Lab 2 Report

Sebastian Rosales, Adwith Malpe, Kent Bourgoing, Jonathan Hsiao

2024-07-20

Introduction

The US Census Bureau published the data for the 2023 Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS). The CPS ASEC 2023 dataset proves a comprehensive view of the US economic and social landscape that is imperative for informed policy-making. This study thoroughly entails monthly labor force statistics, work experience, income, noncash benefits, health insurance, and migration patterns. This topic is of interest as it considers economic well-being in understanding income distribution and employment trends for addressing economic disparities, social equity in identifying the most vulnerable populations, guiding social welfare programs to assist those in need, and public health where health insurance coverage statistics are vital for assessing healthcare accessibility and shaping health policies to ensure comprehensive coverage for citizens. Given the information regarding income and labor force statistics, a large interest lies in how citizens can improve their earnings based on their work ethic and results in the following question at hand.

The **research question** in observance states is there a statistically significant relationship between the number of hours worked per week and the salary earned among the US civilian noninstitutionalized population?

Description of the Data Source

The data was collected from the 2023 Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS). It was jointly sponsored by the US Census Bureau and the US Bureau of Labor Statistics and conducted over three months: February, March, and April. Multistage probability sampling was conducted based on the decennial census results and a sample covers all 50 states and the District of Columbia. There were approximately 89,000 addresses sampled, where 78,000 were eligible for interviews and 57,000 interviews were conducted. The units of observation include individuals, families, and households. Additionally, the civilian noninstitutionalized population of the US that includes Armed Forces members are units as well. Employment and unemployment data reflects the work experience, employment status, occupation, and any industry data for people above the age of 15 years. Income and Noncash benefits reflect total income, income components, non-cash income (i.e. food stamps, health insurance), health insurance coverage that reflects health insurance status and coverage types, demographic characteristics that reflect age, sex, household relationships; geographic mobility and migration that reflects data on household and family characteristics, and geographic coverage that reflects the area including states, regions, divisions, counties, and principal cities. Given the provided data, the analysis of how salary relates to time worked can be analyzed and potential relationships may be identified.

Operationalization

To describe the relationship between the number of hours worked and one's salary, we first need to define and understand the key variables in our dataset. The variable WSAL_VAL represents an individual's total wage and salary earnings, while the variable HRSWK indicates the typical number of hours worked per week. Both variables are important for analyzing the relationship between salary and work hours. Our dataset comprises of 146,133 observations. However, we excluded individuals who reported a "0" for their salary, as this may indicated that the individual was either unemployment in 2023 or a refused to answer the question. This exclusion left us with 69,148 observations after removing 76,985 responses. Distribution plots (see Figure 3 in Appendix) were generated to display the data range for both variables. The weekly hours worked by our respondents vary from 1 to 99. The majority, over 40,000 respondents, reported working between 30 and 40 hours per week. However, more than 10,000 respondents indicated they work more than 40 hours per week. Regarding salary, the range spans from $1 to $1,549,999 annually.

Model Specification

To comprehensively understand the relationship between total earnings (Y) and hours worked per week (X), we employ two regression models:

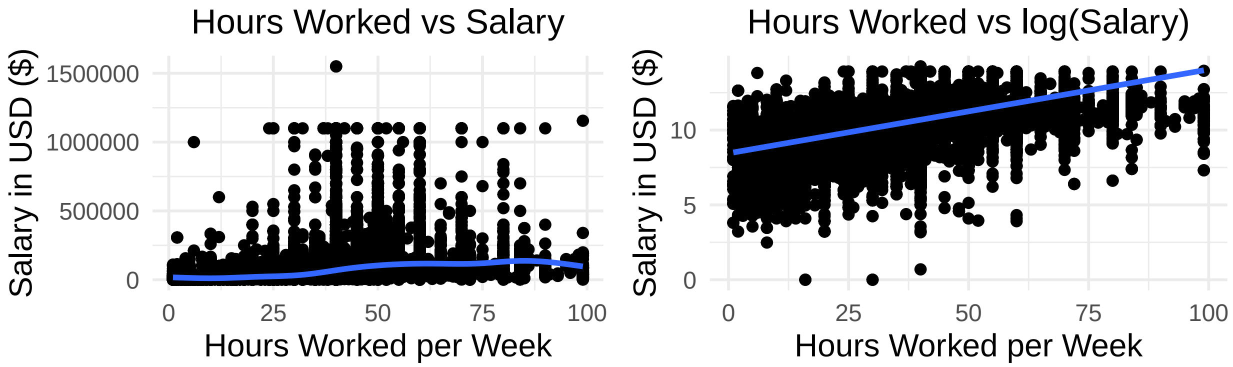

Figure 1: Scatter plot of Hours Worked Per Week vs Salary and Hours Worked Per Week vs log of Salary

