Week 13: Interpreting Model Results: Marginal Effects and the margins Command

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019
Why do we really need marginal effects?
Using derivatives to interpret more complex models
Derivatives definition, redux
Numerical derivatives using the estimated model, two ways
A collection of terms
  1. Average Marginal Effects (AME)
  2. Marginal Effect at the Mean (MEM)
  3. Marginal Effects at Representative values (MER)
Interactions
Examples
Why do we need marginal effects?

- In a simple linear model, say, $y = \beta_0 + \beta_1 age + \beta_2 male$, we could easily interpret the coefficients.

- It became more difficult when we had non-linear terms, for example:
  
  \[
  y = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 male
  \]

- We saw that using the **analytical derivative** made interpretation a lot easier:
  \[
  \frac{\partial E[y|age, male]}{\partial age} = \beta_1 + 2\beta_2 age
  \]

- No single effect of age, the effect depends on age; that is, one effect at age 20, another at 50, etc. Just plug in numbers for age in the above expression to get the effect at different ages.
Why do we need marginal effects?

- With interactions, it was even more complicated:
  
  \[ y = \beta_0 + \beta_1 \text{age} + \beta_2 \text{male} + \beta_3 \text{male} \times \text{age} \]

- But similar in the sense that the effect of age now depends on sex; or the other way around, the effect of sex depends on age

- With simple models, taking the derivative still helps with interpretation

- Centering also helps with parameter interpretation:

  \[ y = \beta_0 + \beta_1 (\text{age} - m) + \beta_2 \text{male} + \beta_3 \text{male} \times (\text{age} - m) \]

- If \( m \) is average age, then \( \beta_2 \) is \( E[Y] \) for males versus females of average age. What is \( \beta_1 \)?
Why do we need marginal effects?

- In the logistic model, things got complicated very quickly:
  \[ \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{male} \]

- We could present odds ratios: \( e^{\beta_1} \) and \( e^{\beta_2} \) and call it a day but we know that odds ratios can be misleading (and EVIL?)

- A simple example as a review (with no covariates): The probability of death in a control group is 0.40. The probability of death in the treatment group is 0.20

- The **odds ratio** is: \( \frac{0.2}{1-0.2} = \frac{0.4}{1-0.4} = 0.375 \). The treatment reduces the odds of death by a factor of 0.375. Or in reverse, the odds of death are 2.67 higher in the control group \( \left( \frac{1}{0.375} \right) \)

- But that’s not the **relative risk**, even though most people, including journalists, would interpret the odds ratio as a relative risk. The relative risk is \( \frac{0.2}{0.4} = 0.5 \). The **probability** of death is reduced by half in the treatment group
Why do we need marginal effects?

- Note something else. With odds ratios and relative risks, we **don’t have a sense of the magnitude**. Same example but now the probability of death in the control group is 0.0004 and 0.0002 in the treatment group.
- The odds ratio is still 0.375 and the relative risk is still 0.5.
- The **magnitudes** are of course quite different.
- A journalist could still say that, for example, eating broccoli sprouts daily reduces the probability of dying of cancer by half. **By half!!!**
- But if you learned that the reduction is \((0.0004 - 0.0002) = 0.0002\) or 0.02 percent points, you probably are not going to run to Whole Foods to get a $4 serving of broccoli sprouts every day.
- On the other hand, a difference of 20 percent point looks quite impressive.
- As we will see, marginal effects is a way of presenting results as **differences in probabilities**, which is more informative than odds ratios and relative risks.
Why do we need marginal effects?: Recap

- Ideally, we want to understand what the model saying in the probability scale and not in the odds scale, much less in the estimation scale, the log-odds.
- In the probability scale, all effects are non-linear because, conditional on covariate values, the probability must be bounded between 0 and 1.
- Here is when numerical methods come to the rescue. If we have interactions in logistic models, we truly need numerical methods.
- We have called them marginal effects but they come in many other names and there are different types.
- Big picture: it’s all about PREDICTION for INTERPRETATION. We are using the estimated model to make predictions so we can better interpret the model in the scale that makes more sense.
- We also saw that we could present adjusted results, which sometimes can be confused with marginal effects so we are going to go over “adjusting,” which you can also do with the margins command.
Big picture: not just for logistic models

- We are using the logistic model to talk about marginal effects because in logistic models we tend to care about effects in the probability scale but coefficients are estimated in the log odds scale.
- But marginal effects are applicable to any other model.
- We will also use them to interpret linear models with more difficult functional forms.
- We will use them with probit models to again use the probability scale.
- Marginal effects are used for Poisson models or any other GLM model or, really, most parametric models.
Big picture: it’s about (numerical) derivatives

- Again, we are going to estimate a model and we are going to use **predictions** to help us **interpret** the model.
- Our goal will be to take **numerical derivatives** of functions for which derivatives are complicated.
- As we saw, in the model \( y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \) we can take the derivative with respect to \( X \):
  \[
  \frac{dE[y_i|X_i]}{dX_i} = \beta_1 + 2\beta_2 X_i
  \]
- In the logistic model, our estimation scale is the **log-odds** but we **would like to interpret our model in the probability** scale.
- We saw that in the probability scale our model is
  \[
  p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}
  \]
Big picture: it’s about (numerical) derivatives

- We know that for a function $y = e^X$ the derivative with respect to $X$ is $\frac{dy}{dX} = e^X$ so the above expression is not going to simplify to a simpler expression.

