1 General statistical approach

1.1 Linear regression

In this thesis, we use binomial mixed effects logistic regression models with crossed random effects (Baayen et al. 2008). These models are, simply speaking, extensions of logistic regression models. A logistic regression models a dependent variable (or an *outcome*, or a *response* variable) as a function of one or more independent predictor variables (or *factors*, or *explanatory* variables). That is, an outcome y is modeled as a function of explanatory variables $x_1, x_2, x_3..., x_n$, and an error term ε.

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n + \varepsilon \tag{1.1}$$

The intercept α, and the regression coefficients β_1, β_2, and β_3 for each explanatory variable are estimated to achieve the model that best fits the data. Analysis of Variance (ANOVA) is a special case of logistic regression (Chatterjee in Jaeger's thesis page 40; Shravan's book and blog) that is one of the most common statistical tools in psychology and psycholinguistics [. . . CITE. . .]. These linear regressions as shown above (Equation 1.1) and ANOVAS however, are not well suited for categorical data like response to multiple choice questions or yes/no questions, confidence ratings, etc. For example, in all the experiments in the current thesis, the

response variables are responce accuracy given correct/incorrect response. Output of linear regression model ranges from $+\infty$ to $-\infty$ while accuracy (or probability) ranges fro, 0 to 1. Additionally, simple regression models do not take into account the variability across individual participants and items. These problems in psychological sciences and psycholinguistics research has been long pointed out as early as 19XX [. . . CITE language as a fixed effects fallacy. . .], and later [. . .]. They are addressed to some extent by binomial logistic regression, and for our purpose by incorporating mixed effects model to binomial logistic regression.

Below we briefly introduce binomial logistic regression and mixed effects model. Then we show a simple example of how binomial logistic mixed effects model is used in the experiments in this thesis.

1.2 Binomial logistic regression

The response variable in this experiments in this thesis are binary. Participants' written response to what they hear are coded as either correct or incorrect. A binomial logistic regression model is best suited for such a categorical data (Jaeger 2008). We will use the term logistic regression model and binomial logistic regression model interchangeably henceforth.

As the name suggests, the output variable in a logistic regression model is in logit scale. The model therefore predicts logits of an outcome variable. Logits are log with base e, i.e. ln.

Probability ranges from 0 to 1 only while *odds* ranges from 0 to $+\infty$. Fitting a linear regression model with probability or odds would assume the range to be between 0 and 1, and between 0 and $+\infty$ respectively. This restricts the range, and is incorrect for a linear model. Therefore, in a binomial logistic regression model, log-odds are used which range from $-\infty$ to $+\infty$.

A simple binomial logistic regression model is shown in Equation 1.2:

$$\ln(\frac{p}{1-p}) = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n + \varepsilon \tag{1.2}$$

This is equivalent to,

$$p = \frac{exp(\alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n + \varepsilon)}{1 + exp(\alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n + \varepsilon)} \tag{1.3}$$

$$= \frac{exp(\ln(\frac{p}{1-p}))}{1 + exp(\ln(\frac{p}{1-p}))} \tag{1.4}$$

where,

$$\ln(\frac{p}{1-p}) = logit(p) \tag{1.5}$$

Log-odds of correct response obtained from Equation 1.2 can be transformed to probability of correct response. Equations 1.4, and 1.5 provide the relationship between probability, logit (or log-odds), and odds ($\frac{p}{1-p}$)).

Some of the assumptions made for binomial logistic regression models are violated in our data. One of them being non-independence of observations, i.e., all data points are independent from one another. This assumption is violated in unbalanced design, and at times even for balanced design. Same participant responds multiple trials of same experimental condition within an experiment. Although the design itself is balanced, after removal of outliers and/or trials which are not appropriate for comprhension measures (see section XXX for details), number of trials in analyses are unequal for each participant, item, and experimental condition. This introduces a bias in the model [Jaeger2008; other papers on GLMM].

Another intrinsic property or feature of logistic regression is that it assumes a common mean for each predictors. It has been shown that this is in fact not true: the effect of a predictor can vary depending on different random variables like participants, or items. To account for these variances, mixed effects models are used. In recent days, these models are frequently used and advocated for by psycholingustis and statisticians [... cite ...].

1.3 Mixed effects modeling

To overcome the limitations of logistic models, like violation of assumption of non-dependence of observations, and to account for the variability in the subject and/or item related parameter, mixed effects models are used. Mixed effects models contain 1) both linear and logistic regressions, and 2) *fixed effects* and *random effects*, hence the name *mixed effects*. Fixed effects term, e.g., levels of degradation assumes that all levels of degradation used in the experiment are independent from one another and they share a common residual variance. The random effects term with only varying intercept, e.g., subject as intercept, assumes that if there are 100 subjects then the mean accuracy of those 100 subjects are only a subset of possible global accuracies drawn from a set of population mean. When a slope, e.g., levels of predictability, is included to the random effects structure in addition to the varying intercept (e.g., subjects), then the model assumes that the effect of predictability on response accuracy varies across subjects.

1.4 Binomial logistic mixed effects modeling

A binomial logistic mixed effects model with varying intercepts and slopes for items and subjects is shown in Equation 1.6 below.

$$\begin{aligned} \ln(\frac{p}{1-p}) = \alpha + u_\alpha + w_\alpha + (\beta_1 + u_{\beta_1} + w_{\beta_1}) \cdot x_1 + \\ (\beta_2 + u_{\beta_2} + w_{\beta_2}) \cdot x_2 + ... + (\beta_n + u_{\beta_n} + w_{\beta_n}) \cdot x_n \end{aligned} \tag{1.6}$$

where,

- α is the Intercept.
- Fixed effects: $\beta_1, \beta_2, ..., \beta_n$ are the coefficients (or effects) of $x_1, x_2, ..., x_n$.
- $\boldsymbol{u} = \langle u_\alpha, u_{\beta_1}, u_{\beta_2}, ..., u_{\beta_n} \rangle$: Varying intercept and slopes for random effect term like, *subject*.
- $\boldsymbol{w} = \langle w_\alpha, w_{\beta_1}, w_{\beta_2}, ..., w_{\beta_n} \rangle$: Varying intercept and slopes for random effect term like, *item*.

In this thesis, statistically, we examine the effect of predictability, speech degradation and speech rate (see section X.X, X.X, X.X) on response accuracy. And hence we use these variables in the fixed effects term. Subjects and items are used as random intercepts with by-subject and by-item slopes. The details of the models fitted to data from each experiment are given in Chapter X, X, X and X.

We therefore use binomial logistic mixed effects model as our primary statistical analysis tool in all the experiments reported in this thesis. We primarily follow the recommendations of Baayen et al. (2008), Barr et al. (2013), and Bates et al. (2015).

Works Cited

Baayen, R. H., D. J. Davidson, and D. M. Bates (2008). "Mixed-effects modeling with crossed random effects for subjects and items". In: *Journal of Memory and Language* 59.4, pp. 390–412. URL: http://dx.doi.org/10.1016/j.jml.2007.12.005.

Barr, Dale J. et al. (Apr. 2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal". In: *Journal of Memory and Language* 68.3, pp. 255–278. URL: http://dx.doi.org/10.1016/j.jml.2012.11.001.

Bates, Douglas et al. (2015). "Parsimonious Mixed Models". In: *arXiv*. arXiv: 1506.04967. URL: http://arxiv.org/abs/1506.04967.

Jaeger, T. Florian (2008). "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models". In: *Journal of Memory and Language* 59.4, pp. 434–446. URL: http://dx.doi.org/10.1016/j.jml.2007.11.007.