Regression Diagnostics:

In the previous chapter, we learned how to do ordinary linear regression with Stata, concluding with methods for examining the distribution of our variables. Without verifying that your data have met the assumptions underlying OLS regression, your results may be misleading. Now we will explore how you can use Stata to check on how well your data meet the assumptions of OLS regression. Particularly we consider the following diagnostics:

1-Checking for unusual or influential data:

A single observation that is substantially different from all other observations can make a large difference in the results of your regression analysis. If a single observation (or small group of observations) substantially changes your results, you would want to know about this and investigate further. There are three ways that an observation can be unusual.

Outliers: In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

Leverage: An observation with an extreme value on a predictor variable is called a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. These leverage points can have an effect on the estimate of regression coefficients.

Influence: An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness.

Example in Stata:

How can we identify the unusual observations in Stata? Let's consider an example dataset called crime. This dataset appears in Statistical Methods for Social Sciences, Third Edition by Alan Agresti and Barbara Finlay (Prentice Hall, 1997). You can access this data file over the web from within Stata with the Stata use command as shown below:

REGRESSION DIAGNOSTICS

```
.use http://www.ats.ucla.edu/stat/stata/webbooks/reg/crime
```

The variables are state id (sid), state name (state), violent crimes per 100,000 people (crime), murders per 1,000,000 (murder), the percent of the population living in metropolitan areas (pctmetro), the percent of the population that is white (pctwhite), percent of population with a high school education or above (pcths), percent of population living under poverty line (poverty), and percent of population that are single parents (single).

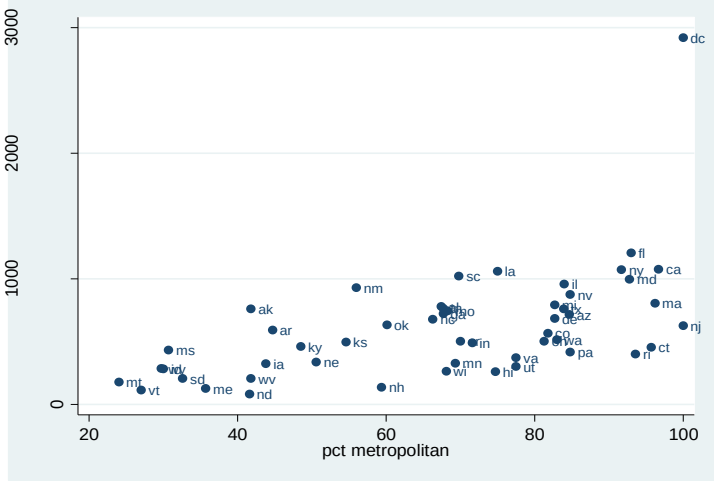
Let's say we want to predict crime by pctmetro, poverty, and single. In other words we want to build a linear regression model between the response variable crime and the independent variables pctmetro, poverty and single. We will first look at the scatter plots of crime against each of the predictor variables before running the regression so we will have some ideas about potential problems by typing this command in Stata:

```
.graph matrix crime pctmetro poverty single
```

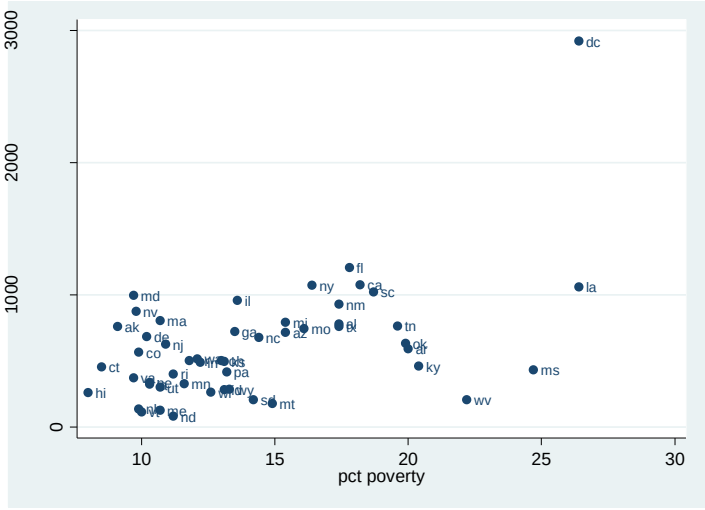
The graphs of crime with other variables, the first row, show some potential problems. In every plot, we see a data point that is far away from the rest of the data points. Let's make individual graphs of crime with pctmetro and poverty and single so we can get a better view of these scatter plots. Also add the mlabel(state) option to label each marker with the state name to identify outlying states.

