Department of Economics
UNIVERSITY OF MELBOURNE

ECOM30001/ECOM90001 Basic Econometrics
Semester 1, 2025

Capstone Project: Final Research Report

The Impact of Educational Attainment on Usual Hours Worked Per Week in the United States

Subject: ECOM30001/ECOM90001		**Subject Name:** Basic Econometrics
Assignment Name or Number: Assignment 2		
	Student ID Number	**Student Name**
1.	1706229	Brendan Cleary
2.	1692993	Jamie McCarville
3.	1681979	Jude Crotty
4.	1737725	Tyler Eagan

PLAGIARISM

Plagiarism is the presentation of an assignment which has in fact been copied in whole or in part from another student's work, or from another group's work, or from any other source (e.g. published books or periodicals), without due acknowledgement.

COLLUSION

Collusion is the presentation by a student or group of students of an assignment as their own which is in fact the result in whole or part of unauthorised collaboration with another person or persons.

DECLARATION

I/We declare that this assignment is my/our own work and does not involve plagiarism or unauthorized collusion.

Electronic Signature (Full Name)	Date
1. Brendan Cleary	26/05/25
2. Jamie McCarville	26/05/25
3. Jude Crotty	26/05/25
4. Tyler Eagan	26/05/25

Summary of Feedback Incorporated

We significantly altered our final report since the progress report in response to feedback. Most notably, we converted our independent variable back into its original categorical form instead of creating a continuous proxy for years of education. This gave us a clearer way to analyse summary statistics and the relationship of interest. We refined our analyses of these descriptive statistics for all variables, now excluding the non-linear AgeSquared. We revised some of our control variables, opting for the more general regional variables, limiting class categories to just public and private sector workers, and collapsing occupational variables into sensible groups.

We elaborated on the conflicting possibilities within our research question by exploring literature on the topic, polished our estimated model with better labelling, and refined our methodology by discussing more opportunities and consequences of bias. Furthermore, we incorporated a more precise interpretation of our main findings from the updated regression, alongside a comprehensive discussion of robustness in our model; we included checks for normality, functional form, heteroskedasticity, and multicollinearity, which helped to inform and revise the limitations of our research.

Research Question

Educational attainment is an influential mechanism that affects labour market outcomes such as earnings, productivity, employment levels, as well as hours worked (Ionescu and Cuza, 2012). Our research question states "what is the relationship between highest levels of educational attainment and the number of hours worked per week". Analysing the returns to education beyond wages provides valuable insights into how individuals allocate their time between labour and leisure, thus maximising utility.

Education can confer substantial economic benefits and longer working hours for graduates (Zhang, 2008), however it is also important to consider the greater flexibility as well as the intersection of socio-economic effects, notably gender in the labour market. These findings are crucial for policymakers in understanding how to best sustain employment when allocating public expenditure and investing in research and development. This information would also aid students' perceptions of the returns to education.

While investigating the relationship between educational attainment and usual hours worked, *it is expected that higher degrees of education will have a positive effect on usual hours worked, relative to those with less than high school education*.

Data Description

The data used for analysis in this project contains 2,304,683 person-level observations from the 2019 American Community Survey. It includes information such as social, demographic, and socioeconomic characteristics of the population of the U.S. The data sample has been restricted to individuals aged between 25 and 64 who are not living in group or institutional quarters. Furthermore, the sample for real hourly wage analysis is restricted to individuals who have worked in the past 12 months, which serves as an intended sample restriction. This helps maintain the integrity of the model and eliminates possible outliers that would not be relevant to the study. Additionally, it is important to note that the hours variable is top coded at 98 hours per week. The purpose of these restrictions is to narrow the data to a more specific topic and reduce heterogeneity.

Empirical Model

To evaluate the research question, the proposed econometric model is given by:

$$uhour_i = \beta_0 + \beta_1 EducationHS_i + \beta_2 EducationSomeCollege_i + \beta_3 EducationBachelors_i + \beta_4 EducationGraduate_i + \beta_5 Age_i + \beta_6 AgeSquared_i + \beta_7 Female_i + \beta_8 Married_i + \beta_9 NChildrenU18_i + \beta_{10} ForeignBorn_i + \beta_{11} Disability_i + \beta_{12} Private_i + \sum_{j=2}^{10} ,\ \beta_{13} Region_{ij} + \sum_{k=2}^{4} ,\ \beta_{14} OccupationGroup_{ik} + \sum_{m=2}^{4} ,\ \beta_{15} EnglishProficiency_{im} + \varepsilon_i \tag{1}$$

Here, **uhour** is a continuous dependent variable that that represents the usual work hours per week in the past 12 months. The explanatory variables of interest – highest level of educational attainment (β_1, β_2, β_3, β_4) modelled as dummy variables for **High School Education, Some College Education, Bachelor's Degree,** and **Graduate Degrees** respectively. These indicators allow assessment of how different education levels are associated with weekly work hours compared to those with less than high school education.

Moreover, there are other key explanatory variables that significantly influence the usual work hours per week and/or are related to an individual's approximate years of education completed:

- **Age** and **AgeSquared** are continuous variables that account for both the linear and non-linear effects of age on work intensity.

- **Female** is an indicator variable that represents the gender of the individual, reference group is males.
- **Married** is an indicator variable for the marriage status of the individual, non-married if not.
- **NChildrenU18** is a continuous variable that represents the number of household members under the age of 18.
- **ForeignBorn** is an indicator variable that represents whether the individual was born outside of the country, native-born if zero.
- **Disability** is an indicator variable that represents whether or not the individual has a disability or not.
- **Private** is an indicator variable that represents if an individual works in the private sector, reference group is those in the public sector.
- **Region** is a set of dummy variables that represent the regions that individuals work in (Midwest, South, West), with Northeast as reference group.
- **OccupationGroup** is a set of dummy variables that represent different types of occupations, grouped by nature of jobs (Manual Labor, Admin, Other), with 'Professional' as refence the reference group.
- **EnglishProficiency** is a full set of indicator variables that represents the levels of proficiency in English (Very well, Well, Not Well, Not at all), with 'Speak English at home' as reference group.

In addition, ε_i is a random error that represents the effects of 'other factors' and is assumed to have zero conditional mean.

