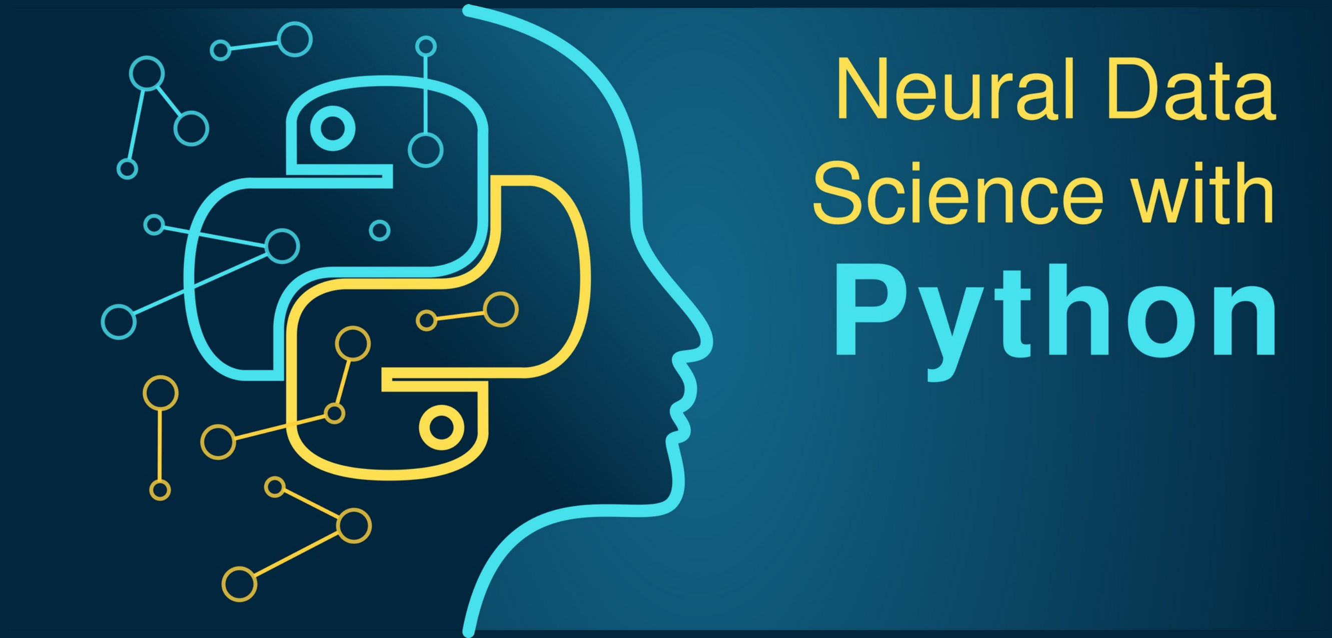


L7 : Regression Analysis

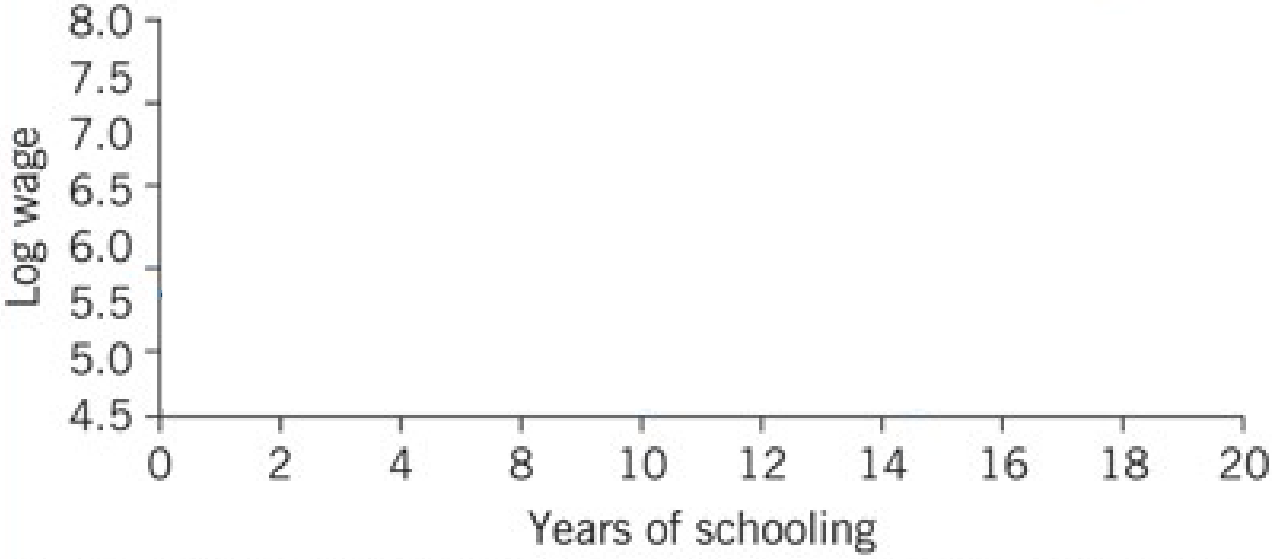
Michael Graupner

SPPIN – Saint-Pères Institute for the Neurosciences

Université de Paris, CNRS

Testing relationships !?

A simple linear regression can investigate the average relationship between two variables

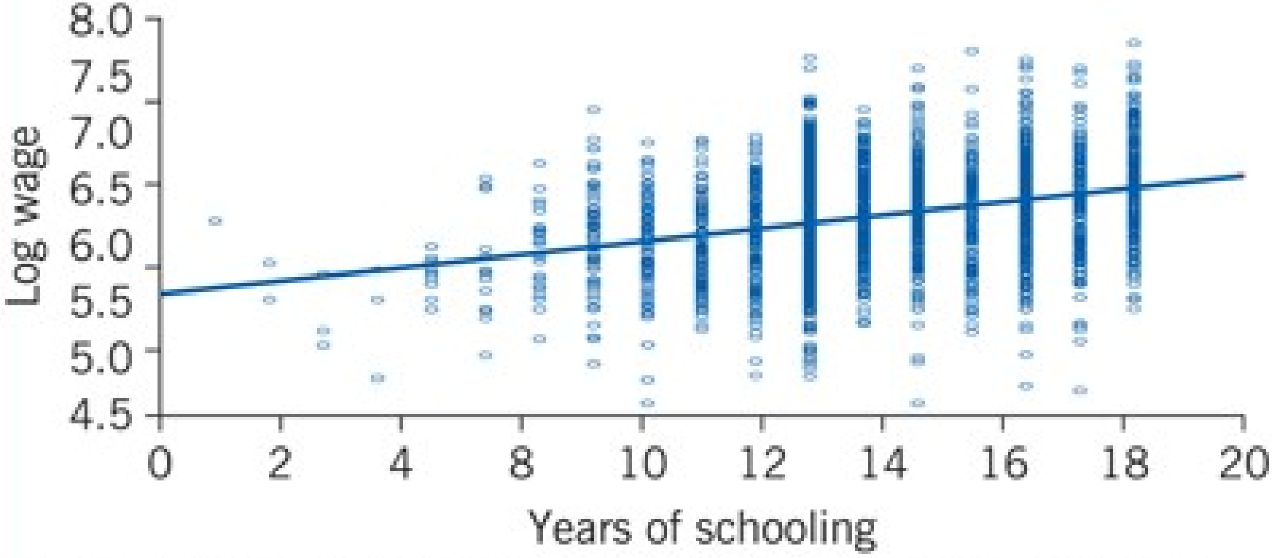


Source: Author's regression using data from [1] on 3,010 men from the US National Longitudinal Survey of Young Men. Online at: http://www.bls.gov/nls/

