Lab 2: Description Using Models

How Many Hours do You Want to Work?

Datasci 203 Team: Yi, Wang, Wu, Yarramreddy

2024-12-13

This report aims to use linear regression modeling to describe how one's occupation, gender, age, household type with or without children under 18, and educational level impacts hours worked per week. The analysis uses a 2023 United States Census Bureau survey from California. Results show that these factors can help describe the amount of hours you work a week.

1 Paper

1.1 Introduction

For many people, one of the most important factors in choosing a career is work-life balance. According to the human resources software provider Cipher, 67% of respondents ranked work-life balance as a top consideration when selecting a job, closely followed by pay, benefits, and job security. A major aspect of work-life balance is the number of hours spent working each week. Some careers demand long hours and overtime, while others offer more predictable 40-hour weeks, or even fewer. This raises an important question:

How does the occupational field impact the number of working hours per week?

Answering this question could provide valuable insights for high school graduates and college students embarking on their careers, helping them choose paths that align with their desired work-life balance. Additionally, it could guide those already in the workforce who are considering a career change, and assist employers in understanding how to improve their employees' work-life balance.

1.2 Description of Data Source

The dataset being used is the Public Use Microdata Sample from the American Community Survey (ACS). The ACS Public Use Microdata Sample (PUMS) is collected by the United State Census Bureau. The population is the United State of America with approximately 1% of the population being sampled. The sampling technique used is stratified sampling where each geographic region is designed to have roughly the same population. The participants will be given the American Community Survey (ACS). Each row in the dataset represents a single person. All ACS responses are confidential and many variables have been added/modified to protect the anonymity and confidentiality of the participants. We chose to analyze survey response from 2023 California as we assumed that California would represent most, if not all of the occupations and industries.

1.3 Data Wrangling

For data wrangling, we used the dplyr package to transform the raw data to data for analysis. We had two raw data files: psam_p_CA.csv (person-level data) and psam_h_CA.csv (househoad-level data). The datasets were joined using an inner join on the SERIALNO column where only matched records were kept. From the merged data, columns WKHP, OCCP, SCHL, HHT2, AGEP, SEX are kept to use as independent and dependent variables for our regression models and renamed to more human-readable names while the rows with missing values in any of these columns were removed to ensure data completeness and avoid bias in the analysis. The cleaned dataset was exported to cleaned_data.csv for analysis. Figure 2.1 shows a histogram of the y independent variable: Hours Worked Weekly. Our sample data had initially 191K rows of data. After filtering some of the variables as we will mention later on, we ended up with 183,391 rows of data that we used for our regression modeling.

1.4 Operationalization

Hour worked weekly is a metric variable that ranges from 1 to 99. The histogram for this variable shows no significant skewness, making it a suitable dependent variable for modeling.

Occupation was converted from a numeric variable (ranging from 0020 to 9920) into a categorical variable to make occupation categories more interpretable. The categories include: "MRG" for Management, "BUS" for Business and Accounting, "FIN" for Finance, "CMM" for Computer, "ENG" for Engineering, "SCI" for Science, "CMS" for Social Services, "LGL" for Legal, "EDU" for Education, "ENT" for Entertainment, "MED" for Medical, "HLS" for Health and Social Services, "PRT" for Public Safety, "EAT" for Food Services, "CLN" for Facilities, "PRS" for Personal Care, "SAL" for Retail and Sales, "OFF" for Office, "FFF" for Agriculture, "CON" for Construction, "EXT" for Extraction and Mining, "RPR" for Repair, "PRD" for Food or Material Production, "TRN" for Transportation, and "MIL" for Military.

Education level was converted from an ordinal variable (ranging from 1 to 24) into a categorical variable to make education levels more intuitive. The categories include: "HS" for High School or below, "AS" for Associate's or Some College, "BD" for Bachelor's degree, "MD" for Master's or other advanced degrees past a Bachelor's, and "PHD" for Doctorate.

Household was converted from a numeric variable into a categorical variable to represent household types. The categories include: "CWC" for Cohabiting or Married with Children under 18, "CWOC" for Cohabiting or Married without Children under 18, "SWOC" for Single without Children under 18, and "SWC" for Single with Children under 18.

