

The plan:

- THIS Week T - Finish MLR Inference  
R - Diagnostics, Remedial Measures, Model Search
- Next Week T - Something on Time Series  
R - Exam on Regression

Confidence Limits for  $\mu_y | x_1, \dots, x_k$

$$\hat{y} \pm t \text{ (SEM)}$$

$$58.704 \pm 2.365 (.6374)$$

A upper 2.5% pt of t  
distn with d.f. = 7

$$58.704 \pm 1.508$$

95% prediction limits for next home  
with  $x_1 = 20$  and  $x_2 = 9$

$$58.704 \pm 2.365 (1.255)$$

$$\pm 2.956$$

$$\sqrt{(1.08)^2 + (.6374)^2}$$

Prediction Limits for  $y_{\text{new}}$

at values  $x_1, x_2, \dots, x_k$

$$\hat{y} \pm t \text{ (SE}_{\hat{y}}) = \sqrt{s^2 + (\text{SEM})^2}$$

Example Real Estate  $x_1 = 20$   
 $x_2 = 9$

$$\hat{y} = 58.704 \quad (y = 57.7)$$

$$\text{SEM} = .6374$$

95% confidence limits for mean  
price of such homes

Exercise For the fake data set  
make 95% confidence limits for  
mean  $y$  when  $x_1 = -2$  and  $x_2 = 0$   
- then make corresponding 95%  
prediction limits - Case #1

$$\hat{y} = 4.4 \quad \text{SEM} = .3464 \quad \text{SE}_{\hat{y}} = .5657$$

(BTW  $s = .4472$  and

$$.5657 = \sqrt{(.4472)^2 + (.3464)^2}$$

Confidence Limits for Mean  $y$ :

$$4.4 \pm (9.303) (.3464)$$

Prediction Limits for new  $y$

$$4.4 \pm (4.303)(.5657)$$

Testing in MLR

t tests for single coefficients I can

test  $H_0: \beta_j = \#$  using

$$T = \frac{b_j - \#}{SE_{b_j}}$$

use  $df = n - k - 1$   
to get  
p-value

most common version is version with  $\# = 0$

Both  $H_0: \beta_1 = 0$  and  $H_0: \beta_2 = 0$  are implausible, both  $x_1$  and  $x_2$  are helpful in predicting/explaining  $y$

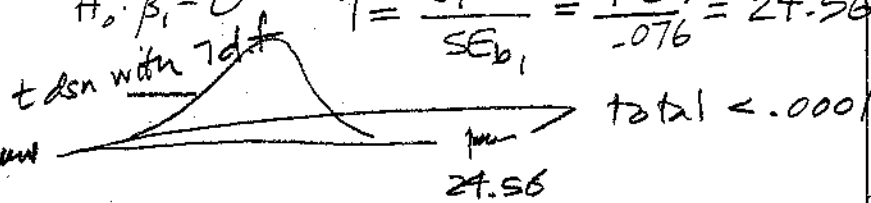
Notice that these tests are not (either numerically or conceptually) equivalent to the tests of  $H_0: \text{slope} = 0$  in SLR

Testing  $H_0: \beta_1 = 0$  in SLR  $y = \beta_0 + \beta_1 x_1 + \epsilon$  asks "if all I have to predict with is  $x_1$ , can I do without it?"

JMP gives p-values for 2-sided tests of  $H_0: \beta_j = 0$

Example Real Estate

$$H_0: \beta_1 = 0 \quad T = \frac{b_1 - 0}{SE_{b_1}} = \frac{1.87}{.076} = 24.56$$



$$H_0: \beta_2 = 0 \quad T = \frac{b_2 - 0}{SE_{b_2}} = \frac{1.28}{.144} = 8.85$$

p-value < .0001

Testing  $H_0: \beta_1 = 0$  in MLR

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

asks "if I have both  $x_1$  and  $x_2$  to predict with, can I do without  $x_1$ ?"

