2.1 → Introduction

→ model fitting steps
- → 1) model specification
- 2) Estimation of parameters
- 3) checking adequacy of model
- 4) Inference

2.2 → Examples

→ Count data →

Location	Count
1	Y_{11}
⋮	
1	Y_{1K}
2	Y_{21}
⋮	⋮
2	Y_{2K}

→ $H_0: \theta_1 = \theta_2 = \theta$

$H_1: \theta_1 \neq \theta_2$

★→ model fitting approach to testing H_0 is to fit 2 models, one assuming H_0 is true & the other assuming H_1

$E(Y_{jk}) = \theta \iff Y_{jk} \sim Poisson(\theta)$ $\qquad$ $E(Y_{jk}) = \theta_j \iff Y_{jk} \sim Poisson(\theta_j)$

→ then we compare how well the data fit each model

→ If H_0 is true → log-likelihood is

$$l_0 = l(\theta|y) = \ln[L(\theta|y)] = \ln\left[\prod_{j=1}^{2}\prod_{k=1}^{K_j} \frac{e^{-\theta}\theta^{y_{jk}}}{y_{jk_j}!}\right]$$
$$= \sum_{j=1}^{2}\sum_{k=1}^{K_j}[-\theta + y_{jk}\ln(\theta) - \ln(y_{jk_j}!)]$$

⋮

→ MLE is $\hat\theta_{MLE} = \frac{\sum\sum y_{jk}}{N}$ → $N = \sum_{j=1}^{2} K_j$

⇒ for this data → $\hat\theta_{MLE} = 1.184$ & max $\hat l_0 = l_0(\hat\theta_{MLE}) = -68.3868$

→ If H_1 is true → log likelihood function is

$$l_1 = l(\theta_1,\theta_2|y) = \ln[L(\theta_1,\theta_2|y)] = \ln\left[\prod_{j=1}^{2}\prod_{k=1}^{K_j}\frac{e^{-\theta_j}\theta_j^{y_{jk}}}{y_{jk_j}!}\right]$$
$$= \sum_{k=1}^{K_1}(-\theta_1 + y_{1k}\ln(\theta_1) - \ln(y_{1k}!)) + \sum_{k=1}^{K_2}(-\theta_2 + y_{2k}\ln(\theta_2) - \ln(y_{2k}!))$$

→ MLEs are $\hat\theta_{j\,MLE} = \sum_k y_{jk}/K_j$

→ for this data → $\hat\theta_{1\,MLE} = 1.423$ $\hat\theta_{2\,MLE} = 0.913$ & max $\hat l_1 = -67.0230$

→ NOTE: max $l_1 \overset{always}{>}$ max l_0 b/c more parameters are fit

→ but looking to see if significantly greater $\overset{?}{>>}$

→ need to know sampling dist of log-likelihood (discussed in ch 4)

→ Instead → look at (standardized) residuals for Poisson dist

→ if $Y \sim Pois(\lambda) \Rightarrow E(Y) = V(Y) = \lambda$

$\Rightarrow r = \frac{Y - \hat Y}{\sigma_Y} = \frac{Y - \hat\theta}{\sqrt{\hat\theta}}$

for straight-up Poisson (no $\lambda(Y) = \beta_0 + \beta_1 Y$) → Every value is predicted to be the mean

→ residuals can be aggregated to produce summary statistics measuring overall adequacy of model

→ Provided expected values θ_i are not too small, the standardized residuals $\overset{approx}{\sim} N(0,1)$, although usually not ⊥

→ Thus if $r_i \sim N(0,1) \Rightarrow r_i^2 \sim \chi_1^2$ &

$\sum r_i^2 = \sum\left(\frac{Y_i - \hat\theta_i}{\sqrt{\hat\theta_i}}\right)^2 \sim \chi_m^2$ → $\neq$ # of r_i in $\sum$