1. A basic linear regression model with untransformed variables:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

2. A linear regression model with the log transformation of Y and the addition of the squared term of X

$$log_{10}(Y) = \beta_0 + \beta_1 X + \epsilon$$

The transformations we selected are based on the relationships we observed in the data. Since the distribution of total earnings exhibits a heavy right tail, we chose to apply a log transformation to reduce skewness and enable a more linear relationship.

Table 1: Linear Regression Model: Total Wage and Salary Earnings

	Dependent variable:	
	Total Wage and Salary Earnings	Log of Total Wage and Salary Earnings
	(1)	(2)
Hours Worked Per Week	2,231.589***	0.024***
	(33.404)	(0.0002)
Intercept	−20,251.660***	3.668***
	(1,181.719)	(0.009)
Observations	69,148	69,148
R^2	0.083	0.307
Adjusted R^2	0.083	0.307
Residual Std. Error (df = 69146)	81,604.940	0.402
F Statistic (df = 1; 69146)	6,249.320***	30,601.390***
Note:		*p<0.1; **p<0.05; ***p<0.01

Model Assumptions

In order for our models to be valid, the large-sample linear model assumptions must be true: data must be identical and independently distributed (IID), and a unique Best Linear Predictor (BLP) must exist.

First, we'll evaluate the IID assumption. Given that our data was collected by the US Census it is possible that multiple entries might be related. For example, there could be multiple respondents who are related to each other, might live in the same community, might work under the same company, etc. However, since the

US Census collects data from US residents all over the country, it is highly unlikely that with a sample size of 69,148, a significant level of dependency will be observed. Additionally, since all of the data is collected from the US resident population it should be identically distributed.

Next, we'll evaluate whether a unique BLP exists. From Figure 3 in the Appendix, we can see that the distribution of salary has a right-tailed skewness. This makes intuitive sense given that wealth is not evenly distributed and there are higher probabilities that we will see salaries that are much larger than the mean. However, the heavy-tail might indicate non-finite variance, which means the BLP may not exist. As for uniqueness, both models will have no problems with colinearity given that hours worked per week is our only independent variable.

Model Results and Interpretation

The regression model with untransformed variables has a coefficient of 2,231.589 for the Hours Worked Per Week variable, which is statistically significant with a p-value < 0.01. This means that for every additional hour worked, the model expects an increase in Total Earnings of about $2,232. The intercept of this model is -20,251.66, which suggests that someone who had zero Hours Worked would have negative Total Earnings. This intercept doesn't make much sense in the real world.

The transformed regression model displayed an Hours Worked Per Week coefficient of 0.024 and a constant of 3.668. Both coefficients were found to be statistically significant at an alpha level of 0.01, as demonstrated by a p-value lower than 0.01. With this model, the intercept is more interpretable as it is a positive number. However, might not apply perfectly to the real world since it doesn't make sense to make money without working a single hour. However, the model does provide us with slightly more realistic numbers when considering working a single hour or higher. For instance, if you work a single hour per week you can expect a wage of 4,920.40$, ($10^{(0.024+3.668)}$). Additionally, for every hour you work, you can expect an increase in wage of around 5.6%, ($10^{.024}$).

