

**UNIVERSITÀ DEGLI STUDI DI SALERNO**



**Dipartimento di Ingegneria dell'Informazione ed  
Elettrica e Matematica applicata**

**Corso di Laurea Magistrale in Ingegneria Informatica**

**APPUNTI DI DATA SCIENCE  
DI FRANCESCO PIO CIRILLO**

<https://github.com/francescopiocirillo>



"Sii sempre forte"

😊 Ehi, un attimo prima di iniziare!

Hai appena aperto una raccolta di appunti che ho deciso di condividere **gratuitamente** su GitHub, se ti sono utili fai **una buona azione digitale**:

-  **Lascia una stellina alla repo:** è gratis, indolore e fa super piacere!
-  **Condividerla con amici**, compagni di corso, o chiunque possa averne bisogno.

Insomma, se questi appunti ti salvano anche solo una giornata di studio... fammelo sapere con una **stellina!**

Grazie di cuore ❤

# 2 - Linear Regression B

## Multiple Linear Regression

La regressione lineare multipla può accomodare più di un preditore.

Supponiamo di avere un vettore di input  $X^T = (X_1, \dots, X_p)$  e vogliamo predire l'output  $Y$ . Allora il modello di regressione lineare prende la forma:

$$Y = f(X) + \epsilon = \beta_0 + \sum_{j=1}^p \beta_j X_j + \epsilon.$$



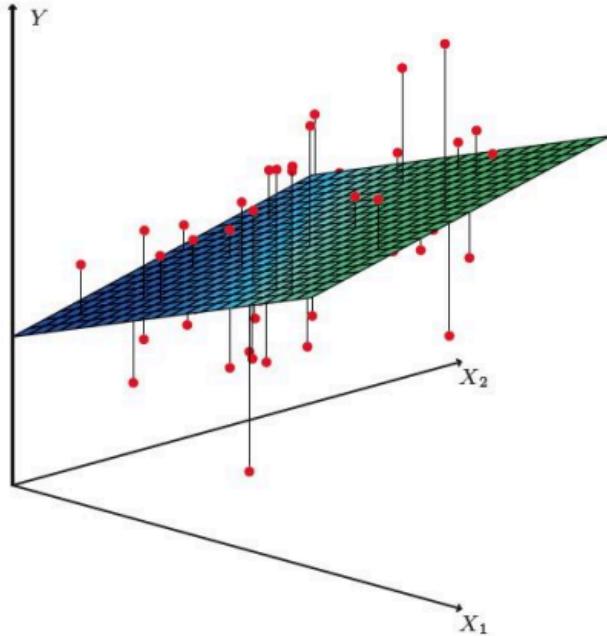
Questo modello è molto semplice da interpretare, ad esempio  $\beta_j$  è l'aumento medio in  $Y$  quando  $X_j$  aumenta di un'unità mantenendo tutti gli altri regressori costanti.

## Least squares

Dati i dati di training  $D_{tr} = \{(x_i, y_i)\}_{i=1}^n$  possiamo selezionare:

$\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$  per minimizzare la residual sum of squares (somma residua dei quadrati):

$$\text{RSS}(\beta) = \sum_{i=1}^n (y_i - f(x_i))^2$$



### Least squares estimates

Considerata  $\mathbf{X}$  la **design matrix** di dimensione  $n \times (p + 1)$  con la  $i$ -esima riga  $(1, x_{i1}, \dots, x_{ip})$  e  $\mathbf{y} = (y_1, \dots, y_n)^T$  allora:

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta).$$

Differenziando rispetto a  $\beta$  otteniamo:

- Gradiente

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta)$$

- Matrice Hessiana

$$\frac{\partial^2 \text{RSS}}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T \mathbf{X}.$$

Assumendo un full column rank (rango pieno di colonna) su  $\mathbf{X}$  e ponendo la derivata prima uguale a 0, otteniamo lo stimatore unbiased di  $\beta$ :

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

I valori fitted rispetto ai training input sono:

$$\hat{\mathbf{y}} = \mathbf{X} \quad \hat{\beta} = \mathbf{X} \quad (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

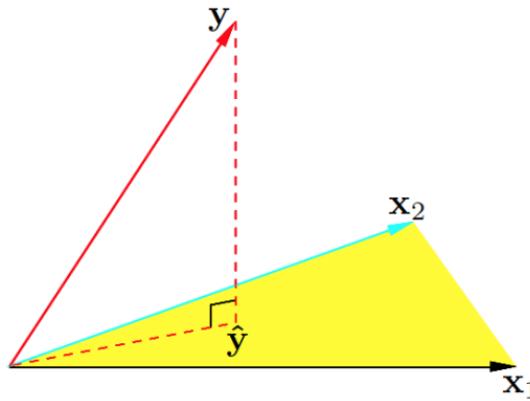
è un'unica equazione

Dove  $H = X(X^T X)^{-1} X^T$  è la Hat Matrix, la matrice cappello, che infatti permette di ottenere  $\hat{y}$  da  $y$ .

### Dimostrazione formula del calcolo di Beta cappello

#### Dimostrazione formula del calcolo di Beta cappello

#### Una vista geometrica dei minimi quadrati



Geometria N-dimensionale della regressione least squares con due predittori.

Il vettore

$y$  risultante è proiettato ortogonalmente sull'hyperplane spanned dai vettori di input  $x_1$  e  $x_2$ . La proiezione  $\hat{y}$  rappresenta il vettore delle predizioni least squares.

### Varianza-Covarianza di uno stimatore least-squares

Il modello lineare per i dati di training è:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n.$$

Per derivare le proprietà di sampling (campionamento) di  $\hat{\beta}$  assumiamo che l'input vector  $x_i$  è fatto da elementi non-random e assumiamo anche che le  $\epsilon_i$  sono indipendentemente distribuite e hanno  $E[\epsilon_i] = 0$  e  $Var(\epsilon_i) = \sigma^2$ .

Detto tutto questo:

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$$

Variance-Covariance  
matrix of the estimator

- ❓ Quando il determinante è piccolo i valori dell'inversa diventano troppo grandi. La varianza cresce.

Una matrice è singolare (non invertibile) quando ci sono righe e colonne proporzionali, i regressori diventano proporzionali.

Aumentare la varianza dello stimatore lo rende impreciso, questo diventa un problema con alta dimensionalità.

Lo stimatore unbiased della varianza  $\sigma^2$  è:

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \text{RSS} = \frac{1}{n - p - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

```
s2 ← sum((Y - X%*%B)^2)/(nrow(X) - ncol(X)) #calcolo della sigma quadro
```

L'esigenza di un modello economico aiuta anche a trovare matrici che sono facilmente invertibili, visto che si considerano meno regressori ( $p$ ), quindi il denominatore cresce e la  $\hat{\sigma}^2$  decresce.

## Alcune domande importanti

- **Almeno uno** dei predittori  $X_1, \dots, X_p$  è **utile nella predizione** della risposta?
- Tutti i predittori spiegano  $Y$  o **solo un sottoinsieme dei predittori è utile?**
- Dato un set di valori di predittori, quale valore di risposta dovremmo prevedere, e **quanto accurata è la nostra predizione?**

Per rispondere a queste domande dobbiamo assumere che il modello lineare è il modello di popolazione corretto, e che:

$$\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

Assunto questo si potrebbe dimostrare (ma non lo dimostriamo) che la distribuzione di  $\hat{\beta}$  è una Normale, ovviamente  $\hat{\beta}$  è un vettore e quindi la distribuzione sarà una MVN (p+1 sono le dimensioni) con media  $\beta$  vero e varianza la matrice var-covar.

Il problema è che come al solito la  $\hat{\sigma}$  non l'abbiamo mai, si dimostra che anche lo stimatore di  $\hat{\sigma}$  a meno di coefficienti è una chi quadratica con  $n-p-1$  gradi di libertà. La formula per la stima di sigma sarà

sempre uguale.

