Chapter 14
Case Study in Ordinal Regression, Data Reduction, and Penalization

This case study is taken from Harrell et al.[272] which described a World Health Organization study[439] in which vital signs and a large number of clinical signs and symptoms were used to develop a predictive model for an ordinal response. This response consists of laboratory assessments of diagnosis and severity of illness related to pneumonia, meningitis, and sepsis. Much of the modeling strategy given in Chapter 4 was used to develop the model, with additional emphasis on penalized maximum likelihood estimation (Section 9.10). The following laboratory data are used in the response: cerebrospinal fluid (CSF) culture from a lumbar puncture (LP), blood culture (BC), arterial oxygen saturation (SaO_2, a measure of lung dysfunction), and chest X-ray (CXR). The sample consisted of 4552 infants aged 90 days or less.

This case study covers these topics:

1. definition of the ordinal response (Section 14.1);
2. scoring and clustering of clinical signs (Section 14.2);
3. testing adequacy of weights specified by subject-matter specialists and assessing the utility of various scoring schemes using a tentative ordinal logistic model (Section 14.3);
4. assessing the basic ordinality assumptions and examining the proportional odds and continuation ratio (PO and CR) assumptions separately for each predictor (Section 14.4);
5. deriving a tentative PO model using cluster scores and regression splines (Section 14.5);
6. using residual plots to check PO, CR, and linearity assumptions (Section 14.6);
7. examining the fit of a CR model (Section 14.7);
8. utilizing an extended CR model to allow some or all of the regression coefficients to vary with cutoffs of the response level as well as to provide formal tests of constant slopes (Section 14.8);

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_14

Table 14.1 Ordinal Outcome Scale

Outcome Level Y	Definition	n	Fraction in Outcome Level		
			BC, CXR Indicated ($n = 2398$)	Not Indicated ($n = 1979$)	Random Sample ($n = 175$)
0	None of the below	3551	0.63	0.96	0.91
1	$90\% \leq SaO_2 < 95\%$ or CXR+	490	0.17	0.04[a]	0.05
2	BC+ or CSF+ or $SaO_2 < 90\%$	511	0.21	0.00[b]	0.03

[a] SaO_2 was measured but CXR was not done

[b] *Assumed* zero since neither BC nor LP were done.

9. using penalized maximum likelihood estimation to improve accuracy (Section 14.9);
10. approximating the full model by a sub-model and drawing a nomogram on the basis of the sub-model (Section 14.10); and
11. validating the ordinal model using the bootstrap (Section 14.11).

14.1 Response Variable

To be a candidate for BC and CXR, an infant had to have a clinical indication for one of the three diseases, according to prespecified criteria in the study protocol ($n = 2398$). Blood work-up (but not necessarily LP) and CXR was also done on a random sample intended to be 10% of infants having no signs or symptoms suggestive of infection ($n = 175$). Infants with signs suggestive of meningitis had LP done. All 4552 infants received a full physical exam and standardized pulse oximetry to measure SaO_2. The vast majority of infants getting CXR had the X-rays interpreted by three independent radiologists.

The analyses that follow are not corrected for verification bias[687] with respect to BC, LP, and CXR, but Section 14.1 has some data describing the extent of the problem, and the problem is reduced by conditioning on a large number of covariates.

Patients were assigned to the worst qualifying outcome category. Table 14.1 shows the definition of the ordinal outcome variable Y and shows the distribution of Y by the lab work-up strategy.

The effect of verification bias is a false negative fraction of 0.03 for $Y = 2$, from comparing the detection fraction of zero for $Y = 2$ in the "Not Indicated" group with the observed positive fraction of 0.03 in the random sample that was fully worked up. The extent of verification bias in $Y = 1$ is $0.05 - 0.04 = 0.01$. These biases are ignored in this analysis.

14.2 Variable Clustering

Forty-seven clinical signs were collected for each infant. Most questionnaire items were scored as a single variable using equally spaced codes, with 0 to 3 representing, for example, sign not present, mild, moderate, severe. The resulting list of clinical signs with their abbreviations is given in Table 14.2. The signs are organized into clusters as discussed later.

Table 14.2 Clinical Signs

Cluster Name	Sign Abbreviation	Name of Sign	Values
bul.conv	abb	bulging fontanel	0-1
	convul	hx convulsion	0-1
hydration	abk	sunken fontanel	0-1
	hdi	hx diarrhoea	0-1
	deh	dehydrated	0-2
	stu	skin turgor	0-2
	dcp	digital capillary refill	0-2
drowsy	hcl	less activity	0-1
	qcr	quality of crying	0-2
	csd	drowsy state	0-2
	slpm	sleeping more	0-1
	wake	wakes less easily	0-1
	aro	arousal	0-2
	mvm	amount of movement	0-2
agitated	hcm	crying more	0-1
	slpl	sleeping less	0-1
	con	consolability	0-2
	csa	agitated state	0-1
crying	hcm	crying more	0-1
	hcs	crying less	0-1
	qcr	quality of crying	0-2
	smi2	smiling ability $\times$ age $>$ 42 days	0-2
reffort	nfl	nasal flaring	0-3
	lcw	lower chest in-drawing	0-3
	gru	grunting	0-2
	ccy	central cyanosis	0-1
stop.breath	hap	hx stop breathing	0-1
	apn	apnea	0-1
ausc	whz	wheezing	0-1
	coh	cough heard	0-1
	crs	crepitation	0-2
hxprob	hfb	fast breathing	0-1
	hdb	difficulty breathing	0-1
	hlt	mother report resp. problems	none, chest, other
feeding	hfa	hx abnormal feeding	0-3
	absu	sucking ability	0-2
	afe	drinking ability	0-2
labor	chi	previous child died	0-1
	fde	fever at delivery	0-1
	ldy	days in labor	1-9
	twb	water broke	0-1
abdominal	adb	abdominal distension	0-4
	jau	jaundice	0-1
	omph	omphalitis	0-1
fever.ill	illd	age-adjusted no. days ill	
	hfe	hx fever	0-1
pustular	conj	conjunctivitis	0-1
	oto	otoscopy impression	0-2
	puskin	pustular skin rash	0-1

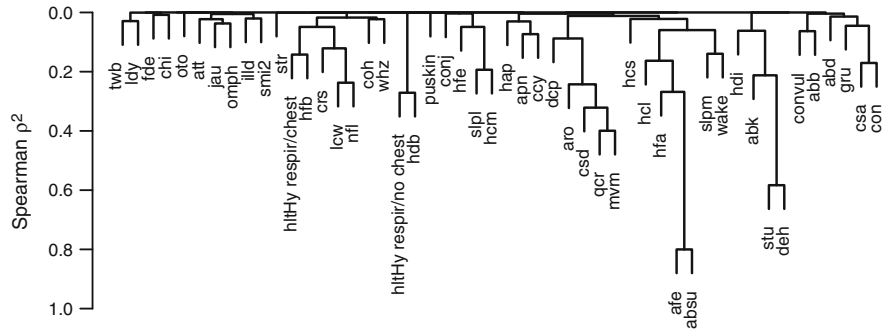

Fig. 14.1 Hierarchical variable clustering using Spearman ρ^2 as a similarity measure for all pairs of variables. Note that since the `hlt` variable was nominal, it is represented by two dummy variables here.

Here, `hx` stands for history, `ausc` for auscultation, and `hxprob` for history of problems. Two signs (`qcr, hcm`) were listed twice since they were later placed into two clusters each.

Next, hierarchical clustering was done using the matrix of squared Spearman rank correlation coefficients as the similarity matrix. The `varclus` R function was used as follows.

```
require(rms)
```

```
getHdata(ari)  # defines ari, Sc, Y, Y.death
vclust ←
  varclus(~ illd + hlt + slpm + slpl + wake + convul + hfa +
            hfb  + hfe + hap  + hcl  + hcm  + hcs    + hdi +
            fde  + chi + twb  + ldy  + apn  + lcw    + nfl +
            str  + gru + coh  + ccy  + jau  + omph   + csd +
            csa  + aro + qcr  + con  + att  + mvm    + afe +
            absu + stu + deh  + dcp  + crs  + abb    + abk +
            whz  + hdb + smi2 + abd  + conj + oto    + puskin,
          data=ari)
plot(vclust)  # Figure 14.1
```

The output appears in Figure 14.1. This output served as a starting point for clinicians to use in constructing more meaningful clinical clusters. The clusters in Table 14.2 were the consensus of the clinicians who were the investigators in the WHO study. Prior subject matter knowledge plays a key role at this stage in the analysis.

14.3 Developing Cluster Summary Scores

The clusters listed in Table 14.2 were first scored by the first principal component of `transcan`-transformed signs, denoted by PC_1. Knowing that the resulting weights may be too complex for clinical use, the primary reasons

Table 14.3 Clinician Combinations, Rankings, and Scorings of Signs

Cluster	Combined/Ranked Signs in Order of Severity	Weights
bul.conv	abb ∪ convul	0–1
drowsy	hcl, qcr>0, csd>0 ∪ slpm ∪ wake, aro>0, mvm>0	0–5
agitated	hcm, slpl, con=1, csa, con=2	0, 1, 2, 7, 8, 10
reffort	nfl>0, lcw>1, gru=1, gru=2, ccy	0–5
ausc	whz, coh, crs>0	0–3
feeding	hfa=1, hfa=2, hfa=3, absu=1 ∪ afe=1, absu=2 ∪ afe=2	0–5
abdominal	jau ∪ abd>0 ∪ omph	0–1

for analyzing the principal components were to see if some of the clusters could be removed from consideration so that the clinicians would not spend time developing scoring rules for them. Let us "peek" at Y to assist in scoring clusters at this point, but to do so in a very structured way that does not involve the examination of a large number of individual coefficients.

To judge any cluster scoring scheme, we must pick a tentative outcome model. For this purpose we chose the PO model. By using the 14 PC_1s corresponding to the 14 clusters, the fitted PO model had a likelihood ratio (LR) χ^2 of 1155 with 14 d.f., and the predictive discrimination of the clusters was quantified by a Somers' D_{xy} rank correlation between $X\hat{\beta}$ and Y of 0.596. The following clusters were not statistically important predictors and we assumed that the lack of importance of the PC_1s in predicting Y (adjusted for the other PC_1s) justified a conclusion that no sign within that cluster was clinically important in predicting Y: `hydration, hxprob, pustular, crying, fever.ill, stop.breath, labor`. This list was identified using a backward step-down procedure on the full model. The total Wald χ^2 for these seven PC_1s was 22.4 ($P = 0.002$). The reduced model had LR $\chi^2 = 1133$ with 7 d.f., $D_{xy} = 0.591$. The bootstrap validation in Section 14.11 penalizes for examining all candidate predictors.

