Module 2A: Categorical Predictors

- So far we've discussed regressions that use continuous variables (e.g., frequency) to predict other continuous variables (e.g., reaction time).
  - What if predictors are categorical?
    - Voiced, voiceless
    - Subject, object
  - (Coming soon: categorical dependent measures)
- To incorporate categorical predictors, we must encode them as numbers
  - E.g., voiced = 0, voiceless = 1
- This can then be incorporated into our regression model
  - E.g., to get the prediction for voiced, enter '0' into the regression equation

Module 2A: Categorical Predictors

- Default coding in R: Treatment Coding.

- One level of the factor is the *baseline*.
  - Other levels are *treatment* levels.

- For a factor with *n* levels, this coding scheme will yield *n* – 1 coefficients in the model, one corresponding to each treatment level.

- Interpretation of resulting model
  - The intercept represents the mean for the baseline level (collapsing across other factors).
  - Each coefficient represents the difference of the corresponding treatment level from the baseline.

Module 2A: Categorical Predictors

- Example: read in votPOA.txt to the variable vot
  - Subset of VOT data from Goldrick & Blumstein (2006), with place of articulation (POA) specified.

- Text labels for levels of factor POA
  - bilabial = stop articulated at lips (p,b)
  - alveolar = closure formed at alveolar ridge (t,d)
  - velar = closure formed at velum (k,g)
- To verify, type

  levels (vot$POA)

Module 2A: Categorical Predictors

- Build a linear model of VOT for voiced stops. Output should be:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.019581	0.001277	15.333	< 2e-16	***
POAbilabial	-0.003802	0.001473	-2.581	0.0102	*
POAvelar	0.008446	0.001733	4.873	1.57e-06	***

- Intercept: VOT for voiced alveolars (baseline)
- Second coefficient: Change in VOT from voiced alveolar to voiced bilabial (difference of treatment 1 from baseline)
- Third coefficient: Change in VOT from voiced alveolar to voiced velar (difference of treatment 2 from baseline)
- **Problem 6: How do the model's predictions line up with the mean VOTs for each place of articulation?**

Module 2A: Categorical Predictors

- You can also specify the ordering of treatment levels
- To define a new factor where bilabial is the baseline: (code chunk 15 in module2.R)

```
vot$POA.Treatment <- factor(vot$POA,levels = c
("bilabial","alveolar","velar"))
```

- **Problem 7. Make a new treatment coding where velar is the baseline. Build a new linear regression model predicting VOT <u>for voiced stops</u> using place of articulation using this new treatment coding [just insert POA.Treatment as the predictor instead of POA]. What does this model predict is the VOT for voiced alveolars? For voiced bilabials? For voiced velars?**

Module 2A: Categorical Predictors

- Alternatively, you can code your factors so that they define *contrasts* between (groups of) levels.

- Example using VOT dataset
  - Contrast 1: Lingual stops—those formed with the tongue (alveolar and velar)—have different VOTs than those formed with lips (bilabial).
  - Contrast 2: The lingual stops have different VOTs (alveolar ≠ velar).

Module 2A: Categorical Predictors

- Contrast coding
  - Contrast 1: Lingual stops—those formed with the tongue (alveolar and velar)—have different VOTs than those formed with lips (bilabial).
  - Define a new factor
    - Assign bilabial value -1/2
    - Assign both alveolar and velar +1/4
  - If bilabials are no different from lingual stops (as a group), the coefficient on this factor should be 0.

Module 2A: Categorical Predictors

- Contrast coding
  - Contrast 2: The lingual stops have different VOTs (alveolar ≠ velar).
  - Define a new factor
    - Assign bilabial value 0 [not relevant]
    - Assign alveolar -1/2
    - Assign velar +1/2

- If alveolars are no different from velars, the coefficient on this factor should be 0.

- Code chunk 16 provides R code

Module 2A: Categorical Predictors

- These contrasts are providing *numeric codes* for each of the alphanumeric labels in your data.

- Think of the contrasts as specifying the "translation" of each of the labels into a number (which a regression model can understand)

Module 2A: Categorical Predictors

- What do the numbers in the output mean?
- Intercept = grand mean
- Coefficient of bilabial vs. lingual: Mean of lingual consonant VOTs – mean of bilabial consonant VOTs
- Coefficient of alveolar vs. velar: Mean of velar consonant VOTs – mean of alveolar consonant VOTs
- **Problem 8. What does the model above predict is the VOT for voiced alveolars? For voiced bilabials? For voiced velars? Make sure to look back at how the different factors are coded when using the coefficients to calculate these values.**

Module 2A: Categorical Predictors

- As seen above, contrasts compare two different sets of levels (set size can equal 1).

- One set of levels is assigned a positive number, the other negative. These sum to 0.
  - The positive/negative assignments just determine what the coefficient means.

- E.g., alveolar vs. velar: alveolar received -1/2, velar +1/2 (bilabial 0). Coefficient is positive when velar has longer VOTs.
  - If we reversed the signs of the contrast, the coefficient would be negative when velars have longer VOTs.

