DADOS ATÍPICOS
(potenciais *outliers*)

Pontos podem ser extremos com relação a X ou Y ou a X e Y

Geralmente, os dados atípicos têm resíduos grandes e podem exercer influencia sobre o modelo ajustado

IMPORTANTE:

- investigar a causa das discrepâncias
- avaliar a influência das observações atípicas no modelo ajustado

Como detectar?

→ análise de resíduos é muito útil na identificação desses pontos

→ resíduos consideravelmente grandes são potenciais *outliers*

3 ou mais desvios padrão distante da média

Qual a razão do seu comportamento não usual?

Se erro de medição ou erro de digitação:

→ corrigir sempre que possível ou então excluir o ponto

Se é um valor particularmente desejável para a resposta (preço baixo, alta produção, etc) ou se o conhecimento da ocorrência dessa resposta pode ser extremamente útil, podendo levar a descoberta de um fenômeno raro:

→ não excluir

PORTANTO, nem sempre o dado atípico deve ser encarado como um valor ruim e muito menos ser automaticamente excluído

Como verificar o efeito de um potencial *outlier*?

→ retirar e reajustar o modelo, comparando as novas estimativas de β_0, β_1 e σ^2

* Em geral, estas estimativas são extremamente sensíveis a dados atípicos

* Dependendo da localização ou do número de pontos, os *outliers* possuem efeito (influência) desde fraco até muito sério no modelo ajustado

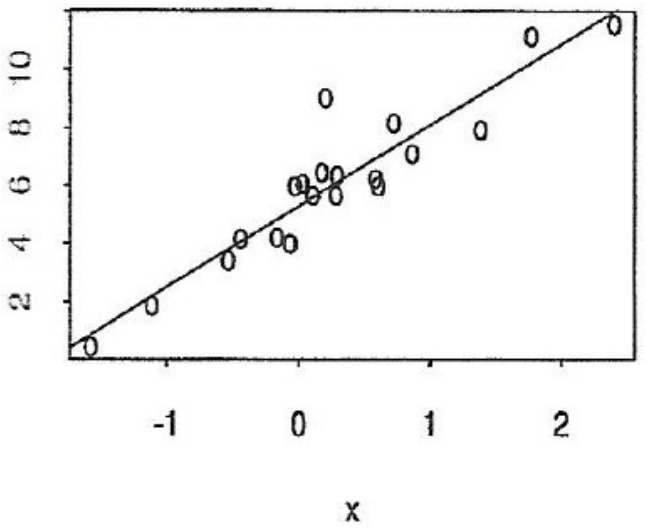
* Uma avaliação mais detalhada dos pontos influentes será vista no modelo de regressão linear múltiplo

Ilustação 1

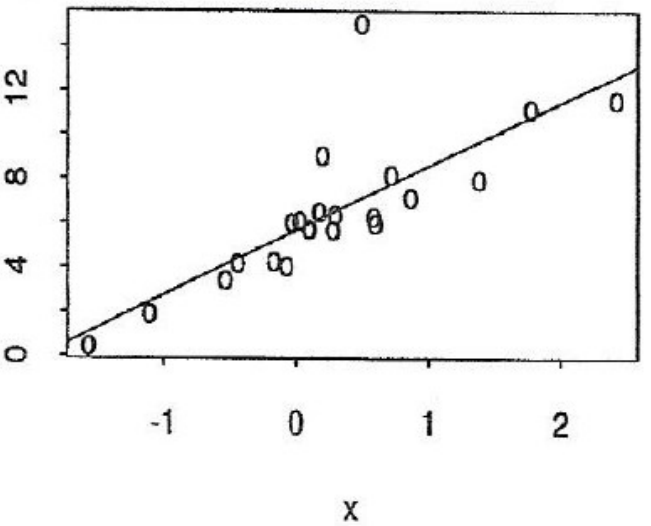
Conjuntos de dados com 20 e 100 observações

Gráfico A → sem ponto atípico
Gráficos B, C e D → iguais a A com mais um ponto atípico

