Computational
Social Science

Logistic Regression Fundamentals .II

Roberto Cerina

22.02.2024

University of Amsterdam

Expressing & Comparing Probabilities

- $\Pr(y_i = 1) = \pi_i \in [0, 1] \leftarrow$ **probability / risk of** $y_i = 1$
- $\frac{\Pr(y_i=1)}{1-\Pr(y_i=1)} = \frac{\pi_i}{1-\pi_i} \in [0, \infty) \leftarrow$ **odds of** $y_i = 1$
- take π_0 to be the probability of the event happening for some 'reference' category, then:
- $\Pr(y_i = 1) - \Pr(y_0 = 1) = \pi_i - \pi_0 \leftarrow$ **risk difference**
- $\frac{\Pr(y_i=1)}{\Pr(y_0=1)} = \frac{\pi_i}{\pi_0} \in [0, \infty) \leftarrow$ **relative risk** (most useful with *rare* $\boldsymbol{\pi}$)
  - $\mathcal{RR} = 1 \leftarrow$ no difference in risk between subject *i* and reference...
- $\left(\frac{\Pr(y_i=1)}{1-\Pr(y_i=1)}\right) / \left(\frac{\Pr(y_0=1)}{1-\Pr(y_0=1)}\right) = \left(\frac{\pi_i}{1-\pi_i}\right) / \left(\frac{\pi_0}{1-\pi_0}\right) \in [0, \infty) \leftarrow$ **odds ratio**
  - $\mathcal{OR} = 1 \leftarrow$ no difference in odds of $y_i = 1$ between subject *i* and reference...

Interpreting Logistic Regression Coefficients

- **Homogeneous Probability - Intercept-only Model** [1]

$$
\begin{aligned}
y_i \sim & \mathrm{Bernoulli}(\pi_i) \\
\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = & \beta_0, \\
\frac{\pi_i}{1-\pi_i} = & \exp(\beta_0) \leftarrow \text{baseline odds;} \\
\pi_i = & \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \leftarrow \text{baseline probability;}
\end{aligned}
$$

[1] The notation here is slightly different from what you have seen, but its meaning is the same: $\mathrm{logit}(\pi_i) = \mu_i \rightarrow \pi_i = \mathrm{logit}^{-1}(\mu_i) = \frac{\exp(\mu_i)}{1+\exp(\mu_i)}$

Interpreting Logistic Regression Coefficients

- **Heterogeneous Probability – Binary Covariate** [2]

$$
\begin{aligned}
y_i \sim & \text{Bernoulli}(\pi_i) \\
\text{logit}(\pi_i) = & \beta_0 + \beta_1 x_{i1}, \qquad\qquad x_{i1} \in \{0, 1\} \\
\exp(\beta_0) = & \text{odds}(y_i = 1 \mid x_{i1} = 0) \\
\exp(\beta_1) = & \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \frac{\text{odds}(y_i = 1 \mid x_{i1} = 1)}{\text{odds}(y_i = 1 \mid x_{i1} = 0)}
\end{aligned}
$$

[2] Here we use the notaiopn '|', which means 'conditional' – so odds$(y_i = 1 \mid x_{i1} = 0)$ means 'the odds of the event happening if $x_{i1} = 0$...

Interpreting Logistic Regression Coefficients

- **Heterogeneous Probability – Continuous Covariate**

$$
\begin{aligned}
y_i \sim & \text{Bernoulli}(\pi_i) \\
\text{logit}(\pi_i) = & \beta_0 + \beta_1 x_{i1}, \qquad\qquad x_{i1} \in (-\infty, \infty) \\
\exp(\beta_0) = & \text{odds}(y_i = 1 \mid x_{i1} = 0) \\
\exp(\beta_1) = & \frac{\exp(\beta_0 + \beta_1(x_{i1} + 1))}{\exp(\beta_0 + \beta_1 x_{i1})} = \frac{\text{odds}(y_i = 1 \mid x_{i1} = z + 1)}{\text{odds}(y_i = 1 \mid x_{i1} = z)}
\end{aligned}
$$

$\exp(\beta_1) \leftarrow$ *the factor by which the odds of $y_i = 1$ are multiplied due to a 1-unit increase in x_{i1}.*

- For each additional unit increase in x_{i1} (e.g., each additional hour spent studying) the odds of the outcome (e.g., passing an exam) are multiplied by $\exp(\beta_1)$.
- $\exp(\beta_1) > 1 \rightarrow$ odds of success increase with each additional hour of study;
- $\exp(\beta_1) < 1 \rightarrow$ the odds of success decrease;
- $\exp(\beta_1) = 1 \rightarrow$ odds of success do not change with additional study time.

