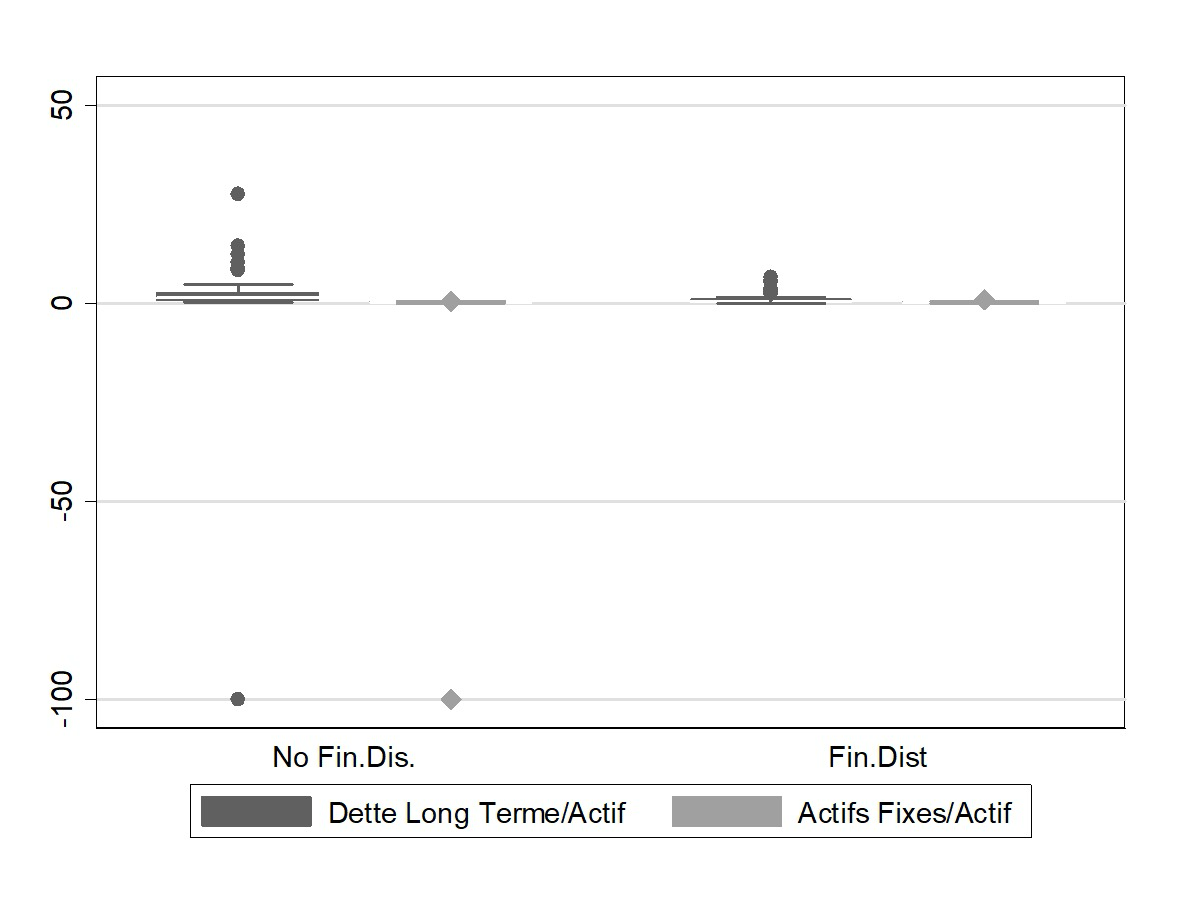
1) What is the use of « label » in SAS et STATA? Why the instruction in the Python does not proceed exactly as in “label” in SAS and STATA? Can you find an equivalent instruction in Python and R?

2) Copy here the image in the slides of the course which signaled some variables had unexpectedly arbitrary value and comment. Copy here the Python code which replaced the observations **-99** by NaN for missing observations for the related ratios.



Comment:

The slide above shows the boxplot comparing for Long-term Debt to Assets and Fixed Assets to Total Assets with both groups of default and non-default firms.

