

## Session 5

### Model diagnostics in STATA 8

	<i>page</i>
The dataset. Cherry tree data	5-3
Checking model formula	5-4
Other omitted variables	5-6
Distributional assumptions	5-8
Independence	5-9
Normality	5-10
Aberrant and influential points	5-11
Leverages	5-12
What do we do about leverages and outliers?	5-14
Box-Cox transformation	5-17

## ***Session 5: Model diagnostics in STATA 8***

How do we know our model is correct?

Assumptions might be violated:

- Normality
- Linearity
- Constant variance
- Model formula
- Choice of transformation of response variate
- Aberrant data points
- Influential data points – points which have too much influence on the regression parameters.

Model diagnostics provide insight into all of these features, which are interrelated.

For example, one aberrant data point can cause the need for a more complex model, and can move the residual distribution away from Normality.

Model diagnostics are usually graphical – it is left to the data analyst to interpret the plot.

The basic building blocks of diagnostics:

- Fitted values  $\hat{y}_i$
- Residuals  $r_i = y_i - \hat{y}_i$
- Leverages - the influence of  $y_i$  on  $\hat{y}_i$
- Deletion quantities— effect of omitting a point from the fit.
- Quantile plots- testing distributional assumptions.

### **The dataset. Cherry tree data**

31 Black cherry trees from the Allegheny national forest.. Data on:

V : Volume of useable wood (Cubic feet)

D : Diameter of tree 4.5 feet above the ground

H: Height of tree

Aim is to predict V from easily measured D and H.

Linear regression, but check model assumptions.

Fit a Normal linear regression model to the tree data. Response V Explanatory D and H

D and H highly significant with positive coefficients.

Can use predict to get many model quantities after model fit:

```
predict newvariable, quantity
```

quantity is

```
residual
```

```
xb
```

```
fitted values
```

```
lev
```

```
leverages
```

## Checking model formula

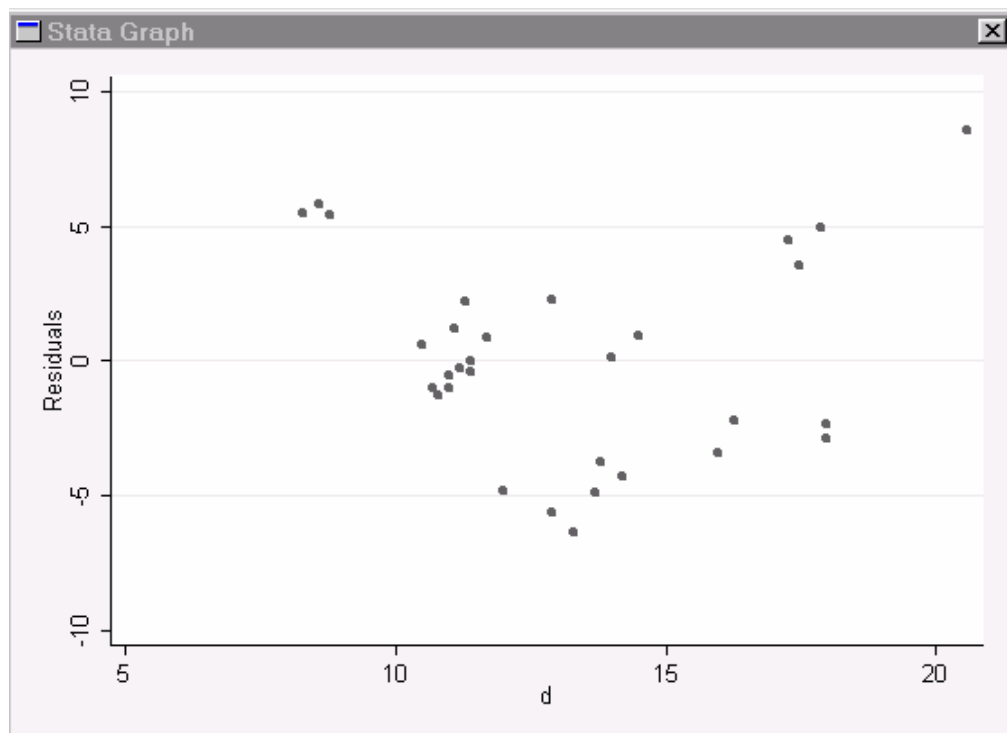
Do we need extra terms in  $D^2$  or  $H^2$ ?

We use residual plots – not all available from graphical menu. Sometimes you need to create plots yourself!

