

Computer Exercise 2. Regression Diagnostics - Answers

```
. u cex2
```

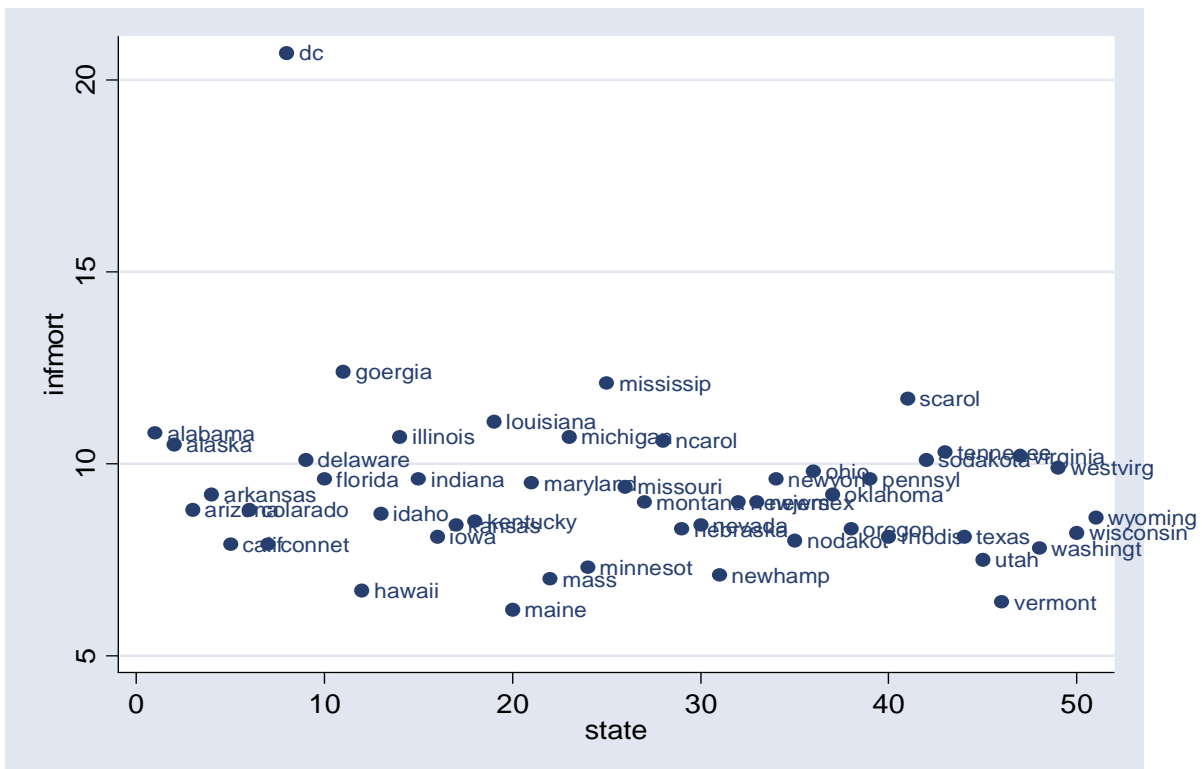
Good idea to summarise the data to check that means, minima and maxima etc look sensible. This can be done with the following stata command

```
summ
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|----------|-----|----------|-----------|----------|----------|
| year | 51 | 1990 | 0 | 1990 | 1990 |
| infmort | 51 | 9.284314 | 2.15057 | 6.2 | 20.7 |
| afdcnum | 51 | 234.4118 | 335.1369 | 16 | 2023 |
| pop | 51 | 4876.647 | 5439.203 | 454 | 29760 |
| pcapinc | 51 | 17836.25 | 2967.524 | 12700 | 25528 |
| docspop | 51 | 205.1569 | 76.06034 | 125 | 615 |
| afdcper | 51 | 4.255785 | 1.462805 | 1.688183 | 8.896211 |
| dc | 51 | .0196078 | .140028 | 0 | 1 |
| state | 51 | 26 | 14.86607 | 1 | 51 |
| ldocs | 51 | 5.278746 | .2800004 | 4.828314 | 6.421622 |
| lpcap | 51 | 9.775915 | .1621253 | 9.449357 | 10.14753 |
| lpop | 51 | 7.995111 | 1.032828 | 6.118097 | 10.30092 |

In small data sets good practice to also plot data to look for evidence of outliers. Above shows that DC appears to have much higher infant mortality rate. Remember that this does is only constitutes cause for concern if the observation has high leverage or the (normalised) residual for this observation from a regression is large.

