# Homework 0

Your Name  Your Partner's Name

> **Note**
>
> The top of this file contains the YAML header, which is used to set the title and author of the document. The YAML header is delimited by three dashes on the top and bottom. The YAML header is not rendered in the final document.
> For your homeworks, you will need to change the title and author fields in the YAML header to your name and your partner's name.
> This file example assumes you use `nbstata` and the full document can be rendered from the file itself. To do this, you need to include the last line -jupyter: nbstata- in the YAML header.
> Since all code will be run from the QMD file, we also add additional options to execution.
>
> ```
> execute:
>     echo: false
>     output: false
> ```
>
> This will prevent the code or ouput from being displayed in the final document. You can change this option to true if you want to display the code in the final document.
> Since the homeworks will use a -research project- format, you should add a title to each section.

## Determinants of Wages: a first look at the data

### Introduction

There is considerable evidence that suggests that wages vary across individuals based on their characteristics.

The human capital theory suggests that wages are determined by the individual's characteristics, such as education, experience, and other characteristics that may not be meaured or may remained unobserved.

As individuals increase their education, they are expected to increase their productivity and, therefore, their wages. On the other hand, it is also possible that more education does not increase productivity, but generates a signal effect that increases wages.

Similarly, as individuals gain more experience, they are expected to increase their productivity and, therefore, their wages. In this case, we are only considering age as proxy for experience.

Finally, we may also consider marital status as a proxy for other characteristics that may affect wages. For example, married individuals may have more responsibilities than single or divorced, and, therefore, may require higher wages. While divorced and single are similar in terms of responsibilities, divorced individuals may still have other responsabilities with the previous marriage like child support. Thus, they may also require higher wages.

In this homework, we will explore the determinants of wages using a sample of 1434 individuals, from the Excerpt from the Swiss Labor Market Survey 1998. The data is available from an online repository, but a copy of the file is available at `data/oaxaca.dta`.

> **Note**
>
> For this particular example, you won't need to provide descriptive statistics, as it was not explicitly requested. However, you should always produce them for your own benefit. Adding details to your homeworks will help you to better understand the data and the results. And may be granted extra points.

### Exploration: Descriptive Statistics

### Exploration: Univariate analysis

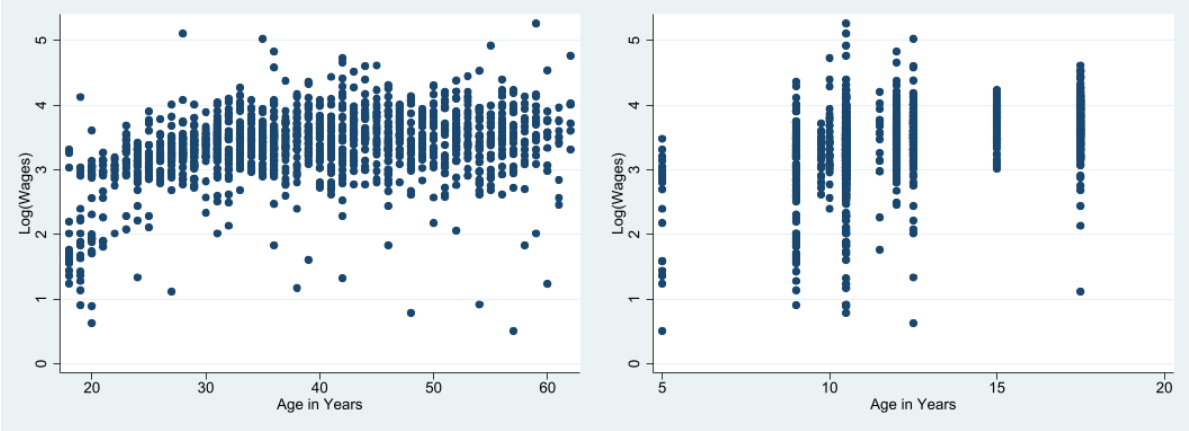
We explore the relationship between wages and the variables of interest by first using a simple correlation analysis.

| | lnwage | educ | age | agesq | single | married | divorced |
|---|---|---|---|---|---|---|---|
| lnwage | 1 | .3649418 | .350274 | .302245 | -.2375836 | .1693033 | .0900732 |

As it can be seen from this table, the correlation between wages and education and age is positive, indicating that as education and age increase, wages may also increase.

With regards to marital status, the correlation is negative for single, and positive for divorced and specially for married. This suggest that married individuals may have higher wages than single or divorced individuals.

