MODELING LANDING DECISIONS USING HIERARCHICAL LOGISTIC REGRESSION

HENGRUI LUO

1. Introduction

The subject I want to investigate are the factors that affects the landing decision of commercial pilots who are landing airplanes during thunderstorms using a existing *observational study* dataset. Various factors can contribute to the response, which is the decision of the pilot whether landing or not in a thunderstorm. The question I primarily consider is:

To what extent does the effect of on-time status on thunderstorm landing decisions vary across airlines, after controlling for other factors related to landing decisions?

2. Methods

2.1. **Exploratory data analysis.** For the exploratory part, we use the *jittered scatterplot matrix* [1] for a first visual exploratory data analysis, in this procedure we get some evidence which can be used for model building procedure. In this step, it is obvious that if we use the normal linear regression we are not taking the non-normality of the data-generating procedure and sample size into our consideration. We must use a generalized linear model in order to model this dataset more accurately.

Therefore *logistic model* is a natural choice for this binary type response. To account both the nature of response and the differences between the blocks arise in the model; a *Bayesian hierarchical model* seems an appropriate for statistical inference because we want to use a *prior distribution* to address the variability in the effect caused by a certain factor `status`. The hierarchical structure can be used to describe block structure.

I use interaction plot to decide whether to include some of the interaction terms in our regression settings, which is a natural method to use for deciding whether there is possibly an interaction between the factors `status` and `order`. From the design of experiment, such an interaction plot indicate that a multiplicative interaction term may be included. But I choose to fit a model without interaction first and then compare the model with multiplicative interaction with this model in discussion. Further, I use other plots like histograms and numerical summaries to aid my model building procedure.

2.2. **Model.** The model I specified for this inference problem is a mixed effect logistic model, which is to say $logit(p_{i,j})$ can be expressed as a linear regression of *treatment variables*[2] [3].

[1] [Weisberg] Chap.1

[2] Again, we emphasize that this dataset is actually from an observational study, but all our treatment variables are carefully chosen to make casual inference.

[3] The EDA supports that we should include an interaction term. Due to the *principle of marginality*, we must include the lower order main effects.
Meanwhile, the interaction plots of `status` × `hours` and `order` × `hours` do not show evidence that we should include these as multiplicative interactions.

FIGURE 2.1. There is a possible interaction between the `status` factor and `order` factor, the slopes in the interaction plot differ significantly.
From the interaction plot of these two factors versus `hours`, we can see that there is not an obvious trend and *a multiplicative form* of interaction term is not reasonable to be included in such a model.

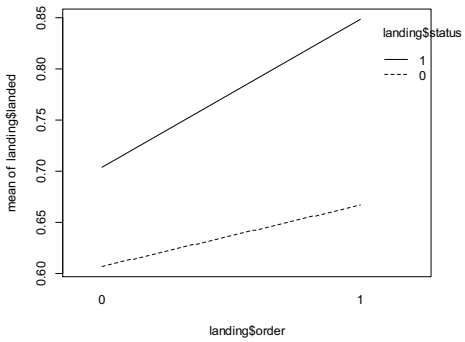


TABLE 1. Definitions of symbols

$logit$	The logistic link function for logistic regression model. $logit(x) \coloneqq log\left(\frac{x}{1-x}\right)$.
$p_{i,j}$	The success probability(or probability of making a landing decision) of i-th pilot in the j-th airline.
$\alpha[j]$	The random effect caused by the `airline` categorical variable of the j-th airline.
μ_α	The average of random effect caused by the `airline`, intercept term. (the only intercept term in this model)
σ^2_α	The variability of random effect caused by the `airline`.
$\beta_{status}[j]$	The random effect caused by the `status` categorical variable, which also depends on the j-th airline.
σ^2_{status}	The variability of random effect caused by the `status` .
β_{order}	The fixed effect caused by the `order` categorical variable, which also depends on the j-th airline.
γ	The fixed effect caused by the `hours` continuous variable.

