The Impact of Age, Education, Marital Status and Sex on Wage and Salary Income

Kai Li[*] *Sicong Wang*[†] *Zeer Kang*[‡]

Abstract

This paper explores the relationship between multiple explanatory variables, marital status, age, sex, education and birthplace (born in the United States or not), and the explained variable income. Motivated by the classical Multiple Linear Regression Models (MLRs) and Multiple Nonlinear Regression Models (MNRs), we analyze the given reliable data and find out the best-fitted model of these variables using the Ordinary Least Squares (OLS) technique by software Stata. The regression results from Stata enable us to easily understand and interpret how these regressors affect the regressand through statistical inference.

Keywords: econometrics, data analysis, MLR, MNR, Stata, statistical inference

1 Introduction

Nowadays, people may wonder about factors that can lead to higher salaries. Economists have already done lots of empirical analysis on this topic. The objective of the paper is to perform similar empirical research among multiple interested predictor variables *Age*[1], *Education*, *Male*, *Nonmarried*, *US*, and a predicted variable *Wage* using a data set from CPS. We treat ordinal variables approximately continuous in our regression model. By performing such a regression analysis, we can analyze and interpret how selected factors can affect salary. More importantly, we, or economists, are able to explain social, economic issues such as gender, birthplace discrimination, income distribution, and the importance of education to future earnings. Thus, this empirical research provides meaningful insights on applications to economics and statistics.

[*]Department of Mathematics, The Ohio State University, USA, email: li.8130@osu.edu
[†]Fisher College of Business, The Ohio State University, USA, email: wang.8810@osu.edu
[‡]Department of Mathematics, Economics, The Ohio State University, USA, email: kang.1063@osu.edu
[1] We *italicize* the variables in our model in the paper.

Variables[2]	Observation	Mean	Standard Deviation	Min	Max
Wage	71800	45570.07	26987.01	6740	120000
Age	71800	42.31	13.92	15	85
Education	71800	89.52	22.17	20	125
Male	71800	0.51	0.50	0	1
Nonmarried	71800	0.30	0.45	0	1
US	71800	0.81	0.39	0	1

Tab. 1: Summary statistics for the research variables.

For the rest of the paper, we first filter out irrelevant variables and observations from the data in Data Section. Then, we shall provide a rough estimate of what the fittest model looks like in Empirical Methodology Section. Starting from assuming an MLR model, we expand our regression model to MNR and measure its goodness of fit. Then, we analyze the data and select a model that represents our data set most effectively and accurately using OLS estimators obatined from Stata, shown in Results Section. Also, we conclude the statistical significance of the coefficients and interpret the results. Before doing the analysis, we predict nonlinear relationships between *Wage* and *Age*, *Education*. Finally, notations, mathematical proofs of the OLS formula and GMT verifications are provided in Appendix B.

2 Data

One critical step is to inspect the reliability of the data set before using the data for analysis. The data set is a subsample of the Community Population Survey (CPS). The CPS, sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics, is the primary source of labor force statistics for the population of the United States. The CPS is one of the oldest, largest, and most well-recognized surveys in the United States [2]. We also examined the statistical sampling techniques and designs of CPS. Therefore, it is reasonable to believe the data set will provide reliable information.

We eliminate *Year* and *Month* in our data to reduce the complexity of working with panel data. Outliers are also narrowed out after drawing the Box-and-Whisker plots in Stata to preserve integrity. After that, we obtain summary statistics (see Table 1) and scatterplots of the ordinal variables in matrix form (see Figure 1). Our data set contains a variety of people whose annual ages are mostly in range 14000 to 78000, aging from 21 to 63 with education level from 5th-grade to doctoral, given the means and the standard deviations. Whether people are male, nonmarried, or born in the US are included for our economic interests as well.

[2] *Wage* and *Education* are ordinal variables. The higher the number, the higher the salary or degree.
[2] *Male*, *Nonmarried* and *US* are indicator variables. See the Code Book for specifics in Reference [2].

