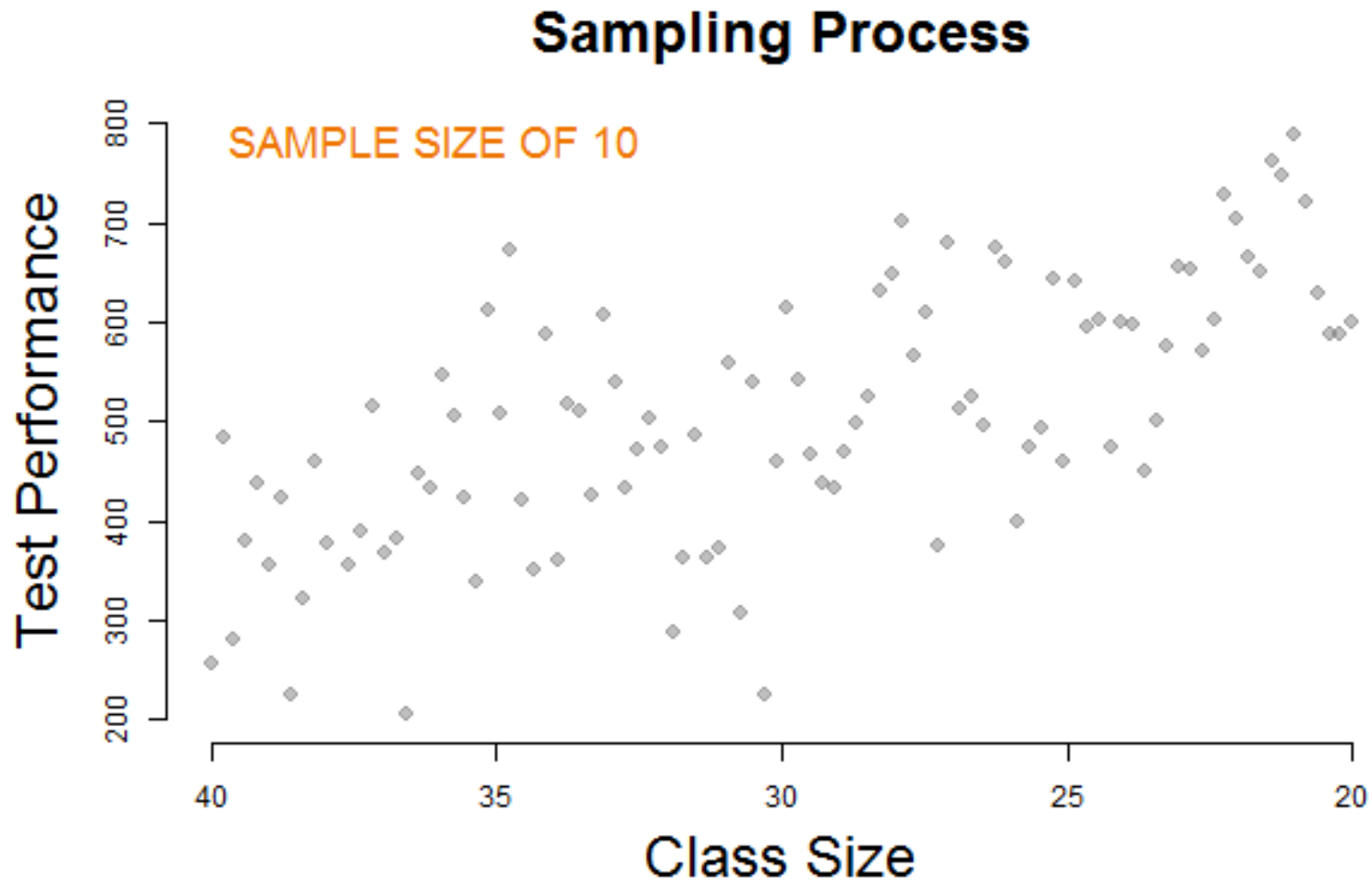
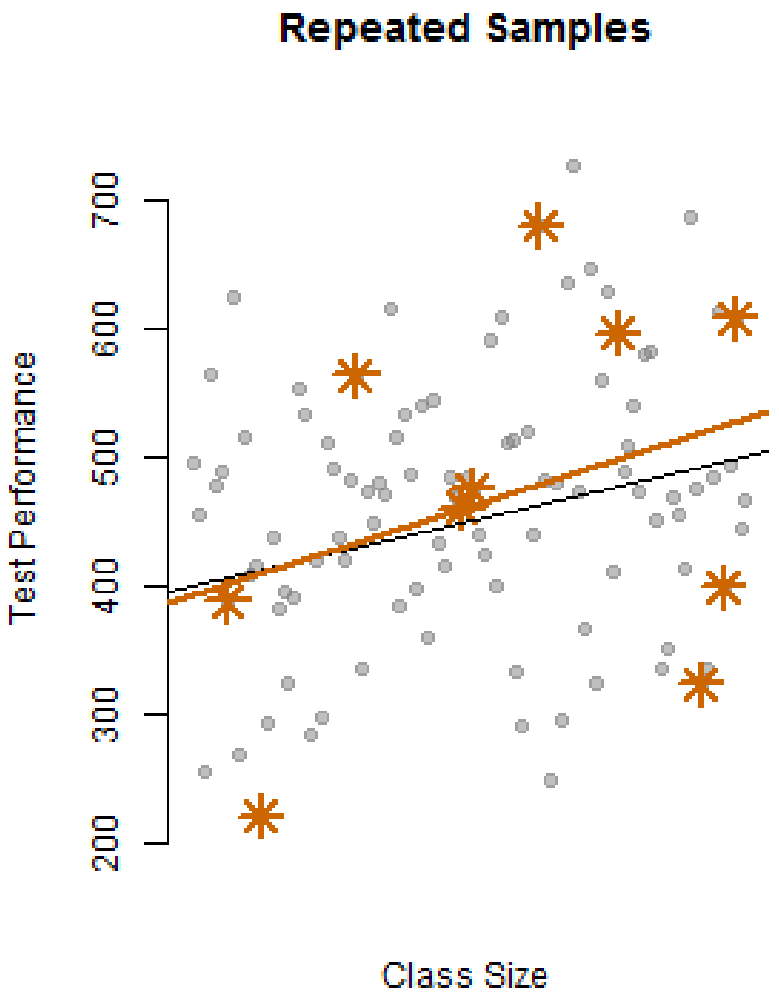
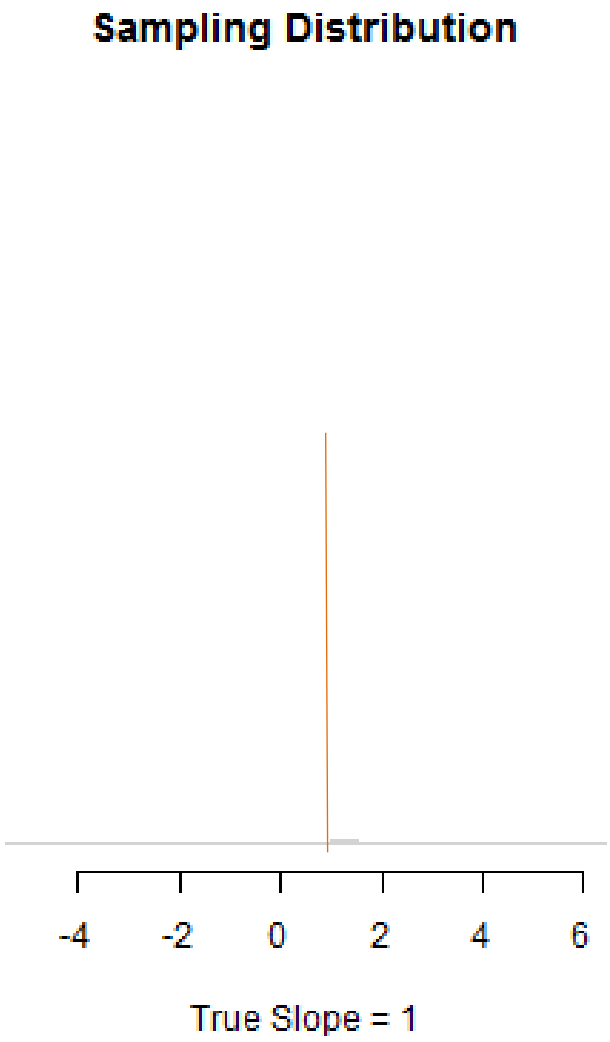
SAMPLING DISTRIBUTION


Sampling Process
SAMPLE SIZE OF 10
Test Performance
200
300
400
500
600
700
800
40
35
30
25
20
Class Size

