Computational
Social Science

Modeling Count Data

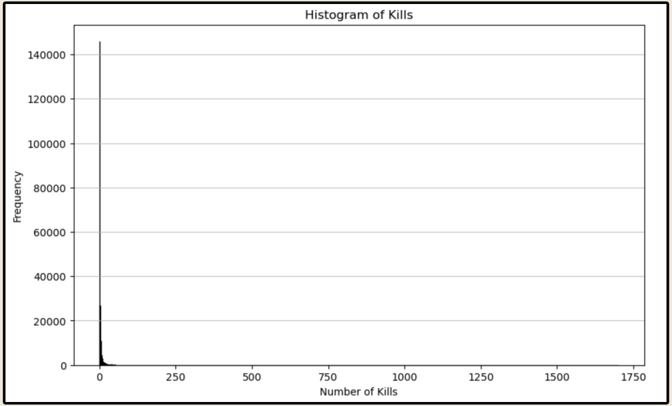
Roberto Cerina

05.03.2024

University of Amsterdam

Count Data

- $y_i \in \mathbb{N}_0$
- E.g $[...26, 1, 1, 0, 1, 0, 0, 2, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 3, ...]$



Modeling Option 1: Log-Linear Regression

- Consider modeling counts as normally distributed...
  - ✖ *a*. predictions can be negative (not an issue if y is very large);
  - ✖ *b*. shape of predictive distribution might be poor due to non-normality (non-symmetry, non-constant variance) of errors;
  - ✖ *c*. predictions will not be integers;
- We can fix *a*. and *b*. by transforming the dependent variable onto the log-scale:

- Empirical posterior distribution of count-scale fitted values:

❶ Learn posterior distribution of $\boldsymbol{\beta}$ from model fit to $log(y)$:
$\log(y_i) \sim N(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \sigma^2)$;

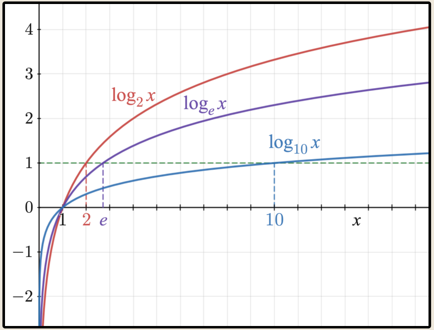
❷ Simulate S values of $\boldsymbol{\beta}$:
$\boldsymbol{\beta}^s \sim MN(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$;

❸ for each simulation, sample values of $\log(y_i)$:
$\tilde{y}_i^s \sim N(\beta_0^s + \beta_1^s x_{i1} + ... + \beta_p^s x_{ip}, \hat{\sigma}^2)$;

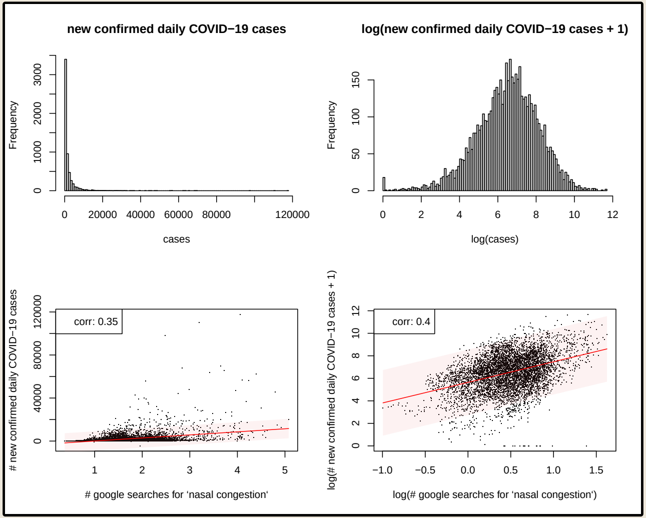
❹ Convert fitted values on the original scale:
$y_i^s = \exp(\tilde{y}^s)$

Modeling Option 1: Log-Linear Regression

- Taking the log a random variable encourages ‘normality’ and ‘constant variance’...
- *log$_e$ stretches* values from 0 to 1 into the negative , and *squishes* values above 1. The further away from 1 the larger the impact of the the log.



