Tests for goodness of fit in ordinal logistic regression models

Morten W. Fagerland\textsuperscript{a} and David W. Hosmer\textsuperscript{b}

\textsuperscript{a}Oslo Centre for Biostatistics and Epidemiology, Research Support Services, Oslo University Hospital, Oslo, Norway; \textsuperscript{b}Department of Public Health, University of Massachusetts, Amherst, MA, USA

**ABSTRACT**

We examine three approaches for testing goodness of fit in ordinal logistic regression models: an ordinal version of the Hosmer–Lemeshow test \(C_g\), the Lipsitz test, and the Pulkstenis–Robinson (PR) tests. The properties of these tests have previously been investigated for the proportional odds model. Here, we extend the tests to two other commonly used models: the adjacent-category and the constrained continuation-ratio models. We use a simulation study to assess null distributions and power. All three tests work well and can detect several types of lack of fit under both the adjacent-category and constrained continuation-ratio models. The \(C_g\) and Lipsitz tests are best suited to detect lack of fit associated with continuous covariates, whereas the PR tests excel at detecting lack of fit associated with categorical covariates. We illustrate the use of the tests with data from a study of aftercare placement of psychiatrically hospitalized adolescents. Based on the results here and previous research, we can make a joint recommendation for testing goodness of fit in proportional odds, adjacent-category, and constrained continuation-ratio logistic regression models: because the tests may detect different types of lack of fit, a thorough assessment of goodness of fit requires use of all three approaches.

**ARTICLE HISTORY**

Received 27 April 2015
Accepted 17 February 2016

**KEYWORDS**

Ordinal models; ordinal regression; adjacent-category; continuation-ratio; goodness of fit; Hosmer–Lemeshow test; calibration; model fit; model testing

**2010 MATHEMATICS SUBJECT CLASSIFICATION CODES**

62J12; 62J20; 62P10; 62P15; 62P25

**1. Introduction**

Several regression models can be used to describe the relationship between an ordinal response variable and one or more explanatory variables.[1,2] The models differ in which and how the response levels are compared. The proportional odds logistic regression model – perhaps the most frequently used ordinal model – compares the probability of an equal or smaller response with the probability of a larger response. The model is constrained by the requirement that the log odds of each explanatory variable does not depend on the response category. Two other constrained models are the adjacent-category and the constrained continuation-ratio models. The adjacent-category model compares the probability of each response with the probability of the next lower response. The constrained continuation-ratio model compares the probability of each response with the probability of all higher responses. We may also use an unconstrained continuation-ratio model, which provides greater flexibility by allowing the effect of each explanatory variable to depend on the response level. The choice between these models should be guided by which comparisons between response levels are most informative for the problem at hand as well as an assessment of model adequacy.
One important step in the assessment of model adequacy is an evaluation of overall goodness of fit. Fagerland and Hosmer [3] considered the proportional odds model and investigated the performances of three methods to test for goodness of fit. One of these methods was based on a generalization of the Hosmer–Lemeshow goodness-of-fit test [4] to multinomial logistic regression. [5,6] The other methods were a likelihood ratio test by Lipsitz et al., [7] partly based on the Hosmer–Lemeshow approach for grouping observations, and an extension of the Pearson $\chi^2$ and deviance statistics to allow for continuous covariates. [8] The recommendation in [3] was to use all three methods to assess goodness of fit.

All three methods are based on a strategy of grouping the observations according to estimated probabilities following model fit and are not based on a particular ordinal model. The derivations of the test statistics are only dependent on an ordinal response variable and a model that can estimate probabilities for each response level. Because different ordinal models impose different structures on the data, it may be that the distributions of the test statistics, both under the null and alternative hypotheses, also differ between the models. So far, goodness-of-fit tests based on the three methods above have only been investigated under the proportional odds model. The purpose of this article is to examine whether these tests can also be used following the fit of adjacent-category and continuation-ratio models. We aim to establish sufficient empirical evidence so that solid recommendations for practical use of goodness-of-fit tests in these models are possible.

Section 2 sets the notation and defines the ordinal models. We derive the test statistics in Section 3. The null distribution (Section 4) and power (Section 5) of the tests are investigated using stochastic simulations. We set the tests to work on data from a study of aftercare placement for psychiatrically hospitalized adolescents in Section 6. We discuss the findings and provide recommendations in Section 7.

2. Four ordinal regression models

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

$$x = (x_1, x_2, \ldots, x_p)'$$

denote a vector of $p$ explanatory variables. The relationship between the response variable and the explanatory variables is described through $c-1$ logits,

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

which relate a set of intercepts ($\alpha$s) and regression coefficients ($\beta$s) to the probability of the response levels. We may interpret $\exp(\beta)$ as the odds ratio for a one unit increase in a given explanatory variable, comparing two response levels or two sets of response levels, depending on the model-specific logit equations. We use the notation OR(2, 1) to denote the odds ratio comparing response level 2 with response level 1. In like manner, OR(4–5, 1–3) denotes the odds ratio comparing response levels 4 and 5 with response levels 1–3.

Assume that we have a sample of $n$ independent observations, denoted by $(x_i, y_i)$, $i = 1, \ldots, n$. Let $\pi_{ij} = \Pr(Y = j \mid x_i)$ denote the conditional probability of a response equal to $j$ for observation $i$ given $x_i$. After the fit of the model, we denote the estimated probabilities by $\hat{\pi}_{ij}$.

In the following, we consider three constrained models and one unconstrained model that differ in which and how the response levels are compared. The constrained models impose more structure on the data than the unconstrained model. They are more parsimonious and simpler to interpret, having a single regression coefficient – when compared with $c-1$ regression coefficients for the unconstrained model – for each explanatory variable.
2.1. The proportional odds model

The proportional odds model, also called the constrained cumulative logit model, compares the probability of an equal or smaller response with the probability of a larger response:

\[
g_j(x) = \log \left[ \frac{\Pr(Y \leq j \mid x)}{\Pr(Y > j \mid x)} \right] = \alpha_j - \beta'x, \quad j = 1, \ldots, c - 1, \tag{1}
\]

where \( \beta = (\beta_1, \beta_2, \ldots, \beta_p)' \) is a vector of \( p \) regression coefficients. The log odds does not depend on the response level, and the regression coefficients \( \beta_1, \ldots, \beta_p \) are constant across the logits. The intercepts satisfy the relationship \( \alpha_1 < \alpha_2 < \cdots < \alpha_{c-1} \). The negative sign of \( \beta'x \) is included to allow for the usual interpretation that a positive value of \( \beta_k \) means that as \( x_k \) increases, the probability of higher values of \( Y \) also increases. This is consistent with the way most statistical software packages implement the proportional odds model. For a four-level response variable, the following interpretations apply:

\[
e^{\beta_k} = \text{OR}(2-4, 1) = \text{OR}(3-4, 1-2) = \text{OR}(4, 1-3),
\]

where \( \beta_k \) is the coefficient of the \( k \)th explanatory variable. If we want to estimate the odds ratio of a smaller response versus a larger response, we can change the sign of the coefficient:

\[
e^{-\beta_k} = \text{OR}(1, 2-4) = \text{OR}(1-2, 3-4) = \text{OR}(1-3, 4). \]

The conditional probability of each response level given \( x \) is derived from Equation (1) by using the equality \( \Pr(Y > j \mid x) = 1 - \Pr(Y \leq j \mid x) \) to obtain

\[
\Pr(Y \leq j \mid x) = \frac{e^{g_j(x)}}{1 + e^{g_j(x)}}. \tag{2}
\]

It follows from Equation (2) that

\[
\Pr(Y = 1 \mid x) = \Pr(Y \leq 1 \mid x), \quad \Pr(Y = j \mid x) = \Pr(Y \leq j \mid x) - \Pr(Y \leq j - 1 \mid x), \quad j = 2, \ldots, c - 1,
\]

and

\[
\Pr(Y = c \mid x) = 1 - \Pr(Y \leq c - 1 \mid x).
\]

2.2. The adjacent-category model

The (constrained) adjacent-category model compares the probability of each response with the probability of the next lower response:

\[
g_j(x) = \log \left[ \frac{\Pr(Y = j + 1 \mid x)}{\Pr(Y = j \mid x)} \right] = \alpha_j + \beta'x, \quad j = 1, \ldots, c - 1. \tag{3}
\]

The regression coefficients \( \beta_1, \ldots, \beta_p \) are constant across the logits, whereas the intercepts are not. As was the case with the proportional odds model, the effect of a particular explanatory variable on the
response can be described by a single coefficient or odds ratio. With a four-level response variable, we can make the following interpretations for the $k$th explanatory variable:

$$e^{\beta_k} = \text{OR}(2, 1) = \text{OR}(3, 2) = \text{OR}(4, 3),$$

and

$$\log[\text{OR}(3, 1)] = 2 \cdot \log[\text{OR}(2, 1)],$$
$$\log[\text{OR}(4, 1)] = 3 \cdot \log[\text{OR}(2, 1)].$$

The conditional probability of each response level given $x$ is derived and shown in Appendix 1.

