Multiple linear regression

We will use a car data set and our dependent variable is mpg (miles per gallon) vs our independent variables car weight and length

Data Editor (Edit) - [regression_auto.dta]

File Edit View Data Tools

make[1] AMC

	make	mpg	weight	weight1	price	foreign	repairs	length
1	AMC	22	2930	2.93	4099	0	3	186
2	AMC	17	3350	3.35	4749	0	3	173
3	AMC	22	2640	2.64	3799	0	3	168
4	Audi	17	2830	2.83	9690	1	5	189
5	Audi	23	2070	2.07	6295	1	3	174
6	BMW	25	2650	2.65	9735	1	4	177
7	Buick	20	3250	3.25	4816	0	3	196
8	Buick	15	4080	4.08	7827	0	4	222
9	Buick	18	3670	3.67	5788	0	3	218
10	Buick	26	2230	2.23	4453	0	3	170
11	Buick	20	3280	3.28	5189	0	3	200
12	Buick	16	3880	3.88	10372	0	3	207

Variables

Filter variables here

Name	Label
make	make of car
mpg	miles per gallon
weight	car weight in pounds
weight1	car weight in 1,000 pounds
price	car price in 1978 dollars
foreign	=1 if car is foreign
repairs	number of car repairs
length	car length

Description of our response variable and predictor variable but let's first create our global variables

```
global ylist mpg
global xlist weight1 length
```

```
. describe $ylist $xlist

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
mpg             byte    %8.0g                 miles per gallon
weight1         float   %9.0g                 car weight in 1,000 pounds
length          int     %8.0g                 car length

. summarize $ylist $xlist

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         mpg |         26    20.92308    4.757504         14         35
     weight1 |         26    3.099231    .6950794       2.02       4.33
      length |         26    190.0769    18.17014        163        222

. summarize $ylist, detail

                      miles per gallon
-------------------------------------------------------------
      Percentiles      Smallest
 1%           14             14
 5%           14             14
10%           15             15       Obs                  26
25%           17             16       Sum of Wgt.          26

50%           21                      Mean           20.92308
                        Largest       Std. Dev.      4.757504
75%           23             25
90%           26             26       Variance       22.63385
95%           29             29       Skewness       .8806144
99%           35             35       Kurtosis       4.243808
```

Question
Can the weight and length statistically significantly predict a car's miles per gallon?

Hypothesis
H0: Car weight and length can not statistically significantly predict a car's miles per gallon
Ha: Car weight and length can statistically significantly predict a car's miles per gallon

The level of significance
alpha = 0.05

ASSUMPTIONS

Determine if data meets requirements to perform a linear regression.

Assumption #1: Your response variable should be measured on a continuous scale.
Assumption #2: You have two or more independent variables, which should be measured at the continuous or categorical level.
Assumption #3: There should be no significant outliers, high leverage points or highly influential points, which represent observations in your data set that are in some way unusual. Lets first plot a box plot:

```
. graph box mpg weight1 length
```

As we can see from our box plot, we have an extreme mpg value. Let's find it and drop the observation.

```
. egen Q1_mpg= pctile(mpg), p(25)

. egen Q3_mpg= pctile(mpg), p(75)

. egen  IC_mpg= iqr(mpg)

. gen touse=1 if (mpg< Q1_mpg-1.5*IC_mpg| mpg> Q3_mpg+1.5*IC_mpg) & missing(mpg)==0
(25 missing values generated)

. recode touse . =0
(touse: 25 changes made)

. tab touse

      touse |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         25       96.15       96.15
          1 |          1        3.85      100.00
------------+-----------------------------------
      Total |         26      100.00

.
. * or use extremes command as follows
. extremes mpg, iqr(1.5)

  +--------------------+
  | obs:    iqr:   mpg |
  |--------------------|
  |  24.   2.000    35 |
  +--------------------+


.
. *Drop the outliers
. drop if (mpg< Q1_mpg-1.5*IC_mpg | mpg> Q3_mpg+1.5*IC_mpg)
(1 observation deleted)
```