```
. twoway (scatter infmort state, mlabel(state)), ytitle(infmort)
ylabel(, labels) xtitle(state) label(,labels)
```



The lin-log regression of the level of infant mortality rate on the logs of doctors per head, population and per income per head produces the following output:

```
. reg infmort lpop ldocs lpcap
```

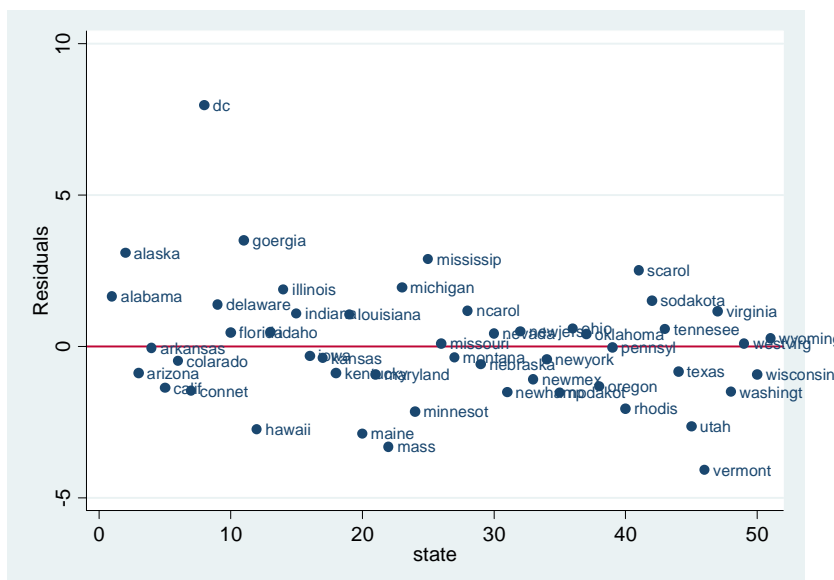
| Source | SS | df | MS | Number of obs = 51 | | |
|----------|------------|-----------|------------|--------------------|----------------------|----------|
| Model | 32.162998 | 3 | 10.7209993 | F(3, 47) | = | 2.53 |
| Residual | 199.084471 | 47 | 4.23583981 | Prob > F | = | 0.0684 |
| ----- | | | | R-squared | = | 0.1391 |
| Total | 231.247469 | 50 | 4.62494938 | Adj R-squared | = | 0.0841 |
| ----- | | | | Root MSE | = | 2.0581 |
| ----- | | | | | | |
| infmort | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
| lpop | -.0878224 | .28725 | -0.306 | 0.761 | -.6656948 | .4900499 |
| ldocs | 4.153261 | 1.512659 | 2.746 | 0.009 | 1.110185 | 7.196338 |
| lpcap | -4.684662 | 2.604124 | -1.799 | 0.078 | -9.923484 | .5541606 |
| _cons | 33.85931 | 20.42785 | 1.658 | 0.104 | -7.236219 | 74.95484 |

