DON’T
PANIC

Structure of Workshop

- Work through slides, with a partner, at your own pace.
  - Each module has accompanying R code (e.g., module1.R) and datasets (these will be referred to in the code).

- You will encounter problems to work on, to check your understanding of the material.
  - After you finish each problem, review answer, in the R file—labeled by module and problem number.
  - E.g., module1-Answer1.R is the answer to problem 1 in module 1.

- **Let me know if you have questions!** Things will probably be confusing at some points; this is a first draft of these materials.

Structure of Workshop

- Module 1: Basics of multiple regression
- Module 2: Categorical predictors+dependent measures in regression
- Module 3: Mixed effects regressions
- Module 4: Significance testing maximal random effects

Structure of Workshop

- Tentative Plan:
  - Module 1: 9- 10:30
  - Module 2: 10:30-12
  - Module 3: 2-3:30
  - Module 4: 3:30-5
- 15 minutes discussion before end of each module's time
- Warning: this might be optimistic, but that's ok; we can skip stuff if need be.
  - Some modules might take longer than others.
  - Module 4 is essentially "application of methods"—this is something you can follow up with on your own. Let's try to get through Modules 1-3!

Structure of Workshop

- When you're ready, get started…
  - Ask me questions!
  - Ask each other questions!!

Module 1A: Simple Linear Regression

- A linear regression characterizes the relationship that holds between continuous variables.

- More precisely, it characterizes a line relating two variables

Module 1A: Simple Linear Models

- Execute Chunk 1 in the code file module1.R
  - Note throughout the code there is text following the # symbol. These are comments to help you understand the code. Make sure to read them!
- This loads in the library languageR and the dataset lexdec.
- This will create a dataframe called lexdec. These are data from 21 subjects from a lexical decision task. The data set contains information on 79 English concrete nouns.
- Focus on the columns specifying the log frequency of the English noun [lexdec$Frequency] and the log reaction time [lexdec$RT].

Module 1A: Simple Linear Models

- Execute code chunk 2.
- This will build a linear model relating frequency to reaction time.
  - Inspect the summary.
  - Note the intercept (6.58). This is the log reaction time when frequency = 0.
  - Note the slope (–0.04). This is how log reaction time changes for 1 unit increase in log frequency.

- Code chunk 2 will also plot the data and add the regression line to the figure.
  - Compare this line to the intercept/slope in the model output.

Module 1A: Simple Linear Models

- Execute code chunk 3. This will *simulate* reaction time data, fitting the assumptions of the linear model.
  - Note: You and your partner will have slightly different data; it's randomly generated!
- It then builds a linear model relating frequency to reaction time.
  - Just like lexdec, we are simulating *log* reaction time. That's why the numbers are between 6 and 8 (400-3000 msec)
  - Inspect the summary; compare it to the code that simulated the reaction times. Does the linear model successfully recover the properties of the model that generated the data?
- Code chunk 3 will also plot the simulated data and add the regression line to the figure.

Module 1A: Simple Linear Models

- What else is in the model summary?

- Information about the distribution of the residuals
  - Residual: Difference between model prediction and actual observation
  - Return to code chunk 3: Note that for each simulated observation, we add a bit of normally distributed random noise—that's what gives rise to the residuals.

Module 1A: Simple Linear Models

- What else is in the model summary?

- For each part of the linear model
  - Coefficient: lm's estimate of the effect
  - Standard error: The error in this estimate
  - A *t* statistic: Calculated for the null hypothesis that the intercept or coefficient is actually equal to 0.
    - the coefficient estimate / standard error
    - N-2 degrees of freedom

- Overall model: F-test
  - Basically: is the amount of variance attributed to model is greater than that attributed to error?

Module 1A: Simple Linear Models

- **Problem 1. The lexdec dataset has a column *Length* that gives the length of each word. Modify Code Chunk 2 to build a simple linear model that predicts reaction times from length. What's the overall relationship between word length and reaction time? What does the model predict should be the reaction time for a word of length 5?**

Module 1B: Multiple Regression Basics

- A multiple regression extends the linear model to include multiple factors.
  - Linear model equation: y = Intercept +Coefficient* x+ε
    - Note ε: Normally distributed random error
  - Multiple regression: allows for multiple predictors (and associated coefficients)
    - y = Intercept + Coefficient$_1$* x_1+Coefficient$_2$* x_2+ε

- These coefficients are estimated by taking into account not only the correlation between each predictor x_i and the dependent measure y (r_{yi}), but the correlations among the multiple predictors (r_{ij}).

Module 1B: Multiple Regression Basics

- Just like the linear model, we can use the *t*-statistic to test, for each individual coefficient, the null hypothesis that it is equal to 0.
  - Degrees of freedom are *N*–number of model parameters (intercept + number of coefficients).

Module 1B: Multiple Regression Basics

- **Problem 2. Build a multiple regression on reaction time using the lexdec dataset. Incorporate two predictors: length and frequency. Compare the coefficients of the predictors in this regression to those of the simple regressions using just frequency or just length. Why do these differences occur?**
- **What does the model predict should be the reaction time for words of length 5, frequency 4.5?**

Module 1B: Multiple Regression Basics

- Execute code chunk 4
  - This extends code chunk 3 (and your answer to problem 2), simulating data combining both length and frequency values.
  - It's a fully balanced design—all levels of frequency appear at all lengths.

