Linear and Logistic Regression

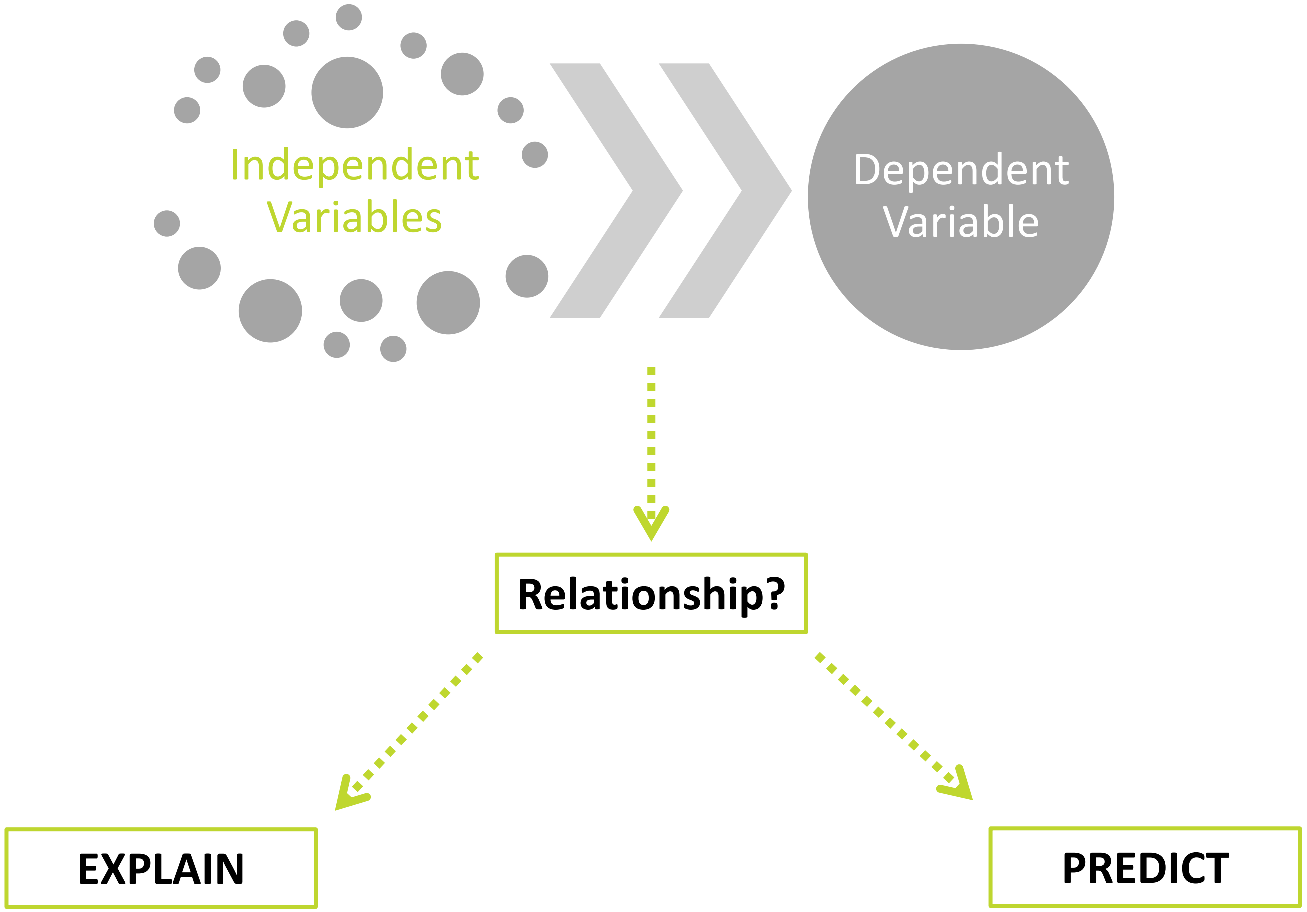
Master in Data Science and Advanced Analytics
BA and DS

Roberto Henriques

NOVA IMS INFORMATION MANAGEMENT SCHOOL

Linear Regression

Regression Analysis



Independent Variables

Dependent Variable

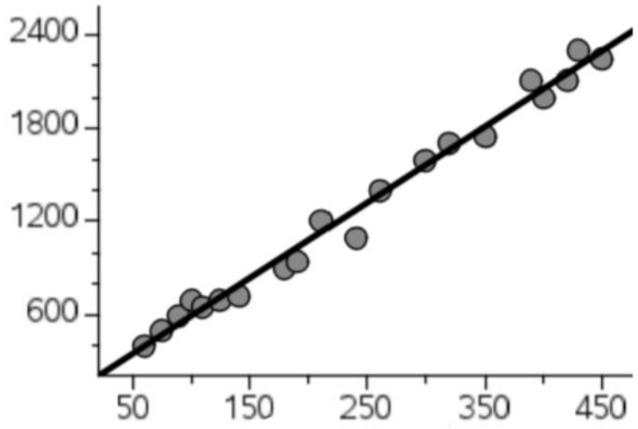
Relationship?

EXPLAIN

PREDICT

Linear Regression

- Use least-squares to fit a line to our data
- Calculate R^2
- Calculate a *p*-value for R^2

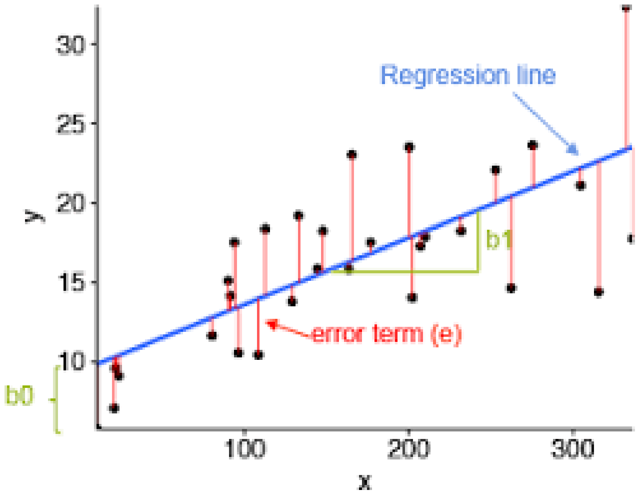


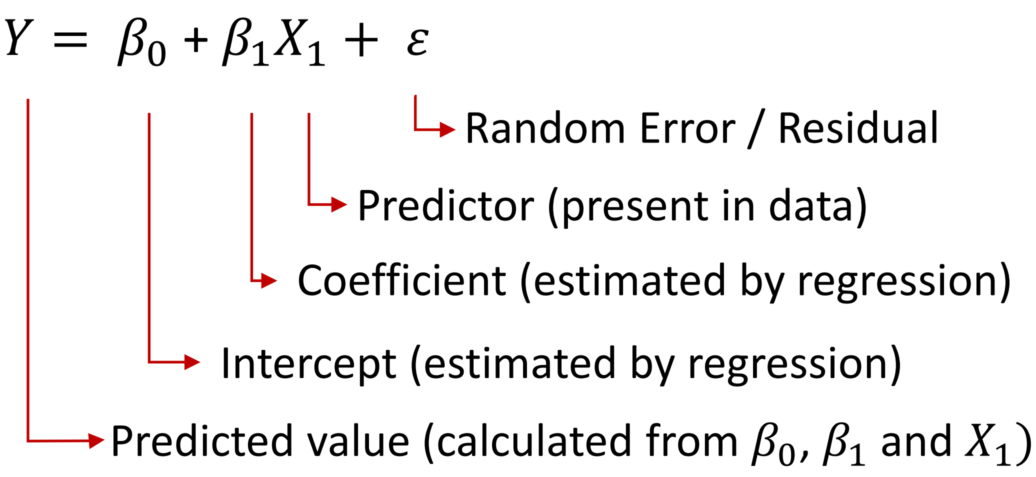
- Exampes:
  - Predict sales amount
  - Predict the growth of the economy
  - Predict the price of a house

Simple and Multiple Linear Regression

Simple Linear Regression

A linear regression model with a single explanatory variable





$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Random Error / Residual

Predictor (present in data)

Coefficient (estimated by regression)

Intercept (estimated by regression)

Predicted value (calculated from β_0, β_1 and X_1)

Multiple Linear Regression

A linear regression model with two or more explanatory variables

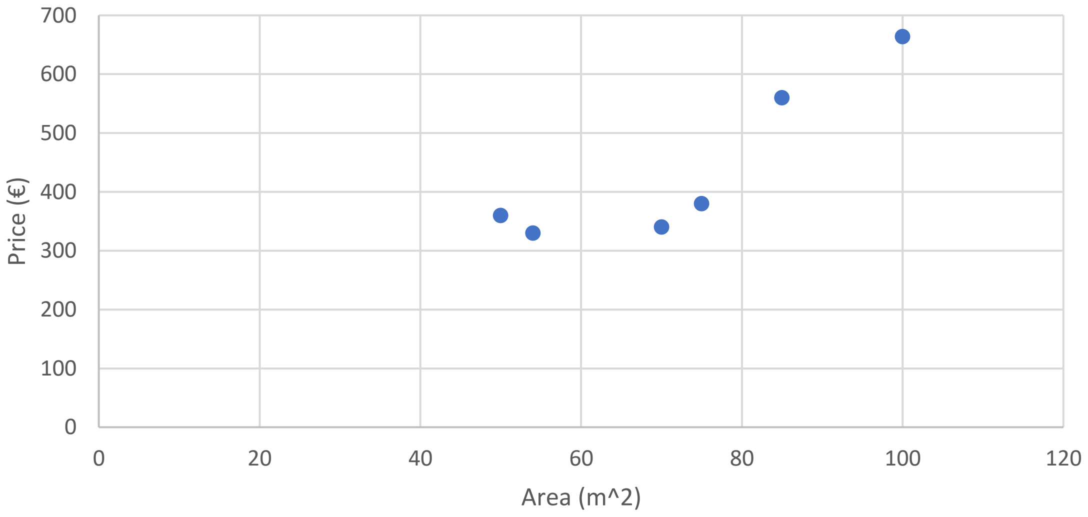
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/