Gender is a categorical variable with two categories: male and female.

Age is a continuous variable, ranging from 18 to 67 years.

1.5 Model Specification

For our first model, we explore a single-variable linear regression model. We set our independent variable as occupation (treated as a categorical variable), and our dependent variable as hours worked per week (treated as a continuous variable). We assumed that one's occupation would best describe the amount of hours worked per week.

For our second model, we explore a multi-variable linear regression model. We set our independent variables as occupation (categorical), education (categorical), household type (categorical), gender (categorical), and age (continuous), and our dependent variable as hours worked per week (continuous). We chose education because we assumed that it would be a good indicator to the amount of hours worked. People with college degrees may at higher rates work in occupations that are set 40 hours a week salary jobs while those without degrees may more frequently work hourly jobs and end up with more or les than 40 hours. Household type was important to capture how having children under 18 may affect the amount of hours worked and if you were a single parent or not mattered. Gender was included as some of the household categories like cohabiting or married did not indicate the gender of the respondent. Lastly, we felt that including age was important as people in different stages of their lives may work different hours.

1.6 Model Assumptions

Model 1: The IID assumption is satisfied as each observation represents an individual or household sampled independently from the population using the PUMS's random sampling methodology, and all observations are drawn from the same population distribution for the survey year. Since there is only one independent variable, collinearity is not a concern. The Residuals vs Occupation plot (Figure 2.3, Plot E) shows no clear pattern, indicating a linear relationship between occupation and weekly work hours. The Residuals vs Fitted Values plot shows no funnel-shaped patterns, meaning the variance of the residuals is relatively constant across different levels of the fitted values indicating homoscedasticity (Figure 2.3, Plot A). Although the Q-Q plot is left-skewed, because we have a large sample size, the normal distribution of errors is still satisfied due to the Central Limit Theorem (Figure 2.3, Plot C).

Model 2: The IID assumption is satisfied for the same reasoning as Model 1. We used the variance inflation factor (Section 2.2), which measures the strength of correlation between predictor variables, to detect multicollinearity in this model. The generalized variance inflation factor for all predictors are well below the commonly accepted threshold of 5, indicating no significant multicollinearity issues. The Residuals vs Occupation plot shows no clear patterns, indicating a linear relationship between occupation and weekly work hours (Figure 2.3, Plot F). The Residuals vs Education Level plot shows no clear patterns, indicating a linear relationship between education level and weekly work hours (Figure 2.3, Plot I). The Residuals vs Household plot shows no clear patterns, indicating a linear relationship between household type and weekly work hours (Figure 2.3, Plot H). The Residuals vs Age plot shows no clear patterns, indicating a linear relationship between age and weekly work hours (Figure 2.3, Plot G). The Residuals vs Gender plot shows no clear patterns, indicating a linear relationship between gender and weekly work hours (Figure 2.3, Plot J). The Residuals vs Fitted Values plot (Figure 2.3, Plot B) shows no funnel-shaped patterns, meaning the variance of the residuals is relatively constant across different levels of the fitted values, indicating homoscedasticity. We see some deviation from the qqline (Figure 2.3, Plot D) likely due to a large amount of outliers in our data (not working/inflation of working hours). We do however have a very large sample size so using the Central Limit Theorem we cautiously satisfy this assumption.

1.7 Model Results and Interpretation

Our modeling results show that while we have P Value less than .05 showing statistical significance, our adjusted R-Squared Value of .07 for Model 1 and .096 for Model 2 means that we really don't have any practical significance. Our model using educational levels, occupational industry, household type, and age only describes 9% of the variability in the data, which is low. While the relationship is statistically significant, it may not be practically meaningful. However, it can still provide us with some meaningful insights.

In Model2, the coefficient is 37.2 hours, and the dummy variables (indicator variables) is a male, cohabiting or married with children under 18, with an associate degree or some amount of college, and in the Business and Accounting occupational group.

In a hypothetical example going off our baseline model, you will work 1.5 more hours if you have a Bachelor's, .5 hours less if you are without child and still married or cohabiting, and 1.5 more hours if you are working in Public Safety.

