UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE
Multiomics Data Science

BIOSTATISTICS

Lecture 11

Linear Regression

Petr Nazarov

Email: petr.nazarov@lih.lu
Skype: pvn.public

30-04-2021

OUTLINE

Lecture 11

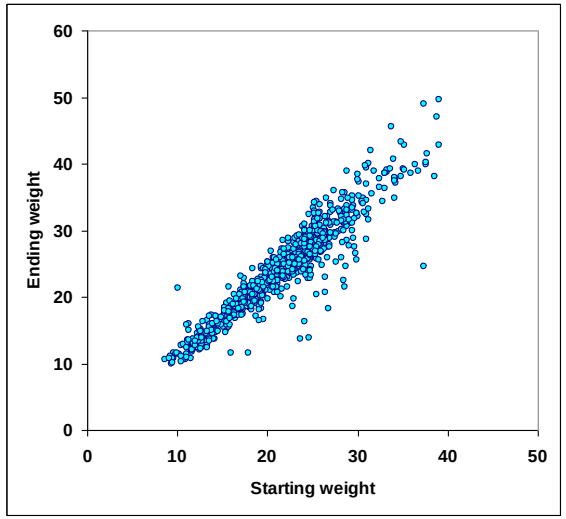
- **Introduction**
  - correlation measures
  - dependent and independent random variables
  - hypotheses about correlation
  - Fisher's transformation
- **Testing for significance**
  - linear models
  - estimation of the noise variance
  - interval estimations for coefficient
  - testing hypothesis about significance
- **Regression Analysis**
  - confidence and prediction
  - multiple linear regression
  - nonlinear regression

CORRELATION

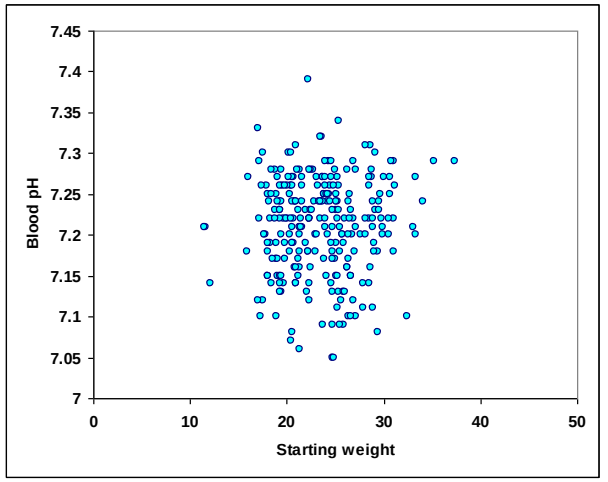
Dependent and Independent Variables

mice.xls

Ending weight vs. Starting weight

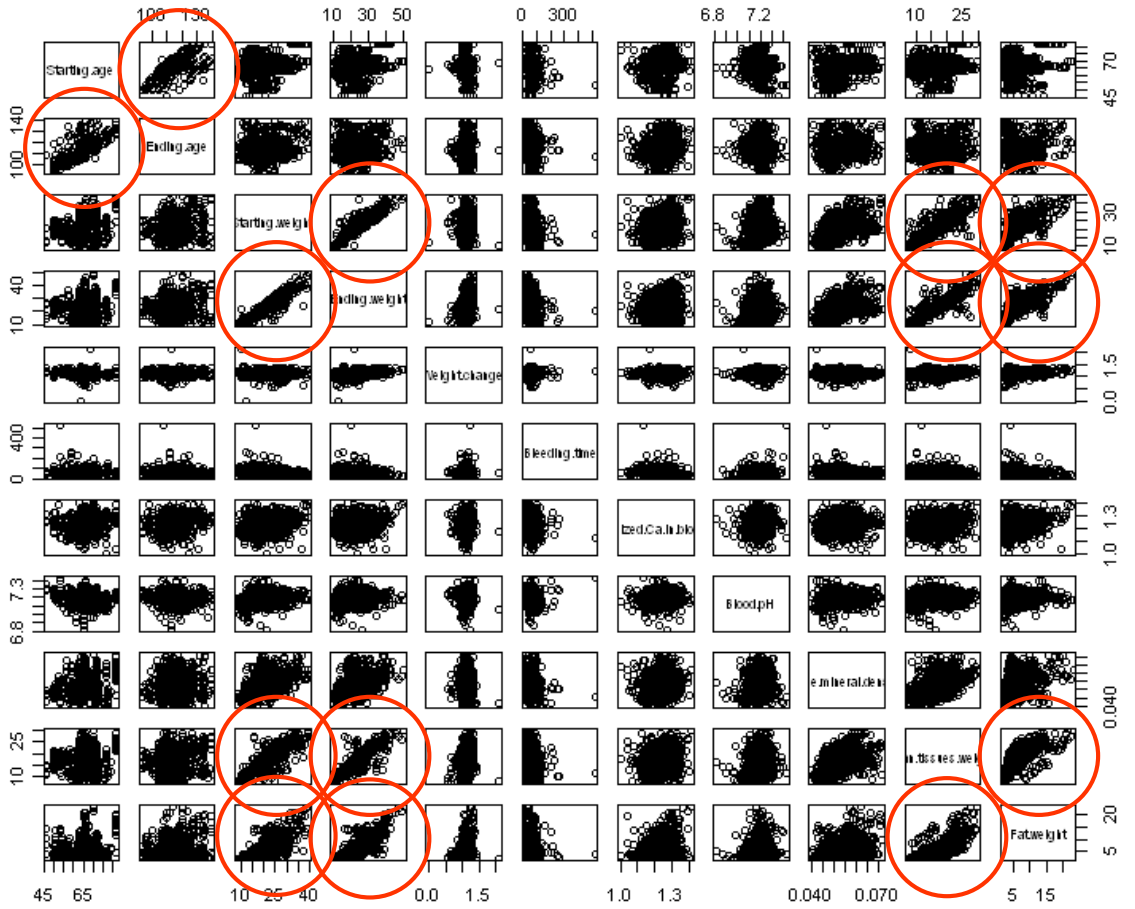


Blood pH vs. Starting weight



CORRELATION

Dependent and Independent Variables



CORRELATION

Measure of Association between 2 Variables

Covariance

A measure of linear association between two variables. Positive values indicate a positive relationship; negative values indicate a negative relationship.

population

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

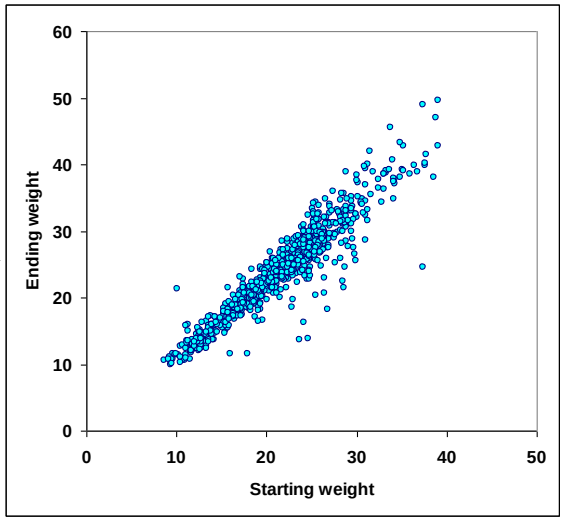
sample

$$s_{xy} = \frac{\sum (x_i - m_x)(y_i - m_y)}{n-1}$$

`= COVAR(data)`

`cov(data)`

mice.xls



Ending weight vs. Starting weight

s_{xy} = 39.8

hard to interpret

CORRELATION

Measure of Association between 2 Variables

Correlation (Pearson product moment correlation coefficient)