Proposed Estimation Methodology

The method of Ordinary Least Squares (OLS) will be used to estimate the proposed econometric model (1). Provided the assumption $E[\varepsilon_i | X_i,$ Education$_i] = 0$ (MR2) is satisfied, the OLS process produces an unbiased estimator of the unknown population parameter β1. Basically, the conditional expected value of the random error is zero.

Furthermore, for the OLS estimator to become the Best Linear Unbiased Estimator, the assumptions required are:

- $\text{Var}[\varepsilon_i | X_i, \text{Education}_i] = \sigma^2$ (MR3) must be true, or else the model (1) exhibits heteroskedasticity, and the standard errors for the OLS estimators create misleading hypothesis tests. In other words, the random error's variance is independent of the explanatory variables and constant. As a robustness check, we used robust standard errors.
- $\text{Cov}(\varepsilon_i, \varepsilon_j | X_i, X_j) = 0$, for all i, j = 1, 2, ... N, i $\neq$ j (MR4) assumes the errors to be uncorrelated across observations. This assumption makes sense because the data is cross-sectional, and each observation refers to a different individual.
- Explanatory variables are not random (MR5a), and that is satisfied because the data is assumed to be drawn from a random sample of the population. It is also assumed to be independently and identically distributed.
- Exact collinearity does not exist (MR5b), or else the model (1) exhibits very large standard errors and covariances for the OLS estimators, resulting in misleading hypothesis tests. That is, no explanatory variable is an exact linear combination of the others. This assumption will be tested when checking for multicollinearity.
- $\text{Cov}(X, \varepsilon) \neq 0$, or else the model suffers from omitted variable bias or endogeneity, leading to bias and inconsistent estimators. This just means that the error term cannot correlate to any of the explanatory variables.

With all that said, as the model (1) includes only a select set of explanatory variables, there remains a risk of omitted variable bias, such as the innate motivation and skills of employees. The consequences of this can lead to biased estimators, inaccurate coefficient estimates, and misleading hypothesis tests. Nonetheless, this risk is mitigated by including a comprehensive set of demographic, regional, and labour-related controls which capture key dimensions of observed heterogeneity that might have otherwise confounded the relationship of interest.

Summary Statistics

Using the restrictions noted in our data description, the sample used for this analysis comprises 1,130,041 observations for individuals aged 25-64 in the United States. From this data, we obtained our summary statistics for individuals in five distinct categories by highest level of educational attainment: below high school, high school, some university, bachelor, and graduate. The average number of hours worked per week for individuals with less than a high school education was 39.41 hours per week, compared to 42.35 hours for those with graduate degrees. Females have a higher educational attainment on average, while age appears to not correlate with the level of educational attainment. The share of women increases from 38% with a less than high school education to 54% for the group with graduate degrees. This point suggests that marital rates also give interesting insights, with a consistent increase from 57% in the below-high school group to 72% with graduate degrees, suggesting marital couples are more likely to work more hours. Martial rates and no. of children also rise with education, with no. of children having a slight decrease for individuals in the high school category. Foreign-born individuals have the lowest level of educational attainment 45%, disability also declines across education groups: 9% for those with less than high school to 3% for those with a graduate degree. Educational attainment levels across regional variables throughout the United States appear to be consistent and randomized. English proficiency levels show that those who speak English very well or well, are more likely to be educated. For occupational industry, manual labour is by far the least educated group: 50% while the graduate degree group is 2%. However, among the professional group educational attainment rises sharply: from 9% for below high school to 88% for the graduate degree group.

Discussion of Results

Figure 2 presents the regression output using OLS and Robust Standard errors that tests our relationship of interest and provides some validation to our assumptions. As expected, the three specifications demonstrate that higher educational attainment is associated with statistically significant increases in usual hours worked per week compared to those with less than high school education, holding all else equal.

This is apparent in the initial basic specification which indicates an increasing positive impact of each degree of education on hours worked, relative to the base group; with a graduate degree corresponding to an average increase of almost 3 weekly hours compared to the reference group, ceteris paribus. While the magnitudes slightly decrease (aside for '*Some College'*) with the addition of control variables in specifications 2 and 3, the positive relationship remains statistically significant. The positive coefficient of 0.637 for Age unsurprisingly suggests that usual hours of worked increases on average as individuals get older, with the negative squared variable indicating that this effect eventually diminishes as they approach closer to retirement, holding all else equal.

Gender exhibits a notable negative and significant effect of 4.898, suggesting that women work on average nearly 5 hours less in their week than men, holding all other variables constant. The interactions between gender and education in specification 3 reveals that the negative effect of the female variable is slightly attenuated at higher education levels; indicating that women with graduate degrees work on average 0.98 more hours than women without, potentially narrowing the gap in female labour supply.

Analysis of Robustness

To detect for omitted variable bias and incorrect functional form in the model a RESET test was carried out for the squared predicted values, as well as both squared and cubed predicted values. From Figures 3 and 4, the p-values for both RESET tests were both below the level of significance; suggesting the functional form of the chosen econometric model is inadequate.

This highlights the potential for omitted unobserved variables such as innate motivation and skill. They are likely positively correlated with educational attainment and hours worked, thus introducing bias and overstating any casual effects from our variables of interest.

Additionally, for the OLS assumptions to hold, the residuals should be normally distributed around zero. In Figure 5, the residuals roughly follow a symmetric distribution centred on zero. The calculated skewness of 0.0151 (Figure 6) is close enough to zero to support little asymmetry. However, the calculated kurtosis value of 6.188 (Figure 6) indicates heavier tails than a normal distribution. Figure 7 shows the results of the Jarque-Bera test which strongly rejects the null hypothesis of normally distributed errors.

We conducted a Breusch-Pagan Test to test for heteroskedasticity in our regression model. The studentized Breusch-Pagan test yielded a p-value of 0.9386, above our 0.05 significance level. Suggesting there is no statistical evidence for heteroskedasticity, and our variance of the residuals appears to be constant (Figure 8); nonetheless we opted to use Robust standard errors as a precaution.

Given the typical progression of academic life cycles, the education dummies were tested for potential multicollinearity with the age variable. Sample correlations between age and the levels of education are all 0.10 and below in Figure 10, giving no indication of collinearity. Furthermore, the highest R^2 from the auxiliary regressions in Figure 11 was 0.0082, providing room to assume no multicollinearity here.

