STA2201H Methods of Applied Statistics II

Monica Alexander

Week 2: Generalized Linear Models

Overview

Over two weeks

Lecture:

- General Linear Models
- Generalized Linear Models
- Exponential family
- Likelihood-based estimation and inference
- Poisson
- Binomial, Multinomial
- Survival analysis

Lab:

- EDA
- GLM in R

The model fitting process

What are we actually trying to achieve? From last week, applied statistics is:

Using statistical methods to answer questions and draw reasonable conclusions from data that have uncertainty and randomness.

The model fitting process

Overview of process

1. Look at the data (EDA, today's lab)
2. Decide on a model
   - Probability distribution for response Y e.g. $Y \sim N(\mu, \sigma^2)$
   - Equation involving explanatory variables (we are trying to explain $E[Y]$)
3. Estimate the parameters
4. Check the model and residuals
5. Inference, interpretation
6. Communication

Motivating examples

Outcomes we may be interested in investigating (in relation to other explanatory variables):

- Police stop and frisks in NYC
- Infant deaths in the US
- Who voted for the Liberal party v other party
- Who voted Liberal, Conservatives, LDP
- Concentration of drug at particular times after ingestion

General linear models

We observe $y_1, y_2, \ldots y_n$ which are realizations of the random variables $Y_1, Y_2, \ldots, Y_n$

In linear models the y_i's have two pieces:

1. A **systematic part**, with the form

$$E(\mathbf{Y}) = \mu = \mathbf{X}\beta$$

2. A **random part**, where errors are assumed to be i.i.d such that $E[\epsilon] = 0$ and $var[\epsilon] = \sigma^2$. We usually further assume that errors are Normal with constant variance σ^2.

General linear models

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \mathbf{x}_i^T \beta \end{aligned}$$

General linear models are not appropriate when

- The range of Y is restricted
- The variance of Y depends on the mean

Generalized Linear Models extend the classical set-up to allow for a wider range of distributions. Introduced by Nelder and Wedderburn (1972) [Later, GAMs in 1990].

Generalized linear models

GLMs have an additional piece on top of the classical linear models:

1. **random component**: $Y_i \sim$ some distribution with $E[Y_i] = \mu_i$
2. **systematic component**: $\mathbf{x}_i^T \beta$
3. The **link function** that links the random and systematic components $g(u_i) = \mathbf{x}_i^T \beta$

- Set-up is almost the same, particularly in terms of specifying a good linear predictor $\mathbf{x}_i^T \beta$
- Just need to think about the link and the distribution of the outcome

GLMs

$$\begin{aligned} Y_i &\sim G(\mu_i, \phi) \\ E[Y_i] &= \mu_i \\ g(\mu_i) &= \mathbf{x}_i^T \beta \end{aligned}$$

- ϕ is the scale parameter.

What can Y be distributed as? In principle, anything. In practice (and original formulation), distributions come from the **exponential family**.

Exponential Family

Exponential Family

The random variable Y belongs to the exponential family of distributions if its support does not depend upon any unknown parameters and its density or probability mass function takes the form

$$p(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

- $\theta = h(\mu)$ depends on the expected value of y and is the **canonical parameter**
- ϕ is the scale parameter (if known: one-parameter family)
- b and c are arbitrary functions

Example: Poisson distribution

$$p(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

Poisson:

$$p(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

Write as

$$p(y|\mu) = \exp\{y \log \mu - \mu - \log y!\}$$

- $\theta = \log \mu$
- $b(\theta) = e^\theta$
- $c(y,\phi) = -\log y!$
- Note that the scale parameter $\phi = 1$ so the variance is entirely determined by the mean

Example: Normal distribution

$$p(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

Normal:

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

Write as

$$f(y|\mu,\sigma^2) = \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\}$$

- $\theta = \mu$
- $b(\theta) = \frac{1}{2}\theta^2$
- $\phi = \sigma^2$
- $c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]$

Other examples

Other common examples:

- Binomial
- Gamma
- Negative binomial
- Inverse Gaussian

Propertites of exponential families

Mean and variance for exponential families

$$p(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

It can be shown that

$$E(Y|\theta,\phi) = b'(\theta) = \mu$$

and

$$Var(Y|\theta,\phi) = \phi b''(\theta) = \phi V(\mu)$$

Note the variance of Y depends not only on the scale parameter but also on a function of the mean.

