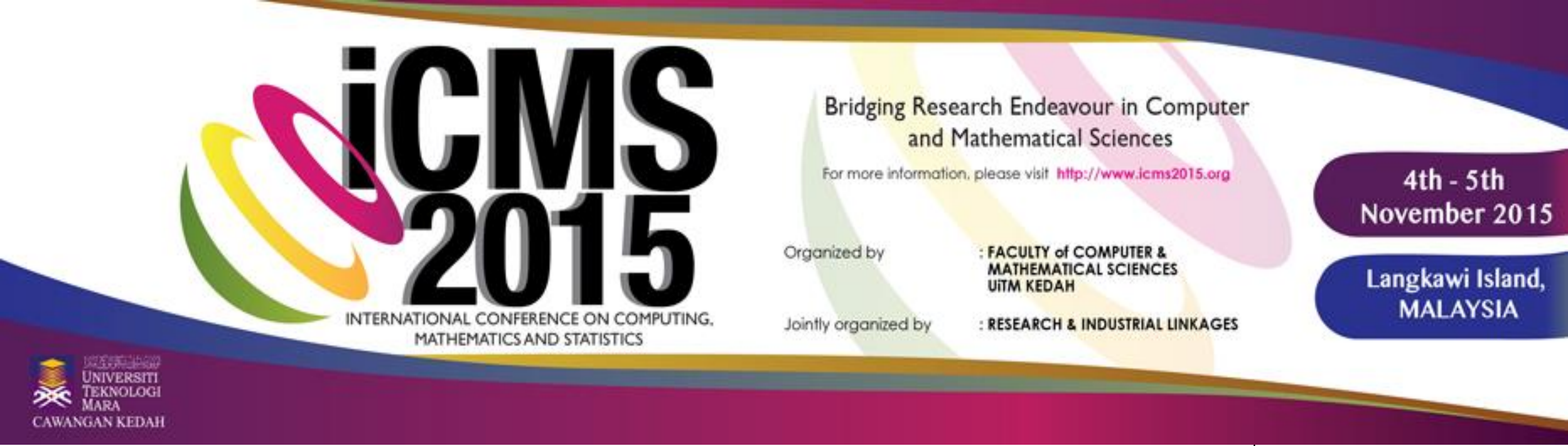


Regression Analysis with R

PRE-CONFERENCE WORKSHOP
"Introduction to R and Data Visualization“
2 - 3 November 2015

Nicoleta Caragea
Department of Economics
Ecological University of Bucharest

Why R for regression analysis?

R is the key solving the complicated regression equation in a common language for:

- academic statisticians,
- scientists,
- engineers,
- data analysts, but also
- less technical individuals with degrees in non-quantitative fields such as the social sciences or business.

Before you start a regression analysis

The formulation of a problem is often more essential than its solution

(Albert Einstein)

To formulate the problem correctly, you must:

- Understand the physical background and the objective of the problem - You may find that simple descriptive statistics is very useful to have a preliminary analysis
- Put the problem into statistical terms. Once the problem is translated into the language of Statistics, the solution is often routine.
  - Difficulties with this step explain why Artificial Intelligence techniques do not yet solve the problems by themselves.

Before you start
a regression analysis

- Understand how the data was collected
- Be sure you are able to “read” the results and evaluate the performance of the model

Regression analyses have several possible objectives

1. *Assessment of the possible correlation between explanatory variables and the response*
2. *Explain the effect of explanatory variables on the response.*
3. *Prediction of future observations.*

When to use Regression Analysis?

Regression analysis is used for explaining or modeling the relationship between

- *a single variable Y, called the response, output or dependent variable, and*
- *one or more explanatory variables* (X_i), named also *predictors, input, or independent variables*

Different regression analysis if the type of variables are different, or the relationship is different

Linear regression	Y – continuous variable	Simple linear	Xi
			i=1
			X – continuous/categorical/both
		Multiple linear	Xi
			i>1
			X – continuous/categorical/both
Logistic regression	Y- discrete variables	Binomial simple logistic	Xi
			X – continuous/categorical/both
			i=1
			Yk - binary variable
			k=2
		Binomial multiple logistic	Xi
			i>1
			X – continuous/categorical/both
			Yk - binary variable
			k=2
		Multinomial multiple logistic	Xi
			i>1
			X – continuous/categorical/both
			Yk - categorical variable
			k>2

Linear regression models

General form for the model:

$$Y = f(X) + \varepsilon$$

Statistical model:

$$y_i = \beta_0 + \beta_1 \times x_{1i} + \beta_2 \times x_{2i} + \ldots + \beta_n \times x_{ni} + \varepsilon_i$$

where β_i , i = 0,1, ...,n - unknown parameters giving the effect of X variables on Y
β_0 – intercept (point in which the regression line intercepts the y-axis)
β_i – slope (for x_i)

ε_i – is named the **Error term** representing that part of Y that cannot be explained using the auxiliary information
or Residuals: The differences between the predicted and observed value of response.

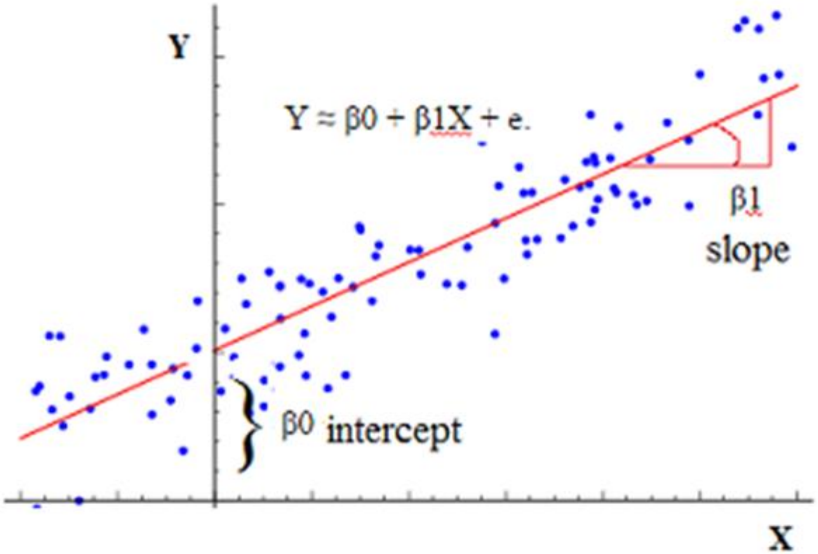
Estimating β – OLS Method

- The problem is to find unknown parameters β such that βX is close to Y (to minimize ε) – Ordinary Least Squares Method

- $\hat{\beta}$ is the best estimate of β within the model when errors have minim values (in fact, residuals sum of squares):

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min$$

Estimating β



The simplest regression model is a linear model with a unique explanatory variable, which takes the following form:

$$y_i = \beta_0 + \beta_1 \times x_{1i} + \varepsilon_i$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \quad \Rightarrow \quad \begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

... R can perform regression quite easily!

lm function in R (stats package)

data should always be inspected for:
•missing values
•outliers

- first stage is to arrange your data (e.g. in a .CSV file). Use a column for each variable and give it a meaningful name.
- second stage is to read your data file into memory
- next stage is to attach your data set so that the individual variables are read into memory.
- finally, we need to define the model and run the analysis.

the attach function will mask objects attached before

Application 1. (NRG_Data.csv)

- Import the data (use R Studio). View dataset.
- Run Summary statistics.
- Preliminary data visualization:
  - histograms for GDP and NRG.
  - Plot NRG versus GDP.
- Estimate parameters for the regression: $NRG = \beta_0 + \beta_1 \times GDP + \varepsilon$
- Interpret the regression results.
- What is the 95% confidence interval for the estimated parameters?
- Plot the residuals.
- Predict NRG

Application 1. (NRG_Data.csv)

- Enter the data

The file contains cross-section data on

- NRG - aggregate energy consumption (thousand Tone of Oil Equivalent -TOE) and
- GDP (Million Euro) for 31 countries, in 2014.

