Interpreting Model Estimates: Marginal Effects

Marcelo Coca Perraillon
marcelo.perraillon@cuanschutz.edu

University of Colorado
Anschutz Medical Campus
—

Nova School of Business and Economics
Lisbon, Portugal
July, 2019

Updated lectures are here: https://clas.ucdenver.edu/marcelo-perraillon/code-and-topics/marginal-effects

These notes are part of a forthcoming book to be published by Cambridge University Press.

Book details:

https://www.perraillon.com/PLH/

Sample Chapter 6 on Marginal Effects

http://www.perraillon.com/PLH/PLH%20marginal%

20effects.pdf

Outline

- Why do we need marginal effects?
- Derivatives to interpret more complex models
- Derivatives definition, 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
- Marginal effects for two-part models
-

Why do we need marginal effects?

- In a simple linear model, say, $y = \beta_0 + \beta_1 age + \beta_2 male$, we can easily interpret the coefficients
- It is less straightforward when there are non-linear terms, for example: $y = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 male$
- Using the **analytical derivative** makes 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. We just plug in numbers for age in the above expression to get the effect at different ages
- Note that **effects** are not the same as **predictions**

Why do we need marginal effects?

- With interactions, it's more complicated: $y = \beta_0 + \beta_1 age + \beta_2 male + \beta_3 male * age$
- 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 interactions, taking the derivative still helps with interpretation: $\frac{\partial E[y|age, male]}{\partial age} = \beta_1 + \beta_3 * male$
- **Centering** also helps with parameter interpretation: $y = \beta_0 + \beta_1(age - m) + \beta_2 male + \beta_3 male * (age - m)$
- If m is average age, then β_2 is $E[Y]$ for males versus females of average age. What is β_1?
- In **non-linear models** interpretation is often more difficult

Logit/probit model reminder

- There are several ways of deriving the logit model. We can assume a latent outcome or assume the observed outcome 1/0 distributes either Binomial or Bernoulli. The latent approach is convenient because it can be used to derive both logit and probit models
- We assume that there is a **latent (unobserved) variable** y^* that is continuous. Think about it as a measure of illness
- If y^* crosses a **threshold**, then the person dies. We only observe if the person died but we can't, by definition, observe the latent variable y^*
- What is the probability of dying? We can write this problem as: $P(y = 1|X) = P(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + u > 0) = P(-u < \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) = F(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$
- F() is the cdf of -u. If we assume logistic distribution, we get logistic regression, if we assume cumulative normal, we get a probit model

Logit/probit model reminder

- Remember that the **cumulative distribution function** (cdf) gives you $P(X < a)$. Remember too that to get the probability you need to integrate the density f(t) from $-\infty$ to a: $\int_{-\infty}^{a} f(t)dt$
- If we assume **standard normal cdf**, our model then becomes $P(y = 1|x) = \int_{-\infty}^{\beta_0+\beta_1 x} \frac{1}{2\pi} e^{(-\frac{t^2}{2})} dt$
- And that's the **probit** model. Note that because we use the cdf, the probability will obviously be constrained between 0 and 1 because, well, it's a cdf
- If we assume that u distributes **standard logistic** then our model becomes $P(y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$
- Remember that there are two different concepts: **logistic response function** and **logistic distribution**. The **standard logistic cdf** happens to have the above formula (the pdf is different)

Logit/probit model reminder

- For now, the most important part to remember is that the **scale of estimation** is not the same as the **scale of interest** (more on this in 30 seconds)
- This is because we use transformations to constraint the probability between 0 and 1
- Keep in mind that the logistic model can be derived in different ways. This tends to confuse students. All ways lead to **same likelihood function and therefore the same parameters**
- Back to why we need marginal effects...

Why do we need marginal effects?

