

5-1-2011

Model Diagnostics for Proportional and Partial Proportional Odds Models

Ann A. O'Connell

The Ohio State University, aoconnell@ehe.osu.edu

Xing Liu

Eastern Connecticut State University, liux@easternct.edu

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

O'Connell, Ann A. and Liu, Xing (2011) "Model Diagnostics for Proportional and Partial Proportional Odds Models," *Journal of Modern Applied Statistical Methods*: Vol. 10 : Iss. 1 , Article 15.

DOI: 10.22237/jmasm/1304223240

Model Diagnostics for Proportional and Partial Proportional Odds Models

Ann A. O’Connell
The Ohio State University,
Columbus, OH USA

Xing Liu
Eastern Connecticut State University,
Willimantic, CT USA

Although widely used to assist in evaluating the prediction quality of linear and logistic regression models, residual diagnostic techniques are not well developed for regression analyses where the outcome is treated as ordinal. The purpose of this article is to review methods of model diagnosis that may be useful in investigating model assumptions and in identifying unusual cases for PO and PPO models, and provide a corresponding application of these diagnostic methods to the prediction of proficiency in early literacy for children drawn from the kindergarten cohort of the Early Childhood Longitudinal Study (ECLS-K; NCES, 2000).

Key words: Model diagnostics, proportional odds models, partial proportional odds models, residual analyses.

Introduction

Although widely used to assist in evaluating the prediction quality of linear and logistic regression models, residual diagnostic techniques are not well developed for regression analyses where the outcome is treated as ordinal. For ordinal regression models, Hosmer and Lemeshow (2000) suggested recombining outcomes according to the ordinal structure of the data and applying residual strategies developed for binary logistic models, such as outlined in Pregibon (1981). This approach is useful in the investigation of the assumption of proportionality as well as for examination of unusual or extreme values via residual diagnostics and this article presents guidelines for proportional odds (PO) and non- or partial-proportional odds (PPO) models.

Residual analyses provide rich opportunities for researchers to examine model fit and misfit, and require going beyond the results obtained through a direct application of a statistical model, the interpretation of parameter estimates or the summary statistics obtained from that model. Studies of residuals are becoming an important analytic process in many research situations, for example, when high-performing or low-performing students or schools are selected for intensive investigation. Despite their importance, however, results are often presented in the research literature with little emphasis on or reference to the model residuals; readers are thus not always provided with a clear understanding of study findings. In the education field, it becomes particularly important to be able to reliably identify children (or schools, or teachers, or program participants, etc.) whose response or outcome may not be adequately represented by a particular derived model, because if such unusual cases can be discerned, attention may be directed to improve desired outcomes.

Extensive outlines of useful residual analyses and diagnostic measures have been provided for logistic (Pregibon, 1981) and linear (Fox, 1991) regression models. In addition, Bender and Benner (2000) suggested some graphical strategies that can be used to examine the feasibility of the proportional odds

Ann A. O’Connell is a Professor in the Program in Quantitative Research, Evaluation, and Measurement (QREM) in the School of Educational Policy and Leadership within the College of Education and Human Ecology. Email her at: aconnell@ehe.osu.edu. Xing Liu is an Associate Professor of Research and Assessment in the Education Department. Email him at: liux@easternct.edu.

MODEL DIAGNOSTICS FOR PO & PPO MODELS

assumption. However, analysis of residuals for PO and PPO models in ordinal logistic regression is not well established; thus, this study was designed to build on the collection of strategies available through logistic and linear approaches. Specifically, these include: Pearson residuals, Deviance residuals, Pregibon leverages, DFBeta's and the use of index plots and other graphical strategies to examine and isolate unusual cases within the logistic framework. The use of Mahalanobis' distance, leverages, SDResiduals, Cook's D and other statistics from the ordinary least-squares framework, when applied to ordinal data are also investigated.

This study contributes to the empirical literature on detection of extreme or unusual cases, investigation of statistical assumptions and validation of ordinal regression models by: reviewing methods of model diagnosis that may be useful in investigating model assumptions, identifying unusual cases for PO and PPO models and providing a corresponding application of these diagnostic methods to the prediction of proficiency in early literacy for children drawn from the kindergarten cohort of the Early Childhood Longitudinal Study (ECLS-K; NCES, 2000). The primary focus is on how outlying or influential cases in ordinal logistic regression models can be reliably detected and on how these strategies can be applied to

proportional and/or partial proportional odds models.

Background

One of the most commonly used models for the analysis of ordinal data comes from the class of logistic models: the PO model. Consider a simple binary model; in a binary logistic model, the data represent two possible ordinal outcomes, success or failure, typically coded 0 for failure and 1 for success. For a K -level ordinal outcome, several different conceptualizations of success can be derived. Table 1 shows the ECLS-K ordinal outcome variable description and the data indicating the proportion of kindergarten children, drawn from a national random sample of kindergarteners followed through the third-grade, attaining mastery of five hierarchical early-literacy skills at the end of the kindergarten year. In this example, $K=6$, and the outcome values are scored as 0, 1, 2, 3, 4 and 5, to represent the highest level of proficiency attained on the ECLS-K literacy mastery test (0 = no mastery at any level; 5 = mastered all 5 levels). For these data, 26.9% of the children were not able to achieve beyond level 1 at the end of the kindergarten year and only 12.8% of these children mastered literacy skills beyond level 3, most students scored in levels 2 and 3 (60.3%).

Table 1: Proficiency Categories and Frequencies (Proportions) for the Study Sample, ECLS-K Measures for Early Literacy and $N = 2687$ Public School Children, End of Kindergarten Year

Proficiency Category	Description	Frequency
0	Did not pass level 1	295 (11.0%)
1	Can identify upper/lowercase letters	427 (15.9%)
2	Can associate letters with sounds at the beginnings of words	618 (23.0%)
3	Can associate letters with sounds at the ends of words	1003 (37.3%)
4	Can recognize sight words	233 (8.7%)
5	Can read words in context	111 (4.1%)

A particular series of questions are of interest when analyzing ordinal outcome data; these involve predicting the likelihood that an observation is at or beyond each specific outcome level given a collection of explanatory variables. For the ECLS-K data, this involves estimating the probability that a child with particular background characteristics or a given set of explanatory variables is at or beyond level 0 (which would always be 1.0); then estimating the probability of that same child being at or beyond level 1, at or beyond level 2, etc., until reaching the probability of the child being at or beyond the last, or K^{th} , outcome category. This series of probabilities are referred to as cumulative probabilities.

The analysis that mimics this method of dichotomizing the outcome, in which the successive dichotomizations are used to form cumulative splits to the data, is referred to as the proportional or cumulative odds model (PO) (Agresti, 2000, 2007; Armstrong & Sloan, 1989; Long, 1997; Long & Freese, 2006; McCullagh, 1980; McCullagh & Nelder, 1989; O'Connell, 2006; Powers & Xie, 2000). The model is defined as:

$$\begin{aligned} \ln(Y_j') &= \text{logit} [\pi(x)] \\ &= \ln \left(\frac{\pi(x)}{1 - \pi(x)} \right) \\ &= \alpha_j + (\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p) \end{aligned} \quad (1)$$

In this logistic model, the prediction represents the expected logit for being in category j or above, conditional on the collection of predictors, and Y_j' represents the odds of being in higher proficiency categories. These predicted logits can be transformed to odds and then to estimated probability:

$$P(Y \geq j) = \frac{\exp(\ln(Y_j'))}{1 + \exp(\ln(Y_j'))}$$

The intercept, α_j , represents the threshold, or cutpoint, for each particular split to the data. Each person thus has $K-1$ predicted values, representing their estimated likelihood of

scoring in category j or beyond, given their explanatory data. Note that in the PO model the effect of each predictor remains the same across each of these $K-1$ prediction models: This means that for each predictor, its effect on the probability of being at or beyond any category is assumed to remain constant within the model; thus, the slope estimate provides a summary of each independent variable's relationship to the outcome across all cutpoints. In this model, β_1 , for example, remains the same for all of the splits, although α_j may change. This restriction is referred to as the assumption of proportional odds.

A model that relaxes the assumption of proportional odds is referred to as a partial-proportional odds (PPO) or non-proportional odds model. This model is given by:

$$\begin{aligned} \ln(Y_j') &= \ln \left(\frac{\pi_j(\underline{x})}{1 - \pi_j(\underline{x})} \right) \\ &= \alpha_j + (\beta_{1j} X_1 + \beta_{2j} X_2 + \dots + \beta_{pj} X_p). \end{aligned} \quad (2)$$

In this expression, all of the effects of the explanatory variables are allowed to vary across each of the cutpoints. If some of the effects are found to be stable, they can be held constant as in the PO model. Thus, partial-proportional odds refers to the case that at least one of the slopes for an explanatory variable varies across splits.

Due to its simplicity and natural correspondence to ordinary logistic regression, the proportional odds model is the most widely used ordinal regression model. Tests for the assumption of proportional odds can be very liberal (Peterson & Harrell, 1990), however, and are strongly affected by sample size and the number of covariate patterns - which will always be large if continuous covariates are used (Allison, 1999; Brant, 1990; Clogg & Shihadeh, 1994). Researchers have argued that if the assumption of proportional odds is rejected, good practice would dictate that the corresponding underlying binary models be fit and compared with the PO results to check for discrepancies or deviations from the general pattern suggested by the PO model (e.g., Allison, 1999; Bender & Grouven, 1998; Brant, 1990; Clogg & Shihadeh, 1994; Long, 1997;

MODEL DIAGNOSTICS FOR PO & PPO MODELS

O'Connell, 2000, 2006). This strategy of considering the PO model as a collection of underlying binary models is an approach that has been found useful not only in qualifying the nature of the proportionality assumption, but also in assessing univariate proportionality, linearity in the logit, and the distribution of residuals from the PO model (O'Connell, Liu, Zhao & Goldstein, 2004).

Methodology

Sample

The data were drawn from the public-use data base Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) (NCES, 2000). The outcome variable of interest was proficiency in early reading, assessed at the end of the kindergarten year. Actual data were used rather than simulated data, because this could help create a realistic context for conducting residual analyses.

The final sample included $n = 2687$ public school children sampled from within 198 public schools across the U.S. The sample contained first-time kindergarten children only, who remained in the same school between kindergarten and first-grade. Analyses intentionally ignored school-level effects in order to focus on residual diagnostics for single-level models (only public schools were selected for this study). Analyses were conducted for the full sample as well as for 10 different randomly selected subsamples of 10% each to facilitate use of casewise statistics and plots.

Residual analyses are often very intensive, thus, for demonstration purposes, two of the smaller subsets were selected to highlight interesting patterns and residual statistics within each data set. No attempt was made to draw inference to the overall sample or population; rather focus was placed on demonstration of diagnostic procedures and strategies for ordinal data.

The proficiency outcomes were obtained from the third-grade release of the ECLS-K data base (prior to that, researchers used a series of dichotomous variables to derive the ordinal proficiency scores). Proficiency is defined as mastery of a cluster of 4 items representing each of the domains outlined in Table 1. The domains are hierarchically structured and theoretically

assumed to follow the Guttman scale (NCES, 2000). Mastery is recognized as students passing 3 out of the 4 items representing each domain.

The selection of explanatory variables was theoretically driven and supported through prior research on factors affecting early childhood literacy. These included: gender (male = 1), minority status (minority = 1), whether the child attended half-day kindergarten (yes = 1), number of family risks (0 to 4), frequency with which parents read books to child (0 to 3), family socio-economic status (continuous), and assessment age (continuous). The data presented were from the end of the Kindergarten year. Tables 2a, 2b and 2c present descriptive statistics for the full-sample and the two sub-samples, respectively.

Data and Models

The data were used to inspect the residuals from the PO model and test the assumptions of equal slopes. Residuals from an OLS regression of the same data as well as from the five corresponding cumulative binary logistic regression models (splits) underlying the proportional odds assumption (i.e., level 0 versus beyond level 0; levels 0 and 1 combined versus beyond level 1; levels 0, 1, and 2 combined versus beyond, etc.) were examined. Logistic regression diagnostics were investigated for each cumulative split to the data; these procedures were repeated for two of the 10% subsamples (referred to as Samples I and II). The study began by investigating the plausibility of the PO assumption in the full- and sub-samples. A PPO model was then fit where the effect of minority was allowed to vary across thresholds.

The SAS (V. 9.1.3), SPSS (V. 15.0) and Stata (V. 9.0) software packages were used for data analyses and graphing. The options for residual diagnostics in logistic regression models were also compared among these packages. SAS PROC LOGISTIC procedure was used for binary logistic models and ordinal logistic models. SPSS was used for descriptive statistics, casewise residual diagnostics in OLS and ordinal regression, index plotting and some scatterplots. Stata was used for residual diagnostics.

Table 2a: Descriptive Statistics at the End of Kindergarten, n = 2,687 (Full-Sample)

	Reading Proficiency Level						Total n = 2,687 100%
	0 n = 295 11.0%	1 n = 427 15.9%	2 n = 618 23.0%	3 n = 1,003 37.3%	4 n = 233 8.7%	5 n = 111 4.1%	
% Male	55%	57%	47%	50%	43%	42%	50%
% Minority	82.03%	51.05%	47.73%	33.40%	38.20%	32.43%	45.22%
Risknum Mean (sd)	1.48 (1.07)	.76 (.90)	.60 (.82)	.44 (.72)	.33 (.62)	.23 (.53)	.62 (.87)
Wksesl Mean (sd)	-0.7080 (0.64)	-0.2678 (0.67)	-0.1017 (0.72)	0.1278 (0.71)	0.2679 (0.72)	0.68 (0.75)	-0.0446 (0.77)
% Halfday	58.98%	47.07%	48.87%	49.55%	41.20%	54.05%	49.50%
% Readbk2	66.10%	74.47%	77.83%	84.85%	87.98%	98.20%	80.35%
r2_kage Mean (sd)	75.18 (4.15)	75.64 (4.69)	75.58 (4.47)	75.23 (4.37)	75.12 (4.43)	75.34 (4.20)	75.37 (4.42)

Table 2b: Descriptive Statistics at the End of Kindergarten, n = 244 (Sample I)

	Reading Proficiency Level						Total n = 244 100%
	0 n = 26 10.7%	1 n = 32 13.1%	2 n = 54 22.1%	3 n = 108 44.3%	4 n = 16 6.6%	5 n = 8 3.3%	
% Male	46%	66%	59%	53%	38%	38%	54%
% Minority	73.08%	50.00%	55.56%	33.33%	37.50%	37.50%	45.08%
Risknum Mean (sd)	1.38 (1.13)	0.75 (0.92)	0.69 (1.01)	0.50 (0.76)	0.25 (0.77)	0.25 (0.46)	0.64 (0.92)
Wksesl Mean (sd)	-0.8108 (0.55)	-0.2256 (0.69)	0.0057 (0.70)	0.1073 (0.76)	0.1019 (0.71)	0.2250 (0.56)	-0.0532 (0.76)
% Halfday	69.23%	46.88%	42.59%	49.07%	56.25%	50.00%	50.00%
% Readbk2	69.23%	71.88%	77.78%	81.48%	93.75%	87.50%	79.10%
r2_kage Mean (sd)	75.74 (4.53)	75.44 (3.24)	75.81 (4.93)	74.57 (4.30)	74.44 (3.94)	77.95 (4.64)	75.18 (4.37)