Cioè:

$$\hat{\beta} \sim N_{p+1}(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2) \quad \text{chi-squared distribution}$$
$$(n - p - 1)\hat{\sigma}^2 \sim \sigma^2 \chi_{n-p-1}^2, \quad \leftarrow \text{with } n - p - 1 \text{ degrees of freedom}$$

and  $\hat{\beta}$  and  $\hat{\sigma}^2$  are statistically independent.

Queste proprietà distribuzionali sono usate per formare i test di ipotesi e gli intervalli di confidenza per i parametri.

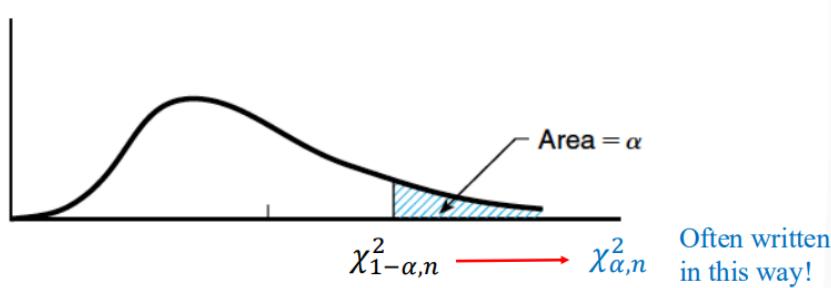
### Definizione della distribuzione Chi-squared

Se  $Z_1, Z_2, \dots, Z_n$  sono variabili aleatorie indipendenti le cui distribuzioni sono normali standard, allora si dice che  $X$ , definita da:

$$X = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

ha una distribuzione Chi-squared con  $n$  gradi di libertà. Si usa per esprimere questo concetto la notazione:

$$X \sim \chi_n^2$$



Quello a destra in azzurro è il **valore soglia**, non il quantile, ma è comunque molto usato.

### Uno specifico predittore $X_j$ è importante? Test di ipotesi

Potremmo rispondere a questa domanda con un test di ipotesi:

$$H_0 : \beta_j = 0, \text{ vs. } H_a : \beta_j \neq 0$$

Calcoliamo la **t-statistic**

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{v_j}}$$

dove  $v_j$  è il j-esimo elemento della diagonale di  $(X^T X)^{-1}$  che è la matrice di var-covar.



Sotto l'ipotesi  $H_0$ ,  $t_j$  segue una  $t$  distribution con  $n - p - 1$  gradi di libertà, quindi il p-value è definito come  $Pr(|t| > |t_j|)$ .

Un p-value minore di 0.05 porta a rigettare  $H_0$  stabilito il livello di significatività 95%.

Quando n è grande, abbiamo:

$$t_{1-\frac{\alpha}{2}, n-p-1} \cong Z_{1-\frac{\alpha}{2}}$$

### Test per la significanza di un gruppo di coefficienti

Supponiamo di testare un sottoinsieme  $q$  dei coefficienti:

$$H_0 : \beta_1 = \dots = \beta_q = 0 \text{ vs. } H_a : \text{at least one is non-zero}$$

Essendo un test diverso la statistica di test cambia e diventa la F-statistic

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)},$$

il denominatore è  $\hat{\sigma}^2$

dove  $RSS_0$  è la residual sum of squares per il modello nell'ipotesi  $H_0$ .

Questo test è poco utilizzato, si fa più spesso il test per una variabile per volta.

SALTARE A ESEMPIO DATI PUBBLICITARI.

Sotto l'ipotesi  $H_0$ ,  $F$  segue una distribuzione  $F_{q, n-p-1}$ , allora

$Pr(F'' > F') = p$  è il p-value, essendo  $F$  i valori (retrieved) recuperati dai dati.

$F_{q, n-p-1}$  è la distribuzione Fisher-Snedecor o **F distribution** con parametri  $q$  e  $n - p - 1$ .

Un p-value inferiore a 0.05 porta a rigettare  $H_0$  al livello di significatività 95%.

### F-statistic (va saltata)

**STIC**

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n - p - 1)}$$

By constraining  $p_1 - p_0$  params. to zero:

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(n - p_1 - 1)}$$

Si può dimostrare che, se le assunzioni del modello lineare sono corrette allora:

$$E\{\text{RSS}/(n - p - 1)\} = \sigma^2$$

- Assumendo  $H_0$  vero, allora:

$$E\{(\text{RSS}_0 - \text{RSS})/q\} = \sigma^2$$

Quindi non c'è relazione tra risposta e predittori, ci si aspetta quindi che la F-statistic assuma un valore vicino ad 1.

- D'altro canto, se  $H_a$  è vero allora:

$$E\{(\text{RSS}_0 - \text{RSS})/q\} > \sigma^2$$

quindi ci si aspetta una F-statistic maggiore di 1.



Se F è alto (nell'esempio pubblicitario F=570 che è un valore molto alto) sotto  $H_0$ , per il quale tutti i coefficienti sono uguali a 0 (la pubblicità non influenza le vendite), è da considerare che  $H_0$  sia molto improbabile (sarebbe vera solo se tutti i coefficienti fossero uguali a 0 per puro caso) e quindi bisognerebbe rigettarla.

## F distribution

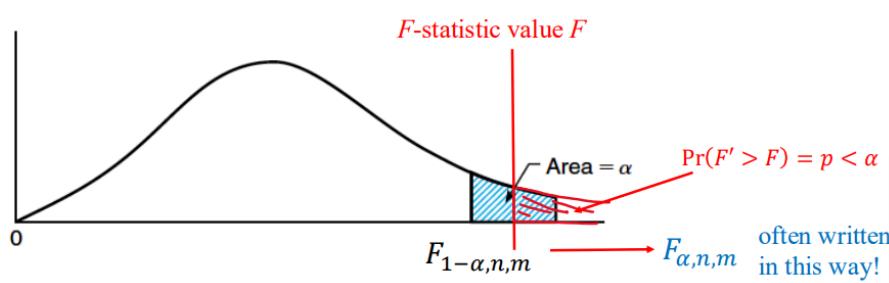
Se  $\chi_n^2$  e  $\chi_m^2$  sono variabili aleatorie con distribuzioni chi-squared indipendenti con rispettivamente n ed m gradi di libertà, allora si dice che la variabile aleatoria  $F_{n,m}$  definita da:

$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$$

ha una F-distribution con n ed m gradi di libertà.

Per un qualsiasi  $\alpha \in (0, 1)$ , sia  $F_{\alpha,n,m}$  tale che:

$$P\{F_{n,m} > F_{\alpha,n,m}\} = \alpha$$



## Esempio: dati pubblicitari — simple linear regression

Simple regression of **sales** on **TV**

	Coefficient	Std. error	t-statistic	p-value
<b>Intercept</b>	7.0325	0.4578	15.36	< 0.0001
<b>TV</b>	0.0475	0.0027	17.67	< 0.0001

Simple regression of **sales** on **radio**

	Coefficient	Std. error	t-statistic	p-value
<b>Intercept</b>	9.312	0.563	16.54	< 0.0001
<b>radio</b>	0.203	0.020	9.92	< 0.0001

Simple regression of **sales** on **newspaper**

	Coefficient	Std. error	t-statistic	p-value
<b>Intercept</b>	12.351	0.621	19.88	< 0.0001
<b>newspaper</b>	0.055	0.017	3.30	0.00115

## Esempio: dati pubblicitari — multiple linear regression

Multiple linear regression:  $\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon$

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

TABLE 3.5. Correlation matrix for TV, radio, newspaper, and sales for the Advertising data.

Quantity	Value
Residual standard error	1.69
R <sup>2</sup>	0.897
F-statistic	570

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)} \sim F_{p, n-p-1}$$

36

## Variable selection in linear regression — da qua si riprende



La **F-statistic** è una delle tecniche adottate per la selezione dei predittori che sono associati con la risposta.



Il task di determinare quei predittori, al fine di fare il fit di un singolo modello comprendente solo quelle variabili indipendenti, è detto **variable selection**.

