Prof. H. D. Vinod, class notes for Econometrics II students, , April 20, 1999.

Gujarati ch 16 Regression on **dummy dependent variables**

LPM= linear prob models. Just regress dummy dep. var. bought house (1,0) on income X.

Why **not** use LPM?: (1) non-normality of dependent variable and of errors is more realistic,

(2) $\exists$ heteroscedasticity $E(u_i^2)=P_i(1-P_i)$

(3) The range of estimated prob. can be outside [0,1], $\Rightarrow$ meaningless (4) R^2 is not meaningful.

Example: How is prob of owning a house related to household income (=X)? Let $Z_i=\beta' X$

LOGIT $P_i=f(\text{income})=\frac{1}{1+\exp(-Z_i)}$=Prob(Y=1 | income X) $\Rightarrow$ $1-P_i=\exp(-Z_i)/[1+\exp(-Z_i)]$

When $Z_i=-\infty$, $\exp(-Z_i)=\exp(\infty) \to \infty \Rightarrow \frac{1}{1+\exp(-Z_i)} \to (1/\infty) \to 0$. Conversely,

when $Z_i=\infty$, $\exp(-Z_i)=\exp(-\infty) \to 0 \Rightarrow \frac{1}{1+\exp(-Z_i)} \to (1/1) \to 1$. For this $P_i=f(\text{income})=\frac{1}{1+\exp(-Z_i)}$

Clearly $\frac{P_i}{1-P_i}=\exp(+Z_i)$. Take log of both sides, then logit $=\ln\frac{P_i}{1-P_i}$=log Odds-ratio$=Z_i=X\beta$

If one writes logit link function $g(\pi_i)=\log[(\pi_i)/(1-\pi_i)]$, then π_i = proportions P_i.

Alternatively, one can also use μ = expectation E(dep. variable) and link $g(\mu)=\log[\mu/(1-\mu)]$.

For each income x find n_i= # who own houses and N_i= total # having that income x, Ratio is $\hat{P}_i=(n_i/N_i)$ Now compute logit from this and then regress $w_i$logit on w_i and w_iX, where $w_i=\text{sqrt}(N_i\hat{P}_i(1-\hat{P}_i))$ The weights w_i correct for heteroscedasticity.

PROBIT: there is a "utility index" variable (unobservable) I_i determined by income X

If the person's utility exceeds a threshold level I_i^* person buys the house. Threshold is also unobservable.

$$P_i=\Pr(Y=1)=\Pr(I_i^* \le I_i)=\text{CDF}(I_i)=\int_{-\infty}^{I_i}\frac{1}{\sqrt{2\pi}}\exp[-t^2/2]dt = F(I_i)$$

$\Rightarrow F^{-1}(P_i)=I_i$ In practice, from $\hat{P}_i=n_i/N_i$= proportion owning house for each income X_i get I_i from inverse CDF of normal (quantile)

Ch 19 of Greene: p. 874 3rd edition

CDF($\beta' x$) goes from 0 to 1 as $\beta' x$ goes from $-\infty$ to ∞. Any CDF will do as Prob(Y=1).

Probit uses $\Phi(\beta' x)=\int_{-\infty}^{\beta' x}\phi(t)\,dt$, where ϕ is N(0,1) density.

Logit uses $\Lambda(\beta' x)=\exp(\beta' x)/[1+\exp(\beta' x)]$ which is the CDF of logistic distribution. (eq. 19-7)

$\exp(-\infty)=0$, so that when $\beta' x=-\infty$, this CDF is 0 and when $\beta' x=\infty$, this CDF is $(\infty/\infty)\to 1$.

Regression model says $E(y|x)=\text{CDF}(\beta' x)$ or $F(\beta' x)$

We want to interpret the regression coefficients as partials of E(y|x) with respect to x.

Note that here the partial is $\partial E(y|x)/\partial x = (\partial F/\partial x)\beta$ = density(at $\beta' x$) β

since derivative of CDF is always the density at that point.

$\partial\Lambda/\partial(\beta' x)=\frac{A}{(1+A)^2}$ where A= exp($\beta' x$). Also, $\partial\Lambda/\partial(\beta' x)=\frac{A}{(1+A)^2}=\Lambda(\beta' x)[1-\Lambda(\beta' x)]$=PQ,

where P =Prob (Y=1) and Q= prob(Y=0)

Hence for the logistic $\partial E(y|x)/\partial x = PQ\beta$, which is convenient.

