Logit, Poisson, and Cox regression models: summary notes

James G. Scott

Spring 2015

1 Logistic regression

Example data sets and scripts: spam, brca, gardasil, cmc, resume

The linear probability model. In many situations, we would like to forecast the outcome of a binary event, given some relevant information:

- Given the pattern of word usage and punctuation in an e-mail, is it likely to be spam?
- Given the temperature and cloud cover on Christmas Eve, is it likely to snow on Christmas?
- Given a person's credit history, is he or she likely to default on a mortgage?

In all of these cases, the y variable is the answer to a yes-or-no question. Nonetheless, we can still use regression for these problems. Let's suppose, for simplicity's sake, that we have only one predictor x, and that we let $y_i = 1$ for a "yes" and $y_i = 0$ for a "no." One naïve way of forecasting y is simply to plunge ahead with the basic, one-variable regression equation:

$$\mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

Since y_i can only take the values 0 or 1, the expected value of y_i is simply a weighted average of these two cases:

$$\begin{aligned} \mathrm{E}(y_i \mid x_i) &= 1 \cdot P(y_i = 1 \mid x_i) + 0 \cdot P(y_i = 0 \mid x_i) \\ &= P(y_i = 1 \mid x_i) \end{aligned}$$

Therefore, the regression equation is just a linear model for the conditional probability that $y_i = 1$, given the predictor x_i:

$$P(y_i = 1 \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

This model allows us to plug in some value of x_i and read off the forecasted probability of a "yes" answer to whatever yes-or-no question is being posed. It is often called the linear probability model, since the probability of a "yes" varies linearly with x.

The logistic link function. The linear probability model is perfectly reasonable in many situations. But suffers from a noticeable problem. The left-hand side of the regression equation, $P(y_i = 1 \mid x_i)$, must be between 0 and 1. But the right-hand side, $\beta_0 + \beta_1 x_i$, can be any real number between $-\infty$ and ∞. We'd be better off with some transformation g that takes an unconstrained number from the right-hand side, and maps it to a constrained number on the left-hand side:

$$P(y_i \mid x_i) = g(\beta_0 + \beta_1 x_i).$$

Such a function g is called a *link function*. A model that incorporates such a link function is called a *generalized linear model*; and the part inside the parentheses $(\beta_0 + \beta_1 x_i)$ is called the *linear predictor*, and is often denoted as ψ_i.

We use link functions and generalized linear models in most situations where we are trying to predict a number that is, for whatever reason, constrained. Here, we're dealing with probabilities, which are constrained to be no smaller than 0 and no larger than 1. Therefore, the function g must map real numbers on $(-\infty, \infty)$ to numbers on $(0, 1)$. It must therefore be shaped a bit like a flattened letter "S," approaching zero for large negative values of ψ_i, and approaching 1 for large positive values.

With multiple regressors $(x_{i1}, \ldots, x_{ip})$, we have

$$\Pr(y_i = 1 \mid x_i) = w_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}. \tag{1}$$

Recall that odds are just a different way of expressing probabilities:

$$(\text{Odds that } y_i \text{ is } 1) = O_i = \frac{w_i}{1 - w_i}.$$

If you churn through the algebra and re-express the logistic-regression equation (1) in terms of odds, you will see that the log-odds of success—or equivalently the *logit transform* of the success probability—are being modeled as a linear function of the predictors:

$$\text{logit}(w_i) = \log O_i = \log\left(\frac{w_i}{1 - w_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

A technical aside: this model cannot be fit by least squares. Instead, it is fit via maximum-likelihood, and requires a nonlinear optimization routine. The most commonly used is a variation on the Newton–Raphson algorithm called *iteratively re-weighted least squares.* This can sometimes break! Thus if you are getting very strange answers

Interpreting the coefficients. For the sake of simplicity, imagine a data set with only a single regressor x_i that can take the values 0 or 1 (a dummy variable). Perhaps, for example, x_i denotes whether someone received the new treatment (as opposed to the control) in a clinical trial.

