INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
sustainable solutions for ending hunger and poverty
Supported by the CGIAR

ETHIOPIAN DEVELOPMENT RESEARCH INSTITUTE

Ethiopian Strategy Support Program II (ESSP II)

Applied Microeconometrics

Course Syllabus

and Exercises

David Stifel

Ethiopia Strategy Support Program II (ESSP II)

ESSP II Training Manual

November 2011

IFPRI-HEADQUARTERS
INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
2033 K Street, NW • Washington, DC 20006-1002 USA
Tel: +1-202-862-5600 • Skype: IFPRIhomeoffice
Fax: +1-202-467-4439 • E-mail: ifpri@cgiar.org

IFPRI-Addis Ababa
http://essp.ifpri.info
IFPRI c/o ILRI
P.O. Box 5689, Addis Ababa, Ethiopia
Tel: +251 11 6 17 25 55 Fax: +251 11 6 46 23 18
E-mail: ifpri-addis@cgiar.org
Contact: Bart Minten, Senior Research Fellow and Program Leader

ETHIOPIAN DEVELOPMENT RESEACH INSTITUTE
http://www.edri.org.et/
Blue Building • Addis Ababa Stadium
P.O. Box 2479 • Addis Ababa, Ethiopia
Tel: +251 11 5 50 60 66; +251 11 5 52 53 15
Fax: +251 11 5 50 55 88
Email: exe-director@edri.org.et

Applied Microeconometrics

Prof. David Stifel

IFPRI – ESSP
stifeld@lafayette.edu

Seminar Description:

The purpose of this intensive two-week seminar is to familiarize participants with the application of econometric techniques commonly used by microeconomists. The emphasis is on specification, estimation, interpretation, and testing of microeconometric models rather than a thorough treatment of asymptotic properties of estimators. Methods considered include instrumental variables estimators, difference-in-differences methods, limited dependent variable models, quantile regressions, non-parametric regressions, and panel data estimators. An emphasis will be placed on application through data-intensive assignments.

Learning Outcomes:

After completing this seminar, you should be able to:

- Understand and use basic econometric tools in the analysis of cross-section and panel data.
- *Appropriately interpret* parametric and nonparametric regression results
- Understand the implications of stratified sampling and the methods used to address them.
- Understand and use econometric tools in the analysis of limited dependent variables
- Identify issues that complicate program evaluations and employ methods of addressing these issues.

Software:

We will use Stata (pronounced *stay-tuh*).

Texts:

1. Angus Deaton, *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Johns Hopkins University Press, 1997.
   (Available as a PDF: http://www-wds.worldbank.org/external/default/main?pagePK=64193027&piPK=64187937&theSitePK=523679&menuPK=64187510&searchMenuPK=64187283&siteName=WDS&entityID=000009265_3980420172958)

2. Joshua Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

The relevant readings for each meeting time will be photocopied and/or provided to participants on a CD.

Empirical Exercises:

There will be daily homework assignments that will include empirical exercises (working with data using Stata). These assignments will give you a hands-on opportunity to apply the concepts discussed in class. This is one of the keys to learning econometrics. The other is discussing econometric issues with your colleagues. Consequently, you are encouraged to collaborate with other participants in the class. But, you must write your own solutions/code on your own, and you should keep in mind that it is in your best interest to not rely too heavily on your study partners/groups. Be sure that you understand and can present the solutions to the problems on your own.

Readings & Seminar Outline:

Day 1: Monday, November 28th:

Topic: **Introduction: Review of Econometrics & Overview of Empirical Research Methods & Design**

Reading: Angrist & Pischke – Chapters 1 & 2 (pp. 3-24)

Exercise: Exercise 1 – Stata Review

Day 2: Tuesday, November 29th:

Topic: **Survey of Microeconometric Data & Issues**

Reading: Deaton – Chapter 1 (pp. 7-18, 22-24, 40-58, 129-131)

Exercise: Exercise 2 – Poverty Estimates

Day 3: Wednesday, November 30th:

Topic: **Statistical Approaches (Nonparametric Regression & Density Estimates)**

Reading: Deaton – Chapter 3.2 (pp. 169-181) and 3.3 (pp. 191-202)

Levinsohn, James, and Margaret McMillan. 2007. "Does Food Aid Harm the Poor? Household Evidence from Ethiopia." In Ann Harrison, ed., *Globalization and Poverty.* Chicago: University of Chicago Press.

