MODS206 DATA ANALYSIS IN ECONOMICS: APPLIED ECONOMETRICS

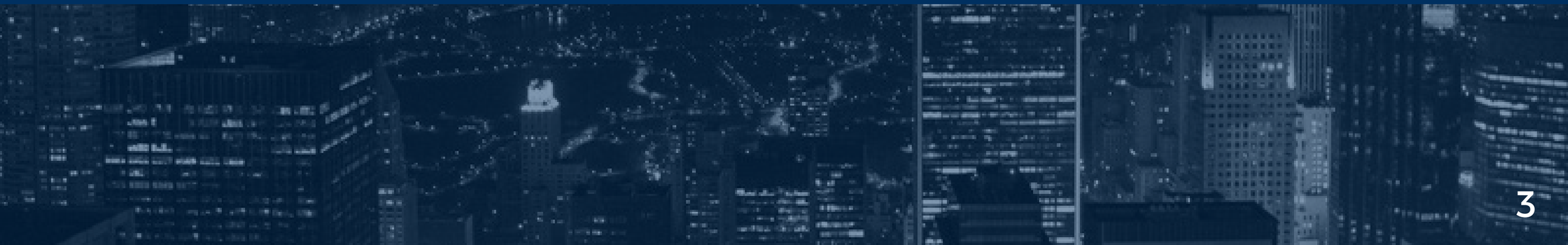
CREDIT CARD FRAUD

Lauryne MOYSE, Rafael MOUROUVIN, Farah JABRI

Summary

1. Recap from interim presentation
2. Empirical strategy
3. Tests and results
4. Conclusion

1. Recap from interim presentation



CONTEXT

DIGITAL PAYMENTS

The number of online payments is increasing significantly with the digitalisation of the economy, which also leads to an increase in cyber fraud.

FIGURES

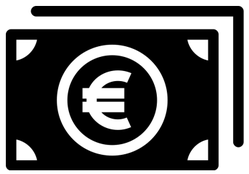
In 2020, there were on average 300 million transactions per day on the Visa card network alone worldwide., detection of fraud is challenging.

GLOBAL LOSSES

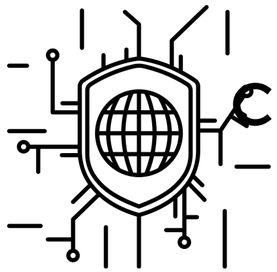
23.97 billion dollars in 2015 and 28.65 billion dollars in 2019

What about legislation ?

European Union’s Payment Services Directive 2 (PSD2)



Fair Credit Billing Act



Cyber Fusion Center

Research question

What is the impact of online transactions on credit card frauds ?

THE DATASET

distance_from_home - the distance from home where the transaction happened.

distance_from_last_transaction - the distance from last transaction happened.

ratio_to_median_purchase_price - Ratio of purchased price transaction to median purchase price.

repeat_retailer - Is the transaction happened from same retailer.

used_chip - Is the transaction through chip (credit card).

used_pin_number - Is the transaction happened by using PIN number.

online_order - Is the transaction an online order.

fraud - Is the transaction fraudulent.

Source : Kaggle

Description of the dataset

```
Contains data
 Observations:     1,000,000
    Variables:             8
```

