ISSN 2616-5902 Vol. 7, Issue 2: 64-71, DOI: 10.25236/AJBM.2025.070209

Education and labor market outcome—Wage equation estimation with QLFS data

Shutong Huang*

University of Hong Kong, Hong Kong, China
shua3533@gmail.com
**Corresponding author*

Abstract: *This paper investigates the impact of education on individual wages, aiming to provide empirical evidence on the causal effect of schooling. The theoretical framework is based on Mincer's human capital model, which is extended to account for health status. The methods involve OLS and IV estimation using recent nationwide survey data. The key results indicate that each additional year of schooling has a positive and significant effect on wages. The magnitudes of OLS and IV estimates are similar, suggesting that ability bias may not be severe. The findings confirm human capital theory that education improves labor market outcomes and drives wage inequality. The conclusion is that expanded access to education, especially for disadvantaged groups, can help reduce income disparities. Implications are discussed for education policy. The paper contributes additional empirical evidence to the literature on the returns to schooling.*

Keywords: *Socioeconomic Inequality; Human Capital; Ability Bias; Labor Market Outcomes; Intervention Efficacy*

1. Introduction

A capstone model of wage equation is proposed in Mincer (1958; 1974)[13][14]. The model pays attention to the importance of human capital accumulation and distinguishes between two forms of human capital: education and working experience. The model is presented as:

$$lnw = lnw_0 + \rho s + \beta_1 x + \beta_2 x^2 \tag{1}$$

Where dependent variable is wage in log form, and independent variables are years of school education, experience, and quadratic term of experience. The model assumes a constant return to education and a diminishing return to working experience.

Later, the model is extended to by Schultz (1961) and Becker (1964)[2][17], in which health status is introduced as an additional form of human capital. The wage equation is extended to the following form, where the last term is a function of health condition.

$$lnw = lnw_0 + \rho s + \beta_1 x + \beta_2 x^2 + f(\theta, h) \tag{2}$$

With the general form of wage equation settled, efforts are then taken to identify the source of wage differentials. A widely adopted framework is introduced by Blinder (1973) and Oaxaca (1973)[7][16], which proposes the decomposition of gender wage gap into difference in productivity, difference in return to human capital, and unexplained difference representing gender discrimination. The model can be denoted as:

$$wage\ gap = \beta_A(\overline{X_A} - \overline{X_B}) + \overline{X_B}(\beta_A - \beta_B) + unexplained \tag{3}$$

This study focuses on the relationship between wage and education background of workers. It is established upon early studies that distinguish between the signaling effect of education and the improvement of productivity from education [4][8][15].

There are four important implications from previous studies. Firstly, education enhances labor market outcomes of workers [11]. Secondly, the impact of education is heterogeneous over the population [5]. Thirdly, education-wage profile is a crucial source of wage inequality in labor market [10]. Lastly, regardless of the possible endogeneity of education due to omitting ability in wage equation, OLS estimation and TSLS estimation are comparable, which suggests that bias of estimation with a simple OLS model is tolerable[4].

The study utilizes data from UK labor survey to detect the effect of university education on labor market outcomes. The rest of the study is divided into seven sections in accordance with the instruction.

2. Model description

Following the tradition of wage equation introduced above, the study determines to adopt a semi-log form, in which the response variable is hourly wage in log transformation. The decision is based on the distribution of hourly wage in its level and log forms presented in Figure 1, since the level term is highly skewed while log transformation mitigates the problem.

