(4) and (5) p-value

p-value : $P(|t_{stat}| > t^*_{(n-1,\alpha)} \mid H_0)$ → two-sided test

$= P(t_{stat} < t_{(n-1)} \mid H_0) + P(t_{stat} > t_{(n-1)} \mid H_0)$

$= pt(q = 42, df = 399, lower.tail = F) \times 2$

$= pt(q = -42, df = 399, lower.tail = T) +$
$pt(q = 42, 399, lower.tail = F)$

(6) Residual = observed − predicted

e_i, $i = 1, \ldots, 400$ $\quad e_i = y_i - \hat{y}_i$

= depths$alt − 3628.5

vector of length 400

residual standard error $= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} e_i^2}$

S = residual Std. Error

ONLY in the intercept only model

$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$

$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2}$

= Standard deviation (s)

= 1730 = sd(depths$alt)

③ $H_0: \mu_0 = \mu_1$ or $\mu_1 - \mu_0 = 0$

$H_a: \mu_0 \neq \mu_1$ or $\mu_1 - \mu_0 \neq 0$

$$t_{stat} = \frac{(\bar{y}_1 - \bar{y}_0) - (\mu_1 - \mu_0)}{SE_{\bar{y}_1 - \bar{y}_0}}$$

$$= \frac{2.11}{1.73} = 1.22$$

n_0: sample size in North

n_1: sample size in South

④ and ⑤

$$\text{p-value} = P(|t_{stat}| > t^*_{(n_0 - 1 + n_1 - 1)/H_0})$$

$df = n - p$

n: sample size

p: # of determinants + 1 Intercept

$pt(q = 1.22, df = 398, lower.tail = F) \times 2$

⑥ Residual std. error

$$= \sqrt{\frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = 1.730$$

② 95% CI for $\Delta\mu$

$(\bar{y}_0 - \bar{y}_1) \pm 1.96 \times SE_{\bar{y}_0 - \bar{y}_1}$

$\rightarrow 211 \pm 1.96 \times 173$

Example 3

$$\log(\mu) = \begin{cases} \log(\mu_0) & \text{if South} = 0 \\ \log(\mu_0) + \log(\theta) & \text{if South} = 1 \end{cases}$$

① (Intercept) Estimate $= \log(\bar{y}_0)$

→ an estimate of the parameter $\log(\mu_0)$

$\log(\bar{y}_0) = 8.167 \Rightarrow \bar{y}_0 = \exp(8.167)$

South estimate: $\widehat{\log(\theta)} = 0.0581$

$\hat{\theta} = \exp(0.0581)$

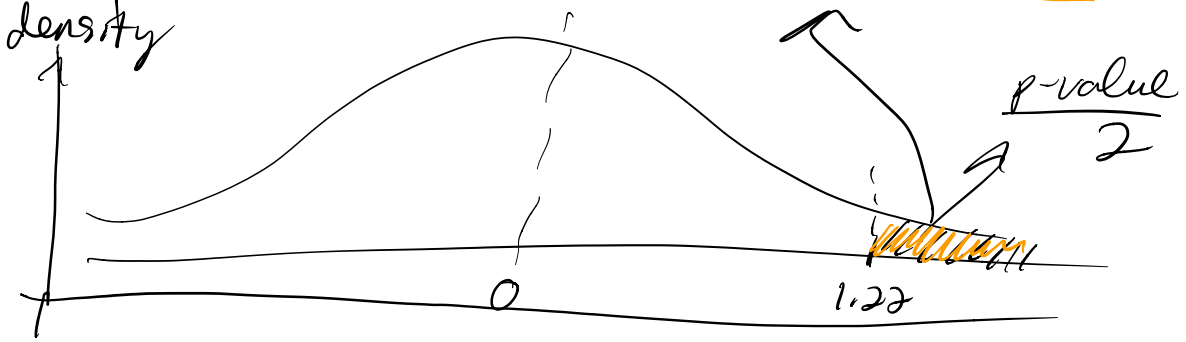
$= 1.06$

② $H_0: \log(\theta) = 0$ $\quad H_a: \log(\theta) \neq 0$

$$t_{stat} = \frac{\widehat{\log(\theta)} - \log(\theta)}{SE_{\widehat{\log(\theta)}}} = \frac{0.0581 - 0}{0.0477} = 1.22$$