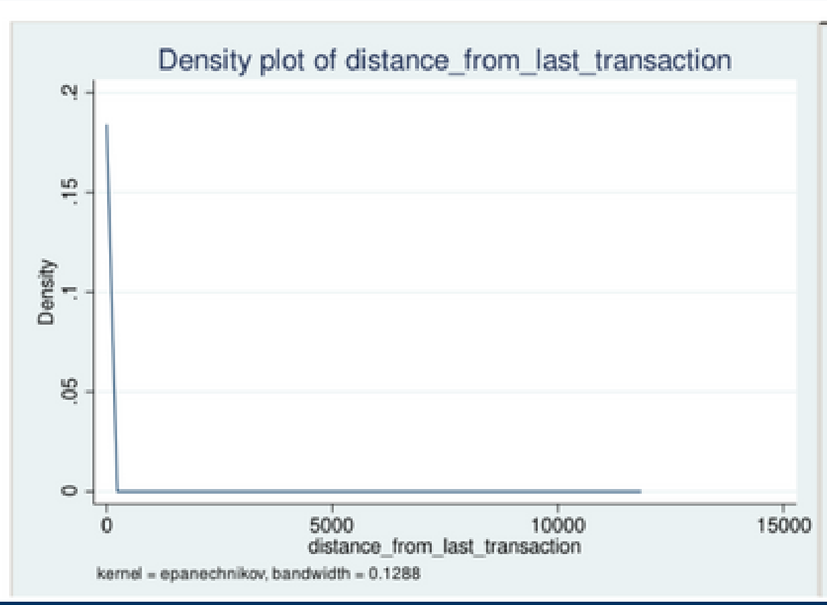
Variable name	Storage type	Display format	Value label	Variable label
distance_from~e	float	%9.0g		
distance_from~n	float	%9.0g		
ratio_to_medi~e	float	%9.0g		
repeat_retailer	byte	%8.0g		
used_chip	byte	%8.0g		
used_pin_number	byte	%8.0g		
online_order	byte	%8.0g		
fraud	byte	%8.0g		

