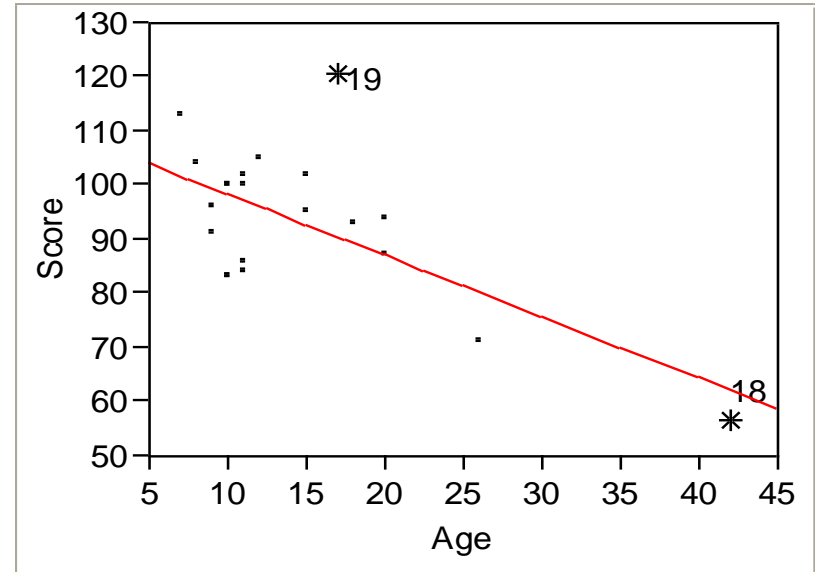


- Outliers and influential observations in simple linear regression (Review)
- Outliers and influential observations in multiple linear regression
- Leverage plots

Outliers and influential points in simple regression

- Does the age at which a child begins to talk predict a score on a test of mental ability at a later age?
- gesell.JMP contains data on the age at first word (x) and their Gesell Adaptive score (y), an ability test taken at a later age.
- Child 18 is an outlier in the x direction, so it is a leverage point and potentially influential.
- Child 19 is a regression outlier.



Outliers in Simple Linear Regression

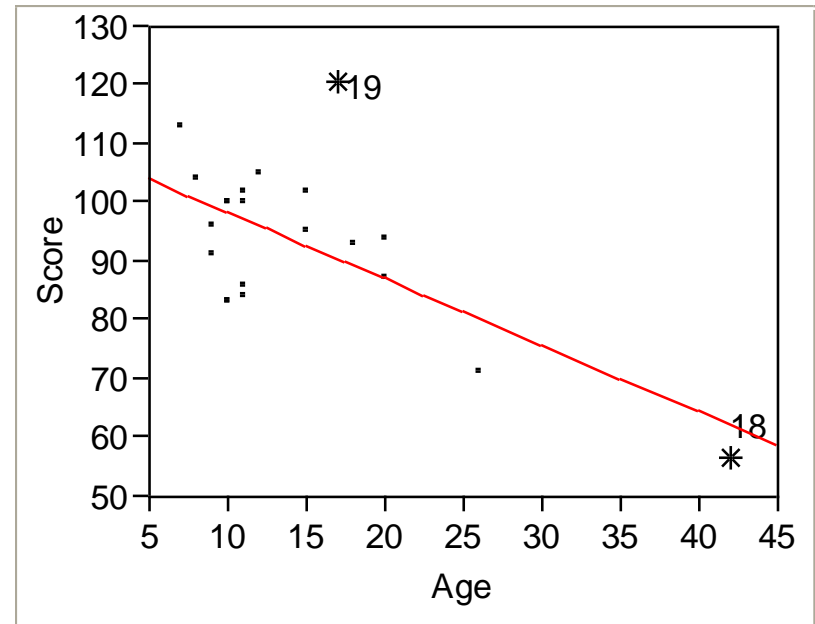
- An outlier is an observation that is unusually small or large.
- Three types of outliers in scatterplots:
 - Outlier in x direction
 - Outlier in y direction
 - Outlier from regression line of scatterplot (residual has large magnitude)
- Several possibilities need to be investigated when an outlier is observed:
 - There was an error in recording the value.
 - The point is not representative of the population of interest.
 - The observation is valid.
- Identify outliers from the scatterplot