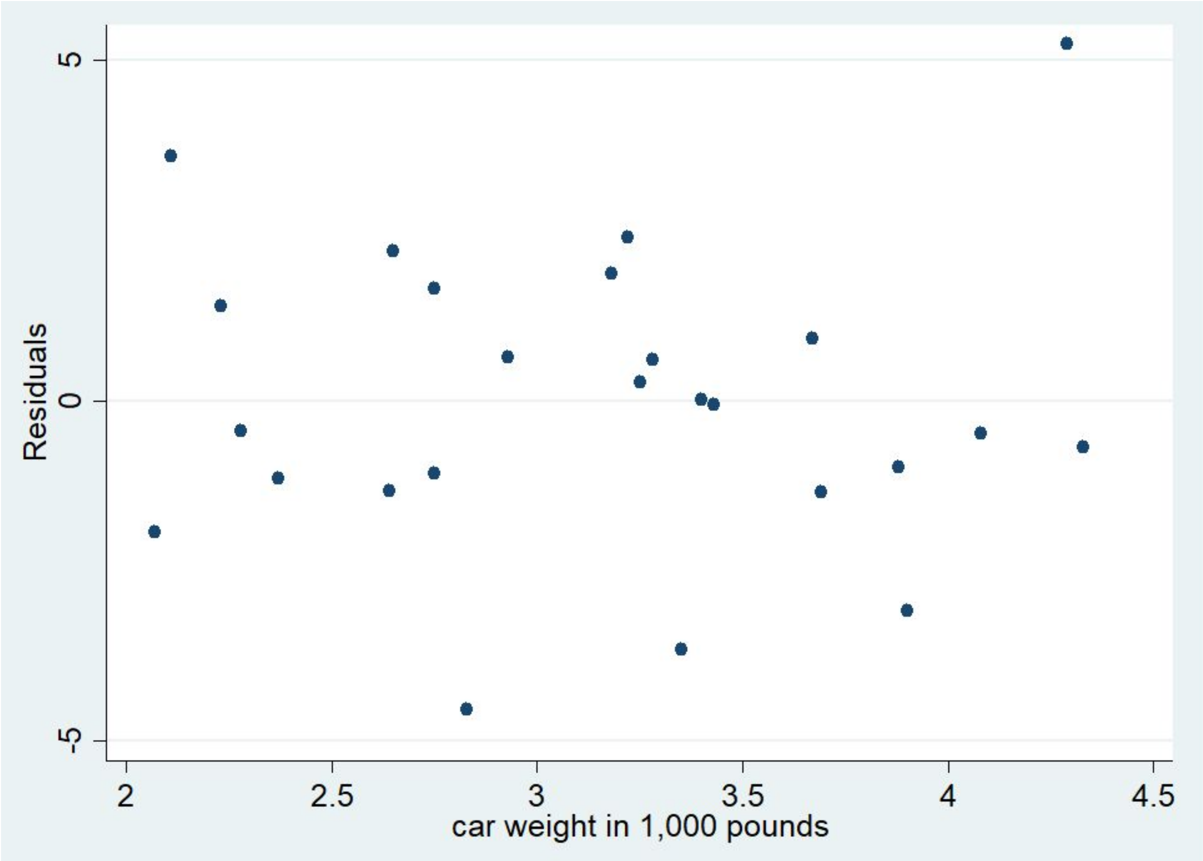
Assumption #4: You should have independence of observations (i.e., independence of residuals), which you can check in Stata using the Durbin-Watson statistic.

```
. //generate time variable
. gen t = _n

. tsset t
        time variable:  t, 1 to 25
                delta:  1 unit

.
. *Run regression first before getting Durbin-Watson statistic as shown below
. * Simple regression
. reg $ylist $xlist

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(2, 22)        =     23.17
       Model |  243.943195         2  121.971597   Prob > F        =    0.0000
    Residual |  115.816805        22  5.26440024   R-squared       =    0.6781
-------------+----------------------------------   Adj R-squared   =    0.6488
       Total |      359.76        24       14.99   Root MSE        =    2.2944

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     weight1 |  -3.382251   1.582302    -2.14   0.044    -6.663745   -.1007576
      length |  -.0552394   .0598333    -0.92   0.366    -.1793262    .0688473
       _cons |   41.54354   7.317217     5.68   0.000     26.36856    56.71852
------------------------------------------------------------------------------

.
. dwstat

Durbin-Watson d-statistic(  3,    25) =  1.616561
```

