Influence Statistics in Linear Regression and SAS® PROC REG

by Bruce Lund

Statistical Modeling Consultant and Trainer, Novi, MI

blund_data@mi.rr.com and blund.data@gmail.com

Announcements

We'll rapidly go through first few slides

These slides will be posted to MSUG site.

I have a paper which includes explanations, examples, and math background on the topic of influence statistics for linear regression

- Send me an email to receive paper.

Good reference books:

Fox, J. (2019) *Regression Diagnostics An Introduction 2nd Ed* (2019), Newbury Park, CA, SAGE Publishing

Hamilton, L. C. (1992). *Regression with Graphics, A Second Course in Applied Statistics*, Belmont, CA: Wadsworth. (short but well written discussion in chapter 4)

Montgomery, D.C., Peck, E. A., Vining, G. G. (2012). *Introduction to Linear Regression Analysis, fifth edition,* Hoboken, NJ: Wiley. (complete discussion)

Linear Regression Model Review

A second class in statistics covers Linear Regression with multiple predictors. Multiple predictors requires use of matrix algebra. Let's begin ...

- k predictors X1(=1), X2, ... Xk AND n cases (sample points) called x1 - xn.

  The design matrix $\mathbf{X}_{nxk}$ has columns X1 – Xk
  - X1 = 1 ... the intercept "predictor".
- Each row of **X** is a case.
- Column matrix Y_{nx1} is the target.
- Assume k < n
- It is assumed: rank(X) = k = # of predictors

This implies $(X^T * X)^{-1}$ exists

cases n	DESIGN X (k predictors)			Y
	X1	X2	X3	
x1	1	-1	0	1
x2	1	1	1	0
x3	1	0	0	2
x4	1	1	0	3
x5	1	-1	-1	1

The linear regression model is: $Y = X*\beta + \varepsilon$ ... where:

β_{kx1} is column matrix of unknown parameters ... k rows

ε_{nx1} is column matrix of errors ... n rows.

Linear Regression Model $Y = X*\beta + \varepsilon$ ... fit by Least Squares

PROC REG computes by Least Squares:

- B ... the estimate of β

$$B = (X^T * X)^{-1} * X^T * Y$$

- $\hat{Y}$ ... estimates / predictions of Y

using the least squares formula's below:

$$\hat{Y} = X * B$$

Alternatively, $\hat{Y} = H * Y$

where $H_{nxn} = X * (X^T * X)^{-1} * X^T$

[John Tukey called H the hat matrix]

$$e_{nx1} = (Y - \hat{Y}) = (Y - H*Y) = (Y - X*B)$$

where e is called the residual.

Standard regression assumptions are made

Population Model $Y = X*\beta + \varepsilon$ assumptions:

- The ε are random variables - identically distributed with mean 0 and common σ^2
- The ε are independent of predictors X_1 to X_K and are independent of one another.
- Either sample size is large or ε's are normal.

In cases where X are random variables, then

$$E(\varepsilon \mid X_j) = 0 \text{ for all j.}$$

Error term: mean zero, constant variance, independent, normal

Goal of this Talk

The goal is to discuss the manner in which a sample point (y, x) influences the value of $\hat{y}$ at x, as well as influences the overall model fit.

Potential of x_i to be influential depends on its position of x_i within the sample … or its Leverage

Leverage of x_i is measured by h_{ii} the i^{th} entry in the diagonal of H, the hat matrix

Leverage (= h_{ii}) depends solely on the hat matrix.

In addition to leverage, the influence of (y, x) depends on whether y is an Outlier.

That is, whether y deviates extremely from the fitted model's predicted value $\hat{y}$ at case x.

It is both figurative and literal that Influence = Leverage x Outlierness

INFLUENCE is measured by: LEVERAGE, RSTUDENT, COOKD, DFBETAS, STUDENT, COVRATIO, DFFITS

Leverages (BLUE) are on main diagonal … called h_{ii}

We use PROC REG but GLM has similar statistics

DESIGN X (k predictors)		
X1	X2	X3
1	-1	0
1	1	1
1	0	0
1	1	0
1	-1	-1

$H = X * (X^T * X)^{-1} * X^T$				
0.70	0.20	0.20	-0.30	0.20
0.20	0.70	0.20	0.20	-0.30
0.20	0.20	0.20	0.20	0.20
-0.30	0.20	0.20	0.70	0.20
0.20	-0.30	0.20	0.20	0.70

Properties of H and h_{ii}

- H is n x n matrix, symmetric, H*H = H, and each row or column of H sums to 1
- Trace(H) = Rank(X) = k
- $1/n \le h_{ii} \le 1$ and $-1/2 < h_{is} \le 1/2$
- If x repeats w times in X, then $1/n \le h \le 1/w$

