Subjective Questions

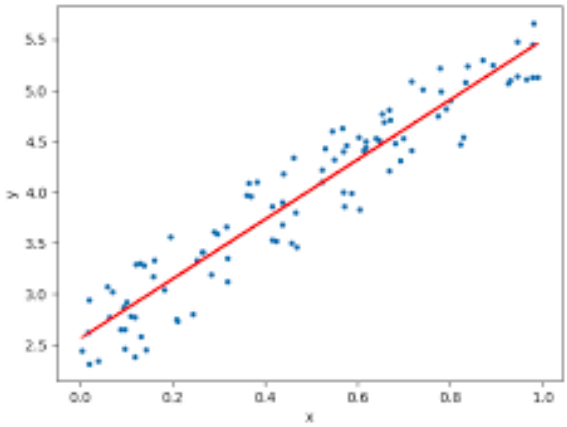
1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression is used to evaluate cause and effect relationships. In these relationships, you are predicting an outcome, or dependent, variable (the "effect") from one or more predictors, or independent, variables ("the cause").

Regression can take many forms, but the most common is least squares linear regression. In least squares regression, the line minimizes the sum of squared differences between the outcome values in the data and the outcome values that are predicted based on the regression equation that is estimated based on the data. Regression is a function that estimates the best fitting line through a set of data. In its simplest form, it looks like this:

$$y = \beta_0 + \beta_1 x$$

we know that β_0 is the y-intercept of the line, and β_1 is the slope, or the average change in the dependent variable for a unit change in the independent variable.



where β_0 is a constant, β_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable. The sample regression line is:

$$\hat{y} = \beta_0 + \beta_1 x$$

2. What are the assumptions of linear regression regarding residuals?

In regression analysis, the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the residual (ϵ). Each data point has one residual.

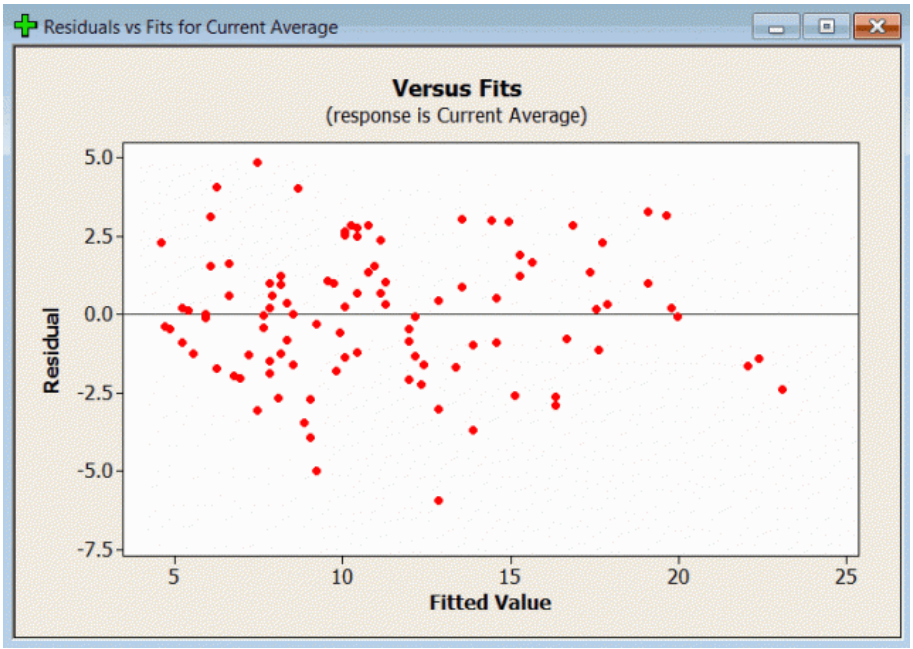
$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$

