Chapter 13

Ordinal Logistic Regression

13.1 Background

Many medical and epidemiologic studies incorporate an ordinal response variable. In some cases an ordinal response Y represents levels of a standard measurement scale such as severity of pain (none, mild, moderate, severe). In other cases, ordinal responses are constructed by specifying a hierarchy of separate endpoints. For example, clinicians may specify an ordering of the severity of several component events and assign patients to the worst event present from among none, heart attack, disabling stroke, and death. Still another use of ordinal response methods is the application of rank-based methods to continuous responses so as to obtain robust inferences. For example, the proportional odds model described later allows for a continuous Y and is really a generalization of the Wilcoxon–Mann–Whitney rank test.

There are many variations of logistic models used for predicting an ordinal response variable Y. All of them have the advantage that they do not assume a spacing between levels of Y. In other words, the same regression coefficients and P-values result from an analysis of a response variable having levels $0, 1, 2$ when the levels are recoded $0, 1, 20$. Thus ordinal models use only the rank-ordering of values of Y.

In this chapter we consider two of the most popular ordinal logistic models, the proportional odds (PO) form of an ordinal logistic model[440] and the forward continuation ratio (CR) ordinal logistic model.[135] 1

F. E. Harrell Jr., *Regression Modeling Strategies*

13.2 Ordinality Assumption

A basic assumption of all commonly used ordinal regression models is that the response variable behaves in an ordinal fashion with respect to each predictor. Assuming that a predictor X is linearly related to the log odds of some appropriate event, a simple way to check for ordinality is to plot the mean of X stratified by levels of Y. These means should be in a consistent order. If for many of the Xs, two adjacent categories of Y do not distinguish the means, that is evidence that those levels of Y should be pooled.

One can also estimate the mean or expected value of $X|Y = j$ ($E(X|Y = j)$) given that the ordinal model assumptions hold. This is a useful tool for checking those assumptions, at least in an unadjusted fashion. For simplicity, assume that X is discrete, and let $P_{jx} = \Pr(Y = j|X = x)$ be the probability that $Y = j$ given $X = x$ that is dictated from the model being fitted, with X being the only predictor in the model. Then

$$\begin{aligned} \Pr(X = x|Y = j) &= \Pr(Y = j|X = x)\Pr(X = x)/\Pr(Y = j) \\ E(X|Y = j) &= \sum_x xP_{jx}\Pr(X = x)/\Pr(Y = j), \end{aligned} \tag{13.1}$$

and the expectation can be estimated by

$$\hat{E}(X|Y = j) = \sum_x x\hat{P}_{jx}f_x/g_j, \tag{13.2}$$

where $\hat{P}_{jx}$ denotes the estimate of P_{jx} from the fitted one-predictor model (for inner values of Y in the PO models, these probabilities are differences between terms given by Equation 13.4 below), f_x is the frequency of $X = x$ in the sample of size n, and g_j is the frequency of $Y = j$ in the sample. This estimate can be computed conveniently without grouping the data by X. For n subjects let the n values of X be $x_1, x_2, \ldots, x_n$. Then

$$\hat{E}(X|Y = j) = \sum_{i=1}^{n} x_i\hat{P}_{jx_i}/g_j. \tag{13.3}$$

Note that if one were to compute differences between conditional means of X and the conditional means of X given PO, and if furthermore the means were conditioned on $Y \geq j$ instead of $Y = j$, the result would be proportional to means of score residuals defined later in Equation 13.6.

13.3 Proportional Odds Model

13.3.1 Model

The most commonly used ordinal logistic model was described in Walker and Duncan[440] and later called the *proportional odds (PO) model* by McCullagh.[309] The PO model is best stated as follows, for a response variable having levels $0, 1, 2, \ldots, k$:

$$\Pr[Y \geq j|X] = \frac{1}{1 + \exp[-(\alpha_j + X\beta)]}, \tag{13.4}$$

where $j = 1, 2, \ldots, k$. Some authors write the model in terms of $Y \leq j$. Our formulation makes the model coefficients consistent with the binary logistic model. There are k intercepts (αs). For fixed j, the model is an ordinary logistic model for the event $Y \geq j$. By using a common vector of regression coefficients β connecting probabilities for varying j, the PO model allows for parsimonious modeling of the distribution of Y.

There is a nice connection between the PO model and the Wilcoxon–Mann–Whitney two-sample test: when there is a single predictor X_1 that is binary, the numerator of the score test for testing $H_0 : \beta_1 = 0$ is proportional to the two-sample test statistic [452, pp. 2258-2259].

