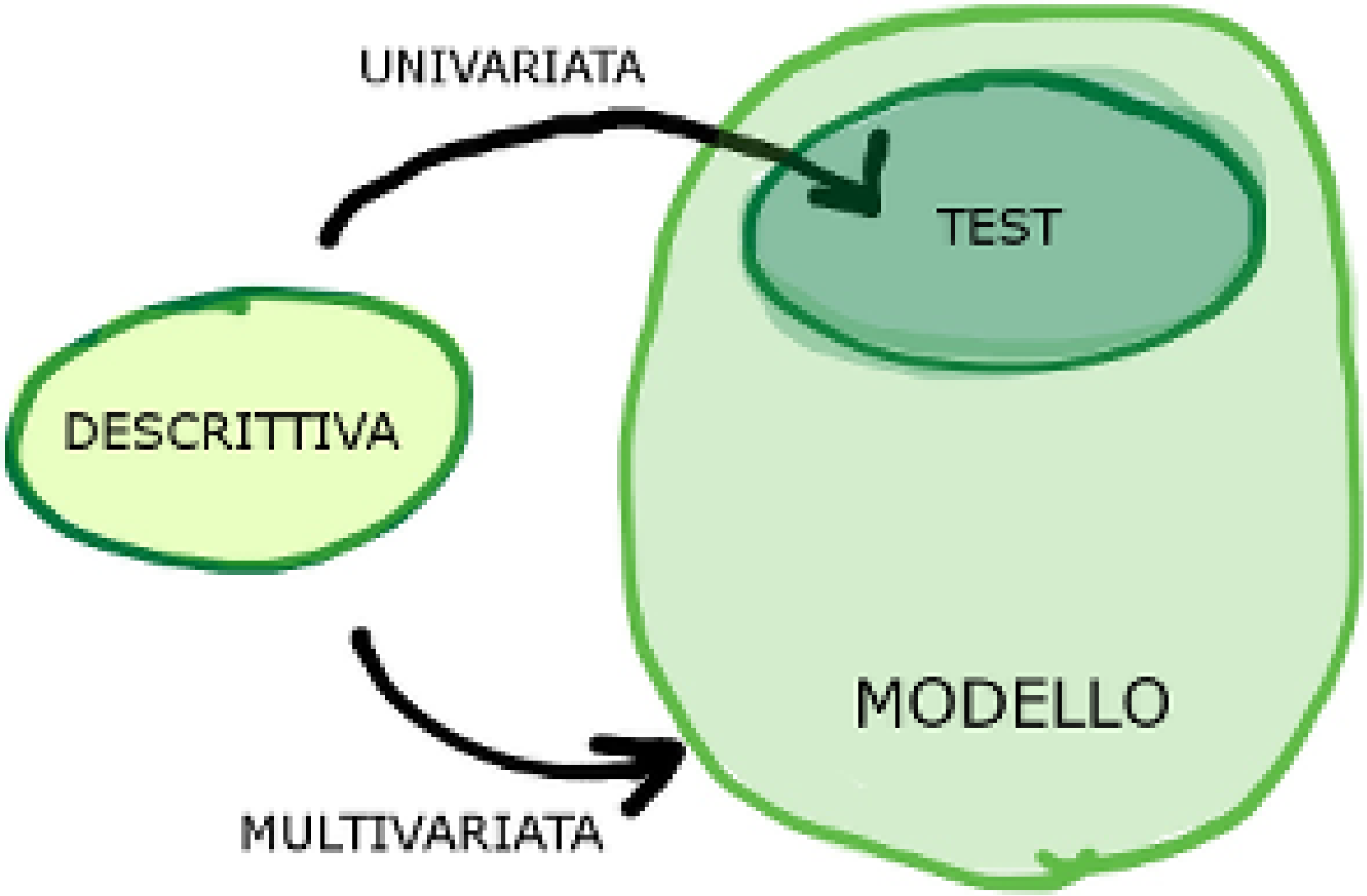
introduzione ai modelli statistici

Massimo Borelli

a.a. 2015-2016

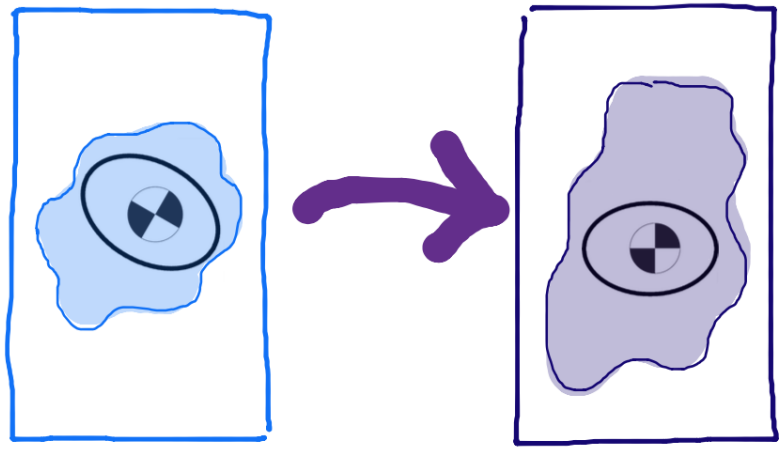
UNIVERSITÀ
DEGLI STUDI DI TRIESTE

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


UNIVARIATA
TEST
DESCRITTIVA
MODELLO
MULTIVARIATA

2.2 Statistical Models

Statistical models are used to describe a sample of data taken from a real or theoretical population. Statistical models can be described using one or more underlying probability distributions. The parameters of the distributions are estimated from the data, and may provide the basis for predicting additional data with the same distributional characteristics of the data being modeled. Models that can be defined in terms of a probability distribution having estimable parameters are called parametric models. We will focus our attention in this text on this type of model.



