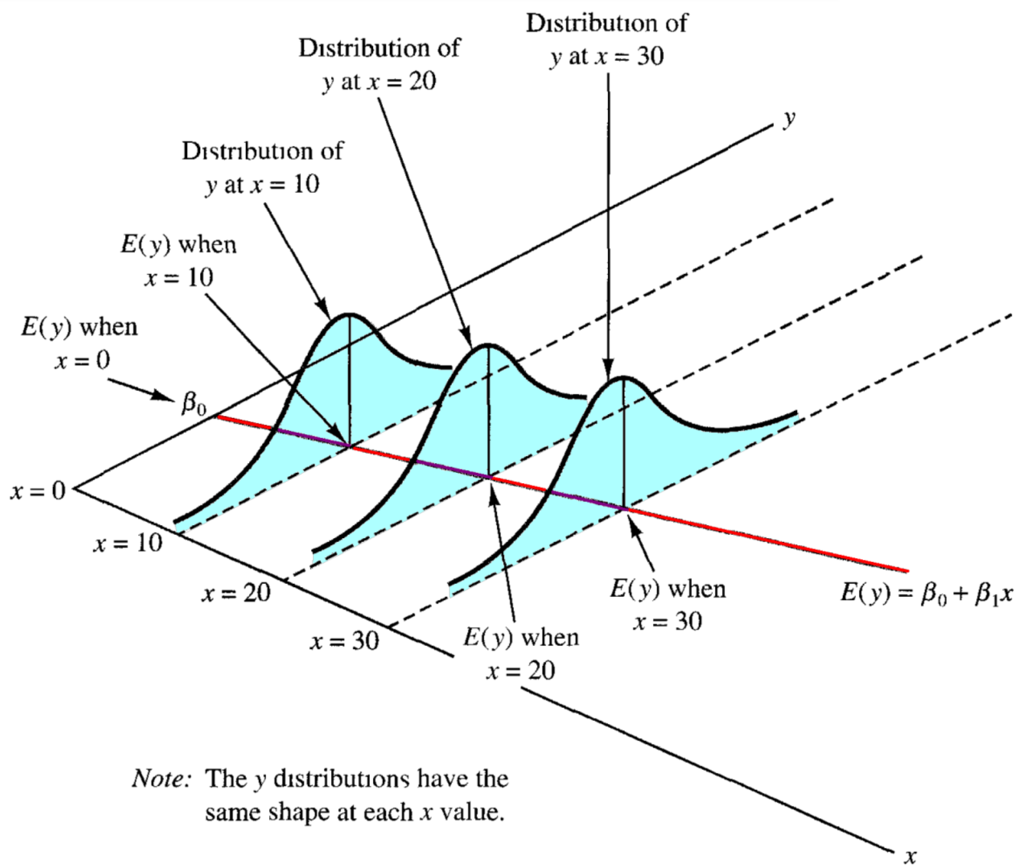
L3.2. Linear Regression

Assumptions

Assumptions for Simple Linear Regression

1. The error term ε is a random variable with 0 mean, i.e. E[ε]=0
2. The variance of ε, denoted by σ^2, is the same for all values of x
3. The values of ε are independent
3. The term ε is a normally distributed variable

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



Note: The y distributions have the same shape at each x value.