Fig. 1: *graph matrix*[3] of approximately continous regression variables

We came up with the following observations from the summary statistics and the figure:

1. *Wage* contains large integer dollar values. According to rules of thumb, it is more appropriate and natural to use logarithms[4] to model *Wage* as labor economists do [4, 6].

2. We observe an implicit quadratic relationship between *Wage* and *Age*, though the width between the zeros of the parabola is large, which hides the potential relationship. Likewise, using a linear or an exponential function to describe *Wage* and *Education* is suitable [6]. To improve the likelihood of getting the most accurate model, we reviewed previous economic literature and concluded that our conjecture on the model would output more satisfactory goodness of fit.

3. For binary variables *Male*, *Nonmarried* and *US*, the original form without any modification is good enough, but some interacted terms for the independent variables are needed for fitness. Based on past literature, we include $Age \times Education$, $Education \times Male$ and so forth for exploration. Thus, we predict a positive linear relationship between *Wage* and *Male*, *US*, and the opposite for *Nonmarried*.

After that, we run regressions and test which model is best given current data. However, there are some constraints regarding the data set. First, compared to time series data or panel data, cross-sectional data provide difficulty to justify if the cause and effect relationship follows exposure in time or exposure results from that effect. Moreover, the chronological ordering of the observations may provide essential information in time series data [6]. In this empirical analysis, cross-sectional data fulfill our goals.

[3] *graph matrix* is a Stata command used to draw a scatterplot matrix to illustrate the bivariate relationships.
[4] We use natural logarithm ln for simplicity throughout the paper.

3 Empirical Methodology

We begin by assuming an MLR because MLR is the most widely used model in econometrics, serving as the baseline to other regression models [1, 6]. The PRF for a cross-sectional sample is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{3.1}$$

In Equation (3.1), the subscript i indicates the i-th of the n observations in the sample[5]. Y_i is the regressand. $X_{1i}, X_{2i}, \ldots, X_{ki}$ are the k regressors. The variable ϵ_i is the error term or stochastic disturbance of the regression, given n observations.

Then, our SRF can be obtained using OLS estimation:

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}, \\ \hat{\epsilon}_i &= Y_i - \hat{Y}_i \end{aligned} \tag{3.2}$$

$\hat{Y}_i$ and $\hat{\epsilon}_i$ are called the OLS predicted values and residuals of the error term ϵ_i. The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and $\hat{\epsilon}_i$ can be computed from a sample of n observations of $(X_{1i}, X_{2i}, \ldots, X_{ki})$ using multivariable calculus and linear algebra. The formula and the proof are shown in Appendix B.1.

Yet, as we have seen in Data Section, MLR may not be the best-fitted model for our data. That is why we introduce MNR in this paper. Here, we focus on two nonlinear functions for a single independent variable: polynomials and logarithms. The polynomial regression model of degree r is written as:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + \epsilon_i \tag{3.3}$$

For logarithms, there are three linear forms after transformed from MNR:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + \epsilon_i \tag{3.4}$$

$$\ln(Y_i) = \beta_0 + \beta_1 (X_i) + \epsilon_i \tag{3.5}$$

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \epsilon_i \tag{3.6}$$

Equation (3.4), (3.5) and (3.6) are referred to as linear-log model, log-linear model and log-log model. As mentioned in Data Section, we model *Age* using quadratic polynomials and *Wage* using logarithms. Hence, we use Equation (3.3) of degree 2 and Equation (3.5) in the regression model. The OLS estimators in the above models are obtained using Stata. We will show if it is reasonable to assume the OLS assumptions in Appendix B.2. For details regarding the notations and mathematical representations, see Appendix A.

[5] We will be using a lot of equations with subscript i having the same meaning. Thus, we will omit $i = 1, 2, \ldots, n$ in the equations for the rest of the paper.