- In fact, the effect of $X_1$ depends on the value of $X_1$ and the values of all other covariates:

$$\frac{\partial p}{\partial X_1} = \frac{\beta_1 e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{(1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)})^2}$$

- We would need to choose values for all the covariates in the model.

- Instead, we will approximate the analytical derivative numerically using the definition of derivative and we will compute the average effect of $X_1$ on $p$. 


Derivative, redux

■ In the review of statistics and math lecture we saw that the analytical derivative is a limit:
\[
\lim_{h \to 0} \frac{f(x+h)-f(x)}{h}
\]

■ All the **formulas for the derivative can be derived** using the definition and taking the limit. For example, an easy one for 
\[
f(x) = x^2
\]
\[
\lim_{h \to 0} \frac{(x+h)^2-x^2}{h} = \frac{x^2+2xh+h^2-x^2}{x} = \frac{2xh+h^2}{h} = 2x + h = 2x
\]

■ Numerically, that is, without finding the analytical formula, we could use the definition plugging in a number for \(h\) that is small enough. In that case:
\[
\lim_{h \to 0} \frac{f(x+h)-f(x)}{h} \approx \frac{f(x+h)-f(x)}{h}
\]

■ Computationally, it’s not trivial to come up with a number \(h\) that is small but “large enough” so that computations are numerically accurate

■ For more on choosing \(h\) see Gould, Pitblado, and Poi (2010), Chapter 1 and Greene (2017) Appendix E
Two-sided derivatives

- The above approximation to the analytical derivative
  \[ f'(x) \approx \frac{f(x+h)-f(x)}{h} \] is not the only way to define the derivative.

- A two-sided derivative calculates a positive and negative small change \( h \):
  \[ f_2'(x) \approx \frac{f(x+h)-f(x-h)}{2h} \]

- The same issues about how to choose a small \( h \) remain.

- We will calculate marginal effects “by hand” and then we will use the `margins` command.

- We will use **both definitions of the derivative** but **Stata uses the two-sided version**.

- Our calculations will be very close but Stata uses an additional iterative procedure that changes the value of \( h \) to achieve numerical accuracy.
Terminology

- As usual, language that originates in one discipline doesn’t translate well to others. The term “marginal affects” is common in economics and is the language of Stata.
- Gelman and Hill (2007) use the term “average predicted probability” to refer to the same concept as marginal effects (in the logit model).
- SAS and R have some procedures that can get marginal effects and are also called marginal effects as well.
- One confusion is that when you tell your statistician friend about marginal effects, your friend imagines an integral because of marginal probability density functions (in a table of joint probabilities, the probabilities “at the margin” are the marginal probabilities).
- In economics, marginal means “additional” or “incremental,” which is a derivative.
- Career advice: When you use marginal effects in a presentation/paper, make sure that you explain what you mean when you show marginal effects.
The margins command must be treated with respect and caution

- Stata’s margins command is worth the price of Stata. It’s truly awesome.
- But it’s very easy to get an answer that is different from what you wanted.
- A small change in syntax produces very different results. Always read the output and be very, very careful.
- “adjust” and “mfx” used to be two separate commands. Now margins does the same (plus more things) but the syntax became more complex.
- Stata is so careful with version control and continuity that you can still use those commands.
- **Bottom line: be careful.**
Data

- We will use birth weight data from Wooldridge (bcuse bwght)
- I created an indicator for low birth weight. Very low birth weight is about 50 ounces (1,500 grams) but here I used 100 ounces
- About 15% of children (mean of lw)

```
bwght          birth weight, ounces
faminc       1988 family income, $1000s
motheduc      mother’s yrs of educ
cigs           cigs smoked per day while preg

gen lw = 0
replace lw = 1 if bwght < 100 & bwght ~= .

sum lw faminc motheduc cigs
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>lw</td>
<td>1,388</td>
<td>.1491354</td>
<td>.3563503</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>faminc</td>
<td>1,388</td>
<td>29.02666</td>
<td>18.73928</td>
<td>.5</td>
<td>65</td>
</tr>
<tr>
<td>motheduc</td>
<td>1,387</td>
<td>12.93583</td>
<td>2.376728</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>cigs</td>
<td>1,388</td>
<td>2.087176</td>
<td>5.972688</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>
Model

- We will model low birth weight as a function of cigarettes, mother’s education, and family income

\[
\log\left(\frac{l_{wi}}{1-l_{wi}}\right) = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 motheduc_i
\]

logit lw cigs faminc motheduc, nolog

Logistic regression

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>z</th>
<th>P&gt;z</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>lw cigs faminc motheduc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cigs</td>
<td>0.049006</td>
<td>0.0104436</td>
<td>4.30</td>
<td>0.000</td>
<td>0.0244316 0.0653696</td>
</tr>
<tr>
<td>faminc</td>
<td>-0.0080855</td>
<td>0.004801</td>
<td>-1.68</td>
<td>0.092</td>
<td>-0.0174953 0.0013243</td>
</tr>
<tr>
<td>motheduc</td>
<td>0.0031552</td>
<td>0.037153</td>
<td>0.08</td>
<td>0.932</td>
<td>-0.0696634 0.0759738</td>
</tr>
<tr>
<td>_cons</td>
<td>-1.678173</td>
<td>0.4497551</td>
<td>-3.73</td>
<td>0.000</td>
<td>-2.559676 -0.796687</td>
</tr>
</tbody>
</table>

Number of obs = 1,387
LR chi2(3) = 24.63
Prob > chi2 = 0.0000
Log likelihood = -572.15891
Pseudo R2 = 0.0211
Model

- We can show odds ratios: An additional cigarette increases the odds of low birth weight by 4.6%
- Of course, we now know that it does NOT mean that an additional cigarette increases the *relative probability* of low birth weight by 4.6% or that it’s even a *difference* in probability, which is what we want

logit, or

Logistic regression

|                | Odds Ratio | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|----------------|------------|-----------|-------|-----|---------------------|
| cigs           | 1.045924   | .0109232  | 4.30  | 0.000| 1.024733            | 1.067554 |
| faminc         | .9919471   | .0047623  | -1.68 | 0.092| .9826569            | 1.001325 |
| motheduc       | 1.00316    | .0372704  | 0.08  | 0.932| .9327077            | 1.078934 |
| _cons          | .1867149   | .083976   | -3.73 | 0.000| .0773298            | .4508283 |
**Model**

- We can also run our trusty linear model with the caveat that SEs are likely not right (but probably close) and that since low birth probability is (relatively) low we should be extra careful.

- Now, in the probability scale, an extra cigarette increases the probability of low birth weight by 0.7%. With 10 cigarettes, 7% (but that assumes a linear effect)

