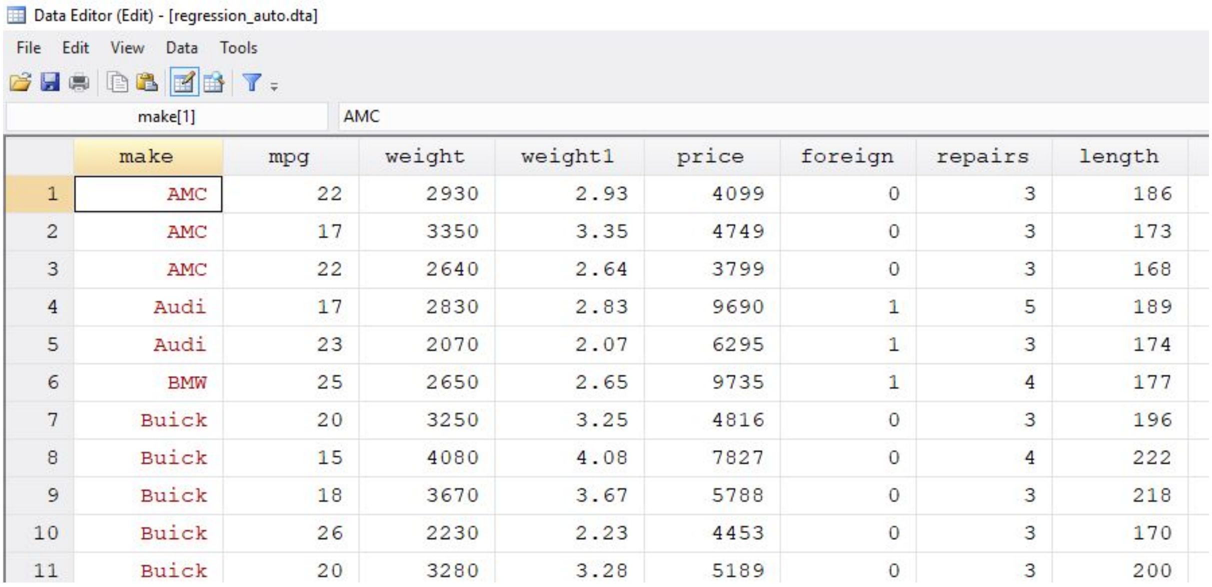
Simple linear regression

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

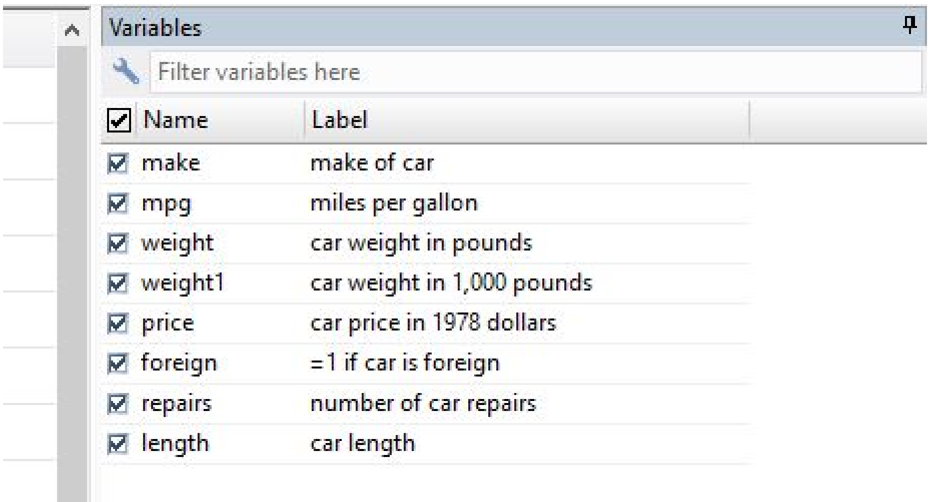


Data Editor (Edit) - [regression_auto.dta]

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



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 lets first create our global variables

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

```
. 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
```