Our model suggests that most occupation do tend to work around 40 hours a week with certain exceptions. Our model also tells us that males tend to work more hours than females in general. Also that you tend to work less hours without a child and more hours with a child for both men and women.

Our model also describes that each level of educational milestone also leads to an increase of an hour worked a week at each step. Next steps in our model refinement could be re-examining how we group different occupations, how age and different stages in life affects how many hours you work, and include additional variables like work/life balance satisfaction metrics. We can also look to remove weekly hours worked value from respondents are working limited hours due to health reasons.

2 Appendix

1. **A Link to your Data Source.** https://data.census.gov/mdat/#/search?ds=ACSPUMS1Y2023
2. **A List of Model Specifications you Tried.** Initial Attempts: We began with a single-variable linear regression using educational level as the predictor. However, this yielded a very low adjusted R-squared value, leading us to switch to occupation as the main variable.

Granular Grouping of Occupations: Initially, we grouped occupations at a more granular level, but this approach did not significantly improve the model's adjusted R-squared value. It also posed challenges by reducing the sample size and risking a narrower scope for our research question.

Incorporating Industries, Household Types, Gender, and age: We explored various combinations of occupation, household type, and industry variables under the assumption that job type and household composition significantly influence hours worked. We tested different approaches for including or excluding gender and experimented with categorizing household types into various groups.

Variable Transformation: As our Y variable (weekly hours worked) had very long tails with values ranging from 0 to 99, we experimented with taking the natural log. Taking the natural log grouped the values too closely so we decided not to transform the y variable.

Final results: Our best results were obtained when we: Retained gender as a variable. Grouped household types into four categories: single (with or without child) and cohabiting (with or without child). Restricted the age range to typical working ages (18–67) and excluded respondents who reported not being in the workforce.

Results and Insights These adjustments increased the model's adjusted R-squared value from approximately 0.08 to 0.096. Although we considered removing gender, it was retained because some household variables lacked sufficient gender-specific information.

2.1 Model Summaries

```
=========================================================================
                                     Dependent variable:
```

```
                    ---------------------------------------------------------
                                       hours_worked_weekly
                                (1)                          (2)
-----------------------------------------------------------------------------
occupationCLN                -5.349***                    -4.685***
                              (0.205)                      (0.207)

occupationCMM                 0.718***                     -0.365**
                              (0.184)                      (0.183)

occupationCMS                -2.865***                    -2.705***
                              (0.250)                      (0.247)

occupationCON                -1.280***                    -1.538***
                              (0.192)                      (0.197)

occupationEAT                -9.344***                    -7.602***
                              (0.183)                      (0.186)

occupationEDU                -4.796***                    -4.783***
                              (0.175)                      (0.173)

occupationENG                 1.486***                      0.004
                              (0.211)                      (0.210)

occupationENT                -3.266***                    -3.293***
                              (0.203)                      (0.200)

occupationEXT                 3.922***                     3.955***
                              (1.456)                      (1.436)

occupationFFF                  -0.249                      0.579**
                              (0.271)                      (0.271)

occupationFIN                 0.971***                     0.512**
                              (0.230)                      (0.227)

occupationHLS                -5.133***                    -3.435***
                              (0.198)                      (0.198)

occupationLGL                 2.476***                     1.405***
                              (0.278)                      (0.277)
```

```
occupationMED                -1.564***                  -1.439***
                              (0.178)                    (0.176)

occupationMIL                 5.187***                   5.383***
                              (0.544)                    (0.537)

occupationMRG                 2.592***                   2.044***
                              (0.158)                    (0.156)

occupationOFF                -3.397***                  -2.045***
                              (0.162)                    (0.162)

occupationPRD                 -0.432**                    -0.013
                              (0.191)                    (0.193)

occupationPRS                -9.289***                  -7.507***
                              (0.216)                    (0.216)

occupationPRT                 1.577***                   1.512***
                              (0.242)                    (0.241)

occupationRPR                  0.408*                     0.166
                              (0.224)                    (0.226)

occupationSAL                -3.812***                  -2.948***
                              (0.166)                    (0.166)

occupationSCI                -0.781***                  -1.711***
                              (0.255)                    (0.257)

occupationTRN                -1.918***                  -1.467***
                              (0.170)                    (0.174)

education_levelBD                                        1.512***
                                                         (0.114)

education_levelHS                                       -0.282***
                                                         (0.107)

education_levelMD                                        2.616***
                                                         (0.126)

education_levelPHD                                       3.565***
```

```
                                                           (0.209)

household_typeCWOC                                        -0.476***
                                                           (0.066)

household_typeSWC                                           -0.083
                                                           (0.139)

household_typeSWOC                                        -0.394***
                                                           (0.073)

age                                                        0.085***
                                                           (0.002)

genderFemale                                              -2.883***
                                                           (0.060)

Constant                     40.093***                    37.226***
                              (0.136)                      (0.196)

-----------------------------------------------------------------------------
Observations                  183,391                      183,391
R2                             0.070                        0.096
Adjusted R2                    0.070                        0.096
Residual Std. Error     11.775 (df = 183366)         11.606 (df = 183357)
F Statistic         574.058*** (df = 24; 183366) 593.295*** (df = 33; 183357)
=============================================================================
Note:                                             *p<0.1; **p<0.05; ***p<0.01
```

