AFRE 835: Introductory Econometrics

Chapter 6: Multiple Regression Analysis: Further Issues

Spring 2017

Introduction

- This chapter reviews a variety of topics related to multiple regression analysis, many of which have already been touched on in earlier chapters, including:
  - the role of data scaling;
  - interpretation of models that are nonlinear in the independent variables; and
  - the choice of variables to include in a model.

Outline

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

Units of Measurement

- As we saw in ch. 2, changes in the units for either the dependent or explanatory variables will impact the corresponding coefficients,
  ... but it will not impact the corresponding t- or F-statistics
- Suppose we use the dataset *smoke.dta* to model cigarette consumption (*cigs*) as a function of *age*, *income* and education (*educ*); i.e.,

$$\widehat{cigs}_i = \hat{\beta}_0 + \hat{\beta}_1 age_i + \hat{\beta}_2 income_i + \hat{\beta}_3 educ_i$$

Units of Measurement

- The following set of results emerge from Stata

```
. reg             cigs age income educ;

      Source |       SS           df       MS      Number of obs   =       807
-------------+----------------------------------   F(3, 803)       =      2.81
       Model |  1578.01597         3  526.005322   Prob > F        =    0.0384
    Residual |  150175.667       803  187.018265   R-squared       =    0.0104
-------------+----------------------------------   Adj R-squared   =    0.0067
       Total |  151753.683       806  188.280003   Root MSE        =    13.675

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0416932   .0287628    -1.45   0.148    -.0981524    .0147659
      income |   .0001171   .0000559     2.09   0.036     7.38e-06    .0002268
        educ |  -.3775954   .1696335    -2.23   0.026    -.7105728   -.044618
       _cons |   12.85394   2.576089     4.99   0.000      7.79728    17.91061
------------------------------------------------------------------------------
```


Units of Measurement - Dependent Variable

- If we instead measure cigarette consumption in packs (*packs*), the corresponding model would become

$$\widehat{packs}_i = \frac{\widehat{cigs}_i}{20} = \frac{\hat{\beta}_0 + \hat{\beta}_1 age_i + \hat{\beta}_2 income_i + \hat{\beta}_3 educ_i}{20}$$
$$= \frac{\hat{\beta}_0}{20} + \frac{\hat{\beta}_1}{20} age_i + \frac{\hat{\beta}_2}{20} income_i + \frac{\hat{\beta}_3}{20} educ_i \qquad (1)$$

- Effectively, all of the coefficients are shrunk by a factor of 20, but all the t- and F-statistics will remain unchanged.

Units of Measurement - Dependent Variable

- Re-estimating the new model using Stata yields

```
. reg            packs age income educ;
```

Source	SS	df	MS
Model	3.94503993	3	1.31501331
Residual	375.439167	803	.467545662
Total	379.384207	806	.470700008

Number of obs	=	807
F(3, 803)	=	2.81
Prob > F	=	0.0384
R-squared	=	0.0104
Adj R-squared	=	0.0067
Root MSE	=	.68377

packs	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0020847	.0014381	-1.45	0.148	-.0049076	.0007383
income	5.86e-06	2.80e-06	2.09	0.036	3.69e-07	.0000113
educ	-.0188798	.0084817	-2.23	0.026	-.0355286	-.0022309
_cons	.6426972	.1288045	4.99	0.000	.389864	.8955304

- Notice that all of the t-statistics, F-statistics, and R^2 are unchanged.

Units of Measurement Independent Variable

- Changes in the units of an explanatory variable only change the parameter on that explanatory variable.
- For example, measuring household income in thousands of dollars (*incthous*), the corresponding coefficient must increase by a factor of 1000.
- We would now have

$$\begin{aligned}
\widehat{cigs}_i &= \hat{\beta}_0 + \hat{\beta}_1 age_i + \hat{\beta}_2 income_i \frac{1000}{1000} + \hat{\beta}_3 educ_i \\
&= \hat{\beta}_0 + \hat{\beta}_1 age_i + (1000 \cdot \hat{\beta}_2) \frac{income_i}{1000} + \hat{\beta}_3 educ_i \\
&= \hat{\beta}_0 + \hat{\beta}_1 age_i + (1000 \cdot \hat{\beta}_2) incthous_i + \hat{\beta}_3 educ_i \qquad (2)
\end{aligned}$$

