STATA TRAINING

Shaheed Bhagat Singh College

Shweta Gupta
Research Analyst
Environment & Production Technology Division (EPTD)
International Food Policy Research Institute

New Delhi | 8th April 2022

Part 5: Data cleaning

Extreme values

- What are extreme values?
  - smallest or largest values of a variable

- How to identify extreme values

  `sum mpg-` use this when you have a large range of values

  `sum foreign` → extreme values are 0 and 1 but this is a dummy variable with only these 2 values

  `tab rep78-` more suitable with discreet data

- Can also use graphs → `kdensity price`
- What to do with extreme values?
  - DO NOT DROP
  - Calculate statistics excluding the extreme values
  - Convert continuous variable into a discreet variable with class intervals

IFPRI

Missing values

- What are missing values?
  - When the value of a variable(s) is missing for 1 or more observations

  NOTE: ZERO is not always a missing value

- How to identify missing values

  `sum`- this gives the N for all variables

  ```
  tab rep78, missing
  count if rep78==.
  ```

  For numeric variable

  - For string variable, missing values appear as . (dot) and (space)

    ```
    tab make, missing
    count if make==“.” | make==“ ”| make==“”
    ```

  ```
  missings report varlist
  ```

Missing values (cont.)

- What to do with missing values?
  - Identify the cause of missing values
  - Is it missing data? (data non-availability)
  - Is that value supposed to be missing? (conditioned data)

- Drop only if the entire observation/variable has missing values

```
missings dropobs, force
missings dropvars varlist
```

- `drop if price==. & mpg==. & make=="  "`
- Replacing missing value by mean of the variable
  - By mean of full variable
  - By mean of the variable by class
  - **Never recommend doing, especially when a large no. of missings**

IFPRI

Example of missing values

name_student	gender	course	year	score_eco	score_english	score_math
A	Female	Eco_hons	2	89	60	80
B	Male	Eco_hons	2	80	40	90
C	.	Eng_hons	3	.	70	.
D	Female	Eng_hons	3	.	77	.
E	.	Maths_hons	1	.	.	100
F	Female	Maths_hons	1	.	.	98
G	Male	Maths_hons	3	.	50	67
H	Male	Maths_hons	3	.	79	59

Scores are out of 100
Year captures the year of college

Loops- forval/foreach commands

- Repeat a set of commands over different values of rep78

tab rep78

```
foreach x in 1 2 3 4 5 {
sum price if rep78==`x'
}

forval x= 1/5{
sum price if rep78==`x'
}
```

Part 6.1: Regression Analysis

2 variable linear regression

- Y is a continuous variable; X is continuous

```
regress y x

reg price mpg
```

```
. reg price mpg

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     20.26
       Model |   139449474         1   139449474   Prob > F        =    0.0000
    Residual |   495615923        72  6883554.48   R-squared       =    0.2196
-------------+----------------------------------   Adj R-squared   =    0.2087
       Total |   635065396        73  8699525.97   Root MSE        =    2623.7

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -238.8943   53.07669    -4.50   0.000    -344.7008   -133.0879
       _cons |   11253.06   1170.813     9.61   0.000     8919.088    13587.03
------------------------------------------------------------------------------
```

ANOVA table

Sq. root of Residual Sum of squares= √(6883554.48)

F-test for joint significance

```
. reg price mpg

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     20.26
       Model |   139449474         1   139449474   Prob > F        =    0.0000
    Residual |   495615923        72  6883554.48   R-squared       =    0.2196
-------------+----------------------------------   Adj R-squared   =    0.2087
       Total |   635065396        73  8699525.97   Root MSE        =    2623.7

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -238.8943   53.07669    -4.50   0.000    -344.7008   -133.0879
       _cons |   11253.06   1170.813     9.61   0.000     8919.088    13587.03
------------------------------------------------------------------------------
```

Dependent var

Constant term

independent var

Coefficients for slope & constant

Standard error of coefficients

T-statistic of coeff.