An adjacent-category model may be fitted via a constrained multinomial model. A rationale is given in [1, p.290–291]. The necessary constraints are $\beta_{jk} = (j - 1) \cdot \beta_{2k}$, for $j = 3, \ldots, c$ and $k = 1, \ldots, p$, where $\beta_{jk}$ is the coefficient of the multinomial logit comparing $Y = j$ with $Y = 1$ (the reference category) for the $k$th explanatory variable. The estimated probabilities, $\hat{\pi}_{ij}$, then derive from the multinomial model. A command for fitting adjacent-category models in Stata has recently been made available.[9]

### 2.3. Continuation-ratio models

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

$$g_j(x) = \log \left[ \frac{\Pr(Y = j \mid x)}{\Pr(Y > j \mid x)} \right] = \alpha_j - \beta'_j x, \quad j = 1, \ldots, c - 1. \quad (4)$$

Again, the effect of a particular explanatory variable on the response can be described by a single coefficient or odds ratio. The interpretations for a four-level response variable are

$$e^{\beta_k} = \text{OR}(2–4, 1) = \text{OR}(3–4, 2) = \text{OR}(4, 3).$$

The conditional probability of each response level given $x$ is derived and shown in Appendix 2. The continuation-ratio model is most useful in situations where $Y$ measures the number of attempts to attain a binary outcome.[1, p.296–297] Also called a sequential process,[2, p.96–97] each subject has to pass through each response level in turn, starting with the lowest one, to reach a higher response level. Constrained continuation-ratio models can be fitted with a generalized linear model formulation. Two commands for fitting constrained continuation-ratio models are available in Stata.[9, 10]

We may modify the model in Equation (4) such that the regression coefficients depend on the logits:

$$g_j(x) = \log \left[ \frac{\Pr(Y = j \mid x)}{\Pr(Y > j \mid x)} \right] = \alpha_j - \beta'_j x, \quad j = 1, \ldots, c - 1, \quad (5)$$

where $\beta_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})'$ for $j = 1, \ldots, c - 1$. The model in Equation (5) is called the unconstrained continuation-ratio model. The effect of a particular explanatory variable on the response depends on the response level and is described by $c-1$ regression coefficients or odds ratios. The
unconstrained continuation-ratio model may be fitted using \( c - 1 \) binary logistic regression models, where the binary response variables are defined as

\[
Y^*_j = \begin{cases} 
1 & \text{if } Y = j \\
0 & \text{if } Y > j \\
\text{missing} & \text{if } Y < j
\end{cases}
\] (6)

for \( j = 1, \ldots, c - 1 \). A command for fitting unconstrained continuation-ratio models in Stata is available.[9] The unconstrained continuation-ratio model has the same number of parameters as the multinomial logistic regression model; however, assessing the fit of unconstrained continuation-ratio models is a more complicated task due to the complexities introduced by the separate binary fits. Here, we do not consider the unconstrained continuation-ratio model further and leave that topic for a future article.

3. Derivation of test statistics

In this section, we present three approaches for testing goodness of fit in ordinal logistic regression models. The test statistics resulting from these approaches are not new. Fagerland and Hosmer [3] showed that these test statistics have satisfactory properties for testing goodness of fit in proportional odds logistic regression models; however, their null distribution and power properties have not yet been investigated for the adjacent-category and constrained continuation-ratio models.

3.1. An ordinal version of the Hosmer–Lemeshow test

We consider a test based on the Hosmer–Lemeshow test [4] for binary logistic regression. A generalization of the Hosmer–Lemeshow test to multinomial logistic regression was suggested by Fagerland et al. [5] and recently adapted to the proportional odds model.[3] Here, we make use of the same test statistic as in [3] to derive tests for use with the adjacent-category and the constrained continuation-ratio models.

Following the fit of the model, calculate the estimated probabilities \( \hat{\pi}_{ij} \). Assign an ordinal score \( s_i \) to each observation:

\[
s_i = \hat{\pi}_{i1} + 2\hat{\pi}_{i2} + \cdots + c\hat{\pi}_{ic}, \quad i = 1, \ldots, n.
\] (7)

Sort the observations according to the ordinal score and arrange them into \( g \) groups, such that group 1 contains the \( n/g \) observations with the lowest ordinal scores and group \( g \) contains the observations with the highest ordinal scores. In theory, any number of groups can be used, although \( g = 10 \) is commonly used for the binary Hosmer–Lemeshow test (the deciles of risk). Denote the sums of the observed and estimated frequencies in each group for each of the \( c \) response levels by \( O_{kj} \) and \( E_{kj} \), respectively. The values of \( O_{kj} \) and \( E_{kj} \) can be summarized in a \( g \times c \) contingency table, as shown in Table 1. The ordinal version of the Hosmer–Lemeshow test statistic is the Pearson \( \chi^2 \) statistic given by

\[
C_g = \sum_{k=1}^{g} \sum_{j=1}^{c} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}.
\] (8)

As shown in [3], the distribution of \( C_g \) is well approximated by the \( \chi^2 \) distribution with \((g - 2)(c - 1) + (c - 2)\) degrees of freedom under a correctly fitted proportional odds model. \( C_g \) was originally derived for use after fitting a multinomial logistic regression model.[5] The degrees of freedom for the multinomial test is \((g - 2)(c - 1)\). This difference in degrees of freedom reflects the reduced number of regression coefficients of the proportional odds model when compared with the multinomial model. The adjacent-category and the constrained continuation-ratio models have the same
Table 1. Observed \((O_{kj})\) and estimated \((E_{kj})\) frequencies sorted and summed into \(g\) groups.

| Group | \(Y = 1\) | \(Y = 2\) | ... | \(Y = c\) | \(Y = c\) |
|-------|-----------|-----------|-----|-----------|---------|
|       | Obs.      | Est.      | ... | Obs.      | Est.    | Sum   |
| 1     | \(O_{11}\) | \(E_{11}\) | ... | \(O_{1c}\) | \(E_{1c}\) | n/g   |
| 2     | \(O_{21}\) | \(E_{21}\) | ... | \(O_{2c}\) | \(E_{2c}\) | n/g   |
| ...   | ...       | ...       | ... | ...       | ...     | ...   |
| \(g\) | \(O_{g1}\) | \(E_{g1}\) | ... | \(O_{gc}\) | \(E_{gc}\) | n/g   |

Note: Obs, observed frequency; Est, estimated frequency.

The number of regression coefficients as the proportional odds model. We thereby posit that the degrees of freedom when we use the adjacent-category and the constrained continuation-ratio models are also equal to \((g - 2)(c - 1) + (c - 2)\).

### 3.2. The Lipsitz test

The test by Lipsitz et al. [7] is partly based on the Hosmer–Lemeshow approach and is suitable for a general ordinal regression model. As in Section 3.1, we sort the observations according to the ordinal score (Equation (7)) and partition them into \(g\) groups. Define \(g - 1\) binary indicator variables \(I_k\) such that

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

for \(i = 1, \ldots, n\) and \(k = 1, \ldots, g - 1\). Fit a new ordinal regression model that includes the indicator variables:

\[
g_j(x) = \alpha_j - \beta'x + \gamma_1 I_1 + \cdots + \gamma_{g-1} I_{g-2}, \quad j = 1, \ldots, c - 1.
\]  

(9)

When the fitted model is the correct model, \(\gamma_1, \ldots, \gamma_{g-1} = 0\). Denote the log likelihoods of the fitted models with and without the indicator variables by \(L_0\) and \(L_1\), respectively. The Lipsitz test statistic is the likelihood ratio statistic \(-2(L_1 - L_0)\). A \(P\)-value is obtained by comparing the observed value of the test statistic with the \(\chi^2\) distribution with \(g - 1\) degrees of freedom. To make sure all expected counts are greater than 1.0, and at least 80% greater than 5.0, Lipsitz et al. [7] propose that \(g\) is chosen such that \(6 \leq g < n/5c\).

