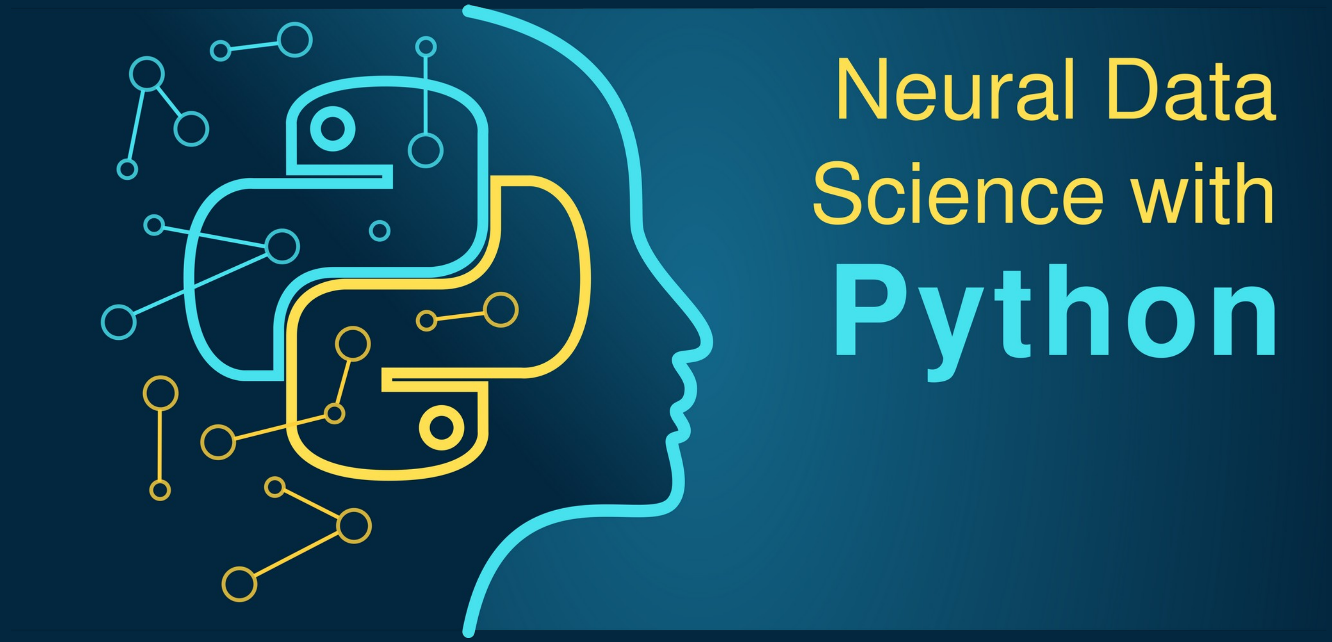


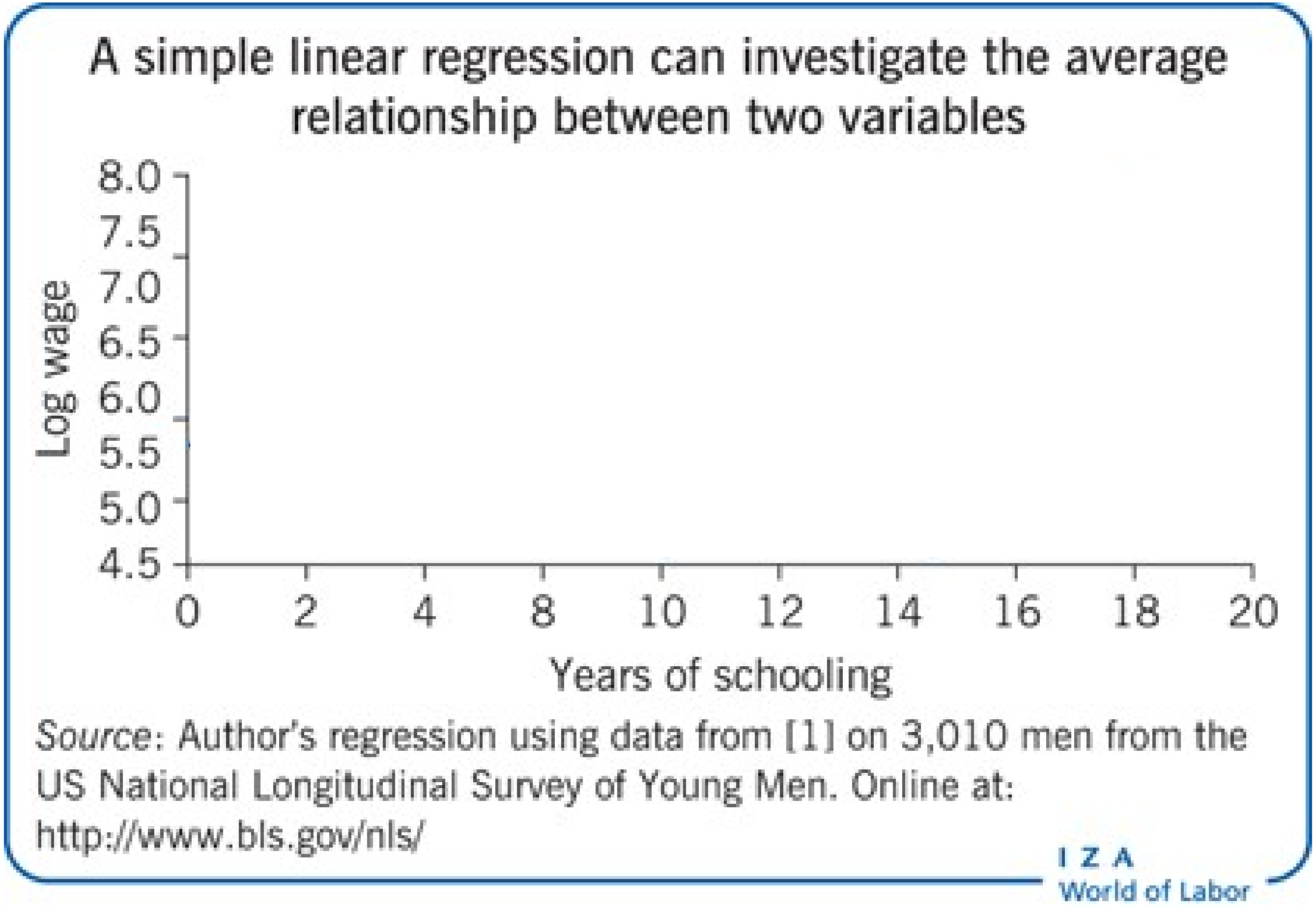
L8 : Regression Analysis

Michael Graupner

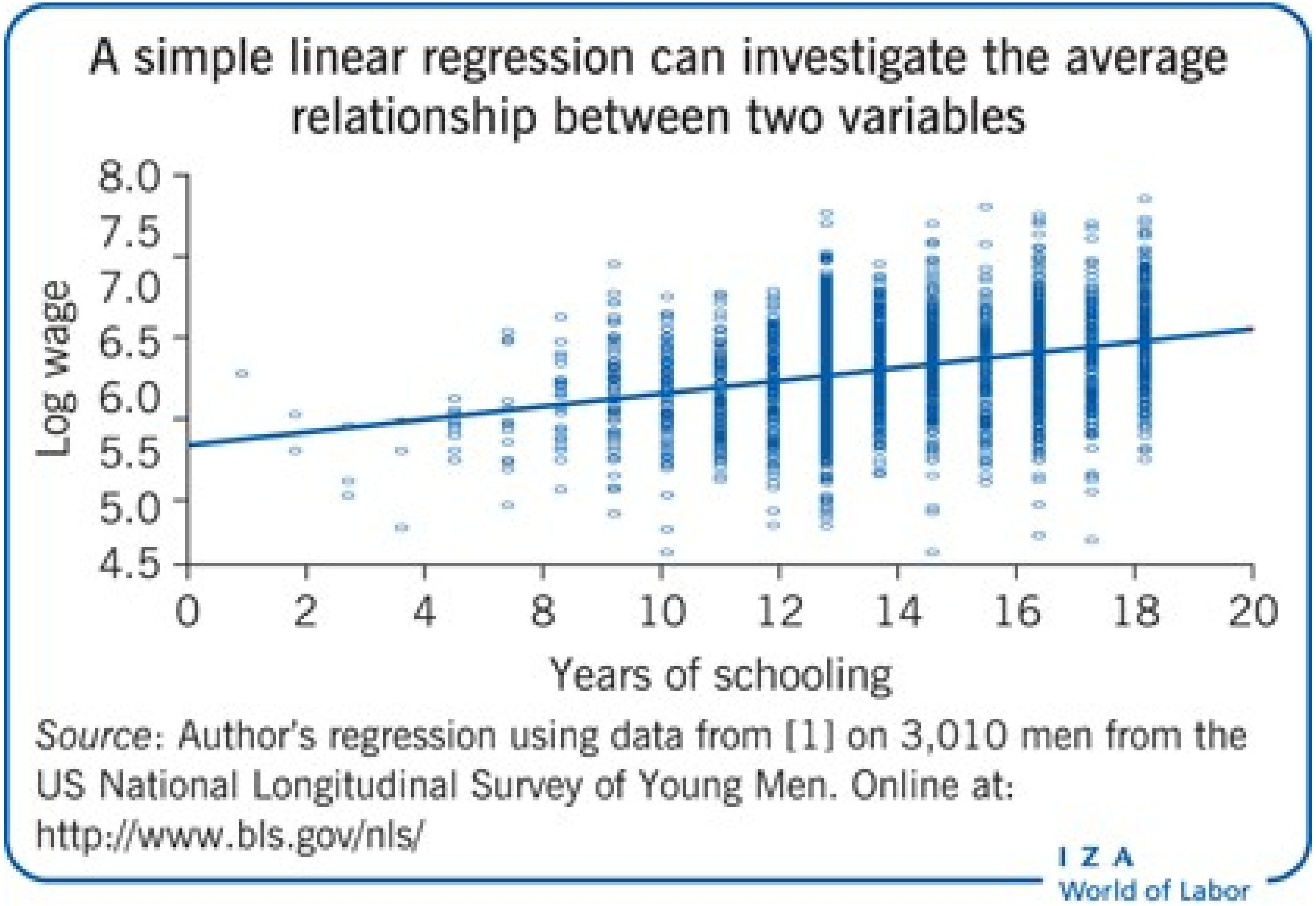
SPPIN – Saint-Pères Institute for the Neurosciences

Université de Paris, CNRS

Testing relationships !?