Several arbitrary or abnormal values are visible, especially for Long-term Debt to Assets, where some firms show extreme negative or very large positive ratios. These outliers could be the result of data entry or scaling mistakes, or from extreme observations of firms that are unusually leveraged or highly indebted. Observations like this are important to note for the purpose of cleaning or winsorizing the data in order to maintain interpretable estimation results.

df = get\_df\_csv('defaut2000.csv', '-99,99')

print(df.head())

df = adjust\_french\_decimal(df)

print(df.head())

df = df.apply(pd.to\_numeric, errors='coerce')

print(df.head())

3) For each explanatory variable in your sample of estimation copy the graphs of distributions and boxplots and report your comments on visual inspection and numbers for skewness, kurtosis, high leverage observations, non-normality for each group, equality of variance, test of equality of means and t-statistics, simple correlation coefficient with the dependent.

**TDTA (Debt/Assets):** Default firms' distribution shape is more right-shifted with a heavier upper tail, while non-defaults peak lower around 0.45. Non-defaults are slightly left-skewed, whereas defaults are right-skewed, and normality is rejected for defaults but not for non-default firms. The distributions differ by KS, and the mean is higher for defaulters, which is consistent with a strong positive correlation of 0.433, as higher overall leverage is associated with default.

**RETA (Retained Earnings/Assets):** Default group distribution is left-shifted with a mean closer to zero, whereas non-defaults center a bit higher. Non-defaults show slight right-skew, while defaults are left-skewed with heavier tails. Normality is rejected for defaults but not for non-defaults, and there is a strong negative correlation of r = -0.451.

**OPITA (Net Income/Assets):** Defaults concentrate at lower and negative shift, while non-defaults are shifted more to the right with a higher and also tighter peak. Both groups show left tails, although it is stronger in defaults (skew = -1.242, kurt = 3.310 vs non-defaults skew = -0.261, kurt = 0.937). JB rejects normality for only defaults, and the KS separation is the largest in the table (D= 0.543), and the coefficient shows a strong negative correlation of r = -0.419, showing that current profitability is a strong separator.

**EBITA (EBIT/Assets):** The default curve is slightly to the left of the non-default curve, which could indicate worse operating profitability for defaulters compared to non-defaulters and also a fatter left tail. JB rejects normality for both groups, KS shows a strong difference in distribution with D = 0.481, and the mean is lower for defaulters (t = -3.759). The correlation is negative and of a significant size, which indicates that weak operating earnings are closely associated with default risk.

**LSLS (Log Sales):** Both groups appear to be similar and symmetric in shape, with only a marginal rightward shift for the non-default group. JB does not reject normality for either group and KS shows no separation. With a correlation of nearly zero, log sales do not appear to distinguish defaults in this sample.

**LTA (Log Assets):** The distribution shapes are similar for both groups. JB does not reject normality and correlation with default is almost zero; therefore, log assets offer little explanatory value in the sample.

**GEMPL (Employee Growth):** Default firms are negative with a left shift relative to non-defaults, which are concentrated slightly positive. JB rejects both groups for normality and KS separation is strong (D = 0.46). Defaulters have lower means (t = -3.487), and the correlation coefficient is negative, indicating that employee growth is a strong signal for default.

**INVSLS (Inventory/Sales):** Both groups are right-skewed, JB rejects for defaulters but not for non-defaults, and the mean difference is only marginally significant (t = -1.799, p = 0.077), as is the correlation coefficient; therefore, higher inventories for defaulters represent a slightly weak signal.

**NWCTA (Working Capital/Assets):** The distributions of both groups overlap and are almost symmetric and the coefficient is negligible (r = 0.102, p = 0.258), representing a weak indicator for default.

**CACL (Current Ratio):** Both defaults and non-defaults are strongly right-skewed with long tails. JB rejects non-defaults and defaults for normality, and KS shows no distributional difference nor a significant mean gap. With a weak correlation coefficient, the current ratio does not separate groups despite non-normal tails.

**QACL (Quick Ratio):** Distributions show right-skew for both groups, as it did for the current ratio, and also similarly statistically fail to separate default and non-default groups.

**FATA (Fixed Assets/Assets):** Both groups are right-skewed, but defaults are slightly higher on average. Heavier tails for defaults (non-defaults: skew = 0.793, kurt = 1.247; defaults: skew = 1.351, kurt = 2.944) and JB rejects normality in both non-defaults and defaults. KS does not reject, and the mean gap is insignificant while the correlation is weakly positive.

