Gender Pay Gap and Educational importance analysis using General Social Survey data

Harish Krishnamoorthy Murali
21068381

Introduction

A person's income is not just impacted by personal criteria like education level and age but also social criteria like gender and occupational prestige score. Gender pay inequality remains a persistent issue globally, transcending geographical boundaries and socioeconomic strata. In this project, we delve into the intricate dynamics of the gender pay gap using empirical evidence drawn from the General Social Survey (GSS) spanning from 1974 to 2018. We would also like to understand how education level affects an individual's salary and the occupation prestige score to examine if higher education improves a person's income and how this trend varies across different occupations.

Dataset:

The dataset, aptly titled "The Gender Pay Gap in the General Social Survey," offers various variables providing insights into the economic landscape over several decades. [1][2]

Key variables: base income, age, occupation code, occupational prestige score, gender, education level.

Variables and summary statistics:

Occrecode: Categorical; Occupational category: 11 categories including Transportation, Professional, Office and Administrative Support, Business/Finance, Service, Production, Sales, Installation, Maintenance, and Repair, Armed Forces, Construction/Extraction, Farming, Fishing, and Forestry

Age: Numerical; variable indicating age; Ranges from 18 to 85

Prestig10: Numerical; Prestige score associated with individuals' occupations, based on measures of income, education, and occupational status; values range from 16 to 80

Gender: Categorical; Two categories: Male and Female; the 1974 dataset contains 348 males (66.8%) and 173 females (33.2%) and the 2018 dataset contains 507 males (51.9%) and 469 females (48.1%).

Educcat: Categorical; Five categories: "Less Than High School", "High School", "Junior College", "Bachelor" and "Graduate"

Realrinc: Numerical; Real annual income of individuals having a mean value of $26882.2 in 1974 and $29126.09 in 2018.

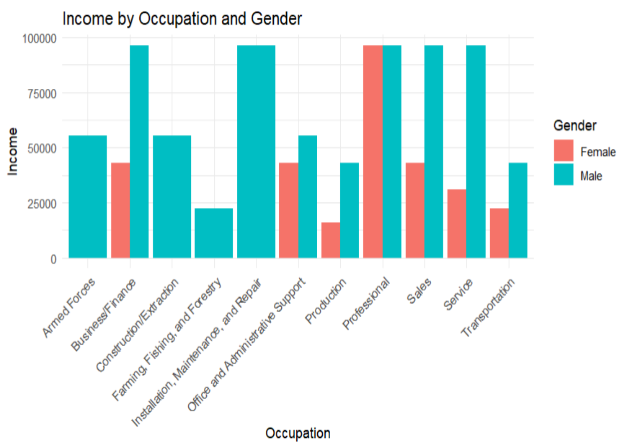
The original dataset contained additional variables like number of children and marital status of individuals which were not included as a part of this study as we intend to address only topics relating to gender, education and pay.

Data preparation:

The original dataset contains 61697 entries collected over a period of 1974 to 2018. By constraining on the year, we reduce and split the dataset by selecting only the data from 1974 and 2018. The data was cleaned by removing the tuples with null values. Entries only with Full-time employment status were selected to give a common ground on hours of work for better analysis. After initial data cleaning, we got 522 entries for the 1974 dataset (348 males and 173 females) and 977 entries for the 2018 dataset (507 males and 469 females).

Exploratory data analysis:

EDA was done on the selected subsets of the data to understand how the data is distributed over different feature categories. 1974 data included almost half as many females as males while the 2018 dataset had an almost equal representation of both genders. Upon plotting bar graphs to observe how the average income of males and females differs across different occupations, we notice that in 1974, except for professional occupations, either the females are not included in the dataset or the gender gap is high in other occupations. Whereas, in 2018 women seem to be included in almost all the domains considered, and the pay gap seems to have gone down in most of the fields, except for the ones like Construction and Production.



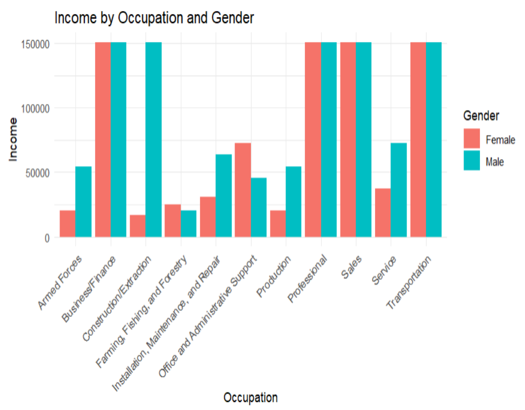


Fig 1: Income by occupation and gender in 1974 and 2018

Research questions:

After observing initial trends during EDA, we bring upon certain research questions that we would like to answer by doing regression analysis on the dataset

1. **Gender-Pay gap:** How has the gender pay gap changed since 1974? Has it gone down?
2. **Impact of education on different occupations:** How important is the level of education in different fields of occupation? Has this importance changed over time?
3. **Impact of education on the relationship between prestige score and income (Causal analysis):** How does education level affect the relationship between prestige score and real income?

Methods

Multiple Linear model was used to analyze the relationships between input and dependent variables. The Gaussian function was used as the link function because the dependent variable, income, is continuous.