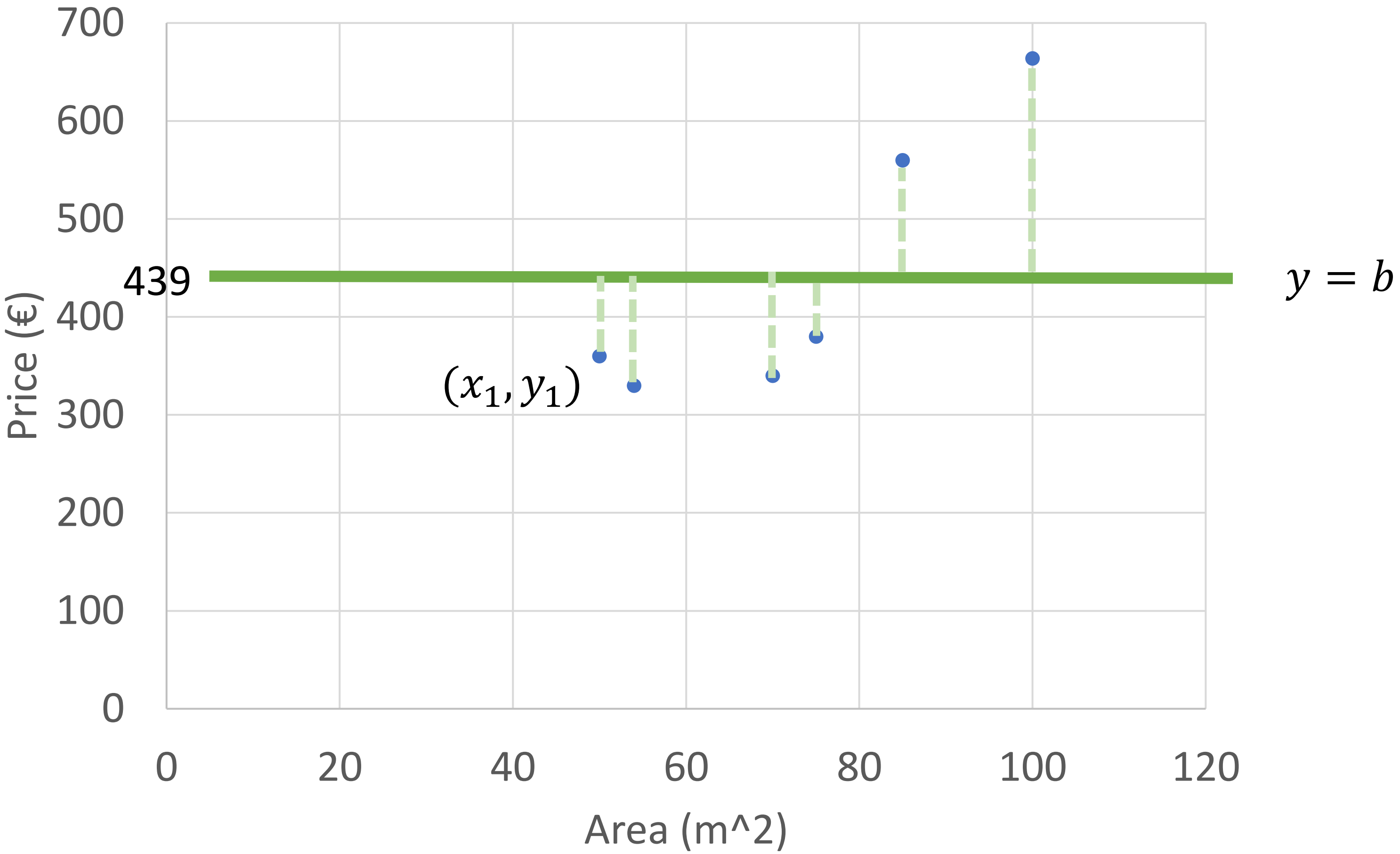
Example

- Let us examine the linear dependency of the house prices based on their size (square meters). With this sample of 6 houses, find the equation of the straight line that best fits the data.

ID	Area (m^2)	Price (€)
1	50	360
2	70	340
3	100	664
4	54	330
5	85	560
6	75	380

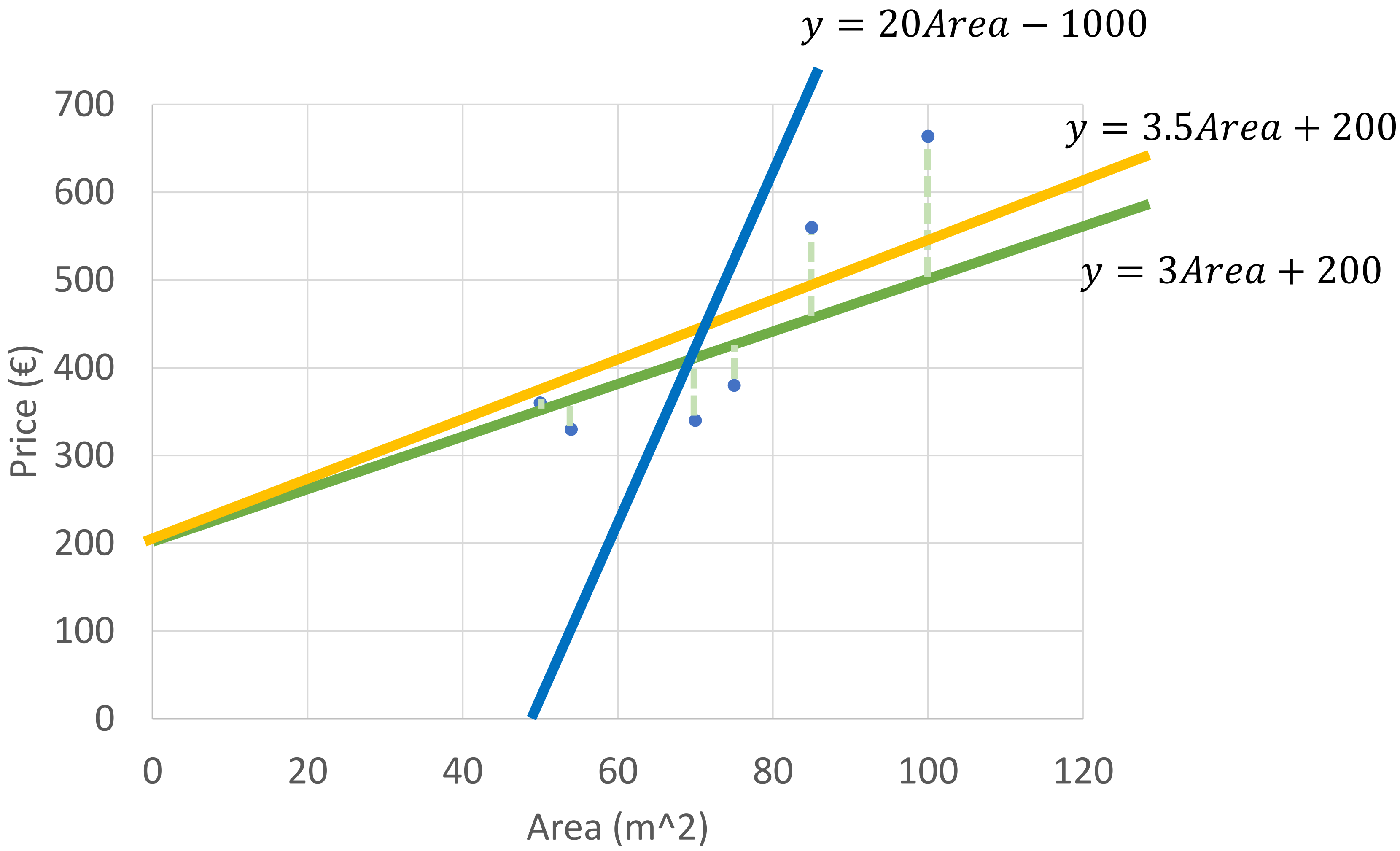


Example



$$sum\ of\ Residuals^2 = (b - y_1)^2 + (b - y_2)^2 + (b - y_3)^2 + (b - y_4)^2 + (b - y_5)^2 + (b - y_6)^2$$
$$sum\ of\ Residuals^2 = 96670$$