Average h_{ii} is k / n

A predicted value $\hat{y}_i$ equals leverage times actual y plus other terms.

$$\hat{y}_i = \sum_{s=1}^{n} h_{is} y_s = h_{ii} y_i + \sum_{s \ne i}^{n} h_{is} y_s$$

X	X1	X2	X3	H					
x1	1	-1	0	0.68	0.24	0.18	-0.32	0.12	0.12
x2	1	1	1	0.24	0.65	0.24	0.24	-0.18	-0.18
x3	1	0	0	0.18	0.24	0.18	0.18	0.12	0.12
x4	1	1	0	-0.32	0.24	0.18	0.68	0.12	0.12
x5	1	-1	-1	0.12	-0.18	0.12	0.12	0.41	0.41
x6	1	-1	-1	0.12	-0.18	0.12	0.12	0.41	0.41

If $h_{ii} = 1/n$, then small leverage on y_i

If h_{ii} is larger, (so h_{is} are smaller, since row sum is 1), then more leverage on y_i

Next Slide ...

Example Dataset (see source below)

Dataset: “concord1” with 496 cases (household) of water use during 1981 in Concord, NH.
Four columns used from “water”:

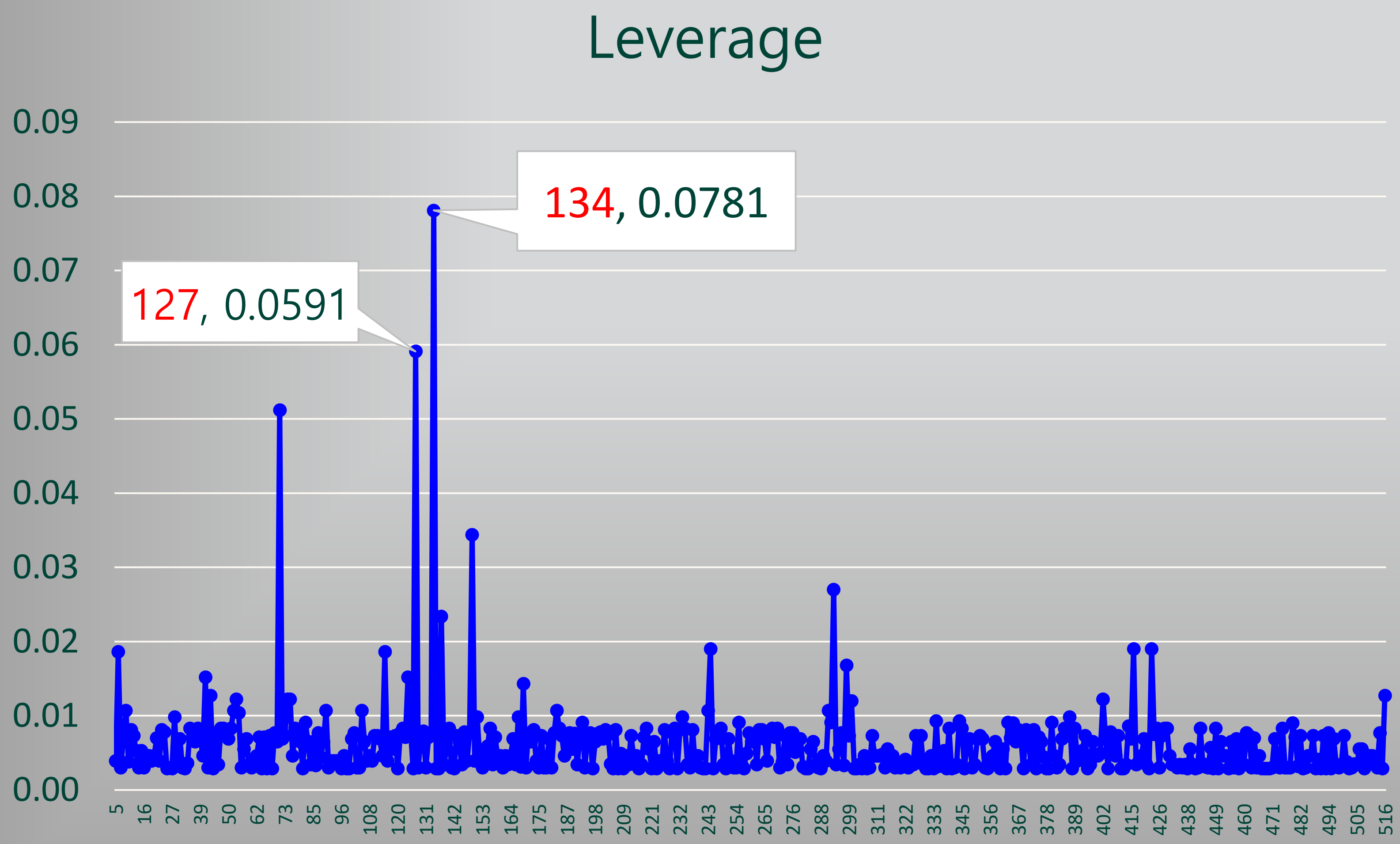
- “Case” is the Case number (ID)
- “water81” is response or target (numeric) … water use in 1981
- “income” is predictor (numeric in 000's)
- "retireN" is predictor (1 = retired, 0 = not retired)