Conclusion

This analysis suggests a positive relationship between hours worked and educational attainment, consistent with some of Zhang's (2008) findings and our initial hypothesis. The relationship is influenced by various socio-economic factors, and suggests that highly educated individuals may possess greater job responsibility, higher returns to education and career-driven motivations. Additionally, education appears to reduce the gap in gender differences in labour supply at higher attainment levels, indicating to policymakers the benefit investing in education has on reducing gender disparities in the labour force

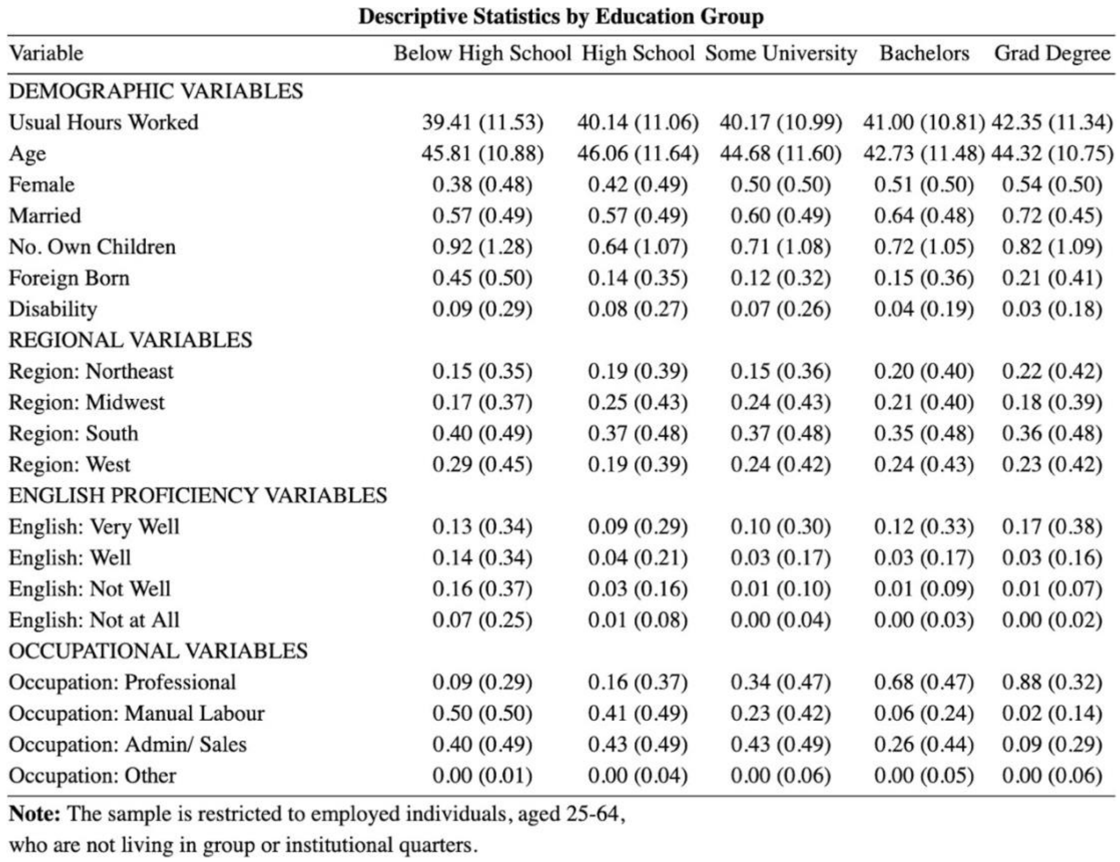
However, limitations to the findings include an overreliance on cross-sectional data and omitted variable bias such as individual motivation or job preferences. Further adjustment of the explanatory variables by including omitted variables and the use of correct functional form would better isolate the true effect of education on labour supply.

References

1. Ionescu, A. M., & Cuza, A. I. (2012). How does education affect labour market outcomes. *Review of Applied Socio-Economic Research*, *4*(2), 130-144.
2. Zhang, L. (2008). The Way to Wealth and the Way to Leisure: The Impact of College Education on Graduates' Earnings and Hours of Work. *Research in Higher Education*, *49*(3), 199–213. http://www.jstor.org/stable/gr25704558

