10. Empirical Linear Regression via The Statistical Hammer

Least Squares Estimator (LSE)

6:47 / 6:47 1.0x

it's just average of the Y's minus b hat average of X.

So this is something that's just exactly written here,

I just have to take my expectations

and replace them by a bar.

OK, so just do it-- maybe a practice

so you don't have to actually just follow

exactly the same thing.

Is that clear for everyone?

End of transcript. Skip to the start.

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

In **empirical linear regression** , we are given a collection of points $\{(x_i, y_i)\}_{i=1}^n$. The goal is to fit a linear model $Y = a + bX + \varepsilon$ by computing the **Least Squares Estimator** , which minimizes the **loss function**

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (a + bx_i))^2.$$

Using the same technique as in the problems on theoretical linear regression, one obtains the solution

$$\hat{a} = \overline{y} - \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2}\overline{x} \qquad \hat{b} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2}.$$

In this particular case, this is precisely what one obtains by taking the least squares solution for the theoretical linear regression problem

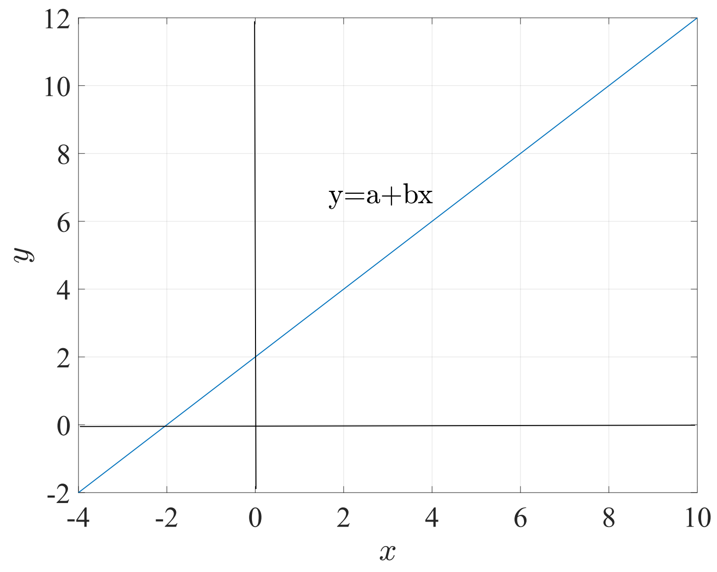
$$a = \mathbb{E}[Y] - \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(X)}\mathbb{E}[X] \qquad b = \frac{\mathsf{Cov}(X,Y)}{\mathsf{Var}(X)}$$

and replacing each term with their empirical counterparts according to the *plug-in principle*, i.e. $\mathbb{E}[X]$ with $\overline{x}$, $\mathsf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ with $\overline{x^2} - \overline{x}^2$, etc. It happens to work nicely in this setting, but this trick does not always work out in general! (See if you can reproduce the proof that this formula is the correct one.)

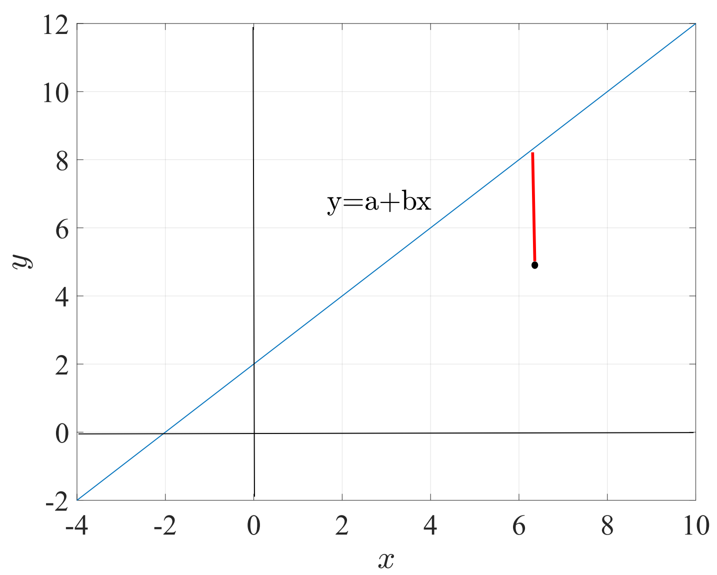
Which Squared Loss?

