UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE
Multiomics Data Science

Regression

from BSc course Biostatistics (UL)

dr. Petr Nazarov

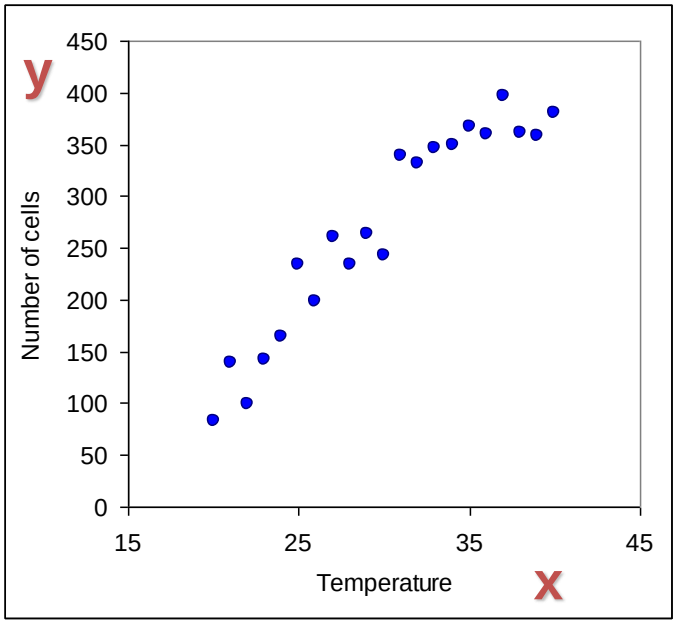
petr.nazarov@lih.lu

2020

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.txt`

Independent variable

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

Dependent variable

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

Linear Regression

Regression Model and Regression Line

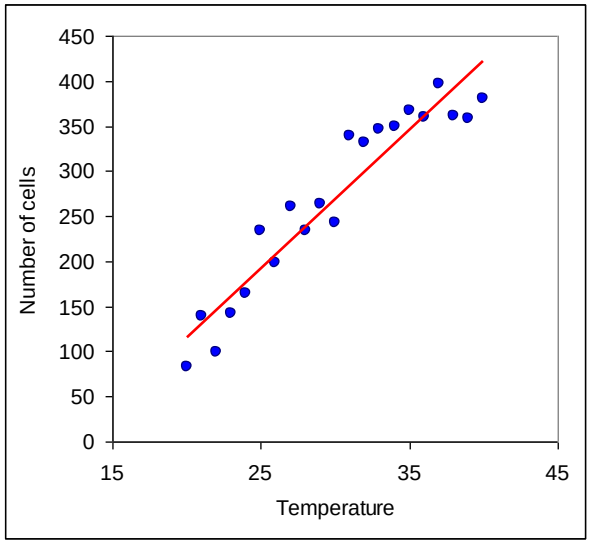
Regression model

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:

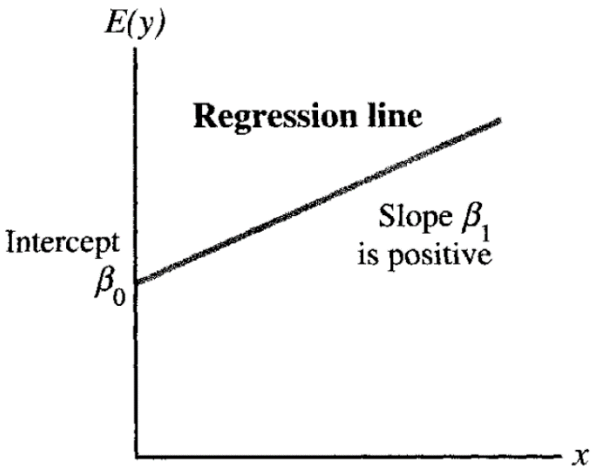
$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$