- We can write the **logistic model** as: $log(\frac{p}{1-p}) = \beta_0 + \beta_1 age + \beta_2 male$
- The estimated parameters are in the **log-odds scale**, which, other than the sign, don't have any useful interpretation
- In the above equation, β_1 is the effect of age on the log-odds of the outcome, **not on the probability**, which is often what were care about
- As an alternative, economists prefer to estimate **Probit** models for binary outcomes:
- $P(y = 1|male, age) = \Phi(\gamma_0 + \gamma_1 age + \gamma_2 male)$
- But still similar problem. In the estimation scale, γ_1 is interpreted as shifts in the standard cumulative normal, which, again, is of little help

Why do we need marginal effects?

- With the logit model we could present odds ratios (e^{β_1} and e^{β_2}) but **odds-ratios are often misinterpreted** as if they were relative risks/probabilities (nonetheless presenting odds-ratios is standard practice in the medical literature)
- A simple example with no covariates: Say that 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: $\frac{\frac{0.2}{1-0.2}}{\frac{0.4}{1-0.4}} = 0.375$. 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 ($\frac{1}{0.375}$)
- 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 Miosotis or Celeiro or Pingo Doce 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
- We call them **marginal effects** in econometrics but they come in many other names and there are different types
- **Big picture**: marginal effects use model **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 (but we are not trying to evaluate how good is the model at predicting...)

Big picture: not just for logit/probit models

- We are going to use the logistic model to introduce marginal effects
- But marginal effects are applicable to any other model
- We will also use them to interpret linear models with more difficult functional forms
- Marginal effects can be use with Poisson models, GLM, two-part models. In fact, most parametric models

It's about derivatives

- 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 **analytical derivatives** are more 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
- In the probability scale, the model is $p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p)}}$
- **Reminder**: To go from log-odds to probabilities, solve for p in $log(\frac{p}{1-p}) = \beta_0 + \beta_1 X$. You get $p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

It's about 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, review

- 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. This is what **slows down** marginal effects calculations

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
- 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 for non-economists, make sure that you explain what you mean when you show marginal effects

Digression: Is it a unit change?

- In a model like $y = \beta_0 + \beta_1 x + \epsilon$ we have have said that β_1 is the change in $E[y]$ for a one unit change in x
- And that's fine when the change is constant (linear)
- In the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, the **marginal effect/change is no longer for a 1 unit change even though most people would interpret it that way when using marginal effects**. Do it by hand:
- Start with $x = x^0$. Then change by one unit to $x^0 + 1$ and compare the two predictions for y. (For simplicity, dropping the "hat" from the betas but the betas are the same)

  $$y^1 - y^0 = \beta_0 + \beta_1(x^0 + 1) + \beta_2(x^0 + 1)^2 - \beta_0 - \beta_1 x^0 - \beta_2(x^0)^2$$
- You will find: $y^1 - y^0 = \beta_1 + 2\beta_2 x^0 + \beta_2$, which is not the same as $dy/dx = \beta_1 + 2\beta_2 x$ evaluated at $x = x^0$. There is the extra β_2 term lingering there

Digression: Is it a unit change?

- Easy to understand why if you use the definition of the derivative: $\lim_{h \to 0} \frac{\beta_0 + \beta_1(x+h) + \beta_2(x+h)^2 - \beta_0 - \beta_1 x - \beta_2 x}{h} = \beta_1 + 2\beta_2 x + \beta_2 h$
- So when h tends to zero both ways will be the same (increasing by one or taking the derivative). **The marginal effect applies to a very small change in x, not to an one unit change in x**
- When are they going to be the same? The above formula provides some intuition: β_2 is the coefficient for x^2, the curvature. The smaller the β_2 the smaller the curvature and closer to a straight line
- So they are close when the curve looks more like a line
- **Bottom line: marginal effects for continuous variables apply to a small change in x when effects are non-linear**. They are not changes by 1 unit, strictly speaking

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 smked per day while preg

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

sum lw faminc motheduc cigs
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          lw |      1,388    .1491354    .3563503          0          1
      faminc |      1,388    29.02666    18.73928         .5         65
    motheduc |      1,387    12.93583    2.376728          2         18
        cigs |      1,388    2.087176    5.972688          0         50
```

Model

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

  $$log(\frac{lw_i}{1-lw_i}) = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 motheduc_i$$

