Chapter 15
Regression Models for Continuous Y and Case Study in Ordinal Regression

This chapter concerns univariate continuous Y. There are many multivariable models for predicting such response variables, such as

- linear models with assumed normal residuals, fitted with ordinary least squares
- generalized linear models and other parametric models based on special distributions such as the gamma
- generalized additive models (GAMs)[277]
- generalization of GAMs to also nonparametrically transform Y (see Chapter 16)
- quantile regression (see Section 15.2)
- other robust regression models that, like quantile regression, use an objective different from minimizing the sum of squared errors[635]
- semiparametric models based on the ranks of Y, such as the Cox proportional hazards model (Chapter 20) and the proportional odds ordinal logistic model (Chapters 13 and 14)
- cumulative probability models (often called *cumulative link models*) which are semiparametric models from a wider class of families than the logistic.

Semiparametric models that treat Y as ordinal but not interval-scaled have many advantages including robustness and freedom from all distributional assumptions for Y conditional on any given set of predictors. Advantages are demonstrated in a case study of a cumulative probability ordinal model. Some of the results are compared to quantile regression and OLS. Many of the methods used in the case study also apply to ordinary linear models.

15.1 The Linear Model

The most popular multivariable model for analyzing a univariate continuous Y is the linear model

F.E. Harrell, Jr., *Regression Modeling Strategies*, Springer Series in Statistics, DOI 10.1007/978-3-319-19425-7_15

$$E(Y|X) = X\beta, \tag{15.1}$$

where β is estimated using ordinary least squares, that is, by solving for $\hat{\beta}$ to minimize $\sum(Y_i - X\hat{\beta})^2$.

To compute P-values and confidence limits using parametric methods we would have to assume that $Y|X$ is normal with mean $X\beta$ and constant variance σ^2[a]. One could estimate conditional means of Y without any distributional assumptions, but least squares estimators are not robust to outliers or high-leverage points, and the model would be inaccurate in estimating conditional quantiles of $Y|X$ or $\text{Prob}[Y \geq c|X]$ unless normality of residuals holds. To be accurate in estimating all quantities, the linear model assumes that the Gaussian distribution of $Y|X_1$ is a simple shift from the distribution of $Y|X_2$.

15.2 Quantile Regression

Quantile regression[355, 357] is a different approach to modeling Y. It makes no distributional assumptions other than continuity of Y, while having all the usual right hand side assumptions. Quantile regression provides essentially the same estimates as sample quantiles if there is only an intercept or a categorical predictor in the model. Quantile regression is transformation invariant — pre-transforming Y is not important.

Quantile regression is a natural generalization of sample quantiles. Let $\rho_\tau(y) = y(\tau - [y < 0])$. The τ^{th} sample quantile is the minimizer q of $\sum_{i-1}^{n} \rho_\tau(y_i - q)$. For a conditional τ^{th} quantile of $Y|X$ the corresponding quantile regression estimator $\hat{\beta}_\tau$ minimizes $\sum_{i=1}^{n} \rho_\tau(Y_i - X\beta)$.

In non-large samples, quantile regression is not as efficient at estimating quantiles as is ordinary least squares at estimating the mean, if the latter's assumptions hold.

Koenker's `quantreg` package in R[356] implements quantile regression, and the `rms` package's `Rq` function provides a front-end that gives rise to various graphics and inference tools.

Using quantile regression, we directly model the median as a function of covariates so that only the $X\beta$ structure need be correct. Other quantiles (e.g., 90^{th} percentile) can be modeled but standard errors will be much larger as it is more difficult to precisely estimate outer quantiles.

[a] The latter assumption may be dispensed with if we use a robust Huber–White or bootstrap covariance matrix estimate. Normality may sometimes be dispensed with by using bootstrap confidence intervals.

15.3 Ordinal Regression Models for Continuous Y

A different robust semiparametric regression approach than quantile regression is the cumulative probability ordinal model. Semiparametric models have several advantages over parametric models such as OLS. While quantile regression has no restriction in the parameters when modeling one quantile versus another[b], ordinal cumulative probability models assume a connection between distributions of Y for different X. Ordinal regression even makes one less assumption than quantile regression about the distribution of Y for a specific X: the distribution need not be continuous.

Applying an increasing 1–1 transformation to Y results in no change to regression coefficient estimates with ordinal regression[c]. Regression coefficient estimates are completely robust to extreme Y values[d]. Estimates of quantiles of Y from ordinal regression are exactly transformation-preserving, e.g., the estimate of the median of $\log Y$ is exactly the log of the estimate of the median Y.

For a general continuous distribution function $F(y)$, an ordinal regression model based on cumulative probabilities may be stated as follows[e]. Let the ordered unique values of Y be denoted by $y_1, y_2, \ldots, y_k$ and let the intercepts associated with $y_1, \ldots, y_k$ be $\alpha_1, \alpha_2, \ldots, \alpha_k$, where $\alpha_1 = \infty$ because $\text{Prob}[Y \geq y_1] = 1$. Let $\alpha_y = \alpha_i, i : y_i = y$. Then

$$\text{Prob}[Y \geq y_i|X] = F(\alpha_i + X\beta) = F(\alpha_{y_i} + X\beta) \tag{15.2}$$

For the OLS fully parametric case, the model may be restated

$$\begin{aligned}\text{Prob}[Y \geq y|X] &= \text{Prob}[\frac{Y - X\beta}{\sigma} \geq \frac{y - X\beta}{\sigma}] &(15.3)\\ &= 1 - \Phi(\frac{y - X\beta}{\sigma}) = \Phi(\frac{-y}{\sigma} + \frac{X\beta}{\sigma}) &(15.4)\end{aligned}$$

[b] Quantile regression allows the estimated value of the 0.5 quantile to be higher than the estimated value of the 0.6 quantile for some values of X. Composite quantile regression[690] removes this possibility by forcing all the X coefficients to be the same across multiple quantiles, a restriction not unlike what cumulative probability ordinal models make.

[c] For symmetric distributions applying a decreasing transformation will negate the coefficients. For asymmetric distributions (e.g., Gumbel), reversing the order of Y will do more than change signs.

[d] Only an estimate of mean Y from these $\hat{\beta}$s is non-robust.

[e] It is more traditional to state the model in terms of $\text{Prob}[Y \leq y|X]$ but we use $\text{Prob}[Y \geq y|X]$ so that higher predicted values are associated with higher Y.

Table 15.1 Distribution families used in ordinal cumulative probability models. Φ denotes the Gaussian cumulative distribution function. For the Connection column, $P_1 = \text{Prob}[Y \geq y|X_1], P_2 = \text{Prob}[Y \geq y|X_2], \Delta = (X_2 - X_1)\beta$. The connection specifies the only distributional assumption if the model is fitted semiparametrically, i.e, contains an intercept for every unique Y value less one. For parametric models, P_1 must be specified absolutely instead of just requiring a relationship between P_1 and P_2. For example, the traditional Gaussian parametric model specifies that $\text{Prob}[Y \geq y|X] = 1 - \Phi(\frac{y - X\beta}{\sigma}) = \Phi(\frac{-y + X\beta}{\sigma})$.

Distribution	F	Inverse (Link Function)	Link Name	Connection
Logistic	$[1 + \exp(-y)]^{-1}$	$\log(\frac{y}{1-y})$	logit	$\frac{P_2}{1-P_2} = \frac{P_1}{1-P_1} \exp(\Delta)$
Gaussian	$\Phi(y)$	$\Phi^{-1}(y)$	probit	$P_2 = \Phi(\Phi^{-1}(P_1) + \Delta)$
Gumbel maximum value	$\exp(-\exp(-y))$	$\log(-\log(y))$	$\log - \log$	$P_2 = P_1^{\exp(\Delta)}$
Gumbel minimum value	$1 - \exp(-\exp(y))$	$\log(-\log(1 - y))$	complementary $\log - \log$	$1 - P_2 = (1 - P_1)^{\exp(\Delta)}$
Cauchy	$\frac{1}{\pi}\tan^{-1}(y) + \frac{1}{2}$	$\tan[\pi(y - \frac{1}{2})]$	cauchit	

so that to within an additive constant[f] $\alpha_y = \frac{-y}{\sigma}$ (intercepts α are linear in y whereas they are arbitrarily descending in the ordinal model), and σ is absorbed in β to put the OLS model into the new notation.

The general ordinal regression model assumes that for fixed X_1, X_2,

$$F^{-1}(\text{Prob}[Y \geq y|X_2]) - F^{-1}(\text{Prob}[Y \geq y|X_1]) \tag{15.5}$$

$$= (X_2 - X_1)\beta \tag{15.6}$$

independent of the αs (parallelism assumption). If $F = [1 + \exp(-y)]^{-1}$, this is the proportional odds assumption.

Common choices of F, implemented in the R `rms` `orm` function, are shown in Table 15.1. The Gumbel maximum value distribution is also called the extreme value type I distribution. This distribution ($\log - \log$ link) also represents a continuous time proportional hazards model. The hazard ratio when X changes from X_1 to X_2 is $\exp(-(X_2 - X_1)\beta)$.

The mean of $Y|X$ is easily estimated from a fitted cumulative probability ordinal model by computing

$$\sum_{i=1}^{n} y_i \widehat{\text{Prob}}[Y = y_i|X] \tag{15.7}$$

and the q^{th} quantile of $Y|X$ is y such that $F^{-1}(1 - q) - X\hat{\beta} = \hat{\alpha}_y$.[g]

[f] $\hat{\alpha_y}$ are unchanged if a constant is added to all y.