Linear Regression

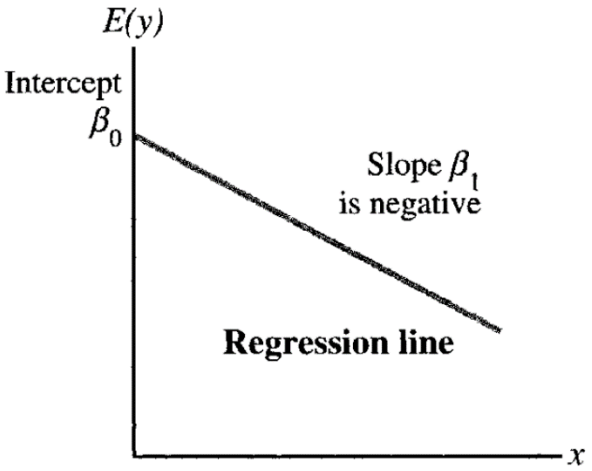
Regression Model and Regression Line

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

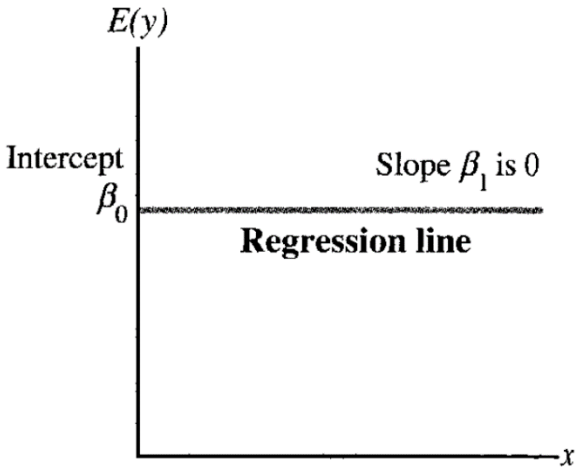
Panel A:
Positive Linear Relationship



Panel B:
Negative Linear Relationship

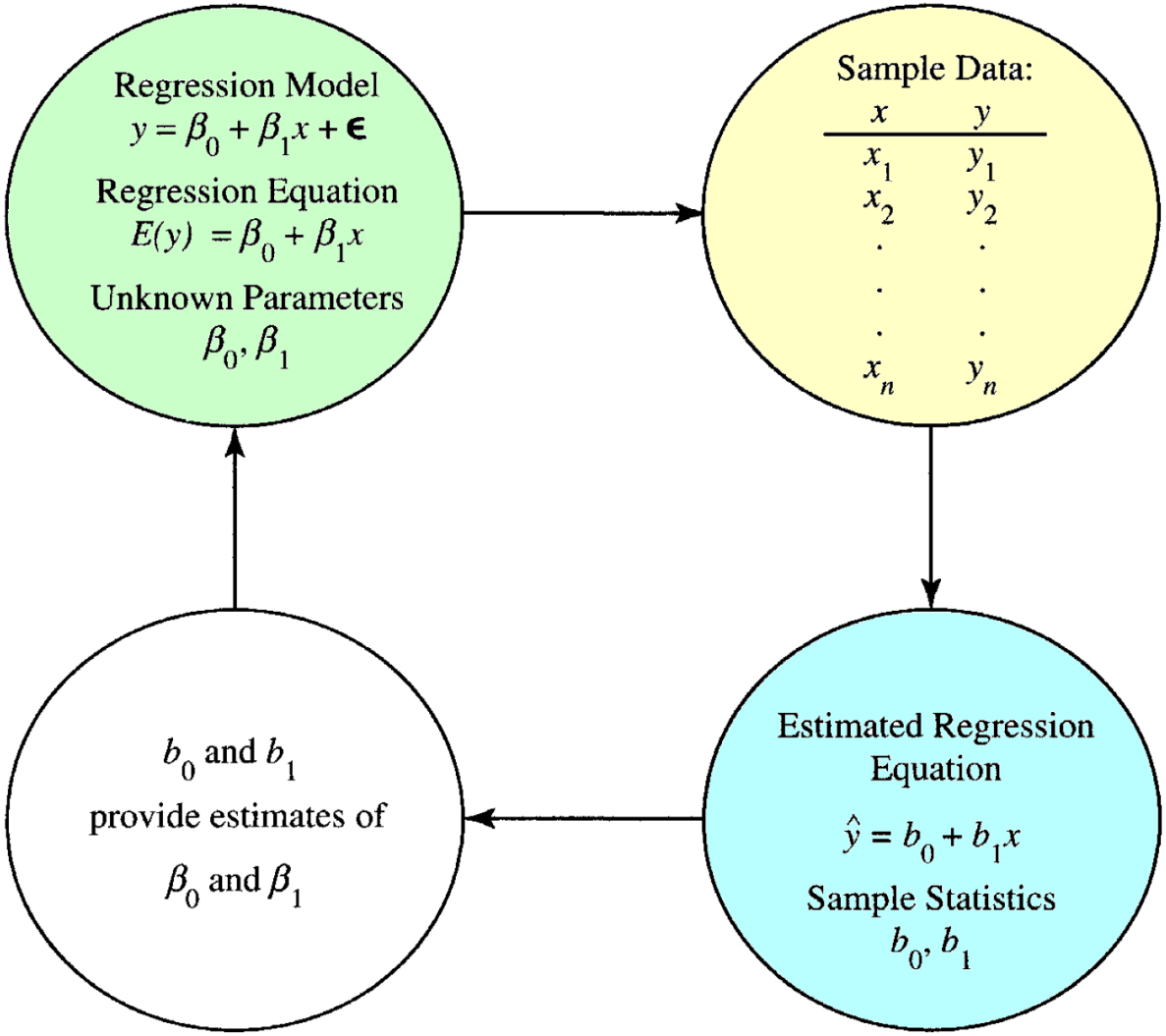


Panel C:
No Relationship



Linear Regression

Overview



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: $$b_0 = m_y - b_1 m_x$$

Linear Regression

Coefficient of Determination

Sum squares due to **error**

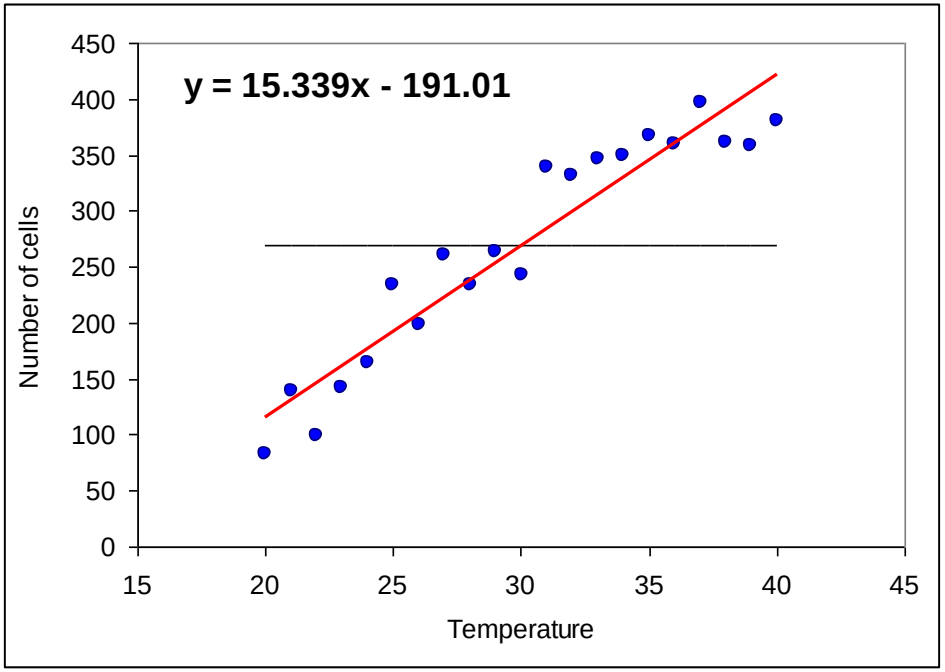
$$SSE = \sum (y_i - \hat{y}_i)^2$$

Sum squares total

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

