Unit 6

Regression

ESIGELEC

Instructor: Federico Perea

Regression analysis

- One of the main problems to be solved in statistics consists of quantifying the relation between different random variables
- For example: a bulb's energy consumption (*Y*) depends on watts *(X_1),* type of material *(X_2),* etc.
- In this section you will model such relation.

- You will predict values of a dependent variable *(Y),* quantify the relation between the dependent variable and a number of explanatory variables *(X_i)*, and measure the importance of each of the explanatory variables.
- Besides, you will identify those variables that do explain the model, and find out how to modify the model so that the real situation is better explained (if necessary).

Examples: Linear regression

- You want to explain fuel consumption of cars (variable *Consumption*) as a function of the horse powers of the car (variable *Power*)

$$Consumption = \beta_0 + \beta_1 Power + U$$

- You want to explain the time a computer will last (variable *Lifespan*) as a function of its price (variable *Price*) and the proportion of the time the computer will be functioning (variable *Use*)

$$Lifespan = \beta_0 + \beta_1 Price + \beta_2 Use + U$$

- U is an error term you need to include in every regression model, because this relationship is not exact!

Assumptions

- The dependent variable (*Y, Consumption, Lifespan*) follows a normal distribution.
- *Y* might depend on another variable *X (Power, Price, Use).*
- All conditional probability distributions are normal. Their mean depend on the explanatory variable value, and their variance is constant

$E(Y/X) = \alpha + \beta X$
$Var(Y/X) = \sigma^2$

The multiple linear regression model

- The proposed model for *k* explanatory variables (also called regressors) is
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$
- Where:
  - *Y* is the dependent variable under study.
  - X_j are the explanatory variables, *j=1,…,k.*
  - β_k are the regression coefficients or model parameters
  - *U* is a random disturbance term, which reflects the effect of other explanatory variables excluded from the model

    For a particular observation *i*, you can write the same model as
    $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + U_i$

Assumptions

- Variable *U* is the *error* or *disturbance*, because it is the difference between the actual value of variable *Y* and its expected value:

$$Y - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$
$$= Y - \hat{Y} = U$$

One would want *U* to be as small as possible. Since *U* is a random variable itself, you want to make:

1. Mean value equal to zero
2. Variance as small as possible

Assumptions

Assumptions, for the models to be correct, regarding the disturbance:

1. All disturbances are random variables with zero mean,
$$E(U_i) = 0 \ \forall \ i$$
2. All disturbances have the same variance,
$$Var(U_i) = \sigma^2 \ \forall \ i$$
3. No serial correlation
$$Cov\left(U_{i_1}, U_{i_2}\right) = 0 \ \ \forall i_1 \neq i_2$$
4. Disturbances follow a normal distribution. This, together with the previous assumptions, conclude that disturbances are independent.
5. Disturbances are independent of the explanatory variables.

Assumptions

Assumptions regarding the explanatory variables

1. Explanatory variables X_j, and dependent variable *Y*, are obtained with no error.
2. Explanatory variables are deterministic (non random)
3. Y_j is the observed value of a linear combination of the values of X_j satisfying:
   $E(Y_j \,/\, X_1 = x_1, \dots, X_k = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$
   $Var(Y_j \,/\, X_1 = x_1, \dots, X_k = x_k)$
   $= Var(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + U_j) = Var(U_j) = \sigma^2$
4. No exact linear relationship between the explanatory variables. If such relationship exists, reduce the number of variables by eliminating one of them.

9

Assumptions

Assumptions regarding the parameters

1. The parameters β are constant for all samples.

- Parameters β quantify the existing relationship between each of the explanatory variables *X* and the dependent variable *Y.*
- In order to know their meaning, it is necessary to take into account the model under consideration

10

Parameters interpretation

If the model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$

1. It can be seen that

$$\beta_j = \frac{E\left(Y \,/\, X_1 = x_1, \ldots, X_j = x_j, \ldots, X_k = x_k\right) - E\left(Y \,/\, X_1 = x_1, \ldots, X_j = x_{j'}, \ldots, X_k = x_k\right)}{x_j - x_{j'}}$$

In words, β_j is the change in the mean value of *Y* per unit change in X_j, holding the other explanatory variables constant.

