Temple University

Department of Economics

Econometrics 8190

Final Exam, Fall 2011

**Answer Key**

**Question 1.** Temple University has hired you as an institutional research consultant to investigate the relationship between Scholastic Aptitude Test score (SAT) and freshman end-of-year grade point average (GPA). The university believes that GPA = g(SAT) is an increasing function and expresses the relationship that interests them. In each of the following situations name, specify, and defend the estimation procedure you would use. In situations where you do not select OLS, be sure to explain (using words, graphs, properties of estimators, and/or mathematics) why you do not do so. In all cases, try to be as clear and intuitive as possible. Except for #3.5, you may ignore non-SAT influences on GPA in your answers.

* 1. Temple University wants you to estimate the population regression function for freshman students in general and provides you with information from a simple random sample (SRS) of size 500 drawn from the population of all freshmen in a given year. This dataset gives GPA and SAT values for all 500 students. In the year in question, no students dropped out.

**A best answer would consider the following items of interest:**

* **Superficially we could just use OLS, but**
* **GPA is bounded to lie in the interval [0, 4.0] and thus we need to think about an efficient estimator.**
* **Just simple regression begs the question of omitted variables attenuation bias.**
* **Are there diminishing marginal returns to a higher SAT score?**
* **Is SAT score endogenous, since it is just a proxy for ability and intelligence and is therefore correlated with the error? This comes up in question 1.5.**
* **There should be some discussion of heteroscedasticity in this cross section. You could deal with this by discussing quantile regression. Note that this doesn’t address the question of diminishing marginal returns.**

**In questions 1.2 – 1.5 all these issues still apply, but you were not expected to bring them up again, except in 1.5.**

* 1. Temple University wants you to estimate the population regression function for freshman students in general and provides you with information from an SRS (simple random sample) of size 500 drawn from the population of all freshmen in a given year. This dataset gives SAT values for all students, but (for privacy purposes) only gives earned GPA values for the 425 students not on academic probation (i.e., those with GPA = or > 2.00). In the dataset, GPA for the 75 students on probation is recorded as 1.99. In the year in question, no students dropped out.

**This is a simple censored regression – Tobit question.**

**1.3** Temple University wants you to estimate the population regression function for

 freshman students in general and provides you with information only on the 425 non-

 probation students in an SRS of size 500 drawn from the population of all freshmen in

 a given year. This dataset gives GPA and SAT values for all 425 students. In the year

 in question, no students dropped out.

**This is a truncated regression problem. The model is** $GPA\_{i}=β\_{0}+β\_{1}SAT\_{i}+γM\left(β\_{0},β\_{1}\right)+u\_{i}$ **where M(.) is the inverse mills ratio.**

**1.4** Temple University wants you to estimate the population regression function for non-

probation freshman students and provides you with information only on the 425 non-

probation students in an SRS of size 500 drawn from the population of all freshmen in

a given year. This dataset gives GPA and SAT values for all 425 students. For the students on probation you have their SAT score, but not their GPA. In the year in question, no students dropped out.

**In this question you know who is omitted and why. The sample selection is exogenous since exclusion has been determined by the university. The appropriate estimator for this truncated regression, in which you have data on the exogenous truncation rule, you would use Heckman’s two step estimator.**

**1.5** Temple University wants you to estimate the population regression function for

 freshmen students in general and provides you with information from an SRS of size

 500 drawn from the population of all freshmen in a year in which there were some

 drop-outs. This dataset gives SAT values for all students, but only gives earned GPA

 values for the 440 students who completed both semesters of their freshman year

 (including both probation and non-probation students).