SAMPLING DISTRIBUTION

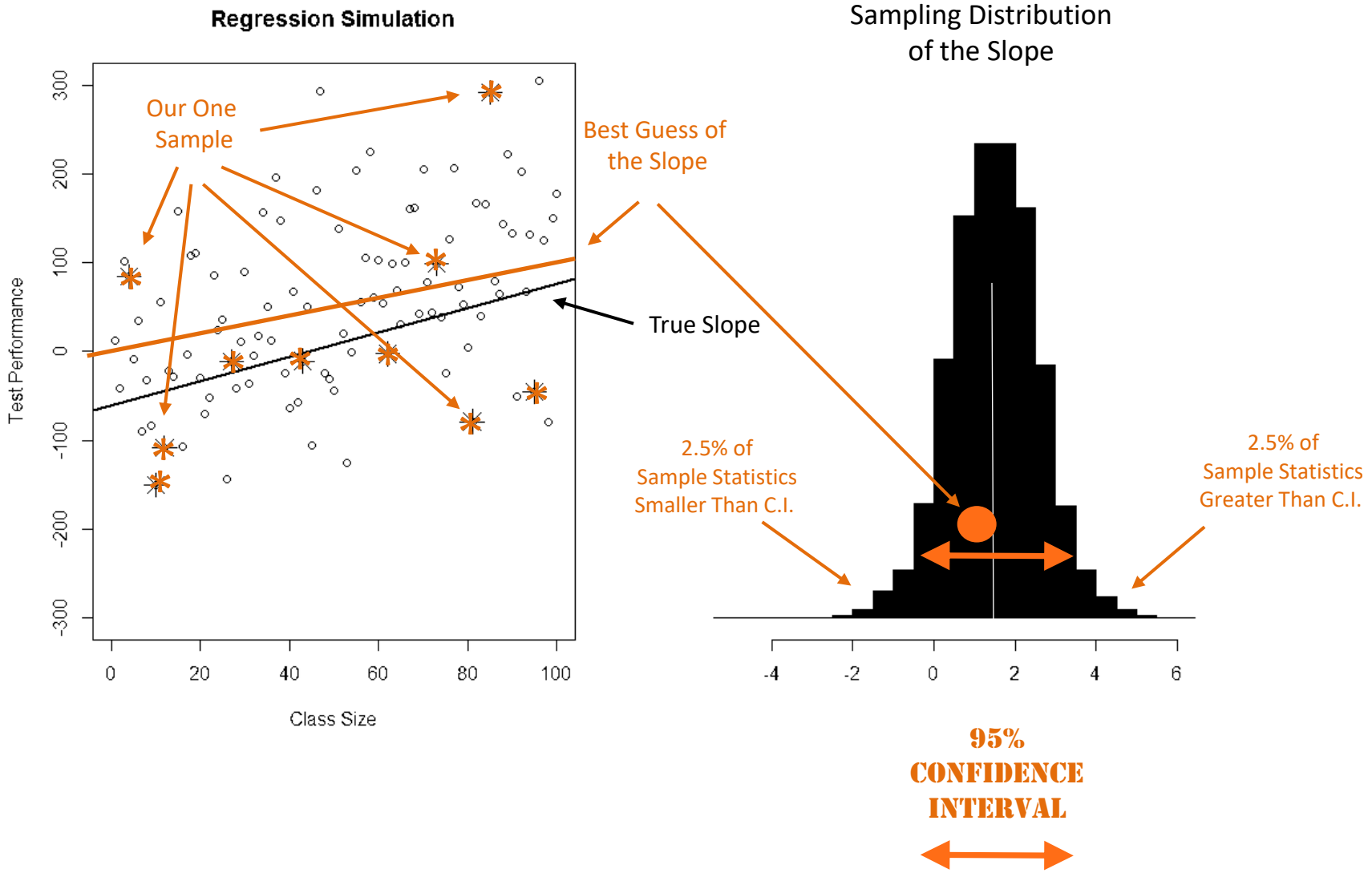


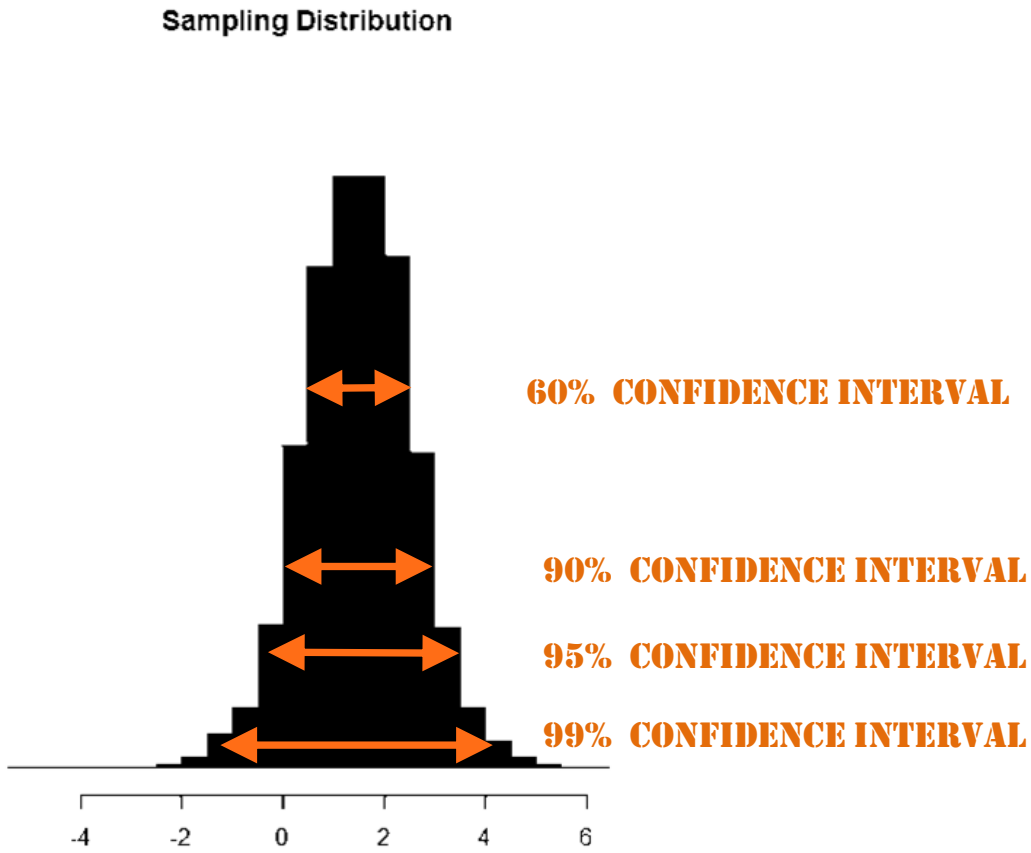


SAMPLE SIZE = 10

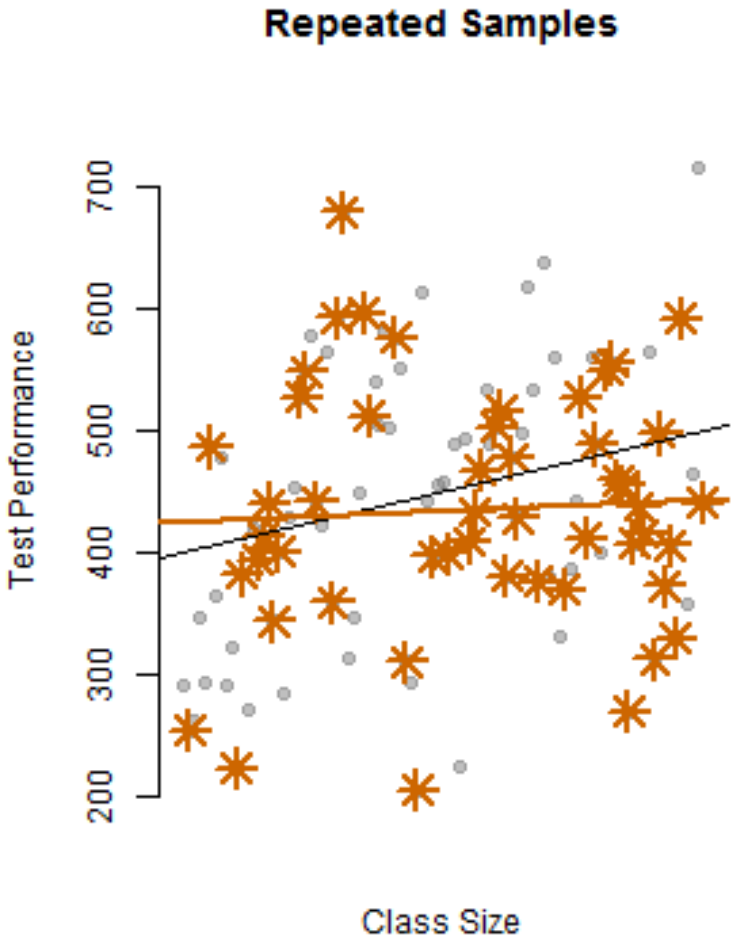
CONFIDENCE INTERVALS

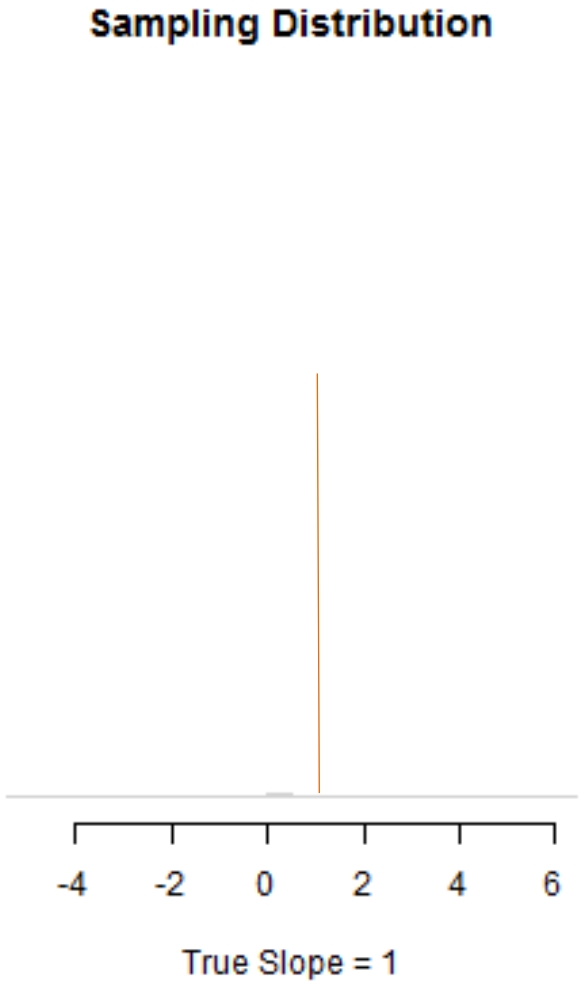
95% CONFIDENCE INTERVAL



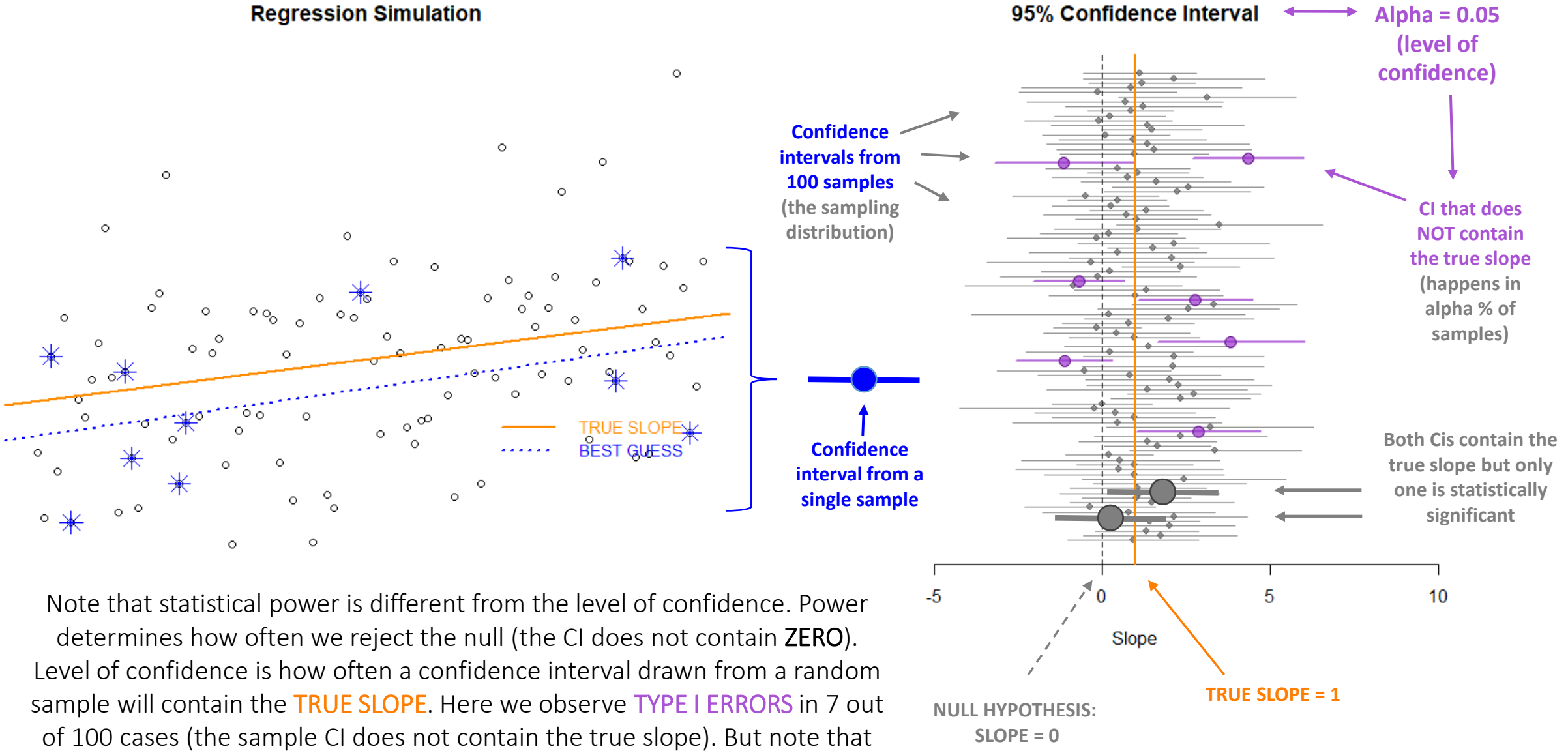

Sampling Distribution
60% CONFIDENCE INTERVAL
90% CONFIDENCE INTERVAL
95% CONFIDENCE INTERVAL
99% CONFIDENCE INTERVAL
-4
-2
0
2
4
6

SAMPLING DISTRIBUTION





SAMPLE SIZE = 50



Note that statistical power is different from the level of confidence. Power determines how often we reject the null (the CI does not contain ZERO). Level of confidence is how often a confidence interval drawn from a random sample will contain the TRUE SLOPE. Here we observe TYPE I ERRORS in 7 out of 100 cases (the sample CI does not contain the true slope). But note that we would fail to