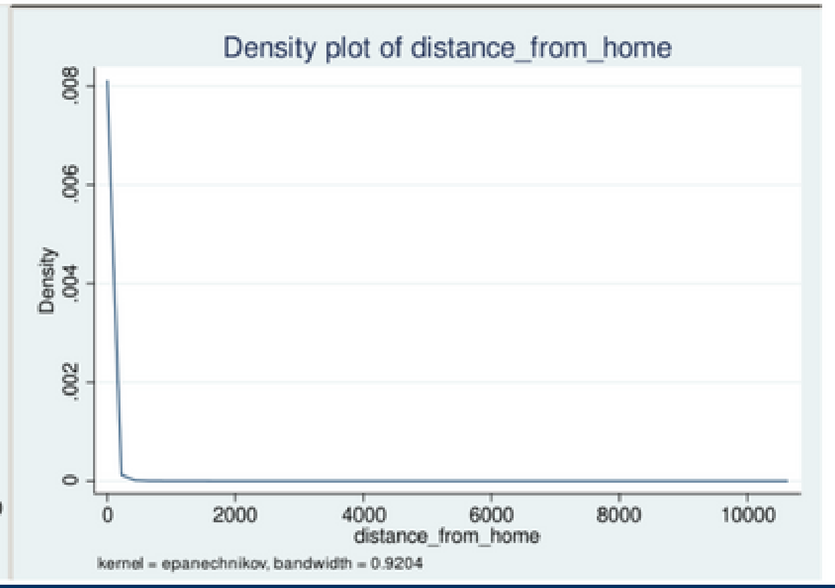
Description of the dataset

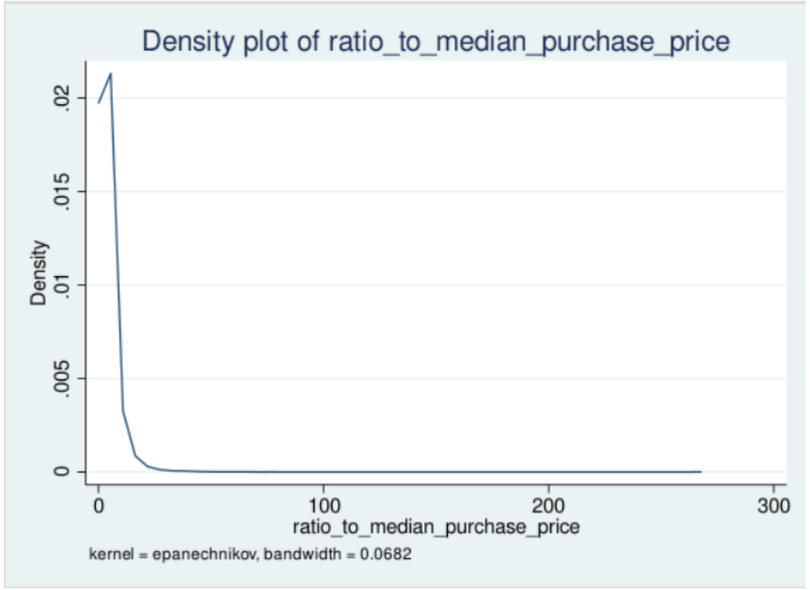
Summary of a few variables

Variable	Obs	Mean	Std. dev.	Min	Max
fraud	1,000,000	.087403	.2824248	0	1
online_order	1,000,000	.650552	.4767959	0	1
distance_f~e	1,000,000	26.62879	65.39078	.0048744	10632.72
distance_f~n	1,000,000	5.036519	25.84309	.0001183	11851.1
ratio_to_m~e	1,000,000	1.824182	2.799589	.0043992	267.8029
repeat_ret~r	1,000,000	.881536	.3231569	0	1
used_chip	1,000,000	.350399	.4770951	0	1
used_pin_n~r	1,000,000	.100608	.3008091	0	1

Continous variables







Dummy variables

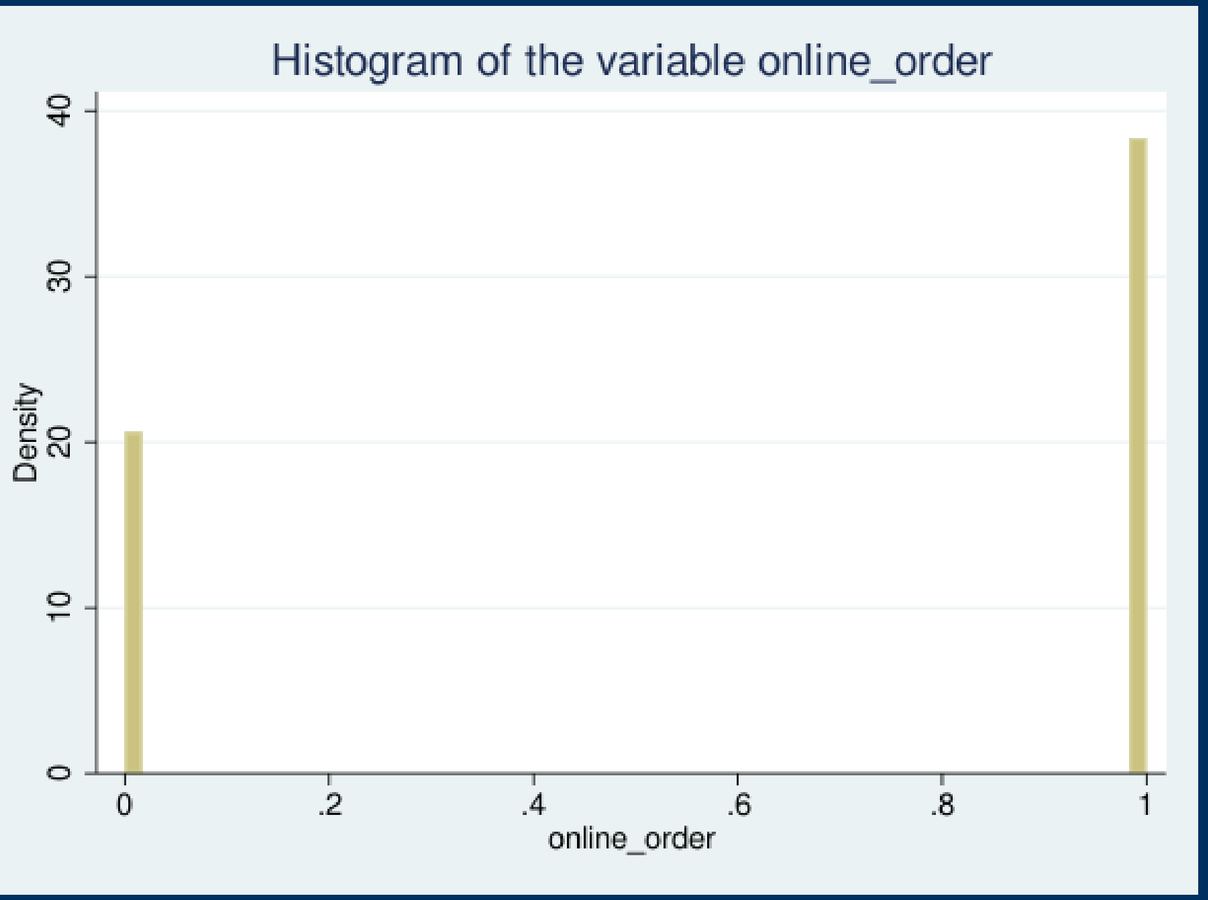
online_order	Freq.	Percent	Cum.
0	349,448	34.94	34.94
1	650,552	65.06	100.00
Total	1,000,000	100.00	