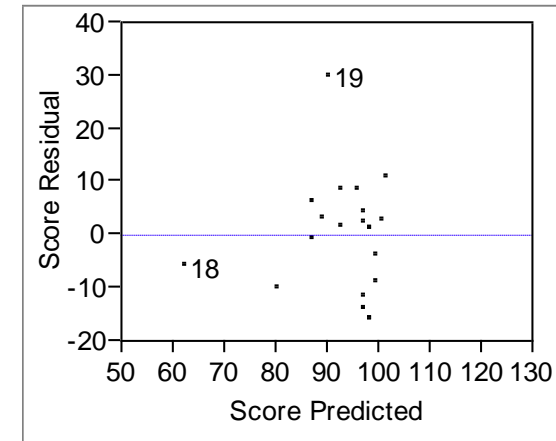
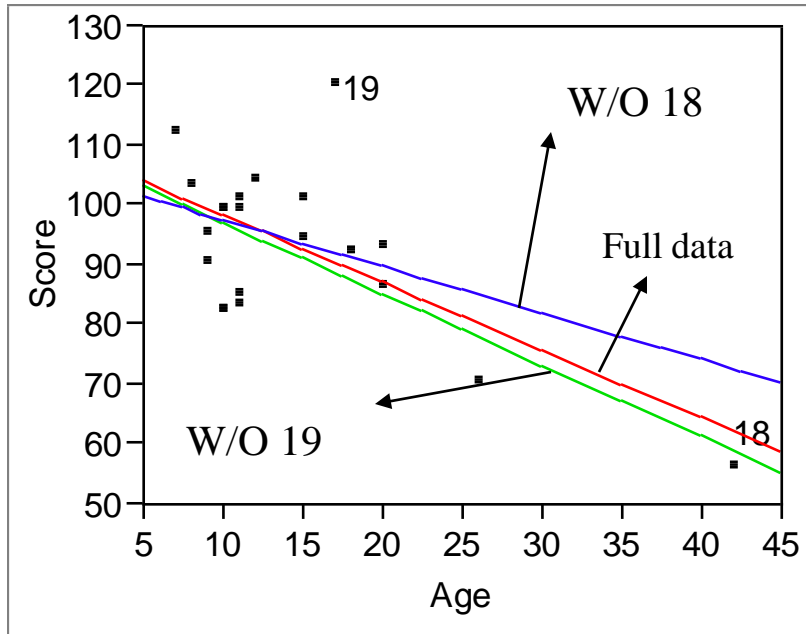
Leverage and Influential Points

- An observation has high leverage if it is an outlier in the x direction.
- An observation is influential if removing it would markedly change the least squares line.
- Observations that have high leverage and are outliers tend to be influential.

Outliers and influential points in simple linear regression

- To assess whether a point is influential, fit the least squares line with and without the point (excluding the row to fit it without the point) and see how much of a difference it makes.
- Child 18 is highly influential; child 19 is not highly influential.





Residuals of full data

Full data: Rsquare=0.41
 Score = 109.87 - 1.127 Age

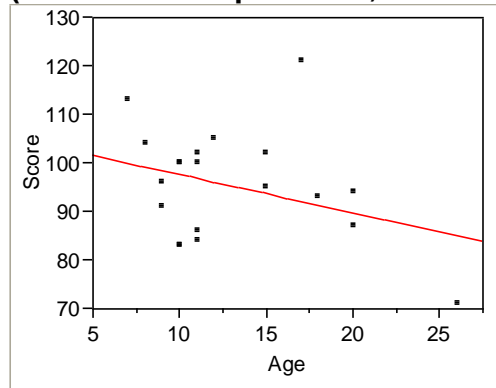
W/O 18: Rsquare=0.11
 Score = 105.63 - 0.779Age

#18: High leverage and influential.

W/O 19: Rsquare=0.57
 Score = 109.30 - 1.193Age

19: possible outlier

**Bivariate Fit of Score By Age
(w/o influential point #18)**



Linear Fit
 $\text{Score} = 105.63 - 0.78\text{Age}$

Summary of Fit
 RSquare 0.1121

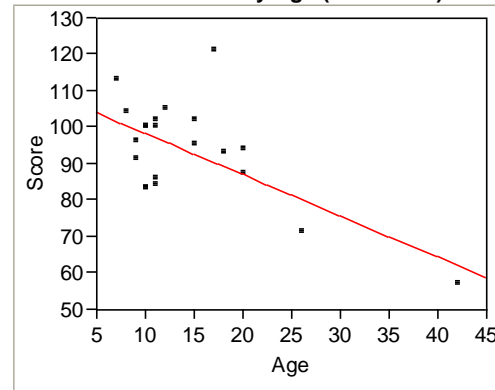
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	280.5195	280.519	2.2740
Error	18	2220.4805	123.360	Prob > F
C. Total	19	2501.0000		0.1489

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	105.62987	7.161928	14.75	<.0001
Age	-0.779221	0.516733	-1.51	0.15

Bivariate Fit of Score By Age (all data)



Linear Fit
 $\text{Score} = 109.87 - 1.13\text{Age}$

Summary of Fit
 RSquare 0.41

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1604.0809	1604.08	13.2018
Error	19	2308.5858	121.50	Prob > F
C. Total	20	3912.6667		0.0018

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	109.87384	5.067802	21.68	<.0001
Age	-1.126989	0.310172	-3.63	0.0018

Conclusion: It is not clear at all that scores and ages are related for normal children

How to identify outliers and influential points
in multiple regression?

