Cross Validated

Derive Variance of regression coefficient in simple linear regression

Asked 5 years, 9 months ago Active 2 years, 7 months ago Viewed 67k times

38

37

In simple linear regression, we have $y = \beta_0 + \beta_1 x + u$, where $u \sim iid\, \mathcal{N}(0, \sigma^2)$. I derived the estimator:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}\ ,$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y.

Now I want to find the variance of $\hat{\beta}_1$. I derived something like the following:

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2 (1 - \frac{1}{n})}{\sum_i (x_i - \bar{x})^2}\ .$$

The derivation is as follow:

$$
\begin{aligned}
&\operatorname{Var}(\hat{\beta}_1) \\
&= \operatorname{Var}\left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}\right) \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \operatorname{Var}\left(\sum_i (x_i - \bar{x})\left(\beta_0 + \beta_1 x_i + u_i - \frac{1}{n}\sum_j (\beta_0 + \beta_1 x_j + u_j)\right)\right) \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \operatorname{Var}\left(\beta_1 \sum_i (x_i - \bar{x})^2 + \sum_i (x_i - \bar{x})\left(u_i - \sum_j \frac{u_j}{n}\right)\right) \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \operatorname{Var}\left(\sum_i (x_i - \bar{x})\left(u_i - \sum_j \frac{u_j}{n}\right)\right) \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \times \\
&\quad E\left[\left(\sum_i (x_i - \bar{x})(u_i - \sum_j \frac{u_j}{n}) - \underbrace{E\left[\sum_i (x_i - \bar{x})(u_i - \sum_j \frac{u_j}{n})\right]}_{=0}\right)^2\right] \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} E\left[\left(\sum_i (x_i - \bar{x})(u_i - \sum_j \frac{u_j}{n})\right)^2\right] \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} E\left[\sum_i (x_i - \bar{x})^2 (u_i - \sum_j \frac{u_j}{n})^2\right] \qquad \text{, since } u_i \text{ 's are iid} \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \sum_i (x_i - \bar{x})^2 E\left(u_i - \sum_j \frac{u_j}{n}\right)^2 \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \sum_i (x_i - \bar{x})^2 \left(E(u_i^2) - 2 \times E\left(u_i \times (\sum_j \frac{u_j}{n})\right) + E\left(\sum_j \frac{u_j}{n}\right)^2\right) \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \sum_i (x_i - \bar{x})^2 \left(\sigma^2 - \frac{2}{n}\sigma^2 + \frac{\sigma^2}{n}\right) \\
&= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\left(1 - \frac{1}{n}\right)
\end{aligned}
$$

Did I do something wrong here?

I know if I do everything in matrix notation, I would get $\operatorname{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$. But I am trying to derive the answer without using the matrix notation just to make sure I understand the concepts.

regression mathematical-statistics variance linear-model regression-coefficients

edited Mar 7 '14 at 2:00
TooTone
3,309 18 31

asked Mar 2 '14 at 15:56
mynameisJEFF
1,231 3 19 26

2 Yes, your formula from matrix notation is correct. Looking at the formula in question, $1-\frac{1}{n}=\frac{n-1}{n}$ so it rather looks as if you might used a sample standard deviation somewhere instead of a population standard deviation? Without seeing the derivation it's hard to say any more. – TooTone Mar 3 '14 at 0:51

General answers have also been posted in the duplicate thread at stats.stackexchange.com/questions/91750. – whuber ♦ Mar 29 '14 at 20:50

3 Answers

35

At the start of your derivation you multiply out the brackets $\sum_i (x_i-\bar{x})(y_i-\bar{y})$, in the process expanding both y_i and $\bar{y}$. The former depends on the sum variable i, whereas the latter doesn't. If you leave $\bar{y}$ as is, the derivation is a lot simpler, because

$$\begin{aligned}\sum_i (x_i-\bar{x})\bar{y} &= \bar{y}\sum_i (x_i-\bar{x})\\ &= \bar{y}\left(\left(\sum_i x_i\right)-n\bar{x}\right)\\ &= \bar{y}\left(n\bar{x}-n\bar{x}\right)\\ &= 0\end{aligned}$$

Hence

$$\begin{aligned}\sum_i (x_i-\bar{x})(y_i-\bar{y}) &= \sum_i (x_i-\bar{x})y_i-\sum_i (x_i-\bar{x})\bar{y}\\ &= \sum_i (x_i-\bar{x})y_i\\ &= \sum_i (x_i-\bar{x})(\beta_0+\beta_1 x_i+u_i)\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2}\right) \\
&= \mathrm{Var}\left(\frac{\sum_i (x_i-\bar{x})(\beta_0+\beta_1 x_i+u_i)}{\sum_i (x_i-\bar{x})^2}\right), \quad \text{substituting in the above} \\
&= \mathrm{Var}\left(\frac{\sum_i (x_i-\bar{x})u_i}{\sum_i (x_i-\bar{x})^2}\right), \quad \text{noting only } u_i \text{ is a random variable} \\
&= \frac{\sum_i (x_i-\bar{x})^2 \mathrm{Var}(u_i)}{\left(\sum_i (x_i-\bar{x})^2\right)^2}, \quad \text{independence of } u_i \text{ and, } \mathrm{Var}(kX)=k^2\mathrm{Var}(X) \\
&= \frac{\sigma^2}{\sum_i (x_i-\bar{x})^2}
\end{aligned}$$

which is the result you want.

