Lab 1. (2 academic hours). Nonparametric statistical tests

L1.1. Exercises

Ex.1.

The file miles.xls contains the data concerning the number of miles per unit of fuel consumed, for 12 pairs of cars of the same class. The fuel for the first car in each pair was "regular" (variable miles1), the fuel for the second car was with special additives (variable miles2). Is there a statistically significant difference between two types of fuel (select your own significance level)? To answer this question, apply the Kolmogorov-Smirnov test, Wilcoxon sign rank, sign test.

Ex.2.

Use the data of the file Change_Job.dta (part of RLMS, Round 15)

Variables:

changejob - answer to the question: "Do you want to change your job?" (0 - no, 1 - yes)

age - age of the respondent in 2006,

sex – sex of the respondent (1 – male, 2 – female),

income - how much money the respondent received in the last 30 days

boss - the answer to the question: "Do you have a subordinates?" (1 - yes, 2 - no)

Using the Wilcoxon rank sum test compare the average income of men and women, preferably one (or close) age (create your own variables).

Ex.3.

The file Youth_unemployment.xls file contains data on the unemployment rate in the group of 20-29 year olds in the seven federal districts of Russia in 1997-2008. Select to study for one year. Applying the Kruskal-Wallis test, check whether the average unemployment rate in all Federal Districts are the same?

L1.2. Non-parametric tests in the package STATA

Test	**Command**
Kolmogorov-Smirnov equality-of-distributions test	ksirnov varname, by (group)
A distribution-free rank sum test (Wilcoxon, Mann and Whitney)	For means equality: ranksum varname, by (groupvar) For medians equality: median varname, by (groupvar)
Wilcoxon signed rank test	signrank X=Y

	(where variables X and Y are determined in advance)
Fisher sign test	signtest X=Y (where variables X and Y are determined in advance)
Kruskal-Wallis equality-of-population (medians) rank test	kwallis varname, by(groupvar)

L1.2.The concept of programming in the R package

1. General information about the statistical package R. Installation. Reading files of various formats.
2. Structure of commands
3. The basic logical operators. Loops.
4. Estimation parameters of linear regression models and logit models.
5. A simple graphiques.

Information resource: http:///www.r-project.org

L1.3. Non-parametric tests in the package R

```
rm(list=ls()) #"cleaning" of all the variables
#if a library is not install, necessary to install it
#install.packages("name _of the library")
library(MASS)
library(Matching)
library(BSDA)
library(exactRankTests)
library(sqldf)
#Path to a file
#Examle of the filepath
filepath="C:/Users/Olga/Desktop/non_param/seminar1/data/"
#reading data from the file
data=read.table(paste(filepath,"miles.txt",sep=""), sep=";", header=TRUE)
#Kolmogorov-Smirnov test
ks.test(data$miles1, data$miles2, exact=FALSE)
#Bootstrap version of Kolmogorov-Smirnov test
ks.boot(Tr=data$miles1,Co=data$miles2, nboots=1000)
#sign test for equality of the median
SIGN.test(data$miles1,data$miles2)
```

```
# sign test for equality of the mean
groupvar binom.test(sum(data$miles1>data$miles2),dim(data)[1])
#Wilcoxon signed rank test without ties
wilcox.test(data$miles1,data$miles2)
# Wilcoxon signed rank test with ties
wilcox.exact(data$miles1,data$miles2)


#-----------------------------------------------------
#Reading of the data
data=read.table(paste(filepath,"change_job.txt",sep=""), sep=";", header=TRUE)
#Choose the data with necessary condition with the help of sql-request
x=sqldf("select income from data where sex=1 and age<=33 and age>=30 and income not like 'NaN'")
y=sqldf("select income from data where sex=2 and age<=33 and age>=30 and income not like 'NaN'")
# Wilcoxon signed rank test
wilcox.test(x$income,y$income)


#-----------------------------------------------------
# Reading of the data
data=read.table(paste(filepath,"unemp.txt",sep=""), sep=";", header=TRUE)
#Forming data for each district with the help of sql-requests
central=sqldf("select youthun from data where central=1")
nordwest=sqldf("select youthun from data where nordwest=1")
south=sqldf("select youthun from data where south=1")
volga=sqldf("select youthun from data where volga=1")
ural=sqldf("select youthun from data where ural=1")
siberia=sqldf("select youthun from data where siberia=1")
fareast=sqldf("select youthun from data where fareast=1")
# Kruskal-Wallis test
kruskal.test(list(central$youthun, nordwest$youthun, south$youthun,
volga$youthun,ural$youthun, siberia$youthun,  fareast$youthun))
```