Modeling Option 1: Log-Linear Regression



Modeling Option 1: Log-Linear Regression

☛ Interpretation:

- **Linear model:**
  $y_i = \beta_0 + \beta_1 x_i + e_i$
  $\frac{\partial y}{\partial x} = \beta_1$
- A one-unit change in x is associated with a β_1 change in y.

Modeling Option 1: Log-Linear Regression

☛ Interpretation:

- **log(x) model:**
  $y_i = \beta_0 + \beta_1 \log(x_i) + e_i$
  $\frac{\partial y}{\partial x} = \frac{1}{x}\beta_1$

✏ $\log(x) + 1 = \log(x) + \log(e) = \log(ex) = \log(2.72x)$

- β is the expected change in y for a 172%[1] increase in x;
- A general $p\%$ increase in x can be expressed as: $x\left(1 + \frac{p}{100}\right)$;
  $$\begin{aligned}\Delta y &= \beta\log(x(1 + \tfrac{p}{100})) - \beta\log(x)\\ &= \beta(\log(x) + \log(1 + \tfrac{p}{100})) - \beta\log(x)\\ &= \beta\log(1 + \tfrac{p}{100})\end{aligned}$$
  $\therefore \beta \times \log\left(1 + \frac{p}{100}\right)$ is the expected change in y for a $p\%$ increase in x.
- For small p, due to Taylor expansion[2], $\log\left(1 + \frac{p}{100}\right) \approx \frac{p}{100}$
  $\therefore$ *if* $p = 1$, $\frac{\beta}{100}$ is approximately the change in y for a 1% change in x.

---

[1] For any factor c multiplied by x, the percentage change in x is $\Delta\% = 100 * (c - 1)$

[2] $\log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$

Modeling Option 1: Log-Linear Regression

☛ Interpretation:

- **log linear model:**
  $\log(y_i) = \beta_0 + \beta_1 x_i + e_i$
  $y_i = \exp(\beta_0 + \beta_1 x_i + e_i)$
  $\frac{\partial y}{\partial x} = \exp(\beta_0 + \beta_1 x_i + e_i)\beta_1 = y_i \beta_1$

✏ The rate of change in y_i for a one-unit change in x is proportional to the current value of y_i;

✑ A one-unit increase in x multiplies the outcome by $\exp(\beta)$:

$$\begin{aligned}
\frac{y(x_i+1)}{y(x_i)} &= \frac{\exp(\beta_0 + \beta_1(x_i+1) + e_i)}{\exp(\beta_0 + \beta_1 x_i + e_i)} \\
&= \frac{\exp(\beta_1(x_i+1))}{\exp(\beta_1(x_i))} = \\
&= \exp(\beta_1(x_i+1) - \beta_1 x_i) = \\
&= \exp(\beta_1) \\
y(x_i+1) &= y(x_i) \times \exp(\beta_1)
\end{aligned}$$

- for any other size of change in x: $y(x_i + c) = y(x_i) \times \exp(c\beta_1)$

Modeling Option 1: Log-Linear Regression

☛ Interpretation:

- **log linear model:**

$$
\begin{aligned}
y(x+1) &= \exp(\beta_0 + \beta_1(x_i + 1) + e_i) \\
&= \exp(\beta_0 + \beta_1 x_i + e_i)\exp(\beta_1) \\
&= y(x_i)\exp(\beta_1) \\
&\approx y(x_i)(\beta_1 + 1)
\end{aligned}
$$

- $\therefore$ a 1 unit increase in x is associated with a $\beta_1\%$ increase in y;
- The approximation above is especially true when β is small;
- Check $\beta^2/2$ – this is the first-order error of the Taylor expansion, and should be ≈ 0, for the approximation above to be reliable;
- As $\Delta x \to 0$, we recover the ‘instantaneous rate of change‘ – or the derivative from the previous slide.