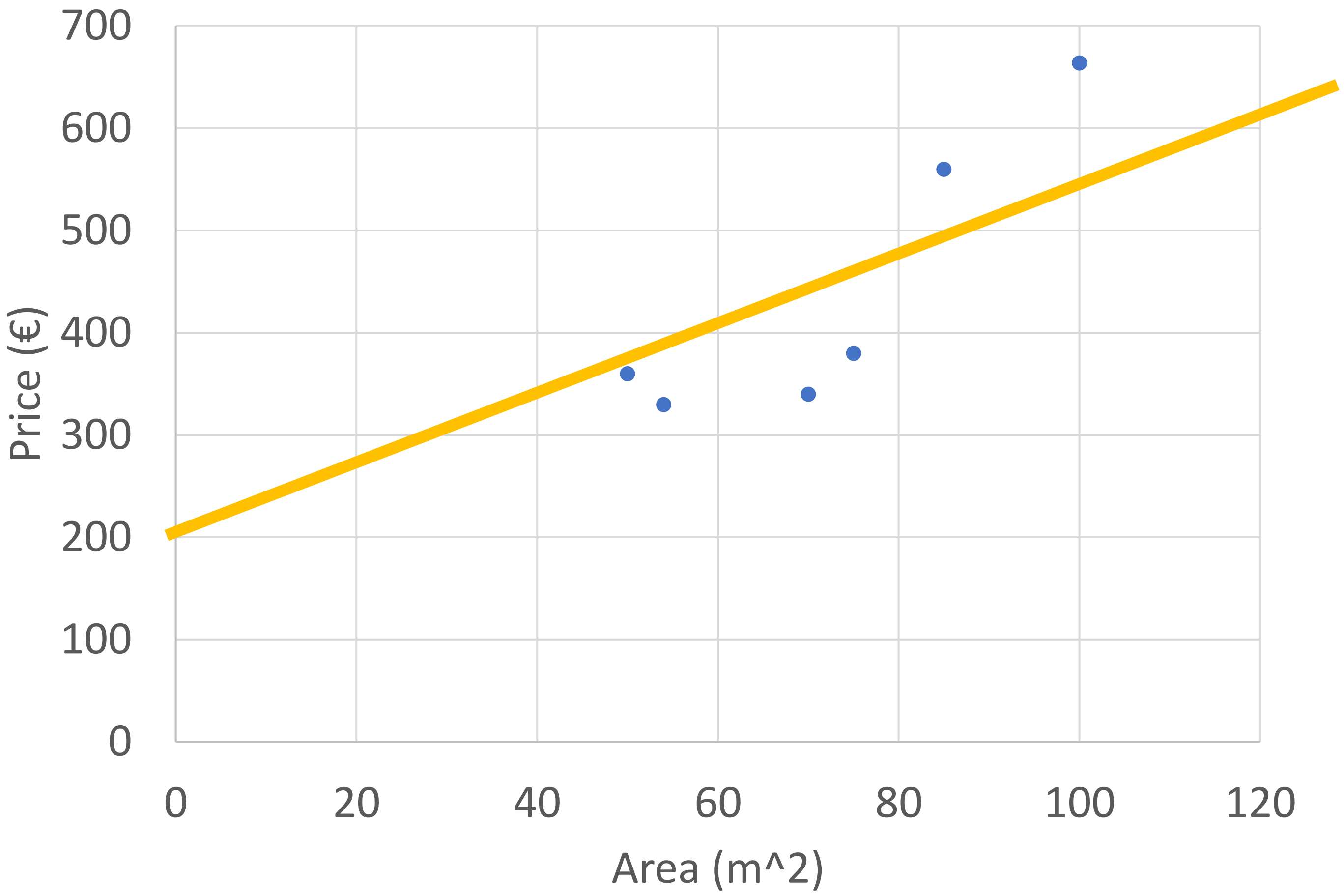
Example



$sum\ of\ Residuals^2 = $ **45970** $sum\ of\ Residuals = $ **342596**

$sum\ of\ Residuals^2 = $ **38439.5**

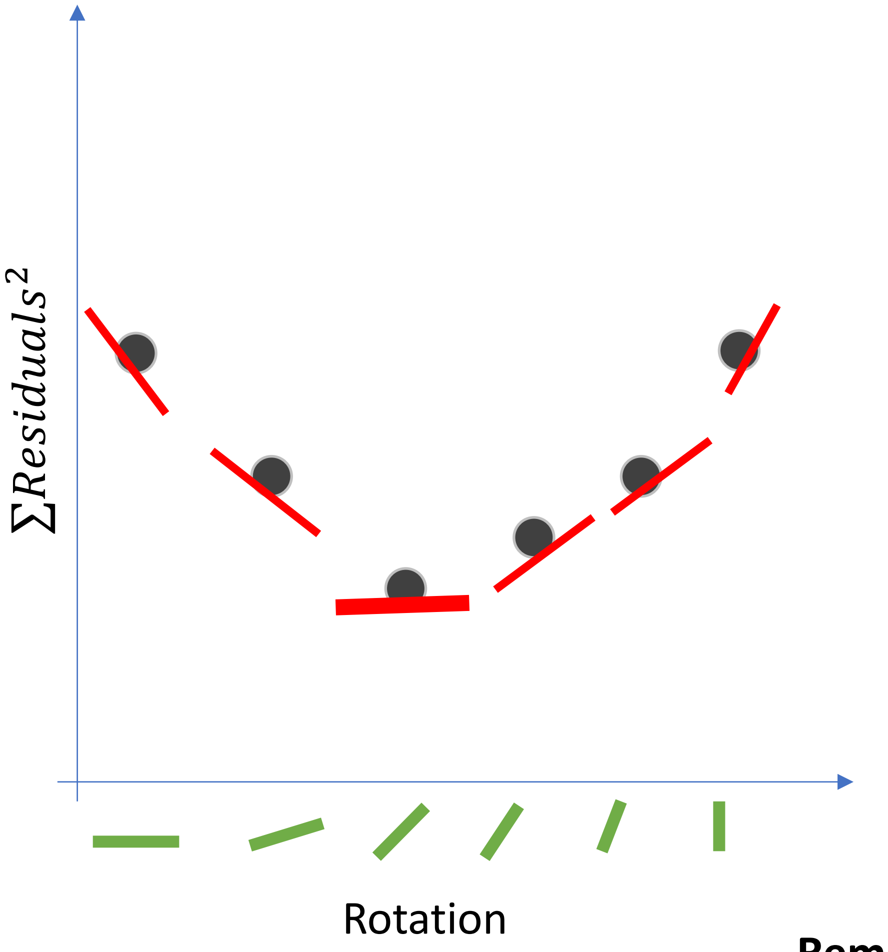
Example



$y = aX + b$

$$\sum Residuals^2 = (\boxed{(a \times x_1 + b)} - \boxed{y_1})^2 + ((a \times x_2 + b) - y_2)^2 + ((a \times x_3 + b) - y_3)^2 + \cdots$$

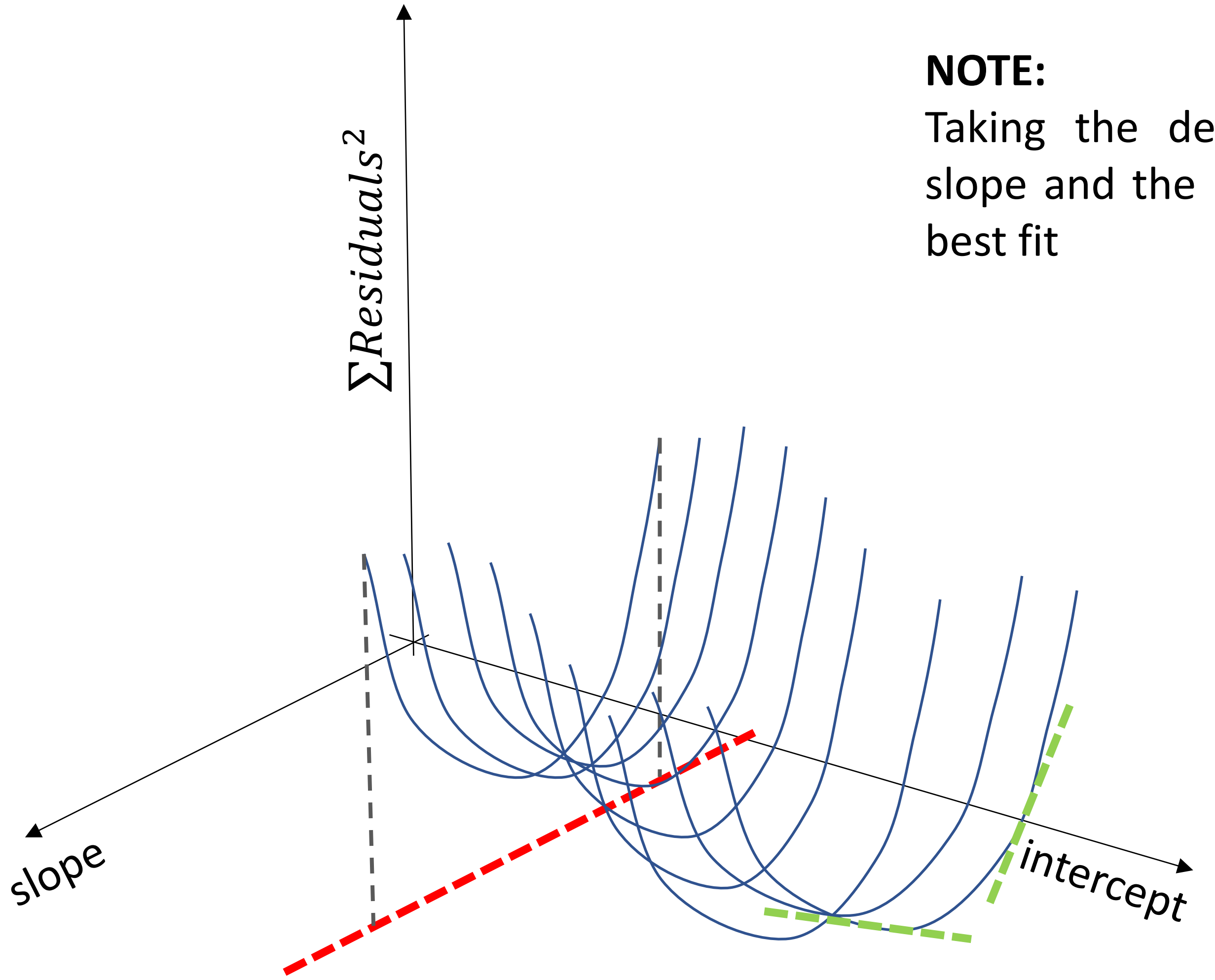
Example



Remember
Different rotations are different values for “*a*” and “*b*”

How do we find the optimal rotation for the line?

Example



NOTE:
Taking the derivatives of both the slope and the intercepts gives us the best fit

How?

$y = \beta_0 + \beta_1 X + \ \epsilon$

$\hat{\beta}_0$ and $\hat{\beta}_1$ estimate β_0 and $\beta_1$from data

$$SumSquaredResid = \sum\left(Y_i - \hat{Y}_i\right)^2 = \sum\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2$$

- We need to minimize a certain function. So:
  - We take the partial derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$
  - We set those partial derivatives equal to zero
  - We solve the equations for $\hat{\beta}_0$ and $\hat{\beta}_1$

How?

Taking the partial derivative with respect to $\hat{\beta}_0$

$$\frac{\partial}{\partial\hat{\beta}_0}\sum\left(Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i)\right)^2=\cdots=-2\sum\left(Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i)\right)$$

Taking the partial derivative with respect to $\hat{\beta}_1$

$$\frac{\partial}{\partial\hat{\beta}_1}\sum\left(Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i)\right)^2=\cdots=-2\sum X_i\left(Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i)\right)$$

How?

Then we set the partial derivative equal to zero

$-2\sum\left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right) = 0$ and $-2\sum X_i\left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right) = 0$

Solving the first equation in order to $\hat{\beta}_0$ we get:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Substituting $\hat{\beta}_0$ by $\bar{Y} - \hat{\beta}_1 \bar{X}$ on the second equation we get:

$$\sum X_i\left(Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i)\right) = 0$$

…

$$\hat{\beta}_1 = \frac{\sum(X_i(Y_i - \overline{Y)}}{\sum(X_i(X_i - \bar{X})}$$

For more details check https://www.youtube.com/watch?v=ewnc1cXJmGA

How?

$\hat{\beta}_1 = \frac{\sum(X_i(Y_i - \overline{Y})}{\sum(X_i(X_i - \bar{X})}$ which is equivalent to $\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \overline{Y})}{\sum(X_i - \bar{X})^2}$

- So, finally we get

$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \overline{Y})}{\sum(X_i - \bar{X})^2}$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

Getting our example

ID	Area	Price (€)	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
1	50	360	-22.33	498.78	-79.00	6241.00	1764.333
2	70	340	-2.33	5.44	-99.00	9801.00	231
3	100	664	27.67	765.44	225.00	50625.00	6225
4	54	330	-18.33	336.11	-109.00	11881.00	1998.333
5	85	560	12.67	160.44	121.00	14641.00	1532.667
6	75	380	2.67	7.11	-59.00	3481.00	-157.333
Sum	434	2634		1773.3		96670	11594
Average	72.33	439.00					

Example

Calculate the coefficient / slope

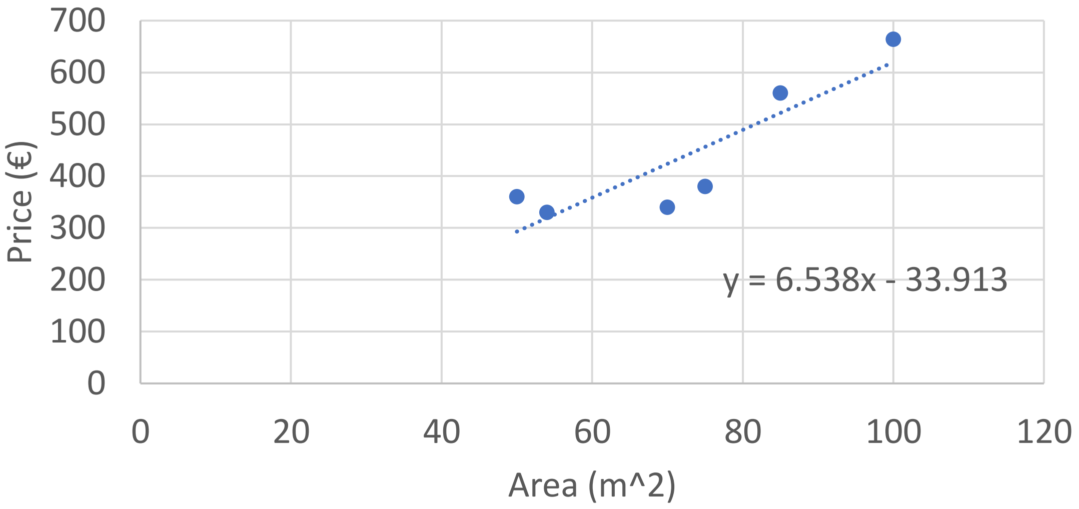
$$\beta_1 = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sum(x_i - \bar{x})^{\,2}} = \frac{11594}{1773.3} = 6.538$$

Find the Intercept

$$\beta_0 = \bar{y} - \beta_1\,\bar{x} = 439 - 6.538 \times 72.33 = -33.912$$

Regression equation

$$y = 6.538x - 33.913$$



Example

The regression equation is: $y = 6.538x - 33.913$

Interpretation of the results:

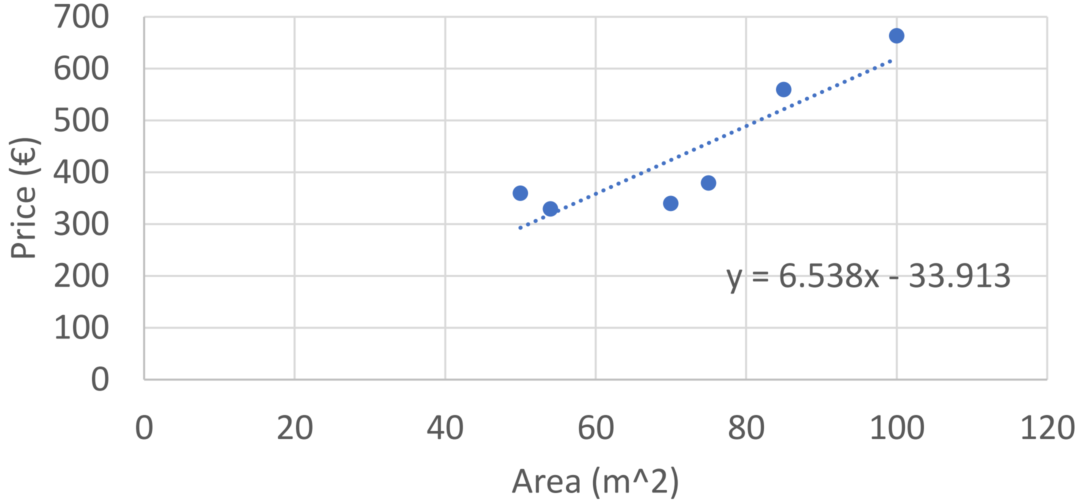
- The slope of 6.538 means that with an increase of one unit in X, we predict Y to increase by an estimated 6.538 units.
- The equation estimates that for each increase of 1 squared meter in the size of the house, the expected price is predicted to increase by 6.538 €

If want to predict the price of a house with 90 $\boldsymbol{m^2}$. Then:
$\hat{y} = 6.538 \times 90 - 33.913$ = 554.50 €

Check how far away is my prediction for house with id 6:
$\hat{y} = 75 - 33.913 =$ 456.43 €
y = 380 €

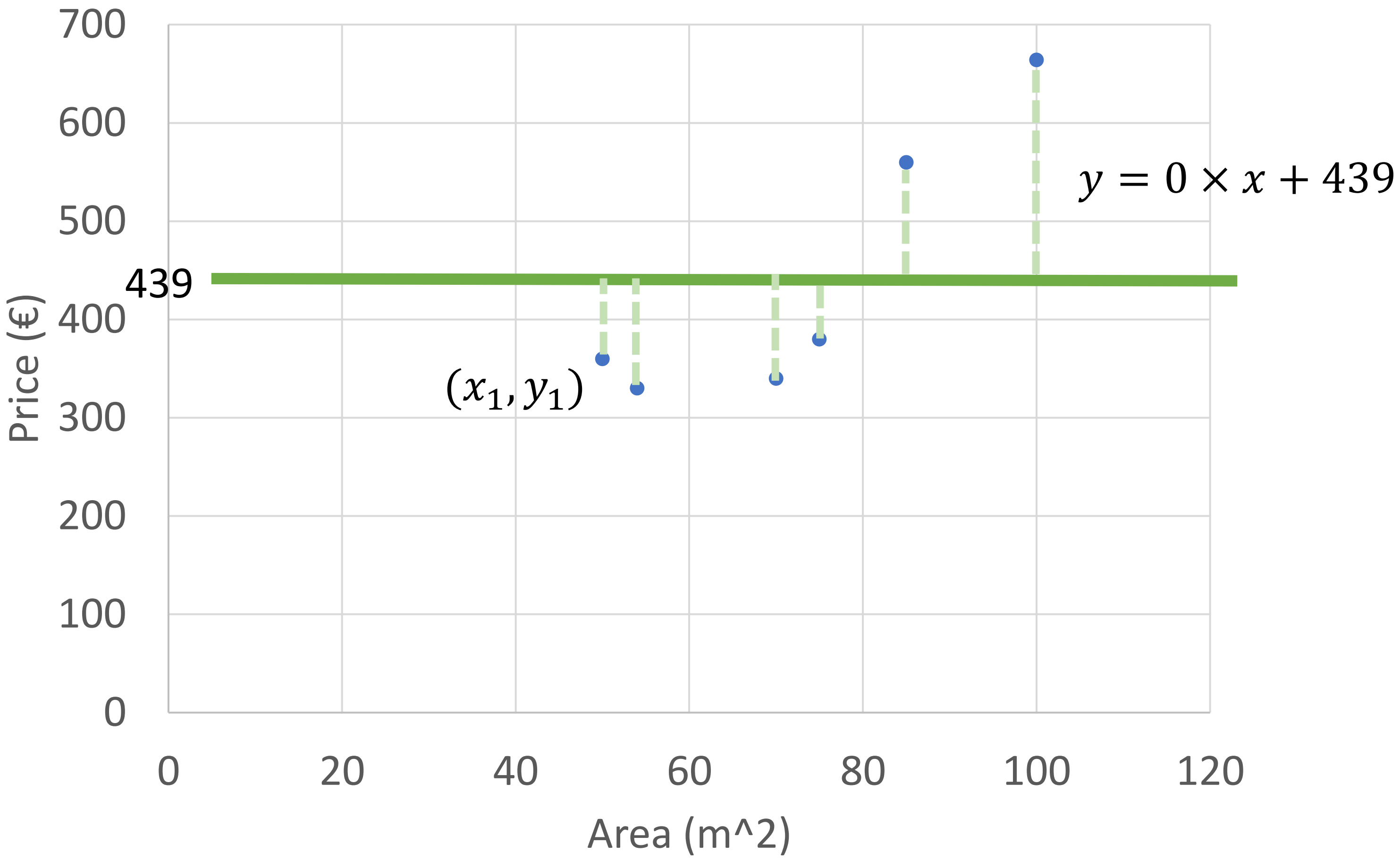
We fitted a line

- Since the slope is not zero, it means the knowing the area of a house will help us making a guess about its price



- But how good is that guess?
  - Calculating R^2 will help us…

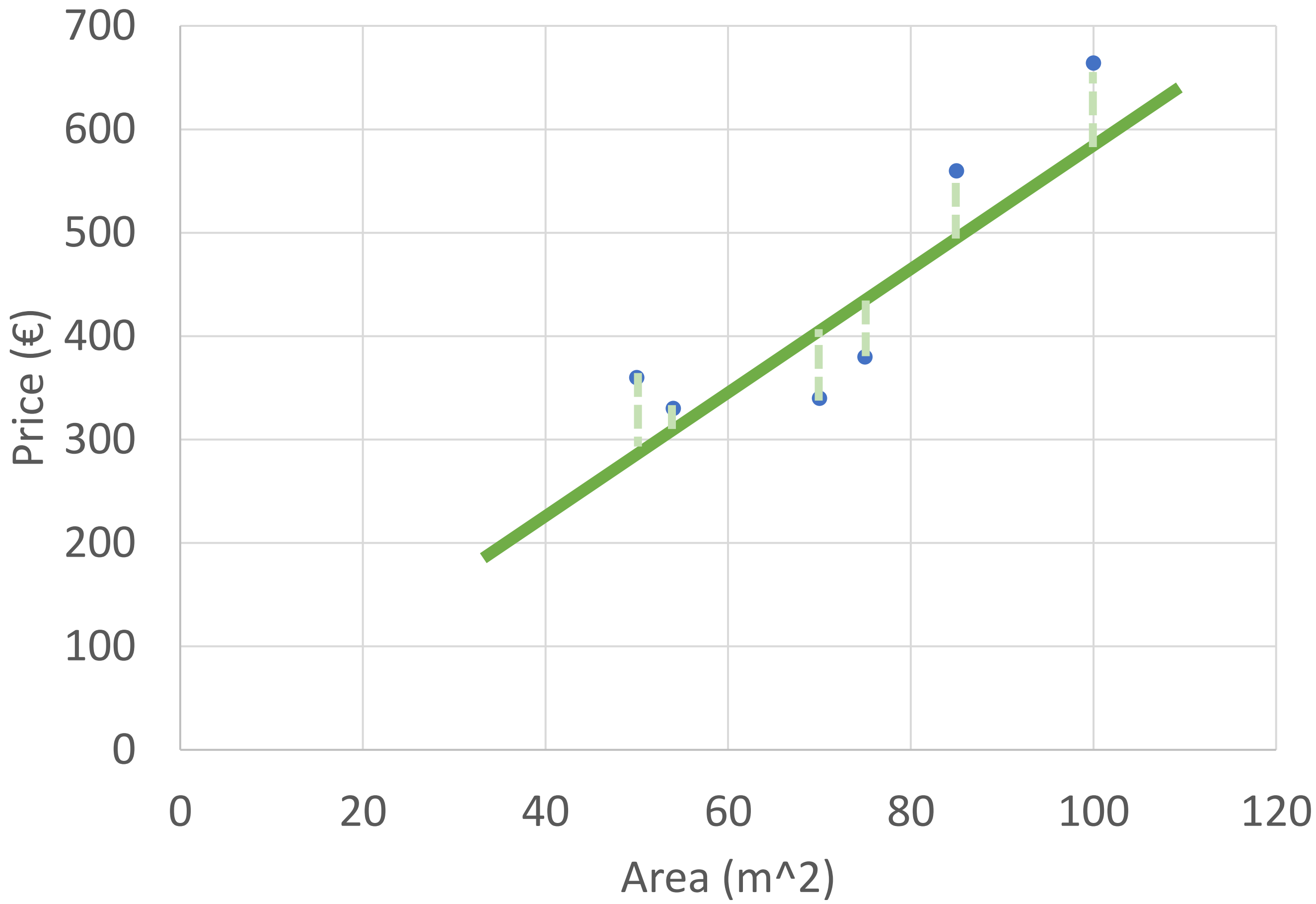
R^2



Sum of squares around the mean (or SST) or $SS(mean) = \sum(y_i \; - \bar{y})^2$

And the variation around the mean or $\text{Var}(mean) = \frac{\sum(y_i - \bar{y})^2}{n} = \frac{SS(mean)}{n}$

