Final Assignment: Predicting Director Compensation

Jasper Ginn (s6100848)

April 8, 2019

Introduction

The gender gap in compensation remains a pervasive problem despite efforts to achieve income parity between men and women in the private and public sector. The boards of American companies consists of *executive* and *independent* directors. Execute directors are employed by the firm and usually have high positions within the company (for example the Chief Executive Officer), while independent directors are concerned primarily with oversight [CITE]. In this paper, I examine the gender gap in compensation among 336 independent directors in three sectors and 52 companies using Bayesian Linear Regression.

The results indicate that we cannot conclude definitively that male independent directors receive a higher remuneration than female independent directors after controlling for age. Model fit statistics and evaluation suggest that linear regression is not an appropriate method to analyse these data because director compensation is highly correlated within companies.

The paper is structured as follows. Section one elaborates on the differences between Frequentist and Bayesian statistical inference. In turn, section two outlines the hypotheses, model specification and reports the results of the analyses. Section three concludes.

Differences in Bayesian and Frequentist Inference

Many differences between Frequentist and Bayesian inference originate in their respective view of uncertainty, how this is captured by probabilities and what causes this uncertainty. In the Frequentist framework, a probability is defined as the number of times an event will occur in the long run[1]. In other words, it is the limiting value of m successes in a sequence of n trials for a particular event, or $p = \lim_{n \to \infty} \frac{m}{n}$.

This definition has two implications. The first is that talking about probability makes sense only in the context of infinite trials. The second implication is that probability converges to some fixed quantity as the number of trials go to infinity, and our uncertainty about the true value of this quantity reduces as we repeatedly take more (or larger) samples. In turn, this implies that the only source of randomness by which our estimate $\hat{p}$ differs from the true value p comes from the data, which may differ from sample to sample due to, for example, sampling error.[2]

The Bayesian framework considers probabilities as a means of quantifying uncertainty about knowledge.[3] Even though the 'true' parameter value may be fixed, we are limited by our knowledge of this value. The uncertainty in our knowledge will change as we collect more information and is represented by the posterior distribution.

Objective and Subjective Knowledge

A Frequentist believes that evidence can only originate 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 certain about our knowledge then we can constrain the parameter space by injecting what we know *a priori*. In other words, a Bayesian

[1] Miller, Miller, and Freund, *John E. Freund's Mathematical Statistics with Applications*, p.21

[2] Miller, Miller, and Freund,

[3] Gelman et al., *Bayesian Data Analysis*, pp. 11-13

looks upon prior beliefs as just another source of knowledge that has been translated into a probability density.

The claim that the Frequentist approach is more objective only holds if one regards the data collection process, (pre-)processing steps and analysis as guaranteed to be objective. This is a tenuous assumption at best in the social sciences; statistics is fraught with subjective decisions during data collection, manipulation and analysis[4], and the charge of subjectivity leveled at Bayesians boils down to the practice of incorporating domain knowledge *explicitly* through the use of a prior distribution.

Methods of Estimation and Hypothesis Testing

Whether a statistician regards the data or the parameters as a random variable determines their choice of estimation method. A Frequentist, believing that variation originates from data, will want to find the most likely combination of parameters $\hat{\theta}$ that explain the data and that provide consistent and asymptotically unbiased estimators. 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 uncertainty in our beliefs after seeing the evidence. The different estimation methods provide different predictions; Maximum Likelihood (ML) estimates yield point estimates and MCMC methods yield posterior predictive distributions. The latter can be used to evaluate, for example, the fit of the model by means of Posterior Predictive Checks (PPP).

