ellipsis
DATA GROUP

Wage Analysis Brief

Descriptive and regression analysis of hourly wages using Stata

Prepared by: Koome Laititi
Date: 10 November 2026
Dataset: NLSW 1988 extract (nlsw88), 2,246 observations and 17 variables
Dataset link: http://www.stata-press.com/data/r12/nlsw88.dta
In Stata: `sysuse nlsw88, clear`

1. Purpose and Data

This brief examines how individual characteristics, educational attainment and employment-related factors explain variation in hourly wages. It uses the National Longitudinal Survey of Women (NLSW) 1988 extract, which contains 2,246 observations and 17 variables for working women in their 30s and early 40s. Table 1 lists the variables.

The **dependent variable** is hourly wage (wage). The **independent variables** fall into four groups: demographic characteristics (age, race and marital status); educational attainment (years of schooling and college graduation status); geographical location (central city residence); and employment characteristics (industry, union membership, usual hours worked, total work experience and job tenure).

The analysis proceeds in two steps. First, each variable is regressed on log hourly wages separately (Models 1 to 11). Second, all variables are entered together in a combined model. All models use heteroskedasticity-robust standard errors.

Table 1: Description of the dataset (Stata describe output)

```
Contains data from C:\Users\ADMIN\OneDrive\Desktop\softwares\Stata17\ado\base/n/nlsw88.dta
 Observations:         2,246                  NLSW, 1988 extract
    Variables:            17                  1 May 2020 22:52
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
idcode          int     %8.0g                 NLS ID
age             byte    %8.0g                 Age in current year
race            byte    %8.0g      racelbl    Race
married         byte    %8.0g      marlbl     Married
never_married   byte    %16.0g     nev_mar    Never married
grade           byte    %8.0g                 Current grade completed
collgrad        byte    %16.0g     gradlbl    College graduate
south           byte    %9.0g      southlbl   Lives in the south
smsa            byte    %9.0g      smsalbl    Lives in SMSA
c_city          byte    %16.0g     ccitylbl   Lives in a central city
industry        byte    %23.0g     indlbl     Industry
occupation      byte    %22.0g     occlbl     Occupation
union           byte    %8.0g      unionlbl   Union worker
wage            float   %9.0g                 Hourly wage
hours           byte    %8.0g                 Usual hours worked
ttl_exp         float   %9.0g                 Total work experience (years)
tenure          float   %9.0g                 Job tenure (years)
```

Source: nlsw88.dta, Stata 17.

2. The Dependent Variable: Hourly Wage

Hourly wage is strongly positively skewed (skewness = 3.10, kurtosis = 15.85), as Figure 1 shows. Most respondents earn less than about USD 15 per hour. The median wage (USD 6.27) is below the mean (USD 7.77), which indicates that a small group of high earners pulls the average upward.

Wages are widely dispersed, ranging from USD 1.00 to USD 40.75, with a standard deviation of USD 5.76. A quarter of workers earn USD 4.26 or less, half earn USD 6.27 or less, and three quarters earn USD 9.60 or less. The 99th percentile is USD 38.71 (Table 2).

Because of this skewness, wages were transformed into natural logarithms. The transformation reduces the influence of extreme values and gives a more suitable functional form for wage

determination. It also allows each coefficient to be read as an approximate percentage difference in hourly wages. Exact percentage changes are calculated as 100 × (e^β − 1).

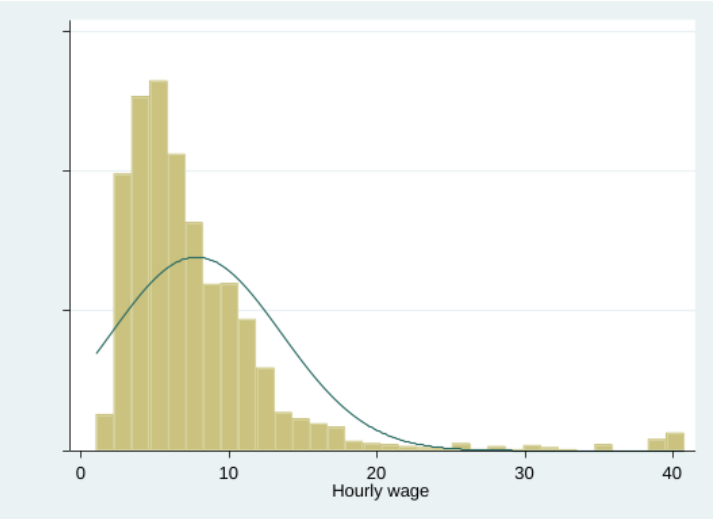


Figure 1: Distribution of hourly wages, with normal curve overlay

Table 2: Summary statistics of hourly wage (Stata summarize, detail output)

```
. summarize wage, detail

                        Hourly wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.930993       1.004952
 5%     2.801002       1.032247
10%     3.220612       1.151368       Obs               2,246
25%     4.259257       1.344605       Sum of wgt.       2,246

50%      6.27227                      Mean           7.766949
                        Largest       Std. dev.      5.755523
75%     9.597424       40.19808
90%     12.77777       40.19808       Variance       33.12604
95%     16.52979       40.19808       Skewness       3.096199
99%     38.70926       40.74659       Kurtosis       15.85446
```

Source: nlsw88.dta, Stata 17.

