5 Sociolinguistics

The main data that we study in sociolinguistics are counts of the number of realizations of sociolinguistic variables. For example, a phonological variable might span the different realizations of a vowel. In some words, like *pen*, I say [ɪ] so that *pen* rhymes with *pin*, while other speakers say [ɛ]. The data that go into a quantitative analysis of this phonological variable are the categorical judgements of the researcher - did the talker say [ɪ] or [ɛ]? Each word of interest gets scored for the different possible pronunciations of /ɛ/ and several factors that might influence the choice of variant are also noted. For example, my choice of [ɪ] in pen is probably influenced by my native dialect of English and the fact that this /ɛ/ occurs with a following /n/ in the syllable coda. Perhaps also, the likelihood that I will say [ɪ] is influenced by my age, socioeconomic status, gender, current peer group, etc.

Other sociolinguistic variables have to do with other domains of language. For example, we can count how many times a person uses a particular verb inflection and try to predict this morphological usage as a function of syntactic environment, social group, etc. Or we could count how many times a person uses a particular syntactic construction, and try to model this aspect of language behavior by noting relevant linguistic and social aspects of the performance.

The key difference between these data and the data that we typically deal with in phonetics and psycholinguistics is that the critical variable - the dependent measure - is nominal. We aren't measuring a property like formant frequency or reaction time on a continuous scale, but instead are noting which of a limited number of possible categorical variants was produced by the speaker. So, in this chapter we turn to a couple of different analytical techniques to find patterns in these nominal response measures.

Of course, other areas of linguistics also deal with nominal data. In phonetics we sometimes count how many times a listener chooses one alternative or another in a listening task. In phonology we may be interested in how often a feature is used in the languages of the world, or how often a "free" variant pronunciation is used. In syntax, as we will see in chapter 7, we analyze counts of the number of times particular constructions are used in different contexts. The methods discussed in this chapter on sociolinguistics are thus applicable in these and other subdisciplines of linguistics.

5.1 When the data are counts - contingency tables.

We can compare the observed frequency of occurrence of an event with its theoretically expected frequency of occurrence using the χ^2 distribution. In some situations you can posit some expected frequency values on the basis of a theory. For example, you might expect that the number of men and women in a statistics class to be about equal because there are about as many

men as there are women in the world. So if the class has a total of 20 students the expected frequency of men is 10 and the expected frequency of women is 10.

In another type of case, if we assume that a set of observations comes from a normal distribution then we should find that most of the observations fall near the mean value and that a histogram of the data should have frequency counts that fits the normal curve defined by the data set's mean and standard deviation.

The difference between the observed counts and counts expected given a particular hypothesis, say that there should be an equal number of men and women in the class or that the data should follow a normal curve, can be measured on the χ^2 distribution. If the difference between observed and expected frequency is much greater than chance you might begin to wonder what is going on. Perhaps an explanation is called for.

To calculate χ^2 from observed and expected frequencies you sum over all of the cells in a contingency table the squared difference of the observed count (o = say 5 men in the class) minus the expected count (e = 10 men) divided by the expected count. For the case in which we have 5 men and 15 women in a class of 20, and we expect 10 men and 10 women, the χ^2 value that tests the accuracy of our expectation is $\chi^2 = (5\text{-}10)^2/10 + (15\text{-}10)^2/10 = 2.5 + 2.5 = 5$.

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$ calculating χ^2 from observed and expected counts

To determine the correctness of the assumption that we used in deriving the expected values, we compare the calculated value of χ^2 with a critical value of χ^2. If the calculated (observed) value of χ^2 is larger than the critical value then the assumption that gives us the expected values is false. Because the distribution is different for different degrees of freedom you will need to identify the degrees of freedom for your test. In the case of gender equity in education, because we have two expected frequencies (one for the number of men and one for the number of women in a class) there is 1 degree of freedom. The probability of getting a χ^2 value of 5 when we have only 1 degree of freedom is only p=0.025, so the assumption that men and women are equally likely to take statistics is probably (97 times in a 100 cases) when there are only 5 men in a class of 20. The remainder of this section explores how χ^2 can be used to analyze count data in contingency tables.

The way that the χ^2 test works is based in the definition of the χ^2 distribution as a sum of squared z scores. In otherwords the χ^2 distribution is just a particular way of looking at random variation. Because the z scores are squared the χ^2 distribution is always positive, and because we expect a certain amount of randomness to be contributed by each z score that is added to the sum

the χ^2 probability density distribution peaks at higher χ^2 values and becomes flatter as the number of z scores increases (see figure 5.1).

$$\chi^2_{(n)} = \sum_i^n z_i^2 = \sum_i^n \frac{(y_i - \mu)^2}{\sigma^2}$$ The definition of the χ^2 distribution

Notice that the expression for how to calculate χ^2 from observed and expected frequencies has exactly the same form as the expression of χ^2 in terms of z scores. This is why you can use the χ^2 distribution to measure how different the expected and observed frequencies are. We let f_e serve as our best estimate of σ^2 and use this to convert the differences between observed and expected frequencies into squared z scores for evaluation with χ^2.

In figure 5.1 you can see that the most highly probable value of χ^2 (the peak of the probability density function) is always just a little below the number of degrees of freedom of the statistic. χ^2 is different from the ratio statistics we discussed in Chapters 3 and 4. No matter what the degrees of freedom were, we expected the *t* and *F* statistics to be approximately equal to 1 if the null hypothesis is true and substantially larger than 1 if the null hypothesis is false. With χ^2 on the other hand, we expect the statistic to be almost equal to the degrees of freedom if the null hypothesis is true and substantially larger than the degrees of freedom if the null hypothesis should be rejected. This is because the expected value of z^2, on average over all y_i, is 1, and χ^2 is a sum of squared *z* scores (this is because the average deviation from the mean in any data set is by definition a "standard" deviation).

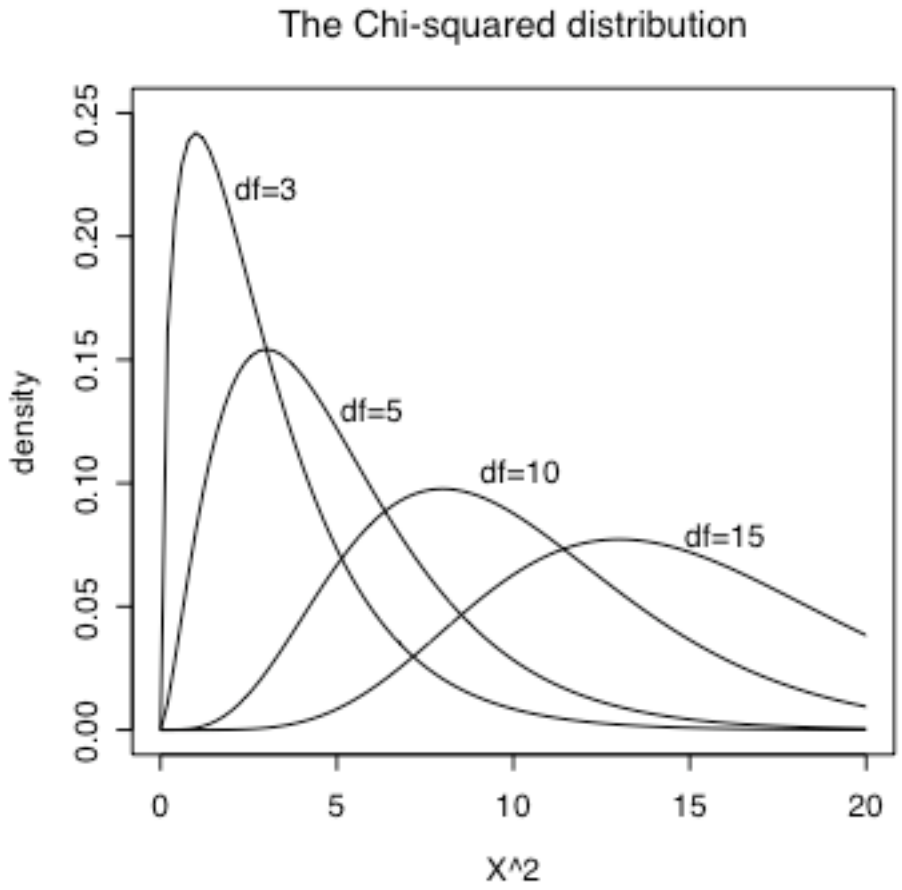

Figure 5.1. The χ^2 distribution at four different degrees of freedom. The peak of the probability density function is at a higher value for versions of the χ^2 distribution with higher degrees of freedom (e.g. the peak of the df=15 distribution is near χ^2=13 while the peak of the df=10 distribution is near χ^2=8)

5.1.2 Frequency in a contingency table.

A contingency table is the count of events classified two ways - on two variables. For example, we might count the number of times that coda /l/ is "vocalized" in the speech of people of different ages. (Robin Dodsworth did this in Worthington, OH and has kindly shared her data for this example. The pronunciation difference under examination here is between /l/ produced with the tongue-tip against the alveolar ridge - usually also with the tongue body raised, a "dark" /l/ - versus a realization in which the tongue tip is not touching the roof of the mouth so that the /l/ sound would be transcribed as [w], or even a vowel off-glide like [ʊ].) The contingency table (see Table 5.1) has columns for each age level, and rows for the two realizations of /l/. If the rate of /l/ vocalization does not differ for people of different ages then we expect the proportion of /l/ vocalization to be the same for each age group. That is, our expectation for each age group is

guided by the overall frequency of /l/ vocalization disregarding age.