2. As for β_0, it can be seen that

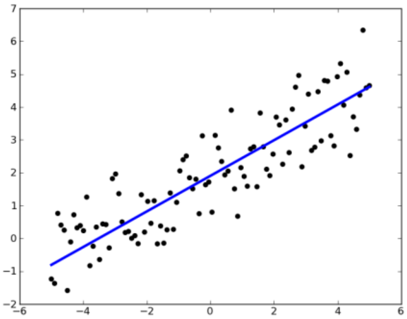
$$\beta_0 = E(Y \,/\, X_1 = 0, \ldots, X_k = 0) .$$

In words, β_0 is the mean value of *Y* when all explanatory variables are zero.

- Interpret the parameters in the following examples:
- $Consumption = \beta_0 + \beta_1 Power + U$
  - β_0 is the average consumption when power is 0
  - β_1 is the increase in average consumption when power increases one unit.
- $Lifespan = \beta_0 + \beta_1 Price + \beta_2 Use + U$
  - β_0 is the average lifespan when price and use are 0
  - β_1 is the increase in average lifespan when price increases one unit and use holds constant.
  - β_2 is the increase in average lifespan when use increases one unit and price holds constant

Estimation of parameters: OLS

- The method of ordinary least squares (OLS) finds the straight line that bests fits the data. A graphical example of Simple Linear regression



Minimizing residuals sum of squares

- $RSS =$ $\sum_{j=1}^{n} e_j^2 \quad = \sum_{j=1}^{n}(Y_j - \hat{Y}_j \ )^2 = \sum_{j=1}^{n}(Y_j - (b_0 + b_1 x_{1,j} + \cdots + b_k x_{k,j}))^2$
- The OLS method finds $b_0, b_1, \ldots, b_k$ minimizing RSS

- Multiple linear regression model:

  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$

- Fitted model:

  $\mathrm{E}[Y] = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$
- The OLS method can be applied if:

1. The number of data must be larger than the number of parameters to be estimated, i.e., *n>k+1.*
2. No exact relationship may exist between the explanatory variables X_j

In the Lifespan example

```
model <- lm(Lifespan ~ Price + Use)
summary(model)
```

```
              Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)   12.0247     3.4202       3.516     0.00265
Price         0.7751      0.2007       3.863     0.00125
Use           4.1768      5.7074       0.732     0.47425
```

Linear regression model: $Lifespan = \beta_0 + \beta_1 Price + \beta_2 Use + U$

Fitted model: $Lifespan = 12.02 + 0.77 \ \cdot Price + 4.17 \cdot Use$

Is the model significant?

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$
- $H_0 : \beta_1 = \cdots = \beta_k = 0; \mathrm{H}_1 : \text{otherwise}$
- If H_0 is true, the model is :
$$Y = \beta_0 + U$$
- We say that the model is “not significant”.
- None of the explanatory variables chosen actually explains the dependent variable.
- The model should be reformulated
- If $p - value$ of regression ANOVA table is less than α, reject H_0, and consider the model significant. Otherwise, the model is not significant.
- In the example of Lifespan, the model is significant, as the p-value is less than 0,05.

```
F-statistic: 8.804 on 2 and 17 DF,  p-value: 0.002376
```

Estimation of parameters

- Punctual or Confidence Interval
- Maybe, some of the parameters (and their corresponding variables), can be deleted from the model.
- This implies that the corresponding independent variable does not explain the dependent variable.
- This is done by performing the hypothesis test: $H_0: \beta_j = 0; H_1: \beta_j \neq 0.$
- If the corresponding p-value (given in the Linear Regression table) is less than α, reject the null hypothesis and accept that the parameter (and its variable) is significant. Otherwise, the parameter could be considered zero (non-significant) and the variable could be removed from the model.
- In the previous example, both β_0, β_1 are significant, but β_2 is not. Should you remove variable *Use* from the model?

