DAY 3 – LECTURE OUTLINE

- Testing for a correlation hypothesis (relationship of variables)
  - Pearson rho analysis (param)
  - Spearman test (no param)
- Measures of association
  - Fisher's Exact Test
  - Chi-Square Test of Independence
- From association to prediction -> Machine learning algorithms
  - Supervised
    - Example: Linear regression models
    - Multiple Linear Regression
    - logistic regression?

Measures of association

Correlation and covariance

For two quantitative variables, the basic statistics of interest are the sample co-variance and/or sample correlation, which correspond to and are estimates of the corresponding population parameters.

The sample **covariance is a measure of how much two variables "co-vary",** i.e., how much (and in what direction) should we expect one variable to change when the other changes. Positive covariance values suggest variables change in the same direction. Negative covariances suggest variables change in the opposite direction. And covariances near zero suggest that the two variables vary independently of each other.

Covariances tend to be hard to interpret, so we often use **correlation** instead. The correlation has the nice property that it is always between -1 and +1, with -1 being a "perfect" negative linear correlation, +1 being a perfect positive linear correlation and 0 indicating that X and Y are uncorrelated.

The symbol r or $r_{x,y}$ is often used for sample correlations.

The general formula for sample covariance is

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

It is worth noting that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

The formula for the sample correlation is

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{s_x s_y}$$

where s_x is the standard deviation of X and s_y is the standard deviation of Y.

Variance = σ^2 = $\Sigma\ (x_i - \mu)^2 / N$

	X	Y	Z
X	5.00	1.77	-2.24
Y	1.77	7.0	3.17
Z	-2.24	3.17	4.0

Table 4.5: A Covariance Matrix

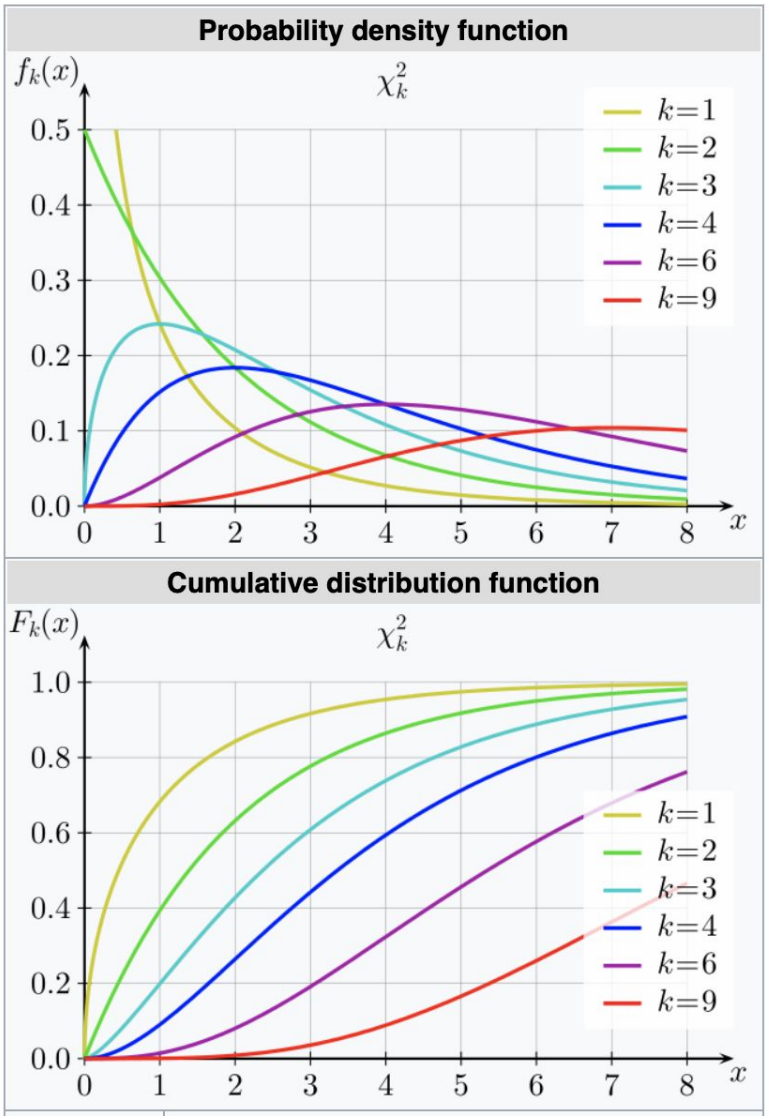
	X	Y	Z
X	1.0	0.3	-0.5
Y	0.3	1.0	0.6
Z	-0.5	0.6	1.0