4 Results

We prove in Appendix B.2.4 that our error term is heteroskedastic. Thus, we use heteroskedasticity-robust t statistics to estimate the parameters after obtaining heteroskedasticity-robust standard errors for $\hat{\beta}_i$s. The estimated regression equations are given in Table 2, where robust standard errors are given in parentheses below coefficient estimates. We test three joint hypotheses using F-statistics in every regression. Our base specification, as well as the best-fitted model, is Regression (5) since the primary variables and interactions follow our economic interests and intuition. We include other regressions as our alternative specification models to see if other combinations of variables are more fitted to the data, based on plausible interactions among all variables. We can tell from R^2 that there is not a big difference among the regressions except for Regression (1), which we have already shown to be the least suitable model for our purposes. Therefore, we are confident that Regression (5) provides the most useful and interesting results. Though B.2.1 assumption may fail in our case, the base specification may not provide a biased interpretation to a large extent. Thus, we are still positive that the regression results provide meaningful application interpretations.

In Regression (5), all variables are statistically significant at 1% level. We reject the joint hypothesis that the coefficients of all variables and interactions are 0. The regression equation explains 26.6 percentage of the variation of *lnWage*. By analyzing the regression coefficients, we find that, on average, as one's *Age* increases before roughly 48 years old, *Wage* will increase at a slower speed, holding other regressors equal. Following that, *Wage* will decrease at a faster rate as *Age* goes up. Also, we observe that the difference in *Wage* between a bachelor's degree and a master's degree is 13.68% for *Nonmarried* individuals, all else equal. Similarly, there is a gender gap for *Nonmarried* individuals in *Wage* difference about 16.8% on average. Interestingly, we find that the interaction coefficient for *Male*×*Nonmarried* is negative, which implies that a *Nonmarried Male*, given a 3-year-college education, has 20.71% less *Wage* compared to another *Male* who is not *Nonmarried*, on average. In that case, there is an economic problem: the marital status gap for *Wage*. One more finding we have is that people who born in the *US*, on average, tend to earn more *Wage* than people who do not, given the same level of *Age*, *Education*, sex and marital status. Finally, we confirm that the statistics result mostly match with our predictions.

For our interest, we hypothesize two individuals for *Wage* prediction using the best-fitted model. One is a 21-year-old nonmarried female born outside the US earning a bachelor's degree and another is a 20-year-old nonmarried male born outside the US. The result is 26622.4 and 30336.3. We also randomly pick a Regression from (2) to (6) to predict, and we obtain a result of 24660.8 and 31839.4. Both results lie in our range of *Wage* estimation. We are optimistic about the difference due to rounding or uncontrollable errors.