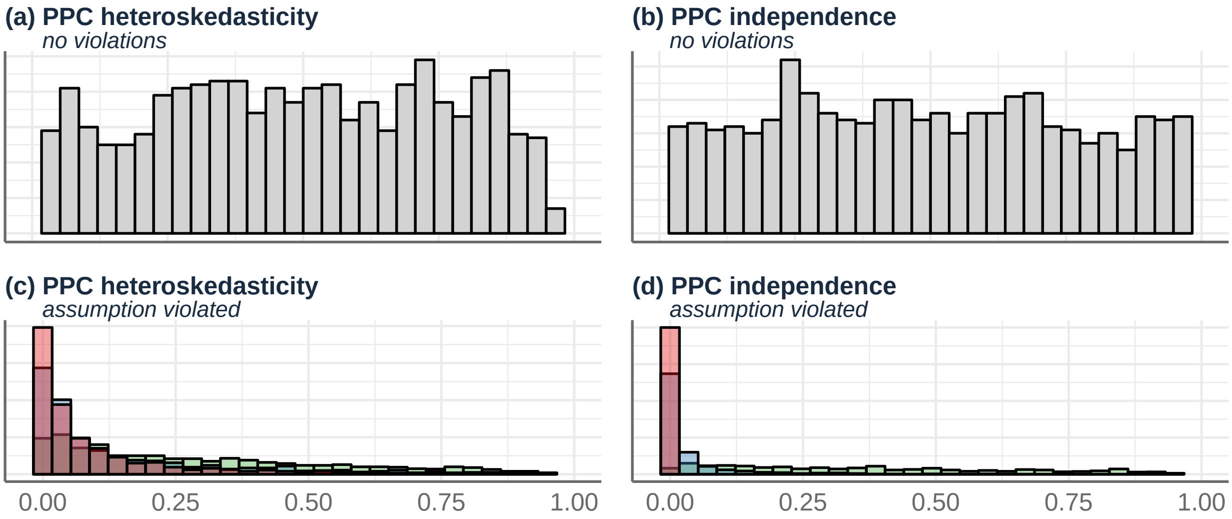

Figure 1: Distributions of posterior predictive p-values for 1.000 simulated data sets. In plots (a) and (b), the simulated data are drawn from a normal without any violations of the linear regression assumptions. In plot (c), the assumption of homoskedasticity is violated in each of the simulations. 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 can be found on GitHub.

The way in which we think about statistical 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, the upper and lower boundaries of the confidence interval are interpreted as random variables due to sampling error. Therefore, we should interpret it as … 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.

[4] Berger and Berry, "Statistical Analysis and the Illusion of Objectivity."

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. We use either confidence intervals or p-values to express support for the alternative hypothesis, but, importantly, we can never quantify support in favor of the null hypothesis. This is not so in the Bayesian Framework. Given that the data can be viewed as a means to update a prior belief, hypothesis testing is a relative statement about the degree to which the evidence found in the data supports this initial belief or contradicts it.

Methods & Results

This section describes the results of the statistical analysis. 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. The variable **Age** has been grand-mean centered to facilitate interpretation and estimation. Some 44% of the data has been deleted due to missingness on one of the variables, which leaves us with 336 observations.

The model is given by the following equation:

$$\log(\widehat{\text{compensation}_i}) = \beta_0 + \beta_1[\text{age}_i - \overline{\text{age}}] + \beta_2\text{male}_i + \epsilon_i$$

Given that we measure the compensation on a log scale, the coefficients we derive from the model must be interpreted as change in percentages. This forces us to rethink the priors we set on the model. We assume that, as age increases, compensation increases as well. Given that we are not certain about the extent to which it increases, we set the the mean to .05 and the standard deviation of this estimate to .2, representing a spread of approximately 20%. 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 [CITE]. Hence, we set this prior a mean of .1 with a spread of .1, which reflects our uncertainty of the estimate.

In accordance with the theory, we hypothesize that male directors earn more than their female counterparts and that compensation only increases with age. H_u is the unconstrained hypothesis.

$$\text{H}_1\text{: } \beta_{\text{Male}} > 0 \qquad \text{H}_2\text{: } \beta_{\text{Age}} > 0 \qquad \text{H}_u\text{: } \beta_{\text{Male}}, \beta_{\text{Age}}$$

The model results are presented in table XX. The convergence diagnostics show that the posterior distributions have forgotten their initial states and have converged to their stable distributions. There is no trace of autocorrelation and the Gelman-Rubin statistic is close to 1, indicating that the chains have converged to the same stationary distribution. The DIC for the model of interest (model 2) is 508 compared to 513 for the intercept-only model (model 1). The difference between the DIC scores indicates that the model fits better than the intercept-only model, but only marginally. The MC error, which should be less than 5% of the standard deviation, is negligible.

The model results indicate that gender is not a predictor of director compensation ($\beta_{\text{male}} = .067; 95\% \text{ CCI} = [-.08, .214]$). The CCI and standard deviation for this coefficient indicate that there is a lot of uncertainty in this estimate. Age is also not a predictor has a small, positive effect on director compensation ($\beta_{\text{age}} = .008; 95\% \text{ CCI} = [-.00, .017]$). The lack of explanatory power provided by the variables that we observed in the parameter estimates and DIC is also reflected by the R^2 value given that the 95% CCI contains values close to 0 (see also figure 2). The median R^2 value for this model is $R^2 = .021; 95\% \text{ CCI} = [.002, .059]$) [CITE GELMAN], indicating that, using these variables, we explain some 2% of the variance in compensation.

