Chapter 6: Further Issues

1. First we want to show what happens if we add a constant to regressor or multiply the regressor by a constant. We have

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u} = (\hat{\beta}_0 - c\hat{\beta}_1) + \hat{\beta}_1 (X + c) + \hat{u} \quad (1)$$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{c}(cX) + \hat{u} \quad (2)$$

   Equation (1) shows that

adding a constant c to regressor does not change slope, but changes intercept

   Equation (2) shows that

multiplying regressor by c does not change intercept, but changes slope

   Furthermore, we can show

t value and p value all remain unchanged

   Remember: the magnitude of coefficient can be manipulated, but t and p values cannot. Do not take seriously the magnitude of a coefficient that is statistically insignificant.

2. Sometimes a key independent variable is measured on a scale that is difficult to interpret. For example, different exams have different full scores. So the scores are not comparable. To avoid the ambiguity we may ask what happens when the test score is one deviation (not one point) higher. What we need to do is to standardize (obtain z score for) all variables, and run regression using the standardized variables. The coefficient in such regression is called beta coefficient.

   (a) The formula for obtaining z score is

$$\texttt{z-score of X} = \frac{X - \mu_X}{\sigma_X}; \ \texttt{z-score of Y} = \frac{Y - \mu_Y}{\sigma_Y} \quad (3)$$

   (b) By construction z score has $E(\texttt{z-score}) = 0, \texttt{se}(\texttt{z-score}) = 1$ so z score is unit free.

   (c) Beta coefficients measures how much deviation of Y will change when X changes

by one deviation.

(d) The beta regression makes the scale of the regressors irrelevant. Put differently, this regression puts the regressors on equal footing.

(e) Therefore comparing the magnitudes of the resulting beta coefficients is more compelling.

(f) The stata command to standardize X is `egen zx = std(x)`.

3. Sometimes we want to use $\log(Y)$ or $\log(X)$ in the regression. The primary motivation is to account for nonconstant marginal effect of X on Y. Using log variable affects how to interpret the coefficient.

(a) One property of log is that

$$\boxed{\text{difference in log approximates percentage change}}$$

Let w be a number close to zero. From calculus we know

$$\log(1+w) \approx w \tag{4}$$

Letting $w = \frac{\Delta Y}{Y}$ in (4) leads to

$$\begin{aligned}
\frac{\Delta Y}{Y} &= \texttt{Percentage change in Y} && (5)\\
&\approx \log\left(1+\frac{\Delta Y}{Y}\right) = \log\left(\frac{Y+\Delta Y}{Y}\right) && (6)\\
&= \log(Y+\Delta Y) - \log(Y) && (7)\\
&= \texttt{difference in log} && (8)
\end{aligned}$$

For example, the difference of log GDP is the growth rate of GDP.

(b) Consider the log-linear regression

$$\log(Y) = \beta_0 + \beta_1 X + u \tag{9}$$

where

$$\beta_1 = \frac{d\log(Y)}{dX} \approx \frac{\Delta \log(Y)}{\Delta X} \tag{10}$$

The interpretation of β_1 is,

$$\boxed{\text{when } X \text{ changes by one unit, } Y \text{ changes by } 100(\beta_1) \text{ percent}}$$

The regression (9) implies that X has increasing marginal effect on Y

$$\frac{dY}{dX} = \beta_1 Y \Rightarrow \frac{dY/Y}{dX} = \beta_1 \tag{11}$$

or equivalently, when X changes, Y changes at constant percentage rate. Another way to understand this is, note regression (9) implies

$$Y = e^{\beta_0 + \beta_1 X + u} \tag{12}$$

so Y changes exponentially when X changes.

(c) Consider the linear log regression

$$Y = \beta_0 + \beta_1 \log(X) + u \tag{13}$$

Note that

$$\beta_1 = \frac{dY}{d\log(X)} = \frac{dY}{\frac{\Delta X}{X}} \Rightarrow \frac{\beta_1}{100} = \frac{dY}{100\left(\frac{\Delta X}{X}\right)}$$

So the interpretation of β_1 is,

$$\boxed{\text{when } X \text{ changes by one percent, } Y \text{ changes by } \beta_1/100 \text{ units}}$$

The regression (13) implies that X has decreasing marginal effect on Y :

$$\frac{dY}{dX} = \frac{\beta_1}{X} \tag{14}$$

(d) For the log-log regression,

$$\log(Y) = \beta_0 + \beta_1 \log(X) + u \tag{15}$$

The interpretation of β_1 is,

$$\boxed{\text{when } X \text{ changes by one percent, } Y \text{ changes by } \beta_1 \text{ percent}}.$$

Therefore β_1 measures the elasticity.

