Comparing the Hosmer-Lemeshow Goodness of Fit Test With Varying Number of Groups to the Calibration Belt in Logistic Regression Models

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Jason Andrew Benedict, BS

Graduate Program in Public Health

The Ohio State University

2016

Master's Examination Committee:

Professor Rebecca Andridge, Advisor

Professor Stanley Lemeshow

Copyright by

Jason Andrew Benedict

2016

Abstract

Logistic regression is a commonly used statistical technique in business and the sciences when an outcome is binary. For example, clinical trials may employ a logistic regression model when an outcome is presence or absence of disease, or a business may use such a model when the outcome is the presence or absence of a customer's purchase of a product. An ideal logistic regression model both discriminates well and is well-calibrated. A well-calibrated model is one where the predicted percentages of success are close to the observed percentages.

The Hosmer-Lemeshow test is a commonly used goodness of fit test that is used to test the calibration of a logistic regression model. The Hosmer-Lemeshow test becomes too powerful as the sample size increases, and an adaptive equation was recently proposed by Paul *et al.* (2013) to recommend the number of groups to use as the sample size increases. A new method to test the calibration of a logistic regression model, the calibration belt, was recently proposed by Nattino *et al.* (2014).

The purpose of this study is to compare the power of the calibration belt with the Hosmer-Lemeshow test through simulations of several models with differing deviations from the true model and various probabilities of success. The Hosmer-Lemeshow test is applied to the models with varying number of groups (from g=6 to g= 5000), including the number of groups recommended through the adaptive equation proposed by Paul e*t*

*al.* (2013). The type 1 error rate of the calibration belt and the Hosmer-Lemeshow test is also assessed in all of these models.

The simulations show that the calibration belt is nearly always the most powerful test, but the type 1 error rate of the calibration belt is often significantly below the nominal rate of 5%. The Hosmer-Lemeshow test does not suffer from this problem. It is also shown that the adaptive group equation proposed by Paul *et al.* (2013) depends largely on the probability of success of each of the models.

## Acknowledgments

I would like to thank my thesis advisor, Dr. Andridge, who has done more for me over the last two years than can possibly be seen in these pages. I could not have asked for a better advisor. I'm very grateful for her instruction and guidance.

Dr. Lemeshow came up with the idea for this project, and he also provided the job opportunity that enabled me to gain experience building logistic regression models. I first learned about logistic regression by taking his class, and I was happy to take on this project after being inspired by his teaching. Of course, this project never would have happened if it weren't for his work developing the Hosmer-Lemeshow test to begin with!

I would also like to thank Giovanni Nattino. He helped me with any questions I had related to his calibration belt. It amazes me that he developed this test so early in his career. He will make an excellent professor in the near future.

Finally, I would like to thank the College of Public Health at The Ohio State University. I have had nothing but excellent experiences in this college, and that mostly stems from the excellent professors I have been fortunate to meet along the way.

# Vita

May 2005 .......................................................Gahanna Lincoln High School

2010.................................................................B.S. Biology, The Ohio State University

2015-2016 .....................................................Graduate Research Associate, College of Public Health, The Ohio State University

# Fields of Study

Major Field: Public Health

Table of Contents

List of Tables

List of Figures

# Chapter 1: Introduction

Logistic regression is one of the most commonly used methods in statistical modeling with binary outcomes. Binary outcomes are found in nearly all areas of study, including the medical, economic, and psychology fields. For example, Witt *et al.* (2004) used logistic regression to identify the association between several demographic factors and cardiac rehabilitation after myocardial infarction. After a logistic regression model is built, model fit is usually assessed. A model fits well if it has both good calibration and discrimination. A model that discriminates well can distinguish successes from failures with high accuracy. A model that is well-calibrated accurately predicts the probabilities of successes.

The Hosmer-Lemeshow test is a commonly used technique for assessing logistic regression model calibration that is included in most statistical software programs. The Hosmer-Lemeshow test first creates groups of observations based on estimated probabilities from the logistic regression model and then compares observed and expected probabilities within these groups. Recently, several papers have looked at issues of type 1 error and power with the Hosmer-Lemeshow test. It is well-known that the Hosmer-Lemeshow test, like all chi-square tests, becomes too powerful as the number of observations increases (Paul 2013). Paul *et al.* (2013) attempted to overcome this limitation by increasing the number of groups in the Hosmer-Lemeshow test through an adaptive equation.  This adjustment standardized the power, so that the test did not always reject when the model is not far from the truth.

Nattino *et al.* (2014) developed another method to assess the calibration of a logistic regression model.  They compared the type 1 error rate of the Hosmer-Lemeshow test versus their newly developed calibration belt.  They found that the Hosmer-Lemeshow test rejected the null hypothesis of good calibration more often than expected in scenarios where the event is rare.  They found the calibration belt, alternatively, to be closer to the nominal type 1 error rate in all scenarios.  Another advantage of the calibration belt is graphical.  While the calibration belt – like the Hosmer-Lemeshow test – is a global test, it can be graphed to identify the probability levels where the model fits imperfectly.  Although the Hosmer-Lemeshow test can also be viewed graphically, it involves collapsing probabilities to see any deviations from the observed and estimated probabilities.  As Hosmer, Lemeshow, and Sturdivant note in their book (Hosmer, Lemeshow, and Sturdivant 2013), as the number of groups increases, it becomes very difficult to distinguish a large departure between estimated and observed probabilities.  Since no collapsing occurs with the calibration belt, it is easier to see the probabilities where the deviations between observed and estimated probabilities occur.  This is a potential advantage of the calibration belt over the Hosmer-Lemeshow test.

The goal of this study is to compare both the power and the type 1 error rate of the Hosmer-Lemeshow test with varying numbers of groups -- including the adaptive model selection procedure proposed by Paul *et al.* (2013) – and the calibration belt.  A simulation study is conducted using the same six models originally used by Paul *et al.*

(2013) as the starting point, however, we also varied the probability of success for each model from .05 to .80 to explore the effect of changing the marginal probability of success. Two of the six models in the original Paul *et al.* (2013) paper differed only in their intercept, thus the number of models we ran was reduced to five. The Hosmer-Lemeshow test was performed using a wide range of number of groups (from 6 to 5000), and these were compared to the calibration belt.

## Chapter 2: Statistical Power and the Hosmer-Lemeshow Test

Statistical power is defined as the probability of rejecting the null hypothesis when the null hypothesis is false. Having high statistical power is a desirable feature of testing, but a test can also quickly become too powerful. A well-known flaw in the Hosmer-Lemeshow test is that the power of the test becomes too high as the number of observations increases (Hosmer, Lemeshow, and Sturdivant 2013) causing the test to always reject the null hypothesis that the model fits even if it does in fact fit well. For example, a statistician may build a model that would have a positive impact in a clinical setting, but it may never be used if tests indicate it fits poorly. This is especially likely to happen if a large data set is used. As Hosmer, Lemeshow, and Sturdivant (2013) note, a model that may be well-calibrated with few observations looks increasingly poorly fit as the number of observations becomes large even with the exact same model. They illustrate this point by starting with a model that fits well with a small sample size. They then duplicated the data multiple times (thus increasing the sample size) until the model no longer fits well according to the Hosmer-Lemeshow test. Thus, using the same model and the same data, a much smaller p-value was produced as the number of data points was increased.

Paul *et al.* (2013) attempted to standardize the power of the Hosmer-Lemeshow test by changing the most mutable part of that test – the number of groups taken. In most software packages, the number of groups defaults to ten. Paul e*t al.* (2013) ran simulations with multiple models and differing numbers of groups to see if the power changes according to the number of groups. They found that the power does in fact change, decreasing as the number of groups increases. They then developed an adaptive equation to standardize the power for sample sizes with up to 25,000 observations. There were two goals of their paper. The first goal was to show the relationship between the power of the test in relation to differing sample sizes, the amount of deviation of the model from perfect fit, and the number of groups. To do this, they simulated binary outcomes, Y, with the model below:

$$logit(P(Y=1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 Z + \beta_5 (Z * X_1)$$

In this scenario, $X_1$ and $X_2$ were standard normal variables, while Z was a binomial variable with n=1 and p=0.5. All three variables were independent of each other. Values of the parameters for each of the six models are listed in Table 1. They then fit the following model to the data,

$$logit(\hat{P}(Y=1)) = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

which differed from the six true models in different ways. Their first model deviated from the fitted model the greatest amount, as it included a quadratic term with $X_1$ and a large interaction between $X_1$ and Z. The second and third models fitted to the data deviated the least from the fitted model, while models four through six were intermediate cases. These models had a probability of success that ranged from 0.055 to 0.874.

| Model # | Model | $p_{\text{success}}$ |
|---|---|---|
| 1 | $logit(P(Y=1)) = -2 + X_1 + .2X_1^2 + Z - 2(Z * X_1)$ | 0.256 |
| 2 | $logit(P(Y=1)) = 2 + X_1 + Z + .5(Z * X_1)$ | 0.874 |
| 3 | $logit(P(Y=1)) = X_1 + Z + .5(Z * X_1)$ | 0.585 |
| 4 | $logit(P(Y=1)) = X_1 + .2X_1^2$ | 0.529 |
| 5 | $logit(P(Y=1)) = X_1 + .2X_1^2 + X_2$ | 0.528 |
| 6 | $logit(P(Y=1)) = -3 - X_1 - .2X_1^2$ | 0.055 |

Table 1. Simulation models used in Paul *et. al.* (2013).

