Academic Journal of Applied Mathematical Sciences
ISSN(e): 2415-2188, ISSN(p): 2415-5225
Vol. 7, Issue. 1, pp: 50-59, 2021
URL: https://arpgweb.com/journal/journal/17
DOI: https://doi.org/10.32861/ajams.71.50.59

Academic Research Publishing Group

Original Research **Open Access**

A Comparison Study of Goodness of Fit Tests of Logistic Regression in R: Simulation and Application to Breast Cancer Data

El-Housainy A. Rady
Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

Mohamed R. Abonazel (Corresponding Author)
Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt
Email: mabonazel@cu.edu.eg

Mariam H. Metawe'e
Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

Article History
Received: 26 October, 2020
Revised: 21 November, 2020
Accepted: 28 November, 2020
Published: 1 December, 2020

Abstract

Goodness of fit (GOF) tests of logistic regression attempt to find out the suitability of the model to the data. The null hypothesis of all GOF tests is the model fit. R as a free software package has many GOF tests in different packages. A Monte Carlo simulation has been conducted to study two situations; the first, studying the ability of each test, under its default settings, to accept the null hypothesis when the model truly fitted. The second, studying the power of these tests when assumptions of sufficient linear combination of the explanatory variables are violated (by omitting linear covariate term, quadratic term, or interaction term). Moreover, checking whether the same test in different R packages had the same results or not. As the sample size supposed to affect simulation results, so the pattern of change of GOF tests results under different sample sizes as well as different model settings was estimated. All tests accept the null hypothesis (more than 95% of simulation trials) when the model truly fitted except modified Hosmer-Lemeshow test in "LogisticDx" package under all different model settings and Osius and Rojek's (OsRo) test when the true model had an interaction term between binary and categorical covariates. In addition, le Cessie-van Houwelingen-Copas-Hosmer unweighted sum of squares (CHCH) test gave unexpected different results under different packages. Concerning the power study, all tests had a very low power when a departure of missing covariate existed. Generally, stukel's test (package 'LogisticDX) and CHCH test (package "RMS") reached a power in detecting a missing quadratic term greater than 80% under lower sample size while OsRo test (package 'LogisticDX') was better in detecting missing interaction term. Beside the simulation study, we evaluated the performance of GOF tests using the breast cancer dataset.

Keywords: Binary logistic regression model; Hosmer-lemeshow test; Misspecification; Power of goodness of fit tests; Pseudo R squared; R packages.

1. Introduction

Many scientific branches (other than statistics) use logistic regression (LR) modeling to get information or to prove or reject specific theory or hypothesis. The important component of any modeling process is an assessment of goodness of fit (GOF) of the model. It reflects whether the model fits the observed data accurately or not [1]. Non statistician scientists use the default of a statistical program to assess GOF. R is one of those programs that its use increased dramatically in previous years especially that it is free and nearly each day statistical tests coded into functions easily and these functions used by non-statistician users. Many GOF tests were already coded in many R packages and sometimes the same test coded in different packages. However choosing which test of which package will assess the fit of a specific data is not a piece of cake. It is supposed that same tests under different R functions should give the same result but there is no scientific work (to our knowledge) checked that. Moreover, some tests didn't reach a satisfied power except when a sample exceeds a specific size. On the other hand, some tests fail to detect the truly specified model. In addition, not all tests can detect all misspecifications. Furthermore, a test can detect a specific misspecification under some settings and fails to detect the same misspecification under different settings.

This paper represented the GOF tests that already coded in R and showed their packages. The function of each test was kept under its default settings as possible because non-statistician users most probably use a function of the test under its default setting. This study also compared the power of same tests (from different packages) and examines whether they had the same simulation results under different settings as expected or there was a variation. Adding to that, The pattern of each test under ten different sample sizes (100, 200, … , 1000) was examined and compared with each other by a simulation (1000 simulation trial) study from two aspects; first is the ability (percentage) of the test to accept the null hypotheses when the true model is fitted. For simplifying, this concept will be abbreviated as "The Null". The second aspect is the ability (percentage) of the test to reject the null hypotheses

when a false model is fitted i.e., "The Power". Both aspects were studied under three Misspecifications; A- Omission of a covariate term, B- Omission of quadratic term, C- Omission of interaction term.

After that, the best test for each misspecification was chosen and finally an application on breast cancer data took place. For this work, the best test means the test which has a power greater than 80% and a null greater than 95% at a lower sample size.

This paper was organized as follows: section (2) represents an overview of a binary LR model, whereas section (3) reviews its GOF tests, settings chosen for the simulation study is at section (4), while the results of it will clarified in section (5), and finally applied statistical analysis and a conclusion will be under section (6) and (7) respectively.

2. Binary LR model

In any regression problem, the key quantity is the mean value of the outcome variable, given the value of the independent variable. This quantity is called the conditional mean and is expressed as "E(Y|x)" where Y denotes the outcome variable and x denotes a specific value of the independent variable [2].

In order to simplify notation, we use the quantity $\pi(x) = E(Y|x)$ to represent the conditional mean of Y given x when the logistic distribution is used. The specific form of the LR model used is:

$$\pi(x) = \frac{e^{X\beta}}{1+e^{X\beta}},$$

where

$$X\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

A transformation of $\pi(x)$ that is central to our study of LR is the logit transformation. This transformation is defined, in terms of $\pi(x)$:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = X\beta.$$

The importance of this transformation is that g(x) has many of the desirable properties of a linear regression model. The logit, g(x), is linear in its parameters, may be continuous, and may range from $-\infty$ to $+\infty$, depending on the range of x [2-4].

3. GOF Tests

Knowing whether the probabilities produced by the model accurately reflect the true outcome experience in the data is referred to as model goodness of fit. The null hypothesis of all GOF tests is that the model truly specified whereas when the alternative hypothesis can't be rejected (i.e. p-value of the test is less than 0.05) indicates that the model is misspecified.

In the LR context, the essential components of fit are listed by the following three assumptions [2]:

A1: If the logit link function is appropriate:$g(x) = X\beta$.

A2: If the linear combination $X\beta$ of the explanatory variables are sufficient; No omission of predictions, transformation of predictors, or interactions of predictors, so the $X\beta$ is sufficient.