(Note that if  $x_1, x_2$  are themselves highly correlated, the answer to 1st question might be NO! while the answer to the second is YES! ----) (see a later discussion of "multicollinearity")

There are also a variety of possible F tests in MLR

"Overall F test" / Model Utility Test

This is a test of  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  based on an overall ANOVA table

Model =  $y_j | x_1, \dots, x_k = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$

Hypothesis =  $y_j | x_1, \dots, x_k = \beta_0$   
 so this is often interpreted a test of whether someplace in  $x_1, x_2, \dots, x_k$  there is predicting power

p-values will come from the F tables with d.f. k and n-k-1

Example Real Estate

ANOVA (for testing  $H_0: \beta_1 = \beta_2 = 0$ )

Source	SS	df	MS	F
Regression ( $x_1, x_2$ )	819.3	2	409.7	350.9
Error	8.2	7	1.2	X
Total	827.5	9	X	X

tiny p-value (consulting F dsb with df = 2, 7)

This is a way of attaching a p-value to  $R^2 = \frac{SSR}{SSTot}$  for MLR

ANOVA table (for MLR)

Source	SS	df	MS	F
Regression	SSR	k	MSR = $\frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	SSE	n-k-1	MSE = $\frac{SSE}{n-k-1}$	X
Total	SSTot	n-1	X	X

from MLR  $S^2$   
 big observed F count as evidence against  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$

Caution: In  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

testing  $\left\{ \begin{matrix} H_0: \beta_1 = 0 \\ H_0: \beta_2 = 0 \end{matrix} \right\}$  separately is NOT

the same thing as testing  $H_0: \beta_1 = \beta_2 = 0$

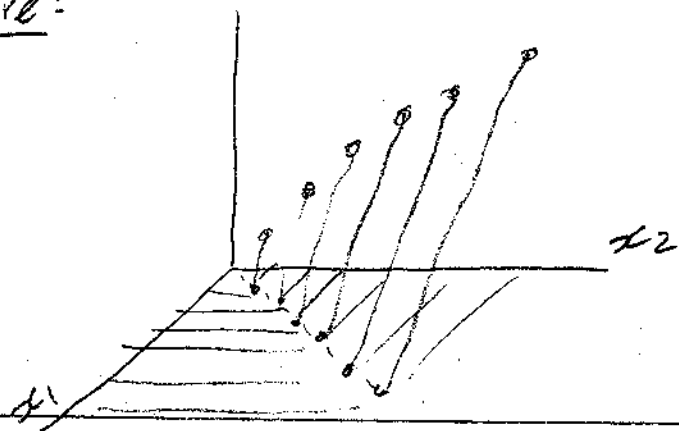
It's possible to get  
 tiny p-value for  $H_0: \beta_1 = \beta_2 = 0$   
 and simultaneously  
 large p-value for  $H_0: \beta_1 = 0$   
 large p-value for  $H_0: \beta_2 = 0$

This kind of thing can happen when predictors are correlated... That brings up the fact that interpretation of MLR is subtle when predictors are correlated — jargon for this circumstance is "multicollinearity"

Example  $x_1, x_2, y$   
 note  $x_2 \approx 2x_1$        $y \approx 1 + 3x_1$   
                                   $\approx 1 + 1.5x_2$

Moral: When  $x$ 's are highly correlated one must be very careful about interpretation of  $b$ 's and tests

Picture:



SLR  $x_1$

$$\hat{y} = 1.194 + 2.896x_1$$

$$R^2 = .9972$$

$$T = 33$$

for testing  $H_0: \beta_1 = 0$

MLR  $x_1, x_2$

$$\hat{y} = 1.03 + 1.239x_1 + .864x_2$$

$$R^2 = .9989$$

$$t\text{'s } 1.26, 1.68$$

SLR  $x_2$

$$\hat{y} = .923 + 1.508x_2$$

$$R^2 = .9979$$

$$T = 38.4$$

for testing  $H_0: \beta_2 = 0$

when  $x_1, x_2$  are highly correlated I can't really separate their effects on  $y$  (and need to be cautious in interpreting  $b$ 's) — because  $x$ 's are highly correlated, while I can pick out a line in 3 dimensions where  $(x_1, x_2, y)$  fall (I can predict  $y$  from  $x_1$  or  $x_2$  or  $(x_1, x_2)$ ) I can't really pick out a plane i.e. I can't separate the effects of  $x_1, x_2$  ----

