CASS
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON
EST 1894

Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Halios Georgios 2. Haw, Kar Whing 3. Sun, Xuehui	4. Turki, Yasmine 5. Yi, Jinze (Eric) 6. Yuan, Ziyun (Isabella) **GROUP NUMBER:** 3
MSc in: Business Analytics	
Module Code: SMM 637	
Module Title: Quantitative Methods	
Lecturer: Radice, Rosalba	**Submission Date:** 8/12/2020

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: %

Table of content

Q1: Justification of the chosen regression model specification

In this report, we are examining the effects of various variables on wage in the US using GLM. All the tables and figures mentioned in this report can be found in the appendix. Wage is a positively continuous data (Figure 1) and by using the minimum function in R we discovered that the minimum value of the wage variable is 50.05, which is greater than 0, therefore we have decided to use a Gamma distribution with a log-link function. By employing the log-link function, this allows us to exponentiate the right-hand side of the equation, which are the explanatory variables. Thus, making sure that our estimates will always be a positive value. There is no need for an offset variable because the wage for all the observations in the data are all measured as weekly data.

The fitted model can be written as:

$$wage_i \sim Gamma\ (\ E(wage_i)\ ,\ \ )$$

$$\begin{aligned} log(\ E(wage_i)\ ) \ = \ & \beta_0 + \beta_1 Education_i + \beta_2 Experience_i + \beta_3 EthnicityCauc_i + \beta_4 Smsa_i + \\ & \beta_5 RegionNorthEast_i + \beta_6 RegionSouth_i + \beta_7 RegionWest_i + \beta_8 Parttime_i \end{aligned}$$

Note: A descriptive summary table explaining the variables can be found in the appendix (Table 1).

Now we are testing the variables in the GLM against the null hypothesis of whether they are significant or not. The hypotheses are

H0 : Variable is not significant in the model
H1 : Variable is significant in the model

According to the p-values (Table 2), all of the variables are significant at the 5% significance level. Hence, we are using the model with all of the variables. In the model selection phase, we employed a stepwise selection method with AIC, using stepAIC function from R, to find the best fit model for our data. The result indicates that the full model has the lowest AIC among all the possible models examined by R (Table 3). Thus, we proceed to the next phase of adding interaction terms with the full model.

a. Education * Smsa: The Smsa variable refers to the residence in Smsa, an area requiring relatively higher living expenses but offering better education facilities including schools and extracurricular activities. People living in Smsa can access to better education and gain advantage in business so we think the education variable may apply various effect on the dependent variable according to the value of Smsa
b. Experience * Smsa: Besides better education facilities, Smsa also offers a better business environment. People living in Smsa can access to professional opportunities with better salary

and welfare so we think the experience variable may apply various effect on the dependent variable according to the value of Smsa

c. Experience * Parttime: The parttime variable refers whether the individual works part-time. In the real world, HR considers different weights for full-time and part-time experiences of candidates so we think the experience variable may apply various effect on the dependent variable according to the value of parttime

Observing the summary of the new model that includes all three of the interaction terms (Table 4), we can see that interaction Experience*Parttime is not significant at the 5% level so we exclude it from our model. The p-value of the Smsa variable indicates that it is not significant at the 5% level, but since it is a variable included in two interaction terms that are significant, therefore, we cannot comment on it alone.

Employing the stepAIC function on the latest model (with interaction terms), we can observe that this model has the lowest AIC among all possible models examined by R (Table 5). As a result we call the model below as our chosen one.

As a result, our final model can be written as below.

$$wage_i \sim Gamma\ (\ E(wage_i)\ ,\ \ )$$

$$\begin{aligned} log(\ E(wage_i)\ ) \ =\ & \beta_0 + \beta_1 Education_i + \beta_2 Experience_i + \beta_3 EthnicityCauc_i + \beta_4 Smsa_i + \\ & \beta_5 RegionNorthEast_i + \beta_6 RegionSouth_i + \beta_7 RegionWest_i + \beta_8 Parttime_i + \\ & \gamma_1 (Education_i * Smsa_i\ ) + \gamma_2 (Experience_i * Smsa_i) \end{aligned}$$

Q2: Using the final model, provide a summary (e.g., using tables and figures) of the empirical findings as well as interpretation of the estimated model parameters

According to the p-values obtained, all of the variables are significant at the 5% level except for Smsa. However, we cannot drop Smsa because it is included in the two interaction terms which are significant at the 5% level. Also, we cannot interpret the coefficient estimates for education, experience, and Smsa because they are part of the interaction terms.

