Gender and Racial Inequality

Introduction

Gender and racial inequality have been an ongoing topic of debate and controversy in the modern century (United Nations, 2021). Throughout history, there have been many instances of gender and racial inequality, such as women being barred from voting in national elections, and bathrooms segregated based on race. Humanity has made great progress fixing these inequalities; women can now vote, drive, work, and do many other things that they were not able to before, notwithstanding some extreme cases like the Taliban-ruled Afghanistan, under which women still struggle to achieve some of their basic human rights (United Nations, 2023). However, while many cases of extreme inequality have been eradicated, lesser forms of inequality are still present, among which the most heavily-discussed being salary-based inequality (Brito, 2022) (Lang & Spitzer, 2020, p. 81). There is debate as to whether women really earn less than men or whether one race earns less than another. In this report, we will evaluate some important variables that determine a person's income, and try to fit a model that allows us to examine the underlying causal relationship between income and other important variables, especially race and sex.

With the provided cross-sectional Australian economic data, we can begin to analyse our variables of interest. Specifically, we are interested in whether income differs across sexes and race, which in this case would be related to indigenous status. One thing to note: The variables in the dataset represent average values, which has implications for estimation.

Mean Average Income by Gender and Indigenous Status

Category	Mean Average Income
Male	786.5619
Female	672.3603
Indigenous	687.3370
Non-Indigenous	745.5046

fig.1

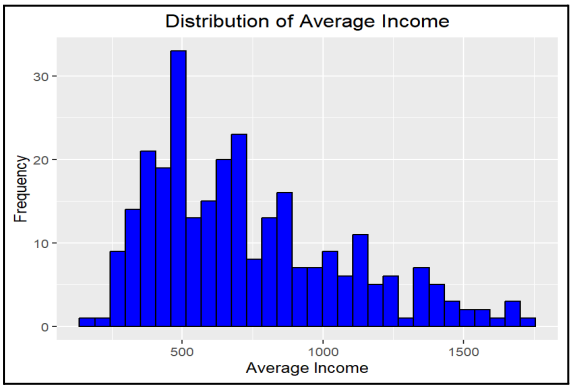


fig.2

Fig.1 shows the mean average income (MAI) achieved by the different sexes and races in Australia. We observe that males have an MAI that is more than AUD 100 higher than females. On the other hand, non-indigenous persons have an MAI that is more than AUD 50 more than indigenous persons. These statistics indicate a possibility of gender-based or racial-based discrimination in the workforce. However, these results alone are not conclusive as we have not yet taken other variables into account. **Fig.2** shows the distribution of the dependent variable, average income. We can see that it is positively skewed, with few observations having average incomes much higher than the rest. This aligns with the reality of income inequality across the world, in which certain demographics have a much higher income compared to others.

Modelling

The linear regression line goes through the mean of variables. If a variable is skewed, then this line may not be a good representation of the data. Logging the dependent variable average income can reduce its skewness and stabilize its variance, making the linear regression line a better fit. In addition, the stabilising of the variance can reduce heteroskedasticity (HSK) in the error terms, which skews the standard errors of the estimated coefficients and makes hypothesis testing unreliable. **Fig 3** shows the distribution of average income after logarithmic transformation. You can observe a more symmetrical distribution and a more scaled x-axis, stabilizing the variance.

