Chapter 6 GLM

Generalized Linear Model

Instead of $\mu_y = \mathbf{X}\beta$ allow some link function: $g(\mu_y) = \mathbf{X}\beta$ or $\mu_y = g^{-1}(\mathbf{X}\beta)$. Where g^{-1} means apply the inverse function.
And view variance as a function of the mean, $v(\mu)$.
Use weighted least squares for estimating β, which depends on $V(\mu)$ which depends on β which $\cdots$
We have used $g =$ logit for bernoulli counts.
$v(\mu) = v(p) = p(1-p)$
Now work with Poisson (log link, variance equals mean), logistic for binomial counts, probit (Normal CDF) for binomial.
Logit and probit also work with multinomials.
Robust Binomial Regression (robit) using a t distribution CDF.

R Tools

- `glm` used for bernoulli, binomial (probit or logit), Poisson
- `bayesglm` for same with identifiability or separation issues (arm package).
- `polr` or `bayespolr` for ordered multinomial categories, (Mass, arm packages)
- `hett` for unordered multinomial categories and robit

Poisson Example

Number of traffic accidents at specific intersections in NYC over 1 year.
$y_i \sim \text{Poisson}(\theta_i)$

$\theta_i = \exp(\mathbf{x}_i\beta)$
$y_i \sim \text{Poisson}[\exp(2.8 + 0.12 Speed - 0.20 Signal)]$

- Skip constant since no speed is 0
- Increasing speed by 1, multiplies predicted response by $e^{0.012} = 1.012$ or 1.2% increase. (12.7% for a 10 mph increase)
- Adding a signal decreases prediction by $e^{-.2} = .82$ for an 18% decrease.

Poisson with rate

Include the number of cars using the intersection.
$y_i \sim \text{Poisson}(u_i\theta_i)$ where u_i is the "exposure".
$\log(u_i)$ is added in as an "offset" – a predictor with coefficient known to be 1.

Variance may be bigger than the mean if we don't have all of the predictors, or there is spatial or other clustering. (Over-dispersion)

Traffic Stops by Ethnicity 1

3 groups, 75 precincts (not intersections)
Just offset and constant:

```
fit.1 <- glm (stops ~ 1, family=poisson, offset=log(arrests))
display(fit.1)

glm (formula = stops ~ 1, family=poisson, offset=log(arrests))
            coef.est coef.se
(Intercept)     -3.4     0.0
 n = 225, k = 1
 resid deviance = 44877, null deviance = 44877
```


Traffic Stops by Ethnicity 2

Add ethnicity

```
fit.2 <- glm (stops ~ eth, family=poisson, offset=log(arrests))
display(fit.2)

glm (formula = stops ~ eth, family=poisson, offset=log(arrests))
             coef.est  coef.se
(Intercept)     -3.30     0.00
eth2             0.06     0.01
eth3            -0.18     0.01

 n = 225, k = 3
 resid deviance = 44133, null deviance = 44877
```

Added 2 parameters, deviance goes down 744.
1 (blacks), 2 (Hispanic) 3 (white)

Traffic Stops by Ethnicity 3

Add precinct

```
fit.2 <- glm (stops ~ eth + precinct, family=poisson, offset=log(arrest
display(fit.2)

glm (formula = stops ~ eth + precinct, family=poisson, offset=log(arres
             coef.est  coef.se
(Intercept)     -4.03     0.05
eth2             0.00     0.01
eth3            -0.42     0.01
precinct2       -0.06     0.07
precinct3        0.54     0.06
...
precinct75       1.41     0.08
 n = 225, k = 77
 resid deviance = 2828.6, null deviance = 44877
 overdispersion parameter = 18.2
```

Add 75 parameters, deviance goes down 42048.
Coefficients on Blacks and Hispanics agree. Whites: $e^{-0.42} = 0.66$ are $34 \pm 2\%$ lower.