quali modelli statistici

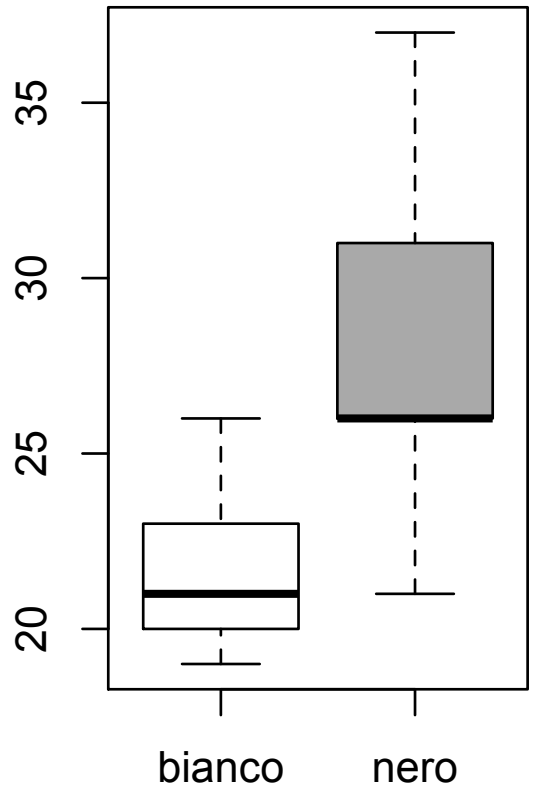
- il **modello lineare**
  - il test t
  - (la anova)
  - la retta di regressione
  - la ancova
- il **modello lineare generalizzato**
  - il test di Fisher
  - la regressione logistica
  - (la regressione di Poisson)
- il **modello di sopravvivenza**
  - (la regressione di Cox)
  - (..)

il dataset

	gruppo	insulina	retinolo
1	bianco	23	19
2	**nero**	**16**	**37**
3	bianco	34	23
4	bianco	29	26
5	bianco	19	20
6	**nero**	**20**	**21**
7	**nero**	**20**	**26**
8	bianco	27	21
9	**nero**	**14**	**26**
10	**nero**	**21**	**31**

visualizziamo il dataset

	gruppo	insulina	retinolo
1	bianco	23	19
2	**nero**	**16**	**37**
3	bianco	34	23
4	bianco	29	26
5	bianco	19	20
6	**nero**	**20**	**21**
7	**nero**	**20**	**26**
8	bianco	27	21
9	**nero**	**14**	**26**
10	**nero**	**21**	**31**



quali modelli statistici (1)

- il **modello lineare**
  - il test t

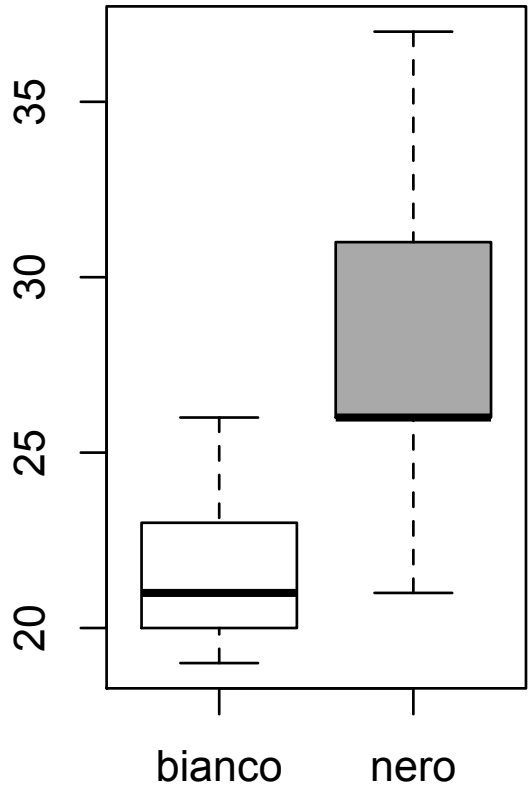
→ input: una variabile aleatoria binomiale

→ output: una variabile aleatoria gaussiana

il test t inteso come modello lineare

	gruppo	retinolo
1	bianco	19
2	**nero**	**37**
3	bianco	23
4	bianco	26
5	bianco	20
6	**nero**	**21**
7	**nero**	**26**
8	bianco	21
9	**nero**	**26**
10	**nero**	**31**

il test t inteso come modello lineare



	gruppo	retinolo
1	bianco	19
2	**nero**	**37**
3	bianco	23
4	bianco	26
5	bianco	20
6	**nero**	**21**
7	**nero**	**26**
8	bianco	21
9	**nero**	**26**
10	**nero**	**31**

il test t inteso come modello lineare

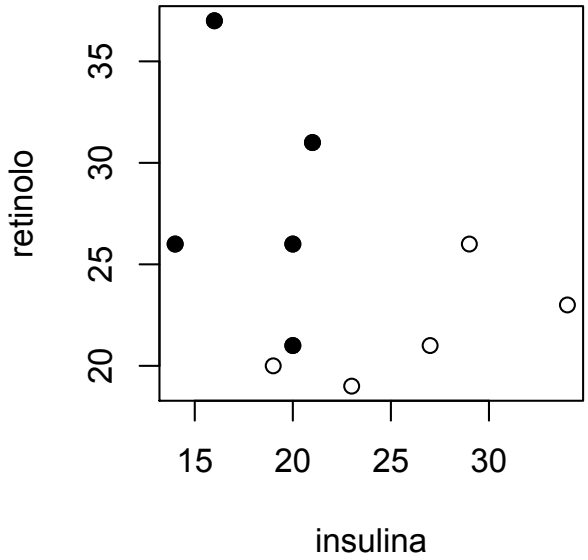
```
modello1 = lm(retinolo ~ 1 + gruppo)
summary(modello)
```