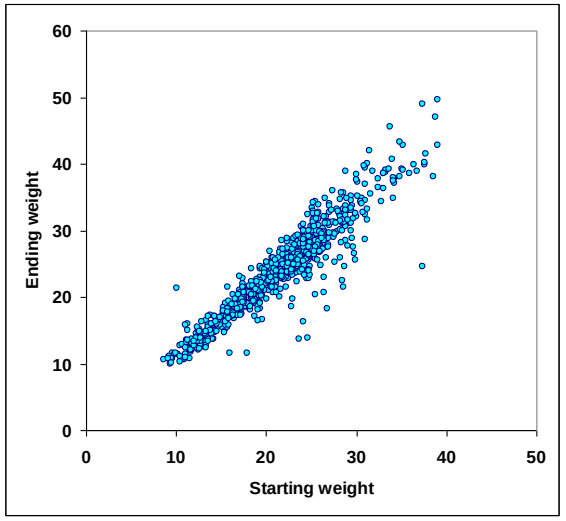
A measure of linear association between two variables that takes on values between -1 and +1. Values near +1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship; and values near zero indicate the lack of a linear relationship.

population

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y N}$$

sample

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - m_x)(y_i - m_y)}{s_x s_y (n-1)}$$



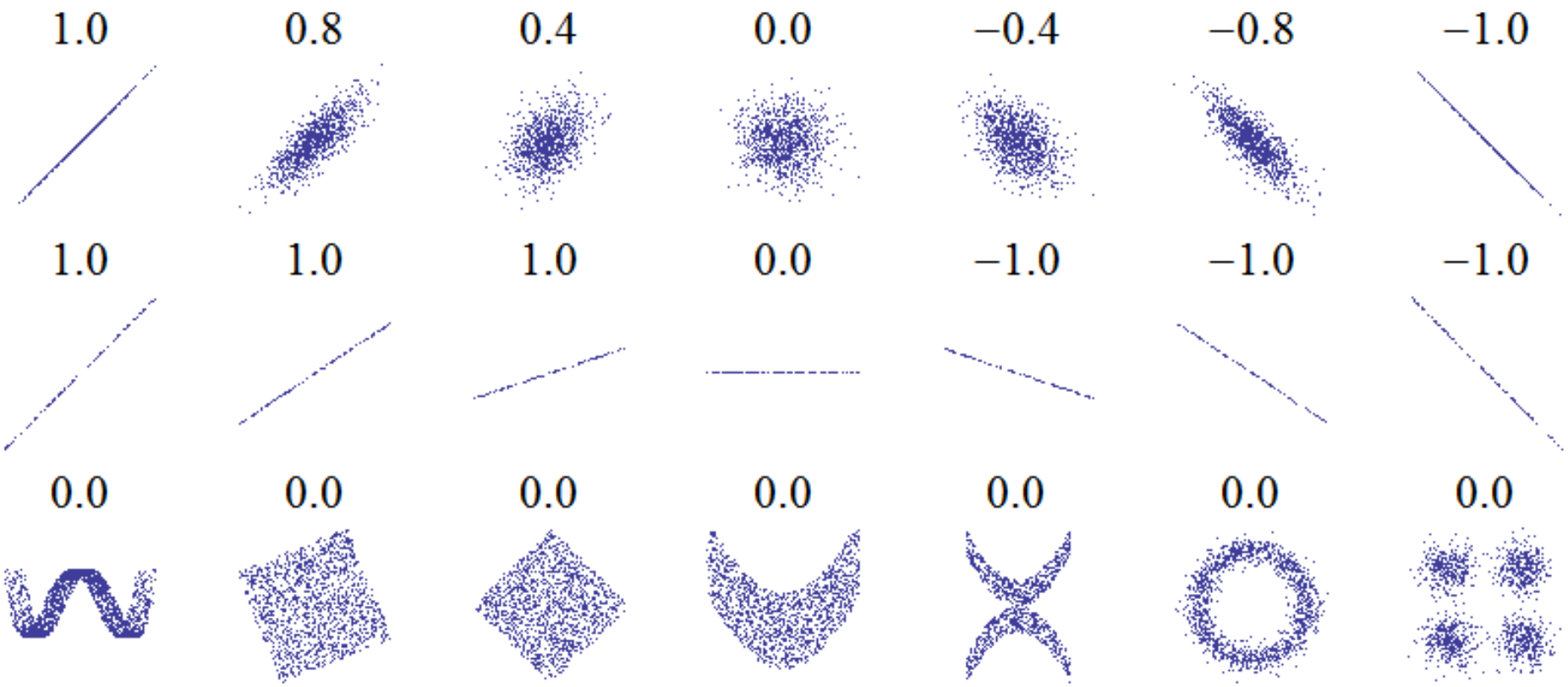
r_{xy} = 0.94

```
= CORREL(data)
```

```
cor(data)
cor(data,method="pearson",
    use="pairwise.complete.obs")
```

CORRELATION

Correlation Coefficient



Wikipedia

? If we have only 2 data points in *x* and y datasets, what values would you expect for correlation b/w *x* and y ?

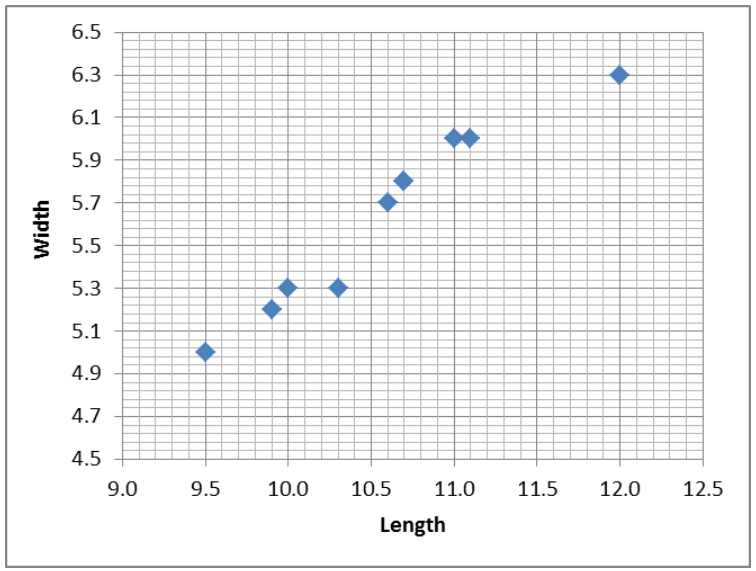
CORRELATION

Test for Significance of Correlation

A malacologist interested in the morphology of West Indian chitons, *Chiton olivaceous*, measured the length and width of the eight overlapping plates composing the shell of 10 of these animals.

chiton

Length	Width
10.7	5.8
11.0	6.0
9.5	5.0
11.1	6.0
10.3	5.3
10.7	5.8
9.9	5.2
10.6	5.7
10.0	5.3
12.0	6.3



r = 0.9692, is it significant?

Test hypotheses:

H_0: $\rho = 0$

H_a: $\rho \neq 0$

Assume x,y has normal distributions, $\rho = 0$, then perform a one sample t-test with following parameters:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Degree of freedom $df = n - 2$

CORRELATION

Test for Significance of Correlation

Test hypotheses:

H_0: ρ = 0
H_a: ρ ≠ 0

r = CORREL(…) = 0.9692

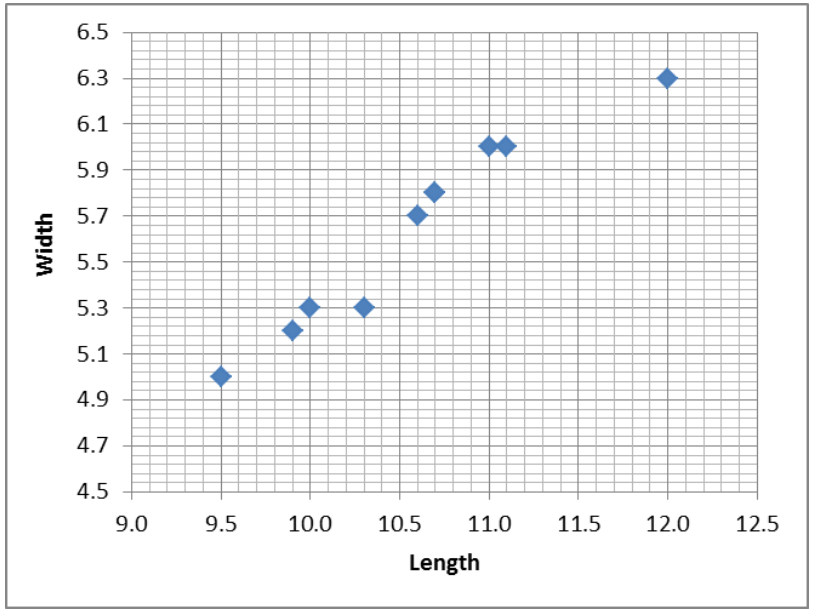
$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Degree of freedom *df* = *n - 2*

$$t = \frac{r-0}{s_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

t = 11.14,

p-value = 4e-6




```
= 2*T.DIST(-abs(t),n-2,TRUE)
```

```
cor.test(data)
```

CORRELATION

Confidence Intervals: Fisher Transformation

Fisher's Z-transformation connects normal z-values and correlation coefficients

$$Z = 0.5 \ln\left(\frac{1+r}{1-r}\right) \quad \Leftrightarrow \quad r = \frac{e^{2Z}-1}{e^{2Z}+1}$$