Leverage Plot!

Outliers, leverage and influential points in multiple regression

Pollution Example

- Data set pollution. JMP provides information about the relationship between pollution and mortality for 60 cities between 1959-1961.
- The variables are
 - y (MORT)=total age adjusted mortality in deaths per 100,000 population;
 - PRECIP=mean annual precipitation (in inches);
 - EDUC=median number of school years completed for persons 25 and older;
 - NONWHITE=percentage of 1960 population that is nonwhite;
 - NOX=relative pollution potential of No_x (related to amount of tons of No_x emitted per day per square kilometer);
 - SO2=relative pollution potential of SO_2

Based on the previous study we will use PRECIP, EDUC, NONWHITE, Log(NOX) and SO2 to predict MORT.

a) We will use stepwise selection guided by the effect tests to add (or delete) predictors into the model.

b) Under Analyze > Fit Model >
MORT > Y

Add PRECIP, EDUC, NONWHITE, Log(NOX) and
SO2 into

Construct Model Effects

c) Choose Stepwise under Personality > Run Model.

We will check (or uncheck) each variable according to the “F-ratio” statistics. The final model is chosen based on R-squares and the p-values. Usually, only variables which are significant should stay in the final model.

Here are the steps:

Stepwise Fit

Response:
MORT

Stepwise Regression Control

Prob to Enter 0.250
Prob to Leave 0.100

Direction:

Current Estimates

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
76259.74	55	1386.5	0.6659	0.6416	6.685792	438.8534

Entered

Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
Intercept	999.316169	1	0	0.000	1.0000
PRECIP	1.61112351	1	8965.8	6.466	0.0138
EDUC	-15.773367	1	7056.097	5.089	0.0281
NONWHITE	3.06092577	1	34458.46	24.852	0.0000
Log(NOX)	.	1	3613.212	2.686	0.1071
SO2	0.3271823	1	21102.28	15.219	0.0003

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	NONWHITE	Entered	0.0000	94595.56	0.4144	43.366	2
2	EDUC	Entered	0.0000	33848.33	0.5627	20.206	3
3	SO2	Entered	0.0030	14603.66	0.6267	11.35	4
4	PRECIP	Entered	0.0138	8965.8	0.6659	6.6858	5

d) Our final model is

Summary of Fit

RSquare 0.665928

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	152013.34	38003.3	27.4087
Error	55	76259.74	1386.5	Prob > F
C. Total	59	228273.08		<.0001

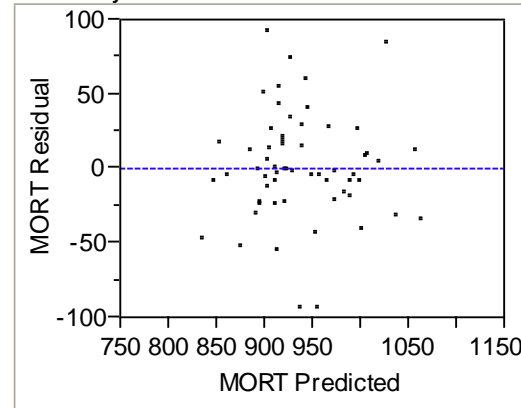
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	999.31617	92.07861	10.85	<.0001
PRECIP	1.6111235	0.633579	2.54	0.0138
EDUC	-15.77337	6.992113	-2.26	0.0281
NONWHITE	3.0609258	0.614004	4.99	<.0001
SO2	0.3271823	0.083867	3.90	0.0003

