

Quasi-Poisson models

Recap: Overdispersion

Overdispersion occurs when the response Y has higher variance than we would expect if Y followed a Poisson distribution.

Formally, let

$$\hookrightarrow \phi = \frac{\text{Variance}}{\text{Mean}}$$

Dispersion parameter

For Poisson, we assume $\phi = 1$

Recap: Estimating overdispersion

The *Pearson residual* for observation i is defined as

$$e_{(P)i} = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

$$\hat{\phi} = \frac{\sum_{i=1}^n e_{(P)i}^2}{n - p}$$

- + p = number of parameters in model

Handling overdispersion

Overdispersion is a problem because our standard errors (for confidence intervals and hypothesis tests) are too low.

If we think there is overdispersion, what should we do?

- option 1: Use $\hat{\phi}$ to adjust standard errors
- option 2: Use a different model / distribution (e.g., negative binomial)

Adjusting the standard error

- + In our data, $\hat{\phi} = 1.83$
- + This means our variance is 1.83 times bigger than it should be
- + So our standard error is $\sqrt{1.83} = 1.35$ times bigger than it should be

$$SE(\hat{\beta})_{\text{overdispersion}} = \sqrt{\phi} \cdot SE(\hat{\beta})_{\text{Poisson}}$$

Poisson : assume that variance = mean
 \Rightarrow estimated standard deviation = $\sqrt{\text{estimated mean}}$

But if there is overdispersion, variance = ϕ mean

$$\Rightarrow \text{standard deviation} \underset{\sqrt{\phi} \sqrt{\text{est. mean}}}{\text{should}} \text{ be} = \sqrt{\phi} (\text{std.dev. for Poisson model})$$

Quasi-Poisson: Fit Poisson model, but inflate SEs by $\sqrt{\hat{\phi}}$

Adjusting the standard error in R

```
m2 <- glm(art ~ ., data = articles,  
          family = quasipoisson)
```

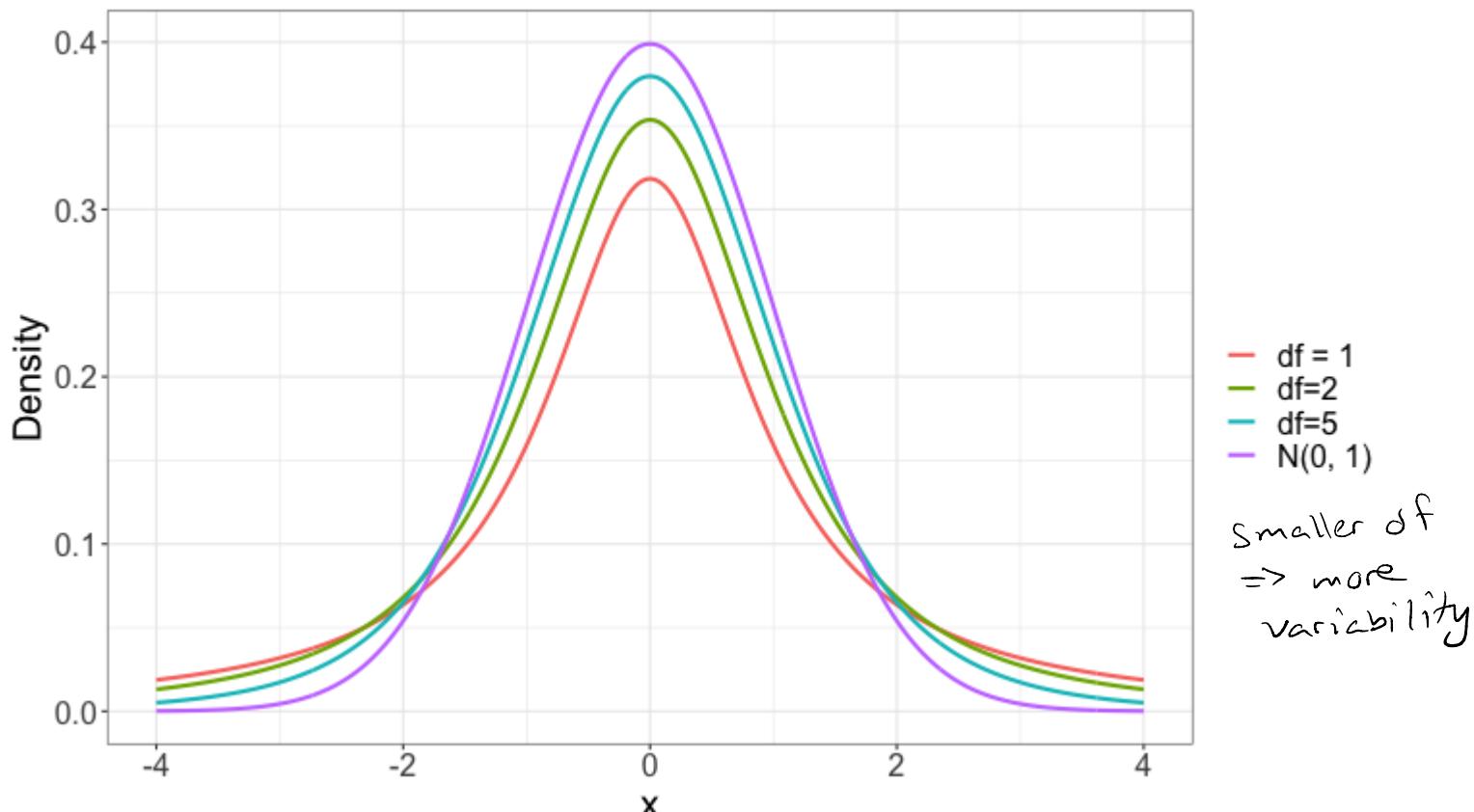
```
...  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.304617  0.139273  2.187 0.028983 *  
## femWomen    -0.224594  0.073860 -3.041 0.002427 **  
## marMarried   0.155243  0.083003  1.870 0.061759 .  
## kid5        -0.184883  0.054268 -3.407 0.000686 ***  
## phd         0.012823  0.035700  0.359 0.719544  
## ment        0.025543  0.002713  9.415 < 2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for quasipoisson family taken to be 1.829006)  
...
```