Reference: *Regression with Graphics* (1992) by L. Hamilton
Source Data: https://stats.oarc.ucla.edu/wp-content/uploads/2016/02/concord1.sas_.txt

Fit a model and then look at influence statistics

$$\widehat{\text{water81}} = \text{B1} + \text{B2} * \text{income} + \text{B3} * \text{retireN}$$

Compute Leverage (diagonals of Hat Matrix)



Leads to question:

What will be the impact of Case 134 on a model?

Maybe it's OK, maybe not

(So far, have not run a model)

EXCEL chart

Regression Model Fit looks Good … But what about Case 134?

Parameter Estimates					
Variable	DF	Estimate	Std Err	t Value	Pr > |t|
Intercept	1	1462.507	149.42819	9.79	<.0001
income	1	41.768	4.98923	8.37	<.0001
retireN	1	-434.784	142.80327	-3.04	0.0025

Declare good fit and stop?

But Case 134 was disturbing.

- Impact on Coefficients if Case 134 deleted?
- How does $\hat{Y}$ change if Case 134 is deleted?

```
ODS OUTPUT OutputStatistics=outStats;
PROC REG DATA = concord1; ID case;
MODEL water81 = income retireN / INFLUENCE;
OUTPUT OUT= outREG predicted= water81hat
cookd=COOKD student=STUDENT press=PRESS;
quit;
```

OUTPUT OUT adds to each Case:

COOKD (and others)

INFLUENCE and ODS OUTPUT adds to each Case:

RESIDUAL, RSTUDENT, LEVERAGE, DFBETAS (and others)

Influence Statistics are based on "Row Deletion"

Notation: "(i)" refers to a model that is fit on data with Case i deleted

$\hat{y}_{(i)s}$ = Estimate of y at Case s … for model with Case **i** deleted

Estimated error variance for the Model with Case i deleted:

$\hat{\sigma}^2_{(i)} = \sum_{s \neq i}^{n} (y_s - \hat{y}_{(i)s})^2$ / (n-k-1) … sum over all cases except Case **i**

$\hat{Y}_{(i)}$ = n x 1 *column* of estimated targets … for model with Case **i** deleted

Next Slide …

RSTUDENT (a standardized residual)

A statistic based on a row deletion:

$$\text{RSTUDENT}_i = \frac{e_i}{\text{sqrt}(\sigma^2_{(i)}(1-h_{ii}))}$$

Full model residual for case i

Full model leverage for case i

It is not necessary to perform 496 regressions in order to compute these 496 !!!

In fact, all the influence statistics are computed from full model results.

Here is the formula that links $\sigma^2_{(i)}$ to full model results:

$$\hat{\sigma}^2_{(i)} = \hat{\sigma}^2 \,(n-k)/(n-k-1) \; - \; e_i^2 \,/\, ((n-k-1)\,(1-\,h_{ii}))$$

All quantities on the RHS are found from the full model.

RSTUDENT is a t-statistic with n-k-1 d.f. … guideline for extreme > | 3 |

(But the RSTUDENT's are not independent of each other)