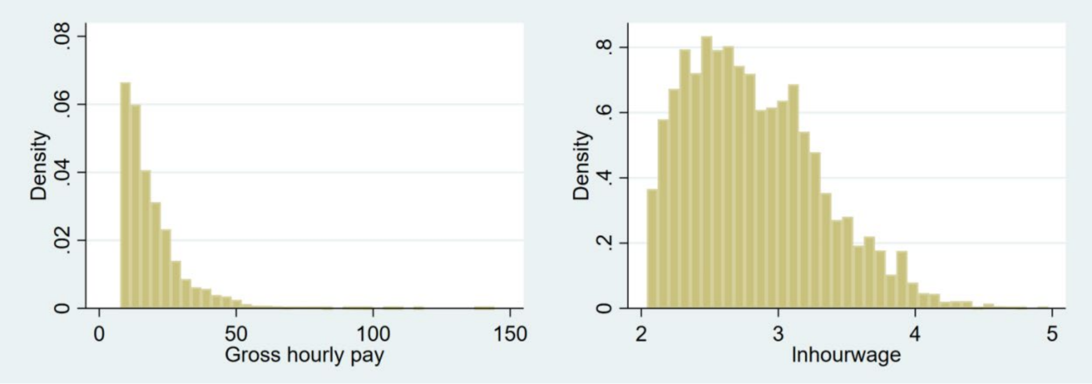


Figure 1: Histogram of hourly wage and log transformation

For the choice of independent variables, the study considers the framework proposed by Schultz (1961) and Becker (1964), in which education, working experience (and its quadratic term), and health status play the crucial roles. However, while the Mincer equation utilizes a continuous measurement of education by years, the data set contains only categorical variables for education level. The study takes "none" as the reference group and includes the binary variables for other three levels of education background. Meanwhile, the study adds in control variables for gender, location, firm type, and position of the worker. The baseline model can be written as:

$$\ln(wage_i) = \beta_0 + \beta_1 female_i + \beta_2 gcse_i + \beta_3 alevel_i + \beta_4 degree_i + \beta_5 potexp_i + \beta_6 potexp_i^2 + \beta_7 healthy_i + \theta X_i + \eta_i \quad (4)$$

Where X_i contains the dummy variables for company location, firm type, and position respectively.

Apart from the baseline estimation, the study digs further into the possible interaction effect of education with other explanatory variables. In detail, the study is interested in whether a college degree affects gender wage gap as well as the return to health and working experience. The framework of Chow break test is adopted for the analysis, where estimation is separated for two subsamples by completion of college education.

3. Hypothesis statement

Then central goal of the study is to check whether there is a positive return to education. The return to different levels of education in comparison with illiterate workers is performed with t test of single coefficient:

$$H_0: \beta_j = 0 \; H_1: \beta_j > 0 \; (j = 2,3,4) \quad (5)$$

In addition, a joint hypothesis test with F statistics is conducted to examine the overall effect of education.

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0 \; H1: At\ least\ one\ of\ the\ coefficients\ is\ non-zero. \quad (6)$$

Moreover, the study examines whether college degree affects the impact of gender, experience, and health condition in wage equation. The hypothesis test is achieved with a Chow break test comparing the estimation between workers with and without a university degree. The associated hypothesis statement is:

$$H_0: \beta_j^{degree} = \beta_j^{nondegree}\ (j = 0,1,\dots,k)\ H1: At\ leaset\ one\ coefficient\ is\ not\ equal. \quad (7)$$

4. Data issues

Quality of data is critical to the internal validity of estimation. From the aspect of Gauss-Markov estimation assumptions [19], the study identifies the following issues of the raw data.

The first problem is sample selection problem. As is reported in the summary statistics, hourly wage ranges from 7.7 to 144.23, which does not include zero or missing values. It indicates that the sample is based on employed workers only. However, as is previously studied, poor education also contributes to a lower labor for participation and a higher unemployment rate [6]. Then the focus on employed sample alone may lead to an underestimation of the educational return.

The second problem is measurement error. The study could suffer from two types of measurement error problems. For one thing, hourly wage could contain measurement error due to rounding and misreport. Specially, for workers whose salary are paid by month or year, the estimated level of hourly payment could be noisy. In addition, the rounding error for data collection also introduces error in measurement. However, as long as the measurement error is random, the issue only leads to a higher variance of estimation while does not affect the unbiasedness of estimators. For the other, measurement error in independent variables such as working experience can be fatal. The variable definition section indicates that experience is derived from age and education level, which does not take into consideration the interim of working due to issues like pregnancy or unemployment. Such a measurement error would bring in the attenuation bias if it is in the form of classical measurement error in regressor.

