LAB: *DO*-SUPPORT AND THE CONSTANT RATE EFFECT

W. GARRETT MITCHENER

1. Linguistic Background

The term *do-support* refers to the use of *do* as an auxiliary verb, as in the examples in Appendix A. This is also called *periphrastic do,* where *periphrastic* means that a special word rather than morphological inflection is used to carry meaning. In Modern English, tense and agreement with the subject are often carried by *do, does*, and *did*, where in French for example they would be carried by endings *-e, -es, e, -ons, -ez, -ent,* ....

Old and Middle English were verb-raising languages, meaning that the main verb could be raised to receive markings for tense, person, and number. If there is an auxiliary verb, such as *be* or *have,* it appears in the higher position and receives tense, person, and number while the main verb remains low and appears as a non-finite form, such as a participle. French, Spanish, Latin, and German are all verb-raising languages.

Modern English does not have verb raising. Instead, the main verb remains low, and tense markings (such as the *-s* in "He knows") move downward. However, if negation intervenes between the tense and the main verb, it blocks affix hopping, leaving the tense stranded. So in sentences with *not* before the main verb, we have to put in a dummy auxiliary verb *do* to carry the tense and subject agreement. When forming questions, the tensed verb moves to the front to a position called C for *complementizer*, and the query word, *who, what, etc.* moves to the specifier position of CP. In forming questions in Modern English, we generally use *do* because the main verb remains low and can't raise to C. By the way, there is no universally accepted explanation of why *do* is forbidden in affirmative questions on subjects. (Just so you'll know, questions on subjects are weird in other situations as well, and across languages.)

English switched from verb-raising to *do*-support over the time period from 1350 to 1650. During that time, writers used a mixture of two grammars, one with verb-raising and one with no verb-raising but with *do*-support. There is evidence that something else began to change at around 1550, so this lab will focus on the years 1350–1550, which is the late Middle English period. Shakespeare and the King James

Date: January 13, 2005.

Bible are from Early Modern English, which still used some features of Middle English syntax.

2. The logistic model and the constant rate effect

Following Kroch [1], we will model the rise of *do*-support with a logistic curve. The logistic curve has the form

$$y(t) = \frac{e^{k+st}}{1+e^{k+st}} = \frac{1}{1+e^{-(k+st)}}. \tag{1}$$

We will call k the *intercept* parameter and s the *slope* parameter.

This function solves the ordinary differential equation

$$y'(t) = s\,y\,(1-y). \tag{2}$$

Deliverable #1: Find all fixed points of (2) and classify them as stable or unstable. Consider the cases of $s > 0$ and $s < 0$. Draw phase portraits for each case.

Deliverable #2: The differential equation (2) is separable. Separate variables and use partial fractions to solve for y as a function of t. Simplify your solution so that it looks like (1).

Deliverable #3: We will use this equation to model the rate at which writers of Middle English used *do*-support. Therefore, $y(t)$ should be a number between 0 and 1 for all time. Prove that if $y(t_0)$ is between 0 and 1, then $y(t)$ is between 0 and 1 for all time.

3. Curve fitting

Our data consists of a table of numbers stating that among particular classes of sentences taken from manuscripts dated t_j, m_j use *do* and n_j do not.

We are interested in how the frequency $m_j/(m_j+n_j)$ of the use of *do* changes over time, and in particular, we would like to determine which sigmoid function $y(t)$ matches the frequency data "best." There are lots of different ways of fitting curves to data, all with different notions of "best." Mathematically, we pick some means of measuring the distance between a function $y(t)$ and the data points $\{(t_j, y_j)\}_{j=1,\dots,n}$.

3.1. Least squares.

A traditional method of fitting linear data is called *least squares* and is based on minimizing the sum of the squares of the differences between the given data and the fitting line. There are two ways of using least squares. The first (and computationally the simplest) is to linearize the data, that is, choose a function f

such that $f(y(t)) = at + b$ then apply the formula for linear regression to the data $(t_j, f(y_j))$ to find a and b. Most scientific calculators (even the non-graphing ones) can perform linear regression, and mathematical software certainly can.

Deliverable #4: Let y be a sigmoid function of t as in (1). Solve for $k + s\,t$ as a function of y. This function of y is known as the *logit* function, and you will often see graphs of $\operatorname{logit} y(t)$ as a function of t because it results in approximately linear plots. The (dimensionless) units of $\operatorname{logit} y(t)$ are also sometimes called *logits.*

Deliverable #5: Using Ellgård's data for intransitive affirmative questions, form the data set $(t_j, \operatorname{logit} m_j/(m_j + n_j))$ giving logit of frequency as a function of time. Instead of years AD, represent time in centuries since 1350 AD. (If you use years AD, you will run into numerical problems later in the lab.) Plot this linearized data. Come up with at least one way to handle the cases of $m_j = 0$ and $n_j = 0$ and discuss the advantages and disadvantages.