```
logit lw cigs faminc motheduc, nolog

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

------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |   .0449006   .0104436     4.30   0.000     .0244316    .0653696
      faminc |  -.0080855    .004801    -1.68   0.092    -.0174953    .0013243
    motheduc |   .0031552    .037153     0.08   0.932    -.0696634    .0759738
       _cons |  -1.678173   .4497551    -3.73   0.000    -2.559676   -.7966687
------------------------------------------------------------------------------
```

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

------------------------------------------------------------------------------
          lw | 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

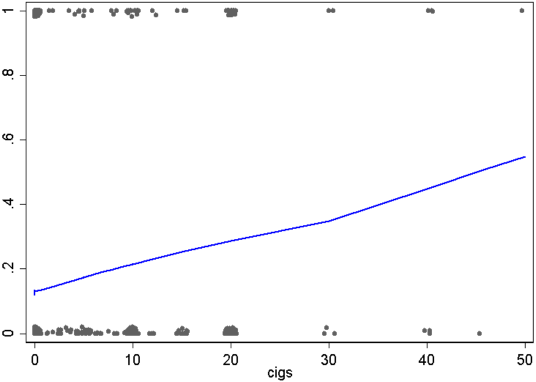
- We can also run our trusty linear model with the caveat that SEs are likely not right (but probably close) and that since low birth probability is (relatively) low we should be extra careful
- Now, in the probability scale, an extra cigarette increases the probability of low birth weight by 0.7%. With 10 cigarettes, 7% (but that assumes a linear effect)

```
 reg lw cigs faminc motheduc, robust

Linear regression                               Number of obs     =      1,387
                                                F(3, 1383)        =       6.51
                                                Prob > F          =     0.0002
                                                R-squared         =     0.0212
                                                Root MSE          =     .35304

------------------------------------------------------------------------------
             |               Robust
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |    .007757   .0020677     3.75   0.000     .0037009    .0118131
      faminc |  -.0009345   .0005785    -1.62   0.106    -.0020693    .0002004
    motheduc |   .0005403   .0042972     0.13   0.900    -.0078895      .00897
       _cons |   .1531912   .0532648     2.88   0.004     .0487027    .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)

- Sketch: we will compute the numerical derivative of a variable X_j for **each observation** using the other covariates X_k **as they were observed**. We will change the value of a predictor by "a litlle 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** $\frac{\partial p}{\partial X_j}$ that starts by computing a small change for each observation
- 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
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dydx |      1,387    .0055768    .0012444   .0040507   .0113006
```

- An additional cigarette increases the probability of low birth weight by 0.56 percent points

Average Marginal Effect (AME)

- Replicate using margins command

```
. margins, dydx(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

   * Duplicte 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
restore
*  .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 a difference $\frac{\Delta Y}{\Delta X}$. What about 10 extra cigarettes?

```
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
   Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dydx |      1,387    .0064608    .0012196   .0048265   .0111532
```

- 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 is 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

```
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
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dydx |      1,387     .118284     .0105063   .0935378   .1363816
```

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)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoked |   .0988076   .0230959     4.28   0.000     .0535405    .1440748
------------------------------------------------------------------------------
qui logit lw i.smoked faminc motheduc, nolog
margins, dydx(smoked)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.smoked |    .118284   .0322576     3.67   0.000     .0550602    .1815078
------------------------------------------------------------------------------
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)

The margins command must be treated with **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**

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 continous 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)
- 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 males the value of the proportion in the sample, 0.6. In another sense, it seems odd because male is a dummy variable
- **Don't waste too much time thinking about this**. When we calculate **marginal effects** (not margins), 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
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dydx |      1,387    .0055624    .0010396   .0051876    .011267
```

- MEM not that different from AME

Marginal Effect at the Mean (MEM)

- Using the margins command

```
 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)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |    .005563   .0012843     4.33   0.000     .0030458    .0080801
------------------------------------------------------------------------------
```

- 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
   Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      dydx10 |      1,387    .0061672    .0010198    .005653   .0112164
      dydx40 |      1,387    .0052304     .001039   .0047327   .0111981
```