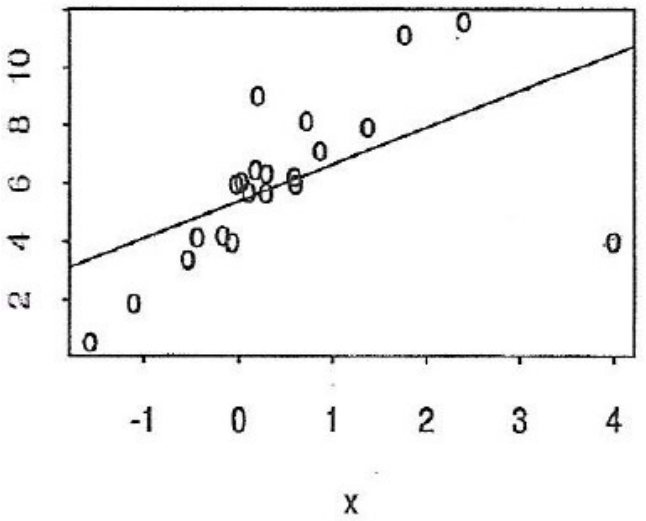
(A) Y= 5,3+2,8X



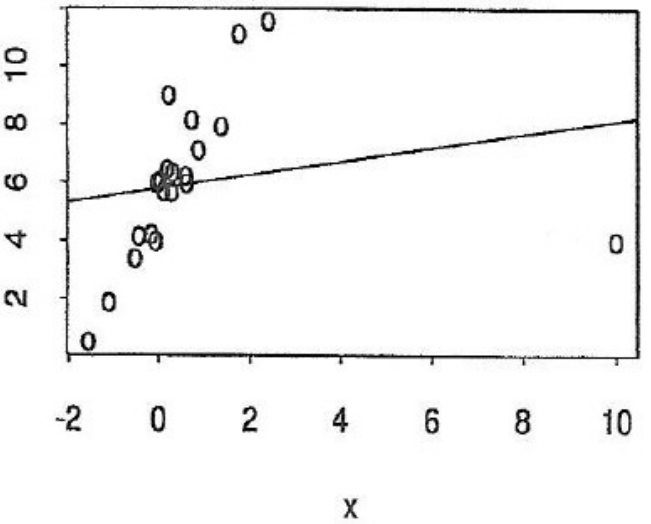
(B) Y= 5,7+2,9X



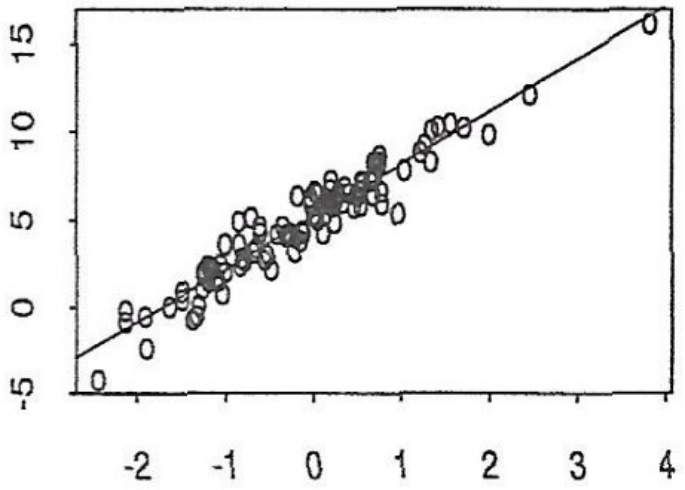
(C) Y= 5,4+1,3X

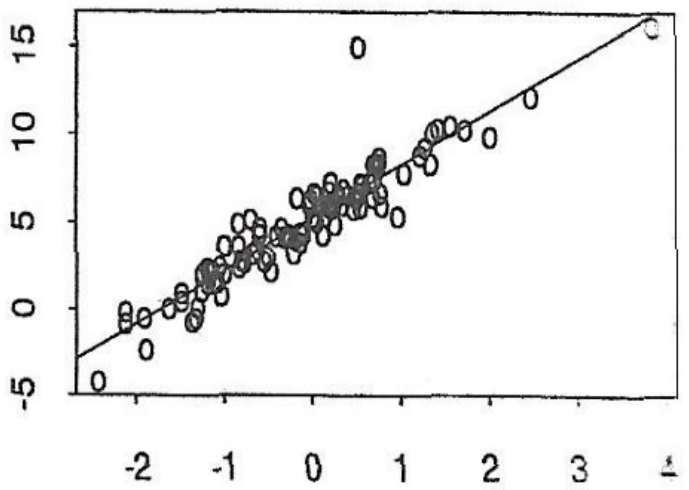


(D) Y= 5,8+0,2X



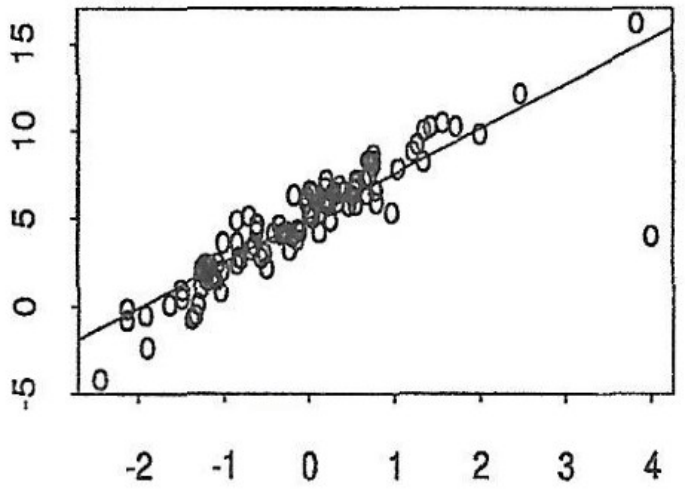
conjunto de 20 pontos

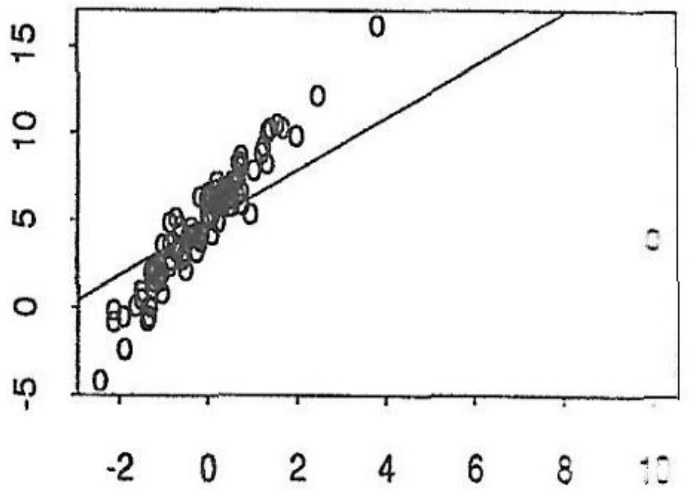




