## Introduction

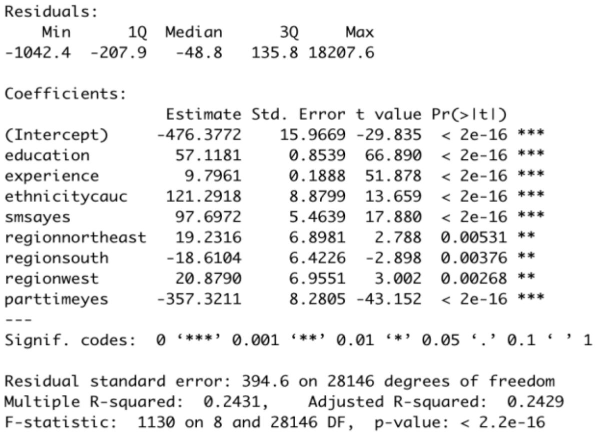
According to the ‘Economic Report of the President 1988’, the US saw the inflation rate move gradually downward from the 4% range, while the average unemployment rate was at its lowest compared to the past 13 years. During this year, minimum wage also continued to be of major issue, however the trend in legislative and administrative actions was to increase the wage above $3.35 per hour (Nelson, 1989); however, this was not guaranteed for every state. Some states such as California, Connecticut and Virgin Islands reached minimum wage of $4.25, making it harder to maintain wage standards across the United States (Nelson, 1989).

In this project the dataset concerning US wages in 1988 was analyzed, consisting of numerical variables - such as education and experience, categorical variables - such as region and ethnicity. Analyzing the effects of variables on wage will help us understand what factors affected the wage most in 1988 and shaped the American economic system as a function of other variables in the context of that year. Particularly, the report will aim to answer how factors such as ethnicity and region impacted an average US citizen’s earnings in 1988 and if there was any statistically significant disparity. The dataset we are using is a pre-loaded dataset available in R, collected by the US Census Bureau in 1988, called CPS1988. Due to this, the data did not require any preprocessing or cleaning; however, ‘id’ values were dropped as they were intended for indexing purposes and did not add any statistical value to the dataset.