Data source: http://ec.europa.eu/eurostat/data/database.

```
> ENERGY <- read.csv(file.choose(), head= TRUE)

> head(ENERGY)
   country      NRG       GDP
1  Belgium  56727.5  395242.0
2 Bulgaria  16763.7   41047.9
3  Czech_R  42191.3  156932.6
4  Denmark  18101.2  252938.9
5 Germany  324271.5 2809480.0
6 Estonia   6702.7   18738.8
```

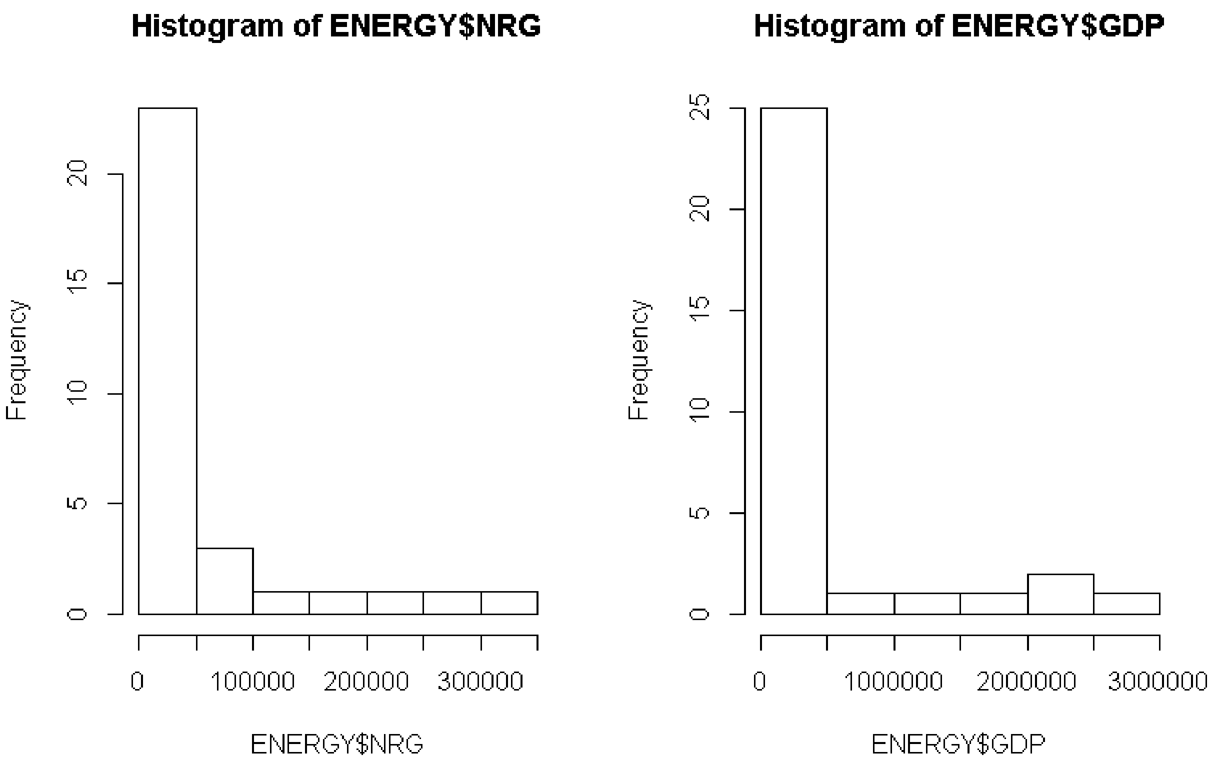
Application 1. (NRG_Data.csv)

- Run Summary statistics.

```
> summary(ENERGY$NRG)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    839    7348   22740   55410   52930  324300 
> summary(ENERGY$GDP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   7508   38600  169400  450500  395700 2809000 
```

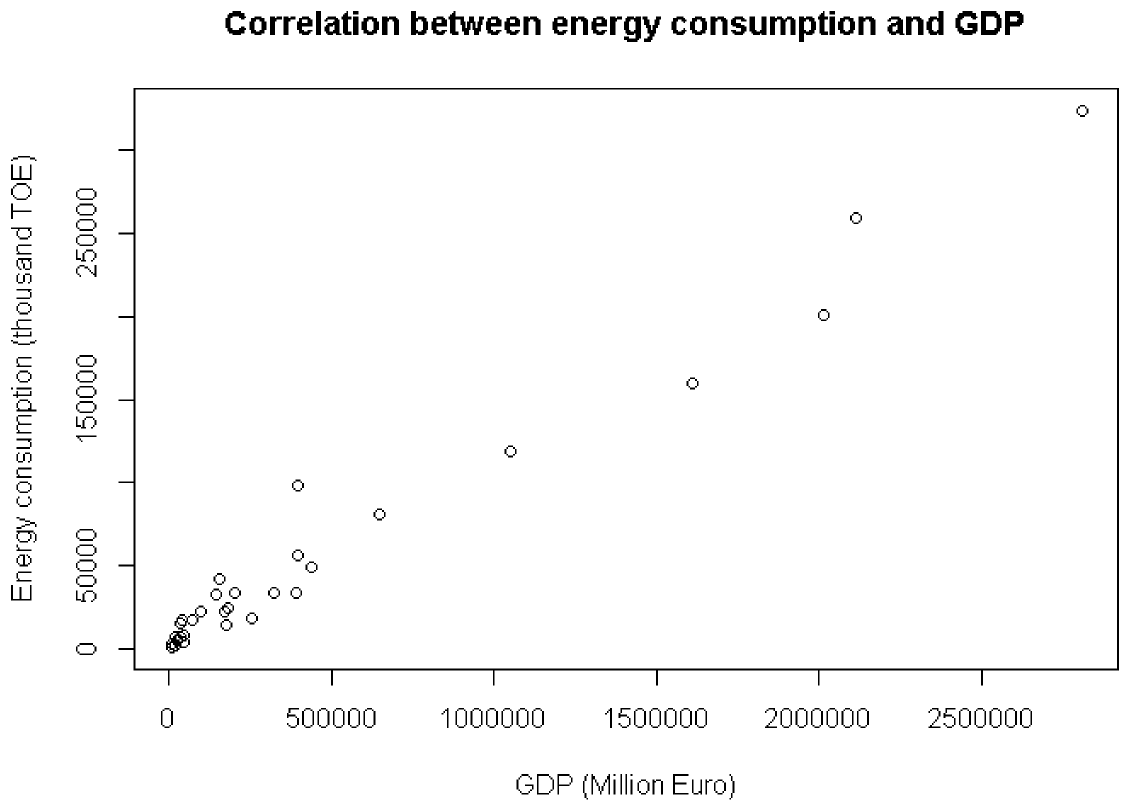
Application 1. (NRG_Data.csv)

- Plot the histograms for GDP and NRG.

