Lab 2

https://github.com/mids-w203/lab-2-team-no-l-s/tree/project

Karan K Patel, Jane Lai, Maxwell Bowman

April 20, 2025

Introduction

In today's technology-driven economy, the modern workplace is increasingly shaped by the electronic tools we use to perform our jobs. But do all workers interact with technology equally? Our analysis aims to answer the following research question:

How does the percentage of time spent using electronic technologies at work vary by age?

Understanding age-related differences in workplace technology usage can illuminate potential gaps in training, inform policy on workforce development, and help businesses design environments that support employees across age groups. While prior research has explored technology use among older adults or teenagers, few studies have examined age differences across the full range of working adults in professional contexts.

Data and Methodology

Data Source:

Our cross-sectional dataset of 1,381 records combines two public sources of data:

1. General Social Survey (GSS): provides wide-ranging population-level data on attitudes, demographics, and behaviors. It is conducted by NORC at the University of Chicago (https://gss.norc.org/us/en/gss.html).

2. U.S. Bureau of Labor Statistics (BLS) – specifically, data from the Occupational Employment Statistics (OES) program and National Compensation Survey (https://www.bls.gov/eci/factsheets/national-compensation-survey-classification-systems-mapping-files.htm), which include job and industry classification info. We're utilizing this dataset's consistent categorization of job types and industries to improve classification of the respondents occupation and industry values from GSS.

Key Variables:

1. *age*: The age of the respondent.
2. *usetech*: The self-reported percentage of total time at work the respondent normally spends using electronic technologies (such as computer, tablets, smart phones, cash registers, etc).
3. *hrs1*: The self-reported number of hours worked in the week before survey by the respondent (given the respondent is employed full-time).
4. *occ10*: The respondent's occupation, coded using a 3-digit numeric scheme based on the 2010 Census Occupation Classification System.
5. *indus10*: The respondent's industry of employment, coded using a 3-digit numeric scheme based on the 2010 Census Industry Classification System.

Other Variables:

1. *wrkstat*: The self-reported status of work of the respondent (such as full-time, part-time, in school, unemployed, retired, etc).
2. *year*: The year in which the response was recorded.

In preparing our data, we filtered for 2022 responses and full-time workers only (based on *year* and *wrkstat* respectively) for consistency in employment patterns. To make sense of industry data, we joined our GSS set with BLS data to classify each respondent's industry and categorize jobs as "white collar", "blue collar" and "service". This approach better allows *age* to operate and enables us to assess whether the relationship between age and technology use differs across broad segments of the labor force. We then segmented 30% of our data into exploratory and 70% of our data into confirmation sets. Our initial model examined *usetech* as the dependent variable, with *age* as the sole independent variable, but found no statistically significant relationship. Adding *hrs1* (total weekly work hours) to test whether work hours influence technology use also did not improve the model, suggesting that neither age nor work hours independently explain variation in workplace technology use within our sample.

Methodology:

Null Hypothesis (H_0): There is no statistically significant relationship between age, weekly hours worked, and workplace technology use, after controlling for industry sector and job type.

Alternative Hypothesis (H_a): At least one of the predictors age or weekly hours worked is significantly associated with workplace technology use, after controlling for industry sector and job type.

Linear Model:

We used OLS linear regression to examine how age and weekly hours worked relate to workplace technology use, while controlling for industry and job type. The outcome variable was the percentage of time individuals reported using electronic technologies at work (*usetech*). The full model included four predictors: age, hours worked per week (hrs1), industry sector, and job type.