Homework

Demonstrate the use of the above non-parametric tests in the analysis of your own data (you can use, for example, the data from RLMS (HSE site) or the site of Federal State Statistics Service (www.gks.ru)).

Lab. 2 (2 academic hours) Bootstraps

L2.1. Exercises

Ex.1.

The file beauty.dta contains data about 1260 Americans.

Variables:

wage - hourly wage,

belavg – indicator of bad appearance (belavg = 1 if the appearance is evaluated by experts below average and 0 otherwise)

abvavg – indicator of good appearance (abvavg = 1 if the appearance is evaluated by experts below average and 0 otherwise)

exper – years of workforse experience,

union – indicator of trade union membership,

goodhlth - indicator of health,

black - indicator of black race,

female- 1 for female and 0 for male,

married - indicator of family status,

south - indicator of living in south,

bigcity - indicator of living in big city,

smallcity - indicator of living in small city,

service - indicator of working in service industry,

educ – years of schooling.

Compare the estimates of mean wage for all individuals (or selected group)

a) sample mean b) bootstrap estimate c) jackknife estimate

Possible commands in STATA

```
summarize wage
bootstrap r(mean): summarize wage
jackknife r(mean): summarize wage
```

Ex.2.

Use the data of the file beauty.dta.

Test the hypothesis about the equity of hourly wage for two selected group of people with the help of

a) ordinary t-test b) bootstrap technique c) jackknife technique

Possible commands in STATA

```
ttest wage, by(abvavg)
bootstrap r(t): ttest wage, by(abvavg)
or
bootstrap t=r(t), rep(1000) strata(abvavg) saving(bsauto, replace): ttest wage, by(abvavg)
jackknife r(t): ttest wage, by(abvavg)
```

Ex.3.

Use the data of the file beauty.dta.

Estimate the dependence of hourly wage from the selected individual characteristics with the help of

a) OLS b) bootstrap technique c) jackknife technique

Possible commands in STATA

```
reg  wage abvavg black female married educ exper
bootstrap, reps(100):reg wage abvavg black female married educ exper
jackknife: reg  wage abvavg black female married educ  exper
```

Ex.4.

Use the models from the previous exercise.

Estimate and compare the different types of bootstrap confidence intervals for all regression coefficients, namely

(N) normal confidence interval

(P) percentile confidence interval

(BC) bias-corrected confidence interval

Possible commands in STATA

```
bootstrap: reg wage abvavg black female married educ exper
estat bootstrap, all
```

L2.2. Bootstrapes estimates in the package R

For ex.1

```
rm(list=ls())
#"cleaning" of all the variables
#if a library is not install, necessary to install it
#install.packages("name _of the library")
library("foreign")
library("boot")
```

```
library("sqldf")
#File path
filepath="C:/Users/Olga/Desktop/non_param/seminar2/data/"
#reading data from the file
data=read.dta(paste(filepath,"beauty.dta",sep=""))
#sample mean
mean(data$wage)
#Bootstrap function
boot.mean=function(data,i)
{
  mean(data[i])
}
#Bootstrap
boot(data=data$wage, statistic=boot.mean, R=10000)

#Bootstrap "by hands"
#Number of replication
R=1000
#Initialize the array with random data
boot_array=c(0,1)
#Set up the technical variable index
data$num=seq(1:dim(data)[1])
# There is a generation of new samples in the cycle, mean calculation and stores the result
for (i in (1:R))
{
  boot_array[i]=mean(data$wage[sample(data$num,dim(data)[2])])
}
#Sample mean
mean(boot_array)
#---------------
```

