Quantitative Methods in Political Science

Recitation

Mai Nguyen

New York University

September 30, 2013

Review from Last Week’s Lab Session

- Summarize data using the *summarize* command

Review from Last Week's Lab Session

- Summarize data using the *summarize* command
- Data visualizations using commands as well as the dropdown menu:

Review from Last Week's Lab Session

- Summarize data using the *summarize* command
- Data visualizations using commands as well as the dropdown menu:
    - Stem and Leaf Plot

Review from Last Week's Lab Session

- Summarize data using the *summarize* command
- Data visualizations using commands as well as the dropdown menu:
  - Stem and Leaf Plot
  - Histogram

Review from Last Week's Lab Session

- Summarize data using the *summarize* command
- Data visualizations using commands as well as the dropdown menu:
    - Stem and Leaf Plot
    - Histogram
    - Pie Charts

Review from Last Week's Lab Session

- Summarize data using the *summarize* command
- Data visualizations using commands as well as the dropdown menu:
    - Stem and Leaf Plot
    - Histogram
    - Pie Charts
    - Bar Charts

Review from Last Week's Lab Session

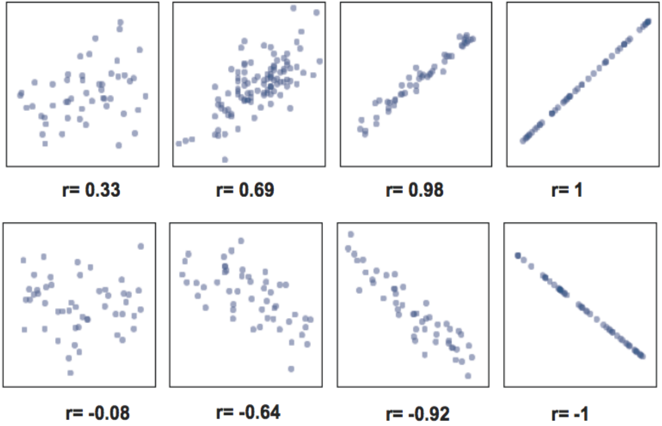
- Summarize data using the *summarize* command
- Data visualizations using commands as well as the dropdown menu:
    - Stem and Leaf Plot
    - Histogram
    - Pie Charts
    - Bar Charts
    - Box Plots

Review from Last Week's Lab Session

- Summarize data using the *summarize* command
- Data visualizations using commands as well as the dropdown menu:
  - Stem and Leaf Plot
  - Histogram
  - Pie Charts
  - Bar Charts
  - Box Plots
- Saving and editing graphs using graph editor

Correlation Analysis

Remember from class...



How do we get here?

Correlation Analysis

- What does correlation tell us? How is it measured?

Correlation Analysis

- What does correlation tell us? How is it measured?
  - Measures the strength of the linear relationship between two variables

Correlation Analysis

- What does correlation tell us? How is it measured?
  - Measures the strength of the linear relationship between two variables
  - Correlation coefficient (r); many properties

Correlation Analysis

- What does correlation tell us? How is it measured?
  - Measures the strength of the linear relationship between two variables
  - Correlation coefficient (r); many properties
- We can perform a correlation analysis in Stata using the *correlate* command:

Correlation Analysis

- What does correlation tell us? How is it measured?
  - Measures the strength of the linear relationship between two variables
  - Correlation coefficient (r); many properties
- We can perform a correlation analysis in Stata using the *correlate* command:
  - Type: *correlate* **variablename1 variablename2**

Correlation Analysis

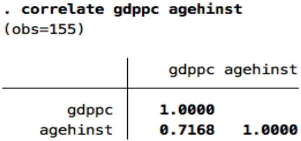
- What does correlation tell us? How is it measured?
    - Measures the strength of the linear relationship between two variables
    - Correlation coefficient (r); many properties
- We can perform a correlation analysis in Stata using the *correlate* command:
    - Type: *correlate* **variablename1 variablename2**
    - Example: *correlate gdppc investment*