Appendices

Figure 1. Summary Statistics

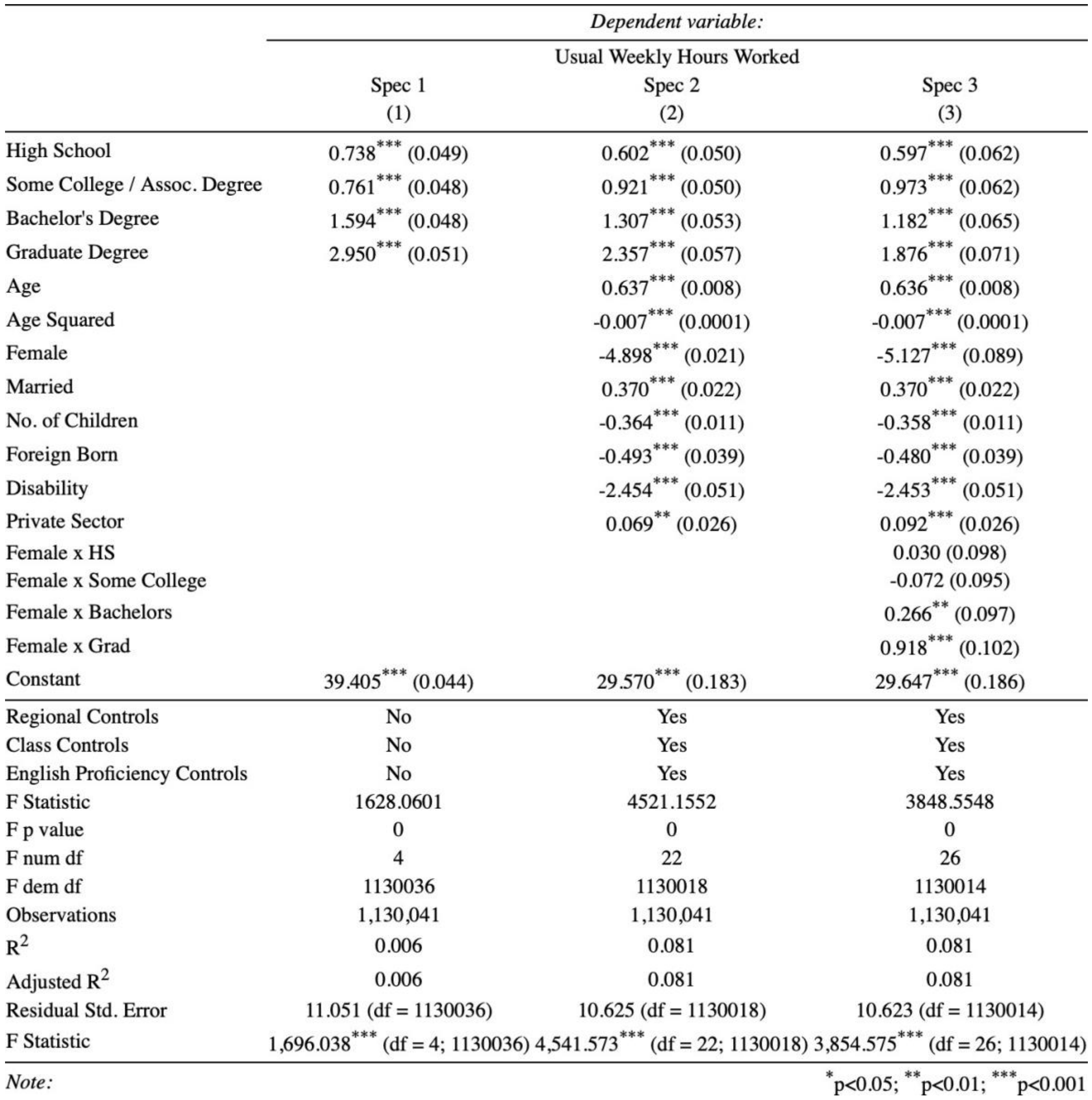


Descriptive Statistics by Education Group

Variable	Below High School	High School	Some University	Bachelors	Grad Degree
DEMOGRAPHIC VARIABLES					
Usual Hours Worked	39.41 (11.53)	40.14 (11.06)	40.17 (10.99)	41.00 (10.81)	42.35 (11.34)
Age	45.81 (10.88)	46.06 (11.64)	44.68 (11.60)	42.73 (11.48)	44.32 (10.75)
Female	0.38 (0.48)	0.42 (0.49)	0.50 (0.50)	0.51 (0.50)	0.54 (0.50)
Married	0.57 (0.49)	0.57 (0.49)	0.60 (0.49)	0.64 (0.48)	0.72 (0.45)
No. Own Children	0.92 (1.28)	0.64 (1.07)	0.71 (1.08)	0.72 (1.05)	0.82 (1.09)
Foreign Born	0.45 (0.50)	0.14 (0.35)	0.12 (0.32)	0.15 (0.36)	0.21 (0.41)
Disability	0.09 (0.29)	0.08 (0.27)	0.07 (0.26)	0.04 (0.19)	0.03 (0.18)
REGIONAL VARIABLES					
Region: Northeast	0.15 (0.35)	0.19 (0.39)	0.15 (0.36)	0.20 (0.40)	0.22 (0.42)
Region: Midwest	0.17 (0.37)	0.25 (0.43)	0.24 (0.43)	0.21 (0.40)	0.18 (0.39)
Region: South	0.40 (0.49)	0.37 (0.48)	0.37 (0.48)	0.35 (0.48)	0.36 (0.48)
Region: West	0.29 (0.45)	0.19 (0.39)	0.24 (0.42)	0.24 (0.43)	0.23 (0.42)
ENGLISH PROFICIENCY VARIABLES					
English: Very Well	0.13 (0.34)	0.09 (0.29)	0.10 (0.30)	0.12 (0.33)	0.17 (0.38)
English: Well	0.14 (0.34)	0.04 (0.21)	0.03 (0.17)	0.03 (0.17)	0.03 (0.16)
English: Not Well	0.16 (0.37)	0.03 (0.16)	0.01 (0.10)	0.01 (0.09)	0.01 (0.07)
English: Not at All	0.07 (0.25)	0.01 (0.08)	0.00 (0.04)	0.00 (0.03)	0.00 (0.02)
OCCUPATIONAL VARIABLES					
Occupation: Professional	0.09 (0.29)	0.16 (0.37)	0.34 (0.47)	0.68 (0.47)	0.88 (0.32)
Occupation: Manual Labour	0.50 (0.50)	0.41 (0.49)	0.23 (0.42)	0.06 (0.24)	0.02 (0.14)
Occupation: Admin/ Sales	0.40 (0.49)	0.43 (0.49)	0.43 (0.49)	0.26 (0.44)	0.09 (0.29)
Occupation: Other	0.00 (0.01)	0.00 (0.04)	0.00 (0.06)	0.00 (0.05)	0.00 (0.06)

Note: The sample is restricted to employed individuals, aged 25-64, who are not living in group or institutional quarters.

Figure 2. Regression Model Specifications



	Dependent variable:		
	Usual Weekly Hours Worked		
	Spec 1	Spec 2	Spec 3
	(1)	(2)	(3)
High School	0.738*** (0.049)	0.602*** (0.050)	0.597*** (0.062)
Some College / Assoc. Degree	0.761*** (0.048)	0.921*** (0.050)	0.973*** (0.062)
Bachelor's Degree	1.594*** (0.048)	1.307*** (0.053)	1.182*** (0.065)
Graduate Degree	2.950*** (0.051)	2.357*** (0.057)	1.876*** (0.071)
Age		0.637*** (0.008)	0.636*** (0.008)
Age Squared		-0.007*** (0.0001)	-0.007*** (0.0001)
Female		-4.898*** (0.021)	-5.127*** (0.089)
Married		0.370*** (0.022)	0.370*** (0.022)
No. of Children		-0.364*** (0.011)	-0.358*** (0.011)
Foreign Born		-0.493*** (0.039)	-0.480*** (0.039)
Disability		-2.454*** (0.051)	-2.453*** (0.051)
Private Sector		0.069** (0.026)	0.092*** (0.026)
Female x HS			0.030 (0.098)
Female x Some College			-0.072 (0.095)
Female x Bachelors			0.266** (0.097)
Female x Grad			0.918*** (0.102)
Constant	39.405*** (0.044)	29.570*** (0.183)	29.647*** (0.186)
Regional Controls	No	Yes	Yes
Class Controls	No	Yes	Yes
English Proficiency Controls	No	Yes	Yes
F Statistic	1628.0601	4521.1552	3848.5548
F p value	0	0	0
F num df	4	22	26
F dem df	1130036	1130018	1130014
Observations	1,130,041	1,130,041	1,130,041
R^2	0.006	0.081	0.081
Adjusted R^2	0.006	0.081	0.081
Residual Std. Error	11.051 (df = 1130036)	10.625 (df = 1130018)	10.623 (df = 1130014)
F Statistic	1,696.038*** (df = 4; 1130036)	4,541.573*** (df = 22; 1130018)	3,854.575*** (df = 26; 1130014)
Note:			*p<0.05; **p<0.01; ***p<0.001