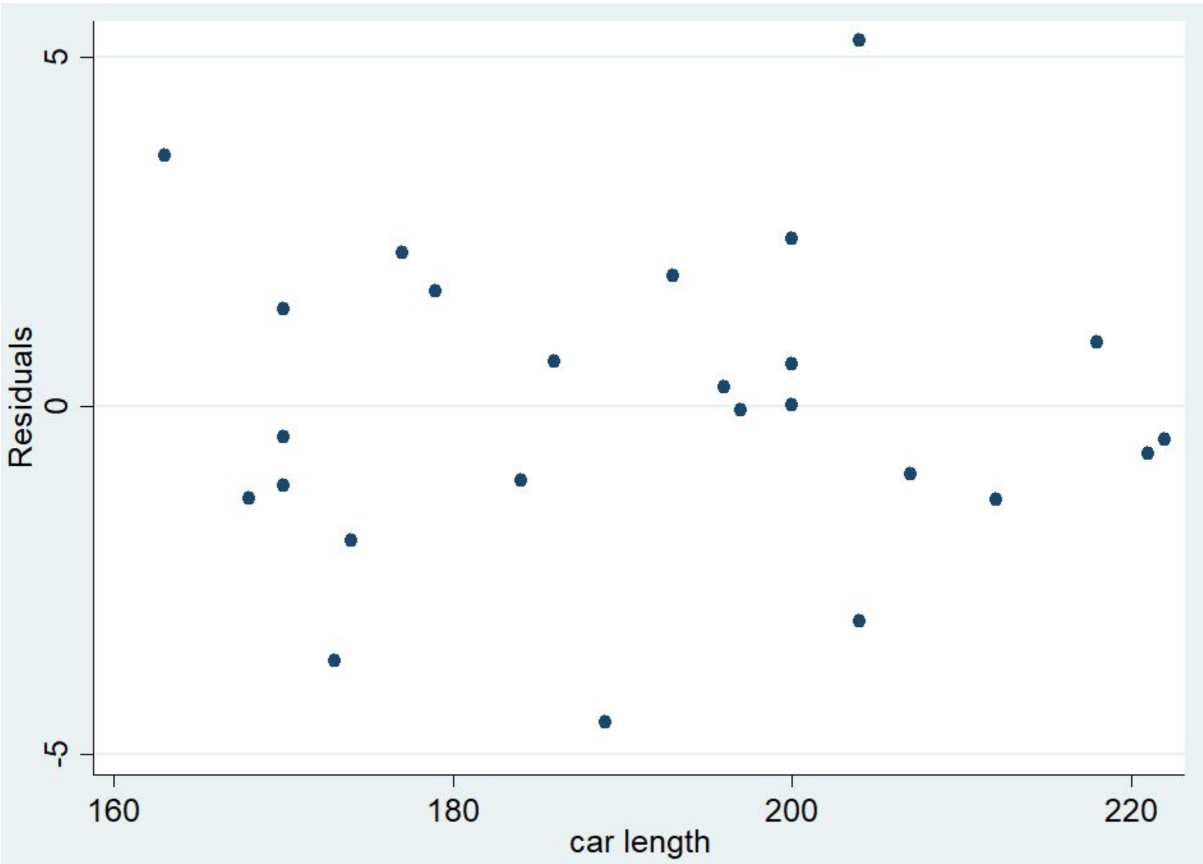
Values of 1.5 < d < 2.5 generally show that there is no autocorrelation in the data while values
0 to 2< means there is positive autocorrelation and values >2 to 4 means there is negative autocorrelation.

Assumption #5: There needs to be a linear relationship between (a) the dependent variable and each of your independent variables, and (b) the dependent variable and the independent variables collectively. Checking the linearity assumption is not so straightforward in the case of multiple regression as is in simple linear regression.
One thing to do is to plot the standardized residuals against each of the predictor variables in the regression model. If there is a clear nonlinear pattern, there is a problem of nonlinearity.

```
. scatter ehat weight1
```