Effect Tests

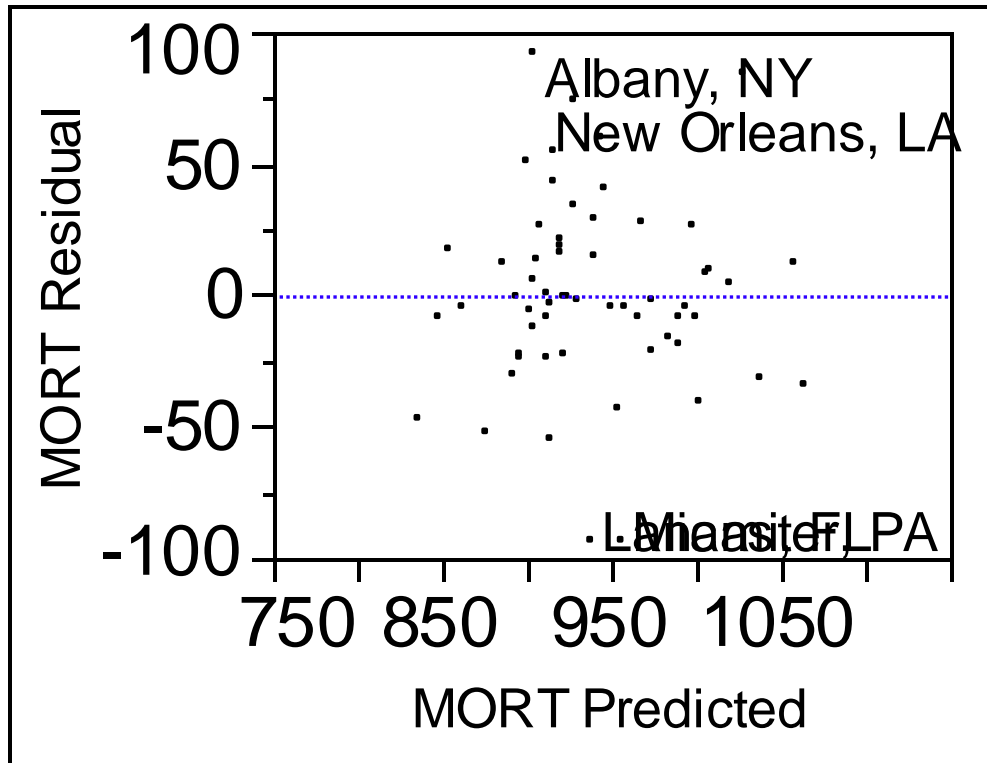
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
PRECIP	1	1	8965.800	6.4663	0.0138
EDUC	1	1	7056.097	5.0890	0.0281
NONWHITE	1	1	34458.462	24.8521	<.0001
SO2	1	1	21102.285	15.2194	0.0003

Residual by Predicted Plot



Outliers in Multiple Regression

- Outliers in terms of multiple regression:
Observations with large residuals.
- If residuals come from normal distribution, then a residual with absolute value larger than about $2.6s_e$ is expected only 1% of the time.
- Investigate observations with residuals of large magnitude.



- Residual plot of MORT vs. PRECIP, EDUC, NONWHITE and SO2
- Four places shown on the plot show some large residuals.
- Notice that residual plots for multiple regression are using residuals vs. predicted values.

Leverage in Multiple Regression

- In a simple regression a point has high leverage if it is an outlier in X .
- In a multiple regression
 - We will identify leverage points for each predictor.
 - We use leverage plots to identify high leverage and influential points for each regression coefficient.
- High leverage observations for a certain x -variable may affect the estimated value of that coefficient.

Leverage Plots

- A “simple regression view” of a multiple regression coefficient. For x_j :
Residual y (w/o x_j) vs. Residual x_j (vs the rest of x 's)
(both axes are recentered at their means)
 - Slope is the
Coefficient for that variable in the multiple regression
 - The p-value: same as the effect test p-value
 - Distances from the points to the LS line are multiple regression residuals.
 - Useful to identify (relative to x_j)
outliers
leverage
influential points
- (Use them the same way as in a simple regression.)

Pollution data: the final model is

Summary of Fit

RSquare 0.665928

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	152013.34	38003.3	27.4087
Error	55	76259.74	1386.5	Prob > F
C. Total	59	228273.08		<.0001

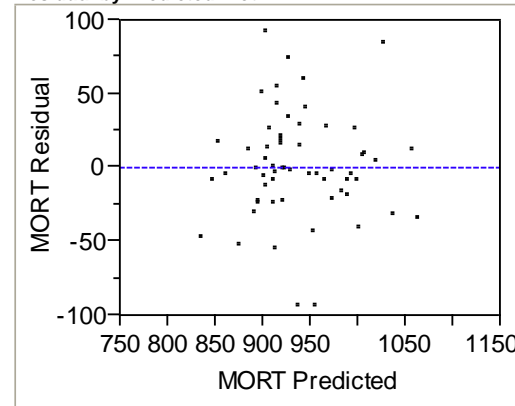
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	999.31617	92.07861	10.85	<.0001
PRECIP	1.6111235	0.633579	2.54	0.0138
EDUC	-15.77337	6.992113	-2.26	0.0281
NONWHITE	3.0609258	0.614004	4.99	<.0001
SO2	0.3271823	0.083867	3.90	0.0003

