Trideux Software Integrates Jérôme Deauvieau's "How to Translate a Logit Model into Probabilities", Bulletin de Méthodologie Sociologique 105, 2010 p.53-60

ONGOING RESEARCH/RECHERCHE EN COURS

TRIDEUX SOFTWARE INTEGRATES JÉRÔME DEAUVIEAU'S "HOW TO TRANSLATE A LOGIT MODEL INTO PROBABILITIES"'

by

Philippe Cibois
(Laboratoire Printemps CNRS / Université de Versailles Saint-Qentin en Yvelines; phcibois@wanadoo.fr)

Résumé:

Dans son article "'Comment traduire sous forme de probabilités les résultats d'une modélisation logit?", Jérôme Deauvieau montre que la présentation sous forme de probabilités ou de pourcentage peu poser des problèmes car les résultats dépendent du choix de la modalité qui a servi de référence; Deauvieau propos deux solutions à ce problème, l'ajustement selon l'écart expérimental et l'ajustement selon l'écart pur. Ces deux méthodes sont maintenant implémentées dans le logiciel Trideux et un exemple est proposé.

Abstract:

In his article "How to Translate a Logit Model into Probabilities?", Jérôme Deauvieau shows that presentation as probabilities or percentages can cause problems because the results depend upon the choice of modalities that serve as a reference category . Deauvieau offers two solutions to this problem, the adjustment under the experimental deviation and the adjustment as pure deviation. Both methods are now implemented in software Trideux and an example is proposed.

***

In his article in this issue of the *BMS*, "How to Translate a Logit Model into Probabilities?", Jérôme Deauvieau stresses the fact that sociologists (in France, in particular) often use discrete variables rather than continuous variables in their data, and therefore in their logistic regressions. For Deauvieau, the reason is that sociologists "very often use categorical variables (sex, PCS, ...) and when in the presence of numerical variables (salary, age, ...), the normal practice is to put them into categories."

This approach has implications for how to present the results of logistic regressions made on these discrete data. Indeed, since the logic behind the usual presentation of discrete variables is the cross tabulation, where differences are marked in percentages, it remains well within the logic of such a presentation to provide the results of logistic regressions in the form of probabilities or percentages. Moreover, as these results are easier to read than the coefficients of a logit model or than as odds ratios, it is possible to communicate results to a wider audience.

THE DIFFICULTIES OF PROBABILITY PRESENTATION

However, this presentation as probabilities or percentages can cause problems because the results depend upon the choice of modalities that serve as a reference category and Deauvieau offers two solutions to this problem.

In what follows, I will show how these methods can now carried out using the Trideux software[1]. To begin with, and to show the difficulties of the usual representation, I will present an example involving data for the choice of taking Latin, an optional course that can be taken in the second year of secondary education in France. The data used are the 1995 Panel of the French Ministry of National Education.

An introduction to Latin takes place in French fifth grade (the second year of secondary education) and is chosen by 27.8% of students. They have usually been good students since their primary education and at their arrival in French sixth grand (first year of secondary education). Their educational level was assessed in four areas (reading, written and spoken French, mathematics), and they received a grade from 0 to 10. Receiving 8 to 10 is considered an “excellent” grade. For the general poppulation of fifth graders, 17% received a grade of “excellent” in all areas, when 34% of Latinists received such grades.

The Latinists can distinguished from others by a number of aspects:

- Choice of first language: they are twice as likely as non-Latinists (18.0% against 9.8%) to have chosen German instead of English as their first language in sixth grade. We know this choice is often a strategy to get into a "good class" (except in regions near Germany).

- Parents are often managers or in higher intellectual professions (28.3% against 10.4%), intermediate professionals (23.1% against 16.1%), but fewer employees (14.8% against 18.7%) or workers (19.9% against 38.4%). Also, parents are more likely to have a higher education (fathers 28.2% against 10.0%; mothers 30.7% against 10.3%).

- Music lessons: the children also regularly take music lessons outside of school twice as often (19.5% against 8.7%).

