Power Analysis of a Logistic Regression Model Used to Predict the Probability of Students Taking Quizzes

Introduction

A course has ten quizzes, and a researcher wishes to study the probability of a student taking each based on a series of factors. The factors considered are two correlated variables (x and z), the quiz taken (denoted by j), and an interaction term between j and x. In particular, the researcher is interested in the interaction term between j and x. The probability is estimated using logistic regression, which results in the following conditional mean equation:

$$E(y|x,z,j) = p(x,z,j) = 1/(1+\exp(-\beta_0 - \beta_1 x - \beta_2 z - \beta_3 j - \beta_4 xj))$$

Note that x and z are separately standardized, $p(x,z,j)$ is expected to decrease as j increases, the indicators y_j are correlated for each student, and each student has probability q_j of dropping the course after taking quiz j. It is expected that q_j increases with j and after a student drops the class, no more information about the student is considered. Overall, it is assumed that a student has approximately a 50% chance of not dropping the course at all.

This memo discusses the statistical power of this setup under different contexts. In particular, it shows what happens to the model's power when varying sample size, the correlation between x and z, the correlation among y_j for each student, and β coefficients.

Generating data

The following sections explain how each of the aforementioned variables was simulated:

Correlated x and z

Since x and z are standardized, I decided to sample from a standard bivariate normal distribution to model them. The corresponding distribution's covariance matrix contained 1 in the diagonal entries and r_{xz} in others. Numpy's multivariate_normal function allowed me to obtain an arbitrary amount of samples from this distribution. Only one pair of x and z was used for each student (different values of these were not used at different values of j for each student).

Probability of taking quiz

Calculating the probability of taking a quiz requires using the logistic function. As per the model's setup, this function depends on x, z, j, and a β vector. For each student, the generated x and z, along with an arbitrarily chosen β, were plugged into the logistic formula to calculate the probability of taking quiz j.

Correlated y_j

Under the assumption that the logistic model defines the true conditional probability of completing a quiz, $p(x, j, z)$ defines the probability of y_j for each student. y_j is then a binomial random variable defined by $\mathrm{Bin}(1, p(x, z, j))$. In order to correlate different y_j, a copula was used as follows:

- Ten correlated samples were generated from a standard normal distribution. Each pair of variables has correlation r_2.
- The ten obtained samples were plugged into the standard normal's cumulative density function. The result was ten correlated uniform variables.
- The uniform variables were plugged into their respective binomial inverse cumulative density functions $\mathrm{CDF}^{-1}_{\mathrm{Bin}(1,p(x,z,j))}(u_j)$.
- The result was ten correlated $y_j \sim \mathrm{Bin}(1, p(x, z, j))$.

Last quiz of a student

The following values of q_j were assumed:

$$[0.01465, 0.0293, 0.04395, 0.0586, 0.07325, 0.0879, 0.10255, 0.1172, 0.13185]$$

And notice that

$$\prod_{j=1}^{9}(1 - q_j) \approx 0.5$$

For each student, nine indicator variables were sampled $I_j \sim \mathrm{Bin}(1, q_j)$. The minimum j so that $I_j = 1$ indicates the student's last quiz. If a student has a last quiz, that means the student dropped the course, and all the student's data was removed after the last quiz.

The following table shows an example of data generated for two students:

Student ID	j	y	x	z	Last Quiz
0	1	0	-0.739155	0.160882	3
0	2	1	-0.739155	0.160882	3
0	3	1	-0.739155	0.160882	3
1	1	0	0.437305	-0.990232	2

1	2	0	0.437305	-0.990232	2

Power Analysis

For the following power estimations, the probability of making a type I error is controlled at $\alpha = 0.05$.

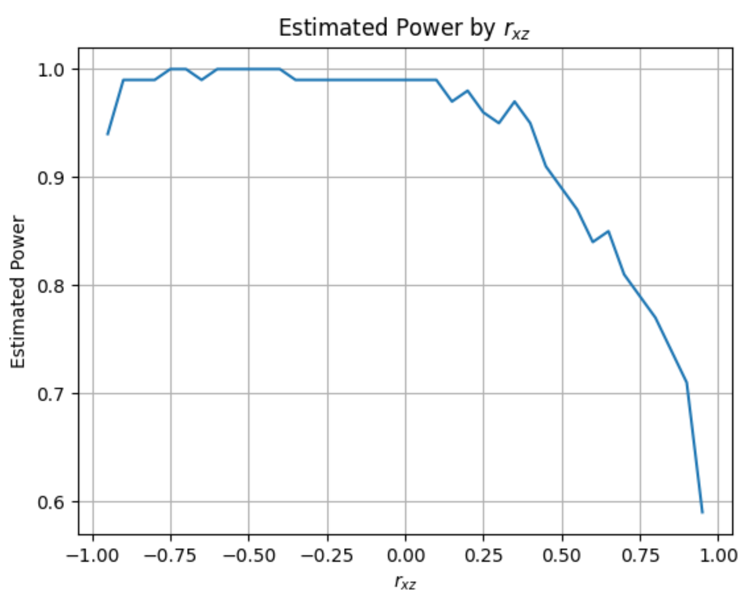
Correlated x and z

In order to study the relationship between r_{xz} and the statistical power of this setup, the power was estimated at different simulated values of r_{xz}. In particular, power was estimated for 39 different values of r_{xz}. These values range from -0.95 to 0.95 in 0.05 increments. In order to estimate the power for a particular value of r_{xz}, 100 samples of data, each simulating data for 100 students, were obtained. For each of these samples, the p-value for the following hypothesis setup was calculated:

$$H_0 : \beta_4 = 0$$
$$H_a : \beta_4 \neq 0$$

If the p-value was greater than α, a type II error was made, and the power was estimated as $1 - P(\text{Type II Error})$, where $P(\text{Type II Error})$ was estimated as the observed type II error rate. One thing to note, the values of the beta coefficients used were $\beta = [1, 2, 3, -0.3, -0.2]$ and the y_j were generated through a correlation of 0.7. Negative values of β_3 and β_4 were used since it is speculated that students are less likely to take quizzes as j increases.

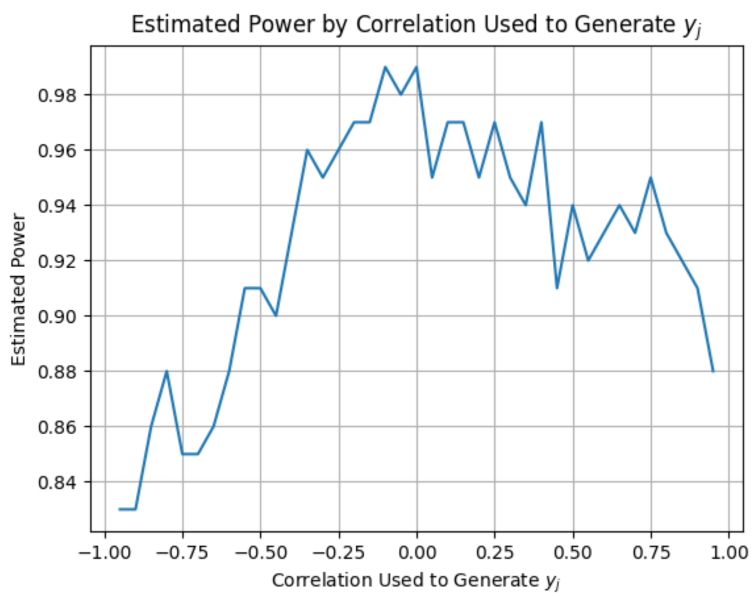
The result is the following:



The above plot shows that as the correlation of r_{xz} increases, the statistical power decreases. Therefore, the researcher should consider a different model setup in case the correlation between x and z is known to be high.

Correlated y_j

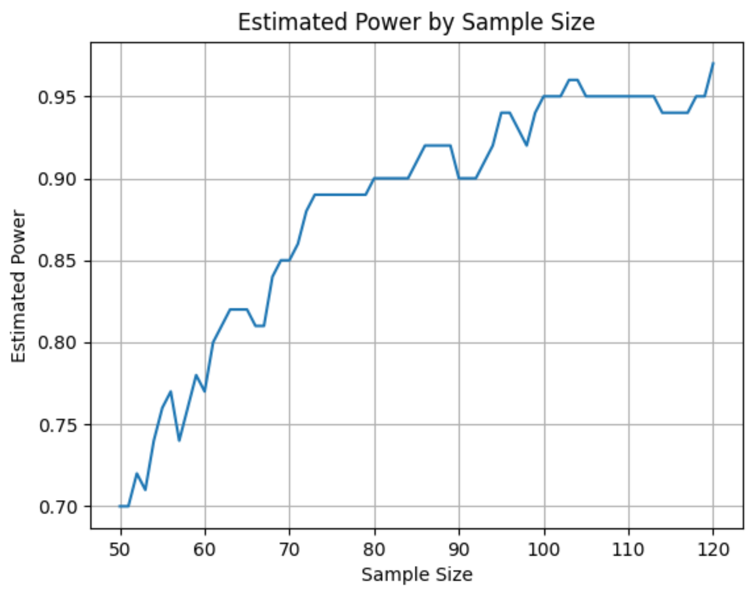
The correlation among different y_j and its relationship to the statistical power of this model setup was studied in the same way r_{xz} was studied. The only difference is that r_{xz} was set to be constant at $r_{xz} = 0.4$ and the y_j were generated at different correlation levels, these ranging from -0.95 to 0.95 in 0.05 increments. In this case, the result is the following:



From the plot, power is maximized whenever the correlation used to generate y_j is close to 0. However, notice that the estimated power in these cases is above 0.82. So, the correlation used to generate y_j does not seem to impact the power of the setup as much as r_{xz} does. If possible, the researcher should consider controlling for the correlation among y_j in order to maximize power.