Correlation Analysis

- What does correlation tell us? How is it measured?
  - Measures the strength of the linear relationship between two variables
  - Correlation coefficient (r); many properties
- We can perform a correlation analysis in Stata using the *correlate* command:
  - Type: *correlate* **variablename1 variablename2**
  - Example: *correlate gdppc investment*
  - As always, you can shorten the Stata command and use *corr*

Correlation Analysis

- What does correlation tell us? How is it measured?
  - Measures the strength of the linear relationship between two variables
  - Correlation coefficient (r); many properties
- We can perform a correlation analysis in Stata using the *correlate* command:
  - Type: *correlate* **variablename1 variablename2**
  - Example: *correlate gdppc investment*
  - As always, you can shorten the Stata command and use *corr*



```
. correlate gdppc agehinst
(obs=155)

             |    gdppc agehinst
-------------+------------------
       gdppc |   1.0000
    agehinst |   0.7168   1.0000
```

Correlation Analysis

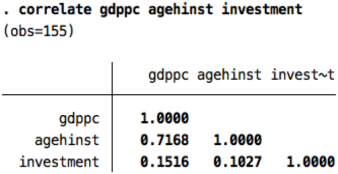
- You can use *correlate* for more than two variables:

Correlation Analysis

- You can use *correlate* for more than two variables:
    - Example: *correlate gdppc agehinst investment*

Correlation Analysis

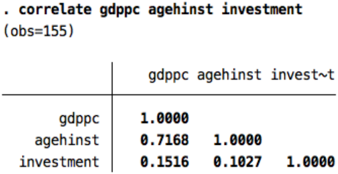
- You can use *correlate* for more than two variables:
  - Example: *correlate gdppc agehinst investment*



```
. correlate gdppc agehinst investment
(obs=155)

             |    gdppc agehinst invest~t
-------------+---------------------------
       gdppc |   1.0000
    agehinst |   0.7168   1.0000
  investment |   0.1516   0.1027   1.0000
```

Correlation Analysis

- You can use *correlate* for more than two variables:
    - Example: *correlate gdppc agehinst investment*



```
. correlate gdppc agehinst investment
(obs=155)

             |    gdppc agehinst invest~t
-------------+---------------------------
       gdppc |   1.0000
    agehinst |   0.7168   1.0000
  investment |   0.1516   0.1027   1.0000
```

- Notice correlation coefficients are still only for each pair of variables.

Scatterplots

Similarly related, we can visualize data using scatterplots:

- We can do this using the *scatter* command in Stata

Scatterplots

Similarly related, we can visualize data using scatterplots:

- We can do this using the *scatter* command in Stata
  - Type *scatter* **variablename1 variablename2**

Scatterplots

Similarly related, we can visualize data using scatterplots:

- We can do this using the *scatter* command in Stata
  - Type *scatter* **variablename1 variablename2**
  - Example: *scatter gdppc agehinst*

Scatterplots

Similarly related, we can visualize data using scatterplots:

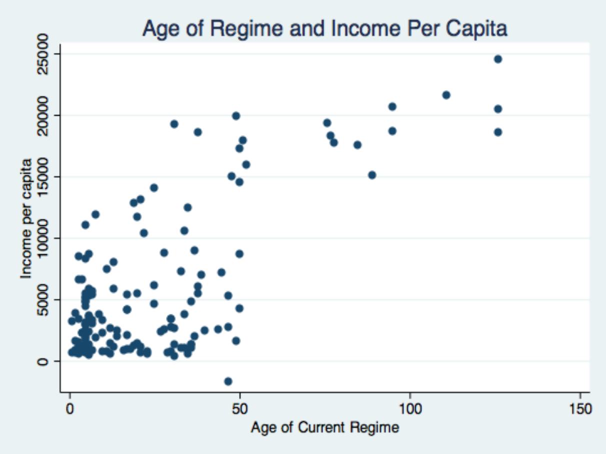
- We can do this using the *scatter* command in Stata
  - Type *scatter* **variablename1 variablename2**
  - Example: *scatter gdppc agehinst*
- Like other forms of data visualization, you can save and edit your scatterplot in the "Graph Editor" window

Scatterplots

Similarly related, we can visualize data using scatterplots:

- We can do this using the *scatter* command in Stata
  - Type *scatter* **variablename1 variablename2**
  - Example: *scatter gdppc agehinst*
- Like other forms of data visualization, you can save and edit your scatterplot in the "Graph Editor" window
- You can also use the dropdown menu: Graphics → Twoway graph

Scatterplots



Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:
    - *generate* **newvariablename**=**something**

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:
  - *generate* **newvariablename**=**something**
  - Example: *generate new=agehinst*

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:
  - *generate* **newvariablename=something**
  - Example: *generate new=agehinst*
  - This creates a variable called "new" that is identitcal to the agehinst variable

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:
  - *generate* **newvariablename=something**
  - Example: *generate new=agehinst*
  - This creates a variable called "new" that is identitcal to the agehinst variable
- You can do a variety of things in creating a new variable:

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:
    - *generate* **newvariablename**=**something**
    - Example: *generate new=agehinst*
    - This creates a variable called "new" that is identitcal to the agehinst variable
- You can do a variety of things in creating a new variable:
    - *generate gdp10=gdppc/10*

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:
  - *generate* **newvariablename=something**
  - Example: *generate new=agehinst*
  - This creates a variable called "new" that is identitcal to the agehinst variable
- You can do a variety of things in creating a new variable:
  - *generate gdp10=gdppc/10*
  - This divides the gdppc variable by 10

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:
  - *generate* **newvariablename=something**
  - Example: *generate new=agehinst*
  - This creates a variable called "new" that is identitcal to the agehinst variable
- You can do a variety of things in creating a new variable:
  - *generate gdp10=gdppc/10*
  - This divides the gdppc variable by 10
  - *generate zero=0*

Generating Variables

We're going to switch gears a little bit now and learn how to create and recode variables in Stata. To create a new variable:

- We could do what we did in week 3 where we manually input data into blank cells in the "Data Editor" window to create a new variable...
- A better option is to use the Stata *generate* command. The general format for creating a new variable is:
  - *generate* **newvariablename=something**
  - Example: *generate new=agehinst*
  - This creates a variable called "new" that is identitcal to the agehinst variable
- You can do a variety of things in creating a new variable:
  - *generate gdp10=gdppc/10*
  - This divides the gdppc variable by 10
  - *generate zero=0*
  - This creates a variable that is all zeros

Recoding Variables

Recode categorical variables:

- Let's take a look at at the *hinst* variable again

Recoding Variables

Recode categorical variables:

- Let's take a look at at the *hinst* variable again
- *tab hinst*

Recoding Variables

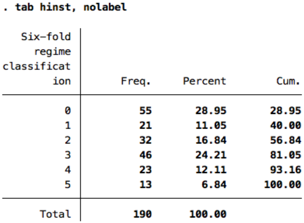
Recode categorical variables:

- Let's take a look at at the *hinst* variable again
- *tab hinst*
- *tab hinst, nolabel*

Recoding Variables

Recode categorical variables:

- Let's take a look at at the *hinst* variable again
- *tab hinst*
- *tab hinst, nolabel*



```
. tab hinst, nolabel

   Six-fold |
     regime |
 classificat |
        ion |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         55       28.95       28.95
          1 |         21       11.05       40.00
          2 |         32       16.84       56.84
          3 |         46       24.21       81.05
          4 |         23       12.11       93.16
          5 |         13        6.84      100.00
------------+-----------------------------------
      Total |        190      100.00
```

Recoding Variables

- We can recode the *hinst* variable use the *recode* command

Recoding Variables

- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**

Recoding Variables

- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**
    - *generate system=hinst*

Recoding Variables

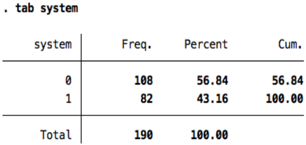
- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**
  - *generate system=hinst*
  - *recode system 0 1 2=0 3 4 5=1*

Recoding Variables

- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**
  - *generate system=hinst*
  - *recode system 0 1 2=0 3 4 5=1*
  - Here we created a variable “system” and recoded it to become a dichotomous variable; notice it is identical to the *regime* variable