### 3.3. The Pulkstenis and Robinson tests

The Pearson \(\chi^2\) and deviance statistics can be used to assess goodness of fit in models where the number of possible covariate patterns is limited. We can then construct a contingency table of observed and estimated frequencies, where the covariate patterns form the rows and the response levels form the columns. When the number of covariate patterns is large, such as when continuous covariates are present, the expected cell counts become too small for the \(\chi^2\) asymptotics to hold.

Pulkstenis and Robinson [8] suggest a grouping of covariate patterns that can be used when continuous covariates are present. Start by determining the covariate patterns using only the categorical covariates (ignore any unobserved patterns). Calculate the ordinal scores (Equation (7)). Split each covariate pattern in two based on the median ordinal score within each pattern. We can now construct a contingency table of observed and estimated frequencies based on the cross-classification of covariate patterns with response levels. The two Pulkstenis and Robinson (PR) test statistics are the
Pearson $\chi^2$ and deviance test statistics on that table:

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

and

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

where $l$ indexes the two subgroups based on the ordinal scores, $K$ is the number of observed covariate patterns due to the categorical covariates, and $c$ is the number of response levels. Let $p_{\text{cat}}$ denote the number of dichotomous variables needed to model all the categorical covariates (i.e. after dummy variables have been substituted for categorical covariates with more than two categories). The reference distribution for both $\text{PR}(\chi^2)$ and $\text{PR}(D^2)$ is the $\chi^2$ distribution with $(2K - 1)(c - 1) - p_{\text{cat}} - 1$ degrees of freedom.

4. Assessing the null distribution of the test statistics

We set up a simulation study to investigate the null distribution of the test statistics. For each ordinal model, we considered scenarios with three, four, and five response levels. Each scenario is characterized by particular values of the intercepts ($\alpha$s) and regression coefficients ($\beta$s), the distribution of the covariates, and the sample size ($n$). Let $x$ denote a continuous covariate, distributed as $U(0, 10)$ (uniform distribution on the interval $[0, 10]$) or $N(5, 3)$ (normal distribution with $\mu = 5$ and $\sigma = 3$). Let $d$ denote a dichotomous covariate distributed as Bernoulli(0.5). We used four sample sizes: $n = 100, 200, 400, 800$. The intercepts and regression coefficients were selected to represent scenarios that one might encounter in practice. We assessed this by plotting the probability distributions of the response levels as functions of the continuous covariate, stratified by the dichotomous covariate. The plots for the final models that were used in the simulation study are shown as Web Figures 1–6 in the supplemental online material (available for downloading at the journal website). The number of simulated replications was 10,000 for all scenarios. We used Stata 12.1 (StatCorp LP, College Station, TX) to fit adjacent-category models with constrained multinomial logistic regression. Constrained continuation-ratio models were fitted with the Stata package `sg86` by Wolfe.[10] We used Matlab R2013a (The Mathworks, Inc., Natick, MA) to perform the simulations.

4.1. Adjacent-category model

We defined three adjacent-category models, one for three response levels (Equation (12)), one for four response levels (Equation (13)), and one for five response levels (Equation (14)):

$$g_j(x) = \alpha_j - 0.15x + 0.5d, \quad \alpha_j = [0.5, 0.25], \quad j = 1, 2,$$

$$g_j(x) = \alpha_j - 0.25x + 0.75d, \quad \alpha_j = [1.25, 1.0, 0.5], \quad j = 1, 2, 3,$$

$$g_j(x) = \alpha_j - 0.2x + 0.6d, \quad \alpha_j = [0.8, 0.8, 0.7, 0.5], \quad j = 1, 2, 3, 4.$$
Table 2. Simulated rejection rates (%) averaged over two continuous distributions (uniform and normal) and three adjacent-category models (Equations (12)–(14)).

| Significance level | n = 100 | n = 200 | n = 400 | n = 800 |
|--------------------|---------|---------|---------|---------|
|                    | 1%      | 5%      | 10%     | 1%      | 5%      | 10%     | 1%      | 5%      | 10%     | 1%      | 5%      | 10%     |
| C₈                 | 1.33    | 5.12    | 9.84    | 1.04    | 4.96    | 9.95    | 1.01    | 5.01    | 9.89    | 1.02    | 4.94    | 9.94    |
| C₁₀                | 1.42    | 5.27    | 9.87    | 1.10    | 5.02    | 9.97    | 1.08    | 4.83    | 9.70    | 1.03    | 4.97    | 9.97    |
| C₁₂                | 1.53    | 5.46    | 10.0    | 1.22    | 5.21    | 9.93    | 1.11    | 4.97    | 9.89    | 1.04    | 5.07    | 10.1    |
| Lipsitz            | 1.50    | 6.58    | 12.4    | 1.36    | 6.28    | 11.8    | 1.19    | 5.58    | 10.8    | 1.10    | 5.27    | 10.4    |
| PR(χ²)             | 1.03    | 5.61    | 11.1    | 1.11    | 5.42    | 11.0    | 1.10    | 5.50    | 11.0    | 1.08    | 5.44    | 10.8    |
| PR(D²)             | 1.67    | 7.69    | 14.5    | 1.36    | 6.76    | 11.8    | 1.30    | 6.25    | 12.2    | 1.17    | 5.69    | 11.2    |

Figure 1. Histograms of test statistics and reference distribution curves for the adjacent-category model in Equation (13) with \( n = 400 \) and \( x \sim N(5, 3) \).

Table 2, which is an average of the results in Web Tables 1–3, stratified by sample size. The rejection rates of the \( C_g \) tests are quite close to the nominal level, whereas the rejection rates of the Lipsitz and PR tests are generally above the nominal level, particularly for small sample sizes. Figure 1 shows an example of the distribution of test statistics and their adherence to the reference distributions. Details of the tail sections of the histograms are given in Figure 2. Figures 1 and 2 indicate excellent fit for the \( C_g \) tests and slightly heavy tails for the Lipsitz and PR tests.

An additional simulation was carried out to assess the performance of the tests on an adjacent-category model with three continuous covariates and no categorical covariates. The number of response levels was four:

\[
g_j(x) = \alpha_j + 0.1x_1 - 0.25x_2 + 0.2x_3, \quad \alpha_j = [-0.5, -0.2, 0.4], \quad j = 1, 2, 3, \quad (15)
\]

where \( x_1, x_2, \) and \( x_3 \) are continuous covariates, distributed as \( U(0, 10) \). We give the full results in Web Table 4 and discuss the results here only briefly. Compared with the results for the models in Equations (12)–(14), the 1% and 5% rejection rates of the \( C_g \) tests for the model in Equation (15) are slightly higher, especially for \( n = 100 \) and \( n = 200 \). For \( n = 400 \) and \( n = 800 \), the 5% and 10% rejection rates are quite close to the nominal level. The rejection rates of the Lipsitz test are generally too high, also for large sample sizes, with 5% rejection rates of 5.98 (\( n = 400 \)) and 5.73 (\( n = 800 \)). No simulations were done for the two PR tests, because the PR tests can only be used when both continuous and categorical covariates are present.
Figure 2. Tail sections of the histograms and reference distribution curves in Figure 1.

Table 3. Simulated rejection rates (%) averaged over two continuous distributions (uniform and normal) and three constrained continuation-ratio models (Equations (16)–(18)).

| Significance level | n = 100 | n = 200 | n = 400 | n = 800 |
|--------------------|---------|---------|---------|---------|
|                    | 1% | 5% | 10% | 1% | 5% | 10% | 1% | 5% | 10% | 1% | 5% | 10% |
| Cₖ                 | 0.89 | 4.60 | 9.48  | 0.91 | 4.96 | 10.0 | 0.98 | 4.83 | 9.67  | 0.99 | 4.95 | 9.91 |
| C₉₀               | 0.90 | 4.58 | 9.52  | 0.92 | 4.89 | 9.98  | 0.90 | 4.79 | 9.76  | 0.97 | 4.83 | 9.88 |
| C₁₂               | 0.83 | 4.52 | 9.40  | 0.93 | 4.80 | 9.67  | 0.91 | 4.87 | 9.65  | 0.95 | 4.77 | 9.66 |
| Lipsitz           | 1.54 | 6.64 | 12.5  | 1.41 | 6.49 | 12.4  | 1.19 | 5.70 | 11.0  | 1.15 | 5.38 | 10.6 |
| PR(χ²)            | 0.99 | 5.55 | 11.5  | 1.09 | 5.73 | 11.4  | 1.12 | 5.72 | 11.2  | 1.22 | 5.79 | 11.4 |
| PR(D²)            | 1.90 | 8.60 | 15.6  | 1.59 | 6.82 | 13.0  | 1.28 | 6.16 | 11.9  | 1.30 | 6.00 | 11.7 |

4.2. Constrained continuation-ratio model

We defined three constrained continuation-ratio models, one for three response levels (Equation (16)), one for four response levels (Equation (17)), and one for five response levels (Equation (18)):

\[
g_j(x) = \alpha_j + 0.2x - 0.5d, \quad \alpha_j = [-2.0, -1.0], \quad j = 1, 2, \quad (16)
\]

\[
g_j(x) = \alpha_j + 0.1x - 0.3d, \quad \alpha_j = [-2.0, -2.0, -0.5], \quad j = 1, 2, 3, \quad (17)
\]

\[
g_j(x) = \alpha_j + 0.25x - 0.4d, \quad \alpha_j = [-2.0, -2.0, -2.0, -1.0], \quad j = 1, 2, 3, 4. \quad (18)
\]

Web Figures 4–6 show the probability distributions for each of the three models as functions of the covariates. We present the full results of the simulations in Web Tables 5–7.