REGRESSION DIAGNOSTICS

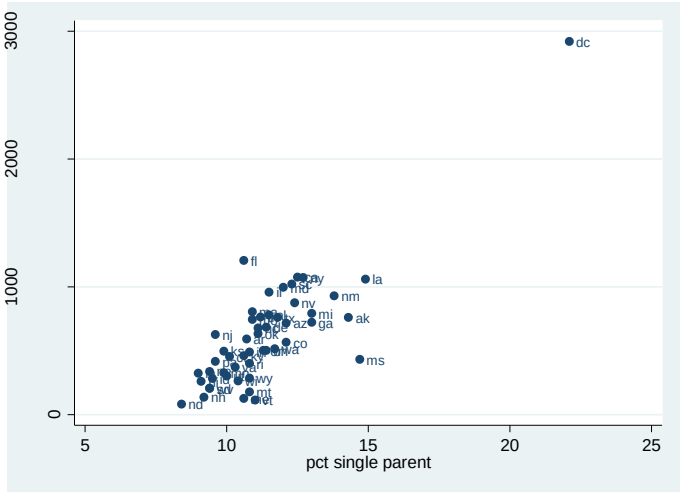
```
.scatter crime pctmetro, mlabel(state)
```




```
.scatter crime poverty, mlabel(state)
```



```
.scatter crime single, mlabel(state)
```



As we observe that all the scatter plots suggest that the observation for `state = dc` is a point that requires extra attention since it stands out away from all of the other points. We will keep it in mind when we do our regression analysis.

1-Checking for Normality of residual:

Many researchers believe that multiple regressions require normality. This is not the case. Normality of residuals is only required for valid hypothesis testing, that is, the normality assumption assures that the p-values for the t-tests and F-test will be valid. Normality is not required in order to obtain unbiased estimates of the regression coefficients. OLS regression merely requires that the residuals (errors) be identically and independently distributed.

After we run a regression analysis, we can use the `predict` command to create residuals and then use commands such as `kdensity`, `qnorm` and `pnorm` to check the normality of the residuals.

Let's use the `elemapi2.dta` data file we used in our regression analysis. You can access this data file over the web from within Stata with the Stata use command as shown below:

REGRESSION DIAGNOSTICS

```
.use http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi2
```

Let's predict academic performance (`api00`) from percent receiving free meals (`meals`), percent of English language learners (`ell`), and percent of teachers with emergency credentials (`emer`).

`regress api00 meals ell emer`

Source	SS	df	MS		
Model	6749782.75	3	2249927.58	Number of obs	= 400
Residual	1323889.25	396	3343.15467	F(3, 396)	= 673.00
				Prob > F	= 0.0000
Total	8073672	399	20234.7669	R-squared	= 0.8360
				Adj R-squared	= 0.8348
				Root MSE	= 57.82

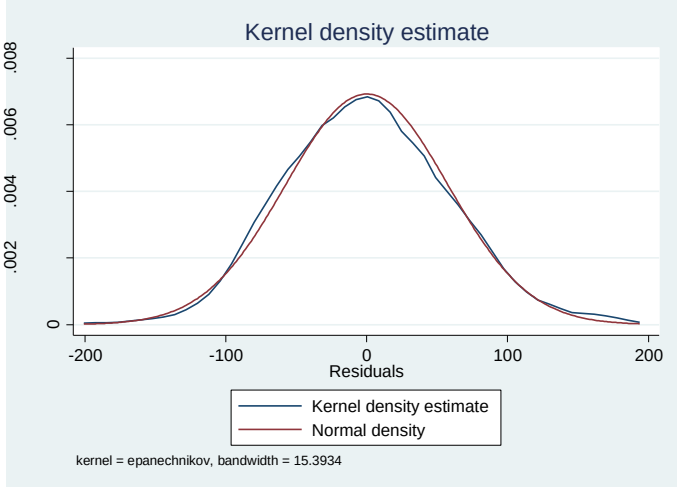
api00	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
meals	-3.159189	.1497371	-21.10	0.000	-3.453568	-2.864809
ell	-.9098732	.1846442	-4.93	0.000	-1.272879	-.5468678
emer	-1.573496	.293112	-5.37	0.000	-2.149746	-.9972456
_cons	886.7033	6.25976	141.65	0.000	874.3967	899.0098

Now use the `predict` command to generate residuals.