To find out what had the most influence on the choice of Latin among the variables listed above, it is possible to compare "all other things being equal " (logistic regression). The results from Trideux are (all effects are significant at 1% level):

```
Paramètres de la régression                          en pourcentages
Régression logistique
Modalité à expliquer : Choice of latin in 5th
Situation de référence No statement of excellence in 6th
                       English in 6th
                       Father without higher education
                       Mother without higher education
                       No music lessons

                           Paramètres    ChancesRef          %
                           -1.5980         0.2023          16.8
Effets marginaux                         Odds-ratio
```

[1] Trideux is free software made by the author that can be obtained at the URL http://pagesperso-orange.fr/cibois/Trideux.html

Stat.of excellence in 6th	**1.2519**	**3.50**	**24.6**
German in 6th	**0.4464**	**1.56**	**7.2**
Father higher education	**0.6261**	**1.87**	**10.6**
Mother higher education	**0.9152**	**2.50**	**16.7**
Music lessons	**0.5365**	**1.71**	**8.9**

As we took as the reference situation, everything not advantageous indicating the choice of the Latin option, the estimated percentage of the reference situation (16.8%) is far the observed situation (27.8%). We see that the most important effect is related to grades of excellence which increase the proportion of Latinist by 25%, and then mother's, then father's higher education degree, and the fact of having taken German in sixth grade or music lessons.

It is quite obvious that this distinguishing of effects is artificial because all these characteristics form the educational logic of a family of higher social origins. However, logistic regression highlights the result of this logic: academic excellence is paramount, even if other elements such as musical culture are part of the training objectives. The fact that music has an "effect" on the choice of Latin reveales that the general orientation of the educational strategy is seeking social excellence.

Following Deauvieau's critics of this logic, we are going to change the reference situation by simply modifying the first variable and taking the fact of having grades of excellence in sixth grade. This corresponds to a situation where the Latin option is more common and therefore this will raise the percentage of choice in the reference situation.

Paramètres de la régression **en pourcentages**
Régression logistique
Modalité à expliquer : Choice of latin in 5th
Situation de référence Statements of excellence in 6th
English in 6th
Father without higher education
Mother without higher education
No music lessons

	Paramètres	**ChancesRef**	**%**
	-0.3461	**0.7074**	**41.4**
Effets marginaux		**Odds-ratio**	
No statement of excellence	**-1.2519**	**0.29**	**-24.6**
German in 6th	**0.4465**	**1.56**	**11.1**
Father higher education	**0.6262**	**1.87**	**15.5**
Mother higher education	**0.9152**	**2.50**	**22.4**
Music lessons	**0.5366**	**1.71**	**13.3**

The reference situation is as expected above the average situation, all odds ratios are identical, except for the first modality which is now the opposite: 0,29 = 1/3.5. For this question, the effect percentage sign has simply changed, but all other effects are different. This awareness of the non-stability effects in percentages led Deauvieau to propose two methods to provide results in the form of probabilities or percentages, whatever the reference situation.

FIRST SOLUTION: THE EXPERIMENTAL DEVIATION

In Deauvieau's own words: "the individual probability of experiencing the event modeled is calculated for each individual in the sample. From these individual probabilities, we calculate "theoretical" probabilities by assuming the "experimental" attitude inherent to multiple regression methods. Indeed, the epistemological basis of multiple regression involves dividing up the individuals

according to their different social characteristics, and measuring the effect of characteristics independently. This come down to mimicking the experimental reasoning found for example in biology: I try to measure the effect of a feature on a given phenomenon, and for that I set up an experience by giving this feature to a group, omitting this feature to a second group, and watching what happens when comparing the two groups. (...) "

"With the play of the set of coefficients estimated by the model, one can enter by calculation into the precise experimental situation. Take for example the variable sex or gender in this situation. The model indicates that women have a logit of 0.59, lower than that of men . To reflect this difference with probability, it is sufficient to perform the following experiment on the sample: if all individuals in my sample were women, what would be the average probability of becoming a manager after five years? Then, if all individuals in my sample were men, what wolud be the average probability of becoming a manager after five years? Then just calculate the difference between these two probabilities. Thus, we obtain a probability measure of the difference between men and women becoming a manager after five years, "all things being equal". "

