Identify the outliers with the highest squared residuals under the Linear regression model in R

Asked 7 years, 1 month ago Modified 7 years, 1 month ago Viewed 5k times

Part of R Language Collective

1

I have a data set [1000 x 80] of 1000 data points each with 80 variable values. I have to linearly regress two variables: price and area, and identify the 5 data points that have highest squared residuals. For these identified data points, I have to display 4 of the 80 variable values.

I do not know how to use the residuals to identify the original data points. All I have at the moment is:

```
model_lm <- lm(log(price) ~ log(area), data = ames)
```

Can I please get some guidance on how I can approach the above problem

R r regression outliers

Share
Improve this question
Follow

edited Oct 27, 2017 at 21:06
MrFlick
206k ● 19 ● 291 ● 316

asked Oct 27, 2017 at 20:26
Paras Joshi
11 ● 1 ● 2

1 Answer

Sorted by: Highest score (default)

1

The model_lm object will contain a variable called 'residuals' that will have the residuals in the same order as the original observations. If I'm understanding the question correctly, then an easy way to do this is base R is:

```
ames$residuals <- model_lm$residuals  ## Add the residuals to the data.frame

o <- order(ames$residuals^2, decreaseing=T)   ## Reorder to put largest first

ames[o[1:5],]   ## Return results
```

Share
Improve this answer
Follow

edited Oct 27, 2017 at 21:07
MrFlick
206k ● 19 ● 291 ● 316

answered Oct 27, 2017 at 20:53
KMcC

typo: `o <- order(ames$residuals^2, decreaseing=T)` should be: `o <- order(ames$residuals^2, decreasing=T)` – Ketil B T Oct 27, 2017 at 21:10

Also, just to address the second part of your problem: If `varNames` is a character vector with your 4 variable names in it, e.g., `varNames = c("var1", "var2", "var3", "var4")`, you can display them with `ames[, varNames]`. Adding this to @KMcC's answer, you get `ames[o[1:5], varNames]`. – Ketil B T Oct 27, 2017 at 21:14