Graphical Diagnostics to Check Model Misspecification for the Proportional Odds Regression Model

Ivy Liu[1], Bhramar Mukherjee[2], Thomas Suesse[1], David Sparrow[3]
and Sung Kyun Park[4]
[1] School of Mathematics, Statistics, and Computer Science
Victoria University of Wellington
Wellington, New Zealand
[2] Department of Biostatistics, University of Michigan, Ann Arbor, MI.
[3] VA Normative Aging Study,
Veterans Affairs Boston Health Care System,
Boston, MA
[4] Department of Environmental Health Sciences
University of Michigan, Ann Arbor, MI

December 17, 2007

SUMMARY

The cumulative logit or the proportional odds regression model is one of the popular choices to study covariate effects on ordinal responses. This paper provides some graphical and numerical methods for checking the adequacy of the proportional odds regression model. The methods focus on evaluating functional misspecification for specific covariate effects, but misspecification of the link function can also be dealt under the same framework. For the logistic regression model with binary responses, Arbogast and Lin [1] developed similar graphical and numerical methods for assessing the adequacy of the model using the cumulative sums of residuals. The paper generalizes their methods to ordinal responses and illustrates them using an example from the VA Normative Aging Study. Simulation studies comparing the performance of the different diagnostic methods indicate that some of the graphical methods are more powerful in detecting model misspecification than the Hosmer-Lemeshow type goodness-of-fit statistics for the class of models studied.

KEYWORDS: Cumulative residuals, Fasting blood glucose, Gaussian Process, Goodness-of-fit, Hosmer-Lemeshow statistic, Normative Aging Study, Ordinal data.

1 Introduction

Categorical responses with an ordinal scale often occur in many applications. For example, migraine severity or degree of pain is often recorded on a scale of ”none”, ”mild”, ”moderate” and ”severe”. Often there are clinical reasons for recording certain continuous measurements in an ordinal scale, for example, fasting blood glucose is often recorded in three categories, clinically defined as normal level, impaired level, and diabetic level. One may want to study the effect of a biomarker or a treatment or other covariates like age, ethnicity on such ordinal responses through a generalized linear model with linear predictors. For ordinal responses, the proportional odds model [2] is currently the most popular model that uses logits of cumulative probabilities. For a c-category ordinal response variable Y and a set of predictors $\mathbf{X}$ with corresponding effect parameters $\boldsymbol{\beta}$, the model has the form

$$\text{logit}[P(Y \leq j \mid \mathbf{X})] = \alpha_j - \boldsymbol{\beta}^T\mathbf{X},\ j = 1, ..., c-1. \tag{1}$$

(The minus sign in the predictor term makes the sign of each component of $\boldsymbol{\beta}$ have the usual interpretation in terms of whether the effect is positive or negative.) The parameters $\{\alpha_j\}$, called *cut points*, are usually nuisance parameters of little interest. This model applies simultaneously to all $c-1$ cumulative probabilities, and it assumes an identical effect of the predictors for each cumulative probability. By collapsing the response into the binary outcome categories ($\leq j$, $> j$), for a fixed j, the proportional odds model reduces to a standard logistic regression model. Model (1) implies that each of the $c-1$ logistic regression models holds with the same set of coefficients $\boldsymbol{\beta}$.

Although there exists many models to analyze ordinal data (see Reference [3], Ch. 7), a major advantage of using the proportional odds model is that to fit this model, it is unnecessary to assign scores to the response categories. So, when the model fits well, different studies using different scales for the response variable should give similar conclusions. Liu and Agresti [4] gave the detailed motivations of using proportional odds models. Agresti [3] discusses model fitting for the proportional odds model. Agresti [3] also describes other possible alternatives for modeling ordinal responses like the adjacent category logit model or the continuation-ratio logits model.

A common method used to test model fit for categorical responses is to compare observed frequencies and estimated expected frequencies under the assumed model, via a chi-squared type goodness of fit statistics. These goodness of fit statistics [5] use the grouping strategy based on the values of estimated probabilities, and compare the observed and the expected responses in these groups. Lipsitz *et al.* [6] generalized the popular Hosmer–Lemeshow statistic proposed for a logistic regression model with binary data to the situation when one has ordinal responses. Toledano and Gatsonis [7] gave a generalization of a receiver operating characteristic (ROC) curve that plots sensitivity against (1 - specificity) for all possible collapsing of c categories. Kim [8] proposed a graphical method for assessing the proportional odds assumption. All of the above methods check the overall adequacy of the proportional odds model. They do not give a close view of model misspecification for the functional form of specific covariates.

Lin *et al.* [9] and Arbogast and Lin [1] developed graphical and numerical methods for assessing the adequacy of the functional form of a covariate in the logistic regression model using the cumulative sums of residuals. In standard linear regression models, the plot of residuals against the explanatory variable X is often viewed as a diagnostic tool to examine model misspecification in X. The residuals for a binary logistic model are typically defined as the difference between observed response, and the estimated probability of the response, conditional on the covariates. The plot of the residuals vs X is hard to interpret in such cases and Arbogast and Lin [1] recommend using cumulative sums of the residuals over the covariate of interest to check for functional misspecification in X. They prove that when the model is correctly specified, the cumulative residual process converges weakly to a zero-mean Gaussian process. Then, they proceed to compare the observed cumulative residuals pattern with the simulated realization based on the limiting Gaussian process under the null hypothesis that the model is correctly specified. When $c = 2$, the proportional odds model is the logistic regression model. The current paper generalizes the methods proposed by Arbogast and Lin [1] for checking model misspecification to the proportional odds model with $c > 2$ using two different routes.

One approach considers the proportional odds model as $c-1$ logistic regression models, where the response categories are collapsed into the binary outcome $(\leq j, > j)$, $j = 1, \ldots, c-1$. The cumulative sums of residuals have the same form as the ones given by Arbogast and Lin [1] for each of the collapsed logistic models and thus convergence to the limiting Gaussian process follows. In the second approach, the proportional odds model (1) is viewed as a member of the class of multivariate generalized linear models (MGLM, see Reference [10] for a detailed definition). The response variable for subject i in a MGLM is a vector of indicator responses $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{i,c-1})^T$, where $Y_{ij} = 1$ if subject i falls in category j and is 0 otherwise. Consequently, the residual, which is the difference between the observed value of the response and the predicted probability of the response for the ith subject is a $(c-1)\times 1$ vector. We then consider a vector of Gaussian processes for the limiting distribution (process) corresponding to the multivariate cumulative residuals and proceed to assess model misspecification.

The remainder of the paper is organized as follows. Sections 2 and 3 discuss the two approaches respectively. Section 4 gives an example of a recent dataset from the Normative Aging Study [11] which studies the effect of two markers of oxidative stress namely, white blood cell count and C-reactive protein on fasting blood glucose (FBG) measurement in men in the age group of 48 to 94 years. FBG is measured into three clinically defined ordinal categories. In Section 5, we evaluate the performance of these approaches through a small-scale simulation study. The last section contains concluding remarks.

2 Binary Approach

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{i,c-1})^T$ be the response for subject i, where $i = 1, \ldots, n$. If the subject responds as level j, then $Y_{ij} = 1$ and $Y_{ih} = 0$ for all $h \neq j = 1, \ldots, c-1$. If the response is at baseline level c, then $\mathbf{Y}_i = (0, 0, \ldots, 0)^T$.

In the binary approach, we first define the collapsed responses as $Y_{ij}^* = \sum_{h=1}^{j} Y_{ih}$, where $j = 1, \ldots, c-1$. That is, Y_{ij}^* is a binary response variable having values 1, or 0. It can be considered as a binary outcome when we collapse the response categories into $(\leq j, > j)$, $j = 1, \ldots, c-1$. If the response category is $\leq j$, then $Y_{ij}^* = 1$. Otherwise, $Y_{ij}^* = 0$. For the jth collapsing, the residual is defined as,

$$r_{ij}^* = Y_{ij}^* - P(Y \leq j \mid \mathbf{X}_i), \tag{2}$$

where $\mathbf{X}_i$ is the covariate vector for the ith subject and $P(Y \leq j \mid \mathbf{X}_i)$ satisfies the proportional odds model (1), which is simply a logistic regression model for a fixed j. Therefore, this approach is equivalent to the method used for the logistic regression model given by Arbogast and Lin [1] for each specific collapsing. Let $\boldsymbol{\delta}^T = (\alpha_1, \alpha_2, \ldots, \alpha_{c-1}, \boldsymbol{\beta}^T)$ and let $\boldsymbol{\delta}_j^T = (\alpha_j, \boldsymbol{\beta}^T)$, which represents the parameters for the jth collapsed model (1). Consider the following stochastic process

$$W_k^{(j)}(t; \hat{\boldsymbol{\delta}}) = n^{-1/2} \sum_{i=1}^{n} \hat{r}_{ij}^* I(X_{ik} \leq t),$$

where X_{ik} is the kth component of $\mathbf{X}_i$ and $\hat{r}_{ij}^* = Y_{ij}^* - \hat{P}(Y \leq j \mid \mathbf{X}_i)$. The form $W_k^{(j)}(t; \hat{\boldsymbol{\delta}})$ uses a cumulative sum of the residuals $\hat{r}_{ij}^*$ over the values of X_{ik}. Following Arbogast and Lin's argument, under the null hypothesis $\mathcal{H}_0$ that model (1) is correct, $W_k^{(j)}(t; \hat{\boldsymbol{\delta}})$ converges weakly to a zero-mean Gaussian process. The distribution of the Gaussian process can be approximated by that of