Paul *et al.* (2013) found that the power of the Hosmer-Lemeshow test increased with sample size and decreased with the number of groups. Of note, when the probability of success was low (Model 6), the Hosmer-Lemeshow test did not follow a chi-square distribution when the number of groups was large. As a result, they believed that the Hosmer-Lemeshow test is not effective in instances when the probability of success is low.

Previous work has shown that the Hosmer-Lemeshow test works best when there are at least five observations per group, and when the number of groups is greater than or equal to six (Hosmer, Lemeshow, and Sturdivant 2013). The test often breaks down as well when the event is rare, as confirmed by Paul *et al.* (2013). Taking all of these into account, Paul e*t al.* (2013) listed recommendations for what group sizes to use in various scenarios. With sample sizes up to 1000, a group size of ten is recommended. This often keeps the power below 70%, which in some scenarios may still be too powerful. For sample sizes between 1,000 and 25,000 observations, they recommend using the following equation to determine the number of groups, *g*, to use:

$$g = \max\left(10, \min\left\{\frac{m}{2}, \frac{n-m}{2}, 2 + 8\left(\frac{n}{1000}\right)^2\right\}\right)$$

where *n* is the sample size and m is the number of successes. This formula is justified by noting that power was kept relatively consistent to a benchmark used with a sample size of 1000 and a group size of 10 in their simulation results when the equation $g = 2 + 8\left(\frac{n}{1000}\right)^2$ was used. Moreover, the assumption is made that the number of groups taken is never below 10. It is also noted that this equation breaks down as the sample size becomes smaller, as it is recommended to have at least five observations per group. Finally, for sample sizes greater than 25,000, this equation breaks down as well, as the equation defaults to the number of successes (m in the equation above) divided by two. This results in a test that is too powerful.

## Chapter 3: Type 1 Error, the Hosmer-Lemeshow Test, and the Calibration Belt

Nattino *et al.*(2014) recently developed a test, the calibration belt, that uses a regression model based on the expected and observed probabilities of the logistic regression model to assess model fit. The predicted probabilities are used as an independent variable in the model, and the observed (binary) outcomes are used as the dependent variable. The calibration belt fits a logistic regression model that is restricted to a polynomial equation up to degree four. If the calibration belt were to exceed a polynomial of degree four, the worry is that non-significant parameters would be included in the model. On the other hand, if the degree were too low, the calibration belt would not be able to accurately identify deviations from the model. A forward selection procedure is used to build the calibration belt's underlying regression model. The first polynomial fit is one of degree two so that a likelihood ratio test can be performed against a polynomial of degree one. This process is continued up to a polynomial of the fourth degree until the most parsimonious model is found. An ideal calibration belt model would have an intercept of 0 and a slope of 1, with no other terms (e.g., no squared term) leading to the bisector of the axes, as this would correspond to perfect calibration.

One of the greatest advantages of the calibration belt is that one can observe the areas where the model is not well-calibrated. After fitting the calibration belt, a graph can be produced that shows the areas of poor calibration, as seen in Figure 1 below.

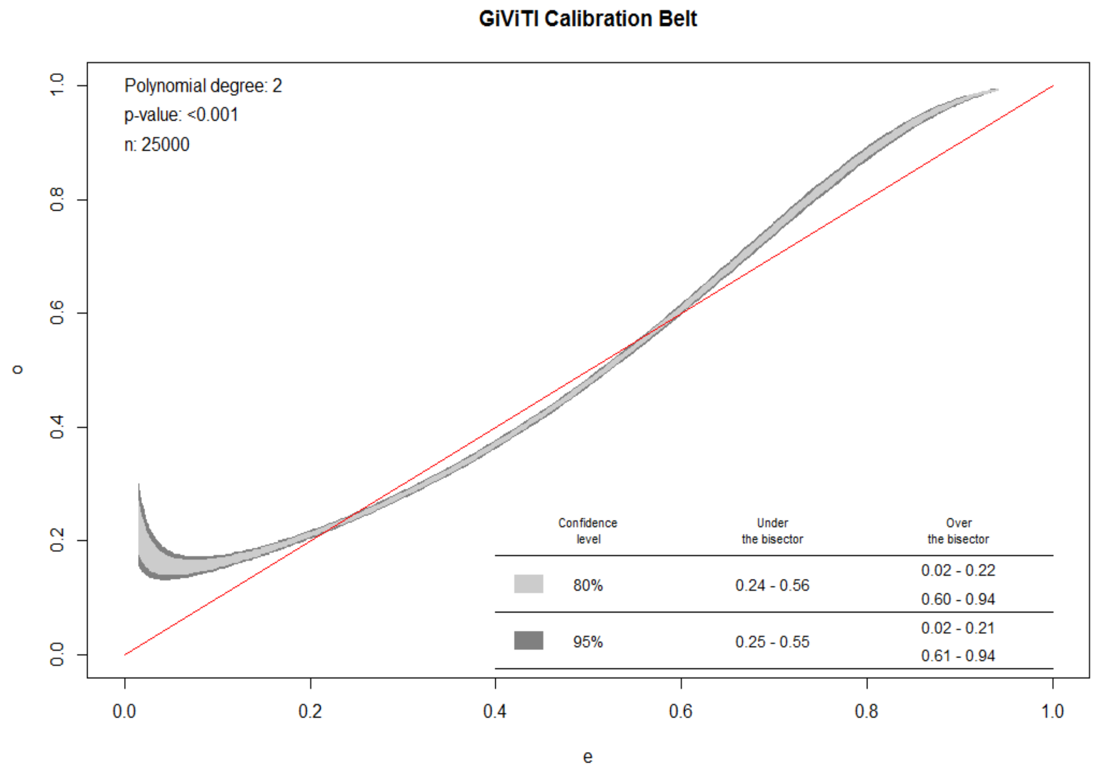


Figure 1. An example of the calibration belt. The sample size is 25,000, the event rate is 0.40, and the calibration belt is fit to model 5.

The advantages of being able to graphically observe where the model fits poorly are numerous. Nattino e*t al.* (2014) give the example of picking a transformation for a continuous variable based off of where the model appears to be poorly calibrated. Additionally, with large sample sizes, the calibration belt will likely reject the null hypothesis that the model is well-calibrated, but one can judge how poorly calibrated the model truly is based on the graph.

Nattino *et al.* (2016) compared the type 1 error rate of the calibration belt with the Hosmer-Lemeshow test in a simulation study with probabilities of success of 0.10, 0.25, and 0.50 with 5, 10 and 50 covariates. In each model, they used 10 as the number of groups for the Hosmer-Lemeshow test. They found that the type 1 error rates for both the calibration belt and the Hosmer-Lemeshow test were generally similar, however, in cases with a rare event the Hosmer-Lemeshow test was more liberal (i.e., had increased type 1 error rates) than the calibration belt (Nattino 2014).

## Chapter 4: Methods

A simulation study was performed using models used in the Paul *et al.* (2013) paper, with the goal of comparing the Hosmer-Lemeshow test with differing group sizes against the calibration belt with respect to both power and the type 1 error rate. The goal was to see if the results reported in Nattino *et al.* (2016) and Paul *et al.* (2013) could be recreated and even expanded upon. Nattino *et al.* (2016) only observed the type 1 error rate between the calibration belt and the Hosmer-Lemeshow test, while Paul *et al.* (2013) compared the power of the Hosmer-Lemeshow test with various number of groups. The methods used in this paper are a synthesis and expansion of these two papers, so that the type 1 error rate and the statistical power could be compared across the calibration belt and the Hosmer-Lemeshow test with and without the adaptive group sizes method proposed by Paul *et al.* (2013).

As in the Paul *et al.* (2013) paper, the model below was used to simulate the binary outcome:

$$logit(P(Y=1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 Z + \beta_5 (Z * X_1)$$

In this scenario, $X_1$ and $X_2$ are standard normal variables, Z follows a binomial distribution with n=1 and a success probability of 0.50, and all three variables are independent of each other. Values of the coefficients $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ were set to the values used in the Paul et al. paper (Table 1), while the $\beta_0$ value was changed so that the probability of success varied. This was done to see if the probability of success of the outcome causes changes in any of the goodness of fit tests. With the only difference between models 2 and 3 in the Paul et al. paper being the value of the intercept, and since we varied the intercept values, these two models were the same for our simulation. Thus results are labeled as “Model 2/3”. The probabilities of successes chosen were 0.05, 0.20, 0.40, 0.60, and 0.80. This created a total of 25 scenarios (five model structures times five probabilities of success). For each model at each different probability of success, data were generated with sample sizes ranging from 100 to 25,000.