Table 5.1. A two way contingency table of the frequency of /l/ vocalization as a function of talker age. The expected table shows the expected values, which were calculated from the marginal proportions, assuming that age groups did not differ from each other.

		teens	twenties	fourties	fifties	total	prop.
Observed	unvocalized	91	47	36	44	218	0.41
	vocalized	110	123	26	54	313	0.59
	total	201	170	62	98	531	
	proportion	0.38	0.32	0.12	0.18		1
Expected	unvocalized	83	70	25	40	218	0.41
	vocalized	118	100	37	58	313	0.59
	total	201	170	62	98	531	
		0.38	0.32	0.12	0.18		
((o-e)^2)/e	unvocalized	0.87	7.44	4.37	0.35		
	vocalized	0.61	5.18	3.04	0.24		
					χ^2		22.18

This way of deriving expectations in a contingency table is exactly analogous to the way that I derived expected frequencies for the gender composition of statistics classes, however the expected values are taken from the overall observed proportion of vocalized and unvocalized utterances in the study rather than from population demographics (50% men).

So, in table 5.1, the expected frequency of unvocalized /l/ for teens is 83 and this expectation comes from taking the total number of observations of /l/ produced by teens multiplied by the overall proportion of unvocalized /l/ (0.41*201 = 83). If you want you can get the same expected number of unvocalized /l/ for teens by taking the total number of unvocalized /l/ and multiply that by the proportion of observations coming from teens (0.38 * 218 = 83). The expected number is calculated on the assumption that teens are just as likely to produce unvocalized /l/ as any of the other age groups. If, for example there was a tendency for teens to produce vocalized /l/ almost all of the time, then the expected value that we derive by assuming that teens are no different would be wrong. χ^2 tests the independence assumption inherent in these expected values by summing up how wrong the expected values are. As table 5.1 shows, our expected values are pretty close to the observed values for teens and fifties,but people in their twenties and fourties differed quite a lot from the expected frequencies (twenties showing less vocalization than expected and fourties showing more). These deviations from the expected counts are enough to

cause the overall x2 to total to 22.1 which with 3 degrees of freedom is large enough to reject the null hypothesis that age doesn't matter for predicting /l/ vocalization. The degrees of freedom for this test is (number of age levels - 1) times (number of /l/ vocalization types -1), which works out in this case to (4-1)*(2-1) = 3.

R note. I recoded Dodsworth's data a bit for this example. She had coded age with 1 for teens, 2 for twenties, and so on. To produce the contingency table, this would have to be converted into a factor anyway, so in the factor statement I added new easy to read labels for the age groups. I did the same thing for Dodsworth's "/l/ vocalization" factor. She scored productions as 1 for "unvocalized", 3 for "vocalized" and 2 for "intermediate". There were only 11 productions scored as "intermediate" (out of 542 total observations) so I decided to exclude them from the dataset.

```
> rd <- read.delim("Robins_data.txt")
> rd$newage <- factor(rd$age,levels=c(1,2,4,5),
     labels=c("teens","twenties","fourties","fifties"))
> rd$lvoc <- factor(rd$l,levels=c(1,3),
     labels=c("unvocalized","vocalized"),exclude=c(2))

    talker age l   gender conscious   newage         lvoc
1       bh    5 3 female     connect  fifties   vocalized
2       bh    5 1 female     connect  fifties unvocalized
3       bh    5 3 female     connect  fifties   vocalized
4       bh    5 3 female     connect  fifties   vocalized
5       bh    5 3 female     connect  fifties   vocalized
6       bh    5 3 female     connect  fifties   vocalized
7       bh    5 1 female     connect  fifties unvocalized
8       bh    5 1 female     connect  fifties unvocalized
9       bh    5 3 female     connect  fifties   vocalized
10      bh    5 2 female     connect  fifties        <NA>
11      bh    5 1 female     connect  fifties unvocalized
...
```

The frequency table that has counts for /l/ vocalization as a function of age is produced by `table()`, and the χ^2 test of the independence of lvoc and newage is given by `summary(table())`.

```
> attach(rd)
> table(lvoc,newage)
            teens twenties fourties fifties
unvocalized    91       47       36       44
```

```
vocalized   110   123       26       54

> summary(table(lvoc,newage))
Number of cases in table: 531
Number of factors: 2
Test for independence of all factors:
Chisq = 22.118, df = 3, p-value = 6.166e-05
-----------------
```

5.2 Working with probabilities - the binomial distribution.

It is often the case in analyzing the frequency of occurrence of a linguistic variable that we are dealing with binomial probabilities. That means that we classify each production as to whether a process applied or not - so we could code the data as 1 (process applied), 0 (process did not apply). There are number of ways of treating binomial probabilities in quantitative analysis. By way of introduction I will take an example from electoral politics, and then we will see how the techniques and concepts from this example extend to an analytic technique with regression analysis.

A warning: In this section and the one to follow on logistic regression I will follow the standard practice in statistics and will use the symbol p to refer to the "population" probability which we are trying to estimate from a sample probability π. This is directly analogous to the use of Greek and Roman letter variables for variance (σ and *s*), however, most of us think of π as a geometric term for the number of radians in a circle - 3.14.... Try to suppress this association and think of π from now on (in this book) as a probability.

5.2.1 Bush or Kerry?

As I write this in the closely contested state of Ohio, it is 2 days before election day 2004. So it is natural to use poll results as an example of how to test hypotheses about binomial data. Responses to a poll can be considered a "success" for Kerry when the respondent says he/she plans to vote for Kerry and a "failure" for Kerry for any other response. For instance, Cleveland, Ohio's largest newspaper, the Plain Dealer, asked 1500 likely voters in Ohio who they plan to vote for and found that 720 said "Bush", 675 said "Kerry" and the rest were either undecided or had determined to vote for other candidates. These poll results can be given in a contingency table (see table 5.2).

Table 5.2. Opinion poll results as "success" or "failure" of a particular candidate. The count of success and failure (in a sample of 1500 voters) is shown in the left table. In the center table these are converted to observed probabilities, and the right table shows the model parameters which these probabilities estimate.

	success	failure	success	failure	success	failure
Kerry	675	825	0.45	0.55	π_k	$(1-\pi_k)$
Bush	720	780	0.48	0.52	π_b	$(1-\pi_b)$

The expected value of the probability of a "Kerry" response in the survey sample is π_k - the proportion in the sample who say they will vote for Kerry. The standard error of this estimate is given by,

$$\sigma(p) = \sqrt{\frac{\pi(1-\pi)}{N}}$$ expected value of the standard error of the parameter π.

It may seem a little mysterious that you can calculate the standard error of a probability from the probability itself (if you want to see a derivation of this result see Hays, 1973). Figure 5.2 illustrates the standard error of a binomial variable for different sample sizes and proportions (p). You can imagine that if everybody said they would vote for Kerry (p=1.0) then there would be no variation at all in the data set. Similarly if everyone planned to not vote for Kerry (p=0.0) the standard error of the estimate would also be zero, and the situation where we have the greatest amount of variation is when the population is evenly split (p=0.5). This is what we see in figure 5.2. The figure also shows that larger samples have smaller standard error of the statistical estimate of π.

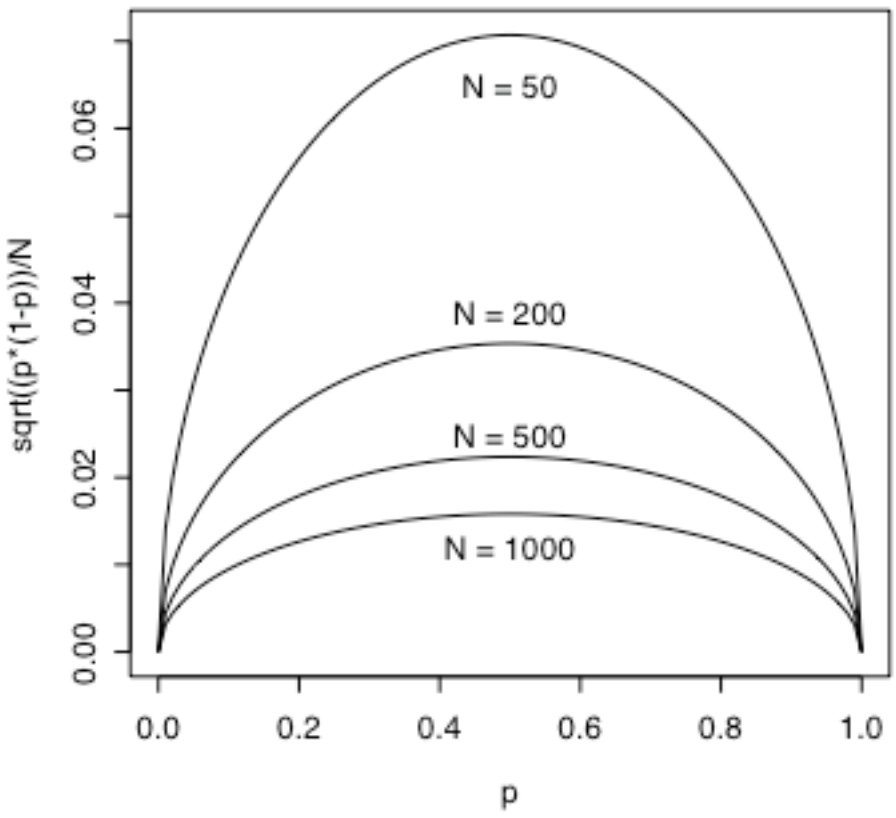

Figure 5.2. The standard error of p, for a samples of size 50, 200, 500, and 1000. Standard error of a probability decreases near 0 and 1, and is also smaller as sample size increases.

So, with a way of estimating the standard error of a proportion we can then test a hypothesis about the Plain Dealer poll. The hypothesis is that Bush and Kerry are tied in Ohio. That is: H_0: $\pi_k = \pi_b$. We'll test this null hypothesis by calculating a z score from the difference between the two probabilities.

$$z = \frac{p_b - p_k}{s(p_b - p_k)}$$

The standard error used in this statistic is related to the method used to calculate standard error for a single probability.

$$s(p_b - p_k) = \sqrt{\frac{p_b(1-p_b)}{n_b} + \frac{p_k(1-p_k)}{n_k}}$$

Entering 0.48 for p_b, 0.45 for p_k, and 720 for n_b and 675 for n_k we get a standard error of 0.027 (this is the "plus or minus 2.7%" that is reported with poll results), and this gives a *z* score of 1.123. This is not a very large *z* score, in fact if we took 100 polls 13 of them would have a larger *z* score even though Bush and Kerry are actually tied in Ohio. That is, this statistical test [z = 1.123, p=0.13] does not lead me to reject the null hypothesis that $\pi_k = \pi_b$.