reject the null in more than half of the samples (p-values > alpha), even though the true slope is not zero. This means the sample framework has low power (ability to detect actual effects).

Point estimate plus the standard error (size of the CI) determines the p-value. The p-value is roughly the probability that the slope = 0

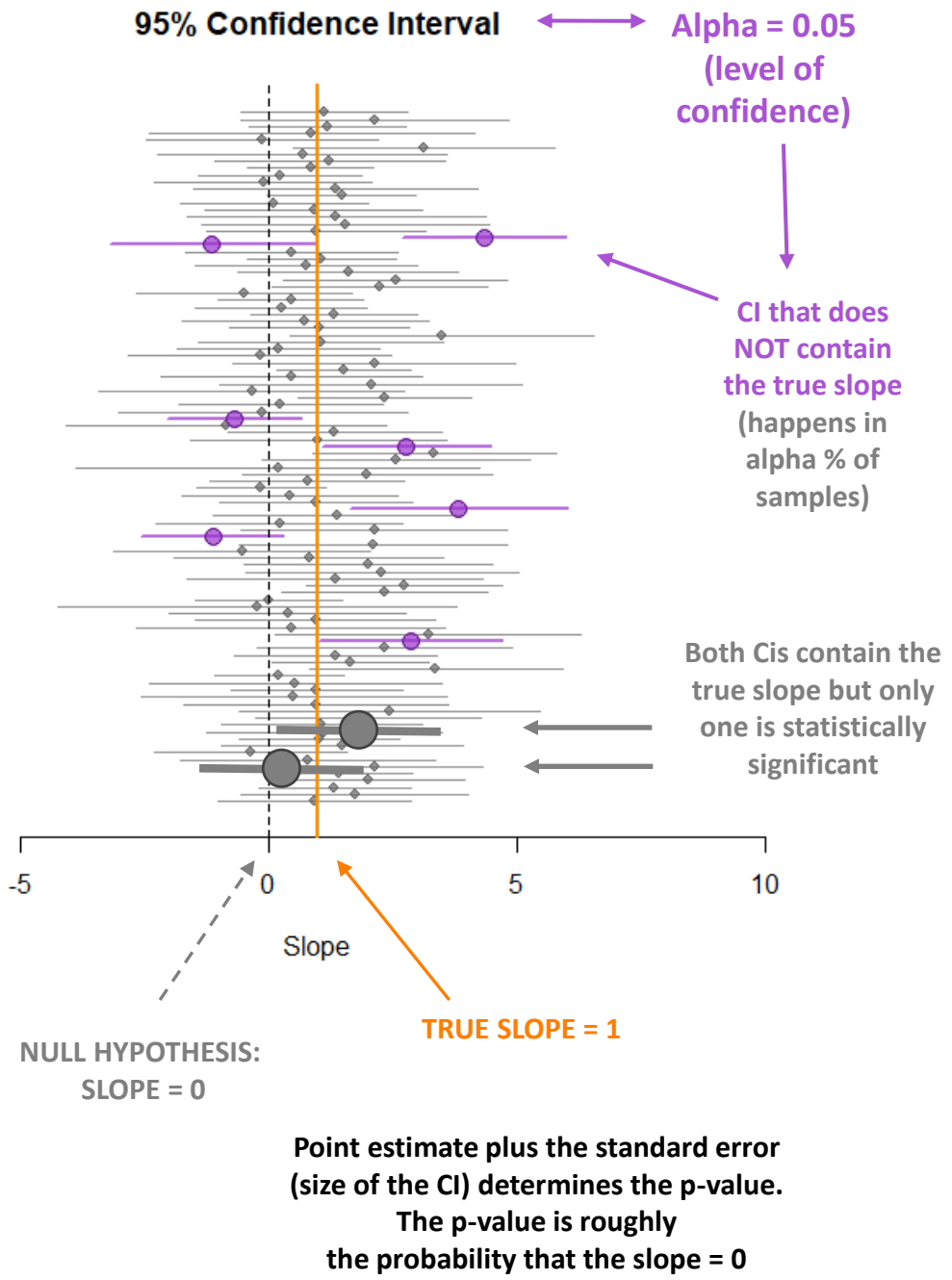
Note that we are using TYPE I ERROR, or a FALSE POSITIVE, in two distinct ways.

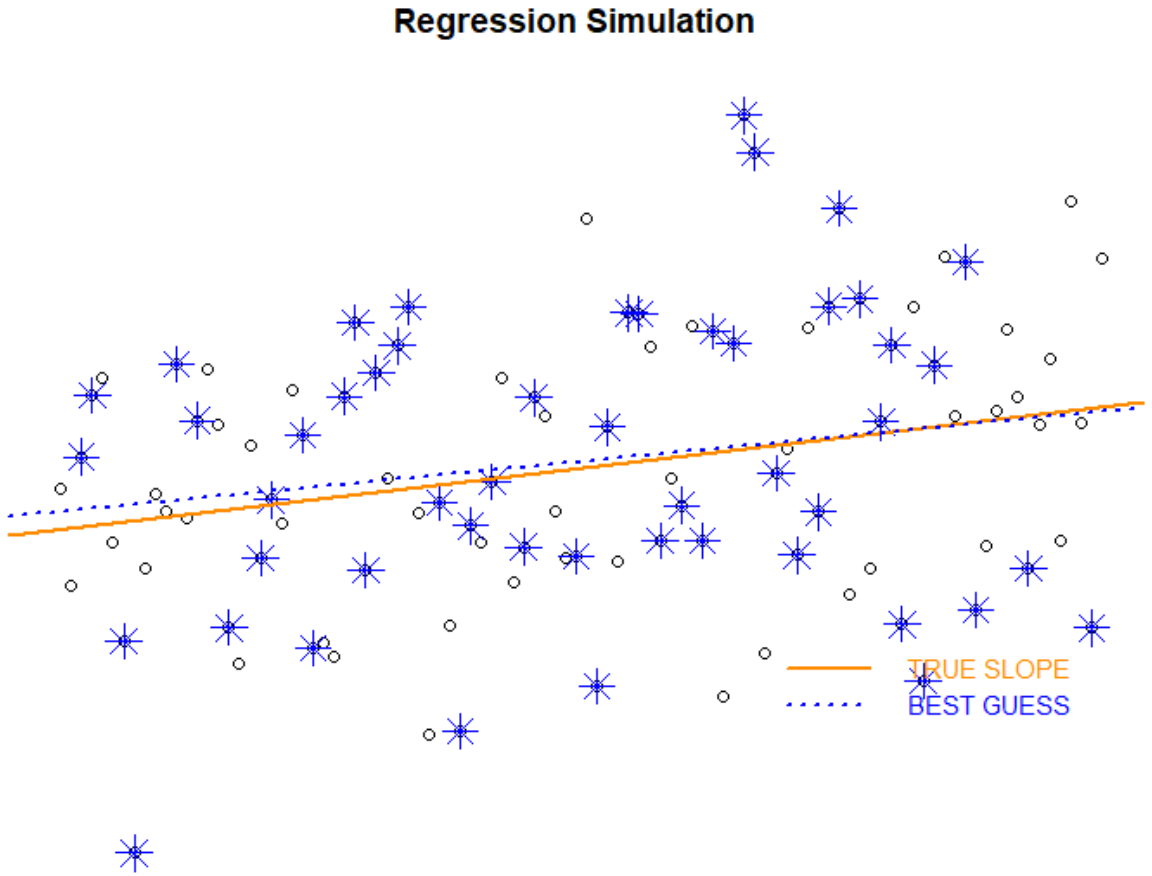
The first type of FALSE POSITIVE is linked to the level of confidence (alpha), which determines how often we will tolerate confidence intervals that fail to contain the true slope. This is the statistical sampling version that uses the **true slope as the NULL**. The purple CIs are all samples that would lead us to conclude that the slope is NOT equal to one (even though it is).

