Interaction/Moderation Effects Using Regression

Details behind interaction models. Use of centered, standardized and raw data in interaction models.

Tomáš Oleš

Department of Economic Policy
Faculty of Economics and Finance

February 1, 2025

Agenda

- Understand details behind interaction models.
- Understand details on use of centered, standardized and raw data in interaction models.
- Develop an interaction/moderation model and estimate it using STATA.

Do attitude towards immigrants differs from a low-unemployment year (2007) to a high-unemployment year (2008)?

Understanding Interaction/Moderation Effects

- **Linear Additive Models:** Assume that the effect of an independent variable on a dependent variable is **constant** across all values of other independent variables.
- **Non-Additive (Interaction) Models:** Allow the effect of one independent variable to vary depending on another variable, providing a more nuanced understanding.

Why Interaction Effects Matter

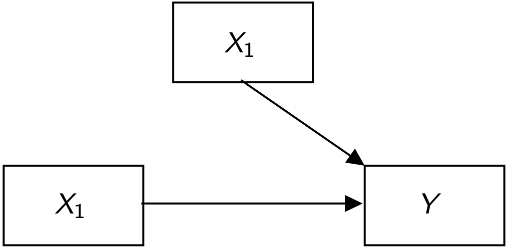
- Standard statistical models often assume **invariance** in relationships between variables.
- Interaction models help identify situations where this assumption **does not hold**, leading to more accurate conclusions.

Defining Interaction/Moderation Effects

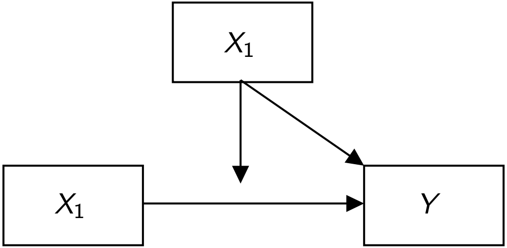
- Occurs when a **moderator variable ($X1$)** influences the relationship between an **independent variable ($X2$)** and a **dependent variable (Y)**.
- This is demonstrated by a **significant change** in the effect size and/or direction of $X2$ on Y at different values of $X1$.
- Interaction effects reveal **conditional relationships** that linear additive models may overlook.
- They are essential in empirical research to **capture complexity** in social, economic, and behavioral studies.

(See next figure for visual representation.)

Interaction/Moderation Effect Diagram



Main Effect



Interaction/Moderation Effect

Product-Term Approach

- The product-term approach involves creating a new variable (X_3) by multiplying two interacting variables ($X_1 \times X_2$) and including it in the regression model alongside X_1 and X_2.
- This results in the following regression model:

$$E[Y_i] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} \tag{1}$$

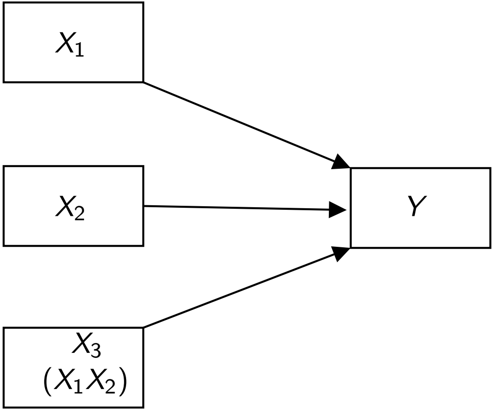
Interpreting the Product-Term Approach

- In an additive model, the equation is:

$$E[Y_i] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{2}$$

- Here, β_1 and β_2 represent the main effects of X_1 and X_2 on Y, assuming their effects are constant.
- In the interaction model (Equation (1)), β_1 and β_2 now represent conditional effects, meaning their impact on Y varies depending on the value of the other variable.