$$\epsilon = y - \hat{y}$$

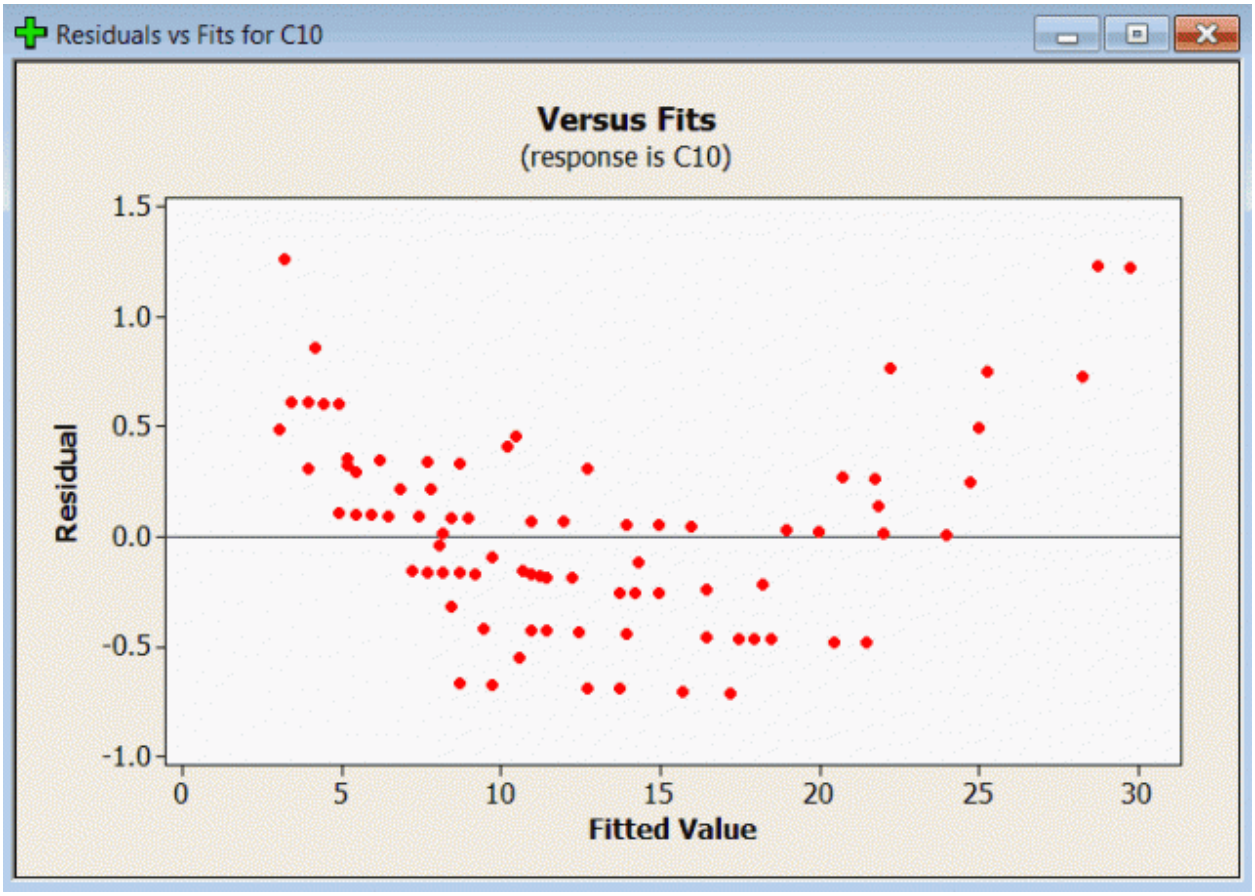
But there are assumptions your data must meet in order for the results to be valid. In this article, I'm going to focus on the assumptions that the error terms (or "residuals") have a mean of zero and constant variance.

When you run a regression analysis, the variance of the error terms must be constant, and they must have a mean of zero. If this isn't the case, your model may not be valid.

To check these assumptions, you should use a residuals versus fitted values plot. The errors have constant variance, with the residuals scattered randomly around zero. If, for example, the residuals increase or decrease with the fitted values in a pattern, the errors may not have constant variance.



