Liaw, Kao-Lee; Lin, Ji-Ping

Working Paper

A shortcoming of the conventional approach to statistical explanation for wage or income variation: Offering a better alternative

QSEP Research Report, No. 453

Provided in Cooperation with:
Research Institute for Quantitative Studies in Economics and Population (QSEP), McMaster University

Suggested Citation: Liaw, Kao-Lee; Lin, Ji-Ping (2014) : A shortcoming of the conventional approach to statistical explanation for wage or income variation: Offering a better alternative, QSEP Research Report, No. 453, McMaster University, Research Institute for Quantitative Studies in Economics and Population (QSEP), Hamilton (Ontario)

This Version is available at:
https://hdl.handle.net/10419/119763

QSEP

RESEARCH INSTITUTE FOR QUANTITATIVE STUDIES IN ECONOMICS AND POPULATION

A SHORTCOMING OF THE CONVENTIONAL APPROACH TO STATISTICAL EXPLANATION FOR WAGE OR INCOME VARIATION: OFFERING A BETTER ALTERNATIVE

KAO-LEE LIAW
JI-PING LIN

QSEP Research Report No. 453

McMaster University
SOCIAL SCIENCES

A SHORTCOMING OF THE CONVENTIONAL APPROACH TO STATISTICAL EXPLANATION FOR WAGE OR INCOME VARIATION: OFFERING A BETTER ALTERNATIVE

KAO-LEE LIAW
JI-PING LIN

QSEP Research Report No. 453

August 2014

Kao-Lee Liaw is a QSEP Research Associate and professor emeritus in the McMaster University School of Geography and Earth Sciences. Ji-Ping Lin is an Associate Research Fellow in the Research Center for Humanities and Social Sciences, Academia Sinica Nankang, Taipei, Taiwan

A Shortcoming of the Conventional Approach to Statistical Explanation for Wage or Income Variation: Offering a Better Alternative

Kao-Lee Liaw, Professor Emeritus

School of Geography and Earth Sciences

McMaster University

Hamilton, Ontario, L8S 4K1

CANADA

E-mail: rliaw@mcmaster.ca or liawkl2001@yahoo.ca

Ji-Ping Lin, Associate Research Fellow

Research Center for Humanities and Social Sciences

Academia Sinica

Nankang, Taipei 115

TAIWAN

E-mails: jplin@gate.sinica.edu.tw

August 2014

A Shortcoming of the Conventional Approach to Statistical Explanation for Wage or Income Variation: Offering a Better Alternative

Kao-Lee Liaw, McMaster University

Ji-Ping Lin, Academia Sinica

Abstract

In explaining wage or income by personal attributes (e.g. educational attainment, age, and ethnicity) in a regression model, many researchers choose to use the log of wage or income as the dependent variable and then to estimate the unknown coefficients by some version of the least-squares method. We call this approach the *conventional approach*.

Using the micro data of the 2005-2007 American Community Survey and Taiwan's 2001-2010 Manpower Utilization Survey, we show that the conventional approach has the serious shortcoming of under-predicting the observed wage structure in the space spanned by the values of the explanatory variables. In addition to revealing the reason for the under-prediction problem and linking the severity of this problem to wage variability, we present a *nonlinear approach* that does not have this shortcoming. We also offer a SAS module for carrying out the estimation task in the nonlinear approach.

Keywords: uneven distortions; log-transformation of dependent variable; regression model; wage structure; income; nonlinear estimation

JEL Code: C18, C51, C87, J31

1. Introduction

In explaining the variation in wage (or other kinds of income) among individuals by a set of explanatory factors such as educational attainment, age (or "experience"), and working time, it is quite common to use the (natural) *log of wage* as the dependent variable and then use a linear-in-coefficient specification of regression model (e.g. Borjas, 1985; Aeberhardt, et al, 2010; El-Araby Aly and Ragan, 2010; Lin, 2013). When cross-sectional data with information on a rich set of relevant explanatory factors are available, the unknown coefficients are conveniently estimated by the weighted or un-weighted ordinary least-squares method, depending on whether there is a weight variable or not.[1] When panel data are available and the distorting effects of

[1] It has been well demonstrated by Aeberhardt, et al (2010) that when well-chosen explanatory factors are used to explain log of wage in a linear-in-coefficient regression model, three different estimation methods (OLS, 2-step Heckman-type procedure, and maximum likelihood method) yield nearly identical estimated coefficients. In other words, when potential confounders of a

unobservable explanatory factors are of concerned, the unknown coefficients are estimated by more elaborate methods (e.g. the fixed effects method). For simplicity, we call this way of statistical analysis the *conventional approach*.

Recently, we discovered that this conventional approach results in serious under-prediction of the underlying *wage structure*, which is defined as the pattern of average wages in the space spanned by the values of the explanatory variables. The extent of under-prediction can be as serious as 20 or 30%.

The main purpose of this paper is to demonstrate this under-prediction problem of the conventional approach and to offer an alternative approach to overcome it, using the micro data of the 2005-2007 American Community Survey (ACS). We will also reveal the mathematical principle beneath this inherent problem of the conventional approach and to show that this problem tends to be particularly serious when the wage variability is great.

To the extent of our knowledge, a careful examination of this problem has not yet been reported in the literature. The main reason seems to be that most researchers tend to focus on testing hypotheses or making assessments, which does not require the examination of the predicted wage structure, and that the indices of goodness-of-fit and trustworthiness (such as the adjusted R-square, t-statistics, and p-values) are incapable of revealing the seriousness of the under-prediction problem.

The organization of the rest of the paper is as follows. In section 2, we present the mathematical formulation of the conventional approach and a better approach (called nonlinear approach). In sections 3 and 4, we reveal and explain the shortcoming of the conventional approach and the superiority of the nonlinear approach. In section 5, we discuss the usefulness of a concise specification of the regression model via the nonlinear approach for dealing effectively with substantively important research questions. In section 6, we summarize the main points.

Two appendices are added to this paper. In Appendix A, we use the micro data of Taiwan's 2001-2010 Manpower Utilization Survey (MUS) to provide additional empirical evidence for substantiating the main points made in the paper. In Appendix B, we present a SAS module for estimating the coefficients of exponential regression models by weighted nonlinear least-squares method, as well as a SAS program that uses this module. The reason for writing this module is that the SAS procedure of nonlinear least-squares estimation (PROC NLIN) does not have the flexibility of allowing the use of a weight variable.

---

causal factor in question are included in the set of explanatory factors, the *conditional independence assumption* (Angrist and Pischke, 2009, p. 53) is valid and *selection bias* disappears, so that OLS is a proper estimation method.

2. Mathematical Formulation

Let Y_i be the wage of the ith individual. Also let $\mathbf{X}_i$ be a vector of explanatory variables, and $\boldsymbol{\beta}$ be a vector of unknown coefficients. The regression model of the conventional approach can be written as:

$$\mathrm{Ln}(Y_i) = f(\boldsymbol{\beta}, \mathbf{X}_i) + \varepsilon_i \qquad (1)$$

where $\mathrm{Ln}(Y_i)$ is the natural log of wage, $f(\boldsymbol{\beta}, \mathbf{X}_i)$ is a *linear-in-coefficient* function of $\boldsymbol{\beta}$ and $\mathbf{X}_i$, and ε_i is an unobservable random error term that is usually assumed to have a normal distribution.[2] It is assumed that the first element of $\mathbf{X}_i$ is 1 so that the first element of $\boldsymbol{\beta}$ is the unknown intercept. Note that when the empirical data are rich enough to allow causal inference, $\mathbf{X}_i$ may contain not only the causal variable in question but all relevant confounders and covariates.

Let w_i be the weight assigned to the ith individual. The unknown coefficients are to be estimated by the weighted least-squares method, which minimizes the sum of

$w_i \{\mathrm{Ln}(Y_i) - f(\mathbf{b}, \mathbf{X}_i)\}^2$ across all individuals in the sample, where $\mathbf{b}$ is the guessed value of $\boldsymbol{\beta}$.[3] Since it will become clear later in the paper that the cause of the under-prediction problem of the conventional approach lies in the log-transformation of the dependent variable rather than the estimation method, we choose the simplest sensible method in our exposition.

In the alternative approach that we propose, the regression model is written as:

$$Y_i = \mathrm{Exp}[f(\boldsymbol{\beta}, \mathbf{X}_i)] + \delta_i \qquad (2)$$

where Exp[ ] is the exponential function, and δ_i is another unobservable random error term. It is important to note that the units of δ_i and ε_i are different. For example, if δ_i =200 \$/week, the corresponding value of ε_i is supposed to be Ln(200)=5.289, with a hard-to-communicate unit of "Ln(\$/week)".

In the alternative approach, the unknown coefficients are estimated by a nonlinear weighted least-squares method that minimizes the sum of

---

[2] With panel data, the subscript i becomes a vector that runs across both individuals and time points. More explicitly, we can replace the single subscript "i" by the double subscripts "i,t" and rewrite the linear-in-coefficient function $f(\boldsymbol{\beta}, \mathbf{X}_i)$ as $f(\alpha_i, \lambda_t, \boldsymbol{\beta}, \mathbf{X}_{it}) = \alpha_i + \lambda_t + \boldsymbol{\beta}'\mathbf{X}_{it}$, where α_i and λ_t are unknown coefficients, and one of the variables in $\mathbf{X}_{it}$ represents a causal factor. The fixed effects method that starts by differencing from the person-specific means or between successive observations will cause α_i to disappear so that there is no longer enough information to compute the predicted value of the dependent variable. In other words, this kind of the fixed effects method can not reveal the under-prediction problem of interest in this paper.

[3] We consider the un-weighted method as a special case of the weighted method in which all weights are identical.

$w_i\{Y_i - Exp[f(\mathbf{B}, \mathbf{X}_i)]\}^2$ across all individuals in the sample, where **B** is another guessed value of **β**. For simplicity, we call this approach the *nonlinear approach*.[4]

Let $\underline{\mathbf{b}}$ and $\underline{\mathbf{B}}$ be the best value of **b** and **B**, respectively. The predicted value of Y_i generated by the conventional approach is

$$\underline{y}_i = Exp[\, f(\underline{\mathbf{b}}, \mathbf{X}_i)], \qquad (3)$$

whereas the predicted value of Y_i generated by the nonlinear approach is

$$\underline{Y}_i = Exp[\, f(\underline{\mathbf{B}}, \mathbf{X}_i)]. \qquad (4)$$

In the next section, we will use the merged 2005-2007 ACS micro data to demonstrate that $\underline{y}_i$ has the tendency of under-predicting the underlying wage structure of the real-world, whereas $\underline{Y}_i$ does not.

3. Demonstrating the Different Predictions between the Conventional and Nonlinear Approaches via the Full Specification

The micro data base of the merged 2005-2007 ACS is ideal for our investigation for several reasons. First, it has an extremely large sample size. Second, the ACS is based on a scientific sampling design that covers the whole country. Third, the responses from the households that received the questionnaire are legally mandatory. Forth, with respect to the households that failed to mail back the completed questionnaire, the staff of the Census Bureau are obliged to contact them or their neighbors so that the response rate is maintained at a high level and the under-coverage bias of low income households is reduced. With these four nice properties, it is highly likely that the observed wage structure can reflect very well the real wage structure of the whole country.

The ACS is a monthly survey on a sample that is representative of all households and individuals of the United States. Its questionnaire is essentially the same as the long-form

[4] Our estimation module computes two sets of standard errors for the estimated coefficients: (1) the conventional standard errors that are based on the assumption that δ_i has a constant variance (i.e. the homoskedasticity assumption); (2) the "robust" standard errors that does not depend on this assumption (i.e. they depend on the heteroskedasticity assumption). The formula for computing the robust standard errors is a weighted nonlinear generalization of equation (3.1.7) in Angrist and Pischke (2009), p. 45. Although the two sets may differ markedly in some cases, the difference between them is in general unimportant for substantive purposes, because for substantively important explanatory variables both sets of standard errors will be quite small, leading to t-statistics with very large magnitudes. The t-statistics shown in this paper are based on the conventional standard errors, because our estimation module did not compute the robust standard errors when the tables in this paper were created.

questionnaire of the 2000 population census. The cumulative sample size over a year is about 1% of the total population of the United States. The micro data set used in this study was created by the U.S. Census Bureau by merging the records of the surveys conducted in 2005, 2006, and 2007. The number of individual records in the data set is more than 8.8 million, which represent about 3% of the country's total population. There is a weight variable that reflects (1) the different sampling intensities of different strata in the sampling design and (2) the actual sizes of the underlying subpopulations. Naturally, its mean is about 33. In other words, on average, each record in the data base represents 33 persons in the real-world.

An important feature of the merged ACS data base is that all values of income variables pertaining to different months and years have been adjusted for inflation so that they reflect the dollar value as of 2007. However, it is useful to keep in mind that it has a shortcoming: all income variables are top-coded (at $666,000 for wage and salary income) so that averages and standard deviations of these variables tend to be understated. For an overview of ACS, see Mather et al. (2005) and U.S. Census Bureau (2009).

The sample selected for our study includes all male and female wage earners who are in the 25-64 age interval and belong to the non-Hispanic White ethnic group. A wage earner is defined as a person whose duration of work in the previous 12 months is at least 10 weeks and whose wage (formally "wage and salary incomes") is positive. We measure wage by "weekly wage", which is obtained by dividing the annual wage by the number of weeks worked. Note that the wage earners include cashiers and dish washers at the low end and hedge fund managers and company CEOs at the high end of the wage scale. The resulting sample size is huge: 1,250,825 male records and 1,167,589 female records.

To facilitate the effective communication of our main points, we focus on only two explanatory factors: educational attainment and age. Using Bachelor's degree as the reference category, educational attainment is represented by four dummy variables: ED_PR (primary), ED_2ND (completion of high school), ED_SC (some college education), and ED_MS (post-graduate degrees, including Master's, doctoral, and professional degrees). The age factor is represented by two explanatory variables: AGE_R45 (current age minus 45), and AGESQ_R45 (the square of AGE_R45). The reason for subtracting 45 from the current age is that we want the estimated intercept to reflect the wage level around mid-career. Underlying the choice of the quadratic function to represent the effect of the age factor is the seemingly sensible assumption that wage tends to rise with age and then decline after reaching a peak. But, this choice also imposes certain rigidity: the predicted age-pattern of wage will always be symmetric and bell-shaped. However, since the peak may be located near or even beyond age 64, the age pattern of the predicted wage structure may not look like a bell at all.

The observed male and female wage structures are shown in Figures 1 and 2, respectively. The wage represented by each point in the graphs is the weighted mean wage of a cell in the space created by crossing two sexes, five educational levels, and 40 single years of age from 25

to 64. Note that all the means mentioned in the rest of the paper are weighted means. For simplicity, we will drop the modifier "weighted".

Figure 1. The Age-by-Education Structure of Weekly Wage of US-born Non-Hispanic White **Male** Wage Earners: based on the micro data of the 2005-2007 ACS.

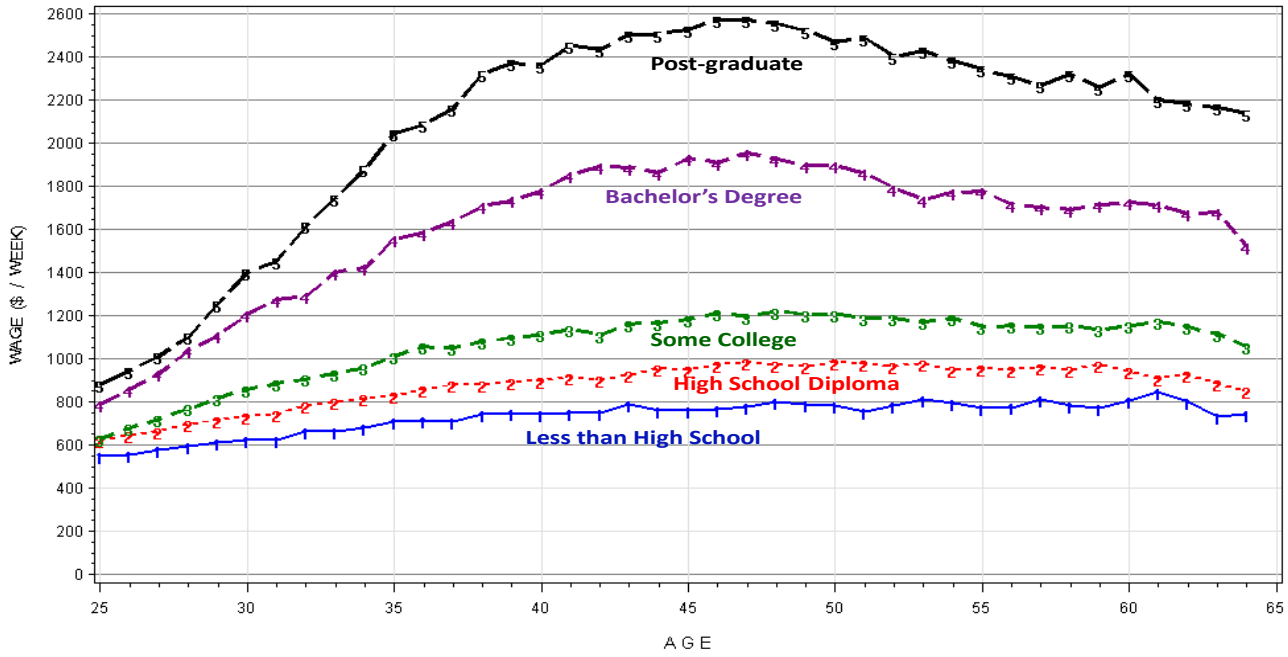
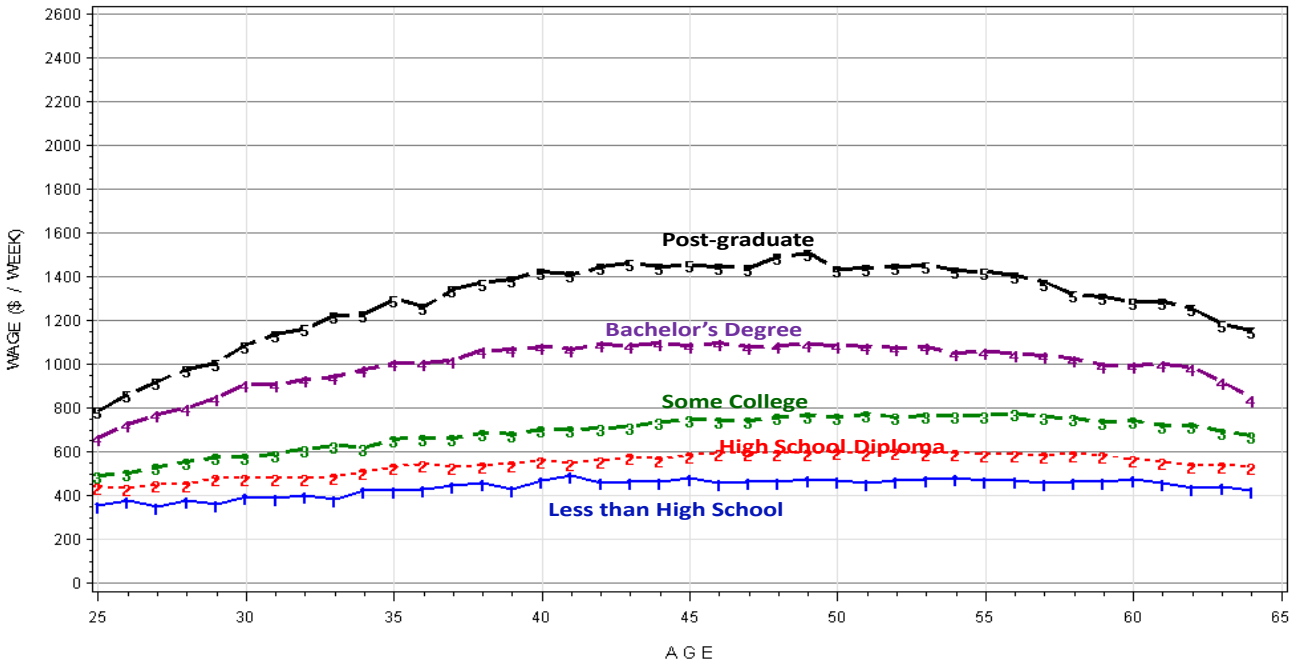


Figure 2. The Age-by-Education Structure of Weekly Wage of US-born Non-Hispanic White **Female** Wage Earners: based on the micro data of the 2005-2007 ACS.



For both males and females, the curves for the five levels of education, as a consequence of the huge sample size, are rather smooth and very well separated in a sensible way--the higher the level of education, the higher the mean wage. At each level of education, the mean wage tends to increase with age to a peak or a plateau in the mid- or late 40s and then decline towards the mid-60s. This pattern is particularly evident at the two highest levels of education. At the post-graduate level, the mean wage of males increases from about $850 at age 25 to a peak of nearly $2,600 at age 46 and then declines to about $2,150 at age 64, whereas the mean wage of females increases from about $800 at age 25 to a broad plateau of about $1,450 between 40 and 55 and then declines to about $950 at age 64. For both sexes, the age pattern of wage becomes flatter as the level of education decreases, implying that the proportion of dead-end jobs is higher at a lower level of education. As our attention moves down the education hierarchy, the peak becomes harder to identify and the plateau becomes wider and wider.

Our task here is to show how well the two approaches can predict the observed sex-specific wage structures. We apply the two approaches to the male and female data separately.

In applying the two approaches to the ACS micro data, we let the linear-in-coefficient function $f(\boldsymbol{\beta}, \mathbf{X}_i)$ in equations (1) and (2) be completely flexible by including the maximum number of interaction terms between the set of education variables and the set of age variables. This specification will allow the age-patterns of the predicted wage structure to differ freely among the five levels of education. For simplicity, we call this specification the *full specification*.