This is actually the chi-squared goodness of fit statistic for count data

$X^2 = \sum \frac{(o_i - e_i)^2}{e_i} \sim \chi_m^2$

→ for large samples, this is a good approximation w/ m = # obs − # parameters

→ for this data & model H_0 → $\sum r_i^2 = 6\times(-1.113)^2 + \dots + 1\times 1.664^2 = 46.759$ $\overset{\approx}{\text{close}}$ Informally Evaluating $E(\chi_m^2) = 48$ (23+26−1 = 48)

→ — — — — H_1 → $\sum r_i^2 = 43.659$ again $H_1 < H_0$ ALWAYS $\overset{\approx}{\text{close}}$ $E(\chi_m^2) = 47$ (23+26−2)

H_0 model

	Y	freq	standardized residual
Location 1	0	6	−1.088
	1	10	−0.169
	2	4	0.750
	3	5	1.669
	4	1	2.589
2	0	9	−1.088
	1	8	−0.169
	2	5	0.750
	3	1	1.669

repeated b/c constant θ

H_1 model

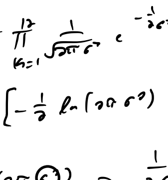
< different b/c $\hat\theta_1$ & $\hat\theta_2$ >

→ $\sum r_i^2$ statistics are each close to the expected values of their respective central chi-squared dist w/ appropriate df

& difference $\sum_{H_0} r_i^2 - \sum_{H_1} r_i^2 = 3.10$ is small ⇒ more complete model may not be better for this data

→ Continuous data →

group 1 o		group 2 *	
$x_{1,1}$	$Y_{1,1}$	$x_{2,1}$	$Y_{2,1}$
⋮		⋮	
$x_{1,12}$	$Y_{1,12}$	$x_{2,12}$	$Y_{2,12}$



Linear trend for both groups

$\bar x_1 \approx \bar x_2$, $\bar Y_1 \overset{?}{>>} \bar Y_2$

→ $H_0: \beta_1 = \beta_2 = \beta$ → $E(Y_{jk}) = \mu_{jk} = \alpha_j + \beta x_{jk}$, $Y_{jk} \sim N(\mu_{jk}, \sigma^2)$ — intercepts by group b/c $\bar Y_j$'s were fairly different

$H_1: \beta_1 \neq \beta_2$ → $E(Y_{jk}) = \mu_{jk} = \alpha_j + \beta_j x_{jk}$, $Y_{jk} \sim N(\mu_{jk}, \sigma^2)$

↳ Goal is to test rate (i.e. slope)

→ If H_1 is true → $l(\alpha_1,\alpha_2,\beta_1,\beta_2) = \ln[L(\alpha_1,\alpha_2,\beta_1,\beta_2)]$

$$= \ln\left[\prod_{j=1}^{2}\prod_{k=1}^{12}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(Y_{jk}-\mu_{jk})^2}\right] \quad \mu_{jk} = \alpha_j - \beta_j x_{jk}$$
$$= \sum_{j=1}^{J}\sum_{k=1}^{K}\left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_{jk} - \mu_{jk})^2\right]$$
$$= -\frac{1}{2}JK\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{jk} - \alpha_j - \beta_j x_{jk})^2$$

↳ nuissance parameter (not estimated)

⋮

→ MLEs are solutions of the simultaneous equations

$$\frac{\partial l_1}{\partial\alpha_j} = \frac{1}{\sigma^2}\sum_k(Y_{jk} - \alpha_j - \beta_j x_{jk}) = 0$$
$$\frac{\partial l_1}{\partial\beta_j} = \frac{1}{\sigma^2}\sum x_{jk}(Y_{jk} - \alpha_j - \beta_j x_{jk}) = 0$$

→ Alternative to MLE is Least Squares Estimation

→ minimize $S_1 = \sum_{j=1}^{J}\sum_{k=1}^{K}(y_{jk} - \mu_{jk})^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{jk} - \alpha_j - \beta_j x_{jk})^2$

