**Regression**

Table of Contents

[Least Squares Regression 4](#_Toc63697883)

[Simple Regression 4](#_Toc63697884)

[Multiple Regression 5](#_Toc63697885)

[Polynomial Line Fitting 6](#_Toc63697886)

[Calculating Probabilities 9](#_Toc63697887)

[Hypothesis Testing 10](#_Toc63697888)

[10](#_Toc63697889)

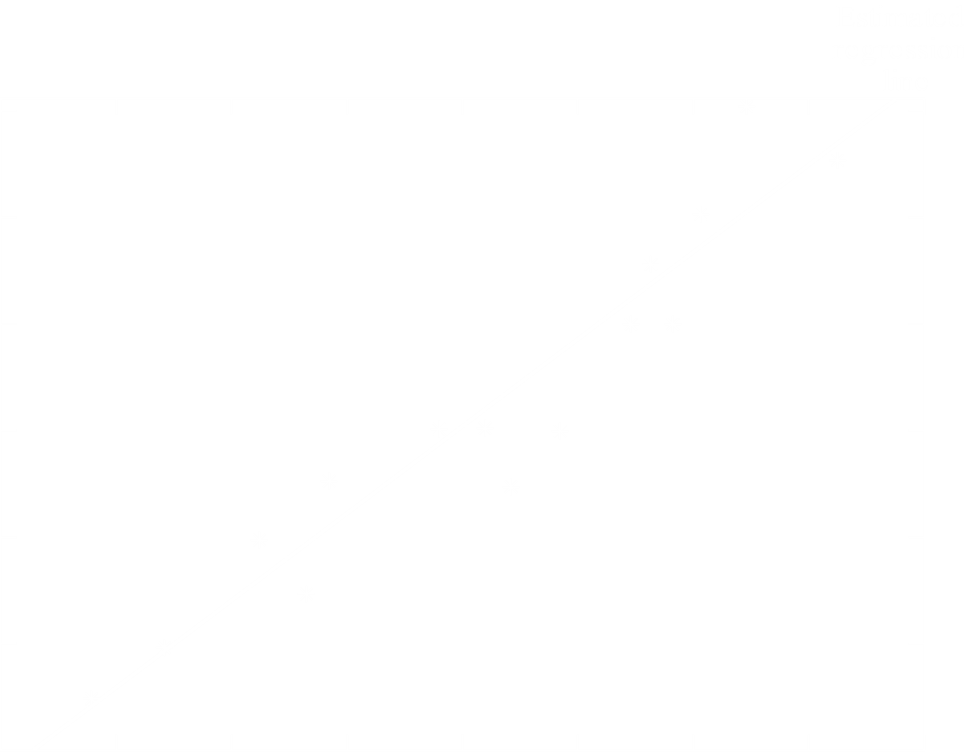
[Interval Estimation 13](#_Toc63697890)

[Working with 14](#_Toc63697891)

When collecting data, we will often find bivariate data, meaning data that comes in pairs. Say we have such pairs of data as . Here, is independent, meaning it can have any value it wants, and is dependent, meaning its value depends on the value of . Our goal is to find a function, , such that

We want to develop this function from the data we collected. Usually, this function is for a line, either a straight line or a polynomial line. The benefit of developing this function is that for a future value of , we can predict the value of .

Say we have a few data points. The easiest way for us to develop the function is to draw a line using the least square method.



Even if this line does not have any of the actual data points on it, it is still a good fit. That is why this process is also sometimes called line fitting. For any line, the points will deviate from the line by some distance. These distances are called the errors. The errors can be both positive or negative, which is why we square it. The purpose of the least square’s method is to find a line such that the sum of the squares of the errors is minimum.

For a specific line, the equation can be written as

We are only dealing with two values here, so for our purposes, and

This case, where , it is called simple regression. When , it is called multiple regression.

