Goodness of fit measures for categorical response models: The CatgoF R-package

Ejike R. Ugba

October 16, 2017

Abstract

A brief overview of the goodness-of-fit measures for some common discrete dependent variable models is presented, with the `CatgoF` package developed to provide a quick and easy computation of the discussed measures. A reproducible example is also provided.

1 Introduction

Statistical models are considered mere simplification or approximation of reality (Burnham & Anderson, 2002). As such, one may want to answer at least two crucial questions about any fitted model. First, how "close" is such to reality? And second, how does it compare to competing models? Answers to such questions and the likes are often sought for via the use of diverse goodness-of-fit routines. However, given that the term 'reality' is not always known, such answers to a very large extent remain probabilistic. Several goodness-of-fit measures already exist in the literature, ranging from the classical to the very recently proposed methods. The coefficient of determination, for instance, provides a quick and direct assessment of fitted continuous response models. Although not directly applicable to the categorical response models (CRMs), a lot of R^2-like measures mimicking the coefficient of determination, often known as pseudo-R^2s, are abundantly available for the CRMs. These, for instance, are studied in Veall & Zimmermann (1992), Menard (2000) and Ugba & Gertheiss (2017). These papers, studied via simulation and real data studies the performance of various goodness-of-fit measures applicable to the CRMs given an underlying latent measure. The third paper, in particular, proposes a category-synchronizing likelihood ratio index for the assessment of both binary and polytomous outcome models.

In addition to the pseudo-R^2 measures, use is also being made of various information criteria measures both for the assessment and comparison of fitted models. Some information criteria yardsticks based on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) provide a lot of valuable insights about fitted models. These include the Akaike weights and differences, the Schwartz weights and differences and the evidence ratios. These measures also aid model selection pretty good. Further methods of model assessments include assessment of prediction errors through loss function optimizations and then the use of various hypothesis test such as the likelihood ratio test, score test, Wald test etc. Although several of these model assessment tools are very widely recognized, their applications are most often ambiguous due to little or no software implementations of such routines. The following packages in R: ADGofTest, AICcmodavg, goftest, and MuMin among others are designed bearing some of these methods in mind. Nevertheless, virtually all of them do not return the pseudo-R^2s for fitted models. Hence, the `CatgoF` package would provide a good supplement to existing packages. It incorporates mainly all the pseudo-R^2s measures discussed in Ugba & Gertheiss (2017) as well as some crucial information measures earlier mentioned for CRMs. As such, the package in question will make it very easy and possible to determine the amount of variations explained in a given CRM as well as how it compares with candidate models. In what follows, a brief overview of the theoretical representations of some of the highlighted methods is presented followed by a detailed description of the `CatgoF` package and then a wrap up with a conclusion.

2 Pseudo-R^2 Measures

Most pseudo-R^2 measures serve as preliminary comparative and diagnostic tools for the CRMs. They tend to furnish a very quick information about the amount of variation present in a given model and the extent

to which the response variable is predicted. Most categorical response models are estimated using the maximum likelihood estimation routine which does always returns the log-likelihood (l_f) at which the model is estimated. This final log-likelihood, when compared to the null log-likelihood (l_0) under different formulations, provides some performance status of a given fitted model in terms of variation explained in the model. Most goodness-of-fit measures depend extensively on these two likelihood values. Such measures are found in McFadden (1973), Cox & Snell (1989), Nagelkerke (1991), Aldrich & Nelson (1984), Veal & Zimmamann (1992) and Ugba & Gertheiss (2017). The most recent paper provides a generalized likelihood ratio index

Measure	Likelihood Measures
Aldrich & Nelson	$R^2_{al} = 2(l_f - l_0)/(2(l_f - l_0) + n)$
Cox & Snell	$R^2_{cs} = 1 - (\frac{L_0}{L_f})^{\frac{2}{n}}$
McFadden	$R^2_{mf} = 1 - (\frac{l_f}{l_0})$
Nagelkerke	$R^2_{nk} = (1 - (L_0/L_f)^{\frac{2}{n}})/(1 - L_0^{\frac{2}{n}})$
Veall & Zimmermann	$R^2_{vz} = [2(l_f - l_0)/(2(l_f - l_0) + n)] \,/\, [-2l_0/(n - 2l_0)]$
Ugba & Gertheiss	$R^2_{ug} = 1 - (\frac{l_f}{l_0})^{\sqrt{2k}}$