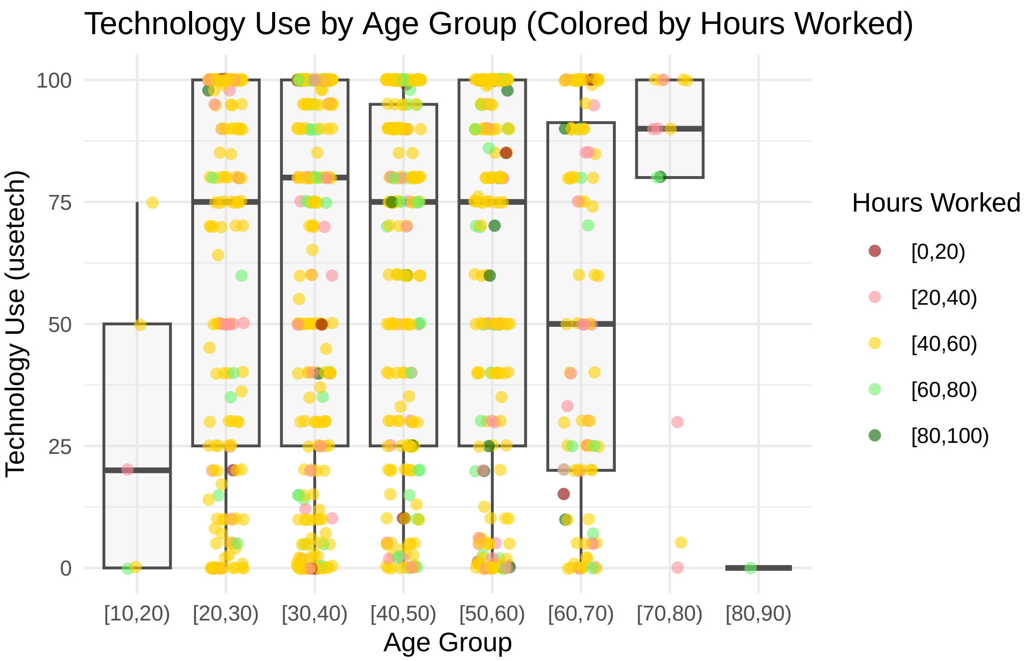


Figure 1: To explore the relationship between age, technology use, and hours worked, we used a jitter plot overlaid with boxplots. Age groups (binned in 10-year increments) are plotted on the x-axis, technology use on the y-axis, and hours worked (grouped in 20-hour buckets) is reflected through the temperature range of color. The jittered points highlight individual-level variation, while the boxplots provide a statistical summary of tech use within each age group.

We observe that the 30–60 age groups show a dense concentration of high technology use (as seen in the cluster of points near 75–100) but also exhibit **tighter interquartile ranges** in the boxplots, suggesting more consistent tech engagement across individuals in these groups. The 10–20 age group shows a much **wider box and greater spread**, reflecting high variability in tech use. Additionally, age groups above 60 show **lower medians and smaller IQRs**, with fewer high outliers—indicating both reduced tech use and a general shift toward lighter or no work hours, likely due to retirement or changing lifestyle demands.

Results

We compared the baseline model (intercept-only) with the full model to examine the explanatory power of the predictors. The results are presented below:

```
Call:
lm(formula = usetech ~ 1, data = ds)
```

```
Residuals:
   Min     1Q Median     3Q    Max 
-60.72 -35.72  14.28  39.28  39.28 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   60.718      1.215   49.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.77 on 966 degrees of freedom



Call:
lm(formula = usetech ~ age + hrs1 + industry + job_ctg, data = ds)

Residuals:
    Min      1Q  Median      3Q     Max 
-82.723 -28.864   9.244  27.912  66.760 

Coefficients:
                                                      Estimate Std. Error
(Intercept)                                           33.10304    6.78249
age                                                   -0.12743    0.08574
hrs1                                                  -0.02603    0.09797
industryEducation and Health Services                 21.47164    4.85129
industryFinancial Activities                          32.31899    5.38991
industryInformation                                   33.03239    6.58440
industryLeisure and Hospitality                       14.24798    6.57361
industryManufacturing                                 23.27426    5.42344
industryNatural Resources and Mining                  13.96301    7.97164
industryOther Services (except Public Administration) 12.53492    7.08201
industryProfessional and Business Services            33.24632    5.43233
industryTrade, Transportation, and Utilities          20.17845    4.57591
job_ctgService                                        -4.34784    4.10325
job_ctgWhite Collar                                   22.29273    2.70660
                                                      t value Pr(>|t|)    
(Intercept)                                             4.881 1.24e-06 ***
age                                                    -1.486   0.1375    
hrs1                                                   -0.266   0.7905    
industryEducation and Health Services                   4.426 1.07e-05 ***
```

```
industryFinancial Activities                             5.996 2.86e-09 ***
industryInformation                                      5.017 6.27e-07 ***
industryLeisure and Hospitality                          2.167   0.0304 *
industryManufacturing                                    4.291 1.96e-05 ***
industryNatural Resources and Mining                     1.752   0.0802 .
industryOther Services (except Public Administration)   1.770   0.0771 .
industryProfessional and Business Services               6.120 1.36e-09 ***
industryTrade, Transportation, and Utilities             4.410 1.15e-05 ***
job_ctgService                                          -1.060   0.2896
job_ctgWhite Collar                                      8.236 5.81e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.18 on 953 degrees of freedom
Multiple R-squared:  0.1923,    Adjusted R-squared:  0.1812
F-statistic: 17.45 on 13 and 953 DF,  p-value: < 2.2e-16
```

The overall model was statistically significant (F(13, 953) = 17.45, $p < 0.001$), indicating that at least one predictor is meaningfully associated with the outcome. However, with an adjusted R-squared of 0.1812 these predictors only explain approximately 18% of the variance in workplace technology use.

