Lecture 13: Introduction to generalized linear models

21 November 2007

1 Introduction

Recall that we've looked at linear models, which specify a conditional probability density $P(Y|X)$ of the form

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \tag{1}$$

Linear models thus assume that the only stochastic part of the data is the normally-distributed noise ϵ around the predicted mean. Yet many (most?) types of data do not meet this assumption at all. These include:

- Continuous data in which noise is not normally distributed;
- Count data, in which the outcome is restricted to non-negative integers;
- Categorical data, where the outcome is one of a number of discrete classes.

One of the important developments in statistical theory over the past several decades has been the broadening of linear models from the classic form given in Equation (1) to encompass a much more diverse class of probabilistic distributions. This is the class of GENERALIZED LINEAR MODELS (GLMs). The next section will describe, step by step, how the generalization from classic linear models is attained.

1.1 Generalizing the classic linear model

The right-hand side of Equation (1) has two components: a deterministic component determining the *predicted mean*, and a stochastic component expressing the *noise distribution* around that mean:

$$Y = \overbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise}(\sim\mathcal{N}(0,\sigma^2))} \tag{2}$$

The first step from classic linear models to generalized linear models is to break these two components apart and specify a more indirect functional relationship between them. In the first step, we start with the idea that for any particular set of predictor variables $\{X_i\}$, there is a predicted mean μ. The probability distribution on the response Y is a function of that μ.[1] We'll review here what this means for linear models, writing both the abbreviated form of the model and the resulting probability density on the response Y:

$$\begin{aligned} Y &= \mu + \epsilon \qquad (3) \\ p(Y = y; \mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{\sigma^2}} \qquad (4) \\ & \qquad (5) \end{aligned}$$

By choosing other functions to map from μ to $p(y)$, we can get to other probability distributions, such as the Poisson distribution over the non-negative integers (see Lecture 2, section 5):

$$P(Y = y; \mu) = \frac{e^{-\mu}}{y!}\mu^y \tag{6}$$

In the second step, we loosen the relationship between the predicted mean and the predictor variables. In the classic linear model of Equation (1), the predicted mean was a linear combination of the predictor variables. In generalized linear models, we call this linear combination η and allow the

[1] There is actually a further constraint on the functional relationship between μ and $f(y)$, which I'm not going into—see McCullagh and Nelder (1989) or Venables and Ripley (2002, Chapter 7) for more details.

predicted mean to be an invertible function of η. We call η the LINEAR PREDICTOR, and call the function relating μ to η the LINK FUNCTION:

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n \qquad \text{(linear predictor)} \qquad (7)$$

$$l(\mu) = \eta \qquad \text{(link function)} \qquad (8)$$

In classic linear regression, the link function is particularly simple: it is the identity function, so that $\eta = \mu$.

Summary: generalized linear models are a broad class of models predicting the outcome of a response as a function of some linear combination of a set of predictors. To define a GLM, you need to choose (a) a link function relating the linear predictor to the predicted mean of the response; and (b) a function defining the "noise" or "error" probability distribution around that mean. For a classical linear model, the link function is the identity function

1.2 Logistic regression as a generalized linear model

Suppose we want a GLM that models binomially distributed data from n trials. We will use a slightly different formulation of the binomial distribution from what we introduced in Lecture 2: instead of viewing the response as the number of successful trials r, we view the response as the *proportion* of successful trials $\frac{r}{n}$; call this Y. Now, the mean number of successes for a binomial distribution is pn; hence the mean proportion is p. Thus p is the predicted mean μ of our GLM. This gives us enough information to specify precisely the resulting model:

$$P(Y = y; \mu) = \binom{n}{yn} \mu^{ny} (1-\mu)^{n(1-y)} \qquad \text{(or equivalently, replace } \mu \text{ with } p) \qquad (9)$$

This should look familiar from Lecture 2, Section 2.

This is part (b) of designing a GLM: choosing the distribution on Y given the mean μ. Having done this means that we have placed ourselves in the BINOMIAL GLM FAMILY. The other part of specifying our GLM is (a): choosing a relationship between the linear predictor η and the mean μ. Unlike the case with the classical linear model, the identity link function is not a possibility, because η can potentially be any real number, whereas the mean proportion μ of successes can only vary between 0 and 1. There are many link functions that can be chosen to make this mapping valid, but here we will use the LOGIT link function (here we replace μ with p for simplicity):[2]

$$\log \frac{p}{1-p} = \eta \tag{10}$$

or equivalently,

$$p = \frac{e^{\eta}}{1+e^{\eta}} \tag{11}$$

When we plunk the full form of the linear predictor from Equation (7) back in, we arrive at the final formula for logistic regression:

Logistic regression formula:

$$p = \frac{e^{\alpha+\beta_1 X_1+\cdots+\beta_n X_n}}{1+e^{\alpha+\beta_1 X_1+\cdots+\beta_n X_n}} \tag{12}$$

This type of model is also called a LOGIT MODEL.

