Analysing Wage Data

Adrija Bhar, Srijani Das, Yenisi Das
Instructor: Soham Sarkar

M. Stat.
Indian Statistical Institute

November 13, 2023

Data Description

We have data on wage, year, age, race, education and jobclass for a group of 3000 male workers in the Mid-Atlantic region.

- **year**: Year that wage information was recorded (**b**).
- **age**: Age of the worker (**a**).
- **race**: A factor with levels 1. White ($\mathbf{x_1}$), 2. Black ($\mathbf{x_2}$), 3. Asian ($\mathbf{x_3}$), and 4. Other ($\mathbf{x_4}$).
- **education**: A factor with levels 1. $<$ HS Grad ($\mathbf{z_1}$), 2. HS Grad ($\mathbf{z_2}$), 3. Some College ($\mathbf{z_3}$), 4. College Grad ($\mathbf{z_4}$), and 5. Advanced Degree ($\mathbf{z_5}$).
- **jobclass**: A factor with levels 1. Industrial ($\mathbf{j_1}$) and 2. Information ($\mathbf{j_2}$).
- **wage**: Worker's raw wage (**W**).

Questions

We will try to answer the following questions by testing different hypotheses:

- Is the effect of education on wage influenced by the levels of race?
- Does wage depend on education level?
- Does communism prevail in America?
  - Does there exist a discrimination of wage based on race?
  - Does there exist a discrimination of wage based on jobclass?

Proposed Model

We start with our proposed model given below:

$$W_i = \alpha + \sum_{j=2}^{5} e_j x_{ji} + j_2 y_{2i} + \sum_{k=2}^{4} r_k z_{ki} + \sum_{j=2}^{5}\sum_{k=2}^{4} \theta_{jk} x_{ji} z_{ki} + \gamma_1 a_i + \gamma_2 b_i + \epsilon_i \quad ; i = 1, .., 3000$$

Also note we can rewrite the model in vector matrix notation as:

$$\underline{W} = Z\underline{\delta} + \mathbf{X}\underline{\beta} + \underline{\epsilon}$$

Assumptions:

- The errors are normally distributed with mean 0.
- The errors are homoscedastic with common variance σ^2 (unknown).
- ϵ_i's are uncorrelated.

Checking whether the continuous covariates are significant

To check whether continuous the covariates are significant or not we test the following hypothesis:

$$H_0 : \begin{bmatrix}\mathbf{0}_{2\times 21} & \mathbf{I}_{2\times 2}\end{bmatrix}\begin{bmatrix}\beta \\ \gamma\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix} \text{ against } H_1 : \text{ Not } H_0$$

$$\equiv H_0 : \mathbf{C}\phi = \mathbf{0} \text{ against } H_1 : \text{ Not } H_0$$

The test statistic is given by

$$\frac{(SSE - SSE_{H_0})/2}{SSE/(3000-23)}$$

which follows $F_{2,2977}$ distribution under null.

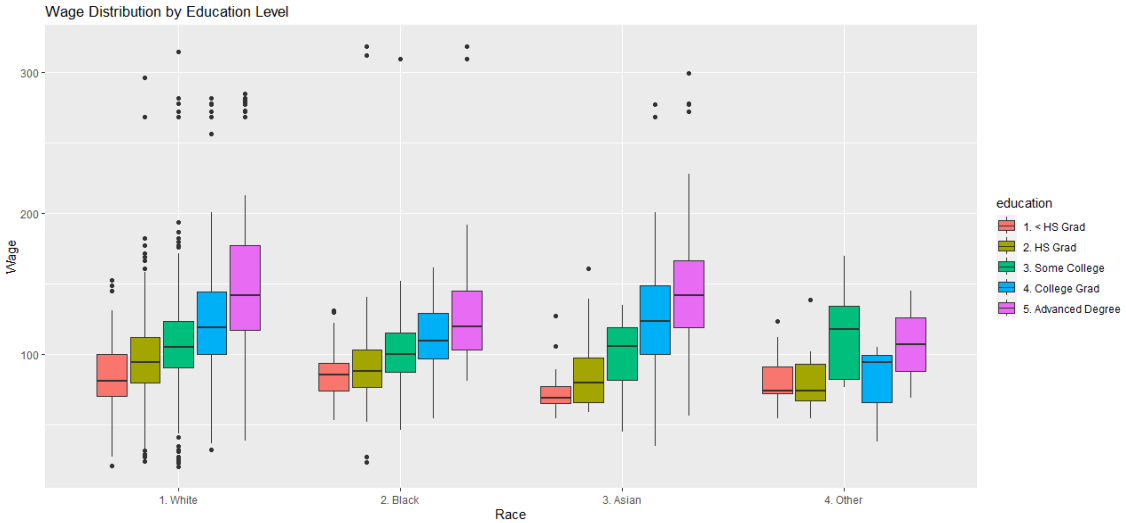
Checking whether the continuous covariates are significant

