good luck

adv. data analysis

Midterm 2

generalized linear models

8.1-8.2 basics

- we want to use non-linear models while still having the benefits of linear regression...

$\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

we use a link function

$\eta = g(\mu)$

- most glm's are made up of responses that are members of exponential family dist.
- almost anything can be a link function, but canonical link functions exist...

family	canonical link	var. func.
normal	$\eta = \mu$	1
poisson	$\eta = \log \mu$	μ
binomial	$\eta = \log(\mu/(1-\mu))$	$\mu(1-\mu)$
gamma	$\eta = \mu^{-1}$	μ^2
inv. gaus.	$\eta = \mu^{-2}$	μ^3

- The distributions also have dispersion parameters ϕ. The poisson and binomial have $\phi = 1$.
- To estimate β's we use an algorithm...

(0) we need an initial estimate $\hat{\mu}^0$ = fitted values, $\hat{\eta}^0 = g(\hat{\mu}^0)$

(1) then we solve for adjusted dependent variable

iteration i: $z^i = \hat{\eta}^i + (y - \hat{\mu}^i)\frac{\partial \eta}{\partial \mu}$

(2) Find weights... $\frac{1}{\omega_i} = \left(\frac{\partial \eta}{\partial \mu}\right)^2 \cdot V(\hat{\mu}^0)$ (variance)

(3) Fit linear model to estimate $\hat{\beta}^i$ and use this as the next iteration of $\eta \to \hat{\eta}^{i+1}$

(4) repeat until convergence.

- This method is good, but the std. errors are wrong. find with $\text{Var}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$ w/ W = diag. weight matrix, X = model matrix

8.3 hypothesis tests

- we can measure goodness of fit by comparing our model to the full model (model w/ all possible var. and interactions). We use an equation to find deviance which can be compared to a χ^2 dist. (w/ df = obs. - param.), $\leq .05$ imply good fit.
- we can also compare two nested models (Ω → larger, ω → smaller) by subtracting their deviances and comparing it to a χ^2 dist...

$D_\omega - D_\Omega \to$ df = diff. in # of params, $\leq .05$ imply bigger model better

8.4 diagnostics

- we can't measure residuals the same way we do for linear models, since variance of response is not uniform for most glm's. So we use Pearson Residuals or Deviance Residuals.
- we can use diagnostics to check Leverage & Influence & Unusual Points → we look at a QQ plot and cook's dist. → we look at a half-norm plot. (NOT checking for normality)

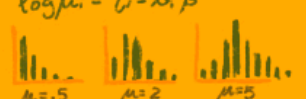
5.1 count reg. & poisson

- Lets take a closer look at one kind of glm → Poisson!

$Y_i \sim \text{Pois}(\mu_i)$

$\log \mu_i = \eta_i = x_i^T \beta$

Pois. dist. are typically skewed right.



- Poisson links work really well for ① unbounded counts (0, 1, 2, ...), ② counts w/ small prob. of success, ③ events where time between them is iid exp. dist.
- For poisson, sum is also poisson and indp!
- Poisson is a type of count reg. model.

5.2-5.5 poisson variations

- Dispersed Poisson → modification of Poisson to allow more var. in resp. We let Y's rate λ also be a RV...

$Y \sim \text{Pois}(\lambda)$, $\lambda \sim \text{Gamma}$, $E[\lambda] = \mu$, $\text{Var}\lambda = \mu/\phi$

$E[Y] = \mu$, $\text{Var}(Y) = \mu(1+\phi)/\phi$ (ϕ = the dispersion param.)

$\phi > 1 \to$ overdispersion, $\phi < 1 \to$ under dispersion

in R → glm(..., family = quasipoisson, ...) (models w/ overdispersion should be compared w/ F test rather than χ^2)

- Rate Models → occurences may depend on a size variable that determines # of opportunities for occurences (ie # of goals scored depends on time played), so we use a rate model → where coef is fixed at 1.

in R → glm(... ~ offset(), ...)

- Hurdle Model → what if we have way more 0's in our data than a Poisson or similar model would predict? Well we have a couple ways to deal w/ that. One way is to imagine there is a hurdle that an observee has to get over, that most people never get over, then when they get 1 it's more likely they'll get more (ie first book published, first robbery)

$P(Y=0) = f_1(0)$ ($f_1(\cdot) \to$ binary response)

$P(Y=j) = \frac{1-f_1(0)}{1-f_2(0)} f_2(j)$, $j > 0$ ($f_2(\cdot) \to$ truncated poisson)

in R → hurdle(y ~ x's, data)

- Zero Inflated Poisson Model → same problem as above, data has lots of 0's! Idea of ZIP model is only a certain population will have > 0. (Compare w/ χ^2 dist.)

$P(Y=0) = \phi + (1-\phi) f(0)$ $P(Y=j) = (1-\phi) f(j)$, $j > 0$

in R → zeroinfl(y ~ x's, data)

- Negative Binomial → series of indp. trials each w/ prob. of success = p. Z = # of trials until kth success. This is a neg. binomial dist. If we let $Y = Z - k$ & $p = (1+\alpha)^{-1}$

$P(Y=y) = \binom{y+k-1}{k-1} \frac{\alpha^y}{(1+\alpha)^{y+k}}$

we end up w/ a glm w/...

$E[Y] = \mu = k\alpha$

$\text{Var}Y = k\alpha + k\alpha^2 = \mu + \frac{\mu^2}{k}$

link function...