Because the dependent variables in the two models have different scales (logged vs unlogged), we can't directly compare R-squared values. Because the two models aren't nested, we can't use an F-test either. To compare the two models, we transformed the predictions from the logged model back to the original scale and then compared RMSE. Using this approach, the transformed model actually performs worse. This is a bit surprising, because the rationale behind our transformations seem to make logical sense.

```
## [1] "RMSE for unlogged model:  81603.7602768842"
```

```
## [1] "RMSE for logged model:  108005.195829584"
```

Overall Effect

Ultimately, both models agree that Hours Worked Per Week has a statistically significant effect on Total Earnings, and both models also show a practically significant effect size that is either greater than $2000 or greater than 5% per additional hour worked, which are both amounts that can make a meaningful impact on a person's financial well-being. However, neither model appears to fit the data particularly well with both models exhibiting RMSE (unlogged) greater than 80000, which is extremely high in the context of annual earnings. This could be due to the fact that earnings is highly skewed and difficult to fit with a linear model, and earnings also depend on many other factors that are omitted from a single-variable model.

The findings from this study can aid labor policymakers in gaining a deeper understanding of the workforce dynamics. This research serves as a foundational step towards further exploring the relationship between earnings and work hours. Ultimately, these insights could potentially guide policymakers in creating a more balanced and equitable labor system.

Appendix

1. Link to Data Source: https://www.census.gov/data/datasets/2023/demo/cps/cps-asec-2023.html
2. List of Model Specifications We Tried:
    - Log-Log: We chose not to use this model because the hours worked variable does not really have a heavy tail, so a log transform didn't seem necessary.
    - Log-Polynomial: We chose not to use this model because the RMSE did not really improve with the squared term, so we felt the additional complexity of interpreting the extra term was not worth it.

$$\log_{10}(Y) = \beta_0 + \beta_1 \log_{10}(X) + \epsilon$$

$$\log_{10}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

3. Residual Plots:

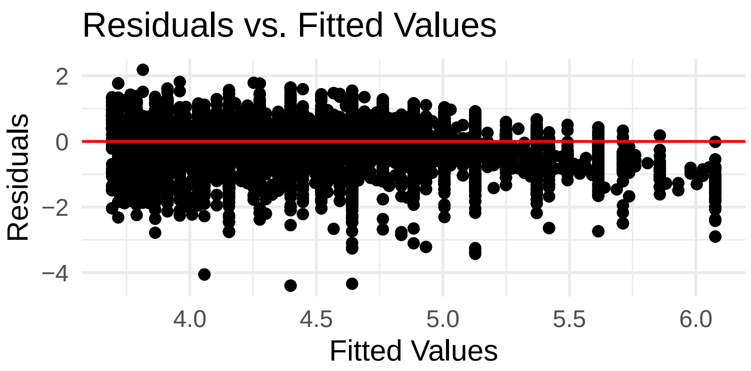

Figure 2: Residuals vs. Fitted for Transformed Model

4. Additional Plots:

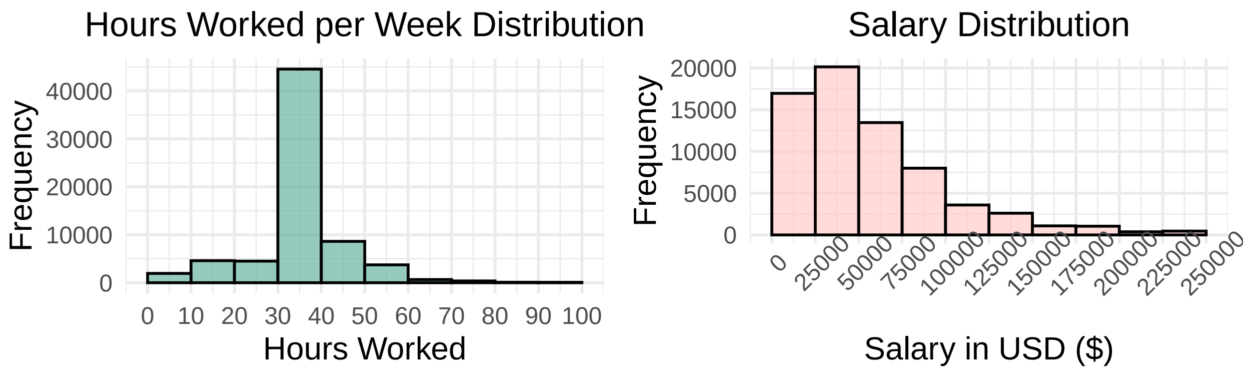

Figure 3: Distribution plots of Hours Worked per Week and Salary