From Andries Coetzee and Pam Beddor:

Phonetics-Phonology Social Hour
Tuesday, July 9, 5-6:45 pm

Dominick's
812 Monroe St.

Just off Tappan Ave, only 2 minutes from Lorch Hall (location of Askwith Auditorium)
Class 5, 7/9/13: Logistic Regression, Significance Testing, Bias I
1. Exercise for next time
   • At end of these slides
2. Reading for next time

- Laura McPherson and Bruce Hayes (ms.) “Relating application frequency to morphological structure: the case of Tommo So vowel harmony”. On course website.
  - This uses a lexical database and maxent to make sense of a nice quantitative pattern discovered by McPherson in her fieldwork.
3. **Where we are**

- Maxent as a controversial but computationally highly viable framework for constraint-based grammar.
- Differential phonotactics, with the Seuss example
- A quick preview of logistic regression, which in our context is merely “binary maxent”
4. For today

- Looking at the Seuss grammar — overall performance
- Redo Seuss with logistic regression
- Testing constraints for statistical significance
- Bring in a new empirical example: Tommo So vowel harmony
5. Looking ahead

• Putting in UG: bias terms in maxent analysis
• One more framework: noisy harmonic grammar
• Comparing the frameworks empirically: Zuraw/Hayes work on intersecting constraint families
• Other stuff I could try to squeeze in if you ask
BACK TO SEUSS
6. Where I went wrong last time

- We found a bunch of environments that favor Seuss.
- These → constraints penalizing non-Seuss status.
- But non-Seuss status is very normal! Odd to just penalize it.
- Solution: a simple, powerful baseline constraint *SEUSS, against which the specific constraints are pitted.
7. **Class procedure for Seuss**

- Seussian environment to the board — *done*
- BH will then attempt to enter them all in with the Phonology Search Program, using our Seussian data file as the source — *done*
- We will edit the draft OTSoft output file so we can run it —
Source file: 1OTSoftFileRaw.txt
Relabel candidates
add *Seuss
Slide down the violations of the last three constraints, so they will penalize the Seuss words.
Grab the completed file, go to MyDocuments, and paste as MaxentInput.txt.
Run MaxentGrammarTool, creating the file MaxentOutput.txt.
Run (for convenience) FormatMaxentOutput.exe, and look at result, FormattedMaxentOutput.txt.
8. **Interpreting the maxent output**

- Put the weights on a separate sheet for convenience.
- Add a “one minus x” column for the predictions and observed — for sorting reasons.
- On the main sheet, add in an empty row at the top.
- Paste in the input file so we can refer to violations.
- Remove the duplicate columns and row.
- Sort the data by candidate, and delete all the NonSeuss cases (they are just Seuss minus 1).
- Save as InterpretedMaxentOutput.txt
SOME WAYS TO EYEBALL THE EFFECTIVENESS OF A STOCHASTIC ANALYSIS OF DIFFERENTIAL PHONOTACTICS
9. **Make a scattergram of predicted vs. observed**

- This is especially helpful when the scattergram is nicely **cloud-shaped** so the points don’t sit on top of each other.
  - Often we can’t get this in phonological modeling: too many points are both predicted/observed zero, or predicted/observed one.
  - Let’s do a scattergram of the Seuss model, which is going to have this problem.
10. Correlations

• Not necessarily a good idea, because they exaggerate model fit where there are huge numbers of zeros.
• We can nevertheless calculate it for our Seuss model.
11. Average error

- Assume a perfect model could assign a probability of 1 to every Seuss word, 0 to every non-Seuss word — zero average error.
- The average error of the mass-Seuss-model is .021.
- This is uninformative, since most forms have observed value 0 and predicted value close to zero.
12. Histograms

- Here is one way to do it.
- In Excel, sort descending by IsSeuss, Predicted
- Make a line graph for Predicted.