The clinicians were asked to rank the clinical severity of signs within each potentially important cluster. During this step, the clinicians also ranked severity levels of some of the component signs, and some cluster scores were simplified, especially when the signs within a cluster occurred infrequently. The clinicians also assessed whether the severity points or weights should be equally spaced, assigning unequally spaced weights for one cluster (`agitated`). The resulting rankings and sign combinations are shown in Table 14.3. The signs or sign combinations separated by a comma are treated as separate categories, whereas some signs were unioned ("or"–ed) when the clinicians deemed them equally important. As an example, if an additive cluster score was to be used for `drowsy`, the scorings would be 0 = none present, 1 = `hcl`, 2 = `qcr>0`, 3 = `csd>0` or `slpm` or `wake`, 4 = `aro>0`, 5 = `mvm>0` and the scores would be added.

This table reflects some data reduction already (unioning some signs and selection of levels of ordinal signs) but more reduction is needed. Even after

Table 14.4 Predictive information of various cluster scoring strategies. AIC is on the likelihood ratio χ^2 scale.

Scoring Method	LR χ^2	d.f.	AIC
PC_1 of each cluster	1133	7	1119
Union of all signs	1045	7	1031
Union of higher categories	1123	7	1109
Hierarchical (worst sign)	1194	7	1180
Additive, equal weights	1155	7	1141
Additive using clinician weights	1183	7	1169
Hierarchical, data-driven weights	1227	25	1177

signs are ranked within a cluster, there are various ways of assigning the cluster scores. We investigated six methods. We started with the purely statistical approach of using PC_1 to summarize each cluster. Second, all sign combinations within a cluster were unioned to represent a 0/1 cluster score. Third, only sign combinations thought by the clinicians to be severe were unioned, resulting in `drowsy=aro>0` or `mvm>0`, `agitated=csa` or `con=2`, `reffort=lcw>1` or `gru>0` or `ccy`, `ausc=crs>0`, and `feeding=absu>0` or `afe>0`. For clusters that are not scored 0/1 in Table 14.3, the fourth summarization method was a hierarchical one that used the weight of the worst applicable category as the cluster score. For example, if `aro=1` but `mvm=0`, `drowsy` would be scored as 4. The fifth method counted the number of positive signs in the cluster. The sixth method summed the weights of all signs or sign combinations present. Finally, the worst sign combination present was again used as in the second method, but the points assigned to the category were data-driven ones obtained by using extra dummy variables. This provided an assessment of the adequacy of the clinician-specified weights. By comparing rows 4 and 7 in Table 14.4 we see that response data-driven sign weights have a slightly worse AIC, indicating that the number of extra β parameters estimated was not justified by the improvement in χ^2. The hierarchical method, using the clinicians' weights, performed quite well. The only cluster with inadequate clinician weights was `ausc`—see below. The PC_1 method, without any guidance, performed well, as in[268]. The only reasons not to use it are that it requires a coefficient for every sign in the cluster and the coefficients are not translatable into simple scores such as $0, 1, \ldots$.

Representation of clusters by a simple union of selected signs or of all signs is inadequate, but otherwise the choice of methods is not very important in terms of explaining variation in Y. We chose the fourth method, a hierarchical severity point assignment (using weights that were prespecified by the clinicians), for its ease of use and of handling missing component variables (in most cases) and potential for speeding up the clinical exam (examining to detect more important signs first). Because of what was learned regarding the relationship between `ausc` and Y, we modified the `ausc` cluster score

by redefining it as `ausc=crs>0` (crepitations present). Note that neither the "tweaking" of `ausc` nor the examination of the seven scoring methods displayed in Table 14.4 is taken into account in the model validation.

14.4 Assessing Ordinality of Y for each X, and Unadjusted Checking of PO and CR Assumptions

Section 13.2 described a graphical method for assessing the ordinality assumption for Y separately with respect to each X, and for assessing PO and CR assumptions individually. Figure 14.2 is an example of such displays. For this dataset we expect strongly nonlinear effects for `temp, rr`, and `hrat`, so for those predictors we plot the mean absolute differences from suitable "normal" values as an approximate solution.

```
Sc ← transform(Sc,
                ausc = 1 * (ausc == 3),
                bul.conv = 1 * (bul.conv == 'TRUE'),
                abdominal = 1 * (abdominal == 'TRUE'))
plot.xmean.ordinaly(Y ~ age + abs(temp-37) + abs(rr-60) +
                    abs(hrat-125) + waz + bul.conv + drowsy +
                    agitated + reffort + ausc + feeding +
                    abdominal, data=Sc, cr=TRUE,
                    subn=FALSE, cex.points=.65) # Figure 14.2
```

The plot is shown in Figure 14.2. Y does not seem to operate in an ordinal fashion with respect to `age`, |`rr`−60|, or `ausc`. For the other variables, ordinality holds, and PO holds reasonably well for the other variables. For heart rate, the PO assumption appears to be satisfied perfectly. CR model assumptions appear to be more tenuous than PO assumptions, when one variable at a time is fitted.

14.5 A Tentative Full Proportional Odds Model

Based on what was determined in Section 14.3, the original list of 47 signs was reduced to seven predictors: two unions of signs (`bul.conv, abdominal`), one single sign (`ausc`), and four "worst category" point assignments (`drowsy, agitated, reffort, feeding`). Seven clusters were dropped for the time being because of weak associations with Y. Such a limited use of variable selection reduces the severe problems inherent with that technique.

At this point in model development add to the model age and vital signs: `temp` (temperature), `rr` (respiratory rate), `hrat` (heart rate), and `waz`, weight-for-age Z-score. Since `age` was expected to modify the interpretation of `temp`,

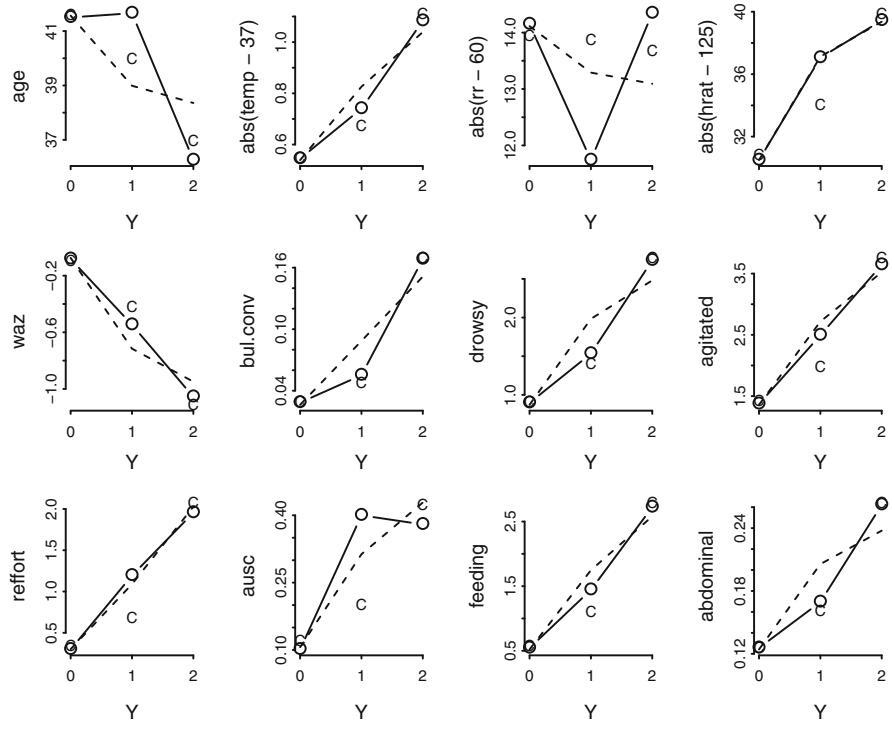

Fig. 14.2 Examination of the ordinality of Y for each predictor by assessing how varying Y relate to the mean X, and whether the trend is monotonic. Solid lines connect the simple stratified means, and dashed lines connect the estimated expected value of $X|Y = j$ given that PO holds. Estimated expected values from the CR model are marked with Cs.

`rr`, and `hrat`, and interactions between continuous variables would be difficult to use in the field, we categorized `age` into three intervals: 0–6 days ($n = 302$), 7–59 days ($n = 3042$), and 60–90 days ($n = 1208$).[a]

```
Sc$ageg ← cut2(Sc$age, c(7, 60))
```

The new variables `temp, rr, hrat, waz` were missing in, respectively, $n =$ 13, 11, 147, and 20 infants. Since the three vital sign variables are somewhat correlated with each other, customized single imputation models were developed to impute all the missing values without assuming linearity or even monotonicity of any of the regressions.

```
vsign.trans ← transcan(~ temp + hrat + rr, data=Sc,
                       imputed=TRUE, pl=FALSE)
```

[a] These age intervals were also found to adequately capture most of the interaction effects.

```
Convergence criterion:2.222 0.643 0.191 0.056 0.016
Convergence in 6 iterations
R^2 achieved in predicting each variable:

 temp  hrat    rr
0.168 0.160 0.066

Adjusted R^2:

 temp  hrat    rr
0.167 0.159 0.064
```

```
Sc ← transform(Sc,
                temp = impute(vsign.trans, temp),
                hrat = impute(vsign.trans, hrat),
                rr   = impute(vsign.trans, rr))
```

After `transcan` estimated optimal restricted cubic spline transformations, `temp` could be predicted with adjusted $R^2 = 0.17$ from `hrat` and `rr`, `hrat` could be predicted with adjusted $R^2 = 0.16$ from `temp` and `rr`, and `rr` could be predicted with adjusted R^2 of only 0.06. The first two R^2, while not large, mean that customized imputations are more efficient than imputing with constants. Imputations on `rr` were closer to the median `rr` of 48/minute as compared with the other two vital signs whose imputations have more variation. In a similar manner, `waz` was imputed using `age`, birth weight, head circumference, body length, and prematurity (adjusted R^2 for predicting `waz` from the others was 0.74). The continuous predictors `temp, hrat, rr` were not assumed to linearly relate to the log odds that $Y \geq j$. Restricted cubic spline functions with five knots for `temp,rr` and four knots for `hrat,waz` were used to model the effects of these variables:

```
f1 ← lrm(Y ~ ageg*(rcs(temp,5)+rcs(rr,5)+rcs(hrat,4)) +
          rcs(waz,4) + bul.conv + drowsy + agitated +
          reffort + ausc + feeding + abdominal,
          data=Sc, x=TRUE, y=TRUE)
# x=TRUE, y=TRUE used by resid() below
print(f1, latex=TRUE, coefs=5)
```

