Final Assignment: Predicting Director Compensation

Jasper Ginn (s6100848)

April 8, 2019

Introduction

Problem statement, literature & hypotheses

Differences in Bayesian and Frequentist inference

https://en.wikipedia.org/wiki/Probability_interpretations

- Interpretation of random variables ==> Bayesian: data fixed but parameters random (what are implication?)

Definition of probability and the source of uncertainty

For a large part, differences between Frequentist and Bayesian inference stems from their respective view of uncertainty and how this is captured by probabilities [CITE]. In the Frequentist framework, probabilities are looked upon as the limiting value of the number of k successes in a sequence of n trials [CITE/CHANGE/DIRECT QUOTE FROM STACKEX], or:

$$p = \lim_{n \to \infty} \frac{k}{n} \tag{1}$$

The implications of this definition are that (1) probabilities make sense only in the context of infinite trials, and (2) the probability is *fixed* in the population and our uncertainty about it reduces as we repeatedly take more (or larger) samples. In particular, this definition implies that the only source of randomness by which our estimate $\hat{p}$ differs from the true value p comes from the data [CITE/BACK UP], which may differ from sample to sample due to, for example, sampling error.

The Bayesian framework looks upon probabilities as a means of quantifying uncertainty about knowledge [CHECK]. Even though the 'true' parameter value may be fixed, we are limited by our knowledge of this value. Hence, the uncertainty by which we make statements about the world changes as we collect more information which is represented by the posterior distribution, which may be looked upon very loosely as the collection 'true' values conditional on how certain we are about the veracity of this knowledge.

Epistemological differences about what constitutes knowledge

In Frequentist statistical inference, evidence may originate only from the data. This is not the case in Bayesian statistical inference, where inferences are based on a mix of domain expertise (*prior* or *belief*) and evidence from the data. This makes perfect sense in the Bayesian framework; if we are uncertain about our knowledge then we can constrain the parameter space by injecting what we *do* know. Hence, a Bayesian looks upon prior beliefs as just another source of knowledge that has been translated into a probability density. The exact definition of this prior, for example through the mean and variance of this prior distribution, affects the results to the degree to which we place stock in its veracity. This is illustrated in the figure below: a strong belief in our prior knowledge (represented by a small variance) leads to a more constrained posterior.

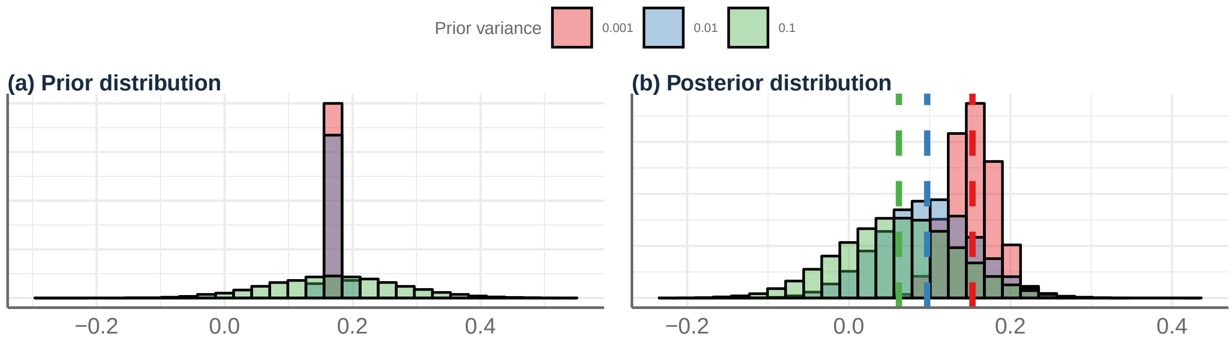

Figure 1: Effect of adjusting the prior variance of the coefficient for variable Male (a) on the posterior distribution (b). The mean is set at M=.17 which represents a 17 percent increase in compensation for males versus females. Reducing the prior variance represents increased certainty about the estimate of our domain knowledge and weights the information from the data less strongly compared to domain knowledge. The dashed lines indicate the value of the posterior means.

The claim that the Frequentist approach is more objective would hold only if the data collection process, (pre-)processing steps and analysis are guaranteed to be objective. This is a tenuous assumption at best in the social sciences (how else would we get by that memorable phrase 'lies, damned lies and statistics'?); the way in which we collect data is fraught with subjective decisions during data collection, manipulation and analysis [CITE]. Data, in and of itself, is not objective [CITE], and the critique leveled at Bayesians boils down to the practice of incorporating domain knowledge *explicitly* through the use of a prior.

