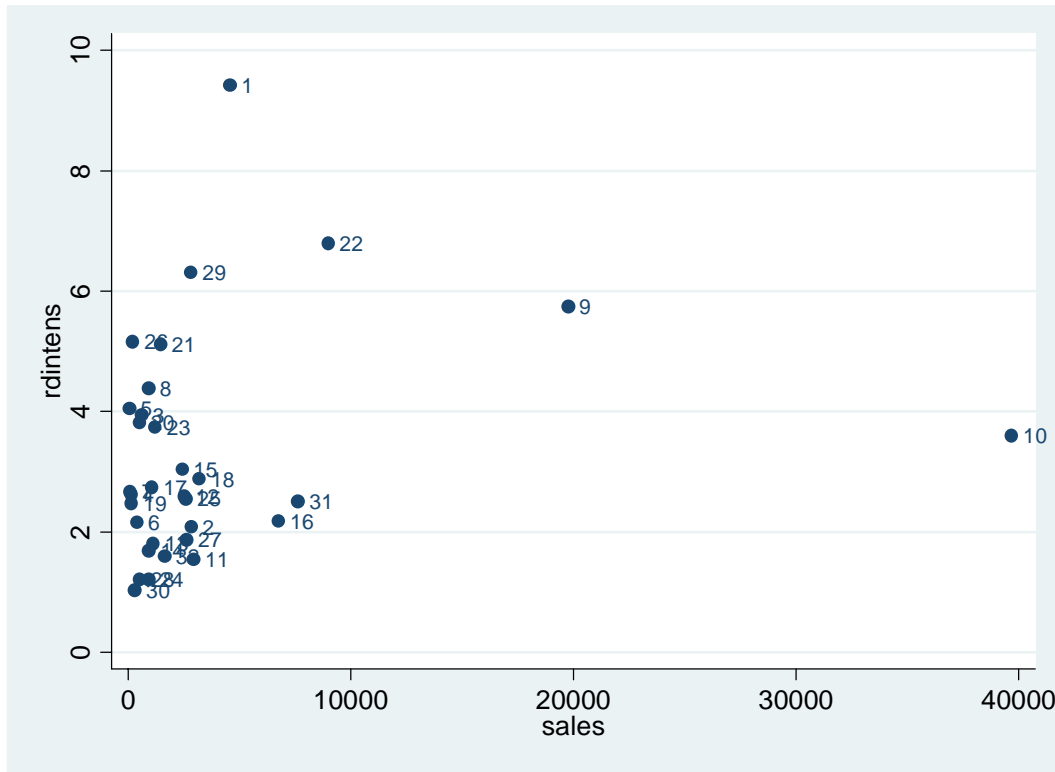


The data set rd (on the course website taken from *Wooldridge: Introductory Econometrics*) has information on the r&d levels and sales of a sample of 32 US firms.

A graph of r&d against sales suggests the presence of an “outlier” (observation number 10) which lies a long way from the main mass of data

twoway (scatter rdintens sales, mlabel(n))



A regression of r&d on sales for the whole sample gives

```
. reg rdintens sales
```

Source	SS	df	MS			
Model	5.07296644	1	5.07296644	Number of obs =	32	
Residual	103.804422	30	3.46014739	F( 1, 30) =	1.47	
Total	108.877388	31	3.51217381	Prob > F =	0.2354	
				R-squared =	0.0466	
				Adj R-squared =	0.0148	
				Root MSE =	1.8601	

rdintens	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sales	.0000533	.000044	1.21	0.235	-.0000366	.0001432
_cons	3.063199	.3688881	8.30	0.000	2.30983	3.816569

and a regression omitting the “outlier” gives

```
. reg rdintens sales
```

Source	SS	df	MS			
Model	15.0989908	1	15.0989908	Number of obs =	31	
Residual	93.6629841	29	3.22975807	F( 1, 29) =	4.67	
Total	108.761975	30	3.62539916	Prob > F =	0.0390	
				R-squared =	0.1388	
				Adj R-squared =	0.1091	
				Root MSE =	1.7972	