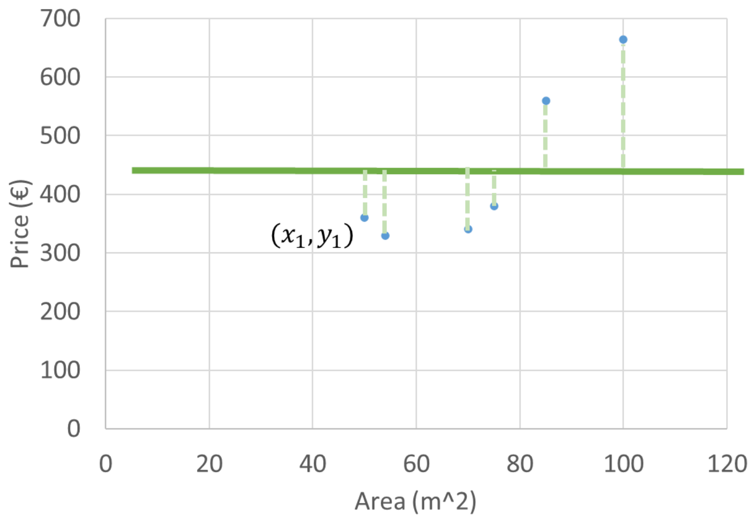
R^2

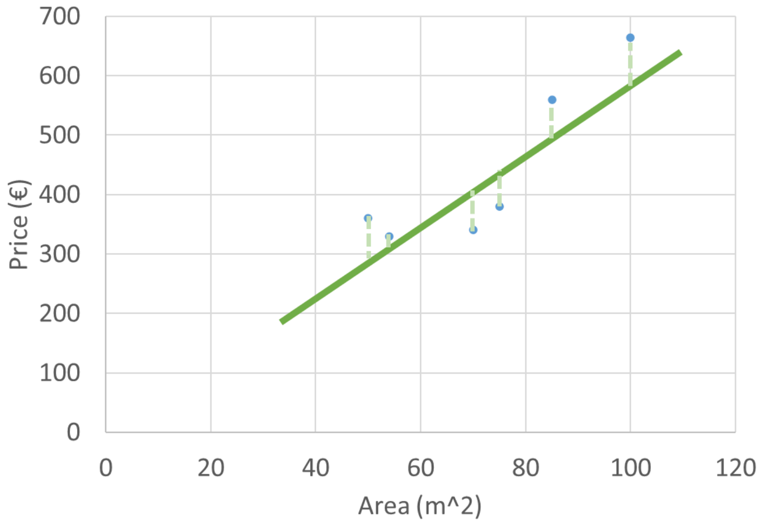


Sum of squares around the least-squares fit or $SS(fit) = \sum(y_i \ - \hat{y}_i\ )^2$

And the variation around the LS or $\mathrm{Var}(fit) = \frac{\sum(y_i - \hat{y}_i\ )^2}{n} = \frac{SS(fit)}{n}$

R^2



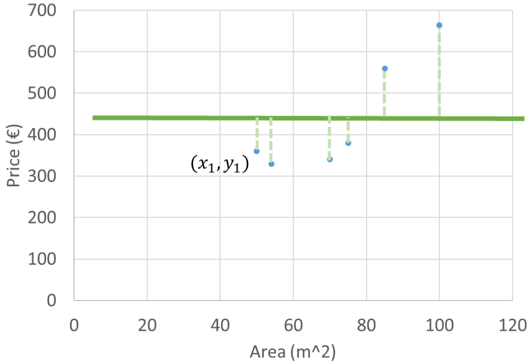


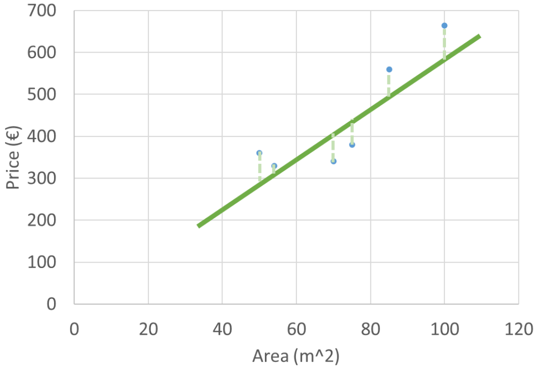
- There is less variation around the LS line compared to the raw variation of prices
- So, we can say that some of the variation in the prices is “explained” by taking house size into account
  - Bigger houses are more expensive and smaller houses are cheaper
- $\boldsymbol{R^2}$ **tells us how much of the variation** in the price can be explained by taking its size into account

R^2

- $\boldsymbol{R^2}$ **tells us how much of the variation** in the price can be explained by taking its size into account

$$R^2 = \frac{Var(mean) - Var(fit)}{Var(mean)} \qquad \text{or} \qquad R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$





$$SS(mean) = 16111.67$$
$$SS(fit) = 3478.1$$
$$R^2 = 0.784$$

So, 78% of the price of the houses can be explained by its size

Adjusted R^2

- Because models with more features always explain more variation we should use an alternative to the R-squared
- The adjusted R-squared value corrects R-squared by penalizing models with a large number of independent variables!

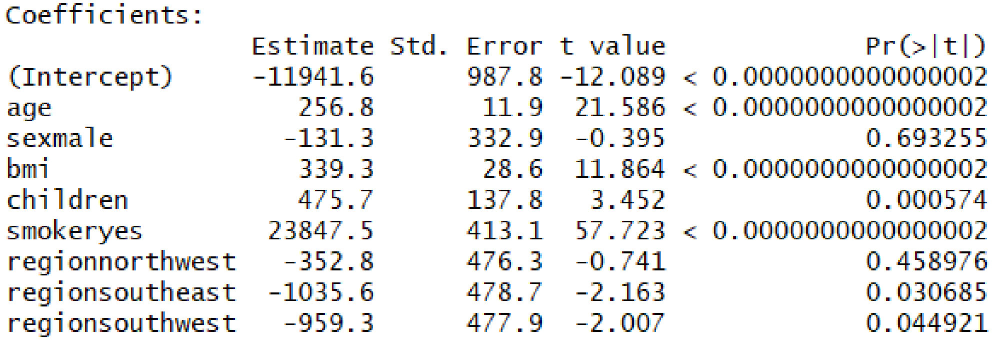
$$Adjusted\ R^2 = 1 - \frac{(1-\ \overset{\text{R-squared}}{\overset{\downarrow}{R^2}})(N-1)}{\underset{\text{\# samples}}{\underset{\uparrow}{N}} - \underset{\text{\# independent variables}}{\underset{\uparrow}{p}} - 1}$$

Example

Imagine we are analyzing a simple dataset of an insurance company, with 1338 observations and 7 variables.

We build a simple regression model that we can use to predict expenses by establishing a statistically significant linear relationship with the independent variables.

Before we use this model, we should ensure that it is **statistically significant**!



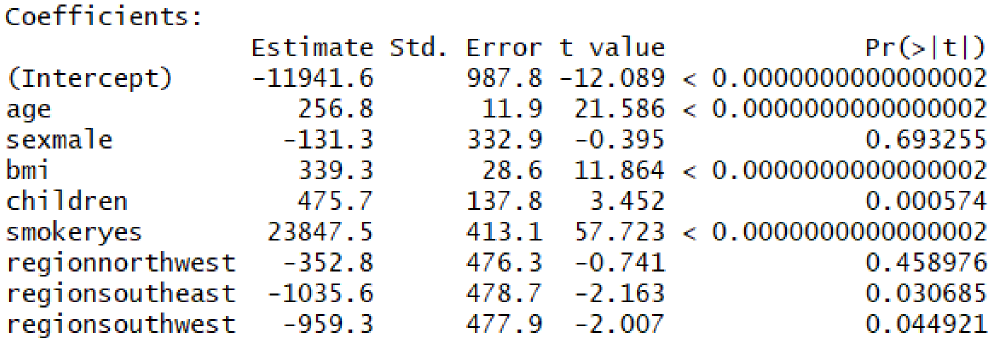
```
Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)      -11941.6      987.8 -12.089 < 0.0000000000000002
age                 256.8       11.9  21.586 < 0.0000000000000002
sexmale            -131.3      332.9  -0.395             0.693255
bmi                 339.3       28.6  11.864 < 0.0000000000000002
children            475.7      137.8   3.452             0.000574
smokeryes         23847.5      413.1  57.723 < 0.0000000000000002
regionnorthwest    -352.8      476.3  -0.741             0.458976
regionsoutheast   -1035.6      478.7  -2.163             0.030685
regionsouthwest    -959.3      477.9  -2.007             0.044921
```

The **standard error** of the coefficient measures the precision of the estimate of the coefficient, how precisely the model estimates the coefficient's unknown value.

The **t-value** is the parameter estimate (aka coefficient) divided by its standard error. The significance of this statistic is given by the **p-value** column.

Example

- **p-value** for each term tests the null hypothesis that the coefficient is equal to zero (no effect).
- A low p-value (< 0.05) indicates that you can reject the null hypothesis.
  - a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.
- **p-values lower than the significance level**, a threshold chosen prior to building the model, **are considered statistically significant**



```
Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)      -11941.6      987.8 -12.089 < 0.0000000000000002
age                 256.8       11.9  21.586 < 0.0000000000000002
sexmale            -131.3      332.9  -0.395             0.693255
bmi                 339.3       28.6  11.864 < 0.0000000000000002
children            475.7      137.8   3.452             0.000574
smokeryes         23847.5      413.1  57.723 < 0.0000000000000002
regionnorthwest    -352.8      476.3  -0.741             0.458976
regionsoutheast   -1035.6      478.7  -2.163             0.030685
regionsouthwest    -959.3      477.9  -2.007             0.044921
```

Example

STATISTIC	CRITERION
R- squared	Higher the better
Adj R-squared	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater than 1.96 for p-value to be less than 0.05
p-value	Should be smaller than the defined threshold

Some references

✓ Greene, W. H. (2012). Econometric Analysis. 7th Edition. Prentice Hall

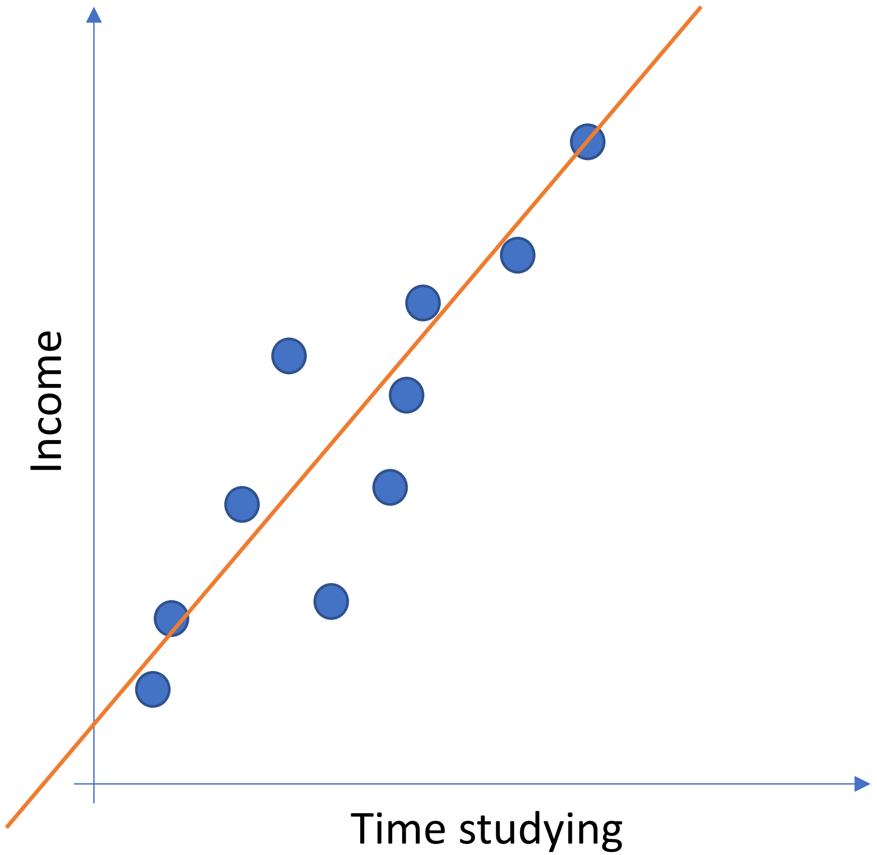
✓ Hill, R. C., Griffiths, W. E. e Lim, G. C. (2012). *Principles of Econometrics*. 4th edition, John Wiley and Sons.

