QUALITATIVE VARIABLES

A. Qualitative Regressors

(Dummy Variables)

— rhs "exogenous" variables for which numerical values are arbitrary.

⇐ * Labor economists use these a lot.

Example 1: $\omega_t = \alpha + \beta_1 E_t + \beta_2 A_t + \beta_3 S_t + \epsilon_t$

(educ.) (age) (sex)

where $S_t = \begin{cases} 1 & \text{if } t^{th} \text{ obs. is male} \\ 0 & \text{if } t^{th} \text{ obs. is female} \end{cases}$

Thus S_t is a "qualitative" variable; expresses a <u>quality</u> (not quantity) that has some impact on the dependent variable (the wage rate).

This model presumes that the intercept is different for males and females.

for females: $\omega_t = \alpha + \beta_1 E_t + \beta_2 A_t + \epsilon_t$

for males: $\omega_t = (\alpha + \beta_3) + \beta_1 E_t + \beta_2 A_t + \epsilon_t$

$\frac{\partial \omega_t}{\partial S_t} = \beta_3 \rightarrow$ amount that mean wage rate differs for males & females [wage differential].

This model specification further presumes that the difference this variable makes (S_t) is "constant." ∄ no interaction with other variables.

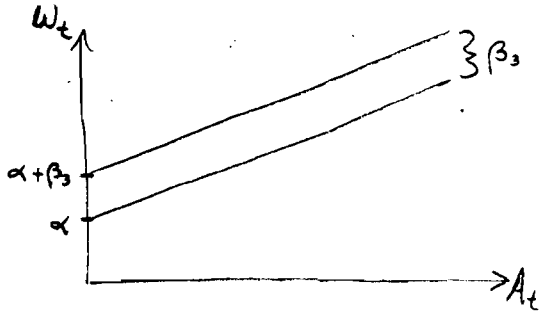
To incorporate interaction, add the variable, $\beta_4 (S_t E_t)$. See example 2.

✶ Note: Numerical values of dummies are irrelevant; results will be the same.

☆ Effect: Residuals will sum to zero for each classification (males & females).

Proof: Recall the Normal Equations,

$$\sum_t e_t X_{it} = 0 \; ; \qquad \sum_t e_t = 0 \quad \left(\text{since } \exists \text{ a constant term}\right)$$



and $\sum_t e_t S_t = 0$

[This splits residuals into the two groups; each of which must sum to zero.]

Implication

☆ Very similar to running 2 separate regressions
- one for males, one for females;
- except β & β_- reflect impact wrt entire sample

(more to follow)

Example 2: Interaction terms

$$w_t = \alpha + \beta_1 E_t + \beta_2 A_t + \beta_3 S_t + \beta_4 (S_t E_t) + \epsilon_t$$

for females, $\frac{\partial w_t}{\partial E_t} = \beta_1$

for males, $\frac{\partial w_t}{\partial E_t} = \beta_1 + \beta_4$

} Education may ↑ w_t for males more than females.

This interaction term allows measurement of the impact of education on the two separate groups in the sample

Interpretation; β_3 = avg. male-female wage differential

β_4 = (differential) the effect of Education on the wage rate of males & females.

If we think Age affects the wage rate of males and females differently, we can add another interaction term;

$$w_t = \alpha + \beta_1 E_t + \beta_2 A_t + \beta_3 S_t + \beta_4 (S_t E_t) + \beta_5 (S_t A_t) + \epsilon_t$$

→ β_5 = the differential effect of Age on the wage rate of males and females.

Note: Running this last Regression, with all possible interaction terms (re: S_t), is (almost) the same as running two separate regressions — one for males & one for females.

(i) The coefficients will be identical.
i.e. $\frac{\partial W_t}{\partial E_t}_{(\text{females})} = \hat{\beta}_1$; $\frac{\partial W_t}{\partial E_t}_{(\text{males})} = \hat{\beta}_1 + \hat{\beta}_4$,

etc. whether run with [illegible] regressions

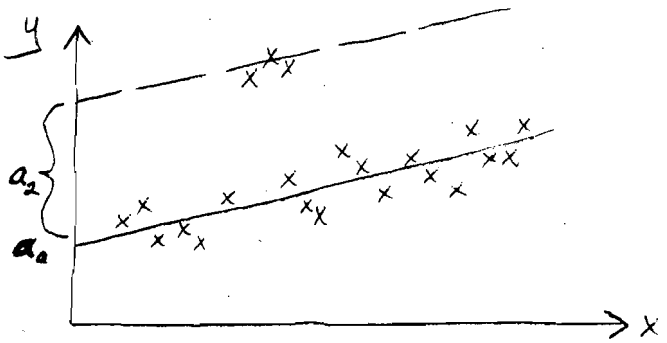
(ii) When running two separate regressions, the estimates of the variance of the residuals (s^2_{male} & s^2_{female}) need not be identical.
Running one regression with all relevant interaction terms presumes that they are identical ($s^2_{male} = s^2_{female}$).

Thus, even though the corresponding estimates will be identical with these two alternatives, the t-ratios may not be.
(since the standard errors depend on s^2!)

These procedures actually represent another method of statistical analysis called Analysis of Variance.

Example 3: Outlier Dummies

e.g. | unemployment rates during a year of a strike;
| commodity prices during a period of an embargo;
| production, ..., during years of a war; etc.



Could include a Dummy with a value of 1 for the outliers. This would result in a constant term that shifts the entire regression line up thru the outliers, when they are effective.

$$y = a_0 + a_1 x + a_2 D + \epsilon_t$$

Alternatively explained, this brings the outliers down about the regression line, so that they are not "outliers" anymore.

R^2 may = .5 w/o Dummies
and may = .9 with Dummies.

Big Deal! Implies, w/o Dummies can "explain" 50% of variation; with Dummies can "explain" 90%, BUT 40% of that is due to (weird) attributes of outliers!

This could be useful, if you know why the outliers should be outliers!
e.g. strike, embargo, war, ...

Note: R^2 is improved immensely with Dummies.
If you "throw out" the obs. that don't fit, then the fit of everything else will improve!

[R^2 bigger; SSE smaller; std. errors smaller, ...]

Note: If I just one outlying observation that is Dummied out, the regression will bring the outlier right down onto the regression line.
This will not affect the slope or position of the fitted line;
∴ effectively throwing out that obs.!
→ Not "explaining" anything new with Dummy

POINT: Use Caution!
Know what you're doing!
Don't "throw out" outliers without knowing why; you had better be able to tell a good story as to why those obs. are outliers (& ∴ should be "thrown out").

Don't use Dummies and then brag about your R^2!

Example 4: Seasonal Dummies

$$C_t = \alpha + \beta y_t + \gamma_1 S_{1t} + \gamma_2 S_{2t} + \gamma_3 S_{3t} + \epsilon_t$$

where $S_{1t} = \begin{cases} 1 & \text{during first quarter} \\ 0 & \text{otherwise} \end{cases}$

$S_{2t} = \begin{cases} 1 & \text{during second quarter} \\ 0 & \text{otherwise} \end{cases}$

$S_{3t} = \begin{cases} 1 & \text{during third quarter} \\ 0 & \text{otherwise} \end{cases}$

This holds the slope, β, constant for all 4 quarters, but allows a different intercept for different quarters.