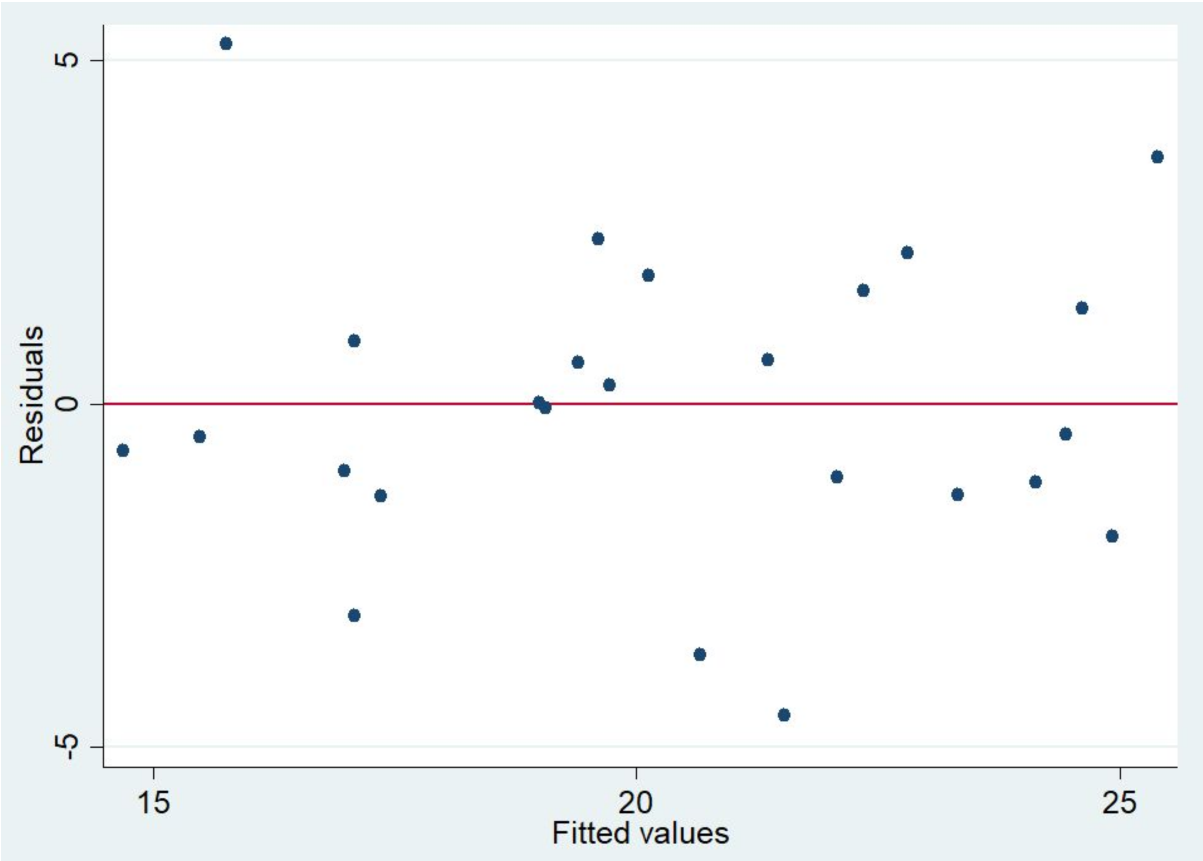

```
. scatter ehat length
```



The two residual versus predictor variable plots above do not indicate strongly a clear departure from linearity

Assumption #5: Your data needs to show homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line. One of the main assumptions for the ordinary least squares regression is the homogeneity of variance of the residuals. A commonly used graphical method is to plot the residuals versus fitted (predicted) values. We do this by issuing the **rvfplot** command.

```
. rvfplot, yline(0)
```



Now let’s look at a couple of commands that test for heteroscedasticity.

```
. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      12.42      5    0.0294
            Skewness |       6.69      2    0.0353
            Kurtosis |       0.12      1    0.7339
---------------------+-----------------------------
               Total |      19.23      8    0.0137
---------------------------------------------------

. estat hettest

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

         chi2(1)      =     0.08
         Prob > chi2  =   0.7777
```

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 not against the null hypothesis that the variance is homogeneous. These tests are very sensitive to model assumptions, such as the assumption of normality. 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.

Assumption #6: Your data must not show multicollinearity, which occurs when you have two or more independent variables that are highly correlated with each other. 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.
We can use the **vif** command after the regression to check for multicollinearity. **vif** stands for *variance inflation factor*.

