




gofcat: An R package for goodness-of-fit of categorical response models

Ejike R. Ugba¹

¹ Department of Mathematics and Statistics, School of Economics and Social Sciences, Helmut Schmidt University, Hamburg, Germany

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 20 April 2022

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

Statistical models are considered simplification or approximation of reality (Burnham & Anderson, 2002). How close to the target a given model is, or how it compares to competing models is always of interest in real-world applications. Answers to such questions are mostly obtained via adequate goodness-of-fit (GOF) procedures. However, while such procedures alongside software implementations are readily available for various continuous outcome models, there are just a handful of open-source implementations available for categorical response models (CRMs).

The gofcat R software package provides a quick means of evaluating some widely used CRMs in empirical studies. Depending on the model of interest, functions are available for the different forms of hypothesis tests associated with CRMs and for computing the summary measures of predictive strength of fits. For instance, the proportional odds assumption in the ordinal regression model can be tested using the Brant or the Likelihood-Ratio tests available in gofcat. Other crucial tests like the Hosmer-Lemeshow, the Lipsitz and the Pulkstenis-Robinson tests are also available for some widely used binary, multinomial and ordinal response models. Moreover, the assessment of prediction errors through error/loss functions and several summary measures of predictive strength of fitted models (pseudo- R^2 s) are also available in gofcat.

Statement of Need

Evaluation of model adequacy via GOF tests and related measures remains a very crucial step in every model-selection procedure. Several forms of GOF tests for regression model analyses already exist in the literature and are very much in use too in empirical studies. However, as earlier noted, not only are there few GOF tests particularly suited for the CRMs, due to little or no open-source software implementations, a lot of users find somewhat ambiguous the application of such methods. R packages, such as goftest (Faraway et al., 2021), ADGofTest (Gil Bellosta, 2011) and AICcmodavg (Mazerolle, 2020), for instance, do provide some GOF tests, but mainly for the continuous outcome models. Other R packages like ResourceSelection (Lele et al., 2019), performance (Lüdtke et al., 2021), generalhoslem (Matthew, 2019) and brant (Schlegel & Steenbergen, 2020), do provide some handful of tests applicable to CRMs. However, available tests in those packages seem restricted to binary models or some unique class of objects when extended beyond the binary settings. For instance, the Hosmer-Lemeshow Test available in the ResourceSelection package supports only binary models, while the Brant Test available in the brant R package supports only objects of class polr().

In contrast, however, while providing several GOF tests that apply to the binary, multinomial and ordinal response models, gofcat also supports several classes of objects, including objects from both s3 and s4 R methods. Currently supported classes include objects of class `serp()` from the `serp` R package (Ugba, 2021), `vglm()` from the `VGAM` R package (Yee, 2010), `clm()` from

the ordinal R package (Christensen, 2019), multinom() from the nnet R package (Venables & Ripley, 2002), polr() from the MASS R package (Venables & Ripley, 2002), mlogit() from the mlogit R package (Croissant, 2020) and glm() from the stats R package (R Core Team, 2021). The last two methods fit only binary models, while the rest fit multinomial/ordinal models. Supported models from these classes can all be directly evaluated using gofcat. However, in situations where the actual model to be evaluated is not available, some tests may still be conducted using instead a data frame of observed and fitted values.

Features and Application

An overview of the main functions of gofcat is given alongside an application to a real-life data example. The data comes from a 6-year longitudinal study on diabetes and retinopathy, with records drawn from 613 diabetic patients (Bender & Grouven, 1998; Jørgens et al., 1993; Mühlhauser et al., 1996). The study aimed to investigate the relationship between retinopathy status and the available risk factors. The outcome variable, retinopathy status (RET), is an ordered factor with three categories: 1 = no retinopathy, 2 = non-proliferative retinopathy, and 3 = advanced retinopathy or blind. The risk factors of interest include smoking (SM), a binary variable with 1 if the patient was a smoker and 0 otherwise, diabetes duration (DIAB) measured in years, glycosylated haemoglobin (GH) measured in percentage, and diastolic blood pressure (BP) measured in mmHg.

A constrained cumulative logit model (also known as proportional odds model) was fit to the data using polr() from the R MASS package (Venables & Ripley, 2002)). The fit is demonstrated in the code chunk below (for brevity, direct code outputs are omitted), with the realized estimates and tests shown in Table 1. It is observed that the effect of smoking is not significant ($p = 0.187$), while the association between retinopathy and the other risk factors are highly significant ($p < 0.0001$).

```
library(gofcat)
library(MASS)

retino <- within(retinopathy, {
  RET <- ordered(RET)
  SM <- factor(SM)
})

RET.fit <- polr(RET ~ SM + DIAB + GH + BP, method = "logistic",
               data = retino)
coefTab <- coef(summary(RET.fit))
pv <- pnorm(abs(coefTab[, "t value"]), lower.tail = FALSE) * 2
```

Table 1: Proportional odds model of the retinopathy dataset. The common significance code "****" indicates values < 0.001 .