3. Single-Variable Regressions

Table 3 summarizes the eleven single-variable models. Each model regresses log hourly wages on one explanatory variable. The numbers of observations differ slightly across models because of missing values.

Table 3: Summary of single-variable regressions (dependent variable: log hourly wages)

Model	Independent variable	Coefficient	% change	Obs.	R^2	F-stat.	Prob > F	Root MSE
1	Union membership (1 = Yes)	0.2249***	25.22%	1,878	0.0350	79.85	0.0000	0.50838
2	Education (years of schooling)	0.0948***	9.95%	2,244	0.1730	508.38	0.0000	0.52279
3	College graduate (1 = Yes)	0.4340***	54.34%	2,246	0.1032	245.08	0.0000	0.54426
4	Age (years)	-0.0042	-0.42%	2,246	0.0005	1.12	0.2897	0.57457
5	Race: Black (ref: White)	-0.1646***	-15.18%	2,246	0.0161	18.36	0.0000	0.57019
	Race: Other (ref: White)	0.0601	6.19%					
6	Married (1 = Yes)	-0.0246	-2.43%	2,246	0.0004	0.89	0.3443	0.57459
7	Central city (1 = Yes)	0.0738***	7.66%	2,246	0.0034	7.65	0.0057	0.57373
8	Industry (11 categories)	See Table A1		2,232	0.1159	32.49	0.0000	0.54089
9	Hours worked (per hour)	0.0112***	1.13%	2,242	0.0420	77.97	0.0000	0.56193
10	Total experience (years)	0.0480***	4.92%	2,246	0.1483	433.15	0.0000	0.53039
11	Job tenure (years)	0.0312***	3.17%	2,231	0.0900	266.32	0.0000	0.54674

*Robust standard errors were used in all models. *** p<0.01, ** p<0.05, * p<0.1. % change = 100 × (e^β − 1). Industry results are reported in Table A1.*

Key findings

- **Education:** Years of schooling has the highest explanatory power of any single variable (R^2 = 0.1730). Each additional year is associated with about 9.95% higher hourly wages. College graduates earn about 54.34% more than non-graduates in this unadjusted comparison (R^2 = 0.1032).
- **Experience and job characteristics:** Total experience (R^2 = 0.1483) is the next strongest single predictor, followed by industry (R^2 = 0.1159) and tenure (R^2 = 0.0900). Each extra year of experience and tenure is associated with about 4.92% and 3.17% higher wages. Union members earn about 25.22% more than non-members, and each additional hour worked is associated with about 1.13% higher wages.
- **Industry:** Relative to the base industry, wages are significantly higher in Transport (111.29%), Public administration (69.56%), Finance/Insurance/Real estate (60.66%), Professional services (39.14%) and Manufacturing (31.25%). Mining and Construction are significant only at the 10% level.
- **Race and location:** Black workers earn about 15.18% less than White workers, while the difference for “Other” is not significant. Central city residents earn about 7.66% more, although location alone explains only 0.34% of the variation in wages.
- **Age and marital status:** Neither is statistically significant on its own.

4. Combined Regression Model

Table 4 reports the combined model, in which all explanatory variables are entered together. The model is jointly significant (F = 68.65, p < 0.001) and has an R-squared of 0.4227. Together, the demographic, educational, locational and employment characteristics explain about 42.27% of the variation in log hourly wages, which is far higher than any single-variable model.