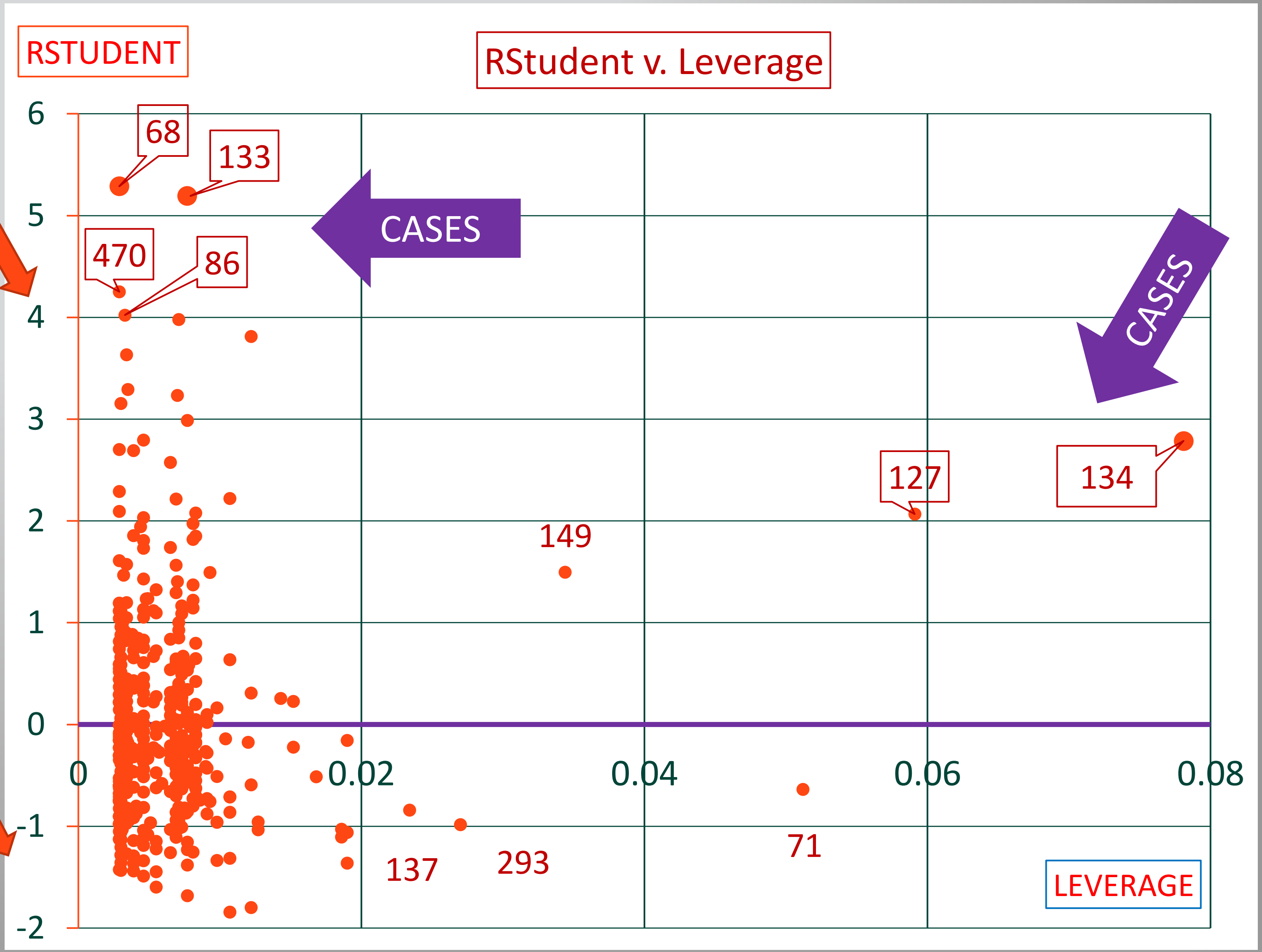
Plot of LEVERAGE (x-axis) v. RSTUDENT (y-axis)

- All extreme RSTUDENT are positive
- PROB (RSTUDENT > 4) = 0.00002
- Case 134: HIGH leverage with fairly high RSTUDENT … examine !
- Case 68, and others … maybe NOT influential because low leverage

What to do?

COOKD gives scaled distance between $\hat{Y}$ and $\hat{Y}_{(i)}$

- It shows "global" effect of deletion of Case i



EXCEL chart

Cook's D (1977)

Delete Case i and fit the model. Obtain estimates $\hat{y}_{(i)s}$ for s=1 to n

Take square of Euclidean distance between:

$\hat{Y}$ and $\hat{Y}_{(i)}$ … this squared distance is $\sum_{s=1}^{n}(\hat{y}_s - \hat{y}_{(i)s})^2$

Standardize by k $\hat{\sigma}^2$ (k=# predictors and $\hat{\sigma}^2$ variance of sample)

$$\text{Cook's } D_i = \sum_{s=1}^{n}(\hat{y}_s - \hat{y}_{(i)s})^2 / k\,\hat{\sigma}^2$$

