Assignment for DAT246, E.Software Engineering

Olof Magnusson, olofmagn@chalmers.se

August 11, 2025

Description of the data

The data frame consists of 140 observations of 4 variable e.g. . *(subject.Category.technique.tp: chr "1;LE;NT;5" "1;LE;OT;6" "2;LE;NT;3" "2;LE;OT;3" ....)* indicating a trust positive (tp) score for each subject classified as experienced or inexperienced (ME,LE), using either a new technique (nt) or old technique (ot). The techniques are classified into three categories: nt, ot and $0T$. We consider $0T$ as a typo and something that we need to change before creating our model. We could standardize true positives (tp), making it easier to traverse the vector space. However, the problem is that we will impact the outcome of the generation process of the model, including data, density, and scale. Therefore, we avoid that since we want to keep the values as original since we are dealing with counts.

We assume that tp increases because it is a new technique, and we code the MCategorical variables where 0 is less experience and 1 is more experience

Create our models

With more knowledge about the data structure, we can start to think about what type of golem we want to create and which approach is most suitable (ontological or epistemological) given the data and task. The model's outcome/effectiveness should be identified by measuring through tp, i.e., the number of found faults classified as true faults using any technique combined with experience. If we want to stick with ontological reasoning, a Poisson(λ) distribution seems to be the most viable option - since TP consists of natural numbers. We want to discover how technique and experience together affect the result. Therefore, it is natural that we use them as predictor variables in our generalized linear model. Hence, our mathematical model can be defined as:

The Poisson expects that the mean and the variance is equal and that the variance is constant.

If we use any likelihood which is not Gaussian, we need to be careful about how to translate the output from the linear regression to the outcome space. Linear regression only thinks about probability, and the outcome space is, in this case, a count.

$$
\begin{aligned}
\mathrm{t}_p &\sim \text{Poission}(\lambda) \\
\log(\lambda) &= \alpha + \beta_T T_i + \beta_C C_i \\
\alpha &\sim \text{Normal}(2, 0.5) \\
\beta_T &\sim \text{Normal}(1, 0.5) \\
\beta_C &\sim \text{Normal}(1, 0.5)
\end{aligned}
$$

Note that we have a wider prior on α, grand mean, and tighter prior on B_t and B_c, that is because they deviate from α and we do not expect them to produce overall different values for tp. We expect small and positive values for tp. We give it some room, the dispersion, to get comprehensive coverage of this expectation since

Ockham's razor: Models with fewer parameters tend to underfit, and we can be more okay with that than overfitting. Adding more dimensions by adding more parameters and priors, which adds more complexity in our models

we know that $log(\lambda) = \alpha$. We will build the models incrementally and make them more complex in each step to see if they make sense to use. For example, one model could be that we only use a technique as a prior B_T, or only categorical B_C as a prior. Another model could mix both technique and categorical B_T, B_C variables to see how they behave together. However, a model could also be null to know the difference without using any predictors. To simplify, we name the models as follows: *MTech*, which uses only technique as predictor β_T and *MCat*, which uses only category β_C as a predictor. *MTechAndCat* uses both technique and category as predictors. *MNull* is the null model focusing on the impact of having no predictors. Finally, we will evaluate the models using information criteria to measure the distance these models have relative to each other and from an out-of-sample prediction.

Dag

As mentioned in *Create our models*, we build different statistical models to understand the causality between the variables. We state them in following way: (1). Tech directly influences TP. (2) Exp directly influences TP. Experience and Technique are totally independent since we are talking about the use of a technique. Exp -> TP and Tech -> TP. Therefore, the experiment is biased, since they don't choose the technique by themselves and the design of the experiment is 2x2.

Discussion

The result from the out-of-sample prediction using WAIC can be shown in Table 1. MTech is the model that performs best from all the models from an epistemological point of view. The experience *MCat*

The iteration is set to 5^{e3} from which we gain a Rhat value of 1 on most of the models. Lowering this will led to ineffective sampling. We do not give MCMC enough time to explore the space.

