Multicolinearity

Bookmark this page

Multicollinearity

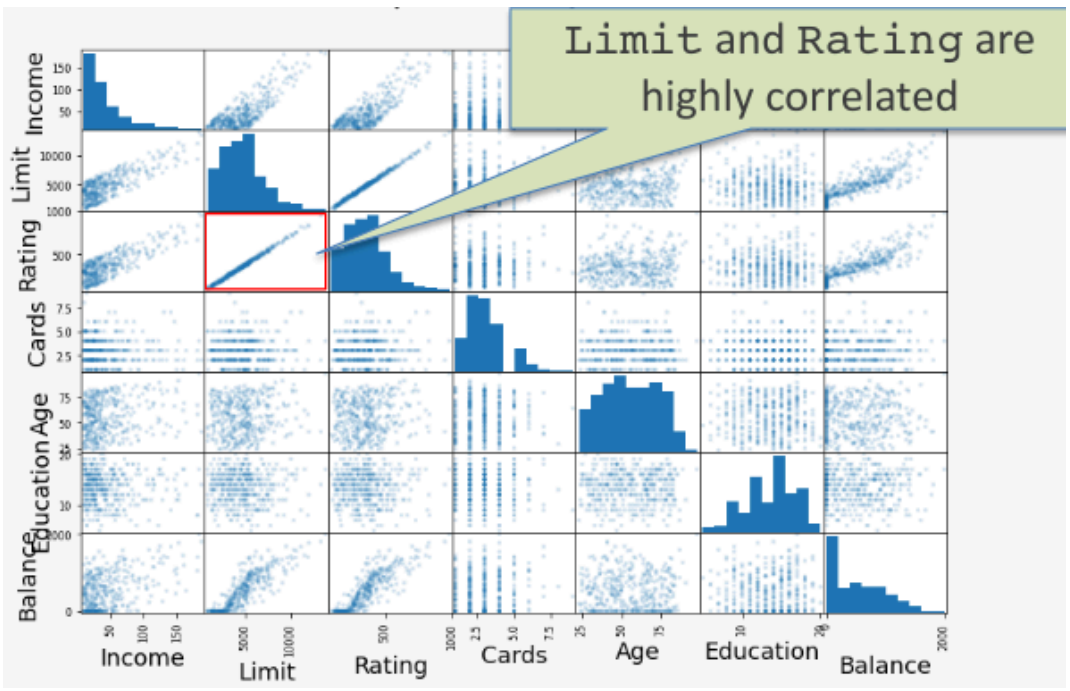
> **Definition**
>
> **Collinearity** and **multicollinearity** refer to the case in which two or more predictors are correlated. As one predictor increases or decreases, the other one always follows it.

One question we may have is, "how are these predictors related to one another?" This is an important to know as high multicolinearity can complicate the interpretation of our model's parameters as we'll see

In this example we're looking at the credit data set. This data set contains many features which may or may not be useful in predicting the risk of extending credit to an individual, though any of the predictors could serve as our response variable. Here we are only looking at the quantitative predictors.

Below we have an example of a **pairplot**, a common visualization method for exploring possible correlations between our predictors.

The off-diagonal subplots contain scatter plots showing how pairs of predictors are related. On the diagonal we have histograms displaying the distribution of values in our data for each predictor.



We can see that Limit and Rating are highly correlated. As Limit increases Rating also increases.

If we fit a regression model to predict Balance from the other predictors we will see why this is a problem. The regression coefficients we get from the fit are **not uniquely determined**! As a result it hurts the **interpretability** of the model.

Consider the coefficient values from two different regression fits below. The model on the left included both Limit and Rating. In the model on the right, Limit was removed.

	Columns	Coefficients
0	Income	-7.802
1	Limit	0.193
2	Rating	1.102
3	Cards	17.923
4	Age	-0.635
5	Education	-1.115
6	Gender	10.407
7	Student	426.469
8	Married	-7.019

	Columns	Coefficients
0	Income	-7.771
1	Rating	3.976
2	Cards	4.031
3	Age	-0.669
4	Education	-0.376
5	Gender	10.369
6	Student	417.417
7	Married	-13.265

Calculator

On the left we see that both `Limit` and `Rating` have positive coefficients. But if we remove `Limit` and refit, then our coefficients change even though both models have roughly the same performance as judged by MSE. As a result it's hard to know if a balance is higher because of an individual's `Rating` or `Limit`.

One way to think about this problem is that several coefficients contain the same information. Diagnostic methods like the pairplot can help us identity multicolinearity and remove redundant predictors, giving us a more interpretable model.

Discussion Board (External resource)

Haga clic en Aceptar para que su nombre de usuario y dirección de correo electrónico se envíen a una aplicación de terceros.

‹ Previous Next ›