"Concretely, it is sufficient to calculate for each individual in the sample the probability of becoming a manager by applying the following two equations:

"First case: the sample is experimentally made up exclusively of women. We therefore omit for all individuals the coefficient related to sex. One calculates the probability of each individual becoming a manager ["cadre" in French] from the following formula:

$$P\,(Y=\text{cadre}) = \frac{1}{1+\exp^{-(-1{,}96+0{,}75*diplome-0{,}29*age2-0{,}63*age3)}}$$ "

"Second case: this time the sample consists only of men. Then we add to each individual in the sample (regardless of sex) the sex-linked factor in the equation. The formula for calculating the individual probability then becomes:

$$P\,(Y=\text{cadre}) = \frac{1}{1+\exp^{-(-1{,}96+0{,}75*diplome+0{,}59-0{,}29*age2-0{,}63*age3)}}$$ "

"The main drawback of this method is that the result is not necessarily equal to the logistic contrast as indicated by the logit coefficient of the model", which is not the case with the method presented below.

SECOND SOLUTION: THE "PURE" DEVIATION

This is the solution proposed by Léridon and Toulemon (1997: 251). The principle is to find the probabilities associated with a predictor that satisfies the following two conditions:

1. The difference between the probabilities respects the logistic contrast between the modalities as defined by the coefficient of the model.

2. The weighted average of probabilities related to the modalities of the explanatory variable is equal to the average probability of the variable to be explained for the whole sample.

Let *p* be the average proportion of the observed variable to be explained.

We suppose we have an explanatory variable with *n* modalities numbered *1* to *n*. Each modality has n_i subjects, with proportion p_i, *i* varying from *1* to *n*.

Let $alpha_i$ be the parameters of the logistic regression for the explanatory variable, where $alpha_n$ corresponds to the reference modality ($alpha_n = 0$, exp ($alpha_n$) = 1).

Let $Pajust_i$ be the sought for adjusted proportions (those of index *n* are those of the reference modality). Condition 1 requires that:

$Pajust_i$ / (1 - $Pajust_i$) / $Pajust_n$ / (1 - $Pajust_n$) = exp ($alpha_i$) from which one derives:

if we set A = exp ($alpha_i$) $Pajust_n$ x / (1 - $Pajust_n$) then $Pajust_i$ = A / (1 + A).

The second condition is that the sum of the n_i *x* $Pajust_i$ equals *p*, the average observed proportion.

The algorithm (box below) chooses $Pajust_n$ arbitrarily, then modifies this value by adding or subtracting successive values of 2^{-k} (k varying from 1 to a value sufficient to reach the desired precision, 30 for example), according to whether the result of the second condition is greater than or less than the average observed ratio.

Pajustn means $Pajust_n$, Effectif(i) designates n_i, alpha(i) denotes $alpha_i$, VarA is (intermediate variable) A, SommeNP means the sum of n_i *x* $Pajust_i$ for which PropSommeNP is the proportion of the total of the modality to be explained totalVarY, PropMoyenneObs is *p*, the general average of the modality to be explained. The value of the adjusted variable for the i^{th} modality is at the end in VarAjust (i)

```
Pajustn = 0
For K = 1 To 30
    SommeNP = 0
    Pajustn = Pajustn + 2 ^ -K
    For i = 1 To nmax
        VarA = exp(alpha(i)) * Pajustn / (1 - Pajustn)
        VarAjust(i) = VarA / (1 + varA)
        SommeNP = SommeNP + VarAjust(i) * Effectif(i)
    Next i
    PropSommeNP = SommeNP / totalVarY
    If PropMoyenneObs - PropSommeNP < 0 Then
         Pajustn = Pajustn - 2 ^ -K
    End If
Next K
```