## Testing Regression Models

*Model0: (wage ~ .)*

At first look, the intercept, education, experience, ethnicity, smsayes, regionwest, regionnortheast, regionwest, and parttimeyes all appear significant (Fig1). They have low p-values and the t-scores are strong as well. However, the Adjusted $R^2$ of 0.2429 here does not indicate a good fit of the model and implies that there may be relatively larger residual terms. The individual p-values of the coefficients cannot be relied upon alone, hence, the collective F-statistic of 1130 with a small p-value does imply that some or all predictors may have strong association. However, the regions in this model have a less significant p-value (especially regionnortheast with p = 0.00531) than other variables so assessing more models that are subsets of this Model0 is essential. After running stepAIC using forward, backward, and both approaches together, Model0 had the lowest AIC value, followed by the model with only region removed.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-1042.4  -207.9   -48.8   135.8 18207.6 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -476.3772    15.9669 -29.835  < 2e-16 ***
education         57.1181     0.8539  66.890  < 2e-16 ***
experience         9.7961     0.1888  51.878  < 2e-16 ***
ethnicitycauc    121.2918     8.8799  13.659  < 2e-16 ***
smsayes           97.6972     5.4639  17.880  < 2e-16 ***
regionnortheast   19.2316     6.8981   2.788  0.00531 ** 
regionsouth      -18.6104     6.4226  -2.898  0.00376 ** 
regionwest        20.8790     6.9551   3.002  0.00268 ** 
parttimeyes     -357.3211     8.2805 -43.152  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 394.6 on 28146 degrees of freedom
Multiple R-squared:  0.2431,	Adjusted R-squared:  0.2429 
F-statistic:  1130 on 8 and 28146 DF,  p-value: < 2.2e-16
```

Fig1: Summary of Model0

*Model1: (wage ~. -region)*

After removing the region variable from Model0, the F-statistic increased to 1796 and the p-value also remained very low. Moreover, the individual coefficients and t-values in this model did not fluctuate much whilst p-values of the coefficients remained near 0. The Adjusted $R^2$ dropped to 0.2417, which is not very drastic, given that the absence of the categorical variable region simplifies the model's interpretability. The ANOVA results between Model0 and Model1 returned a very small p-value, implying that the variance in the two models is statistically significant - thus rejecting the null hypothesis in favour of the more complex model. Looking at the trade-off between choosing a simpler model that has a higher association between the response and predictors (Model1) and a more complex model that has a statistically significant and a slightly better fit (Model0), further exploration of Model0 may show how the linear regression assumptions stand.

Summary of the limitations from the analysis of assumptions in linear regression for Model0:

- The model has a low $R^2$.

-Although Education satisfied the linearity with wage, Experience as predictor follows quadratic curve shape when plotted with wage (Appendix 1). The remaining variables were categorical.

-The residuals do not follow a normal distribution as the points on Q-Q plot deviate upwards in the end (Fig2). Moreover, the wage is highly skewed as seen in the histogram (Appendix 2).

-There is heterodescacity seen in the fitted values vs residual plots - as the fitted values increase, so does the magnitude of residuals.

-We have many outliers (exceeding the range of -3 and 3 standard errors) and high leverage points (increasing threshold of 0.000213).

Otherwise, the observations overall are independent; residuals and independent variables (Education and Experience) do not have a recognisable correlation; the VIF values for all the variables were very close to

1, eliminating the risk of multicollinearity; and no observations fall outside the 0.5 level of Cook's Distance.

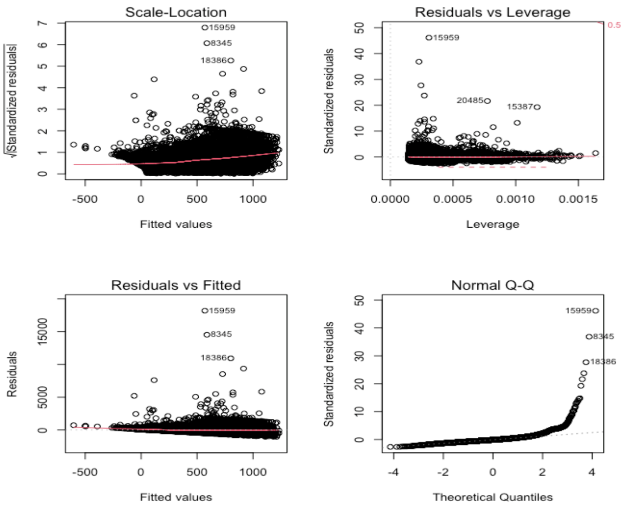


Fig2: Model1 plot

## Proposed Regression Model

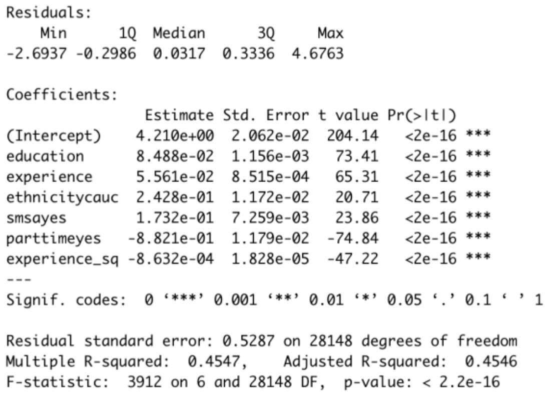
After testing different regression models (Model0 and Model1), the non-linearity between experience and wage was reduced by transforming the experience variable into its squared form to add to the model. Then, to curtail the heterodescacity, high magnitudes of outliers, and the non-normal distribution of the residuals, the wage was log-transformed.

*Model2: (wage_log ~ education + experience + experience$^2$ + region + parttime + ethnicity + smsa)*

The model's Adjusted R$^2$ almost doubled to 0.457.All the coefficients are statistically significant other than regionwest (p-value = 0.564). This increased R$^2$ at the cost of a less significant independent variable is not worth the improvement. Therefore, a more parsimonious model without the region was evaluated.

*Model3: (wage_log ~ education + experience + experience$^2$ + parttime + ethnicity + smsa)*

After dropping the region variable, there was a slight drop in the Adjusted R$^2$ to 0.4546 (Fig3). This indicates that removing the region still maintains a good fit,  the other predictors will have a strong association, and a simpler model will enhance interpretability. Hence, Model3 is now assessed on the summary statistics and assumptions as the final model.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.6937 -0.2986  0.0317  0.3336  4.6763 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     4.210e+00  2.062e-02  204.14   <2e-16 ***
education       8.488e-02  1.156e-03   73.41   <2e-16 ***
experience      5.561e-02  8.515e-04   65.31   <2e-16 ***
ethnicitycauc   2.428e-01  1.172e-02   20.71   <2e-16 ***
smsayes         1.732e-01  7.259e-03   23.86   <2e-16 ***
parttimeyes    -8.821e-01  1.179e-02  -74.84   <2e-16 ***
experience_sq -8.632e-04  1.828e-05  -47.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5287 on 28148 degrees of freedom
Multiple R-squared:  0.4547,	Adjusted R-squared:  0.4546 
F-statistic:  3912 on 6 and 28148 DF,  p-value: < 2.2e-16
```