Effect Tests

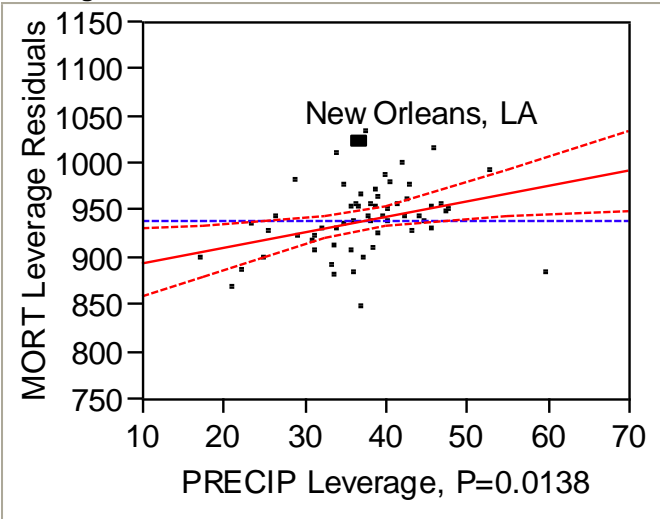
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
PRECIP	1	1	8965.800	6.4663	0.0138
EDUC	1	1	7056.097	5.0890	0.0281
NONWHITE	1	1	34458.462	24.8521	<.0001
SO2	1	1	21102.285	15.2194	0.0003

Residual by Predicted Plot

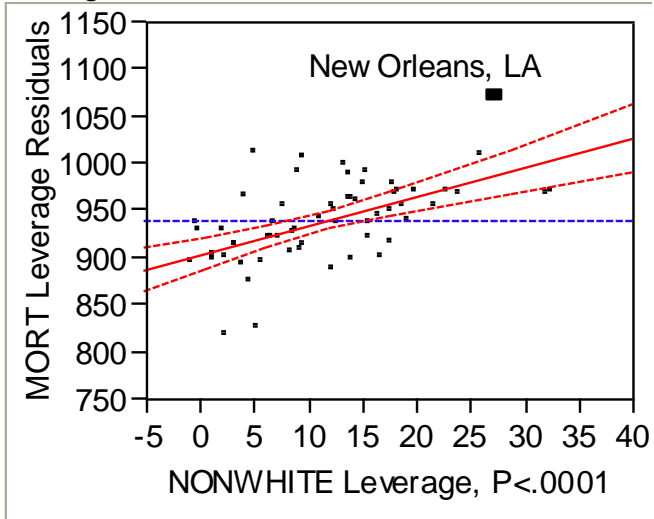


Leverage plots:

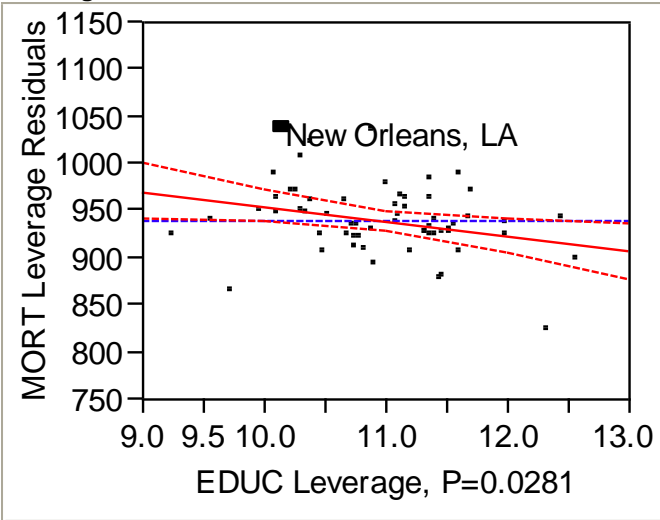
PRECIP
Leverage Plot



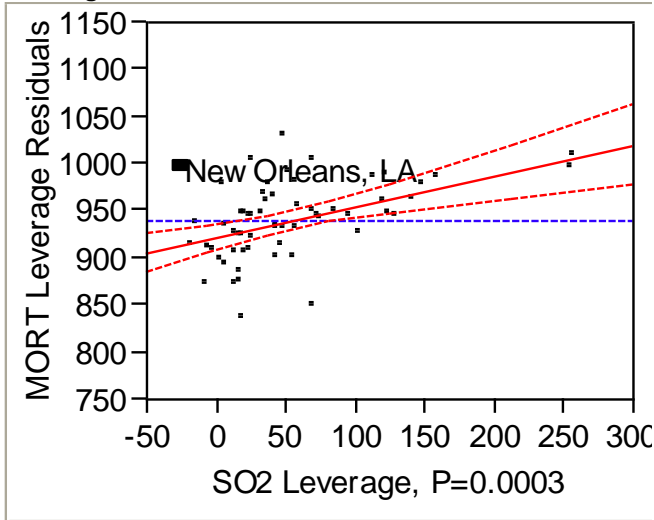
NONWHITE
Leverage Plot



EDUC
Leverage Plot



SO2
Leverage Plot



**Whole Model
Actual by Predicted Plot**

Summary of Fit

RSquare	0.665928
RSquare Adj	0.641631
Root Mean Square Error	37.23628
Mean of Response	940.3568
Observations (or Sum Wgts)	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	152013.34	38003.3	27.4087
Error	55	76259.74	1386.5	Prob > F
C. Total	59	228273.08		<.0001

