



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

The Stata Journal (2017)
17, Number 3, pp. 668–686

How to test for goodness of fit in ordinal logistic regression models

Morten W. Fagerland
Oslo Centre for Biostatistics and Epidemiology
Research Support Services
Oslo University Hospital
Oslo, Norway
morten.fagerland@medisin.uio.no

David W. Hosmer
Department of Mathematics and Statistics
University of Vermont
Burlington, VT

Abstract. Ordinal regression models are used to describe the relationship between an ordered categorical response variable and one or more explanatory variables. Several ordinal logistic models are available in Stata, such as the proportional odds, adjacent-category, and constrained continuation-ratio models. In this article, we present a command (`ologitgof`) that calculates four goodness-of-fit tests for assessing the overall adequacy of these models. These tests include an ordinal version of the Hosmer–Lemeshow test, the Pulkstenis–Robinson chi-squared and deviance tests, and the Lipsitz likelihood-ratio test. Together, these tests can detect several different types of lack of fit, including wrongly specified continuous terms, omission of different types of interaction terms, and an unordered response variable.

Keywords: st0491, `ologitgof`, Hosmer–Lemeshow test, Pulkstenis–Robinson chi-squared and deviance tests, Lipsitz likelihood-ratio test, ordinal models, proportional odds, adjacent category, continuation ratio

1 Background

An ordinal variable is a categorical variable with a natural ordering to the categories, such as level of pain, which is measured as none, mild, moderate, or severe. An ordinal response regression model describes the relationship between an ordinal response variable and one or more explanatory variables (covariates). Ordinal logistic models are of particular interest because of their conceptual similarity to the commonly used binary logistic regression model. One such model—the proportional odds (logistic regression) model—can be fit in Stata with the `ologit` command. Two other logistic models are available via a user-written package by Fagerland (2014): the adjacent-category model (`adjcatlogit`) and the constrained continuation-ratio model (`ccrlogit`). The three models differ in which response categories are compared and how. We choose a model based on which comparisons between responses are most informative for the problem at hand and an assessment of model fit.

For a binary logistic regression model, the Hosmer–Lemeshow (HL) goodness-of-fit test (Hosmer and Lemeshow 1980) can be calculated in Stata by the postestimation command `estat gof`. A generalization of the HL test to multinomial logistic regression models was suggested by Fagerland, Hosmer, and Bofin (2008) and made available through the `mlogitgof` command (Fagerland and Hosmer 2012). For ordinal response models, however, no goodness-of-fit test is available in Stata. Recently, two articles investigated goodness-of-fit tests for proportional odds models (Fagerland and Hosmer 2013) and adjacent-category and continuation-ratio models (Fagerland and Hosmer 2016). Both articles recommend combining three approaches: an ordinal version of the multinomial HL test, the Pulkstenis and Robinson (2004) (PR) tests, and the Lipsitz test (Lipsitz, Fitzmaurice, and Molenberghs 1996).

The purpose of this article is to describe the `ologitgof` command and show how it can be used to test for goodness of fit in proportional odds, adjacent-category, and constrained continuation-ratio models using the HL, PR, and Lipsitz tests.

2 Three ordinal logistic regression models

2.1 Notation

Let Y denote an ordinal response variable with c levels $(1, \dots, c)$, and let

$$\mathbf{x} = (x_1, x_2, \dots, x_p)'$$

be a vector of p explanatory variables (covariates). An ordinal logistic regression model describes the relationship between Y and \mathbf{x} via $c - 1$ logit equations (logits):

$$g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_{c-1}(\mathbf{x})$$

The logits relate a set of intercepts (α s) and regression coefficients (β s) to the probability of the response categories. Let β_k be the regression coefficient of an arbitrary explanatory variable x_k . Then, $\exp(\beta_k)$ can be interpreted as the odds ratio (OR) for a one-unit increase in x_k , comparing two response categories or two sets of response categories, depending on the particular ordinal model used (see sections 2.2–2.4). We write $\text{OR}(2, 1)$ to denote the OR comparing response category 2 with response category 1 and $\text{OR}(3 - 4, 1 - 2)$ to denote the OR comparing response categories 3 and 4 with response categories 1 and 2.

A dataset consisting of n independent observations is denoted by (\mathbf{x}_i, y_i) , $i = 1, \dots, n$. Let $\pi_{ij} = P(Y = j | \mathbf{x}_i)$, $j = 1, \dots, c$, denote the conditional probability of a response equal to category j for observation i given the explanatory variables \mathbf{x}_i . Following model fit, we denote the estimated probabilities by $\hat{\pi}_{ij}$.