Data Summary

```
. 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
```

```
. 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 of a car statistically significantly predict a car's miles per gallon?

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

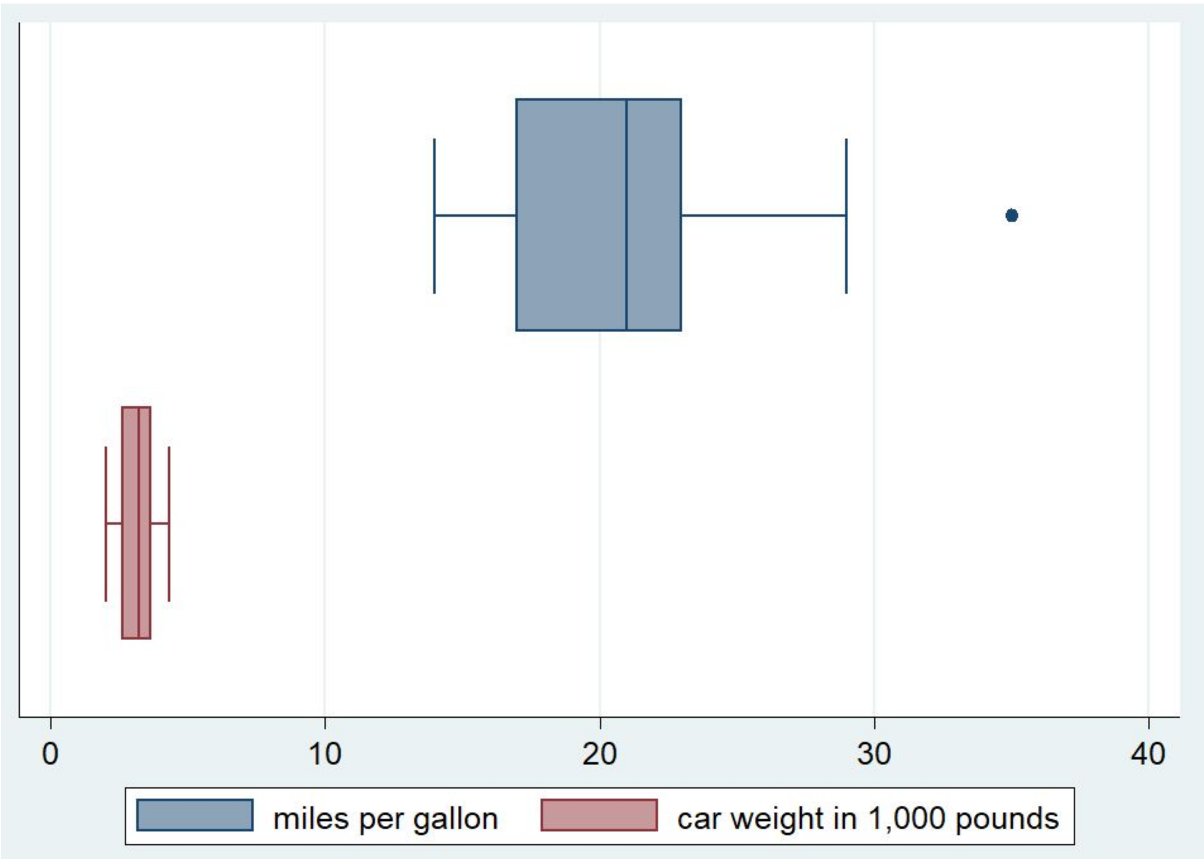
The level of significance
alpha = 0.05

<u>ASSUMPTIONS</u>
Determine if data meets requirements to perform a linear regression.

Assumption #1: Your response variable should be measured on a continuous scale.
Assumption #2: Your independent variable should be measured at the continuous or categorical level.

Assumption #5: There should be no significant outliers. We can use box plot

```
. graph hbox mpg weight1
```

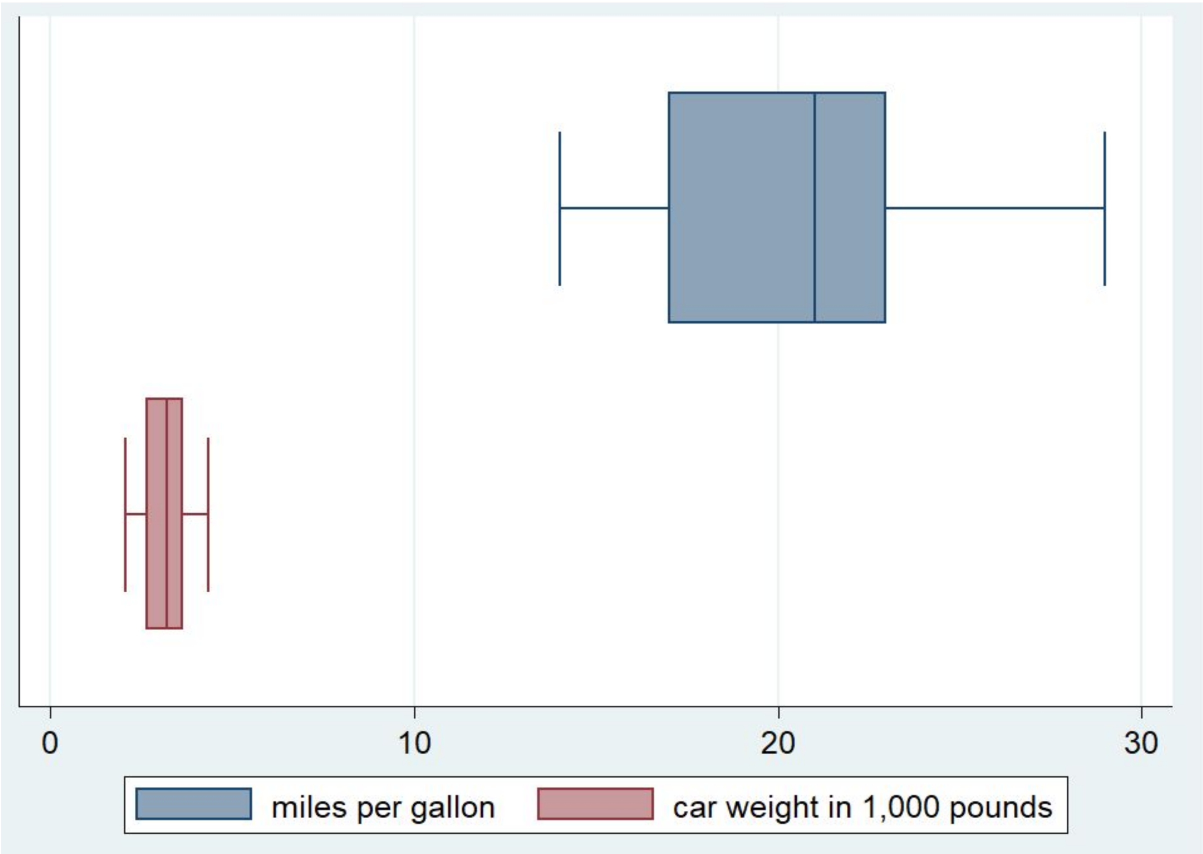


We can see that our response variable mpg has an extreme value, let's drop it.

