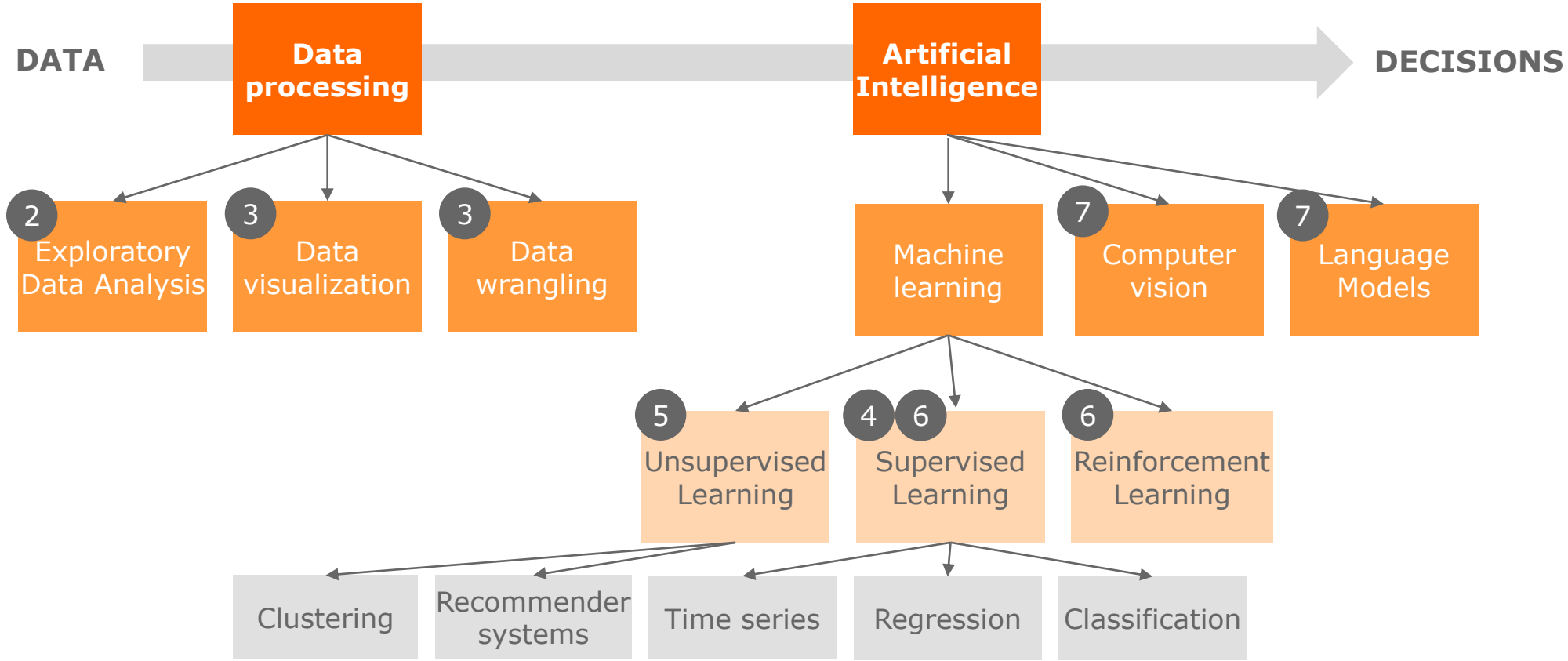




Supervised Learning: Regression, Classification & Time Series

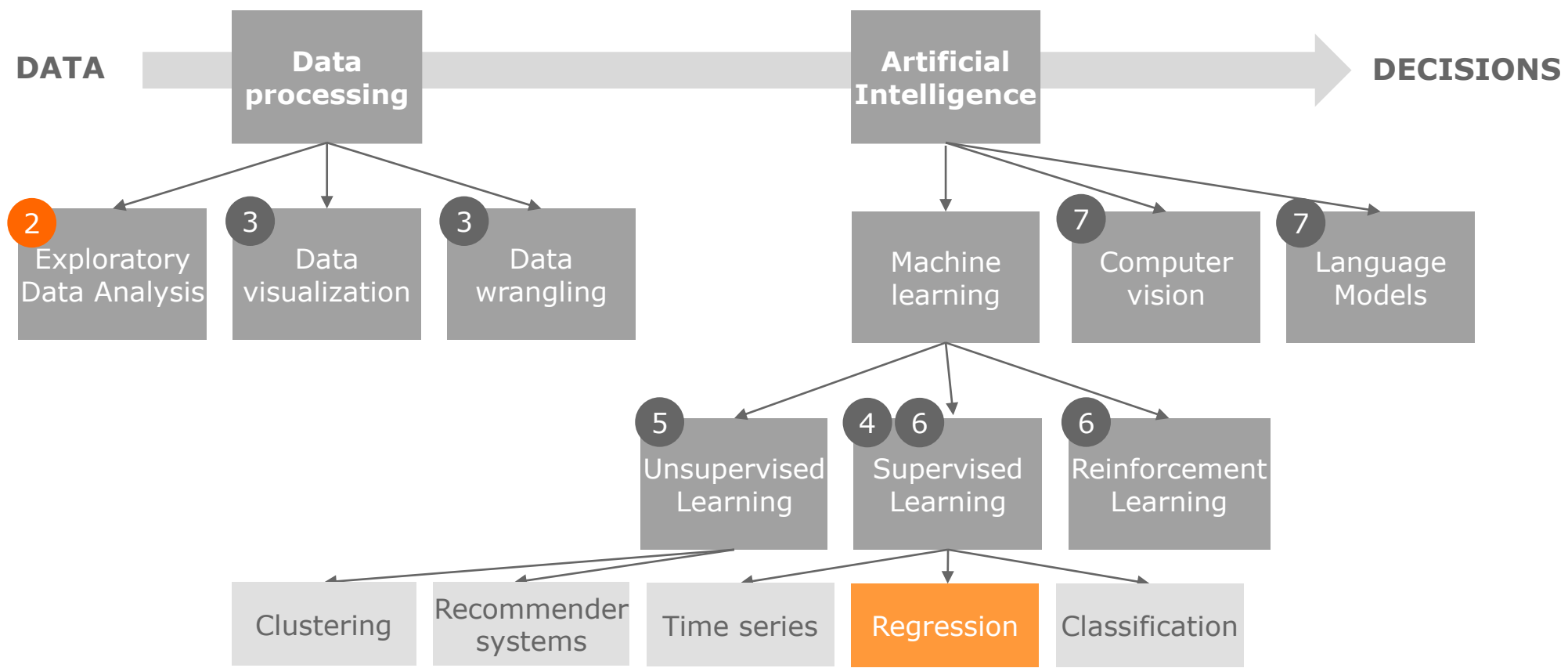
Session 2

Bernardo Almada-Lobo, Daniel Pereira,
Pedro Amorim
Lisboa, 9 a 12 de outubro 2023



Fundamentals

012



Google Cloud

8

Linear Regression

Multiple Linear
Regression

Conditions for Regression
validity

Variable selection

Logistic Regression

Linear Mixed Models



In praise of regression models¹

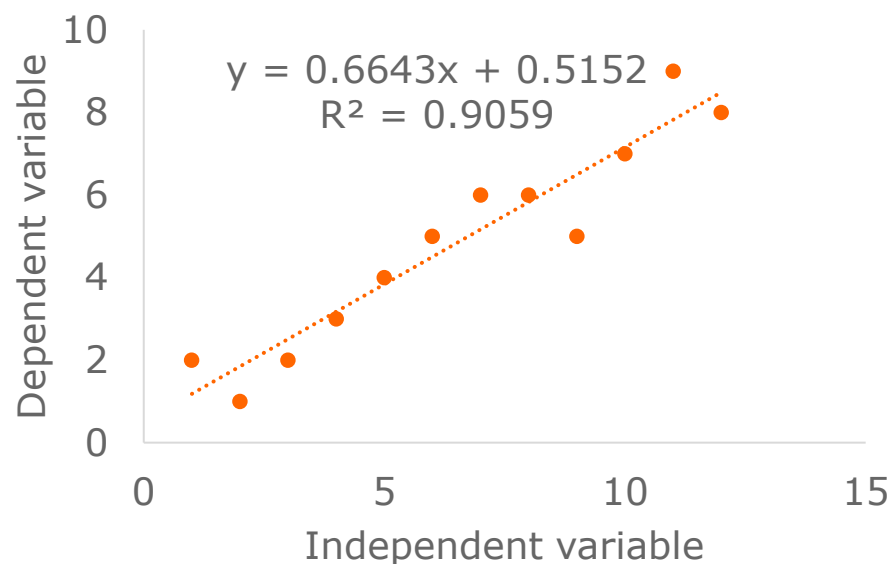
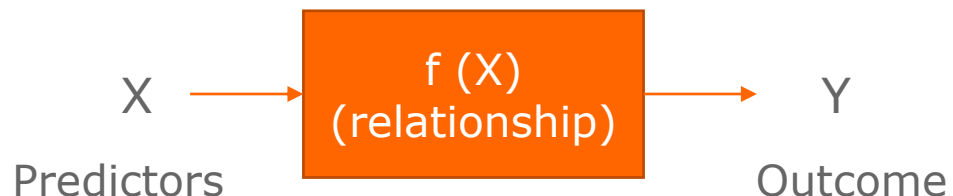
Regression analysis is versatile and has wide applicability

- Regression analysis allows you to understand the strength of relationships between variables
- Regression analysis tells you what predictors in a model are statistically significant and which are not
- Regression analysis can give a confidence interval for each regression coefficient that it estimates

Regression analysis is less of a black box and is easier to communicate

**Regression analysis is a subset of the statistic inference field of study.
Understanding regression techniques will give you a better understanding of statistics overall**

Linear models are among the most used predictive models



- In some circumstances, data can be valuable in helping to determine the parameters in a relationship or its structural form
- The process of using data to formulate relationships is known as **regression analysis**
- In this approach, we identify one variable as the **response variable**, which means that it can be predicted from the values of other variables
- Those other variables are called **explanatory variables**

Simple Linear Regression

$$y = a + bx + e$$

y - dependent variable

x - independent variable

Constants a and b represent the intercept and slope, respectively, of the regression line

e - an "error" term.