I conducted the same analysis for a number of different polls, and pooled results of them, from Ohio, Florida, and Pennsylvania. These results are shown in table 5.3. The *z* score for the sum of polls in Ohio is clearly not significant (p=0.39) indicating that the poll results can't predict the winner in this state. The Florida sum of polls result shows Bush leading Kerry 48% to 47% and this difference is not significant (p=0.088). The difference between Bush and Kerry in the Pennsylvania sum of polls is marginally reliable (p<0.03).

What we saw on election night was an indication of the strengths and limitations of statistical inference. In all three states the candidate who was leading in the sum of polls ended up winning the state, though in two cases - Ohio and Florida - we couldn't have confidently predicted the outcome. However, we see the odd result that in Ohio more of the undecided or non-responsive people in the polls seem to have voted for Bush than for Kerry (about 65% vs. 35%). This discrepancy was even more pronounced in Florida (about 80% vs. 20%). The situation in Pennsylvania is more like we would expect given the nearly even split in the poll results - about 53% of the undecided or non-responsive voters ended up counting for Bush and 47% for Kerry. The discrepancy between the poll results and the election results indicates that there was either a bias toward Kerry in how the polls were taken, or a bias toward Bush in how the election was conducted.

Table 5.3 Poll results and signficance tests for three "battleground" states in the 2004 US presidential election.

	Counts			Proportions		H0: Bush=Kerry	
	Total	Bush	Kerry	Bush	Kerry	error	z
Ohio							
2004 results				0.51	0.49		
Sum of 7 polls	5279	2508	2487	0.475	0.471	0.014	0.28
Florida							
2004 results				0.52	0.47		
Sum of 10 polls	6332	3068	2958	0.48	0.47	0.013	1.35
Pennsylvania							
2004 results				0.49	0.51		
Sum of 9 polls	6880	3216	3377	0.467	0.49	0.012	-1.90

R note. Instead of looking up the probability of *z* scores like those in table 5.3, you can use the `pnorm()` function in R.

```
> pnorm(0.28,lower.tail=F)    # ohio pooled polls
[1] 0.3897388
> pnorm(1.35,lower.tail=F)    # florida pooled polls
[1] 0.088508
> pnorm(-1.9)                 # pennsylvania pooled polls
[1] 0.02871656
```

One thing that I learned from this is why pollsters use samples of 600 people. If you look at the standard error values for the different polls, it is apparent that the polls with larger samples have lower standard error, but not substantially so. For instance, to go from standard error of 4.2% to 1.2% the sample size had to increase from 600 to 6000. Probably, in most instances the extra effort and expense needed to interview 10 times as many people is not worth the extra accuracy.

The trade off between accuracy [$s(p_a\text{-}p_b)$] and sample size in a two alternative poll is shown in figure 5.3. As you can see in the figure, the pollster's preferred sample size, 600 is just below the inflection point in this curve where increased sample size begins to result in sharply diminishing improved accuracy. I thought that that was an interesting thing to learn about binomial distributions and error in estimating a population proportion from a sample proportion.

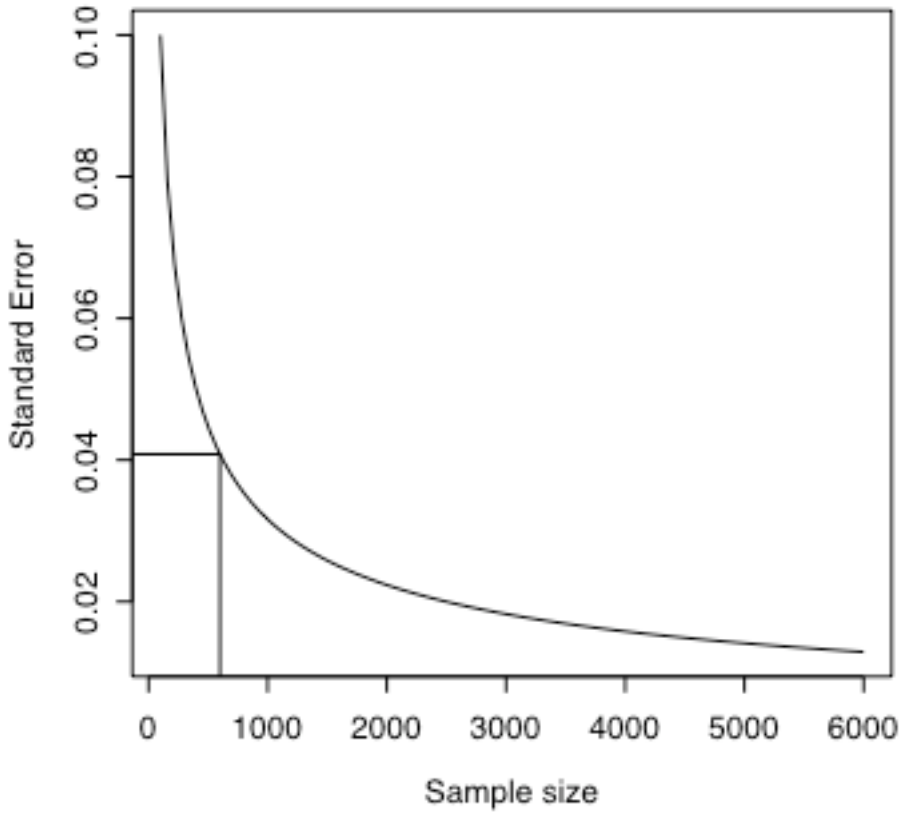

Figure 5.3. Standard error as a function of sample size for a two alternative poll with $\pi_a = \pi_b = 0.5$.

5.3 An aside about Maximum Likelihood Estimation.

Maximum likelihood (ML) estimation of model parameters is a key building block of the main statistical analysis technique used in modeling sociolinguistic data - logistic regression. So before we dive into logistic regression we'll briefly touch on ML estimation.

It is perhaps intuitive to understand that the best possible estimate of a population parameter π - the probability of an event - is the observed probability that is found in the data. Indeed, the expected value of p (the observed probability) is π (the population probability). However, when you deal with multifactor models using logistic regression, direct estimation of model parameters from data is more complicated.

We saw in chapter 4 that model parameters in analysis of variance and linear regression are estimated using the least squares (LS) criterion. The best model parameter is the one that produces the smallest sum of squared deviations from the observed data. This approach is very

powerful because exact solutions are possible. For instance, the mean is the least squares estimate of central tendency and there is only one way to calculate the mean, with no guess work. You calculate it and there you have it - the best fitting (least squares) estimate of central tendency.

The main limitations of LS estimates are that we must have homogeneous variance across conditions (remember the equal variance test for t-test, section 3.1.2?) and it must be reasonable to assume that the data fall in a normal distribution. Neither of these assumptions is true of count data. We usually have different numbers of observations in each cell of a table and maybe even some empty cells, so the variance is not homogeneous. Additionally, dichotomous, "success/failure" responses do not fall on a normal distribution.

Maximum likelihood to the rescue. The idea with maximum likelihood estimation is that we determine a likelihood function for the parameter values being estimated and then find the peak in the likelihood function. For example, the likelihood function for estimating π from a binomial probability distribution with y successes and N total observations is:

$$\lambda = \binom{N}{y} \pi^{y} (1-\pi)^{N-y} \qquad \text{binomial probability density function}$$

In this function, the notation $\binom{N}{y}$ is the number of sets of size y from a list of N elements - the "choose" y from N. This can be calculated as the ratio of factorials $\binom{N}{y} = \frac{N!}{y!(N-y)!}$, where $N!$ is equal to 1*2*...*N. You can use this binomial probability function in a number of ways, for example to rate the probability of finding 20 heads out of 30 tosses of a true ($\pi = 0.5$) coin (2.8% of samples).

R note. The binomial distribution is handled by a family of functions in R, just as the normal distribution, the *t* distribution, *F*, χ^2 and others. For example, to examine the coin toss example I used `dbinom()` to calcuate the probability of throwing 20 heads in 30 tosses of a fair coin. You can also calculate this directly utilizing the `choose()` function in the binomial probability formula.

```
> dbinom(20,30,.5)
[1] 0.0279816

> choose(30,20) * .5^20 * (1-.5)^(30-20)
[1] 0.0279816
```

Yet another way to get the same answer is to subtract the probability of getting 19 or fewer heads from the probability of getting 20 or fewer heads using the `pbinom()` function.

```
> pbinom(20,30,0.5)-pbinom(19,30,0.5)
[1] 0.0279816
---------------------------
```

In maximum likelihood estimation we know the values of our sample observations (the N and y) and we would like to find the most likely value of π - the one that produces the maximum value of λ. But, there is no direct way to calculate the maximum of the likelihood function, so instead the peak must be found using an iterative search. This is illustrated in figure 5.4, where the likelihood function (from the binomial probability density function, above) is shown for a sample that has 5 successes in 50 trials. The horizontal axis shows different values of π while the vertical axis shows the resulting likelihood values λ. The aim of maximum likelihood estimation is to find the value of π that has the highest likelihood. The vertical line drawn at $\pi = 0.1$ is the peak of this likelihood function, the point that is chosen as the best estimate of the population parameter π. In this case, it is simply the probability of a success (5/50 = 0.1) in the sample. As we will see in the next section, this method can also be applied to find the model parameters of complicated regression models (of course I haven't actually said anything about the gradient ascent peak finding methods used in the search for the maximum likelihood, and I won't either).

