UNIVERSITÉ PARIS 1

Applied Econometrics Project

Groupe 21 case 4 part 2

November 9, 2025

Students: Lecomte Thomas, Taieb Migueres Ellis, Tessema Daniel

First-year master's degree MBFA - Université Paris 1 Panthéon-Sorbonne

---

1 Part II/ Questions

1.1) Data description

As a starting point of this second part we are going to provide a first descriptive overview of the dataset and to introduce two key variables of interest that represent the economic and demographic dimensions of the analysis. This directly relates to our research question: *What is the impact of economic and demographic determinants on the likelihood of defaulting?*

In line with this objective, we focus on *Age* and *Taxable Income*. These variables capture complementary aspects of credit default risk. Demographically, age may affect financial stability across the life cycle—young clients tend to have less experience and accumulated wealth, while older individuals often display greater financial discipline. Economically, taxable income directly reflects repayment capacity and remains central in credit risk assessment. We expect some correlation between age and income, which is economically reasonable and unproblematic at this stage, as both jointly shape the borrower's financial profile.

Table 1 reports the mean and standard deviation of these two variables across the two default groups. This provides a preliminary insight into whether defaulting clients differ systematically in age or income compared to non-defaulting clients.

Table 1: Summary Statistics by Default Group for Age and Taxable Income

Variable	Mean	SD	Min	Max
Age	61.2	8.7	34	99
Taxable income	90.1*	15.3	27.4	198.2
Age	54.5	10.4	35	89
Taxable income	74.8	18.2	30.5	165.7

Note: *Taxable income is expressed in thousands of currency units. Summary statistics are computed by default group (*Default* vs. *Not_default*).

The descriptive results show clear differences between the two groups. Defaulting clients tend to be younger on average (54.5 years compared to 61.2 for non-defaulters), but the most striking contrast appears in *taxable income*. Clients who default earn notably less on average (74.8 vs. 90.1 thousand units), and their income levels are also more dispersed. This suggests that economic factors, particularly income, may have a stronger influence on default risk than demographic characteristics such as age.

1.2) Data transformations for econometric analysis

So, in the previous part we did some transformations to create variables that are both meaningful economically and suitable for regression analysis.

The *amountdef* variable contains many zeros and few large values. To reduce skewness and stabilize coefficients, we apply a logarithmic transformation $\log(1 + amountdef)$, which preserves interpretability by making effects proportional to default size. Similarly, *healthexp* is transformed into *log_health_exp* to stabilize variance and linearize relationships. We also construct *liq_to_totalwealth*, the ratio of liquid to total wealth, to capture differences in liquidity among clients regarding their total wealth because we assume that liquidity part of wealth may be a better proxy to assess client's ability to pay for their credit.

Table 2: Illustrative subset of the dataset (variables as rows, individuals as columns)

Individuals/Variable	Ind1	Ind2	Ind3
Age	41	47	46
Female	FALSE	FALSE	TRUE
Taxable income	48.43	63.25	48.19
Total wealth	61.46	116.18	81.63
Liquid wealth	5.80	10.61	7.59
Health exp	3.50	9.45	6.55
Log amount def	2.11	2.55	2.20
Log health exp	1.25	2.25	1.88
Liq / Tot wealth	0.094	0.091	0.093

1.3) The outcome variable

In this part, we focus on the variable *amountdef*, which represents the amount of credit default for each client. This variable is of type *double* and expressed in thousands of monetary units. It takes the value zero for non-defaulting individuals and a positive value for those who have defaulted. Hence, the distribution is semi-continuous — most observations are equal to zero, while the rest are strictly positive. This structure reflects the fact that default is a rare event in the dataset, occurring only for a subset of clients. The choice of *amountdef* as our outcome variable is consistent with our overall objective of understanding the determinants of credit default risk. It directly quantifies the financial magnitude of default and allows us to assess how economic and demographic characteristics influence this amount.