C Y=5,02+2,57X

D Y=4,83+1,51X





conjunto de 100 pontos

Gráfico B: ponto afastado acima, com valor de x próximo a $\overline{x}$ e valor muito alto de y

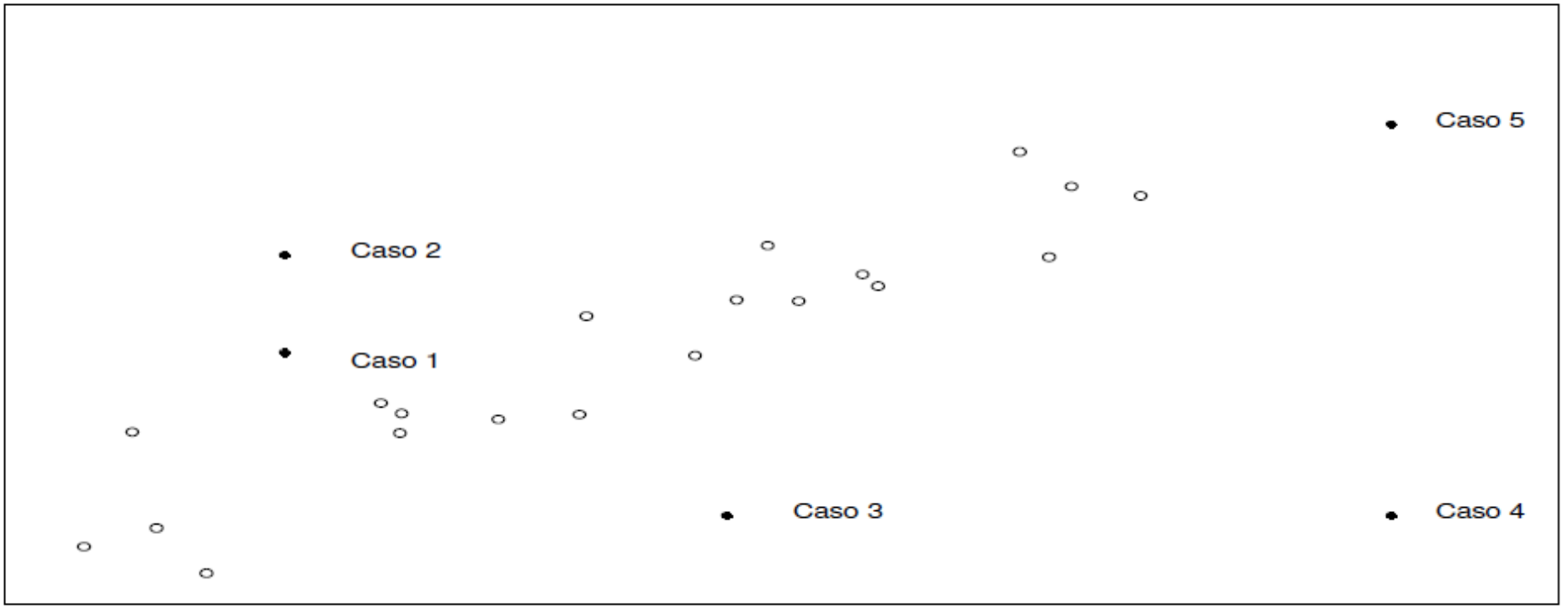
Gráfico C: ponto afastado à direita, com valor alto de x (ponto de alavanca) e valor de y próximo a $\overline{y}$