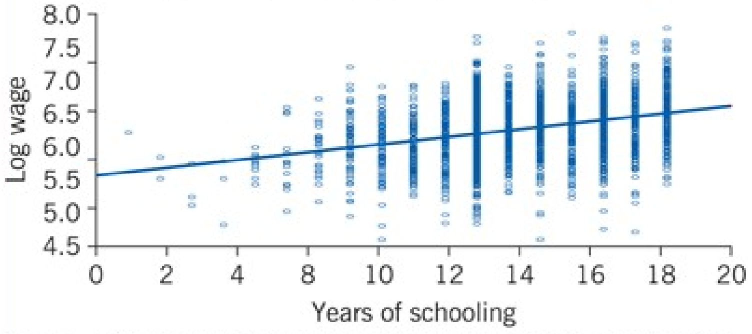


Using linear regression to establish relationships



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/

IZA World of Labor

Figure 1. Alternative regression models explaining log wages for males

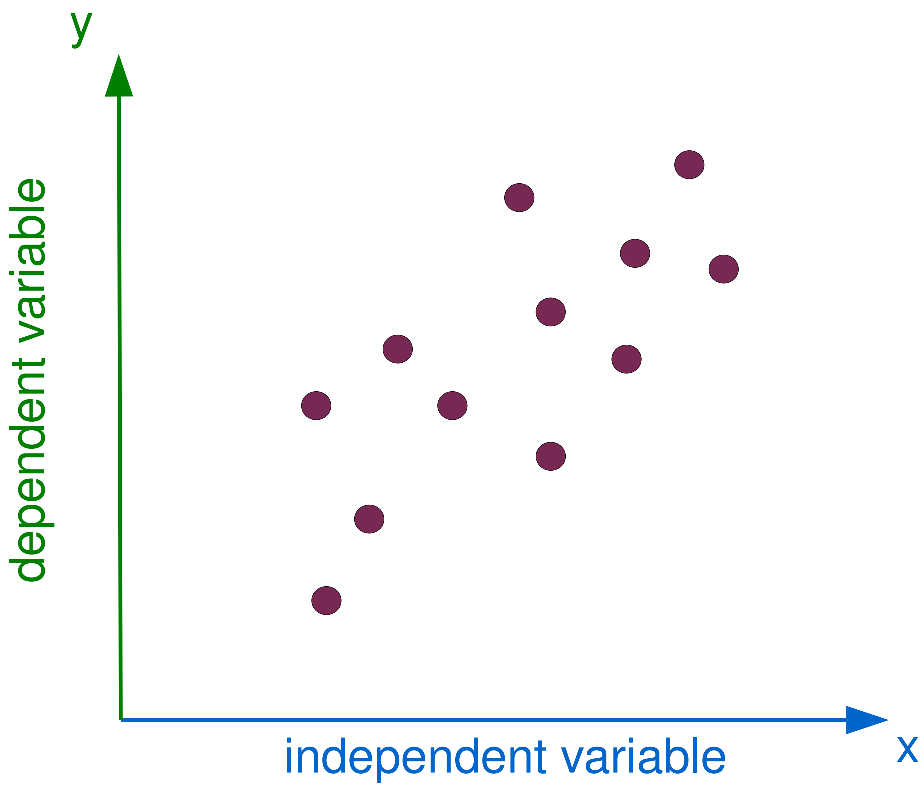
	Specification 1		*Specification 2*		*Specification 3*	
Variable	*Estimated coefficient*	*Standard error*	*Estimated coefficient*	*Standard error*	*Estimated coefficient*	*Standard error*
Intercept	5.571	0.039	4.469	0.069	4.734	0.068
Schooling	0.0521	0.0029	0.0932	0.0036	0.0740	0.0035
Experience			0.0898	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.

IZA World of Labor

What is regression analysis ?

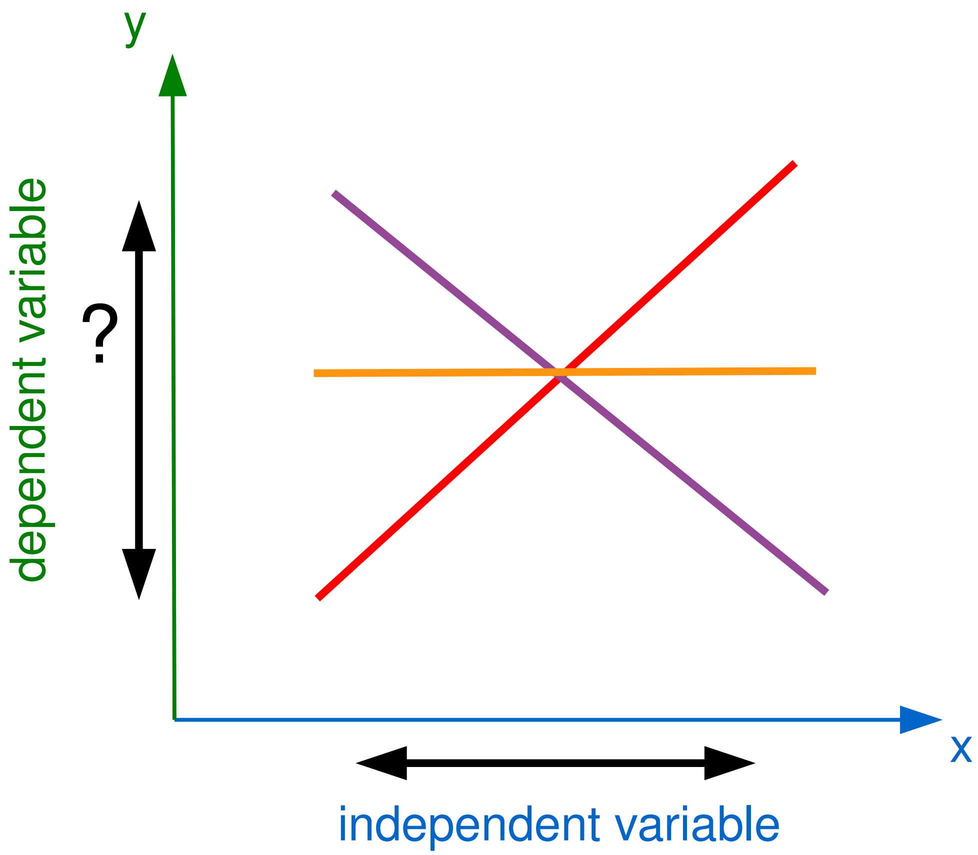


Dependent Variable: This is the main factor that we are trying to understand or predict.

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.

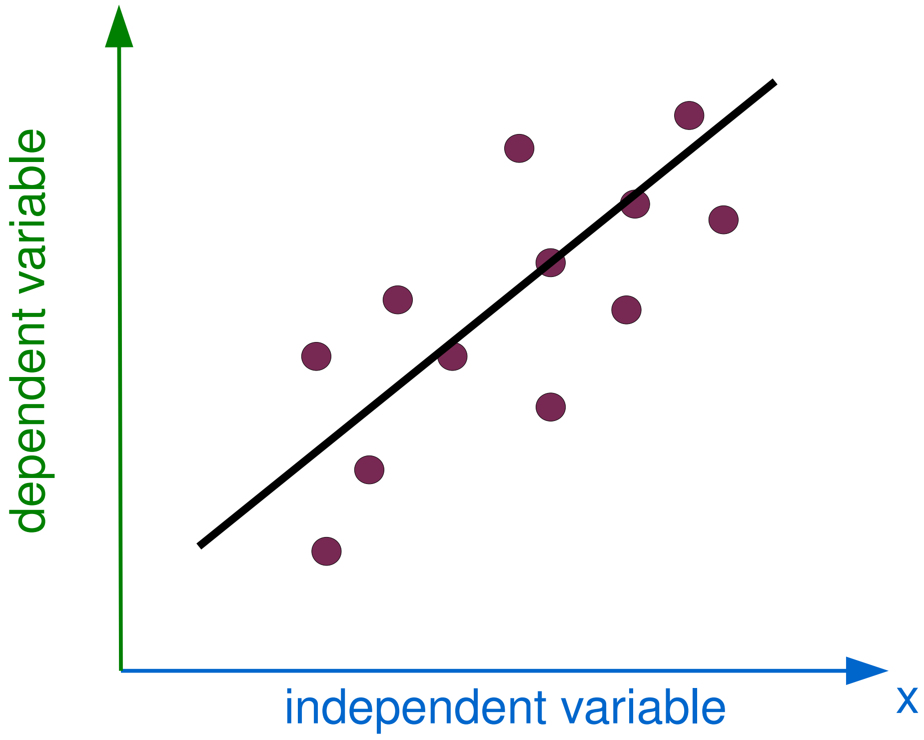
Aim of regression analysis : predict change



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

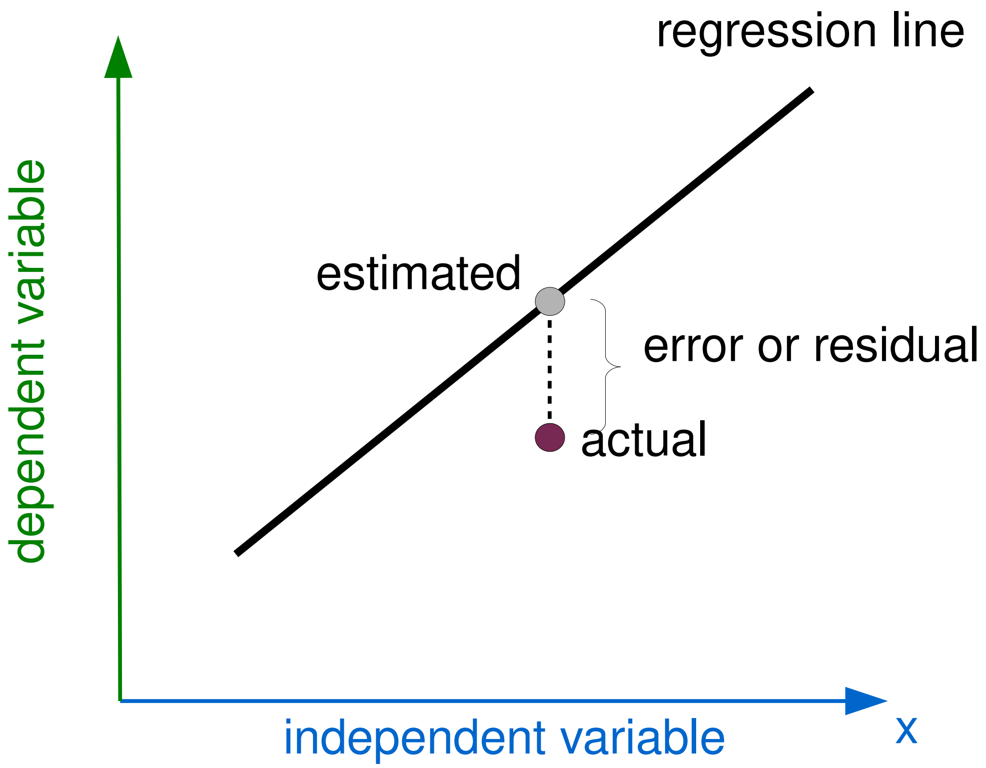


Linear regression finds the best fit line to a cloud of points were we are trying to predict one variable of interest from the know value of another variable.