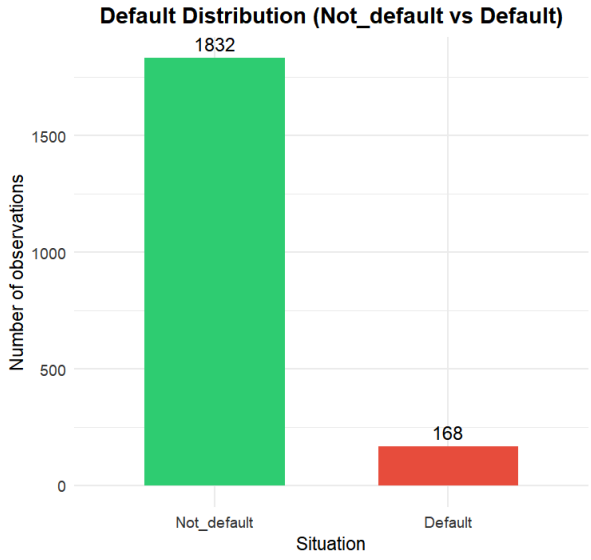

Figure 1: Default group distribution

Figure 1 shows a histogram distinguishing our two groups of individuals: those in red have a strictly positive default amount, while those in green have a default amount equal to zero. We can notice that 91.6 percent of the individuals in our sample are not in default. Therefore, in the next part of the analysis, we plan to create a new dataset including only the individuals with a strictly positive default amount.

1.4) Scatter plot

Therefore, in the next part of the analysis, we plan to create a new dataset including only the individuals with a strictly positive default amount. On the Figure 2 just bellow, our attention focuses on the amount of default as a function of taxable income, analyzed separately according to gender, since income constitutes a central determinant of repayment capacity. We first notice a strong concentration of points on the horizontal axis at zero which we removed to have a better sens of the distribution: the majority of clients show no

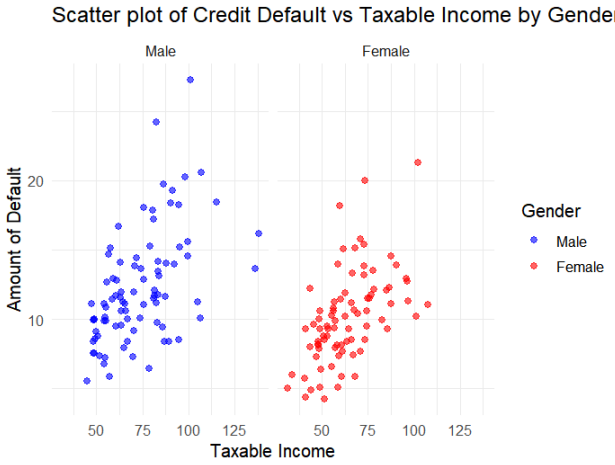

Figure 2: Amount default against Taxable income

default, regardless of their level of taxable income. Among those displaying a strictly positive default amount, the dispersion of points remains substantial: at equivalent taxable income, some record low default amounts, while others present higher values, without a decreasing relationship appearing linearly. We also notice that taxable income mainly exerts an influence on the amout of default occurrence, whereas the amount of default remains marked by strong heterogeneity. As a result, the distinction according to gender highlights a globally similar pattern between the two sub-populations.

1.5) Probit result

We define a dummy variable capturing the occurrence of default based on our initial *amountdef* variable. From that, the theoretical Probit model is written as:

$$\Pr(\text{Default}_i = 1 \mid X_i) = \Phi\Big(\beta_0 + \beta_1 \text{DummyGender}_i\Big) \tag{1}$$

Table 3: Average Marginal Effects from Probit Regression (Dummy Gender)

Statistic	AME	SE	z	p	lower	upper
dummy_gender	-0.006	0.012	-0.458	0.647	-0.030	0.019

Note: The AME (highlighted in red) is the coefficient to interpret. SE = standard error, z = z-statistic, p = p-value, lower/upper = 95% confidence interval.

Here, we choose to use *gender* as an explanatory variable for credit default risk, building on the study published by the Urban Institute in 2016. As the authors report: "Holding all else constant, a female-only borrower has a 0.2% lower probability of default than a male-only borrower, and the t-statistic is a very significant -25.9" (Goodman et al.(2016)), which suggests that, all else equal, women tend to exhibit a lower default risk than men.

