(Fast) Introduction to R - Class 2

Jump into a notebook

Joana Cima

05 dezembro 2023

My beamer

BlaBlaBla

Outline

1. Motivation
2. Data
3. Conceptual discussion

3. Import data (from an excel file)

```
nlswork <- as.data.frame(read_excel("nlswork.xlsx"))
```

3.1 Drop missing values

```
nlswork_no_na <- drop_na(nlswork)
```

4. Descriptive statistics

(. . .)

5. Regression analysis

5.1 Regression analysis: OLS

5.1.1. Variable Selection

Selecting appropriate variables for our model is critical to derive accurate and meaningful results.

```
##            Overall
## age      12.564946
## collgrad 40.420145
## union    24.539365
## hours     6.894666
```

```
## Start:  AIC=-23546.23
## ln_wage ~ age + collgrad + union + hours
```

```
##
##            Df Sum of Sq    RSS    AIC
## <none>                   2335.0 -23546
## - hours     1     8.254 2343.2 -23501
## - age       1    27.414 2362.4 -23391
## - union     1   104.564 2439.5 -22959
## - collgrad  1   283.694 2618.7 -22006
```

```
##
## Call:
## lm(formula = ln_wage ~ age + collgrad + union + hours, data = db_ols)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2458 -0.2650 -0.0113  0.2544  3.4018
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.2878849  0.0226062  56.971  < 2e-16 ***
## age         0.0071387  0.0005681  12.565  < 2e-16 ***
## collgrad    0.3774243  0.0093375  40.420  < 2e-16 ***
## union       0.2115787  0.0086220  24.539  < 2e-16 ***
## hours       0.0024992  0.0003625   6.895 5.64e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4167 on 13447 degrees of freedom
## Multiple R-squared:  0.178,  Adjusted R-squared:  0.1777
## F-statistic: 727.9 on 4 and 13447 DF,  p-value: < 2.2e-16
```

5.1.2. Our regression Table

Exemplo de texto.

Table 1: Regression analysis

	Model (1)	Model (2)
age	0.010***	0.007***
	(0.001)	(0.001)
collgrad		0.377***
		(0.009)
union		0.212***
		(0.009)
hours		0.002***
		(0.0004)
N	13,452	13,452
R^2	0.021	0.178

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Standard errors in parentheses.

5.1.3. HYPOTHESIS TESTING: automatic command

Next, we will be testing specific hypotheses about the coefficients in our regression model.

1. **Testing if the Coefficient for `age` is Zero:**
   We aim to test the null hypothesis that:

$$H_0 : \beta_{\text{age}} = 0 \tag{1}$$

   This tests if `age` has any influence on the dependent variable, after adjusting for other variables in the model.

2. **Testing if the Coefficients for `union` and `collgrad` are Equal:**
   We will test the null hypothesis that:

$$H_0 : \beta_{\text{union}} = \beta_{\text{collgrad}} \tag{2}$$

   This checks if the effect of being in a union on the dependent variable is the same as the effect of being a college graduate, when other factors are held constant.

```
## Linear hypothesis test
##
## Hypothesis:
## - collgrad  + union = 0
##
## Model 1: restricted model
## Model 2: ln_wage ~ age + collgrad + union + hours
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  13448 2362.5
## 2  13447 2335.0  1    27.529 158.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.1.4. Additional quality measures: AIC & BIC

In both the AIC and BIC criteria, the model with the lower values is favored as it suggests a better balance between model fit and model complexity.