We have performed the above mentioned test and observed the below results:

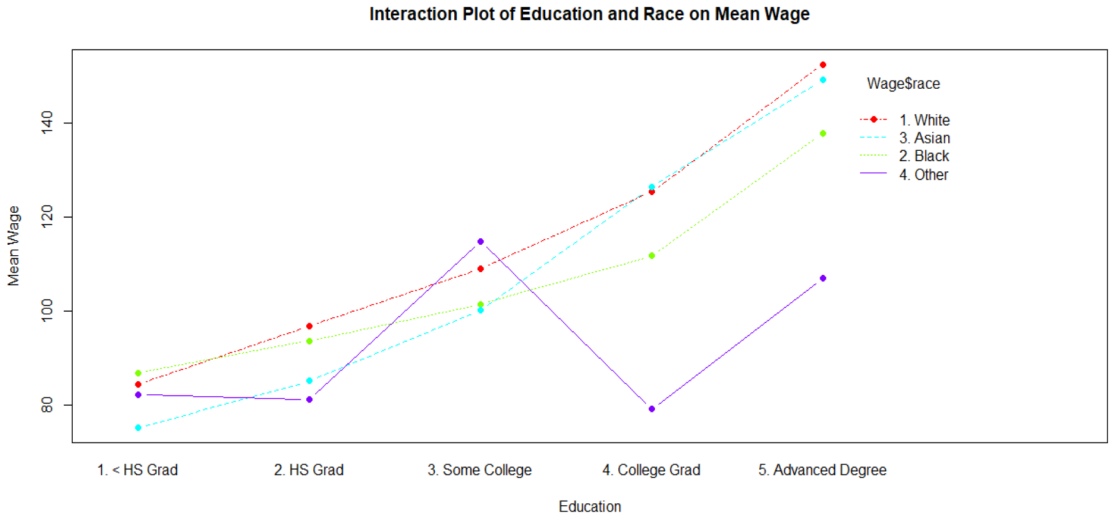
```
1
2           Test of General Linear Hypothesis
3  Call:
4  glh.test(reg = model, cm = C, d = d)
5  F = 1.9369, df1 =    2, df2 = 2977, p-value = 0.1443
```

Since we have a large p-value we fail to reject the null hypothesis at 0.05 level of significance. Hence, we proceed without the continous (concomitant) variables. Therefore, we shall go for Analysis of Variance (ANOVA) model.

Is the effect of education on wage influenced by the levels of race?



Is the effect of education on wage influenced by the levels of race?



Is the effect of education on wage influenced by the levels of race?

We are to test if there is no interaction between the levels of education and the different races under consideration, i.e., we are to test: $H_0 : \theta_{jk} = 0 \ \forall \ j = 2, .., 5, k = 2, .., 4$ against H_1 : not H_0

```
Analysis of Variance Table

Response: wage
                  Df  Sum Sq Mean Sq  F value    Pr(>F)
education          4 1226364  306591 232.0629 < 2.2e-16 ***
jobclass           1   20273   20273  15.3448 9.158e-05 ***
race               3   20389    6796   5.1443  0.001508 **
education:race    12   19337    1611   1.2197  0.262595
Residuals       2979 3935722    1321
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
       1
```

