Neural Networks for Signal Processing – I

Prof. Shayan Srinivasa Garani

Department of Electronic Systems Engineering

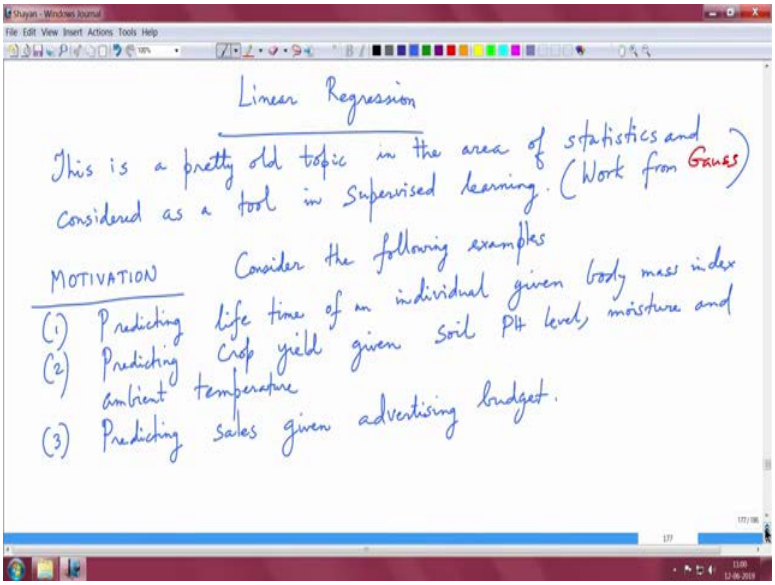
Indian Institute of Science, Bengaluru

Lecture – 12

Linear Regression 1

Welcome to this module, everyone! In this session, we will be delving into regression models. To begin with, we will explore Linear Regression for multiple variables. Following that, we will move on to Logistic Regression.

(Refer Slide Time: 00:45)



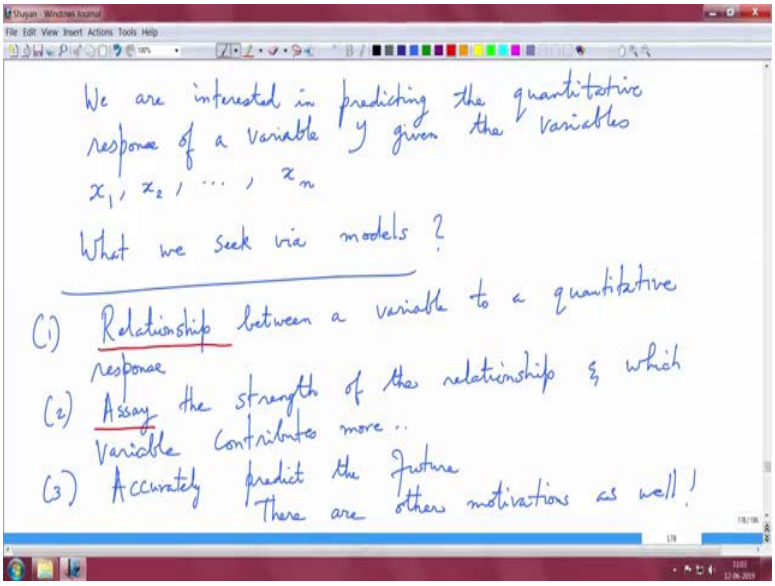
This is a well-established topic in statistics, widely regarded as a fundamental tool in supervised learning. The origins of linear regression can be traced back to the work of Carl Friedrich Gauss, who introduced the concept of least squares for achieving a linear fit.

Let's consider some motivating examples. Suppose we have a list of body mass indices (BMIs) for individuals, and we want to predict their lifespans. In another scenario, imagine we're working in agriculture and wish to predict crop yield based on various parameters such as soil pH level, moisture, and ambient temperature. Can we develop a model to forecast crop yield with this information?

Similarly, think about predicting sales given an advertising budget. We might have expenditures for radio, television, newspapers, and social media advertising. By considering all these parameters, can we accurately predict sales based on our advertising investments across various media platforms?

These examples illustrate common problems in statistics. Since we are also dealing with machine learning and neural networks, which are essentially subtopics within machine learning, it is crucial to understand the fundamental concepts behind regression to progress effectively.