P-value of t-statistic

95% CI for each coefficient

Multiple variable linear regression

- Y is a continuous variable; X is continuous

regress y x1 x2

reg price mpg rep78

```
. reg price mpg rep78

      Source |       SS           df       MS      Number of obs   =        69
-------------+----------------------------------   F(2, 66)        =     11.06
       Model |  144754063         2  72377031.7   Prob > F        =    0.0001
    Residual |  432042896        66  6546104.48   R-squared       =    0.2510
-------------+----------------------------------   Adj R-squared   =    0.2283
       Total |  576796959        68  8482308.22   Root MSE        =    2558.5

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -271.6425   57.77115    -4.70   0.000    -386.9864   -156.2987
       rep78 |   666.9568   342.3559     1.95   0.056    -16.5789    1350.492
       _cons |   9657.754    1346.54     7.17   0.000      6969.3    12346.21
------------------------------------------------------------------------------
```

Linear regression

- T statistic-
  - t= (b_k -0)/b_k(SE)
  - Ho: B_k=0
- Calculate t-statistic for mpg from the output.

```
di “t=” (-238.8943-0)/(53.07669)
```

- Calculate t-statistic for mpg under the following Null hypothesis:
  - Ho: B_1 =250
  - t= (-238.8943-250)/(-238.8943)

```
di “t=” (-238.8943-250)/(53.07669)
```

IFPRI

Linear regression

- Change level of significance in confidence interval

```
reg price mpg, level(99)
```

- ANOVA table
  - MS=SS/df
  - Model: Explained sum of squares (ESS)
  - Residual: Residual sum of squares (RSS)
  - Total: Total sum of squares (TSS)

Source	SS	df	MS
Model	139449474	1	139449474
Residual	495615923	72	6883554.48
Total	635065396	73	8699525.97

IFPRI

Linear regression

```
Number of obs   =        74
F(1, 72)        =     20.26
Prob > F        =    0.0000
R-squared       =    0.2196
Adj R-squared   =    0.2087
Root MSE        =    2623.7
```

- Root MSE= root of RSS
- R-squared = Coefficient of determination= (Coefficient of correlation)^2

$$R^2 = r^2$$

- What is r here?

```
reg price mpg
di (0.0234)^(1/2)
pwcorr price mpg
```

Linear regression-Things to look for

- Beta coefficients- their signs and coefficient value- does it make sense?
- P value of each coefficient- significant at various alpha
- F test- the joint significance of all covariates
- R squared
- N- is N less than the actual no. of observations in data?-
    - Stata omits that observation entirely where the value of 1/more variable(s) is missing

IFPRI

Linear regression- multiple regressions in one frame

```
reg y x1
eststo model1
reg y x1 x2
eststo model2
esttab model1 model2
```

Example:

```
reg price mpg
eststo model1
reg price mpg rep78
eststo model2
esttab model1 model2
```

- By default- beta, t stats, N, and significance (*).
- To add options- r2 se/p ar2
- NOTE: for displaying multiple regressions, eststo is fine, but to **compare R-square** of 2 models, dependent variable must be same

Getting predicted values

- `reg y x1 x2`
- Get predicted values of y→ `predict yhat`
- Get predicted values of error→ `predict ehat, residuals`

Put any name

Put any name

```
reg price mpg
predict pricehat
predict ehat, residuals
```

- Go to Data browser to see the variables pricehat and ehat

Forecasting

- yhat= b0 + b1*x
- **pricehat=11253.06 + (-238.8943)*mpg** → regression equation
- The predicted values of y make more sense ONLY IF the x values lie close to their sample range.
- `sum mpg` → gives the range of mpg

Forecasting- example

- **pricehat=11253.06 + (-238.8943)*mpg**

- **Exercise: A) Predict yhat when x lies in sample range**
  mpg=25, pricehat=?

- **B) Predict yhat when x lies outside sample range (extrapolating)**
  mpg =50, pricehat=? ; mpg=100, pricehat=?