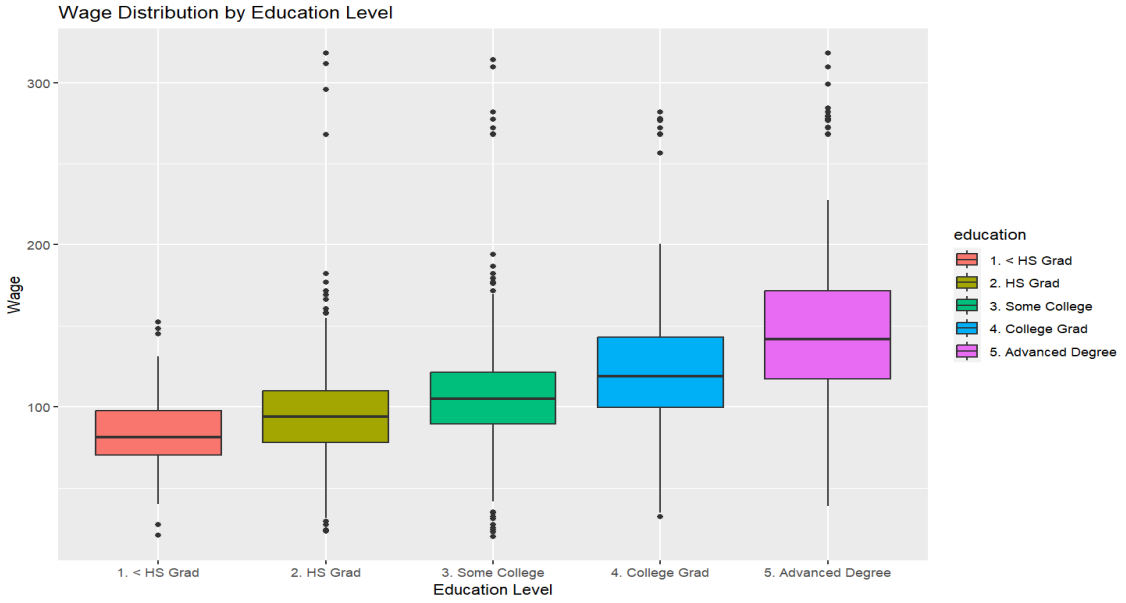
From the table, we conclude that θ_{jk}'s are not significant. Thus, we shall consider the model without interaction terms for answering the rest of the questions.

New Model

Model2:

$$W_i = \alpha + \sum_{j=2}^{5} e_j x_{ji} + j_2 y_{2i} + \sum_{k=2}^{4} r_k z_{ki} + \epsilon_i \quad ; i = 1, .., 3000$$

Does wage depend on education level?



Does wage depend on education level?

We are to test H_0 : $\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ j_2 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ against H_1 : not H_0. The test statistic is given by

$$\frac{(SSE - SSE_{H_0})/4}{SSE/(3000 - 9)}$$

which follows $F_{4,2991}$ distribution under null.

Does wage depend on education level?

We have performed the above mentioned test and observed the below results:

```
1
2           Test of General Linear Hypothesis
3 Call:
4 glh.test(reg = model2, cm = C1, d = d1)
5 F = 183.0594, df1 =    4, df2 = 2991, p-value = < 2.2e-16
```

Since we have very small p-value, in fact p-value $= 2.2e^{-16} < 0.05$, we reject the null hypothesis at 0.05 level of significance. Hence, in the light of the given data, it seems that there is significant difference between the wages for different education levels, *i.e.*, wage seems to be dependent on education level.

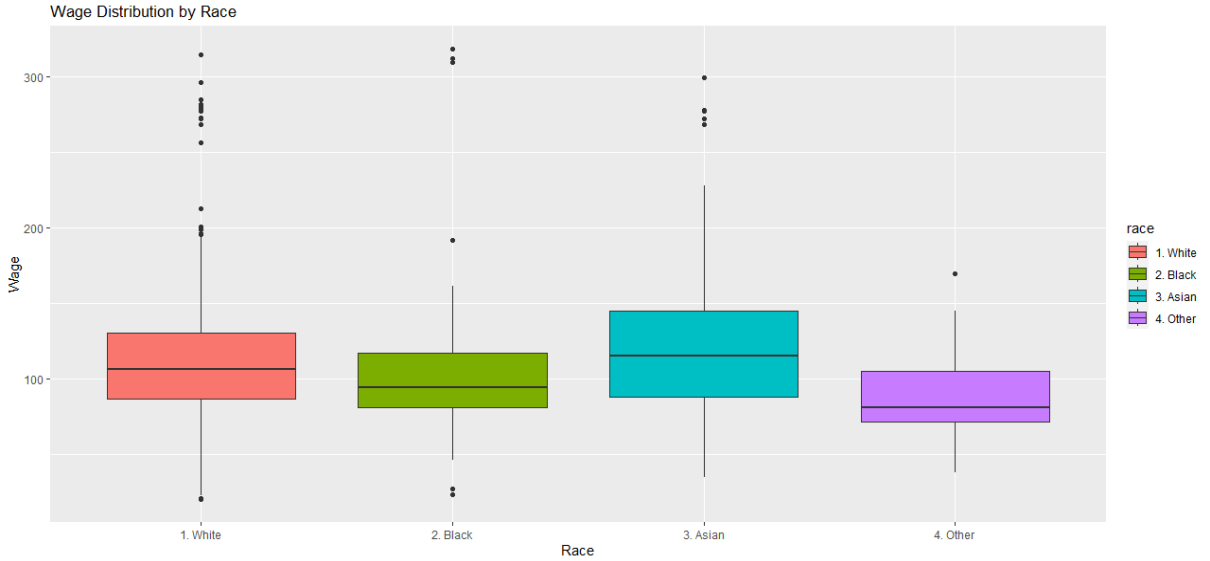
Does communism prevail in America?

