# Linear Regression for Machine Learning

by **Jason Brownlee** on March 25, 2016 in **Machine Learning Algorithms**



Last Updated on August 12, 2019

Linear regression is perhaps one of the most well known and well understood algorithms in statistics and machine learning.

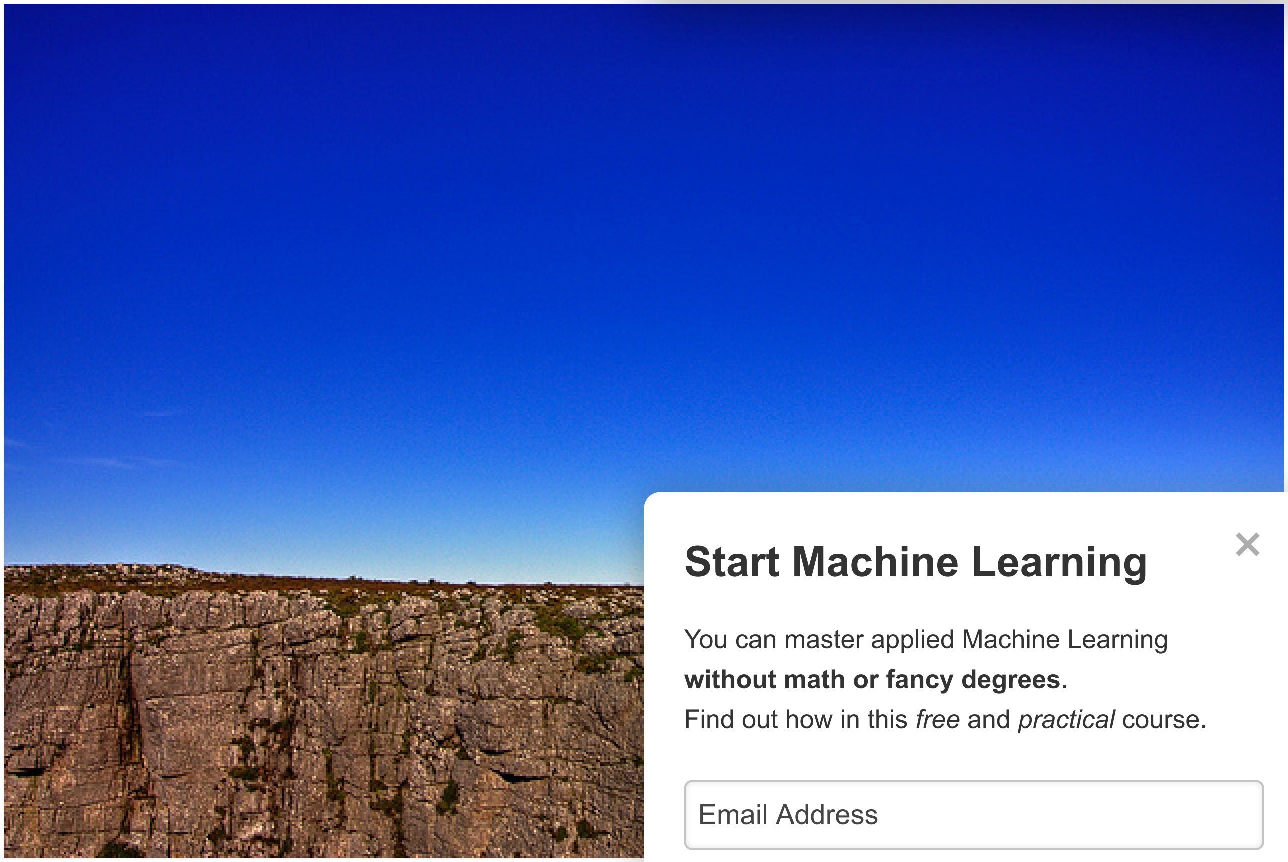
In this post you will discover the linear regression algorithm, how it works and how you can best use it in on your machine learning projects. In this post you will learn:

- Why linear regression belongs to both statistics and machine learning.
- The many names by which linear regression is known.
- The representation and learning algorithms used to create a linear regression model.
- How to best prepare your data when modeling using linear regression.