MODEL DIAGNOSTICS FOR PO & PPO MODELS

Table 2c: Descriptive Statistics at the End of Kindergarten, n = 278 (Sample II)

	Reading Proficiency Level						Total n = 278 100%
	0 n = 23 8.3%	1 n = 47 16.9%	2 n = 65 23.4%	3 n = 102 36.7%	4 n = 27 9.7%	5 n = 14 5.0%	
% Male	57%	68%	58%	46%	56%	29%	54%
% Minority	86.96%	44.68%	44.62%	25.49%	33.33%	42.86%	39.93%
Risknum Mean (sd)	1.39 (1.12)	0.66 (0.87)	0.41 (0.66)	0.45 (0.78)	0.19 (0.40)	0.07 (0.27)	0.51 (0.81)
Wksesl Mean (sd)	-0.8191 (0.50)	-0.2296 (0.53)	0.0534 (0.66)	0.1488 (0.70)	0.2315 (0.75)	0.9864 (0.81)	0.0327 (0.75)
% Halfday	56.52%	51.06%	55.38%	50.00%	40.74%	42.86%	50.72%
% Readbk2	65.22%	78.72%	78.46%	85.29%	85.19%	100.00%	81.65%
r2_kage Mean (sd)	74.62 (3.84)	75.12 (4.92)	75.20 (4.16)	75.81 (4.07)	74.94 (4.83)	75.20 (3.77)	75.34 (4.27)

Results

Interpretation of Proportional Odds for Full-Sample (n=2678)

The model including all seven predictors identified in Tables 2a, 2b, and 2c is referred to the full model. Table 3a provides a results summary for the fitted PO model with seven explanatory variables for the full-sample. The results were obtained using SAS with the descending option (see O’Connell (2006) for details on fitting ordinal regression models). The score test yielded $\chi^2_{32} = 131.53$ ($p < 0.0001$), indicating that the proportional odds assumptions for the full-model was not upheld. This suggested that the effect of one or more of the explanatory variables was likely to differ across separate binary models fit to the cumulative cutpoints. The Cox & Snell $R^2 = 0.210$, Nagelkerke $R^2 = 0.219$, and the likelihood ratio $R^2_L = 0.074$ all suggested that the relationship between the response variable, proficiency and the seven predictors is small.

However, the model fit statistic $\chi^2_7 = 633.55$ ($p < 0.0001$), indicated that the full model provided a better fit than the null model with no independent variables in predicting cumulative probability for proficiency.

Because the proficiency was measured through six categories with outcomes as 0, 1, 2, 3, 4 or 5, with the descending option in SAS, α_5 corresponds to the intercept for the cumulative logit model for $Y \geq 5$, α_4 corresponds to the intercept for the cumulative logit model for $Y \geq 4$, and so on. The effects of the seven independent variables can be interpreted as how variables contribute to the log of the odds of being at or beyond a particular category. In terms of odds ratios, boys were less likely than girls to be at or beyond a particular category (OR=0.73). Being in a minority category (OR=0.649), the presence of any family risk factor (OR=0.649) and attending half-day kindergarten rather than full-day kindergarten (OR=0.695) all had significant negative coefficients in the model and corresponding

Table 3a: Proportional Odds Model and OLS Regression Model for Full-Sample Data, n = 2687 Public School Children (Y > cat.j)

Variable	Proportional Odds Model	OR	OLS Model
	b (se(b))		b (se(b))
α_5	-2.88 (0.62)		
α_4	-1.59 (0.61)		
α_3	0.58 (0.61)		
α_2	1.74 (0.61)		
α_1	3.02 (0.61)		2.64 (0.38)
Gender ^δ	-0.31 (0.07)**	0.730	-0.21 (0.04)**
Minority	-0.43 (0.08)**	0.649	-0.26 (0.05)**
RiskNum	-0.36 (0.05)**	0.695	-0.23 (0.03)**
Halfday	-0.48 (0.07)**	0.619	-0.27 (0.04)**
Readbk2	0.35 (0.09)**	1.422	0.22 (0.06)**
Wksesl	0.78 (0.06)**	2.183	0.48 (0.03)**
R2_kage	-0.001 (0.01)	0.999	0 (0.01)
R^2	$R^2_L = .074$		$R^2 = .218$
Cox & Snell R^2	.210		
Nagelkerke R^2	.219		
Somer's D	.386		
Model Fit ^a	$\chi^2_7 = 633.55$ (p < 0.0001)		F(7, 2679) = 106.76**
Deviance	7869.82 (df = 13398)	0.5874 ^c	
Pearson X^2	13032.47 (df = 13398)	0.9727 ^c	
Score Test ^b	$\chi^2_{32} = 131.53$ (p < 0.0001)		

Notes: ^δgender: male=1; ^aLikelihood ratio test; ^bFor the proportional odds assumption; ^cValue/df; *Significant at p < 0.05; ** p < 0.01

OR's were significantly less than 1.0. These characteristics were associated with a child being in lower proficiency categories rather than in higher categories. Conversely, increasing frequency of parents reading to their children (OR=1.422) and family SES (OR=2.183) had positive effects on children being in higher

proficiency categories. The slopes for both variables were positive and significantly different from zero in the model; child's assessment age was not associated with proficiency in this model because the slope was almost 0 and the OR is close to 1.0.

MODEL DIAGNOSTICS FOR PO & PPO MODELS

Interpretation of Proportional Odds Models for Sub-samples I and II

Table 3b shows the results summary for the fitted PO model for sub-samples I ($n = 244$) and II ($n = 278$): both results were similar to that of the full-sample. An $\alpha = 0.05$ was used to assess the hypothesis of proportionality. The score test for sub-sample I, $\chi^2_{28} = 46.67$ ($p = 0.0148$), and the score test for sub-sample II $\chi^2_{28} = 44.41$ ($p = 0.0253$), were both statistically significant, indicating that the proportional odds assumptions for both sub-models were violated.

The model fit statistic for sub-sample I, $\chi^2_7 = 38.61$ ($p < 0.0001$), and the model fit statistic for sub-sample II, $\chi^2_7 = 71.67$ ($p < 0.0001$) indicated that the models with seven predictors provided a better fit than the null model with no independent variables. The Odds Ratios (OR) for the seven explanatory variables for sub-sample I and II looked similar, and were also similar to those of the proportional odds model for the full-sample. However, in the PO model for sub-sample I, it was noticeable that the effects of gender, minority and frequency of being read to by parents were not statistically significant; and the p value of the slopes of number of family risks and attendance at half-day kindergarten were slightly larger than the 0.05 level. In the CO model for sub-sample II, the effects of minority, half-day kindergarten and having parents read to their children were not significant.

Assumption of Proportional Odds

Table 4a shows the results of five separate binary logistic regression analyses for the full sample, where the data were dichotomized according to the cumulative probability pattern described earlier for the proportional odds model. Each logistic regression model estimates the probability of being at or beyond proficiency level j . In the data set, the grouping of categories coded 1 corresponded to children who were at or beyond each proficiency category and 0 was coded for children below each successive category. The model χ^2 for each separate logistic model was statistically significant, indicating that each model fit well compared to its corresponding null model. The Hosmer-Lemeshow tests were

all not statistically significant, indicating that observed and predicted probabilities were consistent.

Examining the patterns of slopes and ORs for each explanatory variable across these five logistic regression models, it was found that the effects of gender, after adjusting for the other predictors directionally and on average, were similar across the five separate logistic regressions. This was also true for family risk, family SES, half-day kindergarten, being read to by parents and the child's assessment age. However, the effect of minority did appear to present a dissimilar pattern across the five separate logistic regressions. The direction of the effect of minority changed between the first three regressions and the last two. For the other explanatory variables, the direction and average magnitude of the slopes and the ORs from the logistic models were similar to those of the PO model.

Because the proportional odds assumption for the full-sample ordinal model was violated, separate score tests unadjusted for the presence of other covariates in the cumulative odds model were examined for each of the explanatory variables, in order to illuminate where non-proportionality might lie. The univariate score tests for the assumption of proportional odds were upheld for gender and child's assessment age. However, the univariate score tests were violated for minority, family risk, family SES, being read to by parents and half-day kindergarten at the 0.05 level of significance. The p -values for these unadjusted tests are presented in the final column of Table 4a; it should be noted that these score tests are simply descriptive, given their univariate nature.

Table 4b presents the results of five separate binary logistic regression analysis for sub-sample I, $n = 244$. The univariate score tests for the assumption of proportional odds were upheld for separate PO models for all the variables, except for the continuous variable of family SES (*wksesl*). The p values for these unadjusted tests are presented in the final column of Table 4b. However, minority as well as gender, half-day kindergarten attendance and frequency of being read to by parents all exhibited inconsistencies in the directional patterns across the binary splits.

Table 3b: Subsamples I (n = 244) and II (n = 278): Proportional Odds Models (Y > cat.j)

Variable	Sample I		Sample II	
	b (se(b))	OR	b (se(b))	OR
α_5	-1.78 (2.11)		-5.68 (2.00)	
α_4	-0.60 (2.09)		-4.41 (1.99)	
α_3	1.96 (2.09)		-2.35 (1.97)	
α_2	3.08 (2.10)		-1.14 (1.97)	
α_1	4.16 (2.11)		0.40 (1.98)	
Gender ^δ	-0.29 (0.24)	0.750	-0.45 (0.22)*	0.638
Minority	-0.40 (0.28)	0.670	-0.38 (0.24)	0.681
RiskNum	-0.29 (0.15) ^a	0.752	-0.31 (0.16) [*]	0.733
Halfday	-0.45 (0.25) ^a	0.638	-0.33 (0.21)	0.716
Readbk2	0.27 (0.30)	1.312	0.24 (0.29)	1.266
Wksesl	0.58 (0.20) ^{**}	1.793	0.94 (0.18) ^{**}	2.552
R2_kage	-0.02 (0.03)	0.983	0.04 (0.03)	1.039
R^2_L	0.053		0.081	
Cox & Snell R^2	0.146		0.227	
Nagelkerke R^2	0.154		0.237	
Somer's D	0.322		0.393	
Model Fit ^b χ^2_7	38.61 (p < 0.0001)		71.67 (p < 0.0001)	
Deviance	688.63 (df = 1208)	0.5701 ^d	813.11 (df = 1378)	0.5901 ^d
Pearson X^2	1190.51 (df = 1208)	0.9855 ^d	1221.86 (df = 1378)	0.8867 ^d
Score Test ^c χ^2_{28}	46.67 (p = 0.0148)		44.41 (p = 0.0253)	

Notes: ^δgender: male=1; ^ap = 0.06 for risknum; p = 0.07 for halfday; ^bLikelihood ratio test; ^cFor the proportional odds assumption; ^dValue/df; *Significant at p < 0.05; ** p < 0.01

Table 4c shows the results of five separate binary logistic regression analysis for sub-sample II, n = 278. Based on $\alpha = 0.05$, the univariate score tests for the assumption of proportional odds were upheld for separate PO models for these variables: gender, attending half-day kindergarten (halfday), having parents

read to their children (readbk2) and child's assessment age (r2_kage). However, the score tests for the assumptions of proportional odds were violated for these three variables: being in a minority category (minority), number of family risks (risknum) and family SES (wksesl). The p-values for these unadjusted tests were presented in the final column of Table 4c.

MODEL DIAGNOSTICS FOR PO & PPO MODELS

Table 4a: Associated Cumulative Binary Models for the CO Analysis (Descending) for the Full-sample, Where CUMSPj Compares $Y < \text{cat. } j$ to $Y > \text{cat. } j$, $n = 2687$

Variable	CUMSP1 b (se(b)) OR	CUMSP2 b (se(b)) OR	CUMSP3 b (se(b)) OR	CUMSP4 b (se(b)) OR	CUMSP5 b (se(b)) OR	Score Test ^b P value
Constant	2.53 (1.22)	1.88 (0.82)	1.10 (0.72)	-1.77 (1.04)	-5.84 (1.86)	
Gender ^δ	-0.33 (0.14) 0.72*	-0.44 (0.10) 0.64**	-0.23 (0.08) 0.80**	-0.40 (0.12) 0.67**	-0.41 (0.20) 0.67*	0.10
Minority	-1.21 (0.18) 0.30**	-0.49 (0.10) 0.61**	-0.47 (0.09) 0.62**	0.11 (0.13) 1.12	0.17 (0.23) 1.18	0.001
Risknum	-0.50 (.08) 0.61**	-0.33 (0.06) 0.72**	-0.23 (0.06) (0.79)**	-0.35 (0.11) 0.70**	-0.32 (0.21) 0.73	0.001
Wksesl	0.94 (0.13) 2.57**	0.83 (0.09) 2.29**	0.73 (0.07) 2.08**	0.79 (0.09) 2.19**	1.12 (0.14) 3.08**	0.001
Halfday	-0.85 (0.14) 0.43**	-0.42 (0.10) 0.65**	-0.37 (0.09) 0.69**	-0.42 (0.12) 0.65**	-0.12 (0.20) 0.89	0.004
Readbk2	0.19 (0.15) 1.21	0.25 (0.11) 1.28*	0.37 (0.11) 1.45**	0.60 (0.21) 1.83**	2.00 (0.72) 7.40**	0.043
r2_kage	0.19 (0.15) 1.02	0.00 (0.01) 1.00	-0.01 (0.01) 0.99	0.00 (0.01) 1.00	0.01 (0.02) 1.01	0.47
R^2_L	0.237	0.134	0.105	0.088	0.134	
Naglekerke R^2	0.30	0.21	0.18	0.12	0.16	
Model χ^2_7	441.26**	418.37**	389.97**	180.89**	124.18**	
H-L ^a χ^2_8	5.80	8.07	7.94	3.81	13.09	

Notes: ^δ gender: male=1; *Significant at $p < 0.05$; $p < 0.01$; ^aHosmer-Lemeshow test; ^bScore test for each IV, unadjusted (no other covariates in the model)

Table 4b: Associated Cumulative Binary Models for the CO Analysis (Descending) for Sub-sample I, Where CUMSP_j Compares Y < cat. j to Y > cat. j, n = 244