2.2 Variance Inflation Factor

```
                    GVIF Df GVIF^(1/(2*Df))
occupation      2.132511 24        1.015902
education_level 1.729289  4        1.070862
household_type  1.047447  3        1.007756
age             1.061302  1        1.030195
gender          1.222551  1        1.105690
```

2.3 Figures

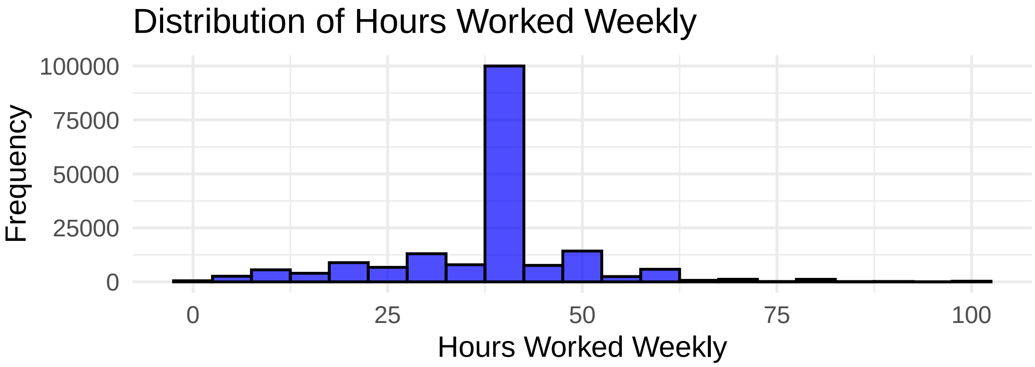

Figure 2.1: Large Sample Model showing normal distribution clustered at 40 hours.

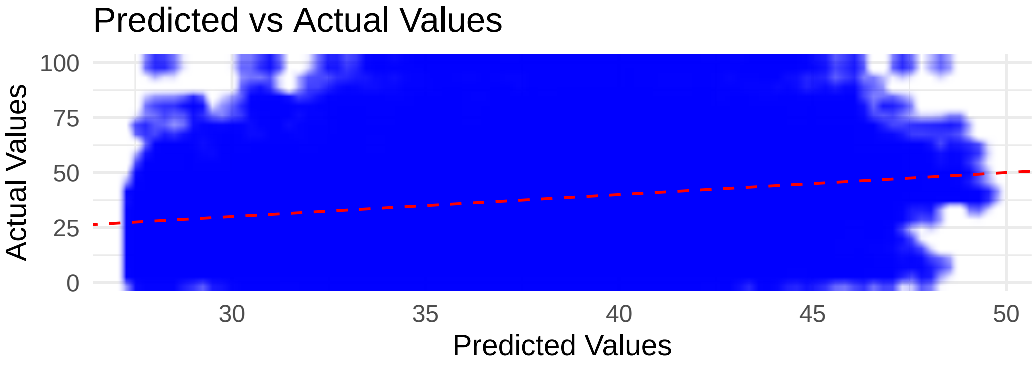

Figure 2.2: Actual vs Predicted for Model2.

