CENTRE DE RECHERCHE PUBLIC SANTÉ

Microarray Center

APPLIED STATISTICS

Lecture 10

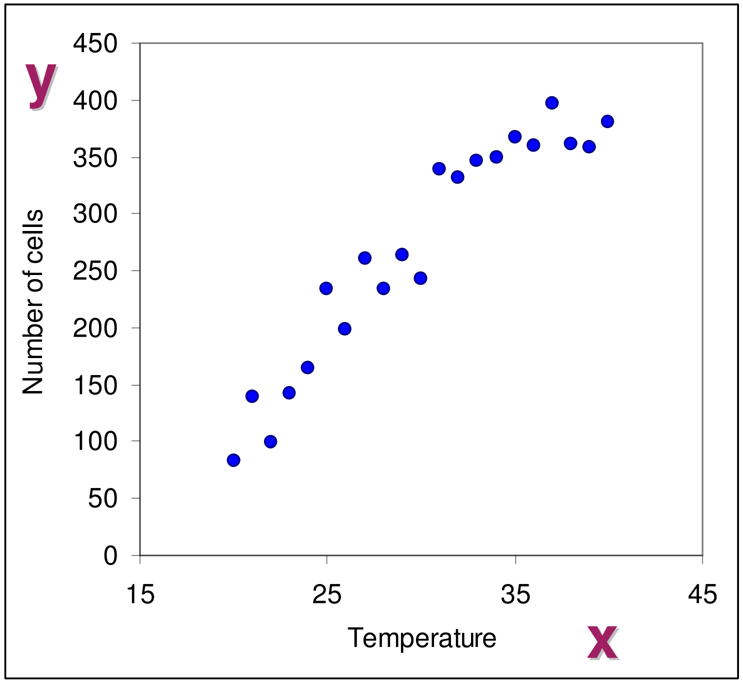
Linear Regression

Petr Nazarov
petr.nazarov@crp-sante.lu

3-11-2009

- **Introduction to linear regression**
  - dependent and independent random variables
  - scatter plot and linear trendline
- **Testing for significance**
  - estimation of the noise variance
  - interval estimations
  - testing hypothesis about significance
- **Regression Analysis**
  - confidence and prediction
  - multiple linear regression
  - nonlinear regression

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



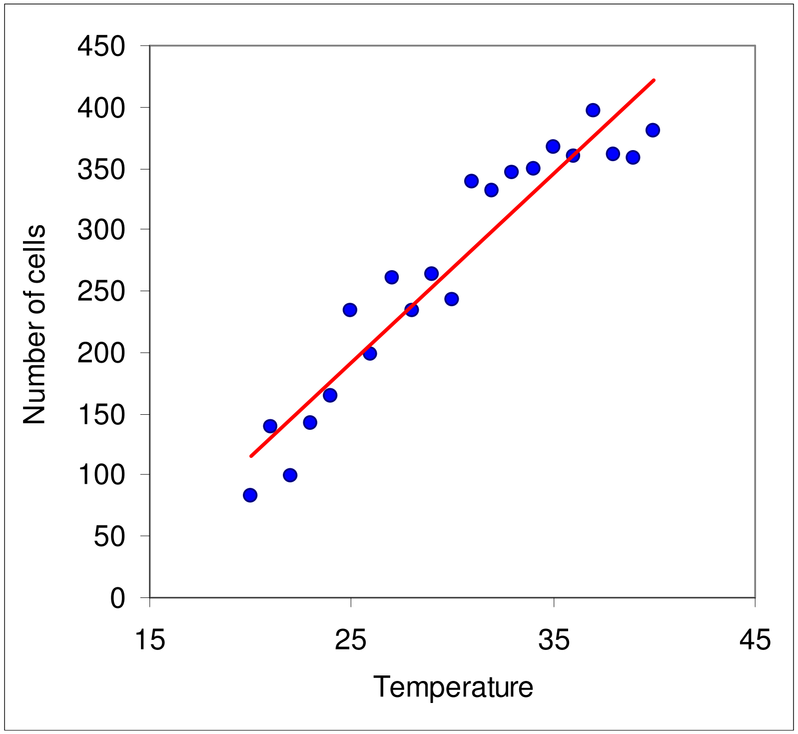
Cells are grown under different temperature conditions from 20° to 40°. A researched would like to find a dependency between T and cell number.

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***.

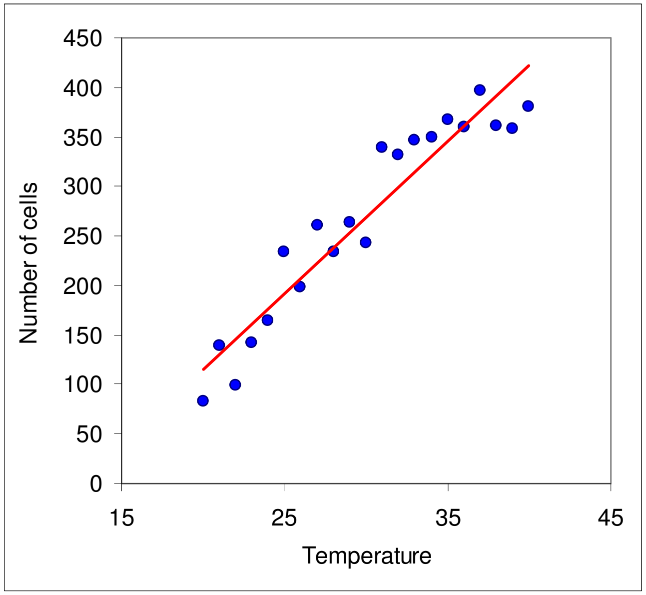
Simple linear regression
Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line.