A final note about maximum likelihood estimation. This method of finding the parameters of a model is not limited to logistic regression. You can also use ML estimation to find the coefficients of a regression equation or the effects of an analysis of variance. The only change is that the method of fitting the statistical model to the data is the maximum likelihood strategy rather than the least squares method. This parallelism is reflected in the R statistical language. The `lm()` function fits models using least squares estimation, while `glm()` uses maximum likelihood.

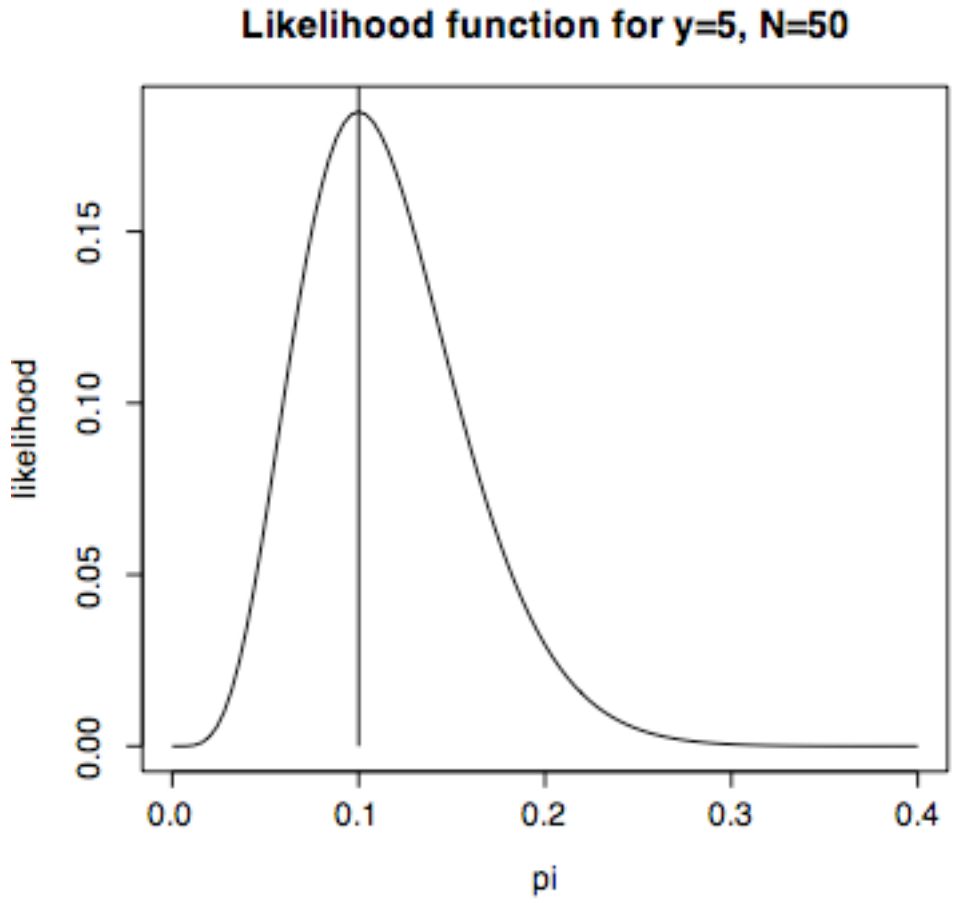

Figure 5.4. The likelihood function used to estimate π, for a sample of 50 trials in which 5 were "successes". The vertical axis is the likelihood of π [$\lambda(\pi)$], and the horizontal axis represents different values of φ that could be considered in the search for the π with the maximum $\lambda(\pi)$. The maximum likelihood estimate is 0.1 = 5/50.

R note. Figure 5.4 is produced by this R command. Note that you could use `factorial()` in this plot statement to spell out the choose function, however the factorial function is undefined in R beyond 170! which is a huge number (7.257416e+306).

```
curve(choose(50,5)*x^5*(1-x)^(50-5),0,0.4,ylab="likelihood",
xlab="p i",main="Likelihood function for y=5, N=50")
```

5.4 Logistic regression.

Maximum likelihood is well-suited to the type of data we usually have in sociolinguistics because it is a method that is nonparametric - it doesn't require equal variance in the cells of a model, and doesn't require that the data be normally distributed.

One other aspect of sociolinguistic data was emphasized in sections 5.1 and 5.2. The data are usually counts of "applications" of some process, so we are dealing with probabilities. You might recall from chapter 1 that we used an "s" shaped tranformation - the arcsine transform - to deal with the reduced variance of probabilities near 0 and 1. In logistic regression we use the logit function for the same purpose (see Cedergren & Sankoff, 1974; Sankoff, 1978, 1988 on the use of logistic regression in sociolinguistics). The logit function is preferable to arcsine because the resulting value has a sensible interpretation as the log value of the odds of application of the process.

$\pi(x)$ the proportion of "applications"

$\frac{\pi(x)}{1-\pi(x)}$ the odds of an application

$\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$ the log odds of an application, the logit

The relationship between probability, odds and logit is shown in table 5.4. There we see that an event that occurs in 80% of the observed cases has an odds ratio of 4 to 1 (the event is 4 times more likely to occur than the non-event. The logit value of 1.386 is not intuitively meaningful if you don't normally deal in log odds (who does?) but it is easy to map logits into odds or probabilities, and the values are centered symmetrically around zero.

Table 5.4. Comparison of probability, odds and log odds for a range of probabilities.

probability	odds	log odds
0.1	0.111	-2.197
0.2	0.25	-1.386
0.3	0.428	-0.847
0.4	0.667	-0.405
0.5	1.00	0.00
0.6	1.5	0.405
0.7	2.33	0.847
0.8	4.00	1.386
0.9	9.00	2.197

The logit function is shown in figure 5.5 as a function of probability. In interpreting the results of a logistic regression some analysts (particularly in medical research) refer directly to the odds or the log odds of of an event, such as recovery from a disease. I find it more understandable to translate the coefficients that are found in logistic regression back into probabilities via the inverse logit function.

$$y = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$ the logit, or logistic, function

$$\pi(x) = \frac{e^y}{1+e^y}$$ the inverse logit (to calculate probabilities from logits)

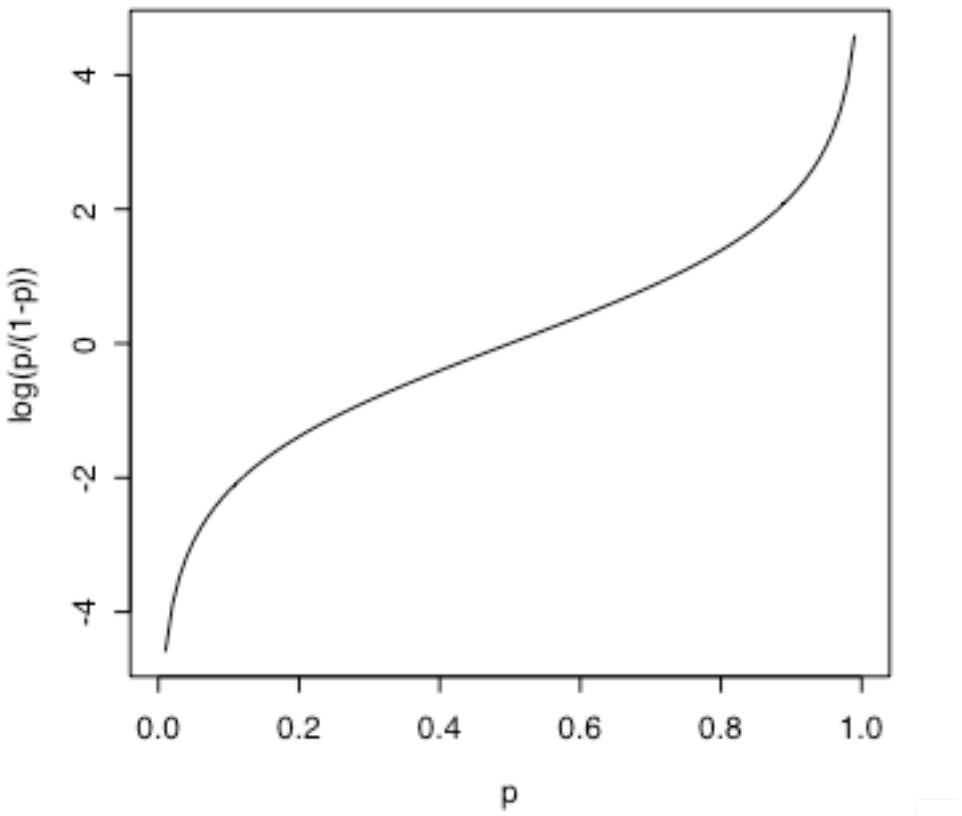

Figure 5.5. The logit transform y = log(p/(1-p). The inverse of this is p = exp(y)/(1+exp(y)).

5.5 An example from the [ʃ]treets of Columbus.

Now, let's turn to an example of a logistic regression analyis of some sociolinguistic data. These data are found in the file "DDRASSTR.txt" and were contributed by David Durian. Durian describes his study this way:

> Data from 120 Caucasian, native-English speaking Columbus, OH store clerks (40 store clerks working in a variety of stores at each of 3 Columbus malls [Easton, Polaris, and City Center]) were elicited via the Rapid Anonymous Survey technique. The entire set of store clerks were also stratified by the following social factors: age (15-30, 35-50, 55-70, 40 speakers in each age group); social class (working class- WC, lower middle class - LMC, and upper middle class - UMC, 40 speakers of each class); and gender (60 males/60 females).

> From each of the 120 speakers, two tokens of words containing word-initial "str" clusters were obtained. The first, in a less emphatic speech environment; the second in a more emphatic one. This leads to a total of 240 tokens (120 more emphatic; 120 less emphatic). The two variant realizations of "str" were rated impressionistically by the researcher as closer to [str] or [ʃtr].
>
> All of the data were elicited by the researcher asking for directions to a known street in the area of the mall that was misidentified as a "road." The speaker then corrected the researcher by saying "you mean X street," and the first "str" token was obtained. Following this first utterance, the researcher said, "excuse me, what did you say" and the speaker would more emphatically repeat the phase "x street" producing the second "str" token. In the case of Polaris Mall no widely identified street is located within proximity to the mall (everything there is a Road, Parkway, Place, etc), and so the researcher asked for directions to a part of the store that would yield the word "straight" in place of "street."
>
> Data were written down on a sheet of paper just after leaving eyesight of the informant. No audio recordings were made. The only judgment made on the sounds were the researher's own impressionistic rating.