[g] The intercepts have to be shifted to the left one position in solving this equation because the quantile is such that $\text{Prob}[Y \leq y] = q$ whereas the model is stated in terms of $\text{Prob}[Y \geq y]$.

The `orm` function in the `rms` package takes advantage of the information matrix being of a sparse tri-band diagonal form for the intercept parameters. This makes the computations efficient even for hundreds of intercepts (i.e., unique values of Y). `orm` is made to handle continuous Y.

Ordinal regression has nice properties in addition to those listed above, allowing for

- estimation of quantiles as efficiently as quantile regression if the parallel slopes assumptions hold
- efficient estimation of mean Y
- direct estimation of Prob$[Y \geq y|X]$
- arbitrary clumping of values of Y, while still estimating β and mean Y efficiently[h]
- solutions for $\hat{\beta}$ using ordinary Newton-Raphson or other popular optimization techniques
- being based on a standard likelihood function, penalized estimation can be straightforward
- Wald, score, and likelihood ratio χ^2 tests that are more powerful than tests from quantile regression.

On the last point, if there is a single predictor in the model and it is binary, the score test from the proportional odds model is essentially the Wilcoxon test, and the score test from the Gumbel log-log cumulative probability model is essentially the log-rank test.

15.3.1 Minimum Sample Size Requirement

When Y is continuous and the purpose of an ordinal model includes semiparametric estimation of probabilities or quantiles, the accuracy of estimates is limited even more by the accuracy of estimating the empirical cumulative distribution of Y than by estimating β. When $\beta = 0$, intercept estimates are transformations of the empirical distribution step function. As described in Section 20.3, the sample size must be 184 to estimate the entire distribution of Y with a global margin of error not exceeding 0.1. For estimating the mean of Y, smaller sample sizes may be needed.

[h] But it is not sensible to estimate quantiles of Y when there are heavy ties in Y in the area containing the quantile.

15.4 Comparison of Assumptions of Various Models

Quantile regression makes the fewest left-hand-side model assumptions except for the assumption that Y be continuous, but can have less estimator precision than other models and has lower power. To summarize how assumptions of parametric models compare to assumptions of semiparametric ordinal models, consider the ordinary linear model or its special case the equal variance two-sample t-test, vs. the probit or logit (proportional odds) ordinal model or their special cases the Van der Waerden (normal-scores) two-sample rank test or the Wilcoxon two-sample test. All the assumptions of the linear model other than independence of residuals are captured in the following, using the more standard $Y \leq y$ notation:

$$F(y|X) = \text{Prob}[Y \leq y|X] = \Phi(\frac{y - X\beta}{\sigma}) \tag{15.8}$$

$$\Phi^{-1}(F(y|X)) = \frac{y - X\beta}{\sigma} \tag{15.9}$$

On the other hand, ordinal models assume the following:

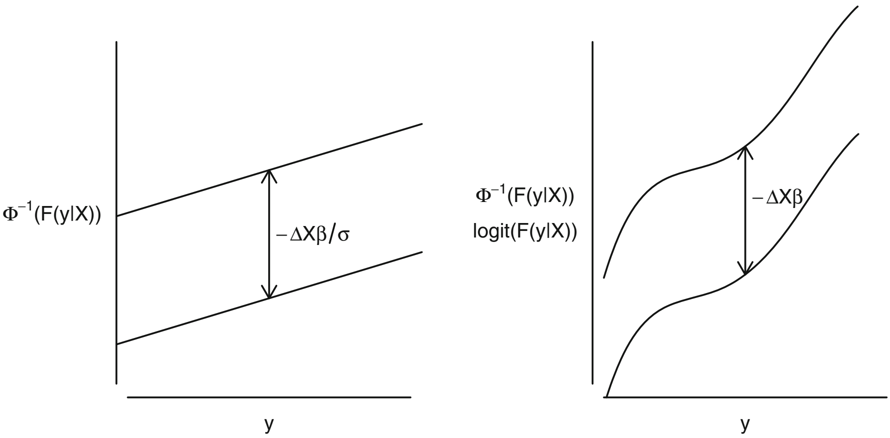


Fig. 15.1 Assumptions of the linear model (left panel) and semiparametric ordinal probit or logit (proportional odds) models (right panel). Ordinal models do not assume any shape for the distribution of Y for a given X; they only assume parallelism. The linear model can relax the parallelism assumption if σ is allowed to vary, but in practice it is difficult to know how to vary it except for the unequal variance two-sample t-test.

$$\text{Prob}[Y \leq y|X] = F(g(y) - X\beta), \tag{15.10}$$

where g is unknown and may be discontinuous. This translates to the parallelism assumption in the right panel of Figure 15.1, whereas the linear model makes the additional strong assumption of linearity of normal inverse cumulative distribution function, which arises from the Gaussian distribution assumption.

15.5 Dataset and Descriptive Statistics

Diabetes Mellitus (DM) type II (adult onset diabetes) is strongly associated with obesity. The currently best laboratory test for diabetes measures glycosylated hemoglobin (HbA$_{1c}$), also called glycated hemoglobin, glycohemoglobin, or hemoglobin A_{1c}. HbA$_{1c}$ reflects average blood glucose for the preceding 60 to 90 days. HbA$_{1c}$ > 7.0 is sometimes taken as a positive diagnosis of diabetes even though there are no data to support the use of a threshold.

The goals of this analyses are to better understand effects of body size measurements on risk of DM and to enhance screening for DM. The best way to develop a model for DM screening is **not** to fit a binary logistic model with HbA$_{1c}$ > 7 as the response variable. There are at least two reasons for this. First, when the relationship between a measurement and its ultimate clinical impact is smooth, all cutpoints are arbitrary. There is no justification for any putative cut on HbA$_{1c}$. Second, such an analysis loses information by treating HbA$_{1c}$=2 the same as HbA$_{1c}$=6.9, and by treating HbA$_{1c}$=7.1 as equal to HbA$_{1c}$=10. Failure to use all available information results in larger standard errors of $\hat{\beta}$, lower power, and wider confidence bands. It is better to predict continuous HbA$_{1c}$ using a continuous response model, then use that model to estimate the probability that HbA$_{1c}$ exceeds any cutoff, or estimate the 0.9 quantile of HbA$_{1c}$.

The data used here are from the National Health and Nutrition Examination Survey (NHANES) 2009–2010 from the U.S. National Center for Health Statistics/Centers for Disease Control. The original data may be obtained from http://www.cdc.gov/nchs/nhanes.htm[94]; the analysis file used here, called **nhgh**, may be obtained from the **DataSets** wiki page, along with R code used to download and create the file. Note that CDC coded age ≥ 80 as 80. We use the subset of subjects with age ≥ 21 who have neither been diagnosed nor treated for DM. Descriptive statistics are shown below.

```
require(rms)
```

```
getHdata(nhgh)
w ← subset(nhgh, age ≥ 21 & dx==0 & tx==0, select=-c(dx,tx))
latex(describe(w), file='')
```

```
                              w
           18 Variables       4629  Observations
```

seqn : Respondent sequence number

```
      n missing unique Info  Mean   .05   .10   .25   .50   .75   .90   .95
   4629       0   4629    1 56902 52136 52633 54284 56930 59495 61079 61641

   lowest : 51624 51629 51630 51645 51647
   highest: 62152 62153 62155 62157 62158
```

sex

```
      n missing unique
   4629       0      2

   male (2259, 49%), female (2370, 51%)
```

age : Age [years]

```
      n missing unique Info  Mean   .05   .10   .25   .50   .75   .90   .95
   4629       0    703    1 48.57 23.33 26.08 33.92 46.83 61.83 74.83 80.00

   lowest : 21.00 21.08 21.17 21.25 21.33
   highest: 79.67 79.75 79.83 79.92 80.00
```

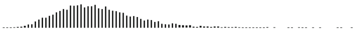
re : Race/Ethnicity

```
      n missing unique
   4629       0      5

   Mexican American (832, 18%), Other Hispanic (474, 10%)
   Non-Hispanic White (2318, 50%), Non-Hispanic Black (756, 16%)
   Other Race Including Multi-Racial (249, 5%)
```

income : Family Income

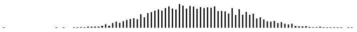
```
      n missing unique
   4389     240     14

   [0,5000) (162, 4%), [5000,10000) (216, 5%), [10000,15000) (371, 8%)
   [15000,20000) (300, 7%), [20000,25000) (374, 9%)
   [25000,35000) (535, 12%), [35000,45000) (421, 10%)
   [45000,55000) (346, 8%), [55000,65000) (257, 6%), [65000,75000) (188, 4%)
   > 20000 (149, 3%), < 20000 (52, 1%), [75000,100000) (399, 9%)
   >= 100000 (619, 14%)
```

wt : Weight [kg]



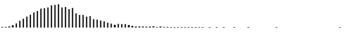
```
      n missing unique Info  Mean   .05   .10   .25   .50   .75    .90    .95
   4629       0    890    1 80.49 52.44 57.18 66.10 77.70 91.40 106.52 118.00

   lowest :  33.2  36.1  37.9  38.5  38.7
   highest: 184.3 186.9 195.3 196.6 203.0
```

ht : Standing Height [cm]



