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. In 2019, the difference in compensation between men and women stood at an estimated median 17% in the general workforce [CITE], and it appears that the glass ceiling may stretch all the way to the top. In 2016, a report found that the average growth in compensation for male execute board members of S&P500 companies[1] was twice that of female executive board members [CITE].

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]. Most of the literature on director compensation concerns executive directors rather than independent directors [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 and sector. 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 2 elaborates on the differences between Frequentist and Bayesian statistical inference. In turn, section 1 outlines the hypotheses, model specification and reports the results of the analyses. Section three concludes.

Differences in Bayesian and Frequentist Inference

For a large part, differences between Frequentist and Bayesian inference stems from their respective view of uncertainty, how this is captured by probabilities and, in particular, 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 (I. Miller, Miller, and Freund (2014), p.21). In other words, it is the limiting value of m successes in a sequence of n trials for a particular event, or:

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

This definition has two important 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 infinity, and our uncertainty about the true value of this quantity reduces as we repeatedly take more (or larger) samples. Hence, 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], which may differ from sample to sample due to, for example, sampling error.

The Bayesian framework considers probabilities as a means of quantifying uncertainty about knowledge (Gelman et al. (2013), pp. 11-13). 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

[1] The S&P500 is a stock market index that tracks the 500 most profitable companies listed on the US stock exchange.

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.

Objective and Subjective Knowledge

By the very definition of probability in the Frequentist framework, evidence can 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 certain about our knowledge then we can constrain the parameter space by injecting what we know *a priori*. Hence, a Bayesian 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 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; statistics is fraught with subjective decisions during data collection, manipulation and analysis Berger and Berry (1988). Hence, the critique leveled at Bayesians boils down to the practice of incorporating domain knowledge *explicitly* through the use of a prior. It should be noted that, if an analyst chooses to use priors such that they represent a complete lack of knowledge (for example, a uniform distribution), results obtained using Frequentist and Bayesian methods often yield comparable results.

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 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 uncertainty in our beliefs after seeing the evidence:

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

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, the boundaries of the confidence intervals are interpreted as random variables due to sampling error since they estimate the frequency of sampling means. 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.

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.

Predictions under the model

Bayesian models yield the posterior probability distribution of the model parameters. This allows us to produce a posterior predictive distribution for each subject in the data set. Unlike Frequentist model results, such distributions can be transformed almost endlessly, and it provides us with a useful way to understand whether the model adheres to its core assumptions by means of *posterior predictive checks* (PPC). PPC require us to simulate data under the model and to compare the simulated outcome variable to the observed outcome variable, which results in a proportion sometimes called the Bayesian p-value. 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. We often find that this assumption does not hold true for posterior predictive checks. To this end, we can simulate data under various conditions to observe how the statistic behaves. This is illustrated in figure XX for two posterior predictive checks included in the R library blm.

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.

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. The variable **Age** has been grand-mean centered to facilitate interpretation and estimation.

The model is given by the following equation:

$$\log(\hat{\text{compensation}}_i) = \beta_0 + \beta_1[\text{age}_i - \bar{\text{age}}] + \beta_2 \text{male}_i + \beta_3 \text{SectorServices}_i + \beta_4 \text{SectorBasicMaterials}_i + \epsilon_i \quad (1)$$

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. 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. But, given that we are not certain about the extent to which it increases, we set the standard deviation of this estimate to .2, representing a spread of approximately 20%[2]. 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 .1 with a spread of .1, which reflects our uncertainty of the estimate.

In accordance with the theory, we hypothesize that there is at most a small difference in compensation between male and female directors compared to the rest of the work force, which we will set at 5%.

$$\mathrm{H}_1\text{: } \beta_{\text{Male}} \leq .05$$

Furthermore, we hypothesize compensation only increases with age. That is:

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

Finally, we specify the unconstrained hypothesis as:

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

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 little to no trace of autocorrelation and the Gelman-Rubin statistic is close to 1, which indicates that the chains have converged to the same stationary distribution. The DIC for the model of interest (model 2) is 488 compared to 512 for the intercept-only model, which indicates it yields a better fit. The MC error, which should be less than 5% of the standard deviation, is negligible.