(Refer Slide Time: 02:31)



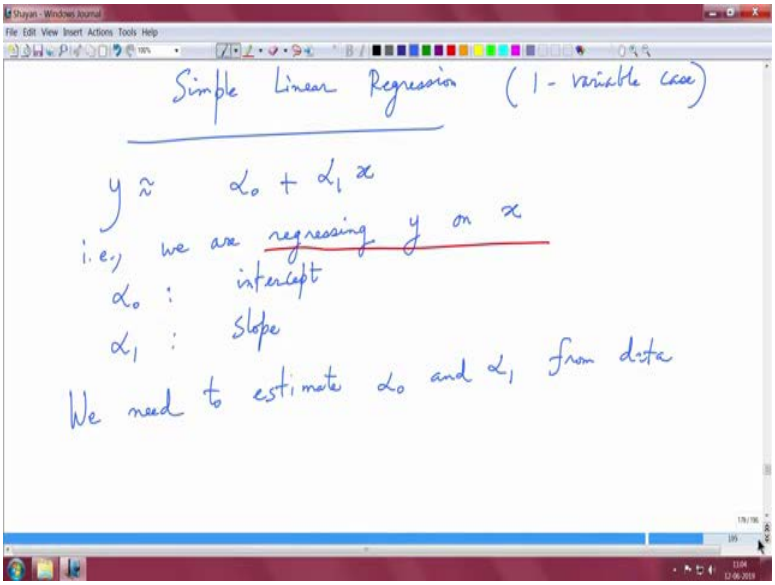
We are now interested in predicting the quantitative response of a variable y given the variables x_1, x_2, …, x_n. For example, in the crop yield scenario, x_1 might represent soil pH level, x_2 could be moisture content, and so on. Our goal is to predict the quantitative response of a specific quantity based on these variables. While this response is typically scalar, it can also be formulated as a vector. For simplicity, we will assume a scalar response for now.

What are we seeking within these models? First, we aim to identify the relationship between a variable and its quantitative response. This is crucial. Second, we want to assess the strength of this relationship and determine which variables are statistically significant. For instance, in the context of advertising, we might find that newspaper advertising has a stronger correlation with sales compared to television advertising, possibly because more

people read newspapers.

We need various indicators to evaluate the strength of the quantitative response, and our objective is to predict future outcomes accurately. This is essential, and there are additional motivations for addressing this problem.

(Refer Slide Time: 04:21)



Let's consider a simple linear regression problem involving one variable. In this case, y is approximately equal to $\alpha_0 + \alpha_1 x$. This is the relationship we aim to establish using our model, where we are regressing y on x. Here, α_0 represents the intercept, and α_1 is the slope. Essentially, this defines a linear relationship.

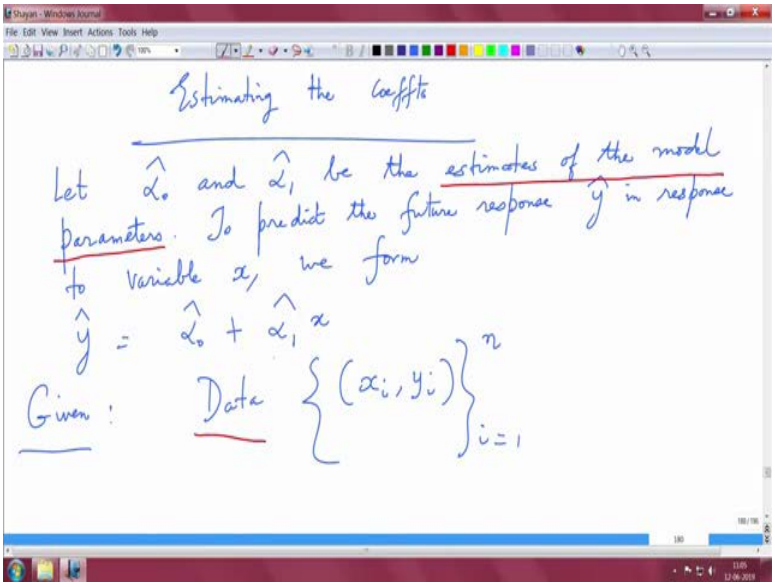
We need to estimate these parameters, α_0 and α_1, from the given data. It's important to note that these are approximations, which is why we use the approximation sign: $y \approx \alpha_0 + \alpha_1 x$. In reality, we do not know the exact nature of the relationship, but we keep the model simple by assuming a linear form.