You do not need to know any statistics or linear algebra to understand linear regression. This is a gentle high-level introduction to the technique to give you enough background to be able to use it effectively on your own problems.

Discover how machine learning algorithms work including kNN, decision trees, naive bayes, SVM, ensembles and much more in my new book, with 22 tutorials and examples in excel.

Let's get started.



Linear Regression fo
Photo by Nicolas Raymor



## Isn’t Linear Regression from Statistics?

Before we dive into the details of linear regression, you may be asking yourself why we are looking at this algorithm.

Isn’t it a technique from statistics?

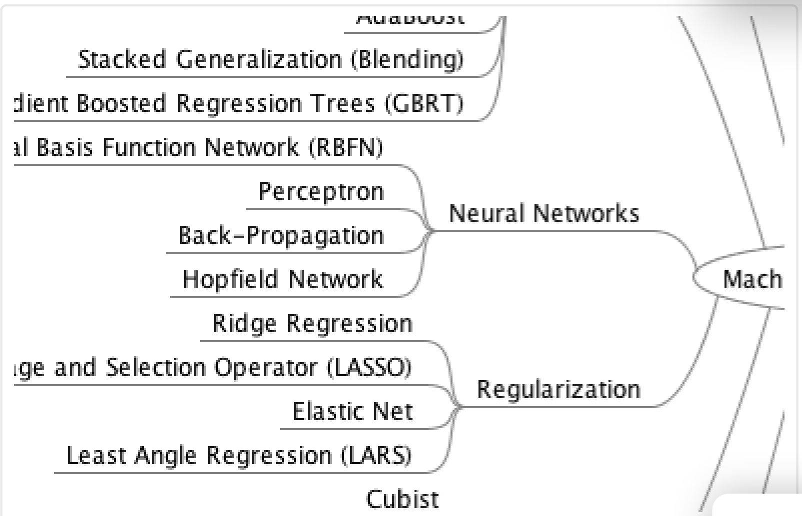
Machine learning, more specifically the field of predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. In applied machine learning we will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.

As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm.

Next, let’s review some of the common names used to refer to a linear regression model.

## Get your FREE Algorithms Mind Map

I've created a handy mind map of 60+ algorithms organized by type.




Sample of the handy machine learning algorithms mind map



When you start looking into linear regression, things c

The reason is because linear regression has been arc n studied from every possible angle and often each ang

Linear regression is a **linear model**, e.g. a model that variables (x) and the single output variable (y). More s combination of the input variables (x).

When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called **Ordinary Least Squares**. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.

Now that we know some names used to describe linear regression, let's take a closer look at the representation used.

## Linear Regression Model Representation

Linear regression is an attractive model because the representation is so simple.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1*x