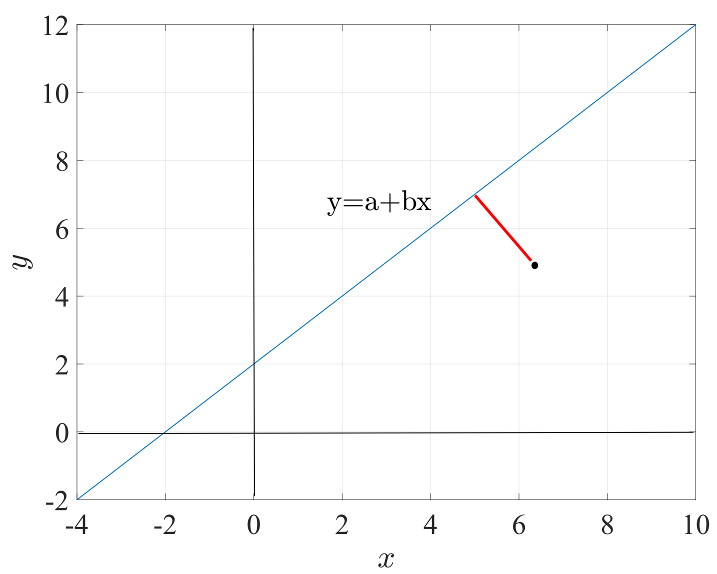
1/1 point (graded)

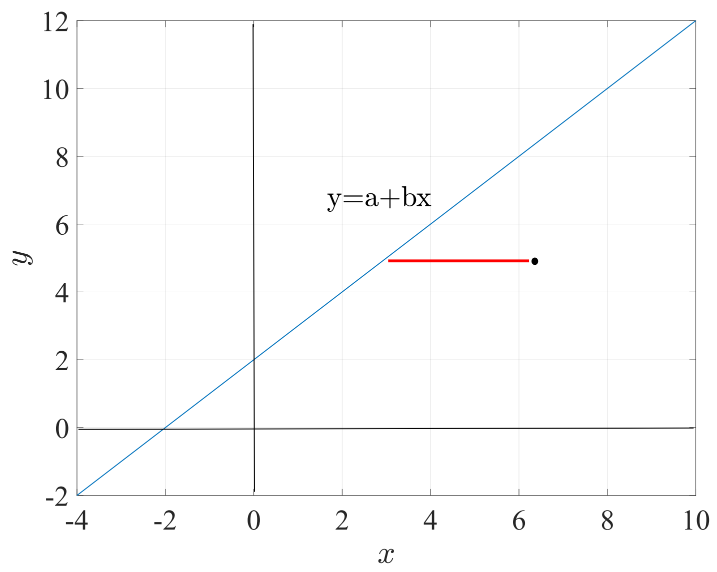
Consider the line $y = a + bx$, illustrated below.



In each of the following choices, (X, Y) is an arbitrary point, not necessarily on the line. Which choice best illustrates, via a line segment highlighted in red, the distance that is squared by the expression $(Y - (a + bX))^2$?







Solution:

The specified quantity $(Y-(a+bX))^2$ is the squared difference between the Y coordinate of the point (X,Y) and that of the point $(X,a+bX)$ predicted by the line. This represents the **vertical** squared distance between these two points.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Assumptions of Linear Regression

0/1 point (graded)

Assume that we are given a collection of n points $\{(x_i,y_i)\}_{i=1}^n$. Which one of the following choices, on its own, provides a **sufficient** condition under which a unique least-squares estimator $(\hat{a},\hat{b})$ exists?