If the line is not straight, but polynomial, the equation will be

Even though the line itself can be polynomial, the models are still called linear models. This is because the variables in this case are not the values of and . Those values come from the input data. What varies are the values of , which we change in an attempt to find the best fitting line. Since the variable is the linear combination of the values of , this is called a linear model.

## Least Squares Regression

### Simple Regression

Let be a random variable that holds the value of the -th data point.

We know that a simple regression line has the equation

For a specific point on the regression line, let the corresponding -value be . Thus, the square of the error from the data point is

The sum of this value for all possible data points is sum of square errors, denoted as .

For least squares regression, our goal is to minimize this value. To do this, we differentiate the equation with respect to our actual variables, and , and set the results to to find the values of each that maximize the equation.

The value we will get for is given by

In many books, this equation is also written as

For the second equation, if we divide the numerator and the denominator by , the values become the covariance of and , , and the variance of , , respectively. Thus, is the ratio of to .

Remember that the covariance of and tells us how varies depending on . We are normalizing this value by dividing it by the covariance of .

From the value of , we can also find the value of .

### Multiple Regression

In multiple regression, is vector.

In this case, the equation for the regression line is

And the square sum will be

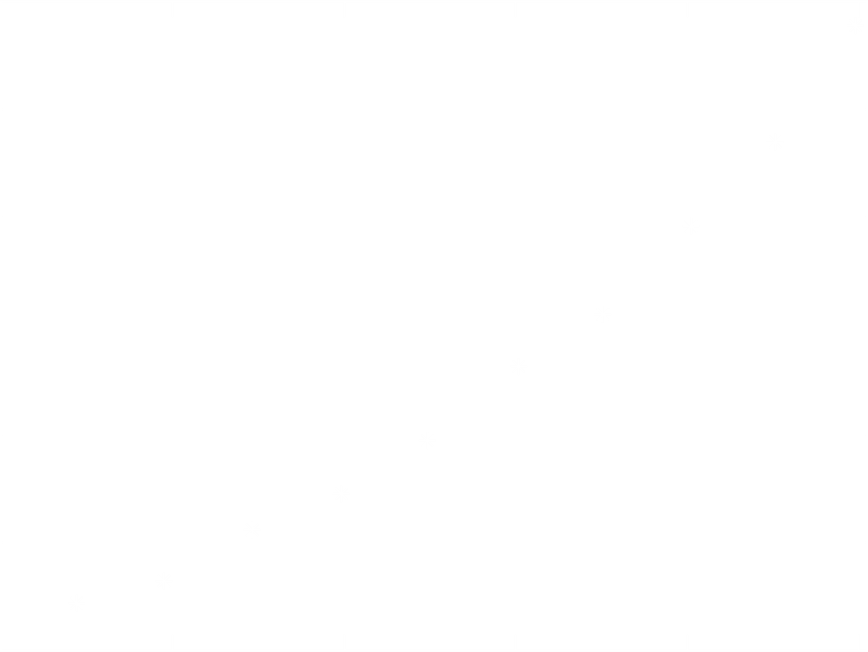
We have parameters here, and we need to find the derivatives of each of them.

If we differentiate these with respect to each of the parameters, we will find the required values. These equations are called Normal equations. Simultaneously solving them will give us the values.

One way to solve this is to represent the left-hand side of the equations as a matrix, ensuring that the determinant is non-zero.