Recoding Variables

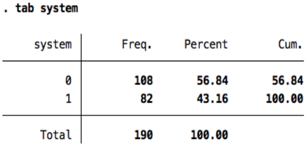
- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**
    - *generate system=hinst*
    - *recode system 0 1 2=0 3 4 5=1*
    - Here we created a variable “system” and recoded it to become a dichotomous variable; notice it is identical to the *regime* variable



```
. tab system

     system |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        108       56.84       56.84
          1 |         82       43.16      100.00
------------+-----------------------------------
      Total |        190      100.00
```

Recoding Variables

- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**
    - *generate system=hinst*
    - *recode system 0 1 2=0 3 4 5=1*
    - Here we created a variable “system” and recoded it to become a dichotomous variable; notice it is identical to the *regime* variable

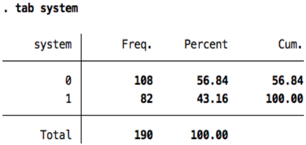


```
. tab system

     system |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        108       56.84       56.84
          1 |         82       43.16      100.00
------------+-----------------------------------
      Total |        190      100.00
```

- We can also recode continuous variables in a similar way

Recoding Variables

- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**
  - *generate system=hinst*
  - *recode system 0 1 2=0 3 4 5=1*
  - Here we created a variable “system” and recoded it to become a dichotomous variable; notice it is identical to the *regime* variable

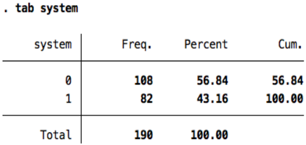


```
. tab system

     system |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        108       56.84       56.84
          1 |         82       43.16      100.00
------------+-----------------------------------
      Total |        190      100.00
```

- We can also recode continuous variables in a similar way
  - *generate majority=govsh*

Recoding Variables

- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**
  - *generate system=hinst*
  - *recode system 0 1 2=0 3 4 5=1*
  - Here we created a variable “system” and recoded it to become a dichotomous variable; notice it is identical to the *regime* variable

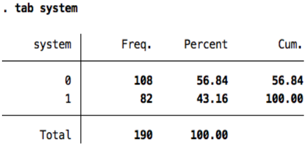


```
. tab system

     system |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        108       56.84       56.84
          1 |         82       43.16      100.00
------------+-----------------------------------
      Total |        190      100.00
```

- We can also recode continuous variables in a similar way
  - *generate majority=govsh*
  - *recode majority 0/0.5=0 0.5/1=1*

Recoding Variables

- We can recode the *hinst* variable use the *recode* command
- Type *recode* **variablename something**
    - *generate system=hinst*
    - *recode system 0 1 2=0 3 4 5=1*
    - Here we created a variable “system” and recoded it to become a dichotomous variable; notice it is identical to the *regime* variable

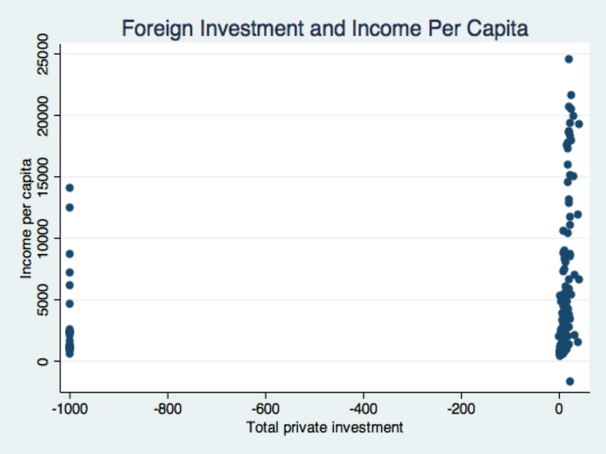


```
. tab system

     system |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        108       56.84       56.84
          1 |         82       43.16      100.00
------------+-----------------------------------
      Total |        190      100.00
```

- We can also recode continuous variables in a similar way
    - *generate majority=govsh*
    - *recode majority 0/0.5=0 0.5/1=1*
    - Here we turned a continous variable into a dichotomous variable