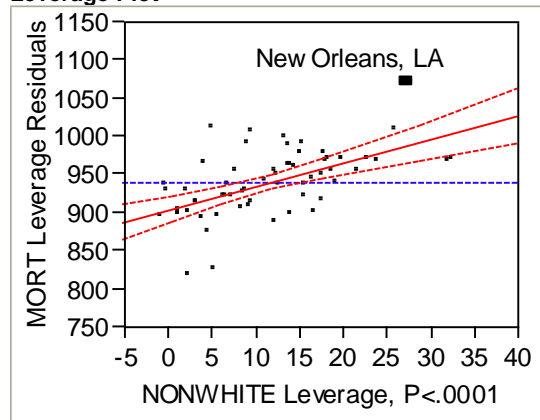
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	999.31617	92.07861	10.85	<.0001
PRECIP	1.6111235	0.633579	2.54	0.0138
EDUC	-15.77337	6.992113	-2.26	0.0281
NONWHITE	3.0609258	0.614004	4.99	<.0001
SO2	0.3271823	0.083867	3.90	0.0003

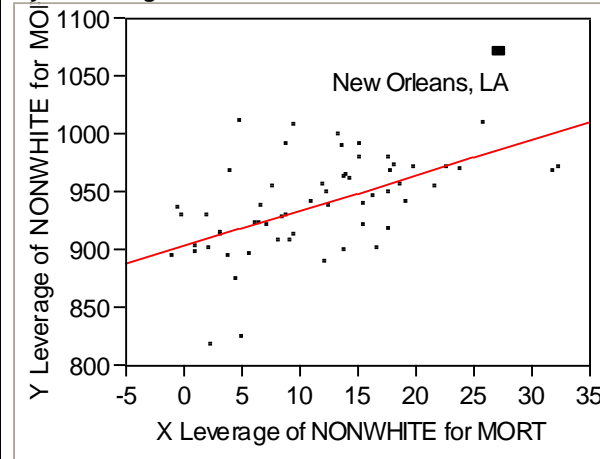
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
PRECIP	1	1	8965.800	6.4663	0.0138
EDUC	1	1	7056.097	5.0890	0.0281
NONWHITE	1	1	34458.462	24.8521	<.0001
SO2	1	1	21102.285	15.2194	0.0003

Leverage Plot



**Bivariate Fit of Y Leverage of NONWHITE for MORT
By X Leverage of NONWH**



Linear Fit

Y Leverage of NONWHITE for MORT =
904.02358 + 3.0609258 X Leverage of NONWHITE for MORT

Summary of Fit

RSquare	0.311227
RSquare Adj	0.299351
Root Mean Square Error	36.26049
Mean of Response	940.3568
Observations (or Sum Wgts)	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	34458.46	34458.5	26.2077
Error	58	76259.74	1314.8	Prob > F
C. Total	59	110718.20		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio
Intercept	904.02358	8.502028	106.33
X Leverage of NONWHITE for MORT	3.0609258	0.597914	5.12

•The output from the whole model fit is **on the left** together with the Leverage plot for NONWHITE

•We can reproduce the leverage plot by

Analyze > Fit Model > Save Columns > Effect Leverage Pairs.

Then fit Y leverage to X leverage in a simple regression , **shown on the right.**

• Notice the coefficients for NONWHITE are the same from both outputs.

Interpretation of Leverage Plots

- The enlarged observation New Orleans is a moderate outlier and it is somewhat leveraged for estimating the coefficient of both SO_2 and NONWHITE and possibly of EDUC. Since New Orleans is both moderately highly leveraged and an outlier, we suspect that it might be influential.

