BAYESIAN LOGISTIC REGRESSION ANALYSIS

N. van Erp* and P. van Gelder†

*TU Delft, Netherlands
h.r.n.vanErp@tudelft.nl
†TU Delft, Netherlands

Abstract. In this paper we present a Bayesian logistic regression analysis. It is found that if one wishes to derive the posterior distribution of the probability of some event, then, together with the traditional Bayes Theorem and the integrating out of nuissance parameters, the Jacobian transformation is an essential added ingredient. The application of the product rule gives the posterior of the unknown logistic regression coefficients. The Jacobian transformation then maps the posterior of these regression coefficients to the posterior of the corresponding probability of some event and some nuisance parameters. Finally, by way of the sumrule the nuissance parameters are integrated out.

Keywords: Regression, Logistic Regression
PACS: 02.50.Ng

INTRODUCTION

A literature search on Bayesian logistic regression models will give one a collection of Monte Carlo schemes. In these schemes the posterior of the beta coefficients of the logistic regression model are constructed and sampled from. Each Monte Carlo realization of a set of regression coefficients corresponds with a probability of some event occurring. So, having obtained a set of realizations of the regression coefficients, we also obtain a corresponding set of realized probabilities. These probabilities then constitute an empirical probability distribution of the probability of some event. These procedures may be viewed as the Monte Carlo implementation of the Jacobian transformation method.

To the best of our knowledge, it is nowhere in the literature mentioned that the Monte Carlo schemes are the solution to a Jacobian transformation problem. Thus, we are of the believe that the here presented approach has some pedagogical merit. By applying the Jacobian transformation to the posterior of the logistic regression coefficients we may obtain an analytical expression of the posterior of the probability of some event. This expression may then be evaluated either numerically or by way of the above described Monte Carlo schemes. That is, we give here the analytical model which the Monte Carlo approach seeks to implement. And, as a rule, analytical models are easier to understand than their corresponding Monte Carlo implementations.

Furthermore, the Jacobian transformation approach may be seen as a general way to derive a class of beta-like distributions whichnot only take into account the number of successes and failures, but also, for example, the values on predictor variables or timeto failures. The former gives the logistic regression analysis and is treated in the main text

Bayesian Inference and Maximum Entropy Methods in Science and Engineering
AIP Conf. Proc. 1553, 147-154 (2013); doi: 10.1063/1.4819994

of this paper. The latter gives a generalization of the third example of Jaynes' [1], which, although already derived in [2] and further generalized in [3], will be given here as an appendix. So as to give the reader a better sense of the overall scope of the here proposed technique.

THE MODEL

Say we have a logistic probability model for a 'success', that is, a certain event happening:

$$\log \frac{\theta}{1-\theta} = \beta_0 + z\beta_1 \tag{1}$$

where z is some given value of some predictor, and β_0, β_1 are unknown regression parameters. Then the probability of a success is

$$\theta = \frac{e^{\beta_0+z\beta_1}}{1+e^{\beta_0+z\beta_1}} \tag{2}$$

Its complement, the probability of a 'failure', that is, a certain event not happening:

$$1-\theta = \frac{1}{1+e^{\beta_0+z\beta_1}} \tag{3}$$

THE LIKELIHOOD, PRIOR, AND POSTERIOR

We observe a sequence of r sucesses having observed predictors x_i, for $i = 1,\ldots,r$, and $n-r$ failures having observed predictors y_j, for $j = 1,\ldots,n-r$. From (2) and (3), it follows that the probability of observing r sucesses and $n-r$ failures, or, equivalently, the likelihood of the unknown parameters β_0 and β_1, is

$$p(D|\beta_0,\beta_1) = \prod_{i=1}^{r} \frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}} \prod_{j=1}^{n-r} \frac{1}{1+e^{\beta_0+y_j\beta_1}} \tag{4}$$

Next, we assign as a prior some uniform distribution to the unknown regression parameters β_0 and β_1

$$p(\beta_0,\beta_1|I) \propto \text{constant} \tag{5}$$

The posterior of β_0 and β_1, then may be found by combining likelihood, (4), with prior, (5):

$$p(\beta_0,\beta_1|D,I) \propto \prod_{i=1}^{r} \frac{e^{\beta_0+x_i\beta_1}}{1+e^{\beta_0+x_i\beta_1}} \prod_{j=1}^{n-r} \frac{1}{1+e^{\beta_0+y_j\beta_1}} \tag{6}$$

Now, we are not that much interested in the regression parameters β_0 and β_1, we want to find the posterior probability distribution of the probability θ. We observe, (2), that, for given z, the value of θ is directlydetermined by the values of both β_0 and β_1. Because of this two-to-one correspondence we may make a Jacobian transformation from β_0 and

β_1 to θ and, so, map the uncertainty regarding the regression parameters β_0 and β_1 unto the parameter of interest, θ, which is the probability of a success given some predictor value z.