From the posterior predictive checks (table XY), we observe that the data are not independent ($p_{\text{independence}} = 0$. 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 $\hat{r} = .573$.[3]

The model results indicate that gender is not a predictor of director compensation ($\beta_{\text{male}} = .062; 95\%$ CCI $= [-.08, .205]$), and credible interval for this coefficient tells us that there is a lot of uncertainty in this estimate. Age has a small, positive effect on director compensation ($\beta_{\text{age}} = .01; 95\%$ CCI $= [.002, .018]$). When holding the other variables constant, every one-year increase above the mean age leads to a 1% increase in compensation. Finally, it appears that directors in the sectors 'basic materials' and 'services' earn some 25% and 34% more than their counterparts in the financial sector. The mean R^2 value for this model is $R^2 = .09; 95\%$ CCI $= [.041, .15]$) [CITE GELMAN], indicating that we are explaining between 4% and 15% of the variance in the data.

Given the hierarchical nature of the data, we next run a linear mixed effects model. 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 ($DIC = 280$), indicating that the multilevel approach

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

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

seems appropriate. The marginal and conditional R-squared values (Nakagawa and Schielzeth (2013)). [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] are $R^2_{\mathrm{M}} = .0056$; 95% CCI $= [.0008, .015]$ and $R^2_{\mathrm{C}} = .5398$; 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 can be confident that director compensation is best modelled using company-specific intercepts, but that we do not have the right variables at either level 1 or 2 that would help us explain the heterogeneity in compensation among directors within boards and between boards.

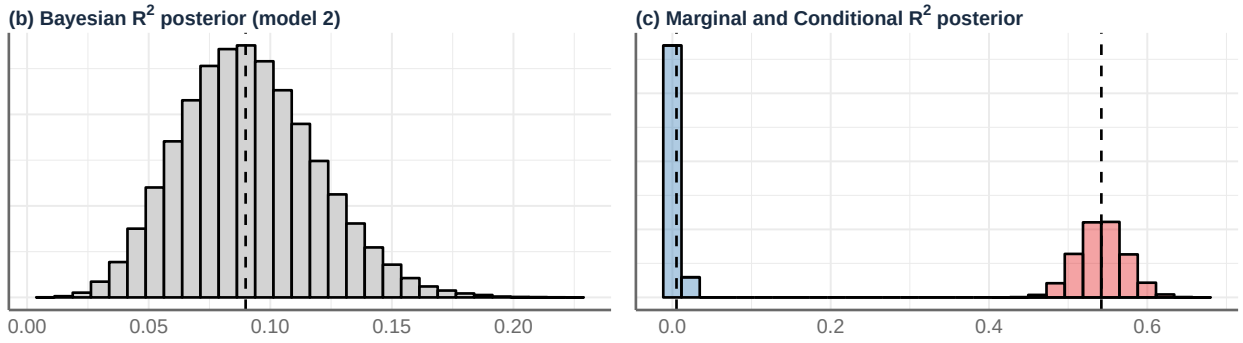

Figure 2: Bayesian R-squared value for model 1 (a) and model 2 (b). The proportion of cases in which the R-squared value of model 2 exceeds that of model 1 is .99. Figure (c) shows the marginal (blue) and conditional (red) R-squared values for the posterior distribution. The marginal R-squared indicates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

	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)
SectorBasic Materials		.225 (.082, .37)	
SectorServices		.294 (.172, .415)	
Male		.062 (−.08, .205)	
Age		.010 (.002, .018)	.009 (.003, .015)
(b) Random			
σ^2_e	.52	.5	.342 (.315, .372)
σ^2_{u0}			.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
$H_1 : \beta_{\text{Male}} \leq .05$	4294.5 (.812, 1)	.382	.291
$H_2 : \beta_{\text{Age}} > 0$	129.8 (.498, .992)	.618	.472
$H_u : \beta_{\text{Male}}, \beta_{\text{Age}}$			.237
Note:		Baseline is sector 'Financials' for models (1) and (2).	

Table 1: Model results

References

Berger, James O, and Donald A Berry. 1988. "Statistical Analysis and the Illusion of Objectivity." *American Scientist* 76 (2). JSTOR: 159–65.

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

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

Nakagawa, Shinichi, and Holger Schielzeth. 2013. "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models." *Methods in Ecology and Evolution* 4 (2). Wiley Online Library: 133–42.