The estimated coefficients and the related statistical indicators obtained by applying the full specification of regression models via the two approaches are shown in Table 1. Although the two approaches use identical specification of the function $f(\boldsymbol{\beta}, \mathbf{X}_i)$, they yield different values for the estimated coefficients. For example, the estimated coefficient of ED_MS (the dummy variable for revealing the effect of changing educational attainment from Bachelor's degree to post-graduate degree) in the male panel is 0.25301 according to the conventional approach and 0.27685 according to the nonlinear approach. These values imply that for a male at age 45, changing educational attainment from Bachelor's level to post-graduate level is expected to increase wage by [Exp(0.25301)-1]*100%=29% according to the conventional approach and [Exp(0.27685)-1]*100%=32% according to the nonlinear approach. The corresponding values of the estimated coefficient in the female panel are 0.32003 and 0.28561, implying a wage increase of 38% and 33%, respectively. Both approaches suggest that pursuing post-graduate education is highly rewarding.[5] By simply looking at these values, it is impossible to tell which of the two approaches is more appropriate.

[5] Because we have not included confounding factors like parental education as part of the explanatory factors, the coefficients obtained from both approaches undoubtedly overstate the causal effect.

Table 1. The estimation results of the *full specification* of the regression model for explaining wage via the application of the conventional and nonlinear approaches to the ACS data.

	Conventional Approach			Nonlinear Approach		
Explanatory Variable	Coefficient	T-statistic	% Effect	Coefficient	T-statistic	% Effect
	Male Panel (Sample Size = 1,250,825)					
Intercept	7.29708	3768.0		7.54916	4154.7	
A. Age Factor						
AGE_R45	0.00611	45.7		0.01021	69.1	
AGESQ_R45	-0.00155	-125.7		-0.00139	-96.3	
B. Education Factor (Ref: Bachelor's deg.)						
ED_PR	-0.89254	-222.0	-59	-0.89604	-113.5	-59
ED_2ND	-0.62392	-245.8	-46	-0.68627	-199.2	-50
ED_SC	-0.41205	-161.0	-34	-0.47301	-155.2	-38
ED_MS	0.25301	77.3	29	0.27685	107.1	32
C. Interaction Terms						
AGE_R45 * ED_PR	-0.00035	-1.3		-0.00280	-4.9	
AGE_R45 * ED_2ND	-0.00053	-3.0		-0.00240	-8.7	
AGE_R45 * ED_SC	0.00013	0.7		-0.00056	-2.4	
AGE_R45 * ED_MS	0.00359	16.0		0.00214	9.7	
AGESQ_R45 * ED_PR	0.00105	42.8		0.00094	18.1	
AGESQ_R45 * ED_2ND	0.00078	48.0		0.00074	28.7	
AGESQ_R45 * ED_SC	0.00046	28.0		0.00051	22.2	
AGESQ_R45 * ED_MS	-0.00015	-6.7		-0.00003	-1.2	
Adj. R-square	0.1881			0.1606		
	Female Panel (Sample Size = 1,167,589)					
Intercept	6.71578	3220.1		7.00964	3714.7	
A. Age Factor						
AGE_R45	0.00118	7.6		0.00391	25.6	
AGESQ_R45	-0.00068	-51.0		-0.00085	-61.8	
B. Education Factor (Ref: Bachelor's deg.)						
ED_PR	-0.84219	-161.4	-57	-0.85716	-84.6	-58
ED_2ND	-0.56677	-202.6	-43	-0.64064	-175.8	-47
ED_SC	-0.34296	-128.1	-29	-0.39944	-136.9	-33
ED_MS	0.32003	91.6	38	0.28561	106.2	33
C. Interaction Terms						
AGE_R45 * ED_PR	0.00339	9.8		0.00116	1.6	
AGE_R45 * ED_2ND	0.00363	18.0		0.00197	7.0	
AGE_R45 * ED_SC	0.00538	27.6		0.00427	18.7	
AGE_R45 * ED_MS	0.00224	9.4		0.00202	9.6	
AGESQ_R45 * ED_PR	0.00022	7.1		0.00037	5.6	
AGESQ_R45 * ED_2ND	0.00014	8.0		0.00035	13.5	
AGESQ_R45 * ED_SC	0.00011	6.4		0.00027	13.0	
AGESQ_R45 * ED_MS	-0.00026	-11.6		-0.00016	-7.8	
Adj. R-square	0.1586			0.1343		

The qualitative aspects of the result generated by the conventional approach seem to be quite reasonable. First, the coefficients of the education variables reveal that education has a very strong monotonic effect on wage. And the huge magnitudes of the associated t-statistics indicate that the very strong effect of educational attainment is extremely trust-worthy. Second, for those

with a Bachelor's degree (the reference education category), the positive coefficient of AGE_R45 and the negative coefficient of AGESQ_R45 imply sensibly that the predicted wage first increases with age and then declines. Third, for each sex, the fact that the value of adjusted R-square generated by the conventional approach turns out to be even greater than the corresponding value generated by the nonlinear approach suggests that the conventional approach is better in terms of overall goodness-of-fit.

But, the sign of the shortcoming of the conventional approach starts to emerge when we notice that for males, the intercept generated by this approach (7.29708) is substantially smaller than the intercept generated by the nonlinear approach (7.54916). Their values implies that for the males at age 45 and with a Bachelor's degree, the predicted wage is Exp(7.29708)=$1,476 according to the conventional approach and Exp(7.54916)=$1,899 according to the nonlinear approach. Compared with the observed mean wage of this group of males ($1,932), the conventional approach under-predicts the target by a hefty 24%, whereas the nonlinear approach under-predicts it only slightly by 2%. For the corresponding group of females, the predicted mean wage is Exp(6.71578)=$825 according to the conventional approach and Exp(7.00964)=$1,107 according to the nonlinear approach. Compared with the corresponding observed mean wage of $1,087, the conventional approach under-predicts the target again by 24%, whereas the nonlinear approach *over*-predicts it slightly by 2%.

To get a broader perspective, we compute the predicted wage structures from the estimated coefficients in Table 1, according to equations (3) and (4). The resulting age patterns are contrasted against the observed patterns for the males and females with post-graduate education in Figures (3) and (4). The shortcoming of the conventional approach becomes more apparent in these figures. Except for the young individuals around 25 years of age, the conventional approach results in very serious under-prediction across the entire age range, whereas the nonlinear approach mostly leads to rather close approximations.

Figure 3. Comparison of Age Patterns of the Wage of US-born Non-Hispanic White **Males** with Post-graduate Degree: (1) Observed, (2) Predicted by Nonlinear Approach, (3) Predicted by Conventional Approach.

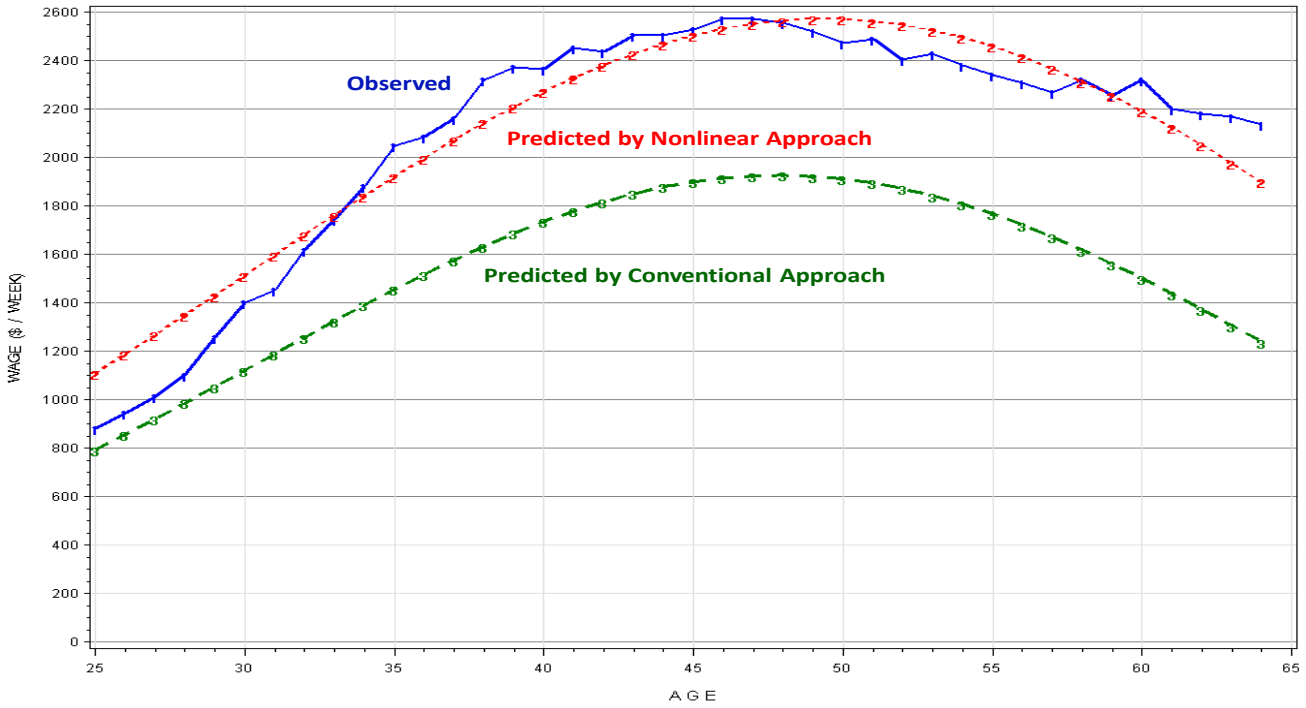
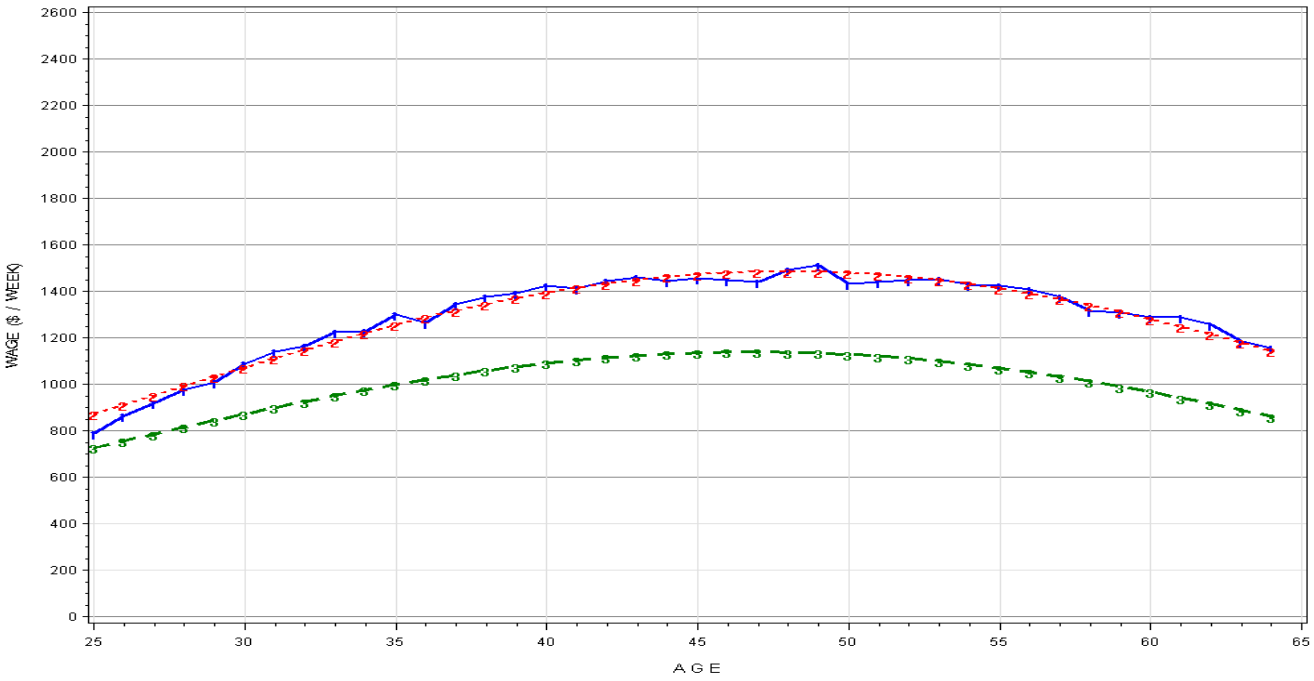


Figure 4. Comparison of Age Patterns of the Wage of US-born Non-Hispanic White **Females** with Post-graduate Degree: (1) Observed, (2) Predicted by Nonlinear Approach, (3) Predicted by Conventional Approach (Green).



To get an overall view without examining too many figures, we divide the age scale into eight 5-year age categories (25-29, 30-34, ... , 60-64) and make the contrasts in terms of these categories in Tables 2.1, 2.2, and 2.3 for males and in Tables 3.1, 3.2, and 3.3 for females at all levels of education. We find that the prediction errors of the conventional approach are all negative and mostly over 20% in magnitude, whereas those of the nonlinear approach are a mixture of positive and negative values that are mostly less than 2% in magnitude. Clearly, the nonlinear approach is superior to the conventional approach.

It is worth noting that for males with a post-graduate degree, even the nonlinear approach yields a few rather large prediction errors (a 21.2% over-prediction for the 25-29 age group, and a 6.6% under-prediction for the 60-64 age group). As shown clearly in Figure 3, the reason for these rather large errors is that there is no simple mathematical function of age such as the quadratic function that can closely represent the shape of the observed wage schedule.

Table 2.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Non-Hispanic White **males** at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	576	462	583	-114	7	-19.8	1.2
30-34	653	515	652	-138	-1	-21.1	-0.2
35-39	726	559	709	-167	-17	-23.0	-2.3
40-44	759	591	755	-168	-4	-22.1	-0.5
45-49	778	610	784	-168	6	-21.6	0.8
50-54	786	614	798	-172	12	-21.9	1.5
55-59	784	602	793	-182	9	-23.2	1.1
60-64	789	577	771	-212	-18	-26.9	-2.3
All	726	565	726	-161	0	-22.2	0.0
	High School Diploma						
25-29	666	557	670	-109	4	-16.4	0.6
30-34	777	646	773	-131	-4	-16.9	-0.5
35-39	868	719	861	-149	-7	-17.2	-0.8
40-44	921	772	928	-149	7	-16.2	0.8
45-49	969	796	967	-173	-2	-17.9	-0.2
50-54	973	792	977	-181	4	-18.6	0.4
55-59	958	757	955	-201	-3	-21.0	-0.3
60-64	909	700	907	-209	-2	-23.0	-0.2
All	882	723	882	-159	0	-18.0	0.0

Table 2.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Non-Hispanic White **males** with some college education and Bachelor's degree.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Some College						
25-29	721	611	746	-110	25	-15.3	3.5
30-34	910	750	900	-160	-10	-17.6	-1.1
35-39	1,060	865	1,034	-195	-26	-18.4	-2.5
40-44	1,140	949	1,140	-191	0	-16.8	0.0
45-49	1,204	983	1,200	-221	-4	-18.4	-0.3
50-54	1,190	966	1,211	-224	21	-18.8	1.8
55-59	1,149	901	1,170	-248	21	-21.6	1.8
60-64	1,137	799	1,086	-338	-51	-29.7	-4.5
All	1,062	861	1,062	-201	0	-18.9	0.0
	Bachelor's Degree						
25-29	939	797	1,003	-142	64	-15.1	6.8
30-34	1,320	1,048	1,314	-272	-6	-20.6	-0.5
35-39	1,643	1,268	1,596	-375	-47	-22.8	-2.9
40-44	1,857	1,425	1,815	-432	-42	-23.3	-2.3
45-49	1,925	1,480	1,922	-445	-3	-23.1	-0.2
50-54	1,814	1,424	1,901	-390	87	-21.5	4.8
55-59	1,724	1,272	1,757	-452	33	-26.2	1.9
60-64	1,677	1,061	1,527	-616	-150	-36.7	-8.9
All	1,609	1,236	1,609	-373	0	-23.2	0.0

Table 2.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Non-Hispanic White **males** with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	1,071	946	1,298	-125	227	-11.7	21.2
30-34	1,634	1,265	1,687	-369	53	-22.6	3.2
35-39	2,195	1,573	2,067	-622	-128	-28.3	-5.8
40-44	2,455	1,814	2,379	-641	-76	-26.1	-3.1
45-49	2,552	1,919	2,546	-633	-6	-24.8	-0.2
50-54	2,435	1,865	2,540	-570	105	-23.4	4.3
55-59	2,301	1,670	2,365	-631	64	-27.4	2.8
60-64	2,214	1,389	2,068	-825	-146	-37.3	-6.6
All	2,228	1,637	2,229	-591	1	-26.5	0.0
	All Levels of Education						
25-29	764	646	802	-118	38	-15.4	5.0
30-34	1,032	832	1,031	-200	-1	-19.4	-0.1
35-39	1,259	985	1,222	-274	-37	-21.8	-2.9
40-44	1,349	1,069	1,333	-280	-16	-20.8	-1.2
45-49	1,404	1,108	1,401	-296	-3	-21.1	-0.2
50-54	1,417	1,123	1,459	-294	42	-20.7	3.0
55-59	1,420	1,072	1,446	-348	26	-24.5	1.8
60-64	1,370	924	1,296	-446	-74	-32.6	-5.4
All	1,253	981	1,254	-272	1	-21.7	0.1

Table 3.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Non-Hispanic White **females** at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	364	281	366	-83	2	-22.8	0.5
30-34	401	310	406	-91	5	-22.7	1.2
35-39	439	333	437	-106	-2	-24.1	-0.5
40-44	470	349	461	-121	-9	-25.7	-1.9
45-49	468	358	473	-110	5	-23.5	1.1
50-54	472	359	475	-113	3	-23.9	0.6
55-59	466	351	465	-115	-1	-24.7	-0.2
60-64	449	337	446	-112	-3	-24.9	-0.7
All	443	336	443	-107	0	-24.2	0.0
	High School Diploma						
25-29	452	360	445	-92	-7	-20.4	-1.5
30-34	489	402	497	-87	8	-17.8	1.6
35-39	539	435	539	-104	0	-19.3	0.0
40-44	563	459	571	-104	8	-18.5	1.4
45-49	592	471	589	-121	-3	-20.4	-0.5
50-54	598	471	593	-127	-5	-21.2	-0.8
55-59	587	459	582	-128	-5	-21.8	-0.9
60-64	549	436	558	-113	9	-20.6	1.6
All	555	444	555	-111	0	-20.0	0.0

Table 3.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Non-Hispanic White **females** with some college education and Bachelor's degree.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Some College						
25-29	530	431	529	-99	-1	-18.7	-0.2
30-34	607	489	606	-118	-1	-19.4	-0.2
35-39	672	535	670	-137	-2	-20.4	-0.3
40-44	714	571	721	-143	7	-20.0	1.0
45-49	753	591	752	-162	-1	-21.5	-0.1
50-54	764	596	763	-168	-1	-22.0	-0.1
55-59	760	583	753	-177	-7	-23.3	-0.9
60-64	716	557	723	-159	7	-22.2	1.0
All	693	547	693	-146	0	-21.1	0.0
	Bachelor's Degree						
25-29	756	644	778	-112	22	-14.8	2.9
30-34	934	723	909	-211	-25	-22.6	-2.7
35-39	1,033	781	1,013	-252	-20	-24.4	-1.9
40-44	1,087	816	1,084	-271	-3	-24.9	-0.3
45-49	1,088	824	1,110	-264	22	-24.3	2.0
50-54	1,075	804	1,090	-271	15	-25.2	1.4
55-59	1,037	760	1,028	-277	-9	-26.7	-0.9
60-64	962	696	932	-266	-30	-27.7	-3.1
All	989	758	990	-231	1	-23.4	0.1

Table 3.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Non-Hispanic White **females** with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	926	794	965	-132	39	-14.3	4.2
30-34	1,169	925	1,149	-244	-20	-20.9	-1.7
35-39	1,332	1,038	1,312	-294	-20	-22.1	-1.5
40-44	1,438	1,114	1,432	-324	-6	-22.5	-0.4
45-49	1,471	1,138	1,482	-333	11	-22.6	0.7
50-54	1,441	1,109	1,458	-332	17	-23.0	1.2
55-59	1,372	1,034	1,369	-338	-3	-24.6	-0.2
60-64	1,247	924	1,227	-323	-20	-25.9	-1.6
All	1,324	1,027	1,324	-297	0	-22.4	0.0
	All Levels of Education						
25-29	617	513	626	-104	9	-16.9	1.5
30-34	748	593	740	-155	-8	-20.7	-1.1
35-39	812	634	804	-178	-8	-21.9	-1.0
40-44	826	646	829	-180	3	-21.8	0.4
45-49	845	657	850	-188	5	-22.2	0.6
50-54	866	667	870	-199	4	-23.0	0.5
55-59	847	642	841	-205	-6	-24.2	-0.7
60-64	754	574	752	-180	-2	-23.9	-0.3
All	796	621	796	-175	0	-22.0	0.0

In the next section, the age factor in f($\underline{\mathbf{b}}$, $\mathbf{X}_i$) will be represented by a set of seven dummy variables. With the 45-49 age group used as the reference category, the seven dummy variables represent the remaining seven 5-year age groups. These dummy variables can help yield a much better prediction for the males with a post-graduate degree than does the quadratic function. Furthermore, they will help us find out the underlying reason for poor performance of the conventional approach.