The points on the plot above appear to be randomly scattered around zero, so assuming that the error terms have a mean of zero is reasonable. The vertical width of the scatter doesn't appear to increase or decrease across the fitted values, so we can assume that the variance in the error terms is constant.



There is definitely a noticeable pattern here! The residuals (error terms) take on positive values with small or large fitted values, and negative values in the middle. The width of the scatter seems consistent, but the points are not randomly scattered around the zero line from left to right. This graph tells us we should not use the regression model that produced these results.

---

3. What is the coefficient of correlation and the coefficient of determination?

In simple linear regression analysis, the coefficient of correlation (or correlation coefficient) is a statistic which indicates an association between the independent variable and the dependent variable. The coefficient of correlation is represented by "r" and it has a range of -1.00 to +1.00.

When the **coefficient of correlation** is a positive amount, such as +0.80, it means the dependent variable is increasing when the independent variable is increasing. It also means that the dependent variable is decreasing when the independent variable is decreasing. However, a high positive correlation does not guarantee there is a cause and effect relationship. (A negative amount indicates an inverse association...the dependent variable is decreasing when the independent variable is increasing and vice versa.)

A **coefficient of correlation** of +0.8 or -0.8 indicates a strong correlation between the independent variable and the dependent variable. An r of +0.20 or -0.20 indicates a weak correlation between the variables. When the coefficient of correlation is 0.00 there is no correlation.

The **coefficient of determination** is a statistic which indicates the percentage change in the amount of the dependent variable that is "explained by" the changes in the independent variables.

The **coefficient of determination** is symbolized by r-squared, where r is the coefficient of correlation. Hence, a coefficient of determination of 0.64 or 64% means that the coefficient of correlation was 0.8 or 80%. (The range for the coefficient of correlation is -1 to +1, and therefore the range for the coefficient of determination is 0 to +1.)

It is important to note that a high **coefficient of determination** does not guarantee that a cause-and-effect relationship exists. However, a cause-and-effect relationship between the independent variable and the dependent variable will result in a high coefficient of determination.

---

4. Explain the Anscombe's quartet in detail.

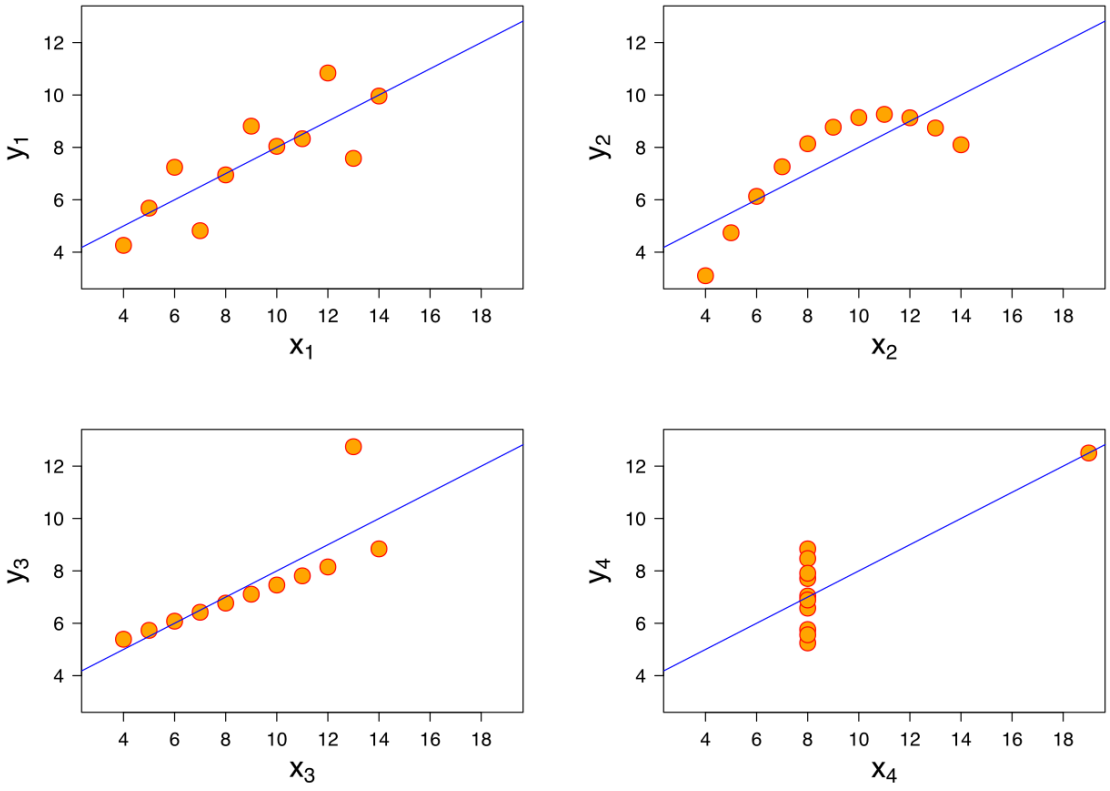
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups :

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

5. What is Pearson's R?

The correlation r measures the strength of the linear relationship between two quantitative variables.

Pearson's R:

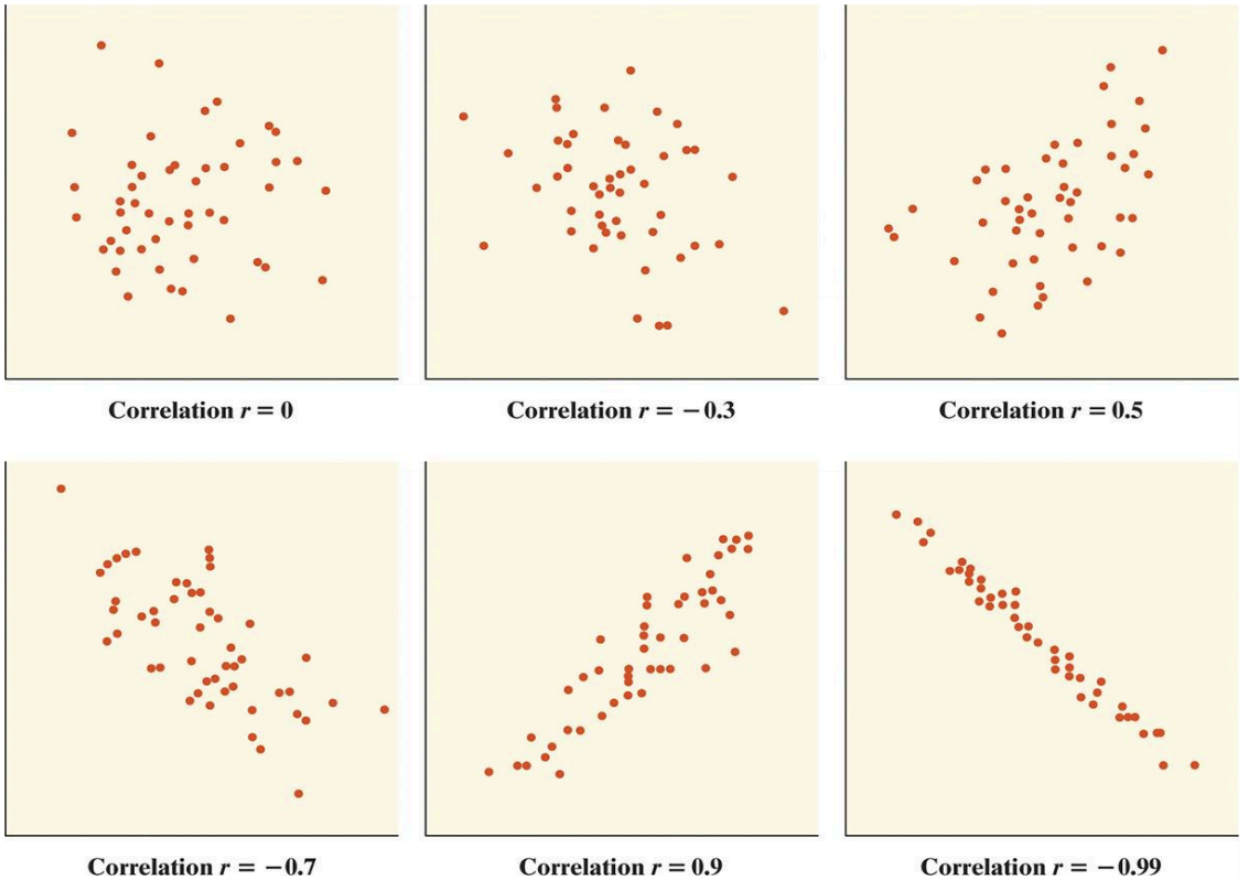
$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- r is always a number between -1 and 1.
- r > 0 indicates a positive association.
- r < 0 indicates a negative association.
- Values of r near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as r moves away from 0 toward -1 or 1.
- The extreme values r = -1 and r = 1 occur only in the case of a perfect linear relationship.