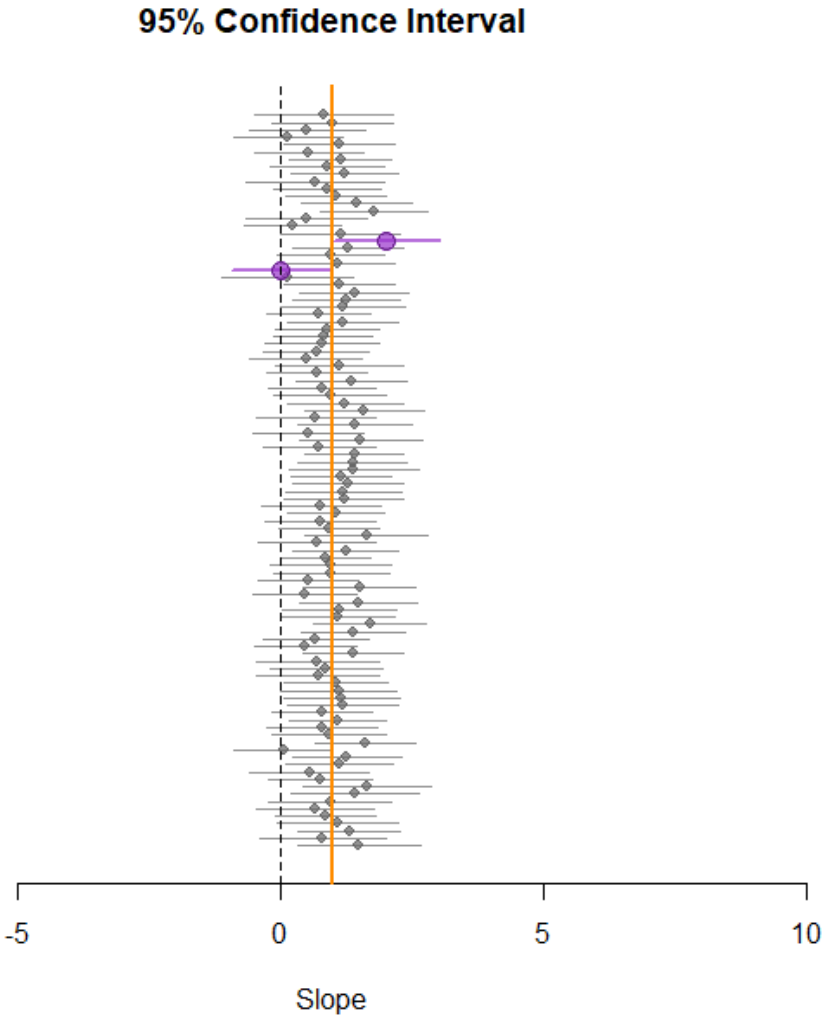
The more important version is the program evaluation context, in which case a TYPE I ERROR or a FALSE POSITIVE means we conclude that a program produced a meaningful change when it did not. This version assumes the **NULL is zero** (no program effect). The p-value produced by a regression model corresponds to this version. Here, since about half of the Cis contain zero, the p-value would be around 0.50.

Similarly, a **TYPE II ERROR** (a **FALSE NEGATIVE**) is when we can't differentiate the true slope from the NULL, even when they differ.

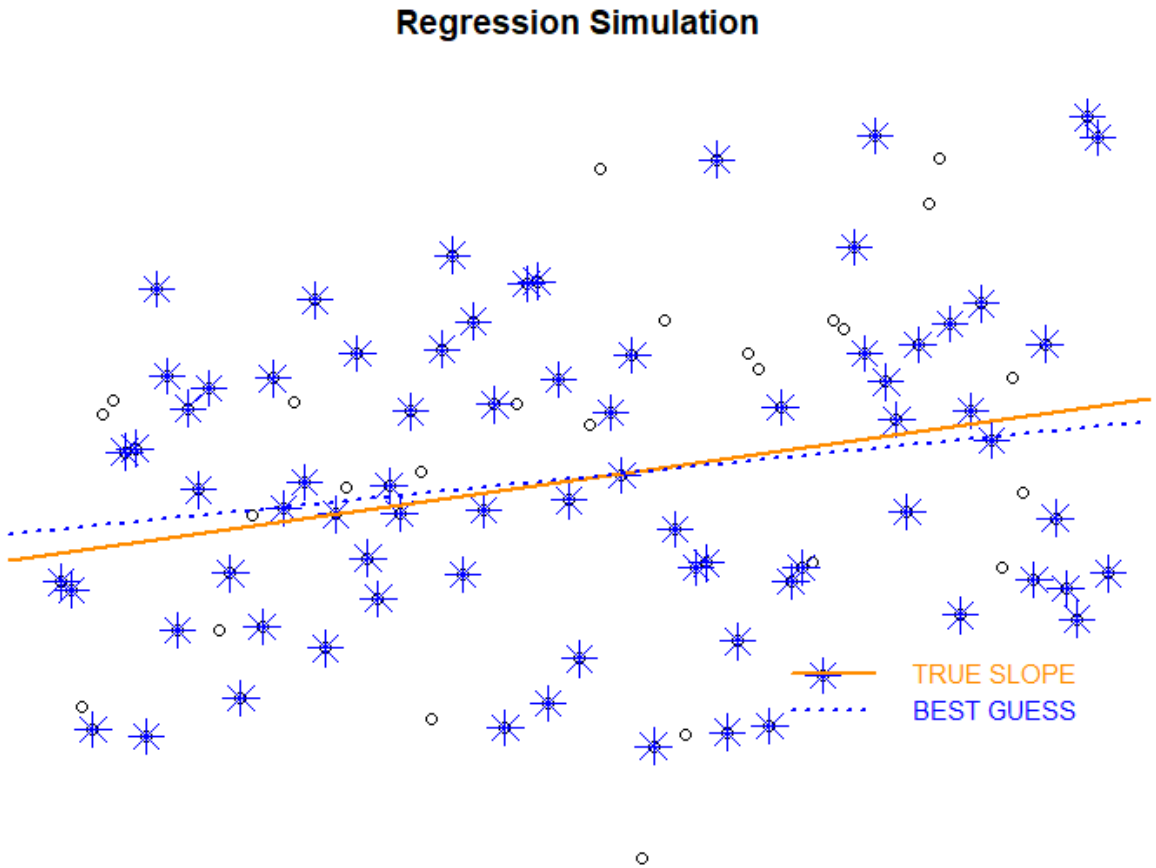
STATISTICAL POWER is the probability of identifying a specific program effect (slope or effect size) using a specific sampling framework. In this case power is very low.