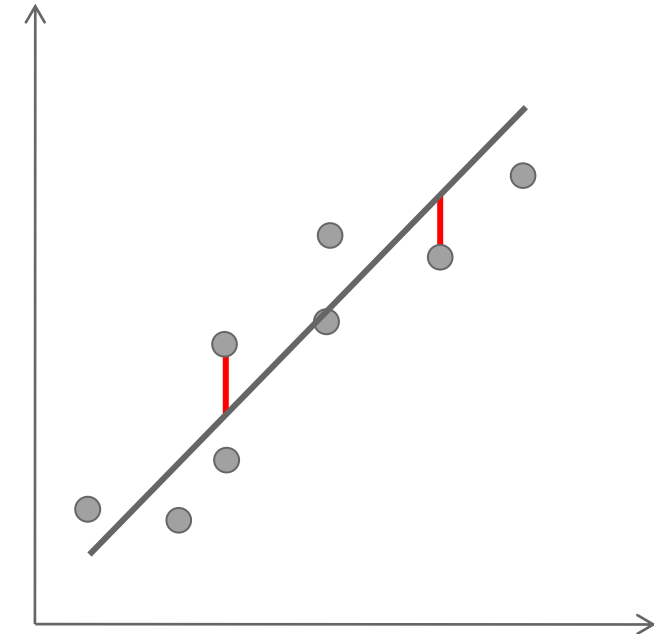
Unexplained "noise" in the relationship

May represent limitations of knowledge

Or may represent random deviations of the dependent variable from its mean, y

The goal is to minimize the estimation error

- Want to find line to most closely match the observed relationship between x and y
- Define “most closely” as minimizing sum of squared differences between observed and model values
 - ✓ Minimizing sum of differences would set y equal to its mean
 - ✓ Penalizes large differences more than small differences



Performing Regression

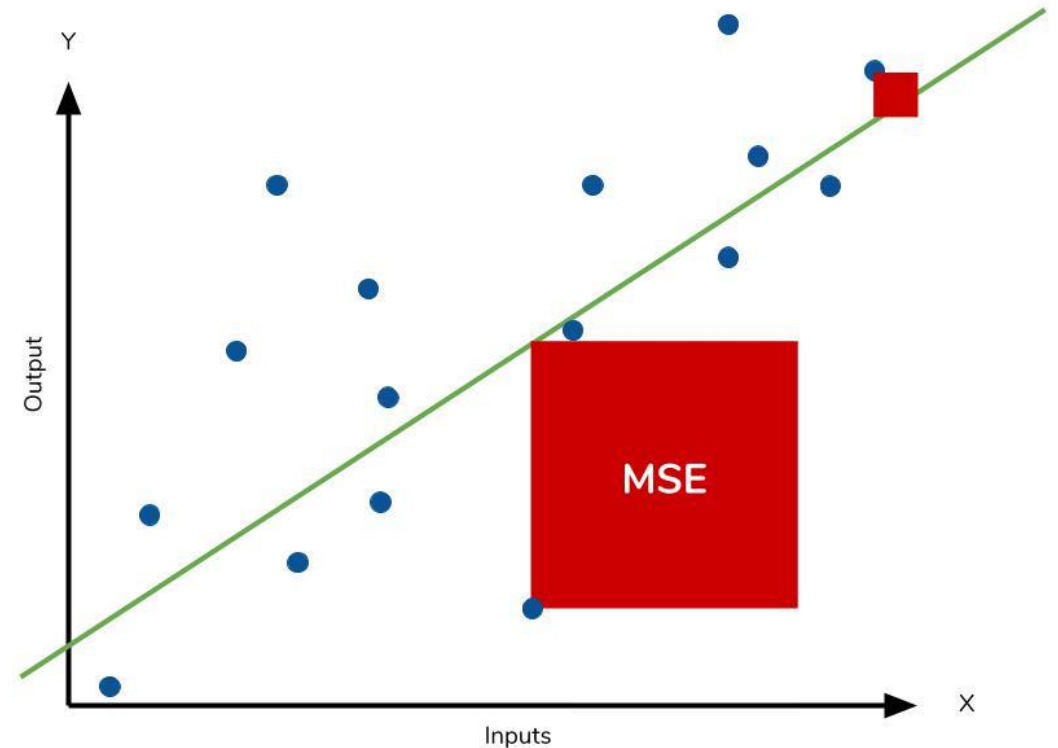
Residuals:

$$e_i = y_i - \hat{y} = y_i - (a + bx_i)$$

Sum of squared differences between observations and model:

$$SS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

The regression problem: choose a and b to minimize SS



Assumes residuals are normally distributed with mean zero
Regression parameters can be calculated directly from the data

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad a = \bar{y} - b\bar{x}$$

Simpler to use Excel's regression tool
(Under Data Analysis tab)

Coefficient of determination: R^2

Lies in range $[0, 1]$

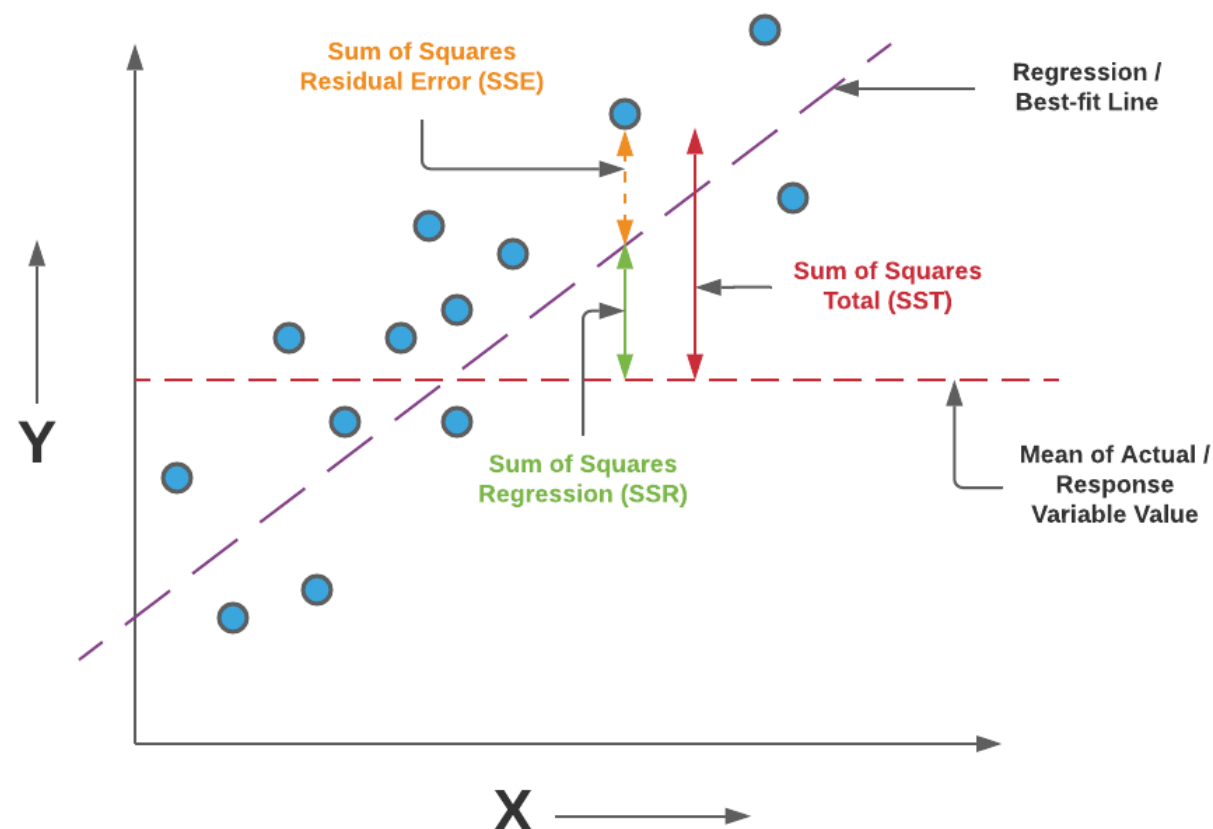
Closer to one – better fit

Measures how much of the variation in y-values is explained by model

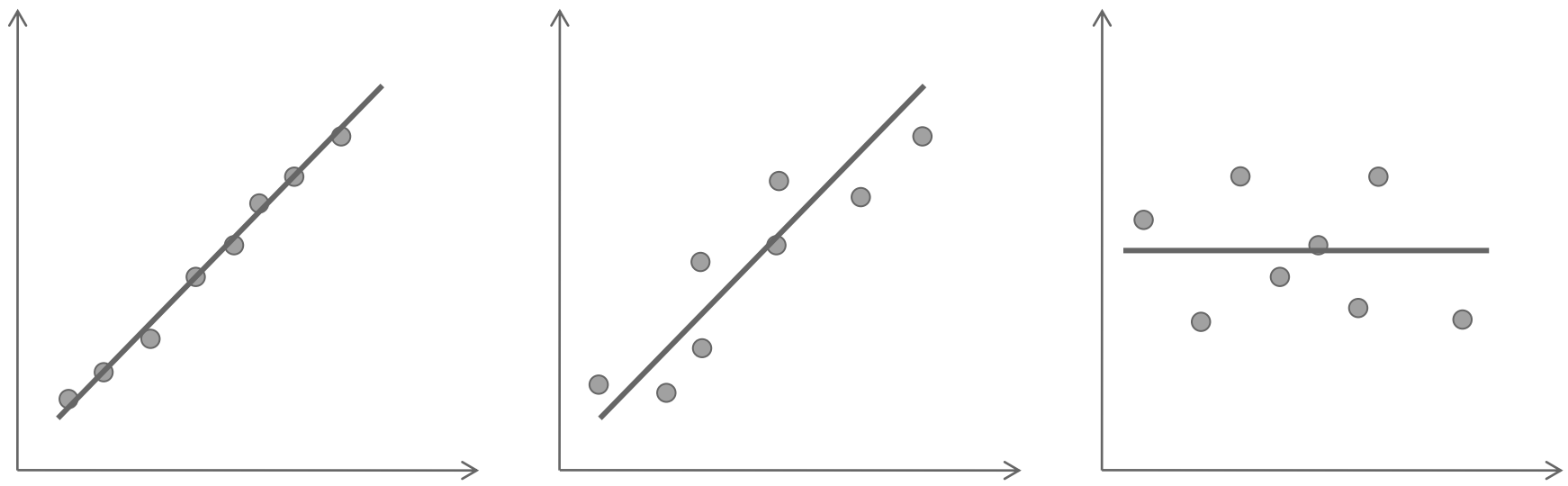
1 – perfect match to model

0 – equation explains none of observed variation

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$



Goodness of Fit



Four measures are used to judge the statistical qualities of a regression:

Regression model
as a whole

R^2 : Measures the percent of variation in the explanatory variable accounted for by the regression model.