```
. 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
```

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

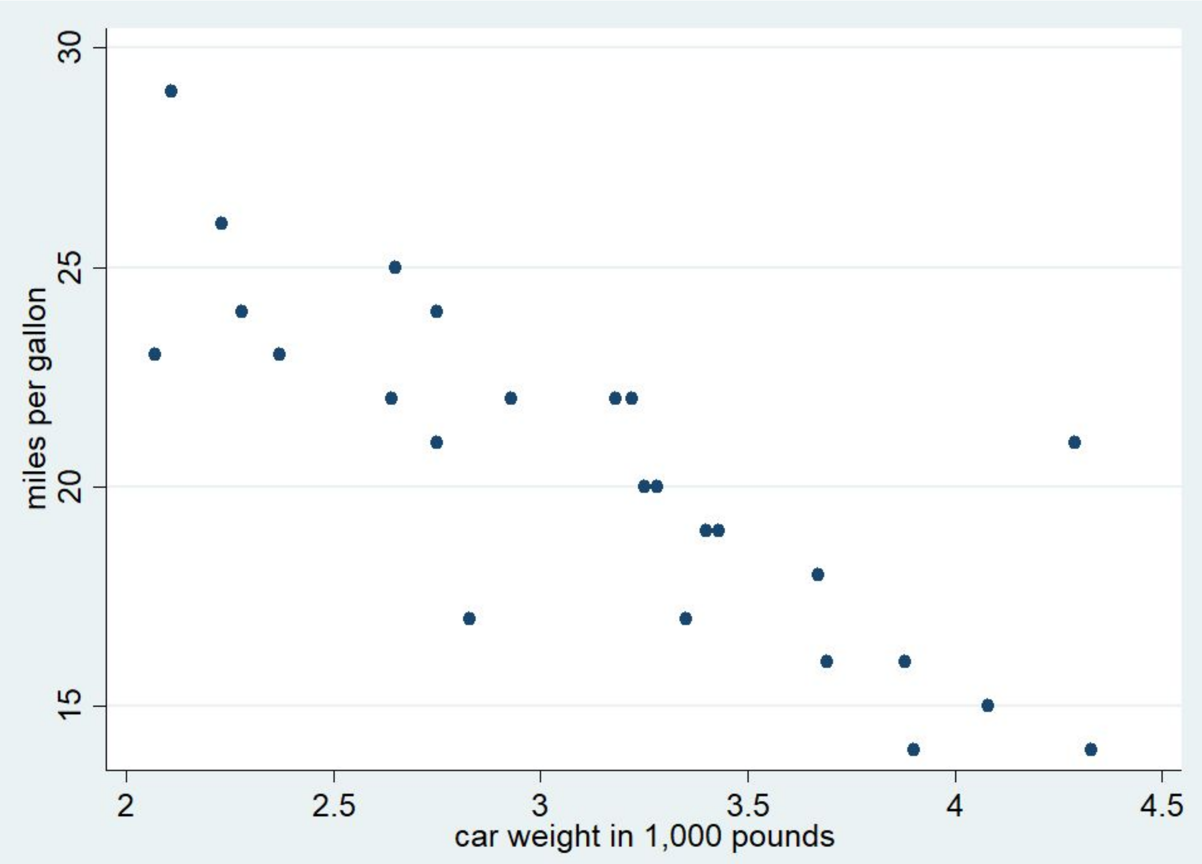
Let us view the box plot again



And now we are good.

Assumption #3: There needs to be a linear relationship between the dependent and independent variables. Let's plot a scatter plot and check