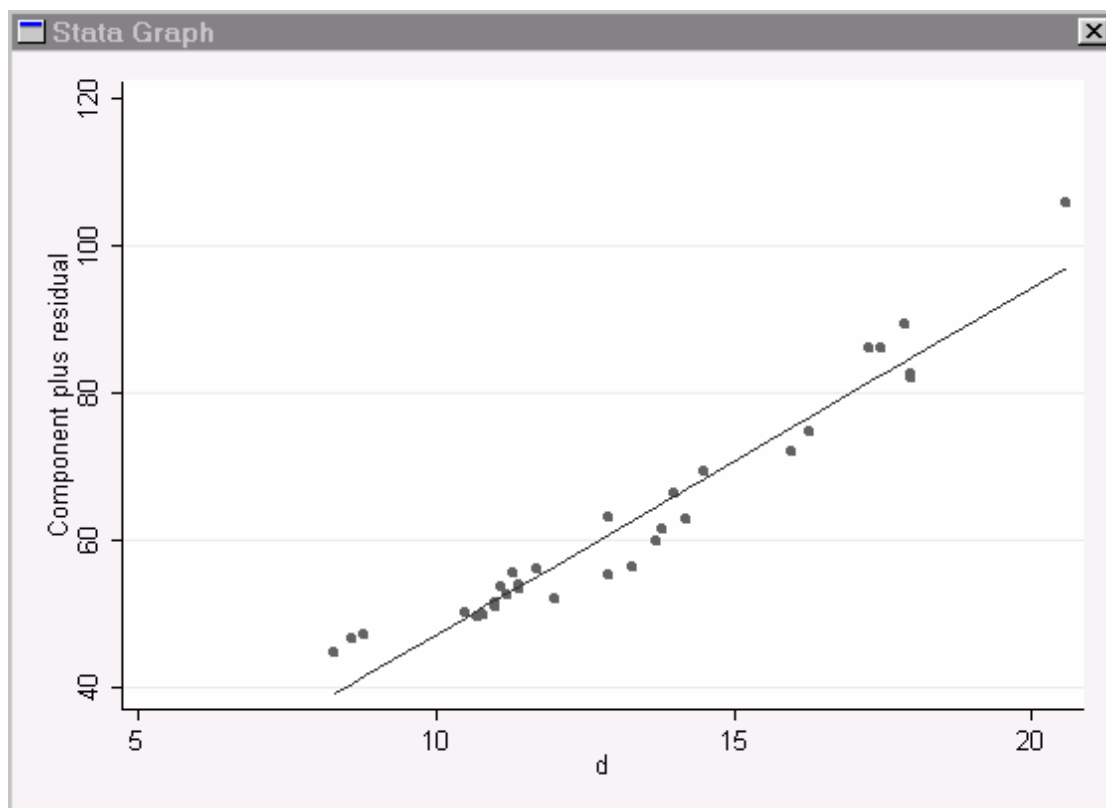
Access in two ways to menu diagnostics – via **graphics** or via **statistics>linear regression**

1. Plot residuals against any included explanatory variables.

```
rvpplot d
```



### Component plus residual plots



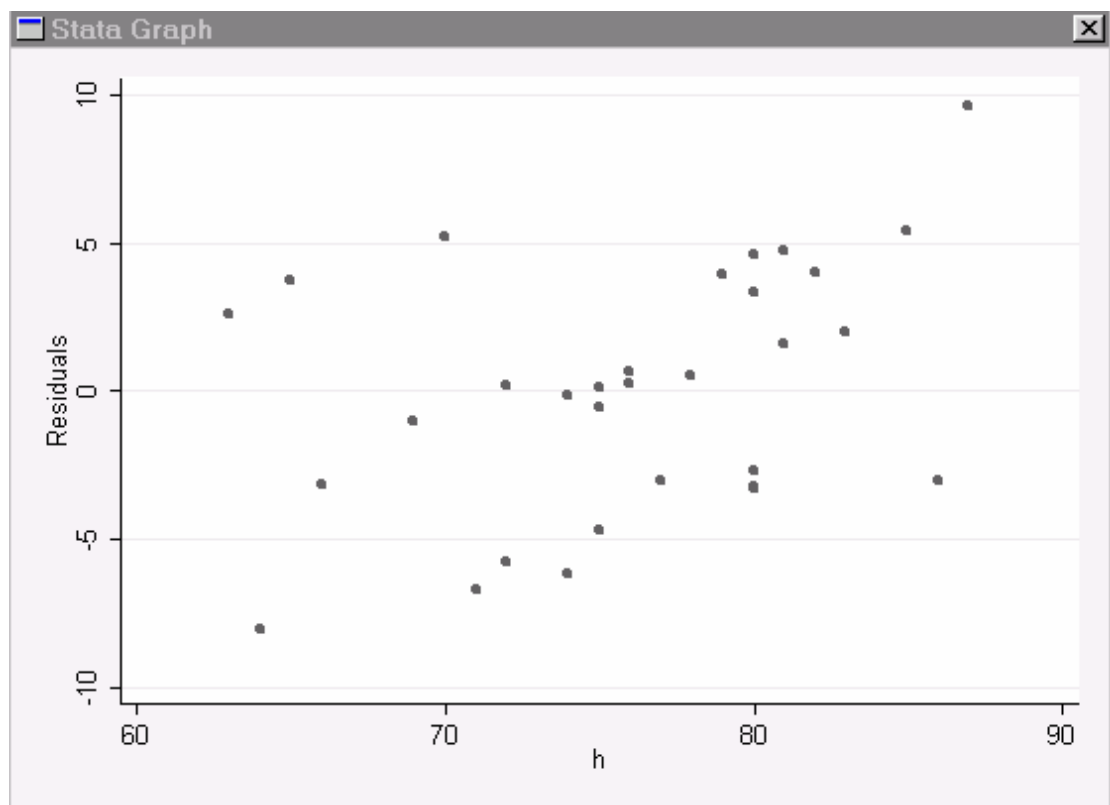
Some evidence of curvature – perhaps term in  $D^2$  needed?

### Other omitted variables

Can produce scatter plots of residuals of current model against omitted variable.