Confidence intervals with Fisher's transformation:

1. Transform correlation r → Z
2. Calculate standard deviation for Z using equation $\sigma_Z = \sqrt{\frac{1}{n-3}}$
3. Calculate upper and lower limits of Z:
   $Z_{min\,/\,max} = Z \pm z_{\alpha/2}\,\sigma_Z = Z \pm 1.96\,\sigma_Z$
4. Transform $Z_{min/max}$ back into $r_{min/max}$

r=	0.969226	
Fisher's Z=	2.079362	
sZ=	0.377964	
	Lower	Upper
Limits Z	1.338552	2.820172
Limits r	0.871324	0.992922

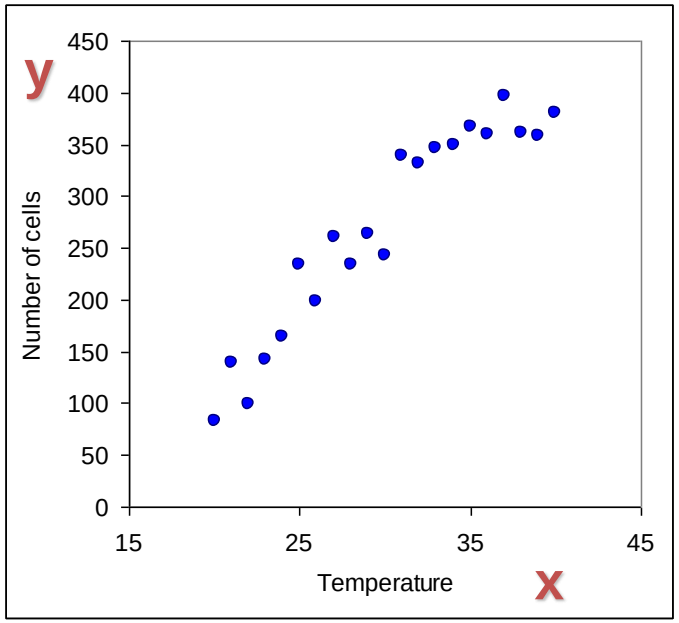
Excel: use steps 1-4

cor.test(data)

SIMPLE LINEAR REGRESSION

Experiments

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

```
Cells = read.table(
        "http://edu.modas.lu/data/txt/cells.txt",
        sep="\t",
        header=TRUE)

str(Cells)

plot(Cells, pch=19)
```

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*.

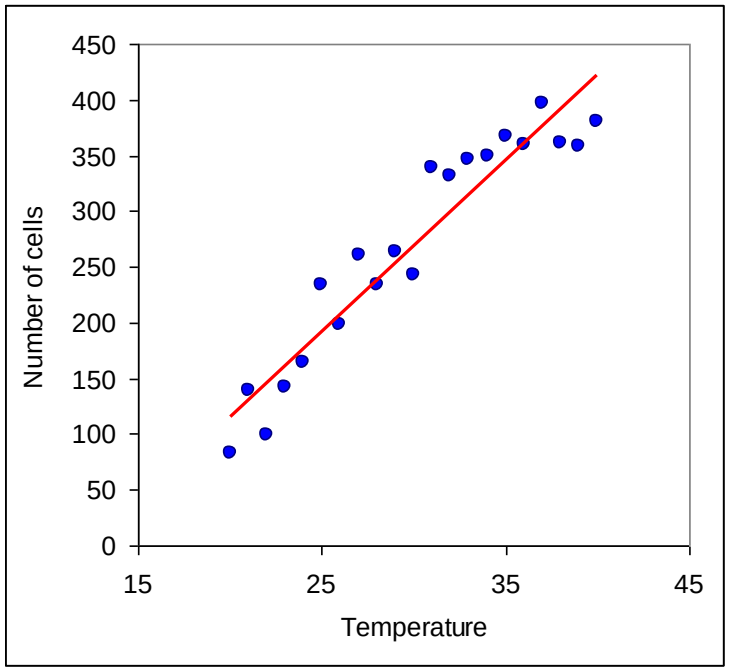
SIMPLE LINEAR REGRESSION

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



SIMPLE LINEAR REGRESSION

Regression Model and Regression Line

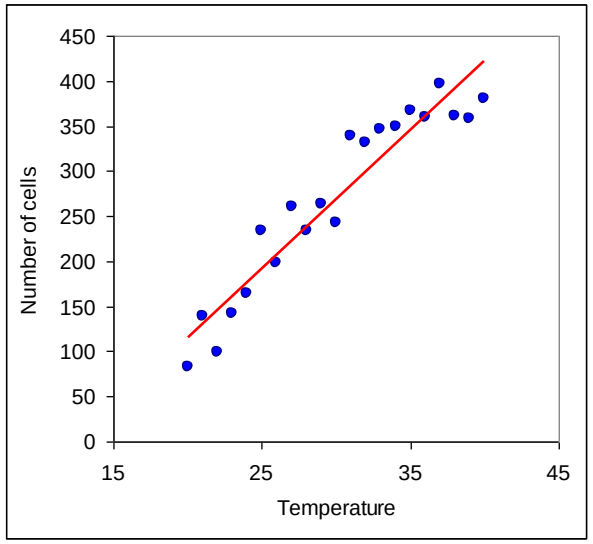
Regression model

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:

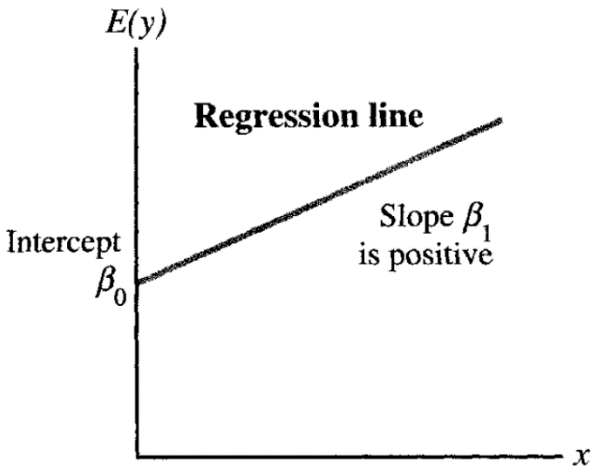
$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