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

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

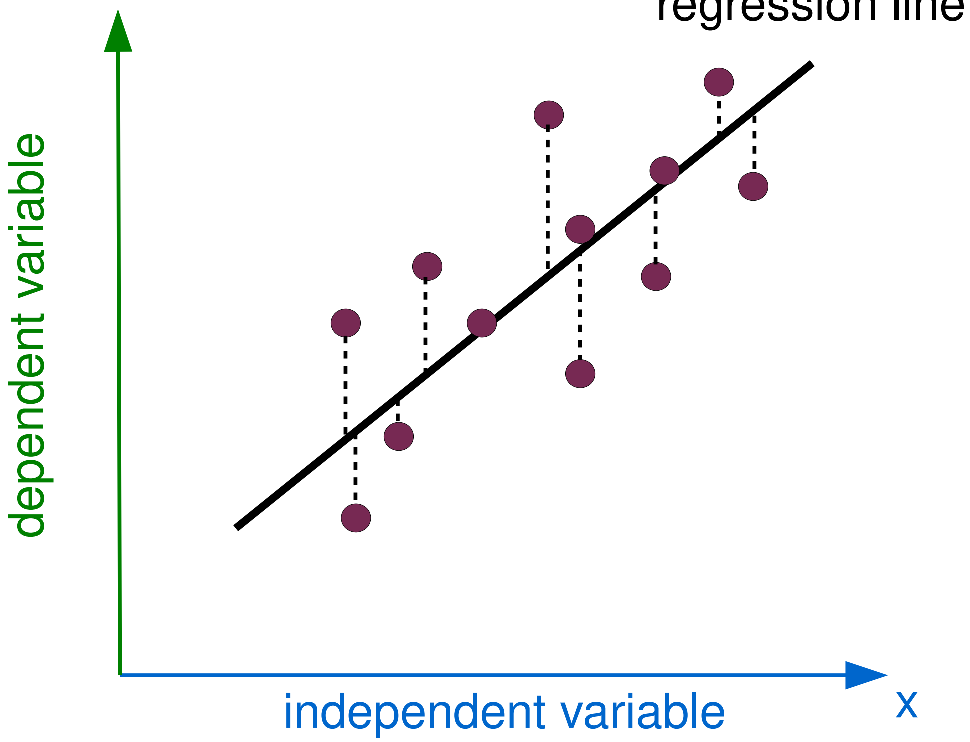
Linear regression – fit a line to the observations



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

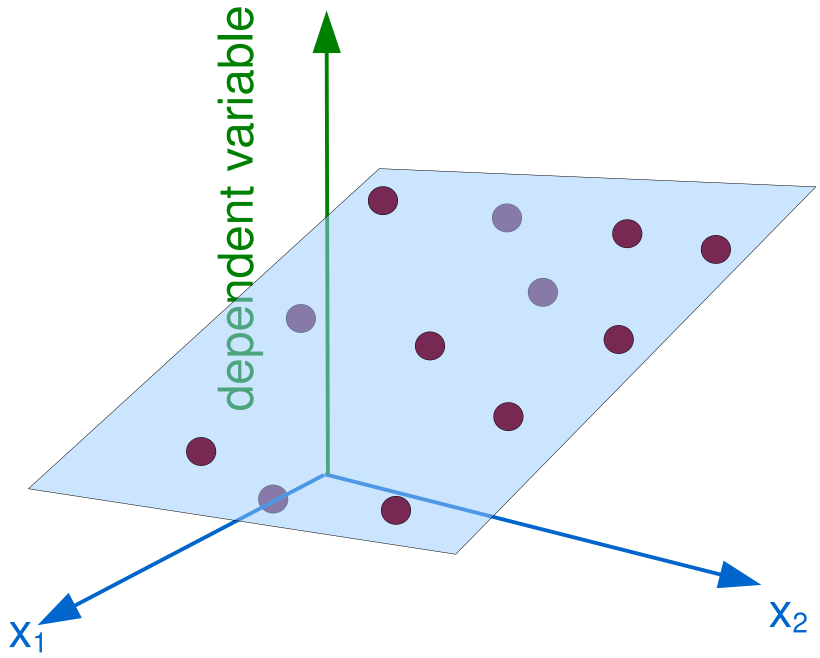


Regression line is found by minimizing the difference between the estimated and the actual value.

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

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

Multiple linear regression

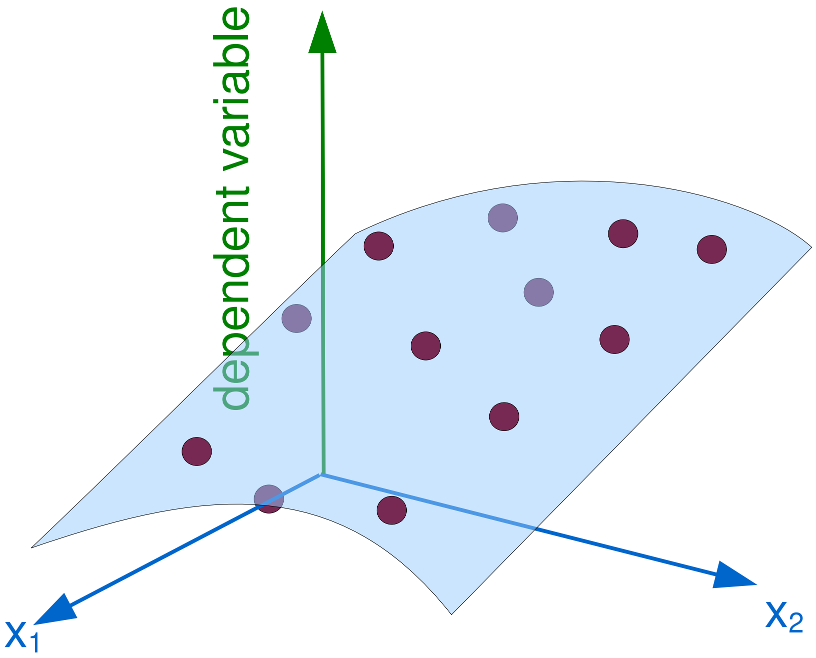


Regression with multiple predictors :
→ multiple linear regression

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

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

Multiple linear regression with interaction term

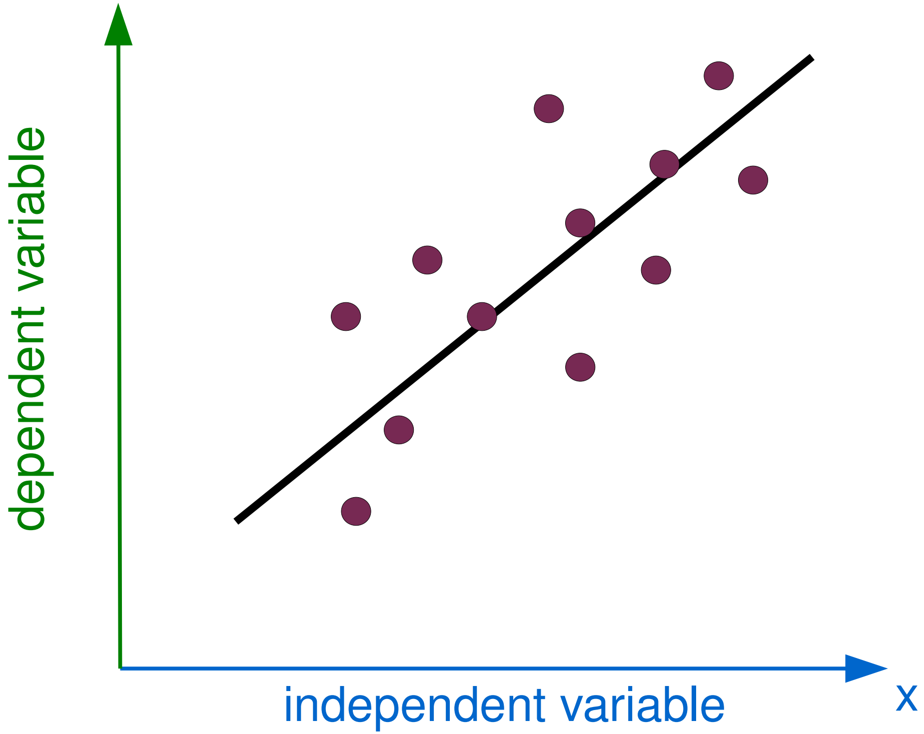


If independent variables are not independent of each other :
→ interaction term

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

- 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.

