APS 425 – Fall 2015

Functional Form and Qualitative Variables

Instructor: G. William Schwert
275-2470
schwert@schwert.ssb.rochester.edu

Topics

- Transformations to linearity
- Dummy variables
- Interaction variables

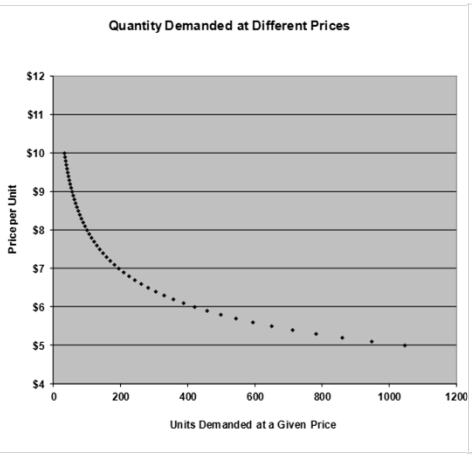
Transformations to Linearity

- Linear regression assumes that Y is linearly related to X_1, X_2, etc. with an additive error term
- Scatter diagrams can help us understand whether the variables we are considering need to be transformed to create a linear relation
  - e.g., if we take logarithms of the variables (assuming that they are all positive numbers), this implies a constant elasticity relation between Y and X

$$\log Y = \beta_0 + \beta_1 \log X + \varepsilon \Rightarrow Y = \beta_0 X^{\beta_1} \exp(\varepsilon)$$

where *log* is the natural logarithm and $\exp(\bullet)$ is the exponential function

Transformations to Linearity



Logs as Percent Changes

- Suppose that log(Y) is linearly related to some variable X

  $log\ Y = \beta_0 + \beta_1 X + \varepsilon$

  Then the slope coefficient β_1 measures the percent change in Y caused by a change in X

  Fact: $\log (1 + r) \approx r$ for $| r | < .15$

  To the extent that the distribution of % changes is more likely to have a constant variance than the distribution of changes, the log transformation can help solve heteroskedasticity problems

Logs as Percent Changes

Excel function ln(•) is the natural logarithm

Logs as approximate percentage

r	ln(1+r)	r	ln(1+r)
-0.150	-0.163	0.150	0.140
-0.140	-0.151	0.140	0.131
-0.130	-0.139	0.130	0.122
-0.120	-0.128	0.120	0.113
-0.110	-0.117	0.110	0.104
-0.100	-0.105	0.100	0.095
-0.090	-0.094	0.090	0.086
-0.080	-0.083	0.080	0.077
-0.070	-0.073	0.070	0.068
-0.060	-0.062	0.060	0.058
-0.050	-0.051	0.050	0.049
-0.040	-0.041	0.040	0.039
-0.030	-0.030	0.030	0.030
-0.020	-0.020	0.020	0.020
-0.010	-0.010	0.010	0.010
0.000	0.000		

Example: Shipping Boxes

- Imagine that you were in charge of a shipping department for an on-line retailing firm. The product you ship via UPS is malleable so that it can fit into virtually any shape of box (think popcorn)
- You want to have a plan for keeping an inventory of boxes on hand that can handle most of your shipping needs (volumes)
- What kinds of factors are likely to affect the volume of boxes?
  - Height (H), Width (W), and Length(L)

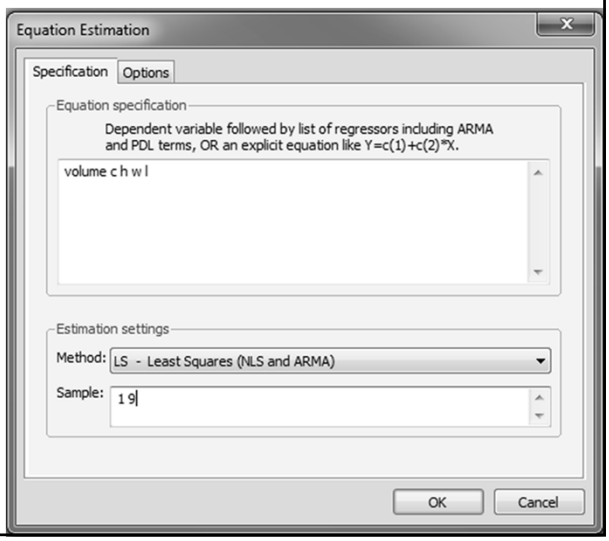
Example: Shipping Boxes

Suppose you have 9 different boxes you currently use with the following dimensions and volumes