From the result of exploratory data analysis, I proposed following candidate hierarchical model whose mean function for this logistic regression is:

$$\begin{gathered} logit(p_{i,j})|\alpha, \beta_{status}, \beta_{order}, \gamma = \alpha[j] + \beta_{status}[j] \cdot Status_{i,j} + \beta_{order} \cdot Order_{i,j} + \gamma \cdot t_{i,j} \\ i = 1, \cdots, N_j; j = 1, 2, \cdots, 10 \end{gathered} \tag{2.1}$$

where the notations can be summarized in the Table 1.

The $Status_{i,j}, Order_{i,j}$ are binary variable values for factor `status`, `order` representing the on-time status and whether there is a recent previous landing airplane; $t_{i,j}$ is the continuous variable for factor `hours` representing the experience of the i-th pilot from the j-th corporation.

The first assumption is that given all other parameters$\alpha[j], \beta_{status}[j], \beta_{order}, \gamma$[4], the landing decision made by two individual pilots in the same airline should be *conditionally independent*. Actually the sampling distribution for a given corporation can be assumed as *exchangeable* Bernoulli trials within groups, which is

$$p\left(y_{1,j}, \cdots, y_{N_j,j}|p_{i,j}, N_j\right) = \prod_{i=1}^{N_j} p\left(y_{i,j}|p_{i,j}\right) \tag{2.2}$$

[4] $\alpha, \beta_{status}, \beta_{order}$ are brief form of vectors whose components are exactly $\alpha[j], \beta_{status}[j], \beta_{order}[j]$.

$$p\left(y_{i,j}|p_{i,j}\right) \sim Bernoulli(p_{i,j}) = Binom(1, p_{i,j}), i = 1, \cdots, N_j \tag{2.3}$$

where N_j is the number of pilots in the j-th airline corporation and $y_{i,j}$ represents the i-th pilot in the j-th corporation's landing decision $y_{i,j} = \begin{cases} 1 & landed \\ 0 & otherwise \end{cases}$ where $p_{i,j}$ is the "success probability"(or probability of making a landing decision) of the i-th pilot from the j-th airline with all other factors controlled. Note that we are not simply assuming $y_{i,j}$ are independent but *condition on* $p_{i,j}$. This is reasonable when the individual probability of making a landing decision is given.

The second assumption is that the hyper-parameters are apriori independent. All the $\to \infty$ *does not mean goes to a limit*, they mean that we choose a large number to plug in in the JAGS model. By doing so we assign to each of parameters a weakly informative but proper prior.

$$p\left(\mu_\alpha, \sigma_\alpha; \mu_{status}, \sigma_{status}\right) = \left[p\left(\mu_\alpha\right) p\left(\sigma_\alpha\right)\right] \left[p(\mu_{status}) p(\sigma_{status})\right] \tag{2.4}$$

$$\begin{gathered} p(\mu_\alpha) \sim N\left(0, C^2\right), C^2 \to \infty, p(\sigma_\alpha) \sim Uniform(0, A), A \to \infty \\ p(\mu_{status}) \sim N\left(0, C^2\right), C^2 \to \infty, p(\sigma_{status}) \sim Uniform(0, A), A \to \infty \end{gathered} \tag{2.5}$$

Assuming these coefficients are conditional independent when the hyper-parameters are conditioned on.

$$\begin{gathered} p(\alpha|\mu_\alpha, \sigma_\alpha) = \prod_{j=1}^{10} p\left(\alpha[j]|\sigma_\alpha\right) \\ p\left(\alpha[j]|\mu_\alpha, \sigma_\alpha\right) \sim N(\mu_\alpha, \sigma_\alpha^2), j = 1, 2, \cdots, 10 \end{gathered} \tag{2.6}$$

$$\begin{gathered} p(\beta|\mu_{status}, \sigma_{status}) = \prod_{j=1}^{10} p\left(\beta_{status}[j] \mid \mu_{status}, \sigma_{status}\right) \\ p\left(\beta_{status}[j] \mid \mu_{status}, \sigma_{status}\right) \sim N\left(\mu_{status}, \sigma_{status}^2\right), j = 1, 2, \cdots, 10 \end{gathered} \tag{2.7}$$

$$p\left(\beta_{order}\right) \sim N\left(0, C^2\right), C^2 \to \infty \quad p\left(\beta_{SO}\right) \sim N\left(0, C^2\right), C^2 \to \infty \quad p\left(\gamma\right) \sim N\left(0, C^2\right), C^2 \to \infty \tag{2.8}$$

These choices on the priors are verified [5] to lead to a proper posterior an hence can be used for Gibbs sampling. This model turns out to be a *varying intercepts and slopes model* mentioned by [Gelman et.al], pp384-390. However, we do not include the correlation parameter ρ for this model[6] .