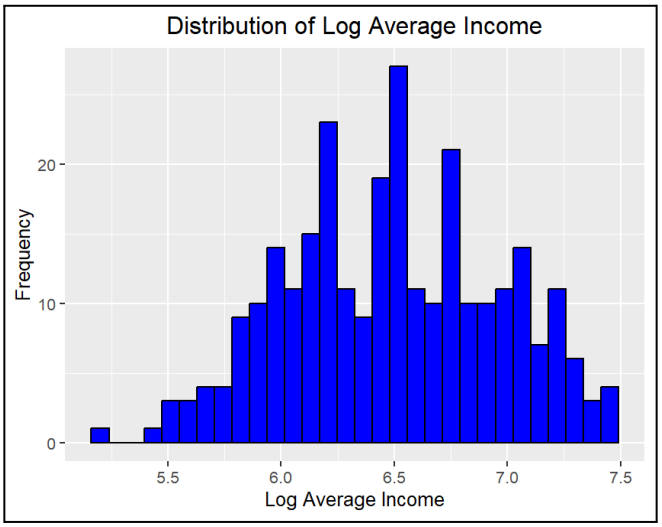


fig. 3

For our first model, all the variables were regressed on log(income) and observed the results (See Appendix A). Hypothesis testing was performed on the coefficients of the variables to determine their significance at $\alpha = 0.05$:

$H_0: \hat{\beta}_k = 0$ $\quad$ $H_1: \hat{\beta}_k \neq 0$

It is observed that most of the region variables are insignificant (p > 0.05), indicating that region may not play a significant role in determining average incomes. For simplicity, parsimony, and generalisation, the regions were transformed into a dummy variable, *urban*, with 1 representing an urban region and 0 representing a non-urban, or rural region. In addition, the *no_people* variable was insignificant. This variable represents the sample size for each of the average values of the other variables and is not a variable in and of itself. Therefore, it is nonsensical to include it in our model, so it was removed in future models. *no_people* represents the sample size for the average values obtained. When dealing with averages, the variance of the average is inversely proportional to the sample size; that is, larger sample sizes produce more precise (lower variance) estimates, while smaller sample sizes result in higher variability. Therefore, the variance of the error terms is proportional to this sample size. As a result, we can use a Weighted Least Squares model to account for this, setting the weightage as 1/no_people.

Thus, for our second model, these changes were implemented. After running the model, the newly-formed *urban* variable was seen to be insignificant (See Appendix A). In addition, a Breusch-Pagan (BP) test was conducted to check for HSK in the residuals. The hypotheses are as follows:

$H_0: Var(u_i) = \sigma^2$ (homoskedasticity) $\quad$ $H_1: Var(u_i) = \sigma_i^2$ (heteroskedasticity)

The BP test showed a p-value > 0.05 indicating failure to reject the null hypothesis. This indicates that the model's residuals exhibit a homoskedastic pattern. This proves the effectiveness of using Weighted Least Squares to account for heteroskedasticity in model residuals. When visualizing a scatterplot of *no_people* against the model's residuals, we can indeed observe a very heteroskedastic pattern (See Appendix B). Smaller values of *no_people* which represents a small sample size, results in much larger variability in the residuals of the model. The transformation from Ordinary Least Squares to Weighted Least Squares resulted in a significant increase in AIC and BIC values of the model, however, this is not related to model adequacy in this case and is just the byproduct of including a weightage in a regression model. The elimination of heteroskedasticity in the residuals means that the standard errors of the coefficients are unbiased, resulting in more accurate hypothesis tests.

By visualizing the scatterplot of no_years_school against the residuals of the fitted model (See Appendix B), we can also observe a heteroskedastic pattern. This makes sense as the variability of income may increase with years of education; a lowly-educated person may have no choice but to work at a minimum wage, or close to a minimum wage job. However, a highly educated person may choose to work in a financially-prosperous private sector industry or he may choose to devote his efforts in an NGO, where the pay can be much lower. Thus, we can also set the weightage as 1/no_years_school. Combining the two weightages, we get a weightage of 1 / (no_people * no_years_school) for our new model. In addition, a quadratic term for age was also included in the new model as there is reason to believe that the relationship between age and income is not strictly linear; a younger person, for example a 16 year old, may be working part-time to support their family or as a way to earn extra income. As they near adulthood, complete high school, and enter higher education, they may forgo part-time work as they focus on university. On the other hand, an older person may have gained valuable skills, completed higher education, and gained industry experience, all of which boost earning potential. The final model can be written in linear form as follows:

$$w(\log(\hat{Income})) = w(\hat{\beta}_0 + \hat{\beta}_1 PartTime + \hat{\beta}_2 EngProfNotWell + \hat{\beta}_3 EngProfWell + \hat{\beta}_4 EngProfVeryWell + \hat{\beta}_5 Male + \hat{\beta}_6 NonIndigenous + \hat{\beta}_7 Age + \hat{\beta}_8 Age^2 + \hat{\beta}_9 No\_Years\_School + \hat{\beta}_{10} Male * NonIndigenous)$$

$$where\ w = 1/(no\_people * no\_years\_school)$$

	Dependent variable:
	log(income)
employstatusEmployed, worked part-time	-0.650***
	(0.036)
englishproficiencyNot well	0.185***
	(0.051)
englishproficiencyVery well	0.935***
	(0.080)
englishproficiencyWell	0.322***
	(0.072)
sexMale	0.106**
	(0.043)
indigenousNon-Indigenous	0.509***
	(0.065)
age	-0.037***
	(0.013)
I(age2)	0.001***
	(0.0002)
no_years_school	0.029*
	(0.015)
sexMale:indigenousNon-Indigenous	-0.007
	(0.067)
Constant	6.167***
	(0.300)
Observations	282
R^2	0.754
Adjusted R^2	0.745
Residual Std. Error	0.015 (df = 271)
F Statistic	83.170*** (df = 10; 271)
Note:	*p<0.1; **p<0.05; ***p<0.01

fig. 4 **fig. 4 (cont.)**

A full comparison of all three models can be found in Appendix A.

Interpretation of Results and Significant Coefficients

Estimator	**Interpretation and Hypothesis Testing (threshold: 0.05)**
Intercept	For a 0 years old indigenous female with 0 years of education who is not at all proficient in english, the mean income is $e^{6.167} \approx$ AUD 476.75. However, this is not realistic as there exists nobody with these characteristics.
$\hat{\beta}_{parttime}$	The mean income of part-time workers is $(e^{0.65} - 1) * 100 \approx 91.55\%$ less than full-time workers, on average, ceteris paribus. With $p < 0.01$, there is strong evidence to conclude that full-time workers have a higher mean income than part-time workers. This is unsurprising as full-time workers work more hours than part-time workers and generally have more responsibilities as well.
$\hat{\beta}_{EngProfNotWell}$	The mean income of individuals whose english proficiency is deemed "Not Well" have a mean income that is $(e^{0.185} - 1) * 100 \approx 20.32\%$ higher than the mean income of individuals who are not at all proficient in English, on average, ceteris paribus. With $p < 0.01$, there is strong evidence to conclude that individuals with 'Not Well' english proficiency have a higher mean income than individuals with no english proficiency at all.
$\hat{\beta}_{EngProfVeryWell}$	The mean income of individuals whose english proficiency is deemed "Very Well" have a mean income that is $(e^{0.935} - 1) * 100 \approx 154.72\%$ higher than the mean income of individuals who are not at all proficient in English, on average, ceteris paribus. With $p < 0.01$, there is strong evidence to conclude that individuals with 'Very Well' english proficiency have a higher mean income than individuals with no english proficiency at all.
$\hat{\beta}_{EngProfWell}$	The mean income of individuals whose english proficiency is deemed "Well" have a mean income that is $(e^{0.322} - 1) * 100 \approx 37.99\%$ higher than the mean income of individuals who are not at all proficient in English, on average, ceteris paribus. With $p < 0.01$, there is strong evidence to conclude that individuals with 'Well' english proficiency have a higher mean income than individuals with no english proficiency at all.
$\hat{\beta}_{Male}$	The mean income of males is $(e^{0.106} - 1) * 100 \approx 11.18\%$ higher than