The third problem is perfect collinearity in regressors. By definition, age, education, and working experience are perfectly collinear, and the same issue is present for the four education indicators and the four location indicators as well. The solution is to leave one out when forming the regression model.

Lastly, the distribution of hourly wage after log transformation still deviates from a normal distribution. The remaining issue of positive skewness could lead to heteroscedasticity, which is handled with the adoption of robust standard error in statistical tests.

5. Model specification test

The estimated baseline model is:

$$\ln(\widehat{wage_i}) = 2.1 - 0.20 female_i + 0.11 gcse_i + 0.18 alevel_i + 0.54 degree_i + 0.038 potexp_i - 0.00067\, potexp_i^2 + 0.06 healthy_i \quad (8)$$

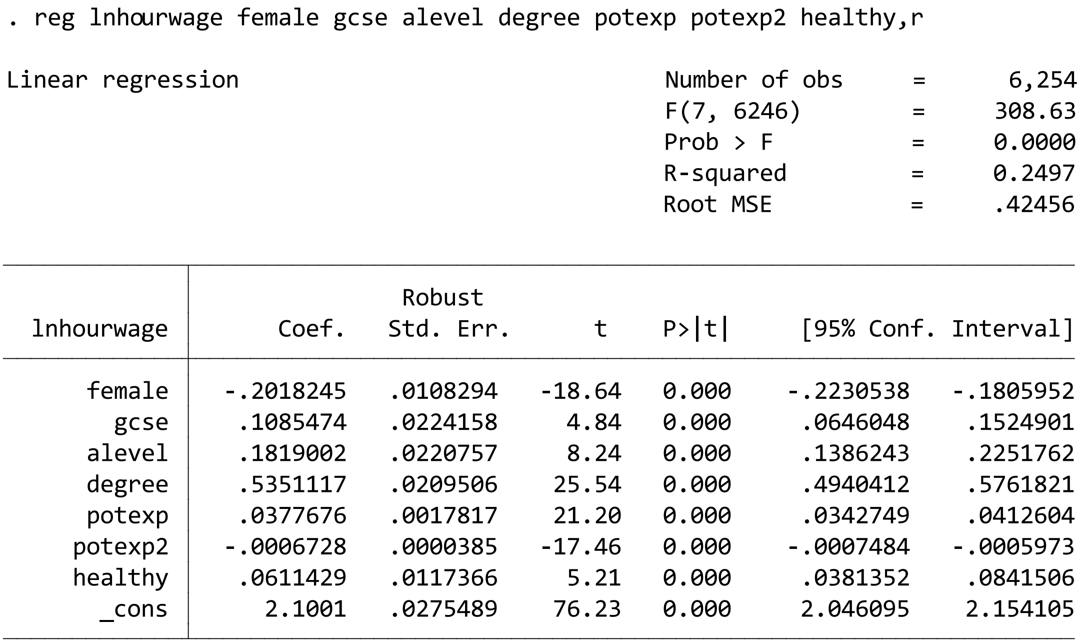
The change in estimated coefficient is not substantial when control variables are included:

$$\ln(\widehat{wage_i}) = 2.06 - 0.14 female_i + 0.10 gcse_i + 0.16 alevel_i + 0.46 degree_i + 0.03 potexp_i - 0.00054\, potexp_i^2 + 0.06 healthy_i + \hat{\theta} X_i \quad (9)$$

The study considers three sets of post estimation diagnostic tests for modeling assumptions. The first is Ramsey RESET test(See Figure 2), which checks whether the model has missing non-linearity by including the power terms of fitted values back into the regression model. The second is White test of heteroscedasticity(See Figure 3), which checks whether constant variance assumption is violated by regressing the squared terms of residual on explanatory variables, quadratic forms, and their interaction terms. The last is VIF test(See Figure 4), which checks whether the model suffers from high multicollinearity problem.

The relevance of these specification tests lies in that violation of linearity assumptioin renders estimators to be biased. In addition, non-constant variance leads to the failure of hypothesis test, while the problem of multicollinearity leads to imprecise estimators and insignificant t test statistics.