(e) See Table 2.3 in textbook for detail about interpretation

4. The nonconstant marginal effect can be captured by other functions. For example we may consider a regression with quadratic (squared) term X^2

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u \tag{16}$$

We can show the marginal effect is given by

$$\frac{dY}{dX} = \beta_1 + 2\beta_2 X \tag{17}$$

So the marginal effect is not constant, and it depends on level of X.

5. We can include interaction term (product of two regressors) if the marginal effect of one regressor depends on another. For example consider the regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_1 X_2) + u \tag{18}$$

where $X_1 X_2$ is the interaction term. We can show the marginal effect is given by

$$\frac{dY}{dX_1} = \beta_1 + \beta_2 X_2 \tag{19}$$

So the marginal effect of X_1 on Y depends on level of X_2.

6. Using R squared as measurement of goodness of fit has one drawback. R^2 never falls when more regressors are added, even though some new regressors are statistically insignificant. The adjusted R squared, defined below, resolves this issue (i.e., adjusted R squared will fall if insignificant or irrelevant regressors are added)

$$\texttt{adjusted } R^2 \equiv 1 - \frac{\texttt{RSS}/(n-k-1)}{\texttt{TSS}/(n-1)} \tag{20}$$

(a) When a regressor is added, the `RSS` becomes smaller, but $n-k-1$ becomes smaller too.

(b) When an insignificant regressor is added, falling $n-k-1$ dominates falling `RSS`, so adjusted R squared falls.

(c) We can use adjusted R squared to select regressions provided the regressions have same dependent variable (so `TSS` is the same). A regression is better if its adjusted R squared is higher.

7. Most often we run several regressions for the same dependent variable, and we want to pick a best one. This is a model selection problem. How to select regressions depends on whether the regressions are nested or not

   (a) Regressions are nested if one is a special case of the other. For example, $Y = \beta_0 + \beta_1 X_1 + u$ is a special case of $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$ (the restriction is $\beta_2 = \beta_3 = 0$). F test can be used to compare nested regressions. For this example, the short regression $Y = \beta_0 + \beta_1 X_1 + u$ is better than long regression if the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ cannot be rejected by the F test. Intuitively the long regression is bad because it uses insignificant (irrelevant) regressors.

   (b) Regressions are non-nested if one is not a special case of others. For example, regression $Y = \beta_0 + \beta_1 X_1 + u$ and regression $Y = \beta_0 + \beta_2 X_2 + u$ are non-nested. The adjusted R squared can be used to tell which regression is better.

8. One motivation for running a regression is to do predication.

   (a) After we obtain the OLS estimates, the predicted value for $X_i = c_i, (i = 1, \ldots, k)$ can be computed as

   $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \ldots + \hat{\beta}_k c_k \tag{21}$$

   (b) $\hat{y}$ is a point prediction, and $\hat{y}$ is a random variable because each OLS estimate $\hat{\beta}_i, (i = 1, \ldots, k)$ is a random variable. We can construct a prediction interval to account for the randomness of $\hat{y}$.

   (c) We need standard error of $\hat{y}$. It turns out we can run an auxiliary regression

   $$Y = \theta_0 + \beta_1 (X_1 - c_1) + \ldots + \beta_k (X_k - c_k) + u \tag{22}$$

   where θ_0 is the intercept term. We can show $\hat{y} = \hat{\theta}_0, \mathtt{se}(\hat{y}) = \mathtt{se}(\hat{\theta}_0)$ and therefore the 95% prediction interval is

   $$(\hat{\theta}_0 - 1.96\mathtt{se}(\hat{\theta}_0), \hat{\theta}_0 + 1.96\mathtt{se}(\hat{\theta}_0)) \tag{23}$$

   (d) With 95% probability the true value is inside the prediction interval.

Example, Chapter 6

1. We want to know which variable is more important in determining house price, baths or area. So first we run a multiple regression of rprice on area and baths. The coefficient is 18.36728 for area and 18602.52 for baths. Even though 18602.52 > 18.36728, we can NOT say baths is more important than area since baths and area have different measurements.

   The slope coefficient depends on measurement and we can manipulate it

   For instance we can divide area by 1000, and then the new coefficient will be 1000 times old coefficient (try it!)

2. Therefore we need to standardize all variables, run a regression using standardized variables and obtain beta coefficients. The stata command to standardize (obtain z score for) Y is `egen zy = std(y)`. It is shown that the beta coefficient is .3854153 for area and .4327977 for baths. We can interpret .3854153 as "rprice will change by .3854153 deviation when area changes one deviation, holding baths constant". Because beta coefficient does not depend on measurement, .3854153 < .4327977 is compelling evidence that baths matters more than area in terms of determining house price.

3. Exercise: Please summarize z score for area using command `sum zarea`. Can you guess what are the mean and standard deviation?