Table 6: Gamma GLM coefficient estimates

Estimates	β_i	exp(β_i)	P-value	Interpretations
Intercept	4.688	108.636	< 0.001	The average weekly wage is 108 USD, when all other variables are set to 0.
Education	0.082	1.085	< 0.001	—
Experience	0.016	1.016	< 0.001	—
EthnicityCauc	0.231	1.260	< 0.001	On average, a Caucasian earns 26% more than an African American on a weekly basis.
Smsa	-0.039	0.962	0.481	—
RegionNorthEast	0.038	1.039	< 0.010	On average, someone living in the northeast earns 3.9% more than someone who lives in the midwest (the base category variable).
RegionSouth	-0.034	0.967	< 0.010	On average, someone living in the south earns 3.3% less than someone who lives in the midwest.
RegionWest	0.043	1.044	< 0.001	On average, someone living in the west earns 4.4% more than someone who lives in the midwest.
Parrttime	-0.904	0.405	< 0.001	On average, someone who works part time earns 59.5% less than someone who works

				full time.
Education*Smsa	0.012	1.012	< 0.010	On average, people living in smsa earn 1.2% more than those who are not on a weekly basis. The effect of being in this category increases as the person's education increases.
Experience*Smsa	0.002	1.002	< 0.001	On average, people living in smsa earn 0.2% more than those who are not on a weekly basis. The effect of being in this category increases as experience increases.
Null deviance: 13092.5 on 28154 degrees of freedom Residual deviance: 8402.4 on 28144 degrees of freedom Number of Fisher Scoring iterations: 6				

In this case, the deviance residuals are used to test for the assumptions. In the Residual Vs Fitted plot (Figure 2), we can detect some sort of pattern. More residuals are located above the 0 horizontal line as compared to below. Also, the residuals located in the negative territory are slightly skewed to the right. Generally, the residuals are not perfectly randomly scattered around 0. We can say that homogeneity of variance and linear relation assumptions hold to a certain extent but not perfectly. In the Normal Q-Q plot we check for the normality of the residuals. At the beginning, all residuals lie on the dotted line and at some point between 2 to 4 along the x-axis there is a sudden exponential increase, causing the curve to move away from the dotted line. As a result, the assumption that the residuals are normally distributed is questionable. From the Scale Location graph, most of the points are evenly scattered around the 0 horizontal line, which suggests the assumption of constant variance holds in this case. Lastly, all the residuals are within the 0.5 cook's distance line in the Residuals vs Leverage plot. Even though there are few high leverage points but their actual influence on the fit is not unduly large.

Q3: Explain why this model is better (worse) than the previous model used in Assignment 1 and compare the results.

The model that we used in assignment 1 is a linear model which can be written as follows.

$$
\begin{aligned}
log(wage_i) = {} & \beta_0 + \beta_1 Education_i + \beta_2 Experience_i + \beta_3 EthnicityCauc_i + \beta_4 Smsa_i + \\
& \beta_5 RegionNorthEast_i + \beta_6 RegionSouth_i + \beta_7 RegionWest_i + \beta_8 Parttime_i + \\
& \gamma_1 (Education_i * Smsa_i) + \gamma_2 (Experience_i * Smsa_i) + \\
& \gamma_3 (Experience_i * Parttime_i) + u_i
\end{aligned}
$$

Note: u_i is the error term which is assumed to have a mean of zero, constant variance, and the error terms are independent of each other. The coefficient estimates can be found in the appendix (Table 7)

In the following, we will explain why our new model is better than the old model in terms of model properties, residual analysis, and AIC. Then, we will compare the coefficient estimates of both models.