Eg Regression of V on D alone – is H needed?

```
regress V D
predict res, residuals
twoway res h
```

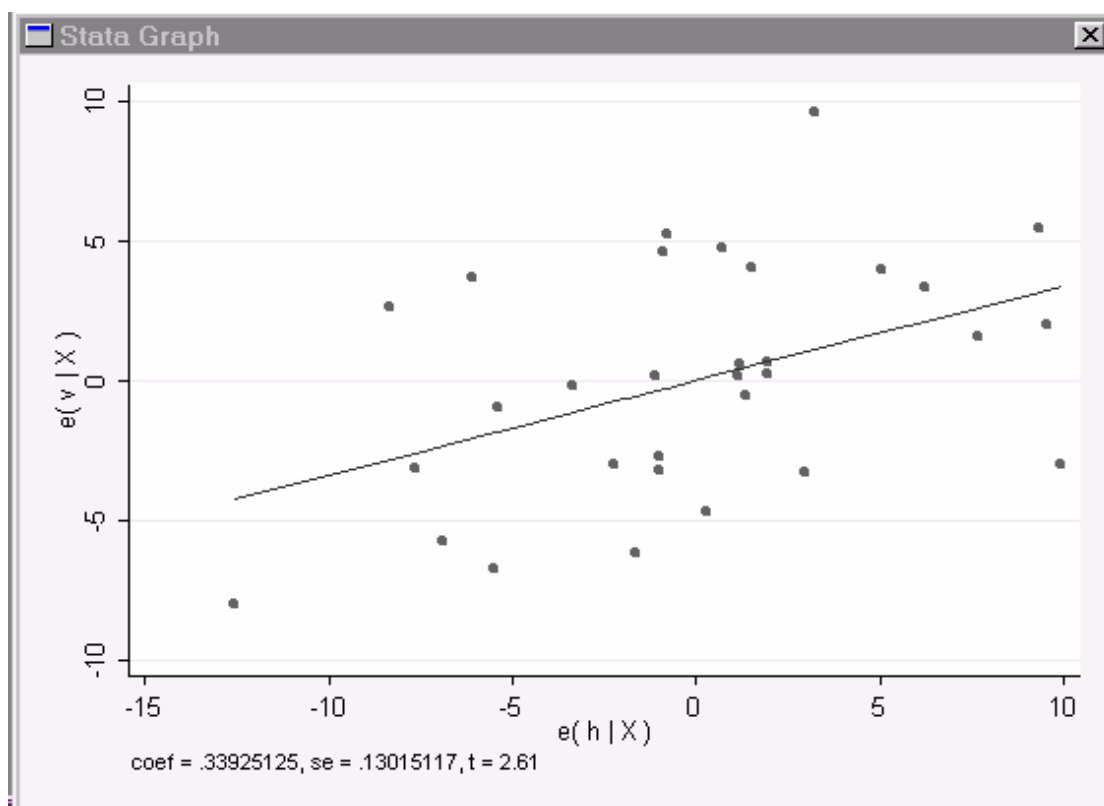


strong linear trend observed.

Also **added variable plots**. Plot residuals against residuals from a model using the new variable as response with the same set of predictors.

So, residuals of V against H are plotted with residuals of D against H. Slope will be regression coefficient in full model. Line goes thru origin.

```
avplot h
```



Need to include H in model.

# Distributional assumptions

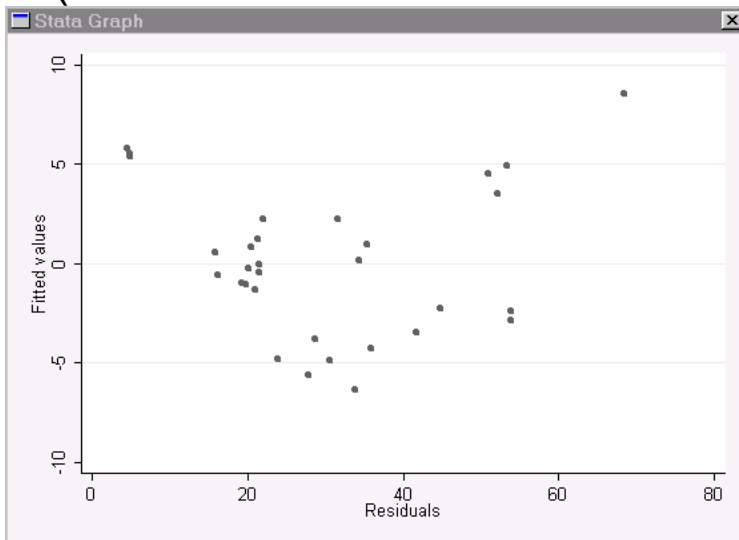
Is the distribution of the residuals Normal, and is there constant variance?