## Exercise For Fake data

- Find T statistics for testing  $H_0: \beta_1 = 0$   $H_0: \beta_2 = 0$  on printout
- Take previous hand work and make up ANOVA table and overall F for testing  $H_0: \beta_1 = \beta_2 = 0$

ANOVA Table

Source	SS	df	MS	F
Regression	15.6	2	7.8	39.0
Error	.4	2	.2	X
Total	16	4	X	X

This is a way of asking "Is reduced model adequate, or is there something in the Full model that is needed in addition to what's in the reduced model?"

Notice

$$SSR_{full} \geq SSR_{reduced}$$

$$R^2_{full} \geq R^2_{reduced}$$

and a "partial F" test is a way of attaching a p-value to the increase in  $SSR$  (or in  $R^2$ ) — usually this is organized in an expanded ANOVA Table

There are other F tests associated with Multiple Linear Regression — "Partial F Tests" — MMD+S do something equivalent in terms of comparing  $R^2$  values — (see their page 661)

These are a way of comparing

$$\text{Full Model: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

$$\text{Reduced Model: } y = \beta_0 + \beta_1 x_1 + \dots + \beta_l x_l + \epsilon$$

for  $l < k$

Source	SS	df	MS	F
Regression	$SSR_{full}$	k		
$x_1, \dots, x_l$	$SSR_{red}$	l		
$x_{l+1}, \dots, x_k   x_1, \dots, x_l$	$SSR_{full} - SSR_{red}$	$k - l$	$\frac{SSR_{full} - SSR_{red}}{k - l}$	$\frac{SSR_{full} - SSR_{red}}{k - l} \cdot \frac{MSE_{full}}{MSE_{full}}$
Error	$SSE_{full}$	$n - k - 1$	$MSE_{full}$	
Total	$SST_{tot}$	$n - 1$		

E.G.  $k = 5, l = 2$

$$\text{Full Model: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

$$\text{Reduced Model: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\text{Tests } H_0: \beta_3 = \beta_4 = \beta_5 = 0$$

### Example Real Estate

$$R^2_{full} = .99 \quad R^2_{SLR} = .88$$

on size

I might want to ask whether this increase in  $R^2$  is "statistically significant"

We've already used a t-test of  $H_0: \beta_2 = 0$  but I could also use a partial F test (when the Full model has one more  $x$  than the Reduced model, the Partial F and t are equivalent)

Note that here (since the full model has 1 more predictor variable than the reduced one) the F statistic is the square of the t statistic for the coefficient potentially being dropped ( $\beta_2$ )

$$78.0 = (8.85)^2$$

Exercise Make the expanded ANOVA table for  $H_0: \beta_1 = 0$  in Real Estate example

(SLR on "condition" gives  $SSR = 115$ )

Expanded ANOVA Table ( $H_0: \beta_2 = 0$ )

Source	SS	df	MS	F
Regression ( $x_1, x_2$ )	819.3	2		
$x_1$	127.9	1		
$x_2   x_1$	91.4	1	91.4	78.0
Error	8.2	7	1.2	
Total	827.5	9	X	

Annotations:  
 -  $SSR_{full}$  points to 819.3  
 -  $SSR_{SLR \text{ with } x_1}$  points to 127.9  
 - "The difference" points to 91.4  
 -  $SSE_{full}$  points to 8.2  
 - "compare to table 17 d.f." points to the F column