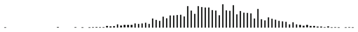
```
      n missing unique Info  Mean   .05   .10   .25   .50   .75   .90   .95
   4629       0    512    1 167.5 151.1 154.4 160.1 167.2 175.0 181.0 184.8

   lowest : 123.3 135.4 137.5 139.4 139.8
   highest: 199.2 199.3 199.6 201.7 202.7
```

bmi : Body Mass Index [kg/m^2]



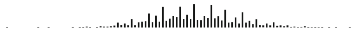
```
   n missing unique Info  Mean   .05   .10   .25   .50   .75   .90   .95
4629       0   1994    1 28.59 20.02 21.35 24.12 27.60 31.88 36.75 40.68

lowest : 13.18 14.59 15.02 15.40 15.49
highest: 61.20 62.81 65.62 71.30 84.87
```

leg : Upper Leg Length [cm]



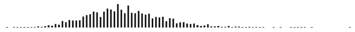
```
   n missing unique Info  Mean  .05  .10  .25  .50  .75  .90  .95
4474     155    216    1 38.39 32.0 33.5 36.0 38.4 41.0 43.3 44.6

lowest : 20.4 24.9 25.0 25.1 26.4, highest: 49.0 49.5 49.8 50.0 50.3
```

arml : Upper Arm Length [cm]



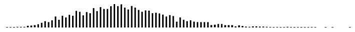
```
   n missing unique Info  Mean  .05  .10  .25  .50  .75  .90  .95
4502     127    156    1 37.01 32.6 33.5 35.0 37.0 39.0 40.6 41.7

lowest : 24.8 27.0 27.5 29.2 29.5, highest: 45.2 45.5 45.6 46.0 47.0
```

armc : Arm Circumference [cm]



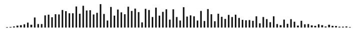
```
   n missing unique Info  Mean  .05  .10  .25  .50  .75  .90  .95
4499     130    290    1 32.87 25.4 26.9 29.5 32.5 35.8 39.1 41.4

lowest : 17.9 19.0 19.3 19.5 19.9, highest: 54.2 54.9 55.3 56.0 61.0
```

waist : Waist Circumference [cm]



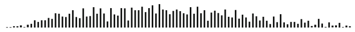
```
   n missing unique Info  Mean  .05  .10  .25  .50   .75   .90   .95
4465     164    716    1 97.62 74.8 78.6 86.9 96.3 107.0 117.8 125.0

lowest :  59.7  60.0  61.5  62.0  62.4
highest: 160.0 160.6 162.2 162.7 168.7
```

tri : Triceps Skinfold [mm]



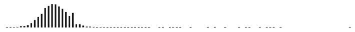
```
   n missing unique Info  Mean .05 .10  .25  .50  .75  .90  .95
4295     334    342    1 18.94 7.2 8.8 12.0 18.0 25.2 31.0 33.8

lowest :  2.6  3.1  3.2  3.3  3.4, highest: 39.6 39.8 40.0 40.2 40.6
```

sub : Subscapular Skinfold [mm]



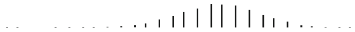
```
   n missing unique Info Mean  .05   .10   .25   .50   .75   .90   .95
3974     655    329    1 20.8 8.60 10.30 14.40 20.30 26.58 32.00 35.00

lowest :  3.8  4.2  4.6  4.8  4.9, highest: 40.0 40.1 40.2 40.3 40.4
```

gh : Glycohemoglobin [%]



```
   n missing unique Info  Mean .05 .10 .25 .50 .75 .90 .95
4629       0     63 0.99 5.533 4.8 5.0 5.2 5.5 5.8 6.0 6.3

lowest :  4.0  4.1  4.2  4.3  4.4, highest: 11.9 12.0 12.1 12.3 14.5
```

albumin : Albumin [g/dL]



```
   n missing unique Info  Mean .05 .10 .25 .50 .75 .90 .95
4576      53     26 0.99 4.261 3.7 3.9 4.1 4.3 4.5 4.7 4.8

lowest : 2.6 2.7 3.0 3.1 3.2, highest: 4.9 5.0 5.1 5.2 5.3
```

bun : Blood urea nitrogen [mg/dL]

```
n missing unique Info  Mean .05 .10 .25 .50 .75 .90 .95
4576     53     50 0.99 13.03   7   8  10  12  15  19  22

lowest :  1  2  3  4  5, highest: 49 53 55 56 63
```

SCr : Creatinine [mg/dL]

```
n missing unique Info   Mean  .05  .10  .25  .50  .75  .90  .95
4576     53    167    1 0.8887 0.58 0.62 0.72 0.84 0.99 1.14 1.25

lowest :  0.34  0.38  0.39  0.40  0.41
highest:  5.98  6.34  9.13 10.98 15.66
```

```
dd ← datadist(w); options(datadist='dd')
```

15.5.1 Checking Assumptions of OLS and Other Models

First let's see if gh would make a Gaussian residuals model fit. Use ordinary regression on four key variables to collapse these into one variable (predicted mean from the OLS model). Stratify the predicted means into six quantile groups. Apply the normal inverse cumulative distribution function Φ^{-1} to the empirical cumulative distribution functions (ECDF) of gh using these strata, and check for normality and constant σ^2. The ECDF estimates Prob$[Y \leq y|X]$ but for ordinal modeling we want to state models in terms of Prob$[Y \geq y|X]$ so take one minus the ECDF before inverse transforming.

```
f ← ols(gh ∼ rcs(age,5) + sex + re + rcs(bmi, 3), data=w)
pgh ← fitted(f)

p ← function(fun, row, col) {
  f ← substitute(fun); g ← function(F) eval(f)
  z ← Ecdf(∼ gh, groups=cut2(pgh, g=6),
           fun=function(F) g(1 - F),
           ylab=as.expression(f), xlim=c(4.5, 7.75), data=w,
           label.curve=FALSE)
  print(z, split=c(col, row, 2, 2), more=row < 2 | col < 2)
}
p(log(F/(1-F)),    1, 1)
p(qnorm(F),        1, 2)
p(-log(-log(F)),   2, 1)
p(log(-log(1-F)),  2, 2)
# Get slopes of pgh for some cutoffs of Y
# Use glm complementary log-log link on Prob(Y < cutoff) to
# get log-log link on Prob(Y ≥ cutoff)
r ← NULL
for(link in c('logit','probit','cloglog'))
  for(k in c(5, 5.5, 6)) {
    co ← coef(glm(gh < k ∼ pgh, data=w, family=binomial(link)))
    r ← rbind(r, data.frame(link=link, cutoff=k,
                                       slope=round(co[2],2)))
}
print(r, row.names=FALSE)
```

```
    link cutoff slope
   logit    5.0 -3.39
   logit    5.5 -4.33
   logit    6.0 -5.62
  probit    5.0 -1.69
  probit    5.5 -2.61
  probit    6.0 -3.07
 cloglog    5.0 -3.18
 cloglog    5.5 -2.97
 cloglog    6.0 -2.51
```

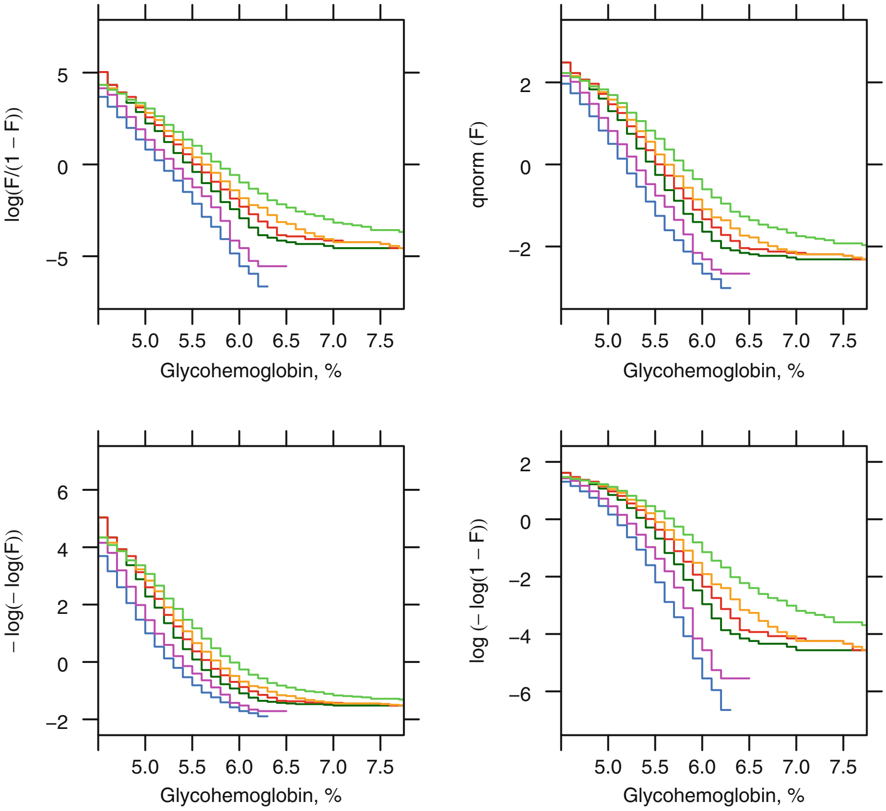


Fig. 15.2 Examination of normality and constant variance assumption, and assumptions for various ordinal models

The upper right curves in Figure 15.2 are not linear, implying that a normal conditional distribution cannot work for `gh`[i] There is non-parallelism for the logit model. The other graphs will be used to guide selection of an ordinal model below.