- There are at least two total observations; i.e. $(n \geq 2)$.
- There are at least two distinct x's in the collection; i.e. $x_1 \neq x_2$. ✔
- There are at least two distinct y's in the collection; i.e. $y_1 \neq y_2$. ✖

Solution:

The correct choice is "**There are at least two distinct x's in the collection; i.e. $x_1 \neq x_2$**". The other two choices allow collections of points where all of the x's are equal: $x_1 = \cdots = x_n$. This makes it so that the empirical variance is equal to zero. Just as in the case of theoretical linear regression, this means that there is an infinite family of lines that minimize the empirical least-squared error. More specifically, any line that crosses $(\bar{x},\bar{y})$, regardless of the slope, is a minimizer.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Linear Least Squares: A Numerical Example

2/2 points (graded)

Consider the four points $(x_1,y_1)=(-5,-10)$, $(x_2,y_2)=(0,3)$, $(x_3,y_3)=(2,11)$ and $(x_4,y_4)=(3,14)$. The line that minimizes the empirical squared error can be expressed as $y=\hat{a}+\hat{b}x$, where

$\hat{a} =$ 4.5 ✔ **Answer:** 4.5

$\hat{b} =$ 3 ✔ Answer: 3

Solution:

We may directly apply the formula

$$\hat{b} = \frac{\frac{1}{4}\sum_{i=1}^{4}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\frac{1}{4}\sum_{i=1}^{4}\left(x_i - \overline{x}\right)^2}$$

$$\hat{a} = \overline{y} - \hat{a}\overline{x}.$$

to obtain the pair $(\hat{a}, \hat{b}) = (4.5, 3)$.
Alternatively, the calculation can be simplified using the observation that $\overline{x} = \frac{1}{4}(-5+0+2+3) = 0$. In fact, recall that Cov $(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ and the numerator of b is the empirical estimate of Cov (X, Y). Since $\overline{x}$ is zero, the numerator of b reduces to the empirical estimate of $\mathbb{E}[XY]$, or $\frac{1}{4}\sum_{i=1}^{4} x_i y_i$. Therefore, the above calculation can be computed as

$$\hat{b} = \frac{\sum_{i=1}^{4} x_i y_i}{\sum_{i=1}^{4} x_i^2} = \frac{(-5\cdot -10) + (0\cdot 3) + (2\cdot 11) + (3\cdot 14)}{(-5)^2 + 0^2 + 2^2 + 3^2} = \frac{114}{38} = 3$$