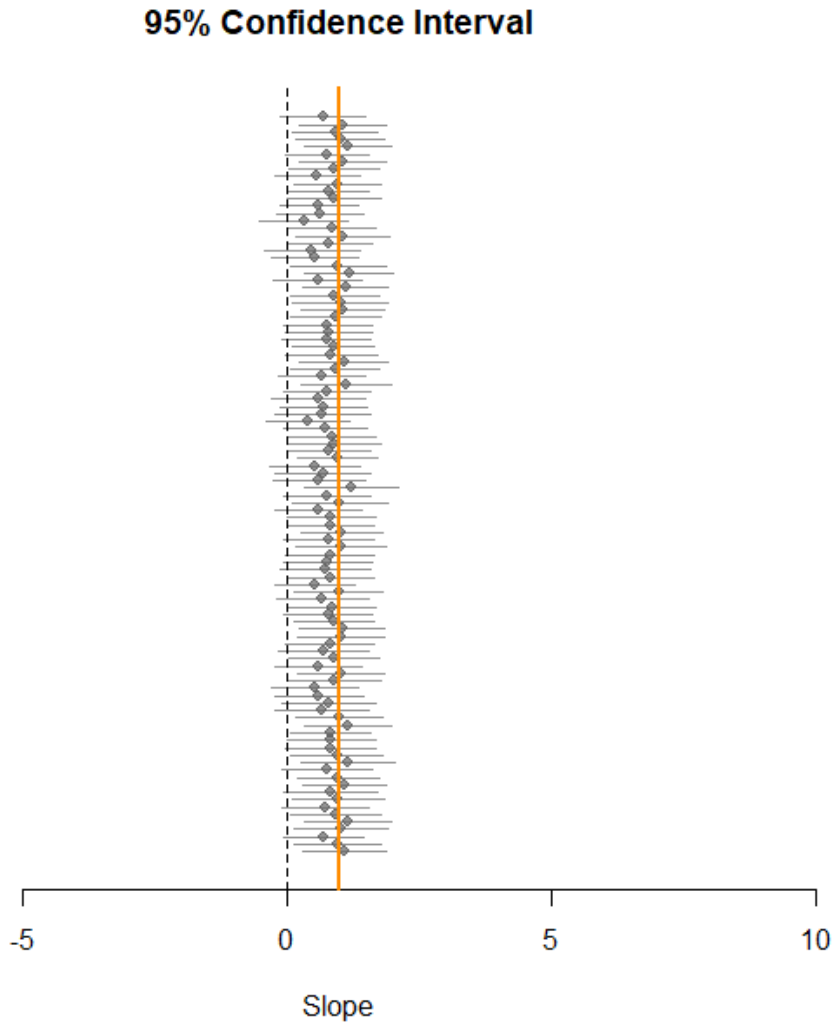


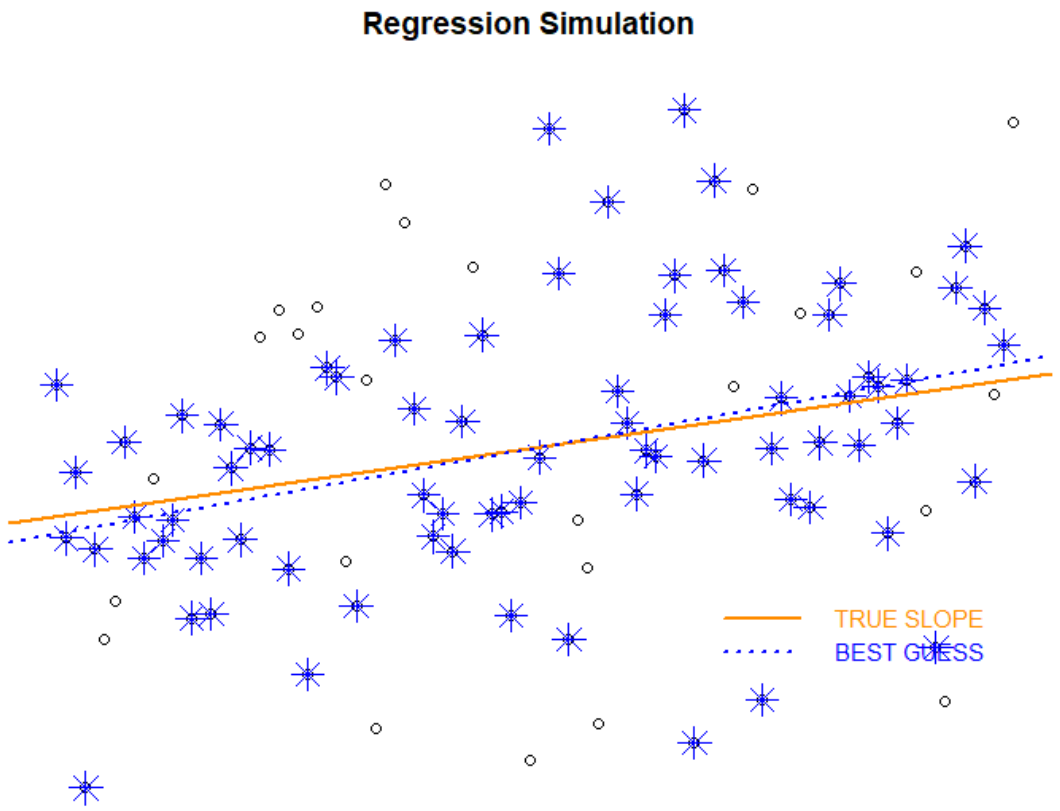


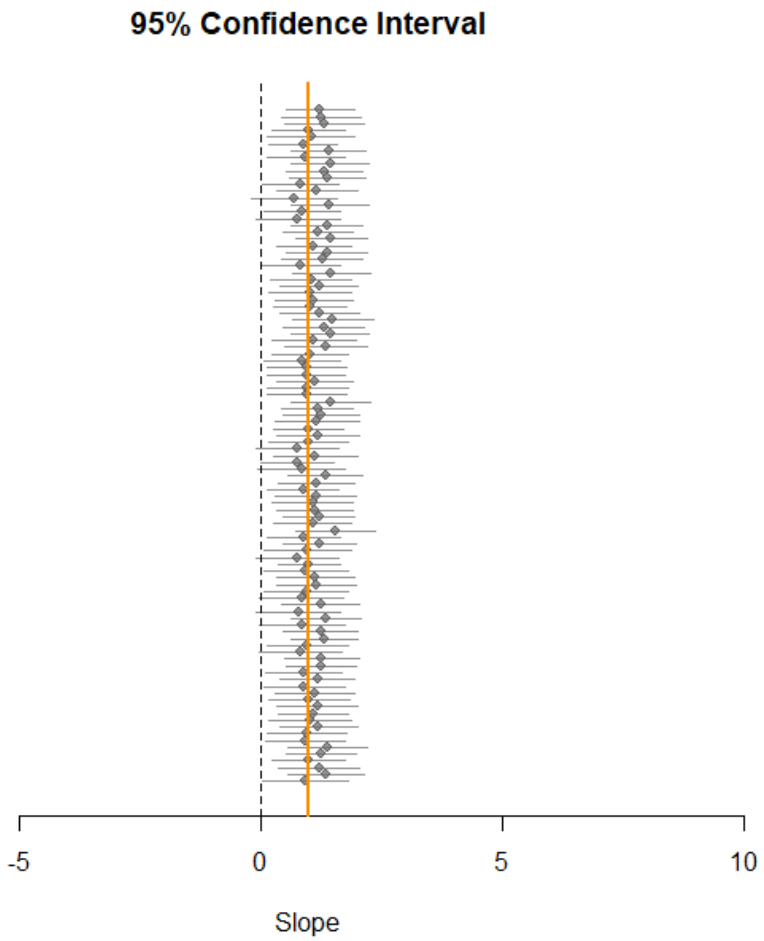


We can increase STATISTICAL POWER by increasing the sample size or adding control variables, thus reducing the standard error and shrinking confidence intervals. Increasing the sample size from 10 to 50 has improved our ability to detect program effects. We are only failing to reject the NULL in 1/3 of cases instead of ½ of cases.