Model for Research Question 1:

Formula: *realrinc ~ age + occrecode + educcat + gender*prestg10*

Model for research question 2:

Formula: *realrinc ~ educcat * occrecode*

Model for research question 3:

Formula: *realrinc ~ educcat * prestg10*

Results and analysis:

Model-1:	*brm (realrinc ~ age + occrecode + educcat + gender*prestg10)*				
Year 1974	**Estimate**	**Std Error**	**l-95% CI**	**U-95% CI**	**r-hat**
genderMale	11031.6	5998.37	-821.2	22477.8	1
Year 2018	**Estimate**	**Std Error**	**l-95% CI**	**U-95% CI**	**r-hat**
genderMale	-2123.09	6022.72	-13945.9	9713.06	1

Model-2:	*lm (realrinc ~ educcat * occrecode)*				
Year 2018	**Estimate**	**Std. Error**	**t-value**	**Pr(>|t|)**	
Less Than High School: Service	51509.3	15608	3.3	0.001	**
Graduate: Transportation	-53676	31302	-1.715	0.0242	.
High School: Transportation	-40015	12969	-3.085	0.0209	**
Junior College: Transportation	-54245	23156	-2.343	0.0193	*

Model -3:	*lm (realrinc ~ educcat * prestg10)*				
Year 1974	**Estimate**	**Std. Error**	**t-value**	**Pr(>|t|)**	
Graduate:prestg10	692.32	327.23	2.116	0.034851	*
High School:prestg10	508.75	217	2.344	0.019436	*
Junior College:prestg10	81.15	560.22	0.145	0.884879	
Less Than High School:prestg10	476.12	273.24	1.742	0.082029	.
Year 2018	**Estimate**	**Std. Error**	**t-value**	**Pr(>|t|)**	
Graduate:prestg10	282.76	244.62	1.156	0.248	
High School:prestg10	-62.58	176.81	-0.354	0.7235	
Junior College:prestg10	29.63	298.54	0.099	0.9209	
Less Than High School:prestg10	-213.2	304.79	-0.699	0.4844	

Fig2: Main columns of model estimate results highlighted for the three models used: Note the following for the models:

1. Model 1 made using brms: the sign of estimate for genderMale switches from +ve to -ve in 1974 and 2018.
2. Model 2 using lm: The interaction terms shown in the image are significant
3. Model 3 using lm: education-prestige score interaction terms have significance in 1974 unlike 2018.

Inferences from these tables are made in the results and discussion below. Full tables we got from R for these models are included as a part of Appendix C

Results of Model 1: *realrinc ~ age + occrecode + educcat + gender*prestg10*

Interaction term on gender was added to understand how gender affects the relationship between prestige score on income.

In 1974:

- Individuals in occupations such as Business/Finance, Sales, and Professional occupations tend to have higher incomes compared to other occupations.

- Higher education levels generally correspond to higher incomes, with Graduate education having the most significant positive effect.
- Being male is associated with higher income, with a positive effect estimate. However, the interaction between gender and prestige score (genderMale:prestg10) is relatively small.

In 2018, the estimated effects show some similarities but also notable differences compared to 1974:

- Occupations such as Business/Finance, Transportation, and Sales exhibit positive effects on income similar to 1974, although the magnitudes vary.
- Higher education levels, particularly Graduate education, continue to be associated with higher incomes.
- Interestingly, the effect of being male on income was negative in 2018, suggesting that females, on average, earn more than males. However, similar to 1974, the interaction between gender and prestige score (genderMale:prestg10) remains relatively small.

Summary on Gender: Comparing these results between 1974 and 2018 reveals significant shifts in the factors influencing income over time. While occupation and education continue to play crucial roles in determining income levels, the effect of gender appears to have reversed from a positive to a negative association with income.

Results of Model 2: *realrinc ~ educcat * occrecode*

To determine the order of occrecode (occupation) in which educcat (education level) influenced realrinc (real income) in 2018, we can examine the coefficients associated with the interaction terms between education level and occupation:

- The interaction term with the largest absolute coefficient is "educcatLess Than High School:occrecodeService" with an estimate of 51509.3. This suggests that among the occupations considered, individuals with less than a high school education working in service had the highest influence on real income in 2018.
- The order of influence decreases gradually as we move to occupations associated with lower absolute coefficient values.
- Some interaction terms have coefficients marked as "NA" due to singularities, indicating that certain combinations of education level and occupation may not have sufficient data for estimation.

- Based on the available data, we can observe a trend where higher education levels generally lead to higher income across different occupations, with some variations in the magnitude of this effect depending on the specific occupation.

Summary: The analysis suggests that in 2018, individuals with less than a high school education working in transportation had the highest influence on real income, followed by other occupations such as construction/extraction and farming, fishing, and forestry.

Results of Model 3: *realrinc ~ educcat * prestg10*

The regression results for the year 1974 indicate that the interaction between education level (educcat) and prestige score (prestg10) significantly influences real income (realrinc):

- This is evident from the statistically significant coefficients for the interaction terms, such as `educcatGraduate:prestg10`, `educcatHigh School:prestg10`, and `educcatLess Than High School:prestg10`.
- These significant coefficients imply that the effect of prestige score on real income varied across different levels of education in 1974. Specifically, the interaction terms show that the relationship between prestige score and real income differs for individuals with different education levels.
- Therefore, in 1974, education level acts as a **moderator** in the relationship between prestige score and real income.