SIMPLE LINEAR REGRESSION

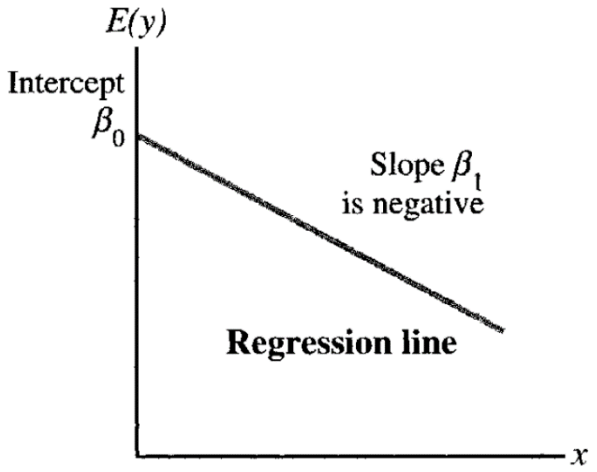
Regression Model and Regression Line

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

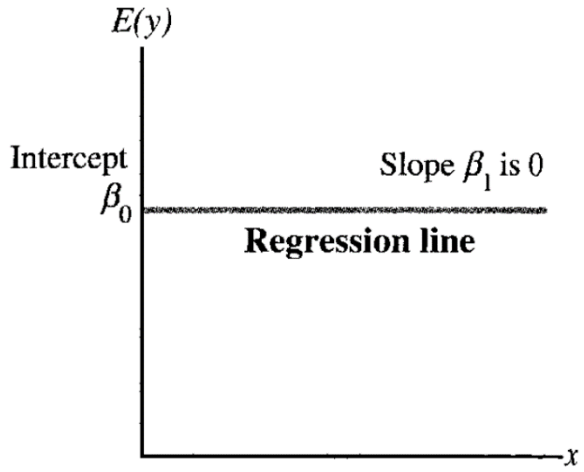
Panel A:
Positive Linear Relationship



Panel B:
Negative Linear Relationship



Panel C:
No Relationship



SIMPLE LINEAR REGRESSION

Experiments

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 $y = b_0 + b_1 x$

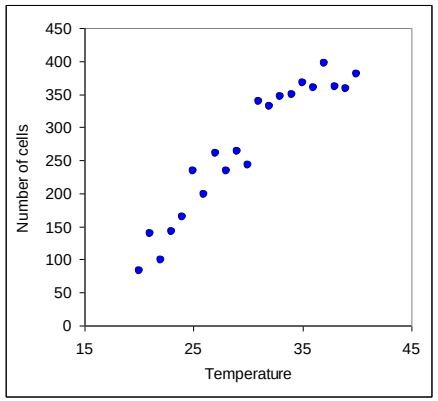
$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

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

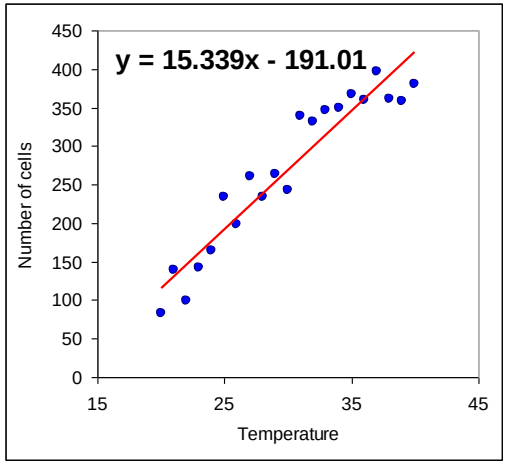
$$E[y(x)] = b_1 x + b_0$$

`cells`

1. Make a scatter plot for the data.



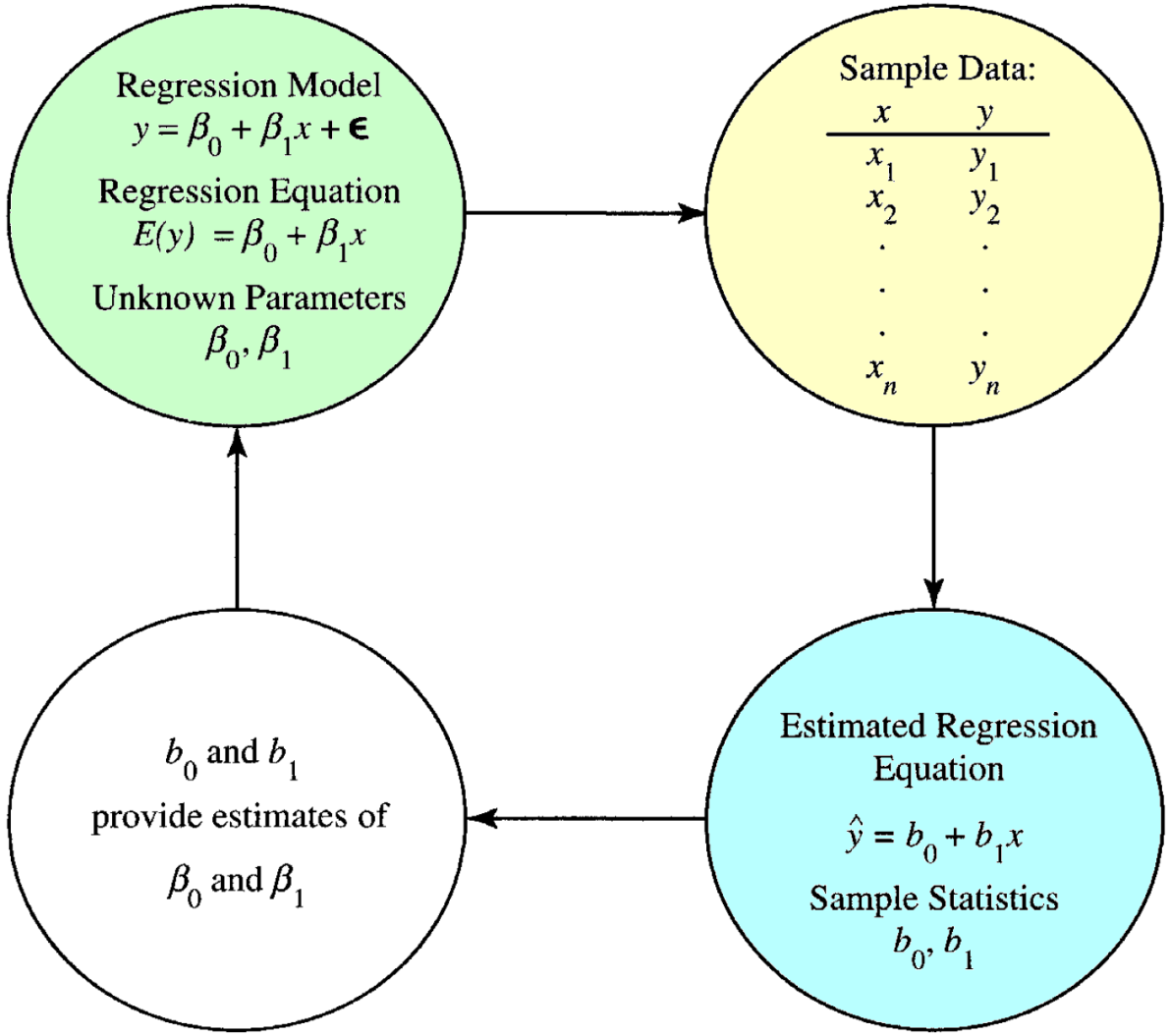
2. Right click to "Add Trendline". Show equation.




```
plot(Cells, pch=19)

abline(a=-191.01,
       b=15.339,
       col=2)
```