4. If we think area has nonconstant marginal effect on rprice, we may try different function form. First we use log of rprice as the dependent variable and run regression

$$\log(rprice) = \beta_0 + \beta_1 area + u$$

   The slope coefficient is .0003735. According to Table 2.3 in textbook, this is a log level model, so we should interpret .0003735 as "rprice will change 100(.0003735) or 0.03735 percent when area changes by one unit (assuming ceteris paribus holds). So this model indicates constant percentage change, and nonconstant absolute change.

5. We can also use quadratic term to capture nonlinearity. The fitted model with $area^2$ is

$$\widehat{rprice} = -7452.548 + 54.49768 area - .0048047 area^2$$

therefore the marginal effect is

$$\frac{\widehat{drprice}}{darea} = \hat{\beta}_1 + 2\hat{\beta}_2(area) = 54.49768 + 2(-.0048047)area$$

The negative coefficient for $area^2$ implies that as area rises, the marginal effect on rprice falls. In other words, rprice rises at decreasing rate. Actually there is a turning point. The house price starts to go down once area exceeds

$$\frac{\beta_1}{-2\beta_2} = \frac{54.49768}{-(2(-.0048047))} = 5671.2885 \tag{24}$$

This result makes sense. Sometimes a house can be too big. (Remember there is something called diminishing marginal utility in economics)

6. We can add interaction term if we believe the marginal effect of area on rprice depends on age. The fitted model with $(area)(age)$ is

$$\widehat{rprice} = 14145.05 + 34.65131area - .092768(area)(age)$$

therefore the marginal effect is

$$\frac{\widehat{drprice}}{darea} = \hat{\beta}_1 + \hat{\beta}_2(age) = 34.65131 - .092768(age)$$

The negative coefficient for the interaction term $(area)(age)$ implies that as house gets old, the marginal effect of area on rprice falls. Again this result makes sense. When the house size gets larger, a new house sees bigger increase in value than an old house.

7. Finally we want to obtain predicted price for rprice for a particular house with baths=2 and age=50. The fitted model is

$$\widehat{rprice} = 19865.82 + 28067.13baths - 100.4649age$$

Let $baths = 2, age = 50$ and we have

$$\widehat{rprice} = 19865.82 + 28067.13(2) - 100.4649(50) = 70976.83.$$

So this house is worth 70976.83 according to his model. Then you can compare this "theoretical" price to the market price, and see if the house is overpriced or underpriced.

8. We know the coefficients are estimated using a specific sample, and we may get different estimates if sample changes. That is why we prefer a prediction interval. We can obtain that interval by running an auxiliary regression

$$\widehat{rprice} = \hat{\theta}_0 + \hat{\beta}_1(baths - 2) + \hat{\beta}_1(age - 50)$$