Examples:

$$E(Y|\theta,\phi) = b'(\theta)$$

and

$$Var(Y|\theta,\phi) = \phi b''(\theta)$$

- Poisson: $E(Y|\theta,\phi) = e^{\theta} = \mu$, $Var(Y|\theta,\phi) = 1 \times e^{\theta} = \mu$
- Normal: $E(Y|\theta,\phi) = \theta = \mu$, $Var(Y|\theta,\phi) = \sigma^2 \times 1 = \sigma^2$

The canonical link

The link function $\eta_i = g(\mu)$ could in theory be any function linking the linear predictor to the distribution of the outcome variable, which is also is **monotonic** and **smooth**.

Recall $\theta = h(\mu)$. If we choose $g = h$, then

$$\theta_i = h(\mu_i) = h(h^{-1}(\eta_i)) = \eta_i = \mathbf{x}_i^T \beta$$

In other words, it ensures that the systematic component of our model is modeling the parameter of interest.

Canonical links

- Normal: identity $\theta = h(\mu) = \mu$
- Poisson: $\theta = h(\mu) = \log \mu$
- Binomial: $\theta = h(\mu) = \log(\frac{\mu}{1-\mu})$

Likelihood-based estimation and inference

Estimation

- Inference is based on MLE, but cannot derive closed form solutions for regression coefficients
- Note we are assuming independence $cov(Y_i, Y_j|\theta_i, \theta_j, \phi) = 0$ for $i \neq j$. (more on dependence later)

The log-likelihood function is:

$$\ell(\theta) = \sum_i \ell(\theta_i) = \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi)$$

What's our usual approach here?

Score function and Information matrix

$$\mathbf{S}(\beta) = \mathbf{D^T V^{-1}} \frac{\mathbf{Y} - \mu(\beta)}{\phi}$$

where D^T is a matrix of the $\partial \mu_i / \partial \beta_j$ and $\mathbf{V}$ is diagonal with ith element $b''(\theta_i)$.

$$\mathbf{I}(\beta) = \mathbf{x^t W}(\beta)\mathbf{x}$$

where $\mathbf{W}$ is diagonal with $w_i = (\frac{\partial \mu_i}{\partial \eta_i})^2 / \phi b''(\theta_i)$.

What about ϕ?

When ϕ is unknown, can estimate it using

$$\hat{\phi} = \frac{1}{n-k-1} \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu})}$$

where $\hat{\mu} = \hat{\mu}(\hat{\beta})$.

Newton-Raphson

Want to find roots such that $\mathbf{S}(\beta) = 0$. First order TS approximation:

$$\mathbf{S}(\beta) \approx \mathbf{S}(\beta^{(\mathbf{0})}) + (\beta - \beta^{(\mathbf{0})})^T \mathbf{S}'(\beta^{(\mathbf{0})})$$

Newton-Raphson iterates the step:

$$\beta^{(\mathbf{t+1})} = \beta^{(\mathbf{t})} - \mathbf{S}'(\beta^{(\mathbf{t})})^{-1} \mathbf{S}(\beta^{(\mathbf{t})})$$

Method of scoring replaces observed information with its expectation $\mathbf{E}[\mathbf{S}'(\beta)] = -\mathbf{I}(\beta)$.

$$\beta^{(\mathbf{t+1})} = \beta^{(\mathbf{t})} + \mathbf{I}(\beta^{(\mathbf{t})})^{-1} \mathbf{S}(\beta^{(\mathbf{t})})$$

Estimation

Can be rewritten in the form:

$$\widehat{\beta}^{(\mathbf{t+1})} = (\mathbf{x}^{\mathbf{T}}\mathbf{W}\mathbf{x})^{-\mathbf{1}}\mathbf{x}^{\mathbf{T}}\mathbf{W}\mathbf{z}$$

where:

$$z_i = x_i\beta + (Y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

- **W** and **z** change depending on $\hat{\beta}$ and vice versa
- Use iteratively weighted least squares (IWLS)
  1. Choose initial value $\hat{\beta}^{(0)}$
  2. Calculate **W** and **z**
  3. Repeat until convergence

Inference

We know that for the MLE

$$\hat{\beta} \sim N(\beta, (\mathbf{X^T W X})^{-1}\phi)$$

$$\hat{\beta} \sim N(\beta, I(\hat{\beta})^{-1})$$

Standard errors are the square roots of the inverse of the information matrix.