Model Properties

In the previous assignment, we have log-transformed the response variable, wage, and there are two drawbacks to the model as compared to a Gamma GLM. Firstly, the values of the response variable have been changed, which implies that we are no longer analyzing the original values. Furthermore, if we only perform log-transformation on the response variable, the model might produce negative value estimation for the response variable, which does not make sense in our case as the response variable is wage and it cannot be negative. Therefore, it is more appropriate to use a Gamma GLM to model for our data because it is easier to interpret and it meets the criteria of having only positive continuous values.

Residual Analysis

Gamma GLM seems to be a better model than the linear model, in terms of normality, if we inspect their residual analysis plots. First of all, if we look at the Residual vs Fitted plot for both models we can observe that the residuals are decreasing across the x-axis. The slope of the residuals from the Gamma GLM (Figure 2) is flatter than that in the linear model (Figure 3). However, we cannot be certain that the linearity has improved because the flatter slope could be caused by the larger y-axis scale for the Gamma GLM Residual vs Fitted plot. Additionally, there are some residuals from the Gamma GLM (Figure 2) that have a value greater than 4, while there are no residuals that have a value greater than 4 from the linear model (Figure 3). Hence, we cannot be conclusive regarding whether the linearity has improved.

On the other hand, the QQ-plot indicates an improvement of residuals normality for Gamma GLM. More residuals are located on the dotted line in the QQ-plot of the Gamma GLM (Figure 2) than the linear model (Figure 3). There are no significant differences when comparing the Scale-Location and Residuals vs Leverage plots between the two models. Overall, Gamma GLM is a better model than the linear model, but it still can be improved.

Coefficient Estimates

Since estimated coefficients are on the same scale, both also employed log, so the results of the two models are comparable. The difference between the models is that in the linear model we log transformed the response variable, while in Gamma GLM the log transformation takes place on the expected value of the response variable.

We observe that the coefficient estimates are very similar for both models (Table 6 and Table 7) and this is suggesting that the explanatory variables in both models have approximately the same effect on the average weekly wage. For example, the intercept of the linear model was 4.62, implying an average weekly wage of 101 USD when all the other explanatory variables are set to 0 (Table 7). The intercept coefficient of the Gamma GLM is 4.69 (Table 6), suggesting an average weekly wage of 108 USD. The difference between the result is only 6 USD. Furthermore, the coefficient estimate of the northeast variable and experience*smsa are the same in both models. Moreover, in the linear model the average weekly wage of part time workers is 63% lower than full time workers. We were worried about that value since it indicates a very large difference in wage. However, the Gamma GLM also produces a similar result for this interaction term. The average weekly wage of a part time worker earns 59.5% lesser than a full time worker.

AIC

AIC is another indicator for model comparison. The lower the AIC, the better the model. The AIC of the Gamma GLM, 394743 (Table 5), is lower than the AIC of the linear model, 45960.26 (Table 8). Therefore, the Gamma GLM is a better model.

Q4: Provide recommendations and limitations of your analysis. How can the analysis be improved?

GLM limitations

A major limitation of the GLM is that they can only have linear components in the systematic component. Moreover, linearity assumption is extremely difficult to be 100% satisfied, as a result the models lose predictive power. As seen from our discussion above according to the Q-Q plots normality assumption was just an approximation.

Limitations of AIC

AIC can only provide a relative test of model quality. AIC does not give us the best model but rather it tells us which one is the least worst among a group of possible models. Therefore, there is certainly room for improvement in the chosen model.

How analysis could be improved

Firstly, we could include more variables that have an effect on wages such as the cost of living, the productivity, and types of occupation. Also, in our data set the variable 'experience' is calculated by *age - education - 6*, which causes some of our data under 'experience' to have negative values. A negative experience does not make any sense. Since negative experience essentially means the individual has no experience at all, maybe the 'experience' data will be more appropriate if we replace all the negative values with 0.

Since there is room of improvement for our models according to the residual analysis (Figure 2), hence, it is worth trying to use another type of positive continuous distribution such as the Weibull distribution, to fit the data to check if the model has been improved.

Since our response variable (wage) follows an exponential family distribution we could have fitted a general additive model (GAM) so we could model non-linear relationships using the smoothing functions. When using GAMs we are not assuming linearity for all of the explanatory variables, instead we allow some of the explanatory variables to be non-linear because not all data in the real world exhibits a linear relationship with our response variable. Alternatively, we could also try using a polynomial regression to capture non-linearity of the explanatory variable.