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      length |      5.17    0.193490
     weight1 |      5.17    0.193490
-------------+----------------------
    Mean VIF |      5.17
```

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.

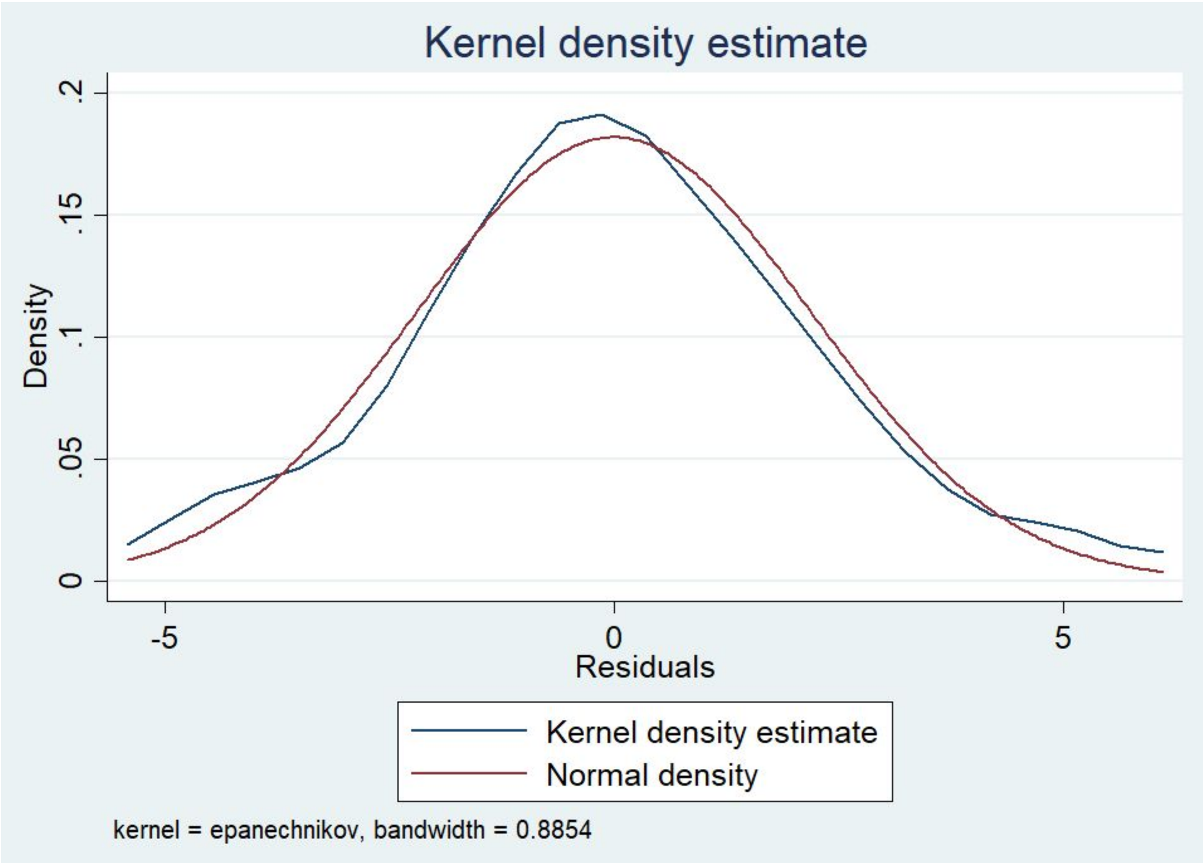
Assumption #8: The residuals (errors) should be approximately normally distributed.

```
. reg $ylist $xlist

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(2, 22)        =     23.17
       Model |  243.943195         2  121.971597   Prob > F        =    0.0000
    Residual |  115.816805        22  5.26440024   R-squared       =    0.6781
-------------+----------------------------------   Adj R-squared   =    0.6488
       Total |      359.76        24       14.99   Root MSE        =    2.2944

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     weight1 |  -3.382251   1.582302    -2.14   0.044    -6.663745   -.1007576
      length |  -.0552394   .0598333    -0.92   0.366    -.1793262    .0688473
       _cons |   41.54354   7.317217     5.68   0.000     26.36856    56.71852
------------------------------------------------------------------------------
```

