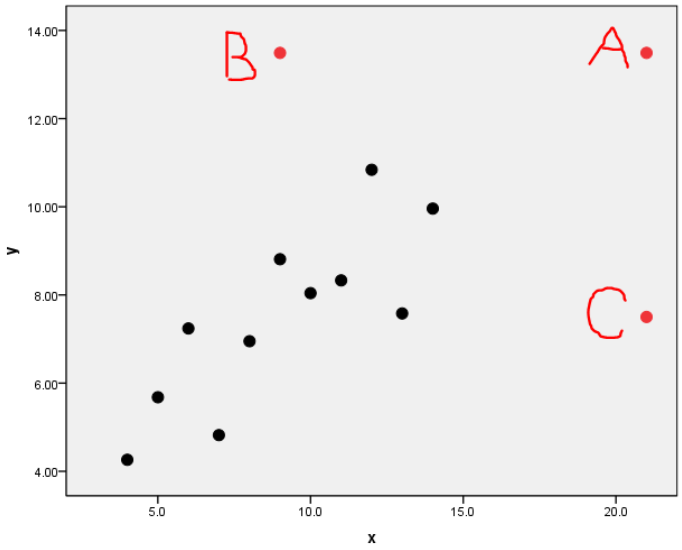
EPRS8550
Regression Diagnostics

Let's consider the following data:



X	Y	YA	YB	YC
10.0	8.04	8.04	8.04	8.04
8.0	6.95	6.95	6.95	6.95
13.0	7.58	7.58	7.58	7.58
9.0	8.81	8.81	8.81	8.81
11.0	8.33	8.33	8.33	8.33
14.0	9.96	9.96	9.96	9.96
6.0	7.24	7.24	7.24	7.24
4.0	4.26	4.26	4.26	4.26
12.0	10.84	10.84	10.84	10.84
7.0	4.82	4.82	4.82	4.82
5.0	5.68	5.68	5.68	5.68
21.0	.	13.49	.	.
9.0	.	.	13.49	.
21.0	.	.	.	7.5

1. Outliers (Residual based)

- Standardized Residuals (ZRESID)

$$ZRESID = \frac{Y - Y'}{S_{y.x}}$$

where $S_{y.x}$ is SE of estimate.

*Criterion |ZRESID| > 2

- Studentized Residuals (SRESID)

$$SRESID = \frac{Y - Y'}{S_{y.x}\sqrt{1 - \left[\frac{1}{N} + \frac{(X - \bar{X})^2}{\sum x^2}\right]}}$$

Note: $\left[\frac{1}{N} + \frac{(X - \bar{X})^2}{\sum x^2}\right]$ is leverage h.

*Criterion |SRESID| > $t_{\alpha, N-k-1}$

- Studentized Deleted Residuals (SDRESID)

Same as above except use $S_{y.x(i)}$ instead of $S_{y.x}$, where $S_{y.x(i)}$ is SE of estimate of data from which i was excluded.

*Criterion |SDRESID| > $t_{\alpha, N-k-2}$

2. Influence Analysis

- Leverage (Related to $X - \bar{X}$)

  h (See above) Max = 1, Min = 1/N

  *Criterion h > 2(k+1)/N

- Cook's D (Related to residual and $X - \bar{X}$)

$$D_i = \left[\frac{SRESID_i^2}{k+1} \right] \left[\frac{h_i}{1-h_i} \right]$$

  *Criterion D_i >4/N, or 1

- DFBETA (Influence on parameters)

  DFBETA0 - change in intercepts
  DFBETA1 - change in slopes

  DFBETAS0 - standardized change in intercepts
  DFBETAS1 - standardized change in slopes

  *Criterion |DFBETAS| > $2/\sqrt{N}$, or $3/\sqrt{N}$ or 1

Regression Diagnostics on SPSS

Analyze
- Regression
  - Linear
    - Dependent y
    - Independent x
      - Save
        - Cook's (Cook's D)
        - Leverage values (h)
        - Standardized (ZRESID)
        - Studentized (SRESID)
        - Studentized deleted (SDRESID)
        - DfBeta (DFBETA)
        - Standardized DfBeta (DFBETAS)

The requested regression diagnostics are saved in your data (see the data window).

Using SPSS

1. Check Maximum value of various outlier statistics.

For example for our example data,

	None	A	B	C
Cook's D	.4892	.2084	.3204	**3.6517**
Leverage	.2273	**.5000**	.2273	**.5000**

2. Examine each observation. For our data, #12 reports the following statistics:

	A	B	C	Criterion
ZRESID[1]	-.00426	**2.54**	-1.47	> 2
SRESID[1]	-.00660	**2.65506**	**-2.2840**	> t_{N-k-1} (Here t_{10}=2.227)
SDRESID[1]	-.00626	**4.63699**	**-3.1330**	> t_{N-k-2}(Here t_9=2.26)
Leverage[2]	**.5000**	.0000	**.5000**	>2(k+1)/n (4/12=.3333)
Cook's D[3]	.00003	.32042	**3.65173**	> 1
DFBETAS0[4]	.00511	.44578	**2.55684**	>1 or $2/\sqrt{n}$ or $3/\sqrt{n}$ (Here .577 or .866)
DFBETAS1[4]	-.00686	.0000	**-3.4320**	>1 or $2/\sqrt{n}$ or $3/\sqrt{n}$ (Here .577 or .866)

[1]Residual based
[2]Related to $X - \bar{X}$
[3]Residual and $X - \bar{X}$
[4]Influence on parameters

A summary of the these analyses follows:

Summary statistics	Y	YA	YB	YC
n	11	12	12	12
$\bar{Y}$	7.5	8.0	8.0	7.5
$\bar{X}$	9.0	10	9.0	10.0
$\Sigma(X-\bar{X})^2$	110.0	242.0	110.0	242.0
b_o	3.0	3.0	3.5	5.23
$b_{Y\|X}$	.5	.5	.5	.23
SSR	27.5	60.39	27.51	12.50
SSE	13.76	13.76	46.64	28.77
MSE	1.53	1.37	4.66	2.88
F	17.98	43.88	5.90	4.34
R^2	~~.82~~ .67	.81	.37	.30