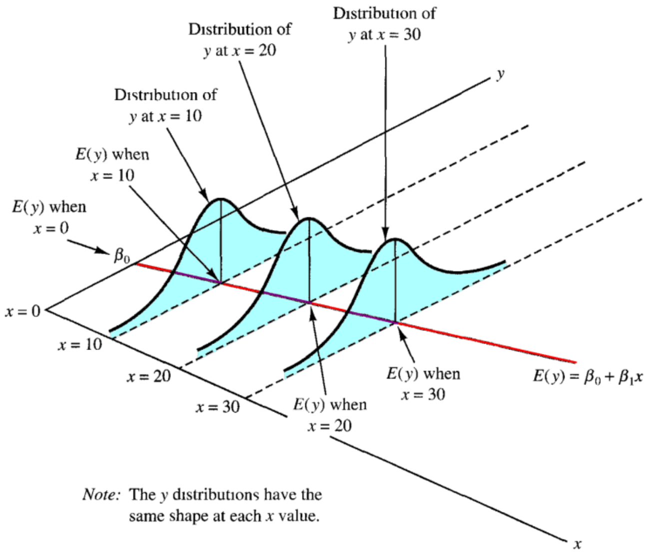
Estimation of σ^2

i-th residual

The difference between the observed value of the dependent variable and the value predicted using the estimated regression equation; for the *i*-th observation the *i*-th residual is: $y_i - \hat{y}_i$

Mean square error

The unbiased estimate of the variance of the error term σ^2. It is denoted by MSE or s^2.
Standard error of the estimate: the square root of the mean square error, denoted by *s*. It is the estimate of σ, the standard deviation of the error term ε.



$$s^2 = MSE = \frac{SSE}{n-2}$$

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

Test for Significance

H_0: $\beta_1 = 0$ *insignificant*

H_a: $\beta_1 \neq 0$

1. Build a t-test statistics.

$$t = \frac{b_1}{\sigma_{b_1}} = \frac{b_1}{s}\sqrt{\sum (x_i - m_x)^2}$$

2. Calculate p-value for *t*

p-value approach: Reject H_0 if p-value $\leq \alpha$

Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

1. Build a F-test statistics.

$$F = \frac{MSR}{MSE}$$

$$\text{MSR} = \frac{\text{SSR}}{\text{Number of independent variables}}$$

2. Calculate a p-value

L3.2. Linear Regression

Example

cells.xls

1. Calculate manually b_1 and b_0

Intercept	b0=	-191.008119
Slope	b1=	15.3385723

In Excel use the function:

- = INTERCEPT(y,x)
- = SLOPE(y,x)

2. Let's do it automatically   Data → Data Analysis → Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950842308
R Square	0.904101095
Adjusted R Square	0.899053784
Standard Error	31.80180903
Observations	21

ANOVA

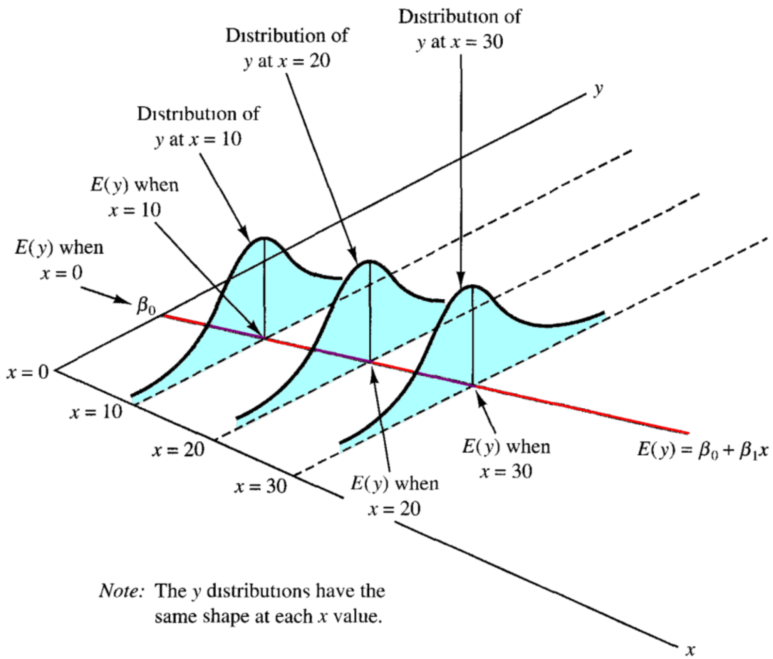
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	181159.2853	181159.3	179.1253	4.01609E-11
Residual	19	19215.7461	1011.355		
Total	20	200375.0314			

