S-043/Stat-151
Analysis for Clustered and Longitudinal Data
(Multilevel & Longitudinal Models)

Unit 5, Lecture 4
Multilevel
Generalized Linear Models

"A key problem with model-based inference—if exploratory analysis is not performed—is that if an inappropriate model is fit to data, it is possible to end up with highly precise, but wrong, inferences."

- A. Gelman on EDA for Complex Models

Final Projects

Final Project Workshop is not the end of everything!

You can continue talking with the teaching staff!

Goals for Today

Continue to solidify generalized multilevel regression

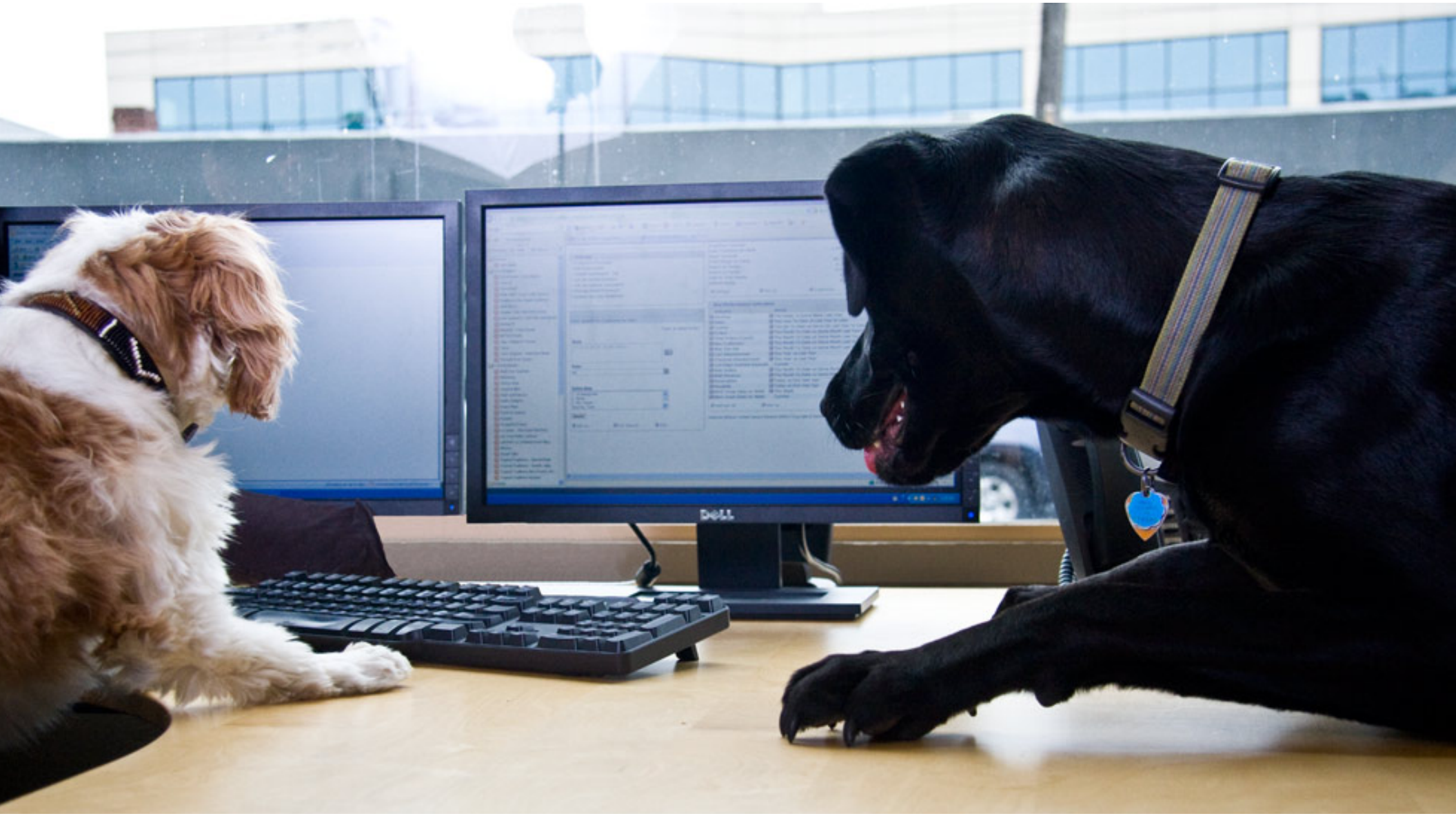
Talk more about Poisson regression (and police stops)

Investigate strategies for over dispersion

Look at sensitivity checks for model building (and how to build and view multiple models)

Main goal: Illustrate how creative use of models can allow for interesting exploration of data.