➤ This is not really very informative because the Seuss points are too compressed
13. Stretched histogram

- Keep the Excel sort just described.
- Make new columns in your spreadsheet that will go gradually from zero to one through each category.
<table>
<thead>
<tr>
<th>Seuss or not</th>
<th>Number the instances</th>
<th>List total number of cases</th>
<th>Divide numbering by total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>.25</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>.5</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>.75</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>6</td>
<td>.17</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>6</td>
<td>.33</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>6</td>
<td>.5</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>6</td>
<td>.67</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>6</td>
<td>.83</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
• Make a scattergram of the predicted values against the column labeled “Divide”.
• For Seuss (with a bit of touch-up), this yields:

![Graph showing scattergram of predicted values against 'Divide' column]

• Socrates: what color would be the diagram for a perfect model?
• Socrates: what would be the appearance of a diagram for the null model (no phonology, just the Intercept constraint\(^1\))

\(^1\) In substantive terms this is DON’T BE BY SEUSS.
14. Seeking improvement: Look at your outliers

- Find them: in Excel, for each datum compute predicted − observed, sort by this, and look at the top and bottom of the lists.
- It is perhaps of interest to look at the most Seussian real words
A BRIEF INTRODUCTION TO LOGISTIC REGRESSION
15. Logistic regression is stripped-down maxent

- As we saw last time, the math, and the predictions, are essentially the same.
- It is easy to do when there are just two candidate types, varying in parallel.
16. Logistic regression has some nice advantages

- It offers various statistical significance tests on your constraints.
  - So you can be confident that the inclusion of a constraint is meaningful.
- It’s available in various statistics software packages, including the ubiquitous R.
- You don’t have to know in advance which type of candidate a constraint penalizes.
17. Swooping into R

- R: a consortium of statisticians making their own package for free
- Source: http://www.r-project.org/
- R is diverse and powerful, and is continually growing as people (e.g. Harald Baayen) invent packages.
- If you save your “scripts”, you can repeatedly do the same things to different data.
- No one thinks R is easy — it’s very unforgiving of random user error, and error is hard to diagnose.
18. Hint for working with R

- Start with an **existing script** if you can.
- Keep your scripts as you do more projects — this will save you a great deal of work.
- Mooch from people who know more than you! Buy them chocolate to keep them happy.
19. A very simple format for R files

- Items (like words) in first column.
- A 1 or 0 in second column for what you’re trying to predict — here “Seuss”
- A bunch of other columns, with headers for the constraints that is violated — by one candidate or the other; it doesn’t matter which.
- Plain text, tab-separated.
- I made one from the OTSoft file; I could also have started with the output of the Search program. [show]
20. Three warnings about input file format

- Be **ultraconservative** in naming your variables. All-letter variables is always safe.\(^2\)
- R is **case-sensitive**
- *Never* put empty cells into your data: if missing values, enter “NA”.

\(^2\) Details: [http://cran.r-project.org/doc/FAQ/R-FAQ.html#What-are-valid-names](http://cran.r-project.org/doc/FAQ/R-FAQ.html#What-are-valid-names)
21. R as software

- This is super-easy compared to OTSoft; go to their website, download, install on your platform.
- To find their website: Googling the single letter r works!
22. What do to when you start R

- **File** menu, **Change dir**, go to where you keep your input file.
- **File** menu, **Open** script, find your script, click to open
23. The R user interface

- Unlike anything else in R, this is user friendly.
- Outline what you want to run in the script with your mouse, hit **Ctr r**, and it runs in the other window.
- I’ll do this for bits of script.
24. Bits of R code from my script

```r
library(languageR)
library(arm)
```

“Grab some libraries you’ll need later to run various commands” (when you learn about the command, you can usually learn about the library it’s in; and the libraries are easily loaded from the Packages menu on the R interface.)

```r
MyData=read.table("6RFileForSeuss.txt", header=T, sep="\t")
```