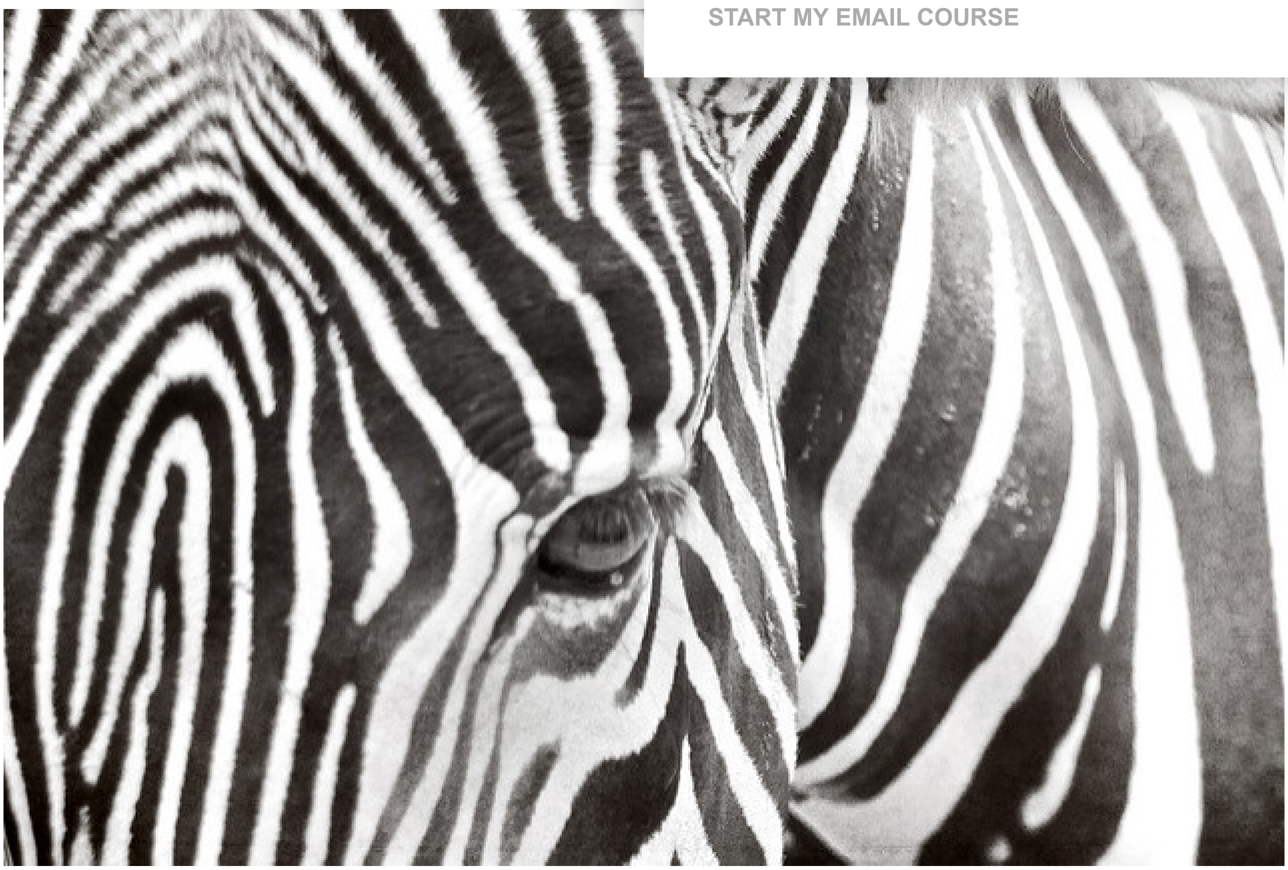
In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively remove
and therefore from the prediction made from the mode
regularization methods that change the learning algori s
by putting pressure on the absolute size of the coeffici



Now that we understand the representation used for a hat we can learn this representation from data.



What is Linear Regression?
Photo by Estitxu Carton, some rights reserved.

# Linear Regression Learning the Model

Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available.

In this section we will take a brief look at four techniques to prepare a linear regression model. This is not enough information to implement them from scratch, but enough to get a flavor of the computation and trade-offs involved.

There are many more techniques because the model is so well studied. Take note of Ordinary Least Squares because it is the most common method used in general. Also take note of Gradient Descent as it is the most common technique taught in machine learning classes.

## 1. Simple Linear Regression

With simple linear regression when we have a single i
coefficients.

This requires that you calculate statistical properties fr
correlations and covariance. All of the data must be av

This is fun as an exercise in excel, but not really usefu



## 2. Ordinary Least Squares

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.

The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.

This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients. It means that all of the data must be available and you must have enough memory to fit the data and perform matrix operations.

It is unusual to implement the Ordinary Least Squares procedure yourself unless as an exercise in linear algebra. It is more likely that you will call a procedure in a linear algebra library. This procedure is very fast to calculate.

