Week 4. Functional Forms

I. Interaction Terms

1. Two-way Interaction terms

   (a)

   $$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$$

   where $x_3 = x_1 \times x_2$. Assume x_1 and x_2 are continuous variables.

   Because $x_3 = x_1 \times x_2$, we can aslo write the regression model as:

   $$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + e$$

   (b) We can do a t-test to see if the interaction term is needed:
   $H_0 : b_3 = 0$
   $H_A : b_3 \neq 0$

   Even though the addition of b_3 makes b_1 and b_2 insignificant, b_1 and/or b_2 were significant in the model of $y = b_0 + b_1x_1 + b_2x_2 + e$, and then b_3 become significant in the model of $y = b_0+b_1x_1+b_2x_2+b_3x_3+e$, then the interaction term (i.e., x_3) should be kept.

   (c) The net effect of x_1 on y:

   $$\frac{\partial \hat{y}}{\partial x_1} = b_1 + b_3x_2$$

   The effect of x_1 on y varies depending on the value of x_2. This is a partial derivative.

   It is common practice to evaluate and report the above formula at the sample mean of x_2.
   $\longrightarrow$ the average effect of x_1 on $y = b_1 + b_3\bar{x}_2$

   (d) The net effect of x_2 on y:

   $$\frac{\partial \hat{y}}{\partial x_2} = b_2 + b_3x_1$$

   The effect of x_2 on y varies depending on the value of x_1. This is a partial derivative.

   t is common practice to evaluate and report the above formula at the sample mean of x_1.
   $\longrightarrow$ the average effect of x_2 on $y = b_2 + b_3\bar{x}_1$

   (e) Another common way to report the effect of x_1 is to draw a graph in which x-axis refers to the value of x_2 and y-axis is the expected value of y. Likewise, the effect of x_2 is reported as a graph in which x-axis refers to the value of x_1 and y-axis is the expected value of y.

2. Three way interaction

(a)

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 + b_5x_1x_3 + b_6x_2x_3 + b_7x_1x_2x_3 + e$$

(b) The net effect of x_1 on y:

$$\frac{\partial \hat{y}}{\partial x_1} = b_1 + b_4x_2 + b_5x_3 + b_7x_2x_3$$

Therefore, the effect of x_1 on y varies depending on the value of x_2 and x_3.

It is common practice to evaluate and report the above formula at the sample mean of x_2 and x_3.
$\longrightarrow$ The average effect of x_1 on $y = b_1 + b_4\bar{x}_2 + b_5\bar{x}_3 + b_7\bar{x}_2\bar{x}_3$

(c) The net effect of x_2 on y:

$$\frac{\partial \hat{y}}{\partial x_2} = b_2 + b_4x_1 + b_6x_3 + b_7x_1x_3$$

Therefore, the effect of x_2 on y varies depending on the value of x_1 and x_3.

It is common practice to evaluate and report the above formula at the sample mean of x_1 and x_3.
$\longrightarrow$ The average effect of x_1 on $y = b_2 + b_4\bar{x}_1 + b_6\bar{x}_3 + b_7\bar{x}_1\bar{x}_3$

(d) The net effect of x_3 on y:

$$\frac{\partial \hat{y}}{\partial x_3} = b_3 + b_5x_1 + b_6x_2 + b_7x_1x_2$$

Therefore, the effect of x_3 on y varies depending on the value of x_1 and x_2.

It is common practice to evaluate and report the above formula at the sample mean of x_1 and x_2.
$\longrightarrow$ The average effect of x_1 on $y = b_3 + b_5\bar{x}_1 + b_6\bar{x}_2 + b_7\bar{x}_1\bar{x}_2$

(e) The statistical significance of the effect of x_1 on y should be done with a F-test:
$H_0 : b_1 = b_4 = b_5 = b_7 = 0$
H_A : At least, one of b_1, b_4, b_5, and b_7 is not zero.

Therefore, even though b_1 is insignificant, that does not necessarily mean that the effect of x_1 is insignificant. Only if $b_1 = b_4 = b_5 = b_7 = 0$, then we can say that the effect of x_1 on y is statistically zero.