Barrett, Christopher, and Paul Dorosh. 1996. "Farmers' Welfare and Changing Food Prices: Nonparametric Evidence from Rice in Madagascar." *American Journal of Agricultural Economics*, 78(3): 656-669.

Exercise: Exercise 3 – Nonparametric and Parametric Estimation

Day 4: Thursday, December 1st:

Topic: **Regression Fundamentals & Model Specification**

Reading: Angrist & Pischke – Chapter 3.1 (pp. 27-40, 48-51), 3.4.1 (pp. 91-94)

Deaton – Chapter 2.1 (pp. 63-73), 2.2 (pp. 73-78)

Exercise: Exercise 4 – Model Specification

Day 5: Friday, December 2nd:

Topic: **Regression & Causality**

Reading: Angrist & Pischke – Chapter 3.2 (pp. 51-68)

Yamano, Takeshi, Harold Alderman, and Luc Christiaensen. 2005. "Child Growth, Shocks, and Food Aid in Rural Ethiopia." *American Journal of Agricultural Economics*, 87(2): 273-288.

Exercise: Exercise 5 – Regression

Day 6: Monday, December 5th:

Topic: **Bootstrapping and Heteroskedasticity & Quantile Regression**

Reading: Deaton – Chapter 1.4 (pp. 58-61) & 2.3 (pp. 78-85)

Angrist & Pischke – Chapter 7.1 (pp. 269-275)

Girma, Sourafel, and Abbi Kedir. 2005. "Heterogeneity in Returns to Schooling: Econometric Evidence from Ethiopia." *Journal of Development Studies*, 41(8): 1405-1416.

Exercise: Exercise 6 - Bootstrapping

Day 7: Tuesday, December 6th:

Topic: **Program Evaluation – Review of Concepts**

Reading: Ravallion, Martin. 2001. "The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluations." *World Bank Economic Review*, 15(1): 115-140.

Exercise: Exercise 7 – Quantile Regression

Day 8: Wednesday, December 7th:

Topic: **Instrumental Variables**

Reading: Angrist & Pischke – Chapter 4 (pp. 113-133)

Deaton – Chapter 2.6 (pp. 111-116)

Murray, Michael. 2006. "Avoiding Invalid Instruments and Coping with Weak Instruments," *Journal of Economic Perspectives,* 20(4): 111-132.

Acemoglu, Daron, Simon Johnson, and James Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91(5): 1369-1401.

Exercise: Exercise 8 – Instrumental Variables

Day 9: Thursday, December 8th:

Topic: **Categorical & Limited Dependent Variables**

Reading: Angrist & Pischke – Chapter 3.4.2 (pp. 94-107)

Exercise: Exercise 9 – Probit and Logit Models

Day 10: Friday, December 9th:

Topic: **Panel Data & Estimators**

Reading: Angrist & Pischke – Chapter 5 (pp. 221-247)

Deaton – Chapter 2.5 (pp. 105-111)

Dercon, Stefan. 2004. "Growth and Shocks: Evidence from Rural Ethiopia," *Journal of Development Economics,* 74: 309-329.

Exercise No. 1
Stata Review

Using the household roster dataset from the 1993 Vietnam LSMS (exercise_1_data.dta), answer the following questions. Note that the household roster dataset has one observation for each person in the household (not all are necessarily household members).

Record all of your Stata commands used to answer the questions in a –do– file.

1. From the household roster data, create a household-level dataset with the following variables (don't forget to keep the household id)…

   a. Number of adult household members (15+ years)
   b. Average age of household members
   c. Gender of the household head.

   Hint: You will want to use the `collapse` *command.*

2. From the household roster data, create an individual-level dataset with a variable indicating if an individual in the household is the most senior male in the household, where senior male is defined as…

   a. HH head if the HH head is male;
   b. Spouse of the HH head if HH head is female;
   c. Oldest male if neither (a) nor (b).

3. Using the anthropometric data from the same survey (exercise_1_data_2.dta), create a series of interactions between the urban dummy in the dataset and the following variables: `age, sex, bord, hhmemb,` and the region dummies.

4. Using the information from the household roster (exercise_1_data.dta), create a variable in the child-level anthropometric data set (exercise_1_data_2.dta) that indicates if the child's parent (father or mother) is the household head.

5. Repeat question 4, except this time create variables that indicate if (a) the mother is the household head, and (b) the father is the household head. Check that these variables are consistent with the parent-head variable that you created in question 4.