	the mean income of females, on average, ceteris paribus. With p < 0.05, there is moderate evidence to conclude that full-time workers have a higher mean income than part-time workers.
$\hat{\beta}_{NonIndigenous}$	The mean income of non-indigenous individuals is $(e^{0.509} - 1) * 100 \approx$ 66.36% higher than the mean income of indigenous individuals, on average, ceteris paribus. With p < 0.01, there is strong evidence to conclude that the mean income of non-indigenous individuals is higher than the mean income of indigenous individuals.
$\hat{\beta}_{age} + \hat{\beta}_{age^2}$	We can find the turning point of age by calculating the minimum point of the following function: $f(ln(y)) = -0.037x + 0.001x^2$ $\rightarrow \frac{dln(y)}{dx} = -0.037 + 2 * 0.001x$ $\rightarrow 0 = -0.037 + 2 * 0.001x$ $\rightarrow 0.002x = 0.037$ [note: $\frac{d^2ln(y)}{dx^2} = +0.002$ indicating $\rightarrow x = \frac{0.037}{0.002}$ $f(ln(y))$ has a relative minimum at the $\rightarrow x = 18.5\ years$ critical point when $\frac{dln(y)}{dx} = 0$ Mean Log(income) decreases as age increases until the minimum point at x = 18.5 when mean log(income) is at its lowest, at which mean log(income) starts to increase with increased values of x. This is logical, as that is generally the age at which one graduates high school and may start working and earning an income. Given that the minimum age in our dataset is 22, we can deduce that increased age only leads to an increase in mean income. With p < 0.01, there is strong evidence that an increase in mean age leads to an increase in mean income, on average, ceteris paribus.

In terms of model fit, both AIC and BIC values went down with the final model compared to the second model. This shows that the final changes we made from our second model to our final model proved to be beneficial. In addition, our final model shows little signs of multicollinearity with no Generalized Variance Inflation Error exceeding 2.15 (with the exception of age and age^2) (See Appendix C). Heteroskedasticity remains eliminated with a now even higher p-value for the final model's Breusch-Pagan test. A full comparison of AIC and BIC values across all 3 models can be found in Appendix D.

The significance of the *Male* variable along with its positive coefficient indicates that, holding other factors constant, **males earn more than females, on average.** This confirms the speculation that men earn more than women in the job market.

In terms of race, **non-indigenous workers earn more than indigenous workers, on average**, holding other factors constant. The significance and positive coefficient of the *Non-Indigenous* variable indicates this. In addition, given the much greater coefficient of *Non-Indigenous* compared to *Male*, the income inequality on the basis of race is much

greater than the one on the basis of gender as shown in our interpretation. Non-indigenous people had an average income that is 66% greater than that of indigenous people, while males had an average income that is 11% greater than that of females.

The interaction term Male * Non-indigenous tests whether there is any additional effect on average income on the basis of being both male and non-indigenous. This variable turned out to be insignificant suggesting that **the interaction of gender and race doesn't provide any additional effects on average income compared to the sum of the effects of gender and race independently.**

This analysis indicates that race may play a much larger part in income inequalities compared to gender. Nevertheless, both are significant factors regarding income discrepancies among demographics.

In terms of English proficiency, we can observe that 'Very Well' english speakers earned significantly more than 'Well' or 'Not Well' speakers when compared to the baseline of people who were 'not at all' proficient in English. 'Very Well' speakers earned 154% higher than 'not at all' speakers, 'Well' speakers earned 38% higher, and 'Not Well' speakers earned 20% higher. This highlights that, although any level of English proficiency would allow, on average, a higher income, being very proficient in English would allow you to earn significantly more. This is logical as working environments generally call for high-level language proficiency due to the technical nature of industry. This may vary from industry to industry, however.

No_years_school proving to be mildly insignificant challenges the traditional notion of the importance of formal education. Formal education may not be as significant as we thought it would be, at least with regards to increasing your income. Education, of course, provides more value than merely increasing your earnings and, thus, is still very important for the greater good of society.

The *age* variable proved to have a significant quadratic effect on average income, however, the negative relationship in the statistical results are deemed to be unreliable as they indicate a negative relationship between age and income up to the age of 18.5, however, the minimum age in the dataset was 22. Therefore, it was concluded that age has a significant positive effect on average income, indicated by the positive coefficient of age^2.