13.3.2 Assumptions and Interpretation of Parameters

There is an implicit assumption in the PO model that the regression coefficients (β) are independent of j, the cutoff level for Y. One could say that there is no $X \times Y$ interaction if PO holds. For a specific Y-cutoff j, the model has the same assumptions as the binary logistic model (Section 10.1.1). That is, the model in its simplest form assumes the log odds that $Y \geq j$ is linearly related to each X and that there is no interaction between the Xs.

In designing clinical studies, one sometimes hears the statement that an ordinal outcome should be avoided since statistical tests of patterns of those outcomes are hard to interpret. In fact, one interprets effects in the PO model using ordinary odds ratios. The difference is that a single odds ratio is assumed to apply equally to *all* events $Y \geq j, j = 1, 2, \ldots, k$. If linearity and additivity hold, the $X_m + 1 : X_m$ odds ratio for $Y \geq j$ is $\exp(\beta_m)$, whatever the cutoff j.

Sometimes it helps in interpreting the model to estimate the mean Y as a function of one or more predictors, even though this assumes a spacing for the Y-levels. 2

13.3.3 Estimation

The PO model is fitted using MLE on a somewhat complex likelihood function that is dependent on differences in logistic model probabilities. The estimation process forces the αs to be in descending order.

13.3.4 Residuals

Schoenfeld residuals[377] are very effective[157] in checking the proportional hazards assumption in the Cox[92] survival model. For the PO model one could analogously compute each subject's contribution to the first derivative of the log likelihood function with respect to β_m, average them separately by levels of Y, and examine trends in the residual plots as in Section 19.5.2. A few examples have shown that such plots are usually hard to interpret. Easily interpreted score residual plots for the PO model can be constructed, however, by using the fitted PO model to predict a series of binary events $Y \geq j, j = 1, 2, \ldots, k$, using the corresponding predicted probabilities

$$\hat{P}_{ij} = \frac{1}{1 + \exp[-(\hat{\alpha}_j + X_i\hat{\beta})]}, \tag{13.5}$$

where X_i stands for a vector of predictors for subject i. Then, after forming an indicator variable for the event currently being predicted ($[Y_i \geq j]$), one computes the score (first derivative) components U_{im} from an ordinary binary logistic model:

$$U_{im} = X_{im}([Y_i \geq j] - \hat{P}_{ij}), \tag{13.6}$$

for the subject i and predictor m. Then, for each column of U, plot the mean $\bar{U}_{\cdot m}$ and confidence limits, with Y (i.e., j) on the x-axis. For each predictor the trend against j should be flat if PO holds. [a] In binary logistic regression, *partial residuals* are very useful as they allow the analyst to fit linear effects for all the predictors but then to nonparametrically estimate the true transformation that each predictor requires (Section 10.4). The partial residual is defined as follows, for the ith subject and mth predictor variable.[79, 256]

$$r_{im} = \hat{\beta}_m X_{im} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)}, \tag{13.7}$$

where

$$\hat{P}_i = \frac{1}{1 + \exp[-(\alpha + X_i\hat{\beta})]}. \tag{13.8}$$

A smoothed plot (e.g., using the moving linear regression algorithm in `loess`[76]) of X_{im} against r_{im} provides a nonparametric estimate of how X_m relates to the log

[a] If $\hat{\beta}$ were derived from separate binary fits, all $\bar{U}_{\cdot m} \equiv 0$.

relative odds that $Y = 1|X_m$. For ordinal Y, we just need to compute binary model partial residuals for all cutoffs j:

$$r_{im} = \hat{\beta}_m X_{im} + \frac{[Y_i \geq j] - \hat{P}_{ij}}{\hat{P}_{ij}(1 - \hat{P}_{ij})}, \tag{13.9}$$

then to make a plot for each m showing smoothed partial residual curves for all j, looking for similar shapes and slopes for a given predictor for all j. Each curve provides an estimate of how X_m relates to the relative log odds that $Y \geq j$. Since partial residuals allow examination of predictor transformations (linearity) while simultaneously allowing examination of PO (parallelism), partial residual plots are generally preferred over score residual plots for ordinal models.