To test whether communism prevails in America, we shall see whether :

- Wage depends on race
- Wage depends on jobclass

Does communism prevail in America?

Does there exist a discrimination of wage based on race?



We are to test if wage changes on the basis of different races.

Thus H_0: $\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ j_2 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ against H_1 : not H_0.

The test statistic is given by

$$\frac{(SSE - SSE_{H_0})/3}{SSE/(3000 - 9)}$$

which follows $F_{3,2991}$ distribution under null.

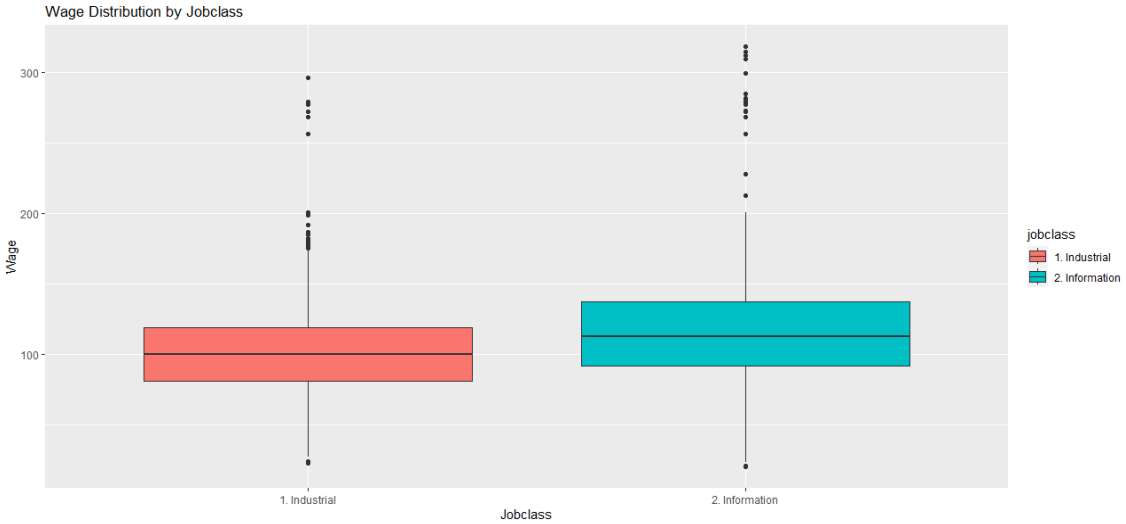
We have performed the above mentioned test and observed the below results:

```
1
2           Test of General Linear Hypothesis
3 Call:
4 glh.test(reg = model2, cm = my_matrix, d = d2)
5 F = 5.1398, df1 =    3, df2 = 2991, p-value = 0.001517
```

Since we have very small p-value, in fact p-value $= 0.001517 < 0.05$, we reject the null hypothesis at 0.05 level of significance. Hence, in the light of the given data, it seems that there is significant difference between the wages for different races.

Does communism prevail in America?

Does there exist a discrimination of wage based on jobclass?



We are to test H_0: $\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ j_2 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = 0$ against H_1 : not H_0.

The test statistic is given by

$$\frac{(SSE - SSE_{H_0})/1}{SSE/(3000 - 9)}$$

which follows $F_{1,2991}$ distribution under null.

```
1
2           Test of General Linear Hypothesis
3 Call:
4 glh.test(reg = model2, cm = my_matrix1, d = d3)
5 F = 18.194, df1 =    1, df2 = 2991, p-value = 2.057e-05
```

Since we have very small p-value, in fact p-value $2.057e-05 < 0.05$, we reject the null hypothesis at 0.05 level of significance. Hence, in the light of the given data, it seems that there is significant difference between the wages for different jobclass.
Thus from the boxplot, we can conclude that the median wage of people working in Information Technology is slightly higher than those working in Industry.

THANK YOU!
ANY QUESTIONS?