Linear Regression Models with Interaction/Moderation

Rose Medeiros
StataCorp LLC
Stata Webinar
March 19, 2019

Contents

1 Introduction
  1.1 Goals ......................................................... 1

2 Estimation
  2.1 Including Categorical Variables .................................. 3
  2.2 Including Interactions ....................................... 5

3 Postestimation
  3.1 About Postestimation ....................................... 10
  3.2 Investigating Categorical by Categorical Interactions .......... 10
  3.3 Investigating Categorical by Continuous Interactions ........... 19
  3.4 Investigating Continuous by Continuous Interactions ........... 24

4 Conclusion
  4.1 Graphing Extras ........................................... 33
  4.2 Conclusion .................................................. 36

1 Introduction

1.1 Goals

Goals

• Learn how to use factor variable notation when fitting models involving
  ◦ Categorical variables
  ◦ Interactions
  ◦ Polynomial terms

• Learn how to use postestimation tools to interpret interactions
  ◦ Tests for group differences
  ◦ Tests of slopes
  ◦ Graphs
A Linear Model

- We'll use data from the National Health and Nutrition Examination Survey (NHANES) for our examples.
  
  \texttt{webuse nhanes2}

- We'll start with a basic model for \texttt{bmi} using \texttt{age} and \texttt{sex (female)}.

- Before we fit the model, let's investigate the variables using \texttt{codebook}.

  \texttt{codebook bmi age female}

<table>
<thead>
<tr>
<th>bmi: Body Mass Index (BMI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>type: numeric (float)</td>
</tr>
<tr>
<td>range: [12.385596, 61.129696] units: 1.000e-07</td>
</tr>
<tr>
<td>unique values: 9,941 missing .: 0/10,351</td>
</tr>
<tr>
<td>mean: 25.5376 std. dev: 4.91497</td>
</tr>
<tr>
<td>percentiles: 10% 25% 50% 75% 90% 20.1037 22.142 24.8181 28.0267 31.7259</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age: age in years</th>
</tr>
</thead>
<tbody>
<tr>
<td>type: numeric (byte)</td>
</tr>
<tr>
<td>range: [20, 74] units: 1</td>
</tr>
<tr>
<td>unique values: 55 missing .: 0/10,351</td>
</tr>
<tr>
<td>mean: 47.5797 std. dev: 17.2148</td>
</tr>
<tr>
<td>percentiles: 10% 25% 50% 75% 90% 24 31 49 63 69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>female: 1=female, 0=male</th>
</tr>
</thead>
<tbody>
<tr>
<td>type: numeric (byte)</td>
</tr>
<tr>
<td>range: [0, 1] units: 1</td>
</tr>
<tr>
<td>unique values: 2 missing .: 0/10,351</td>
</tr>
<tr>
<td>tabulation: Freq. Value</td>
</tr>
<tr>
<td>4,915 0</td>
</tr>
<tr>
<td>5,436 1</td>
</tr>
</tbody>
</table>

- Now we can fit the model.

  \texttt{regress bmi age female}

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 10,351</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7330.98402</td>
<td>2</td>
<td>3665.49201</td>
<td>F(2, 10348) = 156.29</td>
</tr>
<tr>
<td>Prob &gt; F</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2 Estimation

2.1 Including Categorical Variables

Working with Categorical Variables

- We would now like to include \texttt{region} in the model, let's take a look at this variable

\texttt{\texttt{. codebook region}}

\textbf{\begin{tabular}{l|l|l} 
\texttt{region}& \texttt{Freq.}& \texttt{Numeric Label} \\
\hline
1& 2,096 & NE \\
2& 2,774 & MW \\
3& 2,853 & S \\
4& 2,628 & W \\
\end{tabular}}

- It cannot simply be added to the list of covariates because it has 4 categories

- To include a categorical variable, put an \texttt{i.} in front of its name—this declares the variable to be a categorical variable, or in Stataese, a \textit{factor variable}

- For example, to add \texttt{region} to our model we use

\texttt{\texttt{. regress bmi age i.female i.region}}

\textbf{\begin{tabular}{l|l|l|l|l|l|l} 
\texttt{bmi}& \texttt{Coef.}& \texttt{Std. Err.}& \texttt{t}& \texttt{P>|t|}& \texttt{[95\% Conf. Interval]} \\
\hline
\texttt{age}& .0488667& .0027653& 17.67& 0.000& .0434462& .0542872 \\
\texttt{female}& .0380616& .0953249& 0.40& 0.690& -.1487936& .2249168 \\
\texttt{_cons}& 23.19255& .1482223& 156.47& 0.000& 22.90201& 23.48309 \\
\end{tabular}}
Niceities

- Value labels associated with factor variables are displayed in the regression table
- We can tell Stata to show the base categories for our factor variables
  
  . set showbaselevels on

Factor Notation as Operators

- The \(i.\) operator can be applied to many variables at once:

  . regress bmi age i.(female region)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 10,351</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7390.19781</td>
<td>5</td>
<td>1478.03956</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>242633.964</td>
<td>10,345</td>
<td>23.4542256</td>
<td>R-squared = 0.0296</td>
</tr>
<tr>
<td>Total</td>
<td>250024.162</td>
<td>10,350</td>
<td>24.1569239</td>
<td>Root MSE = 4.843</td>
</tr>
</tbody>
</table>

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

  | bmi | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
  |-----|-------|-----------|---|-----|----------------|
  | age | 0.0488851 | 0.0027674 | 17.66 | 0.000 | 0.0434605 | 0.0543097 |
  | female | | | | |
  | 0 | 0 (base) | | | |
  | 1 | 0.0372717 | 0.0953357 | 0.39 | 0.696 | -0.1496047 | 0.2241481 |
  | region | | | | |
  | NE | 0 (base) | | | |
  | MW | 0.0064779 | 0.1402121 | 0.05 | 0.963 | -0.268365 | 0.2813207 |
  | S | 0.0387957 | 0.1393383 | 0.28 | 0.781 | -0.2343342 | 0.3119256 |
  | W | -0.1537648 | 0.1418286 | -1.08 | 0.278 | -0.4317762 | 0.1242466 |
  | _cons | 23.2187 | 0.1760452 | 131.89 | 0.000 | 22.87362 | 23.56378 |

- In other words, it understands the distributive property
  - This is useful when using variable ranges, for example
- For the curious, factor variable notation works with wildcards
  - If there were many variables starting with \(u\), then \(i.u*\) would include them all as factor variables
Using Different Base Categories

- By default, the smallest-valued category is the base category
- This can be overridden within commands
  - \( b\# \) specifies the value \( \# \) as the base
  - \( b(\#) \) specifies the \( \# \)'th largest value as the base
  - \( b(first) \) specifies the smallest value as the base
  - \( b(last) \) specifies the largest value as the base
  - \( b(freq) \) specifies the most prevalent value as the base
  - \( bn \) specifies there should be no base

Playing with the Base

- We can use \( region=3 \) as the base class on the fly:
  ```
  . regress bmi age i.female b3.region
  ```
- We can use the most prevalent category as the base
  ```
  . regress bmi age i.female b(freq).region
  ```
- Factor variables can be distributed across many variables
  ```
  . regress bmi age b(freq).(female region)
  ```
- The base category can be omitted (with some care here)
  ```
  . regress bmi age i.female bn.region, noconstant
  ```
- We can also include a term for \( region=4 \) only
  ```
  . regress bmi age i.female 4.region
  ```