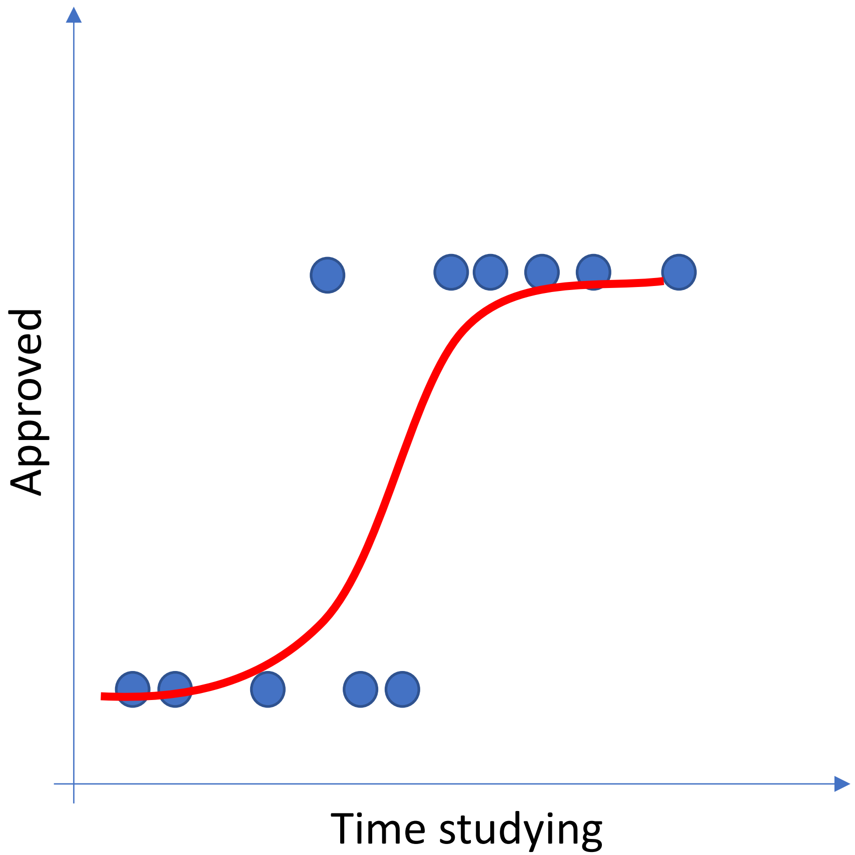
✓ Menard, S. (2010). *Logistic Regression – From Introductory to Advanced Concepts and Applications.* SAGE Publications, Inc..

✓ Wooldridge, J.M. (2012): Introductory Econometrics: A Modern Approach, 5th Edition, South- Western Cengage Learning.

Logistic Regression

Logistic regression





- Logistic regression predicts whether something is true or false, instead of a continuous variable such as income

Linear vs. Logistic Regression

Linear Regression

Outcome: The dependent variable is quantitative/ continuous

Example: height, weight, price…

Coefficient interpretation: straightforward (holding all the other variables constant, with a unit increase in a variable, the dependent variable is expected to increase/decrease by x)

Error minimization technique: uses OLS (Ordinary Least Squares) to minimize the errors

Error term: follows a normal distribution

Logistic Regression

Outcome: The dependent variable is qualitative/ limited to a specific number of possible values

Example: yes/no, true/false, red/green/blue…

Coefficient interpretation: not as straightforward

Error minimization technique: uses MLE (Maximum Likelihood Estimation)

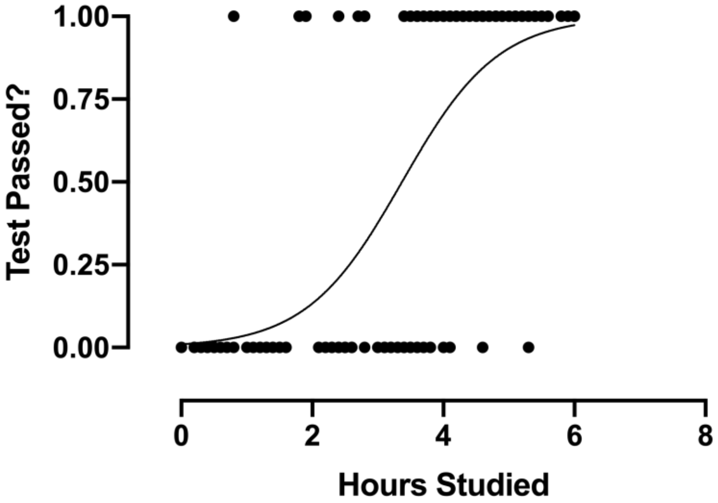
Error term: does not follow a normal distribution

Linear vs. Logistic Regression
Qualitative Target

Let’s consider we want to estimate a model where the dependent variable is limited to a binary response

$$y_i = \begin{cases} 1, if\ student\ passed\ the\ test \\ 0, otherwise \end{cases}$$

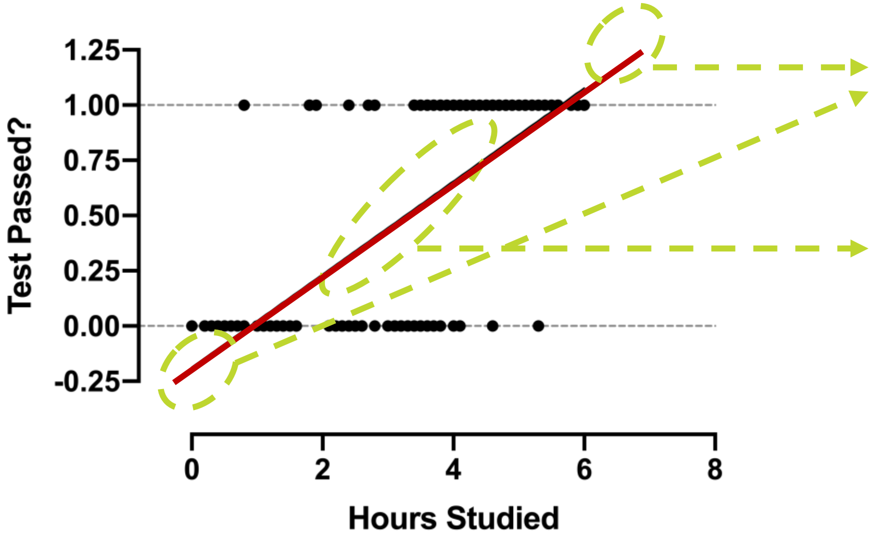
If we apply a scatter diagram to this dataset, we are going to obtain a completely distinct visualization from the ones where the response is continuous!



Reflects that the dependent variable is limited to two possible outcomes

Linear vs. Logistic Regression
Qualitative Target

How to determine the best fit model through data where the dependent variable is limited to a set of outcomes?



Predictions above 1 and bellow 0 don’t make sense

OLS performs well in average observations, but it does a poor job of “fitting” or “describing” the data points

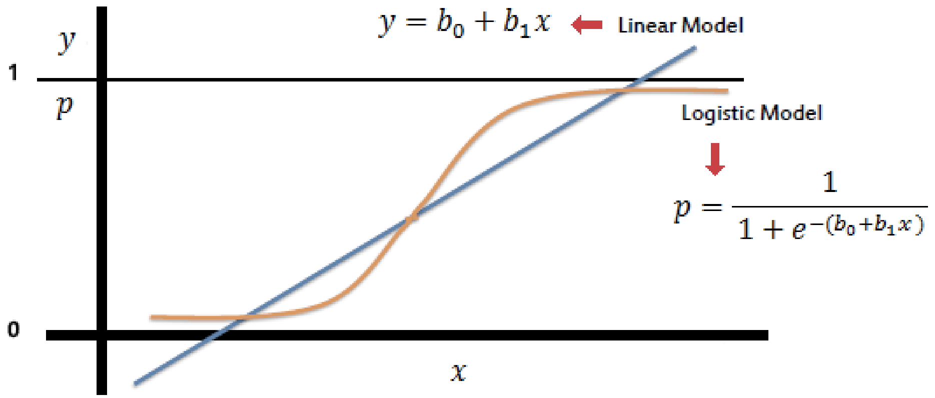
OLS is not an appropriated estimation method if our target is qualitative

Main drawbacks of using linear models in these cases:

1. Having probabilities above 1 or bellow 0, which are impossible outcomes
2. Assuming that the probability changes linearly with the explanatory variables

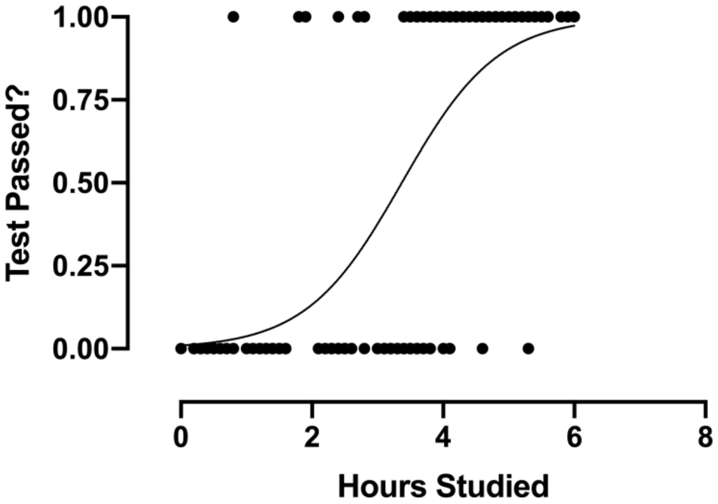
Logistic Regression

- In logistic regression, input features are linearly scaled just as with linear regression however
  - the result are then fed as an input to the logistic function
- This function provides a nonlinear transformation on its input and ensures that the range of the output, which is interpreted as the probability of the input belonging to class 1, lies in the interval [0,1]



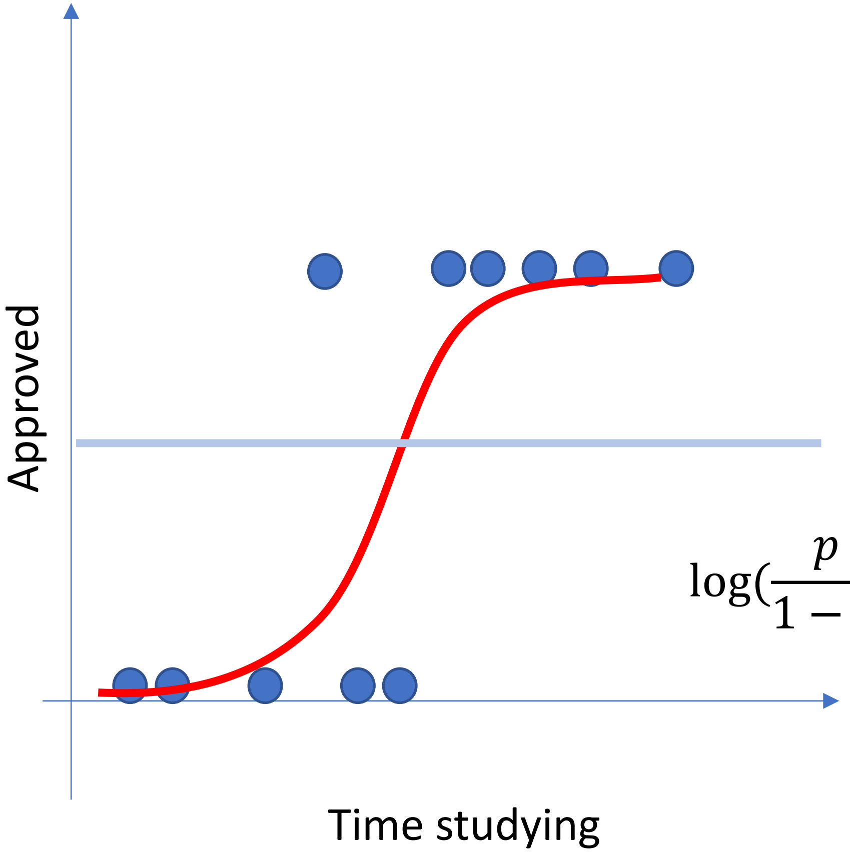
Maximum Likelihood Estimation

- To obtain the estimated vector of parameters we use **Maximum Likelihood Estimation** (MLE)
- This means that we need to find the estimates for β (the vector of the values $\hat{\beta}$) that **maximize the probability, or likelihood, of observing the sample**