Logistic Regression Model

```
lrm(formula = Y ~ ageg * (rcs(temp, 5) + rcs(rr, 5) + rcs(hrat,
    4)) + rcs(waz, 4) + bul.conv + drowsy + agitated + reffort +
    ausc + feeding + abdominal, data = Sc, x = TRUE, y = TRUE)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	4552	LR χ^2	1393.18	R^2	0.355	C	0.826
0	3551	d.f.	45	g	1.485	D_{xy}	0.653
1	490	$\Pr(>\chi^2)$	< 0.0001	g_r	4.414	γ	0.654
2	511			g_p	0.225	τ_a	0.240
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	2×10^{-6}			Brier	0.120		

Table 14.5 Wald statistics from the proportional odds model

	χ^2	d.f.	P
ageg (Factor+Higher Order Factors)	41.49	24	0.0147
All Interactions	40.48	22	0.0095
temp (Factor+Higher Order Factors)	37.08	12	0.0002
All Interactions	6.77	8	0.5617
Nonlinear (Factor+Higher Order Factors)	31.08	9	0.0003
rr (Factor+Higher Order Factors)	81.16	12	< 0.0001
All Interactions	27.37	8	0.0006
Nonlinear (Factor+Higher Order Factors)	27.36	9	0.0012
hrat (Factor+Higher Order Factors)	19.00	9	0.0252
All Interactions	8.83	6	0.1836
Nonlinear (Factor+Higher Order Factors)	7.35	6	0.2901
waz	35.82	3	< 0.0001
Nonlinear	13.21	2	0.0014
bul.conv	12.16	1	0.0005
drowsy	17.79	1	< 0.0001
agitated	8.25	1	0.0041
reffort	63.39	1	< 0.0001
ausc	105.82	1	< 0.0001
feeding	30.38	1	< 0.0001
abdominal	0.74	1	0.3895
ageg × temp (Factor+Higher Order Factors)	6.77	8	0.5617
Nonlinear	6.40	6	0.3801
Nonlinear Interaction : f(A,B) vs. AB	6.40	6	0.3801
ageg × rr (Factor+Higher Order Factors)	27.37	8	0.0006
Nonlinear	14.85	6	0.0214
Nonlinear Interaction : f(A,B) vs. AB	14.85	6	0.0214
ageg × hrat (Factor+Higher Order Factors)	8.83	6	0.1836
Nonlinear	2.42	4	0.6587
Nonlinear Interaction : f(A,B) vs. AB	2.42	4	0.6587
TOTAL NONLINEAR	78.20	26	< 0.0001
TOTAL INTERACTION	40.48	22	0.0095
TOTAL NONLINEAR + INTERACTION	96.31	32	< 0.0001
TOTAL	1073.78	45	< 0.0001

	Coef	S.E.	Wald Z	Pr($>\|Z\|$)
y≥1	0.0653	7.6563	0.01	0.9932
y≥2	-1.0646	7.6563	-0.14	0.8894
ageg=[7,60)	9.5590	9.9071	0.96	0.3346
ageg=[60,90]	29.1376	15.8915	1.83	0.0667
temp	-0.0694	0.2160	-0.32	0.7480
...				

Wald tests of nonlinearity and interaction are shown in Table 14.5.

```
latex(anova(f1), file='', label='ordinal-anova.f1',
  caption='Wald statistics from the proportional odds model',
  size='smaller')   # Table 14.5
```

The bottom four lines of the table are the most important. First, there is strong evidence that some associations with Y exist (45 d.f. test) and very strong evidence of nonlinearity in one of the vital signs or in `waz` (26 d.f. test). There is moderately strong evidence for an interaction effect somewhere in the model (22 d.f. test). We see that the grouped age variable `ageg` is predictive of Y, but mainly as an effect modifier for `rr`, and `hrat. temp` is extremely nonlinear, and `rr` is moderately so. `hrat`, a difficult variable to measure reliably in young infants, is perhaps not important enough ($\chi^2 = 19, 9$ d.f.) to keep in the final model.

14.6 Residual Plots

Section 13.3.4 defined binary logistic score residuals for isolating the PO assumption in an ordinal model. For the tentative PO model, score residuals for four of the variables were plotted using

```
resid(f1, 'score.binary', pl=TRUE, which=c(17,18,20,21))
## Figure 14.3
```

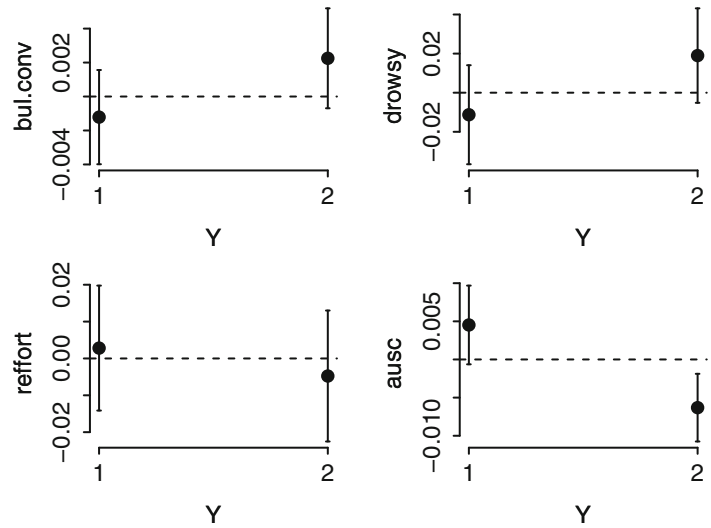

Fig. 14.3 Binary logistic model score residuals for binary events derived from two cutoffs of the ordinal response Y. Note that the mean residuals, marked with closed circles, correspond closely to differences between solid and dashed lines at $Y = 1, 2$ in Figure 14.2. Score residual assessments for spline-expanded variables such as `rr` would have required one plot per d.f.

The result is shown in Figure 14.3. We see strong evidence of non-PO for `ausc` and moderate evidence for `drowsy` and `bul.conv`, in agreement with Figure 14.2.

Partial residuals computed separately for each Y-cutoff (Section 13.3.4) are the most useful residuals for ordinal models as they simultaneously check linearity, find needed transformations, and check PO. In Figure 14.4, smoothed partial residual plots were obtained for all predictors, after first fitting a simple model in which every predictor was assumed to operate linearly. Interactions were temporarily ignored and `age` was used as a continuous variable.

```
f2 ← lrm(Y ∼ age + temp + rr + hrat + waz +
          bul.conv + drowsy + agitated + reffort + ausc +
          feeding + abdominal, data=Sc, x=TRUE, y=TRUE)
resid(f2, 'partial', pl=TRUE, label.curves=FALSE) # Figure 14.4
```

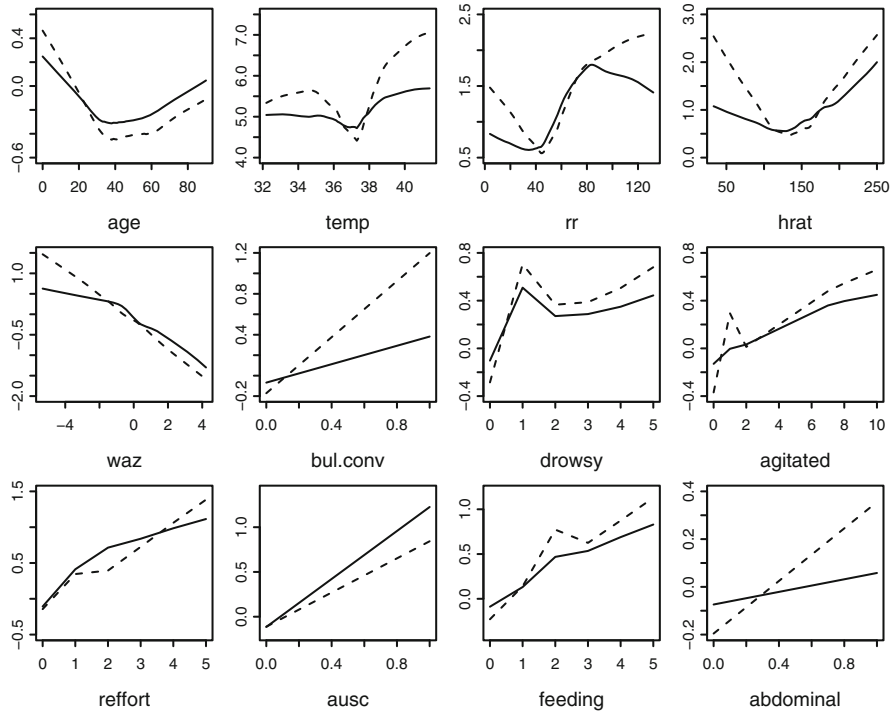

Fig. 14.4 Smoothed partial residuals corresponding to two cutoffs of Y, from a model in which all predictors were assumed to operate linearly and additively. The smoothed curves estimate the actual predictor transformations needed, and parallelism relates to the PO assumption. Solid lines denote $Y \geq 1$ while dashed lines denote $Y \geq 2$.

The degree of non-parallelism generally agreed with the degree of non-flatness in Figure 14.3 and with the other score residual plots that were not shown. The partial residuals show that `temp` is highly nonlinear and that it is much more useful in predicting $Y = 2$. For the cluster scores, the linearity assumption appears reasonable, except possibly for `drowsy`. Other nonlinear effects are taken into account using splines as before (except for `age`, which is categorized).

A model can have significant lack of fit with respect to some of the predictors and still yield quite accurate predictions. To see if that is the case for this PO model, we computed predicted probabilities of $Y = 2$ for all infants from the model and compared these with predictions from a customized binary logistic model derived to predict $\Pr(Y = 2)$. The mean absolute difference in predicted probabilities between the two models is only 0.02, but the 0.90 quantile of that difference is 0.059. For high-risk infants, discrepancies of 0.2 were common. Therefore we elected to consider a different model.

14.7 Graphical Assessment of Fit of CR Model

In order to take a first look at the fit of a CR model, let us consider the two binary events that need to be predicted, and assess linearity and parallelism over Y-cutoffs. Here we fit a sequence of binary fits and then use the `plot.lrm.partial` function, which assembles partial residuals for a sequence of fits and constructs one graph per predictor.