- Compare the output of the linear model based on the data you simulated. These should be close to the intercept and coefficients of the process that generated the data.
  - Hint: Look for these in the code: 6.6, –0.04, 0.02

- Re-run the code a few times. Does the regression model tend to recover the right intercepts and coefficients?

Module 1B: Multiple Regression Basics

- Our simulated data nicely obeys the assumptions of the linear model—in particular, that residual error is normally distributed.

- How can we verify this?
- Execute code chunk 5
  - This plots the distribution of residuals (deviation of observations from model predictions).
  - Looks pretty normal!

Module 1B: Multiple Regression Basics

- Additional, more precise visualization
- Execute code chunk 6
  - A normal distribution quantile-quantile plot, with a superimposed line passing through what should be the first and third quartiles.
- Quartile: Quarter of the data
- Quantile: Any regular division of a probability distribution.
  - Here: Quantiles are based on the normal distribution
  - (see next slide for explanation)

Module 1B: Multiple Regression Basics

- Baayen (2008: 77)
- [this plot] "…graphs the quantiles of the standard normal distribution (displayed on the horizontal axis) against the quantiles of the empirical distribution (displayed on the vertical axis). If the empirical distribution is normal (irrespective of mean or variance), its quantiles should be identical to those of the standard normal, and the quantile-quantile plot should produce a straight line."

- This is what we see here.

Module 1B: Multiple Regression Basics

- Johnson's chapter explained how multiple regression deals with inter-correlations among predictors. As it turns out, this solution is imperfect.
  - As the correlations between predictors get larger, the linear model does a much worse job at recovering the properties of the process that generated the data.

- To illustrate this, code chunk 7 simulates an unbalanced dataset. In this dataset, someone has foolishly designed an experiment where all of the low frequency words are long, and high frequency words are short.
  - (A correlation we saw in the actual dataset—just a lot stronger.)

Module 1B: Multiple Regression Basics

- Execute chunks 7+8.
  - Chunk 8 shows you that the simulated data have high correlations between our predictor variables.

- Execute chunk 9. Compare the output of the regression model to the properties of the process that generated the data.
  - Execute chunks 7+9 repeatedly (say 5 times). Notice how the estimates of the effects of frequency and length are highly unstable.
- **Problem 3. Recall from slide 12 that this wasn't the case with the dataset simulated in chunk 4. Why is that the case?**

Module 1B: Multiple Regression Basics

- Examine what happens when inter-correlations are lower. Vary the noiseSD parameter in code chunk 7. This changes the standard deviation of the random, normally distributed variable used to generate frequencies. As this goes up, frequencies will be less correlated with length.
  - Note: at high levels, simulated frequencies will be negative. Don't worry about that, it's just a simulation!

- **Problem 4. Re-run chunks 7, 8 and 9; 5 times at each of the following levels of the noise SD parameter (1, 5, 10). How does intercorrelation relate to the success of regression in recovering the process that generated the predictions?**

Module 1B: Multiple Regression Basics

- Another issue for regressions is outliers—extreme observations that don't reflect the overall trends in the data.

- Execute code chunk 10. This builds a dataset with a few outliers—extremely long reaction times—built in.

- Execute code chunk 11. This graphs the reaction times. Note there is a small number of very long reaction times.

Module 1B: Multiple Regression Basics

- Execute code chunk 12. Examine the outputs of the linear model and the structure of the quantile-quantile plot.
- Note that the plot provides clear evidence of outliers; there are many points on the right edge of the plot that are far away from the quantile-quantile line.

- Execute code chunks 10+12 5 times. How well does the linear model do at recovering the properties of the process that generated the data?

Module 1B: Multiple Regression Basics

- Execute code chunk 13. This examines a subset of the data, removing RTs that are longer than 7.5 (recall this is log RT, so this is roughly > 1800 msec). Examine the outputs of the linear model and the structure of the quantile-quantile plot.

- Execute code chunks 10+13 5 times. How well does the linear model using trimmed data do at recovering the properties of the process that generated the data?

Module 1B: Multiple Regression Basics

- In code chunk 10, outlierProb controls the probability that an outlier will be generated.
- **Problem 5. Re-run code chunks 10 and 12 multiple times, varying the probability that an outlier will be produced (try 0.005 and 0.10). How does the probability of an outlier being produced influence the error in the linear model's estimate?**

Module 1B: Multiple Regression Basics

- Why are outliers bad?
  - When your dependent measure is highly skewed, the fit of the regression can be strongly biased by a few extreme measurements.
    - Note: this can also happen not just in the case of extreme outliers, but if the dependent measure is itself skewed
    - Ex: Speech sound duration typically follows gamma distributions; see code chunk 13
  - A similar problem can arise with skewed predictors; a small set of observations with an extreme value for a predictor can also skew your linear model's estimate of coefficients.

Module 1B: Multiple Regression Basics

- What can we do?
  - If outliers represent a few isolated cases, they can be removed manually (as in example code)
  - For cases where the measure is "naturally" skewed, transformation of dependent measures/predictors can also help (e.g., using log RT rather than raw RT)
  - BUT
    - Not always
    - Need to be careful about motivations behind particular transformation you're using.
    - Verify that the transformation actually helps! If observations are still highly skewed, this is not useful.

INTERIM SUMMARY

- Multiple regressions: Extension of simple models to multiple variables.

- Significance of predictors assessed via t-tests (individual coefficients), f-tests (whole model)

- Sensitive to inter-correlations among predictors, outliers
  - These issues can be examined through correlation tables, graphs of dependent measures and predictors, quantile-quantile plots of residuals