Additionally, we should be aware that the data might be noisy, and we will address this issue subsequently.

Now, how do we estimate the coefficients? We start with the data, which consists of pairs (x_i, y_i) for i = 1 to n, giving us n data points. The parameters $\widehat{\alpha_0}$ and $\widehat{\alpha_1}$ are our estimates of the model coefficients. To predict the future response $\hat{y}$ for a given variable x, we use

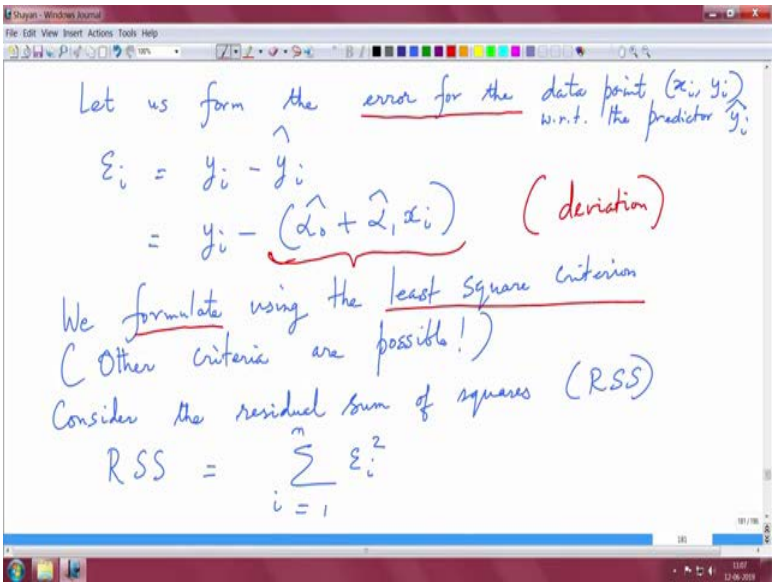
the formula $\hat{y} = \widehat{\alpha_0} + \widehat{\alpha_1} x$.

(Refer Slide Time: 05:44)



This formula serves as our predictor. Using this predictor, our goal is to minimize the deviation between the observed values y_i and the predicted values $\hat{y}_i$ according to our model. This minimization process ensures our model provides the best fit to the data.

(Refer Slide Time: 06:54)



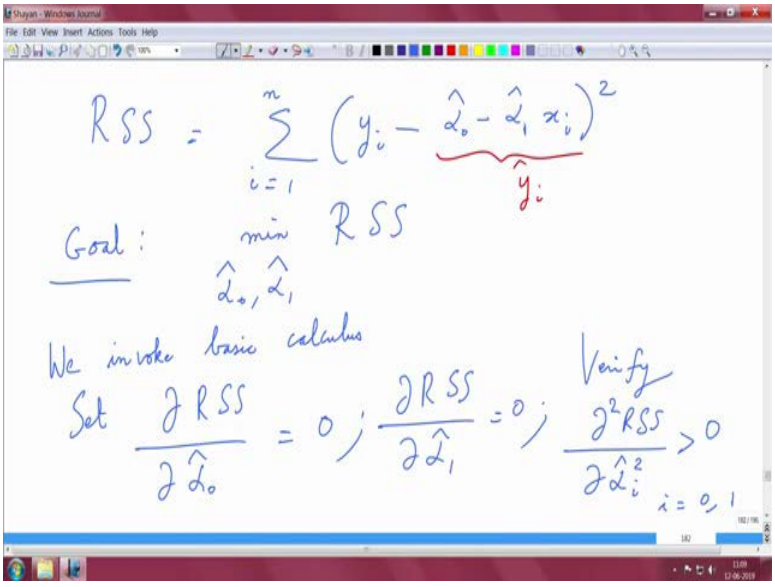
To advance further, we need to formulate the error for each data point (x_i, y_i) with respect to the predictor $\hat{y}_i$. Given the data points (x_i, y_i) for every x_i, we calculate the predictor $\hat{y}_i$

based on the linear relationship. The error, ϵ_i, which is y_i - $\hat{y}_i$, can be simplified as follows: $\widehat{y_i} = \widehat{\alpha_0} + \widehat{\alpha_1} x_i$. Thus, the error, $\epsilon_i = y_i - (\widehat{\alpha_0} + \widehat{\alpha_1} x_i)$, represents the deviation.

