# Metodi Statistici per le decisioni

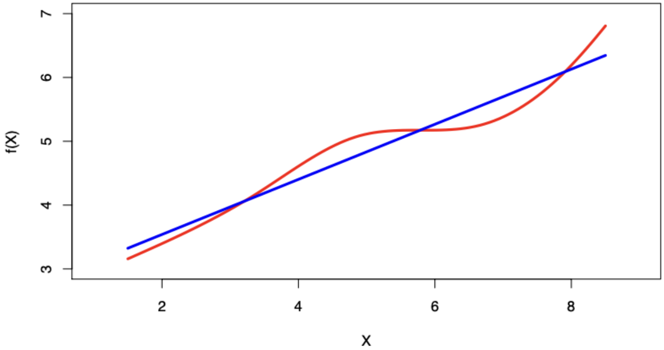
2024-2025

Vincenzo Nardelli

vincenzo.nardelli@unicatt.it

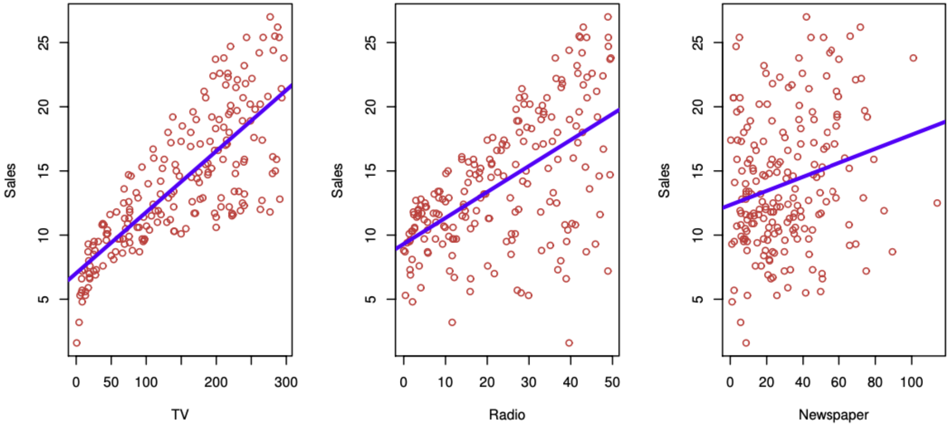
# Regressione Lineare

- La regressione lineare è un approccio semplice per l'apprendimento supervisionato. Assume che la dipendenza di Y da $X_1, X_2, \ldots, X_p$ sia lineare.




- Anche se può sembrare troppo semplicistico, la regressione lineare è estremamente utile sia concettualmente che praticamente.

# Dati pubblicitari



# Regressione lineare per i dati pubblicitari

Domande che potremmo porci:

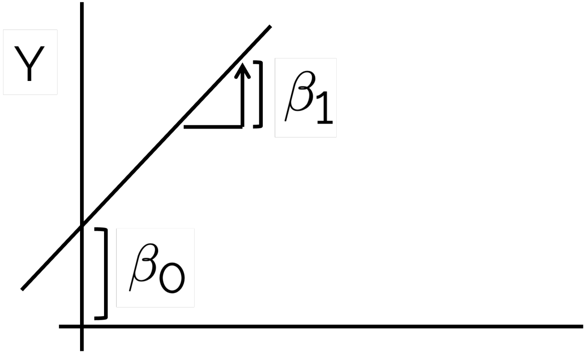
- Esiste una relazione tra budget pubblicitario e vendite?
- Quanto è forte la relazione tra budget pubblicitario e vendite?
- Quali mezzi contribuiscono alle vendite?
- Con quanta accuratezza possiamo prevedere le vendite future?
- La relazione è lineare?
- Esiste sinergia tra i mezzi pubblicitari?

# Regressione lineare semplice con un singolo predittore X

Assumiamo un modello

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

dove $\beta_0$ e $\beta_1$ sono due costanti sconosciute che rappresentano l'*intercetta* e la *pendenza*, noti anche come *coefficienti* o *parametri*, e $\epsilon$ è il termine di errore.