Table 4.6: A Correlation Matrix

The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

Chi-squared distribution

chi-squared



Notation	$\chi^2(k)$ or χ^2_k
Parameters	$k \in \mathbb{N}^*$ (known as "degrees of freedom")
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)}\, \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k
Median	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Mode	$\max(k-2, 0)$
Variance	$2k$

The simplest chi-squared distribution is the square of a standard normal distribution.

Notation	$\chi^2(k)$ or χ^2_k

Chi-Square Goodness of Fit Test

A **Chi-Square goodness of fit test** is used to determine whether or not a categorical variable follows a hypothesized distribution.

Example:

We want to know if the percentage of M&M's that come in a bag are as follows: 20% yellow, 30% blue, 30% red, 20% other. To test this, we open a random bag of M&M's and count how many of each color appear.

We can use a Chi-Square goodness of fit test to determine if there is a statistically significant difference in the number of expected counts for each level of a variable compared to the observed counts.

Chi-Square Goodness of Fit Test: Formula

A Chi-Square goodness of fit test uses the following null and alternative hypotheses:

- **H_0: (null hypothesis)** A variable follows a hypothesized distribution.
- **H_1: (alternative hypothesis)** A variable does not follow a hypothesized distribution.

We use the following formula to calculate the Chi-Square test statistic X^2:

$X^2 = \Sigma(O-E)^2 / E$

where:

Σ: means "sum"

O: observed value

E: expected value

If the p-value that corresponds to the test statistic X^2 with n-1 degrees of freedom (where n is the number of categories) is less than your chosen significance level (common choices are 0.10, 0.05, and 0.01) then you can reject the null hypothesis.

Chi-Square Goodness of Fit Test: Example

A shop owner claims that an equal number of customers come into his shop each weekday. To test this hypothesis, an independent researcher records the number of customers that come into the shop on a given week and finds the following:
Monday: 50 customers
Tuesday: 60 customers
Wednesday: 40 customers
Thursday: 47 customers
Friday: 53 customers

Step 1: Define the hypotheses.
We will perform the Chi-Square goodness of fit test using the following hypotheses:
H_0: An equal number of customers come into the shop each day.
H_1: An equal number of customers do not come into the shop each day.

Chi-Square Goodness of Fit Test: Example

Step 2: Calculate $(O-E)^2 / E$ for each day.
There were a total of 250 customers that came into the shop during the week. Thus, if we expected an equal amount to come in each day then the expected value "E" for each day would be 50.

Step 3: Calculate the test statistic X^2.
$\mathbf{X^2} = \Sigma(O-E)^2 / E = 0 + 2 + 2 + 0.18 + 0.18 = \mathbf{4.36}$
Step 4: Calculate the p-value of the test statistic X^2.
According to the Chi-Square Score to P Value Calculator, the p-value associated with $X^2 = 4.36$ and $n-1 = 5-1 = 4$ degrees of freedom is **0.359472**.

Step 5: Draw a conclusion.
Since this p-value is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that the true distribution of customers is different from the distribution that the shop owner claimed.

Chi-Square Test of Independence

A **Chi-Square Test of Independence** is used to determine whether or not there is a significant association between two categorical variables.

Examples

- We want to know if education level and marital status are associated so we collect data about these two variables on a simple random sample of 50 people.

Chi-Square Test of Independence: Formula

A Chi-Square test of independence uses the following null and alternative hypotheses:

H_0: (null hypothesis) The two variables are independent.

H_1: (alternative hypothesis) The two variables are *not*independent. (i.e. they are associated)

We use the following formula to calculate the Chi-Square test statistic X^2:

$X^2 = \Sigma(O-E)^2 / E$

where:

Σ: means “sum”

O: observed value

E: expected value

If the p-value that corresponds to the test statistic X^2 with (#rows-1)*(#columns-1) degrees of freedom is less than your chosen significance level then you can reject the null hypothesis.