Variable	CUMSP1 b (se(b)) OR	CUMSP2 b (se(b)) OR	CUMSP3 b (se(b)) OR	CUMSP4 b (se(b)) OR	CUMSP5 b (se(b)) OR	Score Test ^b P value
Constant	3.43 (4.29)	2.52 (.2.84)	4.26 (2.45)	-4.46 (4.05)	-15.79 (6.68)	
Gender ^δ	0.37 (0.48) 1.45	-0.22 (0.33) 0.80	-0.40 (0.28) 0.67	-0.71 (0.46) 0.49	-0.69 (0.78) 0.50	0.38
Minority	-0.61 (0.58) 0.54	-0.23 (0.38) 0.79	-0.65 (0.31) 0.52*	0.18 (0.50) 1.20	0.27 (0.84) 1.31	0.36
Risknum	-0.24 (0.26) 0.78	-0.16 (0.19) 0.85	-0.22 (0.18) (0.80)	-0.76 (0.42) 0.47	-0.71 (0.71) 0.49	0.47
Wksesl	1.74 (0.53) 5.71**	1.09 (0.32) 2.98**	0.38 (0.23) 1.47	0.13 (0.36) 1.24	0.55 (0.59) 1.73	0.01
Halfday	-1.66 (0.54) 0.19**	-0.74 (0.34) 0.48*	-0.30 (0.28) 0.74	0.11 (0.46) 1.11	0.01 (0.77) 1.01	0.26
Readbk2	-0.11 (0.59) 0.89	0.18 (0.40) 1.19	0.23 (0.35) 1.26	0.78 (0.78) 2.18	0.07 (1.14) 1.07	0.90
r2_kage	0.01 (0.06) 1.01	-0.01 (0.04) 0.99	-0.05 (0.01) 0.96	0.03 (0.05) 1.03	0.17 (0.08) 1.18	0.13
R ² _L	0.265	0.127	0.082	0.072	0.104	
NaglekerkeR ²	0.33	0.19	0.14	0.10	0.12	
Model χ^2_7	43.83**	33.89**	27.73**	11.34	7.29	
H-L ^a χ^2_8	14.31	14.59	11.25	8.89	4.14	

Notes: ^δ gender: male=1; *Significant at p < 0.05; ^aHosmer-Lemeshow test; ^bScore test for each IV, unadjusted (no other covariates in the model)

MODEL DIAGNOSTICS FOR PO & PPO MODELS

Table 4c: Associated Cumulative Binary Models for the CO Analysis (Descending)
for Sub-Sample II, Where CUMSPj Compares Y < cat. j to Y > cat. j, n = 278

Variable	CUMSP1 b (se(b)) OR	CUMSP2 b (se(b)) OR	CUMSP3 b (se(b)) OR	CUMSP4 b (se(b)) OR	CUMSP5 b (se(b)) OR	Score Test ^b P value
Constant	-4.16 (4.50)	-2.03 (2.79)	-2.09 (2.26)	-0.66 (3.20)	-13.56 (284.9)	
Gender ^δ	-0.02 (0.53) 0.98	-0.63 (0.32) 0.53	-0.62 (0.26) 0.54	-0.26 (0.36) 0.77	-0.97 (0.67) 0.38	0.33
Minority	-1.41 (0.71) 0.24	-0.42 (0.34) 0.66	-0.71 (0.29) 0.49*	0.42 (0.39) 1.53	0.81 (0.64) 2.24	0.00
Risknum	-0.14 (0.29) 0.87	-0.23 (0.21) 0.80	-0.08 (0.20) (0.93)	-1.00 (0.44) 0.37*	-1.65 (1.10) 0.19	0.04
Wksesl	2.60 (0.73) 13.43**	1.24 (0.32) 3.44**	0.71 (0.21) 2.04**	0.82 (0.25) 2.23**	1.59 (0.42) 4.91**	0.00
Halfday	-0.49 (0.54) 0.61	-0.23 (0.31) 0.98*	-0.37 (0.26) 0.69	-0.53 (0.36) 0.59	-0.47 (0.62) 0.62	0.93
Readbk2	0.17 (0.57) 1.19	-0.02 (0.38) 0.98	0.25 (0.35) 1.28	0.32 (0.58) 1.38	11.60 (284.8) 1000.00	0.65
R2_Kage	0.12 (0.06) 1.13	0.05 (0.04) 1.06	0.04 (0.03) 1.04	-0.01 (0.04) 1.00	-0.01 (0.08) 1.00	0.77
R^2_L	0.351	0.163	0.112	0.121	0.297	
NaglekerkeR ²	0.42	0.25	0.19	0.17	0.34	
Model χ^2_7	55.70**	51.25**	43.08**	28.11**	32.94**	
H-L ^a χ^2_8	4.17	5.07	6.99	2.63	2.22	

Notes: ^δ gender: male=1; *Significant at p < 0.05; ^aHosmer-Lemeshow test; ^bScore test for each IV, unadjusted (no other covariates in the model)

For comparison, Table 4d presents the p-values for univariate score tests of proportionality for each explanatory variable analyzed separately in a single variable model for the full-sample and for sub-samples I and II provided for SPSS, SAS and Stata. SPSS performs an approximation to the score test in their PLUM procedure (Nichols, 2004), so analysts should be aware of the possibility for discrepancies and differences in results between software packages.

Given the large sample size, $\alpha = 0.01$ was used to evaluate the assumption of proportionality for these univariate tests for the full-sample. Consistent results were found across the three software packages for all explanatory variables except the frequency with which parents read books to children, for which $p = 0.006$ (SPSS), $p = 0.0426$ (SAS) and $p = 0.078$ (Stata). However, for the two smaller sub-samples, and using $\alpha = 0.05$, it was found that the results of these univariate score tests (using SAS) varied across the sub-samples and were also inconsistent with the full-sample results. For example, the p-values of the score test for the minority variable was 0.3562 in the model for sub-sample I and 0.0031 for sub-sample II; the hypothesis of proportionality was rejected for the minority variable within the full sample ($p < 0.0001$). In addition, the effect of attending

half-day kindergarten was found to deviate from proportionality in the full sample, but this assumption was upheld in both of the smaller samples.

Results of the Partial Proportional Odds (PPO) Model

The discrepancy of results of the score tests between the full-sample and sub-sample analyses and across statistical packages presents a disheartening situation for the analyst attempting to assess the plausibility of the proportional odds model through score tests alone. These results support the view that investigation of proportional odds may be more reasonably investigated through visual examination of the variable effects and odds ratios of the binary models underlying the ordinal progression of the outcome data. Consequently, it was decided to fit a PPO model that relaxes the assumption of proportionality for the minority variable because this effect changed direction across cutpoints in all three analyses (full-sample, I and II). Results using SAS PROC GENMOD are shown in Table 4e for the full-sample data.

The outcome being modeled in this PPO analysis was the probability that a child was at or beyond category j , with the effect of minority being allowed to vary across the $K-1 = 5$

Table 4d: Score tests for Proportion Odds for each IV in the Model: Single Variable Models for the Full Sample and for Subsamples I and II.

IV	Full Sample, p-value SPSS	Full Sample, p-value SAS	Full Sample, p-value Stata	Sample I p-value SAS	Sample II p-value SAS
Gender ^δ	0.101	0.1013	0.102	0.3792	0.3313
Minority	0.000	< 0.0001	0.000	0.3562	0.0031
Risknum	0.000	< 0.0001	0.000	0.4698	0.0440
Halfday	0.004	0.0041	0.005	0.2643	0.8821
Readbk2	0.006	0.0426	0.078	0.8969	0.6471
Wksesl	0.000	< 0.0001	0.000	0.0119	0.0004
R2_Kage	0.467	0.4700	0.445	0.1263	0.7706

Notes: ^δgender: male=1

MODEL DIAGNOSTICS FOR PO & PPO MODELS

Table 4e: Partial Proportional Odds (PPO) Model for Full-sample, SAS (Descending) ($Y \geq \text{cat. } j$), N = 2687

Variable	b (se(b))	OR
Intercept	-3.56 (0.63)**	0.03
Gender (0)	0.34 (0.07)**	1.40
Minority (0)	0.06 (0.21)	
Risknum	-0.33 (0.05)**	0.72
Wksesl	0.83 (0.06)**	2.29
Halfday (0)	0.44 (0.07)**	1.55
Readbk2	-0.34 (0.09)**	0.71
R2_Kage	0.00 (0.00)	1.00
Split 1	5.46 (0.18)**	6.69
Split 2	4.42 (0.17)**	2.36
Split 3	3.26 (0.17)**	0.74
Split 4	1.39 (0.15)**	0.11
Split 5	0.00 (0.00)**	0.03
Split 1* Minority (0)	1.23 (0.26)**	
Split 2* Minority (0)	0.42 (0.22)	
Split 3* Minority (0)	0.35 (0.21)	
Split 4* Minority (0)	-0.15 (0.18)	
Split 5* Minority (0)	0.00 (0.00) (see score test below)	

*p < 0.05; **p < 0.01; Gender: female = 0; Minority: no = 0; Halfday: no = 0

Score Statistics for Type 3 GEE Analysis

Variable	Chi-Square	p
Gender	$\chi^2_1 = 22.54$	< 0.0001
Minority	$\chi^2_1 = 22.69$	< 0.0001
Risknum	$\chi^2_1 = 43.08$	< 0.0001
Wksesl	$\chi^2_1 = 158.61$	< 0.0001
Halfday	$\chi^2_1 = 35.83$	< 0.0001
Readbk2	$\chi^2_1 = 14.38$	0.0001
R2_kage	$\chi^2_1 = 0.00$	0.9734
Split	$\chi^2_4 = 2132.14$	< 0.0001
Split *Minority	$\chi^2_4 = 60.86$	< 0.0001

cutpoints while holding the other variable effects constant. Overall, only one intercept parameter was estimated (for the log-odds of a value being at or beyond proficiency level 5), but different estimates for each split were obtained which were used to modify the intercept value; and the split by minority interaction terms were used to determine how much the effect of minority changed across splits. In this analysis, the coding of categorical variables was internal to GENMOD; effects are shown in Table 4e for the lower value of the coded variables (for example, the effect for gender, $I = 0.34$, was the change in slope for females; this was the opposite of the slope in the PO model, where $I = -0.31$ for the change in slope for males).

According to these results, the split by minority interaction terms were statistically significant for the first logit comparison (corresponding to $P(Y \geq 1)$) and the last (corresponding to $P(Y \geq 5)$), but not for the other splits. These results suggested that reliable differences existed in the effect of minority across the outcome levels of the ordinal model.

Overall, when coding was taken into consideration for the effect of minority, the predicted logits were similar to those obtained from the separate binary regressions shown in Table 4a for the full-sample. Additionally, the slope effects for the remaining variables, which were held proportional in the analysis, reflected the values obtained through the PO analysis in Table 3a (after accounting for coding reversals). Thus, using the separate binary models to investigate the presence and impact of extreme or unusual scores made sense for both the PO and PPO model, as these binary models reflect what was expected in the data for each of the separate splits.

Residuals in Ordinal Logistic Regression

SAS, SPSS and Stata do not provide residual diagnostics for ordinal models. Hosmer and Lemeshow (2000) suggested considering each of the underlying models separately and applying residual methods for these binary logistic models in order to identify unusual or extreme observations. This approach mirrors the aptness of model investigations for the proportional odds assumption presented previously.

Consider the residuals for these underlying cumulative binary models and in addition the OLS strategies that were used for preliminary analyses to examine whether that approach could assist in identifying unusual cases. Under the OLS framework, there are several commonly used measures to identify unusual cases. Mahalanobis' distance is the distance for each case to the centroid of remaining cases (multivariate outliers). Leverages are the diagonal elements of the hat matrix in OLS; they are a transformation of Mahalanobis distance. Leverages flag cases are considered extreme in the X or explanatory variable space, where two or three times the average leverage can be considered large.

However, cases with large leverage values may or may not be influential; that is, an observation may be unusual in terms of being outside an acceptable range relative to the other X values, but it may not affect the shape or direction of the regression function. Cook's D assesses how influential each case is to the fit of the overall model. This measure considers what happens to the model when each case is removed, one at a time, from the overall model. A large Cook's D value is determined in relation to values obtained from all the other cases. Generally Cook's D statistics are plotted to identify any large jumps in the measures. Finally, when assessing outliers on Y , the outcome variable, a common statistic is the studentized deleted residual, or SDRESID. For SDRESID the change in residuals was examined when each case is removed, one at a time, from the model.

Adjustments to the OLS statistics are required for logistic regression. Logistic models predict the probability that $Y = 1$ for a dichotomous dependent variable. The residuals obtained through a logistic regression are heteroscedastic (variance = $\pi_i(1 - \pi_i)$). Techniques similar to those used in OLS models have been developed for logistic regression in order to detect unusual or influential observations (Pregibon, 1981). The Pearson residual, deviance residual and Pregibon leverages are three main types of residual statistics commonly used for logistic regression diagnostics; however, many choices are available. Table 5 presents the types of residual

MODEL DIAGNOSTICS FOR PO & PPO MODELS

diagnostics for logistic regression that are available in the three statistical software packages discussed here: SPSS, SAS and Stata. The corresponding mathematical form for each type of residual is also included in the table.

The Pearson residual is the standardized difference between the observed and fitted values. Pearson residuals, r_i are components of the summary χ^2 . SAS labels these residual components as `reschi`; SPSS labels it as `zresid` (normalized residuals); Stata provides residuals (Pearson residuals) and `rstandard` (standardized Pearson residuals). Pearson residuals can be computed using:

$$r_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}} \quad (3)$$

Where y_i is the observed number of success; n_i is the number of observations with explanatory variable x_i ; $\hat{\pi}_i$ is estimated probability at x_i . When the number of observations is 1, that is, $n_i = 1$ (assuming a Bernoulli rather than binomial model), the Pearson residuals can be simplified as:

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i (1 - \hat{\pi}_i)}} \quad (4)$$

Deviance residuals capture the difference between the maxima of the observed and fitted log likelihood functions. Dev_i are components of the summary model deviance, $D = -2LL$. SAS labels this as `resdev`; SPSS labels it as `dev`; and Stata labels it as `deviance`. Deviance residuals are defined as:

$$d_i = \pm \sqrt{2 \left[y_i \log \frac{y_i}{n_i \hat{\pi}_i} + (n_i - y_i) \log \frac{(n_i - y_i)}{n_i (1 - \hat{\pi}_i)} \right]} \quad (5)$$

Also when $n_i = 1$, the formula can be simplified as:

$$d_i = \pm \sqrt{2 \left[y_i \log \frac{y_i}{\hat{\pi}_i} + (1 - y_i) \log \frac{(1 - y_i)}{(1 - \hat{\pi}_i)} \right]} \quad (6)$$

Large values of r_i or d_i suggest that the model does not fit that case well. The r_i and dev_i are components of alternate tests for same null hypothesis tested based on the Chi-square distribution; that is, does the model fit as well as a saturated model of the data (i.e., one that perfectly reproduces the original data). Under the null hypothesis, the individual components are approximately normally distributed and it is expected that the summary value/df will be less than 1.0. However, neither the summary Pearson residual statistic nor the summary deviance residual statistic follows a Chi-square distribution when continuous explanatory variables are included in the model. Thus, the summary statistics are not appropriately used in that situation. When the data are sparse (i.e., with continuous IVs), this Chi-square distributional assumption is not upheld.

Pregibon leverages (`hat`) are the diagonal elements of the hat matrix in logistic regression. They are used to measure the relative influence of an observation on the model fit. SAS labels these as `h`; SPSS labels them as `lever`; and Stata labels them as `hat`. Pregibon hats tend to be small when the estimated probability of observations is outside of the 0.1 to 0.9 interval, because most extreme cases may also have small leverages (Hosmer & Lemeshow, 2000). Therefore, Pregibon hats may not provide a good assessment of influential cases when the estimated probability of an observation is too small or too large. Pregibon leverages are defined as:

$$H = W^{1/2} X (X' W X)^{-1} X' W^{1/2} \quad (7)$$

$$h_{ii} = \hat{\pi}_i (1 - \hat{\pi}_i) x_i' \hat{Var}(\hat{\beta}) x_i$$

where W is the diagonal weight matrix; and h_{ii} is the leverage or diagonal element of the hat matrix, H .