THE JACOBIAN TRANSFORMATION

We have that, (2),

$$\theta = \frac{e^{\beta_0 + z\beta_1}}{1 + e^{\beta_0 + z\beta_1}}$$

So, a possible transformation would be

$$\beta_0 = -\log\left(\frac{1-\theta}{\theta} e^{z\beta_1}\right), \quad \beta_1 = b_1 \tag{7}$$

The corresponding Jacobian is

$$J = \begin{vmatrix} \frac{\partial}{\partial\theta}\beta_0 & \frac{\partial}{\partial b_1}\beta_0 \\ \frac{\partial}{\partial\theta}\beta_1 & \frac{\partial}{\partial b_1}\beta_1 \end{vmatrix} = \begin{vmatrix} \frac{1}{\theta(1-\theta)} & -z \\ 0 & 1 \end{vmatrix} = \frac{1}{\theta\,(1-\theta)} \tag{8}$$

Substituting (7) into the posterior (6) and multiplying it with the Jacobian (8) gives us the transformedposterior

$$p\,(\,\theta, b_1 | z, D, I) \propto \frac{1}{\theta\,(1-\theta)} \prod_{i=1}^{r} \frac{\frac{\theta}{1-\theta} e^{(x_i - z) b_1}}{1 + \frac{\theta}{1-\theta} e^{(x_i - z) b_1}} \prod_{j=1}^{n-r} \frac{1}{1 + \frac{\theta}{1-\theta} e^{(y_j - z) b_1}} \tag{9}$$

If we (numerically) integrate the unwanted parameter b_1 out of (9), we get the posterior of the probability θ, (2), given some predictor value z, and we have the Bayesian logistic regression model we are looking for

$$p\,(\,\theta | z, D, I) = \int p\,(\,\theta, b_1 | z, D, I)\, db_1 \tag{10}$$

A SPECIAL CASE

For non-informative data, that is, for predictors which all have the same value, $z = x_i = y_j$, for $i = 1, \ldots, r$ and $j = 1, \ldots, n-r$, the terms in the exponentials in (9) all become 0, and the posterior distribution for θ collapsesto the ordinary beta-distribution:

$$\begin{aligned} p\,(\,\theta | z, D, I) &\propto \frac{1}{\theta\,(1-\theta)} \prod_{i=1}^{r} \frac{\frac{\theta}{1-\theta}}{1 + \frac{\theta}{1-\theta}} \prod_{j=1}^{n-r} \frac{1}{1 + \frac{\theta}{1-\theta}} \int db_1 \\ &\propto \theta^{r-1} (1-\theta)^{n-r-1} \end{aligned} \tag{11}$$

This is in nice correspondence with our intuition. If the predictors are non-informative, in that they 'flat-line', then the onlypertinent aspect of our data D which remains is the number of successes, r, and the number of failures, $n-r$, and these are just the sufficient statistics of the beta-distribution (11).

DISCUSSION

We have presented here a Bayesian logistic regression analysis. It is found that if one wishes to derive the posterior distribution of the probability of some event, then, together with the traditional Bayes Theorem and the integrating out of the nuisance parameters, the Jacobian transformation is an essential added ingredient. Furthermore, the beta-distribution may be derived as a special case of this Bayesian logistic regression analysis, where the predictors are non-informative, in that they flat-line.

SOME ENCOUNTERED CRITICISMS

Now, once seen, the analytical solution of the Bayesian logistic regression model may seem too trivial to mention. We can only besympathetic to the fact that for those who are under this impression the following criticisms will be quick to come to mind. And we will try to defend our position on these issues as best we can.

One of the criticisms heard during the presentation of this article was that this Bayesian logistic regression analysis had already been derived. Though what was actually meant was that the posterior (6) for the unknown logistic regression coefficients has been derived many times over. But this misses the point. We do not propose a to derive a new kind of posterior for the logistic regression coefficients. Rather, we wish to show how, given the posterior (6), we may come to an analytical expressionof the Bayesian logistic regression model; (7) through (10). It is our belief that until now the Monte Carlo schemes were solutions to a problem which had not yet been properly articulated. Once we have established the analytical model we wishto implement it is easy to see that the Monte Carlo schemes, as described in the introduction, are just one of three ways to implement the model; the second way being a direct evaluation of (10) by way of numerical integration; the third way being an evaluation of the first four moments of (2), by way of (6), which then may be substituted in an maximum entropy distribution by way of the Jondeau algorithm, [4] and [5].