Overdispersion

With Poisson, variance = mean. Check with Standardized residuals.
$z_i = \frac{y_i - \hat{y}_i}{\mathrm{sd}(\hat{y}_i)} = \frac{y_i - u_i\hat{\theta}_i}{\sqrt{u_i\hat{\theta}_i}}$
Watch out for large estimated overdispersion: $\sum z_i^2/(n-k)$
Above, we had overdispersion estimate of 18, so all SE's are too small by a factor of $4.3 = \sqrt{18.2}$
Approximate 50% CI for reduction for whites is
$e^{-0.42 \pm 0.043 \times 2/3} = [0.64, 0.68]$,
95% CI: $e^{-0.42 \pm 0.043 \times 2} = [0.60, 0.72]$.
Poisson likelihood is incorrect under overdispersion.
Use a negative binomial instead of overdispersed Poisson.

Counts of visits to MSU Student Health Service

Can we predict the onset of an epidemic?
If visits to SHS suddenly go up, can we say that an epidemic is just starting?
We have data on the number of students who come into SHS each day with particular symptoms: Upper Respiratory Infection (URI) or Acute Gastro-Enteritis (AGE).
How would we model this traffic?
Argue for

- raw counts of visits with URI or AGE symptoms
- proportion of visits with URI or AGE symptoms

§6.3 Logistic-Binomial

Very similar to Bernoulli data we did in Chapter 5, just grouping together rows for which all predictors match.

$$y_i \sim \text{Binomial}(n_i, p_i) \qquad p_i = \text{logit}^{-1}(\mathbf{X}_i\boldsymbol{\beta})$$

Example: From 1973 to 1995, 34 states used the death penalty and had appeals. Let n_i = number of appeals, of which y_i were overturned.

Model 1: intercept, indicators for 33 of 34 states, a time trend over years 1 (1973) to 23 (1995). Alternative notation:

$$y_{st} \sim \text{Binomial}(n_{st}, p_{st}) \qquad p_{st} = \text{logit}^{-1}(\mu + \alpha_s + \beta t)$$

Overdispersion

As with Poisson, Binomial variance is determined by the mean, $\text{var}(y_i) = \sqrt{n_i p_i(1-p_i)}$. If unmeasured variables cause unexplained changes in y_i, the binomial model can't describe them. Again use standardized residuals to estimate overdispersion.

$$z_i = \frac{y_i - n_i\widehat{p}_i}{\text{sd}(\hat{y}_i)} = \frac{y_i - n_i\widehat{p}_i}{\sqrt{n_i\widehat{p}_i(1-\widehat{p}_i)}}$$

Estimated overdispersion $= \sum z_i^2/(n-k)$ where n is number of rows of data, k is rank of $\mathbf{X}$. Correct SE's by multiplying each by $\sqrt{\text{overdispersion}}$. Using family = `quasibinomial` does this for us.

§6.4 Probit Regression

Let $\Phi(\cdot)$ be the std normal CDF function

$$P(y_i = 1|\mathbf{X}_i) = \Phi(\mathbf{X}_i\boldsymbol{\beta})$$

or using z as unknown latent variable

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \qquad z_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i \quad \epsilon_i \sim \text{iid N}(0,1)$$

Std logistic distribution matches $N(0, 1.6^2)$ so multiply probit coefficients by 1.6 to get logistic coefficients.

Logit vs Probit

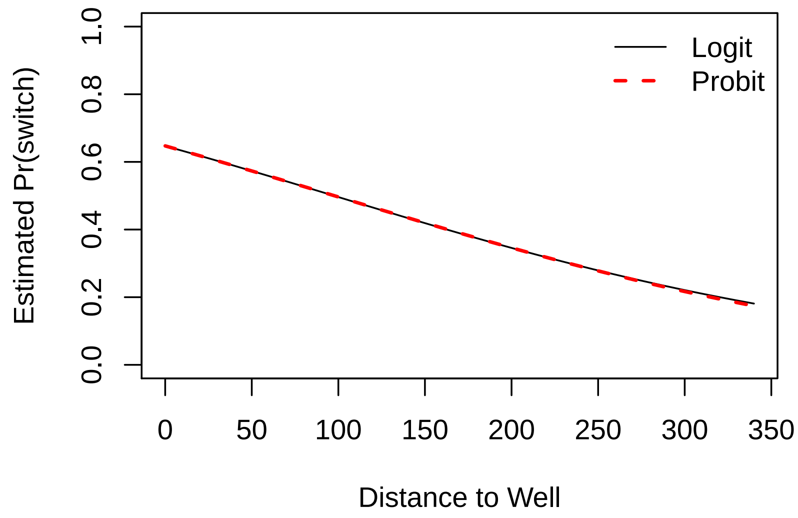
```
wells <- read.table("http://www.math.montana.edu/~jimrc/cla
wells$dist100 <- wells$dist/100
logitfit <- glm(switch ~ dist100, wells, family = binomial)
xtable( summary(logitfit)$coef, digits=3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.606	0.060	10.047	0.000
dist100	-0.622	0.097	-6.383	0.000