## 3. Gradient Descent

When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

This operation is called Gradient Descent and works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

When using this method, you must select a learning rate (alpha) parameter that determines the size of the improvement step to take on each iteration of the procedure.

Gradient descent is often taught using a linear regression model because it is relatively straightforward to understand. In practice, it is useful when you have a very large dataset either in the number of rows or the number of columns that may not fit into memory.

### 4. Regularization

There are extensions of the training of the linear mode … th minimize the sum of the squared error of the model or … ut also to reduce the complexity of the model (like the nu … s in the model).

Two popular examples of regularization procedures fo …

- Lasso Regression: where Ordinary Least Square… the coefficients (called L1 regularization).
- Ridge Regression: where Ordinary Least Squares … sum of the coefficients (called L2 regularization).



These methods are effective to use when there is collinearity in your input values and ordinary least squares would overfit the training data.

Now that you know some techniques to learn the coefficients in a linear regression model, let's look at how we can use a model to make predictions on new data.

## Making Predictions with Linear Regression

Given the representation is a linear equation, making predictions is as simple as solving the equation for a specific set of inputs.

Let's make this concrete with an example. Imagine we are predicting weight (y) from height (x). Our linear regression model representation for this problem would be:

y = B0 + B1 * x1

or

weight =B0 +B1 * height

Where B0 is the bias coefficient and B1 is the coefficient for the height column. We use a learning technique to find a good set of coefficient values. Once found, we can plug in different height values to predict the weight.

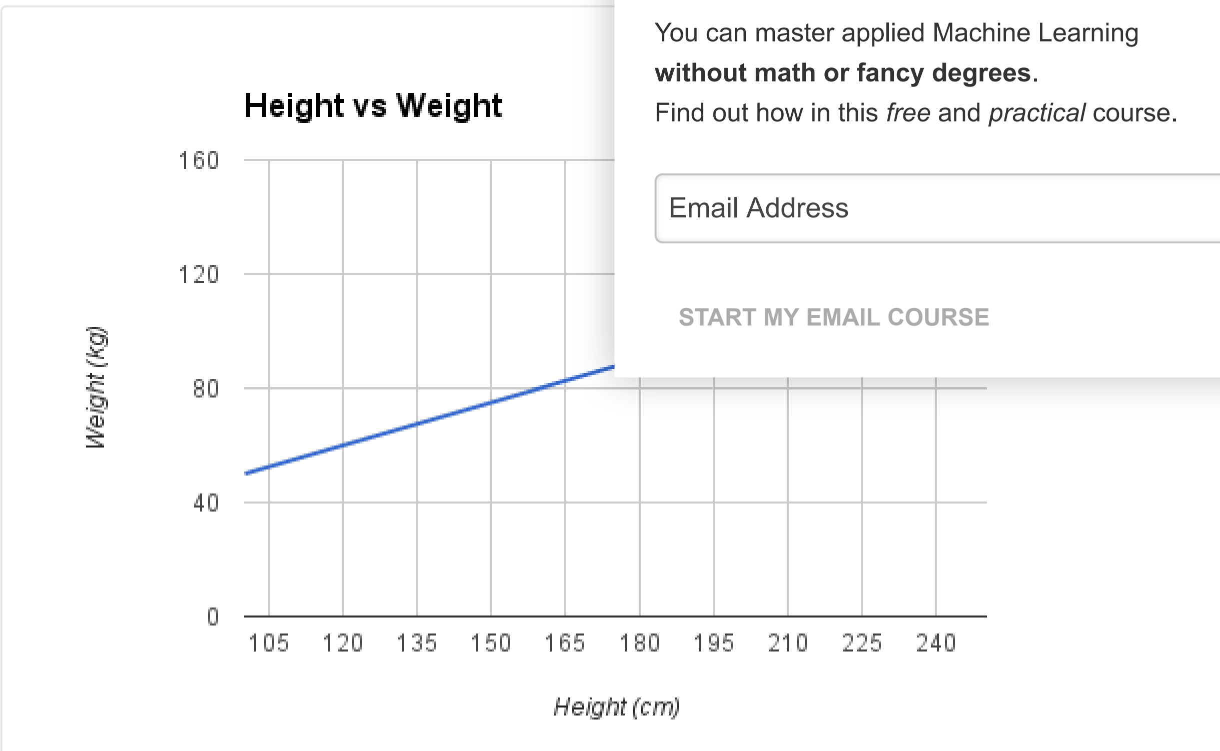
For example, lets use B0 = 0.1 and B1 = 0.5. Let's plug them in and calculate the weight (in kilograms) for a person with the height of 182 centimeters.