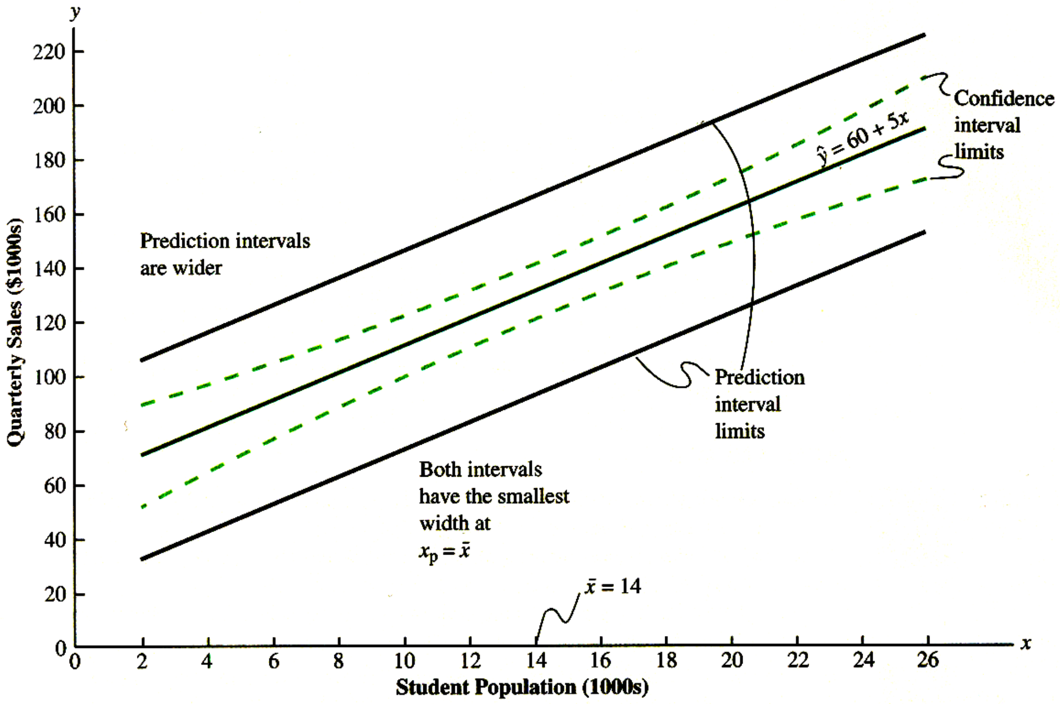
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-191.0081194	35.07510626	-5.445689	2.97E-05	-264.4211603	-117.5950784	-264.4211603	-117.5950784
X Variable 1	15.33857226	1.146057646	13.38377	4.02E-11	12.93984605	17.73729848	12.93984605	17.73729848

Confidence and Prediction

Confidence interval
The interval estimate of the mean value of y for a given value of x.

Prediction interval
The interval estimate of an individual value of y for a given value of x.