The RHS variables are in logs, the dependent variable in levels so coefficients are semi-elasticities ($dy/d\log x_i = b_i$ so $dy = b_i * dx_i / x_i$ so $b_i/100$ is the unit change in the dependent variable when x_i changes by 1%.)

Regression suggests more doctors leads to a **rise** in infant mortality - strange. (a 1% rise in doctors appears to increase infant mortality by 4 in every 100,000 **not** 4 in every 1000 - be careful to divide the estimated coefficient by 100).

Unintuitive results could suggest presence of outliers in the data.

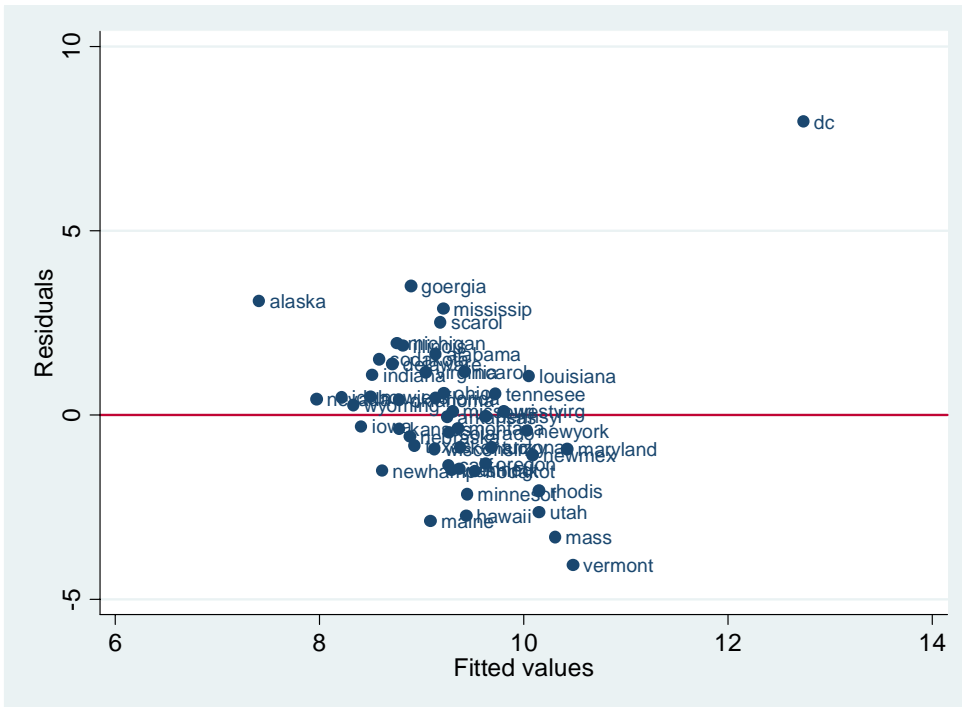
```
. predict res, resid
. scatter res state, mlabel(state) yline(0)
```



In a well-fitted model there should be no systematic pattern to the residuals. Graph below suggests DC may be an outlier, (though absolute size of residual alone may not always be best way to judge, may just tell you model is a bad fit)

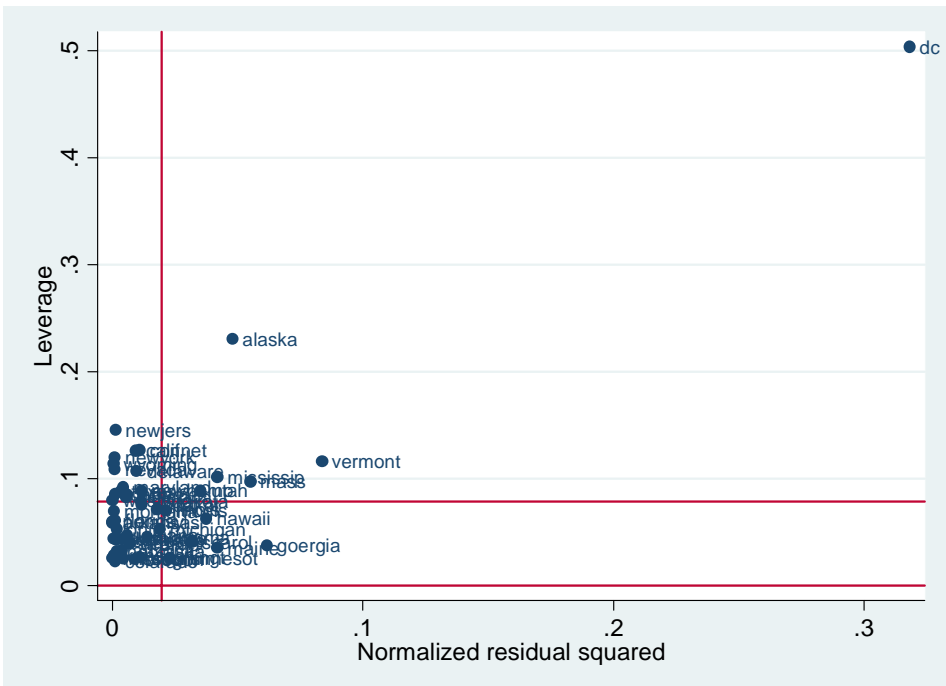
Since in 4 variable model can't plot actual and fitted values against a single X variable. Best can do is graph of residuals against fitted values

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



Plot of leverage against normalised residuals is good way of assessing observations visually. We might consider looking at observations with high leverage and high residuals