Likewise, even though b_4 is insiginfanct, it doesn't necessarily indicate that the interaction between x_1 and x_2 is insiginfanct. Only if $b_4 = b_7 = 0$, then we can say that the interaction effect between x_1 and x_2 is statistically zero.

(f) The statistical significance of the effect of x_2 on y should be done with a F-test:
$H_0 : b_2 = b_4 = b_6 = b_7 = 0$

H_A : At least, one of b_2, b_4, b_6, and b_7 is not zero.

(g) The statistical significance of the effect of x_3 on y should be done with a F-test:
$H_0 : b_3 = b_5 = b_6 = b_7 = 0$
H_A : At least, one of b_3, b_5, b_6, and b_7 is not zero.

(h) The statistical significance of the effect of $x_1 x_2 x_3$ on y can be tested with a t-test.

II. Quadratic Terms

1. Sometimes an independent variable is squared and then entered into the regression model. The association between x_1 and y is not linear but curvilinear. One of the most common examples is the effect of age on earnings. As ages, earnings increases up to a certain point, and then decreases.
$$y = b_0 + b_1 x_1 + b_2 x_1^2 + e$$

2. To test whether the quadratic (or squared) term is needed, you can use a t-test.
$H_0 : b_2 = 0$
$H_A : b_2 \neq 0$

   Or you can draw a residual plot after estimating $y = b_0 + b_1 x_1 + e$. If the residual plot shows a curvilinear pattern, then add the squared term.

3. The net effect of x_1 on y
$$\frac{\partial \hat{y}}{\partial x_1} = b_1 + 2b_2 x_1$$
The net effect of x_1 on y is dependent on x_1. For example, how much earnings increases as a worker becomes 1 year older depends on the worker's current age.

   It is common practice to evaluate and report the above formula at the sample mean of x_1.
$\longrightarrow$ The average effect of x_1 on $y = b_1 + 2b_2 \bar{x}_1$

4. The reflection point

   is the value of x_1 at which $b_1 + 2b_2 x_1 = 0$. It is the highest point of y for an inverted U-shaped curve, and the lowest point of y for a U-shaped curve.

   That is $x_1 = -\dfrac{b_1}{2b_2}$
For example, if years of experiecne (x_1) is regressed on earnings (y), we estimate $y = b_0 + b_1 x_1 + b_2 x_1^2 + e$. Suppose $b_1 > 0$ and $b_2 < 0$. The curve is an inverted U-shaped line. The reflection point, $-\frac{b_1}{2b_2}$, indicates at what age the expected earnings would be highest on average.

5. An example

   $Wage = -4.817 + .706SCH + .392EXP - .006EXP^2 + e$

   The net effect of SCH = .706. As SCH increases by 1, the expected wage increases by .706 dollars on aveage after controlling for EXP.

   The net effect of $EXP = .392 + 2(-.006)EXP = .392 - .012EXP$

   The net effect of $EXP = .392 - .012EXP$. As EXP increases by 1, the expected wage increases by $.392 - .012EXP$ dollars on aveage after controlling for SCH. Therefore, the net effect of EXP differ by the level of EXP. When $E\bar{X}P = 16.29$, the effect of EXP on Wage is $.392 - .012(.16.29) = .197$.

III. Transformations of Variables

1. Principles

   (a) Given the assumptions for homoscedasticity and linearity, the Gauss-Markov Theorem states that OLS estimates are BLUE (best linear unbiased estimate). This is true regardless of normality. In practice, however, when using sociological data, it is often the case (though not always the case) that the assumptions of homoscedasticity and linearity appear to be more valid when the dependent variable has a distribution that is less skewed and more normal (or at least more symmetric).

   (b) In general, when the dependent variable has a strong positive skew, a log transformation of the dependent variable is sometimes (but not always or necessarily) desirable because in doing so the assumptions of linearity and/or homoscedasticity seem more plausible.