4. Learning More about the Shortcoming of the Conventional Approach via the Saturated Specification

In addition to using dummy variables for the age factor, we let the function f($\underline{\mathbf{b}}$, $\mathbf{X}_i$) in equations (1) and (2) include all possible interaction terms between the education dummy variables and the age dummy variables. Borrowing the terminology in the field of categorical analysis, this specification is called the *saturated specification*. A nice property of this specification is that the observed mean value of the dependent variable for any combination of

education level and age group category can be perfectly predicted by a regression model via a least-squares estimation method (For more discussions about the nice properties of the saturated specification, see Angrist and Pischke, 2009, pp. 48-51).

The estimated coefficients based on the saturated specification are shown in Table 4 for males and Table 5 for females. Comparing these tables with Table 1, we see that changing the representation of the dependency of wage on age from a quadratic function to a step function turns out to have practically no effect on the estimated coefficients of the education dummy variables. In other words, as far as the inference about the effects of educational attainment on wage is concerned, it does not really matter which of the two ways of quantifying age is used. We also see that the values of adjusted R-square remain practically unchanged after the replacement of the quadratic function by the step function.

Table 4. The estimated coefficients of the saturated specification of regression model for explaining wage via the application of the conventional and nonlinear approach to the ACS data of males.

	Conventional Approach				Nonlinear Approach		
Explanatory Variable	Coefficient	T-statistic	% Effect		Coefficient	T-statistic	% Effect
Intercept	7.28975	2027.1			7.56252	2461.8	
A. Age Factor (Ref: A4549)							
A2529	-0.63793	-124.4	-47		-0.71786	-101.1	-51
A3034	-0.31366	-60.8	-27		-0.37684	-68.1	-31
A3539	-0.12310	-24.4	-12		-0.15838	-33.8	-15
A4044	-0.02168	-4.3	-2		-0.03598	-8.2	-4
A5054	-0.06199	-12.0	-6		-0.05910	-13.0	-6
A5559	-0.15309	-28.6	-14		-0.11037	-22.6	-10
A6064	-0.27105	-41.0	-24		-0.13798	-22.1	-13
B. Education Factor (Ref: Bachelor's deg.)							
ED_PR	-0.88948	-123.7	-59		-0.90516	-67.1	-60
ED_2ND	-0.60823	-132.4	-46		-0.68591	-119.4	-50
ED_SC	-0.40566	-86.5	-33		-0.46875	-91.4	-37
ED_MS	0.25794	43.3	29		0.28197	65.1	33
C. Interaction Terms							
A2529_E1	0.36380	35.3			0.41625	17.6	
A2529_E2	0.28078	41.8			0.34282	29.1	
A2529_E3	0.14963	22.2			0.20450	18.7	
A2529_E5	-0.14562	-13.8			-0.14995	-10.4	
A3034_E1	0.17012	15.6			0.20062	8.6	
A3034_E2	0.10138	14.8			0.15587	14.9	
A3034_E3	0.07442	10.9			0.09655	10.6	
A3034_E5	-0.07287	-8.0			-0.06869	-7.9	
A3539_E1	0.06589	6.2			0.08901	4.2	
A3539_E2	0.02245	3.4			0.04807	5.3	
A3539_E3	0.01684	2.5			0.03051	3.8	
A3539_E5	-0.01186	-1.4			0.00797	1.2	
A4044_E1	-0.00578	-0.6			0.01048	0.5	
A4044_E2	-0.02614	-4.0			-0.01464	-1.7	
A4044_E3	-0.02064	-3.1			-0.01927	-2.6	
A4044_E5	-0.01391	-1.6			-0.00276	-0.4	
A5054_E1	0.07748	7.2			0.06856	3.3	
A5054_E2	0.05959	8.9			0.06288	7.3	
A5054_E3	0.04136	6.1			0.04680	6.2	
A5054_E5	0.00704	0.8			0.01234	2.0	
A5559_E1	0.16278	14.1			0.11719	5.3	
A5559_E2	0.11778	16.3			0.09811	10.2	
A5559_E3	0.07412	10.5			0.06306	7.8	
A5559_E5	0.01112	1.3			0.00697	1.1	
A6064_E1	0.23689	18.7			0.15130	6.4	
A6064_E2	0.12337	14.2			0.07378	6.3	
A6064_E3	0.08724	10.1			0.08065	8.0	
A6064_E5	-0.00095	-0.1			-0.00413	-0.5	
Adj. R-square	0.1859				0.1608		

Table 5. The estimation results of the saturated specification of regression model for explaining wage via the application of the conventional and nonlinear approach to the ACS data of **females**.

	Conventional Approach				Nonlinear Approach		
Explanatory Variable	Coefficient	T-statistic	% Effect		Coefficient	T-statistic	% Effect
Intercept	6.69418	1751.2			6.99235	2087.9	
A. Age Factor (Ref: A4549)							
A2529	-0.23491	-45.5	-21		-0.36437	-66.1	-31
A3034	-0.07757	-14.2	-7		-0.15299	-29.4	-14
A3539	-0.03629	-6.7	-4		-0.05224	-10.7	-5
A4044	-0.01445	-2.7	-1		-0.00107	-0.2	0
A5054	0.01621	2.9	2		-0.01183	-2.4	-1
A5559	-0.03664	-6.0	-4		-0.04801	-8.8	-5
A6064	-0.17928	-22.8	-16		-0.12327	-16.2	-12
B. Education Factor (Ref: Bachelor's deg.)							
ED_PR	-0.83857	-89.5	-57		-0.84489	-47.6	-57
ED_2ND	-0.52780	-106.0	-41		-0.60811	-99.1	-46
ED_SC	-0.30966	-64.1	-27		-0.36822	-73.4	-31
ED_MS	0.33409	52.3	40		0.30115	63.9	35
C. Interaction Terms							
A2529_E1	0.01966	1.4			0.11381	3.7	
A2529_E2	-0.01884	-2.5			0.09414	8.0	
A2529_E3	-0.07337	-10.7			0.01343	1.5	
A2529_E5	-0.13034	-13.6			-0.09789	-10.8	
A3034_E1	-0.05192	-3.5			-0.00152	-0.1	
A3034_E2	-0.11303	-14.5			-0.03873	-3.4	
A3034_E3	-0.11419	-15.9			-0.06277	-7.4	
A3034_E5	-0.08600	-9.3			-0.07631	-10.1	
A3539_E1	-0.00720	-0.5			-0.00966	-0.4	
A3539_E2	-0.07560	-10.1			-0.04266	-4.3	
A3539_E3	-0.07958	-11.3			-0.06194	-8.0	
A3539_E5	-0.05805	-6.4			-0.04687	-6.7	
A4044_E1	0.02297	1.7			0.00624	0.3	
A4044_E2	-0.04410	-6.2			-0.05046	-5.6	
A4044_E3	-0.04165	-6.1			-0.05271	-7.2	
A4044_E5	-0.02999	-3.3			-0.02113	-3.1	
A5054_E1	0.02495	1.8			0.02196	0.8	
A5054_E2	-0.00582	-0.8			0.02047	2.3	
A5054_E3	0.00807	1.1			0.02646	3.6	
A5054_E5	-0.01779	-2.0			-0.00864	-1.3	
A5559_E1	0.05522	3.9			0.04406	1.6	
A5559_E2	0.01981	2.5			0.03804	3.9	
A5559_E3	0.04173	5.5			0.05758	7.2	
A5559_E5	-0.03090	-3.3			-0.02178	-3.0	
A6064_E1	0.13447	8.5			0.08213	2.8	
A6064_E2	0.05169	5.4			0.04639	3.8	
A6064_E3	0.07733	8.0			0.07278	6.9	
A6064_E5	-0.04911	-4.2			-0.04178	-4.2	
Adj. R-square	0.1585				0.1338		

The predicted means of the saturated specifications via the two approaches are compared with the corresponding observed means in Tables 6.1, 6.2, and 6.3 for males and Tables 7.1, 7.2, and 7.3 for females at all levels of education. The superiority of the nonlinear approach over the conventional approach is even clearer here. While the former generates perfect predictions, the latter still leads to rather serious under-predictions, mostly off by more than 20%.

An important point is that the conventional approach can also yield perfect predictions: it allows the saturated specification to predict perfectly the observed means of the *log of wage* for any combination of education level and age group. From this property of the conventional approach, we start to realize that the fundamental reason for the shortcoming of this approach is the following *mathematical principle*:

For a set of positive numbers $\{Z_i$, for i=1,2, ... $\}$ in which at least two are unequal, Exp[mean of $Ln(Z_i)$] is always less than the mean of Z_i.

Since Exp[mean of $Ln(Z_i)$] is actually the *geometric mean* of Z_i, this principle is the same as the fact that for a set of positive numbers in which at least two are unequal, the arithmetic mean is always greater than the geometric mean. The proof of this fact can be found in Wikipedia (http://en.wikipedia.org/wiki/ Jensen%27s_inequality). Our empirical finding demonstrates that weighted geometric mean should also be smaller than weighted arithmetic mean.

Thus, the truth is that the least-squares method, linear or nonlinear, is a wonderful thing. The reason for the problem of the conventional approach is that the targets for the estimation method are the geometric means that are incapable of reflecting properly the observed the wage structure. In short, *the log transformation of the dependent variable is the ultimate source of the under-prediction problem of the conventional approach*.[6]

[6] Although the log function appears to a simple monotonic transformation, it can lead to a contradiction. For example, we find in Lin (2013) that according to the micro data of the 2000 US census and the 2010 ACS, the average wage of the immigrants from Taiwan increased from 2000 to 2010, whereas their average log of wage decreased in the same period.

Table 6.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Non-Hispanic White **males** at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	576	458	576	-118	0	-20.5	0.0
30-34	653	522	653	-131	0	-20.1	0.0
35-39	726	569	726	-157	0	-21.6	0.0
40-44	759	586	759	-173	0	-22.8	0.0
45-49	778	602	778	-176	0	-22.6	0.0
50-54	786	611	786	-175	0	-22.3	0.0
55-59	784	608	784	-176	0	-22.4	0.0
60-64	789	582	789	-207	0	-26.2	0.0
All	726	564	726	-162	0	-22.3	0.0
	High School Diploma						
25-29	666	558	666	-108	0	-16.2	0.0
30-34	777	645	777	-132	0	-17.0	0.0
35-39	868	721	868	-147	0	-16.9	0.0
40-44	921	760	921	-161	0	-17.5	0.0
45-49	969	798	969	-171	0	-17.6	0.0
50-54	973	796	973	-177	0	-18.2	0.0
55-59	958	770	958	-188	0	-19.6	0.0
60-64	909	688	909	-221	0	-24.3	0.0
All	882	723	882	-159	0	-18.0	0.0

Table 6.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Non-Hispanic White **males** with some college education and Bachelor's degree.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Some College						
25-29	721	599	721	-122	0	-16.9	0.0
30-34	910	769	910	-141	0	-15.5	0.0
35-39	1,060	878	1,060	-182	0	-17.2	0.0
40-44	1,140	936	1,140	-204	0	-17.9	0.0
45-49	1,204	977	1,204	-227	0	-18.9	0.0
50-54	1,190	957	1,190	-233	0	-19.6	0.0
55-59	1,149	902	1,149	-247	0	-21.5	0.0
60-64	1,137	813	1,137	-324	0	-28.5	0.0
All	1,062	860	1,062	-202	0	-19.0	0.0
	Bachelor's Degree						
25-29	939	774	939	-165	0	-17.6	0.0
30-34	1,320	1,071	1,320	-249	0	-18.9	0.0
35-39	1,643	1,296	1,643	-347	0	-21.1	0.0
40-44	1,857	1,434	1,857	-423	0	-22.8	0.0
45-49	1,925	1,465	1,925	-460	0	-23.9	0.0
50-54	1,814	1,377	1,814	-437	0	-24.1	0.0
55-59	1,724	1,257	1,724	-467	0	-27.1	0.0
60-64	1,677	1,117	1,677	-560	0	-33.4	0.0
All	1,609	1,235	1,609	-374	0	-23.2	0.0

Table 6.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Non-Hispanic White **males** with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	1,071	946	1,071	-125	0	-11.7	0.0
30-34	1,634	1,265	1,634	-369	0	-22.6	0.0
35-39	2,195	1,573	2,195	-622	0	-28.3	0.0
40-44	2,455	1,814	2,455	-641	0	-26.1	0.0
45-49	2,552	1,919	2,552	-633	0	-24.8	0.0
50-54	2,435	1,865	2,435	-570	0	-23.4	0.0
55-59	2,301	1,670	2,301	-631	0	-27.4	0.0
60-64	2,214	1,389	2,214	-825	0	-37.3	0.0
All	2,228	1,637	2,228	-591	0	-26.5	0.0
	All Levels of Education						
25-29	764	632	764	-132	0	-17.3	0.0
30-34	1,032	846	1,032	-186	0	-18.0	0.0
35-39	1,259	1,007	1,259	-252	0	-20.0	0.0
40-44	1,349	1,066	1,349	-283	0	-21.0	0.0
45-49	1,404	1,100	1,404	-304	0	-21.7	0.0
50-54	1,417	1,101	1,417	-316	0	-22.3	0.0
55-59	1,420	1,068	1,420	-352	0	-24.8	0.0
60-64	1,370	947	1,370	-423	0	-30.9	0.0
All	1,253	980	1,253	-273	0	-21.8	0.0

Table 7.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the *saturated* specification: Non-Hispanic White **females** at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	364	282	364	-82	0	-22.5	0.0
30-34	401	307	401	-94	0	-23.4	0.0
35-39	439	334	439	-105	0	-23.9	0.0
40-44	470	352	470	-118	0	-25.1	0.0
45-49	468	349	468	-119	0	-25.4	0.0
50-54	472	364	472	-108	0	-22.9	0.0
55-59	466	356	466	-110	0	-23.6	0.0
60-64	449	334	449	-115	0	-25.6	0.0
All	443	336	443	-107	0	-24.2	0.0
	High School Diploma						
25-29	452	370	452	-82	0	-18.1	0.0
30-34	489	394	489	-95	0	-19.4	0.0
35-39	539	426	539	-113	0	-21.0	0.0
40-44	563	449	563	-114	0	-20.2	0.0
45-49	592	476	592	-116	0	-19.6	0.0
50-54	598	481	598	-117	0	-19.6	0.0
55-59	587	469	587	-118	0	-20.1	0.0
60-64	549	419	549	-130	0	-23.7	0.0
All	555	444	555	-111	0	-20.0	0.0

Table 7.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the *saturated* specification: Non-Hispanic White **females** with some college education and Bachelor's degree.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Some College						
25-29	530	435	530	-95	0	-17.9	0.0
30-34	607	489	607	-118	0	-19.4	0.0
35-39	672	528	672	-144	0	-21.4	0.0
40-44	714	560	714	-154	0	-21.6	0.0
45-49	753	593	753	-160	0	-21.2	0.0
50-54	764	607	764	-157	0	-20.5	0.0
55-59	760	596	760	-164	0	-21.6	0.0
60-64	716	535	716	-181	0	-25.3	0.0
All	693	547	693	-146	0	-21.1	0.0
	Bachelor's Degree						
25-29	756	639	756	-117	0	-15.5	0.0
30-34	934	747	934	-187	0	-20.0	0.0
35-39	1,033	779	1,033	-254	0	-24.6	0.0
40-44	1,087	796	1,087	-291	0	-26.8	0.0
45-49	1,088	808	1,088	-280	0	-25.7	0.0
50-54	1,075	821	1,075	-254	0	-23.6	0.0
55-59	1,037	779	1,037	-258	0	-24.9	0.0
60-64	962	675	962	-287	0	-29.8	0.0
All	989	757	989	-232	0	-23.5	0.0

Table 7.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the *saturated* specification: Non-Hispanic White **females** with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	926	783	926	-143	0	-15.4	0.0
30-34	1,169	958	1,169	-211	0	-18.0	0.0
35-39	1,332	1,027	1,332	-305	0	-22.9	0.0
40-44	1,438	1,079	1,438	-359	0	-25.0	0.0
45-49	1,471	1,128	1,471	-343	0	-23.3	0.0
50-54	1,441	1,126	1,441	-315	0	-21.9	0.0
55-59	1,372	1,054	1,372	-318	0	-23.2	0.0
60-64	1,247	898	1,247	-349	0	-28.0	0.0
All	1,324	1,027	1,324	-297	0	-22.4	0.0
	All Levels of Education						
25-29	617	513	617	-104	0	-16.9	0.0
30-34	748	603	748	-145	0	-19.4	0.0
35-39	812	627	812	-185	0	-22.8	0.0
40-44	826	631	826	-195	0	-23.6	0.0
45-49	845	654	845	-191	0	-22.6	0.0
50-54	866	680	866	-186	0	-21.5	0.0
55-59	847	656	847	-191	0	-22.6	0.0
60-64	754	555	754	-199	0	-26.4	0.0
All	796	621	796	-175	0	-22.0	0.0

Why do the under-predictions of the conventional approach turn out to be so large? To answer this question, we first compute for each gender the *coefficient of variation*[7], *skewness*[8], and *kurtosis* of wage distribution for the set of individuals in each of 40 age-by-education cells. Using the population size in each cell as the weight, we then perform many weighted linear regressions of the severity of the under-prediction on various combinations of these three indices of wage distribution. In each cell, the severity of the under-prediction is defined as the magnitude of the under-prediction, expressed as a percentage of the observed mean wage.

The main results of our exploration are shown in Table 8. Here we see that the severity of the under-prediction depends *very strongly* on the coefficient of variation, *moderately* on the skewness, and *modestly* on the kurtosis. With respect to the directions of the effects, the severity of the under-prediction tends to increase with the coefficient of variation, to decrease with the skewness, and to increase with kurtosis. Thus, we claim that *the serious under-prediction of the*

[7] Coefficient of variation is defined as the ratio of mean to standard deviation, multiplied by 100%.

[8] Let Z_i be the standard score of Y_i. Skewness is defined as [sum of Z_i^3 across all observations]*n/[(n-1)(n-2)], where n is the sample size. Kurtosis is defined as {[sum of Z_i^4 across all observations] * n(n+1)/[(n-1)(n-2)(n-3)]} - {3(n-1)^2/[(n-2)(n-3)]}, where n is the sample size. Kurtosis measures the heaviness of tails.

observed wage structure by the conventional approach is mainly due to the fact that the wage distributions of the US-born non-Hispanic Whites at most educational levels and in most age groups are highly unequal.

Table 8. The results of regressing the severity of the under-estimation by the conventional approach on the coefficient of variation, skewness, and kurtosis of the wage distribution of US-born Non-Hispanic Whites.

Explanatory Variable	Specification 1		Specification 2		Specification 3	
	Coefficient	T-Satistic	Coefficient	T-Satistic	Coefficient	T-Satistic
	Male Panel					
INTERCEPT	-15.754	-6.7	-11.058	-5.2	-7.071	-2.5
CV	0.443	15.6	0.418	17.8	0.397	16.2
SKEWNESS			-0.485	-4.6	-1.101	-3.5
KURTOSIS					0.014	2.1
Adj. R-Square	0.86		0.91		0.92	
Additional Contribution to Adj. R-Square			0.05		0.01	
	Female Panel					
INTERCEPT	-2.561	-1.1	-3.268	-1.8	-2.095	-1.3
CV	0.291	10.7	0.326	14.1	0.354	15.9
SKEWNESS			-0.289	-4.7	-1.050	-4.4
KURTOSIS					0.012	3.3
Adj. R-Square	0.74		0.83		0.87	
Additional Contribution to Adj. R-Square			0.09		0.03	

Note: For each of the 40 combinations of 5 education levels and 8 age groups in the saturated specification of $f(\beta, X_i)$ in equation (1), the dependent variable in this linear regression is $[(y-\underline{y})/y]*100\%$, where y is the observed mean wage, and $\underline{y}$ is the mean wage predicted by the conventional method.
For each of the 40 combinations of 5 education levels and 8 age groups, the values of CV, skewness, and kurtosis are computed from the from the wage distribution in the combination.
The linear regression uses the population size in each of the 40 combinations as the weight variable.

This claim is further substantiated by our parallel work on Taiwan's MUS data reported in Appendix A. Figures 5 and 6 show (1) that both in the US and in Taiwan, the severity of the under-prediction by the conventional approach tends to increase with the coefficient of variation of the wage distribution in each combination of education and age, and (2) that both the under-prediction problem of the conventional approach and the coefficient of variation tend to be much higher in the US than in Taiwan.

The rather serious shortcoming of the conventional approach demonstrated by our analysis of the ACS data clearly indicates that researchers who are interested in *predicting* wage by a set of explanatory variables should stop using the conventional approach. It is better to use either the nonlinear approach or the *direct linear approach*, which is a simplification of the conventional approach by replacing $Ln(Y_i)$ in equation (1) by Y_i.

What should be the criteria for the selection between the nonlinear approach and the direct linear approach? One criterion is predictive accuracy. Our additional analysis of the ACS data has revealed that the predictive accuracy of the direct linear approach is essentially as good as that of the nonlinear approach. Another criterion is more conceptual: Should the joint effect of changes in two explanatory variables be assumed to be a *multiplicative* or an *additive* function of the effects of the two variables?[9] To the extent that the former is more likely to reflect the situation in the real-world, the nonlinear approach should be chosen.

[9] To demonstrate what we mean by "multiplicative function", we use the example of the joint effect of changing educational attainment from Bachelor's level to post-graduate level and changing age group from 45-49 to 50-54, based on the saturated specification via the nonlinear approach. From Table 5, we see that the coefficient of the dummy variable representing the post-graduate level is 0.30115, the coefficient of the dummy variable representing the 50-54 age group is -0.01183, and the coefficient of the interaction between these two dummy variables is -0.00864. The joint effect in question is then [exp(0.30115) * exp(-0.01183) * exp(0.00864) -1] * 100% = [(1.35141) * (0.98824) * (0.99140) - 1]*100% = 32%. Note that the effect of changing education alone is [1.35141 - 1] * 100% = 35%, whereas the effect of changing age group alone is [0.98824 - 1] * 100% = -1%. Clearly, much of the effect of the joint change comes from the change in education.