2.2 Including Interactions

Specifying Interactions

- Factor variables are also used for specifying interactions
  - This is where they really shine
- To include both main effects and interaction terms in a model, put \#\# between the variables
- To include only the interaction terms, put \# between the terms
Categorical by Categorical Interactions

• For example, to fit a model that includes main effects for age, female, and region, as well as the interaction of female and region:

\[ \text{regress \ bmi \ age \ female##region} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 10,351</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7559.19099</td>
<td>8</td>
<td>944.898874</td>
<td>F(8, 10342) = 40.30</td>
</tr>
<tr>
<td>Residual</td>
<td>242464.971</td>
<td>10,342</td>
<td>23.4446888</td>
<td>R-squared = 0.0302</td>
</tr>
<tr>
<td>Total</td>
<td>250024.162</td>
<td>10,350</td>
<td>24.1569239</td>
<td>Root MSE = 4.842</td>
</tr>
</tbody>
</table>

| bmi | Coef. | Std. Err. | t     | P>|t|   [95% Conf. Interval] |
|-----|-------|-----------|-------|-------|------------------------|
| age | .0488087 | .0027671 | 17.64 | 0.000 | .0433846 .0542328 |
| female | 0 | 0 (base) |     |       |                        |
| 1 | -.2939562 | .2116093 | -1.39 | 0.165 | -.7087514 .1208389 |
| region | NE | 0 (base) |     |       |                        |
| MW | -.1420836 | .2023593 | -0.70 | 0.483 | -.538747 .2545798 |
| S | -.3347762 | .2015721 | -1.66 | 0.097 | -.7298965 .0603441 |
| W | -.2694841 | .204234 | -1.32 | 0.187 | -.6698222 .1308541 |
| female#region | 1#MW | .2897474 | .280525 | 1.03 | 0.302 | -.2601358 .8396306 |
| 1#S | .7124639 | .2789251 | 2.55 | 0.011 | .1657169 1.259211 |
| 1#W | .2266557 | .2837887 | 0.80 | 0.424 | -.3296251 .7829365 |
| _cons | 23.39271 | .2013939 | 116.15 | 0.000 | 22.99793 23.78748 |

• Variables involved in interactions are assumed to be categorical, so no i. is needed

• To see all the omitted terms we can add the allbaselevels option:

\[ \text{regress \ bmi \ age \ female##region, allbaselevels} \]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 10,351</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7559.19099</td>
<td>8</td>
<td>944.898874</td>
<td>F(8, 10342) = 40.30</td>
</tr>
<tr>
<td>Residual</td>
<td>242464.971</td>
<td>10,342</td>
<td>23.4446888</td>
<td>R-squared = 0.0302</td>
</tr>
<tr>
<td>Total</td>
<td>250024.162</td>
<td>10,350</td>
<td>24.1569239</td>
<td>Root MSE = 4.842</td>
</tr>
</tbody>
</table>

| bmi | Coef. | Std. Err. | t     | P>|t|   [95% Conf. Interval] |
|-----|-------|-----------|-------|-------|------------------------|
| age | .0488087 | .0027671 | 17.64 | 0.000 | .0433846 .0542328 |
| female | 0 | 0 (base) |     |       |                        |
| 1 | -.2939562 | .2116093 | -1.39 | 0.165 | -.7087514 .1208389 |
| region | NE | 0 (base) |     |       |                        |
| MW | -.1420836 | .2023593 | -0.70 | 0.483 | -.538747 .2545798 |
| S | -.3347762 | .2015721 | -1.66 | 0.097 | -.7298965 .0603441 |
| W | -.2694841 | .204234 | -1.32 | 0.187 | -.6698222 .1308541 |
| female#region | 1#MW | .2897474 | .280525 | 1.03 | 0.302 | -.2601358 .8396306 |
| 1#S | .7124639 | .2789251 | 2.55 | 0.011 | .1657169 1.259211 |
| 1#W | .2266557 | .2837887 | 0.80 | 0.424 | -.3296251 .7829365 |
| _cons | 23.39271 | .2013939 | 116.15 | 0.000 | 22.99793 23.78748 |
Categorical by Continuous Interactions

- To include continuous variables in interactions use `c.` to specify that a variable is continuous
- Otherwise it will be assumed to be categorical

Here is our model with an interaction between age and region:

```
.regress bmi c.age##region i.female
```

```
Source | SS df MS Number of obs = 10,351
-------------+---------------------------------- F(8, 10342) = 40.35
Model | 7568.54189 8 946.067737 Prob > F = 0.0000
Residual | 242455.62 10,342 23.4437846 R-squared = 0.0303
-------------+---------------------------------- Adj R-squared = 0.0295
Total | 250024.162 10,350 24.1569239 Root MSE = 4.8419

bmi | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
age | .0607829 .0062164 9.78 0.000 .0485975 .0729683
region | NE | 0 (base)
| MW | .3951518 .4106204 0.96 0.336 -.4097436 1.200047
| S | 1.051668 .4181868 2.51 0.012 .2319407 1.871395
| W | .5921285 .4181932 1.42 0.157 -.227611 1.411868
region#c.age | MW | -.0080245 .0081638 -0.98 0.326 -.0240272 .0079782
| S | -.0211109 .008219 -2.57 0.010 -.0372217 -.0050002
| W | -.0155977 .0082261 -1.90 0.058 -.0317225 .000527
female | 0 | 0 (base)
| 1 | .038259 .0953259 0.40 0.688 -.1485982 .2251161
_cons | 22.64929 .3193208 70.93 0.000 22.02336 23.27522
```

Linear Regression Models with Interaction/Moderation © StataCorp LLC
Continuous by Continuous Interactions

- Prefix both variables in the interaction with \( c \) to fit models with continuous by continuous variable interactions.

- For example, we can interact age with serum vitamin c levels (\texttt{vitaminc})

\[
\texttt{. regress bmi c.age##c.vitaminc i.female i.region}
\]

\[
\begin{array}{llllll}
\text{bmi} & \text{Coeff.} & \text{Std. Err.} & \text{t} & \text{P>|t|} & \text{[95% Conf. Interval]} \\
\hline
\text{age} & .0220407 & .0059366 & 3.71 & 0.000 & .0104038 \text{ - } .0336777 \\
\text{vitaminc} & -2.331426 & .2717928 & -8.58 & 0.000 & -2.864194 \text{ - } -1.798657 \\
\text{c.age#c.vitaminc} & .029107 & .0050017 & 5.82 & 0.000 & .0193026 \text{ - } .0389115 \\
\text{female} & & & & & \\
0 & 0 (base) & & & & \\
1 & .1858965 & .0982311 & 1.89 & 0.058 & -.0066564 \text{ - } .3784494 \\
\text{region} & & & & & \\
NE & 0 (base) & & & & \\
MW & -.0936871 & .1412331 & -0.66 & 0.507 & -.3705326 \text{ - } .1831584 \\
S & -.2137082 & .1431247 & -1.49 & 0.135 & -.4942615 \text{ - } .0668451 \\
W & -.1626738 & .1430181 & -1.14 & 0.255 & -.4430182 \text{ - } .1176706 \\
\_cons & 25.45695 & .3293507 & 77.29 & 0.000 & 24.81136 \text{ - } 26.10255
\end{array}
\]

- To include polynomial terms, interact a variable with itself.