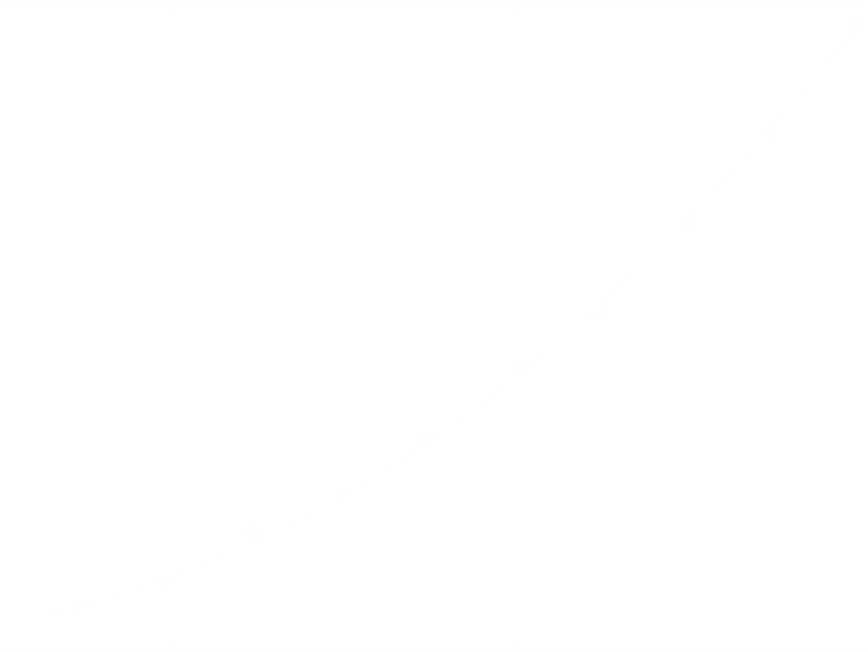
### Polynomial Line Fitting

Say we have data that is spread out like this:



In this case, if we try to find a straight line that fits the data well, we might have some difficulty.

A polynomial line on the other hand, will fit the data much better.



Another way to think about this is that the data was originally a sine wave, but got distorted by some noise, thus causing the errors.

For a polynomial line, the equation for the line of best fit is

And the corresponding square sum will be

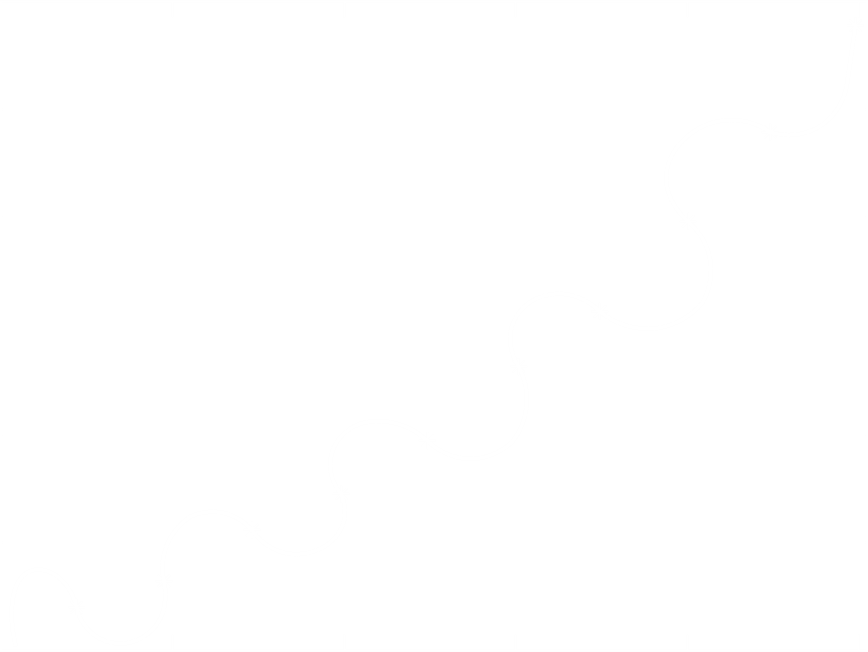
And as always, we can differentiate this to find the values of each parameter.

In the case where and we only have two variables, we will end up with a parabola.

Whenever we have the same number of parameters as the number of data points, it is possible to have a line that passes through all of the data points exact, with no errors. This means that we could end up with a line like this:



In reality though, if we actually tried to do this with a sine wave, we would get a rapidly oscillating wave that passes through all the data points at some point of time. It would look something like this:



Due to the rapid oscillations, the predicated values we will find will have a large error.