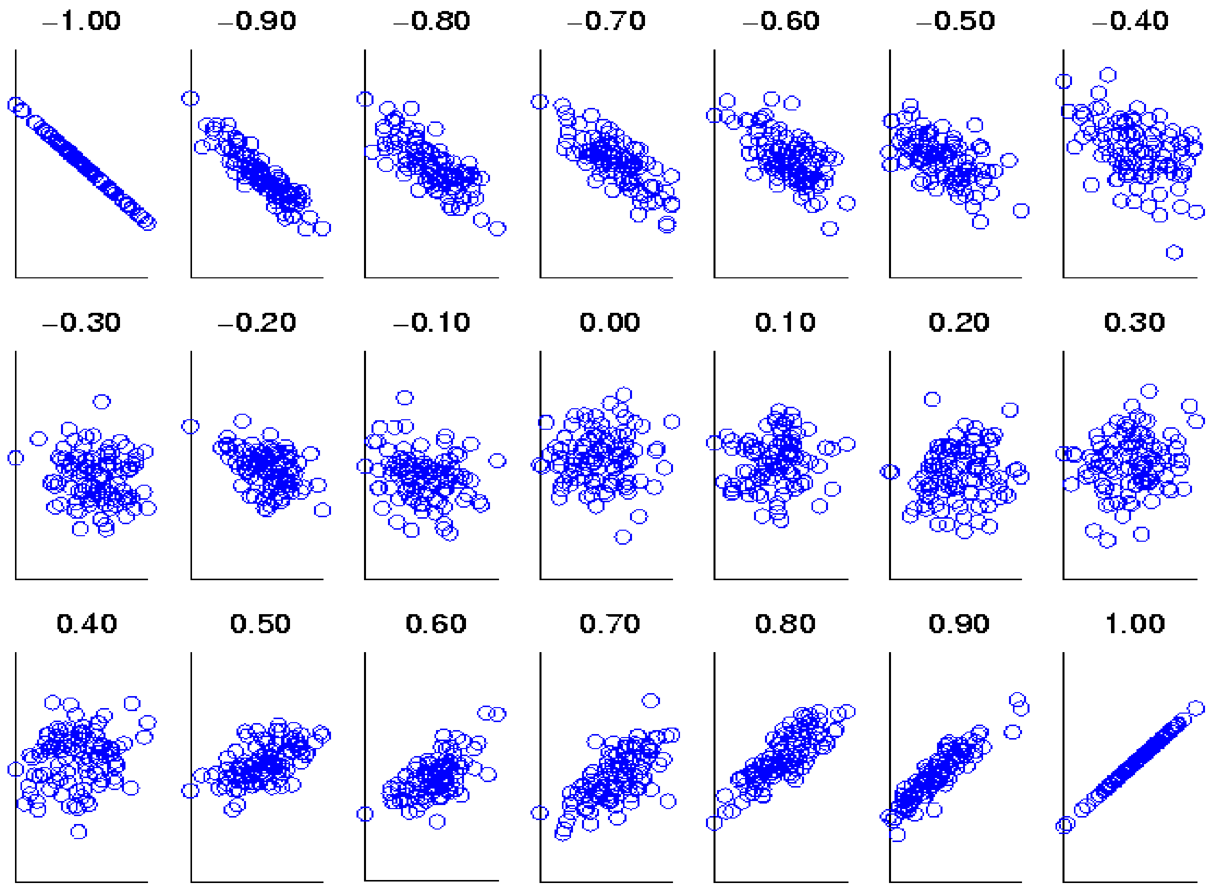


Application 1. (NRG_Data.csv)

- Plot NRG versus GDP



Correlation



$$r_{yx} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Correlation:
- **less than or equal to 0.20 is characterized as very weak;**
- **greater than 0.20 and less than or equal to 0.40 is weak;**
- **greater than 0.40 and less than or equal to 0.60 is moderate;**
- **greater than 0.60 and less than or equal to 0.80 is strong;**
- **greater than 0.80 is very strong.**

Application 1. (NRG_Data.csv)

- Estimate the regression:

$$NRG = \beta_0 + \beta_1 \times GDP + \varepsilon$$

```
> regression <- lm(NRG ~ GDP, data = ENERGY)
> summary(regression)
Call:
lm(formula = NRG ~ GDP, data = ENERGY)
Residuals:
   Min     1Q Median     3Q    Max 
-25696  -6026  -2104   5610  48699 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.139e+03  2.987e+03   2.056   0.0489 *  
GDP         1.094e-01  3.579e-03  30.559   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 14000 on 29 degrees of freedom
Multiple R-squared:  0.9699,    Adjusted R-squared:  0.9688 
F-statistic: 933.8 on 1 and 29 DF,  p-value: < 2.2e-16
```

Application 1. (NRG_Data.csv)

- Interpret the regression results

```
> regression <- lm(NRG ~ GDP, data = ENERGY)
> summary(regression)
Call:
lm(formula = NRG ~ GDP, data = ENERGY)
Residuals:
   Min     1Q Median     3Q    Max
-25696  -6026  -2104   5610  48699

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.139e+03  2.987e+03   2.056   0.0489 *
GDP         1.094e-01  3.579e-03  30.559   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 14000 on 29 degrees of freedom
Multiple R-squared:  0.9699,    Adjusted R-squared:  0.9688
F-statistic: 933.8 on 1 and 29 DF,  p-value: < 2.2e-16
```

Slope: estimated difference in the NRG with one unit increase of GDP (positive relationship)

Level of significance for the estimated value of the coefficient

the explanatory variable (GDP) explains 96.99% of response variation (NRG).

F-statistic - the probability of the F statistic for the overall regression relationship is <0.05. We reject the null hypothesis that there is no relationship between the independent variable and the dependent variable ($R^2 = 0$). We support the research hypothesis that there is a statistically significant relationship between the variables.