Appendix

Figures

Figure 1: Density plot for wage

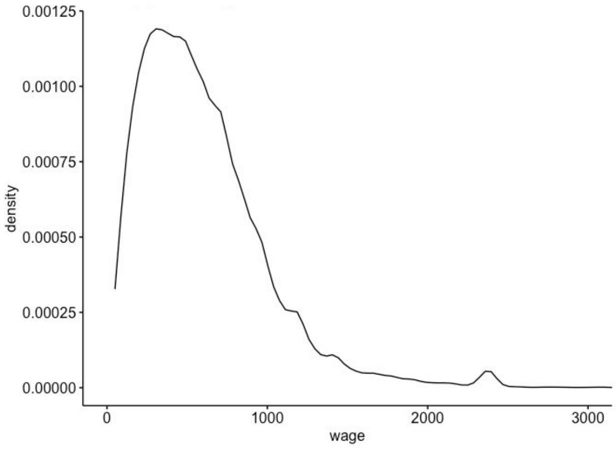


Figure 2: Residual diagnostics plots for Gamma GLM

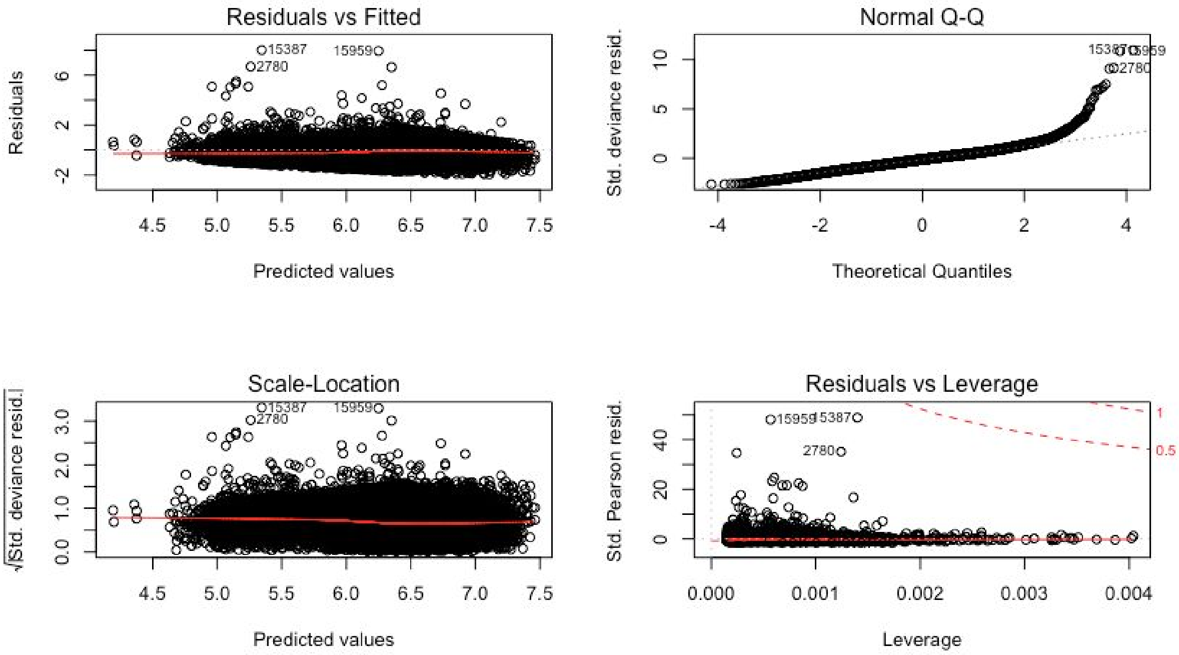
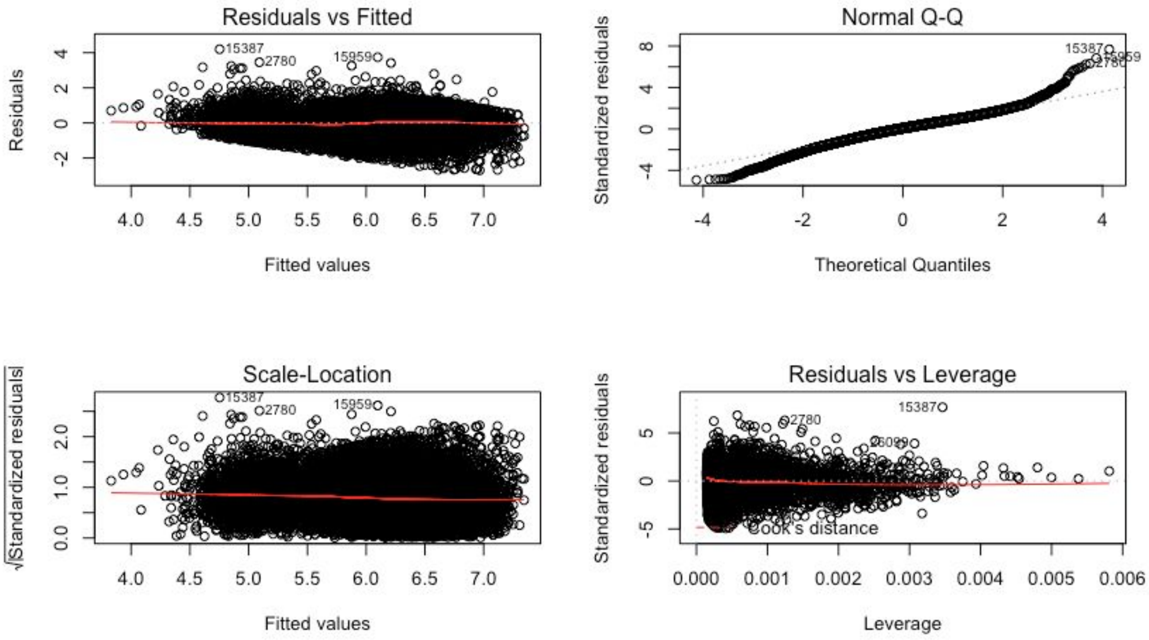


