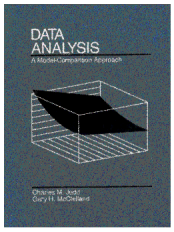
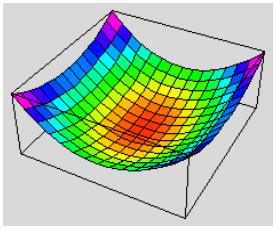


Handout from Psych 5741/5751
University of Colorado
used with
Judd, C.M., & McClelland, G.H. (1989).
Data Analysis: A Model Comparison Approach. HBJ.



Brief Lecture Notes
Chapter 9: Outliers

Until now, DATA have been well-behaved
In Chapt 16 we will deal with ill-behaved data with heterogeneous variances, non-normal distributions, etc. Here:

We noted in Chapt 2 that SSE and estimators which minimize SSE are very sensitive to outliers or wild observations. We had best make sure we don't have any outliers. With outliers, regression estimates can be *very* misleading.

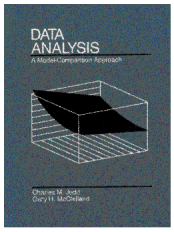
Outliers are extreme observations that for one reason or another do not belong with the other observations in DATA. (vague!)

Why they are a problem:

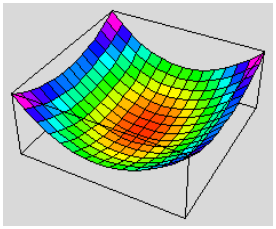
bias or "grab" parameter estimates
inflate SSE, thereby making it difficult to detect reductions in SSE due to other factors
often not obvious that this has happened

example from Chapter 2:

1 3 5 9 14 mean = 6.4, MSE=s^2 = 26.8 [0, 12.8]
1 3 5 9 140 mean=31.6, MSE=s^2=3680.8
[-43.7,106.9]



Handout from Psych 5741/5751
University of Colorado
used with
Judd, C.M., & McClelland, G.H. (1989).
Data Analysis: A Model Comparison Approach. HBJ.



Causes:

1. "Klinkers" (Abelson) data recording or data entry errors. use of computers make these more likely. Lou's energy study example. Should always be fixed. Need computers to help look for them.

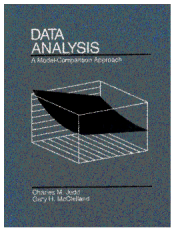
2. Two kinds of cases. (or errors from two bags of error tickets) Math score example from Ex 5.2, p. 74 (typo, book says Ex 4.2). Outliers can provide clues to better MODELs. Need techniques for finding outliers so they can be examined with great care.

3. Thick tails of error distributions. Robust to non-normality but not thick tails, Ex. 9.1, p. 210. With thick tails, extreme observations occur more frequently than they should.

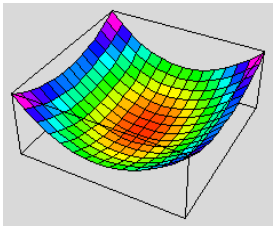
What to do about outliers?
CONTROVERSIAL!

ignoring them is never acceptable. to do nothing is the equivalent of making a decision about their appropriateness in the analysis. will often end up with MODEL that describes essentially none of the DATA—neither the outliers nor the bulk of the DATA

report MODEL with and without outliers included
do analysis to see if outliers significantly different from others in MODEL



Handout from Psych 5741/5751
University of Colorado
used with
Judd, C.M., & McClelland, G.H. (1989).
Data Analysis: A Model Comparison Approach. HBJ.



Examples:
do Abelson example (if not done earlier)

HSRANK and SAT example

Outlier Questions:

1. Is X (or the set of predictors) unusual?

2. Is Y unusual (relative to MODEL of other DATA)?

3. Does Y have a big impact on predictions of other Y's? I.e., does it have big impact on parameter estimates?

Do questions in order:

Is X (or the set of X predictors) unusual?

leverage
illustrate with X-Y graph

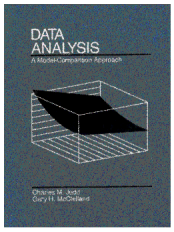
we usually write:

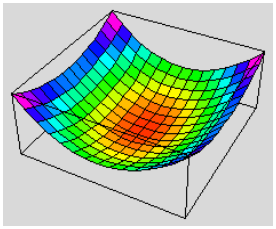
$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots$$

but alternatively and equivalently, could write

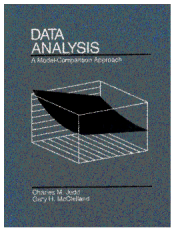
$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

(Note: this is really why called linear model)
The h's are entirely determined by the X's. If we know X's, we can compute h's even before we collect

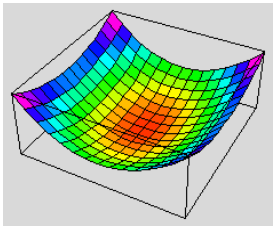




the DATA! Equation separates info about X from info about DATA.



Handout from Psych 5741/5751
University of Colorado
used with
Judd, C.M., & McClelland, G.H. (1989).
Data Analysis: A Model Comparison Approach. HBJ.



LEVERAGE is how much an observation influences its own prediction. LEVER = h_{ii}.