In contrast, the regression results for the year **2018 do not show significant interaction** effects between education level and prestige score on real income:

- None of the interaction terms have statistically significant coefficients, indicating that the relationship between prestige score and real income does not vary significantly across different education levels in 2018.

This implies that the influence of education level on the relationship between prestige score and real income may have changed over time, with education level playing a more significant moderating role in 1974 compared to 2018.

10-fold cross-validation to compare the three models:

CV Results	**RMSE**	**R-squared**	**MAE**
Model-1	24983.54	0.2879351	15358.25

Model-2	27235.67	0.1726562	16773.04
Model-3	26992.78	0.1737815	16477.7

Table 1: Cross-validation results for considered 3 models

Model-1 which includes 5 predictors performs better than model-2 and 3 which include only two predictors each.

Discussion

By doing regression analysis on the datasets, after understanding the causal relationship among them, we were able to draw the following conclusions as solutions to the research questions that we started with:

1. **Gender-Pay gap**: Comparing 1974 and 2018 income results we see that the gender gap has gone down significantly and the effect of gender appears to have reversed from a positive to a negative association with income. Though this may not be true in all occupations, this certainly seems to be the case with the considered dataset. This can be attributed to the societal improvements in terms of gender equality and the changes in the occupational distribution of men and women that happened during the time period considered [3].
2. **Impact of education on different occupations:** The broader pattern observed in the data is that higher education levels tend to correlate with higher incomes across various occupations. Comparing the trends in 2018 to the trend in 1974, it's likely that the influence of education on income has become more pronounced over time, reflecting the increasing importance of education in accessing higher-paying jobs in modern economies.
3. **Impact of education on the relationship between prestige score and income:** In 1974, the effect of prestige score on real income is moderated or influenced by education level. In contrast, in 2018, education's influence on this relationship didn't seem to be statistically significant.

Even though the conclusions drawn from the available datasets show an improvement in problems like gender gap and education-pay gaps over time, we also need to take into consideration that the datasets from 1974 and 2018 are not equally distributed among different genders and different education levels. As a part of future work, some more analysis can be done on these datasets like how having kids impacts a woman's income and how marital status influences how much one earns.

References

1. NORC at the University of Chicago, "GSS Data Explorer," NORC at the University of Chicago. [Online]. Available: https://gssdataexplorer.norc.org/about. [Accessed: 19 Mar 2024].
2. V. Arel-Bundock, "gss_wages," Rdatasets Project, Vincent Arel-Bundock. [Online]. Available: https://vincentarelbundock.github.io/Rdatasets/doc/stevedata/gss_wages.html. [Accessed: 19 Mar 2024].
3. Statistics Canada, "Labour Market Indicators," Statistics Canada. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/75-004-m/75-004-m2019004-eng.htm. [Accessed: 21 Mar 2024].
4. LinkedIn, "What is the best way to use regression analysis to test hypotheses?," LinkedIn. [Online]. Available: https://www.linkedin.com/advice/1/what-best-way-use-regression-analysis-test-hypotheses-a08qf. [Accessed: 21 Mar 2024].
5. Theoretical Ecology Blog, "Mediators, Confounders, Colliders: A Crash Course in Causal Inference," Theoretical Ecology Blog. [Online]. Available: https://theoreticalecology.wordpress.com/2019/04/14/mediators-confounders-colliders-a-crash-course-in-causal-inference/. [Accessed: 22 Mar 2024].

Supplementary Information - Appendix

Contents:

A.Some of the causal relationships examined:

1974:

Gender as a Moderator in prestg10 to income relationship:

- In the first model, the interaction term `genderMale:prestg10` is included to test if gender moderates the relationship between `prestg10` and `realrinc`.
- The coefficient for the interaction term is 217.9 with a p-value of 0.10485, which is not statistically significant at conventional levels ($p > 0.05$). This suggests that there's no strong evidence to support the moderation effect of gender in the relationship between occupational prestige score and income.

Educcat as a Confounder:

- In the third model, `educcat` is included alongside `prestg10` to test if it acts as a confounder in the relationship between `prestg10` and `realrinc`.
- The coefficients for the different levels of education (`educcat`) are statistically significant for some categories, indicating that education level has a direct effect on income.
- However, the coefficient for `prestg10` remains statistically significant, suggesting that occupational prestige score still has an independent effect on income even after controlling for education level.

2018:

Let's interpret the results for each case:

Gender as a Moderator:

- The interaction term `genderMale:prestg10` is included to test if gender moderates the relationship between `prestg10` and `realrinc`.
- The coefficient for the interaction term is 335.94 with a p-value of 0.00838, which is statistically significant at conventional levels ($p < 0.05$). This suggests that gender moderates the relationship between occupational prestige score and income in the year 2018.