Another criticism, in the same vein, was that Jacobian transformations are performed routinely in a Bayesian context, [6] and [7]. But then again, Jacobian transformations are also performed routinely in an orthodox context, [8] and [9]. And this then, we believe, misses the point that the necessity of having to make a change of variables will probably elude those who try their hand at a Bayesian logistic regression analysis for the first time. Just as it has managed to elude these authors for the past ten years, and, for that matter, so it may seem, many others. Seeing that a google search on the words “Jacobian transformation” and “Bayesian logistic regression analysis” did not produce any articles with the here presented change of variables procedure.

So, what we have endeavored to do here is to write down the derivation of the Bayesian logistic regression analysis in such manner as we ourselves would have liked to find it in the text books some ten years ago, when the need for such an analysis first arose; that is, short and sweet.

ANOTHER BETA-LIKE DISTRIBUTION

The posterior (6) is a beta-like distribution in that it takes into account the number of observed successes and failures, (4), and its domain is constricted to $0 \leq \theta \leq 1$. More beta-like distributions may be derived by making either a simple change of variable or a Jacobian transformation, [2] and [3]. We will now proceed to give the derivation of [2] in terms of [1]; the article that started it all.

The Problem

Jaynes gave in [1] as of his worked out examples the Bayesian solution to the following problem: "The probability thata certain machine will operate without failure for a time t is, by hypothesis, $e^{-\lambda t}$, $0 \leq t \leq \infty$. We test n units for a time t, and observe r failures; what assurance do we have that the mean life τ exceeds a preassigned value τ_0?"

The Model

By assumption, the probability of a failure exactly at time t_i is

$$p(\tau = t_i) = \lambda e^{-\lambda t_i} dt \tag{12}$$

and the probability of no failure until time s_j is

$$p(\tau \geq s_j) = \int_{s_j}^{\infty} \lambda e^{-\lambda \tau} d\tau = e^{-\lambda s_j} \tag{13}$$

Note that in Jaynes' problem definition all the $s_j = t$, for $j = 1, \ldots, n-r$.

The Likelihood, Prior, and Posterior

We observe a sequence of r failures having observed failure times t_i, for $i = 1, \ldots, r$, and $n - r$ non-failures having observed failure-free times s_j, for $j = 1, \ldots, n - r$. From (12) and (13), it follows that the probability of r failures and $n - r$ non-failures at the observed times, or, equivalently, the likelihood of the unknown parameter λ, is

$$p(D|\lambda) = \prod_{i=1}^{r} \lambda e^{-\lambda t_i} dt \prod_{j=1}^{n-r} e^{-\lambda s_j} \tag{14}$$

As a prior for the failure rate λ, Jaynes proposes two priors. First the "ridiculously pessimistic" prior

$$p(\lambda|I) \propto \text{constant} \tag{15}$$

which, through a change of variable to the failure time $\tau = \lambda^{-1}$, $d\tau = \left|-\lambda^{-2}\right| d\lambda = \lambda^{-2} d\lambda$, can be seen to correspond with the prior

$$p(\tau|I)\, d\tau \propto \lambda^2 d\tau = \tau^{-2} d\tau \tag{16}$$

Inspecting (16), we can see why Jaynes dubbed (15) to be ridiculously pessimistic. Through the second power in (16), small failure times are overly probable, relative to the standard uninformative Jeffreys' prior for τ:

$$p(\tau|I) \propto \tau^{-1} \tag{17}$$

Note, as an aside, that this uninformative prior (17) would have followed automatically, had we taken for λ the equally uninformative (Jeffreys') prior:

$$p(\lambda|I) \propto \lambda^{-1} \tag{18}$$

Such is the internal consistency of the Jeffreys' prior. Uninformativeness regarding λ automatically implies uninformativeness for its transformation $\tau = \lambda^{-1}$. However, Jaynes takes as his second prior not (18). Rather he insteadgoes for the "reasonable prior":

$$p(\lambda|I^*) = t^* e^{-\lambda t^*} \tag{19}$$

where t^* is the prior expected mean life of the units under consideration.

The rational for this prior is as follows, [1]: "In 'real life' we usually have excellent grounds based on previous experience and theoretical analyses, for predicting the general order of magnitude of the lifetime in advance of the test. It would be inconsistent from the standpoint of inductive logic, and wasteful economically, for us to fail to take this information into account. Suppose that initially, we have grounds for expecting a mean life of the order t^*; or a failure rate of about $\lambda^* = (t^*)^{-1}$. However the prior information does not justify our being to dogmatic about it; to assign a prior centered sharply about λ^* would be to assert so much prior information that we scarcely need a test. Thus, we should assign a prior that, while incorporating the number t^*, is till as 'spread out' as possible in some sense. Using the criterion of maximum entropy, we choose that prior density $p(\lambda)$ which, while yielding an expectation equal to λ^*, maximizes the 'measure of ignorance' $H = -\int p(\lambda) \log p(\lambda)\, d\lambda$. The solution is: $p(\lambda) = t^* e^{-\lambda t^*}$."

