how are people getting inflicted by COVID-19

Bradley Robert Wong

27 April 2022

Abstract

For decades, the world has been through different financial crises and different events for example, housing crises, brexit vote etc. there have been a lot of events affecting the world's ecnomony and all these would be lead to affecting peoples salary income. to understand the living standards of the past, i used a linear regression to do an analysis on how the wages in 1985 was affected what affected them.

Keywords: Wages, Age, Education, Gender, Linear Regression, Ethnicity

Contents

1 Introduction

The Current Population Survey is a statistical survey that is jointly sponsored by the U.S. Bureau of Labor Statistics (BLS) and the U.S. Census Bureau (Census) on a monthly basis. It is the primary source for U.S. labor force statistics. The survey includes a representative sample of about 60,000 homes and focuses on those individuals who are 15 years and older to make an inferential assumption about the U.S. population as a whole.

The data i have selected today is the data from May 1985. i have decided to pick a month in the middle of the year since jobs are mostly stable and without the start and end of year bonuses, the wages in the middle month of the year would better and a more accurate estimation of peoples income.

The CPS is a very important survey as the information gathered are from individuals of all demographics and used to estimate the unemployment rate of the country. The information gathered every month and analyzed on is to ensure the wellbeing of people living in the US, as the government would be able to implement different policies and react to sudden changes of the living qualities of citizens.

Though during COVID-19 the response rate of the survey has decreased to 69.9% in april 2020, the data we are using today is from 1985 which has a response rate of up to 90% back in the days.

The dataset would provide valuable insights. the model would indicate the factors affect the wages of people in 1985 and . the paper would first provide an overview of the dataset i have collected and the variables relevant to the study i would want to conduct. the method of data collection would also be stated and given a run through. next, the methodology section would discuss the pros and cons of the linear regression model i would be using. i would also address how the weaknesses can be further improved and make the study more all rounded. the results section would then display my findings on the models and graphs i have used in the process, with explaining the interpretation of concepts in my results. finally, the discussion section would include subsections of my findings, ethics, limitations and improvements.

2 Data

This report, including all necessary data cleaning, analysis and visualization, was produced with the R statistical programming language (R Core Team 2020) in the R Markdown file format. It uses features from several packages. The dataset is accessed through the CPR dataset. The tidyverse package is used for data cleaning and manipulation (Wickham et al. 2019), as well as janitor (Firke 2021). Tables and graphs are generated with kableExtra (Zhu 2020), modelsummary (Arel-Bundock 2021), patchwork (Pedersen 2020), and ggplot2 (Wickham 2016). A map iscreated with ggmap (Kahle and Wickham 2013), osmdata (Padgham et al. 2017) and sf (Pebesma 2018).External data is accessed, imported, and exported with RCurl (Temple Lang 2021), readxl (Wickham and Bryan 2019), xml2 (Wickham, Hester, and Ooms 2020). The models are assessed with performance (Lüdecke et al. 2021). To facilitate a reproducible workflow, here (Müller 2020) is used to reference file locations, knitr (Xie 2021) to format the report into a PDF style.

The way to recruit respondents and answers to surveys are not specified and that we have assumed it to be randomly sampled.

The dataset contains 547 observations taken from the CPS with 11 variables of the wages, education, experience, age, ethnicity, region, gender, occupation, sector, union and married. the variables are a good representation of the living of people as income is one of the more direct ways of testing the living conditions of citizens.

One thing to notice is that we need to understand shortcomings of the survey, one of which is the source didnt state the number of peoples approached. we only know that there are 547 respondents in this dataset. This means that we would not be able to calculate and estimate the effects of response bias to this paper. Second thing to note is that the income stated as hourly wages doesnt specify whether the amount is before tax, after tax, including side business or only job salary. this would be a slight factor that may lead to instability of the linear regression model. However, since the respondents are taken randomly, we can still assume there wouldnt be a big bias on the paper.the dataset doesnt include options for respondents to not answer to the questions such as answers like dont know, do not wish to responds. therefore, in the dataset, there werent any responses that needed to removed nor cancelled.

After going through the data, there doesnt seem to have any missing data nor an any patterns in missing datas.

3 Variables

Some key variables to note is that wages is calculated in dollars per hour.

Religion would only include whether the person lives in the south of otherwise.

Experience is only a rough estimation using the formula of age - education - 6.

Education is the number of years of education.

In constructing our model, we have decided to take out several variables and use 2 key variables as our predictor variable which is age and education.

4 Methodology

Since this paper aims to find the factors affecting income, we have decided to use a linear regression model to showcase our findings.

The model assumes the wages as the response variable.

The four other predictors i will be using would be age, education, gender and ethnicity.

$$y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_5$$

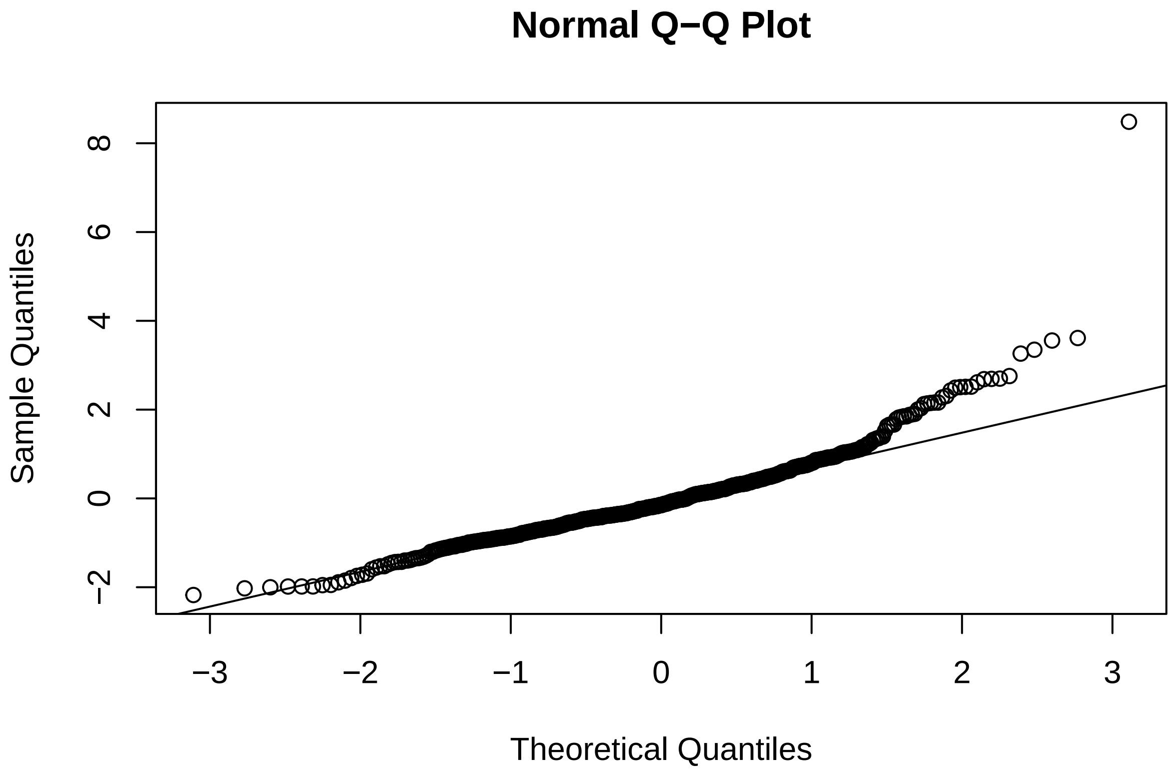
In the above equation, y would the response variable wages, b1 is predictor variable age, b2 is predictor variable education, b3 is predictor variable gender and b4 and b5 are both predictor variable ethnicity.

