MPC Course - Part 2

Regression : predicting a variable using other variable(s)

Simon Malinowski

M1 Miage, Univ. Rennes 1

What is regression ?

From data to predictions

Regression example

House	House size (sq. feet)	Year built	House Price (target, in k€)
1	80	1985	120
2	92	2010	180
3	75	2008	145
4	110	2015	220
5	85	2000	140
6	150	1975	225
7	55	1992	100
8	105	1999	? ?
9	95	2018	? ?

Other regression examples

- How many people will retweet your tweet ? (y)
  $\rightarrow$ depends on x : number of followers, popularity, subject of the tweet, ...
- What will be your next salary ? (y)
  $\rightarrow$ depends on x : your degree, your experience, ...
- How many points will have a football team at the end of the season ?
  $\rightarrow$ depends on x : statistics about the team performance (goals, shoots, possession,...)

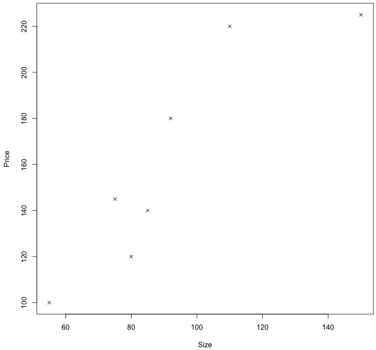
Different kinds of regression

Regression $\rightarrow$ target variable is numerical

- Simple VS multiple
  - $\rightarrow$ simple regression : one feature variable (X) to predict Y
  - $\rightarrow$ multiple regression : several feature variables $(X_1, \ldots, X_p)$ to predict Y
- Linear VS Non-linear

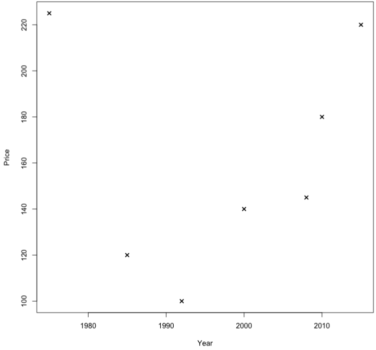
Simple linear regression in a nutshell

Size	Price
80	120
92	180
75	145
110	220
85	140
150	225
55	100
105	? ?
95	? ?



Simple linear regression in a nutshell

Year	Price
1985	120
2010	180
2008	145
2015	220
2000	140
1975	225
1992	100
1999	? ?
2018	? ?



What we'll see about simple regression

- How to find the model ? (i.e. coefficients of the regression line)
- How to evaluate the performance of a model ?
  - $\rightarrow$ will the model be good to predict new inputs ?
  - $\rightarrow$ if I have more than one variable, which one seems the best ?

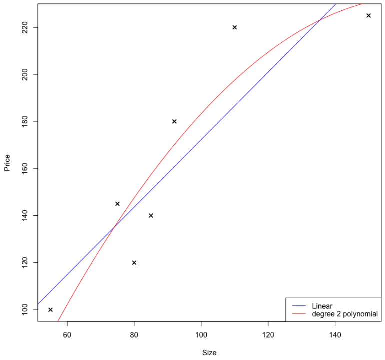
Multiple linear regression in a nutshell

Size	Year	Price
80	1985	120
92	2010	180
75	2008	145
110	2015	220
85	2000	140
150	1975	225
55	1992	100
105	1999	? ?
95	2018	? ?

$$P = -2872 + 1.65 \times S + 1.44 \times Y$$

Multiple linear regression for non-linear regression

Size	$Size^2$	Price
80	6400	120
92	8464	180
75	5625	145
110	12100	220
85	7225	140
150	22500	225
55	3025	100
105	11025	? ?
95	9025	? ?



$$P = -91.3 + 3.94 \times S - 0.012 \times S^2$$

What we'll see about multiple regression

- How to find the model ? (i.e. coefficients of the model)
- How to evaluate the performance of a model ?
  - $\rightarrow$ will the model be good to predict new inputs ?
  - $\rightarrow$ how to compare models with different number of variables
- Non-linear regression with multiple regression
- Variable selection
  - $\rightarrow$ do I really need all the variables I have, or a subset might be better ?
  - $\rightarrow$ how to perform variable selection ? (criteria, methods)

Regression with non-numerical variables

Size	District	Price
80	Center	120
92	Bourg-Lesveque	180
75	Center	145
110	Longchamps	220
85	Longchamps	140
150	Bourg-Lesveque	225
55	Center	100
105	Longchamps	? ?
95	Center	? ?