F-statistic (Significance F): Measures the probability of observing the given R^2 (or higher) when all the true regression coefficients are zero.

Individual
coefficients

p-value: Measures the probability of observing the given estimate of the regression coefficient (or a larger value, positive or negative) when the true coefficient is zero.

Confidence interval: Gives a range within which the true regression coefficient lies with given probability.

Regression Output

Estimate for a

Estimate for b 's

OLS Regression Results						
Dep. Variable:	PRODUCTS	R-squared:	0.376			
Model:	OLS	Adj. R-squared:	0.366			
Method:	Least Squares	F-statistic:	39.62			
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	1.11e-37			
Time:	23:23:16	Log-Likelihood:	-1004.8			
No. Observations:	402	AIC:	2024.			
Df Residuals:	395	BIC:	2052.			
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-1.2437	1.094	-1.137	0.256	-3.395	0.908
SEX	-0.5618	0.314	-1.789	0.074	-1.179	0.055
AGE	-0.0041	0.011	-0.357	0.722	-0.026	0.018
INCOME	-0.0837	0.021	-4.019	0.000	-0.125	-0.043
WEIGHT_KG	0.0589	0.012	5.089	0.000	0.036	0.082
PRICE	0.1034	0.012	8.446	0.000	0.079	0.127
SATISFACTION	0.9350	0.248	3.775	0.000	0.448	1.422
Omnibus:	311.782	Durbin-Watson:	1.819			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5259.574			
Skew:	3.202	Prob(JB):	0.00			
...						
Notes:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						
Output is truncated. View as a scrollable element or open in a text editor . Adjust cell output settings...						

R Squared

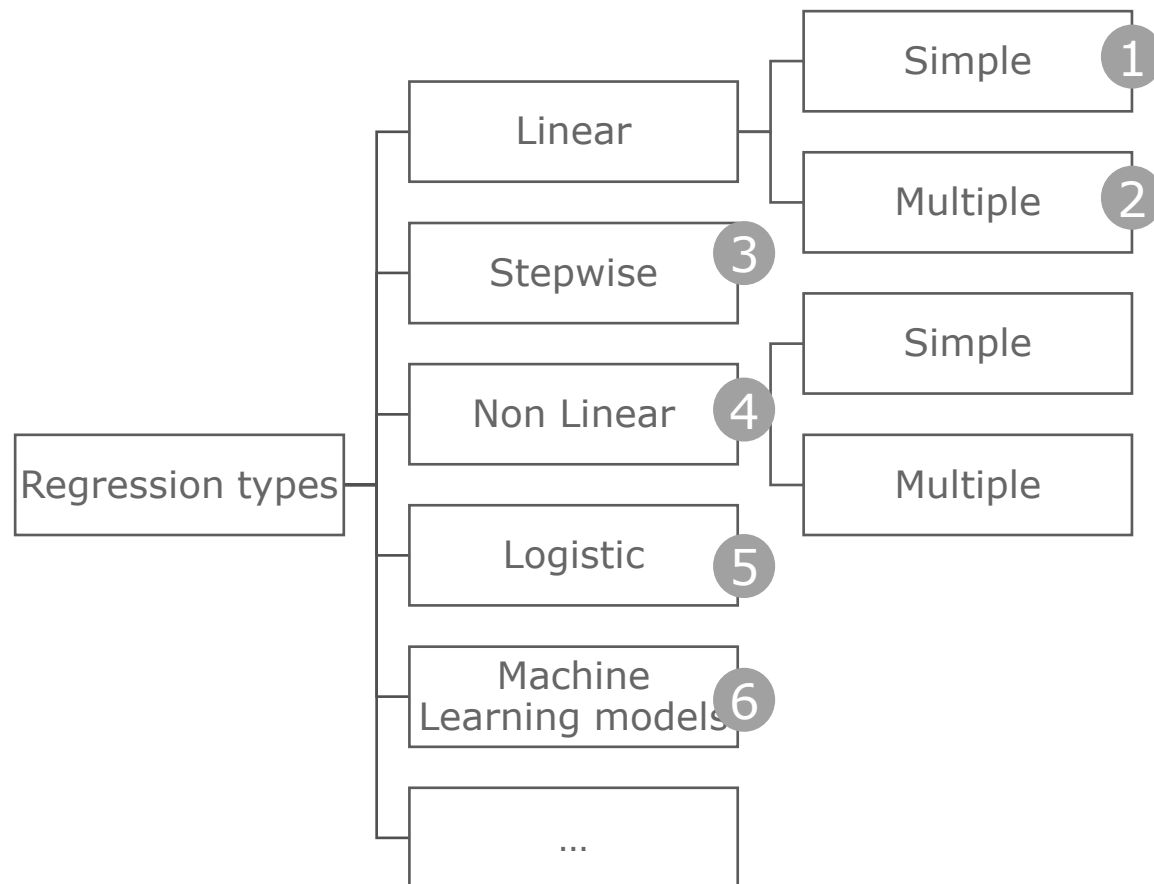
Degree of
significance
(under 0.05 is
significant)

P values of
under 0.1
are
statistically
significant

The choice of the most suitable regression type is critical for successful and robust results

Broad regression type selection

- Inspect data and evaluate the type of relationship being modeled
- Verify data for colinearity of variables
- Distinguish between **continuous and discrete variables**
- Test the **goodness of fit** with different methods (OLS, AIC, BIC, Mallow's Cp)
- Evaluate different approaches with practice and test data according to comparable error metrics
- Trade-off **performance**, ease of **implementation** and **clarity of results**



Linear Regression

**Multiple Linear
Regression**

Conditions for Regression
validity

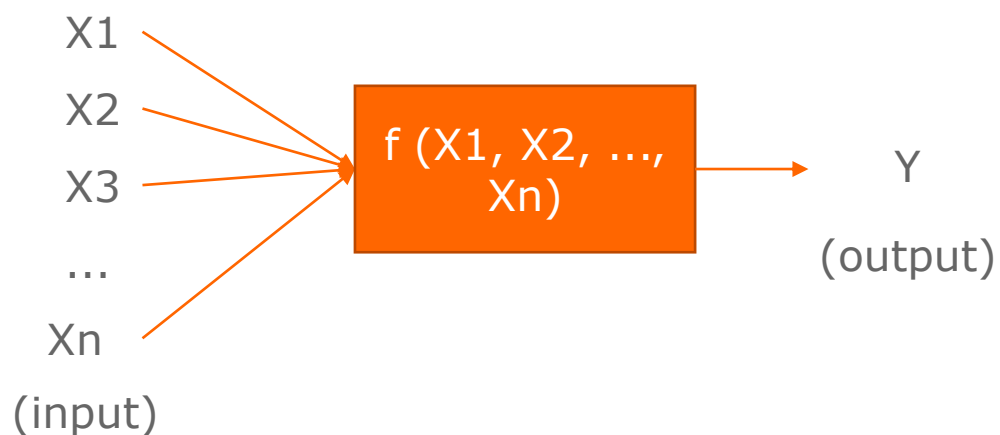
Variable selection

Logistic Regression

Linear Mixed Models



Multiple linear regression is an extension of the linear regression method with multiple predictors



$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

Work with n observations – each has:

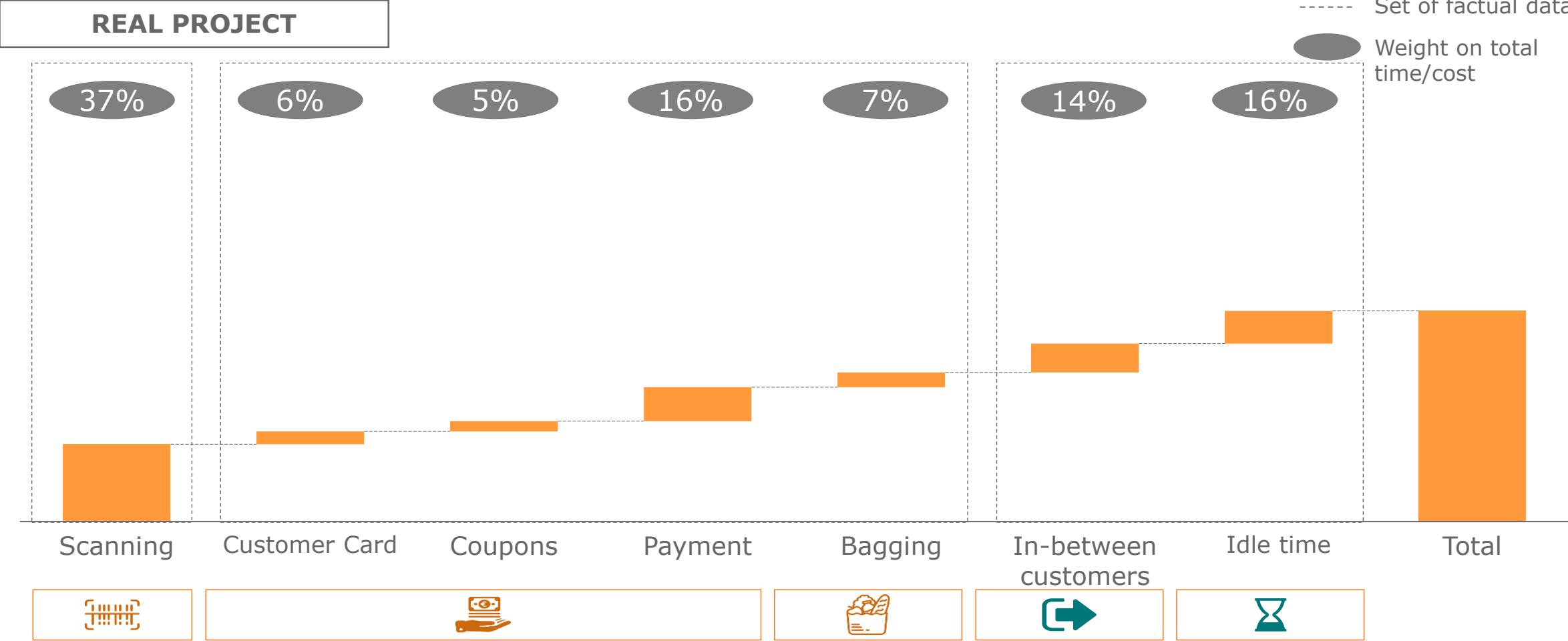
One observation of dependent variable

One observation each of the m independent variables

Seek to minimize the sum of squared differences

- First applied by Pearson in 1908: “*earn more about the relationship between several independent or predictor variables and a dependent or criterion variable*”
- Requires all admission conditions and assumptions as a simple linear regression model
- Variables may be continuous (eg. price), discrete (eg. number of promotions) or categorical (eg. discount type)

A multiple regression on the checkout time allowed to divide the total cost of operating the checkouts



Multiple Linear Regression: Example

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + s,$$

- We interpret β_j as the *average* effect on Y of a one unit increase in X_j , *holding all other predictors fixed*. In a given advertising example, the model becomes

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

Interactions:

- In our previous analysis of the **Advertising** data, we assumed that the effect on **sales** of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

states that the average effect on **sales** of a one-unit increase in **TV** is always β_1 , regardless of the amount spent on **radio**.

Interactions — continued

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for **TV** should increase as **radio** increases.
- In this situation, given a fixed budget of \$100 000, spending half on **radio** and half on **TV** may increase **sales** more than allocating the entire amount to either **TV** or to **radio**.
- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.
- Model takes the form: $\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV})$

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, **TV** and **radio**) do not.
- The *hierarchy principle*:
 - ***If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.***
- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
- Specifically, the interaction terms also contain main effects, if the model has no main effect terms.

Linear Regression

Multiple Linear
Regression

**Conditions for Regression
validity**

Variable selection

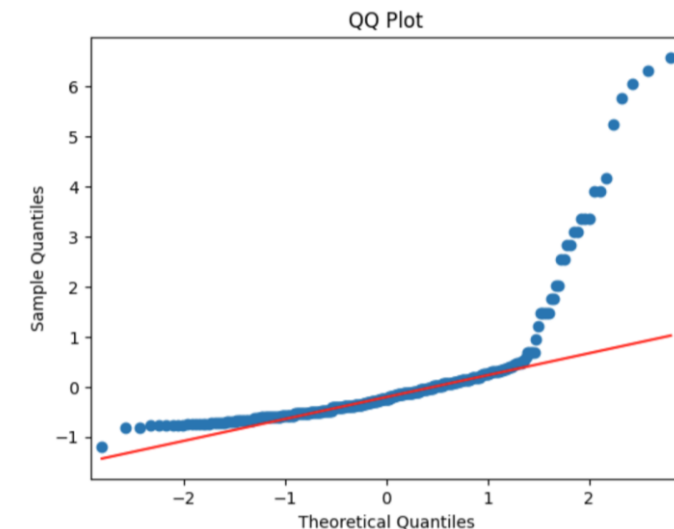
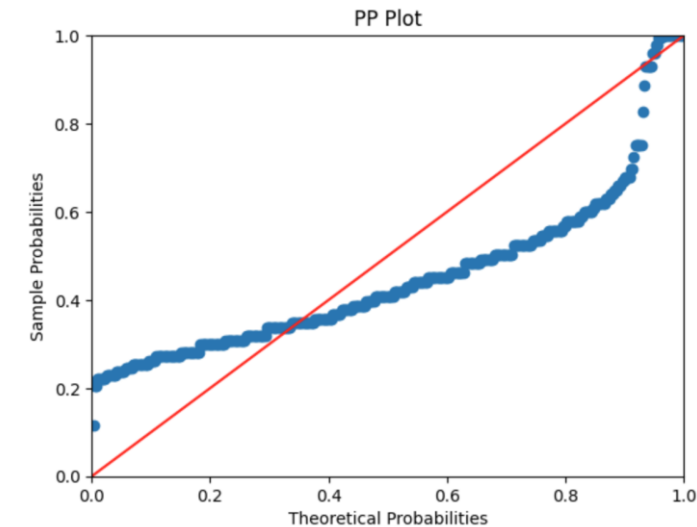
Logistic Regression

Linear Mixed Models



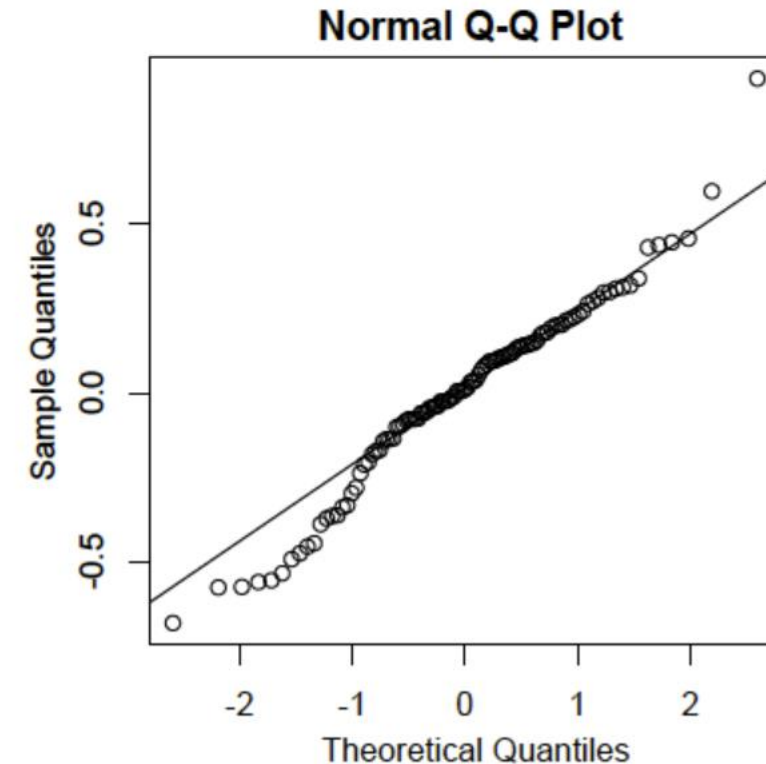
1. Residuals are normally distributed

- A **P-P plot** compares the empirical cumulative distribution function of a data set with a specified theoretical cumulative distribution function
- A **Q-Q plot** compares the quantiles of a data distribution with the quantiles of a standardized theoretical distribution from a specified family of distributions
- For the most part, the normal P-P plot is better at finding deviations from normality in the center of the distribution, and the normal Q-Q plot is better at finding deviations in the tails
- A **Kolmogorov-Smirnov test** (K-S test) can be used as a complement, to compare the sample against the normal distribution



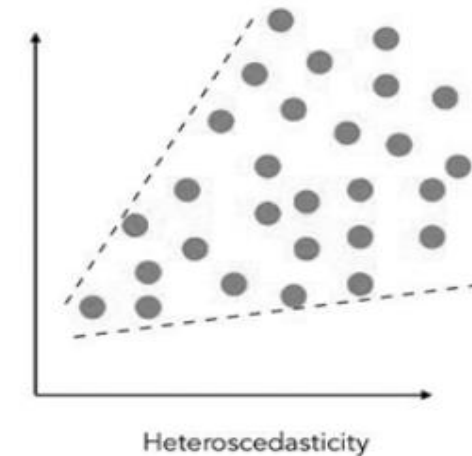
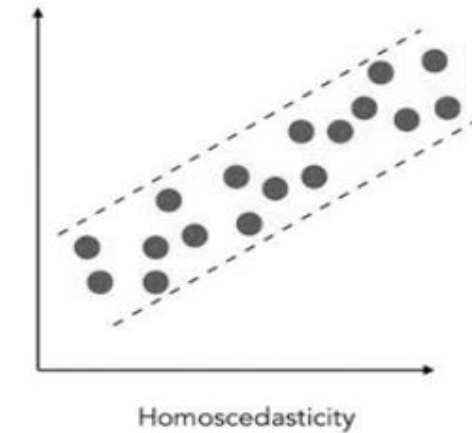
1. Residuals are normally distributed

- A **Q-Q plot** for the residuals can be used to assess this condition
- For the example, the plot reveals that the actual data values at the lower end of the distribution do not increase as much as would be expected for a normal distribution.
- It also reveals that the highest value in the data is higher than would be expected for the highest value in a sample of this size from a normal distribution.
- Nonetheless, the **distribution does not deviate greatly from normality**