in which the regressors are (baths-2) and (age-50). Then the confidence interval for the intercept term $\hat{\theta}_0$ is the prediction interval. In this example, the prediction interval is $(67285.34, 74668.33)$ and with 95% probability, the true "theoretical" price is inside that interval. You may forget a house (with 2 bathrooms and 50 years of age) if its market price is above the upper limit 74668.33. You want to immediately buy a house with market price below the lower limit 67285.34.

```
. reg rprice area baths

      Source |       SS       df       MS              Number of obs =     321
-------------+------------------------------           F(  2,   318) =  199.90
       Model |  1.9550e+11     2  9.7749e+10           Prob > F      =  0.0000
    Residual |  1.5550e+11   318   488980887           R-squared     =  0.5570
-------------+------------------------------           Adj R-squared =  0.5542
       Total |  3.5099e+11   320  1.0969e+09           Root MSE      =   22113

------------------------------------------------------------------------------
      rprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   18.36728   2.375508     7.73   0.000     13.69358    23.04098
       baths |   18602.52   2142.533     8.68   0.000     14387.19    22817.85
       _cons |   1504.686   4294.234     0.35   0.726    -6944.013    9953.386
------------------------------------------------------------------------------

* Beta regression
. reg zrprice zarea zbaths

      Source |       SS       df       MS              Number of obs =     321
-------------+------------------------------           F(  2,   318) =  199.90
       Model |  178.234613     2  89.1173067           Prob > F      =  0.0000
    Residual |  141.765385   318  .445803097           R-squared     =  0.5570
-------------+------------------------------           Adj R-squared =  0.5542
       Total |  319.999998   320  .999999995           Root MSE      =  .66768

------------------------------------------------------------------------------
     zrprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       zarea |   .3854153   .0498472     7.73   0.000     .2873434    .4834873
      zbaths |   .4327977   .0498472     8.68   0.000     .3347258    .5308697
       _cons |   2.14e-09   .0372665     0.00   1.000    -.0733201    .0733201
------------------------------------------------------------------------------

* log(rprice) is dependent variable
. reg lrprice area

      Source |       SS       df       MS              Number of obs =     321
-------------+------------------------------           F(  1,   319) =  258.60
       Model |  21.5568712     1  21.5568712           Prob > F      =  0.0000
    Residual |  26.5922721   319  .083361355           R-squared     =  0.4477
-------------+------------------------------           Adj R-squared =  0.4460
       Total |  48.1491434   320  .150466073           Root MSE      =  .28872

------------------------------------------------------------------------------
     lrprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   .0003735   .0000232    16.08   0.000     .0003278    .0004192
       _cons |   10.47457   .0515135   203.34   0.000     10.37322    10.57592
------------------------------------------------------------------------------
```

```
.
. * quadratice term
. gen area2 = area^2

. reg rprice area area2

      Source |       SS       df       MS              Number of obs =     321
-------------+------------------------------           F(  2,   318) =  138.60
       Model |  1.6347e+11     2  8.1734e+10           Prob > F      =  0.0000
    Residual |  1.8752e+11   318   589699897           R-squared     =  0.4657
-------------+------------------------------           Adj R-squared =  0.4624
       Total |  3.5099e+11   320  1.0969e+09           Root MSE      =   24284

------------------------------------------------------------------------------
      rprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   54.49768   8.084438     6.74   0.000     38.59194    70.40343
       area2 |  -.0048047   .0016782    -2.86   0.004    -.0081065   -.0015029
       _cons |  -7452.548   9336.746    -0.80   0.425    -25822.15    10917.05
------------------------------------------------------------------------------

.
. * interaction term
. gen areaage = area*age

. reg rprice area areaage

      Source |       SS       df       MS              Number of obs =     321
-------------+------------------------------           F(  2,   318) =  169.34
       Model |  1.8103e+11     2  9.0513e+10           Prob > F      =  0.0000
    Residual |  1.6997e+11   318   534488648           R-squared     =  0.5158
-------------+------------------------------           Adj R-squared =  0.5127
       Total |  3.5099e+11   320  1.0969e+09           Root MSE      =   23119

------------------------------------------------------------------------------
      rprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        area |   34.65131   1.902994    18.21   0.000     30.90727    38.39536
     areaage |   -.092768   .0143329    -6.47   0.000    -.1209672   -.0645687
       _cons |   14145.05   4137.359     3.42   0.001     6004.996    22285.11
------------------------------------------------------------------------------
```

```
* 95 prediction interval for rprice when baths=2, age = 50
reg rprice baths age

      Source |       SS       df       MS              Number of obs =     321
-------------+------------------------------           F(  2,   318) =  148.08
       Model |  1.6925e+11     2  8.4627e+10           Prob > F      =  0.0000
    Residual |  1.8174e+11   318   571508284           R-squared     =  0.4822
-------------+------------------------------           Adj R-squared =  0.4790
       Total |  3.5099e+11   320  1.0969e+09           Root MSE      =   23906

------------------------------------------------------------------------------
      rprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       baths |   28067.13   1856.694    15.12   0.000     24414.18    31720.09
         age |  -100.4649   43.93043    -2.29   0.023    -186.8959   -14.03386
       _cons |   19865.82   4871.209     4.08   0.000     10281.95    29449.69
------------------------------------------------------------------------------

gen baths2 = baths-2

gen age50 = age-50

reg rprice baths2 age50

      Source |       SS       df       MS              Number of obs =     321
-------------+------------------------------           F(  2,   318) =  148.08
       Model |  1.6925e+11     2  8.4627e+10           Prob > F      =  0.0000
    Residual |  1.8174e+11   318   571508284           R-squared     =  0.4822
-------------+------------------------------           Adj R-squared =  0.4790
       Total |  3.5099e+11   320  1.0969e+09           Root MSE      =   23906

------------------------------------------------------------------------------
      rprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      baths2 |   28067.13   1856.694    15.12   0.000     24414.18    31720.09
       age50 |  -100.4649   43.93043    -2.29   0.023    -186.8959   -14.03386
       _cons |   70976.83   1876.282    37.83   0.000     67285.34    74668.33
------------------------------------------------------------------------------
```

Do File

```
* Do file for chapter 6
set more off
clear
capture log close
cd "I:\311"
log using 311log.txt, text replace
use 311_house.dta, clear
* standard regression
reg rprice area baths
* standardize rprice,  area and baths
egen zrprice = std(rprice)
egen zarea = std(area)
egen zbaths = std(baths)
* beta coefficient
reg zrprice zarea zbaths
* use log(rprice) as Y
gen lrprice = log(rprice)
reg lrprice area
* quadratic term
gen area2 = area^2
reg rprice area area2
* interaction term
gen areaage = area*age
reg rprice area areaage
* 95 prediction interval for rprice when baths=2, age = 50
reg rprice baths age
gen baths2 = baths-2
gen age50 = age-50
reg rprice baths2 age50
log close
exit
```