I edited Durian's description to simplify the description of how he assigned talkers to different class levels. After collecting the data from a number of different stores, in the initial design with 40 talkers from each of three presumed classes, he had students at Ohio State University estimate the average class background of the shoppers in the stores. These judgements were used to estimate the clerks' class for the logistic analysis.

5.5.1 On the relationship between χ^2 and G^2

Logistic regression results in a statistic called G^2 which is the log likelihood analog of χ^2. So, in this section we'll demonstrate the similarity between χ^2 and G^2 and talk about where this G^2 statistic comes from.

Table 5.5 shows the number of [ʃtr] and [str] productions in Durian's study as a function of whether the utterance was produced without emphasis or in a relatively more emphatic utterance. It looks as if people were much more likely to say [ʃtr] in the less emphatic case (35% versus 12%). The value of χ^2 for this table, found using the R statement `summary(table(str,emphatic))` is 19.35, which is signficantly greater than chance for a χ^2 with one degree of freedom.

Table 5.5. Number [ʃtr] and [str] productions as a function of emphasis.

	emphasis	
	less	more
ʃtr	43	14
str	77	106

Now the logistic regression analysis of these same data is done using the general linear model - `glm()` - which fits model parameters using maximum likelihood estimation instead of least squares. Results of the glm model fit can be printed as an analysis of deviance table which is analogous to the analysis of variance table. This analysis of deviance table for the [ʃtr] by emphasis data is shown in table 5.6. The G^2 value (labeled "Deviance") is 20.08 - almost the same as the χ^2 value.

Table 5.6. Analysis of Deviance Table for a logistic regression analysis of the data shown in table 5.5.

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			239	263.13	
emphatic	1	20.08	238	243.04	7.416e-06

Although, in the paragraphs to follow we will discuss how G^2 is derived from likelihood ratios in model fitting, it may be interesting to some readers to note that like χ^2, G^2 can be calculated directly from the observed and expected frequencies of a contengency table. Using the same counts of observed and expected frequency that are used in calculating χ^2,

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

G^2 is calculated from the natural log of the ratio of observed and expected frequency. One of the exercises at the end of the chapter has you derive 19.35 as the χ^2 and 20.08 as the G^2 of the data in table 5.5.

$$G^2 = -2\sum_i o_i \log\left(\frac{o_i}{e_i}\right)$$

R note. The command that produced the analysis of deviance table shown in figure 5.6 was:

```
> dd<-read.delim("DDRASSTR.txt")
> attach(dd)
> anova(glm(str~emphatic,family=binomial,data=dd),test="Chisq")
```

Parts of this command are familiar from previous chapters. For example, we specify a model `str~emphatic` such that the response variable, whether the person said [str] or [ʃtr], is a function of whether the utterance was emphatic or not. This model is evaluated by the `glm()` function much as we used the `lm()` function to produce analysis of variance tables in chapter 4. Because the response measure is binomial rather than a measurement on a continuous scale, we specify that the model should be evaluated with `family=binomial`. This results in the use of the logistic scores. Finally, `anova()` requests that the results be printed as an analysis of deviance table and `test="Chisq"` specifies that the deviance score should be evaluated on the χ^2 distribution.

In logistic regression, since we are using maximum likelihood parameter estimation to find the best fitting model of the data, the measure that is used to test the degree of fit of a model is the likelihood ratio. The idea here is to compare the maximum likelihood found in a model that includes a possible explanatory factor (like emphasis) with the maximum likelihood of a model that does not include that explanatory factor. In the NULL model, which includes no explanatory factors at all, the only basis for making a prediction is the average probability of [ʃtr] productions. This serves as a baseline and then we add a possible explanatory factor and measure how much the likelihood of the model improves.

To compare models with and without a factor we take the likelihood ratio. The idea here is analogous to the use of ratio measures in all of the other statistical hypothesis tests that we've discussed so far (*t* test, *F* test, *z* score). If the ratio is close to unity (1) then the improved fit offered by the added factor is insubstantial and considered a nonsignificant predictor of the criterion variable.

Likelihood ratios fall on a different distribution than variance ratios, so we use a different statistic called G^2 to test their significance. G^2 has the same distribution as χ^2, so in looking up significance of G^2 you use the χ^2 tables, but because it is calculated a different way we give it a different name. G^2 is called Deviance in table 5.6 and can be calculated from the sum of deviance scores ($y_i - \hat{y}_i$) when the data are normally distributed an the true σ is known.

$$G^2 = -2\log\left(\frac{l_m}{l_{m-1}}\right)$$ comparing the likelihoods of two models.

The likelihood ratio in this formula compares two models, *m* and *m-1*, where *m* has one additional predictive factor that was not included in *m-1*. This value in table 5.6 was 20.08. In practice it is convenient to calculate G^2 from two other G^2 (deviance) scores. The highest possible likelihood value can be obtained when there is a parameter for every observation in the data set. This one-paremeter-per-data-value model is called the saturated model. Any other model can be compared to the saturated model with a G^2 value to measure the improvement in fit. This gives a "residual Deviance" that indicates how much better the saturated model predicts the data compared to a smaller (perhaps more explanatory) model.

$$G^2(m) = -2\log\left(\frac{l_m}{l_s}\right) = -2\left[\log(l_m) - \log(l_s)\right] = -2\left[L_m - L_s\right] \qquad \text{Deviance of model } m.$$

Note that in this statement we take advantage of the fact that the log of a ratio is equal to the difference of the log values. So to calculate the log of the ratio we can take the difference of the log likelihoods. Now to calculate G^2 comparing model m and model m-1 we can simply take the difference between deviance scores.

$$\begin{aligned} G^2(m \mid m-1) &= -2\left[L_m - L_{m-1}\right] = -2\left[L_m - L_s\right] - -2\left[L_{m-1} - L_s\right] \\ &= G^2(m) - G^2(m-1) \end{aligned}$$

This is distributed on the χ^2 distribution with degrees of freedom equal to the difference between the residual degrees of freedom for the two models (the number of coefficients added to the model in going from *m-1* to *m*.

This has been a fairly long explanation of the central observation of this section. The G^2 that we get in logistic regression analysis of deviance is simply a different way to measure the same thing that we measured with χ^2. In the case of our example from the [ʃ]treets of Columbus, Ohio the question is: "does emphasis play a role in the production of [str]"? With a Pearson's $\chi 2$ analysis of the tabulated data we found a significant χ^2 value of 19.35 (with one degree of freedom). With logistic regression we found a significant G^2 value of 20.08 (again with one degree of freedom). This is supposed to illustrate that these two analyses are two ways of testing the same hypothesis.

Logistic regression has several advantages some of which we will explore in the remaining sections of this chapter.

5.5.2 More than one predictor.

One of the main advantages of logistic regression over Pearson's χ^2 is that we can fit complicated models to the data using logistic regression. As an example of this we will consider a set of four predictor variables and their interactions in the Durian [ʃtr] data. Recall that in addition to recording productions in emphatic and non-emphatic context, he collected data from an equal number of men and women, from people in three different age ranges, and classified his talkers according to the economic/social class of their customers. With logistic regression, as with analysis of variance, we can test all of these factors at once.

I want to emphasize that this analysis is possible, not only because logistic regression is a great analysis tool, but also because Durian collected enough data to provide a relatively balanced model with about 10 observations in each cell of the model. This involved collecting data from 120 people. Less than this and he would have had to give up an experimental variable, like age or class, for lack of statistical power to examine the interactions among the factors. It is often necessary in sociolinguistics to study many fewer people because ethnographically careful data collection requires a substantial investment of the investigator's time for each subject in a research study. The trade off that we have to keep in mind in these cases is that there may not be enough data to permit investigation of more than one or two research variables, and particularly not the interaction of variables. This is too bad because interactions are often much more informative than main effects.

The interaction that we will be exploring in the [ʃtr] data is between age and gender. The conclusion that young women are much more likely to use [ʃtr] than either young men or older women, i.e. that young women are leading a sound change in Columbus, Ohio. We'll look at two strategies for approaching this analysis.

The first analysis is exactly like the one shown above in table 5.6, except here we specified a full model which includes fifteen predictive factors - four main effects (age, gender, emphasis, and class) plus all of the possible interactions of these effects. As the bold print in table 5.7 indicates, all four of the main effects were found to have a significant effect on the frequency of [ʃtr] production, while only one interaction seems to matter. That interaction is between age and gender.

Table 5.7. Analysis of Deviance table for a full model of the [ʃtr] data. Statistically reliable main effects and interactions are printed in bold face.

Model: binomial, link: logit

Response: str

Terms added sequentially (first to last)

		Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
	NULL			239	263.127	
Main effects	**age**	2	17.371	237	245.756	1.690e-04
	gender	1	9.019	236	236.737	0.003
	emphatic	1	22.537	235	214.200	2.062e-06
	class	2	21.481	233	192.719	2.165e-05
2-way	**age:gender**	2	8.187	231	184.532	0.017
	age:emphatic	2	0.151	229	184.381	0.927
	age:class	4	7.656	225	176.725	0.105
	gender:emphatic	1	1.463	224	175.262	0.226
	gender:class	2	3.542	222	171.719	0.170
	emphatic:class	2	1.528	220	170.191	0.466
3-way	age:gender:emphatic	2	4.960	218	165.231	0.084
	age:gender:class	4	5.356	214	159.876	0.253
	age:emphatic:class	4	3.333	210	156.543	0.504
	gender:emphatic:class	2	0.001	208	156.542	1.000
4-way	age:gender:emphatic:class	4	0.001	204	156.541	1.000

Notice in table 5.7 the statement at the top that the terms were added sequentially. What this means is that unlike analysis of variance, in logistic regression the order in which the factors are mentioned in our model statement (when invoking `glm()`) has an impact on the statistical test of the factors. In particular, if two predictive factors T1 and T2 are highly correlated with each other, if we enter T1 first then T2 will probably not show up as significant, because to test the effect of T2 the regression tests whether adding T2 to a model that already includes T1 is an improvement over the model with T1. When T2 is correlated with T1 then it provides very little improvement over a model that includes T1, and the nature of the model fitting procedure never tests T2 alone without T1 already in the model. The situation is reversed if T2 is mentioned first in the list of predictor variables.