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



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

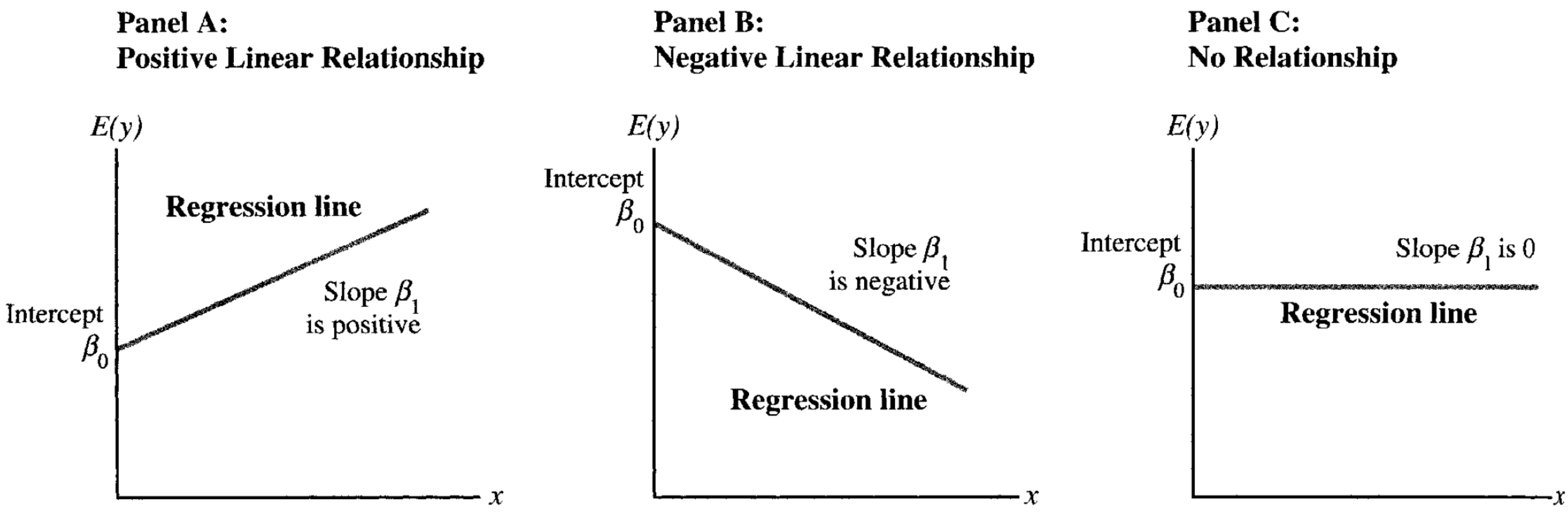
Regression equation
The equation that describes how the mean or expected value of the dependent variable is related to the independent variable; in simple linear regression,
$E(y) = \beta_0 + \beta_1 x$



- Model for a simple linear regression:

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

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



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 $y = b_0 + b_1x$

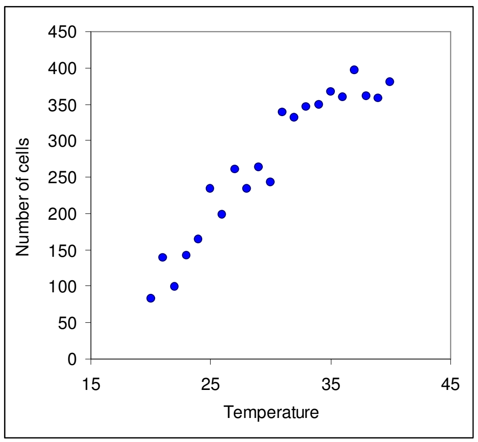
$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