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

Average marginal effects                        Number of obs     =      1,387
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         |
         _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

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
cigs         |
         _at |
          1  |    .007757    .001631     4.76   0.000     .0045574    .0109566
          2  |    .007757    .001631     4.76   0.000     .0045574    .0109566
          3  |    .007757    .001631     4.76   0.000     .0045574    .0109566
          4  |    .007757    .001631     4.76   0.000     .0045574    .0109566
------------------------------------------------------------------------------
```

With interactions the effect should be more noticiable

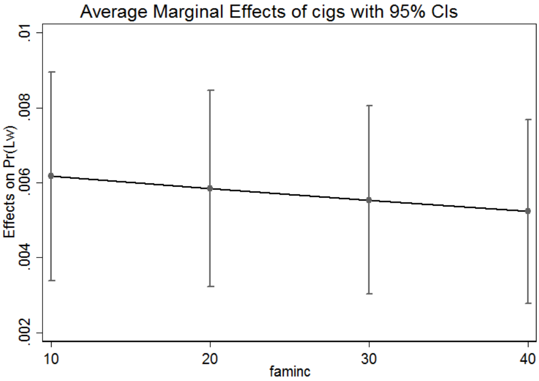
- Now adding interactions between cigarettes and income. This is the right way of making the effect of cigs depend on income
- Note how the conclusion is different. However, the **interaction is not statistically significant**

```
qui logit lw c.cigs##c.faminc motheduc, nolog
margins, dydx(cigs) at(faminc=(10 20 30 40)) vsquish
Average marginal effects                        Number of obs     =      1,387
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         |
         _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
------------------------------------------------------------------------------
```

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 $log(\frac{lw_i}{1-lw_i}) = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 motheduc_i$
- 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 and same intercept
- If we add interactions, the margins command is a life saver because effects are hard to interpret otherwise
- In particular, the **magnitude of effects** is hard to make sense when a logistic model has interactions and we use odds ratios for interpretations

Interactions

- As usual, interactions are easier to understand with two indicator variables. Let's say more than high school (hsp) and male as in previous examples: $log(\frac{p}{1-p}) = \beta_0 + \beta_1 hsp + \beta_2 male + \beta_3 hsp * male$
- 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 (What's β_1? β_2?)
- For males: $log(\frac{p_m}{1-p_m}) = \beta_0 + \beta_1 hsp + \beta_2 + \beta_3 hsp$
- For females: $log(\frac{p_f}{1-p_f}) = \beta_0 + \beta_1 hsp$
- Difference males - females: $log(\frac{p_m}{1-p_m}) - log(\frac{p_f}{1-p_f}) = \beta_2 + \beta_3 hsp$
- 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: $log(\frac{p_m}{1-p_m}) - log(\frac{p_f}{1-p_f}) = \beta_2 + \beta_3 hsp$
- Difference male - female for educated: $log(\frac{p_{me}}{1-p_{me}}) - log(\frac{p_{fe}}{1-p_{fe}}) = \beta_2 + \beta_3$
- Difference male - female for uneducated: $log(\frac{p_{mu}}{1-p_{mu}}) - log(\frac{p_{fu}}{1-p_{fu}}) = \beta_2$
- Difference in difference: $log(\frac{p_{me}}{1-p_{me}}) - log(\frac{p_{fe}}{1-p_{fe}}) - [log(\frac{p_{mu}}{1-p_{mu}}) - log(\frac{p_{fu}}{1-p_{fu}})] = \beta_3$
- So same as with linear model. In the log-odds scale, it is a **difference-in-difference**

Interactions in the odds scale

- $log(\frac{p_{me}}{1-p_{me}}) - log(\frac{p_{fe}}{1-p_{fe}}) - [log(\frac{p_{mu}}{1-p_{mu}}) - log(\frac{p_{fu}}{1-p_{fu}})] = \beta_3$
- We can apply the rules of logs and take $e^{()}$ on both sides: $\frac{\frac{P_{me}}{1-P_{me}}}{\frac{P_{fe}}{1-P_{fe}}} / \frac{\frac{P_{mu}}{1-P_{mu}}}{\frac{P_{fu}}{1-P_{fu}}} = e^{\beta_3}$
- 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 (see the numerator–it's the odds ratio for males females for those "educated" or hsp=1. The denominator is same but for those with hsp=0)
- **Interpreting the magnitude in the probability scale is a lost cause**. A lot easier with margins or stratifying the models

Interactions and stratification

- 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. **Back low birth babies**
- Make sure you understand this (below, you can replicate the 0.947913–the interaction–running two stratified models and using odds ratios)