**LTDTA (Long-term Debt/Assets):** Distributions mainly differ through extreme right-tail outliers among non-defaults, giving them a much higher mean despite most mass near zero for both groups. Outliers among non-defaults also produce very high kurtosis (kurt = 23.362) and strong right-skew (skew = 4.602) as the clearest high-leverage observation in the sample. JB rejects both groups for normality, and KS rejects equality, where the mean is actually higher for non-defaults (t = -2.589) to create a negative correlation with default (r = -0.257). Healthier firms are indicated to hold a few with notably large long-term debt, while distressed firms rely more on short-term liabilities.

**MVELTD (Market Value Equity/Long-term Debt):** The two distribution curves are broadly similar. JB does not reject, and KS and mean-gap show no difference or significance. Paired with a weak correlation coefficient, this market-based leverage ratio seems to show little explanatory power.

4) To which extent normality matters for explanatory variables in this study?

Normality of the explanatory variables is not required for the OLS, logit, or Probit models used in this study. Their assumptions don’t concern the marginal distribution of a variable, but rather concern for the normality of disturbances. Therefore, the fact that many ratios are skewed or heavy-tailed does not invalidate estimation or out-of-sample classification. However, normality is important for classical tests and discriminant analysis. The two-sample t-test is exact only under the conditions of normal and equal variances, so with non-normal data it’s important to also use tests and checks that don’t assume Gaussian X, like the Kolmogorov–Smirnov test, correlations, and visual densities.

5) Why is it useful to rank the explanatory variables by their correlation coefficient and t-statistics? Why is the order the same when using absolute values of t-statistics or using absolute values of correlation coefficients?

It is useful to rank the explanatory variables by correlation coefficient and t-statistics because this approach can show which indicators best separate defaulters from non-defaulters. By ranking, the model can prioritize strong indicators and drop weak ones, which would just add noise and redundancy to a multivariate model, saving efficiency and lowering the risk of overfitting.

The absolute t-stat and the absolute correlation give the same ranking because they measure the same relationship in different units: correlation captures effect size; the t-stat is the same effect scaled by its uncertainty. With a fixed sample, scaling doesn’t change the ordering, so sorting by |t| or by |corr| produces the same list since the t-stat is just a monotone rescaling of the correlation here.

6) Which are the explanatory variables most correlated with the dummy of default? Are there some explanatory variables with a simple correlation with the default dummy below 0.1 in absolute value? What would you expect in this case for predicting default in multivariate analysis?

The explanatory variables most correlated with the default dummy are RETA, TDTA , OPITA, and EBITA, all with absolute correlations above 0.35 - 0.45. These variables are strongly linked to default risk as they capture leverage and profitability, which are key indicators of financial stress.

In contrast, several variables show very low correlations with default for example: LSLS, LTA, MVETLD, and NWCTA, all below 0.1 in absolute value. Variables with weak individual correlations contribute little predictive power on their own. In a multivariate model, they are unlikely to be statistically significant once stronger predictors are included, unless they capture a complementary dimension not explained by the strong financial ratios.

7) Explain why the t-statistics is the same obtained for 4 different tests with equivalent null hypothesis (cf. slides of the course), including the test of equality of means?

The tests of a two-group comparison between the default dummy and a continuous variable can be framed four equivalent ways, which all testing the same null hypothesis of no linear association between the dummy and the variable. First, there is the the two-sample test of equal means, to find whether the variable’s average is the same for defaulters and non-defaulters, second, test whether the correlation between the dummy and the variable is zero, and third, you can estimate a one-way analysis of variance with two groups by regressing the variable on the the dummy. Fourth, the reverse simple regression, a linear probability model regressing the dummy on the variable and test whether the slope is zero. All four methods are different specifications of the same sample information: the between-group difference standardized by within-group variability under the pooled-variance assumption. Because they are algebraicly the same expressions of the same covariance standardized by the shared sample data, they result in the same t-statistics and p-values.

8) Bivariate correlation between explanatory variables: List explanatory variables with simple correlation coefficient with other explanatory variables larger than 0.8? For each case, why is it expected and explained by accounting items included into each ratio?