13.3.5 Assessment of Model Fit

Peterson and Harrell[335] developed score and likelihood ratio tests for testing the PO assumption. The score test is used in the SAS LOGISTIC procedure,[363] but its extreme anticonservatism in many cases can make it unreliable.[335] 3

For determining whether the PO assumption is likely to be satisfied for each predictor separately, there are several graphics that are useful. One is the graph comparing means of $X|Y$ with and without assuming PO, as described in Section 13.2 (see Figure 14.2 for an example). Another is the simple method of stratifying on each predictor and computing the logits of all proportions of the form $Y \geq j, j = 1, 2, \ldots, k$. When proportional odds holds, the differences in logits between different values of j should be the same at all levels of X, because the model dictates that $\text{logit}(Y \geq j|X) - \text{logit}(Y \geq i|X) = \alpha_j - \alpha_i$, for any constant X. An example of this is in Figure 13.1.

Chapter 14 has many examples of graphics for assessing fit of PO models. Regarding assessment of linearity and additivity assumptions, splines, partial residual plots, and interaction tests are among the best tools.

13.3.6 Quantifying Predictive Ability

The R^2_N coefficient is really computed from the model LR χ^2 (χ^2 added to a model containing only the k intercept parameters) to describe the model's predictive power. The Somers' D_{xy} rank correlation between $X\hat{\beta}$ and Y is an easily interpreted measure of predictive discrimination. Since it is a rank measure, it does not matter which intercept α is used in the calculation. The probability of concordance, c, is also a useful measure. Here one takes all possible pairs of subjects having differing Y values and computes the fraction of such pairs for which the values of $X\hat{\beta}$ are in the same direction as the two Y values. c could be called a generalized ROC area in this setting. As before, $D_{xy} = 2(c - 0.5)$. Note that D_{xy},

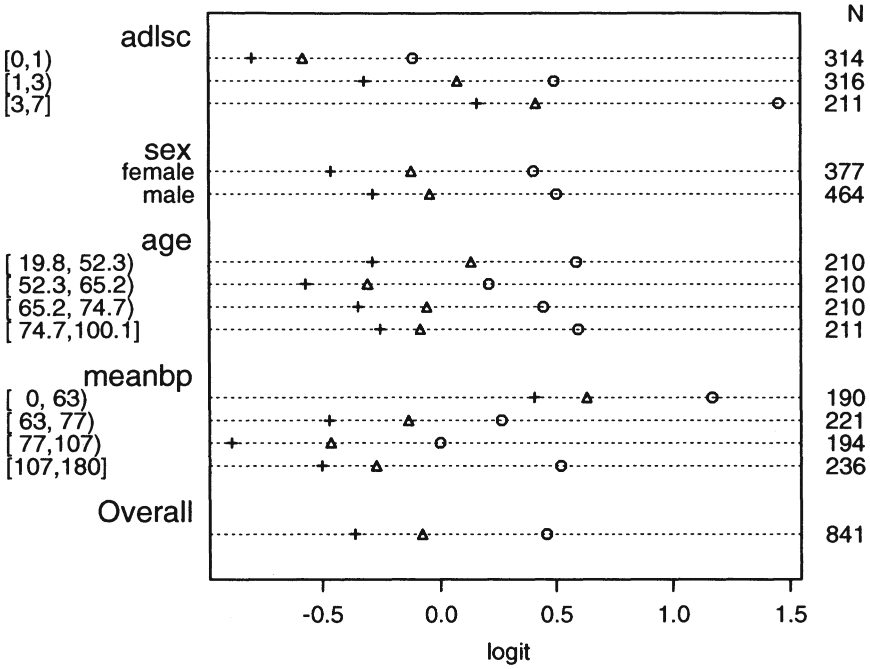

FIGURE 13.1: Checking PO assumption separately for a series of predictors. The circle, triangle, and plus sign correspond to $Y \geq 1, 2, 3$, respectively. PO is checked by examining the vertical constancy of distances between any two of these three symbols. Response variable is the severe functional disability scale `sfdm2` from the 1000-patient SUPPORT dataset, with the last two categories combined because of low frequency of coma/intubation.

c, and the Brier score B can easily be computed for various dichotomizations of Y, to investigate predictive ability in more detail.

13.3.7 Validating the Fitted Model

The PO model is validated much the same way as the binary logistic model (see Section 10.9). For estimating an overfitting-corrected calibration curve (Section 10.11) one estimates $\Pr(Y \geq j|X)$ using one j at a time.

