ACADEMIC YEAR 2017/2018
Università degli Studi di Milano
GRADUATE SCHOOL IN SOCIAL AND POLITICAL SCIENCES
APPLIED MULTIVARIATE ANALYSIS
Luigi Curini
luigi.curini@unimi.it

Regression with a binary dependent variable: Logistic regression diagnostic

Logistic regression is popular in part because it enables the researcher to overcome many of the restrictive assumptions of OLS regression:

1. Logistic regression **does not assume a linear relationship** between the dependent and the independents. It may handle nonlinear effects even when exponential and polynomial terms are not explicitly added as additional independents because the logit function on the left-hand side of the logistic regression equation is non-linear. However, it is also possible and permitted to add explicit interaction and power terms as variables on the right-hand side of the logistic equation, as in OLS regression (as we already discussed!).
2. The dependent variable **need not be normally distributed** (but it assumes its distribution is within the range of the exponential family of distributions, such as normal, logistic, etc.)
3. The dependent variable **need not be homoscedastic** for each level of the independents; that is, there is no homogeneity of variance assumption: variances need not to be the same within categories.
4. **Normally distributed error** terms are not assumed.

However, **other assumptions** still apply:

1. **Multicollinearity**

Multicollinearity (or collinearity for short), as already discussed with OLS, occurs when two or more independent variables in the model are approximately determined by a **linear combination** of other independent variables in the model. For example, we would have a problem with multicollinearity if we had both height measured in inches and height measured in feet in the same model. The degree of multicollinearity can vary and can have different effects on the model. When perfect collinearity occurs, that is, when one independent variable is a perfect linear combination of the others, it is impossible to obtain a unique estimate of regression coefficients with all the independent variables in the model.

```
logit vote_2004 educ age income_hh
findit collin
collin educ age income_hh
```

All the measures in the above output are measures of the strength of the interrelationships among the variables. Two commonly used measures are **tolerance** (an indicator of how much collinearity that a regression analysis can tolerate) and **VIF** (variance inflation factor - an indicator of how much of the inflation of the standard error could be caused by collinearity). The tolerance for a particular variable is 1 minus the R^2 that results from the regression of the other variables on that variable. The corresponding VIF is simply 1/tolerance. If all of the variables are orthogonal to each other, in other words, completely uncorrelated with each other, both the tolerance and VIF are 1. If a variable is very closely related to another variable(s), the tolerance goes to 0, and the variance inflation gets very large.

As a rule of thumb, a tolerance of 0.1 or less (equivalently VIF of 10 or greater) is a cause for concern.

Notice that the R^2 for education is 0.2022. Therefore, the tolerance is 1-0.2022 = 0.7978. The VIF is 1/0.7978 = 1.25.

We can reproduce these results by doing the corresponding regression:

```
reg educ age income_hh
```

2. Model specification

When we build a probit or logit regression model, we assume that **we have included** all the relevant variables and that we have **not** included any variables that should not be in the model. This is always true for any statistical model out there!

Proper specification of the model is particularly crucial; parameters may change magnitude and even direction when variables are added to or removed from the model.

- *Inclusion of all relevant variables in the model*: If relevant variables are omitted, the common variance they share with included variables may be wrongly attributed to those variables, or the error term may be inflated.
- *Exclusion of all irrelevant variables*: If causally irrelevant variables are included in the model, the common variance they share with included variables may be wrongly attributed to the irrelevant variables. The more the correlation of the irrelevant variable(s) with other independents, the greater the standard errors of the regression coefficients for these independents.

The Stata command **linktest** that we have already discussed can be used to detect a specification error, and it is issued after the **logit** command. The idea behind **linktest** is that if the model is properly specified, one should not be able to find any additional predictors that are statistically significant except by chance. After the regression command (in our case, **logit**), **linktest** uses the linear predicted value (**_hat**) and linear predicted value squared (**_hatsq**) as

the predictors to rebuild the model. The variable **_hat** should be a statistically significant predictor, since it is the predicted value from the model. This will be the case unless the model is completely misspecified. On the other hand, if our model is properly specified, variable **_hatsq** shouldn't have much predictive power except by chance. Therefore, if **_hatsq** is significant, then the **linktest** is significant. This usually means that we have omitted relevant variable(s). We need to keep in mind that **linkest** is simply a tool that assists in checking our model. It has its limits. It is better if we have a theory in mind to guide our model building, that we check our model against our theory, and that we validate our model based on our theory.

