**\_\_\_**

**\_\_\_\_**

**\_\_\_\_**

**\_\_\_\_**

**\_\_\_\_ (R)**

**/\_\_ /**

**\_\_\_\_/ /**

**\_\_\_\_/**

**\_\_\_/ / /\_\_\_/ / /\_\_\_/ 14.2** Copyright 1985-2015 StataCorp LLC

**Statistics/Data Analysis** StataCorp

4905 Lakeway Drive

College Station, Texas 77845 USA

800-STATA-PC [http://www.stata.com](http://www.stata.com/)

979-696-4600 [stata@stata.com](mailto:stata@stata.com)

979-696-4601 (fax)

Single-user Stata perpetual license: Serial number:

Licensed to:

Notes:

1. Unicode is supported; see help unicode\_advice.

Functional Form Misspecification (Using Stata)

Let us say our model's functional form is not correct, what will happen then?

The assumption of a "0" conditional mean will be violated, and the estimates will be inconsistent.

This assumption is only broken when we have problems related to:

* Improper model specification: When the regression model itself is flawed.
* Endogeneity: When one or more regressors are correlated with the error term.
* Measurement errors: When the behavior of one or more regressors cannot be accurately measured.

To check the functional form of the regression, we use various measures.

Let us try to cover them using Stata.

## Data Set & Graph Matrix

**Download dataset for working with this tutorial -** [**https://bit.ly/46qOnsx**](https://bit.ly/46qOnsx)

### Upload the dataset and use describe - to get variables details.

. use "C:\Users\ASUS\Desktop\hprice2a.dta",clear (Housing price data for Boston-area communities)

. describe

Contains data from **C:\Users\filepath.dta**

obs: **506 Housing price data for Boston-area communities**

vars: **13 5 Oct 2004 09:50**

size: **26,312**

storage display value

variable name type format label variable label

**price** float %9.0g **median housing price, $**

**crime** float %9.0g **crimes committed per capita**

**nox** float %9.0g **nitrous oxide, parts per 100m**

**rooms** float %9.0g **avg number of rooms per house**

**dist** float %9.0g **weighted dist. to 5 employ centers**

**radial** float %9.0g **accessibiliy index to radial hghwys**

**proptax** float %9.0g **property tax per $1000**

**stratio** float %9.0g **average student-teacher ratio**

**lowstat** float %9.0g **% of people 'lower status'**

**lprice** float %9.0g **log(price)**

**lnox** float %9.0g **log(nox)**

**lproptax** float %9.0g **log(proptax)**

**ldist** float %9.0g **log(dist)**

Sorted by:

### Let's do graph matrix to find the relationship b/w lprice and causal factors.

. graph matrix lprice lnox ldist rooms stratio, ms(0h) msize(tiny)name(scatter\_matrix) title("Scatterplot Matrix of Variables")

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |
| (note: | named | style | 0h | not | found | in | class | symbol, | default | attributes | used) |

A picture containing text, diagram, screenshot, line

Description automatically generated

Here,

* We use ms(0h) to focus on the lines or other graphical elements connecting the points rather than the points themselves.
* We use Msize(tiny) to set the size of the markers to "tiny".
* We use name(scatter\_matrix) to specify a prefix for the name of the graph.
* We use title(scatterplot matrix of variable) to add a title to the graph matrix.

*Note: The graphs below the main diagonal can be used to determine if there is a high intercorrelation between the regressors or if there is a problem of collinearity.*

The scatter points between lnox and ldist appear to be linear. Let us explore this relationship further.

### Using the correlated command to check relationship between lnox and ldist

corr lnox ldist (obs=506)

lnox

ldist

lnox ldist

**1.0000**

**-0.8607 1.0000**

*Note: The two variables have a simple correlation of -0.86*

## Added Variable Plot

The added variable plots are used to assess the relationship between an independent variable and the response variable while controlling for other variables in the model.

*Note: In an added variable plot, the residuals of the response variable are plotted on the y-axis, while the residuals of the independent variable are plotted on the x-axis. This allows us to examine whether there is a linear association or not.*

Added variable plots are useful for model validation in regression analysis. They help in detecting potential issues such as non-linearity or outliers.