```
. predict ehat, resid
```

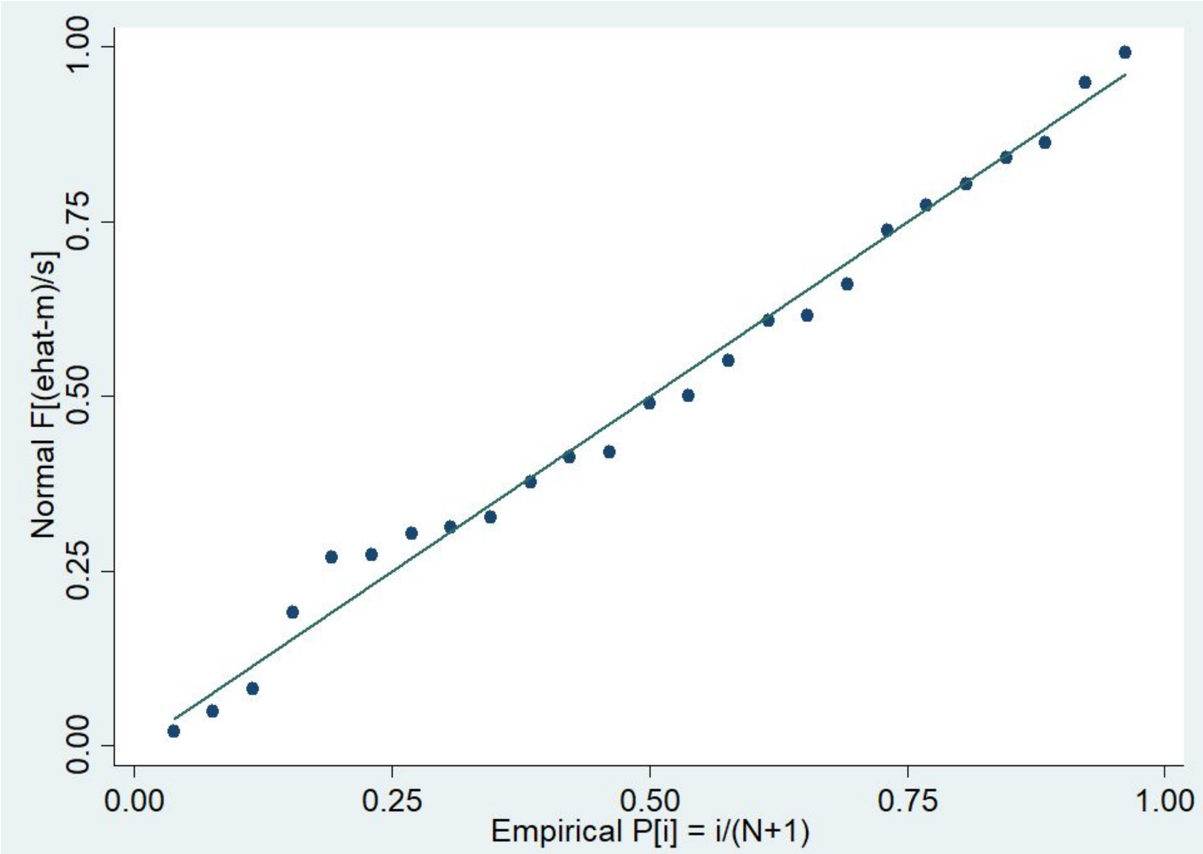
Below we use the **kdensity** command to produce a kernel density plot with the **normal** option requesting that a normal density be overlaid on the plot. **kdensity** stands for kernel density estimate. It can be thought of as a histogram with narrow bins and moving average.