Figure 3: Residual analysis plots for the linear model after log-transforming the response variable

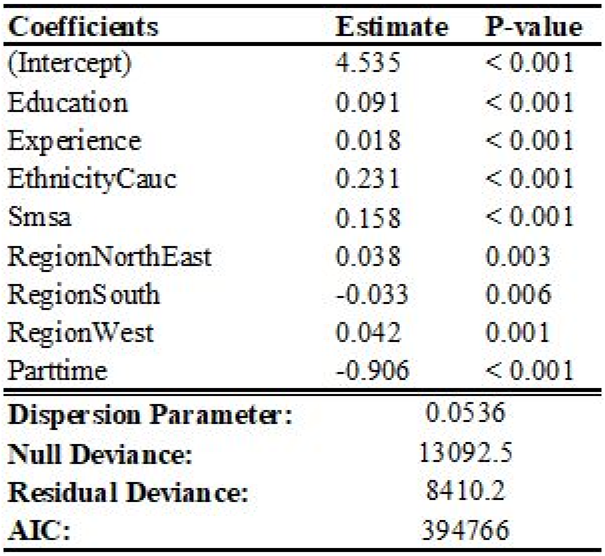


Tables

Table 1: Descriptive summary table of the variables in the dataset

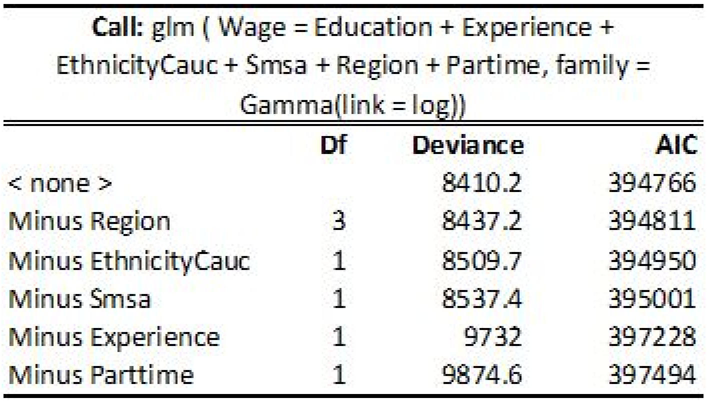
Variable	Description
Wage	Wage in US dollars per week.
Education	Years of education.
Experience	Years of working experience.
Ethnicity	A factor with levels Caucasion (Cauc) & African-American (afam).
Smsa	Residence in a standard metropolitan statistical area.
Region	The region within the United States of America. A factor with levels midwest (baseline category), northeast, south, and west.
Parttime	Whether the individual works part-time.