- **C) Comment on the predicted yhat**

Part 6.2: Different functional forms

Log-linear/constant elasticity/double-log model

- $Y = AX^{b1}$
- lny= bo + b1*lnx + u
- b1→ elasticity of y wrt. X

```
gen lny= ln(y)
gen lnx= ln(x)
regress lny lnx
```

- $Y = AX1^{b1}X2^{b2}$
- Lny= b0 + b1*lnx1 + b2*lnx2
- b1→ partial elasticity of y wrt. X1
- b2→ partial elasticity of y wrt x2

```
gen lny= ln(y)
gen lnx1= ln(x1)
Gen lnx2= ln(x2)
regress lny lnx1 lnx2
```

```
Example:
gen lnprice=ln(price)
gen lnmpg= ln(mpg)
reg lnprice lnmpg
```

IFPRI

Lin-log model

- y= b0 + b1*lnx + u

```
regress y lnx
```

- b1→ change in y when lnx changes by 1 unit
- b1/100→ change in y when x increases by 1%

```
Example:
gen lnmpg= ln(mpg)
reg price lnmpg
```

IFPRI

Semilog/ growth rate model

- $Y= y_0(1+r)^t$; r=compound rate of growth
- $\ln y= \ln(y_0) + t^* \ln(1+r)$
- lny= b0 + b1*t +u
- t is the time variable
- B1→ rate at which lny increases per time period

```
gen lny= ln(y)
reg lny t
```

- Same method if in place of t we have any x variable
- B1*100→ percentage change in Y due to 1 unit change in X→ instantaneous rog

```
Example:
gen lnprice=ln(price)
reg lnprice t
```

IFPRI

Linear trend model

- y= b0+ b1.t + u

```
reg y t
```

- T→ trend variable

```
Example:
reg price t
```

Other models

- Reciprocal model
  - y= b0 + b1*(1/x) + u

```
gen x2= 1/x
reg y x2
```

```
Example:
gen mpg_rec= mpg^(-1)
reg price mpg_rec
```

- Polynomial model
  - $y= b0 + b1*x1 + b2*x1^2 + b3*x1^3$ → 3rd degree polynomial
  - `gen x2= x1^2`
  - `gen x3= x1^3`
  - `reg y x1 x2 x3`

```
Example:
gen mpg2= mpg^2
gen mpg3= mpg^3
reg price mpg mpg2 mpg3
```

- Zero-intercept model
  - y= b1*x
  - `reg y x, noconstant`

Standardized regression

- Y*= b0 +b1.X* + u → Y*= b1.X*+ u → reg through origin
- Y*= (y-mean(y))/SD(y)
- X*= (x-mean(x))/SD(x)

`sum y x` →gives mean and SD for x and y

`gen ystar= (y-mean`$_y$`)/Sd`$_y$

`gen xstar= (x-mean`$_x$`)/SD`$_x$

`reg ystar xstar`

- b1→ if x1 changes by 1 standard deviation unit, average value of y1 changes by b1 standard deviation units

```
Example:
gen pricest=(price-6165.25)
/2949.49
gen mpgst=(mpg-21.29) /5.78
reg pricest mpgst

OR

egen pricemean= mean(price)
egen pricesd= sd(price)
gen pricest=(price-
pricemean)/pricesd
egen mpgmean= mean(mpg)
egen mpgsd= sd(mpg)
gen mpgst=(mpg-mpgmean)/mpgsd
reg pricest mpgst
```

IFPRI

Linear regression- types of independent variable

- Continuous- `reg y x` or `reg y c.x`

`reg price foreign→ all cont`

- Categorical (with >2 categories)-
  - Shortcut- `reg y i.x` –automatically takes the first category (lowest value) as the base category. `reg price i.rep78`
  - Long way-
    - `tab x, gen(x_dummy)` →generate dummy variable for each category and include (n-1) categories, n=total no. of categories;
    - `tab rep78, gen(rep)`
    - `reg price rep2 rep3 rep4 rep5`
- Dummy (categorical with 2 categories)- `reg y x` or `reg y i.x`

Base category

- How to change base category
  - Shortcut- `reg y ib#.x` where # is the category number you want to be base
  - Example- make 3rd category as base
  - `reg price ib3.rep78`
  - Long way- just omit the dummy of the new category but keep all other category dummies.
  - `reg price rep1 rep2 rep4 rep5`

- Include all categories- dummy variable trap due to multicollinearity

IFPRI

Interaction terms

- Interaction terms-
  - x3= x1* x2

```
gen x3=x1*x2
reg y x1 x2 x3
```

  - Shortcut
  - x1 and x2 are both continuous: `reg y c.x1##c.x2`
  - `price~mpg, weight;`
  - `reg price c.mpg##c.weight`
  - x1 is cont. x2 is categorical: `reg y c.x1##i.x2`
  - `Price~mpg, foreign; reg price c.mpg##i.rep78`
  - x1 and x2 are both categorical: `reg y i.x1##i.x2`
- The above regs include x1, x2 and product of x1 and x2 as independent variables