Methods of estimation and hypothesis testing

Given that, for a Frequentist, the data are a random variable and the source of variation, it makes sense to optimize the 'likelihood' of the data conditional on the parameters. That is, we find the *most likely combination* of parameters $\hat{\theta}$ that explain the data and that provide consistent and asymptotically unbiased estimators, or:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \text{Log Likelihood}(\theta|\text{data}) \tag{1}$$

Conversely, given that a Bayesian thinks of the data as fixed and the parameters as random variables, they are interested in finding the distribution of the parameters and hence the source of variance:

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})} \tag{2}$$

The question of how the definition of probability and the methods of estimation impact inference under these frameworks is illustrated by the difference in interpretation of the confidence interval and the credible interval. When we calculate the confidence interval [REFER TO CENTRAL LIMIT THEOREM/FREQUENCY OF SAMPLING MEANS], the boundaries of the confidence intervals are interpreted as random variables due to sampling error since they estimate the frequency of sampling means. With a credible interval, we do not have this source of variability, which gives rise to the definition that the parameter is contained in the credible interval with some probability because the parameter space is assumed to be known under the assumptions by which we arrived at the posterior distribution [CHECK].

Further implications are to be found in the way we test hypotheses in these frameworks. In the Frequentist framework, we usually partition parameter space into an acceptance and a rejection region based on some null and alternative hypothesis. On the basis of a test statistic, computed from the data, we then decide whether or not the result we observe is likely to occur due to chance.

- Type I && II
- Test statistic
- probability of a hypothesis given the data

This is not so in the Bayesian Framework. Given that the data or the evidence can be viewed as a means to update a prior belief, hypothesis testing focuses on the degree to which the evidence found in the data supports this initial belief.

–> Probability of the null hypothesis (p 161, 162 Berger & Benny). "Flipping" of the hypothesis

- Endless transformations on posterior distribution ==> not possible in frequentist except when calculating mean/variances ==> see freq cartoon book

- Frequentists obscure the interpretation process, Bayesian obscure the posterior distribution collection process

Frequentist statistics obscures the interpretation of critical statistics to the point where students learn the heuristics ('p-value is significant') before truly understanding what that heuristic means. Conversely, Bayesian statistics provides an explicit description of a model and its assumptions and has an intuitive interpretation of statistical results. However, the approach obfuscates the estimation method by using the arcane process of Markov Chain Monte Carlo (MCMC) sampling.

The effect of changing the prior variance can be summarized using a posterior shrinking factor (CITE BETANCOURT) and posterior z-score. The shrinking factor shows us the factor by which the variance of the posterior distribution shrinks or expands compared to the prior variance. The posterior z-score tells us the direction and magnitude by which the posterior mean shifts compared to the prior mean; if we are confident in the precision of our domain knowledge (resulting in small prior variance), the resulting posterior weights this information strongly and we end up with a posterior mean that lies closer to the prior mean (represented by the z-score). However, given that we have small variance, and if this is not corroborated by the data, then the shrinkage factor will be large.

If we carry this argument to its logical endpoint, it follows that we would next arrive at a means to compare these posteriors relative to each other. This, in essence, is the Bayes Factor.

Methods & Results

Table XX below shows the descriptive statistics for each variable. The outcome variable **compensation** is given in thousands of Great British pounds and has been log-transformed.

Model 1 is given by the following equation:

$$\log(\widehat{\text{compensation}}_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{male}_i + \epsilon_i \tag{1}$$

Model 2 is given by the following equation:

$$\log(\widehat{\text{compensation}}_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{male}_i + \beta_3 \text{SectorServices}_i + \beta_4 \text{SectorBasicMaterials}_i + \epsilon_i$$

Given that we measure the compensation on a log scale, the coefficients we derive from the model must be interpreted as percentages. This forces us to rethink the priors we set on the model (default is mu=0, sd=1000). Additionally, even if we know very little of the effect of age on compensation (or gender for that matter), we can set some reasonable assumptions in terms of upper and lower bounds. For age, we now assume that, as age increases, compensation increases as well (not a crazy assumption). But, given that we are not certain about the extent to which it increases, we set the standard deviation of this estimate to .1, representing a spread of approximately 10%[1]. For gender, we know from earlier studies that the gender gap is 17%. However, theory suggests that the gender gap is lower or non-existent at top-tier firms. Hence, we set this prior a mean of .05 with a spread of .03, which reflects our uncertainty of the estimate.