A3: If the underlying distribution for the outcome variable is Bernoulli; the variance is Bernoulli: $Var\,(Y_i|x_i) = \pi_i(1-\pi_i)$.

The consequences of violation of all the above assumptions (A1 to A3) are serious; the estimated coefficients and corresponding standard errors will be biased thus significance tests and confidence interval may be misleading. Besides, it could also lead to poor estimation of other quantities. For example, the inaccurate odds ratio estimation will influence the interpretation of the treatment effect. If the model is misspecified, it could reduce the accuracy of the prediction and classification for the new subjects. The interpretation of the relationship between the response variable and independent variables could also be inaccurate.

In this paper, six GOF tests coded in R (see Table 1) were chosen to detect the violation of the second assumption (insufficiency of $X\beta$).

It is worth noting that Stukel's tests in "LogisticDx" package referred to six different tests:

- Score test for addition of vector z1 (S1)
- Score test for addition of vector z2 (S2)
- Score test for addition of vector z1 and z2 (S3)
- Log-likelihood test for addition of vector z1 (S4)
- Log-likelihood test for addition of vector z2 (S5)
- Log-likelihood test for addition of vectors z1 and z2(S6)

Many other studies simulate GOF tests to check this violation. Hosmer-Lemeshow tests [5, 6] were nearly tested in all simulation studies against most of departures as they considered the base line of goodness of fit tests as well as they are the default of most software packages [1, 7-10].

The power of Osius and Rojek X^2_{OR} [11] where examined when detecting missing covariate as well as wrong functional form of the covariate under different dispersion levels of data [12].

Unweighted sum of squares test [13], Stukel's score test [14] as well as a likelihood ratio test of Stukel's suggested by [15] were tested against omission of a quadratic term in a continuous variable, and the omission of the main effect for a dichotomous variable and its interaction with a continuous variable [9] and omission of log term [15].

It deserves mentioning that all the previous simulation studies chose two or maximum three sample sizes (most probably 100, 500 and/or 1000) and none of simulation studies - to our knowledge – studied a detailed pattern of tests power or the optimum point at which the power of the test can exceed 80%.

4. Simulation Settings

The settings of the coefficients of the Monte Carlo simulation study[1] were according to Canary, *et al.* [8], based on the simulation settings that show in Table 2.[2] In true models, setting 1 had three linear covariates, while settings 2 and 3 contained a covariate and it quadratic term but the latter had an extra addition chi-square distributed term, whereas settings 4 and 6 have an interaction term between continuous and categorical covariate and again the latter had and extra addition Bernoulli distributed covariate. Finally, setting 5 had also and interaction term between two continuous covariates.

True models under each setting were examined 1000 times by GOF tests and the percentage of p-value of each GOF test greater than 0.05 was calculated and in this study we referred to this percentage as "The Null" of the test.

For testing the power, independent variable for each setting were generated under the true model terms and then a false model were fitted to the generated independent variable after omitting the last term of the true model. Thus, the ability of GOF tests to detect a missing covariate (misspecification A) were tested under the false model of setting 1 whereas false model of settings 2 and 3 was for misspecification B (missing quadratic term) and for mispsecifcation C (missing interaction term), false models of settings 4, 5 and 6 were used.

Table-1. GOF tests under study

Test	Package	Function	Abbreviation
Hosmer-Lemeshow's "C statistic"	LogisticDx [19]	gof( )	HLc-LDx
	DescTools [20]	HosmerLemeshowTest ( )	HLc-DT
	MKmisc [21]	HLgof.test( )	HLc-MK
Hosmer-Lemeshow's "H statistic"	DescTools	HosmerLemeshowTest ( )	HLh-DT
	MKmisc	HLgof.test ( )	HLh-MK
Modified Hosmer-Lemeshow	LogisticDx	gof ( )	mHL
Osius and Rojek's test of the link function	LogisticDx	gof ( )	OsRo
le Cessie-van Houwelingen-Copas-Hosmer unweighted sum of squares test (CHCH)	DescTools	HosmerLemeshowTest ( )	CHCH-DT
	MKmisc	HLgof.test( )	CHCH-MK
	Rms [22]	Resid ( )	CHCH-Rms
Stukel's tests	LogisticDx	gof ( )	S1:S6

Table-2. Settings used to examine the GOF tests for detecting misspecifications

Setting	True model	x_1	x_2	x_3	β_0	β_1	β_2	β_3	β_4	False model
1	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$	U(-6,6)	N(0,2.25)	$\chi^2(4)$	0	0.267	0.267	0.217	___	$\beta_0+\beta_1x_1+\beta_2x_2$
2	$\beta_0+\beta_1x_1+\beta_2x_1^2$	U(-3,3)	x_1^2	___	-1.963	0.982	0.218	___	___	$\beta_0+\beta_1x_1$
3	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_1^2$	U(-3,3)	N(0,2.25)	$\chi^2(4)$	-2	0.3	0.3	0.3	0.3	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$
4	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	Ber(0.5)	x_1x_2	-1.792	0.135	1.117	0.372	___	$\beta_0+\beta_1x_1+\beta_2x_2$
5	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$	U(-3,3)	U(-3,3)	x_1x_2	0	0.3	0.3	0.3	___	$\beta_0+\beta_1x_1+\beta_2x_2$
6	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_1x_2$	U(1,3)	Ber(0.5)	Ber(0.5)	-1.25	0.5	-2.75	0.75	1.25	$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$

5. Simulation Results

Results were classified to three parts: null results, power results of same tests, and power results of different GOF tests.

5.1. Null Results

All tests under all settings showed percentage of null greater than 92% except:

- The mHL test under all settings showed very weird results where the null decrease as long as the sample size increases as shown in Fig 1.
- The OsRO under Setting 4 and 6 showed a low null results which directly proportional with the sample size BUT it never reaches 90% as shown in Fig 2.
- The S1 and S3 tests had a non-applicable (NA) results under setting 4 in sample sizes 100 up till 500. In addition, S1 and S2 had a NA results under setting 1 for the lowest sample size.