Computer Problem 4: Seasonality Study

Given below is monthly data on Visits to National Parks (V), Personal Income (Y), U. S. Population (N), and the Consumer Price Index (P), for the years 1971 through 1976.

1. From this data you will need to construct the following time series:

Log (V/N); Log (Y/(N*P));

$$CMA_t = (.5*V_{t-6} + V_{t-5} + V_{t-4} + V_{t-3} + V_{t-2} + V_{t-1} + V_t + V_{t+1} + V_{t+2} + V_{t+3} + V_{t+4} + V_{t+5} + .5*V_{t+6})/12$$

(Note that this series will run from July, 1971 through June, 1976).

$$RCMA_t = \frac{V_t}{CMA_t}$$ (over the same time period)

From RCMA, calculate by hand the seasonal adjustment factors, $\overline{RCMA}_i$;

i = Jan., Feb.,..., Dec.

Then construct the Seasonally Adjusted Visits series;

$SAV_t = V_t/\overline{RCMA}_t$ (from July, 1971 through June, 1976)

2. Next run the following regressions:

(1) Log (V/N) = $a_o + a_1$ Log (Y/(P*N)) + ϵ_1

(2) Log (V/N) = $a_o + a_1$ Log (Y/(P*N)) + $\sum_{i=1}^{11} b_i S_i + \epsilon_2$

where $S_i = \begin{cases} 1 & \text{in month } i; \\ 0 & \text{otherwise} \end{cases}$

(3) Log (SAV/N) = $a_o + a_1$ Log (Y/(P*N)) + ϵ_3

Note: Run equations (1) and (2) twice; once over the entire sample, and once over the abridged sample, so that they may be compared with the third regression.

Compare and contrast the results (SSE, R^2, $\hat{a}_1$, t-ratios, etc.) for these three equations, and explain why the results vary.

3. In all cases plot the Actual values $(\frac{V}{N})$ over the Fitted values $(\frac{\hat{V}}{N})$, and the residuals against time.

4. Compare the seasonal adjustment factors computed by hand ($\overline{RCMA}_i$, i = Jan., Feb.,..., Dec.) with those implied by equation (2), the dummy regression. Note that you must calculate the latter ($e^{\hat{b}_i}$ = seasonal adjustment factor for month i; i = Jan., Feb.,..., Dec.).