The results of the null distribution simulations for the constrained continuation-ratio model were similar to the ones for the adjacent-category model (Section 4.1). A summary is given in Table 3, which is an average of the results in Web Tables 5–7. The rejection rates of the \( C_k \) tests were slightly lower than the nominal level, but not by much. As in Section 4.1, the rejection rates of the Lipsitz and the PR tests were above the nominal level. Figure 3 shows histograms of the six test statistics for one of the scenarios with four response levels (Equation (17)), with details of the tail sections shown in Figure 4. Figures 3 and 4 give the same impression as Figures 1 and 2: that of a very good fit for the \( C_k \) tests and somewhat heavy tails for the Lipsitz and PR tests.

An additional simulation was carried out to assess the performance of the tests on a constrained continuation-ratio model with three continuous covariates and no categorical covariates. The number
of response levels was four:

\[ g_j(x) = \alpha_j - 0.2x_1 - 0.1x_2 + 0.2x_3, \quad \alpha_j = [-1.0, -0.5, 1.25], \quad j = 1, 2, 3, \]  

(19)

where \( x_1, x_2, \) and \( x_3 \) are continuous covariates, distributed as \( U(0, 10) \). We give the full results in Web Table 8 and discuss the results here only briefly. The rejection rates of the \( C_g \) and Lipsitz tests are quite similar to those for the models in Equation (16)–(18), with slightly lower rejection rates for the \( C_g \) tests and higher rejection rates for the Lipsitz test, compared with the nominal level. No simulations were done for the two PR tests, because the PR tests can only be used when both continuous and categorical covariates are present.
5. Assessing the power of the tests

The previous section showed that while the rejection rates of the $C_g$ tests were close to the nominal level, the rejection rates of the Lipsitz and PR tests were, in general, above the nominal level. We thereby expect the power of the Lipsitz and PR tests to be greater than that of the $C_g$ test. One might argue that it is unreasonable to compare the power of tests that do not have similar null distribution approximations; however, in the case of goodness-of-fit tests, we are not too concerned about small inflations of the significance level. A small increase in the probability of false positive results from goodness-of-fit tests merely means that the occasional model is investigated for possible lack of fit in more detail than necessary. Goodness-of-fit tests are tools to detect lack of fit, and as such, their power properties are more important than strict control of type I error rates.

Sections 5.1–5.8 describe and report simulations used to estimate the power of the tests to detect several different types of lack of fit. All scenarios used a four-level response variable. Let $x$ be a continuous covariate, distributed as $U(0,10)$, and let $d$ be a dichotomous covariate, distributed as Bernoulli(0.5). As in Section 4, we used four sample sizes: $n = 100,200,400,800$. The number of replications was 5000.

5.1. Omission of a quadratic term in a continuous covariate

For the adjacent-category model, the correct model was defined as

$$g_j(x) = \alpha_j - 0.25x - 0.75d + \beta_3x^2, \quad \alpha_j = [1.5, 1.0, -0.5], \quad j = 1, 2, 3, \quad (20)$$

and for the constrained continuation-ratio model, the correct model was

$$g_j(x) = \alpha_j + 0.25x + 0.5d - \beta_3x^2, \quad \alpha_j = [-2.0, -1.0, -1.0], \quad j = 1, 2, 3. \quad (21)$$

In both cases, the fitted model excluded the quadratic term ($\beta_3 = 0$). The coefficient of the quadratic term was $\beta_3 = 0.02, 0.04, 0.06$, corresponding to three scenarios with increasing difference between the correct and fitted models. Web Figures 7 and 8 show the probability distributions of the responses as functions of the covariates using $\beta_3 = 0.04$ and the models in Equations (20) and (21), respectively. We give the full results of the simulations in Web Tables 9 and 10. An excerpt is shown in Table 4.

The $C_g$ tests had low to moderate power to detect a missing quadratic term. The power was greater for the adjacent-category model than the constrained continuation-ratio model; however, that could be attributed to differences in the definitions of the actual models (i.e. the values of the parameters and their relation to the response) used in the simulations (Equations (20) and (21)). The power of the Lipsitz test was noticeably greater than that of the $C_g$ tests, whereas the PR tests had poor power.

5.2. Omission of an interaction term between a continuous and a dichotomous covariate

For the adjacent-category model, the correct model was defined as

$$g_j(x) = \alpha_j - 0.15x - 0.6d + \beta_3xd, \quad \alpha_j = [1.5, 1.25, 0.75], \quad j = 1, 2, 3, \quad (22)$$

and for the constrained continuation-ratio model, the correct model was

$$g_j(x) = \alpha_j + 0.15x - 0.6d - \beta_3xd, \quad \alpha_j = [-1.5, 0.5, 1.0], \quad j = 1, 2, 3. \quad (23)$$

In both cases, the fitted model excluded the interaction term ($\beta_3 = 0$). The coefficient of the interaction term was $\beta_3 = 0.2, 0.4, 0.6$, corresponding to three scenarios with increasing difference between the correct and fitted models. Web Figures 9 and 10 show the probability distributions of the responses as functions of the covariates using $\beta_3 = 0.4$ and the models in Equations (22) and (23),
Table 4. Simulated power (%) for four scenarios using \( n = 400 \) and selected parameter values. The full results can be found in Web Tables 9–16.

| Significance level | AC model \( \beta_3 = 0.04 \) | CCR model \( \beta_3 = 0.04 \) |
|--------------------|---------------------------------|---------------------------------|
| Ommission of a quadratic term in a continuous covariate | | |
| \( C_8 \) | 12.4 | 29.4 | 40.8 | 6.64 | 18.6 | 29.9 |
| \( C_{10} \) | 13.3 | 30.4 | 43.0 | 6.68 | 18.7 | 29.7 |
| \( C_{12} \) | 12.0 | 28.8 | 41.3 | 5.56 | 17.5 | 28.0 |
| Lipsitz | 31.5 | 53.7 | 66.1 | 15.3 | 33.5 | 45.5 |
| \( PR(\chi^2) \) | 2.34 | 8.82 | 16.2 | 1.36 | 5.94 | 12.5 |
| \( PR(D^2) \) | 2.46 | 9.78 | 17.5 | 1.52 | 6.56 | 13.4 |
| Ommission of an interaction term between a continuous and a dichotomous covariate \( \beta_3 = 0.4 \) | | |
| \( C_8 \) | 99.6 | 99.9 | 100 | 93.9 | 98.4 | 99.3 |
| \( C_{10} \) | 99.2 | 99.9 | 100 | 90.9 | 97.6 | 98.9 |
| \( C_{12} \) | 98.4 | 99.7 | 99.9 | 88.3 | 96.3 | 98.5 |
| Lipsitz | 96.1 | 96.1 | 96.1 | 95.5 | 99.9 | 100 |
| \( PR(\chi^2) \) | 99.7 | 100 | 100 | 97.9 | 99.5 | 99.8 |
| \( PR(D^2) \) | 99.8 | 100 | 100 | 98.1 | 99.6 | 99.8 |
| Ommission of an interaction term between two dichotomous covariates \( \beta_4 = 2.0 \) | | |
| \( C_8 \) | 42.3 | 56.6 | 65.1 | 2.50 | 8.20 | 13.9 |
| \( C_{10} \) | 41.2 | 54.1 | 62.1 | 2.74 | 8.54 | 15.0 |
| \( C_{12} \) | 41.5 | 53.9 | 60.8 | 3.14 | 9.08 | 15.2 |
| Lipsitz | 47.1 | 68.2 | 78.0 | 3.22 | 10.6 | 18.2 |
| \( PR(\chi^2) \) | 67.3 | 79.9 | 85.9 | 92.1 | 98.1 | 99.0 |
| \( PR(D^2) \) | 43.6 | 65.1 | 76.1 | 93.7 | 98.4 | 99.2 |
| Wrong functional form of a continuous covariate | | |
| \( C_8 \) | 6.90 | 20.3 | 30.9 | 2.92 | 10.4 | 18.0 |
| \( C_{10} \) | 7.70 | 20.5 | 31.9 | 3.22 | 11.1 | 18.9 |
| \( C_{12} \) | 7.60 | 20.2 | 30.7 | 3.12 | 11.3 | 19.7 |
| Lipsitz | 21.1 | 41.3 | 53.9 | 8.60 | 21.5 | 32.5 |
| \( PR(\chi^2) \) | 2.48 | 9.82 | 17.4 | 1.16 | 6.42 | 12.7 |
| \( PR(D^2) \) | 2.80 | 10.5 | 18.3 | 1.40 | 7.36 | 13.6 |