### Let us generate a new variable room^2 and create an added variable plot of it.

. gen room2 = room^2

. regress lprice lnox ldist rooms room2 stratio lproptax

|  |  |  |  |
| --- | --- | --- | --- |
| Source | SS | df | MS |
| Model | **52.8357813** | **6** | **8.80596356** |
| Residual | **31.7464896** | **499** | **.06362022** |
| Total | **84.5822709** | **505** | **.167489645** |

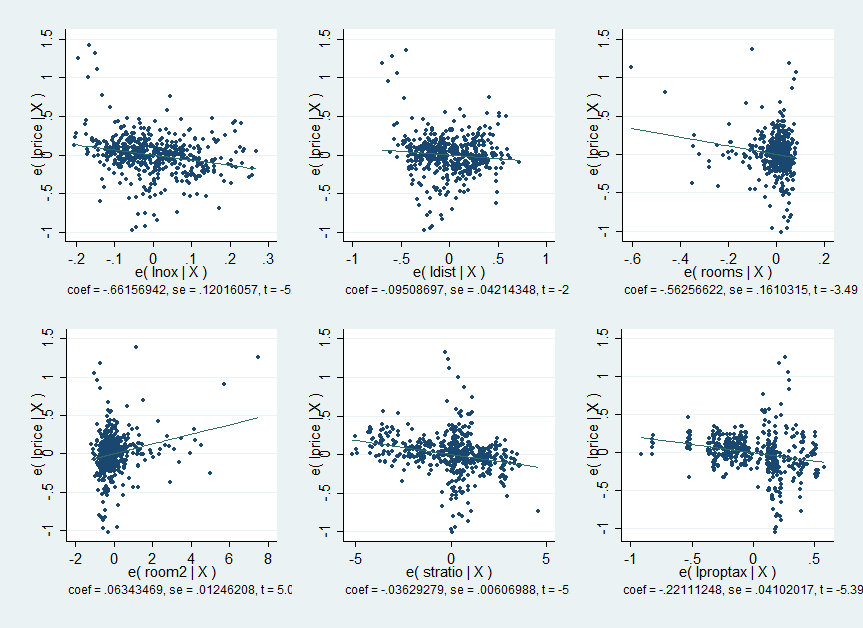
|  |  |  |
| --- | --- | --- |
| Number of obs | = | **506** |
| F(6, 499) | = | **138.41** |
| Prob > F | = | **0.0000** |
| R-squared | = | **0.6247** |
| Adj R-squared | = | **0.6202** |
| Root MSE | = | **.25223** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| lprice | Coef. | Std. Err. | t | P>|t| | [95% Conf. | Interval] |
| lnox | **-.6615694** | **.1201606** | **-5.51** | **0.000** | **-.8976524** | **-.4254864** |
| ldist | **-.095087** | **.0421435** | **-2.26** | **0.024** | **-.1778875** | **-.0122864** |
| rooms | **-.5625662** | **.1610315** | **-3.49** | **0.001** | **-.8789496** | **-.2461829** |
| room2 | **.0634347** | **.0124621** | **5.09** | **0.000** | **.0389501** | **.0879193** |
| stratio | **-.0362928** | **.0060699** | **-5.98** | **0.000** | **-.0482185** | **-.0243671** |
| lproptax | **-.2211125** | **.0410202** | **-5.39** | **0.000** | **-.301706** | **-.1405189** |
| \_cons | **14.15454** | **.5693846** | **24.86** | **0.000** | **13.03585** | **15.27323** |

. avplots, ms(0h) msize(small) col(3)

(note: named style 0h not found in class symbol, default attributes used) (note: named style 0h not found in class symbol, default attributes used) (note: named style 0h not found in class symbol, default attributes used) (note: named style 0h not found in class symbol, default attributes used) (note: named style 0h not found in class symbol, default attributes used) (note: named style 0h not found in class symbol, default attributes used)

*Note: Here, we use col(3) to arrange the plots in three columns.*



*Note: Here, we use col(3) to arrange the plots in three columns. Also, in each pane several observations are far from the straight linking the dependent & the independent variables. Thus, we shall now do Ramsey's RESET.*

### Ramsey's RESET (Regression Equation Specification Error Test)

Ramsey RESET is used to detect potential functional form misspecification. They help determine if the inclusion of additional powers improves the model's fit.

The RESET test statistic, typically an F-statistic, is used to make an inference about the correctness of the model specification. If the p-value associated with the RESET test is below a pre-defined significance level (e.g., 0.05), it suggests that the null hypothesis of correct specification is rejected, indicating a potential misspecification.