It is also often more informative to consider how each case affects fit of the overall model. There are several approaches in logistic regression to measure the change in Chi-square fit, deviance fit, and in the estimated parameters when a single observation is removed from the model. These measures are similar to Cook's D in ordinary least-squares regression. SPSS

provides a measure which is an analog of Cook's D, and labels it Cook; SPSS only provides change in estimated parameters and labels it as dfbeta. SAS and Stata provide all three options. SAS labels the standardized difference in estimated parameters dfbetas, which is different from SPSS. SAS also provides two measures of the change in the confidence interval for the regression estimates when an

observation is deleted, and labels them C and CBAR. SAS labels the change in Chi-square fit as difchisq and change in deviance fit as dfdev. Stata labels the standardized change in regression coefficient as dbeta, the change in Chi-square fit as dx2, and the change in deviance fit as ddeviance. Descriptions of several of these statistics in SPSS, SAS and Stata are listed in Table 5.

Table 5: Residual Diagnostics of Logistic Regression in SPSS, SAS and Stata

Types Of Residuals	Mathematical Formula	SPSS	SAS	Stata
Pearson Residuals	$r_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$	✓	✓	✓
Studentized Residuals	$studentized\ rs_i = \frac{r_i}{\sqrt{1 - h_{ii}}}$	✓		✓
Logit Residuals	$logit\ r_i = \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i (1 - \hat{\pi}_i)}$	✓		
Deviance Residual	$d_i = \pm \sqrt{2 \left(y_i \log \frac{y_i}{n_i \hat{\pi}_i} + (n_i - y_i) \log \frac{(n_i - y_i)}{n_i (1 - \hat{\pi}_i)} \right)}$	✓	✓	✓
Leverage (Hat)	$H = W^{1/2} X (X' W X)^{-1} X' W^{1/2}$ $h_{ii} = \hat{\pi}_i (1 - \hat{\pi}_i) x_i' \hat{Var}(\hat{\beta}) x_i'$	✓	✓	✓
Cook's D	$C_i = \frac{r_i^2 h_{ii}}{(1 - h_{ii})^2}$	✓	✓	✓
Change in Pearson Chi-square Residuals	$\Delta \chi_i^2 = \frac{r_i^2}{1 - h_{ii}}$		✓	✓
Change in Deviance Residuals	$\Delta D_i = d_i^2 + C_i$		✓	
	$\Delta D_i = \frac{d_i^2}{(1 - h_{ii})}$			✓
Change in Betas	$\Delta b_i = \frac{r_i^2 h_{ii}}{1 - h_{ii}}$	✓	✓	✓

MODEL DIAGNOSTICS FOR PO & PPO MODELS

The Hosmer-Lemeshow test does not yield a residual measure, but it does inform about the fit of a model, particularly when continuous explanatory variables are included. The H-L test attempts to compensate for the presence of continuous variables and resulting sparseness of the data by aggregating across deciles of risk, which are formed by collapsing over observations with similar covariate patterns. However, this same strategy of aggregating by covariate pattern for residual diagnostic purposes may mask influential or poorly fit cases; therefore, researchers tend to rely on visual assessment rather than specific concrete measures or approaches, for identification of unusual cases. In addition, given the lack of informative distributional theory regarding many of the residual statistics discussed above, it may be more valuable to apply graphical strategies for observing and identifying unusual or influential cases within logistic and ordinal models.

Many of these graphical approaches mirror what is available through OLS. For ordinary least squares regression, a graph of the observed dependent variable, y , versus the predicted dependent variable, \hat{y} , can be plotted. Observed and predicted outcomes in logistic regression are dichotomous; thus modifications are required – usually the resulting graphs are done at a casewise level. For logistic regression, index plots are enormously useful. Index plots display each residual diagnostic for a case against the case number. The resulting output can be unwieldy for samples with many observations and when multiple residual statistics are investigated; but visually, extreme or unusual cases can be readily detected.

As a minimum, Hosmer and Lemeshow (2000) recommended plotting the change in Chi-square fit versus the predicted probability (\hat{p}) of the dependent variable; change in deviance fit versus \hat{p} ; and change in regression estimates versus \hat{p} . They pointed out that using the summary change statistic rather than the individual component values (r_i or d_i , above) for each case visually emphasized the poorly fit cases. Because not all statistics are available in each statistical package, choices among possible

graphs or plots have to be made depending on the options available.

Residual Diagnostics Results

With large data sets, the amount of output involved in graphical displays for diagnostic statistics can become unworkable very quickly. Thus, two smaller data sets are employed to demonstrate the use and interpretation of regression diagnostic statistics. OLS regression of the ordinal outcome was used on the collection of explanatory variables as an initial strategy to identification of unusual or poorly fit cases for both Samples I and II. A series of logistic regression models were then run for Samples I and II corresponding to the cumulative splits; residual statistics obtained from these five logistic models were used to identify unusual cases. Next, the use of index plots to visually display the residual statistics for a collection of diagnostic values derived from the logistic splits for Sample I (one sample was selected for demonstration of the index plots) was demonstrated. After reviewing the residual strategies and identifying the poorly fit cases, those cases identified as unusual by inspecting the original data were explored along with characteristics of the children in order to better understand who is potentially being poorly fit by the model.

Ordinal Residual Diagnostics for Sample I

Table 6a presents casewise statistics based on unusual cases identified through an OLS scatterplot of observed versus predicted values, OLS casewise diagnostics, and additional cases identified through the five sequential (cumulative) logistic models and corresponding index plots for Sample I ($n = 244$). Figure 1 displays the OLS scatterplot of observed versus predicted values: several potentially unusual cases or outliers were observed in proficiency level 0, 3 and 5. From the scatterplot, nine cases were identified as unusual or poorly fit. The absolute values of the studentized deleted residuals (SDRESID) of these cases were mostly larger than or close to 2. For case number 1698 (ID = 0731010C) its Mahalanobis Distance was 39.23, which is very large, and the corresponding leverage value and Cook's D for this case are also the largest among

those cases identified as outliers. This child was predicted to be in proficiency level 0.28 (via OLS), yet had actually scored in proficiency level 3. Visually, this case can be clearly identified as an outlier in Figure 1 for proficiency level 3.

Table 6a shows statistics for two additional cases identified through the OLS

residual diagnostics options as having standardized deleted residuals (ZRESID) larger than our setting of 2.5 (2.5 was used instead of the default value of 3.0 to maintain consistency with the values used in the logistic regression procedure). Both children were in proficiency level 5 but were predicted to be in levels 2.16 and 2.15 respectively.

Table 6a: Casewise Diagnostics for OLS Model Sample I

From Residual Diagnostics

Case No.	Child ID	ProfrdK2	Predicted	Residual	SDRESID	Lever	Mahal	Cook's D
2103	0936002C	5	2.16	2.84	2.60	.02	4.86	.02
3414	3113018C	5	2.15	2.85	2.62	.03	7.71	.03

From Scatterplot

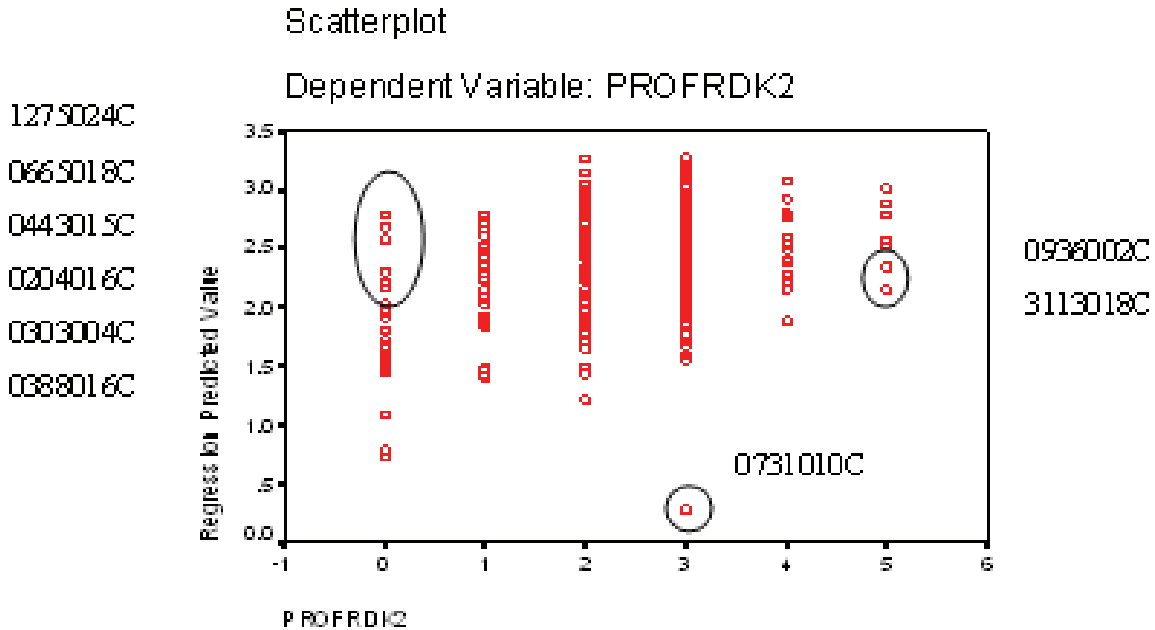
Case No.	Child ID	ProfrdK2	Predicted	Residual	SDRESID	Lever	Mahal	Cook's D
351	0204016C	0	2.30	-2.30	-2.12	.04	10.87	.03
665	0303004C	0	2.23	-2.23	-2.02	.02	4.81	.01
946	0388016C	0	2.21	-2.21	-1.99	.02	3.83	.01
1059	0443015C	0	2.58	-2.58	-2.35	.02	4.36	.02
1571	0665018C	0	2.68	-2.68	-2.44	.02	3.77	.01
1698	0731010C	3	0.28	2.72	2.69	.16	39.23	.17
2343	1049011C	0	2.17	-2.17	-1.97	.02	4.82	.01
2889	1271005C	0	2.04	-2.04	-1.85	.03	6.48	.01
2952	1275024C	0	2.78	-2.78	-2.53	.02	4.20	.02

Additional from Logistic Diagnostics and Plots

Case No.	Child ID	ProfrdK2	Predicted	Residual	SDRESID	Lever	Mahal	Cook's D
1917	0833003C	2	3.06	-1.06	-0.96	.02	4.15	.00
3392	3105020C	4	1.89	2.11	1.96	.06	15.23	.03
3114	3002012C	2	2.90	-0.90	-0.81	.02	5.39	.00
1384	0609022C	3	1.70	1.30	1.21	.08	18.84	.02
3078	2115003C	5	2.88	2.12	1.92	.02	3.88	.01

MODEL DIAGNOSTICS FOR PO & PPO MODELS

Figure 1: OLS Observed Versus Predicted Values for Sample I, n = 244 Public School Children



For comparison, the summary statistics of Table 6a include OLS for cases that were identified through the binary logistic models as being potential outliers: five additional cases were detected. The OLS statistics for these cases, however, do not indicate that these cases are unusual, with the exception of one large Mahalanobis' distance value of 18.84 for Case No. 1384 (ID = 0609022C).

Table 6b presents casewise diagnostics for the five cumulative binary logistic regression models (splits) for Sample I. Four cases (Case No. 1059, 1571, 1698 and 2952) were flagged as outliers for the first cumulative split based on SRESID greater than 2.5 (set greater than the 2.0 default setting); one (Case No. 1698) was flagged as an outlier for the second cumulative split; no cases were identified in the analysis for the third cumulative split; two (Case No. 1917, and 3414) were flagged as outliers for the fourth cumulative split; and three (Case No. 2103, 3392, and 3414) were flagged as outliers for the fifth cumulative split.

Ordinal Residual Diagnostics for Sample II

Table 7a presents casewise statistics based on unusual cases identified through an

OLS scatterplot of observed versus predicted values, OLS casewise diagnostics, and additional cases identified through the five sequential (cumulative) logistic models and corresponding index plots for Sample II (n = 278). Figure 2 displays the OLS scatterplot of observed versus predicted values; eleven potentially unusual cases or outliers can be observed in proficiency level 0, 4 and 5. Referring to Table 7a, only one of these eleven cases had a fairly large OLS Mahalanobis' distance of 11.85; this child scored in proficiency level 4 but was predicted into level 1.95. The statistics for the other 10 cases identified did not seem unusual. The absolute values of the studentized deleted residuals (SDRESID) of these cases were slightly larger than or close to 2.

After running the OLS regression to request residual diagnostics, no additional unusual observations were identified. Only one case, Case 5, had already been identified through the review of the scatterplot; this case had a ZRESID value greater than 2.5.

Table 7b presents casewise diagnostics for the five cumulative binary logistic regression models (splits) for Sample II. One case (Case No. 19) was flagged as an outlier for the first

Table 6b: Casewise Diagnostics for Cumulative Splits, Sample I, n = 244

CUMSP1: P(Y ge 1)					CUMSP2: P(Y ge 2)	
Case	1059	1571	1698	2952	Case	1698
CUMSP4	0**	0**	1**	0**	CUMSP5	1**
Resid	-0.959	-0.990	0.993	-0.978	Resid	0.973
SResid	-2.546	-3.044	3.197	-2.770	SResid	2.746
Dev	-2.525	-3.037	3.155	-2.756	Dev	2.689
ZResid	-4.823	-9.980	11.997	-6.607	ZResid	6.014
DFB0	-0.401	-0.009	2.442	-0.706	DFB0	0.932
DFB1	0.101	-0.137	-0.165	0.098	DFB1	-0.057
DFB2	0.119	0.168	-0.270	0.185	DFB2	-0.113
DFB3	-0.064	0.017	-0.055	0.025	DFB3	-0.017
DFB4	0.002	0.258	0.026	0.241	DFB4	0.002
DFB5	-0.008	-0.061	-0.060	-0.037	DFB5	-0.040
DFB6	-0.204	-0.104	-0.970	-0.021	DFB6	-0.369
DFB7	0.003	-0.004	-0.034	0.004	DFB7	-0.012

CUMSP4: P(Y ge 4)			CUMSP5: P(Y ge 5)			
Case	1917	3414	Case	2103	3392	3414
CUMSP1	1**	1**	CUMSP2	1**	1**	1**
Resid	0.982	0.954	Resid	0.961	0.966	0.985
SResid	2.883	2.521	SResid	2.595	2.628	2.943
Dev	2.844	2.480	Dev	2.552	2.596	2.907
ZResid	7.493	4.547	ZResid	4.996	5.296	8.204
DFB0	0.671	0.594	DFB0	-0.018	3.052	2.649
DFB1	-0.073	0.091	DFB1	-0.215	-0.207	0.331
DFB2	-0.280	-0.081	DFB2	0.396	-0.353	-0.244
DFB3	0.456	-0.066	DFB3	0.305	-0.043	-0.173
DFB4	-0.145	0.115	DFB4	0.369	-0.388	0.313
DFB5	0.136	-0.540	DFB5	0.146	0.183	-1.050
DFB6	-0.043	-0.065	DFB6	0.101	-0.114	-0.229
DFB7	-0.009	-0.001	DFB7	-0.004	-0.034	-0.022

a S = Selected, U = Unselected cases, ** = Misclassified cases; b Cases with studentized residuals greater than 2.500 are listed; For CUMSP3: P(Y ge 3), the casewise plot is not produced because no outliers were found

