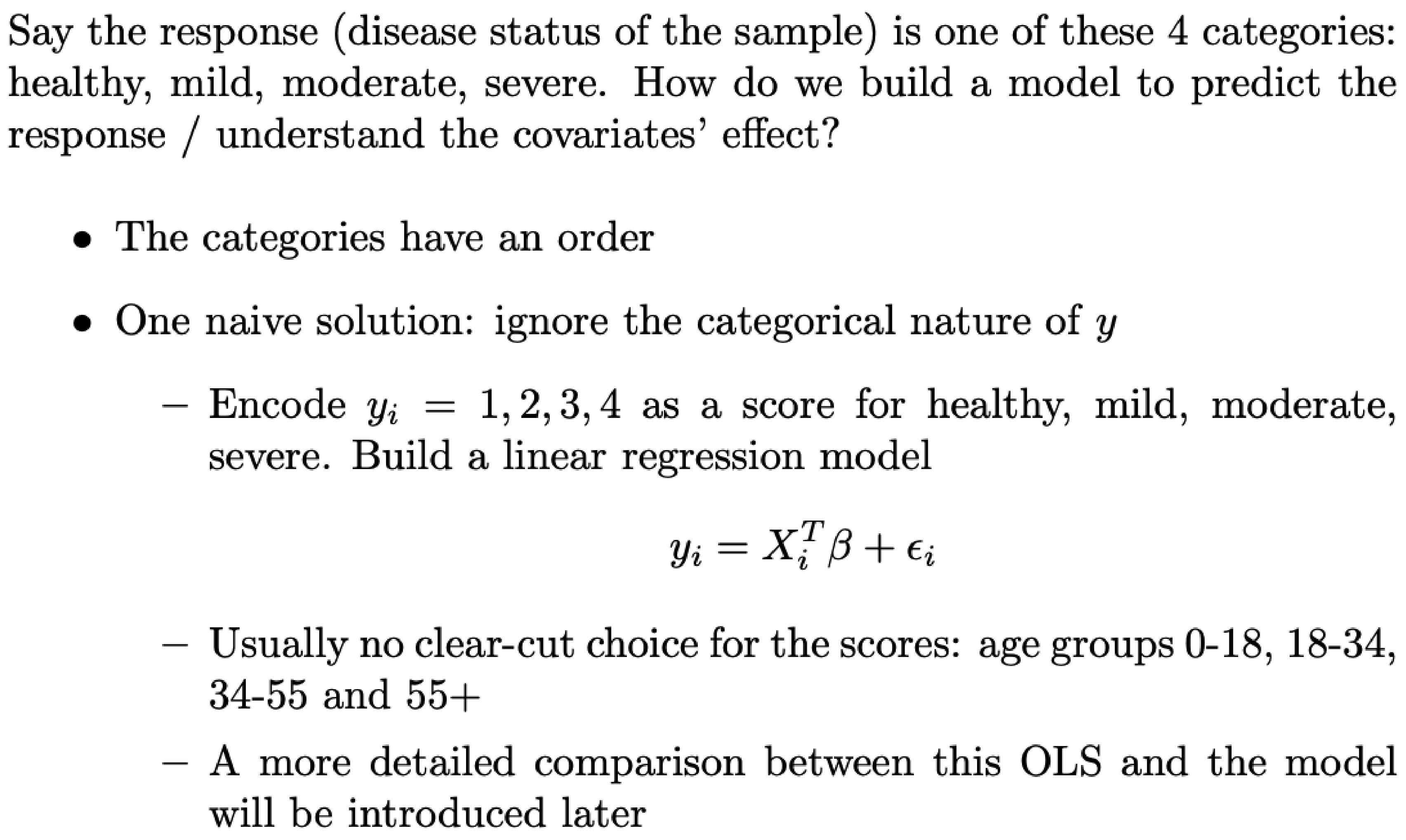
Lecture 9
GLM for ordinal responses

Today’s topics:

- Ordinal response models
- Examples of multinomial GLM

Ordinal response



Say the response (disease status of the sample) is one of these 4 categories: healthy, mild, moderate, severe. How do we build a model to predict the response / understand the covariates' effect?