Units of Measurement - Independent Variable

- The following set of results emerge from Stata

```
. reg              cigs age incthous educ;

      Source |       SS           df       MS      Number of obs   =       807
-------------+----------------------------------   F(3, 803)       =      2.81
       Model |  1578.01597         3  526.005322   Prob > F        =    0.0384
    Residual |  150175.667       803  187.018265   R-squared       =    0.0104
-------------+----------------------------------   Adj R-squared   =    0.0067
       Total |  151753.683       806  188.280003   Root MSE        =    13.675

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0416932   .0287628    -1.45   0.148    -.0981524    .0147659
    incthous |   .1171126   .0559031     2.09   0.036     .0073791     .226846
        educ |  -.3775954   .1696335    -2.23   0.026    -.7105728    -.044618
       _cons |   12.85394   2.576089     4.99   0.000      7.79728    17.91061
------------------------------------------------------------------------------
```

- Again, all our test statistics remain unchanged.

Re-Centering

- In the standard multiple regression model specification, we usually write

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{3}$$

which, under the zero conditional mean assumption, yields

$$E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{4}$$

- In this setting, the intercept has a typically useless interpretation as

$$\beta_0 = E(y|x_1 = 0, \ldots, x_k = 0) \tag{5}$$

Re-Centering (cont'd)

- An alternative is to re-center each of the regressors around a "type" of individual of interest (e.g., $x_1 = c_1, \ldots, x_k = c_k$), such as the mean individual.

$$\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \\
&\quad +(\beta_1 c_1 + \cdots + \beta_k c_k) \\
&\quad -(\beta_1 c_1 + \cdots + \beta_k c_k) \\
y &= (\beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k) \\
&\quad + \beta_1 (x_1 - c_1) + \cdots + \beta_k (x_k - c_k) + u \\
y &= \theta_0 + \beta_1 \tilde{x}_1 + \cdots + \beta_k \tilde{x}_k + u \qquad (6)
\end{aligned}$$

where $\tilde{x}_j \equiv (x_j - c_j)$.

- Now

$$\theta_0 = E(y|\tilde{x}_1 = 0, \ldots, \tilde{x}_k = 0) = E(y|x_1 = c_1, \ldots, x_k = c_k) \qquad (7)$$

- Wooldridge (pp. 189-191) also talks about re-scaling.

Linear in Parameters

- As noted before, when we talk about our regression model being linear, we mean *linear in parameters*.
- It is often convenient to use nonlinear transformations of the dependent and/or independent variables.
- In the case of a simple regression model, we might have

$$g(y) = \beta_0 + \beta_1 h(x) + u \qquad (8)$$

so that

$$\beta_0 = E[g(y)|h(x) = 0] \qquad (9)$$

and

$$\beta_1 = \frac{\partial E[g(y)|h(x)]}{\partial h(x)} \qquad (10)$$

The Level-Level Specification

- In the *level-level* specification, we have $g(y) = y$ and $h(x) = x$ so that
$$y = \beta_0 + \beta_1 x + u \tag{11}$$
- In this case
$$\beta_0 = E[y|x = 0] \tag{12}$$
and
$$\beta_1 = \frac{\partial E[y|x]}{\partial x} \tag{13}$$

The Log-Log Specification

- In the *log-log* specification, we have $g(y) = ln(y)$ and $h(x) = ln(x)$ so that
$$ln(y) = \beta_0 + \beta_1 ln(x) + u \tag{14}$$
- In this case
$$\beta_0 = E[ln(y)|ln(x) = 0] \tag{15}$$
and
$$\beta_1 = \frac{\partial E[ln(y)|ln(x)]}{\partial ln(x)} \tag{16}$$
- β_1 has an *elasticity* interpretation, giving the percentage change in y for each percentage change in x.
- This percentage change interpretation holds best for small changes in x.
- Wooldridge (eq. 6.8, p. 192) provides the appropriate calculation for discrete shifts in x.

Example of Log-Log Specification

- Suppose we wish to find out the elasticity of per capita net income with respect to per capita land base, with

$$ln(ynetpc) = \beta_0 + \beta_1 ln(apc) + u \quad (17)$$

where
  - $ln(ynetpc) = ln(\frac{ynet}{hhsize})$ denotes the log of household net income per capita.
  - $ln(apc) = ln(\frac{landcu}{hhsize})$ denotes household cultivated land per capita.