La variable selection è svolta principalmente per **due ragioni**:

- **Prediction accuracy:** le stime a minimi quadrati hanno spesso piccolo bias ma grande varianza. L'accuratezza della previsione può talvolta essere migliorata attraverso lo shrinking di alcuni coefficienti, o anche settando a zero alcuni coefficienti;
- **Interpretazione:** con un numero grande di predittori, spesso sarebbe conveniente determinare un sottoinsieme più piccolo che mostri l'effetto più forte. Al fin di ottenere la "visione di insieme" siamo disposti a sacrificare alcuni dei piccoli dettagli.

## R<sup>2</sup> e RSE



Si ricorda che nelle regressioni lineari **semplici**  $R^2$  è il quadrato della correlazione tra la risposta e la variabile.

Nelle regressioni lineari **multiple**,  $R^2 = Cor(Y, \hat{Y})^2$ , cioè  $R^2$  corrisponde al quadrato della correlazione tra la risposta e il modello lineare adattato (fitted linear model).



Il modello lineare adattato massimizza la correlazione tra tutti i possibili modelli lineari.

Un  $R^2$  vicino ad 1 indica che il modello "spiega" una grande porzione della varianza nella risposta della variabile.



Il valore  $R^2$  aumenterà **sempre** quando vengono aggiunte più variabili al modello, anche se sono associate alla risposta molto debolmente.

La  $R^2$  statistic è calcolata sui dati di training ed indica un miglior fit su quei dati ma non necessariamente sui dati di test.

## Esempio advertising

### Valutazioni su $R^2$

Nell'esempio Advertising il modello che usa tutti e tre gli advertising media per predire le vendite ha un  $R^2$  di 0.8972, il modello che usa invece solo TV e radio ha un  $R^2$  di 0.89719, in altre parole l'inclusione di newspaper nel modello che contiene TV e radio implica una crescita molto piccola di  $R^2$ , il  $p$ -value per newspaper non è significativo.

Si conclude che newspaper può essere rimosso dal modello, questo aiuta anche ad evitare l'**overfitting**.

Un modello contenente solo TV come predittore ha un  $R^2$  di 0.61, quindi l'aggiunta di radio è rilevante, potremmo quantificare il miglioramento in questione osservando il  $p$ -value per il coefficiente radio in questo modello contenente TV e radio come predittori.

### Valutazioni sull'RSE

In termini di RSE notiamo che:

- il modello contenente solo TV e radio ha RSE = 1.68;
- il modello contenente solo TV, radio e newspaper ha RSE = 1.69;
- il modello contenente solo TV ha RSE = 3.26.

Questo ulteriormente corrobora la nostra conclusione che un modello che usa la spesa pubblicitaria in TV e radio per predire le vendite è molto più accurato di uno che usa solo la spesa in TV. Non ha senso introdurre newspaper viste le conclusioni precedenti.

## La crescita dell'RSE

Come è possibile che l'RSE cresce aggiungendo newspaper visto che l'RSS **dove** decrescere per forza e si trova a numeratore della formula dell'RSE?

In generale l'RSE è definito come:

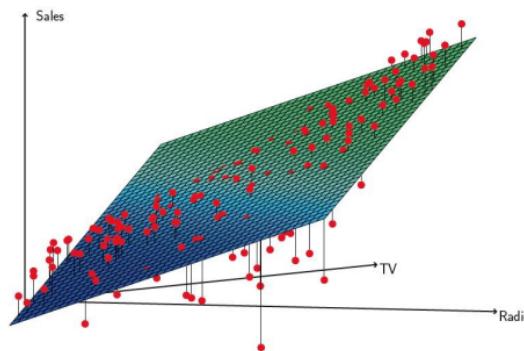
$$RSE = \sqrt{\frac{1}{n-p-1} RSS},$$

e i modelli con più variabili possono avere un RSE più alto se la decrescita di RSS è più piccola della crescita di p.

## Graphical summaries

Oltre all'osservazione dell'RSE e della  $R^2$  statistics può essere utile fare il plot dei dati (se possibile). I Graphical summaries possono rivelare problemi con un modello che sono non visibili dalle statistiche numeriche.

Nel plot 3D di TV e radio vs sales alcune osservazioni sono al di sopra e altre al di sotto del piano di regressione a minimi quadrati (least squares regression plane).



Il modello lineare sembra sovrastimare le vendite quando il budget pubblicitario è speso esclusivamente su TV o radio, mentre sembra sottostimare le vendite quando il budget è diviso tra i due mezzi pubblicitari.

Si nota un effetto di interazione tra i mezzi pubblicitari presenti, la combinazione di diversi mezzi risulta in un aumento di vendite più grande rispetto all'utilizzo di un unico mezzo.

## Confidence and Prediction Interval

Dopo aver fatto il fit di un modello di regressione multipla è possibile predire Y sulla base di un insieme di valori per i predittori  $X_1, X_2, \dots, X_p$ .

Questa previsione è però associata a tre diversi tipi di incertezza.

### Primo tipo di incertezza (non trovi il vero piano, ci vai vicino)

Gli stimatori dei coefficienti  $\hat{\beta}_0, \dots, \hat{\beta}_p$  sono stimatori di  $\beta_0, \dots, \beta_p$ .

Cioè, il least squares plane (il piano a minimi quadrati):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

è solo una stima del vero population regression plane (piano di regressione della popolazione):

$$f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

L'inaccuratezza nella stima dei coefficienti è legata all'errore **riducibile**.

$$E(y_0 - \hat{f}(x_0))^2 = \underbrace{[\text{Bias}(\hat{f}(x_0))]^2}_{\text{Reducible}} + \text{Var}(\hat{f}(x_0)) + \text{Var}(\epsilon)$$

Possiamo calcolare l'intervalllo di confidenza per determinare quanto

$\hat{Y} = \hat{f}(x_0)$  sarà vicina a  $f(x_0)$ , per  $X = x_0$  di una nuova osservazione  $(x_0, y_0)$ .

### Secondo tipo di incertezza (la realtà non è un piano)

In pratica, assumere un modello lineare per  $f(X)$  è quasi sempre una approssimazione della realtà, quindi c'è una sorgente di errore potenzialmente riducibile in quello che chiamiamo **model bias**.

Quindi quando usiamo un modello lineare stiamo stimando la migliore approssimazione lineare (hyperplane) della vera superficie, in ogni caso questa discrepanza sarà ignorata e lavoreremo come se il modello lineare fosse quello corretto.

### Terzo tipo di incertezza (i punti non cadono sul piano perché l'errore li sposta)

Anche se conoscessimo  $f(X)$ , cioè se conoscessimo i veri valori di  $\beta_0, \dots, \beta_p$ , il valore della risposta non può essere predetto perfettamente a causa dell'errore casuale  $\epsilon$  nel modello, ci siamo precedentemente riferiti a questo errore con il termine **errore irriducibile**.

Questo errore esprime la domanda: quanto varierà  $Y$  rispetto ad  $\hat{Y}$ ?

Per rispondere a questa domanda si usano gli intervalli di previsione (prediction intervals).

Gli intervalli di predizione sono sempre più ampi (wider) degli intervalli di confidenza in quanto incorporano sia l'errore nella stima dei coefficienti di  $f(X)$  (l'errore riducibile) che l'incertezza relativa a quanto un singolo punto si distanzierà dal piano di regressione della popolazione (l'errore irriducibile).

### Intro Confidence vs Prediction Interval

L'intervalllo di confidenza sui coefficienti sarà un intervallo di confidenza sul modello di regressione.

Ora voglio un intervallo che ci dia la variabilità della  $\hat{Y}$ , cioè della predizione.

L'intervalllo di confidenza è sulle stime dei coefficienti, dove c'è l'errore riducibile, cioè noi vogliamo trovare il compromesso perfetto tra bias e varianza per riuscire ad avvicinarci alla population

regression line il più possibile.

Ma c'è anche un errore irriducibile, la varianza dell'errore, questo errore emerge quando si fanno le predizioni e quindi l'intervallo di Predizione sarà più ampio.