$$\widehat{W}_k^{(j)}(t; \hat{\boldsymbol{\delta}}) = n^{-1/2} \sum_{i=1}^{n} \left\{ I(X_{ik} \leq t) + \hat{\boldsymbol{\eta}}^T(t, \hat{\boldsymbol{\delta}}_j) \left[n^{-1} \mathcal{I}(\hat{\boldsymbol{\delta}}_j) \right]^{-1} \left[\begin{array}{c} 1 \\ \mathbf{X}_i \end{array} \right] \right\} Z_i \, \hat{r}_{ij}^*, \tag{3}$$

where

$$\begin{aligned}\hat{\boldsymbol{\eta}}(t, \hat{\boldsymbol{\delta}}_j) &= n^{-1/2} \partial W_k^{(j)}(t; \boldsymbol{\delta}) / \partial \boldsymbol{\delta}_j \mid_{\hat{\boldsymbol{\delta}}_j} \\ &= -n^{-1} \sum_{i=1}^{n} \hat{P}(Y \leq j \mid \mathbf{X}_i) \left[1 - \hat{P}(Y \leq j \mid \mathbf{X}_i)\right] I(X_{ik} \leq t) \begin{bmatrix} 1 \\ \mathbf{X}_i \end{bmatrix},\end{aligned}$$

and, where $\mathcal{I}(\hat{\boldsymbol{\delta}}_j)$ is the information matrix, and $\{Z_i,\, i = 1, \ldots, n\}$ are independent standard normal random variables. The proof of this result was given in Arbogast and Lin [1].

To check model misspecification for covariate X_{ik}, we plot the observed cumulative residuals along with a large number of simulated realizations based on the Gaussian process (3) to compare their relative patterns. Arbogast and Lin [1] used the Kolmogorov-type supremum statistic $G_{W_k} := \sup_{t \in \mathbb{R}} |W_k(t; \hat{\boldsymbol{\delta}})|$, where $\mathbb{R}$ denotes the real line and W_k stands for $W_k^{(j)}$, $j = 1, \ldots, c-1$ in our case. Let g_{W_k} denote the observed value of the supremum statistic G_{W_k}. We cannot compute the p-value $\mathrm{P}(G_{W_k} \geq g_{W_k})$ of the test directly, but $\mathrm{P}(G_{W_k} \geq g_{W_k})$ can be approximated by $\mathrm{P}(G_{\widehat{W}_k} \geq g_{W_k})$, where $G_{\widehat{W}_k} = \sup_{t \in \mathbb{R}} |\widehat{W}_k(t; \hat{\boldsymbol{\delta}})|$. The $\mathrm{P}(G_{\widehat{W}_k} \geq g_{W_k})$ is estimated by generating a large number (≥ 1000) of realizations $\widehat{W}_k(t; \hat{\boldsymbol{\delta}})$. That is, the p-value of the test is obtained by computing the proportion of the simulated realizations greater than the largest value of $|W_k^{(j)}(t; \hat{\boldsymbol{\delta}})|$ over t, because the extreme values of $W_k^{(j)}(t; \hat{\boldsymbol{\delta}})$ would suggest that functional misspecification exists for covariate X_{ik}. For each collapsed response, it results in a single plot and a single p-value. In total, there are $c-1$ plots denoted by $B_1, \ldots, B_{c-1}$. One might use the Bonferroni method to adjust for the significance level while combining inference from all these plots, so that the overall Type I error rate is less than or equal to the sum of the individual error rates for all $c-1$ plots. The Bonferroni adjusted significance level is thus the significance level divided by $c-1$. Later, we refer to it as Bonf(B).

Remark: The same method for assessing model misspecification in terms of covariates can be used to judge adequacy of the proportional odds link function where the cumulative residuals are summed over the linear predictor $\left[1,\, \mathbf{X}_i^T\right] \boldsymbol{\delta}_j$ instead of the covariate X_{ik}.

Graphical tools can be constructed exactly as in the same way as above by checking whether the observed realization is a random happenstance under the null model or an occurrence beyond chance.

3 Multivariate Approach

For the multivariate approach, we assume that $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{i,c-1})^T$ is a multinomially distributed random variable with parameter $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{i,c-1})^T$, where $\pi_{ij} = P(Y = j \mid \mathbf{X}_i)$. For the proportional odds model (1), $\pi_{ij} = P(Y \leq j \mid \mathbf{X}_i) - P(Y \leq j-1 \mid \mathbf{X}_i)$. The multivariate residuals can be written as a vector

$$\mathbf{r}_i = \mathbf{Y}_i - \boldsymbol{\pi}_i \,.$$

We consider a vector of stochastic processes

$$\mathbf{W}_k^m(t; \hat{\boldsymbol{\delta}}) = n^{-1/2} \sum_{i=1}^{n} I(X_{ik} \leq t) \hat{\mathbf{r}}_i \,.$$

If the model holds, $\mathbf{W}_k^m(t; \hat{\boldsymbol{\delta}})$ converges weakly to a vector of zero-mean Gaussian processes. The distribution of the processes can be approximated by

$$\widehat{\mathbf{W}}_k^m(t; \hat{\boldsymbol{\delta}}) = n^{-1/2} \sum_{i=1}^{n} \left[I(X_{ik} \leq t) \hat{\mathbf{r}}_i + \hat{\boldsymbol{\eta}}^T(t, \hat{\boldsymbol{\delta}}) \hat{\boldsymbol{\Omega}}^{-1} \hat{\mathbf{U}}_i \right] Z_i$$

where Z_i are independent standard normal random variables, $\boldsymbol{\eta}(t, \boldsymbol{\delta}) = n^{-1/2} \partial \mathbf{W}_k^m / \partial \boldsymbol{\delta} = -n^{-1} \sum_i I(X_{ik} \leq t) \frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\delta}}$, $\boldsymbol{\Omega} = n^{-1} \times$ the scaled Information Matrix, and $\mathbf{U}_i$ is the score function for subject i. The proof of this result is furnished in the appendix.

Instead of using the multivariate residuals $\mathbf{r}$, we can use the multivariate cumulative residuals $\mathbf{r}^*$ defined by

$$\mathbf{r}_i^* = \mathbf{Y}_i^* - \boldsymbol{\pi}_i^*,$$

where $\mathbf{r}_i^* = (r_{i1}^*, r_{i2}^*, \ldots, r_{i(c-1)}^*)^T$, $\mathbf{Y}_i^* = (Y_{i1}^*, Y_{i2}^*, \ldots, Y_{i(c-1)}^*)^T$, and $\boldsymbol{\pi}_i^* = (P(Y \leq 1|\mathbf{X}_i), P(Y \leq 2|\mathbf{X}_i), \ldots, P(Y \leq c-1|\mathbf{X_i}))^T$. Section 2 defined the notations r_{ij}^* and

Y_{ij}^* in (2). We consider a vector of stochastic processes

$$\mathbf{W}_k^*(t;\hat{\boldsymbol{\delta}}) = n^{-1/2}\sum_{i=1}^{n} I(X_{ik} \le t)\hat{\mathbf{r}}_i^*.$$

Similarly, if the model holds, $\mathbf{W}_k^*(t;\hat{\boldsymbol{\delta}})$ converges weakly to a vector of zero-mean Gaussian processes. The distribution of the processes can be approximated by $\widehat{\mathbf{W}}_k^*(t;\hat{\boldsymbol{\delta}})$, which has the same form as $\widehat{\mathbf{W}}_k^m(t;\hat{\boldsymbol{\delta}})$ but replacing $\hat{\mathbf{r}}_i$ with $\hat{\mathbf{r}}_i^*$ and in $\boldsymbol{\eta}$ replacing $\hat{\boldsymbol{\pi}}_i$ with $\hat{\boldsymbol{\pi}}_i^*$.

Unlike the binary approach, we can not plot the observed multivariate residuals directly, because both $\mathbf{W}_k^m(t;\hat{\boldsymbol{\delta}})$ and $\mathbf{W}_k^*(t;\hat{\boldsymbol{\delta}})$ are vectors. We consider a continuous function $f(\cdot)$