Module 2A: Categorical Predictors

- As Problem 6-8 showed, coding up contrasts as opposed to treatment coding doesn't actually change the predictions of the model.

- So why choose one vs. the other?

Module 2A: Categorical Predictors

- The contrasts above are *orthogonal* = not correlated. In Module 1B, we saw how correlations make regression analysis problematic.
- In order for contrasts to be orthogonal:
  - You can only have N-1 of them (where N = number of levels to the factor)
  - The pairwise products of the contrast terms must sum to 0.
    - Contrasts above: (1/4*-1/2) + (-1/2*0) + (1/4*1/2) = 0
- This differs from treatment coding, which necessarily has correlations
  - For each predictor in the regression, the baseline has value 0—so the predictors are partially correlated.

Module 2A: Categorical Predictors

- Additionally, if you want the contrast coefficient to correspond to the difference in group means, you have to make sure that the sum of absolute values of levels within the contrast = 1.
  - E.g., alveolar vs. velar: |-1/2| + |0| + |+1/2| = 1

Module 2B: Interactions

- In addition to incorporating the simple effect of a factor, a linear model can include terms that reflect the joint influence of two factors.

- For example, in the lexdec dataset, the effect of frequency appears to be different for natives vs. nonnatives.
  - **Question 9a [9b comes in a few slides]. To provide some initial quantitative support for this observation, build two linear models of the lexdec data, predicting reaction time using frequency. Build separate models for native speakers (NativeLanguage == English) vs. nonnative speakers (NativeLanguage == Other). Compare the slopes of the two regression lines.**

Module 2B: Interactions

- This suggests that the effect of frequency is not independent of your language background.
  - = interaction between two factors

- Our current regression models can't describe this.
  - They can include a single overall slope for frequency;
  - Plus another overall slope for language background;
    - But they cannot let language background alter the effect of frequency

Module 2B: Interactions

- To specify, incorporate an additional term in the model, defined as factor 1 * factor 2
  - Quite literally, represented as the product of the coding of the levels of each factor.
  - (in fact, if you just type factor 1 * factor 2, R will automatically incorporate the main effects of both factors in your model).
- Example with dataset lexdec
  - NativeLanguage == Other is treatment coded as 1:
  - The NativeLanguage coefficient is the effect of non-native speakers relative to English;
  - The coefficient of frequency is the effect of frequency for English speakers
  - The coefficient of the frequency * NativeLanguage interaction is the effect of frequency for Non-native speakers.

Module 2B: Interactions

- **Question 9b. Build a multiple regression model incorporating main effects of frequency and language background and an interaction of language background and frequency.**
  - Note: the lexdec dataset uses treatment coding, with NativeLanguage as the baseline and Other as the treatment; you don't need to specify anything else!

- **Question 9c. Compare this model to the simple linear models in Question 9a. Does the model's estimate of the frequency effects for English vs. NonNative speakers fit the estimates from the simple models?**

Module 2C: Categorical Dependent Measures

- Categorical dependent measures: Correct/incorrect; yes/no; grammatical/ungrammatical…

- Most commonly modeled using the binomial distribution
  - Assumption: Two possible outcomes (categorical measure)
  - Independent, identically distributed trials—with a constant probability *p*

Module 2C: Categorical Dependent Measures

- Proportions are bounded (0,1)

- The binomial distribution has two other critical properties:
  - As the mean proportion gets closer to 0,1, the variance goes down; variance is dependent on the mean.
  - As the number of observations goes up, the variance goes down; variance is dependent on number of observations

Module 2C: Categorical Dependent Measures

- Our regression models assume errors are normally distributed.

- Normal distributions are bad at modeling proportions:
  - Normal distributions are unbounded
  - Normal distribution variance is independent of the mean
  - Representing proportions as numbers from 0,1 doesn't provide information about the number of observations—doesn't allow model to understand sample size/error link

Module 2C: Categorical Dependent Measures

- Partial solution: Transform proportions
  - Rather than a number bounded between 0 and 1, represent proportions as a number ranging from negative to positive infinity
  - Fits assumption of linear models
    - Really general property of linear models (not just those that assume the error is normally distributed): They are functions that relate real numbers to real numbers.

Module 2C: Categorical Dependent Measures

- Step 1: Odds of one outcome occurring vs. another
  - Odds(outcome 1) = probability outcome 1/ probability outcome 2
  - Relative probability of outcome 1
  - Range from 0 (outcome 1 never happens) to positive infinity (strictly undefined when outcome 1 always happens; then you're dividing by 0)
- Step 2: Log odds
  - Take natural log of the odds; resulting number is called a *logit*
  - Ranges from negative to positive infinity, centered at 0
    - When pr(outcome 1) = pr (outcome 2), odds are 1
    - log(1) = 0
  - Note: undefined when odds are 0

Module 2C: Categorical Dependent Measures

- Logistic regression refers to a type of regression model that uses logits as the dependent measure.
  - Note: rather than assuming residual error is normally distributed, model assumes it's distributed following a binomial.