“Open the file TommoSoForR.txt, which has a header row, and is tab-separated. Put what you read into an object, called MyData, which may be referred to later as such.”
colnames(MyData)

“On the normal output screen, tell me all the column headers in order.” Thus:

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&quot;Word&quot;</td>
<td>&quot;Seuss&quot;</td>
<td>&quot;InitialVoicedObs&quot;</td>
<td>&quot;J&quot;</td>
<td></td>
</tr>
<tr>
<td>[5]</td>
<td>&quot;VoicedObLiquid&quot;</td>
<td>&quot;InitialZ&quot;</td>
<td>&quot;InitialY&quot;</td>
<td>&quot;StressedER&quot;</td>
<td></td>
</tr>
<tr>
<td>[9]</td>
<td>&quot;StresslessER&quot;</td>
<td>&quot;StressedUW&quot;</td>
<td>&quot;InitialFricNasal&quot;</td>
<td>&quot;StressedCaret&quot;</td>
<td></td>
</tr>
<tr>
<td>[13]</td>
<td>&quot;FinalSchwaL&quot;</td>
<td>&quot;TAtataTA&quot;</td>
<td>&quot;InitialR&quot;</td>
<td>&quot;StressedEY&quot;</td>
<td></td>
</tr>
<tr>
<td>[17]</td>
<td>&quot;VCCCV&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now the big move:
MyModel = bayesglm(Seuss ~ + InitialVoicedObs + VoicedObLiquid + InitialZ + InitialY + StressedER + StresslessER + StressedUW + InitialFricNasal + StressedCaret + FinalSchwaL + TAtataTA + InitialR + StressedEY, data = MyData, family="binomial")
“Use the bayesglm function to do regression.\(^3\)  

<table>
<thead>
<tr>
<th>Try to predict as best as you can the value of Seuss on the basis of 13 independent variables</th>
<th>bayesglm()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 for Seuss, 0 for not InitialVoicedObs etc.</td>
<td></td>
</tr>
<tr>
<td>The data you should fit is in the object MyData</td>
<td></td>
</tr>
</tbody>
</table>

\(^3\) You can also just say “glm”, General Linear Model. The Bayes version (http://www.stat.columbia.edu/~gelman/research/unpublished/priors7.pdf) is designed to be more accurate when some of your constraints prefer only winners.
| The particular kind of regression to be used is binomial logistic regression | family="binomial" |
25. Results

summary(MyBayesModel)

Puts a summary on the screen:
|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -5.05974 | 0.12897    | -39.231 | < 2e-16  *** |
| InitialVoicedObs | 0.72878 | 0.20773    | 3.508   | 0.000451 *** |
| J              | 0.59121  | 0.39852    | 1.484   | 0.137932 |
| VoicedObLiquid | 0.53608  | 0.33357    | 1.607   | 0.108028 |
| InitialZ       | 3.17286  | 0.34510    | 9.194   | < 2e-16  *** |
| InitialY       | 2.13120  | 0.36586    | 5.825   | 5.71e-09 *** |
| StressedER     | 0.22876  | 0.39016    | 0.586   | 0.557658 |
| StresslessER   | -0.30089 | 0.24592    | -1.224  | 0.221124 |
| StressedUW     | 0.81472  | 0.25799    | 3.158   | 0.001589 ** |
| InitialFricNasal | 3.24560 | 0.32711    | 9.922   | < 2e-16  *** |
| StressedCaret  | 1.43239  | 0.20439    | 7.008   | 2.41e-12 *** |
| FinalSchwaL    | 0.01132  | 0.28374    | 0.040   | 0.968175 |
| TAtataTA       | 1.93705  | 0.35953    | 5.388   | 7.14e-08 *** |
| InitialR       | -1.11585 | 0.63532    | -1.756  | 0.079025 . |
| StressedEY     | -2.48775 | 0.81136    | -3.066  | 0.002168 ** |
| VCCCV          | -1.27930 | 0.63210    | -2.024  | 0.042980 * |
• “Intercept” is a constraint that R puts in by default. It favors one of the two outcomes.
➢ To leave it out, put “-1” after the last variable and before the comma.
• Socrates: what does positive or negative sign for weights mean?
• The last column is informative: there are constraints with positive weights, but a test known as the Wald test tells us not to take them seriously anyway.
• Socrates: look at the system and find some cases of constraint ganging.
26. bayesglm() vs. plain glm()