To fit the model parameters, we need to establish a criterion for the error. One of the most popular choices is the least squares criterion, where we consider the residual sum of squares. Specifically, we take the squared errors for i = 1 to n and sum them up. You might wonder why we use the square of the errors instead of the cube, the fourth power, or the absolute value.

We prefer quadratic optimization because it yields a unique solution and simplifies the mathematical handling. Therefore, considering these advantages, we stick to quadratic optimization for fitting our model parameters.

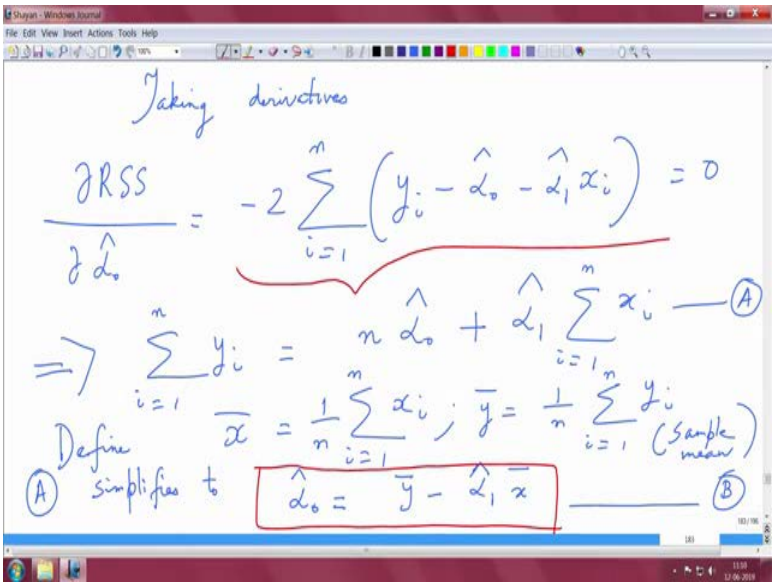
(Refer Slide Time: 09:11)



Additionally, one might consider using maximum likelihood estimates (MLE) if we have a parametric model for the data. Now, I formulate the residual sum of squares (RSS) according to this equation by plugging in $\widehat{y_i}$.

The goal is to minimize this residual sum of squares over the parameters $\hat{\alpha}_0$ and $\hat{\alpha}_1$. The procedure is straightforward: we invoke basic calculus, take the derivative of RSS with respect to $\hat{\alpha}_0$ and $\hat{\alpha}_1$, set these derivatives equal to zero, and then verify that the second derivatives with respect to these variables are positive to ensure we are indeed minimizing the RSS.

(Refer Slide Time: 10:21)



Once we formulate this, the process becomes straightforward. We take the derivatives and set them equal to zero. By taking the partial derivative of the RSS with respect to $\hat{\alpha}_0$, we arrive at the following equation. Since there is a square term, a negative sign appears before $\hat{\alpha}_0$.

Thus, we have -2 $\sum_{i=1}^{n}(y_i - \widehat{\alpha_0} - \widehat{\alpha_1}x_i) = 0$. Removing the -2 factor, we obtain the equation $\sum_{i=1}^{n} y_i = n\widehat{\alpha_0} + \widehat{\alpha_1}\sum_{i=1}^{n} x_i$. This is achieved by transposing the terms to ensure positive quantities on both sides of the equation.

Next, we define $\bar{x}$ and $\bar{y}$ as the sample means, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Using these definitions, we can simplify to find $\widehat{\alpha_0} = \bar{y} - \widehat{\alpha_1}\bar{x}$.