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

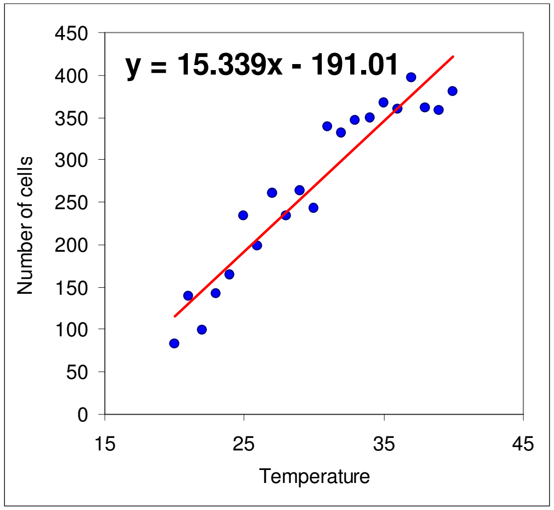
$$E[y(x)] = b_1 x + b_0$$

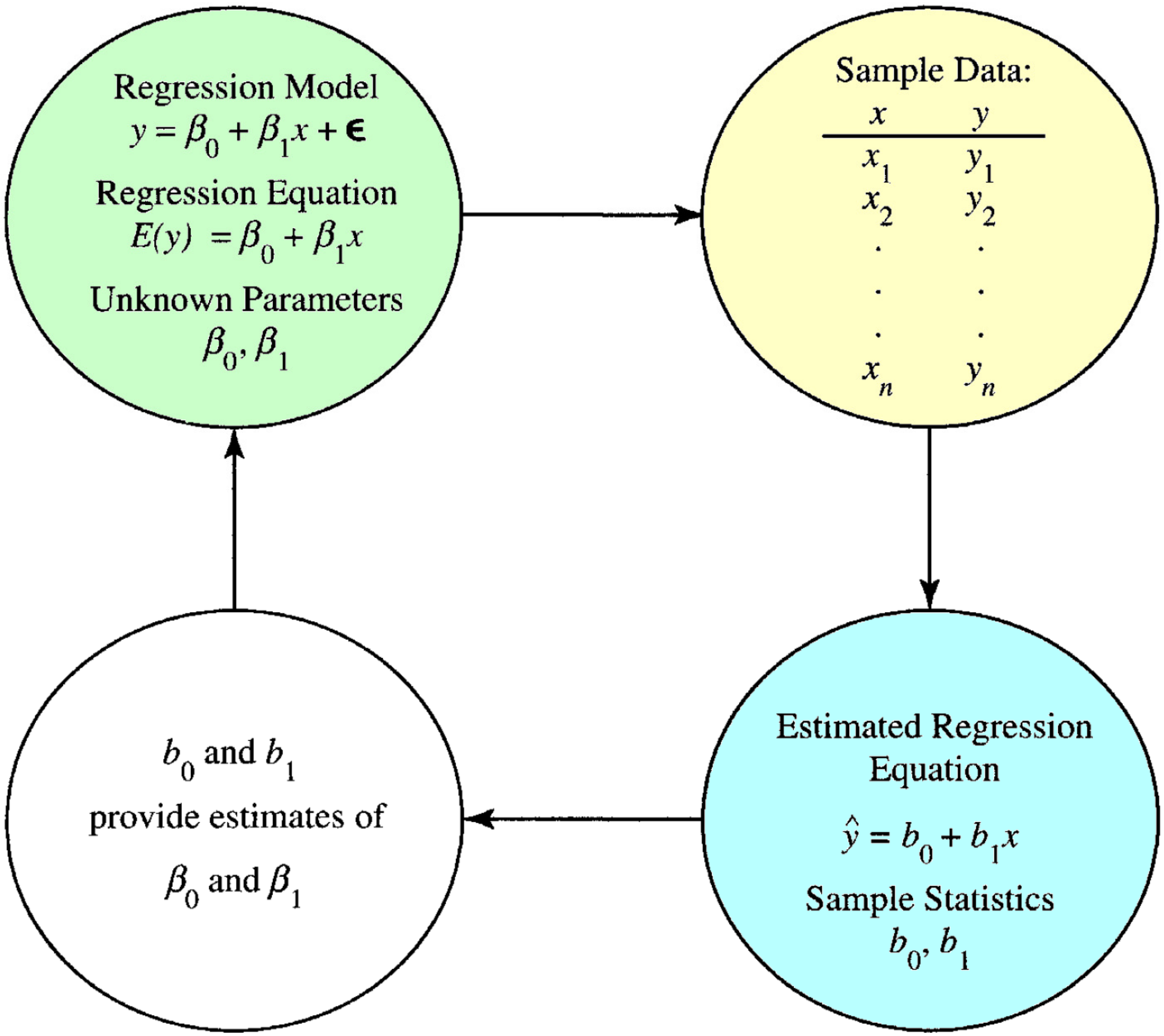
`cells.xls`

1. Make a scatter plot for the data.



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




Regression Model
$y = \beta_0 + \beta_1 x + \epsilon$
Regression Equation
$E(y) = \beta_0 + \beta_1 x$
Unknown Parameters
β_0, β_1
Sample Data:
x y
x_1 y_1
x_2 y_2
. .
. .
. .
x_n y_n
Estimated Regression Equation
$\hat{y} = b_0 + b_1 x$
Sample Statistics
b_0, b_1
b_0 and b_1
provide estimates of
β_0 and β_1

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

Intersect: $$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(x_1 - \bar{x})^2}$$