## Confidence Interval

Come in un caso di regressione semplice, bisogna fornire un modello della distribuzione del termine di errore  $\epsilon$  al fine di determinare gli intervalli di confidenza. Di conseguenza assumiamo:

$$\epsilon \sim N(0, \sigma^2)$$

Dato un nuovo vettore di input  $x_0 = (x_{01}, \dots, x_{0p})^T$ , vogliamo calcolare l'intervallo di confidenza per:

$f(x_0) = E(Y|X = x_0) = \mathbf{x}_{new}^T \beta$ , where  
 $\mathbf{x}_{new} = (1, x_{01}, \dots, x_{0p})^T$ . Then,

$$\mathbf{x}_{new}^T \hat{\beta} \sim N(\mathbf{x}_{new}^T \beta, \mathbf{x}_{new}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{new} \sigma^2)$$

Un intervallo di confidenza  $1 - \alpha$  per  $f(x_0)$  è:

$$\mathbf{x}_{new}^T \hat{\beta} \pm z_{1-\alpha/2} \sqrt{\mathbf{x}_{new}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{new} \hat{\sigma}^2}$$

Dove  $z_{1-\alpha/2}$  è l'1-( $\alpha/2$ ) percentile della distribuzione normale standard (ad esempio  $z_{1-0.05/2} = 1.96$ ).

## Prediction Interval

Per calcolare l'intervallo di previsione assumiamo che

$$\epsilon \sim N(0, \sigma^2)$$

sia indipendente da  $x$ .

Dato un nuovo vettore di input  $x_0 = (x_{01}, \dots, x_{0p})^T$ , vogliamo calcolare l'intervallo di previsione per la vera risposta  $y_0$  associata ad  $x_0$ .

Si nota che:

$$y_0 = f(x_0) + \epsilon = \mathbf{x}_{new}^T \beta + \epsilon$$

dove

$$\mathbf{x}_{new} = (1, x_{01}, \dots, x_{0p})^T$$

Allora:

$$\mathbf{x}_{new}^T \hat{\beta} + \epsilon \sim N(\mathbf{x}_{new}^T \beta, \mathbf{x}_{new}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{new} \sigma^2 + \sigma^2)$$

Un intervallo di previsione  $1 - \alpha$  per  $y_0$  è:

$$\mathbf{x}_{new}^T \hat{\beta} \pm z_{1-\alpha/2} \sqrt{(1 + \mathbf{x}_{new}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{new}) \hat{\sigma}^2}$$

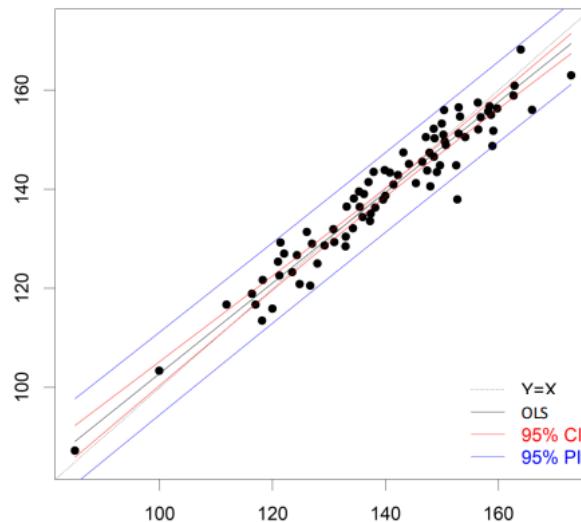
Dove  $z_{1-\alpha/2}$  è l'1-( $\alpha/2$ ) percentile della distribuzione normale standard (ad esempio  $z_{1-0.05/2} = 1.96$ ).

### Confidence and Prediction Interval (simple linear regression)

Nel caso di una regressione lineare semplice, gli intervalli di confidenza e previsione sono (quando n è non troppo grande, ad esempio  $n \leq 30$ ):

$$1 - \alpha \text{ confidence interval: } \hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{1-\frac{\alpha}{2}, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}}$$

$$1 - \alpha \text{ prediction interval: } \hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{1-\frac{\alpha}{2}, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}}$$



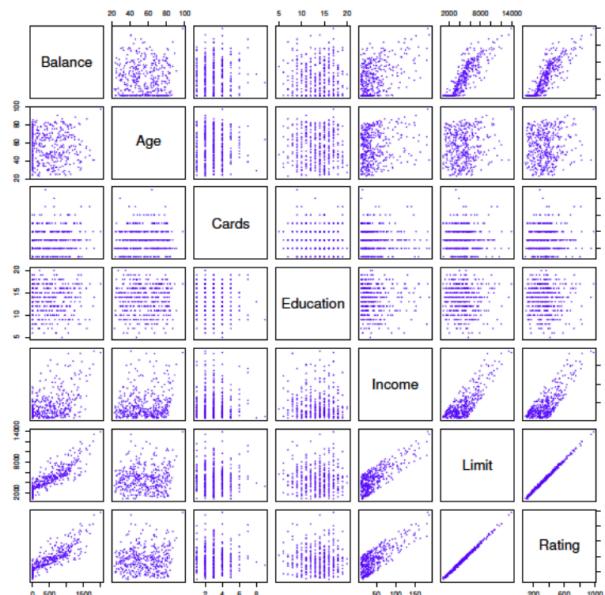
### Diverse provenienze dei regressori

Il modello lineare è relativamente flessibile nel senso che  $X_j$  può venire da diverse fonti:

- trasformazione di input quantitativi, come il log;
- la base di un'espansione, come  $X_2 = X_1^2$ ,  $X_3 = X_1^3$ , portando ad un fit polinomiale;
- dummy coding dei livelli di input qualitativi;
- interazioni tra variabili, e.g.  $X_3 = X_1 * X_2$ .

## Qualitative Predictors

Alcuni predittori sono non quantitativi ma qualitativi, prendendo un set di valori discreti. I predittori qualitativi sono detti **categorical predictors** o **factor variables**.



Nello scatterplot matrix dell'esempio della carta di credito è possibile vedere oltre a 7 variabili quantitative anche 4 qualitative (gender, student, status, ethnicity).

### Predictors con soli due livelli

Supponiamo di voler investigare delle differenze nel bilancio delle carte di credito tra maschi e femmine, ignorando tutte le altre variabili.

Se un predittore qualitativo (anche noto come factor) ha solo due livelli, o possibili valori, allora incorporarlo nel modello di regressione è molto semplice: basta creare una dummy variable che assume due possibili valori numerici.

Creiamo una nuova variabile:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Il modello risultante è:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Con questa equazione  $\beta_0$  può essere interpretato come il bilancio medio di una carta di credito (ignorando l'effetto del genere) mentre  $\beta_1$  è la quantità che indica quanto le femmine hanno un bilancio più alto, e i maschi più basso, della media.

I risultati per il modello del genere sono i seguenti:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender [Female]	19.73	46.05	0.429	0.6690

### Qualitative Predictors con più di due livelli

Con più di due livelli si crea una dummy variable addizionale. Per esempio per la variabile ethnicity si creano due dummy variables.

La prima potrebbe essere:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

e la seconda:

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Allora entrambe queste variabili possono essere utilizzate nell'equazione di regressione al fine di ottenere il modello.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

Con questa equazione  $\beta_0$  (l'intercetta) può essere interpretata come il bilancio medio di una carta di credito di un Afro-americano,  $\beta_1$  può essere interpretata come la differenza nel bilancio medio tra le categorie Asiatici e Afro-americani, infine  $\beta_2$  può essere interpretata come la difference di bilancio medio tra Caucasici e Afro-americani.



Ci sarà sempre una dummy variable in meno rispetto al numero di livelli.

Il livello con nessuna dummy variable, in questo caso l'etnia African American, è noto come **baseline**.

La ragione per la quale si devono usare più dummy variables quando abbiamo più di due livelli è perché non è detto che ci sia un ordinamento, perché asiatico dovrebbe essere 0 e caucasico 1?

Inoltre mettendoli tutti in serie come valori numerici si imporrebbe che la distanza tra asiatico (0) e caucasico (1) è la metà rispetto a quella tra asiatico (0) e afroamericano (2), questo non ha senso quasi mai.