Dependent variable: Wage 71800 observations	(1) Wage	(2) lnWage	(3) lnWage	(4) lnWage	(5) lnWage	(6) lnWage
Age	150.6** (7.56)	0.0657** (0.0011)	0.0648** (0.0011)	0.0556** (0.0044)	0.0650** (0.0011)	0.0436** (0.0047)
Age^2		-0.000679** (<0.0001)	-0.000671** (<0.0001)	-0.000476** (<0.0001)	-0.000673** (<0.0001)	-0.000367** (0.0001)
Education	460.1** (4.22)	0.0100** (0.0001)	0.00945** (0.0001)	0.0101** (0.0011)	0.0096** (0.0001)	0.00750** (0.0012)
Male	12108.0** (177.66)	0.291** (0.0042)	0.291** (0.0042)	0.293** (0.0042)	0.341** (0.0050)	0.628** (0.0180)
Nonmarried	-9550.9** (216.35)	-0.147** (0.0056)	-0.348** (0.0212)	-0.146** (0.0056)	-0.2151** (0.0223)	-0.154** (0.0242)
US	1764.7** (229.35)	0.0750** (0.0055)	0.0768** (0.0055)	0.0756** (0.0055)	0.0764** (0.0055)	-0.00900 (0.0181)
Age×Education				0.000101* (0.0001)		0.000233** (0.0001)
Age^2×Education				-0.00000215** (<0.0001)		-0.00000333** (<0.0001)
Education×Nonmarried			0.00226** (0.0002)		0.0018** (0.0002)	0.00117** (0.0003)
Education×Male						-0.00313** (0.0002)
Education×US						0.000995** (0.0002)
Male×Nonmarried					-0.173** (0.0094)	-0.186** (0.0094)
Constant	-6775.9** (533.90)	8.039** (0.0255)	8.109** (0.0263)	8.050** (0.0964)	8.071** (0.0263)	8.251** (0.1091)
F-Statistics						
(a) All variables and interactions=0	4299.5 (<0.0001)	4367.69 (<0.0001)	3719.21 (<0.0001)	3409.94 (<0.0001)	3307.16 (<0.0001)	2329.64 (<0.0001)
(b) Age×Education, Age^2×Education=0				75.05 (<0.0001)		52.63 (<0.0001)
(c) All interactions=0			95.26 (<0.0001)	75.05 (<0.0001)	214.06 (<0.0001)	144.83 (<0.0001)
R^2	0.2280	0.2614	0.2625	0.2635	0.2660	0.2706

*These regressions are estimated using Stata. Standard errors are given in parentheses under coefficients, and p-values are given in parentheses under F-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

Tab. 2: Regression results

5 Conclusion

In this paper, we present a regression analysis using classical econometrics methods. Namely, we use MLR and MNR models as our functional form based on economic theories, literature and intuition. We estimate the parameters using the OLS estimation method. Though not all OLS assumptions are satisfied, as shown in Appendix B.2, we believe the drawbacks of the regression models will not interfere with our interpretations significantly because OLS estimators are not BLUE in a lot of economics applications. As long as the equations are supported by sound theory, free of major econometric problems, used theoretically logical functional form and no obvious important variables omitted, we can evaluate the regression results as useful [5]. This proves our model accuracy when we interpret the results in the best-fitted model Regression (5).

In econometrics, subfields of economics are involved such as macro, labor, social economics and government finance [6]. In this paper, we focus on only limited subjects in economics, but more interesting research can be done in other fields too. Another direction of future investigation will be to consider an empirical analysis of time series data. Especially, we are interested in real GDP and inflation in an economy over a given time period. Since the methods in this paper cannot be used in time series regression, expanding our knowledge to analyze time series data appears to be a natural next step in the near future.

A List of Notations[6]

Notation	Meaning	Mathematical Expression[7]
RSS	Residual Sum of Squares	$\text{RSS} = \sum(Y_i - \hat{Y}_i)^2$
ESS	Explained Sum of Squares	$\text{ESS} = \sum(\hat{Y}_i - \bar{Y}_i)^2$
TSS	Total Sum of Squares	$\text{TSS} = \sum(Y_i - \bar{Y}_i)^2$
SER	Standard Error of Regression	$\text{SER} = \sqrt{\frac{\text{RSS}}{(n-k-1)}}$
$R, \mathsf{Corr}(X_i, Y_i)$	Pearson correlation coefficient	$R = \frac{\sum(X_i-\bar{X}_i)(Y_i-\bar{Y}_i)}{(n-1)s_X s_Y}$
R^2	Coefficient of determination	$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$
$\bar{R}^2, R^2_{adj}$	The adjusted R^2	$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\text{RSS}}{\text{TSS}}$
VIF	Variance Inflation Factor	$\text{VIF}_i = \frac{1}{1-R_i^2}$

Tab. 3: List of important notations

[6] From now on, summation will be used extensively. Hence, we will omit the bounds of the summation when it includes all the measurements in the sample, *i.e.*, we will refer $\sum$ as $\sum_{i=1}^{n}$.