Measure	Non Likelihood Measures
Efron	$R^2_{ef} = 1 - (\sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2)$
McKelvey & Zavoina	$R^2_{mz} = \hat{V}ar(\hat{y}^*)/(\hat{V}ar(\hat{y}^*) + Var(\epsilon))$
Tjur	$R^2_{tj} = \bar{\hat{\pi}}_1 - \bar{\hat{\pi}}_0$

Table 1. *Likelihood and non likelihood-based pseudo-R2 measures for assessment of categorical response model. L_0 and L_f are the null and full model likelihoods, n and k the sample size and the number of response category respectively*

unifying several likelihood-based goodness-of-fit measures and also proposed a likelihood-based measure capable of handling both binary and polytomous outcome models. The authors found the proposed measure very useful as it competes very favorably with other measures and also enjoys a lot of advantage over many. On the other hand, the non-likelihood-based measures do not utilize the information furnished by the likelihood function, but are rather formulated based on the observed values y_i and estimated outcomes $\hat{y}$ or the predicted probabilities $\hat{\pi}$. Such measures are found in Mckelvey & Zavoina (1976), Efron (1978) and Tjur (2009). Apart from the Tjur measure, the rest of the measures were formulated mimicking the coefficient of determination in terms of explained variations. The Tjur measure, in particular, uses only the observed versus the predicted values in its formulation. The main limitation of some of these non-likelihood measures is their inability to extend to polytomous models like the others and also interpretability problem. The different formulations for all the measures referred to are given in Table 1. See Ugba & Gertheiss (2017) for more discussion on these measures.

3 Information Criteria Measures

Given two continuous densities $\xi^{'}$ & ξ, where $\xi^{'}$ represents the true density and ξ a parameterized candidate density. Suppose $E^{'}$ denotes an expectation of the true density $\xi^{'}$, it follows then that the Kullback-Leibler distance between the two continuous densities is given by

$$KL(\xi^{'}, \xi) = E^{'} loga(\frac{\xi^{'}(r))}{\xi(r)}) = E^{'} log(\xi^{'}(r)) - E^{'} log(\xi(r)). \tag{1}$$

$KL(\xi^{'}, \xi)$ is seen as the information loss in using ξ to approximate $\xi^{'}$ and also the distance between two statistical expectations. However, since the first expectation depends on the unknown true distribution it is

often treated as a constant which eventually vanishes while comparing two candidate models leaving only the amount of discrepancy between the duo. For simplicity sake, this discrepancy is often stated as $-E^{'}log(\xi(r))$, see Tutz (2011). Akaike (1973) proposed an information criterion that relies upon this measure by showing that for model selection one needs to maximize the expectation of the log-likelihood across several competitive models. He suggested using an approximation of $E_yE_r(loga\xi(r|\hat{\theta}(y)))$which is given by $l(\hat{\theta}) - \tau$. Where $\hat{\theta}(y)$ is the estimate of parameter θ estimated from sample y, $l(\hat{\theta})$ the log-likelihood evaluated at θ and τ the dimensionality of parameter θ. The Akaike Information Criteria is defined as

Δ_i	ER	Level of empirical support for Model i
0 - 2	1.0 - 2.7	Substantial
4 - 7	7.4 - 33.1	Considerably less
> 10	> 148	Essentially none

Table 2. *Benchmark for determining the level of empirical support for a candidate model.*

$$AIC = -2(l(\hat{\theta})) - \tau) \tag{2}$$

Hurvich & Tsai (1989) suggested a bias-corrected AIC that is more appropriate for small sample sizes given by

$$AIC_c = AIC + \frac{2\tau(\tau+1)}{(n-\tau-1)}. \tag{3}$$

Assuming one considers $exp(-\frac{AIC}{2})$ as likelihood of a model, given data, for two candidate models M_j and M_i one may ascertain how much likelier M_j is than M_i with the evidence ratio (ER) given by

$$ER = exp(-\frac{AIC_j}{2})/exp(-\frac{AIC_i}{2}) \tag{4}$$

To determine the level of empirical support for a candidate model, one may also consider the ER or the AIC differences (Δ_i) given by

$$\Delta_i = AIC_i - AIC_{min} \tag{5}$$

Burnham & Anderson (2002) suggest some benchmarks for determining if an empirical support for a candidate model is substantial or not. This is summarized in Table 2. In general, the larger Δ_i is, the less plausible it is that the fitted model is the KL best model, given the data. See also Murtaugh (2014).