- Calculating a Pearson correlation coefficient requires the assumption that the relationship between the two variables is linear.
- There is a rule of thumb for interpreting the strength of a relationship based on its r value (use the absolute value of the r value to make all values positive):

Absolute Value of r	Strength of Relationship
r < 0.3	None or very weak
0.3 < r < 0.5	Weak
0.5< r< 0.7	Moderate
r>0.7	Strong

The relationship between two variables is generally considered strong when their r value is larger than 0.7.



- For a correlation coefficient of zero, the points have no direction, the shape is almost round, and a line does not fit to the points on the graph.
- As the correlation coefficient increases, the observations group closer together in a linear shape.
- The line is difficult to detect when the relationship is weak (e.g., r = -0.3), but becomes more clear as relationships become stronger (e.g., r = -0.99)

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. **Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).**

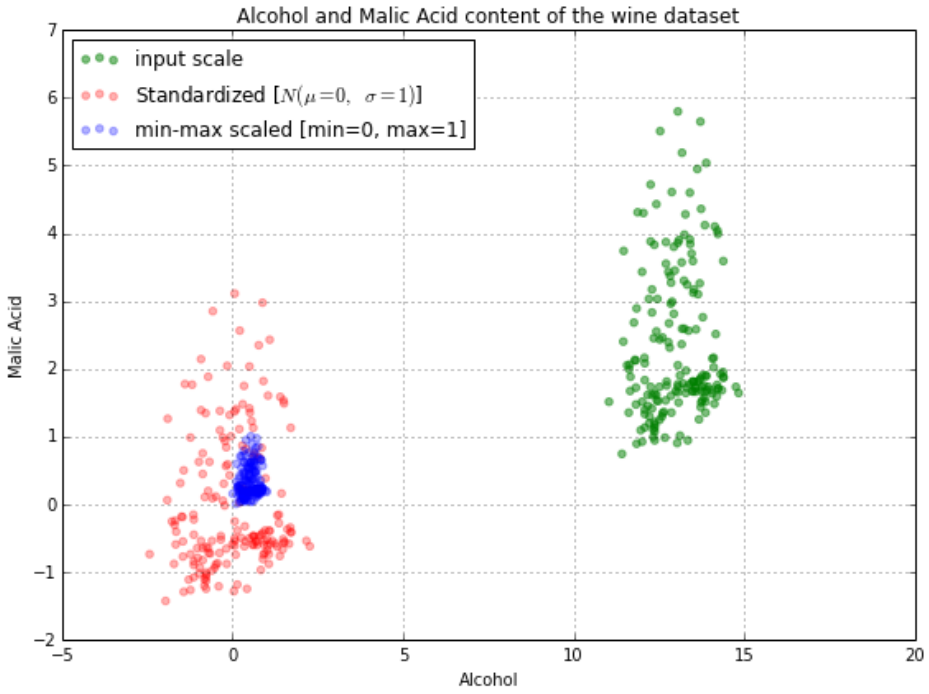
Two methods are usually well known for rescaling data. **Normalization, which scales all numeric variables in the range [0,1]**. One possible formula is given below:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

On the other hand, you can use standardization on your data set. It will then transform it to have zero mean and unit variance, for example using the equation below:

$$x_{new} = \frac{x - \mu}{\sigma}$$

Both of these techniques have their drawbacks. If you have outliers in your data set, normalizing your data will certainly scale the "normal" data to a very small interval. And generally, most of data sets have outliers. When using standardization, your new data aren't bounded (unlike normalization).