The stata output below reports the three test statistics. The RESET test indicates the issue of omitted variables (p value=0.00), which should be solved with the introduction of higher order terms like quadratic or interaction terms. The White test indicates the problem of heteroscedasticity (p value=0.00), which is solved by reporting the robust standard errors for hypothesis test instead. The VIF statistics is less than 10, which refuses the problem of high multicollinearity.

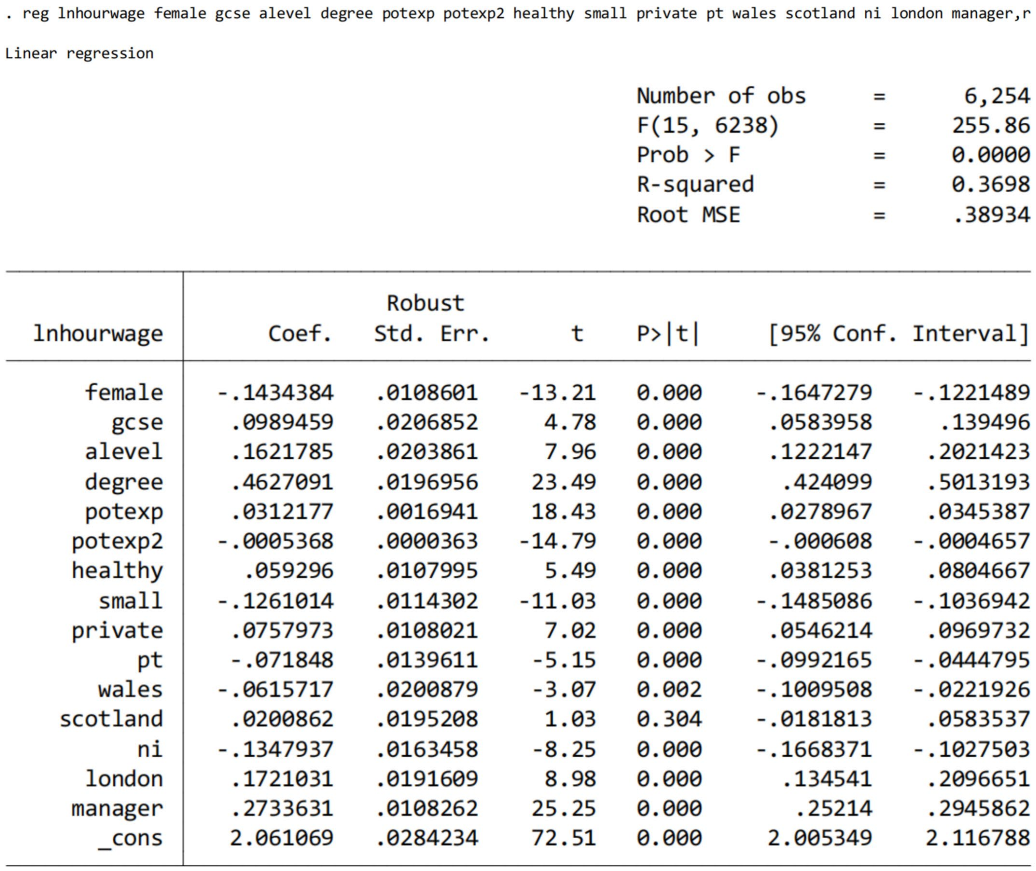


```
. reg lnhourwage female gcse alevel degree potexp potexp2 healthy,r

Linear regression                               Number of obs     =      6,254
                                                F(7, 6246)        =     308.63
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2497
                                                Root MSE          =     .42456

------------------------------------------------------------------------------
             |               Robust
  lnhourwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2018245   .0108294   -18.64   0.000    -.2230538   -.1805952
        gcse |   .1085474   .0224158     4.84   0.000     .0646048    .1524901
      alevel |   .1819002   .0220757     8.24   0.000     .1386243    .2251762
      degree |   .5351117   .0209506    25.54   0.000     .4940412    .5761821
      potexp |   .0377676   .0017817    21.20   0.000     .0342749    .0412604
     potexp2 |  -.0006728   .0000385   -17.46   0.000    -.0007484   -.0005973
     healthy |   .0611429   .0117366     5.21   0.000     .0381352    .0841506
       _cons |     2.1001   .0275489    76.23   0.000     2.046095    2.154105
------------------------------------------------------------------------------
```

