ANALISI DEI DATI

LEZIONE 1 - INTRODUZIONE

Essendo ogni fenomeno reale caratterizzato da variabilità (quotazione dei titoli, comportamento dei debitori, ecc.) risulta necessario spiegare tale variabilità, ovvero trovare una spiegazione di essa tramite l'individuazione di fattori a cui è associata tale variabilità. Dobbiamo quindi utilizzare dei modelli di regressione (lineare, ecc.) differenziati in base alla natura della variabile:

- **Qualitative**: continue / discrete
- **Quantitative**: nominali / ordinali (scala di valori)

Bisogna inoltre distinguere tra variabili discrete (nel senso che non ha tanti valori distinti), oppure continua nel caso in cui abbia tanti valori distinti. Tra i modelli che utilizzeremo abbiamo:

- **Regressione lineare**: per variabili quantitative continue
- **Regressione logistica**: per variabili qualitative dicotomiche (dummy)
- **Regressione logistica multinomiale**: per variabili qualitative nominali con più modalità

L'equazione del modello di regressione lineare multipla è $Y = B_0 + B_1x_1 + B_2x_2 + \ldots + B_nx_n + e$ dove Y è la variabile dipendente, ovvero la variabile da spiegare tramite il modello: è di conseguenza la variabile aleatoria, ovvero una variabile che cambia in base ai valori stimati del modello. Le altre variabili che compaiono nel modello sono:

- **Variabili indipendenti x** (esplicative/covariate) che determinano il valore finale di Y e per questo non risultano aleatorie ma deterministiche (nel caso in cui ci sia solo una x allora avremo un modello lineare semplice)
- **Parametri incogniti B** (coefficienti delle x che determinano l'effetto di x su Y, ovvero gli effetti marginali di x su Y) da stimare con il modello in questione
- **Termine di errore e** (comprende le variabili non incluse nel modello) che rende diverso il vero valore di Y da quello stimato dal modello lineare (è una variabile aleatoria, essendo dipendente dal numero di variabili che riusciamo ad inserire nel modello)

ESERCITAZIONE 1 - INTRODUZIONE R

Iniziando con una variabile qualitativa vediamo come fare una tabella di frequenza ci permetta di ottenere informazioni utili sulla nostra matrice. Fare una distribuzione per frequenze relative con la funzione $prop.table$ ci permetterà di ottenere altre utili informazioni. Lo stesso non si può dire per la tabella di frequenza per variabili "continue" quantitative come il salario. Otterremmo delle frequenze bassissime per tanti tipi di modalità. Possiamo quindi usare degli istogrammi, per cui potremo anche mettere sull'asse y non le frequenze ma i valori delle modalità con la funzione $hist(salary, freq = 0)$. Possiamo vedere come la distribuzione sia inclinata verso destra, ovvero che la maggior parte dei valori si concentra nella fascia bassa dei salari.

Con il boxplot possiamo invece vedere la posizione delle modalità nei quartili della distribuzione e gli outlier: nel caso in cui la distribuzione sia perfettamente simmetrica, media e mediana saranno uguali. I dati del QQplot di permettono di osservare quanto la distribuzione sia distante da quella normale: più è la distanza dalla retta e maggiore sarà la distanza dalla normalità (non confondere

con la simmetria della distribuzione, siccome potrei avere una normale asimmetrica). Il boxplot è permette di osservare la vicinanza alla normale della distribuzione, in modo limitato: una normale avrà una scatola molto piccola e ampi baffi, siccome la maggior parte dei dati si concentrerebbe in una piccola sezione centrale della distribuzione. La correlazione determina la relazione lineare tra le due variabili calcolate: maggiore è la correlazione e maggiore sarà l'associazione lineare tra le variabili considerate, ovvero formeranno una linea retta a 45°. Possiamo quindi osservare adesso la relazione tra una variabile qualitativa (MBA) e una quantitativa (salario): dobbiamo vedere quale sia la distribuzione della variabile salari nei due sottocampioni con le due modalità della variabile MBA. Se le distribuzioni sono simili significa che la presenza di un MBA non avrà effetto sul mio salario come CEO, mentre una differenza significativa determinerà una influenza. Vediamo il box plot, uno ognuna delle due distribuzioni dei due sottocampioni, siccome una funzione di R riporta il paragone immediato dei due. Utilizzo quindi *boxplot(forbes94$salary~forbes94$MBA)* che mette due boxplot uno di fianco all'altro. L'unica cosa che rende differenti i due boxplot è la presenza di un maggior numero di outlier nel caso di assenza dell'MBA, mentre media/mediana risultano in sostanza identiche. Tramite la funzione tapply vediamo come la media risulti diversa a causa dei diversi outlier trovati nei boxplot (funzione che non va se ci sono degli NA).

LEZIONE 2 - MODELLO DI REGRESSIONE LINEARE MULTIPLA

Per utilizzare il modello dobbiamo disporre di un campione di N unità in cui vogliamo osservare la variabile Y, tutte le variabili x del modello e gli effetti B sulla variabile dipendente. Supponendo che Y sia il prezzo al m² di immobili di una città, con x_1 l'età dell'immobile e x_2 la distanza dalla stazione metropolitana. Avremo quindi il modello $Y = B_0 + B_1x_1 + B_2x_2 + e$ che esprime l'effetto di ogni variabile esplicativa sul prezzo al m² degli immobili. Per sfruttare tale funzione utilizzo il campione, nel quale troverò i dati necessari relativi alla Y e alle varie x in modo da stimare il valore dei parametri riferiti ad ogni variabile esplicativa. Avremo che la regressione sarà approssimabile tramite la funzione $Y_i = C_i + e_i$ per la quale ogni Y è data da una costante (che contiene le variabili esplicative ed i parametri, motivo per cui risulta deterministica ma anche ignota dato che non si conoscono i valori dei parametri), più il termine di errore con le variabili che non determinano una relazione lineare con il prezzo dell'immobile. Sarà necessario fare assunzioni relative al modello ed alle sue variabili qualitative/qualitative in relazione alle singole osservazioni (l'assunzione di e_1 si fa su Y_1, ecc.), per cui per le varie e_i avremo:

- **Assunzione di linearità** - $E(e_i) = 0$ per ogni x del modello, siccome per una proprietà del valore atteso (per la quale la parte costante a viene tolta dalla formula del valore atteso - $E(X + a) = E(X) + a$) che ci farà trovare un $E(Y_i) = C + E(e_i)$ ma siccome $E(Y_i) = 0$ allora anche $E(e_i) = 0$
- **Assunzione di indipendenza** - e_i , e_j sono indipendenti per ogni $i \neq j$ per cui avremo che la covarianza $cov(e_i, e_j) = 0$, dato che vi è assenza di correlazione a causa dell'assenza di ogni legame lineare
- **Assunzione di omoschedasticità** - $var(e_i) = sigma^2$ per ogni i = 1,2,… per cui la varianza non dipende dalle varie osservazioni del modello
- **Assunzione di normalità** - $e_i \sim N(0 , sigma^2)$ formula che formalmente contiene la prima e terza assunzione
- **Assunzione di assenza di perfetta multicollinearità** - in relazione alle variabili esplicative, ovvero nessuna delle variabili esplicative deve essere una perfetta combinazione lineare delle altre presenti nel modello (ovvero non potremo avere $x_3 = x_1 + x_2$ siccome avremmo perfetta multicollinearità ed avremmo grandi problemi di stima dei parametri B)

Le corrispondenti assunzioni sulle Y_i saranno:

- **Effetto della assunzione di linearità** - $E(Y_i) = B_0 + B_1x_1 + \ldots + B_nx_n$ che ci fa capire come il prezzo dell'immobile sia dato in media come funzione lineare delle due variabili x inserite nel modello (in media, siccome sul singolo immobile non vale l'assunzione di linearità per cui avremo anche e nella funzione, siccome non tutti gli immobili con uguale valore delle variabili esplicative avranno anche lo stesso prezzo finale)
- **Effetto della assunzione di indipendenza** - avremo che Y_i , Y_j sono indipendenti
- **Effetto della assunzione di omoschedasticità** - $var(Y_i) = sigma^2$
- **Effetto dell'assunzione di normalità** - $Y_i \sim N(B_0 + B_1x_1 + \ldots + B_nx_n , sigma^2)$

La prima assunzione ci permette di scrivere il modello sotto forma di $E(Y_i) = B_0 + B_1x_1 + \ldots + B_nx_n$ che corrisponde alla forma classica in cui troviamo anche il termine di errore. Per effettuare una inferenza dobbiamo utilizzare la stima dei parametri, non solo di B_0 ma anche degli altri parametri del modello, i quali rappresentano i coefficienti angolari delle rette di regressione. Il primo metodo per stimare i parametri è quello dei minimi quadrati ordinari (OLS) che consiste nella stima dei B tramite i valori b_0, b_1, ecc. che minimizzano la quantità $\sum[y_i - \hat{y}_i]^2$ dove $\hat{y}_i = b_0 + b_1x_1 + \ldots$ o nel caso della regressione semplice la stima $y_i = b_0 + b_1x$ che minimizza $\sum[y_i - (b_0 + b_1x_i)]^2$. Si tratta quindi di minimizzare la differenza tra i valori effettivamente osservati nel campione y_i e quello stimato dal modello di regressione con il valore atteso di Y (ovvero usiamo la stima $b_0 + b_ix_i$ della media di Y effettivamente reale $E(Y_i) = B_0 + B_1x_1 + \ldots + B_nx_n$). Vogliamo quindi ridurre la distanza tra i valori stimati dal modello e quelli ottenuti dal campione, stessa cosa che facciamo anche nella regressione multipla, dove il modello stimato $\hat{y}_i = b_0 + b_1x_1 + \ldots + b_nx_n$ rappresenta la stima della relazione in realtà esistente espressa dalla funzione $E(Y_i) = B_0 + B_1x_1 + \ldots + B_nx_n$. Per interpretare le varie stime dei parametri del modello usiamo b_1 come esempio generale: tale valore rappresenta la pendenza della retta di regressione per ogni valore assunto da x nel modello. Si tratta quindi della variazione della stima della media della variabile dipendente E(Y) data dalla variazione unitaria di x relativa al parametro, tenendo fisse le altre variabili (non importa il livello assunto da ogni variabile, solo che esse siano considerate costanti, per misurare l'influenza di una singola variabile su Y). Dentro questo modello, il valore di b_0 rappresenta una sorta di coefficiente di adattamento, il valore base da cui deve partire la retta di regressione influenzata dalle variabili del modello. Per realizzare un modello lineare in R utilizziamo la formula x <- lm (hosp_exp~age+days).

Ma ciò che abbiamo osservato nel campione tramite la regressione stimata $\hat{y}_i = b_0 + b_1x_1 + \ldots + b_nx_n$ posso estenderlo alla popolazione? Possiamo fare inferenza in relazione ai coefficienti B_0, B_1, ecc. in relazione anche alla creazione di verifiche d'ipotesi e intervalli di confidenza? Partendo con test di verifica d'ipotesi vorremo osservare se $b_1 = 0$ (ipotesi nulla) oppure se $b_1 \neq 0$ (ipotesi alternativa che dice come x_1 abbia un effetto su E(Y) ceteris paribus). Per vedere se l'effetto ottenuto dal campione sia un caso, oppure se sia effettivamente esistente, utilizzo tale test con le due ipotesi riferite alla popolazione in questione (ovvero ai parametri B, non alle stime b). Di conseguenza, confermare la ipotesi nulla porterà a definire come x_n non abbia alcun effetto su E(Y), secondo il campione preso. Tuttavia, anche se il test mi conferma che $b_1 = 10$ non significa che anche $B_1 = 10$, dato che si usa un campione con n osservazioni, motivo per cui sarà necessario utilizzare il quoziente $t_{oss} = b_1/s_1$ dove s_1 rappresenta lo standard error di b_1 ovvero un indicatore della variabilità della stima b_1 dentro il campione in questione. Di conseguenza t_{oss} rappresenterà la statistica test necessaria a vedere come la distribuzione t di Student con *n-p-1* gradi di libertà (dove p è il numero di variabili esplicative in modello). Questo perché il quoziente b_1/s_1 mi permette di eliminare l'effetto della variabilità delle stime in modo da poter paragonare b_1 con l'ipotesi nulla (se avessi $b_1 = 10$ con una variabilità pari a 5, ovvero b_1 compreso tra 5 e 15, è ovvio che non potrei mai arrivare ad un valore di $b_1 = 0$ per cui devo dividerlo per la variabilità).

Utilizzando la distribuzione t di Student dovremo rifiutare H_0 nel caso in cui, per un livello totale di rischio di rifiutare H_0 quando essa è vera (a), avremo un t_{oss} così alto da portarci a pensare che il valore b_1 sia così diverso da 0 da impedire di pensare che possa essere uguale a zero (H_0). In modo più preciso dovremo rifiutare H_0 nel caso in cui $|t_{oss}| \geq$ quantile di ordine 1 - a/2 della t di Student con n-p-1 gradi di libertà. Allo stesso modo dovremo rifiutare H_0 se il p-value $\leq$ a, dove il p-value si riferisce a $2P(T \geq |t_{oss}|)$ dato che devo considerarlo per entrambe le parti della distribuzione t di Student. Il rifiuto dell'ipotesi nulla non significa che sia falsa, ma solo che non vi sia significatività sufficiente per approvarla (quindi x_1 non è significativamente in grado di influenzare Y). Tali valori possono essere trovati con la funzione summary di R. Tuttavia, l'obiettivo sarebbe individuare non solo se la stima del B_n sia corrispondente ad una delle due ipotesi, ma anche capire quale intervallo definisca i limiti della stima nei campioni, ovviamente con confidenza pari ad 1-a. Avremo quindi che l'intervallo di confidenza sarà (b_1 - quantile x s_1 ; b_1 + quantile x s_1): secondo questo intervallo, concludiamo come con fiducia pari a 1-a, il parametro B_1 che sto stimando con b_1 apparterrà a tale intervallo. Se l'intervallo contiene lo zero, significa che con fiducia 1-a il parametro B_1 sia dentro il mio intervallo, per cui dovremo concludere che H_0 non sia rifiutata (perlomeno con confidenza del livello 1-a per la popolazione). Vediamo l'intervallo di confidenza con la funzione confint (), in cui se voglio mettere una diversa confidenza useremo confint (… , level = 0.9).

Dobbiamo però considerare eventuali problemi di previsione, ovvero eventuali problemi di stima dell'effetto della variabile esplicativa su quella dipendente. Abbiamo quindi problemi di:

- Previsione di E(Y) in corrispondenza di valori specificati delle variabili esplicative (es: se ci fissiamo dei valori di x_1 potremmo avere errori di stima) secondo la solita formula usata in precedenza $E(Y) = b_0 + b_1x_1 + \ldots + b_nx_n$
- Previsione di Y in corrispondenza di valori specificati delle variabili esplicative (in questo caso non ci riferiamo al valore medio di Y, ma al valore stesso di Y per ogni singolo soggetto della popolazione di riferimento) secondo $Y = B_0 + B_1x_1 + \ldots + B_nx_n + e$, che posso riscrivere come $E(Y) = B_0 + B_1x_1 + \ldots + B_nx_n$ siccome dovrò assumere *e* come un valore nullo, motivo per cui avremo E(Y) = Y e quindi la stessa previsione per il valore medio e il valore di ogni singola osservazione (due concetti ovviamente diversi).