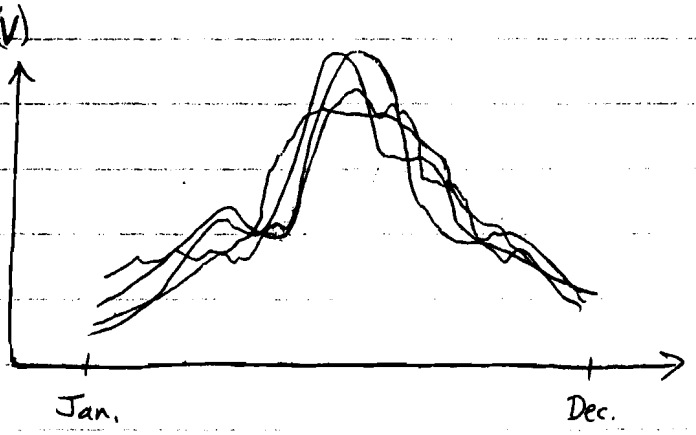
	V		SAV	Y	N	P	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	
Jan	1198		3583.4	831.8	206076	119.2	1	0	0	0	0	0	0	0	0	0	0	0
Feb	1347		3558.5	834	206242	119.4	0	1	0	0	0	0	0	0	0	0	0	0
Mar	1689		3364.2	840.7	206393	119.8	0	0	1	0	0	0	0	0	0	0	0	0
Apr	2609		4051.0	845.7	206567	120.2	0	0	0	1	0	0	0	0	0	0	0	0
May	3653		3766.9	850.1	206726	120.8	0	0	0	0	1	0	0	0	0	0	0	0
June	6725		4130.1	869.1	206891	121.5	0	0	0	0	0	1	0	0	0	0	0	0
July	10266	4075.04	4121.2	860	207053	121.8	0	0	0	0	0	0	1	0	0	0	0	0
Aug	9602	4086.75	4210.9	865.6	207237	122.1	0	0	0	0	0	0	0	1	0	0	0	0
Sept	4978	4115.96	3977.5	868.6	207433	122.2	0	0	0	0	0	0	0	0	1	0	0	0
Oct	3417	4148.63	3664.2	872.6	207627	122.4	0	0	0	0	0	0	0	0	0	1	0	0
Nov	1931	4191.38	3710.4	880.3	207800	122.6	0	0	0	0	0	0	0	0	0	0	1	0
Dec	1246	4244.29	2458.8	891.3	207949	123.1	0	0	0	0	0	0	0	0	0	0	0	0
Jan	1273	4289.46	3607.7	902.4	208088	123.2	1	0	0	0	0	0	0	0	0	0	0	0
Feb	1553	4337.04	4102.7	914.6	208196	123.8	0	1	0	0	0	0	0	0	0	0	0	0
Mar	2184	4389.71	4350.2	920.1	208310	124	0	0	1	0	0	0	0	0	0	0	0	0
Apr	2898	4437.33	4499.7	927	208447	124.3	0	0	0	1	0	0	0	0	0	0	0	0
May	4390	4482.08	4528.6	931.7	208569	124.7	0	0	0	0	1	0	0	0	0	0	0	0
June	7258	4488.83	4457.4	923	208712	125	0	0	0	0	0	1	0	0	0	0	0	0
July	10819	4522.36	4342.4	940.3	208846	125.5	0	0	0	0	0	0	1	0	0	0	0	0
Aug	10393	4550.63	4464.6	949.7	208966	125.7	0	0	0	0	0	0	0	1	0	0	0	0
Sept	5651	4585.75	4515.2	954	209153	126.2	0	0	0	0	0	0	0	0	1	0	0	0
Oct	3667	4567.67	4188.2	971.7	209317	126.6	0	0	0	0	0	0	0	0	0	1	0	0
Nov	2055	4624.92	3948.7	984.1	209457	128.9	0	0	0	0	0	0	0	0	0	0	1	0
Dec	1716	4658.08	3386.2	992	209584	127.3	0	0	0	0	0	0	0	0	0	0	0	0
Jan	1656	4640.21	4953.3	1000.2	209711	127.7	1	0	0	0	0	0	0	0	0	0	0	0
Feb	1646	4803.29	4882.1	1012.8	209809	128.6	0	1	0	0	0	0	0	0	0	0	0	0
Mar	2252	4597.79	4485.6	1022.4	209905	129.6	0	0	1	0	0	0	0	0	0	0	0	0
Apr	3356	4607.67	5210.9	1031.1	210034	130.7	0	0	0	1	0	0	0	0	0	0	0	0
May	4626	4627.36	4976.4	1037.8	210154	131.5	0	0	0	0	1	0	0	0	0	0	0	0
June	7618	4626.46	4678.5	1045.7	210266	132.4	0	0	0	0	0	1	0	0	0	0	0	0
July	10030	4602.63	4025.7	1054.1	210410	132.7	0	0	0	0	0	0	1	0	0	0	0	0
Aug	10296	4571.46	4423.2	1064	210556	135.1	0	0	0	0	0	0	0	1	0	0	0	0
Sept	5616	4544.00	4487.3	1074.8	210715	135.5	0	0	0	0	0	0	0	0	1	0	0	0
Oct	4159	4512.13	4459.9	1086.2	210863	136.6	0	0	0	0	0	0	0	0	0	1	0	0
Nov	2258	4462.75	4334.9	1096.7	210984	137.6	0	0	0	0	0	0	0	0	0	0	1	0
Dec	1493	4399.38	2946.2	1103.6	211097	138.5	0	0	0	0	0	0	0	0	0	0	0	0
Jan	1307	4358.08	3909.4	1104.2	211207	139.7	1	0	0	0	0	0	0	0	0	0	0	0
Feb	1449	4351.50	3828.0	1109.6	211311	141.5	0	1	0	0	0	0	0	0	0	0	0	0
Mar	1992	4357.00	3967.7	1117	211411	143.1	0	0	1	0	0	0	0	0	0	0	0	0
Apr	2851	4369.67	4428.7	1127	211522	143.9	0	0	0	1	0	0	0	0	0	0	0	0
May	4148	4365.71	4275.2	1141.1	211637	145.5	0	0	0	0	1	0	0	0	0	0	0	0
June	6777	4397.86	4162.0	1152.9	211772	146.9	0	0	0	0	0	1	0	0	0	0	0	0
July	9832	4423.56	3946.3	1168.1	211901	146	0	0	0	0	0	0	1	0	0	0	0	0
Aug	10364	4450.08	4481.0	1173.8	212051	149.9	0	0	0	0	0	0	0	1	0	0	0	0
Sept	5660	4476.86	4522.4	1180.9	212216	151.7	0	0	0	0	0	0	0	0	1	0	0	0
Oct	4419	4495.08	4738.7	1191.9	212383	153	0	0	0	0	0	0	0	0	0	1	0	0
Nov	2381	4502.38	4575.1	1152.9	212518	154.3	0	0	0	0	0	0	0	0	0	0	1	0
Dec	1860	4563.29	3275.7	1199.1	212637	155.4	0	0	0	0	0	0	0	0	0	0	0	0
Jan	1757	4757.42	5255.44	1199.6	212746	156.1	1	0	0	0	0	0	0	0	0	0	0	0
Feb	1835	4946.96	4319.37	1204.7	212844	157.2	0	1	0	0	0	0	0	0	0	0	0	0
Mar	2497	4996.68	4973.62	1211	212939	157.8	0	0	1	0	0	0	0	0	0	0	0	0
Apr	2735	5014.00	4246.83	1216.8	213056	158.6	0	0	0	1	0	0	0	0	0	0	0	0
May	4437	5030.36	4575.30	1226.9	213167	159.3	0	0	0	0	1	0	0	0	0	0	0	0
June	7946	5125.00	4881.18	1258.8	213393	160.8	0	0	0	0	0	1	0	0	0	0	0	0
July	13320	5213.87	5348.22	1256.2	213559	162.3	0	0	0	0	0	0	1	0	0	0	0	0
Aug	11445	5240.04	4916.77	1271.3	213741	162.8	0	0	0	0	0	0	0	1	0	0	0	0
Sept	5845	5256.00	4670.22	1281.5	213900	163.8	0	0	0	0	0	0	0	0	1	0	0	0
Oct	4597	5276.88	4929.58	1294.9	214055	164.8	0	0	0	0	0	0	0	0	0	1	0	0
Nov	2596	5325.38	4986.22	1304.4	214200	165.8	0	0	0	0	0	0	0	0	0	0	1	0
Dec	3718	5334.58	7332.68	1312.9	214321	166.3	0	0	0	0	0	0	0	0	0	0	0	0
Jan	1829	5246.88	5470.80	1326.9	214448	166.7	1	0	0	0	0	0	0	0	0	0	0	0
Feb	2196	5144.42	5801.42	1336.9	214562	167.1	0	1	0	0	0	0	0	0	0	0	0	0
Mar	2319	5149.88	4619.07	1346.3	214657	167.5	0	0	1	0	0	0	0	0	0	0	0	0
Apr	3462	5167.50	5375.45	1359.5	214765	168.2	0	0	0	1	0	0	0	0	0	0	0	0
May	4826	5196.42	4976.42	1367.9	214877	169.2	0	0	0	0	1	0	0	0	0	0	0	0
June	7780	5121.13	4777.99	1372.7	215014	170.1	0	0	0	0	0	1	0	0	0	0	0	0
July	11383	4967.13	4568.77	1386.2	215142	171.1	0	0	0	0	0	0	1	0	0	0	0	0
Aug	10923	4799.42	4892.52	1393.7	215300	171.9	0	0	0	0	0	0	0	1	0	0	0	0
Sept	6498	4811.29	5191.98	1401.8	215458	172.6	0	0	0	0	0	0	0	0	1	0	0	0
Oct	4647	4370.42	5197.66	1414.2	215623	173.3	0	0	0	0	0	0	0	0	0	1	0	0
Nov	2606	4025.08	5011.28	1432.1	215764	173.8	0	0	0	0	0	0	0	0	0	0	1	0
Dec	1849	3499.83	3648.68	1450.2	215893	174.3	0	0	0	0	0	0	0	0	0	0	0	0

Problem #5

SEASONAL ADJUSTMENT

Consider the series, Visits to National Parks.
— monthly data, 6 yrs.; Bus. STATISTICS (S.C Bus.)

If you plot variations during each year, a seasonal pattern appears.



During any year, ∃ systematic variation in the series which is not due to any explanatory variable such as income or population, but is due, rather, to seasonality.

Want to consider the relationship between $\left(\frac{\text{Visits}}{N}\right)$ and Real Personal Income Per Capita

If we regress:

$$\ln\left(\frac{V}{N}\right) = a_0 + a_1 \ln\left(\frac{Y}{CPI * N}\right)$$

we will observe the relationship between variations in $\frac{Y}{CPI*N}$ and $\frac{V}{N}$;
$\frac{V}{N}$ will display much noise during each year, not due to variations in $\frac{Y}{CPI*N}$ but due to seasonality.

We are really interested in the relationship between the number of Visits in the typical, avg. month and income, without regard to seasonality (or, holding seasonality constant).