Chi-Square Test of Independence: Example

Suppose we want to know whether or not gender is associated with political party preference. We take a simple random sample of 500 voters and survey them on their political party preference. The following table shows the results of the survey:

	Republican	Democrat	Independent	Total
Male	120	90	40	250
Female	110	95	45	250
Total	230	185	85	500

Step 1: Define the hypotheses.

We will perform the Chi-Square test of independence using the following hypotheses:

- **H_0:** Gender and political party preference are independent.
- **H_1:** Gender and political party preference are *not* independent

Step 2: Calculate the expected values.

Next, we will calculate the expected values for each cell in the contingency table using the following formula:

Expected value = (row sum * column sum) / table sum.

For example, the expected value for Male Republicans is: (230*250) / 500 = **115**.

We can repeat this formula to obtain the expected value for each cell in the table:

	Republican	Democrat	Independent	Total
Male	115	92.5	42.5	250
Female	115	92.5	42.5	250
Total	230	185	85	500

Step 3: Calculate $(O-E)^2 / E$ for each cell in the table.

Next we will calculate **$(O-E)^2 / E$** for each cell in the table where:

- **O:** observed value
- **E:** expected value

For example, Male Republicans would have a value of: $(120-115)^2 /115 =$ **0.2174**.

We can repeat this formula for each cell in the table:

	Republican	Democrat	Independent
Male	0.2174	0.0676	0.1471
Female	0.2174	0.0676	0.1471

Step 4: Calculate the test statistic X^2 and the corresponding p-value.

X^2 $= \Sigma(O-E)^2 / E = 0.2174 + 0.2174 + 0.0676 + 0.0676 + 0.1471 + 0.1471 =$ **0.8642**

According to the Chi-Square Score to P Value Calculator, the p-value associated with $X^2 =$ 0.8642 and $(2-1)*(3-1) = 2$ degrees of freedom is **0.649198**.

Step 5: Draw a conclusion.

Since this p-value is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that there is an association between gender and political party preference.

Fisher's Exact Test

Fisher's Exact Test is used to determine whether or not there is a significant association between two categorical variables. It is typically used as an alternative to the Chi-Square Test of Independence when one or more of the cell counts in a 2×2 table is less than 5.

Fisher's Exact Test uses the following null and alternative hypotheses:

- **H_0: (null hypothesis)** The two variables are independent.
- **H_1: (alternative hypothesis)** The two variables are *not* independent.

Three Ways to Calculate Effect Size for a Chi-Square Test

In statistics, there are two commonly used Chi-Square tests:

Chi-Square Test for Goodness of Fit: Used to determine whether or not a categorical variable follows a hypothesized distribution.

Chi-Square Test for Independence: Used to determine whether or not there is a significant association between two categorical variables from a single population.

F

or both of these tests, we end up with a p-value that tells us whether or not we should reject the null hypothesis of the test. The p-value tells us whether or not the results of the test are significant, but it doesn't tell us the effect size of the test.

There are three ways to measure effect size: Phi (φ), Cramer's V (V), and odds ratio (OR).

Odds Ratio (OR)

Given the following 2 x2 table:

Effect Size	# Successes	# Failures
Treatment Group	A	B
Control Group	C	D

The odds ratio would be calculated as:
Odds ratio = (AD) / (BC)

When to Use

It's appropriate to calculate the odds ratio only when you're working with a 2 x 2 contingency table. Typically the odds ratio is calculated when you're interested in studying the odds of success in a treatment group relative to the odds of success in a control group.

How to Interpret

There is no specific value at which we deem an odds ratio be a small, medium, or large effect, but the further away the odds ratio is from 1, the higher the likelihood that the treatment has an actual effect.

It's best to use domain specific expertise to determine if a given odds ratio should be considered small, medium, or large.

From association to prediction - Machine learning

: Linear regression models

A Quick Introduction to Supervised vs. Unsupervised Learning

The field of machine learning contains a massive set of algorithms that can be used for understanding data. These algorithms can be classified as:

1. **Supervised Learning Algorithms:**

building a model to estimate or predict an output based on one or more inputs.