```bash
reg lw cigs faminc motheduc, robust
```

### Linear regression

|                  | Coef.    | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------------|----------|-----------|-------|-----|----------------------|
| lw               |          |           |       |     |                      |
| cigs             | 0.007757 | 0.0020677 | 3.75  | 0.000 | 0.0037009 - 0.0118131 |
| faminc           | -0.0009345 | 0.0005785 | -1.62 | 0.106 | -0.0020693 - 0.0002004 |
| motheduc         | 0.0005403 | 0.0042972 | 0.13  | 0.900 | -0.0078895 - 0.00897 |
| _cons            | 0.1531912 | 0.0532648 | 2.88  | 0.004 | 0.0487027 - 0.2576797 |
A plot is always helpful

- A plot will help you understand the shape of the relationship of interest but remember that other variables may change the shape

```
lowess lw cigs, gen(lw_c)
scatter lw cigs, jitter(3) msize(small) || ///
    line lw_c cigs, color(blue) sort legend(off) saving(l.gph, replace)
graph export l.png, replace
```
Average Marginal Effect (AME)

- We saw this last class: we compute the (numerical) marginal effect of a variable $X_j$ for each observation using the other covariates $X_k$ as they were observed. We change the value of a predictor by “a little bit” and then we compute the average change across observations.

- In other words, the numerical derivative. Think of marginal effects as getting an average derivative: $\frac{\partial p}{\partial X_j}$

- When using the `margins` command, make sure that 1) you use the option `dydx(varname)` and 2) make sure you use factor syntax so Stata knows that variables are continuous or dummy. Also, Stata will figure out if there are interactions.

- If you coded interactions by hand (say: `gen agemale = age*male`), you will get wrong results using margins without factor syntax.
Average Marginal Effect (AME)

- Let's calculate AME for the cigarette variable using the typical formula for the analytical derivative

```
* Get the "small change"
qui sum cigs
scalar h = (abs(r(mean))+.0001)*.0001
di h
*.00020873
preserve
qui logit lw cigs faminc motheduc, nolog
* as is
predict double lw_0 if e(sample)
* Change cigs by a bit
replace cigs = cigs + scalar(h)
predict lw_1 if e(sample)
* For each obs
gen double dydx = (lw_1-lw_0)/scalar(h)
* Average
sum dydx
restore
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>dydx</td>
<td>1387</td>
<td>.0055768</td>
<td>.0012444</td>
<td>.0040507</td>
<td>.0113006</td>
</tr>
</tbody>
</table>

- An additional cigarette increases the probability of low birth weight by 0.56%
Average Marginal Effect (AME)

- Replicate using margins command

```
. margins, dydx(cigs)
```

| Expression : Pr(lw), predict() |
| dy/dx w.r.t. : cigs |

```
Average marginal effects                       Number of obs  =  1,387
Model VCE       : OIM

Expression : Pr(lw), predict()
dy/dx w.r.t. : cigs

+--------------------------------------------------+
| Delta-method                                      |
| dy/dx   Std. Err.   z    P>|z|  [95% Conf. Interval] |
|--------------------------------------------------|
| cigs    | .0055782  .0012814 | 4.35 | 0.000 | .0030666 | .0080898 |
+--------------------------------------------------+

- Very close but Stata uses the two-sided derivative
- The formula \( h = (|\bar{x}| + .0001) \times .0001 \) is how Stata calculates the initial small change. In this case, .00020873
Average Marginal Effect (AME) the (almost) Stata way

* Two-sided derivative
preserve
    qui logit lw cigs faminc motheduc
    * Define small change for cigs
    qui sum cigs
    scalar h = (abs(r(mean))+0.0001)*0.0001

    * Duplicate variable
    clonevar cigs_c = cigs

    * Small negative change
    replace cigs = cigs_c - scalar(h)
    predict double lw_0 if e(sample)
    * Small positive change change
    replace cigs = cigs_c + scalar(h)
    predict double lw_1 if e(sample)

    gen double dydx = (lw_1-lw_0)/(2*scalar(h))
    sum dydx
resto
* .0055771

Close. Remember, Stata has an iterative procedure to ensure numerical accuracy. This is one reason the margins command for some complex models can be very slow. Stata also calculates the variance using the delta method.
Average Marginal Effect (AME)

- One advantage of doing it “by hand” rather than using margins is that you can change the “small” change to any change or just calculated the difference rather than $\frac{\Delta Y}{\Delta X}$. What about 10 extra cigarettes?