Notice that interactions are always added into the analysis after the main effects or smaller interactions (in both the full model and in the step-wise procedure below). This is entirely appropriate because we want to test if there is any variance to account for by the interaction

after removing variance due to the higher-level effects.

But to determine whether it would be better to use T1 or T2 to predict our data we might want to use a step-wise procedure that will test them independently in order to determine their best order in the model statement. We saw stepwise regression procedures earlier in chapter 3 and here will use the step() function with logistic regression just as with did with linear regression.

As Table 5.8 shows, the stepwise procedure selected the main effects in a different order than I had entered them for the full model analysis in table 5.7, but the results are unchanged - all four main effects, as well as the age by gender interaction, are selected as significant.

Table 5.8. Analysis of Deviance table for the step-wise analysis of the [ʃtr] data.

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			239	263.127	
emphatic	1	20.083	238	243.044	7.416e-06
class	2	23.583	236	219.462	7.570e-06
age	2	16.153	234	203.308	3.107e-04
gender	1	10.589	233	192.719	0.001
age:gender	2	8.187	231	184.532	0.017

We've seen the effect of emphasis - namely that the rate of [ʃtr] is 35% in less emphatic context and drops to 12% in more careful speech. Class had a similarly dramatic influence on [ʃtr] rates with 37% for working class, 27% for lower middle class and only 8% for upper working class averaged across levels of all other factors. By the way, averaging across all other factors is justified by the lack of any interactions with other factors in the logistic regression. Younger speakers had a greater proportion of [ʃtr] than did older speakers (34% versus 9%). Similarly, women were more likely to use [ʃtr] than were men (32% versus 16%). However, the age by gender interaction indicates that it is misleading to average over gender when we evaluate the age effect, or to average over age when we evaluate the gender effect. Figure 5.6, which shows the age by gender interaction, illustrates why this is so. Men had a relatively constant rate of [ʃtr] production regardless of age, while the young and middle age women had much higher rates of [ʃtr] than did the older women. Thus, the age effect is confined primarily to women and the gender effect is confined primarily to the young and mid aged speakers.

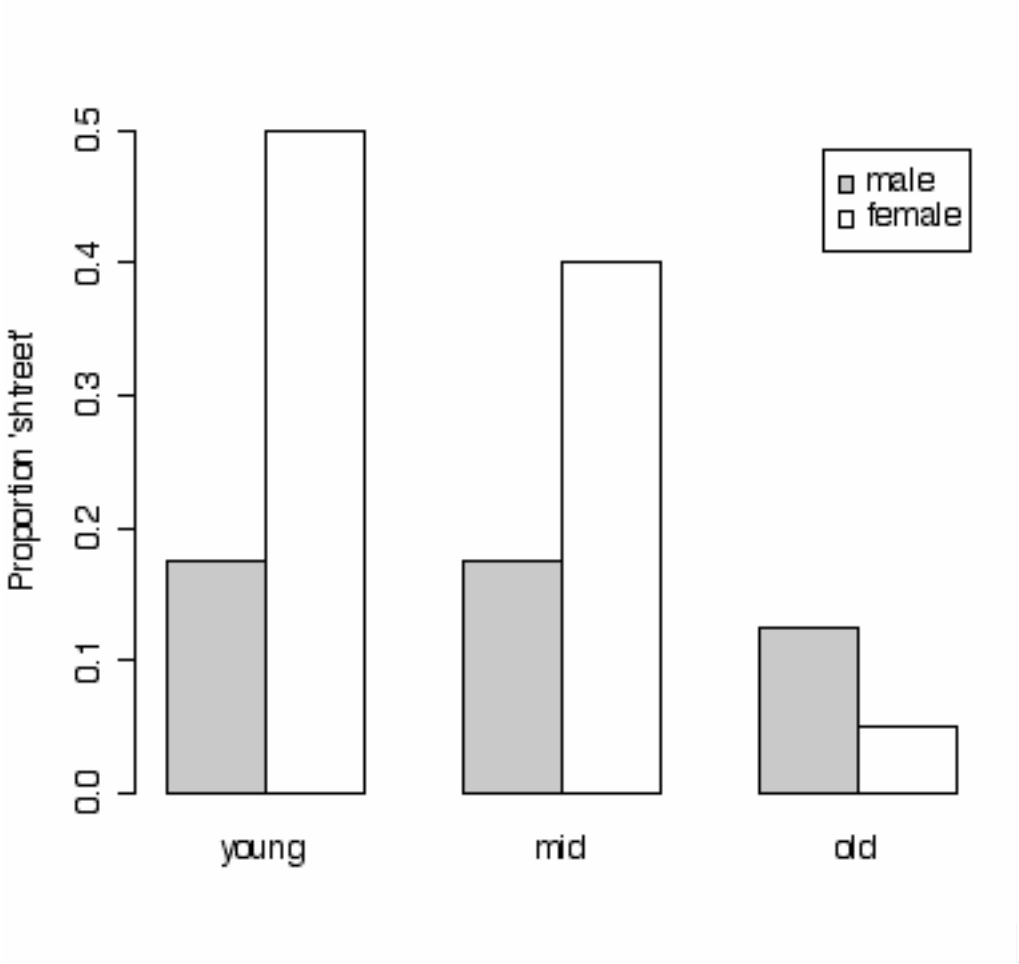

Figure 5.6. The age by gender interaction in Durian's [ʃtr] study. Data for male talkers is shaded gray, while the bars showing data for female talkers are white.

R note. The first analysis was called the "full" analysis, because I included all four factors and all of their their interactions. The order in which the factors are listed in the model statement - here with age first and class last - matters in the statistical test.

```
anova(glm(str~age*gender*emphatic*class, family=binomial, data=dd), test =
"Chisq")
```

The stepwise logistic regression was performed with this statement:

```
dd.glm <- step(glm(str~1, family=binomial, data=dd), str ~ ag e * gender *
class * emphatic)
```

Table 5.9 can then be produded by `anova(dd.glm,test="Chisq")`. The stepwise procedure gives you a blow-by-blow print out as it builds up the model from the NULL model to the model

that contains only those factors that are determined to be statistically reliable.

Figure 5.6 was produced with the `barplot()` command, which requires that the input be a matrix. I constructed the matrix from a vector of the proportion [ʃtr] responses.

```
v <- c(7,7,5,20,16,2)/40
m <- matrix(v,nrow=2,byrow=T)
barplot(m,beside=T,names.arg=c("young","mid","old"),
legend=c("male","female"),col=c("gray","white"), ylab="Proportion 'shtreet'")
---------------
```

5.6 Logistic regression as regression: An ordinal effect - age.

Just briefly here it makes sense to consider how an ordinal effect can be treated in a regression model - this applies both to the least squares models that we discussed in previous chapters and to the maximum likelihood models that we are considering in this chapter. For instance, in Dodsworth's /l/ vocalization data set (section 5.1) we have a factor "age" that takes one of four levels - teens, twenties, fourties, or fifties. The levels of this factor are obviously ordered from youngest to oldest, so it would make sense to treat age as an ordinal variable. This is done by using "polynomial" coding to convert the levels of the age factor into a numeric code. The three variables used to encode ordered trends as age relates to /l/ vocalization. For example, if /l/ vocalization occurs more frequently in the speech of older talkers than in younger talkers, we would expect the linear encoding of age to be significant, meaning that as age increases /l/ vocalization increases just as the linear coefficient values increase linearly from -0.67 to 0.67. The encoding scheme for age from table 5.9 is shown also in figure 5.7.

Table 5.9 Polynomical coding of an ordinal factor.

	linear	quadratic	cubic
teens	-0.67	0.5	-0.22
twenties	-0.22	-0.5	0.67
fourties	0.22	-0.5	-0.67
fifties	0.67	0.5	0.22

Just as with other coding schemes (treatment coding and effects coding) there are three variables in the regression formula for this four-level factor - one variable less than the number of levels. However in this case the first variable encodes a linear increase over the four levels, the second factor encodes the possibility of a dip or bulge in the middle of the age range, while the third factor encodes the more complicated possibility that /l/ vocalization "zigzags" among the age levels.