[1] See[16, 17] For using R software to create the Monte Carlo simulation studies in different regression models.
[2] For more details about how generate the logistic model in simulation studies, see [18]

Fig-1. Null results of mHL test (in package 'LogisticDx') for all settings at ten different sample sizes

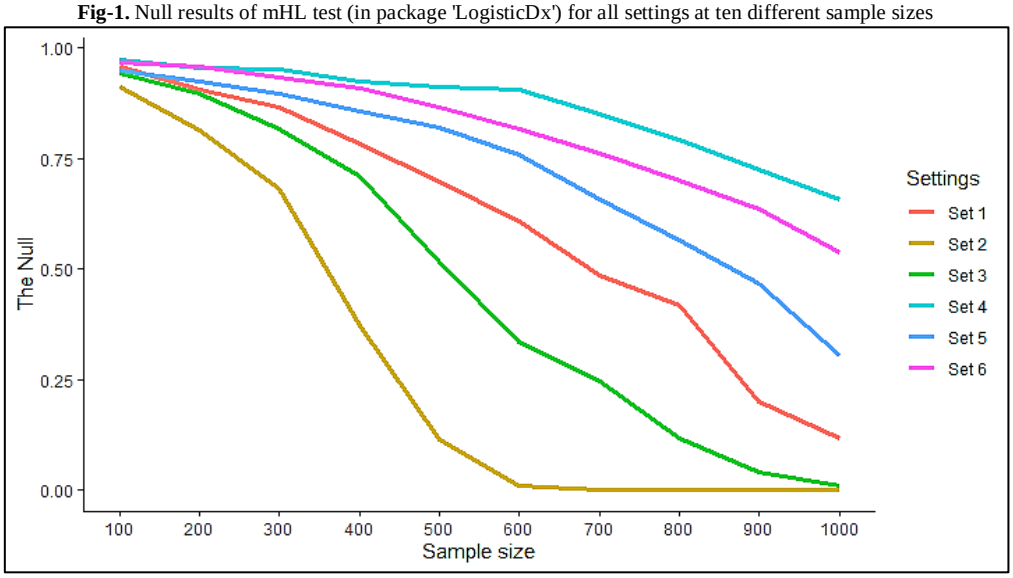
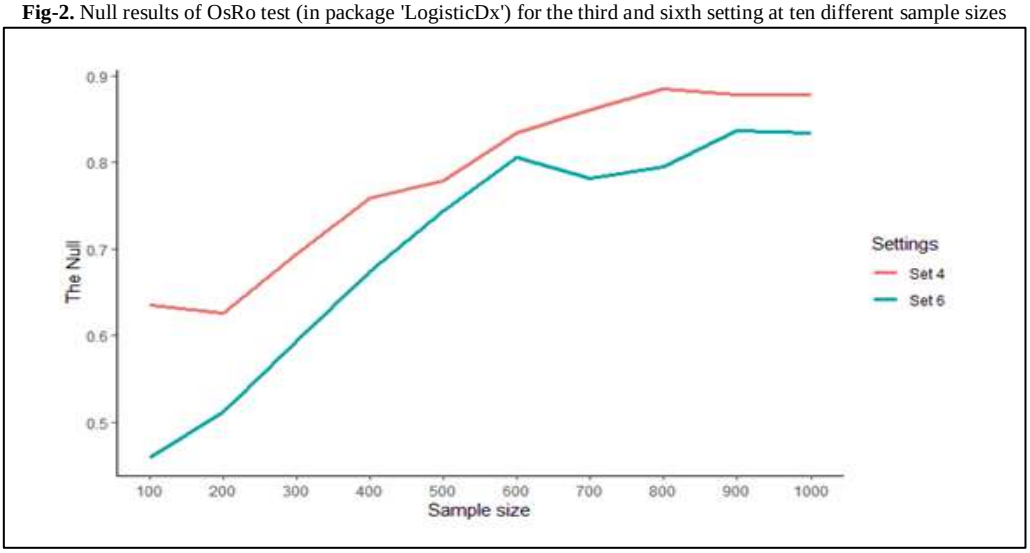


Fig-2. Null results of OsRo test (in package 'LogisticDx') for the third and sixth setting at ten different sample sizes



5.2. Power Results of Same Tests

Results of Hosmer-Lemeshow's C statistic and H statistic were the same under different packages. Whereas le Cessie-van Houwelingen-Copas-Hosmer unweighted sum of squares test (CHCH) had a very weird results. Under the second setting the power of CHCH-MK and CHCH-Rms was the same whereas CHCH-DT was totally different as shown in Fig 3. However, at the rest of settings CHCH-DT and CHCH-MK had a zero power which was different than CHCH-Rms as illustrated in Fig 4 under the largest sample size ONLY for simplification.

Fig-3. Power results of CHCH test of different packages for all sample sizes at setting 2