$$f : \mathbb{R}^{c-1} \rightarrow \mathbb{R},$$

where $\mathbb{R}^{c-1}$ denotes the $(c-1)-$dimensional real plane. Applying function $f(\cdot)$ to the above stochastic processes, the continuous mapping theorem implies that $f(\mathbf{W}_k)$ and $f(\widehat{\mathbf{W}}_k)$ converge to the same functional applied to the limiting stochastic process (not necessarily Gaussian), where $\mathbf{W}_k$ stands for either $\mathbf{W}_k^m$ or $\mathbf{W}_k^*$. Define the Kolmogorov-Smirnov type supremum statistic as $G_{f(\mathbf{W}_k)} := \sup_{t\in\mathbb{R}}|f(\mathbf{W}_k(t;\hat{\boldsymbol{\delta}}))|$. Let $g_{f(\mathbf{W}_k)}$ denote the observed value of $G_{f(\mathbf{W}_k)}$. Similar to the G_{W_k} in the binary approach, we cannot compute the exact p-value $\mathrm{P}(G_{f(\mathbf{W}_k)} \ge g_{f(\mathbf{W}_k)})$ of the test directly, but $\mathrm{P}(G_{f(\mathbf{W}_k)} \ge g_{f(\mathbf{W}_k)})$ can be approximated by $\mathrm{P}(G_{f(\widehat{\mathbf{W}}_k)} \ge g_{f(\mathbf{W}_k)})$, where $G_{f(\widehat{\mathbf{W}}_k)} = \sup_{t\in\mathbb{R}}|f(\widehat{\mathbf{W}}_k(t;\hat{\boldsymbol{\delta}}))|$. The p-value $\mathrm{P}(G_{f(\widehat{\mathbf{W}}_k)} \ge g_{f(\mathbf{W}_k)})$ is estimated by generating a large number ($\ge$ 1000) of realizations $\widehat{\mathbf{W}}_k(t;\hat{\boldsymbol{\delta}})$ and by computing the proportion of the $G_{f(\widehat{\mathbf{W}}_k)}$ greater than the largest value of $|f(\mathbf{W}_k(t;\hat{\boldsymbol{\delta}}))|$ over t. The function $f(\cdot)$ needs to satisfy that $f(\mathbf{0}) = 0$ and a monotonicity condition of the form, for every $|\mathbf{a}| < |\mathbf{b}|$, $|f(\mathbf{a})| < |f(\mathbf{b})|$. For multivariate comparisons "$<$" stands for the product order: $(a_1,\ldots,a_{c-1}) = \mathbf{a} < \mathbf{b} = (b_1,\ldots,b_{c-1})$ iff $a_1 < b_1,\ldots,a_{c-1} < b_{c-1}$, similarly $|\mathbf{a}|$ stands for $(|a_1|,\ldots,|a_{c-1}|)$. Details of the proof for the asymptotical equivalence of $f(\mathbf{W}_k)$ and $f(\widehat{\mathbf{W}}_k)$ is relegated to the appendix. We also show the proof of consistency of the supremum tests against any departures from Model (1).

There are several options available for the choice of function $f(\cdot)$. This article suggests the following simple choices:

$$\begin{aligned} \mathrm{sum}(\mathbf{W}) :=& \quad f(\mathbf{W}) = \textstyle\sum_{j=1}^{c-1}(W)_j \\ \max(\mathbf{W}) :=& \quad f(\mathbf{W}) = \max|\mathbf{W}| \\ \mathrm{prod}(\mathbf{W}) :=& \quad f(\mathbf{W}) = \textstyle\prod_{j=1}^{c-1}(W)_j \end{aligned}$$

where $(W)_j$ is the jth component in the vector $\mathbf{W}$.

In addition, the p-value of the test can be calculated in the same way as in the binary approach using a Bonferroni adjustment. We plot the observed multivariate residuals $\mathbf{r}$ (or $\mathbf{r}^*$) with the simulated realizations separated by rows to create $c-1$ plots, denoted by $(\mathbf{W}^m)_1$, …, $(\mathbf{W}^m)_{c-1}$ (or $(\mathbf{W}^*)_1$, …, $(\mathbf{W}^*)_{c-1}$). If the model is correct, the null hypothesis is accepted for each of the plots. We can adjust the significance level so that the overall Type I error rate is less than or equal to the sum of the individual error rates for all $c-1$ plots. It leads to another two diagnostic method denoted by Bonf($\mathbf{W}^m$) and Bonf($\mathbf{W}^*$). Table 1 gives a summary of all graphical diagnostic methods corresponding to the two approaches. Details of the simulation process and proofs are relegated to the appendix.

The multivariate generalization of the diagnostic approach proposed by Arbogast and Lin [1] and the associated results are new contributions of this article. The extension of the results to MGLM has not previously been developed in the literature. In the following, we discuss an example and conduct a simulation study illustrating the different diagnostics proposed in Sections 2 and 3.

4 Example

The Normative Aging Study (NAS) is a multidisciplinary longitudinal study of aging in men established by the Veteran's Administration in 1963. NAS subjects have reported for medical examination every 3 to 5 years. Though the study records data on a wide spectrum of variables, including several health related measures, dietary and behavioral exposures, exposure to metals, and, psychosocial events, our analysis focuses on exploring the relationship of fasting blood glucose (FBG) with two markers of systemic inflammation, namely, white blood cell count (*wbc*) and blood levels of C-reactive protein (*crp*) after controlling for age and smoking status. The measurements were taken during January, 2000 to December, 2004 and we consider only the last complete observation available on the subject in case multiple measurements were available on the same subject. The current dataset contains observations on 682 men in the age range of 48 to 93 years. FBG was categorized into three categories according to clinical definition of diabetes [12], with FBG $<$ 110mg/dl termed as normal (category 1), between 110 and 126 mg/dl termed as impaired fasting glucose (category 2) and $\geq$ 126mg/dl termed as diabetes (category 3). It has been suggested in the literature that oxidative stress-induced inflammatory response increases insulin resistance, resulting in hyperglycemia or elevated levels of FBG which in turn causes oxidative stress again [13]. Inflammation is known to be a risk factor for diabetes [14]. White blood cell count and C-reactive protein can be viewed as biomarkers of systemic inflammation and thus could potentially be associated with FBG levels, leading to this analysis.

We first try to fit a simple model that includes linear terms of the covariates *wbc*, *crp*, *age* and *smoking*. In this analysis the effect of *wbc* on FBG turns out to be marginally significant with p-value 0.0857 with fitted estimate of β as 0.041, *crp* is not significant with p-value 0.27 and fitted estimate of β as 0.094 (see Table 4). The interpretation of the fitted model, for example, in terms of the *wbc* effect is that given fixed values of all other covariates in the model, the odds of having fasting blood glucose towards higher end of the FBG scale with one unit increase in WBC are estimated to be $e^{0.041}$ or 1.04 times higher than having values on the lower end of the FBG scale. Neither age, nor smoking status was found to be associated with FBG levels. So there appears to be a positive association between FBG and *wbc* and *crp*, but none of them are statistically significant.

We used different diagnostic tools to check the model misspecification for *age*, *smoking*, *wbc* and *crp*. Table 2 shows the p-value corresponding to each of the graphical methods.

Figure 1 gives the plot using the method $(\mathbf{W}^m)_1$ for *wbc* while Figure 2 gives the same for *crp*. The dark black dashed line indicates the observed process and the fine solid lines indicate the simulated realizations. We calculate the p-value using 1000 simulated realizations, while the figure only show 100 of them due to the capacity of the image file. The p-value for testing that the model has a correct functional form in *wbc* is 0.055 whereas the p-value corresponding to right model specification in terms of *crp* is given by 0.108. The results suggest that there is certain degree of model misspecification for the proportional odds model with the covariate *wbc* and *crp* but not with the covariates *age* and *smoking*. The raw scatter plots of actual FBG measurements on a continuous scale (not included in the text) also indicated a non-linear relationship between FBG and *wbc* and *crp*. Since the correlation between *wbc* and *crp* in the original dataset were very weak (0.10) we treat the model specification issue in each predictor separately, which may not be optimal in every situation. We discuss joint multivariate extensions of the proposed method in our concluding discussion.

We re-fit the proportional odds model including a quadratic and cubic term of *wbc* and a quadratic term in *crp* in Table 4. The linear and quadratic terms are significant in *wbc* with the cubic term marginally significant. The linear term in *crp* is also significant in the new model. The results corresponding to *age* and *smoking* remain almost unchanged in the second model, with both being non-significant. Table 3 shows the p-value of each of the graphical diagnostic for the model including higher order powers of *wbc* and *crp*. The graphic diagnostics do not show model misspecification for the new model. Figure 3 shows the plot using the method $(\mathbf{W}^m)_1$ for the new model for *wbc* and Figure 4 shows the same for *crp* . The p-values are 0.51 and 0.761 respectively, indicating that the functional terms chosen in the final model is satisfactory. In terms of the actual FBG data on a continuous scale, it appears that there is a positive association between FBG and *crp* and *wbc* values for lower values of *crp* and *wbc*, below a certain threshold, but the relationship actually reverses or becomes less pronounced for higher extreme levels of these biomarkers, thus overall showing a non-linear pattern. There appears to be a non-linear threshold effect in the association between FBP with both *crp* and *wbc* when we analyzed the continuous FBG data as well.

5 Simulations

In the previous section, the article proposes two approaches including 9 graphical diagnostic methods to detect model inadequacy in the proportional odds model. To compare the performances of these methods, in this section we undertake a small-scale simulation study for investigating the power under a fixed alternative $\mathcal{H}_1$ and the Type I error rate under $\mathcal{H}_0$. We investigate two forms of functional misspecification in a single covariate X. We consider discrete X in one scenario and continuous in the other. For each situation, the empirical Type I error rate and powers are estimated based on the proportion of rejected null hypotheses in 10,000 simulated datasets.

Scenario 1: Let $c = 3$. We consider the true model as follows:

$$\text{logit}[P(Y \leq j \mid X)] = \alpha_j - \beta_1 X - \beta_2 X^2, \ j = 1, 2. \tag{4}$$

We first generate grouped categorical X observations with values ranging from -5 to +5 with equal probability, representing a discrete uniform distribution. Conditional on the X-values Y values are generated from Model (4) by choosing $\alpha_1 = -2$, $\alpha_2 = -1$, $\beta_1 = +0.25$ and $\beta_2 = 0.0, -0.05, -0.1$, and then simulating multinomial random variables with three categories. We generate 110 observations in each dataset, rendering approximately 10 occurrences for each distinct X-value on an average.

