Final Assignment: Predicting Director Compensation

Jasper Ginn (s6100848)

April 8, 2019

Introduction

Problem statement, literature & hypotheses

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.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Compensation	336	4.991	0.518	2.079	4.769	5.017	5.283	8.294
Age	336	−0.000	6.829	−21.363	−4.363	0.637	4.637	20.637
Male	336	0.000	0.384	−0.821	0.179	0.179	0.179	0.179

Table 1

Model 1 is given by the following equation:

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

Model 2 is given by the following equation:

$$\begin{aligned}\log(\hat{\text{compensation}_i}) = &\beta_0 + \beta_1 \text{age}_i + \beta_2 \text{male}_i + \\ &\beta_3 \text{SectorServices}_i + \beta_4 \text{SectorBasicMaterials}_i + \epsilon_i\end{aligned} \tag{2}$$

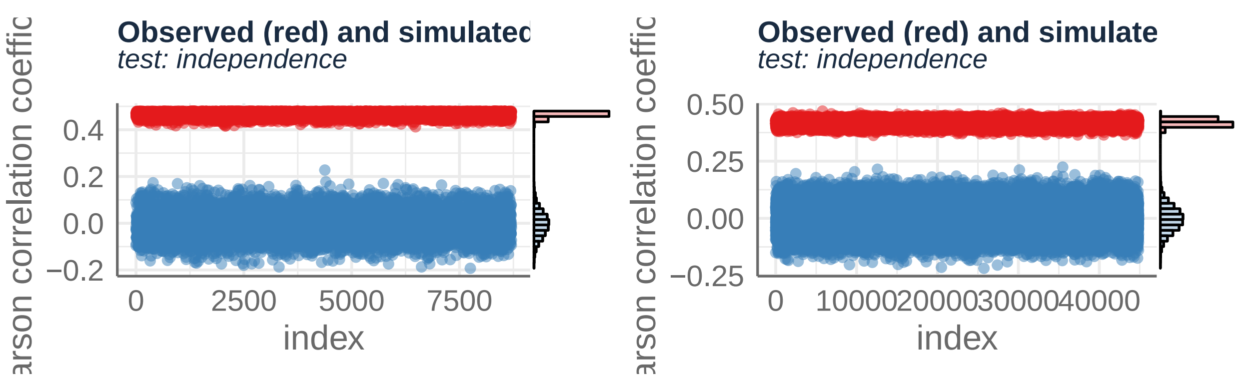

Figure 1: Observed and simulated correlation coefficients for model 1 (left) and model 2 (right). The figure shows that the observed residuals (in red) are much more correlated than is reasonable under the model (represented by the blue, simulated values).

From figure XX and the posterior predictive checks (table XY), we observe that the data are not independent ($\hat{r}_{\text{PPC, Model1}} = .466$). 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 $\hat{r} = .573$.[1]

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

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

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% CCI $= [.0008, .015]$ and $R^2_{\text{C}} = .5398$; 95% 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 (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)	0.001 (.0035, .0155)
(b) Random			
σ^2_e	.516	.501	.3423
σ^2_{u0}			.3992
(c) Model Fit			
Observations	336	336	336
Companies			52
DIC	508	488	280
Penalty.	4	6	47
R^2	.021	.098	.0056 (M), .5362 (C)
(d) Post. Pred. Checks			
Normality	.3609	.3534	
Homoskedasticity	.3261	.607	
Independence	0	0	
Note:	Baseline is sector 'Financials' for models (1) and (2)		

Table 2: Model results

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

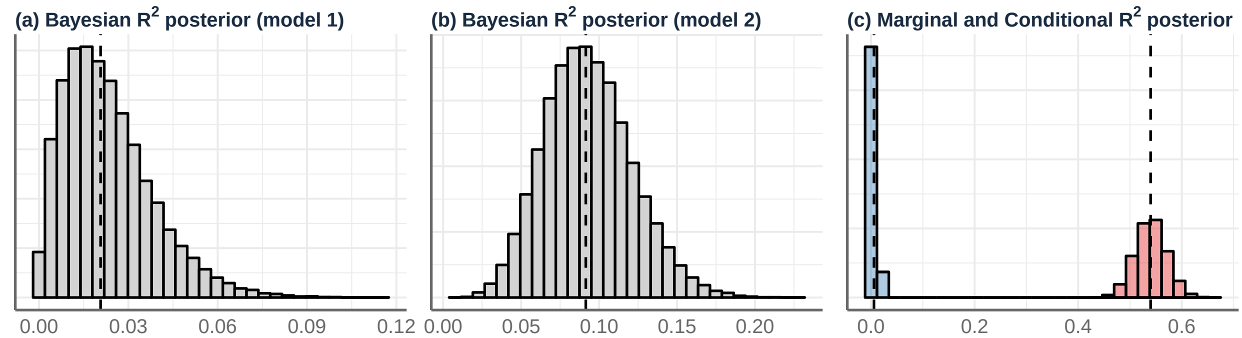

Figure 2: Bayesian R-squared value for model 1 (a) and model 2 (b). The proportion of cases in which the R-squared value of model 2 exceeds that of model 1 is .99. Figure (c) shows the marginal (blue) and conditional (red) R-squared values for the posterior distribution. The marginal R-squared indicates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

Differences in Bayesian and Frequentist inference

- Interpretation of random variables
- Same rules for probability, but different approach
- Can answer the same questions, but Bayesian more flexible
- Endless transformations on posterior distribution
- Incorporating domain knowledge

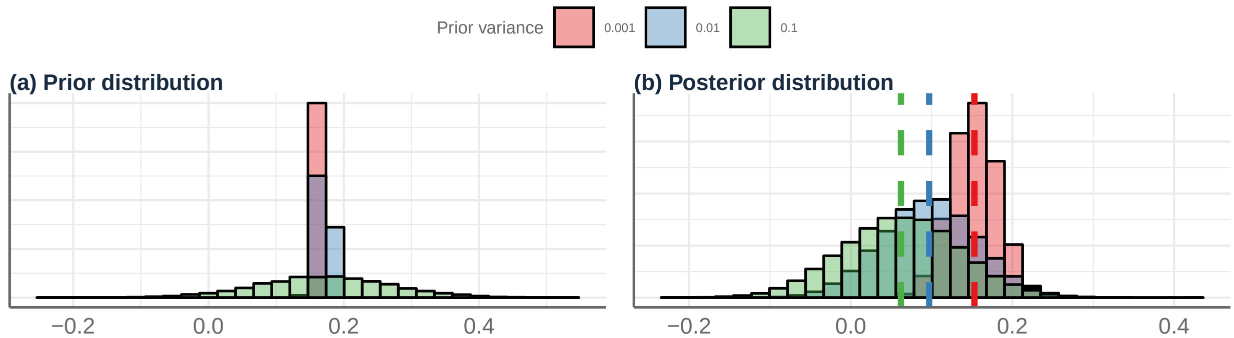

Figure 3: Effect of adjusting the prior variance of the coefficient for variable Male (a) on the posterior distribution (b). The mean is set at M=.17 which represents a 17 percent increase in compensation for males versus females. Reducing the prior variance represents increased certainty about the estimate 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 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.

Other material

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.