Product-Term Interaction Diagram



```
*Load the dataset
use workout.dta, clear

*Generate the interaction term manually
gen healthage = health *age

*Regression with manually created interaction term
reg whours health age healthage

*Alternative: Using factor variable notation
reg whours c.health c.age c.health#c.age

*Simplified notation using ##
reg whours c.health##c.age
```

Interaction Between a Continuous Predictor and a Continuous Moderator

Population Regression Model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + \varepsilon_i \tag{3}$$

Model Specification:

- X_1: Predictor (Health Motivation)
- X_2: Continuous moderator (Age)
- $E[Y_i]$: Expected number of hours spent working out in a month
- β_0: Expected workout hours when both health motivation and age are zero
- β_1: Effect of health motivation when age is zero
- β_2: Effect of age when health motivation is zero
- β_3: Change in the effect of health motivation on workout hours as age increases

Interpretation:

- The interaction term β_3 determines whether the relationship between health motivation and workout hours varies across different age levels.
- A positive β_3 suggests that the impact of health motivation on workout hours increases with age, while a negative β_3 suggests a decreasing effect.

Conditional Effects at Different Ages

$$E[Y_i] = (\beta_0 + a\beta_2) + (\beta_1 + a\beta_3)X_{1i} \tag{4}$$

- When $X_2 = 20$:
$$E[Y_i] = (\beta_0 + 20\beta_2) + (\beta_1 + 20\beta_3)X_{1i} \tag{5}$$
- When $X_2 = 30$:
$$E[Y_i] = (\beta_0 + 30\beta_2) + (\beta_1 + 30\beta_3)X_{1i} \tag{6}$$
- When $X_2 = 40$:
$$E[Y_i] = (\beta_0 + 40\beta_2) + (\beta_1 + 40\beta_3)X_{1i} \tag{7}$$
- When $X_2 = 50$:
$$E[Y_i] = (\beta_0 + 50\beta_2) + (\beta_1 + 50\beta_3)X_{1i} \tag{8}$$
- When $X_2 = 60$:
$$E[Y_i] = (\beta_0 + 60\beta_2) + (\beta_1 + 60\beta_3)X_{1i} \tag{9}$$

```
. *Regression with manually created interaction term
. reg whours health age healthage

      Source |       SS           df       MS      Number of obs   =       210
-------------+----------------------------------   F(3, 206)       =      5.48
       Model |  808.220699         3  269.4069     Prob > F        =    0.0012
    Residual |  10119.7031       206  49.1247724   R-squared       =    0.0740
-------------+----------------------------------   Adj R-squared   =    0.0605
       Total |  10927.9238       209  52.2867168   Root MSE        =    7.0089

------------------------------------------------------------------------------
      whours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      health |  -1.796974   1.361392    -1.32   0.188    -4.481022    .8870748
         age |  -.4756232    .191777    -2.48   0.014    -.8537204    -.097526
   healthage |   .0639482   .0368331     1.74   0.084      -.00867    .1365665
       _cons |    27.6244   6.912041     4.00   0.000     13.99699    41.25181
------------------------------------------------------------------------------
```

Conditional Effect of Health on Workout Hours at Different Ages

- We examine the conditional effect of health motivation on workout hours for individuals aged 16, 26, 36, 46, 56, 66, and 76.

```
margins, dydx(health) at(age=(16(10)76))
```