- Use this for the classic Wald Tests e.g. $\sqrt{W} = \frac{\hat{\beta}-\beta_0}{se(\hat{\beta})}$ follows z distribution.

Likelihood ratio test

Testing nested models, ω_1 and ω_2, $\omega_1 \in \omega_2$ and number of parameters $p_2 > p1$

$$2[\log \mathbf{L}(\widehat{\beta_1}|\mathbf{y}) - \log \mathbf{L}(\widehat{\beta_2}|\mathbf{y})] \sim \chi_{p_1 - p_2}$$

Compared to Wald: doesn't assume symmetry in confidence intervals, but requires you to run two models.

GLM in R

- `glm()`
- same set up as `lm()`; additional `family` argument with a link
- e.g. `glm(y~x, family = binomial(link = 'logit')`

Poisson regression

Review

- mean ?
- variance ?
- link: ?

What's a problem with just looking at counts?

Offsets

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(\lambda_i) \\
\text{or } Y_i &\sim \text{Poisson}(\mu_i O_i) \\
\implies \log \mu_i &= \mathbf{x_i}^T \beta
\end{aligned}
$$

Offset controls for exposure to risk/making inferences to some baseline. e.g.

- population size
- age
- time since exposed

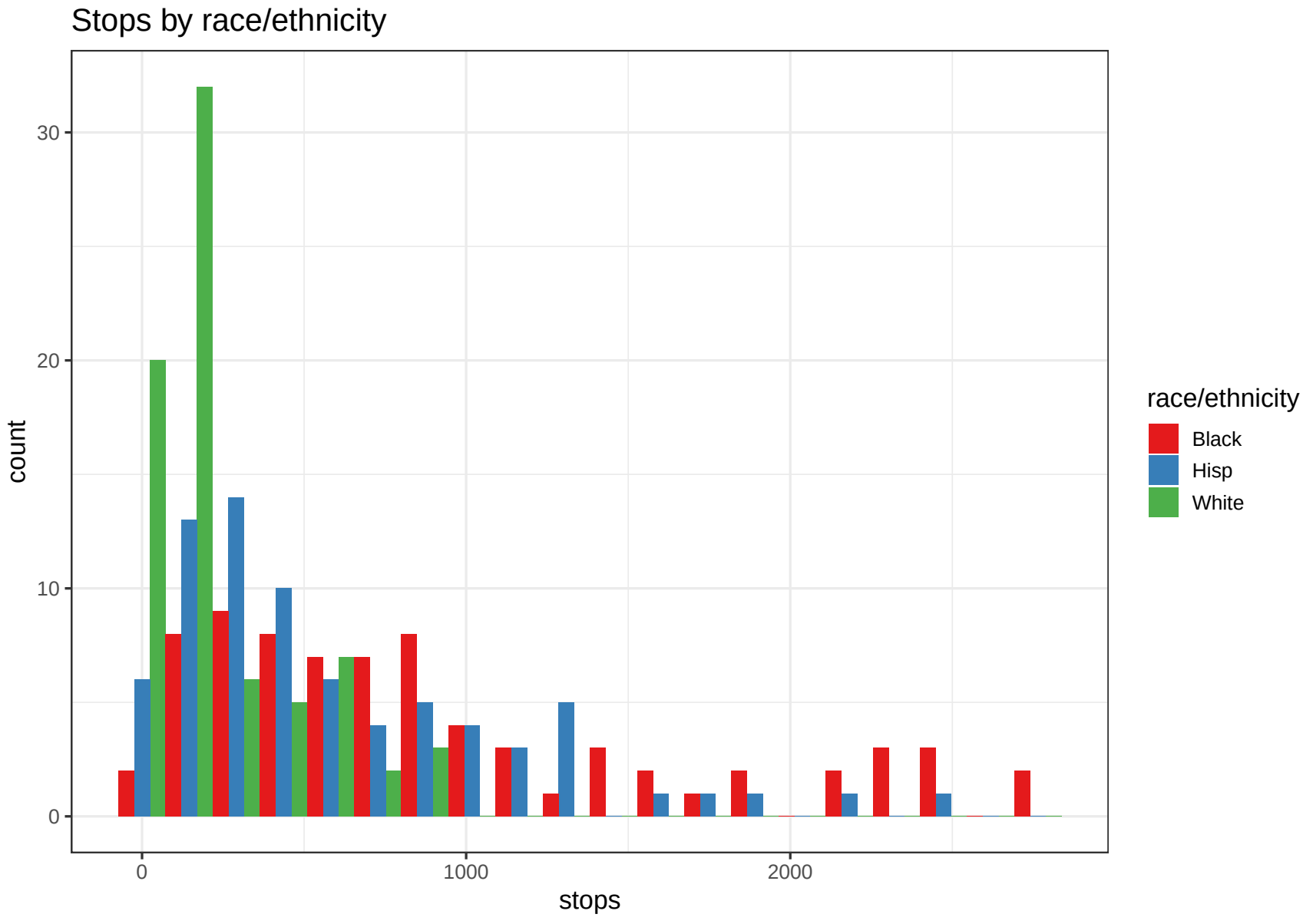
Example: Police stops

Police stop and frisks in NYC (Gelman Hill Chapter 6). Is there a difference in the number of stops by race/ethnicity?

The data look like:

precinct	stops	arrests	race_eth
1	202	980	Black
1	102	295	Hisp
1	81	381	White
2	132	753	Black
2	144	557	Hisp
2	71	431	White
3	752	2188	Black
3	441	627	Hisp
3	410	1238	White
4	385	471	Black

Distribution



GLM

Use arrests as exposure

```
mod1 <- glm(stops~race_eth,family=poisson,offset=log(arrests),data=d)
summary(mod1)