```stata
preserve
    qui logit lw cigs faminc motheduc
    predict double lw_0 if e(sample)

    replace cigs = cigs + 10
    predict double lw_1 if e(sample)

    gen double dydx = (lw_1-lw_0)/10
    * Could be
    * gen dy = (lw_1-lw_0)
    sum dydx
restore
```

```
<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>dydx</td>
<td>1,387</td>
<td>0.0064608</td>
<td>0.0012196</td>
<td>0.0048265</td>
<td>0.0111532</td>
</tr>
</tbody>
</table>
```

- In some cases we may not be interested in an approximation for a small change. You have more flexibility doing it by hand. On the other hand, we could just change the scale of cigs so changes are for 10 units.
AME for indicator variables

- Understanding marginal effects it’s easier with dummy variables; that’s why I have focused on continuous variables
- With dummy variables **we don’t have to do a “small” change.** We change from 0 to 1

```stata
preserve
qui logit lw smoked faminc motheduc
* Nobody smoked
replace smoked = 0
predict double lw_0 if e(sample)

* Everybody smokes
replace smoked = 1
predict double lw_1 if e(sample)

gen double dydx = (lw_1-lw_0)
sum dydx
restore
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>dydx</td>
<td>1,387</td>
<td>0.118284</td>
<td>0.0105063</td>
<td>0.0935378</td>
<td>0.1363816</td>
</tr>
</tbody>
</table>
AME for indicator variables

- We can of course also use the margins command with caution (!)

```
qui logit lw smoked faminc motheduc, nolog
margins, dydx(smoked)
```

```
<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td>dy/dx</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>smoked</td>
<td>.0988076</td>
</tr>
</tbody>
</table>
```

```
qui logit lw i.smoked faminc motheduc, nolog
margins, dydx(smoked)
```

```
<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td>dy/dx</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>1.smoked</td>
<td>.118284</td>
</tr>
</tbody>
</table>
```

Note: dy/dx for factor levels is the discrete change from the base level.

- Even though same margins statement, different results. The first one is not what we wanted. We did not use the factor syntax in the first model so Stata didn’t go from 0 to 1; instead it used a “small” change
- Smoking increases the probability of low birth weight by almost 12% points (yikes)
AME for indicator variables

- With indicator variables, we can also get what Stata calls predictive margins (not marginal effects). Marginal effects are their difference.
- We can also use the results to go from margins to relative risk and to odds ratios.

```
qui logit lw i.smoked faminc motheduc, nolog
margins smoked
```

|        | Margin | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|--------|--------|-----------|-----|------|----------------------|
| smoked |        |           |     |      |                      |
| 0      | .1305183 | .0099014 | 13.18 | 0.000 | .1111118 \(\) .1499248 |
| 1      | .2488023 | .0304311 | 8.18  | 0.000 | .1891584 \(\) .3084461 |

* marginal effects from predictive margins

```
. di .2488023 \(-\) .1305183
.118284
. di .2488023 \(/\) .1305183
1.9062637
. di exp(_b[1.smoked])
2.2115807
```

* Same as

```
. di (.2488023/(1-.2488023))/((0.1305183)/(1-0.1305183))
2.206425
```

- Odds ratio: 2.21, relative probability: 1.90. Difference: 0.118. Remember, the rarer the event the closer they are...
AME for indicator variables

- No predictive margins for continuous variables
  * No margins for continuous variables (marginal effects instead)
  
  ```
  qui logit lw cigs faminc motheduc, nolog
  margins cigs
  factor 'cigs' not found in list of covariates
  r(322);
  ```

- No margins for continuous variables because they take too many values
Marginal Effect at the Mean (MEM)

- We have left the values of the covariates as they were observed rather than holding them fixed at a certain value.
- We can also calculate marginal effects at the mean (of each covariate), much like what we did when we adjusted predictions.
- There is some discussion about which way is better (see Williams, 2012).
- For example, does it make sense to hold male at 0.6 male? In a sense, yes. We are giving makes the value of the proportion in the sample, 0.6. In another sense, it seems odd.
- **Don’t waste too much time thinking about this.** When we calculate marginal effects (not marings), it doesn’t really matter at which value we hold the other covariates constant because we are taking differences in effects. There could some differences in small samples.
- In general, the difference will be so small that it is better to spend mental resources somewhere else.
Marginal Effect at the Mean (MEM)

- Keep covariates at mean values instead

```
preserve
    qui sum cigs
    scalar h = (abs(r(mean)) + 0.0001)*0.0001
    qui logit lw cigs faminc motheduc, nolog
    clonevar cigs_c = cigs

    * At mean
    replace faminc = 29.02666
    replace motheduc = 12.93583

    * Small negative change
    replace cigs = cigs_c - scalar(h)
    predict double lw_0 if e(sample)
    * Small positive change change
    replace cigs = cigs_c + scalar(h)
    predict double lw_1 if e(sample)
    gen double dydx = (lw_1 - lw_0)/(2*scalar(h))
    sum dydx
restore
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>dydx</td>
<td>1,387</td>
<td>0.0055624</td>
<td>0.0010396</td>
<td>0.0051876</td>
<td>0.011267</td>
</tr>
</tbody>
</table>

- MEM not that different from AME
Marginal Effect at the Mean (MEM)

- Using the margins command

