Multiple Regression Analysis

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

4. Further Issues

Ch.6 Multiple Regression: Further Issues

1. Effects of Data Scaling on OLS Statistics
2. More on Functional Form
3. More on Goodness-of-Fit & Selection of Regressors
4. Prediction & Residual Analysis*

6.1 Effects of Data Scaling

Redefining Variables

Changing the scale of the variable will lead to a corresponding change in the scale of the coefficients and standard errors, but ***no change in the significance or interpretation***.

→ *see* table 6.1.

TABLE 6.1

Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Independent Variables		= *bwght*/16	
cigs	−.4634 (.0916)	−.0289 (.0057)	—
packs	—	—	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	.0298	.0298	.0298
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

1/16

TABLE 6.1

Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Independent Variables		= *bwght*/16	
cigs	−.4634 (.0916)	−.0289 (.0057)	—
packs = *cigs*/20	—	—	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	.0298	.0298	.0298
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

×20

Cont. Redefining Variables

- If x_j is multiplied by c, its coefficient is divided by c.
- If y is multiplied by c, ***all*** OLS coefficients is multiplied by c.
- Neither t nor F statistics are affected by changing the units of measurement of any variables.
- If the variables appears in logarithmic form, changing unit of measurement does not affect the slope coefficient.

Effects of Data Scaling

Dependent / Independent	y	cy	y
x_1	β_1 (se_1)	$c\beta_1$ ($c*se_1$)	——
dx_1	——	——	β_1/d (se_1/d)
x_2	β_2 (se_2)	$c\beta_2$ ($c*se_2$)	β_2 (se_2)
Intercept	β_0 (se_0)	$c\beta_0$ ($c*se_0$)	β_0 (se_0)
R-squared	R^2	R^2	R^2
SSR	SSR	c^2*SSR	SSR

Standard errors in parentheses

Beta Coefficients

- Idea is to replace y and each x variable with a standardized version – subtract mean and divide by standard deviation.
- Coefficient reflects standard deviation of y for a one standard deviation change in x.
  - We can compare the magnitudes of the resulting beta coefficients and conclude that "which variable is most important," etc.
  - Whether we use standardized or unstandardized variables does not affect statistical significance.

6.2 More on Functional Form

- Functional Forms
  - OLS can be used for relationships that are not strictly linear in x and y by using nonlinear functions of x and y – will still be linear in the parameters.
    - the natural log of x, y or both
    - quadratic forms of x
    - interactions of x variables

Interpretation of Log Models

- 6-2a Logarithmic Functional Forms
  - $\ln(y) = \beta_0 + \beta_1 \ln(x) + u$
    - β_1 is the elasticity of y with respect to x
  - $\ln(y) = \beta_0 + \beta_1 x + u$
    - β_1 is approximately the percentage change in y given a 1 unit change in x
  - $y = \beta_0 + \beta_1 \ln(x) + u$
    - β_1 is approximately the change in y for a 100 percent change in x

Functional forms with Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\partial y/\partial x$
level-log	y	$\log(x)$	$\partial y/(\partial x/x)$
log-level	$\log(y)$	x	$(\partial y/y)/\partial x$
log-log	$\log(y)$	$\log(x)$	$(\partial y/y)/(\partial x/x)$

Why use log models?

- Log models are invariant to the scale of the variables since measuring percent changes.
- They give a direct estimate of elasticity.
- For models with $y > 0$, the conditional distribution is often heteroskedastic or skewed, while $\ln(y)$ is much less so.
- The distribution of $\ln(y)$ is more narrow, limiting the effect of outliers.

Some Rules of Thumb for "log"

- What types of variables are often used in log form?
  - Dollar amounts that must be positive
  - Very large variables, such as population
- What types of variables are often used in level form?
  - Variables measured in years
  - Variables that are a proportion or percent

Quadratic Models

- 6-2b Models with Quadratics
  - For a model of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u,$$

we can't interpret β_1 alone as measuring the change in y with respect to x, we need to take into account β_2 as well, since

$$\frac{\partial \hat{y}}{\partial x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$$

More on Quadratic Models

- For the case of the coefficient on $x > 0$ and the coefficient on $x^2 < 0$, y is increasing in x at first, but will eventually turn around and be decreasing in x (see fig.6.1).
- For the case of the coefficient on $x < 0$ and the coefficient on $x^2 > 0$, y is decreasing in x at first, but will eventually turn around and be increasing in x (see fig.6.2).
- For both case, the turning point will be at

$$x^* = \left|\hat{\beta}_1 / \left(2\hat{\beta}_2\right)\right|.$$

Interaction Terms

- 6-2c Models with Interaction Terms
  - For a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

we can't interpret β_1 alone as measuring the change in y with respect to x_1, we need to take into account β_3 as well, since

$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$$

- To summarize the effect of x_1 on y, we typically evaluate the above at average of x_2.

6.3 More of Goodness-of-Fit

<u>Adjusted *R*-Squared</u>

- Recall that the R^2 will always increase as more variables are added to the model.
- The adjusted R^2 takes into account the number of variables in a model, and may decrease.

$$\bar{R}^2 \equiv 1 - \frac{[SSR/(n-k-1)]}{[SST/(n-1)]} = 1 - (1 - R^2)\frac{n-1}{n-k-1} \quad (6.22)$$

Cont. Adjusted *R*-Squared

- Most PC packages will give you both R^2 and adj-R^2.
- You can compare the fit of 2 models (with the same y) by comparing the adj-R^2 .
- However, you cannot use the adj-R^2 to compare models with different y's.
  - *e.g.* y vs. $\ln(y)$.

R^2 & Selection of regressors

- It is important *not to fixate too much on adj-R^2 and lose sight of theory and common sense.*
  - If economic theory clearly predicts a variable belongs, generally leave it in the model.
  - Don't want to include a variable that prohibits a sensible interpretation of the variable of interest – remember ceteris paribus interpretation of multiple regression.

  e.g. housing price = f (# of rooms, square footage)

Econometrics 19

Selection of Regressors

- 6-3c Controlling for too many factors
  - Ex.1 Trafic fatalities & Beer taxes

Traffic fatalities
↑
Drunk driving
↑
Beer consumption
↑
State beer taxes

In a regression of traffic fatalities on state beer taxes (and other factors), one should not directly control for beer consumption.

fatalities = f (tax, ~~beercons~~, ...)

Econometrics 20

Cont. Selection of Regressors

- 6-3c Controlling for too many factors
  - Ex.2 Health expenditures & Pesticide usage

Health expenditures
↑
Doctor visits
↑
Pesticide usage

In a regression of family health expenditures on pesticide usage among farmers, one should not control for doctor visits.

hexpend = f (pestusage, ~~docvisit~~, ...)

Econometrics 21

Cont. Selection of Regressors

- 6-3c Controlling for too many factors
  - Ex.3 House prices & house characteristics

House prices
↑
Price assessments
↑
House characteristics
- Size of the lot
- Square footage
- # of bedrooms

In a regression of house prices on house characteristics, one would include price assessments if the purpose of the regression is to study their validity; otherwise one would not include them.

1. *hprice = f (bdrms, assess, ...)*
2. *hprice = f (bdrms, ~~assess~~, ...)*

Econometrics 22