Introduzione alla chemiometria e disegno sperimentale

Modulo 6: Metodi di regressione in R software

Docente: Dr. Sabina Licen (slicen@units.it)

Regressione lineare bivariata in R

Model<-lm(vettoreY~vettoreX)

```
> attributes(Model)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"
```

Model$coefficients → Vettore contenente valori di intercetta e pendenza

Model$fitted.values → Vettore contenente i valori modellati di Y ($\hat{Y}$)

Model$residuals → Vettore contenente i valori dei residui (e)

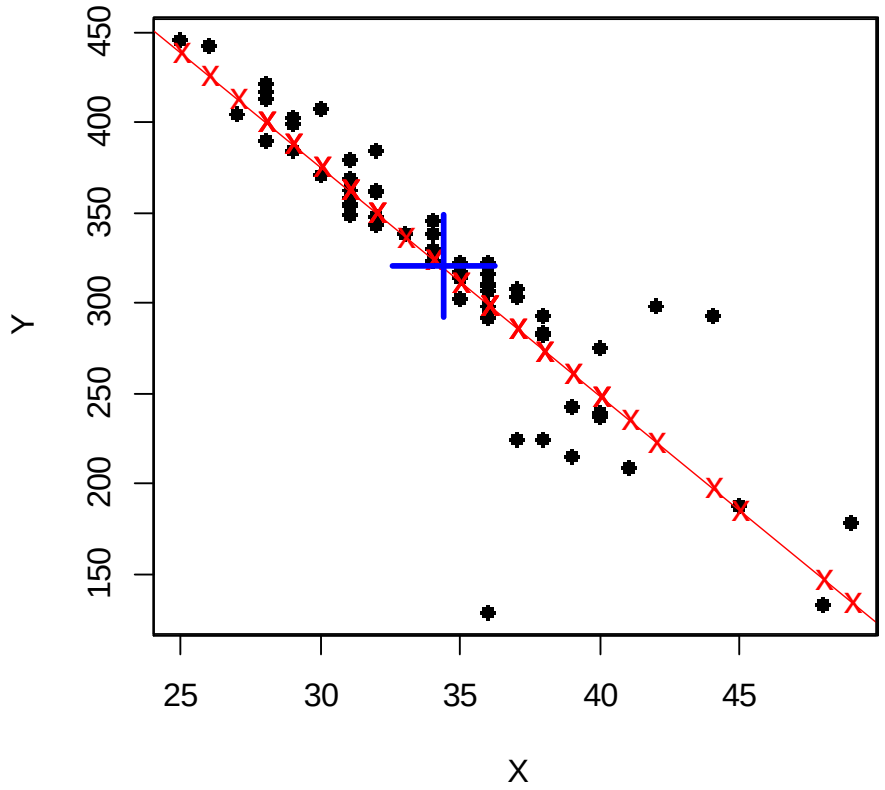
Ymod<-Model$fitted.values

Coef<-Model$coefficients

Res<-Model$residuals

Regressione lineare bivariata in R - grafico

```
plot(originale$X, originale$Y, type="n")
points(originale$X, originale$Y, pch=16,cex=1)
points(originale$X,Ymod,pch="x",cex=1,col="red")
abline(a=Coef[1],b=Coef[2],col="red")
points(mean(originale$X),mean(originale$Y),pch=3,cex=5,lwd=3,col="blue")
```



Regressione lineare bivariata in R - predizione

```
Prediction<-predict.lm(Model, newdata, interval="prediction",level=0.95)
```

	fit	lwr	upr
1	211.8279	134.03456	289.6213
2	173.7473	94.66046	252.8342
3	161.0538	81.45595	240.6517

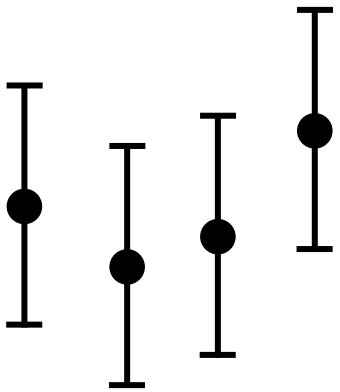
Y predetto per ogni X presente in newdata