Note: AC, adjacent-category; CCR, constrained continuation-ratio.

respectively. We give the full results of the simulations in Web Tables 11 and 12. An excerpt is shown in Table 4.

All tests had excellent power to detect a missing interaction term between a continuous and a dichotomous covariate under both the adjacent-category and the constrained continuation-ratio models. For small sample sizes or small values of \( \beta_3 \), the Lipsitz test had greater power than the \( PR \) tests, which, in turn, had greater power than the \( C_g \) tests.

5.3. Omission of an interaction term between two dichotomous covariates

Let \( d_1 \) and \( d_2 \) denote two dichotomous covariates, distributed as Bernoulli(0.5). For the adjacent-category model, the correct model was defined as

\[
g_j(x) = \alpha_j - 0.15x - 0.75d_1 + 1.25d_2 + \beta_4 d_1 d_2, \quad \alpha_j = [-1.25, -1.75, -1.5], \quad j = 1, 2, 3, \quad (24)
\]

and for the constrained continuation-ratio model, the correct model was

\[
g_j(x) = \alpha_j - 0.2x - 1.0d_1 + 1.25d_2 - \beta_4 d_1 d_2, \quad \alpha_j = [0.5, 1.5, 3.0], \quad j = 1, 2, 3. \quad (25)
\]

In both cases, the fitted model excluded the interaction term (\( \beta_4 = 0 \)). The coefficient of the interaction term was \( \beta_4 = 1.0, 2.0, \) and 3.0, corresponding to three scenarios with increasing difference between the correct and fitted models. Web Figures 11 and 12 show the probability distributions of the responses as functions of the covariates using \( \beta_4 = 2.0 \) and the models in Equations (24) and (25),
respectively. We give the full results of the simulations in Web Tables 13 and 14. An excerpt is shown in Table 4.

Under the adjacent-category model, all tests had good – but not great – power to detect a missing interaction term between two dichotomous covariates. The PR($\chi^2$) test had highest power followed by the Lipsitz test. The PR($D^2$) test was sometimes better and sometimes worse than the $C_g$ tests. Under the constrained continuation-ratio model, we observed excellent power with both PR tests, whereas some unexpected results were observed with the $C_g$ and Lipsitz tests: the power for $\beta_4 \neq 1.0$ was sometimes smaller than the power for $\beta_4 = 1.0$ (Web Table 14). The greatest power was observed for $\beta_4 = 3.0$ (as expected), although neither the $C_g$ tests nor the Lipsitz test had power nearly at the level of the PR tests.

### 5.4. Wrong functional form of a continuous covariate

For the adjacent-category model, the correct model was defined as

$$g_j(x) = \alpha_j + 0.5 \log(x) + 1.25d, \quad \alpha_j = [-1.0, -1.25, -1.5], \quad j = 1, 2, 3, \quad (26)$$

and for the constrained continuation-ratio model, the correct model was

$$g_j(x) = \alpha_j + 0.5 \log(x) - 1.25d, \quad \alpha_j = [-2.0, -0.5, 0.5], \quad j = 1, 2, 3. \quad (27)$$

In both cases, the fitted model included the term $x$ instead of $\log(x)$. Web Figures 13 and 14 show the probability distributions of the responses as functions of the covariates using the models in Equations (26) and (27), respectively. We give the full results of the simulations in Web Tables 15 and 16. An excerpt is shown in Table 4.

Under the adjacent-category model, the $C_g$ tests had low to moderate power, the Lipsitz test had moderate to good power, and the PR tests had poor power to detect the wrongly fitted functional form of $x$. For the constrained continuation-ratio model, the $C_g$ tests had low power, the Lipsitz test had low to moderate power, whereas the PR tests had power slightly above the nominal significance level.

### 5.5. Detection of a nominal (unordered) response variable

We generated a four-level response variable based on the following multinomial logits:

$$g_1(x) = -0.75 + 0.25x - 1.25d,$$  \quad (28)

$$g_2(x) = -0.50 + 0.30x - 0.50d,$$  \quad (29)

$$g_3(x) = -1.25 + 0.35x - 1.00d. \quad (30)$$

We then fitted the two ordinal models with $x$ and $d$ as covariates. Web Figure 15 shows the probability distributions of the model in Equations (28)–(30) as functions of the covariates. We give the full results of the simulations in Web Tables 17 and 18. An excerpt is shown in Table 5.

The PR tests had excellent power to detect an unordered response variable under both the adjacent-category and the constrained continuation-ratio models. The $C_g$ tests had good power for both models, whereas the power of the Lipsitz test was poor.

### 5.6. Omission of an independent covariate

In this scenario, we generated data from models with two continuous covariates ($x_1$ and $x_2$) and one dichotomous covariate ($d$), then fitted models without one of the continuous covariates ($x_2$). The
Table 5. Simulated power (%) for four scenarios using \( n = 400 \) and selected parameter values.

| Significance level | AC model | CCR model |
|--------------------|----------|-----------|
|                    | 1%       | 5%        | 10%      | 1%       | 5%        | 10%      |
| Detection of a nominal (unordered) response variable |          |           |          |          |           |          |
| \( C_8 \)          | 20.8     | 43.2      | 56.3     | 55.7     | 78.0      | 86.9     |
| \( C_{10} \)       | 17.4     | 37.6      | 50.6     | 48.7     | 72.3      | 82.5     |
| \( C_{12} \)       | 15.4     | 34.3      | 46.6     | 43.2     | 67.2      | 78.6     |
| Lipsitz            | 2.18     | 8.58      | 15.3     | 3.58     | 11.2      | 19.0     |
| PR(\( \chi^2 \))  | 72.5     | 88.8      | 94.1     | 92.5     | 97.9      | 99.1     |
| PR(D^2)            | 73.1     | 89.4      | 94.4     | 93.3     | 98.0      | 99.2     |
| Omission of an independent covariate |          |           |          |          |           |          |
| \( C_8 \)          | 1.56     | 5.66      | 11.6     | 2.06     | 8.32      | 15.0     |
| \( C_{10} \)       | 1.26     | 6.00      | 11.3     | 1.76     | 7.66      | 13.8     |
| \( C_{12} \)       | 1.38     | 6.04      | 11.4     | 1.82     | 7.74      | 14.3     |
| Lipsitz            | 1.70     | 7.38      | 14.0     | 1.50     | 6.18      | 12.1     |
| PR(\( \chi^2 \))  | 1.64     | 7.64      | 13.4     | 2.76     | 11.0      | 19.1     |
| PR(D^2)            | 2.00     | 8.66      | 14.6     | 3.20     | 12.1      | 20.4     |
| Detection of CCR data by the AC model, \( \beta_1 = 0.25, \beta_2 = 0.5 \) |          |           |          |          |           |          |
| \( C_8 \)          | 25.2     | 48.7      | 61.9     |          |           |          |
| \( C_{10} \)       | 21.2     | 42.8      | 56.6     |          |           |          |
| \( C_{12} \)       | 18.0     | 38.7      | 51.1     |          |           |          |
| Lipsitz            | 2.90     | 10.6      | 18.0     |          |           |          |
| PR(\( \chi^2 \))  | 31.9     | 56.6      | 69.6     |          |           |          |
| PR(D^2)            | 37.5     | 62.8      | 74.7     |          |           |          |
| Detection of AC data by the CCR model, \( \beta_1 = 0.25, \beta_2 = 0.75 \) |          |           |          |          |           |          |
| \( C_8 \)          | 11.2     | 26.5      | 37.7     |          |           |          |
| \( C_{10} \)       | 9.96     | 24.8      | 35.9     |          |           |          |
| \( C_{12} \)       | 8.88     | 23.7      | 34.3     |          |           |          |
| Lipsitz            | 1.34     | 6.42      | 12.5     |          |           |          |
| PR(\( \chi^2 \))  | 14.6     | 32.0      | 44.5     |          |           |          |
| PR(D^2)            | 13.4     | 31.3      | 43.1     |          |           |          |

Notes: The full results can be found in Web Tables 17–22. AC, adjacent-category; CCR, constrained continuation-ratio.

