# The use of Uber

Guillaume BONIN, Pierre ETAIX, Xavier VALENDUC

ENAC, IENAC15 PREV

April 7, 2017

# Table of contents

# Motivations

- How much time people are taking Uber a month ?
- What brought people to take Uber ?
- Where people are taking Uber ?
- With whom people are taking this sharing service ?

# Table of contents

# Our objectives

- Gathering data about the use of Uber
- Estimate with whom and when someone takes Uber
- Create an Econometrics model about Uber
- Try to answer our problematic

# Table of contents

# The survey

## Database

- Use of a Google Form
- n = 267 answers collected in 6 days
- Majority of students (75%)
- A total of k = 31 variables

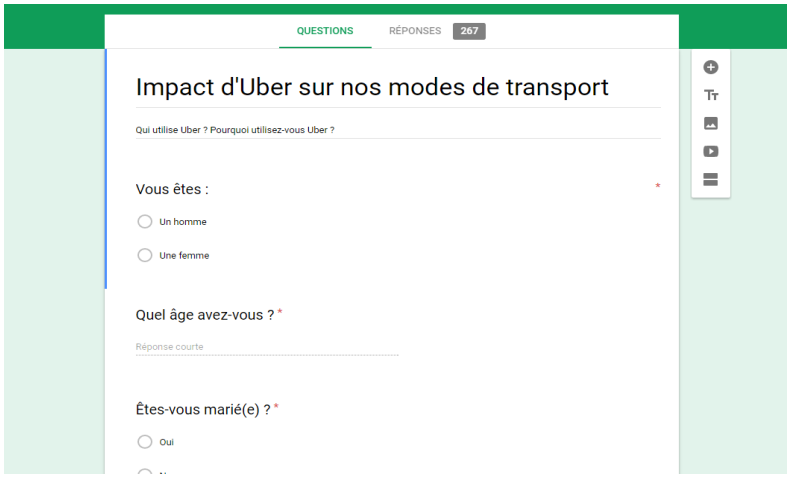


Figure 1: Our survey.

# Table of contents

# Age

- Majority of young adults (median = 22)
- More likely to take Uber
- Model might be biaised

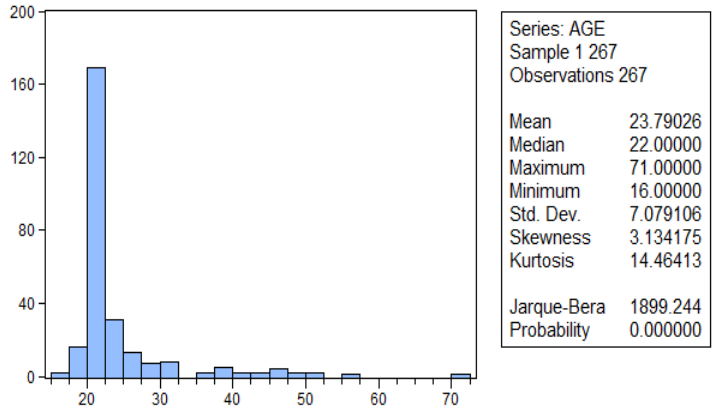


Figure 2: Age histogram.

# Occupation

- People had to categorize them between 5 choices (1 = Student, 2 = Executive, 3 = Employee, 4 = Jobless and 5 = Other)
- Majority of students

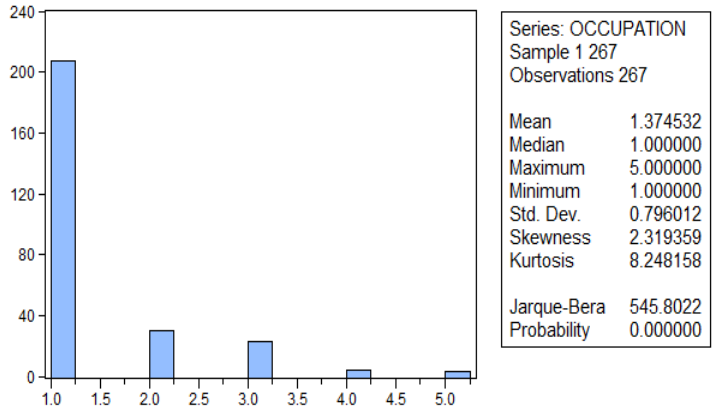


Figure 3: Occupation histogram.