For ex.2

```
#Ordinary t-test
t.test(wage~married, data=data)
#Bootstrap t-test
R=1000
boot_array=c(0,1)
```

```
for (i in (1:R))
{
  boot_array[i]=t.test(wage~married, data=data[sample(data$num, dim(data)[1], replace=TRUE),])$statistic
}
mean(boot_array)

#----------------------------
```

For ex.3

```
#OLS
coef(lm(wage~female+married+educ+exper, data=data))
#Create a storage for regression coefficients after each bootstrap replication
R=999
res=data.frame(t(coef(lm(wage~female+married+educ+exper, data=data[sample(data$num, dim(data)[1], replace=TRUE),]))))[1,]
for (i in (1:R))
{
  res=rbind(res,data.frame(t(coef(lm(wage~female+married+educ+exper, data=data[sample(data$num, dim(data)[1], replace=TRUE),]))))[1,])
}
#Bootstrap estimates of the coefficients
colMeans(res)
#--------------------------------------
```

For ex.4

```
#Create special function
rsq=function(formula, data, indices) {
  d=data[indices,] # allows boot to select sample
  fit=lm(formula, data=d)
  return(coef(fit)[-1])
}
#Save the bootstrap results
results=boot(data=data, statistic=rsq, R=1000, formula=wage~female+married+educ+exper)
#Three types of confidential intervals
boot.ci(results, type=c("norm", "basic", "perc"))
```

Homework

Demonstrate the use of the bootstrap and jackknife method of estimation in the analysis of your own data.

Lab. 3 (2 academic hours) Kernel density estimation

L3.1. Exercises

Ex.1.

The file clothing.dta contains data about 400 Dutch men's fashion stores. . These data were taken from the site http://wileyeurope.com/go/verbeek2ed.

Variables:

tsales – annual sales in Dutch guilders,

sales - sales per square meter,

margin – gross-profit-margin,

nown – number of owners (managers),

nfull – number of full-time workers,

npart - number of part-timers,

naux – number of helpers (temporary workers),

hoursw – total number of hours worked,

hourspw – number of hours worked per worker,

inv1 – investment in shop-premises,

inv2 - investment in automation,

ssize – sales floorspace of the store (in m2),

start – year start of business.

Estimate the density function of the variable sales per square meter

Possible commands in STATA

kdensity sales

Ex.2.

In the previous section epanechnikov kernel was selected by default. Estimate the density function using the other kernels.

Possible commands in STATA

kdensity sales, kernel(biweight)

Possible kernels

epanechnikov - Epanechnikov kernel function; the default

epan2 - alternative Epanechnikov kernel function

biweight - biweight kernel function

cosine - cosine trace kernel function

gaussian - Gaussian kernel function

parzen - Parzen kernel function

rectangle - rectangle kernel function

triangle - triangle kernel function

For the comparison:

Possible commands in STATA

```
kdensity sales, kernel(epanechnikov) nograph generate(x epan)
kdensity sales, kernel(parzen) nograph generate(x2 parzen)
kdensity sales, kernel(gaussian) nograph generate(x3 gaussian)

label var epan "Epanechnikov density estimate"
label var parzen "Parzen density estimate"
label var gaussian "Gaussian density estimate"

line epan parzen gaussian x, sort ytitle(Density) legend(cols(1))
```

Ex.3.