We try to fit a simple model with just the linear term to the simulated data with X^2 omitted, namely,

$$\text{logit}[P(Y \leq j \mid X)] = \alpha_j - \beta_1 X, \ j = 1, 2. \tag{5}$$

When $\beta_2 = 0.0$, the model is correctly specified and we can estimate the rejection rate under this H_0 and compare this estimate of Type I error rate with the significance level (α) which was always set at 0.05. When $\beta_2 = -0.05$, or -0.1, we evaluate the performance of the different graphical diagnostic methods by their power to detect departures from the correct model. Table 5 shows the results for this scenario in the first 3 columns. Among all the methods compared, the naive binary collapsing approach exhibits the worst performance. It fails to maintain the nominal Type I error level and the estimated Type I error rate is about twice the desired level of significance $\alpha\,(= 0.05)$. The multivariate approaches based on the residuals and the cumulative residuals produce better results. Both of the multivariate residuals ($\mathbf{r}$) and multivariate cumulative residuals ($\mathbf{r}^*$) maintain the correct level of significance under a correctly specified model with $\beta_2 = 0$. The power for the multivariate methods based on the functionals sum($\mathbf{W}^m$) and sum ($\mathbf{W}^*$) appear to be the best.

Scenario 2 The second scenario represents a situation where the cumulative logit probabilities associated with the response are related in a non-linear manner with X, but has a linear form in $\cos(X)$. The correct model is as follows:

$$\text{logit}(\Pr(Y \leq j \mid X)) = \alpha_j - \beta \cos X,\ \ j = 1,\, 2. \tag{6}$$

with $\alpha_1 = -1,\ \alpha_2 = 1$ and $\beta = 0, -1, -3$. We simulated X from a standard normal distribution and conditional on X simulated Y from the multinomial distribution with probabilities defined via (6). Again we fit each simulated dataset using the model (5) with a linear term of X. Table 5 shows the results in the last 3 columns. Similar to the first scenario, the binary collapsing approach gives a overly liberal result that rejects the null hypothesis more often than we expect and consequently has high power values. Among the methods in the multivariate approach, the sum($\mathbf{W}^*$) has the best performance in terms of maintaining Type I error and attaining high power values.

A goodness-of-fit statistics as proposed in Lipsitz *et al.* [6] based on the mean score are also included in the simulation study for comparison purposes. According to percentiles of the predicted mean score, subjects are partitioned into G regions as defined in Lipsitz *et al.* [6]. Given the partition of the data, the following model is fitted

$$\operatorname{logit}[\Pr(Y \leq j \mid \mathbf{X})] = \alpha_j - \boldsymbol{\beta}^T\mathbf{X} + \sum_{g=1}^{G-1} I_{ig}\gamma_g \tag{7}$$

where I_{ig} are group indicators with $I_{ig} = 1$, if $\mathbf{s}^T\hat{\boldsymbol{\pi}}_i$ is in region g and $I_{ig} = 0$ otherwise for some score $\mathbf{s}$. If model (5) is correct, then $\gamma_1 = \gamma_2 = \cdots = \gamma_{G-1} = 0$ independently of the chosen regions and scores. We simply test $\mathcal{H}_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_{G-1} = 0$ and compute a likelihood-ratio (LR), Wald and a score statistic. We refer to this statistic as Hosmer-Lemeshow (HL)-type statistic, because the idea stems from the HL statistic developed for logistic regression as extended to ordinal responses. The LR test, the Wald-test and the score test in this case are asymptotically equivalent and showed quite similar power values, hence, Table 5 only lists the result of the HL-type score tests.

For the first scenario, Table 5 also gives the Wald test on the null hypothesis $\mathcal{H}_0$: $\beta_2 = 0$. If we do know that the correct model includes the X^2 term, this test is optimal as one would expect, but the Wald test is not applicable when the true functional form is unknown. Thus in situation 2, we cannot formulate an appropriate Wald test to compare the two models in terms of a single parameter.

Summary of Simulation Results: In general, the graphical diagnostic methods based on sum($\mathbf{W}$) and prod($\mathbf{W}$) have good power properties. We do expect the graphical diagnostic methods to provide a lower power compared to the Wald test when the true model contains the term X^2 as in Scenario 1. Unlike the Wald test, the graphical diagnostic methods do not focus on any specific term. It checks model misspecification for a wide range of the misspecification in a non-parametric manner (e. g., the functional form could be anything like X^2, $\log X$, X^3, cos X etc). Arbogast and Lin [1] also pointed out that the Wald test cannot be used to check whether the chosen functional term is satisfactory, which can be achieved in our graphical approach. Remarkably, some of the graphical diagnostic methods are very comparable with the optimal Wald test in terms of power for

Scenario 1, when one is testing for the missing term in the true model, with a true model known. For example, the sum($\mathbf{W}^*$) gives the power of 0.947 when the true coefficient of X^2 is -0.10. The Wald test gives a power of 0.994 in comparison. On the other hand, the graphical methods of "Bonf", "sum" and "prod" using the cumulative residuals ($\mathbf{r}^*$) in the multivariate approach have higher power than the overall Hosmer-Lemeshow test in scenario 1. The methods with "sum" and "prod" using the cumulative residuals ($\mathbf{r}^*$) still give higher power than the overall Hosmer-Lemeshow test in Scenario 2. The diagnostic based on sum($\mathbf{W}^*$) appear to be the best choice based on our limited simulation settings.

6 Discussion

This article proposes graphical diagnostic methods based on two approaches to test model misspecification for the proportional odds regression models. In the naive binary approach, we treat the proportional odds model as $c-1$ collapsed logistic regression models. Using the cumulative sums of residuals, the graphical diagnostic method extends previously introduced techniques by Arbogast and Lin [1]. However, according to the simulations, it is more appropriate to treat the residuals in a multivariate format as in the second approach and then consider a vector of stochastic processes to represent the limiting behavior of the residuals. In this way, the asymptotic Gaussian processes ($\widehat{\mathbf{W}}_k$) take the correlation between the ordinal responses into account which is ignored in the binary approach.

In the multivariate approach, both the multivariate residuals ($\mathbf{r}$) and the cumulative residuals ($\mathbf{r}^*$) perform better than the binary approach but cumulative residuals outperform the multivariate residuals in our simulation study. For instance, in both simulation scenarios, the methods based on $\mathbf{r}^*$ are better in terms of power than the ones based on $\mathbf{r}$, while maintaining nominal error levels. Furthermore, among the different choices for the functions to combine the components of a vector, $f(\cdot)$, the "sum" tends to be the best among the ones we considered, in most of our simulations.

Lin *et al.* [9] noted that the tests are slightly more powerful when the process has the form

$$W_k^{(j)}(t;\hat{\boldsymbol{\delta}}) = n^{-1/2}\sum_{i=1}^{n} \hat{r}_{ij}^{*} I(t-b < X_{ik} \le t),$$

where b covers the lower half-plane of the covariates. In our large amount of simulations that there is not space to report, including b doesn't give consistently higher power. In general, we suggest taking $b=\infty$.

Following Lin *et al.* [9], we can extend our method and consider a vector-valued stochastic process to check the functional form of a set of multivariate covariates

$$\mathbf{W}_o^m(t;\hat{\boldsymbol{\delta}}) = n^{-1/2}\sum_{i=1}^{n} \mathbf{I}(\mathbf{X}_i \le \mathbf{t})\hat{\mathbf{r}}_i,$$

where $\mathbf{I}(\mathbf{X}_i \le \mathbf{t})$ is a diagonal matrix with $I(\mathbf{X}_{ij} \le \mathbf{t})$ as the jth entry on the diagonal. The $\mathbf{W}_o^m(t;\hat{\boldsymbol{\delta}})$ converges weakly to a vector of zero-mean Gaussian processes. The distribution of the processes can be approximated by $\widehat{\mathbf{W}}_o^m(t;\hat{\boldsymbol{\delta}})$, which has the same form as $\widehat{\mathbf{W}}_k^m(t;\hat{\boldsymbol{\delta}})$ by replacing $I(X_{ik} \le t)$ with $\mathbf{I}(\mathbf{X}_i \le \mathbf{t})$. Similarly, we can consider $\mathbf{W}_o^*(t;\hat{\boldsymbol{\delta}})$ as well.

For a broad range of applications, we can use $\widehat{\mathbf{W}}_k(t;\hat{\boldsymbol{\delta}})$ to a general multivariate generalized linear model and then use a function to combine the components of the multivariate residuals (or processes). These methods provide a good alternative to check the model fit and whether the chosen functional term is satisfactory. Simulation studies indicate they have power advantages compared to standard Hosmer-Lemeshow type partition-based statistic.

To conclude, in clinical investigations, as in the Normative Aging Study example, investigators are often misled about the true nature of association between a predictor and a response due to fitting an incorrect model. For categorical responses, the task is even more daunting as there is no clear mandate about a single goodness of fit statistic. These simple graphical tools may provide us better insight into the inadequacies of the fitted model in such situations. The pattern in these plots may suggest alternative functional terms to include. How to extend these tools to correlated ordinal responses is an interesting avenue for possible research [19].

The appendix describes the computational details for simulating observations from the limiting Gaussian Processes. R-Codes for creating the diagnostic plots and simulation study is available at http://www.sph.umich.edu/bhramar/public_html/research.

ACKNOWLEDGMENTS:

We thank Drs. Howard Hu and Joel Schwartz for their comments. The work of Ivy Liu and Thomas Suesse was supported by University Research Grant from Victoria University. This work was also supported in part by NIEHS R01-ES05257, NIEHS R01-ES07821, and NIEHS R01-ES015172. The VA Normative Aging Study is supported by the Cooperative Studies Program/Epidemiology Research and Information Center of the US Department of Veterans Affairs and is a component of the Massachusetts Veterans Epidemiology Research and Information Center, Boston, Massachusetts. The research of Bhramar Mukherjee was partially supported by NSF DMS grant 07-06935.