MODEL DIAGNOSTICS FOR PO & PPO MODELS

Figure 2: OLS Observed Versus Predicted Values for Sample II, n = 278 Public School Children

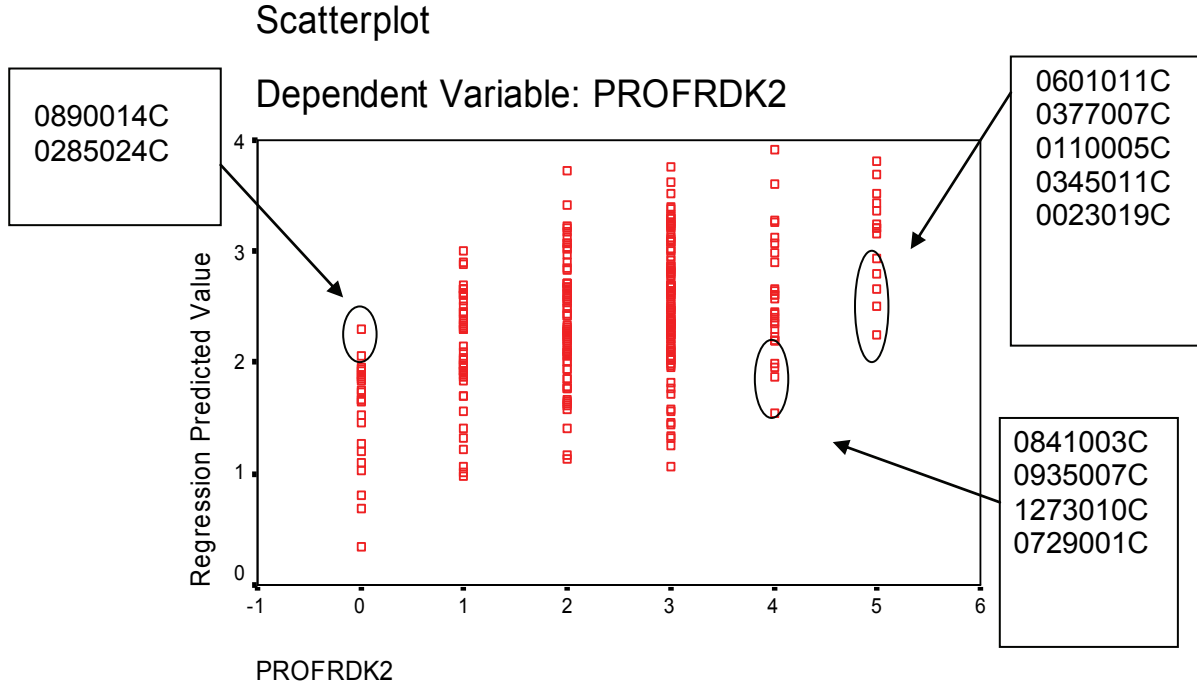


Table 7a: Casewise Diagnostics for OLS Model Sample II
From Scatterplot

Case No.	Child ID	ProfprdK2	Predicted	Residual	SDRESID	Lever	Mahal	Cook's D
5	0023019C	5	2.25	2.75	2.56	.02	6.18	.02
23	0110005C	5	2.66	2.34	2.17	.02	6.82	.02
44	0285024C	0	2.06	-2.06	-1.91	.03	7.55	.01
53	0345011C	5	2.50	2.50	2.32	.02	5.45	.02
58	0377007C	5	2.80	2.20	2.03	.01	3.91	.01
93	0601011C	5	2.94	2.06	1.90	.01	4.10	.01
118	0729001C	4	1.55	2.45	2.27	.02	6.90	.02
142	0841003C	4	1.98	2.01	1.87	.03	8.08	.01
153	0890014C	0	2.29	-2.29	-2.11	.01	3.93	.01
155	0935007C	4	1.95	2.05	1.91	.04	11.85	.02
226	1273010C	4	1.87	2.13	1.97	.02	6.08	.01

From Residual Diagnostics: None

Additional from Logistic Diagnostics and Plots

Case No.	Child ID	ProfprdK2	Predicted	Residual	SDRESID	Lever	Mahal	Cook's D
19	0057012C	0	1.94	-1.94	-1.79	.03	7.66	.01

Table 7b: Casewise Diagnostics for Cumulative Splits, Sample II Public School Children, n = 278

CUMSP1: P(Y ge 1)		CUMSP4: P(Y ge 4)		CUMSP5: P(Y ge 5)	
Case	19	Case	118	Case	5
CUMSP1	0**	CUMSP4	1**	CUMSP5	1**
Resid	-0.977	Resid	0.957	Resid	0.983
SResid	-2.776	SResid	2.532	SResid	2.880
Dev	-2.750	Dev	2.507	Dev	2.862
ZResid	-6.547	ZResid	4.706	ZResid	7.690
DFB0	-0.984	DFB0	0.119	DFB0	1.714
DFB1	-0.129	DFB1	0.052	DFB1	-0.112
DFB2	0.370	DFB2	0.046	DFB2	-0.174
DFB3	-0.052	DFB3	0.113	DFB3	-0.071
DFB4	-0.127	DFB4	-0.055	DFB4	0.245
DFB5	0.275	DFB5	-0.249	DFB5	0.188
DFB6	-0.247	DFB6	-0.031	DFB6	-0.238
DFB7	0.006	DFB7	0.000	DFB7	-0.021

Notes: a S = Selected, U = Unselected cases, ** = Misclassified cases; b Cases with studentized residuals greater than 2.500 are listed; For CUMSP2: P(Y ge 2) and CUMSP3: P(Y ge 3) the casewise plots are not produced because no outliers were found

cumulative split; this case was overpredicted – that is, the observation was in level 0 but predicted to be at or beyond level 1. Case No. 118 was flagged as an outlier for the fourth cumulative split, and was underpredicted (i.e., child scored in a higher category but was predicted into a category below 4). The third outlier identified was Case No. 5 at the fifth cumulative split; this child was also underpredicted. For comparison purposes, the OLS statistics for the one new case, Case No. 19, is included in the bottom section of Table 7a. No additional unusual values were determined from the index plots for Sample II (demonstrated below for Sample I).

Index Plots

Using the separate binary logistic models to mimic the data patterns of an ordinal model yields five different regressions for a K = 6 level ordinal outcome variable. Plotting each

model’s residuals or diagnostic summary values against the case number can enhance the search for extreme or unusual values; however, there will necessarily be multiple plots for each binary split.

Figure 3 contains a display of index plots for 8 residual measures: normalized residual, residual, deviance residual, logit residual, Cook’s analog, leverages, difference in deviance statistic and difference in the Pearson Chi-square statistic. These displays contain the value of the residual statistic on the vertical axis, and the case number horizontally. Strong peaks indicate extreme value in the residual score. In the first plot, it is easy to detect unusual scores, as noted in the Figure. The three marked cases were previously identified in the residual analyses for Sample I.

Figure 4 contains two of the plots recommended by Hosmer and Lemeshow (2000), namely, the change in Chi-square and

MODEL DIAGNOSTICS FOR PO & PPO MODELS

deviance fit statistics against \hat{p} . These graphs are a compilation of two curves, one representing $Y = 1$ (the downward curve, so that outliers are in the top left corner), and one representing $Y = 0$ (the upward curve, so outliers are in the top right corner). As illustrated in the first graph in Figure 4, previously identified cases are again indicated through

these index plots. The plots of change in the regression coefficients versus \hat{p} are not shown (with seven predictors and the intercept, there are eight graphs for each of the five logistic regression splits). Figures 5 through 12 provide the same plots as above for the remaining four logistic regression splits.

Figure 3: Index Plots for Diagnostic Statistics, Split 1 for Sample I

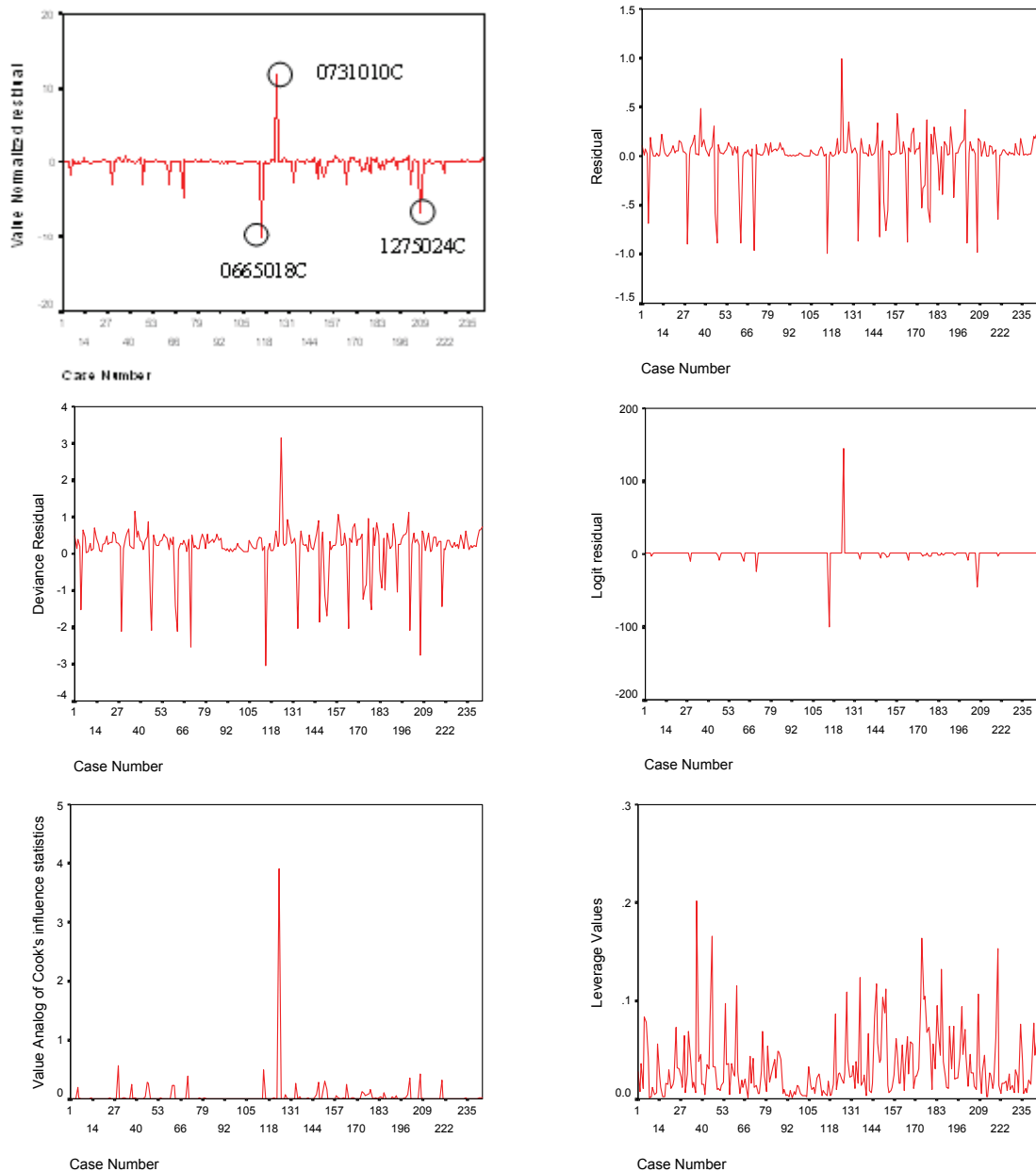


Figure 3 (continued): Index Plots for Diagnostic Statistics, Split 1 for Sample I

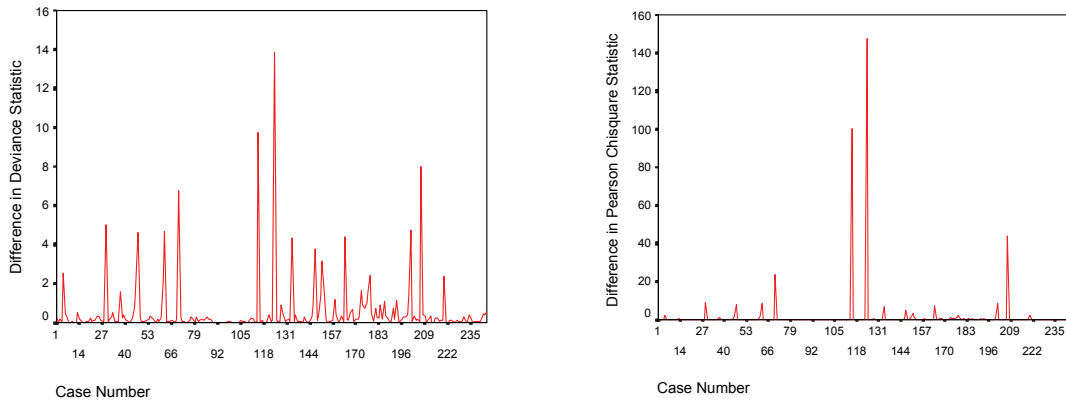
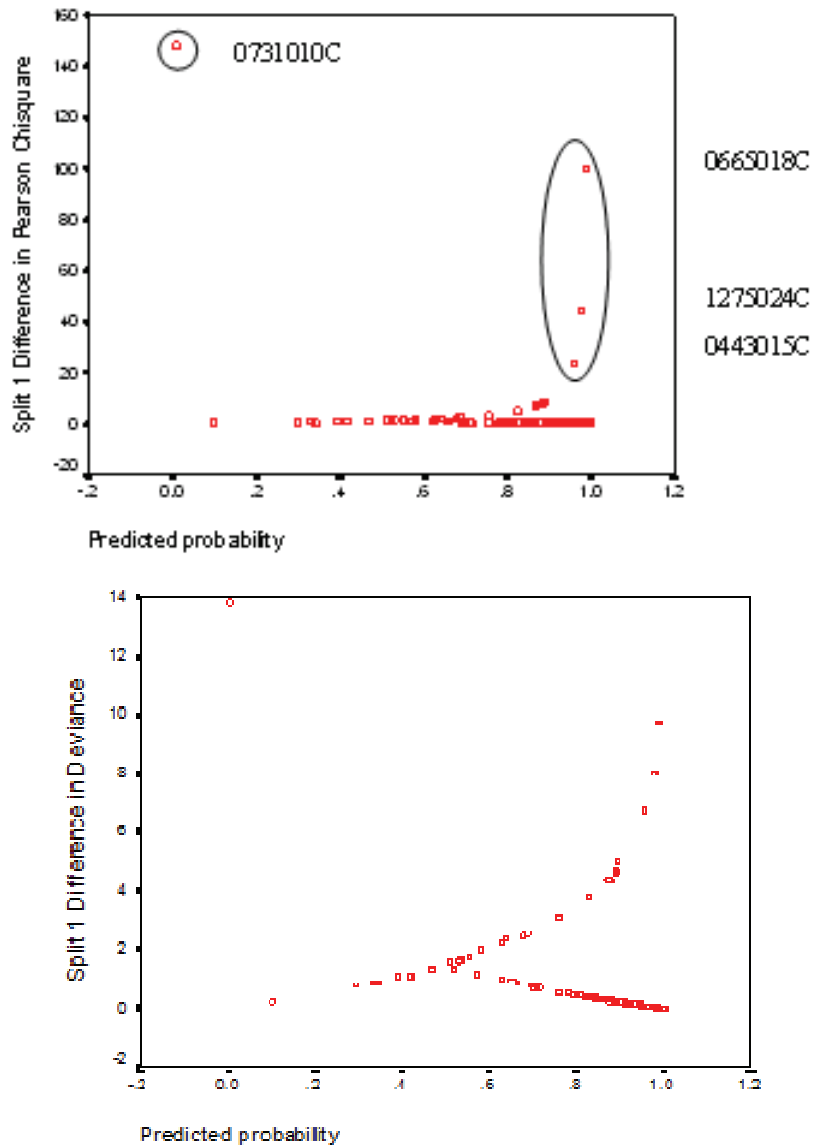