Linear regression model was selected to predict the value of a variable(response variable) based on other variables (predictor variables). By knowing the correlation between the variables we would be able to conclude how the response variable reacts to the change of the predictor variable.

5 Assumptions

Before we can conclude anything and proceed to making our linear regression model, there are several assumptions we need to meet and makes sure our model would not violate them.

The 4 assumptions would be linear relationship, independence, homoscedasticity and normality.



In the above plot, what we see is the normal qqplot. as we can see the end seems to be up and far away from the line. the means that the residuals doesnt full follow the line and it isnt normally distributed. one more thing to note is that there is an outlier on the end of the graph. these all would lead the violation of normality, meaning we would not be able to successfully and accurately implement our linear model.

In this case, we have decided to undergo a transformation for our variables. A transformation in linear regression model would be able to help us stabalize the variance or improve the normality, which in this case we need to improve the normality of our model. This can improve our variables and avoid violating our assumptions.

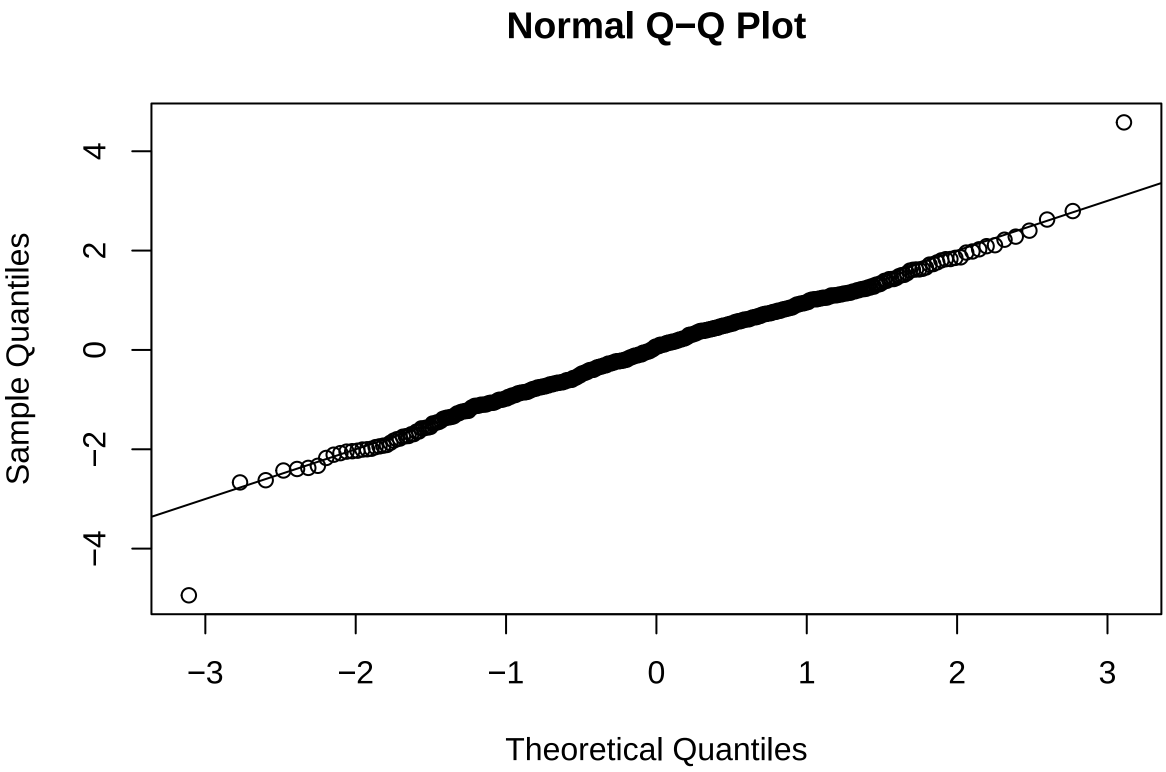
One thing to note about is when we were plotting the residual to variable graph, the graph of education and residual has a slight trend of a fanning shape, however, the trend wasnt very obvious and we have decided that the effect of it on the model would not be very big therefore, we have decided it would be fine for our to continue. however, in the future, if we can add more observations to the dataset and dim out the trend it would be even better.

6 transformation

```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.0253        0.00      -0.1519       0.1012
## Y2   -0.0494        0.00      -0.3114       0.2126
## Y3    1.2649        1.26       1.0075       1.5223
```

```
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                LRT df       pval
## LR test, lambda = (0 0 0) 142.1675  3 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                LRT df       pval
## LR test, lambda = (1 1 1) 302.0455  3 < 2.22e-16
```

As you can see from the table , there are 3 variables that we have decided that would undergo transformation. since the other variables are categorical, only numerical variables are able to do transformation. we can see the first 2 variables is 0. this means we would do a logarithmic of the variables wage and age. for Education, since the rounded power is very close to 1.00, we have decided to keep it as the same without transforming it.