## Conclusioni

	Coefficient	Std. error	t-statistic	p-value
<b>Intercept</b>	531.00	46.32	11.464	< 0.0001
<b>ethnicity[Asian]</b>	-18.69	65.02	-0.287	0.7740
<b>ethnicity[Caucasian]</b>	-12.50	56.68	-0.221	0.8260

**TABLE 3.8.** Least squares coefficient estimates associated with the regression of balance onto ethnicity in the Credit data set. The linear model is given in (3.30). That is, ethnicity is encoded via two dummy variables (3.28) and (3.29).

I p-values associati alle stime dei coefficienti per le due dummy variables sono molto grandi, il che suggerisce una mancanza di una vera prova statistica di una reale differenza nel bilancio delle carte di credito sulla base dell'entità.

Invece di affidarsi ai coefficienti individuali è possibile usare un F-test per testare  $H_0 : \beta_1 = \beta_2 = 0$ , questo non dipende dal coding (dalla codifica). Questo F-test ha un p-value di 0.96, indicando che non possiamo rigettare la null hypothesis.

## Predictori Qualitativi in R

```

Call:
lm(formula = Balance ~ factor(Ethnicity), data = credit)

Residuals:
    Min      1Q   Median      3Q     Max 
-531.00 -457.08 -63.25  339.25 1480.50 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 531.00     46.32 11.464 <2e-16 ***
factor(Ethnicity)Asian -18.69     65.02 -0.287   0.774  
factor(Ethnicity)Caucasian -12.50     56.68 -0.221   0.826  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

## Interactions

### Removing the Additive Assumption

Nella precedente analisi dei dati pubblicitari è stato concluso che sia TV che radio sembrano associati a sales. Il modello lineare che ha posto le basi per questa conclusione partiva dal presupposto che l'effetto su sales dell'aumento di uno dei mezzi pubblicitari è indipendente dalla quantità di denaro spesa per gli altri mezzi.

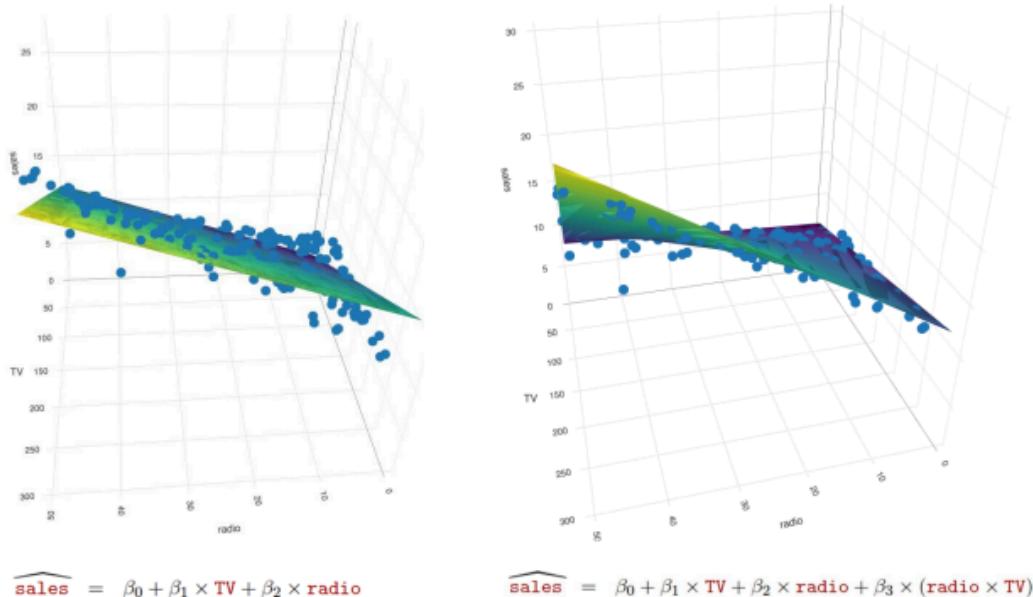
Ad esempio, il modello lineare

$$\widehat{\text{sales}} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

afferma che l'effetto medio su sales di un aumento di una unità in TV è SEMPRE  $\beta_1$ , indipendentemente dalla quantità di denaro spesa in radio.

Comunque, questo semplice modello può essere incorretto. Si può supporre ad esempio che l'aumento di spesa sulla pubblicità radiofonica renda più efficace la pubblicità televisiva. In marketing questo è noto come synergy effect mentre in statistica è noto come interaction effect.

Per questo prima avevamo notato che il modello lineare sembra sovrastimare le vendite quando il budget pubblicitario è speso esclusivamente su TV o radio, mentre sembra sottostimare le vendite quando il budget è diviso tra i due mezzi pubblicitari.



Per rilassare l'assunzione di indipendenza possiamo includere una interazione.

Il modello prende quindi la forma

$$\begin{aligned} \text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{radio} \times \text{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \epsilon. \end{aligned}$$

## Conclusioni

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