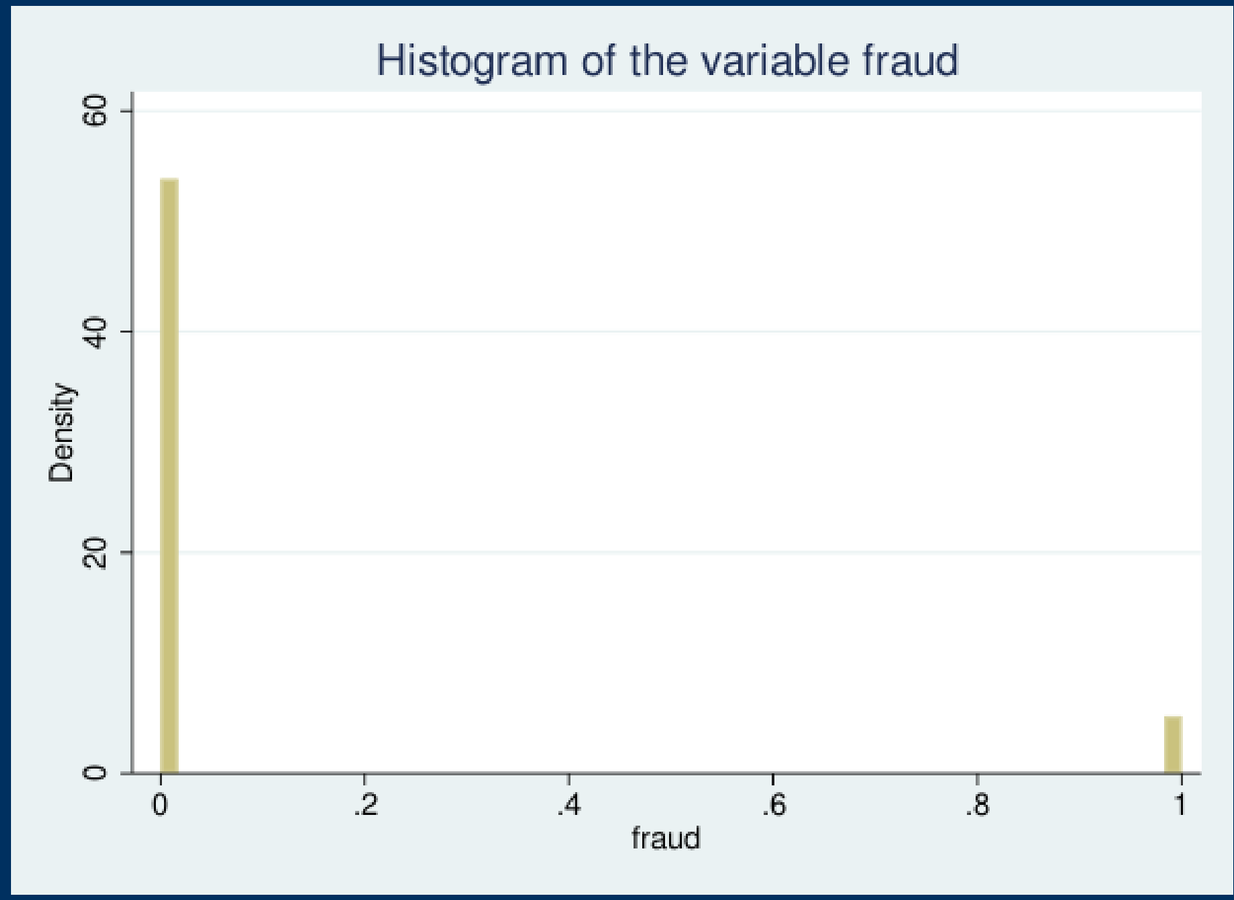
fraud	Freq.	Percent	Cum.
0	912,597	91.26	91.26
1	87,403	8.74	100.00
Total	1,000,000	100.00	