weight = 0.1 + 0.5 * 182

weight = 91.1

You can see that the above equation could be plotted as a line in two-dimensions. The B0 is our starting point regardless of what height we have. We can run t
centimeters and plug them to the equation and get we






Sample Height vs Weight Linear Regression

Now that we know how to make predictions given a learned linear regression model, let's look at some rules of thumb for preparing our data to make the most of this type of model.

# Preparing Data For Linear Regression

Linear regression is been studied at great length, and there is a lot of literature on how your data must be structured to make best use of the model.

As such, there is a lot of sophistication when talking about these requirements and expectations which can be intimidating. In practice, you can uses these rules more as rules of thumb when using Ordinary Least Squares Regression, the most common implementatic

Try different preparations of your data using these heuristics and see what works best for your problem.

- **Linear Assumption**. Linear regression assumes that the relationship between your input and output is linear. It does not support anything else. This may be obvious, but it is good to remember when you have a lot of attributes. You may need to transform data to make the relationship linear (e.g. log transform for an exponential relationship).
- **Remove Noise**. Linear regression assumes that your input and output variables are not noisy. Consider using data cleaning operations that let you better expose and clarify the signal in your data. This is most important for the output variable and you want to remove outliers in the output variable (y) if possible.
- **Remove Collinearity**. Linear regression will over-fit your data when you have highly correlated input variables. Consider calculating pairwise correlations for your input data and removing the most correlated.
- **Gaussian Distributions**. Linear regression will m put variables have a Gaussian distribution. You may g BoxCox) on you variables to make their distributio
- **Rescale Inputs**: Linear regression will often mak variables using standardization or normalization.

See the Wikipedia article on Linear Regression for an del. There's also a great list of assumptions on the Ordina



# Further Reading

There's plenty more out there to read on linear regression. Start using it before you do more reading, but when you want to dive deeper, below are some references you could use.

## Machine Learning Books that Mention Linear Regression

These are some machine learning books that you might own or have access to that describe linear regression in the context of machine learning.

- A First Course in Machine Learning, Chapter 1.
- An Introduction to Statistical Learning: with Applications in R, Chapter 3.
- Applied Predictive Modeling, Chapter 6.
- Machine Learning in Action, Chapter 8.
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Chapter 3.

## Posts on Linear Regression

Below are some interesting essays and blog posts on linear regression that I have come across.

- Ordinary Least Squares Regression: Explained Visually
- Ordinary Least Squares Linear Regression: Flaws, Problems and Pitfalls
- Introduction to linear regression analysis

- Four Assumptions Of Multiple Regression That Researchers Should Always Test

Know any more good references on linear regression with a bent towards machine learning and predictive modeling? Leave a comment and let me know.

## Summary

In this post you discovered the linear regression algorithm for machine learning.

You covered a lot of ground including:

- The common names used when describing linear regression models.
- The representation used by the model.
- Learning algorithms used to estimate the coefficie
- Rules of thumb to consider when preparing data f

Try out linear regression and get comfortable with it.

Do you have any questions about linear regression or
Leave a comment and ask, I will do my best to answe





### About Jason Brownlee

Jason Brownlee, PhD is a machine learning specialist who teaches developers how to get results with modern machine learning methods via hands-on tutorials.