We can capture this desired result by:

i) Seasonally Adjusting $\frac{V}{N}$ by hand, or

ii) Fitting $\ln \frac{V}{N} = a_0 + a_1 \ln \frac{Y}{CPI * N} + b_1 D_1 + b_2 D_2 + \cdots + b_{11} D_{11} + b_{12} I$

i) By Hand.

We want the Index; I_i = Visits in the i^{th} month as % age of avg month

Step 1: Construct a 12-month moving avg.

e.g. observation for June;

Dec. Jan. Feb. Mar Apr. May [Je.] July Aug Sept. Oct. Nov. Dec.

$CMA_{Je.} = \frac{1}{2}$ Dec. + Jan. + Feb. + ... + Oct. + Nov. + $\frac{1}{2}$ Dec.

SMPL 7, 66 $ → [72 obs., 1st six & last six used up in MA]

⇒ GENR CMA = [.5 * V(-6) + V(-5) + ... + V(4) + V(5) + .5 * V(6)] / 12. $

(centered moving avg)

Step 2: Then

GENR RCMA = $\frac{U}{CMA}$ $

↳ Ratio of Visits in any month, to the CMA.
RCMA should vary around 1;
$RCMA_{Je}$ should be > $RCMA_{Jan}$!

We have 5 years of observations. [lose 6 mo. at front & 6 mo. at end]
→ ∴ 5 obs. of $RCMA_{Jan.}$
5 " " $RCMA_{Feb.}$
etc.

✳ We are interested in the Average # of Monthly Visits
in each month,
as a percentage of the avg # of monthly visits
in the years centered by that month.

Year	1	2	3	4	5	6	Averages ← Step 3
Jan.	~	~				~	→ $\overline{RCMA}_{Jan}$
Feb.	~	~				~	→ $\overline{RCMA}_{Feb}$
Mar.	~	~				~	→ $\overline{RCMA}_{Mar}$
Apr.	~	~				~	→ $\overline{RCMA}_{Apr}$
May	~	~				~	→ $\overline{RCMA}_{May}$
Je.	~	~				~	→ $\overline{RCMA}_{Je.}$
July	~	~					→ $\overline{RCMA}_{July}$
Aug.	~	~					→ $\overline{RCMA}_{Aug.}$
Sept.	~	~					→ $\overline{RCMA}_{Sept.}$
Oct.	~	~					→ $\overline{RCMA}_{Oct.}$
Nov.	~	~					→ $\overline{RCMA}_{Nov}$
Dec.	~	~					→ $\overline{RCMA}_{Dec.}$

Month

[We want these Avgs]

These tell us the relative seasonal expectation of visits for [illegible] month

These $\overline{RCMA_i}$ should again fluxuate around 1.0

Furthermore, $\sum_{i=1}^{12} \overline{RCMA_i}$ should $\approx$ 12. (or 1200 if $\times 100$)

Note: $\sum_{i=1}^{12} \overline{RCMA_i}$ will not exactly $= 12$, because of the small sample, and the gaps at the front and back.

Suppose $\sum_{i=1}^{12} \overline{RCMA_i} = 15$;

Then adjust each $\overline{RCMA_i}$ by $*\left(\frac{12.00}{15.00}\right)$ so that they do sum to 12.

These numbers so constructed are the seasonal adjustment factors.

Step 4: We now construct the seasonally adjusted series, I_t, as follows:

$$\boxed{S.A.V.} = \frac{V_t}{\overline{RCMA}_t} \qquad t = \text{Jan, Feb, } \cdots \text{, Dec.;}$$

I_t will now display movements in visits to parks, adjusted for seasonality!
Shows movements in Visits over time (in each month) relative to the # of visits that are expected during that month of the year.

e.g. Suppose seasonal adj. factors: $\overline{RCMA}_{Jan} = .5$, $\overline{RCMA}_{Je} = 1.2$

$V_{Jan} = 50\,000$ visits;
$V_{Je} = 120\,000$ visits;

Seasonally Adjusted;

$$I_{Jan} = \frac{50\,000}{.5} = 10{,}000;$$

$$I_{Je} = \frac{120000}{1.2} = 100\,000$$

This, suggests that January and June experienced the same number of monthly visits, after adjusting for seasonality, i.e. wrt the average number of monthly visits expected in those months.

Alternatively, suppose $V_{Jan} = 60\,000$;
$V_{Je} = 100\,000$;

Then; $I_{Jan} = \frac{60\,000}{.5} = 120\,000$;

$$I_{Je} = \frac{100\,000}{1.2} = 85\,000$$

This suggests that relative to seasonal expectations, June is getting fewer visits than January. [even though the total # of visits is 40000 more !!]

Seasonal Adjustment
ii) by Dummy Regression

Case 1: $V = a_0 + a_1 Y + b_1 S_1 + b_2 S_2 + \cdots + b_{12} S_{12}$

where $S_i = 1$ in month i ; 0 elsewhere

PROBLEM: ∃ perfect multicollinearity here!
a_0 Perfectly correlated with $\sum S_i$.
If ∃ one solution, ∃ an infinite # of solutions!
" " " "
)
I can add any number, k, to $b_1, b_2, \ldots,$ & b_{12},
and subtract k from a_0,
and I'll get the same SSE!

Can solve problem by imposing some constraint.
Which constraint to impose?
$b_{12} = 0$, or $b_i = 0$ $i = 1, \ldots, 11$ [or $a_0 = 0$]
— leave one out.

- impose the constraint easiest to use.
- ⇒ It doesn't matter; results all the same.

BUT be careful; since you must explain the results to someone else!

Story; You do a study with dummies
for the 50 states,
leaving out Texas.
You present the results to
a Senate Subcommittee.
The distinguished senator from Texas says,
"What do you mean son - by leaving out (the great state of) Texas!"

You respond, "I actually didn't leave Texas out,
but rather, I made Texas the basis
for the entire study!"

Then the senator from Massachusetts says,
"What do you mean, sir!"

It's no use trying to explain that it doesn't matter.
All they see is a 1 or a 0.

~~[illegible]~~

Or: Use a sex dummy,
and they ask, "Why is Male = 1 & Female = 0 ?!"

It's no use trying to explain it doesn't matter.

* Can ~~solve~~ avoid this problem by simply imposing
a different constraint. [any relevant constraint will eliminate multicoll.]
e.g. that the sum of all the coefficients of
the dummies must be 0

$$V = a_0 + a_1 Y + b_1 S_1 + b_2 S_2 + \cdots + b_{12} S_{12}$$
$$\rightarrow \text{ s.t. } \Sigma b_i = 0$$

Anyone can understand this constraint!
No problems explaining.

Easiest procedure to impose this constraint:
Run regression with $b_{12} = 0$;
get $\hat{a}_0, \hat{a}_1, \hat{b}_1, \hat{b}_2, \ldots \hat{b}_{11}$, and $\hat{b}_{12} = 0$

Can manipulate seasonal dummies;
convert them into deviations from the mean seasonal coeff.