$\eta = \log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{\mu+k}$

in R → glm(..., negative.binomial(k), ...) or glm.nb()

contingency tables

6.1-6.2 two-by-two tables

- Contingency Table → a way to show cross classified data on 2 or more categorical variables. Variables can be nominal (no natural order) or ordinal (has a natural order). A great way to start observing data and dependence.
- A two-by-two table has two variables...
- We can judge indp. w/ a χ^2 test on the xtab.
- We can also fit a poisson w/ count as the response and test the deviance to test the dependence (very similar conclusion as xtabs)
- We can also use Fisher Test on our table (fisher.test()) this will return an odds ratio that is a measure of association between variables.
- We can also visually judge dependence w/ Mosaic Plots (mosaicplot(..., color=T, main=NULL, las=1))

var 1	outcome1	outcome 2	total
var 2 a.1	count		
var 2 a.2			
total			

in R → xtabs(y ~ var1 + var2)

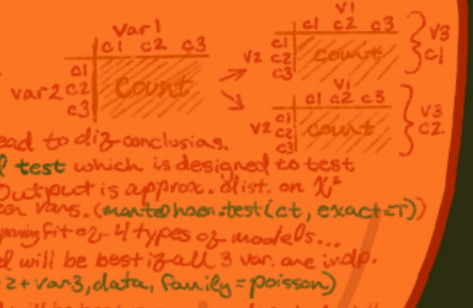
6.4 matched pairs

- w/ matched pairs we build a contigency table where we observe one measure on two matched objects.
- We can also build a glm and then build a matched pairs contingency table of residuals.
- These matched pair contingency tables can be really useful for testing symmetry. We can use a quasi-symmetry model w/ a normal model in anova() test w/ Chi dist. If the test result is sig. there is evidence of lack of marginal homogeneity.

teacher	student A	B	C	D
A				
B	count			
C				
D				

6.5 3-way contingency table

- what if we produce a two way contingency table that shows/doesn't show a relationship we don't/do expect to be there? Well maybe we need to condition upon a third variable, this is how we make a 3-way contingency table
- marginal association → association between 2 var.
- conditional association → association obsv. within each category in the 3rd var.
- Simpson's paradox → when the marginal association and the conditional association lead to diff. conclusions.



- We can also use the Mantel-Haenszel test which is designed to test indp. in 2x2 tables across K strata. Output is approx. dist. on χ^2 dist. If stat. sign ⇒ association between vars. (mantelhaen.test(ct, exact=T))
- We can also test association by comparing fit of 4 types of models...
- Mutual Independence → This model will be best if all 3 var. are indp.
  in R → glm(count ~ var1 + var2 + var3, data, family = poisson)
- Joint Independence → This model will be best if 2 vars are dep. but jointly independent of the 3rd. In R → glm(count ~ var1 * var2 + var3, ...)
- Conditional Independence → This model will be best if the first and second var. are indp. given the third. In R → glm(count ~ var1 * var3 + var2 * var3, ...)
- Uniform Association → This model will be best if each of the variables have some association with both other vars. In R → glm(count ~ (var1 + var2 + var3)^2, ...)

random effects

dealing the parameter (fixed effect).

- mixed effect → when we have random effect and fixed effects in one model...
- simplest design of random effect model...

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

α_i: mean = μ, $V = \sigma_\alpha^2$; ε_{ij}: $V = \sigma_\varepsilon^2$

Intraclass correlation coef → $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}$ → $\rho = 0$ ⇒ no var, $\rho \to 1$ ⇒ lots of var

- in R → lme4::lmer(response ~ 1 + (1|re), ...)
- in our model it doesn't make sense to try and estimate the re, instead we estimate it's func....
- to test two RE models against each other ① create a null model w/o component your testing ② test either likelihood ratio statistic (REML) of nested models against each other, it should be approx. χ^2 dist unless parameters of null mod are on boundary of param. space.
- in R → RLRsim::exactLRT(mod, nullmod), for one mod → RLRsim::exactRLRT(mod)
- for balanced data (same # of obsv. in all levels) we can use sum of squares in anova decomposition for hypothesis testing.
- for fixed effects we should test w/ F test.
- we can also use parametric bootstrap to find pvalues for LRT: To do this we generate data under null mod using fitted param estimates. we repeat this many times to find pvalues. in R → confint(mod, method = "boot")
- we can use AIC to compare many models.

10.1-10.3 basics

- Sometimes we recognize that obsv. being in a certain group causes correlation so we can't assume indp. In these cases we can use random effects → when a rv describes a parameter of our data, rather than constant

$y_{ijk} = \mu + \tau_i + \nu_j + \varepsilon_{ijk}$ (fixed, random, error)

variances were originally found using MSE and MSA, but this is mathematically very complicated and can return negative values, so now we use MLE or the less biased REML

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $\alpha_i \sim N(0, \sigma_\alpha^2)$ $\hat{\alpha} = D Z^T V^{-1}(y - X\beta)$

10.4 prediction

- best linear unbiased predictors → w/ operators we can derive BLUPs to predict new values.
- in R → predict(mod, re.form = ~0)
  or if we know operator a...
  predict(mod, newdata = df(operator = "a"))
- func. does not give us intervals or std errors cause math is to hard.
- for simple models we can bootstrap to get pred. intv.
  in R → lme4::bootMer()

10.5 diagnostics

- assumptions for our mixed lin mods are very similar to linear models. Differences...
  - mixed effect models have more than one kind of fitted residual → degen if is estimated random effect
  - random effects mods are extra sensitive to outliers
  - you can look at normality in group levels.

10.6 blocks as re's

- blocks → properties of the experimental units, groupings. typically we are not interested in the block's effects specifically, but we need to account for their effects, so we use REs!
- in R → lmer(response ~ fixed + (1|block)
- we can then test if our mod to see if blocking is helpful or if we should stick to fixed.