Furthermore, the AIC weights w_i also proves very useful in comparing a list of r number of models. w_i is considered the weight of evidence in favor of M_i being the actual KL best model in the set of candidate models.This is given by

$$w_i = \frac{exp(\Delta_i/2)}{(\sum_r exp(\Delta_r/2))} \tag{6}$$

Alternatively, one may consider the Schwarz's Bayesian Information Criterion (BIC) more appropriate than the AIC since it provides a more stringent measure than the AIC in terms of variable selection. It is given by

$$BIC = -2(l(\hat{\theta})) + \tau loga(n). \tag{7}$$

When the true model is among candidate models, BIC identifies the true model as the number of observation $n \to \infty$ making it a consistent model selector. To obtain the Schwarz differences, weights and evidence ratios one replaces AIC with BIC in (4), (5) and (6). Similarly, should $AICc$ instead of AIC be preferred one replaces instead with (7). For more discussion on these measures see Burnham & Anderson (2002).

4 The {CatgoF} Package

The `CatgoF` package is designed to compute and return the above-discussed algorithms for categorical response models. Virtually all class of model building routines in R for categorical outcome models return the null & fitted model log-likelihoods, the deviance, the predicted values and many more useful objects needed to obtain the goodness-of-fit of such models. The `CatgoF` package relies extensively on such generated objects in computing the various goodness-of-fit measures for categorical response models.

4.1 goFit function

The `goFit` function, in particular, accepts either a single or list of objects of class *glm, vglm, clm2, polr & multinom.* It further extracts, computes and returns an object of class `goF` comprising of various tables of goodness-of-fit measures. The glm families supported by the `goFit` function include: *gaussian, binomial, Gamma, poisson, inverse.gaussian and quasi* together with their individual link functions. From the vglm function, the following methods together with their individual link functions are also supported: *acat, cumulative, propodds, cratio, sratio, multinomial, brat, bratt and ordpoisson.* The different varieties of methods available in polr and clm2 have to do with different link functions. They share the following link functions in common; logistic, probit, cloglog and cauchit with clm2 having additionally *Aranda-ordaz and log-gamma.* All these methods together with those of the multinom function are all supported by `goFit`. The computational routine of `goFit` is as illustrated in the flowchart of Figure 1, alongside the input-output arguments. goFit output includes an information criterion table, an evidence ratio table and a pseudo-R^2 table. The comparison measures are of course not returned if just a single object (*m1*) is passed on to `goFit`, two or more objects are required for such measures to be returned. These could either be multiple objects (*m1,m2,...*) or a list of such objects *list*(*m1,m2,...*).

The returned information criteria table is ranked with the best model on top of the table. In case of a large number of models being compared, the *display* argument offers the option of seeing the *best*, *top* two or *all* the computed results. When not supplied, the best is returned by default. This is also true for the evidence ratio results. The *criterion* argument offers the option of using either of the *AICc*, *AIC* or *BIC* for computing the information measures with *AICc* serving as default. The pseudo-R^2 output of `goFit` contains R^2 measures supporting the model being assessed. For instance, since not all measures available for the binary outcome models are applicable to the polytomous outcome models, such measures are not returned for the latter. Furthermore, `goFit` replaces the initial object names with some internally generated labels (*fit1*, *fit2*, ...), following the ordering of objects in the function or in the list of objects supplied to the function. For instance, given the list (*m3, m4, jr, m1*), `goFit` re-labels the objects as follows:*fit1*$\to m3$, *fit2*$\to m4$, *fit3*$\to jr$, and *fit4*$\to m1$, thus respecting the ordering of objects but ignoring their initial labels.

4.2 plotgoF function

The `plotgoF` function in `catgoF` provides visualizations of the computed pseudo-R^2s and the information weights of fitted models. It accepts an inputted object of class goF with an extra argument that selects if either of the pseudo-R^2s or the information weights should be returned. It returns the pseudo-R^2s by default. However, though the plotgoF function returns the Pseudo-R^2s for every object of class `goF`, the same is not the case for the information measures. For such measures to be returned, it is required that the supplied object results from a prior comparison of at least two or more models. Furthermore, in situations where many models are compared, plotgoF displays the results for the top five models. The argument specifications of the `plotgoF` function are shown in the example to follow, where *obj* denotes an object of class `goF` and *type* the required goodness-of-fit measures.