In previous cases, the optimal window bandwidths were chosen from the condition of minimizing the integral squared error, provided that the initial data are normally distributed and the kernel is a normal (Gaussian). Check, for example, using the Kolmogorov-Smirnov (or Shapiro - Wilk) test whether you can take your variable normally distributed.

Possible commands in STATA

```
summarize sales
ksmirnov sales = normal((sales -r(mean))/r(sd))
```

Ex.4.

Experiment with the width of the window, first by setting the bandwidth automatically, and then increase and decrease it.

Possible commands in STATA

```
kdensity sales, kernel(biweight)
kdensity sales, kernel(biweight) bw(800)
kdensity sales, kernel(biweight) bw(600)
```

L3.2. Kernel density estimation in the package R

```
#"cleaning" of all the variables
rm(list=ls())
#if a library is not install, necessary to install it
#install.packages("name _of the library")
```

```
library(foreign)
library(stats)
library(sqldf)
#File path
filepath="C:/Users/Olga/Desktop/non_param/seminar3/data/"
#reading data from the file
data=read.dta(paste(filepath,"clothing.dta", sep=""))
#Graph of the density function with different kernels
plot(density(data$sales, kernel="epanechnikov"))
plot(density(data$sales, kernel="gaussian"))
plot(density(data$sales, kernel="rectangular"))
plot(density(data$sales, kernel="triangular"))
plot(density(data$sales, kernel="biweight"))
plot(density(data$sales, kernel="cosine"))
plot(density(data$sales, kernel="optcosine"))
#Test the hypothesis about normal distribution for variable sales
shapiro.test(data$sales)
the density function using the other kernels.
# Graph of the density function with different bandwidth
plot(density(data$sales, kernel="biweight"))
plot(density(data$sales, kernel="biweight", bw=10000))
plot(density(data$sales, kernel="biweight", bw=400))
```

Homework

Demonstrate the use of kernel density estimation in the analysis of your own data.

Lab. 4 (2 academic hours) Kernel regression

L4.1. Exercises in Stata

Ex.1.

The file clothing.dta contains data about 400 Dutch men’s fashion stores. . These data were taken from the site http://wileyeurope.com/go/verbeek2ed.

Variables:

tsales – annual sales in Dutch guilders,

sales - sales per square meter,

margin – gross-profit-margin,

nown – number of owners (managers),

nfull – number of full-time workers,

npart - number of part-timers,

naux – number of helpers (temporary workers),

hoursw – total number of hours worked,

hourspw – number of hours worked per worker,

inv1 – investment in shop-premises,

inv2 - investment in automation,

ssize – sales floorspace of the store (in m2),

start – year start of business.

Estimate the dependence of sales per square meter from the total number of hours worked or the size of the store with the help of local-polynomial regression.

Possible commands in STATA

lpoly sales hoursw

Ex.2.

In the previous paragraph the kernel (epanechnikov), the degree of the polynomial (0), the width of the window were selected automatically. Experiment with the choice of kernel, the degree of the polynomial, the width of the window.

Possible commands in STATA

lpoly sales hoursw

lpoly sales hoursw, kernel (gaussian) degree(1)

lpoly sales hoursw, kernel (gaussian) degree(2)

lpoly sales hoursw, kernel (gaussian) degree(3)

lpoly sales hoursw, kernel (gaussian) degree(1) bw(20)

Ex.3.

Estimate the dependence of sales per square meter from the total number of hours worked or the size of the store with the help of lowess regression.

Possible commands in STATA

lowess sales hoursw

Ex.4.

Experiment with the choice of span (share of the points for estimation, by default 0.8)

Possible commands in STATA

lowess sales hoursw, bwidth(0.2)

Ex.5.

Compare the results of the polynomial regression and lowess graphically

Possible commands in STATA

graph twoway (scatter sales hoursw) (lowess sales hoursw, bwidth(0.4)) (lpoly sales hoursw, kernel (gaussian) degree(1))

L4.2. Exercises in R

The data and some parts of this exercise were taken from the book Keele L. "Semiparametric Regression for the Social Sciences", John Wiley&Sons, Ltd., 2008.