ANOVA

```
> anova(regression)
Analysis of Variance Table

Response: NRG
          Df     Sum Sq    Mean Sq F value    Pr(>F)    
GDP        1 1.8296e+11 1.8296e+11  933.84 < 2.2e-16 ***
Residuals 29 5.6817e+09 1.9592e+08                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA - Partitioning of the sum of squares

Source of Variation	Sum of Squares	df	Mean Square	F-test
Regression (variation explained)	$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	k=1	$MSR=\frac{SSR}{k}$	$\frac{MSR}{MSE}$
Errors/Residuals (variation not explained)	$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}e_i^{\,2}$	n-k-1	$MSE=\frac{SSE}{n-k-1}$	
Total	$SST=\sum_{i}(y_i-\bar{y})^2$	n-1	$\frac{SST}{n-1}$	

R^2 represents the proportion of the total sample variability explained by the regression model.

$$R^2=\frac{SSR}{SSR+SSE}$$

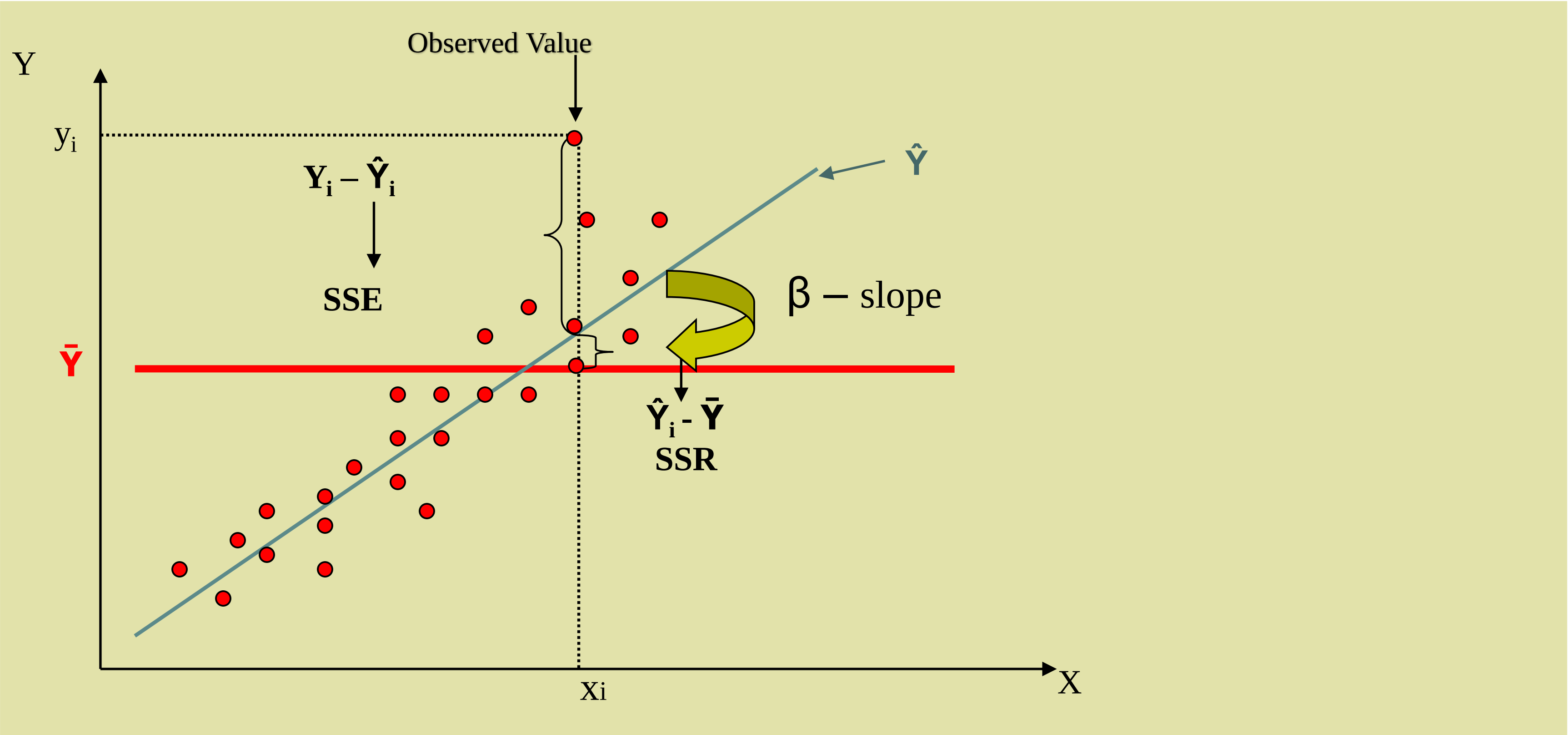
n – number of observations (=31)
k – number of explanatory variables (=1)

$$AdjustedR^2=1-\frac{n-1}{n-k-1}\times(1-R^2)$$

or

$$AdjustedR^2=1-\frac{SSE/df_{Error}}{(SSR+SSE)/df_{Total}}$$

SSE?
SSR?



Total sample Variability = Variability explained by the model + Unexplained (or error) variability

(SST) = (SSR) + (SSE)

There are some other functions in R that allow you to extract elements from a linear model fit.