Figure 3. RESET test - squares

```
RESET test

data:  reg2_full

RESET = 182.79, df1 = 1, df2 = 1130017, p-value < 0.00000000000000022
```

Figure 4. RESET test – squares and cubes

```
RESET test

data:  reg2_full
```

RESET = 209, df1 = 2, df2 = 1130016, p-value < 0.00000000000000022

Figure 5. Histogram of Residuals

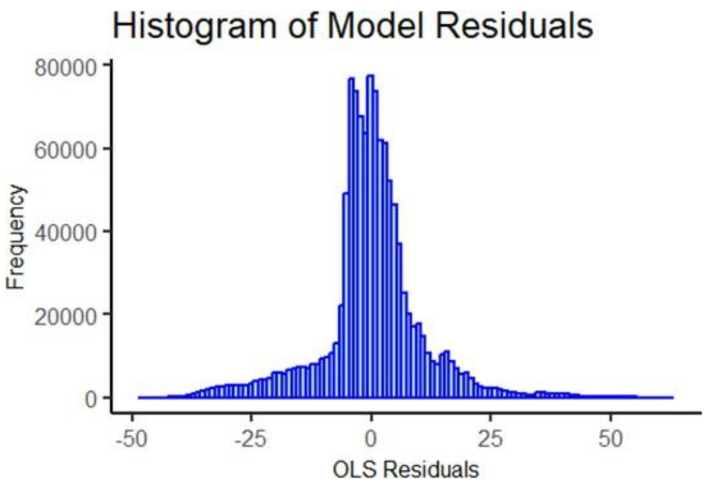


Figure 6. Skewness and Kurtosis

Skewness of residuals: 0.0151

Kurtosis of residuals: 6.188

Figure 7. Jarque Bera Test

data: resids

X-squared = 478570, df = 2, p-value < 0.00000000000000022

Figure 8. Plotting OLS Residuals

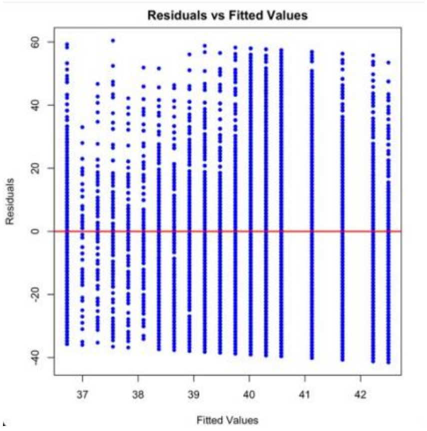
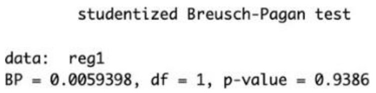


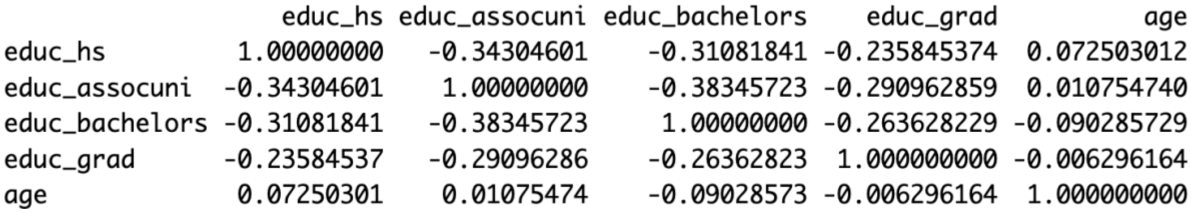
Figure 9. Breusch-Pagan Test



```
	studentized Breusch-Pagan test

data:  reg1
BP = 0.0059398, df = 1, p-value = 0.9386
```

Figure 10. Sample correlations



```
                   educ_hs educ_assocuni educ_bachelors    educ_grad          age
educ_hs         1.00000000   -0.34304601    -0.31081841 -0.235845374  0.072503012
educ_assocuni  -0.34304601    1.00000000    -0.38345723 -0.290962859  0.010754740
educ_bachelors -0.31081841   -0.38345723     1.00000000 -0.263628229 -0.090285729
educ_grad      -0.23584537   -0.29096286    -0.26362823  1.000000000 -0.006296164
age             0.07250301    0.01075474    -0.09028573 -0.006296164  1.000000000
```

Figure 11. Auxiliary Regressions

```
> print(R2_bachelors_age)

[1] 0.008151513

> print(R2_hs)

[1] 0.005256687

> print(R2_assocuni)

[1] 0.0001156644

> print(R2_grad)

[1] 0.00003964168
```