After obtaining the average marginal effect from our probit regression (with *Default_dummy* as the dependent variable and *Gender_dummy* as the explanatory variable, equal to 1 if the individual is female and 0 if male), we can now interpret this coefficient. We obtain a value of -0.006, which means that, on average, being female reduces the probability of default by approximately 0.6 percentage points, holding all

other factors constant. Although this effect appears small, it is consistent with the findings of the study cited above.So, about the randomness, the dummy variable is defined as 1 for female and 0 for male. In the sample, there are 1010 females and 990 males, indicating a roughly balanced distribution. The balance between groups ensures sufficient variability to estimate the effect of gender on default probability in a probit model. However, the probit regression must be estimated on the full dataset, which includes 1,832 out of 2,000 individuals (91.6 percent) with no default. It is hardly feasible to correctly estimate the probability of default—the central objective of the analysis—if the sample is restricted to individuals with strictly positive default amounts. Consequently, low coefficients are expected, and caution should be exercised regarding the robustness and interpretation of the obtained results.

1.6) OLS results

For this question, we considered that it might be more relevant to estimate the regression on a subsample restricted to individuals with a strictly positive amount of default, in order to obtain more significant coefficients. To do so, we replaced our original dependent variable, *Default_dummy*, with *amountdef* (the amount of default). This yields a linear regression line. However, since our explanatory variable (*Gender_dummy*) is binary, the regression line has no meaningful interpretation for continuous values between 0 and 1 ($\forall$Gender_dummy $\in]0, 1[$).

$$\text{AmountDef}_i = \alpha_0 + \alpha_1 \text{DummyGender}_i + \varepsilon_i \tag{2}$$

Table 4: OLS regression results

	Dependent variable:
	amountdef
Gender_dummy	-2.114^{***}
	(0.576)
Constant	12.266^{***}
	(0.403)
Observations	168
R^2	0.075
Adjusted R^2	0.069
Residual Std. Error	3.735 (df = 166)
F Statistic	13.453^{***} (df = 1; 166)
Note:	*p<0.1; **p<0.05; ***p<0.01

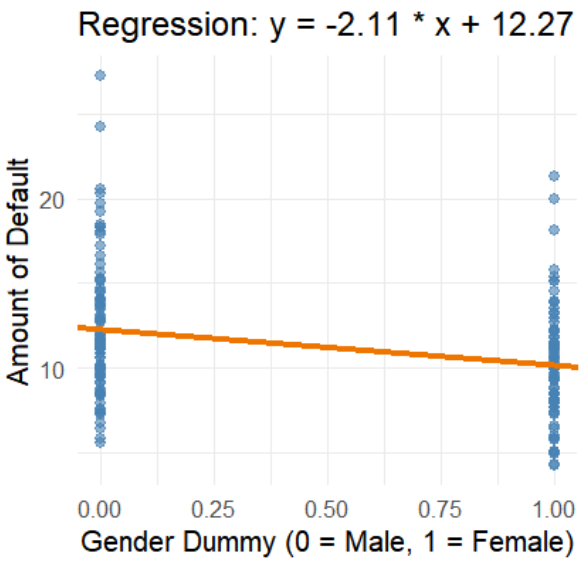

Figure 3: OLS results scatter plot

It can only be interpreted for the two actual categories of the variable. Thus, only two points on this line are interpretable, namely the points $(0, 12.661)$ and $(1, 10.5468)$, which correspond respectively to the average default amount for men and the average default amount for women. Accordingly, the estimated coefficient $\alpha_0 = 12,661$ indicates that the average default amount for men is \$12,661, which is significantly different from zero. Moreover, the coefficient $\alpha_1 = -2,114.2$ implies that, all else being equal, the average default amount for women is approximately \$2,114.2 lower than that of men. In other words, the average default amount for women equals \$10,546.8.

1.7) OLS extended results

Table 5: OLS Regression of *amountdef* on Gender and Controls, with Correlations

	Dependent variable:	Correlation
	amountdef	
Gender_dummy	−0.557	
	(0.573)	cor(Gender_dummy, amountdef) = 0.000
taxincome	0.109***	cor(amountdef, taxincome) = 0.571
	(0.013)	cor(Gender_dummy, taxincome) = −0.256
liq_to_totalwealth	−14.744*	cor(amountdef, liq_to_totalwealth) = −0.255
	(8.276)	cor(Gender_dummy, liq_to_totalwealth) = 0.508
Constant	5.637***	
	(1.329)	
Observations	168	
R^2	0.356	
Adjusted R^2	0.344	
Residual Std. Error	3.135 (df = 164)	
F Statistic	30.235*** (df = 3; 164)	
Note:		*p<0.1; **p<0.05; ***p<0.01

$$\text{AmountDef}_i = \alpha_0 + \alpha_1 \text{DummyGender}_i + \alpha_2 \text{TaxIncome}_i + \alpha_3 \text{LiqToTotalWealth}_i + \varepsilon_i \quad (3)$$