B Mathematical Proof and Verification

B.1 The OLS Estimator of $\hat{\beta}$

In this appendix, we introduce the basic mathematics to obtain a formula for the OLS estimators $\hat{\beta}_i$ in MLR. We reach the goal through minimizing the RSS, given Y_i and $\hat{Y}_i$ in Equation (3.2). *i.e.* We take partial derivatives of RSS with respect to each $\hat{\beta}_i$, set those derivatives equal to zero, and solve for $\hat{\beta}_i$s:

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = 0 \quad \frac{\partial RSS}{\partial \hat{\beta}_1} = 0 \quad \cdots \quad \frac{\partial RSS}{\partial \hat{\beta}_k} = 0$$

To show the properties of this system of linear equations, let us check the equation for $\hat{\beta}_0$:

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = (-2) \sum \left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_n X_{ki}) \right] = 0$$

i.e.

$$n\hat{\beta}_0 + (\sum X_{1i})\hat{\beta}_1 + (\sum X_{2i})\hat{\beta}_2 + \cdots + (\sum X_{ki})\hat{\beta}_k = \sum Y_i$$

Similarly, we get a system of linear equations for other $\hat{\beta}_i$s:

$$\begin{aligned}
(\sum X_{1i})\hat{\beta}_0 + (\sum X_{1i}^2)\hat{\beta}_1 + (\sum X_{1i}X_{2i})\hat{\beta}_2 + \cdots + (\sum X_{1i}X_{ki})\hat{\beta}_k &= \sum X_{1i}Y_i \\
(\sum X_{2i})\hat{\beta}_0 + (\sum X_{1i}X_{2i})\hat{\beta}_1 + (\sum X_{2i}^2)\hat{\beta}_2 + \cdots + (\sum X_{2i}X_{ki})\hat{\beta}_k &= \sum X_{2i}Y_i \\
&\vdots \\
(\sum X_{ki})\hat{\beta}_0 + (\sum X_{1i}X_{ki})\hat{\beta}_1 + (\sum X_{2i}X_{ki})\hat{\beta}_2 + \cdots + (\sum X_{ki}^2)\hat{\beta}_k &= \sum X_{ki}Y_i
\end{aligned} \tag{B.1}$$

A simple way to solve Equations (B.1) is by expressing them in matrix form. Before moving forward, it is essential to know the matrix representation of our PRF in Equation (3.1):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{B.2}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ 1 & x_{13} & x_{23} & \dots & x_{k3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

[7] k is the number of slope coefficients, given n observations.

Thus, by similar transformation, the system of linear Equations (B.1) can be written as:

$$(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y} \tag{B.3}$$

Thus, the solutions of $\hat{\beta}_i$ in Equation (B.3) can be obtained if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible[8]:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{B.4}$$

B.2 The OLS Assumptions Verification

In this appendix, we show whether the extended least squares assumptions (*i.e.* CLM assumptions) are satisfied in our model. If B.2.1 to B.2.5 are satisfied, by GMT, the OLS estimator is BLUE and MVUE [4, 6]. However, in practice, the economic applications are often heteroscedastic, but that does not mean the estimators are less efficient than OLS [1, 4]. B.2.6 is not required for asymptotic analysis [6].

B.2.1 ϵ_i has conditional mean zero

Equivalently, the expectation of the stochastic error term ϵ is zero given all values of the independent variables. Mathematically speaking, we have

$$\mathbb{E}[\,\epsilon \mid X_{1i}, \ldots, X_{ki}\,] = 0 \tag{B.5}$$

To verify Equation (B.5), we check whether our model has functional form misspecification, OVB, measurement error in an explanatory variable [6]. These tests are sufficient, but not necessary conditions to prove the assumption [5]. We use Ramsey's RESET as a general test for functional form misspecification, and it has been proven to be useful and can be made robust to heteroskedasticity [1, 5, 6]. In our model, though RESET was rejected, we built our model based on economic literature and intuition, and thus we believe our model provides satisfactory interpretations in Results. For OVB, our multiple regression analysis has a lower probability of OVB compared to single regression analysis [6]. Also, adding one or more regressors not necessarily affects the adequacy of the model, observed from Table 2. Therefore, in our case, we included the hidden biases and measurement error in the stochastic error term ϵ [5].