### ANOVA Table

Source	SS	df	MS	F
Regression ( $x_1, x_2$ )	819.5	2		
$x_2$	115	1		
$x_1   x_2$	704.3	1	704.3	603
Error	8.2	7	"1.2"	
Total	827.5	9		

JMP

Tomte: Vardeman's Version of Sect 11.3

"Model Building"

1. Making "new" predictors (and responses)
  - "interaction"
  - incorporating qualitative info
2. Searching
  - "algorithm(s)"
  - criteria
3. Model "checking" / Diagnosis
  - plots
  - "statistics" / measures

I can also take

$$\begin{matrix} x_1 \\ x_2 \end{matrix} \longrightarrow x'$$

i.e. make up a new predictor from several existing ones - This brings up a piece of terminology - "interaction"

if I start with  $k$  predictors and make up new ones but never use more than 1 at a time, then the response is said to be "additive" in the predictors / original predictors "don't interact" - this gives "parallel response traces"

1. Making "New"  $x$ 's +  $y$ 's

$y \rightarrow y'$  (like on SLR homework)  
 model  $y'$ , do inferences and then  $y' \rightarrow y$  and state results in the original units - This possibility makes MLR more flexible and widely applicable

$x \rightarrow x'$  gives me things like on SLR homework and gives me things like fitting  

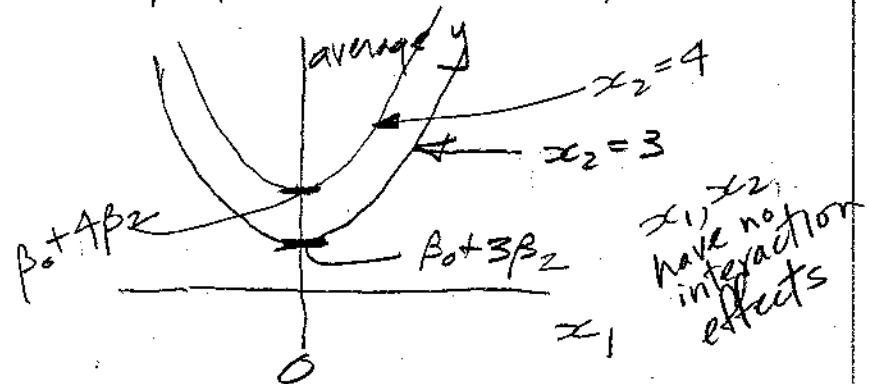
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Example

$$\underbrace{x_1, x_2, y}_{k=2}$$

$$x_1 \rightarrow x'$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \epsilon$$



## Example Real Estate



An alternative model would be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

How to incorporate "qualitative" information in a MLR model? Use "dummy variables" / "indicator variables"