a) Constant variance

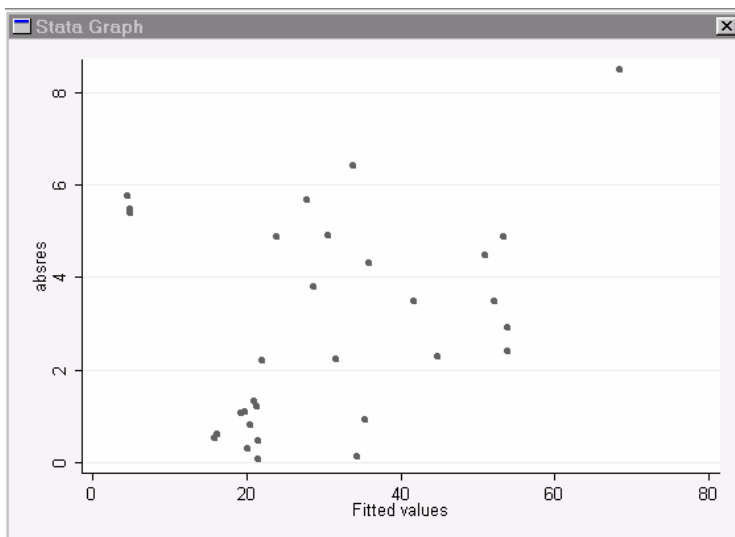
Plot residuals against fitted values and look at spread.

`rvfplot`

(or menu choice **residual vs fitted plot**)



What is the mistake in this graph?



Or use  
absolute  
residuals  $|r_i|$

No real  
evidence in  
either plot of  
non-constant  
variance

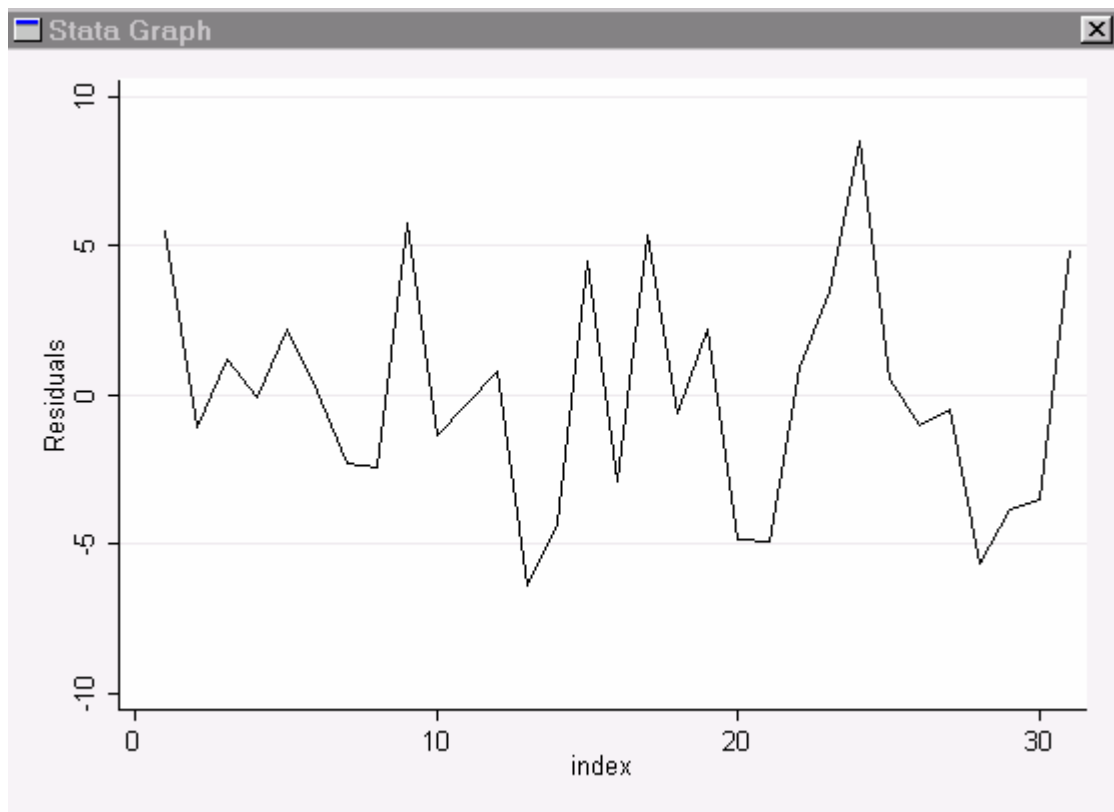
## Independence

Plot residuals against order of data. (index plot)  
Assumes that data points are listed in order of collection, and that dependence might be introduced by this route.

Other sources – clustered data, interviewer effects, time effects, learning effects.

```
generate index=sum(1)
twoway line res index
```

Look for clusters of positive or negative residuals and then relate these clusters to what you know about the data. Will lead to more complex model which incorporates this extra knowledge.