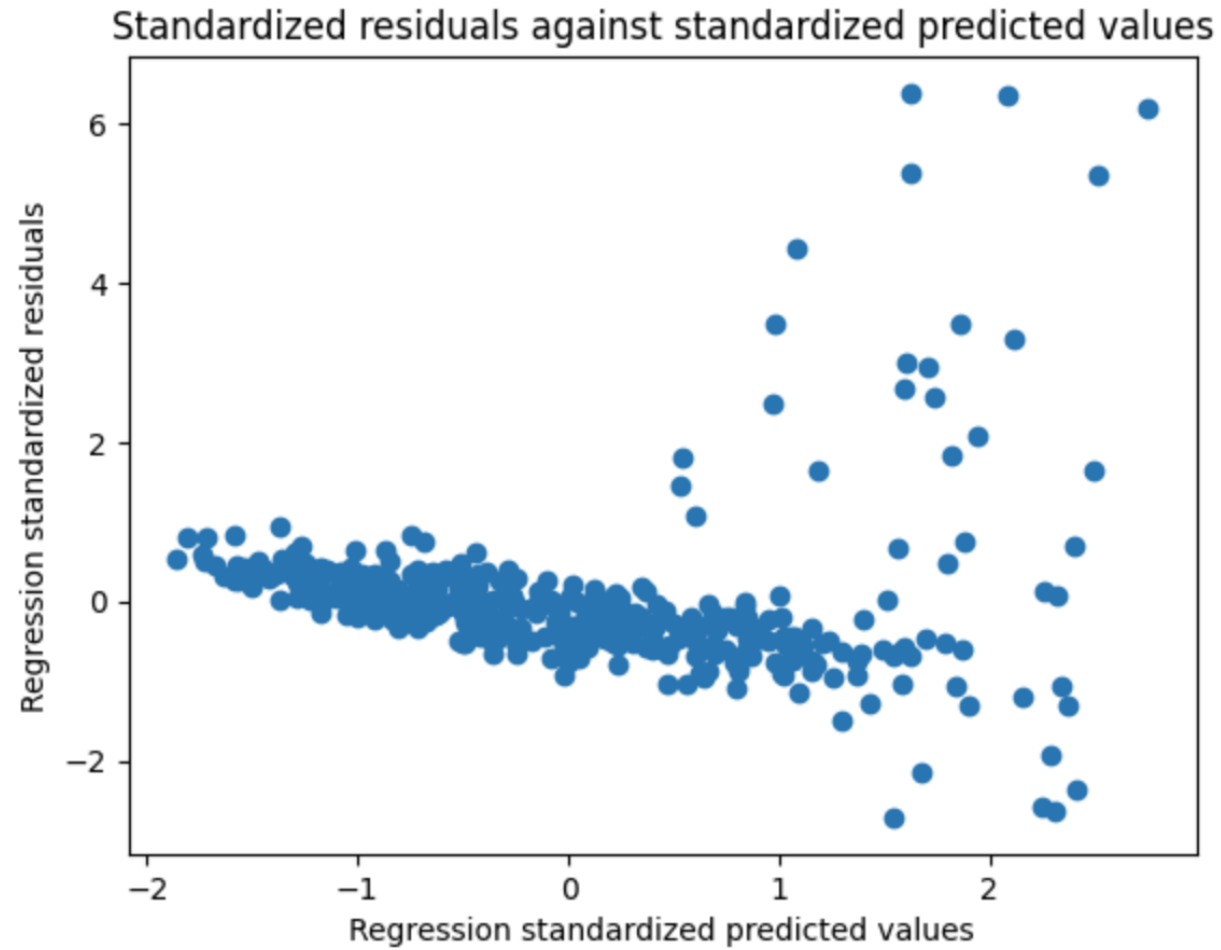


2. Homoscedasticity

- **Homoscedasticity** describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables
- **Heteroscedasticity** (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable. The impact of violating the assumption of homoscedasticity is a matter of degree, increasing as heteroscedasticity increases

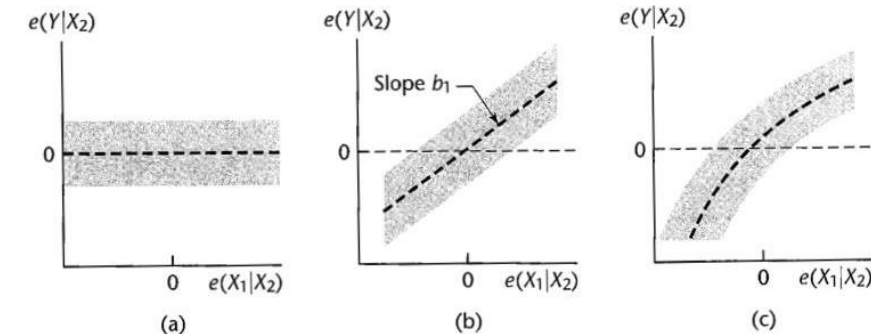


2. Homoscedasticity



3. Linearity

- The relationship between each **predictor variable** and the **criterion variable** is **linear**
- If this assumption is not met, then the predictions may systematically overestimate the actual values for one range of values on a predictor variable and underestimate them for another
- Independent variables and the dependent variables could be **transformed so that the relationship between them is linear**
- Plotting **partial regression plots** is the way-to-go to assess this condition



4. Absence of severe multicollinearity

- Absence of severe multicollinearity
Multicollinearity is the condition of strong correlation between predictors
- It is acceptable to have it as long as it is not so severe to the extent of impeding independent variation in the predictors
- If it takes place the estimates will be biased
- Variance Inflation Factor (VIF) is a test to assess this metric. Severe multicollinearity when $VIF > 10$
- Calculating the correlation coefficient of the predictors is a good complement to VIF

Regression Analysis: Femoral Neck versus %Fat S, Weight S, Activity S

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.55578	0.138946	27.95	0.000
%Fat S	1	0.04786	0.047863	9.63	0.003
Weight S	1	0.30473	0.304728	61.29	0.000
Activity S	1	0.04703	0.047027	9.46	0.003
%Fat S*Weight S	1	0.04175	0.041745	8.40	0.005
Error	87	0.43256	0.004972		
Total	91	0.98834			

Model Summary

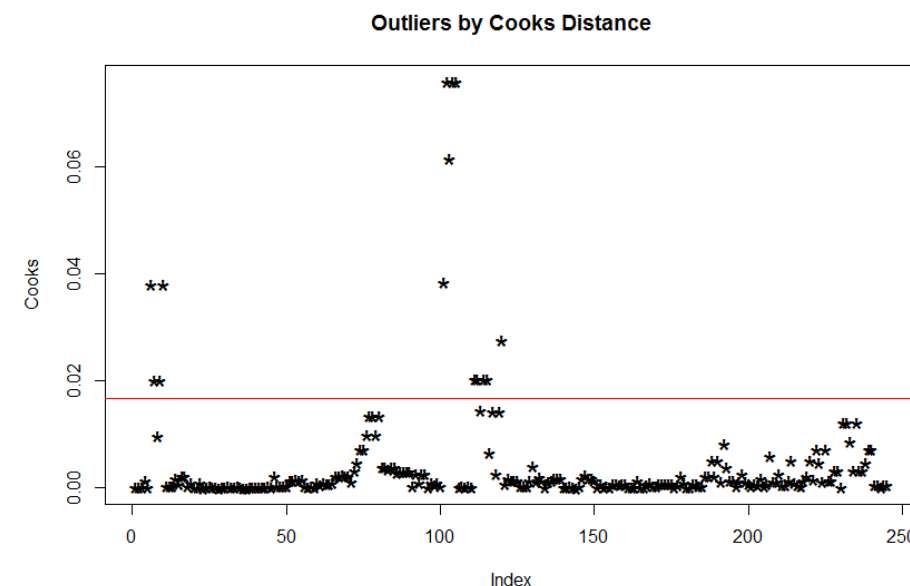
S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.82161	0.00973	84.40	0.000	
%Fat S	-0.00598	0.00193	-3.10	0.003	3.32
Weight S	0.00835	0.00107	7.83	0.000	4.75
Activity S	0.000022	0.000007	3.08	0.003	1.05
%Fat S*Weight S	-0.000214	0.000074	-2.90	0.005	1.99

5. Absence of influential outliers

- More important than finding outliers is to detect **influential data points**, that is, points that when deleted produce a substantial change in at least one of the regression coefficients
- **Cook's distance** or **Cook's D** is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis (practical threshold = 0.25)
- Cook's distance can be used to indicate influential data points that are particularly worth checking for validity or to indicate regions of the design space where it would be good to be able to obtain more data points