epeat_retailer	Freq.	Percent	Cum.
0	118,464	11.85	11.85
1	881,536	88.15	100.00
Total	1,000,000	100.00	

used_chip	Freq.	Percent	Cum.
0	649,601	64.96	64.96
1	350,399	35.04	100.00
Total	1,000,000	100.00	

used_pin_number	Freq.	Percent	Cum.
0	899,392	89.94	89.94
1	100,608	10.06	100.00
Total	1,000,000	100.00	


Histogram of the variable online_order
Density
0
10
20
30
40
0
.2
.4
.6
.8
1
online_order


Histogram of the variable fraud
Density
0
20
40
60
0
.2
.4
.6
.8
1
fraud

Logit regression of fraud and online_order

```
. logit fraud online_order

Iteration 0:   log likelihood = -296487.78
Iteration 1:   log likelihood = -276045.76
Iteration 2:   log likelihood = -272768.07
Iteration 3:   log likelihood = -272688.98
Iteration 4:   log likelihood = -272688.94
Iteration 5:   log likelihood = -272688.94

Logistic regression                                     Number of obs =  1,000,000
                                                        LR chi2(1)    =   47597.66
                                                        Prob > chi2   =     0.0000
Log likelihood = -272688.94                             Pseudo R2     =     0.0803

------------------------------------------------------------------------------
       fraud | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
online_order |   2.370489   .0151618   156.35   0.000     2.340773    2.400206
       _cons |  -4.296978   .0146979  -292.35   0.000    -4.325785   -4.268171
------------------------------------------------------------------------------
```

Percentage of online_order according to the fraud state of a transaction

```
. tabulate online_order fraud, row
```

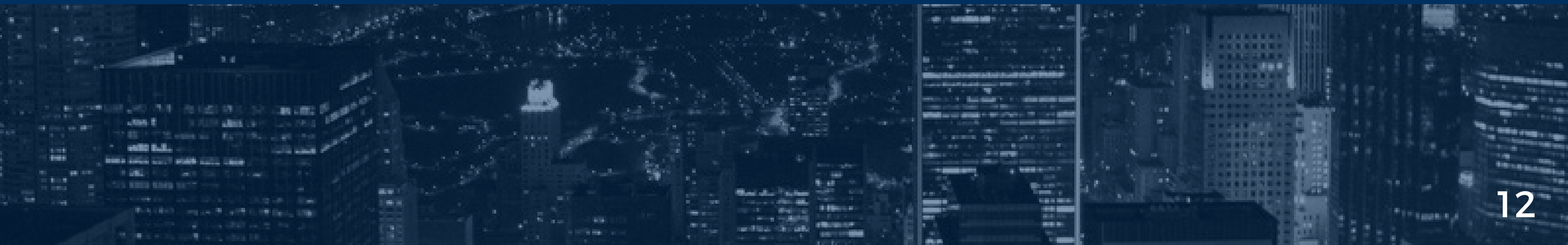
Key
frequency *row percentage*

online_order	fraud 0	1	Total
0	344,756 98.66	4,692 1.34	349,448 100.00
1	567,841 87.29	82,711 12.71	650,552 100.00
Total	912,597 91.26	87,403 8.74	1,000,000 100.00