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

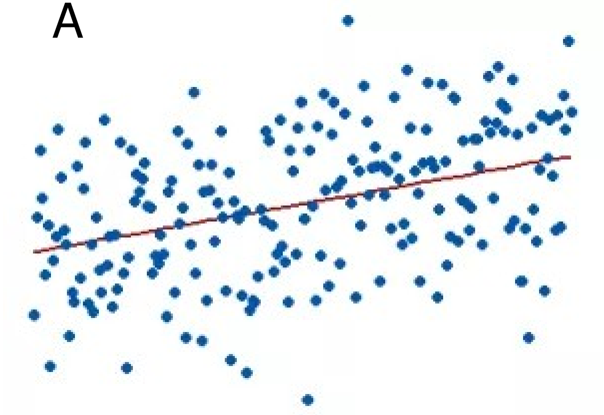
- $R^2 = 1$: we can perfectly predict all values in the data (suspicious)
- $R^2 = 0$: model fails to predict any of the variability in the data

Linear regression : R squared value

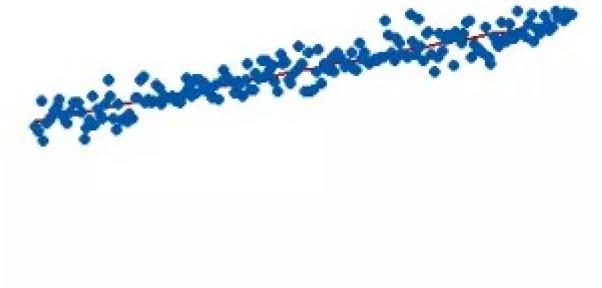
$R_A^2 = 15\,\%$

$R_B^2 = 85\,\%$

A



B



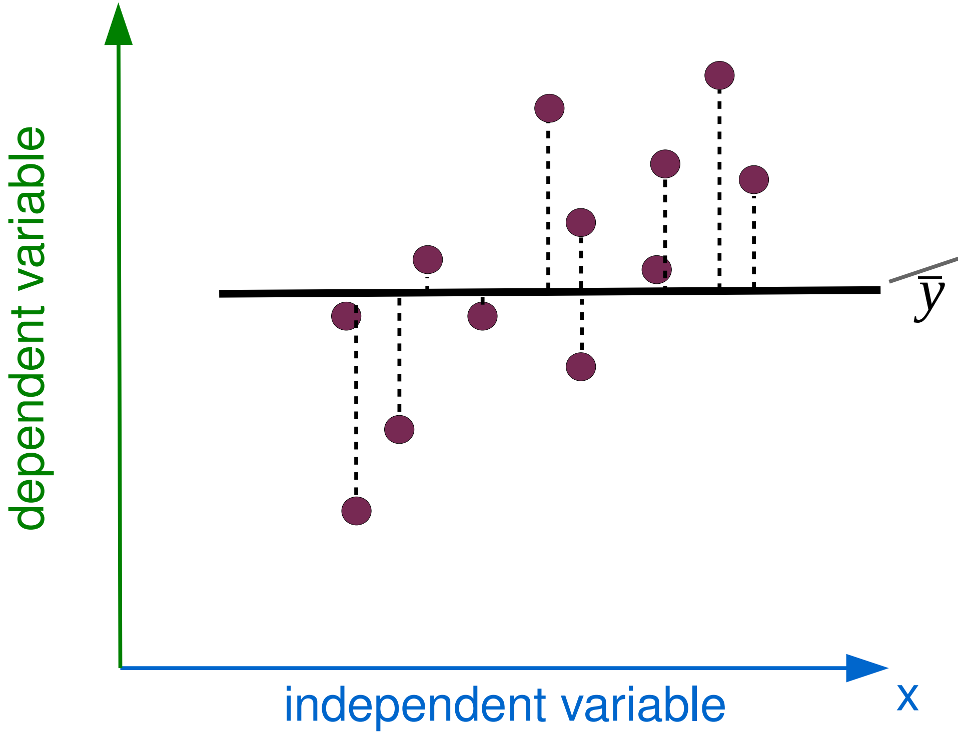
When a regression model accounts for more of the variance, the data points are closer to the regression line.

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

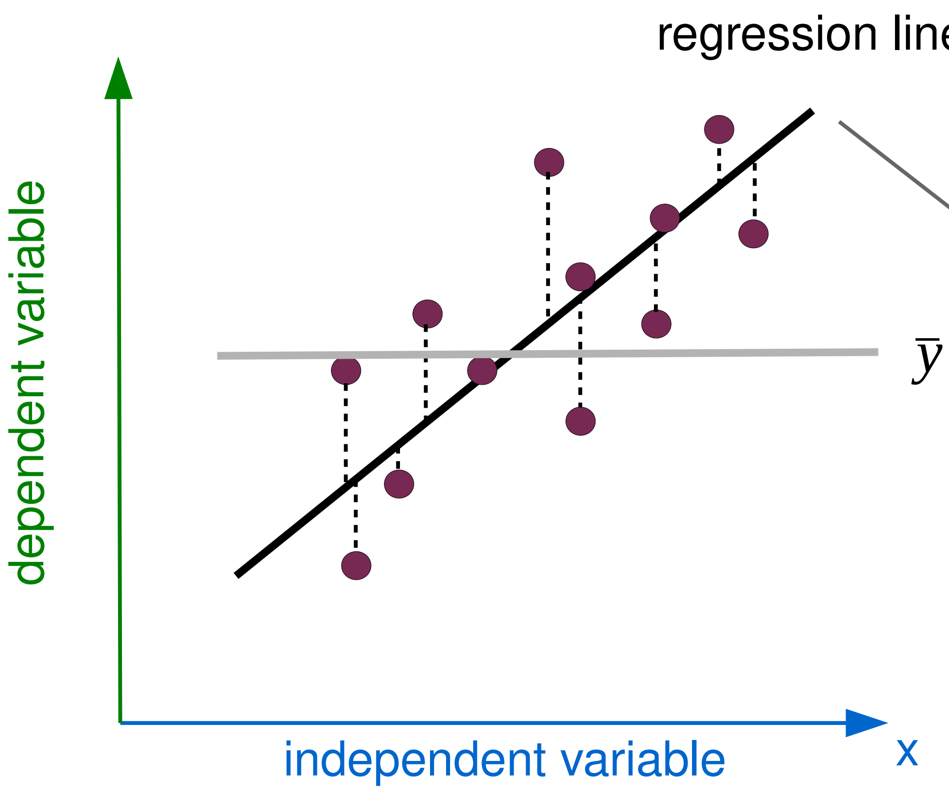
Linear regression : calculate R squared value

- 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):

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

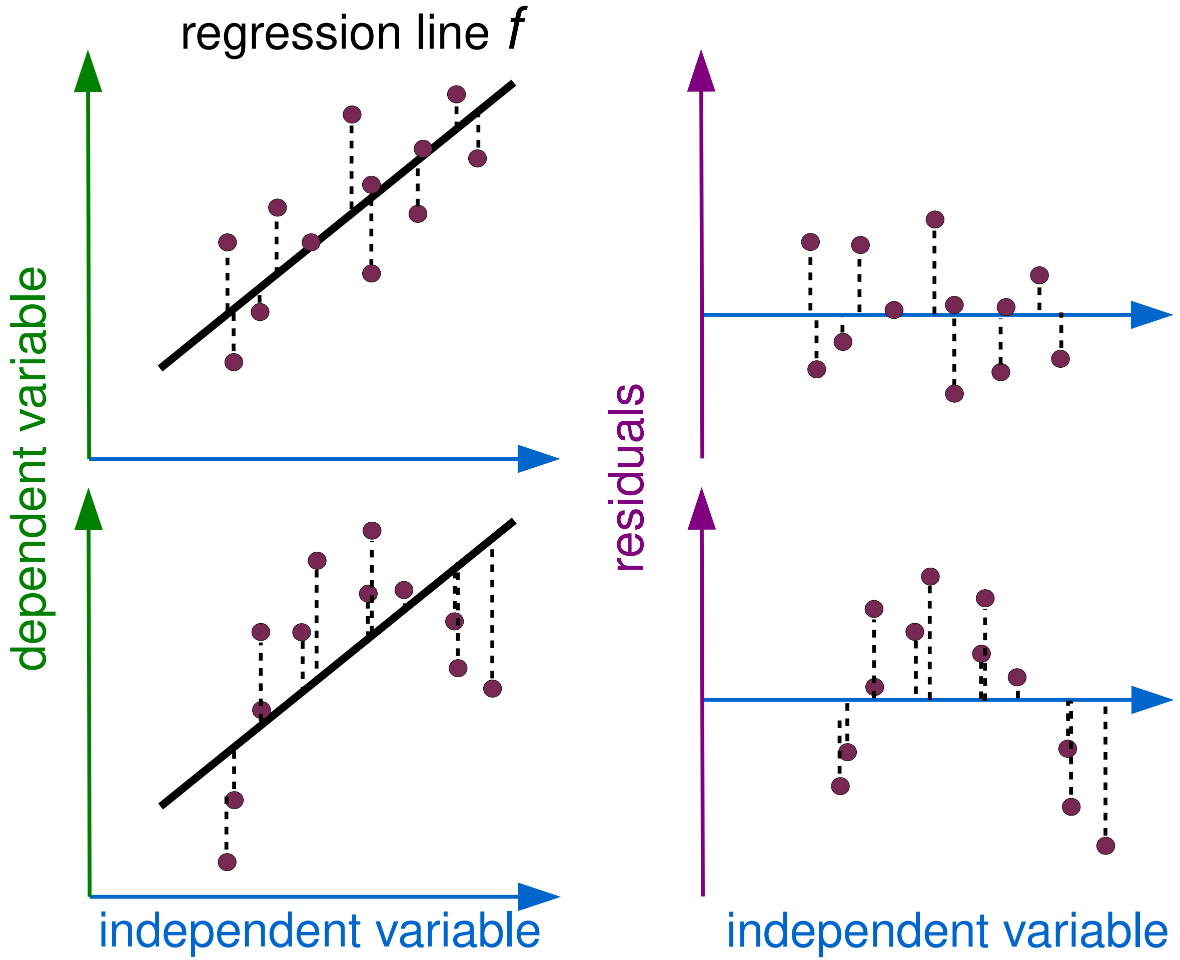
- The sum of squares of residuals, also called the residual sum of squares:

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

- general definition of R^2

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

Linear regression : residuals

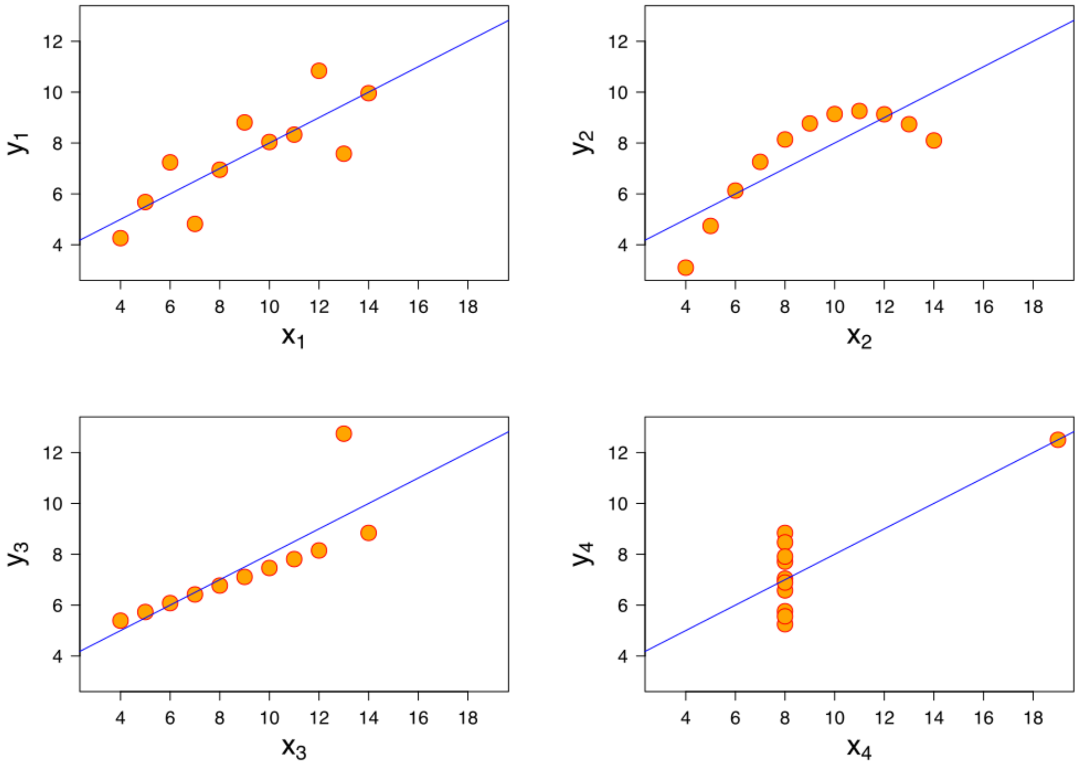


Residuals are the differences between the observations and the regression line (the function *f*)

$$residuals = (y_i - f_i)$$

- 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

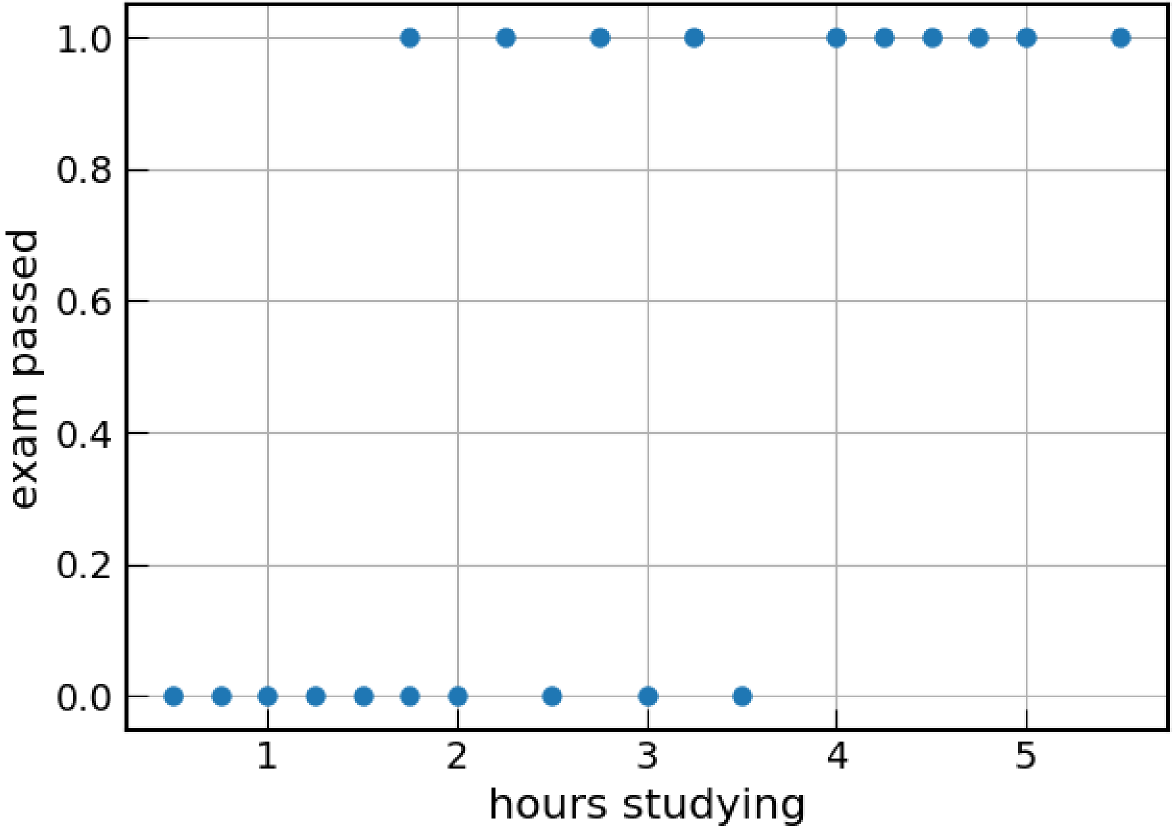
Pitfalls of linear regression



- 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

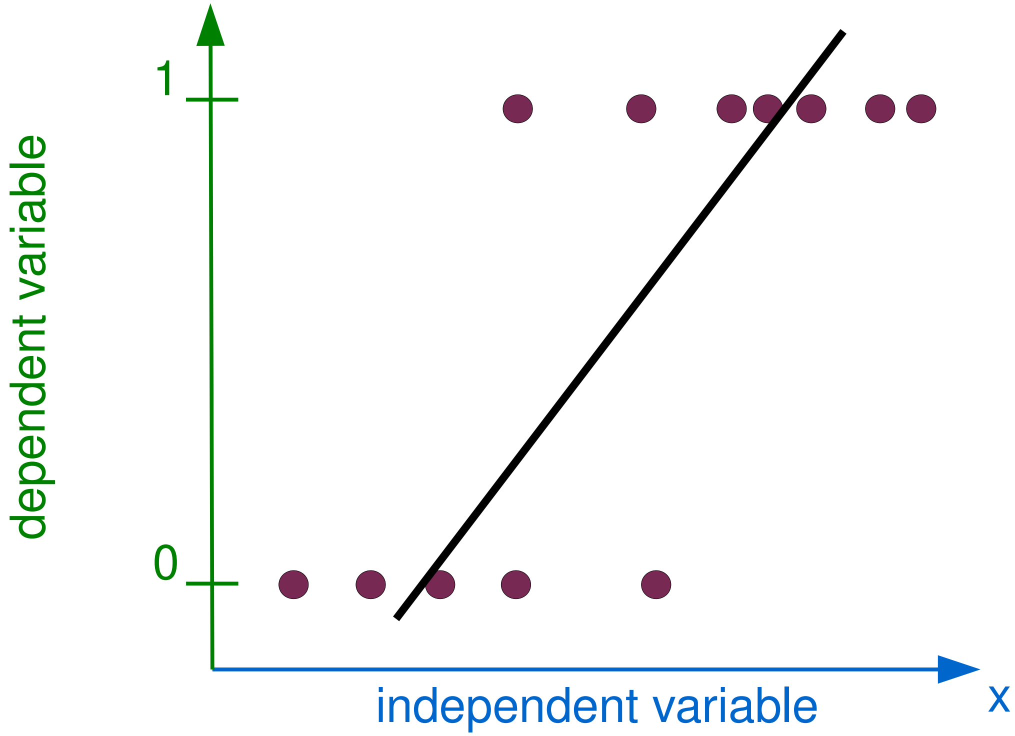


Example

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?

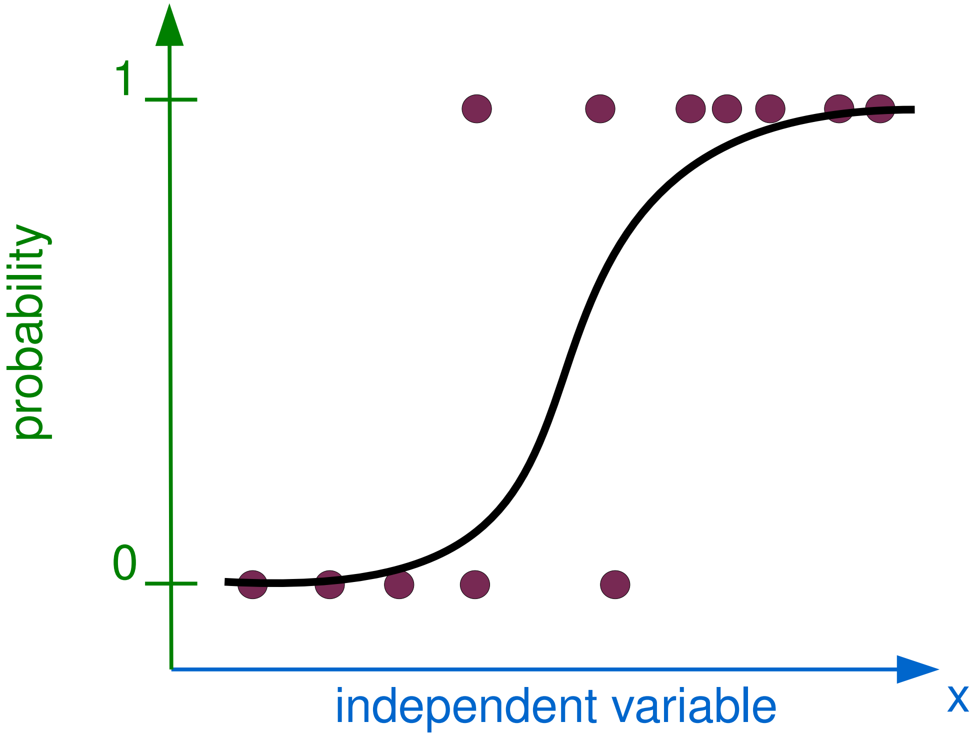
Regression with binary outcomes



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 ... .