Sum squares due to regression

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

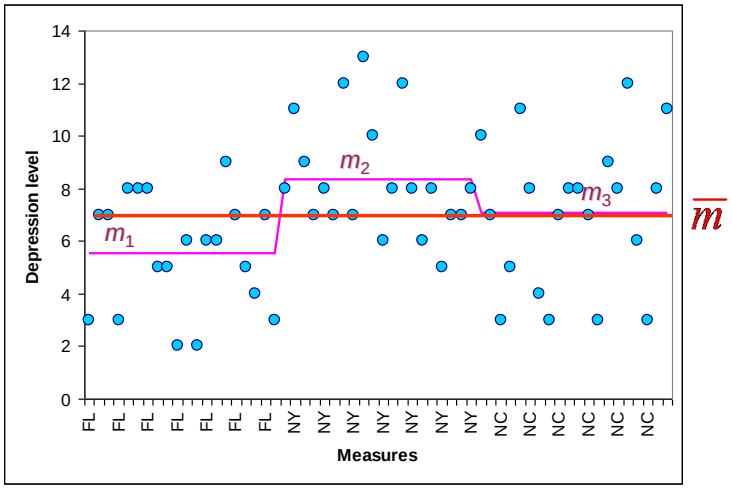


The Main Equation

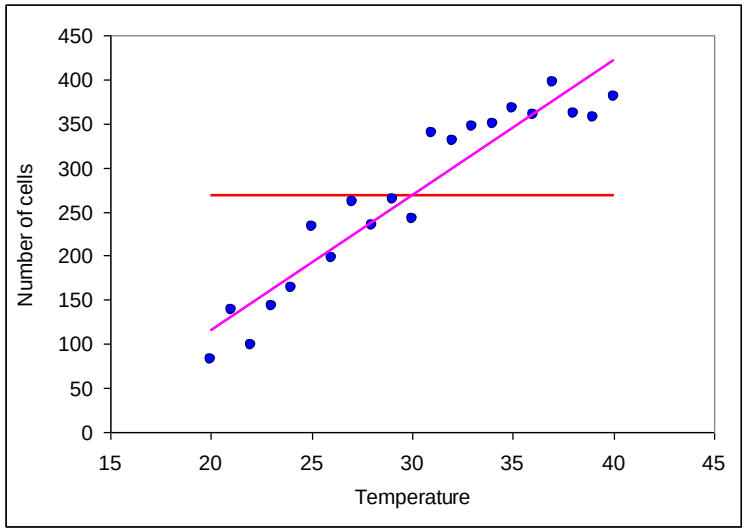
$$SST = SSR + SSE$$

Linear Regression

ANOVA and Regression



$$SST = SSTR + SSE$$



$$SST = SSR + SSE$$

Linear Regression

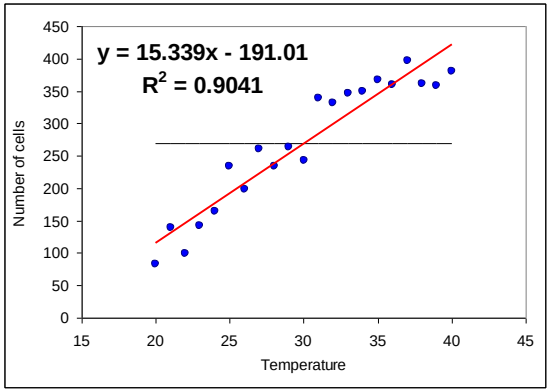
Coefficient of Determination

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

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

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

$$SST = SSR + SSE$$



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).