**On the surface this is also a truncated regression problem, but now there is a twist mentioned above. There is sample selection, but now it is endogenous. The student makes the exclusion decision, not the university. The decision to withdraw is usually made but low ability – low motivation students. Therefore, SAT is correlated with the error in the base model. Now you have to deal with the endogeneity of sample selection in your implementation of the Heckit.**

**Question 2:** In this question you are to revisit the homework on the productivity of Phuzicists. The academic labor market is unusual in many respects, one of which is that there are wide disparities in (research) productivity over the life cycle (within heterogeneity) and across individuals (between heterogeneity) and considerable concomitant salary dispersion. In this problem we consider detailed data on phuzicists to determine the nature of these life cycle effects. Since phuzics is a rather arcane field, there are only a few research journals, all of which are, essentially, equally irreputable. Therefore we will measure research productivity by the annual number of pages published. To evaluate joint publications we employ Thumb's First Rule that a coauthor of a joint paper of p pages with n authors deserves credit for $^{p}/\_{√n}$ pages.

In the sample we have phuzicists of several vintages employed in positions with a considerable range of research expectations. It is also plausible that the vintage of a phuzicist's PhD training may be an important determinant of productivity due to subsequent periods of rapid paradigm-shift. Fortunately, we observe productivity over a rather long period {the oldest phuzicist in the sample has a 1953 PhD}, and hence we are potentially able to distinguish vintage and life-cycle effects.

The data consists of a **panel** on individual phuzicists' research productivity, salary, and other characteristics. The raw data consists of observations on individuals; there are 11 variables:

i ~ person identifier

t0i  ~ year of Phd - 1900

t ~ current year - 1900

xi ~ gender (==1 if female)

dit ~ indicator of employment in research university (1 if person i is in a research position in year t, and 0 otherwise.)

yit ~ page equivalents in current year

ri ~ rank of PhD granting University

Yit ~ discounted cumulative page equivalents: We use Thumb's Second Rule that phuzics research depreciates at 8% per year.

sit ~ current annual salary in $1,000/yr

A relatively new aspect of the data is the availability of the variable ri which is the rank of the ith individual's PhD granting University at the time of the PhD. While such rankings are notoriously flawed, it is of obvious interest to investigate the extent to which individuals productivity depends upon the perceived quality of their PhD program. Note also however that since unobserved individual effects on productivity are likely to be also correlated with these ranks, this variable should probably not be treated as exogenous.

1. In the second paragraph of the question reference is made to vintage and lifecycle effects. Regardless of your persuasion, salt- or freshwater, your macro theory class included a discussion of vintage effects when considering capital in a growth model. Similarly, when discussing consumption expenditures in the aggregate or for a representative economic agent you discussed life cycle effects. Explain how these discussions inform your anticipated modeling o f scholarly productivity.

**Economics is a messy business. Stories in one sub-field often have relevance elsewhere and our thinking should be so informed. We don’t take a class, learn and retain the material long enough to pass an exam, and then forget the material; this is the undergrad student paradigm.**

**Vintage effects - Like physical capital in a macro growth model, human capital depreciates; the machinery wears out and there must be ongoing investment to keep it running. The same can be said for academic scholars. Hence, a young scholar may publish only 2 articles per year and an old scholar may publish 2 articles per year for quite different reasons. Additionally, a 2010 car is like a 1980 car only to the extent that both have four wheels and a steering wheel. There have been many exogenous and endogenous technology shocks over the last 30 years. Again, the same can be said in scholarship. At the start of my career a computer the size of a 3-ring binder and with many times the power of U of I’s IBM 360 was just fantasy.**

**Life cycle effects – The model of life cycle consumption is all about intertemporal substitution of consumption. In the context of this question there is a question of theintertemporal substitution of output. Relative to one’s peers, one’s output in, say, the fifth year of one’s career shouldn’t depend on the year in which you received your degree, ceteris paribus. Unlike a widget, your scholarly article is not purchased and consumed by a customer. Instead it enters the domain of infinitely lived public goods. In scholarship you have to decide when and how often to publish. With time your human capital depreciates, the marginal efficiency of your human capital declines; it gets harder to come up with an idea and execute it. Moreover, due to discounting, an article published early in your career is worth a LOT more than one published late in your career.**

1. In the initial homework assignment you considered a model of the form

$log\left(y\_{it}\right)=\sum\_{s=1}^{q}ρ\_{i,s}log\left(y\_{i,t-s}\right)+f\left(t,t\_{0,i},t-t\_{0,i},r\_{i}\right)+u\_{it}$.

