STATISTICS 538, Lecture #9

Ordinal Regression Models

November 22, 2010

Order suggests underlying continuous latent variable

Response $Y_i \in \{1, 2, \ldots, J\}$

e.g., strongly disagree, disagree, neutral, agree, strongly agree. Likert scale

e.g., no recovery, partial recovery, full recovery

For *thresholds* $\underbrace{\theta_1, \ldots, \theta_{J-1}}$, think:

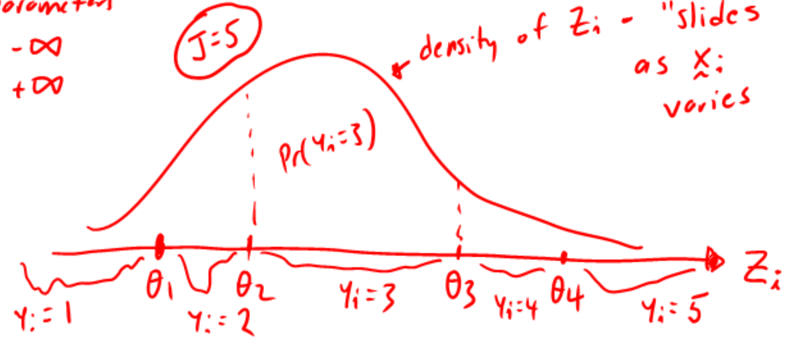
unknown parameters

$$Y_i = j \quad \leftrightarrow \quad \theta_{j-1} < Z_i < \theta_j$$

$\theta_0 = -\infty$

$\theta_J = +\infty$

J=5



Then model continuous variable with linear model

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

$$\epsilon_i \sim F_0$$

cumulative dist function of the error term

So $Pr(Y_i = j) = Pr(\theta_{j-1} < Z_i < \theta_j)$

$= Pr(\theta_{j-1} - \beta^T X_i < \underbrace{Z_i - \beta^T X_i}_{\varepsilon_i} < \theta_j - \beta^T X_i)$

$= F_0(\theta_j - \beta^T X) - F_0(\theta_{j-1} - \beta^T X)$

- Note: no intercept, no unknown scale parameter in F_0.
- Choice of F_0 like choice of link function.

  $F_0 =$ Normal may not be most interpretable!

since $\underset{\sim}{\theta}$ describes the dist of Y when predictors are zero

scale addressed via θ's being concentrated or spread out

Logistic dist: $F_0(s) = 1/\{1 + \exp(-s)\}$ cdf ✓

$$\text{Odds}(Y > j \mid \underset{\sim}{x}) = \frac{\Pr(Z > \theta_j \mid \underset{\sim}{x})}{\Pr(Z \le \theta_j \mid \underset{\sim}{x})} = \frac{\Pr(\varepsilon > \theta_j - \underset{\sim}{\beta}^T \underset{\sim}{x})}{\Pr(\varepsilon \le \theta_j - \underset{\sim}{\beta}^T \underset{\sim}{x})}$$

$$= \frac{1 - F_0(\theta_j - \underset{\sim}{\beta}^T \underset{\sim}{x})}{F_0(\theta_j - \underset{\sim}{\beta}^T \underset{\sim}{x})} = e^{-[\theta_j - \underset{\sim}{\beta}^T \underset{\sim}{x}]}$$

$$\frac{\text{Odds}(Y > j \mid \underset{\sim}{x} = \underset{\sim}{a})}{\text{Odds}(Y > j \mid \underset{\sim}{x} = \underset{\sim}{b})} = e^{\underset{\sim}{\beta}^T (\underset{\sim}{a} - \underset{\sim}{b})} \leftarrow \boxed{\text{same for each } j}$$

familiar logistic-regression flavour

"PROPORTIONAL ODDS MODEL"

nice interpretation

Extreme value dist: $F_0(s) = 1 - \exp(-\exp(s))$ cdf ✓

$$\Pr(Y > j \mid \underset{\sim}{x}) = \Pr(Z > \theta_j \mid \underset{\sim}{x})$$