Educcat as a Confounder:

- In the third model, `educcat` is included alongside `prestg10` to test if it acts as a confounder in the relationship between `prestg10` and `realrinc`.
- The coefficients for the different levels of education (`educcat`) are statistically significant for some categories, indicating that education level has a direct effect on income.
- However, the coefficient for `prestg10` remains statistically significant, suggesting that occupational prestige score still has an independent effect on income even after controlling for education level.

B. The ordered list of `occrecode` categories based on the magnitude of the coefficient of the interaction term in 1974:

- Installation, Maintenance, and Repair
- Sales
- Business/Finance
- Professional
- Office and Administrative Support
- Production
- Construction/Extraction
- Service
- Farming, Fishing, and Forestry
- Transportation

2018:

- Service
- Installation, Maintenance, and Repair
- Transportation
- Construction/Extraction
- Production
- Farming, Fishing, and Forestry

- Professional, Sales, Business/Finance, Office and Administrative Support

C. Full result tables

C1. Hypothesis 1

1974 results:

```
Family: gaussian
  Links: mu = identity; sigma = identity
Formula: realrinc ~ age + occrecode + gender * prestg10
   Data: income_1974 (Number of observations: 521)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
                                          Estimate Est.Error  l-95% CI u-95%
CI Rhat
Intercept                                -10338.50   8697.93 -27355.65
7037.06 1.00
age                                         239.74     57.90    128.81
356.58 1.00
occrecodeBusinessDFinance                 18665.07   6066.26   6957.11
30286.24 1.00
occrecodeConstructionDExtraction           2947.83   6592.84 -10064.09
15517.02 1.00
occrecodeFarmingFishingandForestry        -5821.19  10929.90 -27760.78
15626.28 1.00
occrecodeInstallationMaintenanceandRepair  8115.22   6738.19  -4879.23
21655.93 1.00
occrecodeOfficeandAdministrativeSupport    8908.51   6230.08  -3215.18
21309.22 1.00
occrecodeProduction                        4685.52   6339.31  -7992.53
16754.06 1.00
occrecodeProfessional                     14715.00   5816.79   3447.37
26133.73 1.00
occrecodeSales                            15169.40   6502.36   2049.66
27818.87 1.00
occrecodeService                           2755.90   6399.55  -9706.63
15305.61 1.00
occrecodeTransportation                    5802.20   6875.71  -8169.98
18687.97 1.00
genderMale                                 6654.50   6115.97  -5428.64
18743.72 1.00
prestg10                                    181.02    127.75    -70.53
431.50 1.00
genderMale:prestg10                         194.70    132.93    -67.88
459.14 1.00
                                         Bulk_ESS Tail_ESS
```

```
Intercept                                    1178     1781
age                                          3824     2924
occrecodeBusinessDFinance                     968     1483
occrecodeConstructionDExtraction             1016     1354
occrecodeFarmingFishingandForestry           1907     2442
occrecodeInstallationMaintenanceandRepair    1124     1677
occrecodeOfficeandAdministrativeSupport      1004     1638
occrecodeProduction                           990     1528
occrecodeProfessional                        1033     1615
occrecodeSales                                997     1509
occrecodeService                             1025     1666
occrecodeTransportation                      1030     1627
genderMale                                   1766     2194
prestg10                                     2014     2505
genderMale:prestg10                          1831     2333

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
```

2018 Results:

```
Family: gaussian
  Links: mu = identity; sigma = identity
Formula: realrinc ~ age + occrecode + gender * prestg10
   Data: income_2018 (Number of observations: 976)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
                                           Estimate Est.Error  l-95% CI u-95%
CI Rhat
Intercept                                 -21028.81  13445.35 -46607.08
5669.31 1.01
age                                          377.96     64.84    250.12
503.63 1.00
occrecodeBusinessDFinance                  20948.65  12158.03  -2907.43
44203.73 1.01
occrecodeConstructionDExtraction            5900.20  12655.61 -18756.90
30242.64 1.01
occrecodeFarmingFishingandForestry         -2006.20  15066.44 -31246.12
27209.60 1.00
occrecodeInstallationMaintenanceandRepair   3682.88  12580.49 -21587.69
27953.98 1.01
occrecodeOfficeandAdministrativeSupport     4253.65  12367.78 -20109.59
27722.07 1.01
occrecodeProduction                         1542.54  12624.10 -23138.50
25409.40 1.01
occrecodeProfessional                       5101.17  12119.39 -18958.21
28029.85 1.01
occrecodeSales                             10970.85  12479.07 -13209.02
34643.50 1.01
occrecodeService                            -751.04  12266.09 -24955.16
22816.25 1.01
```

```
occrecodeTransportation                       17507.54  12707.07  -7392.26
41490.46 1.01
genderMale                                    -2824.84   6284.29 -14720.95
9922.51 1.00
prestg10                                        445.54    110.13    227.63
656.92 1.00
genderMale:prestg10                             279.45    126.82     25.90
519.65 1.00
                                           Bulk_ESS Tail_ESS
Intercept                                       750     1453
age                                            4384     2721
occrecodeBusinessDFinance                       690     1276
occrecodeConstructionDExtraction                719     1399
occrecodeFarmingFishingandForestry              866     1723
occrecodeInstallationMaintenanceandRepair       703     1384
occrecodeOfficeandAdministrativeSupport         671     1286
occrecodeProduction                             719     1388
occrecodeProfessional                           705     1252
occrecodeSales                                  705     1374
occrecodeService                                674     1300
occrecodeTransportation                         727     1332
genderMale                                     2669     2378
prestg10                                       3092     2759
genderMale:prestg10                            2554     2400

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 26182.48    595.58 25040.53 27421.38 1.00     4322     2513
```