Use COOK's D to rank cases

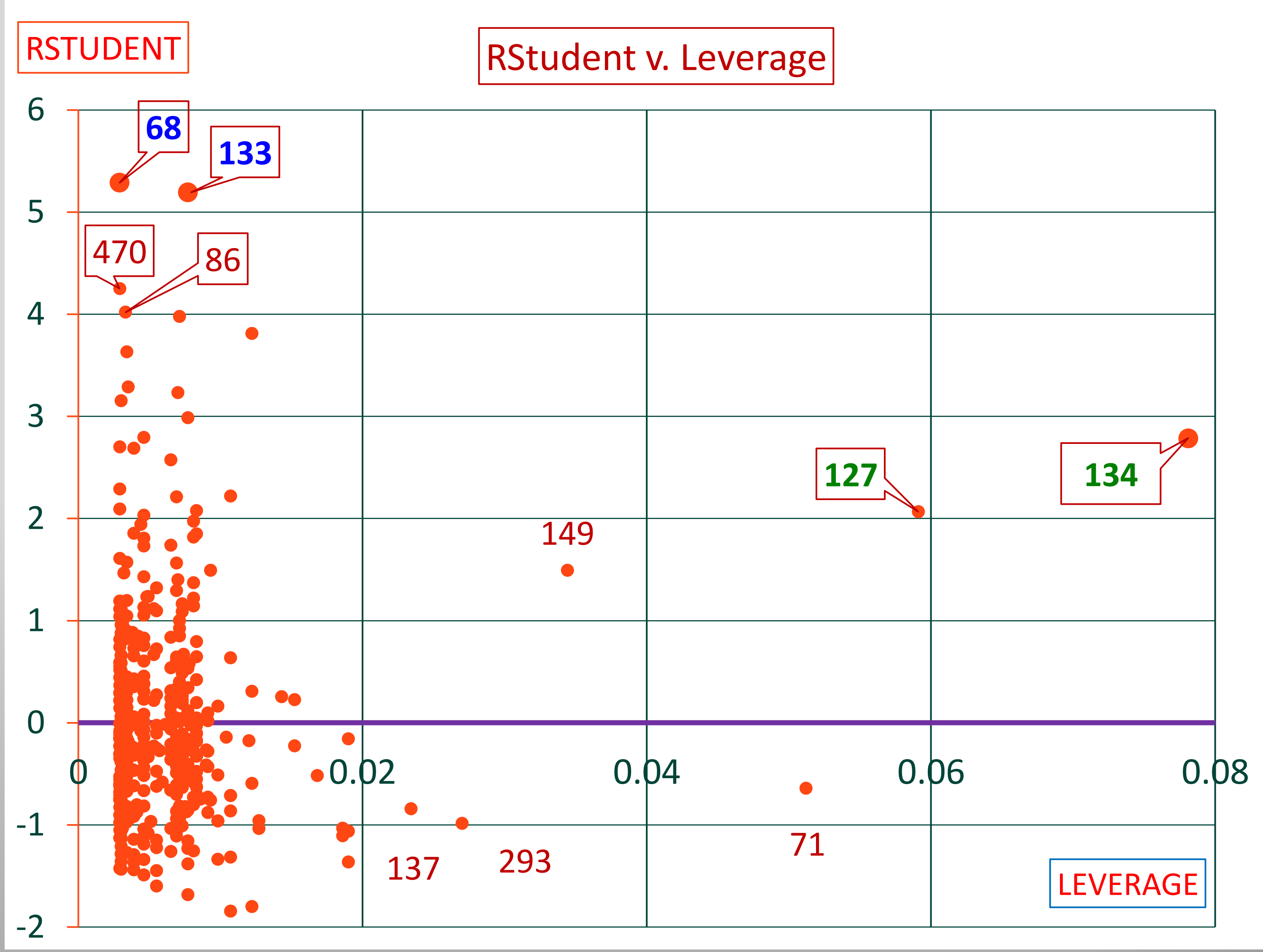
… high ranked COOK's D suggests (y_i, x_i) has high influence.

$$\text{Cook's } D_i = h_{ii} / (1 - h_{ii})^2 \times (1/k)\,(e_i/\hat{\sigma})^2$$

Influence = Leverage x Outlier

Cook's D ... Ranked

Case i	COOKD	RStudent	Leverage
134	**0.216**	2.79	0.078
127	0.089	2.07	0.059
133	0.066	5.19	0.008
74	0.058	3.81	0.012
117	0.037	3.98	0.007
149	0.026	1.49	0.034
68	0.026	5.29	0.003
371	0.024	3.23	0.007



Based on Cook's D
– look only at Case 134

Case	COOKD	RStudent	Leverage	Residual	water81	income	retireN
134	0.216	2.79	0.078	3561	9200	100	0
127	0.089	2.07	0.059	2678	7900	90	0