Example of Log-Log Specification

- The following set of results emerge from Stata

reg lnynet lnapc

Source	SS	df	MS
Model	20.9072029	1	20.9072029
Residual	557.985632	1823	.306080983
Total	578.892835	1824	.317375458

Number of obs =	1825
F(1, 1823) =	68.31
Prob > F =	0.0000
R-squared =	0.0361
Adj R-squared =	0.0356
Root MSE =	.55325

lnynetpc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnapc	.1439027	.0174116	8.26	0.000	.1097539	.1780515
_cons	7.705952	.0133034	579.24	0.000	7.67986	7.732044

- The coefficient on *lnapc* (0.144) is the elasticity of per capita net income w.r.t. per capita land base.
- 0.144 implies that a 1% increase in rural households per capita land endowment would lead to an increase of 0.144% in per capita net income.

The Log-Level (or Semi-Log) Specification

- In the *log-level* specification, we have $g(y) = ln(y)$ and $h(x) = x$ so that
$$ln(y) = \beta_0 + \beta_1 x + u \tag{18}$$
- In this case
$$\beta_0 = E[ln(y)|x = 0] \tag{19}$$
and
$$\beta_1 = \frac{\partial E[ln(y)|x]}{\partial x} \tag{20}$$
- $(100 \cdot \beta_1)$ gives the percentage change in y for each unit change in x.
- As in the case of the log-log specification, this percentage change interpretation holds best for small changes in x.

Example of Log-Level Specification

- Suppose we wish to use a log-level model of wages as a function of a individual's education level (Wooldridge, example 2.10), with
$$ln(wage) = \beta_0 + \beta_1 educ + u \tag{21}$$
where
  - *wage* denotes the individual's wage rate.
  - *educ* denotes individual's years of education.

Example of Log-Level Specification

- The following set of results emerge from Stata

```
. reg lwage educ

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(1, 524)       =    119.58
       Model |  27.5606296         1  27.5606296   Prob > F        =    0.0000
    Residual |  120.769132       524  .230475443   R-squared       =    0.1858
-------------+----------------------------------   Adj R-squared   =    0.1843
       Total |  148.329762       525   .28253288   Root MSE        =    .48008

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0827444   .0075667    10.94   0.000     .0678796    .0976092
       _cons |   .5837726   .0973358     6.00   0.000     .3925562     .774989
------------------------------------------------------------------------------
```

- The coefficient on *educ* (0.083) implies that an additional year of education would (on average) lead to an 8.3% increase in the individual's wage rate.

The Level-Log Specification

- In the *level-log* specification, we have $g(y) = y$ and $h(x) = ln(x)$ so that

$$y = \beta_0 + \beta_1 ln(x) + u \tag{22}$$

- In this case

$$\beta_0 = E[y|ln(x) = 0] \tag{23}$$

and

$$\beta_1 = \frac{\partial E[y|ln(x)]}{\partial ln(x)} \tag{24}$$

- $\frac{\beta_1}{100}$ gives the change in y for each percentage change in x.
- As in the case of the log-log specification, this percentage change interpretation holds best for small changes in x.

Example of Level-Log Specification

- Suppose we wish to use a level-log model of how food consumption changes with a given percentage change in income, with

$$xfdconpc = \beta_0 + \beta_1 ln(ynetpc) + u \tag{25}$$

where

- *xfdconpc* denotes food expenditure per capita (in Yuan).
- $ln(ynetpc) = ln(\frac{ynet}{hhsize})$ denotes the log of household net income per capita.

Example of Level-Log Specification

- The following set of results emerge from Stata

```
. reg xfdconpc lnynet

      Source |       SS       df       MS              Number of obs =    1836
-------------+------------------------------           F(  1,  1834) = 1126.27
       Model |  194788218     1  194788218           Prob > F      =  0.0000
    Residual |  317190209  1834  172949.95           R-squared     =  0.3805
-------------+------------------------------           Adj R-squared =  0.3801
       Total |  511978427  1835  279007.317           Root MSE      =  415.87

------------------------------------------------------------------------------
    xfdconpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnynetpc |   578.4513   17.23637    33.56   0.000     544.6464    612.2563
       _cons |   -3271.21   133.6503   -24.48   0.000    -3533.333   -3009.087
------------------------------------------------------------------------------
```

- The coefficient on *lnynetpc* (578) implies that a 1% increase in income would lead to an increase in food expenditures of 5.78 Yuan.