# Region

- People had to categorize them between 5 choices (1 = West, 2 = East, 3 = Parisian, 4 = North and 5 = Foreign)
- Strong representation of the population of France

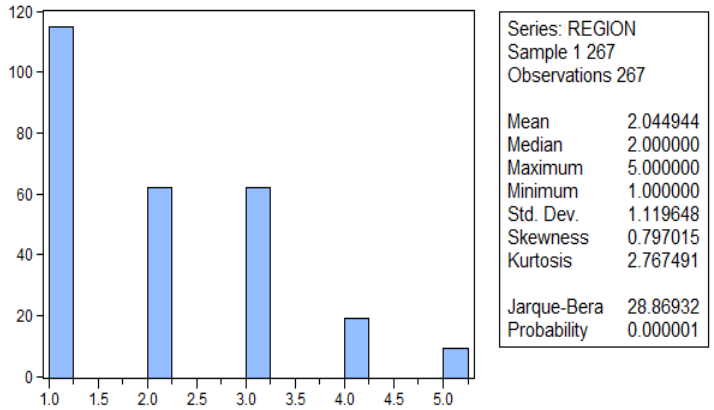


Figure 4: Region histogram.

# Table of contents

# How would you describe a Uber user ?

- How often do they take Uber per month ? $\Rightarrow$ **UBER_FREQ**
- Where are they taking Uber ? $\Rightarrow$ **LOCATION**
- How much kilometers traveled during a Uber fare ? $\Rightarrow$ **DISTANCE**
- When are they taking Uber ? $\Rightarrow$ **TIME_OF_DAY**
- Why are they taking Uber ? $\Rightarrow$ **PRIVATE_CAUSE**

# Table of contents

# Variables for the Uber frequency model

## How do we choose the variables :

- Analysis of the correlation matrix
- Suppression of some variables to avoid multicollinearity
- Non-significant variables (with p $>$ 0.1)
- Try to increase $\overline{R^2}$

# Analysis of the correlation matrix

| **Variable** | **Correlated with** | **Coefficient** |
|---|---|---|
| Age | Income | 0.53 |
| - | Married | 0.61 |
| - | Occupation | 0.54 |
| Public transport frequency | Region | 0.43 |
| - | Vehicle | -0.51 |
| Smartphone | Children | -0.68 |
| - | Taxi frequency | -0.39 |
| Income | Occupation | 0.58 |

Figure 5: Main significant coefficients in the correlation matrix.

# Choice of our dependent variable

## Dependent variable

- Previously, we would choose the dummy variable USE_UBER (0 or 1) for our Y, however the model was not good enough.
- Consequently, UBER_FREQ is our most likely dependent variable to answer at our problematic.

# Table of contents

# First linear regression

## Observations

- $\overline{R^2}$=0.34
- Prob(F-stat) = 0
- A lot of insignificant variables

Dependent Variable: UBER_FREQ
Method: Least Squares
Date: 04/04/17 Time: 18:06
Sample: 1 267
Included observations: 267