Fig3: Summary of Model3

## Results and Interpretation

**Intercept** is the expected mean value of wage when all variables are 0. In our model, the mean wage would be 4.2 (log form) or $.23 (Fig3).

**Coefficients** are the estimates of wage and describe the relationship between the predictors and the response variable.

Education: 1 year increase leads to on average 0.084 ($11.78) increase in wage.

Experience: 1 year increase leads to on average 0.056 ($19.98) increase in wage.

Ethnicity: Caucasion makes on average 0.024 ($4.12) more than an AfricanAmerican

SMSA: A resident of SMSA makes on average 0.017 ($5.77) more than a non-resident.

Parttime: A part-time worker on average earns 0.882 ($1.133) less than a full time worker.

**Standard error** shows how various sample estimated values of a statistic vary from the population value. The largest standard error is for ethnicity at 0.01172 and smallest standard error is for experience at 0.0008515.

**T-value** measures the size of the difference relative to the variation on the data. The bigger the magnitude, the higher the chances of rejecting the null hypothesis. Parttime (-74.84) and Education (73.41) have the largest magnitude.

**P-value** higher values suggest that changes in wage are not associated with the coefficients - Model3 coefficients are all very close to zero.

**RSE** the smaller the value, the closer to the regression line. According to the model, actual wage will deviate from the true regression line by approximately $0.53 on average.

**$R^2$** value implies that the set of predictors in the model can explain about 45% of the variance in Log(Wage) and therefore the wage itself. High $R^2$ is not sufficient by itself to judge a model.

**F-statistic** value is used to assess significance in multiple regression, and compares the hypothesis. Our model's F-test value of 3912 implies that the model with independent variables is better than the one without. The low p-value associated with F-statistics confirms that this isn't by coincidence.

**Confidence Intervals** The narrower the interval, the more reliable is the estimate of the mean of the future values. For Model3, the confidence intervals are very narrow and all values fit within the range, making estimates more dependable (Appendix 3).

From the above statistical analysis, the null hypothesis that the variables have 0 impact on the wage can be rejected in the favour of the alternative hypothesis that they have a non-zero impact.