- What does a logistic regression mean? How should we interpret the output?

Module 2C: Categorical Dependent Measures

- Coefficient on each predictor *x* = *x* change in log-odds
  - Transforming back, multiplying odds by e^x
- Suppose you're regressing odds of making an error and have treatment-coded native language background (1= non-native).
  - A positive coefficient means the odds of making an error increase for non-natives relative to natives.
    - Example: coefficient 2.3; $e^{2.3}$ = multiple odds by 9.97 = nearly ten-fold increase in odds of an error
  - A negative coefficient means the odds of making an error go down for non-natives vs. natives.
    - Ex: coefficient –2.3; $e^{-2.3}$ = multiple odds by 0.1 = *decrease* odds of making an error by 10

Module 2C: Categorical Dependent Measures

- Output of linear model = sum of logits
  - Sum of logit coefficients = log of *product* of coefficients
    - Model therefore assumes that odds combine *multiplicatively*

- Suppose we have a logistic regression on odds of making an error with two factors, native language status and word length. Both factors have positive coefficients.
  - Model claim: the odds of making an error is a product of
    - the change in odds for non-natives
    - the change in odds for longer words

Module 2C: Categorical Dependent Measures

- Last issue: variance in proportions is dependent on number of trials
  - R functions for logistic regression use variety of methods to correct for this
  - glm function in R uses by default a type of *weighted least-squares* method.
    - During the process of fitting the regression model, instead of simply minimizing least squared error, weight the contribution of each observation to your error estimate by how certain you are about the observations.
    - Underweight proportions that come from fewer numbers of trials

Module 2C: Categorical Dependent Measures

- For the next few questions, let's examine some data on the dative alternation in English.
  - In English, the dative construction can be realized in at least two ways:
  - Recipient is a noun phrase: Mary gave John the book.
  - Recipient is a prepositional phrase: Mary gave the book to John.
- The dataset *verbs* in languageR has some data from Bresnan et al. (2007), reporting frequencies of dative uses (from recorded telephone conversations and digitized versions of the Wall Street Journal).

Module 2C: Categorical Dependent Measures

- Dataset *verbs* from languageR consists of one line for each occurrence of a dative construction. Columns are:
  - Realization of recipient: NP,PP
  - Identity of verb
  - Animacy of recipient (animate, inanimate)
  - Animacy of theme (animate, inanimate)
  - Length of theme (a transformed coding of length in words of the theme)

Module 2C: Categorical Dependent Measures

- Code chunk 17 focus on cases where the the theme (e.g., the book) is inanimate, and then calculates proportions of PP vs. NP use depending on the animacy of the recipient

- **Question 10. Calculate the odds of using a prepositional phrase in each condition.**

Module 2C: Categorical Dependent Measures

- To build a logistic regression, we use the function glm [generalized linear model] specifying family = “binomial”; this means that the glm function should assume the data are binomially distributed

- Code chunk 18 builds a logistic regression examining how the odds of using a prepositional phrase is influenced by recipient animacy.
- For the dependent measure: glm will use the contrast coding specified by your categorical variable. Make sure to use treatment coding (already done for you here, PP = 1, NP = 0)
- Note: AnimacyOfRec: inanimate = 1, animate = 0
  - Can check this using contrasts() command.

Module 2C: Categorical Dependent Measures

- **Question 11. Examine the output of the model using summary(verb.glm). What does the model predict is the odds of using a PP for inanimate recipients? Animate recipients? Remember: the model is specifying logits, not odds. Compare these to the empirical odds calculated in question 10.**

Module 2C: Categorical Dependent Measures

- For continuous measures, multiple regressions provide t-statistics for coefficients.
- Logistic regression provides the Wald z statistic.
  - It's a type of standardized score; like the t used in linear regression, it's computed by dividing the coefficient estimate by the standard error of the estimate.
  - If the sample size is large enough, then the square of this standardized score is distributed as chi square (χ^2) with 1 degree of freedom.

Module 2C: Categorical Dependent Measures

- However, these have been argued to be biased statistics
  - In particular, Wald z is argued to be very sensitive to small sample sizes

- We'll discuss an alternative later today.

Module 2C: Categorical Dependent Measures

- Note: logistic regression can run into trouble when you have lots of cases where probabilities are 0% and 100%.
  - Logits are undefined in these cases, approaching $\pm\infty$
  - Essentially--can act like outliers, inflating parameter estimates
- Is this a problem?
  - In many cases, algorithms for fitting logistic regressions will be able to cope with parameter estimates.
  - Likelihood ratio tests are fairly robust to these cases.
- Solutions
  - Only analyze participants that are not at ceiling/floor (Florian Jaeger has adopted this in some papers).
  - "Empirical logit" analyses (see work by Dale Barr); these introduce various corrections of the logit to avoid infinities.

INTERIM SUMMARY

- Categorical predictors must be “translated” into numerical codes in order to become part of a regression equation.

- Interactions can be easily encoded in regressions.

- Categorical dependent measures must be “translated” into logits for regression analysis.