```
probitfit <- glm(switch ~ dist100, wells,
                 family = binomial(link="probit"))
xtable( summary(probitfit)$coef, digits=3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.378	0.037	10.128	0.000
dist100	-0.387	0.060	-6.420	0.000

Logit vs Probit Plot



§6.5 Categorical Response

Response has more than 2 outcomes.

Ordered categories: military ranks, [Never, Sometimes, Always], [strongly favor , favor, neutral, oppose, strongly oppose]. Are they equally spaced?
Unordered: hair color, ethnicity, sports

Ordered Categories: $1, \ldots, K$

Assume the same logit function (same slope coefficients) takes us up one level from any given level.

$$
\begin{aligned}
\Pr(y > 1) &= \text{logit}^{-1}(\mathbf{X}\beta) \\
\Pr(y > 2) &= \text{logit}^{-1}(\mathbf{X}\beta - c_2) \\
\Pr(y > 3) &= \text{logit}^{-1}(\mathbf{X}\beta - c_3) \\
\vdots &= \vdots \\
\Pr(y > K-1) &= \text{logit}^{-1}(\mathbf{X}\beta - c_{K-1})
\end{aligned}
$$

c_i are ordered cutpoints, c_1 set to 0 (binary outcomes, $K = 2$) estimated via maximizing likelihood.

Outcomes

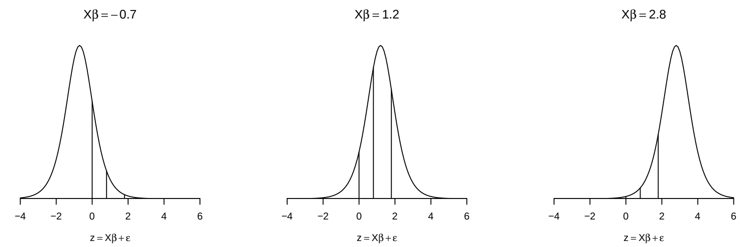
At level k

$$\begin{aligned}\Pr(y == k) &= \Pr(y > k-1) - \Pr(y > k)\\ &= \operatorname{logit}^{-1}(\mathbf{X}\boldsymbol{\beta} - c_{k-1}) - \operatorname{logit}^{-1}(\mathbf{X}\boldsymbol{\beta} - c_k)\end{aligned}$$

Latent variable:
Cut $z_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$ at points $c_1 = 0, c_2, \ldots, c_{K-1}, c_K = \infty$.
Errors ϵ_i are iid, logistic. Let $y_i = k$ if $z_i \in (c_{k-1}, c_k)$

4 Categories Example

As $\mathbf{X}\boldsymbol{\beta}$ increases, the higher categories become more likely.