## Model Limitations

Wage and Experience$^2$ share a more visibly linear relation, especially because all experience values are positive. Log(Wage) and Experience$^2$ do not have a quadratic relationship either. (Appendix 4)

Log(Wage) although became more linear but the skewness still persists (Appendix 5). The Q-Q plot (Fig4) has points diverting away from the line on both the ends but not as far off as for Model0. Although many outliers still exist, most standardised residuals are within the range of -5 and 5 but in model0 they were further away (many exceeding 10).

The log transformation rendered constant variance of residuals for part of the fitted values, but the variance still differs as the magnitude of the fitted values increases. However, heterodescacity seems much better than Model0 in which variance in residuals increases more sharply.

The high leverage points are still there as many exceed the accepted value of 0.000213 [(p+1)/n = 5+1/28155]. However, as previously seen, the magnitude of standardised residuals has dropped overall. There still are some points that have high residuals and high leverage statistics.

Lastly, the model does not account for interactions between categorical and/or continuous variables. For example, how an individual from SMSA may have a different education than someone from outside; same for experience. The model fails to appreciate how these granular changes can affect its prediction of wages.

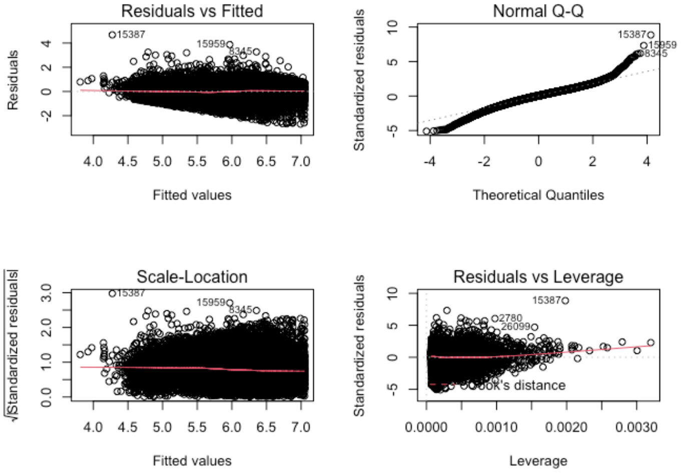


Fig4: Model3 plot

## Conclusion

The analyses showed that region is not a significant predictor for wages whereas the regression model does explain the relationship between education and experience and wages in the American economic standard in 1988 - the more education and experience one has, increases the chances of earning more (Alsulami, 2018). Part time employment intuitively and evidently from the model will pay less compared to a full time one. As cited previously, not all states had equal wages, which is corroborated by our analyses of SMSA.

Moreover, the wages were highly skewed at that time and there seemed to exist various concentrations of earnings at different levels. The earning distribution table from the ‘Earnings inequality accelerates in the 1980’s’ article (Appendix 6) and understand that a white individual earns more than that compared to a black individual. The model also explains this disparity in wages stemming from ethnicity. The importance of meeting the linear regression assumptions were appreciated through the analyses; especially in the context of transforming certain variables to attain linearity, diminish variance, and enhance the explanatory power of the regression; and of determining whether linear regression is indeed the correct technique for the given dataset. It was also observed that $R^2$ is not the absolute metric as it is also important to factor in the quality of association that predictors offer - more particularly to drop predictors in favour of interpretability. It is recommended that the analysis may be improved using interaction terms given that the dataset contains many categorical variables or perhaps deploying a technique other than linear regression that amplifies the value of such variables.

# References

President & Council of Economic Advisers, Economic Report of the President., 1988. *FRAZER, Federal Reserve Bank of St. Louis*. Available at: https://fraser.stlouisfed.org/title/economic-report-president-45/1988-8159 [Accessed November 13, 2021].