## Normality

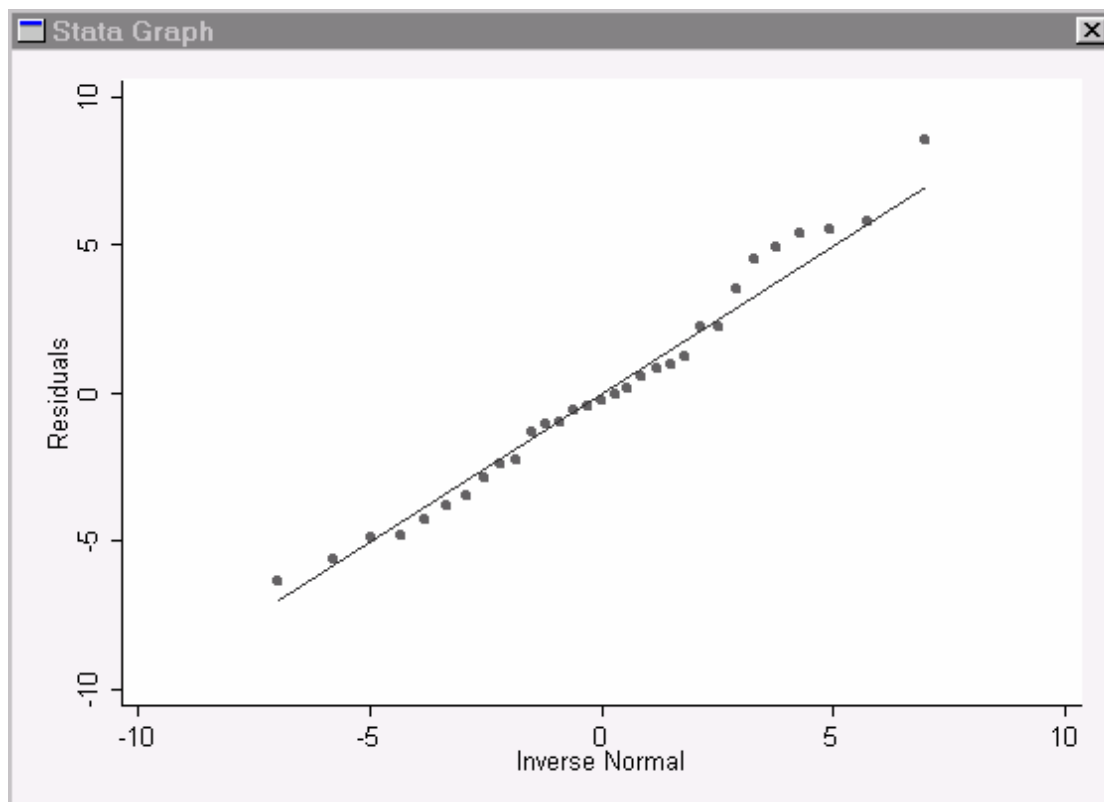
Plot the ordered residuals against a set of typical residuals from a normal distribution. These are obtained using Normal quantiles, so this plot is known as a quantile –quantile plot (or Q-Q plot).

A straight line gives Normality. The points on the graph show your data- the line is the perfect answer.

**graphics>distributional graphs>normal quantile**

```
qnorm res
```

We use the residuals `res` as the variable in this command. This plot looks good.

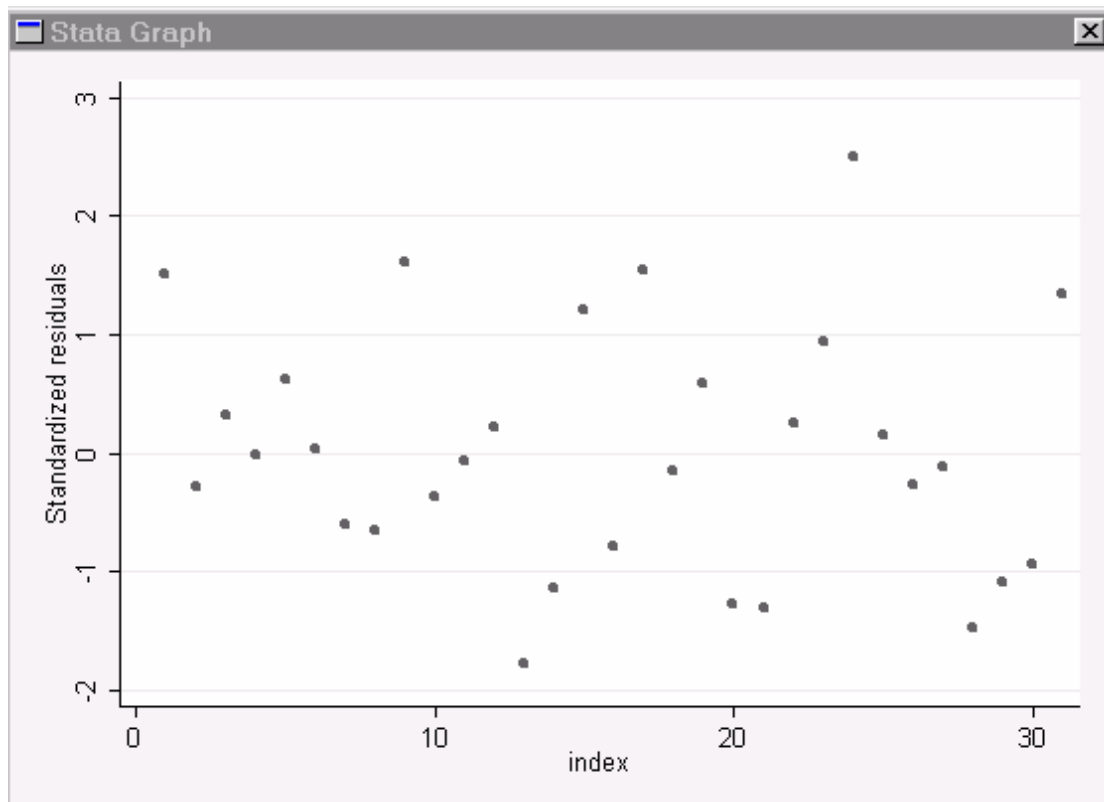


## Aberrant and influential points

Identify outliers by examining points with large standardised residuals. Plot against index vector

Look for large standardised residuals greater than two in absolute value, taking into account that one in twenty residuals will be above 2 or less than  $-2$ .

```
predict sres, rstandard  
twoway scatter sres index
```



Point 24 has a residual of 2.5, but this is the only large point in the dataset. Ignore.