Combining the likelihood (14) with either prior (15) or prior (19), the posterior for λ is found to be

$$p(\lambda|D,I) = T \frac{(\lambda T)^r}{r!} e^{-\lambda T} \tag{20}$$

For the ridiculously pessimistic prior (15) we have that T is defined as

$$T = \sum_{i=1}^{r} t_i + \sum_{j=1}^{n-r} s_j \tag{21}$$

the actual observed total unit-time of failure free operation. Whereas for the reasonable prior (19) we have that T is defined as

$$T = \sum_{i=1}^{r} t_i + \sum_{j=1}^{n-r} s_j + t^* \tag{22}$$

the observed plus prior expected total unit-time of failure free operation.

Jaynes' Solution of the Problem

We quote Jaynes: "...we note that if λ were known, then by our original hypothesis [in the problem statement] the probability that the lifetime τ of a given unit is at least τ_0, is

$$p(\tau \geq \tau_0 | \lambda) = e^{-\lambda \tau_0} \tag{23}$$

"The probability that $\tau \geq \tau_0$, conditional on the evidence of the test, is therefore

$$p(\tau \geq \tau_0 | , D, I) = \int_0^\infty e^{-\lambda \tau_0} p(\lambda | D, I)\, d\lambda = \left(\frac{T}{T + \tau_0} \right)^{r+1} \tag{24}$$

"...a result which is simple, sensible, and as far as I can see, utterly beyond the reach of orthodox statistics."

Now, the idea for the Jacobian transformation, or, in this case, the change of variable, procedure was directly inspired by (24). Looking at this equation it was felt that the probability $p(\tau \geq \tau_0 | , D, I)$ had the form of the expectation value $E\left(e^{-\lambda \tau_0}\right)$. This then begged the question if there also was a variance $\mathrm{var}\left(e^{-\lambda \tau_0}\right)$. Having established that this was indeed the case, it followed automatically that $\theta = e^{-\lambda \tau_0}$ should admit its own probability distribution. Once this was realized, it was just a small step to find the explicit distribution of θ by way of a change of variable.

The Change of Variable Solution

The probability of interest is, (23):

$$\theta = e^{-\lambda \tau_0} \tag{25}$$

In order to find the explicit beta-like posterior distribution of θ we make the following change of variable

$$\lambda = -\frac{\log \theta}{\tau_0}, \quad d\lambda = \left| -\frac{1}{\theta \tau_0} \right| d\theta = \frac{1}{\theta \tau_0} d\theta \tag{26}$$

Substituting (26) in (20), we find

$$p(\theta | \tau_0, D, I) = \left(\frac{T}{\tau_0} \right)^{r+1} \frac{(-\log \theta)^r}{r!} \theta^{(T - \tau_0)/\tau_0} \tag{27}$$

It may be checked that the mean of (27) is (24)

$$E(\theta) = \int_0^1 \theta\, p(\theta | \tau_0, D, I)\, d\theta = \left(\frac{T}{T + \tau_0} \right)^{r+1} \tag{28}$$

Now, seeing that Jaynes himself, the modern father of all things Bayesian, stopped at (24), instead of forging ahead to (27), seems to us an indication that the whole change of variable argument is not that trivial. It is not earth shattering either. It is just a pointer to the usefulness of transformations when we wish to determine the beta-like posteriors of probabilities $\theta(\{\phi\})$, which are a function of a set of unknown parameters $\{\phi\}$ for which we have some posterior distribution, $p(\{\phi\}|D,I)$.

Some Closing Thoughts

Now, if we try the change of variable procedure on a Poisson probability of observing m events in a given period,

$$\theta = p(m|\lambda) = \frac{\lambda^m}{m!}e^{-\lambda} \tag{29}$$

then we will find that no change of variable can be made, as λ cannot be rewritten as a closed expression of θ. However, what we can do, if we have some posterior $p(\lambda|D,I)$, is compute the first four moments of (29) and substitute these moments into a maximum entropy distribution, by way of the Jondeau algorithm, [4] and [5]; thus,approximating the intractable change of variable distribution.

REFERENCES

1. E.T. Jaynes, *Confidence intervals vs Bayesian intervals*, in W.L. Harper & C.A. Hooker, eds., Foundations of Probability Theorey, Statistical Inference, and Statistical Theories of Science, 1976.
2. H.R.N. van Erp and P.H.A.J.M. van Gelder, *Deriving a Beta-Like Distribution for Reliability Problems*, Proceedings of IPW (International Probabilistic Workshop), Editors: Van Gelder, Gucma, and Proske, Szczeecin, Poland, 2010.
3. H.R.N. van Erp and P.H.A.J.M. van Gelder, *Generalizing the Beta-Like Distribution*, Proceedings ofIPW (International Probabilistic Workshop) Editors: Van Gelder, Gucma, and Proske, Szczeecin, Braunschweig, 2011.