Banked Votes

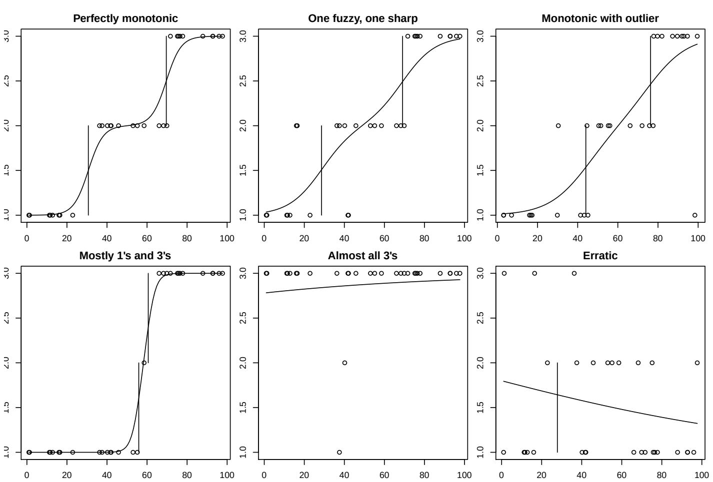


Figure 6.4 Voting patterns

Models

Variance (6.11)	Slope-Intercept (6.12)
$y_i = \begin{cases} 1 & \textit{if } z_i < c_{1.5} \\ 2 & \textit{if } z_i \in (c_{1.5}, c_{2.5}) \\ 3 & \textit{if } z_i > c_{2.5} \end{cases}$ $\mathbf{z}_i \sim \text{logistic}(x_i, \sigma^2)$	$y_i = \begin{cases} 1 & \textit{if } z_i < 0 \\ 2 & \textit{if } z_i \in (0, c_2) \\ 3 & \textit{if } z_i > c_2 \end{cases}$ $\mathbf{z}_i \sim \text{logistic}(\alpha + \beta x_i, 1)$

Slope (6.13)	Equivalences
$y_i = \begin{cases} 1 & \textit{if } z_i < c_{1\|2} \\ 2 & \textit{if } z_i \in (c_{1\|2}, c_{2\|3}) \\ 3 & \textit{if } z_i > c_{2\|3} \end{cases}$ $\mathbf{z}_i \sim \text{logistic}(\beta x_i, 1)$	$\begin{array}{rcc} c_{1.5} = & -\alpha/\beta = & -c_{1\|2}/\beta \\ c_{2.5} = & (c_2 - \alpha)/\beta = & -c_{2\|3}/\beta \\ \sigma = & 1/\beta = & 1/\beta \end{array}$

Fit with Proportional Odds

Proportional Odds Logistic Regression

```
vote.fit1 <- bayespolr(factor(y) ~ x)
display(vote.fit1)

     coef.est  coef.se
 x    0.10      0.04
 1|2 3.46       1.53
 2|3 7.03       2.44
    n=20 k=3 (including 2 intercepts)
    residual deviance = 32.2, null deviance is not computed by polr
```

Equivalently: $\hat{\sigma} = 10$, $\hat{c}_{1.5} = 3.46/.10 = 34.6$, $\hat{c}_{2.5} = 7.03/.10 = 70.3$

Zero-truncated Poisson

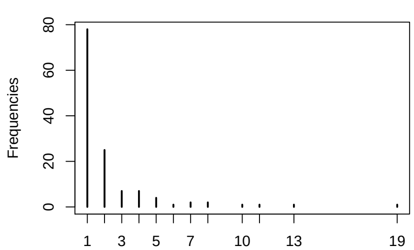
Ref: Zuur, Walker, Saveliev & Smith (2009) *Mixed Effects Models and Extensions in Ecology with R* give examples and analysis of zero-truncated models in Chapter 11.

- If the mean of the Poisson process is large, zeros will be rare, and are not expected.
- Occasionally zeroes are missing values:

Ecologists in Portugal studied snakes killed on roadways and recorded the number of days a snake carcass lay on the road. A zero means the snake crossed the road, and is not part of the dataset. Even if the carcass is only hours old, a 1 was recorded.