To refine our estimation, we add two additional explanatory variables: *taxincome* (taxable income) and *liq_to_totalwealth* (the share of liquid wealth in total wealth). These variables are potentially correlated with *Gender_dummy* as well as with the dependent variable *amountdef*. Logically, higher income is expected to increase an individual's repayment capacity, which in turn may reduce both the probability and the amount of default. Furthermore, the composition of wealth—particularly the liquid share—affects financial flexibility in case of payment difficulties. Just looking at Table 5 the results obtained are as follows: The coefficient $\alpha_0 = 5.63718$ indicates that, all else being equal, the average default amount for a man with no liquid assets and a zero share of taxable income is \$5,637.18. The coefficient $\alpha_1 = -0.55669$ shows that, for a woman with the same characteristics (same share of taxable income and same proportion of liquid wealth) as a man, the average default amount is approximately \$556.69 lower. Furthermore, $\alpha_2 = 0.10912$ implies that an increase of \$1,000 in the share of taxable income leads, on average, to an increase of \$109.12 in the default amount, while $\alpha_3 = -14.74450$ indicates that moving from no liquid assets to having all wealth in liquid form reduces the average default amount by approximately \$14,744.50, all else being equal.

We note that the coefficient α_1 associated with *Gender_dummy* is much less significant (closer to zero) than in our previous regression (−2.1142). This result can be explained by the presence of an omitted variable bias in the previous model (Question 6). Indeed, there is a negative correlation between being a woman and the share of taxable income (corr(*Gender_dummy*, *taxincome*) = −0.2557). In other words, among individuals with strictly positive default amounts, men have, on average, a higher share of taxable income than women. Moreover, the share of taxable income is positively correlated with the default amount (corr(*amountdef*, *taxincome*) = 0.5711), meaning that as the share of taxable income increases, the default amount tends to rise. Therefore, by not initially including this variable, we underestimated the effect of gender on the default amount.

Similarly, there is a positive correlation between being a woman and the ratio of liquid wealth to total wealth (corr(*Gender_dummy*, *liq_to_totalwealth*) = 0.5080), indicating that, on average, women hold a higher proportion of liquid wealth than men. This ratio, in turn, is negatively correlated with the default amount (corr(*amountdef*, *liq_to_totalwealth*) = −0.2548), meaning that as the proportion of liquid wealth increases,

the default amount decreases. Overall, these relationships suggest that omitting these variables in the previous regression biased the estimate of the gender coefficient. Once these variables are included, the effect of gender on the default amount is adjusted upward (from -2.1142 to -0.55669), which explains why the coefficient β_1 becomes closer to zero and statistically non-significant in this new estimation.

1.8) Conclusion

From an economic standpoint, the results obtained are broadly consistent with our initial expectations. We observe that women, on average, exhibit a slightly lower default amount than men, although this relationship loses its statistical significance once we control for other explanatory variables such as taxable income or the ratio of liquid wealth to total wealth. This suggests that gender alone is not a major determinant of default risk, but rather that the observed differences are partly explained by disparities in income and wealth composition. However, the estimated relationship is likely not free from bias, notably because the conditional independence assumption is not satisfied. Indeed, for the effect of gender to be interpreted as causal, the "treatment" (here, the *Gender_dummy* variable) would need to be randomly distributed conditional on the explanatory variables included in the model. In other words, if we controlled for all individual differences other than gender that could influence the default amount (*amountdef*), the estimated coefficient on gender would truly reflect its own effect, independent of any omitted variables.