③ and ④ $p\text{-value} = P(|t_{stat}| > t_{(398)} | H_0)$

$pt(q = 1.22, df = 398, lower.tail = F) \times 2$



⑤ 95% CI:

$0.0581 \pm qt(c(0.025, 0.975), df = 398) \times 0.0477$

$= [\text{Lower}, \text{upper}]$

95% CI for the $\log(\theta)$

a 95% CI for θ

$= [\exp(\text{lower}), \exp(\text{upper})]$

This is incorrect:

$\exp(\log(\theta)) \pm 1.96 \cdot \exp(SE_{\log(\theta)})$

This is correct:

$\exp[\log(\hat{\theta}) \pm 1.96 \cdot SE(\log\hat{\theta})]$

(7) pseudo R^2

$= 1 - \frac{\text{residual deviance}}{\text{null deviance}}$

residual deviance → for the model being fit

null deviance → measure of fit from an intercept only model

lm → is alway family=gaussian

glm → family = gaussian(link = identity)

confint("glm object")

↳ 95% CI for log(θ)

exp(confint(glm object))

Poisson Regression

Nov 6, 2020

Example 1

(1) Coefficient estimate for PT is $\hat{\lambda}_0 = 3.39$ cases per infant year.

(2) Estimte for $\Delta\lambda \rightarrow$ rate difference

$= \lambda_1 - \lambda_0$

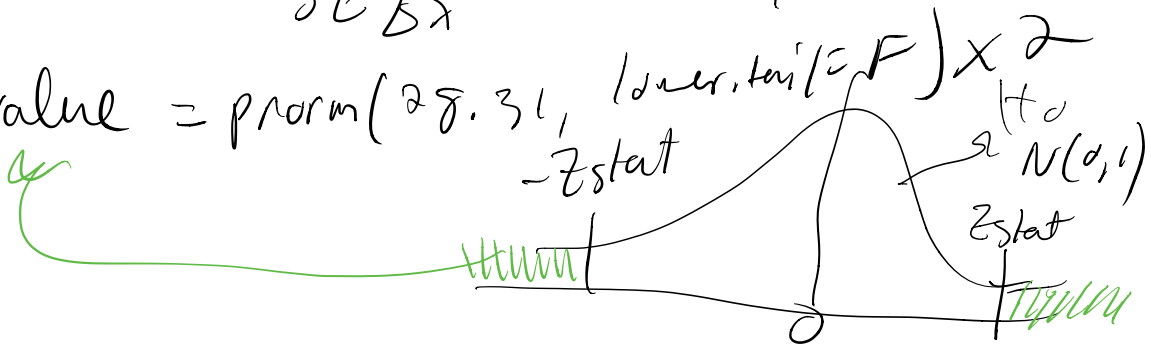
$= 0.31 \rightarrow \hat{\lambda}_1 = 3.7$

Cases per infant year.

(3) $H_0: \Delta\lambda = 0$ $\quad H_a: \Delta\lambda \neq 0$

$$Z\,stat = \frac{\hat{\Delta\lambda} - 0}{SE_{\hat{\Delta\lambda}}} = \frac{0.31 - 0}{0.010951} = 28.31$$

p-value = pnorm(28.31, lower.tail = F) x 2



95% CI for $\Delta\lambda$ based on Normal approximation

qnorm(p = c(0.025, 0.975),
mean = 0.31,
sd = 0.010951)

OR

$0.31 \pm 1.96 \cdot 0.010951$

$SE_{\Delta\lambda}$

confint(fit)

Nov 11, 2020

$$\log(A \cdot B) = \log(A) + \log(B)$$

$$\log(A^X) = X \log(A)$$

properties (rules)

$$\ln(\text{depth} \sim \text{South}) = \beta_1$$

ln(depth ~ 1 + South)

$$\mu = \mu_0 + \Delta\mu \cdot \text{South}$$

(Intercept) → $\hat{\mu_0}$

South → $\hat{\Delta\mu}$

Location	Depth	South
1	1000	0
2	200	1
⋮	⋮	0

$$\mu = \lambda_0 \cdot \theta^{NBF} \cdot PT \quad (1)$$

glm does not know how to fit this model.