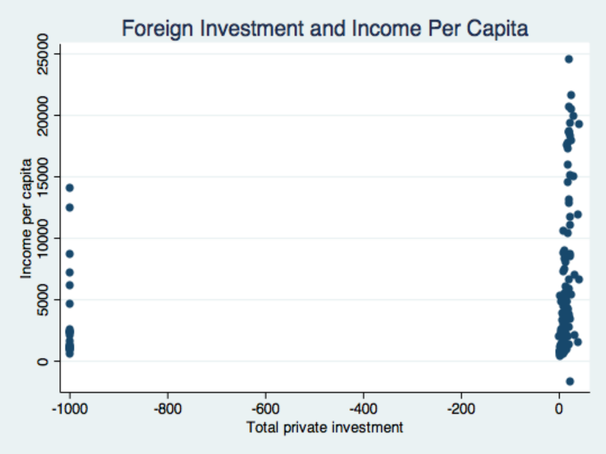
Missing Data

Let's take a look at a scatterplot of income per capita and investment: *scatter gdppc investment*



Missing Data

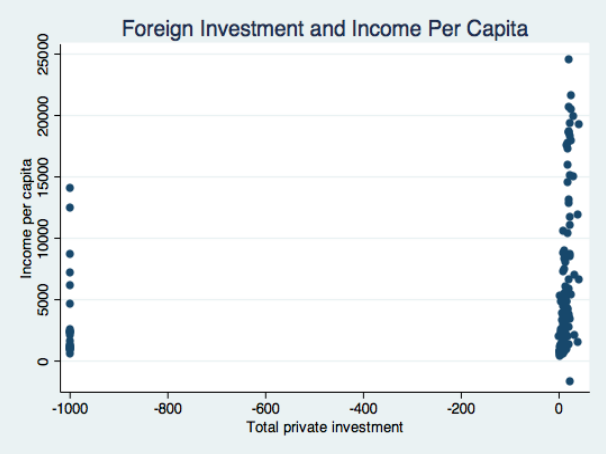
Let's take a look at a scatterplot of income per capita and investment: *scatter gdppc investment*



- correlation = 0.1516 (found by *correlate gdppc investment*)

Missing Data

Let's take a look at a scatterplot of income per capita and investment: *scatter gdppc investment*



- correlation = 0.1516 (found by *correlate gdppc investment*)
- Does something look a little weird?

Missing Data

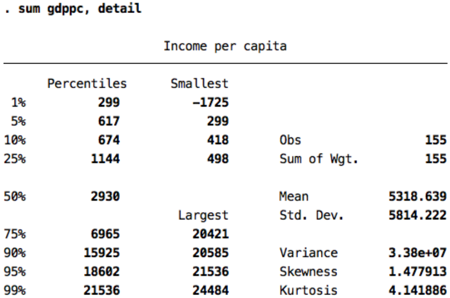
- Let’s take a look at our data:

Missing Data

- Let’s take a look at our data:
  - *sum gdppc, detail*

Missing Data

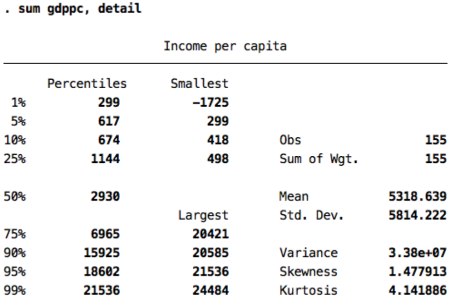
- Let’s take a look at our data:
    - *sum gdppc, detail*



```
. sum gdppc, detail

                      Income per capita
-------------------------------------------------------------
      Percentiles      Smallest
 1%          299          -1725
 5%          617            299
10%          674            418       Obs                 155
25%         1144            498       Sum of Wgt.         155

50%         2930                      Mean           5318.639
                        Largest       Std. Dev.      5814.222
75%         6965          20421
90%        15925          20585       Variance       3.38e+07
95%        18602          21536       Skewness       1.477913
99%        21536          24484       Kurtosis       4.141886
```