- For example, a model that includes both \texttt{age} and \texttt{age}^2

\[
\texttt{. regress bmi c.age##c.age i.female i.region}
\]

\[
\begin{array}{llllllll}
\text{bmi} & \text{Coeff.} & \text{Std. Err.} & \text{t} & \text{P>|t|} & \text{[95% Conf. Interval]} \\
\hline
\text{age} & .2731368 & .0203077 & 13.45 & 0.000 & .2333297 \text{ - } .3129439 \\
\text{c.age#c.age} & -.0024099 & .0002162 & -11.15 & 0.000 & -.0028337 \text{ - } -.0019861 \\
\text{female} & & & & & \\
0 & 0 (base) & & & & \\
1 & .0462855 & .0947764 & 0.49 & 0.625 & -.1394945 \text{ - } .2320656 \\
\text{region} & & & & & \\
NE & 0 (base) & & & & \\
\end{array}
\]
Higher Order Interactions

- Factor variable syntax can be used to specify higher order interactions
- If the interactions are specified using `##` all lower order terms are included
- For example, here we fit a model for `bmi` using a model that includes the three-way interaction of continuous variables `age` and `vitaminc` and categorical variable `female`

```
. regress bmi c.age##c.vitaminc##female
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 9,973</th>
</tr>
</thead>
<tbody>
<tr>
<td>F(7, 9965) = 76.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>12294.4386</td>
<td>7</td>
<td>1756.34837</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>228483.691</td>
<td>9,965</td>
<td>22.9286193</td>
<td>R-squared = 0.0511</td>
</tr>
<tr>
<td>Total</td>
<td>240778.13</td>
<td>9,972</td>
<td>24.1454201</td>
<td>Adj R-squared = 0.0504</td>
</tr>
<tr>
<td>F(7, 9965) = 76.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| bmi | Coef. Std. Err. t P>|t| 95% Conf. Interval |
|-----|------------------|-----|------|-------------------|
| age | -.0038595 .0084263 -0.46 0.647 -.0203767 .0126578 |
| vitaminc | -2.008713 .4231851 -4.75 0.000 -2.838241 -1.179185 |
| c.age#c.vitaminc | .0313728 .0078481 4.00 0.000 .0159889 .0467566 |
| female | 0 | 0 (base) |
| 1 | -2.098183 .6208318 -3.38 0.001 -3.315138 -.8812268 |
| female#c.age | 1 | .0646392 .0119517 5.41 0.000 .0412115 .0880668 |
| female#c.vitaminc | 1 | .0314475 .5539279 0.06 0.955 -1.054363 1.117258 |
| female#c.age#c.vitaminc | 1 | -.0166002 .0102645 -1.62 0.106 -.0367206 .0035203 |
| _cons | 26.16464 .4416624 59.24 0.000 25.29889 27.03039 |

Some Factor Variable Notes

- If you plan to look at marginal effects of any kind, it is best to
  - Explicitly mark all categorical variables with `i`
  - Specify all interactions using `#` or `##`
3 Postestimation

3.1 About Postestimation

Introduction to Postestimation

- In Stata jargon, postestimation commands are commands that can be run after a model is fit, for example
  - Predictions
  - Additional hypothesis tests
  - Checks of assumptions

- We’ll explore postestimation tools that can be used to help interpret the results of models that include interactions
- The usefulness of specific tools will depend on the types of hypotheses you wish to examine

3.2 Investigating Categorical by Categorical Interactions

Estimating a Model

- Let’s begin by running a model with main effects for age, female, and region, and the interaction of female and region

  . regress bmi age female##region

  Source | SS        df       MS              Number of obs = 10,351
  ------ | --------- | -------- | ------------------ -------------
  Model  | 7559.1910 | 8        944.898874 Prob > F    = 0.0000
  Residual| 242464.97 | 10,342   23.4446888 R-squared = 0.0302
  Total  | 250024.16 | 10,350   24.1569239 Root MSE = 4.842

  -------------------------------------------------------------------------------
  bmi | Coef.   Std. Err.     t    P>|t|     [95% Conf. Interval]
  --- | ------- | -------- | ------ | -------- |------------------|---------------
  age | .0488087 .0027671 17.64 0.000 .0433846 .0542328
  female | 0 (base)
  1 | -.2939562 .2116093 -1.39 0.165 -.7087514 .1208389
  region | NE | 0 (base)
  MW | -.1420836 .2023593 -.070 0.483 -.538747 .2545798
  S | -.3347762 .2015721 -1.66 0.097 -.7298965 .0603441
  W | -.2694841 .204234 -1.32 0.187 -.6698222 .1308541
  female#region | 1#MW | .2897474 .280525 1.03 0.302 -.2601358 .8396306
  1#S | .7124639 .2789251 2.55 0.011 .1657169 1.259211

Linear Regression Models with Interaction/Moderation © StataCorp LLC
How might we begin?

- Perform joint tests of coefficients
- Estimate and test hypotheses about group differences

Finding the Coefficient Names

- Some postestimation commands require that you know the names used to store the coefficients
- To see these names we can replay the model and showing the *coefficient legend*