Quick code tips on factors and complete cases



Code Tip: factors, characters, numbers

If you have something as a character (e.g., “Yes” and “No”) and you want the proportion of it in your group, try this:

```
# make fake data
> a = data.frame( G = sample( letters[1:5], 30, replace=TRUE ),
        Y = sample( c("Yes","No","Maybe"), 30, replace=TRUE ) )

# summarize it!
> a %>% group_by( G) %>% summarize(
        mn = mean( Y == "Yes" ) )
  G         mn
1 a 0.0000000
2 b 0.7500000
3 c 0.1111111
4 d 0.3333333
5 e 0.3333333
```

Factors are numbers, sort of.

```
> a$Y = as.factor( a$Y )
> levels( a$Y )
[1] "Maybe" "No"    "Yes"

> table( a$Y == 3 )
FALSE
   30

> table( as.numeric( a$Y ) == 3 )
FALSE  TRUE
   21     9

> table( a$Y == "Yes" )
FALSE  TRUE
   21     9

> B = as.numeric( a$Y == "Yes" )
> table( B )
B
 0  1
21  9
```

complete.cases() for dropping cases

```
> a = data.frame( A=1:5, B=letters[1:5], C=LETTERS[1:5] )
> a$A[c(1,3)] = NA
> a$C[3:4] = NA
> a
   A B    C
1 NA a    A
2  2 b    B
3 NA c <NA>
4  4 d <NA>
5  5 e    E
>
> complete.cases( a )
[1] FALSE  TRUE FALSE FALSE  TRUE
> keep = complete.cases( a[c("B","C")] )
> keep
[1]  TRUE  TRUE FALSE FALSE  TRUE
>
> b = filter( a, keep )
> b
   A B C
1 NA a A
2  2 b B
5  5 e E
```

This method tells you which rows are complete.

Here we look at only the listed columns

If keep is a bunch of TRUE and FALSE, select all rows corresponding to TRUE.

pop quiz on prior lecture!

Pop quiz: calculating probabilities

```
> M1 = glmer( outcome ~ treatment * month + (1|patient),
              family=binomial,
              data=toes )
>
>
> display( M1 )
                  coef.est coef.se
(Intercept)        -2.51     0.76
treatment          -0.30     0.69
month              -0.40     0.05
treatment:month    -0.14     0.07

Error terms:
 Groups   Name        Std.Dev.
 patient  (Intercept) 4.56
---
number of obs: 1908, groups: patient, 294
AIC = 1265.6, DIC = -26
deviance = 615.0
```

Question:
What is the probability of a treated patient having a detachment at month 3 if their initial proclivity for detachment is 1 standard deviation above average?

Briefly, how do I pick a model?

A tiny taxonomy

Count data

- ★ # events in a given period of time
- ★ Do you want to have multiplicative effects?
  - Yes: Use Poisson
  - No: Use linear

Binary data

- ★ 0/1 outcome (Bernoulli) or # events out of # trials (Binomial)
- ★ Do you have rare events?
  - Yes: Use logit (or probit)
  - No: Use logit/probit or linear

You can almost always use Linear Regression with robust SEs

GLMs give more nuanced interpretation of coefficients, which is often useful for rare events.

Stop-and-Frisk revisited (Multilevel Poisson Models)

Example taken from G&H 15.1

Data: Police stops by ethnic group

- ★ Units *i* are precincts and ethnic group combos (i = 1, ..., n = 3 x 75)
- ★ *exposure* u_i is the number of **arrests** by people of that group in that precinct in the previous year (as recorded by DCJS)
  - Why arrests? We want to (attempt to) control for baseline level of crime and baseline number of people from that group in that precinct.

1=black, 2=hispanic, 3=white

```
> head( stops )
  precinct eth past.arrests stops   pop
1        1   1          980   202  1720
2        1   2          295   102  1368
3        1   3          381    81 23854
4        2   1          753   132  2596
```

Outcome yi is number of stops of that group in the precinct

Reminder of prior analysis

In Lecture 5.1 we used a Poisson generalized linear model to analyze these data.

We had precinct as a fixed effect:

```
> fit.4 <- glm (stops ~ factor(eth) + factor(precinct),
          family=quasipoisson,
          offset=log(past.arrests), data=stops )
```

We were able to see if the different ethnic groups were stopped at different rates, controlling for precinct and for prior rate of arrest (i.e., population and "true" criminal activity). (They were.)

The precinct fixed effects made our comparisons *within precinct comparisons*. (I.e. we asked, in a given place, was there imbalance?)