The adjacent-category model was

\[
g_j(x) = \alpha_j - 0.25x_1 - 1.25d + 0.5x_2, \quad \alpha_j = [-1.0, 0.25, 0.5], \quad j = 1, 2, 3, \quad (31)
\]

and the constrained continuation-ratio model was

\[
g_j(x) = \alpha_j - 0.25x_1 - 1.25d + 0.5x_2, \quad \alpha_j = [-2.0, -1.0, 0], \quad j = 1, 2, 3. \quad (32)
\]

We give the full results of the simulations in Web Tables 19 and 20. An excerpt is shown in Table 5. The power of all tests was low. The highest power was observed for the two PR tests under the constrained continuation-ratio model when \( n = 800 \); however, even with such a large sample size, the 5% rejection rates were below 20%.

5.7. Detection of constrained continuation-ratio data by the adjacent-category model

In this section, we generated data from the constrained continuation-ratio model and fit them with the adjacent-category model. The models consisted of one continuous covariate (\( x \)) and one dichotomous covariate (\( d \)). We considered two scenarios, one in which the continuous covariate was the most influential, and one in which the dichotomous covariate was the most influential. The two constrained continuation-ratio models used to generate the data that was fitted by the adjacent-category model
were
\[ g_j(x) = \alpha_j + 0.25x + 0.5d, \quad \alpha_j = [-1.5, -1.0, 0], \quad j = 1, 2, 3, \quad (33) \]
and
\[ g_j(x) = \alpha_j + 0.05x + 1.0d, \quad \alpha_j = [-1.5, -1.0, 0], \quad j = 1, 2, 3. \quad (34) \]

Web Figure 16 shows the probability distributions of the models in Equations (33) and (34) as functions of the covariates. We give the full results of the simulations in Web Tables 21. An excerpt is shown in Table 5. When the continuous covariate was more influential than the dichotomous covariate, the \( C_g \) tests had good power and the PR tests had very good power to detect that the wrong model had been fit. The power of the PR tests was quite good also for the scenario in which the dichotomous covariate was the most influential, whereas the power of the \( C_g \) tests was low in this scenario. The Lipsitz test had poor power in both scenarios.

### 5.8. Detection of adjacent-category data by the constrained continuation-ratio model

This section contains a similar scenario as the previous section; however, we now generated data by the adjacent-category model and fit them with the constrained continuation-ratio model. Again, two scenarios were considered: one in which the continuous covariate was the most influential, and one in which the dichotomous covariate was the most influential. The two adjacent-category models used to generate the data that was fitted by the constrained continuation-ratio model were

\[ g_j(x) = \alpha_j + 0.25x + 0.75d, \quad \alpha_j = [-0.5, 0.5, 1.0], \quad j = 1, 2, 3, \quad (35) \]

and

\[ g_j(x) = \alpha_j + 0.075x + 1.75d, \quad \alpha_j = [-0.5, 0.5, 1.0], \quad j = 1, 2, 3. \quad (36) \]

Web Figure 17 shows the probability distributions of the models in Equations (35) and (36) as functions of the covariates. We give the full results of the simulations in Web Tables 22. An excerpt is shown in Table 5. We observe a similar pattern of results here as in the previous section: the PR tests have good power in both scenarios; the \( C_g \) tests have better power when the continuous covariate is the most influential than when the dichotomous covariate is the most influential, although this difference is only noticeably for small sample sizes; and the Lipsitz test has poor power in both scenarios. The simulated power values are higher in Web Table 21 (detection of constrained continuation-ratio data by the adjacent-category model) than in Web Table 22 (detection of adjacent-category data by the constrained continuation-ratio model), which can be seen also in Table 5; however, this may be due to the choices of model parameters. It seems that the effect of the continuous covariate is stronger in Web Figure 16 (detection of constrained continuation-ratio data by the adjacent-category model) than in Web Figure 17 (detection of adjacent-category data by the constrained continuation-ratio model), which may affect the power values.

### 6. Example

In [3], the proportional odds model was fitted to data from a study of determinants of aftercare services for psychiatrically hospitalized adolescents,[11] and goodness of fit was assessed with the \( C_g \), Lipsitz, and PR tests. The data set for that study, called the Adolescent Placement Study, is available from the website of the University of Massachusetts (http://www.umass.edu/statdata/statdata). In short, the data set consists of the medical records of 508 adolescents admitted to three psychiatric hospitals and includes sociodemographics, clinical and family characteristics, service history, and treatment characteristics. A description of the variables in the data set and details on fitting, interpretation of model parameters, and model-building strategies can be found in [1]. Here, we expand
Table 6. Results of fitting an adjacent-category model of Danger on Weeks, Behav, Gender, and Custd; \( n = 508 \).

| Variable | Coefficient | Standard error | \( z \) | \( Pr > |z| \) | 95% CI |
|----------|-------------|---------------|------|-------------|-------|
| Weeks    | 0.028       | 0.014         | 2.04 | 0.041       | (0.0011, 0.055) |
| Behav    | 0.36        | 0.037         | 9.60 | < 0.001     | (0.28, 0.43)    |
| Gender   | 0.42        | 0.12          | 3.62 | < 0.001     | (0.19, 0.65)    |
| Custd    | -0.32       | 0.11          | -2.75| 0.006       | (-0.54, -0.091) |
| \( \hat{\alpha}_1 \) | -0.88       | 0.21          |      |             |       |
| \( \hat{\alpha}_2 \) | -1.72       | 0.42          |      |             |       |
| \( \hat{\alpha}_3 \) | -2.10       | 0.68          |      |             |       |

Table 7. Results of fitting a constrained continuation-ratio model of Danger on Weeks, Behav, Gender, and Custd; \( n = 508 \).

| Variable | Coefficient | Standard error | \( z \) | \( Pr > |z| \) | 95% CI |
|----------|-------------|---------------|------|-------------|-------|
| Weeks    | 0.041       | 0.018         | 2.26 | 0.024       | (0.0054, 0.076) |
| Behav    | 0.46        | 0.046         | 9.94 | < 0.001     | (0.37, 0.55)    |
| Gender   | 0.53        | 0.16          | 3.39 | 0.001       | (0.23, 0.84)    |
| Custd    | -0.50       | 0.15          | -3.22| 0.001       | (-0.80, -0.19) |
| \( \hat{\alpha}_1 \) | 0.31        | 0.25          |      |             |       |
| \( \hat{\alpha}_2 \) | 1.60        | 0.28          |      |             |       |
| \( \hat{\alpha}_3 \) | 2.83        | 0.31          |      |             |       |

on the analyses in [1,3] to illustrate testing goodness of fit for the adjacent-category and constrained continuation-ratio models.

We first consider the response variable Neuro (neuropsychiatric disturbance: \( 1 = \) none, \( 2 = \) mild, \( 3 = \) moderate, \( 4 = \) severe). As described in [3], there was no evidence of lack of fit for a proportional odds model with Age (in years, centred about the sample mean age of 14.3 years), \( \text{Age}^2 \), Gender (\( 0 = \) female, \( 1 = \) male), Race (\( 0 = \) white, \( 1 = \) non-white), Emot (emotional disturbance: \( 0 = \) mild, \( 1 = \) severe), and Custd (state custody: \( 0 = \) no, \( 1 = \) yes) as covariates (\( p \geq .39 \) for all tests). We now fit adjacent-category and constrained continuation-ratio models of Neuro on the six covariates and calculate the \( C_8 \), \( C_{10} \), \( C_{12} \), Lipsitz, \( PR(\chi^2) \), and \( PR(D^2) \) tests. Web Tables 23 and 24 show the results of the model fits, which are similar to the fit of the proportional odds model in [3]. We observe no evidence of lack of fit for any of the models with any of the tests: \( p \)-values from .39 to .98 for the adjacent-category model and \( p \)-values from .18 to .92 for the constrained continuation-ratio model (Web Table 25).