```
lm(formula = Sales ~ TV + Radio + TV * Radio, data = ad)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00 2.479e-01 27.233 <2e-16 ***
TV          1.910e-02 1.504e-03 12.699 <2e-16 ***
Radio       2.886e-02 8.905e-03 3.241 0.0014 **
TV:Radio    1.086e-03 5.242e-05 20.727 <2e-16 ***
Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
```

I risultati nella tabella suggeriscono che le interazioni sono importanti, il p-value per l'interazione  $TV \times radio$  è molto basso, il che indica che c'è una prova forte che  $H_A : \beta_3 \neq 0$ .

La  $R^2$  per il modello di interazione è 96.8%, più alto dell'87.9% del modello che non includeva il termine di interazione.



Questo vuol dire che il  $(96.8 - 89.7) / 100 = 6.9\%$  della variabilità in sales che resta dopo aver fatto il fitting del modello additivo viene spiegato dal termine di interazione.

Le stime dei coefficienti nella tabella suggeriscono che un aumento in pubblicità televisiva di 1000 dollari è associato ad un aumento in vendite di:

$$(\hat{\beta}_1 + \hat{\beta}_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio} \text{ units.}$$

### Hierarchy principle

Talvolta un interaction term ha un p-value molto piccolo ma i main effects a esso associati no (succede in questo caso con TV e radio).

Questo ci porta al **principio gerarchico**:



Se includiamo una interazione in un modello dovremmo anche includere gli effetti principali a essa corrispondenti, anche se i p-values associati ai coefficienti degli effetti principali non sono significativi.

La logica dietro questo principio è che le interazioni sono difficili da interpretare in un modello senza i corrispondenti effetti principali, il loro significato viene cambiato.

Specificatamente, il termine di interazione contiene anche i main effects se il modello non ha i termini dei main effects.

### Interazioni tra variabili qualitative e quantitative

Tornando all'esempio del dataset sulle Carte di Credito, immaginiamo di voler predire il valore balance usando income (quantitativo) e student (qualitativo).

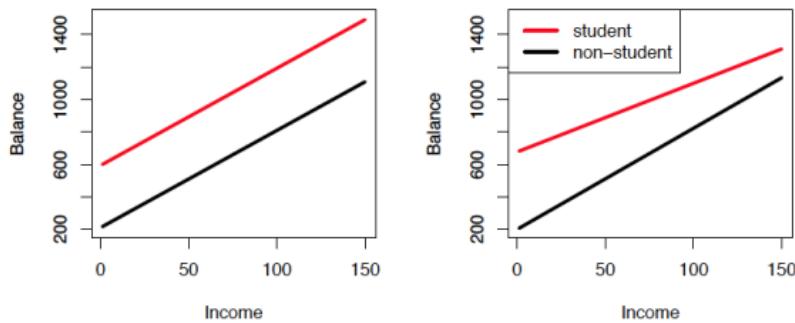
Senza un termine di interazione, il modello prende la forma:

$$\begin{aligned}\text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}\end{aligned}$$

Mentre con l'interazione prende la forma:

$$\begin{aligned}\text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}_i & \text{if not student} \end{cases}\end{aligned}$$

Graficamente:



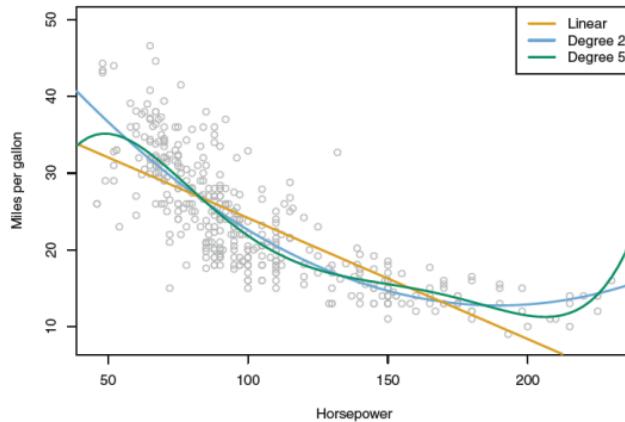
Credit data; Left: no interaction between `income` and `student`.  
Right: with an interaction term between `income` and `student`.

### Effetti non lineari dei predittori

Quando la relazione tra la risposta e i predittori sembra non lineare, il linear model può essere esteso per far fronte alle relazioni non lineari per mezzo dell'utilizzo della **polynomial regression**, una regressione lineare multipla nella quale i regressori sono:  $X_1 = X, X_2 = X^2, X_3 = X^3, \dots, X_p = X^p$ .

$$Y = f(X) + \epsilon = \beta_0 + \sum_{k=1}^p \beta_k X^k + \epsilon$$

- Example:  
Auto data set



La figura suggerisce che si potrebbe ottenere un fit migliore con:

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \epsilon$$

Controlliamo:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower <sup>2</sup>	0.0012	0.0001	10.1	< 0.0001

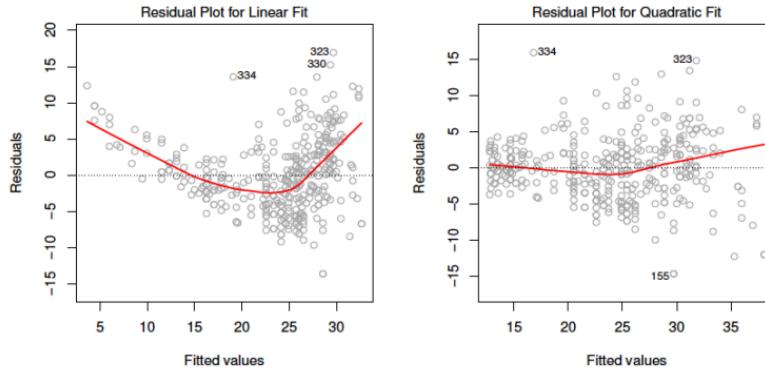
## Potenziali problemi

Quando facciamo il fit di un modello di regressione lineare rispetto ad uno specifico dataset possono verificarsi diversi problemi:

- non linearità della relazione response-predictor;
- correlazione dei termini di errore;
- varianza dei termini di errore non costante (eteroschedasticità);
- outliers (valori anomali, eccezioni);
- high-leverage point;
- collinearità.

### Non linearità della relazione response-predictor

La regressione lineare parte dall'assunzione che ci sia una relazione lineare tra predittori e risposta ma se questo non è vero le nostre previsioni saranno inaccurate.



LEFT: A linear regression of *mpg* on *horsepower*.

RIGHT: A linear regression of *mpg* on *horsepower* and  $horsepower^2$ .

Il **Residual Plot** sono uno strumento grafico per identificare la non linearità, in questi plot i residui  $e_i = y_i - \hat{y}_i$  sono riportati contro le  $x_i$  (per i modelli di regressione semplice) o contro i valori fitted  $\hat{y}_i$  (per i modelli di regressione multipli).

Idealmente il residual plot non dovrebbe mostrare alcun pattern, in quanto un pattern potrebbe indicare un problema con il modello lineare.



Se il Residual Plot indica che ci sono associazioni non lineari nei dati le trasformazioni non lineari dei predittori possono essere utili, è il caso di  $\log(X)$ ,  $\sqrt{X}$  e  $X^2$ .

## Correlazione dei termini di errore

Nei modelli di regressione lineare una importante assunzione è che gli errori siano incorrelati, questa ipotesi ad esempio è usata per la stima dei coefficienti di regressione o dei fitted values.

Se è presente una correlazione tra i termini di errore allora gli errori standard stimati tenderanno a **sottostimare** i veri errori standard.

A causa di ciò gli intervalli di confidenza e di previsione saranno più stretti di quanto dovrebbero essere.

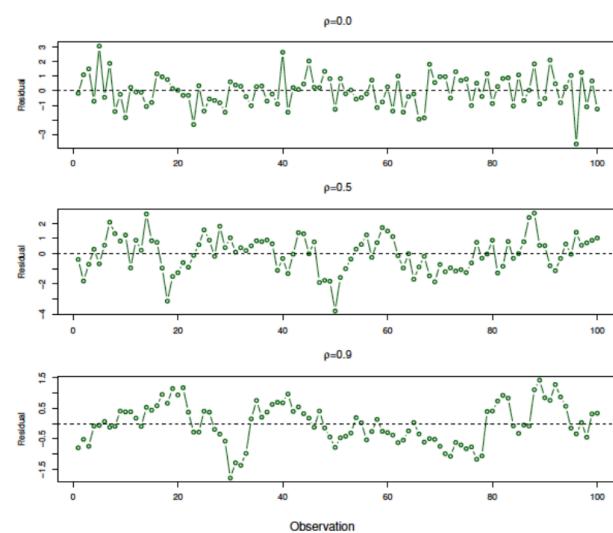
Ad esempio un intervallo di confidenza al 95% potrebbe in realtà avere una probabilità di contenere il valore vero più bassa di 0.95.

In aggiunta i p-values associati con il modello saranno più bassi del dovuto, questo potrebbe causare l'errata conclusione che il parametro sia statisticamente significativo.

In breve la correlazione tra gli errori porta ad un ingiustificato senso di sicurezza nei confronti del nostro modello.

Nel top panel (errore non correlato) non c'è evidenza di un trend nei residui.

In contrasto, i residui nel bottom panel (correlazione pari a 0.9), c'è un pattern chiaro: i residui adiacenti tendono ad assumere valori simili.



