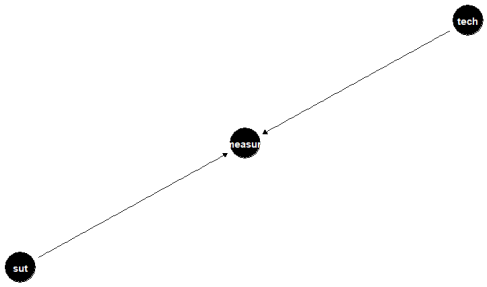
Assignment for DAT246

Lakshmi Sri Rupa Kurukuri, 950328-4888

November 10, 2023



The DAG, confounders, analysis:

The directed acyclic graph (DAG) illustrates how 'tech,' 'sut,' and 'fmeasure' are connected in the model. We also visualized the DAG representing the relationships between the variables in the Generalized Linear Model with Exponential Likelihood. In the context of Generalized Linear Model with Gamma Likelihood, the predictors 'tech' and 'sut' influence the model parameters in a linear fashion, like the DAG for the Exponential Likelihood.

The DAG for Generalized Linear Models (Exponential and Gamma Likelihoods) helps us assess unknown confounders. These unseen variables can bias our inferences. Conducting sensitivity analyses is crucial to test the robustness of our findings when dealing with these confounders. While our DAGs illustrate assumed causal relationships, recognizing unknown confounders is vital as a significant analysis limitation.

Description of the data:

Initial observations revealed specific dataset patterns with six distinct "tech" values and 12 unique "sut" values. The dataset consists of 72 observations, and the "fmeasure" variable, indicating software performance, ranges from 299.8 to 3491.9 with a median of 1093.0, slightly positively skewed.

Descriptive statistics indicate that 25th percentile "fmeasure" is 640.8, and the 75th percentile is 1737.0, offering valuable insights into data distribution. These statistics serve as a foundation for subsequent modelling decisions.

Despite the wide range of "fmeasure," there are no extreme outliers. Given the Generalized Linear Models' ability to handle varying data scales effectively, standardization or normalization may not be necessary.

Likelihood defense

Generalized Linear Model with Exponential Likelihood (m1):

We have chosen the Exponential Likelihood for modelling 'fmeasure' because it aligns with our ontological understanding that 'fmeasure' mirrors an exponential distribution due to its inherent properties in the data-generating process, showing an exponential decay pattern over time. This choice is further supported by epistemological reasoning as Exponential distributions conform to established statistical principles for modelling continuous variables with positive values and right-skewed distributions.

$$
\begin{aligned}
y_i &\sim \text{Exponential}(\lambda) \\
\log(\lambda_i) &= \alpha tech_i + \beta sut_i \\
\alpha &\sim \text{Normal}(10, 0.5) \\
\beta &\sim \text{Normal}(10, 0.5)
\end{aligned}
$$

Generalized Linear Model with Gamma Likelihood (m2):

From an ontological perspective, it aligns with the nature of 'fmeasure,' which is expected to exhibit a right-skewed distribution, mirroring the inherent characteristics of the response variable in the data generating process. Epistemologically, our decision to use the Gamma Likelihood is in harmony with established statistical knowledge, as the gamma distribution is well-suited for modelling non-negative, right-skewed continuous data.

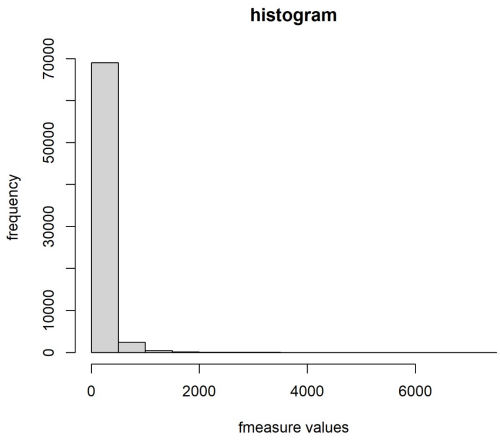
$$
\begin{aligned}
y_i &\sim \text{gamma}(\lambda, k) \\
\log(\lambda_i) &= \alpha tech_i + \beta sut_i \\
\alpha &\sim \text{Normal}(100, 10) \\
\beta &\sim \text{Normal}(100, 10) \\
k &\sim \text{exponential}(1)
\end{aligned}
$$

In summary, employing both Generalized Linear Models with Exponential and Gamma Likelihoods reflects a thoughtful and data-centric approach. It aligns with principles of objectivity, interpretability, and robustness, ensuring our models suit the given data's unique characteristics.

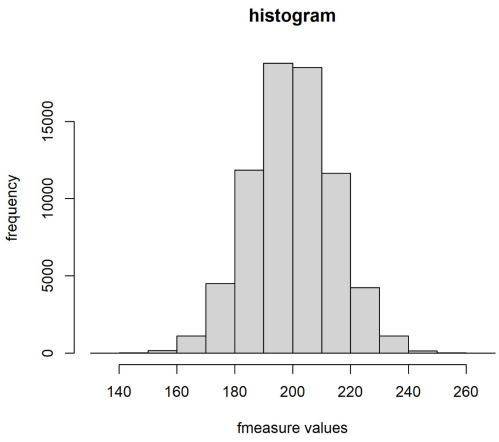
Prior defense with prior predictive check

Prior Predictive Plot for Exponential Likelihood:

In this analysis, we employed an Exponential Likelihood model to predict the response variable 'fmeasure' based on the predictor variables 'tech' and 'sut'. To inform our model, we incorporated the following priors: Normal (10, 0.5). These priors have been chosen to capture the characteristics of the data and guide the modelling process effectively. While the priors may seem wide, they are still within the plausible range of the outcome space, allowing for meaningful inferences.