Figure 5. The severity of the under-prediction of the observed mean wage by the conventional approach versus the **coefficient of variation** of the wage distribution, based on the **male** data of the 2005-2007 ACS and Taiwan's 2001-2010 MPS. Each point represents a combination of an education level and an age group.

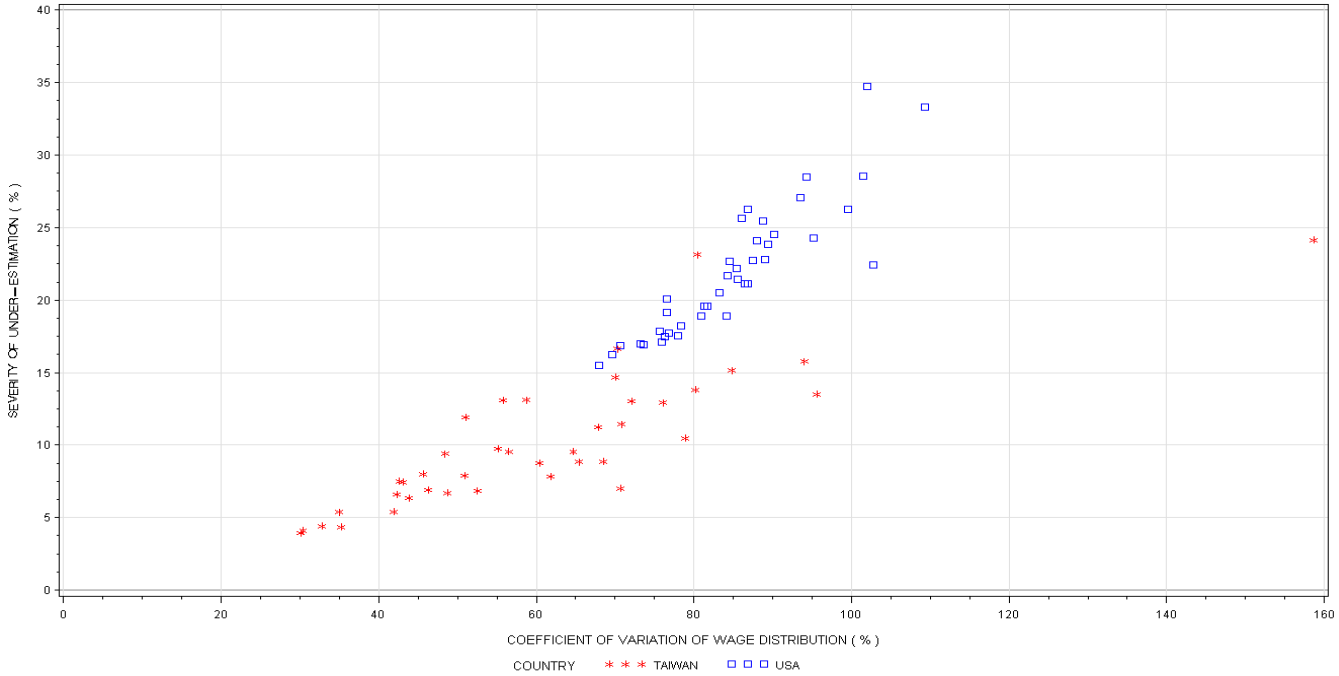
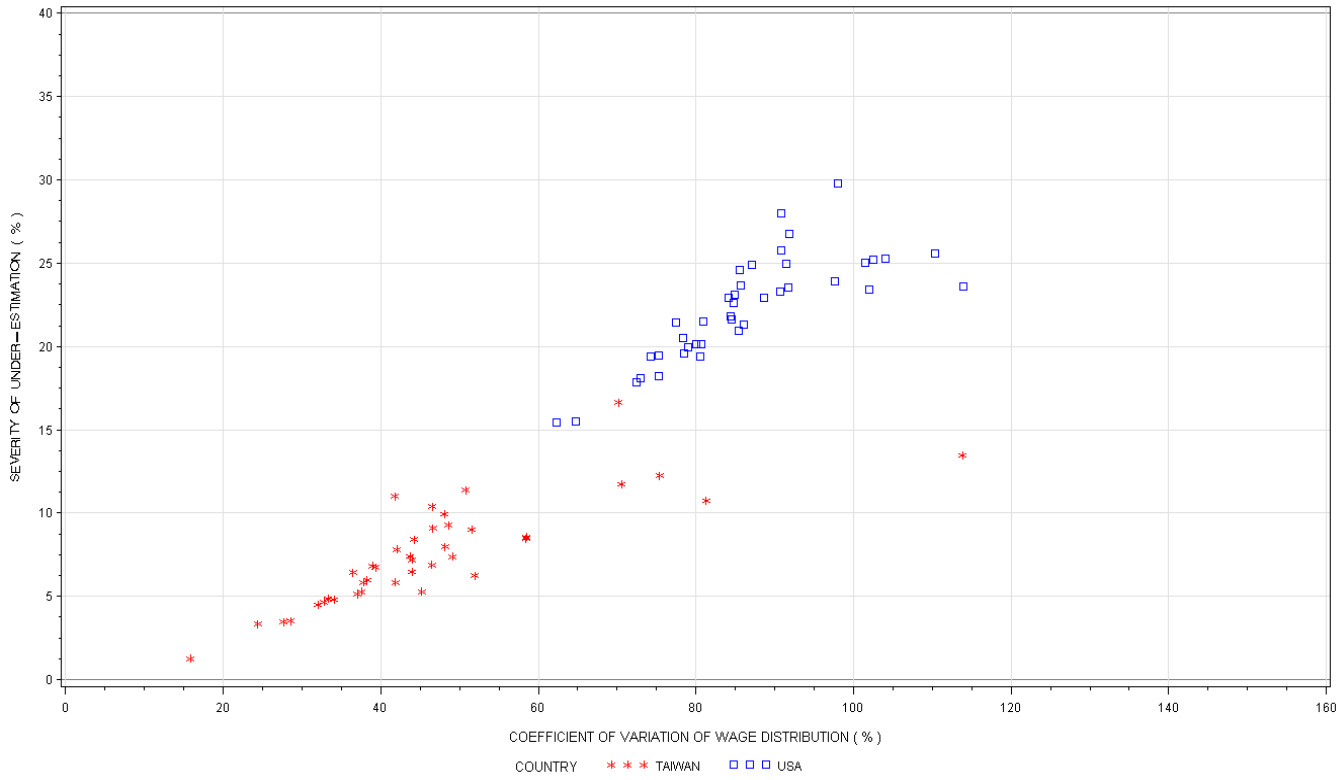


Figure 6. The severity of the under-prediction of the observed mean wage by the conventional approach versus the **coefficient of variation** of the wage distribution, based on the **female** data of the 2005-2007 ACS and Taiwan's 2001-2010 MPS. Each point represents a combination of an education level and an age group.



What is the reason for the fact that the conventional approach has been much more popular than the direct linear approach? We believe the reason is that most researchers are motivated by the question: What would be the proportional (or percentage) change in wage if an explanatory variable is increased by one unit? The estimated coefficients of the conventional method can be easily transformed into the proportional changes in question, whereas those of the direct linear approach alone can not.[10] Thus, we believe that most researchers would prefer the nonlinear approach to the direct linear approach.

It is worth noting that econometricians whose research focus is on the estimated coefficient of a *causal* variable (especially those who have strong preference for the fixed effects estimation method) may not be interested in the predicted wage structure at all. However, we want to draw the attention of these econometricians to our finding that *the estimated coefficient of an explanatory variable via the conventional approach can be quite far off*. For example, we see in Table 4that according to the conventional approach, the estimated coefficient for the dummy variable representing some college education turns out to be -0.40566, which is rather different from the corresponding value generated via the nonlinear approach (-0.46875).

5. A Concise Specification that Works Well for Practical Purpose

In our ongoing research on the wage structures of foreign-born wage earners from many countries, we found some cases in which the quadratic function is not suitable for representing the age pattern of wage. Thus, we prefer step function over quadratic function for representing the age factor in designing our regression model.

Comparing the full and saturated specifications shown in the previous two sections, we see that the step function representation can lead to an annoyingly large number of explanatory variables. Beyond being annoying, in the case of the wage earners from a minor source country like Taiwan or Japan, the sample size can be too small to yield trustworthy estimated coefficients for most of the interaction terms.

Using the male data, we want to demonstrate in this section that the removal of a large number of interaction terms in the step-function representation can have a relatively small effect on the regression model's ability to predict the underlying wage structure, if the researcher has a good idea about the key features of the wage structure.

[10] Since the coefficient of an explanatory variable in the conventional approach is a partial derivative of wage divided by wage, many economists like Borjas (1985) and Lin (2003) take the coefficient as representing the proportional change in wage due to a unit increase in the explanatory variable. Thus, this practice makes the conventional approach even more appealing than the direct linear approach. But, this practice is increasing incorrect as the estimated coefficient deviates more and more from the range between -0.03 and 0.03. For the estimated coefficients of the education dummy variables in our work, this practice is simply wrong.

In Figure 1, we see that a key feature of the male wage structure is that the range of observed mean wages among the five educational levels is much smaller in the 25-29 age group than in other age groups. Thus, we decide to keep only the interaction terms between the four education variables and the dummy variable for representing the 25-29 age group. For simplicity, we call this specification a *concise specification*.

Based on the male data, the estimated coefficients of the regression models with the concise specification via the conventional and nonlinear approaches are shown in Table 9. Comparing Table 9 and Table 4, we see that the general pattern of the estimated coefficients is not much affected by the deletion of many interaction terms from the saturated specification. The deletion, on the other hand, has resulted in large increases in the magnitude of most of the associated t-statistics, which is partly a consequence of the increased degrees of freedom.

The predicted wage structures via the two approaches are shown for males in Tables 10.1, 10.2, and 10.3. We see in these tables that with the concise specification, the prediction errors are still much smaller for the nonlinear approach than for the conventional approach.

In our ongoing research that attempts to explain the huge wage gap of the Taiwan-born wage earners between the US and Canada, we have expanded this concise specification by including two other explanatory factors: (1) the age at entry into the host country, represented by a dummy variable that assumes the value of 1 if the entry age in question is too old (i.e. >=35); and (2) the recency of the entry, represented by another dummy variable that assumes the value of 1 if the time of the entry is in the 2000s (Liaw, Lin, and Liu, 2014). Applied to the 2005-2007ACS data and the long-form records of the 2006 Canadian census via the nonlinear approach, this expanded concise specification helps us see the big picture effectively. We find that much of the observed wage gap can be accounted for by two factors: educational attainment and age of entry. With respect to the former, the wage structure of the US, especially at the post-graduate level, is much better than that of Canada, and those who went to the US have a much better educational composition than those who ended up in Canada. With respect to the latter, the wage disadvantage of entering at age 35 or older is much greater in Canada than in the US, and a much higher proportion of those who entered Canada have such entry age than their counterparts who entered the US. In short, the use of the concise specification of the regression model with dummy explanatory variables via the nonlinear approach is an effective way to conduct substantively meaningful research.

Table 9. The estimation results of a concise specification of regression models for explaining the wages of the US-born Non-Hispanic White males via the application of the conventional and nonlinear approach to the 2005-2007 ACS data.

	Conventional Approach			Nonlinear Approach		
Explanatory Variable	Coefficient	T-statistic	% Effect	Coefficient	T-statistic	% Effect
Intercept	7.26134	3481.5		7.54952	3584.8	
A. Age Factor (Ref: A4549)						
A2529	-0.60951	-144.8	-46	-0.70487	-104.6	-51
A3034	-0.25742	-104.5	-23	-0.35127	-104.9	-30
A3539	-0.10699	-44.9	-10	-0.14178	-51.1	-13
A4044	-0.03742	-16.1	-4	-0.04189	-16.2	-4
A5054	-0.02452	-10.4	-2	-0.03590	-14.0	-4
A5559	-0.08658	-34.5	-8	-0.08288	-30.3	-8
A6064	-0.19376	-65.6	-18	-0.11400	-34.6	-11
B. Education Factor (Ref: Bachelor's deg.)						
ED_PR	-0.80524	-262.3	-55	-0.83510	-137.7	-57
ED_2ND	-0.56269	-291.4	-43	-0.64172	-240.3	-47
ED_SC	-0.37301	-194.3	-31	-0.43702	-189.2	-35
ED_MS	0.24569	103.2	28	0.28024	144.9	32
C. Interaction Terms						
A2529_E1	0.27956	34.9		0.34619	17.0	
A2529_E2	0.23524	44.6		0.29862	28.0	
A2529_E3	0.11698	22.4		0.17277	17.3	
A2529_E5	-0.13337	-14.8		-0.14821	-10.7	
Adj. R-square	0.1849			0.1602		

Table 10.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on a concise specification: Non-Hispanic White **males** at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	576	458	576	-118	0	-20.5	0.0
30-34	653	492	580	-161	-73	-24.7	-11.2
35-39	726	572	715	-154	-11	-21.2	-1.5
40-44	759	613	790	-146	31	-19.2	4.1
45-49	778	637	824	-141	46	-18.1	5.9
50-54	786	621	795	-165	9	-21.0	1.1
55-59	784	584	759	-200	-25	-25.5	-3.2
60-64	789	524	735	-265	-54	-33.6	-6.8
All	726	566	723	-160	-3	-22.0	-0.4
	High School Diploma						
25-29	666	558	666	-108	0	-16.2	0.0
30-34	777	627	704	-150	-73	-19.3	-9.4
35-39	868	729	868	-139	0	-16.0	0.0
40-44	921	782	959	-139	38	-15.1	4.1
45-49	969	811	1000	-158	31	-16.3	3.2
50-54	973	792	965	-181	-8	-18.6	-0.8
55-59	958	744	921	-214	-37	-22.3	-3.9
60-64	909	668	892	-241	-17	-26.5	-1.9
All	882	724	880	-158	-2	-17.9	-0.2

Table 10.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on a concise specification: Non-Hispanic White **males** with some college education and Bachelor's degree.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Some College						
25-29	721	599	721	-122	0	-16.9	0.0
30-34	910	758	864	-152	-46	-16.7	-5.1
35-39	1,060	881	1,065	-179	5	-16.9	0.5
40-44	1,140	945	1,177	-195	37	-17.1	3.2
45-49	1,204	981	1,227	-223	23	-18.5	1.9
50-54	1,190	957	1,184	-233	-6	-19.6	-0.5
55-59	1,149	899	1,130	-250	-19	-21.8	-1.7
60-64	1,137	808	1,095	-329	-42	-28.9	-3.7
All	1,062	860	1,060	-202	-2	-19.0	-0.2
	Bachelor's Degree						
25-29	939	774	939	-165	0	-17.6	0.0
30-34	1,320	1,101	1,337	-219	17	-16.6	1.3
35-39	1,643	1,280	1,649	-363	6	-22.1	0.4
40-44	1,857	1,372	1,822	-485	-35	-26.1	-1.9
45-49	1,925	1,424	1,900	-501	-25	-26.0	-1.3
50-54	1,814	1,390	1,833	-424	19	-23.4	1.0
55-59	1,724	1,306	1,749	-418	25	-24.2	1.5
60-64	1,677	1,173	1,695	-504	18	-30.1	1.1
All	1,609	1,232	1,610	-377	1	-23.4	0.1

Table 10.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on a concise specification: Non-Hispanic White **males** with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	1,071	866	1,071	-205	0	-19.1	0.0
30-34	1,634	1,408	1,770	-226	136	-13.8	8.3
35-39	2,195	1,636	2,182	-559	-13	-25.5	-0.6
40-44	2,455	1,754	2,411	-701	-44	-28.6	-1.8
45-49	2,552	1,821	2,514	-731	-38	-28.6	-1.5
50-54	2,435	1,777	2,426	-658	-9	-27.0	-0.4
55-59	2,301	1,670	2,314	-631	13	-27.4	0.6
60-64	2,214	1,500	2,243	-714	29	-32.2	1.3
All	2,228	1,630	2,232	-598	4	-26.8	0.2
	All Levels of Education						
25-29	764	632	764	-132	0	-17.3	0.0
30-34	1,032	855	1,010	-177	-22	-17.2	-2.1
35-39	1,259	1,004	1,260	-255	1	-20.3	0.1
40-44	1,349	1,054	1,361	-295	12	-21.9	0.9
45-49	1,404	1,091	1,414	-313	10	-22.3	0.7
50-54	1,417	1,101	1,416	-316	-1	-22.3	-0.1
55-59	1,420	1,076	1,413	-344	-7	-24.2	-0.5
60-64	1,370	959	1,360	-411	-10	-30.0	-0.7
All	1,253	979	1,253	-274	0	-21.9	0.0

6. Conclusion

In explaining wage or income by personal attributes (e.g. educational attainment, age, and ethnicity) in a regression model, many researchers choose to use the natural log of wage or income as the dependent variable and then to estimate the unknown coefficients by a reasonable version of the least-squares method. We call this approach the *conventional approach*.

Using the micro data of the 2005-2007 ACS on the Non-Hispanic White male and female wage earners, we have shown that the conventional approach has the shortcoming of seriously under-predicting the observed wage structure in the space spanned by the values of the explanatory variables. We have also offered a *nonlinear approach* to remedy this shortcoming.

Based on the *saturated specification* of the regression function, we have revealed the *mathematical principle* that accounts for the under-prediction problem of the conventional approach:

For a set of positive numbers $\{Z_i$, for i=1,2, ... $\}$ in which at least two numbers are unequal, Exp[mean of $Ln(Z_i)$] is always less than the mean of Z_i.

To get some insights into the severity of the under-prediction problem of the conventional approach, we have regressed the severity of the under-prediction on (1) the *coefficient of variation*, (2) the *skewness*, and (3) the *kurtosis* of the wage distributions. We found that the severity depends *strongly* on the coefficient of variation, *moderately* on the skewness, and *modestly* on the kurtosis. With respect to the directions of the effects, the severity of the under-prediction tends to *increase* with coefficient of variation, to *decrease* with skewness, and to *increase* with kurtosis. These findings are clearly substantiated by our parallel analysis of the micro data of Taiwan's 2001-2010 Manpower Utilization Survey.

In light of the annoyingly large numbers of coefficients in the saturated specification of the regression function, we have demonstrated that the nonlinear approach can yield a very good prediction of the observed wage structure, even after most interaction terms are deleted. Thus, with a concise specification of the regression function, the nonlinear approach is an effective way to generate substantively meaningful research results.

Since the procedure for nonlinear least-squares estimation in SAS (PROC NLIN) does not have the flexibility to allow the use of a weight variable, we offer a more flexible SAS module that can take in a weight variable, as well as a SAS program that uses this module. A useful feature of our module is that it computes "robust" standard errors for the estimators of the unknown coefficients. These standard errors are robust in the sense that they do not depend on the restrictive homoskedasticity assumption about the error term. To the extent that the use of the conventional approach is partially motivated by the desire to avoid violating this assumption, our module has further weakened the justification for sticking to the conventional approach.

Finally, in light of the fact that economists have a strong tendency to log-transform the dependent variable of their linear-in-coefficient regression models in explaining not only income but also other variables like employment (Angrist and Pischke, 2009), we note that the under-prediction problem that we try to remedy has been an extensive problem in empirical research.

References

Aeberhardt, R., Fougère, D., Pouget, J., and Rathelot, R. (2010). Wages and employment of French workers with African origin. *Journal of Population Economics*, 23:881–905. DOI 10.1007/s00148-009-0266-3. DOI 10.1007/s00148-009-0245-8.

Angrist, J. D. and Pischke, J-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Borjas, George J. (1985). Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants. *Journal of Labor Economics* 3:463–89.

El-Araby Aly, A. and Ragan, J. F. Jr. (2010). Arab immigrants in the United States: how and why do returns to education vary by country of origin? *Journal of Population Economics*, 23: 519-538. DOI 10.1007/s00148-009-0245-8.

Liaw, K-L., Lin, J-P., and Liu, C-C. (2014). Uneven Performance of Taiwan-born Wage Earners in the American and Canadian Job Markets: Important Roles of Educational Attainment and Entry Age. (not published yet)

Lin, C. (2013). Earnings Gap, Cohort Effect and Economic Assimilation of Immigrants from Mainland China, Hong Kong, and Taiwan in the United States. *Review of International Economics*, 21(2): 249-265.

Lin, J-P. (2007). Involuntary Job Turnover in Taiwan: 1996-2000. In Joseph S. Lee and Edward Elgar (ed), *The Labor Market and Economic Development of Taiwan*, 211-237.

Mather, M., Rivers, K. L., and Jacobsen, L. A. (2005). The American Community Survey. *Population Bulletin* 60(3): 1-20.

U.S. Census Bureau. (2009). *A compass for understanding and using American Community Survey data: What PUMS data users need to know.* Washington, DC: U.S. Government Printing Office.

Appendix A. Evidence Based on Taiwan's Data

Here we use the micro data of Taiwan's 2001-2010 Manpower Utilization Survey (MUS) to conduct parallel analysis. The wage variable is the monthly employment income[11], measured in terms of the Taiwanese dollar as of 2010. The sample selected for this study only includes the workers who was in the 25-64 age interval, worked for at least 20 hours per week, and had a reported employment income of more than one dollar.[12] The resulting sample includes 154,827 male records and 92,325 female records. The average weight is 319 for males and 337 for females, implying that the survey represents about 0.3% of the underlying population. For more information about the MUS, see Lin (2007). For simplicity, we call the monthly employment income as wage.