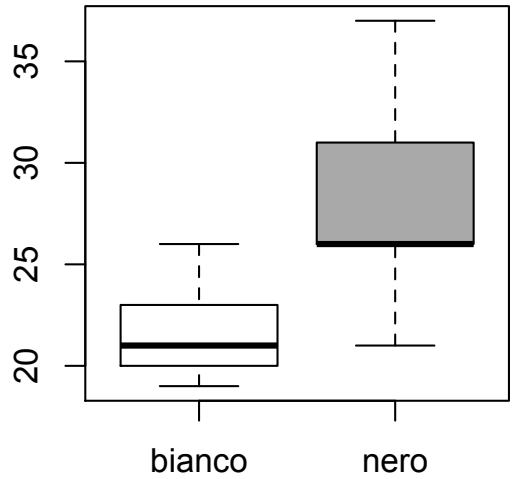
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.80	2.11	10.35	0.00
grupponero	6.40	2.98	2.15	0.06

Residual standard error: 4.7 on 8 degrees of freedom
Multiple R-squared: 0.37
F-statistic: 4.6 on 1 and 8 DF, p-value: 0.06

interpretazione del test t inteso come modello lineare

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.80	2.11	10.35	0.00
grupponero	6.40	2.98	2.15	0.06





quali modelli statistici (2)

- il **modello lineare**
  - la retta di regressione

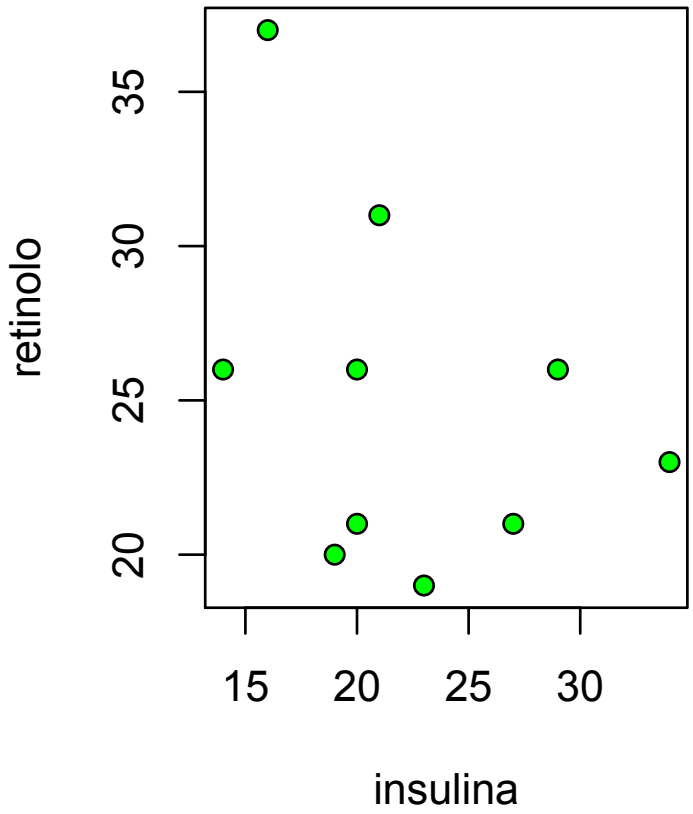
→ input: una variabile aleatoria gaussiana

→ output: una variabile aleatoria gaussiana

la retta di regressione intesa come modello lineare

	insulina	retinolo
1	23	19
2	16	37
3	34	23
4	29	26
5	19	20
6	20	21
7	20	26
8	27	21
9	14	26
10	21	31

la retta di regressione intesa come modello lineare



	insulina	retinolo
1	23	19
2	16	37
3	34	23
4	29	26
5	19	20
6	20	21
7	20	26
8	27	21
9	14	26
10	21	31

la retta di regressione intesa come modello lineare

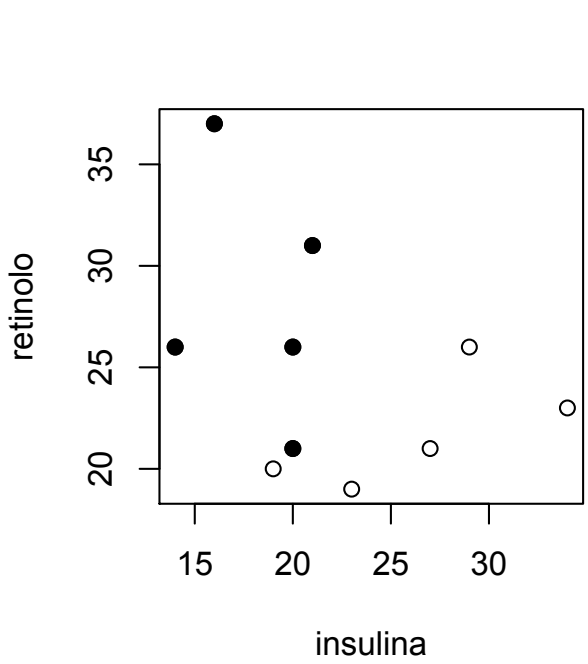
```
modello2 = lm(retinolo ~ 1 + insulina)
summary(modello2)
```