Take "log" both sides of (1)

$$\log(\mu) = \log(\lambda_0 \cdot \theta^{NBF} \cdot PT)$$

$$\overset{\text{property of logs}}{=} \log(\lambda_0) + \log(\theta^{NBF}) + \log(PT)$$

$$= \log(\lambda_0) + NBF \cdot \log(\theta) + \log(PT)$$

$\log(\hat{\lambda}_2) = 1.220832 \Rightarrow \hat{\lambda}_2 = \exp(1.22)$

$= 3.39$

$\log(\hat{\theta}) = 0.087505 \Rightarrow \hat{\theta} = \exp(0.087)$

$= 1.09$

1.09 is the incidence rate ratio the rate of respiratory infections among not breastfed infants is 9% higher than those who were breastfed.

95% CI for $\log(\theta)$

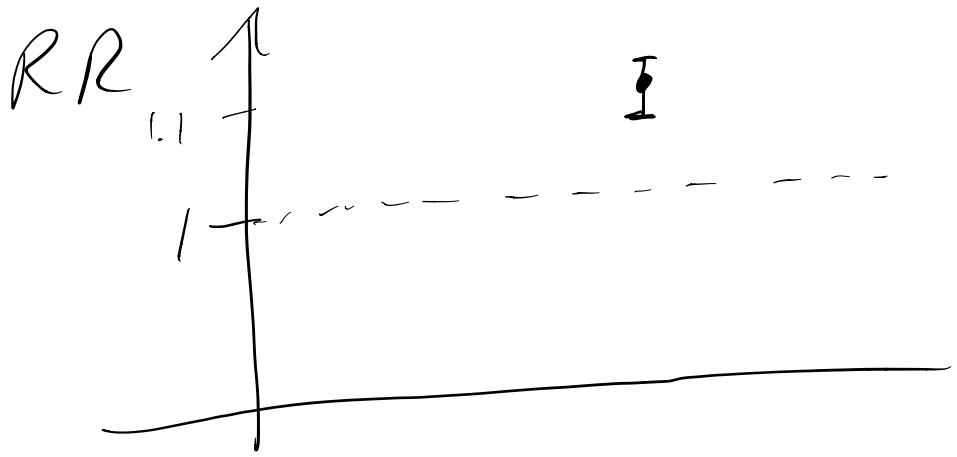
$0.08 \pm 1.96 \cdot 0.003012$

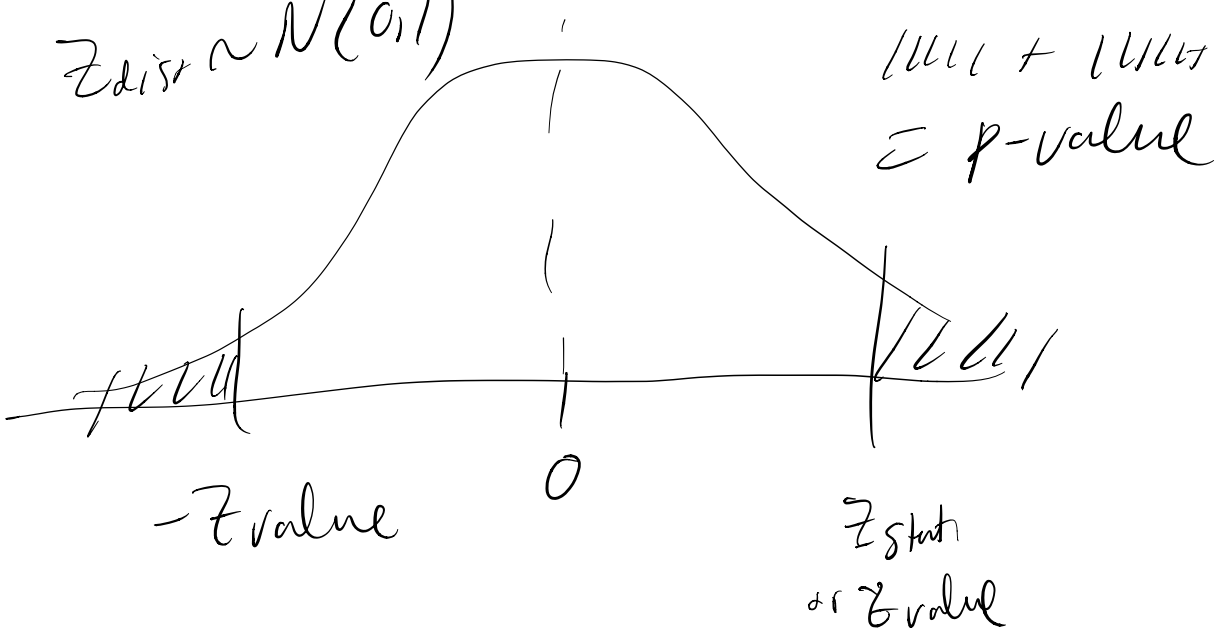
95% CI for θ

$\exp(0.08 \pm 1.96 \cdot 0.003012)$

exp(confint(fit))