```
> coef(regression)
 (Intercept)          GDP
6139.1508517    0.1093652
```

```
> resids <- resid(regression)
> head(resids)
        1          2          3          4          5          6
  7362.637   6135.338  18889.187 -15700.659  10873.064  -1485.823
```

Application 1. (NRG_Data.csv)

- What is the 95% confidence interval for the estimated parameters?

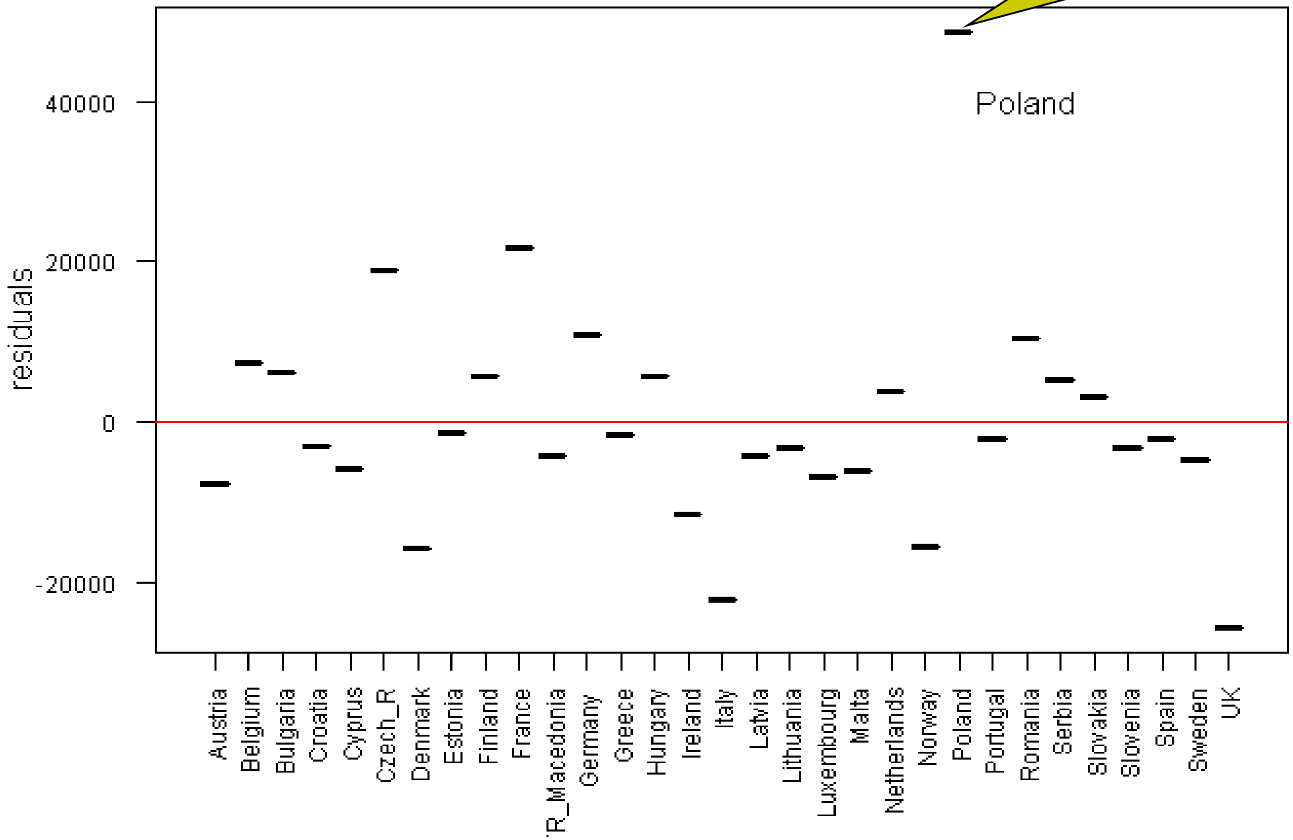
```
> confint(regression)
                   2.5 %       97.5 %
(Intercept) 30.9065066 1.224740e+04
GDP          0.1020456 1.166848e-01
```

The slope of the regression (0.1093652) line is in the range 0.1020456 – 0.116848.

Application 1. (NRG_Data.csv)

- Plot the residuals

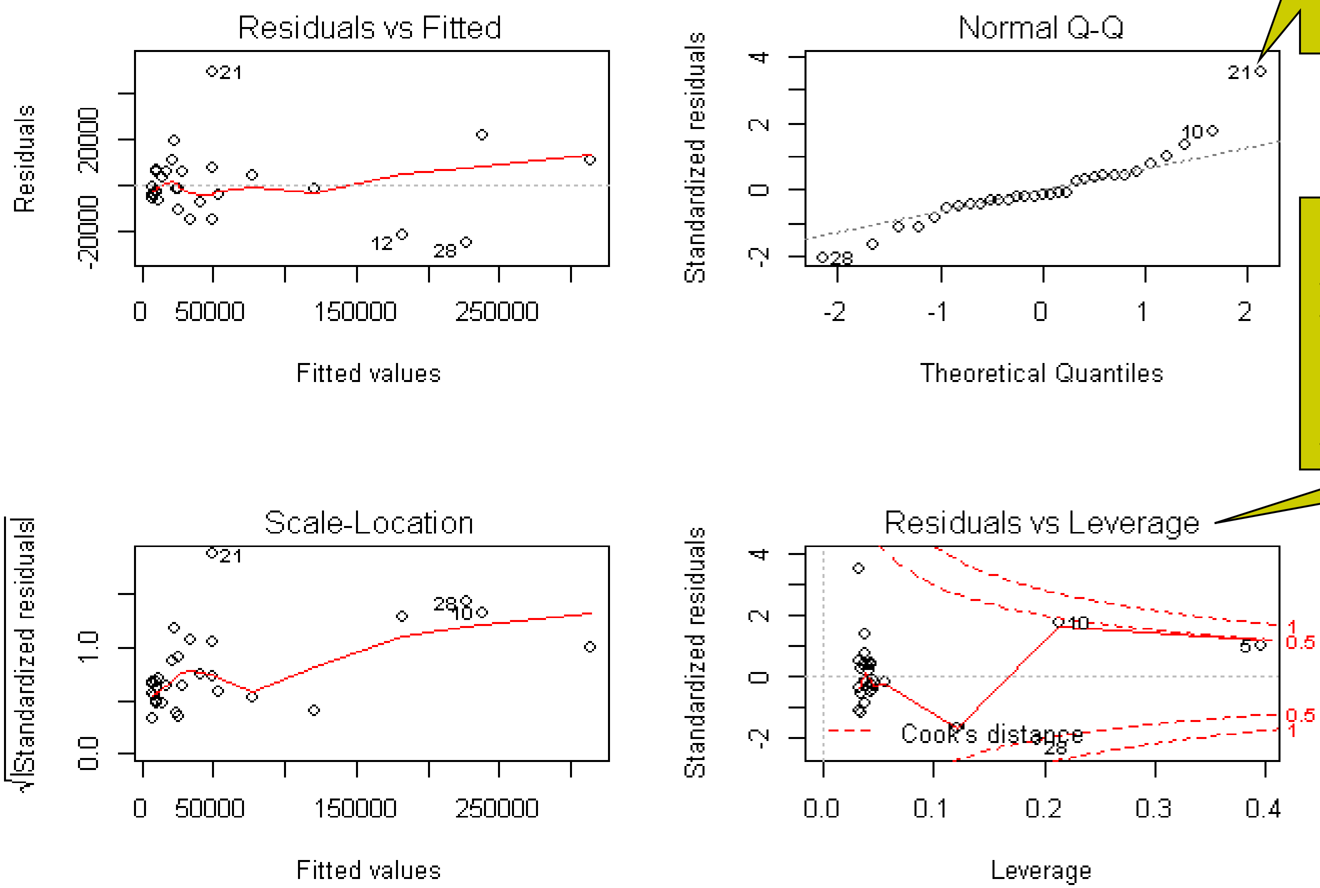
a good linear relationship is compromised by a single outlier.



Application 1. (NRG_Data.csv)

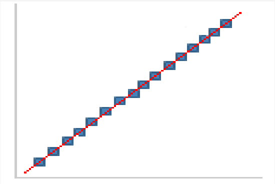
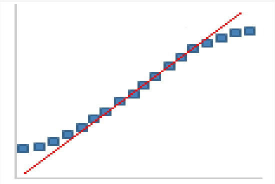
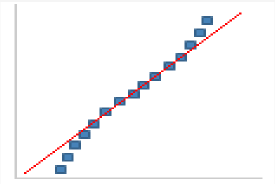
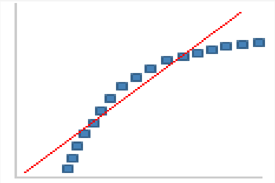
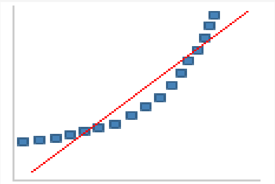
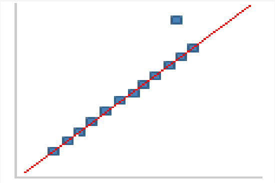
- Plot the regression

Approximately a normal distribution (but more variance than expected)
Outlier - (21. Poland)
This point radically influences the slope on the regression, and we'd expect this to show up as a large Cook's distance.



If the Cooks distance line encompasses a data point, it suggests that the analysis may be very sensitive to that point and it may be prudent to repeat the analysis with those data excluded. A leverage point is defined as an observation that has a value of x that is far away from the mean of x.

Q-Q plot shapes (examples)

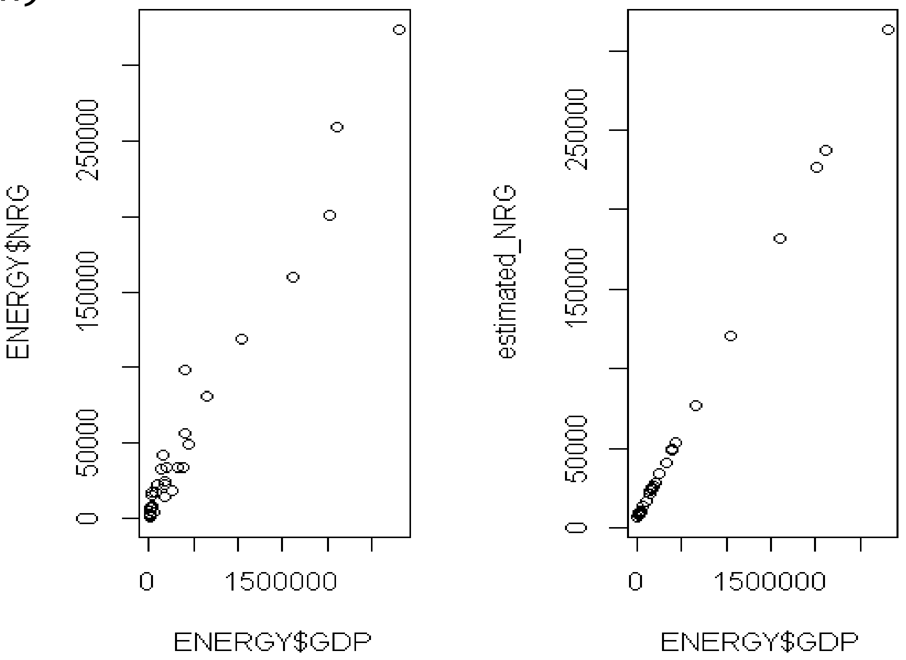
Shape (exaggerated)	Conclusion
	Approximately normal distribution.
	Less variance than expected. While this distribution differs from the normal, it seldom presents any problems in statistical calculations.
	More variance than you would expect in a normal distribution.
	Left skew in the distribution.
	Right skew in the distribution.
	Outlier. Outliers can disturb statistical analyses and should always be thoroughly investigated. If the outliers are due to known errors, they should be removed from the data before a more detailed analysis is performed.

Application 1. (NRG_Data.csv)

- Predict NRG

predictions with functions *fitted* or *predict*