```

```
##
## Call:
## glm(formula = stops ~ race_eth, family = poisson, data = d, offset = log(arrests))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -47.327   -7.740   -0.182   10.241   39.140
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.588086   0.003784 -155.40   <2e-16 ***
## race_ethHisp   0.070208   0.006061   11.58   <2e-16 ***
## race_ethWhite -0.161581   0.008558  -18.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 46120  on 224  degrees of freedom
## Residual deviance: 45437  on 222  degrees of freedom
## AIC: 47150
##
## Number of Fisher Scoring iterations: 5
```

GLM

Add in factors for precinct

```
mod2 <- glm(stops~race_eth + factor(precinct), family=poisson,offset=log(arrests),data=d)
summary(mod2)[["coefficients"]][1:10,]
```

```
##                       Estimate  Std. Error    z value      Pr(>|z|)
## (Intercept)       -1.37886803 0.051019006 -27.026556 7.205634e-161
## race_ethHisp       0.01018798 0.006802045   1.497782  1.341899e-01
## race_ethWhite     -0.41900122 0.009434996 -44.409261  0.000000e+00
## factor(precinct)2 -0.14904964 0.074030344  -2.013359  4.407691e-02
## factor(precinct)3  0.55995498 0.056758425   9.865583  5.869222e-23
## factor(precinct)4  1.21063605 0.057548994  21.036615  3.032678e-98
## factor(precinct)5  0.28286532 0.056794015   4.980548  6.340447e-07
## factor(precinct)6  1.14420375 0.058047383  19.711547  1.716374e-86
## factor(precinct)7  0.21817307 0.064335032   3.391202  6.958688e-04
## factor(precinct)8 -0.39056473 0.056867814  -6.867940  6.513564e-12
```