```
cr0 ← lrm(Y==0 ∼ age + temp + rr + hrat + waz +
          bul.conv + drowsy + agitated + reffort + ausc +
          feeding + abdominal, data=Sc, x=TRUE, y=TRUE)
# Use the update function to save repeating model right-
# hand side.  An indicator variable for Y=1 is the
# response variable below
cr1 ← update(cr0, Y==1 ∼ ., subset=Y ≥ 1)
plot.lrm.partial(cr0, cr1, center=TRUE)  # Figure 14.5
```

The output is in Figure 14.5. There is not much more parallelism here than in Figure 14.4. For the two most important predictors, `ausc` and `rr`, there are strongly differing effects for the different events being predicted (e.g., $Y = 0$ or $Y = 1|Y \geq 1$). As is often the case, there is no one constant β model that satisfies assumptions with respect to all predictors simultaneously, especially when there is evidence for non-ordinality for `ausc` in Figure 14.2. The CR model will need to be generalized to adequately fit this dataset.

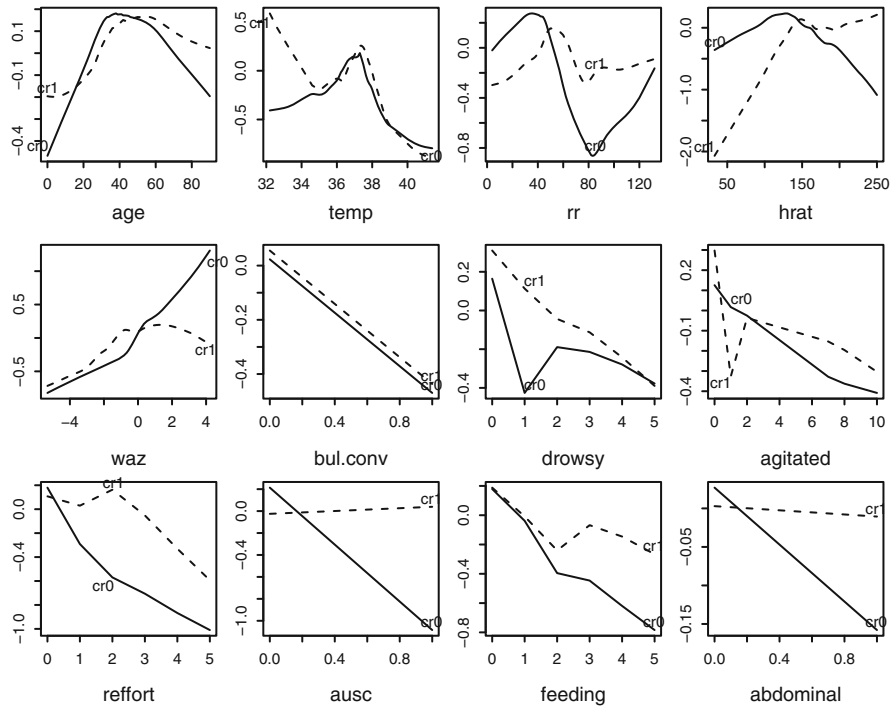

Fig. 14.5 `loess` smoothed partial residual plots for binary models that are components of an ordinal continuation ratio model. Solid lines correspond to a model for $Y = 0$, and dotted lines correspond to a model for $Y = 1|Y \geq 1$.

14.8 Extended Continuation Ratio Model

The CR model in its ordinary form has no advantage over the PO model for this dataset. But Section 13.4.6 discussed how the CR model can easily be extended to relax any of its assumptions. First we use the `cr.setup` function to set up the data for fitting a CR model using the binary logistic trick.

```
u ← cr.setup(Y)
Sc.expanded ← Sc[u$subs, ]
y       ← u$y
cohort  ← u$cohort
```

Here the `cohort` variable has values `'all', 'Y>=1'` corresponding to the conditioning events in Equation 13.10. Once the data frame is expanded to include the different risk cohorts, vectors such as `age` are lengthened (to 5553 records). Now we fit a fully extended CR model that makes no equal slopes assumptions; that is, the model *has* to fit Y assuming the covariables are linear and

additive. At this point, we omit `hrat` but add back all variables that were deleted by examining their association with Y. Recall that most of these seven cluster scores were summarized using PC_1. Adding back "insignificant" variables will allow us to validate the model fairly using the bootstrap, as well as to obtain confidence intervals that are not falsely narrow.[16]

```
full ←
  lrm(y ∼ cohort*(ageg*(rcs(temp,5) + rcs(rr,5)) +
      rcs(waz,4) + bul.conv + drowsy + agitated + reffort +
      ausc + feeding + abdominal + hydration + hxprob +
      pustular + crying + fever.ill + stop.breath + labor),
      data=Sc.expanded, x=TRUE, y=TRUE)
# x=TRUE, y=TRUE are for pentrace, validate, calibrate below
perf ← function(fit) {  # model performance for Y=0
  pr ← predict(fit, type='fitted')[cohort == 'all']
  s ← round(somers2(pr, y[cohort == 'all']), 3)
  pr ← 1 - pr   # Predict Prob[Y > 0] instead of Prob[Y = 0]
  f ← round(c(mean(pr < .05), mean(pr > .25),
              mean(pr > .5)), 2)
  f ← paste(f[1], ', ', f[2], ', and ', f[3], '.', sep='')
  list(somers=s, fractions=f)
}
perf.unpen ← perf(full)
print(full, latex=TRUE, coefs=5)
```

Logistic Regression Model

```
lrm(formula = y ~ cohort * (ageg * (rcs(temp, 5) +
    rcs(rr, 5)) + rcs(waz, 4) + bul.conv + drowsy +
    agitated + reffort + ausc + feeding + abdominal +
    hydration + hxprob + pustular + crying + fever.ill +
    stop.breath + labor), data = Sc.expanded, x = TRUE,
    y = TRUE)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	5553	LR χ^2	1824.33	R^2	0.406	C	0.843
0	1512	d.f.	87	g	1.677	D_{xy}	0.685
1	4041	$\Pr(>\chi^2) < 0.0001$		g_r	5.350	γ	0.687
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	8×10^{-7}			g_p	0.269	τ_a	0.272
				Brier	0.135		

Table 14.6 Wald statistics for `cohort` in the CR model

	χ^2	d.f.	P
cohort (Factor+Higher Order Factors)	199.47	44	< 0.0001
All Interactions	172.12	43	< 0.0001
TOTAL	199.47	44	< 0.0001

	Coef	S.E.	Wald Z	Pr(> $\|Z\|$)
Intercept	1.3966	9.0827	0.15	0.8778
cohort=Y≥1	1.5077	14.6443	0.10	0.9180
ageg=[7,60)	-9.3715	11.4104	-0.82	0.4115
ageg=[60,90]	-26.4502	17.2188	-1.54	0.1245
temp	-0.0049	0.2551	-0.02	0.9846
...				

```
latex(anova(full, cohort), file='',  # Table 14.6
  caption='Wald statistics for \\co{cohort} in the CR model',
  size='smaller[2]', label='ordinal-anova.cohort')
```

```
an ← anova(full, india=FALSE, indnl=FALSE)
```

```
latex(an, file='', label='ordinal-anova.full',
  caption='Wald statistics for the continuation ratio model.
   Interactions with \\co{cohort} assess non-proportional
   hazards', caption.lot='Wald statistics for $Y$ in the
                  continuation ratio model',
      size='smaller[2]')    # Table 14.7
```

This model has LR $\chi^2 = 1824$ with 87 d.f. Wald statistics are in Tables 14.6 and 14.7. The global test of the constant slopes assumption in the CR model (test of all interactions involving `cohort`) has Wald $\chi^2 = 172$ with 43 d.f., $P < 0.0001$. Consistent with Figure 14.5, the formal tests indicate that `ausc` is the biggest violator, followed by `waz` and `rr`.

14.9 Penalized Estimation

We know that the CR model must be extended to fit these data adequately. If the model is fully extended to allow for all `cohort` × predictor interactions, we have not gained any precision or power in using an ordinal model over using a polytomous logistic model. Therefore we seek some restrictions on the model's parameters. The `lrm` and `pentrace` functions allow for differing λ for shrinking different types of terms in the model. Here we do a grid search to determine the optimum penalty for simple main effect (non-interaction) terms and the penalty for interaction terms, most of which are terms interacting with `cohort`

Table 14.7 Wald statistics for the continuation ratio model. Interactions with `cohort` assess non-proportional hazards

