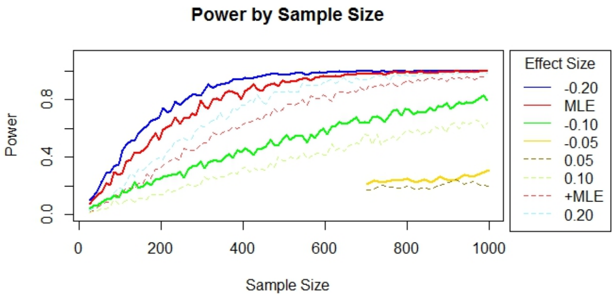
STA 723 HW 2 **Zach White**

Power Analysis

Power $(1-\beta)$ refers the probability that a study can correctly reject a false null hypothesis. The goal of a power analysis is to determine the necessary sample size to achieve a certain level of power given an α-level and effect size (the desired coefficient for distance under a logistic model). The α-level refers to the probability of rejecting a true null hypothesis. We perform a power analysis under the same construct as last week's project, but we will focus specifically on the effect of distance while holding arsenic and education levels constant.

We take advantage of the fact that we have previous data from the project last week. We set $\alpha = 0.05$, and we will scale the distance by 20 meters. We begin by determining a generative model based on the variables that are given. We use the variables distance, arsenic, and education to model whether or not someone will switch through logistic regression. We begin by fitting a logistic model to find the maximum likelihood estimates for the coefficients for $\beta_{intercept}$, $\beta_{arsenic}$, β_{dist}, and β_{educ}. We then fix the coefficients for all of the β values except distance, which we will replace with the effect size and then test to find the required sample size to detect that effect size. This describes our generative model.

To find the required sample size, we sample R datasets with replacement of size n from the overall dataset. We then use our generative model to find p_i for each of the n individuals and then simulate values for switching wells by drawing from $switch_i \sim Bern(p_i)$ for all n. We now have R pseudo-datasets, which represent a location of interest, and since this is actual data, we can preserve the covariance structure well. We then fit a logistic model to each of these datasets. We test to see if the $\beta_{distance}$ is significant. When we repeat this R times, we can simulate the power for the given effect size by calculating power$_n = \frac{\sum_{i=1}^{R} 1_{significant}}{R}$. This value is the power because we are testing whether our pseudo data can detect an effect that we determined. The plot shows the required n for the effect sizes: -0.20, -0.17, -0.10, -0.05, 0.05, 0.10, 0.17, and 0.20. The plot makes demonstrates that as effect size decreases, we need larger n to detect it. Also, as effect size increases, we require a lower sample size to detect this effect. Both of these phenomena show that this power analysis produces results that we would anticipate. An interesting result from this simulation is that the power curve isn't symmetric about $\beta_{distance} = 0$, which is clear by looking at these curves.



Conclusion

If we have great foresight and predict the MLE exactly, then it would require $n \approx 325$. Smaller effect size would require higher sample size n. There are two very clear limitations with performing power analysis this way. The first is that we require a similar study where we can use this bootstrap technique. This is certainly not a guaranteed assumption because most likely we won't have population representative of our area of interest. The second limitation is that we have an underlying assumption that the true model is a logistic because that is our generative model. Thus if the true model isn't logistic, we are overestimating our power significantly, and our estimations through simulation are completely wrong. An alternative to this approach is using finding someway to estimate the covariance matrix of the underlying data so we can maintain the relational structure of the data. If we can do this and are reasonably certain the variables are multivariate normal, then we can generate draws from this distribution and follow a similar process. This doesn't address the issue of assuming the wrong model, but it does address the issue of not having data a priori. There is another possible approach using a normal approximation for the power curves.