- The categories have an order
- One naive solution: ignore the categorical nature of y
  - Encode $y_i = 1, 2, 3, 4$ as a score for healthy, mild, moderate, severe. Build a linear regression model

    $$y_i = X_i^T \beta + \epsilon_i$$

  - Usually no clear-cut choice for the scores: age groups 0-18, 18-34, 34-55 and 55+
  - A more detailed comparison between this OLS and the model will be introduced later

A latent variable motivation to model ordinal response

- Assume that there a continuous latent variable y_i^* that satisfy

$$y_i^* = X_i^T \beta + \epsilon_i$$

  where ϵ_i are i.i.d. with cdf function $F(\cdot)$
- Assume that the observed response satisfy

$$y_i = k \quad \text{if } \alpha_{k-1} < y_i^* \leq \alpha_k$$

  where $-\infty = \alpha_0 \leq \alpha_1 \leq ... \leq \alpha_c = \infty$ are cutoff points

- Then we have

$$P(y_i \leq k) = P(y_i^* \leq \alpha_k) = F(\alpha_k - X_i^T \beta)$$

A latent variable motivation to model ordinal response

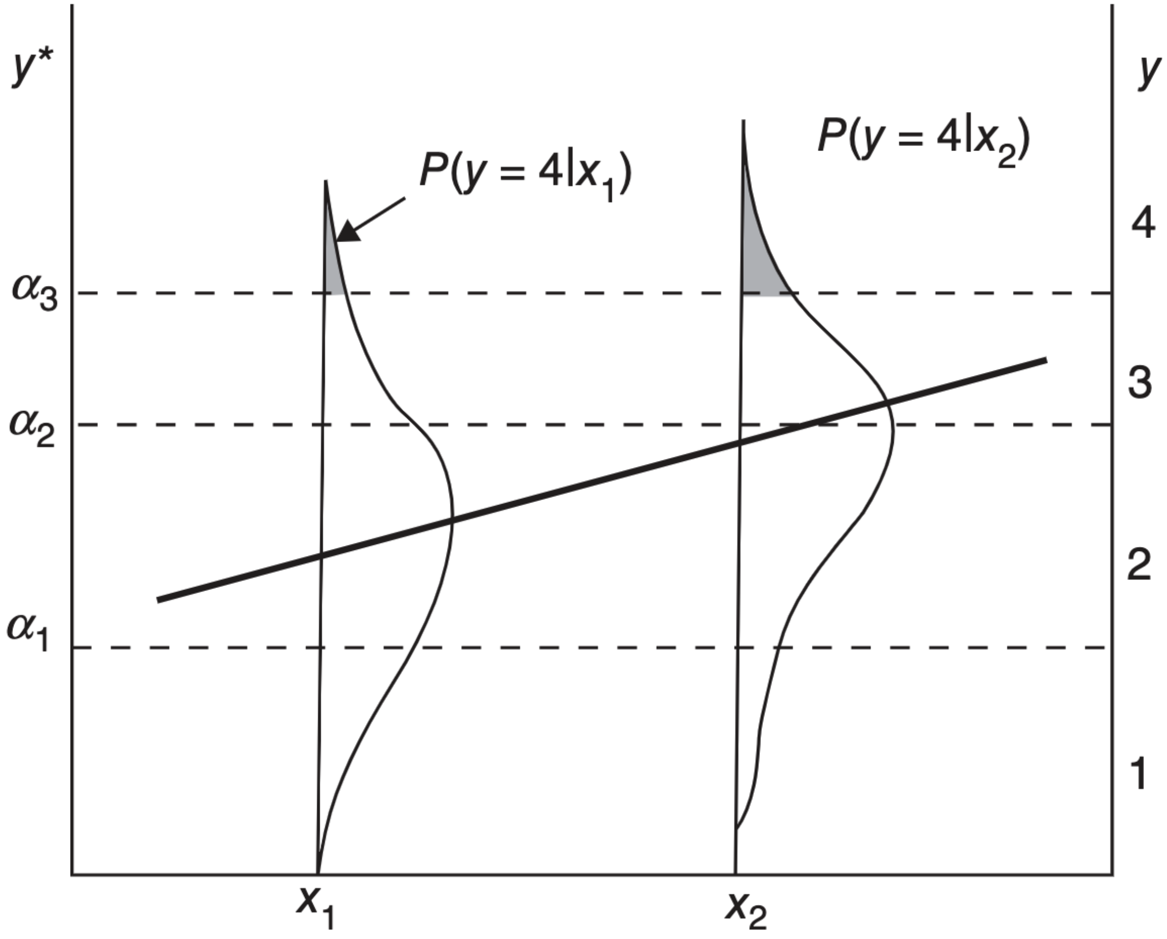


Figure 6.2 Ordinal measurement and underlying linear model for a latent variable.

Cumulative logit/probit models