This phenomenon is called line overfitting. When we face this issue, we basically have to set some of the variables to to get rid of it. There are mathematical concepts like Regularization which can help us do this.

## Calculating Probabilities

For a set of random samples, , we can say that is a random variable. Thus, we can write the equation for simple regression as

For this, the predicated value can be found from

Here, is the error. We assume that .

is a constant, and adding a constant to a normal random variable does not change the distribution, so

Mathematically, it is possible to show from here that

## Hypothesis Testing

We shall now perform hypothesis testing on the least squares regression results we got.

The value of chosen here is not important. It can be any value.

Now, we need the test statistic and a null distribution. The test statistic must be defined in such a way that a null distribution for it exists.

In the last class, we saw that

It can be proven that (though we do not need to know how this can be proven). Looking at this, one might assume that we can just standardize this and use the standard normal distribution like we always have. But the problem is, we might not know . If we do not, then we would have to find the sample variance.

If we do have the value of , then by standardizing it we get

### 

We now need to define two terms. Firstly, the residual. This is the difference between the original value and the estimated value.

The square sum of the residual, or , is thus

From this, it can be proven that

Again, we do not need to know why this is true.

In the first lecture of statistics, we saw that

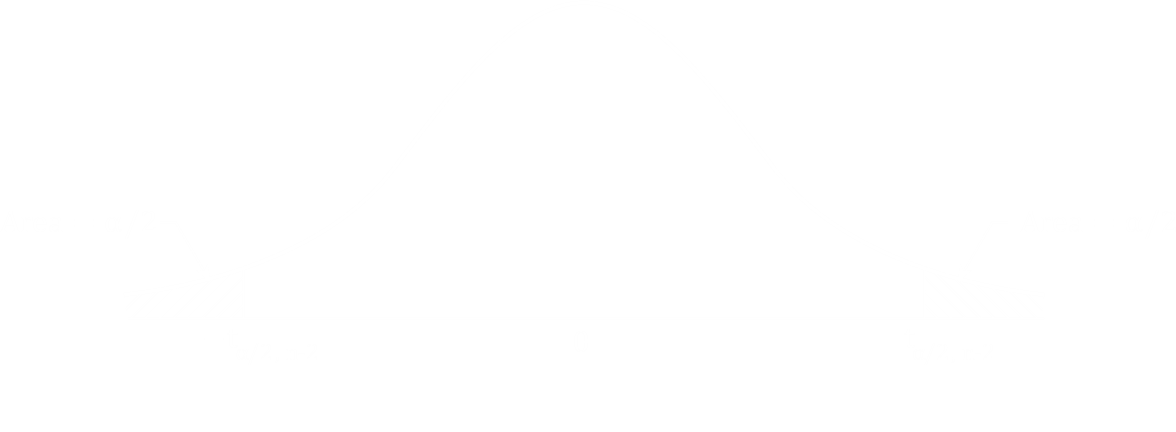
Extending on this,

Thus, we got rid of . Even if we do not know it, we can still perform hypothesis testing.

The test statistic is

And the null distribution is .

If we want to test the hypothesis at significance level, we will reject if the absolute value of the test statistic is beyond the rejection region, i.e. . We can of course get these values from the given table.



The -value is

-value

## Interval Estimation

Regardless of whether we use the least squares method or MLE to find the value of , if the random variables of the sample data were from a normal distribution, we will get the same value. However, this value is a point estimation. As we discussed earlier, if we use an interval estimation, we can have greater confidence in our answer. Specifically, for a significance level of , we will have a confidence level of .

If we now replace by , the final equation we will get is

Thus, the interval is

and the confidence level for this interval is .

## Working with

We can do similar things with .

Again, we now know the test statistics and the null distribution.

-value