```
> fitted_NRG <- fitted (regression)
> estimated_NRG <- predict(regression)
```



Predictions - NRG for a target value of GDP

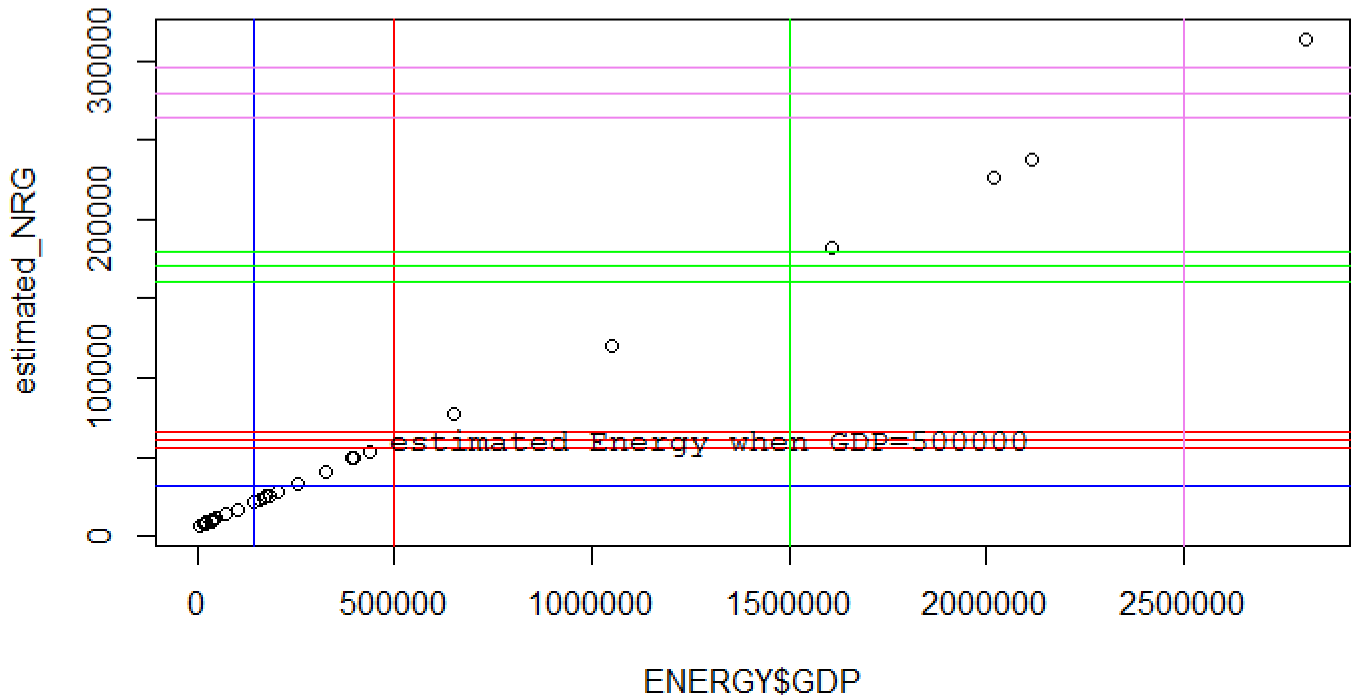
The observed values for NRG and GDP in Romania are:

NRG	GDP
32346.0	144282.2

the predicted value of NRG in the case of GDP=500000?

$$NRG = \hat{\beta}_0 + \hat{\beta}_1 \times GDP$$

$$= 6139.1508517 + 0.1093652 \times 500000 = 60821.75$$



Plot predicted NRG and Confidence interval

```
# plot Romanian observation: GDP=144282.2 and ENERGY=32346.0
> plot(ENERGY$GDP, estimated_NRG)
> abline(v=144282.2,col="blue")
> abline(h=32346.0,col="blue")

#predict NRG value for GDP=500000
> estimated_NRG3 <- predict(regression, newdata =
data.frame(GDP=c(500000)), interval = "confidence")

#plot estimated ENERGY
> abline(v=500000,col="red")
> abline(h=estimated_NRG3,col="red")
```

Application 2 (HBS_youth.csv)

HBS – Household Budget Suvey (Romania, 2013)

- Import the data.
- Plot the histograms/density for expenditure and income.
- Plot expenditure versus income.
- Estimate the regressions:
  - expenditure = f (income)
  - expenditure =f (income, age)
- Interpret the regression results.
- What is the 95% confidence interval for the estimated parameters?
- Plot the residuals.
- Predict expenditure

Generalized Linear Models – GLM
Logistic Regression

Why use logistic regression?

To predict a non-numerical value of dependent variable

(y – is a categorical or qualitative output)

1. **in the simplest case scenario y is binary - the model is "binomial logistic regression"**

For example:

- *voting (vote/not vote)*
- *unemployment (unemployed/employed)*
- *smoking (yes/not)*
- *poverty (*rich/poor)

2. **if y assumes more than 2 categories – the model is named "multinomial logistic regression"**

For example:

- multiple response (yes/no/don't know/refuse).

- The predictors (xi) can be continuous, categorical or a mix of both

Binomial logistic regression

- Model:

$$Y = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

- *p* is the probability that the event Y occurs, p(Y=1)
- *1-p* is the probability that no event occurs, p(Y=0)

- p/(1-p) is the "odds"
- ln[p/(1-p)] is the log odds, or "logit"

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Multiple Logistic Regression

- Extension to more than one predictor variable
  - With *k* predictors, the model is written:

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

…some explanation

- Suppose we are interested in estimating the proportion of unemployed persons in a population. Naturally, we know that entire population do not have equal probability of ‘success’ (i.e. being employed). Lower educated people is more likely to be unemployed. Consider the predictor variable X to be any of the risk factor that might contribute to the unemployed status. Probability of success will depend on the levels of the risk factors.

Odds Ratio

- Interpretation of Regression Coefficient (β):
  - If in linear regression, the slope coefficient is the change in the mean response as *x* increases by 1 unit
  - In logistic regression, we can show that:

$$OR = \frac{odds(x+1)}{odds(x)} = e^{\beta} \qquad \left(odds(x) = \frac{p(x)}{1 - p(x)} \right)$$

Thus e^{β} represents <u>the change in the odds of the outcome</u> by increasing x by 1 unit (holding all other predictors constant)

Interpretation of OR= e^{β}

- If $\beta = 0$, the probability is the same at all x levels ($e^{\beta}=1$)
- If $\beta > 0$, the probability increases as x increases ($e^{\beta}>1$)
- If $\beta < 0$, the probability decreases as x increases ($e^{\beta}<1$)

Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the coefficients of a model.

MLE involves:

- finding the coefficients (βk) that makes the log of the likelihood function (LL < 0) as large as possible (maximize the probability that event to occur)
- or, finds the coefficients that make -2 times the log of the likelihood function (-2LL) as small as possible

Logistic regression implementation in R

- R makes it very easy to fit a logistic regression model.
- The function to be called is glm()
  - nlme package

Application 3. (EUSILC.csv)

fitting a binary logistic regression model

- EU_SILC-European Survey on Income and Living Condition
- Predict the probability of poor vs. rich people, as a function of some characteristics of population (income, occupational status, age, sex, family size, education level, civil status, regional distribution, residence area).

Dataset and model description