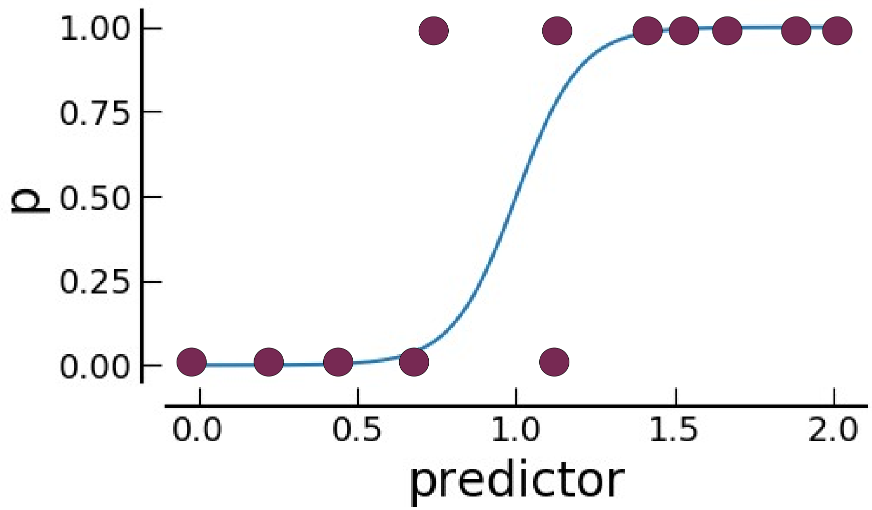
So far, we considered cases where the dependent variable is continuous, for which case we used linear regression.

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).

Logistic regression

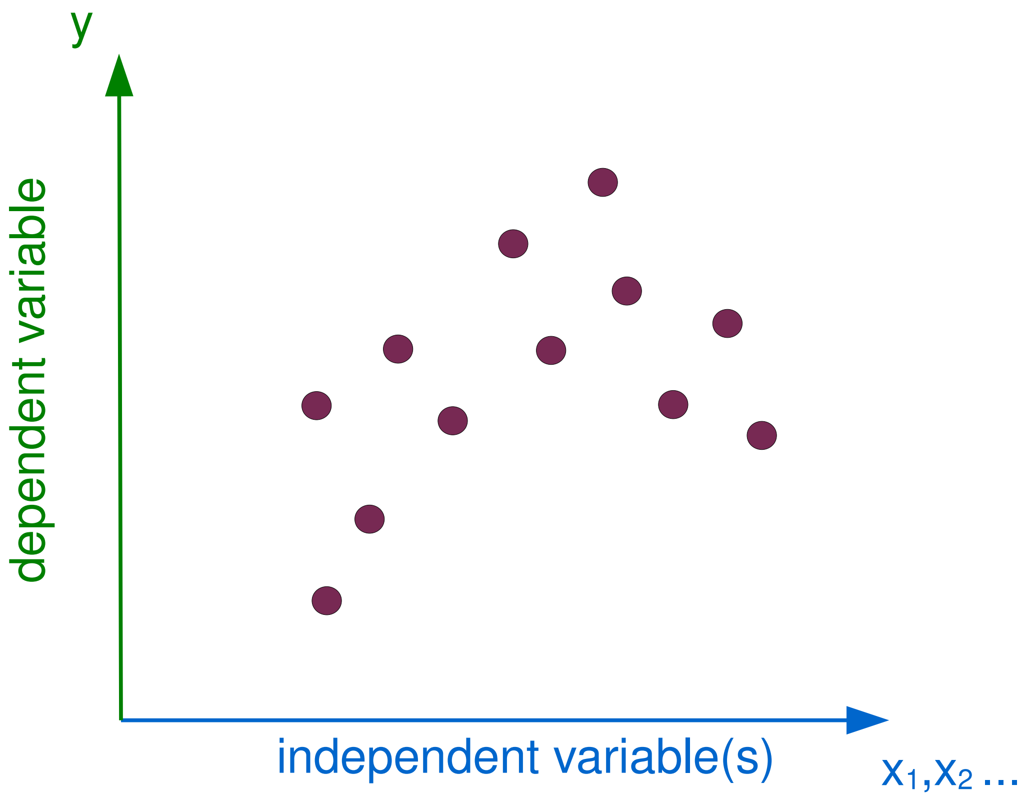


$$p = \frac{e^{\beta_0 + \beta_1 X_1}}{\left(1 + e^{\beta_0 + \beta_1 X_1}\right)}$$

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

→ provides probability given the predictor variable

Assumption : General relationship between X and Y



Regression Analysis : Approaches to examine the relationship between the variables, *i.e.*, to estimate f.

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

X ... independent variables or predictors

Y ... dependent or response variable

f ... fixed but unknown function of $X_1, \ldots X_p$; represents the systematic information that X provides about Y

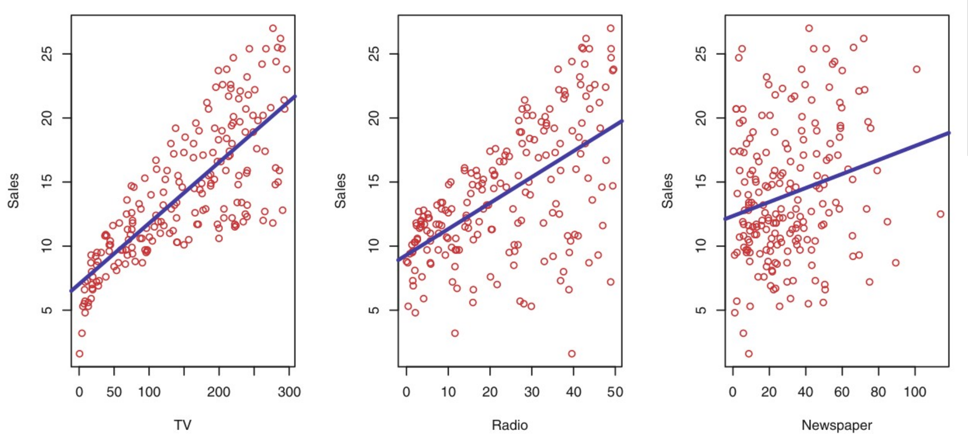
ε ... additive error term

Why estimate *f* ? : prediction and inference

prediction	**inference**
X is readily available; output *Y* cannot easily be obtained : requires to predict *Y*	want to understand relationship between X an Y (how *Y* changes of function of *X*); not necessarily to make predictions
exact form of *f* is not of interest; provided it yields accurate prediction of *Y*	*f* cannot be treated as black box, we need to know exact form : - Which predictors are associated with the response? - What is the relationship between each response and the predictors? - Can the relationship btw. Y and each predictor adequately summarized using a linear equation?

Examples for prediction and inference

prediction	inference
$X_1, \ldots, X_p$ are characteristics of a patient's blood sample that can be easily measured in a lab, and Y is a variable encoding the patient's risk for a severe adverse reaction to a particular drug. Prediction stock price in the future	Advertising data set consists of the sales of a product in different markets, along with advertising budgets for the product for three different media: **TV , radio, and newspaper.** - Which media contribute to sales? - Which media generate the biggest boost in sales?



Prediction : reducible and irreducible error

$$Y = f(X) + \epsilon \xrightarrow{\text{estimation}} \hat{Y} = \hat{f}(X)$$

$\hat{f}$ ... represents the estimate of f

$\hat{Y}$ ... represents the estimate of Y

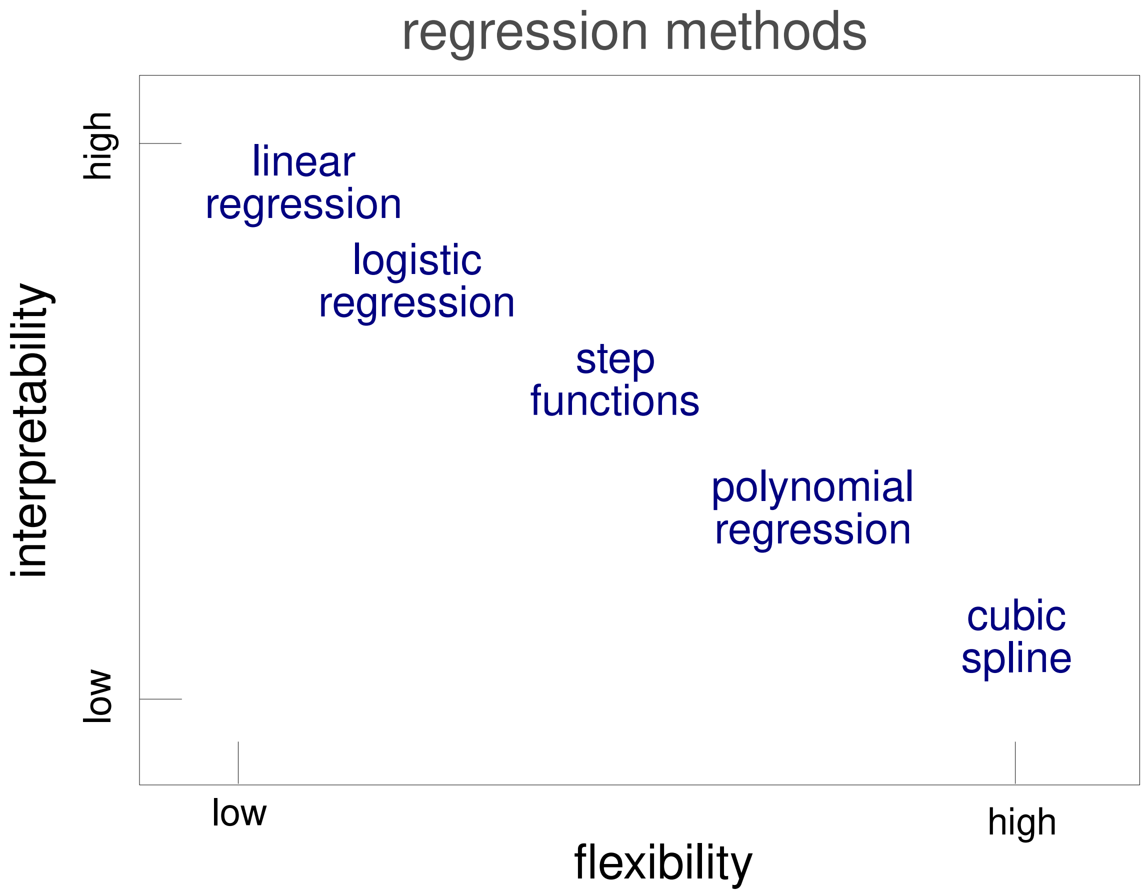
- accuracy of $\hat{Y}$ depends on the *reducible* and the *irreducible* error :
  - **reducible error :** $\hat{f}$ will not be the perfect estimate of f; finding the most appropriate estimate for f improves the accuracy of $\hat{f}$
  - **irreducible error :** even for the perfect estimate of f, we still have the error ϵ ; this error cannot be predicted by using X (e.g. unmeasured variables, unmeasurable variations)

Which method of estimating f ?