Extreme Incomes

$DFBETAS_i$

Belsey, D.A., Kuh, E., and Welsch, R.E., (1980)

$DFBETAS_i$ is change in B when Case i is deleted
... But then divided by a scaling factor

	DFBETAS		
Case	Intercept	income	retireN
134	-0.6243	0.7961	0.2252

abs(DFBETAS) > 2 / sqrt(n) is cause for examination.
2 / sqrt(496) = 0.09 ... all 3 DFBETAS are well over guide

B for Full Model	
Intercept	1462.511
income	41.768
retireN	-434.784

B(134): Model w/o Case 134	
Intercept	1555.155
income	37.823
retireN	-466.718

Change in B's	
Intercept	-92.644
income	3.945
retireN	31.934

PROC REG does not give the "Change in B's" directly.
A SAS program is needed to get "Change in B's"
... see "Extra Slides" for this code.

What is effect of deleting Case 134 on estimates?

For income ≥ 60 (1%), the % change in estimates v. full model exceeds 3%.

For income ≤ 8 (10%), the % change in estimates v. full model exceeds |3%|

Do these differences matter?
This is a subject matter decision.

If YES, then consider …

- New model with more X's
- Robust regression methods
- Transform INCOME or WATER81?
- … or …
- Remove Case 134

Modeler has information to decide on next steps.

Income Percentile	Income	RetireN	Estimates Full	Estimates no 134	Change	% Chg v. Full
99%	60	0	3969	3825	144	3.6%
	60	1	3534	3358	176	5.0%
75%	30	0	2716	2690	26	0.9%
	30	1	2281	2223	58	2.5%
50%	21	0	2340	2349	-10	-0.4%
	21	1	1905	1883	22	1.2%
25%	15	0	2089	2123	-33	-1.6%
	15	1	1654	1656	-2	-0.1%
10%	8	0	1797	1858	-61	-3.4%
	8	1	1362	1391	-29	-2.1%
5%	5	0	1671	1744	-73	-4.4%
	5	1	1237	1278	-41	-3.3%
1%	3	0	1588	1669	-81	-5.1%
	3	1	1153	1202	-49	-4.2%

A new scenario … weighted regression and influence

Now suppose there are 10 households … all with income 100 and not retired

Suppose water usage is high for all 10 and the average is 9200.

This is simply Case 134 repeated 10 times.

NEW FACT: These 10 are in a new subdivision.
There was extra water use in 1981 due to landscaping.

- Maybe the entire subdivision should be deleted?
- What is the impact on the model of deleting Case 134 (with 10 repetitions)?

Can answer the question by weighted regression with weights w.

$w_{134} = 10$, else $w_i = 1$.

The least squares solution for the parameters B of a weighted regression is:

$B = (X^T*W*X)^{-1}*X^T*W*Y$ … where is diagonal with entries w_i

How are the influence statistics changed?

Influence Statistics for weighted Case 134

Parameter Estimates			
Variable	DF	Estimate	Pr > |t|
Intercept	1	1011.23	<.0001
income	1	60.98	<.0001
retireN	1	-279.24	0.0559

w	Case	COOKD (sorted)	RStudent	Leverage	DFB_ Intercept	DFB_ income	DFB_ retireN
10	134	11.81	6.75	0.459	-4.22	6.02	1.34
1	74	0.04	3.39	0.010	-0.16	0.19	0.29
1	133	0.04	4.67	0.005	-0.06	0.23	-0.04
1	117	0.03	3.77	0.007	-0.04	0.05	0.26
1	68	0.02	5.07	0.003	0.18	-0.05	-0.15

```
ods output OutputStatistics=outStatsx;
PROC REG DATA = concord1;
WEIGHT w; /* Do not use FREQ */
ID case;
MODEL water81 = income retireN / INFLUENCE;
OUTPUT OUT= outREGx predicted= water81hat
cookd= COOKD student= STUDENT press= PRESS;
quit;
```

- Regression is significant … but
- HIGH influence for Case 134
- Estimates $\hat{Y}$ are **much** changed (but not shown)

WHAT TO DO?

- income_sq = income**2 is highly significant
  … Investigate a new model ??
- delete Case 134 (all 10) and analyze separately
  Sensible, households are a new subdivision

Influence Statistics and Weighting

Influence Statistics with weights make sense when …

- Survey sample with sampling weights
- True repeats as in a designed experiment
  - Must average the Y's. Weight is the number of repeats
- Pseudo repeats. Group cases with very similar X's
  - Must average the Y's. Weight is the number in the group

Weights are also used to correct for unequal error variances (heteroscedasticity)

This is a different situation …

Here, the weights don't represent a count or projection of cases.