Gráfico D: ponto muito afastado à direita, com valor muito alto de x (ponto de alavanca)e valor de y próximo a $\overline{y}$

20 dados:
Gráfico B → leve acréscimo no intercepto
Gráfico C → mudança na inclinação
Gráfico D → grande mudança no intercepto e na inclinação

100 dados:
Gráfico B → praticamente não teve efeito sobre o ajuste
Gráfico C → leve mudança no ajuste
Gráfico D → mudança maior no ajuste (mais na inclinação)

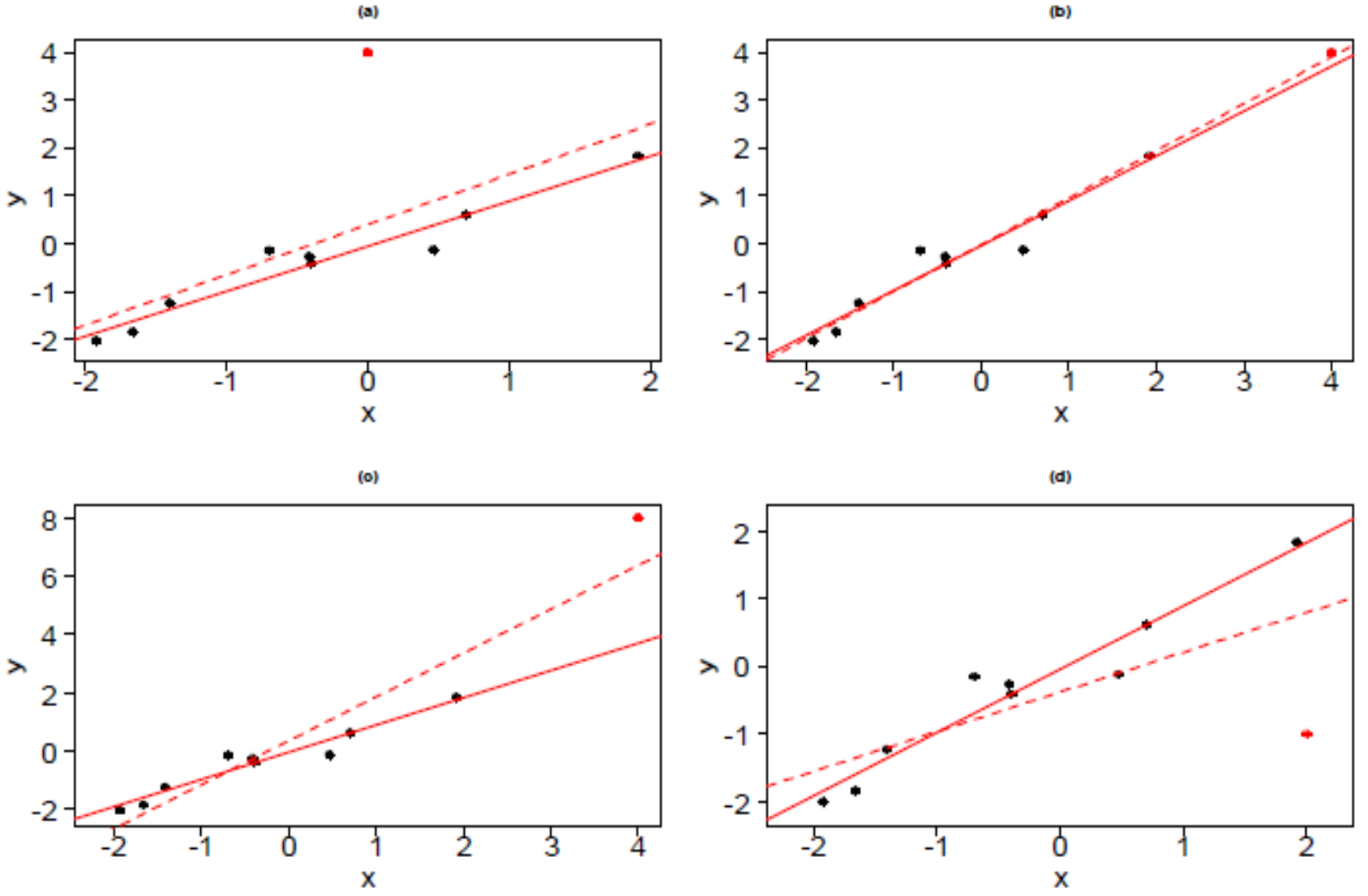
Ilustração 2



Casos 1, 2 e 3: discrepantes com relação a Y
Caso 5: discrepante com relação a X
Caso 4: discrepante com relação a X e Y

Casos 2, 3 e 4: maior influência sobre o ajuste da reta
Caso 1: menor influência, dado sua maior proximidade ao conjunto de pontos
Caso 5: pode ser consistente, dada sua relação com os demais pontos

Ilustração 3



Diagramas de dispersão com observações atípicas (a reta pontilhada representa a regressão linear simples ajustada com todas as observações e, a contínua, sem a observação atípica).

(a) ponto destacado tem resíduo elevado; é **inconsistente**, pois destoa da tendência dos demais pontos e é **influente**, pelo impacto produzido na reta ajustada

(b) ponto destacado não tem resíduo elevado; é **ponto de alavanca**, por ter um valor extremo de *x*; é **consistente**, pois não destoa da tendência dos demais pontos e é **não influente**