To assess power, the following (incorrect) model was then fit to the data,

$$logit(\hat{P}(Y = 1)) = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

which differed from each of the six parent models in various ways. The calibration belt and the Hosmer-Lemeshow test with varying numbers of groups (6 to 130) were then performed; for sample sizes of 25,000, the Hosmer-Lemeshow test with number of groups chosen using the adaptive group selection method proposed by Paul *et al.* (2013) was also conducted. A total of 5,000 replicates were made for each of the five models listed below, and empirical power was estimated as the percentage of replicates where the test rejected the null hypothesis of good model fit.

To assess type 1 error, the same simulation design was used, but the corresponding model-generating equation was fit to the data. Again, both the Hosmer-

Lemeshow test with different numbers of groups and the calibration belt were performed. Empirical type 1 error was estimated as the percentage of replicates where the test rejected the null hypothesis of good fit.

# Chapter 5: Results

With few exceptions, the results show that the calibration belt was more powerful than the Hosmer-Lemeshow test in all models run and at all probability levels and group sizes. This can be seen in Table 2 where the event rate is 0.40. The results, however, are typical of those seen at all event rates.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 5000 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | .0980 | .0800 | .0638 | .0326 | .0018 | N/A | | .1762 |
| | 500 | .3700 | .3058 | .2316 | .1638 | .1094 | .0624 | | .5948 |
| | 1000 | .6706 | .6222 | .5008 | .3690 | .2462 | .1512 | | .8678 |
| | 2000 | .9448 | .9334 | .8776 | .7498 | .5744 | .3766 | | .9924 |
| | 4000 | .9998 | .9988 | .9980 | .9904 | .9542 | .8274 | | 1 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .7958 | 1 |
| 2/3 | 100 | .0438 | .0404 | .0418 | .0284 | .0154 | N/A | | .056 |
| | 500 | .0560 | .0514 | .0524 | .0436 | .0398 | .0328 | | .0686 |
| | 1000 | .0634 | .0600 | .0564 | .0508 | .0514 | .0406 | | .1054 |
| | 2000 | .0860 | .0758 | .0704 | .0618 | .0562 | .0502 | | .156 |
| | 4000 | .1242 | .1050 | .0950 | .0754 | .0640 | .0476 | | .2604 |
| | 25000 | .6376 | .5940 | .4900 | .3756 | .2626 | .1698 | .0270 | .904 |
| 4 | 100 | .0708 | .0736 | .0658 | .0620 | .0368 | N/A | | .1442 |
| | 500 | .2022 | .1924 | .1740 | .1432 | .1218 | .1110 | | .5266 |
| | 1000 | .4024 | .3874 | .3326 | .2706 | .2346 | .1870 | | .8088 |
| | 2000 | .7224 | .7252 | .6700 | .5682 | .4762 | .3636 | | .979 |
| | 4000 | .9674 | .9724 | .9654 | .9188 | .8364 | .6968 | | .9998 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .9732 | 1 |
| 5 | 100 | .0692 | .0672 | .0582 | .0512 | .0150 | N/A | | .135 |
| | 500 | .1758 | .1680 | .1464 | .1212 | .0996 | .0822 | | .449 |
| | 1000 | .3262 | .3292 | .2708 | .2116 | .1692 | .1366 | | .7316 |
| | 2000 | .6218 | .6164 | .5602 | .4542 | .3428 | .2642 | | .9596 |
| | 4000 | .9194 | .9346 | .9108 | .8400 | .7206 | .5530 | | .9996 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .8285 | 1 |
| 6 | 100 | .0664 | .0578 | .0460 | .0282 | .0040 | N/A | | .1606 |
| | 500 | .2140 | .1728 | .1368 | .1042 | .0754 | .0462 | | .5432 |
| | 1000 | .3974 | .3678 | .3094 | .2286 | .1588 | .0912 | | .8342 |
| | 2000 | .7328 | .7166 | .6576 | .5334 | .3868 | .2398 | | .9816 |
| | 4000 | .9694 | .9776 | .9624 | .9066 | .7996 | .6000 | | 1 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .37 | 1 |

Table 2. Empirical power for the five data generation models with an event success rate of 0.40 and varying sample sizes.

In regards to the adaptive group number equation proposed by Paul *et al.* (2013), the probability of success appeared to have a large impact on the power. Although the models from the original paper were used, the power did not standardize well as the probability of success changed.  For example, as can be seen in Figure 2 below, the power using the adaptive group number equation changed dramatically as the probability of success within each model changed as well.

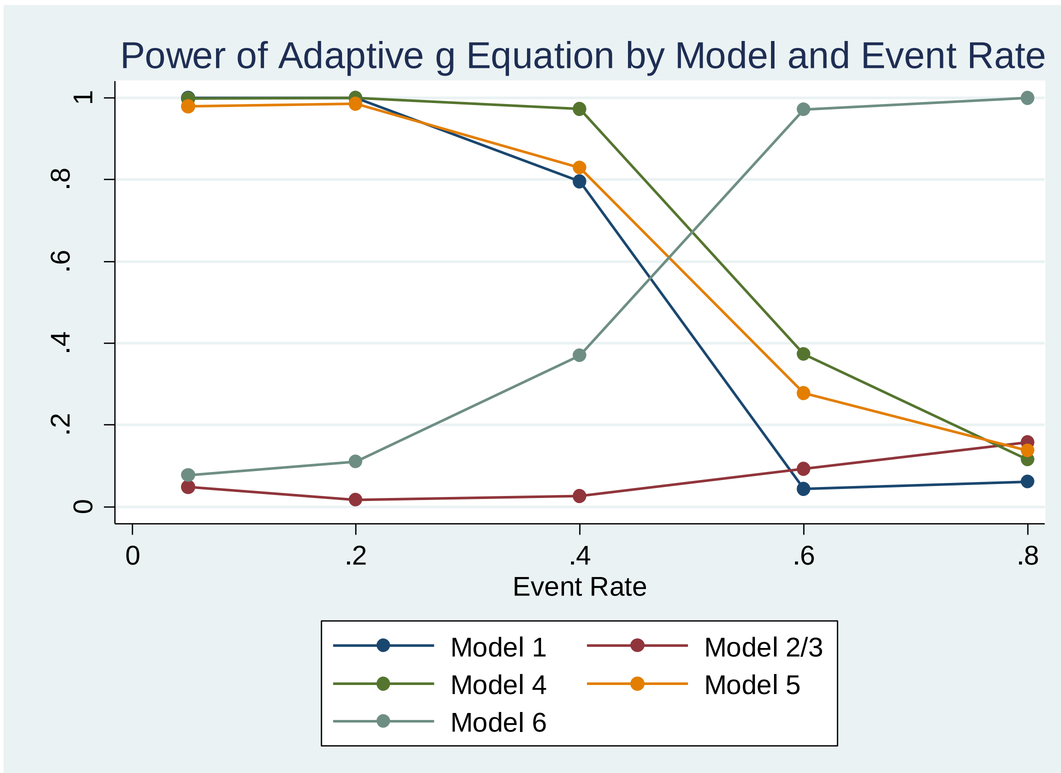


Figure 2.  Statistical power by event rate for each of the five models.  The sample size is 25,000, and the number of groups recommended by the adaptive group equation varies by event rate.

The Hosmer-Lemeshow test and the calibration belt were both very conservative, with the calibration belt's type 1 error being substantially lower than that of the Hosmer-Lemeshow test.  Only in models 2/3 and 5 did the type 1 error rate of the calibration belt ever appear to be close to the ideal type 1 error rate of 5%.  In all other cases, the type 1 error rate was often an order of magnitude or more too conservative, as can be seen in