Data, problem

Simple regression : 1 numerical target variable, 1 numerical explanatory variable

Individuals	X	Y
1	x_1	y_1
	$\vdots$	
i	x_i	y_i
	$\vdots$	
n	x_n	y_n

Multiple regression : 1 numerical target variable, several numerical explanatory variables

Individuals	X_1	$\ldots$	X_j	$\ldots$	X_p	Y
1	x_{11}	$\ldots$	x_{1j}	$\ldots$	x_{1p}	y_1
	$\vdots$		$\vdots$		$\vdots$	
i	x_{i1}	$\ldots$	x_{ij}	$\ldots$	x_{ip}	y_i
	$\vdots$		$\vdots$		$\vdots$	
n	x_{n1}	$\ldots$	x_{nj}	$\ldots$	x_{np}	y_n

Quantity of information in the target variable

$$I_t = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Linear model

We assume that Y is a linear function of the explanatory variables X_i

$$Y \simeq \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Estimation of the coefficients

$\beta_0, \ldots, \beta_p$ are the "optimal" parameters (that we would find if we knew all the data about the problem)

Aim : find $\hat{\beta}_0, \ldots, \hat{\beta}_p$, estimations of $\beta_0, \ldots \beta_p$ that are as accurate as possible

$\rightarrow$ Least-square regression

Least square regression (simple regression case)

$\rightarrow$ Objective : minimizing the sum of squared residuals

$$I_r(\beta_0, \beta_1) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$I_r(\beta_0, \beta_1) = \sum_{i=1}^{n} ((\beta_0 + \beta_1 x_i) - y_i)^2$$

Partial derivatives needs to be equal to 0 :

$$\begin{cases} \beta_1 & = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \\ \beta_0 & = \overline{y} - \beta_1 \overline{x} \end{cases}$$

We can also compute $\hat{\sigma}_{\beta_1}$: uncertainty about the value of β_1

Coefficient of determination

We have seen that the quantity of information in the target variable is :

$$I_t = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

This quantity of information can be split into two parts :

- the quantity of information explained by the model I_m
- the quantity of information not explained by the model I_r

We have $I_t = I_m + I_r$

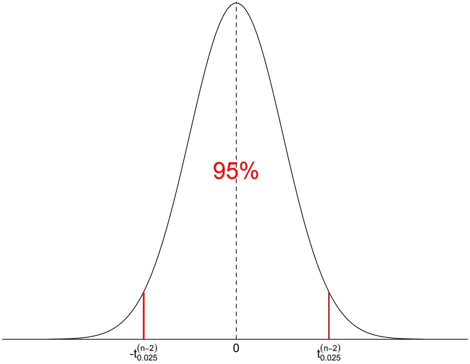
The R-squared is defined as :

$$R^2 = \frac{I_m}{I_t}$$

Simple regression : Test for the significancy of β_1

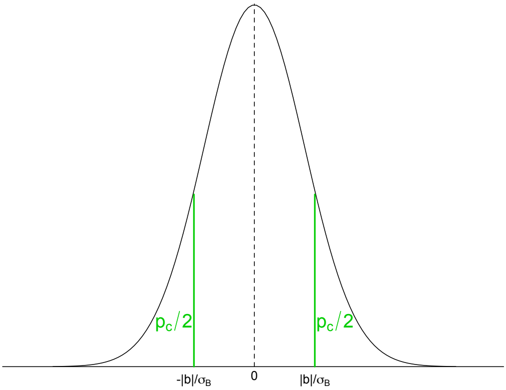
Hypothesis 0 (H_0) : X does not influence Y $\rightarrow \beta_1 = 0$
Hypothesis 1 (H_1) : X has an influence on Y $\rightarrow \beta_1 \neq 0$

Under H_0, $\frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}$ is generated by a Student's law with $n-2$ degrees of freedom



Simple regression : Test for the significancy of β_1

Under H_0, $\frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}$ is generated by a Student's law with $n-2$ degrees of freedom



p_c : probabilty of making a wrong decision when deciding H_1

$p_c < 0.05 \Rightarrow$ decide H_1

Multiple reg. : Test of significancy of one coefficient

Hypothesis $H_0 : \beta_j = 0$ (x_j is not significant **in the presence of other variables**)

Hypothesis $H_1 : \beta_j \neq 0$ (x_j is significant **in the presence of other variables**)

Test value : $T = \frac{\hat{\beta}_j}{\hat{\sigma}_j}$

Under H_0, T is generated by a Student's law with $(n-p-1)$ degrees of freedom.