Use the dataset jacob.dta (on behalf of the author, whose data are used in the exercise.).

History of the problem. "The example databcome from Jacobson and Dimock's (1994) study of the 1992 US House elections. In the study of American politics, one well known regularity is that Congressional incumbents tend to be reelected. In the 1992 House elections, many incumbents were defeated, and Jacobson and Dimock explore the factors that contributed to the unusually high number of incumbents that were defeated that year. They argue that dissatisfaction with Congress was high due to a weak economy and a number of Congressional scandals. While they focus on a number of different predictors for the success of challengers, one indicator that they use to measure dissatisfaction with Congress is the percentage of the vote for H.Ross Perot in the 1992 presidential election. The district level vote between the president and members of the House is highly correlated, and Jacobson and Dimock test whether the level of support for Perot in this election increased support for challengers. For many voters, Perot served as a protest vote which may indicate dissatisfaction with incumbents of either party". "Jacobson and Dimock also try to isolate whether the House Bank scandal also contributed to the strong showing by challengers in 1992. In the House bank scandal, it was found that members of the House were often overdrawing their accounts at the House Bank. To measure the effect of the House Bank scandal, Jacobson and Dimock use a measure that records the number of overdrafts by each member of the House".

Variables:

chal_vote – the challenger's vote share Perot in each Congressional district in the 1992 general election,

perotvote – the percentage of the vote for H.Ross Perot in each Congressional district in the 1992 general election,

checks_raw – number of overdrafts,

exp_chal – 1 if the challenger had held elective public office, 0 otherwise, chal_spend – challenger's spending,

inc_spend – imcumbent's spending,

pres_vote – challenger's party's presidential vote,

marginal – marginal in 1990,

partisan_redist – partisan redistriction.

Use perotvote as the dependent variable in all models.

1) Use a scatterplot to study the relationship between variables perotvote and chal_vote. Does their appear to be a statistical relationship between the two variables? Does it appear to be linear or nonlinear?
2) Estimate linear regression.
3) Estimate loess models with different spans и various kernels.
4) Calculate 95% confidence interval bands and plot the relationship with CI bands for the estimated loess models.
5) Test whether the relationship is statistically significant and whether there is a significant amount of nonlinearity.
6) Is the nonparametric fit better than either a logarithmic or quadratic transformation of chal_vote ?
7) Add to the loess model variable checks_raw and plot the joint nonparametric effect.

Possible commands in R could be found in the script Smoothing R.

Homework

Demonstrate the use of a kernel regression in the analysis of your own data .

Lab. 5 (2 academic hours) Splines and Cross-Validation

L5.1. Exercises with Splines

Use the dataset jacob.dta. The description of data and variables is given in Lab 4 (L4.2).

Ex.1.

Estimate cubic B – splines and natural spline with four knots. Experiment with the number of knots. Select the optimal number of knots using AIC.

Ex.2.

Estimate smoothing splines with different degrees of freedom.

Ex.3.

Experiment with the number of knots for selected degrees of freedom.

Ex.4.

Compare a smoothing spline fit to a lowess fit. Next, compare a smoothing spline model to a natural cubic spline a natural spline model. Do the results differ essentially?

Ex.5.

Calculate 95% confidence interval bands for natural cubic splines and smoothing splines.

Ex.6.

Test the significance of natural cubic spline. Is the natural cubic spline better than either a linear, quadratic or logarithmic or transformation?

Possible commands in R could be found in script Splines.

L5.2. Exercises with Cross-Validation

Ex.7.

Use generalized cross-validation to select the optimal span in lowess regression and fit the recommended model.

Ex.8.

Use generalized cross-validation to choose the optimal smoothing parameter for spline and fit the recommended model.

Possible commands in R could be found in the script Cross-Validation.R.