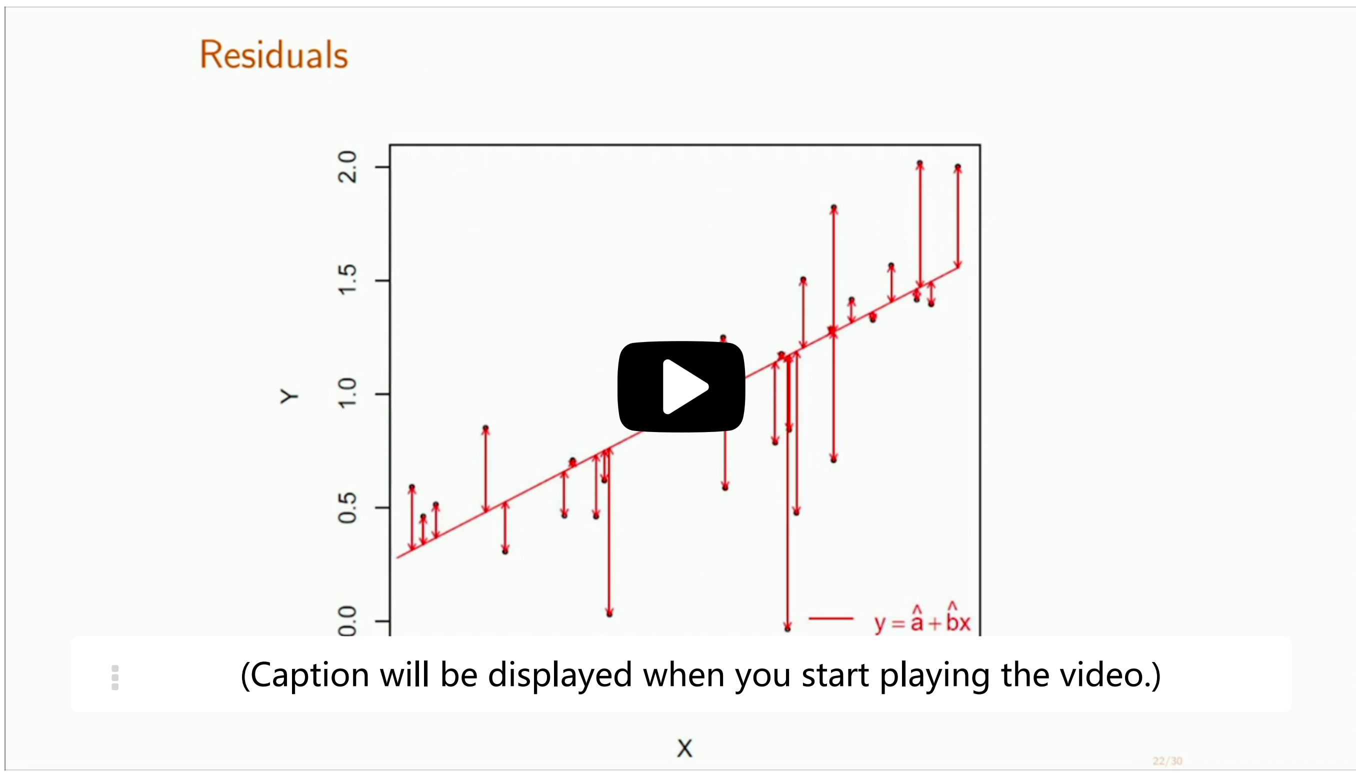
$$\hat{a} = \overline{y} = \frac{-10+3+11+14}{4} = \frac{18}{4} = 4.5.$$

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Residuals

Start of transcript. Skip to the end.



(Caption will be displayed when you start playing the video.)

0:00 / 0:00 1.0x

So these are what my residuals look like.

So here-- so why do a plot those residuals?

This is the red line.

The red line is the line that has the estimators there.

If I really wanted to know what my epsilon Is were,

I would have to draw the blue line with the A star and B star

and then see how far my points were from

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Discussion

Topic: Unit 6 Linear Regression:Lectures 19: Linear Regression 1 / 10. Empirical Linear Regression via The Statistical Hammer

Hide Discussion

Add a Post

All Posts

noise and residuals

question posted 5 days ago by michael_x

What's the difference ?

Noise is the distance from the theoretical line to each data point, while residual is the distance from the empirical fitting line to each data point ?

This post is visible to everyone.

sudarsanvsr_mit (Staff)
about 7 hours ago - marked as answer about 4 hours ago by **michael_x**

@michael_X: Your understanding is correct. Noise (in magnitude) is the distance between a given point and the linear regression line given by the true a^* and b^* while a residual (in magnitude) is the distance between the point and the estimated regression line given by $\hat{a}$ and $\hat{b}$.

Show Comment (1)

Add a Response

1 other response

kunapalli
5 days ago

https://en.wikipedia.org/wiki/Errors_and_residuals

The "disturbance" or "error" is the difference between the population mean and the observed value.

The "residual" is the difference between the sample mean and the observed value.

The sum of the residuals is necessarily zero. The sum of the disturbances is, with probability 1, not zero.

The disturbances are independent. The residuals cannot be independent since they must sum to zero. (So, for example, if you add up all but one of the residuals and the sum is +8, then the remaining residual must be −8. It can be predicted based on the other residuals, and so cannot be independent of them.)

Are ("disturbance" or "error") and noise the same thing here? As far as I understand, noise is a random variable, which should have nothing to do with observed value.

posted about 14 hours ago by bunnybunny

Add a comment

Showing all responses

Add a response:

Preview