For this hypothetical case, let's consider the ratio of two quantities: the odds of success for

person i with $x_i = 1$, versus the odds of success for person j with $x_j = 0$. Denote this ratio by R_{ij}. We can write this as

$$\begin{aligned} R_{ij} &= \frac{O_i}{O_j} \\ &= \frac{\exp\{\log(O_i)\}}{\exp\{\log(O_j)\}} \\ &= \frac{\exp\{\beta_0 + \beta_1 \cdot 1\}}{\exp\{\beta_0 + \beta_1 \cdot 0\}} \\ &= \exp\{\beta_0 + \beta_1 - \beta_0 - 0\} \\ &= \exp(\beta_1). \end{aligned}$$

Therefore, we can interpret the quantity e^{β_1} as an *odds ratio*. Since $R_{ij} = O_i/O_j$, we can also write this as:

$$O_i = e^{\beta_1} \cdot O_j .$$

In words: if we start with $x = 0$ and move to $x = 1$, our odds of success ($y = 1$) will change by a multiplicative factor of e^{β_1}.

The ordinal logit model. We can modify the logistic regression model to handle ordinal responses. The hallmark of ordinal variables is that they are measured on a scale that can't easily be associated with a numerical magnitude, but that does imply an ordering: employee evaluations, survey responses, bond ratings, and so forth.

There are several varieties of ordinal logit model. Here we consider the *proportional-odds* model, which is most easily understood as a family of related logistic regression models. Label the categories as $1, \ldots, K$, ordered in the obvious way. Consider the probability $c_{ik} = P(y_i \leq k)$: the probability that the outcome for the ith case falls in category k *or any lower category.* (We call it c_{ik} because it is a cumulative probability of events at least as "low" as k.) The proportional-odds logit model assumes that the logit transform of c_{ik} is a linear function of predictors:

$$\text{logit}(c_{ik}) = \log\left(\frac{c_{ik}}{1 - c_{ik}}\right) = \eta_k + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} .$$

Crucially, this relationship is assumed to hold for all categories at once. Because $c_{iK} = 1$ for the highest category K, we have specified $K - 1$ separate binary logit models that all share the same predictors x_j and the same coefficients β_j. The only thing that differs among the models are the intercepts η_k; these are commonly referred to as the *cutpoints*. Since the log odds differ only by an additive constant for different categories, the odds differ by a multiplicative factor—thus the term "proportional odds."

To interpret the ordinal-logit model, I find it easiest to re-express individual fitted values in

terms of covariate-specific category probabilities $w_{ik} = P(y_i = k)$:

$$w_{ik} = P(y_i \leq k) - P(y_i \leq k-1) = c_{ik} - c_{i,k-1},$$

with the convention that $c_{i0} = 0$. Good software makes it fairly painless to do this.

The multinomial logit model. Another generalization of the binary logit model is the multinomial logit model. This is intended for describing *unordered* categorical responses: PC/Mac/Linux, Ford/Toyota/Chevy, plane/train/automobile, and so forth. Without a natural ordering to the categories, the quantity $P(y_i \leq k)$ ceases to be meaningful, and we must take a different approach.

Suppose there are K possible outcomes ("choices"), again labeled as $1, \ldots, K$ (but without the implied ordering). As before, let $w_{ik} = P(y_i = k)$. For every observation, and for each of the K choices, we imagine that there is a linear predictor ψ_{ik} that measures the preference of subject i for choice k. Intuitively, the higher ψ_{ik}, the more likely that $y_i = k$.

The specific mathematical relationship between the linear predictors and the probabilities w_{ik} is given the multinomial logit transform:

$$\begin{aligned} w_{ik} &= \frac{\exp(\psi_{ik})}{\sum_{l=1}^{K} \exp(\psi_{il})} \\ \psi_{ik} &= \beta_0^{(k)} + \beta_1^{(k)} x_{i1} + \cdots \beta_p^{(k)} x_{ip}. \end{aligned}$$

Each category gets its own set of coefficients, but the same set of predictors x_1 through x_p.