13.3.8 S-Plus Functions

The `Design` library's `lrm` function fits the PO model directly, assuming that the levels of the response variable (e.g., the `levels` of a `factor` variable) are listed in the proper order. If the response is numeric, `lrm` assumes the numeric codes properly order the responses. If it is a character vector and is not a `factor`, `lrm` assumes the correct ordering is alphabetic. Of course `ordered` variables in S-Plus are appropriate response variables for ordinal regression.

The S-Plus functions `popower` and `posamsize` (in the `Hmisc` library) compute power and sample size estimates for ordinal responses using the proportional odds model.

The function `plot.xmean.ordinaly` in `Design` computes and graphs the quantities described in Section 13.2. It plots simple Y-stratified means overlaid with $\hat{E}(X|Y = j)$, with j on the x-axis. The $\hat{E}$s are computed for both PO and continuation ratio ordinal logistic models.

The `Hmisc` library's `summary.formula` function is also useful for assessing the PO assumption. Figure 13.1 was produced using the following code.

```
attach(support)
sfdm ← as.integer(sfdm2) - 1
sf ← function(y)
  c('Y>=1'=qlogis(mean(y >= 1)), 'Y>=2'=qlogis(mean(y >= 2)),
    'Y>=3'=qlogis(mean(y >= 3)))
s ← summary(sfdm ~ adlsc + sex + age + meanbp, fun=sf)
plot(s, which=1:3, pch=1:3, xlab='logit', vnames='names',
     main='', width.factor=1.5)
```

Generic `Design` functions such as `validate`, `calibrate`, and `nomogram` work with PO model fits from `lrm` as long as the analyst specifies which intercept(s) to use.

13.4 Continuation Ratio Model

13.4.1 Model

Unlike the PO model, which is based on *cumulative* probabilities, the continuation ratio (CR) model is based on *conditional* probabilities. The (forward) CR model[24, 36, 135] is stated as follows for $Y = 0, \ldots, k$.

$$\begin{aligned}
\Pr(Y = j|Y \geq j, X) &= \frac{1}{1 + \exp[-(\theta_j + X\gamma)]} \\
\operatorname{logit}(Y = 0|Y \geq 0, X) &= \operatorname{logit}(Y = 0|X) \\
&= \theta_0 + X\gamma \qquad (13.10) \\
\operatorname{logit}(Y = 1|Y \geq 1, X) &= \theta_1 + X\gamma \\
&\ldots \\
\operatorname{logit}(Y = k - 1|Y \geq k - 1, X) &= \theta_{k-1} + X\gamma.
\end{aligned}$$

The CR model has been said to be likely to fit ordinal responses when subjects have to "pass through" one category to get to the next. The CR model is a discrete version of the Cox proportional hazards model. The discrete hazard function is defined as $\Pr(Y = j|Y \geq j)$.

13.4.2 Assumptions and Interpretation of Parameters

The CR model assumes that the vector of regression coefficients, γ, is the same regardless of which conditional probability is being computed.

One could say that there is no $X\times$ condition interaction if the CR model holds. For a specific condition $Y \geq j$, the model has the same assumptions as the binary logistic model (Section 10.1.1). That is, the model in its simplest form assumes that the log odds that $Y = j$ conditional on $Y \geq j$ is linearly related to each X and that there is no interaction between the Xs.

A single odds ratio is assumed to apply equally to *all* conditions $Y \geq j, j = 0, 1, 2, \ldots, k - 1$. If linearity and additivity hold, the $X_m + 1 : X_m$ odds ratio for $Y = j$ is $\exp(\beta_m)$, whatever the conditioning event $Y \geq j$.

To compute $\Pr(Y > 0|X)$ from the CR model, one only needs to take one minus $\Pr(Y = 0|X)$. To compute other unconditional probabilities from the CR model, one must multiply the conditional probabilities. For example, $\Pr(Y > 1|X) = \Pr(Y > 1|X, Y \geq 1) \times \Pr(Y \geq 1|X) = [1 - \Pr(Y = 1|Y \geq 1, X)][1 - \Pr(Y = 0|X)] = [1 - 1/(1 + \exp[-(\theta_1 + X\gamma)])][1 - 1/(1 + \exp[-(\theta_0 + X\gamma)])]$.