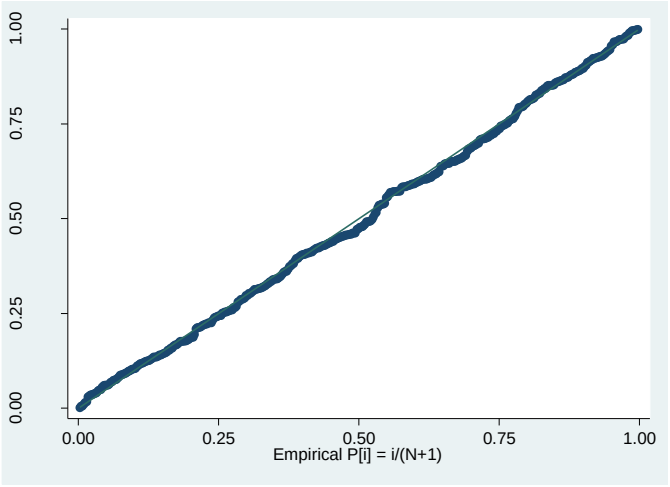
```
.predict r, resid
```

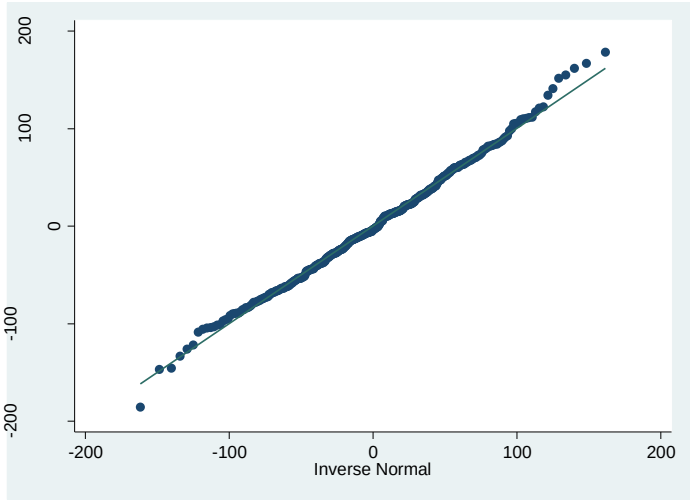
After we store the residuals from the above regression in variable r we use the `kdensity` command to produce a kernel density plot with the `normal` option requesting that a normal density be overlaid on the plot. `kdensity` stands for kernel density estimate. It can be thought of as a histogram with narrow bins and moving average. By typing the following command:

```
.kdensity r, normal
```



We can also use `pnorm` command to graph the standardized normal probability (P-P) plot and `qnorm` plots the quantiles of a variable against the quantiles of a normal distribution.

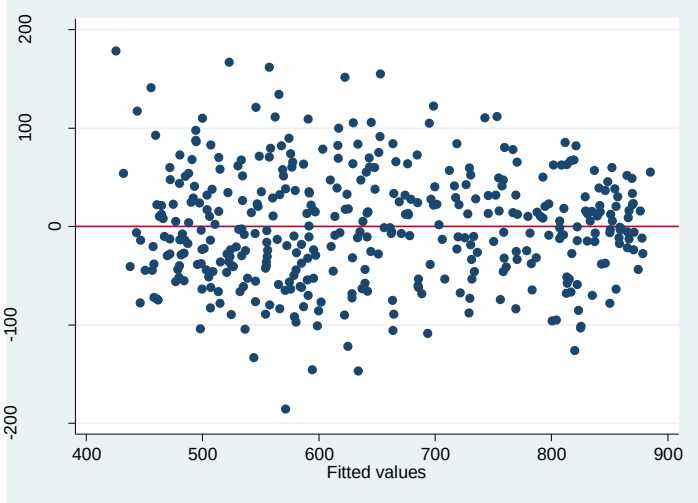




Nevertheless, this seems to be a minor and trivial deviation from normality in that case. Therefore, we can accept that the residuals are close to a normal distribution.