2. Unsupervised Learning Algorithms:

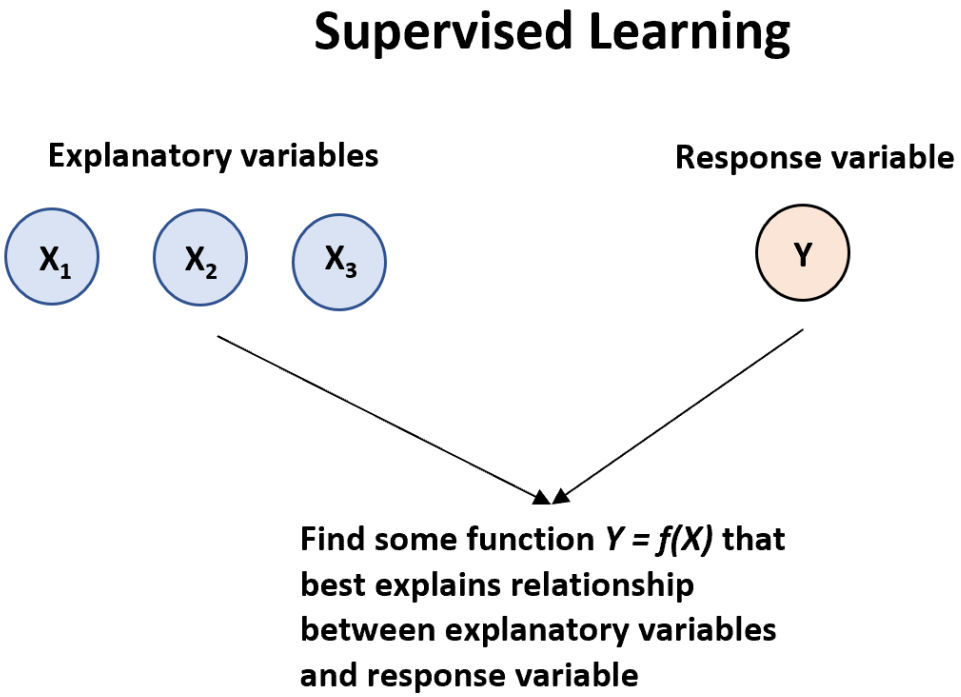
finding structure and relationships among inputs. There is no “supervising” output.

Supervised Learning Algorithms

A **supervised learning algorithm** can be used when we have one or more explanatory variables (X_1, X_2, X_3, …, X_p) and a response variable (Y) and we would like to find some function that describes the relationship between the explanatory variables and the response variable:

$Y = f(X) + \varepsilon$

where *f* represents systematic information that X provides about Y and where ε is a random error term



There are two main types of supervised learning algorithms:

1. **Regression:** The output variable is continuous (e.g. weight, height, time, etc.)

2. **Classification:** The output variable is categorical (e.g. male or female, pass or fail, benign or malignant, etc.)

There are two main reasons that we use supervised learning algorithms:

1. **Prediction:** We often use a set of explanatory variables to predict the value of some response variable (e.g. using *square footage* and *number of bedrooms* to predict *home price*)

2. **Inference:** We may be interested in understanding the way that a response variable is affected as the value of the explanatory variables change (e.g. how much does home price increase, on average, when the number of bedrooms increases by one?)

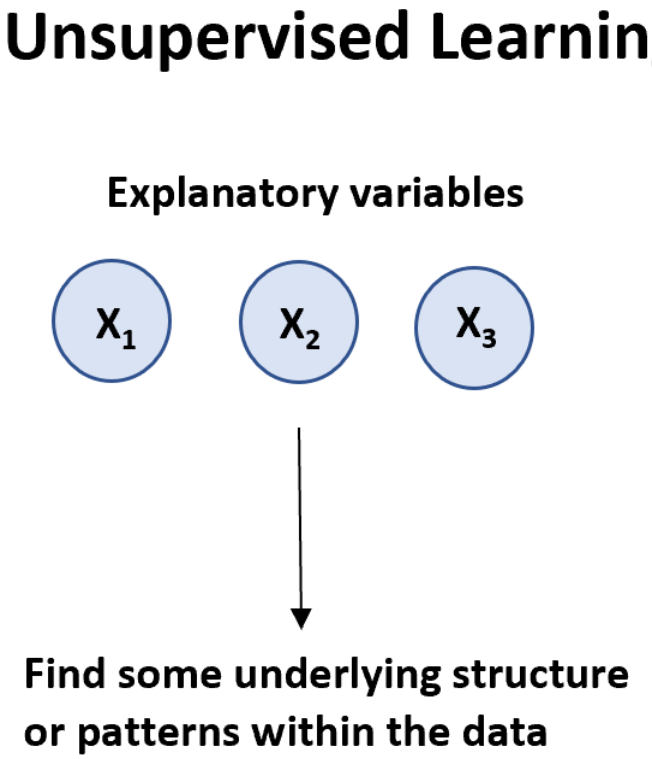
Depending on whether our goal is inference or prediction (or a mix of both), we may use different methods for estimating the function f. For example, linear models offer easier interpretation but non-linear models that are difficult to interpret may offer more accurate prediction.