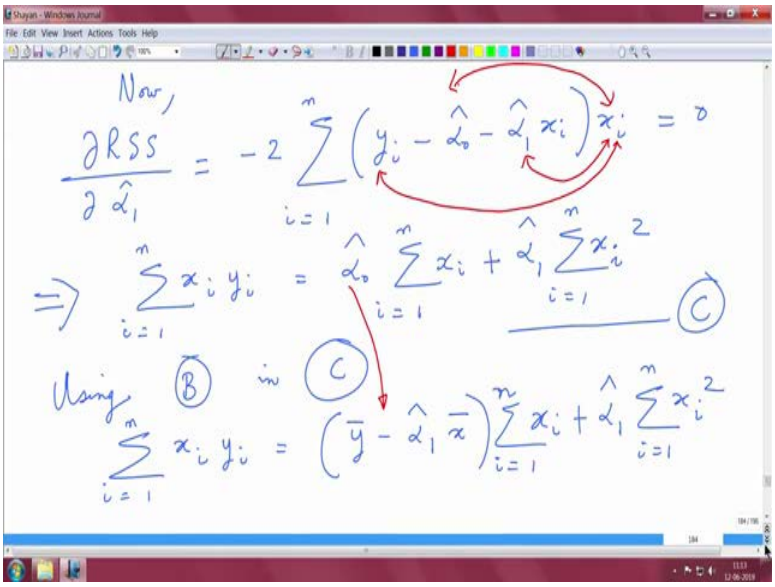
From this equation, we can compute $\hat{\alpha}_0$ and $\hat{\alpha}_1$ given the sample means $\bar{x}$ and $\bar{y}$, leading us to the final equations for our model parameters.

Once we grasp the method, we apply the same process to $\hat{\alpha}_1$. We take the partial derivative of the RSS with respect to $\hat{\alpha}_1$. Since there's a square term, we get a factor of 2, and due to the negative sign in front of $\hat{\alpha}_1$, there's also a minus sign.

Taking the partial derivative of α_1 x_i with respect to $\hat{\alpha}_1$, we get x_i. Thus, the expression simplifies to:

$$2\sum_{i=1}^{n} x_i(y_i - \widehat{\alpha_0} - \widehat{\alpha_1}x_i) = 0$$

(Refer Slide Time: 12:22)



Rearranging this equation, we factor out the -2 to get:

$$\sum_{i=1}^{n} x_i y_i = \widehat{\alpha_0}\sum_{i=1}^{n} x_i + \widehat{\alpha_1}\sum_{i=1}^{n} x_i^2$$

Using the previously derived relation $\widehat{\alpha_0} = \bar{y} - \widehat{\alpha_1}\bar{x}$ from equation B, we can further simplify equation C.

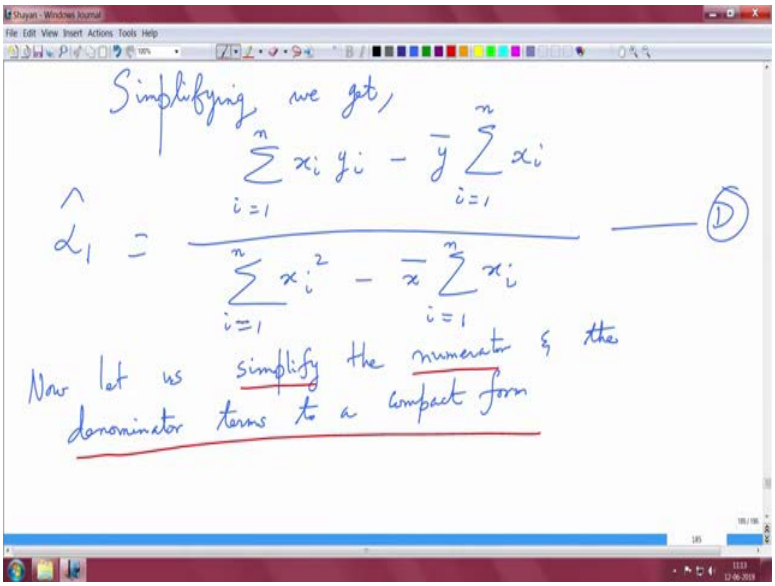
Substituting $\hat{\alpha}_0$ into the equation, we get:

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - \widehat{\alpha_1}\bar{x})\sum_{i=1}^{n} x_i + \widehat{\alpha_1}\sum_{i=1}^{n} x_i^2$$

This appears complex, so we will rearrange it by grouping terms involving $\hat{\alpha}_1$ and those that do not. Simplifying this will lead us to a more manageable form for $\hat{\alpha}_1$.