```
. regress, coeflegend
```

```
Source | SS       df       MS              Number of obs = 10,351
--------+------------------------------------- F(8, 10342) = 40.30
Model   | 7559.19099     8  944.898874        Prob > F = 0.0000
Residual| 242464.971 10,342 23.4446888        R-squared = 0.0302
        |                      Adj R-squared = 0.0295
Total   | 250024.162 10,350 24.1569239        Root MSE = 4.842
--------+-------------------------------------
bmi     | Coef. Legend
--------+-------------------------------------
age    | .0488087 _b[age]
female | 0    | 0 _b[0b.female]
       | 1    | -.2939562 _b[1.female]
region | NE   | 0 _b[1b.region]
       | MW   | -.1420836 _b[2.region]
       | S    | -.3347762 _b[3.region]
       | W    | -.2694841 _b[4.region]
female#region | 1#MW | .2897474 _b[1.female#2.region]
            | 1#S  | .7124639 _b[1.female#3.region]
            | 1#W  | .2266557 _b[1.female#4.region]
_cons   | 23.39271 _b[_cons]
--------+-------------------------------------
```

- From here, we can see the full specification of the factor levels:
  - `_b[2.region]` corresponds to `region=2` which is “MW” or midwest
  - `_b[3.region]` corresponds to `region=3` which is “S” or south
- We can also see the terms for the interaction:
  - `_b[1.female#2.region]` corresponds to the term for the interaction of `region=2` and `female=1`
  - `_b[1.female#3.region]` corresponds to the term for the interaction of `region=3` and `female=1`
Joint Tests

- The test command performs a Wald test of the specified null hypothesis
  - The default test is that the listed terms are equal to 0
- test takes a list of terms, which may be variable names, but can also be terms associated with factor variables
- To perform a joint test of the null hypothesis that the coefficients for the levels of region are all equal to 0

  . test 2.region 3.region 4.region
  \( (1) \ 2.\text{region} = 0 \)
  \( (2) \ 3.\text{region} = 0 \)
  \( (3) \ 4.\text{region} = 0 \)

  \[ F(3, 10342) = 1.07 \]
  Prob > F = 0.3600

  - Since the model contains an interaction, this is a test of the effect of region when female=0

Testing Sets of Coefficients

- To test that all of the coefficients associated with the interaction of female and region we would need to give the full name of all the coefficients

  . test 1.\text{female}\#2.\text{region} 1.\text{female}\#3.\text{region} 1.\text{female}\#4.\text{region}

- testparm also performs Wald tests, but it accepts lists of variables, rather than coefficients in the model
- So we can perform joint tests with less typing, for example

  . testparm i.\text{region}\#i.\text{female}
  \( (1) \ 1.\text{female}\#2.\text{region} = 0 \)
  \( (2) \ 1.\text{female}\#3.\text{region} = 0 \)
  \( (3) \ 1.\text{female}\#4.\text{region} = 0 \)

  \[ F(3, 10342) = 2.40 \]
  Prob > F = 0.0656

An Alternative Test

- Likelihood ratio tests provide an alternative method of testing sets of coefficients
- To test the coefficients associated with the interaction of female and region we need to store our model results. The name is arbitrary, we’ll call them m1

  . estimates store m1

- Now we can rerun our model without region

  . regress bmi age i.\text{female} i.\text{region}
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 10,351</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7390.1978</td>
<td>5</td>
<td>1478.0396</td>
<td>F(5, 10345) = 63.02</td>
</tr>
<tr>
<td>Residual</td>
<td>242633.964</td>
<td>10,345</td>
<td>23.4542256</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>250024.162</td>
<td>10,350</td>
<td>24.1569239</td>
<td>Adj R-squared = 0.0291</td>
</tr>
</tbody>
</table>

| bmi | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----|-------|-----------|-------|-------|-----------------------|
| age | .0488851 | .0027674 | 17.66 | 0.000 | .0434605 .0543097 |
| female | 0 | 0 (base) | | | |
| 1 | .0372717 | .0953357 | 0.39 | 0.696 | -.1496047 .2241481 |
| region | 0 | 0 (base) | | | |
| NE | | | | | |
| MW | .0064779 | .1402121 | 0.05 | 0.963 | -.268365 .2813207 |
| S | .0387957 | .1393383 | 0.28 | 0.781 | -.2343342 .3119256 |
| W | -.1537648 | .1418286 | -1.08 | 0.278 | -.4317762 .1242466 |
| _cons | 23.2187 | .1760452 | 131.89 | 0.000 | 22.87362 23.56378 |

• If we were removing one of these variables entirely, we would want to add if e(sample) to makes sure the same sample, what Stata calls the estimation sample, is used for both models.

Likelihood Ratio Tests (Continued)

• Now we store the second set of estimates

  . estimates store m2

• And use the lrtest command to perform the likelihood ratio test

  . lrtest m1 m2

  Likelihood-ratio test LR chi2(3) = 7.21
  (Assumption: m2 nested in m1) Prob > chi2 = 0.0654

• We’ll restore the results from m1

  . estimates restore m1

  (results m1 are active now)

• Now it’s as if we just ran the model stored in m1

Tests of Differences

• test can also be used to the equality of coefficients

  . test 3.region#1.female = 4.region#1.female

( 1) 1.female#3.region - 1.female#4.region = 0

  F( 1, 10342) = 3.43
  Prob > F = 0.0640
• A likelihood ratio test can also be used; see help constraint for information on setting the necessary constraints.

• The lincom command can be used to calculate linear combinations of coefficients, along with standard errors, hypothesis tests, and confidence intervals.

• For example, to obtain the difference in coefficients

  . lincom 3.region#1.female - 4.region#1.female

  ( 1) 1.female#3.region - 1.female#4.region = 0

  ________________________________________________________________
  bmi | Coef. Std. Err. t P>|t| [95% Conf. Interval]
  -------------+--------------------------------------------------
  (1) | 0.4858082 0.2622654 1.85 0.064 -0.0282827 .9998991
  ________________________________________________________________

Contrasts

• The contrast command allows us to test a wide variety of comparisons across groups.

• For example comparing regions separately for men and women

  . contrast region#female, effects

  Contrasts of marginal linear predictions

  Margins : asbalanced

  ________________________________________________________________
   | df  F  P>F
  -------------+--------------------------------------------------
  region#female |         
  0 | 3 1.07 0.3600
  1 | 3 2.17 0.0890
  Joint | 6 1.62 0.1364
  Denominator | 10342
  ________________________________________________________________

  | Contrast Std. Err. t P>|t| [95% Conf. Interval]
  -------------+--------------------------------------------------
  region#female |         
  (MW vs base) 0 | -0.1420836 0.2023593 -0.70 0.483 -0.538747 .2545798
  (MW vs base) 1 | 0.1476637 0.1943419 0.76 0.447 -0.2332839 .5286114
  (S vs base) 0 | -0.3347762 0.2015721 -1.66 0.097 -0.7298965 .0603441
  (S vs base) 1 | 0.3778878 0.1927872 1.96 0.050 -0.0002125 .755588
  (W vs base) 0 | -0.2694841 0.204234 -1.32 0.187 -0.6698222 .1308541
  (W vs base) 1 | -0.0428284 0.1970381 -0.22 0.828 -0.4290612 .3434044

  ⋄ The @ symbol requests comparisons of the levels of region at each value of female.
  ⋄ The effects option requests that individual contrasts be displayed along with their standard errors, hypothesis tests, and confidence intervals.
Adjusting for Multiple Comparisons

- Use of contrast can result in a large number of hypothesis tests
- The `mcompare()` option can be used to adjust p-values and confidence intervals for multiple comparisons within factor variable terms
- The available methods are
  - `noadjust`
  - `bonferroni`
  - `sidak`
  - `scheffe`
- To apply Bonferroni's adjustment to our previous contrast
  
  ```
  . contrast region@female, effects mcompare(bonferroni)
  
  Contrasts of marginal linear predictions
  ```

  **Margins**: asbalanced

<table>
<thead>
<tr>
<th>df</th>
<th>F</th>
<th>P&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3 1.07</td>
<td>0.3600</td>
</tr>
<tr>
<td>1</td>
<td>3 2.17</td>
<td>0.0890</td>
</tr>
<tr>
<td>Joint</td>
<td>6 1.62</td>
<td>0.1364</td>
</tr>
</tbody>
</table>

  **Denominator**: 10342

  Note: Bonferroni-adjusted p-values are reported for tests on individual contrasts only.

<table>
<thead>
<tr>
<th>Number of Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>region@female</td>
</tr>
</tbody>
</table>

  | Contrast          | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
  |-------------------|-----------|------|------|----------------------|
  | region@female     |           |      |      |                      |
  | (MW vs base) 0    | 0  | - .1420836 | .2023593 | -0.70  | 1.000  | - .6760623 | .391895 |
  | (MW vs base) 1    | 0  | .1476637  | .1943419 | 0.76   | 1.000  | - .3651588 | .6604863 |
  | (S vs base) 0     | 0  | - .3347762 | .2015721 | -1.66  | 0.581  | - .8666776 | .1971252 |
  | (S vs base) 1     | 0  | .3778878  | .1927872 | 1.96   | 0.301  | - .1310325 | .886408  |
  | (W vs base) 0     | 0  | - .2694841 | .204234  | -1.32  | 1.000  | - .8084096 | .2694415 |
  | (W vs base) 1     | 0  | - .0428284 | .1970381 | -0.22  | 1.000  | - .5627657 | .4771089 |
Average Predicted Values

- We might want to explore predictions based on our model and data
- Predictions for individual observations can be made using the `predict` command, see `help predict`
- To find out about our model more generally, we may be more interested in average predicted values
  - Also known as predictive margins or recycled predictions
- To obtain the average predicted value of bmi
  ```
  . margins
  Predictive margins                  Number of obs  =  10,351
  Model VCE : OLS
  Expression : Linear prediction, predict()

  Delta-method                      Margin  Std. Err.  t  P>|t|    [95% Conf. Interval]
  _cons  25.5376  .0475917  536.60  0.000   25.44431   25.63089