Here is a list of the most commonly used **supervised** learning algorithms:

- **Linear regression**
- Logistic regression
- Linear discriminant analysis
- Quadratic discriminant analysis
- Decision trees
- Naive bayes
- Support vector machines
- Neural networks

Unsupervised Learning Algorithms

An **unsupervised learning algorithm** can be used when we have a list of variables (X_1, X_2, X_3, …, X_p) and we would simply like to find underlying structure or patterns within the data.



There are two main types of unsupervised learning algorithms:

1.Clustering: Using these types of algorithms, we attempt to find “clusters” of observations in a dataset that are similar to each other.

2. Association: Using these types of algorithms, we attempt to find “rules” that can be used to draw associations. For example, if a patient has a high biomarker X, he will have a low biomarker Y.

Here is a list of the most commonly used unsupervised learning algorithms:

- **Principal component analysis**
- K-means clustering
- K-medoids clustering
- **Hierarchical clustering**
- Apriori algorithm

Summary: Supervised vs. Unsupervised Learning

The following table summarizes the differences between supervised and unsupervised learning algorithms:

	Supervised Learning	Unsupervised Learning
Description	Involves building a model to estimate or predict an output based on one or more inputs.	Involves finding structure and relationships from inputs. There is no "supervising" output.
Variables	Explanatory and Response variables	Explanatory variables only
End goal	Develop model to **(1)** predict new values or **(2)** understand existing relationship between explanatory and response variables	Develop model to **(1)** place observations from a dataset into a specific cluster or to **(2)** create rules to identify associations between variables.
Types of algorithms	**(1)** Regression and **(2)** Classification	**(1)** Clustering and **(2)** Association