As a consequence of being different from the ACS in questionnaire design, the education factor for the MUS data is represented by the following categories: (1) less than high school, (2)

[11] In Taiwan's MUS, the questionnaire does not make a distinction between "wage and salary income" and "self-employment income". Thus, here we let "wage" be employment income. We found that the very few individuals who reported $1 as their employment income were mostly self-employed and partly employers.

[12] Those with a reported employment income being exactly 1 dollar turned out to be either employers or self-employed persons. Clearly, the reported value is fictitious.

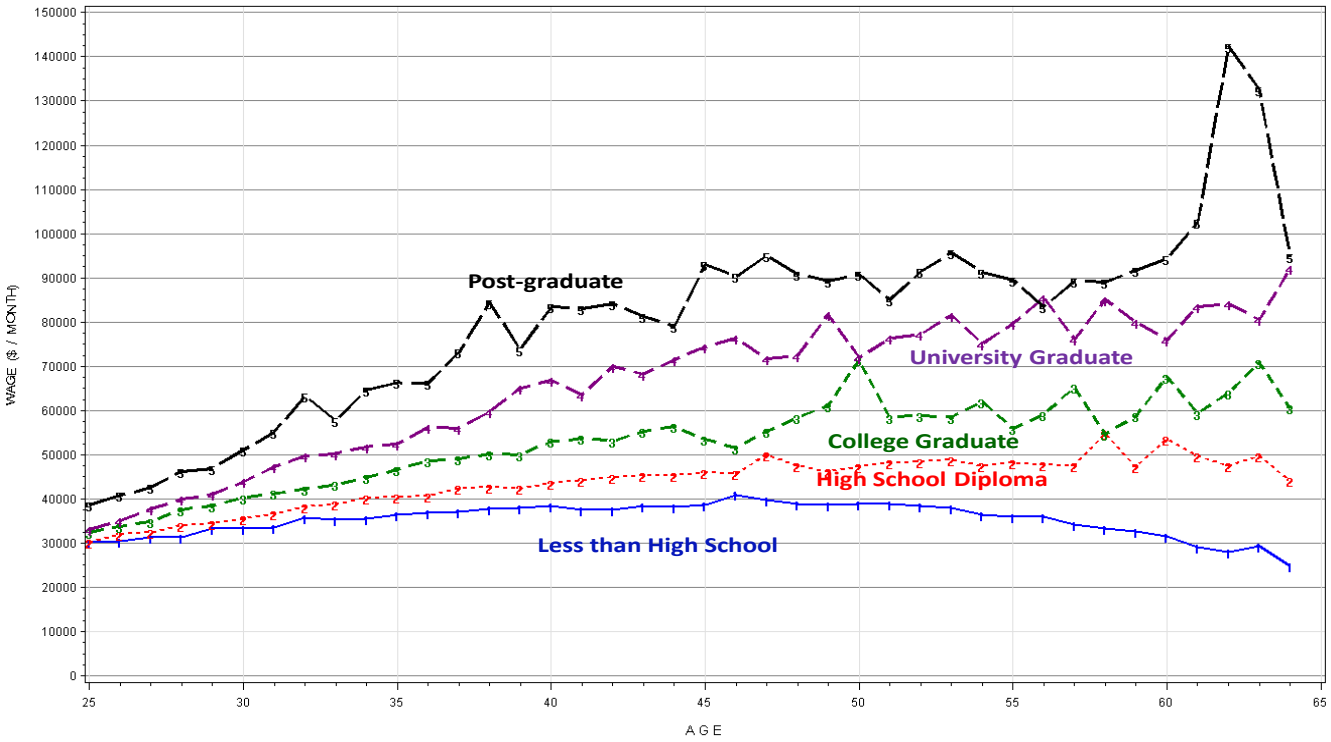
high school graduate, (3) college graduate, (4) university graduate, and (5) post-graduate degree. The categories for the age factor remain identical to those used for the ACS data.

The observed education-by-age wage structures are shown in Appendix Figures 1 and 2 for males and females, respectively. In light of the fact that the sample size is much smaller in the MUS than in the ACS, it is not surprising that the curves in these figures are less smooth than those in Figures 1 and 2, although there is clear evident that the curve for a higher education level tends to be substantially higher than the curve for a lower education level.

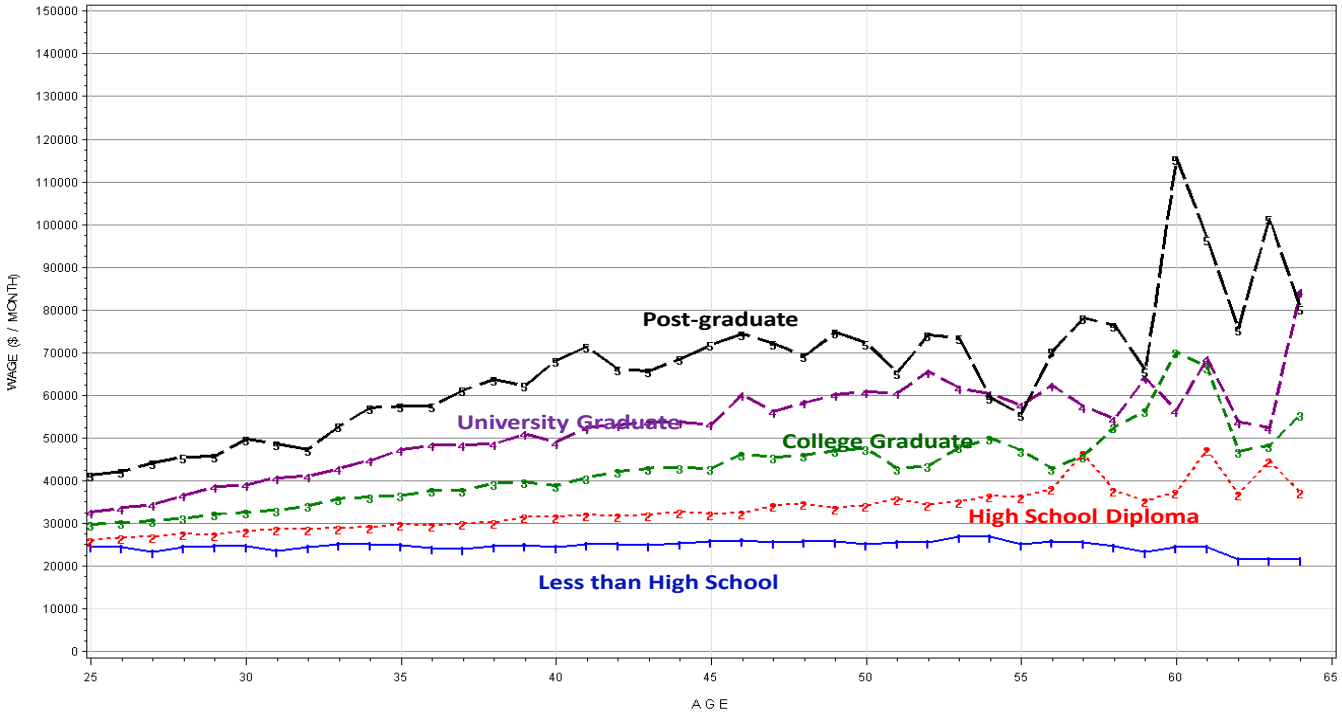
Appendix Table 1 shows the estimation results of the *full specification* of regression models for explaining wage variation via the application of the conventional and nonlinear approaches to Taiwan's MUS data. We see in the table that the estimated coefficients differ between the two approaches. For example, in the male panel, the estimated coefficient of the dummy variable "Post-graduate Degree" turns out to be 0.22574 according to the conventional approach and 0.19531 according to the nonlinear approach, implying that for a male at age 45, a change from a university graduate to a post-graduate is expected to raise wage by [exp(0.22574)-1]*100%=25% according to the conventional approach, and by [exp(0.19531)*100%]= 22% according to the nonlinear approach. The values of Adjusted R-square (0.2383 versus 0.1546) give the misleading impression that for males, the conventional approach is better than the nonlinear approach. For females, the values of Adjusted R-square happen to be nearly identical (0.2411 versus 0.2429), suggesting that the two approaches perform similarly well.

Appendix Figure 3 (for males) and 4 (for females) show clearly that for the university graduates, the observed wage structure is well predicted by the nonlinear approach but is substantially under-predicted by the conventional approach.

Appendix Figure 1. The Age-by-Education Structure of Monthly Wage of **Males** in Taiwan : based on the micro data of the 2001-2010 Manpower Utilization Survey.



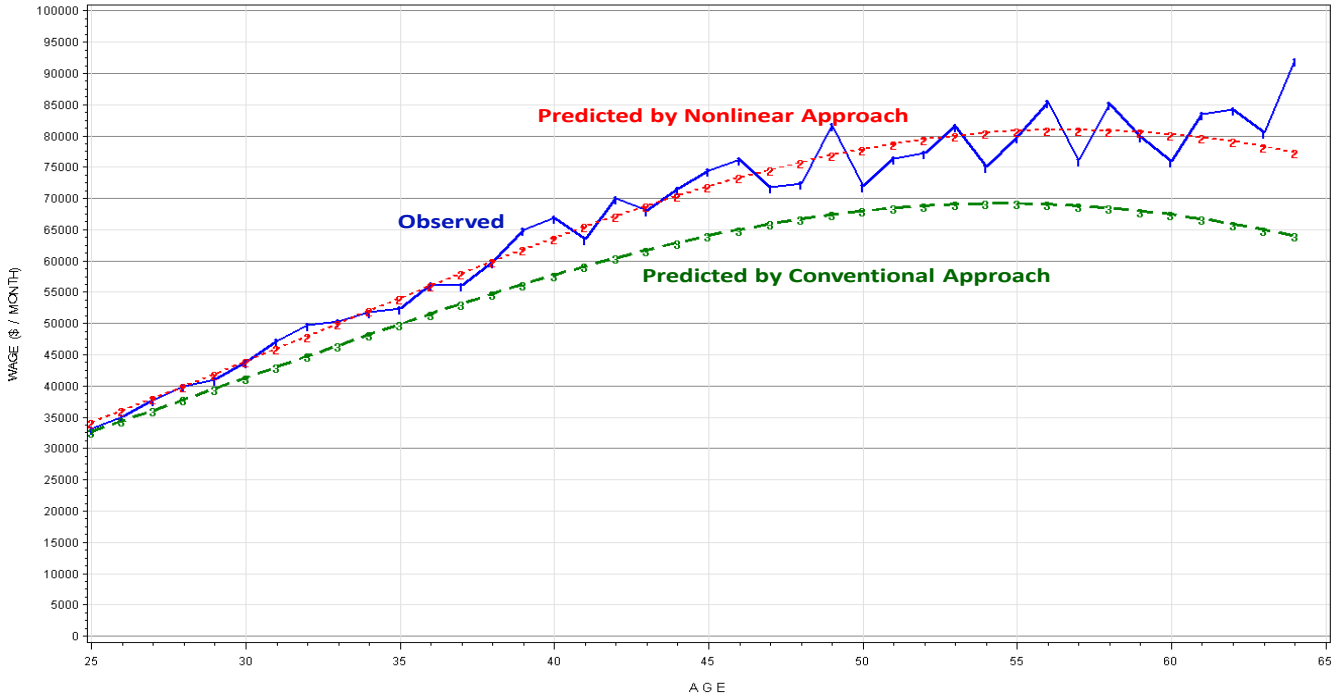
Appendix Figure 2. The Age-by-Education Structure of Monthly Wage of **Female**s in Taiwan : based on the micro data of the 2001-2010 Manpower Utilization Survey.



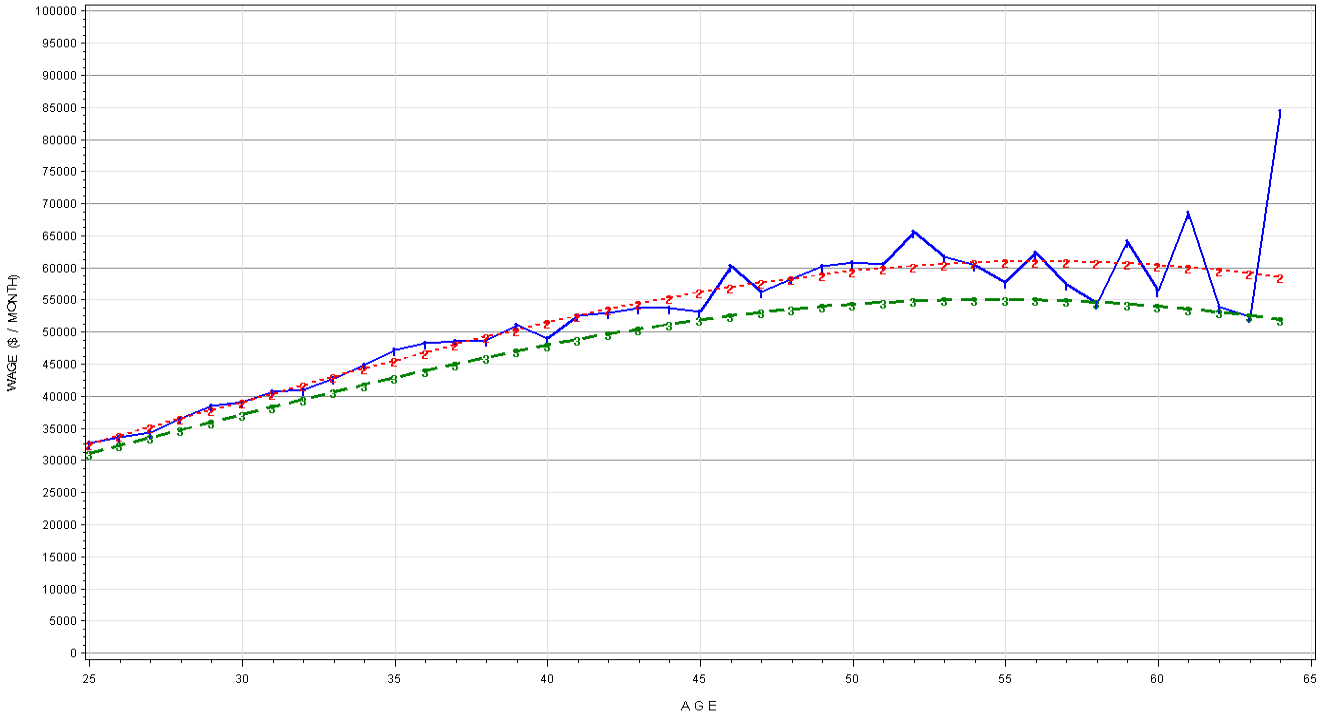
Appendix Table 1. The estimation results of the full specification of regression models for explaining wage via the application of the conventional and nonlinear approach to Taiwan's MUS data.

	Conventional Approach			Nonlinear Approach		
Explanatory Variable	Coefficient	T-statistic	% Effect	Coefficient	T-statistic	% Effect
	Male Panel (Sample Size = 154,827)					
Intercept	11.06754	2435.8		11.18379	2431.3	
A. Age Factor						
AGE_R45	0.01633	41.2		0.02008	53.9	
AGESQ_R45	-0.00086	-27.4		-0.00085	-24.5	
B. Education Factor (Ref: University Grad.)						
Less than High School	-0.59345	-111.8	-45	-0.60385	-90.2	-45
High School Graduate	-0.42163	-80.1	-34	-0.43830	-71.8	-35
College Graduate	-0.22261	-37.3	-20	-0.24260	-36.1	-22
Post-graduate Degree	0.22574	25.9	25	0.19531	25.4	22
C. Interaction Terms						
AGE_R45 * Less than High School	-0.02322	-51.6		-0.02141	-37.8	
AGE_R45 * High School Graduate	-0.01098	-22.8		-0.00991	-18.2	
AGE_R45 * College Graduate	-0.00611	-10.4		-0.00618	-9.8	
AGE_R45 * Post-graduate Degree	0.00179	2.2		-0.00065	-1.0	
AGESQ_R45 * Less than High School	-0.00018	-4.9		-0.00002	-0.3	
AGESQ_R45 * High School Graduate	0.00025	6.6		0.00037	7.6	
AGESQ_R45 * College Graduate	0.00019	4.1		0.00022	3.9	
AGESQ_R45 * Post-graduate Degree	-0.00011	-1.7		-0.00007	-1.1	
Adj. R-square	0.2383			0.1546		
	Female Panel (Sample Size = 92,325)					
Intercept	10.85736	2301.0		10.93709	2726.8	
A. Age Factor						
AGE_R45	0.01247	21.0		0.01447	31.1	
AGESQ_R45	-0.00065	-17.5		-0.00064	-19.0	
B. Education Factor (Ref: University Grad.)						
Less than High School	-0.78831	-136.7	-55	-0.78697	-111.5	-54
High School Graduate	-0.54712	-98.8	-42	-0.53415	-94.1	-41
College Graduate	-0.25777	-39.7	-23	-0.25574	-41.0	-23
Post-graduate Degree	0.23756	20.3	27	0.22149	26.7	25
C. Interaction Terms						
AGE_R45 * Less than High School	-0.01546	-23.3		-0.01394	-19.6	
AGE_R45 * High School Graduate	-0.00384	-5.4		-0.00308	-4.6	
AGE_R45 * College Graduate	-0.00011	-0.1		0.00018	0.2	
AGE_R45 * Post-graduate Degree	0.00033	0.2		-0.00176	-1.7	
AGESQ_R45 * Less than High School	0.00027	5.8		0.00040	6.5	
AGESQ_R45 * High School Graduate	0.00069	15.1		0.00066	13.5	
AGESQ_R45 * College Graduate	0.00041	7.6		0.00039	7.1	
AGESQ_R45 * Post-graduate Degree	-0.00005	-0.5		-0.00013	-1.8	
Adj. R-square	0.2411			0.2429		

Appendix Figure 3. Comparison of Age Patterns of the Wage of Taiwan's **Males** who were university graduates: (1) Observed, (2) Predicted by Nonlinear Approach, (3) Predicted by Conventional Approach.



Appendix Figure 4. Comparison of Age Patterns of the Wage of Taiwan's **Females** who were university graduates: (1) Observed, (2) Predicted by Nonlinear Approach, (3) Predicted by Conventional Approach.



The comparisons of the predictive capacities between these two approaches for the full specification of the regression function are shown for all age groups and all education levels in Appendix Tables 2.1, 2.2, and 2.3 for males, and 3.1, 3.2, and 3.3 for females. We see from these tables that the observed wage structure is mostly well predicted by the nonlinear approach but is all under-predicted by the conventional approach, with an overall prediction error of 9.3% for males and 6.8% for females.