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.523646 | 1.181974 | 1.289069 | 0.1986 |
| AGE | -0.027178 | 0.015151 | -1.793798 | 0.0741 |
| CHILDREN | 0.135198 | 0.148979 | 0.907495 | 0.3650 |
| GENDER | 0.331845 | 0.117361 | 2.827555 | 0.0051 |
| INCOME=1 | 0.199093 | 0.421208 | 0.472672 | 0.6369 |
| INCOME=2 | 0.280356 | 0.422624 | 0.663371 | 0.5077 |
| INCOME=3 | 0.488857 | 0.354367 | 1.379521 | 0.1690 |
| LOCATION=1 | 0.317679 | 0.378943 | 0.838328 | 0.4027 |
| LOCATION=2 | 0.077762 | 0.369377 | 0.210521 | 0.8334 |
| LOCATION=3 | 0.283178 | 0.369193 | 0.767019 | 0.4438 |
| LOCATION=4 | -0.487645 | 0.432204 | -1.128274 | 0.2603 |
| MARRIED | -0.048873 | 0.392795 | -0.124424 | 0.9011 |
| OCCUPATION=1 | -0.079836 | 0.597620 | -0.133589 | 0.8938 |
| OCCUPATION=2 | -0.449610 | 0.560686 | -0.801893 | 0.4234 |
| OCCUPATION=3 | -0.169349 | 0.578729 | -0.292622 | 0.7701 |
| OCCUPATION=4 | -1.000514 | 0.723862 | -1.382189 | 0.1682 |
| PT_FREQ | 0.098597 | 0.051363 | 1.919599 | 0.0561 |
| REGION=1 | -0.760711 | 0.330812 | -2.299525 | 0.0223 |
| REGION=2 | -0.873080 | 0.326000 | -2.678161 | 0.0079 |
| REGION=3 | -0.285515 | 0.331097 | -0.862332 | 0.3894 |
| REGION=4 | -0.872315 | 0.371186 | -2.350074 | 0.0196 |
| SMARTPHONE | -0.035913 | 0.712523 | -0.050403 | 0.9598 |
| TAXI_FREQ | 0.464743 | 0.097316 | 4.775616 | 0.0000 |
| VEHICLE | 0.040973 | 0.133471 | 0.306983 | 0.7591 |

| | | | |
|---|---|---|---|
| R-squared | 0.393097 | Mean dependent var | 0.988764 |
| Adjusted R-squared | 0.335654 | S.D. dependent var | 1.063697 |
| S.E. of regression | 0.866993 | Akaike info criterion | 2.638015 |
| Sum squared resid | 182.6573 | Schwarz criterion | 2.960465 |
| Log likelihood | -328.1751 | Hannan-Quinn criter. | 2.767541 |
| F-statistic | 6.843198 | Durbin-Watson stat | 1.767095 |
| Prob(F-statistic) | 0.000000 | | |

Figure 6: EQ01.

# Table of contents

# Model analysis

- Linear regression : $Y = X\beta + u$
- F-statistic $= 6.84 > F_{0.01}(23; 267 - 23) = 2.26$
- So we do reject the null, it means that our model explains something statistically

## Observations

- Too many insignificant variables ?
- Surprising positive effect of taking a taxi($\widehat{\beta} = 0.47$)
- Gender has a positive marginal effect, we actually expect that result ($\widehat{\beta} = 0.33$)
- The marginal effect of different REGION is not the same. For example people are more likely to take Uber in Paris than in the provinces ($\widehat{\beta_{PARIS}} > \widehat{\beta_{PROVINCES}}$)
- $\overline{R^2}$=0.34 is good but actually it depends on the context !

## We need to improve the model !

- Reduce the number of insignificant variables
- Maybe improve $\overline{R^2}$ to find better model of who is taking Uber

# Table of contents

# The equation

## Linear equation of the model

$$\widehat{UBER\_FREQ} = \widehat{\beta_0} + \widehat{\beta_1} DISTANCE = 1 + \widehat{\beta_2} DISTANCE = 2 + \widehat{\beta_3} PASSENGERS = 1 + \widehat{\beta_4} PASSENGERS = 2 + \widehat{\beta_5} PRIVATE\_CAUSE + \widehat{\beta_6} TIME\_OF\_DAY = 1 + \widehat{\beta_7} TIME\_OF\_DAY = 3 + \widehat{\beta_8} QUALITY + \widehat{\beta_9} SECURITY + \widehat{\beta_{10}} SIMPLICITY + \widehat{\beta_{11}} WAITING\_TIME + \widehat{\beta_{12}} PRICE + \widehat{\beta_{13}} PAYMENT$$

# The linear regression

## Observations

- $\overline{R^2}$=0.13
- Prob(F-stat) = 0.001
- Many insignificant variables
- k = 13 variables

Dependent Variable: UBER_FREQ
Method: Least Squares
Date: 04/06/17 Time: 14:39
Sample: 1 161
Included observations: 158