15.6 Ordinal Regression Applied to HbA$_{1c}$

In the upper left panel of Figure 15.2, logit inverse curves are not parallel so the proportional odds assumption does not hold when predicting HbA$_{1c}$. The log-log link yields the highest degree of parallelism and most constant regression coefficients across cutoffs of `gh`, so we use this link in an ordinal regression model (linearity of the curves is not required).

15.6.1 Checking Fit for Various Models Using Age

Another way to examine model fit is to flexibly fit the single most important predictor (age) using a variety of methods, and compare predictions to sample quantiles and means based on subsets on age. We use overlapping subsets to gain resolution, with each subset composed of those subjects having age within five years of the point being predicted by the models. Here we predict the 0.5, 0.75, and 0.9 quantiles and the mean. For quantiles we can compare to quantile regression (discussed below) and for means we compare to OLS.

```
ag ← 25:75
lag ← length(ag)
q2 ← q3 ← p90 ← means ← numeric(lag)
for(i in 1:lag) {
  s ← which(abs(w$age - ag[i]) < 5)
  y ← w$gh[s]
  a ← quantile(y, probs=c(.5, .75, .9))
  q2[i]    ← a[1]
  q3[i]    ← a[2]
  p90[i]   ← a[3]
  means[i] ← mean(y)
}
fams ← c('logistic', 'probit', 'loglog', 'cloglog')
fe   ← function(pred, target) mean(abs(pred$yhat - target))
mod  ← gh ~ rcs(age,6)
P    ← Er ← list()
for(est in c('q2', 'q3', 'p90', 'mean')) {
  meth ← if(est == 'mean') 'ols' else 'QR'
  p ← list()
  er ← rep(NA, 5)
  names(er) ← c(fams, meth)
  for(family in fams) {
    h  ← orm(mod, family=family, data=w)
    fun ← if(est == 'mean') Mean(h)
    else {
      qu ← Quantile(h)
```

[i] They are not parallel either.

```
      switch(est, q2  = function(x) qu(.5,  x),
                  q3  = function(x) qu(.75, x),
                  p90 = function(x) qu(.9,  x))
    }
    p[[family]] ← z ← Predict(h, age=ag, fun=fun, conf.int=FALSE)
    er[family] ← fe(z, switch(est, mean=means, q2=q2, q3=q3, p90=p90))
  }
  h ← switch(est,
             mean= ols(mod, data=w),
             q2  = Rq (mod, data=w),
             q3  = Rq (mod, tau=0.75, data=w),
             p90 = Rq (mod, tau=0.90, data=w))
  p[[meth]] ← z ← Predict(h, age=ag, conf.int=FALSE)
  er[meth] ← fe(z, switch(est, mean=means, q2=q2, q3=q3, p90=p90))

  Er[[est]] ← er
  pr ← do.call('rbind', p)
  pr$est ← est
  P ← rbind.data.frame(P, pr)
}

xyplot(yhat ~ age | est, groups=.set., data=P, type='l', # Figure 15.3
       auto.key=list(x=.75, y=.2, points=FALSE, lines=TRUE),
       panel=function(..., subscripts) {
         panel.xyplot(..., subscripts=subscripts)
         est ← P$est[subscripts[1]]
         lpoints(ag, switch(est, mean=means, q2=q2, q3=q3, p90=p90),
                 col=gray(.7))
         er ← format(round(Er[[est]],3), nsmall=3)
         ltext(26, 6.15, paste(names(er), collapse='\n'),
               cex=.7, adj=0)
         ltext(40, 6.15, paste(er, collapse='\n'),
               cex=.7, adj=1)})
```

It can be seen in Figure 15.3 that models dedicated to a specific task (quantile regression for quantiles and OLS for means) were best for those tasks. Although the log-log ordinal cumulative probability model did not estimate the median as accurately as some other methods, it does well for the 0.75 and 0.9 quantiles and is the best compromise overall because of its ability to also directly predict the mean as well as quantities such as $\mathrm{Prob}[\mathrm{HbA_{1c}} > 7|X]$.

From here on we focus on the log-log ordinal model. Returning to the bottom left of Figure 15.2, let's look at quantile groups of predicted $\mathrm{HbA_{1c}}$ by OLS and plot predicted distributions of actual $\mathrm{HbA_{1c}}$ against empirical distributions.

```
w$pghg ← cut2(pgh, g=6)
f  ← orm(gh ~ pghg, data=w)
lp ← predict(f, newdata=data.frame(pghg=levels(w$pghg)))
ep ← ExProb(f)  # Exceedance prob. functn. generator in rms
z  ← ep(lp)
j  ← order(w$pghg)  # puts in order of lp (levels of pghg)
plot(z, xlim=c(4, 7.5), data=w[j,c('pghg', 'gh')]) # Fig. 15.4
```

Agreement between predicted and observed exceedance probability distributions is excellent in Figure 15.4.

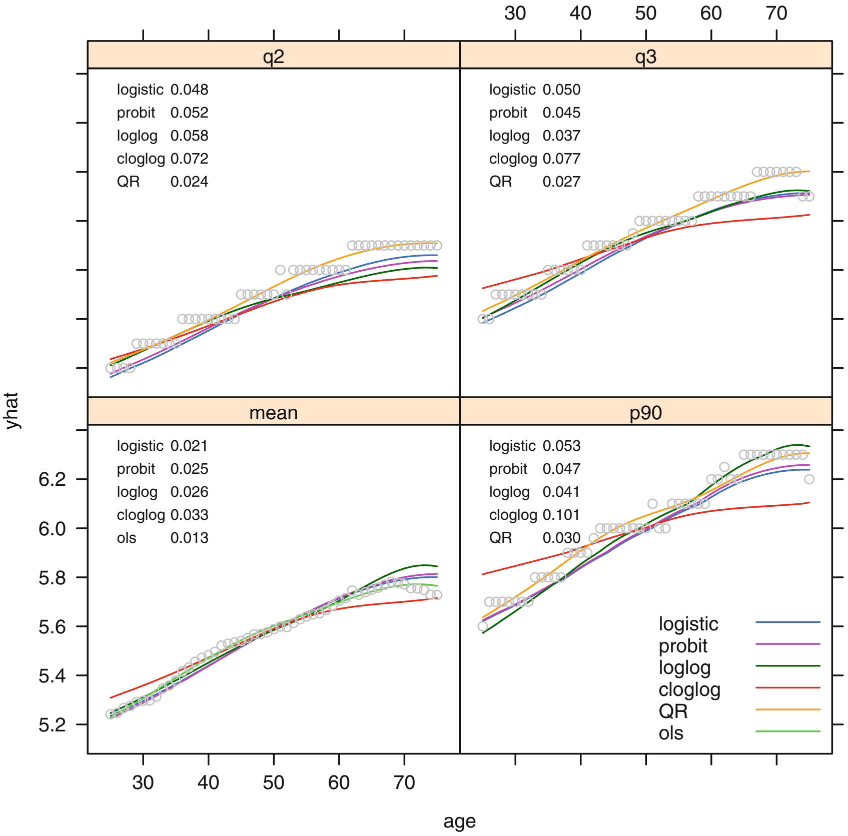


Fig. 15.3 Three estimated quantiles and estimated mean using 6 methods, compared against caliper-matched sample quantiles/means (circles). Numbers are mean absolute differences between predicted and sample quantiles using overlapping intervals of age and caliper matching. QR:quantile regression.

To return to the initial look at a linear model with assumed Gaussian residuals, fit a probit ordinal model and compare the estimated intercepts to the linear relationship with `gh` that is assumed by the normal distribution.

```
f ← orm(gh ∼ rcs(age,6), family=probit, data=w)
g ← ols(gh ∼ rcs(age,6), data=w)
s ← g$stats['Sigma']
yu ← f$yunique[-1]
r ← quantile(w$gh, c(.005, .995))
alphas ← coef(f)[1:num.intercepts(f)]
plot(-yu / s, alphas, type='l', xlim=rev(- r / s), # Fig. 15.5
   xlab=expression(-y/hat(sigma)), ylab=expression(alpha[y]))
```

Figure 15.5 depicts a significant departure from the linear form implied by Gaussian residuals (Eq. 15.4).

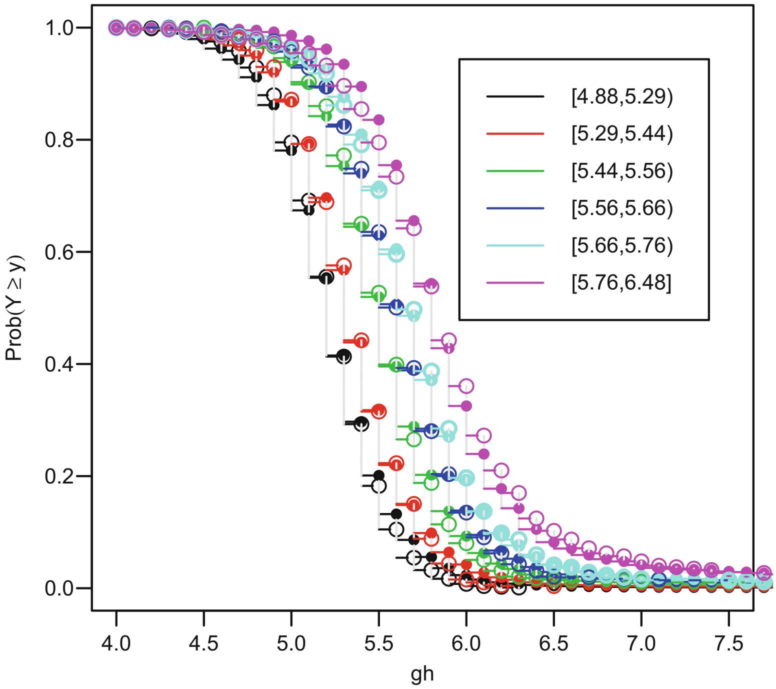


Fig. 15.4 Observed (dashed lines, open circles) and predicted (solid lines, closed circles) exceedance probability distributions from a model using 6-tiles of OLS-predicted HbA$_{1c}$. Key shows quantile group intervals of predicted mean HbA$_{1c}$.