Next, we consider the response variable Danger (danger to others: \( 1 = \) unlikely, \( 2 = \) possibly, \( 3 = \) probably, and \( 4 = \) likely). In [1], the purposeful selection procedure was used to obtain a proportional

Table 8. Results of applying the goodness-of-fit tests to adjacent-category and constrained continuation-ratio models of Danger on Weeks, Behav, Gender, and Custd; \( n = 508 \).

| Model/test | Value of test statistic | DOF | \( p \)-Value |
|------------|------------------------|-----|-------------|
| Adjacent-category model | | | |
| \( C_8 \) | 35.7 | 20 | .017 |
| \( C_{10} \) | 41.2 | 26 | .030 |
| \( C_{12} \) | 54.9 | 32 | .0071 |
| Lipsitz | 6.96 | 9 | .64 |
| \( PR(\chi^2) \) | 38.8 | 18 | .0031 |
| \( PR(D^2) \) | 40.1 | 18 | .0021 |

| Constrained continuation-ratio model | | | |
| \( C_8 \) | 56.5 | 20 | < .0001 |
| \( C_{10} \) | 72.3 | 26 | < .0001 |
| \( C_{12} \) | 87.1 | 32 | < .0001 |
| Lipsitz | 5.71 | 9 | .77 |
| \( PR(\chi^2) \) | 53.4 | 18 | < .0001 |
| \( PR(D^2) \) | 56.8 | 18 | < .0001 |

Note: DOF = \( \chi^2 \) degrees of freedom.
Table 9. Observed and estimated frequencies sorted and summed into 10 groups, following the fit of the adjacent-category model of Danger on Weeks, Behav, Gender, and Custd.

| Group | Danger = 1 | | Danger = 2 | | Danger = 3 | | Danger = 4 | | Sum |
|-------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-------|
|       | Obs.  | Est. | Obs.  | Est. | Obs.  | Est. | Obs.  | Est. |       |
| 1     | 27    | 24.0 | 15    | 18.1 | 1     | 6.91 | 8     | 1.98  | 51    |
| 2     | 13    | 11.6 | 20    | 18.9 | 10    | 13.7 | 8     | 6.85  | 51    |
| 3     | 7     | 6.99 | 20    | 16.1 | 13    | 16.4 | 11    | 11.5  | 51    |
| 4     | 5     | 4.74 | 13    | 13.7 | 19    | 17.4 | 14    | 15.2  | 51    |
| 5     | 1     | 2.71 | 11    | 10.4 | 20    | 17.3 | 18    | 19.6  | 50    |
| 6     | 1     | 1.70 | 9     | 8.09 | 18    | 16.9 | 23    | 24.3  | 51    |
| 7     | 0     | 1.18 | 8     | 6.62 | 18    | 16.2 | 25    | 27.0  | 51    |
| 8     | 0     | 0.59 | 3     | 4.37 | 12    | 14.3 | 36    | 31.8  | 51    |
| 9     | 0     | 0.36 | 2     | 3.23 | 20    | 12.8 | 29    | 34.7  | 51    |
| 10    | 0     | 0.12 | 0     | 1.54 | 15    | 10.9 | 40    | 39.2  | 50    |

Note: Obs, observed frequency; Est, estimated frequency.

odds model for Danger, which resulted in four significant covariates (at the 5% level): Weeks (length of stay in hospital, measured in weeks), Behave (behavioural symptom score 0–9), Gender, and Custd. The final model was evaluated for fit, using Brant’s test [12] \( p < .001 \), an approximate likelihood ratio test \( p < .001 \), and the \( C_{10} \) test \( p = .001 \), which indicated that the proportional odds assumption was violated. Using the same covariates, we fit adjacent-category (Table 6) and constrained continuation-ratio (Table 7) models to Danger and calculate the goodness-of-fit tests (Table 8). All three \( C_{g} \) tests and both PR tests show clear signs of lack of fit for both the adjacent-category and constrained continuation-ratio models. The Lipsitz test, on the other hand, does not indicate lack of fit with any of the two models.

A secondary tool to examine lack of fit is to tabulate the observed and estimated frequencies, as in Table 1. Using \( g = 10 \) groups, corresponding to the \( C_{10} \) test, the contingency table for the adjacent-category model of Danger on Weeks, Behav, Gender, and Custd is shown in Table 9. The agreement between the observed and estimated frequencies is mostly fair, although there are some large discrepancies, particularly in the row for group 1 and in the columns corresponding to response levels Danger = 3 and Danger = 4 that suggest lack of fit. Because we are not able to relate these discrepancies to specific values of the covariates, and because case-wise diagnostic tools have not been developed for ordinal regression models, this is as far as we can go with the assessment of fit for this particular model.

So far, we have considered proportional odds, adjacent-category, and constrained continuation-ratio models for the relationship between Danger and Weeks, Behav, Gender, and Custd. These three models all constrain the regression coefficients to be constant across the logits, and they all exhibit lack of fit. The best way forward may be to use a model that allows the regression coefficients to vary with the logits. The unconstrained continuation-ratio model does this without ignoring the ordinal nature of the response. Unfortunately, there are as yet no methods available to assess the goodness of fit of such models. An alternative solution is to use a multinomial logistic regression model, for which a multinomial \( C_{g} \) test and a normalized version of the ungrouped Pearson \( \chi^2 \) test are available.[5] The disadvantage with this approach is that the response is treated as unordered, thus ignoring the inherent information in the ordinal response.

7. Discussion

This article has investigated the null distribution and power properties of three approaches for testing goodness of fit under two ordinal logistic regression models: the adjacent-category and the constrained continuation-ratio models. An ordinal version of the Hosmer–Lemeshow test \( (C_{g}) \) had excellent null distribution results and good, but not great, power. The Lipsitz test [7] had null rejection rates slightly above the nominal level and higher power than the \( C_{g} \) test for four out of eight of
the considered scenarios (Table 10). Two tests by Pulkstenis and Robinson [8] had null rejection rates somewhat above the nominal level under both ordinal models. The power of the PR tests was better than the $C_g$ test for five of the eight considered scenarios and better than the Lipsitz test for four of the eight scenarios (Table 10). The $C_g$ and Lipsitz tests work best when lack of fit is associated with continuous covariates, whereas the PR tests excel if categorical covariates drive the lack of fit, such as when an interaction term that includes a categorical covariate is missing. Another useful feature of the PR tests is the ability to assess which covariate patterns contribute to the lack of fit by way of a contingency table of observed and estimated frequencies. An example where scrutiny of the contingency table suggests inclusion of an interaction term between two dichotomous covariates can be seen in the Example section of [8].

The simulation results for the adjacent-category and the constrained continuation-ratio models in this article resemble those in Fagerland and Hosmer,[3] who evaluated the $C_g$, Lipsitz, and PR approaches for testing goodness of fit under the proportional odds regression model. The only noticeable difference was that the power to detect that a continuous covariate was modelled as $x$ instead of $\log(x)$ was higher under the adjacent-category model than under the proportional odds model. The consistency of results across this article and Fagerland and Hosmer [3] allows us to make a joint recommendation for testing goodness of fit in these three ordinal logistic regression models. Because the different tests may detect different types of lack of fit, we recommend use of all three approaches.

Three versions of the $C_g$ test were assessed: $C_8$, $C_{10}$, and $C_{12}$, which correspond to grouping observations into 8, 10, and 12 groups. The three tests had similar null distribution properties; however, the power of $C_8$ was usually higher than that of $C_{10}$, which in turn was higher than that of $C_{12}$. This result is consistent with the results for the multinomial $C_g$ test in [5] and the $C_g$ test for the proportional odds model in [3]. Even though the power of the $C_8$ test seems to be higher than the $C_{10}$ test, we recommend the $C_{10}$ test because using 10 groups is an established practice for the binary Hosmer–Lemeshow test.

In some cases, regardless of the type of ordinal model used, one or more of the tests may not be computable. The Lipsitz test is not recommended if the number of groups ($g$) cannot satisfy $6 \leq g < n/5c$, where $n$ is the sample size and $c$ is the number of response levels. The PR tests can only be calculated if both continuous and categorical covariates are present. If the number of covariates patterns is small, which happens when there are no continuous covariates and only a few categorical covariates present, neither the $C_g$ nor the PR tests can be calculated. In this case, the standard Pearson $\chi^2$ statistic can be used to assess goodness of fit.