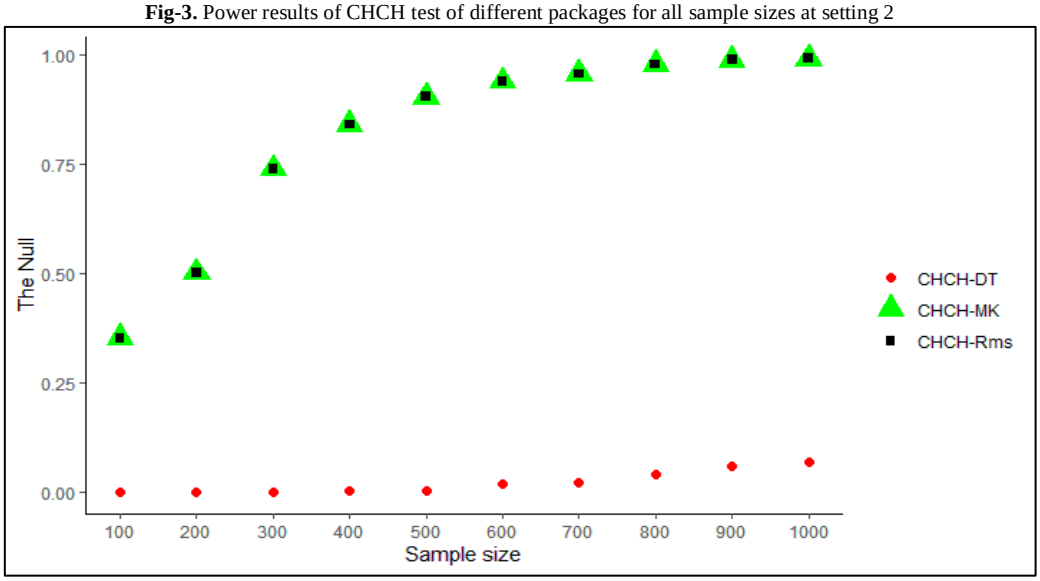
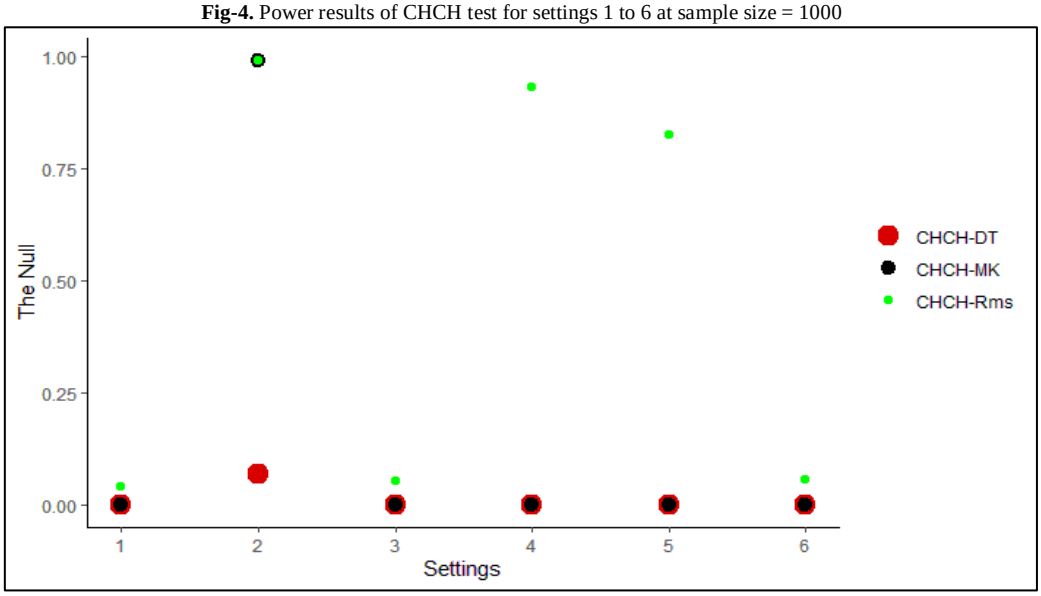


Fig-4. Power results of CHCH test for settings 1 to 6 at sample size = 1000



5.3. Power Results of Different Tests

In general, different tests gave different power pattern under different misspecification. In addition, patterns varied a lot under different settings of same misspecification. For each setting, the Best test would refer to the test which gave a power greater than 0.80 at a lower sample size.

5.3.1. Power Results of Misspecification A

Under setting 1, all the tests, except mHL, under all sample sizes studied had low power less than 0.12% to detect the omission of a covariate with a $\chi 2$ distribution (Table 3).

This result was compatible with [8] as HL-C statistic didn't exceed a power of 0.05. In addition, [12] showed that OsRo and HL-C power at sample size of 100 didn't exceed 0.45 while the power of the former exceed the 0.80 under lower degree of sparseness at sample size 500.

Moreover, Kuss [12] stated that the test which has the best power in his study when a missing covariate exists, was Farrington test where the satisfied power was achieved at more sparse data although this test has the structural deficiency of never rejecting the null hypothesis with extreme sparseness data (when each covariate pattern contains single observation).

It is observable that most of simulation studies of goodness of fit tests didn't include the misspecification of omitting a covariate term as this title is considered mainly under the title of model selection.

On the other hand, the power of mHL test is increased as long as the sample size increased till reached its maximum (71%) at the highest sample size (Table 3). it is expected that, this test may exceed 0.80 at larger sample sizes.

Furthermore, S2 and S3 had a NA results (p-value failed to be calculated) under setting 1 of sample size 100.

Table-3. Power results of all tests for setting 1 (missing covariate term) for all sample sizes

	HL-C	mHL	OsRo	S1	S2	S3	S4	S5	S6	HL-H	CHCH-Rms
n=100	0.052	0.066	0.114	0.033	NA	NA	0.043	0.073	0.102	0.039	0.032
n=200	0.042	0.068	0.086	0.052	0.03	0.05	0.044	0.039	0.049	0.035	0.056
n=300	0.067	0.112	0.057	0.051	0.04	0.052	0.059	0.054	0.071	0.056	0.046
n=400	0.046	0.18	0.083	0.042	0.044	0.054	0.041	0.051	0.043	0.046	0.059
n=500	0.038	0.206	0.076	0.053	0.076	0.087	0.057	0.072	0.061	0.05	0.049
n=600	0.05	0.275	0.052	0.052	0.063	0.084	0.059	0.058	0.054	0.047	0.045
n=700	0.044	0.346	0.066	0.046	0.028	0.052	0.041	0.041	0.047	0.036	0.046
n=800	0.062	0.474	0.057	0.051	0.047	0.074	0.056	0.061	0.069	0.047	0.05
n=900	0.052	0.614	0.053	0.054	0.065	0.075	0.059	0.064	0.064	0.056	0.05
n=1000	0.045	0.712	0.055	0.042	0.036	0.051	0.042	0.044	0.037	0.059	0.042

5.3.2. Power Results of Misspecification B

The misspecification of quadratic term omission was tested under two settings (2 & 3) as shown in Table 4 and Table 5 respectively.

Under the second setting, The Best test S3 where it gave the satisfied power at a sample size =300 whereas the second best tests were OsRo, S2, S5 and CHCH-Rms at sample size 400. On the other hand, the power of Hosmer C and H statistics (which are the default of many packages) didn't exceed 0.80 except at sample size 600 and 1000.