Lacking an interaction term could cause a **model specification problem**. Similarly, we could also have a model specification problem if some of the predictor variables are not **properly transformed**.

To address this, a Stata program called **boxtid** could be used. It is a user-written program that you can download over the internet by typing "**findit boxtid**". **boxtid** stands for Box-Tidwell model, which transforms a predictor using power transformations and finds the best power for model fit based on maximal likelihood estimate. More precisely, a predictor *x* is transformed into $B_1 + B_2x^p$ and the best p is found using maximal likelihood estimate. Besides estimating the power transformation, **boxtid** also estimates exponential transformations, which can be viewed as power functions on the exponential scale. Of course, you theory should be the main factor here…that is, you should know ex-ante if the relationship between the DV and a given IV is linear or a more complex one…

Now let's look at an example.

```
logit hiqual yr_rnd meals
linktest
```

(`yr_rnd`: *year-round school.* Year-round education is actually an approach that gives schools a variety of options to arrange the 180-day school calendar to better support student learning. Instead of containing a three month vacation, as a traditional school calendar does, it evenly spaces several "mini" vacations into the twelve month school calendar. During these twelve months, learning time may be extended or spaced; `meals`: percentage of students on free or reduced-priced meal)

The **linktest** is significant, indicating problem with model specification. We then use **boxtid**, and it displays the best transformation of the predictor variables, if needed.

```
boxtid logit hiqual yr_rnd meals
```

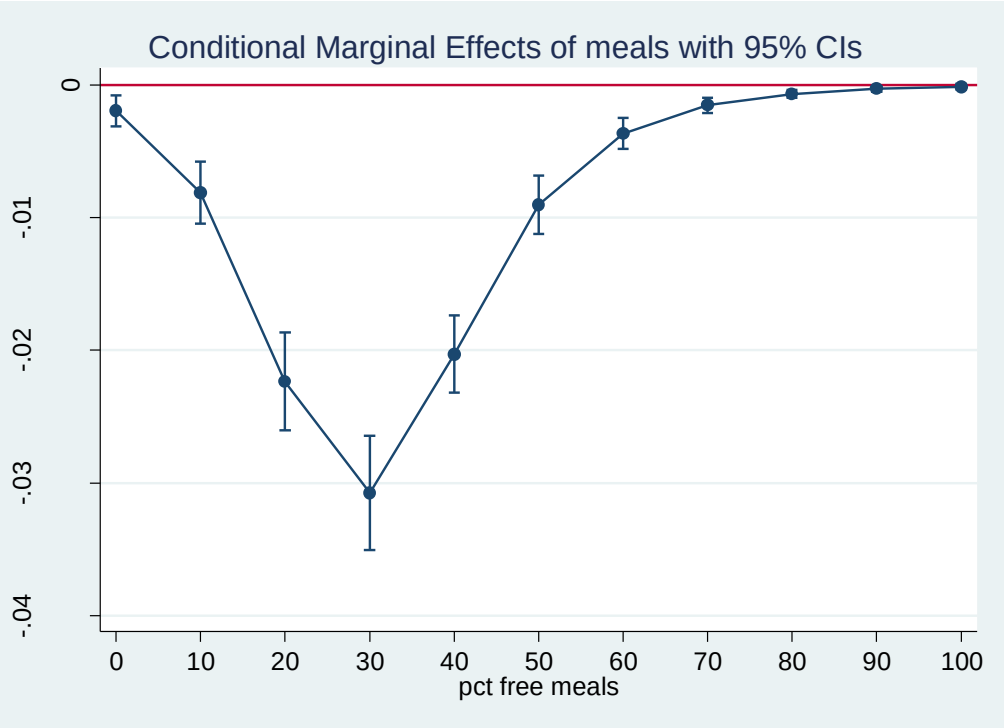
The test of nonlinearity for the variable **meals** is statistically significant with p-value =.005. The null hypothesis is that the predictor variable **meals** is of a linear term, or, equivalently, p1 = 1. But it shows that p1 is around .55 to be optimal. This suggests a square-root transformation of the variable **meals**. So let's try this approach and replace the variable **meals** with the square-root of itself. This might be consistent with a theory that the effect of the variable meals will attenuate at the end.

```
gen m2=meals^.5

logit hiqual yr_rnd m2
linktest
```

As an alternative, we could have suspected a quadratic relationship between meals and DV (being poor is not that bad for the quality of a school, if everyone is already poor!)

```
logit hiqual yr_rnd c.meals##c.meals
linktest
margins, dydx(meals) at(meals=(0 (10) 100) (mean)_all) vsquish
marginsplot, yline(0)
```



This shows that sometimes the logit of the outcome variable may not be a linear combination of the predictors variables, but a linear combination of transformed predictor variables, possibly with interaction terms.