Figure 4: Change in Chi-Square and Deviance Fit against P-hat (Split 1), Sample I



MODEL DIAGNOSTICS FOR PO & PPO MODELS

Figure 5: Index Plots for Diagnostic Statistics, Split 2 for Sample I

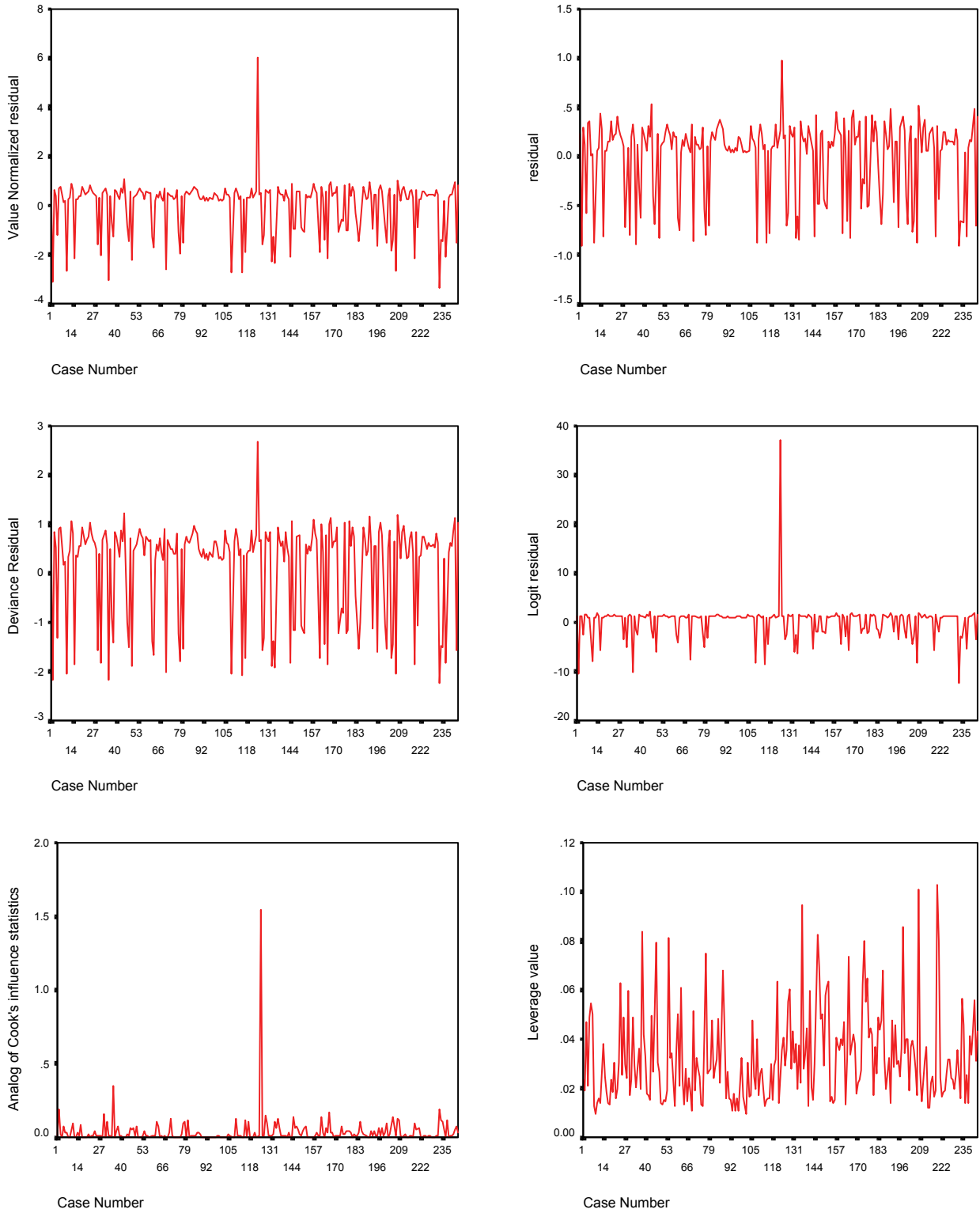
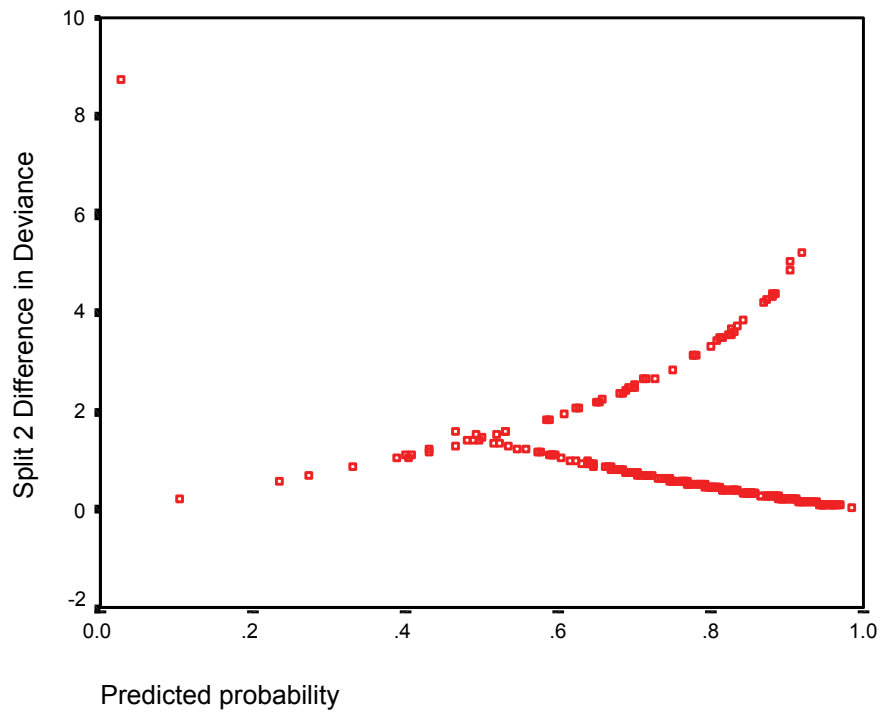
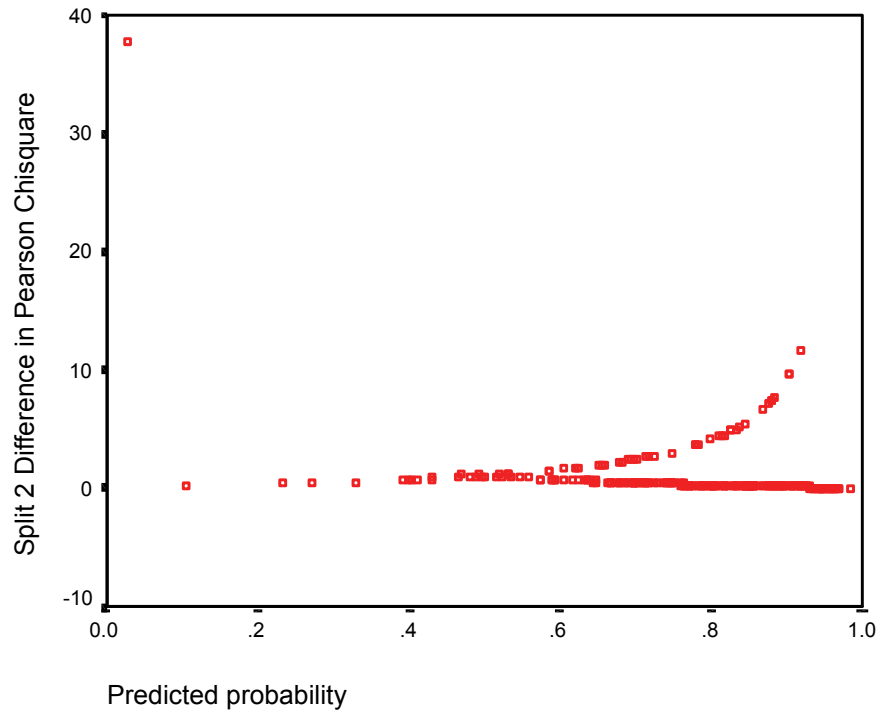


Figure 6: Change in Chi-Square and Deviance Fit against P-hat, Split 2, Sample I



MODEL DIAGNOSTICS FOR PO & PPO MODELS

Figure 7: Index Plots for Diagnostic Statistics, Split 3 for Sample I

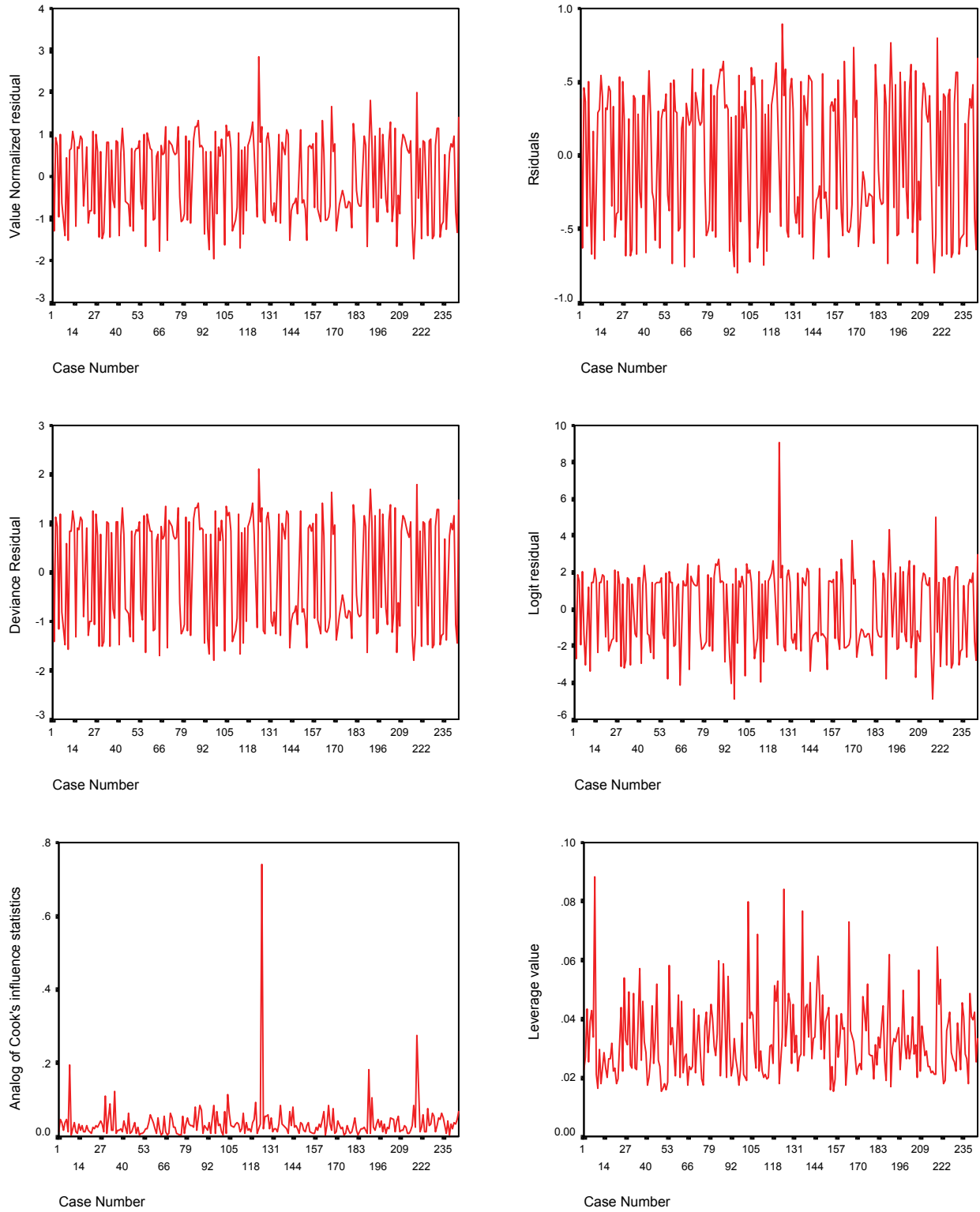
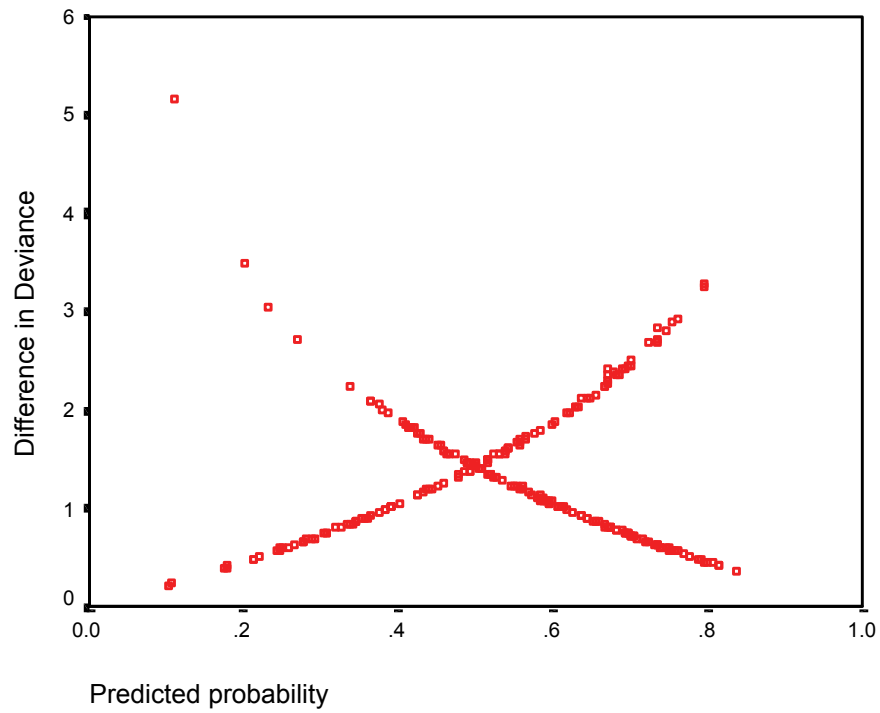
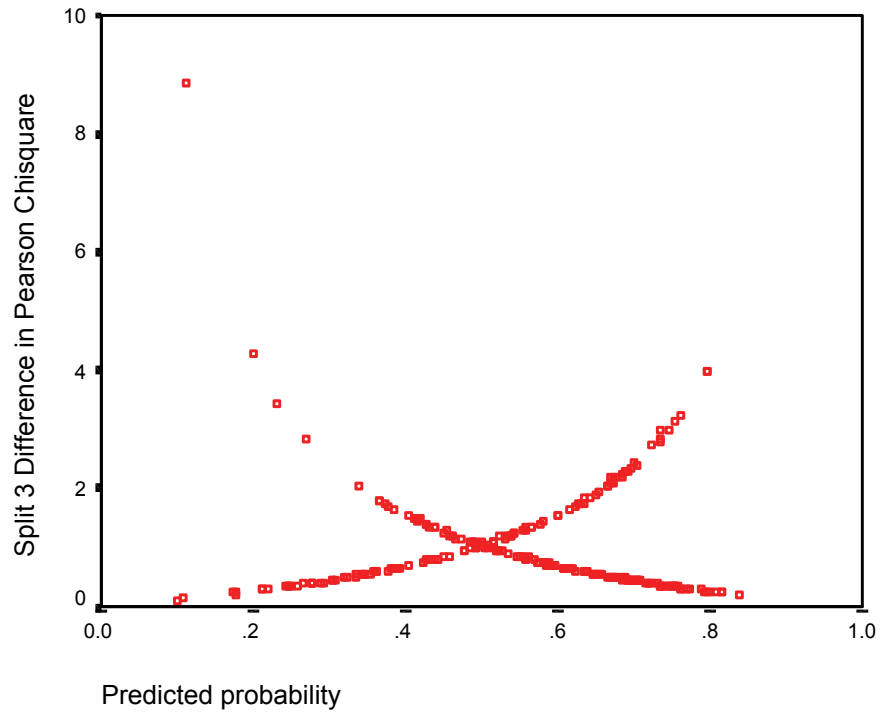