Sample size

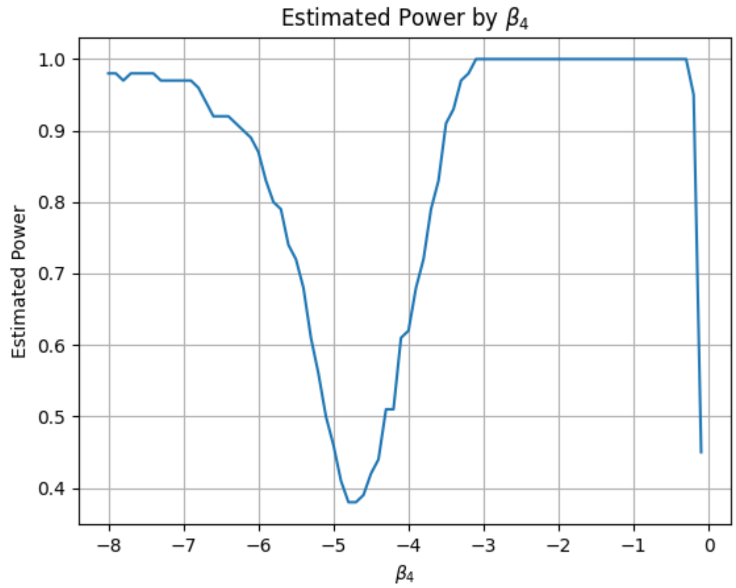
Sample size and its relation to the statistical power of this model setup was studied in the same way r_{xz} was studied. However, in this case r_{xz} was set to 0.4 and the power for samples sizes 50 to 120 were estimated; the result is the following:



It is natural to see that the higher the sample size, the higher the statistical power. However, the above plot could help researchers determine appropriate sample sizes that achieve specific levels of power. In particular, the values above are for $\beta = [1, 2, 3, -0.3, -0.2]$, $r_{xz} = 0.4$, $\alpha = 0.05$, and 0.7 used as a correlation to generate y_j. Using these values as reference, and considering the negative effects that correlations (r_{xz} and among y_j) have over power, they could get an idea of the sample sizes needed to obtain a particular desired level of power.

Different β_4 values

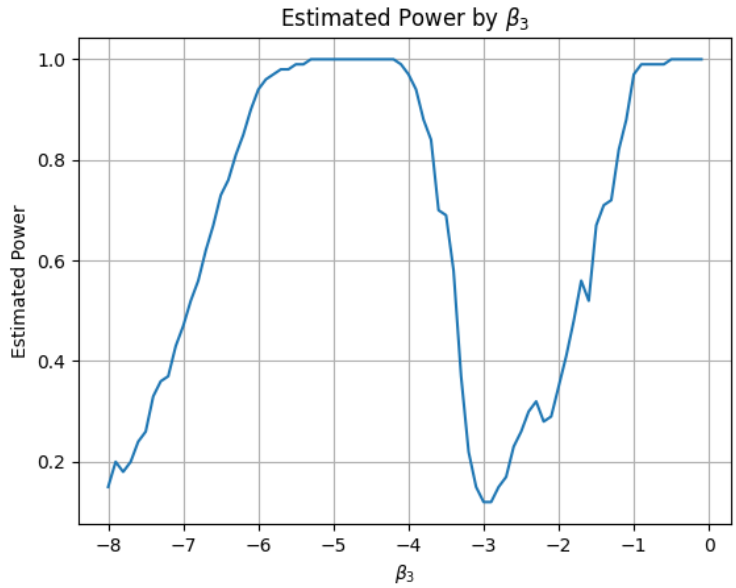
β_4 and its relation to the statistical power of this model setup was studied in the same way r_{xz} was studied. In this case, $r_{xz} = 0.4$ across all different values of β_4. Additionally, power was estimated for values of β_4 between -8 and -0.1 in 0.1 increments. The result is the following:



For some reason, values of β_4 close to -4.9 have low power compared to values away from it. The effect of β_4 over power is so significant that it might be worth investigating further in case the researcher has reasons to believe that the true value of β_4 is close to -4.9 and the other values in the model are close to those assumed in the plot generated above. Another thing to consider about the plot above is that it is understandable that values close to 0 have low power since the effect of β_4 becomes weaker as β_4 approaches 0. It should also be noted that power increases as the magnitude of β_4 increases since the effect of β_4 becomes more and more important for the data.

Different β_3 values

β_3 was studied in the same way β_4 was studied. In this case, $\beta_4 = -0.4$ across all different values of β_3. The result is the following:



The above plot shows that the effect of β_3 over power is significant. Depending on the values of β_3, it is possible to achieve expected powers smaller than 0.1 and greater than 0.95. The above plot shows that, under this model setup, very negative values and values close to -3 have relatively low statistical power.

Conclusion

Data was simulated according to the model's specification. From the data, statistical power was estimated for different combinations of parameters. It was concluded that sample size, r_{xz}, β_3, β_4, and the correlation among y_j can significantly affect the statistical power of the logistic regression model assumed.

Code

The code for this analysis can be found [here].