$\bar{\hat{b}}_\bullet$ = Mean Seasonal Coeff. = $\dfrac{\sum_{i=1}^{11} \hat{b}_i + 0}{12}$

Then let $\tilde{b}_1 = \hat{b}_1 - \bar{\hat{b}}_\bullet$, $\tilde{b}_2 = \hat{b}_2 - \bar{\hat{b}}_\bullet$, $\ldots$, $\tilde{b}_{12} = 0 - \bar{\hat{b}}_\bullet$;

so that $\sum_{i=1}^{12} \tilde{b}_i = 0$.

$\Rightarrow$ Same effective results;
easy to interpret. (For anyone! even a senator!)

DO THIS!!

Converting Results to seasonal coefficients:

[for 2nd regression – seasonality with constant slope]

$$\ln \frac{V}{N} = a_0 + a_1 \ln \frac{Y}{CPI*N} + b_1 S_1 + b_2 S_2 + \cdots + b_{12} S_{12}$$

$$\text{s.t. } \sum_{i=1}^{12} b_i = 0$$

take antilog!

$$\Rightarrow \frac{V}{N} = a_0 \left[\frac{Y}{CPI*N}\right]^{a_1} \left\{ e^{b_1 S_1} e^{b_2 S_2} e^{b_3 S_3} \cdots e^{b_{12} S_{12}} \right\}$$

Then in Jan., $b_1 S_1 = b_1 1$, and $S_2 = S_3 = \cdots = 0$;

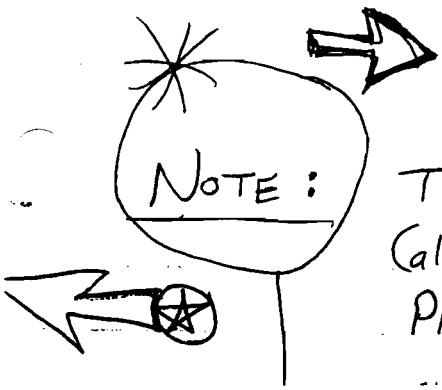
and $\frac{V}{N} = a_0 \left[\frac{Y}{CPI*N}\right]^{a_1} e^{b_1}$. → gets you estimate of actual $\frac{V}{N}$

Interpretation; $\hat{a}_0 \left[\frac{Y}{CPI*N}\right]^{\hat{a}_1}$ is the ("fitted") average # of visits, over all observations (regardless of seasonality, or holding constant for seasonality) Seasonally adjusted!!

and $\left[e^{\hat{b}_1}\right]$ is the seasonal coefficient for Jan.

∴ Seasonal coeff.'s are $(e^{\hat{b}_i})$, $i = 1, \ldots, 12$

$\hat{b}_i$ must be constrained to sum to zero here!!



These seasonal coefficients should be (almost) identical to those derived by hand, ($\overline{RCMA_i}$)

Plot each set of seasonal adjustment factors

B. Binary Dependent Variables

The lhs variable is a Dummy; attempting to explain a qualitative variable with a regression equation.

Models 1.-3. go from poor models to good models.

1. Linear Model (poor)

$$y_t = \underset{T\times K}{X_t} \underset{K\times 1}{\beta} + \epsilon_t \; ; \quad y_t = \begin{cases} 1 & \text{if event occurs} \\ 0 & \text{if not} \end{cases}$$

Interpretation: This model implies that

$$X_t\beta = E(y_t) = P(y_t = 1)$$
$$= \text{probability that the event occurs}$$

[Note: $E(y_t) = p(y_t=1)(1) + p(y_t=0)(0)$.]

Example: Might want to attempt to measure the factors affecting an individual's decision to work.

$$y_t = \begin{cases} 1 & \text{if indiv. supplies labor} \\ 0 & \text{if not} \end{cases}$$

We wish to measure β, the impact of the variables involved (X_t) on the individual's decision to wk

With the Linear Model specification,

$$\beta = \frac{\partial E(y_t)}{\partial X_t} = \frac{\partial P(y_t = 1)}{\partial X_t}$$

Problem 1: ϵ_t is strange; it has just 2 values.

We have $y_t = P_t + \epsilon_t$; where $P_t = P(y_t = 1) = X_t\beta$

Thus, since y_t has a binary distribution, ϵ_t has a binary distribution.

ϵ_t:	*	**		OR:		
possible values:	$1 - P_t$	$-P_t$			$1 - X_t\beta$	$-X_t\beta$
probability:	P_t	$1 - P_t$			$X_t\beta$	$1 - X_t\beta$

* If $P(y_t = 1) = X_t\beta = P_t$,
then $y_t = 1 = X_t\beta + \epsilon_t$
$= P_t + (1 - P_t)$.

** Analagously,
if $P(y_t = 0) = 1 - P_t$,
then $y_t = 0 = X_t\beta + \epsilon_t$
$= P_t + (-P_t)$.

Consider the first two moments of ϵ_t.

$$E(\epsilon_t) = (1 - X_t\beta)(X_t\beta) + (-X_t\beta)(1 - X_t\beta)$$
$$= 0$$

$$\begin{aligned} Var(\epsilon_t) &= E(\epsilon_t^2) \\ &= (1-P_t)^2(P_t) + (-P_t)^2(1-P_t) \\ &= (1-X_t\beta)^2(X_t\beta) + (-X_t\beta)^2(1-X_t\beta) \\ &\quad \text{factoring} \\ &= (1-X_t\beta)\left[(1-X_t\beta)(X_t\beta) + (-X_t\beta)^2\right] \\ &= (1-X_t\beta)\,X_t\beta \end{aligned}$$

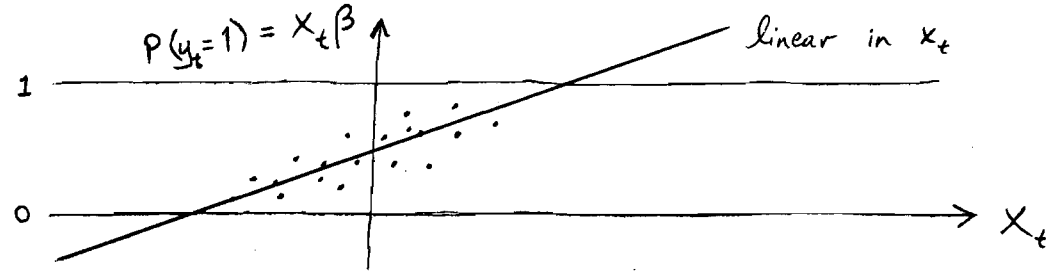
Observe that $Var(\epsilon_t)$ depends on X_t; it is not constant; ∄ <u>heteroskedasticity</u>.

This is not a very serious problem; <u>can</u> be corrected with GLS-type transformation of the model.

<u>Problem 2</u>: (big problem)

What if $X_t\beta < 0$ or $X_t\beta > 1$?

Can't be! It is a probability!



Since $y_t = X_t\beta + \epsilon_t$ is linear in X_t, for <u>some</u> X's $X_t\beta$ may imply probabilities outside the range, $(0, 1)$.