```stata
margins, dydx(cigs) at((mean) faminc motheduc)
```

Average marginal effects                   Number of obs  =  1,387
Model VCE       : OIM
Expression      : Pr(lw), predict()
dy/dx w.r.t.    : cigs
at              : faminc        = 29.04218 (mean)
               : motheduc      = 12.93583 (mean)

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dy/dx     Std. Err.</td>
</tr>
<tr>
<td>cigs</td>
<td>.005563    .0012843</td>
</tr>
</tbody>
</table>

- Always read Stata’s output; **remember, be afraid of the margins command!**
Marginal Effect at the Mean (MEM)

- Not the same as using the atmeans option

margins, dydx(cigs) atmeans

Conditional marginal effects Number of obs = 1,387
Model VCE : OIM
Expression : Pr(lw), predict()
dy/dx w.r.t. : cigs
at cigs = 2.088681 (mean) faminc = 29.04218 (mean) motheduc = 12.93583 (mean)

| Delta-method | dy/dx Std. Err. z P>|z| [95% Conf. Interval]
|--------------|-----------------|----------|--------|------------------|
cigs | .0055506 .0012879 4.31 0.000 .0030264 .0080749 |

In this one, cigarettes were held at its mean, 2.088. Not a big deal in this example because in this example the effect of cigs is relatively linear (see lowess plot above) but you could have gotten a very different answer

- One more time: please be careful with the margins command
Marginal effects at representative values (MER)

- We can hold values at observed values (AME) or at mean values (MEM)
- We could also choose representative values; values that are of interest
- For example, what is the marginal effect of an additional cigarette on the probability of low birth weight at different levels of income, say 10K, 20K, 30K and 40K?
- Leave other covariates as observed
Marginal effects at representative values (MER)

We will do it “by hand” for low income (10K) and higher income (40K) using the one-sided version to make the code shorter

preserve
qui logit lw cigs faminc motheduc, nolog
* income 10k
replace faminc = 10
predict double lw_0_10 if e(sample)
replace cigs = cigs + .00597269
predict double lw_1_10 if e(sample)
gen double dydx10 = (lw_1_10-lw_0_10)/.00597269

* income 40k
replace faminc = 40
predict double lw_0_40 if e(sample)
replace cigs = cigs + .00597269
predict double lw_1_40 if e(sample)
gen double dydx40 = (lw_1_40-lw_0_40)/.00597269

sum dydx*
restore

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>dydx10</td>
<td>1,387</td>
<td>.0061672</td>
<td>.0010198</td>
<td>.005653</td>
<td>.0112164</td>
</tr>
<tr>
<td>dydx40</td>
<td>1,387</td>
<td>.0052304</td>
<td>.001039</td>
<td>.0047327</td>
<td>.0111981</td>
</tr>
</tbody>
</table>
Marginal effects at representative values (MER)

Below, income reduces the effect of smoking. Better access to health care? So income is a modifier of the effect?

```
qui logit lw cigs faminc motheduc, nolog
margins, dydx(cigs) at(faminc=(10 20 30 40)) vsquish
```

<table>
<thead>
<tr>
<th></th>
<th>Number of obs</th>
<th>1,387</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model VCE</td>
<td>OIM</td>
<td></td>
</tr>
<tr>
<td>Expression</td>
<td>Pr(lw), predict()</td>
<td></td>
</tr>
<tr>
<td>dy/dx w.r.t.</td>
<td>cigs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>faminc = 10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>faminc = 20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>faminc = 30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>faminc = 40</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dy/dx w.r.t.</th>
<th>faminc</th>
<th>cigs</th>
</tr>
</thead>
<tbody>
<tr>
<td>_at</td>
<td></td>
<td>_at</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>faminc = 10</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>faminc = 20</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>faminc = 30</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>faminc = 40</td>
</tr>
</tbody>
</table>

| dy/dx         | Std. Err.  | z      | P>|z|   | [95% Conf. Interval] |
|---------------|------------|--------|-------|----------------------|
| cigs          |            |        |       |                      |
| _at           |            |        |       |                      |
| 1             | .0061667   | .0014205 | 4.34  | 0.000                | .0033825 .0089508 |
| 2             | .0058463   | .001337 | 4.37  | 0.000                | .0032258 .0084668 |
| 3             | .0055334   | .0012803 | 4.32  | 0.000                | .0030239 .0080428 |
| 4             | .0052289   | .0012501 | 4.18  | 0.000                | .0027788 .0076789 |
Marginal effects at representative values (MER)

- **But wait.** Why different effects by income? We did not add **interactions between income and cigs and income** in the model.
- Because in the probability scale the effect cigs depends on the values of income and all other covariates!!!!
- This won’t happen in the linear probability model.
- If we think that the effect of cigs depend on income, we **should add interactions** to make this explicit.
Same but with LPM

- Since there are no interactions, the marginal effect doesn’t depend on the value of income

```
qui reg lw cigs faminc motheduc
margins, dydx(cigs) at(faminc=(10 20 30 40)) vsquish
```

Average marginal effects  Number of obs  =  1,387
Model VCE : OLS

Expression : Linear prediction, predict()
dy/dx w.r.t. : cigs
1._at : faminc  =  10
2._at : faminc  =  20
3._at : faminc  =  30
4._at : faminc  =  40

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dy/dx</td>
<td>Std. Err.</td>
<td>t</td>
<td>P&gt;</td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
<td>-------------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>cigs</td>
<td></td>
<td></td>
<td></td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>_at</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.007757</td>
<td>.001631</td>
<td>4.76</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>.007757</td>
<td>.001631</td>
<td>4.76</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>.007757</td>
<td>.001631</td>
<td>4.76</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>.007757</td>
<td>.001631</td>
<td>4.76</td>
<td>0.000</td>
</tr>
</tbody>
</table>
With interactions the effect should be more noticeable

- Now adding interactions between cigarettes and income. This is the right way of making the effect of cigs depend on income.
- Note how the conclusion is different. However, the interaction is not statistically significant.