when we use $\hat{\phi}$,
we get t distributions
instead $N(0,1)$

- + Allowing ϕ to be different from 1 means we are using a *quasi-likelihood* (in this case, a *quasi-Poisson*)

$$\hat{\phi} = 1.83$$

Intuition: we're estimating ϕ . The estimate $\hat{\phi}$ has variability
⇒ more variability in test statistics {
confidence intervals}



Family of distributions t_v parameterized by degrees of freedom v

As $v \rightarrow \infty$, $t_v \rightarrow N(0, 1)$

Calculating a confidence interval

```
...  
## (Intercept) 0.304617  0.139273  2.187 0.028983 *  
## femWomen    -0.224594  0.073860  -3.041 0.002427 **  
## marMarried   0.155243  0.083003  1.870 0.061759 .  
## kid5        -0.184883  0.054268  -3.407 0.000686 ***  
## phd          0.012823  0.035700  0.359 0.719544  
## ment         0.025543  0.002713  9.415 < 2e-16 ***  
...
```

New confidence interval for β_4 :

$$\hat{\beta}_4 \pm t_{n-p}^* \text{SE}(\hat{\beta}_4)$$

$$0.0128 \pm t_{n-p}^* \cdot 0.0357$$

t_{n-p}^* = critical value for t distribution with $n-p$ df

```
qt(0.025, df=909, lower.tail=F) ← t*90%
```

```
## [1] 1.962577
```

$$0.0128 \pm 1.96 \cdot 0.0357 = (-0.0572, 0.0828)$$

Adjusting the standard error in R

Poisson:

$$0.04 (\sqrt{1.83}) \approx 0.054$$

...

	Estimate	Std. Error	z value	Pr(> z)	
## (Intercept)	0.304617	0.102981	2.958	0.0031	**
## femWomen	-0.224594	0.054613	-4.112	3.92e-05	***
## marMarried	0.155243	0.061374	2.529	0.0114	*
## kid5	-0.184883	0.040127	-4.607	4.08e-06	***

...

Quasi-Poisson:

$$\text{some } \hat{\beta} \text{s} \quad \text{SE}(\hat{\beta})_{QP} = \sqrt{\hat{\phi}} \text{ SE}(\hat{\beta})_{Poisson}$$

...

	Estimate	Std. Error	t value	Pr(> t)	
## (Intercept)	0.304617	0.139273	2.187	0.028983	*
## femWomen	-0.224594	0.073860	-3.041	0.002427	**
## marMarried	0.155243	0.083003	1.870	0.061759	.
## kid5	-0.184883	0.054268	-3.407	0.000686	***

...

Back to simulations

```
n <- 1000
nsim <- 500
contains_beta <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 0.5)
  y2 <- rnbinom(n, size=0.5, mu=exp(x))
  m2 <- glm(y2 ~ x, family = poisson) ← change to quasipoisson
  upper <- summary(m2)$coefficients[2,1] +
    ↗ 1.96*summary(m2)$coefficients[2,2]
  lower <- summary(m2)$coefficients[2,1] -
    1.96*summary(m2)$coefficients[2,2]
  contains_beta[i] <- upper > 1 && lower < 1
}
mean(contains_beta)

## [1] 0.642
```

Adjusting for overdispersion

```
n <- 1000
nsim <- 500
contains_beta <- rep(0, nsim)
for(i in 1:nsim){
  x <- rnorm(n, sd = 0.5)
  y2 <- rnbinom(n, size=0.5, mu=exp(x))
  m2 <- glm(y2 ~ x, family = quasipoisson)
  upper <- summary(m2)$coefficients[2,1] +
    qt(0.025, n-2, lower.tail = F)*summary(m2)$coefficients[2,2]
  lower <- summary(m2)$coefficients[2,1] -
    qt(0.025, n-2, lower.tail = F)*summary(m2)$coefficients[2,2]
  contains_beta[i] <- upper > 1 && lower < 1
}
mean(contains_beta)
```

[1] 0.926

coverage improves when we use the
quasi-Poisson model

Class activity

https://sta214-s23.github.io/class_activities/ca_lecture_22.html

Class activity

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
## (Intercept)	-1.30445	0.34161	-3.818	0.000274	***
## regionMW	0.09754	0.48893	0.199	0.842417	
## regionNE	0.76268	0.42117	1.811	0.074167	.
## regionSE	0.87237	0.42175	2.068	0.042044	*
## regionSW	0.50708	0.50973	0.995	0.323027	
## regionW	0.20934	0.51242	0.409	0.684055	

...