2.2 Proportional odds logistic regression

Each of the $c - 1$ logits of the proportional odds model compares the probabilities of two sets of response categories: an equal or smaller response versus a larger response,

$$\begin{aligned} g_j(\mathbf{x}) &= \log \left\{ \frac{P(Y \leq j|\mathbf{x})}{P(Y > j|\mathbf{x})} \right\} \\ &= \alpha_j - \beta' \mathbf{x} \quad j = 1, \dots, c - 1 \end{aligned}$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ is a vector of p regression coefficients. The regression coefficients are constant across the logits. Thus a single coefficient or OR is sufficient to describe the effect of an explanatory variable on the response. We include the negative sign of $\beta' \mathbf{x}$ so that we may interpret a positive value of β_k to mean that as x_k increases, the probability of higher values of Y also increases.

2.3 Adjacent-category logistic regression

The logits of the adjacent-category model compare the probability of each response category (except the first) with the probability of the next larger response category:

$$\begin{aligned} g_j(\mathbf{x}) &= \log \left\{ \frac{P(Y = j + 1|\mathbf{x})}{P(Y = j|\mathbf{x})} \right\} \\ &= \alpha_j + \beta' \mathbf{x} \quad j = 1, \dots, c - 1 \end{aligned}$$

As was the case with the proportional odds model, the regression coefficients are constant across the logits, and the effect of a particular explanatory variable on the response can be described by a single coefficient or OR.

2.4 Continuation-ratio logistic regression

The constrained continuation-ratio model compares the probability of each response with the probability of all higher responses:

$$\begin{aligned} g_j(\mathbf{x}) &= \log \left\{ \frac{P(Y = j|\mathbf{x})}{P(Y > j|\mathbf{x})} \right\} \\ &= \alpha_j - \beta' \mathbf{x} \quad j = 1, \dots, c - 1 \end{aligned} \tag{1}$$

Again we can describe the effect of each explanatory variable using one coefficient or OR.

3 The goodness-of-fit tests

The null hypothesis for the goodness-of-fit tests is that the model fits the data well. The alternative hypothesis is that there is some (unspecific) problem with the fit, which we usually refer to as lack of fit. A small p -value is thus an indication that something is wrong with the model.

3.1 An ordinal version of the HL test

The ordinal HL test (Fagerland and Hosmer 2013, 2016) is based on the multinomial HL test (Fagerland, Hosmer, and Bofin 2008; Fagerland and Hosmer 2012), which in turn is based on the original (binary) HL test (Hosmer and Lemeshow 1980). In all three cases, one groups the observations according to model-predicted response probabilities, usually into $g = 10$ groups. Observed and estimated frequencies for each group in each response category can be tabulated in a $g \times c$ contingency table. The goodness-of-fit test is obtained by calculating the Pearson chi-squared statistic from the table. The binary, multinomial, and ordinal tests differ in the particular grouping strategy used and the number of degrees of freedom for the chi-squared reference distribution. Here we give only the details of the ordinal test.

After fitting the model, calculate the estimated (model-predicted) probabilities $\hat{\pi}_{ij}$ derived from the fit ordinal model. Each observation can now be assigned an ordinal score (OS) (Lipsitz, Fitzmaurice, and Molenberghs 1996):

$$OS_i = \hat{\pi}_{i1} + 2\hat{\pi}_{i2} + \cdots + c\hat{\pi}_{ic} \quad i = 1, \dots, n \quad (2)$$

The OS is the predicted mean score or the “fit” score for each observation. In (2), we have used equally spaced integer scores for the response categories, which is a commonly recommended approach unless there is information about the categories that clearly points to a different set of scores. Another justification for using equally spaced scores is an argument based on the equivalence of a linear combination of cumulative probabilities and (2). See Lipsitz, Fitzmaurice, and Molenberghs (1996) for details.

Sort the observations according to the OS, and create g groups so that group 1 contains the n/g observations with the lowest OSs, group 2 contains the n/g observations with the next lowest score, and so on. Observations with tied OSs are further sorted according to their observed responses (y_i). The observations are allocated to each group so that the group sizes are as similar as possible. Thus observations with equal OSs and equal observed responses may be allocated to different groups if their ranks (after sorting) are on separate sides of the optimal cutoff rank for dividing the groups into similar sizes. Using $g = 10$ groups is recommended (Fagerland and Hosmer 2013, 2016), although any number of groups can be used in principle. If the number of groups is too low, say, below six, the power of the test may be poor because of the heterogeneity within groups. If the number of groups is too big, the contingency table may be sparsely populated, and the distribution of the test statistic may not adhere well to the reference chi-squared distribution.

The ordinal HL test statistic is

$$C_g = \sum_{k=1}^g \sum_{j=1}^c \left(O_{kj} - \hat{E}_{kj} \right)^2 / \hat{E}_{kj}$$

where O_{kj} and \hat{E}_{kj} denote the sums of the observed and estimated frequencies in each group for each response category, respectively,

$$O_{kj} = \sum_{l \in \Omega_k} \tilde{y}_{lj} \quad (3)$$

$$\hat{E}_{kj} = \sum_{l \in \Omega_k} \hat{\pi}_{lj} \quad (4)$$

where \tilde{y}_{ij} is a binary indicator variable with $\tilde{y}_{ij} = 1$ when $y_i = j$ and $\tilde{y}_{ij} = 0$ otherwise, and Ω_k denotes the set of indices of the n/g observations in group k .