7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the squared multiple correlation of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes infinite.

The VIF can be conceived as related to the R-squared of a particular predictor variable regressed on all other includes predictor variables.:

$$VIF = \frac{1}{1 - R^2}$$

If you only have 1 for X or that X is orthogonal with all the other X's then $VIF = \frac{1}{1-0} = 1$ so no variance inflation

If two X's are perfectly correlated, then $\boldsymbol{VIF = \frac{1}{1-1} = \frac{1}{0} = \infty}$ that is the estimate is as imprecise as it can be

8. What is the Gauss-Markov theorem?

We start with estimation of the linear (in the parameters) model y = Xβ + ε,

where we assume that:

1. E(ε|X) = 0 for all X (mean independence)
2. V AR(ε|X) = E(εε0|X) = σ2εIN (homoskedasticity) The Gauss-Markov Theorem states that

$$\hat{\beta} = (X0X)-1X0y$$

is the Best Linear Unbiased Estimator (BLUE) if ε satisfies (1) and (2).

Proof: An estimator is "best" in a class if it has smaller variance than others estimators in the same class. We are restricting our search for estimators to the class of linear, unbiased ones. Since the data are the y (not the X), we are looking at estimators that are linear

where β is a k×1 parameter vector, m is a k×1 vector of constants, M is a k×n matrix of constants, and the data vector y is n × 1.

Second, we are restricting attention to the class of unbiased estimators, that is we require that

$$E(\tilde{\beta}) = \beta,$$

for any "valid" possible value β could take, i.e., for all β in the parameter space Ωβ.

First note that if $\tilde{\beta}$ is to be unbiased, then

$$E(\tilde{\beta}|X) = m + ME(y|X)$$

$$= m+ME(X\beta+\varepsilon|X)$$

$$= m+MX\beta+ME(\varepsilon|X) = m+MX\beta,$$

where the last line follows from the mean independence assumption. To be unbiased for any possible value of β then requires m=0

Now the matrix CC0 is a k×k "cross products" matrix, which by construction cannot be negative definite. The best estimator in a class of estimators is the one with the "smallest" covariance matrix, where by small we mean that the covariance matrix associated with any other estimator in the class (that is, linear and unbiased in the current context) minus the covariance matrix of the best estimator is a positive definite matrix. Formally, the matrix difference

MM0 − COV (best estimator)

is positive definite. Since MM0 is minimized when we set the matrix C equal to 0 (that is, it contains k × n 00s), the best estimator in the class $\hat{\beta}$. Any other estimator M in this class (in which the C matrix does not contain 00s in every row and column) has a strictly "larger" covariance matrix. We conclude that the OLS estimator $\hat{\beta}$ is BLUE under the two conditions set forth (mean independence and homoskedastic).

---

9. Explain the gradient descent algorithm in detail.

Gradient descent optimization algorithms, while increasingly popular, are often used as black-box optimizers, as practical explanations of their strengths and weaknesses are hard to come by. This article aims to provide the reader with intuitions with regard to the behavior of different algorithms that will allow her to put them to use. In the course of this overview, we look at different variants of gradient descent, summarize challenges, introduce the most common optimization algorithms, review architectures in a parallel and distributed setting, and investigate additional strategies for optimizing gradient descent.

They can be classified by two methods mainly:

- On the basis of data ingestion
    1. Full Batch Gradient Descent Algorithm
    2. Stochastic Gradient Descent Algorithm

In full batch gradient descent algorithms, you use whole data at once to compute the gradient, whereas in stochastic you take a sample while computing the gradient.