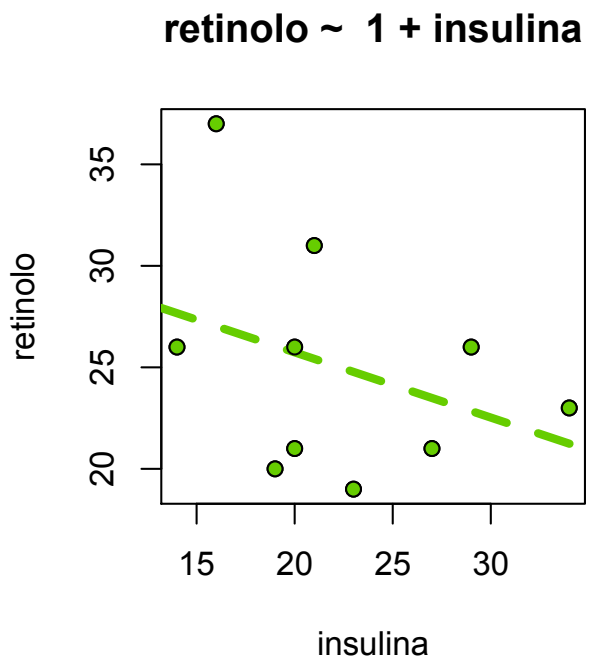
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.1657	6.9594	4.62	0.0017
insulina	-0.3213	0.3020	-1.06	0.3184

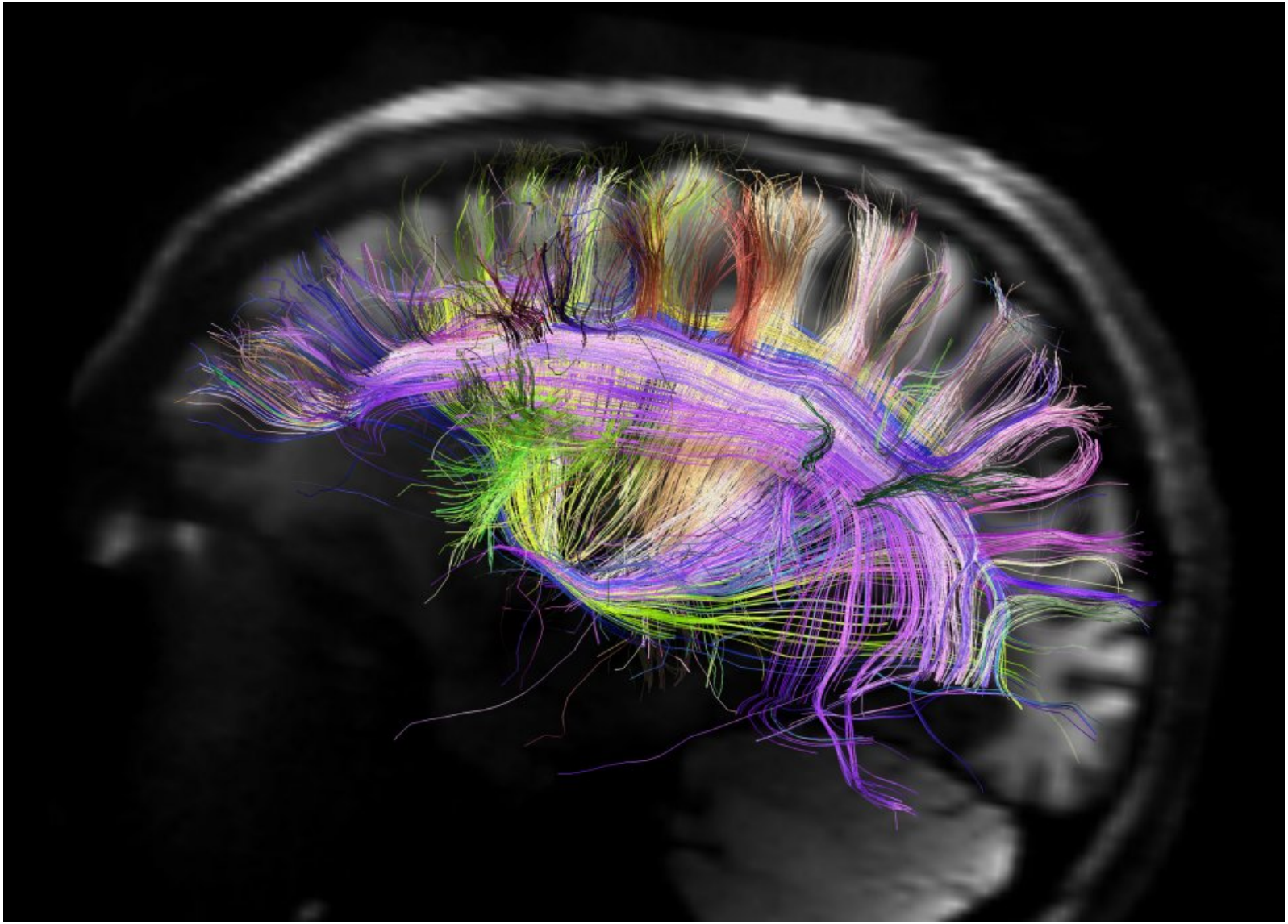
Residual standard error: 5.537 on 8 degrees of freedom
Multiple R-squared: 0.1239
F-statistic: 1.132 on 1 and 8 DF, p-value: 0.3184

interpretazione della retta di regressione

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.1657	6.9594	4.62	0.0017
insulina	-0.3213	0.3020	-1.06	0.3184