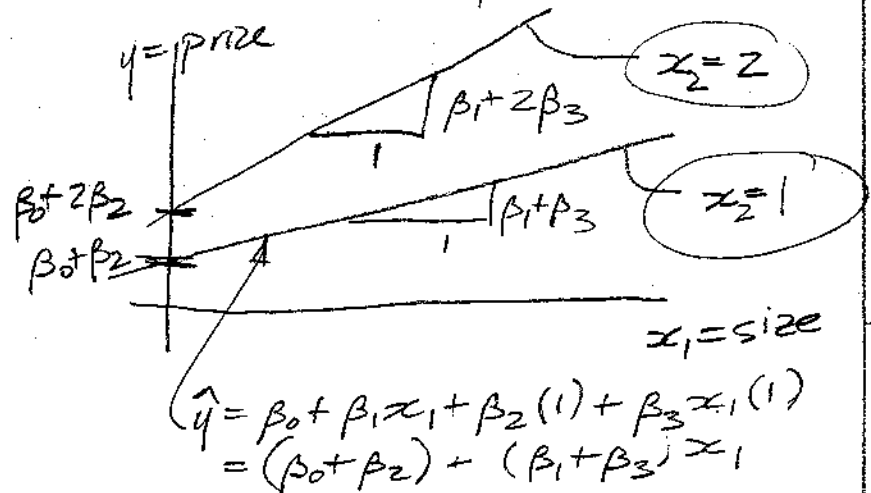
### Hypothetical Real Estate Scenario

$y = \text{price}$   
 $x_1 = \text{size}$   
 $x_2 = \text{condition}$

#2	#1
NW	NE
#3 SW	#4 SE

I'd like build "region" information into a MLR model

This leads to a plot "with interactions" / non-parallel lines



a naive attempt to use region info might be to invent

$x_3 = \text{region \#}$

and model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

but that is silly... it says that I get  $\beta_3$  increase in price

region #1  $\rightarrow$  region #2  
 region #2  $\rightarrow$  region #3  
 region #3  $\rightarrow$  region #4

I need to be more clever

"More clever" is using "Dummy Variables" — If a qualitative factor A has I levels 1, 2, ..., I (A = region I = 4) I define I-1 dummies

$$x_{A1} = \begin{cases} 1 & \text{if observation is from level 1 of A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{A2} = \begin{cases} 1 & \text{if observation is from level 2 of A} \\ 0 & \text{otherwise} \end{cases}$$

⋮  
 $x_{AI-1}$

y	$x_1$	$x_2$	$x_{A1}$	$x_{A2}$	$x_{A3}$	
—	—	—	1	0	0	NE homes
⋮	⋮	⋮	⋮	⋮	⋮	
—	—	—	1	0	0	NW homes
⋮	⋮	⋮	⋮	⋮	⋮	
—	—	—	0	1	0	SW homes
⋮	⋮	⋮	⋮	⋮	⋮	
—	—	—	0	0	1	SE homes
⋮	⋮	⋮	⋮	⋮	⋮	
—	—	—	0	0	0	SE homes
⋮	⋮	⋮	⋮	⋮	⋮	

a MLR involving  $x_{A1}, x_{A2}, \dots, x_{AI}$  will allow for a shift for each different level of A

Example Hypothetical Real Estate



I = 4 quadrants  
 define I-1 = 3  
 dummies

What does this do for me?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{A1} + \beta_4 x_{A2} + \beta_5 x_{A3} + \epsilon$$

Says that

NE homes are modeled

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 + \epsilon$$

$$= (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

NW homes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 + \epsilon$$

$$= (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

## SW homes

$$y = (\beta_0 + \beta_5) + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

## SE homes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

SE is a baseline and coefficients on dummy variables account for level I shifts up or down in price as we look at other quadrants

note as we've got it thus far, price / sq ft is the same in each region, if I want it

Model Searching (Mucking around in the possibilities looking for something that's plausible)

Qualitatively: one wants a model that reproduces  $y$ 's faithfully and is

small/simple

for simple understanding and avoiding "overfitting" (where because it wiggles a lot an equation describes data in hand, but does a bad job for interpolating or extrapolating)

change region to region, I could make up new predictors

$$x_1, x_{A1}$$

$$x_1, x_{A2}$$

$$x_1, x_{A3}$$

This allows price per square foot to change — this would be called "interaction" between region and size

JMP does something like this automatically (coding 1, 0, -1 rather than 1, 0) for "nominal" variables

Criteria (numerical) for comparing models?

$$\begin{cases} R^2 & \text{(generally we like big } R^2 \text{)} \\ S & \text{(generally we like small } S \text{)} \end{cases}$$

(there are cases where  $R_1^2 > R_2^2$  but  $S_2 < S_1$ )

There are other possibilities

Mallovs Cp

If I have  $k$  predictors possible

$$C_p = \frac{SSE_{red}}{MSE_{full}} + 2p - n$$

for a reduced model with  $p-1$  predictors, this should be about  $p$  (skip the rationale for why this is a happy circumstance)

AIC

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2(p+1)$$

(for a model with  $p$  predictors)  
small AIC are desirable

JMP will do this under "stepwise" "personality" for "Fit Model" . . . .