Table 3 below. Additionally, it is clear from Figure 3 that under the null hypothesis, p-values from the calibration belt do not follow a uniform distribution as would be expected in models 1, 4, and 6, while they did for the Hosmer-Lemeshow test.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 2500 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | .0302 | .0310 | .0364 | .0352 | .0370 | .0398 | | .003 |
| | 1000 | .0280 | .0300 | .0344 | .0406 | .0414 | .0374 | | .0018 |
| | 2000 | .0300 | .0282 | .0318 | .0352 | .0360 | .0396 | | .0006 |
| | 4000 | .0262 | .0262 | .0366 | .0384 | .0378 | .0368 | | .0012 |
| | 25000 | .0296 | .0322 | .0346 | .0396 | .0446 | .0468 | .0366 | .0008 |
| 2/3 | 500 | .0396 | .0420 | .0426 | .0484 | .0558 | .0614 | | .0216 |
| | 1000 | .0392 | .0426 | .0432 | .0442 | .0454 | .0562 | | .0268 |
| | 2000 | .0342 | .0440 | .0412 | .0454 | .0514 | .0532 | | .0264 |
| | 4000 | .0426 | .0386 | .0444 | .0456 | .0518 | .0572 | | .0248 |
| | 25000 | .0384 | .0458 | .0418 | .0436 | .0486 | .0476 | .0676 | .0278 |
| 4 | 500 | .0346 | .0370 | .0392 | .0410 | .0426 | .0280 | | .0038 |
| | 1000 | .0268 | .0294 | .0322 | .0364 | .0392 | .0340 | | .0022 |
| | 2000 | .0324 | .0346 | .0380 | .0408 | .0434 | .0440 | | .0032 |
| | 4000 | .0272 | .0316 | .0330 | .0370 | .0372 | .0412 | | .0014 |
| | 25000 | .0306 | .0328 | .0394 | .0378 | .0426 | .0442 | .0324 | .0012 |
| 5 | 500 | .0448 | .0482 | .0478 | .0514 | .0578 | .0724 | | .0332 |
| | 1000 | .0468 | .0434 | .0470 | .0490 | .0554 | .0632 | | .032 |
| | 2000 | .0476 | .0416 | .0514 | .0528 | .0526 | .0598 | | .0316 |
| | 4000 | .0502 | .0452 | .0488 | .0474 | .0524 | .0536 | | .0352 |
| | 25000 | .0454 | .0424 | .0494 | .0424 | .0458 | .0520 | .0716 | .0328 |
| 6 | 500 | .0292 | .0366 | .0372 | .0404 | .0454 | .0488 | | .0084 |
| | 1000 | .0292 | .0326 | .0368 | .0410 | .0428 | .0428 | | .0028 |
| | 2000 | .0304 | .0314 | .0368 | .0392 | .0382 | .0458 | | .0044 |
| | 4000 | .0314 | .0352 | .0408 | .0384 | .0372 | .0450 | | .0022 |
| | 25000 | .0312 | .0366 | .0354 | .0398 | .0410 | .0454 | .0648 | .0036 |

Table 3. Type 1 error rates for each of the five data generation models where the event rate is 0.20 with varying sample sizes.

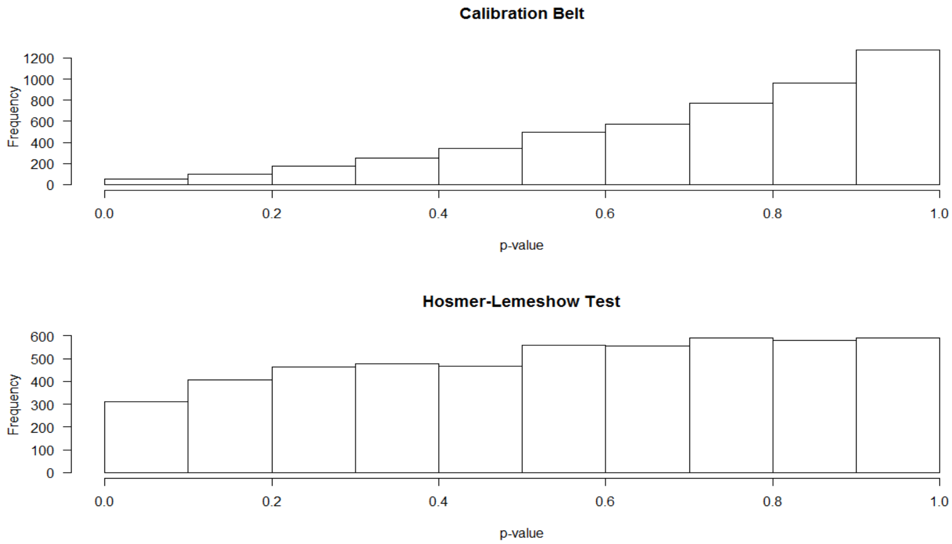


Figure 3. Histograms of the p-values for the calibration belt and Hosmer-Lemeshow test for model 4 with an event probability of 0.20 and a sample size of 1,000 after N=5,000 replications.

## Chapter 6: Discussion

It appears that none of the goodness of fit measures tested in this paper are perfect, yet all seem to be useful. The Hosmer-Lemeshow test, a staple of measuring goodness of fit in logistic regression, is still an extremely effective test. Although it is not as powerful as the calibration belt, it remains a useful technique for evaluating the fit of a logistic regression model. Paul *et al.* (2013) proposed an adaptive equation for the group sizes, but this equation appears to fall short, as it is dependent on the success probability of the model. Perhaps an ideal solution would be to take multiple group sizes for a large sample, and then to judge whether one believes the model fits well based on the results. These multiple group sizes would ideally span from the default size of ten, up to what is recommended by Paul *et al*.'s (2013) adaptive equation.

One could also use the calibration belt, which was found to be more powerful than the Hosmer-Lemeshow test in nearly all cases. Additionally, the ability to observe where the model fits imperfectly with a graph is a large boon for this test. Although one is capable of seeing at what expected probability levels the Hosmer-Lemeshow test also imperfectly fits, the results must first be collapsed into groups. With the calibration belt proposed by Nattino e*t al.* (2014), collapsing of the groups is no longer a necessary step. A problem, however, occurred when trying to recreate the type 1 error levels that were nominally stated for the test. In the Nattino *et al.* (2015) paper, it was found that the type 1 error rate was close to the ideal level of 5% in both the calibration belt and the Hosmer-

Lemeshow test, with the Hosmer-Lemeshow test appearing to be slightly liberal with its type 1 error rate when the marginal success probability was low. The results of this simulation study show that both the Hosmer-Lemeshow test and the calibration belt are generally conservative, but the calibration belt is often an order of magnitude or more too conservative. It is unknown why this would be the case.

Based on the results of this simulation study, ideally any logistic regression model fit to data would be checked with both the Hosmer-Lemeshow test with several group sizes and the calibration belt. These results would then be further analyzed to see if there is any apparent issue with the model. With large sample sizes, the calibration belt may be the best pick, as one can clearly see at what probabilities the model deviates. Unfortunately, it is likely that the calibration belt and the Hosmer-Lemeshow test used with large sample sizes will reject the null hypothesis of a well-calibrated model, but one can observe where this lack of fit occurs better with the calibration belt. One could similarly plot the observed and expected probabilities produced by the Hosmer-Lemeshow test to look for deviance. The outcome would be similar to the calibration belt, but not quite as smooth. With the models fit in this simulation study, it is clear that both the calibration belt and the Hosmer-Lemeshow test are useful for assessing the calibration of a logistic regression model.

## Future Work

As was done in the Nattino *et al.* (2016) paper, more covariates could be tested when fitting the models. It is possible that the models produced by the calibration belt in this paper are over-fitting the data. This could be explored further in future analyses.

References

[1] Paul, Prabasaj, Michael L. Pennell, and Stanley Lemeshow. "Standardizing The Power Of The Hosmer-Lemeshow Goodness Of Fit Test In Large Data Sets". *Statist. Med.* 32.1 (2013): 67-80.

[2] Nattino, Giovanni, Stefano Finazzi, and Guido Bertolini. "A New Test And Graphical Tool To Assess The Goodness Of Fit Of Logistic Regression Models". *Statist. Med.* 35.5 (2016): 709-720.

[3] Nattino, Giovanni, Stefano Finazzi, and Guido Bertolini. "A New Calibration Test And A Reappraisal Of The Calibration Belt For The Assessment Of Prediction Models Based On Dichotomous Outcomes". *Statist. Med.* 33.14 (2014): 2390-2407.

[4] Hosmer, David W, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*. (2013). Print.

[5] Witt, Brandi, et al. "Cardiac Rehabilitation After Myocardial Infarction In The Community". *J Am Coll Cardiol.* 44.5 (2004): 996-998.