   (c) Ideally, theory would tell us about which functional form is appropriate. In practice, theory is almost never so precise. Therefore, in practice, examination of residual plots and outliers often suggests whether or not a transformation of the dependent variable is useful. Ceteris paribus, simplicity is desirable because it is clearer, but the transformation is desirable usually if it it improves linearity or homoscedasticity.

2. Log Transformation of the Dependent Variables

   (a) Suppose that the relationship between the dependent and independent variables is exponential as follows:

$$\begin{aligned} Y &= e^{b_0+b_1X_1+b_2X_2}\varepsilon \\ &= e^{b_0}e^{b_1X_1}e^{b_2X_2}\varepsilon \end{aligned}$$

   OLS assumes linearity. It means the dependent variable is an additive function of independent variables. Therefore the above equation cannot be estimated with OLS, because, in the above equation, the dependent variable is a multiplicative function of independent variables.

However, by taking log on both sides of the equation, we can convert the multiplicative function to an addictive function:

$$logY = b_0 + b_1 X_1 + b_2 X_2 + \log \varepsilon$$

Antilog (or exponentiation) of the logged Y is Y, thus,

$$e^{b_0 + b_1 X_1 + b_2 X_2} = \hat{Y}$$

(b) Interpretation 1

Suppose we estimated the following model,

$$logY = b_0 + b_1 EDU + b_2 AGE + b_3 MALE + \log \varepsilon$$

where Y is annual earnings. EDU and AGE are continuous variables and MALE is a dummy variable (i.e., female worker is set as a reference group).

As EDU increases by 1 year, the expected annual earnings will increase by $[(e^{b_1} - 1) \times 100]\%$, holding AGE and gender constant. That is, $[(e^{b_1} - 1) \times 100]\%$ increase is the net effect of EDU on annual earnings after controlling for age and gender.

As AGE increases by 1 year, the expected annual earnings will increase by $[(e^{b_2} - 1) \times 100]\%$, holding EDU and gender constant.

For male workers, the expected annual earnings is $[(e^{b_3} - 1) \times 100]\%$ higher than female workers, other things being equal.

(c) Interpretation 2 (not recommended)

As EDU increases by 1 year, the expected annual earnings is multiplied by e^{b_1}, holding AGE and gender constant.

From the above equation, $\hat{Y} = e^{b_0} e^{b_1 EDU} e^{b_2 AGE} e^{b_3 MALE}$.

Suppose $\hat{b}_1 = .07032$.
$e^{.07032} = 1.07$

Thus, when EDU increases by 1 unit then predicted annual earnings is multiplied by 1.07. It is better to say that as EDU increases by 1 year, the expected earnings increases by 7% (1.07 - 1 = .07).

(d) Interpretation 3

The expected annual earnings at EDU=16 and AGE=30 for male $= e^{b_0 + b_1(16) + b_2(30) + b_3}$

(e) Interpretation 4

Suppose we estimated the following model,

$$logY = b_0 + b_1 EDU + b_2 BLK + b_3 HISP + \log \varepsilon$$

where Y is annual earnings. BLK and HISP are dummy variables, referring to African Americans and Hispanics respectively. Whites are set as a reference group in this model.

b_2 refers to the expected log annual earnings gap between whites and blacks, net of EDU. Suppose b_2 and b_3 are negative. Afro-Americans' annual earnings is $[(1 - e^{b_2}) \times 100]\%$ lower than whites, after controlling for education.

For example, suppose $b_2 = -.22$.

$e^{-.22} = .803$.

African Americans annual earnings is (Whites' annual earnings * .803). That is, African Americans annual earnings is only 80.3% of that of whites, net of education.

Or you can say that African Americans annual earnings is (1-.803 = .197)*100 = 19.7% lower than whites, net of education.

(f) Interpretation 5

The gap between blacks and Hispanics can be computed as $e^{b_2 - b_3}$.

Suppose $b_2 = -.22$ and $b_2 = -.30$

The gap between blacks and Hispanics is $-.22 - (-.30) = .08$. Thus, net of education, Afro-Americans' annual earnings is on average $[e^{-.22-(-.30)} \times 100]\%$ higher than that of Hispanics.