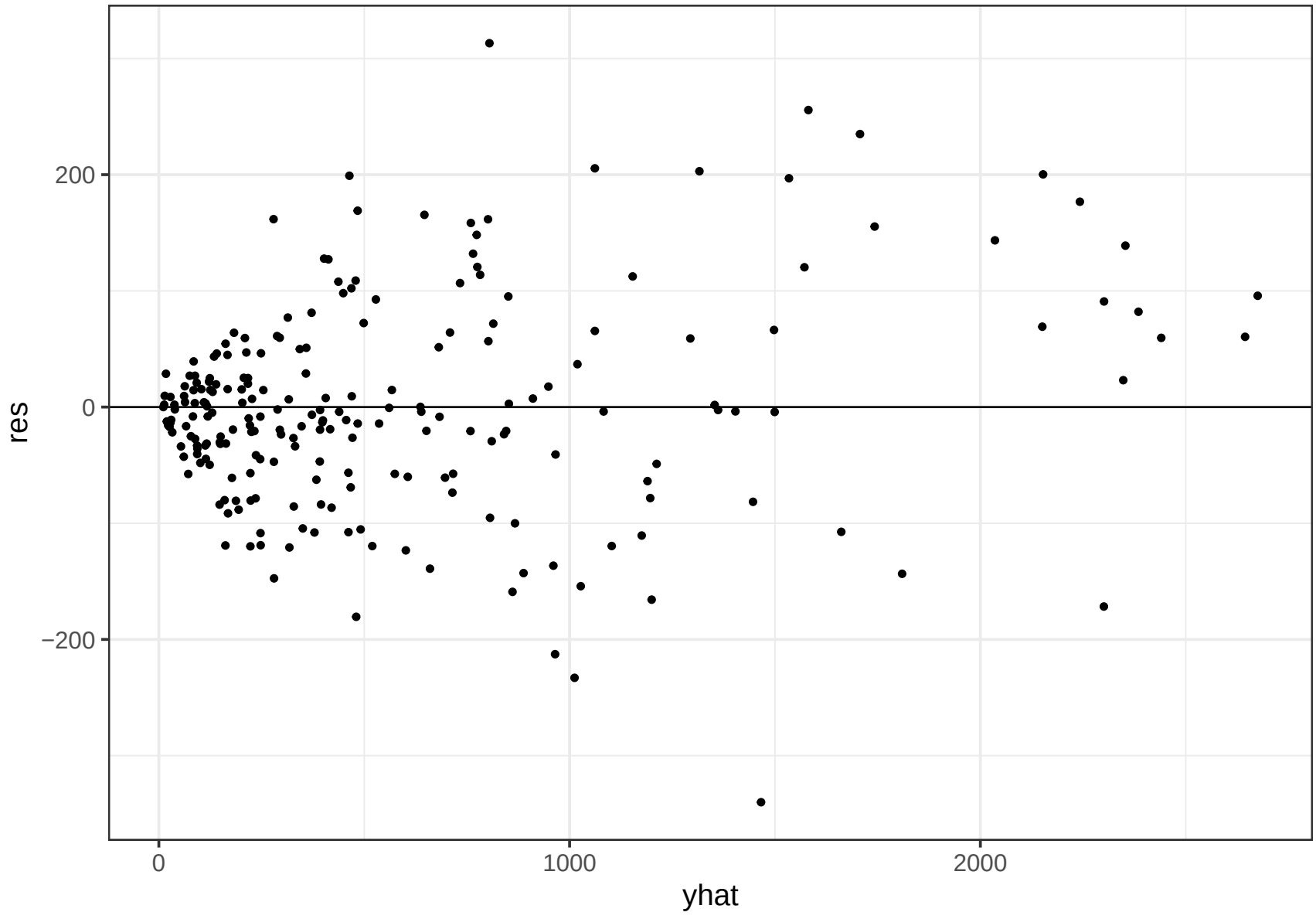
Coefficient interpretation

- e.g. after controlling for precinct, compared to blacks, whites have $1 - exp(-0.42) = 34\%$ less chance of being stopped.
- be wary of exposure variable: stops are compared to the number of arrests in the previous year
- so that the coefficient 'whites' will be less than 1 if the people in that group are stopped disproportionately less than their rates of arrest, as compared to blacks.
- would be different if we had population as exposure variable

Is this a reasonable model?