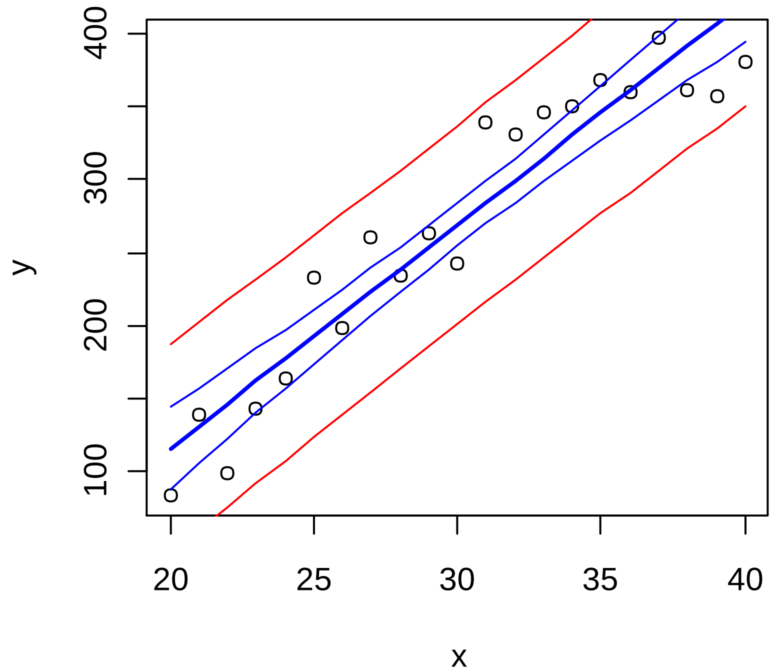




L3.2. Linear Regression

Example

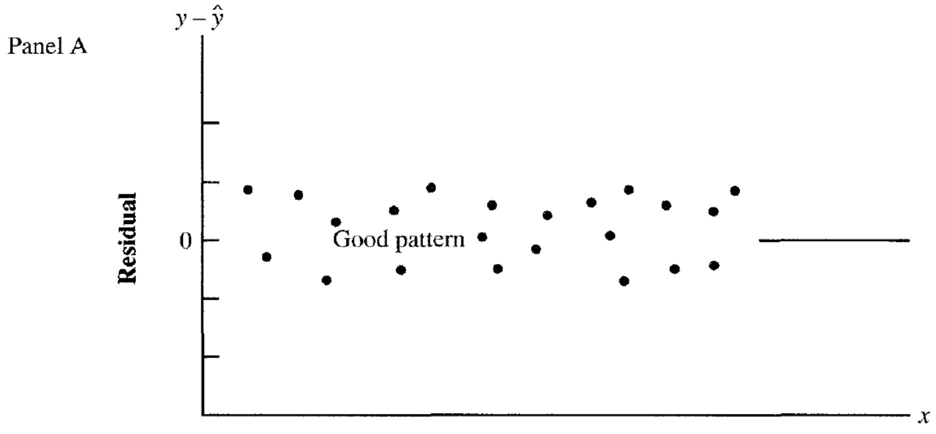
cells.txt

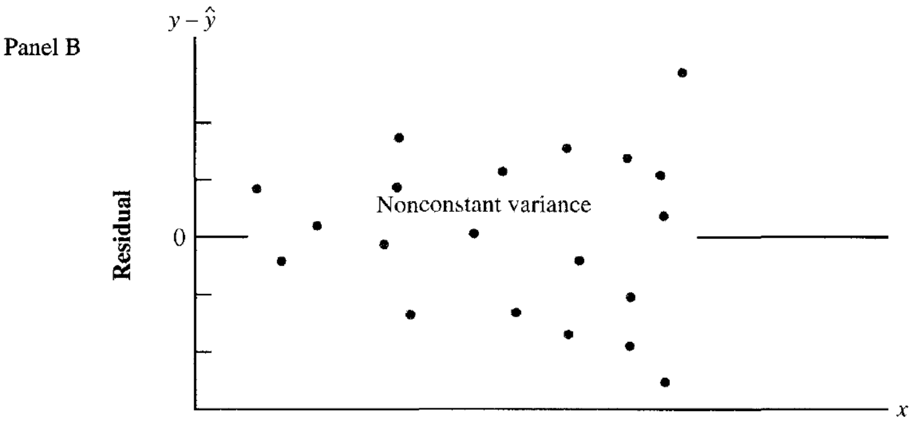


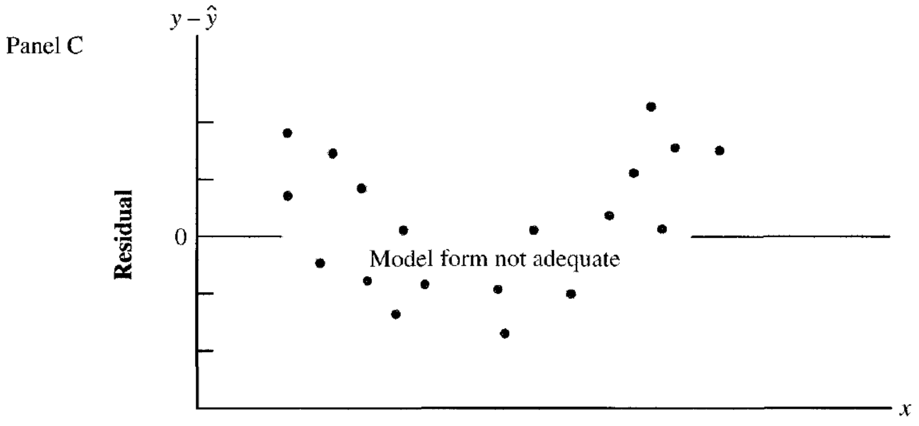

```
x = data$Temperature
y = data$Cell.Number
res = lm(y~x)
res
summary(res)

# draw the data
x11()
plot(x,y)
# draw the regression and its confidence (95%)
lines(x, predict(res,int = "confidence")[,1],col=4,lwd=2)
lines(x, predict(res,int = "confidence")[,2],col=4)
lines(x, predict(res,int = "confidence")[,3],col=4)
# draw the prediction for the values (95%)
lines(x, predict(res,int = "pred")[,2],col=2)
lines(x, predict(res,int = "pred")[,3],col=2)
```