From the posterior predictive checks (table 1), we observe that the data violate the independence assumption of the model ($p_{\text{independence}} = 0$. 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 $\hat{r} = .573$.[5]

Given the hierarchical nature of the data, we next run a linear mixed effects model. This model corresponds to the following equation.

$$\log(\widehat{\text{compensation}}_i) = \gamma_{00} + u_{0j} + \gamma_{10} \cdot [\text{Age}_{ij} - \overline{\text{Age}}]) + e_{ij} \tag{3}$$

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 ($DIC = 280$), indicating that the multilevel approach seems appropriate. The marginal and conditional R-squared values (Nakagawa and Schielzeth[6])[7] 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 can be confident that director compensation is best modelled using company-specific intercepts, but that we do not have the right variables at either the company or individual level that would help us explain the heterogeneity in compensation among directors within boards and between boards.

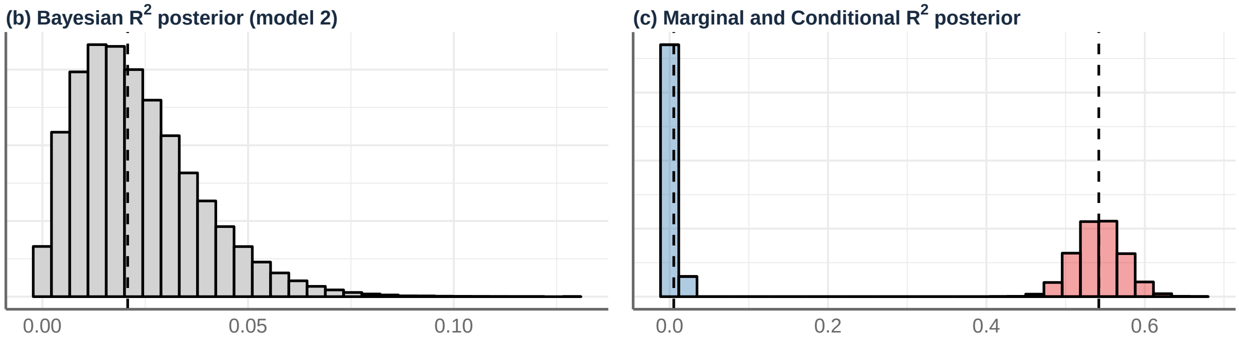

Figure 2: Bayesian R-squared value for model 1 (a) and model 2 (b). The proportion of cases in which the R-squared value of model 2 exceeds that of model 1 is .99. Figure (c) shows the marginal (blue) and conditional (red) R-squared values for the posterior distribution. The marginal R-squared indicates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

[5] The complete multilevel analysis of this data can be found here.

[6] "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models."

[7] The marginal R-squared takes into account only the fixed part of the variance, while the conditional R-squared takes into account the fixed and random parts

	Dependent variable:		
	Compensation (GBR ’000, logged)		
	(1)	(2)	(3)
	Intercept-only	Full model	Linear mixed effects
	(blm)	(blm)	(JAGS)
(a) Fixed			
Constant	4.991 (4.935, 5.048)	4.820 (4.731, 4.901)	4.993 (4.875, 5.112)
Male		.062 ($-$.08, .205)	
Age		.010 (.002, .018)	.009 (.003, .015)
(b) Random			
σ_e^2	.52	.5	.342 (.315, .372)
σ_{u0}^2			.399 (.322, .504)
(c) Model Fit			
Observations	336	336	336
Companies			52
DIC	513	488	281
Penalty.	2	6	49
R^2	0	.09	.006 (M), .536 (C)
(d) Post. Pred. Checks			
Normality		.357	
Homoskedasticity		.612	
Independence		0	
(e) Bayes’ Factors (model 2 only)			
Hypothesis	BF (complexity, fit)	PMPa	PMPb
$\text{H}_1 : \beta_{\text{Male}} \leq .05$	4294.5 (.812, 1)	.382	.291
$\text{H}_2 : \beta_{\text{Age}} > 0$	129.8 (.498, .992)	.618	.472
$\text{H}_u : \beta_{\text{Male}}, \beta_{\text{Age}}$			.237

Table 1: Model results

Conclusion

Berger, James O, and Donald A Berry. “Statistical Analysis and the Illusion of Objectivity.” *American Scientist* 76, no. 2 (1988): 159–65.

Gelman, Andrew, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman; Hall/CRC, 2013.

Miller, Irwin, Marylees Miller, and John E Freund. *John E. Freund’s Mathematical Statistics with Applications*. Boston: Pearson, 2014.

Nakagawa, Shinichi, and Holger Schielzeth. “A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models.” *Methods in Ecology and Evolution* 4, no. 2 (2013): 133–42.