C2. Hypothesis 2 results:

```
Call:
lm(formula = realrinc ~ educcat * occrecode, data = income_2018)

Residuals:
   Min      1Q Median      3Q    Max
-64884 -12838  -3794   4921 126292

Coefficients: (7 not defined because of singularities)
                                                   Estimate Std. Error
t value Pr(>|t|)
(Intercept)                                           27575       5664
4.869 1.32e-06 ***
educcatG                                             -12911      13162
-0.981 0.326866
educcatHS                                             -9918       6422
-1.544 0.122853
```

```
educcatJunior College                              -12115        9082
-1.334 0.182565
educcatL                                           -58529       12454
-4.699 3.00e-06 ***
occrecodeArmed Forces                               -4024       19620
-0.205 0.837540
occrecodeBusiness                                   21619        6504
3.324 0.000923 ***
occrecodeConstruction                               47745       16350
2.920 0.003583 **
occrecodeFarming                                    -2605       27163
-0.096 0.923609
occrecodeMaintenance                                -4894       12235
-0.400 0.689243
occrecodeProduction                                 -3882       14440
-0.269 0.788106
occrecodeProfessional                                7069        6268
1.128 0.259644
occrecodeSales                                      12137        8208
1.479 0.139540
occrecodeService                                   -10592        9810
-1.080 0.280575
occrecodeTransportation                             49794       11528
4.319 1.73e-05 ***
educcatG:occrecodeArmed Forces                      43840       35098
1.249 0.211952
educcatHS:occrecodeArmed Forces                      6797       33164
0.205 0.837663
educcatJunior College:occrecodeArmed Forces         13534       33781
0.401 0.688778
educcatL:occrecodeArmed Forces                         NA          NA
NA       NA
educcatG:occrecodeBusiness                          32181       14312
2.248 0.024780 *
educcatHS:occrecodeBusiness                         -1188        8154
-0.146 0.884210
educcatJunior College:occrecodeBusiness             -9946       11697
-0.850 0.395399
educcatL:occrecodeBusiness                          21592       20015
1.079 0.280954
educcatG:occrecodeConstruction                         NA          NA
NA       NA
educcatHS:occrecodeConstruction                    -38837       17259
-2.250 0.024669 *
educcatJunior College:occrecodeConstruction        -35682       25896
-1.378 0.168579
educcatL:occrecodeConstruction                      -2821       21320
-0.132 0.894762
```

```
educcatG:occrecodeFarming                                    NA          NA
NA       NA
educcatHS:occrecodeFarming                               -12139       31341
-0.387 0.698610
educcatJunior College:occrecodeFarming                       NA          NA
NA       NA
educcatL:occrecodeFarming                                 43615       31655
1.378 0.168589
educcatG:occrecodeMaintenance                                NA          NA
NA       NA
educcatHS:occrecodeMaintenance                            12960       13678
0.948 0.343624
educcatJunior College:occrecodeMaintenance                19795       16453
1.203 0.229222
educcatL:occrecodeMaintenance                             60817       31281
1.944 0.052168 .
educcatG:occrecodeProduction                                 NA          NA
NA       NA
educcatHS:occrecodeProduction                              9385       15371
0.611 0.541634
educcatJunior College:occrecodeProduction                 -6300       24735
-0.255 0.798998
educcatL:occrecodeProduction                              44645       20794
2.147 0.032047 *
educcatG:occrecodeProfessional                            23793       13745
1.731 0.083782 .
educcatHS:occrecodeProfessional                           -2935        7813
-0.376 0.707313
educcatJunior College:occrecodeProfessional                1091       11353
0.096 0.923462
educcatL:occrecodeProfessional                            45260       19939
2.270 0.023442 *
educcatG:occrecodeSales                                   13680       21066
0.649 0.516244
educcatHS:occrecodeSales                                  -5036        9704
-0.519 0.603955
educcatJunior College:occrecodeSales                      -6884       17153
-0.401 0.688266
educcatL:occrecodeSales                                   33117       18209
1.819 0.069271 .
educcatG:occrecodeService                                 33950       20343
1.669 0.095476 .
educcatHS:occrecodeService                                 5200       10626
0.489 0.624733
educcatJunior College:occrecodeService                    11830       13812
0.857 0.391928
educcatL:occrecodeService                                 51509       15608
3.300 0.001003 **
```

```
educcatG:occrecodeTransportation                   -53676      31302
-1.715 0.086719 .
educcatHS:occrecodeTransportation                  -40015      12969
-3.085 0.002093 **
educcatJunior College:occrecodeTransportation   -54245      23156
-2.343 0.019360 *
educcatL:occrecodeTransportation                       NA          NA
NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26570 on 928 degrees of freedom
Multiple R-squared:  0.244,     Adjusted R-squared:  0.2057
F-statistic: 6.372 on 47 and 928 DF,  p-value: < 2.2e-16
```

