704SM Biostatistica

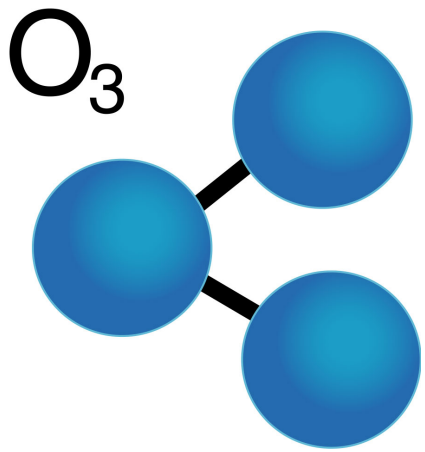
Massimo Borelli

dicembre 2016

UNIVERSITÀ DEGLI STUDI DI TRIESTE
Dipartimento di Scienze della Vita

SOCIETA' DEI MATEMATICI
E NATURALISTI DI MODENA
www.socnatmatmo.unimore.it

questioni da approfondire



La regressione multipla

il dataset `airquality`

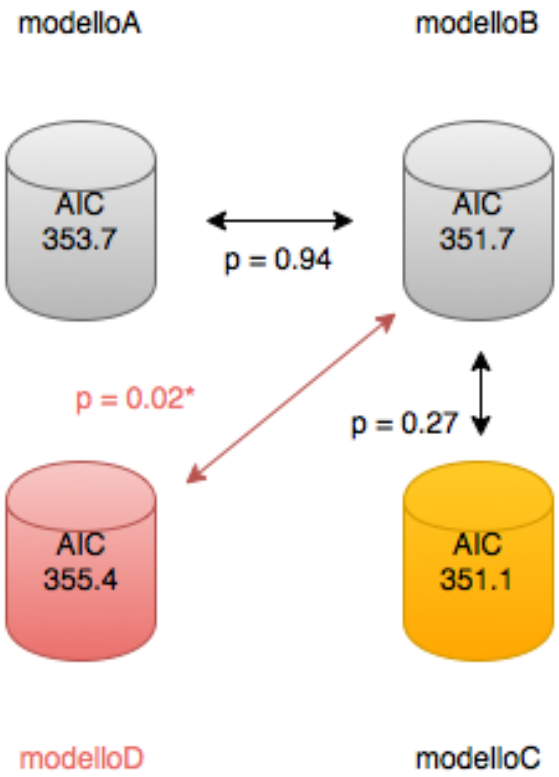
	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.40	67	5	1
2	36	118	8.00	72	5	2
3	12	149	12.60	74	5	3
4	18	313	11.50	62	5	4
5	NA	NA	14.30	56	5	5
6	28	NA	14.90	66	5	6
..	..	..	..	..	..	..
152	18	131	8.00	76	9	29
153	20	223	11.50	68	9	30

si tratta di una serie temporale, non di un design *cross-section*

questioni da approfondire

domande

come si sceglie il 'migliore' modello statistico?

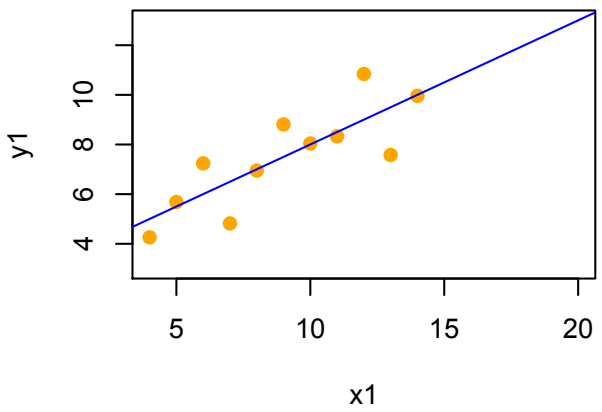


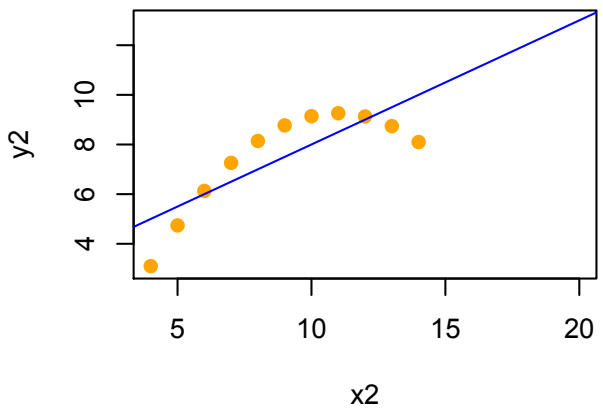
domande

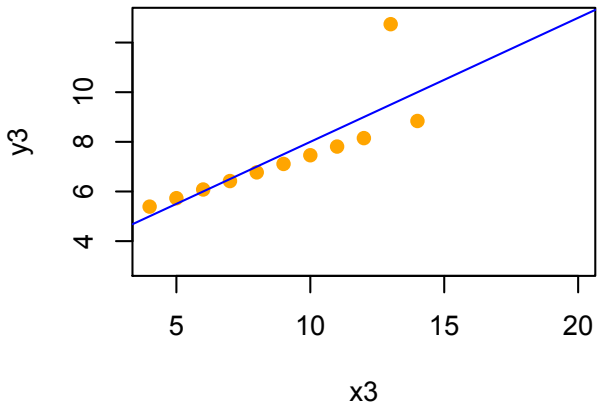
come si sceglie il 'migliore' modello statistico

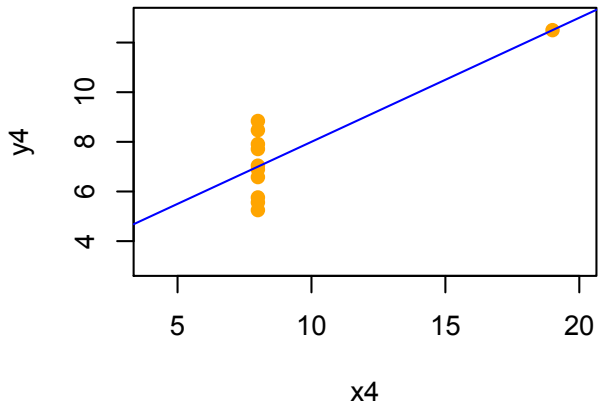
- mediante la **analisi della devianza**
- confrontando i **criteri di informazione**
- simulando dati in base allo *residual standard error*

che cosa è la devianza?