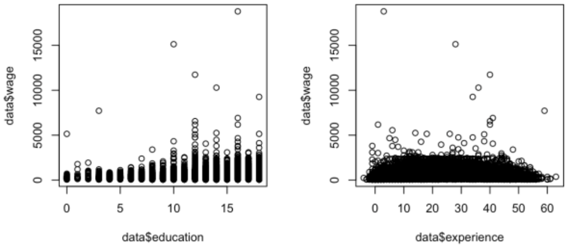
Nelson, R.R., 1989. Sate labor legislation enacted in 1988. *Monthly Labor Review*, *112*(1), 40–58. Available at: https://www.jstor.org/stable/41843200 [Accessed November 13, 2021].

Ryscavage, P., & Henle, P., 1990. Earnings inequality accelerates in the 1980's. *Monthly Labor Review*, *113*(12), 3–16. Available at: http://www.jstor.org/stable/41843401 [Accessed November 13, 2021]

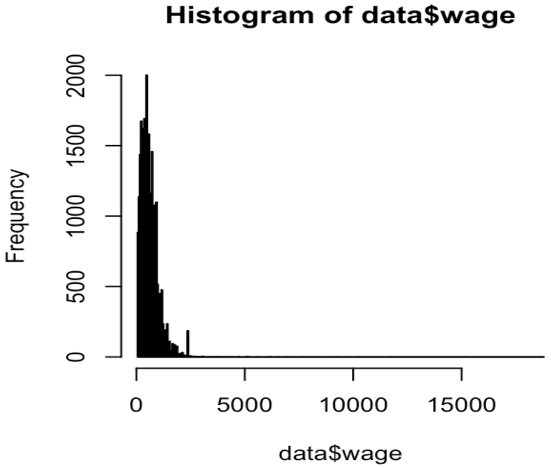
Alsulami, H., 2018. The effect of education and experience on wages: The case study of saudi arabia. *American Journal of Industrial and Business Management*. Available at: https://www.scirp.org/journal/paperinformation.aspx?paperid=81882 [Accessed November 14, 2021].

# Appendix

1. Education and Experience plotted against Wage.



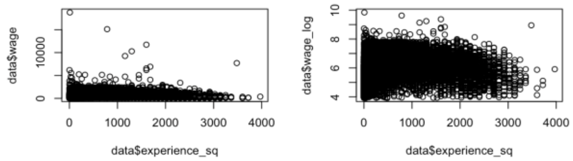
2. Histogram of Wage.



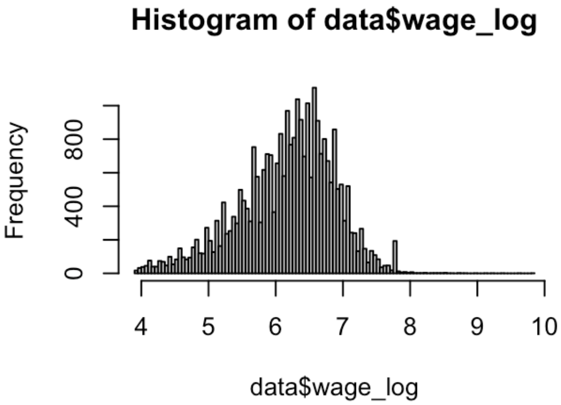
3. Model3 confidence intervals.

```
                    2.5 %        97.5 %
(Intercept)     4.1691424371  4.2499783928
education       0.0826154792  0.0871483851
experience      0.0539453946  0.0572835509
ethnicitycauc   0.2197816294  0.2657230038
smsayes         0.1589844418  0.1874384537
parttimeyes    -0.9051536920 -0.8589513576
experience_sq  -0.0008989846 -0.0008273319
```