However, the power of all tests failed to reach the satisfied ratio to detect the omission of quadratic term when the model contains a second covariate as in setting 3 (where the power didn't exceed 40%) except mHL at the largest sample size.

Table-4. Power results of all tests for Setting 2 [missing quadratic term] for all sample sizes

	HL-C	mHL	OsRo	S1	S2	S3	S4	S5	S6	HL-H	CHCH-Rms
n=100	0.352	0.255	0.356	0.215	0.397	0.412	0.273	0.305	0.287	0.082	0.353
n=200	0.447	0.597	0.512	0.356	0.555	0.587	0.41	0.49	0.463	0.143	0.503
n=300	0.627	0.887	0.709	0.544	0.79	0.807	0.592	0.743	0.683	0.267	0.74
n=400	0.754	0.979	0.806	0.663	0.875	0.893	0.689	0.852	0.796	0.373	0.842
n=500	0.797	0.998	0.872	0.765	0.934	0.937	0.792	0.915	0.874	0.433	0.905
n=600	0.871	1	0.921	0.829	0.959	0.966	0.849	0.952	0.93	0.539	0.941
n=700	0.9	1	0.937	0.865	0.976	0.977	0.878	0.966	0.949	0.621	0.958
n=800	0.948	1	0.968	0.914	0.99	0.992	0.921	0.99	0.977	0.677	0.98
n=900	0.953	1	0.982	0.949	0.994	0.995	0.953	0.993	0.982	0.765	0.99
n=1000	0.973	1	0.984	0.951	0.995	0.996	0.955	0.994	0.993	0.803	0.993

Table-5. Power results of all tests for Setting 3 [missing quadratic term] for all sample sizes with extra covariate exist in the model

	HL-C	mHL	OsRo	S1	S2	S3	S4	S5	S6	HL-H	CHCH-Rms
n=100	0.077	0.093	0.102	0.058	0.192	0.179	0.144	0.131	0.123	0.125	0.063
n=200	0.073	0.113	0.025	0.033	0.154	0.139	0.109	0.134	0.106	0.058	0.053
n=300	0.051	0.129	0.031	0.041	0.099	0.09	0.06	0.077	0.063	0.061	0.048
n=400	0.174	0.385	0.065	0.173	0.392	0.399	0.275	0.347	0.303	0.189	0.069
n=500	0.099	0.435	0.039	0.093	0.254	0.236	0.136	0.224	0.17	0.097	0.073
n=600	0.068	0.376	0.123	0.065	0.126	0.135	0.078	0.113	0.095	0.084	0.062
n=700	0.068	0.442	0.056	0.12	0.185	0.209	0.164	0.16	0.145	0.062	0.037
n=800	0.118	0.564	0.078	0.162	0.226	0.254	0.202	0.199	0.185	0.118	0.047
n=900	0.084	0.707	0.043	0.145	0.229	0.28	0.208	0.215	0.178	0.094	0.041
n=1000	0.216	0.858	0.076	0.1	0.376	0.349	0.155	0.353	0.256	0.208	0.054

These results agreed with [8] where HL gave a power of 80% only when there was a single covariate and its square in the true model (as in setting 2) at sample size of 500 whereas when there existed other covariates than the squared one, the test fails to detect the misspecification at all the sample sizes under the study.

In addition, at the study of Hosmer and Hjort [23] the HL-C gave the same result as this work. However, the partial sum of resedual test gave a better result for this misspecification.

5.3.3. Power Results of Misspecification C

Misspecificaion C was expressed under three settings (4, 5 and 6). For the missing interaction term between categorical and continuous covarites (Set 4), the OsRo was the best to detect that departure when sample size reach 600 and the second best was CHCH-Rms and S5 whereas the worst tests were Hl-C, HL-H and S4 as even at the largest sample size, the power didn't reach a satisfied percent. Furthermore, under this setting, S1 and S3 failed to be calculated at all sample sizes (Table 6).

Concerning setting 5, where the ommited interaction term was between two continuous covariates, the best test was of OsRo at sample size 200 even at sample size 100 the power was very close to 80%. All the tests gave a power over 90% at sample size 300 Except CHCH-Rms where the its power exeed 80% at sample size 800 (Table 7). It was worth mentioning that S2 and S3 had a NA results (failed to be calculated) under setting 5 of sample size 200. for the last setting (Table 8), unfortunately all the tests (except OsRo) had a very low power (less than 30%) to detect missing interaction term of binary and continuous covariates when an extra covariate exist in the model. a weired unexpected power results gained from OsRo where the power decreased while the sample size increased.

Table-6. Power results of all tests for Setting 4 [missing interaction term between categorical and continuous covariates] for all sample sizes

	HL-C	mHL	OsRo	S1	S2	S3	S4	S5	S6	HL-H	CHCH-Rms
n=100	0.106	0.085	0.426	NA	0.239	NA	0.109	0.198	0.16	0.054	0.237
n=200	0.192	0.201	0.527	NA	0.398	NA	0.143	0.379	0.287	0.149	0.397
n=300	0.247	0.356	0.644	NA	0.552	NA	0.164	0.532	0.415	0.208	0.555
n=400	0.352	0.513	0.695	NA	0.631	NA	0.192	0.607	0.533	0.311	0.64
n=500	0.422	0.633	0.756	NA	0.741	NA	0.232	0.718	0.624	0.368	0.743
n=600	0.488	0.707	0.814	NA	0.786	NA	0.253	0.774	0.69	0.457	0.793
n=700	0.574	0.853	0.884	NA	0.875	NA	0.28	0.866	0.799	0.536	0.881
n=800	0.63	0.892	0.895	NA	0.891	NA	0.263	0.878	0.799	0.612	0.902
n=900	0.689	0.945	0.925	NA	0.913	NA	0.33	0.912	0.86	0.654	0.929
n=1000	0.702	0.949	0.933	NA	0.929	NA	0.321	0.924	0.873	0.671	0.931

Table-7. Power results of all tests for Setting 5 [missing interaction term between two continuous covariates] for all sample sizes