Si $|T| > t_{0.025}^{n-p-1}$, we refuse H_0.

Examples of the interest of this test

```
Call:
lm(formula = y ~ x1 + x10, data = ozo)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.940 -11.255   0.369   9.500  55.273 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -28.96058   10.15508  -2.852  0.00521 ** 
x1            4.47827    0.66801   6.704 9.58e-10 ***
x10           0.40446    0.07219   5.602 1.63e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.44 on 108 degrees of freedom
Multiple R-squared:  0.6275,	Adjusted R-squared:  0.6206 
F-statistic: 90.98 on 2 and 108 DF,  p-value: < 2.2e-16
```

Examples of the interest of this test

```
Call:
lm(formula = y ~ x1 + x2, data = ozo)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.822 -11.896   0.239  11.338  46.672 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -31.1244    10.2426  -3.039  0.00298 ** 
x1            0.8023     1.2252   0.655  0.51398    
x2            4.9473     0.9254   5.346 5.06e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.61 on 108 degrees of freedom
Multiple R-squared:  0.6199,	Adjusted R-squared:  0.6128 
F-statistic: 88.06 on 2 and 108 DF,  p-value: < 2.2e-16
```

Examples of the interest of this test

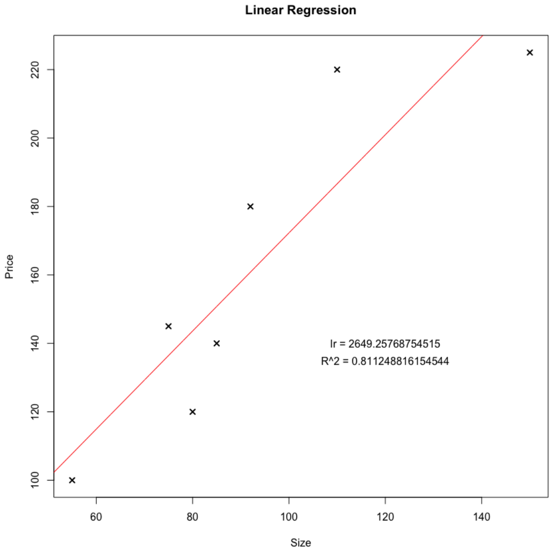
```
Call:
lm(formula = y ~ x1 + x10 + x2, data = ozo)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.982  -8.149   0.438   9.916  38.693 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -28.11029    9.04994  -3.106  0.00243 ** 
x1           -0.55031    1.10696  -0.497  0.62011    
x10           0.36521    0.06474   5.641 1.39e-07 ***
x2            4.42587    0.82145   5.388 4.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.54 on 107 degrees of freedom
Multiple R-squared:  0.707, Adjusted R-squared:  0.6988 
F-statistic: 86.07 on 3 and 107 DF,  p-value: < 2.2e-16
```

Is a given model good to make predictions ?

Are the determination coefficient or the I_r (or the statistical tests) adapted to tell me if a given model is good to make predictions ?