1.3 Fitting a simple logistic regression model

The most common criterion by which a logistic regression model for a dataset is exactly the way that we chose the parameter estimates for a linear regression model: the method of maximum likelihood. That is, we choose the parameter estimates that give our dataset the highest likelihood.

We will give a simple example using the `dative` dataset. The response variable here is whether the recipient was realized as an NP (i.e., the double-object construction) or as a PP (i.e., the prepositional object construction). This corresponds to the `RealizationOfRecipient` variable in the dataset. There are several options in `R` for fitting basic logistic regression models, including `glm()` in the `stats` package and `lrm()` in the `Design` package. In this case we will use `lrm()`. We will start with a simple study of the effect of recipient pronominality on the dative alternation. Before fitting a model, we examine a contingency table of the outcomes of the two factors:

[2] Two other popular link functions for binomial GLMs are the PROBIT link and the COMPLEMENTARY LOG-LOG link. See Venables and Ripley (2002, Chapter 7) for more details.

```
> xtabs(~ PronomOfRec + RealizationOfRecipient,dative)
               RealizationOfRecipient
PronomOfRec       NP   PP
  nonpronominal  600  629
  pronominal    1814  220
```

So sentences with nonpronominal recipients are realized roughly equally often with DO and PO constructions; but sentences with pronominal recipients are recognized nearly 90% of the time with the DO construction. We expect our model to be able to encode these findings.

It is now time to construct the model. To be totally explicit, we will choose ourselves which realization of the recipient counts as a "success" and which counts as a "failure" (although `lrm()` will silently make its own decision if given a factor as a response). In addition, our predictor variable is a factor, so we need to use dummy-variable encoding; we will satisfice with the `R` default of taking the alphabetically first factor level, `nonpronominal`, as the baseline level.

```
> response <- ifelse(dative$RealizationOfRecipient=="PP",
                1,0) # code PO realization as success, DO as failure
> lrm(response ~ PronomOfRec, dative)

Logistic Regression Model

lrm(formula = response ~ PronomOfRec, data = dative)


Frequencies of Responses
   0    1
2414  849

       Obs  Max Deriv Model L.R.       d.f.          P          C        Dxy
      3263      2e-12     644.08          1          0      0.746      0.492
     Tau-a         R2      Brier
      0.19      0.263      0.154

                       Coef    S.E.    Wald Z P
Intercept               0.0472 0.05707   0.83 0.4082
PronomOfRec=pronominal -2.1569 0.09140 -23.60 0.0000
```

The thing to pay attention to for now is the estimated coefficients for the intercept and the dummy indicator variable for a pronominal recipient. We can use these coefficients to determine the values of the linear predictor η and the predicted mean success rate p using Equations (7) and (12):

$$\eta_{--} = 0.0472 + (-2.1569) \times 0 \quad = 0.0472 \quad \text{(non-pronominal receipient)} \tag{13}$$

$$\eta_{+} = 0.0472 + (-2.1569) \times 1 \quad = -2.1097 \quad \text{(pronominal recipient)} \tag{14}$$

$$p_{\text{nonpron}} = \frac{e^{0.0472}}{1 + e^{0.0472}} \quad = 0.512 \tag{15}$$

$$p_{\text{pron}} = \frac{e^{-2.1097}}{1 + e^{-2.1097}} \quad = 0.108 \tag{16}$$

When we check these predicted probabilities of PO realization for nonpronominal and pronominal recipients, we see that they are equal to the proportions seen in the corresponding rows of the cross-tabulation we calculated above: $\frac{629}{629+600} = 0.518$ and $\frac{220}{220+1814} = 0.108$. This is exactly the expected behavior, because (a) we have two parameters in our model, α and β_1, which is enough to encode an arbitrary predicted mean for each of the cells in our current representation of the dataset; and (b) as we have seen before (Lecture 5, Section 2), the maximum-likelihood estimate for a binomial distribution is the relative-frequency estimate—that is, the observed proportion of successes.

1.4 Multiple logistic regression

Just as we were able to perform multiple linear regression for a linear model with multiple predictors, we can perform multiple logistic regression. Suppose that we want to take into account pronominality of both recipient and theme. First we conduct a complete cross-tabulation and get proportions of PO realization for each combination of pronominality status:

apply()

```
> tab <- xtabs(~ RealizationOfRecipient + PronomOfRec + PronomOfTheme, dative)
> tab
, , PronomOfTheme = nonpronominal

                      PronomOfRec
```

```
RealizationOfRecipient nonpronominal pronominal
                    NP           583       1676
                    PP           512         71

, , PronomOfTheme = pronominal

                      PronomOfRec
RealizationOfRecipient nonpronominal pronominal
                    NP            17        138
                    PP           117        149
> apply(tab,c(2,3),function(x) x[2] / sum(x))
               PronomOfTheme
PronomOfRec     nonpronominal pronominal
  nonpronominal     0.4675799  0.8731343
  pronominal        0.0406411  0.5191638
```

Pronominality of the theme consistently increases the probability of PO realization; pronominality of the recipient consistently increases the probability of DO realization.