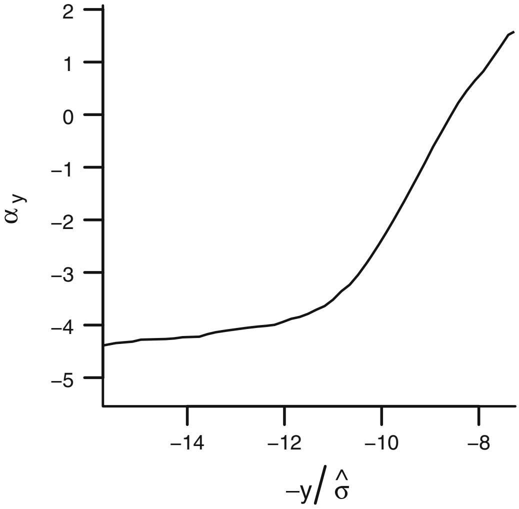


Fig. 15.5 Estimated intercepts from probit model. Linearity would have indicated Gaussian residuals.

15.6.2 Examination of BMI

Body mass index (BMI, weight divided by height^2) is commonly used as an obesity measure because it is well correlated with abdominal visceral fat. But it is not obvious that BMI is the correct summary of height and weight for predicting pre-clinical diabetes, and it may be the case that body size measures other than height and weight are better predictors.

Use the log-log ordinal model to check the adequacy of BMI, adjusting for age (without assuming linearity). This can be done by examining the ratio of coefficients of log height and log weight, and also by using AIC to judge whether BMI is an adequate summary of height and weight when compared to nonlinear functions of the logs, and to a tensor spline interaction surface.

```
f ← orm(gh ∼ rcs(age,5) + log(ht) + log(wt),
        family=loglog, data=w)
print(f, latex=TRUE)
```

-log-log Ordinal Regression Model

```
orm(formula = gh ~ rcs(age, 5) + log(ht) + log(wt), data = w,
    family = loglog)
```

	Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs 4629	LR χ^2 1126.94	R^2 0.217	ρ 0.486
Unique Y 63	d.f. 6	g 0.627	
$Y_{0.5}$ 5.5	Pr(> χ^2) < 0.0001	g_r 1.872	
max $\lvert\frac{\partial \log L}{\partial \beta}\rvert$	Score χ^2 1262.81	$\lvert\Pr(Y \geq Y_{0.5}) - \frac{1}{2}\rvert$ 0.153	
1×10^{-6}	Pr(> χ^2) < 0.0001		

	Coef	S.E.	Wald Z	Pr(> $\lvert Z\rvert$)
age	0.0398	0.0055	7.29	< 0.0001
age’	-0.0158	0.0275	-0.57	0.5657
age”	-0.0072	0.0866	-0.08	0.9333
age”’	0.0309	0.1135	0.27	0.7853
ht	-3.0680	0.2789	-11.00	< 0.0001
wt	1.2748	0.0704	18.10	< 0.0001

```
aic ← NULL
for(mod in list(gh ∼ rcs(age,5) + rcs(log(bmi),5),
                gh ∼ rcs(age,5) + rcs(log(ht),5) + rcs(log(wt),5),
                gh ∼ rcs(age,5) + rcs(log(ht),4) * rcs(log(wt),4)))
  aic ← c(aic, AIC(orm(mod, family=loglog, data=w)))
print(aic)
```

```
[1] 25910.77 25910.17 25906.03
```

The ratio of the coefficient of log height to the coefficient of log weight is - 2.4, which is between what BMI uses and the more dimensionally reasonable weight / height^3. By AIC, a spline interaction surface between height and weight does slightly better than BMI in predicting HbA$_{1c}$, but a nonlinear function of BMI is barely worse. It will require other body size measures to displace BMI as a predictor.

As an aside, compare this model fit to that from the Cox proportional hazards model. The Cox model uses a conditioning argument to obtain a partial likelihood free of the intercepts α (and requires a second step to estimate these log discrete hazard components) whereas we are using a full marginal likelihood of the ranks of Y^{330}.

```
print(cph(Surv(gh) ~ rcs(age,5) + log(ht) + log(wt), data=w),
      latex=TRUE)
```

Cox Proportional Hazards Model

```
cph(formula = Surv(gh) ~ rcs(age, 5) + log(ht)
              + log(wt), data = w)
```

		Model Tests		Discrimination Indexes	
Obs	4629	LR χ^2	1120.20	R^2	0.215
Events	4629	d.f.	6	D_{xy}	0.359
Center	8.3792	Pr($>\chi^2$)	0.0000	g	0.622
		Score χ^2	1258.07	g_r	1.863
		Pr($>\chi^2$)	0.0000		

	Coef	S.E.	Wald Z	Pr($>\|Z\|$)
age	-0.0392	0.0054	-7.24	$<$ 0.0001
age'	0.0148	0.0274	0.54	0.5888
age''	0.0093	0.0862	0.11	0.9144
age'''	-0.0321	0.1131	-0.28	0.7767
ht	3.0477	0.2779	10.97	$<$ 0.0001
wt	-1.2653	0.0701	-18.04	$<$ 0.0001

Close agreement of the two is seen, as expected.

15.6.3 Consideration of All Body Size Measurements

Next we examine all body size measures, and check their redundancies.

```
v ← varclus(~ wt + ht + bmi + leg + arml + armc + waist +
              tri + sub + age + sex + re, data=w)
plot(v)
```

```
# Omit wt so it won't be removed before bmi
redun(∼ ht + bmi + leg + arml + armc + waist + tri + sub,
      data=w, r2=.75)
```

```
Redundancy Analysis

redun(formula = ∼ht + bmi + leg + arml + armc + waist + tri +
    sub, data = w, r2 = 0.75)

n: 3853          p: 8    nk: 3

Number of NAs:   776
Frequencies of Missing Values Due to Each Variable
   ht    bmi    leg   arml   armc waist    tri    sub
    0      0    155    127    130   164    334    655


Transformation of target variables forced to be linear

R^2 cutoff: 0.75          Type: ordinary

R^2 with which each variable can be predicted from all other variables:

   ht    bmi   leg   arml  armc waist   tri   sub
0.829 0.924 0.682 0.748 0.843 0.864 0.531 0.594

Rendundant variables:

bmi ht

Predicted from variables:

leg arml armc waist tri sub

  Variable Deleted   R^2 R^2 after later deletions
1              bmi 0.924                     0.909
2               ht 0.792
```

Six size measures adequately capture the entire set. Height and BMI are removed (Figure 15.6). An advantage of removing height is that it is age-dependent due to vertebral compression in the elderly:

```
f ← orm(ht ∼ rcs(age,4)*sex, data=w)  # Prop. odds model
qu ← Quantile(f); med ← function(x) qu(.5, x)
ggplot(Predict(f, age, sex, fun=med, conf.int=FALSE),
       ylab='Predicted Median Height, cm')
```

However, upper leg length has the same declining trend, implying a survival bias or birth year effect.

In preparing to create a multivariable model, degrees of freedom are allocated according to the generalized Spearman ρ^2(Figure 15.7)[j].

```
s ← spearman2(gh ∼ age + sex + re + wt + leg + arml + armc +
                   waist + tri + sub, data=w, p=2)
plot(s)
```

Parameters will be allocated in descending order of ρ^2. But note that subscapular skinfold has a large number of NAs and other predictors also have NAs. Suboptimal casewise deletion will be used until the final model is fitted (Figure 15.8).

[j] Competition between collinear size measures hurts interpretation of partial tests of association in a saturated additive model.

