GLMs for Data Types

1. Continuous

- Range : $(-\infty, \infty)$
- Distribution : $y \sim Normal(\mu, \sigma^2)$
- Link: identity

$\mu_i = X_i^T \beta$

Coefficient Interpretation:

- Standard interpretation

2. Binary

- Range : $[0, 1]$
- Distribution : $y \sim Bernoulli(\pi)$

Logistic Link Coefficient Interpretation:

- Link: logistic / logit

$$\log(\frac{\pi_i}{1 - \pi_i}) = X_i^T \beta$$

- β_0 : log-odds of the event happening with all other predictors set to 0
- Continuous β_k : The change in the log-odds of response, comparing two populations whose value of X differs by 1 unit.
- Categorical β_k : Each β_k represents the change in the log-odds of the event happening when the predictor is in category k compared to the reference level.
- $\exp\{\beta_0\}$: odds of response with all other predictors set to 0
- Continuous $\exp\{\beta_k\}$:
  - ChatGPT: A one-unit increase in the continuous predictor gives a $\exp\{\beta_k\}$ increase in the odds of the event happening.
  - Slides: the ratio of the odds of response when X = 1 to that when X = 0

Note that an odds ratio > 1 indicates an increase in the odds of the event happening, while < 1 indicates a decrease in the odds.

Probit Link Coefficient Interpretation

$$\text{probit}(\mu_i) = \Phi^{-1}(\mu_i) = X_i^T \beta$$

- β_0 : the probit of probability of response when X = 0
- β_1 : the change in the probit of the probability of response, comparing two populations whose value of X differs by 1 unit.

3. Polytomous

- Range : $[0, ..., K]$
- Distribution : $y_i \sim Multinomial(1, \pi_i), \quad \pi_i = (\pi_{i0}, ..., \pi_{iK})$
- Notation: i refers to study unit i, and we have n study units, so 1 to n total. k is the "treatment" index variable, in class examples k was psycho, wtloss, etc.
- π_i: probability of success in unit i.

3.1 Nominal: no ordering of response

Three model types:

1. Collapsing: combine response levels so we have a new binary response.
2. Separate Regressions

- each with level '0' as baseline.
- have 5 logistic regression models: π_{ki} vs. π_{0i}. That is, take each pair from π_i and compare to baseline.
- $logit(\pi_{i1}) = \log(\frac{\pi_{i1}}{\pi_{i0}}) = X_i^T \beta$
- π_{ik}/π_{i0} is the relative risk of response level k to 0.
- β_{kj} is the difference in log-relative risks between two populations whose value of X_j differ by one unit. The intercept is compared to the reference level but with covariates zero'd out.
- $\exp(\beta_{kj})$ is the RRR for outcome level k vs. 0 comparing two pops whose values of X_j differ by one unit. RRR_{kj}. Also called the odds ratio.

3. Simultaneous Regression

- Same interpretations as before just running everything at once.

3.2 Ordinal: natural ordering of response

Two types of models:

1. cumulative logits

$$\log(\frac{P(Y \leq k|X=x)}{P(Y > k|X=x)})$$

- which are the log odds of being at or below response level k. These are the cumulative logits.
- β_{kj}: usual interpretation of a log-odds ratio, but is a log-cumulative odds ratio. If we exponentiate, get the cumulative odds ratio. **But**, is different for each level of k. Each k includes every level up to that one.
- Also have the proportional odds model that assumes all slopes are same. Have:

$$H_0 : \beta_{1j} = \beta_{2j} = \beta_{3j} = ...H_A : all\ else$$

2. adjacent categories model:

- model probabilities against each other. For example, Complete vs. Partial, which is $log(\pi_{i3}/\pi_{i2})$. In cumulative model, we compare everything to the baseline, $log(\pi_{i3}/\pi_{i0})$.

4. Count

All models seem to use log link.

4.1 $y \sim Binomial(n, \pi)$

- Range : $[0, 1, ..., n]$
- This is out of n trials, unlike a Poisson model.

4.2 $y \sim Poisson(\mu)$

- Range : $[0, 1, ...]$

4.3 $Y|X \sim NBIN(\mu_i, \alpha)$

- Same as Poisson but α controls dispersion. Always a better modeling choice than Poisson.

Log Link Coefficient Interpretation

$$\log(\mu_i) = X_i^T \beta$$

Recall that μ_i is just an expected count for study unit i.

- β_0 : The intercept term represents the expected count of the event (e.g., number of occurrences) when all predictor variables are set to zero.
- β_k : represent the relative change in the expected count of the event for a one-unit change in the predictor variable, holding all other variables constant.
- $\exp(\beta_k)$: This is the rate ratio. The rate ratio indicates how the expected count changes for a one-unit change in the predictor variable.

5. Miscellaneous Notes

odds: $\frac{p}{1-p}$

- $>$1, event is more likely to happen than not.

log-odds: $\log(\frac{p}{1-p})$

odds ratio: $\log(\frac{\pi_{i1}/(1-\pi_{i1}))}{\pi_{i0}/(1-\pi_{i0}))})$

- an OR > 1 indicates increased odds of the event in the denominator. OR < 1 indicates increased odds of event in the numerator.
- Same as RRR.