Deliverable #6: Use linear regression to fit a line to the linearized data. Plot the resulting sigmoid function against the original non-linear data.

Deliverable #7: How sensitive is the fit? What would happen if someone discovered that Ellegård missed a sentence that used *do* in the first time period?

3.2. **Non-linear least squares.** An alternative is to use least squares directly on the non-linear data. This is a much more difficult computation, but it has the advantage of bypassing the problem with $\log 0$ that came up when we tried to linearize the frequency data. We define the error for a particular function $f(t)$ to be

$$E(f) = \sum_{j=1}^{n} |f(t_j) - y_j|^2 \tag{3}$$

and use a numerical method to adjust the parameters of f so as to minimize this error.

Deliverable #8: Using Ellgård's data for intransitive affirmative questions, form the data set $(t_j, m_j/(m_j + n_j))$ giving frequency as a function of time. As before, represent time a centuries since 1350 AD. Plot this non-linear data.

Deliverable #9: Define $f(t) = \frac{1}{1+e^{-(k+st)}}$. Use a numerical method to select s and k so as to minimize $E(f)$ for this data. Plot the resulting logistic curve on top of the data points.

Deliverable #10: How sensitive is the fit? What would happen if someone discovered that Ellegård missed a sentence that used *do* in the first time period?

3.3. **Maximum likelihood curve fitting.** The linear and non-linear least squares fits only use the usage frequency $m_j/(m_j+n_j)$ and throw away the magnitude of m_j and n_j. That is, they do not take into account all the available information. Specifically, they treat a data point $m = 2, n = 2$ exactly the same as $m = 200, n = 200$ because both yield a usage frequency of 0.5. However, there is much more uncertainty in that 0.5 in the case $m = 2, n = 2$

Deliverable #11: Why is there more uncertainty in the frequency if m and n are small?

The method used by Kroch is called *maximum likelihood* and has a direct interpretation in terms of probability. Furthermore, it makes use of all the data, incorporating the uncertainty that comes from the varying sizes of samples at different times.

Let's restrict our attention to one class of sentences, say, intransitive affirmative. At time t we suppose that speakers of Middle English use *do*-support in such sentences with probability $y(t)$.

Deliverable #12: Suppose we select intransitive affirmative sentences at random at time t. What expression represents the probability that our sample consists of m sentences that use *do* and n sentences that do not? (Where $n+m$ is the sample size.)

Deliverable #13: Suppose we look at sentences selected at random at times $t_1, t_2, \ldots, t_k$. What expression represents the probability that our sample consists of m_j sentences with *do* and n_j sentences without for time t_j's? (Where $n_j + m_j$ is the sample size at t_j.) This is called the *likelihood* of getting the data, given $y(t)$.

For maximum likelihood curve fitting, we define $p(s,k)$ to be the likelihood of getting the data given $y(t) = \frac{1}{1+e^{-(k+st)}}$, and choose $\hat{s}$ and $\hat{k}$ so as to maximize p:

$$(\hat{s}, \hat{k}) = \arg\max_{(s,k)} p(s,k) \tag{4}$$

Deliverable #14: Why does maximizing $\log p(s,k)$ give the same values of $\hat{s}$ and $\hat{k}$ as maximizing $p(s,k)$? (It is good to do this because p tends to be extremely small in magnitude, and the numerical methods work much better on $\log p$.)

Deliverable #15: Why can you discard the binomial coefficients from $\log p(s,k)$ when searching for $\hat{s}$ and $\hat{k}$? (It's good to do this because it cuts down on the amount of calculations we have to do in finding $\hat{s}$ and $\hat{k}$.)

Deliverable #16: Using the data for intransitive affirmative sentences, find $\hat{s}$ and $\hat{k}$.

Deliverable #17: How sensitive is the fit? What would happen if someone discovered that Ellegård missed a sentence that used *do* in the first time period?

Deliverable #18: Use this method to find $\hat{s}$ and $\hat{k}$ for all classes of sentence types. For each class, plot the data points and maximum likelihood sigmoid.

4. The constant rate effect

Part of the point of doing all this is to confirm that the rate parameter s is actually the same for all classes of sentences. This means that the rise of *do*-support follows the same dynamics (2) across all classes of sentences, which would suggest that they are all processed similarly in the brain. An alternative hypothesis is that the rate constant in (2) is different for different classes of sentences, which would suggest that different classes of sentences are processed differently in the brain.