Interpreting Logistic Regression Coefficients

- **Heterogeneous Probability – Continuous Covariate**

$$
\begin{aligned}
\frac{\pi_i}{1-\pi_i} =& \exp(\beta_0 + \beta_1 x_{i1}) = \\
=& \exp(\beta_0)\exp(\beta_1 x_{i1}) = \\
=& \text{odds}(y_i = 1 \mid x_{i1} = 0) \times \exp(\beta_1 x_{i1})
\end{aligned}
$$

$\exp(\beta_1 x_{i1}) \leftarrow$ the factor by which the baseline odds multiply when the predictor variable increases by x_{i1} units (e.g. from 0 to 30 hours of study).

- by how much the odds of the outcome (e.g., passing an exam) are multiplied due to studying for x_{i1} hours, relative to not studying at all.

Interpreting Logistic Regression Coefficients

- **Heterogeneous Probability – Binary & Continuous Covariates** [3]

$$
\begin{aligned}
y_i \sim& \text{Bernoulli}(\pi_i) \\
\text{logit}(\pi_i) =& \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \qquad x_{i1} \in \{0,1\},\ x_{i2} \in (-\infty, \infty) \\
\exp(\beta_0) =& \text{odds}(y_i = 1 \mid x_{i1} = 0, x_{i2} = 0) \\
\exp(\beta_1) =& \frac{\text{odds}(y_i = 1 \mid x_{i1} = 1, x_{i2} = c)}{\text{odds}(y_i = 1 \mid x_{i1} = 0, x_{i2} = c)} \\
\exp(\beta_2) =& \frac{\text{odds}(y_i = 1 \mid x_{i1} = c, x_{i2} = z+1)}{\text{odds}(y_i = 1 \mid x_{i1} = c, x_{i2} = z)}
\end{aligned}
$$

[3] Here we use the notation $x = c$ to indicate the variable is being held at a constant value (i.e. *ceteris paribus*, we are not looking at the effects of variables as they 'change together' , but rather one at the time...)

Interpreting Logistic Regression Coefficients

- **Heterogeneous Probability – Interactions**

$$
\begin{aligned}
y_i \sim & \text{Bernoulli}(\pi_i) \\
\text{logit}(\pi_i) = & \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i1} x_{i2}), \qquad x_{i1} \in \{0, 1\},\ x_{i2} \in (-\infty, \infty) \\
\exp(\beta_3) = & \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i2} + 1) + \beta_3 x_{i1} (x_{i2} + 1))}{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2})} = \\
= & \frac{\exp(\beta_2 + \beta_3)}{\exp(\beta_2)} = \\
= & \frac{\text{odds}(y_i = 1 \mid x_{i1} = 1, x_{i2} = z + 1)}{\text{odds}(y_i = 1 \mid x_{i1} = 0, x_{i2} = z)}
\end{aligned}
$$

$\exp(\beta_3) \leftarrow$ *an additional factor by which the odds of y_i are multiplied due to a 1-unit increase in x_{i12}, when $x_{i1} = 1$.*

Interpreting Logistic Regression Coefficients

- **Heterogeneous Probability – Interactions**

$$
\begin{aligned}
\frac{\pi_i}{1-\pi_i} =& \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i1} x_{i2}) \\
=& \text{exp}(\beta_0)\text{exp}(\beta_1 x_{i1})\text{exp}(\beta_2 x_{i2})\text{exp}(\beta_3 x_{i1} x_{i2}) = \\
=& \text{odds}(y_i = 1 \mid x_{i1} = 0, x_{i2} = 0) \times \text{exp}(\beta_1 x_{i1})\text{exp}(\beta_2 x_{i2})\text{exp}(\beta_3 x_{i1} x_{i2})
\end{aligned}
$$

Problems with Logit Coefficients Interpretation

❶ odds and odds ratios are difficult to interpret – multiplicative quantities rather than additive;

❷ probabilities / risk easier to interpret:

- only when π is rare ($\pi \to 0$, or smaller), odds $\approx$ risk...

❸ logistic regression coefficients are *not collapsable*[4]:

- suppose you estimate a logit model with one regression coefficient: $\text{logit}(\hat{\pi}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$;
- I now add another covariate $\boldsymbol{x}_2$, which is fully *independent* (i.e. does not affect in any way, notation: $\boldsymbol{x}_1 \perp\!\!\!\perp \boldsymbol{x}_2$));
- because of independence, the introduction of a coefficient for $\boldsymbol{x}_2$ should not affect the value of $\hat{\beta}_1$.
- In linear regression, this holds, and $\hat{\beta}_1$ is unchanged...
- ... but in logistic regression, this does not hold and $\hat{\beta}_1$ will be different, even if in theory the effect $\boldsymbol{x}_1$ should not be disturbed by the introduction of $\boldsymbol{x}_2$!

[4] Norton, E. C., Dowd, B. E., Maciejewski, M. L. (2018). Odds ratios—current best practice and use. Jama, 320(1), 84–85.