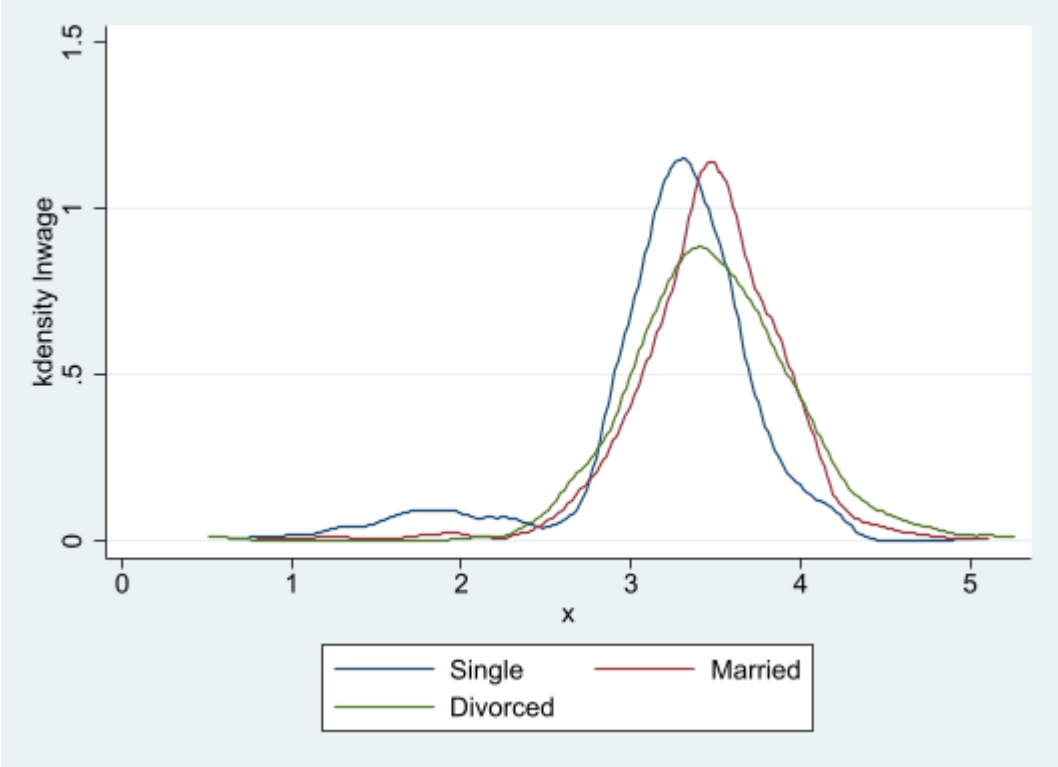
In addition to this, we can also construct scatter plots and density plots to visualize the relationships between these variables and wages:




In both cases, we see an increasing trend between wages and age and wages and education. However, we also see that there is a lot of dispersion in the data.

With respect to age, there is a sharp slope for younger individuals, but the slope is flatter for middle age and older invidivuals. In terms of education, there isn't a clear pattern because of the data dispersion and little variation in education. Nevertheless, there is some signs of a positive relationship between wages and education.

Because marital status is a categorical variable, we can't construct a scatter plot. However, we can construct a density plot to visualize the relationship with respect to wages:




This plot confirms what we suspected earlier, that singles have the lowest wages, followed by divorced, and married individuals have the highest wages. Among singles, there may be a cluster of low wage individuals.

In the next section, we will explore the relationship between wages and the variables of interest using a multivariate regression analysis.

## Regression analysis

We estimate a simple regression model to explore the relationship between wages and the variables of interest, as a function of age, education and martial status:

| | (1) | (2) |
|---|---|---|
| age of respondent | 0.014*** | 0.106*** |
| | (0.001) | (0.008) |
| years of education | 0.077*** | 0.067*** |
| | (0.005) | (0.005) |
| Marital Status=1 | 0.000 | 0.000 |
| | (.) | (.) |
| Marital Status=2 | 0.087** | -0.003 |

| | (1) | (2) |
|---|---|---|
| | (0.030) | (0.030) |
| Marital Status=3 | 0.113** | 0.038 |
| | (0.043) | (0.042) |
| age squared | | -0.001*** |
| | | (0.000) |
| Constant | 1.892*** | 0.319* |
| | (0.072) | (0.151) |
| Observations | 1434 | 1434 |
| $R2$ | 0.247 | 0.312 |

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results of interest are in column 1. Based on this, age and education have a positive effect on wages, while being married and divorced have a positive effect on wages.

Each year of age may increase wages in 1.4%, while each year of education may increase wages in 7.7%. Which is consistent with the results from the correlation analysis and our previous discussion.

Regarding marital status, married and divorced earn wages that are about 8.7% and 11.3% higher than singles. This is not consistent with the previous dicussion. Looking at the density plot, however, it may be that the wage distribution of divorcees is larger than married or single, explaining the higher mean wage.

In column 2, we add a quadratic term for age. We do this because is not likely that people will continue to earn higher wages as they become older. Instead, it is likely that wages will increase until a certain age, and then decrease. The coefficient for age is now larger than before at 0.11, while the coefficient for age squared is small, negative, but highly significant.

This result suggests that the original model may suffer from misspecification, since its not capturing the nonlinear relationship between wages and age. Also, that the "age" coefficient cannot be interpreted on its own. If we estimate the marginal effects for someone age 40 (about the average age in the sample), the effect its 0.0147. Just slighly higher than our main model.

## Testing for model-misspecification and heteroskedasticity

We test for model misspecification using the RESET test. They are based on the model that includes age squared in the specification.

The results are included below:

**RESET test for model misspecification**

Ramsey RESET test for omitted variables Omitted: Powers of fitted values of lnwage
H0: Model has no omitted variables
F(3, 1425) = 25.55 Prob > F = 0.0000

Similarly, we test for heteroskedasticity using the default options in Stata. The results are included below:

**Heteroskedasticity test**

Breusch–Pagan/Cook–Weisberg test for heteroskedasticity Assumption: Normal error terms Variable: Fitted values of lnwage
H0: Constant variance

```
chi2(1) =  54.58
```

Prob > chi2 = 0.0000

Based on the results, we conclude that we reject the null hypothesis of correct specification of the model. This means that we should consider other specifications for the model.

Similarly, we reject the null hypothesis of homoskedasticity. Because of this, statistical inference of the model should be done using robust standard errors, or Generalized Least Squares.

## Conclusions

In this homework, we explored the relationship between education, age and marital status as determinants of wages. We found that education and age have a positive effect on wages. In regards to marital status, once all factors are controlled for, singles earned the lowest wages, followed by married and divorced individuals earned the highest wages.

The specification test, however, suggest that the model is misspecified with heteroskedastic errors. This means that the results should be interpreted with caution, and that other specifications should be considered.