H	W	L	Volume
4	6	6	144
4	8	4	128
4	12	8	384
6	6	6	216
6	8	4	192
6	12	8	576
8	6	6	288
8	8	4	256
8	12	8	768

Eviews Regression

As a first step, regress the variable you are interested in modeling/forecasting, VOLUME, on the explanatory variables, H, W, and L



Eviews Regression

It looks like we can create a pretty good regression with these data, explaining over 91% of the variation in volume

Dependent Variable: VOLUME
Method: Least Squares
Sample: 1 9
Included observations: 9

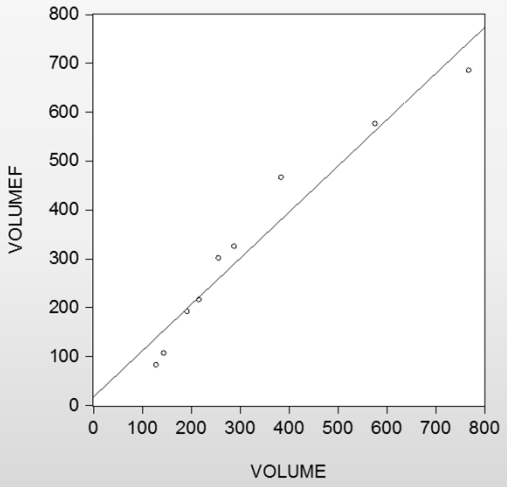
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-688.0000	117.0907	-5.875789	0.0020
H	54.66667	13.09113	4.175854	0.0087
W	42.00000	11.33725	3.704601	0.0139
L	54.00000	17.31794	3.118153	0.0263

R-squared	0.944406	Mean dependent var	328.0000
Adjusted R-squared	0.911050	S.D. dependent var	215.0349
S.E. of regression	64.13319	Akaike info criterion	11.46090
Sum squared resid	20565.33	Schwarz criterion	11.54856
Log likelihood	-47.57407	Hannan-Quinn criter.	11.27174
F-statistic	28.31259	Durbin-Watson stat	1.999654
Prob(F-statistic)	0.001455		

Eviews Regression

What's wrong with this regression?

- •The constant term implies large significant negative volume if you have a box where H=L=W=0!
- •Plot of predicted volume (volumef) versus volume shows a curved relation: very small and very large volume boxes are under-predicted by the model



Eviews Regression: Logs

Of course, if we estimate the model in logs, we get a perfect fit, since:

Volume = H*W*L

Dependent Variable: LOG(VOLUME)
Method: Least Squares
Sample: 1 9
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.91E-14	3.21E-14	-1.217305	0.2778
LOG(H)	1.000000	1.12E-14	8.96E+13	0.0000
LOG(W)	1.000000	1.29E-14	7.76E+13	0.0000
LOG(L)	1.000000	1.29E-14	7.76E+13	0.0000

R-squared	1.000000	Mean dependent var	5.623699
Adjusted R-squared	1.000000	S.D. dependent var	0.603165
S.E. of regression	9.52E-15	Sum squared resid	4.53E-28
F-statistic	1.07E+28	Durbin-Watson stat	2.175433
Prob(F-statistic)	0.000000		

Message: It Is Important to *Think* Before Computing!

In this case, recollecting a simple formula (from 7th grade) where the volume of a rectangular prism (aka "box") is just the height times the width times the length:

Volume = H x W x L

would give a perfect solution to the problem.