```
## [1] 16981.05
```

```
## [1] 17003.57
```

```
## [1] 14630.89
```

```
## [1] 14675.94
```

5.1.5. COLINEARITY: VIF

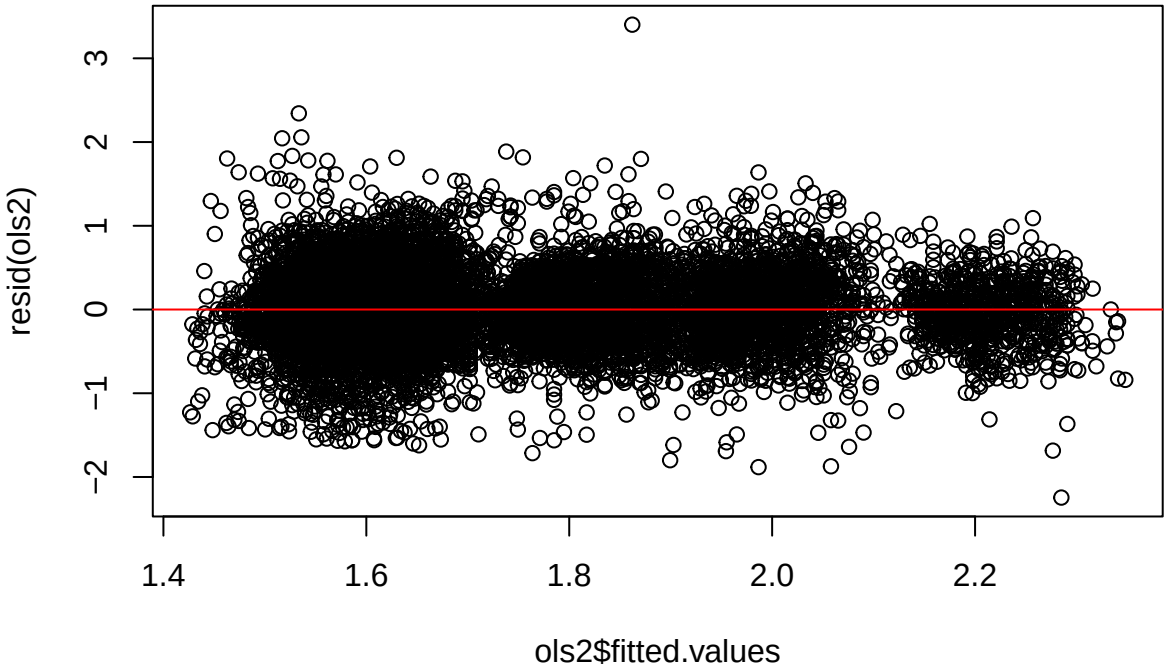
The Variance Inflation Factor (VIF) assesses the severity of multicollinearity in a regression, with values greater than 10 suggesting high correlation between predictors (Wooldrige; Verbeek).

```
##      age collgrad    union    hours
## 1.028617 1.034259 1.015531 1.024762
```

5.1.6. HETEROSKEDASTICITY

Homoscedasticity is a fundamental assumption underlying standard linear regression models, stipulating that the variance of the residuals remains constant across levels of the independent variables. This property ensures that the ordinary least squares (OLS) estimator remains the best linear unbiased estimator (BLUE), providing

minimum variance. Violations of homoscedasticity, known as heteroscedasticity, can lead to inefficient and potentially biased coefficient estimates, as well as unreliable standard errors. To rigorously assess the presence of homoscedasticity, researchers often employ diagnostic tests, such as the Breusch-Pagan test.



Graphical analysis

Breusch-Pagan test

```
##
##  studentized Breusch-Pagan test
##
## data:  ols2
## BP = 284.11, df = 4, p-value < 2.2e-16
```

The results from the Breusch-Pagan test for the `ols2` model suggest the presence of heteroscedasticity. The test statistic is $BP = 199.54$ with a degree of freedom (df) of 4. The low p-value, effectively zero at $p < 2.2e-16$, leads us to reject the null hypothesis of homoscedasticity.

5.1.7. Robust estimation

```
robust_ols2 <- coeftest(ols2, vcov. = vcovHC(ols2, type="HC1"))
print(robust_ols2)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 1.28788491 0.02783552 46.2677 < 2.2e-16 ***
```

```
## age         0.00713867 0.00058433 12.2168 < 2.2e-16 ***
## collgrad    0.37742427 0.00956496 39.4590 < 2.2e-16 ***
## union       0.21157873 0.00842208 25.1219 < 2.2e-16 ***
## hours       0.00249920 0.00053996  4.6285 3.718e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2: Comparison of OLS and Robust Regression Models

	ln_wage	
	OLS	Robust OLS
	(1)	(2)
Constant	1.288***	1.288***
	(0.023)	(0.028)
age	0.007***	0.007***
	(0.001)	(0.001)
collgrad	0.377***	0.377***
	(0.009)	(0.010)
union	0.212***	0.212***
	(0.009)	(0.008)
hours	0.002***	0.002***
	(0.0004)	(0.001)
N	13,452	
R^2	0.178	
Adjusted R^2	0.178	
Residual Std. Error	0.417 (df = 13447)	
F Statistic	727.942*** (df = 4; 13447)	
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	

5.1.8. Coefficient interpretation with a full model (Model 2)

Given that our dependent variable is the natural logarithm of the salary, the coefficients should be interpreted accordingly. For every additional year in age, the salary is expected to increase by approximately 0.7024557%, holding all other factors constant. Being a college graduate is associated with an estimated increase of about 45.7904309% in the salary, compared to not being a college graduate. Being a member of a union is associated with an approximate 23.6147885% increase in salary, holding everything else constant. All variables are significant at the 1% level.