this gives a way to screen a huge # of possible models down a few to examine carefully (to see whether the normal MLR model is appropriate . . . we care because all of our inference formulas hang on the appropriateness of that model)

Model Checking / Diagnostics

Basic Tool here is the idea of residuals

## "Search Algorithms"

Ancient History:

Forward Selection

Backward Elimination

These produce lists of candidate sets of predictors . . . at any given step they are not guaranteed to give best  $R^2$  for a model of that size

Obsolete because there is an "all possible  $R^2$ 's" algorithm -

Recall

$$e_i = y_i - \hat{y}_i = \text{residual for } i\text{th case}$$

These are empirical approximations for

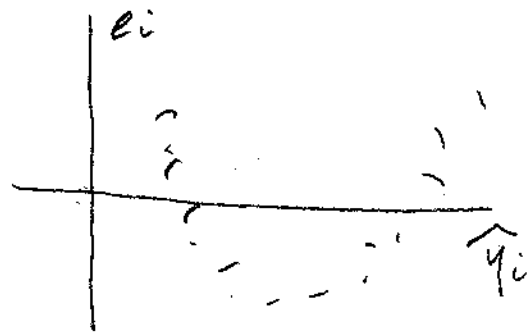
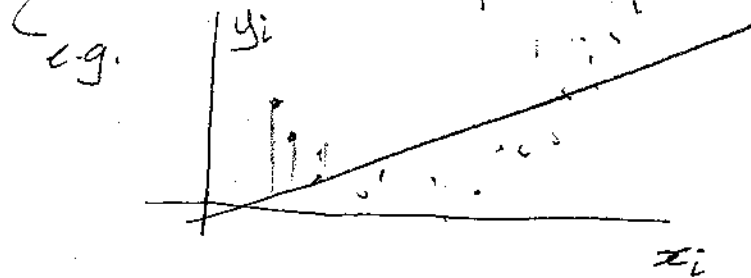
$$y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}) = \epsilon_i$$

and the model says these are random draws from a normal dsu with mean 0 and std dev  $\sigma$  -

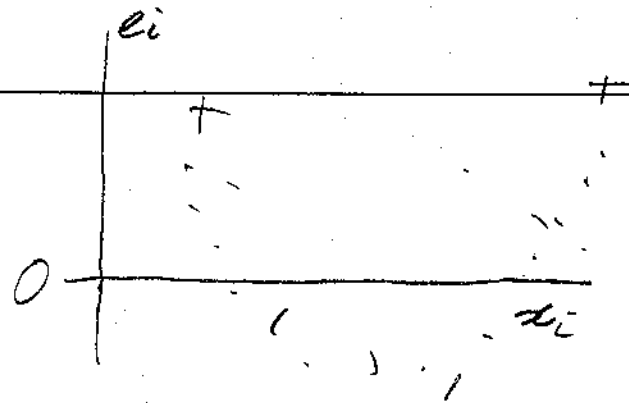
so one would hope that residuals look like "noise" i.e. random variation

i.e. I expect/hope (if MLR model is sensible) that  $e_i$

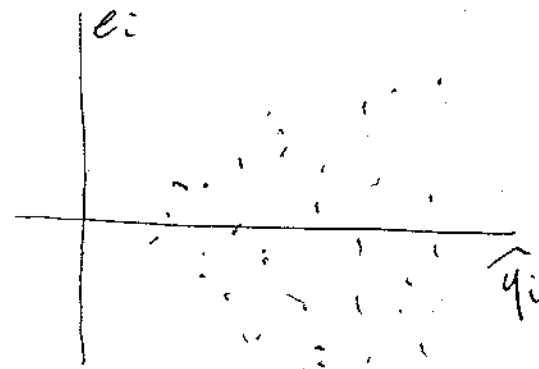
- 1) have a bell-shaped histogram
- 2) are "patternless" when plotted against any sensible variable