Linear regression

Supervised regression

Linear /multiple linear models

Supervised classification

Discriminant analysis

Unsupervised clustering

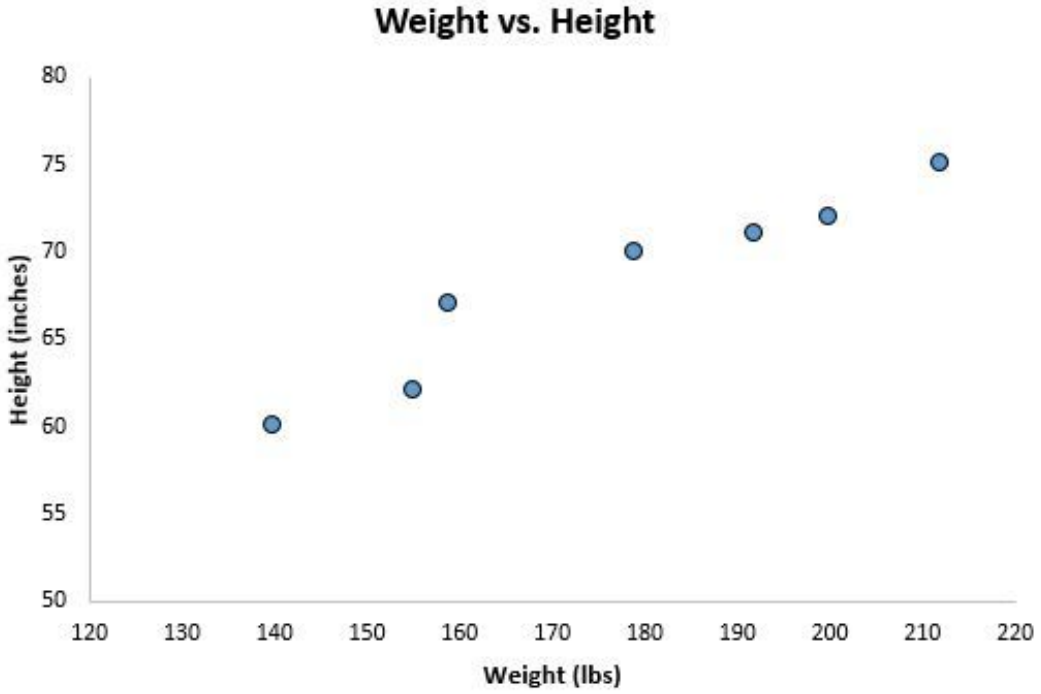
Hierarchical clustering

Unsupervised association

PCA

Simple linear regression is a statistical method you can use to understand the relationship between two variables, x and y.
One variable, **x**, is known as the **predictor variable**.
The other variable, **y**, is known as the **response variable**.

Weight (lbs)	Height (inches)
140	60
155	62
159	67
179	70
192	71
200	72
212	75



The formula for the line of best fit is written as:

$$\hat{y} = b_0 + b_1 x$$

where $\hat{y}$ is the predicted value of the response variable, b_0 is the y-intercept, b_1 is the regression coefficient, and x is the value of the predictor variable.

least squares regression line:

$$\hat{y} = 32.7830 + 0.2001x$$

b_0 = 32.7830. This means when the predictor variable *weight* is zero pounds, the predicted height is 32.7830 inches. Sometimes the value for b_0 can be useful to know, but not in this specific example

b_1 = 0.2001. This means that a one unit increase in *x* is associated with a 0.2001 unit increase in *y*. In this case, a one pound increase in weight is associated with a 0.2001 inch increase in height.

The Coefficient of Determination

One way to measure how well the least squares regression line "fits" the data is using the **coefficient of determination**, denoted as R^2.

The coefficient of determination is the proportion of the variance in the response variable that can be explained by the predictor variable.
The coefficient of determination can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all.
A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

An R^2 between 0 and 1 indicates just how well the response variable can be explained by the predictor variable.
For example, an R^2 of 0.2 indicates that 20% of the variance in the response variable can be explained by the predictor variable; an R^2 of 0.77 indicates that 77% of the variance in the response variable can be explained by the predictor variable.

Assumptions of Linear Regression

For the results of a linear regression model to be valid and reliable, we need to check that the following four assumptions are met:

1.Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.

2. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

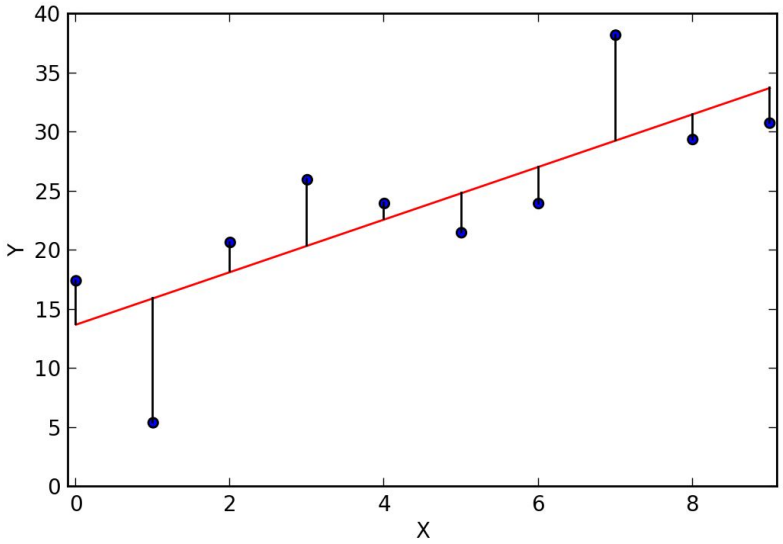
This is mostly relevant when working with time series data. Ideally, we don't want there to be a pattern among consecutive residuals.

A **residual** is the vertical distance between a data point and the regression line.

