PROBLEM SET 5

General Instructions

Complete the following assignments in a sweave file that shows your code, output, and discussion. Hand in the compiled (pdf) version. You can use this file to get you started. You can work with others, but please hand in your own version. Please report any glitches as soon as you discover them - thanks!

For the theory questions, you may want to work out your answer on paper first, then type up the key steps in LaTeX. You can insert your LaTeXcode directly in in your sweave file, or - if you prefer, in a separate file or files in TeXnicCenter. For full credit, typed answers are required.

1. Q1: Serial correlation

(See scripts `mod4s3a` and `mod4s3b` for guidance) Consider Greene's gasoline consumption data (on the web under "gasoline" in tab-delimited .txt format). The variables are as follows:

(1) Year = Year, 1953-2004,
(2) GasExp = Total U.S. gasoline expenditure, in thousands
(3) Pop = U.S. total population in thousands
(4) GasP = Price index for gasoline,
(5) Income = Per capita disposable income,
(6) Pnc = Price index for new cars,
(7) Puc = Price index for used cars,
(8) Ppt = Price index for public transportation,
(9) Pd = Aggregate price index for consumer durables,
(10) Pn = Aggregate price index for consumer nondurables,
(11) Ps = Aggregate price index for consumer services.

The textbook analyzes a model using these data in the context of autocorrelation on pp. 649-650. Load the data into R, and specify Greene's model on p. 649 (6th edition), p. 927 (7th edition). Your dependent variable should be log[(GasExp)/(Pop*GasP)]. Your regressors should be:

(1) constant
(2) income = log(Per capita disposable income)
(3) GasP = log(Price index for gasoline)
(4) Pnc = log(Price index for new cars)
(5) Puc = log(Price index for used cars)
(6) time index

To create the last regressor (time index), you need to translate the year - variable into a running index from 1:52. Label this variable "t_i".

(1) Run a simple OLS model. (Note: Greene's results on p. 650 / 927 are a bit off, but close). Comment on the significance levels of each regressor (ignore the constant term). Are the signs of significant regressors as expected? Explain.

(2) Generate OLS residuals (call them "**e**") and plot them against time (year). The pattern should look a lot like the graph on p. 650 / 928. Does it indicate autocorrelation - why or

why not? If so, is it suggestive of positive or negative autocorrelation - explain.

(3) Perform a Breusch-Godfrey multiplier test for AR(1). State the null and alternative hypotheses, the computed p-value and your decision regarding the null (at 5% level of significance).

(4) Compute the Durbin-Watson statistic. State the null and alternative hypotheses, the appropriate degrees of freedom, the appropriate critical values from the DW Table at $\alpha = 0.05$ (textbook or google on the web), and your decision regarding the null. (the DW value should be the same as the one mentioned in Greene, up two the first 2 decimals).

(5) Estimate robust OLS with Newey-West corrected standard errors. (In `R`, you can use

```
L<-ceiling(n^(1/4))
```

for the lag indicator, where n is the sample size).

(6) Compute the Prais-Winsten FGLS estimator. (The results will be a bit different than those given in Greene - he uses a slightly different estimation approach).

(7) Compare your original OLS estimates, the robust estimates, and the FGLS results and elaborate:

   (a) Compare the s.e.'s and t-values between OLS and robust OLS. Are there any noteworthy changes in significance levels?
   (b) Compare the s.e.'s and t-values between the robust OLS and the FGLS model. Are there any noteworthy changes in significance levels?
   (c) Assume the main focus of your research is on the effect of "gas prices" on "gas consumption". Overall, which model would you choose?

2. Q2: Estimation of treatment effects via regression

This question uses home sales data from Connecticut (CT) for 1991-1999. All properties are single-family residential homes located within 0.25 miles of the coastline. The total sample size is 6,327. Of these, 2,439 are located in a spezial flood hazard area (SFHA), which means they have been declared to be at higher risk of flooding than the remaining 3,888 properties outside the SFHA. However, these homes are alse enjoying overall nicer amenities, such as proximity to beach, water, and views. Thus, it is ex-ante not clear which effect will dominate - the flood risk effect or the amenity effect.

The outcome variable of interest is sale price, in $1,000. The main objective of this exercise is to estimate the combined SFHA & amenity effect on home prices, and check which effect is stronger.

The data are sorted by treatment (SFHA=1 properties first, followed by SFHA=0 properties), and by sales date within treatment. The variables are as follows:

(1) DQid original property ID
(2) saleyr year of sale
(3) salemonth month of sale
(4) saleday calendar day of sale
(5) saledateE sale date in elapsed days since 1/1/1960
(6) price000 sale price in 1000's of 2014 dollars
(7) SFHA 1= located in SFHA zone
(8) age age of structure, years
(9) sqft00 square footage, in 100's
(10) lot000 lot size, in 1000 sqft
(11) bedrooms total number of bedrooms
(12) bathrooms total number of baths
(13) elev10 elevation in meters at 10 meter resolution
(14) ISMi distance to nearest Interstate, miles
(15) PAMi distance to nearest principal artery, miles
(16) beaMi distance to nearest beach, miles
(17) hidMi distance to nearest high-density development, miles
(18) coaestMi miles to nearest coast or estuary
(19) reslkMi miles to nearest lake, pond, or reservoir
(20) ag10 acreage of ag land w/in 1000m, most current
(21) ind10 acreage of ind. land w/in 1000m, most current
(22) op10 acreage of open land w/in 1000m, most current

The following loads in the data (including all variable names), and saves it immediately in R's internal ("rda") format:

```
R> data<- read.table('c:/Klaus/AAEC5126/R/data/CTfloodzones.txt',
 sep="\t", header=TRUE)
R> save(data, file = "c:/Klaus/AAEC5126/R/data/CTfloodzones.rda")
```

(1) Let the dependent variable be "price000," the treatment variable "SFHA,", and let the explanatory data $\mathbf{X}$ include all variables listed above from "age" to "op10" (15 variables), in

addition to a constant term where needed.

Check for overlap and show results in a table - which explanatory variables raise red flags by exceeding the recommended overlap core of 0.25 (in absolute terms)? Explain in words which group, treated or controls, has relatively larger or smaller values for these red flag variables.

(2) Estimate the ATT via difference in means, pooled regression adjustment, and regression adjustment using separate equations, deriving standard errors and t-values as in script `mod5s1`. Show all results in a combined table, as in the lecture script.

(3) Comment on your results - are they similar or not? Which effect appears to be stronger - the risk effect or the amenity effect? Which estimate would you pick if you had to choose among those three? Provide some rationale.

3. Q3: Estimation of treatment effects via matching

Use the same data as for Q2, and lecture script `mod5s3` for guidance. Use 1-neighbor matching ($M = 1$) without forcing exact matches ($Me = 0$), and use all variables in $\mathbf{X}$ described above for both matching and the regression adjustment.

(1) Find matches and check for overlap (="balance") - show the overlap results in a table. How do these scores compare to those from the unmatched data in Q2? Are there any noteworthy improvements (lower overlap score, in absolute terms) for some variables that would indicate that the data are now better balanced?

(2) Compute the uncorrected and corrected ATTs, along with consistent standard errors and t-values following script `mod5s3`. Report this output in a table.

(3) How does the corrected ATT compare to its uncorrected counterpart? Which one would you choose and why?

(4) Across all ATT estimates from Q2 and Q3, which one would you choose and why?

(5) In sum, what can you conclude regarding the challenge of estimating flood risk effects on home prices in presence of (unobserved) coastal amenities? Which effect is likely going to dominate? What additional data might help to directly control for amenities in the econometric model?