However, the same influence statistics are produced,

but what do they mean?

Supplementary Slides at end of Deck

Influence Statistics in Linear Regression and SAS® PROC REG

by Bruce Lund

Statistical Modeling Consultant and Trainer, Novi, MI

blund_data@mi.rr.com and blund.data@gmail.com

How weights appear in Influence Statistics

In weighted regression the coefficients are found by minimizing the weighted squared residuals $\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} w_i (e_i)^2$ and $\sigma^2 = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 / (n-k)$

```
DATA test2;
INPUT X Y W;
DATALINES;
1 -1 1
1  1 1
0 -1 1
0  1 1
2  2 4
;
```

Weighted hat matrix: $H_w = X*(X^T*W*X)^{-1}*X^T*W$

H_W				
0.136	0.136	0.182	0.182	0.364
0.136	0.136	0.182	0.182	0.364
0.182	0.182	0.409	0.409	-0.182
0.182	0.182	0.409	0.409	-0.182
0.091	0.091	-0.046	-0.046	**0.909**

Weight increases leverage.
$H_{55} = 0.909$

$\hat{\sigma}^2_{(i)} = \hat{\sigma}^2 (n-k)/(n-k-1) - e_i^2 w_i / ((n-k-1)(1- h_{ii}))$

- $\hat{\sigma}^2$ and h_{ii} are weighted
- Notice: "$e_i^2 w_i$" ... squared residual is weighted

Idempotent: $H_w*H_w = H_w$

Projection of X onto X: $H_w*X = X$

As a result: Hw*1 = 1 ... $\sum_j^n (hw)_{ij} = 1$

$\text{Trace}(H_w) = \text{Rank}(H_w) = k$

H_w not symmetric. Column sums $\sum_i^n h_{ij} \neq 1$

PROC REG gives DFBETAS but not DEBETA or $B_{(i)}$

Annoying ….

PROC REG gives DFBETAS (standardized) but not the simple DFBETA (difference of B and $B_{(i)}$)

With programming and using the formula below, the DFBETA can be computed for each case:

$$DFBETAS_{(i)j} = (b_j - b_{(i)j}) / (\hat{\sigma}_{(i)} \, sqrt((X^T*X)^{-1}{}_{jj})) \ \ldots \text{ for } j = 1 \text{ to } k$$

where $(X^T*X)^{-1}{}_{jj}$ is the jth entry on the diagonal of $(X^T*X)^{-1}$

The plan is to multiply $DFBETAS_{(i)j}$ (reported by SAS) by ($\hat{\sigma}_{(i)}$ $sqrt((X^T*X)^{-1}{}_{jj})$) for j = 1 to k

Needed: $\hat{\sigma}_{(i)}$ and $(X^T*X)^{-1}{}_{jj}$

(1) Solve for $\hat{\sigma}_{(i)}$ in equation: $RSTUDENT_i = e_i / (\hat{\sigma}_{(i)} (1 - h_{ii})^{1/2})$ … e_i and h_{ii} are reported by SAS.

(2) Obtain $(X^T*X)^{-1}$ from ODS OUTPUT InvXPX = InvXPX … from PROC REG

This ODS OUTPUT requires that option I be included on the MODEL statement:

```
MODEL water81 = income retireN / INFLUENCE I;
```

SAS code is given on the next slide

Compute DFBETA with SAS program

```
ods graphics off;
ods output InvXPX = InvXPX;
ods output OutputStatistics=outStats;
PROC REG DATA = concord1;
ID case;
MODEL water81 = income retireN / INFLUENCE I;
OUTPUT OUT= outREG predicted= water81hat
cookd= COOKD student= STUDENT press= PRESS;
quit;
DATA InvXPX_diag; SET InvXPX;
keep
invXX_intercept invXX_income invXX_retireN;
retain
invXX_intercept invXX_income invXX_retireN;
if _N_ = 1 then invXX_intercept = intercept;
if _N_ = 2 then invXX_income = income;
if _N_ = 3 then invXX_retireN = retireN;
if _N_ = 3 then output;
run;
```

```
%LET keep1 =
case RESIDUAL RSTUDENT HatDiagonal DFB_intercept DFB_income DFB_retireN;
DATA DFBETA; SET OUTstats(keep=&keep1);
retain invXX_intercept invXX_income invXX_retireN;
if _N_ = 1 then SET InvXPX_diag;
SIGMA_i = RESIDUAL/(RSTUDENT * sqrt(1 - HatDiagonal));
DFBETA_intercept = DFB_intercept * SIGMA_i * sqrt(invXX_intercept);
DFBETA_income = DFB_income * SIGMA_i * sqrt(invXX_income);
DFBETA_retireN = DFB_retireN * SIGMA_i * sqrt(invXX_retireN);
run;
PROC PRINT DATA = DFBETA(obs=2);
var case DFBETA_intercept DFBETA_income DFBETA_retireN;
run;
```

Covariance Ratio

From least squares, the covariance matrix V of estimated coefficients is:

$$V = \sigma^2 (X^T * X)^{-1}$$

Let $X_{(i)}$ be the design matrix with Case x_i deleted.