- bayesglm, devised by the statistician Andrew Gelman, is usually a better choice for linguists than classical glm()
- It’s devised not to freak out when there is unanimity
  - Commonly and rightly *expected* by linguists, problematic for classical glm()
# 27. Weights of maxent vs. logistic regression

<table>
<thead>
<tr>
<th></th>
<th>bayesglm() in R</th>
<th>Maxent Grammar Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept/*Seuss</td>
<td>-5.06</td>
<td>5.32</td>
</tr>
<tr>
<td>InitialVoicedObs</td>
<td>0.729</td>
<td>0.787</td>
</tr>
<tr>
<td>J</td>
<td>0.591</td>
<td>0.649</td>
</tr>
<tr>
<td>VoicedObLiquid</td>
<td>0.536</td>
<td>0.579</td>
</tr>
<tr>
<td>InitialZ</td>
<td>3.173</td>
<td>3.370</td>
</tr>
<tr>
<td>InitialY</td>
<td>2.131</td>
<td>2.230</td>
</tr>
<tr>
<td>StressedER</td>
<td>0.229</td>
<td>0.409</td>
</tr>
<tr>
<td>StresslessER</td>
<td>-0.301</td>
<td>0</td>
</tr>
<tr>
<td>StressedUW</td>
<td>0.815</td>
<td>0.945</td>
</tr>
<tr>
<td>InitialFricNasal</td>
<td>3.246</td>
<td>3.446</td>
</tr>
<tr>
<td>StressedCaret</td>
<td>1.432</td>
<td>1.517</td>
</tr>
<tr>
<td>FinalSchwaL</td>
<td>0.011</td>
<td>0.093</td>
</tr>
<tr>
<td>TAtataTA</td>
<td>1.937</td>
<td>1.596</td>
</tr>
<tr>
<td>InitialR</td>
<td>-1.116</td>
<td>0</td>
</tr>
<tr>
<td>StressedEY</td>
<td>-2.488</td>
<td>0</td>
</tr>
<tr>
<td>VCCCCV</td>
<td>-1.279</td>
<td>0</td>
</tr>
</tbody>
</table>
- Why is Maxent Grammar Tool giving zeros?
28. How can we interpret a weight?

- In maxent/logistic regression, we can look at a pure case: what if it were the only constraint in the grammar?
- Let’s go back to the maxent-cum-logistic regression formula:

\[
\frac{1}{1 + e^{(\text{HarmonyOfB} - \text{HarmonyOfA})}}
\]

- We’ll imagine just one single constraint, with weight \( w \), violated by \( B \).
So have just:

\[
\frac{1}{1 + e^w}
\]

Where \( w \) is zero, this comes out as one half: even odds — the maximum entropy answer with no effect of grammar.

Scaling up \( w \), we get a greater penalty for \( B \), with half of a sigmoid curve (logit function).
29. The log-odds principle

- In the same spread sheet, we can take \( P(B)/P(A) \) (A is the other binary candidate), which a gambler would call the **odds** of B.
- A nice theorem tells us that the weight of a single-constraint grammar is exactly the same as the log odds — see spreadsheet for this.
30. Getting out of R

- You get two prompts:
- Never save your workspace — it will produce phantom values in later runs without telling you!
- Save your script according to your judgment.
A BIT OF HUNGARIAN
31. Like maxent, logistic regression doesn’t use as many constraints