C3. Hypothesis 3 Results:

```
Call:
lm(formula = realrinc ~ educcat * prestg10, data = income_1974)

Residuals:
   Min     1Q Median     3Q    Max 
-46407 -11215  -3575   7278  75308 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     38721.13   10395.98   3.725 0.000217 ***
educcatG                       -26892.51   19697.78  -1.365 0.172773    
educcatHS                      -31762.47   11377.15  -2.792 0.005438 ** 
educcatJunior College          -14796.24   26452.01  -0.559 0.576160    
educcatL                       -32495.42   12707.56  -2.557 0.010840 *  
prestg10                          -93.64     190.06  -0.493 0.622458    
educcatG:prestg10                 692.32     327.23   2.116 0.034851 *  
educcatHS:prestg10                508.75     217.00   2.344 0.019436 *  
educcatJunior College:prestg10     81.15     560.22   0.145 0.884879    
educcatL:prestg10                 476.12     273.24   1.742 0.082029 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18850 on 511 degrees of freedom
Multiple R-squared:  0.1879,   Adjusted R-squared:  0.1736 
F-statistic: 13.14 on 9 and 511 DF,  p-value: < 2.2e-16


Call:
lm(formula = realrinc ~ educcat * prestg10, data = income_2018)

Residuals:
   Min     1Q Median     3Q    Max 
-50615 -13528  -5111   5155 135452 
```

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     18589.09    7468.52   2.489   0.0130 *
educcatG                        -8755.61   14319.35  -0.611   0.5410
educcatHS                      -10125.22    8854.53  -1.144   0.2531
educcatJunior College          -16025.27   14551.25  -1.101   0.2710
educcatL                       -11838.42   12729.18  -0.930   0.3526
prestg10                          386.90     138.44   2.795   0.0053 **
educcatG:prestg10                 282.76     244.62   1.156   0.2480
educcatHS:prestg10                -62.58     176.81  -0.354   0.7235
educcatJunior College:prestg10     29.63     298.54   0.099   0.9209
educcatL:prestg10                -213.20     304.79  -0.699   0.4844
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27280 on 966 degrees of freedom
Multiple R-squared:  0.1704,   Adjusted R-squared:  0.1627
F-statistic: 22.05 on 9 and 966 DF,  p-value: < 2.2e-16
```

C4. Hypothesis 2 results using brms:

```
Call:
lm(formula = realrinc ~ educcat * occrecode, data = income_2018)

Residuals:
   Min     1Q Median     3Q    Max
-64884 -12838  -3794   4921 126292

Coefficients: (7 not defined because of singularities)
                                              Estimate Std. Error t value
Pr(>|t|)
(Intercept)                                      27575       5664   4.869
1.32e-06 ***
educcatG                                        -12911      13162  -0.981
0.326866
educcatHS                                        -9918       6422  -1.544
0.122853
educcatJunior College                           -12115       9082  -1.334
0.182565
educcatL                                        -58529      12454  -4.699
3.00e-06 ***
occrecodeArmed Forces                            -4024      19620  -0.205
0.837540
occrecodeBusiness                                21619       6504   3.324
0.000923 ***
occrecodeConstruction                            47745      16350   2.920
0.003583 **
occrecodeFarming                                 -2605      27163  -0.096
0.923609
```

```
occrecodeMaintenance                            -4894      12235  -0.400
0.689243
occrecodeProduction                             -3882      14440  -0.269
0.788106
occrecodeProfessional                            7069       6268   1.128
0.259644
occrecodeSales                                  12137       8208   1.479
0.139540
occrecodeService                               -10592       9810  -1.080
0.280575
occrecodeTransportation                         49794      11528   4.319
1.73e-05 ***
educcatG:occrecodeArmed Forces                  43840      35098   1.249
0.211952
educcatHS:occrecodeArmed Forces                  6797      33164   0.205
0.837663
educcatJunior College:occrecodeArmed Forces     13534      33781   0.401
0.688778
educcatL:occrecodeArmed Forces                     NA         NA      NA
NA
educcatG:occrecodeBusiness                      32181      14312   2.248
0.024780 *
educcatHS:occrecodeBusiness                     -1188       8154  -0.146
0.884210
educcatJunior College:occrecodeBusiness         -9946      11697  -0.850
0.395399
educcatL:occrecodeBusiness                      21592      20015   1.079
0.280954
educcatG:occrecodeConstruction                     NA         NA      NA
NA
educcatHS:occrecodeConstruction                -38837      17259  -2.250
0.024669 *
educcatJunior College:occrecodeConstruction    -35682      25896  -1.378
0.168579
educcatL:occrecodeConstruction                  -2821      21320  -0.132
0.894762
educcatG:occrecodeFarming                          NA         NA      NA
NA
educcatHS:occrecodeFarming                     -12139      31341  -0.387
0.698610
educcatJunior College:occrecodeFarming             NA         NA      NA
NA
educcatL:occrecodeFarming                       43615      31655   1.378
0.168589
educcatG:occrecodeMaintenance                      NA         NA      NA
NA
educcatHS:occrecodeMaintenance                  12960      13678   0.948
0.343624
educcatJunior College:occrecodeMaintenance      19795      16453   1.203
0.229222
educcatL:occrecodeMaintenance                   60817      31281   1.944
0.052168 .
```