Whole Model
Actual by Predicted Plot

Summary of Fit

RSquare	0.665928
RSquare Adj	0.641631
Root Mean Square Error	37.23628
Mean of Response	940.3568
Observations (or Sum Wgts)	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	152013.34	38003.3	27.4087
Error	55	76259.74	1386.5	Prob > F
C. Total	59	228273.08		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	999.31617	92.07861	10.85	<.0001
PRECIP	1.6111235	0.633579	2.54	0.0138
EDUC	-15.77337	6.992113	-2.26	0.0281
NONWHITE	3.0609258	0.614004	4.99	<.0001
SO2	0.3271823	0.083867	3.90	0.0003

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
PRECIP	1	1	8965.800	6.4663	0.0138
EDUC	1	1	7056.097	5.0890	0.0281
NONWHITE	1	1	34458.462	24.8521	<.0001
SO2	1	1	21102.285	15.2194	0.0003

Whole Model
Actual by Predicted Plot

Summary of Fit

RSquare	0.656139
RSquare Adj	0.630668
Root Mean Square Error	35.50292
Mean of Response	937.4297
Observations (or Sum Wgts)	59

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	129877.83	32469.5	25.7601
Error	54	68064.71	1260.5	Prob > F
C. Total	58	197942.54		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	967.76738	88.65992	10.92	<.0001
PRECIP	1.6309136	0.604135	2.70	0.0093
EDUC	-12.8732	6.762958	-1.90	0.0623
NONWHITE	2.6542287	0.606761	4.37	<.0001
SO2	0.3675916	0.081518	4.51	<.0001

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
PRECIP	1	1	9185.898	7.2877	0.0093
EDUC	1	1	4566.969	3.6233	0.0623
NONWHITE	1	1	24119.568	19.1356	<.0001
SO2	1	1	25630.014	20.3339	<.0001

- The output on the right is for the data without New Orleans
- Removing New Orleans has some impact on the coefficients for SO2, EDUC and NONWHITE. The difference is quite noticeable for the EDUC coefficient, less so for the others. We will stop here for this model.

The influential points can have extreme impact on the analysis

- We might have used an alternative model
- Because of the importance of NOX and SO2, we might have chosen the final model to be:
MORT vs. PRECIP, NONWHITE, EDUC and log Nox and log SO2
- Notice that log Nox is not significant. One could still leave it in the model so that we can better see whether it has an effect.

Whole Model Summary of Fit

RSquare 0.688278

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	157115.28	31423.1	23.8462
Error	54	71157.80	1317.7	Prob > F
C. Total	59	228273.08		<.0001

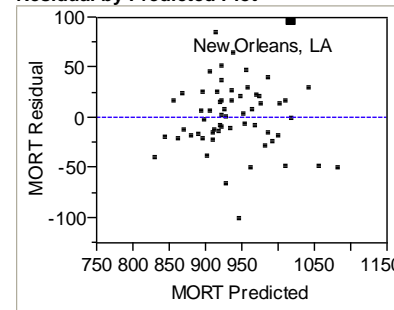
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	940.6541	94.05424	10.00	<.0001
PRECIP	1.9467286	0.700696	2.78	0.0075
EDUC	-14.66406	6.937846	-2.11	0.0392
NONWHITE	3.028953	0.668519	4.53	<.0001
Log(NOX)	6.7159712	7.39895	0.91	0.3681
Log(SO2)	11.35814	5.295487	2.14	0.0365

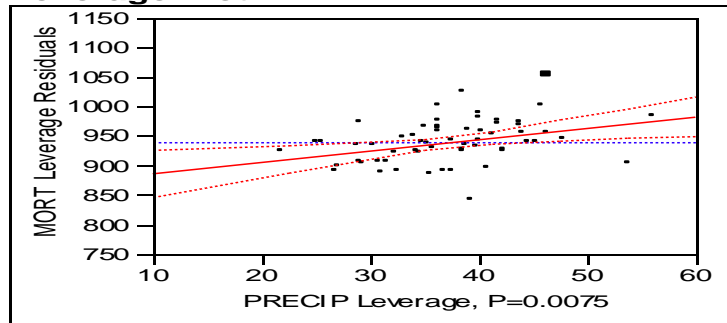
Effect Tests

Source	Sum of Squares	F Ratio	Prob > F
PRECIP	10171.388	7.7188	0.0075
EDUC	5886.913	4.4674	0.0392
NONWHITE	27051.227	20.5285	<.0001
Log(NOX)	1085.691	0.8239	0.3681
Log(SO2)	6062.217	4.6005	0.0365