After the transformation, we have plotted the normal qqplot again, we can see that the residuals have lies on the line and that normality violation has been fixed. with normality violation fixed, all the 4 assumptions have been satisfied and we can hence move on and plot our linear regression model.

Since we did a transformation on some of the variables, there will also be slight amendments to the final equation for the linear regression model.

$$logy = \beta_1 + \beta_2 logx_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_5$$

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

From the summary table above, we can see the values of each variables in the linear regression model. for variables log(age) and education. with a p value of 0, it means the two variables are very statistically

Table 1: summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.9287507	0.2547555	-3.6456551	0.0002931
data_tran$age_trans	0.5050543	0.0616423	8.1933062	0.0000000
data_tran$education	0.0820851	0.0075408	10.8855200	0.0000000
data_tran$gendermale	0.2587125	0.0389616	6.6401846	0.0000000
data_tran$ethnicityhispanic	-0.0887915	0.0897198	-0.9896534	0.3227969
data_tran$ethnicityother	-0.0983500	0.0587810	-1.6731584	0.0948884

significant towards the model and that they are the most important variables.However, ethnicity and gender has a fairly high p value meaning they would not be able accurately reflect the relation between them and wage.

Our final model will be modeled as below.

$$logy = -0.9287507 + 0.5050543logx_1 + 0.0820851x_2 + 0.2587125x_3 - 0.0887915x_4 - 0.0983500x_5$$

# Limitations The first limitation would be the sample size being a little small. as we can see the proportion of people in ethnicity seems to have a very large portion of caucasian people in the dataset with over 70%. this problems seems to have appeared in several other variables including occupation and sector. there dataset seems to contain a certain proportion more than others and i dont think this is the proportion of the real society back in the days. To make the model a more comprehensive model, it is important for us in the future to take more respondents and include more genres of people to ensure the observations we have collected is a good representation of the society. only in this case, the model would be more persuasive.

The second limitation would be the dataset only containing 2 numeric variable and the lack of variables. for our model, we have concluded that there are only 2 variables that are statistically significant to the model, however, there are a lot of more factors affecting the wage of people. we can add some open ended questions to the survey, this can help us have more different observations and know maybe different factors are affecting different levels of income of people or specific occupations of people.

One last limitation is that when we were plotting an age to wage and education to wage scatterplots, we realized that there were an exceptionally high outlier in the plot, in future investigations, we should look into it, understanding why such an observation has appeared. since we didnt have too many observation, we had to make sure the outlier wouldnt affect our model significantly.

Discussion

Unsurprisingly, with more years of Education, people tend to have high amount of wages. this is very normal since more years of education meaning more learning and more knowledge, able to be proficient in a certain aspect of the field definitely brings more value of the company and hence will have a higher wage. This concept is the same as the person's age, if a person is older, the person has more experience. With more experience, they will have higher wages.

We chose income as our topic because we believe income can represent a lot of things not only the individual but also the society. With higher average wages, it may mean better economy in the city and better living quality of individuals in the city