	HL-C	mHL	OsRo	S1	S2	S3	S4	S5	S6	HL-H	CHCH-Rms
n=100	0.292	0.269	0.788	0.619	0.731	0.786	0.703	0.691	0.691	0.454	0.276
n=200	0.67	0.633	0.994	0.788	NA	NA	0.804	0.863	0.843	0.676	0.595
n=300	0.921	0.924	0.913	0.977	0.982	0.988	0.979	0.98	0.976	0.932	0.427
n=400	0.992	0.995	0.931	1	0.997	1	1	0.997	0.998	0.993	0.554
n=500	1	1	0.978	1	1	1	1	1	1	1	0.726
n=600	0.999	0.999	0.989	1	1	1	1	1	1	0.999	0.772
n=700	1	1	0.979	1	1	1	1	1	1	1	0.785
n=800	1	1	0.987	1	1	1	1	1	1	1	0.848
n=900	1	1	0.84	1	1	1	1	1	1	1	0.569
n=1000	1	1	0.97	1	1	1	1	1	1	1	0.825

Table-8. Power results of all tests for Setting 6 [missing interaction term between categorical and continuous covariates] for all sample sizes with extra covariate exist in the model

	HL-C	mHL	OsRo	S1	S2	S3	S4	S5	S6	HL-H	CHCH-Rms
n=100	0.053	0.053	0.804	0.093	0.077	0.115	0.099	0.082	0.098	0.061	0.076
n=200	0.063	0.078	0.811	0.101	0.094	0.135	0.095	0.108	0.089	0.068	0.051
n=300	0.084	0.114	0.767	0.158	0.115	0.193	0.146	0.128	0.116	0.085	0.071
n=400	0.063	0.132	0.694	0.116	0.08	0.141	0.114	0.089	0.094	0.083	0.063
n=500	0.068	0.153	0.592	0.128	0.088	0.164	0.126	0.099	0.095	0.074	0.059
n=600	0.101	0.252	0.501	0.185	0.171	0.241	0.178	0.199	0.154	0.134	0.061
n=700	0.101	0.317	0.661	0.177	0.128	0.214	0.175	0.136	0.132	0.118	0.08
n=800	0.101	0.371	0.433	0.229	0.205	0.288	0.222	0.222	0.196	0.137	0.054
n=900	0.094	0.407	0.559	0.216	0.139	0.235	0.213	0.152	0.163	0.139	0.069
n=1000	0.091	0.506	0.415	0.213	0.173	0.251	0.206	0.188	0.164	0.132	0.056

6. An Application to Breast Cancer Data

It was discovered that the age of patients and the location of cancer (LOC) whether in right or left breast, or both, contributes significantly to the survival of patients [24]. In this study the dataset of Oguntunde, *et al.* [24] was reanalyzed and two models were represented; the first model is the same as Oguntunde, *et al.* [24] but without categorizing the age where the specified linear predictor of model A was:

$$\eta_A = \beta_0 + \beta_1(Age) + \beta_2(LOC)$$

This linear predictor is used to compute the predicted risk of death (PRD) in model A, defined as

$$\text{PRD}_A = \pi(\eta_A) = \frac{e^{\eta_A}}{1 + e^{\eta_A}}$$

The result of the model A was showing in Table 9. It is clear that the magnitude of the intercept is higher than other betas and it was significant which may indicate a missing term in the model.

In this study model B was proposed where,

$$\eta_B = \beta_0 + \beta_1(Age) + \beta_2(LOC) + \beta_3(Age)^2$$

The PRD in model B, $\text{PRD}_B = \pi(\eta_B)$, is defined using the standard logistic function in the same manner as in model A. the result of model B was shown in Table 10 where the magnitude of the intercept decreased nearly four times and became not significant which may indicate a stronger effect of the new term.

Table-9. Logistic regression result of model A

Variable		Coefficient	Standard deviation	z-value	p-value	Odds (95% CI)
Intercept		-3.565	0.577	-6.175	<0.001	0.0283(0.009-0.084)
Age		0.046	0.0101	4.606	<0.0001	1.047 (1.028- 1.069)
LOC (Ref.: Right)	Left	0.704	0.278	2.532	<0.05	2.023 (1.17-3.517)
	Both	1.328	0.478	2.778	<0.01	3.773(1.466-9.692)

Table-10. Logistic regression result of model B

Variable		Coefficient	Standard deviation	z-value	p-value	Odds(95% CI)
Intercept		0.914	1.598	0.572	0.567	2.494 (0.110-59.721)
Age		-0.136	0.064	-2.118	<0.05	0.873 (0.766-0.986)
LOC (Ref.: Right)	Left	0.738	0.286	2.585	<0.01	2.092 (1.203-3.697)
	Both	1.236	0.474	2.610	<0.01	3.443 (1.345-8.742)
Age^2		0.0017	0.0006	2.773	<0.01	1.0017(1.00058-1.0030)

Assessing the fit of both models took place using different R^2 as in Fig 5, Akaike information criterion (AIC) and Bayesian information criterion (BIC) as in Table 11, as well as goodness of fit tests under this study as in Table