Passando dalle stime puntuali agli intervalli di previsione, come tutti gli intervalli di confidenza si compongono con ($\hat{y}$ - ME ; $\hat{y}$ + ME) dove ME rappresenta il margine di errore, il quale dipende dal livello 1 - a, per cui maggiore il livello di confidenza 1 - a, maggiore sarà l'ampiezza dell'intervallo e quindi anche la probabilità di avere all'interno di esso la stima $\hat{y}$. Tuttavia, è ovvio come avere un intervallo infinito sia inutile, siccome l'obiettivo è trovare il minor intervallo possibile che contenga la stima voluta. L'unico modo per aumentare la probabilità senza dover aumentare la confidenza è quello di avere un maggior numero di osservazioni, in modo da ridurre la variabilità e avere degli intervalli più piccoli con uguale 1 - a. La considerazione importante è che l'intervallo di previsione per E(Y) è più piccolo di quello di Y a parità di tutto il resto (numero di osservazioni, 1 - a, ecc.): ciò deriva dal fatto che risulta molto più difficile prevedere il valore di singole osservazioni rispetto a quella della media di numerose osservazioni.

Possiamo anche condurre una valutazione dell'errore di previsione, che possiamo considerare solo attraverso l'utilizzo di un campione con previsione $E(Y) = b_0 + b_1x_1 + \ldots + b_nx_n$ dove applico criteri di previsione alle singole unità del campione, ottenendo $\hat{y} = b_0 + b_1x_1 + \ldots + b_nx_n$ per ogni singola osservazione. Possiamo quindi paragonare i valori previsti dalla funzione di previsione (la prima) con quelli ottenuti sostituendo le variabili reali esplicative: possiamo quindi fare la differenza tra i valori ottenuti per ottenere l'errore di previsione $\sum[y_i - \hat{y}_i]^2$.

Possiamo quindi introdurre nel modello delle variabili qualitative nel modello, prima di tutto nel caso in cui la variabile sia *dicotomica* (dummy), ovvero variabili che rappresentano la presenza o la

assenza di una certa modalità del soggetto. Questa tipologia potrà essere inserita nel modello come una variabile quantitativa, considerando ovviamente il possibile cambiamento di valore 0 e 1. Tale considerazione mi permette quindi di definire un effetto su Y diverso da quello delle quantitative, dato che dovremo considerare anche la variazione di valore (dobbiamo vedere quale sia l'impatto su Y della presenza della variabile qualitativa). La stima del modello sarà la stessa di prima, ovvero con la formula $\hat{y} = b_0 + b_1x_1 + \ldots + b_nx_n$. Considerando una formula contenente $x_3 = 1$ e una $x_3 = 0$, avremo che la differenza b_3x_3 rappresenterà la differenza media tra le due funzioni.

È quindi ovvio come b_3 non rappresenti l'effetto singolo di x_3 su Y ceteris paribus, siccome non c'è mai una situazione in cui avremo x_3 diverso da 0 oppure 1 (es: non avremo mai $x_3 = 2$ e quindi un effetto di x_3 su Y pari a $2 \times b_3$). Potremo tuttavia utilizzare lo stesso test di significatività per vedere se ci sia un effetto di x_3 su Y, per cui mi sto semplicemente chiedendo se la presenza di x_3 abbia un effetto su Y.

L'altra tipologia di variabile qualitativa è quello delle variabili non dicotomiche, ovvero che abbia un numero di modalità $k \geq 3$. L'unico modo per inserire un numero superiore di modalità dentro la mia equazione è quello di inserire un numero di dummy pari a k - 1 variabili qualitative. Nel caso in cui io voglia inserire tutte le tre modalità dovrò per forza togliere b_0 come variabile del modello. Anche in questo caso, l'effetto delle varie x_n riferite alle variabili dummy rappresenterà differenze verticali tra equazioni del modello (trovo quindi l'effetto su Y medio della presenza delle dummy nel mio modello stimato). Anche in questo caso potremo usare un test di verifica d'ipotesi mirato a controllare se sia significativa la differenza tra le Y derivanti dal modello contenente le dummy ed ogni modello senza la dummy in questione.

ESERCITAZIONE 2 - MODELLO DI REGRESSIONE LINEARE MULTIPLA

Riferendosi al file *Hospital* possiamo creare un modello tramite la *mod1<-lm(hosp_exp~days+surgery, data=HE)*, formula che restituisce il seguente risultato.

Questa formula risulta quindi uguale indipendentemente dalla presenza di una variabile qualitativa nel modello. In questo caso vediamo quindi come *surgeryyes*, la variabile qualitativa con x = 1, suggerisca un effetto su Y pari ad un totale statisticamente significativo di b = 568,66 con una valutazione del p-value = 7,50e-05. Importante sapere che R restituisce sempre l'effetto della variabile con un valore pari a 1, per cui non vedremo mai *surgeryno*. Devo anche sapere che secondo questo modello l'effetto di *days* sarà di 342,80 indipendentemente dal fatto che surgery sia 0/1.

Questo perché sappiamo come l'effetto di una variabile quantitativa su Y sia determinato solo alla assenza di variazioni di altre variabili esplicative, ma siccome la dummy può assumere un valore di 0/1 non avremo nessuna differenza che tale valore sia 0/1. Siccome ciò non risulta spesso vero, dovremo inserire delle *trasformazioni di variabili*, ma ciò lo vedremo più avanti. Importante sapere, inoltre, che gli effetti delle variabili su Y sono sempre considerate in media (es: se ho $b_1 = 10$ allora l'effetto di x_1 su Y sarà in media pari a tale valore), oltre al fatto che l'effetto considerato va rilevato sul totale della popolazione e non sul singolo campione. Passando ad un altro modello possiamo inserire le variabili del modello con *mod2<-lm(hosp_exp~days+age+surgery+pat_cond, data=HE)*.

Vediamo come R abbia eliminato una delle tre dummy, in modo da eliminare problemi di perfetta collinearità, cosa che ha portato alla presenza di due dummy nel modello in questione. In particolare, la variabile inserita *pat_cond* è in realtà formata da due dummy, per cui il totale di variabili del modello è 5 (non quattro come nella formula). Noto la significatività delle condizioni severe (*condsevere*), ma non per le condizioni medie di gravità (*condaverage*). In caso di variabili non

significative potrei eliminare le variabili, così da semplificare il modello tramite l'introduzione di una singola dummy che raggruppi le due modalità di severe.