SIMPLE LINEAR REGRESSION

Overview



SIMPLE LINEAR REGRESSION

Experiments

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

Intersect: $$b_0 = m_y - b_1 m_x$$

SIMPLE LINEAR REGRESSION

Coefficient of Determination

Sum squares due to **error**

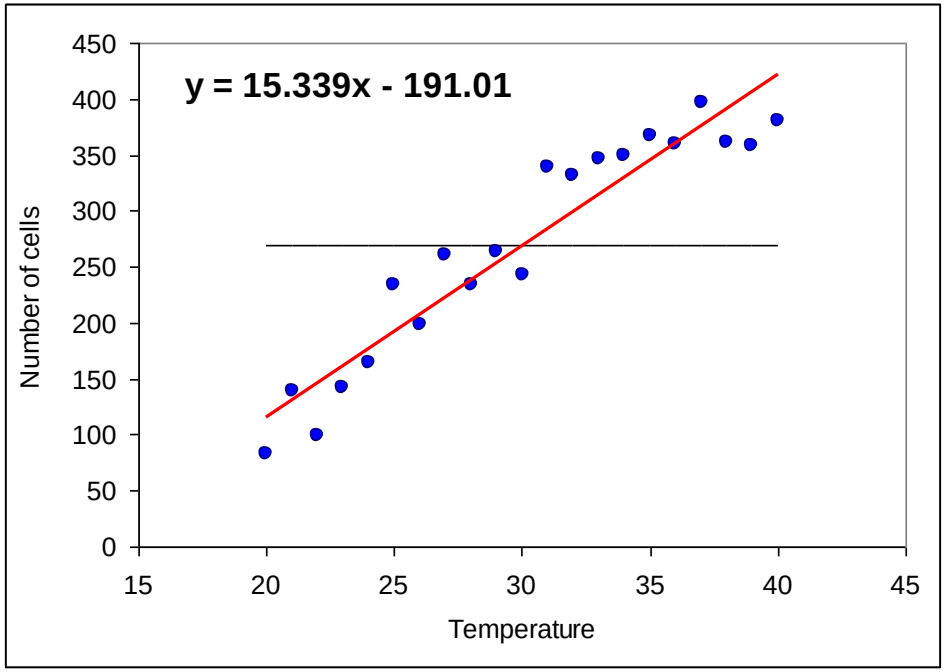
$$SSE = \sum (y_i - \hat{y}_i)^2$$

Sum squares total

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

Sum squares due to regression

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

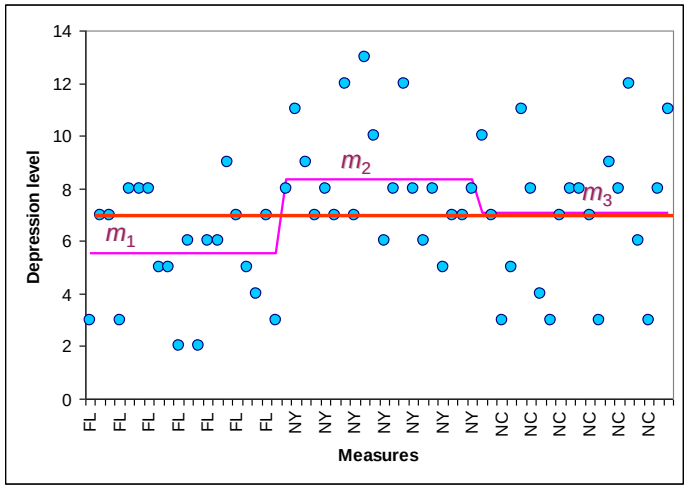


The Main Equation

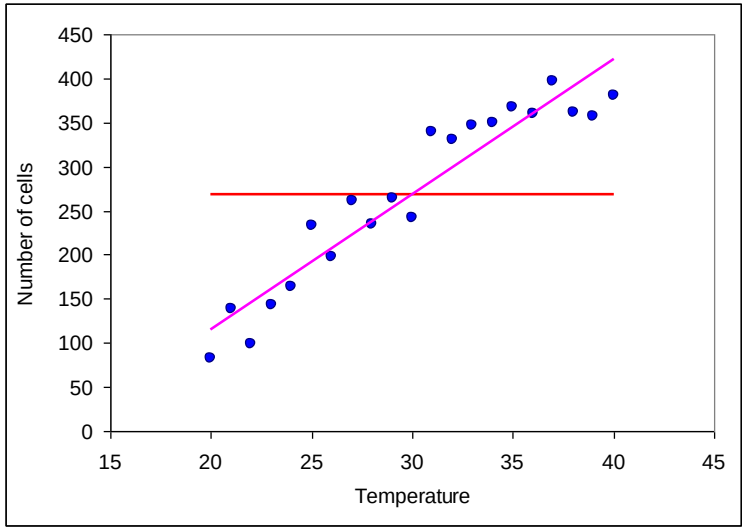
$$SST = SSR + SSE$$

SIMPLE LINEAR REGRESSION

ANOVA and Regression



$$SST = SSTR + SSE$$



$$SST = SSR + SSE$$

SIMPLE LINEAR REGRESSION

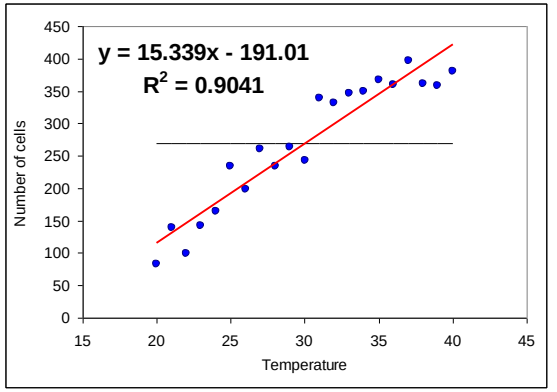
Coefficient of Determination

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

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

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

$$SST = SSR + SSE$$



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).

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

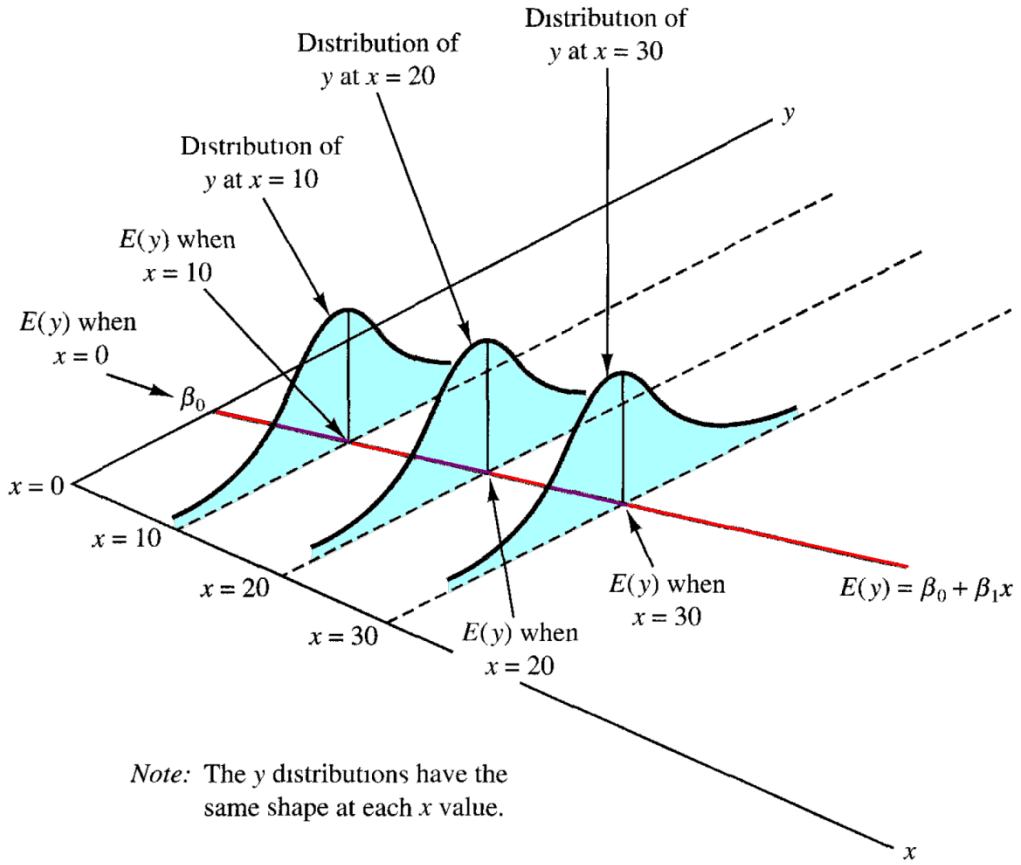
LINEAR REGRESSION

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.