Coefficients	B	SE-B	Pr(> z)
(Intercept):1	12.303	1.294	0.000 ***
(Intercept):2	13.673	1.321	0.000 ***
SM1	0.255	0.193	0.187
DIAB	0.140	0.014	0.000 ***
GH	0.460	0.076	0.000 ***
BP	0.072	0.014	0.000 ***

66 Post Estimation Tests

67 Applicable GOF tests for the fitted model (RET.fit) together with other yardsticks of model
68 evaluation were further obtained using the available gofcat functions. A brief overview of
69 what each function does is given, followed by an application to the model under consideration.

70 hosmerlem()

71 This function performs the Hosmer-Lemeshow (HL) test for CRMs. Details of the test for the
72 binary outcome model are given in Hosmer & Lemeshow (1980). An extension to multinomial
73 models is found in Fagerland et al. (2008) and Fagerland & Hosmer (2012), while the extension
74 to ordinal models is found in Fagerland and Hosmer (2013, 2016, 2017). In each of the three
75 settings, one splits the original observations into k groups (10 by default), after ranking by
76 ordinal scores (OS). A Pearson Chi-square Statistic is then obtained from the expected and
77 observed frequency tables. For a good chi-square approximation, it is recommended to have at
78 least 80% of the estimated frequencies > 1 (Fagerland & Hosmer, 2017). About 97% of the
79 estimated frequencies from RET.fit meet this criterion. However, as observed in Table 2, the
80 HL test is significant ($p = 0.014$), indicating a possible lack of fit.

```
hosmerlem(RET.fit, tables = TRUE)
```

81 lipsitz()

82 This function computes the Lipsitz test for the ordinal models (Lipsitz et al., 1996). For
83 this test, one also splits the observations into k separate groups using the ordinal scores
84 of the estimated values. An indicator variable denotes the observations belonging to each
85 group, producing additional pseudo-variables with which the original model could be updated.
86 Supposing the original model fits correctly, then the coefficients of the pseudo-variables all
87 equal zero. The likelihood ratio statistic obtained from the log-likelihoods of the original and
88 the refitted models is subsequently compared with the chi-square distribution having $k - 1$
89 degrees of freedom. In contrast to the LH test (Table 2), the Lipsitz test for the RET.fit is not
90 statistically significant ($p = 0.459$), implying that no lack of fit was detected. More discussion
91 on this follows shortly.

```
lipsitz(RET.fit)
```

92 pulkroben()

93 This function performs the Pulkstenis-Robinson (PR) test for ordinal models (Pulkstenis and
94 Robinson 2004). It particularly forms groups of observations using covariate patterns obtained
95 from the categorical covariates. Each covariate pattern is subsequently split in two based
96 on the median ordinal scores. The test statistic (chi-sq or deviance) is obtained using the
97 tabulated observed and estimated frequencies. Let's assume that c is the number of covariate
98 patterns, r the number of response categories and k the number of categorical variables in
99 the model, the test statistic approximates the chi-sq distribution with $(2c - 1)(r - 1) - k - 1$
100 degrees of freedom. Considering the RET.fit once again (Table 2), the conducted PR test
101 turns out significant ($p = 0.011$), supporting the initial idea of lack of fit.

```
pulkroben(RET.fit, test = "chisq", tables = TRUE)
```

102 So far, two out of the three GOF tests (HL and PR) for the RET.fit suggest a possible lack of
103 fit. The Lipsitz test, in particular, detected no lack of fit. However, as observed in Fagerland &
104 Hosmer (2017), the Lipsitz test (together with the HL test) is best suited to detect lack of fit
105 associated with continuous covariates, while the PR test works well when lack of fit is associated
106 with categorical predictors. So, a potential explanation of the insignificant result of the Lipsitz
107 test is that the lack of fit is driven by the binary/categorical variable smoking. Meanwhile, it
108 is particularly recommended to compare the results of the ordinal Hosmer-Lemeshow test with
109 the Lipsitz and the Pulkstenis-Robinson tests (Fagerland & Hosmer, 2016, 2017).