13.4.3 Estimation

Armstrong and Sloan[24] and Berridge and Whitehead[36] showed how the CR model can be fitted using an ordinary binary logistic model likelihood function, after certain rows of the X matrix are duplicated and a new binary Y vector is constructed. For each subject, one constructs separate records by considering successive conditions $Y \geq 0, Y \geq 1, \ldots, Y \geq k-1$ for a response variable with values $0, 1, \ldots, k$. The binary response for each applicable condition or "cohort" is set to 1 if the subject failed at the current "cohort" or "risk set," that is, if $Y = j$ where the cohort being considered is $Y \geq j$. The constructed cohort variable is carried along with the new X and Y. This variable is considered to be categorical and its coefficients are fitted by adding $k-1$ dummy variables to the binary logistic model. For ease of computation, the CR model is restated as follows, with the first cohort used as the reference cell.

$$\Pr(Y = j | Y \geq j, X) = \frac{1}{1 + \exp[-(\alpha + \theta_j + X\gamma)]}. \tag{13.11}$$

Here α is an overall intercept, $\theta_0 \equiv 0$, and $\theta_1, \ldots, \theta_{k-1}$ are increments from α.

13.4.4 Residuals

To check CR model assumptions, binary logistic model partial residuals are again valuable. We separately fit a sequence of binary logistic models using a series of binary events and the corresponding applicable (increasingly small) subsets of subjects, and plot smoothed partial residuals against X for all of the binary events. Parallelism in these plots indicates that the CR model's constant γ assumptions are satisfied.

13.4.5 Assessment of Model Fit

The partial residual plots just described are very useful for checking the constant slope assumption of the CR model. The next section shows how to test this assumption formally. Linearity can be assessed visually using the smoothed partial residual plot, and interactions between predictors can be tested as usual.

13.4.6 Extended CR Model

The PO model has been extended by Peterson and Harrell[335] to allow for unequal slopes for some or all of the Xs for some or all levels of Y. This partial PO model requires specialized software, and with the demise of SAS Version 5 `PROC LOGIST`, software is not currently available. The CR model can be extended similarly. In S-PLUS notation, the ordinary CR model is specified as

```
y ~ cohort + X1 + X2 + X3 + ... ,
```

with `cohort` denoting a polytomous variable. The CR model can be extended to allow for some or all of the βs to change with the cohort or Y-cutoff.[24] Suppose that nonconstant slope is allowed for `X1` and `X2`. The S-PLUS notation for the extended model would be

```
y ~ cohort*(X1 + X2) + X3
```

The extended CR model is a discrete version of the Cox survival model with time-dependent covariables.

There is nothing about the CR model that makes it fit a given dataset better than other ordinal models such as the PO model. The real benefit of the CR model is that using standard binary logistic model software one can flexibly specify how the equal-slopes assumption can be relaxed.

13.4.7 Role of Penalization in Extended CR Model

As demonstrated in the upcoming case study, penalized MLE is invaluable in allowing the model to be extended into an unequal-slopes model insofar as the information content in the data will support. Faraway[134] has demonstrated how all data-driven steps of the modeling process increase the real variance in "final" parameter estimates, when one estimates variances without assuming that the final model was prespecified. For ordinal regression modeling, the most important modeling steps are (1) choice of predictor variables, (2) selecting or modeling predictor transformations, and (3) allowance for unequal slopes across Y-cutoffs (i.e., non-PO or non-CR). Regarding Steps (2) and (3) one is tempted to rely on graphical methods such as residual plots to make detours in the strategy, but it is very difficult to estimate variances or to properly penalize assessments of predictive accuracy for subjective modeling decisions. Regarding (1), shrinkage has been proven to work better than stepwise variable selection when one is attempting to build a main-effects model. Choosing a shrinkage factor is a well-defined, smooth, and often a unique process as opposed to binary decisions on whether variables are "in" or "out" of the model. Likewise, instead of using arbitrary subjective (residual plots) or objective (χ^2 due to `cohort` × covariable interactions, i.e., nonconstant covariable effects), shrinkage can systematically allow model enhancements insofar as the information content in the data will support, through the use of differential penalization. Shrinkage is a solution to the dilemma faced when the analyst attempts to choose between a parsimonious model and a more complex one that fits the data. Penalization does not require the analyst to make a binary decision, and it is a process that can be validated using the bootstrap.