Interaction terms

- Only interaction term:
  - `reg y x1#x2` along with operators c or i.
  - **Example: Run regression of price on mpg and mpg*rep78 treating rep78 as categorical.**

- When x1=x2, then its squared term;
  - polynomial regression- `reg y x1 x1#x1 x1#x1#x1` 3rd order polynomial regression
  - reg price mpg mpg#mpg mpg#mpg#mpg

Part 6.3: Exploring CLRM assumptions

Exploring relationships

`pwcorr y x1 x2, star(0.05) sig` → does estimating a regression even make sense?

`scatter y x` -linear/quardratic or cubic r/s?

- Depending on the plot include single, squared or cube terms

- reg price mpg

`acprplot mpg, lowess` (run right after regression)

- Shows whether to include squared terms or not

IFPRI

Outliers

`sum y x1 x2..` → check if variables have high standard deviations

```
kdensity x1
```

- Added Variable plots
- Useful in Multiple regression- Plots the partial regression of Y on X1 keeping X2 constant

```
avplot x1
```

`avplots` – gives plots for all independent variables

- Solution: Don’t include those observations that include outlier value

```
kdensity price
reg price mpg if price<10000
```

Multicollinearity

- Independent variables should not be perfectly collinear
- gen mpg2= 2*weight
- reg price mpg2 weight
- How to suspect?
  - High SEs
  - Wrong signs of coefficient
  - High R-sq but many coefficients insignificant

Multicollinearity

- Stata automatically drops a variable if it is **perfectly** collinear with the other
- Example:

```
gen mpg2= mpg + 2
reg price mpg mpg2
```

```
. reg price mpg mpg2
note: mpg2 omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     20.26
       Model |   139449474         1   139449474   Prob > F        =    0.0000
    Residual |   495615923        72  6883554.48   R-squared       =    0.2196
-------------+----------------------------------   Adj R-squared   =    0.2087
       Total |   635065396        73  8699525.97   Root MSE        =    2623.7

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -238.8943   53.07669    -4.50   0.000    -344.7008   -133.0879
        mpg2 |          0  (omitted)
       _cons |   11253.06   1170.813     9.61   0.000     8919.088    13587.03
------------------------------------------------------------------------------
```

Multicollinearity

- To check for high (<1) correlation:
  - `pwcorr`
  - `pwcorr x1 x2`… → >0.5 then high correlation

  - `vif` (run right after regression)- gives variance inflation factor
  - A vif>10 means problem
  - scatter x1 x2

- Solution→ drop the collinear variable, rethink the model

Multicollinearity- example

```
reg price mpg weight
scatter mpg weight, sort
pwcorr mpg weight
```

```
. reg price mpg weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(2, 71)        =     14.74
       Model |   186321280         2  93160639.9   Prob > F        =    0.0000
    Residual |   448744116        71  6320339.67   R-squared       =    0.2934
-------------+----------------------------------   Adj R-squared   =    0.2735
       Total |   635065396        73  8699525.97   Root MSE        =      2514

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -49.51222   86.15604    -0.57   0.567    -221.3025     122.278
      weight |   1.746559   .6413538     2.72   0.008      .467736    3.025382
       _cons |   1946.069    3597.05     0.54   0.590    -5226.245    9118.382
------------------------------------------------------------------------------
```

IFPRI

Homoskedasticity/heteroskedasticity

- Residuals should not vary by values of X or yhat; leads to inefficient betas→ High SEs
- Graphically-
  - `rvfplot` (run right after regression)
  - Gives scatterplot of residuals on y-axis and predicted values of dependent variable on x-axis
  - If pattern, then homoskedasticity violated

Example:

```
reg price weight if price<10000
rvfplot
reg price weight rep78 if price<10000
rvfplot
```

Homoskedasticity/heteroskedasticity

- Tests of homoskedasticity, **park test, Glejser test**

1)Breusch-Pagan test

- Ho- Residuals are homosk.
- Reject ho if prob>chi is less than alpha

```
reg y x
estat hettest
```