3. **Error terms are assumed to be independent**

Of course, one can also have a problem with model specification (i.e., omission bias) if the model is violating the issue of independence assumption (remember our previous discussion with OLS). Violations of this assumption can have serious effects. Violations will occur, for instance, in cluster sampling, or time-series data. All our previous discussion on cluster standard error, fixed effects, random models apply here!

Let's see an example with the Union Dataset:

```
logit union age  south year
estimates store logit
logit union age  south year, cluster(id)
estimates store cluster
xtlogit union age  south year , i(id) re
```
the lnsig2u reported in the table is just the log of the variance at the second level, and in fact di exp(1.76131)^0.5 = 2.41248
remember that in a logit, the variance at the level-1 is fixed and equals to $\pi^2/3$. Therefore the rho in this case is equals to: 2.41248^2 /(2.41248^2 + (3.14^2/3))
```
estimates store re
xtlogit union age  south year , i(id) fe
estimates store fe

estimates table logit cluster re fe
hausman fe re, eq(1:1)
```

Addendum:

Note that after running the fixed effect model you get:

```
note: multiple positive outcomes within groups encountered.
note: 2744 groups (14165 obs) dropped because of all positive
or all negative outcomes
```

What's the meaning?

Suppose that for individual 1, there is no variation in the dependent variable over time (Y = 0 in every year). A fixed effect for this individual will perfectly predict the outcome (Y = 0). Consequently, the first individual will be dropped from the estimation sample. In fact, the fixed-effects logit model will drop all individuals that exhibit no variation in the dependent variable over time.

REMEMBER: the fixed-effects logit model is **not** equivalent to logit + dummy variables as it happens with a continuous dependent variable. When the dependent variable is binary, the required transformation is different and more complicated. If you are interested in the derivation, see the Baltagi textbook (pages 178-180). In the fixed-effects logit, the fixed effects (u_j) are not actually estimated, instead they are "conditioned" out of the model.

Addendum:

Estimating margins after xtlogit is a bit more tricky:

```
xtlogit union age  south year , i(id) re

di 1/(1+exp(-( -2.825538 +  30.43221*.016448 +0*(-1.006305)+
79.47137*.0039759 )))
di 1/(1+exp(-( -2.825538 +  30.43221*.016448 +1*(-1.006305)+
79.47137*.0039759 )))

# or alternatively (assuming that u_j =0):

margins, at(south=(0 1) (mean)_all) predict(pu0)

# otherwise:

margins, at(south=(0 1) (mean)_all)
```

4. Influential Observations

So far, we have seen how to detect potential problems in model building. We will focus now on detecting **potential observations** that have a significant impact on the model. In OLS regression, we have several types of residuals and influence measures that help us understand how each observation behaves in the model, such as if the observation is too far away from the rest of the observations, or if the observation has too much leverage on the regression line. Similar techniques have been developed for logit regression.

Standardized Pearson residuals is one type of residual. Pearson residuals are defined to be the standardized difference between the observed frequency and the predicted frequency. They measure the **relative deviations between the observed and fitted values** (only for logit models).

Deviance residual is another type of residual. It measures the disagreement between the maxima of the observed and the fitted log likelihood functions. Since logistic regression uses the maximal likelihood principle, the goal in logistic regression is to minimize the sum of the deviance residuals. Therefore, this residual is parallel to the raw residual in OLS regression, where the goal is to minimize the sum of squared residuals (both for logit and probit models).

Another statistic measures the **leverage of an observation.** An observation with an extreme value on a predictor variable is called a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. These leverage points can have an effect on the estimate of regression coefficients. Large values indicate covariate patterns far from the average covariate pattern that can have a large effect on the fitted model even if the corresponding residual is small.