Modeling Option 1: Log-Linear Regression

☛ Interpretation:

- **log-log model:**
  $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + e_i$
  $\frac{\partial y}{\partial x} = \exp(\beta_0 + \beta_1 \log(x_i) + e_i)\frac{1}{x_i}\beta_1 = \frac{y_i}{x_i}\beta_1$
- Looking at the instantaneous rate of change above, we can see the impact of a one-unit-change in x onto y can be interpreted as a mix of the partial models considered up to now:
- a $p\%$ increase in x is associated with an $\exp\left(\log\left(1 + \frac{p}{100}\right)\beta_1\right)$ percentage increase in y
- When β and p are small:
  $\approx \exp(\frac{p}{100}\beta_1) \approx \frac{p}{100}\beta_1 + 1$

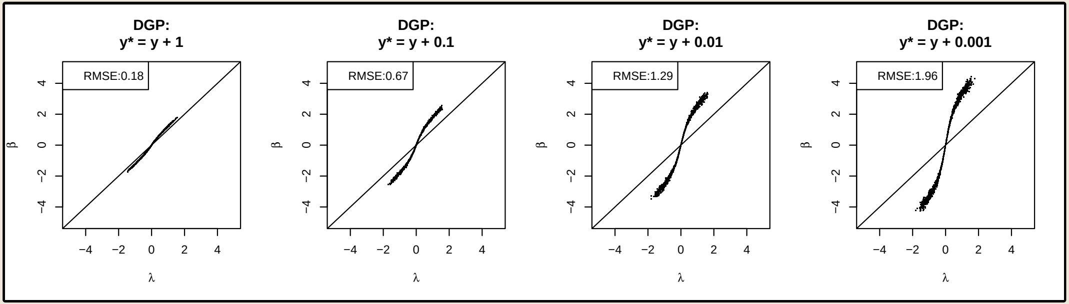
✔ You get approximately a $p\beta_1\%$ increase in y for a $p\%$ increase in x.

Problems with *log-plus-1* Regression

- ✏ To analyse count data witha linear regresison model with have to take $\log(y)$...
- ✖ ... $\log(0)$ is not defined ...
- ... need to add a constant to y to shift it away from 0 ...
- ... but this can substantial bias to regression coefficients:

$$\begin{aligned} \log(y_i) &= \beta_0 + \beta_1 x_i + e_i \\ \log(y_i + 1) &= \lambda_0 + \lambda_1 x_i + e_i \\ \hat{\lambda}_1 &\approx \frac{\hat{y}(x = \bar{x})}{1 + \hat{y}(x = \bar{x})} \hat{\beta}_1 \end{aligned}$$

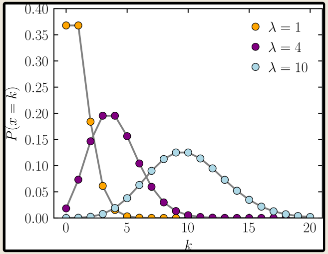
Problems with *log-plus-1* Regression



Modeling Option 2: Poisson Regression

- ✏ Log-linear regression uses a mathematical expedient to 'massage' the data into a convenient form, to analyse with methods we are familiar with...
- ✏ arguably the log-linear model does not represent the 'true' data generating process –
- ✖ counts from skewed samples are typically not 'really' normally distributed....

- ✏ The default distribution for count data is the Poisson distribution !

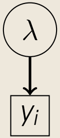


The ‘Homogeneous Rate’ Poisson Model

$$y_i \sim \text{Poisson}(\lambda) \qquad \forall\, y_i \in \mathbb{N}_0,\; i \in \{1, ..., n\}$$

- $\lambda \in (0, \infty)$ – *rate* parameter
  - expected number of events per ‘interval‘ (e.g. per time- or space-units)
  - the Poisson model assumes the variance of the data is equal to the expected value:
    $E[y] = Var[y] = \lambda$;
  - this is different from the normal (and log-normal) models, which allow for estimation of separate mean and variance...
    $E[y] = \mu$
    $Var[y] = \sigma^2$

The ‘Homogeneous Rate’ Poisson Model



Note:

- ‘square’ nodes indicate ‘observed / known values’;
- ‘circular’ nodes indicate ‘unknown parameters’;
- ‘solid’ arrows indicate ‘stochastic’ relationships – i.e. subject to random variability;
- ‘dotted’ arrows indicate ‘deterministic’ relationships – i.e. a given input will always provide the same output;

Estimation (Point Estimate)

Define a loss function, and minimise!