Missing Data

- Let's take a look at our data:
  - *sum gdppc, detail*



```
. sum gdppc, detail

                      Income per capita
-------------------------------------------------------------
      Percentiles      Smallest
 1%          299          -1725
 5%          617            299
10%          674            418       Obs                 155
25%         1144            498       Sum of Wgt.         155

50%         2930                      Mean           5318.639
                        Largest       Std. Dev.      5814.222
75%         6965          20421
90%        15925          20585       Variance       3.38e+07
95%        18602          21536       Skewness       1.477913
99%        21536          24484       Kurtosis       4.141886
```

- We can see from the scatterplot and summarize output that we have some values of income per capita that are negative. Does this make sense?

Missing Data

We can change this...

- First let's find the observations that have negative income per capita:

Missing Data

We can change this...

- First let's find the observations that have negative income per capita:
  - *list name country gdppc if gdppc<0*

Missing Data

We can change this...

- First let's find the observations that have negative income per capita:
  - *list name country gdppc if gdppc<0*
  - We find that Germany has negative income per capita

Missing Data

We can change this...

- First let's find the observations that have negative income per capita:
  - *list name country gdppc if gdppc<0*
  - We find that Germany has negative income per capita
- We can fix this using the *recode* command:

Missing Data

We can change this...

- First let's find the observations that have negative income per capita:
    - *list name country gdppc if gdppc<0*
    - We find that Germany has negative income per capita
- We can fix this using the *recode* command:
    - *recode gdppc min/0=.*

Missing Data

We can change this...

- First let's find the observations that have negative income per capita:
    - *list name country gdppc if gdppc<0*
    - We find that Germany has negative income per capita
- We can fix this using the *recode* command:
    - *recode gdppc min/0=.*
    - The period (.) is used to signify missing data in Stata

Missing Data

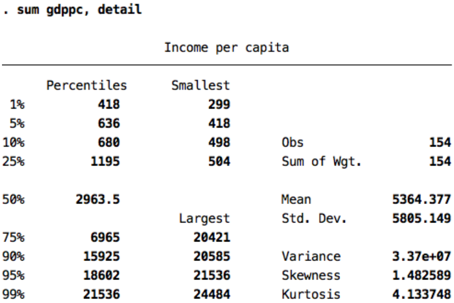
- Let’s take a look at our data again:

Missing Data

- Let's take a look at our data again:
  - *sum gdppc, detail*

Missing Data

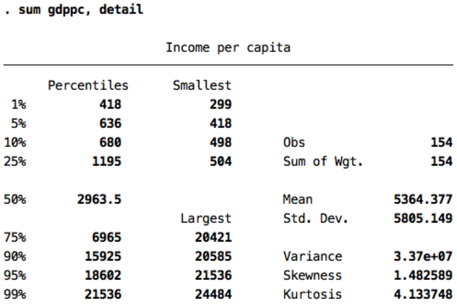
- Let's take a look at our data again:
    - *sum gdppc, detail*



```
. sum gdppc, detail

                       Income per capita
-------------------------------------------------------------
      Percentiles      Smallest
 1%          418            299
 5%          636            418
10%          680            498       Obs                 154
25%         1195            504       Sum of Wgt.         154

50%       2963.5                      Mean           5364.377
                        Largest       Std. Dev.      5805.149
75%         6965          20421
90%        15925          20585       Variance       3.37e+07
95%        18602          21536       Skewness       1.482589
99%        21536          24484       Kurtosis       4.133748
```

Missing Data

- Let's take a look at our data again:
  - *sum gdppc, detail*



```
. sum gdppc, detail

                      Income per capita
-------------------------------------------------------------
      Percentiles      Smallest
 1%          418            299
 5%          636            418
10%          680            498       Obs                 154
25%         1195            504       Sum of Wgt.         154

50%       2963.5                      Mean           5364.377
                        Largest       Std. Dev.      5805.149
75%         6965          20421
90%        15925          20585       Variance       3.37e+07
95%        18602          21536       Skewness       1.482589
99%        21536          24484       Kurtosis       4.133748
```

- Now everything is positive. Also notice how the mean, percentiles, variance and standard deviation changes as well.