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

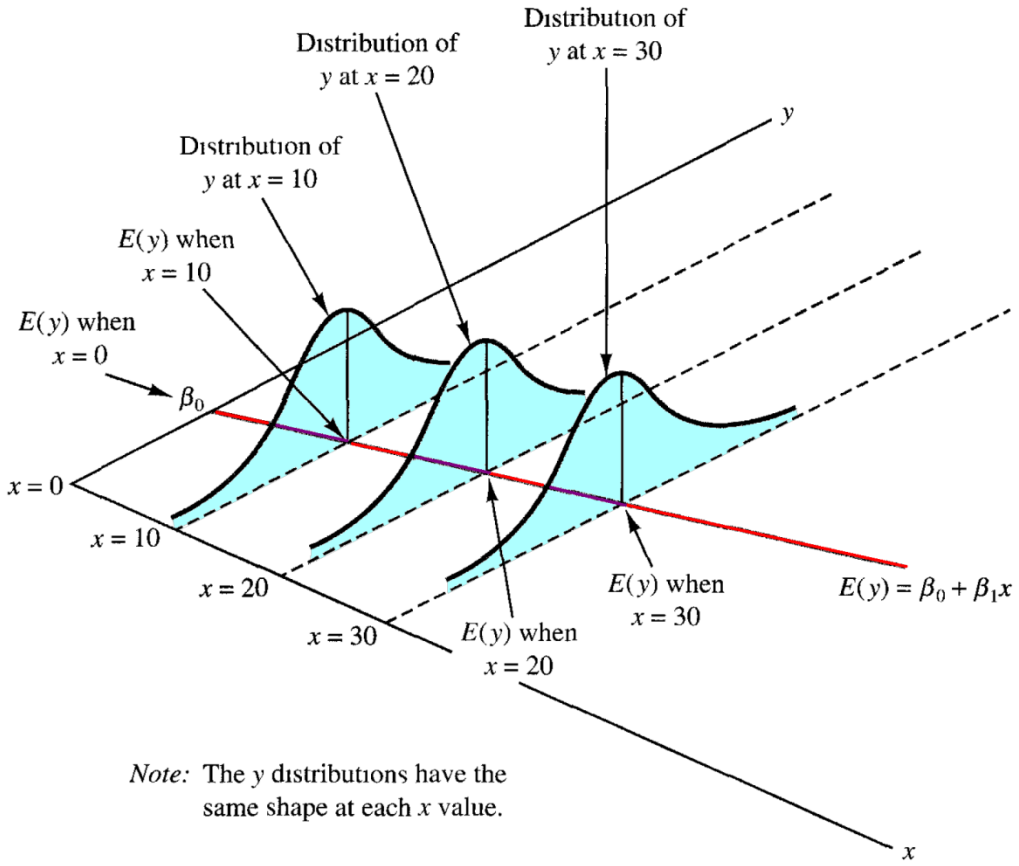
Linear Regression

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 term ε is a normally distributed variable
4. **The values of ε are independent**

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



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

Linear Regression

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.

Linear Regression

Example


```
mod = lm(Cell.Number~Temperature,data=cells)
summary(mod)

    Call:
    lm(formula = Cell.Number ~ Temperature, data = cells)

    Residuals:
        Min      1Q  Median      3Q     Max
    -49.183 -26.190  -1.185  22.147  54.477

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept) -190.784     35.032  -5.446 2.96e-05 ***
    Temperature   15.332      1.145  13.395 3.96e-11 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 31.76 on 19 degrees of freedom
    Multiple R-squared:  0.9042,    Adjusted R-squared:  0.8992
    F-statistic: 179.4 on 1 and 19 DF,  p-value: 3.958e-11


coef(mod)
(Intercept) Temperature
 -190.78355    15.33247
```

In R use the function:

- `mod = lm(x,y)`
- `summary(mod)`

`cells.txt`

Linear Regression

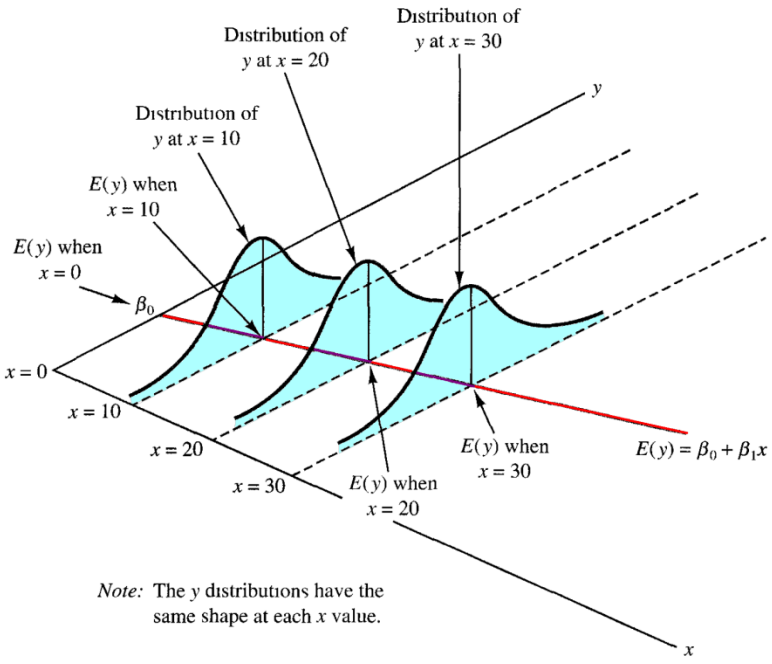
Confidence and Prediction

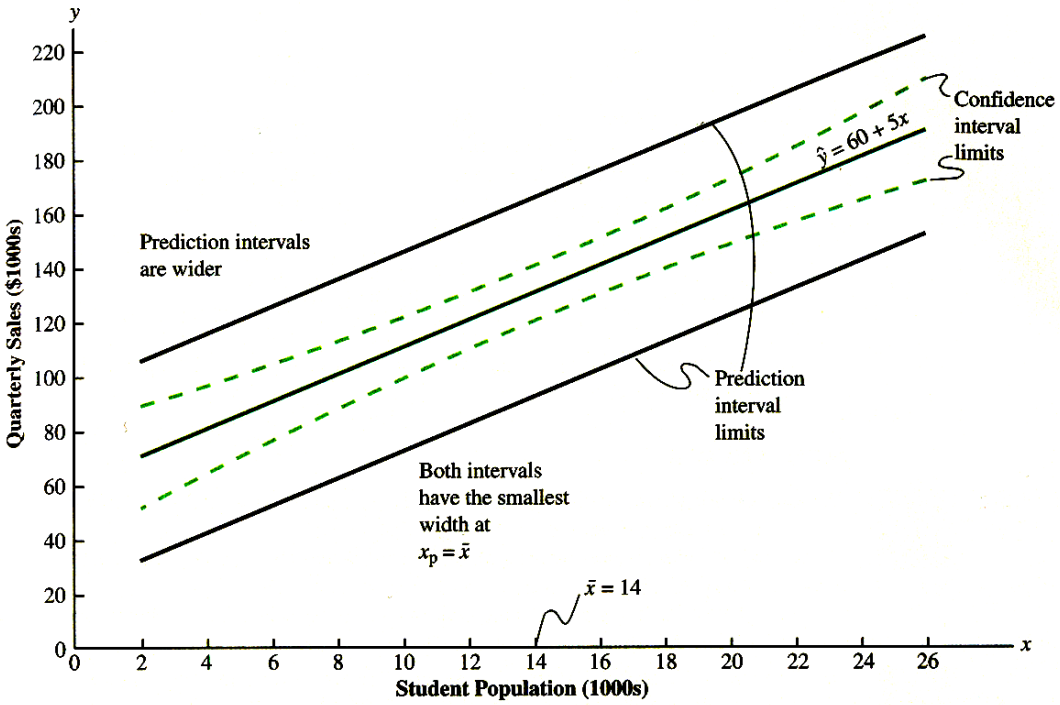
Confidence interval

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.