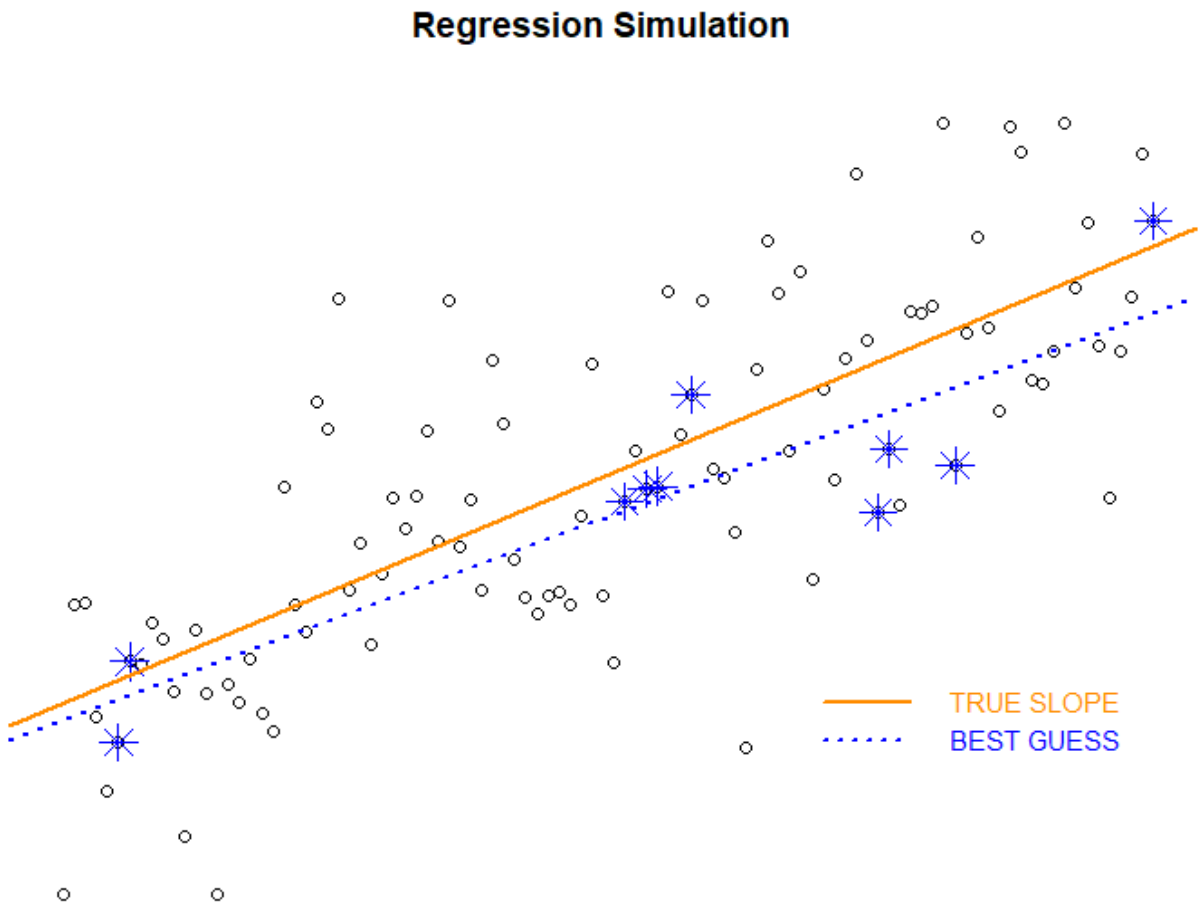

Regression Simulation
TRUE SLOPE
BEST GUESS

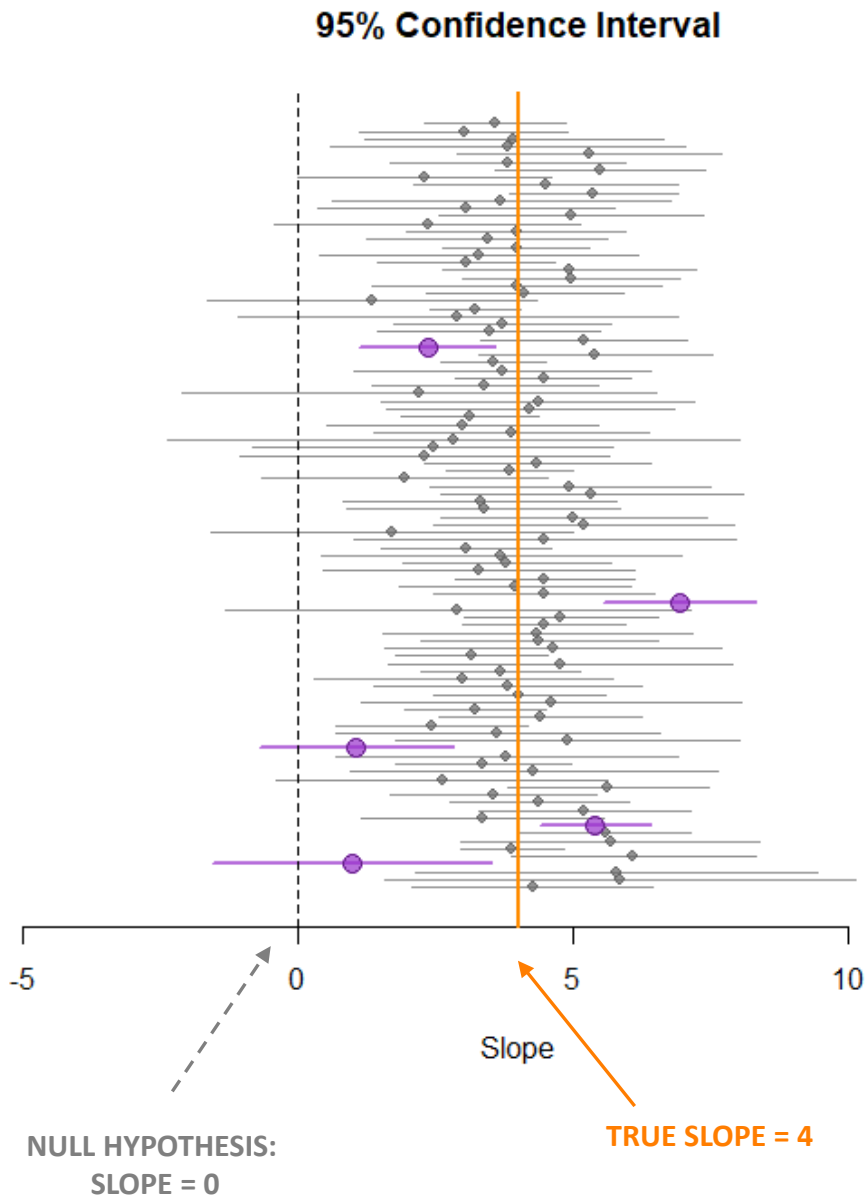

95% Confidence Interval
-5
0
5
10
Slope





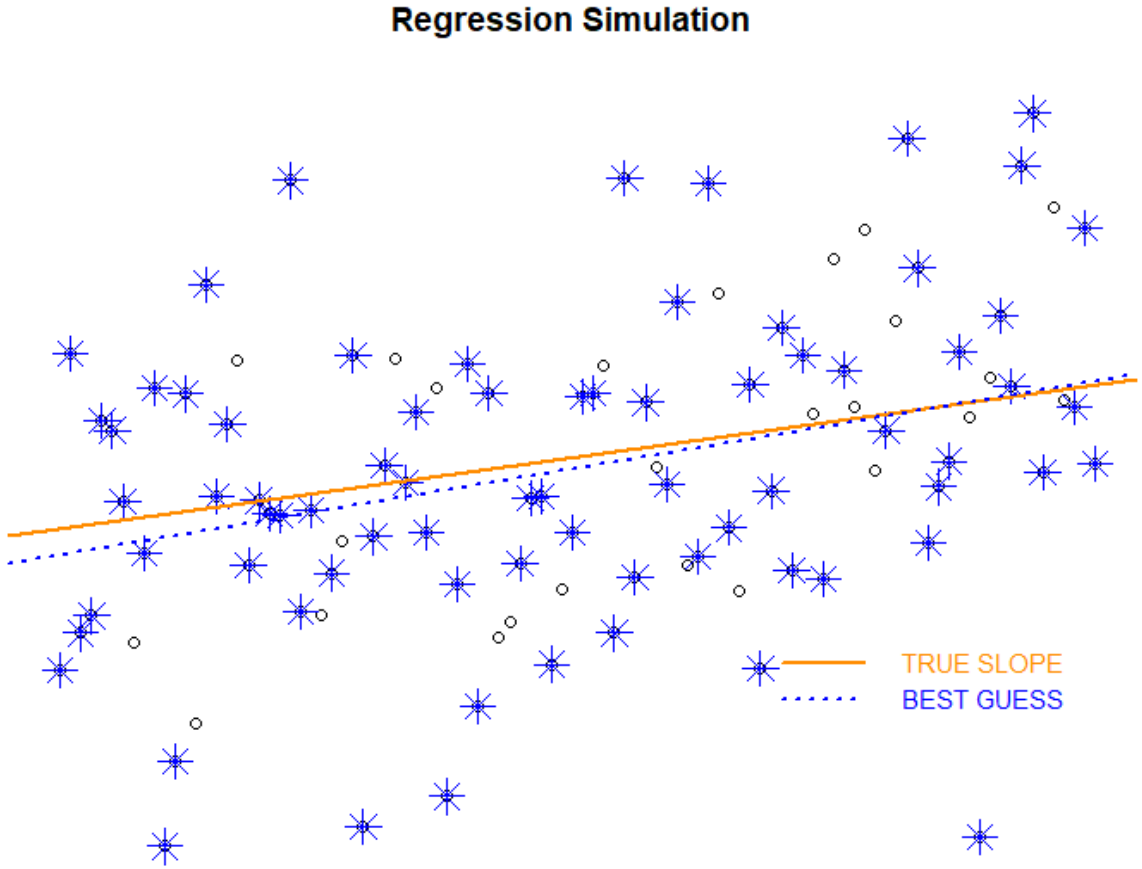
We do even better when the sample size is increased to 75. Now we only fail to identify program effects around 5% to 10% of the time in these case (about 5 or 10 CIs still contain zero).