(2.a) In the model as suggested, ability and the rank of the institution one attended are likely to be correlated. What is the implication for your OLS-type results in your homework? How would you overcome the problem? **This is an IV problem, NOT a fixed effects problem. The set of dummies you would need for fixed effects would not differ in a statistical way from the rank variable. If the dummies are school dummies, then the effect is identical. If the dummies are person dummies and more able people go to higher ranked schools then indirectly you are still have the same problem as the original one.**

(2.b) You were then asked to consider the shape of research productivity over the life cycle. Given the data as described, how might you use quantile regression to describe the shape of research productivity. **For this you would need the quantile regressions of Yt on the other variables, including experience. Then you would need to look at the graph of τ versus** $β\left(τ\right)$ **for the experience coefficient. People who are further into their career will correspond to the higher τ by construction.**

(2.c) What additional information would you learn about productivity from quantile regression that you did not learn from the least squares model you used in the homework? **Conditional on vintage we could construct the pdf for annual number of articles from quantile regression. OLS doesn’t show the shape of this distribution.**

(2.d) In the original homework you were asked about the 1980 paradigm shift. How would your quantile regression model of the previous two parts address this question? **Insert a dummy variable for pre- versus post 1980 degree and look at the** $β\left(τ\right)$ **for this variable. Alternatively, you could look at** $β\left(τ\right)$ **on the year of degree variable to with the corresponding OLS coefficient.**

1. In your homework you were asked to consider an earnings equation for phuzicists. The suggested equation was

$$log\left({s\_{i,t}}/{s\_{i,t-1}}\right)=γlog\left({Y\_{i,t}}/{Y\_{i,t-1}}\right)+other effects+v\_{i,t}$$

(3.a) What is the interpretation of γ? Think carefully before you start writing. **I put this question in because I didn’t like your answers to the question about productivity and lifetime earnings in the homework. The answer to this question is necessary in order to have answered the hwk question. The answer to this is that γ is the elasticity of salary GROWTH with respect to the growth of cumulative output.**

(3.b) How would you use kernel density (not regression) estimation to provide information about salary and vintage effects? **This is again a question about the density conditional on vintage versus the shape of the regression line (productivity over the life cycle).**

(3.c) What would be the advantage(s) of applying kernel regression methods to the above earnings equation? **You can see how your salary growth depends on the growth of the stock of published articles. An increase in the growth rate for low stock of articles growth scholars may differ from that for scholars whose discounted stock of articles is already growing rapidly. Incidentally, Dean Porat seems to have been aware of this all along. Under his regime persons whose growth in articles slowed suffered great penalties, but those whose stock articles was already high and published another never got much of a reward. From discussions with friends around the country this is an increasingly commonplace policy.**

(4) The distinguished statistician Emmanuel Parzen has argued that economics is gradually becoming more scientific because publication is becoming more concentrated in the hands of fewer researchers. Outline a “test” for this hypothesis for phuzics using either kernel density or kernel regression estimation.

**This is a question about the concentration of scholarship in the hands of the few. There are several ways to do this. The ones that come to mind right away are:**

1. **For each year rank order the data by number of articles published. Now construct the percent of articles published and percent of authors responsible for those articles. Plot this data in the % of authors - % of articles plane. Fit a kernel regression to these data pairs and you should get something that looks like a Lorenze curve. Do it for each year, then compare your fitted Lorenz curves across the years.**
2. **Aggregate the data to annual observations. Construct the empirical joint density over number of articles in year t and number of working PhDs in year t. Since each articles-scholars pair has a date attached to it you can see what is happening over time.**
3. **For each year compute the Herfindahl index. Fit a kernel regression of the sort Herfindahl = ao +b1\*time+u**