The Connection between Cumulative Link Models and Linear Models

Rune Haubo B Christensen

April 29, 2011

Abstract

Pending...

1 Introduction

A cumulative link model can be motivated by linear model for a latent variable. Suppose a latent variable S_i follows the linear model:

$$S_i = \alpha + \mu_i + \varepsilon_i, \quad \varepsilon \sim N(0, \sigma^2)$$

or equivalently: $S_i \sim N(\alpha + \mu_i, \sigma^2)$. The observed variable, Y_i is then generated as a coarsened version of S_i in J groups where $Y_i = j$ is observed if $\theta_j < S_i \leq \theta_{j-1}$ and $\{\theta_j\}$ for $j = 0, \ldots, J$ are strictly increasing with $\theta_0 \equiv -\infty$ and $\theta_J \equiv \infty$. The cumulative probability of an observation falling in category j or below is then:

$$\gamma_{ij} = P(Y_i \leq j) = P(S_i \leq \theta_j) = P\left(Z_i \leq \frac{\theta_j - \alpha - \mu_i}{\sigma}\right) = \Phi\left(\frac{\theta_j - \alpha - \mu_i}{\sigma}\right) \tag{1}$$

where $Z_i = (S_i - \alpha - \mu_i)/\sigma \sim N(0, 1)$ and Φ is the standard normal CDF.

Since the absolute location and scale of the latent variable, α and σ respectively, are not identifiable from ordinal observations, an identifiable model is

$$\gamma_{ij} = \Phi(\theta_j^* - \mu_i^*) \tag{2}$$

with identifiable parameter functions: $\theta_j^* = (\theta_j - \alpha)/\sigma$ and $\mu_i^* = \mu_i/\sigma$. The latter can therefore be thought of as signal-to-noise ratios.

Thus, from a CLM we obtain estimates of θ_j^* and μ_i^* while if S_i were available we would from a linear model obtain estimates of α, μ_i and σ. If Y_i is just a coarsened version of S_i, we could attempt to model S_i directly and assume that the coarsening does not distort or remove too much information. Indeed in many cases ordinal data is modeled using linear models, so the connection between the two approaches is interesting.

If a model for S_i is fitted and the following estimates obtained: $\hat{\alpha}^{lm}$, $\hat{\mu}_i^{lm}$ and $\hat{\sigma}^{lm}$, then the corresponding relative estimates are μ_i^{*lm}/σ^{lm}. Estimates of θ_j can be obtained by the following: Let n_j denote the frequencies of observations in the response categories and n

the total number of observations. Then $p_j = n_j/n$ are the observed proportions and g_j are the cumulative proportions. An overall estimate of θ_j is then given as $\theta_j = \Phi^{-1}(g_j)$ for $j = 1, \ldots, J$. If μ_i represents, say two temperature levels, then, for identifiability, we may take the reference level $\mu_1 = 0$ such that the mean of S_i is α for the first level of temperature and $\alpha + \mu_2$ for the second level of temperature. We should then estimate the cut-points as $\theta_j^{*lm} = (\Phi^{-1}(g_j)\sigma^{\mathrm{lm}_0} + \mu_2^{lm}/2)/\sigma^{\mathrm{lm}}$ where $\mu_2^{*lm} = \mu_2^{lm}/\sigma^{lm}$. This corresponds to estimating $\{\theta_j^{*lm}\}$ by pluging in g_j for γ_j and μ^{*lm} in (2), i.e., estimating θ_j^{*lm} from:

$$g_j = \Phi(\theta_j^{*lm} - \mu_i^{*lm})$$

Naturally we can also find μ_i from a CLM if α and σ are available.

To see how the plug-in estimator works, consider a simple model for the wine data presented by Randall (1989). We will consider how ratings depend on a temperature variable:

```
> S <- as.numeric(wine$rating)
> temp <- wine$temp
> (freq <- table(temp, S))

      S
temp    1  2  3  4  5
  cold  5 16 13  2  0
  warm  0  6 13 10  7
```

A CLM gives the following estimates and predictions:

```
> clm.temp <- clm(rating ~ temp, data = wine, link = "probit")
> coef(clm.temp)

       1|2        2|3        3|4        4|5   tempwarm
-1.1087829  0.2831333  1.4651529  2.2889608  1.3722905

> gamma.cold <- c(0, pnorm(clm.temp$alpha), 1)
> p.cold <- diff(gamma.cold)
> freq.cold <- p.cold * 36
> round(freq.cold, 2)

  1|2   2|3   3|4   4|5
 4.82 17.20 11.42  2.17  0.40

> gamma.warm <- c(0, pnorm(clm.temp$alpha - clm.temp$beta), 1)
> p.warm <- diff(gamma.warm)
> freq.warm <- p.warm * 36
> round(freq.warm, 2)

  1|2   2|3   3|4   4|5
 0.24  4.73 14.36 10.20  6.47
```

With a linear model we get the following estimate of α, the coefficient for temperature and σ:

```
> lm1 <- lm(as.numeric(rating) ~ temp, data = wine)
> coef(lm1)

(Intercept)    tempwarm
   2.333333    1.166667

> (sd.S <- summary(lm1)$sigma)
```

```
[1] 0.9023778
```

The relative coefficient, $\hat{\mu}^*$ is therefore

```
> coef(lm1)[2]/sd.S
```

```
tempwarm
1.292880
```

The plug-in estimates of the cut-points are

```
> g.j <- cumsum(colSums(freq)/sum(freq))
> (theta <- (qnorm(g.j[-5]) * sd(S) + coef(lm1)[2]/2)/sd.S)
```

```
         1          2          3          4
-1.1106972  0.2681191  1.3961050  2.1870156
```

This plug-in estimate can be used to provide starting values for a Newton-Raphson estimation of a CLM and will generally save an iteration compared to simpler alternatives.

Questions:

- How does the plug-in estimator depend on the numbers attached to the categories? Can they be shifted or scaled?
-

From (1) we see that if σ descrease, the regression parameter estimates increase: if noise is reduced and the signal unchanged, the signal-to-noise ratio is higher. This highlights an important difference between a CLM and the LM for the underlying variable.

Using the wine data for illustation: If we add the variable `contact` to the LM, the estimate for `temp` remain unchanged (balanced, orthogonal design), but σ decreases, since variation is removed from the noise term:

```
> lm2 <- lm(as.numeric(rating) ~ contact + temp, data = wine)
> coef(lm2)
```

```
(Intercept)  contactyes    tempwarm
  2.0000000   0.6666667   1.1666667
```

```
> (sd.S2 <- summary(lm2)$sigma)
```

```
[1] 0.842701
```

According to (1) we should therefore expect the coefficient for `temp` to increase to $\mu^*_{(1)}\sigma_{(1)}/\sigma_{(2)} \approx \mu^*_{(2)}$ in a corresponding CLM:

```
> clm.temp$beta * sd.S/sd.S2
```

```
tempwarm
1.469471
```

```
> clm2 <- clm(rating ~ contact + temp, data = wine, link = "probit")
> coef(clm2)
```

```
       1|2        2|3        3|4        4|5 contactyes   tempwarm
-0.7732627  0.7360215  2.0446805  2.9413450  0.8677435  1.4993746
```

References

Randall, J. (1989). The analysis of sensory data by generalised linear model. *Biometrical journal 7*, pp. 781–793.