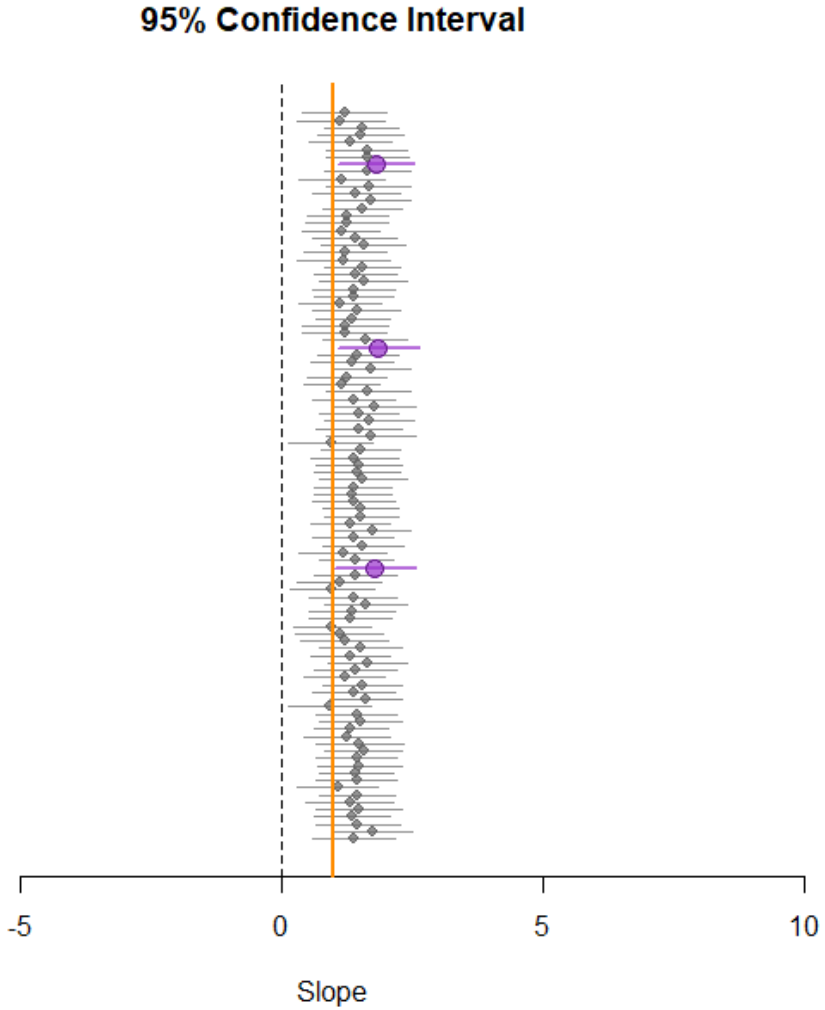




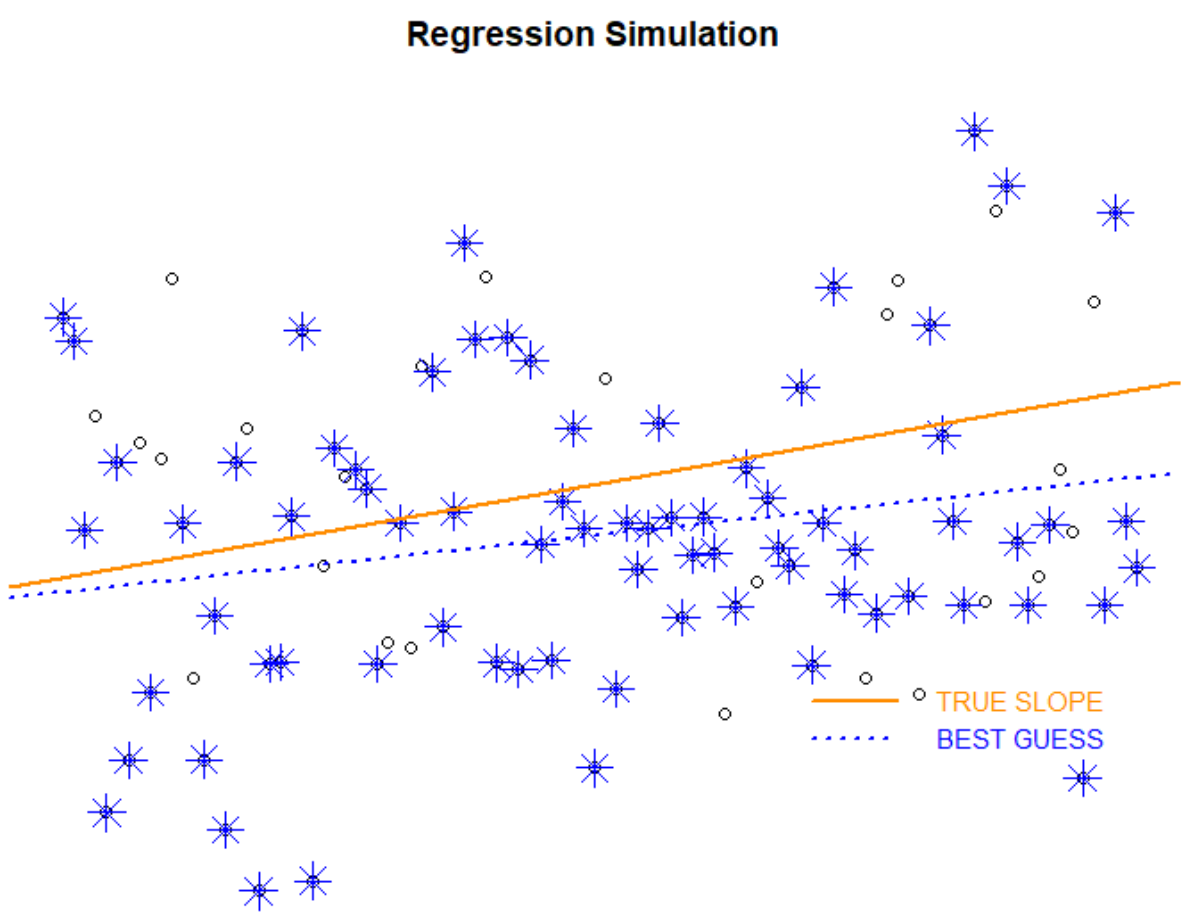
If the sample size remains low (n=10) but the program impact is large (slope of 4 instead of slope of 1) we achieve a lot more statistical power.

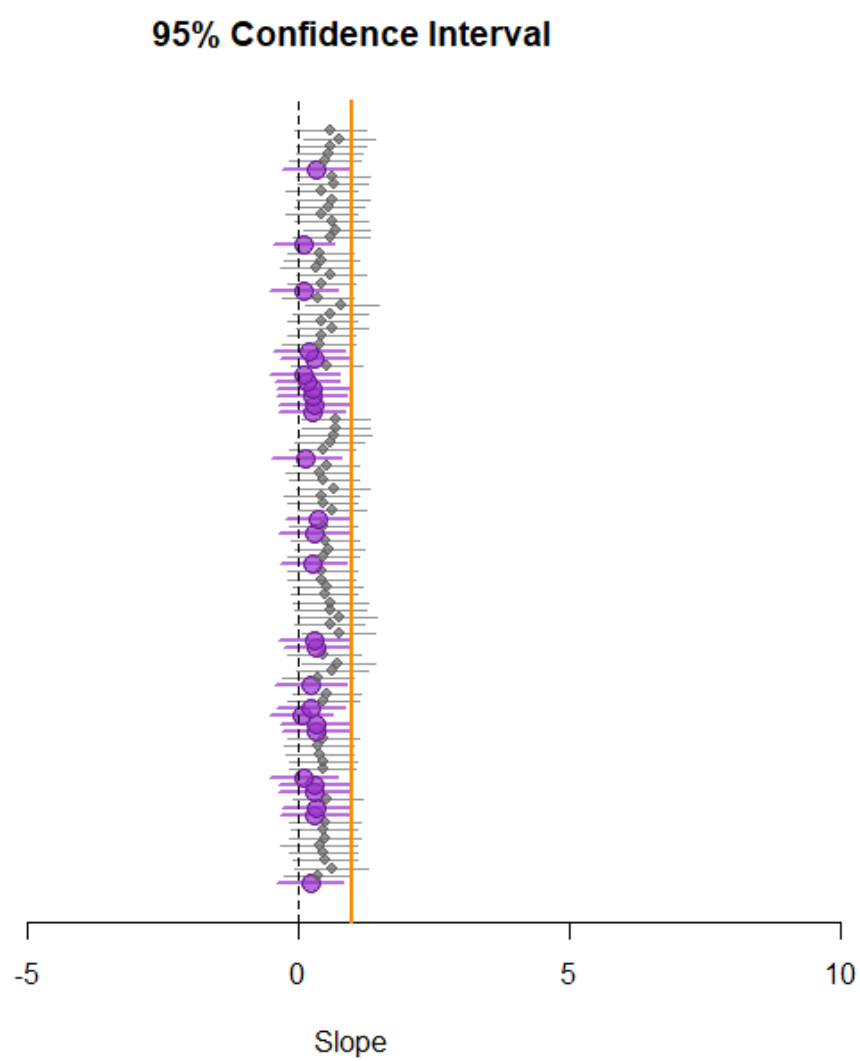
This should be somewhat intuitive - it is easier to detect large changes than small changes.