	χ^2	d.f.	P
cohort	199.47	44	< 0.0001
ageg	48.89	36	0.0742
temp	59.37	24	0.0001
rr	93.77	24	< 0.0001
waz	39.69	6	< 0.0001
bul.conv	10.80	2	0.0045
drowsy	15.19	2	0.0005
agitated	13.55	2	0.0011
reffort	51.85	2	< 0.0001
ausc	109.80	2	< 0.0001
feeding	27.47	2	< 0.0001
abdominal	1.78	2	0.4106
hydration	4.47	2	0.1069
hxprob	6.62	2	0.0364
pustular	3.03	2	0.2194
crying	1.55	2	0.4604
fever.ill	3.63	2	0.1630
stop.breath	5.34	2	0.0693
labor	5.35	2	0.0690
ageg × temp	8.18	16	0.9432
ageg × rr	38.11	16	0.0015
cohort × ageg	14.88	18	0.6701
cohort × temp	8.77	12	0.7225
cohort × rr	19.67	12	0.0736
cohort × waz	9.04	3	0.0288
cohort × bul.conv	0.33	1	0.5658
cohort × drowsy	0.57	1	0.4489
cohort × agitated	0.55	1	0.4593
cohort × reffort	2.29	1	0.1298
cohort × ausc	38.11	1	< 0.0001
cohort × feeding	2.48	1	0.1152
cohort × abdominal	0.09	1	0.7696
cohort × hydration	0.53	1	0.4682
cohort × hxprob	2.54	1	0.1109
cohort × pustular	2.40	1	0.1210
cohort × crying	0.39	1	0.5310
cohort × fever.ill	3.17	1	0.0749
cohort × stop.breath	2.99	1	0.0839
cohort × labor	0.05	1	0.8309
cohort × ageg × temp	2.22	8	0.9736
cohort × ageg × rr	10.22	8	0.2500
TOTAL NONLINEAR	93.36	40	< 0.0001
TOTAL INTERACTION	203.10	59	< 0.0001
TOTAL NONLINEAR + INTERACTION	257.70	67	< 0.0001
TOTAL	1211.73	87	< 0.0001

to allow for unequal slopes. The following code uses `pentrace` on the full extended CR model fit to find the optimum penalty factors. All combinations of the `simple` and `interaction` λs for which the interaction penalty $\geq$ the penalty for the simple parameters are examined.

```
d ← options(digits=4)
pentrace(full,
         list(simple=c(0,.025,.05,.075,.1),
              interaction=c(0,10,50,100,125,150)))
```

```
Best penalty:

 simple interaction    df
   0.05         125 49.75

 simple interaction    df  aic  bic aic.c
  0.000           0 87.00 1650 1074  1648
  0.000          10 60.63 1671 1269  1669
  0.025          10 60.11 1672 1274  1670
  0.050          10 59.80 1672 1276  1670
  0.075          10 59.58 1671 1277  1670
  0.100          10 59.42 1671 1278  1670
  0.000          50 54.64 1671 1309  1670
  0.025          50 54.14 1672 1313  1671
  0.050          50 53.83 1672 1316  1671
  0.075          50 53.62 1672 1317  1671
  0.100          50 53.46 1672 1318  1671
  0.000         100 51.61 1672 1330  1671
  0.025         100 51.11 1673 1334  1672
  0.050         100 50.81 1673 1336  1672
  0.075         100 50.60 1672 1337  1671
  0.100         100 50.44 1672 1338  1671
  0.000         125 50.55 1672 1337  1671
  0.025         125 50.05 1673 1341  1672
  0.050         125 49.75 1673 1343  1672
  0.075         125 49.54 1672 1344  1672
  0.100         125 49.39 1672 1345  1671
  0.000         150 49.65 1672 1343  1671
  0.025         150 49.15 1672 1347  1672
  0.050         150 48.85 1673 1349  1672
  0.075         150 48.64 1672 1350  1671
  0.100         150 48.49 1672 1351  1671
```

```
options(d)
```

We see that shrinkage from 87 d.f. down to 49.75 effective d.f. results in an improvement in χ^2–scaled AIC of 23. The optimum penalty factors were 0.05 for simple terms and 125 for interaction terms.

Let us now store a penalized version of the full fit, find where the effective d.f. were reduced, and compute χ^2 for each factor in the model. We take the effective d.f. for a collection of model parameters to be the sum of the

diagonals of the matrix product defined underneath Gray's Equation 2.9[237] that correspond to those parameters.

```
full.pen ←
  update(full,
         penalty=list(simple=.05, interaction=125))
print(full.pen, latex=TRUE, coefs=FALSE)
```

Logistic Regression Model

```
lrm(formula = y ~ cohort * (ageg * (rcs(temp, 5) + rcs(rr, 5)) +
    rcs(waz, 4) + bul.conv + drowsy + agitated + reffort + ausc +
    feeding + abdominal + hydration + hxprob + pustular + crying +
    fever.ill + stop.breath + labor), data = Sc.expanded, x = TRUE,
    y = TRUE, penalty = list(simple = 0.05, interaction = 125))
```

Penalty factors

```
  simple nonlinear interaction nonlinear.interaction
    0.05      0.05         125                    125
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	5553	LR χ^2	1772.11	R^2	0.392	C	0.840
0	1512	d.f.	49.75	g	1.594	D_{xy}	0.679
1	4041	$\Pr(>\chi^2)$	< 0.0001	g_r	4.924	γ	0.681
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	1×10^{-7}	Penalty	21.48	g_p	0.263	τ_a	0.269
				Brier	0.136		

```
effective.df(full.pen)
```

```
Original and Effective Degrees of Freedom

                          Original Penalized
All                             87     49.75
Simple Terms                    20     19.98
Interaction or Nonlinear        67     29.77
Nonlinear                       40     16.82
Interaction                     59     22.57
Nonlinear Interaction           32      9.62
```

```
## Compute discrimination for Y=0 vs. Y>0
perf.pen ← perf(full.pen)   # Figure 14.6
# Exclude interactions and cohort effects from plot
plot(anova(full.pen), cex.labels=0.75, rm.ia=TRUE,
     rm.other='cohort  (Factor+Higher Order Factors)')
```

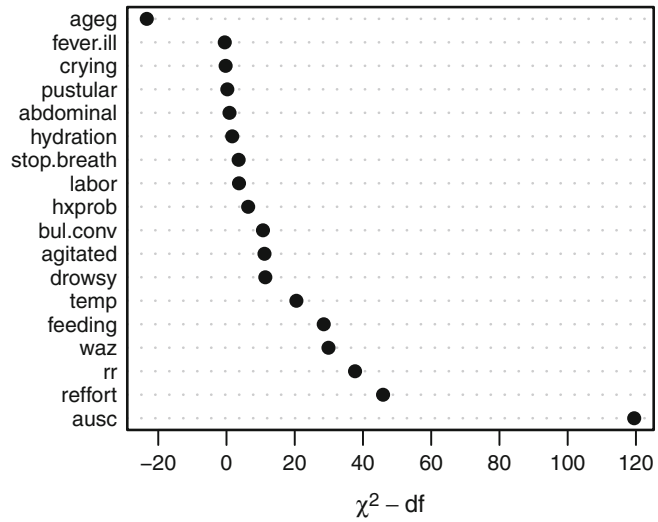

Fig. 14.6 Importance of predictors in full penalized model, as judged by partial Wald χ^2 minus the predictor d.f. The Wald χ^2 values for each line in the dot plot include contributions from all higher-order effects. Interaction effects by themselves have been removed as has the `cohort` effect.

This will be the final model except for the model used in Section 14.10. The model has LR $\chi^2 = 1772$. The output of `effective.df` shows that non-interaction terms have barely been penalized, and coefficients of interaction terms have been shrunken from 59 d.f. to effectively 22.6 d.f. Predictive discrimination was assessed by computing the Somers' D_{xy} rank correlation between $X\hat{\beta}$ and whether $Y = 0$, in the subset of records for which $Y = 0$ is what was being predicted. Here $D_{xy} = 0.672$, and the ROC area is 0.838 (the unpenalized model had an apparent $D_{xy} = 0.676$). To summarize in another way the effectiveness of this model in screening infants for risks of any abnormality, the fraction of infants with predicted probabilities that $Y > 0$ being $< 0.05, > 0.25$, and > 0.5 are, respectively, 0.1, 0.28, and 0.14. `anova` output is plotted in Figure 14.6 to give a snapshot of the importance of the various predictors. The Wald statistics used here are computed on a variance–covariance matrix which is adjusted for penalization (using Gray Equation 2.6[237] before it was determined that the sandwich covariance estimator performs less well than the inverse of the penalized information matrix—see p. 211).

The full equation for the fitted model is below. Only the part of the equation used for predicting $\Pr(Y = 0)$ is shown, other than an intercept for $Y \geq 1$ that does not apply when $Y = 0$.

```
latex(full.pen, which=1:21, file='')
```

$$
\begin{aligned}
X\hat{\beta} = & \\
& -1.337435[\text{Y} >= 1] \\
& +0.1074525[\text{ageg} \in [\ 7, 60)] + 0.1971287[\text{ageg} \in [60, 90]] \\
& +0.1978706\text{temp} + 0.1091831(\text{temp} - 36.19998)_+^3 - 2.833442(\text{temp} - 37)_+^3 \\
& +5.07114(\text{temp} - 37.29999)_+^3 - 2.507527(\text{temp} - 37.69998)_+^3 \\
& +0.1606456(\text{temp} - 39)_+^3 \\
& +0.02090741\text{rr} - 6.336873 \times 10^{-5}(\text{rr} - 32)_+^3 + 8.405441 \times 10^{-5}(\text{rr} - 42)_+^3 \\
& +6.152416 \times 10^{-5}(\text{rr} - 49)_+^3 - 0.0001018105(\text{rr} - 59)_+^3 + 1.960063 \times 10^{-5}(\text{rr} - 76)_+^3 \\
& -0.07589699\text{waz} + 0.02508918(\text{waz} + 2.9)_+^3 - 0.1185068(\text{waz} + 0.75)_+^3 \\
& +0.1225752(\text{waz} - 0.28)_+^3 - 0.02915754(\text{waz} - 1.73)_+^3 - 0.4418073\ \text{bul.conv} \\
& -0.08185088\ \text{drowsy} - 0.05327209\ \text{agitated} - 0.2304409\ \text{reffort} \\
& -1.158604\ \text{ausc} - 0.1599588\ \text{feeding} - 0.1608684\ \text{abdominal} \\
& -0.05409718\ \text{hydration} + 0.08086387\ \text{hxprob} + 0.007519746\ \text{pustular} \\
& +0.04712091\ \text{crying} + 0.004298725\ \text{fever.ill} - 0.3519033\ \text{stop.breath} \\
& +0.06863879\ \text{labor} \\
& +[\text{ageg} \in [\ 7, 60)][6.499592 \times 10^{-5}\ \text{temp} - 0.00279976(\text{temp} - 36.19998)_+^3 \\
& -0.008691166(\text{temp} - 37)_+^3 - 0.004987871(\text{temp} - 37.29999)_+^3 \\
& +0.0259236(\text{temp} - 37.69998)_+^3 - 0.009444801(\text{temp} - 39)_+^3] \\
& +[\text{ageg} \in [60, 90]][0.0001320368\text{temp} - 0.00182639(\text{temp} - 36.19998)_+^3 \\
& -0.01640406(\text{temp} - 37)_+^3 - 0.0476041(\text{temp} - 37.29999)_+^3 \\
& +0.09142148(\text{temp} - 37.69998)_+^3 - 0.02558693(\text{temp} - 39)_+^3] \\
& +[\text{ageg} \in [\ 7, 60)][-0.0009437598\text{rr} - 1.044673 \times 10^{-6}(\text{rr} - 32)_+^3 \\
& -1.670499 \times 10^{-6}(\text{rr} - 42)_+^3 - 5.189082 \times 10^{-6}(\text{rr} - 49)_+^3 + 1.428634 \times 10^{-5}(\text{rr} - 59)_+^3 \\
& -6.382087 \times 10^{-6}(\text{rr} - 76)_+^3] \\
& +[\text{ageg} \in [60, 90]][-0.001920811\text{rr} - 5.52134 \times 10^{-6}(\text{rr} - 32)_+^3 \\
& -8.628392 \times 10^{-6}(\text{rr} - 42)_+^3 - 4.147347 \times 10^{-6}(\text{rr} - 49)_+^3 + 3.813427 \times 10^{-5}(\text{rr} - 59)_+^3 \\
& -1.98372 \times 10^{-5}(\text{rr} - 76)_+^3]
\end{aligned}
$$

where $[c] = 1$ if subject is in group c, 0 otherwise; $(x)_+ = x$ if $x > 0$, 0 otherwise.

Now consider displays of the shapes of effects of the predictors. For the continuous variables `temp` and `rr` that interact with age group, we show the effects for all three age groups separately for each Y cutoff. All effects have been centered so that the log odds at the median predictor value is zero when `cohort='all'`, so these plots actually show log odds relative to reference values. The patterns in Figures 14.9 and 14.8 are in agreement with those in Figure 14.5.