L3.2. Linear Regression

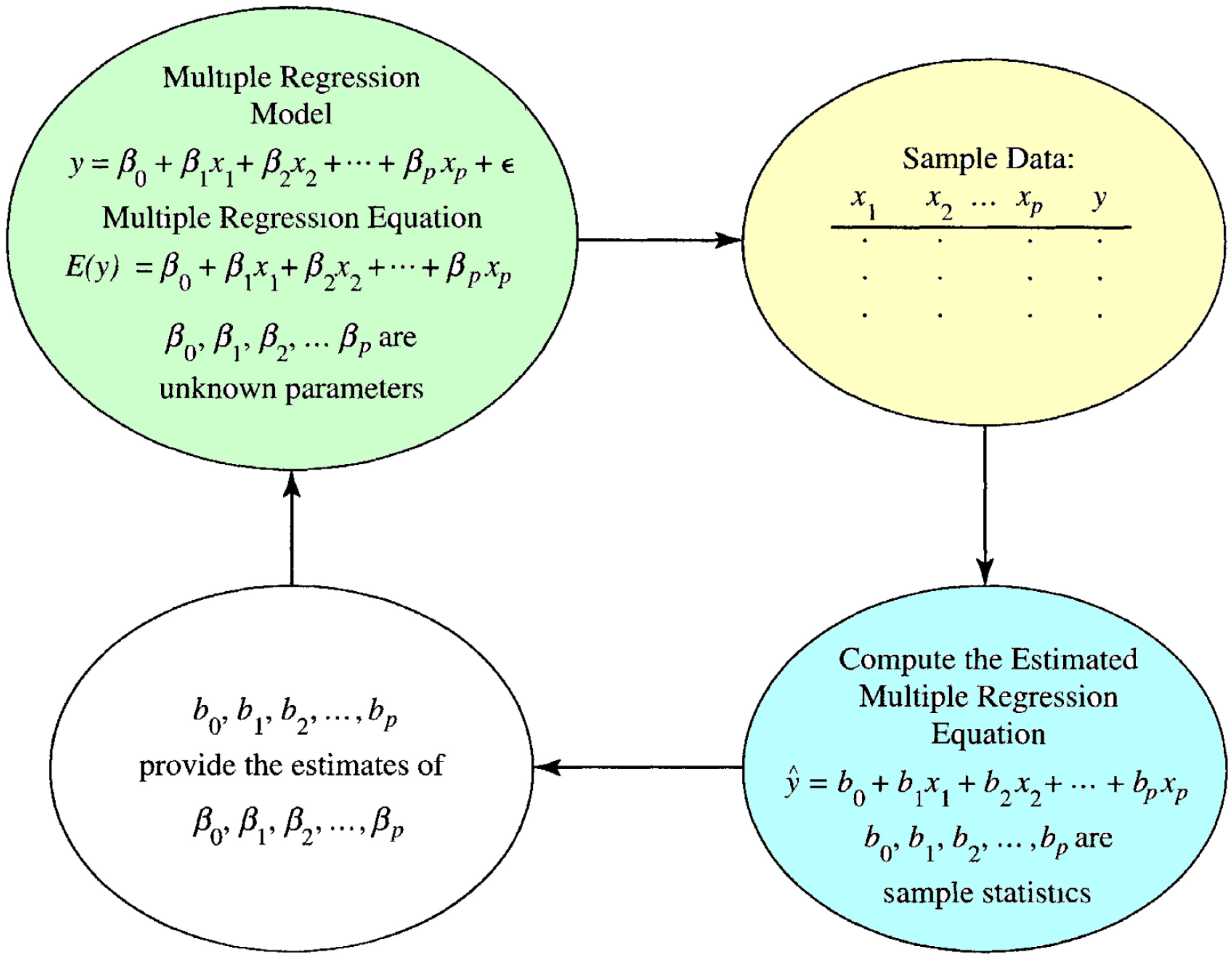
Residuals





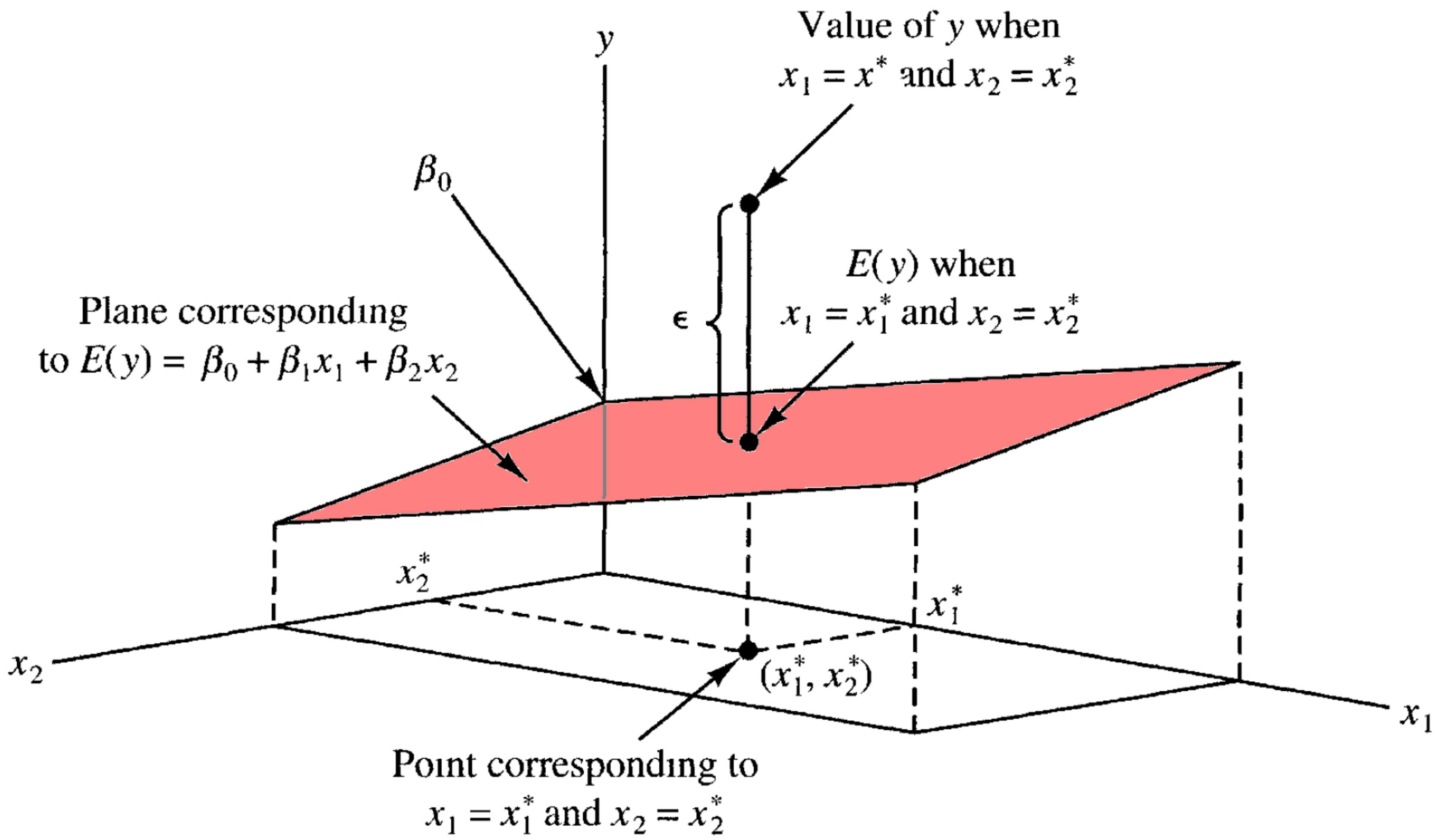


Multiple Regression



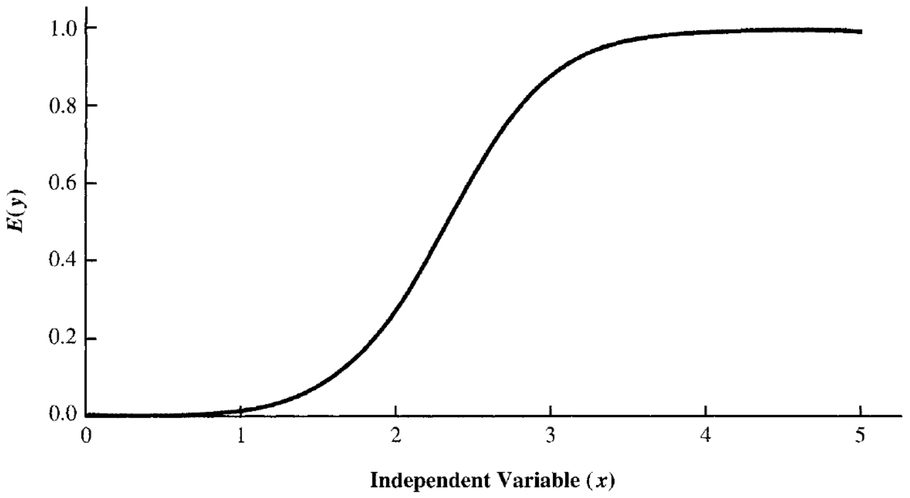
L3.2. Linear Regression

Multiple Regression



Non-Linear Regression

FIGURE 15.12 LOGISTIC REGRESSION EQUATION FOR $\beta_0 = -7$ AND $\beta_1 = 3$



$$E(y) = P(y = 1 \mid x_1, x_2, \ldots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}$$

UNIVERSITÉ DU LUXEMBOURG

L3.2. Linear Regression

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Please go through the code at:

http://edu.sablab.net/abs2017/scripts3.html

Section 3.2

Do Exercises 3.2

Questions ?

Thank you for your attention

to be continued...