You may run this and get implied probabilities within the range (0, 1). Then you're OK.

But you may get some outliers that are uninterprettable!

Some people do this and just wave their hands and say, "it is obvious that this implies a probability of 1 (or of 0)."
It is not obvious!
This is poor procedure.

[Linear Model is poor, for this reason.]

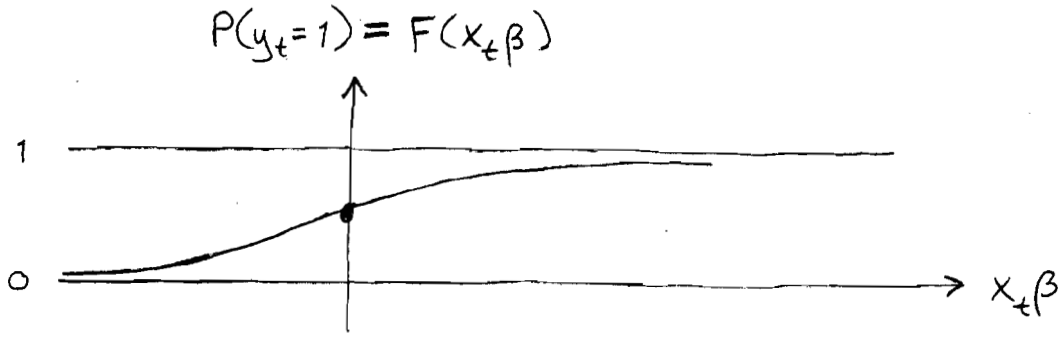
2. Probit Model

How can this second problem be dealt with?

We are attempting to fit a curve to the data, where the data represent the relationship between a probability (that must range between 0 and 1) and factors that influence it, X_t.

It is not a linear relationship that we are trying to fit!

It is actually something like
a cumulative distribution function!



The Probit Model simply replaces the linear specification with a nonlinear cumulative distribution function.

Which d.f. to use?

Heroic Assumption:

The Probit model uses as F, the Standard Normal c. d. f.
[$N(0,1)$; Expected Value = 0, Variance = 1]

The Model:

$$y_t^* = x_t\beta + \epsilon_t \qquad \text{with the } \underline{\epsilon_t \text{ iid } N(0,1)}$$

where y_t^* is unobservable, but y_t is observed;

$$y_t = \begin{cases} 1 & \text{if } y_t^* > 0 \\ 0 & \text{if } y_t^* \le 0. \end{cases}$$

Then $E(y_t) = P(y_t=1) = P(y_t^* > 0) = P(x_t\beta + \epsilon_t > 0)$

$= P(\epsilon_t > -x_t\beta)$

(by symmetry of Standard Normal Dist.) $= P(\epsilon_t < x_t\beta)$

$= F(x_t\beta)$ ✓

Effect: We generate a probability that is nice.

Comments:

1) Can add a constant term, with no change in the implications;

$$y_t = \begin{cases} 1 & \text{if } y_t^* = \alpha + \beta x_t + \epsilon_t > 0 \\ 0 & \text{if } y_t^* = \alpha + \beta x_t + \epsilon_t \leq 0 \end{cases}$$

2) We are estimating β; the impact of x_t on the probability of the event occurring, $P(y_t = 1)$.
— more explanation below.

3) The Standard Normal c.d.f. is used for good reason. Consider an alternative assumption that is more general.
Let ϵ_t be iid $N(0, \sigma^2)$.

$$\Rightarrow E(y_t) = P(y_t = 1) = P(y_t^* > 0) = P(x_t\beta + \epsilon_t > 0)$$
$$= P(\epsilon_t > -x_t\beta)$$
$$= P(\epsilon_t < x_t\beta)$$
$$= P(\epsilon_t/\sigma < x_t \beta/\sigma)$$
$$= F(x_t \tfrac{\beta}{\sigma})$$

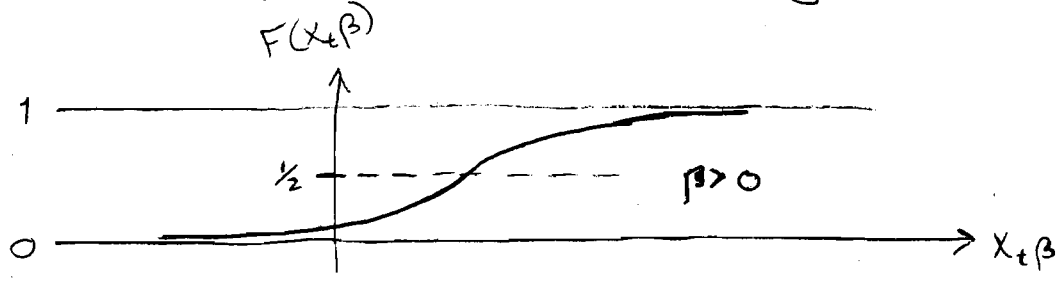
($\div \sigma$)

(Must standardize to obtain the cdf, F) → ↖ for $N(0,1)$

Here we can estimate $\frac{\beta}{\sigma}$, but we cannot obtain β & σ.
∃ an identification problem.

$\sigma^2 = 1$ is a "normalization" assumption.

Comments, cont. 4) A drawback (for some models) is that there is symmetry above and below the probability of ½, due to the assumption of Normality.



Estimation of Probit Models: by MLE

Here, $E(y_t) = P(y_t = 1) = P_t = F(X_t\beta)$

Binary Dist.

and $P(y_t = 0) = 1 - P_t = 1 - F(X_t\beta)$

$$\mathcal{L} = P(y_1)\, P(y_2)\, P(y_3) \cdots P(y_T)$$

$$= \prod_{t \to y_t = 1} P_t \prod_{t \to y_t = 0} (1 - P_t)$$

$$= \prod_{t=1}^{T} P_t^{y_t} (1 - P_t)^{(1 - y_t)}$$

These powers will mechanically include a factor, P_t, if $y_t = 1$, and $(1 - P_t)$ if $y_t = 0$.

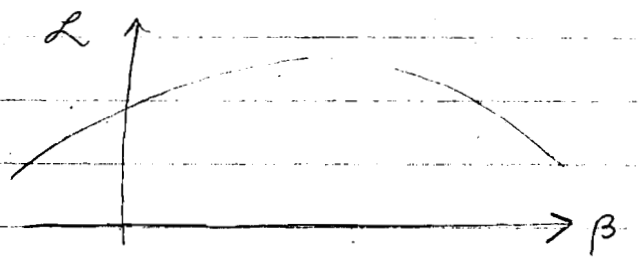
$$= \prod_{t=1}^{T} [F(X_t\beta)]^{y_t} [1 - F(X_t\beta)]^{(1 - y_t)}$$

Thus, $\log \mathcal{L} = \sum_{t=1}^{T} \left\{ y_t \log\left[F(X_t\beta)\right] + (1-y_t) \log\left[1 - F(X_t\beta)\right] \right\}$

Find the values of the parameters, β, that maximize this; MLE ($\hat{\beta}$).