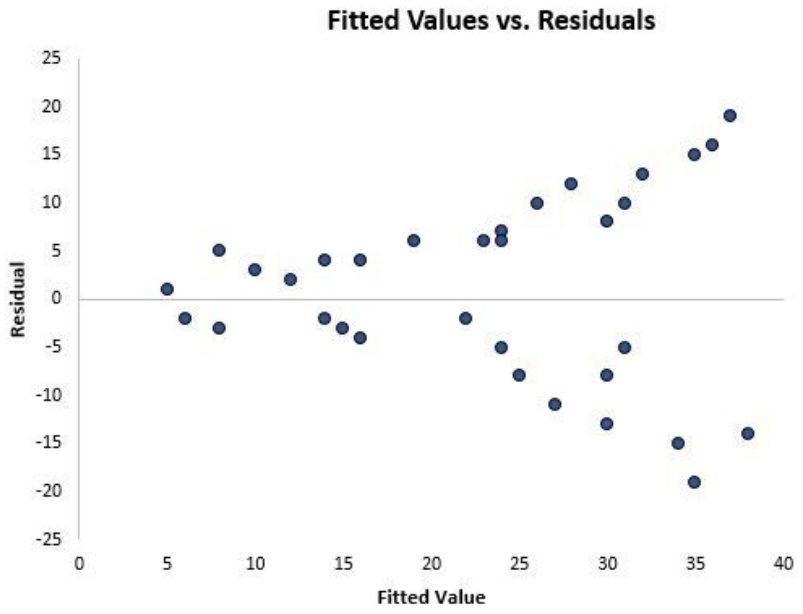
$y_i - \hat{y}_i$

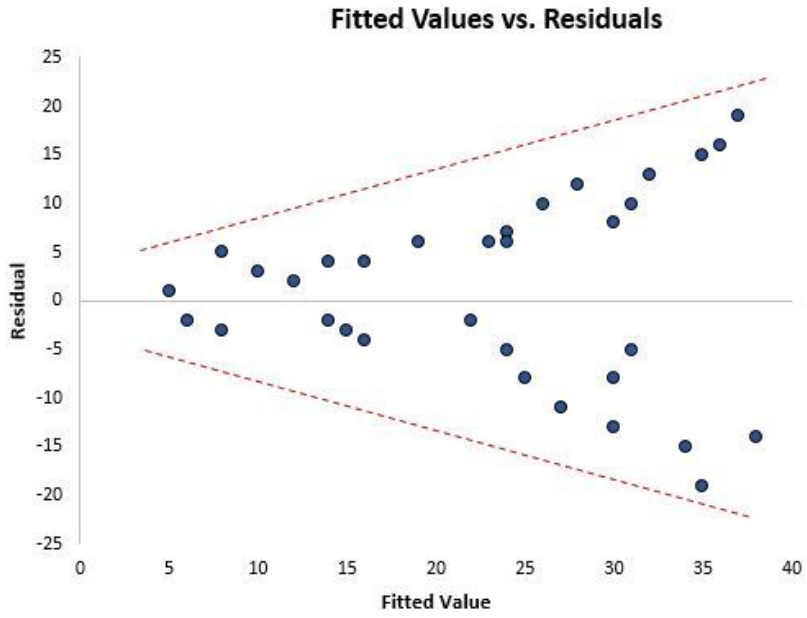
y_i: The actual response value for the i^{th} observation

$\hat{y}_i$: The predicted response value based on the multiple linear regression model



3. Homoscedasticity: The residuals have constant variance at every level of x.





3. Homoscedasticity: The residuals have constant variance at every level of x.

I verify the homogeneity of the variances (homoskedasticity) of the residuals with the Breusch Pagan test implemented in the bptest () command of the package lmtest

```
install.packages("lmtest")
library(lmtest)

bptest(model)

#Result
studentized Breusch-Pagan test

data:  model
BP = ..., df = 1, p-value = ...
```

If p-value > 0.05 I consider respected the assumption of Homoscedasticity

4. Normality: The residuals of the model are normally distributed.
Check normality (OF RESIDUALS) with the known methods (QQplot, Shapiro-Wilk, Kolmogorov Smirnov)