(c) ponto destacado tem resíduo elevado; é classificado como **ponto de alavanca**, por ter um valor extremo de *x*; é **inconsistente** e **influente**

(d) ponto destacado tem resíduo elevado; é **inconsistente** e **influente**

Exemplo 1 (cont.): 12 meninas de uma escola de dança

i	y_i	x_i	e_i	z_i	z_i^*	$\hat{\sigma}^2_{(i)}$	$\hat{\beta}_0 - \hat{\beta}_{0(i)}$	$\hat{\beta}_1 - \hat{\beta}_{1(i)}$
1	34	1,35	3,89	0,71	0,70	38,05	9,99	-6,77
2	34	1,35	3,89	0,71	0,70	38,05	9,99	-6,77
3	29	1,35	-1,11	-0,20	-0,19	39,94	-2,84	1,93
4	27	1,35	-3,11	-0,57	-0,55	38,80	-7,97	5,40
5	40	1,40	3,91	0,68	0,66	38,24	2,53	-1,53
6	25	1,40	-11,09	-1,93	-2,32	25,10	-7,18	4,35
7	40	1,40	3,91	0,68	0,66	38,24	2,53	-1,53
8	34	1,40	-2,09	-0,36	-0,35	39,57	-1,35	0,82
9	46	1,50	-2,05	-0,39	-0,37	39,49	6,54	-4,78
10	*42*	*1,50*	*-6,05*	*-1,15*	*-1,17*	*34,80*	*19,29*	*-14,08*
11	47	1,50	-1,05	-0,20	-0,19	39,94	3,36	-2,45
12	*59*	*1,50*	*10,95*	*2,08*	*2,62*	*22,76*	*-34,88*	*25,46*

```
###dados - Exemplo 1
x <- c(rep(1.35,4),rep(1.4,4),rep(1.5,4))
y <- c(34,34,29,27,40,25,40,34,46,42,47,59)
# formando a base de dados
dados <- cbind(x,y)

#ajuste de MQ
reta<- lm(y~x)

#resíduos
residuo <- reta$res
z <- rstandard(reta)    # residuos padronizados
zstudent <- rstudent(reta)    # residuos studentizados

#Analise de influencia
infl<-influence(reta)
names(infl)
sigma2i <- infl$sigma^2
difbeta0i <- infl$coef[,1]
difbeta1i <- infl$coef[,2]
cbind(y,x,residuo,z,zstudent,sigma2i,difbeta0i,difbeta1i)
```