As shown in [Fagerland and Hosmer \(2013, 2016\)](#), the distribution of C_g adheres well to the chi-squared distribution with $(g - 2)(c - 1) + (c - 2)$ degrees of freedom under a correctly fit proportional odds, adjacent-category, or constrained continuation-ratio model.

3.2 The PR tests

A simple method of assessing goodness of fit is to calculate the Pearson chi-squared and deviance statistics on the cross-classification of covariate patterns with observed and estimated response frequencies. This strategy works well if the number of covariate patterns is small compared with the number of observations. When the number of covariate patterns is large—for instance, when continuous covariates are present—the estimated frequencies in the cross-classification will be too small for the chi-squared asymptotics to hold. [Pulkstenis and Robinson \(2004\)](#) suggest an approach that starts by grouping the observations according to the covariate patterns using the categorical covariates only. To account for the continuous covariates, one must split each covariate pattern in two based on the median OS (2) within each pattern. The PR test statistics are the Pearson chi-squared and deviance statistics on the contingency table formed from tabulating covariate patterns with response categories

$$\text{PR}(\chi^2) = \sum_{l=1}^2 \sum_{k=1}^K \sum_{j=1}^c \frac{(O_{lkj} - \hat{E}_{lkj})^2}{\hat{E}_{lkj}}$$

and

$$\text{PR}(D^2) = 2 \sum_{l=1}^2 \sum_{k=1}^K \sum_{j=1}^c O_{lkj} \log \frac{O_{lkj}}{\hat{E}_{lkj}}$$

where l indexes the two subgroups based on the OSSs, K is the number of observed covariate patterns because of the categorical covariates, and c is the number of response categories. The sums of the observed and expected frequencies, O_{lkj} and E_{lkj} , are defined as in (3) and (4), only now with an additional partition because of the two subgroups based on the OSSs. The reference distribution for both $\text{PR}(\chi^2)$ and $\text{PR}(D^2)$ is the chi-squared distribution with $(2K - 1)(c - 1) - p_{\text{cat}} - 1$ degrees of freedom, where p_{cat} denotes the number of dichotomous variables needed to model all the categorical covariates (substitute dummy variables for categorical covariates with more than

two categories). This reference distribution adheres well to the distribution of $\text{PR}(\chi^2)$ and $\text{PR}(D^2)$ under both the proportional odds model (Pulkstenis and Robinson 2004; Fagerland and Hosmer 2013) and the adjacent-category and constrained continuation-ratio models (Fagerland and Hosmer 2016).

3.3 The Lipsitz test

To calculate the Lipsitz test, we start by grouping the observations into g groups based on the OS (2), as in section 3.1. Lipsitz, Fitzmaurice, and Molenberghs (1996) suggest that the number of groups is chosen such that $6 \leq g < n/5c$. Next, we define $g - 1$ indicator variables

$$I_{ik} = \begin{cases} 1 & \text{if observation } i \text{ is in group } k \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \dots, n$ and $k = 1, \dots, g - 1$. Define a new ordinal regression model that includes the indicator variables:

$$g_j(\mathbf{x}) = \alpha_j \pm \beta' \mathbf{x} + \gamma_1 I_1 + \dots + \gamma_{g-1} I_{g-1} \quad j = 1, \dots, c - 1$$

The \pm sign is used because we have defined the proportional odds and continuation-ratio models with a minus sign ($\alpha_j - \beta' \mathbf{x}$) and the adjacent-category model with a plus sign ($\alpha_j + \beta' \mathbf{x}$). If the original fit model is the correct model, $\gamma_1, \dots, \gamma_{g-1} = 0$. We can test this proposition by fitting the new model and comparing the log likelihoods of the models with (L_0) and without (L_1) the indicator variables using the likelihood-ratio statistic $-2(L_1 - L_0)$. The observed value of the test statistic can be compared with the chi-squared distribution with $g - 1$ degrees of freedom. This approximation to the distribution of the likelihood-ratio statistic holds for both the proportional odds model (Fagerland and Hosmer 2013) and the adjacent-category and constrained continuation-ratio models (Fagerland and Hosmer 2016).

4 The ologitgof command

4.1 Syntax

```
ologitgof [varlist] [if] [in] [, group(#) all outsample osvar(newvar)
          groupvar(newvar) patternvar(newvar) tableHL tablePR]
```

4.2 Description

`ologitgof` is a postestimation command that calculates the ordinal HL, PR, and Lipsitz goodness-of-fit tests. The command can be used after proportional odds logistic regression (`ologit`), adjacent-category logistic regression (`adjcatlogit`), or constrained continuation-ratio logistic regression (`ccrlogit`). The PR tests will be calculated only if the categorical covariates from the estimation command are specified in *varlist*.