```
yl ← c(-3, 1)     # put all plots on common y-axis scale

# Plot predictors that interact with another predictor
# Vary ageg over all age groups, then vary temp over its
# default range (10th smallest to 10th largest values in
# data).  Make a separate plot for each 'cohort'
# ref.zero centers effects using median x

dd ← datadist(Sc.expanded); dd ← datadist(dd, cohort)
options(datadist='dd')

p1 ← Predict(full.pen, temp, ageg, cohort,
             ref.zero=TRUE, conf.int=FALSE)
p2 ← Predict(full.pen, rr,   ageg, cohort,
              ref.zero=TRUE, conf.int=FALSE)
p ← rbind(temp=p1, rr=p2)   # Figure 14.7:
source(paste('http://biostat.mc.vanderbilt.edu/wiki/pub/Main',
             'RConfiguration/graphicsSet.r', sep='/'))
ggplot(p, ~ cohort, groups='ageg', varypred=TRUE,
       ylim=yl, layout=c(2, 1), legend.position=c(.85,.8),
       addlayer=ltheme(width=3, height=3, text=2.5, title=2.5),
       adj.subtitle=FALSE)   # ltheme defined with source()
```

```
# For each predictor that only interacts with cohort, show
# the differing effects of the predictor for predicting
# Pr(Y=0) and Pr(Y=1 given Y exceeds 0) on the same graph

dd$limits['Adjust to','cohort'] ← 'Y≥1'
v ← Cs(waz, bul.conv, drowsy, agitated, reffort, ausc,
       feeding, abdominal, hydration, hxprob, pustular,
       crying)
yeq1 ← Predict(full.pen, name=v, ref.zero=TRUE)
yl ← c(-1.5, 1.5)
ggplot(yeq1, ylim=yl, sepdiscrete='vertical')   # Figure 14.8
```

```
dd$limits['Adjust to','cohort'] ← 'all'  # original default
all ← Predict(full.pen, name=v, ref.zero=TRUE)
ggplot(all, ylim=yl, sepdiscrete='vertical')  # Figure 14.9
```

1

14.10 Using Approximations to Simplify the Model

Parsimonious models can be developed by approximating predictions from the model to any desired level of accuracy. Let $\hat{L} = X\hat{\beta}$ denote the predicted log odds from the full penalized ordinal model, including multiple records for subjects with $Y > 0$. Then we can use a variety of techniques to approximate $\hat{L}$ from a subset of the predictors (in their raw form). With this approach one can immediately see what is lost over the full model by computing, for

example, the mean absolute error in predicting $\hat{L}$. Another advantage to full model approximation is that shrinkage used in computing $\hat{L}$ is inherited by any model that predicts $\hat{L}$. In contrast, the usual stepwise methods result in $\hat{\beta}$ that are too large since the final coefficients are estimated as if the model structure were prespecified. [2]

CART would be particularly useful as a model approximator as it would result in a prediction tree that would be easy for health workers to use.

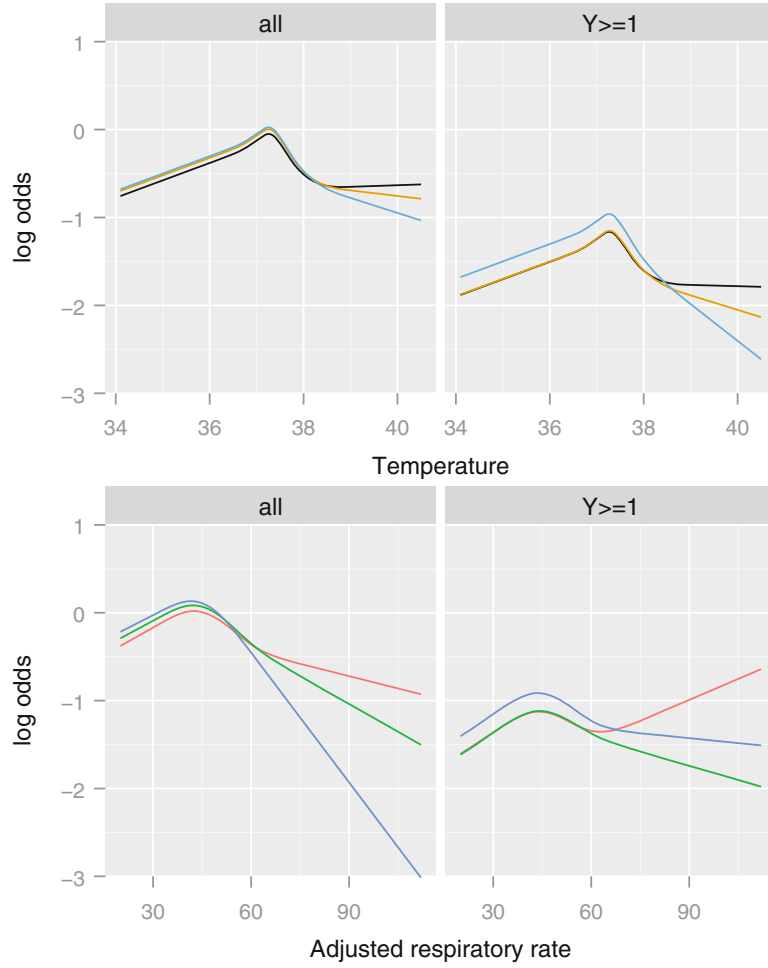

Fig. 14.7 Centered effects of predictors on the log odds, showing the effects of two predictors with interaction effects for the `age` intervals noted. The title `all` refers to the prediction of $Y = 0|Y \geq 0$, that is, $Y = 0$. `Y>=1` refers to predicting the probability of $Y = 1|Y \geq 1$.

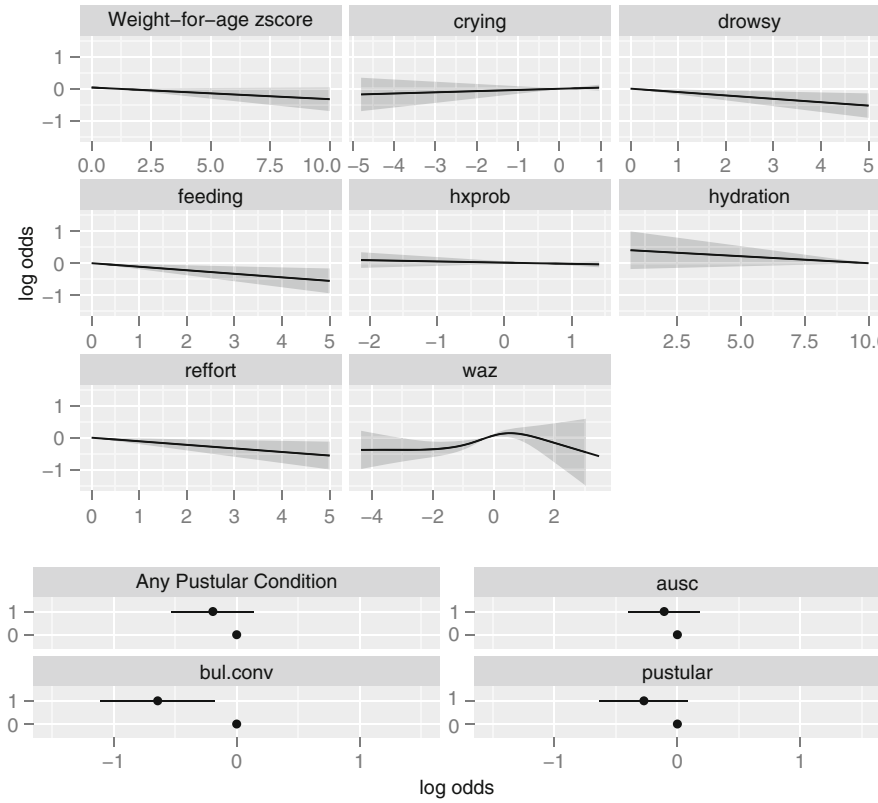

Fig. 14.8 Centered effects of predictors on the log odds, for predicting $Y = 1|Y \geq 1$

Unfortunately, a 50-node CART was required to predict $\hat{L}$ with an $R^2 \geq 0.9$, and the mean absolute error in the predicted logit was still 0.4. This will happen when the model contains many important continuous variables.

Let's approximate the full model using its important components, by using a step-down technique predicting $\hat{L}$ from all of the component variables using ordinary least squares. In using step-down with the least squares function `ols` in `rms` there is a problem when the initial $R^2 = 1.0$ as in that case the estimate of $\sigma = 0$. This can be circumvented by specifying an arbitrary nonzero value of σ to `ols` (here 1.0), as we are not using the variance–covariance matrix from `ols` anyway. Since `cohort` interacts with the predictors, separate approximations can be developed for each level of Y. For this example we approximate the log odds that $Y = 0$ using the cohort of patients used for determining $Y = 0$, that is, $Y \geq 0$ or `cohort='all'`.

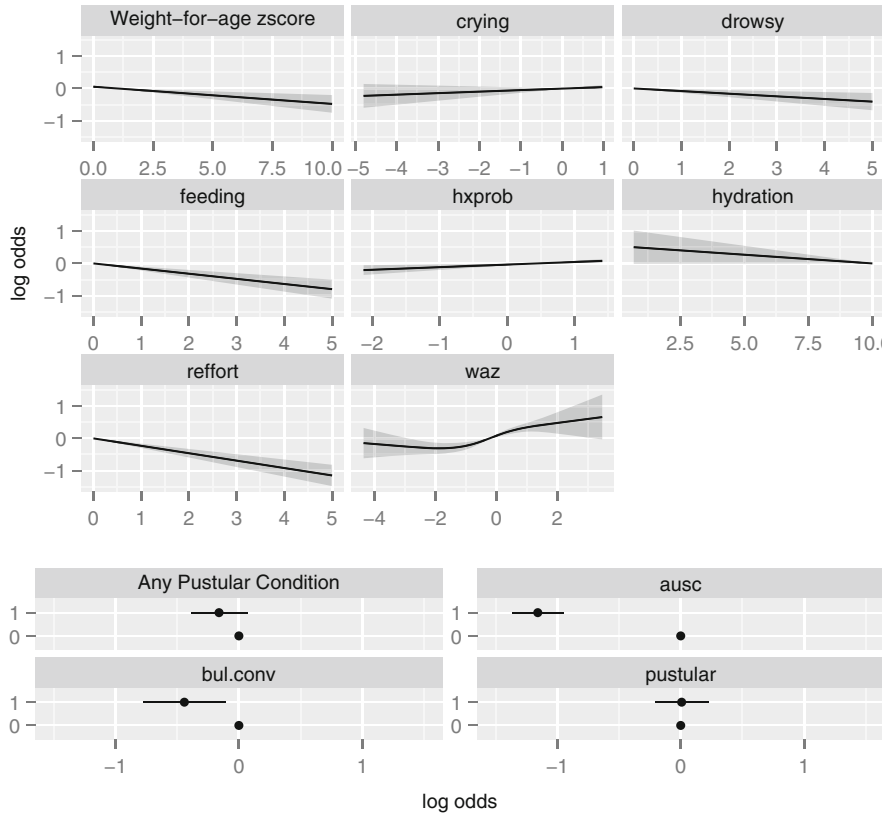

Fig. 14.9 Centered effects of predictors on the log odds, for predicting $Y \geq 1$. No plot was made for the `fever.ill, stop.breath`. or `labor` cluster scores.