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 1.279870 | 0.711425 | 1.799022 | 0.0741 |
| DISTANCE=1 | 0.114723 | 0.326118 | 0.351783 | 0.7255 |
| DISTANCE=2 | 0.241101 | 0.301770 | 0.798958 | 0.4256 |
| PASSENGERS=1 | 0.535499 | 0.216272 | 2.476048 | 0.0144 |
| PASSENGERS=2 | -0.011306 | 0.163972 | -0.068949 | 0.9451 |
| PRIVATE_CAUSE | -0.981251 | 0.384640 | -2.551087 | 0.0118 |
| TIME_OF_DAY=1 | -1.317948 | 0.366104 | -3.599930 | 0.0004 |
| TIME_OF_DAY=3 | 0.145824 | 0.214194 | 0.680805 | 0.4971 |
| QUALITY | -0.021853 | 0.053934 | -0.405181 | 0.6859 |
| SECURITY | 0.076133 | 0.042842 | 1.777062 | 0.0777 |
| SIMPLICITY | 0.148050 | 0.049941 | 2.964487 | 0.0035 |
| WAITING_TIME | -0.010114 | 0.044270 | -0.228459 | 0.8196 |
| PRICE | -0.026931 | 0.031356 | -0.858875 | 0.3918 |
| PAYMENT | -0.050313 | 0.041676 | -1.207231 | 0.2293 |

| | | | |
|---|---|---|---|
| R-squared | 0.204297 | Mean dependent var | 1.626582 |
| Adjusted R-squared | 0.132462 | S.D. dependent var | 0.899286 |
| S.E. of regression | 0.837610 | Akaike info criterion | 2.567906 |
| Sum squared resid | 101.0291 | Schwarz criterion | 2.839275 |
| Log likelihood | -188.8646 | Hannan-Quinn criter. | 2.678113 |
| F-statistic | 2.843998 | Durbin-Watson stat | 2.066918 |
| Prob(F-statistic) | 0.001150 | | |

Figure 7: EQ02.

# Table of contents

# Variable Observations

## Observations

- DISTANCE, who represents the distance of an Uber trip, does not really matter
- People are more likely to take Uber during night than during day ($\widehat{\beta_{NIGHT}} > 0 > \widehat{\beta_{DAY}}$)
- Regulars users of Uber are mainly satisfied by SECURITY and SIMPLICITY of the service.

# Analysis of the marginal effects

| Variable | Expected Effect | Real effect | Significance |
|---|---|---|---|
| DISTANCE=1 | - | + | No |
| DISTANCE=2 | + | + | No |
| PASSENGERS=1 | - | + | Yes |
| PASSENGERS=2 | + | + | No |
| PRIVATE_CAUSE | + | + | Yes |
| TIME_OF_DAY=1 | - | - | Yes |
| TIME_OF_DAY=3 | + | + | No |

Figure 8: Expected and real marginal effects(1).

# Analysis of the marginal effects

| Variable | Expected Effect | Real effect | Significance |
|---|---|---|---|
| QUALITY | + | - | No |
| SECURITY | + | + | Yes |
| SIMPLICITY | + | + | Yes |
| WAITING_TIME | + | - | No |
| PRICE | + | - | No |
| PAYMENT | + | - | No |

Figure 9: Expected and real marginal effects(2).