In questo caso la nuova variabile raggrupperebbe la condizione *severa* e quella non severa, cosa che possiamo inserire tramite la funzione *HE$cond_severe<-factor(ifelse(HE$pat_cond=="severe", "severe", "not severe"))* che ci farà ottenere una variabile cond_severe che assuma il valore severe se pat_cond assume il valore severe, mentre assumerà il valore not severe se anche pat_cond fa lo stesso. Tale variabile cond_severe possiamo inserirla nel modello al posto della precedente, ottenendo modelli semplificati che contengano solo 4 variabili (dove cond_severe risulta in media significativa). Posso quindi usare *predict(mod3, data.frame(days=10, age=50, surgery="yes", cond_severe="not severe")* così da ottenere la predizione del modello per il valore medio di Y in base ai valori inseriti per variabili esplicative del modello. Inserendo anche un altro termine con *interval="confidence"* otterremo ogni valore relativo all'intervallo di confidenza per la popolazione, mentre con *interval="prediction"* ne otterremo l'intervallo di confidenza per i valori specificati delle variabili esplicative. Devo pensare inoltre a come la previsione per days=100 non avremo delle stime utili, siccome risulta un periodo di tempo eccessivamente lungo per realizzare stime utili: posso fare previsioni utili solo se i valori della previsione risultano abbastanza vicini ai valori registrati nel campione (non posso ottenere le stime utili per days=100 se days non è mai stata maggiore di 7 dentro il campione). Questo perché non posso essere sicuro che la previsione a lungo termine mantenga la stessa forma della funzione che ho teorizzato per il modello (per days=100 la funzione potrebbe essere non lineare).

LEZIONE 3 - BONTÀ DEL MODELLO DI REGRESSIONE

L'obiettivo è la valutazione della capacità esplicativa della variabilità di un modello di regressione lineare multipla, anche in relazione alla capacità previsionale. Abbiamo già accennato il modello teorico (Y) e il modello stimato riferito al campione ($\hat{y}$), oltre ad aver già visto le previsioni del mio modello in base alle variabili esplicative sostituite all'interno del modello. Possiamo quindi vedere le seguenti misure in grado di valutare la capacità esplicativa della variabilità:

- **Devianza residua (non spiegata)** - $\sum[y_i - \hat{y}_i]^2$
- **Devianza totale** - $\sum[y_i - \bar{y}_i]^2$
- **Devianza spiegata** - $\sum[\hat{y}_i - \bar{y}_i]^2$

In quanto a tali formule possiamo dire che vale la $\sum[y_i - \bar{y}_i]^2 = \sum[y_i - \hat{y}_i]^2 + \sum[\hat{y}_i - \bar{y}_i]^2$ che deriva dalle tre formule specificate. La devianza residua rappresenta un errore di previsione commesso a causa dell'utilizzo del modello in questione. La devianza totale rappresenta l'errore di previsione che si commette nel caso in cui si preveda il modello utilizzando sempre la media delle osservazioni, per cui senza utilizzare il modello di regressione (uso il modello nullo $Y = B_0 + e$). La terza rappresenta infine la differenza tra la devianza totale e residua, ovvero la riduzione dell'errore di previsione ottenuto usando le variabili della regressione (deve essere per forza positiva). Se quest'ultima sia molto ampia significa che l'utilizzo del modello permette di spiegare molto della variabilità della variabile dipendente Y (passaggio dal modello senza variabili a quello con variabili). Possiamo ora fare il coefficiente di determinazione $R^2 = \sum[\hat{y}_i - \bar{y}_i]^2 / \sum[y_i - \bar{y}_i]^2$ = dev spieg/dev tot = 1 - dev res/dev tot, che definisce quanta della variabilità totale sia stata spiegata dal modello utilizzato (dalle variabili utilizzate), oppure può essere vista come la riduzione proporzionale dell'errore previsione ottenuto utilizzando le variabili del modello (se considero la formula 1 - dev res/dev tot). Nel caso di un $R^2 = 0{,}2$ avremo che il modello spiega il 20% della variabilità totale, oppure che l'utilizzo del modello ha permesso di ridurre del 20% l'errore di previsione del modello rispetto al modello con la sola media. Le proprietà di R^2 sono diverse:

- Compreso tra 0-1 (siccome non posso spiegare meno di 0)
- $R^2 = 0$ se la dev spiegata è uguale a zero, per cui il modello non serve (dev tot = dev res), dato che la retta di previsione si colloca sull'asse delle x, mentre la media ricalca in modo perfetto le osservazioni del campione
- $R^2 = 1$ se la dev spiegata è uguale a uno, per cui il modello prevede perfettamente tutta la variabilità voluta (le osservazioni ricalcano perfettamente la retta di previsione che abbiamo generato con il modello)
- Nel caso di regressione semplice (p=1) avremo R^2 = coefficiente di correlazione lineare *r* (nel caso generale di regressione multipla $R^2 = r^2$ tra le y_i e $\hat{y}_i$ per cui non vi è differenza tra fare la correlazione tra y_i e x_i oppure tra $\hat{y}_i$ e x_i)
- R^2 non diminuisce se al modello si aggiungono variabili esplicative (anzi aumenta, per cui potrebbe essere utile prendere un R^2 aggiustato per evitare di inserire eccessive variabili solo per aumentare la variabilità spiegata, cosa che potrebbe aumentare la complessità dei calcoli e portare a problemi di collinearità). Avremo quindi un $R^2_{adj} = 1 - [\text{dev res}/(n-p-1)]/[\text{dev tot}/(n-1)]$ dove n = ampiezza del campione e p = numero di variabili esplicative nel modello in questione. Quindi se p aumenta avremo una riduzione del R^2_{adj} che risulta una misura del compromesso tra capacità esplicativa e semplicità del modello (molto utile per confrontare dal punto di vista descrittivo modelli con numero diverso di variabili). Parlo di modelli dal punto di vista descrittivo perché qui siamo ancora nella statistica descrittiva e non inferenziale (vedremo poi come confrontare modelli dal punto di vista inferenziale).
- $R^2_{adj} \leq R^2$ (sarà uguale solo nel caso di $R^2 = 0$, mentre è possibile che $R^2_{adj} < 0$ in specifiche situazioni molto particolari)

Importante sapere come il valore di R^2 dipenda dal contesto in cui si applica il modello, motivo per cui non è possibile utilizzare R^2 per valutare l'utilità del modello (in contesti finanziari abbiamo un valore molto alto per mancanza di effetti esterni sulla variabile in questione). Per valutare l'utilità del modello si usa il test F che vediamo ora. Ad un modello sono associate due quantità: devianza residua del modello $\sum[y_i - \hat{y}_i]^2$ e gradi di libertà (n-p-1).

Possiamo quindi definire il test F, il quale serve per valutare la significatività di gruppi di variabili, ovvero per il confronto inferenziale tra modelli *annidati* (*nested*, in cui uno sta dentro l'altro, ovvero nel caso in cui uno dei due modelli contenga le variabili dell'altro sommate ad altre). Si tratta per cui si capire se la differenza data dalle variabili extra tra due modelli permetta di ottenere risultati migliori rispetto al modello abbreviato (se le variabili extra sono significative, il modello completo sarà migliore di quello ridotto anche per un numero q di variabili inferiori al numero p di variabili per il modello completo). Questo test si riferisce all'ipotesi nulla per cui effetto delle variabili extra sia nullo, ovvero che non abbiano alcun effetto su Y:

- $H_0 : B_n = B_{n+1} = \ldots = 0$
- $H_1 : B_n \neq 0$ per almeno uno dei coefficienti

Il test F parziale è basato sulla statistica test $F = [(\text{dev res}_r - \text{dev res}_c)/(p-q)]/[\text{dev res}_c/(n-p-1)]$ che misura la differenza di capacità esplicativa tra il modello ridotto e completo. Se questa differenza è significativa allora guadagneremo molto nell'utilizzo del modello completo, motivo per cui valori grandi della statistica test porteranno al rifiuto dell'ipotesi nulla (ovvero dovremmo utilizzare quel modello completo). In particolare, si rifiuta H_0 a livello *a* se la statistica test è superiore al quantile di ordine *a* derivante dalla distribuzione F secondo un test unilaterale ($F_{oss} \geq F_a$). Esattamente come la distribuzione t di Student la distribuzione F non ha una sola distribuzione, ma dipende da valori

di *p-q* e *n-p-1* che rappresentano i gradi di libertà. Allo stesso modo rifiuto H_0 se p-value ≤ *a*, dove il p-value è definito dalla probabilità $P(F \geq F_{oss})$. Per farlo in R dobbiamo fare quei due modelli con la funzione *lm*, uno ridotto e uno completo, mentre il calcolo della statistica F deriva dalla funzione *anova(mod2, mod1)*, dove mod2 rappresenta quello ridotto (altrimenti fa calcoli al contrario): con la formula otteniamo anche il RSS (residual sum of squares), oltre che *gdl* e *somma dei quadrati*. Quindi il test F, osservando la formula, ci permette di osservare se la dev resid del modello ristretto risulti statisticamente superiore a quella del modello completo.