6. Using the dataset you created in question 5, run an OLS regression of child height-for-age z-scores (haz) on the child's age (age), gender (sex), indicators for mother or father being the household head, household per capita expenditure (pcexpend), an urban dummy (urban), and region dummies. Predict the fitted values (use `predict` after `regress`) and graph the haz-scores (scatter) and fitted values (line) against the fitted values.

Applied Microeconometrics IFPRI / EDRI
November 2011 Prof. David Stifel

Exercise No. 2
Poverty Estimates

Using the constructed dataset (exercise_2_data.dta) from the 1993 South Africa LSMS (South Africa Integrated Household Survey) along with the "Overview of the South African Integrated Household Survey" (exercise_2_data_overview.pdf), answer the following question.

1. The South Africa LSMS sample was designed to be self-weighted. What does this imply about the use of sampling weights (expansion factors) to estimates of means of variables in these data? *In practice, do you need to use sampling weights with these data?*

2. The sampling for this survey was also done in two stages (clusters and households). What does this imply about estimating means of variables and the standard errors of these means (not the standard deviations of the variables) in these data? *Do you need to do anything differently in estimating the mean of a variable, or the standard error of the mean?*

3. Estimate the mean of per capita monthly expenditures ($\overline{x}$) and its standard error (SE($\overline{x}$)) for the population of individuals. Show how controlling for two-stage sampling affects theses estimates. Provide some intuition for these differences/similarities.

4. Estimate the mean of per capita monthly expenditures ($\overline{x}$) and its standard error (SE($\overline{x}$)) for the sample of individuals. Explain how this is different from the mean and standard error for the population of individuals. *Do not worry about controlling for two-stage sampling for this question.*

5. Calculate the percent of the population that is poor using a poverty line of Rand 115.5 per month. Do you need to use sampling weights (expansion factors) to get an estimate of the population poverty rate? Why? If so, what is the appropriate weight?

6. Calculate the poverty rate using (a) household expansion factors and (b) individual expansion factors. What do your different estimates tell you about poor households compared to nonpoor households?

7. Show that an appropriately weighted mean of cluster means of per capita monthly expenditures is equal to the mean of per capita monthly expenditures for the population of individuals. In other words show that the average of group averages is equivalent to the overall average as long as you use the appropriate weights. Are the standard deviations of per capita monthly expenditures the same? Why?

   A key concept here is that means are additively separable. In other words, for example…

$$\frac{1}{4}(1+2+3+4) = \frac{1}{2}\left(\frac{1+2}{2}\right) + \frac{1}{2}\left(\frac{3+4}{2}\right)$$

Exercise No. 3
Nonparametric & Parametric Estimation

Using the constructed dataset on rice yields (exercise_3_data.dta) from the 2001 nationally representative Madagascar household survey, answer the following question.
Note: Do not worry about clustering for this assignment.

1. Estimate the mean and median rice yields among the population of households in the country and within each province. Put them all in a table.
   [*Hint: tabstat is a useful Stata command for this question. Type "help tabstat" in the Stata command window for syntax. Remember to think about the appropriate weights.*]

2. Estimate the kernel densities of log rice yields for the population of households in the country, in province 1, and in province 7. Plot them in a single graph. What additional information does this graph provide to your table from question 1?

3. Estimate a parametric (linear) regression for rice yields with travel time (a measure of remoteness) as the covariate.

4. Estimate and plot a nonparametric regression for rice yields with travel time (a measure of remoteness) as the covariate. Is this nonparametric regression consistent with the linear model you estimated in question 3? [*Note: "lowess" does not support weights.*]

Exercise No. 4
Model Specification

Using the constructed dataset on rice yields (exercise_4_data.dta) from the 2001 nationally representative Madagascar household survey from exercise 3, answer the following question.
Note: Do not worry about clustering for this assignment.

1. As we saw in question 4 of Exercise 3, the conditional expectation function (CEF) that we want to estimate might not always appear to be linear. One way to deal with nonlinearities without assuming a particular functional form, and while maintaining a model that is linear in the parameters, is to estimate a *saturated* model (see Angrist & Pischke, pp. 48-51). Estimate a regression for rice yields with a set of dummies representing travel-time quintiles. Is this model consistent with the nonparametric regression that you estimated in Exercise 3 question 4? [*Note: You may want to plot the predicted values (as a function of travel time – don't forget to sort the data by travel time). You may also want to determine the range of travel times in each quintile to help answer this last part of this question. "*`tabstat`*" may be a useful command for this purpose.*]

2. Another way to handle nonlinearities while maintaining a model that is linear in the parameters is to use a quadratic, or even a cubic. Estimate two additional regressions for rice yields – one in which travel time enters as a quadratic (X and X^2), and one in which travel time enters as a cubic (X, X^2 and X^3). Are either of these consistent with the nonparametric regression from Exercise 3 question 4?