The model we constructed above is a ***hierarchical model*** since the $p_{i,j}$ linearly depends on the parameters which depends on the hyper-parameters which is further distributed as non-informative prior.

3. Results

I run JAGS 4.2.0 with the interface provided by `rjags` library in `R` 3.2.3 to simulate my model proposed above[7]. On each of the trace plots[8], the trace of Gibbs sampler's trace ends up with a repetitive pattern yet no obvious trend occur. The trace oscillates around a certain value, which is a *sign of convergence* of the algorithm. Combined with the theoretic result that the priors are proper, we can trust our result of simulation and based

[5][Gelman et.al] Chap.5, Sec 5.4

[6]There is a practical reason why I did not choose another prior, the scaled Wishart, proposed by [Gelman et.al] p.390. Neither JAGS and R have a good realized sampler from a scaled Wishart distribution and my own experiments showed that when sampling from a scaled Wishart, there is a numerical problem when the ratio of scale vector components go too large.

[7]See Appendix for all of codes and outputs

[8]See Appendix for all of trace plots and numerical summaries

FIGURE 3.1. Part of trace plots of the simulation

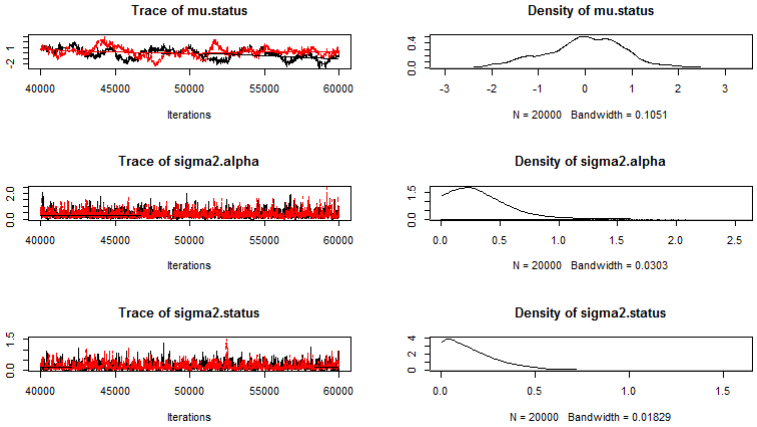
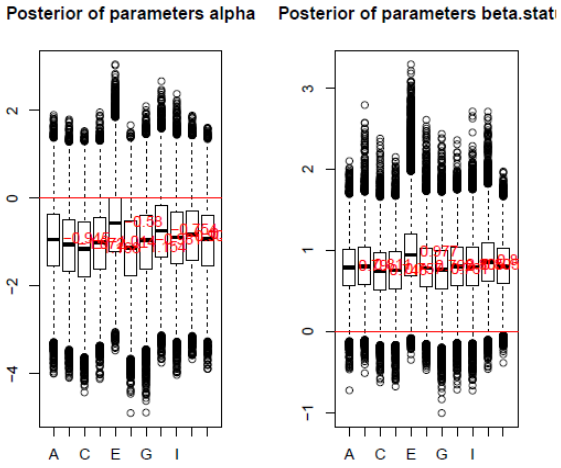


FIGURE 3.2. The boxplots of $\alpha[j]$ and $\beta_{status}[j]$ from simulated samples, no need to standardize the data because they are of the same factor scale.



our inference on them. Based on the samples obtained from Gibbs sampler, we can plot the sampling density distribution. The *simulated posterior mean* is a reliable estimator of corresponding parameters.

For the varying coefficient of slope $\beta_{status}[j]$, the boxplots shows that the `status` have greatest influence on airline E's pilots compared to other airlines. But except for airline E, there seems little difference between the effects of `status` on probability of landing decision making.