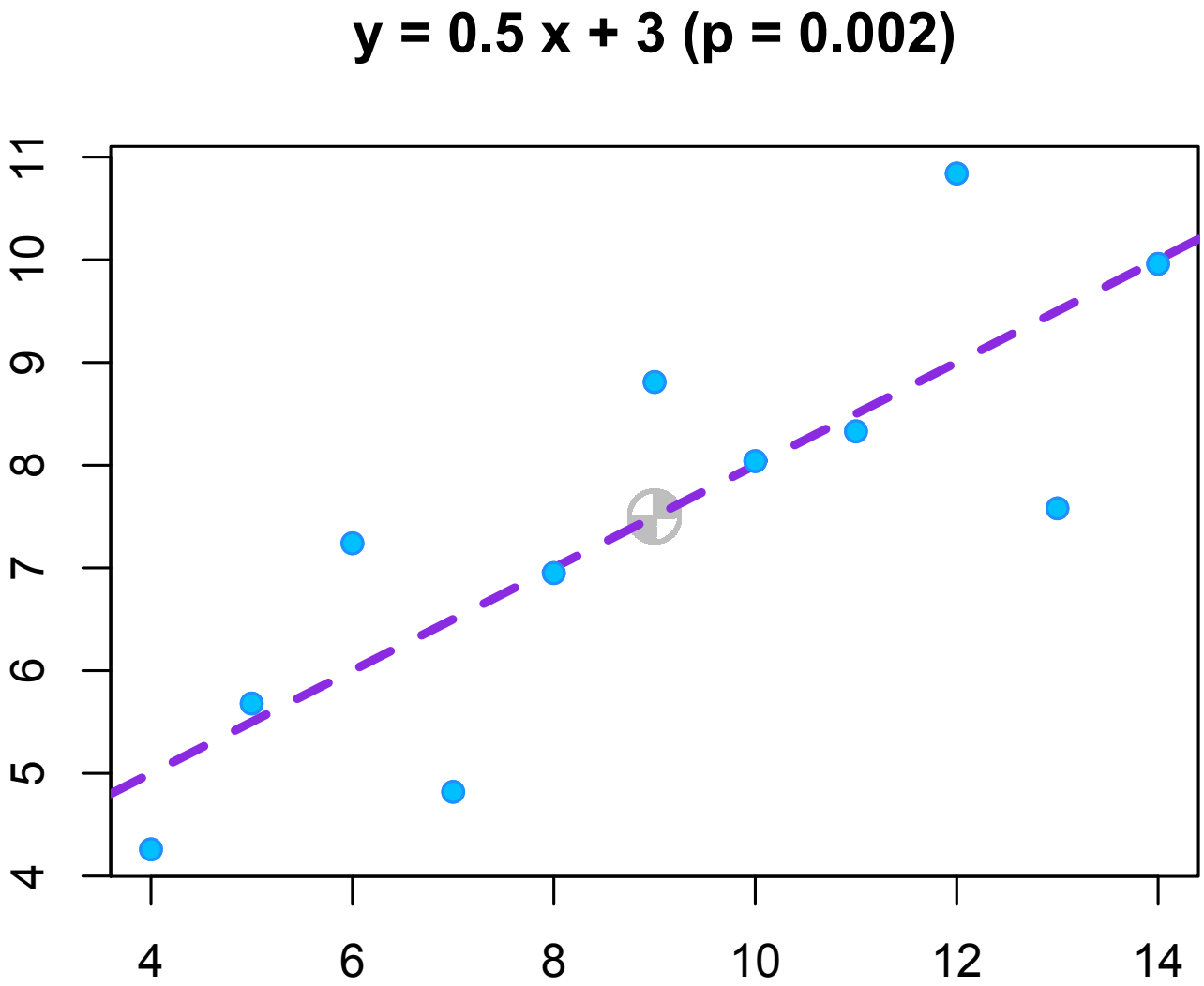




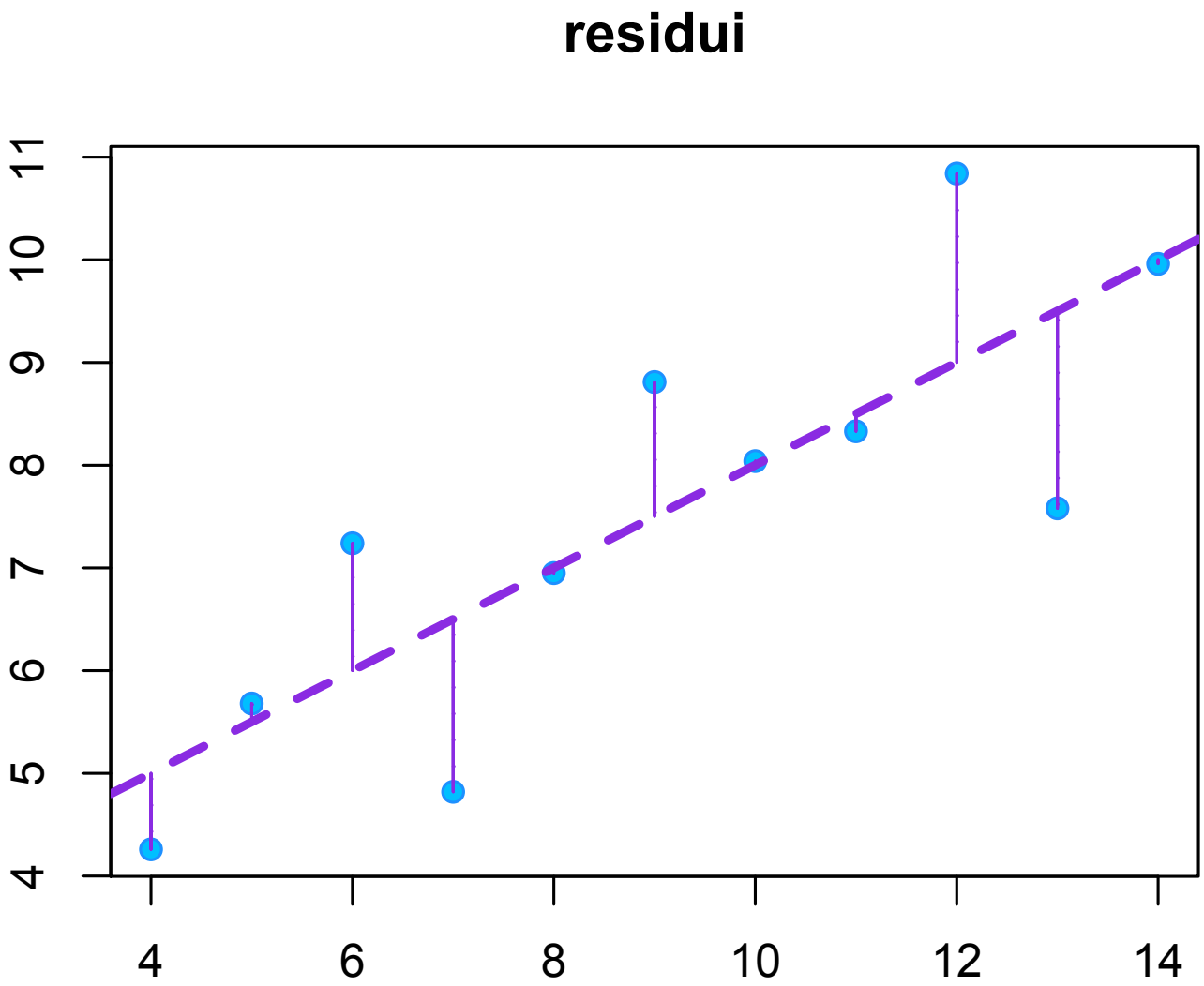


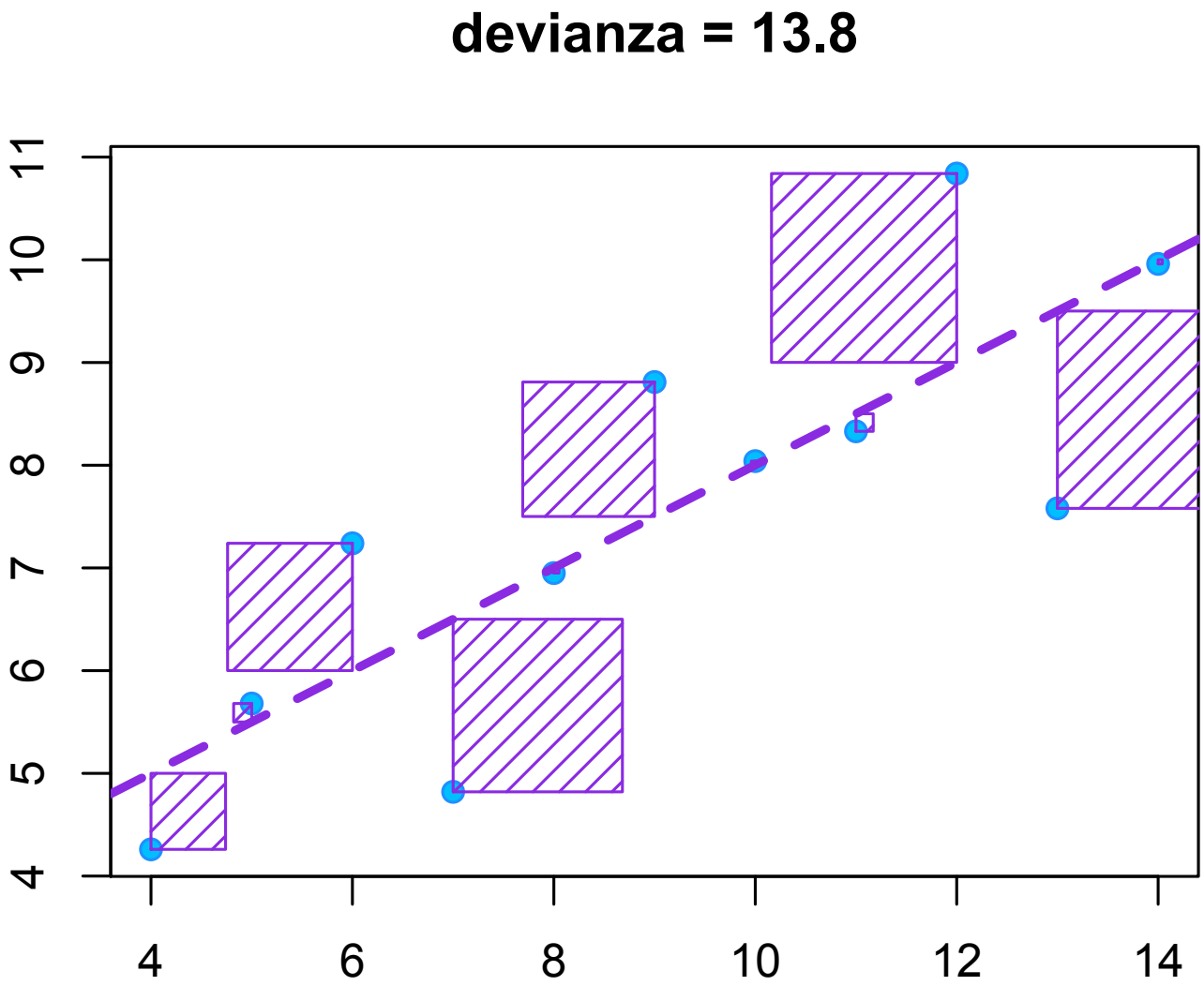


che cosa è la devianza

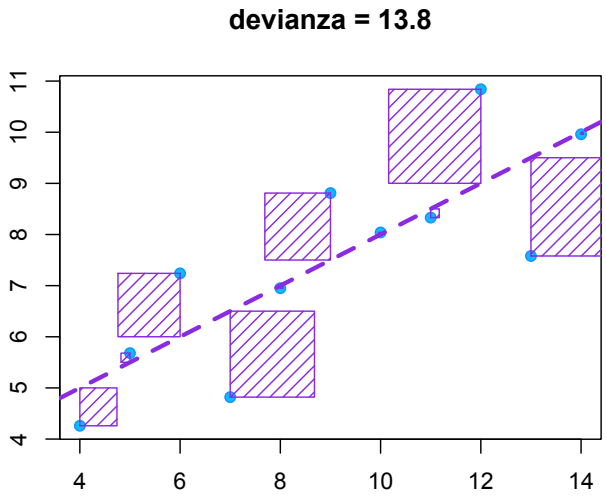


che cosa è la devianza




devianza = 13.8
4
5
6
7
8
9
10
11
4
6
8
10
12
14

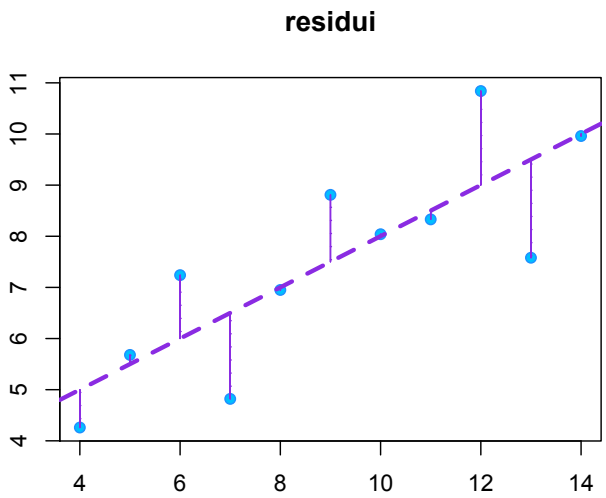
analisi della devianza

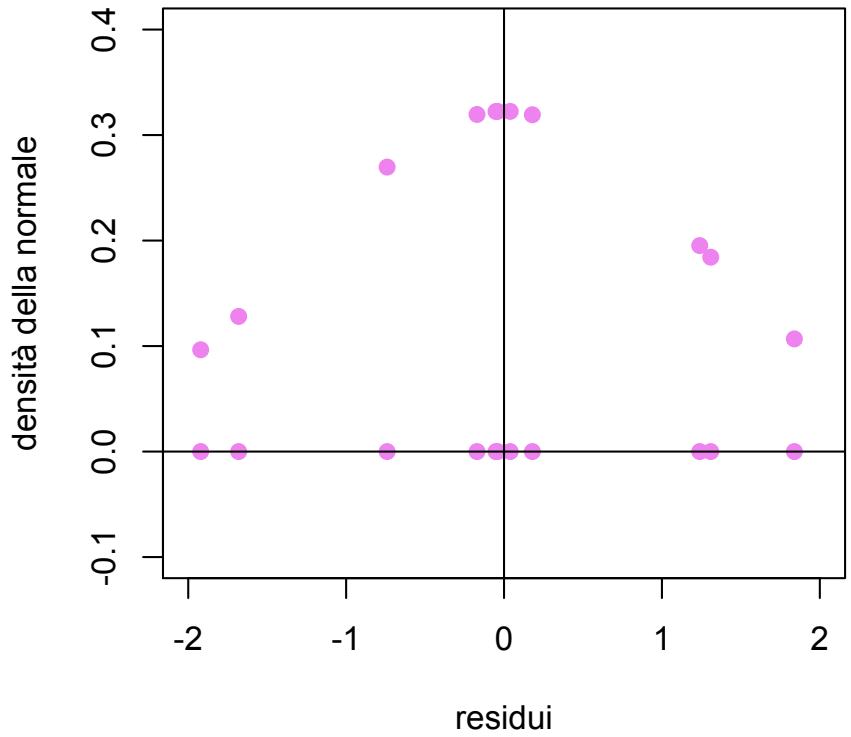


- la somma dei quadrati dei residui (RSS) si comporta approssimativamente come una variabile aleatoria Chi Quadrato
- possiamo testare le RSS dei modelli:**analisi della devianza**

```
anova(modelloA, modelloB)
```

che cosa è la log-verosimiglianza?






```
plot(resid(m2), dnorm(resid(m2), 0, sigma))
```

che cosa sono i criteri di informazione?

che cosa è la log-verosimiglianza?

```
> dim(model.matrix(m1))[2]     #  intercept + slope
[1] 2
> sigma.ML = sigma*sqrt((11-dim(model.matrix(m1))[2])/11)
> sigma.ML
[1] 1.11855
> sum(log(dnorm(resid(m2), mean = 0, sd = sigma.ML)))
[1] -16.84069
> logLik(m1)
'log Lik.' -16.84069 (df=3)
```

che cosa sono i criteri di informazione?

Definition

Suppose that we have a statistical model of some data. Let L be the maximum value of the likelihood function for the model; let k be the number of estimated parameters in the model. Then the AIC value of the model is the following.[1][2]

$$\mathrm{AIC} = 2k - 2\ln(L)$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages

```
> logLik(m2)
'log Lik.' -16.84069 (df=3)
> 2 * 3 - 2 * logLik(m1)
'log Lik.' 39.68137 (df=3)
> AIC(m2)
[1] 39.68137
```