Conclusion

Overall, the evaluation has provided interesting results regarding income inequality on the basis of gender and/or race. After controlling for several key indicators of income, there is indeed an income discrepancy between gender and race, with the latter proving to have a much larger variation. The latter falls in line with (Gale & Mills, 2015, para. 2) which says that indigenous people are often marginalized. In addition, there was also insufficient evidence of any additional benefit/marginalization due to the interaction of gender and race beyond the sum of the individual effects of the respective gender and race.

While the final model was robust, there were still some limitations. Firstly, there are several omitted variables that were not controlled for in our model, with the most significant one being job sector/title. It is argued that women may opt for jobs in lower-paying sectors while men may opt for jobs in higher-paying ones (Olson, 2012). If this is the case, then a higher average income for males may be a result of the lucrative industry they tend to choose to work in, and not as a result of their gender. Another potentially significant variable not considered is intelligence. Intelligent people may bring more value to the workforce, potentially driving up their value and their income (Drasgow, 2003). While hard to measure, some metrics can be used as a proxy for intelligence, such as IQ. By not taking these two critical variables into account, the final model's ability to determine causality diminishes.

Other potential considerations that may have been taken into account are other interaction terms such as no_years_school * Female. It may have shown potential discrepancies between the effects of additional years of schooling between males and females. Perhaps females felt discouraged to pursue higher education due to the perceived lesser marginal benefit of an extra year of schooling for them compared to a male, ceteris paribus. These considerations would significantly aid the prescriptive power of our model.