Choosing Logs versus Levels

- Logs are typically used for variables measured in positive currency amounts, such as wages, salaries, sales, or firm market values.
- Rationales for doing so include
  - Doing so provides the convenient elasticity interpretation;
  - Currency metrics are often positively skewed (i.e., with long right-hand tails) and a logarithmic transformation creates a more symmetric distribution.

    ...which is more consistent with the CLM's normality assumption.
- The log transformation is problematic if the variable can take on a zero value.
- One can use R^2 to guide choosing between *level* vs. *log* versions of an independent variable.

  ...but not between a *level* vs. *log* version of an independent variable.

Models with Quadratics

- It is often desirable to incorporate quadtratic terms into a model to allow for increasing or decreasing effects of a variable.
- For example, while one might expect wage rates to increase with experience, the *marginal* impact of experience is likely to decrease with experience.
- In the case of a single independent variable, this would be captured by setting

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u \tag{26}$$

- In this case the marginal effect of x becomes:

$$\frac{\partial E[y|x]}{\partial x} = \beta_1 + 2\beta_2 x \tag{27}$$

- In this case, the marginal (or partial) effect of x on y is no longer constant, but depends on the level of x.

The Wage Example

- Wooldridge (eq. 6.12) presents an example using wages and experience, with

$$\widehat{wage} = \underset{(0.35)}{3.73} + \underset{(0.041)}{0.298}exper - \underset{(0.009)}{0.0061}exper^2 \tag{28}$$

- In this case,

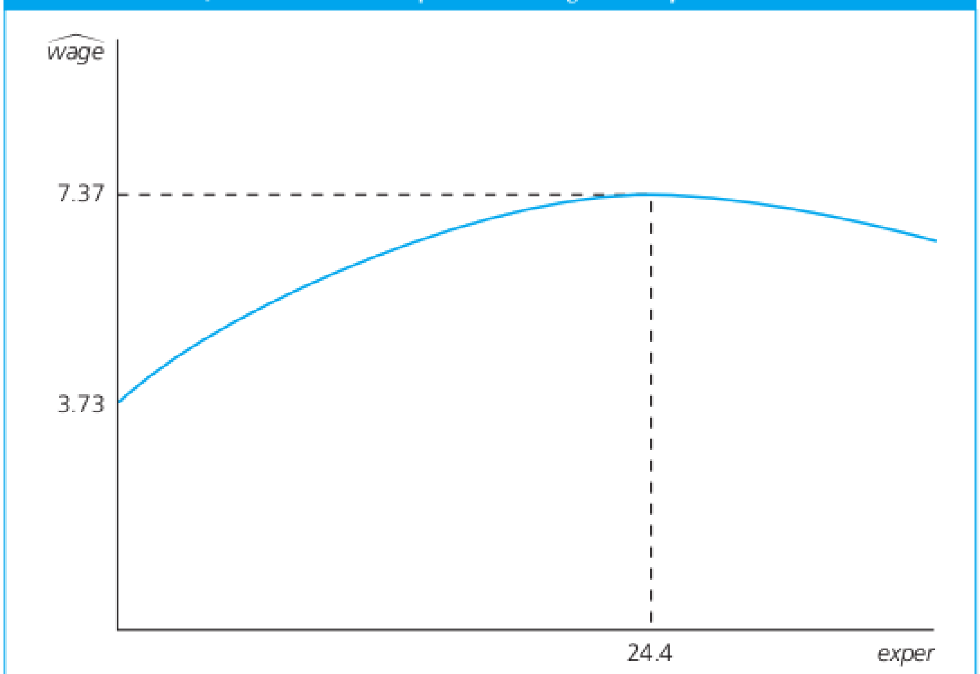
$$\frac{\partial E[wage|exper]}{\partial exper} = 0.298 - 0.0122exper \tag{29}$$

  with experience increasing expected wage at a diminishing rate.
- The turning point in this relationship occurs at:

$$x^* = -\frac{\hat{\beta}_1}{2\hat{\beta}_2} \tag{30}$$

- In the case of the wage example, this occurs at $x^* = 24.4$.