Homework

Demonstrate the use of splines and cross-validation in the analysis of your own data.

Lab.6 (2 academic hours) Semiparametric regression models estimation

L6.1. Exercises with semiparametric regression

Use the dataset jacob.dta. The description of data and variables is given in Lab 4 (L4.2).

Ex.1.

Choose chal_vote as the dependent variable and perotvote, checks_raw as independent variables.

Fit a semiparametric regression model to the data using spline fits for both continuous predictors. Plot all the nonparametric estimates. Which appear to be nonlinear?

Ex.2.

Compare model with two splines and

a) linear model, b) model with one spline, c) model with power or log transformation

and choose the best model.

Ex.3.

Estimate the linear model with the addition of variables exp_chal, chal_spend, inc_spend , pres_vote, marginal, partisan_redist.

Test for each independent variable which functional form is better: a) spline or linear, b) spline or power transformation, c) spline or log transformation

Ex.4.

Choose the best model and plot all the nonparametric estimates.

Possible commands in R could be found in the script Semiparametric Models.R.

Homework

Demonstrate the use of semiparametric regression in the analysis of your own data.

Lab. 7. (2 academic hours) Robust Regression

Ex. Robust Regression in Stata

Use the dataset inequality.dta. Thisfile containes data for 49 countries in 1990.

Variables:

secpay –the name of the variable associated with a payment of two secretaries who perform roughly the same job but have different wages because of their different abilities. Secpay -1 is a percentage of surveyed people in the country who do not agree with the fact that at the same position can receive different wages. This variable is a measure of attitudes toward equality in pay.

gini – Gini index for a country,

gdp – GDP per capita in 1000's of dollars

democ – a dummy variable that is 1 if the country has been a democracy for longer than 10 years and 0 otherwise

Select secpay as a dependent variable in all models.

Ex.1.

Estimate an OLS regression with independent variables gini, gdp, democ described above. Do a visual diagnostic test for outliers in this model. Save the results.

Possible commands in STATA

```
reg secpay gini gdp democ
est store regols
```

Ex.2.

Calculate and plot the studentized residuals

Possible commands in STATA

```
predict residst, rstudent
twoway (scatter  residst  Number, sort)
```

Ex.3.

Make sure that the observations 25 and 49 are potential outliers (what are these countries?). Try to estimate the regression equation without these observations. Compare the results with those obtained previously.

Possible commands in STATA

```
reg secpay  gini gdp democ if Number !=25 & Number!=49
```

Ex.4.

Which observations have large leverage? Having identified these countries try to estimate the regression without them and compare the coefficient estimates.

Possible commands in STATA

```
predict xdist, leverage
sum xdist, detail
list country if  xdist > 0.2
```

Ex.5.

Estimate a relationship between leverage and normalized residuals. What observations seem outliers?

Possible commands in STATA

```
lvr2plot, mlabel(country)
```

Ex.6.

Calculate DFBETA for each variable, select those observations for which $DFBETA > 2/\sqrt{n}$.

Possible commands in STATA

```
dfbeta
list  country if  _dfbeta_1 > 2/7
list  country if  _dfbeta_2 > 2/7
list  country if  _dfbeta_3 > 2/7
```

Estimate the regression without these observations and compare coefficient estimates with the previous ones.

Ex.7.

Calculate DFITS for each variable, select those observations for which $DFITS > 2\sqrt{\frac{k}{n}}$

Possible commands in STATA

predict dfits, dfits

list country if dfits >2*sqrt(4/49)

Estimate the regression without these observations and compare coefficient estimates with the previous ones.

Ex.8.

Calculate Cook's Distance (D_i) for each variable, select those observations for which $D_i > 4/n$

Possible commands in STATA

predict cooksd, cooksd

list country if cooksd > 4/49

Estimate the regression without these observations and compare coefficient estimates with the previous ones.

Ex.9.