4.3 Options

group(#) specifies the number of quantiles to be used to group the observations (HL and Lipsitz tests). The default is **group(10)**.

all requests that the goodness-of-fit test be computed for all observations in the dataset, ignoring any **if** or **in** qualifier specified with the estimation command.

outsample adjusts the degrees of freedom for the chi-squared reference distribution for samples outside the estimation sample (HL test).

osvar(newvar) generates *newvar* containing the OS.

groupvar(newvar) generates *newvar* containing a group identifier.

patternvar(newvar) generates *newvar* containing a covariate pattern identifier.

tableHL displays a contingency table for the HL test, where the groups form the rows and the columns consist of the cutoff values of the OS, observed and estimated frequencies, and totals for each group.

tablePR displays a contingency table for the PR tests, where the covariate patterns form the rows and the columns consist of the observed and estimated frequencies and totals for each pattern.

4.4 Stored results

ologitgof stores the following in **r()**:

Scalars

r(N)	number of observations
e(k_cat)	number of categories
r(g)	number of groups
r(numpatterns)	number of covariate patterns
r(chi2_HL)	chi-squared statistic; HL test
r(df_HL)	degrees of freedom; HL test
r(P_HL)	probability > chi-squared; HL test
r(chi2_PR)	chi-squared statistic; PR test
r(D2)	deviance statistic; PR test
r(df_PR)	degrees of freedom; PR tests
r(P_chi2)	probability > chi-squared; PR test
r(P_D2)	probability > chi-squared; PR test
r(chi2_L)	chi-squared statistic; Lipsitz test
r(df_L)	degrees of freedom; Lipsitz test
r(P_L)	probability > chi-squared; Lipsitz test

Macros

r(cmd)	ologitgof
r(cmdline)	command as typed
r(title)	title in estimation output
r(ecmd)	ologit , adjcatlogit , or ccrlogit ; estimation command

Matrices

r(cat)	category values
r(HLtable)	entire HL contingency table
r(HLtableOE)	observed and estimated frequencies from the HL contingency table
r(PRtable)	observed and estimated frequencies from the PR contingency table

5 Example

5.1 The Low Birth Weight study

We use the Low Birth Weight study ([Hosmer, Lemeshow, and Sturdivant 2013, 24](#)) to illustrate how to fit the models and assess their goodness of fit. The following Stata commands will load and describe the dataset:

```
. webuse lbw
(Hosmer & Lemeshow data)
. describe
(output omitted)
```

The dataset contains the birthweight of 189 children born at the Baystate Medical Center in Springfield, Massachusetts. Also included in the dataset are 8 potential risk factors for low birthweight—defined as birthweight less than 2,500 grams—including age, mother’s weight, and smoking status during pregnancy. Low birthweight is associated with increased risk of infant mortality and birth defects.

We may regard an ordinal variable to be of one of two distinct types: “grouped continuous”, that is, explicitly derived from the categorization of a continuous variable; and “assessed”, such as symptoms of disease (none, some, severe), which may be related to an underlying continuous variable but measured only on a categorical scale ([Anderson 1984](#)). Here we use the continuous variable `bwt` (birthweight, measured in grams) to form a 4-category ordinal variable `bwt4` using the cutpoints 2,500 grams, 3,000 grams, and 3,500 grams:

```
. gen bwt4 = .
(189 missing values generated)
. replace bwt4 = 1 if bwt > 3500 & bwt != .
(46 real changes made)
. replace bwt4 = 2 if bwt <= 3500 & bwt > 3000 & bwt != .
(46 real changes made)
. replace bwt4 = 3 if bwt <= 3000 & bwt > 2500 & bwt != .
(38 real changes made)
. replace bwt4 = 4 if bwt <= 2500 & bwt != .
(59 real changes made)
. tabulate bwt4
```

bwt4	Freq.	Percent	Cum.
1	46	24.34	24.34
2	46	24.34	48.68
3	38	20.11	68.78
4	59	31.22	100.00
Total	189	100.00	

We choose this coding so that the heaviest births are the reference category (`bwt4 = 1`) and higher responses represent increased risk of an unfavorable outcome. As explanatory variables, we use the following:

- **smoke**: smoking status during pregnancy; 0 = no, 1 = yes
- **lwt**: weight (in pounds) of mother at last menstrual period
- **race**: 1 = white, 2 = black, 3 = other
- **ptl**: history of premature labor; number

The example models in this article should be interpreted as illustrations and not taken as fully developed models for the subject matter problem. For details and model-building strategies for ordinal regression models, we refer the reader to [Agresti \(2010\)](#) and Hosmer, Lemeshow, and Sturdivant (2013).