Depending on goal – prediction, inference or combination of both – different methods for estimating f might be appropriate

- **linear models :** allow simple and interpretable inference; but may not yield accurate predictions
- **highly non-linear approaches :** can provide accurate predictions of Y ; less interpretable model for which inference is challenging

Trade-off: prediction accuracy vs model intepretability



- less flexible : produce small range of shapes to estimate *f* (e.g. linear regression → lines,planes)
- more flexible : can generate much wider range of shapes to estimate *f* (e.g. splines)

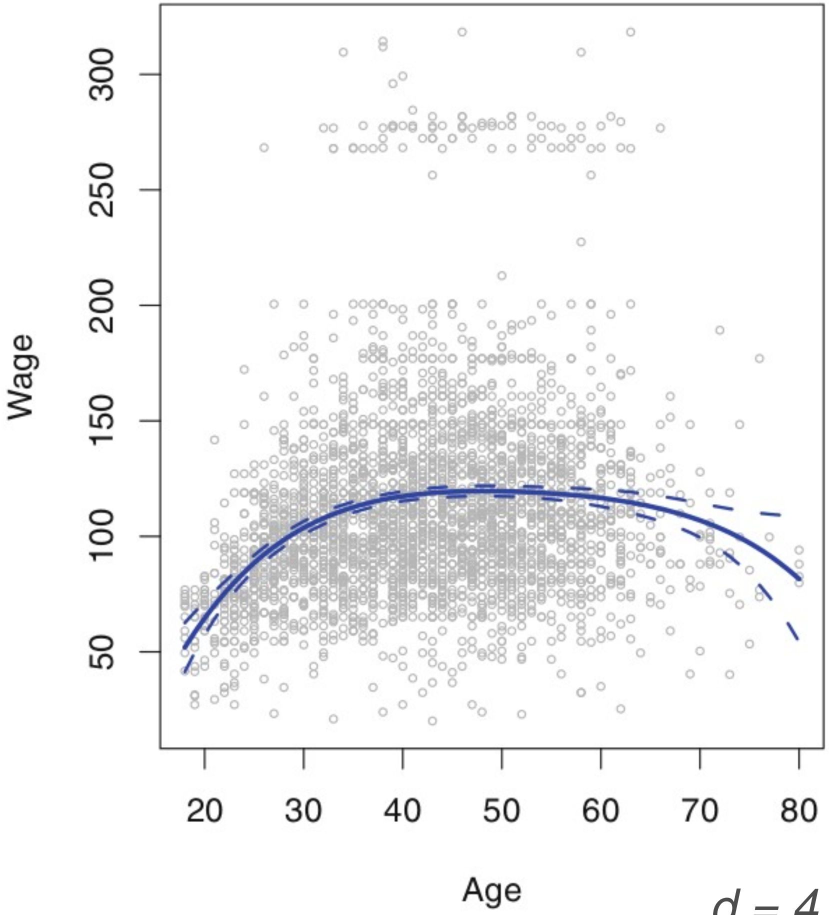
Regression methods : beyond linearity

- *Polynomial regression* extends the linear model by adding extra predictors, obtained by raising each of the original predictors to a power.

- *Step functions* cut the range of a variable into K distinct regions in order to produce a qualitative variable. This has the effect of fitting a piecewise constant function.

- *Regression splines* are more flexible than polynomials and step functions, and in fact are an extension of the two. They involve dividing the range of X into K distinct regions. Within each region, a polynomial function is fit to the data. However, these polynomials are constrained so that they join smoothly at the region boundaries, or knots. Provided that the interval is divided into enough regions, this can produce an extremely flexible fit.

Polynomial Regression



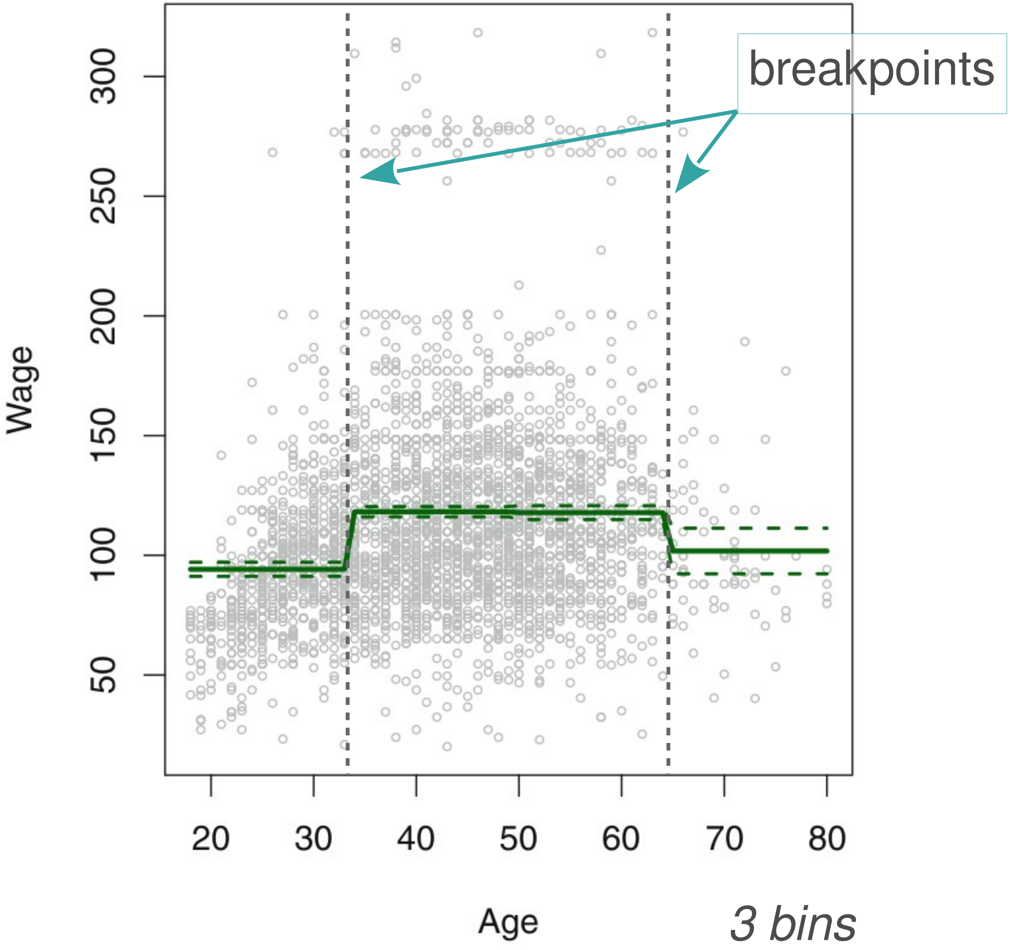
- replaces the linear model with a polynomial function

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots \beta_d X^d + \epsilon$$

- degree *d* controls the non-linearity of the curve
- unusual to use d greater than 3 or 4 : curve becomes very flexible for d>4 and take strange shapes

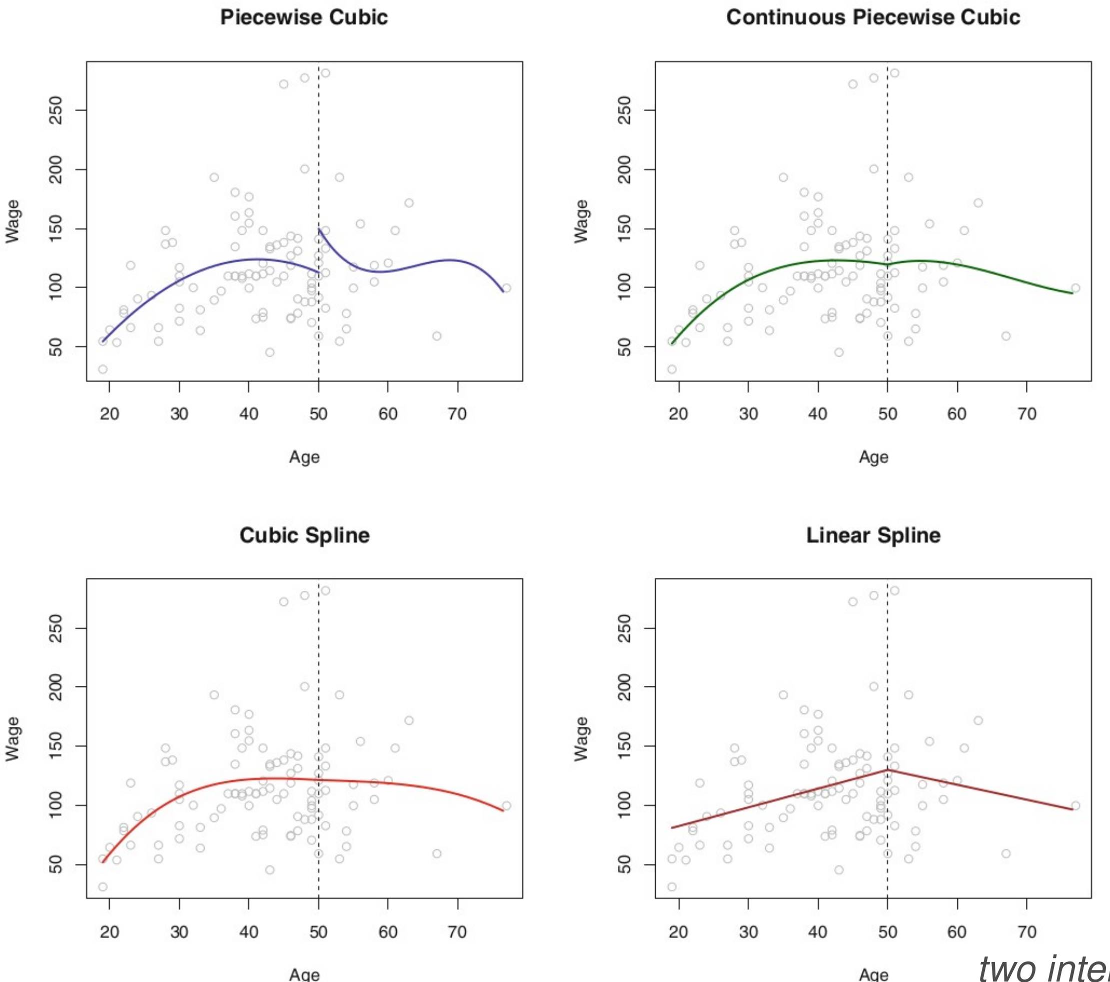
d = 4 (dashed curve 95 % confidence interval)