Figure 2: Ramsey RESET test



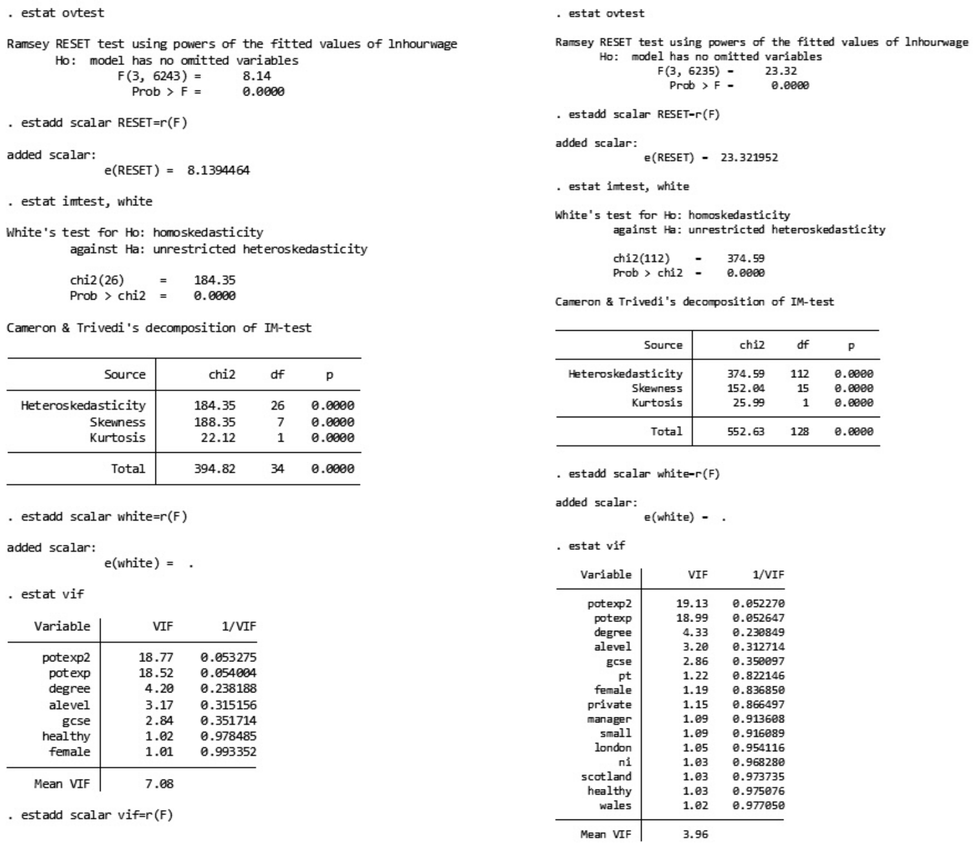
```
. reg lnhourwage female gcse alevel degree potexp potexp2 healthy small private pt wales scotland ni london manager,r

Linear regression

                                                Number of obs     =      6,254
                                                F(15, 6238)       =     255.86
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3698
                                                Root MSE          =     .38934

------------------------------------------------------------------------------
             |               Robust
  lnhourwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.1434384   .0108601   -13.21   0.000    -.1647279   -.1221489
        gcse |   .0989459   .0206852     4.78   0.000     .0583958     .139496
      alevel |   .1621785   .0203861     7.96   0.000     .1222147    .2021423
      degree |   .4627091   .0196956    23.49   0.000      .424099    .5013193
      potexp |   .0312177   .0016941    18.43   0.000     .0278967    .0345387
     potexp2 |  -.0005368   .0000363   -14.79   0.000     -.000608   -.0004657
     healthy |    .059296   .0107995     5.49   0.000     .0381253    .0804667
       small |  -.1261014   .0114302   -11.03   0.000    -.1485086   -.1036942
     private |   .0757973   .0108021     7.02   0.000     .0546214    .0969732
          pt |   -.071848   .0139611    -5.15   0.000    -.0992165   -.0444795
       wales |  -.0615717   .0200879    -3.07   0.002    -.1009508   -.0221926
    scotland |   .0200862   .0195208     1.03   0.304    -.0181813    .0583537
          ni |  -.1347937   .0163458    -8.25   0.000    -.1668371   -.1027503
      london |   .1721031   .0191609     8.98   0.000      .134541    .2096651
     manager |   .2733631   .0108262    25.25   0.000       .25214    .2945862
       _cons |   2.061069   .0284234    72.51   0.000     2.005349    2.116788
------------------------------------------------------------------------------
```