Prior Predictive Plot for Gamma Likelihood:



In this analysis, we employed an Gamma Likelihood model to predict the response variable 'fmeasure' based on the predictor variables 'tech' and 'sut'. To inform our model, we incorporated priors as follows Normal (100, 10). The choice of these priors is made to capture the characteristics of the data and guide the modelling process effectively.

Defense of Selected Priors:

Our analysis involved defining priors and conducting prior predictive checks, affirming the validity of our prior choices. The selected priors aim for a balance between objectivity and data-driven inference, as altering priors significantly impacts model outcomes.

In our pursuit of the most effective technique, lower fmeasure values indicate better performance, leading us to favour the 'final model' with Exponential Likelihood.

Model comparison

Discussion of 'Final Model' Results with ulam and Comparative Analysis:

The results of our 'final model,' Generalized Linear Model with Exponential Likelihood using ulam(), provide a strong endorsement for this approach. This model effectively captures the relationship between the response variable, fmeasure, and its predictors, tech and sut, within a dataset of 72 observations. These results affirm the model's suitability for explaining fmeasure variability when utilizing ulam().

```
Compare (m1,m2, func = "LOO")
```

```
     PSIS    SE dPSIS   dSE pPSIS weight
m1 1175.8 12.51   0.0    NA   3.5      1
m2 1581.5 91.13 405.6 80.78   7.3      0
```

In our comparative analysis, we evaluated the performance and model complexity of Generalized Linear Models using different likelihood functions, specifically Exponential and Gamma Likelihoods. To assess their predictive capabilities and model complexity, we utilized LOO (Leave-One-Out Cross-Validation) and the outcomes are:

- Generalized Linear Model with Exponential Likelihood exhibits a slight advantage in predictive accuracy when compared to the Gamma Likelihood model.
- It also presents a simpler model, characterized by a more favourable looic (Leave-One-Out Information Criterion).

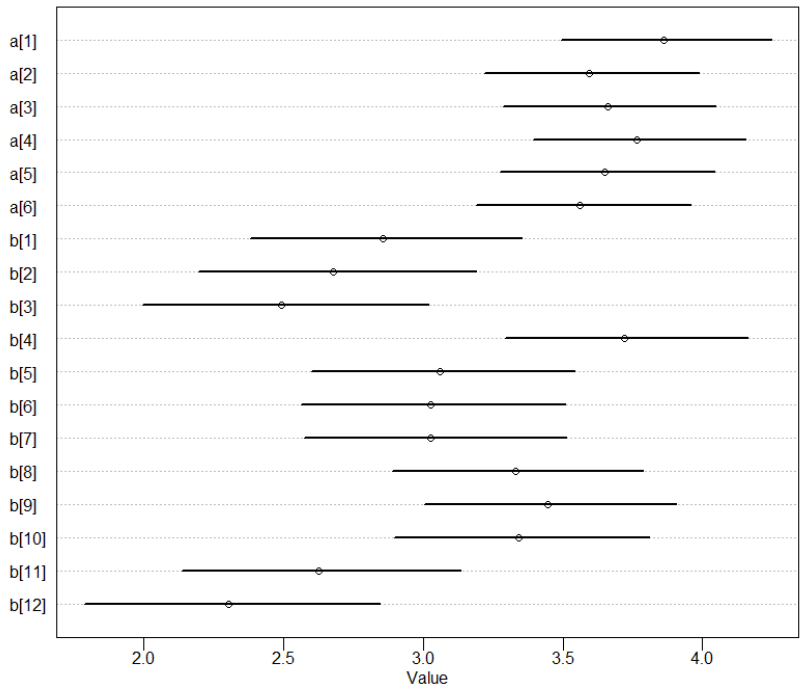
Model diagnostics:

```
precis( m1, 2 )
```

The *Rhat* value is equal to 1, indicating that the chains are converging. By examining the *n_eff* value which exceeded 2000 obs. confirming the efficiency of the sampling process.

Practical interpretation of results:

```
precis( m1, 2 )
plot(precis( m1 , 2 ))
```



Practical considerations highlight Technique 6 is superior performance and the Exponential Likelihood model's crucial role in enhancing predictive accuracy.

The choice of System Under Test (SUT) significantly influences model results, the Exponential Likelihood model is considered better for modelling SUT-influenced data. It is noted that the Gamma Likelihood model allows more variability in parameter estimates, which may introduce uncertainty or instability, whereas the Exponential Likelihood model is comparatively more rigid, potentially resulting in more reliable and consistent results.

	mean	sd	5.5%	94.5%	n_eff	Rhat4
a[1]	3.86	0.23	3.49	4.24	19802	1
a[2]	3.60	0.24	3.22	3.99	18163	1
a[3]	3.66	0.24	3.29	4.05	17829	1
a[4]	3.77	0.23	3.40	4.15	17546	1
a[5]	3.65	0.24	3.28	4.04	17385	1
a[6]	3.56	0.24	3.18	3.95	17293	1
b[1]	2.85	0.30	2.38	3.35	20410	1
b[2]	2.68	0.31	2.20	3.19	19937	1
b[3]	2.49	0.32	2.00	3.02	20095	1
b[4]	3.72	0.27	3.30	4.16	18483	1
b[5]	3.06	0.29	2.60	3.54	20690	1
b[6]	3.03	0.30	2.56	3.51	20208	1
b[7]	3.03	0.30	2.57	3.51	20479	1
b[8]	3.33	0.28	2.89	3.79	19272	1
b[9]	3.45	0.28	3.00	3.91	18987	1
b[10]	3.34	0.29	2.90	3.81	18229	1
b[11]	2.63	0.31	2.14	3.14	19994	1
b[12]	2.31	0.33	1.79	2.85	20344	1