Degrees of Freedom, Parameters and Model Choice

In many cases, data is best analysed using a general or generalized linear model, including any of

1. Multiple linear regression
2. Logistic regression
3. Poisson regression
4. Negative binomial regression
5. Multinomial regression
6. Ordinal regression
7. Exponential regression
8. Weibull regression
9. Cox proportional hazards models
10. Accelerated failure time models.

For each of these types of model, it is still necessary to consider exactly what terms will appear on the right hand side of the model equation. In particular, it is necessary to consider what predictor variables (main effects) to include, and how, and what interaction terms to include. A useful approach to making these choices is due to Frank Harrell, in his book Regression Modelling Strategies. We summarise one of the main ideas of the book here.

This main idea is that **the size of your dataset and the types of responses you observe determine how many parameters you can put into your model** and still obtain reasonably stable estimates of them through the modelling process. If you have a huge dataset, with >10000 cases, e.g., you might as well include any term in your model that you think might be interesting. However, if you have only a moderate sized dataset, you will need to make some decisions about how many terms you can include, and this should be done BEFORE the modelling is carried out.

<u>How many parameters can you afford?</u>

If you are doing any of the following regression types, you can afford n/15 parameters, where n is your sample size:

- Multiple linear regression
- Poisson regression
- Negative binomial regression

If you are doing logistic or multinomial regression, you can afford m/15 parameters where m is the number of responses in the smallest category.

If you are doing any time to event analysis method, including those listed below, you can afford m/15 degrees of freedom, where m is the number of observed events.

- Exponential regression models
- Weibull regression models
- Cox proportional hazards models
- Accelerated failure time models.

If you are doing ordinal logistic regression, you can afford m/15 parameters, where m is calculated by the following formula:

$$m = n - \frac{1}{n^2} \sum_{i=1}^{k} n_i^3,$$

where n is the sample size, k is the number of levels in the ordinal response variable, and n_i is the number of datapoints at the ith level of the ordinal response variable.

How many parameters do various terms in the model "cost"?

For any general or generalised linear model EXCEPT multinomial or ordinal regression, the numbers of parameters needed for various terms on the right hand side are given below:

Main effects:

Type of predictor	Number of parameters
Quantitative or ordinal treated as quantitative	1
Nominal treated as fixed effect or ordinal treated as fixed effect nominal with k levels	k - 1
Nominal treated as random effect or ordinal treated as random effect nominal	1

Interaction effects:

First predictor	Second predictor	Number of parameters in interaction term
Quantitative	Quantitative	1
Quantitative	Fixed effect nominal with k levels	k-1
Quantitative	Random effect nominal	1
Fixed effect nominal with k levels	Fixed effect nominal with j levels	(k-1)(j-1)
Fixed effect nominal with k levels	Random effect nominal	k-1
Random effect nominal	Random effect nominal	1

For **multinomial** or **ordinal regression** models, the number of necessary parameters must be multiplied by the number of levels of the response variable minus 1.

Consequences of this for modelling

We can see from the above chart that nominal variables, whether response variables or fixed effect nominal predictors, are very costly in terms of available parameters. This means that if you have some categories in your nominal variables with very few datapoints, you may want to consider if it would be better to combine some of these categories to use fewer parameters or to treat a nominal predictor variable as a random effect. Of course, if the distinctions between categories are critical to the research question, this should not be done. But this may mean that either it is necessary to collect more data, or that other variables will need to be excluded from the analysis.

Another consequence is that it not advisable to turn quantitative variables into categorical variables by grouping, as this greatly increases the necessary number of parameters in the model. For instance, it would be better to keep BMI as a quantitative

measurement than group measurements into “underweight” “healthy weight” and “overweight” categories.