Figure 3: White test of heteroscedasticity



```
. estat ovtest

Ramsey RESET test using powers of the fitted values of lnhourwage
       Ho:  model has no omitted variables
                F(3, 6243) =      8.14
                  Prob > F =      0.0000

. estadd scalar RESET=r(F)

added scalar:
              e(RESET) =  8.1394464

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(26)     =    184.35
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |     184.35     26    0.0000
            Skewness |     188.35      7    0.0000
            Kurtosis |      22.12      1    0.0000
---------------------+-----------------------------
               Total |     394.82     34    0.0000
---------------------------------------------------

. estadd scalar white=r(F)

added scalar:
              e(white) =  .

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
     potexp2 |     18.77    0.053275
      potexp |     18.52    0.054004
      degree |      4.20    0.238188
      alevel |      3.17    0.315156
        gcse |      2.84    0.351714
     healthy |      1.02    0.978485
      female |      1.01    0.993352
-------------+----------------------
    Mean VIF |      7.08

. estadd scalar vif=r(F)
```

```
. estat ovtest

Ramsey RESET test using powers of the fitted values of lnhourwage
       Ho:  model has no omitted variables
                F(3, 6235) =     23.32
                  Prob > F =      0.0000

. estadd scalar RESET=r(F)

added scalar:
              e(RESET) =  23.321952

. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(112)    =    374.59
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |     374.59    112    0.0000
            Skewness |     152.04     15    0.0000
            Kurtosis |      25.99      1    0.0000
---------------------+-----------------------------
               Total |     552.63    128    0.0000
---------------------------------------------------

. estadd scalar white=r(F)

added scalar:
              e(white) =  .

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
     potexp2 |     19.13    0.052270
      potexp |     18.99    0.052647
      degree |      4.33    0.230849
      alevel |      3.20    0.312714
        gcse |      2.86    0.350097
          pt |      1.22    0.822146
      female |      1.19    0.836850
     private |      1.15    0.866497
     manager |      1.09    0.913608
       small |      1.09    0.916089
      london |      1.05    0.954116
          ni |      1.03    0.968280
    scotland |      1.03    0.973735
     healthy |      1.03    0.975076
       wales |      1.02    0.977050
-------------+----------------------
    Mean VIF |      3.96
```

Figure 4: VIF test

6. Research findings

Table 1 illustrates baseline estimation, with first column considering only key explanatory variables and second column including full set of control variables.

Table 1: Baseline estimation

	(1) lnhourwage	(2) lnhourwage
female	-0.20*** (0.011)	-0.14*** (0.011)
gcse	0.11*** (0.022)	0.099*** (0.021)
alevel	0.18*** (0.022)	0.16*** (0.020)
degree	0.54*** (0.021)	0.46*** (0.020)
potexp	0.038*** (0.0018)	0.031*** (0.0017)
potexp2	-0.00067*** (0.000039)	-0.00054*** (0.000036)
healthy	0.061*** (0.012)	0.059*** (0.011)
_cons	2.10*** (0.028)	2.06** (0.028)
Control variables		
N	6254	6254
R2	0.25	0.37
white	184.35	374.59
RESET	8.14	23.3
vif	7.08	3.96

*Standard errors in parentheses;*p<0.05,**p<0.01,***P<0.001*

For the key research question regarding educational return, the estimation suggests an increase of earnings with education. In detail, compared with illiterate workers, GCSE qualification increases wage by a tenth, A level qualification increases wage by a sixth, and college degree increase wage by more than a half. The effect is statistically significant at 1% level for single variable t test, and jointly significant at the level of 1% as well.