Appendix A: Additional Simulation Results

| Model | Sampl e Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 625 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | N/A | N/A | N/A | N/A | N/A | N/A | | * |
| | 500 | .5294 | .5704 | .5520 | .5230 | .4750 | .4474 | | .8516 |
| | 1000 | .8658 | .8906 | .8878 | .8562 | .7994 | .7364 | | .9844 |
| | 2000 | .9956 | .9990 | .9980 | .9950 | .9874 | .9692 | | 1 |
| | 4000 | 1 | 1 | 1 | 1 | 1 | 1 | | 1 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2/3 | 100 | N/A | N/A | N/A | N/A | N/A | N/A | | * |
| | 500 | .0492 | .0512 | .0586 | .0704 | .0914 | .1140 | | .07 |
| | 1000 | .0412 | .0462 | .0506 | .0594 | .0760 | .1034 | | .07 |
| | 2000 | .0502 | .0488 | .0432 | .0540 | .0630 | .0804 | | .079 |
| | 4000 | .0468 | .0414 | .0430 | .0510 | .0496 | .0666 | | .1116 |
| | 25000 | .0684 | .0672 | .0626 | .0662 | .0592 | .0504 | .0478 | .3244 |
| 4 | 100 | N/A | N/A | N/A | N/A | N/A | N/A | | * |
| | 500 | .1538 | .1822 | .2162 | .2704 | .3308 | .3958 | | .1706 |
| | 1000 | .1948 | .2154 | .2668 | .3306 | .4086 | .4902 | | .3006 |
| | 2000 | .2922 | .3244 | .3518 | .4204 | .4994 | .5966 | | .5556 |
| | 4000 | .4952 | .5154 | .5380 | .5648 | .6290 | .7238 | | .8426 |
| | 25000 | .9988 | .9984 | .9988 | .9984 | .9960 | .9952 | .9988 | 1 |
| 5 | 100 | N/A | N/A | N/A | N/A | N/A | N/A | | * |
| | 500 | .1128 | .1288 | .1630 | .1968 | .2538 | .3156 | | .1312 |
| | 1000 | .1278 | .1558 | .1860 | .2418 | .2958 | .3804 | | .2212 |
| | 2000 | .2218 | .2446 | .2644 | .3012 | .3752 | .4632 | | .4176 |
| | 4000 | .3606 | .3676 | .3836 | .3992 | .4496 | .5296 | | .6848 |
| | 25000 | .9854 | .9856 | .9836 | .9762 | .9632 | .9460 | .9800 | 1 |
| 6 | 100 | N/A | N/A | N/A | N/A | N/A | N/A | | * |
| | 500 | .0614 | .0570 | .0426 | .0360 | .0294 | .0244 | | .222 |
| | 1000 | .0894 | .0834 | .0588 | .0434 | .0360 | .0218 | | .3594 |
| | 2000 | .1616 | .1388 | .1062 | .0698 | .0446 | .0264 | | .5966 |
| | 4000 | .3324 | .2982 | .2206 | .1364 | .0818 | .0400 | | .8718 |
| | 25000 | .9980 | .9994 | .9976 | .9878 | .9398 | .7572 | .0764 | 1 |

Table 4. Incorrect models with a success rate of 0.05.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 2500 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | .1456 | .1386 | .0968 | .0652 | .0238 | N/A | | .35 |
| | 500 | .7486 | .736 | .6588 | .5208 | .362 | .2048 | | .9388 |
| | 1000 | .973 | .9738 | .9552 | .8986 | .7742 | .5674 | | .9982 |
| | 2000 | 1 | 1 | .9994 | .9988 | .9936 | .9572 | | 1 |
| | 4000 | 1 | 1 | 1 | 1 | 1 | 1 | | 1 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2/3 | 100 | .0456 | .0414 | .044 | .046 | .057 | N/A | | .0694 |
| | 500 | .052 | .0494 | .0358 | .0428 | .0458 | .0374 | | .0796 |
| | 1000 | .0578 | .0498 | .0482 | .0422 | .0444 | .039 | | .107 |
| | 2000 | .068 | .0584 | .0574 | .049 | .0422 | .0354 | | .1606 |
| | 4000 | .0972 | .0898 | .0748 | .0662 | .0542 | .0460 | | .271 |
| | 25000 | .4496 | .4722 | .4362 | .3464 | .2376 | .1560 | .0168 | .9048 |
| 4 | 100 | .0730 | .0884 | .0956 | .1240 | .1550 | N/A | | .111 |
| | 500 | .1826 | .1912 | .1862 | .2028 | .2308 | .2752 | | .3926 |
| | 1000 | .2968 | .3146 | .3022 | .2858 | .3004 | .3324 | | .6446 |
| | 2000 | .5708 | .5888 | .5524 | .5066 | .4792 | .4734 | | .9158 |
| | 4000 | .8738 | .8886 | .8682 | .8106 | .7488 | .6978 | | .9976 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5 | 100 | .0662 | .0694 | .0742 | .0826 | .1062 | N/A | | .1004 |
| | 500 | .1328 | .1328 | .1408 | .1386 | .1560 | .1722 | | .301 |
| | 1000 | .2166 | .2188 | .2056 | .1912 | .2002 | .2000 | | .5312 |
| | 2000 | .4286 | .4240 | .3962 | .3556 | .3288 | .3010 | | .8222 |
| | 4000 | .7318 | .7514 | .7150 | .6410 | .5632 | .4980 | | .9862 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .9852 | 1 |
| 6 | 100 | .0562 | .0468 | .0402 | .0246 | .0130 | N/A | | .1526 |
| | 500 | .1382 | .1186 | .0900 | .0638 | .0402 | .0204 | | .4614 |
| | 1000 | .2682 | .2404 | .1838 | .1174 | .0728 | .0338 | | .7446 |
| | 2000 | .5538 | .5106 | .4210 | .3010 | .1800 | .0902 | | .9578 |
| | 4000 | .8734 | .8784 | .8210 | .7020 | .4976 | .2860 | | .9984 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .1106 | 1 |

Table 5. Incorrect models with a success rate of 0.20.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 5000 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | .0980 | .0800 | .0638 | .0326 | .0018 | N/A | | .1762 |
| | 500 | .3700 | .3058 | .2316 | .1638 | .1094 | .0624 | | .5948 |
| | 1000 | .6706 | .6222 | .5008 | .3690 | .2462 | .1512 | | .8678 |
| | 2000 | .9448 | .9334 | .8776 | .7498 | .5744 | .3766 | | .9924 |
| | 4000 | .9998 | .9988 | .9980 | .9904 | .9542 | .8274 | | 1 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .7958 | 1 |
| 2/3 | 100 | .0438 | .0404 | .0418 | .0284 | .0154 | N/A | | .056 |
| | 500 | .0560 | .0514 | .0524 | .0436 | .0398 | .0328 | | .0686 |
| | 1000 | .0634 | .0600 | .0564 | .0508 | .0514 | .0406 | | .1054 |
| | 2000 | .0860 | .0758 | .0704 | .0618 | .0562 | .0502 | | .156 |
| | 4000 | .1242 | .1050 | .0950 | .0754 | .0640 | .0476 | | .2604 |
| | 25000 | .6376 | .5940 | .4900 | .3756 | .2626 | .1698 | .0270 | .904 |
| 4 | 100 | .0708 | .0736 | .0658 | .0620 | .0368 | N/A | | .1442 |
| | 500 | .2022 | .1924 | .1740 | .1432 | .1218 | .1110 | | .5266 |
| | 1000 | .4024 | .3874 | .3326 | .2706 | .2346 | .1870 | | .8088 |
| | 2000 | .7224 | .7252 | .6700 | .5682 | .4762 | .3636 | | .979 |
| | 4000 | .9674 | .9724 | .9654 | .9188 | .8364 | .6968 | | .9998 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .9732 | 1 |
| 5 | 100 | .0692 | .0672 | .0582 | .0512 | .0150 | N/A | | .135 |
| | 500 | .1758 | .1680 | .1464 | .1212 | .0996 | .0822 | | .449 |
| | 1000 | .3262 | .3292 | .2708 | .2116 | .1692 | .1366 | | .7316 |
| | 2000 | .6218 | .6164 | .5602 | .4542 | .3428 | .2642 | | .9596 |
| | 4000 | .9194 | .9346 | .9108 | .8400 | .7206 | .5530 | | .9996 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .8285 | 1 |
| 6 | 100 | .0664 | .0578 | .0460 | .0282 | .0040 | N/A | | .1606 |
| | 500 | .2140 | .1728 | .1368 | .1042 | .0754 | .0462 | | .5432 |
| | 1000 | .3974 | .3678 | .3094 | .2286 | .1588 | .0912 | | .8342 |
| | 2000 | .7328 | .7166 | .6576 | .5334 | .3868 | .2398 | | .9816 |
| | 4000 | .9694 | .9776 | .9624 | .9066 | .7996 | .6000 | | 1 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .37 | 1 |

