Working in the margins to plot a clear course

Bill Rising

StataCorp LP

10th German Stata Users Group Meeting
Berlin, Germany
1. Jun 2012

Goals

- Learn about predictive margins and marginal effects
- Learn about making nice graphs to help explanations

Getting Ready

- We'll be doing a lot of work with categorical variables
- By default, Stata does not show the base reference class when using factor variables
- Let's fix this

```
. set showbaselevels on, permanently
```

Starting Simple

- We will start with the low birthweight dataset from Hosmer and Lemeshow's book on logistic regression

```
. webuse lbw
```

- Let's start simple with an easy linear regression for birthweights

```
. regress bwt lwt i.smoke i.race
```

- We know what the coefficients mean
  - The coefficient for `1.smoke` says: all other things being equal (i.e. weight and race), we think smokers' babies are about 400g lighter than non-smokers' babies

Predicting at Specific Values

- Suppose we would like to predict the average birthweight for babys whose mothers weigh 130 lbs, are white and who smoke
- We could use `lincom` by building the regression equation

```
. lincom 130*lwt + 1.race + 1.smoke + _cons
```

What If We Have Partial Information?

- Now suppose that all we know about a mother-to-be is that she weighs 130 pounds
- What should we guess for her average baby's weight?
- There are two paths here:
  - We could average all the other covariates, and plug in the averages as above
    - Order: average, then predict with partial info
  - We could plug in 130 for the weight for all the women in our sample, predict the birthweights, and average the results
    - Order: predict with partial info, then average
- The latter path is called 'predictive margins' or 'average predictive margins'—it is the one we will take

Starting Predictive Margins

- Stata implements predictive margins using the `margins` command
- Here is what we would guess if all we knew was that the woman weighed 130 pounds:

```
. margins, at(lwt==130)
```

- Computationally, the point estimate could be computed by
  - Changing `lwt` to 130 everywhere
  - Using `predict` to get predicted values
  - Finding the mean of those values
- `margins` does more work so that it can find proper standard errors

Adding Information

- Now suppose that we had the full information that we had earlier
- We can get a predicted value here, too

```
. margins, at(lwt==130 race==1 smoke==1)
```

- The results look very similar to the results from `lincom`

Margins Across Multiple Values

- One of the nice things about `margins` is that it can be used get average predicted values over a range
  - This, in some way, is a way to see a variable's effect over a range of values
- So, for example, if we wanted to show someone how birthweights change by weight, we could make the following table

```
. margins, at(lwt==(100(10)200))
```

- This is a little underwhelming, because it is not very readable
  - Still, you can see that the differences in the margins are about 40g, which is 10 times the coefficient of `lwt`

Graphing Predictive Margins

- Stata has a command which can draw graphs arising from the `margins` command: `marginsplot`
- Here is a simple example

  ```
  . marginsplot
  ```

- This being linear regression the graph is a straight line with a slope of about 4g/lb matching the coefficient of `lwt`

Some `marginsplot` Details

- By default `marginsplot` draws confidence bars for the mean
  - These can be shut off with the `noci` option
- By default `marginsplot` draws the predicted points
  - These can be shut off with the `recast(line)` option
- We'll see these below when the graphs become busy

What About Comparing Groups

- Suppose we would like to compare weights across the three race categories
- This is done by including race in the `varlist` for the `margins` command:

```
. margins race
```

  - Aside: this is equivalent to the following

```
. margins, at(race==1) at(race==2) at(race==3)
```

- Once again, these values can be interpreted as best-guesses for partial information for your sample

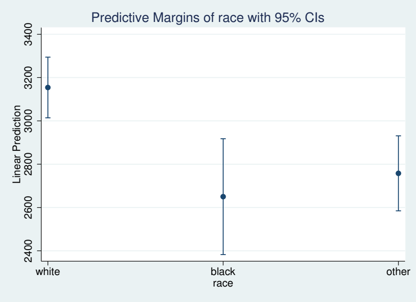
Differences in Margins and Coefficients

- For a linear model, there should be an easy correspondence between differences in the margins and the coefficients
- Here, the difference between the margin for white and the margin for black is roughly 500g, which matches the coefficient for `2.race`
- Likewise, the difference between the margin for white and the margin for "other" is roughly 400g, which matches the coefficient for `3.race`