Yet, our data clearly show that this condition is not met. For example, there is a negative correlation between *Gender_dummy* and *taxincome*, indicating that men, on average, have a higher share of taxable income than women. Conversely, we observe a positive correlation between *Gender_dummy* and *liq_to_totalwealth*, suggesting that women hold proportionally more liquid wealth than men. If the conditional independence assumption were satisfied, we would expect these correlation coefficients to be close to zero.These cross-correlations indicate that the *Gender_dummy* variable is not randomly distributed, even after controlling for the variables included in the model. Consequently, the estimated gender effect is likely biased, as it also captures the indirect influence of unobserved or partially controlled socio-economic characteristics.

More broadly, our model does not lie fully within the domain of the observable, since the explanatory variables available in our dataset cover only a subset of the relevant determinants of default risk. This limitation reduces the model's ability to accurately represent the underlying economic reality and prevents a truly causal interpretation.

In summary, although the results are consistent with the idea that women tend to have a slightly lower default risk, aligning with the evidence we found on that precise point (Goodman et al.(2016)), the estimated relationship should be interpreted as correlational rather than causal, due to the non-random assignment of the treatment, the violation of the conditional independence assumption, and potential biases arising from omitted variables.

Reference

GOODMAN, L., ZHU, J., BAI, B., & Urban Institute. (2016). *Women Are Better than Men at Paying Their Mortgages*. Housing Finance Policy Center Research Report. Available at : https://www.urban.org/sites/default/files/publication/84206/2000930-Women-Are-Better-Than-Men-At-Paying-Their-Mortgages.pdf

Annexe

Table 6: Sample of the raw dataset

ID	Age	Female	Taxable income	Total wealth	Liquid wealth	Health expenditure	Education	Amount default
1932	41	FALSE	48.43	61.46	5.80	3.50	7	7.54
158	47	FALSE	63.25	116.18	10.61	9.45	6	11.95
598	46	TRUE	48.19	81.63	7.59	6.55	8	8.10
1397	62	TRUE	67.70	109.36	18.58	14.82	8	10.65
534	43	FALSE	47.54	90.65	11.58	6.42	6	11.12
1368	70	FALSE	114.31	142.17	18.01	21.07	6	0.00
1384	60	TRUE	85.96	124.46	20.44	6.77	7	0.00
1134	59	TRUE	69.65	101.72	15.30	8.07	9	0.00
1524	51	TRUE	66.34	115.34	12.70	8.79	8	0.00
724	83	FALSE	175.87	249.84	21.85	34.36	9	0.00

Note: Taxable income, total wealth, liquid wealth, and health expenditure are expressed in thousands of currency units.