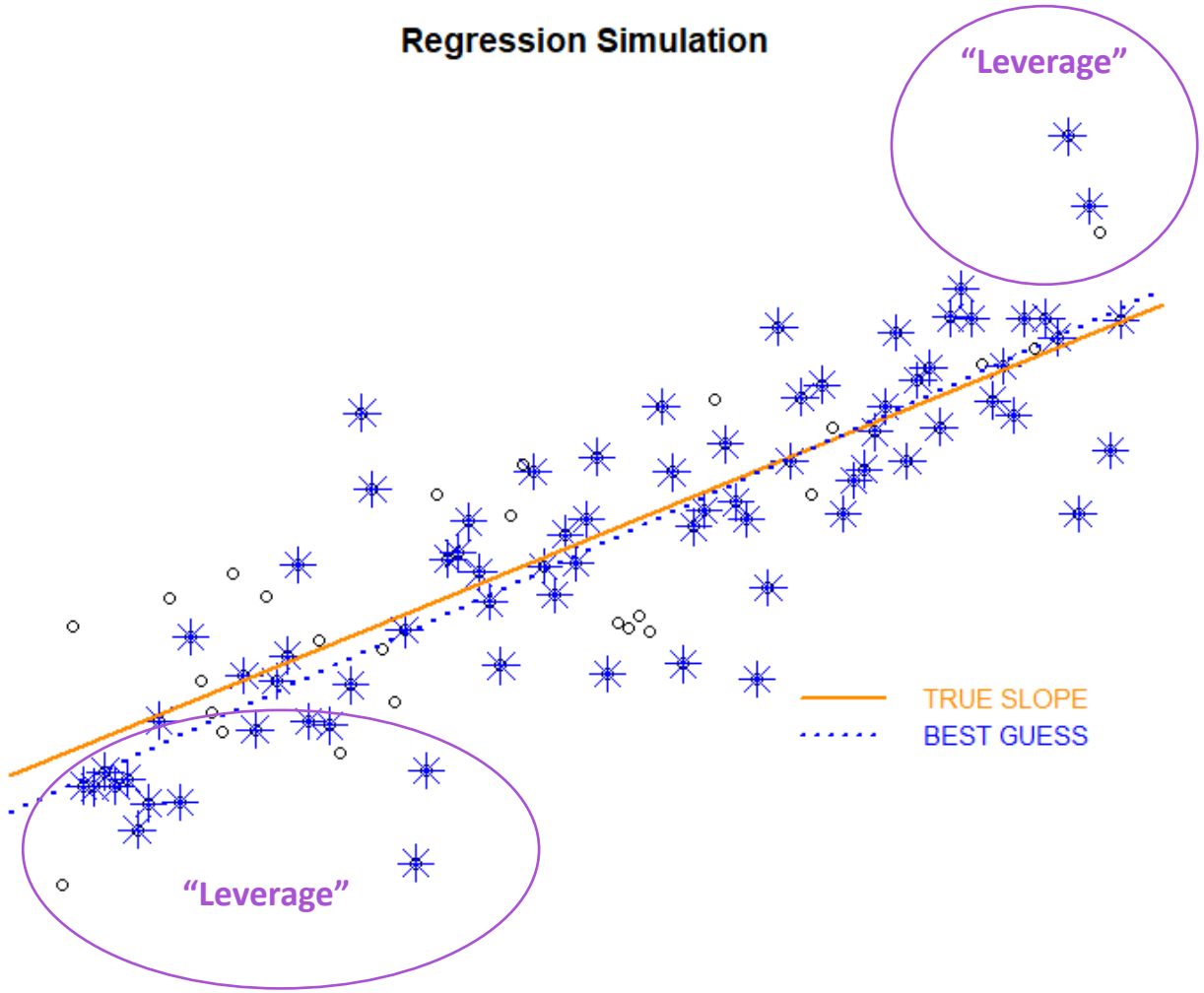


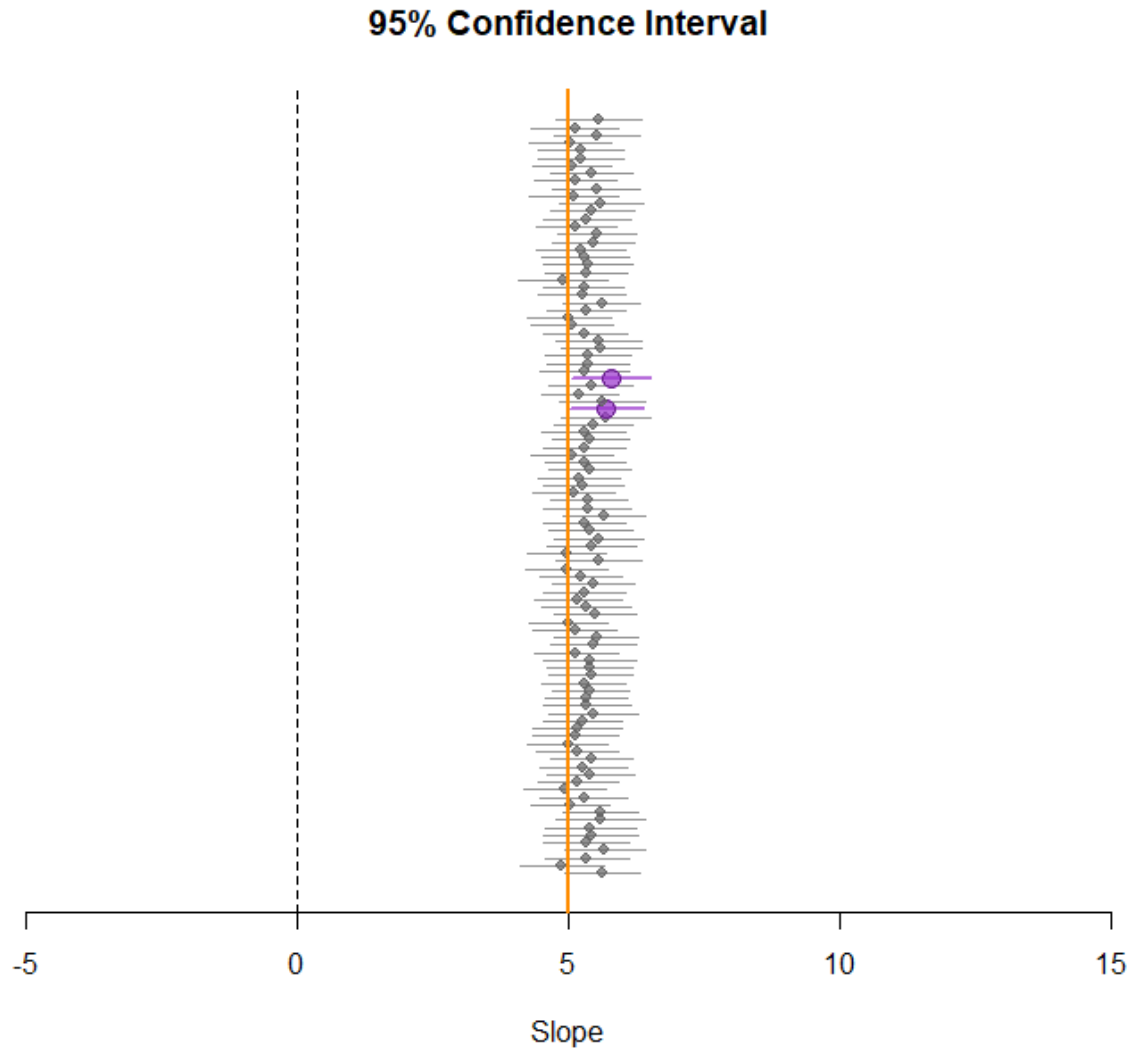
Any idea what is happening here? The sample slope estimates are systematically too large, even though they represent random samples? Hint, it is a type of specification bias.

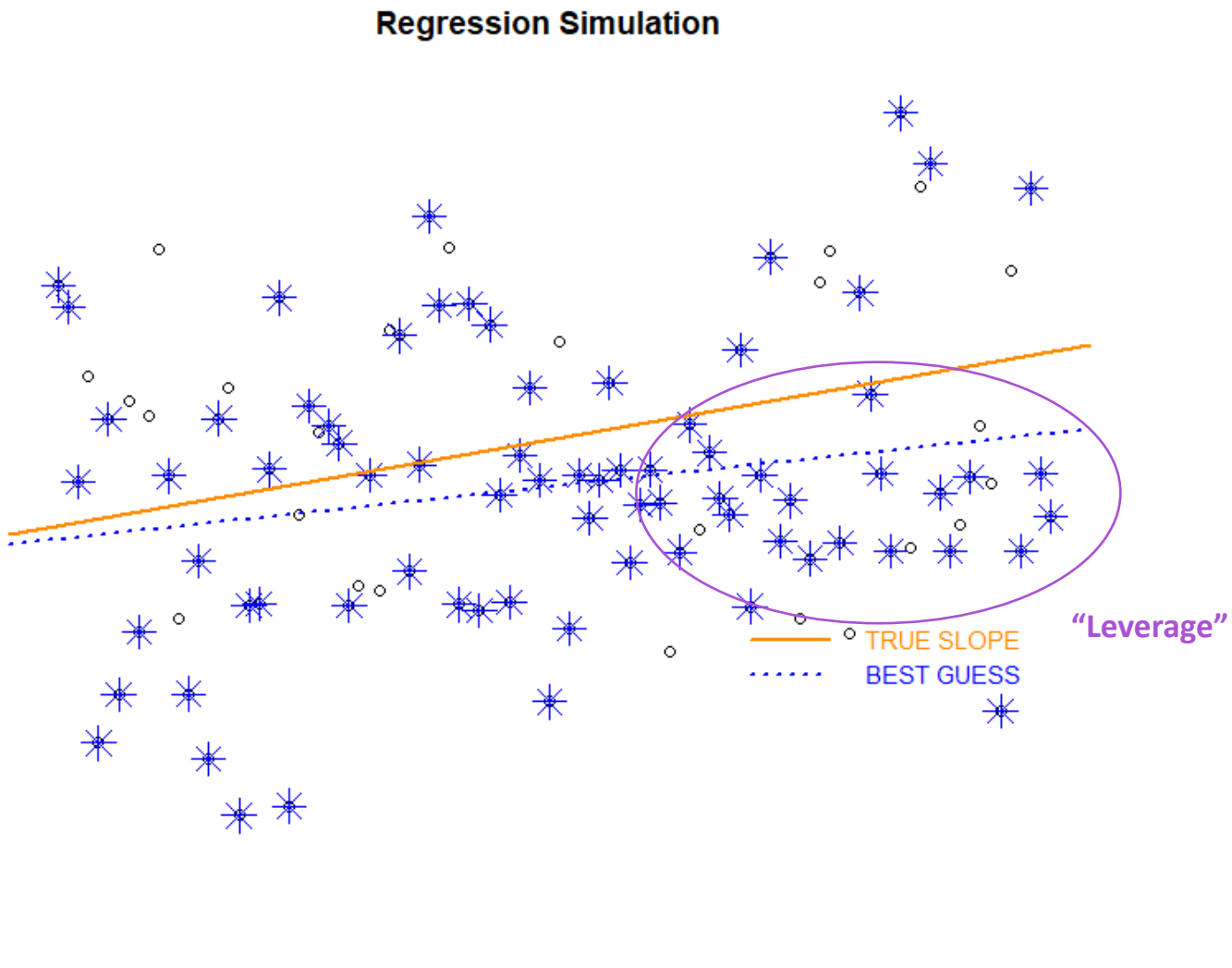


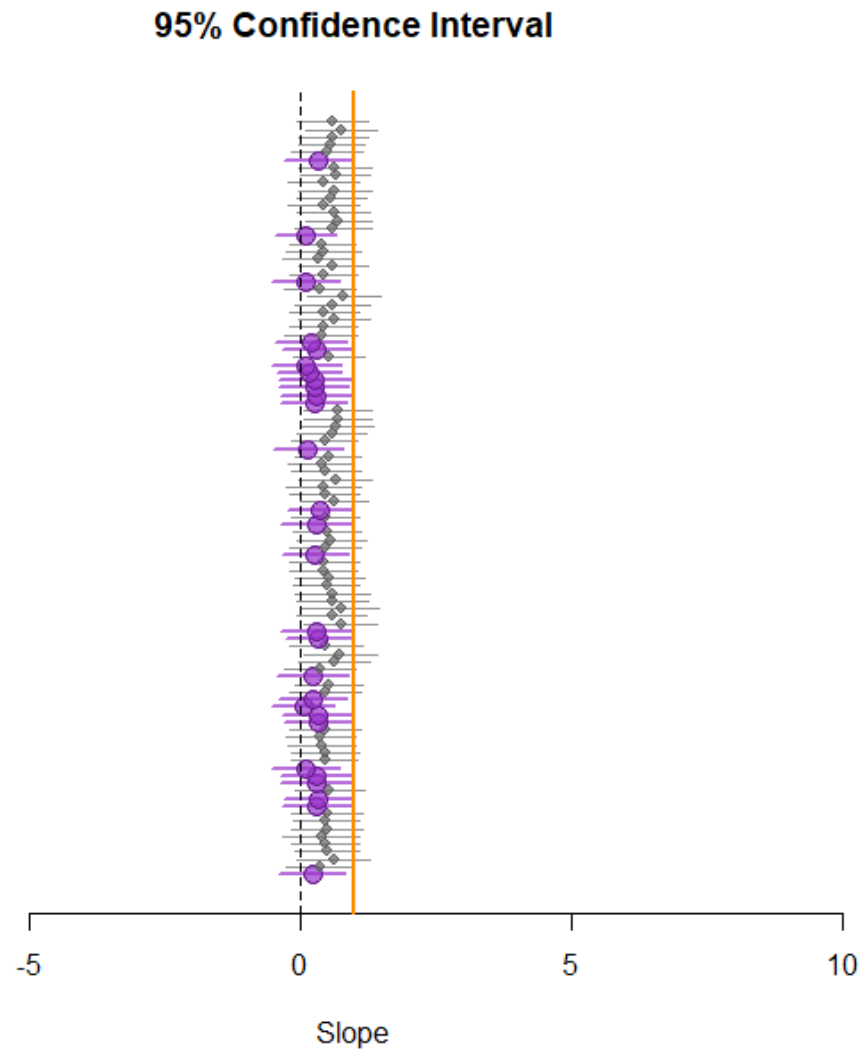


And similarly, here?


Regression Simulation
“Leverage”
TRUE SLOPE
BEST GUESS
“Leverage”


95% Confidence Interval
-5
0
5
10
15
Slope


Regression Simulation
TRUE SLOPE
BEST GUESS
“Leverage”


95% Confidence Interval
-5
0
5
10
Slope