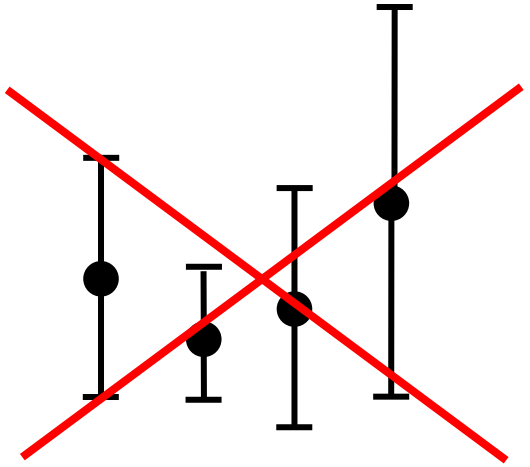
Intervallo di confidenza in predizione per ogni Y predetto

Fonti: https://www.statmethods.net/stats/regression.html; cran.r-project.org › doc › contrib › Frascati-FormularioStatisticaR

Regressione multilineare: assunzioni

- ✓ INDIPENDENZA degli errori (cioè un errore non deve causarne un altro, es. mancata pulizia tra una analisi ed un'altra);
- ✓ Distribuzione NORMALE degli errori;
- ✓ OMOSCEDASTICITA': la varianza degli errori è costante:





olsrr package:
https://cran.r-project.org/web/packages/olsrr/vignettes/heteroskedasticity.html

Regressione multilineare: sequenza applicativa

1) Valutare separatamente le relazioni tra ogni predittore e la variabile dipendente (grafici, correlazioni, etc...);
2) Valutare possibili correlazioni tra i predittori (multicollinearità);
3) Valutare modelli di regressione lineare bivariata dei singoli predittori vs. la variabile dipendente;
4) Utilizzare i predittori non collineari per costruire un modello di regressione multilineare;
5) Valutare tutti i risultati ottenuti;
6) Utilizzare, per scopi di predizione, il miglior modello ottenuto.

Adjusted R^2

gradi di libertà

$$R_{Adj}^2 = 1 - \frac{SSE / (d.f. - 1)}{SST / (n - 1)}$$

numero di osservazioni

Adjusted R^2 corrisponde al coefficiente di determinazione R^2 aggiustato secondo i gradi di libertà (d.f. = n - p).

R^2, per sua formulazione, aumenta se si aggiungono predittori al modello.

Al contrario R_{Adj}^2 che è "pesato" sui gradi di libertà non subisce questo effetto, quindi in regressione multilineare è un migliore indicatore di *best-fit* rispetto a R^2.

Inoltre la diversità tra R^2 e R_{Adj}^2 può dare quindi una indicazione sul reale effetto che ha l'aggiunta di un predittore sulla qualità del modello. Più i valori sono lontani più è peggiorativa l'aggiunta di un predittore sulla qualità del modello (quindi la possibilità di *overfitting*)

VIF – Variance Inflation Factor

E' un parametro che viene calcolato per consentire di capire se nel modello esistono problemi di multicollinearità tra variabili indipendenti (predittori).

Viene calcolato per ogni predittore, secondo una regola empirica, se è > 10 c'è una probabilità molto alta di multicollinearità. Se è > 5 bisogna verificare le correlazioni tra i predittori e decidere se usarli o meno.