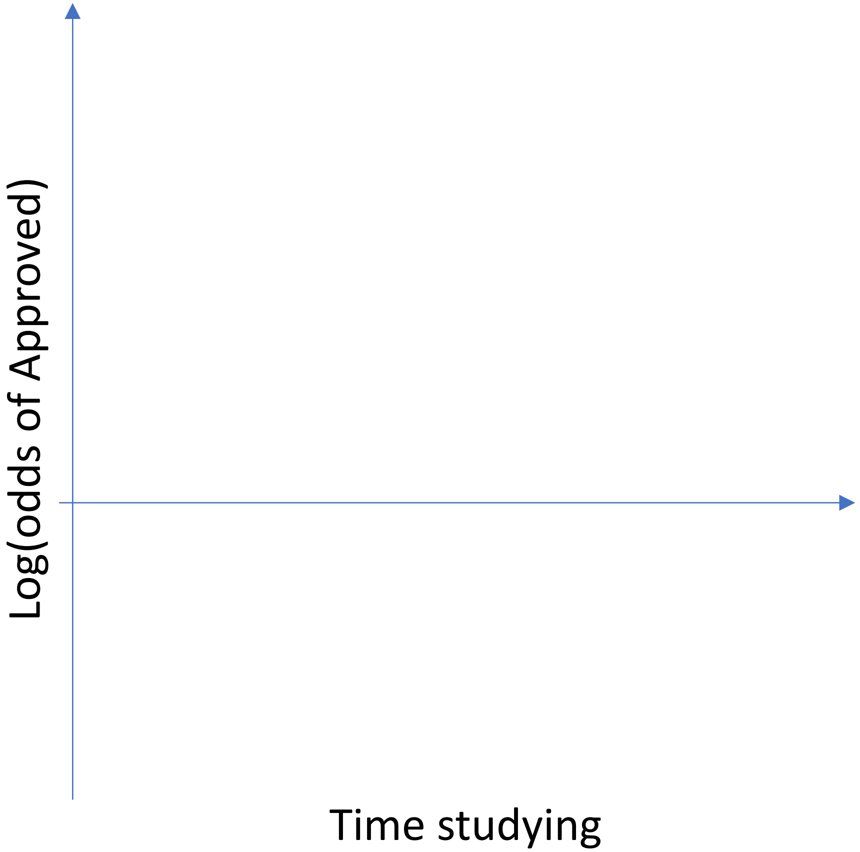


We are trying to find the best "S-shaped" function for our data!

ML



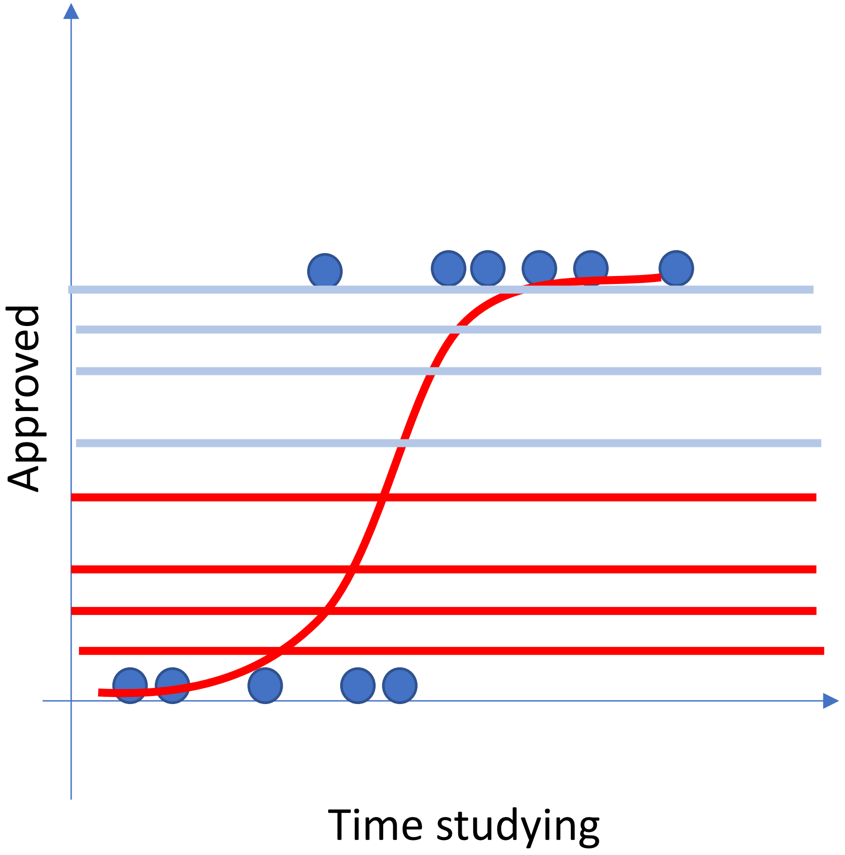
$$\log(\frac{p}{1-p})$$

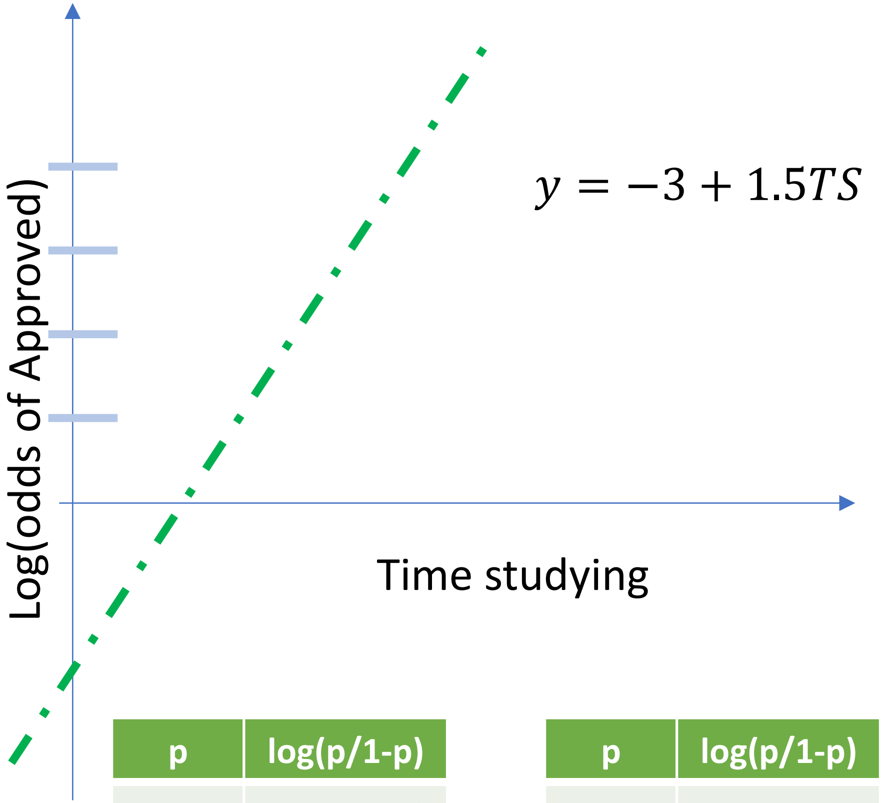


The y axis in a logistic regression is transformed from the probability of Approved to the log(odds of approval)

P is the provbability of a student being approved

ML





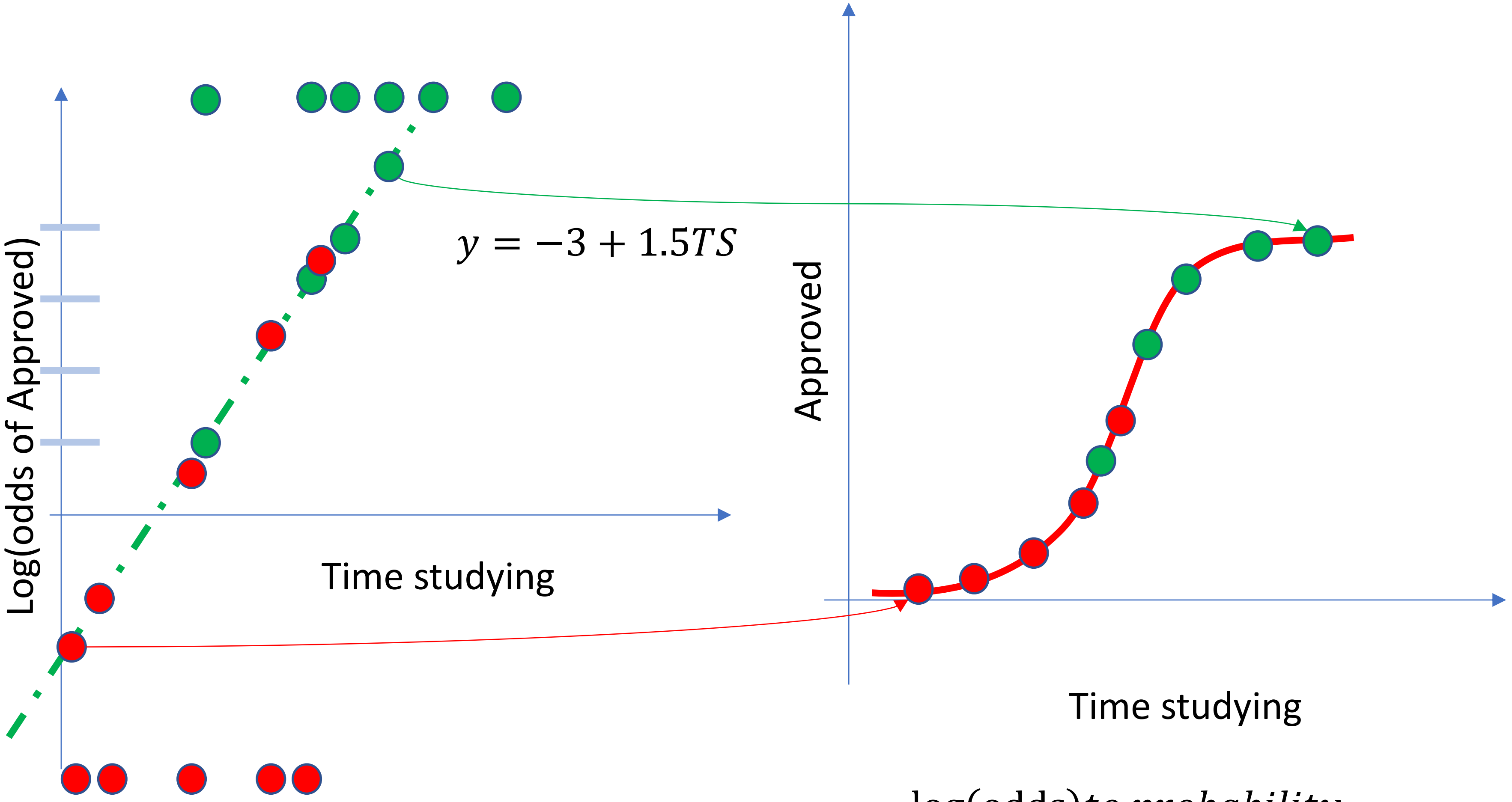
$$y = -3 + 1.5TS$$

probability to log(odds) → $\log(\frac{p}{1-p})$

p	log(p/1-p)
0.5	0
0.731	0.999702
0.88	1.99243
0.95	2.944439
1	#DIV/0!

p	log(p/1-p)
0.5	0
0.27	-0.99462
0.119	-2.00193
0.047	-3.00947
0	#NUM!