TESTING FOR SIGNIFICANCE

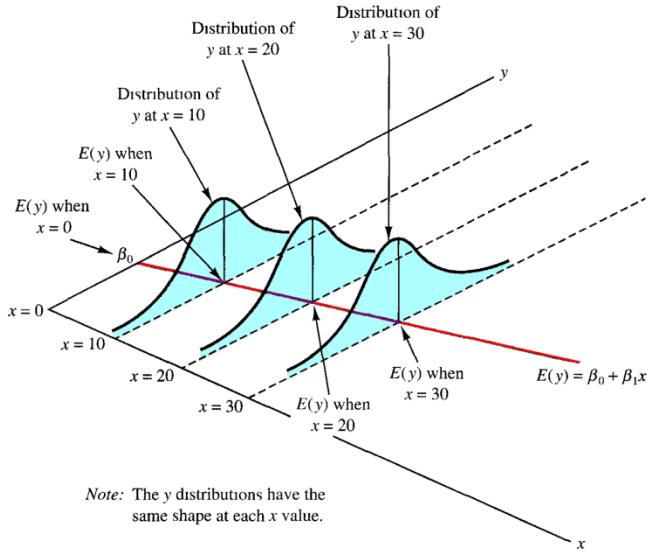
Estimation of σ^2

_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.



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

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

TESTING FOR SIGNIFICANCE

Sampling Distribution for b_1

If assumptions for ε are fulfilled, then the sampling distribution for b_1 is as follows:

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

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

Expected value $E[b_1] = \beta_1$

St.deviatiation $\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - m_x)^2}}$ **= Standard Error**

Distribution: ***normal***

Interval Estimation for β_1

$$\beta_1 = b_1 \pm t_{\alpha/2}^{(n-2)} \frac{\sigma}{\sqrt{\sum (x_i - m_x)^2}}$$

$$\beta_1 = b_1 \pm t_{\alpha/2}^{(n-2)} SE$$

TESTING FOR SIGNIFICANCE

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

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

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.

REGRESSION ANALYSIS

Example: Excel and R

`cells`

1. Calculate manually b_1 and b_0

Intercept	b0=	-191.008119
Slope	b1=	15.3385723

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

In R you should run the complete analysis:

```
model=lm(Cell.Number~Temperature, data=Cells)
```

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

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95091908
R Square	0.9042471
Adjusted R Square	0.89920747
Standard Error	31.7623796
Observations	21

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	181015.1117	181015.11	179.4274	3.95809E-11
Residual	19	19168.12641	1008.8488		
Total	20	200183.2381			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-190.783550	35.031618	-5.446039	2.96E-05	-264.10557	-117.46153	-264.10557	-117.46153
Temperature	15.332468	1.144637	13.395051	3.96E-11	12.93671537	17.7282197	12.93671537	17.7282197

```
# Regression table
summary(model)

# ANOVA table
anova(model)

# intercept/slope
model$coefficients
```

REGRESSION ANALYSIS

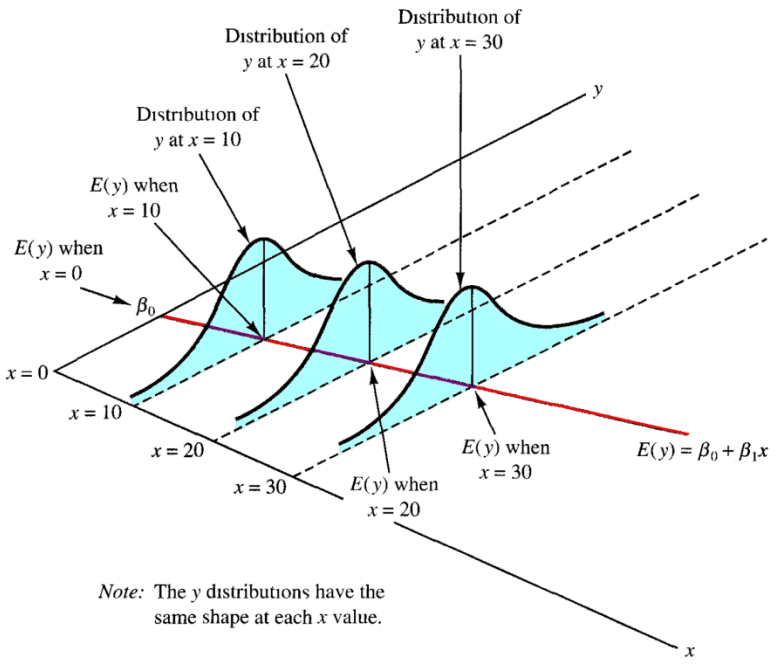
Confidence and Prediction

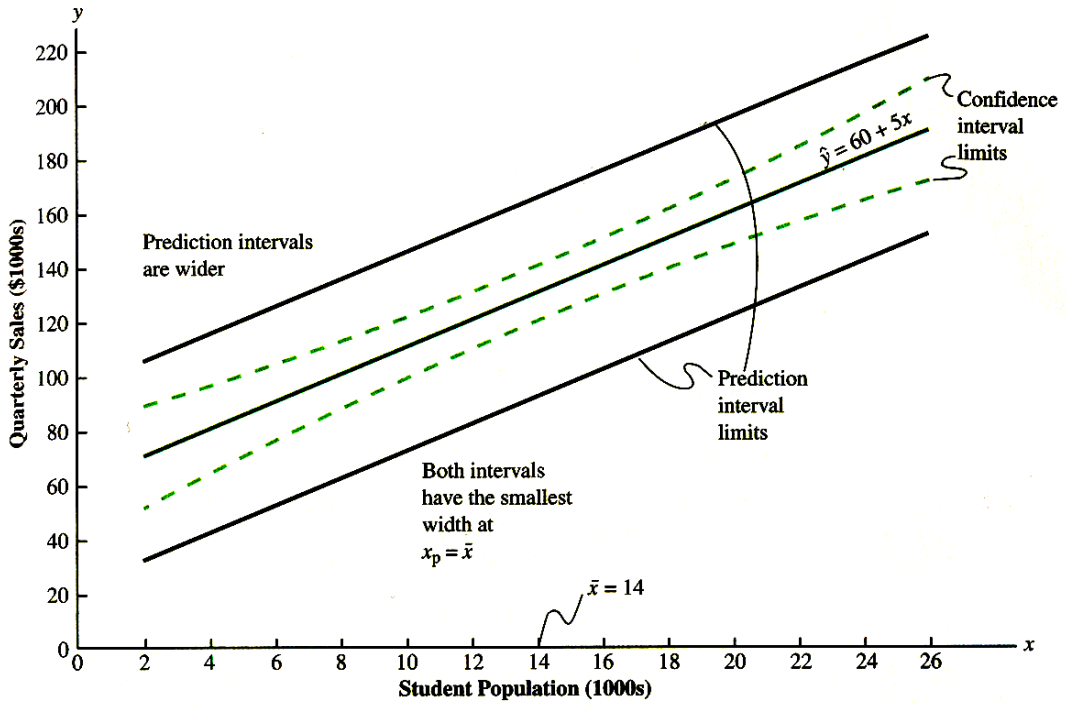
Confidence interval

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.