We can construct a logit model with independent effects of theme and recipient pronominality as follows:

```
> dative.lrm < - lrm(response ~ PronomOfRec + PronomOfTheme, dative)
> dative.lrm

Logistic Regression Model

lrm(formula = response ~ PronomOfRec + PronomOfTheme, data = dative)


Frequencies of Responses
   0    1
2414  849

       Obs  Max Deriv Model L.R.       d.f.          P           C        Dxy
      3263      1e-12    1122.32          2          0       0.827      0.654
     Tau-a         R2      Brier
     0.252      0.427      0.131
```

```
                            Coef    S.E.    Wald Z P
Intercept                   -0.1644 0.05999  -2.74 0.0061
PronomOfRec=pronominal      -2.8670 0.12278 -23.35 0.0000
PronomOfTheme=pronominal     2.9769 0.15069  19.75 0.0000
```

And once again, we can calculate the predicted mean success rates for each of the four combinations of predictor variables:

Recipient	Theme	η	$\hat{p}$
nonpron	nonpron	-0.1644	0.459
pron	nonpron	-3.0314	0.046
nonpron	pron	2.8125	0.943
pron	pron	-0.0545	0.486

In this case, note the predicted proportions of success are not the same as the observed proportions in each of the four cells. This is sensible – we cannot fit four arbitrary means with only three parameters. If we added in an interactive term, we would be able to fit four arbitrary means, and the resulting predicted proportions would be the observed proportions for the four different cells.

1.5 Multiplicativity of the odds

Let us consider the case of a dative construction in which both the recipient and theme are encoded with pronouns. In this situation, both the dummy indicator variables (indicating that the theme and recipient are pronouns) have a value of 1, and thus the linear predictor consists of the sum of three terms. From Equation (10) we can write

$$\frac{p}{1-p} = e^{\alpha+\beta_1+\beta_2} \tag{17}$$

$$= e^{\alpha}e^{\beta_1}e^{\beta_2} \tag{18}$$

The ratio $\frac{p}{1-p}$ is the odds of success, and in logit models the effect of any predictor variable on the response variable is multiplicative in the odds of success. If a predictor has coefficent β in a logit model, then a unit of that predictor has a multiplicative effect of e^{β} on the odds of success.

Unlike the raw coefficient β, the quantity e^{β} is not linearly symmetric—it falls in the range $(0, \infty)$. However, we can also perform the full reverse

			Multiplicative
Predictor	Coefficient	Factor Weight	effect on odds
Intercept	-0.1644	0.4590	0.8484
Pronominal Recipient	-2.8670	0.0538	0.0569
Pronominal Theme	2.9769	0.9515	19.627

Table 1: Logistic regression coefficients and corresponding factor weights for each predictor variable in the `dative` dataset.

Recip.	Theme	Linear Predictor	Multiplicative odds	P(PO)
–pron	–pron	-0.16	0.8484	0.46
+pron	–pron	$-0.16 - 2.87 = -3.03$	$0.85 \times 0.06 = 0.049$	0.046
–pron	+pron	$-0.16 + 2.98 = 2.81$	$0.85 \times 19.6 = 16.7$	0.94
+pron	+pron	$-0.16 - 2.87 + 2.98 = -0.05$	$0.85 \times 0.06 \times 19.63 = 0.947$	0.49

Table 2: Linear predictor, multiplicative odds, and predicted values for each combination of recipient and theme pronominality in the `dative` dataset. In each case, the linear predictor is the log of the multiplicative odds.

logit transform of Equation (11), mapping β to $\frac{e^\beta}{1+e^\beta}$ which ranges between zero and 1, and is linearly symmetric around 0.5. The use of logistic regression with the reverse logit transform has been used in quantitative sociolinguistics since Cedergren and Sankoff (1974) (see also Sankoff and Labov, 1979), and is still in widespread use in that field. In quantitative sociolinguistics, the use of logistic regression is often called VARBRUL (variable rule) analysis, and the parameter estimates are reported in the reverse logit transform, typically being called factor weights.

Tables 1 and 2 show the relationship between the components of the linear predictor, the components of the multiplicative odds, and the resulting predictions for each possible combination of our predictor variables.

2 Confidence intervals and model comparison in logit models

We'll close our introduction to logistic regression with discussion of confidence intervals and model comparison.