```
qui logit lw c.cigs##c.faminc motheduc, nolog
margins, dydx(cigs) at(faminc=(10 20 30 40)) vsquish
Average marginal effects
Model VCE : OIM
Expression : Pr(lw), predict()
dy/dx w.r.t. : cigs
1._at : faminc = 10
2._at : faminc = 20
3._at : faminc = 30
4._at : faminc = 40
------------------------------------------------------------------------------
| Delta-method
| dy/dx Std. Err. z P>|z| [95% Conf. Interval]
|-----|-----|-----|-----|-----------------|-----------------|
cigs  |     |     |     |     |     |     |
     |     |     |     |     |     |     |
_1.at |     |     |     |     |     |     |
   1  | .0054953 .0016702 3.29 0.001 .0022217 .0087689
   2  | .0059858 .0013512 4.43 0.000 .0033374 .0086342
   3  | .0064006 .0016609 3.85 0.000 .0031453 .0096558
   4  | .0067452 .0022472 3.00 0.003 .0023408 .0111497
------------------------------------------------------------------------------
```

38
Marginsplot

- You can visualize changes using marginsplot. This is a way to get adjusted plots using the margins command

```
marginsplot, saving(mp.gph, replace)
graph export mp.png, replace
```
Interactions

- We have estimated the model
  \[ \log\left( \frac{l_{w_i}}{1-l_{w_i}} \right) = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 motheduc_i \]
- We didn’t use interactions between cigarettes and income so we have assumed the same effect regardless of income in the **log-odds scale**
- In other words, same slope and same intercept
- We could add interactions. And that’s when the margins command is a life saver because effects are hard to compute otherwise
Interactions

Effects across the values of income

```
logit lw c.cigs##c.faminc motheduc, nolog
margins, dydx(cigs)
```

Average marginal effects

```
Model VCE : OIM
Expression : Pr(lw), predict()
dy/dx w.r.t. : cigs
```

|                      | Delta-method | dy/dx | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|----------------------|-------------|-------|-----------|-------|-----|---------------------|
| cigs                 | .0061543    | .0014573 | 4.22     | 0.000 |     | .0032981  .0090106  |

Wait, why one effect?? **Careful with something like this!** Margins is computing predicted $P(lw = 1)$ for a small change in cigs regardless of the value of family income.

But if we know that the effect of cigs depends on faminc, it would be better to compute the marginal effects at some value of faminc. The margins command is giving you the average effect.

**One more time, careful with the margins command!!!**
Interactions

- As usual, interactions are easier to understand with two indicator variables. Let’s say more than high school (hsp) and male
- Our model is \( \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 hsp + \beta_2 male + \beta_3 hsp \times male \)
- In the log-odds scale, we interpret them exactly in the same way as in the linear model and we can use derivatives and the other tricks we learned
- For males: \( \log\left(\frac{p_m}{1-p_m}\right) = \beta_0 + \beta_1 hsp + \beta_2 + \beta_3 hsp \)
- For females: \( \log\left(\frac{p_f}{1-p_f}\right) = \beta_0 + \beta_1 hsp \)
- Difference males - females: \( \log\left(\frac{p_m}{1-p_m}\right) - \log\left(\frac{p_f}{1-p_f}\right) = \beta_2 + \beta_3 hsp \)
- So the effect depends on the level of education
- Note that we can write it in terms of odds: \( \frac{\frac{p_m}{1-p_m}}{\frac{p_f}{1-p_f}} = e^{\beta_2 + \beta_3 hsp} \)
Interactions

- Difference males - females: \( \log\left(\frac{p_m}{1-p_m}\right) - \log\left(\frac{p_f}{1-p_f}\right) = \beta_2 + \beta_3 hsp \)
- Difference male - female for educated:
  \( \log\left(\frac{p_{me}}{1-p_{me}}\right) - \log\left(\frac{p_{fe}}{1-p_{fe}}\right) = \beta_2 + \beta_3 \)
- Difference male - female for uneducated:
  \( \log\left(\frac{p_{mu}}{1-p_{mu}}\right) - \log\left(\frac{p_{fu}}{1-p_{fu}}\right) = \beta_2 \)
- Difference in difference:
  \( \log\left(\frac{p_{me}}{1-p_{me}}\right) - \log\left(\frac{p_{fe}}{1-p_{fe}}\right) - \left[\log\left(\frac{p_{mu}}{1-p_{mu}}\right) - \log\left(\frac{p_{fu}}{1-p_{fu}}\right)\right] = \beta_3 \)
- So same as with linear model. In the log-odds scale, it is a difference-in-difference
Interactions in the odds scale

- \( \log\left( \frac{p_{me}}{1-p_{me}} \right) - \log\left( \frac{p_{fe}}{1-p_{fe}} \right) - \left[ \log\left( \frac{p_{mu}}{1-p_{mu}} \right) - \log\left( \frac{p_{fu}}{1-p_{fu}} \right) \right] = \beta_3 \)

- We can apply the rules of logs and take \( e() \) on both sides:

\[
\frac{\frac{p_{me}}{1-p_{me}}}{\frac{p_{mu}}{1-p_{mu}}} = e^{\beta_3}
\]

- In the odds scale, an interaction is not a difference-in-difference but rather a **ratio of odds ratios**

- The interpretation is... well, good luck with that...

- The interpretation is not that hard, actually. If greater than one, the odds of outcome for educated are higher than for the uneducated

- Interpreting the magnitude is a lost cause. A lot easier with margins or stratifying the models
Interactions

- Remember, a fully interacted model is the same as a stratified model; it helps to think about it when verifying that interactions in the odds scale is a ratio of odds ratios

- Make sure you understand this