Here are the values of $\hat{s}$ and $\hat{k}$ as found by Kroch:

- negative declarative: $\hat{s} = 3.74, \hat{k} = -8.33$
- negative questions: $\hat{s} = 3.45, \hat{k} = -5.57$
- affirmative transitive and yes/no questions: $\hat{s} = 3.62, \hat{k} = -6.58$
- affirmative intransitive yes/no questions: $\hat{s} = 3.77, \hat{k} = -8.08$
- affirmative object questions: $\hat{s} = 4.01, \hat{k} = -9.26$

Deliverable #19: Chances are, you didn't find exactly these numbers in your maximum likelihood calculations. In fact, when I asked Kroch for a copy of his data, he re-ran his spreadsheet and got different numbers that were off by a few hundredths from what is reported in his article. There is no numerical uncertainty about the counts, unlike measuring a length with a ruler where the measurement is only accurate to within a millimeter or so. The differences in results are due to numerical instability in finding the maximum of $p(s, k)$. Identify the source of this instability. How severe is it?

Deliverable #20: Assume for the moment that our numerical instability issues have negligible influence on the values of $\hat{s}$. Here's an open ended question: We observe that $\hat{s}$ ranges from about 3.45 to 4.01 over the different contexts. Is this variation due to random chance, or are those differences significant? That is, do the maximum likelihood calculations support the constant rate hypothesis or not? Justify your answer carefully and thoroughly. (As in this should be a pretty long answer.) Include whatever calculations and mathematical arguments are necessary.

Appendix A. *Do*-support examples

(5) He knows the muffin man.
(Affirmative declarative, *do* forbidden.)

(6) He *does* know the muffin man.
(Emphatic affirmative declarative, *do* allowed only when contradicting a claim that he doesn't know the muffin man.)

(7) He doesn't know the muffin man.
(Negative declarative, *do* required.)

(8) He can't know the muffin man.

(9) *He not knows the muffin man.
(Negation must follow tense.)

(10) *He knows not the muffin man.
(Verb raising, acceptable in Shakespeare's era, not acceptable today.)

(11) Do you know the muffin man?
(Affirmative yes/no question, *do* required.)

(12) Don't you know the muffin man?
(Negative yes/no question, *do* required.)

(13) *Know you the muffin man?
(Verb raising, acceptable in Shakespeare's era, not acceptable today.)

(14) *Know you not the muffin man?
(Verb raising, acceptable in Shakespeare's era, not acceptable today.)

(15) Who knows the muffin man?
(Affirmative question on subject, *do* forbidden.)

(16) Who *does* know the muffin man?
(Emphatic affirmative question, *do* allowed only for exceptional meaning, as in a rhetorical question suggesting that no one knows the muffin man.)

(17) Who doesn't know the muffin man?
(Negative question on subject, *do* required.)

(18) *Who knows not the muffin man?

(19) Who do you know?
(Affirmative question on object, *do* required.)

(20) *Who know you?

(21) Who don't you know?
(Negative question on object, *do* required.)

(22) *Who know you not?

Appendix B. Ellegård's data as reported by Kroch

Date	Neg Decl	Neg Q	Aff Trans Q	Aff Intrans Q	Aff Obj Q
1412.5	0, 177	2, 15	0, 3	0, 7	0, 1
1450	11, 892	2, 23	6, 50	0, 86	0, 27
1487.5	33, 660	3, 24	10, 64	0, 68	1, 50
1512.5	47, 558	46, 32	22, 69	19, 71	7, 55
1530	89, 562	34, 22	18, 8	15, 61	6, 57
1542.5	205, 530	63, 21	56, 35	37, 79	8, 65
1562.5	119, 194	41, 7	42, 15	30, 41	27, 48
1587.5	150, 479	83, 45	137, 36	91, 114	46, 74
1612.5	102, 176	89, 6	214, 63	192, 118	51, 120
1637.5	109, 235	32, 6	60, 6	56, 18	35, 31
1675	126, 148	48, 4	72, 4	92, 39	28, 23
1725	61, 9	16, 0	33, 0	20, 3	11, 4

Ellegård's data consists of counts of sentences found in manuscripts grouped into time periods, and each row represents one period. The dates of most manuscripts are approximate, so the date for time period is the midpoint of the estimated dates for manuscripts within that group. The actual periods are 1400–1425, 1425–1475, 1475–1500, 1500–1525, 1525–1535, 1535–1550, 1550–1575, 1575–1600, 1600–1625, 1625–1650, and 1650–1700. The use of *do*-support increases throughout. Something different happened around 1550, so for the purposes of this project, use only the data from the first seven rows of this chart, that is, up to and including the 1562.5 row where the line is.