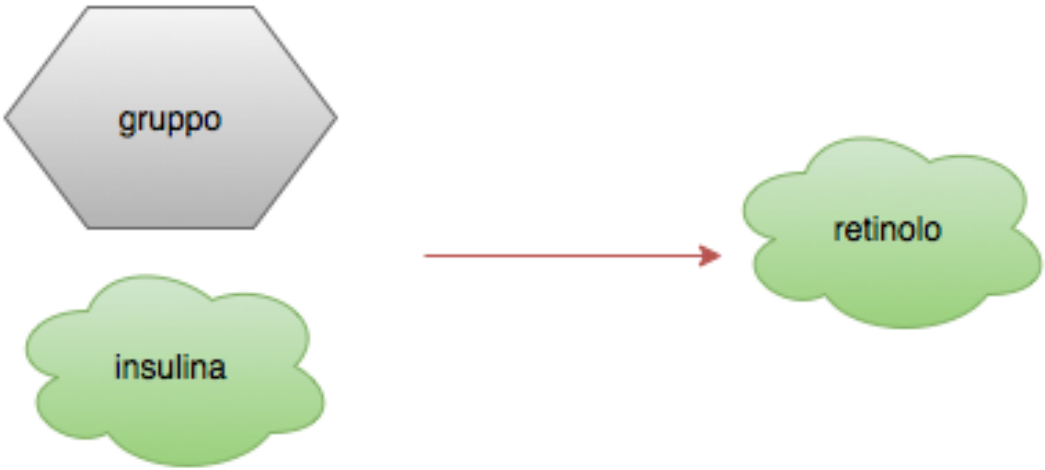






quali modelli statistici (3)

- il **modello lineare**
  - la ancova (analisi della co-varianza)



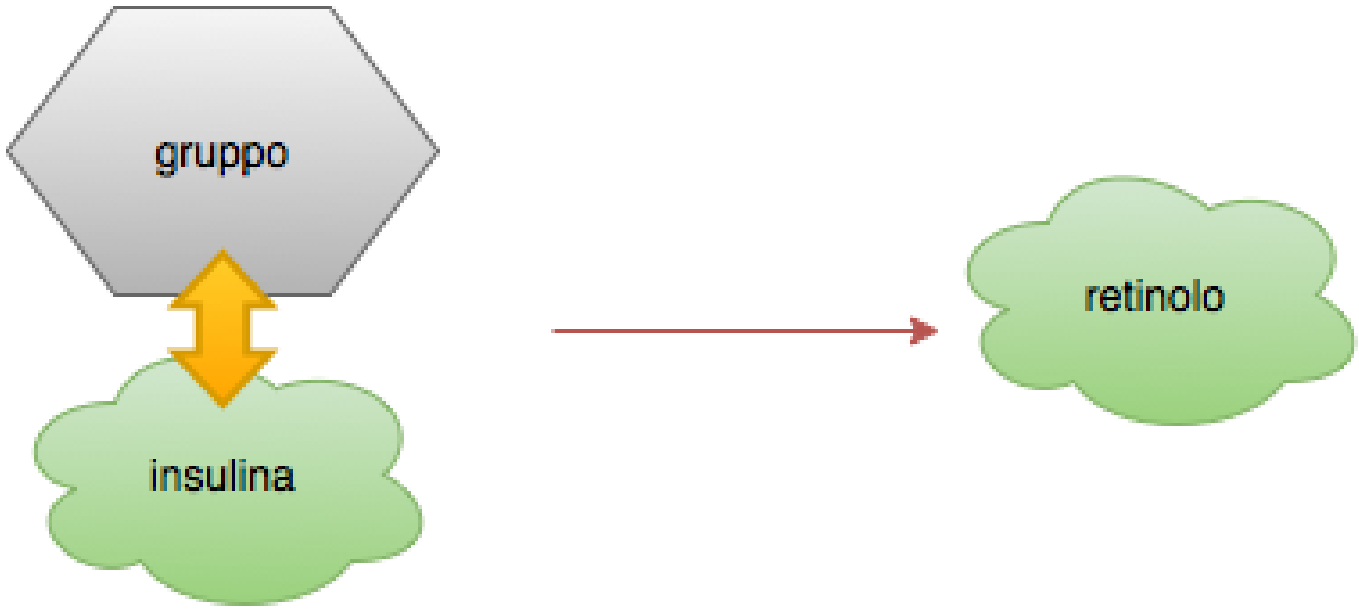
→ input: una variabile aleatoria binomiale

→ input: una variabile aleatoria gaussiana

→ output: una variabile aleatoria gaussiana

prima possibilità

Ancova with interaction, modello moltiplicativo $*$
le due variabili di input interagiscono tra di loro