We can change precinct to a random effect

Benefits:

- ★ Estimate how much the precincts vary
  - This shows us if there is differential rates of stop-and-frisk geographically, controlling for criminal activity.

- ★ Obtain superior estimates of individual precincts via *partial pooling*.
  - This allows us to estimate which precincts are more or less "active."

- ★ Provides a platform to do some fancy and cool things
  - We will see this soon.

Fitting a multi-level poisson model

```
> Mall = glmer( stops ~ 1 + eth + (1|precinct),
+                 offset = log(past.arrests),
+                 family = poisson(link="log"),
+                 data=stops )

> display(Mall)
             coef.est coef.se
(Intercept) -0.45     0.07
ethHispanic  0.01     0.01
ethWhite    -0.42     0.01

Error terms:
 Groups   Name        Std.Dev.
 precinct (Intercept) 0.60
 Residual             1.00
---
number of obs: 225, groups: precinct, 75
AIC = 5690.4, DIC = 1172
deviance = 3427.3
```

Recall past arrests is our proxy for expected number of legitimately arrestable people of that group in that precinct

How much precincts vary (in the log rate space)

Quick Aside: Over-dispersion in single level count data

Problem: Accounting for overdispersion

We have seen two ways of handling over dispersion:

- ★ Quasipoisson (easy to do)
- ★ Negative binomial (similar to quasipoisson)

We can also model over dispersion by adding a random effect for each individual case.

- ★ This allows some cases to have more events than expected and some less, given exposure and observed covariates.

This works because we can think of the individual case as, in some sense, multiple observations (the individual counted events).

Enhancement: Modeling overdispersion

We can model over dispersion as a *random intercept* at the observation level (in level 1 data)

$$y_i \sim \text{Poisson}\left(u_i e^{X_i\beta}\right)$$

$$\begin{aligned} y_i &\sim \text{Poisson}\left(u_i e^{X_i\beta+\epsilon_i}\right) \\ \epsilon_i &\sim \text{N}(0, \sigma^2_\epsilon). \end{aligned}$$

This residual is our unit-level random effect.

We add some noise to the individual rates. This corresponds to other, missing predictors and uncertainty.

This is different from using a negative binomial!

Random intercepts provide over-dispersion (even in "Single Level Data"!!)

We can add random intercepts where each observation gets its own intercept

- ★ This is quite different from our general approach: normally we need some clusters with multiple observations

BUT, because the Poisson dictates what the variance (residuals) "should be" we can get our random intercepts with only 1 observation per "group"

See RH&S 13.9 & 13.10 for further discussion

Over-dispersion *beyond* precinct

This is the model from the past lecture on Police Stops

```
> stops$precinct = as.factor( stops$precinct )

> fit.4 <- glm (stops ~ eth + precinct,
                family=quasipoisson,
                offset=log(past.arrests),
                data=stops )

> display( fit.4 )

             coef.est coef.se
(Intercept)  -1.38     0.24
ethHispanic   0.01     0.03
ethWhite     -0.42     0.04
precinct2    -0.15     0.35
precinct3     0.56     0.27
…
  overdispersion parameter = 21.9
```

Our quasipoisson model allows for overdispersion. (But is a bit mysterious in what it is up to.)

It does estimate how much overdispersion we have, however.

We still have over-dispersion BEYOND the precinct variation.

This is variation predicted by omitted variables.

Our new approach for overdispersion: "Single Level" data with random intercept

```
> stops$id = with( stops,
             paste( precinct, eth, sep="-" ) )
```

We make a new "group" id for each observation

```
> head( stops$id )
[1] "1-Black"    "1-Hispanic" "1-White"    "2-Black"    "2-Hisp

> fit.4b = glmer( stops ~ 1 + eth + precinct + (1|id),
                  offset = log(past.arrests),
                  family = poisson(link="log"),
                  data=stops )

> display( fit.4b )

             coef.est coef.se
(Intercept)  -1.25     0.14
ethHispanic   0.02     0.04
ethWhite     -0.50     0.04
precinct2    -0.23     0.20
precinct3     0.57     0.20
...
```

Alternatively, to make a new group id, you could say:

```
stops$id = 1:nrow(stops)
```

The prior way makes "nice looking" ids, however.

Now back to our multi-level Poisson model

We want random Precinct effects but we also want to allow for over-dispersion *beyond what can explained by precinct variation*

How do we do it?

Our overdispersed Poisson multilevel model

e = ethnic/racial group
p = precinct

"race/ethnic effects" (pooled across precinct)

$$y_{ep} \sim Poisson\left(\frac{15}{12} n_{ep} e^{\beta_{0p} + \beta_1 W_{ep} + \beta_2 H_{ep} + \epsilon_{ep}}\right)$$

$$\epsilon_{ep} \sim N(0, \sigma^2)$$

"residual noise" to decouple variance and count

$$\beta_{0p} = \gamma_{00} + u_p \text{ with } u_p \sim N(0, \tau)$$

random precinct effect

n_{ep}: # arrests recorded by DCJS for that group in that precinct in that year (multiplied by 15/12 to get to a 15-month period for comparability).