For mean LEVER = $h_{ii} = 1/n$

For simple regression:

$$h_{ij} = \frac{1}{n} + \frac{(X_{i1} - \bar{X}_1)(X_{j1} - \bar{X}_1)}{SSX}$$

So LEVER =

$$h_{ii} = \frac{1}{n} + \frac{(X_{i1} - \bar{X}_1)^2}{SSX}$$

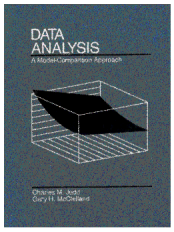
For Multiple Regression with two predictors LEVER =

$$h_{ii} = \frac{1}{n} + \frac{X_{1i}^2 \sum X_2^2 - X_{1i} X_{2i} \sum X_1 X_2}{\sum X_1^2 \sum X_2^2 - \left(\sum X_1 X_2\right)^2} + \frac{X_{2i}^2 \sum X_1^2 - X_{1i} X_{2i} \sum X_1 X_2}{\sum X_1^2 \sum X_2^2 - \left(\sum X_1 X_2\right)^2}$$

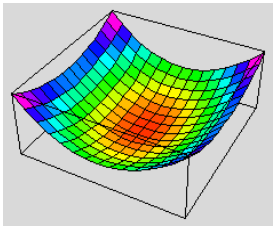
(assuming mean deviation form for both predictors). If there is no redundancy, then this reduces to

$$h_{ii} = \frac{1}{n} + \frac{X_{1i}^2}{\sum X_1^2} + \frac{X_{2i}^2}{\sum X_2^2}$$

(Illustrate with two-sample t-test?)



Handout from Psych 5741/5751
University of Colorado
used with
Judd, C.M., & McClelland, G.H. (1989).
Data Analysis: A Model Comparison Approach. HBJ.



Evaluating LEVERs

$$0 \le h_{ii} \le 1$$

$$\sum_{i=1}^{n} h_{ii} = PA \quad \Leftrightarrow \quad \bar{h}_{ii} = \frac{PA}{n}$$

Tells us how much of a parameter is dedicated to the prediction of a single observation!

1/h "equivalent number of observations" involved in the determination of Yhat.

(e.g., for two-sample t-test, half obs for Yhats from one group and half the obs for Yhats in the other group)

Is Y_i Unusual?

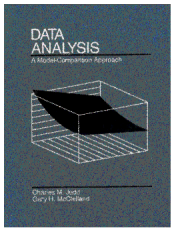
$$e_i = Y_i - \hat{Y}_i$$

Difficult to interpret

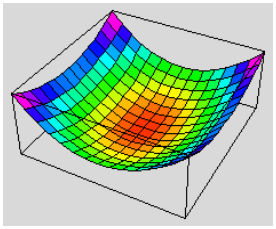
1. **need standardization**
2. **paradox in allowing outlier to determie model**

really want to ask if Y_k is unusual WRT to a MODEL based on all the other observations

such a statistic is the *studentized deleted residual*



Handout from Psych 5741/5751
University of Colorado
used with
Judd, C.M., & McClelland, G.H. (1989).
Data Analysis: A Model Comparison Approach. HBJ.



rationale: Outlier Model

MODEL A: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i \neq k$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 + \varepsilon_i \quad i = k$$

MODEL C: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i$

OR, equivalently,

$X_2 = 1$, if k-th observation, 0 otherwise

A: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$

C: $Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$

Example:
leaving out 6th obs

SAT = 6.71 + .50 HSRANK + 55.49 X[6]
SAT = 96.55 - .50 HSRANK

PRE = .68, F*[1,10] = 21.4, p < .01

----
leaving out 1st obs

SAT = 95.71 - .48 HSRANK - 10.67 X[1]
SAT = 96.55 - .50 HSRANK

PRE = .096, F*[1,10] = 1.06, n.s.
(see Ex 9.6, p. 223)

don't have to do separate regressions

$$F^* = \frac{e_i^2 (n - PA - 1)}{SSE(1 - h_{ii}) - e_i^2}$$

won't do by hand, but just remember that all the information is available from the original regression so they are cheap to get---so look at them!

cutoffs:
-2, +2 deserve a look
-3,+3 require a check
-4,+4 all alarm bells!

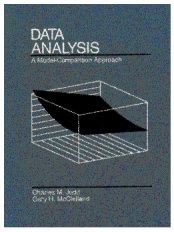
---

3. Does Y_k affect other predictions? (i.e., the parameters?)

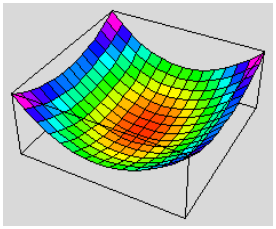
Cook's D

$$D_k = \frac{\sum_i (\hat{Y}_i - \hat{Y}_{i,[k]})^2}{PA(MSE)}$$

Again, everything we need is available so don't have to re-do regression n times, each time leaving out a different variable:



Handout from Psych 5741/5751
University of Colorado
used with
Judd, C.M., & McClelland, G.H. (1989).
Data Analysis: A Model Comparison Approach. HBJ.



$$D_k = \frac{e_k^2}{PA(MSE)}\left[\frac{h_{kk}}{(1-h_{kk})^2}\right]$$

Interaction interpretation:

big effect when both X unusual AND Y unusual
little effect if either X or Y is very usual

cutoffs:

gaps
bigger than 1 or 2

do Anscombe's example

p. 232

partial regression plots