FIGURE 6.1 Quadratic relationship between $\widehat{wage}$ and *exper*.



Interaction Terms

- It is often of interest to allow for **interaction** effects between two variables.
- With two independent variables, this would involve a model such as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u. \quad (31)$$

- In this case, the marginal effect of x_1 now depends on the level of x_2; i.e.,

$$\frac{\partial E[y|x_1]}{\partial x_1} = \beta_1 + \beta_3 x_2. \quad (32)$$

- Note that it is also the case that the marginal effect of x_2 now depends on the level of x_1; i.e.,

$$\frac{\partial E[y|x_2]}{\partial x_2} = \beta_2 + \beta_3 x_1. \quad (33)$$

Housing Example of an Interaction Effect

- Suppose that we are modeling housing prices as a function of house size (sqft) in a log-log model, but we want the marginal effect of house size to depend on the age of the home.
- One specification would be

$$ln(price) = \beta_0 + \beta_1 ln(area) + \beta_2 age + \beta_3 [ln(area) \cdot age] + u. \quad (34)$$

where *area* denotes the house's square footage.
- In this case:

$$\frac{\partial E[ln(price)|ln(area)]}{\partial ln(area)} = \beta_1 + \beta_3 age. \quad (35)$$

- In the Stata results on the next page, we find that there is a significant interaction effect, with the elasticity of price with respect to square footage increasing with house age.

```
. gen agexlarea=age*larea

. reg lprice larea age agexlarea

      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(3, 317)       =    129.60
       Model |  33.8450131         3  11.281671   Prob > F        =    0.0000
    Residual |  27.5939722       317  .087047231   R-squared       =    0.5509
-------------+----------------------------------   Adj R-squared   =    0.5466
       Total |  61.4389853       320  .191996829   Root MSE        =    .29504

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .7204672   .0614051    11.73   0.000     .5996543    .8412802
         age |  -.0261485   .0101517    -2.58   0.010    -.0461218   -.0061753
   agexlarea |   .0027749   .0013039     2.13   0.034     .0002094    .0053403
       _cons |   5.998851   .4710447    12.74   0.000     5.072082     6.92562
------------------------------------------------------------------------------
```


Adjusted R^2

- One limitation of R^2 as a measure of model fit is that there is no penalty for adding variables to a model, even if they provide relatively little explanatory power.
- The adjusted-R^2, denoted $\bar{R}^2$, does such an adjustment, taking into account the loss of degrees of freedom by adding variables.
- Specifically,

$$\bar{R}^2 = \frac{\left(\frac{SSR}{n-k-1}\right)}{\left(\frac{SST}{n-1}\right)} \tag{36}$$

- Essentially, this adjustment argues for simpler models, all else equal.

Including Too Many Factors in a Model

- As Wooldridge points out on pp. 205-206, one has to be careful not to include too many factors in a model.
- In particular, we want to be sure that we are not controlling for the very effect we want to capture.
- In modeling the impact of a change in liquor taxes on highway fatalities, we are trying to capture the fact that taxes discourage liquor consumption and, hence, liquor related road fatalities.
- We would *not* want to specific a model such as

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 beercons + \cdots \tag{37}$$

because then β_1 would be measuring the effect of the beer tax on fatalities holding beer consumption constant.
- This is sometimes referred to as **over controlling**.

Prediction

- We are often interested in using our estimated model for prediction purposes.
- Specifically, we might want to estimate what the expected value of our dependent variable might be for a given "type" of individual, with say $x_1 = c_1, \ldots, x_k = c_k$.
- But we know that

$$\theta_0 \equiv E(y|x_1 = c_1, \ldots, x_k = c_k) = \beta_0 + \beta_1 c_1 + \cdot + \beta_k c_k \tag{38}$$

- A natural estimator would be

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdot + \hat{\beta}_k c_k \tag{39}$$

- But, if you recall from the earlier discussion in this chapter about re-centering, $\hat{\theta}_0$ is just the OLS estimated intercept from the model:

$$y = \theta_0 + \beta_1 \tilde{x}_1 + \cdots + \beta_k \tilde{x}_k + u \tag{40}$$

where $\tilde{x}_j \equiv (x_j - c_j)$, providing also $se(\hat{\theta})$.