  Predictions at Specified Values of Factor Variables

- Stata calls the list of variables that follow the `margins` command the `marginslist`
  - To appear in the `marginslist` a variable must have been specified as factor variable in the model
- To obtain the average predicted value of bmi at different values of region
  ```
  . margins region
  Predictive margins                  Number of obs  =  10,351
  Model VCE : OLS
  Expression : Linear prediction, predict()

  Delta-method                      Margin  Std. Err.  t  P>|t|    [95% Conf. Interval]
  region
   NE  25.56063  .1057882  241.62  0.000   25.35327   25.768
   MW  25.57071  .09198  278.00  0.000   25.39042   25.75101
    S  25.60002  .0906777  282.32  0.000   25.42227   25.77776
    W  25.41018  .0944557  269.02  0.000   25.22503   25.59533

  How were these values generated?
  1. Calculate the predicted value of bmi setting region=1 and using each case’s observed values of female and age
  2. Find the mean of the predicted values
  3. Repeat steps 1 and 2 for each value of region
Predicted Values with Multiple Factor Variables

- We can obtain margins for multiple variables

  . margins region female

  Predictive margins Number of obs = 10,351
  Model VCE : OLS

  Expression : Linear prediction, predict()

  +--------------------------------------------------+
  | Delta-method                                     |
  | Margin  Std. Err.      t     P>|t|   [95% Conf. Interval] |
  |--------------------------------------------------|
  | region  |                                   |
  | NE      | 25.56063   .1057882  241.62  0.000  25.35327  25.7682 |
  | MW      | 25.57071   .09198   278.00  0.000  25.39042  25.75101 |
  | S       | 25.60002   .0906777  282.32  0.000  25.42227  25.77776 |
  | W       | 25.41018   .0944557  269.02  0.000  25.22503  25.59533 |
  | female  |                                   |
  | 0       | 25.51624   .0690736  369.41  0.000  25.38084  25.65164 |
  | 1       | 25.55385   .0656788  389.07  0.000  25.42511  25.68259 |
  +--------------------------------------------------+

- Or we can obtain predicted values of bmi at each combination of region and female

  . margins region#female

  Predictive margins Number of obs = 10,351
  Model VCE : OLS

  Expression : Linear prediction, predict()

  +--------------------------------------------------+
  | Delta-method                                     |
  | Margin  Std. Err.      t     P>|t|   [95% Conf. Interval] |
  |--------------------------------------------------|
  | region#female                                   |
  | NE#0    | 25.71501   .1517587  169.45  0.000  25.41753  26.01248 |
  | NE#1    | 25.42105   .1474742  172.38  0.000  25.13197  25.71013 |
  | MW#0    | 25.57292   .1338383  191.07  0.000  25.31058  25.83527 |
  | MW#1    | 25.56872   .1265618  202.03  0.000  25.32063  25.81686 |
  | S#0     | 25.38023   .1326702  191.30  0.000  25.12017  25.64029 |
  | S#1     | 25.79874   .1241829  207.75  0.000  25.55532  26.04216 |
  | W#0     | 25.44552   .1366851  186.16  0.000  25.17759  25.71345 |
  | W#1     | 25.37822   .1306734  194.21  0.000  25.12208  25.63437 |
  +--------------------------------------------------+

- We might prefer to graph these results, we can do so using the marginsplot command

Graphing Predicted Values

  . marginsplot
Predictive Margins of region#female with 95% CIs

If our model did not include a region by female interaction, the lines would be parallel.

Predicted Values for Specific Groups

- When we specify the variables in the `marginslist` Stata calculates predicted values treating each case as though it belonged to each group.

- The `over()` option allows us to obtain predictions separately for each group, for example:
  ```stata
  . margins, over(female)
  ```

```
Predictive margins

  Number of obs = 10,351
  Model VCE : OLS
  Expression : Linear prediction, predict()
  over : female

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>25.50999</td>
<td>.0690654</td>
<td>369.36</td>
<td>0.000</td>
<td>25.37461 25.64538</td>
</tr>
<tr>
<td>1</td>
<td>25.56256</td>
<td>.0656723</td>
<td>389.24</td>
<td>0.000</td>
<td>25.43383 25.69129</td>
</tr>
</tbody>
</table>
```

- This time the table shows
  - The average predicted value of `bmi` for cases where `female=0` using each case’s observed values of `age` and `region`
  - The average predicted value of `bmi` for cases where `female=1` using each case’s observed values of `age` and `region`

- This can be useful when we want to compare groups
### 3.3 Investigating Categorical by Continuous Interactions

**A Categorical by Continuous Interaction**

- For this set of examples, we'll fit a model that includes an interaction between the continuous variable `age` and the categorical variable `region`.

```
.regress bmi c.age##region i.female
```

```
            Source |  SS     df    MS       Number of obs  =  10,351  
---------------------------------------------------------------------
            Model | 7568.54189     8   946.06774 7 Prob > F    =  0.0000  
            Residual | 242455.62   10,342  23.4437846   R-squared  =  0.0303  
---------------------------------------------------------------------
            Total | 250024.162   10,350  24.1569239   Root MSE  =  4.8419
```

```
            bmi | Coef.  Std. Err.  t   P>|t|  [95% Conf. Interval]
---------------------------------------------------------------------
            age |    .0607829   .0062164   9.78 0.000    .0485975   .0729683  
            region |           
            NE | 0   (base)  
            MW |    .3951518   .4106204   0.96 0.336  -.4097436   1.200047  
            S |    1.051668   .4181868   2.51 0.012    .2319407   1.871395  
            W |    .5921285   .4181932   1.42 0.157  -.2276110   1.411868  
            region#c.age |           
            MW |  -.0080245   .0081638  -.09 0.929  -.0240272   .0080782  
            S |  -.0211109   .0082190  -.25 0.799  -.0372217  -.0050002  
            W |  -.0155977   .0082261  -.19 0.847  -.0317225   .0005272  
            female |           
            0 |   0   (base)  
            1 |    .038259   .0953259   0.40 0.688  -.1485982   .2251161  
            _cons |    22.64929   .3193208  70.93 0.000    22.02336   23.27522  
```

- Let's take a look at how the coefficients are stored.

```
.regress, coeflegend
```

```
            Source |  SS     df    MS       Number of obs  =  10,351  
---------------------------------------------------------------------
            Model | 7568.54189     8   946.06774 7 Prob > F    =  0.0000  
            Residual | 242455.62   10,342  23.4437846   R-squared  =  0.0303  
---------------------------------------------------------------------
            Total | 250024.162   10,350  24.1569239   Root MSE  =  4.8419
```