Linear Regression

Multiple Linear
Regression

Conditions for Regression
validity

Variable selection

Logistic Regression

Linear Mixed Models



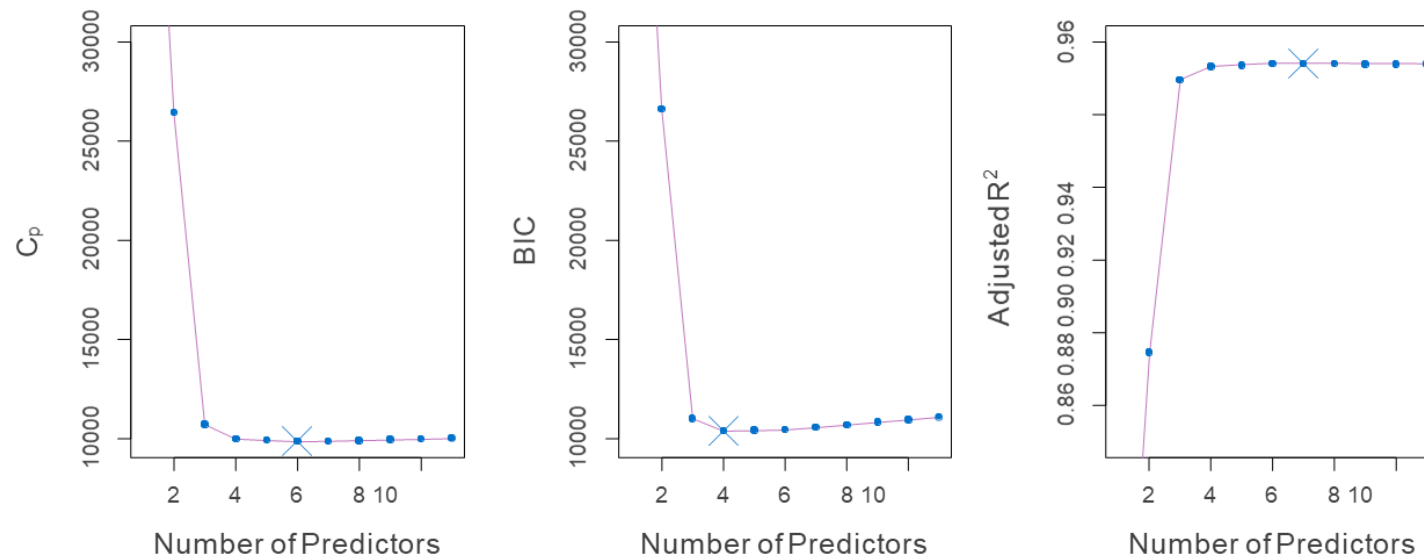
1. Let \mathcal{M}_0 denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \dots, p$:
 - (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - (b) Pick the best among these $\binom{p}{k}$ models, and call it \mathcal{M}_k . Here *best* is defined as having the smallest RSS, or equivalently largest R^2 .
3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large p .
- Best subset selection may also suffer from statistical problems when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus, an enormous search space can lead to *overfitting* and high variance of the coefficient estimates.
- For both of these reasons, *stepwise* methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.
 - *Forward stepwise selection* begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
 - *Backward stepwise selection* begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Estimating test error: two approaches

- We can indirectly estimate test error by making an *adjustment* to the training error to account for the bias due to overfitting.
- We can *directly* estimate the test error, using either a validation set approach or a cross-validation approach
- Let us deep-dive on the adjustment approach. The next figure displays C_p , BIC, and adjusted R^2 for the best model of each size produced by best subset selection



Estimating test error metrics

- *Mallow's C_p* :

$$C_p = \frac{1}{n} (\text{RSS} + 2d\hat{\sigma}^2),$$

where d is the total # of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement.

- The *AIC* criterion is defined for a large class of models fit by maximum likelihood:

$$\text{AIC} = -2 \log L + 2 \cdot d$$

where L is the maximized value of the likelihood function for the estimated model.

- In the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.

Estimating test error metrics

$$\text{BIC} = \frac{1}{n} (\text{RSS} + \log(n)d\hat{\sigma}^2) .$$

- Like C_p , the BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.
- Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ term, where n is the number of observations.
- Since $\log n > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p . See Figure on slide 19.

Estimating test error metrics

- For a least squares model with d variables, the adjusted R^2 statistic is calculated as

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n - d - 1)}{\text{TSS}/(n - 1)}.$$

where TSS is the total sum of squares.

- Unlike C_p , AIC, and BIC, for which a *small* value indicates a model with a low test error, a *large* value of adjusted R^2 indicates a model with a small test error.
- Maximizing the adjusted R^2 is equivalent to minimizing $\frac{\text{RSS}}{n-d-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{\text{RSS}}{n-d-1}$ may increase or decrease, due to the presence of d in the denominator.
- Unlike the R^2 statistic, the adjusted R^2 statistic *pays a price* for the inclusion of unnecessary variables in the model.

Linear Regression

Multiple Linear
Regression

Conditions for Regression
validity

Variable selection

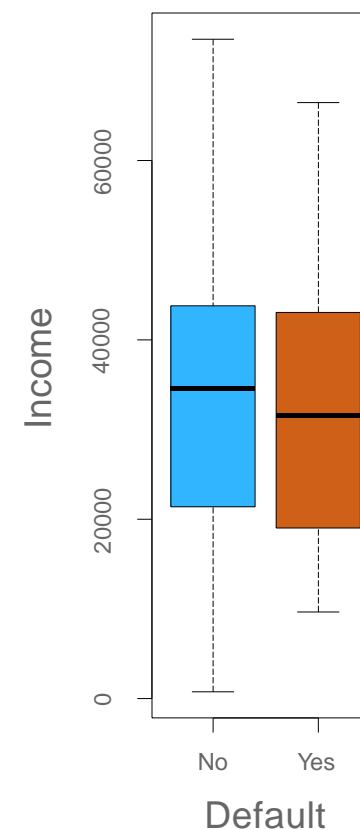
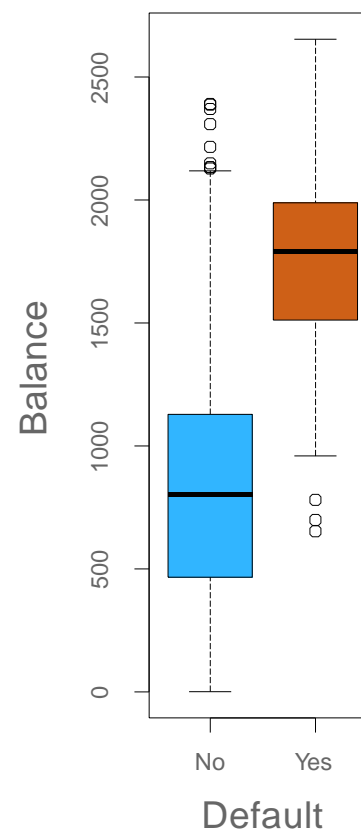
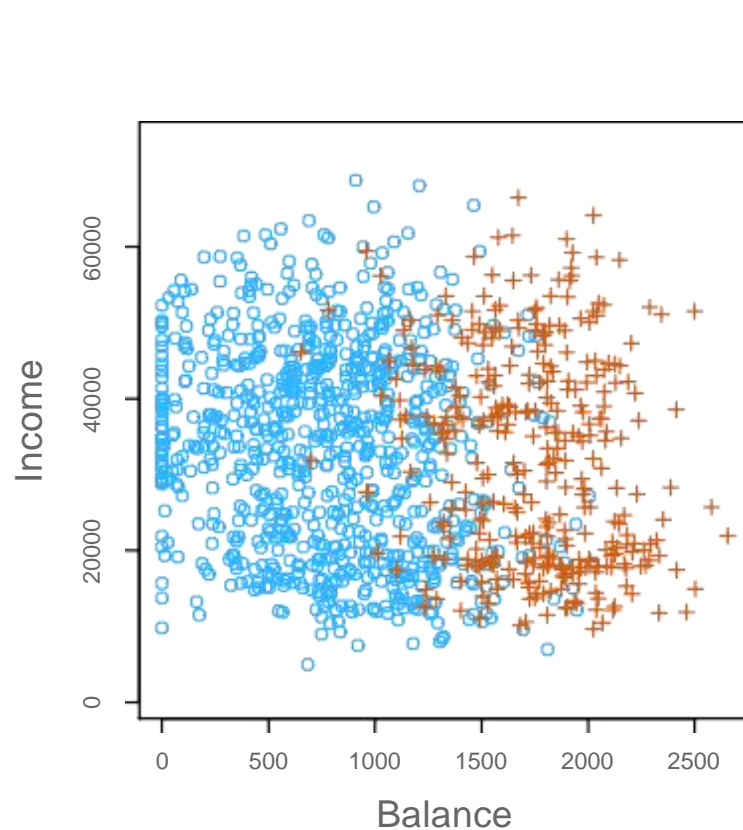
Logistic Regression

Linear Mixed Models



- Qualitative variables take values in an unordered set C , such as:
 $\text{eye color} \in \{\text{brown, blue, green}\}$
 $\text{email} \in \{\text{spam, ham}\}.$
- Given a feature vector X and a qualitative response Y taking values in the set C , the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y ; i.e. $C(X) \in C$.
- Often, we are more interested in estimating the *probabilities* that X belongs to each category in C . For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

Credit Card Default (classic example)



Can we use Linear Regression?

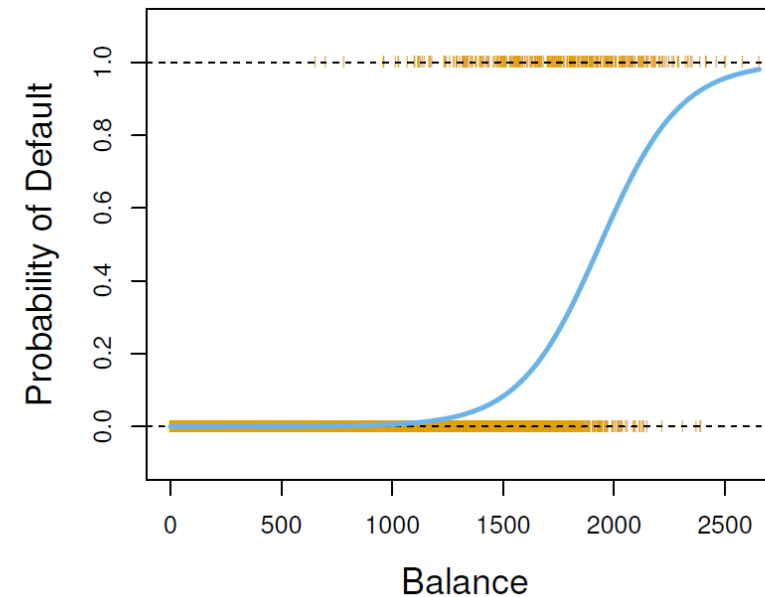
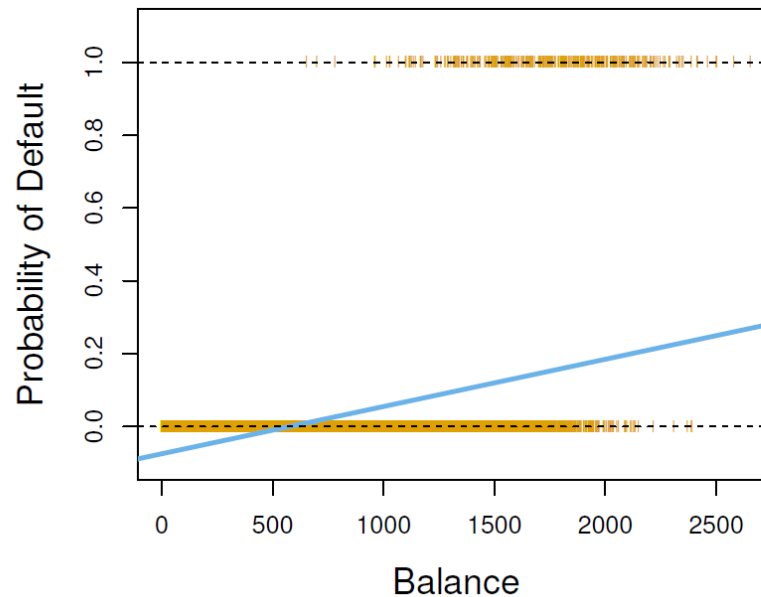
Suppose for the **Default** classification task that we code

