V. DUMMY/QUALITATIVE/BINARY VARIABLES

Dummy/Qualitative/Binary Dependent Variable

Example: labor force participation, home ownership

Y_i = 1 if the person is in labor force
0 otherwise

Y_i = 1 if the person owns a home
0 otherwise

where $Y_i = \beta_1 + \beta_2 X_i + u_i$ (1)

These are examples of *choices that are binary* (either-or) in nature.

Interpretation of the model?

Assuming $E(u_i | X_i) = 0$, $E(Y_i | X_i) = \beta_1 + \beta_2 X_i$ (2)

Suppose P_i is the probability that Y_i=1 (success) and, therefore, (1-P_i) is the probability that Y_i=0 (failure). => $E(Y_i) = (1) P_i + (0) (1-P_i) = P_i$

So (2) => $E(Y_i | X_i) = P(Y_i=1 | X_i) = \beta_1 + \beta_2 X_i$ (3)
=> the expected value is a probability => **Linear Probability Model (LPM)**.

β_2 here measures the change in the probability of success for a change in X (holding all other factors constant, in multiple regressors situation).
Note that since $0 \leq P_i \leq 1$, we have the restriction $0 \leq E(Y_i | X_i) \leq 1$.

Problems with LPM:

- The resulting $\hat{Y}_i$ is now the estimated **probability** that the ith observation is a success. => values of $\hat{Y}_i$ should satisfy $0 \leq \hat{Y}_i \leq 1$. However, there is no guarantee that the $\hat{Y}_i$ resulting from regressing Y on X would meet this restriction.

- The above problem arises due to the fact that X affects the dependent variable – the probability of success - *linearly*. That is, the marginal or incremental effect of X on probability of success remains constant for all values of X. The linearity assumption is not very realistic either.

Other problems (surmountable)

- Non-normality of u_i => problem with the application of the usual t and F tests.
- Heteroscedasticity

  Since Y_i can only take values 1 or 0, u_i can only take two values: $1- \beta_1 - \beta_2 X_i$ and $- \beta_1 - \beta_2 X_i$. => the pdf for u_i is

$\mathbf{u_i}$	**Probability**
$1- \beta_1 - \beta_2 X_i$	$P(Y_i=1 \mid X) = \beta_1 + \beta_2 X_i$
$- \beta_1 - \beta_2 X_i$	$1- P(Y_i=1 \mid X) = 1- \beta_1 - \beta_2 X_i$

=> $v(u_i)$

$= (1- \beta_1 - \beta_2 X_i)^2 (\beta_1 + \beta_2 X_i) + (- \beta_1 - \beta_2 X_i)^2(1- \beta_1 - \beta_2 X_i)$

$= (1- \beta_1 - \beta_2 X_i) (\beta_1 + \beta_2 X_i) = (1-(E(Y_i \mid X_i)) E(Y_i \mid X_i)$ => heteroscedasticity

One can estimate a LPM using EGLS.

1. Estimate (2) by OLS
2. Compute $\hat{Y}_i$
3. Define $w_i = \sqrt{\hat{Y}_i(1-\hat{Y}_i)}$
4. Regress Y_i/w_i on $X_i/\ w_i$

Logit and Probit Models

To avoid the limitations of the LPM consider

$P(Y=1 \mid X) = F(\beta_1 + \beta_2 X)$ where F is a function where:

- F approaches 1 at a slower and slower rate as X gets very large and F approaches 0 at slower and slower rate as X gets very small, and

- $0 \leq F(z) \leq 1$ for all real numbers z.

<= Cumulative Distribution Function (cdf) of a random variable.

Various forms of F have been suggested. The most popular are:

- $F(z) = \dfrac{e^z}{1+e^z}$

  which is known as *logistic distribution*. Writing $z_i = \beta_1 + \beta_2 X_i$, we get

  $$P_i = P(Y_i=1 \mid X_i) = \frac{e^{\beta_1+\beta_2 X_i}}{1+e^{\beta_1+\beta_2 X_i}} \leftarrow \text{the } \textbf{Logit Model}.$$

- F(z) = $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_i} e^{-t^2/2}dt$

  which is a *standard normal* cdf.

  $P_i = P(Y_i\!=\!1 \mid X_i) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\beta_1+\beta_2 X_i} e^{-t^2/2}dt$ ← the **Probit Model**.

Logit and Probit models can be obtained from an underlying *latent variable model*. Let Y* be an unobserved or latent variable (quantitative/continuous, not qualitative/dummy/discrete – can think in terms of an *utility index (or propensity or ability to "succeed"*). We assume that

$Y_i^* = \beta_1 + \beta_2 X_i + u_i$

What we observe, however, is the dummy variable Y_i - a binary realization of the underlying latent variable - where

$Y_i = 1$ if $Y_i^* > 0$
$\quad = 0$ otherwise

We assume that *u is distributed either standard normal or standard logistic*. => Probability density function of u is symmetric. => 1-F(-z) = F(z) for all real number z.

