Kolmogorov-Smirnov Test in R

Having imputed our missing values, we might want to make sure that these missing values are likely to be coming from the same distribution as the observed data or whether our two imputation methods are producing data that's from a different distribution that might help us pick which one we might prefer to use. As we've seen, there might be some differences in the EDFs or maybe the density functions. And one imputation method might look like it's following the observed data more closely.

And so if those two distributions of the imputations are different, then that might mean we should probably pick the one that's more closely following to it and discard our other one because it is from a different distribution. So we'll do this with the Kolmogorov-Smirnov test.

And the function for this is ks.test, into which we simply specify the two variables that we want to compare. I'm going to be comparing the five-year haemoglobin variable, having gone through regression imputation and predictive mean matching to see if there is a noticeable difference between the two methods on this occasion. So once I run the function, it will give me my D-statistic and my p-value. We can see here that the p-value is extremely high.

It's actually recorded as 1, which means that we can be very confident that these two data sets under the different imputation methods have produced similar results from the same distribution here. And so statistically, there is not likely to be a huge difference in which one that we use.

We might want to also check that our observed values are consistent with our imputed values. And we can do that by doing a Kolmogorov-Smirnov test using the original data set and the imputed data set to see if the imputations have shifted our distribution away from what it originally was. So I'm going to use the predictive mean matching data set for demonstration here. And again, we can see our D-statistics are extremely low.

Our p-values are extremely high. So we haven't shifted our distribution of donations to somewhere else. We don't want to do this, for example, if we have missing completely at random or missing at random data because those two patterns of missingness do state that

the missingness of the observed variable does not depend on what that actual missing value of the variable was. Although, for missing at random data it might depend on other variables in your data set.

Either way, because it doesn't depend on the actual value that was missing, we should expect to see that these distributions are similar.