3. Some analysts argue that agroecological conditions in the provinces differ considerably, and as such rice growing conditions and yields also differ by province. If you include dummy variables for the provinces, how does this affect your regression coefficients from the quadratic model in question 2? Based on intuition alone (we'll address tests later), does this permit you to interpret the time coefficients in a causal manner? If so, why? If not, why?

4. Graph the effects of travel time for each province from question 3 in one graph.

5. Note how adding province dummies in question 4 only results in shifts of the travel time-yield relationship (i.e. the only thing that changes is the constant). Suppose, however, that we believe that not only do the intercepts differ by province, but that the slopes differ as well. One way to check this is to run separate regressions (quadratic) for each province. Please do this. Another way to do this, and one that facilitates testing of differences in parameters, is to use interaction terms. Please estimate one regression in which you interact all of the province dummies with travel time and travel time squared. Do we get the same estimates with the one regression as we do with the province level regressions? Was our intuition about differing slopes correct?

Exercise No. 5
Regression

Using Stata's auto data (exercise_5_data.dta) that comes with the Stata program, answer the following questions.

In this exercise, we want to investigate how the prices of cars are determined. Our basic economic hypotheses are that *ceteris paribus* (a) larger cars are likely to be more expensive (they are more comfortable, etc.), (b) more fuel-efficient cars can be more attractive to customers and thus more expensive, and (c) foreign cars might have different pricings due to tariffs and/or strategic pricing.

1. Estimate an OLS model of car price (price) on size (weight), fuel efficiency (mpg – *miles per gallon*), and a dummy variable for foreign cars (foreign). Are our hypotheses correct?

There are a number of issues that might influence the validity of our statistical inference. These include (i) heteroskedasticity (inefficient estimates), (ii) collinearity/ multicollinearity (increases the variance of the estimates), and (iii) latent nonlinearity (affects normality of residuals and may lead to heteroskedasticity). Let's test for them.

Note that Stata has numerous regression diagnostic tools. You can check them out by typing...

```
help regress postestimation
```

2. Heteroskedasticity: Start by plotting the residuals versus fitted values using `rvfplot` (note that the option `yline(0)` draws a horizontal line at zero). Now conduct a Breusch-Pagan (`estat hettest`). Is this consistent with the graph? Finally, estimate the model again with the `robust` option and compare your results to the model estimated in question 1.

3. Collinearity/multicollinearity: Multicollinearity can be a problem when the regressors are dependent (at least, statistically). In our case, we might suspect that heavier cars consume more fuel (in absolute terms), and thus fuel efficiency (mpg) is likely to be correlated with weight. Using `correlate` or `pwcorr` (the latter lets you test the significance of the correlation), get a sense of whether this is a problem. Estimate the variance inflation factors (`estat vif`) to assess the influence of collinearity on the model estimates.

4. Nonlinearity & Model Specification Tests: Now let's consider nonlinearity and potential omitted variables using Ramsey's RESET test (`estat ov`). This tests whether non-linear combinations of the estimated values help explain the endogenous variable. The intuition behind the test is that, if non-linear combinations of the explanatory variables

have any power in explaining the endogenous variable, then the model is mis-specified. In addition, plot the prices against the predicted values (use `predict` after `regress`). Is this model mis-specified? Why might this be the case?

5. Now let's consider the residuals and the assumption of normality of the residuals (which can be obtained using `predict` after `regress`). Estimate the density of the residuals and compare it to the density of a normal distribution (type `help kdensity` to figure out how to plot the normal distribution). Does the distribution look normal? Conduct a formal test for normality (`sktest`). What do you find?

6. One issue might be the distribution of the dependent variable. Estimate the density of price and compare it to the normal distribution. Is there a way to transform the dependent variable to address the problem? Try doing so and see what happens to the residuals. Are there any other transformations to consider?

Applied Microeconometrics IFPRI / EDRI
November 2011 Prof. David Stifel

Exercise No. 6
Bootstrapping

Using the constructed dataset (exercise_6_data.dta) from the United States High School and Beyond survey answer the following question.