Table 6. Incorrect models with a success rate of 0.40.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 5000 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | .0598 | .0484 | .0466 | .0286 | .0012 | N/A | | .0798 |
| | 500 | .0820 | .0710 | .0684 | .0580 | .0506 | .0328 | | .11 |
| | 1000 | .1084 | .0922 | .0740 | .0612 | .0574 | .0446 | | .183 |
| | 2000 | .1680 | .1356 | .1020 | .0902 | .0748 | .0586 | | .2908 |
| | 4000 | .3112 | .2480 | .1892 | .1358 | .1142 | .0884 | | .5294 |
| | 25000 | .9852 | .9794 | .9478 | .8566 | .7078 | .5168 | .0440 | .9976 |
| 2/3 | 100 | .0518 | .0456 | .0420 | .0364 | .0246 | N/A | | .0588 |
| | 500 | .0576 | .0570 | .0514 | .0516 | .0506 | .0450 | | .063 |
| | 1000 | .0636 | .0614 | .0642 | .0572 | .0570 | .0456 | | .0958 |
| | 2000 | .0772 | .0698 | .0598 | .0628 | .0620 | .0532 | | .1246 |
| | 4000 | .1024 | .0922 | .0830 | .0686 | .0672 | .0628 | | .1764 |
| | 25000 | .4362 | .3882 | .3042 | .2258 | .1762 | .1336 | .0920 | .7544 |
| 4 | 100 | .0676 | .0572 | .0422 | .0284 | .0046 | N/A | | .1606 |
| | 500 | .2144 | .1728 | .1380 | .1060 | .0744 | .0464 | | .5432 |
| | 1000 | .3944 | .3692 | .3054 | .2274 | .1588 | .0914 | | .8362 |
| | 2000 | .7334 | .7178 | .6588 | .5292 | .3884 | .2420 | | .9808 |
| | 4000 | .9704 | .9778 | .9634 | .9042 | .7994 | .5974 | | 1 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .3725 | 1 |
| 5 | 100 | .0722 | .0516 | .0444 | .0252 | .0020 | N/A | | .1606 |
| | 500 | .1934 | .1680 | .1334 | .0984 | .0684 | .0360 | | .5126 |
| | 1000 | .3562 | .3268 | .2636 | .2020 | .1400 | .0844 | | .7976 |
| | 2000 | .6890 | .6650 | .5958 | .4618 | .3156 | .2038 | | .9788 |
| | 4000 | .9566 | .9564 | .9344 | .8602 | .7198 | .5216 | | .9998 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .2785 | 1 |
| 6 | 100 | .0742 | .0738 | .0704 | .0622 | .0402 | N/A | | .144 |
| | 500 | .2028 | .1932 | .1734 | .1430 | .1220 | .1106 | | .5264 |
| | 1000 | .4014 | .3860 | .3314 | .2706 | .2348 | .1860 | | .809 |
| | 2000 | .7218 | .7248 | .6694 | .5704 | .4786 | .3638 | | .9782 |
| | 4000 | .9668 | .9722 | .9644 | .9188 | .8360 | .6966 | | 1 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .971 | 1 |

Table 7. Incorrect models with a success rate of 0.60.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 2500 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | .0534 | .0456 | .0348 | .0278 | .0078 | N/A | | .0738 |
| | 500 | .0624 | .0514 | .0516 | .0502 | .0472 | .0370 | | .0644 |
| | 1000 | .0562 | .0558 | .0590 | .0534 | .0444 | .0348 | | .0684 |
| | 2000 | .0872 | .0822 | .0710 | .0658 | .0594 | .0502 | | .0842 |
| | 4000 | .1208 | .1104 | .1026 | .0880 | .0734 | .0690 | | .1348 |
| | 25000 | .6290 | .6348 | .5808 | .4744 | .3424 | .2342 | .0605 | .676 |
| 2/3 | 100 | .0544 | .0490 | .0544 | .0606 | .0766 | N/A | | .0598 |
| | 500 | .0572 | .0570 | .0556 | .0614 | .0680 | .0868 | | .0546 |
| | 1000 | .0568 | .0562 | .0624 | .0620 | .0706 | .0722 | | .0634 |
| | 2000 | .0636 | .0626 | .0618 | .0646 | .0658 | .0738 | | .0764 |
| | 4000 | .0678 | .0648 | .0622 | .0672 | .0650 | .0760 | | .1012 |
| | 25000 | .1922 | .1746 | .1516 | .1256 | .1094 | .0880 | .1580 | .4246 |
| 4 | 100 | .0562 | .0470 | .0362 | .0248 | .0126 | N/A | | .1532 |
| | 500 | .1384 | .1180 | .0890 | .0624 | .0408 | .0202 | | .4616 |
| | 1000 | .2666 | .2412 | .1808 | .1174 | .0736 | .0340 | | .7446 |
| | 2000 | .5544 | .5110 | .4212 | .3010 | .1798 | .0908 | | .9584 |
| | 4000 | .8730 | .8780 | .8196 | .7020 | .4964 | .2856 | | .9984 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .1155 | 1 |
| 5 | 100 | .0622 | .0472 | .0366 | .0248 | .0086 | N/A | | .1552 |
| | 500 | .1406 | .1220 | .0922 | .0752 | .0430 | .0264 | | .442 |
| | 1000 | .2858 | .2422 | .1870 | .1254 | .0744 | .0400 | | .7198 |
| | 2000 | .5612 | .5122 | .3992 | .2906 | .1762 | .0952 | | .95 |
| | 4000 | .8844 | .8794 | .8194 | .6840 | .4822 | .2794 | | .9996 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | .1375 | 1 |
| 6 | 100 | .0708 | .0878 | .1030 | .1236 | .1606 | N/A | | .1106 |
| | 500 | .1824 | .1902 | .1862 | .2022 | .2300 | .2746 | | .3922 |
| | 1000 | .2982 | .3130 | .2984 | .2866 | .3030 | .3324 | | .6448 |
| | 2000 | .5688 | .5902 | .5518 | .5068 | .4778 | .4732 | | .9154 |
| | 4000 | .8740 | .8890 | .8662 | .8114 | .7514 | .6982 | | .9978 |
| | 25000 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Table 8.  Incorrect models with a success rate of 0.80.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 625 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | .0286 | .0312 | .0324 | .0446 | .0572 | .0722 | | .0056 |
| | 1000 | .0288 | .0318 | .0348 | .0412 | .0462 | .0588 | | .0016 |
| | 2000 | .0300 | .0288 | .0334 | .0426 | .0448 | .0562 | | .0016 |
| | 4000 | .0306 | .0362 | .0390 | .0464 | .0460 | .0528 | | .002 |
| | 25000 | .0300 | .0328 | .0346 | .0400 | .0448 | .0426 | .0494 | <.00001 |
| 2/3 | 500 | .0328 | .0408 | .0442 | .0602 | .0748 | .1042 | | .0262 |
| | 1000 | .0364 | .0426 | .0534 | .0652 | .0772 | .1054 | | .0246 |
| | 2000 | .0408 | .0408 | .0450 | .0584 | .0654 | .0852 | | .0192 |
| | 4000 | .0428 | .0424 | .0438 | .0500 | .0564 | .0792 | | .0234 |
| | 25000 | .0458 | .0480 | .0456 | .0432 | .0486 | .0532 | .1920 | .0178 |
| 4 | 500 | .0368 | .0382 | .0374 | .0450 | .0504 | .0614 | | .0088 |
| | 1000 | .0290 | .0340 | .0382 | .0422 | .0450 | .0578 | | .0066 |
| | 2000 | .0354 | .0350 | .0322 | .0416 | .0434 | .0486 | | .0024 |
| | 4000 | .0318 | .0334 | .0312 | .0406 | .0442 | .0436 | | .0018 |
| | 25000 | .0352 | .0400 | .0392 | .0440 | .0460 | .0444 | .0404 | .001 |
| 5 | 500 | .0472 | .0538 | .0674 | .0862 | .1098 | .1468 | | .0264 |
| | 1000 | .0458 | .0504 | .0570 | .0820 | .1034 | .1454 | | .0242 |
| | 2000 | .0494 | .0428 | .0550 | .0660 | .0866 | .1198 | | .0266 |
| | 4000 | .0486 | .0470 | .0512 | .0582 | .0746 | .0964 | | .0282 |
| | 25000 | .0554 | .0508 | .0514 | .0524 | .0612 | .0636 | .2266 | .024 |
| 6 | 500 | .0346 | .0396 | .0352 | .0400 | .0486 | .0564 | | .0168 |
| | 1000 | .0322 | .0372 | .0416 | .0460 | .0572 | .0626 | | .0106 |
| | 2000 | .0368 | .0330 | .0410 | .0420 | .0508 | .0572 | | .0074 |
| | 4000 | .0314 | .0352 | .0348 | .0416 | .0478 | .0536 | | .0066 |
| | 25000 | .0302 | .0376 | .0364 | .0412 | .0416 | .0482 | .1156 | .0028 |