Let's proceed to use the Ramsey RESET test.

### To do Ramsey we need regression results - here will run regression quietly.

. quietly regress lprice lnox ldist rooms stratio

*Note: Using quietly will suppress the regression result. Next, we will use the command estat ovtest for Ramsey.*

. estat ovtest, rhs

Ramsey RESET test using powers of the independent variables Ho: model has no omitted variables

F(12, 489) = **11.79**

Prob > F = **0.0000**

*Note: estat ovtest command is used to perform the overidentification test. We write "rhs" to specify that the test should be performed on the right-hand side (exogenous) variables only. Here, we have p-value = 0, means there is specification error.*

## Residual-versus-fitted-plot

It is used to access the specification of the model.

### Let us graph the residuals versus the predicted values for ldist

. quietly regress lprice lnox ldist rooms stratio

. rvpplot ldist, ms(0h) yline(0)

(note: named style 0h not found in class symbol, default attributes used)

A picture containing text, screenshot, diagram, line

Description automatically generated

*Note: In this graph, the residuals appear much more variable for low levels versus high levels of log of distance (ldist), it seems there is constant variance issue.*

## Interaction Terms

To address the issue of specification we sometime introduce interaction terms. For example, we have lproptax and stratio - we can use their interaction. The notion here is, people like to pay lower property tax and prefer schools with low student teacher ratio - so interaction can be - lproptax\*stratio.

. gen interaction = lproptax\*stratio

. regress lprice lnox ldist stratio lproptax interaction

|  |  |  |  |
| --- | --- | --- | --- |
| Source | SS | df | MS |
| Model | **38.7301562** | **5** | **7.74603123** |
| Residual | **45.8521148** | **500** | **.09170423** |
| Total | **84.5822709** | **505** | **.167489645** |

|  |  |  |
| --- | --- | --- |
| Number of obs | = | **506** |
| F(5, 500) | = | **84.47** |
| Prob > F | = | **0.0000** |
| R-squared | = | **0.4579** |
| Adj R-squared | = | **0.4525** |
| Root MSE | = | **.30283** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| lprice | Coef. | Std. Err. | t | P>|t| | [95% Conf. | Interval] |
| lnox | **-.9041103** | **.1441253** | **-6.27** | **0.000** | **-1.187276** | **-.6209444** |
| ldist | **-.1430541** | **.0501831** | **-2.85** | **0.005** | **-.2416499** | **-.0444583** |
| stratio | **-.4388722** | **.1538321** | **-2.85** | **0.005** | **-.7411093** | **-.1366351** |
| lproptax | **-1.48103** | **.5163117** | **-2.87** | **0.004** | **-2.495438** | **-.4666219** |
| interaction | **.0641648** | **.026406** | **2.43** | **0.015** | **.0122843** | **.1160452** |
| \_cons | **21.47905** | **2.952307** | **7.28** | **0.000** | **15.6786** | **27.27951** |

*Note: Here the interaction term is coming out to be significant, however we have not included rooms in the regression model.*

## Outliers

Sometimes the presence of unusual points can change the entire results because they can change the actual coefficient values.

### We can address the problem of outliers by identifying influential data. Let us try this.

. quietly regress lprice lnox ldist rooms room2 stratio lproptax

### Let us generate a new variable "town" with unique identifier for each obs.

. generate town = \_n

### Let us use the predict command to calculate the leverage values and store them in the variable "lev" for the observations in the sample.

. predict double lev if e(sample), leverage

*Note: We use "double" to specify the data type (it is optional). The unusual data point is a problem for least-square regression fit because it alters the estimated coefficients by a sizable amount.*

*Sometime the data points with large residuals have an unusual leverage, we can identify these unusual points using leverage.*

*Also "if e(sample)" this condition ensures that the residuals are only calculated for observations that were part of the estimation sample.*

### Let us use predict command to calculate the residuals and store them in the variable "eps" for the observations in the sample.