The columns represent different classes of sentences: negative declarative, negative yes/no questions, affirmative transitive yes/no questions, affirmative intransitive yes/no questions, and affirmative object questions. (Transitive verbs take a direct object, where intransitive verbs do not.)

The entries in each cell are m, n where m sentences were found that use *do* and n were found that do not.

References

[1] Anthony Kroch. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244, 1989.

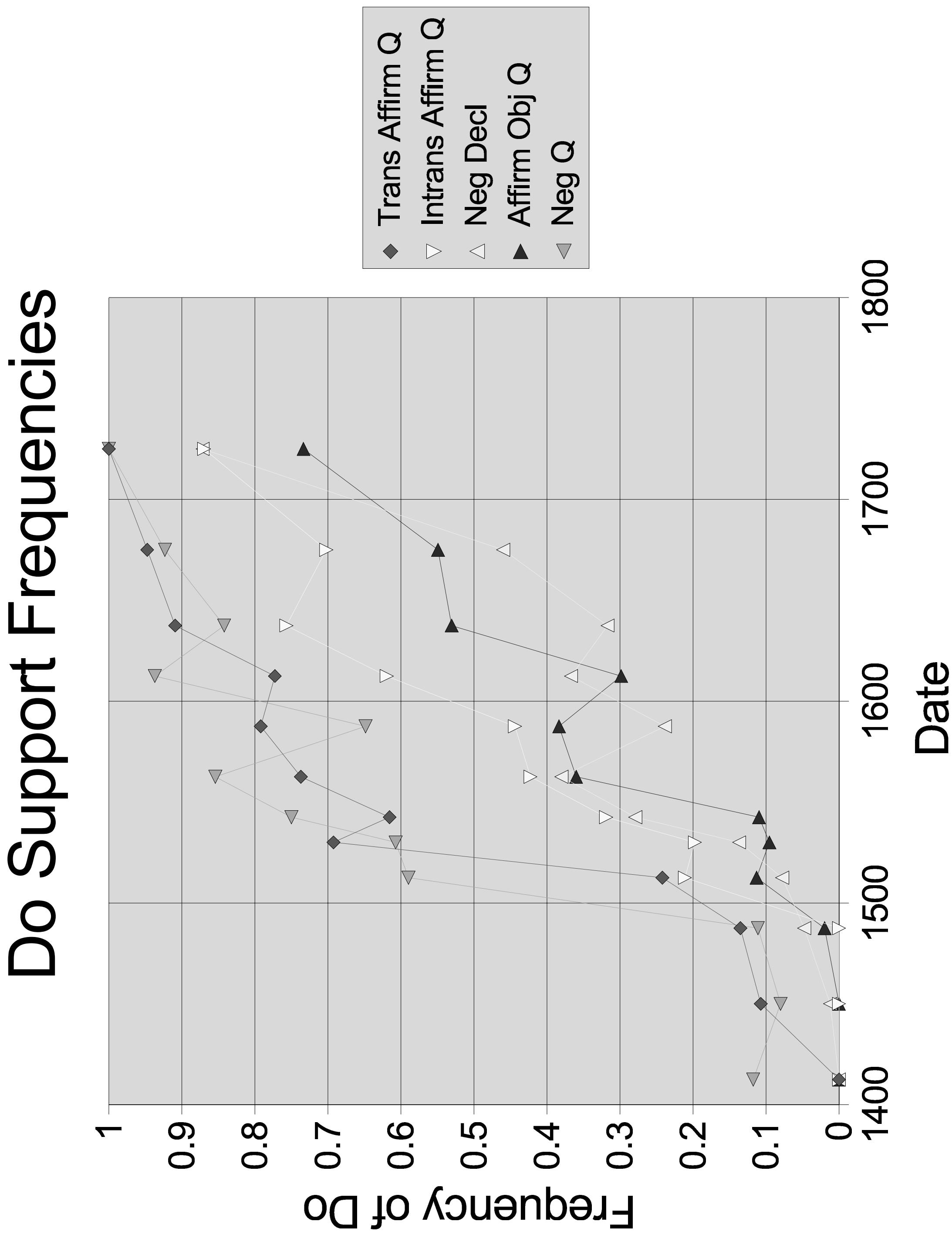

Figure 1. Point plot of *do*-support data.