Slope: $$b_0 = \bar{y} - b_1\bar{x}$$

Sum squares due to **error**

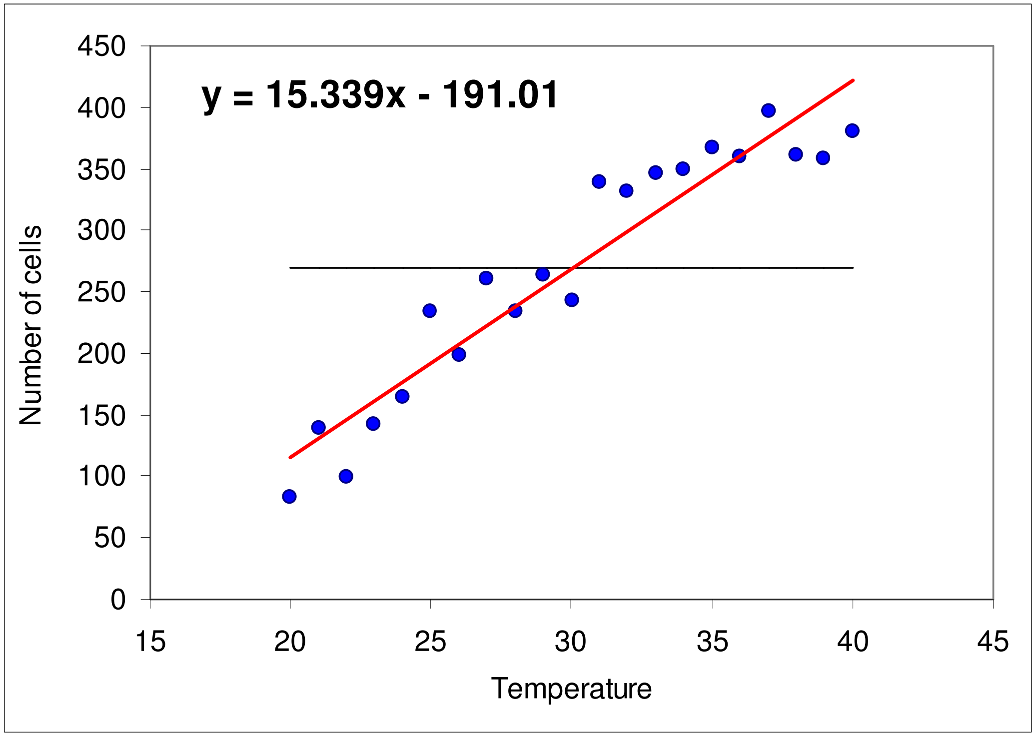
$$SSE = \sum (y_i - \hat{y}_i)^2$$

Sum squares total

$$SST = \sum (y_i - \overline{y})^2$$

Sum squares due to regression

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



The Main Equation

$$SST = SSR + SSE$$

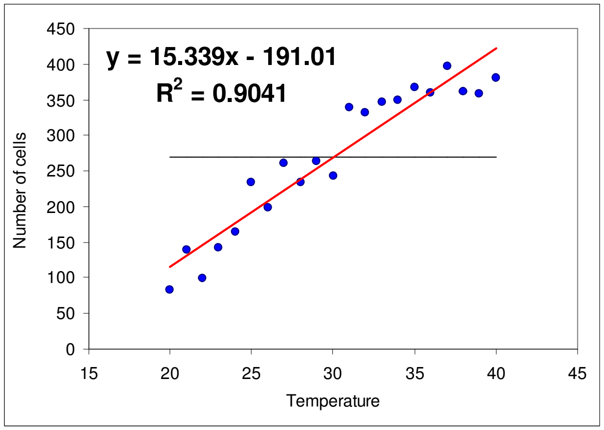
Coefficient of Determination

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

$$SST = \sum (y_i - \bar{y})^2$$

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

$$SST = SSR + SSE$$



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

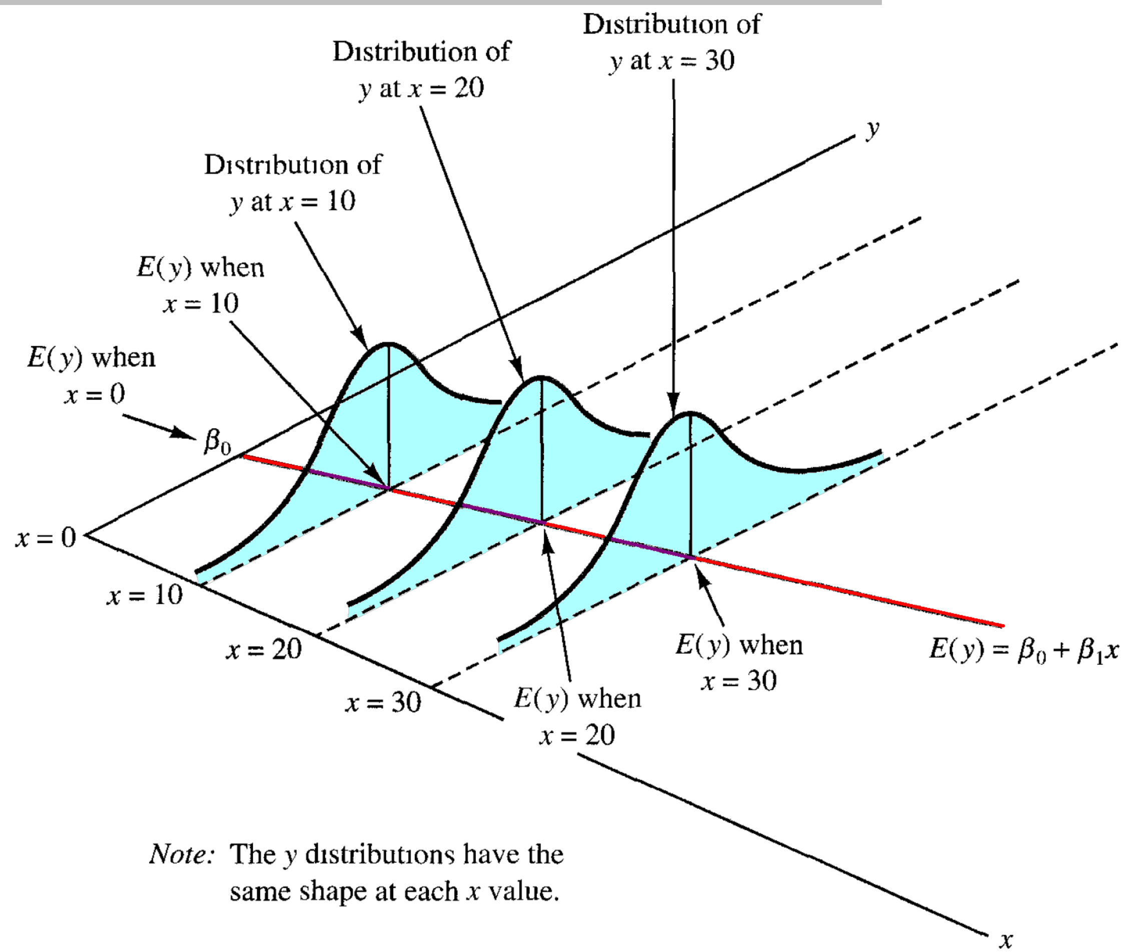
A measure of the strength of the linear relationship between two variables (previously discussed in Lecture 1).

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

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

Assumptions for Simple Linear Regression

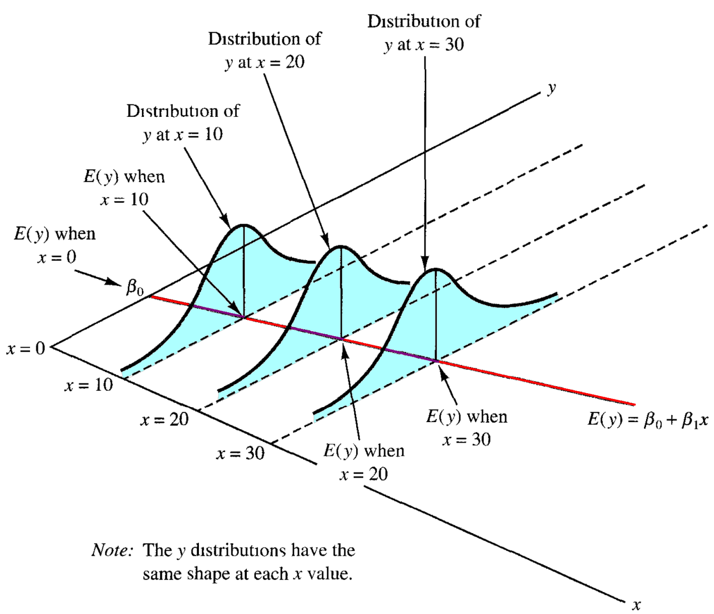
1. The error term ε is a random variable with 0 mean, i.e. E[ε]=0
2. The variance of ε, denoted by σ^2, is the same for all values of x
3. The values of ε are independent
3. The term ε is a normally distributed variable



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

_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.