B.2.2 $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$ are i.i.d drawn from their joint distribution

We check this assumption by considering if the data are collected by simple random sampling [4]. In our case, it is very plausible to have this assumption based on CPS's sampling designs and techniques. Moreover,

[8] In linear algebra, an n-by-n matrix $\boldsymbol{A}$ is called invertible if there exists an n-by-n matrix $\boldsymbol{B}$ such that $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I_n}$, where $\boldsymbol{I_n}$ is n-by-n identity matrix. We say that $\boldsymbol{B}$ is the inverse of $\boldsymbol{A}$, denoted by $\boldsymbol{A^{-1}}$.

Variable	VIF	1/VIF
Nonmarried	1.31	0.76
Age	1.30	0.77
Education	1.05	0.96
US	1.03	0.97
Male	1.02	0.98
Mean VIF	1.14	

Tab. 4: *estat vif*[11] for explanatory variables

an important feature of cross-sectional data is that we can assume the samples are randomly selected from the population [6].

B.2.3 No perfect multicollinearity

To show no perfect multicollinearity between predictor variables, we show that the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular[9] so that the solution of $\hat{\boldsymbol{\beta}}$ is unique in Equation (B.4) [1, 3, 4]. We verified the matrix is nonsingular using Stata in our regression results[10].

For imperfect multicollinearity, we check by calculating the VIFs of the variables. The results are shown in Table 4. As a rule of thumb, we are able to conclude that there is high collinearity between the explanatory variable and explained variables when the largest VIF is bigger than 10 [1, 3, 6]. Yet, the VIFs test is a sufficient but not a necessary condition to conclude multicollinearity [5].

B.2.4 Homoskedasticity

Homoskedasticity means that the error term ϵ has a finite constant variance. *i.e.*

$$\mathrm{Var}[\,\epsilon \mid X_{1i}, \ldots, X_{ki}\,] = \sigma^2 < \infty \tag{B.6}$$

We check homoskedasticity by graphing the fitted values and residuals, shown in Figure 2. It is not reasonable to assume homoskedasticity since more and more residuals of data points are negative towards the right tail, which indicates a sign of heteroscedasticity [5]. Or, we use Cameron & Trivedi's decomposition of IM-test to see whether the null hypothesis, the variance of the residuals is constant, can be rejected. The result is shown in Table 5. Thus, we have sufficient evidence to say that our OLS estimator is heteroskedastic. Heteroskedasticity typically causes OLS to no longer be an MVUE [5].

[9] A matrix is nonsingular is equivalent to that a matrix is invertible or a matrix has full rank.
[10] Invertible matrix theorems and properties help us to determine multicollinearity by hand.
[11] *estat vif* is a Stata command used to find VIFs for the independent variables.

Source	chi2	df	p
Heteroskedasticity	1504.45	31	0.00
Skewness	3952.01	8	0.00
Kurtosis	222.82	1	0.00
Total	5679.28	40	0.00

