GLM Study

Matt McCarthy

2024-25 AY

Contents

Factor vs. Covariate

In **statistical modelling**, these two terms are used to distinguish between **categorical** vs. **numeric** predictors:

- A **factor** (in R terminology) is a **categorical** variable that takes on one of a finite set of "levels". The model will typically estimate one parameter for each level (minus one reference level).
- A **covariate** is typically a **numeric** (continuous or integer) predictor. The model treats these as quantitative variables for which we estimate a slope (change in log-odds, or outcome, per unit increase).

Concept of an Interaction

- An **interaction** occurs when two or more **explanatory** variables do **not act independently** on the outcome/response variable, i.e., the combined effect of the explanatory variables is not just the sum of their separate effects.

Formally Specifying a Logistic Regression Model

We need to give the Systematic Component, and the Random Component.
Then just define the dummy variables.

- Systematic Component:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Random Component:

$$y_i | x_1, x_2, \ldots, x_n \sim Bin(n_i, p_i)$$

where:

$$x_1 = \ldots$$
$$x_2 = \ldots$$
$$\ldots$$
$$x_n = \ldots$$

and $i = 1$ to $\ldots$

Notes re Logistic Regression Model

Say we have the following GLM in R:
y is a factor of successes and failures for each cross-classification.

Aggregated vs Un-Aggregated Data

Explain the advantages of working with aggregated binary data rather than un-aggregated binary data.

- Allows us to test the goodness of fit (deviance) of the model as the approximate distribution is known.
- The residuals are easier to interpret.

Overdispersion

What is Over-dispersion?

Over-dispersion refers to **greater variability in the response variable** than would be expected under a Binomial distribution. This often occurs due to:

1. Unmodeled heterogeneity,
2. Clustering of data, or
3. Omitted variables.

How to Check for Over-dispersion in Aggregated Binary Data?

1. **Calculate the Mean Deviance:**

$$\text{Mean Deviance} = \frac{\text{Residual Deviance}}{\text{Degrees of Freedom}}$$

2. **Alternative Check: Pearson Chi-Square Statistic:**

$$\text{Mean Pearson Chi-Square} = \frac{\text{Pearson Chi-Square}}{\text{Degrees of Freedom}}$$

3. **Interpretation:**
   - If Mean Deviance ≤ 1: No evidence of overdispersion.
   - If Mean Deviance > 1: Overdispersion is present, indicating that the observed variance exceeds the expected variance.

Why Does This Work?

For aggregated binary data, the residual deviance (or Pearson Chi-Square statistic) approximately follows a **Chi-Square distribution** under the null hypothesis of no overdispersion.

What to Do if Overdispersion is Detected?

If overdispersion is present, consider the following remedies:

- Use a **quasi-binomial model** to introduce a dispersion parameter ($\phi > 1$).
- Consider a **beta-binomial model** to account for extra variability in success probabilities.
- Use a **generalized linear mixed model (GLMM)** to account for random effects or clustering.

Backward Elimination and Issues to Consider

What is Backward Elimination?

Backward elimination is a stepwise regression method where:

- The process starts with a **complex model** that includes all candidate predictors.
- An **algorithm sequentially deletes predictors** based on their **p-value**, starting with the predictor that has the largest (least statistically significant) p-value.
- The process continues until all remaining predictors are below a specified significance threshold (e.g., $p < 0.05$).

Issues to Consider When Conducting Backward Elimination

When applying backward elimination, several issues should be considered to ensure the validity and interpretability of the final model:

1. **Interactions with Predictors:**
    - Predictors that are part of interaction terms should **not be deleted individually**, as doing so may render the interaction term invalid or meaningless.
2. **Dummy Variables:**
    - **Dummy variables** representing categorical variables should be **treated as a group** and not eliminated individually, as this may result in incomplete representation of the categorical variable.
3. **Statistical vs. Practical Significance:**
    - A predictor may be statistically significant but not **practically significant**, especially in large sample sizes where even small effects can result in low p-values.
4. **Meaningfulness of the Model:**
    - Relying solely on statistical significance may result in a **model that lacks meaningful or actionable insights**.

Conclusion

Backward elimination is a useful tool for simplifying models, but careful attention must be paid to interactions, grouping of dummy variables, and the distinction between statistical and practical significance to avoid invalid conclusions or misinterpretation of the final model.