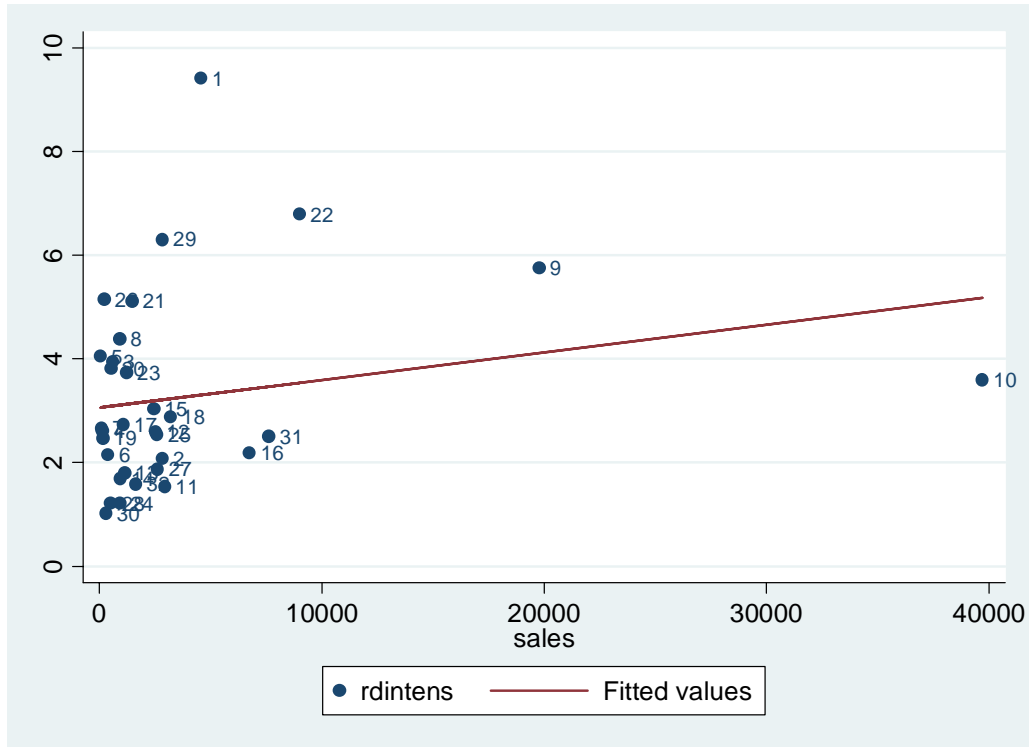
  

rdintens	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sales	.0001824	.0000844	2.16	0.039	9.87e-06	.000355
_cons	2.773458	.3921142	7.07	0.000	1.971495	3.575422

so the estimated coefficient is around 3 times as large in the absence of the observation.

```
predict yhat /* this gets fitted values and saves under variable name “yhat” */
```

However inspection of the residuals from the original regression does not reveal this  
 twoway (scatter rdintens sales, mlabel(n)) (line yhat sales)

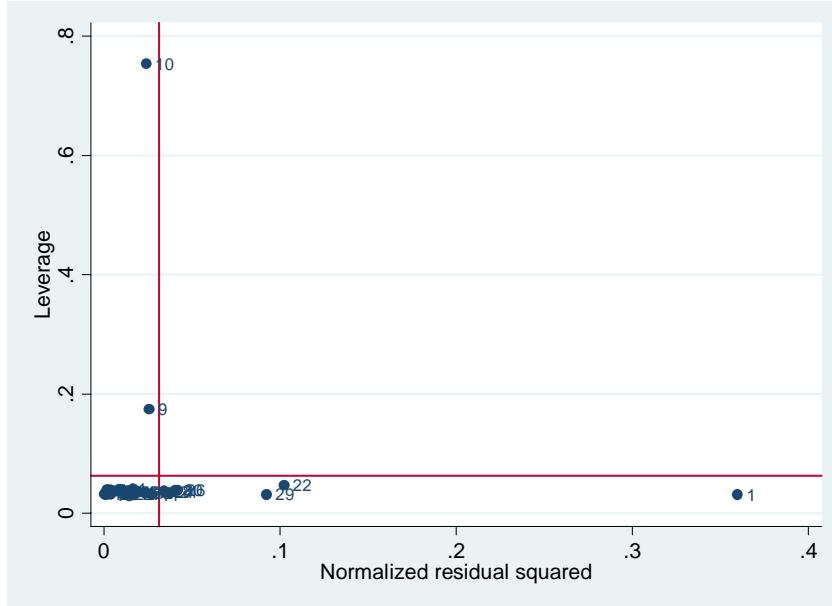


The observation with the highest residual is observation 1 not 10. Observation 10 therefore has high “leverage”. Since it lies away from the main mass of X observations on sales it exerts lots of weight in the OLS prediction, in effect forcing the regression line to pass close to it.

Moral: care has to be taken when deciding which variables to include in the data. Use of tests such as standardised residuals or DFITS can help detect observations that are worthy of further investigation, but do not in themselves say that these observations should be dropped.

The command "lvr2plot" plots the estimated leverage value against the standardised residuals squared

lvr2plot, mlabel(n)



The DFITS command issued after the regression command

predict, dfit, dfit

will estimate the DFITS value for each observation in the data set

.list rdintens sales dfit

```
+-----+
rdintens  sales    dfit
-----+-----+
1.   9.42  4570.2  .7481167
2.   2.08   2830  -.1110674
3.   3.94   596.8  .0895145
4.   2.62   133.6 -.0488057
5.   4.05    42   .1076869
-----+-----+
6.   2.15    390  -.100122
7.   2.66    93.9 -.044334
8.   4.39   907.9 .133946
9.   5.75  19773  .4431768
10.   3.6  39709  -3.10056
```

In both cases observation 10 looks to be worth inspecting in detail.