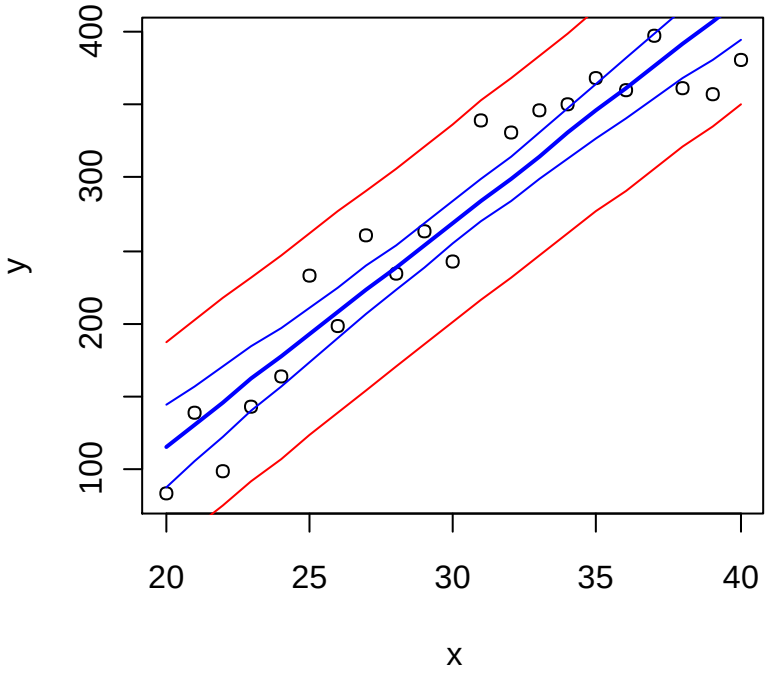




Linear Regression

Example

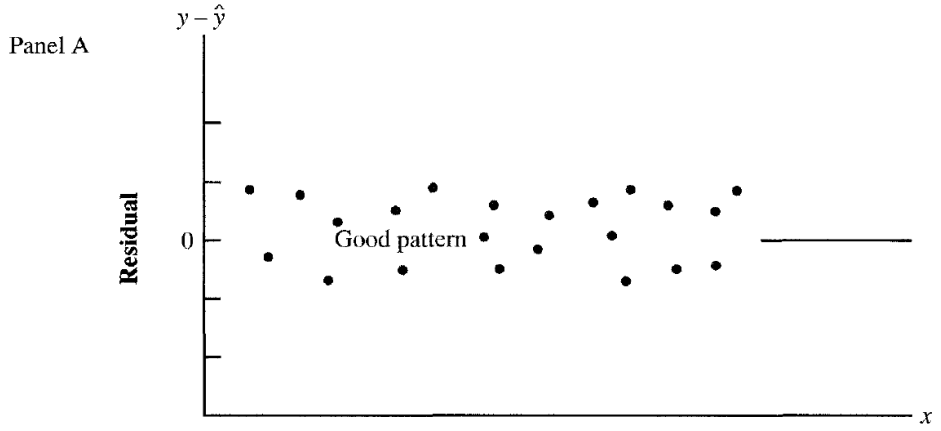
cells.txt

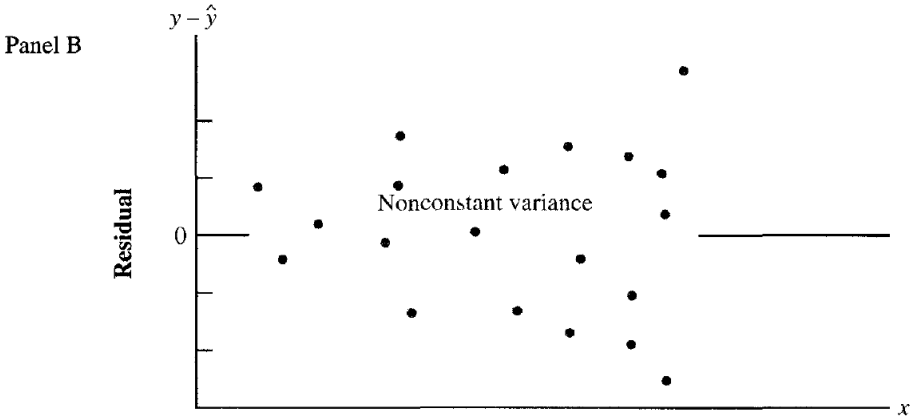


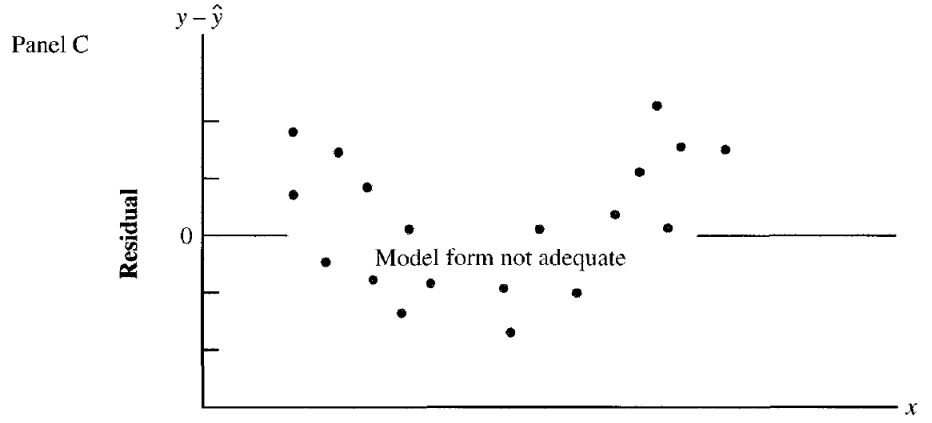

```
x = data$Temperature
y = data$Cell.Number
res = lm(y~x)
res
summary(res)

# 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)
```