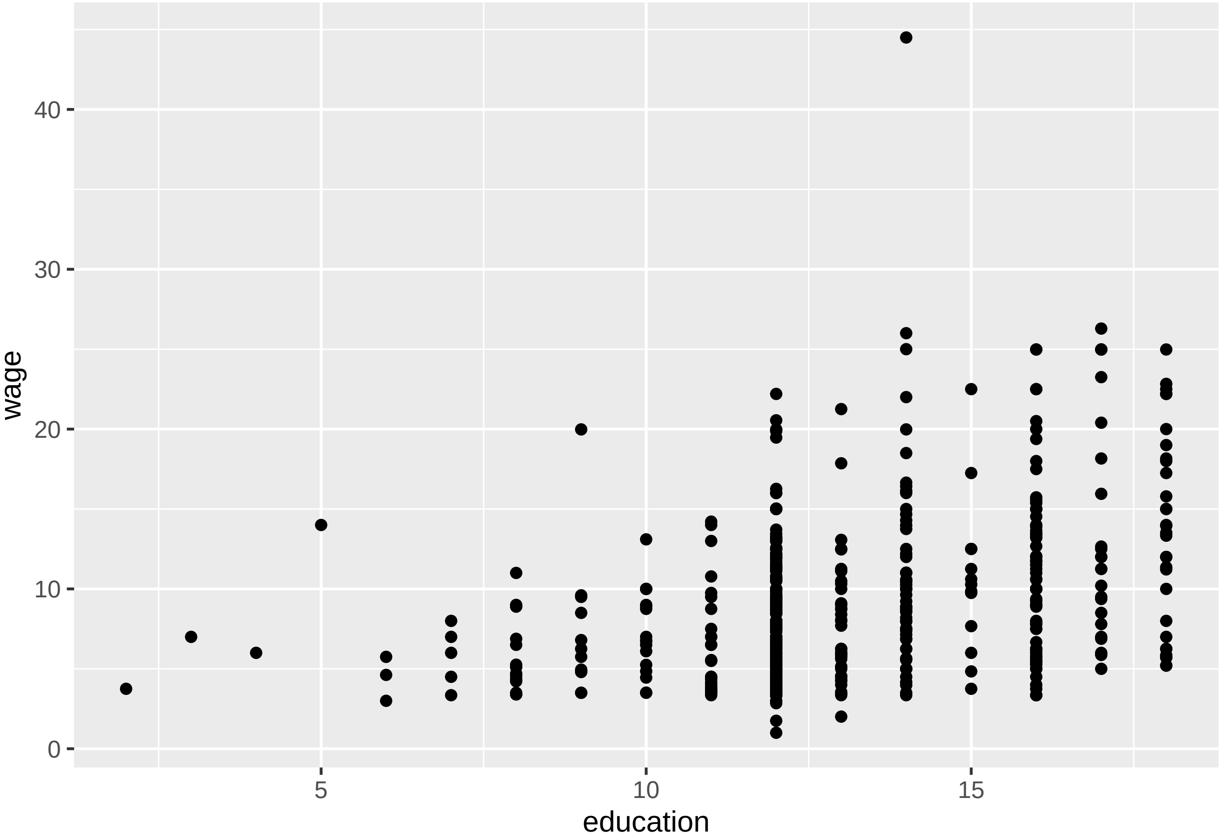
Enhancements

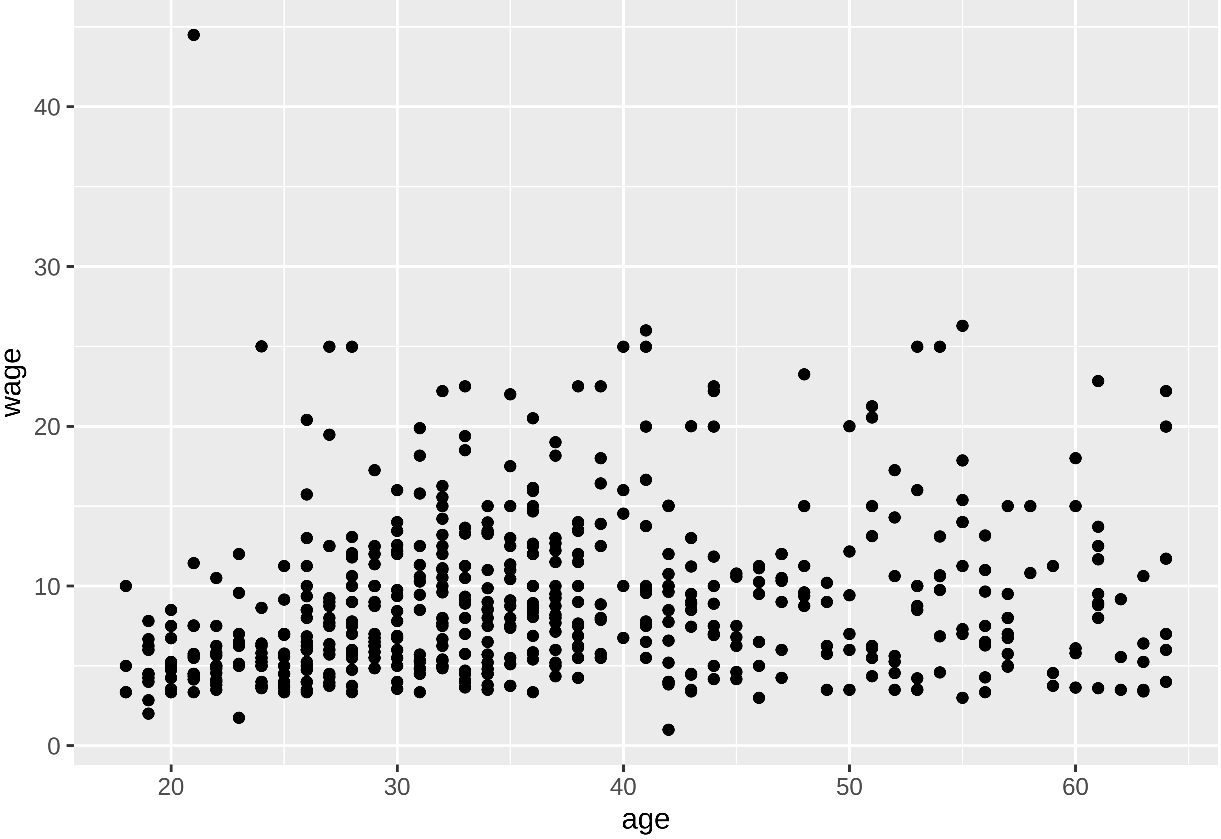
Ethics

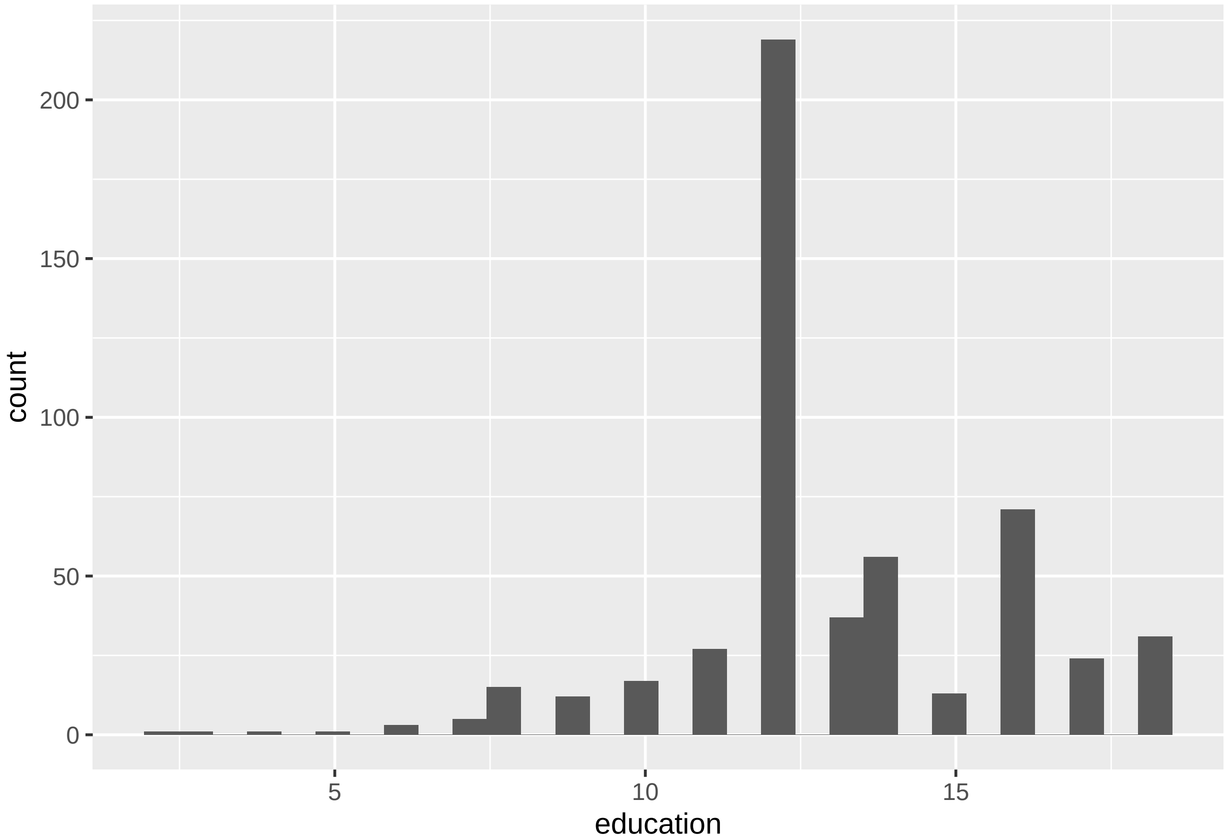
The survey didnt state whether it was a annoynymous survey or not. confidentiality wasnt included in the survey as well. respondents to the survey also isnt clearly stated that they were informed that information they give would be confidential or not. Even though except income, there werent any specifically private information collected but information to contact respondents should be collected in order to appropriately inform measures and uses of the respones they give will be used for.