R Code

```
# Research Proposal - Group 4

rm(list=ls())                    # clear memory

#----------------------------------------

# set  working directory

# Method 2 using rStudioapi package

setwd(dirname(rstudioapi::getSourceEditorContext()$path))

#---------------------------------------

options(scipen=999)

library(ggplot2)

library(car)                  # Package for linear hypothesis tests
library(stargazer)
library(sandwich)
library(car)
library(lmtest)
library(rio)
library(fastDummies)
library(margins)
#-------------------------------------
acs <- import("acs_data_2019_NEW.csv")
# check variable listing for categorical variable codes
table(acs$educ)
# create data subset
# Sample restrictions:
# Keep individuals aged 25–64 with valid hours worked
acs_subset1 <- subset(acs, age >= 25 & age <= 64)
acs_subset2 <- subset(acs_subset1, uhours > 0 & uhours <= 98)


#----------------------------- creating required variables for model------------------------------------
# analysing education variable
table(acs_subset2$educ)
# appears that group 3 is the most frequent, should definitely be included in the model

# create dummy variables for education categories
acs_subset2$educ_lths       <- as.numeric(acs_subset2$educ == 1)
acs_subset2$educ_hs         <- as.numeric(acs_subset2$educ == 2)
acs_subset2$educ_assocuni  <- as.numeric(acs_subset2$educ == 3)
acs_subset2$educ_bachelors <- as.numeric(acs_subset2$educ == 4)
```

```
acs_subset2$educ_grad       <- as.numeric(acs_subset2$educ == 5)


# 'age squared' variable

acs_subset2$age_squared <- acs_subset2$age^2




# 'class of worker'

# creating 'private' and 'public' dummies

# intentionally excluding 'self employed' individuals


# Public sector includes: Federal (2), State (3), Local (4)

acs_subset2$public_sector <- as.numeric(acs_subset2$class %in% c(2, 3, 4))


# Private sector includes: Private for-profit (1), Private non-profit (8)

acs_subset2$private_sector <- as.numeric(acs_subset2$class %in% c(1, 8))


# create dummy variables for 'region'

# will likely be used as a control

acs_subset2$region_1 <- as.numeric(acs_subset2$region == 1) # using 1 as reference group

acs_subset2$region_2 <- as.numeric(acs_subset2$region == 2)

acs_subset2$region_3 <- as.numeric(acs_subset2$region == 3)

acs_subset2$region_4 <- as.numeric(acs_subset2$region == 4)




# create dummy variables for 'engprof'

# will also be used as control

acs_subset2$engprof_1 <- as.numeric(acs_subset2$engprof == 1)

acs_subset2$engprof_2 <- as.numeric(acs_subset2$engprof == 2)
```

```
acs_subset2$engprof_3 <- as.numeric(acs_subset2$engprof == 3)

acs_subset2$engprof_4 <- as.numeric(acs_subset2$engprof == 4)

acs_subset2$engprof_9 <- as.numeric(acs_subset2$engprof == 9) # using 9 as reference group


# collapse occupation into groups, categorised by broader qualifications

acs_subset2$occupation_group <- with(acs_subset2, ifelse(occupation %in% 1:4, "Professional",

                                          ifelse(occupation %in% c(5,6,7), "Administrative.Sales",

                                               ifelse(occupation %in% 8:12, "Manual.Labour",

                                                    ifelse(occupation == 13, "Other", NA)))))


acs_subset2$occ_prof         <- as.numeric(acs_subset2$occupation_group == "Professional")

acs_subset2$occ_adminsales   <- as.numeric(acs_subset2$occupation_group ==
"Administrative.Sales")

acs_subset2$occ_manuallab    <- as.numeric(acs_subset2$occupation_group == "Manual.Labour")

acs_subset2$occ_other        <- as.numeric(acs_subset2$occupation_group == "Other")



# ------_______-------_______--------_______-- summary statistics -- _______-------________----
________----_______------_______-#


# STARGAZER method

# defining variables

vars_to_sum <- c("uhours", "age", "female", "married", "nchild18", "forborn", "dis",

            "region_1", "region_2", "region_3", "region_4",

            "engprof_1", "engprof_2", "engprof_3", "engprof_4",

            "private_sector", "public_sector",

            "occ_prof", "occ_manuallab", "occ_adminsales", "occ_other")


labels <- c(
  uhours = "Usual Hours Worked", age = "Age", female = "Female", married = "Married",
  nchild18 = "No. Own Children", forborn = "Foreign Born", dis = "Disability",
  region_1 = "Region: Northeast", region_2 = "Region: Midwest", region_3 = "Region: South",
region_4 = "Region: West",
  engprof_1 = "English: Very Well", engprof_2 = "English: Well", engprof_3 = "English: Not Well",
engprof_4 = "English: Not at All",
  private_sector = "Private Sector", public_sector = "Public Sector",
  occ_prof = "Occupation: Professional", occ_manuallab = "Occupation: Manual Labour",
  occ_adminsales = "Occupation: Admin/ Sales", occ_other = "Occupation: Other"
)


# subsetting by education group
df_list <- list(
  "Below High School" = subset(acs_subset2, educ_lths == 1),
  "High School" = subset(acs_subset2, educ_hs == 1),
  "Some University" = subset(acs_subset2, educ_assocuni == 1),
  "Bachelors" = subset(acs_subset2, educ_bachelors == 1),
  "Grad Degree" = subset(acs_subset2, educ_grad == 1)
)


# computing summaries of mean and SD for each group
summary_combined <- sapply(df_list, function(df) {
sapply(vars_to_sum, function(v) {
    m <- mean(df[[v]], na.rm = TRUE)
    s <- sd(df[[v]], na.rm = TRUE)
    sprintf("%.2f (%.2f)", m, s)
  })
})
```

```
# converting to data frame and adding variable names
summary_df <- as.data.frame(summary_combined)
summary_df$Variable <- labels[vars_to_sum]
summary_df <- summary_df[, c("Variable", names(df_list))]  # Reorder columns


# adding section headers
section_rows <- data.frame(matrix(NA, nrow = 4, ncol = ncol(summary_df)))
colnames(section_rows) <- colnames(summary_df)
section_rows$Variable <- c("DEMOGRAPHIC VARIABLES",
                           "REGIONAL VARIABLES",
                           "ENGLISH PROFICIENCY VARIABLES",
                           "OCCUPATIONAL VARIABLES")


# indexing sections and splitting existing table
demographic_vars <- 1:7
region_vars      <- 8:11
engprof_vars     <- 12:15
occupation_vars  <- 18:21