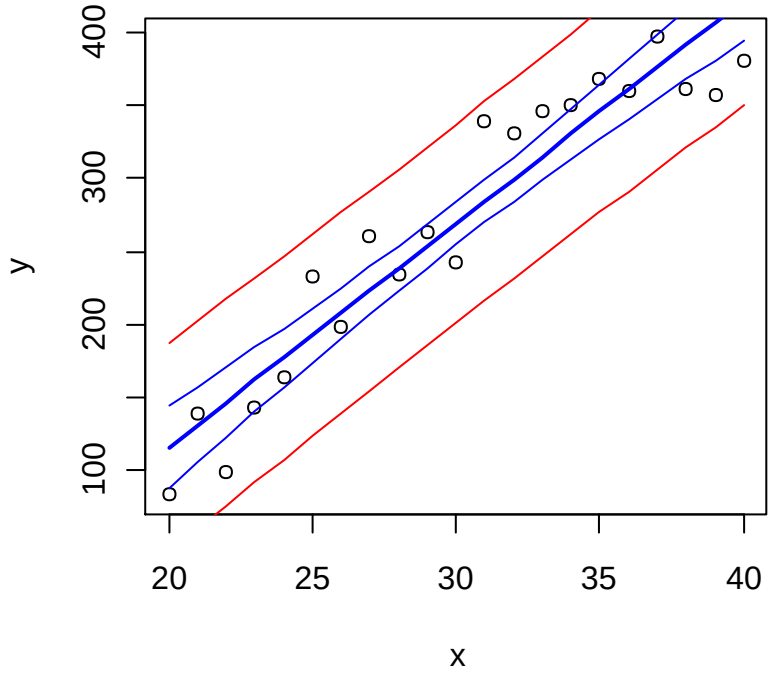


REGRESSION ANALYSIS

Example

cells.txt

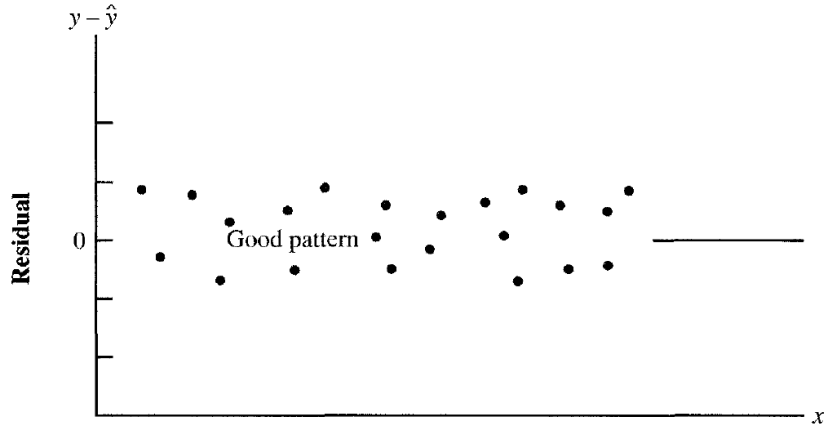
```
x = data$Temperature
y = data$Cell.Number
res = lm(y~x)
res
summary(res)

# 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)
```



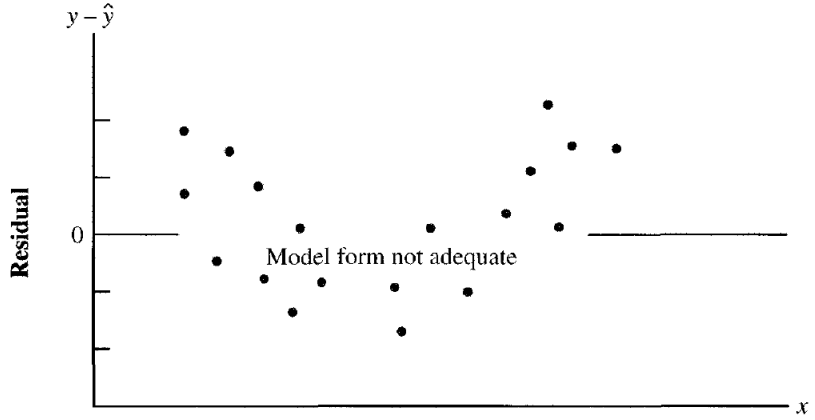
REGRESSION ANALYSIS

Residuals

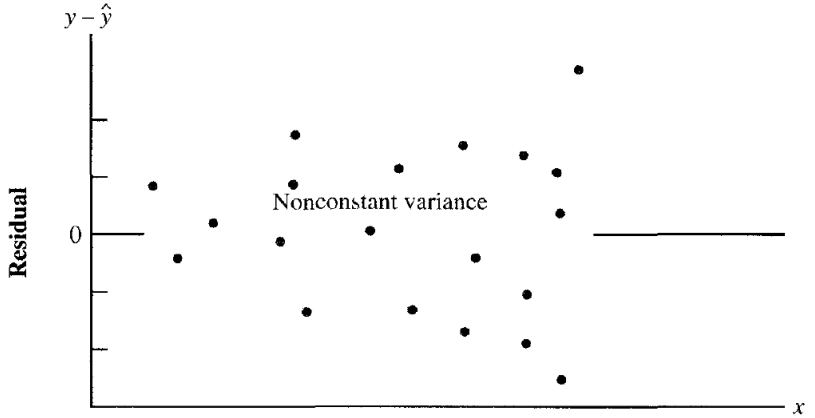
Panel A



Panel C



Panel B



REGRESSION ANALYSIS

Example 2

rana

A biology student wishes to determine the relationship between temperature and heart rate in leopard frog, *Rana pipiens*. He manipulates the temperature in 2° increment ranging from 2 to 18°C and records the heart rate at each interval. His data are presented in table rana.txt

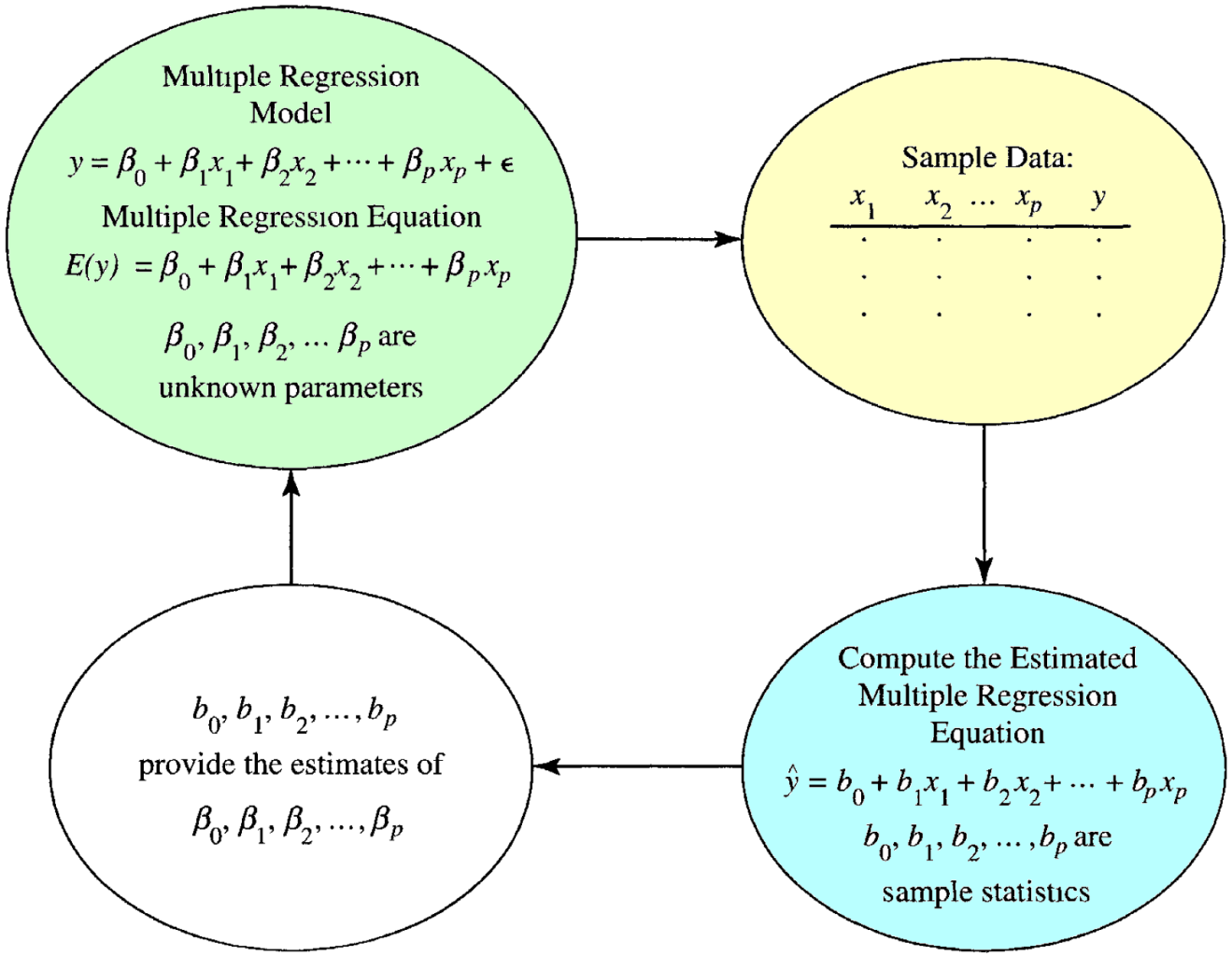
1) Build the model and provide the p-value for linear dependency