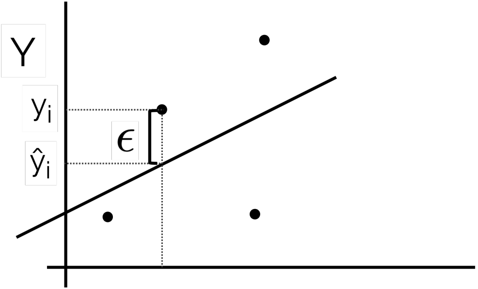
# Regressione lineare semplice con un singolo predittore X

Dati alcuni valori stimati $\hat{\beta}_0$ e $\hat{\beta}_1$ per i coefficienti del modello, prevediamo le vendite future usando

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

dove $\hat{y}$ indica una previsione di Y sulla base di X = x. Il simbolo con il cappello denota un valore stimato.

Sia $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la previsione per Y basata sul valore i-esimo di X. Allora $e_i = y_i - \hat{y}_i$ rappresenta il residuo i-esimo.



# Stima dei parametri tramite minimi quadrati

- Definiamo la *somma dei quadrati dei residui* (RSS) come

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2,$$

o equivalentemente come

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

# Stima dei parametri tramite minimi quadrati

- Definiamo la *somma dei quadrati dei residui* (RSS) come

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

o equivalentemente come

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- L'approccio dei minimi quadrati sceglie $\hat{\beta}_0$ e $\hat{\beta}_1$ per minimizzare l'RSS. I valori che minimizzano possono essere mostrati come

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

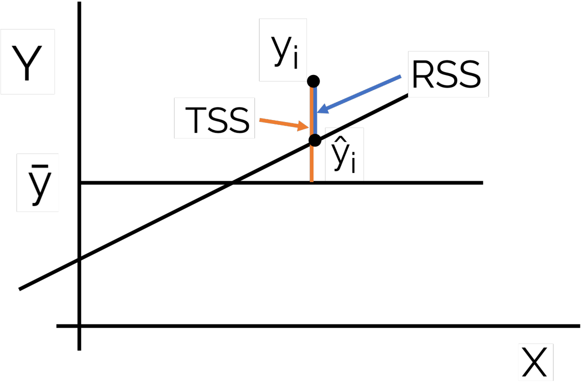
dove $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ e $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ sono le medie campionarie.

# Valutazione dell'Accuratezza Complessiva del Modello

- Il *R-quadrato* o frazione della varianza spiegata è

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

dove $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ è la *somma totale dei quadrati*.



# LAB Consumo di Sigarette

Un ente sanitario sta studiando come il prezzo e la tassazione influenzino il consumo di sigarette a livello statale. Analizzando i dati storici, l'obiettivo è comprendere l'efficacia delle politiche di aumento dei prezzi e delle tasse sul controllo del consumo di tabacco. Questa analisi fornirà basi solide per formulare raccomandazioni a favore della salute pubblica, con l'intento di ridurre il consumo di sigarette e i rischi correlati alla salute.

Attraverso lo studio di queste relazioni, l'ente sanitario intende valutare l'effetto delle variazioni di prezzo e tassazione per pianificare strategie che possano incentivare un decremento nel consumo di sigarette.

# Descrizione delle Variabili - Dataset Cigarette (Pacchetto Ecdat)

Il dataset contiene informazioni raccolte a livello statale negli Stati Uniti e include variabili chiave per l'analisi dell'effetto dei prezzi e della tassazione sul consumo di sigarette. Le principali variabili sono:

- **state**: Stato in cui sono stati raccolti i dati.
- **year**: Anno della raccolta dei dati.
- **avgprs**: Prezzo medio di un pacchetto di sigarette (in dollari).
- **packpc**: Consumo di sigarette (pacchetti pro capite).
- **taxs**: Totale delle tasse su un pacchetto di sigarette (in dollari).

# LAB Consumo di Sigarette: Domande di Analisi

- Esiste una relazione tra il prezzo medio di un pacchetto di sigarette (`avgprs`) e il consumo di sigarette (`packpc`)? La relazione è positiva o negativa?
- Qual è la correlazione tra prezzo medio e consumo di sigarette? Come possiamo interpretare questo valore per valutare l'influenza del prezzo sul consumo?
- In che modo la tassazione totale (`taxs`) influenza il consumo di sigarette? Analizza questa relazione utilizzando uno scatterplot e calcola la correlazione.
- Tramite un modello di regressione lineare, quale sarebbe l'impatto sui consumi di sigarette se la tassazione in uno stato passasse da 50$ a 100$?