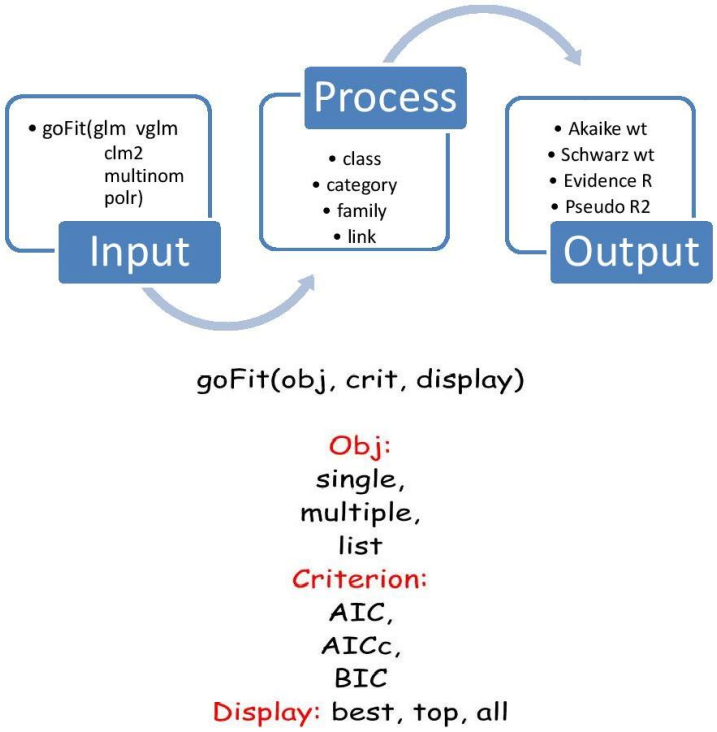

Figure 1: Flowchart illustration of the goFit function and specification of its function call

5 Data Example

To illustrate the use of `CatgoF` functions, let's first simulate some couple of categorical outcome models. Four sets of explanatory variables of length $n = 200$ are drawn thus: $x_1 \sim N(n, 0, 1)$, $x_2 \sim (n, 2, 3)$, $x_3 \sim N(n, 5, 2)$, and $x_4 \sim U(n, 2, 5)$. Two different sets of dependent variables are obtained by splitting the distribution of the continuous model y^* at the median for the binary outcomes (bdv) and at the first, second and third quartiles for the polytmous outcomes (pdv).

$$y^* = 0.2x_1 - 0.5x_2 + 0.4x_3 - 0.3x_4 + 2 + N(n, 0, 1)$$

The simulated dataset is used in building different forms of categorical response models supported by `CatgoF`. The fitted models were subsequently passed on to the `goFit` function for goodness-of-fit assessment and to also generate an object of class `goF`. The different fitted models are given below.

```
library(CatgoF)
require("MASS") ## for polr method
```

```
## Loading required package: MASS
```

```
### Binary & polytmous outcome models
set.seed(63)
x1  <- rnorm(200,0,1); x2 <- rnorm(200,2,3)
x3  <- rnorm(200,5,2); x4 <- runif(200,2,5)
cnt <- 0.2*x1 - 0.5*x2 + 0.4*x3 - 0.3*x4 + 2 + rnorm(200, 0, 1)

pdv <- as.factor(cut(cnt, breaks=quantile(cnt), include.lowest=TRUE, labels=c(1, 2, 3, 4)))
```

```
qnt <- quantile(cnt)
bdv <- cut(cnt, breaks = c(-Inf,qnt[3],Inf), include.lowest=TRUE, labels = c(0,1))

m1 <-  glm(bdv ~ x1 + x2, family = binomial(link = "probit"))
m2 <- polr(pdv ~ x1 * x2 + x3, method = c("logistic"))
m3 <- polr(pdv ~ x1 + x2 * x3, method = c("logistic"))
m4 <- polr(pdv ~ x1 + x2 + x3, method = c("logistic"))
m5 <- polr(pdv ~ x1 * x2 * x3, method = c("logistic"))
m6 <- polr(pdv ~ x1 * x2 + x3 * x4, method = c("logistic"))
mlst <- list(m2, m3, m4, m5, m6)
```