4. Experience$^2$ against Log(Wage) and Wage.



5. Histogram of Log(Wage)



6. Distribution of earnings, from 'Earnings inequality accelerates in the 1980's' article.

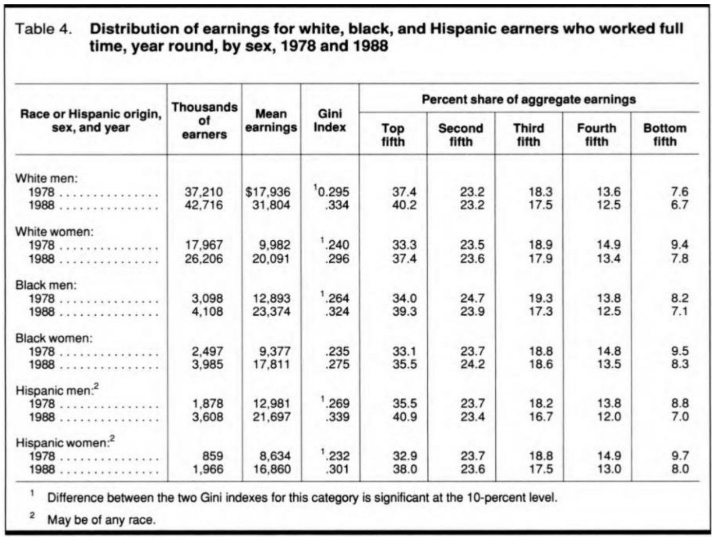


Table 4. **Distribution of earnings for white, black, and Hispanic earners who worked full time, year round, by sex, 1978 and 1988**

| Race or Hispanic origin, sex, and year | Thousands of earners | Mean earnings | Gini Index | Percent share of aggregate earnings | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Top fifth | Second fifth | Third fifth | Fourth fifth | Bottom fifth |
| White men: | | | | | | | | |
| 1978 . . . . . . . . . . . . . . . | 37,210 | $17,936 | [1]0.295 | 37.4 | 23.2 | 18.3 | 13.6 | 7.6 |
| 1988 . . . . . . . . . . . . . . . | 42,716 | 31,804 | .334 | 40.2 | 23.2 | 17.5 | 12.5 | 6.7 |
| White women: | | | | | | | | |
| 1978 . . . . . . . . . . . . . . . | 17,967 | 9,982 | [1].240 | 33.3 | 23.5 | 18.9 | 14.9 | 9.4 |
| 1988 . . . . . . . . . . . . . . . | 26,206 | 20,091 | .296 | 37.4 | 23.6 | 17.9 | 13.4 | 7.8 |
| Black men: | | | | | | | | |
| 1978 . . . . . . . . . . . . . . . | 3,098 | 12,893 | [1].264 | 34.0 | 24.7 | 19.3 | 13.8 | 8.2 |
| 1988 . . . . . . . . . . . . . . . | 4,108 | 23,374 | .324 | 39.3 | 23.9 | 17.3 | 12.5 | 7.1 |
| Black women: | | | | | | | | |
| 1978 . . . . . . . . . . . . . . . | 2,497 | 9,377 | .235 | 33.1 | 23.7 | 18.8 | 14.8 | 9.5 |
| 1988 . . . . . . . . . . . . . . . | 3,985 | 17,811 | .275 | 35.5 | 24.2 | 18.6 | 13.5 | 8.3 |
| Hispanic men:[2] | | | | | | | | |
| 1978 . . . . . . . . . . . . . . . | 1,878 | 12,981 | [1].269 | 35.5 | 23.7 | 18.2 | 13.8 | 8.8 |
| 1988 . . . . . . . . . . . . . . . | 3,608 | 21,697 | .339 | 40.9 | 23.4 | 16.7 | 12.0 | 7.0 |
| Hispanic women:[2] | | | | | | | | |
| 1978 . . . . . . . . . . . . . . . | 859 | 8,634 | [1].232 | 32.9 | 23.7 | 18.8 | 14.9 | 9.7 |
| 1988 . . . . . . . . . . . . . . . | 1,966 | 16,860 | .301 | 38.0 | 23.6 | 17.5 | 13.0 | 8.0 |

[1] Difference between the two Gini indexes for this category is significant at the 10-percent level.

[2] May be of any race.