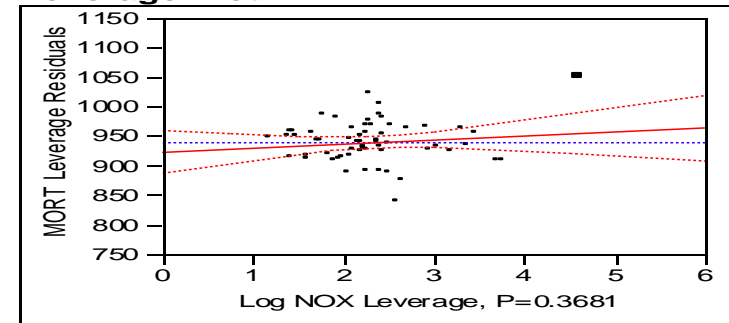
Residual by Predicted Plot



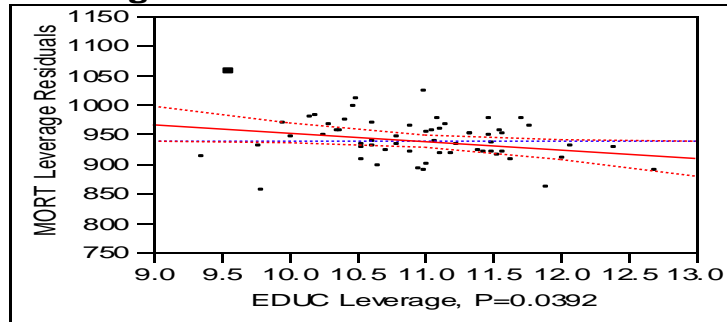
**PRECIP
Leverage Plot**



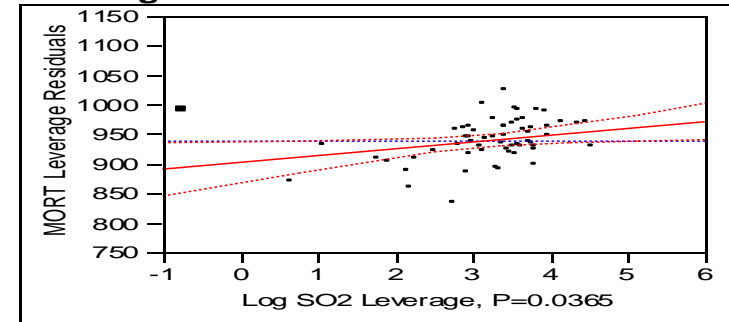
**Log NOX
Leverage Plot**



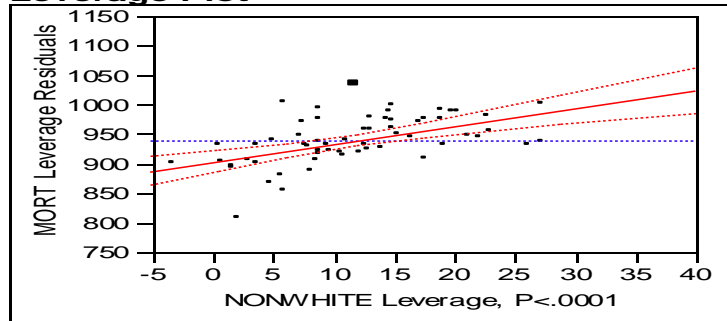
**EDUC
Leverage Plot**



**Log SO2
Leverage Plot**



**NONWHITE
Leverage Plot**



The enlarged observation New Orleans is an outlier for estimating each coefficient and is highly leveraged for estimating the coefficients of interest on log Nox and log SO2. Since New Orleans is both highly leveraged and an outlier, we expect it to be influential.

Multiple Regression with New Orleans

Summary of Fit

RSquare	0.688278
RSquare Adj	0.659415
Root Mean Square Error	36.30065
Mean of Response	940.3568
Observations (or Sum Wgts)	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	157115.28	31423.1	23.8462
Error	54	71157.80	1317.7	Prob > F
C. Total	59	228273.08		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	940.6541	94.05424	10.00	<.0001
PRECIP	1.9467286	0.700696	2.78	0.0075
EDUC	-14.66406	6.937846	-2.11	0.0392
NONWHITE	3.028953	0.668519	4.53	<.0001
Log NOX	6.7159712	7.39895	0.91	0.3681
Log SO2	11.35814	5.295487	2.14	0.0365

Multiple Regression without New Orleans

Summary of Fit

RSquare	0.724661
RSquare Adj	0.698686
Root Mean Square Error	32.06752
Mean of Response	937.4297
Observations (or Sum Wgts)	59

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	143441.28	28688.3	27.8980
Error	53	54501.26	1028.3	Prob > F
C. Total	58	197942.54		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	852.3761	85.9328	9.92	<.0001
PRECIP	1.3633298	0.635732	2.14	0.0366
EDUC	-5.666948	6.52378	-0.87	0.3889
NONWHITE	3.0396794	0.590566	5.15	<.0001
Log NOX	-9.898442	7.730645	-1.28	0.2060
Log SO2	26.032584	5.931083	4.39	<.0001

Removing New Orleans has a large impact on the coefficients of log NOX , EDUC and log SO2, in particular, it reverses the sign of log NOX.