$$P(y_i \leq k) = P(y_i^* \leq \alpha_k) = F(\alpha_k - X_i^T \beta)$$

- Take $F(\cdot)$ to be the cdf of a standard logistic/Gaussian distribution, we get the cumulative logit/probit models
- For identifiability, X_i here does not include the intercept term
- We assume constant β across categories

- Another equivalent way to define the cumulative logit model

$$\text{logit}[\mathbb{P}(y_i \leq k)] = \log \frac{p_{i1} + \cdots + p_{ik}}{p_{i,k+1} + \cdots + p_{ic}} = \alpha_k + X_i^T \tilde{\beta}$$

where $\tilde{\beta} = -\beta$.

Proportional odds

$$\begin{aligned}
&\operatorname{logit}[\mathbb{P}(y_i \leq k|X_i = u)] - \operatorname{logit}[\mathbb{P}(y_i \leq k|X_i = v)] \\
=&\log \frac{\mathbb{P}(y_i \leq k|X_i = u)/\mathbb{P}(y_i > k|X_i = u)}{\mathbb{P}(y_i \leq k|X_i = v)/\mathbb{P}(y_i > k|X_i = v)} \\
=&(u-v)^T \tilde{\beta}
\end{aligned}$$

So the cumulative odds ratio between two samples keeps the same for all k.

- The cumulative odds ratio is proportional to the distance between u and v

- Settings are stochastically ordered on the response
  If $X_i^T \tilde{\beta} \geq X_{i'}^T \tilde{\beta}$ then we have $P(y_i \leq k) \geq P(y_{i'} \leq k)$ for ALL k.