I Z A
World of Labor

Using linear regression to establish relationships

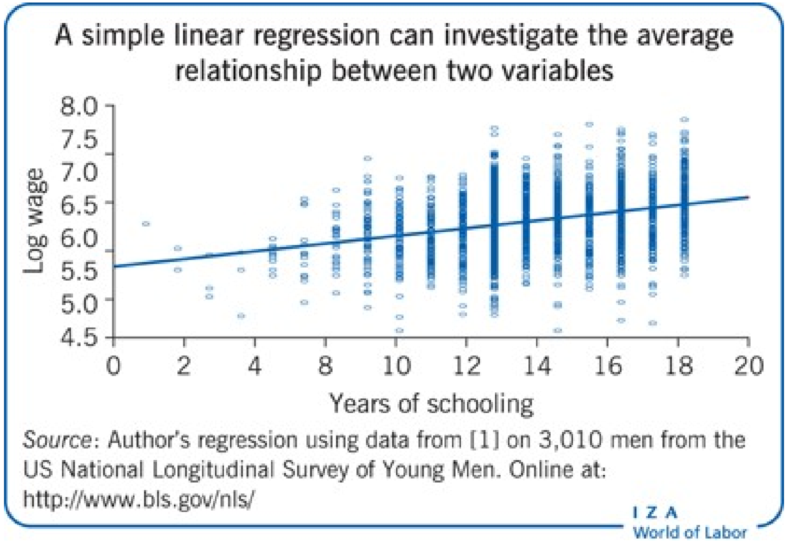
A simple linear regression can investigate the average relationship between two variables



Source: Author's regression using data from [1] on 3,010 men from the US National Longitudinal Survey of Young Men. Online at: http://www.bls.gov/nls/

I Z A
World of Labor

Using linear regression to establish relationships


A simple linear regression can investigate the average relationship between two variables

Source: Author's regression using data from [1] on 3,010 men from the US National Longitudinal Survey of Young Men. Online at: http://www.bls.gov/nls/

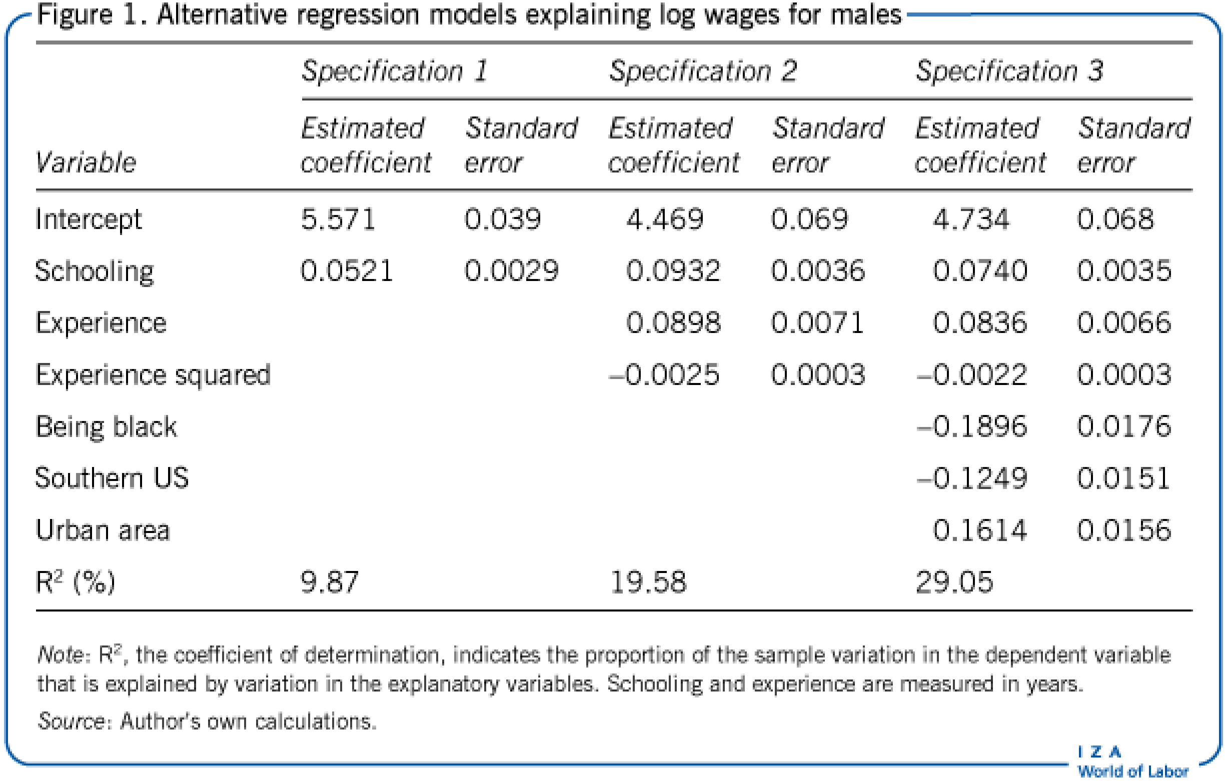


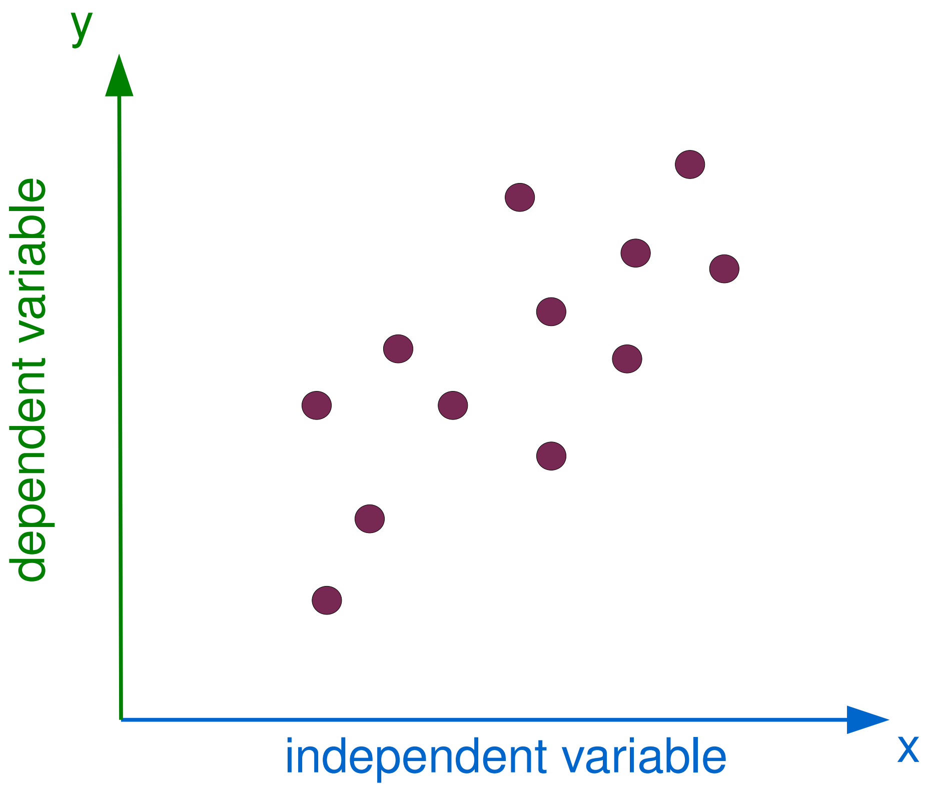
Figure 1. Alternative regression models explaining log wages for males

	Specification 1		*Specification 2*		*Specification 3*	
Variable	*Estimated coefficient*	*Standard error*	*Estimated coefficient*	*Standard error*	*Estimated coefficient*	*Standard error*
Intercept	5.571	0.039	4.469	0.069	4.734	0.068
Schooling	0.0521	0.0029	0.0932	0.0036	0.0740	0.0035
Experience			0.0898	0.0071	0.0836	0.0066
Experience squared			–0.0025	0.0003	–0.0022	0.0003
Being black					–0.1896	0.0176
Southern US					–0.1249	0.0151
Urban area					0.1614	0.0156
R^2 (%)	9.87		19.58		29.05	

Note: R^2, the coefficient of determination, indicates the proportion of the sample variation in the dependent variable that is explained by variation in the explanatory variables. Schooling and experience are measured in years.

Source: Author's own calculations.

What is regression analysis ?