Remark 1: The generalized least squares (GLS) is extended into the general linear model (GLM) in three steps, McCullagh and Nelder (1989). Let y be the dependent variable and X matrix of regressors.

(i) Instead of $y \sim N(\mu, \sigma^2 I)$ we allow non-normal distributions with various relations between mean and variance functions. Non-normality permits the expectation $E(y)=\mu$ to take on values only in a meaningful restricted range (*e.g.*, nonnegative integer counts or binary (0,1) outcomes).

(ii) Define a systematic component $\eta = X\beta = \sum_{j=1}^{p} x_j\beta_j$, as a linear predictor of the dependent variable. $X\beta \in (-\infty, \infty)$ is a linear function of regressors in columns of X matrix.

(iii) A monotonic differentiable link function $\eta = g(\mu)$ relates the mean $\mu = E(y)$ to the systematic component $X\beta$. The *t-th* observation satisfies $\eta_t = g(\mu_t)$. For GLS, the link function is identity, or $\eta = \mu$, since $y \in (-\infty, \infty)$. When *y* data are counts of something, we need a link function which makes sure that $X\beta = \mu > 0$. Similarly, for *y* as binary (dummy variable) outcomes, $y \in [0,1]$, we need a link function $g(\mu)$ which maps the interval [0,1] for *y* or for its mean μ on the unrestricted interval $(-\infty, \infty)$ for $X\beta$. For example, y can be binary or dummy dependent variable taking only two values, 0 or 1 or success/failure of the Binomial family. Then $g(\mu)=\log[\mu/(1-\mu)]$ is the logit link function

Remark 2: To obtain generality, the normal distribution is often replaced by a member of the exponential family of distributions, which includes Poisson, binomial, gamma, inverse-Gaussian, etc. It is well known that "sufficient statistics" are available for the exponential family. In our context, $X'y$ which is a p×1 vector similar to β, is a sufficient statistic. A "canonical" link function is one for which a sufficient statistic of p×1 dimension exists. Some well known canonical link functions for distributions in the exponential family are: $g(\mu)=\mu$ for the Normal, $g(\mu)=\log\mu$ for the Poisson, $g(\mu)=\log[\mu/(1-\mu)]$ for the Binomial, and $g(\mu) = -1/\mu$ is negative for the gamma distribution, respectively.

Venables & Ripley page 184

What is a Link Function? Function of μ which equals $X'\beta$ or similar convenient function of regressors
if link $g(\mu) = X'\beta$ Then g is the link function
If $\mu = h(X'\beta)$ then link function is the inverse of h
If the distribution of Y has a density from exponential family:

$$f(y,\theta,\varphi) = \exp\left[A_i\{y_i\theta_i - \gamma(\theta_i)\}/\varphi + \tau(y_i, \varphi/A_i)\right]$$ where A_i are known prior weights and parameter

θ_i controls the distribution of y_i and is also an invertible function of μ. In fact $\theta=(\gamma')^{-1}(\mu)$

Binomial family $\Rightarrow$ Canonical link $\log(\mu/(1-\mu))$ or logit with variance $\mu(1-\mu)$
link can be probit or cloglog, but they are not canonical.
Canonical link means it is minimal sufficient statistic. e.g. for Gaussian family, $X'y$ is minimal sufficient for β link function for which $\theta=X'\beta$ is called canonical

Greene ch. 14 on panel data has continuous dependent variable y_{it}. here we have limited dep. var.
In the tradition of econometrics a Hausman-type specification test for fixed versus random effects has been suggested ln C = 0.67 lnY +.... 6 dummy variables for 6 groups gives least sq. dummy var. (LSDV) model. Thus b=0.67 Its SE=0.61
In a random effects model error become $u_i + \varepsilon_{it}$ and covariances are assumed zero
(WHY zero covar is n9t clear) page 624 of Greene has Ω mtx spelled out with diag $=\sigma^2_\varepsilon + \sigma^2_u$ all and off-diag all elements $= \sigma^2_u$
I we define $\theta = 1 - \sigma_\varepsilon[T\sigma^2_u + \sigma^2_\varepsilon]^{-0.5}$ and use weighted differencing from grp mean $y_{i1} - \theta\bar{y}_{i.}$ from within group means is regressed on similar quantity for regressor Y
weighted diff lnC = intercept + 0.796 wtd diff of Y from group mean. Thus $\hat{\beta}$=0.6796 (SE=0.042)

Hausman st'c = (b • s)^2 , [var(b) • var(s)] = 7.7 so reject Hausman by Chi-sq with df=1