2)White Test

Ho: Residuals are homoscedastic

Reject Ho if prob>chi is too small

```
reg y x
estat imtest, white
```

Homoskedasticity/heteroskedasticity

- Solution
  - Add/remove some variables
  - Robust standard errors- `reg y x1, robust` → Rule of thumb is to assume heteroscedasticity in your model and put this option
  - Weighted least squares- not covered in this session

Omitted variable bias

- E(e|x)=0)- errors and independent variable not correlated;
- leads to biased & inconsistent beta coefficients→ large sample doesn't reduce bias

- Test for OV bias-
  - **Ramsey RESET test**
  - Ho- no omitted variables in model
  - Reject Ho if prob>F is less than alpha
  - Command- `ovtest` (run right after regression)

- `linktest` (run right after regression)
  - regresses actual y against yhat and yhat-squared
  - Ho- no specification error
  - Look for significance of yhat-squared
  - yhat-squared insignificant→ reject Ho → no specification error

Irrelevant variable bias

- Leads to inefficient beta coefficients→ high variance

- Check the significance of irrelevant variable
- See if regression output is affected by removing the variable
- Theory should be the guide
- Sometimes important variables are also insignificant→ even then keep it→ called as control variables

Specification error- Functional form bias

- Should we run linear or log-linear regression?
- Mackinnon-White-Davidson (MWD) test

```
reg price mpg
predict yhat1

gen lnprice= ln(price)
gen lnmpg= ln(mpg)
reg lnprice lnmpg
predict yhat2

gen lnyhat1= ln(yhat1)
gen Z= lnyhat1- yhat2
reg price mpg Z
```

Ho: Linear model is correct
H1: Log-linear model is correct

Reject Ho if p-value of Z variable is less than alpha, i.e., if coefficient of Z is statistically sig.

IFPRI

Normality of errors- graphically

- Errors are normally distributed; if violated then all problems arise
- First estimate ehat

```
regress y x1 x2
predict ehat, resid
```

1. Density curves:
   - `kdensity ehat, normal` → plots ehat and a normal distribution for comparison
   - `histogram ehat, kdensity normal` → above plot + a histogram of residuals
2. Normal probability plots- plots residuals on x-axis against E(residuals|normal) on y-axis
   - If normal, then the residuals lie on straight line
   - `pnorm ehat` → non-normality in the middle range of residuals
   - `qnorm ehat` → non-normality in extreme values of data

Normality of errors- tests

1. Jarque Bera test (large sample test)
   - Ho: Residuals are normally distributed
   - Reject Ho if p-value of computed chi-sq is lower than alpha

a) Manually construct JB statistic and compare it with chi-square value

```
regress y x1 x2
predict ehat, resid
sum ehat,d
```
→get N, skewness, kurtosis from this output