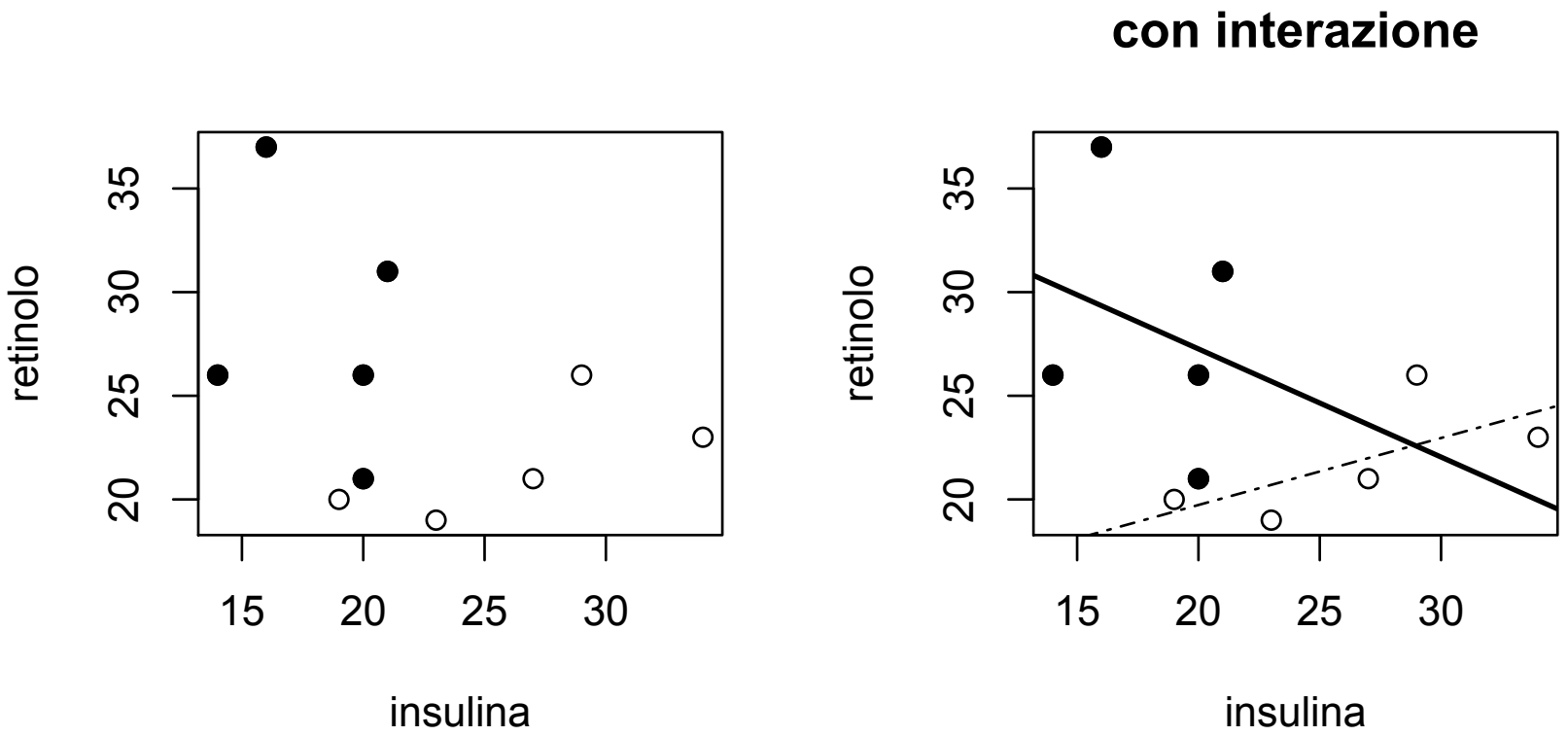
`retinolo` $\sim$ `1 + insulina + gruppo + insulina:gruppo`

prima possibilità

Ancova with interaction

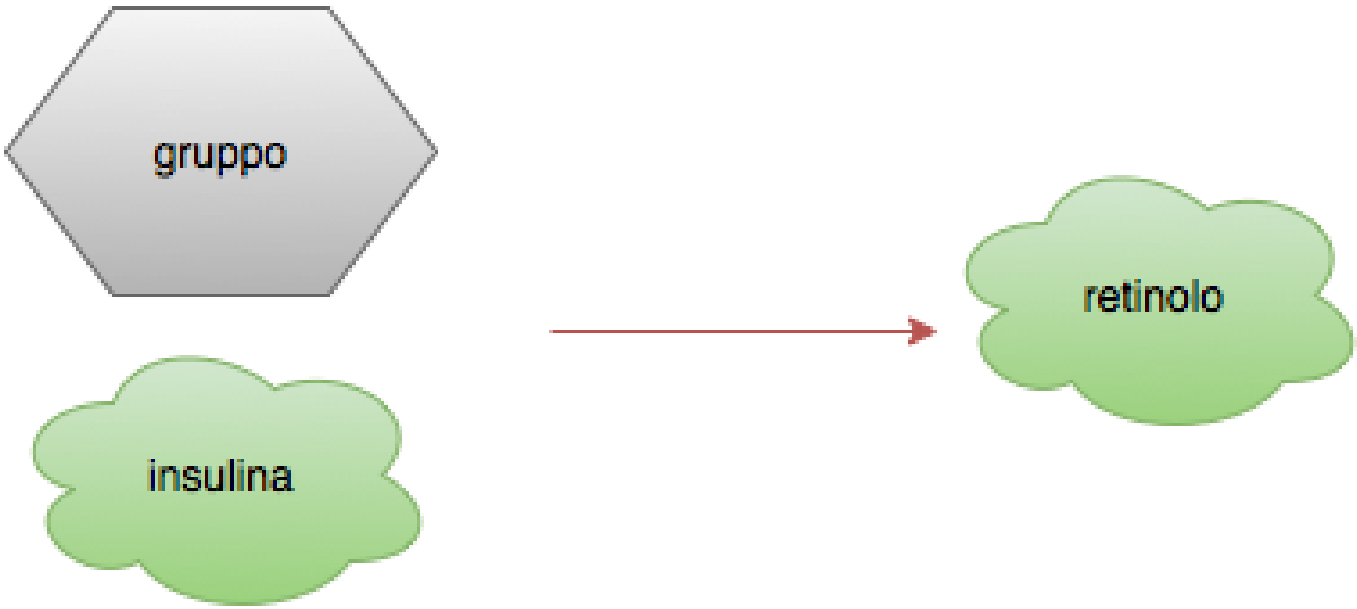
le due variabili di input interagiscono tra di loro