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

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

Sampling Distribution for b_1

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

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

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

Expected value $E[b_1] = \beta_1$

Variance $$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Distribution: ***normal***

Interval Estimation for b_1

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

Test for Significance

H_0: $\beta_1 = 0$ *insignificant*

H_a: $\beta_1 \neq 0$

1. Build a t-test statistics.

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

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.

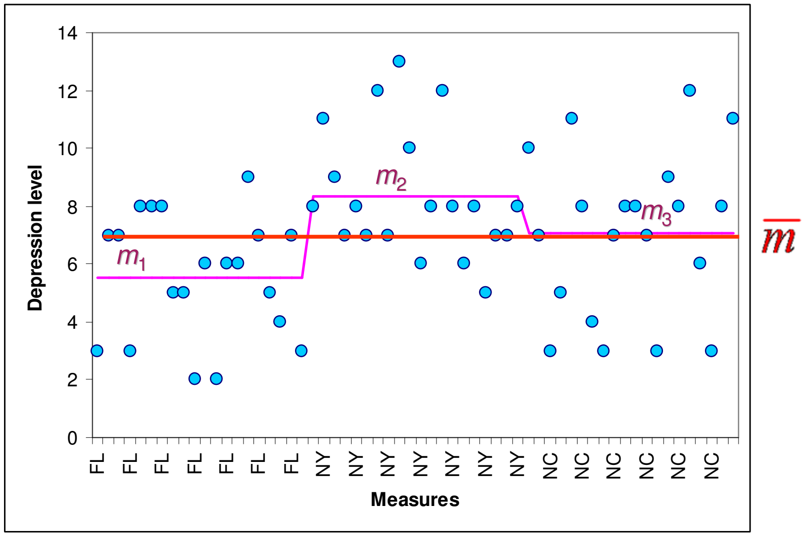
$$F = \frac{MSR}{MSE}$$

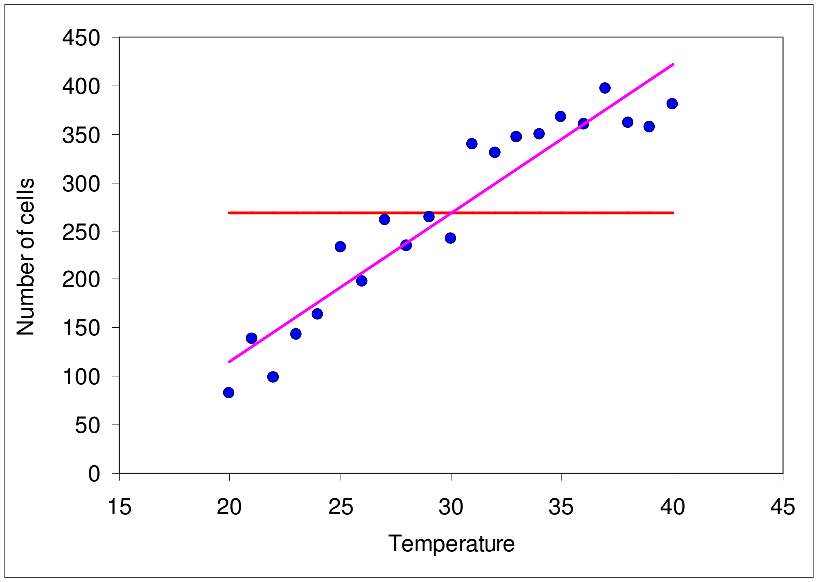
$$\text{MSR} = \frac{\text{SSR}}{\text{Number of independent variables}}$$

2. Calculate a p-value

TESTING FOR SIGNIFICANCE

ANOVA and Regression





$$SST = SSTR + SSE$$

$$SST = SSR + SSE$$

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 Tools → 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

