

## Handling Influential Observations

What do you do with influential observations? Here are some suggestions. First, recheck for data entry errors and correct them if possible. If your records indicate that a house sold for \$10 and you can't locate the true price, it's reasonable to throw out that data and document the deletion.

Second, if the data appears to be valid, consider whether you have an adequate model. A different model might fit the data better. For example, stand-alone garages and storage areas were initially included in the Ames housing data. These created unusual conditions, such as home sales with no living space. A researcher might decide to change the model to accommodate these unusual observations, perhaps by including a categorical predictor that distinguished residential homes from garage only and storage area only units. Our purpose is limited to modeling residential home sales, so these were excluded from the data because they were contradictory to our goal. Including them would add noise to the estimated relationship between Above Ground Living Area and SalePrice for residential homes.

Third, determine whether the influential observation is valid, and only unusual. If you had a larger sample size, there might be more observations that are similar to the unusual one. You might need to collect more data to confirm the relationship that's suggested by the influential observation.

As a general rule, you should not exclude data. In many circumstances, some of the unusual observations contain important information. If you choose to exclude some observations, include in your report a description of the types of observations that were excluded and why. As part of your report or presentation, you should discuss the limitation of your conclusions, given the exclusions.