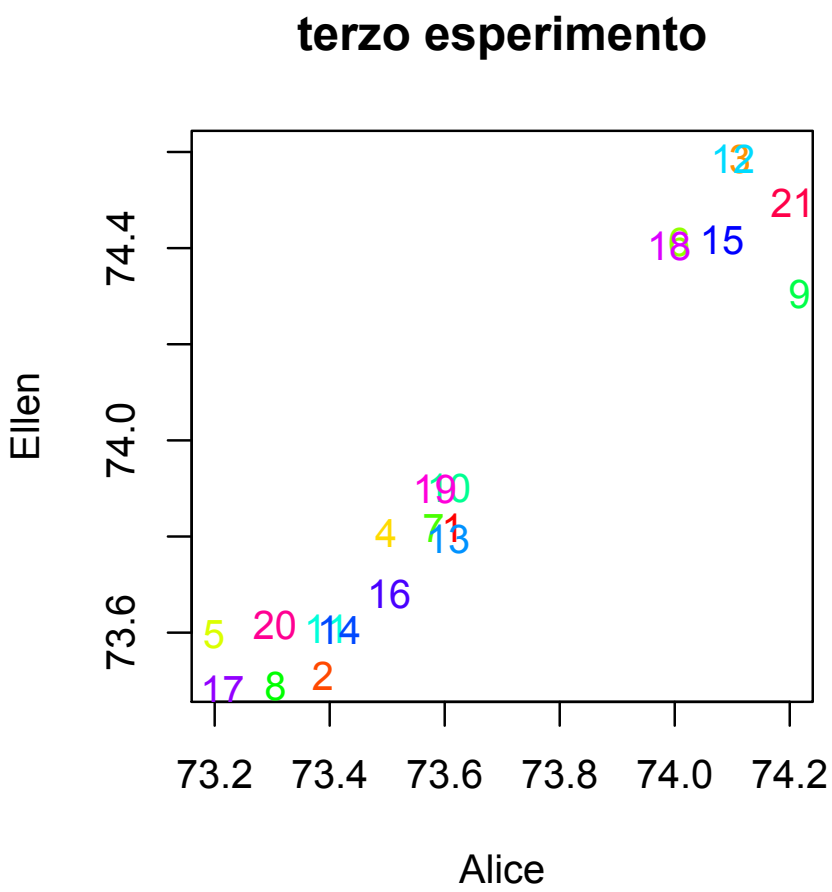
come si simulano i dati in base allo *residual standard error*?

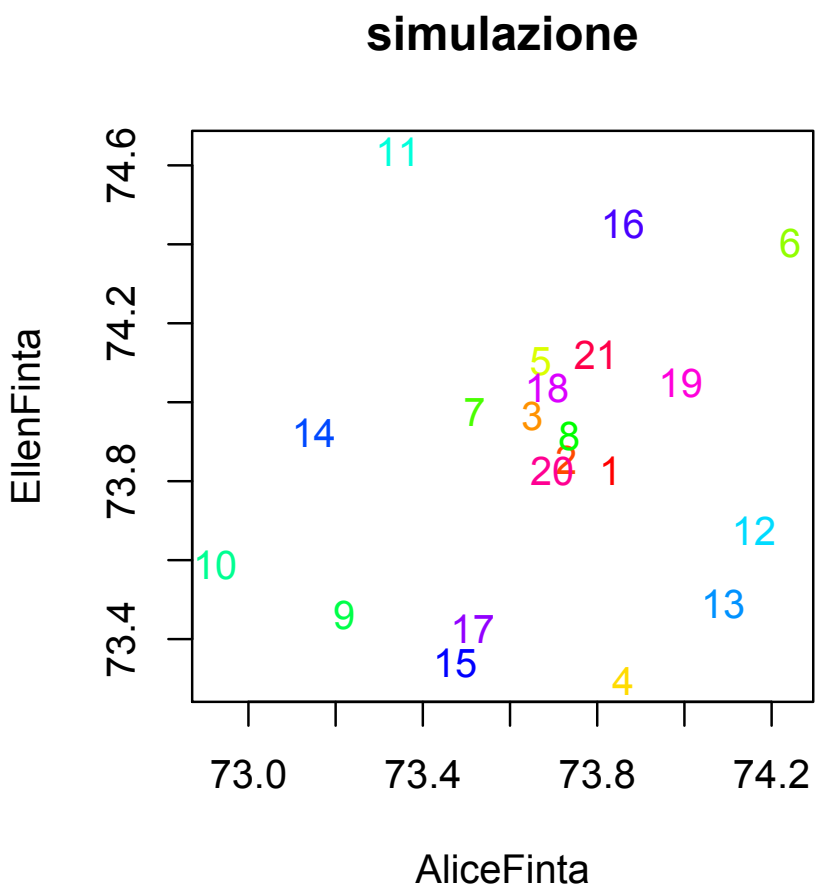
	Alice	Ellen		Alice	Ellen
1	73.60	73.80	12	74.10	74.60
2	73.40	73.50	13	73.60	73.80
3	74.10	74.60	14	73.40	73.60
4	73.50	73.80	15	74.10	74.40
5	73.20	73.60	16	73.50	73.70
6	74.00	74.40	17	73.20	73.50
7	73.60	73.80	18	74.00	74.40
8	73.30	73.50	19	73.60	73.90
9	74.20	74.30	20	73.30	73.60
10	73.60	73.90	21	74.20	74.50
11	73.40	73.60	-	-	-

come si simulano i dati in base allo *residual standard error*?

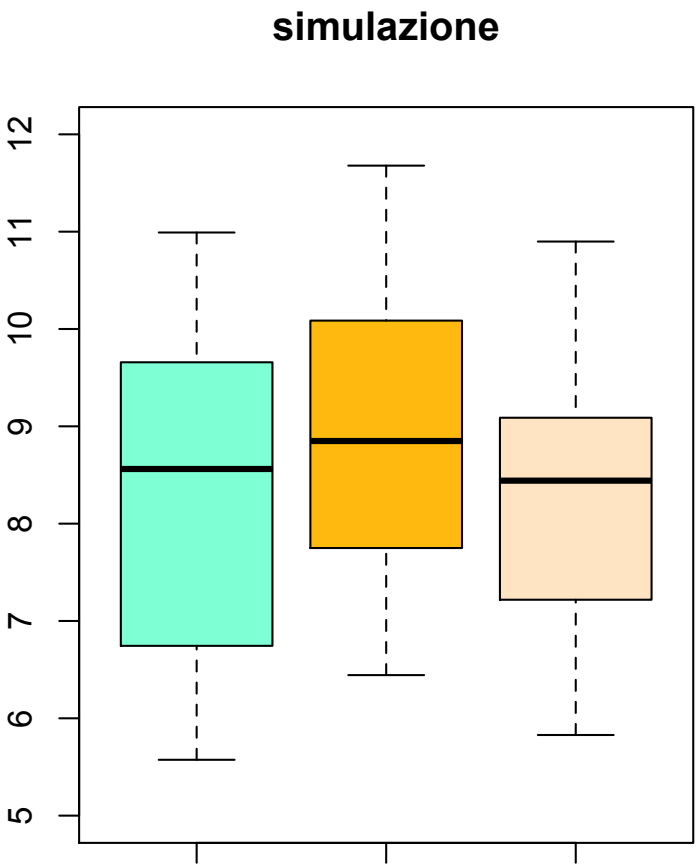
```
Call:
lm(formula = peso ~ gemella)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4619 -0.3226 -0.1024  0.4179  0.6571

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.66190    0.08078 911.915   <2e-16 ***
gemellaellen  0.28095    0.11424   2.459   0.0183 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3702 on 40 degrees of freedom
Multiple R-squared:  0.1314, Adjusted R-squared:  0.1096
F-statistic: 6.049 on 1 and 40 DF,  p-value: 0.01834
```