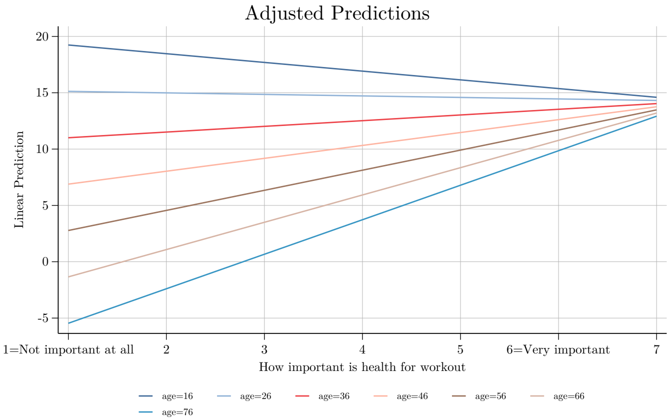
- The dydx(health) refers to the change-in-Y/change-in-X ratio.
- The (10) between 16 and 76 specifies the increment for age.

```
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
health       |
         _at |
          1  |  -.7738018   .8284528    -0.93   0.351    -2.407135    .8595316
          2  |  -.1343195   .5564723    -0.24   0.810    -1.231431    .9627917
          3  |   .5051629    .452023     1.12   0.265    -.3860216    1.396347
          4  |   1.144645   .6085426     1.88   0.061    -.0551248    2.344415
          5  |   1.784128   .8986986     1.99   0.048     .0123013    3.555954
          6  |    2.42361   1.231394     1.97   0.050    -.0041401     4.85136
          7  |   3.063092    1.57998     1.94   0.054    -.0519124    6.178097
------------------------------------------------------------------------------
```

Results:

- The effect of health on workout hours is not statistically significant for ages 16, 26, and 36.
- The effect is statistically significant ($p < 0.1$) for ages 46, 56, 66, and 76.
- The magnitude of this effect increases with age.
- At age 76, each unit increase in health motivation leads to an average increase of 3 workout hours per month.

```
*Calculate the mean-Y at all six values of health for each age 16, 26,
    36, 46, 56, 66 and 76
margins, at(health=(1(1)6) age=(16(10)76))
marginsplot, noci x(health) recast(line) xtitle("How important is health
     for workout")
```



Interaction Between a Continuous Predictor and a Dummy Moderator

Population Regression Model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + \varepsilon_i \tag{10}$$

Model Specification:

- X_1: Predictor (Age, measured in years)
- X_2: Dummy-variable moderator (Gender, $0 =$ Women, $1 =$ Men)
- $E[Y_i]$: Expected number of hours spent working out in a month
- β_1: Coefficient on age for women
- β_2: Coefficient on gender when age is zero
- β_3: Difference in slope of age between men and women

Interpretation:

- β_3 determines whether the effect of age on workout hours is stronger or weaker for men compared to women.
- Moving from $X_2 = 0$ (women) to $X_2 = 1$ (men), the slope coefficient on age changes by β_3.

Conditional Effects:

$$E[Y_i] = \beta_0 + \beta_1 X_{1i} + \beta_2(0) + \beta_3(X_{1i} \times 0) \tag{11}$$

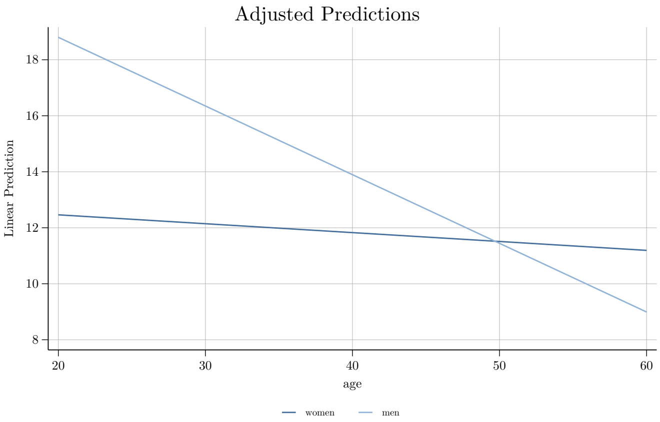
$$E[Y_i] = \beta_0 + \beta_1 X_{1i} \tag{12}$$