- Likelihood: $\mathcal{L}_i(\boldsymbol{y} \mid \lambda)$
- for an observation i: $\mathcal{L}_i = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$
- for the entire sample: $\boldsymbol{\mathcal{L}} = \Pi_i^n \mathcal{L}_i$
- log-likelihood: $\log(\boldsymbol{\mathcal{L}}) = \log(\Pi_i^n \mathcal{L}_i) = \sum_i^n \log(\mathcal{L}_i) \leftarrow$ this we want to maximise
- negative log-likelihood:
  $L = -\sum_i^n \log(\mathcal{L}_i) = -\sum_i^n (y_i \log(\lambda) - \lambda - \log(y_i!))$

✏ This has an analytical solution:
$\hat{\lambda}_{MLE} = \frac{1}{n}\sum_i^n y_i = \bar{y} \leftarrow$ the sample mean is the MLE of λ.

Uncertainty

- by asymptotic normality ($n \to \infty$) / Central Limit Theorem / Laplace approximation:

  $$\lambda \sim N\left(\hat{\lambda}_{MLE}, \frac{\hat{\lambda}_{MLE}}{n}\right)$$

- This is a good approximation of the true posterior distribution when:
  1. n is large enough;
  2. the true rate is substantially above 0.
- When the approximation fails, we risk producing samples estimates below 0...

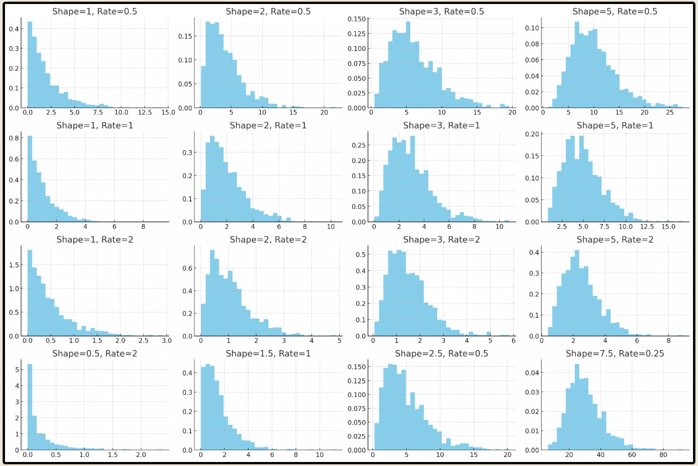
Uncertainty

- Alternative distribution by Bayesian empirical posterior distribution:

  $\lambda \sim \text{Gamma}(\alpha = \sum_i^n y_i, \beta = n)$;

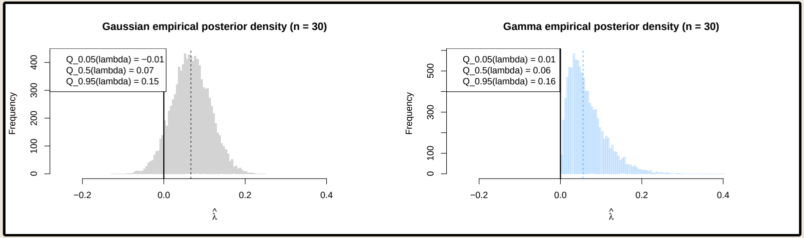
  - α: ‘shape’ – n. of observed events;
  - β: ‘rate’ – n. of intervals under consideration ;

Gamma Distribution



Simulation-Based (Monte-Carlo) Inference

- What statements can be made about λ ?
- draw S values from the empirical posterior:
  $\lambda_s \sim N\left(\hat{\lambda}_{MLE}, \frac{\hat{\lambda}_{MLE}}{n}\right)$;
  $\lambda_s \sim \text{Gamma}(\alpha = \sum_i^n y_i, \beta = n)$;
  $\lambda_{1:S} = \{0.24, 0.15, 0.12, 0.03, 0.04...\}$[3];
- Plot the distribution of λ and infer its properties:



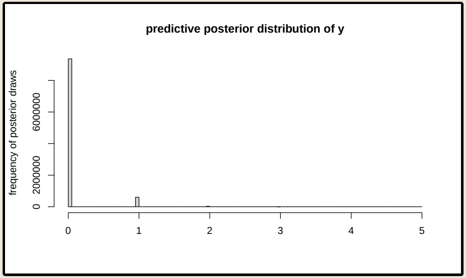
[3] These are all below 0 because for this example I chose a very rare event rate to display when the normal approximation can go wrong, but in principle they need not be !