Table 2: Summary table of the fitted Gamma GLM before adding interaction terms



Coefficients	Estimate	P-value
(Intercept)	4.535	< 0.001
Education	0.091	< 0.001
Experience	0.018	< 0.001
EthnicityCauc	0.231	< 0.001
Smsa	0.158	< 0.001
RegionNorthEast	0.038	0.003
RegionSouth	-0.033	0.006
RegionWest	0.042	0.001
Parttime	-0.906	< 0.001
Dispersion Parameter:	0.0536	
Null Deviance:	13092.5	
Residual Deviance:	8410.2	
AIC:	394766	

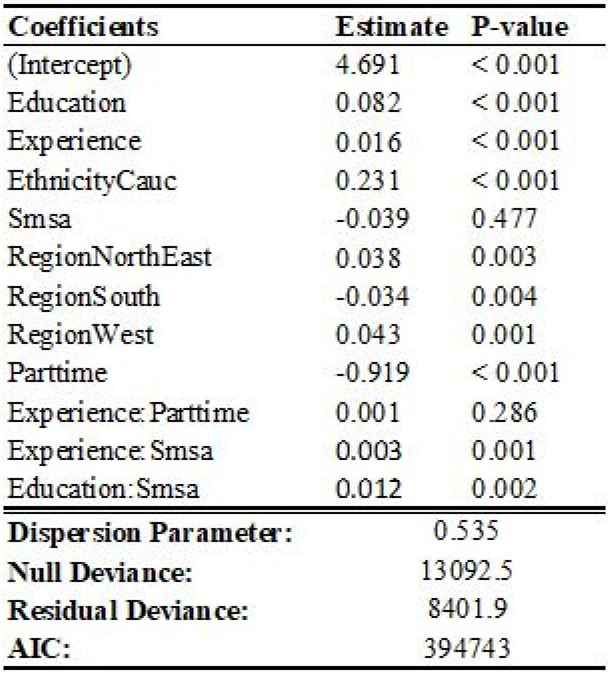
Table 3: stepAIC table of the fitted Gamma GLM before adding interaction terms



Call: glm (Wage = Education + Experience + EthnicityCauc + Smsa + Region + Partime, family = Gamma(link = log))

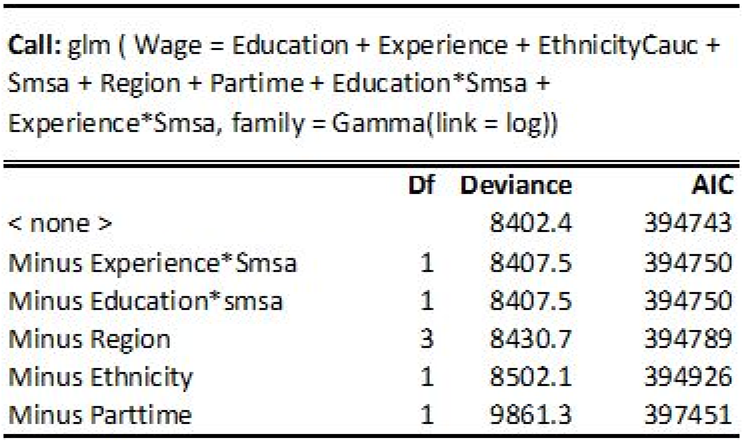
	Df	Deviance	AIC
< none >		8410.2	394766
Minus Region	3	8437.2	394811
Minus EthnicityCauc	1	8509.7	394950
Minus Smsa	1	8537.4	395001
Minus Experience	1	9732	397228
Minus Parttime	1	9874.6	397494