Look at predicted values versus residuals $(y_i - \hat{y}_i)$. What do we expect?

Predicted values versus residuals



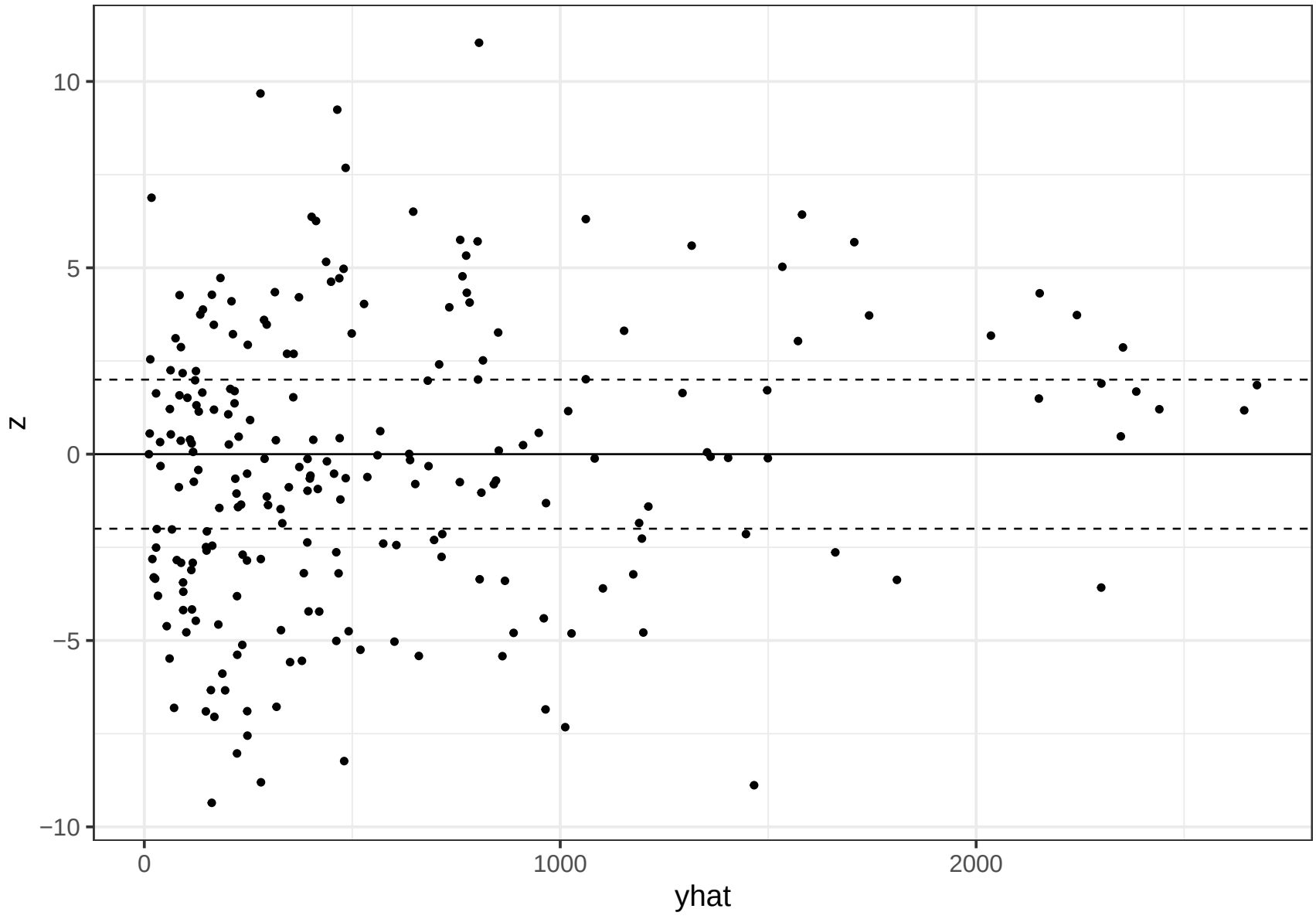
Is this a reasonable model?