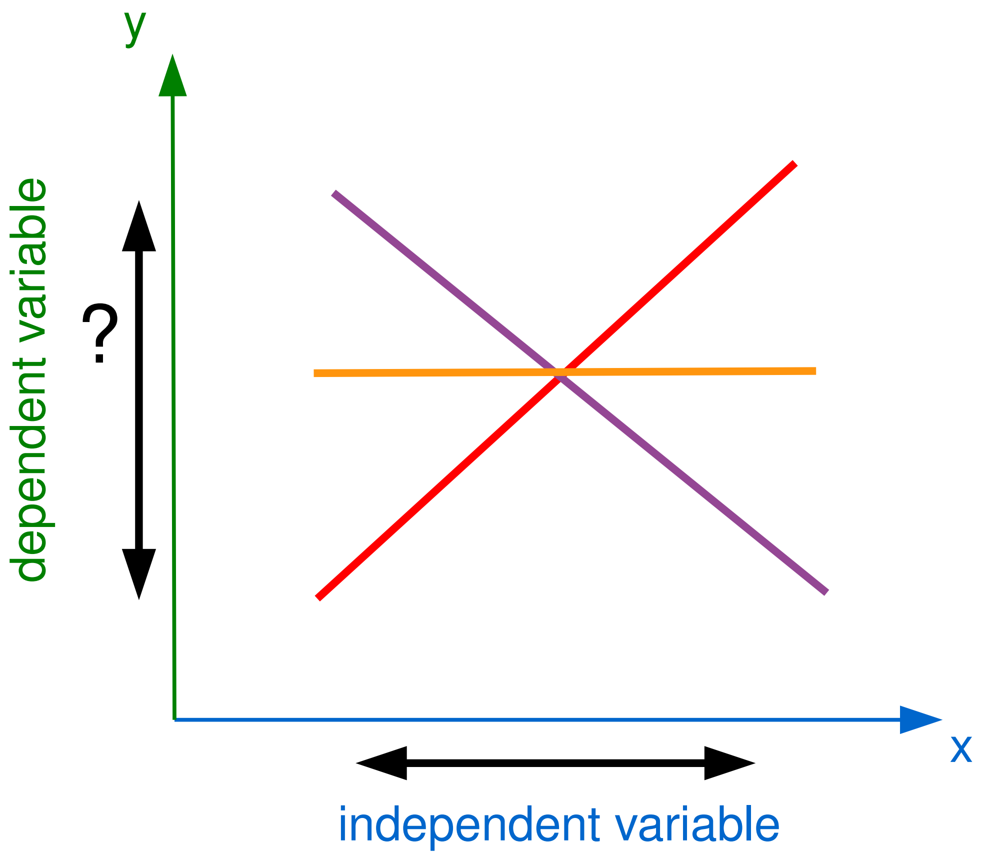


Dependent Variable: This is the main factor that we are trying to understand or predict.

Independent Variables (predictor): These are the factors that we hypothesize have an impact on your dependent variable.

Observations: Data points -> measured relations between independent and dependent variable.

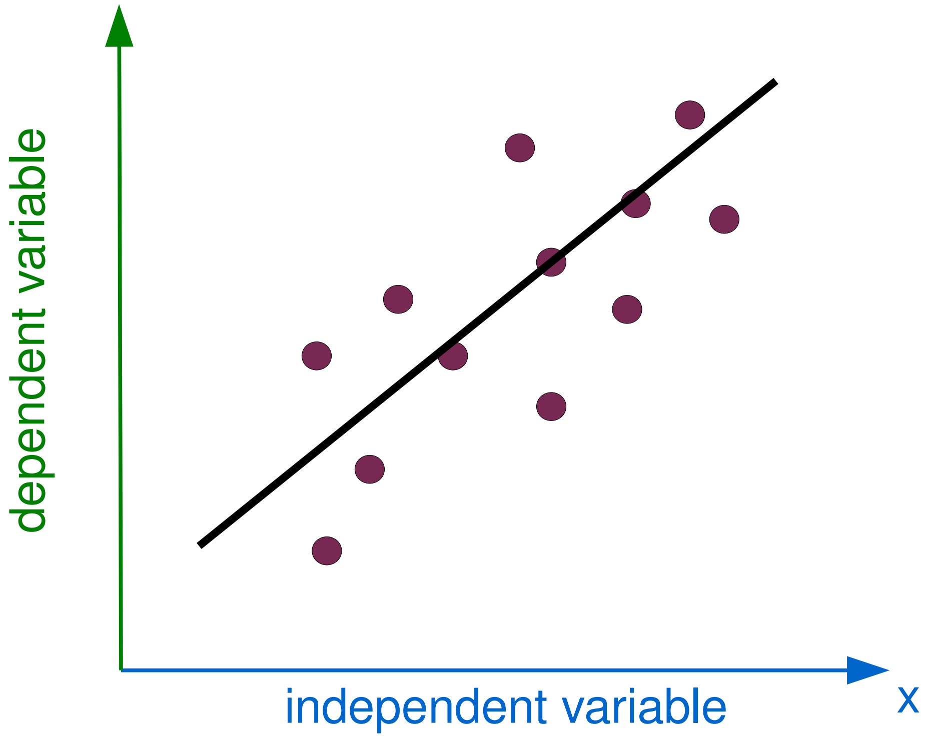
Aim of regression analysis : predict change



As the independent variable is changing, what happens to the dependent variable ?

1. positive relationship / positive correlation :
   independent var. ↗ → dependent var. ↗
2. negative relationship / negative correlation :
   independent var. ↗ → dependent var. ↘
3. no relationship/no correlation/uncorrelated :
   independent var. ↗ or ↘ → no effect on dependent var.

Linear regression – fit a line to the observations

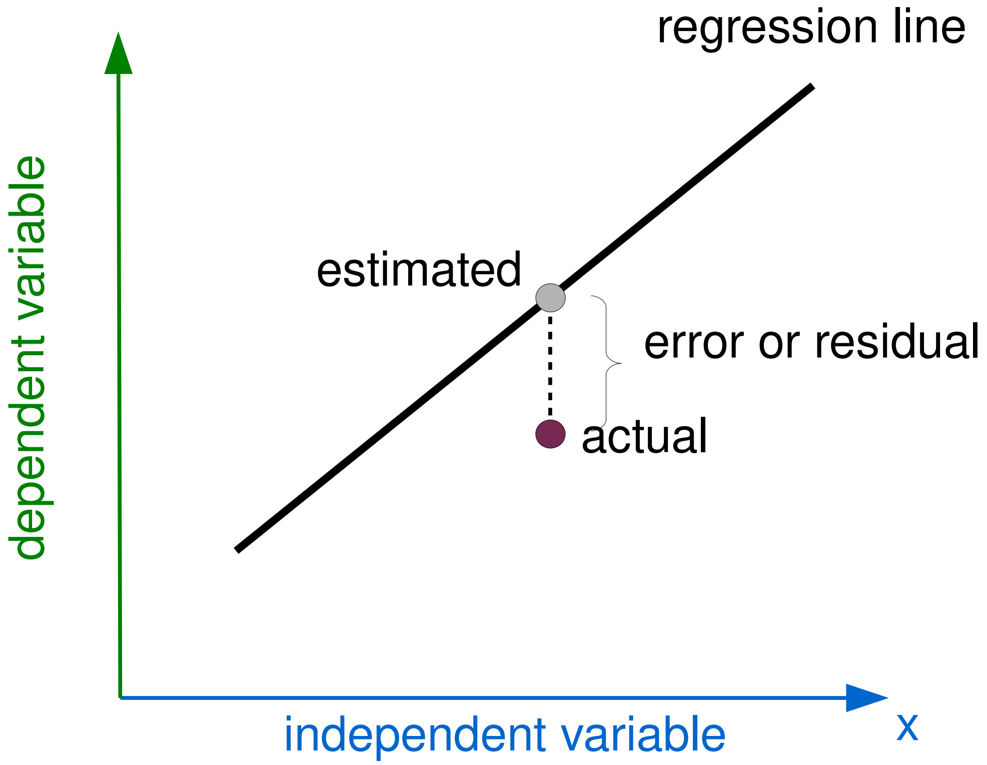


Linear regression finds the best fit line to a cloud of points were we are trying to predict one variable of interest from the know value of another variable.