`retinolo` $\sim$ `1 + insulina + gruppo + insulina:gruppo`



seconda possibilità

Ancova without interaction, modello additivo +
le due variabili di input non interagiscono tra di loro



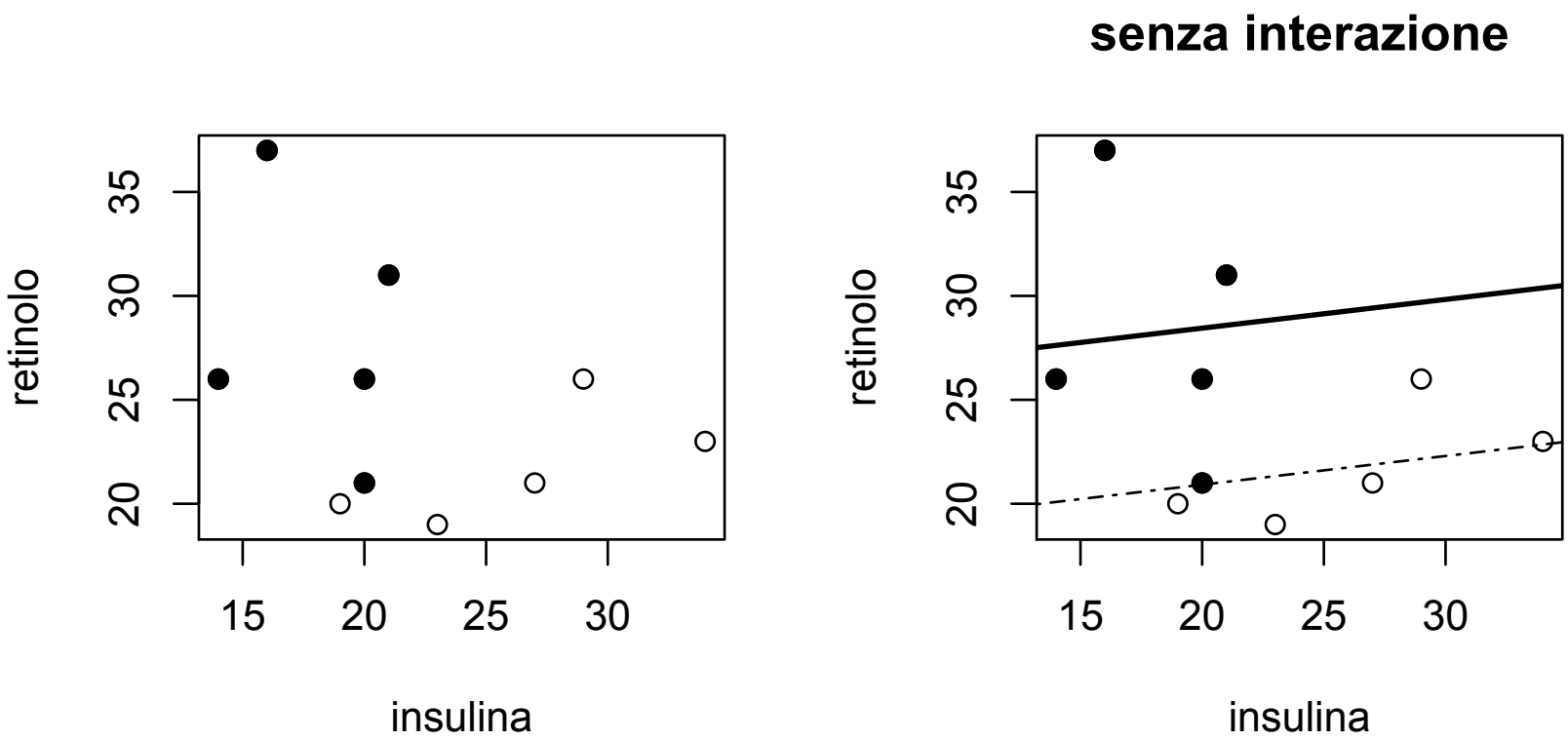
`retinolo` $\sim$ `1 + insulina + gruppo`