Possiamo quindi vedere il test F globale (caso particolare di quello parziale), che permette di fare il paragone tra il modello completo e ristretto. In sostanza è un test F parziale con q = 0, motivo per cui l'ipotesi nulla dovrà prevedere che tutti i coefficienti del modello completo siano uguali a zero, mentre quella alternativa che almeno uno sia diverso da zero (siccome il modello ridotto sarà dato dal modello nullo $Y = B_0 + e$). Per fare il test avremo $F = [(\text{dev res}_r - \text{dev res}_c)/p]/[\text{dev res}_c /(n-p-1)]$ che rappresenterà la statistica test in grado di determinare il risultato. In particolare, siccome quel modello ridotto non contiene variabili explicative, la varianza residua = varianza totale, cosa che ci porterà ad avere la formula $F = [(\text{dev tot} - \text{dev res})/p]/[\text{dev res}/(n-p-1)]$, ma siccome dev tot - dev res = dev spiegata allora avremo la formula finale per il test F globale che risponde alle ipotesi fatte finora $F = [\text{dev spieg}/p]/[\text{dev res}/(n-p-1)]$. Come sempre rifiutiamo H_0 se $F_{oss} \geq$ quantile della mia distribuzione F con p e n-p-1 gradi di libertà, oppure quando p-value < *a*. In questo caso avremo le variabili inserite nel modello completo come statisticamente significative: per vederlo con R dovrò solo fare la funzione *summary* della mia funzione di regressione.

Ovviamente, nel caso di regressione semplice il risultato del test F e test t sarà identico, siccome si tratterà di paragonare il modello nullo con quello contenente una sola variabile. Questo ci risulta ovvio se pensiamo che il test t permette di valutare le singole variabili esplicative, mentre il test F paragona l'intera regressione, ma siccome abbiamo una sola variabile avremo lo stesso risultato in entrambi i casi. In particolare, avremo una statistica $F = t^2$ nel caso di una variabile.

LEZIONE 4 - TRASFORMAZIONI DI VARIABILI

Passiamo ora alle trasformazioni di variabili esplicative in un modello di regressione lineare, in cui inseriamo anche le interazioni (che sono delle trasformazioni). Partiamo da due considerazioni che ci serviranno: prima di tutto l'effetto di una variabile x_i è lo stesso per qualunque livello fissato di tutte le variabili rimanenti nel modello, mentre la seconda considerazione richiede che l'effetto di una variabile x_i sia lo stesso a prescindere dal livello di partenza di x_i (se incremento x_i di 1 non ne avremo un effetto sul parametro B_i). Queste due ipotesi sono fondamentali nei modelli lineari, cosa che non sarà uguale per modelli non lineari ovviamente: il fatto che queste due considerazioni del modello lineare valgano per qualunque variabile esplicativa rende il modello lineare inadatto per la spiegazione di un gran numero di relazioni tra variabili (è improbabile avere relazioni lineari che soddisfino sempre queste caratteristiche, per cui potremo usare poche volte tali modelli). Nel caso in cui valgano entrambe le ipotesi, infatti, se consideriamo un dummy che rappresenti un prodotto nuovo con x = 1 ed uno non nuovo con x = 0, indipendentemente dal prodotto vecchio/nuovo per forza avrò lo stesso effetto sul fatturato Y (non è ovviamente realistica come teoria). Vogliamo per cui consentire effetti diversi di x_i per valori diversi delle altre variabili esplicative del modello, per cui utilizzeremo delle interazioni. Utilizziamo x_1 quantitativa e x_2 dummy, oltre che inserire dentro il modello un prodotto tra le due variabili pari a $B_3x_1x_2$ che ci permetterà di avere due casi:

- Se $x_2 = 0$ avremo $Y = B_0 + B_1x_1 + e$
- Se $x_2 = 1$ avremo $Y = B_0 + B_1x_1 + B_2 + B_3x_1 + e = (B_0 + B_2) + (B_1 + B_3)x_1 + e$

Quindi se $x_2 = 0$ avremo che l'effetto di x_1 su Y sarà dato solo da B_1 mentre se $x_2 = 1$ avremo l'effetto di x_1 su Y pari ad un valore di $(B_1 + B_3)$. Dall'altra parte, avere $x_2 = 1$ genererà un effetto B_2 su Y, ma noi ci interessiamo solo come x_2 sia in grado di generare effetti su x_1 tramite l'interazione. Dentro il grafico non avremo quindi solo una variazione dell'intercetta ma anche dell'inclinazione data dalla modifica dell'effetto di x_1 su Y rispetto al caso con $x_2 = 0$ (senza l'interazione avremmo solamente la modifica dell'intercetta): mentre l'intercetta rappresenta l'effetto di x_2 su Y, il coefficiente angolare rappresenta l'influenza di x_1 su Y in base ai due casi di x_2. Considerando il modello descritto avrò quindi tre variabili, dato che l'interazione conta come variabile ($p = 3$). In più potremmo sottoporre il modello al test t con $H_0 : B_3 = 0$, così da vedere se l'interazione risulti statisticamente significativa, ovvero se l'effetto di x_1 sulla Y cambi in base al valore ricoperto da x_2. Se non risulta significativa, potremo tornare al modello senza l'interazione. Per inserire tale interazione in R devo inserire una formula dentro il modello data da x_1*x_2, togliendo dalla formula della regressione le singole x_1 e x_2 che potremo valutare con la solita funzione summary. Se non è significatività è probabile che anche una delle altre due variabili relative all'intenzione diventi non significativa a causa di problemi di collinearità. Considerando il file Hospital, utilizzando le variabili *days* e *surgeryyes*, avendo valore della interazione pari a 46,11 significa che un giorno in più di ricovero per chi ha subito operazioni avrà un costo superiore pari a 46,11 rispetto a chi non ha avuto l'operazione (differenza tra l'effetto di x_1 con l'intervento e senza intervento).

Per valutare la significatività legata all'interazione posso usare il test t, mentre per quella di x_1 devo usare un test F parziale (l'effetto totale è dato da $B_1 + B_3$) con $H_0 : B_1 = B_3 = 0$. Infine, avendo visto casi con interazione tra variabili quantitative e dummy, ricordiamoci che potrebbe essere anche tra due variabili quantitative/dummy. Nel caso di due quantitative il processo sarà identico (vogliamo l'effetto di un aumento unitario di x_1 su Y per valori invariati di x_2): utilizzo il modello $\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$ che per un incremento unitario $x_1 + 1$ avremo $\hat{y} = b_0 + b_1x_1 + b_1 + b_2x_2 + b_3x_1x_2 + b_3x_2$ per cui avremo una differenza tra i due data da $b_1 + b_3x_2$ per cui lasciando fisso x_2 avremo avuto un effetto di x_1 su Y dato proprio da tale differenza influenzata da x_2 (differenza $b_1 + b_3x_2$). Nel caso di due dummy la situazione sarà ancora identica.