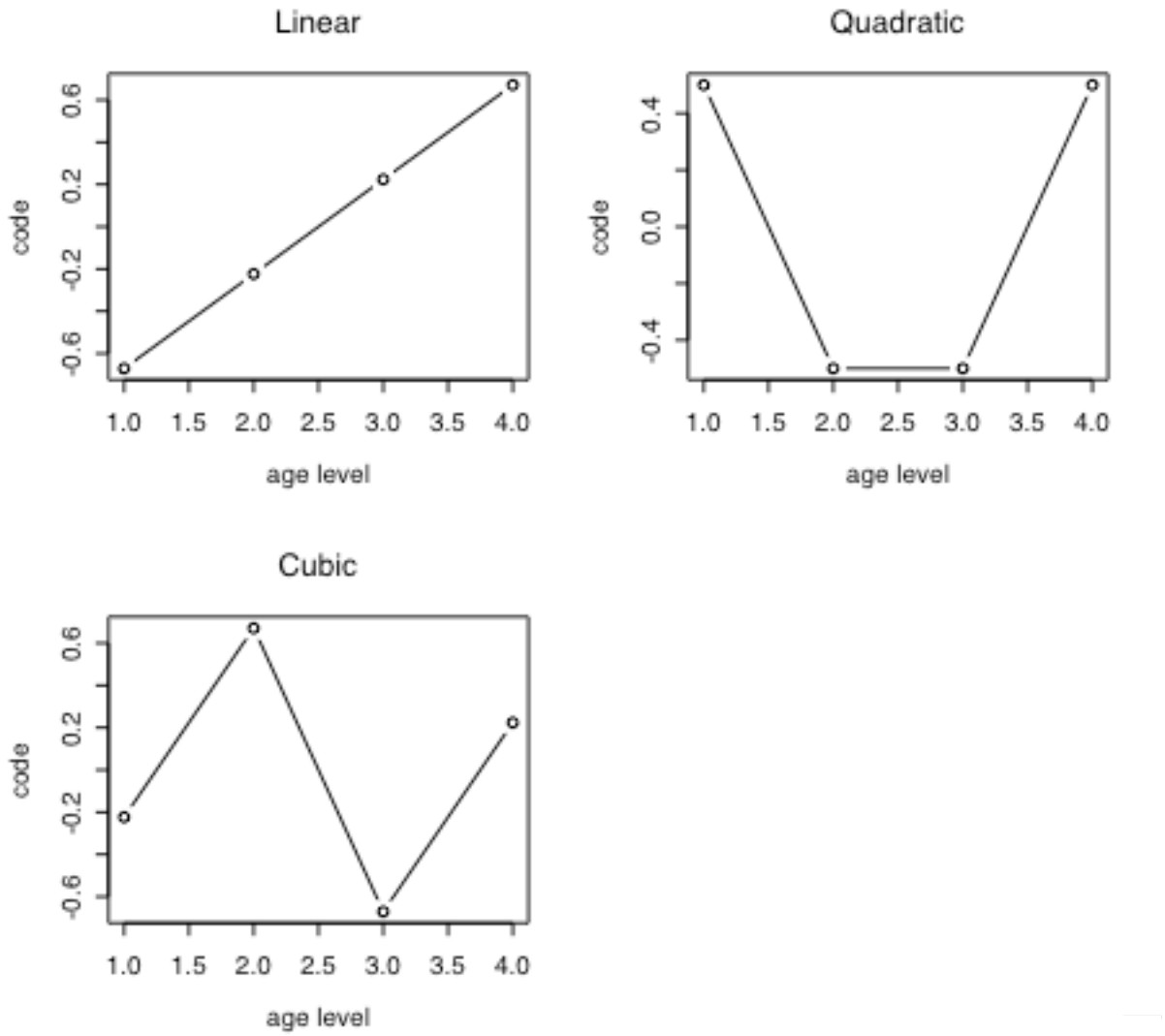

Figure 5.7. The polynomial encoding scheme for the different levels of "age" - "teens" (1.0 on the x axis), "twenties" (2.0 on the x axis), etc. The linear, quadratic and cubic encoding schemes are shown.

Now, when we conduct the logistic regression with age as an ordinal factor, the regression coefficient (Table 5.10) for the cubic ordered pattern `age.C` is larger than either the linear or the quadratic coefficients, and the cubic coefficient is also reliably different from zero. This means that the pattern of /l/ vocalization as a function of the age of the talker was more like the "zig zag" pattern shown by the cubic encoding scheme shown in figure 5.6. That this is in fact the pattern of of /l/ vocalization is shown in the graph of proportion of /l/ vocalization shown in Figure 5.8.

Table 5.10. Regression coefficients from a logistic regression analysis of /l/

vocalization. This analysis focussed on the ordinal effect of age.

```
Coefficients:
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)  0.25776    0.09905      2.602     0.00926**
age.L        -0.27771   0.17994      -1.543    0.12275
age.Q        -0.12110   0.19810      -0.611    0.54099
age.C        0.86705    0.21473      4.038     5.39e-05***
```

So, it is possible, by treating an ordinal factor as ordinal to determine whether the relationship between that factor and the predicted variable is basically linear, or if there are quadratic, or cubic (or higher powers, if more levels are considered). The overall independence of /l/ vocalization from age is tested in exactly the same way whether we treat age as ordinal or as nominal, and the success of the regression model does not depend on this - effects coding of age, and ordinal coding of age captures exactly the same amount of variation. However, being able to specify that a factor is ordinal lets us determine the type of trend that relates this ordered factor to the response variable. I guess in most situations you would hope for something pretty simple like a linear or quadratic trend. This cubic trend requires a special kind of theory to explain it.

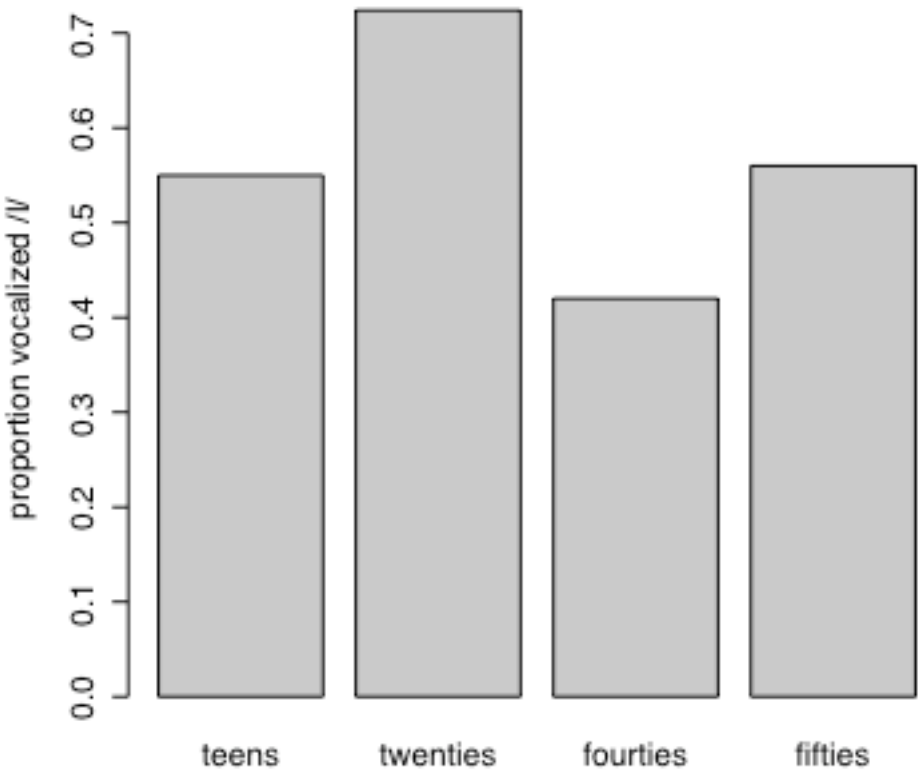

Figure 5.8. Proportion of /l/ vocalization for each of four different age groups. The pattern here is most like the "cubic" ordinal effect.

R note. I mentioned earlier that I recoded Dodsworth's data a little for this chapter. To encode the fact that age is an ordinal variable I added "`ordered = T`" to the `factor()` statement that I had earlier used to recode "age" into "newage". Additionally, you can use `ordered()` to indicate that an existing factor should be treated as an ordered factor - this using the `levels` option to list the factor levels in order.

```
rd$newage <- factor(rd$age,levels=c(1,2,4,5),   labels= c("teens", "twenties",
"fourties", "fifties"), ordered = T)

rd$newage <- ordered(rd$newage, levels = c("teens", "twenties", "fourties",
"fifties"))
```

Table 5.9 is the output from `contrasts(rd$newage)`, and table 5.8 was produced by the following command.

```
> summary(glm(lvoc ~ newage, data=rd, family=binomial))
----------------------
```

For example, when we consider whether age influences /l/ vocalization the analysis of deviance table from the logistic regression analysis shows that the NULL model has a deviance, or G^2, value of 719.03. This is $G^2(m\text{-}1)$. When we add age as a predictor variable the deviance of this new model m is 696.47. This is $G^2(m)$. The difference between these two gives the G^2 value that we use in testing the hypothesis that age influences /l/ vocalization. This difference value $G^2(m|m\text{-}1)$ is 22.56, which on the χ^2 distribution with 3 degrees of freedom is a large enough value that is unlikely to have occured by chance. we conclude that age is an important predictor of /l/ vocalization. One thing to recall is that the χ^2 value testing the independence of /l/ vocalization and age was also 22. In other words these are two ways of calculating the same thing.

```
Analysis of Deviance Table
          Df Deviance Resid. Df Resid. Dev
NULL                      530     719.03
newage    3    22.56       527     696.47
```

This illustrates the similarity between G^2 and χ^2 in the case where a single factor is being tested. Now if you think about it a bit you may realize that the order in which you add factors to a model will have an impact on the G^2 value that you find for each factor in the model. This is because G2 for a particular factor is not being tested against an overall residual error variance but against the likelihood of a model that differs only by not having the factor in the model.

Consider for example a two factor model that has age and conscious as factors for /l/ vocalization. When we add age first the deviance is as we found in the one factor model, but when we add age second the G2 value drops to 13.68. This is still significant but highlights the importance of the order in which the factors are added to a model.

```
Analysis of Deviance Table
             Df Deviance Resid. Df Resid. Dev
NULL                         530     719.03
newage       3    22.56       527     696.47
conscious    3     9.41       524     687.06

Analysis of Deviance Table
             Df Deviance Resid. Df Resid. Dev
```

```
NULL                          530     719.03
conscious   3    18.29        527     700.74
newage      3    13.68        524     687.06
```

There are a couple of ways that you could choose to order factors in a model. Cohen and Cohen (1986) recommend adding prior factors first. That would mean putting factors that describe unchanging aspects of the situation before adding factors that describe aspects that might change from observation to observation. So factors that describe people come first and factors that describe the words they say come last. However, in this case both age and social consciousness describe the people under study.

A second approach is to add factors according to their "importance", in a stepwise model selection procedure. We've seen this before in Chapter 3.