(Nowadays, studies usually employ much more complicated statistical models than OLS. Nonetheless, high quality empirical research can still be conducted with OLS. An example is Kim and Sakamoto (2010, ASR). Read this paper to check how to interpret the log transformed dependent variable.)

3. Log-log Model

   (a) Both the dependent variable and an independent variable are log-transformed.

   $$\log Y = b_0 + b_1 \log X_1 + b_2 X_2 + b_3 X_3 + e$$

   A good example is that Y is children's income and X_1 is parents' income.

   (b) How to interpret:
   As x increases by 1%, y is expected to rise by b_1%. This is "elasticity."
   Be cautious not to multiply b_1 by 100.

   (c) Another caution: As x increases by g%, the expected change in y is
   $\big(exp(\beta \cdot \ln\big(\frac{100+g}{100}\big)) - 1\big) \times 100\big)$%. Do not multiply b_1 by g when g is more than 10.

4. Multiplicative Model

   (a) Multiplicative models can be estimated with OLS if we log-transform both sides as follows:

   $$Y = \delta_0 X_1^{\delta_1} X_2^{\delta_2} X_3^{\delta_3} \varepsilon$$
   $$\log Y = \log \delta_0 + \delta_1 \log X_1 + \delta_2 \log X_2 + \delta_3 \log X_3 + \log \varepsilon$$

   (b) How to interpret:
   $$\frac{\frac{\partial \hat{y}}{\hat{y}}}{\frac{\partial x}{x}} = \frac{\text{\% change in expected y}}{\text{\% change in x}} = \delta$$
   As X changes by 1%, the expected change in Y is δ %. In economics, this kind of effect is called elasticity.

   The exact interpretation is that when x increases by g%, the expected change is Y is $\big(exp(\delta \cdot \log\big(\frac{100+g}{100}\big)) - 1\big) \times 100$ percentage.

   For example, let's say, we have the following estimate:

   log(retirement-savings) = 10 + 1.2·log(income) + e

   In this case, we can say that as income increases by 1%, the contribution to retirement savings increases by 1.2%.

   Then, how much retirement savings will rise if income is increased by 40%?

   The answer is

   i. $exp(1.2 \cdot log(1.4)) = 1.4975$
   ii. (1.4975 - 1)* 100 = 49.7%. Note that it is not 48%.

   When δ is very close to 1, you can do $\delta \cdot$ (percent change in x). In all other situation, you should be very careful.
   (Multiplicative models are not so common in sociology. One exception is Sakamoto and Kim (2010, Sociological Perspective).)

(c) An example: Population Growth

$$\begin{aligned} PopulationSize &= Ae^{b_1 TIME}\varepsilon \\ \log(PopulationSize) &= \log A + b_1 TIME + log\varepsilon \end{aligned} \tag{1}$$

where A is the population size at time 0; b_1 refers to the population growth rate (percentage effect of another year of time).

Suppose,

Pop = .2 billion $\times e^{.02T}$.
Or $\log(\hat{Pop}) = \log(.2 \text{ billion}) + .02(TIME)$

That is, the current population is .2 billion and it supposes to grow 2% each year. 100 years later, the expected US population is .2 billion $\times e^{.02(100)} = 1.48$ billion.

The US population in 1917 was 103.3 million and in 2017, it is expected to be 320.0 million. What is the average population growth rate over the last 100 years?

Growth rate (%) = $\Big(exp\Big(\dfrac{log(Pop_{time_1}) - log(Pop_{time_0})}{\text{number of years}}\Big) - 1\Big) \times 100$
Thus,

i. $\dfrac{log(320.0) - log(103.3)}{100} = .0113$
ii. $exp(.0113) = 1.011371$
iii. 1.01137 - 1 = .01137
iv. .01137 × 100 = 1.14%

5. Transformation to standardized scores

The absolute magnitude of the estimated coefficients vary, in part, by the scale of variables. Suppose we would like to estimate the following model where Y refers to annual earnings:

$Y = b_0 + b_1 EDU + b_2 HEIGHT + e$

As the unit of earnings change from US dollars to Korean won, the estimated coefficients will change. Furthermore, if we are interested in comparing the effects of education and heights on earnings, the comparison of b_1 and b_2 is not so useful because the unit change in education and height cannot be measured equivalently.