Equations are the same ⇒ maximizing l_1 is equivalent minimizing S_1

→ LSEs are solutions of the equations

$$\frac{\partial S_1}{\partial\alpha_j} = -2\sum_{k=1}^{K}(Y_{jk} - \alpha_j - \beta_j x_{jk}) = 0$$
$$\frac{\partial S_1}{\partial\beta_j} = -2\sum_{k=1}^{K}x_{jk}(Y_{jk} - \alpha_j - \beta_j x_{jk}) = 0$$

simplify to

$$\Rightarrow \sum_{k=1}^{K}y_{jk} - K\alpha_j - \beta_j\sum_{k=1}^{K}x_{jk} = 0$$
$$\sum_{k=1}^{K}x_{jk}y_{jk} - K\alpha_j\sum_{k=1}^{K}x_{jk} - \beta_j\sum_{k=1}^{K}x_{jk}^2 = 0$$

normal equations

→ solutions are → $\hat\beta_j = \frac{K\sum_k x_{jk}y_{jk} - (\sum_k x_{jk})(\sum_k y_{jk})}{K\sum_k x_{jk}^2 - (\sum_k x_{jk})^2}$

$\hat\alpha_j = \bar Y_j - b_j\bar x_j$

→ numerical minimum for S_1 is found by substituting parameter estimates & data into S_1

→ If H_0 is true → same process, but use $E(Y_{jk}) = \mu_{jk} = \alpha_j + \beta x_{jk}$ instead

⋮

< parameter estimates >

→ Get residuals of fitted values to check model adequacy

→ for this example → very little difference between models

→ Formally → test H_0 vs H_1 → if H_0 is correct $\hat S_0 \approx \hat S_1$ ⇒ accept simpler model

if H_1 is correct $\hat S_0 > \hat S_1$ ⇒ accept complex model

again → to decide this, need sampling dist of $\hat S_1$ & $\hat S_0$

→ It can be shown $\hat S_1 = \sum_{j=1}^{J}\sum_{k=1}^{K}[Y_{jk} - \underbrace{(\alpha_j + \beta_j x_{jk})}_{\mu_{Y_{jk}}}]^2 - K\sum_{j=1}^{J}(\bar Y_j - \underbrace{\alpha_j - \beta_j\bar x_j}_{\mu_{\bar Y_j}})^2 - \sum_{j=1}^{J}(\hat\beta_j - \underbrace{\beta_j}_{\mu_{\hat\beta_j}})^2\left(\sum_{k=1}^{K}x_{jk}^2 - K\bar x_j^2\right)$

where $Y_{jk} \sim N(\alpha_j + \beta_j x_{jk}, \sigma^2)$

$\bar Y_j \sim N(\alpha_j + \beta_j x_{jk}, \frac{\sigma^2}{K})$

$\hat\beta_j \sim N(\beta_j, \frac{\sigma^2}{S_{x_jx_j}})$ → $S_{x_jx_j} = \sum_{k=1}^{K}(x_{jk}^2 - K\bar x_j^2)$

⇒ $\frac{\hat S_1}{\sigma^2}$ is a linear combo of sum of squares of normal random variables

↳ sum of square centers them & ÷ σ^2/(extra factors) scales them appropriately to be Z variables

↳ ⇒ * on top

→ in general → JK RVs $\frac{(Y_{jk} - \alpha_j - \beta_j x_{jk})^2}{\sigma^2}$

J RVs $\frac{(\bar Y_j - \alpha_j - \beta_j x_{jk})^2}{\sigma^2/K}$

J RVs $\frac{(\hat\beta_j - \beta_j)^2}{\sigma^2/S_{x_jx_j}}$

⇒ $\frac{\hat S_1}{\sigma^2} \sim \chi^2_{JK-2J}$ → N − n(estimated parameters)