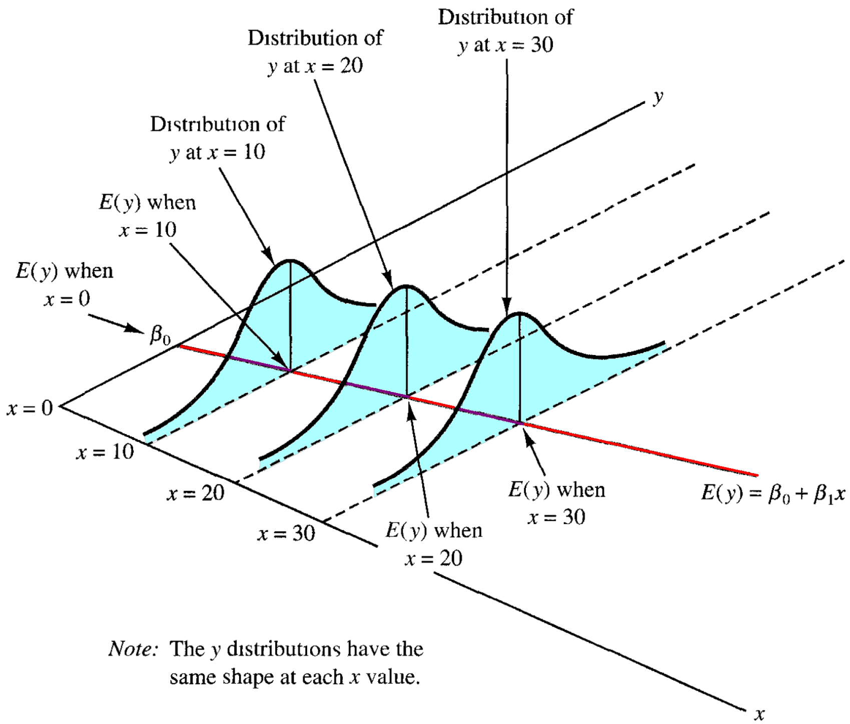
ANOVA

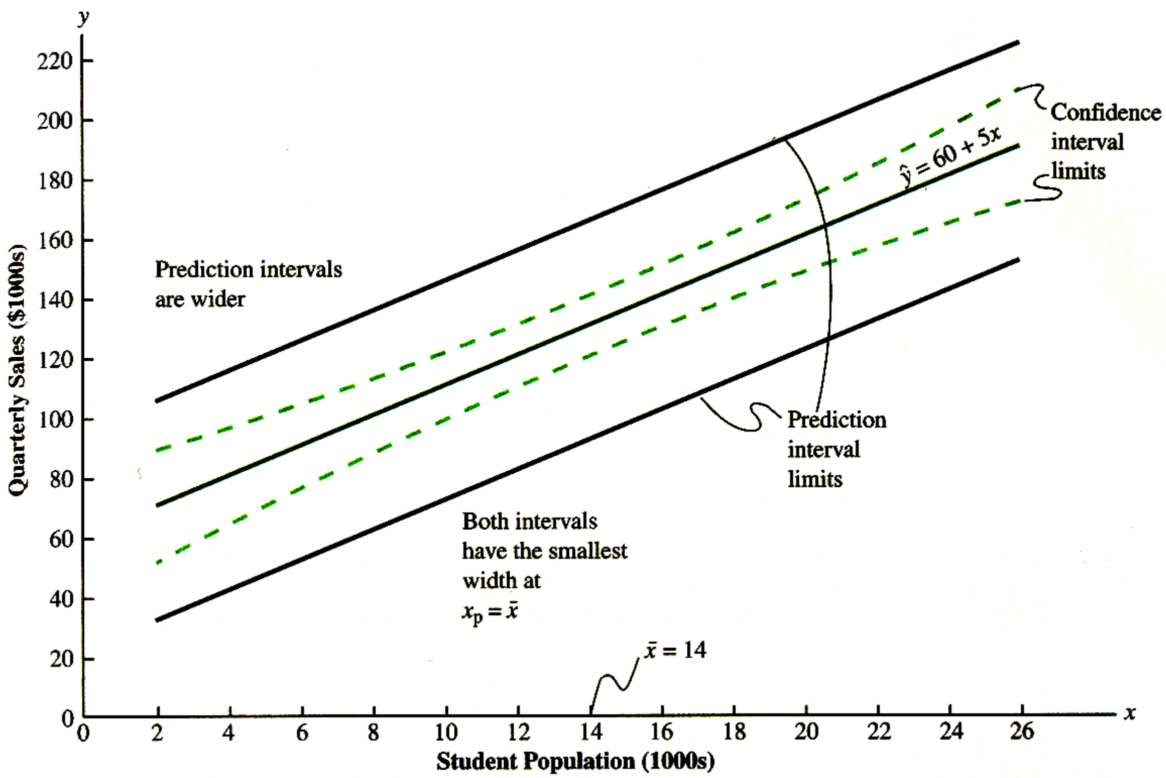
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	181159.2853	181159.3	179.1253	4.01609E-11
Residual	19	19215.7461	1011.355		
Total	20	200375.0314			

	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

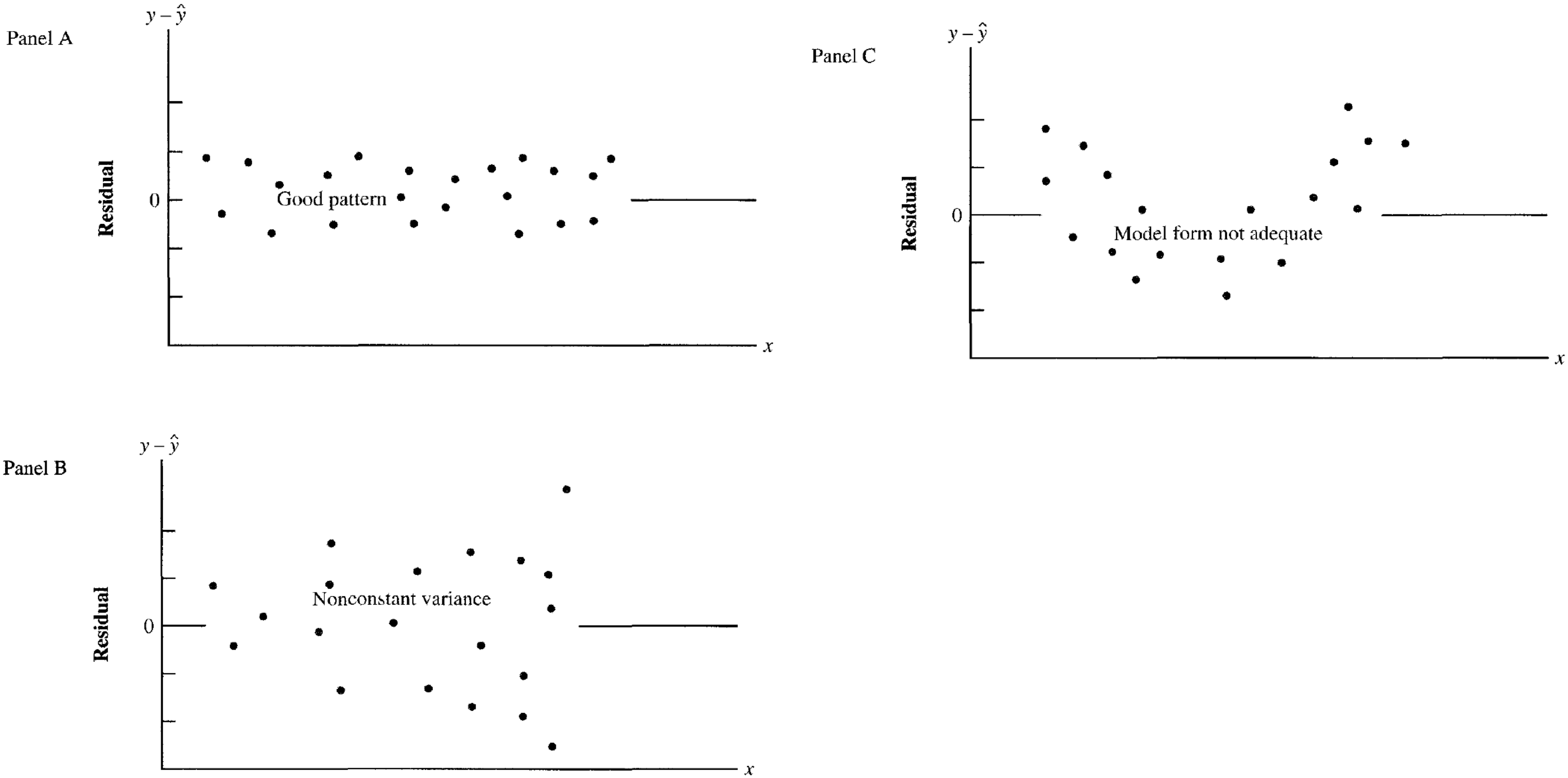
Confidence interval
The interval estimate of the mean value of y for a given value of x.