Note: Our observation period is 15 months, but prior arrests was for a single year.

Same model as prior slide, broken into components (for reference)

Observed outcome is Poisson with mean of exposure times rate

$$y_{ep} \sim Poisson\left(\left(\frac{15}{12} n_{ep}\right) \cdot \mu_{ep}\right)$$

our rate of stops/unit exposure

$$\mu_{ep} = exp(\eta_{ep})$$

exp() is our link from our linear predictor to our rate

linear predictor

$$\eta_{ep} = \beta_{0p} + \beta_1 W_{ep} + \beta_2 H_{ep} + \epsilon_{ep}$$

Look, we have residuals again! The residuals allow for variation BEYOND what the Poisson says we should have

$$\epsilon_{ep} \sim N(0, \sigma^2)$$

Our level-2 model (random intercept)

$$\beta_{0p} = \gamma_{00} + u_p \text{ with } u_p \sim N(0, \tau)$$

Fit our model to see differential rates of stops

```
> Mall = glmer( stops ~ 1 + eth + (1|id) + (1|precinct),
                offset = log(past.arrests),
                family = poisson(link="log"),
                data=stops )
>
display( Mall )
             coef.est coef.se
(Intercept) -0.49     0.08
ethHispanic  0.02     0.05
ethWhite    -0.50     0.05

Error terms:
 Groups   Name        Std.Dev.
 id       (Intercept) 0.29
 precinct (Intercept) 0.60
 Residual             1.00
---
number of obs: 225, groups: id, 225; precinct, 75
AIC = 2899.8, DIC = -2848.2
deviance = 20.8
```

'eth' is our fixed part. We only have random precinct and overdispersion random intercepts

Black is baseline. White, Hispanic is relative to this.

Individual variation from unobserved covariates (overdispersion) substantial.

Precinct variation is larger than the black-white difference

NOTE: We are modeling total stops across the crime types so far.
See R file for how aggregation across crime types is done.

Looking at shifts relative to grand mean instead of shifts relative to baseline group

```
> fixef( Mall )
 (Intercept) ethHispanic    ethWhite
-0.49320098  0.01833415 -0.49749359

> fes = fixef( Mall )
> fes[2:3] = fes[1] + fes[2:3]

> fes
 (Intercept) ethHispanic    ethWhite
  -0.4932010  -0.4748668  -0.9906946

> mean( fes )
[1] -0.6529208

> fes - mean( fes )
 (Intercept) ethHispanic    ethWhite
   0.1597198   0.1780540  -0.3377738
```

We add in the intercept to the other groups to get mean (log) rates for each group.

Mean rate averaged across three groups

Our three "ethnic shifts" relative to overall mean

Compare to G&H Numbers:
0.1549558 0.1736464 -0.3286022

Question:
if we didn't want to do this, how change our original model?

What have we learned so far?

There is considerable precinct variation!

- ★ The variation in amount of stop-and-frisk in precincts swamps that of different ethnic groups.
- ★ We might then ask: what precincts have higher rates of stop-and-frisk, beyond baseline rates of crime? (Any guesses?)

There is a lot left to explain!

- ★ The residual variation within precinct is around 0.3 (half the precinct variation, and remember this is above and beyond the inherent randomness of the Poisson.
- ★ There are things causing higher and lower rates of stop-and-frisk for specific ethnic groups to be different that our model does not account for.

Getting a more detailed look: Subgroup analysis and including predictive covariates

Are these trends different across precinct type and crime type?

Precinct type:

Some precincts are predominately black, some are predominately not black.

How do rates of stops might vary based on precinct?

Reason for stop:

Police record why they stopped someone (violent crime, weapons crime, property crime, drug crime)

Are there different patterns of race/ethnic gaps for different crime (and for different precinct types)?

Exploring variation across precincts

We have 4 crime types ("suspected charges") as listed on the official stop form.

We divide the precincts into three groups

< 10% black, 10-40% black, > 40% black

and run our model for each crime type for each precinct subgroup

This gives us 4 x 3 = 12 runs of our model with different outcomes (crime type) and different groups of precincts.

This allows the full estimation to vary:

Our 12 models are completely unpooled.