The goFit output for a single, multiple and list of input objects are given below. For the first result (single object) the extra arguments of goFit are not necessary and hence not included in the function call. Moreover, only the basic information measures and the pseudo-R^2s are returned. The second and third examples include the information differences and weights as well as the evidence ratios. Judging by the benchmark provided in Table 2, the models identified as the best among candidate models are substantially different from the rest. The values of both Akaike and Schwarz weights also support this position.

```
### Single object
goFit(m1)
```

```
## $Info.Criterion
##   Categ  logLik   Param     BIC     AIC    AICc
##   2.000 -84.056   3.000 184.008 174.113 174.235
##
## $Pseudo.R2
##                     value
## Cox & Snell         0.421
## Nagelkerke          0.561
## McFadden            0.394
## Aldrich & Nelson    0.353
## Veal & Zimmermann   0.608
## Ugba & Gertheiss    0.632
## McKlevey & Zavoina  0.651
## Efron               0.456
## Tjur                0.455
##
## attr(,"class")
## [1] "goF"
```

```
### list of objects
goFit(mlst, crit = "BIC",display="all")
```

```
## $Info.Criterion
##      categ  logLik Param    BIC BIC.df Schwarz.wt
## fit3     4 -167.93     6 367.64   0.00       0.84
## fit1     4 -167.62     7 372.32   4.68       0.08
## fit2     4 -167.88     7 372.86   5.22       0.06
## fit5     4 -163.75     9 375.19   7.55       0.02
## fit4     4 -166.92    10 386.83  19.19       0.00
##
## $Evidence.Ratio
```

```
##       Model       ER
## 1 fit3|fit1    10.38
## 2 fit3|fit2    13.60
## 3 fit3|fit5    43.60
## 4 fit3|fit4 14691.14
##
## $Pseudo.R2
##                     fit3
## Cox & Snell        0.665
## Nagelkerke         0.709
## McFadden           0.394
## Aldrich & Nelson   0.522
## Veal & Zimmermann 0.711
## Ugba & Gertheiss   0.758
##
## attr(,"class")
## [1] "goF"
```

Figure 2 shows the `plotgoF` output for a single object of class goF while Figures 3 shows the information weights of some nested categorical outcome models. In terms of the returned pseudo-R^2s, the Mckelvey & Zavoina, Ugba & Gertheiss and Veal & Zimmermann R^2s report more effect than the rest of the measures for the assessed model (*m1*).

```
### list of objects
sr <- goFit(m1)
plotgoF(sr)
```

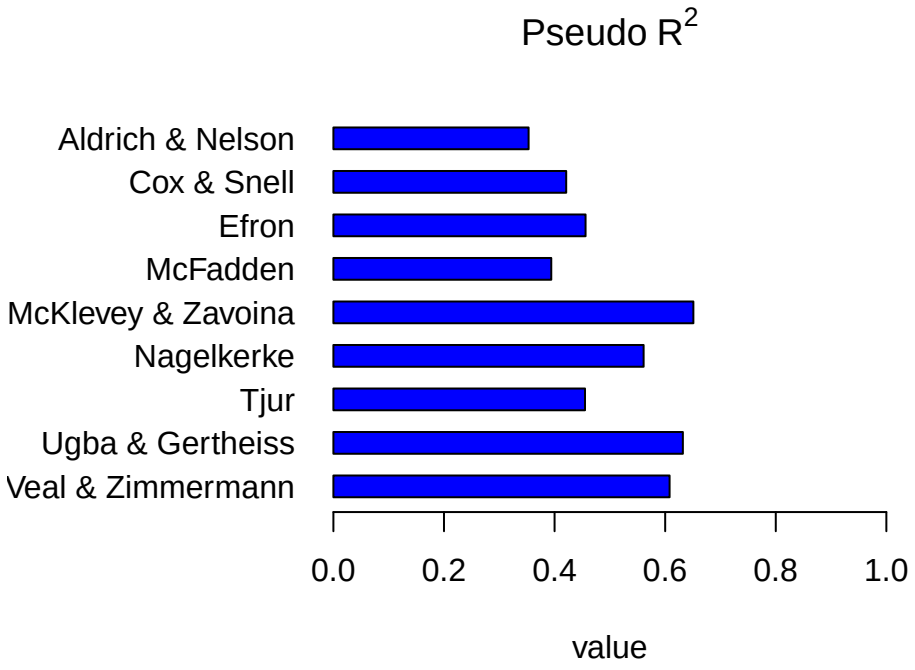

Figure 2. plotgoF display of pseudo-R^2s from an object of class goF.

```
### list of objects
hr <- goFit(mlst, crit = "AIC",display="all")
plotgoF(hr,type = "wt")
```