Consider standardized residuals

$$z_i = \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)}$$

If Poisson is a good model then these should have mean 0 and sd 1.

Predicted values versus standardized residuals



Overdispersion

- Extra variation in the data beyond what is allowed for in statistical model
- Poisson does not have independent variance parameter

Test for overdispersion: compare sum of squares of standardized residuals to χ^2_{n-k} distribution.

Estimated overdispersion factor is

$$\frac{1}{n-k}\sum_i z_i^2$$

Overdispersion

overdispersion factor is

```
sum(res_df$z^2)/(n-k)
```

```
## [1] 21.88505
```

P-value of test is

```
pchisq(sum(res_df$z^2), n-k)
```

```
## [1] 1
```

But what's a problem here?

Fit overdispersed Poisson

- General form includes extra dispersion parameter θ
- Assume variance is proportion to the mean, rather than equal to the mean $E[Y] = \mu\theta$

```
mod3 <- glm(stops~race_eth + factor(precinct), family=quasipoisson,offset=log(arrests),data=d)
summary(mod3)[["coefficients"]][1:10,]
```

```
##                       Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       -1.37886803 0.23867441 -5.7771925 4.326149e-08
## race_ethHisp       0.01018798 0.03182097  0.3201657 7.492943e-01
## race_ethWhite     -0.41900122 0.04413830 -9.4929170 5.489337e-17
## factor(precinct)2 -0.14904964 0.34632483 -0.4303753 6.675488e-01
## factor(precinct)3  0.55995498 0.26552425  2.1088656 3.664011e-02
## factor(precinct)4  1.21063605 0.26922265  4.4967837 1.384310e-05
## factor(precinct)5  0.28286532 0.26569075  1.0646412 2.887722e-01
## factor(precinct)6  1.14420375 0.27155419  4.2135374 4.352372e-05
## factor(precinct)7  0.21817307 0.30096874  0.7249028 4.696562e-01
## factor(precinct)8 -0.39056473 0.26603599 -1.4680898 1.442019e-01
```

Notice

```
summary(mod3)[["dispersion"]]
```

```
## [1] 21.88506
```

. . . and the SEs are inflated $\sim \sqrt{21.9}$.

Overdisperson

Downside to quasi-Poisson it's not true MLE so you don't get likelihood etc to compare models.

Alternative:

- Could also add a multiplicative random effect θ to represent unobserved heterogeneity.
- Conditional distribution is Poisson $E[Y|\theta] \sim Pois(\mu\theta)$
- Leads to unconditional distribution being Negative Binomial distribution
- Can choose parameters so $E(Y) = \mu$ and $Var(Y) = \mu(1 + \sigma^2\mu)$

Overdispersion

Fit Negative Binomial

```
library(MASS)
mod4 <- glm.nb(stops~race_eth + factor(precinct), data = d)
summary(mod3)[["coefficients"]][1:10,]
```

```
##                       Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       -1.37886803 0.23867441 -5.7771925 4.326149e-08
## race_ethHisp       0.01018798 0.03182097  0.3201657 7.492943e-01
## race_ethWhite     -0.41900122 0.04413830 -9.4929170 5.489337e-17
## factor(precinct)2 -0.14904964 0.34632483 -0.4303753 6.675488e-01
## factor(precinct)3  0.55995498 0.26552425  2.1088656 3.664011e-02
## factor(precinct)4  1.21063605 0.26922265  4.4967837 1.384310e-05
## factor(precinct)5  0.28286532 0.26569075  1.0646412 2.887722e-01
## factor(precinct)6  1.14420375 0.27155419  4.2135374 4.352372e-05
## factor(precinct)7  0.21817307 0.30096874  0.7249028 4.696562e-01
## factor(precinct)8 -0.39056473 0.26603599 -1.4680898 1.442019e-01
```

Lab

- Using data from Open Data Portal in Toronto
  - `opendatatoronto` package by Sharla Gelfand
- EDA
- Questions at end need to be handed in via GitHub