```
> head(EUSILC)
  year Age Sex poverty Income_EQ civil_status Education Occup area_resid family_size  REG
1 2011  31   1       1      8843            2         7     1          1           2 RO12
2 2011  88   2       1      8843            4         3     6          1           2 RO12
3 2011  86   2       0     16190            4         2     6          1           1 RO12
4 2011  68   1       1     12433            1         3     6          1           2 RO12
5 2011  63   2       1     12433            1         6     1          1           2 RO12
6 2011  54   2       1      6560            5         6     1          1           2 RO12
```

Dataset – number of observations: 17370

Y – poverty (0 = rich, 1=poor)

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times Income\_EQ + \varepsilon$$

The simplest Logistic Model:

p - probability to be poor, p(Y=1)

1-p - probability to be rich, p(Y=0)

results

```
> mylogit <- glm(poverty ~ Income_EQ, data=EUSILC)
> summary(mylogit)
glm(formula = poverty ~ Income_EQ + Sex + Age + Education, family =
"binomial",
    data = EUSILC)
Deviance Residuals:
    Min        1Q    Median        3Q       Max
-3.9225   0.0098   0.0399   0.1563   3.6779

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.350e+01  3.134e-01  43.079  < 2e-16 ***
Income_EQ   -6.904e-04  1.525e-05 -45.279  < 2e-16 ***
Sex         -7.915e-02  7.676e-02  -1.031    0.302
Age          1.370e-02  2.074e-03   6.607 3.93e-11 ***
Education   -6.698e-01  1.989e-02 -33.668  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

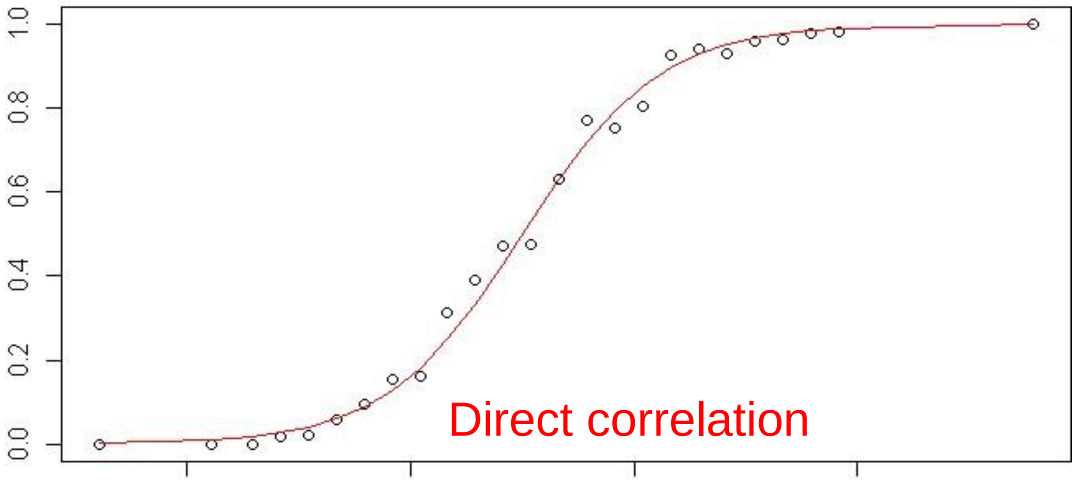
Interpreting the results of logistic regression model

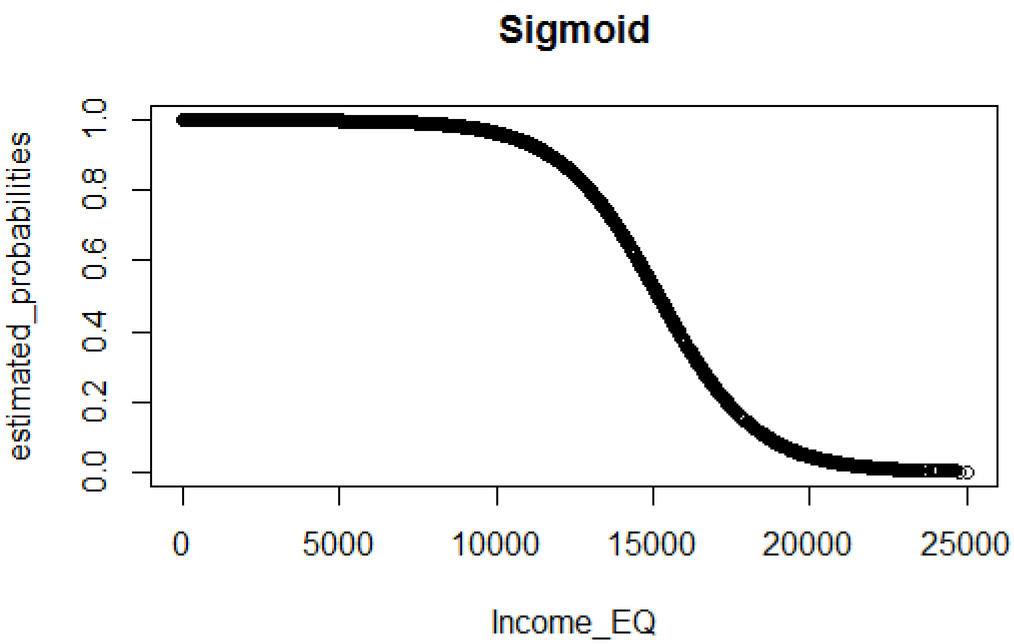
The regression coefficients for each term are the log of the odds ratio for that term, so that the estimated odds ratio is *e* raised to the power of the regression coefficient.

```
> exp(coef(mylogit))
 (Intercept)    Income_EQ
1.544615e+04 9.993647e-01
```

- One-unit increase of income produces a decrease of probability to be in poverty risk to 99.99%