Predictions

- One of the objectives of the regression model is to predict values of the dependent variable

Given the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$,

the prediction of Y when $X_i = x_i \; \forall \, i$ is given by

$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$

which coincides with its mean value, that is

$\hat{Y} = E(Y / X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k).$

predict.lm(model,data.frame(Price = 10, Use = 0.5))

The following also gives you the confidence interval of the average dependent value of all observations with the given values of the explanatory vriables.

predict.lm(model,data.frame(Price = 10, Use = 0.5), interval = "confidence")

The following also gives you the confidence interval of one observation with the explanatory variables as given

predict.lm(model,data.frame(Price = 10, Use = 0.5), interval = "prediction")

Is the model good? R^2

The *Total Sum of Squares* (*TSS*) can be divided into two parts: the sum of squares due to the explanatory variables (or the *Explained Sum of Squares (ESS)*) and the *Residual Sum of Squares (RSS)*.

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

$$TSS=ESS+RSS$$

RSS=0→Perfect model
ESS=0→Horrible model

ESS is also known as *Model Sum of Squares* and *Sum of Squares due to Regression*

Coefficient of determination and adjusted coefficient of determination

The **coefficient of determination** measures the goodness of fit of a regression model:

$$R^2=\frac{ESS}{TSS}=1-\frac{RSS}{TSS},\quad 0\le R^2\le 1$$

In words, the coefficient of determination measures the proportion (or percentage) of the total variation in *Y* explained by the regression model. R^2 = 1 implies a model perfectly explained, R^2=0 implies no explanation at all.

The **adjusted coefficient of determination** is used to compare different models with different number of explanatory variables that explain the same dependent variable. It is defined as:

$$\overline{R}^2=1-\frac{RSS/(n-k-1)}{TSS/(n-1)}=1-(1-R^2)\frac{n-1}{n-k-1}$$

It is good practice to use $\bar{R}^2$ rather than R^2, because R^2 tends to give an overly optimistic picture of the fit of regression, particularly when the number of explanatory variables is not very small compared with the number of observations (Theil, 1978).

$R^2\ and\ \bar{R}^2$ in the lifespan example

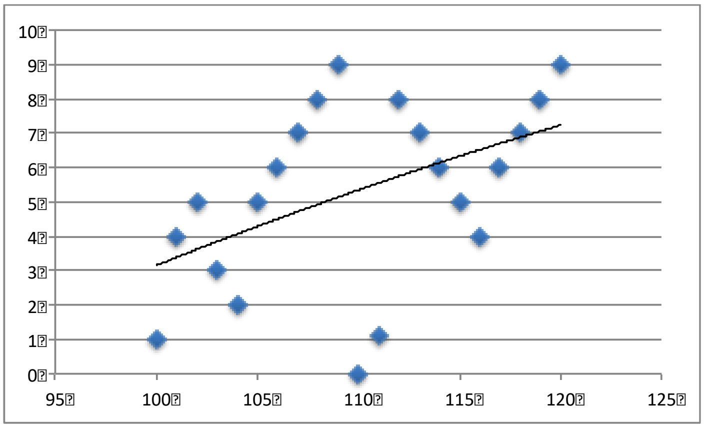
```
summary(model)$r.squared
summary(model)$adj.r.squared
```

```
Multiple R-squared:  0.5088,    Adjusted R-squared:  0.451
```