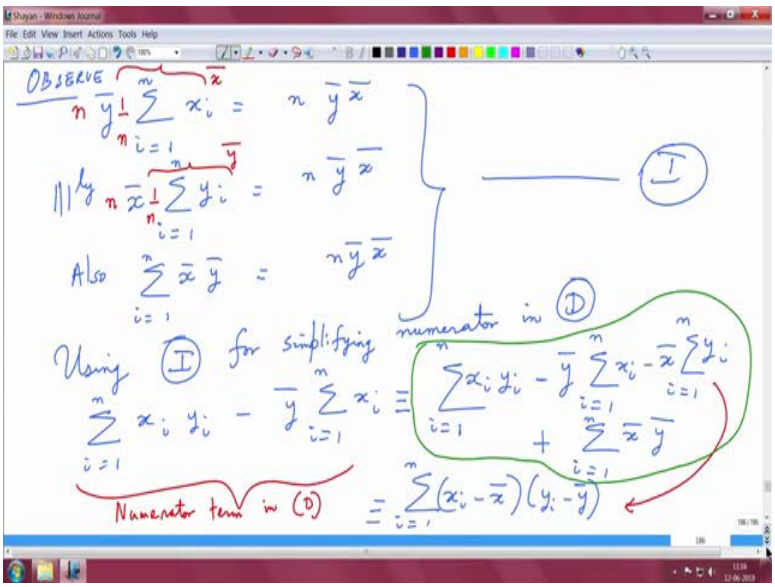
After some algebra and simplification, we arrive at the following equation. At first glance, it may not seem intuitive or easy to remember, as there isn't much apparent symmetry between the numerator and the denominator.

(Refer Slide Time: 14:32)



Let's try to express the numerator and the denominator in a more compact and simplified form.

(Refer Slide Time: 15:10)



We make the following observations, which are fairly straightforward: $\bar{y}\sum_{i=1}^{n} x_i$ simplifies to $n\bar{y}\bar{x}$. This is because if you multiply and then divide by n on the left side of the equation, you get n $\bar{y}$ times $\bar{x}$. Similarly, $\bar{x}\sum_{i=1}^{n} y_i$ also simplifies to $n\bar{y}\bar{x}$.

Using this same approach, $\frac{1}{n}\sum_{i=1}^{n} \bar{x}\,\bar{y}$ simplifies to n $\bar{y}$ $\bar{x}$ because there are n terms. All

these quantities are essentially equivalent. From these observations, which I will denote as equation 1, we can simplify the numerator of equation D.

To simplify, I add and subtract n $\bar{x}\,\bar{y}$ in a way that makes the expression more manageable. So, observe that the numerator term from equation D can be simplified by adding $\sum_{i=1}^{n} \bar{x}\,\bar{y}$ and subtracting $\bar{x}\sum_{i=1}^{n} y_i$.

Therefore, the numerator can be expressed as:

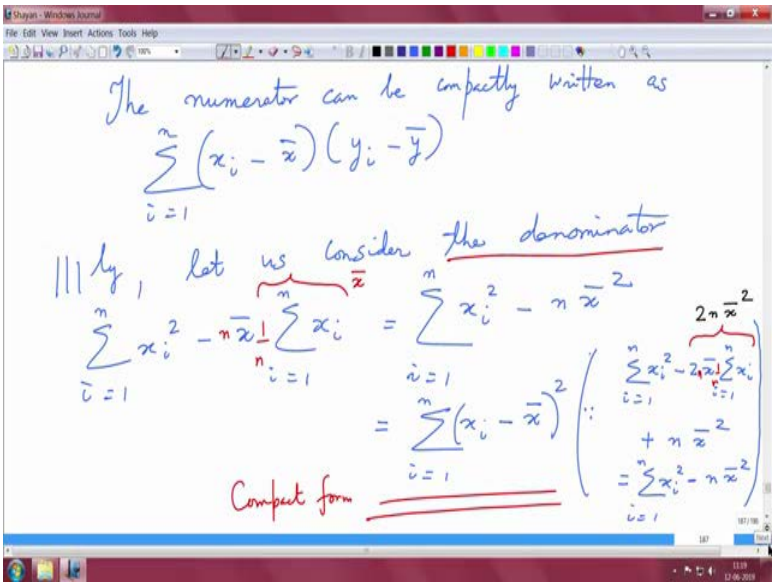
$$\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}$$

Simplifying this equation, we get:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

This compact form is much easier to work with. To use it, compute the sample means, subtract these means from each x_i and y_i, multiply the results, and then sum these products across all data points.