View all posts by Jason Brownlee →

‹ Gradient Descent For Machine Learning

Simple Linear Regression Tutorial for Machine Learning ›

## 50 Responses to *Linear Regression for M*

**Amit P Jagtap** October 22, 2016 at 12:19 am #

HI
Thanks for good article.
I have a doubt about Linear regression hypothesis.
I feel in single variable linear regression equationY= W
W1*X term. If E > W1*X them it means other variables
Kindly, add and correct me if I am wrong.



**Jason Brownlee** October 22, 2016 at 7:00 am #

REPLY

Hi Amit, I’m not sure what you mean.

The simple linear regression equation is:

```
1 y = B0 + B1*x
```

**Prof Narain Sinha** January 8, 2019 at 8:03 pm #

REPLY

If E>w1*x it does not mean any thing.In fact E is assumed to take any value between – infinite and + infinity

**Jason Brownlee** January 9, 2019 at 8:43 am #

REPLY

Thanks.

**sagar dnyane** July 5, 2019 at 10:22 pm # REPLY

y=mx+c
linear regression equation .

**Nuwan C** May 3, 2017 at 11:53 am # REPLY

Hi Amith,Jason,
I think Amith trying to say that the ERROR regarding n linear regression is a part of linear equation?correct me ig I wrong



**Terry John** September 4, 2017 at 4:41 pm #

hi Jason
This is article is good
This help me to complete linear regression project in m
But still i feel bit confussing in linear algebra concept

**Jason Brownlee** September 7, 2017 at 12:33 pm # REPLY

Hang in there.

**Sureshbabu** September 7, 2017 at 8:54 pm # REPLY

Nice Explanation about Linear Regression.

**Jason Brownlee** September 9, 2017 at 11:43 am # REPLY

Thanks.

**Tom** September 12, 2017 at 1:40 pm # REPLY

Can someone please explain the time complexity for this algorithm?

**Joe** November 26, 2019 at 5:29 am # REPLY

The time complexity for training simple Linear regression is O(p^2n+p^3) and O(p) for predictions.

**Jason Brownlee** November 26, 2019 at 6:16 am # REPLY

Thanks for sharing Joe.

**luky** October 31, 2017 at 10:11 am #

Hi Jason, what if there is multiple values Y for
for each X is nonsense isn't it? Eg 10 different Y value
meaning only if there is only on Y value for each X, or
Because with multiple Y values you will never hit the c
average value than trying to do some magic with linear



**Jason Brownlee** October 31, 2017 at 2:50 pm

If your problem is to predict a class label, then you can use multi-label classification to predict multiple y values for a given X.

**luky** November 14, 2017 at 10:02 am # REPLY

Hi Jason, thank you for your reply. I just looked into Linear Regression a little bit more and now it is bit more clear to me. I have to improve my lacking mathematical and statistical (and of course also ML) skills. So it goes quite slowly :). When i was looking into linear equations recently i noticed there is same formula as here in LR (slope – intercept form) :). Quite surprising, but then the LR formula is more familiar to one. Thank you and best regards, and later i will look also into class label thing. Luky

**sonth** December 11, 2017 at 3:26 pm # REPLY

Thank you so much Jason. I was looking for linear regression applied on datasets in weka to get a clear understanding. Could you please let me know where I can find them like how you explained the boston housing prices dataset.

**Jason Brownlee** December 11, 2017 at 4:53 pm # REPLY

Here is an example:
https://machinelearningmastery.com/regression-machine-learning-tutorial-weka/

**Riccardo** December 16, 2017 at 6:22 am # REPLY

Thanks for this good article! Please, I need some more help with a project I'm doing at university: I have to apply (nothing too difficult, I'm not an expert) a machine learning algorithm to a financial dataset, using R. I chose a linear regression where the daily price of the asset is the y and daily Open/High/Low are the x. I just used the command lm to fit, analysed the r
variables are obviously correlated, and if I plot the orig
proceed like a straight line. Can I conclude there's a lir
Maybe it's obvious, but I asking cause I'm not sure all t
Thank you again, regard from Italy ☺



**Jason Brownlee** December 16, 2017 at 9:20

I have some help with time series here tha
https://machinelearningmastery.com/start-here/#tir

**Sam** January 11, 2018 at 9:01 am # REPLY

Hi,

I really love your articles, very comprehensive yet simple to understand. I'm trying to wrap my head around machine learning and i'm watching tutorials on regression. So my question is, with a given data set, before i build the model, should i be doing feature extraction – using either forward selection or backward elimination or bidirectional elimination. After I get the features, that's when i build the model, Ordinary least squares is used to build the model. Is my understanding correct? I'm looking for a sequence as to what is done first.