5.2 The proportional odds model

Fitting the model with ologit

The proportional odds model is available in official Stata through the `ologit` command. Consider the following model:

```
. ologit bwt4 smoke lwt i.race ptl, nolog
```

Ordered logistic regression		Number of obs	=	189
		LR chi2(5)	=	35.52
		Prob > chi2	=	0.0000
Log likelihood = -241.89265		Pseudo R2	=	0.0684

bwt4	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke	1.038034	.3096776	3.35	0.001	.4310766	1.644991
lwt	-.0124006	.0045665	-2.72	0.007	-.0213508	-.0034505
race						
black	1.496324	.4220844	3.55	0.000	.6690534	2.323594
other	.9485401	.3273635	2.90	0.004	.3069195	1.590161
ptl	.4500006	.3121285	1.44	0.149	-.16176	1.061761
/cut1	-1.861848	.6781887			-3.191073	-.5326227
/cut2	-.6196977	.662932			-1.919021	.6796251
/cut3	.3332177	.6596246			-.9596229	1.626058

The coefficient for **smoke** is positive, which indicates that smoking during pregnancy is associated with increased risk of low birthweight, as are being black or other race. Mother's weight (**lwt**), on the other hand, has a negative coefficient, which means that heavier mothers tend to give birth to heavier children. A history of premature labor might increase the risk of low birthweight, although the coefficient for **ptl** is not significantly different from zero.

By giving the `or` option, `ologit` reports ORs instead of coefficients:

<code>. ologit, or</code>					
Ordered logistic regression					
				Number of obs	= 189
				LR chi2(5)	= 35.52
				Prob > chi2	= 0.0000
Log likelihood = -241.89265				Pseudo R2	= 0.0684
bwt4	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
smoke	2.823659	.8744241	3.35	0.001	1.538913 5.180961
lwt	.987676	.0045102	-2.72	0.007	.9788756 .9965555
race					
black	4.465243	1.884709	3.55	0.000	1.952388 10.21231
other	2.581938	.845232	2.90	0.004	1.359232 4.904537
pt1	1.568313	.4895152	1.44	0.149	.8506454 2.891459
/cut1	-1.861848	.6781887			-3.191073 -.5326227
/cut2	-.6196977	.662932			-1.919021 .6796251
/cut3	.3332177	.6596246			-.9596229 1.626058

For smoking during pregnancy, the following interpretations apply:

$$\widehat{\text{OR}}(2-4, 1) = \widehat{\text{OR}}(3-4, 1-2) = \widehat{\text{OR}}(4, 1-3) = 2.82$$

Pregnant smokers have 2.82 times the odds of nonsmokers of giving birth to babies with birthweight below versus above 3,500 grams. The OR is the same for comparing birthweight below versus above 3,000 grams and for comparing birthweight below versus above 2,500 grams.

Testing goodness of fit

We calculate the HL, PR, and Lipsitz tests with a single `ologitgof` command. We may specify the number of groups (for the HL and Lipsitz tests) with the `group(#)` option or use the default $g = 10$. The options `tableHL` and `tablePR` draw up the contingency tables of observed and estimated frequencies for the HL and PR tests, respectively. To calculate the PR tests, `ologitgof` requires that we specify a list of the categorical covariates from the estimation command.

```
. ologitgof smoke race, tableHL tablePR
```

Goodness-of-fit tests for ordinal logistic regression models

Table: observed and estimated frequencies for the HL test

Group	Ordinal score	Obs_1	Est_1	Obs_2	Est_2	Obs_3	Est_3
1	1.9236	8	10.15	6	4.98	4	2.15
2	2.0556	13	8.09	4	5.58	1	2.84
3	2.3723	7	6.49	4	5.68	3	3.44
4	2.5440	6	4.65	5	5.39	3	4.09
5	2.6120	5	4.07	5	5.16	5	4.26
6	2.7121	2	3.73	5	4.97	5	4.34
7	2.7957	2	3.28	2	4.69	6	4.42
8	2.9549	1	2.78	6	4.29	4	4.43
9	3.2344	1	2.10	4	3.60	6	4.28
10	3.6576	1	0.99	5	2.00	1	3.09

Group	Obs_4	Est_4	Total
1	1	1.71	19
2	1	2.49	19
3	5	3.40	19
4	5	4.88	19
5	4	5.51	19
6	7	5.96	19
7	9	6.61	19
8	8	7.50	19
9	8	9.01	19
10	11	11.92	18

Table: observed and estimated frequencies for the PR tests

Covariate pattern	Obs_1	Est_1	Obs_2	Est_2	Obs_3	Est_3
1 <= median OS	10	11.44	7	5.87	4	2.60
1 > median OS	13	8.72	4	6.52	2	3.53
2 <= median OS	2	2.01	2	2.15	1	1.66
2 > median OS	0	0.98	2	1.62	4	1.84
3 <= median OS	7	6.58	8	7.80	7	6.10
3 > median OS	3	5.12	9	7.26	4	6.87
4 <= median OS	9	6.61	3	7.19	7	5.47
4 > median OS	1	3.72	6	5.68	7	5.89
5 <= median OS	0	0.47	1	0.85	1	1.07
5 > median OS	0	0.21	1	0.45	1	0.75
6 <= median OS	1	0.75	4	1.09	0	1.23
6 > median OS	0	0.32	2	0.66	0	1.03