Note: $\log \mathcal{L}$ has a global maximum with the Probit model specification.

Thus, MLE ($\hat{\beta}$) is "easy" to find with any iterative technique.



– find the top of a hill.

What β means:

$$\frac{\partial P(y_t=1)}{\partial X_{ti}} = \frac{\partial F(X_t\beta)}{\partial X_{ti}} = f(X_t\beta) \frac{\partial X_t\beta}{\partial X_{ti}} = f(X_t\beta)\,\beta_i$$

where f is the Standard Normal density fn.

Observe that X_{ti} affects $P(y_t=1)$ by an amount, $f(X_t\beta)\,\beta_i$.

→ If $\beta_i > 0$, then if X_{ti} ↑es, $P(y_t=1)$ ↑es.

Attended meetings;

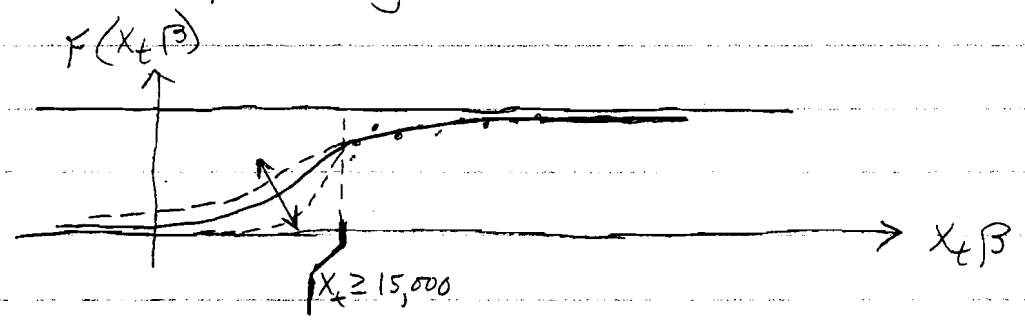
Paper presented, studying the likelihood that loan applicants would default on loans;
This probability of default might be influenced by factors such as the applicants' income, age, education, ...

Had data from Credit Unions on people with Incomes $\geq$ 15,000; data on Income, age, education, ..., and whether or not they defaulted.

Fit probit model; got interesting results, claimed ability to "predict" the likelihood that an applicant would default, based on these factors.
— Potentially Valuable for Credit Unions, and all financial institutions!

De Min Wu, U. of Kansas, discussant;
argued that sample was not random; truncated at income levels $\geq$ 15,000.

Thus, the fitted curve is only one possibility; many other curves might be consistent with the given data. Thus, results ($\hat{\beta}$) not necessarily reliable for predicting this likelihood.



3. Logit Model

$$\log\left[\frac{P(y_t=1)}{P(y_t=0)}\right] = X_t\beta$$

X_t affects the (log of) relative probabilities of events, $y_t=1$ and $y_t=0$, by the amount, β.

Note: $\exists$ no error term here.

This is a description of the probability of an event occurring, and this probability is deterministic, as the model is specified.

It is then the actual occurrance or non-occurrance of the event which is random.

Observe: If $P(y_t=1) = P_t$, then Logit is

$$\log\left(\frac{P_t}{1-P_t}\right) = X_t\beta$$

$$\Longrightarrow \quad \frac{P_t}{1-P_t} = e^{X_t\beta}$$

$$\Longrightarrow \quad P_t = (1-P_t)\,e^{X_t\beta}$$

$$\Longrightarrow \quad P_t + P_t e^{X_t\beta} = e^{X_t\beta}$$

$P(y_t=1)$ $$\Longrightarrow \quad P_t = \frac{e^{X_t\beta}}{1+e^{X_t\beta}} = \frac{1}{e^{-X_t\beta}(1+e^{X_t\beta})} = \frac{1}{1+e^{-X_t\beta}}$$

$$= G(X_t\beta)$$

[cdf for the Logistic Distribution]

Probit uses the Standard Normal cdf; Logit uses the Logistic cdf.

Note: $G(x) = \frac{1}{1+e^{-x}}$ is the Logistic Dist.

SKIP

Note: $P(y_t = 0) = 1 - P(y_t = 1)$

$$= 1 - \frac{e^{x_t\beta}}{1+e^{x_t\beta}} \quad \Big\| \quad = 1 - \frac{1}{1+e^{-x_t\beta}}$$

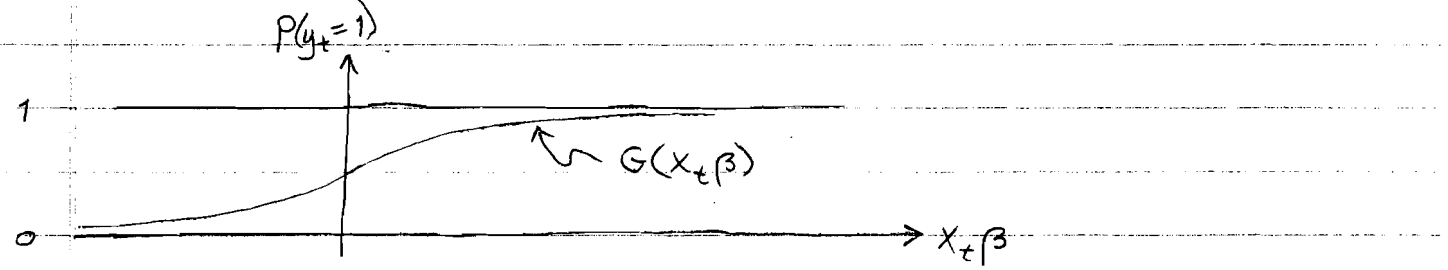
$$= \frac{1}{1+e^{x_t\beta}} \quad \Big\| \quad = \frac{e^{-x_t\beta}}{1+e^{-x_t\beta}}$$

Advantage of Logit:

Although $P(y_t = 1)$ and $[1 - P(y_t = 1)]$ must be in the interval $(0, 1)$,

$\left[\frac{P(y_t=1)}{P(y_t=0)}\right]$ may range from $[0$ to $\infty]$,

and thus, $\log\left[\frac{P(y_t=1)}{P(y_t=0)}\right] = x_t\beta$ [what we're trying to explain w/ x_t] may range from $[-\infty$ to $\infty]$.



The Logistic Dist., G, is much like the Standard Normal cdf used in Probit. The difference amounts to slightly thicker tails in one of the two.

Estimation of Logit: by MLE

$$\mathcal{L} = \prod_{t=1}^{T} G(X_t\beta)^{y_t} [1 - G(X_t\beta)]^{(1-y_t)}$$

$$= \prod_{t \ni y_t=1} (P_t) \prod_{t \ni y_t=0} (1-P_t)$$

$$= \prod_{t \ni y_t=1} \left(\frac{e^{X_t\beta}}{1+e^{X_t\beta}}\right) \prod_{t \ni y_t=0} \left(\frac{1}{1+e^{X_t\beta}}\right)$$

- Same as Binary Probit, with $G(X_t\beta)$ replacing $F(X_t\beta)$. Like Probit, this $\mathcal{L}$ has a global maximum.