```
gen   inc = 0 if faminc ~=.
replace inc = 1 if faminc > 40 & faminc ~=.
logit lw c.cigs##i.inc, nolog or

------------------------------------------------------------------------------
lw |   Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+--------------------------------------------------------------
cigs |   1.050702   .0109996    4.72  0.000     1.029363   1.072483
1.inc |    .904556   .1740226   -0.52  0.602     .6204083   1.318844
inc#c.cigs |
   1 |    .947913   .0553484   -0.92  0.360     .8454093   1.062845
   _cons |   .1583417   .0151181  -19.30  0.000     .1313182   .1909263
------------------------------------------------------------------------------
qui logit lw cigs if inc == 0, nolog
    . di exp(_b[cigs])
    1.0507018
qui logit lw cigs if inc == 1, nolog
    . di exp(_b[cigs])
    .99597388
    . di .99597388/1.0507018
    .94791299
```
We can of course estimate marginal effects for linear models without nonlinear terms but they will be boring

```stata
digression
sysuse auto, clear
reg price mpg turn

price | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
mpg | -259.6967  76.84886  -3.38  0.001  -412.929   -106.4645
turn | -38.03857  101.0624  -0.38  0.708  -239.5513   163.4742
_cons | 13204.27   5316.186   2.48  0.015     2604.1  23804.45
-------------+----------------------------------------------------------------
margins, dydx(turn)

| Delta-method
| dy/dx Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
turn | -38.03857  101.0624  -0.38  0.708  -239.5513   163.4742
-------------+----------------------------------------------------------------
```

46
Margins are predictions

- The essence of margins and marginal effects is that they are predictions
- We are using our estimated model to make predictions when we change a continuous variable by a small amount or when we change an indicator variable from 0 to 1
- They are extremely useful because they allow us to interpret our models
- They are truly indispensable when the scale of estimation is not the same as the scale of interest (logit, Poisson, etc) or when we have non-linear terms
“Adjusting”

- We saw that **adjusting** is a good way of graphically presenting results
- We make predictions “holding” other covariates at a constant value
- The value at which we hold the covariates constant **matters** (but we will see shortly cases in which it doesn’t matter that much)
- Thankfully, adjusting is the common term in many areas so no confusion, but...

**Confusion**: In epi, when incidence or prevalence rates are adjusted, they are adjusted at some particular distribution (similar idea; different way to do it)
Example

- Say, we have trends in hip fractures that are increasing. It could be that hip fractures are going up just because the population is getting older.
- So we want to “adjust” for the aging population and present adjusted trends; in many cases, we don’t have individual level data so we can’t run a regression model.
- So calculate hip fractures by age group and year. Then hold age distribution at one particular year and weight the rates using those weights.
Excel example

- Go over the Excel example (adjust.xlsx)
- I held the distribution to be the same as Year 1’s distribution

<table>
<thead>
<tr>
<th>Age</th>
<th>Incidence Year 1</th>
<th>Incidence Year 2</th>
<th>Incidence Year 3</th>
<th>Age distribution Year 1</th>
<th>Age distribution Year 2</th>
<th>Age distribution Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>65-75</td>
<td>0.100</td>
<td>0.105</td>
<td>0.112</td>
<td>0.500</td>
<td>0.475</td>
<td>0.449</td>
</tr>
<tr>
<td>75-85</td>
<td>0.200</td>
<td>0.210</td>
<td>0.225</td>
<td>0.200</td>
<td>0.210</td>
<td>0.220</td>
</tr>
<tr>
<td>85+</td>
<td>0.350</td>
<td>0.368</td>
<td>0.393</td>
<td>0.300</td>
<td>0.315</td>
<td>0.331</td>
</tr>
<tr>
<td></td>
<td>Unadjusted</td>
<td></td>
<td></td>
<td>Adjusted</td>
<td>0.195</td>
<td>0.205</td>
</tr>
</tbody>
</table>
Adjusting with models

- We can also adjust with models rather than keeping the distribution constant.
- For example: what is the probability of low birth weight as a function of cigarettes holding the other covariates constant?
- Or what is the probability of hip fracture by year holding age constant?
- We can hold the values constant at different values and compare adjusted trends or just keep them constant at their mean.
- No hard rules; most often adjusted at means.
Model adjusted

- For example, keep income and education at their mean values
- Then plot the adjusted trend (you would describe this as “adjusted for income and education”)

```
preserve
    qui logit lw cigs faminc motheduc, nolog
    replace faminc = 30
    replace motheduc = 13
    predict plw_adj if e(sample)

    sum plw_adj
    line plw_adj cigs, sort color(red) saving(adj.gph, replace)
    graph export adj.png, replace
restore
```
Adjusted probability of lw and number of cigarettes

- Adjusted probability (we entered cigs as linear in the log-odds model, but it’s non-linear in the probability scale)
Predictions in logistic models

- We saw that we can easily make predictions in the probability scale