Proportional odds

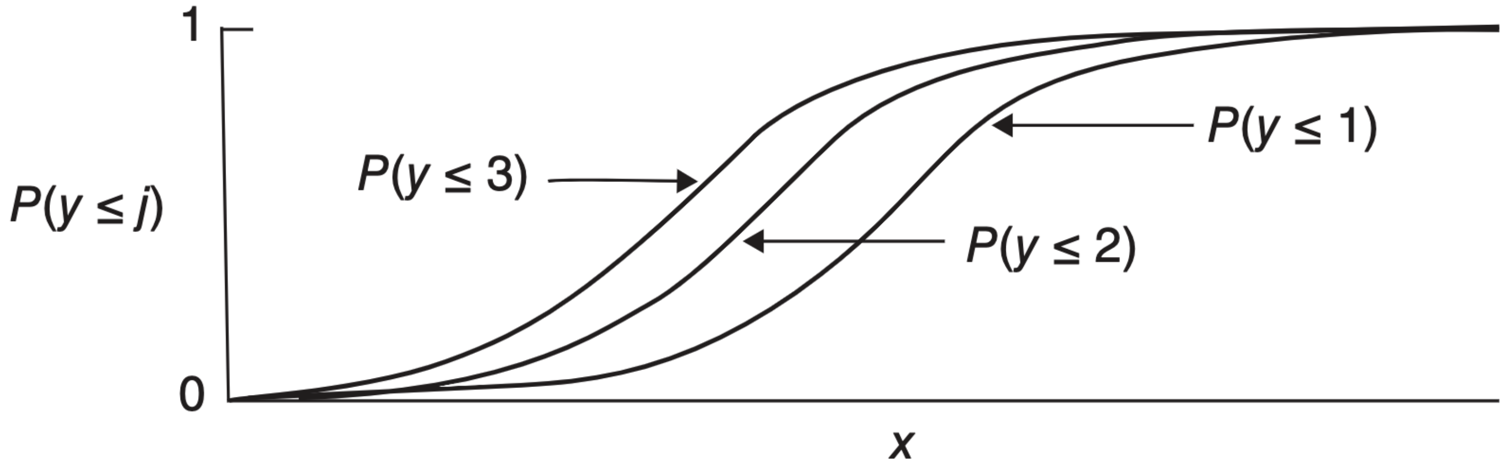
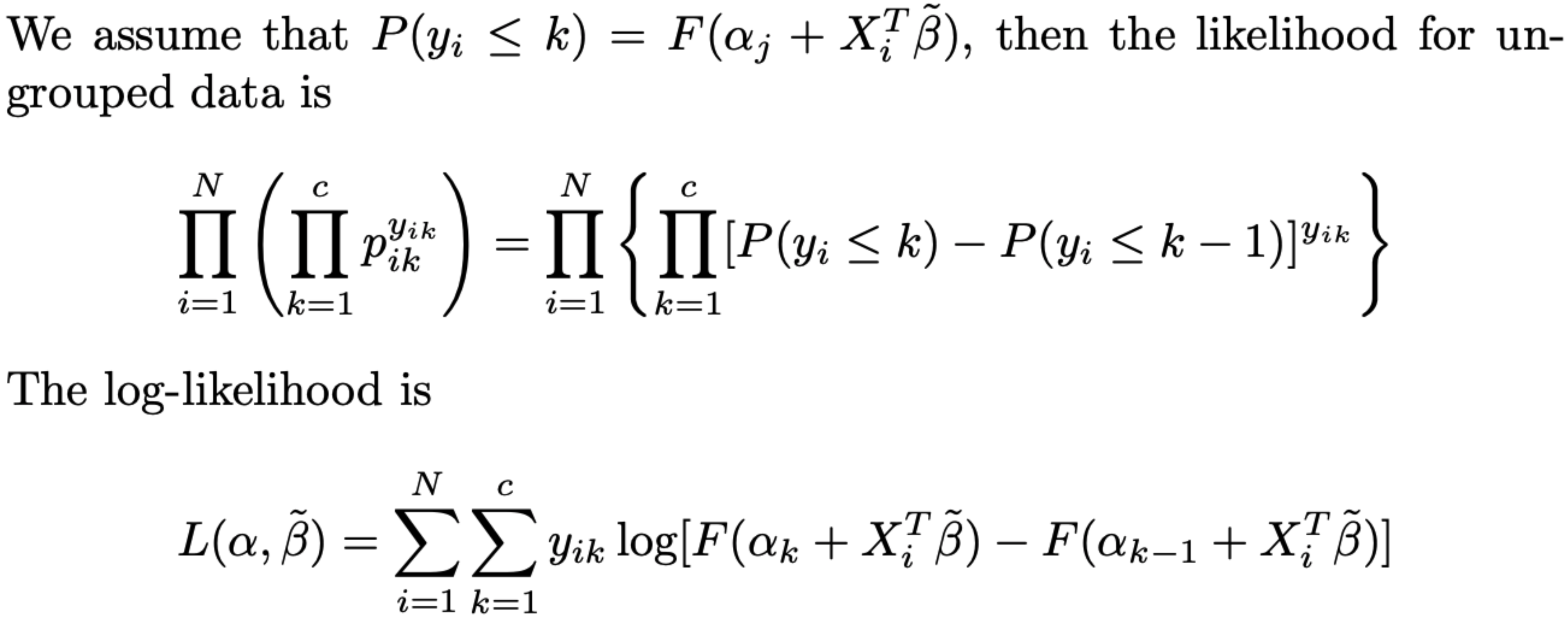