So $P_i = P(Y_i\!=\!1 \mid X_i) = P\,(Y_i^* > 0_i\,) = P\,(u_i > -\,(\beta_1 + \beta_2 X_i\,))$
$= 1 - F(-\,(\beta_1 + \beta_2 X_i\,)) = F(\beta_1 + \beta_2 X_i\,)$ (4)

If u_i has a logistic distribution, we have

$$P_i = P(Y_i=1 \mid X_i) = F(\beta_1 + \beta_2 X_i) = \frac{e^{\beta_1 + \beta_2 X_i}}{1 + e^{\beta_1 + \beta_2 X_i}} \qquad (5)$$

the **Logit Model**.

If u_i is standard normal ($\sigma^2=1$), we have

$$P_i = P(Y_i=1 \mid X_i) = F(\beta_1 + \beta_2 X_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_1+\beta_2 X_i} e^{-u^2/2} du \qquad (6)$$

the **Probit Model**.

(Ceteris Paribus/Partial) Effect of X on P(Y=1 | X) in Logit or Probit Models

$$\frac{dP}{dX} = \frac{dF(\beta_1 + \beta_2 X)}{dX} = \frac{dF}{dz}\frac{dz}{dX} = f(\beta_1 + \beta_2 X)\beta_2$$

Since $P(Y_i=1 \mid X) = F(\beta_1 + \beta_2 X)$ is a cdf, $f(\beta_1 + \beta_2 X)$ is a probability density function (pdf). In Probit or Logit case F(z) is strictly increasing => f(z) >0 for all z. *Effect of X on P will always have the same sign as β_2.*

The partial effect of X on $P(Y=1 | X)$ is the highest when $f(\beta_1 + \beta_2 X)$ is maximum. For the Probit and Logit case, where the pdfs are symmetric about zero with a unique mode at zero, the largest effect occurs when $\beta_1 + \beta_2 X=0$. (These maximum values are: $f(0) = .25$ for Logit because $f(z) = e^z/(1+e^z)^2$, and $f(0) \approx .40$ for Probit because $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$.) Here $P(Y_i=1 | X_i) = F(\beta_1 + \beta_2 X_i) = .5$ and there is equal chance that the outcome will be a success or a failure. It makes sense then in this case the change in X will have the greatest effect since the decision maker is on the "border line" of decision making. Contrast the situation with situations where $\beta_1 + \beta_2 X_i$ is either very large or very small => $f(\beta_1 + \beta_2 X)$ is very small => Effect of a change of X on $P(Y_i=1 | X_i)$ is small. Consistent with the notion that when decision maker is "set" in his/her way with $P(Y_i=1 | X_i) = F(\beta_1 + \beta_2 X_i)$ near to 0 or 1, the effect of a small change in X on P will be negligible.

The same results holds true for multiple regressor situation.

Further Observations on Logit Model

Note:

1-P_i $= \dfrac{1}{1 + e^{\beta_1 + \beta_2 X_i}}$

Therefore, $\dfrac{P_i}{1-P_i} = e^{\beta_1 + \beta_2 X}$ ← **odds ratio** in favor of "success"

Taking log of both sides we get

$L_i =$ $\ln(\dfrac{P_i}{1-P_i}) = \beta_1 + \beta_2 X_i$

L_i is the log of the odds ratio ← **logit**.

- Although P goes from 0 to 1, L varies from -∞ to +∞.
- Although L is a linear function of X, P is not.

Estimation of Logit and Probit Models:

Note that both Logit and Probit models are non-linear in parameters. OLS does not work. Have to use :

- Maximum Likelihood Method

If grouped/replicated data are available, EGLS can be used.

Maximum Likelihood Method

Since the n observed Y_i are just realization of a binomial process with probabilities given by (4), we can write the likelihood function as:

$$L = \prod_{1}^{n} P_i^{Y_i}(1-P_i)^{(1-Y_i)}$$

Maximization of L or ln(L) w.r.t. β_1 and β_2 yield maximum likelihood estimators (MLE) of β_1 and β_2 . Because of the nonlinear nature of the maximization problem (even after taking log of the likelihood function) we cannot have formulas for the Logit and Probit MLE. These MLEs are generally consistent, asymptotically normal and asymptotically efficient. Asymptotic standard errors of the estimates can also be obtained. We can, therefore, construct (asymptotic) t tests and confidence intervals.

(*See SAS example. Note Proc Probit in SAS estimates the model for the event for which Y=0. Since $P(Y=0 / X) = 1-P(Y=1 / X)=1-F(\beta_1 + \beta_2 X)=F(-\beta_1 - \beta_2 X)$, to get the estimates of the coefficients of our model (that explains $P(Y=1 / X)$) we need to multiply the SAS coefficients by −1.*

Calculate the partial effects of the regressors at the mean value of the regressors)

EGLS with grouped data

Measuring Goodness of Fit

Note here the predicted value is a probability and the actual value of the dependent variable is either 0 or 1. R^2 is not a good measure of goodness of fit. Generally found low.

Alternative measures:

- Proportion of correct prediction or Count R^2
- Various Pseudo R^2, e.g., McFadden's $R^2 = 1 - \frac{\ln L_{UR}}{\ln L_R}$

Testing Multiple Hypotheses

Likelihood Ratio Test is used.

$LR = 2\ (\ln L_{UR} - \ln L_R) \sim \chi^2$ asymptotically with r d.o.f. where r is the number of restrictions.

Reject the null at $\alpha(100)\%$ level of significance if $LR > \chi^2_\alpha$