```
. lvr2plot, yline(0) mlabel(state)
```



Plot suggests DC has all the symptoms of classic outlier - away from main body of data **and** high residual.

More formal tests given by

```
. predict dfit, dfit
. predict cooks, cooksd

. l state dfit if dfit>2*sqrt(4/51)           (since k=4 & N=51)
```

| | state | dfit |
|-----|--------|----------|
| 24. | dc | 9.130011 |
| 50. | alaska | .956772 |

```
. l state cooks if cooks>4/51
```

| | state | cooks |
|-----|---------|----------|
| 3. | vermont | .1462254 |
| 24. | dc | 7.643065 |
| 50. | alaska | .2192566 |

Both suggest DC is a problem.

```
. tab state if dfit>2*sqrt(4/51)
```

| state | Freq. | Percent | Cum. |
|---------|-------|---------|-------|
| alabama | 1 | 1.96 | 1.96 |
| dc | 1 | 1.96 | 15.69 |
| Total | 51 | 100.00 | |

```
. tab state if dfit>2*sqrt(4/51), nolabel
```

| state | Freq. | Percent | Cum. |
|-------|-------|---------|-------|
| 1 | 1 | 1.96 | 1.96 |
| 8 | 1 | 1.96 | 15.69 |
| Total | 51 | 100.00 | |

So DC has value 8 - exclude

```
. reg infmort lpop ldocs lpcap if state~8
```

| Source | SS | df | MS | Number of obs = | 50 |
|----------|------------|----|------------|-----------------|--------|
| Model | 26.8600265 | 3 | 8.95334216 | F(3, 46) = | 5.76 |
| Residual | 71.4631754 | 46 | 1.55354729 | Prob > F = | 0.0020 |
| | | | | R-squared = | 0.2732 |
| | | | | Adj R-squared = | 0.2258 |
| Total | 98.3232019 | 49 | 2.00659596 | Root MSE = | 1.2464 |

```
-----+-----
```

| infmort | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|---------|-----------|-----------|--------|-------|----------------------|
| lpop | .6292349 | .1911062 | 3.293 | 0.002 | .2445581 1.013912 |
| ldocs | -2.741837 | 1.190773 | -2.303 | 0.026 | -5.138739 -.3449347 |
| lpcap | -.5669275 | 1.641216 | -0.345 | 0.731 | -3.870524 2.736669 |
| _cons | 23.95479 | 12.41946 | 1.929 | 0.060 | -1.044287 48.95388 |

```
-----+-----
```

Now doctor variable is negative and significant (makes more intuitive sense)
 A 1% rise in the number of doctors (per 100,000 inhabitants) leads to a fall in infant mortality of 2.7 per 100,000. The t value also suggests that the effect is statistically significantly different from zero (at the 5% level). - The