Multiple Linear Regression

to understand the relationship between *multiple* predictor variables and a response variable

a multiple linear regression model takes the form:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon}$$

where:

- **Y**: The response variable
- $\mathbf{X_j}$: The j^{th} predictor variable
- $\boldsymbol{\beta_j}$: The average effect on Y of a one unit increase in X_j, holding all other predictors fixed
- **ε**: The error term

The values for β_0, β_1, B_2, … , β_p are chosen using **the least square method**, which minimizes the sum of squared residuals (RSS):

$$\mathbf{RSS = \Sigma(y_i - \hat{y}_i)^2}$$

where:

Σ: A greek symbol that means *sum*

$\mathbf{y_i}$: The actual response value for the i^{th} observation

$\hat{\mathbf{y}}_i$: The predicted response value based on the multiple linear regression model

Multiple linear regression

D	E	F	G	H	I	J	K
SUMMARY OUTPUT							
Regression Statistics							
Multiple R	0.857						
R Square	0.734						
Adjusted R Square	0.703						
Standard Error	5.366						
Observations	20						
ANOVA							
	df	*SS*	*MS*	*F*	*Significance F*		
Regression	2	1350.76	675.38	23.46	0.00		
Residual	17	489.44	28.79				
Total	19	1840.20					

Regression SS is the total variation in the dependent variable that is explained by the regression model.

$\sum (\hat{y} - \bar{y})^2$

Residual SS — is the total variation in the dependent variable that is left unexplained by the regression model.

$\sum (y - \hat{y})^2$

Total SS — is the sum of both, regression and residual SS

D	E	F	G	H	I	J	K
SUMMARY OUTPUT							
Regression Statistics							
Multiple R	0.857						
R Square	0.734						
Adjusted R Square	0.703						
Standard Error	5.366						
Observations	20						
ANOVA							
	df	*SS*	*MS*	*F*	*Significance F*		
Regression	2	1350.76	675.38	23.46	0.00		
Residual	17	489.44	28.79				
Total	19	1840.20					

Mean Squared Errors (MS) — are the mean of the sum of squares or the sum of squares divided by the degrees of freedom for both, regression and residuals.
Regression MS = $\sum (\hat{y} - \bar{y})^2$/Reg. dfResidual MS = $\sum (y - \hat{y})^2$/Res. df

F — is used to test the hypothesis that the slope of the independent variable is zero. ($\beta j = 0$ means the explanatory variable is useless)
Mathematically, it can also be calculated as
F = Regression MS / Residual MS

Significance F — is nothing but the p-value for the null hypothesis that the coefficient of the independent variable is zero

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	67.67	2.82	24.03	0.00	61.73	73.61
hours	5.56	0.90	6.18	0.00	3.66	7.45
prep_exams	-0.60	0.91	-0.66	0.52	-2.53	1.33

Predictors parameters

A p-value below 0.05 indicates 95% confidence that the slope of the regression line is not zero and hence there is a significant linear relationship between the dependent and independent variables.

A p-value greater than 0.05 indicates that the slope of the regression line may be zero and that there is not sufficient evidence at the 95% confidence level that a significant linear relationship exists between the dependent and independent variables.

Regression Statistics	
Multiple R	0.810350635
R Square	0.656668152
Adjusted R Square	0.655978731
Standard Error	0.082783542
Observations	500

R^2 (R Square) — represents the power of a model. It shows the amount of variation in the dependent variable the independent variable explains and always lies between values 0 and 1. As the R^2 increases, more variation in the data is explained by the model and better the model gets at prediction. A low R^2 would indicate that the model doesn't fit the data well and that an independent variable doesn't explain the variation in the dependent variable well.

R^2 = Regression Sum of Squares/Total Sum of Squares

Adjusted R^2 — is R^2 multiplied by an adjustment factor. This is used while comparing different regression models with different independent variables. This number comes in handy while deciding on the right independent variables in multiple regression models.

Standard Error —This is the estimated standard deviation of the error of the regression equation and is a good measure of the accuracy of the regression line. It is the square root of the residual mean squared errors.

Std. Error = √(Res.MS)

Supervised regression
Linear /multiple linear models

Supervised classification
Discriminant analysis

Unsupervised clustering
Hierarchical clustering

Unsupervised association
PCA