```
. graph twoway (scatter $ylist $xlist)
```




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

	mpg	weight1
mpg	1.0000	
weight1	-0.8158	1.0000

As seen above the two variables appear to be linearly related

Assumption #4: You should have independence of observations.   which you can easily check using the **Durbin-Watson statistic**, which is a simple test to run using Stata.

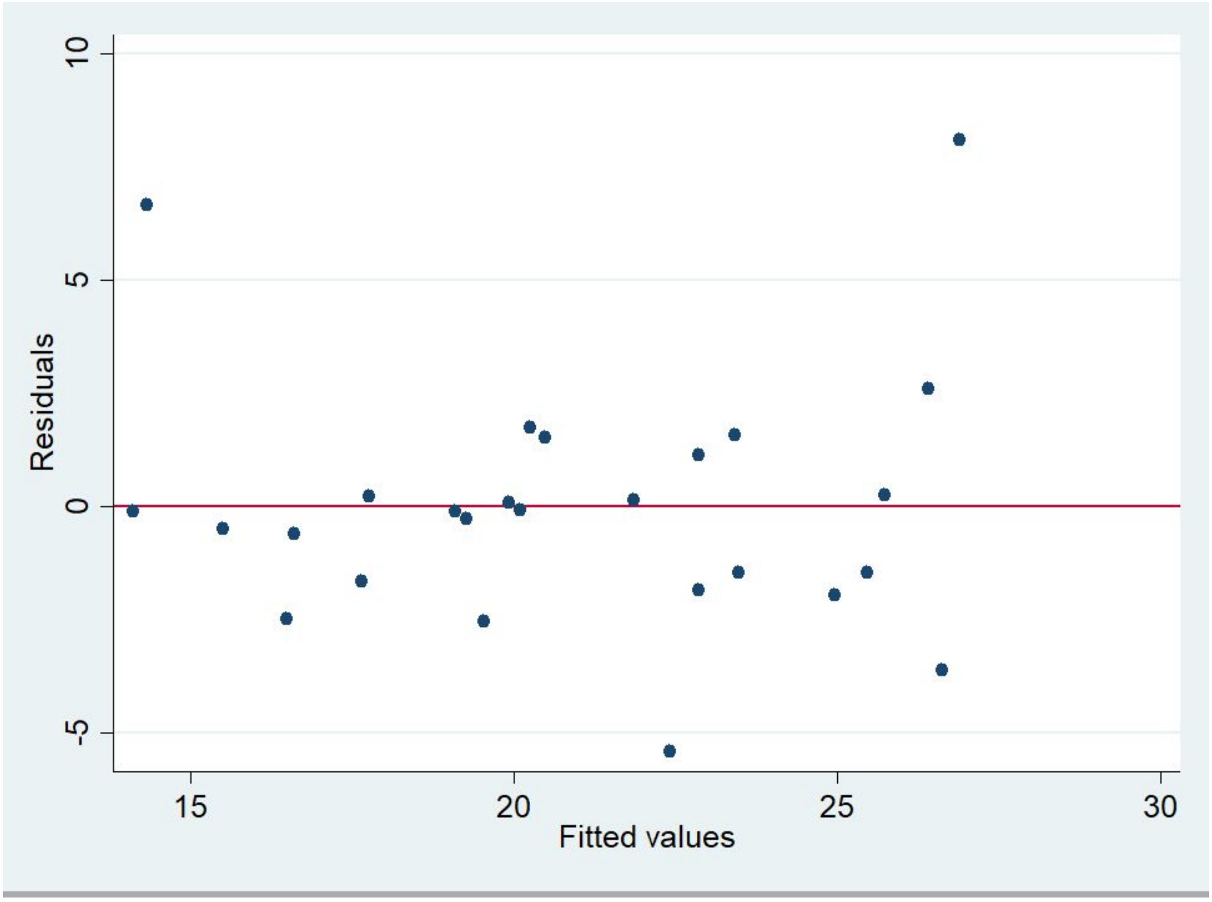
```
. //generate time variable
. gen t = _n

. 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(1, 23)        =     45.78
       Model |  239.456142         1  239.456142   Prob > F        =    0.0000
    Residual |  120.303858        23  5.23060252   R-squared       =    0.6656
-------------+----------------------------------   Adj R-squared   =    0.6511
       Total |      359.76        24       14.99   Root MSE        =    2.2871

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     weight1 |  -4.694151    .693777    -6.77   0.000    -6.129338   -3.258964
       _cons |    35.1109   2.227593    15.76   0.000     30.50277    39.71903
------------------------------------------------------------------------------

.
. dwstat

Durbin-Watson d-statistic(  2,    25) =  1.709989
```

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 #6: 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 major assumptions given for type ordinary least squares regression is the homogeneity in the case of variance of the residuals. In the case of a well-fitted model, if you plot residual values versus fitted values, you should not see any particular pattern. Now, what if the variance given by the residuals is not a constant? In this case, the **residual variance** is called **heteroscedastic**. The most commonly used way to detect heteroscedasticity is by plotting residuals versus predicted values. We should not have any side narrower than the other.

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



We can also use non-graphical commands as shown below. The first test on heteroscedasticity given by imtest is the White's test and the second one given by hettest is the Breusch-pagan test.

```
. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |       3.37      2    0.1853
            Skewness |       5.09      1    0.0241
            Kurtosis |       1.45      1    0.2278
---------------------+-----------------------------
               Total |       9.91      4    0.0419
---------------------------------------------------