12. It was clear that R^2 was always higher in model B than model A whereas AIC and BIC was lower in the second model.

Fig-5. Different types of pseudo R squared calculated by "DescTools" package

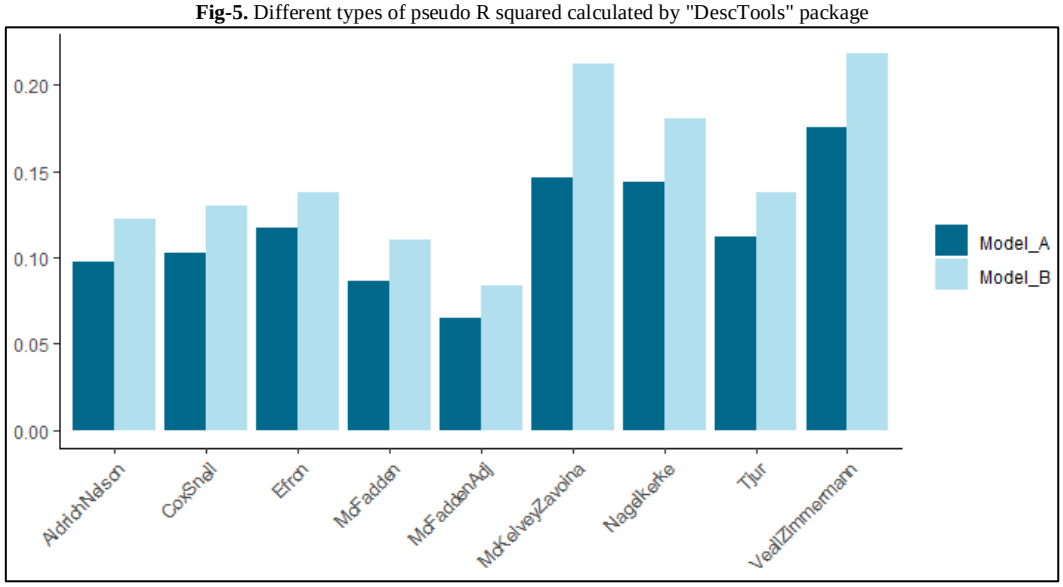


Table-11. Measures of assessing the model fit

Measure	Model A	Model B
AIC	353.10	345.96
BIC	367.92	364.48

Note: AIC calculated by "DescTools" package, while BIC calculated by "blorr" package [25]

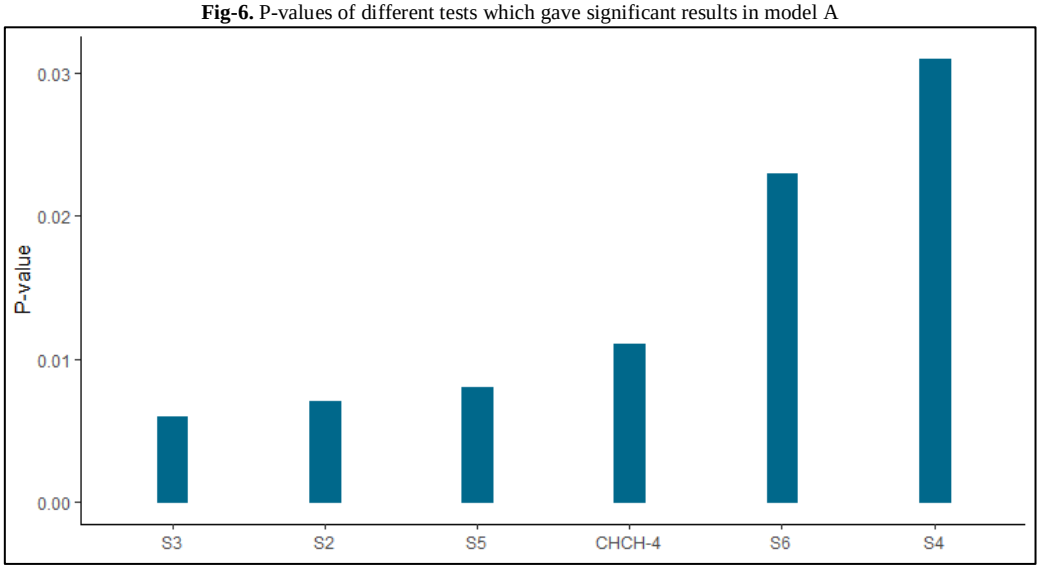
In general, the p-value of most of GOF tests increased in model B which indicated that this model fitted better. However, not all tests detect the missed quadratic term (i.e. not all tests gave significant p-value in model A).

Table-12. Goodness of fit tests for models A and B

Test	Statistic	df	Model A		Model B	
			test value	p-value	test value	p-value
HL-C	chiSq	8	10.114	0.257	4.816	0.777
HL-H	chiSq	8	11.221	0.189	5.038	0.753
mHL	F	9	1.725	0.093	0.636	0.764
OsRo	Z	NA	0.297	0.766	-0.655	0.512
S1	Z	NA	1.692	0.091	1.126	0.260
S2	Z	NA	2.713	0.007	1.466	0.143
S3	chiSq	2	10.225	0.006	3.419	0.181
S4	chiSq	1	4.630	0.031	3.848	0.050
S5	chiSq	1	7.052	0.008	2.250	0.134
S6	chiSq	2	7.530	0.023	3.910	0.142
CHCH-DT	Z	NA	-0.257	0.797	0.000	0.995
CHCH-MK	Z	NA	-2.551	0.011	-0.021	0.983
CHCH-Rms	Z	NA	-2.551	0.011	-0.021	0.983

Furthermore, from all the tests that reject the null hypothesis in model A, the lowest p-value was from S3 test (Fig 6) which is compatible with the simulation study. Moreover, CHCH-MK and CHCH-Rms had the same results as in the simulation study and both detect the omission of quadratic term whereas CHCH-DT had a different result (under the default settings of the function) and failed to reject the null hypothesis in model A.

Fig-6. P-values of different tests which gave significant results in model A



Now the model had two coefficients for the age; a negative one for linear and a positive coefficient for quadratic term. That is mean that, when other variables are fixed, the PRD decreases as long as the age increases till a specific age equal to $\frac{0.1356283}{0.0017289} \approx 78$ years old where the PRD start to increase.

7. Conclusion

Functions of goodness of fit tests existed in R were examined from different aspects. The first aspect was to check the ability of each function to accept the null hypotheses when the model is true (the null). All tests showed a null over 92% except mHL (from package "LogisticDx") because it had a strange pattern that it failed to accept the null hypotheses when the sample size get larger. Furthermore, the null of OsRo test didn't reach 90% at all sample sizes under two different settings where there was an interaction term between binary and categorical covariate existed in the model.

The second aspect was comparing the power of same tests under different packages in R; Hosmer Lemeshow tests (C and H statistics) displayed the same simulation results at all settings whereas the functions of le Cessie-van Houwelingen-Copas-Hosmer unweighted sum of squares test (CHCH) under different packages revealed different power results except when quadratic term present (under setting 2) "MKmisk" and "Rms" were the same. Only the CHCH test of package "Rms" exhibited reasonable results for all settings.