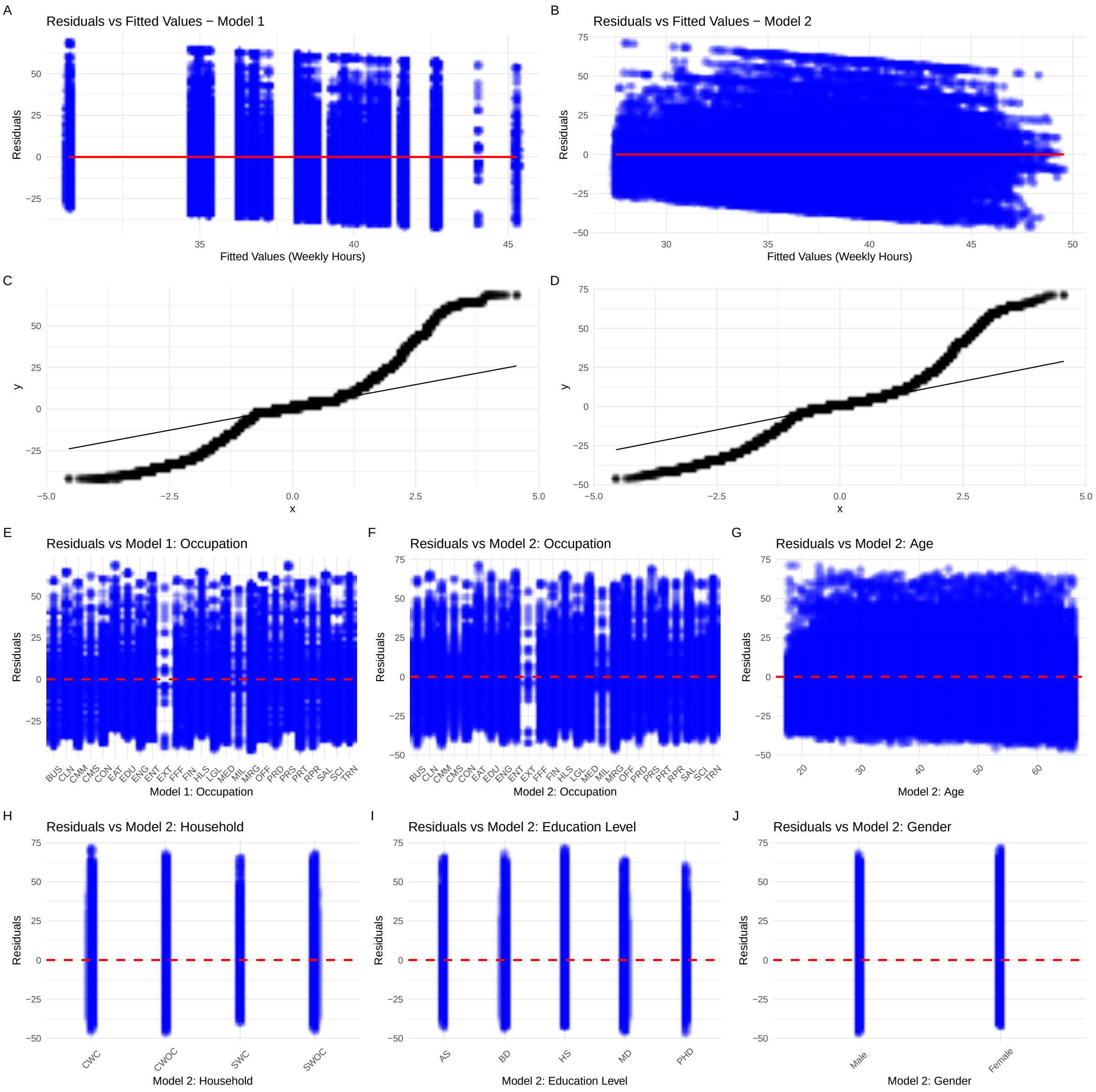

Figure 2.3: Graphs of Assumptions