Missing Data

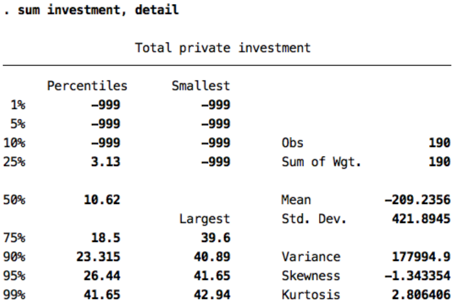
- Now we can look at the *investment* variable:

Missing Data

- Now we can look at the *investment* variable:
  - *sum investment, detail*

Missing Data

- Now we can look at the *investment* variable:
    - *sum investment, detail*



```
. sum investment, detail

                  Total private investment
-------------------------------------------------------------
      Percentiles      Smallest
 1%         -999           -999
 5%         -999           -999
10%         -999           -999       Obs                 190
25%         3.13           -999       Sum of Wgt.         190

50%        10.62                      Mean          -209.2356
                        Largest       Std. Dev.      421.8945
75%         18.5           39.6
90%       23.315          40.89       Variance       177994.9
95%        26.44          41.65       Skewness      -1.343354
99%        41.65          42.94       Kurtosis       2.806406
```

Missing Data

- We have quite a bit of data that is listed as -999. Does this seem right?

Missing Data

- We have quite a bit of data that is listed as -999. Does this seem right?
- Oftentimes when merging data from outside sources, -999 will be used to represent missing data. However, we need to change this in Stata.

Missing Data

- We have quite a bit of data that is listed as -999. Does this seem right?
- Oftentimes when merging data from outside sources, -999 will be used to represent missing data. However, we need to change this in Stata.
- Again, we can fix this using the *recode* command:

Missing Data

- We have quite a bit of data that is listed as -999. Does this seem right?
- Oftentimes when merging data from outside sources, -999 will be used to represent missing data. However, we need to change this in Stata.
- Again, we can fix this using the *recode* command:
  - *recode investment -999=.*

Missing Data

- We have quite a bit of data that is listed as -999. Does this seem right?
- Oftentimes when merging data from outside sources, -999 will be used to represent missing data. However, we need to change this in Stata.
- Again, we can fix this using the *recode* command:
  - *recode investment -999=.*
  - The missing data is now coded with a " . "

Missing Data

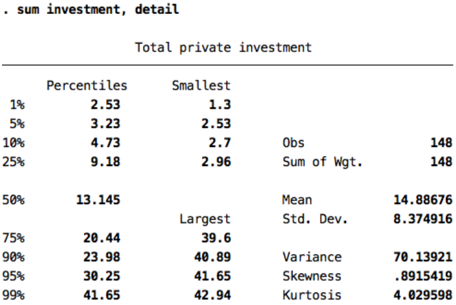
- Let's look at the newly recoded *investment* variable:

Missing Data

- Let's look at the newly recoded *investment* variable:
    - *sum investment, detail*

Missing Data

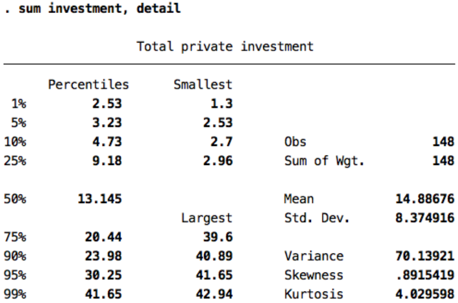
- Let’s look at the newly recoded *investment* variable:
  - *sum investment, detail*



```
. sum investment, detail

                  Total private investment
-------------------------------------------------------------
      Percentiles      Smallest
 1%         2.53            1.3
 5%         3.23           2.53
10%         4.73            2.7       Obs                 148
25%         9.18           2.96       Sum of Wgt.         148

50%       13.145                      Mean           14.88676
                        Largest       Std. Dev.      8.374916
75%        20.44           39.6
90%        23.98          40.89       Variance       70.13921
95%        30.25          41.65       Skewness       .8915419
99%        41.65          42.94       Kurtosis       4.029598
```