4. Discussion

4.1. **Conclusion.** From the density plots we can see that the *posterior means* can serve as a good estimate for corresponding parameters.

The posterior $\mu_{status} = 0.80$. This means that if the plane is behind the schedule and all other factors are controlled, then the probability of making a landing decision will increase by $logit^{-1}(0.80) - 0.5 \approx 19\%$ on average regardless factor `airline`.

The posterior $\sigma_{status} = 0.197$. This means that the variation of $\beta_{status}[j]$ will be about 0.19 (units) across different airlines. This variability caused that if the plane is delayed, the probability of making landing decision will

TABLE 2. Comparison of DIC results from two models' simulations

	Model without interaction	Model with interaction
Mean deviance	432.9	433.5
penalty	9.393	10.66
Penalized deviance	442.3	444.1

increase by $\Delta, \Delta \in \left[logit^{-1}(0.74) - 0.5, logit^{-1}(0.98) - 0.5\right] \approx [17\%, 22\%]$. Considering that average increase is 19%, we cannot say that the effect of on-time status varies *much* across different airlines.

There is some differences in the effect of on-time status (variable `status`) on landing probability across the airlines. This also means that pilots from different airlines perceive the influence of `status` factor differently, however, this effect of$\beta_{status}[j]$ does not vary as much as the intercept term $\alpha[j]$

Therefore I can say that after controlling for other factors related to landing decision, there is some variability of the effect of on-time status $\beta_{status}[j]$ across airlines. But such a variability is small compared to the variability of the general effect caused by corporation culture $\alpha[j]$.

One thing to be noted is that all simulated posterior distributions are *heavy-tailed*, partially because the relatively small sample size in each airline. In such a relative sparse case, we cannot draw a very strong conclusion about a certain factor's effect.

4.2. **Model Comparison.** As the interaction plot shown, it seems that we should include an interaction term accounting for `status` $\times$ `order` effect.

I suspect that there is a *colinearity* (or *aliased effect*) between factor `order` and `status` $\times$ `order`. The term β_{SO} denotes the fixed effect caused by the `status` $\times$ `order` interaction. Then I dropped the interaction term in our model and refitted the Bayesian model with logistic mean function

(4.1)

$$logit(p_{i,j})|\alpha, \beta_{status}, \beta_{order}, \beta_{SO}, \gamma = \alpha[j] + \beta_{status}[j] \cdot Status_{i,j} + \beta_{order} \cdot Order_{i,j} + \beta_{SO} \cdot Status_{i,j} \cdot Order_{i,j} + \gamma \cdot t_{i,j}$$
$$i = 1, \cdots, N_j; j = 1, 2, \cdots, 10$$

Then I used the DIC criterion (Penalized deviance) from [Spiegelhalter et.al] to evaluate these two fitted Bayesian logistic models.

Both models differ little in terms of DIC criterion, yet the model without interaction term performs a bit better. This may indicate that multiplicative interaction is not an appropriate term to be included.

However, the more serious consequence arises after including the `status` $\times$ `order` interaction, the posterior coefficients $\beta_{status}[j]$ all become negative and hence hard to interpret and go against our exploratory result.

Therefore I decide that we should still use our ***model without interaction term*** in order to get a more understandable fitted model.

5. Appendix

References

[Gelman et.al] Gelman, Andrew, et al. Bayesian data analysis. 3ed. USA: Chapman & Hall/CRC, 2014.

[Genkin et.al] Genkin, Alexander, David D. Lewis, and David Madigan. "Large-scale Bayesian logistic regression for text categorization." Technometrics 49.3 (2007): 291-304.

[Wong et.al] Wong, George Y., and William M. Mason. "The hierarchical logistic regression model for multilevel analysis." Journal of the American Statistical Association 80.391 (1985): 513-524.

[Weisberg] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons.

[Spiegelhalter et.al] Spiegelhalter, David J., et al. "Bayesian measures of model complexity and fit." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64.4 (2002): 583-639.

[Park&Casella] Park, Trevor, and George Casella. "The Bayesian lasso." Journal of the American Statistical Association 103.482 (2008): 681-686.