```
. *kdensity command to produce a kernel density plot
. kdensity ehat, normal
(n() set to 25)
```

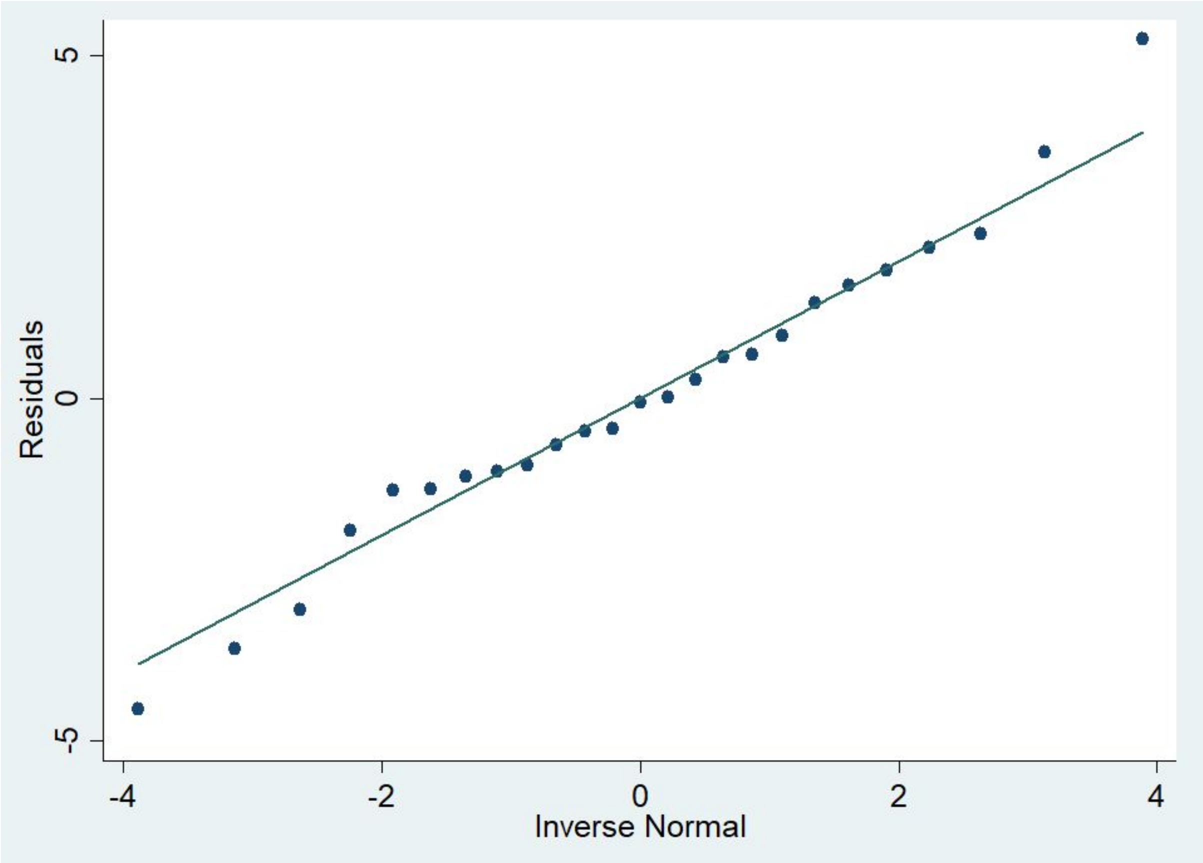


The **pnorm** command graphs a standardized normal probability (P-P) plot while **qnorm** plots the quantiles of a variable against the quantiles of a normal distribution. **pnorm** is sensitive to non-normality in the middle range of data and **qnorm** is sensitive to non-normality near the tails. As you see below, the results from **pnorm** show no indications of non-normality, while the **qnorm** command shows a slight deviation from normal at the upper tail, as can be seen in the **kdensity** above. Nevertheless, this seems to be a minor and trivial deviation from normality. We can accept that the residuals are close to a normal distribution.

```
. pnorm ehat
```




```
. qnorm ehat
```



There are also numerical tests for testing normality.

```
. swilk ehat

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
        ehat |         25    0.98588      0.392    -1.913    0.97214
```

As seen above Shapiro-Wilk W test is not significant which means the residuals (errors) are approximately normally distributed.

Multiple Linear Regression

```
. reg $ylist $xlist

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(2, 22)        =     23.17
       Model |  243.943195         2  121.971597   Prob > F        =    0.0000
    Residual |  115.816805        22  5.26440024   R-squared       =    0.6781
-------------+----------------------------------   Adj R-squared   =    0.6488
       Total |      359.76        24       14.99   Root MSE        =    2.2944

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     weight1 |  -3.382251   1.582302    -2.14   0.044    -6.663745   -.1007576
      length |  -.0552394   .0598333    -0.92   0.366    -.1793262    .0688473
       _cons |   41.54354   7.317217     5.68   0.000     26.36856    56.71852
------------------------------------------------------------------------------
```

A multiple regression was run to predict mpg (miles covered by the car per gallon) from car weight, its price and whether it's foreign or not. These variables statistically significantly predicted mpg (miles covered by the car per gallon) , $F(2, 22) = 23.17$, $p < .0005$, $R^2 = .678$. But we can see that car weight is the only significant predictor. The regression equation was: predicted miles covered by the car per gallon = 41.544 - 3.382*(weight of the car)-0.055*(Car Length)

```
. correlate $ylist $xlist
(obs=25)

             |      mpg  weight1   length
-------------+---------------------------
         mpg |   1.0000
     weight1 |  -0.8158   1.0000
      length |  -0.7818   0.8981   1.0000
```