Siamo quindi stati in grado di risolvere il problema per cui l'effetto di una variabile doveva essere lo stesso indipendentemente dal livello fissato per le altre. Dobbiamo quindi risolvere la seconda, per la quale l'effetto di una x_i non cambia in base al livello di partenza di x_i stessa. Dobbiamo per cui rendere il modello non lineare, in modo da accomodare eventuali differenze all'interno del mio modello lineare, tramite trasformazioni di variabili quantitative. Ad esempio, posso utilizzare un logaritmo di x_1, il quadrato di x_1 o molto altro (se considero delle potenze, queste si aggiungono alla variabile originaria, per cui non si tratta di sostituire come per le altre trasformazioni come un logaritmo). Avendo eseguito la trasformazione, un incremento della variabile trasformata non sarà più diretto su Y, ma influenzato dalla trasformazione sulla variabile: dovremo quindi concentrarci sull'interpretazione dell'effetto della variabile trasformata. Considerando un modello con una sola trasformazione logaritmica della x (modello semi-logaritmico), ovvero senza trasformazione della variabile dipendente (modello logaritmico), con la funzione $\hat{y} = b_0 + b_1 \log x$ ci chiediamo quale sia l'effetto di x su $\hat{y}$. Possiamo farlo in due modi:

- Considerare logx come una nuova variabile, per cui un incremento unitario di logx avrà un aumento di b_1 su $\hat{y}$ (metodo poco utile)
- Chiamiamo l'incremento Δx, per cui avremo $\hat{y} = b_0 + b_1 \log (x + \Delta x)$, la cui differenza con la formula normale $\hat{y} = b_0 + b_1 \log x$ sarà data da $b_1 \log (1 + \Delta x/x)$ che per una proprietà della matematica con x relativamente piccolo potrà essere riscritto come $b_1 \Delta x/x$, dove $\Delta x/x$ sarà l'incremento % di x (questo valore $b_1 \Delta x/x$ rappresenta l'incremento di Y derivante dal mio incremento % per un valore di $\Delta x/x$, ma solo se l'incremento $\Delta x/x$ risulta piccolo). Essendo un cambiamento %, un incremento unitario su un totale di 10 sarà il 10%, mentre su totale di 100 sarà 1% per cui abbiamo risolto il secondo problema (ovviamente posso applicarlo in

caso di valori positivi di x dato che stiamo utilizzando logaritmi).

Utilizzando il logx non avremo più un grafico lineare su x, ma un grafico con trend logaritmico che potrebbe essere anche applicato alla Y (non alla x). In questo caso avremmo un grafico logaritmico che determini un effetto diretto da x su logY pari ad un valore di b_1 (se lo voglio dire su Y e non sul logY allora avremo che un incremento di x porterà ad un incremento % di Y pari a b_1). Infine, un modello con logY e anche logx porterà ad un incremento % di x rispetto ad una variazione % di Y pari ad un valore di $b_1 \Delta x/x$. Facendo una considerazione finale, se in un modello una variabile x_i entra in diversi termini, per valutarne la significatività dovremo fare un test F parziale (come detto prima per le interazioni). Nel caso di un modello come $Y = B_0 + B_1x_1 + B_2x_1^2 + B_3x_1^3 + B_2x_2 + e$ devo fare un test che comprenda $H_0 : B_1 = B_2 = B_3 = 0$. Utilizzando test parziali come $H_0 : B_2 = B_3 = 0$ mi permette di capire se il modello risulti lineare e quindi migliori/peggiori all'aggiunta di elementi non lineari (se l'effetto quadratico è significativo allora aggiungere queste variabili porterà ad una migliore previsione sulla popolazione), mentre un test t permette di capire l'influenza della singola variabile sulla Y. Per fare una trasformazione in R dobbiamo semplicemente inserire Log/potenza devo semplicemente creare una seconda variabile in cui inserisco quella vecchia trasformata con la formula *log(variabile)*, trasformazione che dovremmo inserire solo se utile (es: posso controllare con un plot se vale la pena fare la trasformazione).

Passiamo quindi all'analisi della multicollinearità, che si riferisce alla relazione lineare tra variabili esplicative. In particolare, tra le assunzioni abbiamo messo che non sia consentita la perfetta multi collinearità, mentre un certo livello di multicollinearità è sempre presente, per cui potremmo avere un livello di relazione lineare (correlazione) pari a 0,91. La situazione ideale sarebbe quella di una relazione lineare nulla tra variabili esplicative (indipendenti) ma fortemente correlate con Y, caso in cui il contributo delle singole variabili alla spiegazione di Y si sommerebbe, dato che nessuna delle altre variabili prenderebbe parte della variabilità spiegata dalle altre x_i (somma della variabilità che ogni variabile spiega di Y). Ci sono diversi elementi da considerare della multicollinearità:

- **Effetti della multicollinearità elevata**: la conseguenza maggiore è l'aumento degli standard error delle stime dei coefficienti delle variabili esplicative coinvolte nella multicollinearità, per cui una minore precisione (*inflation of variance*). Avendo un maggior SE avremo inferiori stime della statistica t di Student, con conseguente maggiore p-value e maggiore probabilità di non rifiutare l'ipotesi nulla anche se essa è falsa. Questo perché essendo correlate, le più variabili esplicative inserite nel modello spiegherebbero la stessa variabilità di Y, con minor guadagno in termini di precisione del modello stimato per l'inserimento di una variabile in più nel modello (è anche possibile che l'altra variabile diventi non significativa).
- **Metodi di rilevazione della multicollinearità elevata**: nel caso di poche variabili correlate potrei usare anche semplicemente un coefficiente di correlazione, ma nel caso in cui le mie variabili siano più di due devo usare altri metodi. Potrei valutare correlazioni a due a due, metodo che risulta impreciso siccome potrebbe esserci collinearità forte tra tre o più, motivo per cui posso usare lo strumento VIF (variance inflation factor). Essi risultano associati agli aumenti degli standard error, uno per ogni variabile esplicativa, per cui un aumento di VIF significherà un aumento degli standard error dovuti ad alta multicollinearità. Utilizzo per cui tanti $VIF_i = 1 / (1 - R_i^2)$ quante sono le x_i del modello, dove R_i^2 rappresenta il coefficiente di determinazione della regressione di x_i sulle rimanenti variabili esplicative. Ovviamente, se la variabile in questione non è quantitativa si utilizzeranno dei generalized VIF in grado d risolvere tale problema. Osservando la formula del VIF, maggiore il coefficiente tra delle variabili esplicative e maggiore sarà il VIF, il quale segnalerà come detto un aumento degli standard error della regressione. In generale, un VIF > 10 segnala una forte multicollinearità tra le esplicative, derivante quindi da $R_i^2 > 0,9$.

- **Rimedi alla multicollinearità elevata**: potremmo aumentare *n* del campione per evitare un aumento degli standard error derivanti dalla multicollinearità. In alternativa posso togliere variabili esplicative fortemente correlate con altre, cosa che però potrebbe togliere parte di quella spiegazione di Y che deriva da tale variabile (trade-off). Posso anche sostituire quella variabile correlata x_i con il relativo x_i - *media* x_i in modo da togliere l'eventuale collinearità che derivi dalla presenza di effetti quadratici (dato che inserire potenze nel modello insieme alla variabile stessa porterà a grande correlazione tra le due, siccome certe parti del grafico delle due saranno molto simili, come nel caso di x e x^2 che risultano avere y molto simili per valori bassi di x). Sostituendo x_i - *media* x_i alla x_i avremo una relazione molto meno lineare tra x_i e x_i^2 in modo da evitare la multicollinearità tra le due. Infine, potremmo utilizzare una analisi preliminare dei fattori: per delle variabili esplicative potrei prendere non le singole variabili, ma delle combinazioni lineari in modo da generarne di nuove (es: non metto x_1 e x_2 ma x_a derivante dalla combinazione lineare). Ciò rimuove la collinearità, ma distrugge il significato delle singole variabili (difficile interpretare il significato di una combinazione).