110 `brant.test()`

111 The Brant test provides the means of testing the parallel regression assumption in ordinal
112 regression models. The test follows the procedures outlined in Brant (1990), also see the
113 explanations in the Section 2 of Ugba et al. (2021). Looking at Table 1, the Brant test of the
114 RET.fit (Omnibus) is significant ($p = 0.035$), indicating that the proportional odds assumption
115 is violated. However, a closer look at the individual variables in the brant test (Table 2) shows
116 that the non-proportionality is primarily driven by the only categorical variable (smoking) in
117 the model, which, as earlier suspected, was also causing the lack of fit in a more general sense.

```
brant.test(RET.fit)
```

118 `Rsquared()`

119 Several summary measures of predictive strength of CRMs (pseudo- R^2) are obtained with
120 this function. These include both likelihood and non-likelihood-based pseudo- R^2 measures.
121 For instance, the recently proposed modification of the McFadden's R^2 for binary and ordinal
122 outcome models can be obtained via this function (McFadden, 1974; Ugba & Gertheiss, 2018,
123 2022). As reported in Table 2, the McFadden's and the Ugba & Gertheiss' R^2 s for the RET.fit
124 are respectively 0.191 and 0.405. See, Ugba & Gertheiss (2022) for further details on the
125 reported measures.

```
Rsquared(RET.fit, measure = "mcfadden")  
Rsquared(RET.fit, measure = "ugba")
```

126 `erroR()`

127 This function calculates some useful error metrics of fitted binary and multi-categorical response
128 models. Available measures include the brier score (Brier, 1950), the cross-entropy loss (log
129 loss) and the misclassification error. Once again, considering the RET.fit, the obtained metrics
130 (Table 2), do not suggest a well performed fit, as more than 30% misclassification on the
131 training data was observed.

```
erroR(RET.fit, type = "brier")  
erroR(RET.fit, type = "logloss")  
erroR(RET.fit, type = "misclass")
```

Table 2: Post-estimation tests for the proportional odds model of the retinopathy dataset, featuring hypothesis tests for lack of fit, tests for proportional odds assumption and summary/error metrics of fit. The significance code "*" indicates values < 0.05 .

Hypothesis Tests	Chi-sq	df	pr(>chi)
Hosmer-Lemeshow (HL)	32.148	17	0.014 *
Lipsitz	8.764	9	0.459
Pulkstenis-Robinson (PR)	13.094	4	0.011 *

Brant Test	chi-sq	df	pr(>chi)
Omnibus	10.38	4	0.035 *
SM1	4.99	1	0.026 *
DIAB	2.21	1	0.137
GH	1.63	1	0.202
BP	1.37	1	0.241

R -squared	value
McFadden's R^2	0.191
Ugba & Gertheiss' R^2	0.405

Error Metrics	value
Brier Score	0.427
LogLoss	0.737
Misclassification Error	0.318

Conclusion & Outlook

The development of the `gofcat` software package is geared towards a stress-free evaluation of some widely used regression models in empirical studies. As shown in this paper, `gofcat` bundles together crucial GOF tests for CRMs, also providing some quick means of assessing their strength and performance. Several classes of objects are supported and can be directly handled by the available functions, yielding the desired test results. The provided example in this article deals with an ordinal outcome model, even though, as earlier hinted, both the binary and the multinomial outcome models can likewise be assessed using the functions in `gofcat`. Tests for the ordinal models, in particular, are currently available for the constrained forms of the proportional odds model, the adjacent-category model and the continuation-ratio model. Future development of `gofcat` will include tests other than those presently covered and tests for the unconstrained ordinal models.

Availability

The released version of `gofcat` is available in the Comprehensive R Archive Network ([CRAN](#)) ([R Core Team, 2021](#)), with details about usage provided in the package [documentation](#). Alternatively, the development version is available from [GitHub](#) with further details found on a pkgdown website on [Github Pages](#) ([Wickham & Hesselberth, 2020](#)).

Acknowledgements

The author would like to thank Jan Gertheiss for his helpful suggestions. This project was partly supported by Deutsche Forschungsgemeinschaft (DFG) under Grant GE2353/2-1.