2-Checking Homoscedasticity of Residuals

One of the main assumptions for the ordinary least squares regression is the homogeneity of variance of the residuals. If the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. If the variance of the residuals is non-constant then the residual variance is said to be "heteroscedastic". There are graphical and non-graphical methods for detecting heteroscedasticity. A commonly used graphical method is to plot the residuals versus fitted (predicted) values. We do this Stata by issuing the `rvfplot` command. Below we use the `rvfplot` command with the `yline(0)` option to put a reference line at y=0.



We see that the pattern of the data points is getting a little narrower towards the right end, which is an indication of heteroscedasticity. Now use other commands that test the heteroscedasticity.

```
.estat imtest

Cameron & Trivedi's decomposition of IM-test
```

Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	18.35	9	0.0313
Skewness	7.78	3	0.0507
Kurtosis	0.27	1	0.6067
Total	26.40	13	0.0150

```
.estat hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of api00

         chi2(1)      =     8.75
         Prob > chi2  =   0.0031
```

The first test on heteroskedasticity given by `imest` is the White's test and the second one given by hettest is the `Breusch-Pagan` test. Both test the null hypothesis that the variance of the residuals is homogenous. Therefore, if the p-value is very small, we would have to reject the hypothesis and accept the alternative hypothesis that the variance is not homogenous. So in this case, the evidence is against the null hypothesis that the variance is homogeneous. Therefore it is a common practice to combine the tests with diagnostic plots to make a judgment on the severity of the heteroscedasticity and to decide if any correction is needed for heteroscedasticity.

3-Checking for Multicollinearity:

When there is a perfect linear relationship among the predictors, the estimates for a regression model cannot be uniquely computed. The term collinearity implies that two variables are near perfect linear combinations of one another. When more than two variables are involved it is often called multicollinearity, although the two terms are often used interchangeably.

The primary concern is that as the degree of multicollinearity increases, the regression model estimates of the coefficients become unstable and the standard errors for the coefficients can get wildly inflated. In this section, we will explore some Stata commands that help to detect multicollinearity.

We can use the `vif` command after the regression to check for multicollinearity. `vif` stands for *variance inflation factor*. As a rule of thumb, a variable whose VIF values are greater than 10 may merit further investigation. Tolerance, defined as 1/VIF, is used by many researchers to check on the degree of collinearity. A tolerance value lower than 0.1 is comparable to a VIF of 10. It means that the variable could be considered as a linear combination of other independent variables. Let's run the `vif` test in Stata:

```
. vif
    Variable |      VIF       1/VIF
-------------+----------------------
       meals |     2.73    0.366965
         ell |     2.51    0.398325
        emer |     1.41    0.706805
-------------+----------------------
    Mean VIF |     2.22
```

The VIFs look fine here so there is no multicolinearity problem in this regression.

The .do file content:

```
. capture log close
. log using RegressionDiagnostics_output.txt, text replace

.//  This dataset appears in Statistical Methods for Social Sciences, Third Edition by Alan Agresti and

.// Barbara Finlay (Prentice Hall, 1997).

.use http://www.ats.ucla.edu/stat/stata/webbooks/reg/crime, clear

.cd "S:\DRC\Help_Sheets\Zeb_Help\STATA_Help"

.save crime, replace

.graph matrix crime pctmetro poverty single

.scatter crime pctmetro, mlabel(state)

.scatter crime poverty, mlabel(state)

.scatter crime single, mlabel(state)
```

```
.use http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi2, clear

.regress api00 meals ell emer


.// Normality test

.predict r, resid

.kdensity r, normal

.// use pnorm command to graph the standardized normal probability (P-P) plot

.pnorm r

.// qnorm plots the quantiles of a variable against the quantiles of a normal distribution.

.qnorm r


.// Homoscedasticity tests

.* residuals vs. fitted plot:

.rvfplot , yline(0)

.estat imtest

.estat hettest


.// Multicollinearity test:
```

```
.* 1. variance inflation factor : a variable whose VIF values are greater than 10 may merit further
.*investigation. Tolerance, defined as 1/VIF, is used by many researchers to check on the degree of
.*collinearity. A tolerance value lower than 0.1 is comparable to a VIF of 10.

.vif

.log close
```