Construct JB as follows and compare it with chi-square

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \sim \chi^2_{(2)}$$

Normality of errors- tests

b) Automatically do JB test

```
regress y x1 x2
predict ehat, resid
jb ehat
```

Ho is rejected is alpha is more than Chi(2) from output

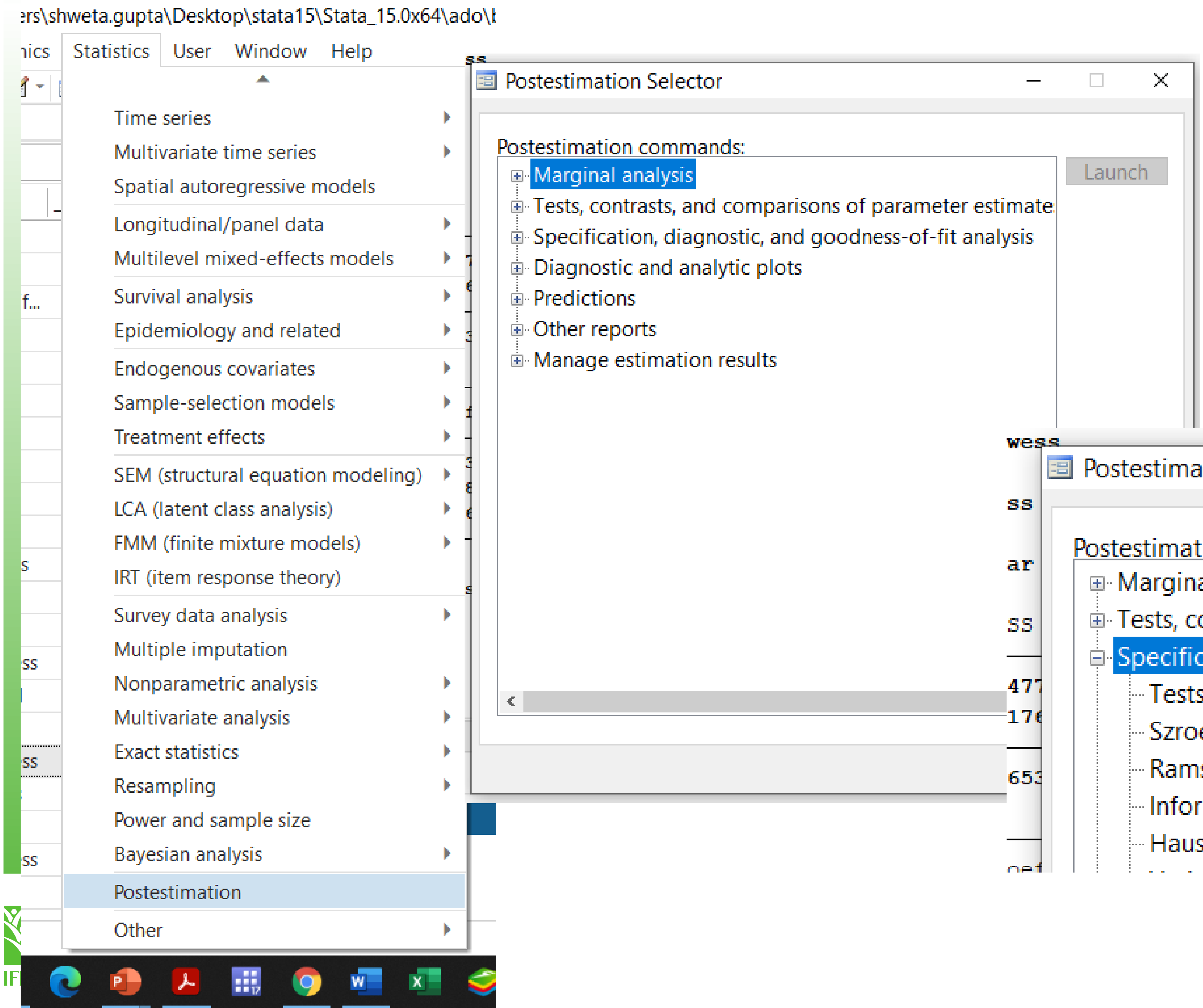
Normality of errors- tests

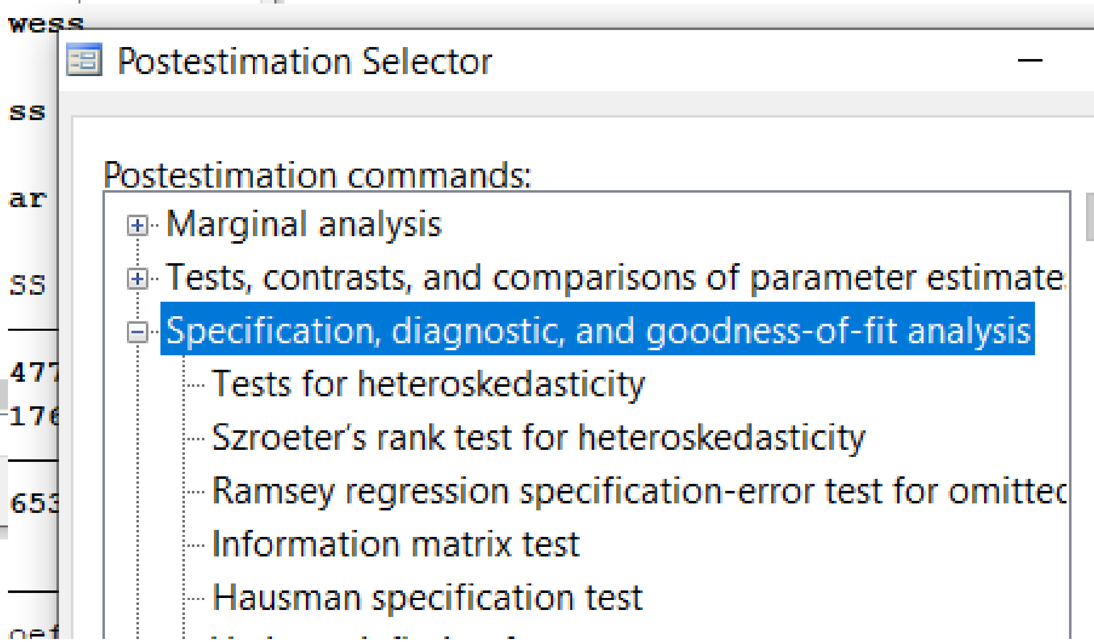
2. Shapiro-Wilk test
   - Ho- normal distribution of residuals
   - Reject Ho if p-value is less than alpha
   - Command:

```
regress y x1 x2
predict ehat, resid
swilk ehat
```