Assume $X_{(i)}{}^T * X_{(i)}$ has rank k. Let $V_{(i)} = \sigma^2_{(i)} (X_{(i)}{}^T * X_{(i)})^{-1}$

$COVRATIO_i = \det(V_{(i)}) / \det(V) = (\hat{\sigma}^2_{(i)} / \hat{\sigma}^2)^k \det((X_{(i)}{}^T * X_{(i)})^{-1}) / \det((X^T * X)^{-1})$

With some razzle-dazzle: $COVRATIO_i = (\hat{\sigma}^2_{(i)} / \hat{\sigma}^2)^k / (1 - h_{ii})$

Motivation underlying COVRATIO comes from equality of absolute value of det(V) and volume of the parallelepiped spanned by the columns of V. A larger det(V) implies that the associated parallelepiped has longer sides with angles between the sides that are closer to 90 degrees instead of 0 or 180. This says that variances of coefficient estimates are larger and covariances are smaller. In this sense a large determinant det(V) indicates less precision in the estimates B. A value of $COVRATIO_i$ not close to 1 indicates the precision of the estimates has changed substantially with deletion of case i.

Extreme influence is indicated when COVRATIO falls outside the interval (1 ± 3k/n)

For concord1 this range = (0.982 1.018) … $COVRATIO_{134}$ = 1.041 … outside the range.

PROC REG with MODEL / INFLUENCE reports COVRATIO.

I think DFBETA, DFBETAS, and COOK'S D are more transparent and useful.

Leverage is determined by the position of x_i in the sample

For k = 2 (intercept X_1 and predictor X_2) and with sample = n, there is this formula for the leverage points:

$$h_{ii} = 1/n + (x_{2i} - \bar{x}_2)^2 / \sum_{s=1}^{n} (x_{2s} - \bar{x}_2)^2$$

If x_{2i} is at the mean, then $(x_{2i} - \bar{x}_2) = 0$ and $h_{ii} = 1/n$

$$\hat{y}_i = \sum_{s=1}^{n} h_{is} y_s = y_i/n + \sum_{s \neq i}^{n} h_{is} y_s$$

The "leverage" of $h_{ii} = 1/n$ on y_i is small

When x_{2i} is extreme, then h_{ii} is large (but ≤ 1)

$$\hat{y}_i = \sum_{s=1}^{n} h_{is} y_i = h_{ii} y_i + \sum_{s \neq i}^{n} h_{is} y_s$$

The "leverage" of h_{ii} on y_i is large

Only situation where $h_{ii} = 1$

X	X1	X2	X3	X_D
x1	1	-1	0	0
x2	1	1	1	0
x3	1	0	0	0
x4	1	1	0	0
x5	1	-1	-1	0
x6	1	-1	-1	0
x_d	0	0	0	1

If predictor X_D is zero for all rows except "d"

and

X_D is independent of the other X's, then $h_{dd} = 1$.

So, if x_d is exceptional, then could add dummy variable X_D to model …

This makes: $h_{dd} = 1$ AND $y_d = \hat{y}_d$

Regression plane at x_d passes thru y_d

See $H * Y = \hat{Y}$ as shown on below

$$\begin{bmatrix} \dots & \dots & 0 \\ \dots & \dots & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \dots \\ y_d \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \dots \\ y_d \end{bmatrix}$$

No HAT

Variance of Residual e_r and STUDENT

The residual e_r has a distribution which depends on $Var(\varepsilon_r) = \sigma^2$ and also on X via H.

Begin with the matrix formula for the residuals:

$e = (I - H)*Y$ … a matrix equation with n x 1 column matrices

… Now, show that $Var(e_i) = \sigma^2(1 - h_{ii})$ … where σ^2 = variance of ε_r

Recall that: σ^2 is estimated by $\hat{\sigma}^2 = \sum_{s=1}^{n}(y_s - \hat{y}_s)^2 / (n-k)$

The STUDENT residual for Case i is given by:

$$STUDENT_i = \frac{e_i}{sqrt(\hat{\sigma}^2(1-h_{ii}))}$$