Table 4: Summary table of the fitted Gamma GLM after adding interaction terms



Coefficients	**Estimate**	**P-value**
(Intercept)	4.691	< 0.001
Education	0.082	< 0.001
Experience	0.016	< 0.001
EthnicityCauc	0.231	< 0.001
Smsa	-0.039	0.477
RegionNorthEast	0.038	0.003
RegionSouth	-0.034	0.004
RegionWest	0.043	0.001
Parttime	-0.919	< 0.001
Experience:Parttime	0.001	0.286
Experience:Smsa	0.003	0.001
Education:Smsa	0.012	0.002
Dispersion Parameter:	0.535	
Null Deviance:	13092.5	
Residual Deviance:	8401.9	
AIC:	394743	

Table 5: stepAIC table of the fitted Gamma GLM after adding interaction terms



Call: glm (Wage = Education + Experience + EthnicityCauc + Smsa + Region + Partime + Education*Smsa + Experience*Smsa, family = Gamma(link = log))

	Df	**Deviance**	**AIC**
< none >		8402.4	394743
Minus Experience*Smsa	1	8407.5	394750
Minus Education*smsa	1	8407.5	394750
Minus Region	3	8430.7	394789
Minus Ethnicity	1	8502.1	394926
Minus Parttime	1	9861.3	397451

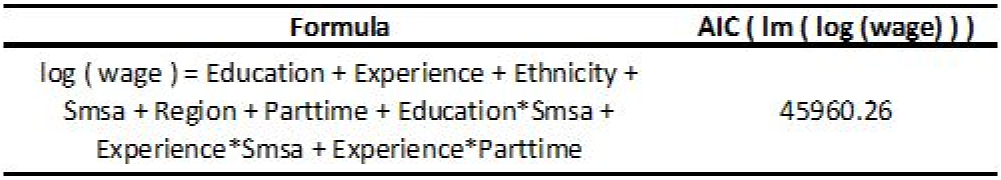
Table 7: Linear model coefficient estimates

Estimates	β_i	exp(β_i)	P-values	Interpretation
Intercept	4.617	101.145	< 0.001	The average wage is 101 USD per week when all the variables are set to 0.
Education	0.079	1.082	< 0.001	—
Experience	0.016	1.016	< 0.001	—
EthnicityCauc	0.216	1.242	< 0.001	The wage will be 24% higher for the caucasian workers than for the african american workers.
Smsa	-0.115	0.89105	0.006	—
RegionNorthEast	0.383	1.039	< 0.001	Having as categorical base region midwest we can say that, when i = northeast, wage increases by 3.9%.
RegionSouth	-0.051	0.950	< 0.001	when i = south, wage decrease by 5%.
RegionWest	0.021	1.021	0.032	when i = west, wage increase by 2.1%
Parttime	-0.991	0.371	< 0.001	The wage will be 63% lower for a worker working part-time compared to a worker who doesn't.

Education*Smsa	0.018	1.0182	< 0.001	Interaction indicates that residents in standard metropolitan statistical areas have 1.8% more wage compared to the ones which are not. The effect of being in this category increases when the years of education increase.
Experience*Smsa	0.002	1.002	9.38e-05 ***	Interaction indicates that residents in standard metropolitan statistical areas have 0.23% more wage compared to the ones which do not. The effect of being in this category increases when the years of experience increase.
Experience*Parttime	-0.005	0.995	8.18e-16 ***	Interaction indicates that individuals that work part time have 0.5% less wage compared to the ones which are not part time. The effect of being in this category decreases when the years of experience increase.

Residual standard error: 0.5474 on 28143 degrees of freedom , Multiple R-squared: 0.4156, Adjusted R-squared: 0.4153 , F-statistic: 1819 on 11 and 28143 DF, p-value: < 2.2e-16

Table 8: AIC of the Linear Model



Formula	AIC (lm (log (wage)))
log (wage) = Education + Experience + Ethnicity + Smsa + Region + Parttime + Education*Smsa + Experience*Smsa + Experience*Parttime	45960.26