```
            bmi | Coef.  Legend  
---------------------------------------------------------------------
            age |  .0607829   _b[age]  
            region |           
            NE | 0   _b[1b.region]  
            MW |  .3951518   _b[2.region]  
            S |  1.051668   _b[3.region]  
            W |  .5921285   _b[4.region]  
            region#c.age |           
```
test and testparm

- As before, we can test the null hypothesis that all of the coefficients associated with the interaction of age and region are equal to 0 using testparm
  
  . testparm c.age#i.region

  ( 1) 2.region#c.age = 0
  ( 2) 3.region#c.age = 0
  ( 3) 4.region#c.age = 0

  F( 3, 10342) = 2.54
  Prob > F = 0.0549

- We could also use lrtest

- We can test specific hypotheses about the slopes

  - For example we might want to test whether the slope of age is significantly different in the south (region=3) versus the west (region=4)
    
    . test 3.region#c.age = 4.region#c.age

    ( 1) 3.region#c.age - 4.region#c.age = 0

    F( 1, 10342) = 0.52
    Prob > F = 0.4689

Estimated Slopes

- We can use lincom to estimate the slope of age for the south (region=3)
  
  . lincom c.age + 3.region#c.age

  ( 1) age + 3.region#c.age = 0

- We can also use margins with the dydx() option to calculate the slope of age for each region
  
  . margins region, dydx(age)
Average marginal effects
Number of obs = 10,351
Model VCE : OLS
Expression : Linear prediction, predict()
dy/dx w.r.t. : age

| Delta-method | dy/dx Std. Err. t P>|t| [95% Conf. Interval] |
|-------------|-----------------|-----------------|-----------------|-----------------|
| age | region |
| NE | .0607829 .0062164 9.78 0.000 .0485975 .0729683 |
| MW | .0527584 .0052919 9.97 0.000 .0423853 .0631315 |
| S | .0396719 .0053765 7.38 0.000 .0291329 .0502109 |
| W | .0451852 .0053875 8.39 0.000 .0346246 .0557457 |

• The dydx() option calculates derivative of the predicted values with respect to the specified variable, also known as the marginal effect

Predictions at Specified Values

• To obtain margins at set values of continuous variables use the at() option

• For example, the predicted value of bmi at each level of region setting age=20

  . margins region, at(age=20) vsquish

  Predictive margins
  Number of obs = 10,351
  Model VCE : OLS
  Expression : Linear prediction, predict()
  at : age = 20

| Delta-method | Margin Std. Err. t P>|t| [95% Conf. Interval] |
|-------------|-----------------|-----------------|-----------------|-----------------|
| region |
| NE | 23.88504 .2026955 117.84 0.000 23.48772 24.28236 |
| MW | 24.1197 .1678019 143.74 0.000 23.79078 24.44862 |
| S | 24.51449 .1766004 138.81 0.000 24.16832 24.86066 |
| W | 24.16521 .1772397 136.34 0.000 23.81779 24.51264 |

• The vsquish option reduces the vertical space in the output

• The at() option accepts numlists so we aren't restricted to a single value of age

  . margins region, at(age=(20(25)70)) vsquish

  Predictive margins
  Number of obs = 10,351
  Model VCE : OLS
  Expression : Linear prediction, predict()
  1._at : age = 20
  2._at : age = 45
  3._at : age = 70
| _at#region | Delta-method | Margin | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|------------|-------------|--------|-----------|---|-----|---------------------|
| 1#NE       | 23.88504    | .2026955 | 117.84    | 0.000 | 23.48772 | 24.28236  |
| 1#MW       | 24.1197     | .1678019 | 143.74    | 0.000 | 23.79078 | 24.44862  |
| 1#S        | 24.51449    | .1766004 | 138.81    | 0.000 | 24.16832 | 24.86066  |
| 1#W        | 24.16521    | .1772397 | 136.34    | 0.000 | 23.81779 | 24.51264  |
| 2#NE       | 25.40461    | .1072029 | 236.98    | 0.000 | 25.19447 | 25.61475  |
| 2#MW       | 25.43866    | .0922856 | 275.65    | 0.000 | 25.25776 | 25.61956  |
| 2#S        | 25.50629    | .0922593 | 276.46    | 0.000 | 25.32544 | 25.68713  |
| 2#W        | 25.29484    | .0956797 | 264.37    | 0.000 | 25.10729 | 25.48239  |
| 3#NE       | 26.92418    | .1737943 | 154.92    | 0.000 | 26.58351 | 27.26485  |
| 3#MW       | 26.75762    | .1545335 | 173.15    | 0.000 | 26.46473 | 27.06054  |
| 3#S        | 26.49809    | .1482211 | 178.77    | 0.000 | 26.20754 | 26.78863  |
| 3#W        | 26.42447    | .1522388 | 173.57    | 0.000 | 26.12605 | 26.72289  |

- The observed values of age are from 20 to 74

Graphing Predicted Values

- And we can plot the results
  
  . marginsplot

Supressing Confidence Intervals

- The confidence intervals can make the graph appear messy; we can suppress them
  
  . marginsplot, noci
○ This is dangerous because it makes the predictions look more precise than they are

**Testing for Differences**

- We might want to perform tests of differences at different levels of the continuous variable
- To obtain tests of differences between levels of region at each level of age

```
. margins region, at(age=(20(10)70)) vsquish contrast
```

Contrasts of predictive margins
Model VCE : OLS
Expression : Linear prediction, predict()

```
  1._at : age = 20
  2._at : age = 30
  3._at : age = 40
  4._at : age = 50
  5._at : age = 60
  6._at : age = 70

<table>
<thead>
<tr>
<th>df</th>
<th>F</th>
<th>P&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.94</td>
<td>0.1200</td>
</tr>
<tr>
<td>2</td>
<td>1.59</td>
<td>0.1884</td>
</tr>
<tr>
<td>3</td>
<td>1.06</td>
<td>0.3642</td>
</tr>
<tr>
<td>4</td>
<td>0.93</td>
<td>0.4251</td>
</tr>
<tr>
<td>5</td>
<td>1.56</td>
<td>0.1974</td>
</tr>
<tr>
<td>6</td>
<td>2.05</td>
<td>0.1041</td>
</tr>
<tr>
<td>Joint</td>
<td>1.69</td>
<td>0.1193</td>
</tr>
</tbody>
</table>
```

Denominator | 10342

Linear Regression Models with Interaction/Moderation © StataCorp LLC
Predicted Values Over Groups

- As with `marginslist`, when we specify `at()` Stata calculates predicted values treating each case as though they belong to each group or combination of values.
- As before, we can use the `over()` option after models with categorical by continuous interactions.
- For example, to obtain predicted values for each region using the observed values of `female` and `age` in that region:

  ```stata
  . margins, over(region)
  Predictive margins
  Number of obs    = 10,351
  Model VCE : OLS
  Expression : Linear prediction, predict()
  over : region
  _________________________________________________________________
  | Delta-method
  | Margin Std. Err.       t     P>|t|     [95% Conf. Interval]
  |-----------------+------------------------------------
  region          |______________________________
  NE | 25.57535 .1057592 241.83 0.000 25.36804 25.78266
  MW | 25.51936 .0919307 277.59 0.000 25.33916 25.69956
  S  | 25.63317 .090649 282.77 0.000 25.45548 25.81086
  W  | 25.42299 .0944498 269.17 0.000 25.23785 25.60813
  _________________________________________________________________
```

3.4 Investigating Continuous by Continuous Interactions

A Continuous by Continuous Interaction

- For this example we'll use a similar model for `bmi` but we'll add a main effect of serum vitamin c (`vitaminc`), and an interaction between `age` and `vitaminc`.
- Before we fit the model, let's take a closer look at `vitaminc`:

  ```stata
  . summ vitaminc, detail
  serum vitamin C (mg/dL)
  _________________________________
  Percentiles Smallest
  1% .2 .1
  5% .3 .1
  10% .3 .1     Obs       9,973
  25% .6 .1     Sum of Wgt. 9,973
  50% 1.034814
  Largest        Std. Dev. .5813791
  75% 1.48     8.3      Variance .3380017
  90% 1.793     9.4      Skewness 4.539869
  99% 2.432     18.1     Kurtosis 108.2617
  _________________________________________________________________
  ○ The distribution has a long tail, but most observations are between .2 and 2.
- Now let's fit the model:

  ```stata
  . regress bmi c.age##c.vitaminc i.female i.region
  ```
Source | SS  df  MS
-------------+----------------------------------
Model | 10298.9223  7  1471.27461
Residual | 230479.207 9,965 23.1288718
-------------+----------------------------------
Total | 240778.13  9,972 24.1454201

| Number of obs = 9,973
| F(7, 9965) = 63.61
| Prob > F = 0.0000
| R-squared = 0.0428
| Adj R-squared = 0.0421
| Root MSE = 4.8092

| bmi | Coef.  Std. Err.  t  P>|t|  [95% Conf. Interval] |
|-----|-------|--------|-----|---------------------|
| age | 0.0220407 0.0059366 3.71 0.000 0.0104038 0.0336777 |
| vitaminc | -2.331426 0.2717928 -8.58 0.000 -2.864194 -1.798657 |
| c.age#c.vitaminc | 0.029107 0.0050017 5.82 0.000 0.0193026 0.0389115 |
| female |       |       |     |                     |
| 0 | 0 (base) |
| 1 | .1858965 0.0982311 1.89 0.058 -.0066564 .3784494 |
| region |       |       |     |                     |
| NE | 0 (base) |
| MW | -.0936871 .1412331 -0.66 0.507 -.3705326 .1831584 |
| S | -.2137082 .1431247 -1.49 0.135 -.4942615 .0668451 |
| W | -.1626738 .1430181 -1.14 0.255 -.4430182 .1176706 |
| _cons | 25.45695 .3293507 77.29 0.000 24.81136 26.10255 |

• We can replay the model using coeflegend
  . regress, coeflegend

Estimating Slopes

• We can use lincom to calculate the slope for vitaminc when age=49 (it’s median)

  . lincom vitaminc + c.vitaminc#c.age*49

  ( 1) vitaminc + 49*c.age#c.vitaminc = 0

| bmi | Coef.  Std. Err.  t  P>|t|  [95% Conf. Interval] |
|-----|-------|--------|-----|---------------------|
| (1) | -.9051806 .0870624 -10.40 0.000 -1.075841 -.7345206 |

• We could also calculate the slope of age when vitaminc=1 (it’s median)

  . lincom age + c.vitaminc#c.age*1

  ( 1) age + c.age#c.vitaminc = 0

| bmi | Coef.  Std. Err.  t  P>|t|  [95% Conf. Interval] |
|-----|-------|--------|-----|---------------------|
| (1) | .0511478 .0028212 18.13 0.000 .0456176 .0566779 |

• margins can produce estimates of the slopes for a range of values
. margins, dydx(vitaminc) at(age=(20(10)70)) vsquish

Average marginal effects Number of obs  =  9,973
Model VCE : OLS
Expression : Linear prediction, predict()
dy/dx w.r.t. : vitaminc
1._at : age  =  20
2._at : age  =  30
3._at : age  =  40
4._at : age  =  50
5._at : age  =  60
6._at : age  =  70

------------------------------------------------------------------------------
| Delta-method
| dy/dx Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
vitaminc _at  
1  | -1.749285  .1797317 -9.73 0.000  -2.101595  -1.396974
2  | -1.458214  .1379304 -10.57 0.000  -1.728586  -1.187843
3  | -1.167144  .1036802 -11.26 0.000  -1.370378  -0.9639098
4  | -0.8760735  .0864746 -10.13 0.000  -1.045581  -0.706569
5  | -0.5850031  .0959667  -6.10 0.000  -0.7731173  -0.396889
6  | -0.2939327  .126273  -2.33 0.020  -0.5414532  -0.0464122
------------------------------------------------------------------------------

Graphing Slopes

• We can graph the slopes of vitaminc across age
  . marginsplot, yline(0)

![Average Marginal Effects of vitaminc with 95% CIs](image-url)
Predicted Values

- Specifying multiple variables in the \texttt{at(\())\) option results in predictions at each combination of values.

\begin{verbatim}
. margins , at(age=(20(25)70) vitaminc=(.2(.6)2)) vsquish
\end{verbatim}

<table>
<thead>
<tr>
<th>Predictive margins</th>
<th>Number of obs = 9,973</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model VCE : OLS</td>
<td></td>
</tr>
<tr>
<td>Expression : Linear prediction, predict()</td>
<td></td>
</tr>
</tbody>
</table>

\begin{verbatim}
1._at : age = 20  vitaminc = .2
2._at : age = 20  vitaminc = .8
3._at : age = 20  vitaminc = 1.4
4._at : age = 20  vitaminc = 2
5._at : age = 45  vitaminc = .2
6._at : age = 45  vitaminc = .8
7._at : age = 45  vitaminc = 1.4
8._at : age = 45  vitaminc = 2
9._at : age = 70  vitaminc = .2
10._at : age = 70  vitaminc = .8
11._at : age = 70  vitaminc = 1.4
12._at : age = 70  vitaminc = 2
\end{verbatim}

\begin{verbatim}
| Delta-method | Marginal | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|--------------|----------|-----------|---|--------|----------------------|
| at           |          |           |   |        |                      |
| 1             | 25.52113 | 0.1698638 | 150.24 | 0.000 | 25.18816 | 25.8541 |
| 2             | 24.47156 | 0.097744  | 250.36 | 0.000 | 24.27996 | 24.66316 |
| 3             | 23.42199 | 0.1162436 | 201.49 | 0.000 | 23.19413 | 23.64985 |
| 4             | 22.37242 | 0.2018162 | 110.86 | 0.000 | 21.97682 | 22.76802 |
| 5             | 26.21768 | 0.0891689 | 294.02 | 0.000 | 26.04289 | 26.39247 |
| 6             | 25.60472 | 0.0525344 | 487.39 | 0.000 | 25.50174 | 25.70769 |
| 7             | 24.99175 | 0.0606647 | 411.97 | 0.000 | 24.87284 | 25.11067 |
| 8             | 24.37879 | 0.1034993 | 235.55 | 0.000 | 24.17591 | 24.58167 |
| 9             | 26.91423 | 0.1388456 | 193.84 | 0.000 | 26.64207 | 27.1864 |
| 10            | 26.73788 | 0.0879343 | 304.07 | 0.000 | 26.56551 | 26.91024 |
| 11            | 26.56152 | 0.0875619 | 303.35 | 0.000 | 26.38988 | 26.73315 |
| 12            | 26.38516 | 0.1381377 | 191.01 | 0.000 | 26.11438 | 26.65593 |
\end{verbatim}

\begin{verbatim}
. marginsplot
\end{verbatim}
Changing the X-axis Variable

- We can select which variable appears on the x-axis using the `xdimension()` option.
  
  `. marginsplot, xdimension(vitaminc)`

Models with Polynomial Terms

- We’ll start by fitting a model that includes `age` and `age^2`
  
  `. regress bmi c.age##c.age i.female i.region`
Graphs can be particularly useful in understanding models with polynomial terms.