$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes.} \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as **Yes** if $\hat{Y} > 0.5$?

- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to *linear discriminant analysis* which we discuss later.
- Since in the population $E(Y|X = x) = \Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- However, *linear* regression might produce probabilities less than zero or bigger than one. *Logistic regression* is more appropriate.

Linear versus Logistic Regression



The orange marks indicate the response Y , either 0 or 1. Linear regression does not estimate $\Pr(Y = 1 | X)$ well. Logistic regression seems well suited to the task.

Multiclass Logistic Regression

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

This coding suggests an ordering, and in fact implies that the difference between **stroke** and **drug overdose** is the same as between **drug overdose** and **epileptic seizure**.

Linear regression is not appropriate here.

Multiclass Logistic Regression or *Discriminant Analysis* are more appropriate.

Logit function

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using **balance** to predict **default**. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

($e \approx 2.71828$ is a mathematical constant [Euler's number].)

It is easy to see that no matter what values β_0 , β_1 or X take, $p(X)$ will have values between 0 and 1.

A bit of rearrangement gives

$$\log \left(\frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X.$$

This monotone transformation is called the *log odds* or *logit* transformation of $p(X)$. (by log we mean *natural log*: \ln .)

We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

Most statistical packages can fit linear logistic regression models by maximum likelihood.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Making Predictions

What is our estimated probability of **default** for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 1000}}{1 + e^{-10.6513 + 0.0055 \times 1000}} = 0.006$$

With a balance of \$2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 2000}}{1 + e^{-10.6513 + 0.0055 \times 2000}} = 0.586$$

Conditions for logistic regression validity

- 1 Dependent variable is dichotomous
- 2 There are no influential outliers (as in linear regression)
- 3 Absence of multicollinearity (as in linear regression)
- 4 Linear relationship between the odds ratio and each independent variable (Box-Tidwell test - used to check whether the logit transform is a linear function of the predictor, effectively by adding the non-linear transform of the original predictor as an interaction term to test if this addition made no better prediction)

Can we highlight the countries that are more likely to legalize cannabis?

REAL PROJECT

Cannabis already
legalized for
adult/medical use



Canada
Uruguay

Cannabis already legalized
for medical use only

Australia	Denmark
Austria	Finland
Belgium	FYR Macedonia
Chile	Georgia
Colombia	Germany
Croatia	Greece
Cyprus	Israel
Czech Republic	Italy



Cannabis already
legalized for
medical use only

Jamaica
Luxembourg
Malta
Mexico
Netherlands
New Zealand
Norway
Peru
Poland
Portugal
Slovenia
South Africa
Sri Lanka
Switzerland
United Kingdom
Zimbabwe

Drivers considered to build the model covered economic, political, religious and legal contexts

REAL PROJECT

Economic and human development

- GDP per capita
- Inequality index (Gini coefficient)
- Human Development Index
- Life expectancy
- Average years of schooling
- Index of economic/press freedom
- Environmental Performance Index

Political context

- Incidence of women in parliament
- Political system
- Parliament's positioning (left vs. right wing)
- Democracy/pluralism index

Religious context

- Religious freedom index
- Incidence of Christians
- Incidence of Muslims
- Incidence of non-believers

Media buzz

- Google searches (from Google trends)
- Reddit mentions

Mindset and behaviors

- Public opinion on same-sex marriage
- Incidence of opioids consumption

Legal framework (and date of approval)

- Same-sex sexual activity
- Recognition of same-sex unions
- Same-sex marriage
- Adoption by same-sex couples
- Anti-discrimination laws for sexual orientation
- Laws concerning gender identity/expression
- Voluntary termination of pregnancy
- Euthanasia
- Prostitution
- Drug consumption

A significant portion of the world's countries were analyzed over a wide set of metrics

REAL PROJECT

145



COUNTRIES SCORED
ON THE LIKELIHOOD
OF CANNABIS
LEGALIZATION

90+



METRICS
GATHERED,
COVERING A WIDE
SET OF DRIVERS

33



COUNTRIES WHERE
CANNABIS IS LEGAL
FOR MEDICAL USE

13



SPECIFIC LAWS
WHOSE COUNTRIES'
STANCE ON WAS
DEEPLY ANALYZED

22



DIFFERENT DATA
SOURCES
COMBINED

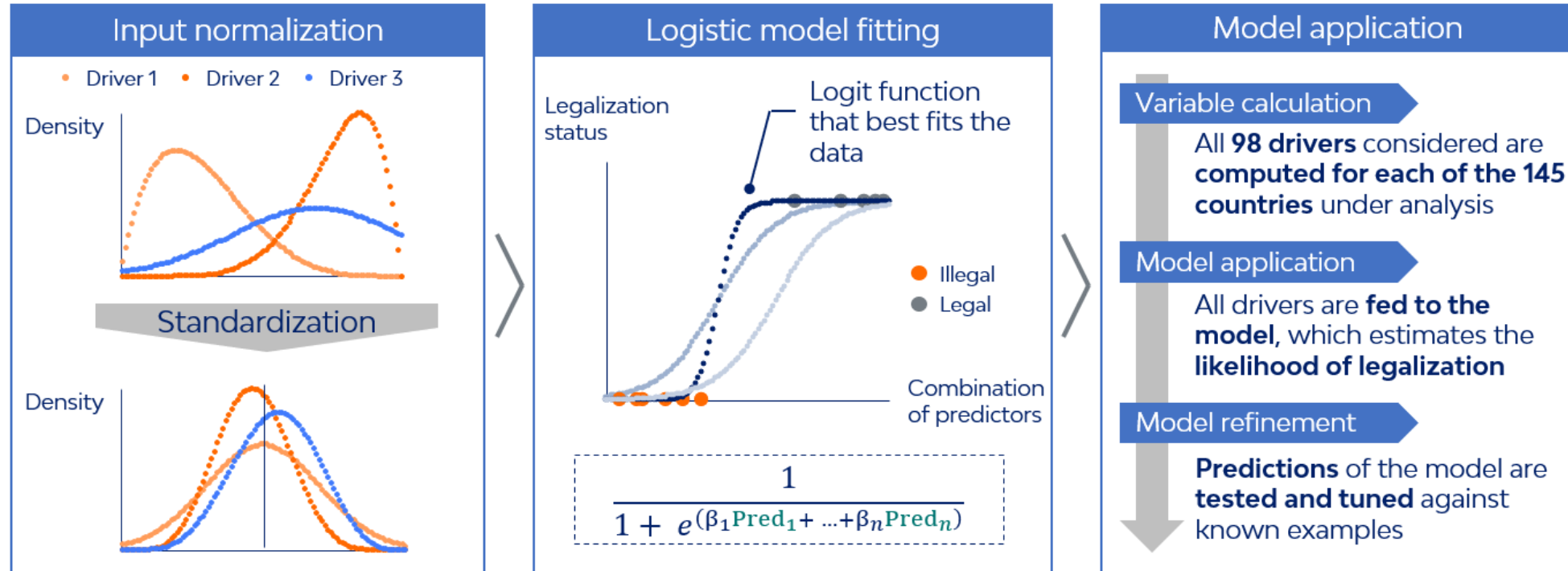
2



COUNTRIES WITH
RECREATIONAL USE
OF CANNABIS
LEGALIZED

A logistic regression was the most successful model at capturing the intricacies of the challenge

REAL PROJECT




Countries more likely to be next to legalize cannabis seem to be Sweden and Australia

REAL PROJECT

Country rank by likelihood to legalize cannabis for: # Medical use # Adult use

Top 10 countries more likely to be the next to legalize cannabis for medical use

 Sweden	1	12	 Hungary	6	33
 Japan	2	18	 France	7	34
 Spain	3	20	 United States	8	35
 Iceland	4	21	 Costa Rica	9	36
 Ireland	5	27	 Estonia	10	38

Top 10 countries more likely to be the next to legalize cannabis for adult use

 Australia	-	1	 Netherlands	-	6
 UK	-	2	 New Zealand	-	7
 Luxembourg	-	3	 Mexico	-	8
 Germany	-	4	 Denmark	-	9
 Finland	-	5	 Malta	-	10

Potential next steps:

- Analyze US at state level
- Split analysis in 2 independent models (one focused on medical use, another focused on adult use)