3. Anderson-Darling test (large sample test)
   - Ho- Normal distribution of residuals
   - Do not reject Ho if p value is more than alpha
   - Command: `lmnad y x1 x2`
   - No need to run regression first

Short cut to locate various tests





Tests of significance after regression

- Ho: B_k=0
  - Regression output directly gives t-statistic and p-value
  - After regression run:
  - `test x1` → gives F statistic(=square of t statistic) and its significance
  - F ~ F(1,n-k)
- Ho: B_k=k
  - `test x1=k` (k!=0) → revise t-statistic
  - `test x1=-k`
  - `reg price mpg`
  - `test mpg=-250`
- Ho:B_1 and B_2 form a linear relationship→ this is not testing for multicollinearity
  - `test x1=x2`
  - `test x1+x2= k`
  - `test x1 x2, mtest` → testing B_1=0 and B_2=0 in one command

IFPRI

Tests of significance

- Joint testing:-
- Ho: $B_1=B_2=0$
  - `test x1 x2`
- Ho: $B_1=B_2=B_3=\ldots\ldots=B_k=0$
  - Test of joint significance of all variables
  - `test x1 x2.. Xk`
  - $F \sim F(k-1,n-k)$
  - verify F-statistic from the regression output

ANOVA table for joint testing

- How is F statistic calculated?
  - F= (ESS/df) / (RSS/df)
- F~ F(k-1,n-k)
- **Exercise: Check regression output for ANOVA table and calculate F**

Part 7: Types of data

Types of data

- **Cross sectional→** multiple variables/individuals/groups data given in a single time point (no time dimension)
- **Time series→** data given in successive order for multiple time points (has a time dimension)
- **Panel data→** same set of individuals are tracked for multiple time points and their data on one/many variables collected for all those times
- **Pooled data→** data on multiple individuals available for multiple time points but not necessarily same individuals tracked overtime

IFPRI

Cross-sectional data

student_name	score	income_annual_lakh	year
A	98	8	2000
B	45	4	2000
C	67	4.5	2000
D	89	6.8	2000
E	45	4	2000
F	34	6	2000
G	90	21.4	2000
H	78	20	2000
I	65	3	2000
J	48	7	2000
K	67	1	2000
L	80	6	2000
M	56	8.9	2000
N	59	3.5	2000
O	67	8	2000

Time series data

year	score_avg	income_annual_avg
2000	98	8
2001	45	4
2002	67	4.5
2003	89	6.8
2004	45	4
2005	34	6
2006	90	21.4
2007	78	20
2008	65	3
2009	48	7
2010	67	1
2011	80	6
2012	56	8.9
2013	59	3.5
2014	67	8

IFPRI

Panel data

student_name	score	income_annual_lakh	year
A	98	8	2000
B	45	4	2000
C	67	4.5	2000
D	89	6.8	2000
E	45	4	2000
A	98	8	2001
B	45	4	2001
C	67	4.5	2001
D	89	6.8	2001
E	45	4	2001
A	98	8	2002
B	45	4	2002
C	67	4.5	2002
D	89	6.8	2002
E	45	4	2002

IFPRI

Pooled data

student_name	score	income_annual_lakh	year
A	98	8	2000
B	45	4	2000
C	67	4.5	2000
D	89	6.8	2000
E	45	4	2000
A	98	8	2001
P	45	4	2001
Q	67	4.5	2001
R	89	6.8	2001
E	45	4	2001
Y	98	8	2002
Z	45	4	2002
H	67	4.5	2002
K	89	6.8	2002
L	45	4	2002

IFPRI

Question

- What type of data is the auto.dta used in the sessions?