= [1.085023, 1.0979]





$= pnorm(q = a, lower.tail = F) \times 2$

Bednets

expected # of cases of malaria $= \text{Rate} \times PT$

$$\mu = \boxed{\lambda} \times PT$$

λ → model for rate

$$\lambda = \lambda_0 \cdot \theta^{\text{exposure}}$$

$$\theta = \frac{\lambda_1}{\lambda_0} \rightarrow \text{rate ratio}$$

Regression Equation

$$\mu = \lambda_0 \cdot \theta^{\text{exposure}} \times PT$$

$$\log(\mu) = \log(\lambda_0) + \text{exposure} \cdot \log(\theta) + \log(PT)$$

Estimates

$\log(\hat{\lambda}_0) = 0.683 \Rightarrow \hat{\lambda}_0 = 1.98$ cases/child year

$\log(\hat{\theta}) = -0.266 \Rightarrow \hat{\theta} = 0.765$

Goodness of fit

we need to compare observed # of cases to expected # of cases.

$$\hat{\mu} = \hat{\lambda} \times PT = \hat{\lambda}_0 \cdot \hat{\theta}^{\text{exposure}} \times PT$$

$$\hat{\mu} = \begin{cases} 1.98 \times PT & \text{if exposed} = 0 \\ 1.98 \times 0.765 \times PT & \text{if exposed} = 1 \end{cases}$$

month	exposure	cases	years	expected
June 2014	0	33	79	1.98×79
July 2014	0	454	123	1.98×123
aug 2014	0	244	103	1.98×103
aug 2014	1	43	23	$1.98 \times 0.765 \times 23$
sep 2014	0	177	79	
Sep 2014	1	66	39	

H_0: no lack of fit H_a: lack of fit

observed $\approx$ expected

$$X^2_{stat} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim X^2_{(k-1)}$$

$$= \frac{(33 - 156.43)^2}{156.43} + \frac{(454 - 243)^2}{243} + \dots$$

compare to a chisq. distribution

# of elements in X^2_{stat}

p-value = pchisq(q = X^2_{stat}, df = k-1, lower.tail = F)

vectorized

cases is a vector

expect " "

$$\text{Chisq.stat} \leftarrow \text{Sum}\left(\frac{(\text{cases} - \text{expected})^2}{\text{expected}}\right)$$

Nov 13, 2020

Age multipliers (M_{75}, M_{80}, M_{85})

years	Age	Female	Male
80-84	70-74	1	1
	75-79	0.0468/0.027 = 1.69	0.0822/0.052 = 1.58
	80-84	0.0808/0.027 = 2.97	2.33
	85-89	0.137/0.027 = 5.07	3.49
2000-2004	70-74	1	1
	75-79	1.57	1.66
	80-84	2.58	2.66
	85-89	4.49	4.22

A best estimate for M_{75} might be

mean (1.69, 1.57, 1.58, 1.66) → equal weight to all cells.

or

median (1.69, 1.57, 1.58, 1.66)

$\mu = \lambda \times PT$

↳ we need a model for lambda

Y: count $\quad Y \sim Poisson(\mu = \lambda \times PT)$

$$\lambda = \lambda_0 \cdot M_{75}^{I_{75}} \times M_{80}^{I_{80}} \times M_{85}^{I_{85}} \times M_{20+}^{I_{20+}} \times M_M^{I_M}$$

I : binary indicator variable.

$$\log(\lambda) = \log(\lambda_0) + I_{75} \cdot \log(M_{75}) + \cdots \leftarrow$$

Agl

"[70, 74)"

"[75-79]"

age → factor variable

factor(age, levels = c("(70,74)", "(75,79)", ---)

labels =

Obs	I_{75}	I_{80}	I_{90}	I_M	I_{20+}
1	0	0	0	0	0
2	1	0	0	0	0
3	1	1	0		

```
glm(cases ~ I75+ I80 + I90 +
            Im + I20y +
            offset(log(PT)),
    family = poisson(link="log"))
glm(cases ~ age + gender + period
    + offset(log(PT)))
```

```
levels(NHANES$variable)
```