Covariate pattern	Obs_4	Est_4	Total
1 <= median OS	1	2.10	22
1 > median OS	3	3.23	22
2 <= median OS	3	2.17	8
2 > median OS	2	3.56	8
3 <= median OS	6	7.52	25
3 > median OS	14	10.75	30
4 <= median OS	7	6.73	26
4 > median OS	12	10.72	26
5 <= median OS	3	2.61	5
5 > median OS	3	3.59	5
6 <= median OS	1	2.93	6
6 > median OS	4	3.99	6

OS = ordinal score

covpatternlabel:

1 nonsmoker white
 2 nonsmoker black
 3 nonsmoker other
 4 smoker white
 5 smoker black
 6 smoker other

Model: proportional odds (ologit)

Dependent variable: bwt4 = [1, 2, 3, 4]

Number of observations = 189

Tests	Number of groups/patterns	Statistic	df	P-value
Ordinal HL	10	24.714	26	0.5352
PR(chi2)	6	36.528	30	0.1913
PR(deviance)	6	38.026	30	0.1491
Lipsitz	10	13.833	9	0.1284

(HL = Hosmer-Lemeshow; PR = Pulkstenis-Robinson)

The bottom table provides us with the p -values for the tests. None of the tests give any evidence of lack of fit. The observed and estimated frequencies in the HL contingency table agree fairly well except for one big discrepancy in group 2 for response **bwt4** = 1. There are somewhat bigger discrepancies between observed and estimated frequencies in the PR table, although not enough to produce p -values below 0.1. The difference in the total number of observations in the nonsmoker other category (PR table) for OSs below and above the median is due to several observations with equal OS. Note that the actual covariate values corresponding to each covariate pattern are defined below the PR table following the header **covpatternlabel**. If labels are not defined for a categorical covariate, the numerical values for that covariate are shown instead of the labels.

A note on the size of the estimated frequencies

As a general rule, it is often stated that for the chi-squared approximation to hold, all estimated frequencies should be greater than 1 and at least 80% should be greater than 5; see, for instance, [Lipsitz, Fitzmaurice, and Molenberghs \(1996\)](#). In the HL table above, 1 of the \hat{E}_{kj} 's is less than 1 and only 33% are greater than 5. Similarly, in the PR table, 9 of the \hat{E}_{kjl} 's are less than 1 and only 40% are greater than 5. We believe this rule is too strict. The results of the simulation studies in [Fagerland and Hosmer \(2013, 2016\)](#) indicate that the test statistics in this article are well approximated by the proposed chi-squared reference distributions even for small sample sizes such as $n = 100$ and $n = 200$. A better rule might be to make sure that at least 80% of the estimated frequencies are greater than 1, although we hasten to point out that our recommendation is based on a limited number of simulations.

An example of lack of fit

We present an (artificial) example of a poorly fit model. Consider a proportional odds model with `bwt4` as the response variable and `smoke`, `age`, `age2`, and their interactions as covariates:

```
. ologit bwt4 smoke##c.age##c.age, nolog
Ordered logistic regression
```

Number of obs	=	189
LR chi2(5)	=	14.26
Prob > chi2	=	0.0141
Pseudo R2	=	0.0275

```
Log likelihood = -252.52312
```

bwt4	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke						
smoker	-2.726678	5.589375	-0.49	0.626	-13.68165	8.228296
age	.0959844	.22932	0.42	0.676	-.3534744	.5454433
smoke#c.age						
smoker	.2099504	.4668952	0.45	0.653	-.7051475	1.125048
c.age#c.age	-.0033024	.0045721	-0.72	0.470	-.0122635	.0056587
smoke#c.age#c.age						
smoker	-.0024751	.0094211	-0.26	0.793	-.02094	.0159899
/cut1	-.5351304	2.794063			-6.011393	4.941133
/cut2	.6035787	2.791269			-4.867208	6.074365
/cut3	1.481455	2.790778			-3.988369	6.951278

```
. ologitgof smoke
Goodness-of-fit tests for ordinal logistic regression models
Model: proportional odds (ologit)
Dependent variable: bwt4 = [1, 2, 3, 4]
Number of observations = 189
```

Tests	Number of groups/patterns	Statistic	df	P-value
Ordinal HL	10	42.237	26	0.0232
PR(chi2)	2	5.030	7	0.6563
PR(deviance)	2	5.362	7	0.6159
Lipsitz	10	17.766	9	0.0380

(HL = Hosmer-Lemeshow; PR = Pulkstenis-Robinson)

The HL and Lipsitz tests indicate lack of fit, whereas the PR tests do not. This is not surprising, because there are only two categorical covariate patterns (defined by `smoke`) and the model is dominated by continuous terms. The PR tests work best when lack of fit is associated with categorical variables, whereas the HL and Lipsitz tests work best when continuous covariates drive lack of fit ([Fagerland and Hosmer 2013, 2016](#)).