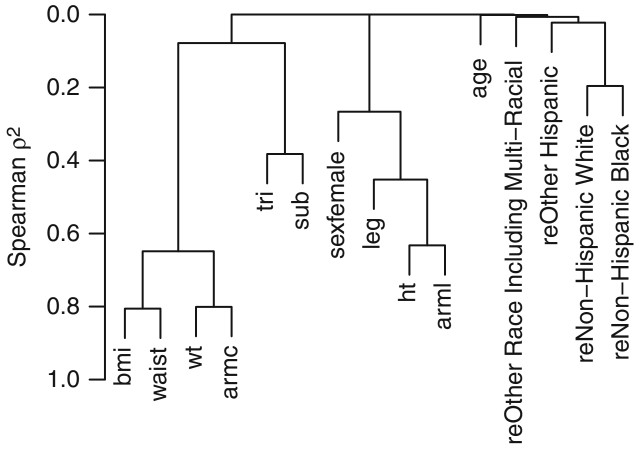


Fig. 15.6 Variable clustering for all potential predictors

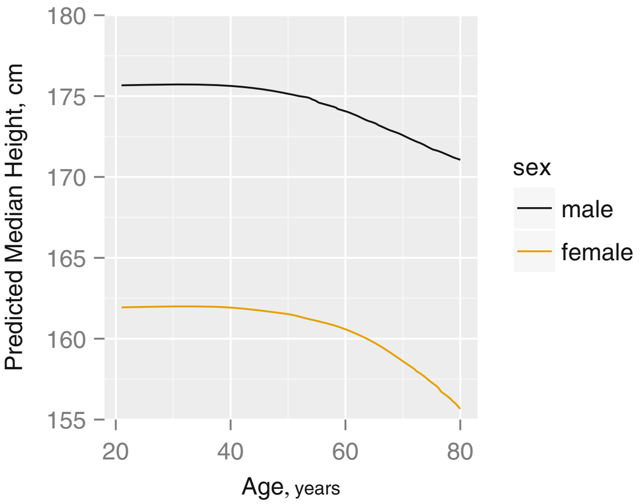


Fig. 15.7 Estimated median height as a smooth function of age, allowing age to interact with sex, from a proportional odds model

Because there are many competing body measures, we use backwards step-down to arrive at a set of predictors. The bootstrap will be used to penalize predictive ability for variable selection. First the full model is fit using casewise deletion, then we do a composite test to assess whether any of the frequently–missing predictors is important.

```
f ← orm(gh ∼ rcs(age,5) + sex + re + rcs(wt,3) + rcs(leg,3) + arml +
         rcs(armc,3) + rcs(waist,4) + tri + rcs(sub,3),
         family='loglog', data=w, x=TRUE, y=TRUE)
print(f, latex=TRUE, coefs=FALSE)
```

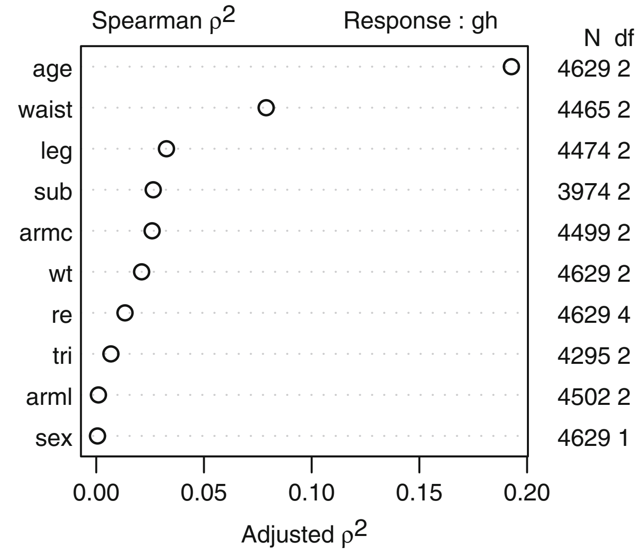
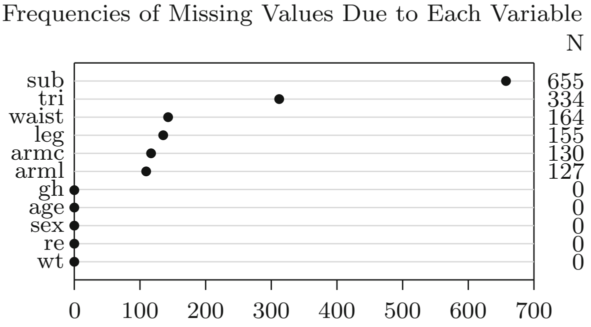


Fig. 15.8 Generalized squared rank correlations

-log-log Ordinal Regression Model

```
orm(formula = gh ~ rcs(age, 5) + sex + re + rcs(wt, 3)
    + rcs(leg, 3) + arml + rcs(armc, 3) + rcs(waist, 4)
    + tri + rcs(sub, 3), data = w, x = TRUE, y = TRUE,
    family = "loglog")
```



	Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs 3853	LR χ^2 1180.13	R^2 0.265	ρ 0.520
Unique Y 60	d.f. 22	g 0.732	
$Y_{0.5}$ 5.5	$\Pr(>\chi^2) < 0.0001$	g_r 2.080	
$\max\lvert\frac{\partial \log L}{\partial \beta}\rvert$	Score χ^2 1298.88	$\lvert\Pr(Y \geq Y_{0.5}) - \frac{1}{2}\rvert$ 0.172	
3×10^{-5}	$\Pr(>\chi^2) < 0.0001$		

```
## Composite test:
lan ← function(a) latex(a, table.env=FALSE, file='')
lan(anova(f, leg, arml, armc, waist, tri, sub))
```

	χ^2	d.f.	P
leg	8.30	2	0.0158
Nonlinear	3.32	1	0.0685
arml	0.16	1	0.6924
armc	6.66	2	0.0358
Nonlinear	3.29	1	0.0695
waist	29.40	3	< 0.0001
Nonlinear	4.29	2	0.1171
tri	16.62	1	< 0.0001
sub	40.75	2	< 0.0001
Nonlinear	4.50	1	0.0340
TOTAL NONLINEAR	14.95	5	0.0106
TOTAL	128.29	11	< 0.0001

The model achieves Spearman $\rho = 0.52$, the rank correlation between predicted and observed HbA$_{1c}$.

We show the predicted mean and median HbA$_{1c}$ as a function of age, adjusting other variables to their median or mode (Figure 15.9). Compare the estimate of the median and 90th percentile with that from quantile regression.

```
M      ← Mean(f)
qu     ← Quantile(f)
med    ← function(x) qu(.5, x)
p90    ← function(x) qu(.9, x)
fq     ← Rq(formula(f), data=w)
fq90   ← Rq(formula(f), data=w, tau=.9)
```

```
pmean  ← Predict(f,    age, fun=M,   conf.int=FALSE)
pmed   ← Predict(f,    age, fun=med, conf.int=FALSE)
p90    ← Predict(f,    age, fun=p90, conf.int=FALSE)
pmedqr ← Predict(fq,   age, conf.int=FALSE)
p90qr  ← Predict(fq90, age, conf.int=FALSE)
z ← rbind('orm mean'=pmean,'orm median'=pmed,'orm P90'=p90,
          'QR median'=pmedqr, 'QR P90'=p90qr)
ggplot(z, groups='.set.',
       adj.subtitle=FALSE, legend.label=FALSE)
```

```
print(fastbw(f, rule='p'), estimates=FALSE)
```

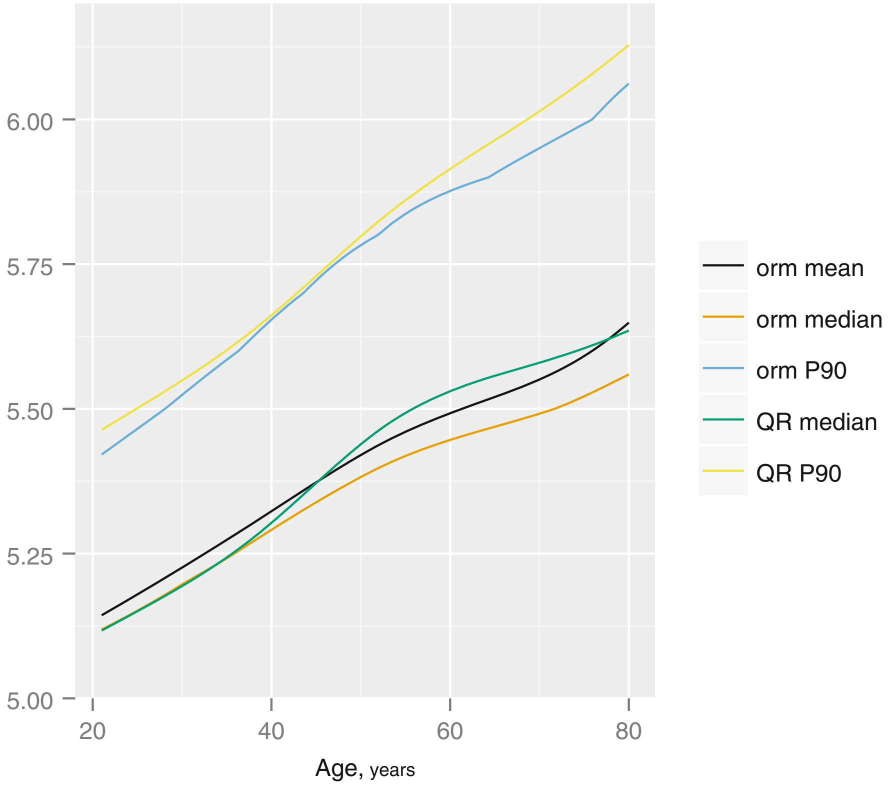


Fig. 15.9 Estimated mean and 0.5 and 0.9 quantiles from the log-log ordinal model using casewise deletion, along with predictions of 0.5 and 0.9 quantiles from quantile regression (QR). Age is varied and other predictors are held constant to medians/-modes.

```
 Deleted Chi-Sq d.f. P      Residual d.f. P      AIC
 arml    0.16   1    0.6924 0.16     1    0.6924 -1.84
 sex     0.45   1    0.5019 0.61     2    0.7381 -3.39
 wt      5.72   2    0.0572 6.33     4    0.1759 -1.67
 armc    3.32   2    0.1897 9.65     6    0.1400 -2.35

Factors in Final Model

[1] age   re    leg   waist tri   sub
```

```
set.seed(13)  # so can reproduce results
v ← validate(f, B=100, bw=TRUE, estimates=FALSE, rule='p')
```

```
                Backwards Step-down - Original Model

 Deleted Chi-Sq d.f. P      Residual d.f. P      AIC
 arml    0.16   1    0.6924 0.16     1    0.6924 -1.84
 sex     0.45   1    0.5019 0.61     2    0.7381 -3.39
 wt      5.72   2    0.0572 6.33     4    0.1759 -1.67
 armc    3.32   2    0.1897 9.65     6    0.1400 -2.35

Factors in Final Model

[1] age   re    leg   waist tri   sub
```

```
# Show number of variables selected in first 30 boots
latex(v, B=30, file='', size='small')
```