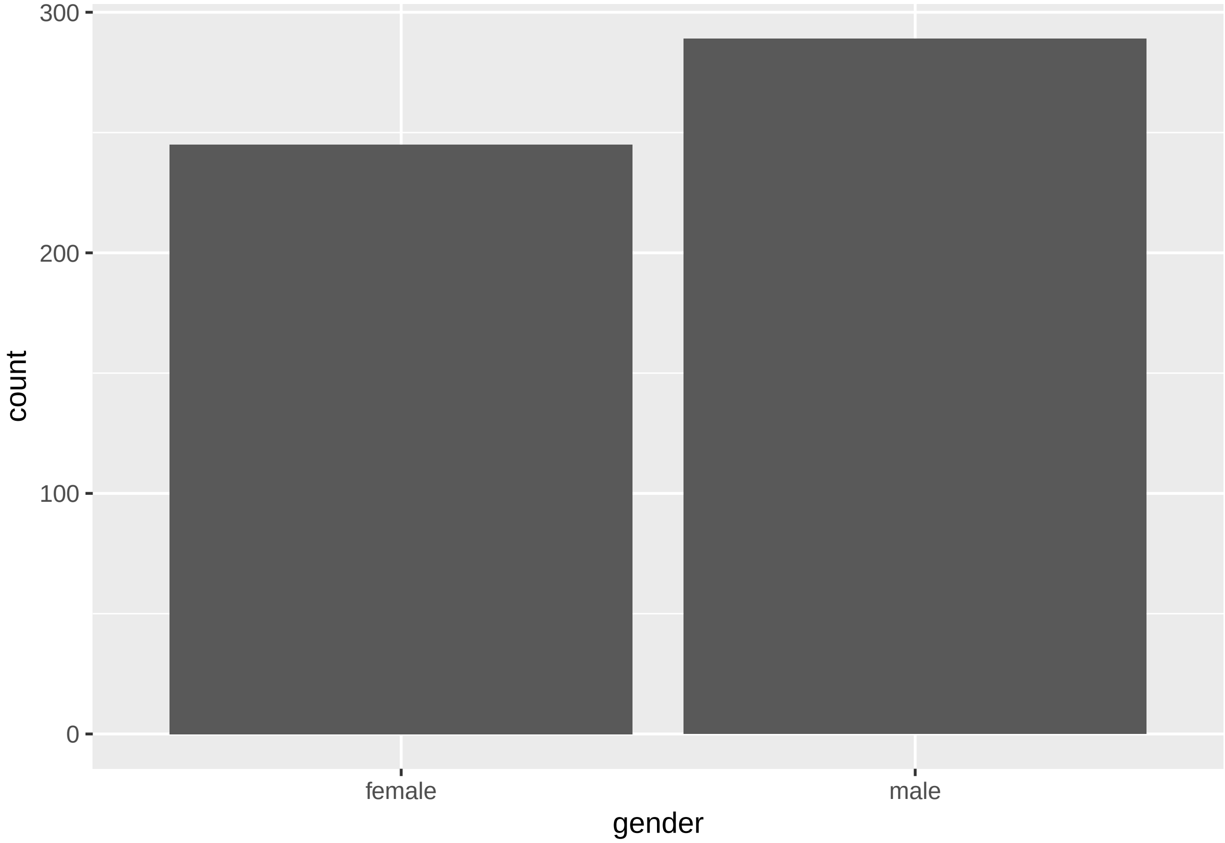
Ethics standards should be assured involving any human activities especially sensitive information like income, person information etc… we should always make sure that appropriate annoucements are given to respondents to ensure their privacy is being well protected.