# LAB Real Estate

Un'agenzia immobiliare è interessata a comprendere come il livello socioeconomico di una zona influenzi il valore medio delle case di Boston. L'obiettivo è valutare l'impatto del livello di povertà sul valore medio delle abitazioni, così da poter offrire raccomandazioni più precise agli investitori e ai pianificatori urbani. L'analisi di questa

relazione è fondamentale per definire strategie di investimento mirate e per prevedere l'andamento del mercato immobiliare in funzione delle variabili socioeconomiche, contribuendo a migliorare l'efficacia delle decisioni di business.

# Descrizione delle Variabili - Dataset Boston (Pacchetto ISLR2)

Il dataset contiene informazioni su vari quartieri, includendo variabili chiave per l'analisi socioeconomica del valore delle abitazioni. Di seguito alcune delle principali variabili utilizzate:

- **lstat**: Percentuale di popolazione con basso livello socioeconomico, un indicatore del livello di povertà nella zona.
- **medv**: Valore medio delle abitazioni in migliaia di dollari, rappresenta il target di interesse per il mercato immobiliare.
- **rm**: Numero medio di stanze per abitazione, una misura della dimensione abitativa media in ciascun quartiere.
- **age**: Percentuale di abitazioni costruite prima del 1940, indica la vetustà del patrimonio immobiliare.
- **dis**: Distanza media dai centri di lavoro di Boston, importante per valutare l'accessibilità ai servizi urbani.

# LAB Real Estate: Domande di Analisi

- Qual è la relazione tra il livello di povertà (`lstat`) e il valore medio delle case (`medv`)? La relazione è positiva o negativa?
- Qual è la covarianza tra `lstat` e `medv`? Che cosa ci indica questo valore?
- Qual è la correlazione tra `lstat` e `medv`? Come si interpreta questa correlazione in termini di influenza del livello di povertà sui prezzi immobiliari?
- Possiamo prevedere il valore delle case in funzione del livello di povertà? Quali sono le previsioni del valore medio delle case per quartieri con livelli di povertà pari al 5%, 10%, e 15%?

# LAB Real Estate: Visualizzazione dei Dati

- Qual è la distribuzione del livello di povertà (`lstat`) tra i quartieri? Visualizza e descrivi la distribuzione.
- Qual è la distribuzione del valore medio delle case (`medv`) tra i quartieri?
- Esiste una relazione lineare tra livello di povertà e valore medio delle case? Mostra il grafico e discuti i risultati.

# LAB Real Estate: Analisi del Modello

- Quali sono i coefficienti di regressione tra livello di povertà e valore medio delle case? Come si interpretano l'intercetta e la pendenza?
- Qual è il livello di bontà di adattamento del modello? È una relazione forte o debole?
- In che modo possiamo utilizzare i risultati del modello per stimare il valore delle case in quartieri con livelli diversi di povertà?
- Prova a stimare altri modelli con le rimanenti variabili. Come si interpretano i risultati?

# Errore Standard

- L'errore standard di un estimatore riflette come varia sotto campionamento ripetuto. Abbiamo:

$$\mathrm{SE}\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, \quad \mathrm{SE}\left(\hat{\beta}_0\right)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]$$

dove $\sigma^2 = \mathrm{Var}(\epsilon)$.

# Intervalli di Confidenza

- Questi errori standard possono essere usati per calcolare intervalli di confidenza.
- Un intervallo di confidenza al 95% è definito come un intervallo che, con il 95% di probabilità, contiene il valore vero sconosciuto del parametro. Ha la forma:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right)$$

# Intervalli di Confidenza - Continuazione

- Cioè, c'è approssimativamente una probabilità del 95% che l'intervallo:

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right)\right]$$

contenga il valore vero di $\beta_1$ (in uno scenario in cui si ottengano campioni ripetuti come quello attuale).

# Test di Ipotesi

- Gli errori standard possono essere utilizzati anche per eseguire test di ipotesi sui coefficienti.
- Il test più comune riguarda l'ipotesi nulla:

$$H_0 : \text{Non c'è relazione tra X e Y}$$

contro l'ipotesi alternativa:

$$H_A : \text{C'è una relazione tra X e Y.}$$

- Matematicamente, ciò corrisponde a testare:

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0$$

poiché se $\beta_1 = 0$, il modello si riduce a $Y = \beta_0 + \epsilon$, e X non è associato a Y.

