

Lab:
Taking Care of Outliers

OBJECTIVES. In this lab you will

- 1) look at some simple regression results
- 2) look at residual output to decide if you have come cases that are strongly influencing the regression
- 3) re-run the regression leaving that case/those cases out, to see if it makes a difference

EXAMPLE 1:

Predicting 1985 State-Level Reported Property Crime Rates Using AVGPAY

In this example we go ahead to do something about problems revealed by our analysis of residuals.

Our independent variable is average yearly pay in a state in 1985. We will use the square root transform of this variable to make it somewhat more normal We are going to use AVGPAY even though it has high skewness (2.76) and an outlier (AK) because the purpose of this exercise is to show you how a single case influences regression results, and how we can use regression diagnostics to fix things.

Original Simple Regression

First, let's run the regression, and save the the following regression results:

- the predicted score
- the residual
- the leverage
- the studentized residual

Here are the results. Let's interpret the b weight and the R squared, and the beta weight.

```
Block Number 1. Method: Enter AVGPAY
Variable(s) Entered on Step Number
  1.. AVGPAY AVERAGE ANNUAL PAY PER WORKER 1983
```

```
Multiple R .46669
R Square .21780
Adjusted R Square .20150
Standard Error 911.49191
```

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	11104017.41392	11104017.41392

```

Residual          48          39879239.86608          830817.49721
F =          13.36517          Signif F =          .0006
----- Variables in the Equation -----
Variable          B          SE B          Beta          T          Sig T
AVGPAY          .206358          .056446          .466688          3.656          .0006
(Constant)      876.530150      955.057691          .918          .3633

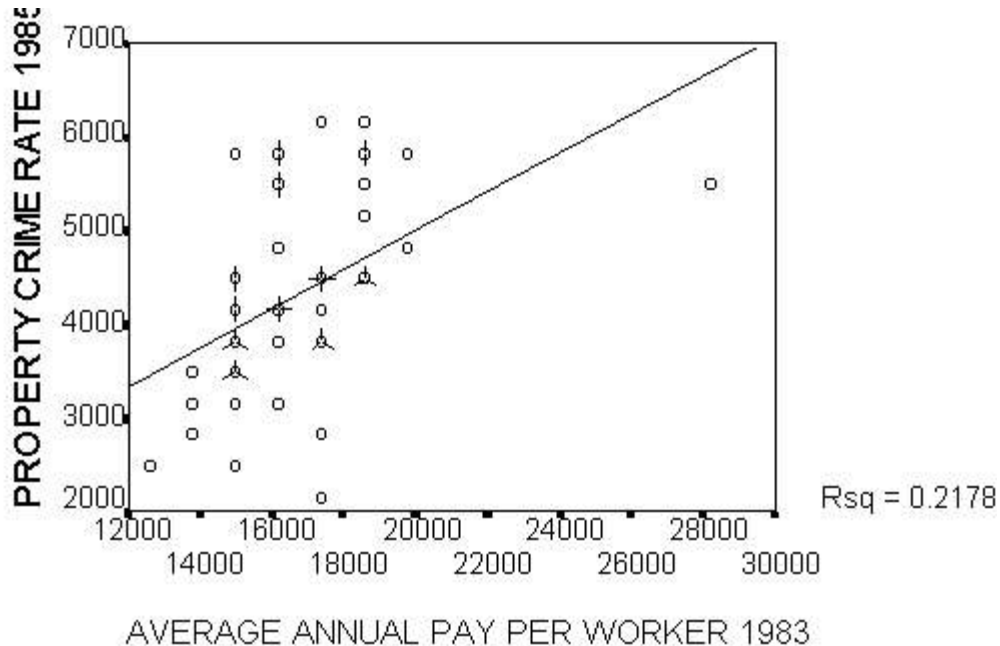
```

Residuals Statistics:

	Min	Max	Mean	Std Dev	N
*PRED	3597.9778	6803.1284	4336.1200	476.0384	50
*ZPRED	-1.5506	5.1824	.0000	1.0000	50
*SEPPRED	128.9094	687.0144	164.2911	79.8029	50
*ADJPRED	3684.7815	8642.1709	4373.4438	687.4895	50
*RESID	-2109.1367	1870.0493	.0000	902.1430	50
*ZRESID	-2.3139	2.0516	.0000	.9897	50
*SRESID	-2.3375	2.0785	-.0163	1.0335	50
*DRESID	-3237.1704	1919.4299	-37.3238	1014.7210	50
*SDRESID	-2.4571	2.1561	-.0161	1.0565	50
*MAHAL	.0001	26.8570	.9800	3.7710	50
*COOK D	.0000	3.5828	.0846	.5051	50
*LEVER	.0000	.5481	.0200	.0770	50

Name	Contents
PRE_1	Predicted Value
RES_1	Residual
SRE_1	Studentized Residual
LEV_1	Leverage

We can see that the maximum leverage of a case is .54. Hamilton tells us that above .2 is possibly risky, and above .5 we should avoid. Let's find this case. Sort your file by the leverage statistic (LEV_1 or LEV_N where N is the number you see in your output), make it a Descending sort, and look at the first case. We see it is ALASKA.