Here we predict values of `bmi` at different values of `age`.

```
margins, at(age=(20(10)70)) vsquish
```

Graphing Predicted Values

```
. margins, at(age=(20(10)70)) vsquish
```

And graph the predictions.
Slopes

- We can also obtain estimates of the slope of age across its range
- To do so we’ll include age in both the dyed() and at() options

```
. margins, dydx(age) at(age=(20(10)70)) vsquish
```

Average marginal effects

Number of obs = 10,351
Model VCE : OLS
Expression : Linear prediction, predict()
dy/dx w.r.t. : age
1._at : age = 20
2._at : age = 30
3._at : age = 40
4._at : age = 50
5._at : age = 60
6._at : age = 70

<table>
<thead>
<tr>
<th>Delta-method</th>
<th>dy/dx Std. Err. t P&gt;</th>
<th>t</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_at</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.1767405</td>
<td>0.0117968</td>
<td>14.98</td>
</tr>
<tr>
<td>2</td>
<td>0.1285424</td>
<td>0.0076583</td>
<td>16.78</td>
</tr>
<tr>
<td>3</td>
<td>0.0803442</td>
<td>0.0039415</td>
<td>20.38</td>
</tr>
<tr>
<td>4</td>
<td>0.0321460</td>
<td>0.0031343</td>
<td>10.26</td>
</tr>
<tr>
<td>5</td>
<td>-0.0160521</td>
<td>0.0064432</td>
<td>-2.49</td>
</tr>
<tr>
<td>6</td>
<td>-0.0642503</td>
<td>0.010517</td>
<td>-6.11</td>
</tr>
</tbody>
</table>
Adding a Cubic Term

- The same process can be used with higher order polynomials, here we add a cubic term for age.

```
.regress bmi c.age##c.age##c.age i.female i.region
```

```
Source | SS        | df  | MS            | Number of obs = 10,351
-------------+---------------------------------- F(7, 10343) = 64.27
Model       | 10422.3157 | 7   | 1488.90224    | Prob > F     = 0.0000
Residual    | 239601.846 | 10,343 | 23.1656044  | R-squared    = 0.0417
-------------+---------------------------------- Adj R-squared = 0.0410
Total       | 250024.162 | 10,350 | 24.1569239  | Root MSE     = 4.8131
-------------+----------------------------------

bmi | Coef.  Std. Err.  t  P>|t|  [95% Conf. Interval]
------------------+--------------------------------------------------------------
age | .5056311  .0927387  5.45  0.000  .3238453  .6874169
| c.age##c.age | -.0077683  .0020967  -3.70  0.000  -.0118782  -.0036583
| c.age##c.age##c.age | .0000383  .0000149  2.57  0.010  9.07e-06  .0000675
female | 0 | 0 (base)
| 1 | .0449127  .0947522  0.47  0.636  -.1408201  .2306454
region | NE | 0 (base)
| MW | .0274302  .1393783  0.20  0.844  -.2457782  .3060386
| S  | .025305  .1384883  0.18  0.855  -.2461589  .2967689
| W | -.1172832  .1410312  -0.83  0.406  -.3937317  .1591653
_cons | 15.6426  1.268785  12.33  0.000  13.15554  18.12967
```

- As before we can predict slopes at specified values of age.

```
.margins, dydx(age) at(age=(20(10)70)) vsquish
```

```
Average marginal effects Number of obs = 10,351
Model VCE : OLS
Expression : Linear prediction, predict()
dy/dx w.r.t. : age
1._at : age = 20
2._at : age = 30
3._at : age = 40
4._at : age = 50
5._at : age = 60
6._at : age = 70

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
</tr>
</thead>
</table>
|                  | dy/dx  Std. Err.  t  P>|t|  [95% Conf. Interval]
| age              |--------------------------------|
|      _at |                      |
| 1 | .2408252  .0275901  8.73  0.000  .1867432  .2949071
| 2 | .1428662  .0094709  15.08  0.000  .1243014  .161431
| 3 | .06787  .0062529  10.85  0.000  .056513  .0801269
| 4 | .0158363  .0070791  2.24  0.025  .0019598  .0297128
```
• Or predict bmi at different values of age

\[
\text{margins, at(age=(20(9)74)) vsquish}
\]

Predictive margins
Number of obs = 10,351
Model VCE : OLS

<table>
<thead>
<tr>
<th>Expression : Linear prediction, predict()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1._at : age = 20</td>
</tr>
<tr>
<td>2._at : age = 29</td>
</tr>
<tr>
<td>3._at : age = 38</td>
</tr>
<tr>
<td>4._at : age = 47</td>
</tr>
<tr>
<td>5._at : age = 56</td>
</tr>
<tr>
<td>6._at : age = 65</td>
</tr>
<tr>
<td>7._at : age = 74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Margin Std. Err. t P&gt;</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

• Here, we get predictions across the full range of ages in the dataset (i.e. 20-74)

Graphing the Cubic Term

• And we can easily graph this as well

\[
\text{marginsplot}
\]
4 Conclusion

4.1 Graphing Extras

Adding Additional Plots

- We can add other types of twoway plots to the plots drawn by marginsplots
  - Continuing with our cubic example
- The addplot option allows us to add additional plots to our marginsplots
- We do want to be careful about the order in which graphs are drawn, we usually want the most dense graphs, for example individual data points, drawn first
  - Specifying addplot(..., below) draws the added plot below the marginsplot

Adding Observed Data

```
. marginsplot, addplot(scatter bmi age, below ///
    legend(order(3 "Observed Values" 2 "Predictions"))) ///
    xlabel(20(9)74))
```
Changing the Plot Type

- We can change the plots drawn by `marginsplot` to another `twoway` plot type
  - See `help twoway` for a list
- The `recast()` option changes the plot for the predictions
  - `recastci()` changes how the CIs are plotted
- Let’s run a simple model to demonstrate

```
. regress bmi i.region
. margins region
```

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs</th>
<th>F(3, 10347)</th>
<th>Prob &gt; F</th>
<th>Adj R-squared</th>
<th>Root MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>64.491028</td>
<td>3</td>
<td>21.4970093</td>
<td>10,351</td>
<td>0.89</td>
<td>0.4455</td>
<td>-0.0003</td>
<td>4.915</td>
</tr>
<tr>
<td>Residual</td>
<td>249959.671</td>
<td>10,347</td>
<td>24.1576951</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>250024.162</td>
<td>10,350</td>
<td>24.1569239</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| bmi     | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|---------|-------|-----------|-------|------|----------------------|
| region  |       |           |       |      |                      |
| NE      | 0 (base) |          |       |      |                      |
| MW      | -.055989 | .1422471  | -0.39 | 0.694 | -.3348208            | .2228428 |
| S       | .0578207 | .1413969  | 0.41  | 0.683 | -.2193446            | .334986  |
| W       | -.1523645 | .1439376  | -1.06 | 0.290 | -.4345101            | .1297811 |
| _cons   | 25.57535 | .1073574  | 238.23 | 0.000 | 25.36491            | 25.78579 |
```
Adjusted predictions

Number of obs = 10,351
Model VCE : OLS
Expression : Linear prediction, predict()

------------------------------------------------------------------------------
| Delta-method
| Margin Std. Err. t P>|t|  [95% Conf. Interval]
------------------------------------------------------------------------------
region |
NE | 25.57535 .1073574 238.23 0.000 25.36491 25.78579
MW | 25.51936 .09332 273.46 0.000 25.33644 25.70229
S | 25.63317 .0920189 278.56 0.000 25.4528 25.81355
W | 25.42299 .0958771 265.16 0.000 25.23505 25.61092
------------------------------------------------------------------------------

Estimates as a Scatterplot

. marginsplot, recast(scatter)

Estimates as a Bar plot

. marginsplot, recast(bar) plotopts(barwidth(.9))
The `plotopt()` option allows you to specify options for the plots.

- `barwidth()` specifies the width of the bars in units of the x variable.

### 4.2 Conclusion

**Conclusion**

- We’ve seen how to fit models that include interactions.
- We’ve learned how to use Stata’s postestimation tools to explore the resulting models.
- We’ve learned how to graph predictions and how to modify those graphs.
Index