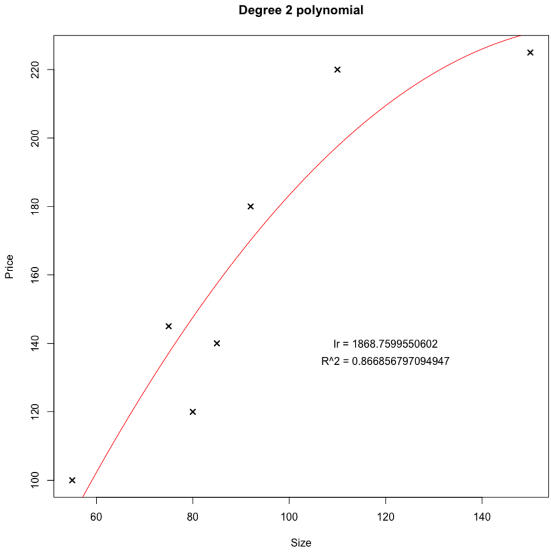
l_r and R^2 are not good indicators

Size	Price
80	120
92	180
75	145
110	220
85	140
150	225
55	100



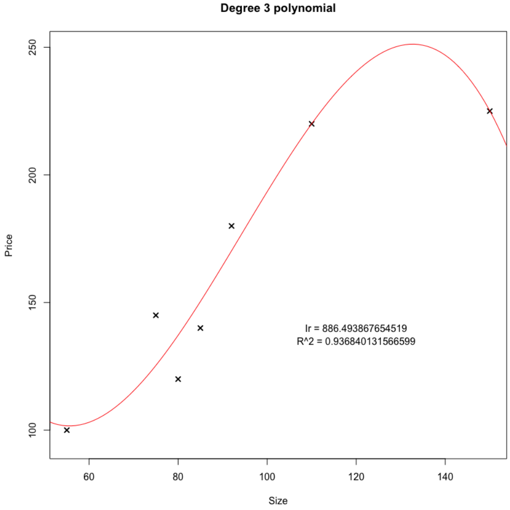
I_r and R^2 are not good indicators

Size	$Size^2$	Price
80	6400	120
92	8464	180
75	5625	145
110	12100	220
85	7225	140
150	22500	225
55	3025	100



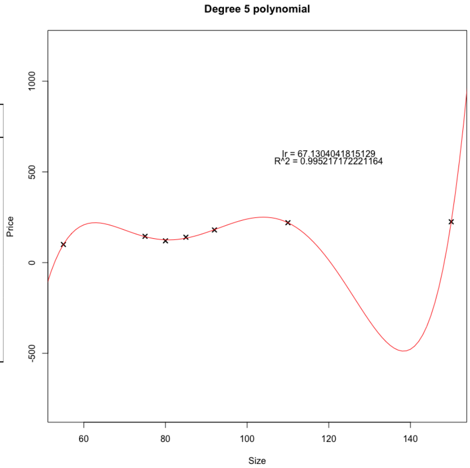
I_r and R^2 are not good indicators

Size	$Size^2$	$Size^3$	Price
80	6400		120
92	8464		180
75	5625		145
110	12100		220
85	7225		140
150	22500		225
55	3025		100



I_r and R^2 are not good indicators

Size	$Size^2$	$Size^3$	$Size^4$	$Size^5$
80	6400			
92	8464			
75	5625			
110	12100			
85	7225			
150	22500			
55	3025			



Generalisation error of a model

Definition : it is the average error that a model would do when predicting any new individuals.

Problem : It is impossible to know exactly as we don't have any new individuals...

$\rightarrow$ we will try to have a fair estimation of this value.

2 main methods :

1. Train/Test split
2. K-fold cross validation

Train/Test split

Individuals	X_1	$\ldots$	X_j	$\ldots$	X_p	Y
1	x_{11}	$\ldots$	x_{1j}	$\ldots$	x_{1p}	y_1
	$\vdots$		$\vdots$		$\vdots$	
i	x_{i1}	$\ldots$	x_{ij}	$\ldots$	x_{ip}	y_i
i +1	$x_{i+1,1}$	$\ldots$	$x_{i+1,j}$	$\ldots$	$x_{i+1,p}$	y_{i+1}
	$\vdots$		$\vdots$		$\vdots$	
n	x_{n1}	$\ldots$	x_{nj}	$\ldots$	x_{np}	y_n

Individuals in blue are used to **learn** a model (or different models)

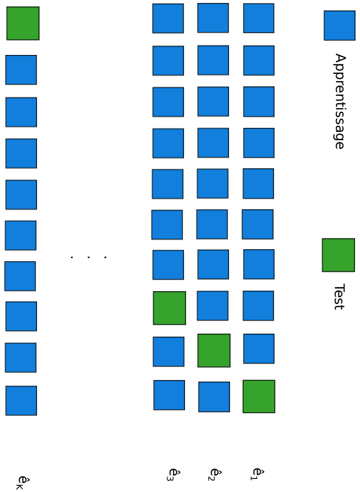
The model is used to predict y for all the individuals in green

The generalization error is estimated by a Mean Square Error of these predictions.

Drawback ?

K-fold cross validation

Principle : the dataset is split into K disjoint subsets



The estimation of the generalization error is the average of the K measure errors.