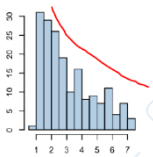
FORMULARIO

- Se il coeff. è segnato come: 8.891e01 = 0.8891
- **RICORDA:** quasi sempre la risposta con "sembra" è giusta!!!!
- Se il dato è in migliaia (es: income 1000 USD) e il parametro è in log(income), l'interpretazione è in %!
- Se media maggiore di mediana: asimmetria a destra. Il grafico è così:



TEST

Pearson's Chi-squared test (test d'indipendenza)
p-value basso = associazione tra le variabili
Per calcolare l'indice di associazione uso CRAMER.
CRAMER: Per calcolare l'intensità della relazione:

$C = \sqrt{\frac{\chi^2}{n \cdot \min\{\text{righe-1, colonne-1}\}}}$ χ^2 è X squared che vedo nel test

Pearson's product-moment correlation
Test di indipendenza lineare (varia tra -1 e +1; 0= no relazione): ci si basa sul coefficiente di correlazione lineare per vedere se la relazione è forte o debole. P-value basso = relazione.

Analysis of Variance table (se voglio vedere la significatività di un evento)
Valore della statistica test alto = p-value basso, quindi c'è significatività

Analysis of Deviance table (è il test LRT delle verosomiglianze)
p-value basso = significatività, quindi accetto il modello completo che include quella variabile.
Resid. Dev. è l'AIC (che deve essere basso!)

Test F
p-value è grande (più di 0.05) prendo il modello RIDOTTO
p-value è Piccolo (meno di 0.05) prendo il modello COMPLETO

Welch two Sample test (confronto delle medie)
p-value basso = significatività

Test Breusch Pagan (omo/eteroschedasticità)
P-value alto = omoschedasticità (UNICO caso in cui vogliamo p-value alto)

DISTRIBUZIONE NORMALE

- Si fissa α (la probabilità di errore) → 5% (0.05)
- Si hanno 4 punti per il test:
  - 1. Impostare H_0 e H_1
  - 2. Calcolo della statistica test (t): misura la distanza da H_0, se è vicino a 0 accetto H_0

    T test: $\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$
  - 3. P-value: probabilità da 0 a 1, più è piccolo meno è probabile che si accetti H_0

    P-value: PV $(t_{df} \geq \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \mid H_0)$ dove $\bar{x}$ è la media
  - 4. P-value $< \alpha$: rifiuto H_0 per H_1

➢ **Linear Correlation p-values (dopo grafico con i pallini rossi, neg. e blu, pos.):** sono maggiormente correlate le variabili più piccole relative a un parametro

p-value tra SALES e PERPERHH

Linear Correlation p-values

	PERPERHH	MEDSCHYR	PRC55P	HHMEDAGE	MEDINC	SALES
PERPERHH	NA	0.1889	0.0001	0.0004	0.0178	0.3454
MEDSCHYR	0.1889	NA	0.0004	0.0004	0.0001	0.0006
PRC55P	0.0001	0.0004	NA	0.0000	0.0000	0.0057
HHMEDAGE	0.0004	0.0004	0.0000	NA	0.0000	0.0092
MEDINC	0.0178	0.0001	0.0000	0.0000	NA	0.0006

PARAMETRI

- AIC = $-2 \log L(M) + 2(p+1)$ deve essere BASSO **AIC = 2p + residual deviance**
  dove L(M) indica il valore della verosomiglianza per il modello M e p=n° predittori (con intercetta)

- **Sensitività (TPR) =** $\frac{TP}{TP+FN}$ → nella Confusion Matrix prendo i numeri di: no-no (TN), yes-yes (TP), no-yes (FN), yes-no (FP). Se aumenta la sensitività, diminuisce la specificità. Un valore alto è vicino a 1.

- **Specificità (TNR) =** $\frac{TN}{TN+FP}$ → Un valore alto è vicino a 1.

- **Accuratezza (accuracy) =** $\frac{TP+TN}{TP+FN+TN+FP}$ → più è ALTA meglio è (vicino a 1), indica la bontà predittiva del modello.

- **ROC (curva) =** numeri nella parentesi indicano (specificità, sensitività)

- **AUC (area sotto la curva)** = deve essere ALTA (vicino a 1)
- Log-Likelihood: verosomiglianza

- **RMSE** (s) → Più basso è meglio è = $\sqrt{MSE}$ (lo trovo nel modello random forest)

- ***R-squared (R2) = 1*** $\frac{SSE\ (somma\ dei\ quadrati\ residui, varianza\ non\ spiegata\ dal\ modello - ultimo\ num\ SumSq)}{SST\ (somma\ dei\ quadrati\ totali, varianza\ della\ variabile\ dipendente)}$

$R^2 = 1 - \frac{n-p-1(1-R^2 adj)}{n-1}$ è compreso tra 0 e 1 (anche chiamato "chi squared")
Più è alto e meglio è → Ci dice la variabilità spiegata

- ***Adj. R-squared (R2) = 1 -*** $\frac{(1-R^2)(n-1)}{n-p-1}$ p=predittori

Più alto è e meglio è → varia da 0 a 1; un valore alto indica migliore adattabilità del modello
Lo uso per confrontare i modelli
Tiene conto della complessità. Dovrebbe essere molto simile all'R^2.