This suggests I've missed something in modeling



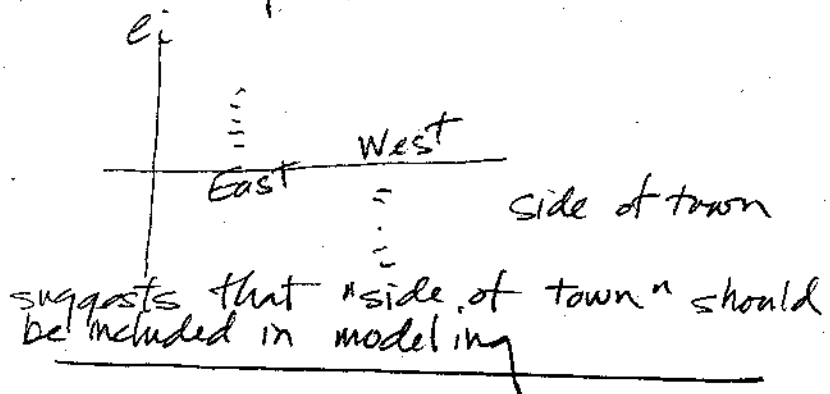
I can plot  $e_i$  vs  $x_{1i}$   
 $x_{2i}$   
 $x_{ki}$   
 $y_i$



This suggests that the "constant  $\sigma$ " part of the model is no good

Plotting against "other" variables not in a model

"Example" Real Estate  
 $\hat{y} = \beta_0 + \beta_1 \text{size} + \beta_2 \text{condition}$



Deleted Residuals

$\hat{y}_{(i)}$  = value predicted by model for case  $i$  when model is fit to only the other  $n-1$  cases

$e_{(i)} = y_i - \hat{y}_{(i)}$  (ith deleted residual)

and one hopes that  $e_{(i)}$  are not much bigger than the  $e_i = y_i - \hat{y}_i$

A way of measuring this is using  
 $\text{PRESS} = \sum e_{(i)}^2 = \sum (y_i - \hat{y}_{(i)})^2$

Other kinds of residuals

$e_i = y_i - \hat{y}_i$  ← ordinary residuals

$e_i^* = \frac{e_i}{SE_{e_i}}$  ← standardized residuals  
 (std error) for  $i$ th residual

these are expected to be between -2 to 2 ... big ones can flag recording errors / odd cases / influential data pts

Recall  $SSE = \sum e_i^2$

$\text{PRESS} \geq SSE$

(but we hope it's not too much bigger)

There are also "partial residuals" (that are origin of the JMP "leverage plots")

put this on hold for a few minutes

Another idea (for identifying important cases in regression is the idea of "hats" (leverage values in non-SAS world) (JMP calls something else The leverage values)

Fact: In a given MLR there are  $n \times n$  constants  $h_{ij}$  so that  $i$ th hat  $\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \dots + h_{ii}y_i + \dots + h_{in}y_n$

That value ( $h_{ii}$ ) in some sense measures how important  $y_i$  is in predicting  $y_i$

$h_{ii}$  flag data points with  $z$ 's that put them on the "edge" of the data set as regards the predictors — we might want a measure that also takes into account a data point's  $y_i$  —

"Cook's Distance"

$$D_i = \frac{h_{ii}}{(k+1)MSE} \left( \frac{e_i}{1-h_{ii}} \right)^2$$

"  $\left( \frac{h_{ii}}{k+1} \right) \left( \frac{e_i}{s} \right)^2$  " i<sup>th</sup> deleted residual

$$0 < h_{ii} < 1$$

$$\sum h_{ii} = k+1 \quad \left. \vphantom{\sum h_{ii}} \right\} \text{this means that } h_{ii}'\text{'s average to } \frac{k+1}{n}$$

a common rule of thumb is that  $h_{ii} > 2 \frac{k+1}{n}$

flags a "high leverage" case, i.e. one that is influential in fitting potentially

To have a big  $D_i$  a data point must have both a big  $h_{ii}$  value (be near the "edge" of data set) and have a big deleted residual (be poorly predicted by a model fit without using it)