To complete this exercise, you will need to use Stata's `bootstrap` *command. The syntax for this command differs slightly depending on which version of Stata you are using. To determine the appropriate syntax for you, type* `help bootstrap` *in the Stata command window and proceed.*

1. Using `summarize, detail`, determine the median reading score (read). Now bootstrap the median using 500 replications (note, use `summarize, detail` with `bootstrap`; note also that `r(p50)` is how the median is saved after running `summarize, detail`). What extra information does this provide you about the median reading score?

2. Without setting the seed, run your bootstrap command from question 1 several times. Do you get the same median ("Observed Coef.") and bootstrapped standard errors? Why or why not? Now set the seed to a fixed number and run your boostrap command several times. What happens now? Why?

3. Estimate an OLS regression of reading scores (read) as a function of gender (female), math scores (math), writing scores (write) and socioeconomic status (ses – treat it as continuous here). Now use the `bootstrap` command to bootstrap the model (500 replications). Compare this with the standard OLS estimates. Finally, bootstrap just the gender parameter estimate.

4. Using the `saving` option in `bootstrap`, bootstrap the OLS regression from question 3 and save the parameter estimates in a temporary dataset. Open this dataset and using `kdensity` plot the distribution of parameter estimates for gender and socioeconomic status. Are these distributions consistent with the output in the table from the linear regression? Do they look normally distributed?

5. Using the Stata command `bsample`, write your own code to bootstrap the estimated parameter and t-statistic for gender (female) in the OLS regression from question 3. *Hint: You will want to use preserve and while-loops in which you bsample the data in each of 500 loops, estimate the model, save the parameter estimate and t-statistic in a dataset, and go on to the next loop.*

6. As in question 5, write your own code to bootstrap the average reading score, the average math score, and the ratio of the average reading score to the average math score. Is the ratio of the average scores in the bootstrapped sample the same as the average of the bootstrapped ratio?

Applied Microeconometrics IFPRI / EDRI
November 2011 Prof. David Stifel

Exercise No. 7
Quantile Regressions

Using the constructed dataset on rice yields (exercise_7_data.dta) from the 2001 nationally representative Madagascar household survey, answer the following question.

1. Estimate a linear regression of yields on a quadratic of travel time (i.e. travel time and squared travel time). How do you interpret the coefficients? What is the marginal effect of travel time on yields (i.e. $\partial y/\partial x$ if x = travel time)? Is the marginal effect statistically different from zero?

2. Estimate a quantile regression of yields on a quadratic of travel time through the 10^{th} percentile of the yield distribution. What is the marginal effect of travel time on yields for this model? Is it statistically different from zero? Is it economically different from zero?

3. Estimate quantile regressions of yields on a quadratic of travel time through the 10^{th}, 50^{th}, and 90^{th} percentiles of the yield distribution. Plot the predicted values for each of the quantile regressions and for the OLS regression (q.1) in a single graph. Provide an interpretation of the results from your regression output and graph.

4. Estimate quantile regressions of yields on a quadratic of travel time through the 10^{th}, 50^{th}, and 90^{th} percentiles of the yield distribution separately for province 2 and for province 7. Plot the predicted values for each of the quantile regressions in province-specific graphs. Provide an interpretation of the results from your regression output and graphs.

5. Consider a school feeding program that is intended to improve the nutritional outcomes of poor children (e.g height-for-age). Assume that the program is well targeted to the poor. Using conditional expectation notation, illustrate how a comparison of mean nutritional outcomes for program participants and nonparticipants can be decomposed into the average treatment effect and bias. Explain why the bias is likely to be negative? What is the source of the bias and how can it be addressed in order to determine the causal impact of the feeding program on nutritional outcomes.

Applied Microeconometrics IFPRI / EDRI
November 2011 Prof. David Stifel

Exercise No. 8
Instrumental Variables

Answer the following questions using the dataset (exercise_8_data.dta) from Angrist & Krueger's (1991) article examining the economic return to schooling in the United States. The emphasis of this exercise is on these returns to schooling.

1. Using OLS, estimate a model of log weekly wage (lwage) as a function of years of education (educ), race (race), marital status (married), area of residence (smsa), year of birth dummies (yob), and region dummies (region). What is the return to schooling? Interpret the coefficient on years of education.

2. Why might we be concerned that education is endogenous in this model? Explain. Can we not simply include as many observables on the right hand side as possible and appeal to our conditional independence assumption (CIA) to assert a causal relationship between schooling and wage earnings?