Our hypotheses are:

$$\mathrm{H}_1\text{: } \beta_{\mathrm{Male}} \approx 0$$

$$\mathrm{H}_2\text{: } \beta_{\mathrm{Age}} > 0$$

$$\mathrm{H}_u\text{: } \beta_{\mathrm{Male}}, \beta_{\mathrm{Age}}$$

These hypotheses will be included in the models we construct. (Hoitjtink p.24).

BF sensitive to outliers (Hoijtink pp.31-32) ==> check for outliers before BF. Also violation of model assumptions using ppc. We should use a robust BF but we don't have that. Anyway, interpret the BF with a grain of salt.

Bayes' Factor is sensitive to the choice of prior ==> something about the size of complexity. (Hoijtink pp.29-30)

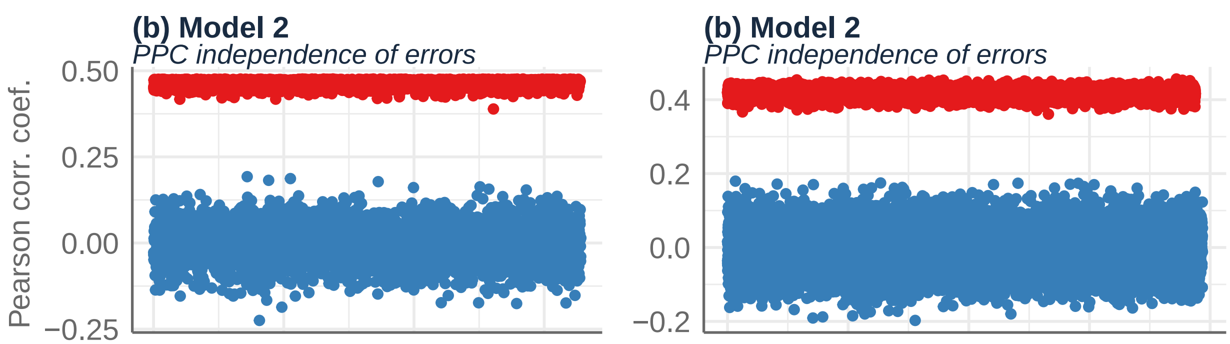

Figure 2: A random subset of observed and simulated correlation coefficients for model 1 (left) and model 2 (right). The figure shows that the observed residuals (in red) are much more correlated than is reasonable under the model (represented by the blue, simulated values).

From figure XX and the posterior predictive checks (table XY), we observe that the data are not independent ($\hat{r}_{\mathrm{PPC,\ Model1}} = .466$). That is, directors who are in the same board tend to are more similar to each other than directors from other boards. This is not unexpected: the data are hierarchical in nature such that individuals are nested in boards. Indeed, the results of the random effects model shows that the intra-class correlation coefficient ρ equals 57%, meaning that the expected correlation of two randomly picked directors from the same company is $\hat{r} = .573$.[2]

The final random effects model is presented in table YY. This model corresponds to the following equation.

[1] For values close to 0, $exp(x) \approx 1 + x$

[2] The results of running all different stages of a multilevel model are presented in another document.

$$\text{compensation}_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} \cdot \text{Age}_{ij} + e_{ij} \tag{3}$$

Where γ_{00} is the overall intercept and u_{0j} is a company-specific error term. Notice that the fit of this model is much better than that of the previous models, indicating that the multilevel approach seems appropriate. The marginal and conditional R-squared values [CITE] are $R^2_{\text{M}} = .0056;\ 95\%\ \text{CCI} = [.0008, .015]$ and $R^2_{\text{C}} = .5398;\ 95\%\ \text{CCI} = [.4807, .6]$ respectively. This indicates that the fixed part of the model (age) explains almost no variation in the data, but the fixed and random parts together explain some 54% of the total variation. Hence, we must conclude that we do not have the right variables at either level 1 or 2 that would help us explain the heterogeneity in comepnsation among directors.