Figure 8: Change in Pearson and Deviance Fit against P-hat, Sample I, Split 3



MODEL DIAGNOSTICS FOR PO & PPO MODELS

Figure 9: Index Plots for Diagnostic Statistics, Sample I Split 4

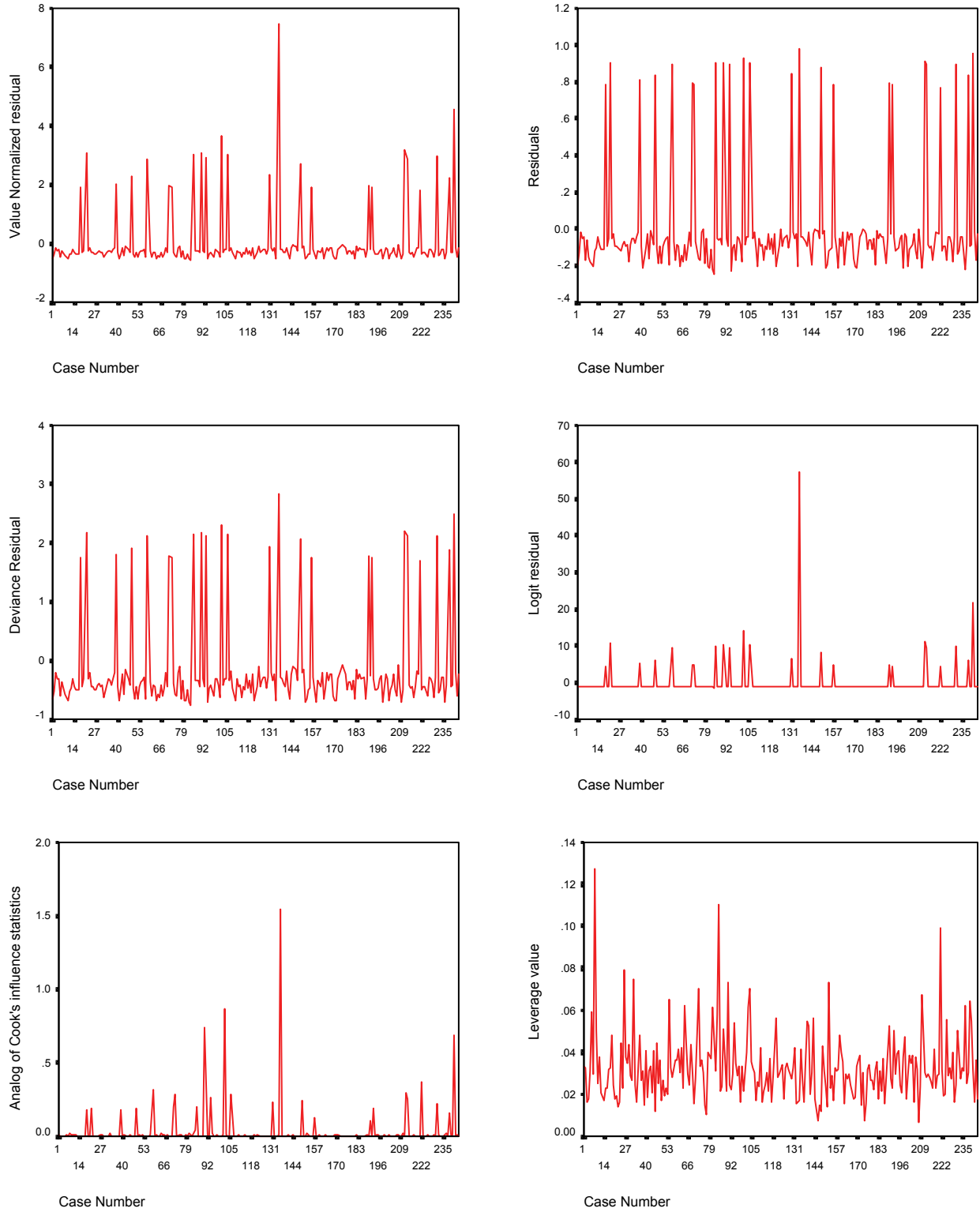
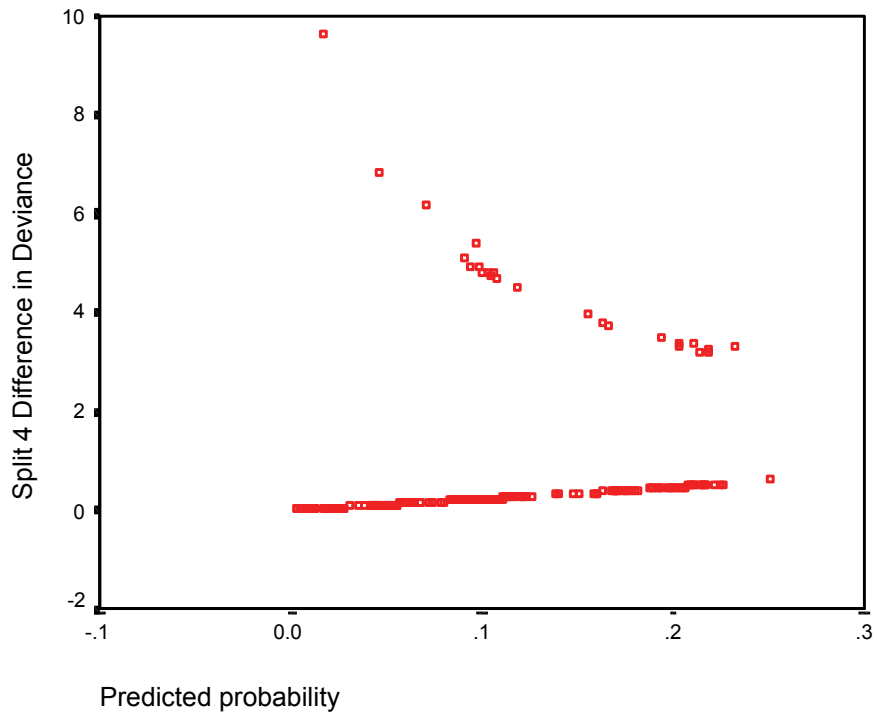
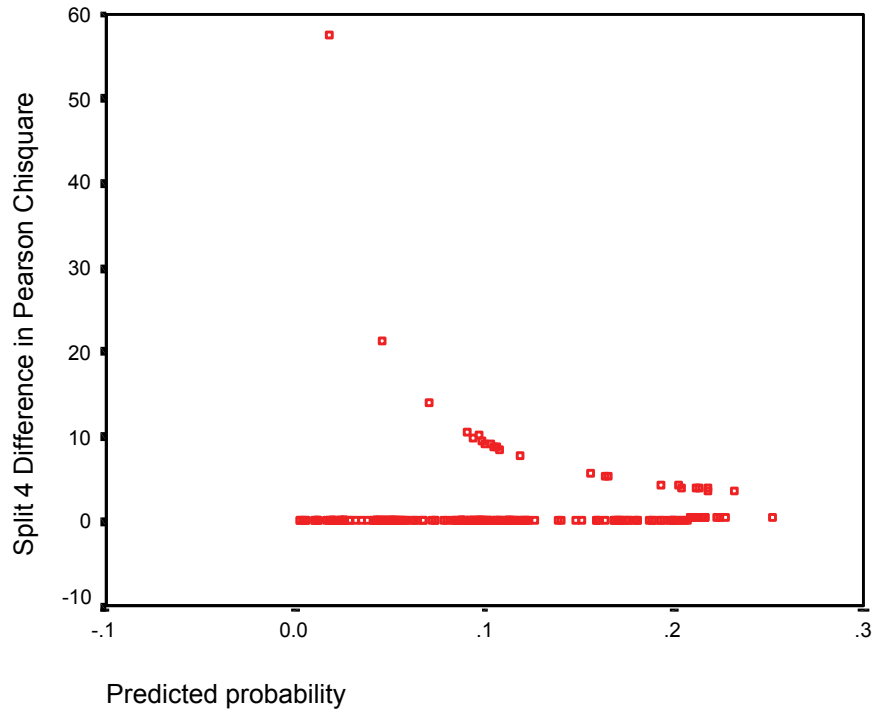


Figure 10: Change in Chi-Square and Deviance Fit, Sample I Split 4



MODEL DIAGNOSTICS FOR PO & PPO MODELS

Figure 11: Index Plots for Diagnostic Statistics, Sample I Split 5

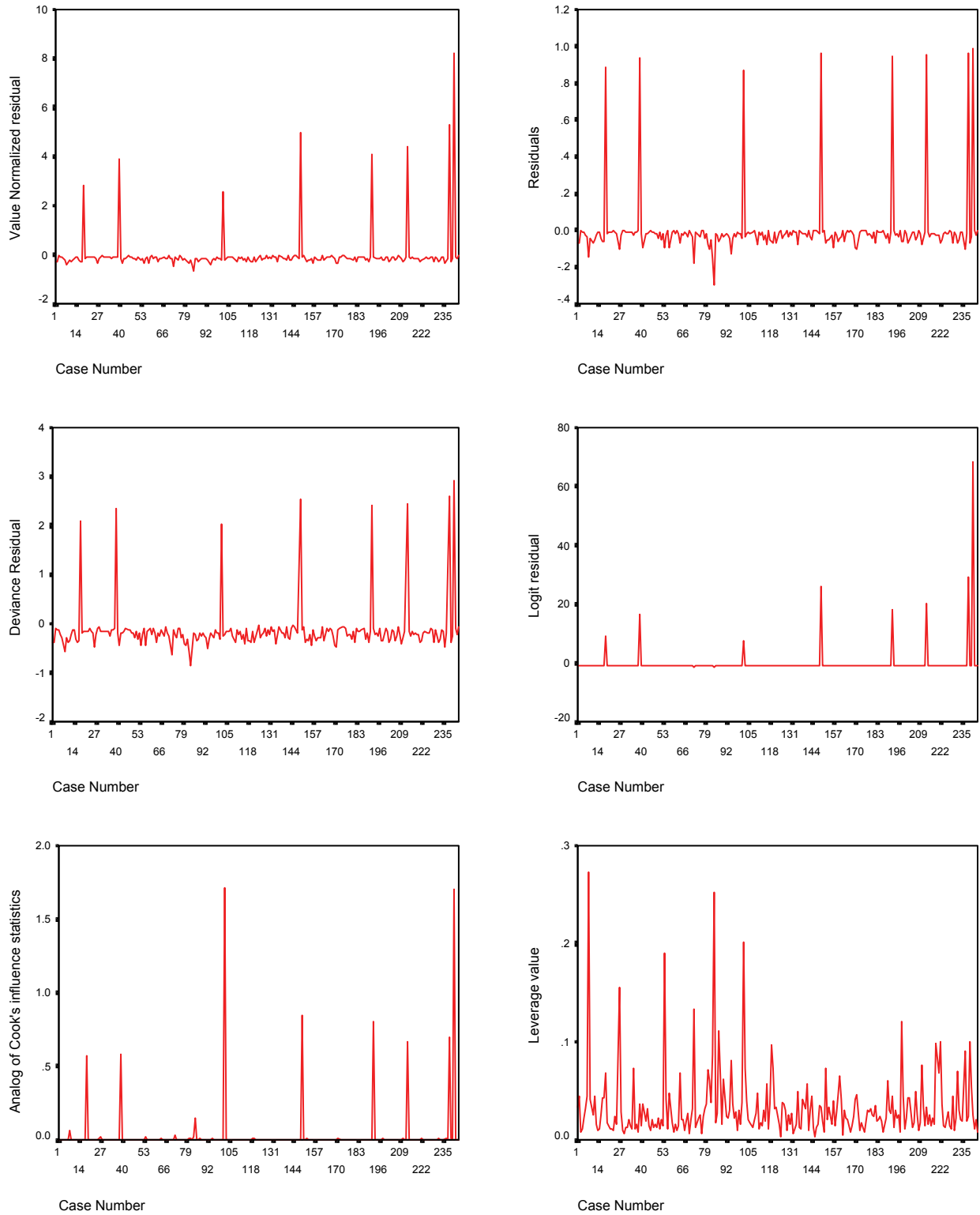
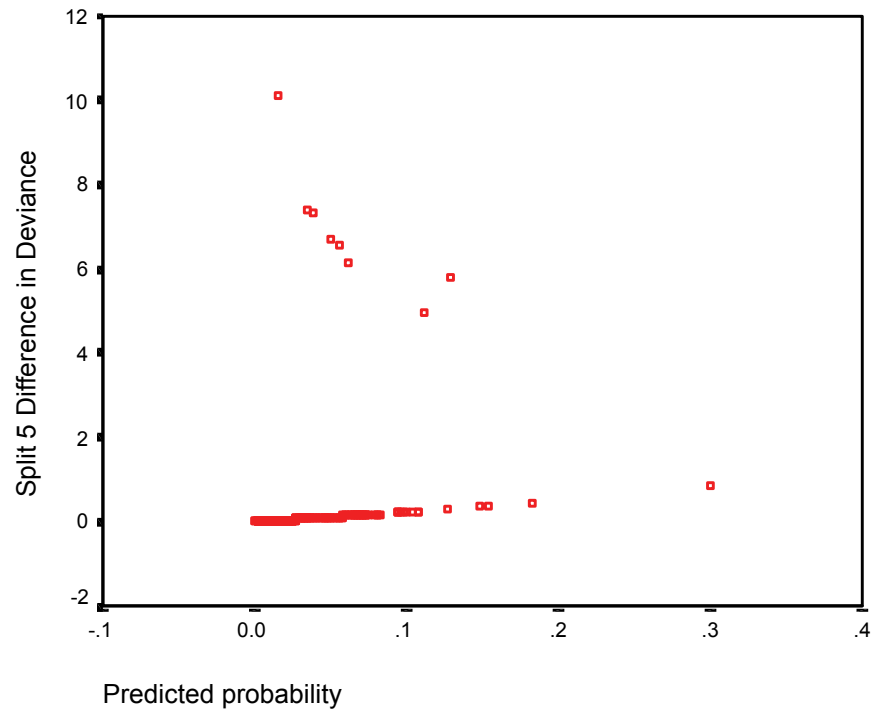
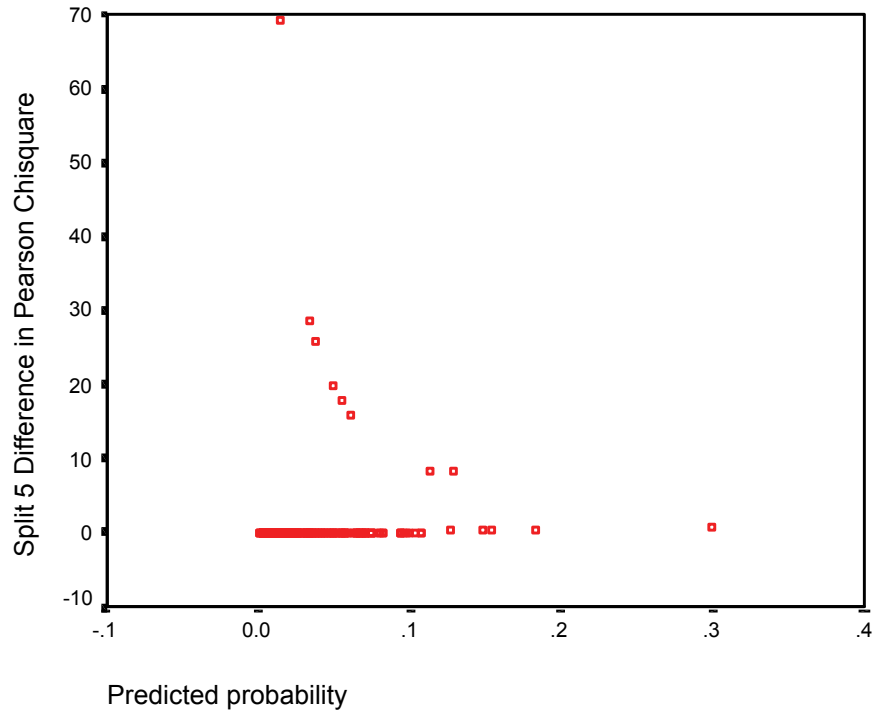


