* Additional predictor variables will make the R-squared stay the same or increase, even if they show no relationship with the response variable.
* Adjust R-squared looks at whether additional predictor variables are contributing to the model.

Studentized residuals (or internally studentized residuals)

MSE = mean square error

= leverage

The major problem with ordinary residuals = their magnitude depends on the units of measurement -> difficult to use the residuals as a way of detecting unusual y values

* **Dividing the residuals by an estimate of their standard deviation => known as studentized residuals**

Distinction between Outliers and High Leverage Observations

* An outlier is a data point whose response y does not follow the general trend of the rest of data
* A data point has high leverage if it has “extreme” predictor x values (high or low)
* A data point can be an outlier and has high leverage

Hypothesis Testing and The Null Hypothesis

p-values: what are they and how to interpret them

* p-values are numbers, between 0 and 1, that, in [this example], quantify how confident we should be that [drugA] is different from [drugB].
* The closer a p-value is to 0, the more confidence we have that [drugA] and [drugB] are different.
* How small does a p-value have to be before we are sufficiently confident that [drugA] is different from [drugB]? (what threshold) -> commonly used threshold is 0.05 -> 5% of experiments would result in the wrong decision.
* Help us decide if we should reject the Null Hypothesis or not.
* While a small p-value helps us decide if drugA is different from drugB, it does not tell us how different they are.

Some observations about KNN:

* Non-parametric approach: no assumptions are made about the shape of the decision boundary
* Requires a lot of observations relative to the number of predictors (n >> p)
* Tends to **reduce the bias** while incurring a lot of variance
* Unlike logistic regressions, KNN does not tell us which predictors are important

t-distribution ? -> the shape is similar to normal distribution, but it has a tendency to yield more extreme points

t ~ s when df increases

the problem is most of the time, we do not know value -> use sample

s: sample standard deviation