Many other outlier detection techniques.

## Leverages

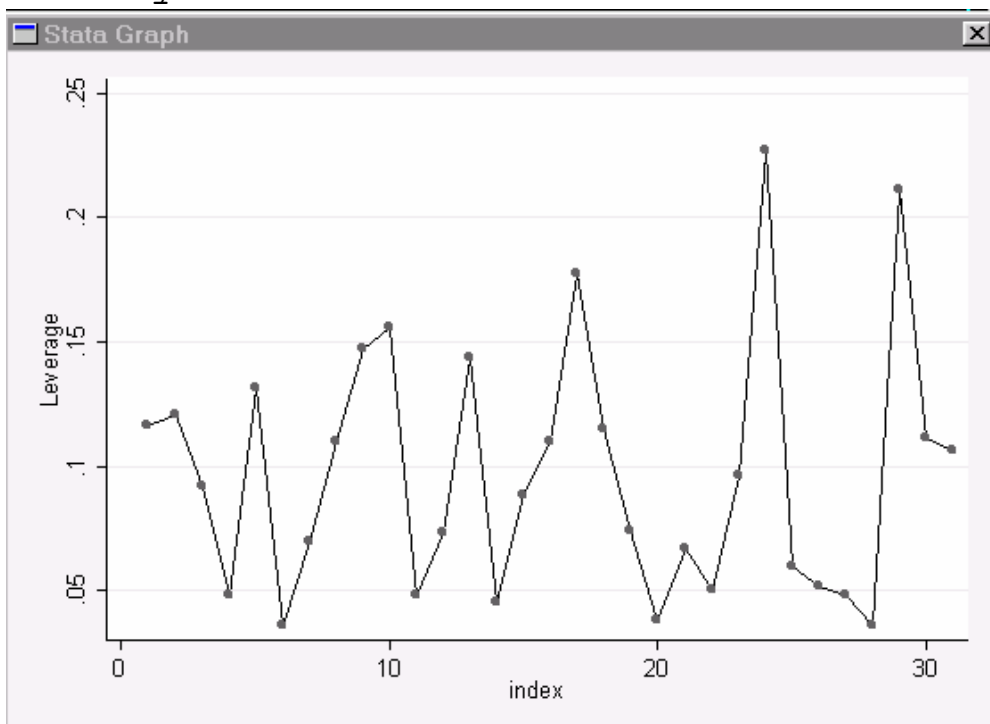
Leverages  $h_{ii}$  are the contribution of the  $i$ th point to the  $i$ th fitted value. Ideally, we would want each point in the regression to contribute equally to each fitted value.

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n$$

Large values of  $h_{ii}$  can be taken to be twice the average value  $2p/n$  where  $p$  is the number of estimated parameters in the model. For our current model,  $p=3$  and so we look for leverages greater than  $6/31 = 0.194$

We plot leverages against the case order (index plot).

```
predict lev, lev  
twoway scatter lev index
```