When $X_2 = 0$ (for women), the effect of age on workout hours is given by β_1.

```
. *Estimate whether the effect of age on whours is different for women and men
. reg whours c.age i.gender c.age#i.gender

      Source |       SS           df       MS      Number of obs   =       210
-------------+----------------------------------   F(3, 206)       =      8.49
       Model |  1202.50438         3  400.834794   Prob > F        =    0.0000
    Residual |  9725.41943       206  47.2107739   R-squared       =    0.1100
-------------+----------------------------------   Adj R-squared   =    0.0971
       Total |  10927.9238       209  52.2867168   Root MSE        =     6.871

------------------------------------------------------------------------------
      whours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0317262   .0546461    -0.58   0.562    -.1394635    .0760111
             |
      gender |
        men  |   10.61049    3.25435     3.26   0.001      4.19439    17.02659
             |
gender#c.age |
        men  |  -.2135282   .0789673    -2.70   0.007    -.3692158   -.0578406
             |
       _cons |   13.09677   2.290885     5.72   0.000     8.580188    17.61336
------------------------------------------------------------------------------
```

```
. *Estimate the margins
. margins, dydx(age) at(gender=(0 1))

Average marginal effects                        Number of obs     =        210
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age

1._at        : gender          =           0

2._at        : gender          =           1

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |  -.0317262   .0546461    -0.58   0.562    -.1394635    .0760111
          2  |  -.2452544   .0570056    -4.30   0.000    -.3576435   -.1328652
------------------------------------------------------------------------------
```

```
*Plot the difference beween these two coeffitiets \beta_3
margins, at(age=(20(10)60) gender=(0 1))
marginsplot, noci x(age) recast(line)
```



Interaction Between a Dummy Predictor and a Dummy Moderator

Population Regression Model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + \varepsilon_i \tag{13}$$

Model Specification:

- X_1: Dummy predictor (Gender, 0 = Women, 1 = Men)
- X_2: Dummy moderator (Marital Status, 0 = Married, 1 = Single)
- $E[Y_i]$: Expected number of hours spent working out in a month
- β_0: Expected workout hours for married women
- β_1: Effect of gender on workout hours for married individuals
- β_2: Effect of marital status on workout hours for women
- β_3: Difference in the effect of gender on workout hours between married and single individuals

Conditional Effects:

- When $X_2 = 0$ (Married):
$$E[Y_i] = \beta_0 + \beta_1 X_{1i} \tag{14}$$
- When $X_2 = 1$ (Single):
$$E[Y_i] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{1i} \tag{15}$$

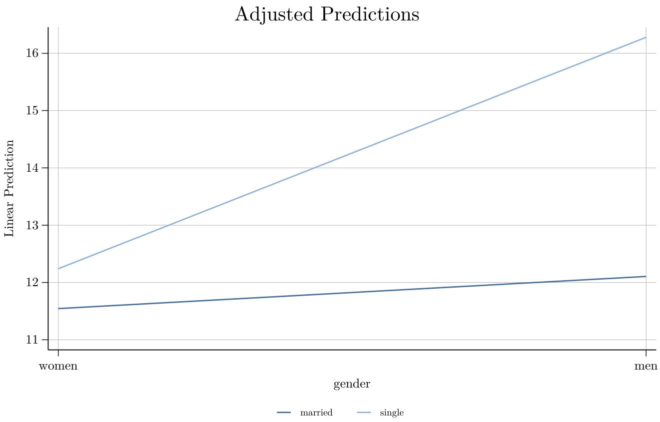
Interpretation:

- The interaction term β_3 determines whether the effect of gender on workout hours differs by marital status.
- A positive β_3 suggests that the effect of gender on workout hours is stronger for single individuals, while a negative β_3 suggests a weaker effect.

```
. *Estimate the interaction between a dummy predictor (gender) and a
   dummy moderator (marital status)
. reg whours i.gender i.marital i.gender#i.marital

      Source |       SS           df       MS      Number of obs   =
         210
-------------+----------------------------------   F(3, 206)       =
        4.55
       Model |  678.978763         3  226.326254   Prob > F        =
      0.0041
    Residual |   10248.945       206  49.7521604   R-squared       =
     0.0621
-------------+----------------------------------   Adj R-squared   =
    0.0485
       Total |  10927.9238       209  52.2867168   Root MSE        =
      7.0535

------------------------------------------------------------------------------

      whours |      Coef.   Std. Err.      t    P>|t|     [95% Conf.
   Interval]
```

```
. *Estimate the margins
. margins, dydx(gender) at(marital=(0 1))

Conditional marginal effects                    Number of obs     =
          210
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : 1.gender

1._at        : marital         =           0

2._at        : marital         =           1

------------------------------------------------------------------------------

             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf.
    Interval]
-------------+----------------------------------------------------------------

0.gender     |  (base outcome)
```

```
*Plot the difference beween these two coeffitiets \beta_3
margins, at(gender=(0 1) marital=(0 1))
marginsplot, noci x(gender) recast(line)
```



Interaction Between a Continuous Predictor and a Polytomous Moderator

Population Regression Model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{1i} \times X_{2i}) + \beta_5 (X_{1i} \times X_{3i}) + \varepsilon_i \tag{16}$$

Model Specification:

- X_1: Continuous predictor (Age)
- X_2: First dummy-variable moderator (University education, $0 =$ No, $1 =$ Yes)
- X_3: Second dummy-variable moderator (More than university education, $0 =$ No, $1 =$ Yes)
- $E[Y_i]$: Expected number of hours spent working out in a month
- β_0: Expected workout hours for individuals with secondary/high school education when age is zero
- β_1: Effect of age on workout hours for individuals with secondary/high school education
- β_2: Difference in workout hours between university-educated individuals and those with secondary/high school education when age is zero
- β_3: Difference in workout hours between individuals with more than university education and those with secondary/high school education when age is zero
- β_4: Difference in the effect of age on workout hours between university-educated individuals and those with secondary/high school education
- β_5: Difference in the effect of age on workout hours between individuals with more than university education and those with secondary/high school education

Conditional Effects:

- When $X_2 = 0$ and $X_3 = 0$ (Secondary/High School):
$$E[Y_i] = \beta_0 + \beta_1 X_{1i} \tag{17}$$
- When $X_2 = 1$ and $X_3 = 0$ (University):
$$E[Y_i] = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_{1i} \tag{18}$$
- When $X_2 = 0$ and $X_3 = 1$ (More than University):
$$E[Y_i] = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_{1i} \tag{19}$$

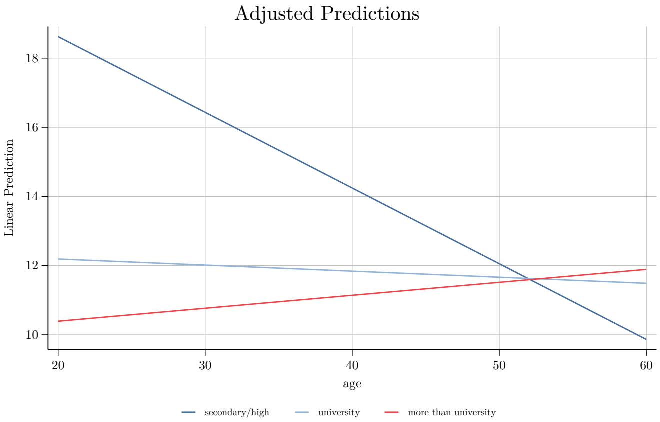
Interpretation:

- The interaction terms β_4 and β_5 determine whether the effect of age on workout hours differs by education level.
- A positive β_4 suggests that age has a stronger effect on workout hours for university-educated individuals compared to those with secondary/high school education.
- A positive β_5 suggests that age has a stronger effect on workout hours for individuals with more than university education compared to those with secondary/high school education.

```
. *Estimate whether the effect of age on whours is dependent on educational level
. reg whours c.age i.educ c.age#i.educ

      Source |       SS           df       MS      Number of obs   =       210
-------------+----------------------------------   F(5, 204)       =      6.27
       Model |  1455.73963         5  291.147926   Prob > F        =    0.0000
    Residual |  9472.18418       204  46.4322754   R-squared       =    0.1332
-------------+----------------------------------   Adj R-squared   =    0.1120
       Total |  10927.9238       209  52.2867168   Root MSE        =    6.8141

-----------------------------------------------------------------------------------------
                 whours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------+----------------------------------------------------------------
                    age |  -.2190014   .0579104    -3.78   0.000    -.3331811   -.1048217
                        |
                   educ |
            university  |  -10.46198   4.363082    -2.40   0.017    -19.06449   -1.859459
more than university    |  -13.36174   4.018536    -3.33   0.001    -21.28493   -5.438548
                        |
             educ#c.age |
            university  |   .2014475    .102947     1.96   0.052     -.001529     .404424
more than university    |   .2564962   .0985739     2.60   0.010      .062142    .4508504
                        |
                  _cons |   23.00372   2.148345    10.71   0.000     18.76791    27.23953
-----------------------------------------------------------------------------------------
```

```
. *Estimate the margins
. margins, dydx(age) at(educ=(1 2 3))

Average marginal effects                          Number of obs     =        210
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age

1._at        : educ            =           1

2._at        : educ            =           2

3._at        : educ            =           3

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |  -.2190014   .0579104    -3.78   0.000    -.3331811   -.1048217
          2  |  -.0175539   .0851144    -0.21   0.837    -.1853707    .1502629
          3  |   .0374948   .0797696     0.47   0.639    -.1197838    .1947734
------------------------------------------------------------------------------
```

```
*Plot the difference beween these three coeffitients
margins, at(age=(20(10)60) educ=(1 2 3))
marginsplot, noci x(age) recast(line)
```



Significant versus Non-Significant Interaction

Consideration of Interaction Terms:

- Up to now, we have assumed that interaction terms are statistically significant.
- However, what should be done if an interaction term is statistically non-significant?

Best Practices in Model Specification:

- If an interaction was hypothesized a priori (before data collection), it should remain in the model even if non-significant.
- In social sciences, particularly in non-experimental research, interactions are often examined a posteriori (after data collection).
- In such cases, exclusion of non-significant interactions is recommended to ensure a parsimonious and less complex model.

Focus on Interpretation:

- Interaction models emphasize the interpretation of interaction terms and simple (conditional) effects.
- Main effects of variables involved in interaction terms should not be the primary focus.

Why Consider Transformations?

- So far, we have worked with raw (untransformed) data to understand interactions.
- In an interaction model, coefficients reflect slopes when the moderator is zero.
- However, if the moderator does not have zero in its scale, interpretation becomes difficult.

Mean-Centring

How Centring Works:

- To improve interpretability, the moderator/predictor variable can be centred.
- The most common approach is centring at the mean: subtracting the mean value from each observation.
- With centred data, coefficients reflect the slope at the mean value of the moderator.
- The coefficient on the interaction term remains unchanged by centring.

Z-Score Standardization

How Standardization Works:

- Another transformation is the z-score standardization.
- This transformation subtracts the mean and divides by the standard deviation.
- Similar to centring, standardized coefficients reflect slopes at the mean.
- Interpretation shifts from raw units to standard deviations.

Choosing Between Raw, Centred, and Standardized Data

General Recommendation:

- Unless there is a specific reason, we recommend working with raw data.
- This approach maintains interpretability in the original metric.
- Raw data also provides flexibility in using prediction equations.
- When using categorical variables, centring or standardization is usually unnecessary.

References I

Laffers, L. (2021). *Draft poznámok k predmetu Moderná Aplikovaná regresia 1*. UMB Banská Bystrica.

Mehmetoglu, M. and Jakobsen, T. G. (2022). *Applied Statistics using Stata: a Guide for the Social Sciences*. Sage.