```
educcatG:occrecodeProduction                            NA         NA       NA
NA
educcatHS:occrecodeProduction                         9385      15371    0.611
0.541634
educcatJunior College:occrecodeProduction            -6300      24735   -0.255
0.798998
educcatL:occrecodeProduction                         44645      20794    2.147
0.032047 *
educcatG:occrecodeProfessional                       23793      13745    1.731
0.083782 .
educcatHS:occrecodeProfessional                      -2935       7813   -0.376
0.707313
educcatJunior College:occrecodeProfessional           1091      11353    0.096
0.923462
educcatL:occrecodeProfessional                       45260      19939    2.270
0.023442 *
educcatG:occrecodeSales                              13680      21066    0.649
0.516244
educcatHS:occrecodeSales                             -5036       9704   -0.519
0.603955
educcatJunior College:occrecodeSales                 -6884      17153   -0.401
0.688266
educcatL:occrecodeSales                              33117      18209    1.819
0.069271 .
educcatG:occrecodeService                            33950      20343    1.669
0.095476 .
educcatHS:occrecodeService                            5200      10626    0.489
0.624733
educcatJunior College:occrecodeService               11830      13812    0.857
0.391928
educcatL:occrecodeService                            51509      15608    3.300
0.001003 **
educcatG:occrecodeTransportation                    -53676      31302   -1.715
0.086719 .
educcatHS:occrecodeTransportation                   -40015      12969   -3.085
0.002093 **
educcatJunior College:occrecodeTransportation       -54245      23156   -2.343
0.019360 *
educcatL:occrecodeTransportation                        NA         NA       NA
NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26570 on 928 degrees of freedom
Multiple R-squared:  0.244,    Adjusted R-squared:  0.2057
F-statistic: 6.372 on 47 and 928 DF,  p-value: < 2.2e-16
```

C5. Hypothesis 3 using brms:

1974:

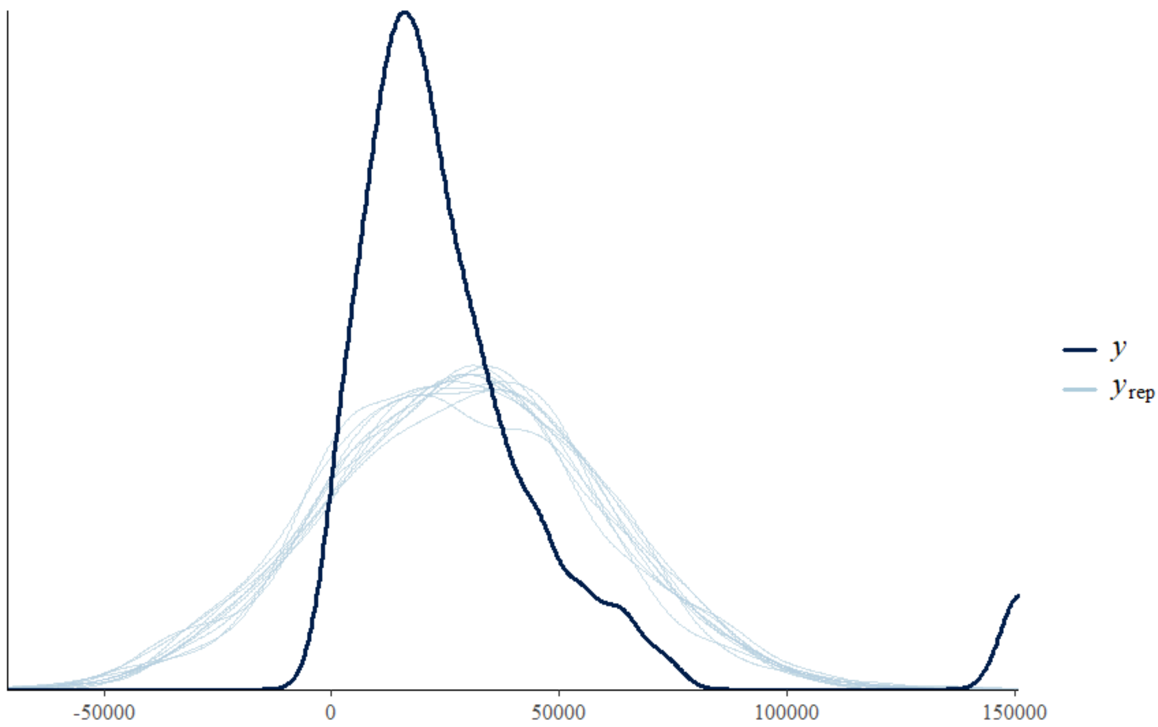
```
Chain 4:
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: realrinc ~ educcat * prestg10
   Data: income_1974 (Number of observations: 521)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
                               Estimate Est.Error  l-95% CI u-95% CI Rhat
Bulk_ESS Tail_ESS
Intercept                      38747.30  10556.01  18046.07 59682.08 1.00
1446     1821
educcatG                      -27383.07  19867.41 -67569.27 10906.26 1.00
1987     2343
educcatHS                     -31747.80  11552.57 -54318.68 -9537.76 1.00
1494     1832
educcatJuniorCollege          -14793.81  25898.78 -65303.09 34572.37 1.00
2290     2586
educcatL                      -32662.26  12730.80 -57478.03 -7887.87 1.00
1572     2135
prestg10                         -94.88    192.86   -472.17   280.73 1.00
1404     1813
educcatG:prestg10                701.34    329.98     71.72  1363.67 1.00
1888     1975
educcatHS:prestg10               509.12    219.75     93.30   927.55 1.00
1512     1923
educcatJuniorCollege:prestg10     83.88    546.63   -969.46  1162.16 1.00
2426     2606
educcatL:prestg10                480.49    274.30    -45.94  1025.91 1.00
1763     2188