Ref: **<u>Johnston and DiNardo</u>** Econometric Methods, 4th ed., McGraw Hill 1997
has excellent chapter on limited dependent variables.

sec 13.7 p 430 extension of basis model for discrete and limited dep. var. models
prob (y_i=1) = CDF F(f(X_i)/5) where f(X)=X" , F is cumulative normal for probit, logistic for logit.
Evaluate at the mean and compare the regression coefficients with caution. Let $\bar{p}$=prop of ones.
" (from Lin Prob. model) ¸ " (from a Logit model) $\bar{p}(1 • \bar{p})$
or compare
9($\bar{X}_i$") " (from probit model) ¸ " (from a Logit) $\bar{p}(1 • \bar{p})$
where 9 is read off the z tables N(0,1) when mean index is calculated.
Better comparison formula is: 9[F^{-1}($\bar{p}$)]" (from probit) ¸ " (from Logit) * $\bar{p}(1 • \bar{p})$
Note that this comparison does not require mean index at all.
<u>Grouped Data</u>:
J classes, where X are constant within a class, y_i is binary, prob(y_i=1)=F(X_i")
ln-lkhd=$D_{i\ in\ N}$ {y_i ln F + (1 • y_i) ln (1 • F)}
Original ln-lkhd simplifies. We can replace y_i by p_j where p_j=prop of ones in j-th class
New ln-lkhd=$D_{j\ in\ J}$ {p_j ln F + (1 • p_j) ln (1 • F)}
where n_1 to n_J denotes # of tems in each relevant class, and where p_j=(1/n_j) $D_j y_i$

Fully saturated model has J parameters (separate for each class) $\$_j$ then replace F by $\$_j$
and multiply by n_j. The ML estimator of $\$_j$ is simply p_j.
new ln-lkhd=$D_{j\ in\ J}$ {p_j ln $\$_j$ + (1 • p_j) ln (1 • $\$_j$)}$n_j$

<u>Likelihood ratio</u> test = • 2(ln-lkhd with F • ln-lkhd with fully saturated model)
1_j=pop prop of those who experienced the event in j-th class
assuming that the no. of items in cell grows at a constant rate n_j/N Ä q_j and N Ä _
E(p_j)=1_j and var(p_j)=(1/n_j)1_j(1 • 1_j), this var is max when 1_j=0.5
For the Linear prob model E(p_j)=X" hetero is obvious.

<u>Minimum Chi-sq</u> method of estimation is calculated by using OLS software
weight are simply the reciprocals of the square roots of variance terms.
<u>Variance formulas:</u>
Linear prob. model: p_j=X" dep. var.=p_j, var=(1/n_j)$p_j q_j$
Log-linear: p_j=exp(X") dep var.=log(p_j), var=(q_j/$n_j p_j$)

Probit: p_j=F(X") dep var.=F^{-1}(p_j), var=[1/n9(p_j)$^{\#}$]$p_j q_j$

Logit: p_j=A(X_j"), dep var=log(p_j/q_j), var=1/($n_j p_j q_j$)

ORDERED PROBIT: y_i^n=1 if person does not work, y_i^p=1 if part-time work, y_i^f=1 if full-time
y* is an indicator variable= choice of work status so that higher y* means he is likely to work.
c_1 threshold below which does not work (e.g. c_1=0)
If y* is between c_1 to c_2 then the person is part time

If y^* is larger than c_2, this means the person is full time.
We write the right hand side of regression as simply $X_i\beta=\alpha+\beta z_i$ with an intercept and slope.
prob $(y_i^n=1)=\Phi(\frac{c_1-\alpha-\beta z_i}{\sigma})$ for non-workers

prob $(y_i^p=1)=\Phi(\frac{c_2-\alpha-\beta z_i}{\sigma}) - \Phi(\frac{c_1-\alpha-\beta z_i}{\sigma})$, for part-time workers

prob $(y_i^f=1)=\Phi(\frac{c_2-\alpha-\beta z_i}{\sigma})$, for full-time workers.

In the above model, without loss of generality (WLOG) we can set $c_1=0$ (why?)
Hence only one threshold needs to be estimated called c and we can identify (c/σ), α/σ and (β/σ).
Again, just like probit, we can identify parameters up to some factor of proportionality.
For example. if we multiply c and σ both by 999, the ratio (c/σ) is the same, similarly for other ratios.
one complication for ordered probit (not for usual probit) is that partial derivative of prob(some y=1) w.r.t. z (the explanatory variable) for part-time case depends on the threshold c.
If z goes up (β is positive) prob. of working should go down, but the derivative has ambiguous sign for part-time work, $\rightarrow$ this is a problem!