Tab. 5: *estat imtest*[12] for heteroskedasticity, skewness and kurtosis

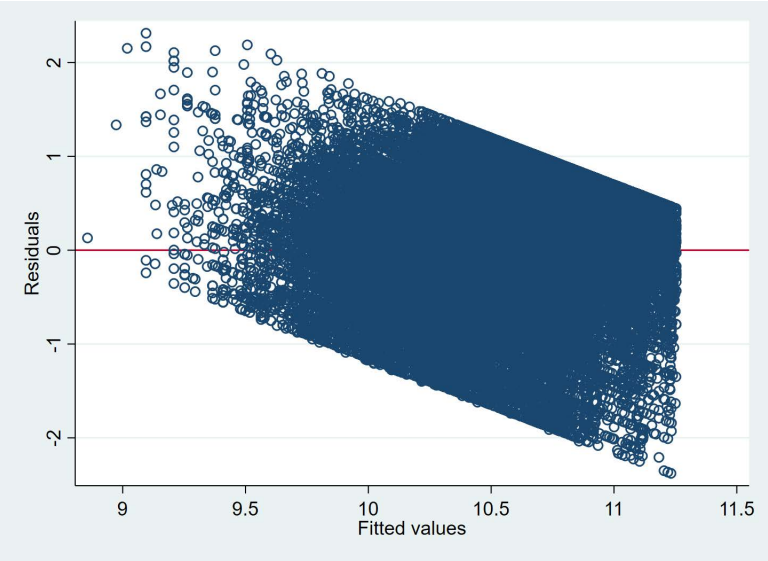

Fig. 2: *rvfplot*[13] of fitted values and residuals

B.2.5 Large outliers are unlikely

Mathematically speaking, we say that it is unlikely to have large outliers if

$$0 < \mathbb{E}[\, X_{1i}^4 \,] < \infty \quad , \cdots , \quad 0 < \mathbb{E}[\, X_{ki}^4 \,] < \infty, \quad \textit{and} \quad 0 < \mathbb{E}[\, Y_i^4 \,] < \infty$$

We already cleaned our data set by using Box-and-Whisker plots as mentioned in Data section. Another way to state this assumption is that X and Y have finite kurtosis [4]. We found that all variables have a finite kurtosis using Stata, though we reject the hypothesis that the kurtosis is normal (See Table 5).

[12] *estat imtest* is a Stata command used to perform an information matrix test for the regression model and an orthogonal decomposition into tests for heteroskedasticity, skewness, and kurtosis due to Cameron and Trivedi (1990).

[13] *rvfplot* is a Stata command used to graph a residual-versus-fitted plot, a graph of the residuals against the fitted values.

B.2.6 Normal errors

Normal errors refer to the normality of our error term ϵ:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

In multivariate regression, CLT applies to the OLS estimators as well [1, 4]. *i.e.* $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2_{\hat{\beta}_i})$. In our case, since the model is heteroscedastic, it is hard to conclude that the errors are normal [6]. However, we can observe the error follows asymptotic normality in some extent in Figure (2). When sample size is large, the distributions of the regression coefficients approach normality under general conditions [3].

C Acronyms

BLUE	Best Linear Unbiased Estimator	**MNR**	Multiple Nonlinear Regression Model
CLM	Classical Linear Model	**MVUE**	Minimum-Variance Unbiased Estimator
CLT	Central Limit Theorem	**OLS**	Ordinary Least Squares
CPS	Community Population Survey	**OVB**	Omitted Variable Bias
GMT	Gauss-Markov Theorem	**PRF**	Population Regression Function
i.i.d	independent and identically distributed	**RESET**	Regression Specification Error Test
MLR	Multiple Linear Regression Model	**SRF**	Sample Regression Function

References

[1] Christopher F. Baum. *An introduction to modern econometrics using stata*. Stata Press, 2006.

[2] Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. Integrated Public Use Microdata Series, Current Population Survey: Version 7.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D030.V7.0.

[3] Michael Kutner, Christopher Nachtsheim, John Neter, and William Li. *Applied linear statistical models*. McGraw-Hill/Irwin, fifth edition, 2005.

[4] James H. Stock and Mark W. Watson. *Introduction to econometrics*. Pearson, fourth edition, 2019.

[5] A. H. Studenmund. *Using econometrics: a practical guide*. Pearson, seventh edition, 2017.

[6] Jeffrey M. Wooldridge. *Introductory econometrics: a modern approach*. Cengage, seventh edition, 2020.