APPENDIX

A1. Proof of Asymptotical Equivalence of $f(\mathbf{W}_k)$ and $f(\widehat{\mathbf{W}}_k)$:

In the Appendix, the notation $\mathbf{W}_k$ might either stand for $\mathbf{W}_k^m$ or $\mathbf{W}_k^*$ and used as a generic representation for both. Let us define the univariate process $W_k := \text{sum}(\mathbf{W}_k)$ and also the estimated process. $\widehat{W}_k := \text{sum}(\widehat{\mathbf{W}}_k)$. The method of generalized estimation equations (GEE) is commonly used for marginal models with dependent observations. Lin *et al.* [9] showed that $W_k(t; \hat{\boldsymbol{\delta}})$ converges under the marginal model for dependent observations to a zero mean Gaussian process and is asymptotically equivalent to $\widehat{W}_k(t; \hat{\boldsymbol{\delta}})$. In the proportional odds model (1), the Y_{ij} is the response on level j for the ith subject. We can re-consider Y_{ij} as the response at the jth occasion for the ith subject in a longitudinal setting and then use the result given by Lin *et al.* [9] to prove the asymptotical equivalence of $f(\mathbf{W}_k)$ and $f(\widehat{\mathbf{W}}_k)$.

According to Lin *et al.* [9], first, we know that $W_k(t; \hat{\boldsymbol{\delta}})$ converges under model (1) to a zero mean Gaussian process and is asymptotically equivalent to $\widehat{W}_k(t;\hat{\boldsymbol{\delta}})$. Note the process W_k can be expressed in terms of the multivariate process $\mathbf{W}_k$ by $W_k(t;\boldsymbol{\delta}) = \mathbf{1}_{c-1}^T \mathbf{W}_k(t;\boldsymbol{\delta})$, where $\mathbf{1}_{c-1}$ is a vector of ones of length $c-1$. Similarly we have $\widehat{W}_k(t;\boldsymbol{\delta}) = \mathbf{1}_{c-1}^T \widehat{\mathbf{W}}_k(t;\boldsymbol{\delta})$

We want to show that $\mathbf{W}_k$ converges to a zero mean Gaussian process under model (1) and that $\widehat{\mathbf{W}}_k$ is asymptotically equivalent to $\mathbf{W}_k$. According to the proposition in Andrews [15] on page 2251, we need to show: (i) for every finite set of indices $\{t_1,\ldots,t_m\}$: $\Big(\mathbf{W}_k(t_1;\hat{\boldsymbol{\delta}}),\ldots,\mathbf{W}_k(t_m;\hat{\boldsymbol{\delta}})\Big)$ (also for $\widehat{\mathbf{W}}_k$) converges to a zero mean multivariate normal distribution, and (ii) $\mathbf{W}_k$ is equicontinuous. Let $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_{c-1})$ be arbitrary but fixed with $\|\boldsymbol{\lambda}\|_2 = 1$, where $\|\cdot\|_2$ is the Euclidean norm. Now let the univariate process W_k be defined in terms of $\tilde{\mathbf{Y}}_i := \mathrm{Diag}(\boldsymbol{\lambda}) \cdot \mathbf{Y}_i$ and not in terms of $\mathbf{Y}_i$. In other words, the underlying random variable of W_k is scaled according to $\boldsymbol{\lambda}$, but the underlying variable of $\mathbf{W}_k$ stays unscaled. Previously (without scaling) we had $W_k(t;\boldsymbol{\delta}) = \mathbf{1}_K^T \mathbf{W}_k(t;\boldsymbol{\delta})$ and $\widehat{W}_k(t;\boldsymbol{\delta}) = \mathbf{1}_K^T \widehat{\mathbf{W}}_k(t;\boldsymbol{\delta})$, whereas now we can show

$$W_k(t;\boldsymbol{\delta}) \equiv \boldsymbol{\lambda}^T \mathbf{W}_k(t;\boldsymbol{\delta}) \text{ and } \widehat{W}_k(t;\boldsymbol{\delta}) \equiv \boldsymbol{\lambda}^T \widehat{\mathbf{W}}_k(t;\boldsymbol{\delta}) \tag{8}$$

for fixed $\boldsymbol{\lambda}$.

Thus $W_k(t;\hat{\boldsymbol{\delta}})$ and $\widehat{W}_k(t;\hat{\boldsymbol{\delta}})$ converge to a zero mean Gaussian process [9] and from the aforementioned proposition (see Reference [15], p. 2251) it follows that $W_k(t;\hat{\boldsymbol{\delta}}) \equiv \boldsymbol{\lambda}^T \mathbf{W}_k$ and $\widehat{W}_k(t;\hat{\boldsymbol{\delta}}) \equiv \boldsymbol{\lambda}^T \widehat{\mathbf{W}}_k$ are equicontinuous and for all finite sets $\{t_1,\ldots,t_m\}$: $\Big(W_k(t_1;\hat{\boldsymbol{\delta}}),\ldots,W_k(t_m;\hat{\boldsymbol{\delta}})\Big) \equiv \Big(\boldsymbol{\lambda}^T \mathbf{W}_k(t_1;\hat{\boldsymbol{\delta}}),\ldots,\boldsymbol{\lambda}^T \mathbf{W}_k(t_m;\hat{\boldsymbol{\delta}})\Big)$ (also for $\hat{\mathbf{W}}_k$) converges to a zero mean multivariate normal distribution. Now we apply the Cramer-Wald theorem (if for fixed $\boldsymbol{\lambda}$ the random variable $\boldsymbol{\lambda}^T \mathbf{W}_k(t_j;\hat{\boldsymbol{\delta}})$ converges in distribution to $\boldsymbol{\lambda}^T \mathbf{W}$, then $\mathbf{W}_k(t_j;\hat{\boldsymbol{\delta}})$ converges in distribution to $\mathbf{W}$) and it immediately follows (i). One can show: if $\boldsymbol{\lambda}^T \mathbf{W}_k$ and $\boldsymbol{\lambda}^T \widehat{\mathbf{W}}_k$ are equicontinuous, then also $\mathbf{W}_k$ and $\widehat{\mathbf{W}}_k$ are equicontinuous, that is (ii). From (i) and (ii), the asymptotic equivalence of W_k and $\widehat{W}_k$ with (8) follows that the multivariate processes $\mathbf{W}_k$ and $\widehat{\mathbf{W}}_k$ converge to the same zero mean multivariate Gaussian process. Applying a continuous function $f(\cdot)$ to these processes, $f(\mathbf{W}_k)$ and $f(\widehat{\mathbf{W}}_k)$ converge to the same stochastic process (not necessarily Gaussian) by the continuous mapping theorem.

A2. Proof of Consistency of the Supremum Tests:

The consistency of similar supremum tests was shown/mentioned in several papers [16, 17, 9, 18, 1]. It was shown that under certain sufficient conditions $n^{-1/2}W_k(t_0;\hat{\boldsymbol{\delta}}) \rightarrow_p c \neq 0$ for at least some t_0, hence, $n^{-1/2}G_{W_k}$ converges to a nonzero constant.

We want to show now the consistency of $G_{f(\mathbf{W}_k)}$. First, we show that $n^{-1/2}(\mathbf{W}_k)_j$ converges to a non-zero constant c_j. As before we use (8) and set $\boldsymbol{\lambda} := \mathbf{e}_j$, where $\mathbf{e}_j$ is the jth unit vector. We now have $W_k \equiv \mathbf{e}_j^T\mathbf{W}_k = (\mathbf{W}_k)_j$. From the above, we can conclude that $n^{-1/2}W_k \rightarrow_p c_j \neq 0$, or equivalently $n^{-1/2}(\mathbf{W}_k)_j \rightarrow_p c_j \neq 0$.

To show that the test $G_{f(\mathbf{W}_k)}$ is consistent, it is sufficient to show $n^{-1/2}f(\mathbf{W}_k)$ converges to a nonzero vector for some t_0 (then $n^{-1/2}G_{f(\mathbf{W}_k)}$ converges to a nonzero constant). We just established that $n^{-1/2}\mathbf{W}_k \rightarrow_p \mathbf{c}$ with $\mathbf{c}$ being nonzero in all components. Thus, $n^{-1/2}f(\mathbf{W}_k) \rightarrow_p f(\mathbf{c})$. We have $\mathbf{0} < |\mathbf{c}|$ and it follows from the monotonicity condition $\mathbf{0} = |f(\mathbf{0})| < |f(\mathbf{c})|$, which was to be shown. □

A3. Simulating observations from the Gaussian Processes: Computational Details

Given the parameter estimates obtained from the dataset after fitting the proportional odds model, the computation of the $\mathbf{W}_k$'s is relatively easy. The vector of residuals $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)^T$ is a byproduct of the fitting process and the computation of the $\mathbf{W}_k$ only requires the computation of the unknown indicator functions $I(X_{ik} \leq t)$. We do not need to compute $I(X_{ik} \leq t)$ for infinitely many t's, but only for the number $m \leq n$ of different values $t_1, \ldots, t_m$ corresponding to the kth covariate. We can store all these $I(X_{ik} \leq t)$ in a $n \times m$ matrix $\mathbf{I}(\mathbf{X}_k)$. For given $\mathbf{r}$ and $\mathbf{I}(\mathbf{X}_k)$, the computation of $\mathbf{W}_k$ simply requires matrix operations.