ESERCITAZIONE 3 - TRASFORMAZIONI

Utilizziamo il file Capm che riporta i rendimenti mensili di diversi titoli in borsa, un indice borsa e una obbligazione risk free, per un totale di 120 mesi in modo da calcolare il prezzo di mercato sulla base del modello Capm, oltre che la dipendenza del rendimento del titolo dall'andamento mercato (premio per il rischio). Dobbiamo quindi costruire prima il premio, trovato come differenza tra un rendimento dei singoli titoli e quello del titolo risk free. Ad esempio, per IBM faremo *capm$ex.IBM <- capm$IBM - capm$RKFREE*. Possiamo quindi provare a vedere la relazione tra i due rendimenti con un grafico, oppure calcolando anche la correlazione con la funzione *cor*. Calcolando il modello vediamo come la variabile sia significativa, sia guardando il test t che il test F (che è ovviamente lo stesso nel caso di una variabile sola). Vediamo comunque come il R^2 sia molto basso, ma ciò non ci interessa siccome le variabili sono significative anche se la variabilità totale è spiegata poco (con la correlazione calcolata prima, sappiamo che in una relazione lineare semplice il nostro $R^2 = r^2$). Con il R^2 trovato significa che il rendimento del titolo sarà dato per il 27,9% dall'andamento del mercato e la parte rimanente da altri elementi. Con la funzione *anova* per il modello avremo la *sum sqr* che rappresenta rappresenta la devianza spiegata da ex.MKT nel modello, oltre che la devianza residua facente parte del modello. La devianza totale sarà la somma dei due valori (0,11 + 0,30). Se osservo la funzione *summary* di prima, invece, vedo il *residua standard* error (la radice quadrata della stima di $sigma^2$). Dalle assunzioni ricordiamo che $sigma^2$ deve essere la stessa per ogni x, ma questa non ci viene data per cui dovremo stimarla con il residua standard error al quadrato.

Possiamo quindi aprire il file Web, in cui su 40 giorni sono rilevate le visite e vendite su un website per spiegare le vendite attraverso il numero di visite, in modo da fare previsioni. Facendo modello e funzione summary troviamo una relazione significativa che spiega il 69% delle vendite. Possiamo poi usare la funzione *predict(lm.web, data.frame(Web.visits = 80)* che restituisce la previsione per 80 visite. Possiamo anche inserire l'intervallo di confidenza per le vendite medie E(Y) inserendo una parte *confidence*, mentre per ottenere l'intervallo di confidenza per le vendite relative ad una sola Y devo usare la parte finale *predict* (dove l'intervallo sarà maggiore rispetto a quello della media).

Passando ora ad alcune considerazioni sull'analisi dei residui, tornando alle assunzioni sui residui abbiamo fatto le assunzioni di linearità, indipendenza, omoschedasticità, normalità e assenza di perfetta multicollinearità. Potremo fare una analisi di diagnostica per controllare quali assunzioni siano state rispettate nel modello, tramite l'analisi dei residui. Sappiamo che il residuo risulta una differenza $y_i - \hat{y}_i = e_i$ per ogni osservazione, quantità molto utile ad esempio nel caso delle devianze residue. Se osserviamo il modello di regressione $Y = B_0 + B_1x_1 + B_2x_2 + e$, osserviamo come e_i sia la differenza $Y_i - (B_0 + B_1x_1 + B_2x_2)$, dove la seconda parte dell'equazione rappresenta quindi $\hat{y}_i$ per cui

avremo stimato il residuo puntuale della nostra regressione. Tramite tali residui $y_i - \hat{y}_i = e_i$ siamo in grado di vedere se l'errore *e* del modello soddisfi le assunzioni dette in precedenza. Importante che tu sappia che i residui vengono spesso standardizzati tramite la divisione per la loro varianza (che in generale richiede di sottrarre prima la media, ma se è zero allora non c'è differenza), in modo da avere media = 0 e varianza = 1 per tutti i residui. Nel nostro caso, facendo un plot del modello vedi come vengano riportati sulle ordine i residui (non standardizzati) per ogni osservazione stimata $\hat{y}_i$ (*fitted values* sulle ascisse) del modello.

Per la condizione di linearità la media degli errori per ogni $\hat{y}_i$ dovrebbe essere nulla, mentre quella di omoschedasticità dice che dovrebbero avere la stessa varianza. Se gli $\hat{y}_i$ hanno tutti media nulla, per qualunque x_i dovremmo avere dei punti dispersi in modo omogeneo sopra e sotto zero, per cui non dovremmo avere livelli di x_i con punti del grafico solo sopra/sotto (se ciò non avviene allora la distribuzione delle osservazioni non rappresenta una retta, motivo per cui si chiama assunzione di linearità). Siccome il grafico riporta una situazione sufficientemente lineare possiamo dire che la assunzione risulta ragionevolmente soddisfatta. Tale grafico serve anche per verificare l'assunzione di omoschedasticità, dato che in ogni settore i punti dovrebbero avere la stessa dispersione attorno alla media (abbastanza soddisfatta).

Possiamo poi vedere il QQplot che confronta i residui della distribuzione standardizzata con quelli della normale, per cui ci dice se sia rispettata l'assunzione di normalità. Se tale assunzione fosse in totale rispettata dovremmo avere i punti sulla retta tratteggiata: nel nostro caso non risulta troppo rispettata, soprattutto a causa della presenza di poche osservazioni.

Passando invece al data frame CPS vediamo dati riferiti ad individui maschi in base a reddito, tipo di lavoro e tanti altri elementi. Vogliamo sapere come la retribuzione dipenda da caratteristiche che siano state include nel modello, per cui partendo con un istogramma vediamo la distribuzione che risulta fortemente asimmetrica verso destra, motivo per cui sarebbe utile inserire la trasformazione logaritmica in modo da ridurre l'importanza dei valori alti (compatta la distribuzione). Posso per cui utilizzare un modello che abbia log.wage come variabile dipendente Y. Facendo un plot riferito alla relazione log.wage ed esperienza risulta un andamento logaritmico, quasi quadratico per cui il livello del reddito debba ridursi per livelli alti di esperienza. Potrebbe quindi essere utile usare un effetto quadratico relativo all'esperienza, cosa che facciamo tramite *experience2 <- experience^2*. Tale formula ci permette poi di costruire il modello includendo le due variabili relative all'esperienza, educazione e provenienza etnica. Troviamo un modello fortemente significativo, per tutte variabili e per il modello nel suo complesso, con un residua standard error = 0,5839. Tutte le variabili sono significative, anche grazie alla grandezza del campione, per cui è molto probabile che esso risulti in realtà significativo per la popolazione. Importante notare come avendo inserito il quadrato della variabile experience, l'effetto di essa sul salario non sia lineare ma quadratica (per cui non avremo uguale variazione del salario per un aumento unitario dell'esperienza tra 20-25 anni o 50-55). Vedo anche che a parità di variabili un nero avrà un salario inferiore al caucasico, relazione che risulta in modo lineare differentemente da quella dell'esperienza.