(Refer Slide Time: 18:15)



Now, let's simplify the denominator using a similar approach. Consider the denominator

term:

$$\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i$$

We can write this in a more compact form by recognizing that:

$$\sum_{i=1}^{n} x_i^2$$

The second term simplifies to n $\bar{x}^2$ because $\sum_{i=1}^{n} x_i$ is $n\bar{x}$, and we multiply by $\bar{x}$ again. To further simplify, we add and subtract n $\bar{x}^2$:

$$\sum_{i=1}^{n} x_i^2 - n\overline{x^2} + n\overline{x^2}$$

Rearranging this, we get:

$$\sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\overline{x^2}$$

Now, simplify this expression. We know:

$$\sum_{i=1}^{n} x_i^2 - n\overline{x^2} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Therefore, the compact form of the denominator is:

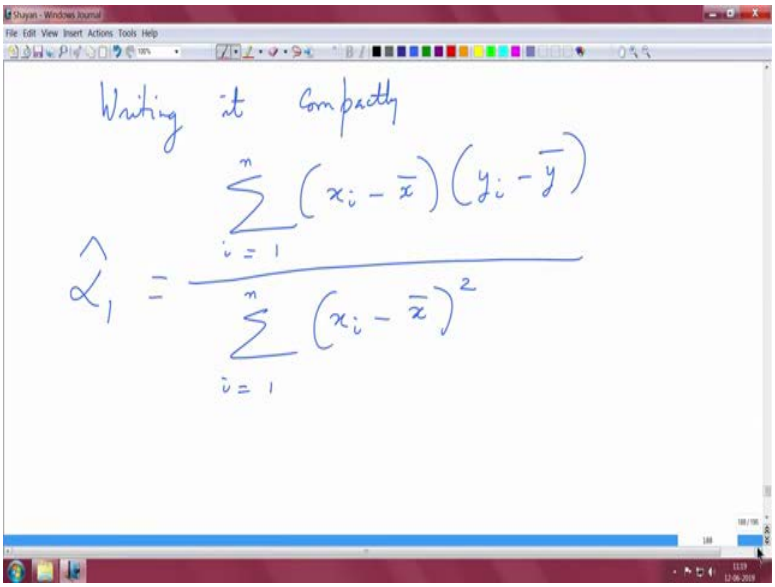
$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

This is a neat and compact expression for the denominator, showing the variance of x_i around the mean $\bar{x}$.

Now, let's simplify the expression for $\hat{\alpha}_1$. The equation for $\hat{\alpha}_1$ is:

$$\widehat{\alpha_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(Refer Slide Time: 20:18)



This simplified form is easier to remember. Essentially, you subtract the sample mean from each x_i and y_i, multiply the results, sum them up, and then divide by the sum of the squared differences of x_i from its mean. This compact form makes the calculations straightforward.

However, generalizing this concept to the multivariable case is more complex. It involves solving a system of simultaneous equations with all the sums, which can become quite intricate. We will explore this challenge in more detail as we continue our investigation.

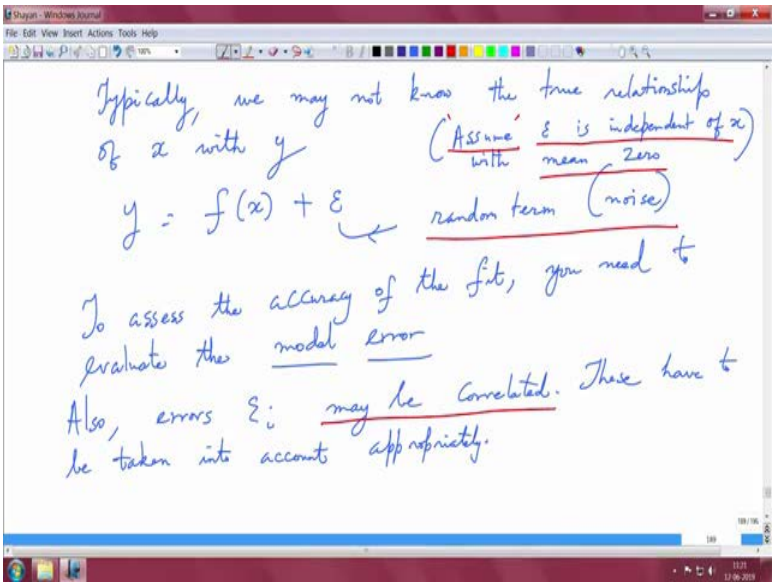
Before diving into that, let's first address the regression problem in the presence of observation noise. Here's the setup for that problem.