The computation of the $\widehat{\mathbf{W}}_k$'s is much more intensive, because we need to resample a large number $M \geq 1000$ of realizations from the $\widehat{\mathbf{W}}_k$'s. Again, as a byproduct from the fitting algorithm we obtain $\boldsymbol{\Omega}$, $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n)^T$ and $\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\delta}} = (\frac{\partial \boldsymbol{\pi}_1}{\partial \boldsymbol{\delta}}, \ldots, \frac{\partial \boldsymbol{\pi}_n}{\partial \boldsymbol{\delta}})$. With

$\mathbf{I}(\mathbf{X}_k)$ and $\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\delta}}$ we can then compute $\boldsymbol{\eta}(t_1), \ldots, \boldsymbol{\eta}(t_m)$. In the definition of the $\widehat{\mathbf{W}}_k$'s, which have the form $\sum_{i=1}^n [\ldots]_i Z_i$, the quantities in the bracket terms $[\ldots]_i$ can be computed by matrix operations and can be stored in a $n \times (c-1) \times m$ array $\mathbf{B}$. Now we generate M times the n realizations $Z_1, \ldots, Z_n$ from $N(0,1)$ and then store in the $M \times n$ matrix $\mathbf{Z}$. Finally, we can compute the $\widehat{\mathbf{W}}_k$'s from $\mathbf{Z}$ and $\mathbf{B}$ by matrix multiplication and to avoid any loops by a tensor product. Also note, that $\mathbf{W}_k^* = \mathbf{A}\mathbf{W}_k^m$ with simple pre multiplication by a given matrix $\mathbf{A}$ and similarly $\widehat{\mathbf{W}}_k^* = \mathbf{A}\widehat{\mathbf{W}}_k^m$. In fact, for the multivariate approach, we only need to compute $\mathbf{W}^m$ and the $\widehat{\mathbf{W}}^m$'s and the analogues with the cumulative residuals are obtained easily. Given these processes, all other quantities involved in our diagnostic methods can be computed easily.

REFERENCES

1. Arbogast P, Lin D. Model-checking techniques for stratified case-control studies. *Statistics in Medicine* 2005: **24**:229-247.

2. McCullagh P. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B* 1980: **42**:109-142.

3. Agresti A. *Categorical Data Analysis, 2nd ed.* Wiley: New Jersey, 2002.

4. Liu I, Agresti A. The analysis of ordered categorical data: an overview and a survey of recent developments. *Test* 2005: **14**:1-73.

5. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics* 1980: **A10**:1043-1069.

6. Lipsitz SR, Fitzmaurice GM, Molenberghs G. Goodness-of-fit tests for ordinal response regression models. *Applied Statistics* 1996: **45**:175-190.

7. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* 1996: **15**:1807-1826.

8. Kim J-H. Assessing practical significance of the proportional odds assumption. *Statistics and Probability Letters* 2003: **65**:233-239.

9. Lin DY, Wei LJ, YING Z. Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika* 1993: **80**: 557-572.

10. Fahrmeir L, Tutz G. *Multivariate Statistical Modeling based on Generalized Linear Models, 2nd ed.* Springer: New York, 2001.

11. Bell B, Rose CL, Damon A. The Veterans Administration longitudinal study of healthy aging. *Gerontologist* 1966: **6**:179-184.

12. The Expert committee on the diagnosis and classification of diabetes mellitus. Report of the Expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care* 1997: **20**:1183-1194.

13. Pliquett RU, Fasshauer M, Blüher M, Paschke R. Neurohumoral stimulation in type-2-diabetes as an emerging disease concept. *Cardiovascular diabetology* 2004: **3**:4.

14. Nakanish S, Yamane K, Kamei N, Okubo M, Kohno N. Elevated C-Reactive protein is a risk factor for the development of type-2 diabetes in Japanese Americans. *Diabetes Care* 2003: **26**:2754-2757.

15. Andrews D. Empirical Process Methods in Econometrics. *Handbook of Econometrics* 1994: **4**:2247-2294.

16. Su JQ, Wei LJ. A lack-of-fit test for the mean function in a generalized linear-model. *Journal Of The American Statistical Association* 1991: **86**:420-426.

17. Lin DY, Wei LJ, Ying Z. Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika* 1993: **80**:557-572.

18. Pan ZY, Lin DY. Goodness-of-fit methods for generalized linear mixed models. *Biometrics* 2005: **61**:1000-1009.

19. Pan, W. Goodness-of-fit tests for GEE with correlated binary data. *Scandinavian Journal of Statistics* 2002: **29**:101-110.

7 Tables

Table 1: Notations used for graphical diagnostic methods

Notation	Approach	Description
B_j	Binary	Collapse the response categories into $(\leq j, > j)$
Bonf(B)	Binary	Bonferroni adjustment: compare the p-value with $\alpha/(c-1)$
$(\mathbf{W}^m)_j$	Multivariate ($\mathbf{r}$)	Using the jth component of residual $\mathbf{r}$
Bonf($\mathbf{W}^m$)	Multivariate ($\mathbf{r}$)	Bonferroni adjustment: compare the p-value with $\alpha/(c-1)$
sum($\mathbf{W}^m$)	Multivariate ($\mathbf{r}$)	Using function $\text{sum}(\mathbf{W}^m) := \sum_{j=1}^{c-1}(W^m)_j$
prod($\mathbf{W}^m$)	Multivariate ($\mathbf{r}$)	Using function $\text{prod}(\mathbf{W}^m) := \prod_{j=1}^{c-1}(W^m)_j$
max($\mathbf{W}^m$)	Multivariate ($\mathbf{r}$)	Using function $\max(\mathbf{W}^m) := \max\lvert\mathbf{W}^m\rvert$
$(\mathbf{W}^*)_j$	Multivariate ($\mathbf{r}^*$)	Using the jth component of residual $\mathbf{r}^*$
Bonf($\mathbf{W}^*$)	Multivariate ($\mathbf{r}^*$)	Bonferroni adjustment: compare the p-value with $\alpha/(c-1)$
sum($\mathbf{W}^*$)	Multivariate ($\mathbf{r}^*$)	Using function $\text{sum}(\mathbf{W}^*) := \sum_{j=1}^{c-1}(W^*)_j$
prod($\mathbf{W}^*$)	Multivariate ($\mathbf{r}^*$)	Using function $\text{prod}(\mathbf{W}^*) := \prod_{j=1}^{c-1}(W^*)_j$
max($\mathbf{W}^*$)	Multivariate ($\mathbf{r}^*$)	Using function $\max(\mathbf{W}^*) := \max\lvert\mathbf{W}^*\rvert$

Table 2: The p-values of testing model misspecification based on graphical diagnostics for model "age+smk+wbc+crp"

Tests	*age*	*smk*	*wbc*	*crp*	α (Bonferroni adjustment)
B_1	0.139	0.864	0.056	0.096	0.05 (0.025)
B_2	0.145	0.191	0.838	0.643	0.05 (0.025)
$(\mathbf{W}^m)_1$	0.175	0.981	0.055	0.108	0.05 (0.025)
$(\mathbf{W}^m)_2$	0.545	0.766	0.133	0.298	0.05 (0.025)
$(\mathbf{W}^*)_1$	0.175	0.981	0.055	0.108	0.05 (0.025)
$(\mathbf{W}^*)_2$	0.352	0.735	0.821	0.791	0.05 (0.025)
sum($\mathbf{W}^m$)	0.352	0.735	0.821	0.791	0.05
max($\mathbf{W}^m$)	0.299	0.799	0.069	0.188	0.05
prod($\mathbf{W}^m$)	0.233	0.898	0.047	0.122	0.05
sum($\mathbf{W}^*$)	0.332	0.866	0.193	0.209	0.05
max($\mathbf{W}^*$)	0.235	0.829	0.059	0.156	0.05
prod($\mathbf{W}^*$)	0.304	0.887	0.323	0.308	0.05

Table 3: The p-values of testing model misspecification based on graphical diagnostics for model "age+smk+wbc+wbc^2+wbc^3+crp+crp^2"

Tests	*age*	*smk*	*wbc*	*crp*	α (Bonferroni adjustment)
B_1	0.262	0.774	0.497	0.662	0.05 (0.025)
B_2	0.249	0.148	0.125	0.071	0.05 (0.025)
$(\mathbf{W}^m)_1$	0.114	0.961	0.510	0.761	0.05 (0.025)
$(\mathbf{W}^m)_2$	0.543	0.875	0.532	0.678	0.05 (0.025)
$(\mathbf{W}^*)_1$	0.114	0.961	0.510	0.760	0.05 (0.025)
$(\mathbf{W}^*)_2$	0.334	0.712	0.347	0.231	0.05 (0.025)
sum($\mathbf{W}^m$)	0.334	0.712	0.347	0.231	0.05
max($\mathbf{W}^m$)	0.235	0.914	0.696	0.811	0.05
prod($\mathbf{W}^m$)	0.196	0.943	0.679	0.699	0.05
sum($\mathbf{W}^*$)	0.344	0.745	0.255	0.267	0.05
max($\mathbf{W}^*$)	0.169	0.818	0.581	0.376	0.05
prod($\mathbf{W}^*$)	0.428	0.940	0.225	0.299	0.05

Table 4: Parameter estimates and p-values for the fitted proportional odds model using covariates "age+smk+wbc+crp" (Model 1) followed by the model "age+smk+wbc+wbc^2+wbc^3+crp+crp^2" (Model 2) in the Normative Aging Study Example.