Index	Original	Training	Test	Optimism	Corrected	n
	Sample	Sample	Sample		Index	
ρ	0.5225	0.5290	0.5208	0.0083	0.5142	100
R^2	0.2712	0.2788	0.2692	0.0095	0.2617	100
Slope	1.0000	1.0000	0.9761	0.0239	0.9761	100
g	1.2276	1.2505	1.2207	0.0298	1.1978	100
$\lvert \overline{\Pr(Y \geq Y_{0.5}) - \frac{1}{2}} \rvert$	0.2007	0.2050	0.1987	0.0064	0.1943	100

Factors Retained in Backwards Elimination
First 30 Resamples

age	sex	re	wt	leg	arml	armc	waist	tri	sub
●	●	●	●	●		●	●		●
●		●		●		●	●	●	●
●		●		●		●	●	●	●
●	●	●	●	●		●	●		●
●	●	●	●	●		●	●		●
●		●		●		●	●	●	●
●	●	●	●			●	●		●
●	●	●	●	●		●	●		
●		●	●	●		●	●	●	●
●	●	●	●	●		●	●		●
●		●	●			●	●	●	●
●		●		●	●	●	●	●	●
●		●	●	●		●	●	●	●
●		●	●	●		●	●	●	●
●		●		●		●	●	●	●
●		●	●	●		●	●	●	●
●	●	●	●	●		●	●	●	●
●		●		●			●	●	●
●		●	●	●		●	●	●	●
●		●	●	●		●	●	●	●
●		●	●	●		●	●	●	●
●		●	●			●	●	●	●
●		●	●	●		●	●	●	●
●	●	●		●			●		●
●		●		●		●		●	●
●		●	●	●		●	●	●	●
●		●	●			●	●	●	●
●		●	●	●		●	●	●	●
●		●	●	●		●	●	●	●
●		●		●		●		●	●

Frequencies of Numbers of Factors Retained

5	6	7	8	9	10
1	19	29	46	4	1

Next we fit the reduced model, using multiple imputation to impute missing predictors (Figure 15.10).

```
a ← aregImpute(∼ gh + wt + ht + bmi + leg + arml + armc + waist +
               tri + sub + age +re, data=w, n.impute=5, pr=FALSE)
g ← fit.mult.impute(gh ∼ rcs(age,5) + re + rcs(leg,3) +
                    rcs(waist,4) + tri + rcs(sub,4),
                    orm, a, family=loglog, data=w, pr=FALSE)
```

```
print(g, latex=TRUE, needspace='1.5in')
```

-log-log Ordinal Regression Model

```
fit.mult.impute(formula = gh ~ rcs(age, 5) + re + rcs(leg, 3)
    + rcs(waist, 4) + tri + rcs(sub, 4), fitter = orm,
    xtrans = a, data = w, pr = FALSE, family = loglog)
```

	Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs 4629	LR χ^2 1448.42	R^2 0.269	ρ 0.513
Unique Y 63	d.f. 17	g 0.743	
$Y_{0.5}$ 5.5	$\Pr(>\chi^2) < 0.0001$	g_r 2.102	
$\max\left\lvert\frac{\partial \log L}{\partial \beta}\right\rvert$	Score χ^2 1569.21	$\lvert\Pr(Y \geq Y_{0.5}) - \frac{1}{2}\rvert$ 0.173	
1×10^{-5}	$\Pr(>\chi^2) < 0.0001$		

	Coef	S.E.	Wald Z	$\Pr(>\lvert Z\rvert)$
age	0.0404	0.0055	7.29	< 0.0001
age'	-0.0228	0.0279	-0.82	0.4137
age"	0.0126	0.0876	0.14	0.8857
age"'	0.0424	0.1148	0.37	0.7116
re=Other Hispanic	-0.0766	0.0597	-1.28	0.1992
re=Non-Hispanic White	-0.4121	0.0449	-9.17	< 0.0001
re=Non-Hispanic Black	0.0645	0.0566	1.14	0.2543
re=Other Race Including Multi-Racial	-0.0555	0.0750	-0.74	0.4593
leg	-0.0339	0.0091	-3.73	0.0002
leg'	0.0153	0.0105	1.46	0.1434
waist	0.0073	0.0050	1.47	0.1428
waist'	0.0304	0.0158	1.93	0.0536
waist"	-0.0910	0.0508	-1.79	0.0732
tri	-0.0163	0.0026	-6.28	< 0.0001
sub	-0.0027	0.0097	-0.28	0.7817
sub'	0.0674	0.0289	2.33	0.0198
sub"	-0.1895	0.0922	-2.06	0.0398

```
an ← anova(g)
lan(an)
```

	χ^2	d.f.	P
age	692.50	4	< 0.0001
Nonlinear	28.47	3	< 0.0001
re	168.91	4	< 0.0001
leg	24.37	2	< 0.0001
Nonlinear	2.14	1	0.1434
waist	128.31	3	< 0.0001
Nonlinear	4.05	2	0.1318
tri	39.44	1	< 0.0001
sub	39.30	3	< 0.0001
Nonlinear	6.63	2	0.0363
TOTAL NONLINEAR	46.80	8	< 0.0001
TOTAL	1464.24	17	< 0.0001

```
b  ← anova(g, leg, waist, tri, sub)
# Add new lines to the plot with combined effect of 4 size var.
s ← rbind(an, size=b['TOTAL', ])
class(s) ← 'anova.rms'
plot(s)
```

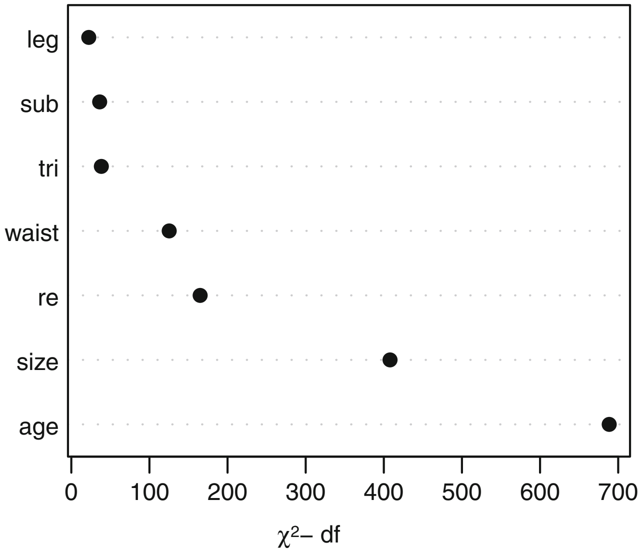


Fig. 15.10 ANOVA for reduced model, after multiple imputation, with addition of a combined effect for four size variables

```
ggplot(Predict(g), abbrev=TRUE, ylab=NULL)   # Figure 15.11
```

Compare the estimated age partial effects and confidence intervals with those from a model using casewise deletion, and with bootstrap nonparametric confidence intervals (also with casewise deletion).

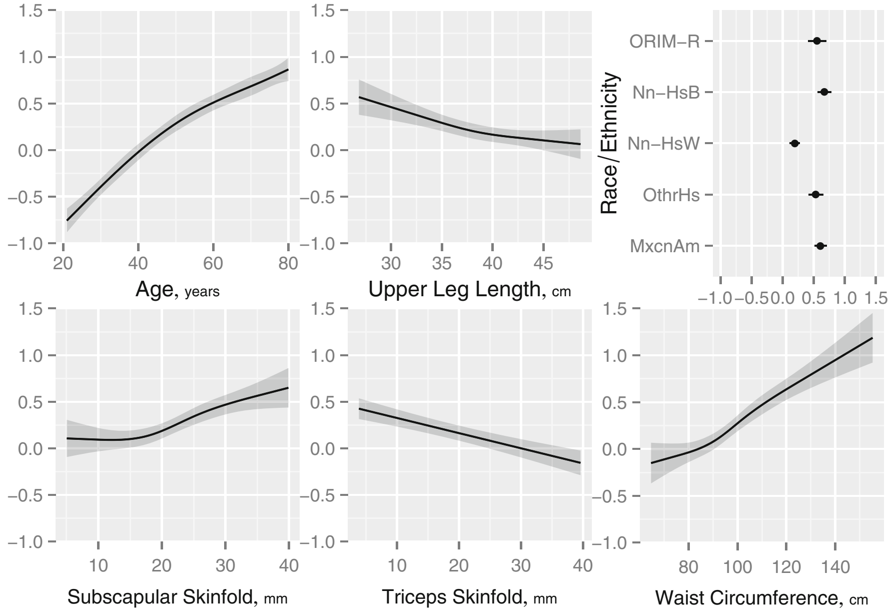


Fig. 15.11 Partial effects (log hazard or log-log cumulative probability scale) of all predictors in reduced model, after multiple imputation

```
gc ← orm(gh ∼ rcs(age,5) + re + rcs(leg,3) +
           rcs(waist,4) + tri + rcs(sub,4),
           family=loglog, data=w, x=TRUE, y=TRUE)
gb ← bootcov(gc, B=300)
```

```
bootclb ← Predict(gb, age, boot.type='basic')
bootclp ← Predict(gb, age, boot.type='percentile')
multimp ← Predict(g,  age)
plot(Predict(gc, age), addpanel=function(...) {
  with(bootclb, {llines(age, lower, col='blue')
                 llines(age, upper, col='blue')})
  with(bootclp, {llines(age, lower, col='blue', lty=2)
                 llines(age, upper, col='blue', lty=2)})
  with(multimp, {llines(age, lower, col='red')
                 llines(age, upper, col='red')
                 llines(age, yhat, col='red')} ) },
    col.fill=gray(.9), adj.subtitle=FALSE)   # Figure 15.12
```