Prediction

Posterior Predictive Distribution:

❶ simulate S 'new' values from λ according to its empirical posterior distribution:
$\tilde{\lambda}_s \sim \text{Gamma}(\alpha = \sum_i^n y_i, \beta = n)$

❷ simulate S new values $\tilde{y}$ according to its likelihood, conditional on the simulated values of λ:
$\tilde{y}_s \sim \text{Poisson}(\tilde{\lambda}_s)$



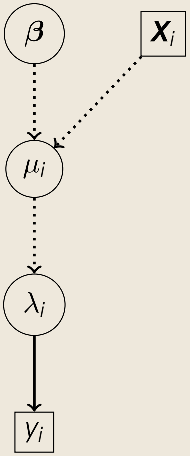
The 'Heterogeneous Rate' Poisson Model

$$
\begin{aligned}
y_i \sim & \text{Poisson}(\lambda_i) & \forall\, y_i \in \mathbb{N}_0,\ i \in \{1, \ldots, n\}, \\
\lambda_i = & e^{\mu_i} \\
\mu_i = & \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}
\end{aligned}
$$

$$\lambda_i = e^{\mu_i} \rightarrow \text{log link function}$$

- ✏ μ_i is transformed to a rate λ:
- ☎ Regression coefficients are interpretable in the same way as log-linear regression !
- ✏ λ is itself interpratable as the expected rate of at which an event occours;
- ✔ the Poisson distribution can natively handle $y = 0$, so no need to transform the data...

The ‘Heterogeneous Rate’ Poisson Model



Estimation (Point Estimate)

Define a loss function, and minimise !

- Likelihood for Poisson regression: $\mathcal{L}_i(y_i \mid \beta_0, \dots, \beta_p)$
- For an observation i: $\mathcal{L}_i = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})^{y_i} \exp\left(-\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})\right)}{y_i!}$
- Minimise:

$$L = -\sum_{i=1}^{n} \log(\mathcal{L}_i) = -\sum_{i=1}^{n} \left[y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) - \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) - \log(y_i!) \right]$$

✏ This **does not have an analytical solution !**

Uncertainty

- by asymptotic normality ($n \to \infty$) / Laplace approximation:

$$\boldsymbol{\beta} \sim MN(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$$

  - software will estimate $\hat{\boldsymbol{\beta}}$ via optimisation...
  - use estimate $\hat{\boldsymbol{\Sigma}}$ by plugging in $\hat{\boldsymbol{\beta}}$ into the *Hessian* matrix...

Posterior Predictive Distribution

❶ define a set of L ‘new subjects‘ characterised by design vector:
$\tilde{x}_l, \forall l \in \{1, ..., L\}$;

❷ simulate S ‘new’ values from $\boldsymbol{\beta}$ according to its empirical (joint) posterior distribution:
$\boldsymbol{\beta}^s \sim N(\hat{\boldsymbol{\beta}}_{MLE}, \hat{\boldsymbol{\Sigma}})$;

❸ Calculate μ_l (log-scale rate) for each simulation round:
$\tilde{\mu}_l^s = \beta_0^s + \beta_1^s \tilde{x}_{l1} + ... + \beta_p^s \tilde{x}_{lp}$

❹ Calculate λ_l (rate / expected count) for each simulation round:
$\tilde{\lambda}_l^s = \exp(\mu_l^s)$

❺ simulate S new values $\tilde{y}$ according to its likelihood, conditional on the simulated values of $\tilde{\lambda}$:
$\tilde{y}_l^s \sim \text{Poisson}(\tilde{\lambda}_l^s)$;

- You can then use Monte Carlo methods to make inference about predictions – the MC median will typically be ‘your best guess’ for the expected count.