Calculate Welsch's Distance (W_i) for each variable, select those observations for which $W_i > 3\sqrt{k}$

Possible commands in STATA

predict wd, welsch

list country if wd > 6

Estimate the regression without these observations and compare coefficient estimates with the previous ones.

Ex.10.

Calculate COVRATIO for each variable, select those observations for which $\left|\operatorname{cov}ratio_i - 1\right| \geq \frac{3k}{n}$

Possible commands in STATA

predict covr, covratio

list country if abs(covr – 1) >=3*4/49

Estimate the regression without these observations and compare standard errors with the previous ones.

Ex.11.

Estimate the median regression and compare the results with the OLS estimates. Save the results.

Possible commands in STATA

```
qreg secpay gini gdp democ
est store regmed
```

Ex.12.

Estimate the M regression (with Hubert kernel) and compare the results with the OLS estimates. Save the results.

Possible commands in STATA

```
mregress secpay gini gdp democ
est store mreg
```

Ex.13.

Estimate M regression (with Hubert kernel for an initial approximation, and with a biweight kernel in the second step) and compare the results with the OLS estimates. Save the results.

Possible commands in STATA

```
rreg secpay gini gdp democ
est store rreg
```

Ex.14.

Estimate MM regressions for different levels of efficiency and compare the results with the OLS estimates. Save the results.

Possible commands in STATA

```
mmregress secpay gini gdp democ, eff(0.95) outlier graph label( country)
est store mmreg95
mmregress secpay gini gdp democ, eff(0.7) outlier graph label( country)
est store mmreg70
```

Ex.15.

Estimate MS regressions for different levels of efficiency and compare the results with the OLS estimates. Save the results.

Possible commands in STATA

msregress secpay gini gdp, dummies(democ) outlier graph

est store msreg

Ex.16.

Compare the results of all models estimation.

Possible commands in STATA

est tab regols regmed mreg rreg mmreg95 mmreg70 msreg,star

Possible commands in R could be found in script Robust Regression.R.

Homework

Demonstrate the use of robust regressions in the analysis of your own data.

Lab. 8. (2 academic hours) Quantile Regression

Ex. Quantile Regression in STATA

Use the dataset Rus_income.dta.

Variables:

psu – describe the place of respondent's residence,

age – respondent's age,

educ – respondent's years of education,

sex – respondent's sex (1 – male, 0 – female),

inc – respondent's monthly income,

exp – respondent's working experience ,

Ex.1.

Estimate the OLS regression with the dependent variable inc and independent variables age , age_square, sex, education. Estimate OLS regression and several quantile regressions.

Possible commands in STATA

qreg lninc age agesq sex educ, quantile(10)

estimates store qreg10

…

qreg lninc age agesq sex educ, quantile(90) estimates store qreg90

Combine the results into one table and compare.

Possible commands in STATA

est table qreg10 … qreg90, b(%6.3f) star

Ex.2.

Estimate quantile regressions with bootstrap standard errors.

Possible commands in STATA

sqreg lninc age agesq sex educ, q(0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9)

Ex.3.

Test the hypothesis about equality of all coefficients

Possible commands in STATA

test [q10=q20=q30=q40=q50=q60=q70=q80=q90]

or hypothesis about equality of selected coefficients

Possible commands in STATA

test [q10=q20=q30=q40=q50=q60=q70=q80=q90] : varlist (.)

Ex.4.

In order to get a graphical representation of quantile regression coefficients, install grgreg

Possible commands in STATA

ssc install grqreg, replace

Ex.5.

Estimate quantile regressions' coefficients and confidence intervals for them, and compare with point and interval OLS estimates.

Possible commands in STATA

sqreg lninc age agesq sex educ, quantile (0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9)

grqreg, ci ols olsci scale(1.1)

Possible commands in R could be found in script Quantile Regression.R.

Homework

Demonstrate the use of quantile regressions in the analysis of your own data .

Content