In terms of other key explanatory variables in the wage equation, table 1 yields the following conclusions. Firstly, there is a gender wage difference after considering the level of productivity. Ceteris paribus, a female worker earns 14% to 20% lower than her male counterparts and the gap is statistically significant at 1% level. Secondly, health status is associated with a wage differential in labor market. Ceteris paribus, healthy workers earn 6% more and the effect is statistically significant at 1% level. Lastly, earnings increase with working experience at a diminishing speed. For a freshman with no working experience, the first year of experience improves hourly earnings by 3.1% to 3.8%. The effect is statistically significant at 1% level. However, the gain from working experience decays over time. Holding all else equal, hourly wage peaks with 28 years of working experience, after which salary begins to decrease with additional experience accumulation.

Besides, for joint hypothesis test, the study finds evidence for wage inequality over location, firm type, and position of workers separately. In detail, the location dummy variables (three regional dummy plus one dummy indicator for London) have an F statistics of 44.42 (p value=0.00), which indicates geographical inequality in UK labor market. Meanwhile, small companies offer a lower wage by 12.6% while private companies offer a higher wage by 7.58%, both individually significant at 1% level. The two are jointly significant at 1% level as well (F statistics=70.07), suggesting the earnings differential across firm types. Lastly, part time workers are found to earn a lower wage by 7.18% and managers are found to outperform others by 27.33%. The two are statistically significant at 1% level in both single variable t test and joint test, indicating the effect of job type on earnings.

For overall model fitness, the first column explains only a fourth of the variation in hourly wage, while the proportion increases to 37% in the second column. It suggests that the control variables are jointly significant in the wage equation and cannot be excluded from estimation.

Then Table 2 reports the estimation for non-degree workers and degree holders. The estimation suggests an expanded gender gap among degree holders. Meanwhile, the return to experience as well as health is lower for non-degree workers. In addition, the intercept term suggests a college premium by 23% after accounting for the changed marginal effect by other explanatory variables with a college degree.

The chow break test statistics is 89.32, which is statistically significant at 1% level. Therefore, at the significance level of 1%, there is sufficient evidence for a structural break of wage equation between non-degree workers and college (or above) graduates.

$$F=\frac{\frac{1110.8-579.69-368.02}{12}}{\frac{579.69+368.02}{6228}}=89.32 \tag{10}$$

Table 2: Subsample estimation

	(1) degree-0	(2) degree-1
female	-0.11*** (0.016)	-0.16*** (0.015)
potexp	0.024*** (0.0027)	0.037*** (0.0023)
potexp2	-0.00039*** (0.000055)	-0.00070*** (0.000053)
healthy	0.058*** (0.015)	0.016*** (0.015)
_cons	2.26*** (0.034)	2.49*** (0.027)
Control variables		
N	2715	3539
R^2	0.203	0.294

*Standard errors in parentheses;*p<0.05,**p<0.01,***p<0.001*

7. Discussion

Ethnical groups would be important regressors to include in the model. For one thing, previous studies have provided evidence for racial inequality in labor market [1][3]. It indicates that race is a crucial explanator variable in wage equation. Specially, race is associated with education attainment as well [12]. Since being black is found to be associated with a lower wage and a poorer education achievement simultaneously, omitting racial groups from the estimation could lead to an overestimation of education return. For the other, racial groups could be related with a structural break in wage equation as well. In detail, return to education as well as gender wage gap is found to differ over racial groups [9]. Therefore, no only the level term of racial groups but also the interaction term of race indicators with other explanatory variables should be included in the analysis.

However, I find the survey question poorly designed as it contains potentially overlapping groups such as Chinese and Asian, Black and Mixed. It has too many categories which could lead the comparison between different ethnical groups to be unclear. Instead of creating a binary indicator for each of the seven categories, I would include only a binary variable of white to distinguish those privileged and potentially discriminated ethnical groups in job market. Such a design makes possible and convenient the identification as well as statistical inference with a chow break test framework.

