Dummy Variables: Reasoning

The next thing to do in our analysis from a statistical perspective will be to drop the first column of our table.

There is a solid mathematical explanation for this step, however dealing with its complexity is beyond the scope of this course.

Nevertheless, if you are interested in proving econometric logic and intuition with math, you can visit this Quora link:

http://qr.ae/TUTz9q

Otherwise, you could simply advance to the next video and leave the raw mathematical proofs aside.

In the Quora article, an explanation is provided for 3 dummy variables that are MECE. In our exercise, the dummy variables will be of the same type, however there will be 28 (and hence the reasoning will be applied not for 3, but n = 28 dummies).

In a nutshell, the motivation for us to drop the first column, reason 0, goes like this.

If a person has been absent due to reason 0, this means they have been away from work for an unknown reason. Hence, this column acts like the baseline, and all the rest are represented in comparison to this.

As a consequence, dropping this column would allow us to only conduct the analysis for the reasons we are aware of. And that's exactly what we want to do - explore whether or not a specific *known* reason for absence induces an individual to be excessively absent from work. That's why we don't really need to keep in our data set information about someone who has been away due to an *unknown* reason.

Otherwise, regarding the values stored in the remaining columns, representing reasons 1 and above, we can solidify our rationale by saying the following: Imagine that by default there's no particular reason for a given individual to be absent from work. If there is any, though, it will be marked with the value of 1 under the corresponding reason number.

Therefore, as explained in the article on Quora, to avoid issues with multicollinearity, we must remove one column with dummy variables. To preserve the logic of our analysis, this will be the column that stands for reason 0.

In our next video, we will go back to Python and proceed from this point on - dropping the unnecessary column.

See you there!