. estat hettest

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

         chi2(1)      =     0.61
         Prob > chi2  =   0.4349
```

Both test the null hypothesis that the variance of the residuals are homogenous.

Assumption #7: Finally, you need to check that the residuals (errors) of the regression line are approximately normally distributed.

First we run the regression as shown below

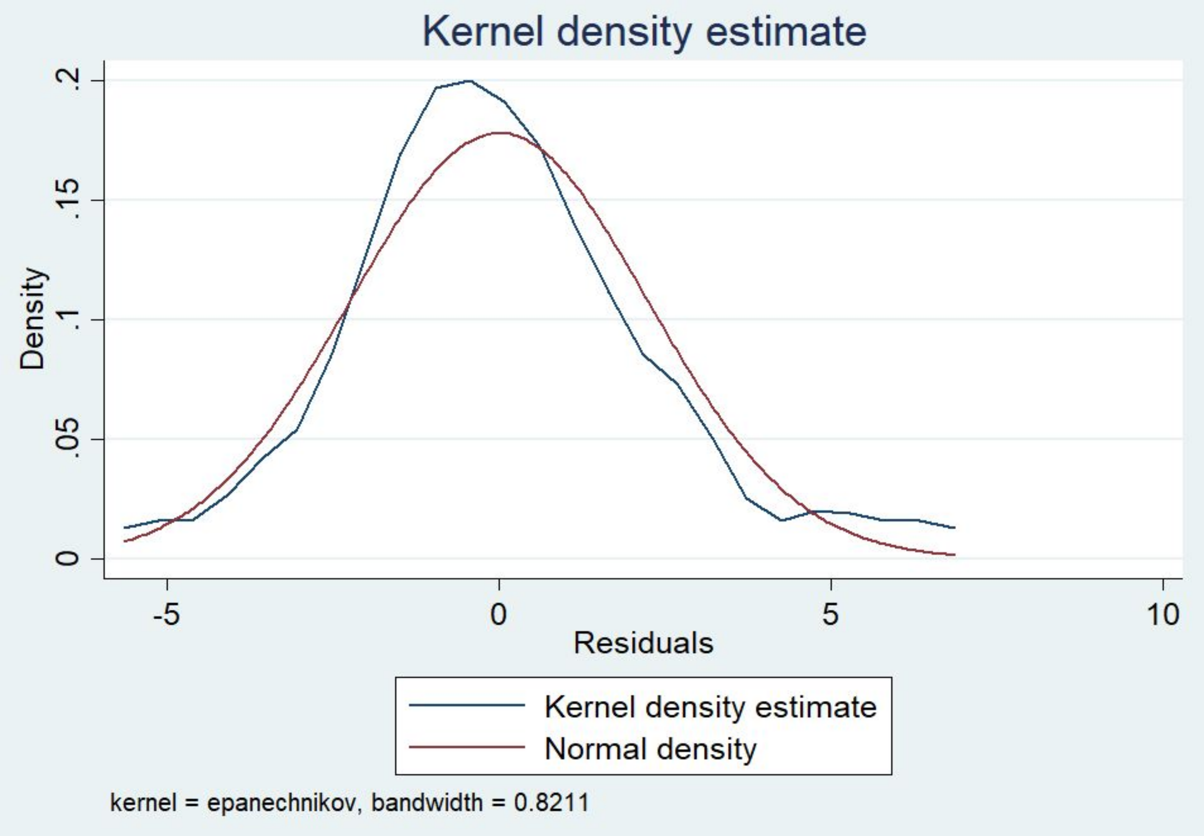
```
. reg $ylist $xlist

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(1, 23)        =     45.78
       Model |  239.456142         1  239.456142   Prob > F        =    0.0000
    Residual |  120.303858        23  5.23060252   R-squared       =    0.6656
-------------+----------------------------------   Adj R-squared   =    0.6511
       Total |      359.76        24       14.99   Root MSE        =    2.2871