Figure 6.1 Cumulative logit model with the same effect of x on each of three cumulative probabilities, for an ordinal response variable with $c = 4$ categories.

Fitting cumulative link model



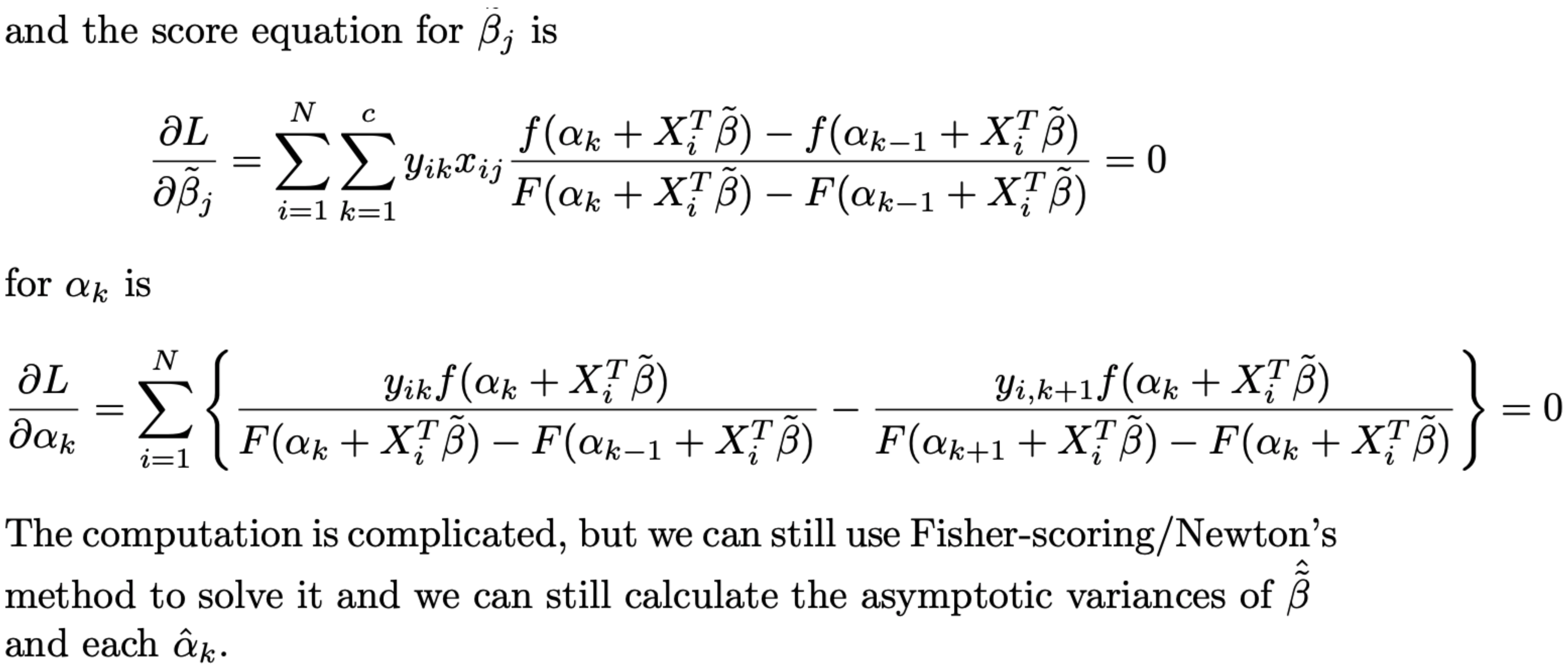
We assume that $P(y_i \leq k) = F(\alpha_j + X_i^T\tilde{\beta})$, then the likelihood for ungrouped data is

$$\prod_{i=1}^{N}\left(\prod_{k=1}^{c} p_{ik}^{y_{ik}}\right) = \prod_{i=1}^{N}\left\{\prod_{k=1}^{c}[P(y_i \leq k) - P(y_i \leq k-1)]^{y_{ik}}\right\}$$