Table 4: Combined regression model (dependent variable: log hourly wages)

Independent variable	Coefficient (robust SE)	% change
Demographic characteristics		
Age (years)	-0.0045 (0.0031)	-0.45%
Race: Black (ref: White)	-0.1388*** (0.0222)	-12.96%
Race: Other (ref: White)	0.0273 (0.0908)	2.77%
Married (1 = Yes)	-0.0041 (0.0210)	-0.41%
Educational attainment		
Education (years of schooling)	0.0665*** (0.0060)	6.88%
College graduate (1 = Yes)	0.0700* (0.0371)	7.25%
Geographical location		
Central city (1 = Yes)	0.0653*** (0.0217)	6.75%
Employment characteristics		
Industry (ref: Agriculture, forestry and fisheries)		
Mining	0.3012 (0.2380)	35.15%
Construction	0.3203** (0.1391)	37.76%
Manufacturing	0.1301 (0.1076)	13.89%
Transport/Communications/Utilities	0.4308*** (0.1154)	53.84%
Wholesale/Retail trade	-0.0272 (0.1077)	-2.68%
Finance/Insurance/Real estate	0.2434** (0.1109)	27.56%
Business/Repair services	0.1432 (0.1186)	15.39%
Personal services	-0.1138 (0.1168)	-10.76%
Entertainment/Recreation	0.1377 (0.1654)	14.77%
Professional services	0.0240 (0.1059)	2.43%
Public administration	0.2286** (0.1082)	25.69%
Union membership (1 = Yes)	0.1191*** (0.0214)	12.65%
Hours worked (per hour)	0.0013 (0.0013)	0.13%
Total experience (years)	0.0286*** (0.0027)	2.90%
Job tenure (years)	0.0078*** (0.0021)	0.79%
Constant	0.6101*** (0.1838)	
Model statistics		
Observations	1,852	
R-squared	0.4227	
F-statistic (22, 1829)	68.65	
Prob > F	0.0000	
Root MSE	0.39336	

*Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. % change = 100 × (e^β − 1). The sample is smaller than in the single-variable models because of missing values.*

Key findings

- **Education remains important.** Holding other factors constant, each additional year of schooling is associated with about 6.88% higher wages. College graduation adds a further 7.25%, but this is significant only at the 10% level (p = 0.059).

- **Union membership, location and experience matter.** Union members earn about 12.65% more and central city residents about 6.75% more. Each additional year of total experience and job tenure is associated with about 2.90% and 0.79% higher wages.
- **The racial wage gap persists.** Black workers earn about 12.96% less than White workers with otherwise similar characteristics.
- **Several effects shrink.** The union premium falls from 25.22% to 12.65% and the college premium from 54.34% to 7.25%, which shows that the single-variable results partly reflect overlapping factors such as experience and industry. Hours worked becomes insignificant, as do age, marital status and “Other” race.
- **Industry:** Only Transport (53.84%), Construction (37.76%), Finance (27.56%) and Public administration (25.69%) differ significantly from the base industry.

5. Conclusions and Limitations

Hourly wages are significantly associated with education, work experience, job tenure, union membership, industry, race and central city residence. Age, marital status and hours worked add little once these other factors are controlled for. The combined model explains about 42% of the variation in log wages, so a large share remains unexplained by the observed characteristics.

Four limitations should be kept in mind when interpreting the results:

- **Association, not causation.** The data are observational, and unobserved factors such as ability and family background may affect both education and wages. Industry, union membership and experience are also partly outcomes of education, so the combined model should not be read as a causal estimate of the return to college. A causal estimate would require an approach such as instrumental variables.
- **Different samples.** The combined model uses 1,852 observations, compared with 2,231 to 2,246 in most single-variable models, so R-squared values are not strictly comparable.
- **Overlapping education measures.** Years of schooling and college graduation are strongly correlated. The weak significance of college graduation in the combined model may partly reflect this, so a variance inflation factor (VIF) check is advisable.
- **Reference categories.** Coefficients for race and industry are relative to White workers and the Agriculture, forestry and fisheries industry respectively.

Appendix A: Industry Regression

Table A1: Industry and hourly wages (single-variable model 8, dependent variable: log hourly wages)

Industry (ref: Agriculture, forestry and fisheries)	Coefficient (robust SE)	% change
Mining	0.7819* (0.4236)	118.56%
Construction	0.3038* (0.1601)	35.50%
Manufacturing	0.2720** (0.1338)	31.25%
Transport/Communications/Utilities	0.7481*** (0.1395)	111.29%
Wholesale/Retail trade	0.0410 (0.1342)	4.19%
Finance/Insurance/Real estate	0.4741*** (0.1387)	60.66%
Business/Repair services	0.2130 (0.1472)	23.74%
Personal services	-0.2105 (0.1387)	-18.98%
Entertainment/Recreation	0.1312 (0.2036)	14.02%
Professional services	0.3303** (0.1322)	39.14%
Public administration	0.5280*** (0.1351)	69.56%
Constant	1.5804*** (0.1308)	
Model statistics		
Observations	2,232	
R-squared	0.1159	
F-statistic (11, 2220)	32.49	
Prob > F	0.0000	
Root MSE	0.54089	

*Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. % change = 100 × (e^β − 1).*