```
plogit  ← predict(full.pen)
f ← ols(plogit ∼ ageg*(rcs(temp,5) + rcs(rr,5)) +
          rcs(waz,4) + bul.conv + drowsy + agitated +
          reffort + ausc + feeding + abdominal + hydration +
          hxprob + pustular + crying + fever.ill +
          stop.breath + labor,
          subset=cohort=='all', data=Sc.expanded, sigma=1)

# Do fast backward stepdown
w ← options(width=120)
fastbw(f, aics=1e10)
```

```
 Deleted      Chi-Sq  d.f. P       Residual d.f. P       AIC      R2
 ageg * temp    1.87  8    0.9848    1.87    8    0.9848  -14.13 1.000
 ageg           0.05  2    0.9740    1.92    10   0.9969  -18.08 1.000
 pustular       0.02  1    0.8778    1.94    11   0.9987  -20.06 1.000
 fever.ill      0.08  1    0.7828    2.02    12   0.9994  -21.98 1.000
 crying         9.47  1    0.0021   11.49    13   0.5698  -14.51 0.999
 abdominal     12.66  1    0.0004   24.15    14   0.0440   -3.85 0.997
 rr            17.90  4    0.0013   42.05    18   0.0011    6.05 0.995
 hydration     13.21  1    0.0003   55.26    19   0.0000   17.26 0.993
 labor         23.48  1    0.0000   78.74    20   0.0000   38.74 0.990
 stop.breath   33.40  1    0.0000  112.14    21   0.0000   70.14 0.986
 bul.conv      51.53  1    0.0000  163.67    22   0.0000  119.67 0.980
 agitated      63.66  1    0.0000  227.33    23   0.0000  181.33 0.972
 hxprob        84.16  1    0.0000  311.49    24   0.0000  263.49 0.962
 drowsy       109.86  1    0.0000  421.35    25   0.0000  371.35 0.948
 temp         295.67  4    0.0000  717.01    29   0.0000  659.01 0.911
 waz          368.86  3    0.0000 1085.87    32   0.0000 1021.87 0.866
 reffort      449.83  1    0.0000 1535.70    33   0.0000 1469.70 0.810
 ageg * rr    751.19  8    0.0000 2286.90    41   0.0000 2204.90 0.717
 ausc        1906.82  1    0.0000 4193.72    42   0.0000 4109.72 0.482
 feeding     3900.33  1    0.0000 8094.04    43   0.0000 8008.04 0.000

Approximate Estimates after Deleting Factors

      Coef    S.E. Wald Z P
[1,] 1.617 0.01482   109.1 0

Factors in Final Model

None
```

```
options(w)
# 1e10 causes all variables to eventually be
# deleted so can see most important ones in order

# Fit an approximation to the full penalized model using
# most important variables
full.approx ←
  ols(plogit ∼ rcs(temp,5) + ageg*rcs(rr,5) +
      rcs(waz,4) + bul.conv + drowsy + reffort +
      ausc + feeding,
      subset=cohort=='all', data=Sc.expanded)
p ← predict(full.approx)
abserr ← mean(abs(p - plogit[cohort == 'all']))
Dxy ← somers2(p, y[cohort == 'all'])['Dxy']
```

The approximate model had R^2 against the full penalized model of 0.972, and the mean absolute error in predicting $\hat{L}$ was 0.17. The D_{xy} rank correlation between the approximate model's predicted logit and the binary event $Y = 0$

is 0.665 as compared with the full model's $D_{xy} = 0.672$. See Section 19.5 for an example of computing correct estimates of variance of the parameters in an approximate model.

Next turn to diagramming this model approximation so that all predicted values can be computed without the use of a computer. We draw a type of nomogram that converts each effect in the model to a 0 to 100 scale which is just proportional to the log odds. These points are added across predictors to derive the "Total Points," which are converted to $\hat{L}$ and then to predicted probabilities. For the interaction between `rr` and `ageg`, `rms`'s `nomogram` function automatically constructs three `rr` axes—only one is added into the total point score for a given subject. Here we draw a nomogram for predicting the probability that $Y > 0$, which is $1 - \Pr(Y = 0)$. This probability is derived by negating $\hat{\beta}$ and $X\hat{\beta}$ in the model derived to predict $\Pr(Y = 0)$.

```
f ← full.approx
f$coefficients      ← -f$coefficients
f$linear.predictors ← -f$linear.predictors

n ← nomogram(f,
             temp=32:41, rr=seq(20,120,by=10),
             waz=seq(-1.5,2,by=.5),
             fun=plogis, funlabel='Pr(Y>0)',
             fun.at=c(.02,.05,seq(.1,.9,by=.1),.95,.98))
# Print n to see point tables
plot(n, lmgp=.2, cex.axis=.6)  # Figure 14.10
newsubject ←
  data.frame(ageg='[ 0, 7)', rr=30, temp=39, waz=0, drowsy=5,
             reffort=2, bul.conv=0, ausc=0, feeding=0)
xb ← predict(f, newsubject)
```

The nomogram is shown in Figure 14.10. As an example in using the nomogram, a six-day-old infant gets approximately 9 points for having a respiration rate of 30/minute, 19 points for having a temperature of 39°C, 11 points for `waz=0`, 14 points for `drowsy=5`, and 15 points for `reffort=2`. Assuming that `bul.conv=ausc=feeding=0`, that infant gets 68 total points. This corresponds to $X\hat{\beta} = -0.68$ and a probability of 0.34. 3

14.11 Validating the Model

For the full CR model that was fitted using penalized maximum likelihood estimation (PMLE), we used 200 bootstrap replications to estimate and then to correct for optimism in various statistical indexes: D_{xy}, generalized R^2, intercept and slope of a linear re-calibration equation for $X\hat{\beta}$, the maximum calibration error for $\Pr(Y = 0)$ based on the linear-logistic re-calibration (`Emax`), and the Brier quadratic probability score `B`. PMLE is used at each of the 200 resamples. During the bootstrap simulations, we sample with

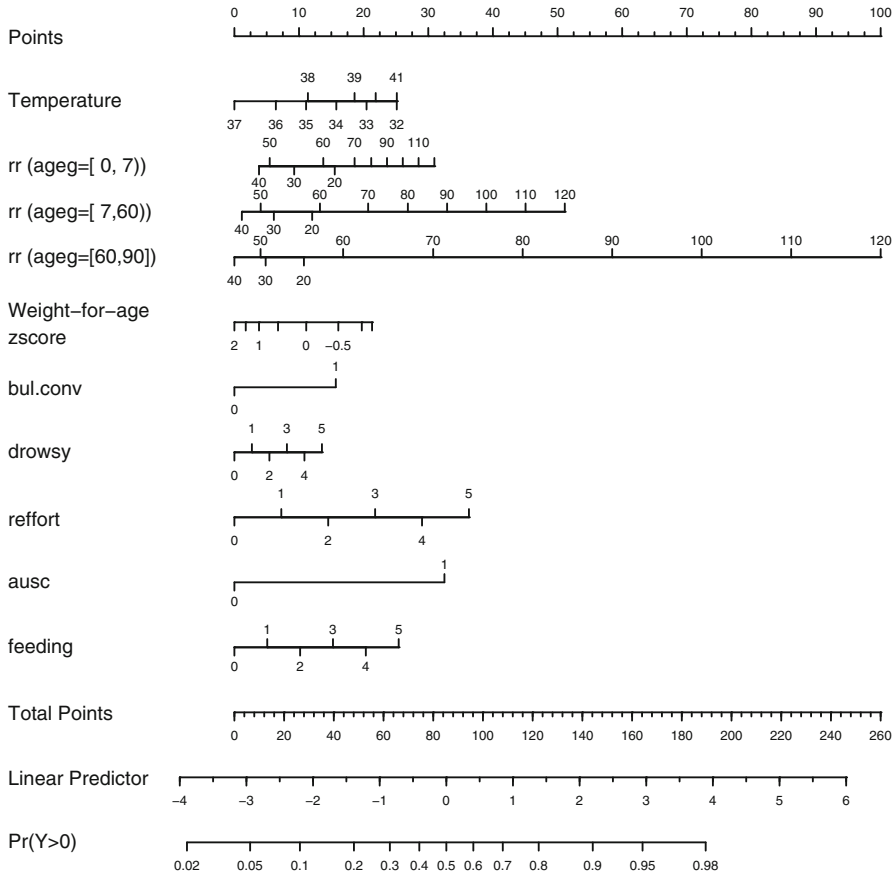

Fig. 14.10 Nomogram for predicting $\Pr(Y > 0)$ from the penalized extended CR model, using an approximate model fitted using ordinary least squares ($R^2 = 0.972$ against the full model's predicted logits).

replacement from the *patients* and not from the 5553 expanded *records*, hence the specification `cluster=u$subs`, where `u$subs` is the vector of sequential patient numbers computed from `cr.setup` above. To be able to assess predictive accuracy of a single predicted probability, the `subset` parameter is specified so that $\Pr(Y = 0)$ is being assessed even though 5553 observations are used to develop each of the 200 models.