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     weight1 |  -4.694151    .693777    -6.77   0.000    -6.129338   -3.258964
       _cons |    35.1109   2.227593    15.76   0.000     30.50277    39.71903
------------------------------------------------------------------------------
```

We the use the predict command to generate residuals

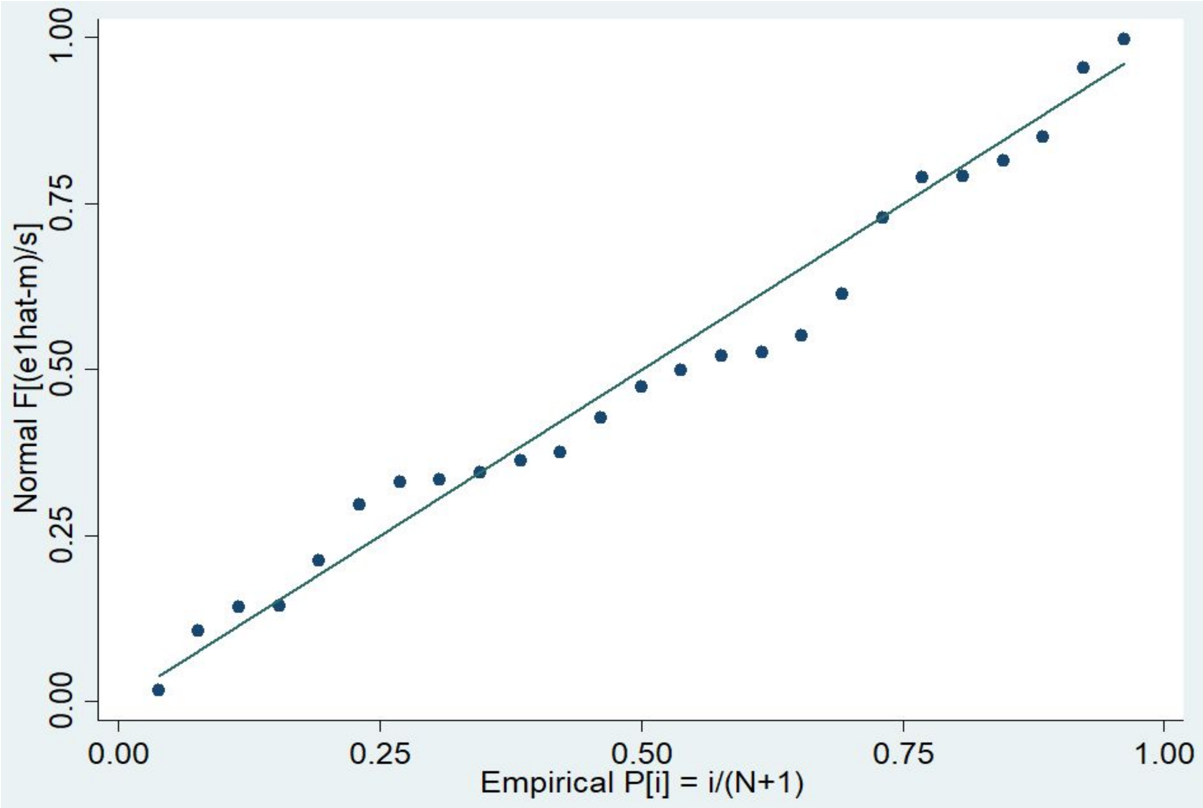
```
. predict elhat, resid
```

We finally use kdensity command to plot a kernel density plot with a normal option requesting that a normal density be overlaid on the plot.

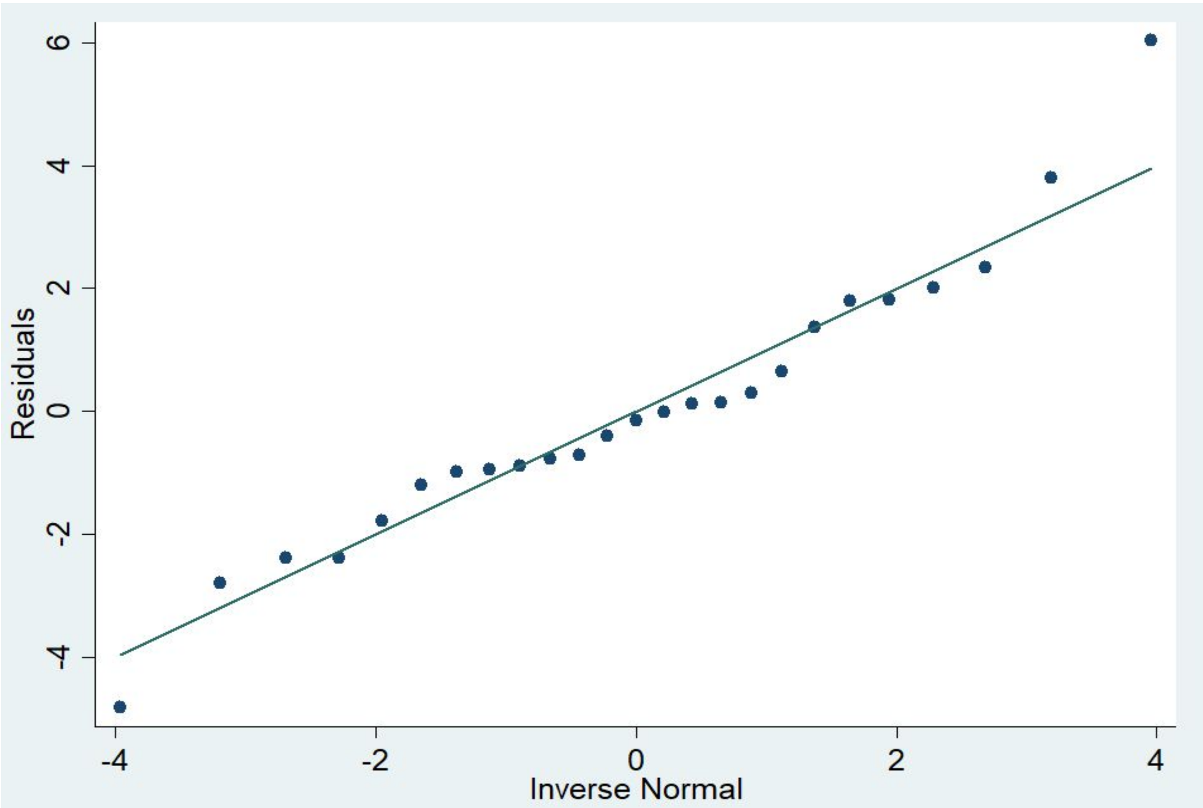
```
. kdensity e1hat, normal
(n() set to 26)
```