```
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 |   .9045562   .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
```

Interactions and marginal effects

- What about if we use marginal effects instead? The model is: The model is $logit(lw) = \beta_0 + \beta_1 cigs + \beta_2 inc + \beta_3 cigs * inc$
- inc is a dummy and equal 1 if income is higher than 40k

```
margins, dydx(*)
Average marginal effects                        Number of obs     =      1,388
Model VCE    : OIM
Expression   : Pr(lw), predict()
dy/dx w.r.t. : cigs 1.inc
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |   .0046941   .0018482     2.54   0.011     .0010717    .0083165
       1.inc |  -.0288422   .0225222    -1.28   0.200    -.0729849    .0153006
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

- Wait, two effects? The model has three coefficients. Where is the interaction?

Interactions and marginal effects

- This may seem confusing but it's not when you remember how Stata calculates marginal effects
- For cigs, a continuous variable, it's using the two-sided derivative increasing cigs by a little bit and calculating predictions. It's **increasing cigs in both the main effect and the interaction**
- Then it takes an average so the marginal effect of cigs is the numerical derivative for both inc=1 and inc=0 combined
- For the marginal effect of inc, it's doing the same going from 0 to 1, averaging over the values of cigs
- To get what we need, which in this case is the marginal effect of cigs separately for inc=1 and inc=0, we have to be more specific

Interactions and marginal effects

- With interactions the effect of cigs depends on income

```
margins, dydx(cigs) at(inc=(0 1)) vsquish
Average marginal effects                          Number of obs     =      1,388
Model VCE    : OIM
Expression   : Pr(lw), predict()
dy/dx w.r.t. : cigs
1._at        : inc             =           0
2._at        : inc             =           1

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
cigs         |
         _at |
          1  |   .0062867   .0012881     4.88   0.000     .0037621    .0088113
          2  |  -.0004394   .0062301    -0.07   0.944    -.0126501    .0117713
------------------------------------------------------------------------------
```

- You need to remember this. **Be careful using the margins command when you have interactions**

Interactions and marginal effects

- Of course, interactions go both ways. So the effect of income depends on the number of cigs. But cigs is continuous; we have to choose some values

```
 margins, dydx(inc) at(cigs=(0 10 20 40)) vsquish
Conditional marginal effects                      Number of obs     =      1,388
Model VCE    : OIM
Expression   : Pr(lw), predict()
dy/dx w.r.t. : 1.inc
1._at        : cigs            =           0
2._at        : cigs            =          10
3._at        : cigs            =          20
4._at        : cigs            =          40
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0.inc        |  (base outcome)
-------------+----------------------------------------------------------------
1.inc        |
         _at |
          1  |  -.0114123   .0214935    -0.53   0.595    -.0535388    .0307143
          2  |  -.0851994   .0622866    -1.37   0.171    -.2072788      .03688
          3  |  -.1819228   .1224851    -1.49   0.137    -.4219893    .0581436
          4  |  -.4251388   .2397838    -1.77   0.076    -.8951064    .0448287
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

- Again, interaction was not significant so...

Digression

- We can of course estimate marginal effects for linear models without nonlinear terms but they will be boring