Shipping boxes example:

http://schwert.ssb.rochester.edu/a425/a425_logs.wf1

Nonlinear Dependent Variables:

Predictions in the multiplicative model

- Multiplicative model: $Y = \beta_0 X^{\beta_1} exp(\varepsilon)$
- Linearized model: $log\ Y = log\ \beta_0 + \beta_1\ log\ X + \varepsilon$
  or, $Y^* = \beta_0^* + \beta_1 X^* + \varepsilon$
- Predictors for linearized model:
  - b_0^* is the OLS estimator for β_0^*
  - b_1 is the OLS estimator for β_1

Nonlinear Dependent Variables:

Predictions in the multiplicative model

- Predictor for multiplicative model:
  - $b_0 = \exp(b_0^*)$ is the estimator for β_0
  - Note that b_0 is a biased estimator
- Similarly,
  - $\log \hat{Y} = b_0^* + b_1 \log X$
  - is an unbiased estimator of $\log(Y)$
- But:
  - $\exp(\log \hat{Y})$ is a biased estimator of Y
    [Jensen's Inequality]

Qualitative Variables

- Sometimes you can observe only qualitative or categorical information that you may want to use
  - Gender of customers
  - College graduate?
  - Resident of a particular region, city, state?

Qualitative Variables

- Define a variable = 1 if female, and 0 if male
  - Coefficient of this variable in the regression measures the difference in the mean for females versus males
  - So, the mean for females is the intercept plus the coefficient of "FEMALE"
  - The mean for males is the intercept

Ciba-Geigy Ritalin Experiment

- Ritalin is tested to see if it helps with Central Auditory Processing Disorder (CAPD)
  - Similar symptoms to ADD/ADHD
- Experiment:
  - "Randomly" select 64 children
  - All receive auditory test
  - 32 (control group) receive no drug (or placebo?)
  - 32 (treatment group) receive varying doses of Ritalin
  - All children are tested a second time

Ciba-Geigy Ritalin Experiment

- $DOSAGE_i$ = amount of Ritalin received by child i
  - Measured as Mg of Ritalin per Kg of body weight
- $IMPROVE_i$ = child's 2^{nd} test score – 1^{st} test score
  - Dataset A425_ritalin.wf1 also contains:
    - AGE of child in months
    - Gender (FEMALE = 1, for girls)

Examples of Dummy Variables: Ritalin Case

- DRUGDUM = 1 if taking Ritalin; 0, otherwise
- FEMALE = 1, when female; 0, otherwise

Ritalin example:

http://schwert.ssb.rochester.edu/a425/a425_ritalin.wf1

Interpreting Coefficients of Dummy Variables

- Females average 4.5 points less improvement than males
- Kids who receive ritalin improve by 6.3 points more than kids who don't

Dependent Variable: IMPROVE
Method: Least Squares
Sample: 1 64
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.547147	9.406639	-1.014937	0.3142
AGE	0.103014	0.089944	1.145317	0.2566
FEMALE	-4.472018	3.433368	-1.302516	0.1977
DRUGDUM	6.285486	3.003482	2.092733	0.0406

R-squared	0.112850	Mean dependent var	3.359375
Adjusted R-squared	0.068492	S.D. dependent var	12.28157
S.E. of regression	11.85352	Akaike info criterion	7.843588
Sum squared resid	8430.353	Schwarz criterion	7.978518
Log likelihood	-246.9948	Hannan-Quinn criter.	7.896744
F-statistic	2.544096	Durbin-Watson stat	1.885493
Prob(F-statistic)	0.064534		

Qualitative Interaction Variables

- Multiply a continuous variable by a dummy variable to allow for differences in slope coefficients
- FEMALE*DOSAGE in the ritalin example allows for the effect of a mg/kg of body weight of ritalin to be different for females than males
- Coefficient of this variable in the regression measures the difference in the slope for females versus males
  - So, the effect of a dose of ritalin for females is the coefficient of DOSAGE plus the coefficient of "FEMALE*DOSAGE"
  - effect of a dose of ritalin for males is the coefficient of DOSAGE

Original Ritalin Regression

- original regression shows 12.2 points of IMPROVE for each mg/kg of ritalin

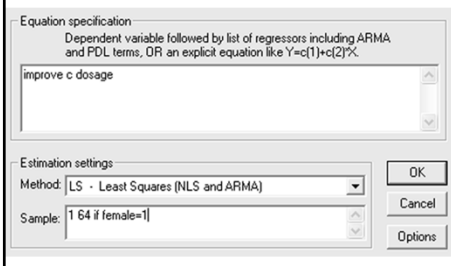
Dependent Variable: IMPROVE
Method: Least Squares
Sample: 1 64
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.225872	2.097651	0.107678	0.9146
DOSAGE	12.17855	5.723027	2.127991	0.0373