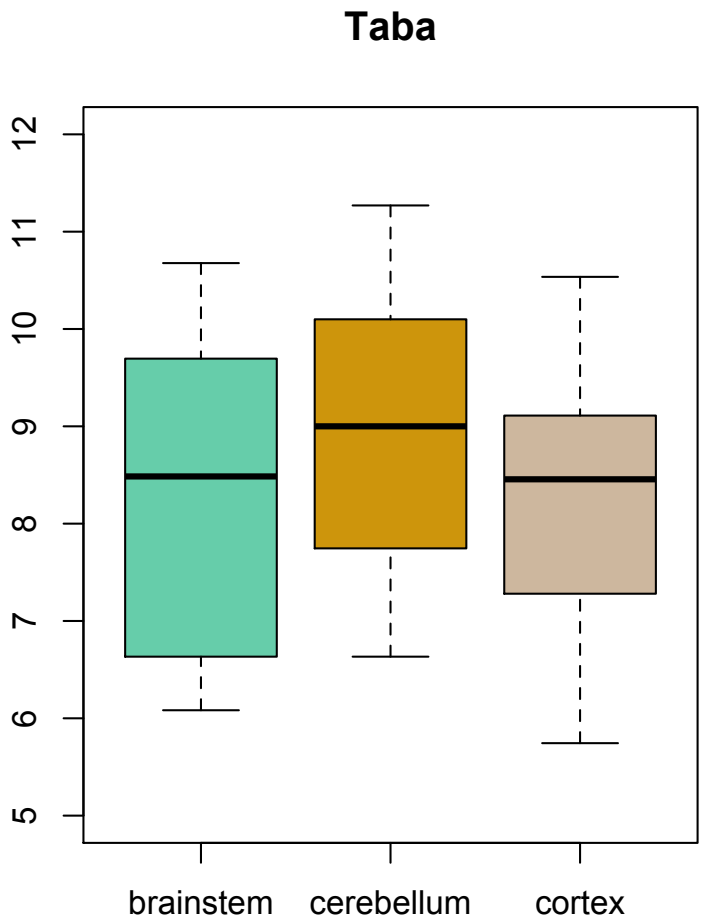
come si simulano i dati in base allo *residual standard error*?



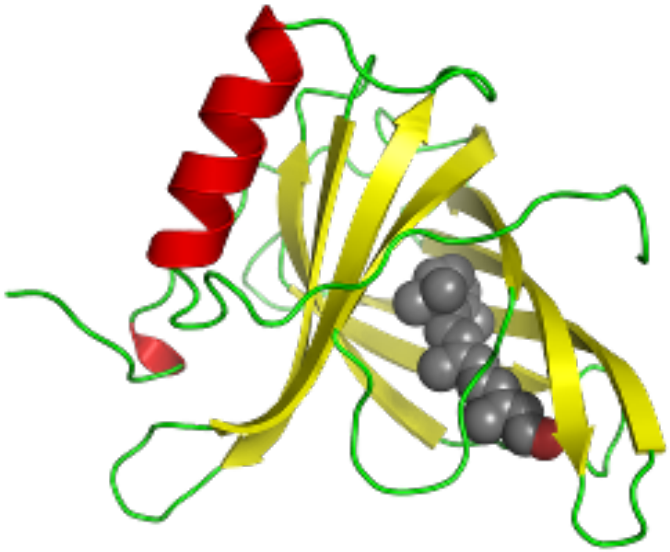


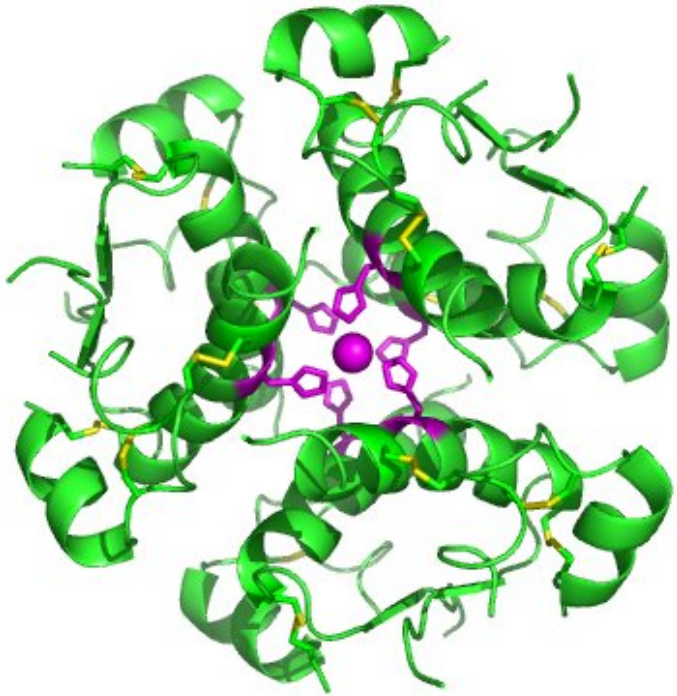
come si simulano i dati in base allo *residual standard error*?





il dataset diabete





0021-972X/07/$15.00/0
Printed in U.S.A.