- On the basis of differentiation techniques
    1. First order Differentiation
    2. Second order Differentiation

Gradient descent requires calculation of gradient by differentiation of cost function. We can either use first order differentiation or second order differentiation.

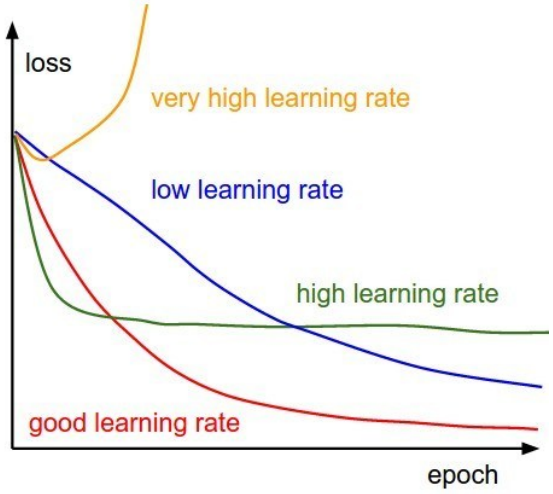
Vanilla Gradient Descent

This is the simplest form of gradient descent technique. Here, vanilla means pure / without any adulteration. Its main feature is that we take small steps in the direction of the minima by taking gradient of the cost function.

```
update = learning_rate * gradient_of_parameters


parameters = parameters - update
```

Here, we see that we make an update to the parameters by taking gradient of the parameters. And multiplying it by a learning rate, which is essentially a constant number suggesting how fast we want to go the minimum. Learning rate is a hyper-parameter and should be treated with care when choosing its value.



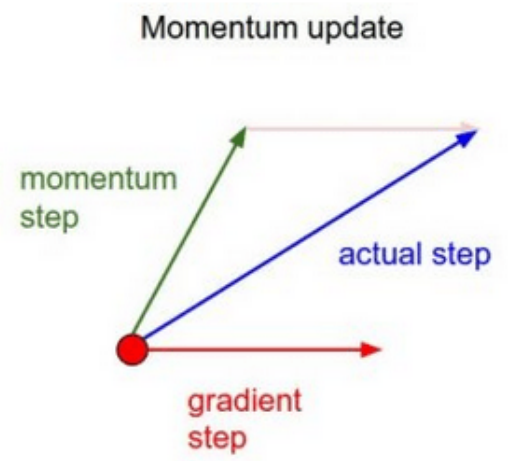
Gradient Descent with Momentum

our update is the same as that of vanilla gradient descent. But we introduce a new term called velocity, which considers the previous update and a constant which is called momentum.

```
update = learning_rate * gradient


velocity = previous_update * momentum


parameter = parameter + velocity - update
```

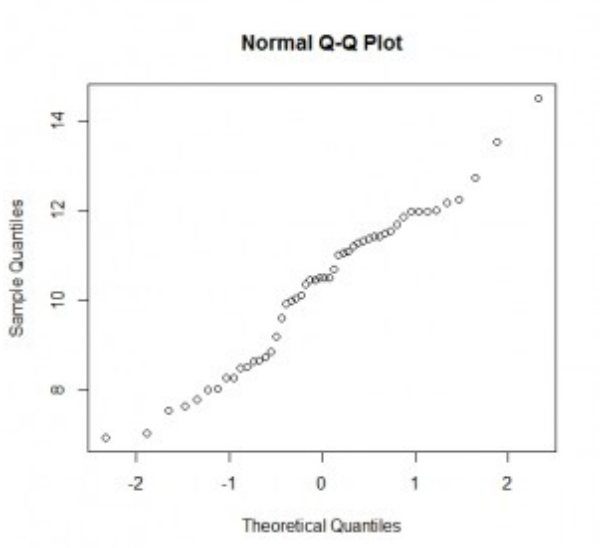


---

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Now what are "quantiles"? These are often referred to as "percentiles". These are points in your data below which a certain proportion of your data fall. For example, imagine the classic bell-curve standard Normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That's the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64. The following R code generates the quantiles for a standard Normal distribution from 0.01 to 0.99 by increments of 0.01