Finally, the power of different tests under various settings and ten sample sizes were compared to select the best test for each setting as the best test defined in this study as the test which gave a power equal to or exceed 80% at the smallest sample size and its null was greater than 90%.

All goodness of fit test unfortunately failed to detect the departure of missing linear covariate term.

For the departure of omitted quadratic term, S3 was the best when sample size is not less than 300 and S2, S3, S5, OsRo, and CHCH-Rms when sample size is equal to or greater than 400 was the best test when the model contained one covariate and its square but in another setting when another covariate added in the true model, all tests failed to reach a satisfied power.

On the other hand, the CHCH test of package "rms", S2 and S5 was the best to detect a missing interaction term between binary and categorical covariates when sample size is not less than 700. However, when another covariate added in the true model, all tests failed to detect such departure.

OsRo, S4, S5, and S6 tests were the best to detect a missing interaction term of two categorical covariates for sample size greater than or equal 200.

Analysis applied on a data of breast cancer where the first model had a categorical and continuous covariate and another proposed model where quadratic term of the continuous covariate added to other covariates. S2, S3, S4, S5, S6, and CHCH-Rms failed to accept the null hypothesis for the first model and the S3 had the lower p-value. In the second model, p-values increased and nearly all of them became not significant. CHCH test of packages "Rms" and "Mkmisc" had same results.

References

[1] Hussain, J. N. and Low, H. C., 2008. "Chi-square goodness-of-fit tests for the logistic regression model: Which one is best? ." In *International Conference on Science and Technology: Applications in Industry and Education.*

[2] Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X., 2013. *Applied logistic regression (third)*. John Wiley and Sons.

[3] Abonazel, M. R. and Ibrahim, M. G., 2018. "On estimation methods for binary logistic regression model with missing values." *International Journal of Mathematics and Computational Science*, vol. 4, pp. 79-85.

[4] Mahdy, S. M., Abonazel, M. R., and Ghallab, M. G., 2020. "A review of ten imputation methods for handling missing values in logistic regression: A medical application, working paper. Faculty of graduate studies for statistical research, Cairo University, Egypt."
[5] Hosmer, D. W. and Lemesbow, S., 1980. "Goodness of fit tests for the multiple logistic regression model." *Communications in Statistics-Theory and Methods,* vol. 9, pp. 1043-1069.
[6] Hosmer, D. W. and Lemeshow, S., 1989. *Applied logistic regression*. New York: Jhon Wiley and Son.
[7] Badi, N. H. S., 2017. "Asymptomatic distribution of goodness-of-fit tests in logistic regression model." *Open Journal of Statistics,* vol. 7, pp. 434-445.
[8] Canary, J. D., Blizzard, L., Barry, R. P., Hosmer, D. W., and Quinn, S. J., 2017. "A comparison of the hosmer--lemeshow, pigeon--heyse, and tsiatis goodness-of-fit tests for binary logistic regression under two grouping methods." *Communications in Statistics-Simulation and Computation,* vol. 46, pp. 1871-1894.
[9] Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S., 1997. "A comparison of goodness-of-fit tests for the logistic regression model." *Statistics in Medicine,* vol. 16, pp. 965-980.
[10] Zhang, J., Ding, J., and Yang, Y., 2019. "A binary regression adaptive goodness-of-fit Test (BAGofT)." Available: https://arxiv.org/abs/1911.03063
[11] Osius, G. and Rojek, D., 1992. "Normal goodness-of-fit tests for multinomial models with large degrees of freedom." *Journal of the American Statistical Association,* vol. 87, pp. 1145-1152.
[12] Kuss, O., 2002. "Global goodness-of-fit tests in logistic regression with sparse data." *Statistics in Medicine,* vol. 21, pp. 3789-3801.
[13] Copas, J. B., 1989. "Unweighted sum of squares test for proportions." *Journal of the Royal Statistical Society: Series C (Applied Statistics),* vol. 38, pp. 71-80.
[14] Stukel, T. A., 1988. "Generalized logistic models." *Journal of the American Statistical Association,* vol. 83, pp. 426-431.
[15] Nygaard, E., 2019. *A simulation study of goodness-of-fit tests for binary regression with applications to norwegian intensive care registry data*. The University of Bergen.
[16] Abonazel, M. R., 2018. "A practical guide for creating Monte Carlo simulation studies using R." *International Journal of Mathematics and Computational Science*, vol. 4, pp. 18-33.
[17] Abonazel, M. R., 2020. "Handling outliers and missing data in regression models using R: Simulation examples." *Academic Journal of Applied Mathematical Sciences,* vol. 6, pp. 187-203.
[18] Abonazel, M. R. and Farghali, R. A., 2019. "Liu-type multinomial logistic estimator." *Sankhya B,* vol. 81, pp. 203-225.
[19] Dardis, C., 2015. "LogisticDx: diagnostic tests for models with a binomial response. R Package Version 0.2." Available: https://rdrr.io/cran/LogisticDx/
[20] Signorell, A., Aho, K., Anderegg, N., Aragon, T., Arppe, A., Baddeley, A., Bolker, B., Caeiro, F., Champely, S., *et al.*, 2018. "DescTools: Tools for descriptive statistics. 2018. R Package Version 0.99, 24." Available: https://andrisignorell.github.io/DescTools/
[21] Kohl, M., 2019. "Mkmisc: Miscellaneous functions from m. Kohl. R package version, 1.6." Available: https://rdrr.io/cran/MKmisc/
[22] Harrell, J. and Frank, E., 2017. *Rms: regression modeling strategies. R package version 5.1-2*. Dept. Biostatist., Vanderbilt Univ., Nashville, TN, USA.
[23] Hosmer, D. W. and Hjort, N. L., 2002. "Goodness-of-fit processes for logistic regression: simulation results." *Statistics in Medicine,* vol. 21, pp. 2723-2738.
[24] Oguntunde, P. E., Adejumo, A. O., and Okagbue, H. I., 2017. "Breast cancer patients in Nigeria: data exploration approach." *Data in Brief,* vol. 15, p. 47.
[25] Hebbali, A., 2019. "Blorr: Tools for developing binary logistic regression models. R package version 0.2.1." Available: https://blorr.rsquaredacademy.com/