nel bottom plot è possibile vedere il "tracking": valori vicini nel tempo dei residui sono simili

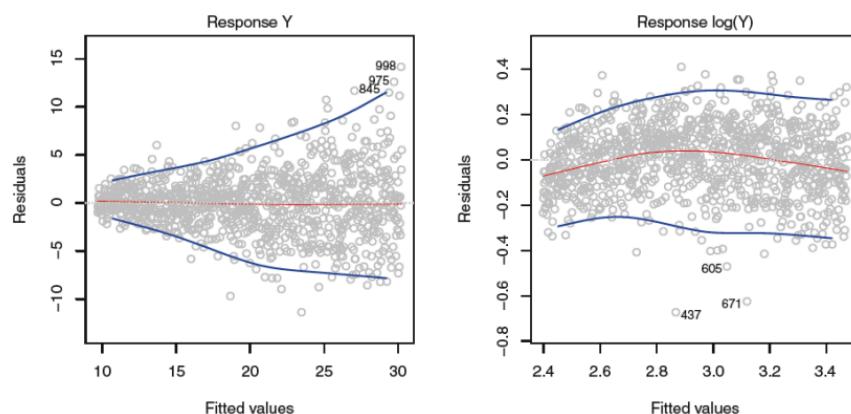
Se la correlazione è temporale (si può riscontrare constatando che ci sono pattern nel plot che confronta i residui del modello in funzione del tempo) prendo i campioni più lontani nel tempo, approccio di progettazione per esperimento, cioè progettare provando ad eliminare le cause di queste correlazioni.

Un buon design sperimentale è cruciale al fine di mitigare il rischio di correlazioni.

## Varianza non costante dei termini di errore

Un'altra assunzione importante dei modelli di regressione lineare è che i termini di errore abbiano varianza costante  $\text{Var}(\varepsilon_i) = \sigma^2$  (homoscedasticity).

I dati reali spesso mostrano varianza degli errori non costante (heteroscedasticity) che può essere identificata dai residual plot a forma di imbuto.



Una possibile soluzione è trasformare la risposta  $Y$  usando una funzione concava come  $\log(Y)$  o  $\sqrt{Y}$ . Queste funzioni introducono una quantità più grande di shrinkage quanto più grande è la risposta,

riducendo l'effetto dell'heteroscedasticity.

A destra possiamo vedere l'effetto della funzione che sembra aver reso la varianza costante, anche se si deve tenere in conto che si può introdurre una leggera relazione non lineare nei dati, come si vede nel grafico.

Un'altra tecnica per rimediare a questo problema è fare il fit del modello con weighted least squares.

## Outliers

Un outlier è un punto per il quale  $y_i$  è lontana dal valore previsto dal modello.

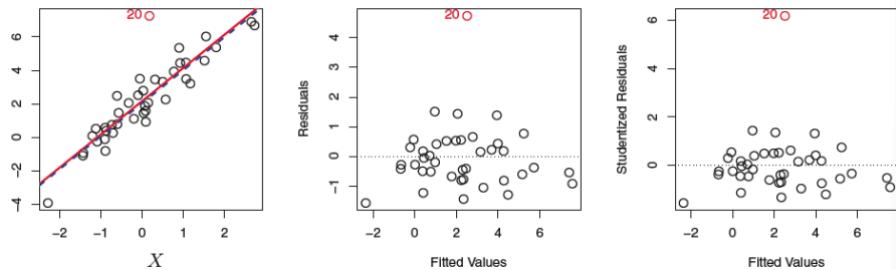
Gli outliers possono emergere per una serie di ragioni, ad esempio per la registrazione sbagliata di una osservazione durante il processo di data collection, se la motivazione è questa rimuovere l'outlier è una semplice soluzione ma bisogna fare attenzione in quanto un outlier potrebbe anche indicare una deficienza del nostro modello, come ad esempio un predittore mancante.

Un outlier che non ha un predittore con un valore inusuale tendenzialmente non ha molto effetto sul fit least squares, nel grafico a sinistra la linea rossa (least squares regression fit) e la linea tratteggiata blu (least squares regression fit dopo la rimozione dell'outlier) sono praticamente uguali.

Tuttavia, un outlier implica un aumento dell'RSE e l'RSE è usato per calcolare tutti gli intervalli di confidenza e i p-values.

L'outlier riduce l' $R^2$  e aumentando il p-value.

Ex: l'RSE è 1.09 con l'outlier e 0.77 senza;  $R^2$  declina da 0.892 a 0.805.



Per identificare gli outliers si può usare un residual plot (al centro), per essere più sicuri quando non è così ovvio chi è un outlier si può fare il plot degli studentized residuals (a destra), nel quale tutti i valori fuori dall'intervallo [-3, 3] sono da considerarsi outlier.

Gli studentized residuals  $\tau_i$ , possono essere calcolati dividendo ogni residuo  $e_i$  per il suo standard error stimato:

$$\tau_i = \frac{e_i}{\hat{\sigma} \sqrt{1-h_{ii}}},$$

dove il leverage  $h_{ii}$  è la  $i$ -esima entry della diagonale della Hat matrix H.

Le osservazioni con  $|\tau_i| > 3$  sono possibili outlier (come regola generale).

Ex: il residuo studentized dell'outlier eccede 6 (punto numero 20), mentre tutte le altre osservazioni hanno residui studentized tra -2 e 2.

## High influential (leverage) points

Gli outliers sono osservazioni per le quali la risposta  $y_i$  è inusuale dato il predittore  $x_i$ . In contrasto, le osservazioni che hanno un'alta leverage  $h_{ii}$  hanno un valore inusuale di  $x_i$ , oppure sono fuori dalla massa dei dati.

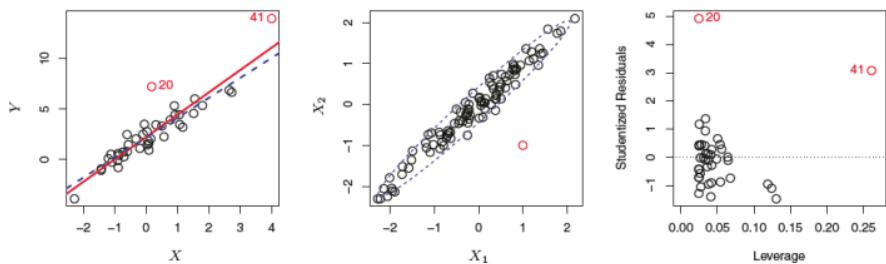
Ad esempio l'osservazione 41 ha valore del predittore grande rispetto alle altre osservazioni, notiamo che la differenza tra la linea rossa e quella blu è molto più grande rispetto a quella relativa al punto 20, in altre parole la rimozione di un High Leverage Point ha un effetto molto più sostanziale sulla least squares line rispetto alla rimozione di un Outlier.



Gli High Leverage Point hanno un impatto significativo sulla estimated regression line.

Di conseguenza è fondamentale rilevare questi punti, in una regressione lineare semplice basta osservare se ci sono osservazioni i cui valori dei predittori sono fuori dal range normale.

Il discorso è più complesso in una Multiple Linear Regression, nel qual caso può essere che una osservazione abbiamo valori dei predittori tutti nei range normali ma che considerati insieme sono inusuali, un esempio di questo è il grafico centrale che ci mostra un punto che ha un valore di  $X_1$  accettato, un valore di  $X_2$  accettato ma una coppia  $(X_1, X_2)$  che comporta un isolamento rispetto alla massa di tutte le altre osservazioni. (il problema è più complesso quando ci sono più di due predittori perché non c'è un modo semplice per tracciare il plot di tutte le dimensioni dei dati simultaneamente)



Per quantificare la leverage di una observation si calcola la leverage statistic. Un valore grande della leverage statistic indica una osservazione con alto leverage.

Per una regressione lineare semplice:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^n (x_{i'} - \bar{x})^2}.$$

NON mostrata in aula

Dall'equazione risulta chiaro che la leverage cresce con l'aumentare della distanza di  $x_i$  dalla sua media campionaria.

C'è un'estensione semplice di  $h_i$  al caso di più predittori ma la formula non è presentata nel libro.

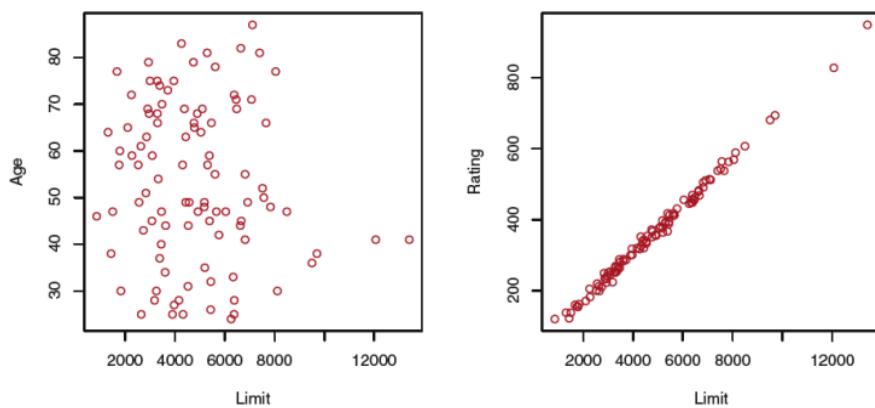
La leverage statistic  $h_i$  è tra  $1/n$  e 1, e la leverage media per tutte le osservazioni è uguale a  $(p+1)/n$ .

Quindi se una data osservazione ha una leverage statistic molto più grande di  $(p+1)/n$  allora il corrispondente punto ha leverage alta.

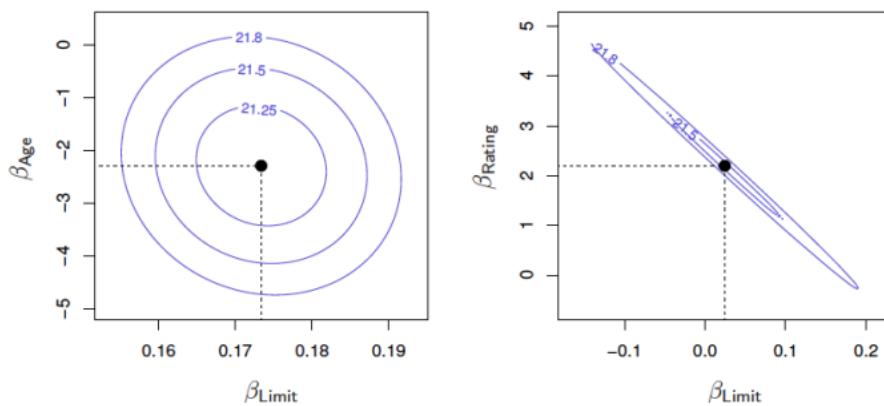