$$= 1 - F_0(\theta_j - \underset{\sim}{\beta}^T \underset{\sim}{x})$$

$$= e^{-e^{\theta_j - \underset{\sim}{\beta}^T \underset{\sim}{x}}}$$

$$= \left\{ e^{-e^{\theta_j}} \right\}^{e^{-\underset{\sim}{\beta}^T \underset{\sim}{x}}}$$

$$= \left\{ \Pr(Y > j \mid \underset{\sim}{x} = \underset{\sim}{0}) \right\}^{e^{-\underset{\sim}{\beta}^T \underset{\sim}{x}}}$$

"proportional hazards model"

Copenhagen Housing Conditions Survey

of 1681 renters

```
> library(MASS); help(housing)
```

y

Sat: Satisfaction of householders with their present housing circumstances, (High, Medium or Low).

J=3

x

Infl: Perceived degree of influence householders have on the management of the property (High, Medium, Low).

Type: Type of rental accommodation, (Tower, Atrium, Apartment, Terrace).

Cont: Contact residents are afforded with other residents, (Low, High).

Freq: the numbers of residents in each class.

Copenhagen Housing Conditions Survey

```
> housing
       Sat   Infl      Type Cont Freq
1      Low    Low     Tower  Low   21
2   Medium    Low     Tower  Low   21
3     High    Low     Tower  Low   28
4      Low Medium     Tower  Low   34
5   Medium Medium     Tower  Low   22
6     High Medium     Tower  Low   36
...
70     Low   High   Terrace High    5
71  Medium   High   Terrace High    6
72    High   High   Terrace High   13
```

y

x

multinomial response with n=70 OR

70 n=1 responses

Fit using polr(), default 'link' is logistic

prop. odds model

y x

```
ft1 <- polr(Sat ~ Infl + Type + Cont,
            weights = Freq, data = housing)

### or equivalently

housing.new <- housing[rep(1:72, times=housing$Freq), -5]

ft2 <- polr(Sat ~ Infl + Type + Cont, data=housing.new)
```

data frame with 1681 rows

SAT INFL TYPE CONT

1681

summary(ft2)

```
Coefficients:
                    Value Std. Error   t value
InflMedium      0.5663924  0.1046528  5.412110
InflHigh        1.2888218  0.1271561 10.135741
TypeApartment  -0.5723552  0.1192380 -4.800107
TypeAtrium     -0.3661912  0.1551733 -2.359885
TypeTerrace    -1.0910195  0.1514860 -7.202113
ContHigh        0.3602834  0.0955358  3.771187

Intercepts:
            Value   Std. Error t value
Low|Medium  -0.4961  0.1248     -3.9739
Medium|High  0.6907  0.1255      5.5049

Residual Deviance: 3479.149
AIC: 3495.149
```

$\hat{\beta}$

i.e SAT and INFL positively associated given TYPE and CONT

$\hat{\theta}_1$ $\hat{\theta}_2$

would need $2 \times 7 = 14$ parameters for analogous nominal model

$+ 2\{6+2\}$

Or try proportional hazards:
ft3 ← polr(Sat ∼ ..., method="cloglog")

"prop. hazards"

```
> predict(ft2,
  newdata=expand.grid(Infl=c("Low","Medium","High"),
           Type="Tower", Cont="Low"), type="p")
```

SAT

INFL

	Low	Medium	High
1	0.3784494	0.2876742	0.3338765
2	0.2568267	0.2742112	0.4689620
3	0.1436921	0.2110823	0.6452255

fitted dist. of (SAT | INFL, Type=tower, Cont=low)

changing link function has some effect

```
> predict(ft3,...)
```

	Low	Medium	High
1	0.3996213	0.2641956	0.3361831
2	0.2660860	0.2874496	0.4464643
3	0.1324311	0.2728674	0.5947016