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 18851.89    589.55 17740.10 20021.14 1.00     3601     2632

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

2018:

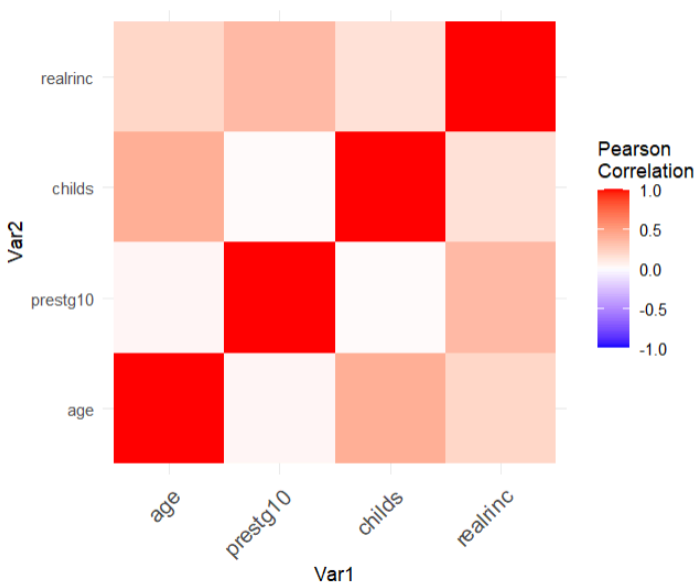
```
Chain 4:
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: realrinc ~ educcat * prestg10
   Data: income_2018 (Number of observations: 976)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
                               Estimate Est.Error  l-95% CI u-95% CI Rhat
Bulk_ESS Tail_ESS
```

```
Intercept                      18680.01   7280.57   4402.93 33127.03 1.01
1323     1941
educcatG                       -8657.60  14112.88 -35774.12 19550.29 1.00
2040     2560
educcatHS                     -10226.13   8725.81 -27580.23  6771.84 1.01
1403     2094
educcatJuniorCollege          -16835.87  14273.72 -45006.82 11660.82 1.00
1800     2008
educcatL                      -12046.11  12720.95 -36196.79 13347.21 1.00
1733     2345
prestg10                        384.33    134.91    117.98   642.23 1.01
1323     1888
educcatG:prestg10               281.74    240.95   -191.15   747.11 1.00
1879     2375
educcatHS:prestg10              -60.60    174.91   -389.99   287.92 1.01
1500     2075
educcatJuniorCollege:prestg10    45.32    292.19   -538.52   617.90 1.00
1898     2102
educcatL:prestg10              -209.77    303.10   -806.38   370.99 1.00
1992     2651

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 27287.32    638.71 26097.79 28562.23 1.00     4616     2932

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

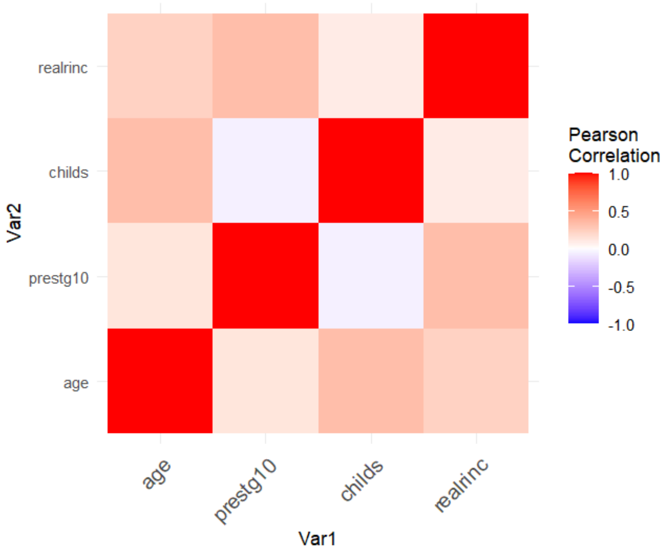
D. pp_check() for model 1 on 2018 dataset



E. **Correlation matrix for numerical columns of 1974 dataset**



F. **Correlation matrix for numerical columns of 2018 dataset**


realrinc
childs
prestg10
age
Var2
age
prestg10
childs
realrinc
Var1
Pearson
Correlation
1.0
0.5
0.0
-0.5
-1.0