13.4.8 Validating the Fitted Model

Validation of statistical indexes such as D_{xy} and model calibration is done using techniques discussed previously, except that certain problems must be addressed. First, when using the bootstrap, the resampling must take into account the existence of multiple records per subject that were created to use the binary logistic likelihood trick. That is, sampling should be done with replacement from *subjects* rather than *records*. Second, the analyst must isolate which event to predict. This is because when observations are expanded in order to use a binary logistic likelihood function to fit the CR model, several different events are being predicted simultaneously. Somers' D_{xy} could be computed by relating $X\hat{\gamma}$ (ignoring intercepts) to the ordinal Y, but other indexes are not defined so easily. The simplest approach here would be to validate a single prediction for $\Pr(Y = j | Y \geq j, X)$, for example. The simplest event to predict is $\Pr(Y = 0|X)$, as this would just require subsetting on all observations in the first cohort level in the validation sample. It would also be easy to validate any one of the later conditional probabilities. The validation functions described in the next section allow for such subsetting, as well as handling the cluster sampling. Specialized calculations would be needed to validate an unconditional probability such as $\Pr(Y \geq 2|X)$.

13.4.9 S-PLUS Functions

The **cr.setup** function in **Design** returns a list of vectors useful in constructing a dataset used to trick a binary logistic function such as **lrm** into fitting CR models. The **subs** vector in this list contains observation numbers in the original data, some of which are repeated. Here is an example.

```
u ← cr.setup(Y)              # Y is original ordinal response vector
attach(mydata[u$subs,])      # mydata is the original dataset
                             # mydata[i,] subscripts the input data,
                             # using duplicate values of i for repeats
y      ← u$y                 # constructed binary responses
cohort ← u$cohort            # cohort or risk set categories

f ← lrm(y ~ cohort*age + sex)
```

Since the **lrm** and **pentrace** functions have the capability to penalize different parts of the model by different amounts, they are valuable for fitting extended CR models in which the **cohort** × predictor interactions are allowed to be only as important as the information content in the data will support. Simple main effects can be unpenalized or slightly penalized as desired.

The **validate** and **calibrate** functions for **lrm** allow specification of subject identifiers when using the bootstrap, so the samples can be constructed with replacement from the original subjects. In other words, cluster sampling is done from the expanded records. This is handled internally by the **predab.resample** function. These functions also allow one to specify a subset of the records to use in the validation, which makes it especially easy to validate the part of the model used to predict $\Pr(Y = 0|X)$.

The **plot.xmean.ordinaly** function is useful for checking the CR assumption for single predictors, as described earlier.

13.5 Further Reading

[1] See [5, 18, 19, 24, 25, 36, 44, 45, 78, 88, 164, 168, 192, 242, 309, 335, 380, 452, 459] for some excellent background references, applications, and extensions to the ordinal models.

[2] Anderson and Philips [19, p. 29] proposed methods for constructing properly spaced response values given a fitted PO model.

[3] The simplest demonstration of this is to consider a model in which there is a single predictor that is totally independent of a nine-level response Y, so PO *must* hold. A PO model is fitted in SAS using:

```
DATA test;
DO i=1 to 50;
y=FLOOR(RANUNI(151)*9);
x=RANNOR(5);
OUTPUT;
END;
PROC LOGISTIC; MODEL y=x;
```

The score test for PO was $\chi^2 = 56$ on 7 d.f., $P < 0.0001$. This problem results from some small cell sizes in the distribution of Y.[335] The P-value for testing the regression effect for X was 0.76.

13.6 Problems

Test for the association between disease group and total hospital cost in SUPPORT, without imputing any missing costs (exclude the one patient having zero cost).

1. Use the Kruskal–Wallis rank test.
2. Use the proportional odds ordinal logistic model generalization of the Wilcoxon–Mann–Whitney Kruskal–Wallis Spearman test. Group total cost into 20 quantile groups so that only 19 intercepts will need to be in the model, not one less than the number of subjects (this would have taken the program too long to fit the model). Use the likelihood ratio χ^2 for this and later steps.
3. Use a binary logistic model to test for association between disease group and whether total cost exceeds the median of total cost. In other words, group total

cost into two quantile groups and use this binary variable as the response. What is wrong with this approach?

4. Instead of using only two cost groups, group cost into 3, 4, 5, 6, 8, 10, and 12 quantile groups. Describe the relationship between the number of intervals used to approximate the continuous response variable and the efficiency of the analysis. How many intervals of total cost, assuming that the ordering of the different intervals is used in the analysis, are required to avoid losing significant information in this continuous variable?

5. If you were selecting one of the rank-based tests for testing the association between disease and cost, which of any of the tests considered would you choose?

6. Why do all of the tests you did have the same number of degrees of freedom for the hypothesis of no association between `dzgroup` and `totcst`?

7. What is the advantage of a rank-based test over a parametric test based on log(cost)?