## Collinearity

La collinearità si riferisce alla situazione nella quale due o più predittori sono molto correlati tra loro. La presenza di collinearità pone problemi nella regressione, rendendo difficile separare gli effetti individuali delle variabili collineari sulla risposta.

Ad esempio, nel dataset della carta di credito limit e rating sono collineari e può essere difficile determinare come ognuna è separatamente associata alla risposta balance.



Possiamo osservare i Contour Plots per i valori di RSS come funzione dei parametri  $\beta$ . In ogni plot, i punti neri rappresentano i valori dei coefficienti corrispondenti al minimo RSS. Ogni ellissi rappresenta un insieme di coefficienti che corrispondono allo stesso RSS, più le ellissi sono vicine al centro e più l'RSS corrispondente è piccolo.



A destra a causa della collinearità ci sono molte coppie  $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$  con un valore di RSS simile.

In presenza di collinearità, a destra, notiamo che il Contour Plot si è ristretto e c'è ora una larga gamma di valori di coefficienti che hanno lo stesso RSS, quindi un cambiamento anche piccolo dei dati può

permettere ai valori dei coefficienti di muoversi ovunque nella valle dei valori equivalenti di RSS.

Questo implica grande incertezza nelle stime dei coefficienti.

La scala del coefficiente limit è aumentato di (circa) otto volte in presenza di collinearità, andando da -0.1 a 0.2 mentre prima andava da 0.16 a 0.19.

Una cosa interessante è che anche se la variabilità dei due coefficienti presi individualmente è aumentata molto questi comunque saranno quasi sicuramente presenti nella valle visibile nel plot, anche se -0.1 e 1 sono valori accettati per, rispettivamente, limit e rating non ci aspettiamo di rilevarli come valori veri dei coefficienti.

Ricordando che:

$$\hat{\beta} \sim N_{p+1}(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$$

La collinearità causa la crescita dell'errore standard per  $\beta_j$ .

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.734e+02	4.383e+01	-3.957	9.01e-05 ***
Age	-2.291e+00	6.725e-01	-3.407	0.000723 ***
Limit	1.734e-01	5.026e-03	34.496	< 2e-16 ***

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-377.53680	45.25418	-8.343	1.21e-15 ***
Rating	2.20167	0.95229	2.312	0.0213 *
Limit	0.02451	0.06383	0.384	0.7012

La collinearità comporta una riduzione della t-statistic (visto che aumenta lo SE che sta a denominatore) e potremmo quindi fallire nel rigettare  $H_0 : \beta_j = 0$ .

Questo vuol dire che la potenza dell'hypothesis test (la probabilità di rilevare correttamente un coefficiente non-zero) è **ridotto** dalla collinearità.

### Come rilevare la Collinearità

Un modo semplice per rilevare la collinearità è osservare la matrice di correlazione dei predittori. Un elemento di questa matrice che è grande in valore assoluto indica una coppia di variabili altamente correlate e di conseguenza un problema di collinearità dei dati.

Sfortunatamente è possibile che esista collinearità tra 3 o più variabili anche se nessuna coppia delle variabili in questione ha una correlazione particolarmente alta, questa situazione è chiamata multicollinearity.

### VIF

Un modo comune di stabilire la multicollinearity è calcolare il variance inflation factor (VIF):

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R_{X_j|X_{-j}}^2},$$

dove  $R_{X_j|X_{-j}}^2$  è il valore  $R^2$  da una regressione di  $X_j$  su tutte gli altri predittori  $X_{-j}$ .

Se  $R_{X_j|X_{-j}}^2$  è vicino a 1 la collinearità è presente e quindi il VIF sarà grande.

Come regola generale un VIF più grande di 5 o 10 indica una quantità problematica di collinearità.

VIF dall'esempio precedente:

Age	Limit
1.010283	1.010283

Rating	Limit
160.4933	160.4933

### Formula del VIF

Il VIF per una variabile indipendente  $X_i$  si calcola come:

$$VIF_i = \frac{1}{1 - R_i^2}$$

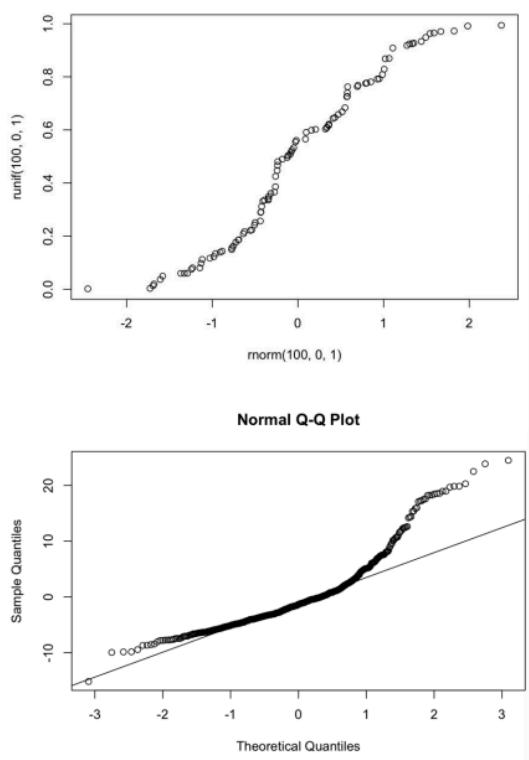
dove:

- $R_i^2$  è il coefficiente di determinazione della regressione lineare di  $X_i$  sulle altre variabili indipendenti del modello.

In altre parole:

1. Si prende una variabile  $X_i$  come dipendente.
2. Si regredisce  $X_i$  rispetto a tutte le altre variabili indipendenti.
3. Si calcola il  $R_i^2$  per questo modello.
4. Si inserisce  $R_i^2$  nella formula del VIF.

## Q-Q plot



Tipicamente un singolo data set è comparato ad una determinata distribuzione teorica, come nel caso di un Normal Q-Q plot dove i quantili del dataset sono comparati a quelli dati da una distribuzione normale.

Un Q-Q plot è un tipo particolare di scatterplot dove due insiemi di quantili (percentili) sono messi in plot l'uno contro l'altro.

Se entrambi i set di quantili vengono dalla stessa distribuzione dovrebbe vedere che i punti tendono ad allinearsi lungo una linea.

Risulta utile per capire se due samples vengono dalla stessa distribuzione.

## **DISCLAIMER**

Questi appunti sono stati realizzati a scopo puramente educativo e di condivisione della conoscenza. Non hanno alcun fine commerciale e non intendono violare alcun diritto d'autore o di proprietà intellettuale.

I contenuti di questo documento sono una rielaborazione personale di lezioni universitarie, materiali di studio e concetti appresi, espressi in modo originale ove possibile. Tuttavia, potrebbero includere riferimenti a fonti esterne, concetti accademici o traduzioni di materiale didattico fornito dai docenti o presente in libri di testo.

Se ritieni che questo documento contenga materiale di tua proprietà intellettuale e desideri richiederne la modifica o la rimozione, ti invito a contattarmi. Sarò disponibile a risolvere la questione nel minor tempo possibile.

In quanto autore di questi appunti non posso garantire l'accuratezza, la completezza o l'aggiornamento dei contenuti e non mi assumo alcuna responsabilità per eventuali errori, omissioni o danni derivanti dall'uso di queste informazioni. L'uso di questo materiale è a totale discrezione e responsabilità dell'utente.