Other models in simple regression

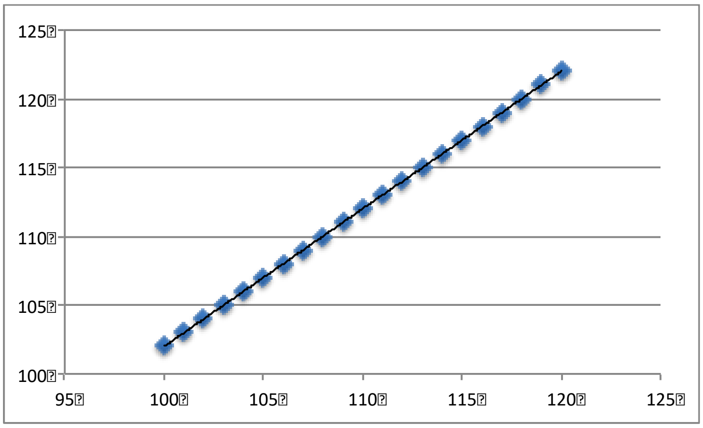
- In simple regression you can use scatter diagrams to guess the relationship between the two variables.
- A scatter (or dispersion) diagram is a graph consisting of a horizontal axis representing the *X* variable, a vertical axis representing the *Y* variable, and a point plotted for each *x*-*y* pair of sample data.

Scatter diagram: no linear relationship



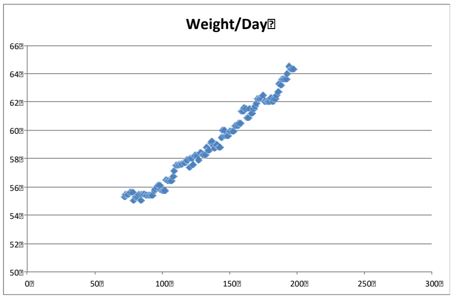
$R^2 \cong 0$

Scatter diagram: functional linear relationship

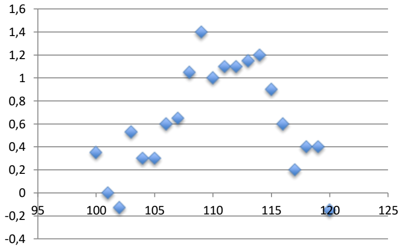


$R^2 \cong 1$

Scatter diagram: statistical relationship



Non-linear relation?



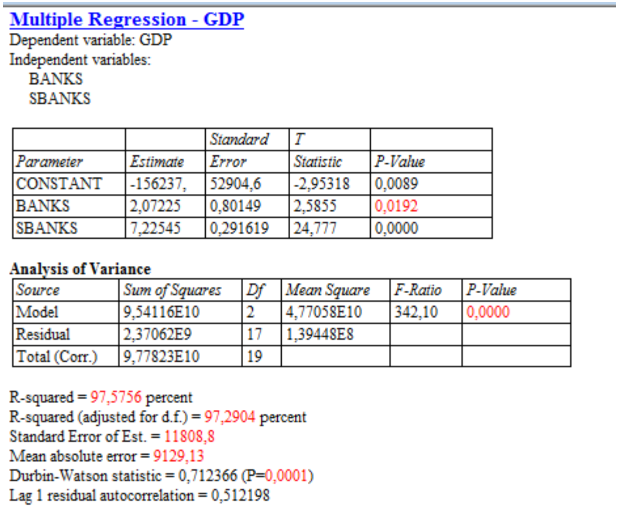
$0'7 \leq R^2 \leq 1$

More questions about regression

- Does a explanatory variable affect the dependent variable differently depending on the values of other explanatory variables? Interaction!
- What to do if (some of) the assumptions of the linear model are not satisfied: multicollinearity, heteroscedasticity, autocorrelation, non-normality of residuals,…
- Confidence intervals for predictions?
- You may transform many non-linear models by doing an appropriate variable change. Example: $Y = \alpha e^{\beta X}$; $Ln(Y) = Ln(\alpha) + \beta X$

Exercises

- The gross domestic product of a country (GDP) wants to be explained as a function of the number of bank offices (BANKS) and number of saving bank offices (SBANKS).
- $GDP = \beta_0 + \beta_1 BANKS + \beta_2 SBANKS + U$
- Are the explanatory variables actually explaining the dependent variable? What would you do to improve the model? What is the percentage of the total variability in GDP explained by this model? Is this a good percentage?
- From this model, and the results in the next slide, predict the GDP for a country in which the number of banks is 70000 and the number of saving banks is 45000.
- Take α=0,05 .