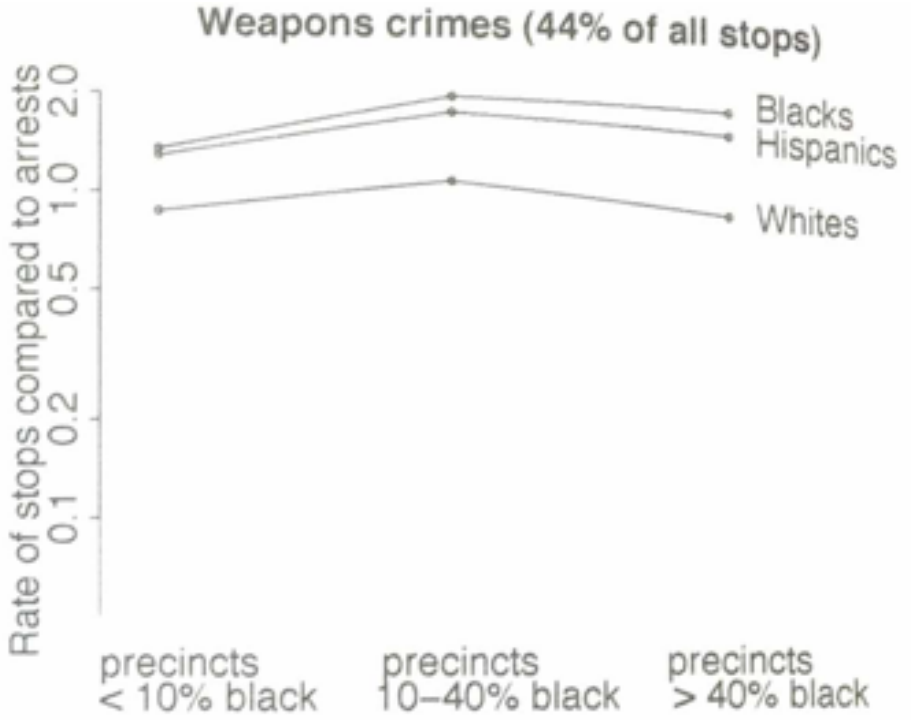
Alternative option: full modeling

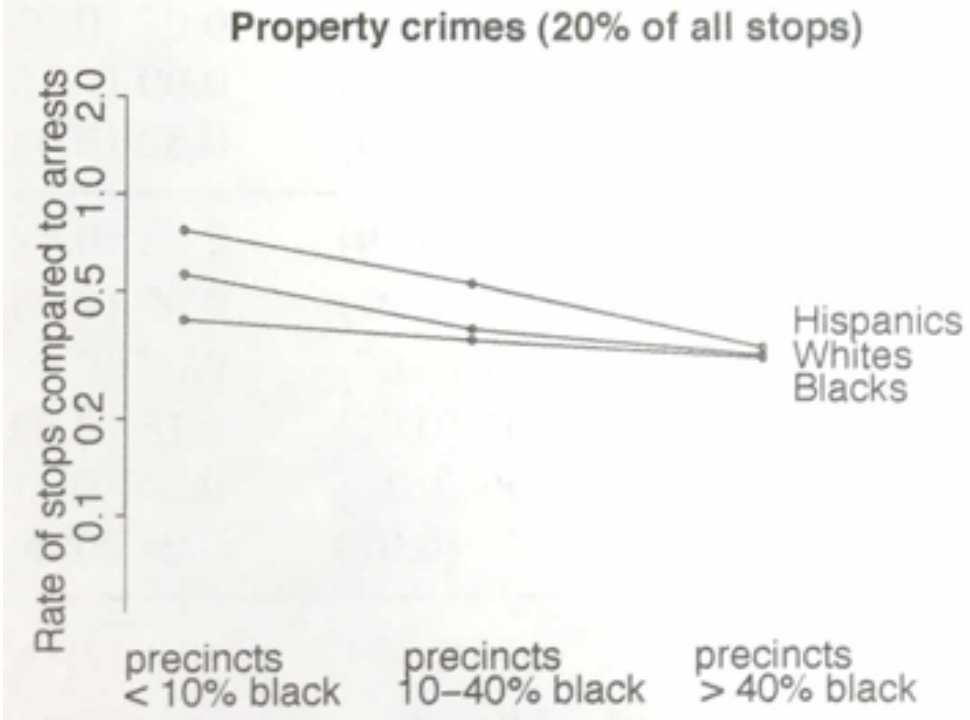
Fit one large model with dummies for all these things. This would pool across models.