Step Functions



- break the range of X into bins and fit different constants in each bin
- breakpoints have to be defined before fitting the constants (e.g. based on percentiles)
- unless there are natural breakpoints, piecewise-constant functions can miss the action
- popular in biostatistics and epidemiology

Regression splines

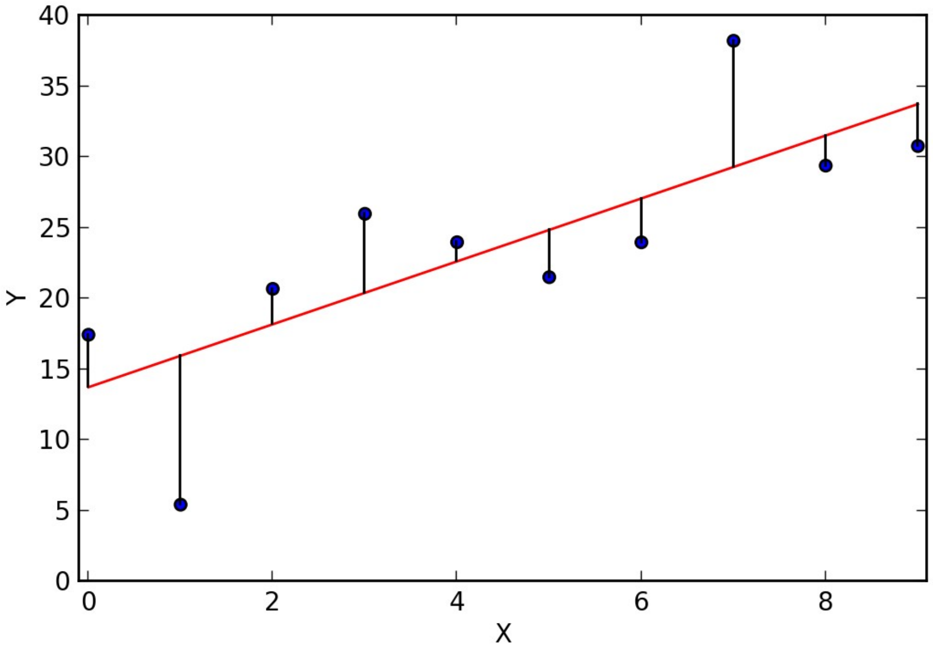


- extends polynomial regression and step function : fitting separate low-degree polynomials over different regions of X
- *additional constraints* are that fitted curve must be *continuous* and *smooth*
- *general definition* : degree-d spline is a piecewise degree-d polynomial with continuity in derivatives up to degree d-1

two intervals, or 1 knot

Measuring quality of fit : mean squared error (MSE)

Example : MSE for linear regression

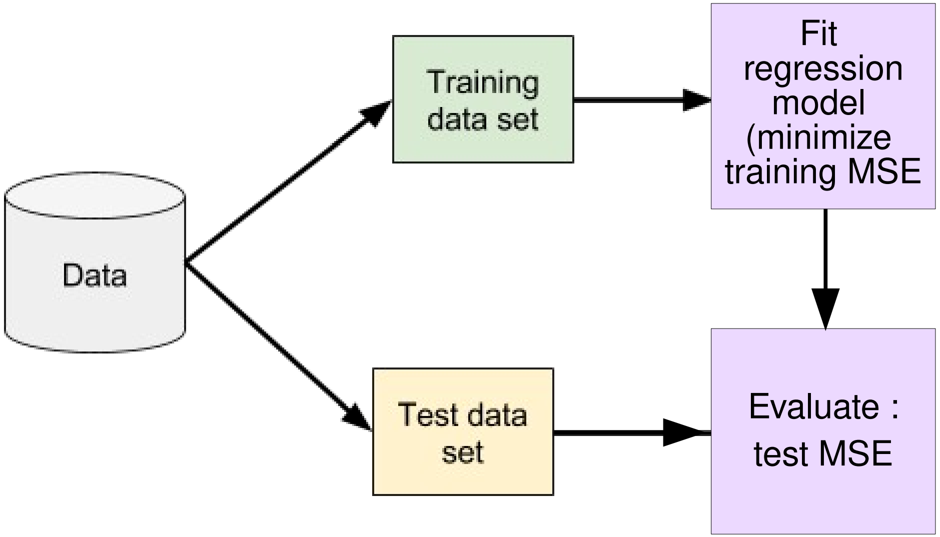


- quantifies the extent to which predicted response value is close to the true response

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

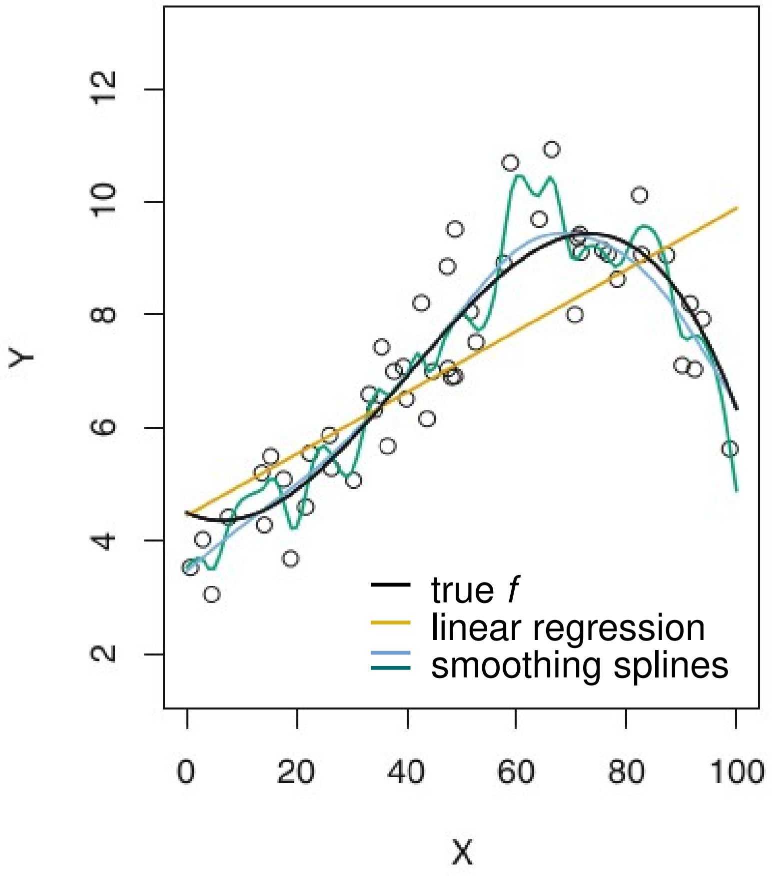
$\hat{f}(x_i)$ ... prediction that $\hat{f}$ gives for the ith observation

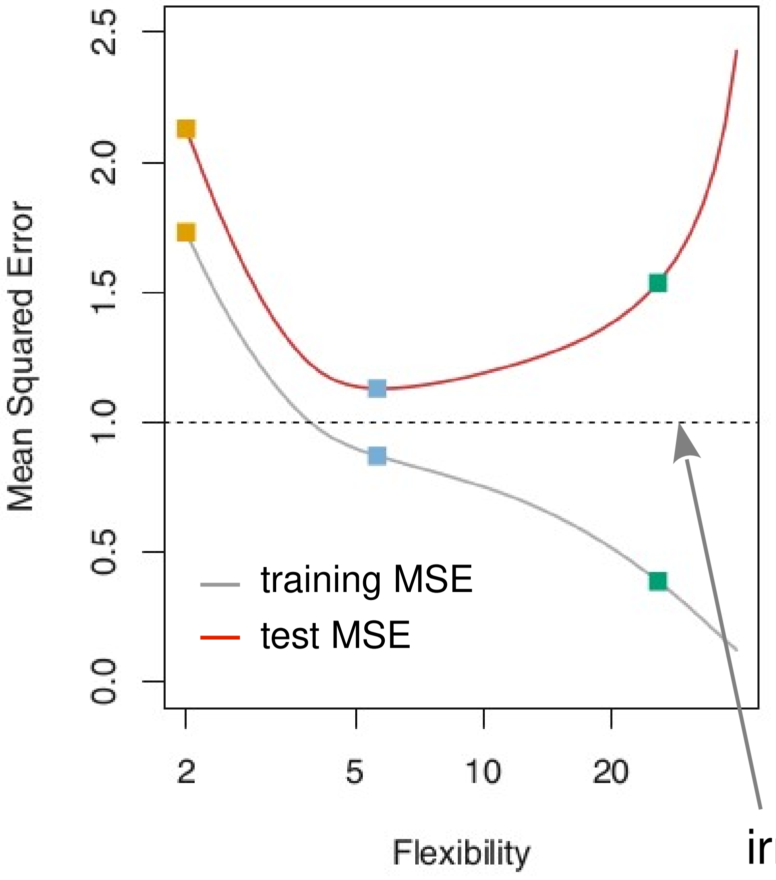
Training vs. Test mean-squared error (MSE)



- MSE computed using the training data is used to fit the model : *training MSE*
- we are interested in the accuracy of the prediction when model is applied to previously unseen test data
- want to chose the model that gives lowest *test MSE* , i.e., the MSE calculated on the previously unseen test data (as opposed to the model with lowest training MSE)
- **Attention :** model with the lowest training MSE is not necessarily the model with the lowest test MSE

Overfitting : small training MSE & large test MSE





overfitting data : small training MSE and large test MSE; training to specific (random) pattern in training data which does not reflect true property of f

irreducible error ϵ