Snake Carcasses

```
snakefit1 <- MASS::glm.nb(N_days ~ PDayRain * Tot_Rain + Road_Loc,
    data = Snakes)
snakefit2 <- vglm(N_days ~ PDayRain * Tot_Rain + Road_Loc, data = Snakes,
    family = posnegbinomial, control = vglm.control(maxit = 100))
plot(table(Snakes$N_days), ylab = "Frequencies")
```



Use negative binomial to account for overdispersion

Compare Snake Fits

```
Z <- cbind(coef(snakefit1), coef(snakefit2)[-2])
ZSE <- cbind(sqrt(diag(vcov(snakefit1))), sqrt(diag(vcov(snakefit2))[-1]))
Comp <- cbind(Z[, 1], Z[, 2], ZSE[, 1], ZSE[, 2])
dimnames(Comp)[[2]] <- c("NB", "Trunc.NB", "SE NB", "SE Trunc.NB")
```

	NB	Trunc.NB	SE NB	SE Trunc.NB
(Intercept)	0.37	-1.57	0.11	1.09
PDayRain	-0.00	0.11	0.19	0.42
Tot_Rain	0.12	0.24	0.02	0.06
Road_LocV	0.45	1.04	0.15	0.35
PDayRain:Tot_Rain	-0.11	-0.22	0.02	0.06

Truncated Negative Binomial has coefficients further from zero with greater SE.

Too Many Zeroes

Go to a site in the Greater Yellowstone Ecosystem and look for grizzly bears.
Why might we see zero bears?

- Habitat is not suitable.
- "Design error" It's winter and bears are all in dens.
- "Observer error" we missed signs which were there.
- "Bear error" good habitat, but bears haven't found it.
- Naughty naughts: bad zeroes from sampling outside the range (downtown Bozeman?) Remove these from the sample.

Zero Altered

Hurdle models:

1. Model zeroes versus non-zeroes as binomial with logistic regression.
2. Given some were observed, use a truncated Poisson or negative binomial to model frequency.

Economists use censored regression models for variables like earnings or "labor supply". Tobit models assume a latent variable with a threshold at which the response becomes positive (probit regression). Above the threshold use a linear model with the same predictors (and β?) for the positive responses. Combined likelihood includes the binomial and linear model for those over threshold.

Zero inflated

Mixture of two models

1. $1 - p$ of the time we observe a "false" zero due to observer or bear error.
2. p of the time we observe a Poisson or negative binomial, which could also give a zero (bear error or non-suitable habitat).

See Zuur et al. for examples.

Other models

- Survival data (log right tailed, censored) modeled with Gamma, Weibull, or proportional hazards models.
- Nonparameteric models (highly parametric?)
  - gam generalized additive models
  - neural networks
  - support vector machines

  for nonlinear trends

Constructive Choice Models

Decision theory assumes there are costs (minimize loss) and gains (maximize utility) for actions described by a *value* function.

- a_i is benefit of switching from unsafe to safe well (in \$?)
- $b_i + c_i x_i$ is the \$ cost of switching when well is $100x_i$m away.

Logit or probit model: switch if $a_i > b_i + c_i x_i$

$$\Pr(y_i = 1) = \Pr\left(\frac{a_i - b_i}{c_i} > x_i\right)$$

Constructive Choice Models 2

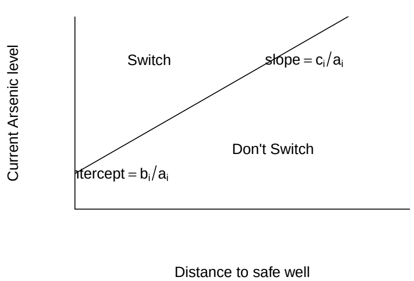
$a_i,\ b_i,\ c_i$ are not identifiable, but let $d_i = \frac{a_i - b_i}{c_i}$ and assume it has a logistic (normal) distribution with mean μ and spread σ.

$$\Pr(y_i = 1) = \Pr(d_i > x_i) = \Pr(\frac{d_i - \mu}{\sigma} > \frac{x_i - \mu}{\sigma}) = \text{logit}^{-1}(\frac{\mu - x}{\sigma})$$

or in probit regression, $\Phi(\frac{\mu - x}{\sigma})$. We need a slope and intercept for x.

Can estimate the population-average model, not individual values for each household.

Individual Choice



Each individual has their own cost and value.
Switch if $a_i(As) > b_i + c_i x_i$.
Analysis shows people mistakenly used $\log(As)$ instead of As.