Appendix Table 2.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Taiwan's **males** at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	31,403	28,657	30,502	-2,746	-901	-8.7	-2.9
30-34	34,733	32,484	34,616	-2,249	-117	-6.5	-0.3
35-39	37,338	34,982	37,651	-2,356	313	-6.3	0.8
40-44	38,052	35,712	39,139	-2,340	1,087	-6.1	2.9
45-49	39,429	34,680	39,026	-4,749	-403	-12.0	-1.0
50-54	38,082	32,014	37,308	-6,068	-774	-15.9	-2.0
55-59	34,540	28,148	34,238	-6,392	-302	-18.5	-0.9
60-64	28,741	23,401	30,007	-5,340	1,266	-18.6	4.4
All	36,381	32,144	36,383	-4,237	2	-11.6	0.0
	High School Diploma						
25-29	32,606	31,295	33,036	-1,311	430	-4.0	1.3
30-34	37,900	35,351	37,489	-2,549	-411	-6.7	-1.1
35-39	41,745	38,692	41,479	-3,053	-266	-7.3	-0.6
40-44	44,627	41,069	44,780	-3,558	153	-8.0	0.3
45-49	47,112	42,305	47,199	-4,807	87	-10.2	0.2
50-54	48,058	42,317	48,589	-5,741	531	-11.9	1.1
55-59	49,082	41,108	48,885	-7,974	-197	-16.2	-0.4
60-64	49,665	38,784	48,052	-10,881	-1,613	-21.9	-3.2
All	41,897	38,289	41,899	-3,608	2	-8.6	0.0

Appendix Table 2.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Taiwan's **males** with college and university degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	College Degree						
25-29	35,597	34,394	35,937	-1,203	340	-3.4	1.0
30-34	42,336	40,002	42,290	-2,334	-46	-5.5	-0.1
35-39	48,983	45,222	48,491	-3,761	-492	-7.7	-1.0
40-44	54,262	49,354	53,772	-4,908	-490	-9.0	-0.9
45-49	55,676	52,068	57,739	-3,608	2,063	-6.5	3.7
50-54	62,103	53,207	60,187	-8,896	-1,916	-14.3	-3.1
55-59	58,601	52,662	60,828	-5,939	2,227	-10.1	3.8
60-64	64,012	50,373	59,617	-13,639	-4,395	-21.3	-6.9
All	48,629	44,812	48,632	-3,817	3	-7.8	0.0
	University Degree						
25-29	37,593	36,289	38,231	-1,304	638	-3	1.7
30-34	48,405	44,593	47,772	-3,812	-633	-8	-1.3
35-39	57,849	53,118	57,945	-4,731	96	-8	0.2
40-44	68,036	60,457	67,170	-7,579	-866	-11	-1.3
45-49	75,158	65,799	74,459	-9,359	-699	-12	-0.9
50-54	76,359	68,703	79,228	-7,656	2,869	-10	3.8
55-59	81,341	68,811	80,844	-12,530	-497	-15	-0.6
60-64	81,998	66,195	79,294	-15,803	-2,704	-19	-3.3
All	57,783	52,251	57,798	-5,532	15	-10	0.0

Appendix Table 2.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Taiwan's **males** with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	44,205	43,542	47,071	-663	2,866	-1.5	6.5
30-34	58,289	53,779	58,094	-4,510	-195	-7.7	-0.3
35-39	72,829	65,288	70,630	-7,541	-2,199	-10.4	-3.0
40-44	82,282	75,236	81,704	-7,046	-578	-8.6	-0.7
45-49	91,773	82,708	90,377	-9,065	-1,396	-9.9	-1.5
50-54	90,593	86,616	95,424	-3,977	4,831	-4.4	5.3
55-59	88,486	86,680	96,575	-1,806	8,089	-2.0	9.1
60-64	113,646	82,505	93,185	-31,141	-20,461	-27.4	-18.0
All	71,849	66,036	71,902	-5,813	53	-8.1	0.1
	All Levels of Education						
25-29	34,838	33,407	35,259	-1,431	421	-4.1	1.2
30-34	41,621	38,810	41,311	-2,811	-310	-6.8	-0.7
35-39	46,008	42,503	45,764	-3,505	-244	-7.6	-0.5
40-44	48,669	44,643	48,852	-4,026	183	-8.3	0.4
45-49	50,305	45,026	50,344	-5,279	39	-10.5	0.1
50-54	49,424	42,963	49,387	-6,461	-37	-13.1	-0.1
55-59	46,588	39,358	46,764	-7,230	176	-15.5	0.4
60-64	39,974	31,922	39,688	-8,052	-286	-20.1	-0.7
All	45,015	40,808	45,021	-4,207	6	-9.3	0.0

Appendix Table 3.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Taiwan's **females** at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	24,345	22,047	23,439	-2,298	-906	-9.4	-3.7
30-34	24,644	23,035	24,418	-1,609	-226	-6.5	-0.9
35-39	24,548	23,591	25,104	-957	556	-3.9	2.3
40-44	25,031	23,712	25,490	-1,319	459	-5.3	1.8
45-49	25,798	23,406	25,584	-2,392	-214	-9.3	-0.8
50-54	25,926	22,695	25,379	-3,231	-547	-12.5	-2.1
55-59	24,952	21,606	24,880	-3,346	-72	-13.4	-0.3
60-64	23,001	20,186	24,098	-2,815	1,097	-12.2	4.8
All	25,139	22,913	25,139	-2,226	0	-8.9	0.0
	High School Diploma						
25-29	26,852	26,011	27,032	-841	180	-3.1	0.7
30-34	28,711	27,014	28,526	-1,697	-185	-5.9	-0.6
35-39	30,226	28,105	30,136	-2,121	-90	-7.0	-0.3
40-44	31,940	29,264	31,831	-2,676	-109	-8.4	-0.3
45-49	33,404	30,523	33,656	-2,881	252	-8.6	0.8
50-54	35,048	31,861	35,581	-3,187	533	-9.1	1.5
55-59	38,893	33,364	37,728	-5,529	-1,165	-14.2	-3.0
60-64	40,364	34,867	39,861	-5,497	-503	-13.6	-1.2
All	30,609	28,442	30,609	-2,167	0	-7.1	0.0

Appendix Table 3.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Taiwan's **females** with college and university degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	College Degree						
25-29	30,798	29,659	30,774	-1,139	-24	-3.7	-0.1
30-34	34,319	32,733	34,410	-1,586	91	-4.6	0.3
35-39	38,167	35,739	38,050	-2,428	-117	-6.4	-0.3
40-44	41,483	38,498	41,478	-2,985	-5	-7.2	0.0
45-49	45,402	40,987	44,666	-4,415	-736	-9.7	-1.6
50-54	46,185	43,116	47,498	-3,069	1,313	-6.6	2.8
55-59	47,714	44,733	49,760	-2,981	2,046	-6.2	4.3
60-64	62,117	45,926	51,562	-16,191	-10,555	-26.1	-17.0
All	36,638	34,493	36,638	-2,145	0	-5.9	0.0
	University Degree						
25-29	34,987	33,415	35,019	-1,572	32	-4.5	0.1
30-34	41,512	39,341	41,577	-2,171	65	-5.2	0.2
35-39	48,674	44,906	47,886	-3,768	-788	-7.7	-1.6
40-44	52,231	49,576	53,368	-2,655	1,137	-5.1	2.2
45-49	57,419	52,985	57,610	-4,434	191	-7.7	0.3
50-54	61,771	54,724	60,084	-7,047	-1,687	-11.4	-2.7
55-59	59,094	54,887	60,911	-4,207	1,817	-7.1	3.1
60-64	59,249	53,486	59,989	-5,763	740	-9.7	1.2
All	44,291	41,539	44,292	-2,752	1	-6.2	0.0

Appendix Table 3.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the full specification: Taiwan's **females** with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	44,092	42,017	43,864	-2,075	-228	-4.7	-0.5
30-34	50,859	49,137	51,784	-1,722	925	-3.4	1.8
35-39	60,523	56,782	60,279	-3,741	-244	-6.2	-0.4
40-44	68,125	62,726	66,812	-5,399	-1,313	-7.9	-1.9
45-49	72,673	66,981	71,350	-5,692	-1,323	-7.8	-1.8
50-54	69,495	69,359	73,649	-136	4,154	-0.2	6.0
55-59	68,569	69,305	73,047	736	4,478	1.1	6.5
60-64	96,637	66,255	68,976	-30,382	-27,661	-31.4	-28.6
All	56,235	53,215	56,235	-3,020	0	-5.4	0.0
	All Levels of Education						
25-29	31,105	29,829	31,111	-1,276	6	-4.1	0.0
30-34	33,739	31,976	33,734	-1,763	-5	-5.2	0.0
35-39	35,210	32,880	35,100	-2,330	-110	-6.6	-0.3
40-44	35,259	32,830	35,449	-2,429	190	-6.9	0.5
45-49	35,608	32,487	35,520	-3,121	-88	-8.8	-0.2
50-54	35,157	31,603	35,068	-3,554	-89	-10.1	-0.3
55-59	33,242	29,432	33,320	-3,810	78	-11.5	0.2
60-64	28,940	25,004	29,241	-3,936	301	-13.6	1.0
All	33,960	31,647	33,960	-2,313	0	-6.8	0.0

Appendix Tables 4 (for males) and 5 (for females) show the estimation results of the *saturated specification* of regression function via the application of the conventional and nonlinear approaches to Taiwan's MUS data. Again, we see that the corresponding coefficients estimated by the two approaches are not identical. The values of Adjusted R-square yield a misleading impression that the two approaches perform similarly well.

The comparisons of the predictive capacities between these two approaches for the saturated specification of the regression function are shown for all age groups and all education levels in Appendix Tables 6.1, 6.2, and 6.3 for males, and 7.1, 7.2, and 7.3 for females. We see from these tables that the observed wage structure is perfectly predicted by the nonlinear approach but is all under-predicted by the conventional approach, with an overall prediction error of 9.4% for males and 6.8% for females.

Appendix Table 4. The estimation results of the saturated specification of regression model via the application of the conventional and nonlinear approach to the Taiwan's MUS data for males.

	Conventional Approach			Nonlinear Approach		
Explanatory Variable	Coefficient	T-statistic	% Effect	Coefficient	T-statistic	% Effect
Intercept	11.08891	1260.2		11.22734	1435.7	
A. Age Factor (Ref: A4549)						
A2529	-0.59950	-55.9	-45	-0.69276	-51.7	-50
A3034	-0.37117	-33.7	-31	-0.43999	-36.5	-36
A3539	-0.21594	-18.5	-19	-0.26176	-22.1	-23
A4044	-0.06151	-5.1	-6	-0.09954	-8.8	-9
A5054	0.01448	1.1	1	0.01586	1.3	2
A5559	0.04578	3.0	5	0.07907	6.0	8
A6064	0.08508	3.7	9	0.08710	4.6	9
B. Education Factor (Ref: University grad.)						
Less than High School	-0.62828	-63.0	-47	-0.64509	-57.7	-48
High School Graduate	-0.43940	-43.4	-36	-0.46705	-44.3	-37
College Graduate	-0.24401	-20.9	-22	-0.30005	-24.9	-26
Post-graduate Degree	0.23548	14.1	27	0.19973	15.4	22
C. Interaction Terms						
A2529_E1	0.43815	30.5		0.46519	19.7	
A2529_E2	0.29755	23.4		0.32471	18.1	
A2529_E3	0.19429	13.3		0.24550	12.1	
A2529_E5	-0.07030	-3.2		-0.03771	-1.5	
A3034_E1	0.29760	21.3		0.31318	15.7	
A3034_E2	0.19870	15.4		0.22240	13.8	
A3034_E3	0.12409	8.5		0.16608	9.3	
A3034_E5	-0.05277	-2.5		-0.01392	-0.7	
A3539_E1	0.19959	14.3		0.20728	11.6	
A3539_E2	0.13400	10.0		0.14081	9.1	
A3539_E3	0.09921	6.5		0.13369	7.8	
A3539_E5	0.00569	0.3		0.03055	1.6	
A4044_E1	0.04881	3.5		0.06401	3.9	
A4044_E2	0.02629	1.9		0.04536	3.0	
A4044_E3	0.02521	1.6		0.07382	4.4	
A4044_E5	-0.04552	-2.0		-0.00962	-0.5	
A5054_E1	-0.06819	-4.5		-0.05061	-3.1	
A5054_E2	-0.00317	-0.2		0.00402	0.3	
A5054_E3	0.03197	1.8		0.09338	5.2	
A5054_E5	-0.00290	-0.1		-0.02881	-1.4	
A5559_E1	-0.23858	-13.9		-0.21143	-11.2	
A5559_E2	-0.05860	-3.1		-0.03810	-2.0	
A5559_E3	-0.06091	-2.7		-0.02786	-1.2	
A5559_E5	-0.05697	-1.8		-0.11554	-4.6	
A6064_E1	-0.54298	-22.3		-0.40328	-15.2	
A6064_E2	-0.19799	-7.1		-0.03435	-1.2	
A6064_E3	0.00984	0.3		0.05243	1.4	
A6064_E5	0.07258	1.6		0.12667	4.1	
Adj. R-square	0.1508			0.1538		

Appendix Table 5. The estimation results of the saturated specification of regression model via the application of the conventional and nonlinear approach to the Taiwan's MUS data for females.

	Conventional Approach			Nonlinear Approach		
Explanatory Variable	Coefficient	T-statistic	% Effect	Coefficient	T-statistic	% Effect
Intercept	10.86940	1116.6		10.95812	1455.3	
A. Age Factor (Ref: A4549)						
A2529	-0.45279	-41.7	-36	-0.49539	-51.2	-39
A3034	-0.29583	-25.7	-26	-0.32439	-32.5	-28
A3539	-0.14114	-11.6	-13	-0.16522	-16.4	-15
A4044	-0.07679	-5.9	-7	-0.09468	-9.1	-9
A5054	0.07373	4.6	8	0.07307	6.2	8
A5559	0.01275	0.6	1	0.02876	1.6	3
A6064	0.01016	0.2	1	0.03138	1.0	3
B. Education Factor (Ref: University grad.)						
Less than High School	-0.79457	-71.4	-55	-0.80009	-66.9	-55
High School Graduate	-0.54747	-48.6	-42	-0.54169	-50.9	-42
College Graduate	-0.22757	-16.6	-20	-0.23481	-19.4	-21
Post-graduate Degree	0.24729	10.0	28	0.23560	14.9	27
C. Interaction Terms						
A2529_E1	0.42517	24.9		0.43744	17.0	
A2529_E2	0.29292	22.3		0.27705	19.1	
A2529_E3	0.11075	7.2		0.10728	7.0	
A2529_E5	-0.01770	-0.6		-0.00430	-0.2	
A3034_E1	0.27299	17.1		0.27865	12.6	
A3034_E2	0.18458	13.5		0.17298	12.0	
A3034_E3	0.04318	2.7		0.04452	2.9	
A3034_E5	-0.03337	-1.2		-0.03253	-1.6	
A3539_E1	0.10474	6.9		0.11558	6.1	
A3539_E2	0.06412	4.5		0.06525	4.6	
A3539_E3	-0.01272	-0.8		-0.00837	-0.5	
A3539_E5	-0.01477	-0.5		-0.01772	-0.9	
A4044_E1	0.05303	3.5		0.06450	3.7	
A4044_E2	0.03681	2.5		0.04986	3.4	
A4044_E3	-0.00671	-0.4		0.00442	0.3	
A4044_E5	0.02209	0.7		0.03007	1.4	
A5054_E1	-0.09934	-5.6		-0.06811	-3.8	
A5054_E2	-0.02876	-1.5		-0.02503	-1.5	
A5054_E3	-0.07064	-3.1		-0.05598	-2.9	
A5054_E5	-0.10798	-2.5		-0.11779	-4.3	
A5559_E1	-0.09359	-3.7		-0.06209	-2.5	
A5559_E2	0.08921	3.2		0.12338	5.0	
A5559_E3	-0.00265	-0.1		0.02090	0.7	
A5559_E5	-0.02805	-0.4		-0.08688	-2.1	
A6064_E1	-0.22373	-5.0		-0.14611	-3.7	
A6064_E2	0.14865	2.9		0.15787	3.5	
A6064_E3	0.26810	3.8		0.28208	5.4	
A6064_E5	0.33898	2.7		0.25362	3.9	
Adj. R-square	0.2357			0.2411		

Appendix Table 6.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Taiwan's males at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	31,403	29,711	31,403	-1,692	0	-5.4	0.0
30-34	34,733	32,437	34,733	-2,296	0	-6.6	0.0
35-39	37,338	34,347	37,338	-2,991	0	-8.0	0.0
40-44	38,052	34,473	38,052	-3,579	0	-9.4	0.0
45-49	39,429	34,913	39,429	-4,516	0	-11.5	0.0
50-54	38,082	33,087	38,082	-4,995	0	-13.1	0.0
55-59	34,540	28,791	34,540	-5,749	0	-16.6	0.0
60-64	28,741	22,087	28,741	-6,654	0	-23.2	0.0
All	36,381	32,137	36,381	-4,244	0	-11.7	0.0
	High School Diploma						
25-29	32,606	31,181	32,606	-1,425	0	-4.4	0.0
30-34	37,900	35,491	37,900	-2,409	0	-6.4	0.0
35-39	41,745	38,854	41,745	-2,891	0	-6.9	0.0
40-44	44,627	40,712	44,627	-3,915	0	-8.8	0.0
45-49	47,112	42,172	47,112	-4,940	0	-10.5	0.0
50-54	48,058	42,651	48,058	-5,407	0	-11.3	0.0
55-59	49,082	41,635	49,082	-7,447	0	-15.2	0.0
60-64	49,665	37,669	49,665	-11,996	0	-24.2	0.0
All	41,897	38,280	41,897	-3,617	0	-8.6	0.0

Appendix Table 6.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Taiwan's males with college degree and Bachelor's degree.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	College Degree						
25-29	35,597	34,190	35,597	-1,407	0	-4.0	0.0
30-34	42,336	40,047	42,336	-2,289	0	-5.4	0.0
35-39	48,983	45,623	48,983	-3,360	0	-6.9	0.0
40-44	54,262	49,444	54,262	-4,818	0	-8.9	0.0
45-49	55,676	51,272	55,676	-4,404	0	-7.9	0.0
50-54	62,103	53,709	62,103	-8,394	0	-13.5	0.0
55-59	58,601	50,502	58,601	-8,099	0	-13.8	0.0
60-64	64,012	56,377	64,012	-7,635	0	-11.9	0.0
All	48,629	44,795	48,629	-3,834	0	-7.9	0.0
	University Degree						
25-29	37,593	35,933	37,593	-1,660	0	-4.4	0.0
30-34	48,405	45,150	48,405	-3,255	0	-6.7	0.0
35-39	57,849	52,732	57,849	-5,117	0	-8.8	0.0
40-44	68,036	61,537	68,036	-6,499	0	-9.6	0.0
45-49	75,158	65,441	75,158	-9,717	0	-12.9	0.0
50-54	76,359	66,395	76,359	-9,964	0	-13.0	0.0
55-59	81,341	68,507	81,341	-12,834	0	-15.8	0.0
60-64	81,998	71,253	81,998	-10,745	0	-13.1	0.0
All	57,783	52,202	57,783	-5,581	0	-9.7	0.0

Appendix Table 6.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Taiwan's males with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	44,205	42,387	44,205	-1,818	0	-4.1	0.0
30-34	58,289	54,201	58,289	-4,088	0	-7.0	0.0
35-39	72,829	67,114	72,829	-5,715	0	-7.8	0.0
40-44	82,282	74,411	82,282	-7,871	0	-9.6	0.0
45-49	91,773	82,817	91,773	-8,956	0	-9.8	0.0
50-54	90,593	83,781	90,593	-6,812	0	-7.5	0.0
55-59	88,486	81,896	88,486	-6,590	0	-7.4	0.0
60-64	113,646	96,960	113,646	-16,686	0	-14.7	0.0
All	71,849	66,008	71,849	-5,841	0	-8.1	0.0
	All Levels of Education						
25-29	34,838	33,309	34,838	-1,529	0	-4.4	0.0
30-34	41,621	38,994	41,621	-2,627	0	-6.3	0.0
35-39	46,008	42,559	46,008	-3,449	0	-7.5	0.0
40-44	48,669	44,225	48,669	-4,444	0	-9.1	0.0
45-49	50,305	44,922	50,305	-5,383	0	-10.7	0.0
50-54	49,424	43,299	49,424	-6,125	0	-12.4	0.0
55-59	46,588	39,458	46,588	-7,130	0	-15.3	0.0
60-64	39,974	31,868	39,974	-8,106	0	-20.3	0.0
All	45,015	40,792	45,015	-4,223	0	-9.4	0.0

Appendix Table 7.1. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Taiwan's females at the two lowest levels of education.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Less Than High School						
25-29	24,345	23,091	24,345	-1,254	0	-5.2	0.0
30-34	24,644	23,202	24,644	-1,442	0	-5.9	0.0
35-39	24,548	22,890	24,548	-1,658	0	-6.8	0.0
40-44	25,031	23,181	25,031	-1,850	0	-7.4	0.0
45-49	25,798	23,738	25,798	-2,060	0	-8.0	0.0
50-54	25,926	23,138	25,926	-2,788	0	-10.8	0.0
55-59	24,952	21,895	24,952	-3,057	0	-12.3	0.0
60-64	23,001	19,173	23,001	-3,828	0	-16.6	0.0
All	25,139	22,917	25,139	-2,222	0	-8.8	0.0
	High School Diploma						
25-29	26,852	25,902	26,852	-950	0	-3.5	0.0
30-34	28,711	27,192	28,711	-1,519	0	-5.3	0.0
35-39	30,226	28,139	30,226	-2,087	0	-6.9	0.0
40-44	31,940	29,201	31,940	-2,739	0	-8.6	0.0
45-49	33,404	30,392	33,404	-3,012	0	-9.0	0.0
50-54	35,048	31,790	35,048	-3,258	0	-9.3	0.0
55-59	38,893	33,654	38,893	-5,239	0	-13.5	0.0
60-64	40,364	35,623	40,364	-4,741	0	-11.7	0.0
All	30,609	28,440	30,609	-2,169	0	-7.1	0.0

Appendix Table 7.2. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Taiwan's females with college degree and Bachelor's degree.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	College Degree						
25-29	30,798	29,726	30,798	-1072	0	-3.5	0.0
30-34	34,319	32,506	34,319	-1813	0	-5.3	0.0
35-39	38,167	35,882	38,167	-2285	0	-6.0	0.0
40-44	41,483	38,497	41,483	-2986	0	-7.2	0.0
45-49	45,402	41,850	45,402	-3552	0	-7.8	0.0
50-54	46,185	41,979	46,185	-4206	0	-9.1	0.0
55-59	47,714	42,274	47,714	-5440	0	-11.4	0.0
60-64	62,117	55,276	62,117	-6841	0	-11.0	0.0
All	36,638	34,485	36,638	-2153	0	-5.9	0.0
	University Degree						
25-29	34,987	33,410	34,987	-1577	0	-4.5	0.0
30-34	41,512	39,088	41,512	-2424	0	-5.8	0.0
35-39	48,674	45,628	48,674	-3046	0	-6.3	0.0
40-44	52,231	48,660	52,231	-3571	0	-6.8	0.0
45-49	57,419	52,544	57,419	-4875	0	-8.5	0.0
50-54	61,771	56,564	61,771	-5207	0	-8.4	0.0
55-59	59,094	53,218	59,094	-5876	0	-9.9	0.0
60-64	59,249	53,080	59,249	-6169	0	-10.4	0.0
All	44,291	41,512	44,291	-2779	0	-6.3	0.0

Appendix Table 7.3. Comparison of the mean wages predicted by the conventional and nonlinear approaches against the correspoding observed mean wages, based on the saturated specification: Taiwan's females with post-graduate degrees.