→ If H_0 is true $\hat S_0 \sim \chi^2_{JK-(J+1)}$ → only 1 β instead of J β's

→ If $\beta_1 \neq \beta_2$ (i.e. H_1) → $\frac{\hat S_0}{\sigma^2} \sim \chi^2_{JK-(J+1),d}$ non central χ^2 b/c mean > that of corresponding central χ^2

& $\frac{\hat S_1}{\sigma^2} \sim \chi^2_{JK-2J}$ central χ^2

→ statistic if H_0 is correct

$\Rightarrow \frac{1}{\sigma^2}(\hat S_0 - \hat S_1) \sim \chi^2_{J-1}$ ↳ difference in # of parameters estimated

= improvement in fit

→ if H_0 is not correct

$\Rightarrow \frac{1}{\sigma^2}(\hat S_0 - \hat S_1) \sim \chi^2_{noncentral}$

BUT σ^2 is unknown ⇒ can't compare directly to χ^2_{J-1} dist

⇒ use ratio $F = \frac{\frac{\hat S_0 - \hat S_1}{\sigma^2}/(J-1)}{\frac{\hat S_1}{\sigma^2}/(JK-2J)} = \frac{(\hat S_0 - \hat S_1)/(J-1)}{\hat S_1/(JK-2J)}$

→ If H_0 is correct $F_0 \sim F(J-1, JK-2J)$ central

if not $F_1 \sim$ non central ⇒ $F_1 > F_0$

2.3 → Some Principles of statistical modelling

→ Model formulation

→ EDA helps pick model

→ model has two components

1) Probability distribution of Y (e.g. $Y \sim N(\mu, \sigma^2)$)

2) Equation linking $E(Y)$ w/ a linear combination of X

(e.g. $E(Y) = \alpha + \beta x$

$l(E(Y)) = \beta_0 + \beta_1 \sin(\alpha x)$)

→ for GLMs → all probability dists belong to exponential family

→ general model form $g(E(Y)) = \underbrace{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m}_{\text{linear component}} = X\beta$

→ Parameter estimation is done w/ MLE or LSE or Bayes

→ Residuals & model checking

↳ always $Y_i - \hat\mu_i$

⇒ Approximate standardized residuals have different definitions / transformations designed to improve normality or independence

→ for normal → $r_i = \frac{Y_i - \hat\mu_i}{\hat\sigma}$ → estimate of σ

⇒ these are slightly correlated (all depend on $\hat\mu_i$ & $\hat\sigma$)

& not exactly normally distributed b/c estimated σ

→ for Poisson → $r_i = \frac{Y_i - \hat\theta}{\sqrt{\hat\theta}}$

→ If model is good → should be little variation in the residuals

⇒ associations captured well by $\hat\mu_i$'s

→ Sum of squared residuals $\sum(y_i - \hat\mu_i)^2$ provides overall statistic for assessing adequacy of model

(it is what is optimized in MLE & LSE during estimation)

2.4 → Notation + coding of Explanatory variables

→ dummy variables strategy (usual way)

→ sum-to-zero constraint (soft max coding) → treats groups symmetrically

→ μ is overall average & α represents group differences

→ $[E(Y_{1k}) - \mu] + [E(Y_{2k}) - \mu] = \alpha + (-\alpha) = 0$

$\beta = \begin{bmatrix}\mu \\ \alpha\end{bmatrix}$ X: Group 1 $[1 \ \ 1]$, Group 2 $[1 \ \ {-1}]$ $\underbrace{}_{\Sigma = 0}$

→ ordinal regression

→ $E(Y_{1k}) = \mu$

$E(Y_{2k}) = \mu + \alpha_1$

$E(Y_{3k}) = \mu + \alpha_1 + \alpha_2$

$\beta = \begin{bmatrix}\beta_1 \\ \alpha_1 \\ \alpha_2\end{bmatrix}$ X: Group 1 $[1 \ 0 \ 0]$, Group 2 $[1 \ 1 \ 0]$, Group 3 $[1 \ 1 \ 1]$