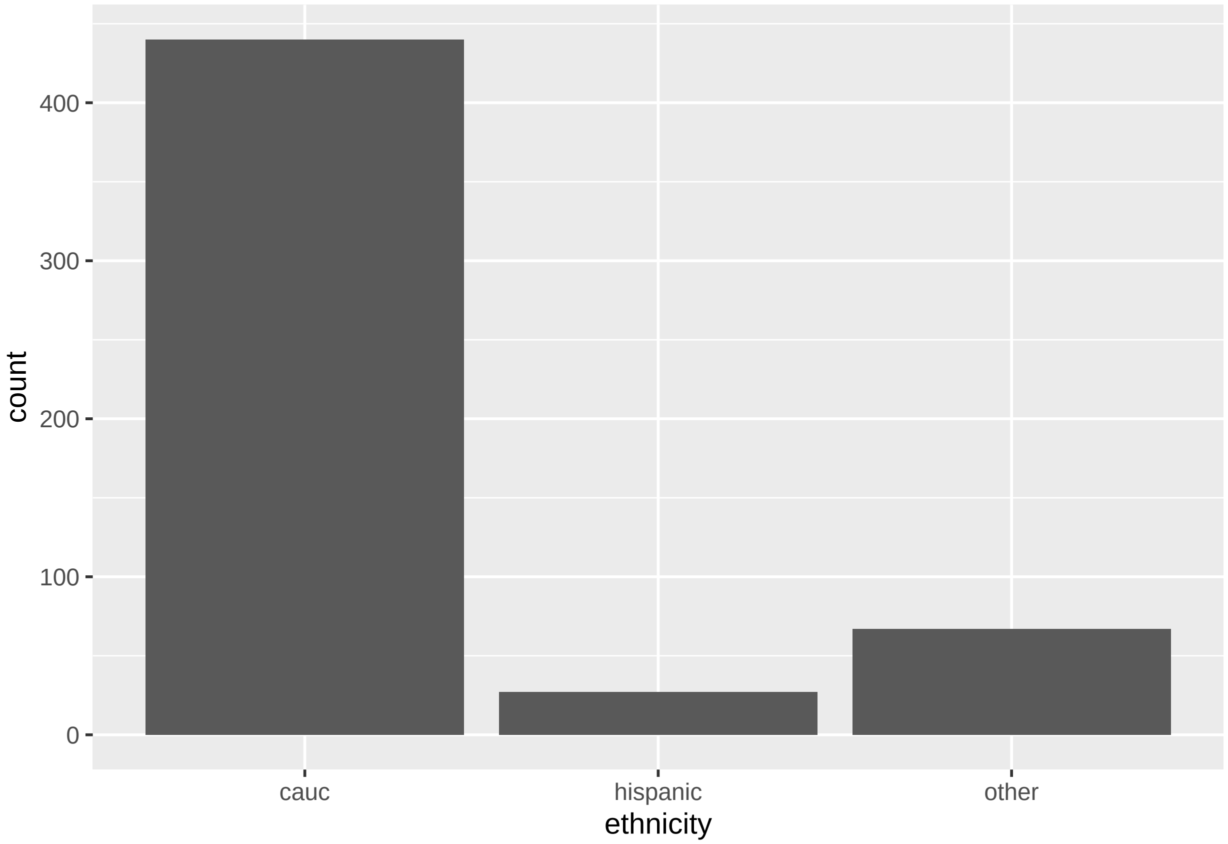
Appendix

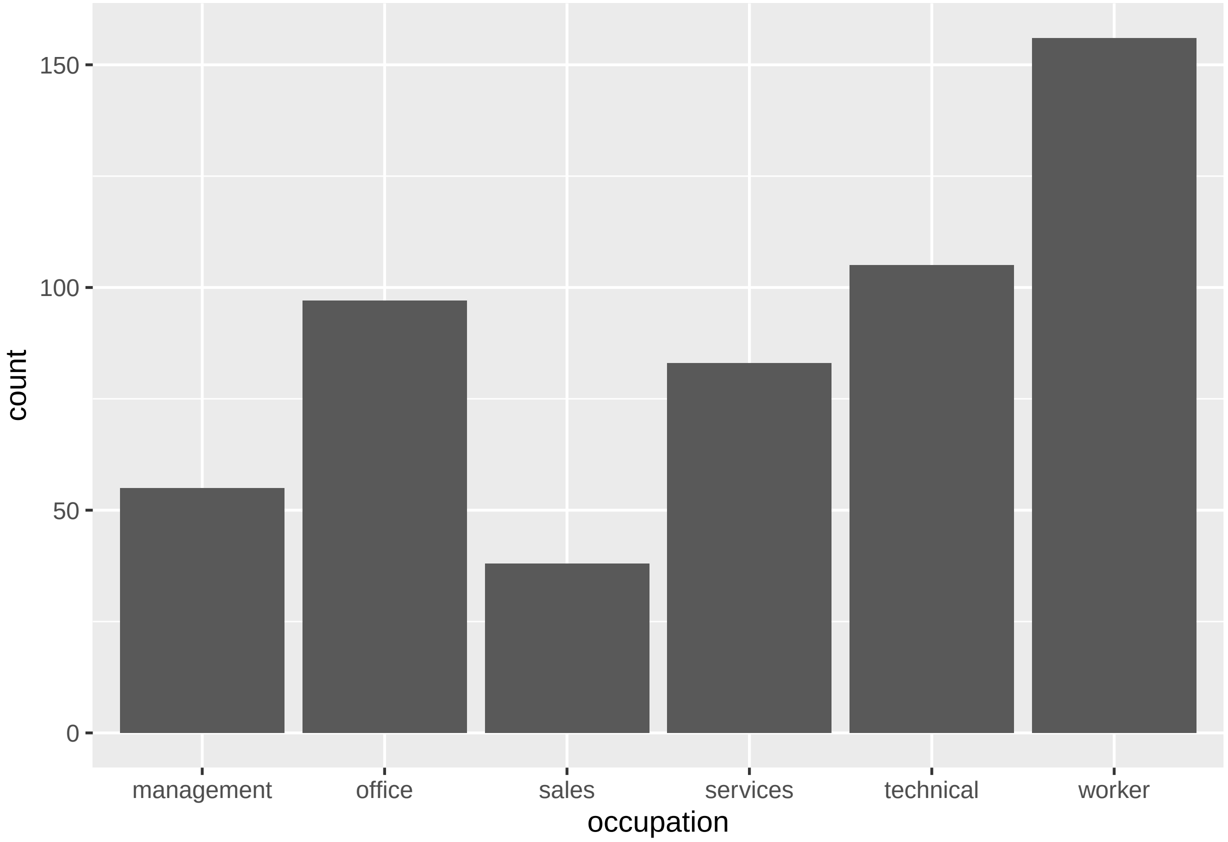

wage
0
10
20
30
40
5
10
15
education

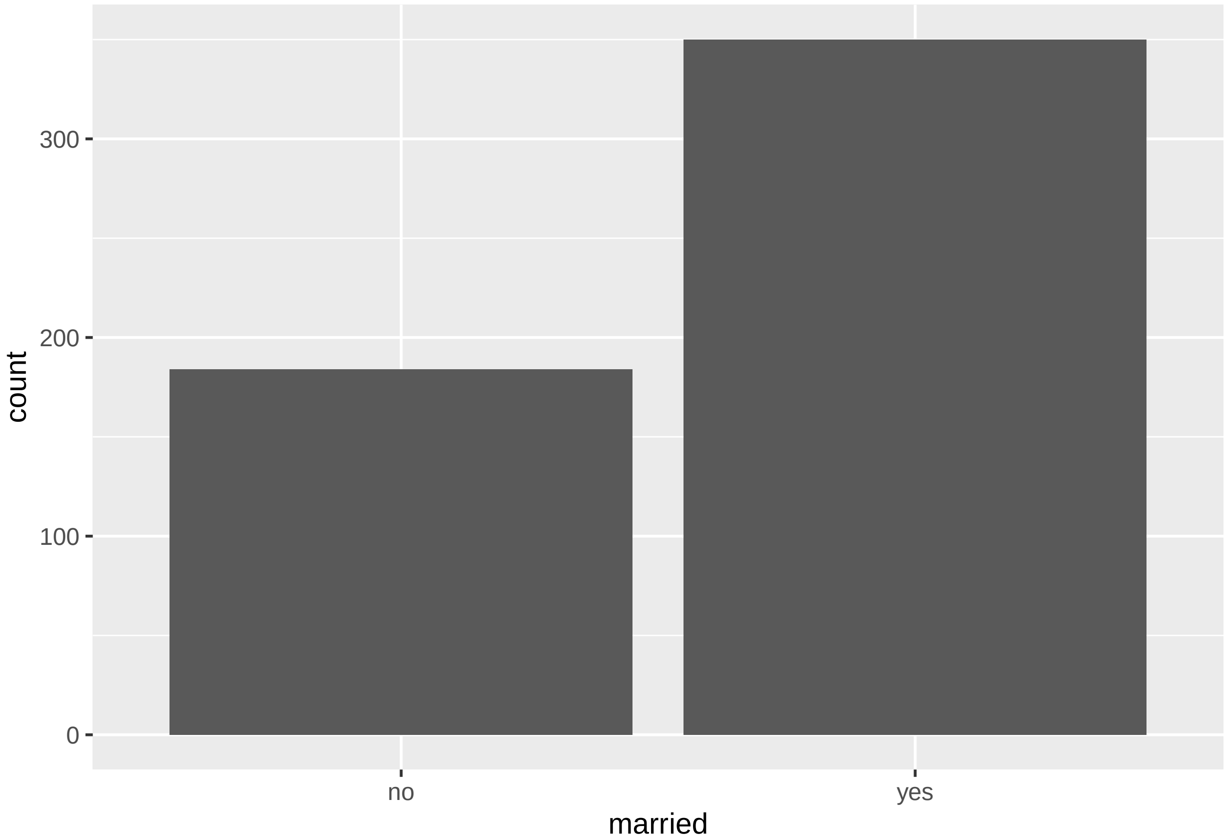

wage
0
10
20
30
40
20
30
40
50
60
age

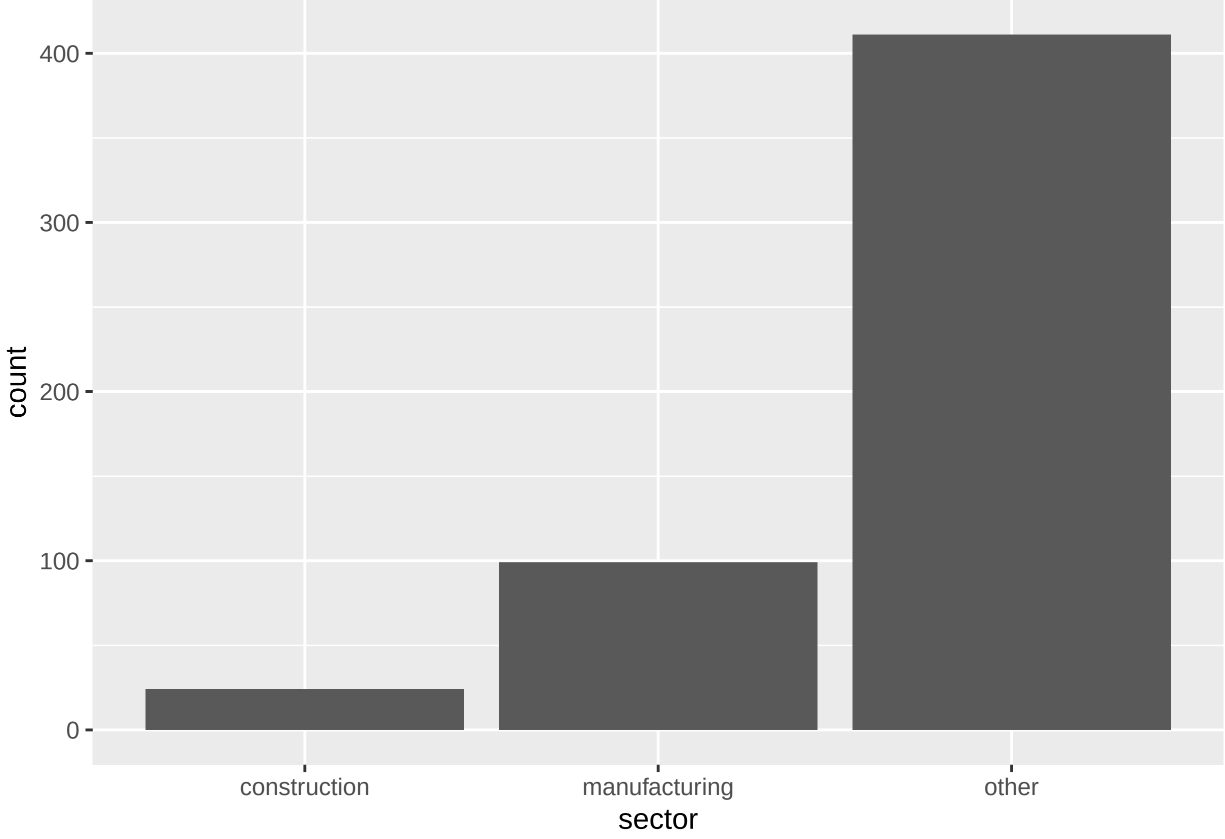

count
200
150