Among the predictors, job category and industry show notable association with usetech. Individuals in white-collar jobs reported significantly higher tech usage compared to those in other job categories ($\beta_1 = 22.29$, $p < 0.001$). Industries such as Information, Manufacturing, and Professional and Business Services also had significantly higher reported tech use relative to the baseline industry category. In contrast, age and weekly working hours worked were not significantly associated with technology use in this model.

Discussion

The regression analysis indicates that age and hours worked per week do not significantly explain variation in workplace technology use, countering the assumption that older individuals are less engaged with technology. The estimated effect of age was close to zero and statistically insignificant, even after accounting for work hours, job category and industry.

In contrast, industry and job category demonstrated stronger and statistically significant associations with technology use. Employees in industries like Information, Manufacturing, and Professional and Business Services, as well as those in white-collar roles, reported higher technology engagement than others.

These findings suggest that workplace technology use is driven more by the nature of the work being performed than by individual demographics such as age. Consequently, we fail to reject the null hypothesis, indicating no statistically significant difference in technology use across age groups.

Appendix

"Without Job Category" - "With Job Category" Linear Model Comparison

In the Reports>Results folder of our repo, stargazer results comparing the Linear Models for "Without Job Category" and "With Job Category" indicates that the second model including the "Job Category" variables provides a better fit and explains more of the variation in Technology Use (usetech) than the model without "Job Category" variables.

Stargazer results providing this indication include the r-squared coefficient of 10.9% in technology use without Job Category being lower than the 19.2% in the model with the Job Category. Additionally, industry categories such as "Education and Health Services" have significant positive coefficients, indicating that this industry, as an example, increases technology use. Lastly, our model without Job Category has an f-statistic of 10.321 compared to the f-statistic of 17.472 provided by the model with Job Category- indicating that the inclusion of the Job Category variable has improved the model.