There is one minor issue here. With a bit of algebra, you could convince yourself that adding a constant factor to each ψ_{ik} would not change the resulting probabilities w_{ik}, as this factor would cancel from both the numerator and denominator of the above expression. To fix this indeterminacy, we choose one of the categories (usually the first or last) to be the reference category, and set its coefficients equal to zero.

2 Models for count outcomes

Example data sets and scripts: springbok, flutrends

The Poisson model. For modeling event-count data (photons, organisms, heart attacks), a useful place to start is the Poisson distribution. The key feature of counts is that they must be non-negative integers. Like the case of logistic regression, where probabilities had to live between 0 and 1, this restriction creates some challenges that take us beyond ordinary least squares.

The Poisson distribution is parametrized by a rate parameter, often written as λ. Let k denote an integer, and y_i denote the event count for subject i. In a Poisson model, we assume that

$$P(y_i = k) = \frac{\lambda_i^k}{k!} e^{-\lambda_i},$$

and we wish to model λ_i in terms of covariates. Because the rate parameter of the Poisson cannot be negative, we must employ the same device of a link function to relate λ_i to covariates. By far the most common is the (natural) log link:

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip},$$

or equivalently,

$$\lambda_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip}\}.$$

As with the case of logistic regression, the model is fit via maximum-likelihood.

Interpreting the coefficients. Because we are fitting a model on the log-rate scale, additive changes to an x variable are associated with multiplicative changes in the y variable. As before, let's consider the ratio of two quantities: the rate of events for person i with $x_1 = x^\star + 1$, versus the rate of events for person j with $x_1 = x^\star$. Let's further imagine that all other covariates are held constant at values x_2 to x_p, respectively. This implies that the only difference between subjects i and j is a one-unit difference in the first predictor, x_1.

We can write their ratio of rates as

$$\begin{aligned} R_{ij} &= \frac{\lambda_i}{\lambda_j} \\ &= \frac{\exp\{\beta_0 + \beta_1 \cdot (x^\star + 1) + \beta_2 x_2 + \cdots \beta_p x_p\}}{\exp\{\beta_0 + \beta_1 \cdot x^\star + \beta_2 x_2 + \cdots \beta_p x_p\}} \\ &= \exp\{\beta_1 (x^\star + 1 - x^\star)\} \\ &= \exp(\beta_1). \end{aligned}$$

Thus person i experiences events events e^{β_1} times as frequently as person j.

Overdispersion. For most data sets outside of particle physics, the Poisson assumption is usually one of convenience. Like the normal distribution, it is familiar and easy to work with. It also has teeth, and may bite if used improperly. One crucial feature of the Poisson is that its mean and variance are equal: that is, if $y_i \sim \text{Pois}(\lambda_i)$, then the expected value of y_i is λ_i, and the standard deviation of y_i is $\sqrt{\lambda_i}$. (Since λ_i depends on covariates, we should really be calling these the *conditional* expected value and standard deviation.)

As a practical matter, this means that if your data satisfy the Poisson assumption, then roughly 95% of observations should fall within $\pm 2\sqrt{\lambda_i}$ of their conditional mean λ_i. This is quite narrow, and many (if not most) data sets exhibit significantly more variability about their mean. If the conditional variance exceeds the conditional mean, the data exhibits *overdispersion with respect to the Poisson*, or just *overdispersion* for short.

Overdispersion can really mess with your standard errors. In other words, if you use (i.e. let

your software use) the Poisson assumption to calculate error bars, but your data are overdispersed, then you will end up overstating your confidence in the model coefficients. Sometimes the effect is dramatic, meaning that the blind use of the Poisson assumption is a recipe for trouble.

There are three common strategies for handling overdispersion:

1. Use a quasi-likelihood approach ("family=quasipoisson" in R's glm function);
2. Fit a different count-data model, such as the negative binomial or Poisson-lognormal, that can accommodate overdispersion;
3. Fit a hierarchical model.