demographics   <- summary_df[demographic_vars, ]
regionals      <- summary_df[region_vars, ]
eng_profs      <- summary_df[engprof_vars, ]
occupations    <- summary_df[occupation_vars, ]

# reassemble with section headers inserted
summary_sectioned <- rbind(
  section_rows[1, ], demographics,

  section_rows[2, ], regionals,

  section_rows[3, ], eng_profs,

  section_rows[4, ], occupations

)


#export as HTML

stargazer(summary_sectioned, type = "html", summary = FALSE, rownames = FALSE,

          title = "Descriptive Statistics by Education Group",

          out = "summary_by_educ_with_sections.html",

          notes = c("<strong>Note:</strong> The sample is restricted to employed individuals, aged 25-64,",

                    "who are not living in group or institutional quarters."))




#--------___________---------__________---------________Regressions_____--------________---------_________------__________-----.________

library(lmtest)

library(sandwich)

library(stargazer)


# Spec 1: Basis Model

reg1_updated <- lm(uhours ~ educ_hs + educ_assocuni + educ_bachelors + educ_grad, data = acs_subset2)


# Robust standard errors

cov1_updated <- vcovHC(reg1_updated, type = "HC1")

reg1_robust_se_updated <- sqrt(diag(cov1_updated))
```

```
# Wald F-test
wald_reg1_updated <- waldtest(reg1_updated, vcov = cov1_updated)

# Extract key F-statistics
fstat1_updated <- round(wald_reg1_updated$"F"[2], 4)
pvalf1_updated <- round(wald_reg1_updated$"Pr(>F)"[2], 4)
numdf1_updated <- abs(wald_reg1_updated$"Df"[2])
demdf1_updated <- df.residual(reg1_updated)

# Stargazer output with labels and robust SEs
stargazer(reg1_updated,
      type = "html",
      se = list(reg1_robust_se_updated),
      out = "regression_table_educ_dummies.html",
      title = "Regression of Hours Worked on Education (Category Dummies)",
      dep.var.labels = "Usual Weekly Hours Worked",
      covariate.labels = c("High School",
                  "Some College / Assoc. Degree",
                  "Bachelor's Degree",
                  "Graduate Degree"),
      digits = 3,
      star.cutoffs = c(0.05, 0.01, 0.001),
      single.row = TRUE)

#------------------------------------------------------------------------------
#Spec 2: including other variables
reg2_full <- lm(uhours ~
         educ_hs + educ_assocuni + educ_bachelors + educ_grad +
         age + age_squared + female + married + nchild18 + forborn + dis +
         private_sector +
         region_2 + region_3 + region_4 +
         engprof_1 + engprof_2 + engprof_3 + engprof_4 +
         occ_adminsales + occ_manuallab + occ_other,
        data = acs_subset2)

# Robust standard errors
cov2_full <- vcovHC(reg2_full, type = "HC1")
reg2_robust_se <- sqrt(diag(cov2_full))

# Wald F-test
wald_reg2 <- waldtest(reg2_full, vcov = cov2_full)

# Extract test statistics
fstat2 <- round(wald_reg2$"F"[2], 4)
pvalf2 <- round(wald_reg2$"Pr(>F)"[2], 4)
numdf2 <- abs(wald_reg2$"Df"[2])
demdf2 <- df.residual(reg2_full)

# Omit dummy controls & label inclusion
omit_controls <- c("region_2", "region_3", "region_4",
                   "engprof_1", "engprof_2", "engprof_3", "engprof_4",
                   "occ_adminsales", "occ_manuallab", "occ_other")

add_lines2 <- list(
  c("Regional Controls", "Yes"),
  c("Class Controls", "Yes"),
  c("English Proficiency Controls", "Yes"),
  c("F Statistic", fstat2),
  c("F p value", pvalf2),
  c("F num df", numdf2),
  c("F dem df", demdf2)
)

# Stargazer Output
stargazer(reg2_full,
          type = "html",
          se = list(reg2_robust_se),
          out = "regression_table_spec2.html",
          title = "Regression of Hours Worked on Education and Controls",
          dep.var.labels = "Usual Weekly Hours Worked",
          covariate.labels = c("High School",
                               "Some College / Assoc. Degree",
                               "Bachelor's Degree",
                               "Graduate Degree",
                               "Age", "Age Squared",
                               "Female", "Married", "No. of Children",
                               "Foreign Born", "Disability", "Private Sector"),
          omit = omit_controls,
          add.lines = add_lines2,
          digits = 3,
          star.cutoffs = c(0.05, 0.01, 0.001),
          single.row = TRUE)
```

```
#-------------------------------------------------------------------------------------------------------------------
--

# Spec 3: with interaction between education and female variable

library(lmtest)
library(sandwich)
library(stargazer)

# create interaction terms
acs_subset2$int_female_hs         <- acs_subset2$female * acs_subset2$educ_hs
acs_subset2$int_female_assocuni  <- acs_subset2$female * acs_subset2$educ_assocuni
acs_subset2$int_female_bachelors <- acs_subset2$female * acs_subset2$educ_bachelors
acs_subset2$int_female_grad       <- acs_subset2$female * acs_subset2$educ_grad

# Run regression with female interactions
reg3_femaleonly <- lm(uhours ~
                        educ_hs + educ_assocuni + educ_bachelors + educ_grad +
                        female + married +
                        age + age_squared + nchild18 + forborn + dis + private_sector +
                        region_2 + region_3 + region_4 +
                        engprof_1 + engprof_2 + engprof_3 + engprof_4 +
                        occ_adminsales + occ_manuallab + occ_other +
                        int_female_hs + int_female_assocuni + int_female_bachelors + int_female_grad,
                      data = acs_subset2)

# Robust standard errors
cov3_femaleonly <- vcovHC(reg3_femaleonly, type = "HC1")
```

```
reg3_robust_se <- sqrt(diag(cov3_femaleonly))

# Wald F-test
wald_reg3 <- waldtest(reg3_femaleonly, vcov = cov3_femaleonly)
fstat3 <- round(wald_reg3$"F"[2], 4)
pvalf3 <- round(wald_reg3$"Pr(>F)"[2], 4)
numdf3 <- abs(wald_reg3$"Df"[2])
demdf3 <- df.residual(reg3_femaleonly)

# Omitted control variables
omit_controls_3 <- c("region_2", "region_3", "region_4",
                     "engprof_1", "engprof_2", "engprof_3", "engprof_4",
                     "occ_adminsales", "occ_manuallab", "occ_other")