So max $\mathcal{L}$ wrt $\beta \rightarrow \hat{\beta}_{MLE}$.

Thus ends discussion of Binary Dependent variables.

- Use Probit or Logit!
- Slightly more expensive than least squares, but much better. ↑ [on Linear Model]

SAS - PROC PROBIT
PROC FUNCAT (?)

PROC LOGIST

C. Qualitative Dependent Variables with more than 2 values.

$$y = f(x's)$$

$$\text{where } y = \begin{cases} 1 & \text{if in classific. I} \\ 2 & \text{if in classific. II} \\ \vdots & \\ k & \text{if in classific. k} \end{cases}$$

e.g. attempting to explain how various factors might influence how individuals are classified into: income groups
education levels
good products - faulty prod.
⋮

Any situation in which you hypothesize that certain factors are influential in discriminating among various classifications.

∃ many possible models to choose from

1. Multinomial Logit
2. Conditional Logit
3. Multinomial Probit (with ordered or unordered responses)
4. Discriminant Analysis

D. Limited Dependent Variables

This is situation in which lhs variable is drawn from or described by a truncated distribution.

1. Tobit Model (by Tobin)

The model; $y_t^* = \underbrace{x_t\beta}_{\text{usual rhs variables}} + \epsilon_t \qquad \epsilon_t \sim N(0,\sigma^2)$

y_t^* is the variable from the entire population of this particular distribution.

However, we don't observe y_t^*.
Instead, we have a sample that is drawn from a subsample or truncated portion of the distribution of y_t^*.

e.g. observe $y_t = \begin{cases} y_t^* & \text{if } y_t^* > 0 \\ 0 & \text{if } y_t^* \leq 0 \end{cases}$

Note: $E(y_t^*) = x_t\beta$; this is the mean of the dist. of y_t^*, but not of the truncated distribution of y_t !

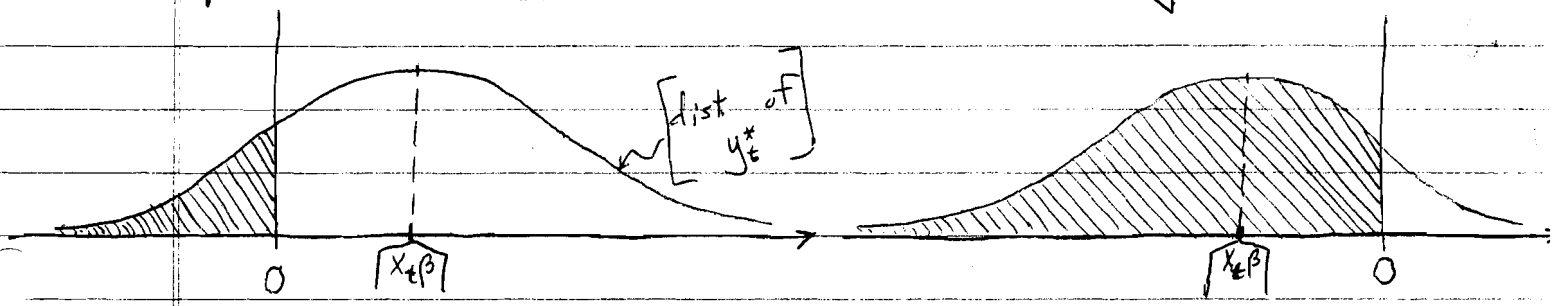
Tobin's example:

The amount of expenditures on cars in one year, by a household.

This is bounded, or truncated at zero.

$$\begin{bmatrix}\text{Car exp.}\\ \text{in 1 yr.}\end{bmatrix} = \begin{cases} 0 & \text{for 70\% of population} \\ 300-50{,}000 & \text{for 30\% of population} \end{cases}$$

2 examples:



Over a 5-yr. pd, most people buy a car

Over a 1-yr pd, most do not buy a car (70% in above ex.)

KEY

The Difference between Tobit & Binary Probit (or Logit):

Probit
$$\text{observed } y_t = \begin{cases} 1 & \text{if } y_t^* > 0 \\ 0 & \text{if } y_t^* \le 0 \end{cases} \quad \text{where } y_t^* = X_t\beta + \epsilon_t \text{ is unobserved}$$

Tobit
$$\text{observed } y_t = \begin{cases} y_t^* & \text{if } y_t^* > 0 \\ 0 & \text{if } y_t^* \le 0 \end{cases}$$

Thus, Tobit is for a truly Limited dependent var.

Interpretation of β:

$$\beta = \frac{\partial E(y_t^*)}{\partial x_t} = \frac{\partial x_t \beta}{\partial x_t}$$

Thus, β is the impact on the mean of the entire (unlimited) population, y_t^*, from a change in x_t.

This interpretation is clouded when we observe y_t and not y_t^*.

Tobin shows how this kind of model can be estimated.

— Somewhat gory, but can be done.

E. Sample Selection Bias

In Tobit model, our sample consisted of all people in the entire population; but only some bought cars this year.

What if you didn't have a random sample, but instead, you only had a sample of people who bought cars (nonzero obs.).

— e.g. 1; only have car registration data.

[e.g. 2; financial data on loan applicants with $y \geq 15{,}000$].

Here, the sample is not random!
— sample selection bias exists.

This is not a Tobit model since we don't have data on people who do not buy cars. We only have data for the truncated distribution, y_t; not on the entire distribution, y_t^*.

☆ See Hausman & Weiss, ECON, 1977.
Heckman, ECON, 1979.

They show how to deal with this model.

— Get something like Max. Likelihood Estimates, only the density functions involved are conditional densities (conditional on $y_t^* > 0$, ...).

F. Simultaneous equation models in which one equation involves a Qualitative Dependent Var.

e.g. Relationship between Wage Rates and the Quit Rate(Q) (Union Membership (U_t))

—— Simultaneously determined. $W_t \leftrightarrow \overset{(U_t)}{Q_t}$.

Cross-sectional Survey data on individuals'
Wage Rate, Income, Education Level, Age, ...,
and whether or not they recently "Quit."
(belong to a Union)

System:

$$W_t = f(Y, E, A, \overset{(U_t)}{Q}) \qquad Q \text{ (or } U_t) \rightarrow W_t$$

$$\overset{(U_t)}{Q_t} = f(W, Y, E, A) \qquad W_t \rightarrow Q \text{ (or } U_t)$$

where $\overset{(U_t)}{Q_t} = \begin{cases} 1 & \text{if } \overset{\text{(union member)}}{\text{quit recently}} \\ 0 & \text{otherwise} \end{cases}$

Can estimate this kind of model.

1. Simultaneous Probit Model
2. Simultaneous Logit Model
3. Simultaneous Tobit Model

⋮