```
. pnorm e1hat
```



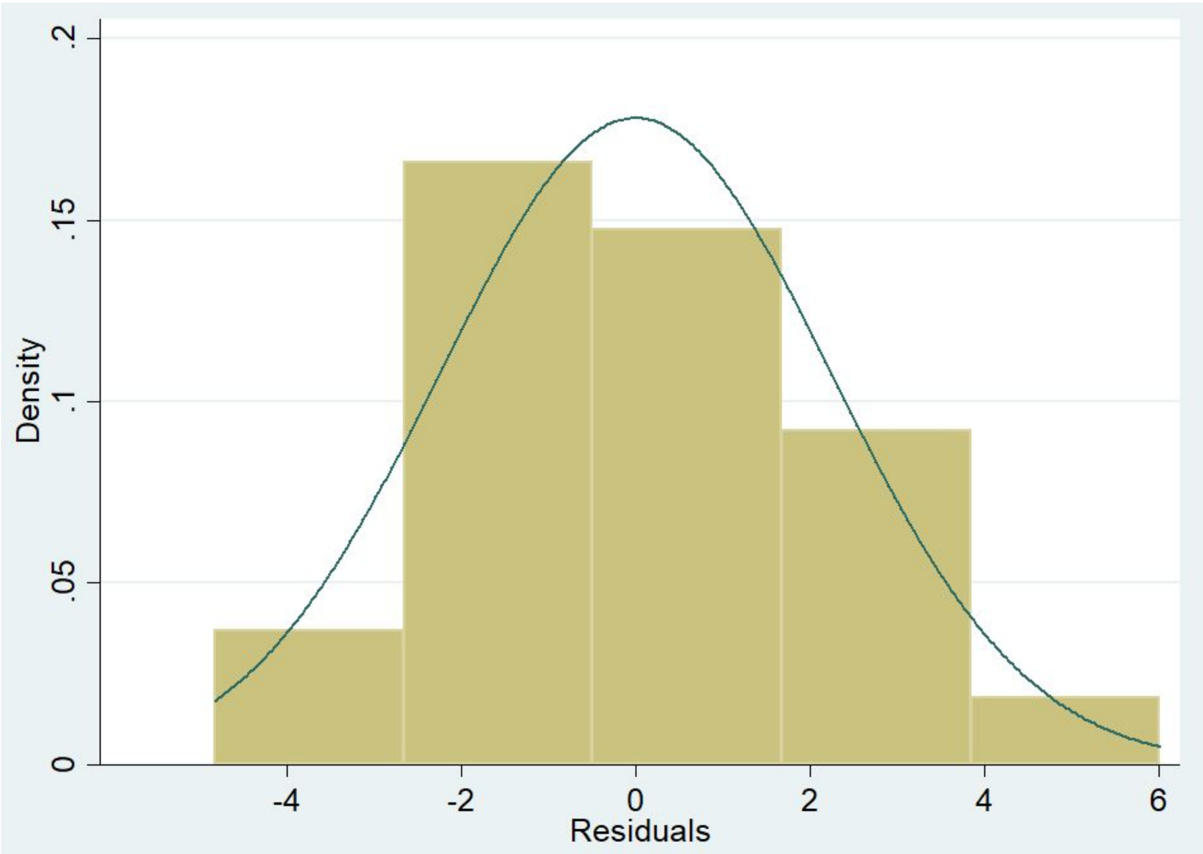

```
. qnorm e1hat
```



We can also use Shapiro-Wilk W test for normal data as shown below. The p value is based on the assumption that the distribution is normal.