# Test di Ipotesi - Continuazione

- Per testare l'ipotesi nulla, calcoliamo una statistica t, data da:

$$t = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)}$$

- Questa statistica ha una distribuzione t con n – 2 gradi di libertà, assumendo che $\beta_1 = 0$.
- Usando software statistici, è facile calcolare la probabilità di osservare un valore uguale o maggiore di |t|. Questa probabilità è chiamata p-value.

# Risultati sui Dati di Pubblicità

| | Coefficiente | Errore Std. | Statistica t | p-value |
|---|---|---|---|---|
| Intercetta | 7.0325 | 0.4578 | 15.36 | < 0.0001 |
| TV | 0.0475 | 0.0027 | 17.67 | < 0.0001 |

# Valutazione dell'Accuratezza Complessiva del Modello

- Calcoliamo l'Errore Standard Residuo (RSE):

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

dove il residuo somma dei quadrati è:

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

# Risultati sui Dati di Pubblicità

| Quantità | Valore |
|---|---|
| Errore Standard Residuo | 3.26 |
| $R^2$ | 0.612 |
| Statistica F | 312.1 |

# Regressione Lineare Multipla

- Il modello è:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Interpretiamo $\beta_j$ come l'effetto medio su Y di un aumento unitario in $X_j$, mantenendo tutti gli altri predittori fissi. Nell'esempio pubblicitario, il modello diventa:

$$\text{vendite} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{giornali} + \epsilon$$

# Interpretazione dei Coefficienti di Regressione

- Lo scenario ideale è quando i predittori sono non correlati:
  - Ogni coefficiente può essere stimato e testato separatamente.
  - Interpretazioni come "un cambiamento unitario in $X_j$ è associato a un cambiamento di $\beta_j$ in Y, mentre tutte le altre variabili rimangono fisse" sono possibili.
- Attenzione alle correlazioni tra i predittori, possono causare problemi!

# Stima e Predizione per la Regressione Multipla

- Date le stime $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, possiamo fare previsioni usando la formula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- Stimiamo $\beta_0, \beta_1, \ldots, \beta_p$ come i valori che minimizzano la somma dei residui al quadrato:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2 \end{aligned}$$

- Questo è fatto usando software statistici standard. I valori $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ che minimizzano RSS sono le stime dei coefficienti di regressione multipla ottenuti con il metodo dei minimi quadrati.

# Risultati per i Dati Pubblicitari

| | Coefficiente | Errore Std. | Statistica t | p-value |
|---|---|---|---|---|
| Intercetta | 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 |
| giornali | -0.001 | 0.0059 | -0.18 | 0.8599 |

Correlazioni

| | TV | radio | giornali | vendite |
|---|---|---|---|---|
| TV | 1.0000 | 0.0548 | 0.0567 | 0.7822 |
| radio | | 1.0000 | 0.3541 | 0.5762 |
| giornali | | | 1.0000 | 0.2283 |
| vendite | | | | 1.0000 |

# Alcune domande importanti

1. Almeno uno dei predittori $X_1, X_2, \ldots, X_p$ è utile per prevedere la risposta?
2. Tutti i predittori aiutano a spiegare Y, o è utile solo un sottoinsieme?
3. Quanto bene il modello si adatta ai dati?
4. Dato un set di valori predittori, quale valore della risposta dobbiamo prevedere, e quanto è accurata la nostra previsione?

# Almeno un predittore è utile?

- Per rispondere alla prima domanda, possiamo usare la statistica F:

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

| Quantità | Valore |
|---|---|
| Errore Standard Residuo | 1.69 |
| $R^2$ | 0.897 |
| Statistica F | 570 |

# Valutazione dell'Accuratezza Complessiva del Modello: Confronto tra $R^2$ e $R^2$ adj

- Il $R^2$ è calcolato come:

$$R^2 = 1 - \frac{RSS}{TSS}$$

  dove:

  - $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ è la somma dei quadrati dei residui.
  - $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ è la somma totale dei quadrati.

- Il $R^2$ *adj* tiene conto del numero di predittori e del numero di osservazioni, ed è calcolato come:

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

  dove:

  - $n$ è il numero di osservazioni.
  - $p$ è il numero di predittori nel modello.