Monte Carlo experiments

- Conducting simulations over an artificial sample
- Help in understanding properties of OLS estimators
- Unbiasedness: $b1 \sim B1$ as $n \rightarrow \infty$
- Consistency: bias $\rightarrow 0$ as $n \rightarrow \infty$
- Normality: $b1 \sim N(B1, \sigma)$
- Efficiency: Var(b1) is the lowest among all b1s

- Show using the monte carlo do-file

IFPRI

Time series data

- Example:-
- GDP and inflation rate for a country over time

year	gdp_perc	inflation_rate
1990	3%	5.01%
1991	4%	5.2%
1992	4.2%	5.89%
1993	4%	4%
1994	3.65%	4.5%
1995	4.5%	4.32%
1996	5%	5%
1997	5.1%	5%
1998	4.98%	5%
1999	3%	6%
2000	3.8%	5.89%
2001	3.9%	5.6%

Time series analysis

- **Minimum no. of times- 20**
- First ensure that the time variable in the correct format
- Are there any gaps in the time variable?
- Then do `tsset datevar`
- If gaps in time variable:
  - A- create a continuous time variable (say, time), and do `tsset time`
  - B- fill gaps in the time variable
    - `tsset datevar`
    - `tsfill` → data will show as missing for times where there was a gap in the time variable

IFPRI

Lagged variables

- Lags- create a variable that takes value of previous year for each year
- `gen x2= L1.x1` → x2(t)= x1(t-1)
- `gen x2= L2.x1` → x2(t)= x1(t-2)
- `gen xk= Lk.x1` → x2(t)= x1(t-k)

- `reg y x1 L1.x1 L2.x1`
- `reg y x1 L(1/5).x1`

- Note- with lags, missing values are created in lagged variables. Those observations get dropped in regression

Lead variables

- Leads/forwards- create a variable that takes value of next year for each year
- `gen x2= F1.x1` → x2(t)= x1(t+1)
- `gen x2= F2.x1` → x2(t)= x1(t+2)
- `gen x2= Fk.x1` → x2(t)= x1(t+k)

- `reg y x1 F1.x1 F2.x1`
- `reg y x1 F(1/5).x1`

- Note- with leads, missing values are created in forward variables. Those observations get dropped in regression

- `reg y x1 L1.x1 F2.x1` →**How many observations will be dropped?**

Difference variables

- Create a variable that takes difference of value between 2 years for each year
- `gen x2= D1.x1` → x2(t)= x1(t) – x1(t-1)
- `gen x2= D2.x1` → x2(t)= x1(t) – x1(t-2)
- `gen x2= Dk.x1` → x2(t)= x1(t) – x1(t-k)

- `reg y x1 D1.x1 D2.x1`
- `reg y x1 D(1/5).x1`

- Note- missing values are created in difference variables. Those observations get dropped in regression

Problems in time series data

- Autocorrelation- when errors at different time points are correlated to each other
  - corrgram x1, lags(k) (interpret output)

- Non-stationarity- when properties of time series depend on time
  - dfuller x1, lag(k)

Panel data

- Example:-

country	year	gdp_perc	inflation_rate
India	1990	3%	5.01%
India	1991	4%	5.2%
India	1992	4.2%	5.89%
USA	1990	4%	4%
USA	1991	3.65%	4.5%
USA	1992	4.5%	4.32%
China	1990	5%	5%
China	1991	5.1%	5%
China	1992	4.98%	5%
Russia	1990	3%	6%
Russia	1991	3.8%	5.89%
Russia	1992	3.9%	5.6%

IFPRI

contact

- shweta.gupta@cgiar.org
- shwetagpt33@gmail.com