As a side note, I spent a long time trying to find an error in your derivation. In the end I decided that discretion was the better part of valour and it was best to try the simpler approach. However for the record I wasn't sure that this step was justified

$$\begin{aligned}
&=. \frac{1}{(\sum_i (x_i-\bar{x})^2)^2} E\left[\left(\sum_i (x_i-\bar{x})(u_i-\sum_j \frac{u_j}{n})\right)^2\right] \\
&= \frac{1}{(\sum_i (x_i-\bar{x})^2)^2} E\left[\sum_i (x_i-\bar{x})^2 (u_i-\sum_j \frac{u_j}{n})^2\right] \quad \text{, since } u_i \text{ 's are iid}
\end{aligned}$$

because it misses out the cross terms due to $\sum_j \frac{u_j}{n}$.

answered Mar 7 '14 at 13:35
TooTone
3,309 18 31

I noticed that I could use the simpler approach long ago, but I was determined to dig deep and come up with the same answer using different approaches, in order to ensure that I understand the concepts. I realise that first $\sum_j \hat{u}_j = 0$ from normal equations (FOC from least square method), so $\bar{\hat{u}} = \frac{\sum_i u_i}{n} = 0$, plus $\bar{\hat{u}} = \bar{y} - \bar{\hat{y}} = 0$, so $\bar{y} = \bar{\hat{y}}$. So there won't be the term $\sum_j \frac{u_j}{n}$ in the first place. – mynameisJEFF Mar 7 '14 at 13:56

ok, in your question the emphasis was on avoiding matrix notation. – TooTone Mar 7 '14 at 14:01

Yes, because I was able to solve it using matrix notation. And notice from my last comment, I did not use any linear algebra. Thanks for your great answer anyway^.^ – mynameisJEFF Mar 7 '14 at 14:04

sorry are we talking at cross-purposes here? I didn't use any matrix notation in my answer either, and I thought that was what you were asking in your question. – TooTone Mar 7 '14 at 14:06

sorry for misunderstanding haha... – mynameisJEFF Mar 7 '14 at 14:21

|

2

Begin from "The derivation is as follow:" The 7th "=" is wrong.

Because

$\sum_i(x_i-\bar{x})(u_i-\bar{u})$

$=\sum_i(x_i-\bar{x})u_i-\sum_i(x_i-\bar{x})\bar{u}$

$=\sum_i(x_i-\bar{x})u_i-\bar{u}\sum_i(x_i-\bar{x})$

$=\sum_i(x_i-\bar{x})u_i-\bar{u}(\sum_i x_i-n\bar{x})$

$=\sum_i(x_i-\bar{x})u_i-\bar{u}(\sum_i x_i-\sum_i x_i)$

$=\sum_i(x_i-\bar{x})u_i-\bar{u}0$

$=\sum_i(x_i-\bar{x})u_i$

So after 7th "=" it should be:

$\frac{1}{(\sum_i(x_i-\bar{x})^2)^2}E\left[(\sum_i(x_i-\bar{x})u_i)^2\right]$

$=\frac{1}{(\sum_i(x_i-\bar{x})^2)^2}E\left(\sum_i(x_i-\bar{x})^2u_i^2+2\sum_{i\neq j}(x_i-\bar{x})(x_j-\bar{x})u_iu_j\right)$

=$\frac{1}{(\sum_i(x_i-\bar{x})^2)^2}E\left(\sum_i(x_i-\bar{x})^2u_i^2\right)+2E\left(\sum_{i\neq j}(x_i-\bar{x})(x_j-\bar{x})u_iu_j\right)$

=$\frac{1}{(\sum_i(x_i-\bar{x})^2)^2}E\left(\sum_i(x_i-\bar{x})^2u_i^2\right)$, because u_i and u_j are independent and mean 0, so $E(u_iu_j)=0$

=$\frac{1}{(\sum_i(x_i-\bar{x})^2)^2}\left(\sum_i(x_i-\bar{x})^2E(u_i^2)\right)$

$\frac{\sigma^2}{(\sum_i(x_i-\bar{x})^2)^2}$

edited Apr 27 '17 at 8:17

answered Apr 24 '17 at 6:10

user158565

6,502 2 6 18

1 It might be helpful if you edited your answer to include the correct line. – mdewey Apr 24 '17 at 7:48

Your answer is being automatically flagged as low quality because it's very short. Please consider expanding on your answer – Glen_b -Reinstate Monica Apr 24 '17 at 12:04

2

I believe the problem in your proof is the step where you take the expected value of the square of $\sum_i (x_i - \bar{x}) \left(u_i - \sum_j \frac{u_j}{n}\right)$. This is of the form $E\left[\left(\sum_i a_i b_i\right)^2\right]$, where $a_i = x_i - \bar{x}; b_i = u_i - \sum_j \frac{u_j}{n}$. So, upon squaring, we get $E\left[\sum_{i,j} a_i a_j b_i b_j\right] = \sum_{i,j} a_i a_j E\left[b_i b_j\right]$. Now, from explicit computation, $E\left[b_i b_j\right] = \sigma^2 \left(\delta_{ij} - \frac{1}{n}\right)$, so $E\left[\sum_{i,j} a_i a_j b_i b_j\right] = \sum_{i,j} a_i a_j \sigma^2 \left(\delta_{ij} - \frac{1}{n}\right) = \sum_i a_i^2 \sigma^2$ as $\sum_i a_i = 0$.

answered May 3 '16 at 13:30

Stefano
21 1