6 Multiple Regression Analysis: Further Issues

Introduction: This chapter is important from the point of view of an applied economist and covers aspects which would be encountered in the first stage of data wrangling and model design. So be careful and internalize these ideas to succeed in empirical work.

This chapter answers the following questions

1. What happens if I scale my data from thousands to millions, or change the level of the independent or the dependent variable?

2. What happens when either the independent or dependent variable in a regression are transformed to logs or quardratics?

3. How much should I trust R-square for a regression?

4. How to obtain confidence intervals for OLS predictions?

6.1 Effects of Data Scaling on OLS Statistics

1. Often, data scaling is used for cosmetic purposes, such as to reduce the number of zeros after a decimal point in an estimated coefficient.

2. What happens to regression statistics when you scale data?
   Case 1: Changing dependent variable: y

   (a) $\hat{\beta}$ gets scaled down proportionately,

   (b) Statistical significance i.e. t-stat identical

   (c) R-squared or a measure of goodness of fit is also identical.

   (d) Standard error of the regression (SER) changes. This only reflects difference in unit of measurement.

3. **Why**: scaling of regression should not affect the relation if all variables are scaled proportionately.

4. *Example*: Consider the model of infant birth weights (lb vs. ounces) explained by mother's smoking habits, controlling for family income

$$bwght = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc$$

   (a) $bwght$ is child birth weight, in ounces where 1 lb = 16 ounces.

   (b) $cigs$ is number of ciggarettes smoked by the mother while pregnant, per day.

   (c) $faminc$ is annual family income, in thousands of dollars.

   (d) Notice $\hat{\beta}$, calculate t-stat, R-squared and SER.

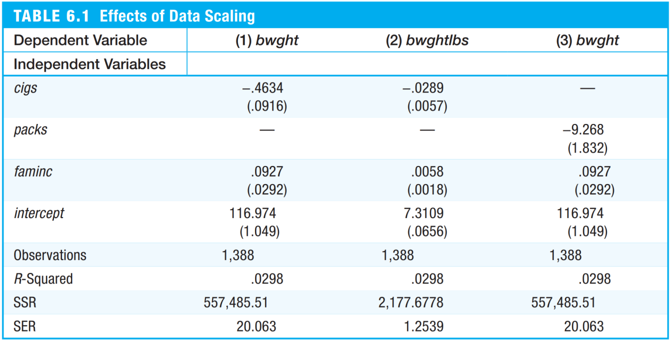

TABLE 6.1 Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Independent Variables			
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

5. **Case 2: changing scale of independent variable:** x - one cigarette pack i.e. $cigs/20$ (table 3)

   (a) Coefficients change cause we alter the rate of change of x-variable.

   (b) The standard error on $packs$ (1.83) is 20 times larger than that on $cigs$ (0.0916) in column 3

6. **Log dependent variable**: In the case of $log(x_i)$, there is no change in slope when the data is transformed. Only the intercept is scaled.

   (a) Estimating elasticities in a log-log model of the following form is invariant to scaling data.

$$\widehat{\ln y} = \hat{\beta}_0 + \hat{\beta}_1 \ln x$$

   Here $\hat{\beta}_1$ is the elasticity estimate.

6.1.1 Beta Coefficients

1. **Level differences in y and x variables**: When the level of independent and dependent variables are very different, then its best to ask *what happens when x is one standard deviation higher or lower*.

   (a) *Standardized variables*: A variable is standardized in the sample by subtracting off its mean and dividing by its standard deviation.

   (b) Also know as the z-score transformation.

   (c) Makes different variables comparable.

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{u}_i$$

When we standardize using σ_y and σ_1 for x_1 we obtain the following equation in their z-scores:

$$\frac{y_i - \bar{y}}{\hat{\sigma}_y} = \frac{\hat{\sigma}_1}{\hat{\sigma}_y}\hat{\beta}_1 \frac{x_{i1} - \bar{x}}{\hat{\sigma}_1} + \frac{\hat{\sigma}_2}{\hat{\sigma}_y}\hat{\beta}_2 \frac{x_{i2} - \bar{x}}{\hat{\sigma}_2} + \hat{u}_i$$

Thus the new standardized coefficients are a transformation of the original $\hat{\beta}$ and are called "standardized coefficients".

2. **Interpreting new** β: if x_1 increases by one standard deviation, then $\hat{y}$ changes by $\hat{b}_1 = \frac{\hat{\sigma}_1}{\hat{\sigma}_y}\hat{\beta}_1$ standard deviations.

   (a) **Note**: In a standard OLS equation, it is not possible to simply look at the size of different coefficients and conclude that the explanatory variable with the largest coefficient is "the most important."

   (b) Even in elasticity regressions, if the variability in x is small and y is large (or vice versa), you could compute standardized coefficients.

(c) *Example*: What is the standardized effect of pollution on house prices? (HPRICE2)

$$\begin{aligned} z\hat{price} = {} & -0.340 znox - 0.143 zcrime + 0.515 zrooms \\ & -0.235 zdist - 0.270 zstratio \end{aligned}$$

This equation shows that a one standard deviation increase in nox decreases price by .34 standard deviation; a one standard deviation increase in crime reduces price by .14 standard deviation. Thus, the same relative movement of pollution in the population has a larger effect on housing prices than crime does.