2.1 Confidence intervals for logistic regression

When there are a relatively large number of observations in comparison with the number of parameters estimated, the standardized deviation of the MLE for a logit model parameter θ is approximately normally distributed:

$$\frac{\hat{\theta} - \theta}{\text{StdErr}(\hat{\theta})} \overbrace{\sim}^{\text{approx}} \mathcal{N}(0, 1) \tag{19}$$

This is called the WALD STATISTIC[3]—note the close similarity with the t statistic that we were able to use for classic linear regression in Lecture 11 (remember that once the t distribution has a fair number of degrees of freedom, it looks a great deal like a standard normal distribution). If we look again at the output of the logit model we fitted in the previous section, we see the standard error, which allows us to construct confidence intervals on our model parameters.

```
                         Coef    S.E.    Wald Z P
Intercept                -0.1644 0.05999  -2.74 0.0061
PronomOfRec=pronominal   -2.8670 0.12278 -23.35 0.0000
PronomOfTheme=pronominal  2.9769 0.15069  19.75 0.0000
```

The Wald statistic can also be used for a frequentist test on the null hypothesis that an individual model parameter is 0. This is the source of the p-values given for the model parameters above.

2.2 Model comparison

Just as in the analysis of variance, we are often interested in conducting tests of the hypothesis that introducing *several* model parameters simultaneously leads to a better overall model. In this case, we cannot simply use a single Wald statistic for hypothesis testing. Instead, the most common approach is to use the LIKELIHOOD-RATIO TEST. A generalized linear model assigns a likelihood to its data as follows:

[3] It is also sometimes called the Wald Z statistic, because of the convention that standard normal variables are often denoted with a Z, and the Wald statistic is distributed approximately as a standard normal.

$$\text{Lik}(\vec{x};\hat{\theta}) = \prod_i P(x_i|\hat{\theta}) \tag{20}$$

Now suppose that we have two classes of models, M_0 and M_1, and M_0 is nested inside of M_1 (that is, the class M_0 is a "special case" of the class M_1). It turns out that if the data are generated from a model M_0 is the correct model, the ratio of the data likelihoods in the ML estimates for M_0 and M_1 is well-behaved. In particular, twice the log of the likelihood ratio is distributed as a χ^2 random variable with degrees of freedom equal to the difference k in the number of free parameters in the two models. This quantity is sometimes called the DEVIANCE:

$$2\log\frac{\text{Lik}_{M_1}(\vec{x})}{\text{Lik}_{M_0}(\vec{x})} = 2\left[\log\text{Lik}_{M_1}(\vec{x}) - \log\text{Lik}_{M_0}(\vec{x})\right] \qquad \sim \chi^2_k \tag{21}$$

As an example of using the likelihood ratio test, we will hypothesize a model in which pronominality of theme and recipient both still have additive effects but that these effects may vary depending on the modality (spoken versus written) of the dataset. We fit this model and our modality-independent model using `glm()`, and use `anova()` to calculate the likelihood ratio:

`glm()`

```
> m.0 <- glm(response ~ PronomOfRec + PronomOfTheme,dative,family="binomial")
> m.A <- glm(response ~ PronomOfRec*Modality + PronomOfTheme*Modality,dative,famil
> anova(m.0,m.A)
Analysis of Deviance Table

Model 1: response ~ PronomOfRec + PronomOfTheme
Model 2: response ~ PronomOfRec * Modality + PronomOfTheme * Modality
  Resid. Df Resid. Dev   Df Deviance
1      3260    2618.74
2      3257    2609.67    3     9.07
```

We can look up the p-value of this deviance result in the χ^2_3 distribution:

```
> 1-pchisq(9.07,3)
[1] 0.02837453
```

Thus there is some evidence that we should reject a model that doesn't include modality-specific effects of recipient and theme pronominality.

3 Further reading

There are many places to go for reading more about generalized linear models and logistic regression in particular. The classic comprehensive reference on generalized linear models is McCullagh and Nelder (1989). For GLMs on categorical data, Agresti (2002) and the more introductory Agresti (2007) are highly recommended. For more information specific to the use of GLMs and logistic regression in R, Venables and Ripley (2002, Section 7), Harrell (2001, Chapters 10–12), and Maindonald and Braun (2007, Section 8.2) are all good places to look.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley, second edition.

Cedergren, H. J. and Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language*, 50(2):333–355.

Harrell, Jr, F. E. (2001). *Regression Modeling Strategies*. Springer.

Maindonald, J. and Braun, J. (2007). *Data Analysis and Graphics using R*. Cambridge, second edition.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, second edition.

Sankoff, D. and Labov, W. (1979). On the uses of variable rules. *Language in Society*, 8:189–222.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, fourth edition.