```
set.seed(1)    # so can reproduce results
v ← validate(full.pen, B=200, cluster=u$subs,
             subset=cohort=='all')
latex(v, file='', digits=2, size='smaller')
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.67	0.68	0.67	0.01	0.66	200
R^2	0.38	0.38	0.37	0.01	0.36	200
Intercept	−0.03	−0.03	0.00	−0.03	0.00	200
Slope	1.03	1.03	1.00	0.03	1.00	200
$E_{\max}$	0.00	0.00	0.00	0.00	0.00	200
D	0.28	0.29	0.28	0.01	0.27	200
U	0.00	0.00	0.00	0.00	0.00	200
Q	0.28	0.29	0.28	0.01	0.27	200
B	0.12	0.12	0.12	0.00	0.12	200
g	1.47	1.50	1.45	0.04	1.42	200
g_p	0.22	0.23	0.22	0.00	0.22	200

```
v ← round(v, 3)
```

We see that for the apparent $D_{xy} = 0.672$ and that the optimism from overfitting was estimated to be 0.011 for the PMLE model, so the bias-corrected estimate of predictive discrimination is 0.661. The intercept and slope needed to re-calibrate $X\hat{\beta}$ to a 45° line are very near (0, 1). The estimate of the maximum calibration error in predicting $\Pr(Y = 0)$ is 0.001, which is quite satisfactory. The corrected Brier score is 0.122.

The simple calibration statistics just listed do not address the issue of whether predicted values from the model are miscalibrated in a nonlinear way, so now we estimate an overfitting-corrected calibration curve nonparametrically.

```
cal ← calibrate(full.pen, B=200, cluster=u$subs,
                subset=cohort=='all')
err ← plot(cal)  # Figure 14.11
```

```
n=5553   Mean absolute error=0.017   Mean squared error=0.00043
0.9 Quantile of absolute error=0.038
```

The results are shown in Figure 14.11. One can see a slightly nonlinear calibration function estimate, but the overfitting-corrected calibration is excellent everywhere, being only slightly worse than the apparent calibration. The estimated maximum calibration error is 0.044. The excellent validation for both predictive discrimination and calibration are a result of the large sample size, frequency distribution of Y, initial data reduction, and PMLE.

14.12 Summary

Clinically guided variable clustering and item weighting resulted in a great reduction in the number of candidate predictor degrees of freedom and hence increased the true predictive accuracy of the model. Scores summarizing clusters of clinical signs, along with temperature, respiration rate, and weight-for-age after suitable nonlinear transformation and allowance for interactions

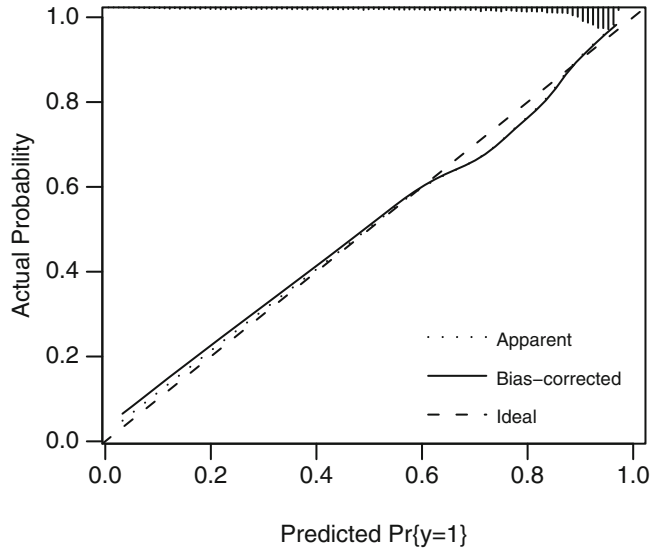

Fig. 14.11 Bootstrap calibration curve for the full penalized extended CR model. 200 bootstrap repetitions were used in conjunction with the `loess` smoother.[111] Also shown is a "rug plot" to demonstrate how effective this model is in discriminating patients into low- and high-risk groups for $\Pr(Y = 0)$ (which corresponds with the derived variable value $y = 1$ when `cohort='all'`).

with age, are powerful predictors of the ordinal response. Graphical methods are effective for detecting lack of fit in the PO and CR models and for diagramming the final model. Model approximation allowed development of parsimonious clinical prediction tools. Approximate models inherit the shrinkage from the full model. For the ordinal model developed here, substantial shrinkage of the full model was needed.

14.13 Further Reading

1. See Moons et al.[462] for another case study in penalized maximum likelihood estimation.
2. The *lasso* method of Tibshirani[608,609] also incorporates shrinkage into variable selection.
3. To see how this compares with predictions using the full model, the extra clinical signs in that model that are not in the approximate model were predicted individually on the basis of $X\hat{\beta}$ from the reduced model along with the signs that are in that model, using ordinary linear regression. The signs not specified when evaluating the approximate model were then set to predicted values based on the values given for the 6-day-old infant above. The resulting $X\hat{\beta}$ for the full model is -0.81 and the predicted probability is 0.31, as compared with -0.68 and 0.34 quoted above.

14.14 Problems

Develop a proportional odds ordinal logistic model predicting the severity of functional disability (`sfdm2`) in SUPPORT. The highest level of this variable corresponds to patients dying before the two-month follow-up interviews. Consider this level as the most severe outcome. Consider the following predictors: `age, sex, dzgroup, num.co, scoma, race` (use all levels), `meanbp, hrt, temp, pafi, alb, adlsc`. The last variable is the baseline level of functional disability from the "activities of daily living scale."

1. For the variables `adlsc, sex, age, meanbp`, and others if you like, make plots of means of predictors stratified by levels of the response, to check for ordinality. On the same plot, show estimates of means assuming the proportional odds relationship between predictors and response holds. Comment on the evidence for ordinality and for proportional odds.
2. To allow for maximum adjustment of baseline functional status, treat this predictor as nominal (after rounding it to the nearest whole number; fractional values are the result of imputation) in remaining steps, so that all dummy variables will be generated. Make a single chart showing proportions of various outcomes stratified (individually) by `adlsc, sex, age, meanbp`. For continuous predictors use quartiles. You can pass the following function to the `summary` (`summary.formula`) function to obtain the proportions of patients having `sfdm2` at or worse than each of its possible levels (other than the first level). An easy way to do this is to use the `cumcategory` function with the `Hmisc` package's `summary.formula` function. `cumcategorysummary.formula` Print estimates to only two significant digits of precision. Manually check the calculations for the `sex` variable using `table(sex, sfdm2)`. Then plot all estimates on a single graph using `plot(object, which=1:4)`, where `object` was created by `summary` (actually `summary.formula`). Note: for printing tables you may want to convert `sfdm2` to a 0–4 variable so that column headers are short and so that later calculations are simpler. You can use for example:

```
sfdm ← as.integer(sfdm2) - 1
```

3. Use an `R` function such as the following to compute the *logits* of the cumulative proportions.

```
sf ← function(y)
  c('Y≥1'=qlogis(mean(y ≥ 1)),
    'Y≥2'=qlogis(mean(y ≥ 2)),
    'Y≥3'=qlogis(mean(y ≥ 3)),
    'Y≥4'=qlogis(mean(y ≥ 4)))
```

   As the $Y = 3$ category is rare, it may be even better to omit the $Y \geq 4$ column above, as was done in Section 13.3.9 and Figure 13.1. For each predictor pick two rows of the `summary` table having reasonable sample sizes, and take the difference between the two rows. Comment on the

validity of the proportional odds assumption by assessing how constant the row differences are across columns. Note: constant differences in log odds (logits) mean constant ratios of odds or constant relative effects of the predictor across outcome levels.

4. Make two plots nonparametrically relating `age` to all of the cumulative proportions or their logits. You can use commands such as the following (to use the R `Hmisc` package).

```
for(i in 1:4)
  plsmo(age, sfdm ≥ i, add=i>1,
        ylim=c(.2,.8), ylab='Proportion Y≥j')
for(i in 1:4)
  plsmo(age, sfdm ≥ i, add=i>1, fun=qlogis,
        ylim=qlogis(c(.2,.8)), ylab='logit')
```

   Comment on the linearity of the `age` effect (which of the two plots do you use?) and on the proportional odds assumption for `age`, by assessing parallelism in the second plot.
5. Impute `race` using the most frequent category and `pafi` and `alb` using "normal" values.
6. Fit a model to predict the ordinal response using all predictors. For continuous ones assume a smooth relationship but allow it to be nonlinear. Quantify the ability of the model to discriminate patients in the five outcomes. Do an overall likelihood ratio test for whether any variables are associated with the level of functional disability.
7. Compute partial tests of association for each predictor and a test of nonlinearity for continuous ones. Compute a global test of nonlinearity. Graphically display the ranking of importance of the predictors.
8. Display the shape of how each predictor relates to the log odds of exceeding any level of `sfdm2` you choose, setting other predictors to typical values (one value per predictor). By default, `Predict` will make predictions for the second response category, which is a satisfactory choice here.
9. Use resampling to validate the Somers' D_{xy} rank correlation between predicted logit and the ordinal outcome. Also validate the generalized R^2, and slope shrinkage coefficient, all using a single R command. Comment on the quality (potential "export-ability") of the model.