# Decidere sulle variabili importanti

- L'approccio più diretto si chiama regressione con tutti i sottoinsiemi o migliori sottoinsiemi: calcoliamo la stima dei minimi quadrati per tutti i sottoinsiemi possibili e poi scegliamo in base a un criterio che bilancia l'errore di training con la dimensione del modello.
- Tuttavia, spesso non possiamo esaminare tutti i modelli possibili, dato che sono $2^p$. Ad esempio, quando $p = 40$, ci sono più di un miliardo di modelli!
- Serve quindi un approccio automatizzato che cerchi tra un sottoinsieme di modelli. Discutiamo due approcci comunemente usati.

# Selezione Avanti (Forward Selection)

- Inizia con il modello nullo: un modello che contiene un'intercetta ma nessun predittore.
- Adatta p regressioni lineari semplici e aggiungi al modello nullo la variabile che produce il minimo RSS.
- Aggiungi al modello la variabile che produce il minimo RSS tra tutti i modelli a due variabili.
- Continua finché non si soddisfa una regola di arresto, ad esempio quando tutte le variabili rimanenti hanno un p-value sopra una certa soglia.

# Selezione Indietro (Backward Selection)

- Inizia con tutte le variabili nel modello.
- Rimuovi la variabile con il p-value più alto, cioè quella meno statisticamente significativa.
- Adatta il nuovo modello con $(p - 1)$ variabili e rimuovi la variabile con il p-value più alto.
- Continua finché non si raggiunge una regola di arresto. Ad esempio, si può arrestare quando tutte le variabili rimanenti hanno un p-value significativo rispetto a una soglia prestabilita.

# Previsioni con il Modello Lineare

**Equazione della previsione**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

**Esempio pratico**:

- Coefficienti stimati: $\hat{\beta}_0 = 5$, $\hat{\beta}_1 = 2$
- Per $x = 10$, si ottiene:

$$\hat{y} = 5 + 2 \cdot 10 = 25$$

**Nota**: La previsione è valida solo nel range dei dati osservati (*interpolazione*).

# Valutazione tramite MSE

**Errore Quadratico Medio (MSE)**:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

**Interpretazione**:

- Misura la distanza media quadratica tra i valori osservati ($y_i$) e quelli previsti ($\hat{y}_i$).
- Un MSE basso indica un buon adattamento del modello ai dati.

# Rischi e Cose da Attenzionare

**Rischi delle Previsioni**:

- **Overfitting**: Il modello si adatta troppo ai dati di training, perdendo capacità predittiva sui nuovi dati.
- **Estrapolazione**: Fare previsioni fuori dal range dei dati osservati può portare a risultati poco accurati.
- **Outlier**: Valori anomali nei dati possono influenzare negativamente il modello e aumentare l'MSE.

**Cose da Attenzionare**:

- **Selezione delle variabili**: Utilizzare solo variabili rilevanti per ridurre il rumore nel modello.
- **Validazione del modello**: Valutare il modello su un set di dati di test per verificare la generalizzabilità.
- **MSE non contestualizzato**: Un MSE basso non garantisce che il modello sia utile; valutare anche il contesto dei dati.

# Altre Considerazioni nel Modello di Regressione

**Predittori Qualitativi**

- Alcuni predittori non sono quantitativi ma qualitativi, assumendo un set discreto di valori.
- Questi sono anche chiamati predittori categorici o variabili fattoriali.
- Per esempio, dati sulle carte di credito.

# Predittori Qualitativi - Continuazione

**Esempio: Differenze nei saldi delle carte di credito tra uomini e donne, ignorando le altre variabili.** Creiamo una nuova variabile:

$$x_i = \begin{cases} 1 & \text{se la persona i è donna} \\ 0 & \text{se la persona i è uomo} \end{cases}$$

Modello risultante:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{se la persona i è donna} \\ \beta_0 + \epsilon_i & \text{se la persona i è uomo} \end{cases}$$

# Dati Carte di Credito - Continuazione

**Risultati per il modello con genere:**

| | Coefficiente | Errore Std. | Statistica t | p-value |
|---|---|---|---|---|
| Intercetta | 509.80 | 33.13 | 15.389 | < 0.0001 |
| genere[Femminile] | 19.73 | 46.05 | 0.429 | 0.6690 |

# Predittori Qualitativi con Più di Due Livelli

- Con più di due livelli, creiamo variabili dummy aggiuntive. Ad esempio, per la variabile etnia possiamo creare due variabili dummy. La prima potrebbe essere:

$$x_{i1} = \begin{cases} 1 & \text{se la persona } i \text{ è Asiatica} \\ 0 & \text{se la persona } i \text{ non è Asiatica} \end{cases}$$

e la seconda potrebbe essere:

$$x_{i2} = \begin{cases} 1 & \text{se la persona } i \text{ è Caucasica} \\ 0 & \text{se la persona } i \text{ non è Caucasica} \end{cases}$$

