(Fast) Introduction to R - Class 2

Jump into a notebook

Joana Cima

13 outubro 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      11.92846
## collgrad 41.15039
## union    25.20145
```

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		0.383***
		(0.009)
union		0.217***
		(0.009)
N	13,452	13,452
R^2	0.021	0.175
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:
## age = 0
##
## Model 1: restricted model
## Model 2: ln_wage ~ age + collgrad + union
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  13449 2368.0
## 2  13448 2343.2  1    24.793 142.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Linear hypothesis test
##
## Hypothesis:
## - collgrad  + union = 0
##
## Model 1: restricted model
## Model 2: ln_wage ~ age + collgrad + union
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  13449 2371.0
## 2  13448 2343.2  1    27.743 159.22 < 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] 14676.36
```

```
## [1] 14713.9
```

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
## 1.018819 1.025643 1.007647
```

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.

Breusch-Pagan test

```
##
##  studentized Breusch-Pagan test
##
## data:  ols2
## BP = 220.76, df = 3, 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

t test of coefficients:

```
         Estimate Std. Error t value  Pr(>|t|)
```

(Intercept) 1.38759270 0.01699578 81.643 < 2.2e-16 ***age 0.00675636 0.00057968 11.655 < 2.2e-16*** collgrad 0.38330049 0.00950024 40.346 < 2.2e-16 ***union 0.21681639 0.00828751 26.162 < 2.2e-16*** — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: sex, out 13, 2023 - 10:38:15

Table 2:

age	0.0068***
	(0.0006)
collgrad	0.3833***
	(0.0095)
union	0.2168***
	(0.0083)
Constant	1.3876***
	(0.0170)
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 in this regression model represent percentage changes in the salary for a one-unit change in the independent variables.For every additional year in age, the salary is expected to increase by approximately 0.6779236%, holding all other factors constant. Being a college graduate is associated with an estimated increase of about 46.7118818% in the salary, compared to not being a college graduate. Being a member of a union is associated with an approximate 24.2116016% increase in salary, holding everything else constant. This effect is highly significant (p-value < 2e-16).

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) -1.963478   0.073502 -26.713   <2e-16 ***
## age         -0.002860   0.001959  -1.459    0.144
## ln_wage      0.738308   0.030169  24.472   <2e-16 ***
## collgrad    -0.001693   0.032350  -0.052    0.958
## ---
## 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: 13733  on 13448  degrees of freedom
## AIC: 13741
##
## Number of Fisher Scoring iterations: 4
```

Table 3: Regression

	Probit	Logit
age	−0.0029	−0.0055
	(0.0020)	(0.0034)
ln_wage	0.7383***	1.2492***
	(0.0302)	(0.0524)
collgrad	−0.0017	−0.0025
	(0.0323)	(0.0545)
N	13,452	13,452
Log Likelihood	−6,866.5360	−6,873.1680
Akaike Inf. Crit.	13,741.0700	13,754.3400

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

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 -0.0008 0.0006 -1.4596 0.1444 -0.0019 0.0003
##  collgrad -0.0005 0.0093 -0.0523 0.9583 -0.0186 0.0177
##   ln_wage  0.2114 0.0082 25.8419 0.0000  0.1954 0.2274
```

5.2.3. Marginal effects - logit

```
##    factor     AME     SE       z      p   lower  upper
##       age -0.0009 0.0006 -1.6289 0.1033 -0.0020 0.0002
##  collgrad -0.0004 0.0091 -0.0458 0.9635 -0.0182 0.0174
##   ln_wage  0.2082 0.0082 25.2995 0.0000  0.1921 0.2244
```

5.2.4. Summary of Marginal Effects for Probit 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 0.08 percentage points. This effect, though statistically significant, is minimal, indicating age may have only a slight influence on union membership.

- **College Graduation (collgrad):** Being a college graduate reduces the probability of being in a union by 0.05 percentage points. The small magnitude of this effect suggests that in this context, educational attainment (specifically having a college degree) has limited influence on union membership.
- **Log of Wage (ln_wage):** A one-unit increase in the logarithm of wage is associated with a 21.14 percentage point increase in the likelihood of a worker being unionized. This significant effect underscores the importance of wage in a worker's decision to join a union.

6. Machine Learning: Linear Regression vs. Random Forest

In this section, we will train both a traditional linear regression model and a Random Forest model to predict `ln_wage` and then compare their performance on a test dataset.

6.1. Splitting Data into Training and Testing Sets

Before we train our models, we need to partition our dataset into a training set, used to train the models, and a test set, used to evaluate their performance.

6.2. Training and Evaluating Linear Regression

```
## [1] 0.3517657
```

6.3. Training and Evaluating Random Forest

```
## [1] 0.3137955
```

6.4. Comparing Model Performances

With the results in hand, let's compare the performance of the two models. RMSE provides a measure of the magnitude of the prediction errors. Lower values of RMSE indicate a better fit of the model to the data.

```
##               Model      RMSE
## 1 Linear Regression 0.3517657
## 2     Random Forest 0.3137955
```

7. 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<-drop_na(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, and educ 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
```