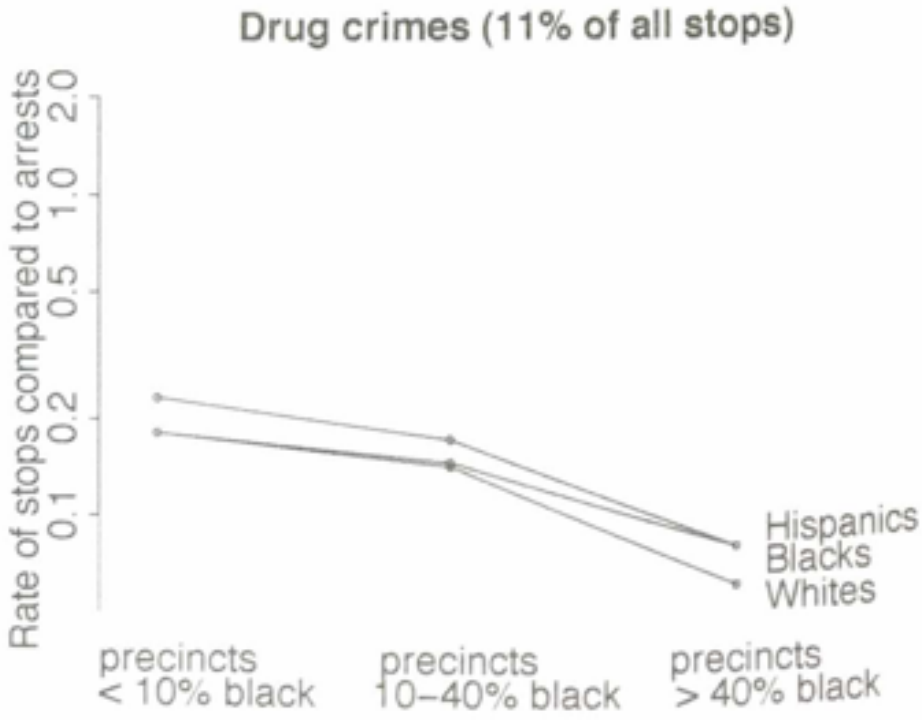
Proportion black in precinct	Parameter	Crim		Drug
$< 10\%$	intercept, μ^{ad}			-1.62 (0.16)
	α_1^{adj} [blacks]			-0.08 (0.09)
	α_2^{adj} [hispani			0.17 (0.10)
	α_3^{adj} [whites]			-0.08 (0.09)
	σ_β			0.87 (0.16)
	σ_ϵ			0.50 (0.07)
10–40%	inte			7 (0.13)
	α_1^{adj}			.05 (0.05)
	α_2^{adj}			0.12 (0.06)
	α_3^{adj}			-0.07 (0.05)
	σ_β			0.90 (0.13)
	σ_ϵ			0.32 (0.04)
$> 40\%$	intercep			2.62 (0.12)
	α_1^{adj} [b			.09 (0.06)
	α_2^{adj} [h		(0.07)	0.09 (0.07)
	α_3^{adj} [wh		1 (0.07)	-0.18 (0.09)
	σ_β	47 (0.11)	0.96 (0.18)	0.54 (0.11)
	σ_ϵ	0.24 (0.37 (0.05)	0.42 (0.07)	0.28 (0.06)

TABLES ARE IMPOSSIBLE TO READ


Violent crimes (25% of all stops)
Rate of stops compared to arrests
0.1
0.2
0.5
1.0
2.0
Blacks
Hispanics
Whites
precincts
< 10% black
precincts
10–40% black
precincts
> 40% black


Weapons crimes (44% of all stops)
Rate of stops compared to arrests
0.1
0.2
0.5
1.0
2.0
Blacks
Hispanics
Whites
precincts
< 10% black
precincts
10–40% black
precincts
> 40% black


Property crimes (20% of all stops)
Rate of stops compared to arrests
0.1
0.2
0.5
1.0
2.0
Hispanics
Whites
Blacks
precincts
< 10% black
precincts
10–40% black
precincts
> 40% black


Drug crimes (11% of all stops)
Rate of stops compared to arrests
0.1
0.2
0.5
1.0
2.0
Hispanics
Blacks
Whites
precincts
< 10% black
precincts
10–40% black
precincts
> 40% black

Model sensitivity checks

So far, we see some trends. We might ask

- ★ Did our results depend on how we divided precincts?
- ★ How much would things change if we tweaked our analysis?

Answering these questions are crucial for separating out what data is saying from what model *makes* the data say.

Alternative model for the same questions: Investigating Variation by Including Level-2 predictors

$$y_{ep} \sim \text{Poisson}\left(\frac{15}{12} n_{ep} e^{\mu + \alpha_e + \zeta_1 z_{1p} + \zeta_2 z_{2p} + \beta_p + \epsilon_{ep}}\right)$$

Z_{1p}, Z_{2p} are proportion of black and hispanic people, respectively, in the precinct.