# Table of contents

# What about the errors ?

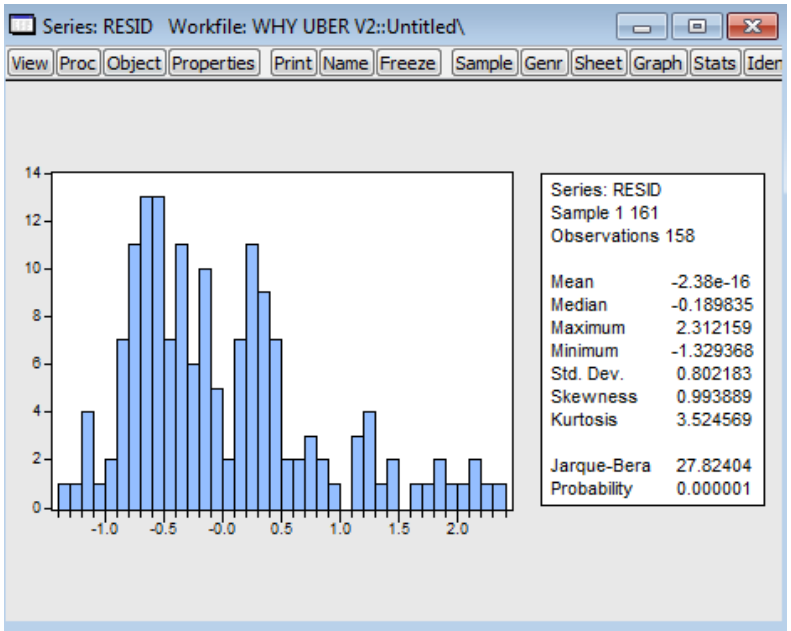


Figure 10: Graph of residuals.

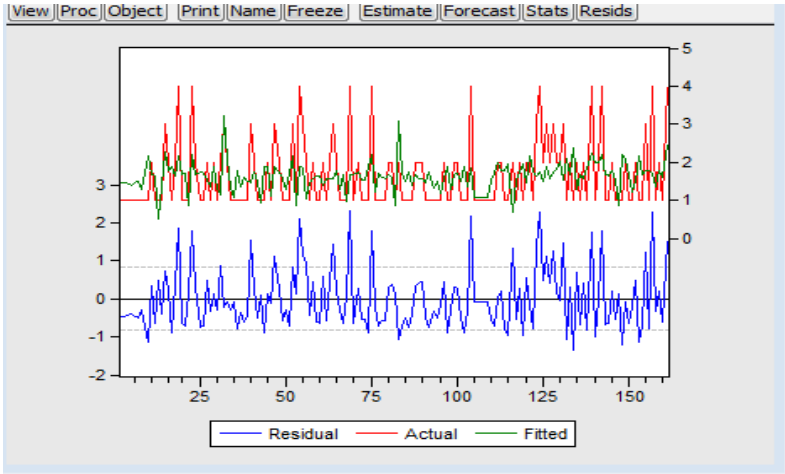


Figure 11: Repartition of the errors.

# Is there heteroscedasticity in our model ?

| Heteroskedasticity Test: White | | | |
|---|---|---|---|
| F-statistic | 1.005526 | Prob. F(77,80) | 0.4897 |
| Obs*R-squared | 77.70805 | Prob. Chi-Square(77) | 0.4560 |
| Scaled explained SS | 81.47682 | Prob. Chi-Square(77) | 0.3418 |

Figure 12: White's test of heteroscedasticity.

- $nR^2 = 77.70 < \chi^2_{0.95}(77) = 101.88$
- So we d.n.r the null
- We assume homoscedasticity : $Var(u) = \sigma^2 I_n$

# What happens with linearity ?

Ramsey RESET Test:

| | | | |
|---|---|---|---|
| F-statistic | 1.809886 | Prob. F(1,143) | 0.1807 |
| Log likelihood ratio | 1.987186 | Prob. Chi-Square(1) | 0.1586 |

Figure 13: Ramsey's test of linearity.

- We do not reject the null
- We assume linearity of the model : $Y = X\beta + u$

# Table of contents

# Conclusion : could we build a better model ?

## Negative points

- Our model is limited : We do not have an homogeneous sample (many students)
- Our survey is limited to 267 answers : there is not enough accuracy in our model
- We include many binary variables in our model and it could be problematic

# Conclusion : could we build a better model ?

## Positive points

- Our first expectations concerning the marginal effects of parameters are often verified by our model.
- Now, we can establish some characteristics of people who are taking Uber , and some reasons for why they are using Uber : **we have a meaningful model.**

# Model improvements

## First model

- We could reduce the model by using some hypotheses tests
- We could use White standard errors to take into account heteroscedasticity

## Second model

- We can use others explanatory variables to improve the quality of the model
- We could ask others questions in our study

# ANY QUESTIONS ?