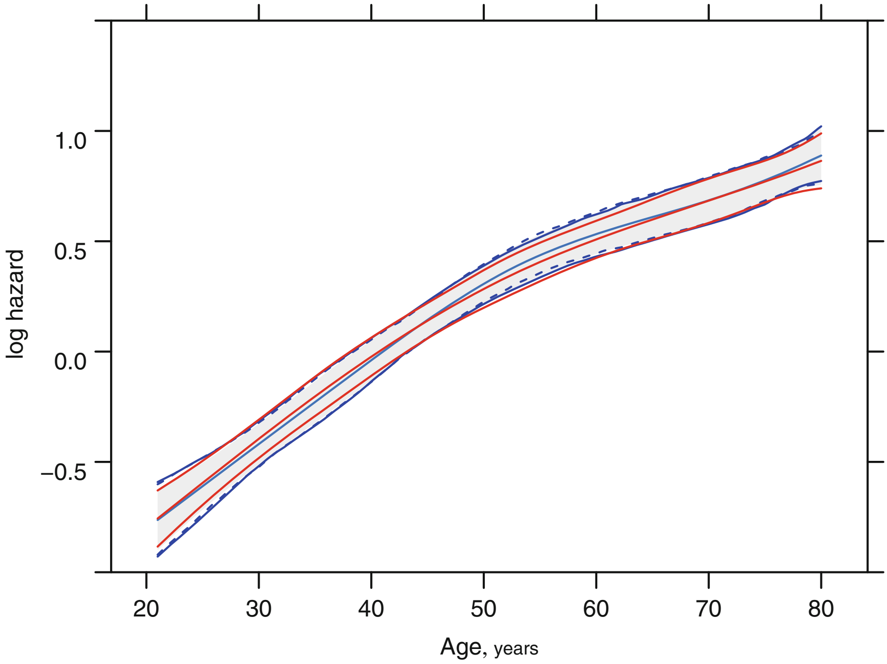


Fig. 15.12 Partial effect for age from multiple imputation (center red line) and casewise deletion (center blue line) with symmetric Wald 0.95 confidence bands using casewise deletion (gray shaded area), basic bootstrap confidence bands using casewise deletion (blue lines), percentile bootstrap confidence bands using casewise deletion (dashed blue lines), and symmetric Wald confidence bands accounting for multiple imputation (red lines).

Figure 15.13 depicts the relationship between various predicted quantities, demonstrating that the ordinal model makes fewer model assumptions that dictate their connections. A Gaussian or log-Gaussian model would have a straight-line relationship between the predicted mean and median.

```
M   ← Mean(g)
qu  ← Quantile(g)
med  ← function(lp) qu(.5, lp)
q90  ← function(lp) qu(.9, lp)
lp   ← predict(g)
lpr  ← quantile(predict(g), c(.002, .998), na.rm=TRUE)
lps  ← seq(lpr[1], lpr[2], length=200)
pmn  ← M(lps)
pme  ← med(lps)
p90  ← q90(lps)
plot(pmn, pme,    # Figure 15.13
     xlab=expression(paste('Predicted Mean ',  HbA["1c"])),
     ylab='Median and 0.9 Quantile', type='l',
     xlim=c(4.75, 8.0), ylim=c(4.75, 8.0), bty='n')
box(col=gray(.8))
```

```
lines(pmn, p90, col='blue')
abline(a=0, b=1, col=gray(.8))
text(6.5, 5.5, 'Median')
text(5.5, 6.3, '0.9', col='blue')
nint ← 350
scat1d(M(lp),   nint=nint)
scat1d(med(lp), side=2, nint=nint)
scat1d(q90(lp), side=4, col='blue', nint=nint)
```

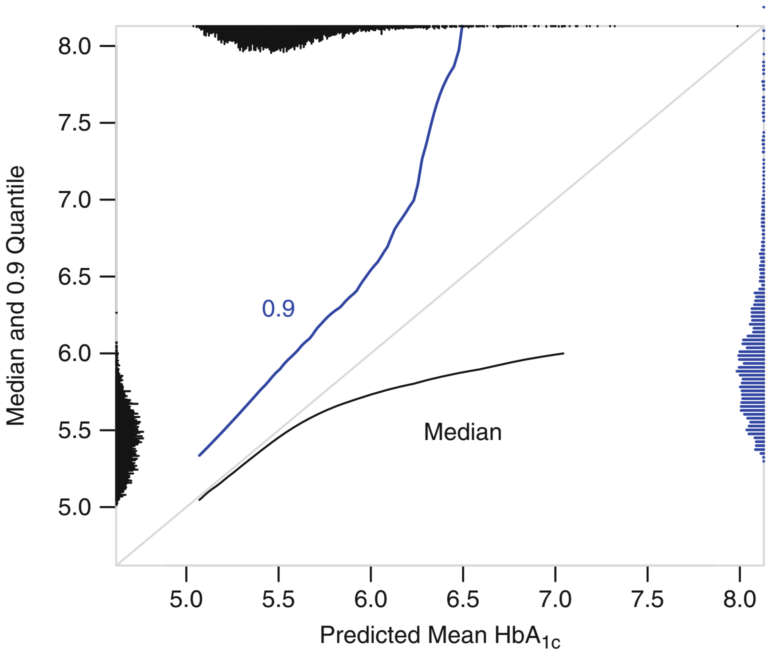


Fig. 15.13 Predicted mean HbA$_{1c}$vs. predicted median and 0.9 quantile along with their marginal distributions

Finally, let us draw a nomogram that shows the full power of ordinal models, by predicting five quantities of interest.

```
g      ← Newlevels(g, list(re=abbreviate(levels(w$re))))
exprob ← ExProb(g)
nom ←
  nomogram(g, fun=list(Mean=M,
                 'Median Glycohemoglobin' = med,
                 '0.9 Quantile'           = q90,
                 'Prob(HbA1c ≥ 6.5)'=
                      function(x) exprob(x, y=6.5),
                 'Prob(HbA1c ≥ 7.0)'=
                      function(x) exprob(x, y=7),
                 'Prob(HbA1c ≥ 7.5)'=
                     function(x) exprob(x, y=7.5)),
           fun.at=list(seq(5, 8, by=.5),
             c(5,5.25,5.5,5.75,6,6.25),
             c(5.5,6,6.5,7,8,10,12,14),
             c(.01,.05,.1,.2,.3,.4),
             c(.01,.05,.1,.2,.3,.4),
             c(.01,.05,.1,.2,.3,.4)))
plot(nom, lmgp=.28)    # Figure 15.14
```

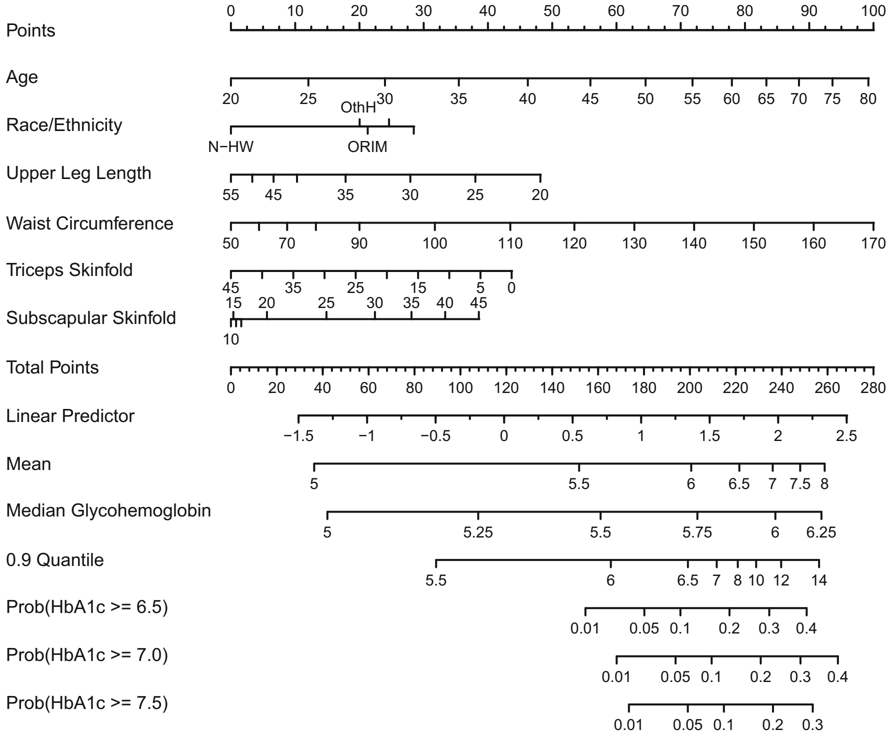


Fig. 15.14 Nomogram for predicting median, mean, and 0.9 quantile of glycohemoglobin, along with the estimated probability that HbA$_{1c} \geq 6.5, 7$, or 7.5, all from the log-log ordinal model