Lecture 15

Motivating example: air pollution data

- Data on Chicago air quality and death between 1987 and 2000
- Variables include:
  - deaths
  - ozone concentration
  - sulphur dioxide concentration
  - temperature

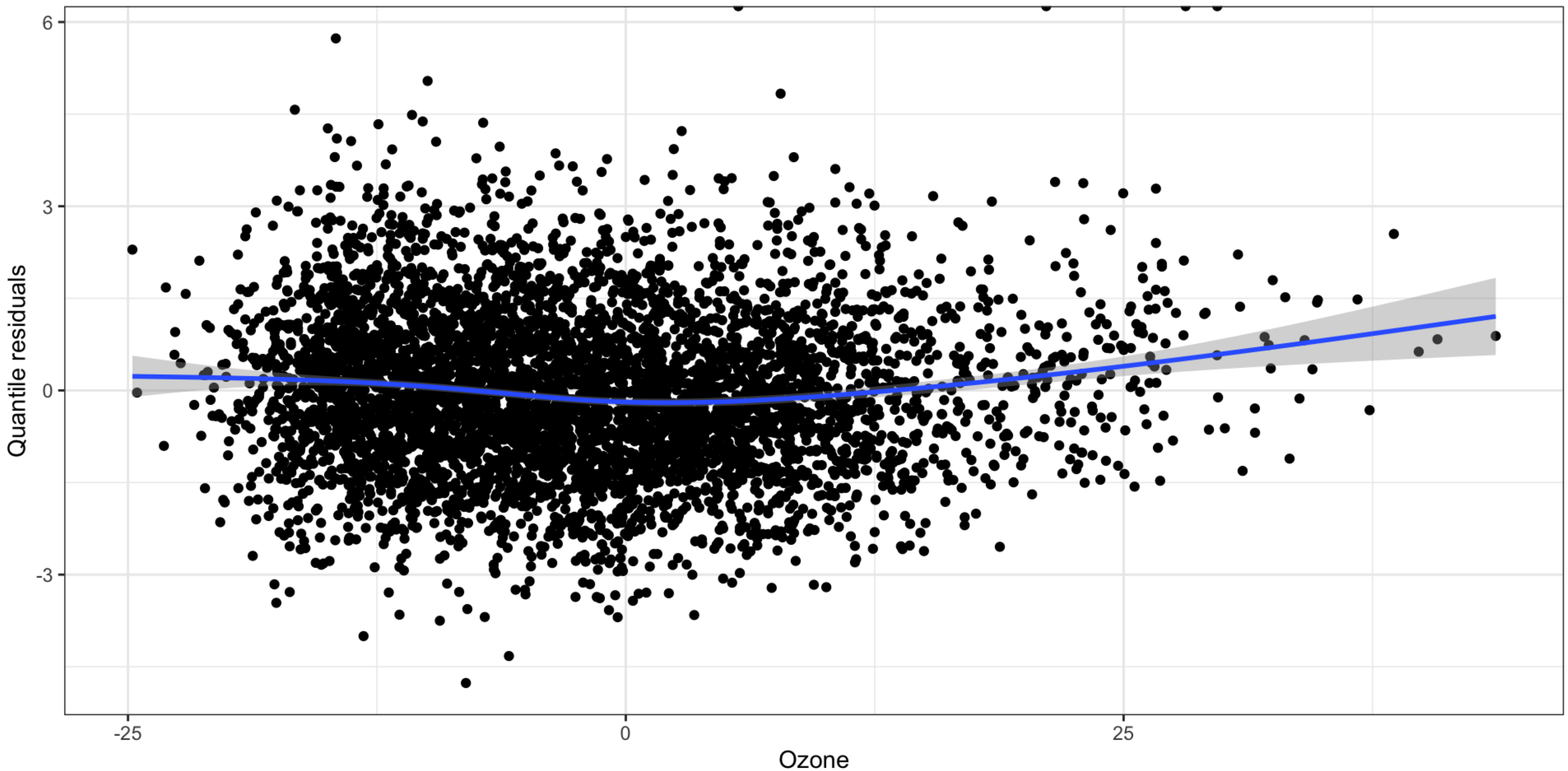
Motivating example: air pollution data

$$\text{Deaths}_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{Ozone}_i$$

Quantile residual plot

- variance of quantile residuals is ≈ constant, but too high



- nice random scatter around 0 (maybe slight pattern for high values Ozone)
- maybe some more variability than we want

GOF test

```
m1$deviance
```

[1] 9551.836

```
m1$df.residual
```

[1] 5112

```
pchisq(m1$deviance, m1$df.residual, lower.tail=F)
```

[1] 6.362106e-273

=> maybe Poisson distribution is not appropriate for Y_i

Overdisperion

Overdispersion occurs when the response variable $\mathrm{Y_i}$ has greater variability than the model accounts for