What confidence interval should I calculate to compare western and central schools?

$$CI \text{ for } \beta_5 : \hat{\beta}_5 \pm SE(\hat{\beta}_5) t_{n-p}^*$$

Class activity

```
...  
## regionSE      0.87237      0.42175      2.068  0.042044 *  
## regionSW      0.50708      0.50973      0.995  0.323027  
## regionW       0.20934      0.51242      0.409  0.684055
```

...

$qt(0.025, 75, \text{lower.tail=F}) \leftarrow t_{n-p}^*$ $n=81, p=6$

```
## [1] 1.992102
```

95% confidence interval for β_5 :

$$0.209 \pm 1.99 \cdot 0.512 = (-0.81, 1.23)$$

95% confidence interval for e^{β_5} :

$$(e^{-0.81}, e^{1.23}) = (0.44, 3.42)$$

Comparing Poisson and quasi-Poisson

Poisson:

- + Mean = λ_i
- + Variance = λ_i

quasi-Poisson:

- + Mean = λ_i
- + Variance = $\phi \lambda_i \Rightarrow SE = \sqrt{\phi} \sqrt{\lambda_i}$
- + Variance is a linear function of the mean

What if we want variance to depend on the mean in a different way?

Introducing the negative binomial

If $Y_i \sim NB(\theta, p)$, then Y_i takes values $y = 0, 1, 2, 3, \dots$ with probabilities

size parameter
↓
 $P(Y_i = y) = \frac{(y + \theta - 1)!}{y!(\theta - 1)!} (1 - p)^\theta p^y$

- + $\theta > 0, p \in [0, 1]$
- + Mean = $\frac{p\theta}{1 - p} = \mu$
- + Variance = $\frac{p\theta}{(1 - p)^2} = \mu + \frac{\mu^2}{\theta}$
- + Variance is a *quadratic* function of the mean

Mean and variance for a negative binomial variable

If $Y_i \sim NB(\theta, p)$, then

- + Mean = $\frac{p\theta}{1 - p} = \mu$
- + Variance = $\frac{p\theta}{(1 - p)^2} = \mu + \frac{\mu^2}{\theta}$

How is θ related to overdispersion?

larger $\theta \Rightarrow$ mean & variance are more similar
smaller $\theta \Rightarrow$ greater variance relative to mean

Negative binomial regression

$$Y_i \sim NB(\theta, p_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 X_i$$

- + $\mu_i = \frac{p_i \theta}{1 - p_i}$
- + Note that θ is the same for all i
- + Note that just like in Poisson regression, we model the average count
 - + Interpretation of β s is the same as in Poisson regression

Comparing Poisson, quasi-Poisson, negative binomial

Poisson:

- + Mean = λ_i
- + Variance = λ_i

quasi-Poisson:

- + Mean = λ_i
- + Variance = $\phi\lambda_i$

negative binomial:

- + Mean = μ_i
- + Variance = $\mu_i + \frac{\mu_i^2}{\theta}$

In R

```
m3 <- glm.nb(art ~ ., data = articles)
```

```
...
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.256144  0.137348  1.865 0.062191 .
## femWomen    -0.216418  0.072636 -2.979 0.002887 **
## marMarried   0.150489  0.082097  1.833 0.066791 .
## kid5        -0.176415  0.052813 -3.340 0.000837 ***
## phd          0.015271  0.035873  0.426 0.670326
## ment         0.029082  0.003214  9.048 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for Negative Binomial(2.2644) family)
##
```

$$\hat{\theta} = 2.264$$

In R

```
...  
##                Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 0.256144  0.137348  1.865 0.062191 .  
## femWomen    -0.216418  0.072636 -2.979 0.002887 **  
## marMarried   0.150489  0.082097  1.833 0.066791 .  
## kid5        -0.176415  0.052813 -3.340 0.000837 ***  
## phd          0.015271  0.035873  0.426 0.670326  
## ment         0.029082  0.003214  9.048 < 2e-16 ***  
...
```

How do I interpret the estimated coefficient -0.176?

quasi-Poisson vs. negative binomial

quasi-Poisson:

- + linear relationship between mean and variance
- + easy to interpret $\hat{\phi}$
- + same as Poisson regression when $\phi = 1$
- + simple adjustment to estimated standard errors
- + estimated coefficients same as in Poisson regression

negative binomial:

- + quadratic relationship between mean and variance
- + we get to use a likelihood, rather than a quasi-likelihood
- + Same as Poisson regression when θ is very large and p is very small