# Predittori qualitativi con più di due livelli - Continuazione

- Entrambe queste variabili possono essere utilizzate nell'equazione di regressione, ottenendo il modello:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{se la persona i è Asiatica} \\ \beta_0 + \beta_2 + \epsilon_i & \text{se la persona i è Caucasica} \\ \beta_0 + \epsilon_i & \text{se la persona i è Afroamericana.} \end{cases}$$

- Ci sarà sempre una variabile dummy in meno rispetto al numero di livelli. Il livello senza variabile dummy (in questo esempio Afroamericano) è noto come **baseline**.

# Risultati per l'etnia

| | Coefficiente | Errore Std. | Statistica t | p-value |
|---|---|---|---|---|
| Intercetta | 531.00 | 46.32 | 11.464 | < 0.0001 |
| etnia [Asiatica] | -18.69 | 65.02 | -0.287 | 0.7740 |
| etnia [Caucasica] | -12.50 | 56.68 | -0.221 | 0.8260 |

# LAB Credit Analysis

Una banca è interessata a comprendere quali fattori influenzano il saldo medio delle carte di credito dei clienti, con l'obiettivo di migliorare le decisioni sulle richieste di prestiti e crediti. L'analisi si focalizza su variabili socioeconomiche e comportamentali, come reddito, limite di credito e status da studente. L'obiettivo finale è fornire

raccomandazioni strategiche per ottimizzare la gestione del credito e ridurre i rischi finanziari, contribuendo a un processo decisionale più efficace.

# Descrizione delle Variabili - Dataset Credit (Pacchetto ISLR2)

Il dataset contiene informazioni su clienti di una banca, incluse variabili chiave per l'analisi delle loro abitudini finanziarie e socioeconomiche. Di seguito alcune delle principali variabili utilizzate:

- **Balance**: Saldo medio della carta di credito, rappresenta la variabile target per l'analisi.
- **Income**: Reddito annuale del cliente (in migliaia di dollari), un indicatore della capacità di spesa.
- **Limit**: Limite massimo della carta di credito, rappresenta la disponibilità economica concessa dalla banca.
- **Rating**: Valutazione del cliente da parte della banca, in base alla sua storia creditizia.
- **Student**: Indica se il cliente è uno studente (variabile categorica).

# LAB Credit Analysis: Analisi descrittiva

- Qual è la distribuzione del reddito (`Income`) tra i clienti? Visualizza e descrivi la distribuzione.
- Qual è la distribuzione del saldo medio delle carte di credito (`Balance`)?
- Qual è la relazione tra reddito (`Income`) e saldo medio della carta di credito (`Balance`)? La relazione è positiva o negativa?
- Qual è la covarianza tra `Income` e `Balance`? Che cosa ci indica questo valore?
- Qual è la correlazione tra `Income` e `Balance`? Come si interpreta questa correlazione in termini di influenza del reddito sul saldo della carta di credito?

# LAB Credit Analysis: Modello

- Quali sono i coefficienti di regressione tra reddito e saldo medio? Come si interpretano l'intercetta e la pendenza?
- Possiamo prevedere il saldo medio della carta in funzione del reddito? Quali sono le previsioni del saldo per redditi pari a 20k, 50k, e 80k dollari?
- Qual è il livello di bontà di adattamento del modello ($R^2$)? È una relazione forte o debole?
- Includendo altre variabili come `Limit` e `Rating`, come cambia il modello? Quali variabili risultano più significative?
- In che modo la variabile categorica `Student` influisce sul saldo medio della carta di credito?

# Cosa non abbiamo trattato

- Outlier
- Varianza non costante dei termini di errore
- Collinearità

# Generalizzazione del Modello Lineare

- **Problemi di classificazione:** regressione logistica, support vector machines.
- **Non-linearità:** kernel smoothing, splines e modelli additivi generalizzati; metodi dei k-vicini più prossimi.
- **Interazioni:** metodi basati su alberi, bagging, foreste casuali e boosting (che catturano anche le non-linearità).
- **Fitting regolarizzato:** regressione Ridge e Lasso.