Percentage of fraud according to the online order state of a transaction

fraud	online_order 0	1	Total
0	344,756 37.78	567,841 62.22	912,597 100.00
1	4,692 5.37	82,711 94.63	87,403 100.00
Total	349,448 34.94	650,552 65.06	1,000,000 100.00

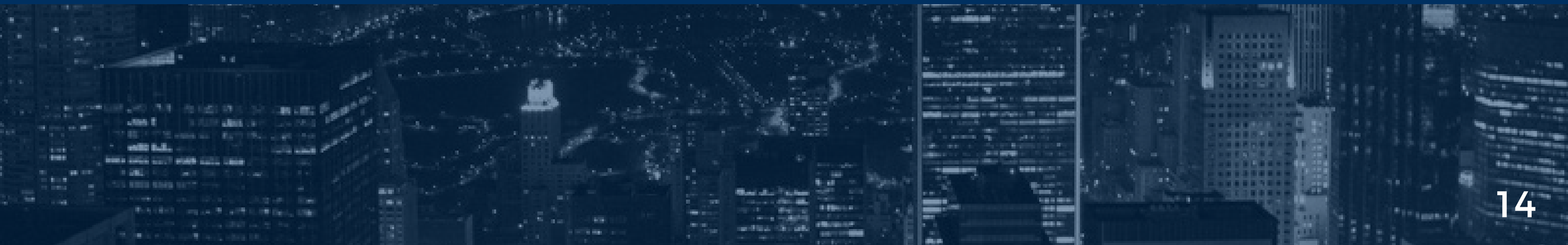
2.Empirical strategy



Empirical strategy

1. FIRST NAIVE ANALYSIS
2. IDENTIFY THE CONTROL VARIABLES OF THE MODEL IN ORDER TO LIMIT THE OMITTED VARIABLE BIAS
3. CHOOSE BETWEEN THE LOGIT AND PROBIT MODEL
4. PERFORM VARIOUS LOGIT REGRESSIONS

3.Tests and results



Qualitative approach

fraud	online_order 0	1	Total
0	344,756 37.78	567,841 62.22	912,597 100.00
1	4,692 5.37	82,711 94.63	87,403 100.00
Total	349,448 34.94	650,552 65.06	1,000,000 100.00

```
. tabulate online_order fraud, row
```

Key
frequency *row percentage*

online_order	fraud 0	1	Tot
0	344,756 98.66	4,692 1.34	349,4 100.
1	567,841 87.29	82,711 12.71	650,5 100.
Total	912,597 91.26	87,403 8.74	1,000,0 100.

Estimation of the conditional propability of 'fraud' using the frequency :

When the transaction is online, its propability of being fraudulent increases of 848.5%

Correlation between data

	fraud	online~r	distan~e	distan~n	ratio_~e	repeat~r	used_c~p
fraud	1.0000						
online_order	0.1920	1.0000					
distance_f~e	0.1876	-0.0013	1.0000				
distance_f~n	0.0919	0.0001	0.0002	1.0000			
ratio_to_m~e	0.4623	-0.0003	-0.0014	0.0010	1.0000		
repeat_ret~r	-0.0014	-0.0005	0.1431	-0.0009	0.0014	1.0000	
used_chip	-0.0610	-0.0002	-0.0007	0.0021	0.0006	-0.0013	1.0000
used_pin_n~r	-0.1003	-0.0003	-0.0016	-0.0009	0.0009	-0.0004	-0.0014