3 Survival analysis

Example data sets and scripts: colon, recid

Survival times. Suppose that we decide to run an epidemiological cohort study, which is a kind way of saying that we follow people and wait until something bad happens to them (an "event"). Let T_i be the time elapsed from the start of the study until the event. The random variable T_i is often called a survival time—even if the event in question isn't an actual death—or alternatively, a failure time.

In most studies of this kind, the goal is to understand how a subject's survival time depends on covariates:

- Under which treatment arm of a clinical trial do people survive longer?
- Does this computer screen last longer under manufacturing process A or B?
- Do criminals who read Nietzsche in prison recidivate at higher rates?

There are many ways to proceed. We could directly model $F(t) = P(T_i \leq t)$, the cumulative distribution function of the random variable T_i. Equivalently, we could model the corresponding probability density $f(t)$, or the *survival curve* $S(t) = 1 - F(t) = P(T_i > t)$. This is the most natural extension of regression analysis—specify a probability model, and describe changes in the model's parameters as a function of covariates. Many approaches to survival analysis involve just this; examples include the Weibull, gamma, and log-normal.

Modeling the hazard function. An alternative approach is to model the *hazard function*, denoted $h(t)$:

$$h_i(t) \approx \frac{P\left(t < T_i < t + \Delta t \mid T_i > t\right)}{\Delta t},$$

for some small time interval of width Δt. We actually define the hazard function using calculus, as the limit of this quantity as Δt approaches 0. Intuitively, the hazard function is the instantaneous

rate of failure at time t, conditional upon having survived up to time t. It turns out that the density $f(t)$ and the hazard function $h(t)$ can be used to give mathematically equivalent specifications of the distribution of the random survival time T_i.

The Cox proportional-hazards model is a model for the hazard function $h(t)$. It is the most popular tool for survival analysis because it is simple, and because it can easily accommodate *right-censoring*: that is, the presence of subjects in the data set who have not yet experienced a failure by the end of the study period. Virtually all survival analyses involve right-censoring, which is not as easily or transparently handled in models for $f(t)$.

The key assumption of the Cox model is proportionality, or separability. Specifically, it assumes that subject i, having covariates x_{i1} through x_{ip}, has the hazard function

$$h_i(t) = h_0(t) \cdot \exp\left\{\beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip}\right\}.$$

Notice that $h_0(t)$ is a function of the time t. We call this the *baseline hazard function.* Everything else on the right-hand side just boils down to a single scalar $\exp(\psi_i)$ that depends on a subject's covariates. This factor uniformly inflates or deflates the baseline hazard across all values of t. The Cox model is therefore *semiparametric*, in that it allows a flexible nonparametric model for the baseline hazard $h_0(t)$, but requires that the effect of covariates enter through a parametric linear model.

Interpreting the coefficients. Consider the ratio of two hazard functions: the hazard for person i with $x_1 = x^\star + 1$, versus the hazard for person j with $x_1 = x^\star$. As before, we imagine that all other covariates are held constant at values x_2 to x_p, respectively. Thus the only difference between subjects i and j is a one-unit difference in the first predictor.

We can write their ratio of hazard functions as

$$\begin{aligned}
\frac{h_i(t)}{h_j(t)} &= \frac{h_0(t)\exp\{\beta_0 + \beta_1 \cdot (x^\star + 1) + \beta_2 x_2 + \cdots \beta_p x_p\}}{h_0(t)\exp\{\beta_0 + \beta_1 x^\star + \beta_2 x_2 + \cdots \beta_p x_p\}} \\
&= \exp\{\beta_1(x^\star + 1 - x^\star)\} \\
&= \exp(\beta_1).
\end{aligned}$$

Thus person i has a hazard function e^{β_1} times higher (or lower) than person j. Crucially, this is assumed to hold across all values of t. This explains why, to summarize the results of a Cox model, people usually exponentiate the coefficients and quote them as *hazard ratios*.