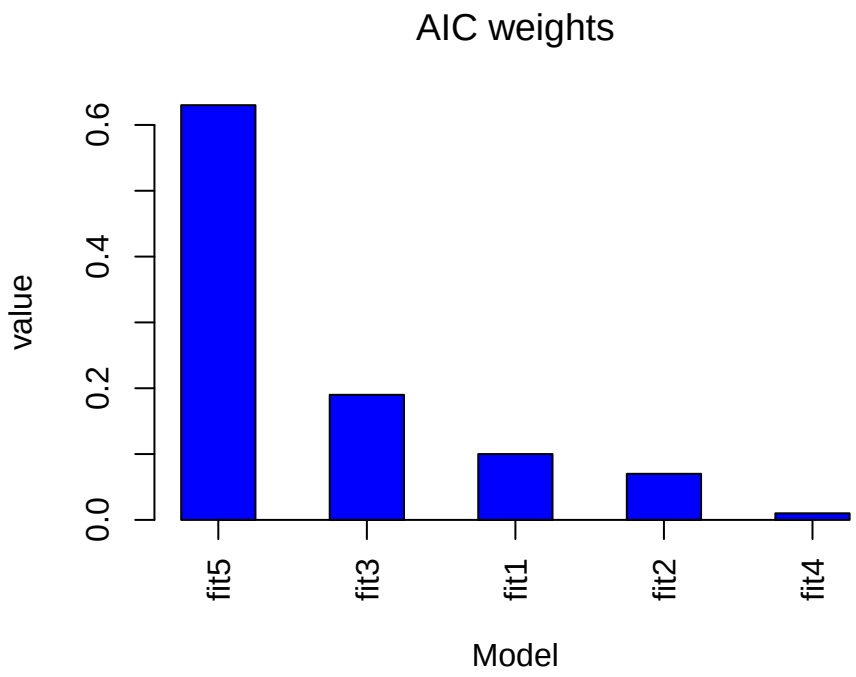

Figure 3. plotgoF display of Akaike and Schwarz weights from an object of class goF.

6 Conclusion & Outlook

The wide acceptance and application of the categorical response models in diverse empirical studies certainly necessitate the need for a thorough and very rational evaluation of such resultant models. In this study, some helpful and applicable routines for such measurement are discussed. To ease the stress and inherent difficulties associated with computing such measures the `CatgoF` package is further developed. This package, as already shown, bundles together some crucial goodness-of-fit measures for categorical outcome models thereby providing a very quick means of assessing their strength and performance. For a start, the `CatgoF` package supports objects from the *glm, vglm, clm2, polr and multinom* methods. These methods are frequently and widely used in many empirical studies utilizing categorical outcome variables. However, the subsequent development of the `CatgoF` package would incorporate some other class of models other than the ones already included, likewise other methods of goodness-of-fit for categorical response models.

References

Akaike, H. (1973). In B. Petrov and F. Caski (Eds.), *Information Theory and the Extension of the Maximum Likelihood Principle*, Second International Symposium on Information theory. Akademia Kiado.

Aldrich J. H., & Nelson F. D., (1984). *Linear probability, logit, and probit models.* Sage University Paper Series on Quantitative Applications in the Social Sciences. London: SAGE Publications.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach.* Springer.

Cox, D. R., & Snell, E.J. (1989) *Analysis of Binary Data*. 2nd ed. London: Chapman & Hall.

Efron, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation. *J. Am Stat. Assoc.*, **73**, 113-121.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *BMA* **76**, 297 – 307.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. California: Sage Publications.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, P. Zarembka (ed.), New York: Academic Press, 105-142.

McKelvey, R. D., & Zavoina, W. (1976). A statistical model for the analysis of ordinal level dependent variables. *J.Math. Sociol.*, **4**, 103–120.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *Am. Stat.*, **54**, 17-24.

Murtaugh, P. A. (2014). In defense of P values. *Ecology*, **95**, 611– 617

Nagelkerke, N. J. D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*, **78**, 691-692.

Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: the coefficient of discrimination. Am. Stat., 63, 366-372.

Tutz, G. (2011). *Regression for Categorical Data*. Cambridge: University Press.

Ugba, E. R. & Gertheiss, J. (2017). A Generalized Likelihood Ratio Index for Discrete Limited Dependent Variable Models — with Emphasis on the Cumulative Link Model. (preprint)

Veall, M.R., & Zimmermann, K.F. (1992). Pseudo-R2's in the ordinal probit model. *J. Math. Sociol.*, **4**, 103-120