The most important limitations in this article are associated with the limited range of the scenarios and parameters of the simulation study: (i) the distribution and number of covariates have been kept few and small; (ii) the relationship between the covariates and the response has been described using a limited set of regression coefficients; and (iii) power has been assessed using eight specific scenarios for lack of fit. These restrictions have been imposed to keep the simulation study manageable to

---

**Table 10.** Qualitative summary of between-test comparisons of power.

| Power scenario | AC model | CCR model |
|----------------|----------|-----------|
| Missing $x^2$  | Lipsitz $> C_g$ $> \text{PR}^a$ | Lipsitz $> C_g$ $> \text{PR}^b$ |
| Missing $x_d$  | Lipsitz $\approx \text{PR} > C_g$ | Lipsitz $\approx \text{PR}(D^2) > \text{PR}(\chi^2) > C_g$ |
| Missing $d_1d_2$ | $\text{PR}(\chi^2)$ $> \text{Lipsitz} > \text{PR}(D^2) \approx C_g$ | $\text{PR}(D^2) > \text{PR}(\chi^2) > \text{Lipsitz} > C_g$ |
| $x$ instead of $\log(x)$ | Lipsitz $> C_g > \text{PR}^a$ | Lipsitz $> C_g > \text{PR}$ |
| Nominal model   | $C_g^a \approx \text{Lipsitz}^b \approx \text{PR}^a$ | $\text{PR}^b \approx C_g^a \approx \text{Lipsitz}^b$ |
| Missing covariate | PR $> C_g > \text{Lipsitz}^a$ | PR $> C_g > \text{Lipsitz}^b$ |
| CCR data       | PR $> C_g > \text{Lipsitz}^a$ | PR $> C_g > \text{Lipsitz}^b$ |
| AC data        | PR $> C_g > \text{Lipsitz}^b$ | PR $> C_g > \text{Lipsitz}^b$ |

Note: AC, adjacent-category; CCR, constrained continuation-ratio. $^a$ Poor power. $^b$ Power barely above the nominal significance level.
perform, interpret, and report. We believe the results show a consistency that allow for general – but cautious – recommendations.

In summary, we recommend that all three approaches (Cg, Lipsitz, and PR) are used whenever possible to assess goodness of fit in ordinal logistic regression models. Together, these tests have reasonable power with moderate to large sample sizes to detect several types of lack of fit: omission of a quadratic term in a continuous covariate, omission of an interaction term between a continuous and a dichotomous covariate, omission of an interaction term between two dichotomous covariates, wrong functional form of a continuous covariate, detection of an unordered response variable, detection of constrained continuation-ratio data by the adjacent-category model, and detection of adjacent-category data by the constrained continuation-ratio model. Because power is generally low for small sample sizes, we recommend using a nominal significance level of 10% whenever the sample size is smaller than, say, 400.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**References**

[1] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken, NJ: Wiley; 2013.

[2] Agresti A. Analysis of ordinal categorical data. 2nd ed. Hoboken, NJ: Wiley; 2010.

[3] Fagerland MW, Hosmer DW. A goodness-of-fit test for the proportional odds regression model. Stat Med. 2013;32:2235–2249.

[4] Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Commun Stat. 1980;A9(10):1043–1069.

[5] Fagerland MW, Hosmer DW, Bofin AM. Multinomial goodness-of-fit tests for logistic regression models. Stat Med. 2008;27:4238–4253.

[6] Fagerland MW, Hosmer DW. A generalized Hosmer-Lemeshow goodness-of-fit test for multinomial logistic regression models. Stat Med. 2012;12(3):447–453.

[7] Lipsitz SR, Fitzmaurice GM, Molenberghs G. Goodness-of-fit tests for ordinal response regression models. Appl Stat J Roy Stat Soc C. 1996;45(2):175–190.

[8] Pulkstenis E, Robinson TJ. Goodness-of-fit tests for ordinal response regression models. Stat Med. 2004;23:999–1014.

[9] Fagerland MW. adjcatlogit, ccrlogit, and ucrlogit: Fitting ordinal logistic regression models. Stata J. 2014;14:947–964.

[10] Wolfe R. Continuation-ratio models for ordinal response data. Stata Tech Bull. 1998;44:18–21.

[11] Fontanella CA, Early TJ, Phillips G. Need or availability? Modeling aftercare decisions for psychiatrically hospitalized adolescents. Child Youth Serv Rev. 2008;30:758–773.

[12] Brant R. Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics. 1990;46:1171–1178.

**Appendix 1. Conditional probabilities for the adjacent-category model**

The logits are defined as

\[ g_j(x) = \log \left( \frac{\Pr(Y = j + 1 \mid x)}{\Pr(Y = j \mid x)} \right) = \alpha_j + \beta_j' x, \quad j = 1, \ldots, c - 1. \]  

Let \( \pi_j = \Pr(Y = j \mid x) \). From Equation (A1), we have that

\[ \pi_{j+1} = \tau_j \cdot e^{g_j(x)} = \left( \tau_{j-1} \cdot e^{g_{j-1}(x)} \right) \cdot e^{g_j(x)} = \tau_1 \cdot e^{g_1(x)} \cdot e^{g_2(x)} \cdots \cdot e^{g_j(x)} = \pi_1 \cdot e^{g_{1}(x)+g_{2}(x)+\ldots+g_{j}(x)} = \pi_1 \cdot \theta_j, \]
where \( \theta_j = e^{\sum_{k=1}^{c} g_k(x)} \). The sum of the probabilities equals one, hence
\[
\pi_1 = 1 - \sum_{k=2}^{c} \pi_k \\
= 1 - \sum_{k=2}^{c} \pi_1 \cdot \theta_{k-1} \\
= 1 - \pi_1 \cdot \sum_{k=1}^{c-1} \theta_k \\
= 1 - \pi_1 \cdot \theta,
\]
where \( \theta = \sum_{k=1}^{c-1} \theta_k \). We can now formulate the following expressions for the conditional probabilities:
\[
\pi_1 = \frac{1}{1 + \theta}, \\
\pi_{j+1} = \frac{\theta_j}{1 + \theta}, \quad j = 1, \ldots, c - 1.
\]

### Appendix 2. Conditional probabilities for the continuation-ratio model

The logits for the constrained continuation-ratio model are defined as
\[
g_j(x) = \log \left( \frac{\Pr(Y = j \mid x)}{\Pr(Y > j \mid x)} \right) = \alpha_j - \beta' x, \quad j = 1, \ldots, c - 1.
\]
Let \( \pi_j = \Pr(Y = j \mid x) \), such that
\[
g_j(x) = \log \left( \frac{\pi_j}{\pi_{j+1} + \cdots + \pi_c} \right), \quad j = 1, \ldots, c - 1. \tag{A2}
\]
We also have that the sum of the probabilities equals one:
\[
\sum_{k=1}^{c} \pi_k = 1. \tag{A3}
\]
From Equation (A3), we get that \( \sum_{k=2}^{c} \pi_k = 1 - \pi_1 \). Combined with Equation (A2), evaluated at \( j = 1 \), we obtain
\[
g_1(x) = \log \left( \frac{\pi_1}{1 - \pi_1} \right).
\]
Solving for \( \pi_1 \):
\[
\pi_1 = \frac{e^{g_1(x)}}{1 + e^{g_1(x)}}. \tag{A4}
\]
We now show how to obtain \( \pi_2 \). We rewrite Equation (A3) such that
\[
\sum_{k=3}^{c} \pi_k = 1 - \pi_1 - \pi_2 \\
= \gamma_1 - \pi_2,
\]
where \( \gamma_1 = 1 - \pi_1 \), which is a known quantity from Equation (A4). Combined with Equation (A2), evaluated at \( j = 2 \), we obtain
\[
g_2(x) = \log \left( \frac{\pi_2}{\gamma_1 - \pi_2} \right).
\]
Solving for \( \pi_2 \):
\[
\pi_2 = \frac{\gamma_1 \cdot e^{g_2(x)}}{1 + e^{g_2(x)}}.
\]
In general, we have
\[
\gamma_j = 1 - \sum_{k=1}^{j} \pi_k, \quad j = 1, \ldots, c - 1,
\]
where

$$\pi_j = \frac{\gamma_j - 1 \cdot e^{g_j(x)}}{1 + e^{g_j(x)}}, \quad j = 2, \ldots, c - 1,$$

and

$$\pi_c = 1 - \sum_{k=1}^{c-1} \pi_k.$$