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

X ... independent variable or predictor
Y ... dependent or predicated variable
β ... weights or parameters
β_0 ... offset, y - intercept
β_1 ... slope
ε ... additive error term

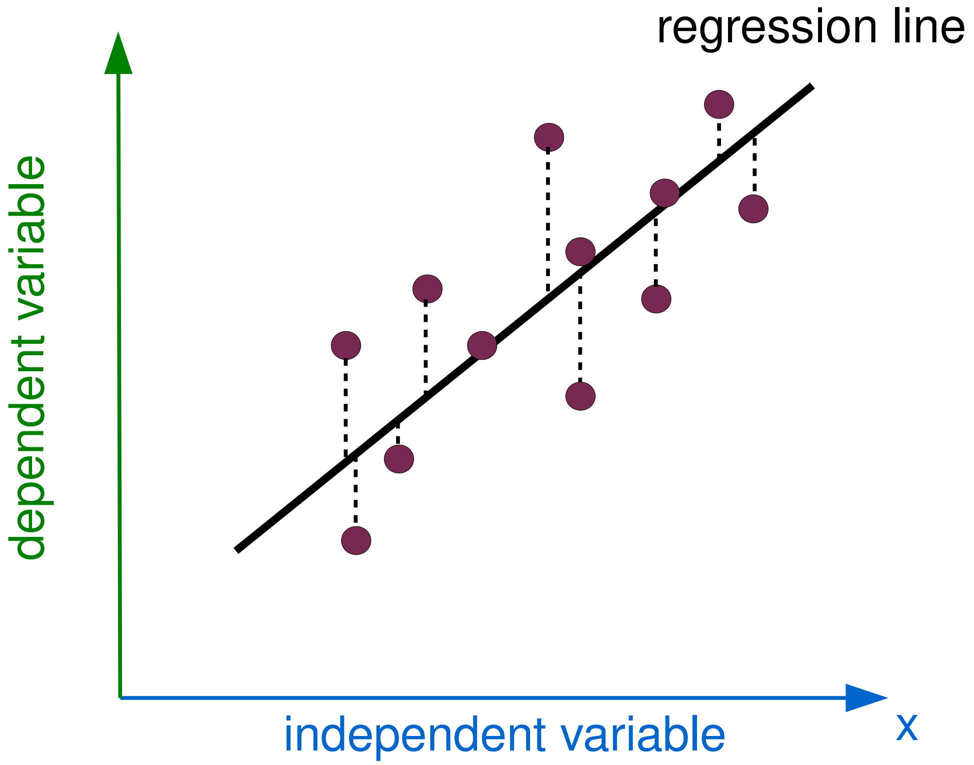
Linear regression – fit a line to the observations



Regression line is found by minimizing the difference between the estimated and the actual value.

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

Linear regression – fit a line to the observations

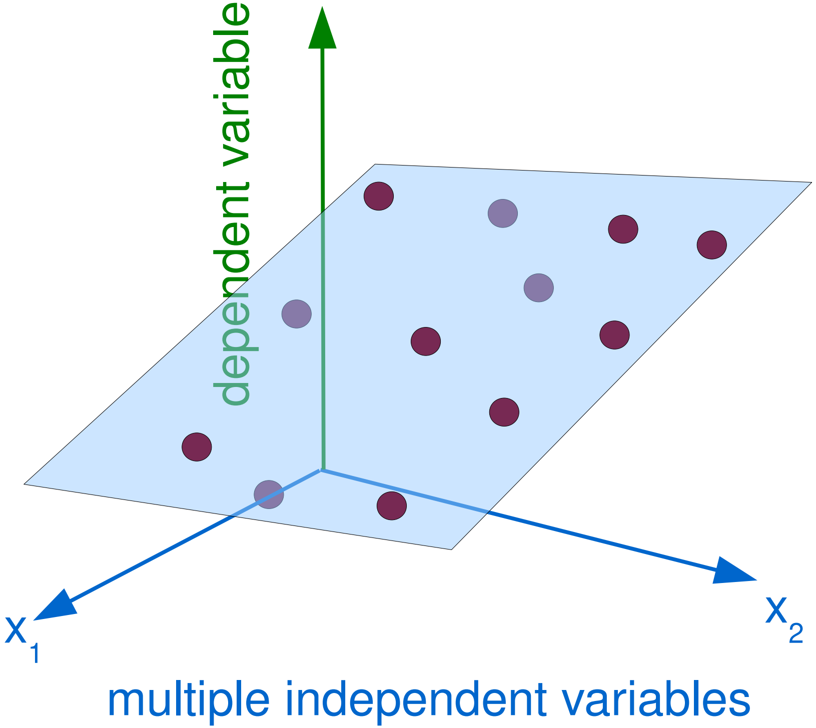


Regression line is found by minimizing the difference between the estimated and the actual value.

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

β_0 and β_1 are optimized by minimizing all errors through a least square method.

Multiple linear regression

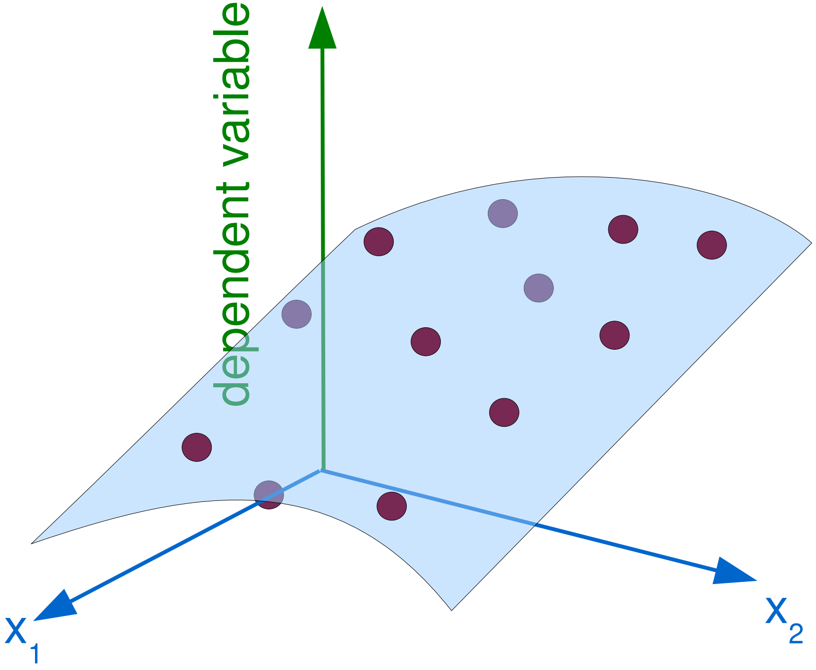


Regression with multiple predictors :
→ multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- becomes the equation of a plane
- β - weights measure relative influence of independent variables on dependent variable

Multiple linear regression with interaction term



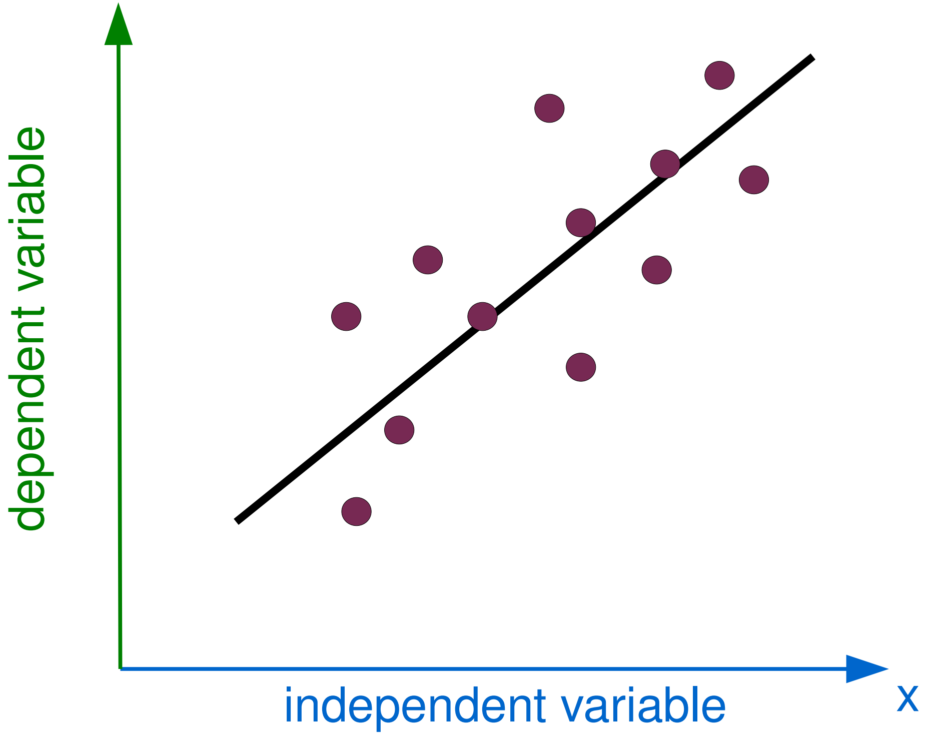
multiple independent variables

If independent variables are not independent of each other :
→ interaction term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- equation of a curved plane
- interaction terms add weights to be estimated by the fitting procedure

Linear regression : R squared value



***r* squared** or r^2 or R^2 (coefficient of determination) denotes the proportion of variation in the dependent variable that can be accounted for by the model/ regression line.