Multiple Regression - GDP

Dependent variable: GDP
Independent variables:
BANKS
SBANKS

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-156237,	52904,6	-2,95318	0,0089
BANKS	2,07225	0,80149	2,5855	0,0192
SBANKS	7,22545	0,291619	24,777	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	9,54116E10	2	4,77058E10	342,10	0,0000
Residual	2,37062E9	17	1,39448E8		
Total (Corr.)	9,77823E10	19			

R-squared = 97,5756 percent
R-squared (adjusted for d.f.) = 97,2904 percent
Standard Error of Est. = 11808,8
Mean absolute error = 9129,13
Durbin-Watson statistic = 0,712366 (P=0,0001)
Lag 1 residual autocorrelation = 0,512198

Example

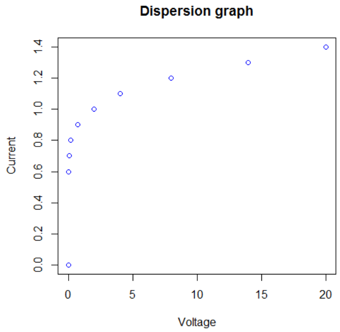
- Consider a simple diode. Its voltage wants to be explained from its current using the following simple linear regression model:

$$Voltage = \beta_0 + \beta_1 Current + U$$

In order to check the validity of this model, the following 10 observations are available:

Voltage	0	0,6	0,7	0,8	0,9	1	1,1	1,2	1,3	1,4
Current	0	0,01	0,05	0,2	0,7	2	4	8	14	20

A scatter plot and the results of the simple linear regression model for these data are:



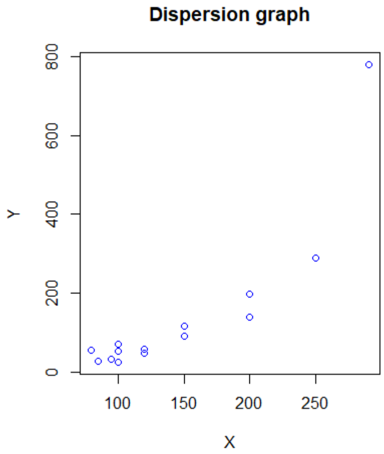
Model	R^2
$Voltage = \beta_0 + \beta_1 Current + U$	0.5441
$Voltage = \beta_0 + \beta_1 Log(Current)$+U	0.8901
$Log(Voltage) = \beta_0 + \beta_1 Log(Current)$+U	0.471

Based on R^2 the best model is the logarithmic model (highest value)

Other models: example

- Decide which of the following models best explains variable Y from X. Estimate its parameters.

X	Y
95	32
120	58
100	25
150	90
200	140
100	51
150	115
200	198
80	54
250	290
100	71
85	28
120	47
290	780



Type	Model	Fitted	R^2
Exponential	$Y = \alpha e^{\beta X}$	$10,815e^{0,014x}$	R^2 = 0,9011
Linear	$Y = \alpha + \beta X$	y = 2,6118x - 239,22	R^2 = 0,7442
Logarithmic	$Y = \alpha Ln(X) + \beta$	y = 378,44ln(x) - 1712,4	R^2 = 0,6172
Quadratic	$Y = \beta_0 + \beta_1 X + \beta_2 X^2$	$y = 0,0231x^2 - 5,5726x + 370,58$	R^2 = 0,9322
Order 3	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$	$y = 0,0002x^3 - 0,0849x^2 + 12,505x - 542,99$	R^2 = 0,9733
…			
Power	$Y = \alpha X^{\beta}$	$y = 0,0019x^{2,1775}$	R^2 = 0,8598