Linear Regression

Residuals

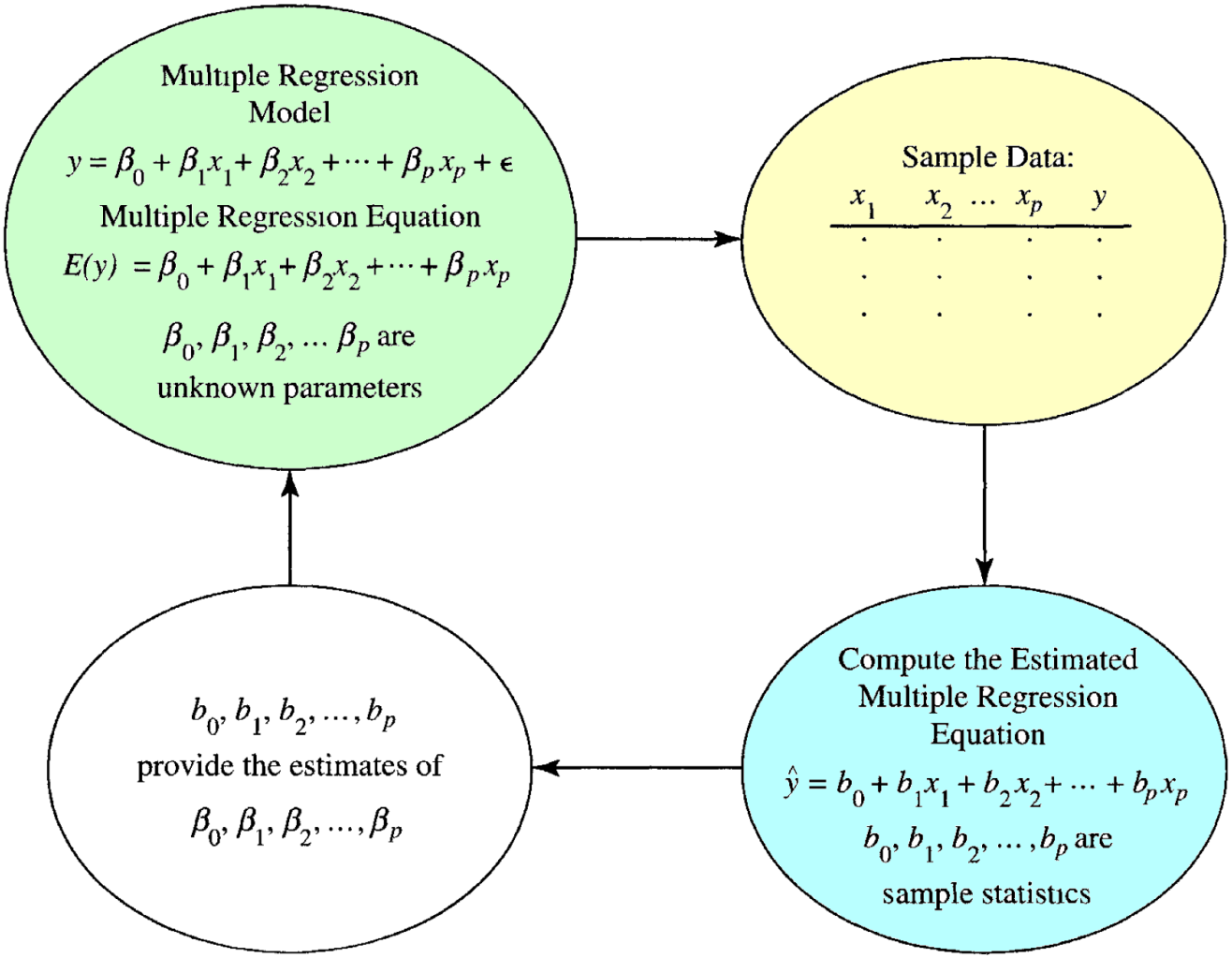






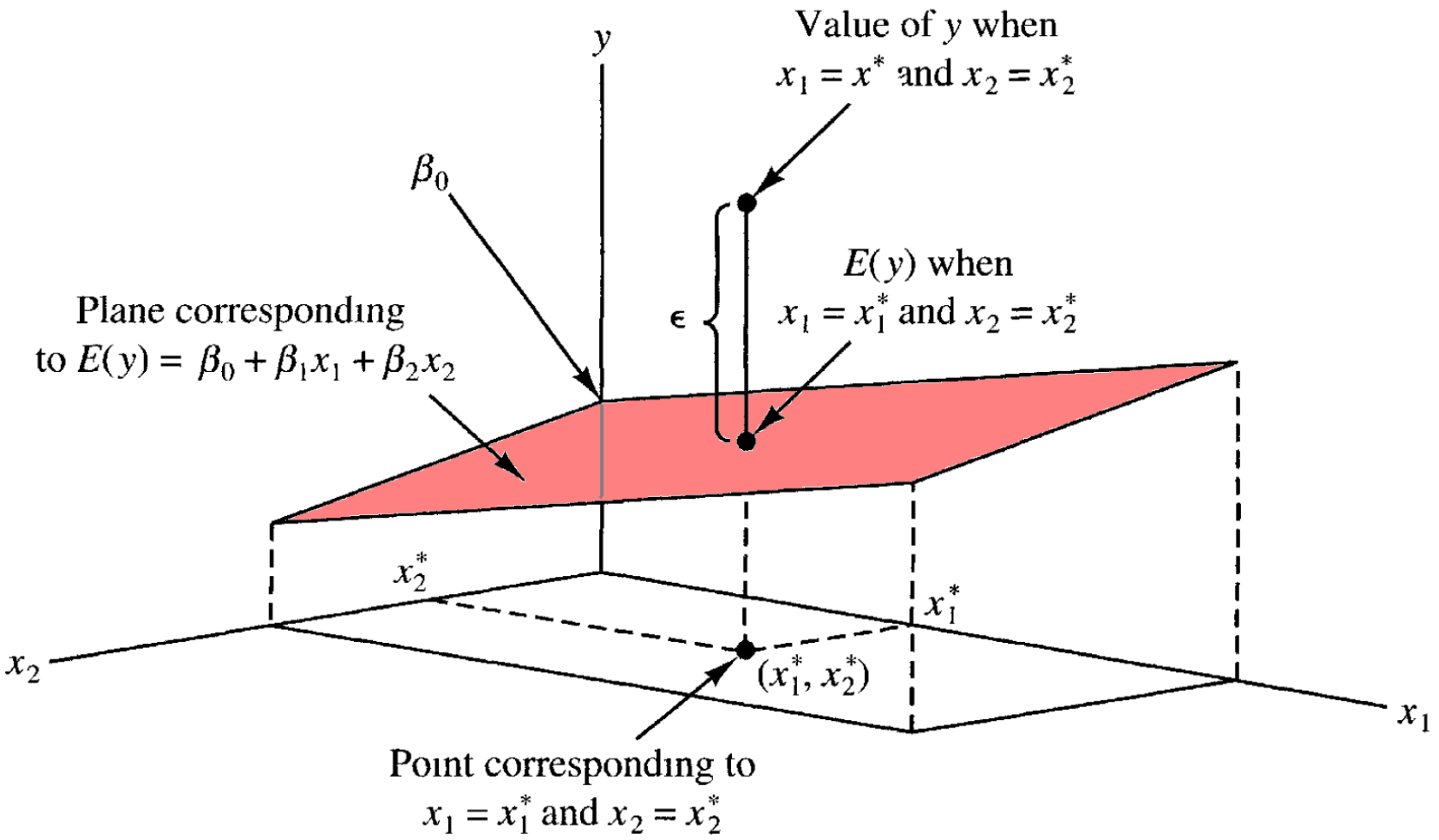
Linear Regression

Multiple Regression



Linear Regression

Multiple Regression



Linear Regression

Multiple Regression


```
pdac = read.table("http://edu.modas.lu/data/txt/pdac.txt",header=TRUE,sep="\t")
> str(pdac)
'data.frame':   253 obs. of  6 variables:
 $ Sample.Code: chr  "ductaladenocarcinoma-Ge0014" "ductaladenocarcinoma-Ge0017" "ductaladenocarcinoma-
Ge0018" "ductaladenocarcinoma-Ge0022" ...
 $ State      : chr  "cancer" "cancer" "cancer" "cancer" ...
 $ TGFBI      : num  12.9 12.6 12.7 12.2 12.6 ...
 $ PEA15      : num  11.9 11.7 11.6 11.8 12 ...
 $ SERPINI2   : num  5.94 11.79 10.54 10.24 6.1 ...
 $ ZD52F10    : num  7.66 8.23 7.15 7.5 7.11 ...

mod2 = lm(PEA15 ~ TGFBI + SERPNI2 + ZD52F10, data = pdac)

summary(mod2)

Call:
lm(formula = PEA15 ~ TGFBI + SERPINI2 + ZD52F10, data = pdac)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05868 -0.27414 -0.00205  0.29654  1.32305

