Day 5: GLMs for Ordinal and Categorical Data

Recap

We split regression models into three main parts:

- Random component: outcome variable Y takes some probability distribution, with $E[Y] = \mu$
- Systematic component: Explanatory variables $X_1, \ldots, X_p$ affect the response through some function $\eta = \phi(X_1, \ldots, X_p)$
- Link Function: The random component and systematic component are related through a link function, $g(\mu) = \eta$

Recap

The distribution used in the random component must be a member of the exponential family:

$$f(y_i|\theta_i, \phi) = \mathsf{exp}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

- One common choice of link function sets $g(\mu) = \theta$
- This is the canonical link function

Recap

Binary data describes outcomes that can take the value 1 with probability π_i or 0 with probability $1 - \pi_i$.

In this situation, we assume that the outcome data is described by the binomial distribution, with likelihood

$$f(y_i|n, \pi_i) = \binom{n}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n - y_i}$$

Categorical Data

Models for categorical data extend the binary data situation to cases where outcomes can take $J > 2$ categories.

There are three main types of categorical data, with each having different modeling requirements:

- Nominal scale: Categories are completely exchangeable, and their order does not matter (e.g. disease type)
- Ordinal scale: Categories are ordered but there is no measure of distance between categories (e.g. Hi, Med, Lo)
- Interval scale: Categories are ordered and there are numeric measures of distance between categories

We will focus on models for nominal and ordinal responses today.

Categorical Data

We assume the outcome follows the multinomial distribution with probabilities $\{\pi_1(x), \ldots, \pi_J(x)\}$.

We have the likelihood

$$f(y_i|n, \pi) = \left(\frac{n!}{y_1! \ldots y_J!}\right) \pi_1^{y_1} \ldots \pi_J^{y_J}.$$

Nominal Responses: Baseline-Category Logit Models

Baseline-covariate models compare the probability of observing each response category with a chosen baseline category (e.g. category J).

$$\log \frac{\pi_j(x)}{\pi_J(x)} = \alpha_j + \beta_j^T x.$$

Compare this to the logistic regression model:

$$\log \frac{\pi}{1-\pi} = \alpha + \beta^T x$$

Nominal Responses: Baseline-Category Logit Models

Given model

$$\log \frac{\pi_j(x)}{\pi_J(x)} = \alpha_j + \beta_j^T x,$$

the left hand side is equal to the logit of probability $P(Y = j | Y = j \text{ or } Y = J)$.

We calculate the effects of explanatory variables x on all applicable logits simultaneously.

$$\begin{aligned}
\log \frac{\pi_a(x)}{\pi_b(x)} &= \log \pi_a(x) - \log \pi_b(x) \\
&= (\log \pi_a(x) - \log \pi_J(x)) - (\log \pi_b(x) - \log \pi_J(x)) \\
&= \log \frac{\pi_a(x)}{\pi_J(x)} - \log \frac{\pi_b(x)}{\pi_J(x)}
\end{aligned}$$

Ordinal Responses: Cumulative Logit Models

To model ordinal responses, we use the ordering of the categories to create logit functions of cumulative probabilities

$$P(Y \leq j|x) = \pi_1(x) + \cdots + \pi_j(x), \quad j = 1, \ldots, J$$

From these, we define cumulative logits:

$$\log \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} = \log \frac{\pi_1(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \cdots + \pi_J(x)}$$

Ordinal Responses: Cumulative Logit Models

By creating the cumulative logits, we use the ordinal outcome to create a series of binary outcomes. We can then fit individual logistic regression models for each binary outcome:

$$\text{logit}[P(Y \leq j|x)] = \alpha_j + \beta_j^T x.$$

Note that we calculate separate coefficients for each response level.

This can make model interpretation difficult.

Ordinal Responses: Proportional Odds Models

A special case of the cumulative logit model is the proportional odds model:

$$\text{logit}[P(Y \leq j|x)] = \alpha_j + \beta^T x.$$

Here, we assume the same β values for every response level, but each level has its own intercept.

The logistic regression curves for each response level have the same shape, but are offset from each other by the changing intercept term.

Ordinal Responses: Proportional Odds Models

To interpret the coefficients in proportional odds models, notice

$$\begin{aligned}\beta^T(x_1 - x_2) &= (\alpha_j + \beta^T x_1) - (\alpha_j + \beta^T x_2)\\ &= \mathsf{logit}[P(Y \leq j|x_1)] - \mathsf{logit}[P(Y \leq j|x_2)]\\ &= \log\frac{P(Y \leq j|x_1)}{P(Y > j|x_1)} - \log\frac{P(Y \leq j|x_2)}{P(Y > j|x_2)}\\ &= \log\frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)}\end{aligned}$$

The odds of observing response $Y \leq j$ at $x = x_1$ are $\exp[\beta^T(x_1 - x_2)]$ times the odds at $x = x_2$.

Ordinal Responses: Proportional Odds Models

How should we check the proportional odds assumption?

- Compare a proportional odds model to one allowing for different coefficients β_j for each response level. If coefficients are not significantly different, use the proportional odds model.

Even if the differences in coefficients are statistically significant, we may consider using the proportional odds model when:

- Coefficients are not significant in practical terms
- Separate logistic regression models for binary collapsed outcomes give similar coefficients

Ordinal Responses: Proportional Odds Models

What if the proportional odds model doesn't fit well?

- Try adding additional terms (e.g. interaction terms) to the regression equation
- Try alternative link functions (e.g. complementary log-log)
- Allow separate effects for some predictors, but not all (partial proportional odds)
- Use the more general cumulative logit model

Alternative Models for Ordinal Responses

It is possible to use link functions other than the logit function for ordinal data:

$$g[P(Y \leq j|x)] = \alpha_j + \beta^T x.$$

For example, the cumulative probit model is given below:

$$\Phi^{-1}[P(Y \leq j|x)] = \alpha_j + \beta^T x.$$

Alternative Models for Ordinal Responses

Adjacent-category models are essentially the baseline-category equivalent for ordinal data:

$$\log \frac{\pi_j}{\pi_{j+1}} = \alpha_j + \beta^T x,$$

where $\log \frac{\pi_j}{\pi_{j+1}} = \text{logit}[P(Y = j | Y = j \text{ or } j + 1)]$.

Fitting Models for Categorical Data

Baseline-category model: use the `multinom()` function in `nnet` package

```r
mod <- multinom(formula = y ~ x1 + x2, data = df)
```

Proportional odds model: use the `polr()` function in `MASS` package or `vglm()` function in `VGAM` package

```r
mod <- polr(
  formula = y ~ x1 + x2,
  data = df, Hess = TRUE
)
mod2 <- vglm(
  formula = y ~ x1 + x2,
  data = df,
  family = cumulative(parallel = TRUE)
)
```

Fitting Models for Categorical Data

Cumulative logit model: use the `vglm()` function in `VGAM` package with `parallel = FALSE`

```r
mod <- vglm(
  formula = y ~ x1 + x2,
  data = df,
  family = cumulative
)
```

Testing proportional odds assumption:

```r
pchisq(
  deviance(mod2) - deviance(mod),
  df = df.residual(mod2) - df.residual(mod),
  lower.tail = FALSE
)
```