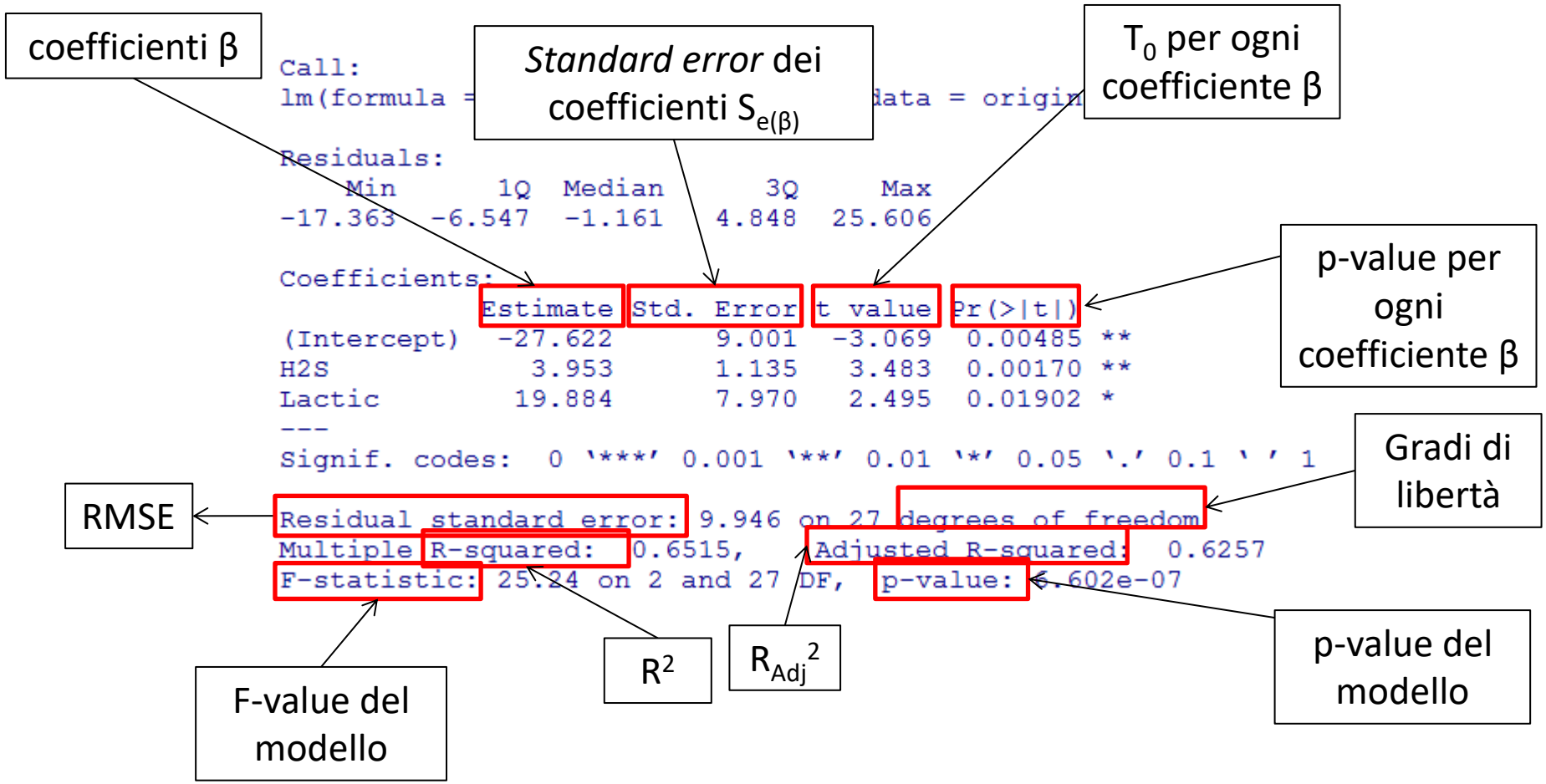
In R:

```
library(car)
vif(Model)
```

Regressione multilineare in R

Model<-lm(vettoreY~vettoreX1+vettoreX2+vettoreX3)

summary.lm(Model)



Regressione multilineare: conclusioni

Il modello ***best-fit*** viene scelto in base a:

- Attinenza del valore dei coefficienti dei predittori con la "realtà" che si sta indagando;
- Alti valori di R^2 e di R_{Adj}^2 (e non troppo dissimili);
- Alto valore di F-value;
- Basso valore di p-value (almeno <0.05, meglio se < 0.01);
- Bassi valori di VIF per i predittori

Attenzione!!!Non dimenticare di confrontare il modello multilineare con i modelli bivariati!!! Uno di questi ultimi potrebbe essere il migliore di tutti!!!