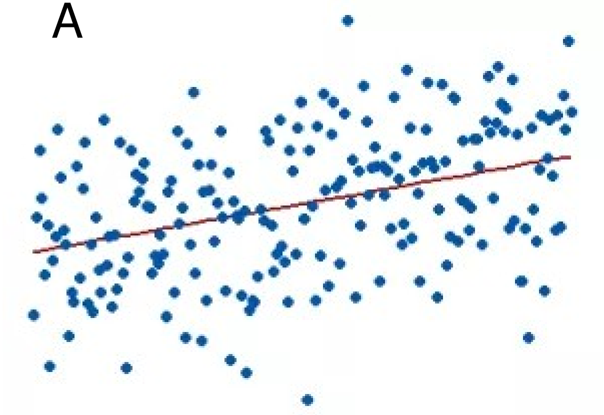
$$R^2 = \frac{\textit{variance explained by the model}}{\textit{total variance}}$$

- $R^2 = 1$: we can perfectly predict all values in the data (suspicious)
- $R^2 = 0$: model fails to predict any of the data

Linear regression : R squared value

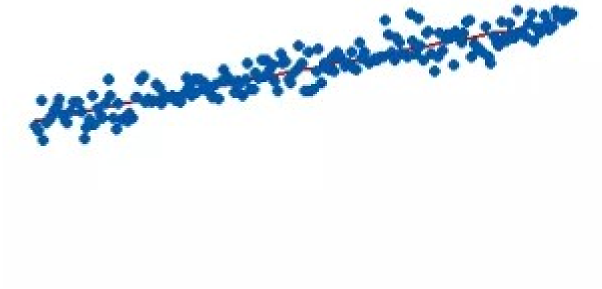
$R_A^2 = 15\,\%$

A



$R_B^2 = 85\,\%$

B

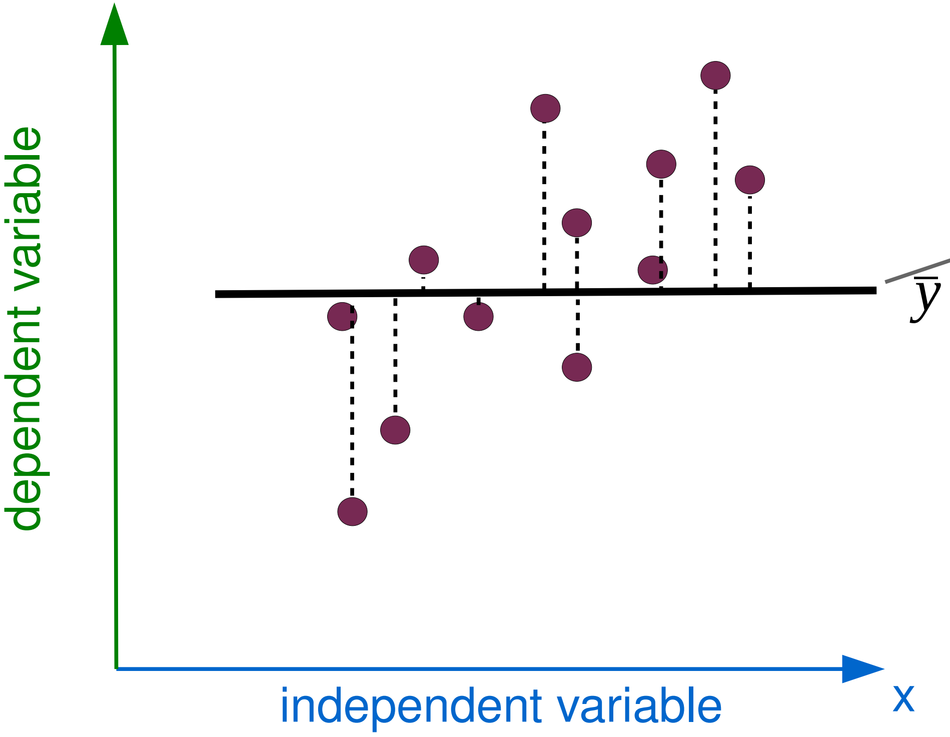


Both data-sets show a positive correlation between independent and dependent variable.

When a regression model accounts for more of the variance, the data points are closer to the regression line.

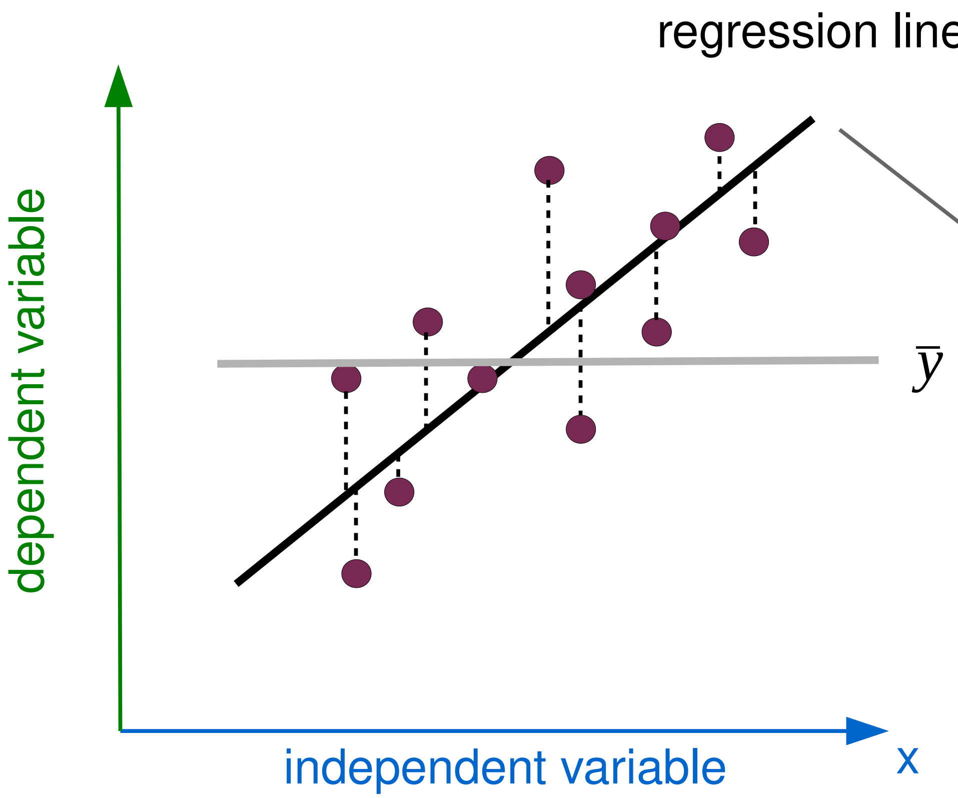
Linear regression : calculate R squared value

- total sum of squares (proportional to the variance of the data):



$$SS_{tot} = \sum_{i} (y_i - \bar{y})^2$$

Linear regression : calculate R squared value



- total sum of squares (proportional to the variance of the data):