ML



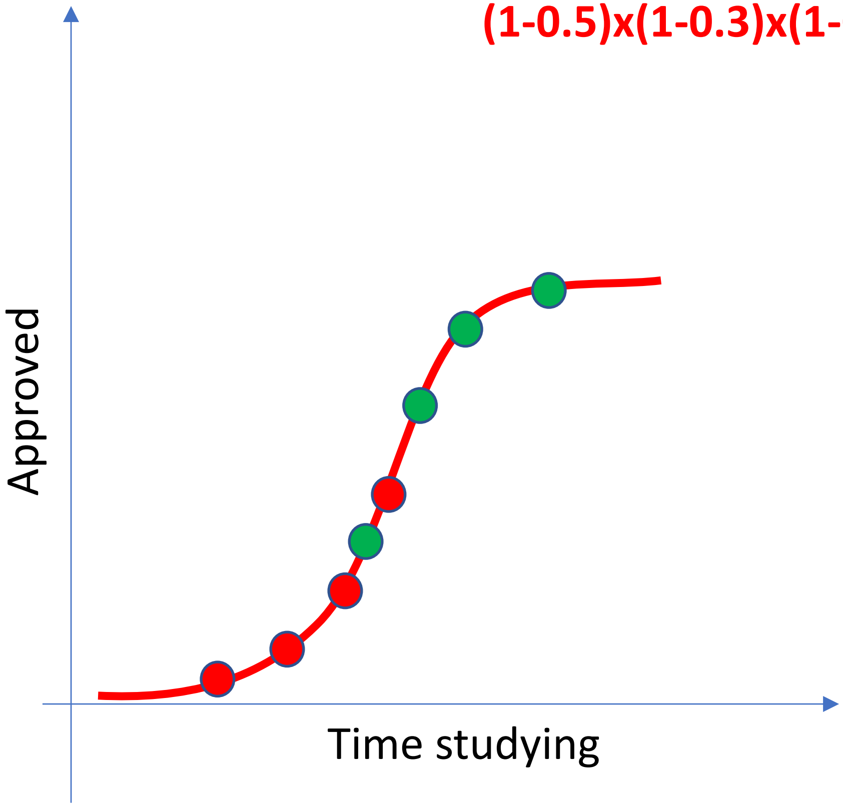
probability to log(odds) → $\log(\frac{p}{1-p})$

$\log(\text{odds})\mathit{to\ probability}$

$$\mathrm{p}\left(\frac{e^{\log(odds)}}{1+e^{\log(odds)}}\right) or\ \mathrm{p}(\frac{1}{1+e^{-\log(odds)}})$$

ML

Likelihood of data given the curve=**0.9x0.8x0.7x0.4 x …**
(1-0.5)x(1-0.3)x(1-0.1)x(1-0.05)



Log(Likelihood of data given the curve)=log(**0.9**)**+** log **(0.8)** +log(**0.7**) +log (**0.4**)
+log**(1-0.5)** +log**(1-0.3)** +log **(1-0.1)** +log **(1-0.05)**
=3.8

Maximum Likelihood Estimation

The parameters are chosen to **maximize the likelihood** of **observing the sample values** rather than minimizing the sum of squared errors

- There are no formulas that give these estimates like there are in least squares estimation of the linear regression model
- One needs to use the computer and techniques from numerical analysis (it is easier to maximize the log-likelihood function, instead of the likelihood)
- ML estimates are then obtained using an iterative algorithm, which starts with arbitrary values and the process is repeated until the log-likelihood doesn’t change significantly

Maximum Likelihood Estimation

This is what the estimated equation will look like:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \ldots + \hat{\beta}_k x_{ki}$$

✓ From the **value of $\boldsymbol{\hat{y}_i}$** one can obtain the estimated probability

$$\hat{P}(y_i = 1|X_i) = \frac{\boldsymbol{e^{\hat{y}_i}}}{\boldsymbol{1 + e^{\hat{y}_i}}}$$

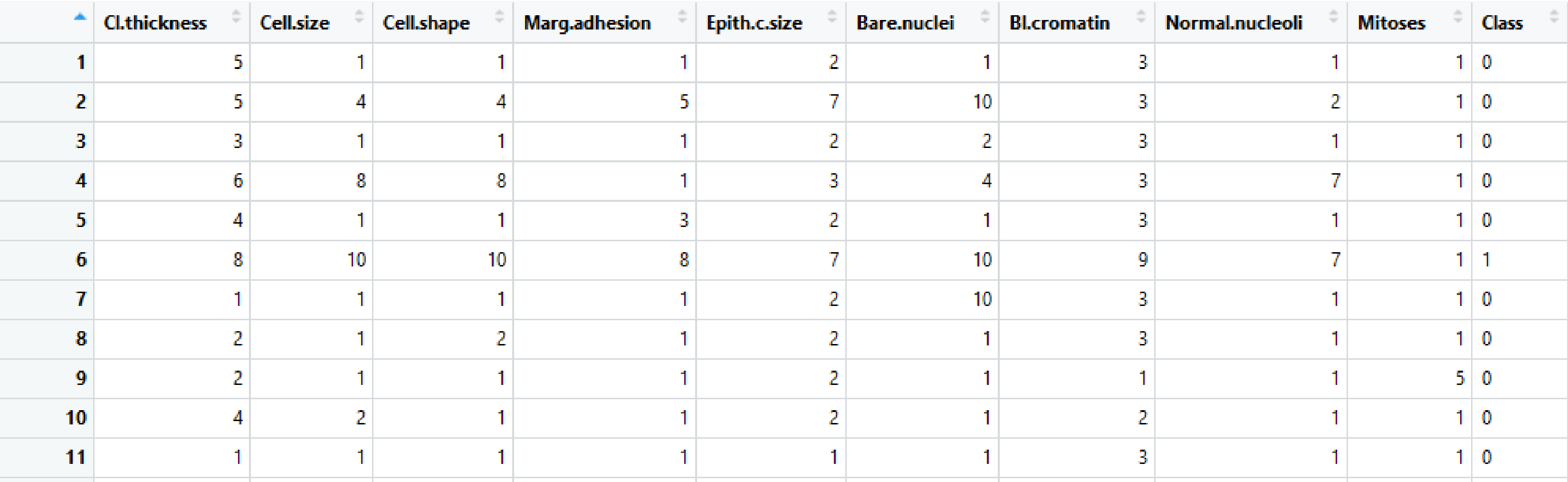
✓ The **parameters $\boldsymbol{\hat{\beta}_k}$** determines the rate of increase or decrease of the S-shaped curve for $\hat{\pi}_i$

- ℹ The sign of $\hat{\beta}_k$ indicates whether the curve ascends $\left(\hat{\beta}_k > 0\right)$ or descends $\left(\hat{\beta}_k < 0\right)$

- ℹ The rate of change in the curve increases as $\left|\hat{\beta}_k\right|$ increases

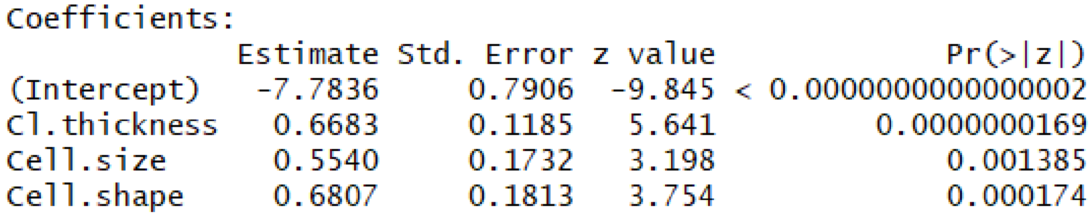
Example

We have a dataset with 683 observations, and the column ‘Class’ is our dependent variable, that tells us if a given tissue is malignant or benign.



	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1	5	1	1	1	2	1	3	1	1	0
2	5	4	4	5	7	10	3	2	1	0
3	3	1	1	1	2	2	3	1	1	0
4	6	8	8	1	3	4	3	7	1	0
5	4	1	1	3	2	1	3	1	1	0
6	8	10	10	8	7	10	9	7	1	1
7	1	1	1	1	2	10	3	1	1	0
8	2	1	2	1	2	1	3	1	1	0
9	2	1	1	1	2	1	1	1	5	0
10	4	2	1	1	2	1	2	1	1	0
11	1	1	1	1	1	1	3	1	1	0

Example



```
Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)    -7.7836     0.7906  -9.845 < 0.0000000000000002
Cl.thickness    0.6683     0.1185   5.641         0.0000000169
Cell.size       0.5540     0.1732   3.198             0.001385
Cell.shape      0.6807     0.1813   3.754             0.000174
```

⚠ **Contrarily to linear regression, we cannot directly interpret the coefficient estimates, but we can say that:**

✓ The positive sign in the thickness, the size and the shape of the cell $\left(\hat{\beta}_k > 0\right)$ indicates that the curve ascends – increase in the probability

✓ The rate of change in the curve is higher for shape than for the size of the cell, since the rate of change in the curve increases as $\left|\hat{\beta}_k\right|$ increases

Example

Here we have our estimated equation:

$$\widehat{malign}_i = -7.7836 + 0.6683 \text{ thickness} + 0.5540 \text{ size} + 0.6807 \text{ shape}$$

If we consider a cell where the **thickness is 5**, the **size is 2** and the **shape is 2**, we can easily compute the predicted probability of the cell being malign:

$$\widehat{malign}_i = -7.7836 + 0.6683 * 5 + 0.5540 * 2 + 0.6807 * 2 = -1.9727$$

$$\Lambda\left(\widehat{malign}_i\right) = \frac{e^{\widehat{malign}_i}}{1 + e^{\widehat{malign}_i}} = \frac{e^{-1.9727}}{1 + e^{-1.9727}} = 0{,}1221$$

- In the scenario described above, there’s an expected **probability of the cell being malign equal to 12%**!

Example

What if a cell has **tickness of 10**, **size of 8** and **shape of 9**?

$\widehat{malign}_i =$ -7.7836 + 0.6683 * 10 + 0.5540 * 8 + 0.6807 * 9 = 9.4577

$$\Lambda\left(\widehat{malign}_i\right) = \frac{e^{\widehat{malign}_i}}{1 + e^{\widehat{malign}_i}} = \frac{e^{9.4577}}{1 + e^{9.4577}} = 0.999$$

💡 In the scenario described above, there's an **expected probability of the cell being malign of 99%**!

NOVA IMS INFORMATION MANAGEMENT SCHOOL