These statistics are considered to be the basic building blocks for logit regression diagnostics. We always want to inspect these first. They can be obtained from Stata after the **logit** command. A good way of looking at them is to graph them against either the predicted probabilities or simply case numbers. Let us see them in an example.

```
logit vote_2004 educ age income_hh
predict p
predict stdres, rstand
scatter stdres p, mlabel(V040001) ylab(-4(2) 16) yline(0)
scatter stdres V040001, mlab(V040001) ylab(-4(2) 16) yline(0)

predict dv, dev
scatter dv p, mlab(V040001) yline(0)
scatter dv V040001, mlab(V040001)

predict hat, hat
scatter hat p, mlab(V040001)  yline(0)
scatter hat V040001, mlab(V040001)
```

As you can see, we have produced two types of plots using these statistics: the plots of the statistics against the predicted values, and the plots of these statistics against the index id (it is therefore also called an index plot) These two types of plots basically convey the same information. The data points seem to be more spread out on index plots, making it easier to see the index for the extreme observations. What do we see from these plots?

We see some observations that **are far away** from most of the other observations. These are the points that need particular attention. Which are the possible characteristics of such observations? They could have a very high Pearson and deviance residual. This could happen for example when the observed outcome hiqual is high but the predicted probability is very, very low (meaning that the model predicts the outcome to be 0). This leads to **large residuals**. Or they could be observations with **high leverage**.

We have seen quite a few logistic regression diagnostic statistics. Now how large does each one have to be, to be considered **influential**? That is to say, that by not including this particular observation, our logistic regression estimate will be quite different from the model that includes this observation. First of all, we always have to make our judgment based on our theory and our analysis. Secondly, there are some **rule-of-thumb cutoffs** when the sample size is large. These are shown below. When the sample size is large, the asymptotic distribution of some of the measures would follow some standard distribution. That is why we have these cutoff values, and why they only apply when the sample size is large enough. Usually, we would look at the relative magnitude of a statistic an observation has compared to others. That is, we look for data points that are farther away from most of the data points.

Measure	Value
leverage (hat value)	>2 or 3 times of the average of leverage
abs(Pearson Residuals)	> 2
abs(Deviance Residuals)	> 2

```
mean  hat
list  V040001 if hat > 3*.0048226 & hat!=.
list  V040001 if abs(stdres) > 2 & stdres!=.
list  V040001 if abs(dv) > 2 & dv!=.

scatter stdres p, yline(2 0 -2) mlabel(V040001) ylab(-4(2) 16)
scatter stdres V040001, yline(2 0 -2) mlabel(V040001) ylab(-
4(2) 16)

scatter dv p , yline(2 0 -2) mlabel(V040001) ylab(-4(2) 16)
scatter dv V040001, yline(2 0 -2) mlabel(V040001) ylab(-4(2)
16)

di 3*.0048226
scatter hat p, mlab(V040001)  yline(0.0144678)
scatter hat V040001, mlab(V040001) yline(0.0144678)
```

There is no lvr2plot command after a logit, but you can still check if you have observations with both high leverage and high deviance!

```
list   V040001 if abs(dv) > 2 & dv!=. & hat > 3*.0048226 &
hat!=.
list  V040001 if abs(stdres) > 2 & dv!=. & hat > 3*.0048226 &
hat!=
```

Pregibon's dbeta provides summary information of influence on parameter estimates of each individual observation (more precisely each covariate pattern). **dbeta** is very similar to Cook's D in ordinary linear regression. We can obtain **dbeta** using the **predict** command after the **logit** command. It is a measure of the change in the coefficient vector that would be caused by deleting an observation (and all others sharing the covariate pattern):

```
predict dbeta, dbeta
scatter dbeta V040001, mlab(V040001)
```

The last type of diagnostic statistics is related to **coefficient sensitivity**. It concerns how much impact each observation has on each parameter estimate. Similar to OLS regression, we also have dfbeta's for logistic regression.

A program called **ldfbeta** is available for download. Like other diagnostic statistics for logistic

regression, **ldfbeta** also uses one-step approximation. After the **logit** command, we can simply issue the **ldfbeta** command. It can be used without any arguments, and in that case, **dfbeta** is calculated for each predictor. It will take some time since it is somewhat computationally intensive. Or we can specify a variable, as shown below. For example, suppose that we want to know how each individual observation affects the parameter estimate for the variable **educ**.

```
logit vote_2004 educ age income_hh
ldfbeta educ
scatter DFeduc V040001, mlab(V040001)
```