Still have random effect for precinct: variation beyond what we can explain with the demographic information.

This assumes linear relationships. We could extend model to have higher order terms, e.g., Z_{1p}^2 or Z_{1p} x Z_{2p}

Results including our predictor model

Parameter	Crime type: Violent	Weapons	Property	Drug
intercept, μ^{adj}	−0.66 (0.08)	0.08 (0.11)	−0.14 (0.24)	−0.98 (0.17)
α_1^{adj} [blacks]	0.41 (0.03)	0.24 (0.03)	−0.19 (0.04)	−0.02 (0.04)
α_2^{adj} [hispanics]	0.10 (0.03)	0.12 (0.03)	0.23 (0.04)	0.15 (0.04)
α_3^{adj} [whites]	−0.51 (0.03)	−0.36 (0.03)	−0.05 (0.04)	−0.13 (0.04)
ζ_1 [coeff for prop. black]	−1.22 (0.18)	0.10 (0.19)	−1.11 (0.45)	−1.71 (0.31)
ζ_2 [coef for prop. hispanic]	−0.33 (0.23)	0.71 (0.27)	−1.50 (0.57)	−1.89 (0.41)
σ_β	0.40 (0.04)	0.43 (0.04)	1.04 (0.09)	0.68 (0.06)
σ_ϵ	0.25 (0.02)	0.27 (0.02)	0.37 (0.03)	0.37 (0.03)

Table shows coefficients with standard errors.

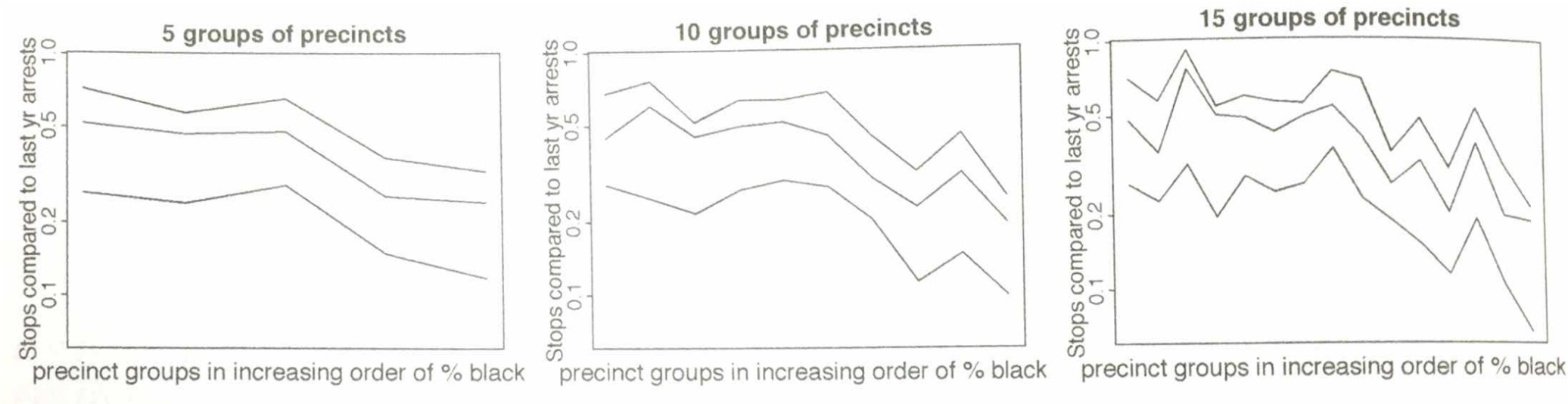
Each column is a different outcome (count of given suspected crime)

The *model* is the same across the four runs.

We see:

- ★ Stop *rate* goes down for violent crime as precinct becomes "more black"
- ★ Black stopped more for violent, weapons. Hispanic more for property, drug.

Another sensitivity check: Looking at different subsetting



- ★Chop precincts into different datasets, and then analyzed those datasets independently.
- ★These are extensions (sensitivity checks) on the three-subset model shown before.
- ★For each analysis, the different analyses are connected with lines to ease comparison.

Different models / exploration

The text (G&H 15.1) explores many different models, all related.

For example, grouping precincts by similarity (e.g., % black) and model different patterns within those groups, **or** including % black as a predictor in a larger model.

If the models all give similar answers, we are comforted that **model misfit/dependence** is not a real concern.

We don't want to think our findings are a product of our **estimation strategy**. We want them to be a product of our **data**

No true crime:
glimpse of an advanced way of handling our lack of true crime rate

How do we fix past arrests not being a good proxy of actual crime?