The Journal of Clinical Endocrinology & Metabolism 92(12):4883–4888

doi: 10.1210/jc.2007-0325

BRIEF REPORT

Serum Retinol-Binding Protein-4, Leptin, and Adiponectin Concentrations Are Related to Ectopic Fat Accumulation

Conclusion: Serum RBP-4 was a robust marker of insulin resistance. Ser… RBP-4, leptin, and adiponectin concentrations reflected ec-to… …ccumulation in humans. (***J Clin Endocrinol Metab* 92:** …4888, 2007)

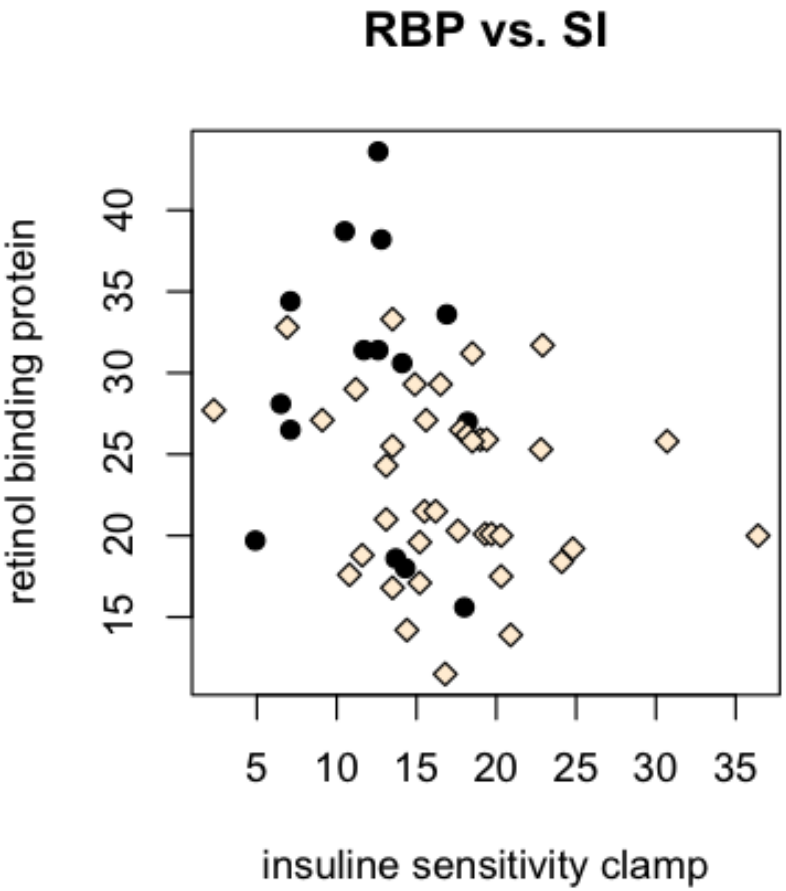
Conclusion: Serum RBP-4 was a robust marker of insulin resistance. Serum RBP-4, leptin, and adiponectin concentrations reflected ectopic fat accumulation in humans. (***J Clin Endocrinol Metab* 92: 4883–4888, 2007)**

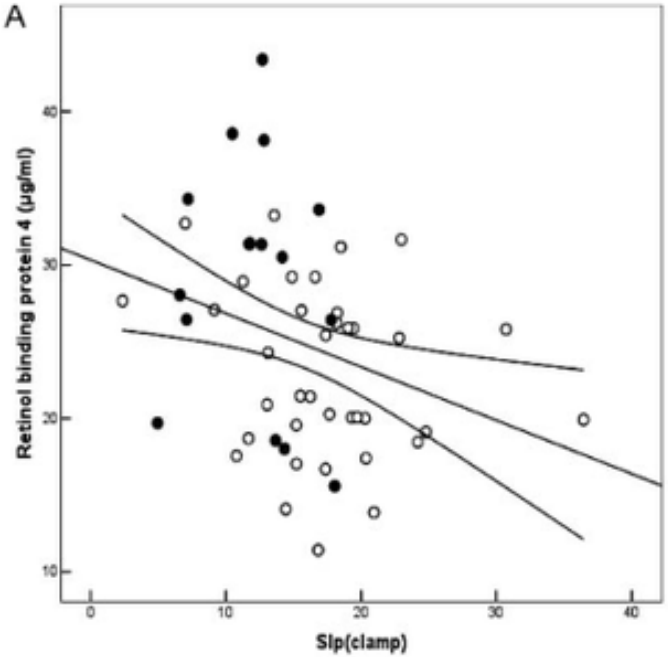
…trations were measured.

il dataset diabete

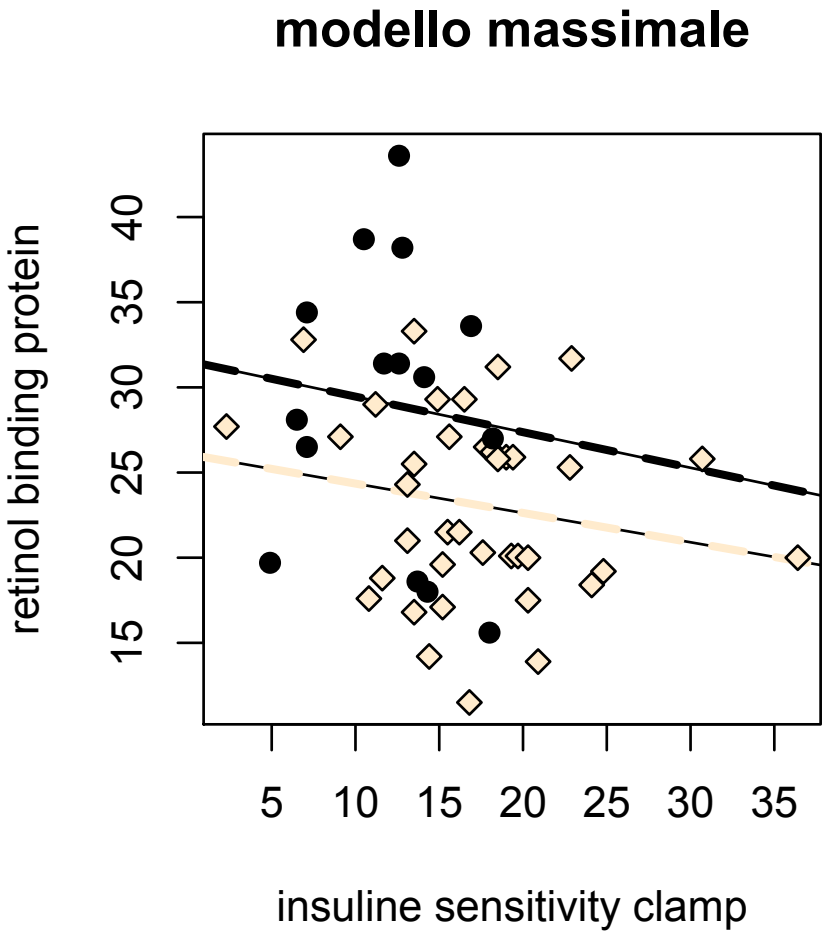
	serumprotein	response	group
1	12.60	43.60	black
2	10.50	38.70	black
3	12.80	38.20	black
4	7.10	34.40	black
5	16.90	33.60	black
6	13.50	33.30	white
...	...	...	...
51	14.40	14.20	white
52	20.90	13.90	white
53	16.80	11.50	white