The log-likelihood is

$$L(\alpha, \tilde{\beta}) = \sum_{i=1}^{N}\sum_{k=1}^{c} y_{ik}\log[F(\alpha_k + X_i^T\tilde{\beta}) - F(\alpha_{k-1} + X_i^T\tilde{\beta})]$$

Fitting cumulative link model



and the score equation for $\tilde{\beta}_j$ is

$$\frac{\partial L}{\partial \tilde{\beta}_j} = \sum_{i=1}^{N} \sum_{k=1}^{c} y_{ik} x_{ij} \frac{f(\alpha_k + X_i^T \tilde{\beta}) - f(\alpha_{k-1} + X_i^T \tilde{\beta})}{F(\alpha_k + X_i^T \tilde{\beta}) - F(\alpha_{k-1} + X_i^T \tilde{\beta})} = 0$$

for α_k is

$$\frac{\partial L}{\partial \alpha_k} = \sum_{i=1}^{N} \left\{ \frac{y_{ik} f(\alpha_k + X_i^T \tilde{\beta})}{F(\alpha_k + X_i^T \tilde{\beta}) - F(\alpha_{k-1} + X_i^T \tilde{\beta})} - \frac{y_{i,k+1} f(\alpha_k + X_i^T \tilde{\beta})}{F(\alpha_{k+1} + X_i^T \tilde{\beta}) - F(\alpha_k + X_i^T \tilde{\beta})} \right\} = 0$$

The computation is complicated, but we can still use Fisher-scoring/Newton's method to solve it and we can still calculate the asymptotic variances of $\hat{\tilde{\beta}}$ and each $\hat{\alpha}_k$.

Limitation of the cumulative link models

- Settings are stochastically ordered:
  If $X_i^T\tilde{\beta} \geq X_{i'}^T\tilde{\beta}$ then we have $P(y_i \leq k) \geq P(y_{i'} \leq k)$ for ALL k.

- When $c = 4$, the model can not allow the probability of each ordered category to be $(0.3, 0.2, 0.2, 0.3)$ for one sample and $(0.1, 0.4, 0.4, 0.1)$ for the other sample.

- More flexible model: replace $\tilde{\beta}$ with $\tilde{\beta}_k$

$$\text{logit}[\Pr(y_i \leq k)] = \log \frac{p_{i1} + \cdots + p_{ik}}{p_{i,k+1} + \cdots + p_{ic}} = \alpha_k + X_i^T\tilde{\beta}_k$$

  - We can perform likelihood ratio test to check if the more flexible model is necessary