Table 9. Correct models with a success rate of 0.05.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 2500 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | .0302 | .0310 | .0364 | .0352 | .0370 | .0398 | | .003 |
| | 1000 | .0280 | .0300 | .0344 | .0406 | .0414 | .0374 | | .0018 |
| | 2000 | .0300 | .0282 | .0318 | .0352 | .0360 | .0396 | | .0006 |
| | 4000 | .0262 | .0262 | .0366 | .0384 | .0378 | .0368 | | .0012 |
| | 25000 | .0296 | .0322 | .0346 | .0396 | .0446 | .0468 | .0366 | .0008 |
| 2/3 | 500 | .0396 | .0420 | .0426 | .0484 | .0558 | .0614 | | .0216 |
| | 1000 | .0392 | .0426 | .0432 | .0442 | .0454 | .0562 | | .0268 |
| | 2000 | .0342 | .0440 | .0412 | .0454 | .0514 | .0532 | | .0264 |
| | 4000 | .0426 | .0386 | .0444 | .0456 | .0518 | .0572 | | .0248 |
| | 25000 | .0384 | .0458 | .0418 | .0436 | .0486 | .0476 | .0676 | .0278 |
| 4 | 500 | .0346 | .0370 | .0392 | .0410 | .0426 | .0280 | | .0038 |
| | 1000 | .0268 | .0294 | .0322 | .0364 | .0392 | .0340 | | .0022 |
| | 2000 | .0324 | .0346 | .0380 | .0408 | .0434 | .0440 | | .0032 |
| | 4000 | .0272 | .0316 | .0330 | .0370 | .0372 | .0412 | | .0014 |
| | 25000 | .0306 | .0328 | .0394 | .0378 | .0426 | .0442 | .0324 | .0012 |
| 5 | 500 | .0448 | .0482 | .0478 | .0514 | .0578 | .0724 | | .0332 |
| | 1000 | .0468 | .0434 | .0470 | .0490 | .0554 | .0632 | | .032 |
| | 2000 | .0476 | .0416 | .0514 | .0528 | .0526 | .0598 | | .0316 |
| | 4000 | .0502 | .0452 | .0488 | .0474 | .0524 | .0536 | | .0352 |
| | 25000 | .0454 | .0424 | .0494 | .0424 | .0458 | .0520 | .0716 | .0328 |
| 6 | 500 | .0292 | .0366 | .0372 | .0404 | .0454 | .0488 | | .0084 |
| | 1000 | .0292 | .0326 | .0368 | .0410 | .0428 | .0428 | | .0028 |
| | 2000 | .0304 | .0314 | .0368 | .0392 | .0382 | .0458 | | .0044 |
| | 4000 | .0314 | .0352 | .0408 | .0384 | .0372 | .0450 | | .0022 |
| | 25000 | .0312 | .0366 | .0354 | .0398 | .0410 | .0454 | .0648 | .0036 |

Table 10. Correct models with a success rate of 0.20.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 5000 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | .0238 | .0260 | .0342 | .0378 | .0350 | .0284 | | .002 |
| | 1000 | .0288 | .0302 | .0356 | .0360 | .0376 | .0368 | | .0022 |
| | 2000 | .0268 | .0294 | .0354 | .0376 | .0414 | .0398 | | .0024 |
| | 4000 | .0254 | .0266 | .0340 | .0382 | .0432 | .0428 | | .0004 |
| | 25000 | .0232 | .0286 | .0360 | .0318 | .0358 | .0382 | .0366 | .0008 |
| 2/3 | 500 | .0392 | .0402 | .0414 | .0422 | .0402 | .0360 | | .0278 |
| | 1000 | .0420 | .0448 | .0458 | .0444 | .0474 | .0404 | | .0266 |
| | 2000 | .0426 | .0458 | .0428 | .0454 | .0526 | .0492 | | .0316 |
| | 4000 | .0398 | .0438 | .0422 | .0444 | .0410 | .0490 | | .0288 |
| | 25000 | .0396 | .0360 | .0450 | .0480 | .0498 | .0486 | .0386 | .032 |
| 4 | 500 | .0306 | .0336 | .0344 | .0380 | .0336 | .0276 | | .0066 |
| | 1000 | .0304 | .0328 | .0408 | .0418 | .0432 | .0310 | | .0036 |
| | 2000 | .0324 | .0336 | .0368 | .0402 | .0472 | .0428 | | .002 |
| | 4000 | .0266 | .0304 | .0340 | .0374 | .0394 | .0456 | | .0014 |
| | 25000 | .0282 | .0308 | .0372 | .0360 | .0444 | .0462 | .0372 | .001 |
| 5 | 500 | .0400 | .0418 | .0444 | .0460 | .0404 | .0384 | | .0312 |
| | 1000 | .0430 | .0480 | .0432 | .0426 | .0438 | .0432 | | .0386 |
| | 2000 | .0480 | .0460 | .0478 | .0458 | .0460 | .0424 | | .0408 |
| | 4000 | .0464 | .0520 | .0462 | .0476 | .0474 | .0464 | | .0356 |
| | 25000 | .0478 | .0458 | .0520 | .0538 | .0494 | .0524 | .0424 | .039 |
| 6 | 500 | .0328 | .0354 | .0362 | .0380 | .0374 | .0352 | | .0052 |
| | 1000 | .0350 | .0342 | .0364 | .0396 | .0438 | .0354 | | .0038 |
| | 2000 | .0312 | .0346 | .0344 | .0466 | .0396 | .0422 | | .002 |
| | 4000 | .0318 | .0358 | .0366 | .0414 | .0476 | .0460 | | .0024 |
| | 25000 | .0292 | .0316 | .0360 | .0380 | .0462 | .0466 | .0418 | .0014 |

Table 11. Correct models with a success rate of 0.40.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 5000 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | .0246 | .0310 | .0294 | .0342 | .0334 | .0326 | | .0014 |
| | 1000 | .0230 | .0276 | .0308 | .0398 | .0394 | .0380 | | .0024 |
| | 2000 | .0260 | .0288 | .0318 | .0376 | .0364 | .0370 | | .0014 |
| | 4000 | .0262 | .0300 | .0322 | .0398 | .0402 | .0402 | | .0004 |
| | 25000 | .0270 | .0324 | .0374 | .0404 | .0392 | .0434 | .0452 | .0006 |
| 2/3 | 500 | .0392 | .0416 | .0412 | .0428 | .0440 | .0410 | | .029 |
| | 1000 | .0444 | .0490 | .0508 | .0464 | .0436 | .0432 | | .0354 |
| | 2000 | .0388 | .0480 | .0444 | .0452 | .0504 | .0554 | | .0396 |
| | 4000 | .0450 | .0442 | .0488 | .0526 | .0520 | .0484 | | .0358 |
| | 25000 | .0438 | .0436 | .0462 | .0494 | .0436 | .0456 | .0484 | .0366 |
| 4 | 500 | .0322 | .0354 | .0366 | .0376 | .0380 | .0368 | | .005 |
| | 1000 | .0348 | .0338 | .0374 | .0412 | .0444 | .0354 | | .0038 |
| | 2000 | .0304 | .0344 | .0340 | .0474 | .0404 | .0428 | | .0022 |
| | 4000 | .0320 | .0352 | .0360 | .0420 | .0472 | .0452 | | .0024 |
| | 25000 | .0308 | .0316 | .0360 | .0366 | .0464 | .0452 | .0422 | .0012 |
| 5 | 500 | .0476 | .0426 | .0454 | .0504 | .0474 | .0462 | | .0336 |
| | 1000 | .0502 | .0436 | .0468 | .0510 | .0440 | .0488 | | .0354 |
| | 2000 | .0474 | .0442 | .0524 | .0504 | .0534 | .0506 | | .04 |
| | 4000 | .0472 | .0482 | .0460 | .0446 | .0492 | .0452 | | .0428 |
| | 25000 | .0436 | .0482 | .0532 | .0512 | .0548 | .0518 | .0546 | .0368 |
| 6 | 500 | .0304 | .0334 | .0348 | .0392 | .0342 | .0284 | | .0066 |
| | 1000 | .0288 | .0328 | .0384 | .0410 | .0436 | .0316 | | .0038 |
| | 2000 | .0324 | .0338 | .0386 | .0404 | .0460 | .0420 | | .0018 |
| | 4000 | .0268 | .0302 | .0348 | .0364 | .0404 | .0468 | | .0012 |
| | 25000 | .0278 | .0306 | .0354 | .0350 | .0432 | .0440 | .0350 | .001 |

Table 12. Correct models with a success rate of 0.60.