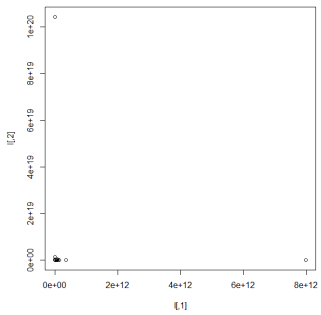
	WAIC	SE	dWAIC	dSE	PWAIC	weight
MTech	590.5	12.63	109.9	11.54	1.6	0
MTechAndCat	591.0	12.45	110.4	11.51	2.3	0
MNull	591.6	13.28	111.0	12.15	0.9	0
MCat	591.8	13.02	111.2	12.01	1.5	0

Table 1: Result of running function `compare` with the models

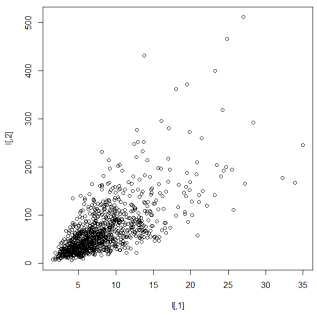
model performs actually worse than the Null model. Technique and experience performance worse than technique alone but it is penalized for using more parameters. *MNull* model and *MTechAndCat* are very close from an out-of-sample prediction. But in reality, the *TechAndCat* performs better. We stick with the most logic model, *MTechandExp*, since it explains things well enough to answer the

question of the assignment.

The figure (a) (b) below shows the difference with choosing non-informative priors or vague priors, such as $\alpha \sim dnorm(0, 10)$, $\beta_t \sim dnorm(0, 10)$ and $\beta_c \sim dnorm(0, 10)$ where MCMC struggle with finding good values in the space. Compared to $\alpha \sim dnorm(2, 0.5)$, $\beta_t \sim dnorm(1, 0.5)$ and $\beta_c \sim dnorm(1, 0.5)$. We do not want to delimit the data too much by setting even lower priors, we want it to speak its language. The figure (b) shows the final model where we get good priors with some room of uncertainty. We allow to a small extent values above 11, which is the max TP value, but the probability density majority is in the same region.



(a) Bad priors where MCMC struggle with finding good values in the probable space given the priors. The tail of the distribution can be shown where 8e+12 is extremely wide distribution



(b) Good priors with some room of uncertainty. We shrink, with more tighter priors, the dimensional space where the MCMC should search.

This shows the importance of doing a prior predictive check before doing anything with the data. Previously, we got much data where $TP \gg 12$ && $TP \ll 0$, which does not make sense regarding the data-set. Now, we only have some anomalies, and we can be okay with that. Since when we are using the priors on an actual model, the chance is lower than we will overfit. The Rhat is < 1.01, the variance within each of the chains compared with the variance between each of these chains. This indicates that the chains have reached a stationary posterior distribution. The neff is around 15% above the sample size, and since we are interested in posterior means/medians and not specific quantiles of the posterior, we are okay with it. By using `traceplot` we get a caterpillar hair which indicates that the four chain has converged in a good way.

Conclusion

Less experience has more influence than experience in the sample distribution, which can be visualized in the table below. This is because the algorithm will go through every sample one at a time, and with more iterations that contain less experience, the lower *bc* will be.

	mean	sd	5.5%	94.5%	n_eff	Rhat4
a	1.43	0.08	1.30	1.55	4166	1
bt	0.14	0.08	0.01	0.26	4638	1
bc	-0.10	0.08	-0.03	0.23	4338	1

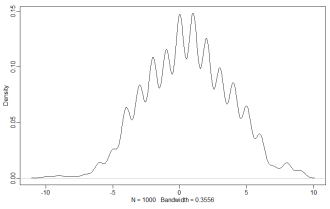
From the experiment result, we can see that new technique and less experience is better 517 of 1000 compared to old technique and less experience.

There is a problem with validating the result of the model due to unbalanced sampling. The LE experience population is more significant than ME experienced. To increase the validity of the experiment, we should introduce partial pooling and hyper predictors. This would increase the validity of the experiment. This could be considered the next step in the experiment.

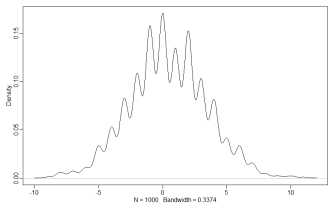
-1	0	1
343	140	517

We also see similarities where new technique and more experience is better 510 of 1000 compared to old technique and more experience. LE influences technique more than ME since the sample contains more LE samples than ME. This experiment concludes that new technique always performs better than an old technique no matter the experience.

-1	0	1
370	120	510



(a) Difference between new technique and less experience and old technique and less experience



(b) Difference between new technique and more experience and old technique and more experience