References

- Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *J. Clin. Epidemiol.*, 51, 809–816. [https://doi.org/10.1016/s0895-4356\(98\)00066-3](https://doi.org/10.1016/s0895-4356(98)00066-3)
- Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46, 1171–1178. <https://doi.org/10.2307/2532457>
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather. Rev.*, 78, 1–3.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretical approach*. Springer. <https://doi.org/10.1007/b97636>

- Christensen, R. H. B. (2019). *Ordinal—regression models for ordinal data*. R package version 2019.12-10. <https://CRAN.R-project.org/package=ordinal>
- Croissant, Y. (2020). Estimation of random utility models in R: The mlogit package. *Journal of Statistical Software*, 95(11), 1–41. <https://doi.org/10.18637/jss.v095.i11>
- Fagerland, M. W., & Hosmer, D. W. (2012). A generalized hosmer-lemeshow goodness-of-fit test for multinomial logistic regression models. *Stata Journal*, 12, 447–453. <https://doi.org/10.22004/ag.econ.229435>
- Fagerland, M. W., & Hosmer, D. W. (2013). A goodness-of-fit test for the proportional odds regression model. *Statistics in Medicine*, 32, 2235–2249. <https://doi.org/10.1002/sim.5645>
- Fagerland, M. W., & Hosmer, D. W. (2016). Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation*, 86, 3398–3418. <https://doi.org/10.1080/00949655.2016.1156682>
- Fagerland, M. W., & Hosmer, D. W. (2017). How to test for goodness of fit in ordinal logistic regression models. *Stata Journal*, 17, 668–686. <https://doi.org/10.1177/1536867X1701700308>
- Fagerland, M. W., Hosmer, D. W., & Bofin, A. M. (2008). Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine*, 27, 4238–4253. <https://doi.org/10.1002/sim.3202>
- Faraway, J., Marsaglia, G., Marsaglia, J., & Baddeley, A. (2021). *GofTest: Classical goodness-of-fit tests for univariate distributions*. R package version 1.2-3. <https://CRAN.R-project.org/package=gofTest>
- Gil Bellosta, C. J. (2011). *ADGofTest: Anderson-darling GoF test*. R package version 0.3. <https://CRAN.R-project.org/package=ADGofTest>
- Hosmer, D. W., & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9, 1043–1069. <https://doi.org/10.1080/03610928008827941>
- Jørgens, V., Grüsser, M., Bott, U., Mühlhauser, I., & Berger, M. (1993). Effective and safe translation of intensified insulin therapy to general internal medicine departments. *Diabetologia*, 36, 99–105. <https://doi.org/10.1007/BF00400688>
- Lele, S. R., Keim, J. L., & Solymos, P. (2019). *ResourceSelection: Resource selection (probability) functions for use-availability data*. R package version 0.3-5. <https://CRAN.R-project.org/package=ResourceSelection>
- Lipsitz, S. R., Garrett, M. F., & M, G. (1996). Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45, 175–190. <https://doi.org/10.2307/2986153>
- Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
- Matthew, J. (2019). *Generalhoslem: Goodness of fit tests for logistic regression models*. R package version 1.3.4. <https://CRAN.R-project.org/package=generalhoslem>
- Mazerolle, M. J. (2020). *AICcmodavg: Model selection and multimodel inference based on (q)AIC(c)*. R package version 2.3-1. <https://cran.r-project.org/package=AICcmodavg>
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics P. Zarembka (Ed.)*, 105–142.
- Mühlhauser, I., Bender, R., Bott, U., Jørgens, V., Grüsser, M., Wagener, W., Overmann, H., & Berger, M. (1996). Cigarette smoking and progression of retinopathy and nephropathy

- 208 in type 1 diabetes. *Diabetic Med*, 13, 536–543.
- 209 R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation
210 for Statistical Computing. <https://www.R-project.org/>
- 211 Schlegel, B., & Steenbergen, M. (2020). *Brant: Test for parallel regression assumption*. R
212 package version 0.3-0. <https://CRAN.R-project.org/package=brant>
- 213 Ugba, E. R. (2021). Serp: An r package for smoothing in ordinal regression. *Journal of Open*
214 *Source Software*, 6(66), 3705. <https://doi.org/10.21105/joss.03705>
- 215 Ugba, E. R., & Gertheiss, J. (2018). An augmented likelihood ratio index for categorical
216 response models. In *Proceedings of 33rd International Workshop on Statistical Modelling*,
217 2, 293–298.
- 218 Ugba, E. R., & Gertheiss, J. (2022). A modification of McFadden's R^2 for binary and ordinal
219 response models. <https://arxiv.org/abs/2204.01301>
- 220 Ugba, E. R., Mörlein, D., & Gertheiss, J. (2021). Smoothing in ordinal regression: An
221 application to sensory data. *Stats*, 4, 616–633. <https://doi.org/10.3390/stats4030037>
- 222 Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). Springer.
223 <https://www.stats.ox.ac.uk/pub/MASS4/>
- 224 Wickham, H., & Hesselberth, J. (2020). *pkgdown: Make static HTML documentation for a*
225 *package*. R package version 1.6.1. <https://CRAN.R-project.org/package=pkgdown>
- 226 Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical*
227 *Software*, 32(10), 1–34. <https://doi.org/10.18637/jss.v032.i10>