Typically, we may not know the true relationship between x and y. Let's assume that y can be expressed as a function f(x) plus a random term ϵ, which represents noise. One of the key assumptions we make is that the noise ϵ is independent of x and has a mean of zero. In practice, the noise may not have a mean of zero, but you can always adjust the data to make the mean zero, effectively removing any bias.

To evaluate the accuracy of the fit, we might need to make some simplifications. In many practical situations, the errors ϵ_i might not be correlated, so it's essential to account for this

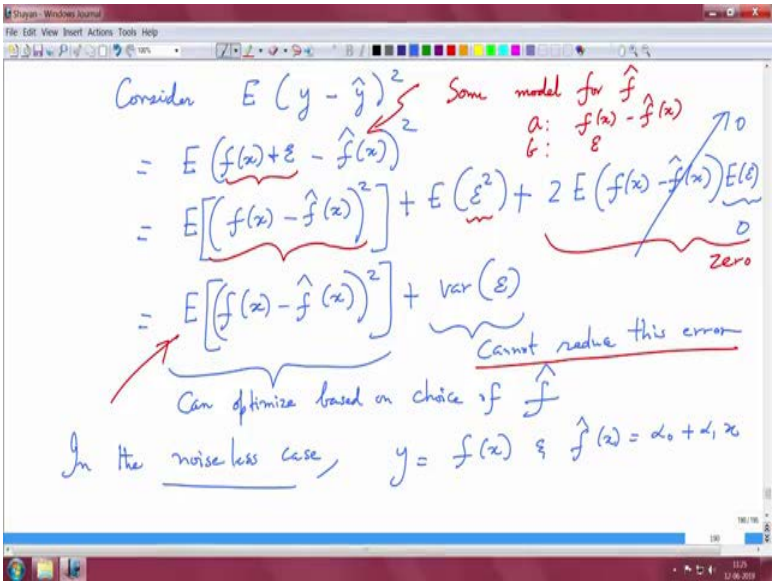
appropriately.

(Refer Slide Time: 21:34)



However, for simplicity, we assume that ϵ is independent of x and has a mean of zero. If the errors are correlated, we can decorrelate them, and if the noise has a non-zero mean, we can adjust it to zero. By making these assumptions, we can manage these issues more effectively.

(Refer Slide Time: 23:06)



Let’s simplify this problem by focusing on the expected value of $(y - \hat{y})^2$. Here, y is

expressed as f(x) + ϵ, where f(x) represents the true model and $\hat{f}$(x) is our approximation model. The goal is to evaluate the fit of $\hat{f}$(x) compared to the true function f(x).

To proceed, we express the squared difference as:

$$(y - \hat{y})^2 = \left(f(x) + \epsilon - \hat{f}(x)\right)^2$$

Expanding this, we have:

$$\left(f(x) - \hat{f}(x) + \epsilon\right)^2$$

Here, let $a = f(x) - \hat{f}(x)$ and b = ϵ. The expression becomes:

$$(a + b)^2 = a^2 + 2ab + b^2$$

Taking the expectation, and using the linearity of expectation, we get:

$$E\left[\left(f(x) - \hat{f}(x) + \epsilon\right)^2\right] = E[a^2] + E[2ab] + E[b^2]$$

Given that ϵ is independent of x, the cross-term E[2ab] simplifies to zero, due to the zero mean assumption of ϵ. Hence:

$$E\left[\left(f(x) - \hat{f}(x) + \epsilon\right)^2\right] = E\left[\left(f(x) - \hat{f}(x)\right)^2\right] + E[\epsilon^2]$$

The term $E[\epsilon^2]$ represents the variance of the noise ϵ. Since the variance of ϵ is a property of the observation noise and is not reducible through model adjustments, the only part we can influence is $E\left[\left(f(x) - \hat{f}(x)\right)^2\right]$.

In a noiseless case, where y = f(x), we assume the model $\hat{f}$(x) to be of the form $\alpha_0 + \alpha_1$ x, and we aim to minimize the squared difference between f(x) and $\hat{f}$(x). By choosing a more complex model, such as $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots$, we can improve the fit. However, the variance of ϵ remains constant and cannot be further reduced. This highlights an important practical consideration: while we can refine our model to reduce the approximation error, the inherent noise variance is an unavoidable part of the data.