TOBIT as an extension of probit model
y^* = index of a man's desire for a car, $y_i=1$ if he buys a car. Note that y^* is not observable.
The observable $y_i=y_i^*$ only if $y_i^*>0$,
$=0$ if $y_i^* \le 0$. If the desire is negative, it is not observed at all.
This is called **censored** regression model, since $y_i=\max(0, X_i\beta+\varepsilon_i)$, censors the observable data on the dependent variable only. The data on RHS regressors need not be missing. More importantly, censored data are reported erroneously, all those with $y_i^* \le 0$ are reported as if they are at 0.

Truncated data are not same as censored. **Truncated** data means that both the dependent variable and regressor variable data (both LHS and RHS of regression) may not be observed.
This is a characteristic of the underlying distribution.
$y_i^*=X_i\beta+\varepsilon_i$, $\varepsilon \sim N(0,\sigma^2)$, $y_i=1$ if $y^*>0$ and $y_i=0$ if $y^* \le 0$.
when is prob$(y^*<0)$?
$y_i^*<0 \Leftrightarrow$ RHS$=X_i\beta+\varepsilon_i<0$ or $-X_i\beta \le \varepsilon_i$
or $(-X_i\beta/\sigma) \le (\varepsilon_i/\sigma)$, since scale does not matter.$=1-\Phi[(-X_i\beta/\sigma)]$
The log-lkhd has 2 parts: (One part similar to probit) times (a part similar to usual OLS).
Normalizing by $\sigma=1$ is not harmless here (it is harmless for probit).
This is also true for OLS regression.
The coefficients of TOBIT are not interpreted the same way as usual regr. coeff due to censoring. It is true that
$\partial E(y_i^*|X_i)/\partial x_k = \beta_k$,

However, if y_i is inserted instead of y_i^* (due to censoring) we have $\Phi(\frac{X_i\beta}{\sigma})*\beta_k$

Further if we impose the condition that $y_i^*>0$ is given, then the partial of $E(y_i \mid X_i$ and $y_i^*>0)$ wrt x_k has 3 terms $\beta_k[1-T2-T3]$,

where $T2=\frac{X_i\beta}{\sigma}\{\phi(\frac{X_i\beta}{\sigma})/\Phi(\frac{X_i\beta}{\sigma})\}$ and where $T3=\{\phi(\frac{X_i\beta}{\sigma})/\Phi(\frac{X_i\beta}{\sigma})\}^2$.

If censoring is more of an annoyance rather than funda. aspect of the relation, then simple β_k is ok.
In gen. McDonald and Moffit decompose the partial into 2 parts:

effect on conditional mean + effect on prob. that an obs. will be positive.

Let x and y* be joint normal, $y^*=X\beta+\varepsilon$. If censoring problem is ignored we have inconsistency. Why? plim β_{ols}=[β times the prob(y*>0)]. Since any prob must be <1, this plim=(a fraction times β) Á β. If we use $\beta_{consistent}=\beta_{ols}(N/n_1)$ this undoes the bias, and becomes consistent only if joint normality holds.

CASE where censoring "problem" can be ignored:
C_i=no of cigarettes smoked
T_i=1 if restrictions on smoking at work exist.
=0 otherwise
Cigarette co. is interested in # of cigarettes smoked $E(C_i|T_i)$ not on condition that $C_i>0$ or in the prob that $C_i>0$ given T_i. here OLS is consistent estimate of average treatment effect.

Average # of cigarettes smoked by those who work at places where there is NO restriction MINUS average # of cigarettes smoked by those who work at places where there IS a restriction.

In more complicated examples, where RHS variables are not binary, etc., things are complicated.
Tobit imposes the condition that the relationship generating the ones and zeros is the same as the process that produces the positive values.
If TOBIT is well specified, ML estimate from TOBIT for $\hat{\beta}/\hat{\sigma}$ should be the same as estimate of the probit coeff. from the same data treating nonzero values as 1 and 0 values as 0.
(any amount of dollars spent on car should be made 1 and probit used)
If probit and Tobit results are very different, something is wrong. This is a specification test.

TOP coding in standard wage eq. If wage is too big, e.g. hourly wage > 999, then for privacy reasons the wage data is omitted. Hence observed wage is
$y_i=\min(999, X_i\beta+\varepsilon_i)$. Of course, this is better than throwing away the data that are top coded.