Inference via Predicted Probabilities

- ✏ Logit coefficients are challenging to interpret and can be inconsistent – this is not the case with predicted probabilities...
- ✏ we can use predicted values to estimate relative risks of a given ‘profile’ v. a reference ‘profile’...
- ✏ e.g. how likely is a 57 year old divorced man, without a college degree, to not able to pay for her medication, relative to an average American ?

Inference via Predicted Probabilities

Step 1: Calculate 'risk' of not being able to pay for medication for the 'average American'

❶ Define the 'average' subject as an individual who has exactly the average value for every attribute:
$\bar{\boldsymbol{x}} = [x_1 = \bar{x_1}, x_2 = \bar{x_2}, ..., x_p = \bar{x_p}]$
- if you standardise your covariates ($\boldsymbol{x}^\star = \frac{x - \bar{x}}{sd(x)}$), then the average individual is simply $\bar{\boldsymbol{x}}^\star = [x_1^\star = 0, x_2^\star = 0, ..., x_p^\star = 0]$;

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 $\bar{\mu}$ for each simulation round:
$\bar{\mu}^s = \beta_0^s + \beta_1^s \bar{x}_1 + ... + \beta_p^s \bar{x}_p$

❹ Calculate $\bar{\pi}$ for each simulated $\bar{\mu}$:
$\bar{\pi}^s = \frac{\exp(\bar{\mu}^s)}{1+\exp(\bar{\mu}^s)}$

Inference via Predicted Probabilities

Step 2: Calculate 'risk' of not being able to pay for medication for a profile of interest...

- e.g. an American who is not college educated ($x_1 = 0$) but otherwise average...

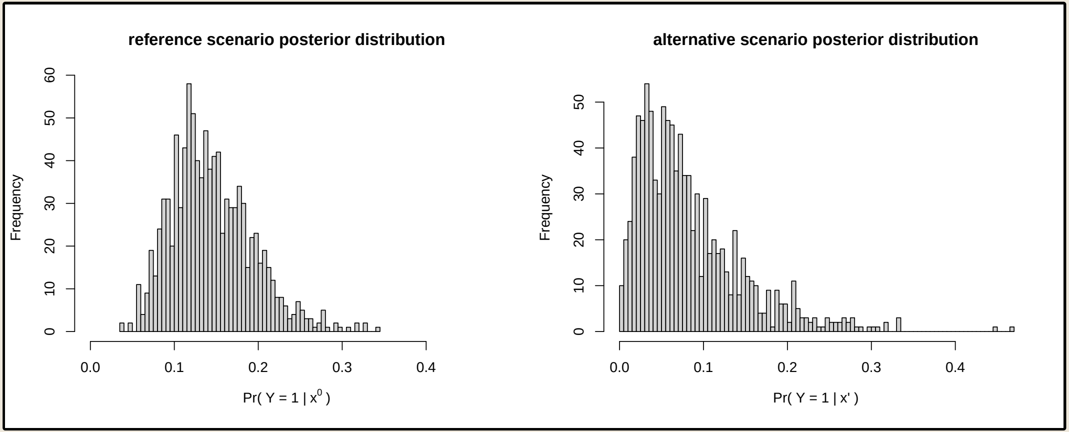
❶ Define the profile of interest:
$\tilde{\boldsymbol{x}} = [x_1 = 0, x_2 = \bar{x_2}, ..., x_p = \bar{x_p}]$

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 $\tilde{\mu}$ for each simulation round:
$\tilde{\mu}^s = \beta_0^s + \beta_1^s \times 0 + ... + \beta_p^s \bar{x}_p$

❹ Calculate $\tilde{\pi}$ for each simulated $\tilde{\mu}$:
$\tilde{\pi}^s = \frac{\exp(\tilde{\mu}^s)}{1+\exp(\tilde{\mu}^s)}$

Inference via Predicted Probabilities



Inference via Predicted Probabilities

Step 3: Calculate 'relative risk' of being Republican for a profile of interest, relavtive to the 'average' profile...

- For each simulated pair of values of $\tilde{\pi}$ and $\bar{\pi}$, calculate the relative risk $\mathcal{RR}^s = \frac{\tilde{\pi}^s}{\bar{\pi}^s}$
  - This results in the empirical distribution of your Relative-Risk, allowing us to quantify uncertainty around this measure.
  - You can use Monte Carlo methods to make inference about the Risks or the Relative risk...
  - This is a direct measure of the impact of changing covariates 'away from the average' - and it consistent across models / complexity.
- Sometimes, risk-differences (the difference between the risk of the profile of interest and the risk of the average profile) are of more interpretable / interesting (especially if π is not rare): $\mathcal{RD}^s = \tilde{\pi}^s - \bar{\pi}^s$

Inference via Predicted Probabilities