Quantitative approach

Regression of “fraud” according to “online_order”

```
Iteration 0:   log pseudolikelihood = -296487.78
Iteration 1:   log pseudolikelihood = -276045.76
Iteration 2:   log pseudolikelihood = -272768.07
Iteration 3:   log pseudolikelihood = -272688.98
Iteration 4:   log pseudolikelihood = -272688.94
Iteration 5:   log pseudolikelihood = -272688.94

Logistic regression                               Number of obs =  1,000,000
                                                  Wald chi2(1)  =   24444.04
                                                  Prob > chi2   =     0.0000
Log pseudolikelihood = -272688.94                 Pseudo R2     =     0.0803
```

fraud	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
online_order	2.370489	.0151618	156.35	0.000	2.340773	2.400206
_cons	-4.296978	.0146979	-292.35	0.000	-4.325785	-4.268171

Quantitative approach

Regression of “fraud” according to “online_order”, "ratio_to_median_purchase_price”

Logistic regression

Log pseudolikelihood = **-184017.62**

Number of obs = **1,000,000**
Wald chi2(**2**) = **21101.33**
Prob > chi2 = **0.0000**
Pseudo R2 = **0.3793**

fraud	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
online_order	4.526214	.0499215	90.67	0.000	4.42837	4.624059
norm_ratio_to_median_purchase	163.6307	1.161496	140.88	0.000	161.3543	165.9072
_cons	-7.939848	.0552018	-143.83	0.000	-8.048042	-7.831655

Note: 0 failures and 376 successes completely determined.

Quantitative approach

Regression of “fraud” according to “online_order”,"ratio_to_median_purchase_price” and “distance_from_home”

Logistic regression

Log pseudolikelihood = **-164240.94**

Number of obs = **1,000,000**
Wald chi2(**3**) = **17019.80**
Prob > chi2 = **0.0000**
Pseudo R2 = **0.4460**

fraud	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
online_order	5.49446	.066317	82.85	0.000	5.364481	5.624439
norm_distance_from_home	131.9205	1.92061	68.69	0.000	128.1562	135.6848
norm_ratio_to_median_purchase	191.3214	1.502591	127.33	0.000	188.3764	194.2664
_cons	-9.600645	.0771921	-124.37	0.000	-9.751939	-9.449351

Note: 0 failures and 598 successes completely determined.

Quantitative approach

Regression of “fraud” according to “online_order”, "ratio_to_median_purchase_price”, “distance_from_home” and “used_pin_number”

Logistic regression

Log pseudolikelihood = **-144742.23**

Number of obs = **1,000,000**
Wald chi2(4) = **16449.00**
Prob > chi2 = **0.0000**
Pseudo R2 = **0.5118**

fraud	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
online_order	6.05083	.0742545	81.49	0.000	5.905294	6.196366
norm_distance_from_home	146.1087	2.176056	67.14	0.000	141.8437	150.3737
norm_ratio_to_median_purchase	215.712	1.714815	125.79	0.000	212.351	219.0729
used_pin_number	-13.41294	.5593944	-23.98	0.000	-14.50933	-12.31654
_cons	-10.24491	.0868139	-118.01	0.000	-10.41506	-10.07475

Note: 33431 failures and 858 successes completely determined.

Quantitative approach

=> General downward bias
Underestimation of the effects of 'online_order' on 'fraud'

Final Z-index : (-10,24 + 6,05*OO)
P(fraud=1|OO=0) = 3,57.10**(-5)
P(fraud=1|OO=1) = 0,014

When the transaction is online,
its propability of being
fraudulent increases of 39115%

4. Conclusion

Qualitative approach

- 95% of fraudulent transactions were occurring online
- 98.7% of non-fraudulent transactions were not online

Quantitative approach

each control variable we added --> the coefficient of the “online_order” feature increased

Limits

- the dataset's source
- the correlation between our control variables and our explanatory variable being very weak