To resolve this problem, we can transform both dependent and independent variables to their standardized scores (or z-scores) as follows:

$Z_Y = b_1^* Z_{EDU} + b_2^* Z_{HEIGHT} + e$

b_1^* and b_2^* are standardized coefficients.

As height increases by 1 standard deviation, the annual earnings is expected to increase by b_2 standard deviation after controlling for education.

As standard deviation of education increases by 1, the expected change in the standard deviation of Y (i.e., annual earnings) will increase by b_1, after controlling for height.

Note that the intercept of the second equation (in which all variables are transformed into z-scores) is zero by definition, because the mean of standardized score is zero by definition.

From the 1st equation,

$$b_0 = \hat{Y} - (b_1 EDU + b_2 HEIGHT) = \bar{Y} - b_1 E\bar{D}U + b_2 HEI\bar{G}HT$$

when all variables are transformed to their standardized scores,

$$b_0^* = Z_{\bar{Y}} - (b_1^* Z_{E\bar{D}U} + b_2^* Z_{HEI\bar{G}HT}) = 0$$

Table 1: Transformations of Variables and Their Effects

y	x	Common name of the effect	Interpretation
Original	Original	Metric, raw, or unstandardized effect	- As x increases by 1 unit, the expected change in y is $\hat{\beta}$.
ln(y)	Original	Proportionate effect; % effect; rate of return to x	- As x increases by 1 unit, the expected change in y is $(\hat{\beta} \times 100)\%$ (Use this simple interpretation when $\lvert\hat{\beta}\rvert$ is less than .08.) - As x increases by 1 unit, the expected change in y is $(e^{\hat{\beta}} - 1)\%$.
Original	ln(x)	Metric or raw effect for a proportionate change of x	- As x increases by 1%, the expected change in y is $\hat{\beta}$. - As x increases by g%, the expected change in y is $\hat{\beta} \times \ln\left(\frac{100+g}{100}\right)$
ln(y)	ln(x)	Elasticity of y with respect to x	- As x increases by 1%, the expected change in y is $\hat{\beta}\%$. - As x increases by g%, the expected change in y is $\big(exp(\beta \cdot \ln\left(\frac{100+g}{100}\right)) - 1\big) \times 100\big)\%$.
Z_y	Z_x	Standardized effect	- As x increases by 1 st.dev., y is expected to increases by $\hat{\beta}$ standard deviation.

Appendix. Extreme Brief Review of Logarithms

1. $Y = a \times b$
   $\log Y = \log a + \log b$

   $100 = 4 \times 25$,
   $\log(100) = \log 4 + \log 25$,
   $4.605 = 1.386 + 3.219$

2. $Y = a \times b^x$
   $\log Y = \log a + x \log b$

   $100 = 4 \times 5^2$
   $\log 100 = \log 4 + 2 \log 5$
   $4.605 = 1.386 + 2(1.609)$

3. $Y = \frac{a}{b} = ab^{-1}$
   $\log Y = \log a - \log b$

   $100 = \frac{4}{.04} = 4(.04)^{-1}$
   log 100 = log 4 - log .04
   4.605 = 1.386 - (-3.219) = 1.386 + 3.219

4. $X = \log Y$
   $e^X = e^{\log Y} = Y$

   $4.605 = \log 100$
   $e^{4.605} = 100$

5. $X = e^Y$
   $\log X = log(e^Y) = Y$

   $100 = e^{4.60517}$
   $\log 100 = \log(e^{4.60517}) = 4.60517$

6. Y = a + b
   log Y = log (a+b)
   i.e., $\log Y \neq \log a + \log b$

7. $e^0 = 1$

8. log 0 = - Inf
   Therefore, the negative numbers cannot be log transformed.