Comparison with OLS

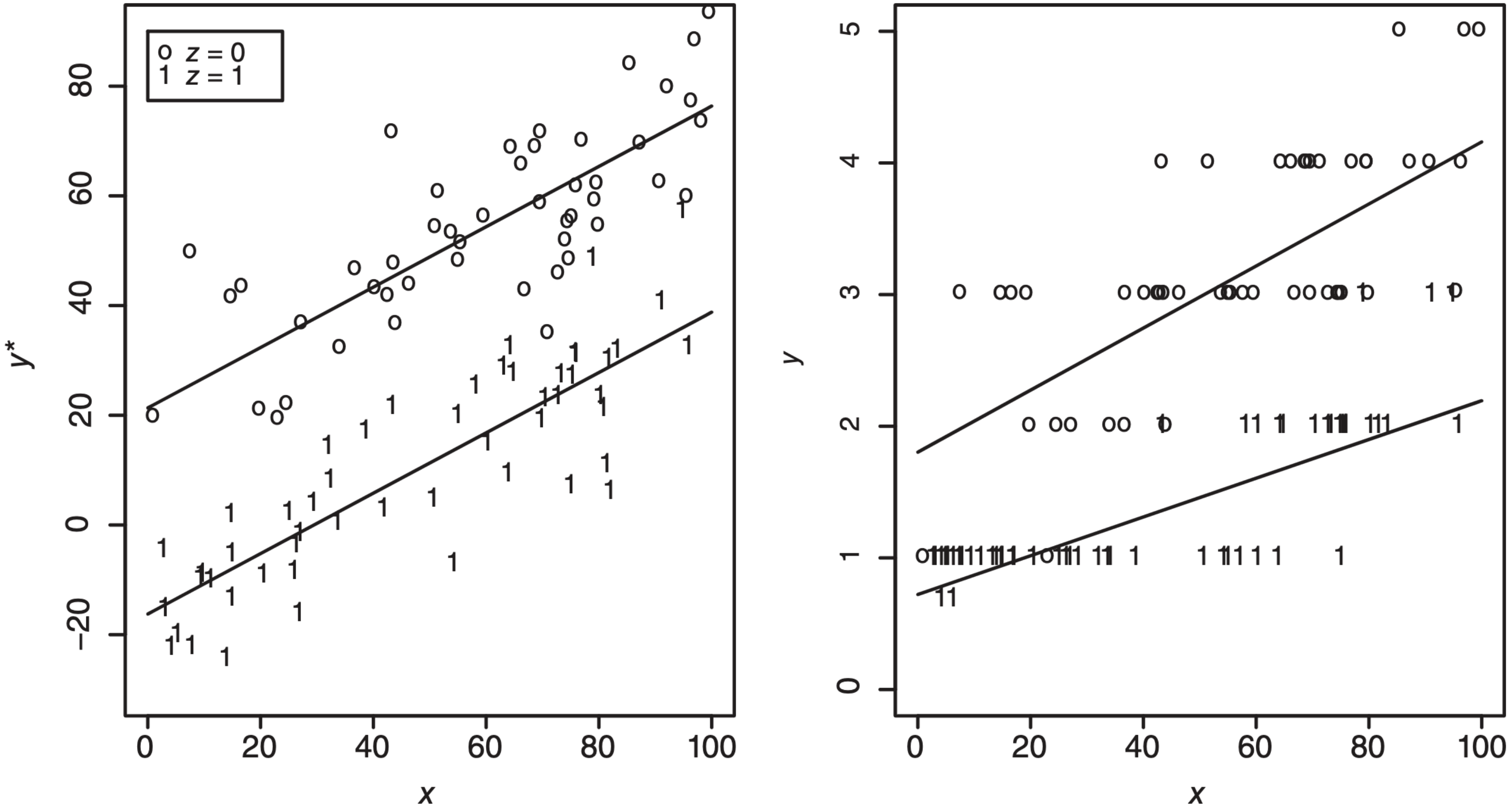
Disadvantages of modeling ordered categories using a linear model:

- Usually no clear cut for the numerical scores
- Linear model does not allow for the measurement error in discretization
- From the linear model you can not get estimated probabilities of each category for a particular sample
- Linear model ignores that the variability in each category can be different

A simulation example

$$y_i^* = 20 + 0.6x_i - 40z_i + \epsilon_i$$

where $x_i \overset{i.i.d.}{\sim}$ Uniform$[0, 100]$, $z_i \overset{i.i.d.}{\sim}$ Bernoulli(0.5) and $\epsilon_i \overset{i.i.d.}{\sim} N(0, 100)$.
Set $\alpha_1 = 20$, $\alpha_2 = 40$, $\alpha_3 = 60$ and $\alpha_4 = 80$.



R data example for ordinal response

- Check Example 4_2 R notebook