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

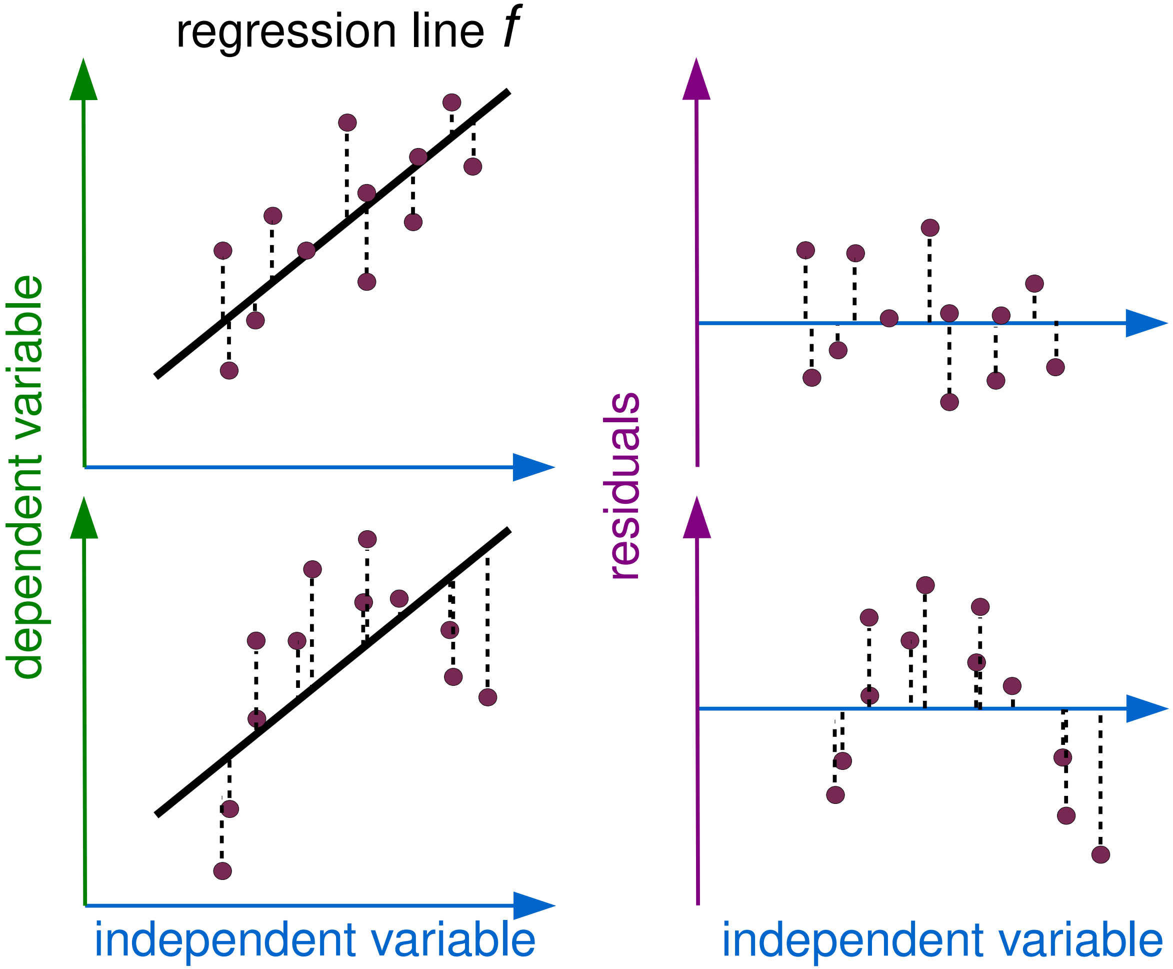
- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{res} = \sum_i (y_i - f_i)^2$$

- general definition of R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Linear regression : residuals

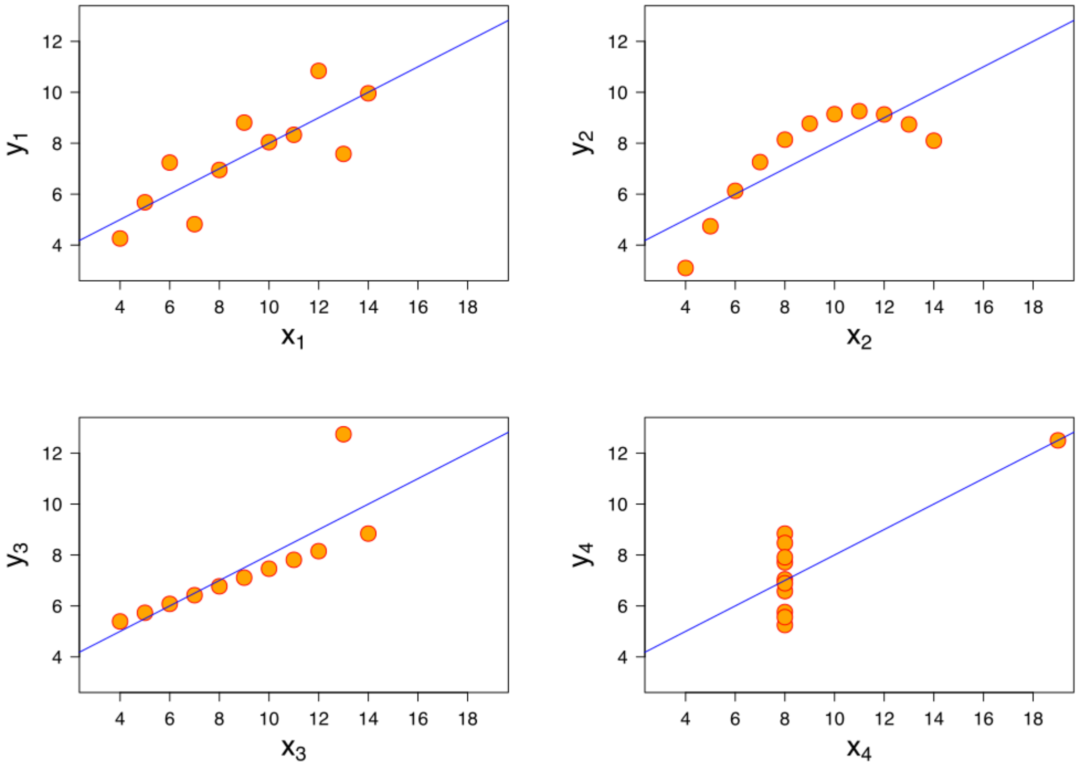


Residuals are the differences between the observations and the regression line (the function *f*)

$$residuals = (y_i - f_i)$$

- residuals versus independent variable plot emphasizes unwanted pattern
- An unbiased model has residuals that are randomly scattered around zero
- Non-random residual patterns indicate a bad fit, a bias or wrong model