# Control flags and F-test info
add_lines_3 <- list(
  c("Regional Controls", "Yes"),
  c("Class Controls", "Yes"),
  c("English Proficiency Controls", "Yes"),
  c("F Statistic", fstat3),
  c("F p value", pvalf3),
  c("F num df", numdf3),
  c("F dem df", demdf3)
)

# Stargazer output
stargazer(reg3_femaleonly,
          type = "html",
          se = list(reg3_robust_se),
          out = "regression_table_spec3_femaleonly.html",
          title = "Regression with Female × Education Interactions",
          dep.var.labels = "Usual Weekly Hours Worked",
          covariate.labels = c("High School",
                          "Some College / Assoc. Degree",
                          "Bachelor's Degree",
                          "Graduate Degree",
                          "Female", "Married",
                          "Age", "Age Squared", "No. of Children",
                          "Foreign Born", "Disability", "Private Sector",
                          "Female x HS", "Female x Some College", "Female x Bachelors", "Female x Grad"),
          omit = omit_controls_3,
          add.lines = add_lines_3,
          digits = 3,
          star.cutoffs = c(0.05, 0.01, 0.001),
          single.row = TRUE)


#--------------------------------------------------------------------------------------------------------------------------------------------

# All 3 Specs together
# Stargazer Combined Output


stargazer(reg1_updated, reg2_full, reg3_femaleonly,
          type = "html",
          se = list(reg1_robust_se_updated, reg2_robust_se, reg3_robust_se),
          out = "regression_combined_spec1_2_3.html",
          title = "Regression of Usual Weekly Hours Worked - Specs 1 to 3",
          dep.var.labels = "Usual Weekly Hours Worked",
          column.labels = c("Robust (White)", "Robust (White)", "Robust (White)"),
          covariate.labels = c("High School",
                          "Some College / Assoc. Degree",
                          "Bachelor's Degree",
                          "Graduate Degree",
                          "Age", "Age Squared",
                          "Female", "Married", "No. of Children",
                          "Foreign Born", "Disability", "Private Sector",
                          "Female x HS", "Female x Some College",
                          "Female x Bachelors", "Female x Grad"),
          omit = omit_controls_3,
          add.lines = list(
            c("Regional Controls", "No", "Yes", "Yes"),
            c("Class Controls", "No", "Yes", "Yes"),
            c("English Proficiency Controls", "No", "Yes", "Yes"),
            c("F Statistic", fstat1_updated, fstat2, fstat3),
            c("F p value", pvalf1_updated, pvalf2, pvalf3),
            c("F num df", numdf1_updated, numdf2, numdf3),
            c("F dem df", demdf1_updated, demdf2, demdf3)
          ),
          digits = 3,
          star.cutoffs = c(0.05, 0.01, 0.001),
          single.row = TRUE)


#------------------------------------------------------------------------


#ROBUSTNESS CHECKS
```

```
#------------------------------------

# test for multicollinearity

#------------------------------------


# assessing education dummy variables and age

expv_educ_age <- subset(acs_subset2, select = c(educ_hs, educ_assocuni, educ_bachelors,
educ_grad, age))


# create correlation matrix

educ_age_cormat <- cor(expv_educ_age, use = "complete.obs")

print(educ_age_cormat)


# auxiliary regressions for each education dummy on age

xreg_bachelors_age <- lm(educ_bachelors ~ age, data = acs_subset2)

xreg_hs <- lm(educ_hs ~ age, data = acs_subset2)

xreg_assocuni <- lm(educ_assocuni ~ age, data = acs_subset2)

xreg_grad <- lm(educ_grad ~ age, data = acs_subset2)


R2_bachelors_age <- summary(xreg_bachelors_age)$r.squared

R2_hs <- summary(xreg_hs)$r.squared

R2_assocuni <- summary(xreg_assocuni)$r.squared

R2_grad <- summary(xreg_grad)$r.squared


print(paste("R-squared for educ_bachelors ~ age:", round(R2_bachelors_age, 4)))

print(paste("R2 educ_hs ~ age:", round(R2_hs, 4)))

print(paste("R2 educ_assocuni ~ age:", round(R2_assocuni, 4)))

print(paste("R2 educ_grad ~ age:", round(R2_grad, 4)))

#--------------------------------------

# functional form and normality
```

```
#--------------------------------------

# Load required package
if (!require(moments)) install.packages("moments")
library(moments)
library(lmtest)
library(sandwich)
library(ggplot2)

# Get residuals from the model
resids <- reg2_full$residuals

#----------------------
# RESET Test

# RESET test for functional form
resettest(reg2_full, power = 2, type = "fitted")       # squared fitted values
resettest(reg2_full, power = 2:3, type = "fitted")     # squared + cubed fitted values

#----------------------
# Jarque-Bera Test

if (!require(tseries)) install.packages("tseries")
library(tseries)

jarque.bera.test(resids)

#----------------------
```

```
#OLS Residuals Plot
#----------------------
# Open window
dev.new()

# set margins
par(mar = c(4, 4, 2, 2))

# plot OLS Residuals
plot(fitted(reg1), reg1$residuals,
    xlab = "Fitted Values",
    ylab = "Residuals",
    main = "Residuals vs Fitted Values",
    pch = 20, col = "blue")

# Add horizontal line
abline(h = 0, col = "red", lwd = 2)
#---------------------
#Test for Heteroskedasticity
#---------------------
#Breusch-Pagan Test
# Model 2
bptest(reg1)
#----------------------
# Skewness and Kurtosis

skew_val <- skewness(resids)
kurt_val <- kurtosis(resids)
```

```
cat("Skewness of residuals:", round(skew_val, 4), "\n")
cat("Kurtosis of residuals:", round(kurt_val, 4), "\n")


#-----------------------
# Histogram of Residuals


ggplot(data = acs_subset2, aes(x = resids)) +
  geom_histogram(breaks = seq(min(resids, na.rm = TRUE),
                              max(resids, na.rm = TRUE), by = 1),
                 fill = 'lightblue', col = "blue") +
  labs(x = "OLS Residuals", y = "Frequency", title = "Histogram of Model Residuals") +
  theme_classic() +
  theme(axis.text = element_text(size = 8), axis.title = element_text(size = 8))


# Save the plot to file
ggsave("histogram_residuals_acs.pdf")
```