Thanks,
Sam

**Jason Brownlee** January 12, 2018 at 5:48 am # REPLY

Yes you can. Try with and without feature selection to ensure it gives a lift in skill.

**Dorukhan Sergin** March 9, 2018 at 3:51 am # REPLY

Hi Jason!

Thank you for the great article summarizing the major concepts. I just wanted to express my doubt about one thing. Are you sure that linear regression assumes Gaussian for the inputs? The inputs should only be giving information about the mean of the output distribution (which is the only Gaussian assumed). To express it in math terms:
Y = beta0 + beta1*X + eps
Since
eps ~ N(0,sigma)
Y|X ~ N(beta0 + beta1*X , sigma)

As you can see, there is no assumption on X. What ma population where X is sampled from, so that we can cl wide range of inputs.

What do you think?

Thanks for the comment!



**Jason Brownlee** March 9, 2018 at 6:27 am #

Yes, learn more here:
https://en.wikipedia.org/wiki/Linear_regression
https://en.wikipedia.org/wiki/Ordinary_least_squares

**Dorukhan Sergin** September 27, 2018 at 4:15 am # REPLY

Under the assumptions section of the first link…
"Weak exogeneity. This essentially means that the predictor variables x can be treated as fixed values, rather than random variables. This means, for example, that the predictor variables are assumed to be error-free—that is, not contaminated with measurement errors. Although this assumption is not realistic in many settings, dropping it leads to significantly more difficult errors-in-variables models."

**Jason Brownlee** September 27, 2018 at 6:05 am # REPLY

Nice.

**programmer** March 31, 2018 at 4:16 am # REPLY

Hi,

in regression problem:

If I fitted a line in my data, that means there is a linear relationship between the value Y and other features.

Now, What else we can conclude. is there is a possibility that the features that have the high weights could have similarity with the value Y?

**Trolliyama** April 25, 2018 at 4:25 am #

let's assume I have three features A, B and C
following hypothesis,
hypothesis = bias + A*W1 + B*W2 + C*W3 + A^2*W4
is the above hypothesis correct? or do I need to make
hypothesis = bias + A*W1 + B*W2 + C*W3 + A^2*W4
please do provide reason as to why one or both are co



**Jason Brownlee** April 25, 2018 at 6:38 am #

No, you have have a mix of normal and squared inputs.

I would recommend carefully experimenting to see what works best for your specific data.

**Miguel** April 25, 2018 at 11:27 pm # REPLY

Correct me if I am wrong.. but all the methods to train/create/make/fit a model to a data set have to do with minimizing the sum of square errors function?

method 1 I believe is also minimizing the SSE but using statistics.
method 2 is minimizing the SSE
method 3 is minimizing the SSE for multi variable functions
method 4 is minimizing the SSE with an additional constraint

am I incorrect?

**Miguel** April 25, 2018 at 11:28 pm # REPLY

method 1: https://en.wikipedia.org/wiki/Sir

**alv** May 2, 2018 at 8:54 pm # REPLY

following data for linear regression problem
X. Y
2104. 460
1416. 232
1534. 314
852. 178

Which of the following set if values will give minimum error from training sample
1. m=4,. c=300
2. m=5,. c=30

**Jason Brownlee** May 3, 2018 at 6:33 am #

Is this a homework question?



**vishal** July 1, 2018 at 6:58 pm #

which book is good to refer for learning the linear regression deeply??