	Dependent variable:		
	Compensation (GBR '000, logged)		
	(1)	(2)	(3)
	Linear	Linear	Linear mixed effects
	(blm)	(blm)	(JAGS)
(a) Fixed			
Constant	4.991 (4.936, 5.046)	4.821 (4.732, 4.910)	4.991 (4.877, 5.110)
SectorBasic Materials		.225 (.079, .370)	
SectorServices		.292 (.169, .414)	
Male	.065 (−.085, .215)	.153 (.095, .209)	
Age	.008 (0.000, .017)	.009 (.001, .017)	.001 (.0035, .0155)
(b) Random			
σ^2_e	.516	.501	.342
σ^2_{u0}			.399
(c) Model Fit			
Observations	336	336	336
Companies			52
DIC	508	488	280
Penalty.	4	6	47
R^2	.021	.098	.006 (M), .536 (C)
BF	.112	37.596	
(d) Post. Pred. Checks			
Normality	.361	.353	
Homoskedasticity	.326	.607	
Independence	0	0	
(e) Bayes' Factors			
H_1	4.5	5.2	
H_2	3.4	6	
H_2	.3	.2	
Note:		Baseline is sector 'Financials' for models (1) and (2)	

Table 1: Model results

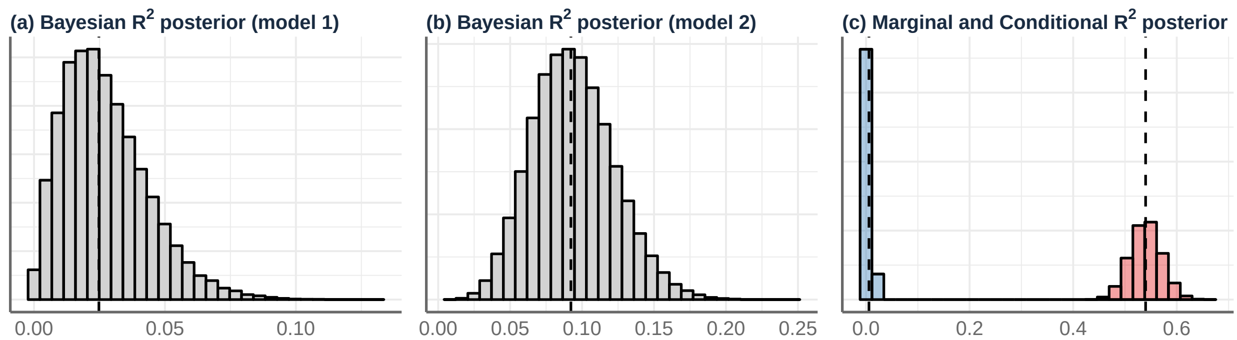

Figure 3: Bayesian R-squared value for model 1 (a) and model 2 (b). The proportion of cases in which the R-squared value of model 2 exceeds that of model 1 is .99. Figure (c) shows the marginal (blue) and conditional (red) R-squared values for the posterior distribution. The marginal R-squared indicates the amount of variance explained by the fixed part of the model; the conditional R-squared indicates the amount of variance explained by the fixed and random part of the model.

Other material

If we desire to interpret this p-value like we do the traditional, Fisherian p-value, then we must assume that, under the assumption that a particular test is not violated, the Bayesian p-value is uniformly distributed. Otherwise, we cannot interpret it as a proportion. Hence, we desire $p_{\text{posterior}} = P(x \leq X)$. We often find that this assumption does not hold true for posterior predictive checks. To this end, we can simulate the assumption when it holds [CHANGE]. This is illustrated in figure XX for the posterior predictive checks included in the R library blm.

Figure 4: Distributions of posterior predictive p-values for 1.000 simulated data sets. In plots (a) and (b), the simulated data are drawn from a normal without any violations of the linear regression assumptions. In plot (c), the assumption of homoskedasticity is violated in each of the simulations. In plot (d), the assumption of independence of errors is violated in each of the simulations. The color indicates the severity of the violation; the green bars indicate mild violation, blue indicates medium violation and red indicates severe violation. The script used to generate the data and run the simulations may be found here.

References

Lynch, S. M. (2007). Introduction to applied Bayesian statistics and estimation for social scientists. Springer Science & Business Media. Chapter 9.2

Aarts, E. (2019). Introduction to multilevel analysis and the basic two-level regression model, week 1 notes [powerpoint presentation]. *Introduction to Multilevel Analysis*, Utrecht University.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 4(2), 133-142.