Graphing the Group Means

- We can graph these using marginsplot

```
. marginsplot, recast(scatter)
```



- The recast(scatter) option squelches the connecting lines

A Richer Dataset

- Now will switch over to the `nhanes2` dataset

```
. webuse nhanes2
```

- It is a nice dataset, because it has good things for both linear and logistic regression
- These are survey data

```
. svyset
```

  - We will need to use the `svy:` prefix for estimation
- We will model body mass index (BMI)

$$\text{bmi} = \frac{\text{weight in kg}}{(\text{height in m})^2}$$

- We will then model the chance of having diabetes

A Simple Model for BMI

- We can start off with a simple model for predicting BMI

```
. svy: regress bmi age sex
```

- We can make this more interesting by using interactions

```
. svy: regress bmi c.age##i.sex
```

- Interpreting the coefficients gets a little more wordy here
  - For men, bmi increase by 0.036 for each year of age
  - For women, bmi is lower than for men by 2.2, but increases more rapidly, at $(0.036 + 0.042) = 0.078$ per year

Using Margins to Visualize

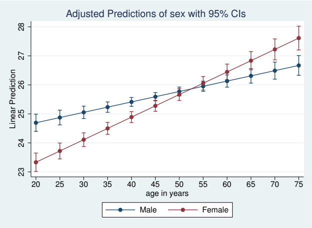
- We can get a better picture of what happens here by looking at margins

```
. margins sex, at(age==(20(5)75)) vce(uncond)
```

  - The `vce(unconditional)` option should be used with survey data

- The graph is now more interesting

```
. marginsplot
```



Making for More Complex Models

- Here is a much more complex model

```
. svy: regress bmi c.age##c.age##i.sex i.sex##i.race
```

  - We can see that age creates a concave-down parabola
  - We can see that the parabolas for women are likely not as sharp
- Let's find the margins here

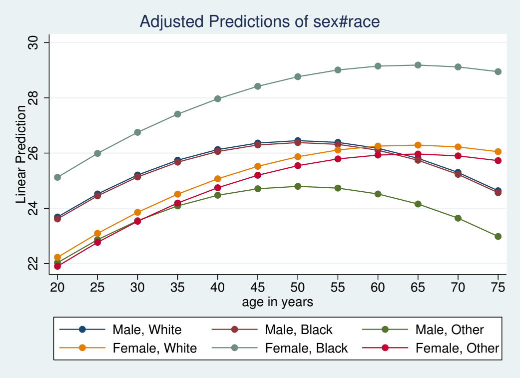
```
. margins sex#race, at(age==(20(5)75)) vce(uncond)
```

- Quite the messy table!

A Nicer Picture

- We can do better with a picture

```
. marginsplot, legend(rows(2)) noci
```



Looking at Diabetes

- Now we would like to look at the chances of having diabetes
- Here is a simple model

```
. svy: logistic diabetes age i.sex i.race bmi
```

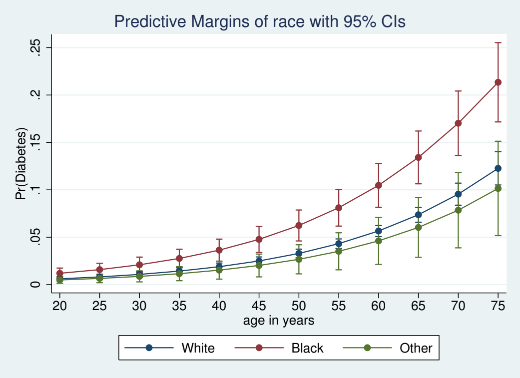
- We can see that age and bmi both increase the odds of diabetes by about 6% for each unit increase
- How does this play out for race and age?

```
. margins race, at(age==(20(5)75)) vce(uncond)
```