3. Angrist and Krueger's approach to this problem is to recognize that there is a relationship between the quarter of the year in which a child was born and the amount of schooling that he/she receives. Angrist & Pisckhe describe this on pp. 117-118, noting that most states require students to enter school in the calendar year in which they turn 6. This means that kids who are born late in the year (4th quarter) are young for their grade, while those born early in the year (1st quarter) are older for their grade. Combine this with schooling being compulsory only until a child's 16th birthday, and we have "a natural experiment in which children are compelled to attend school for different lengths of time, depending on their birthdays" (Angrist & Pischke, p. 118).

   So now estimate the model in question 1 using 2SLS (ivregress) in which quarter of birth (qob) dummies are the instruments. What does this model indicate the returns to schooling to be? Does your answer depend on whether you allow for heteroskedasticity (i.e. you use robust or not)? (Be sure to show the first stage regression as well.)

4. Angrist and Krueger actually use the interactions of quarter of birth dummies with year of birth dummies. They have many more instruments. Repeat what you did for question 3, but now use the larger set of instruments. Compare your results for the OLS model and your IV models. Is this what you would expect in terms of point estimates and standard errors? What do we expect to happen to the IV point estimates for the troublesome variable as we add more instruments? Is this what we see happening here?

5. Using the model from question 3, test for endogeneity of schooling (estat endogenous). Test the overidentifying restrictions (estat overid) and for weak instruments (estat firststage). What do you find?

6. Finally, to illustrate the consequence of weak instruments, we will use a dataset from the NLSY79 (exercise_8_data_2.dta). Start by estimating a simple OLS model of math test scores (pmath) as a function of an extreme obesity dummy (ex_obese). Then include a full set of covariates that are listed in the data between the gender dummy (girl) and a dummy for the year 2002 (yr_2002) (i.e. all variables girl-yr_2002).

   One might be concerned that even after including all of these observables, obesity may still be endogenous in the math test score model (there might be an unobserved characteristic of the child or mother that codetermines weight and test scores, or there might be a two-way causal relationship). So let's try estimating an IV model in which we use as instruments, an indicator for a sibling who is overweight (sib_over), an indicator for a sibling who is underweight (sib_under), mother's BMI prior to 1981 as a quadratic (bmi_mom_81 & bmi_mom_81_2), and fast food prices in the county as a quadratic (cpi_ffood & cpi_ffood_2). Are these legitimate instruments? How do they affect your point estimate for the effect of obesity on math test scores? Why? (Try testing for weak instruments.)

Exercise No. 9
Probit & Logit Models

Using the constructed dataset on child weight status (exercise_9_data.dta) from the 1998 round of the NLSY79 answer the following question.

An over-riding question in this application is the effect of fast food prices and breast feeding on child (ages 6-13) obesity.

1. Regress (OLS) the obesity dummy on fast food prices and the dummy variable indicating if the child was breastfed. Do we need to worry about either of these variables being endogenous? Why? If so, assume that there is selection on observables (i.e. other variables in the dataset) and appeal to the conditional independence assumption (CIA). Does this affect the point estimates for the two variables of interest (fast-food prices and breast feeding).

2. Estimate the same full model as in question 1 (assuming CIA), but now use probit and logit estimators. Does Amemiya's Rule roughly hold? Why is this so? Compare the marginal effects for the OLS regression from question 1 with the marginal effects from the probit and the logit models *evaluated at the mean* (Hint: use <mfx>). What is the general message that comes out of this comparison?

3. Compare the *average* marginal effects (Hint: use <margeff>) from the probit and logit (question 2) with the OLS regression (question 1). What is the general message that comes out of this comparison?

4. Including a dummy variable for girls in an OLS regression serves as a shifter and does not affect the marginal effects for fast-food prices and breast feeding (slopes). Because the probit model is non-linear, however, the girl dummy does affect the marginal effects in this model. Show how (a) the marginal effects evaluated at the mean of all other variables and (b) the average marginal effects in the probit model differ for boys and girls when a girl dummy is included in the model. (Hint: When you use mfx to estimate the marginal effects at the mean of all the other variables, use the at(…) option. When you use margeff to estimate the average marginal effect, do it separately for boys and for girls.)

   How does this differ from estimating the probit model separately for boys and for girls?

5. Using the sample of girls and the probit model that you estimated in question 4, plot the graph of observed y-values (obese) and predicted probabilities on the estimated aggregator index ($X'_i\hat{\theta}$). Does the graph look the way you expected it to? Why or why not?