Pitfalls of linear regression



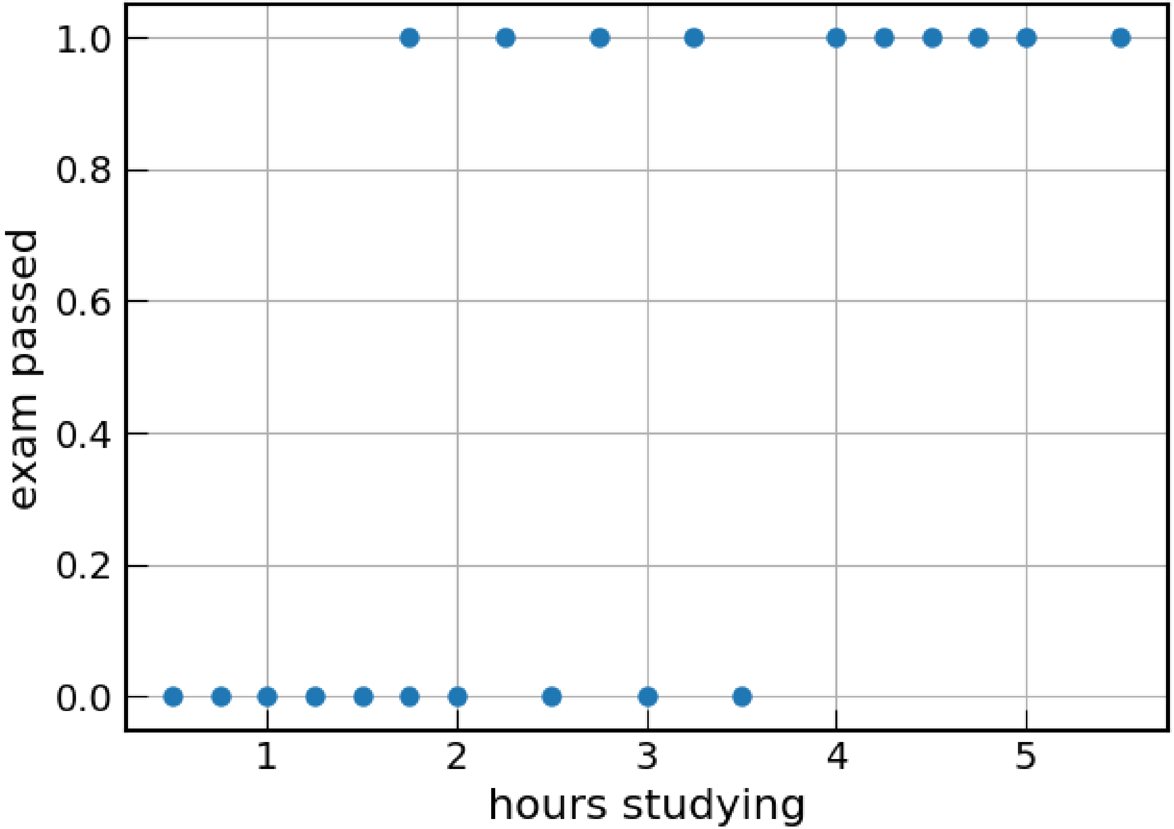
- data with similar regression lines and R^2 values
- observations are graphically very different

→ always inspect data and regression line visually

→ check residuals

→ linear model might not be enough

Regression with binary outcomes

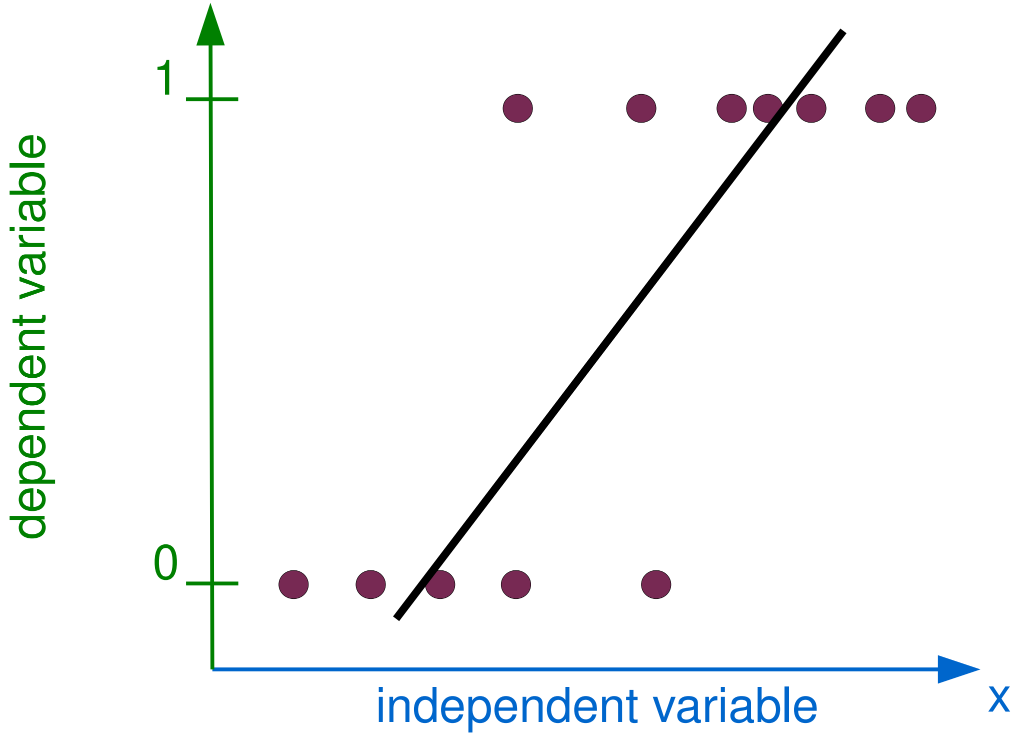


Example

A group of 20 students spend between 0 and 6 hours studying for an exam.

How does the number of hours spent studying affect the probability that the student will pass the exam?

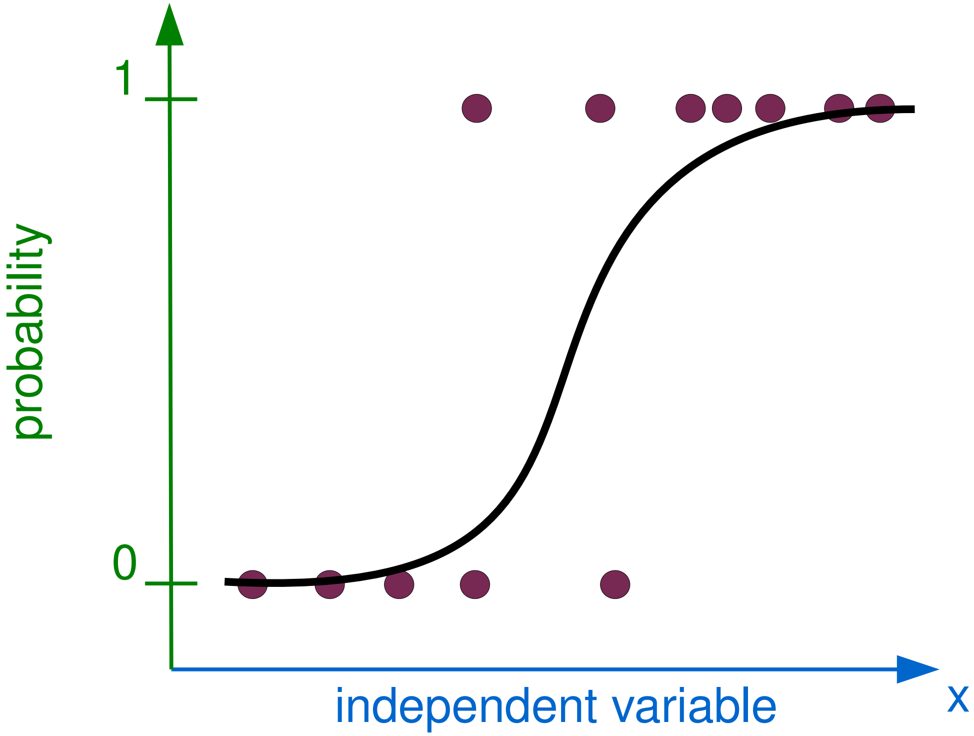
Regression with binary outcomes



In many cases outcomes are binary, in turn we desire to predict : win or loss, up or down votes, buy or sell decisions, life or death, approach or avoid, stay or fight, fight or flight ... .

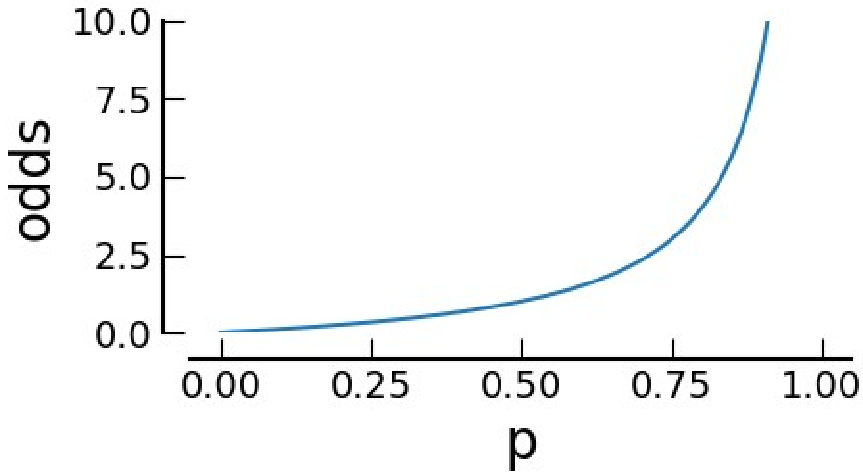
So far, we considered cases where the dependent variable is continuous, for which case we used linear regression.