- **MAE (mean absolute error, media dei residui in valore assoluto)**

Più basso è, meglio è

ETEROGENEO/OMOGENEO

Confronto dei coefficienti di variazione (CV): la variabile con la distribuzione più dispersa/eterogenea è quella con il CV più grande. Posso anche guardare l'IQR più alto = più eterogeneità oppure Std. Dev.
Se le variabili hanno un'unità di misura diversa, non possiamo confrontare la loro eterogeneità attraverso le deviazioni standard campionarie. Per risolvere questo problema dobbiamo calcolare i rispettivi coefficienti di variazione, ovvero i rapporti tra le deviazioni standard con le medie.

- il valore di MEDSCHYR quante deviazioni standard si trova sopra la rispettiva media? $\frac{\text{max - mean}}{\text{Std.Dev.}}$

Dal boxplot e dalla tabella con gli indici di sintesi possiamo vedere che il valore di MEDSCHYR per l'osservazione di Shorewood è pari a 15.9 anni. Poiché la media campionaria di MEDSCHYR è pari a 12.696 anni e la sua deviazione standard è 0.5514, il valore di MEDSCHYR per Shorewood dista dalla media esattamente (15.9 – 12.696)/0.551 = 5.8149 deviazioni standard.

OMOSCHEDASTICITÀ E ETEROSCHEDASTICITÀ (VARIABILITÀ ERRORI):

Guardiamo grafico $\sqrt{stud.residual}$ vs. Fitted Value e vediamo se range dei valori è costante.
Se c'è Test BREUSCH - PAGAN: UNICO CASO DOVE VOGLIAMO **p-value ALTO (=omosched.)**
Se p-value alto NON c'è problema di eteroschedasticità.

OUTLIERS

StudRes: maggiore di 2 o minore -2 (residui standardizzati)
CookD: se >1 ci dice che il valore è influente
Hat values: più si avvicina a 1 e più è osservazione ad alto leverage
Hat value medio: si guardano le linee verticali nel grafico, la prima che incontro indica l'hat value medio
VIF (multicollinearità) = se maggiore di 10 c'è collinearità (almeno tra 2 predittori).

INTERVALLO DI CONFIDENZA per una media campionaria $\bar{x}$ è data da: $\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$

$\bar{x}$: media campionaria

t_{n-1}: grado di libertà (DF)

1 - $\frac{\alpha}{2}$: percentile (di solito α è 95% oppure 0,95)

s: deviazione standard → più è alta meno è precisa la stima

$\frac{s}{\sqrt{n}}$: errore standard

$t_{n-1,1-\frac{\alpha}{2}}$: valore critico della distribuzione

GRAFICO

- Intercetta: linea nera (ci interessa solo come punto di riferimento)
- Parametro 1: indica la pendenza
- Si va in ordine in base alla grandezza dei coefficienti (se +, la curva è sopra l'intercetta; se -, la curva è sotto)
- Se sotto "Call" c'è scritto "glm" (modello logistico) allora devono essere delle CURVE
- "lm" = modello lineare, rette

INTERPRETAZIONI

Modello lineare e semi-log: è un numero
Modello log-log: è una percentuale, si interpreta come l'elasticità di y (interc) rispetto a x (parametro).
$\frac{e}{1+e}$ → risposta probabilistica di un **modello logistico glm** (si usa solo qui), che spiega appunto una probabilità

*LOT.SIZE è in piedi quadrati.
Lm: è un numero!
Il coefficiente stimato di LOT.SIZE (0.0196) si interpreta come variazione nel prezzo medio (INTERCETTA) per l'incremento di LOT.SIZE di 1 piede quadrato.
Lm semi-log:
Approccio di lm standard: un incremento unitario di x è associato ad una variazione pari a $\beta 1$ di log y .
Glm (logit) no log: NO PERCENTUALE!
Il coefficiente stimato di LOT.SIZE (0.0002) è interpretato come la variazione nei log-odds stimati che una casa abbia una taverna se aumentiamo di 1 piede quadrato LOT.SIZE, assumendo costante gli altri parametri.
Equivalentemente, l'esponenziale exp(0.0002) = 1.00018609, è interpretato come l'odds ratio stimato che una casa abbia una taverna se aumentiamo di 1 piede quadrato LOT.SIZE, assumendo costante NUM.BED.
In altre parole, se aumentiamo la dimensione di una casa, è più probabile che la casa abbia una taverna.
Variabile dummy
il coefficiente stimato per GENDER(dummy) rappresenta la differenza tra i valori medi stimati per la variabile FACE per i maschi e le femmine, tenendo sotto controllo le altre variabili incluse nel modello.

Non nel programma per l'esame:

OVERFITTING

Scelta del parametro di Tuning (SOLO 1 - costo per la complessità - PER I MODELLI AD ALBERO):

1. Scegliere il miglior valore del parametro adottato (RMSE= più basso è, meglio è)
2. Regola di 1 volta lo standard error (1-se):
   - Prendere valore scelto al punto 1
   - +/- il suo standard error
   - Nel RANGE ottenuto, prendere il valore del parametro meno performante e che abbia lo standard error più basso.

Call: Ranger → random forest