RESULTS OF BOTH METHODS

Let's take the data presented above. This time, no matter which situation is taken as a reference, the adjusted proportions are identical, and that is true for both methods.

```
Sorties en probabilité selon méthode de l'effet expérimental
(Deauvieau, BMS)
Modalité à expliquer : Choice of latin in 5th, moyenne générale
0.2777

MoyExp : Moyenne expérimentale
Ecart1 : Ecart de la moyenne expérimentale à la moyenne générale en
%
Ecart2 : Ecart de la moyenne observée à la moyenne générale en %
                          MoyExp    Ecart1   Ecart2
Stat.of excellence in 6th 0.4913    21.4     28.6
```

```
No statement of excel.    0.2335    -4.4    -5.8
German in 6th             0.3496     7.2    16.3
English in 6th            0.2679    -1.0    -2.1
Father higher education   0.3787    10.1    28.1
Father no higher educ.    0.2593    -1.8    -4.6
Mother higher education   0.4309    15.3    29.5
Mother no higher educ.    0.2487    -2.9    -5.2
Music lessons             0.3660     8.8    21.1
No music lessons          0.2660    -1.2    -2.6

Sorties en probabilité selon méthode de l'effet pur
(Leridon/Toulemon 1997 p.252)
Modalité à expliquer : Choice of latin in 5th, moyenne générale
0.2777

MoyAjust : Moyenne ajustée selon l'effet pur
Ecart1 : Ecart de la moyenne ajustée à la moyenne générale en %
Ecart2 : Ecart de la moyenne observée à la moyenne générale en %
                         MoyAjust Ecart1   Ecart2
Stat.of excellence in  6th0.4490   17.1    28.6
No statement of excel.    0.1890   -8.9    -5.8
German in 6th             0.3552    7.8    16.3
English in 6th            0.2606   -1.7    -2.1
Father higher education   0.3721    9.4    28.1
Father no higher educ.    0.2406   -3.7    -4.6
Mother higher education   0.4111   13.3    29.5
Mother no higher educ.    0.2185   -5.9    -5.2
Music lessons             0.3698    9.2    21.1
No music lessons          0.2554   -2.2    -2.6
```

For both methods and for all explanatory modalities, the experimental means and adjusted average are given in proportions. In the column "Ecart1", the percentage of deviation from the overall average, here the proportion of choising Latin in fifth grade, which is 27.7% of the population. One immediately finds that the two methods give similar results.

The column marked "Ecart2", which is identical for both methods, gives the result of simple cross tabulation between the in line difference the percentage of a modality and the percentage all lines together; that is to say, the same overall average. As for all the modalities of an question, the differences given by logistic regression and by cross tabulations are very similar, which signifies that the procedure "all things equal elsewhere" does not provide better result than cross tabulations. This is not the case here because differences can be important.

In a comparison between, on one hand, the two previous analyses where the reference situations were different, and, on the other, the pure ajusted average method, we can classify the modalities by the importance of the effect.

1. First experiment (see above)

```
Statements of excellence in 6th     24.6
Mother higher education             16.7
Father higher education             10.6
Music lessons                        8.9
German in 6th                        7.2
```

2. Second experiment (*idem*. Here we classify by absolute value)

```
No statements of excellence in 6th  24.6
Mother higher education             22.4
Father higher education             15.5
```

Music lessons		**13.3**
German in 6th		**11.1**

3. Method according to the pure ajusted average: one class by taking the differences between modalities.

Stat.of excellence in 6th	**0.4490**	
No statement of excel.	**.1890**	
Difference		**26.0**
Mother higher education	**0.4111**	
Mother no higher educ.	**0.2185**	
Difference		**19.3**
Father higher education	**0.3721**	
Father no higher educ.	**0.2406**	
Difference		**13.2**
Music lessons	**0.3698**	
No music lessons	**0.2554**	
Difference		**11.4**
German in 6th	**0.3552**	
English in 6th	**0.2606**	
Difference		**9,5**

We find the same order, and similar orders of magnitude, but we now have a method with which we no longer ask the question of choice of reference, which is particularly reassuring.

REFERENCES

Deauvieau J (2010) Comment traduire sous forme de probabilités les résultats d'une modélisation logit ? *Bulletin de Méthodologie Sociologique*, 105 : 5-15.

Leridon H, Toulemon L (1997) *Démographie, approche statistique et dynamique des populations*. Paris : Economica.