Logistic regression



Logistic regression is a nonlinear model to link predictors and outcomes through a *sigmoidal* function. It gives the *odds* that an outcome happens – vs. it not happening for a given value of the independent variable (predictor value).

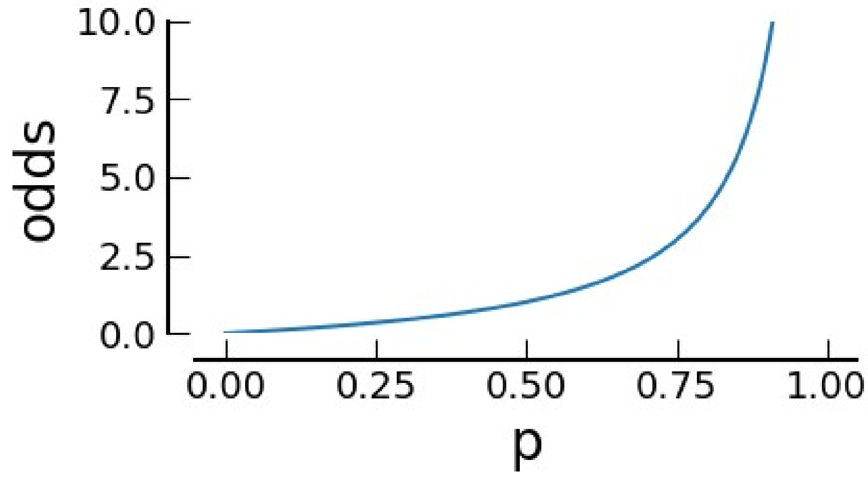
What are the odds ?



Odds is the ratio between the likelihood of an event occurring vs. it not occurring.

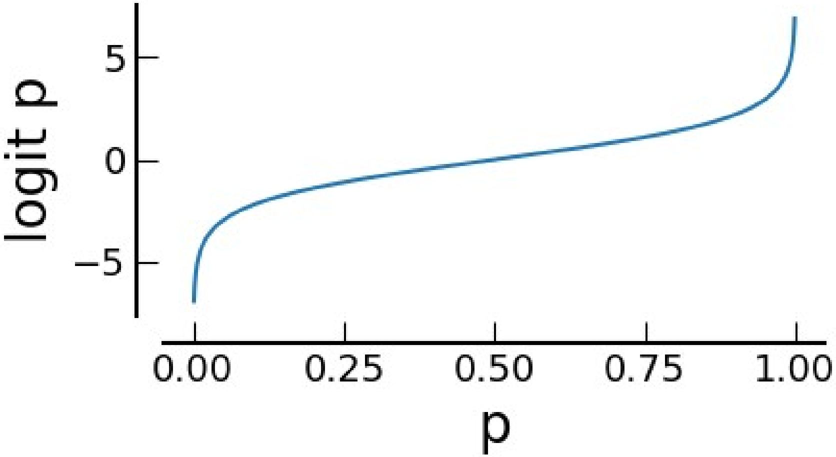
$$Odds = \frac{p}{1-p}$$

What are the odds ?



Odds is the ratio between the likelihood of an event occurring vs. it not occurring.

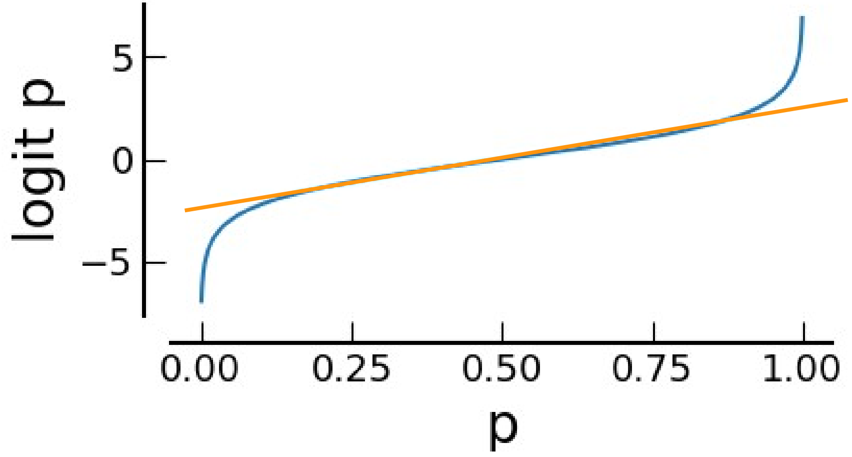
$$Odds = \frac{p}{1-p}$$



The logit or the log-odds function is a type of function that creates a map of probability values from [0,1] to {-∞,+∞}, and it is symmetric

$$logit(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right)$$

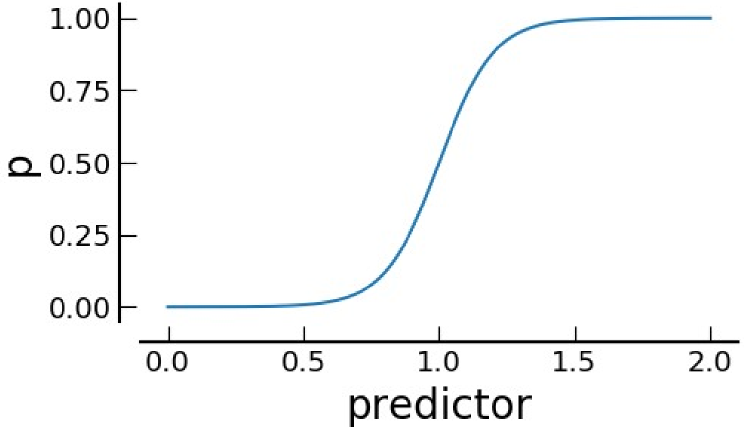
Logistic regression: fit line to logit function



regression line is fitted to the logit function :

$$logit(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

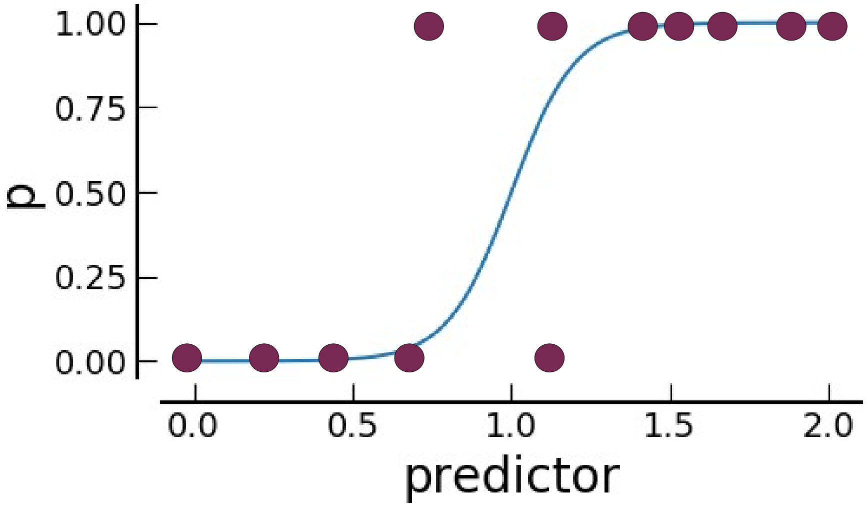
→ solve for p



logistic function : links predictor (x) to the probability of outcome (p)

$$p = \frac{e^{\beta_0 + \beta_1 X_1}}{\left(1 + e^{\beta_0 + \beta_1 X_1}\right)}$$

Logistic regression



$$p = \frac{e^{\beta_0 + \beta_1 X_1}}{\left(1 + e^{\beta_0 + \beta_1 X_1}\right)}$$

logistic regression: estimates the weight that best link predictor to outcomes in a maximum likelihood estimation.

→ provides probability given the predictor variable