```
. swilk e1hat

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
       e1hat |         25    0.96556      0.957    -0.090    0.53576
```

As seen above this assumption is satisfied as the Shapiro-Wilk W test is not significant.

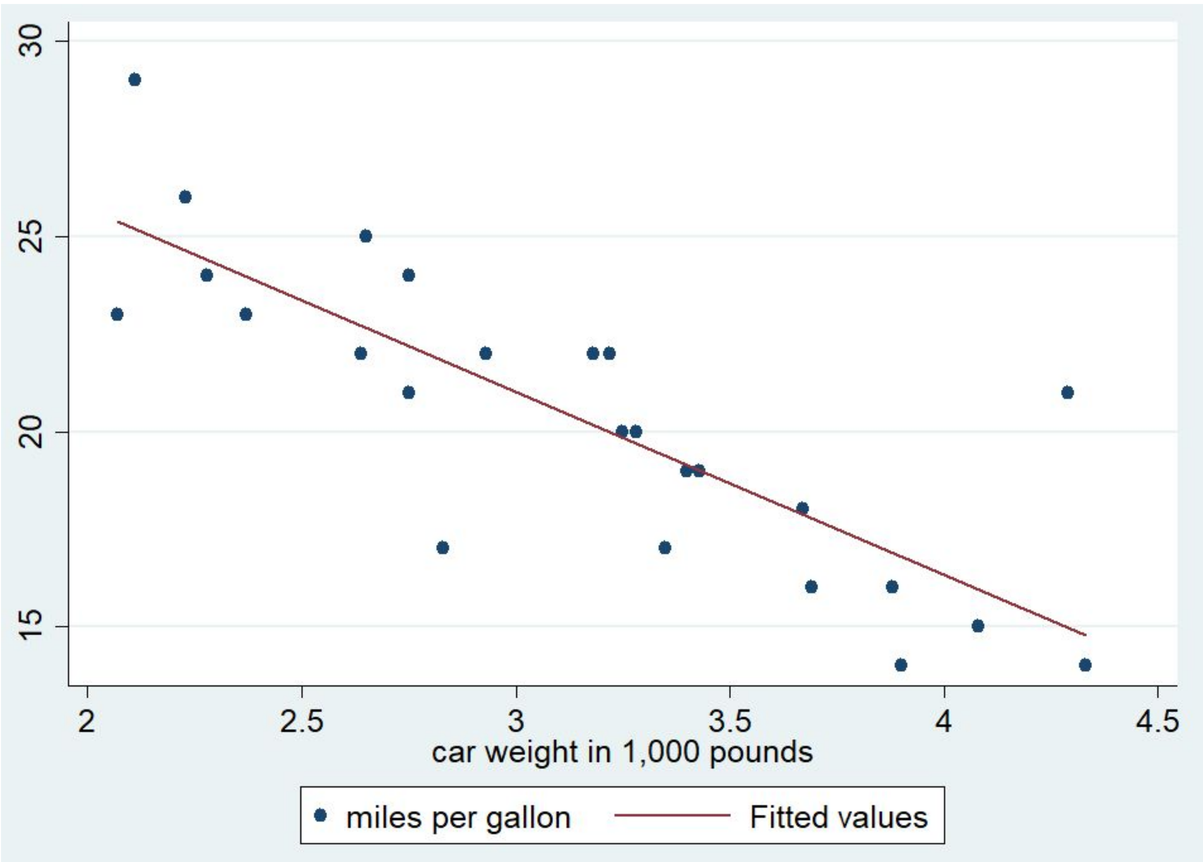


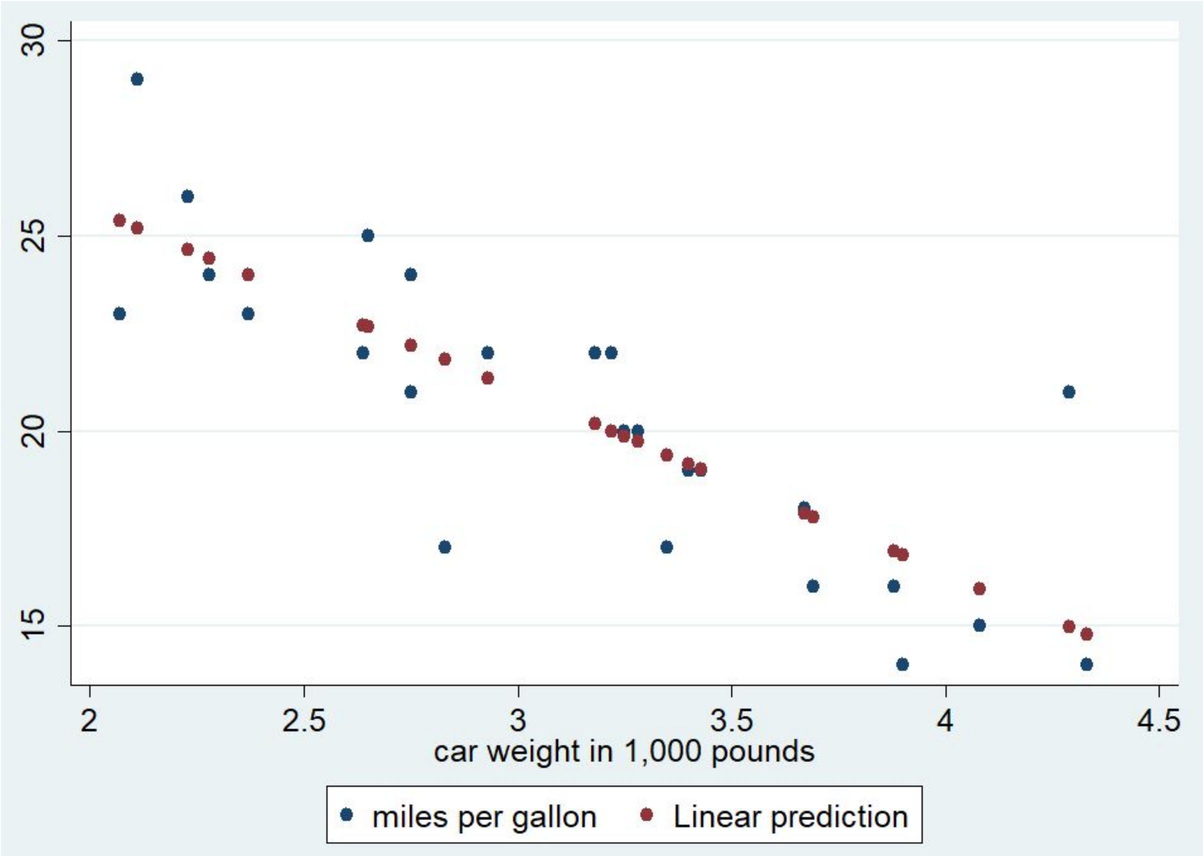
<u>Simple linear interpretation</u>

```
. reg $ylist $xlist

      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(1, 23)        =     45.78
       Model |  239.456142         1  239.456142   Prob > F        =    0.0000
    Residual |  120.303858        23  5.23060252   R-squared       =    0.6656
-------------+----------------------------------   Adj R-squared   =    0.6511
       Total |      359.76        24       14.99   Root MSE        =    2.2871

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     weight1 |  -4.694151    .693777    -6.77   0.000    -6.129338   -3.258964
       _cons |    35.1109   2.227593    15.76   0.000     30.50277    39.71903
------------------------------------------------------------------------------
```

A linear regression established that weight of a car could statistically significantly predict mpg (miles covered by the car per gallon), $F(1, 23) = 45.78$, $p < .05$ and the weight of a car accounted for 66.56% of the explained variability in miles covered by the car per gallon. The regression equation was: predicted miles covered by the car per gallon = 35.1109 - 4.694 x (weight of the car).

Plotting a regression line




30
25
20
15
2
2.5
3
3.5
4
4.5
car weight in 1,000 pounds
miles per gallon
Linear prediction