100
50
0
5
10
15
education

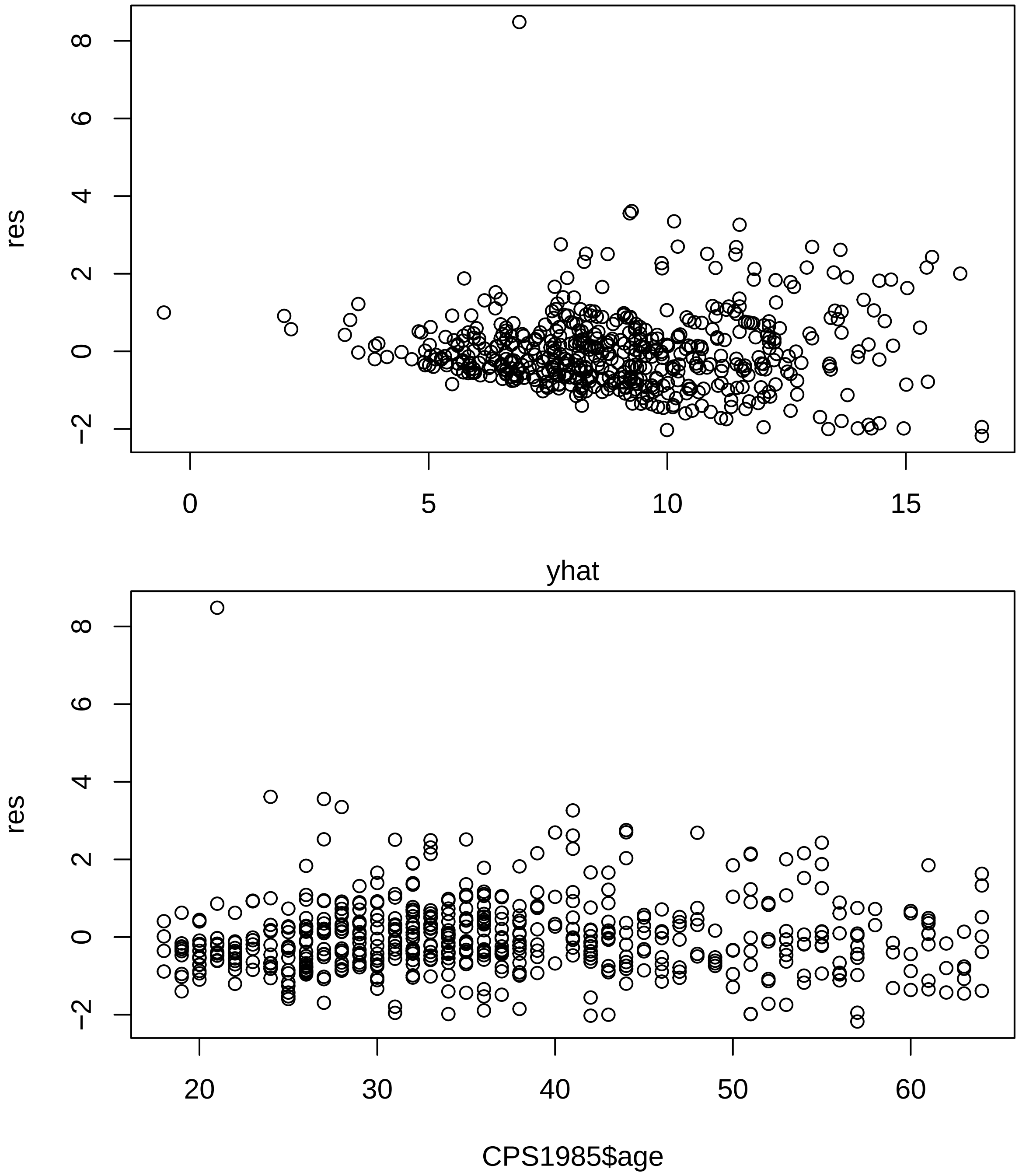

count
300
200
100
0
female
male
gender

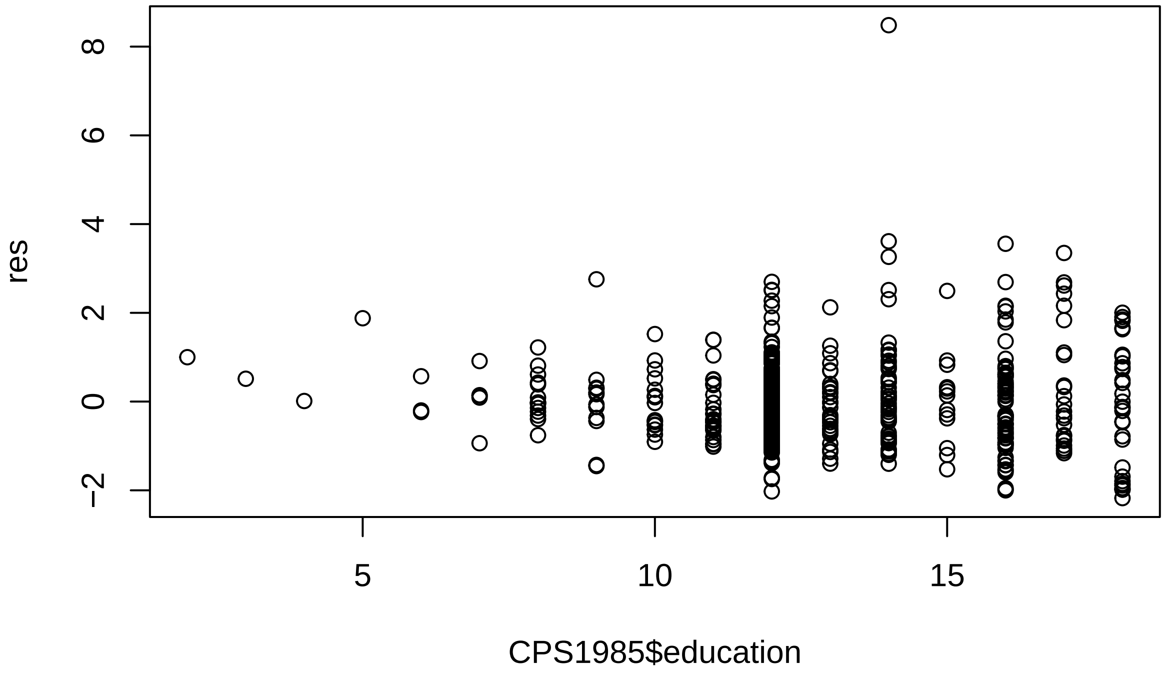

count
400
300
200
100
0
cauc
hispanic
other
ethnicity


count
150
100
50
0
management
office
sales
services
technical
worker
occupation


count
0
100
200
300
no
yes
married


count
0
100
200
300
400
construction
manufacturing
other
sector


res
8
6
4
2
0
−2
0
5
10
15
yhat
res
8
6
4
2
0
−2
20
30
40
50
60
CPS1985$age



References

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.