Linear Regression

Multiple Linear
Regression

Conditions for Regression
validity

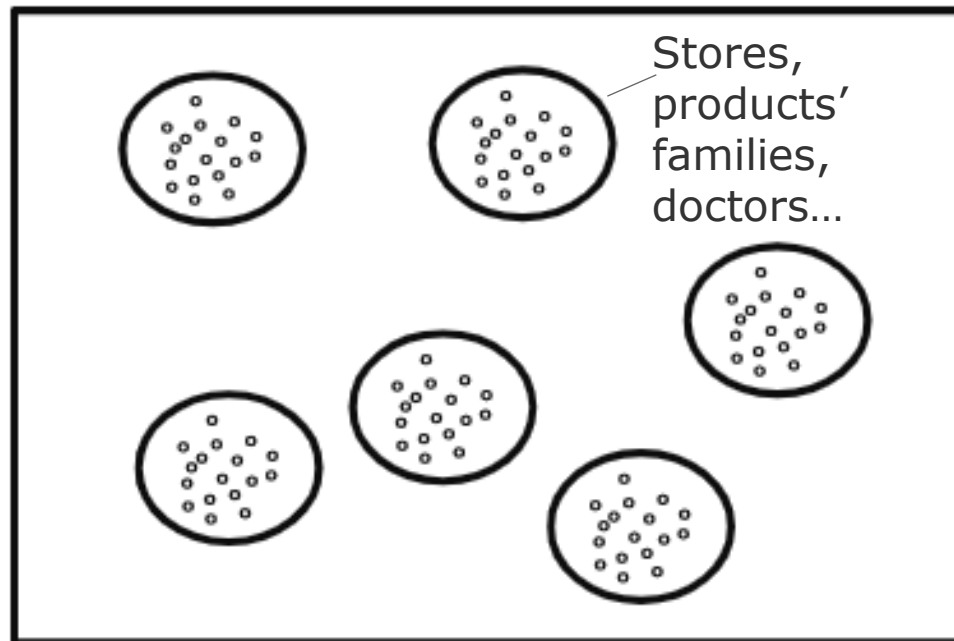
Variable selection

Logistic Regression

Linear Mixed Models



Linear Mixed Models: motivation



- Linear mixed models are an extension of simple linear models
- Particularly relevant when there is non independence in the data arising from a hierarchical structure
- Variability in the outcome can be thought of as being either within group or between group

Approaches:

- Aggregation (does not take advantage of all the data – data is averaged)
- Individual regressions (does not take advantage of the information in data from other groups)
- **Multilevel models**

Some theory behind! (and example)

Matrix of p predictor variables

Design matrix for the q random effects and J groups

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

Outcomes

Vector of the **fixed-effects** regression coefficients

Vector of q **random effects** (the random complement to the fixed $\boldsymbol{\beta}$)

Fixed-effect: Parameter that does not vary

Random-effects: Parameters that are themselves random variables

Example:

- $J=407$ doctors; $N=8525$ patients
- y – continuous variable, mobility scores
- $y \sim f(\text{Age, Married, Sex, Red Blood Cell count, White Blood Cell count, intercept})$ – 6 fixed effects predictors

$$\underbrace{\mathbf{y}}_{8525 \times 1} = \underbrace{\underbrace{\mathbf{X}}_{8525 \times 6} \underbrace{\boldsymbol{\beta}}_{6 \times 1}}_{8525 \times 1} + \underbrace{\underbrace{\mathbf{Z}}_{8525 \times 407} \underbrace{\mathbf{u}}_{407 \times 1}}_{8525 \times 1} + \underbrace{\boldsymbol{\epsilon}}_{8525 \times 1}$$

Some theory behind! (and example)

Matricial view of the fixed effects coefficients

$$\mathbf{y} = \begin{bmatrix} \text{mobility} \\ 2 \\ 2 \\ \dots \\ 3 \end{bmatrix} \quad \mathbf{n}_{ij} = \begin{bmatrix} 1 \\ 2 \\ \dots \\ 8525 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \text{Intercept} & \text{Age} & \text{Married} & \text{Sex} & \text{WBC} & \text{RBC} \\ 1 & 64.97 & 0 & 1 & 6087 & 4.87 \\ 1 & 53.92 & 0 & 0 & 6700 & 4.68 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 56.07 & 0 & 1 & 6430 & 4.73 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} 4.782 \\ .025 \\ .011 \\ .012 \\ 0 \\ -.009 \end{bmatrix}$$

Final equation (with random effects inclusion)

- i – index for patient
- j – index for doctor

$$L1: Y_{ij} = \beta_{0j} + \beta_{1j}Age_{ij} + \beta_{2j}Married_{ij} + \beta_{3j}Sex_{ij} + \beta_{4j}WBC_{ij} + \beta_{5j}RBC_{ij} + e_{ij}$$

$$L2: \beta_{0j} = \gamma_{00} + u_{0j}$$

$$L2: \beta_{1j} = \gamma_{10}$$

$$L2: \beta_{2j} = \gamma_{20}$$

$$L2: \beta_{3j} = \gamma_{30}$$

$$L2: \beta_{4j} = \gamma_{40}$$

$$L2: \beta_{5j} = \gamma_{50}$$

Random coefficient
to capture the
difference between
doctors (in this
case, intercept
only)

Model evaluation

$$R_{\text{GLMM}(m)}^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^u \sigma_l^2 + \sigma_e^2 + \sigma_d^2}$$

Marginal R²: Proportion of variance explained by the fixed factor(s) alone.

Denominator, in order: (1) fixed-effects variance; (2) random variance (partitioned by level l); (3)(4) residual variance

$$R_{\text{GLMM}(c)}^2 = \frac{\sigma_f^2 + \sum_{l=1}^u \sigma_l^2}{\sigma_f^2 + \sum_{l=1}^u \sigma_l^2 + \sigma_e^2 + \sigma_d^2}$$

Conditional R²: Proportion of variance explained by both the fixed and random factors

WINNERS: example of application

REAL PROJECT

Linear mixed model fit by maximum likelihood ['lmerMod']

Formula: `net_sales_va_pc ~ (1 | loja) + ano + mes_num + q1_avg_price_dif + valor_total_ind_efe_pc + valor_total_col_efe_pc + sum_pontos_valor_pc + Data: incentives_data_nout_loja_pc`

`q2_avg_price_dif + q3_avg_price_dif + q4_avg_price_dif + flag_sem_pontos + avg_antiguidade_aswas + avg_peso_variavel_perc`

AIC	BIC	logLik	deviance	df.resid
114704.6	114877.7	-57326.3	114652.6	5748

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.7932	-0.6114	-0.0563	0.5447	4.0563

Random effects:

Groups	Name	Variance	Std.Dev.
loja	(Intercept)	135053837	11621
Residual		20718912	4552

Number of obs: 5774, groups: loja, 189

Random effects: intercept at the store level

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-29518.717	7644.051	-3.862
ano2018	2013.784	283.064	7.114
ano2019	4107.429	460.124	8.927
mes_num2	-5144.621	340.892	-15.092
mes_num3	-2088.057	358.096	-5.831
mes_num4	-3650.615	354.765	-10.290
mes_num5	-3088.596	325.332	-9.494
mes_num6	-1220.657	335.539	-3.638
mes_num7	1851.741	336.911	5.496
mes_num8	3035.791	335.777	9.041
mes_num9	192.454	341.418	0.564
mes_num10	904.931	336.492	2.689
mes_num11	7704.962	436.102	17.668
mes_num12	13596.047	369.123	36.833
q1_avg_price_dif	139.773	38.817	3.601
q2_avg_price_dif	-244.083	44.988	-5.425
q3_avg_price_dif	-123.204	36.246	-3.399
q4_avg_price_dif	190.032	29.168	6.515
valor_total_ind_efe_pc	36.914	5.249	7.033
valor_total_col_efe_pc	19.846	3.289	6.035
sum_pontos_valor_pc	58.692	6.063	9.680
flag_sem_pontos1	826.928	174.129	4.749
avg_antiguidade_aswas	785.611	61.940	12.684
avg_peso_variavel_perc	2884.113	431.887	6.678

Fixed effects: time effects

Fixed effects: incentives (and other) variables

WINNERS: example of application – continued

REAL PROJECT

```
Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: net_sales_va_pc ~ (1 | loja) + ano + mes_num + (1 + valor_total_ind_efe_pc +
  valor_total_col_efe_pc | formato) + q1_avg_price_dif + q2_avg_price_dif + q3_avg_price_dif + q4_avg_price_dif + sum_pontos_valor_pc +
  flag_sem_pontos + avg_antiguidade_aswas1 + avg_antiguidade_aswas2 + valor_aval_qual_pc
Data: incentives_data_nout_loja_pc
```

```
      AIC      BIC    logLik deviance df.resid
114580.2 114786.7 -57259.1 114518.2     5743
```

Scaled residuals:

```
      Min       1Q   Median       3Q      Max
-3.7939 -0.5927 -0.0645  0.5470  4.3612
```

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
loja	(Intercept)	5.784e+07	7605.54	
formato	(Intercept)	8.655e+07	9303.38	
	valor_total_ind_efe_pc	3.956e+03	62.90	0.62
	valor_total_col_efe_pc	6.793e+02	26.06	0.96 0.82
Residual		2.071e+07	4550.84	

Number of obs: 5774, groups: loja, 189; formato, 3

Random effects: extension to calculate **incentives' impact at 'formato' level**

```
$formato
      valor_total_ind_efe_pc valor_total_col_efe_pc
Mega          101.007195          45.427462
Mobile         54.824422           4.113389
Super           7.658933          23.216804
```

Make change happen

Calculating Odds Ratio

Calculating the Odds Ratio (OR)

	Disease (Case)	No Disease (Control)
Exposed	A	B
Unexposed	C	D

$$\text{OR} = \frac{\text{Odds that a case was exposed (A/C)}}{\text{Odds that a control was exposed (B/D)}} = \frac{AD}{BC}$$