In our original models we used past arrests as a proxy for the actual crime rate for the different groups.

We have not addressed uncertainty in this proxy.

To do so:

- ★ We model it as an outcome dependent on an *unobserved* crime rate.
- ★ We then use the true unobserved crime rate in our original model.

This looks like this...

A two-stage model

$$
\begin{aligned}
y_{ep} &\sim \text{Poisson}\left(\tfrac{15}{12}\theta_{ep} e^{\mu+\alpha_e+\beta_p+\epsilon_{ep}}\right) \\
n_{ep} &\sim \text{Poisson}(\theta_{ep}) \\
\log \theta_{ep} &= \log N_{ep} + \tilde{\alpha}_e + \tilde{\beta}_p + \tilde{\epsilon}_{ep}.
\end{aligned}
$$

This models rate of stops given "rate of crime."

Number of arrests depends on rate of crime.

This models "rate of crime" which gives past arrests

Our **exposure** is now a **latent variable**.

Our observed *proxy* for exposure, n_{ep}, is also Poisson and a function of our latent variable.

We can model our latent variable with their own error terms for ethnic group and precinct.

(N_{ep} is the total population of the area)

How would you fit such a model?

These models are not “classic” in that we have latent parameters where we typically put observed data.

This means:

- ★ glmer() does not apply anymore.
- ★ We need new methods for fitting.

This is where general modeling packages come in. In particular, STAN!

We will briefly discuss **fitting** such models later in the course.

Looking back

Take-away from G&H Section 15.1

We described themes and patterns using models to interrogate whether there was evidence of elevated rates of policing for some demographic groups.

A thorough analysis will explore many different models of the same (or related) thing.

You can subset data in different ways to

1. examine model dependence or
2. explore different facets of your data.

Check-In
http://cs179.org/lec54

This case study illustrates creative model building, and is worth scrutiny for how one might

- customize models for a given context and
- use models descriptively.

Supplement (if time)
The Gelman & Hill Way:
"Let's partially pool everything."

Our overdispersed Poisson multilevel model with ethnicity partially pooled

$$
\begin{aligned}
y_{ep} &\sim \text{Poisson}\left(\tfrac{15}{12} n_{ep} e^{\mu+\alpha_e+\beta_p+\epsilon_{ep}}\right) \\
\alpha_e &\sim \text{N}(0, \sigma_\alpha^2) \\
\beta_p &\sim \text{N}(0, \sigma_\beta^2) \\
\epsilon_{ep} &\sim \text{N}(0, \sigma_\epsilon^2),
\end{aligned}
$$

"ethnic effect"

random precinct effect

"residual noise" to decouple variance and count

n_{ep}: # arrests recorded by DCJS for that group in that precinct in that year (multiplied by 15/12 to get to a 15-month period for comparability).

With this model, our intercept is overall arrest rate (across all groups). We have no baseline group.

Gelman & Hill love to partially pool everything. Here they even do the ethnic groups! I wouldn't.

Fitting the G&H Model

```
> Mall = glmer( stops ~ 1 + (1|eth) + (1|id)
                        + (1|precinct),
                offset = log(past.arrests),
                family = poisson(link="log"),
                data=stops )
>
> display( Mall )
coef.est  coef.se
   -0.65      0.16

Error terms:
 Groups    Name          Std.Dev.
 id        (Intercept)   0.29
 precinct  (Intercept)   0.61
 eth       (Intercept)   0.24
 Residual                1.00
---
number of obs: 225, groups: id, 225; precinct, 75; eth, 3
AIC = 2910, DIC = -2860.9
deviance = 20.5
```

But our parameters of interest for ethnic group are missing!

We need to extract them as they are random effects. (Only three of them, but still...)

This is our overdispersion measure now.

Extracting ethnic group effects when they are random effects

```
> ranef( Mall )$eth
          (Intercept)
Black       0.1571785
Hispanic    0.1758692
White      -0.3263795

> fixef( Mall )
(Intercept)
 -0.6528433

> alphas = ranef( Mall )$eth$`(Intercept)`

> alpha.bar = mean( alphas )
> alpha.adj = alphas - mean( alphas )
> alpha.adj
[1]  0.1549558  0.1736464 -0.3286022

> mu.adj = fixef(Mall) + alpha.bar
> mu.adj
(Intercept)
 -0.6506205
```

This code recenters our random intercepts and our grand mean so our groups deviations all average out to zero
Our model says they do on average. This enforces it.

Why make race/ethnicity a random effect?

Partially pooling coefficients brings them closer together.

Ideally, it reduces noise, making them easier to interpret.

We are comforted that they are not far apart due to measurement error, but rather something structural.

This is the Bayesian approach in spirit.