R-squared	0.068066	Mean dependent var	3.359375
Adjusted R-squared	0.053035	S.D. dependent var	12.28157
S.E. of regression	11.95146	Akaike info criterion	7.830335
Sum squared resid	8855.917	Schwarz criterion	7.897800
Log likelihood	-248.5707	Hannan-Quinn criter.	7.856913
F-statistic	4.528347	Durbin-Watson stat	1.975687
Prob(F-statistic)	0.037321		

Ritalin Regression for Females

- For the females, there is a 0.6 point increase in test score for each mg/kg of ritalin
  - Lower than the base result and not significantly different from 0
  - Note that there are only 17 females in the sample
  - Prob(17 or less females out of 64) = .0001! => not a random sample



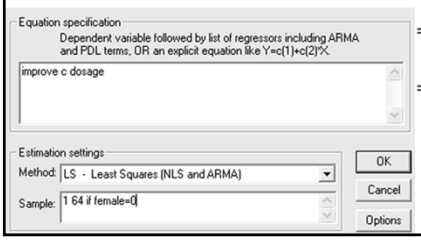
Dependent Variable: IMPROVE
Method: Least Squares
Sample: 1 64 IF FEMALE=1
Included observations: 17

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.717771	2.429588	0.295429	0.7717
DOSAGE	0.604690	6.482739	0.093277	0.9269

R-squared	0.000580	Mean dependent var	0.882353
Adjusted R-squared	-0.066048	S.D. dependent var	6.669730
S.E. of regression	6.886470	Akaike info criterion	6.807125
Sum squared resid	711.3521	Schwarz criterion	6.905150
Log likelihood	-55.86057	Hannan-Quinn criter.	6.816869
F-statistic	0.008701	Durbin-Watson stat	1.197790
Prob(F-statistic)	0.926918		

Ritalin Regression for Males

- For the males, there is a 16.5 increase in test score for each mg/kg of ritalin
  - Thus, it seems that the benefits of ritalin are limited to boys



Dependent Variable: IMPROVE
Method: Least Squares
Sample: 1 64 IF FEMALE=0
Included observations: 47

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.096493	2.658614	0.036295	0.9712
DOSAGE	16.50885	7.313969	2.257168	0.0289

R-squared	0.101703	Mean dependent var	4.255319
Adjusted R-squared	0.081741	S.D. dependent var	13.71205
S.E. of regression	13.13969	Akaike info criterion	8.030773
Sum squared resid	7769.311	Schwarz criterion	8.109502
Log likelihood	-186.7232	Hannan-Quinn criter.	8.060399
F-statistic	5.094805	Durbin-Watson stat	1.829841
Prob(F-statistic)	0.028899		

Interpreting Coefficients of Interactive Dummy Variables

- Note the constant, C, and the coefficient of DOSAGE are the same as for the males only regression
- Note the constant plus the coefficient of FEMALE equals the constant for the females only regression
- Note the sum of the coefficients of DOSAGE and FEMALEDOSAGE equals the coefficient of DOSAGE for the females only regression

Dependent Variable: IMPROVE
Method: Least Squares
Sample: 1 64
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.096493	2.405524	0.040113	0.9681
FEMALE	0.621277	4.835284	0.128488	0.8982
DOSAGE	16.50885	6.617705	2.494649	0.0154
FEMALEDOSAGE	-15.90416	13.00196	-1.223212	0.2260

R-squared	0.107555	Mean dependent var	3.359375
Adjusted R-squared	0.062933	S.D. dependent var	12.28157
S.E. of regression	11.88883	Akaike info criterion	7.849538
Sum squared resid	8480.663	Schwarz criterion	7.984468
Log likelihood	-247.1852	Hannan-Quinn criter.	7.902694
F-statistic	2.410357	Durbin-Watson stat	1.889101
Prob(F-statistic)	0.075708		

Links

Return to APS 425 Home Page:

http://schwert.ssb.rochester.edu/a425/a425main.htm