5.4 Significance of features

In the same vein, which features are significant? A quick observation shows that numerical values of the coordinates in $\mathbf{x}$ depend on the units used to measure the associated features. So simply the fact that the first coordinate x_1 of $\mathbf{x}$ is larger in magnitude than, say the second coordinate x_2 of $\mathbf{x}$ is no indication that feature 1 is more significant than feature 2.

Traditionally, significance was answered by $t-$ or $F-$ tests. The OLS program from the statmodel api in python will print out significance values for the $t-$test for each feature as the notebook shows.

The idea behind them is as follows—suppose we want to know if feature 1 (the first column of B) is significant. We adopt a hypothesis testing framework, where in the *null* hypothesis, we assume that the first coordinate of $\mathbf{x}$, $x_1 = 0$ (*i.e.,* the first feature should not be used at all). The alternate hypothesis is that $x_1 \neq 0$, the first feature forms part of the ground truth.

Under the null hypothesis ($x_1 = 0$), conditioned on the measurement matrix B and other coordinates of $\mathbf{x}$, the $t-$test estimates, using the probability model for Z, the error with and without using the first feature. The ratio of the errors is then used to estimate the probability that the calculated first coordinate of $\mathbf{x}_{OLS}$ could have happened purely by chance under the null model, and this is presented as the significance value. A significance value of .01 therefore means that there is only a .01 probability the first value of $\mathbf{x}_{OLS}$ could have magnitude at least as much as we have computed. So smaller the significance value, the less likely the observed value is purely by chance, the more we are inclined to disbelieve the null model, and more likely that the feature is significant. The $F-$test tests for multiple features together. See the module on $t-$ and $F-$ tests.

In the classical setting where we may suspect only a few features, the $t-$ or $F-$ is a fine strategy and has been the workhorse of significance testing. In modern problems, particularly in genetics, we have thousands or even tens of thousands of protein concentrations or gene expressions in a cell (the number of participants in these studies is at best in hundreds). In these problems, most of these genes and proteins are irrelevant to the target (perhaps the response to a drug), so we suspect *most* features. In these settings, the $t-$ or $F-$ test approaches are not very meaningful. For example, in the $t-$ test, if 100 features have significance at the .01 level. Each of the 100 events (a feature shows up purely by chance) is rare, that is each event happens at most .01 probability. But there are 100 of them, and the probability at least one of them occurs may actually be quite high. For this reason, we need more effective ways, and we briefly examined the LASSO approach.