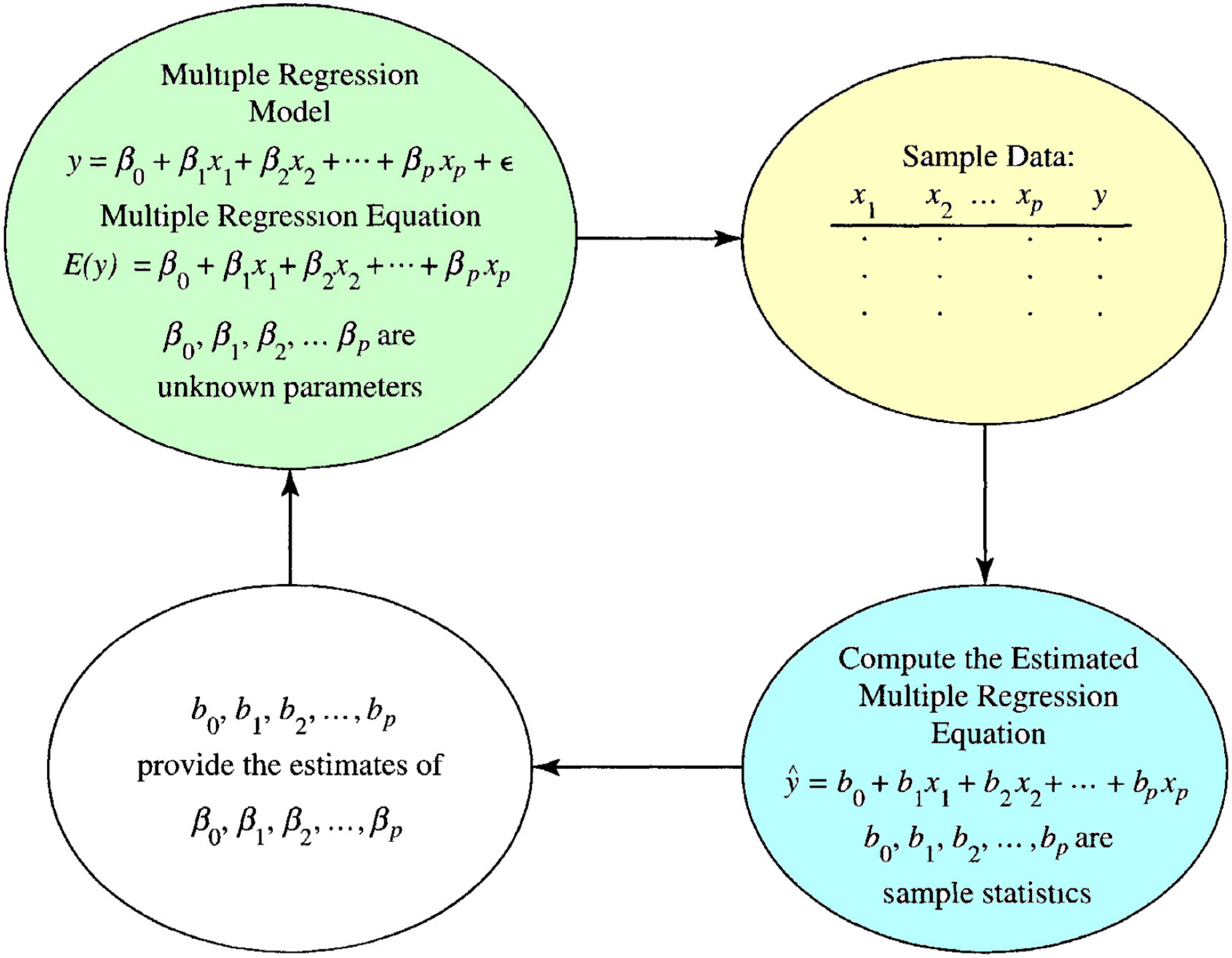
Prediction interval
The interval estimate of an individual value of y for a given value of x.