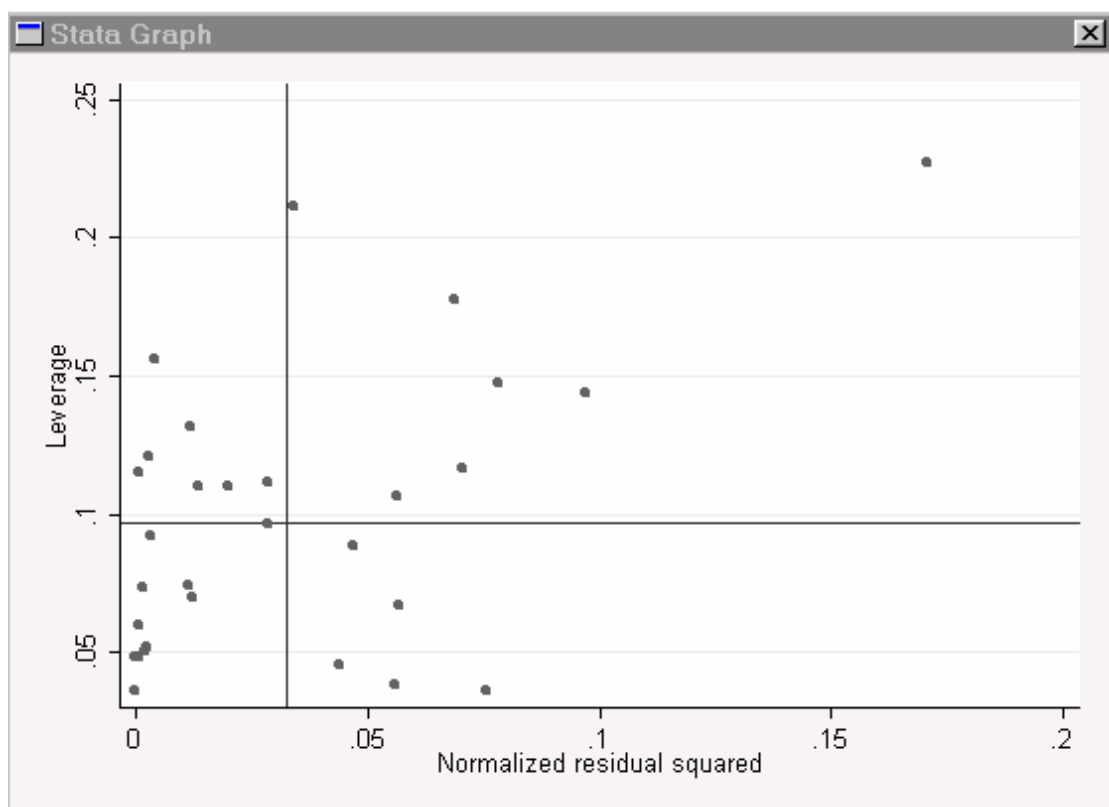


Two points have high leverage – point 24 and point 29. These trees are more influential than other points in determining the regression coefficients.

We can also plot leverages against the squared residuals. and look for points with high leverage and high residual.

### regression diagnostics> leverage-versus-squared-residual

```
lvr2plot
```



Point 24 is seen to have a high residual and a high leverage. Point 29 has a smaller residual and high leverage.

## What do we do about leverages and outliers?

Look at the effect of deleting the point. Two procedures can be followed. We can look at the effect on the parameter estimates, or look at the effect on the fitted values.

a) effect on parameter estimates.

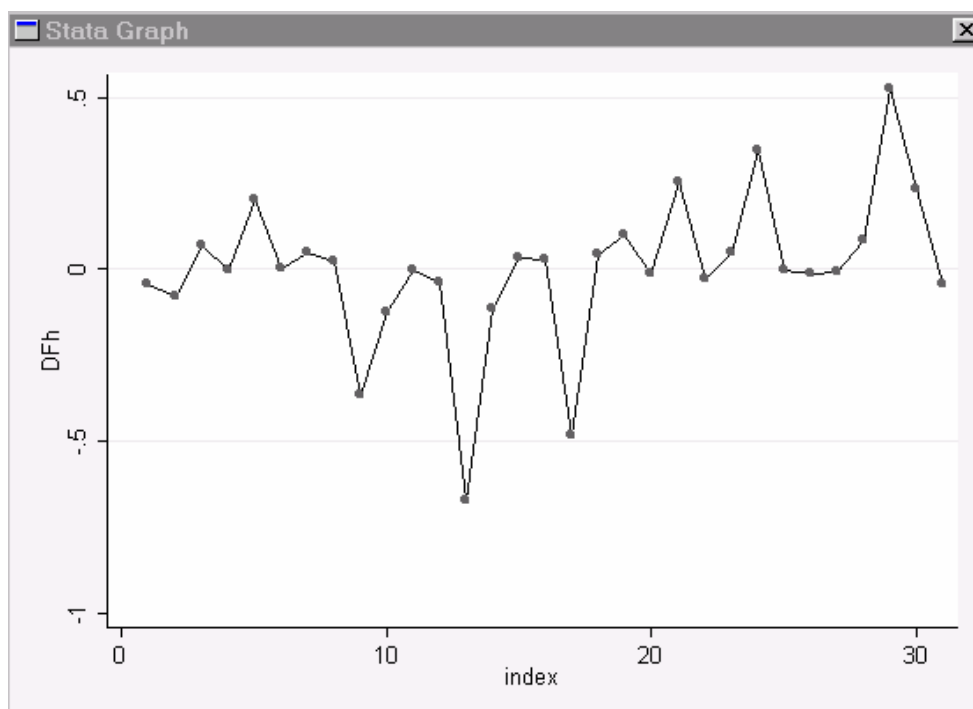
We use the `dfbeta` command , or **regression diagnostics>Dfbeta**

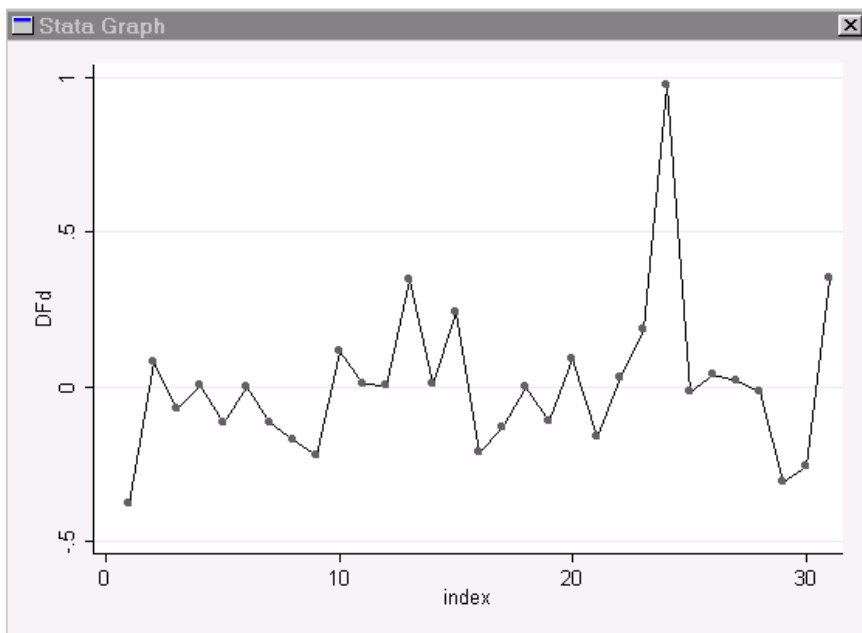
We need to specify the estimate of interest.

```
dfbeta h
```