If $Y_i \sim \text{Poisson}(\lambda_i)$ then $Var(Y_i) = \lambda_i$

But if $Y_i \nsim$ Poisson, and we incorrectly assume it is Poisson, then $Var(Y_i)$ may be larger than the model accounts for

Recap: sandwich estimator for GLMs

$\hat{\beta}$ solves $u(\beta) = \dfrac{X^T(Y-\mu)}{\phi} = \dfrac{1}{\phi}\sum_{i=1}^{n}(Y_i-\mu_i)X_i = 0$

β^* solves $\mathbb{E}\left[\frac{1}{\phi}(Y_i-\mu_i)X_i\right] = 0$

$\sqrt{n}(\hat{\beta}-\beta^*) \xrightarrow{d} N(0, \underbrace{J_1(\beta^*)^{-1}}_{\text{bread}} \underbrace{V_1(\beta^*)}_{\text{meat}} J_1(\beta^*)^{-1})$

$\Rightarrow \hat{\beta} \approx N(\beta^*, J_n(\beta^*)^{-1} V_n(\beta^*) J_n(\beta^*)^{-1})$

$J_n(\beta^*) = -\mathbb{E}[U'(\beta^*)]$ $\quad V_n(\beta^*) = \mathrm{Var}(U(\beta^*))$

- If model is correct (both shape <u>and</u> distribution) then $J_n(\beta^*) = V_n(\beta^*)$
- If model is incorrect, $\hat{\beta}$ still has this asymptotic variance
- $\hat{\beta} \xrightarrow{p} \beta^*$

β^* = the coefficients for relationship <u>if</u> mean $\mu_i = g^{-1}(\beta^T X_i)$ is correct

Assumptions about both mean and variance

$$\frac{\partial \mu}{\partial \theta} = V(\mu)$$

Suppose we assume

$$\mathbb{E}[Y_i] = \mu_i = g^{-1}(\beta^T X_i)$$

$$Var(Y_i) = \phi V(\mu_i)$$

true if $Y_i \sim EDM(\mu_i, \phi)$

But we don't have to assume a distribution for Y_i, just the first 2 moments

$\hat{\beta}$ solves $u(\beta) = \frac{X^T(Y-\mu)}{\phi} = \frac{1}{\phi}\sum_{i=1}^{n}(Y_i-\mu_i)X_i = 0$

$$Var(u(\beta)) = \sum_{i=1}^{n} \frac{1}{\phi^2} \cdot Var(Y_i) X_i X_i^T$$

$$\| \quad = \sum_{i=1}^{n} \frac{1}{\phi} \cdot V(\mu_i) X_i X_i^T$$

$$-\mathbb{E}[u'(\beta)] = -\frac{1}{\phi}\sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial}{\partial \beta}(Y_i-\mu_i)X_i\right] = \frac{1}{\phi}\sum_{i=1}^{n} V(\mu_i) X_i X_i^T$$

If we correctly assume

$$\mathbb{E}[Y_i] = \mu_i = g^{-1}(\beta^T X_i)$$

$$\mathrm{var}(Y_i) = \phi V(\mu_i)$$

then

$$\hat{\beta} \approx N\left(\beta^*, \quad \phi\,(X^T W X)^{-1}\right)$$

$$W = \mathrm{diag}(V(\mu_i))$$

E.g. for

$$\log(\mu_i) = \beta^T X_i \qquad \text{(like in Poisson)}$$

$$\mathrm{Var}(Y_i) = \phi\,\mu_i \qquad \text{(like in Poisson)}$$

then

$$\hat{\beta} \approx N\left(\beta^*, \quad \underbrace{\phi\,(X^T W X)^{-1}}_{=\ \phi\,\mathrm{Var}(\hat{\beta})_{\text{Poisson model}}}\right)$$

$$W = \mathrm{diag}(\mu_i)$$

Quasi-Poisson models

$\log(\mu_i) = \beta^T X_i$ ← same as Poisson regression

$Var(Y_i) = \phi \mu_i$ ϕ is allowed to be > 1

$\phi \, Var_{poisson}$

$\Rightarrow \hat{\beta}_{QP} = \hat{\beta}_{Poisson}$

$\hat{Var}(\hat{\beta}_{QP}) = \hat{\phi} \, \hat{Var}(\hat{\beta}_{Poisson})$

↑

need to estimate ϕ...

Example: Chicago air quality

Poisson model:

```
               Estimate   Std. Error     z value      Pr(>|z|)
(Intercept)  4.743277988 0.0013382057 3544.50583 0.000000e+00
o3median    -0.002301345 0.0001285909   -17.89664 1.252641e-71
```

Quasi-Poisson model:

identical!

```
Call:
glm(formula = death ~ o3median, family = quasipoisson, data = chicago)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7432780  0.0018822 2520.02   <2e-16 ***
o3median    -0.0023013  0.0001809  -12.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.978347)

    Null deviance: 9873.8  on 5113  degrees of freedom
```

$\hat{\phi}$

$0.0001809 = 0.00012859\sqrt{1.978}$

Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_15.html