**Jason Brownlee** July 2, 2018 at 6:23 am # REPLY

Any good textbook on machine learning.

I list some books here:
https://machinelearningmastery.com/faq/single-faq/what-other-machine-learning-books-do-you-recommend

**Asma Rahman** August 19, 2018 at 12:48 am # REPLY

How or From Where i can get the value of B0 and B1?I am from the beginner levels .

**Jason Brownlee** August 19, 2018 at 6:25 am # REPLY

An algorithm will estimate them, learn them from examples.

For example, an algorithm implemented and provic

**JG** October 2, 2018 at 1:56 am # REPLY

Hi jason,
Thanks.

"linear" regression word terminology is often misused (due to language issues). For example I can use (for sure) linear regression to approach dependent variable (Y) of multinomial independents variables (x). How? adding more layers and 'relu' activation of the output layers, I calculated; cubic, quadratic and some other polynomials (Y=x^3, or Y = x^2, etc.).
But in case you want to be more "orthodox" or canonical, you simply added more features to your inputs, such as X`2 (as a new input feature from your current
only mean that all the features are added (beta0 * x + l
can accomplish non-linear dependencies Y= f(x) deper
regards
thank you very much for all yours tutorials !



**Jason Brownlee** October 2, 2018 at 6:28 am

Adding non-linear transforms of input vari
It is still a weighted sum of inputs.

You can choose where the complexity is managed, in the transforms or in the model.

**JG** October 2, 2018 at 8:25 am # REPLY

OK. Thks

**kotrappa SIRBI** October 16, 2018 at 11:57 pm # REPLY

Nice article, Thank you so much, I am more interested in Machine Learning applications .Please give references like books or web links , Thank you

**Jason Brownlee** October 17, 2018 at 6:53 am # REPLY

I have many examples of applications. What are you looking for exactly?

**F.R.** November 28, 2018 at 1:13 am # REPLY

There is a mistake under "Making Predictions with Linear Regression".

There should be 0.5 instead of 0.05 in "weight = 0.1 + 0.05 * 182"

**Jason Brownlee** November 28, 2018 at 7:43 am # REPLY

Thanks, fixed.

**Ada** January 25, 2019 at 4:52 pm #

hey can you please guide me for # training da
which method is to be used??

**Jason Brownlee** January 26, 2019 at 6:08 am

Sorry, I don't understand, can you please



**Prateek Anand** June 16, 2019 at 7:19 pm # REPLY

Hey ,
i wanted to ask which data set is the best and which one is the worst for linear regression ? and also if you could suggest a book or some articles about similar theoretical information on other algorithms like logistic regression and SVM.
Thanks

**Jason Brownlee** June 17, 2019 at 8:20 am # REPLY

A dataset that has a linear relationship between inputs and outputs is a good fit for linear regression.

One that has a nonlinear relationship is probably a bad fit.

**Prashanth** September 6, 2019 at 1:24 am # REPLY

I have read the above article, it is good. Sir, How much knowledge one should have to implement Linear Regression algorithm from scratch?

**Jason Brownlee** September 6, 2019 at 5:05 am #

REPLY

No prior knowledge, go right ahead!

## Leave a Reply

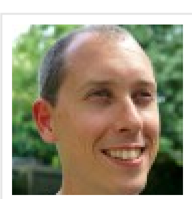


Name (required)

Email (will not be published) (required)

Website

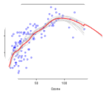
SUBMIT COMMENT



**Welcome!**
My name is *Jason Brownlee* PhD, and I **help developers** get results with **machine learning**.
Read more

**Never miss a tutorial:**

**Picked for you:**


[A Tour of Machine Learning Algorithms]


[Supervised and Unsupervised Machine Learning Algorithms]


[Logistic Regression Tutorial for Machine Learning]


[Simple Linear Regression Tutorial for Machine Lear


[Logistic Regression for Machine Learning]



**Loving the**

The Machine Learnir
is where I keep the

SEE WHAT'S INSIDE