The command creates a new variable DFh, and we can produce an index plot of this variable.

```
twoway connected DFh index
```



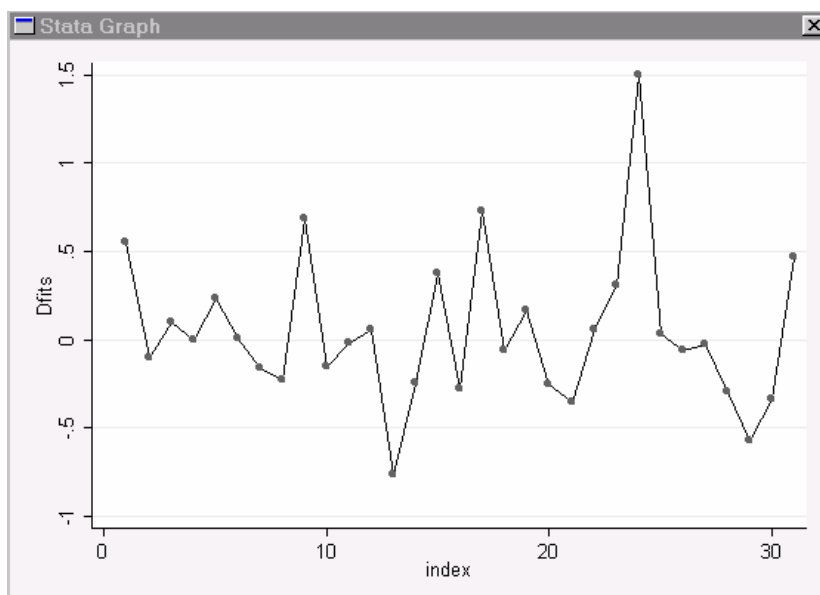


Point 24 has a large influence on the estimate for D, changing it by 1 unit.

b) effect on fitted values.

Use `predict` to get **dfits** for each observation, and produce an index plot.

```
predict dfi, dfits
twoway connected dfi index
```



Again, point 24 is identified.

So, what have we learnt? We have found one or possibly two influential points, and there is a suggestion that we need to add a term in  $D$  squared. If we add this term, then we will need to repeat these diagnostic tests again - the process is iterative.

However, before we do this, we have not considered transformations of the  $Y$ -variable or the explanatory variables. We can investigate this through the Box Cox procedure.

## Box –Cox transformation

A family of power transformations for the response variable.

$$T(y) = \begin{cases} \frac{y^\theta - 1}{\theta} & \theta \neq 0 \\ \log(y) & \theta = 0 \end{cases}$$

We assume that there is some value of  $\theta$  which transforms to Normality, gives homogeneous variance, and simple model structure.

We find  $\theta$  by maximum likelihood. We are interested in “sensible” values of  $\theta$  –

$\theta=2$	square transformation
$\theta=1$	(no transformation)
$\theta=1/2$	Square root transformation
$\theta=0$	log transformation
$\theta=-1$	reciprocal transformation – etc

Use Box Cox regression to do this.

### **Statistics>Linear regression>Box Cox regression**

Specify transformation on LHS only (Response variate)

```
boxcox v d h, model(lhsonly)
```

## Relevant part of output

v	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/theta	.3065849	.0929172	3.30	0.001	.1244706 .4886992

Test	Restricted	LR statistic	P-Value
H0:	log likelihood	chi2	Prob > chi2
theta = -1	-100.54818	67.42	0.000
theta = 0	-71.462357	9.24	0.002
theta = 1	-84.454985	35.23	0.000

The estimate of theta is 0.306. However, the tests below indicate that this value of theta is not consistent with a sensible value of -1, 0 or 1.

However this value is consistent with  $\theta = 1/3$ . This is sensible from a dimension point of view – Volume is a cubic measure and height and diameter are linear.

Another possibility is to consider transformations of both the response and explanatory variables. We choose the option '**both sides with the same parameter**' and repeat.

```

-----
              v |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
    /lambda |   -.1113739   .1372059   -0.81   0.417   - .3802925   .1575447
-----

```

```

-----
Test              Restricted      LR statistic      P-Value
H0:              log likelihood      chi2      Prob > chi2
-----
lambda = -1      -81.063148          30.57          0.000
lambda =  0      -66.099057           0.65          0.422
lambda =  1      -84.454985          37.36          0.000
-----

```

The procedure now estimates  $\theta = -0.11$ .

This is very close to zero, and the likelihood ratio tests in the later part of the output indicates that this value of  $\theta$  is consistent with  $\theta = 0$ .

So we have two possibilities to investigate.

1. Take a cube root transformation of  $V$  and assess the effect of  $D$  and  $H$
2. Take logs of all variables, and consider modelling  $\log V$  in terms of  $\log D$  and  $\log H$ .

Both of these are sensible ways of proceeding. Diagnostic plots can be carried out as before, but the Box Cox procedure has suggested something useful.