S-Shape





$\beta < 0$, the probability to be poor decreases as *Income* increases ($e^{\beta}<1$)

More factors of poverty

```
Call:
glm(formula = poverty ~ Income_EQ + Sex + Age + Education, family = "binomial",
    data = EUSILC)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9225   0.0098   0.0399   0.1563   3.6779

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.350e+01  3.134e-01  43.079  < 2e-16 ***
Income_EQ   -6.904e-04  1.525e-05 -45.279  < 2e-16 ***
Sex         -7.915e-02  7.676e-02  -1.031    0.302
Age          1.370e-02  2.074e-03   6.607 3.93e-11 ***
Education   -6.698e-01  1.989e-02 -33.668  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> exp(coef(mylogit2))
  (Intercept)    Income_EQ          Sex          Age    Education
7.288401e+05 9.993098e-01 9.238991e-01 1.013798e+00 5.118176e-01
```

Probability to be poor decreases for more educated people

Predicted probabilities to be poor vs. rich

```
> probabilities <- fitted(mylogit2)
> head(probabilities)
        1         2         3         4         5         6 
0.9548332 0.9983938 0.8810402 0.9772202 0.8322706 0.9960622 
```

```
> head(EUSILC)
  year Age Sex poverty Income_EQ civil_status Education Occup area_resid family_size  REG
1 2011  31   1       1      8843            2         7     1          1           2 RO12
2 2011  88   2       1      8843            4         3     6          1           2 RO12
3 2011  86   2       0     16190            4         2     6          1           1 RO12
4 2011  68   1       1     12433            1         3     6          1           2 RO12
5 2011  63   2       1     12433            1         6     1          1           2 RO12
6 2011  54   2       1      6560            5         6     1          1           2 RO12
```

The threshold for Income_EQ=15000, so that sampled person who have 16190 has associated a probability to be poor 0.88

Multinomial logistic regression

y assumes more than 2 categories

One simple way to keep in mind a multinomial logit model is to imagine, for J possible outcomes, running J-1 independent binary logistic regression models

Odds in multinomial regression

$$\Omega = \frac{p_{ij}}{1 - p_{iJ}} = e^{\beta_{j0} + \beta_{j1} x_{i1} + \beta_{j2} x_{i2} + ... + \beta_{jk} x_{ik}}$$

J – number of output categories

j = 1,2,…,(J-1)

summary output

- The model output has a block of coefficients and a block of standard errors.
  - Separate coefficients are computed for independent variables for each category of response.
- Before running our model, we then choose the level of the outcome that we wish to use as our *baseline/reference category* and specify this in the relevel function.

Application 4. (Census.csv)
fitting a binary logistic regression model

- Predict the probability of privileged social class vs. other social categories (middle & disadvantaged social classes)
  - as a function of some characteristics of population (income, occupational status, economic activity, education level, age, sex).

Variables description

CATEGORIA – Professional category:

1= Employers and leaders in big economic and social units

2= Employers and leaders in medium and small economic and social units

3= Persons with public prestige

4= Specialists in technical fields

5= Specialists in services

6= Specialists in traditional occupations

7= Workers and labourers with high skills and vocational training

8= Workers and labourers with medium skills and vocational training

9= Workers and labourers with low skills and vocational training

10= Own account workers in subsistence agriculture

11= (un-qualified) Workers and labourers without skills or vocational training

- SEX

1= male

2= female

EDUC – *Level of education*

1= no school graduated

2= primary education

3= gymnasium

4= professional or apprendiship

5= high-school

6= post high-school or technical foreman

7= tertiary education

STAO - *Occupational status*

1= employee

2= own-account worker (including employer)

3= retired

4= unemployed

5= pupil/student

6= housewife

7= othe inactive person

ACTP – *Activity of national economy*

1= agriculture, silviculture and fishing

2= industry

3= construction

4= transports

5= commercial services

6= social servicies

7= other activities of national economy

Income – *Gross annual income of individuals*

CLASA – *Social classes*

1= privileged class (CATEGORIA=1, 2 and 3)

2= middle class (CATEGORIA= 4-8)

3= disadvantaged class (CATEGORIA= 9-11)

Computing the multinomial logistic regression in R

- Use the multinom function from the nnet package in R.
- There are other functions in other R packages capable of multinomial regression.

```
multinom(formula = CLASA ~ INCOME + SEX + AGE + EDUC + STAO + ACTP, data = census)
```

Interpreting multinomial regression results

The logistic coefficient is the expected amount of change in the logit for each one unit change in the predictor

```
> summary(multinomial_regression)
Call:
multinom(formula = CLASA ~ INCOME + SEX + AGE + EDUC + STAO + ACTP, data = census)

Coefficients:
  (Intercept)        INCOME       SEX        AGE      EDUC      STAO       ACTP
2    16.51398 -7.463532e-05 0.3742069 0.02052921 -2.059134 0.2164003 0.07040719
3    21.25786 -1.660704e-04 0.7702002 0.01689434 -2.824972 0.2264809 0.07704389
```

The closer a logistic coefficient is to zero, the less influence the predictor has in predicting the logit

```
Std. Errors:
   (Intercept)       INCOME          SEX         AGE         EDUC         STAO        ACTP
2 7.979139e-05 3.568506e-06 1.387133e-04 0.002587334 0.0006893810 7.427057e-05 0.001811720
3 3.904086e-05 4.229925e-06 8.510392e-05 0.002771283 0.0002180411 4.707955e-05 0.001810409

Residual Deviance: 37683.02
AIC: 37711.02
```

The Akaike Information Criterion (AIC) is a measure of the relative quality of a statistical model for a given set of data.

```
## odds ratios
> exp(coef(multinomial_regression))
  (Intercept)    INCOME      SEX      AGE       EDUC     STAO     ACTP
2    14856946 0.9999254 1.453838 1.020741 0.12756434 1.241599 1.072945
3  1706753894 0.9998339 2.160199 1.017038 0.05931031 1.254179 1.080089
```

Predicted probabilities

```
> pp <- fitted(multinomial_regression)
> head(pp)
             1         2          3
1 4.271765e-05 0.3536245 0.64633274
2 2.347877e-04 0.4012185 0.59854668
3 8.333801e-02 0.8322419 0.08442006
4 7.804750e-02 0.8252126 0.09673992
5 3.307917e-04 0.4697502 0.52991899
6 2.391351e-04 0.4003440 0.59941687
```

Probability to be in **privileged class**
is almost zero – for individuals having the profile of the first 6 persons in the sample

```
> head(census)
  CATEGORIA CLASA SEX AGE AGROUP EDUC OCUP STAP STAO ACTP   INCOME
1         9     3   1  47      4    4 8331    1    1    4 14646.82
2         9     3   2  43      3    5 5151    1    1    5 13054.57
3         2     1   1  39      3    7 1324    1    1    6 21605.16
4         6     2   2  37      3    7 2631    1    1    5 24360.37
5         9     3   1  47      4    5 5151    1    1    5 11621.23
6         9     3   2  42      3    5 5151    1    1    5 13054.57
```

And…
Why R for regression analysis? Because…

"If you wanted to do research in statistics in the mid-twentieth century, you had to be bit of a mathematician, whether you wanted to or not . . .

If you want to do statistical research at the turn of the twenty-first century, you have to be a computer programmer."

Andrew Gelman

Department of Statistics,Columbia University