```
 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(*)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   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
------------------------------------------------------------------------------
```

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

A very brief summary of margins and the margins command

- Most common uses: **estimate “effects”** in the scale of interest. That is a 1) a numerical derivative for continuous covariates or 2) incremental effects for dummy variables. Syntax is “margins, dydx(varname)”
- Another possibility is to use margins to obtain “**predictive margins**” of dummy variables–and if you fix a continuous covariate at some values, for continuous variable as well. Syntax is “margins varname” or “margins varname, at(...)”
- With the previous syntax you can use margins to **obtain predictions**, Just specify values for all covariates: “margins, at(var1=10 var2=20 var3=...)”
- You can also use margins to obtain “**adjusted predictions**,” which is essentially the same ideas as previous point. You need to fix covariates at some values: margins, at(cigs=(0(1)50) faminc=50 motheduc=13)

A very brief summary of margins and the margins command

- **Confusion alert**: Make sure you understand the difference between marginal effects and predictive margins. I **guarantee** you are going to get confused
- Marginal effects (dydx) is about effects; the other is about calculating predictions but not effects. Yet, part of the confusion is that in order to calculate effects you also use predictions BUT changing values by a “small” amount or from 0 to 1
- We haven’t discussed other features but you can use the margins command to express effects as elasticities, for example
- The marginsplot has many options. It’s specially helpful to display interactions and understand the model
- You can produce adjusted plots as in the example in Excel using a reference population...

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
------------------------------------------------------------------------------
di exp(_b[_cons] + _b[smoked] + _b[faminc]*30) / (1+exp(_b[_cons] + _b[smoked] + _b[faminc]*30))
.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)

What about other models?

- Once you understand the logic of numerical derivatives is not difficult to understand what the margins command does when applied to other models
- By default, Stata calculates marginal effects in the **same units as the post-estimation command predict**
- For example, if you use the poisson command for counts, the default is to obtain predictions as counts, not rates
- So the margins command by default will provide effects in the counts scale

More custom: Two-part models

- Two-part models are often used to estimate cost data with a large proportion of costs that are zero
- The idea is to estimate the model in two parts:

  1) Estimate the probability that the cost is greater than zero conditional on covariates: $P(y_i > 0|X_i)$ (using logit, probit, complementary log-log models)

  2) For those observations with non-zero costs, estimate the expected costs conditional on covariates: $E(y_i|y_i > 0, X_i)$ (using Poisson, linear models, Gamma, $log(y)$, Box-Cox, etc)
- Predictions are obtaining combining both parts (multiplication)

Two-part models marginal effects

- If you know how to get predictions, you now how to calculate marginal effects
- First, using the twopm user-written command and the margins command

```
* Get data from example in twopm command
webuse womenwk
replace wage = 0 if wage==.


* First part logit, second part GLM with Gamma family
qui twopm wage i.married children, firstpart(logit) secondpart(glm, family(gamma) link(log))
margins, dydx(married)

Average marginal effects                        Number of obs     =      2,000
Expression   : twopm combined expected values, predict()
dy/dx w.r.t. : 1.married
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.married |   6.615493   .5323105    12.43   0.000     5.572183    7.658802
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Two-part models marginal effects “by hand”

```
gen     nozeroc = 0
replace nozeroc = 1 if wage >0
preserve
   clonevar marr  = married
   * not married
   qui logit nozeroc i.married children
   replace married = 0
   predict double fp0
   replace married = marr
   qui glm wage i.married children if wage > 0, f(gamma) l(log)
   replace married = 0
   predict double c0
   gen chat0 = fp0*c0

   * married
   replace married = marr
   qui logit nozeroc i.married children
   replace married = 1
   predict double fp1
   replace married = marr
   qui glm wage i.married children if wage > 0, f(gamma) l(log)
   replace married = 1
   predict double c1
   gen chat1 = fp1*c1

   gen dif = chat1 - chat0
   sum dif
restore
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         dif |      2,000    6.615493    1.660084   2.485561   8.001396
```

Summary

- Marginal effects come to the rescue when we need to interpret non-linear models
- Different terms for these types of effects
- What we did today was about **PREDICTION**, 100 percent. We use predictions to interpret models using numerical derivatives
- 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**