- When there are two opposing constraints, it can give the weaker one a zero weight.
- Why so?
  - This is a use of the ability of maxent grammars to assign positive probability to harmonically bounded candidates
- So: if you use a “standard” set of OT constraints, with opposing pairs, you may get zero weights/nonsignificant results for certain constraints
- Trimback may be needed
32. Another form of trimback

- Unless you put “-1” after the last constraint, R will include an intercept term; i.e. “prefer on particular kind of candidate”
- If this job is done in your analysis by a sensible set of conflicting constraints, you would want to not use this intercept.
33. Demo: Hungarian

- We’ll turn our OTSoft file into an R file, run logistic regression, and do trimback
- We can use the Akaike Information Criterion (model evaluation statistic penalizing redundancy) to show that we’re getting better by removing intercept and redundant constraints
INTRODUCTION TO BIAS
34. What is bias (in this context)?

- Any principle that keeps the learning system from achieving its objective function
- = something that wants the grammar to be some particular way, even at the cost of pure empirical accuracy
35. Why would we want to impose bias?

- More than one reason:
  - Proper skepticism when data are few — don’t leave to conclusions just because that would be the best-fit model.
  - Substantive biases — the child expects the ambient language to have certain properties, and resists learning pure data-fit grammars when these properties are contradicted.
36. Skepticism under few data

- When you hear, say, 5 examples of output A and 0 examples of output B, you can, in principle, set the weight of PREFERA extremely high, and get a superb fit.

- Let’s try this in two softwares, differing in their biases
  - Maxent in OTSoft goes immediately for a weight of 50, and achieves a huge number of decimal places of accuracy. Hooray!
  - Maxent Grammar Tool, in its default settings, computes weight of 10.1, and accuracy is to five decimal places: 0.99996.

- Many people would prefer the more conservative value, or even more conservative.

- Suppose instead that there are one million examples of A, 0 of B.
Maxent in OTSoft: same as before.
Maxent Grammar Tool now computes weight of 21.6; accuracy is to ten decimal places.

So it seems sensible to want weights to be small, pending enough evidence to raise them high.
37. **A common way to express suitable skepticism in maxent: the Gaussian Prior**

- For every constraint we stipulate two values

\[ \mu = \text{the preferred value} \]
\[ \sigma = \text{the degree to which the weight is “willing” to diverge from } \mu \text{ under the pressure of learning data} \]

- For conservativity against few data, we would posit \( \mu = 0 \).
- For a constraint that we think is a priori ranked highly in UG (Markedness? OO-Faithfulness?), we posit a high value of \( \mu \).
• Alternative approaches might express a “potentially strong” constraint with a low \( \sigma \).
38. How to specify a prior in the Maxent Grammar Tool

- Make a little file with a name like Prior.txt.
- Each line: Constraints name, μ, σ, separated by tabs.
- Minor bug: you have to put a minus sign in from of your mu’s if you want them to be positive; sorry.
- When you run the program, access Prior.txt with the middle button, labeled “Open constraints.”
39. Making maxent more skeptical with a prior

- We’ll do a million learning cases again.
- But we’ll set sigma at 1, instead of the program default of 100000.
- Accuracy descends to just four decimal places — still high, since a million examples is a great deal of data.
40. Coming up

- Applications of the prior
  - Evidence that human learners are skeptical in the face of insufficient data.
  - Evidence that human learners actively prefer to give stronger weights to certain constraints.
41. Exercise

- Go to the course web site and download the R input file used today for Seuss.
- Think of some simple constraint you can assess the violations of using just the spelling, and assess them in Excel, adding it to the R file.
  - A really simple constraint would simply assess the length of words in letters.
  - Or there are various other constraints that are accurately assessable from spelling, e.g. “Seuss if starts with th”
- Save your work in text format.
- Install R on your machine. Open the script from the course web site.
• Modify the script so that the logistic regression includes your new constraint.
• Run the script, line by line, and use the result to assess the merits of the new constraint you proposed.