5.5 Varbrul/R comparison

Varbrul is an implementation of logistic regression that is used by many sociolinguists (Cedergren & Sankoff, 1974; Sankoff, 1978, 1988). This implementation of logistic regression has been very important in the history of sociolinguistics because it conveniently made logistic regression available to researchers before the standard statistical packages included logistic regression. At this point Varbrul is a bit of a "legacy" program because most major statistical packages do now provide logistic regression. There are several reasons to use a general purpose software package rather than a specialized implementation like Varbrul. For instance, data handling, graphics, and model specification are additionally supplied in the general purpose package, as are other data analysis techniques (such as repeated measures logistic regression, which will be discussed in chapter 7).

One hurdle though, in using a general purpose statistics package is that the analytic procedures (such as stepwise regression) are quite flexible, presenting a wide range of possible analysis strategies. This can be confusing at times. Unfortunately I will not be able to explore the range of analysis strategies that one might employ using the R glm() function. We have seen that step() implements a stepwise regression strategy that is familiar to Varbrul users. We will also see a training/testing strategy in chapter 7, which should be seriously considered for sociolinguistic data analysis.

Another hurdle we face with a general purpose software package is that one doesn't know which of the sometimes many values reported in summary and print statements should be used. In this section we will compare a Varbrul analysis of the Durian str/shtr data to a logistic regression in R. My aim here is to show the interested reader how to compute the familiar Varbrul results table.

David Durian provides a printout of a Varbrul analysis of his data which is shown here as Table 5.11.

Table 5.11. Results from a Varbrul analysis of the Durian str/shtr data.

Log Likelihood -86.816				
Significance .001				
Input 0.102				
Chi-square/Cell .7579				
Group	Factor	Weight	App/Total	Input & Weight
1: Gender	M	0.334	0.016	0.05
	W	0.666	0.032	0.18
2: Environment	Less Emphatic	0.731	0.36	0.24
	More Emphatic	0.269	0.12	0.04
3: Age	15-55*	0.663	0.31	0.18
	55-70	0.206	0.09	0.03
4: Region	Polaris/Easton*	0.562	0.27	0.13
	City Center	0.376	0.16	0.06
5: Social Class	MWC-UWC	0.863	0.47	0.42
	LWC-MWC	0.711	0.26	0.22
	UWC-LMC	0.740	0.37	0.24
	LMC-MMC	0.584	0.22	0.14
	UMC-UC	0.061	0.02	0.01

R note. In conducting the analyses described in this section I used a couple of helpful functions that are described in more detail in this note.

We start the R analysis then with the R commands to read in the data and produce a logistic regression analysis of these data.

```
> dd<-read.delim("DDRASSTR.txt")
> dd.glm <- glm(str~emphatic+class,data=dd,family=binomial)
```

In preliminary analyses I noticed that because of the order of the levels "str" and "shtr" in the description of the data, my analyses were treating "str" as a "success" or an "application" of the process. I wanted to have the opposite be true, that "shtr" be the focus of the investigation. So I used relevel() to stipulate that "str" is the default value of response.

```
> dd$response <- relevel(dd$response,"str")
```

I also found, in presenting the results in this section that it was useful to be able to know how the different factor coding schemes work. For this, the `contrasts()` is invaluable. The output from `contrasts()` shows the mapping between nominal variables, listed in rows, and the numerical codes used in the regression. For example, the columns below correspond to classUMC and classWC in the regression formula.

```
> contrasts(class)
    UMC WC
LMC   0  0
UMC   1  0
WC    0  1
```

Similarly, when the coding scheme is `contr.sum`, `contrasts()` shows how the three levels of "class" will be coded. The first column shows the coding scheme for the class1 variable in the regression formula, and the second column shows the coding for the class2 variable.

```
> contrasts(class)
     [,1] [,2]
LMC    1    0
UMC    0    1
WC    -1   -1
```

There are two different ways to specify that you want "sum" coding instead of "treatment" coding. One is to add a list of contrasts to the glm command. In the list you name the contrast coding scheme for each variable.

```
> dd.glm <- glm(str~emphatic+class, data=dd, family=binomial,
      contrasts=list(emphatic="contr.sum", class="contr.sum"))
```

An alternative method is to change the global options used in R. There are two default contrasts specifications in R. The first indicates how nominal variables will be treated and the second indicates how ordinal variables will be treated. So, using options() we can specify that we want all nominal variables to be coded using the "contr.sum" coding scheme.

```
> options(contrasts = c("contr.sum","contr.poly"))
```

Finally, to find the inverse of the logit function (to translate from coefficients in the logistic model to predicted probabilities) you can use the `inv.logit()` function. Naturally enough there is also a `logit()` function to calculate the log odds of a probability.

```
> library(gtools)   # this library has the inv.logit()  and logit() functions
> inv.logit(0.1889)
[1] 0.54708
```

Two factors here were compressed into two levels from three. Here's how I did that in R.

```
> summary(dd$age)
   mid   old young
    80    80    80
> levels(dd$age) <- c("15-55","55-70","15-55")

> summary(dd$Mall)
     CityCenter     Easton    Polaris
           79         80         81
> levels(dd$Mall) <- c("CityCenter","EastonPolaris","EastonPolaris")
```

The logistic regression is then:

```
> dd.glm <- glm(str~gender+emphatic+age+Mall+bank,family=binomial, data=dd)

> anova(dd.glm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: str

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       239    263.127
gender    1    8.432       238    254.695    0.004
emphatic  1   20.837       237    233.858 5.001e-06
age       1   19.102       236    214.755 1.239e-05
Mall      1    5.171       235    209.584    0.023
bank      4   38.469       231    171.115 8.965e-08
```

Recall that Deviance is -2[L] where L is the log likelihood of the current model versus the saturated model. So, the log likelihood of the final model in the list (the one that has all of the terms added is:

```
> 171.115/-2
[1] -85.5575
```

This corresponds to the Varbrul overall log likelihood which is reported as -86.8 for this analysis

(the R results are close but not exactly identical to the Varbrul output - perhaps a difference in the search algorithm being used or in the details of the data?).

From the table of coefficients we can calculate the varbrul weights.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3724     0.3682  -6.443 1.17e-10 ***
gender1      -0.6937     0.2017  -3.439 0.000585 ***
emphatic1     1.0044     0.2110   4.760 1.94e-06 ***
age1          1.0168     0.2656   3.829 0.000129 ***
Mall1        -0.3778     0.2426  -1.557 0.119370
bank1         0.7714     0.4397   1.754 0.079353 .
bank2         0.0633     0.3584   0.177 0.859800
bank3         0.6262     0.4830   1.296 0.194813
bank4        -3.0241     0.8539  -3.542 0.000398 ***
```

Now to calculate the weights we can take the inverse logits of the logistic regression coefficients. For example the weights for gender in the Varbrul analysis were 0.33 for men and 0.66 for women. We see from the R analysis that the contrast for gender `contrasts(dd$gender)`has men coded as 1 and women coded as -1. So, taking the inverse logit of the gender coefficient times the gender codes gives the same values that are in the Varbrul output.

```
inv.logit(-.6937 * 1) = 0.3332105   # for men
inv.logit(-.6937 * -1) = 0.6667895  # for women
```

Similarly, the Varbrul weights for more and less emphatic productions are given by noting (from `contrasts(dd$emphatic)`) that less emphatic was coded with 1 and more emphatic was coded with -1. The inverse logit of the coefficient gives the Varbrul weights.

```
inv.logit(1.004 * 1) = 0.7318443    # less emphatic
inv.logit(1.004 * -1) = 0.2681557   # more emphatic
```

The App/Total column in the Varbrul output is the actual probability of "shtr" in the dataset. In R we compute this from the contingency table. For example, the proportion of utterances with "shtr" in speech produced by men was 19 out of 120 total observations, for a proportion of 0.158. This is the value shown in the App/Total for men in the Varbrul printout.

```
> table(dd$str,dd$gender)
      m   w
str  101  82
shtr  19  38
```

Finally, the Input & Weight column is the predicted proportion of "shtr" given by the model parameters. In this case we find the same degree of mismatch between predicted and actual that

was found in Varbrul. The predicted proportion of "shtr" productions is the inverse logit of the intercept plus the factor coefficient.

```
inv.logit(-2.3724 - 0.6937) = 0.04452746  # for men
inv.logit(-2.3724 + 0.6937) = 0.1572677   # for women
```

The logistic regression model is predicting that men will produce "shtr" 4% of the time, when actually they say "shtr" in almost 16% of their utterances.

Exercises

1. Calculate χ^2 to test the hypothesis that one of your classes has an equal number of men and women.

2. In section 5.5.1, I said that the χ^2 for the data in table 5.5 (testing whether emphasis influences [Str] production) was 19.35. Table 5.6 also shows that the G^2 value from a logistic regression of these data is 20.08. Compute these two numbers using the formulas in section 5.5.1.

3. Using "Robins_data.txt". Produce a contingency table of the variables lvoc (which you will have to create as in the R-note above), and gender. Show how to calculate the χ^2 to test whether /l/ vocalization depends on the gender of the talker - this means that you produce a table like table 5.1. Then check your answer with summary(table(lvoc,gender)).

4. Two variables in the DDRASSTR.txt data set code economic class. One is called "class" and has three levels (Working Class, Lower Middle Class, and Upper Middle Class). The other is called "bank" and has five levels (middle and upper working class, and three levels of middle class). Compare logistic regression models that include one or the other of these economic class variables evaluating their overall significance and the significance of their interactions with other predictor variables. Also examine the model coefficients that correspond to different levels of these factors. Last, but not least, look at the data! Examine the actual probability of "shtr" as a function of economic class and other variables. Your goal in all of this activity is to decide which of the two methods of encoding economic class seems to be more sensible or valuable in analyzing the data (if such a determination can be made).

```
dd.glm.class <- glm(str~age*gender*class*emphatic, family=binomial, data=dd)
dd.glm.bank <- glm(str~age*gender*bank*emphatic, family=binomial, data=dd)
```