5.3 The adjacent-category model

Fitting the model with `adjcatlogit`

There is no specific command in official Stata that fits adjacent-category models. However, a user-written command `adjcatlogit` was provided by [Fagerland \(2014\)](#). It may be installed by typing

```
. search adjcatlogit
```

and following the on-screen instructions. (This action will also install `ccrlogit`, which is used in section 5.4.) Once `adjcatlogit` is installed, we fit the adjacent-category model of `bwt4` on `smoke`, `lwt`, `race`, and `ptl` by typing

```
. adjcatlogit bwt4 smoke lwt i.race ptl, or
Adjacent-category logistic regression
```

```
Number of obs =    189
LR chi2( 5)    =   36.23
Prob < chi2    =   0.0000
Pseudo R2     =   0.0698
```

```
Log likelihood = -241.53716
```

	bwt4	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
bwt4	smoke	1.702946	.2778756	3.26	0.001	1.236823	2.344739
	lwt	.9929517	.0024911	-2.82	0.005	.9880812	.9978462
	race						
	black	2.289208	.5397237	3.51	0.000	1.44211	3.633891
	other	1.582235	.267944	2.71	0.007	1.135335	2.205047
	ptl	1.189998	.1818554	1.14	0.255	.8819943	1.605561
_anc	cons1	1.76541	.7428428	1.35	0.177	.7738912	4.027273
	cons2	1.247386	.7670759	0.36	0.719	.3737293	4.163365
	cons3	2.005744	1.532941	0.91	0.362	.4484592	8.970734

The effect of **smoke** is contained in one coefficient or OR:

$$\widehat{\text{OR}}(2, 1) = \widehat{\text{OR}}(3, 2) = \widehat{\text{OR}}(4, 3) = 1.70$$

Pregnant women who smoke have 1.70 times the odds of nonsmokers of having a baby with birthweight in the next lower weight category.

Testing goodness of fit

As in section 5.2, we calculate the goodness-of-fit tests using a single `ologitgof` command. For brevity, we omit the contingency tables for the HL and PR tests, though we recommend that these always be displayed in real-life applications.

```
. ologitgof smoke race
Goodness-of-fit tests for ordinal logistic regression models
Model: adjacent-category (adjcatlogit)
Dependent variable: bwt4 = [1, 2, 3, 4]
Number of observations = 189
```

Tests	Number of groups/patterns	Statistic	df	P-value
Ordinal HL	10	24.707	26	0.5356
PR(chi2)	6	33.827	30	0.2878
PR(deviance)	6	35.608	30	0.2212
Lipsitz	10	12.632	9	0.1800

```
(HL = Hosmer-Lemeshow; PR = Pulkstenis-Robinson)
```

We find no evidence of lack of fit for this model.

5.4 The continuation-ratio model

Fitting the model with ccrlogit

To fit a constrained continuation-ratio model in Stata, we use the `ccrlogit` command, which is available with the same package as `adjcatlogit` (Fagerland 2014). If the `adjcatlogit` command is already installed, the `ccrlogit` command was installed with it. If not, type

```
. search ccrlogit
```

and follow the on-screen instructions to install both the `adjcatlogit` and `ccrlogit` commands. We fit the constrained continuation-ratio model in the usual manner:

```
. ccrlogit bwt4 smoke lwt i.race ptl, or
Constrained continuation-ratio logistic regression      Number of obs =      189
                                                         LR chi2( 5)      =    35.69
                                                         Prob < chi2      =    0.0000
Log likelihood = -241.80470                             Pseudo R2       =    0.0687
```

	bwt4	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
bwt4							
	smoke	2.281733	.5644419	3.33	0.001	1.405075	3.705357
	lwt	.9893324	.0038353	-2.77	0.006	.9818439	.996878
	race						
	black	3.943965	1.402531	3.86	0.000	1.964417	7.918311
	other	2.072409	.5353397	2.82	0.005	1.249094	3.438395
	ptl	1.240664	.2755414	0.97	0.332	.8028031	1.917339
_anc							
	cons1	2.284023	1.370738	1.38	0.169	.7044638	7.405293
	cons2	.7863597	.4506991	-0.42	0.675	.2557162	2.418156
	cons3	.7496437	.4189136	-0.52	0.606	.25072	2.241407

Women who smoke during pregnancy have 2.28 times the odds of nonsmokers of having birthweight in any of the next lower weight categories:

$$\widehat{\text{OR}}(2-4, 1) = \widehat{\text{OR}}(3-4, 2) = \widehat{\text{OR}}(4, 3) = 2.28$$