2) Provide interval estimation for the slope of the dependency

3) Estimate 95% prediction interval for heart rate at 15°

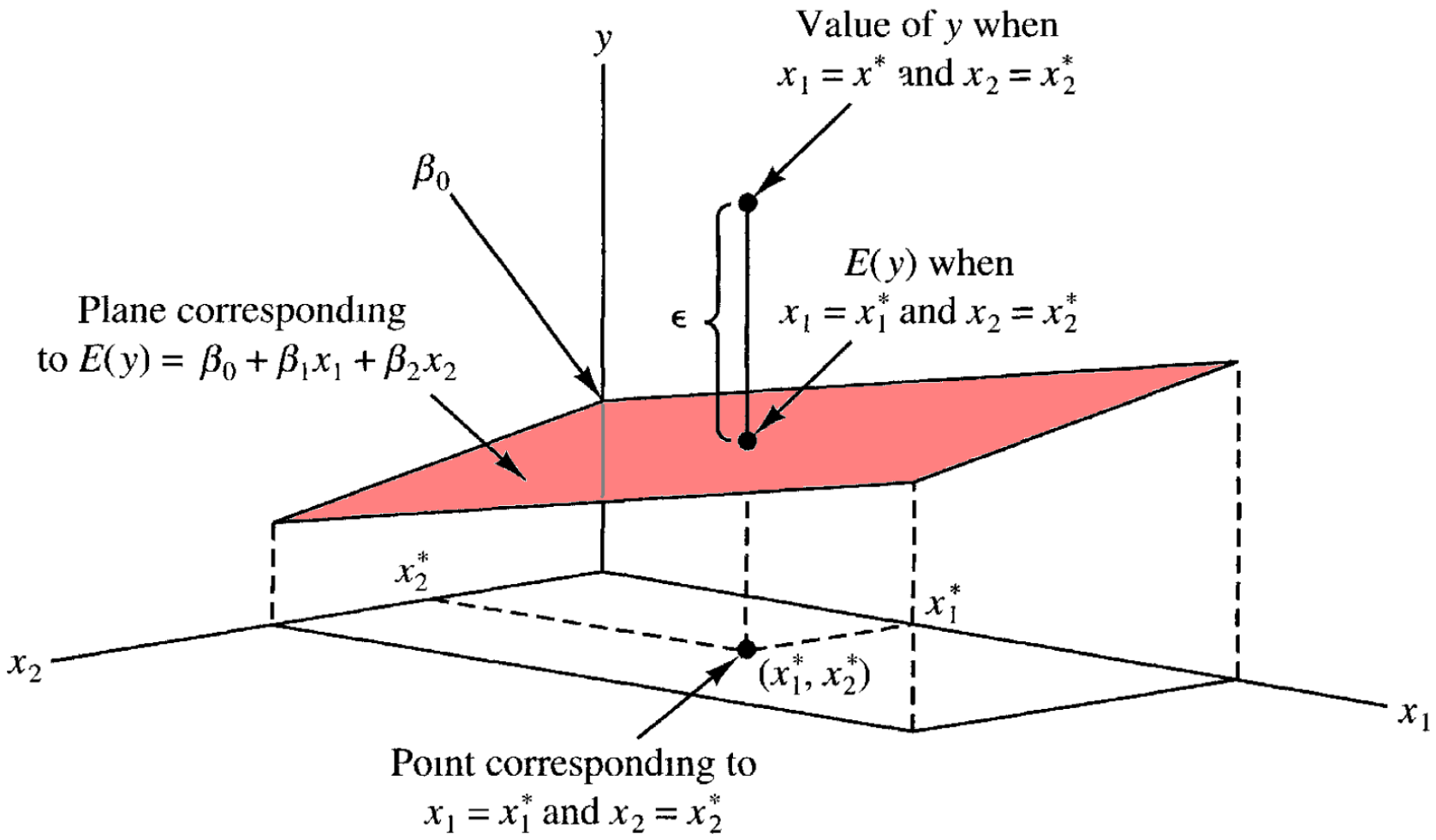
REGRESSION ANALYSIS

Multiple Regression



REGRESSION ANALYSIS

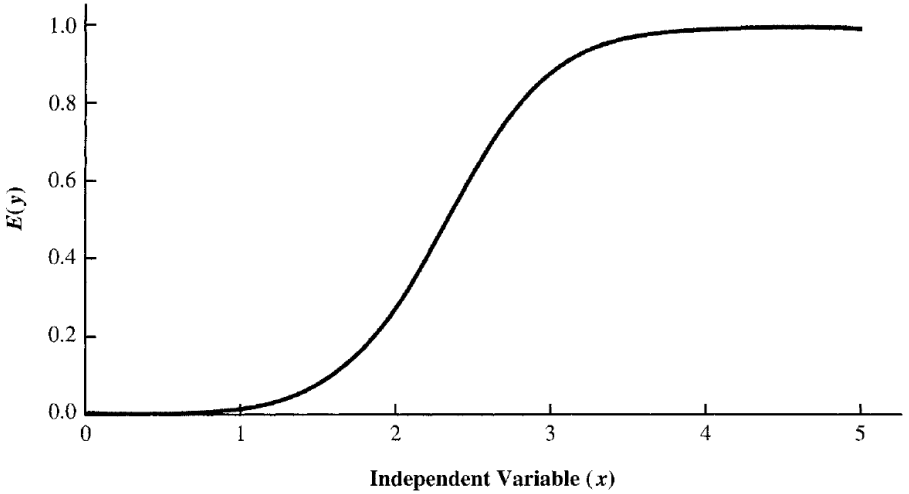
Multiple Regression



REGRESSION ANALYSIS

Non-Linear Regression

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



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

in **R**: glm(..., family="binomial")

Example:

```
Mice = read.table(
"http://edu.modas.lu/data/txt/mice.txt",
header=T,sep="\t",as.is=FALSE)
str(Mice)

model = glm( Sex ~ Blood.pH +
Bone.mineral.density + Lean.tissues.weight +
Ending.weight,
          data = Mice[ikeep,],
          family = "binomial")

summary(model)
```

http://edu.modas.lu/modas_pm/part2.html

CASE STUDY

Correlation Analysis of Transcriptomic Data

Gene regulatory networks (GRN) in living cells can be considered as extremely complex information processing systems. Despite their complexity, the main feature of the GRN is their robustness and ability to form a proper biochemical respond to a wide range of extracellular conditions. The knowledge about the part of GRN related to a specific bio-function of cellular process is of extreme importance for controlling them. Another important aspect of understanding cell functionality is linked to knowledge about the regulatory effect of small non-coding micro-RNA (**miRNA**). miRNAs influence most fundamental biological processes by ultimately altering the expression levels of proteins either through degradation of **mRNA** or through interference with mRNA translation. miRNAs tend to have long half lives and therefore represent promising candidates to be used as disease markers and therapeutic targets.

Being a reverse-engineering task, the GRN reconstruction is highly challenging, and requires analysis of large sets of experimental data. One of the straightest ways to reconstruct GRN is based on co-expression (CE) analysis of transcriptomic data from cDNA microarrays. Two significantly co-expressed genes or a gene and miRNA have the same or inverted expression profile over a number of samples. Biologically this is a good evidence for either a direct interaction between the genes or their mutual participation in the same biological function.

The performance of the software was tested using public mRNA and miRNA expression data from 14 various cell lines (A498, ACHN, CAKI1, CCRFCEM, HCT15, HL60, K562, MALME3M, MCF7, MOLT4, NCIH226, NCIH522, RPMI8226, SKOV3). Data from 42 Affymetrix® HGU133plus2 arrays and 14 miRNA custom microarray experiments were downloaded from public repositories (ref. E-MTAB-37 and E-MEXP-1029, http://www.ebi.ac.uk) , normalized and analyzed.

Tool: http://edu.sablab.net/biostat2/coexpress.zip

Data: http://edu.sablab.net/biostat2/data-mir-mrna_14cl.zip

QUESTIONS ?

Thank you for your attention