il dataset diabete

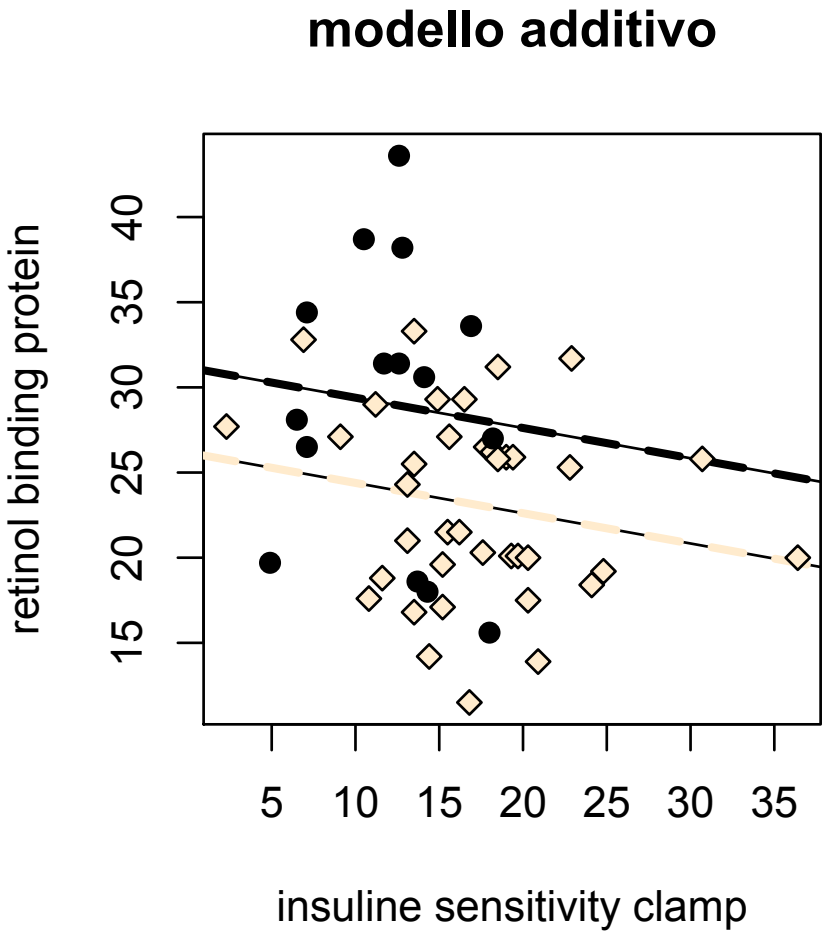




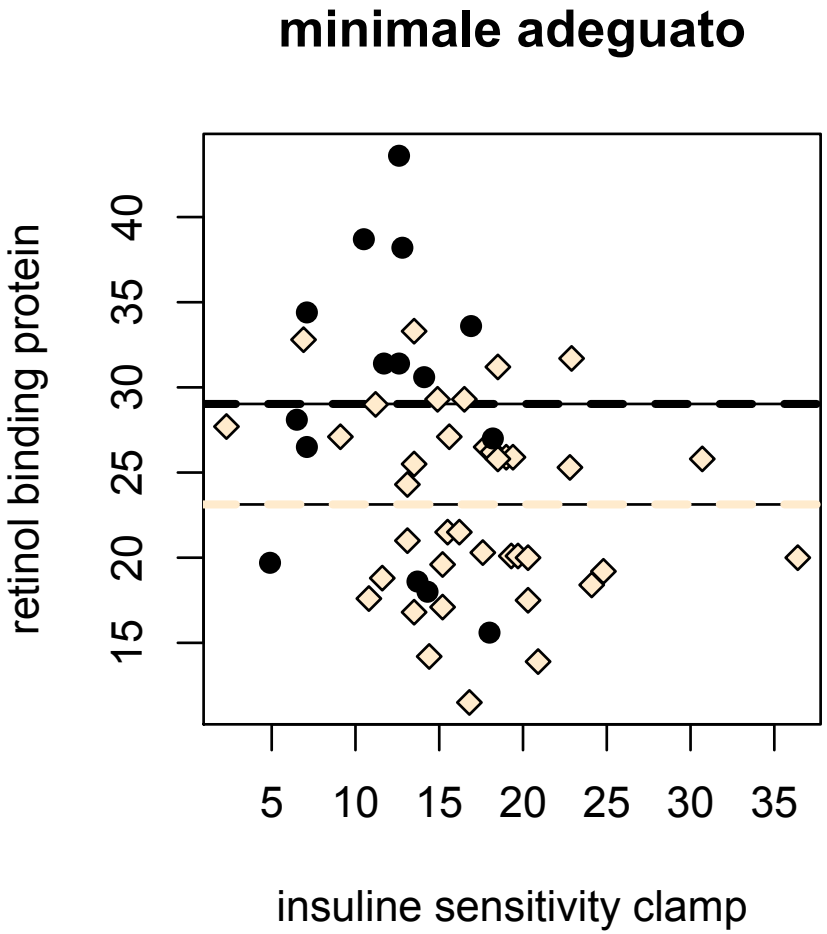
il dataset diabete



il dataset diabete



il dataset diabete



il dataset diabete