AGE	Mean Wage ($ / week)			Difference from Observed ($)		Difference from Observed (%)	
	Observed	Conventional	Nonlinear	Conventional	Nonlinear	Conventional	Nonlinear
	Post-graduate Degrees						
25-29	44,092	42,033	44,092	-2,059	0	-4.7	0.0
30-34	50,859	48,412	50,859	-2,447	0	-4.8	0.0
35-39	60,523	57,572	60,523	-2,951	0	-4.9	0.0
40-44	68,125	63,704	68,125	-4,421	0	-6.5	0.0
45-49	72,673	67,285	72,673	-5,388	0	-7.4	0.0
50-54	69,495	65,020	69,495	-4,475	0	-6.4	0.0
55-59	68,569	66,264	68,569	-2,305	0	-3.4	0.0
60-64	96,637	95,400	96,637	-1,237	0	-1.3	0.0
All	56,235	53,210	56,235	-3,025	0	-5.4	0.0
	All Levels of Education						
25-29	31,105	29,863	31,105	-1,242	0	-4.0	0.0
30-34	33,739	31,906	33,739	-1,833	0	-5.4	0.0
35-39	35,210	32,960	35,210	-2,250	0	-6.4	0.0
40-44	35,259	32,568	35,259	-2,691	0	-7.6	0.0
45-49	35,608	32,627	35,608	-2,981	0	-8.4	0.0
50-54	35,157	31,802	35,157	-3,355	0	-9.5	0.0
55-59	33,242	29,299	33,242	-3,943	0	-11.9	0.0
60-64	28,940	24,790	28,940	-4,150	0	-14.3	0.0
All	33,960	31,640	33,960	-2,320	0	-6.8	0.0

To account for the severity of the under-prediction by the conventional approach in the saturated specification, we make a weighted linear regression of the severity of the under-prediction on the *coefficient of variation*, the *skewness*, and the *kurtosis* of the wage distribution within each of the 40 education-by-age cells, based on the MUS data. The results are shown in the male and female panels of Appendix Table 8. Here we find again that the severity of the under-prediction depends *very strongly* on coefficient of variation, *moderately* on skewness, and *modestly* on kurtosis. Actually, for females, kurtosis turns out to have practically no explanatory power. With respect to the directions of the effects, we find again that the severity of the under-prediction tends to *increase* with coefficient of variation, to *decrease* with skewness, and to *increase* with kurtosis.

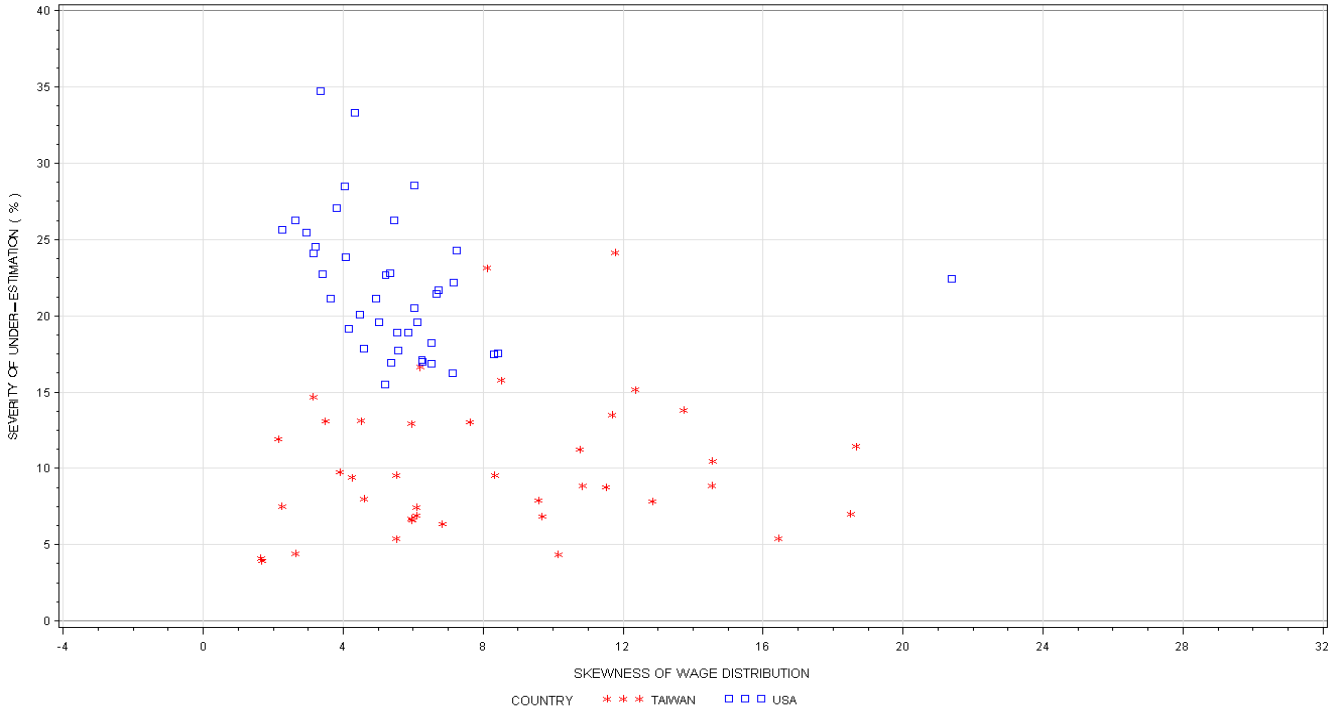
Appendix Table 8. The results of regressing the severity of the under-estimation by the conventional approach on the coefficient of variation, skewness, and kurtosis of the wage distribution of workers in Taiwan.

Explanatory Variable	Specification 1		Specification 2		Specification 3	
	Coefficient	T-Satistic	Coefficient	T-Satistic	Coefficient	T-Satistic
	Male Panel					
INTERCEPT	-1.460	-1.1	-0.937	-0.8	0.572	0.5
CV	0.189	8.7	0.227	10.5	0.255	11.4
SKEWNESS			-0.300	-3.6	-1.109	-3.7
KURTOSIS					0.019	2.8
Adj. R-Square	0.65		0.74		0.78	
Additional Contribution to Adj. R-Square			0.08		0.04	
	Female Panel					
INTERCEPT	-0.030	-0.1	-1.929	-3.2	-1.726	-2.7
CV	0.151	11.3	0.221	13.2	0.223	13.3
SKEWNESS			-0.185	-5.3	-0.295	-2.9
KURTOSIS					0.003	1.2
Adj. R-Square	0.76		0.86		0.86	
Additional Contribution to Adj. R-Square			0.10		0.00	

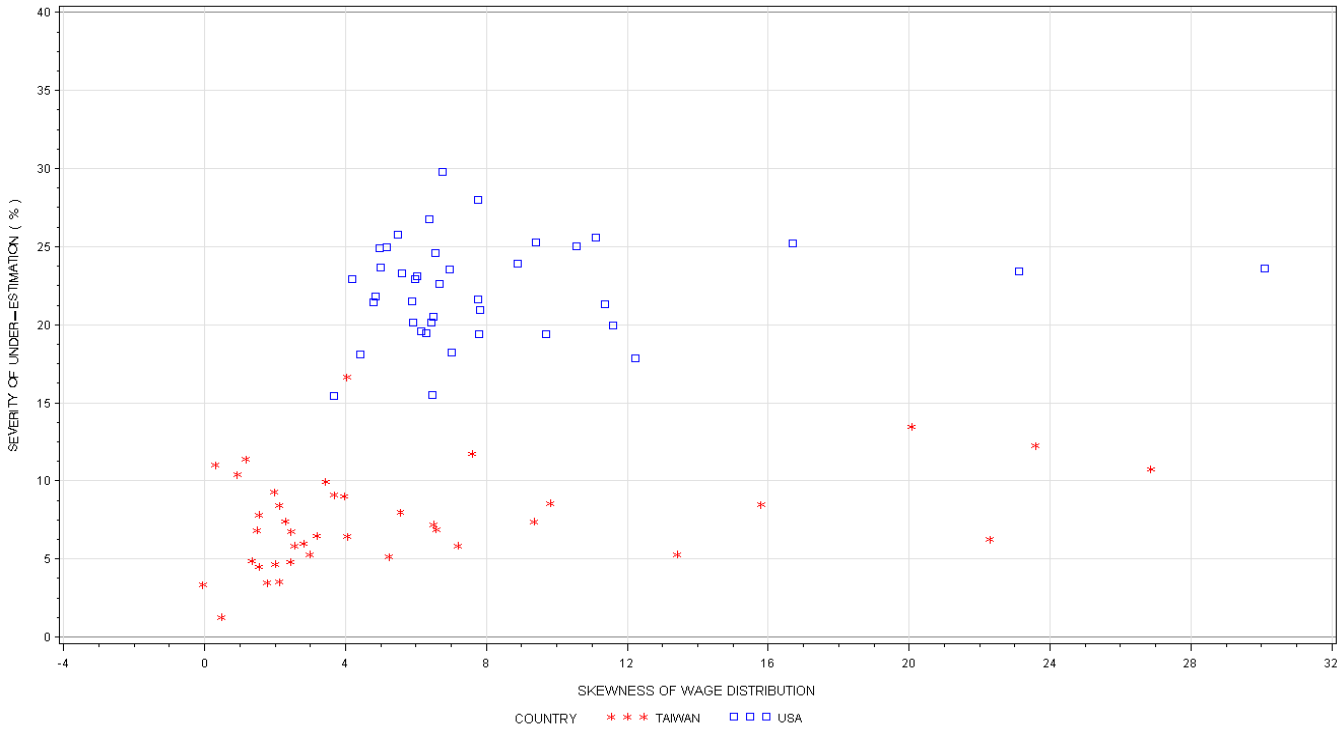
Note: See fottnote in Table 8.

In Appendix Figures 5 to 6, we plot, for each gender, the severity of the under-prediction by the conventional approach against the corresponding skewness of income distribution. In Appendix Figures 7 to 8, we plot, for each gender, the severity of the under-prediction by the conventional approach against the corresponding kurtosis of income distribution. From these figures, we see that for both the ACS and PMS data, there is absence of any strong relationships. The rather weak effects of these two attributes of wage distribution on the severity of the conventional approach's under-prediction can only be revealed by a multiple regression model.

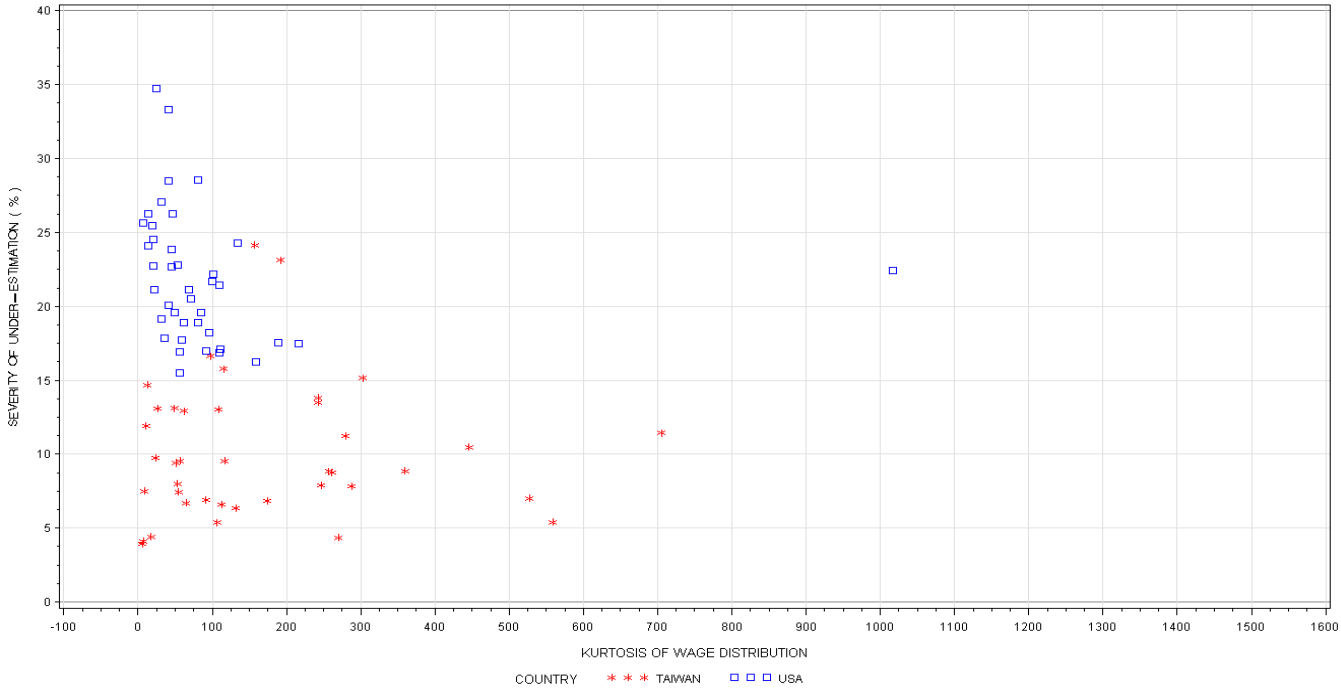
Appendix Figure 5. The severity of the under-prediction of the observed mean wage by the conventional approach versus (2) the **skewness** of the wage distribution, based on the **male** data of the 2005-2007 ACS and Taiwan's 2001-2010 MUS. Each point represents a combination of an education level and an age group.



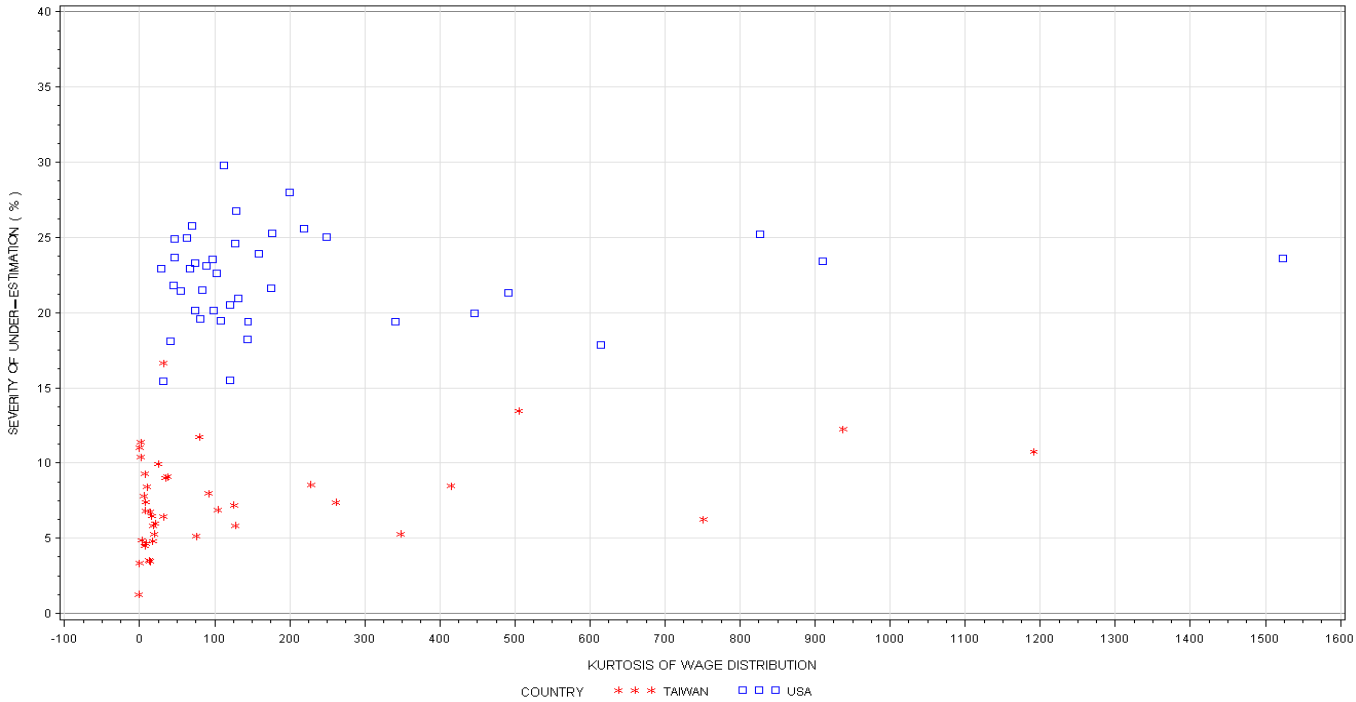
Appendix Figure 6. The severity of the under-prediction of the observed mean wage by the conventional approach versus (2) the **skewness** of the wage distribution, based on the **female** data of the 2005-2007 ACS and Taiwan's 2001-2010 MUS. Each point represents a combination of an education level and an age group.



Appendix Figure 7. The severity of the under-prediction of the observed mean wage by the conventional approach versus (2) the **kurtosis** of the wage distribution, based on the **male** data of the 2005-2007 ACS and Taiwan's 2001-2010 MUS. Each point represents a combination of an education level and an age group.



Appendix Figure 8. The severity of the under-prediction of the observed mean wage by the conventional approach versus (2) the **kurtosis** of the wage distribution, based on the **female** data of the 2005-2007 ACS and Taiwan's 2001-2010 MUS. Each point represents a combination of an education level and an age group.



We conclude by stating (1) that both in the US and in Taiwan, the severity of the under-prediction by the conventional approach is strongly dependent on the variability in the wage distribution, and (2) that the under-prediction problem of the conventional approach is much more serious in the US than in Taiwan, mainly because wage variability is much greater in the US than in Taiwan.

Appendix B. The SAS Module and a SAS Program for Carry Out Weighted Nonlinear Least-square Estimation According to the Nonlinear Approach