. predict double eps if e(sample), res

### Create a variable "eps2" that contains the squared residuals.

. generate double eps2 = eps^2

### Get summary statistics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| . summarize price  Variable | lprice  Obs | Mean | Std. Dev. | Min | Max |
| price | **506** | **22511.51** | **9208.856** | **5000** | **50001** |
| lprice | **506** | **9.941057** | **.409255** | **8.517193** | **10.8198** |

### Let us use sort the dataset in descending order based on the variable lev to produce the descending-sort order.

. gsort -lev

### Let us list down the top 5 observations in terms of lev along with the variable’s town, price, lprice lev and eps2.

. list town price lprice lev eps2 in 1/5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **town** | **price** | **lprice** | **lev** | **eps2** |
| **366** | **27499** | **10.2219** | **.17039262** | **.61813718** |
| **368** | **23100** | **10.04759** | **.11272637** | **.30022048** |
| **365** | **21900** | **9.994242** | **.10947853** | **.33088957** |
| **258** | **50001** | **10.8198** | **.08036068** | **.06047061** |
| **226** | **50001** | **10.8198** | **.0799096** | **.03382768** |

1.

2.

3.

4.

5.

### Let us also get the town with largest squared errors.

. gsort -eps2

. list town price lprice lev eps2 in 1/5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **town** | **price** | **lprice** | **lev** | **eps2** |
| **369** | **50001** | **10.8198** | **.02250047** | **1.7181195** |
| **373** | **50001** | **10.8198** | **.01609848** | **1.4894088** |
| **372** | **50001** | **10.8198** | **.02056901** | **1.2421055** |
| **370** | **50001** | **10.8198** | **.0172083** | **1.0224558** |
| **406** | **5000** | **8.517193** | **.00854955** | **1.0063662** |

1.

2.

3.

4.

5.

Note: Both the results differ, thus a large value of leverage does not imply a large, squared residual and vice-versa.

## DFITS (Difference in Fits)

DFITS values assess how much the predicted values change when a particular observation is removed from the analysis.

Note: Here, we will create a cutoff value to compare with DFITS statistics

absolute value. The cut off value = 1 if it is large and 0 otherwise.

### Let us find DFITS for our dataset.

. predict double dfits if e(sample), dfits

### Let us sort the calculated DFITS statistics in descending order and check the results.

. gsort -dfits

### Create a cut off value based on the degrees of freedom (e(df\_m)), the number of observations (e(N)), and a multiplication factor of 2.

### It uses the formula 2 \* sqrt ((df\_m + 1) / N).

. quietly generate cutoff = abs(dfits) > 2\*sqrt((e(df\_m)+1)/e(N)) & e(sample)

. list town price lprice dfits if cutoff

|  |  |  |  |
| --- | --- | --- | --- |
| **town** | **price** | **lprice** | **dfits** |
| **366** | **27499** | **10.2219** | **1.5679033** |
| **368** | **23100** | **10.04759** | **.82559867** |
| **369** | **50001** | **10.8198** | **.8196735** |
| **372** | **50001** | **10.8198** | **.65967704** |
| **373** | **50001** | **10.8198** | **.63873964** |
| **371** | **50001** | **10.8198** | **.55639311** |
| **370** | **50001** | **10.8198** | **.54354054** |
| **361** | **24999** | **10.12659** | **.32184327** |
| **359** | **22700** | **10.03012** | **.31516743** |
| **408** | **27901** | **10.23642** | **.31281326** |
| **367** | **21900** | **9.994242** | **.31060611** |
| **360** | **22600** | **10.02571** | **.28892457** |
| **363** | **20800** | **9.942708** | **.27393758** |
| **358** | **21700** | **9.985067** | **.24312885** |
| **386** | **7200** | **8.881836** | **-.23838749** |
| **388** | **7400** | **8.909235** | **-.25909393** |
| **491** | **8100** | **8.999619** | **-.26584795** |
| **400** | **6300** | **8.748305** | **-.28782824** |
| **416** | **7200** | **8.881836** | **-.29288953** |
| **402** | **7200** | **8.881836** | **-.29595696** |
| **381** | **10400** | **9.249561** | **-.29668364** |
| **258** | **50001** | **10.8198** | **-.30053391** |
| **385** | **8800** | **9.082507** | **-.302916** |
| **420** | **8400** | **9.035987** | **-.30843965** |

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11.

12.

13.

14.

490.

491.

492.

493.

494.

495.

496.

497.

498.

499.

|  |  |  |  |
| --- | --- | --- | --- |
| **490** | **7000** | **8.853665** | **-.3142718** |
| **401** | **5600** | **8.630522** | **-.33273658** |
| **417** | **7500** | **8.922658** | **-.34950136** |
| **399** | **5000** | **8.517193** | **-.36618139** |
| **406** | **5000** | **8.517193** | **-.37661853** |
| **415** | **7012** | **8.855378** | **-.43879798** |
| **365** | **21900** | **9.994242** | **-.85150064** |

500.