```
logit lw smoked faminc, nolog
Logistic regression
Number of obs = 1,388
LR chi2(2) = 24.30
Prob > chi2 = 0.0000
Log likelihood = -572.48309 Pseudo R2 = 0.0208
------------------------------------------------------------------------------
lw | Coef. Std. Err. z P>|z| [95% Conf. Interval]
-------------+----------------------------------------------------------------
smoked | .7898437 .1830874 4.31 0.000 .4309989 1.148688
faminc | -.0076573 .0043414 -1.76 0.078 -.0161662 .0008516
_cons | -1.681833 .151239 -11.12 0.000 -1.978256 -1.38541
------------------------------------------------------------------------------
.24569462
```

- A mother who smoked during pregnancy with 30K in family income has 25% probability of having a child weighting less than 100 ounces at birth

- Pseudo $R^2$ is low but LRT test tell us that the two explanatory variables are better than a model with no explanatory variables (never a surprise)
Predictions in logistic models

- One way to evaluate the predictive ability of our models is to compare predictors and observed values.
- We did so with linear models. We can use the root mean square error (RMSE) or the $R^2$ because it is also the square of the correlation between observed and predicted values.
- In logistic models, the observed value is a 1/0 variable but predicted values are either in the log odds scale or in the probability scale.
- We can transform probabilities into 1/0 values. If the predicted probabilities is $\geq 0.5$, then that means that the observation is more likely than not to experience the event.
One way to evaluate the predictive ability of our models is to compare predictors and observed values.

We did so with linear models. We can use the root mean square error (RMSE) or the $R^2$ because it is also the square of the correlation between observed and predicted values.

In logistic models, the observed value is a 1/0 variable but predicted values are either in the log odds scale or in the probability scale.

We can transform probabilities into 1/0 values. If the predicted probabilities is $\geq 0.5$, then that means that the observation is more likely than not to experience the event.
Correctly predicted

■ Calculating the observations correctly predicted

```stata
qui logit lw smoked faminc, nolog
predict phat if e(sample)
gen    hatlw = 0 if phat ~= .
replace hatlw = 1 if phat >= 0.5 & phat ~= .

tab lw hatlw, row col
```

<table>
<thead>
<tr>
<th></th>
<th>lw</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>lw</td>
<td>hatlw</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1,180</td>
<td>1,180</td>
</tr>
<tr>
<td>1</td>
<td>207</td>
<td>207</td>
</tr>
<tr>
<td>Total</td>
<td>1,387</td>
<td>1,387</td>
</tr>
</tbody>
</table>

■ But... low birth weight is not that common so using 0.5 as the cut off point doesn’t make much sense
Remembering epi

- From Wiki:
- **Sensitivity**: True positive; proportion of positives that are correctly identified as such (i.e. the percentage of sick people who are correctly identified as having the condition).
- **Specificity**: True negative; proportion of negatives that are correctly identified as such (i.e., the percentage of healthy people who are correctly identified as not having the condition).
- False positive: 1-specificity
- Sensitivity and specificity are both correct predictions, either positive (1) or negative (0)
- **We can just focus on whether we get the 1s right.** We will use true positives (sensitivity) and false positives (1-specificity)
Calculating sensitivity and specificity

- We need to come up with a cut-off point; we saw that if the cut-off point is 0.5 our rate of false positives (1-specificity) is 1 because we don’t classify anybody as 1.
- If we lower the cut-off too much, everybody will be a one: our model is too sensitive but not specific.
- Of course, there is a command for that and graph: the post-estimation command `lsens`.
Sensitivity and specificity

- It’s a trade off. If we lower the cut-off, we call too many observations a 1 so we make more false positive mistakes (1-specificity)

```
qui logit lw smoked faminc motheduc, nolog
lsens, saving(sens.gph, replace)
graph export sens.png, replace
```
How can we evaluate predictions?

- Remember, the outcome is a 1/0 variable. If we just try to guess randomly, we have 50/50 change to get it right.
- So our model should be at least better than chance.
- One way to calculate and graph this is by using the **Receiver Operating Characteristic (ROC)** (has its origins in signal detection theory).
- Essentially, it plots sensitivity and 1-specificity (true positives, false positives) using different cut-off points to determine if the observation is a 1 (going from cut-off point 0 to 1).
- The **area under the curve** is a measure of how good is the model at discriminating 1s.
- It also called **c-statistic** or **concordance** statistic (higher is better).
There is, of course, a command for that: \texttt{lroc}

\begin{verbatim}
qui logit lw smoked faminc, nolog
lroc, saving(lrocm1.gph, replace)

qui logit lw smoked faminc motheduc parity, nolog
lroc, saving(lrocm2.gph, replace)

graph combine lrocm1.gph lrocm2.gph, xsize(15) ysize(10)
graph export lrocm2.png, replace
\end{verbatim}
A random model (bad)

- This is how a model with no predictive power looks like

```
gen x1 = rnormal(0,10)
quilogit lw x1
lroc, saving(noise.gph, replace)
graph export noise.png, replace
```
What happened?

- In our linear model, adding more variables was always better ($R^2$ won’t go down)
- When you are in the world of models in which the mean also determines the variance, adding more variables is not always better
- We just made the model worse: area under the curve went from 0.6 to 0.59. Not terrible, but adding more variables was not better
- Parsimony in these models is a good thing and we must be careful about adding unnecessary variables
Summary

- Main difficulty with logistic models is to interpret parameters
- Marginal effects come to the rescue
- Different terms for these types of effects. AMEs are usually called **average predicted comparisons**
- What we did today was about **PREDICTION**, 100 percent. We use predictions to understand what our models are saying
- The existence of the margins command has unified some of the terminology
- But if you talk to your friendly statistician, **you need to explain what you mean by marginal effects**. They start thinking about integrals in their heads when we mean derivatives...