ANOVA Test

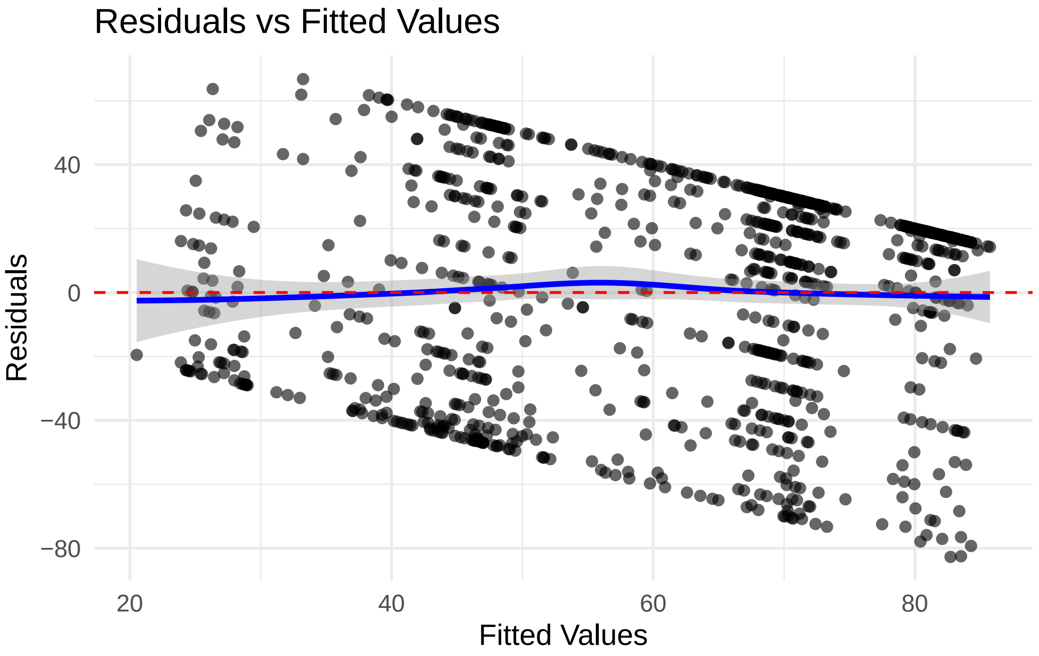
```
Analysis of Variance Table

Model 1: usetech ~ 1
Model 2: usetech ~ age
Model 3: usetech ~ age + hrs1
Model 4: usetech ~ age + hrs1 + industry
Model 5: usetech ~ age + hrs1 + industry + job_ctg
  Res.Df     RSS Df Sum of Sq       F Pr(>F)
1    966 1378224
2    965 1377427  1       797  0.6826 0.4089
3    964 1376310  1      1117  0.9560 0.3284
4    955 1227594  9    148716 14.1453 <2e-16 ***
5    953 1113256  2    114339 48.9396 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conducted a sequential ANOVA test to evaluate how well different predictors explain variation in usetech. Five nested models were compared, beginning with a null model (Model 1) and progressively adding predictors.

- Model 2 adds age, but does not significantly improve model fit (F = 0.6826, p = 0.4089).

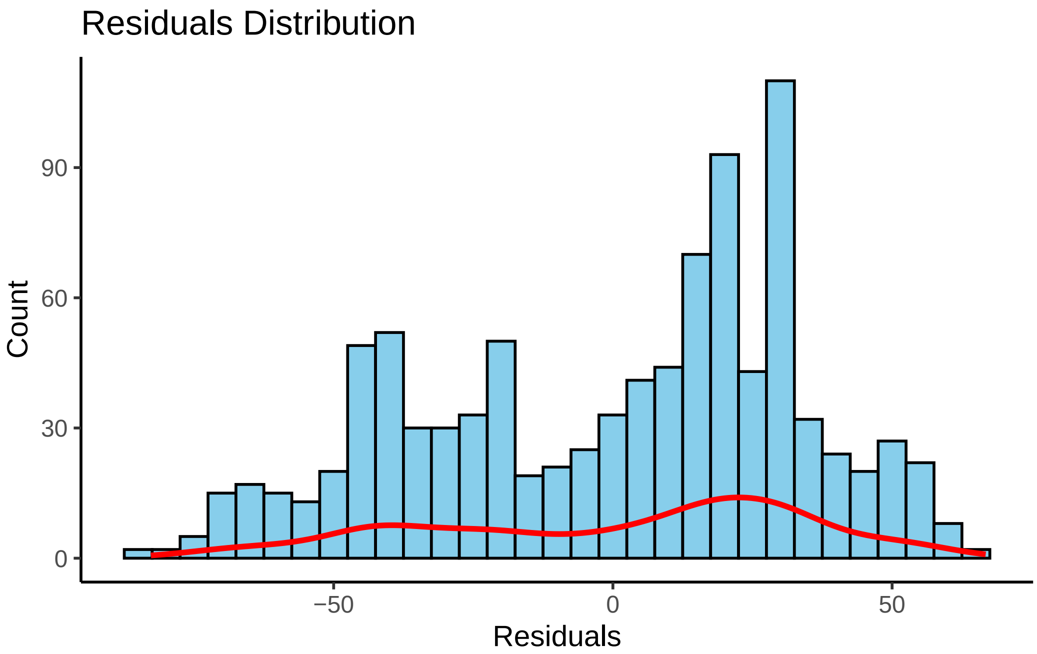
- Model 3 adds hrs1 (hours worked per week), resulting in a small, non-significant improvement (F = 0.9560, p = 0.3284).
- Model 4, which includes industry, shows a substantial and statistically significant improvement in fit (F = 14.14, p < 0.001), suggesting that industry is an important predictor of technology use.
- Model 5 adds job category (job_ctg) and also significantly improves model fit (F = 48.94, p < 0.001), indicating job category adds additional explanatory value after accounting for age, hours, and industry.

Fitted values Vs Residuals plot



This is the Residuals vs Fitted Values plot, which helps evaluate whether the model's prediction errors are randomly distributed. It is commonly used to assess key linear regression assumptions, such as homoscedasticity (constant variance) and the correctness of the model's functional form.

The plot reveals two key issues: heteroscedasticity and non-normality of residuals. First, the residuals show non-constant variance across fitted values, violating the homoscedasticity assumption. This undermines the model's reliability and can lead to biased standard errors and p-values. Second, the residuals are not normally distributed, as seen in the histogram. While normality isn't strictly required for descriptive regression, it affects the validity of inferential statistics such as p-values and confidence intervals, reducing the credibility of hypothesis test results.



This pattern is primarily due to the discrete nature of both the response variable (usetech) and predictors like age and hrs1. Since these variables take on repeated, limited values rather than a continuous spread, the model generates fitted values and residuals that are not smoothly distributed but instead form visible bands or blocks. Additionally, there maybe issues with model specification like omitted variables, non-linear relationships, or inappropriate use of a linear model on data that might be better suited to other appropriate regression techniques.

An alternative regression technique called *beta regression* can be used. Beta regression is well-suited for continuous percentages bounded between 0% and 100%. Unlike linear regression, it accounts for the heteroscedasticity. This makes it particularly appropriate when the residuals in linear regression exhibit non-constant variance.