Model	Predictor	Coef	S.E.	Wald Z	P-value
Model 1	age	-0.00747	0.01255	-0.60	0.5516
	smk	0.03331	0.06186	0.54	0.5902
	wbc	0.04134	0.02406	1.72	0.0857
	crp	0.09408	0.08572	1.10	0.2724
Model 2	age	-0.0080846	0.0126358	-0.64	0.5223
	smk	0.0442334	0.0624535	0.71	0.4788
	wbc	0.5628662	0.2464199	2.28	0.0224
	wbc^2	-0.0376317	0.0192671	-1.95	0.0508
	wbc^3	0.0005244	0.0002956	1.77	0.0760
	crp	0.3960383	0.1821148	2.17	0.0297
	crp^2	-0.0297128	0.0198322	-1.50	0.1341

Table 5: Simulation results

	The functional form of $\boldsymbol{\beta}^T\mathbf{X}$ in the true model					
	$0.25X + \beta X^2$			$\beta\cos(X)$		
Methods	$\beta = 0.00$	$\beta = -0.05$	$\beta = -0.10$	$\beta = 0.0$	$\beta = -1.0$	$\beta = -3.0$
B_1	0.148	0.435	0.934	0.168	0.393	0.981
B_2	0.097	0.482	0.959	0.155	0.407	0.822
Bonf(B)	0.126	0.491	0.964	0.180	0.433	0.969
Bonf($\mathbf{W}^m$)	0.042	0.220	0.811	0.046	0.129	0.879
sum($\mathbf{W}^m$)	0.051	0.285	0.855	0.052	0.179	0.591
prod($\mathbf{W}^m$)	0.054	0.113	0.386	0.058	0.112	0.704
max($\mathbf{W}^m$)	0.035	0.102	0.543	0.056	0.086	0.836
Bonf($\mathbf{W}^*$)	0.043	0.292	0.895	0.049	0.191	0.906
sum($\mathbf{W}^*$)	0.048	0.357	0.947	0.049	0.344	0.974
prod($\mathbf{W}^*$)	0.047	0.340	0.941	0.049	0.270	0.958
max($\mathbf{W}^*$)	0.041	0.266	0.874	0.051	0.203	0.939
Wald-$\beta = 0$	0.050	0.568	0.994	-	-	-
HL (G=5)	0.049	0.278	0.894	0.046	0.255	0.949

Figure 1: Plot of residuals against *wbc* using the method $(\mathbf{W}^m)_1$ to check the model misspecification for *wbc* in the model of "age+smk+wbc+crp". The dark black line indicates the observed process and the fine lines indicate the simulated realizations.

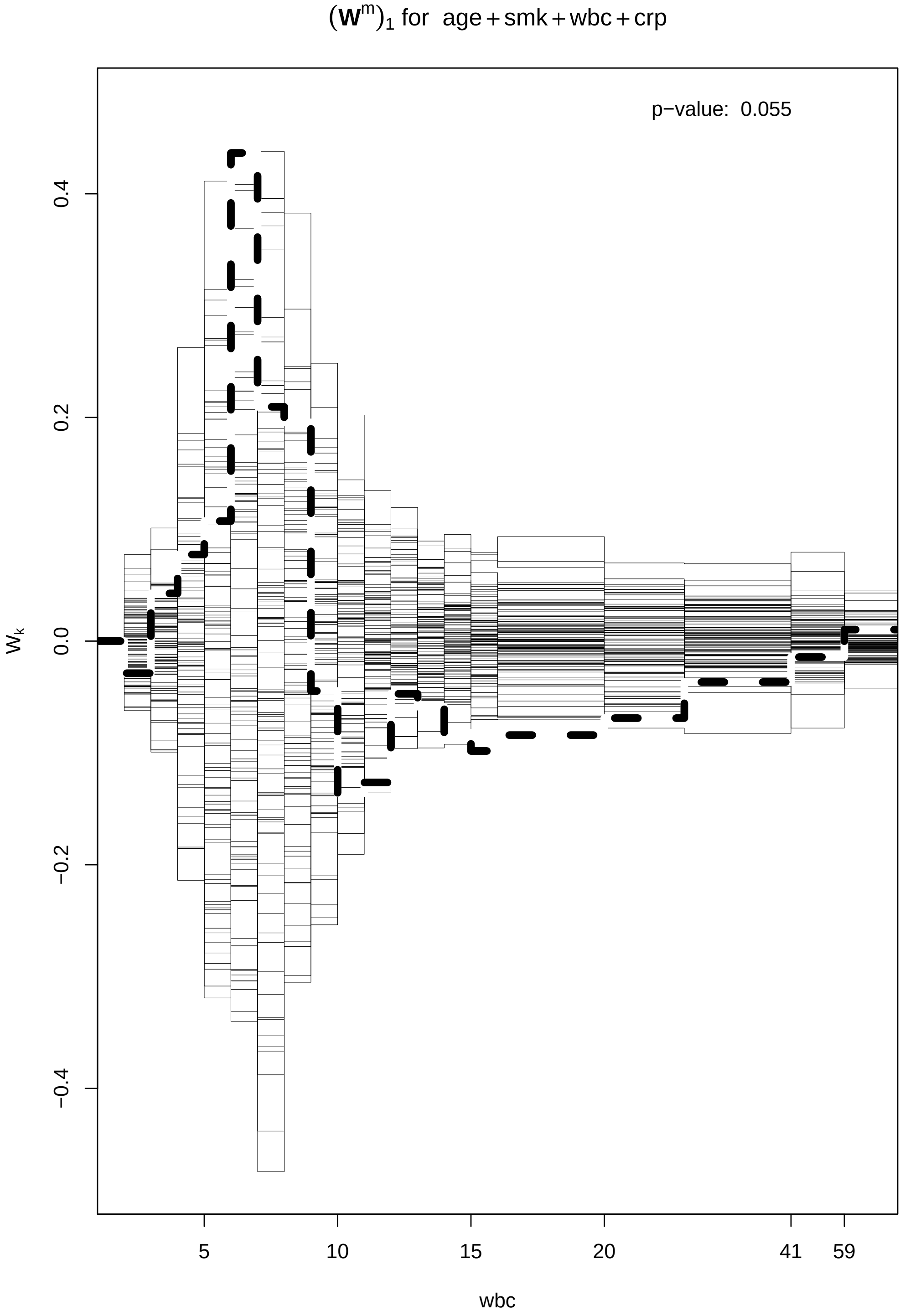


Figure 2: Plot of residuals against *crp* using the method $(\mathbf{W}^m)_1$ to check the model misspecification for *crp* in the model of “age+smk+wbc+crp”. The dark black line indicates the observed process and the fine lines indicate the simulated realizations.

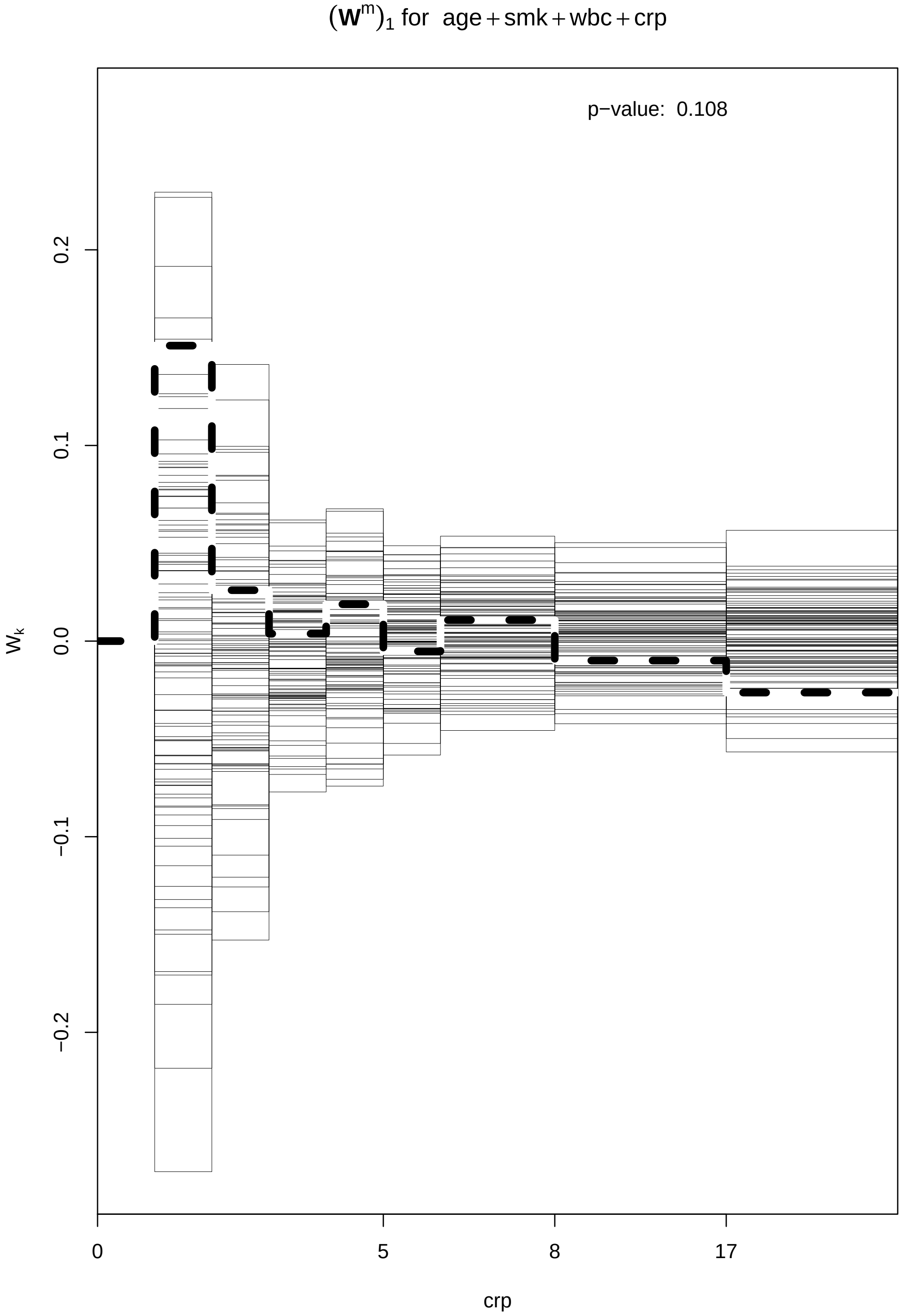


Figure 3: Plot of residuals against *wbc* using the method $(\mathbf{W}^m)_1$ to check the model misspecification for *wbc* in the model of "age+smk+wbc+wbc^2+wbc^3+crp+crp^2". The dark black line indicates the observed process and the fine lines indicate the simulated realizations.