The following SAS text is the module for carrying out the iterative procedure for the weighted nonlinear least-squares estimation of the unknown coefficients in an exponential regression model. Note that the colors in the text of the module is determined by the program editor of SAS 9.2, which uses the red color to indicate syntactical mistakes. Because this program editor is not yet fully developed for SAS/IML, it incorrectly assigns red color to some valid words.

```
* NAME OF THE NONLINEAR ESTIMATION MODULE: NNL_REG_ROBUST_MODULE.SAS;
START NNL_REG ;
/*****************************************************************/
/* THE FOLLOWING MODULE IS FOR ESTIMATING THE PARAMETERS OF AN      */
/* EXPONENTIAL MODEL, USING WEIGHTED NONLIEAR LEAST-SQUARES METHOD. */
/* WRITTEN BY KAO-LEE LIAW IN 2013.                                 */
/*                                                                  */
/* INPUT VARIABLES:                                                 */
/* N_ITER=NO. OF DESIRED ITERATIONS. (SCALAR).                      */
/* DETAIL= 1 (IF YOU WANT TO SEE THE INFORMATION AT EACH ITERATION).*/
/* BL_SIZE= NO. OF OBSERVATIONS PER BLOCK (OR BUNCH).               */
/* (BL_SIZE SPECIFIES THE NUMBER OF OBSERVATIONS TO BE USED AT      */
/*  EACH STEP IN THE CONSTRUCTION OF THE INFORMATION MATTIX.        */
/*  TRY 500. IF THE MEMORY IS TOO SMALL, REDUCE IT TO A SMALLER     */
/*  NUMBER.)                                                        */
/* STEPSIZE= A SCALAR TO ADJUST THE SIZE OF THE CHANGE IN THE       */
/*       PARAMETER VECTOR FROM ONE ITERATION TO THE NEXT. IT MUST   */
/*       BE A POSITIVE VALUE LESS THAN OR EQUAL TO 1.               */
/*       WHEN IT IS SET TO 1, THE COMPUTATION TAKES THE LEAST       */
/*       AMOUNT OF TIME BUT THE RISK OF DIVERGENCE IS THE HIGHEST.  */
/*       WHEN CONVERGENCE FAILED, TRY USING A SMALLER STEP SIZE.    */
/* MI= COLUMN VECTOR OF THE DEPENDENT VARIABLE.                     */
/* POPRISK= COLUMN VECTOR OF AT-RISK POPULATION (WEIGHT VARIABLE).  */
/* INDEP= COLUMN VECTOR WITH NAMES OF EXPLANATORY VARIABLES,        */
/*        WITH THE FIRST VARIABLE BEING THE CONSTANT TERM.          */
/* DETAIL= 1 (IF YOU WANT TO SEE THE DETAILED INFORMATION AT EACH   */
/*            ITERATION).                                           */
/*         0 (IF YOU DON'T WANT TO SEE THE DETAILED INFORMATION     */
/*            AT EACH ITERATION).                                   */
/*  Technical Advice:                                               */
/*   IF YOU GET AN "OVERFLOW" ERROR MESSAGE, YOU SHOULD CHANGE THE  */
/*   SCALE OF YOUR DEPENDENT VARIABLE                               */
/*    BY DIVIDING A LARGE NUMBER (E.G. 1000) INTO IT.               */
/*****************************************************************/
```

```
   read all var{&DEP_VAR} into MI;
   READ ALL VAR{&POP_RISK} INTO POPRISK;
   read all var{&indep} into X;
   NOBS=NROW(MI);
   WEIGHT_TOT=SUM(POPRISK);
   WT=POPRISK/(WEIGHT_TOT/NOBS);/* THIS WEIGHT VARIABLE IS SCALED SO THAT ITS
SUM = THE SAMPLE SIZE. */
   LEFTOVER=MOD(NOBS,&BL_SIZE);
   NBUNCH=INT(NOBS/&BL_SIZE);
   Y_MEAN =SUM(MI#WT)/SUM(WT);

   B = REPEAT(0,NCOL(X),1); OLDB=B+1; /* STARTING VALUES */

* BEGINNING OF ITERATIONS;
DO ITER=1 TO &N_ITER;
      IF MAX(ABS(B-OLDB))<1E-8 THEN GOTO LAB1;
      OLDB=B;
      XPX=REPEAT(0,NCOL(X),NCOL(X));
      XPY=REPEAT(0,NCOL(X),1);

     XPX_RB=XPX;

      WSSQ=0;   WSSQ0=0;
      RSRMSQ=0;
      DO IBUNCH=1 TO NBUNCH;
        N1=(IBUNCH-1)*&BL_SIZE+1;
        N2=N1+&BL_SIZE-1;
 P_NULL=REPEAT(Y_MEAN, &BL_SIZE);
       PI=EXP(X[N1:N2, ]*B);
      DER=X[N1:N2, ]#PI;
      * DER is part of the Jacobian;
     DIFF_I=MI[N1:N2] - PI;
       DIFF_0=MI[N1:N2] - P_NULL;
      WSSQ = WSSQ + SUM(WT[N1:N2] # DIFF_I # DIFF_I);
   WSSQ0 = WSSQ0 + SUM(WT[N1:N2] # DIFF_0 # DIFF_0);
     DERTW=(DER#WT[N1:N2])`;

        DERTW_RB=(DER#WT[N1:N2]#DIFF_I#DIFF_I)`;
            XPX_RB= XPX_RB + DERTW_RB*DER;

        XPX= XPX + DERTW*DER; /* BUILDING UP THE INFORMATION MATRIX */
        XPY= XPY + DERTW*DIFF_I;
      END;
    IF LEFTOVER > 0 THEN DO;
        N1=NBUNCH*&BL_SIZE+1;
        N2=N1+LEFTOVER-1;
 P_NULL=REPEAT(Y_MEAN, LEFTOVER);
       PI=EXP(X[N1:N2, ]*B);
      DER=X[N1:N2, ]#PI;
     DIFF_I=MI[N1:N2] - PI;
       DIFF_0=MI[N1:N2] - P_NULL;
      WSSQ = WSSQ + SUM(WT[N1:N2, ] # DIFF_I # DIFF_I);
WSSQ0 = WSSQ0 + SUM(WT[N1:N2, ] # DIFF_0 # DIFF_0);
        DERTW=(DER#WT[N1:N2])`;

        DERTW_RB=(DER#WT[N1:N2]#DIFF_I#DIFF_I)`;
        XPX_RB= XPX_RB + DERTW_RB*DER;
```

```
        XPX= XPX + DERTW*DER; /* BUILDING UP THE INFORMATION MATRIX */
        XPY= XPY + DERTW*DIFF_I;
                            END; /* At this point, the construction of the
Information Matrix is completed */
      btransp = b`;
      IF &DETAIL = 1 THEN print iter WSSQ btransp;
      XPX = INV(XPX);/* NOW XPX IS THE INVERSE OF INFORMATION MATRIX  */
      B = B + &STEPSIZE*( XPX * XPY);/* REDUCE THE STEP SIZE, IF THE MODULE
DOES NOT CONVERGE */
   END; /*  THE MAXIMUM NUMBER OF ITERATIONS IS REACHED HERE  */
   IF ITER >= &N_ITER THEN DO;
      PRINT "!!! WARNING!!!: THE ESTIMATED PARAMETERS MAY NOT BE
MEANINGFUL,";
      PRINT "BECAUSE THE MAXIMUM NO. OF ITERATIONS IS REACHED.";
                             END;
LAB1: DF_ESS    = NOBS - NCOL(X) ;
   NVAR=NCOL(X)-1;
   RSRMSQ=SQRT(WSSQ/DF_ESS);/* WEIGHTED RESIDUAL ROOT_MEAN_SQUARE */
   RSRMSQ0=SQRT(WSSQ0/(NOBS-1));/* WEIGHTED RESIDUAL ROOT_MEAN_SQUARE OF THE
NULL MODEL */
   R_SQUARE = (WSSQ0 - WSSQ) / WSSQ0;
   ADJ_R_SQ= 1 - (RSRMSQ/RSRMSQ0)**2;
   PRINT  NOBS NVAR WEIGHT_TOT Y_MEAN ITER ;
   PRINT  WSSQ WSSQ0 RSRMSQ DF_ESS  RSRMSQ0 R_SQUARE ADJ_R_SQ;

   CV_RB=XPX * XPX_RB * XPX;
   STDERR_RB = SQRT(VECDIAG(CV_RB));
   TRATIO_RB = B/STDERR_RB;

   STDERR = SQRT(VECDIAG(XPX))*RSRMSQ;
   TRATIO = B/STDERR;
   EXP_MS_1 = EXP(B) - 1;
   V_NAME={&INDEP}`;
   PRINT "ESTIMATED COEFFICIENTS: B,   STDERR_RB: ROBUST STANDARD ERROR,
TRATIO_RB: ROBUST T-RATIO.";
   PRINT V_NAME B[FORMAT=13.6] TRATIO[FORMAT=10.2]  STDERR[FORMAT=13.6]
         EXP_MS_1[FORMAT=13.6]  TRATIO_RB[FORMAT=10.2]
STDERR_RB[FORMAT=13.6] ;
/* CREATE THE DATA SET CONTAINING THE ESTIMATED PARAMETERS AND RELATED
STATISTICS */
   CREATE SD1.&PARM_FL VAR{V_NAME B  STDERR TRATIO  STDERR_RB TRATIO_RB };
   APPEND;
   CLOSE SD1.&PARM_FL;
/* CREATE THE DATA SET CONTAINING THE PREDICTED VALUES OF THE DEPENDENT
VARIBLE */
         CREATE PI VAR{PI};
DO IBUNCH=1 TO NBUNCH;
        N1=(IBUNCH-1)*&BL_SIZE+1;
        N2=N1+&BL_SIZE-1;
         PI=EXP(X[N1:N2, ]*B);
             APPEND;
 END;
 IF LEFTOVER > 0 THEN DO;
        N1=NBUNCH*&BL_SIZE+1;
        N2=N1+LEFTOVER-1;
        PI=EXP(X[N1:N2, ]*B);
```

```
            APPEND;
                                        END;
         CLOSE PI;
FINISH /* NNL_REG_ROBUST */ ;
/********************** END OF THE MODULE ***********************/

/*************************************************************************/
/*  OUTPUT VARIABLES:                                                    */
/*          NOBS=NO. OF OBSERVATIONS.                                    */
/*          NVAR=NO. OF SUBSTATIVE EXPLANATORY VARIABLES.                */
/*         WEIGHT_TOT=THE SUM OF ORIGINAL WEIGHT VARIABLE.               */
/*         Y_MEAN=THE WEIGHTED MEAN OF THE DEPENDENT VARIABLE.           */
/*         ITER=THE NUNBER OF ITERATIONS AT CONVERGENCE.                 */
/*         WSSQ=THE WEIGHTED RESIDUAL SUM OF SQUARES.                    */
/*         WSSQ0=  THE WEIGHTED TOTAL SUM OF SQUARES.                    */
/*         RSRMSQ= WEIGHTED RESIDUAL ROOT_MEAN_SQUARE.                   */
/*         RSRMSQ0=  WEIGHTED RESIDUAL ROOT_MEAN_SQUARE OF THE NULL MODEL. */
/*         R_SQUARE.                                                     */
/*         ADJ_R_SQ= ADJUSTED R_SQUARE.                                  */
/*                                                                       */
/* OUTPUT DATA SETS:                                                     */
/*     (1)    SD1.&PARM_FL:                                              */
/*       V_NAME=A COLUMN VETCTOR CONTAINING THE VARIABLE NAMES.          */
/*       B = A COLUMN VECTOR CONTAINING THE ESTIMATED PARAMETERS.        */
/*       STDERR = A COLUMN VECTOR CONTAINING THE STANDARD ERRORS OF THE  */
/*                 PARAMETER ESTIMATOR.                                  */
/*       TRATIO = T-RATIO.                                               */
/*       STDERR_RB = A COLUMN VECTOR CONTAINING THE ROBUST STANDARD ERRORS */
/*                   OF THE PARAMETER ESTIMATORS.                        */
/*       TRATIO_RB = ROBUST T-RATIO.                                     */
/*      (2)  PI: THIS DATA SET CONTAINS THE PREDICTED AVLUES OF THE      */
/*               DEPENDENT VARIABLE.                                     */
/*                                                                       */
/*  NOTE: IN THE PROGRAM THAT USES THIS MODULE, PLEASE REMEMBER TO USE A */
/*        "LIBNAME" STATEMENT TO ASSOCIATE THE NAME "SD1" TO A FOLDER    */
/*        IN YOUR HARDDISK FOR STORING THE PARAMETER FILE.               */
/*************************************************************************/
```

The following is a SAS program that uses the above module to estimate the coefficients of the *full specification* of the regression model via the nonlinear approach, based on Taiwan's MUS data.

```
*MPTW_WAGE_2564_NNL_REG_ROBUST_QDT_FULL.SAS ;
*THIS PROGRAM APPLIES THE NONLINEAR LEAST-SQUARES ESTIMATION METHOD TO
* AN EXPONENTIAL MODEL OF THE WAGES OF WAGE EARNERS, AGED 25-64;
* DATA SOURCE: THE MICRO DATA OF THE 2001-2010 MANPOWER SURVEY OF TAIWAN;

ods html
body='C:\D\TWN_2011\TWN_2014\MPTW_WAGE_2564_NNL_REG_ROBUST_QDT_FULL.html';
OPTIONS LS=130 PS=10000 NOCENTER;
LIBNAME SD1 "C:\D\TWN_2011\TWN_2014\";

PROC FORMAT;
VALUE SEX 1="MALE" 2="FEMALE";
value AGE_X
1="00-04"
2="05-09"
3="10-14"
4="15-19"
5="20-24"
6="25-29"
7="30-34"
8="35-39"
9="40-44"
10="45-49"
11="50-54"
12="55-59"
13="60-64"
14="65-69"
15="70+";
value EDUC_A
 1="LT HS"
 2="HS Grad"
 3="College Drad"
 4="University Grad"  5="Master+";

DATA MPTW_M MPTW_F  ;
TITLE "MPTW2001_2010, AGED 25-64.";
SET SD1.POOLEDMP_1979_2010 ;

IF(AGE GE 25 AND AGE LE 64) AND YEAR GE 2001 AND INCOME GT 1 AND WORKHOUR GE
20;

* EXPRESS INCOME OF ALL YEARS IN TERMS OF THE 2013 $;
IF YEAR=2001  THEN  WAGE_P_MNTH = INCOME * 1.1444223;
IF YEAR=2002  THEN  WAGE_P_MNTH = INCOME * 1.1464143;
IF YEAR=2003  THEN  WAGE_P_MNTH = INCOME * 1.1494024;
IF YEAR=2004  THEN  WAGE_P_MNTH = INCOME * 1.1314741;
IF YEAR=2005  THEN  WAGE_P_MNTH = INCOME * 1.1065737;
IF YEAR=2006  THEN  WAGE_P_MNTH = INCOME * 1.0996016;
IF YEAR=2007  THEN  WAGE_P_MNTH = INCOME * 1.0796813;
IF YEAR=2008  THEN  WAGE_P_MNTH = INCOME * 1.0438247;
IF YEAR=2009  THEN  WAGE_P_MNTH = INCOME * 1.0527888;
```

```
IF YEAR=2010  THEN  WAGE_P_MNTH = INCOME * 1.0428287;
IF YEAR=2011  THEN  WAGE_P_MNTH = INCOME * 1.0278884;
IF YEAR=2012  THEN  WAGE_P_MNTH = INCOME * 1.0079681;

IF EDU LE 4 THEN EDUC_A=1;
ELSE IF EDU LE 5 OR EDU LE 6 THEN EDUC_A=2;
ELSE IF EDU LE 7 THEN EDUC_A=3;
ELSE IF EDU LE 8 THEN EDUC_A=4;
ELSE  EDUC_A=5;

ED_PR=0; ED_2ND=0;  ED_SC=0;  ED_MS=0;
IF EDUC_A=1 THEN ED_PR=1;
ELSE IF EDUC_A=2 THEN ED_2ND=1;
ELSE IF EDUC_A=3 THEN ED_SC=1;
ELSE IF EDUC_A=5 THEN ED_MS=1;

A2529=0;   A3034=0;  A3539=0;   A4044=0;   A4549=0;   A5054=0;   A5559=0;
A6064=0;
IF AGE GE 25 AND AGE LT 30 THEN DO; A2529=1; AGE_X=6; END;
ELSE IF AGE GE 30 AND AGE LT 35 THEN DO; A3034=1; AGE_X=7; END;
ELSE IF AGE GE 35 AND AGE LT 40 THEN  DO; A3539=1; AGE_X=8; END;
ELSE IF AGE GE 40 AND AGE LT 45 THEN  DO; A4044=1; AGE_X=9; END;
ELSE IF AGE GE 45 AND AGE LT 50 THEN  DO; A4549=1; AGE_X=10; END;
ELSE IF AGE GE 50 AND AGE LT 55 THEN  DO; A5054=1; AGE_X=11; END;
ELSE IF AGE GE 55 AND AGE LT 60 THEN  DO; A5559=1; AGE_X=12; END;
ELSE IF AGE GE 60 AND AGE LT 65 THEN  DO; A6064=1; AGE_X=13; END;

AGE_R45= AGE -45;
AGESQ_R45 = AGE_R45 * AGE_R45;

AGE_R45_E1 =AGE_R45 * ED_PR;
AGE_R45_E2= AGE_R45 * ED_2ND;
AGE_R45_E3= AGE_R45 * ED_SC;
AGE_R45_E5 =AGE_R45 * ED_MS;

AGESQ_R45_E1= AGESQ_R45 * ED_PR;
AGESQ_R45_E2= AGESQ_R45 * ED_2ND;
AGESQ_R45_E3= AGESQ_R45 * ED_SC;
AGESQ_R45_E5= AGESQ_R45 * ED_MS;

CONST =1;

IF SEX=1 THEN OUTPUT MPTW_M;
ELSE OUTPUT MPTW_F;

KEEP  WAGE_P_MNTH    SEX AGE AGE_X EDUC_A
        CONST    AGE_R45 AGESQ_R45  ED_PR  ED_2ND  ED_SC  ED_MS
        AGE_R45_E1 AGE_R45_E2 AGE_R45_E3 AGE_R45_E5
        AGESQ_R45_E1 AGESQ_R45_E2 AGESQ_R45_E3 AGESQ_R45_E5
         WEIGHT  ;

RUN;

PROC MEANS DATA=MPTW_M  N MEAN MEDIAN STD MIN MAX;
TITLE "MPTW2001_2010:  MALES, AGED 25-64.";
TITLE2 "RESTRICTED TO THOSE WITH WORKHOUR GE 20.";
TITLE3 "WEIGHT:";
```

```
TITLE4 "NOTE!!!: STD generated from weighted PROC MEANS must be divided by
the square root of the average weight.";
VAR  WAGE_P_MNTH   SEX AGE AGE_X EDUC_A   ED_PR  ED_2ND  ED_SC  ED_MS ;
WEIGHT WEIGHT ;
PROC MEANS DATA=MPTW_F  N MEAN MEDIAN STD MIN MAX;
TITLE "MPTW2001_2010:  FEMALES, AGED 25-64.";
TITLE2 "RESTRICTED TO THOSE WITH WORKHOUR GE 20.";
TITLE3 "UN-WEIGHT:";
TITLE4 "NOTE!!!: STD generated from weighted PROC MEANS must be divided by
the square root of the average weight.";
VAR  WAGE_P_MNTH    SEX AGE AGE_X EDUC_A   ED_PR  ED_2ND  ED_SC  ED_MS ;
WEIGHT WEIGHT ;

PROC MEANS DATA=MPTW_M  N MEAN MEDIAN STD CV  MIN MAX;
TITLE "MPTW2001_2010:  MALES, AGED 25-64.";
TITLE2 "RESTRICTED TO THOSE WITH WORKHOUR GE 20.";
TITLE3 "UN-WEIGHT:";
VAR  WAGE_P_MNTH   SEX AGE AGE_X EDUC_A   ED_PR  ED_2ND  ED_SC  ED_MS WEIGHT;
PROC MEANS DATA=MPTW_F  N MEAN MEDIAN STD CV  MIN MAX;
TITLE "MPTW2001_2010:  FEMALES, AGED 25-64.";
TITLE2 "RESTRICTED TO THOSE WITH WORKHOUR GE 20.";
TITLE3 "UN-WEIGHT:";
VAR  WAGE_P_MNTH   SEX AGE AGE_X EDUC_A   ED_PR  ED_2ND  ED_SC  ED_MS WEIGHT;

%LET BL_SIZE=50; /* BLOCK SIZE FOR CONSERVING MEMORY SPACE */
* IF THE SAMPLE SIZE IS SO LARGE THAT THERE IS NOT ENOUGH MEMORY SPACE,
PLEASE REDUCE THE BLOCK SIZE;
%LET STEPSIZE=0.5; /* IF CONVERGENCE FAILED, REDUCE THE STEPSIZE. */
%LET N_ITER=200;/* NO. OF ITERATIONS */
* IF DETAILS OF ITERATIONS ARE TO BE PRINTED, SET "DETAIL" TO "1";
%LET DETAIL=0;
* SPECIFY THE NAME OF THE DEPENDENT VARIABLE;
%LET DEP_VAR= WAGE_P_MNTH;
* SPECIFY THE NAME THE VARIABLE REPRESENTING THE SIZE OF THE AT-RISK
POPULATION (I.E. THE WEIGHT VARIABLE);
%LET POP_RISK= WEIGHT;
* DON'T FORGET THE CONSTANT TERM;
%LET INDEP=CONST  AGE_R45 AGESQ_R45   ED_PR  ED_2ND  ED_SC  ED_MS
                    AGE_R45_E1 AGE_R45_E2 AGE_R45_E3 AGE_R45_E5
       AGESQ_R45_E1 AGESQ_R45_E2 AGESQ_R45_E3 AGESQ_R45_E5;
*************************** MALE ***********************;
%LET RUN_NO= 1M;
%LET PARM_FL=NNL_PAR_&RUN_NO;

DATA INF&RUN_NO;
  SET MPTW_M  (KEEP= WAGE_P_MNTH &POP_RISK &INDEP);
WAGE_P_MNTH= WAGE_P_MNTH /1000;
* HERE WE CHANGE THE UNIT OF WAGE TO $1000/MONTH IN ORDER TO AVOID OVERFLOW
PROBLEM;
PROC IML ;
use INF&RUN_NO;
TITLE "MPTW2001_2010:  MALES, AGED 25-64.";
TITLE2 "RESTRICTED TO THOSE WITH WORKHOUR GE 20.";
 TITLE3 'DEPENDENT VAR = WAGE_P_MNTH (IN $1000 PER MONTH).   FULL QUADRATIC
SPECIFICATION.';
TITLE4 "#RUN &RUN_NO :    EXPONENTIAL MODEL.   WEIGHTED LEAST-SQUARES METHOD.
SUM OF SCALED WEIGHTS = SAMPLE SIZE.";
```

```
%INCLUDE "C:\D\TWN_2011\NNL_REG_ROBUST_MODULE.SAS";
RUN  NNL_REG;

DATA SD1.MPTW_M_NNL_QDT_FULL;
MERGE MPTW_M  (KEEP= WAGE_P_MNTH &POP_RISK EDUC_A AGE_X AGE )  PI;
P_WAGE_M = PI * 1000;
DIFF= WAGE_P_MNTH -P_WAGE_M;
PROC TABULATE DATA=SD1.MPTW_M_NNL_QDT_FULL;
TITLE5 "OBSERVED VERSUS PREDICTED WEEKLY WAGE:";
VAR WAGE_P_MNTH P_WAGE_M DIFF;
CLASS EDUC_A AGE_X;
TABLE (AGE_X ALL) , (EDUC_A ALL) *(WAGE_P_MNTH*MEAN*F=COMMA7.0
P_WAGE_M*MEAN*F=COMMA7.0   DIFF*MEAN*F=COMMA7.0);
  WEIGHT WEIGHT;
FORMAT EDUC_A EDUC_A. AGE_X AGE_X. ;
RUN;

************************** FEMALE ***********************;
%LET RUN_NO= 1F;
%LET PARM_FL=NNL_PAR_&RUN_NO;
DATA INF&RUN_NO;
  SET MPTW_F  (KEEP= WAGE_P_MNTH &POP_RISK &INDEP);
WAGE_P_MNTH= WAGE_P_MNTH /1000;
* HERE WE CHANGE THE UNIT OF WAGE TO $1000/MONTH IN ORDER TO AVOID OVERFLOW
PROBLEM;
PROC IML ;
use INF&RUN_NO;
TITLE "MPTW2001_2010:  FEMALES, AGED 25-64.";
TITLE2 "RESTRICTED TO THOSE WITH WORKHOUR GE 20.";
 TITLE3 'DEPENDENT VAR = WAGE_P_MNTH (IN $1000 PER MONTH).       FULL
QUADRATIC SPECIFICATION.';
TITLE4 "#RUN &RUN_NO :    EXPONENTIAL MODEL.   WEIGHTED LEAST-SQUARES METHOD.
SUM OF SCALED WEIGHTS = SAMPLE SIZE.";
%INCLUDE "C:\D\TWN_2011\NNL_REG_ROBUST_MODULE.SAS";
RUN  NNL_REG;

DATA SD1.MPTW_F_NNL_QDT_FULL;
MERGE MPTW_F  (KEEP= WAGE_P_MNTH &POP_RISK EDUC_A AGE_X  AGE)  PI;
P_WAGE_F = PI * 1000;
DIFF= WAGE_P_MNTH -P_WAGE_F;
PROC TABULATE DATA=SD1.MPTW_F_NNL_QDT_FULL;
TITLE5 "OBSERVED VERSUS PREDICTED WEEKLY WAGE:";
VAR WAGE_P_MNTH P_WAGE_F DIFF;
CLASS EDUC_A AGE_X;
TABLE (AGE_X ALL) , (EDUC_A ALL) *(WAGE_P_MNTH*MEAN*F=COMMA7.0
P_WAGE_F*MEAN*F=COMMA7.0   DIFF*MEAN*F=COMMA7.0);
  WEIGHT WEIGHT;
FORMAT EDUC_A EDUC_A. AGE_X AGE_X. ;
RUN;
```