Here Is the Picture

- We can get a nice picture

```
. marginsplot, legend(rows(1))
```



This is Better than Odds Ratios

- This type of graph is something that makes explaining a logistic model much easier than via odds ratios
- It is as applicable to the general population as much as your belief that your sample is representative of the general population
    - Which is important for the odds ratios also
- Here, a picture is worth a thousand hard words

For Probit Fans

- If you prefer probit models, we can use the same type of logic

```
. svy: probit diabetes age i.sex i.race bmi
```

  - Now the coefficients are not very interpretable

- We can still get margins

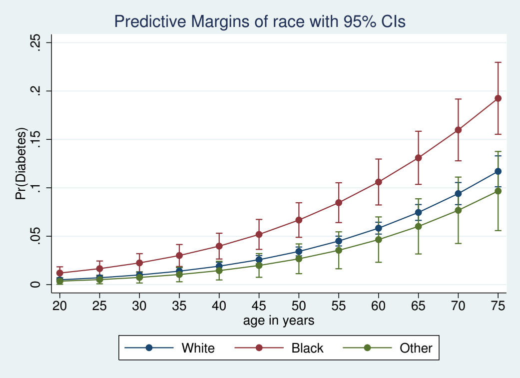
```
. margins race, at(age==(20(5)75)) vce(uncond)
```

- Creating the predictive margins still works the same

Picturing a Probit

- We can still get a very similar nice picture

```
. marginsplot, legend(rows(1))
```



A Very Messy Model

- Just to drive the point home, here is a very messy model

```
. svy: logistic diabetes c.age##c.age##(sex race) c.bmi
```

- We still can get the same margins

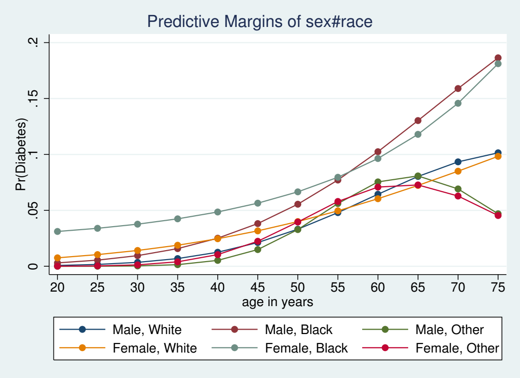
```
. margins sex#race, at(age==(20(5)75)) vce(uncond)
```

- Again, nothing really new

The Complex Picture

- The graph now is more complex, but still viewable

```
. marginsplot, noci legend(rows(2))
```



Marginal Effects

- Sometimes changes are nicer to talk about than the predicted values
  - The effect of smoking is..., instead of showing predictive margins for smokers and non-smokers
- For linear regression, this is simple—look at the coefficients
- For a non-linear model, it is less simple
  - We now need derivatives
- This is the concept of `marginal effects`

Looking at Marginal Effects

- The `margins` command can be used for marginal effects as well as for predictive margins
    - These are truly average marginal effects
- Mechanically, there is not much different—all we need to do is to specify what derivatives we want

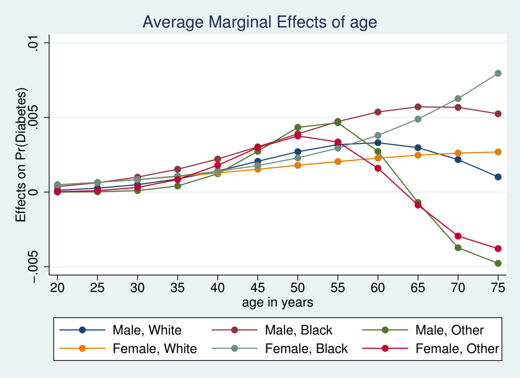
One Marginal Effect Example

- Here is one example based on the preceding example

```
. margins sex#race, at(age==(20(5)75)) vce(uncond) dydx(age)
```

- Here is the graph

```
. marginsplot, noci legend(rows(2))
```



Conclusion

- `marginsplot` is very good for showing how models work
- This can be used to good effect when explaining even simple non-linear models
- This can be used to good effect when explaining any type of interactions