Testing goodness of fit

As in sections 5.2 and 5.3, we calculate the goodness-of-fit tests using a single `ologitgof` command:

```
. ologitgof smoke race
Goodness-of-fit tests for ordinal logistic regression models
Model: constrained continuation-ratio (ccrlogit)
Dependent variable: bwt4 = [1, 2, 3, 4]
Number of observations = 189
```

Tests	Number of groups/patterns	Statistic	df	P-value
Ordinal HL	10	26.647	26	0.4280
PR(chi2)	6	33.421	30	0.3046
PR(deviance)	6	35.571	30	0.2225
Lipsitz	10	16.728	9	0.0531

(HL = Hosmer-Lemeshow; PR = Pulkstenis-Robinson)

The Lipsitz test—unlike the HL and PR tests—suggests there might be problems with the fit for this model. At this point, we should examine the model content as well as the possibility that one of the other ordinal models might better fit the data.

6 Discussion

Evaluating goodness of fit is an important step in the assessment of the adequacy of a regression model. The `ologitgof` command presented in this article can be used to test the goodness of fit of three ordinal logistic regression models: the proportional odds, adjacent-category, and constrained continuation-ratio models. Two of the tests provided with `ologitgof`, the ordinal HL and Lipsitz tests, are best suited to detect lack of fit associated with continuous covariates, whereas the two PR tests work best when lack of fit is related to categorical covariates. Together, the four tests have good power with moderate to large sample sizes to detect several types of lack of fit, including omission of a quadratic term in a continuous covariate, omission of different types of interaction terms, wrong functional form of a continuous covariate, and detection of an unordered response variable (Fagerland and Hosmer 2013, 2016).

Goodness-of-fit tests are tools to detect lack of fit. They are not designed to provide proof that a model is well fit to the data. In that perspective, we recommend that a $p < 0.10$ with any of the tests should lead to further investigation into the nature of the lack of fit, except for large sample sizes, say, $n > 400$, where a 5% significance level can be used. Ideally, tests for goodness of fit should be augmented by casewise diagnostic tools. Unfortunately, casewise diagnostics for ordinal models are not widely available.

This article has considered only constrained ordinal models, in which the regression coefficients are constant across the logit equations for each response category. When this assumption is not realistic—or if none of the constrained models fit the data well—an

unconstrained continuation-ratio model may be fit (`ucrlogit` by Fagerland [2014]). This model is equal to the constrained continuation-ratio model in (1) with the exception that we substitute β_j for β . This model has $c - 1$ regression coefficients for each explanatory variable and thus allows for more flexible models. No goodness-of-fit test for the unconstrained continuation-ratio model currently exists.

A review of ordinal response regression models that go beyond the models considered in this article is given in Ananth and Kleinbaum (1997).

7 References

- Agresti, A. 2010. *Analysis of Ordinal Categorical Data*. 2nd ed. Hoboken, NJ: Wiley.
- Ananth, C. V., and D. G. Kleinbaum. 1997. Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology* 26: 1323–1333.
- Anderson, J. A. 1984. Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B* 46: 1–30.
- Fagerland, M. W. 2014. `adjcatlogit`, `ccrlogit`, and `ucrlogit`: Fitting ordinal logistic regression models. *Stata Journal* 14: 947–964.
- Fagerland, M. W., and D. W. Hosmer. 2012. A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models. *Stata Journal* 12: 447–453.
- . 2013. A goodness-of-fit test for the proportional odds regression model. *Statistics in Medicine* 32: 2235–2249.
- . 2016. Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation* 86: 3398–3418.
- Fagerland, M. W., D. W. Hosmer, and A. M. Bofin. 2008. Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine* 27: 4238–4253.
- Hosmer, D. W., Jr., and S. Lemeshow. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics—Theory and Methods* 9: 1043–1069.
- Hosmer, D. W., Jr., S. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.
- Lipsitz, S. R., G. M. Fitzmaurice, and G. Molenberghs. 1996. Goodness-of-fit tests for ordinal response regression models. *Applied Statistics* 45: 175–190.
- Pulkstenis, E., and T. J. Robinson. 2004. Goodness-of-fit tests for ordinal response regression models. *Statistics in Medicine* 23: 999–1014.

About the authors

Morten W. Fagerland is Head of the Section for Biostatistics, Epidemiology, and Health Economics at the Oslo Centre for Biostatistics and Epidemiology. His research interests include the application of statistical methods in medical research, analysis of categorical data and contingency tables, and comparisons of statistical methods using Monte Carlo simulations. He is one of the authors of the recently published book *Statistical Analysis of Contingency Tables*.

David W. Hosmer is a professor (emeritus) of biostatistics at the University of Massachusetts–Amherst and an adjunct professor of statistics at the University of Vermont. He is a coauthor of *Applied Logistic Regression*, of which a third edition has recently been published, and *Applied Survival Analysis*. His current research includes nonlogit link modeling of binary data and studying factors related to mortality following hospitalization for a gunshot wound and admission to an emergency department for a concussion.