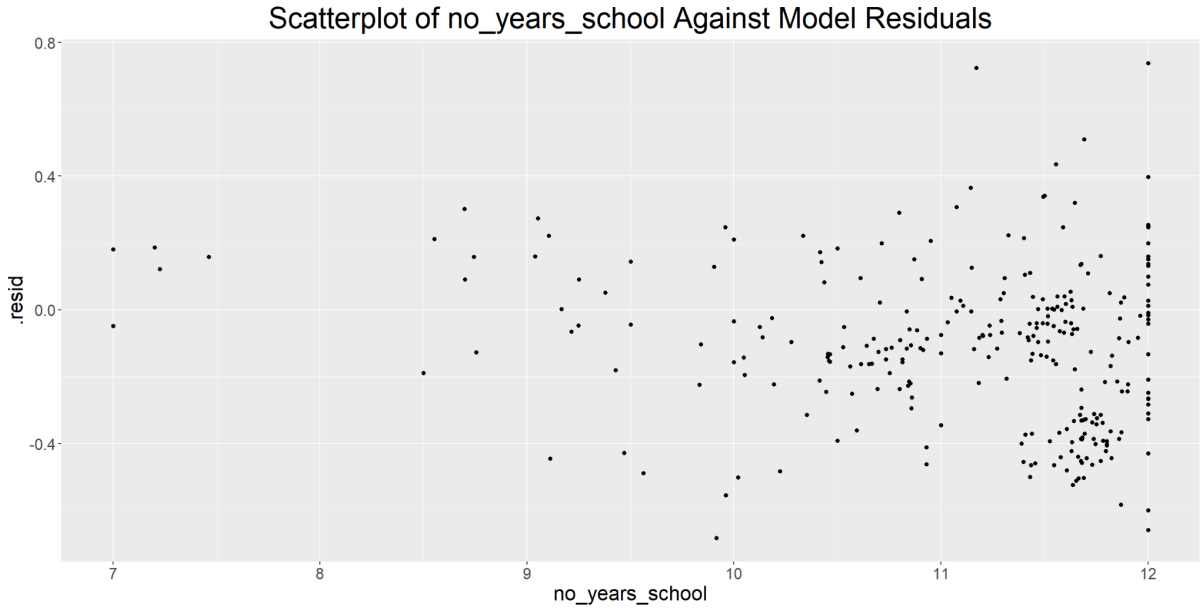
Appendix A

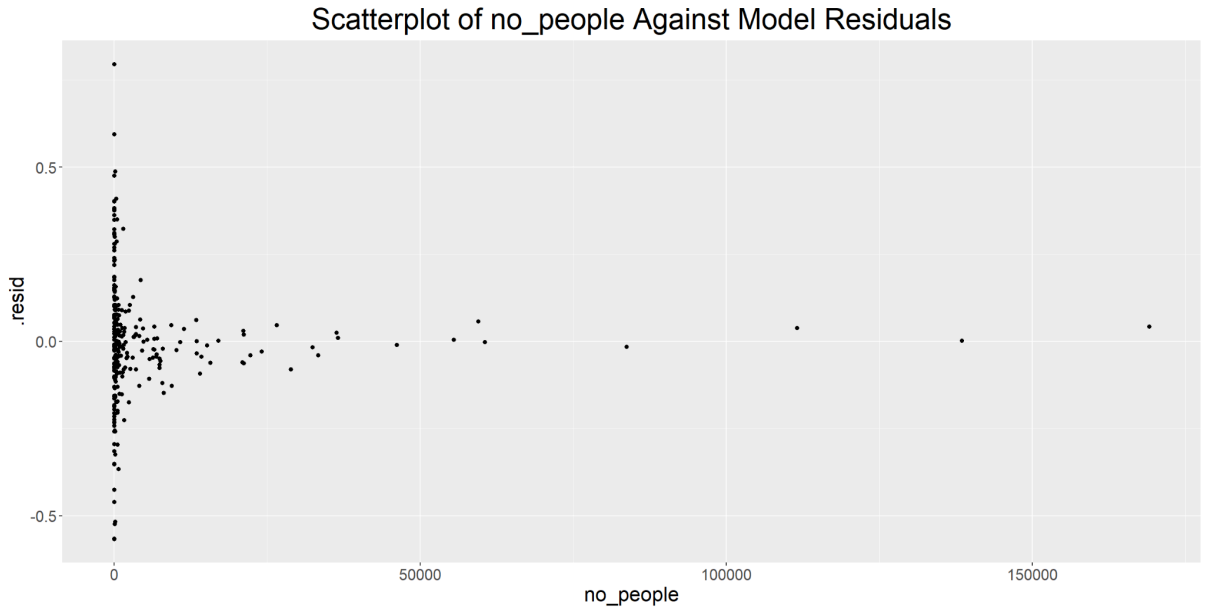
Full Comparison of all Three Models

	Dependent variable:		
	log(income)		
	(1)	(2)	(3)
regionGreater Adelaide	-0.121*		
	(0.066)		
regionGreater Brisbane	-0.033		
	(0.064)		
regionGreater Darwin	0.115*		
	(0.068)		
regionGreater Hobart	-0.206***		
	(0.073)		
regionGreater Melbourne	-0.071		
	(0.067)		
regionGreater Perth	0.018		
	(0.065)		
regionGreater Sydney	-0.047		
	(0.069)		
regionRest of NSW	-0.033		
	(0.065)		
regionRest of NT	-0.075		
	(0.065)		
regionRest of Qld	-0.072		
	(0.061)		
regionRest of SA	-0.123*		
	(0.066)		
regionRest of Tas.	-0.145**		
	(0.072)		
regionRest of Vic.	-0.092		
	(0.068)		
regionRest of WA	-0.025		
	(0.064)		
urban		0.045	
		(0.039)	
employstatusEmployed, worked part-time	-0.655***	-0.652***	-0.650***
	(0.023)	(0.037)	(0.036)
englishproficiencyNot well	0.049	0.177***	0.185***
	(0.040)	(0.051)	(0.051)
englishproficiencyVery well	0.470***	0.821***	0.935***
	(0.050)	(0.086)	(0.080)
englishproficiencyWell	0.181***	0.256***	0.322***
	(0.043)	(0.073)	(0.072)
sexMale	0.128***	0.119***	0.106**
	(0.022)	(0.033)	(0.043)
indigenousNon-Indigenous	0.225***	0.395***	0.509***
	(0.040)	(0.061)	(0.065)
age	0.012***	0.010***	-0.037***
	(0.003)	(0.002)	(0.013)
I(age2)			0.001***
			(0.0002)
no_years_school	0.099***	0.047***	0.029*
	(0.018)	(0.017)	(0.015)
no_people	-0.00000		
	(0.00000)		
sexMale:indigenousNon-Indigenous			-0.007
			(0.067)
Constant	4.847***	5.215***	6.167***
	(0.246)	(0.218)	(0.300)
Observations	282	282	282
R^2	0.859	0.735	0.754
Adjusted R^2	0.846	0.726	0.745
Residual Std. Error	0.183 (df = 258)	0.053 (df = 272)	0.015 (df = 271)
F Statistic	68.190*** (df = 23; 258)	83.800*** (df = 9; 272)	83.170*** (df = 10; 271)
Note:			*p<0.1; **p<0.05; ***p<0.01