5.2. Binary choice models

When the dependent variable is binary, ordinary least squares regression is not suitable due to the non-continuous nature of the outcome variable. Models such as Logit and Probit are specifically designed to handle such binary outcomes. In our analysis, where we aim to understand the probability of a worker being unionized, these models are appropriate choices as they provide insights into the factors influencing this

binary decision.These models show the direction of the relationship between independent variables and the dependent variable but don't quantify the magnitude.

```
##
## Call:
## glm(formula = union ~ ., family = binomial(link = "probit"),
##     data = db_ols)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.467689   0.091163 -27.069   <2e-16 ***
## age         -0.001280   0.001970  -0.650    0.516
## ln_wage      0.734203   0.030386  24.162   <2e-16 ***
## collgrad    -0.032086   0.032593  -0.984    0.325
## hours        0.012740   0.001326   9.608   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14463  on 13451  degrees of freedom
## Residual deviance: 13640  on 13447  degrees of freedom
## AIC: 13650
##
## Number of Fisher Scoring iterations: 4
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: ter, dez 05, 2023 - 21:45:30

5.2.1. Marginal effects

Marginal effects are crucial in non-linear models like Logit and Probit. While the model's coefficients tell us about the direction of effects, they don't show the actual change in probability for a one-unit change in the predictor. Marginal effects provide this information, making it easier to understand the real-world impact of each variable on the outcome.

5.2.2. Marginal effects - probit

```
##    factor     AME     SE       z      p   lower  upper
##       age -0.0004 0.0006 -0.6497 0.5159 -0.0015 0.0007
##  collgrad -0.0091 0.0093 -0.9845 0.3249 -0.0273 0.0090
##     hours  0.0036 0.0004  9.6734 0.0000  0.0029 0.0044
##   ln_wage  0.2089 0.0082 25.4688 0.0000  0.1928 0.2250
```

5.2.3. Marginal effects - logit

```
##    factor     AME     SE       z      p   lower  upper
##       age -0.0005 0.0006 -0.9483 0.3430 -0.0016 0.0006
##  collgrad -0.0088 0.0091 -0.9667 0.3337 -0.0266 0.0090
##     hours  0.0037 0.0004  9.6196 0.0000  0.0030 0.0045
##   ln_wage  0.2076 0.0083 25.1307 0.0000  0.1914 0.2238
```

5.2.4. Summary of Marginal Effects for Logit Model on Union Membership:

- **Age:** A one-year increase in age is linked to a decrease in the likelihood of a worker being unionized by 0.05 percentage points. However, the effect is not statistically significant at the 5% level

Table 3: Comparative Table of Logit and Probit Models

	union	
	logistic	*probit*
	(1)	(2)
age	−0.003	−0.001
	(0.003)	(0.002)
ln_wage	1.253***	0.734***
	(0.053)	(0.030)
collgrad	−0.053	−0.032
	(0.055)	(0.033)
hours	0.022***	0.013***
	(0.002)	(0.001)
Constant	−4.167***	−2.468***
	(0.160)	(0.091)
N	13,452	13,452
Log Likelihood	−6,825.528	−6,820.092
Akaike Inf. Crit.	13,661.060	13,650.180

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

- **College Graduation (collgrad):** Being a college graduate reduces the probability of being in a union by 0.88 percentage points. However, the effect is not statistically significant at the 5% level
- **Hours:** A one-unit increase in hours is associated with a 0.37 percentage point increase in the likelihood of a worker being unionized.
- **Log of Wage (ln_wage):** A one-unit increase in the logarithm of wage is associated with a 20.76 percentage point increase in the likelihood of a worker being unionized.

6 Assessment

Problem 1: Data Importing

Import the "card" dataset.

```
card<-as.data.frame(read_excel("card.xlsx"))
```

Problem 2: Drop the missing data

```
#BEGIN SOLUTION

card_no_na <- na.omit(card)


#END SOLUTION
```

Problem 3

Estimate a linear regression model that uses the log of the salary as the dependent variable and IQ, married, age, educ, and the log of weight as independent variables.

```
#BEGIN SOLUTION


#END SOLUTION
```

Problem 4:

Which variables are the most important in the model? Explain

```
#BEGIN SOLUTION


#END SOLUTION
```

Problem 5: What can you conclude regarding homoscedasticity?

```
#BEGIN SOLUTION


#END SOLUTION
```

Problem 6:

Interpret the coefficients of the variables

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 7:

Create a binary variable that takes the value 1 if the salary is above the average and 0 otherwise

```
average_wage <- mean(card_no_na$wage)

# Create a binary variable (high wage): 1 if wage is above average, 0 otherwise
card_no_na <- card_no_na %>%
  mutate(high_wage = ifelse(wage > average_wage, 1, 0))
```

Problem 8:

Estimate a logit model to explain the probability of an individual having a salary above the average, using the same independent variables as in the linear regression mode.

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 9:

Discuss how the independent variables are positively/negatively related to the probability of the salary being above the average.

```
#BEGIN SOLUTION

#END SOLUTION
```

Problem 10:

Compute the marginal effects of the logit model

```
#BEGIN SOLUTION

#END SOLUTION
```