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.76233    0.62810  15.543  < 2e-16 ***
TGFBI        0.23927    0.03469   6.896 4.38e-11 ***
SERPINI2    -0.03654    0.01612  -2.268   0.0242 *
ZD52F10     -0.10680    0.04231  -2.524   0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4474 on 249 degrees of freedom
Multiple R-squared:  0.4491,    Adjusted R-squared:  0.4425
F-statistic: 67.67 on 3 and 249 DF,  p-value: < 2.2e-16
```

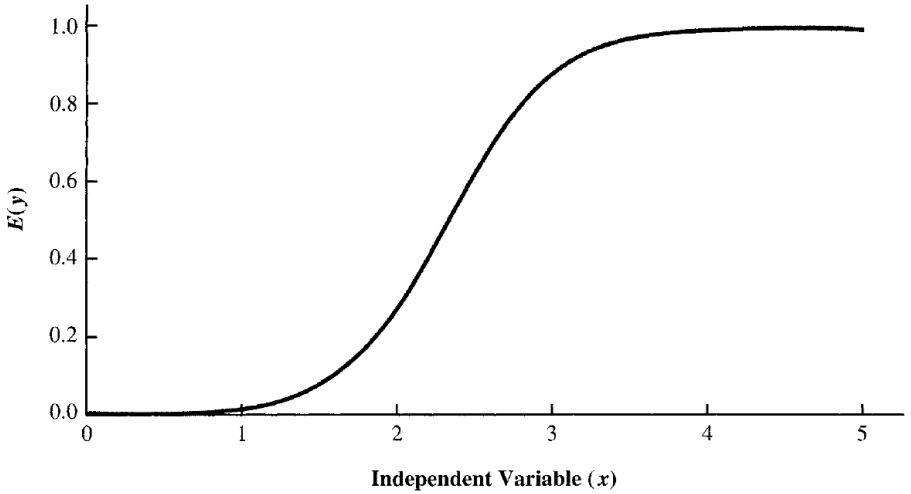
PEA15

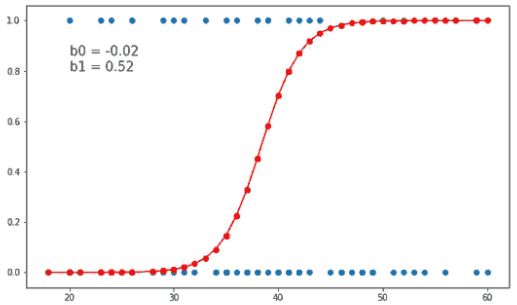
This gene encodes a death effector domain-containing protein that functions as a negative regulator of apoptosis.

Regression Analysis

Logistic Regression

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





$$E\left[y\left(x_1, x_2, \ldots, x_p\right)\right] = P\left(y = 1 \middle| x_1, x_2, \ldots, x_p\right) = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right)}}$$

This is often used for **classification**