Appendix B

Scatterplots showing heteroskedastic patterns of 2 variables, which were later used as weightages in a Weighted Least Squares final model





Appendix C

Generalized Variance Inflation Factors along with each variable's degrees of freedom and the adjusted generalized variance inflation factors, adjusting for the degrees of freedom. Age and Age^2 exhibit a higher GVIF value to the highly correlated nature of these 2 variables.

Variance Inflation Factors (VIF) for Final Model			
Variable	GVIF	Df	GVIF^(1/(2*Df))
employstatus	1.432603	1	1.196914
englishproficiency	5.937759	3	1.345665
sex	1.985936	1	1.409232
indigenous	4.639394	1	2.153925
age	64.894282	1	8.055699
I(age^2)	64.786006	1	8.048975
no_years_school	2.433799	1	1.560064
sex:indigenous	3.647076	1	1.909732

Appendix D

The increase in AIC and BIC from model 1 to model 2 is due to the addition of a weightage in the model. The decrease in AIC and BIC from model 2 to model 3 is due to the beneficial final changes made, including the addition of another variable in the weightage of the model (no_people), which aided model fit.

AIC and BIC of all Models		
Model	AIC	BIC
Model 1	-133.3071	-42.25945
Model 2	764.6401	804.70105
Final Model	744.1340	787.83693

References

Brito, C. (2022, February 11). *Serena Williams says "it takes time" to address gender pay equality in the sports world.* CBS News. Retrieved September 17, 2024, from https://www.cbsnews.com/news/serena-williams-sports-gender-inequality

Drasgow, F. (2003). Intelligence and the Workplace. , 107-130. https://doi.org/10.1002/0471264385.WEI1206.

Gale, T., & Mills, C. (2015, December 15). Creating Spaces in Higher Education for Marginalised Australians: Principles for Socially Inclusive Pedagogies. *Enhancing Learning in the Social Sciences*, 5(2), 7-19. https://doi.org/10.11120/elss.2013.00008

Lang, K., & Spitzer, A. K.-L. (2020, Spring). Race Discrimination: An Economic Perspective. *Journal of Economic Perspectives,* 34(2), 68-89. https://doi.org/10.1257/jep.34.2.68

Olson, K. (2012). Our Choices, Our Wage Gap?. *Philosophical Topics*, 40, 45 - 61. https://doi.org/10.5840/PHILTOPICS20124014.

United Nations. (2021, March 29). *UN News*. Gender equality, the ‘unfinished human rights struggle of this century’: UN chief. Retrieved September 17, 2024, from https://news.un.org/en/story/2021/03/1088512

United Nations. (2023, June 19). *OHCHR*. Taliban edicts suffocating women and girls in Afghanistan: UN experts. Retrieved September 17, 2024, from https://www.ohchr.org/en/press-releases/2023/06/taliban-edicts-suffocating-women-and-girls-afghanistan-un-expert