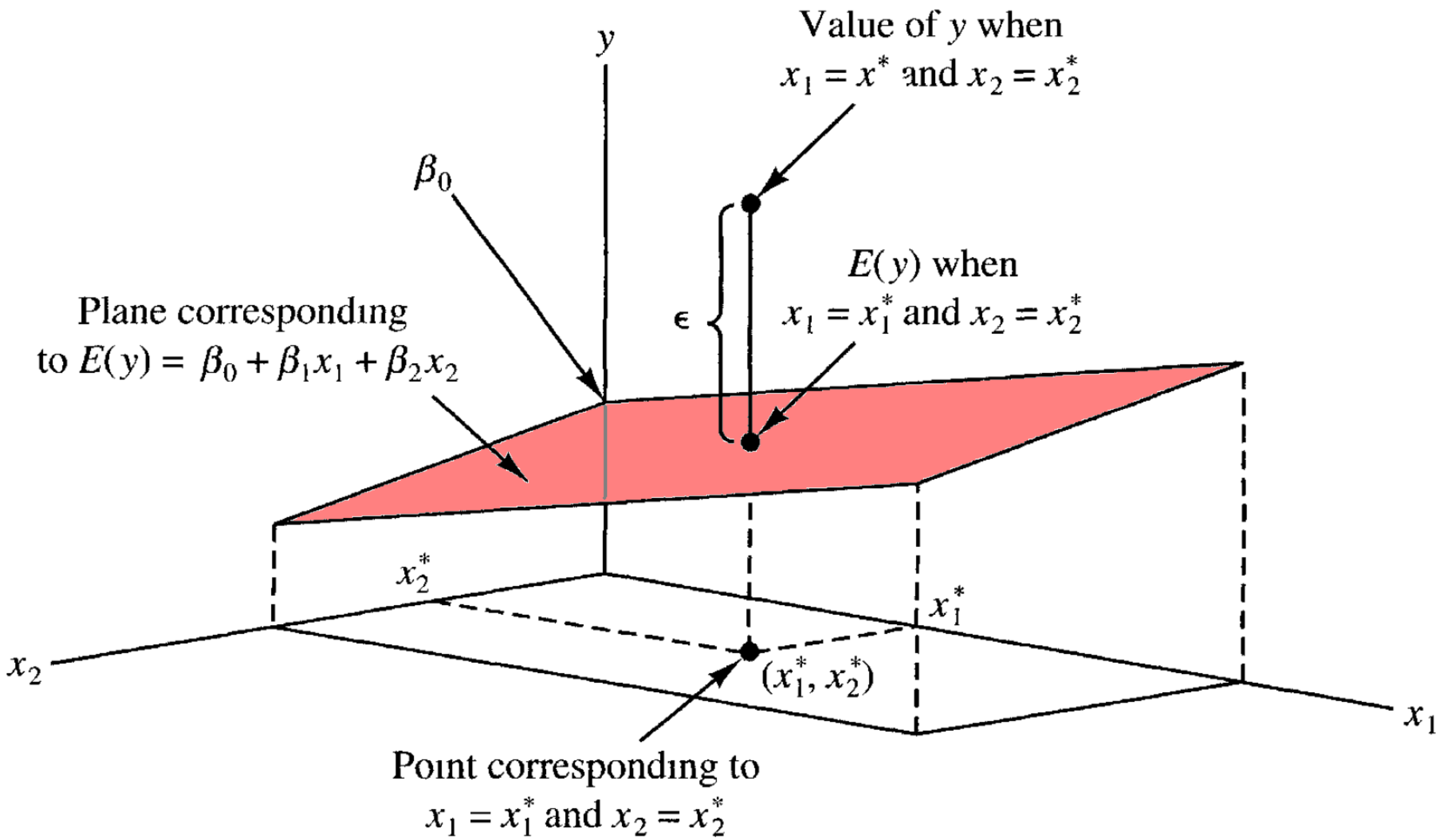




Residuals

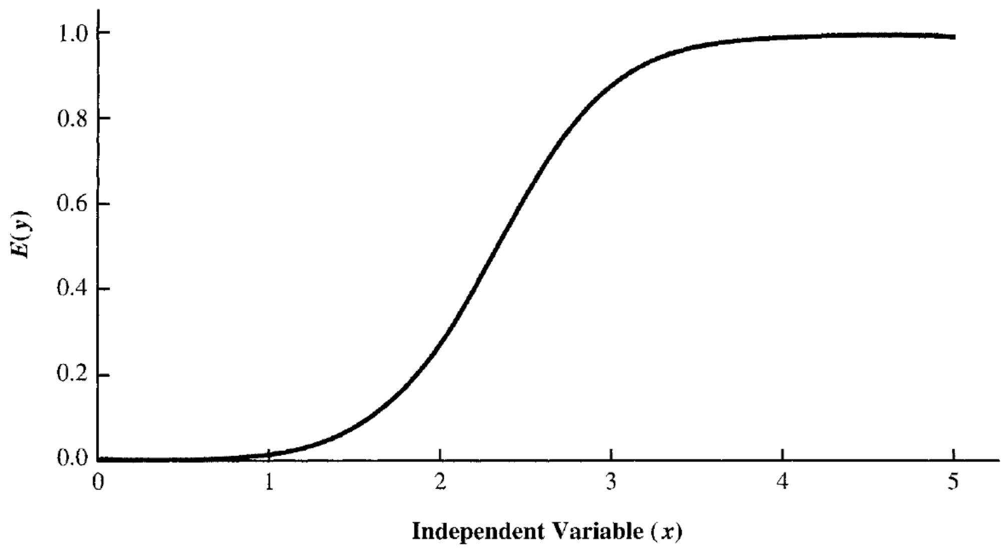

Panel A
$y - \hat{y}$
Residual
0
Good pattern
x
Panel C
$y - \hat{y}$
Residual
0
Model form not adequate
x
Panel B
$y - \hat{y}$
Residual
0
Nonconstant variance
x


Multiple Regression Model
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$
Multiple Regression Equation
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
$\beta_0, \beta_1, \beta_2, \dots \beta_p$ are unknown parameters
Sample Data:
x_1 x_2 $\dots$ x_p y
Compute the Estimated Multiple Regression Equation
$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$
$b_0, b_1, b_2, \dots, b_p$ are sample statistics
$b_0, b_1, b_2, \dots, b_p$
provide the estimates of
$\beta_0, \beta_1, \beta_2, \dots, \beta_p$



Value of y when
$x_1 = x^*$ and $x_2 = x_2^*$
y
β_0
$E(y)$ when
$x_1 = x_1^*$ and $x_2 = x_2^*$
Plane corresponding
to $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
ϵ
x_2^*
x_1^*
x_2
(x_1^*, x_2^*)
x_1
Point corresponding to
$x_1 = x_1^*$ and $x_2 = x_2^*$

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



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

Thank you for your attention

to be continued…