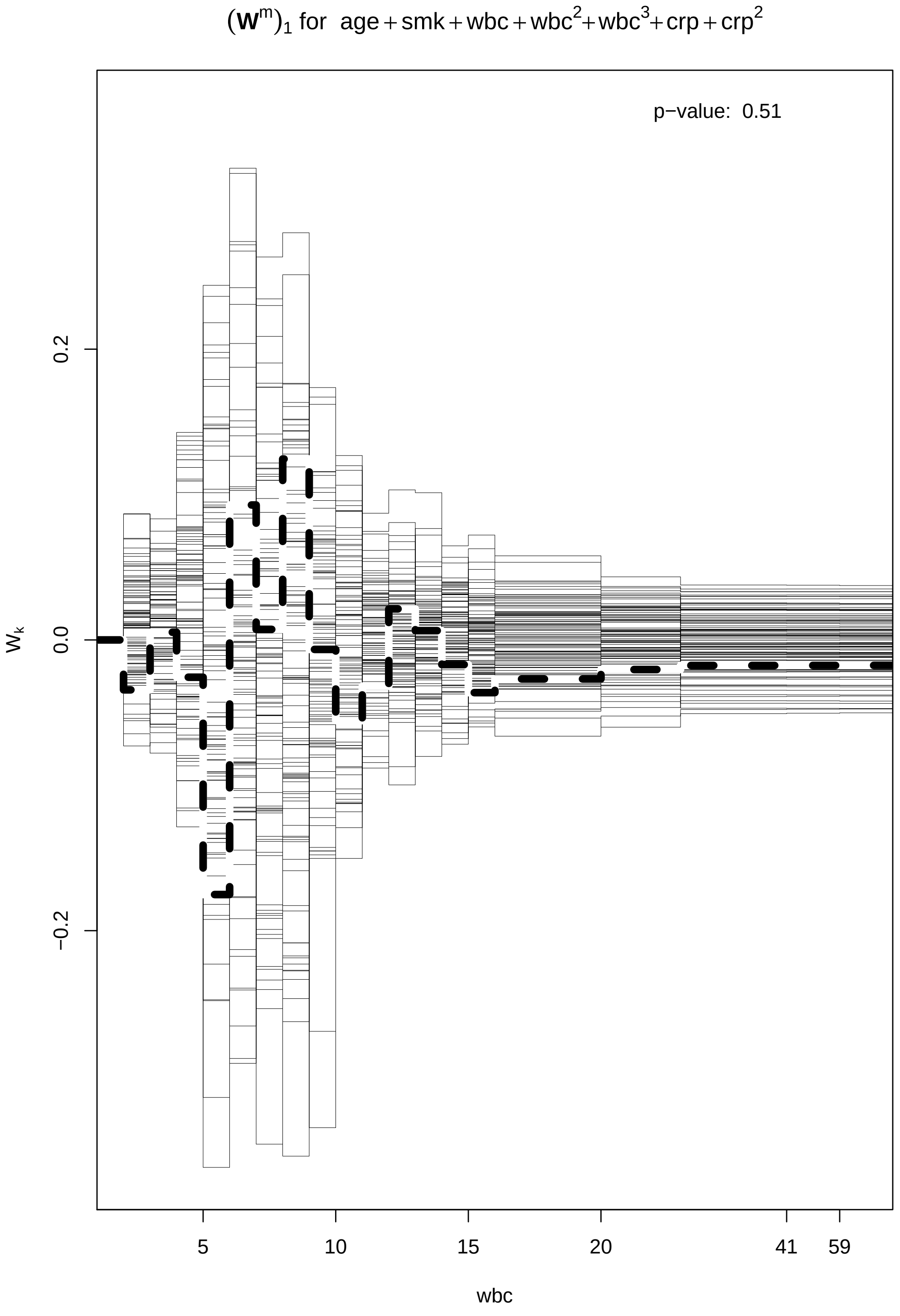
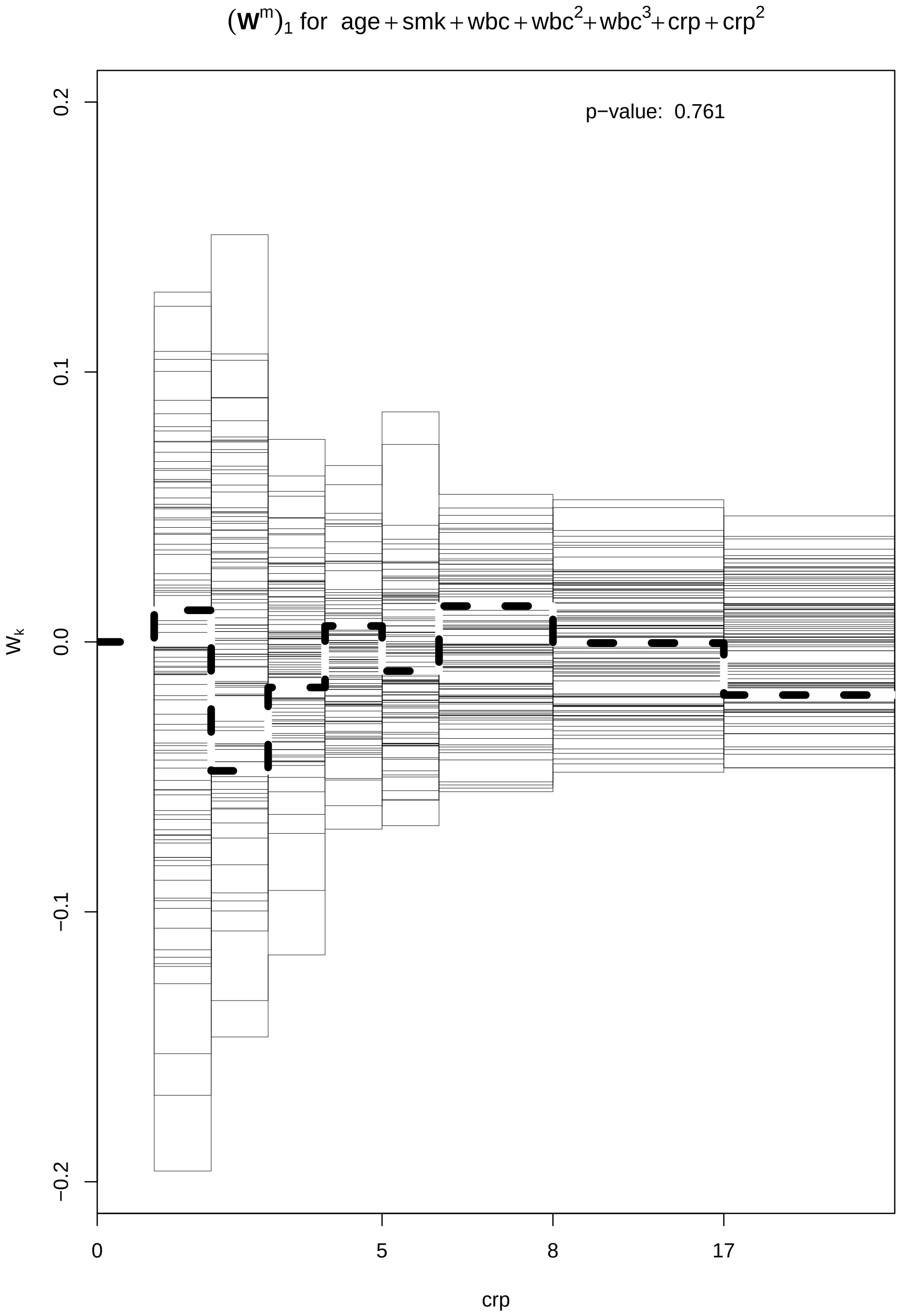


Figure 4: Plot of residuals against *crp* using the method $(\mathbf{W}^m)_1$ to check the model misspecification for *crp* in the model of "age+smk+wbc+wbc^2+wbc^3+crp+crp^2". The dark black line indicates the observed process and the fine lines indicate the simulated realizations.


$(\mathbf{W}^m)_1$ for age + smk + wbc + wbc^2+wbc^3+ crp + crp^2
p−value: 0.761
W_k
crp