8. Project Extension

There are two potential extensions to conduct. The first caters for the issue of sample selection discussed in the third section. Instead of collecting data only from those employed workers, the extension should be a population census data that covers inactive and unemployed workers as well. Then through a probit/logit model, an extension analysis can reveal how health status affects the opportunity of getting a job. In addition, by estimating a tobit model, the extension analysis can reveal the true causal effect of being unhealthy on wage income, which accounts for the loss of job opportunities as well. The second extension will involve the usage of a panel data set. In the current cross-sectional data set, the estimation is rather descriptive rather than causal due to the possible endogeneity of education and the issue of cohort effect [18]. If a pooled cross sectional data set with a possible policy shock (eg. staggered expansion of college enrollment in China) is available, a two way fixed effect estimation could be useful.

9. Conclusion

The study estimates wage equation using UK labor force survey.

The baseline estimation indicates a rise of earning with education attainment. Specially, education return gets higher with increased years of education. Further analysis with subsample estimation suggests that college degree affects wage both in level form and in its interaction with other regressors. In detail, while there is a pure college premium by 23%, the effect of college degree is also illustrated by an expanded gender gap as well as a higher return to experience and health. Besides, the study points out the limitations due to truncation, measurement error, and endogeneity, while proposes associated solutions to be performed in future studies.

References

[1] Altonji, J.G. and Blank, R.M., 1999. Race and gender in the labor market. Handbook of labor economics, 3, pp.3143-3259.

[2] Becker, G.S., 1964. Human capital: A theoretical and empirical analysis, with special reference to education. University of Chicago press.

[3] Cain, G.G., 1986. The economic analysis of labor market discrimination: A survey. Handbook of labor economics, 1, pp.693-785.

[4] Angrist, J.D. and Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics, 106(4), pp.979-1014.

[5] Balestra, S. and Backes-Gellner, U., 2017. Heterogeneous returns to education over the wage distribution: Who profits the most? Labour Economics, 44, pp.89-105.

[6] Biagi, F. and Lucifora, C., 2008. Demographic and education effects on unemployment in Europe. Labour Economics, 15(5), pp.1076-1101.

[7] Blinder, A.S., 1973. Wage discrimination: reduced form and structural estimates. Journal of Human resources, pp.436-455.

[8] Brown, S. and Sessions, J.G., 2004. Signalling and screening. International handbook on the economics of education, 9, pp.58-100.

[9] Browne, I. and Misra, J., 2005. Labor-market inequality: intersections of gender, race, and class (pp. 165-189). Londres, Blackwell Publishing.

[10] Dos Reis, J.G.A. and De Barros, R.P., 1991. Wage inequality and the distribution of education: A study of the evolution of regional differences in inequality in metropolitan Brazil. Journal of Development Economics, 36(1), pp.117-143.

[11] Griliches, Z., 1997. Education, human capital, and growth: a personal perspective. Journal of Labor Economics, 15(1, Part 2), pp.S330-S344.

[12] Lang, K. and Manove, M., 2011. Education and labor market discrimination. American Economic Review, 101(4), pp.1467-1496.

[13] Mincer, J., 1958. Investment in human capital and personal income distribution. Journal of political economy, 66(4), pp.281-302.

[14] Mincer, J. and Polachek, S., 1974. Family investments in human capital: Earnings of women. Journal of political Economy, 82(2, Part 2), pp.S76-S108.

[15] Noldeke, G. and Van Damme, E., 1990. Signalling in a dynamic labour market. The Review of Economic Studies, 57(1), pp.1-23.

[16] Oaxaca, R., 1973. Male-female wage differentials in urban labor markets. International economic review, pp.693-709.

[17] Schultz, T.W., 1961. Investment in human capital. The American economic review, 51(1), pp.1-17.

[18] Stapleton, D.C. and Young, D.J., 1988. Educational attainment and cohort size. Journal of Labor Economics, 6(3), pp.330-361.

[19] Wooldridge, J.M., 2015. Introductory econometrics: A modern approach. Cengage learning.