| Model | Sample Size | g=6 | g=10 | g=18 | g=34 | g=66 | g=130 | Adaptive g = 2500 | Calibration Belt |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | .0258 | .0296 | .0334 | .0342 | .0410 | .0488 | | .0038 |
| | 1000 | .0298 | .0296 | .0324 | .0342 | .0404 | .0440 | | .002 |
| | 2000 | .0284 | .0294 | .0308 | .0324 | .0426 | .0478 | | .002 |
| | 4000 | .0342 | .0326 | .0350 | .0326 | .0404 | .0448 | | .0014 |
| | 25000 | .0294 | .0302 | .0324 | .0370 | .0394 | .0432 | .0662 | .002 |
| 2/3 | 500 | .0452 | .0462 | .0462 | .0472 | .0630 | .0714 | | .0356 |
| | 1000 | .0420 | .0464 | .0454 | .0464 | .0616 | .0666 | | .0358 |
| | 2000 | .0444 | .0438 | .0462 | .0446 | .0510 | .0604 | | .039 |
| | 4000 | .0418 | .0460 | .0466 | .0526 | .0558 | .0584 | | .0358 |
| | 25000 | .0470 | .0448 | .0424 | .0468 | .0506 | .0490 | .0906 | .0346 |
| 4 | 500 | .0290 | .0364 | .0370 | .0398 | .0450 | .0480 | | .008 |
| | 1000 | .0276 | .0332 | .0354 | .0420 | .0426 | .0418 | | .0028 |
| | 2000 | .0312 | .0328 | .0372 | .0380 | .0378 | .0440 | | .0044 |
| | 4000 | .0316 | .0354 | .0406 | .0380 | .0376 | .0440 | | .002 |
| | 25000 | .0310 | .0366 | .0348 | .0396 | .0418 | .0446 | .0658 | .0036 |
| 5 | 500 | .0472 | .0540 | .0468 | .0516 | .0614 | .0718 | | .037 |
| | 1000 | .0496 | .0506 | .0514 | .0580 | .0616 | .0746 | | .0388 |
| | 2000 | .0518 | .0468 | .0546 | .0538 | .0602 | .0696 | | .0388 |
| | 4000 | .0436 | .0440 | .0438 | .0458 | .0518 | .0634 | | .0462 |
| | 25000 | .0564 | .0576 | .0528 | .0504 | .0536 | .0540 | .1102 | .042 |
| 6 | 500 | .0350 | .0368 | .0392 | .0414 | .0422 | .0282 | | .0038 |
| | 1000 | .0272 | .0288 | .0310 | .0368 | .0394 | .0336 | | .0022 |
| | 2000 | .0324 | .0342 | .0380 | .0410 | .0432 | .0444 | | .0032 |
| | 4000 | .0272 | .0314 | .0338 | .0360 | .0378 | .0418 | | .0014 |
| | 25000 | .0306 | .0332 | .0394 | .0382 | .0420 | .0436 | .0312 | .0014 |

Table 13. Correct models with a success rate of 0.80.

Appendix B:  R Code

```
require(ResourceSelection)  # package containing the H-L test
require(givitiR) # Package containing Giovanni's test

runsim <- function(NREPS, n, B0, B1, B2, B3, B4, B5, x){
  set.seed(6320489)
  # vectors to hold results of the replicates
  hl.stat <- hl.pval <- matrix(NA,nrow=NREPS,ncol=length(x))
  giovanni.pval <- vector(length=NREPS)
  mean_y <- vector(length=NREPS)

for (i in 1:NREPS)
{
  if(i %% 100 == 0) print(paste("Replicate",i))
  ###########i############
  ###### GENERATE DATA
  ########################
  ## generate covariates (all are independent of each other)
  x1 <- rnorm(n,0,1)
  x2 <- rnorm(n,0,1)
  z <- rbinom(n,1,0.5)
  ## generate binary outcome
  # linear predictor (XB)
```

```
linpred <- B0 + B1*x1 + B2*x1^2 + B3*x2 + B4*z + B5*z*x1

# P(Y=1)

prob <- 1 - 1/(1+exp(linpred))

# draw Y

y <- rbinom(n,1,prob)


########################

###### FIT MODEL

########################

#To test alpha for model 1

#fit <- glm(y ~ x1 + I(x1^2) + z + z*x1, family=binomial)


#To test alpha for model 2/3

#fit <- glm(y ~ x1 + z + z*x1, family=binomial)


#To test alpha for model 4

#fit <- glm(y ~ x1 + I(x1^2), family=binomial)


#To test alpha for model 5

#fit <- glm(y ~ x1 + I(x1^2) + x2, family=binomial)


#To test alpha for model 6
```

```
#fit <- glm(y ~ x1 + I(x1^2), family=binomial)



#To test power
fit <- glm(y ~ x1, family=binomial)  # may not match the data generation model,
depending on the Bs

########################
###### H-L TEST
########################
# perform test
for (j in 1:length(x))
{
  G <- x[j]
  hl <- hoslem.test(fit$y, fitted(fit), g=G)
  # save the test statistic and p-value
  hl.stat[i,j] <- hl$statistic
  hl.pval[i,j] <- hl$p.value
}
ctest <- givitiCalibrationTest(fit$y, fitted(fit), "internal")
giovanni.pval[i] <- ctest$p.value
mean_y[i] <- mean(y)
```

```
}
  par(mfrow=c(2,1))
  hist(giovanni.pval, main="Calibration Belt", xlab = "p-value", las=1)
  hist(hl.pval[,1], main = "Hosmer-Lemeshow Test", xlab = "p-value", las=1)

hl.power <- apply(hl.pval, 2, function(X) sum(X<.05)/NREPS)
Final <-
list(yprob=mean(mean_y),hl=hl.power,gv=sum(giovanni.pval<.05)/NREPS,G=x)
return(Final)
return(fit)
}

##For model 1
#set1.100 <- runsim(5000, 100, 1.024, 1, .2, 0, 1, -2, c(6,10,18,34,66,130))
set1.500 <- runsim(100, 5000, 1.024, 1, .2, 0, 1, -2, c(6,10,18,34,66,130))
set1.1000 <- runsim(5000, 1000, 1.024, 1, .2, 0, 1, -2, c(6,10,18,34,66,130))
set1.2000 <- runsim(5000, 2000, 1.024, 1, .2, 0, 1, -2, c(6,10,18,34,66,130))
set1.4000 <- runsim(5000, 4000, 1.024, 1, .2, 0, 1, -2, c(6,10,18,34,66,130))
set1.25000 <- runsim(2000, 25000, 1.024, 1, .2, 0, 1, -2, c(6,10,18,34,66,130,2500))

#################################
##For model 2/3
```

```
#set2.100 <- runsim(5000, 100, 1.345, 1, 0, 0, 1, .5, c(6,10,18,34,66,130))

set2.500 <- runsim(5000, 500, 1.345, 1, 0, 0, 1, .5, c(6,10,18,34,66,130))

set2.1000 <- runsim(5000, 1000, 1.345, 1, 0, 0, 1, .5, c(6,10,18,34,66,130))

set2.2000 <- runsim(5000, 2000,  1.345, 1, 0, 0, 1, .5, c(6,10,18,34,66,130))

set2.4000 <- runsim(5000, 4000, 1.345, 1, 0, 0, 1, .5, c(6,10,18,34,66,130))

set2.25000 <- runsim(2000, 25000, 1.345, 1, 0, 0, 1, .5, c(6,10,18,34,66,130,2500))


#################################

##For model 4

#set4.100 <- runsim(5000, 100, 1.462, 1, .2, 0, 0, 0, c(6,10,18,34,66,130))

set4.500 <- runsim(5000, 500, 1.462, 1, .2, 0, 0, 0, c(6,10,18,34,66,130))

set4.1000 <- runsim(5000, 1000, -1.843, 1, .2, 0, 0, 0, c(6,10,18,34,66,130))

set4.2000 <- runsim(5000, 2000, 1.462, 1, .2, 0, 0, 0, c(6,10,18,34,66,130))

set4.4000 <- runsim(5000, 4000, 1.462, 1, .2, 0, 0, 0, c(6,10,18,34,66,130))

set4.25000 <- runsim(2000, 25000, 1.462, 1, .2, 0, 0, 0, c(6,10,18,34,66,130,2500))


#################################

##For model  5

#set5.100 <- runsim(5000, 100, 1.678, 1, .2, 1, 0, 0, c(6,10,18,34,66,130))

set5.500 <- runsim(5000, 500, 1.678, 1, .2, 1, 0, 0, c(6,10,18,34,66,130))

set5.1000 <- runsim(5000, 1000, 1.678, 1, .2, 1, 0, 0, c(6,10,18,34,66,130))

set5.2000 <- runsim(5000, 2000, 1.678, 1, .2, 1, 0, 0, c(6,10,18,34,66,130))
```

```
set5.4000 <- runsim(5000, 4000, 1.678, 1, .2, 1, 0, 0, c(6,10,18,34,66,130))

set5.25000 <- runsim(2000, 25000, 1.678, 1, .2, 1, 0, 0, c(6,10,18,34,66,130,2500))


#################################

##For model  6

#set6.100 <- runsim(5000, 100, 1.844, -1, -.2, 0, 0, 0, c(6,10,18,34,66,130))

set6.500 <- runsim(50, 500, 1.844, -1, -.2, 0, 0, 0, c(6,10,18,34,66,130))

set6.1000 <- runsim(5000, 1000, 1.844, -1, -.2, 0, 0, 0, c(6,10,18,34,66,130))

set6.2000 <- runsim(5000, 2000, 1.844, -1, -.2, 0, 0, 0, c(6,10,18,34,66,130))

set6.4000 <- runsim(5000, 4000, 1.844, -1, -.2, 0, 0, 0, c(6,10,18,34,66,130))

set6.25000 <- runsim(2000, 25000, 1.844, -1, -.2, 0, 0, 0, c(6,10,18,34,66,130,2500))









mean(mean_y)

sum(hl.pval<.05)/NREPS

sum(giovanni.pval<.05)/NREPS


#Creating the calibration belt
```

```
cb <- givitiCalibrationBelt(fit$y, fitted(fit), "internal")

plotGivitiCalibrationBelt(cb)
```