Figure 12: Change in Chi-Square and Deviance Fit, Sample I Split 5



MODEL DIAGNOSTICS FOR PO & PPO MODELS

Reviewing plots for residual analyses can be informative, but to match extreme or unusual observations in a plot to the original data can be time-consuming. In identifying the outliers in plots, SPSS does not label the cases directly; the analyst must use a two-step procedure (isolating the extreme value and matching that back to the original case). However, Stata has an option of adding labels to the points using the `mlabel` command. Figure 13 displays the plots of change in Pearson Chi-square against \hat{p} for five cumulative splits for Sample I using Stata. This option can facilitate easy identification of unusual cases from the plots.

Characteristics of Children Misfit by the Model

Investigation of the characteristics of the children who were identified as unusual, in terms of not being well represented by the model through at least one of the strategies employed, can ultimately help with understanding who the model is not providing a good fit for and why, and thus lead to development of better models that demonstrate stronger knowledge of the outcome for all persons. Tables 8 and 9 contain the values of the explanatory variables of interest for the collection of cases identified as unusual through the diagnostics applied to Sample I and Sample II, respectively. To summarize, for both samples, most identified children were female with SES below the standardized average of 0.

In general, the model tended to under-predict proficiency for some children who have theoretically assumed strikes against them, such as low SES or a large number of family risk characteristics. These cases tended to perform as well as or - in many cases - better than their peers. For example, child ID 0936002C (Case No. 2103) in sub-sample I is a female minority student who does not speak English at home and who attended half-day kindergarten rather than full-day kindergarten: this child's parents read to her at least three times per week and her actual reading proficiency level is 5. The logistic models predicted her to be in a lower category, as did the OLS model (level 2). Thus, some high achieving children do not have their reading proficiency adequately captured by the current model or current set of predictors.

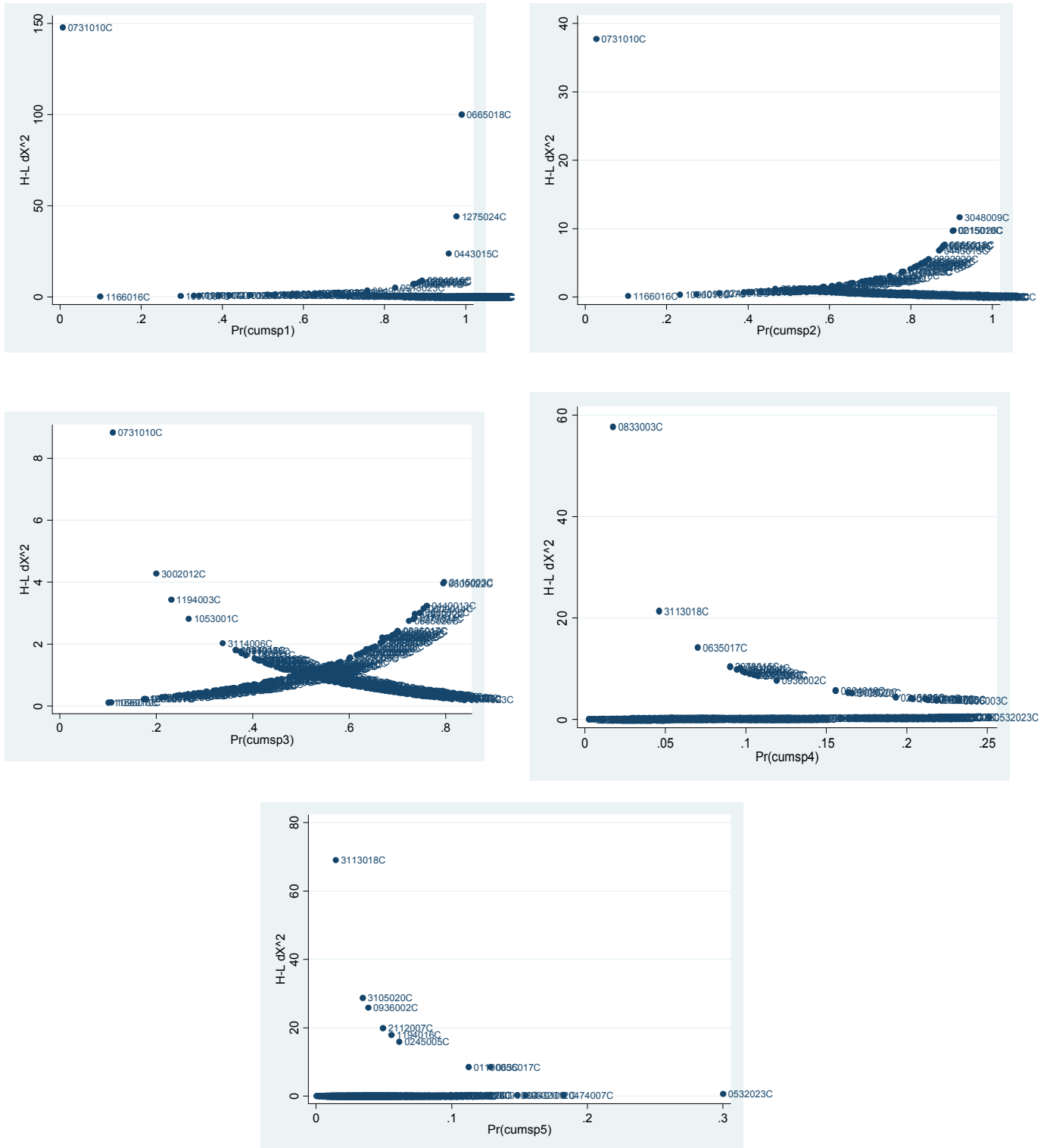
Conversely, some children who have perceived theoretical benefits in their favor, such as higher SES or no family risk characteristics, performed less well than the model predicts. For example, child ID 1275024C (Case No. 2952) in sub-sample I is a female non-minority student without any family risk factor, who attends full-day kindergarten, had parents read to her at least three times per week, and speaks English at home: this student's actual reading proficiency level is 0, but she is predicted to be in level 3.

Conclusion

Although poor predictions are inevitable in any modeling situation, the concern is that the typically limited range of the dependent variable for ordinal (or logistic) regression models may lead to more systematic under- or over-predictions relative to what might be expected with a continuous outcome. Identification of unusual cases or cases that are poorly fit by a particular model is only the first step in a residual analysis. After the cases are identified, the process turns towards understanding what characteristics of the collection of identified cases are associated with their corresponding under- or over-prediction from the model. This study has only examined the collection of data for the variables included in the models; it could be that a variable external to the model would better explain why some children are so strongly under- or over-predicted by the model.

For the proportional odds and partial-proportional odds model, the analysis of residuals was split into two components, given that residual analyses for ordinal regression analyses are not directly available. These were the OLS analysis of the ordinal outcome (treated as interval/continuous) and the analysis of residuals from the separate binary regression models forming the progression of the cumulative logit model. The latter approach has been advocated by many researchers (Hosmer & Lemeshow; 2000; Harrell, 2001; Long & Freese, 2006), but it remains unclear how well these cases may be fit by a specific ordinal model. That particular model was only approximated in the approach taken herein. Other work has investigated quality of goodness-of-fit tests for ordinal and multinomial data (Fagerland, Hosmer & Bofin, 2008; Pulkstenis & Robinson,

Figure13: Change in Pearson versus P-hat using Stata, Sample I, 5 Cumulative Splits



2004). However, this study focused specifically on data diagnostics for ordinal outcomes, which constitutes a preliminary but necessary step in the determination of model fit.

This article reviewed proportional odds model and partial proportional odds model and the use of some valuable strategies were demonstrated for identifying influential or unusual cases after fitting ordinal regression models. In particular, OLS strategies for preliminary residual detection were applied, followed by application of logistic regression diagnostics approaches for each cumulative binary model in the cumulative odds ordinal series. Results of these methods confirm that investigation of casewise fit to the proportional odds model can be very intensive.

OLS strategies can be a good first step in the diagnostic process, but does it not capture some extreme or influential cases. Differences across statistical packages in terms of availability of options for residual diagnostics can limit the kinds of analyses a researcher has available, thus choice of statistical package should be made with full understanding of the procedures and statistics available. Further, it was demonstrated that index and change plots can yield important information about who is not being fit well by these model and these plots can augment findings from other residual strategies.

Overall, it may be concluded that reliance on one method or approach to understanding residuals from an ordinal regression model can be very misleading to the researcher. Results from this study clearly make a case for the need to consider multiple strategies in determining quality of model fit – not just overall, but for individuals as well. Further studies and extensions to this research should consider the residuals obtained from the PO model based on predicted category (i.e., classification accuracy). In addition, it would be worthwhile to pursue the use of Monte Carlo techniques to examine residuals and to control over the nature of departure from the assumption of proportional odds or other model characteristics; for example, one question not answered in the current study is the degree to which outliers or extreme values affects the determination of the assumption of proportional odds.

Substantively, it may be reasonable to consider separate models for cases that are not well-represented by a general population-type ordinal regression model. From a policy perspective, findings suggest that individual student performance or proficiency can be easily misunderstood, and in the current climate of accountability, the repercussions from such a situation can have great impact. The approaches and strategies presented here could be used to effectively argue for support for intervention programs, including gifted-education programs, or to support improved funding for special education or second-language acquisition programs. Only through residual diagnostics would the children who are left far behind, or who score far beyond their peers, be readily identified.

Proposed Diagnostic Guidelines for Ordinal Regression Models

1. Residuals from both OLS and Binary Logistic Models provide a good first look at the potential for influential or unusual cases from an ordinal model.
2. OLS does not capture all the unique unusual values; neither do the corresponding binary logistic analyses. Thus, researchers need to be aware of the potential to miss important misfit cases and counteract this possibility by viewing/plotting as many different diagnostic statistics as possible, particularly for the binary models corresponding to a given ordinal approach (e.g., proportional odds)
3. Graphical strategies, while extensive and often time-consuming, can tell the researcher more about their data than a single summary statistic.
4. Researchers should make a commitment early on to include residual diagnostics in all their presented or published papers. It is easy to mislead oneself, one's audience, and various research stakeholders when residual diagnostic strategies are ignored.

References

- Agresti, A. (2002). *Categorical data analysis (2nd Ed.)*. New York: John Wiley & Sons.
- Agresti, A. (2007). *An introduction to categorical data analysis (2nd Ed.)*. New York: John Wiley & Sons.
- Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute, Inc.
- Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, 26, 1323-1333.
- Armstrong, B. B., & Sloan, M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology*, 129(1), 191-204.
- Bender, R., & Benner, A. (2000). Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal*, 42(6), 677-699.
- Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology*, 51(10), 809-816.
- Brant, (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46, 1171-1178.
- Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.
- Fagerland, M. W., Hosmer, D. W., & Bofin, A. M. (2008). Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine*, 27, 4238-4253.
- Fox, J. (1991). *Regression diagnostics*. Thousand Oaks, CA: Sage.
- Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer Verlag.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression (2nd Ed.)*. New York: John Wiley & Sons.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata (2nd Ed.)*. Texas: Stata Press.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models (2nd Ed.)*. London: Chapman and Hall.
- Menard, S. (1995). *Applied Logistic Regression Analysis*. Thousand Oaks, CA: Sage.
- NCES (2000). *America's kindergarteners*. Retrieved August 10, 2009, from <http://www.nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2000070>.
- O'Connell, A. A. (2000). Methods for modeling ordinal outcome variables. *Measurement and Evaluation in Counseling and Development*, 33(3), 170-193.
- O'Connell, A. A., Liu, X., Zhao, J., & Goldstein, J. (2004). *Understanding residual diagnostics of ordinal regression models*. Paper presented at the 35th Northeastern Educational Research Association annual meeting, October 2004, Kerhonkson, New York.
- O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks: SAGE.
- Peterson, B. L., & Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39(3), 205-217.
- Powers, D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. San Diego, CA: Academic Press.
- Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705-724.
- Pulkstenis, E., & Robinson, T. J. (2004). Goodness-of-fit tests for ordinal response regression models. *Statistics in Medicine*, 23, 999-1014.