The highly correlated pairs (|r| ≥ 0.8) are: OPITA and EBITA (~0.97), LSLS and LTA (~0.96), CACL–QACL (~0.86), and TDTA–RETA (~0.86). These correlations are expected from the given their financial definitions: OPITA (net income/assets) and EBITA (EBIT/assets) are profitability ratios sharing the same denominator and closely related numerators as net income is EBIT after interest and taxes, so they move together. LSLS (log sales) and LTA (log assets) are both firm-size proxies, and larger firms tend to have both higher sales and asset bases. CACL (current ratio) and QACL (quick ratio) share the same denominator of current liabilities, and nearly the same numerator as quick assets are current assets minus inventories. TDTA (debt/assets) versus RETA (retained earnings/assets) are both scaled by assets. These overlaps of shared denominators, and similar numerators could signal potential multicollinearity in multivariate models.

9) Will you include all of them in your regressions? Explain what are the consequences of highly correlated explanatory variables for regression analysis.

A multivariate regression should not include all of the highly correlated variables together as severe multicollinearity inflates standard errors and variance inflation factors, making individual coefficients imprecise, unstable in sign and magnitude, and highly sensitive to small sample changes. It can also lead to low t-stats even when the model’s overall fit looks strong, and it undermines causal interpretation of any one variable because their effects are not possible to intentify separately. In-sample prediction may change little, but out-of-sample forecasts become innacurate. It is better practice to keep one explanatory variable from each cluster of correlated variables.

10) Comment the table of the bivariate clouds of points for six variables: yd, total debt and its highly correlated companion ratio, operating income / total assets and its highly correlated companion ratio, and growth of employees. Does the shape of the clouds matches with the simple correlation coefficients?

The scatter cloud shapes align with the simple correlation coefficients. The TDTA and RETA panel shows a tight, downward-sloping line, confirming expectations of the strong negative correlation as high leverage is linked with low retained earnings. The OPITA and EBITA panel is almost a 45 degree line, matching the very strong positive correlation as both profitability ratios share the assets denominator and closely related numerators. In contrast, relationships with GEMPL are light with only slight slopes, consistent with weaker correlations with both leverage and profitability. Looking across columns by defaults and non-default groups, defaulters cluster at higher TDTA and lower RETA/OPITA/EBITA, which matches the sign of their correlations with the default dummy.

11) Comment the items and compare the default output table in the Python code with the alternative single table with a column for each of the estimations: for linear probability model, Logit and Probit when total debt/total assets ratio is the only explanatory variable.

With TDTA as the only predictor, all three models show that higher leverage strongly increases the probability of default, and the effect is significant. In the LPM, the slope is positive and sizable (0.94), meaning a 0.10 increase in TDTA is associated with about a 9-10 percentage points higher default probability. In Logit and Probit, the TDTA coefficients are also positive and highly significant (5.25 and 3.20), and the marginal effects show the expected nonlinearity, roughly 0.15 around a low baseline probability of default (PD=0.25) and 0.40 near the middle (PD=0.50), therefore so leverage has a sharper effect when firms are already at intermediate risk. Despite different scales, the three deliver virtually the same ranking (AUC = 0.73), because each is a monotone transform of the same index with a single regressor.

12) Explain how you compute the percentage of concordant pairs when the dependent variable is binary.

To compute the percentage of concordant pairs with a binary outcome, take all possible pairs that contain one positive case (y=1) and one negative case (y=0), then compare their model scores through predicted probabilities or risk scores. A pair is concordant if the positive case is assigned the higher score, and discordant if the negative case is assigned the higher score, and a draw if the scores are equal. All observations can be sorted by score, and for each positive, by counting how many negatives lie below it, those pairs are concordant for that positive. These counts are then added over all positives to get the total number of concordant pairs, divided by the total number of all y=1 vs y=0 pairs, and then multipled by 100 to get the percentage. This measure captures how well the model orders cases by risk which is also the same ranking system behind ROC/AUC and the Mann–Whitney U viewpoint. Draws reflect indifference and are often treated as half credit when converting to AUC.

Document T:

Python to Stata:

|  | Initial error by the LLM translation | Correction using LLM |
| --- | --- | --- |
| 1 | Estimation Sample was empty | Preserved sample |
| 2 | Histogram bin widths were not accepted | Precomputed bin width |
| 3 | Invalid Syntax error in |  |