Missing Data

- Let's look at the newly recoded *investment* variable:
    - *sum investment, detail*

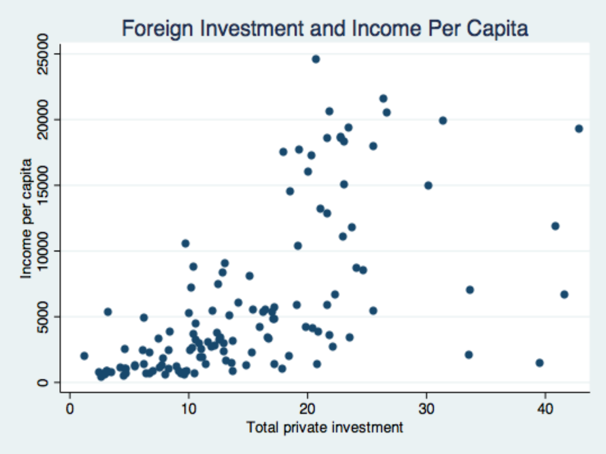


```
. sum investment, detail

                  Total private investment
-------------------------------------------------------------
      Percentiles      Smallest
 1%         2.53            1.3
 5%         3.23           2.53
10%         4.73            2.7       Obs                 148
25%         9.18           2.96       Sum of Wgt.         148

50%       13.145                      Mean           14.88676
                        Largest       Std. Dev.      8.374916
75%        20.44           39.6
90%        23.98          40.89       Variance       70.13921
95%        30.25          41.65       Skewness       .8915419
99%        41.65          42.94       Kurtosis       4.029598
```

- Again notice how many of our descriptive statistics have changed.

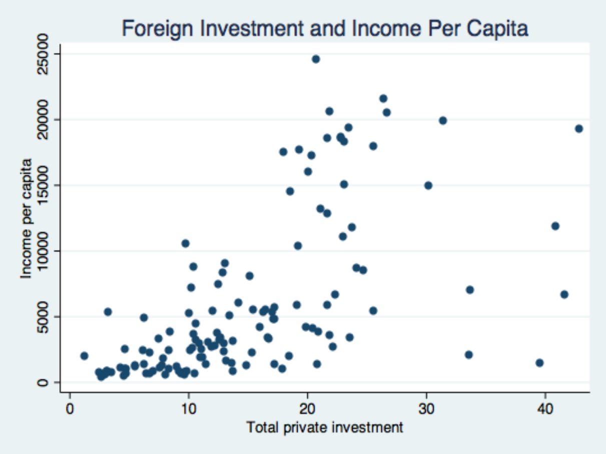
Missing Data

Let's go all the way back and take a look at the scatterplot of income per capita and investment again: *scatter gdppc investment*



Missing Data

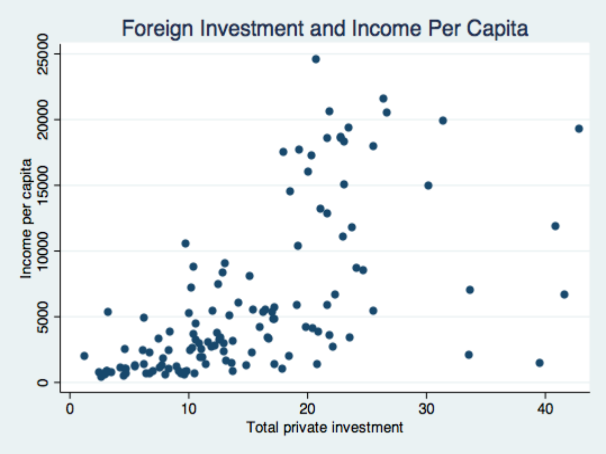
Let's go all the way back and take a look at the scatterplot of income per capita and investment again: *scatter gdppc investment*



- Much better looking scatterplot

Missing Data

Let's go all the way back and take a look at the scatterplot of income per capita and investment again: *scatter gdppc investment*



- Much better looking scatterplot
- correlation = 0.6071 (found by *correlate gdppc investment*)