seconda possibilità

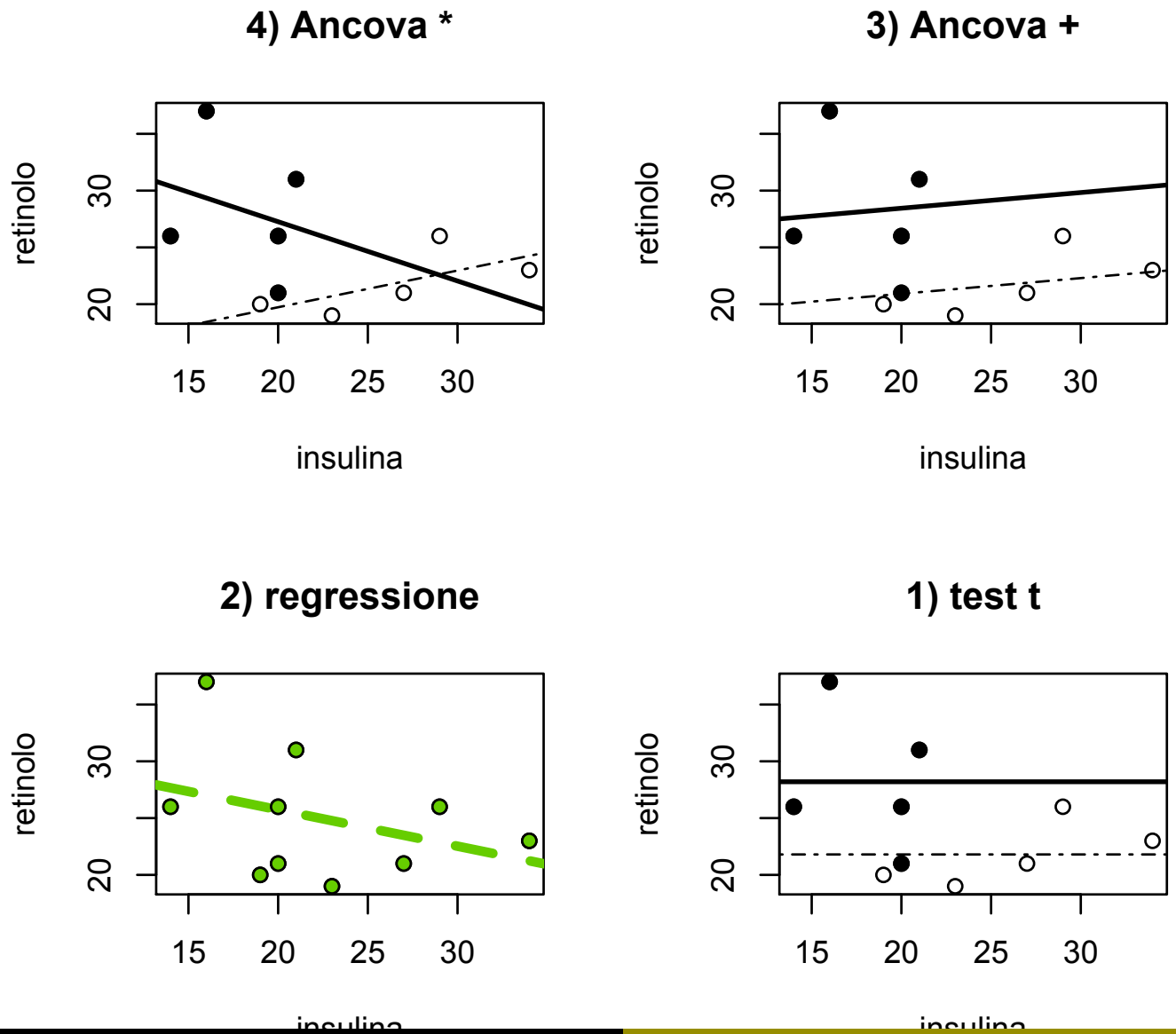
Ancova with interaction

le due variabili di input non interagiscono tra di loro

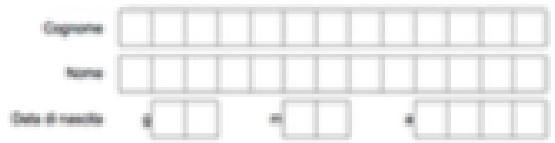
`retinolo` $\sim$ `1 + insulina + gruppo`



decisione? qual è il modello 'giusto'?



dall'esercizio alla dura realtà



1 Selezione del modello tramite p-values

1.1 modello4: Ancova con interazione

Le due rette hanno pendenze significativamente diverse? ____
Il modello 4 è il modello minimale adeguato? ____

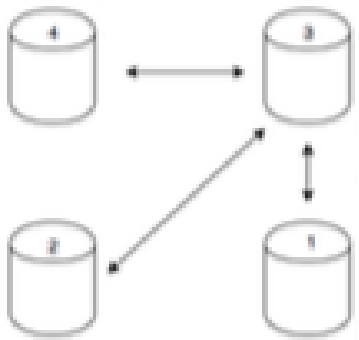
1.2 modello3: Ancova senza interazione

Le due rette hanno pendenze significativamente diverse? ____
Il modello 3 è il modello minimale adeguato? ____

1.3 modello1: t test, differenza tra gruppi

Le due rette hanno quote significativamente diverse? ____
Il modello 1 è il modello minimale adeguato? ____

2 Selezione tramite devianza / informazione



il dataset

- il **modello lineare generalizzato**
  - il test di Fisher
  - la regressione logistica

	logHE4	MENOPAUSA	OUTCOME
1	8.18	PRE	BENIGNO
2	8.02	PRE	BENIGNO
3	10.29	POST	MALIGNO
4	8.75	POST	MALIGNO
5	8.17	POST	BENIGNO
6	8.31	PRE	BENIGNO
..	..	..	..
209	8.67	POST	BENIGNO
210	8.56	POST	MALIGNO

quali modelli statistici (4)

- il **modello lineare generalizzato**
  - il test esatto di Fisher

→ input: una variabile aleatoria binomiale

→ output: una variabile aleatoria binomiale

il test esatto di Fisher inteso come modello lineare generalizzato

	BENIGNO	MALIGNO
POST	65	27
PRE	106	12

il test esatto di Fisher inteso come modello lineare generalizzato

```
modello = glm(OUTCOME ~ 1 + MENOPAUSA, family =
binomial))
summary(modello)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8786	0.2290	-3.84	0.0001
MENOPAUSAPRE	-1.3000	0.3810	-3.41	0.0006

quali modelli statistici (5)

- il **modello lineare generalizzato**
  - la regressione logistica

→ input: una variabile aleatoria gaussiana

→ output: una variabile aleatoria binomiale

la regressione logistica intesa come modello lineare generalizzato

```
modello = glm(OUTCOME ~ 1 + logHE4, family =
binomial))
summary(modello)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-28.3658	4.9976	-5.68	0.0000
logHE4	3.0618	0.5697	5.37	0.0000

contatti

Massimo Borelli
borelli@units.it

www.dmi.units.it/borelli/