Group21_case4_Part2

```
# All the package that we're using for this second part.

library(corrplot)
library(lmtest)
library(stargazer)
library(psych)
library(arm)
library(ggplot2)
library(car)
library(rgl)
library(dplyr)

#-----------------------------------------------
#1.1) Data description
#-----------------------------------------------

data<- read.csv('group21.csv')
data
df<- data.frame(data)

# Just for the annexe to have a global sense of the data and not just have a look on only amountdef ==0

set.seed(123)

sample_default <- df[df$amountdef > 0, ]
sample_nodefault <- df[df$amountdef == 0, ]

sample_final <- rbind(
  sample_default[sample(1:nrow(sample_default), 5), ],
  sample_nodefault[sample(1:nrow(sample_nodefault), 5), ]
```

```
)

print(sample_final)

summary(df)

subset_data <- df[, c("age", "taxincome", "totwealth", "edu", "amountdef_group")]
subset_data

# For Overleaf only :)

stargazer(subset_data)

#------------------------------------------------
# 1.2 ) Transformation on the dataset.
#------------------------------------------------

df$log_amount_def<- log(1+df$amountdef)

df$log_health_exp<- log(df$healthexp)

df$liq_to_totalwealth<- (df$liqwealth)/(df$totwealth)

#------------------------------------------------
# 1.3. The outcome variable
#------------------------------------------------

# Outcome variable --> amount_def.

typeof(df$amountdef)
```

```
hist(df$amountdef)

  # Let's just a better histogram using grouping so that we can have a better sens of the data.

table(df$amountdef_group) # Count de number of observations for each group.

df$amountdef_group <- ifelse(df$amountdef == 0, "Not_default", "Default")
df$amountdef_group

df$amountdef_group <- factor(ifelse(df$amountdef == 0, "0", ">0"),
                             levels = c("0", ">0"),
                             labels = c("Not_default", "Default"))

ggplot(df, aes(x = amountdef_group, fill = amountdef_group)) +
  geom_bar(width = 0.8, show.legend = FALSE) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, size = 5) +
  scale_fill_manual(values = c("Not_default" = "#2ECC71",  # vert
                               "Default" = "#E74C3C")) +   # rouge
  labs(title = "Default Distribution (Not_default vs Default)",
       x = "Situation",
       y = "Number of observations") +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text = element_text(color = "gray20")
  )

#-----------------------------------------------
# 1.4) Scatter plot
#-----------------------------------------------
# Subset for amountdef ==0
```

```
df_positive <- subset(df, amountdef > 0)

df_positive %>%
  ggplot(aes(x = taxincome, y = amountdef, colour = factor(Gender_dummy))) +
  geom_point(alpha = 0.6, size = 2) +
  facet_wrap(~Gender_dummy,
             labeller = as_labeller(c("0" = "Male", "1" = "Female"))) +
  labs(title = "Scatter plot of Credit Default vs Taxable Income by Gender",
       x = "Taxable Income",
       y = "Amount of Default",
       colour = "Gender") +
  scale_color_manual(values = c("0" = "blue", "1" = "red"),
                     labels = c("Male", "Female")) +
  theme_minimal(base_size = 12)
#-----------------------------------------------
# 1.5) Probit results
#-----------------------------------------------

df$Gender_dummy<-ifelse(df$female, 1, 0)

df$Default_dummy <- ifelse(df$amountdef > 0, 1, 0)

Probit1<- glm(Default_dummy~Gender_dummy,
              data = df,
              family = binomial(link = "probit"))

res_probit<-summary(Probit1)
res_probit

# Marginal effect computation
```

```
marginal_effect <- margins(Probit1)
marginal_effect

# Probit summary

res_marginal_effect<- summary(marginal_effect)

# For Overleaf only :)

stargazer(res_marginal_effect)

# About randomness
# It anable us to see how many women and men are present in the overall 2000 observations on the dataset.
table(df$female)
# Then we can conclude

#------------------------------------------------
# 1.6) OLS results
#------------------------------------------------

# We tailor our OLS model for only the individuals with amountdef>0 so we just use the dataset just created at question 4(see justification in main document pdf).

df_positive$Gender_dummy <- ifelse(df_positive$female, 1, 0)

# Linear model (OLS):

ols_positive <- lm(amountdef ~ Gender_dummy, data = df_positive)

# OLS summary
```

```
res_ols_postitive<-summary(ols_positive)

ggplot(df_positive, aes(x = Gender_dummy, y = amountdef)) +
  geom_point(alpha = 0.6, color = "steelblue", size = 2) +
  geom_abline(intercept = coef(ols_positive)[1],
              slope = coef(ols_positive)[2],
              color = "darkorange2",
              linewidth = 1.2) +
  labs(title = paste0("Regression: y = ", round(coef(ols_positive)[2], 2),
                      " * x + ", round(coef(ols_positive)[1], 2)),
       x = "Gender Dummy (0 = Male, 1 = Female)",
       y = "Amount of Default") +
  theme_minimal(base_size = 12)

# For Overleaf only :)

stargazer(ols_positive)

#-----------------------------------------------
# 1.7) OLS extended results
#-----------------------------------------------



df_positive$log_taxincome<- log(df_positive$taxincome)
df_positive$liq_to_totalwealth<- df_positive$liqwealth/df_positive$totwealth

ols_positive2 <- lm(amountdef ~ Gender_dummy+taxincome+liq_to_totalwealth, data = df_positive)

summary(ols_positive2)
```

```
# For overleaf ony :)


stargazer(ols_positive2)


# We compute the correlation coefficient just here
cor(df_positive$Gender_dummy, df_positive$taxincome, use = "complete.obs")
cor(df_positive$Gender_dummy, df_positive$liq_to_totalwealth, use = "complete.obs")
# Also there..
cor(df_positive$amountdef, df_positive$taxincome, use = "complete.obs")
cor(df_positive$amountdef, df_positive$liq_to_totalwealth, use = "complete.obs")
#-----------------------------------------------
# 8) only text ...
#-----------------------------------------------
```