Why does Alaska have such huge leverage? Look at the scatterplot and see if you can see why.

OK, so let's drop Alaska from the file. Just put your blinker on farthest left part of row, click to highlight and delete. Be sure to save the file as something with a different name, so you do NOT confuse your 49 state file with your 50 state file. Also, let's manually drop the variables just added from the regression results.

If you run the regression results again, they look like this:

Listwise Deletion of Missing Data

Equation Number 1 Dependent Variable.. PROPRA85 PROPERTY
CRIME RATE 1985

Block Number 1. Method: Enter AVGPAY

Variable(s) Entered on Step Number

1.. AVGPAY AVERAGE ANNUAL PAY PER WORKER 1983

Multiple R .53884
R Square .29034
Adjusted R Square .27524
Standard Error 867.29292

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	14464176.63749	14464176.63749
Residual	47	35353259.77068	752197.01640

F = 19.22924 Signif F = .0001

----- Variables in the Equation -----

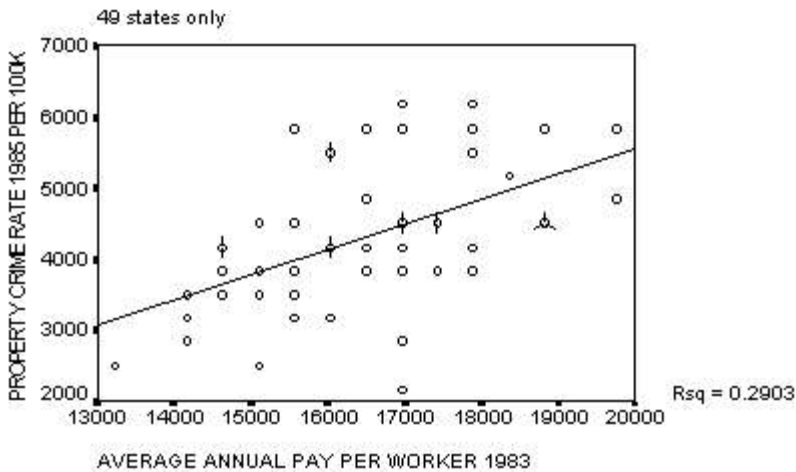
Variable	B	SE B	Beta	T	Sig T
AVGPAY	.354773	.080904	.538835	4.385	.0001
(Constant)	-1546.900075	1342.344011		-1.152	.2550

End Block Number 1 All requested variables entered.

* * * * M U L T I P L E R E G R E S S I O N * *

Equation Number 1 Dependent Variable.. PROPRA85 PROPERTY
CRIME RATE 1985

Check your residual output? Do you see any high leverage cases?



```

* COMMANDS
* YOUR DRIVE DESIGNATIONS WILL BE DIFFERENT
* .
GET
  FILE='C:\PCW\GRSTAT96\USDANU12.SAV' .
EXECUTE .
*****
* NOW DOING THE FIRST REGRESSION
* .
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT propa85
  /METHOD=ENTER avgpay
  /SAVE PRED LEVER RESID SRESID .
*****
* SORTING CASES BY LEVERAGE
* .
SORT CASES BY
  lev_1 (D) .
*****
* GRAPHING THE REGRESSION
* .
GRAPH
  /SCATTERPLOT(BIVAR)=avgpay WITH propa85
  /MISSING=LISTWISE .
* Here we manually dropped Alaska, and
* then re-saved the file with different name.
* Also suggest you drop the four variables at the
* right hand side of the file, that were just added
* SAVE FILE W/ A DIFFERENT NAME
* .
SAVE OUTFILE='C:\PCW\GRSTAT96\US_49.SAV'
  /COMPRESSED.
*****
* NOW RUN THE REGRESSION AGAIN
* .
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT propa85
  /METHOD=ENTER avgpay
  /SAVE PRED LEVER RESID SRESID .
*****
* LOOK AT RESULTS AGAIN
* .

GRAPH

```

```
/SCATTERPLOT(BIVAR)=avgpay WITH propa85  
/MISSING=LISTWISE .
```