501.

502.

503.

504.

505.

506.

*Note: Above are the observations that satisfy the cutoff criterion. Most of the observations associated with large positive DFITS are those which have a top-coded value of $50,001 for median housing price.*

## DFBETA (Deletion Residuals or change-in-estimate statistics)

It assesses how much the estimated coefficients change when a particular observation is removed from the analysis.

Like DFITS Statistics, let us follow the same steps to generate DFBETA.

. quietly regress lprice lnox ldist rooms room2 stratio lproptax

### Calculate the DFBETA statistic for the variable lnox and store the values it in the variable \_dfbeta\_1.

. dfbeta lnox

\_dfbeta\_1: dfbeta(lnox)

### Create a binary variable named dfcut based on a condition involving the absolute value of \_dfbeta\_1 and a threshold value.

### Note: The above condition checks if the absolute value of \_dfbeta\_1 is greater than 2/sqrt(e(N)) and if the observation is part of the sample.

. quietly generate dfcut = abs(\_dfbeta\_1) > 2/sqrt(e(N)) & e(sample)

. sort \_dfbeta\_

. summarize lnox

Variable

Obs

Mean

Std. Dev.

Min

Max

lnox

**506**

**1.693091**

**.2014102 1.348073 2.164472**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **town** | **price** | **lprice** | **lnox** | **\_dfbeta\_1** |
| **369** | **50001** | **10.8198** | **1.842136** | **-.4316933** |
| **372** | **50001** | **10.8198** | **1.842136** | **-.4257791** |
| **373** | **50001** | **10.8198** | **1.899118** | **-.3631822** |
| **371** | **50001** | **10.8198** | **1.842136** | **-.2938702** |
| **370** | **50001** | **10.8198** | **1.842136** | **-.2841335** |
| **365** | **21900** | **9.994242** | **1.971299** | **-.2107066** |
| **408** | **27901** | **10.23642** | **1.885553** | **-.1728729** |
| **368** | **23100** | **10.04759** | **1.842136** | **-.1309522** |
| **11** | **15000** | **9.615806** | **1.656321** | **-.1172723** |
| **410** | **27499** | **10.2219** | **1.786747** | **-.1117743** |
| **413** | **17900** | **9.792556** | **1.786747** | **-.0959273** |
| **437** | **9600** | **9.169518** | **2.00148** | **-.0955826** |
| **146** | **13800** | **9.532424** | **2.164472** | **-.0914387** |
| **154** | **19400** | **9.873029** | **2.164472** | **.0910494** |
| **463** | **19500** | **9.87817** | **1.964311** | **.0941472** |
| **464** | **20200** | **9.913438** | **1.964311** | **.0974507** |
| **427** | **10200** | **9.230143** | **1.764731** | **.1007114** |
| **406** | **5000** | **8.517193** | **1.93586** | **.1024767** |
| **151** | **21500** | **9.975808** | **2.164472** | **.1047597** |
| **152** | **19600** | **9.883285** | **2.164472** | **.1120427** |
| **460** | **20000** | **9.903487** | **1.964311** | **.1142668** |
| **160** | **23300** | **10.05621** | **2.164472** | **.1165014** |
| **491** | **8100** | **8.999619** | **1.806648** | **.1222368** |
| **362** | **19900** | **9.898475** | **2.04122** | **.1376445** |
| **363** | **20800** | **9.942708** | **2.04122** | **.1707894** |
| **490** | **7000** | **8.853665** | **1.806648** | **.1791869** |
| **358** | **21700** | **9.985067** | **2.04122** | **.1827834** |
| **360** | **22600** | **10.02571** | **2.04122** | **.2209745** |
| **361** | **24999** | **10.12659** | **2.04122** | **.2422512** |
| **359** | **22700** | **10.03012** | **2.04122** | **.2483543** |

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11.

12.

13.

490.

491.

492.

493.

494.

495.

496.

497.

498.

499.

500.

501.

502.

503.

504.

505.

506.

*Note: Just like DFITS we have similar patterns for the DFBETA for lnox. The sample here, exhibiting large values for $50,001 of median housing price which confirm outliers.*

#### PS - The problem of this type - removing the bottom and top observations from the sample can be done using censoring (coding extreme value) with Tobit model.