LEZIONE 5 - ANALISI DEI RESIDUI e VALORI ANOMALI

Possiamo quindi passare ad una analisi dei residui in modo da verificare eventuali violazioni delle assunzioni fatte inizialmente. Utilizziamo una analisi grafica dei residui e_i rappresentanti stime del valore reale dell'errore del modello utilizzato (non sono vere stime, ma sono elementi che ci danno delle informazioni sull'errore). In un modello di regressione lineare semplice potremmo usare una analisi grafica basata su un semplice *plot* tra la variabile esplicativa e gli eventuali residui delle mie osservazioni, per capire le variazioni di quest'ultimi al variare della variabile esplicative (mentre in un modello di regressione multipla non possiamo, dato che dovremmo verificare le variazioni dei

residui in base ai valori di tutte le variabili esplicative). In un modello multiplo, il valore $\hat{y}_i$ riporta valori di tutte le variabili esplicative in base alle varie osservazioni, per cui solo un grafico con $\hat{y}_i$ ci permetterebbe di verificare gli effetti sui residui dati da vari livelli di $\hat{y}_i$ (plot). In modelli semplici, fare un grafico con x_i oppure con $\hat{y}_i$ non fa alcuna differenza, siccome sono legate da una relazione lineare semplice (siccome è la stessa cosa, tutti i grafici sui residui si fanno con $\hat{y}_i$). Possiamo quindi vedere un grafico dei residui che rispetti assunzioni di linearità, omoschedasticità e normalità (non usiamo quella di indipendenza per ora). Dovrebbe avere:

- Media settoriale dei residui pari a zero (non globale, per cui non mi riferisco all'interità del grafico ma per ogni valore/settore di $\hat{y}_i$)
- Varianza uguale per ogni $\hat{y}_i$ (in caso contrario avremmo dipendenza tra varianza e le varie osservazioni *i* del modello)
- Linearità nella relazione tra Y e x (se faccio un grafico con Y sulle ordinate e x sulle ascisse, siccome mettere $\hat{y}_i$ sarà uguale a mettere x_i in un modello lineare, dovremmo avere solo una relazione lineare tra le due, ovvero una retta a 45° con pendenza positiva). In particolare, la realizzazione della condizione di linearità dovrebbe portare un grafico dei residui con una retta orizzontale, siccome dovremmo avere un andamento lineare dei residui in base a ogni possibile valore di $\hat{y}_i$.

Per eliminare le eventuali violazioni delle assunzioni si possono applicare trasformazioni di tutte le variabili esplicative/dipendente. Come visto nell'esercitazione, usare il logaritmo permette di fare una modifica della distribuzione in modo da avvicinarsi alla distribuzione normale, oltre che per la violazione di eteroschedasticità siccome, come nel caso degli ultimi due grafici, potrei far ridurre la variabilità generale del grafico (riduco la distanza tra i punti per ogni valore di $\hat{y}_i$). Infine, sistema la violazione della linearità, spostandomi da una relazione quadratica ad una lineare, ecc.

Considerando i valori anomali, invece, ci stiamo riferendo ai valori che si discostano in modo tanto rilevante dagli altri punti. In particolare ci interessa sapere quanto tali punti siano significativi sui risultati del nostro modello (se ho un punto che rientra in più di una delle tipologie che vedremo è probabile che abbia un effetto significato sulla regressione). Si distinguono in:

- **Outlier**: osservazioni con alto residuo rispetto agli altri punti (generalmente con un valore di e_i standardizzata ≥ 3, in modo da rientrare dentro certi limiti della normale standard) che per questo si discostano tanto dalla regressione che abbiamo fatto
- **Punto ad alto leverage**: osservazioni con alto leverage (misura della *distanza* delle variabili esplicative dalla loro media), ovvero che si distanzia molto dalla media delle osservazioni riferite alla specifica variabile esplicativa
- **Punti influenti**: punti outlier/leva in grado di generare effetti sulla regressione, in generale punti che risultano sia leva che outlier (ma anche in misura inferiore outlier non leva che mi modificano solo la stima di b_0, mentre punti solo leva non avranno effetti sull'inclinazione b_1 della retta di regressione, per cui non devo per forza toglierli). Si parla di osservazioni in grado di generare effetti sulla regressione nel momento in cui vengono eliminati.

Possiamo quindi vedere quattro misure di influenza, le quali ci dicono se un punto risulta avere un certo livello di influenza sulla regressione. Ovviamente si basano sui residui e la media relativa alle osservazioni del campione:

- **Distanza di Cook**: ne abbiamo una per ogni osservazione, la quale misura il cambiamento complessivo delle stime e del comportamento del modello dovuto alla rimozione del punto (la distanza di Cook si basa sul leverage e residuo, mettendoli insieme). Se ho distanza > 1, generalmente il punto risulta influente.

- **DF Beta**: ne abbiamo uno per ogni variabile esplicativa e per ogni osservazione, sistema che misura quanto cambia la stima del B di ogni variabile esplicativa in seguito alla rimozione dell'osservazione potenzialmente influente (misura quanto cambia il B cambiando modello tramite la rimozione della variabile, che in R viene segnalato con un asterisco).

LEZIONE 6 - REGRESSIONE LOGISTICA

Tale modello è come sempre interessato alla spiegazione/previsione della Y attraverso le variabili esplicative, ma la regressione logistica ha una natura diversa della variabile dipendente: mentre in quello lineare rappresenta una variabile quantitativa, nel modello logistico Y risulta una variabile dicotomica riferita al successo (1) e insuccesso (0). Ad esempio, potremmo spiegare la variabile Y relativa al fallimento/non fallimento di una azienda, in modo da conoscere gli effetti delle singole variabili sul risultato. Considerando un modello di regressione lineare multipla, l'insieme delle x_i riporterà l'effetto delle singole variabili sulla E(Y), ovvero sulla media dei risultati derivanti dalla variabile dipendente Y. Non è possibile procedere allo stesso modo per la regressione logistica dato che sono disponibili solo due risultati opposti con probabilità *p* e *1-p*, per cui E(Y) = p, impossibile dato che il modello $E(Y) = B_0 + B_1x_1 + \dots + B_px_p$ potrebbe facilmente uscire dall'intervallo 0-1. Non è quindi possibile usare il modello lineare anche per le variabili dipendenti dicotomiche.

Possiamo quindi usare i modelli lineari generalizzati (MLG) che generalizzano il modello lineare così da non esprimere $\mu = E(Y)$ come funzione lineare delle x_i ma una funzione di μ come funzione lineare di x_i. Avremo quindi la funzione $g(\mu) = B_0 + B_1x_1 + \dots + B_px_p$ chiamata *link function* del mio modello, la quale risulta $g(\mu) = g(p) = \log(p/(1 - p))$ per cui anche $\log(p/(1 - p)) = B_0 + B_1x_1 + \dots + B_px_p$. La funzione g(p) viene detta funzione logit del modello per *p*, la quale risolve il problema del modello lineare incompatibile con una Y dicotomica dato che la funzione logit si muove dentro un intervallo $(-\infty , +\infty)$ per *p* compreso tra 0-1. Possiamo anche invertire la funzione logit:

Cercheremo di vedere che effetti avranno le variabili esplicative sulla variabile dipendente, oppure sulla media p, ma sempre in termini di probabilità. In particolare, parliamo di odds = p/(1 - p) che rappresentano la "probabilità" di ottenere un certo valore di Y. Ad esempio, se Y = 1 avremo che la probabilità p sarà data da p(Y = 1), con cui potremo calcolare l'odds di fallimento dell'azienda. Per cui posso esprimere la probabilità di tale evento in termini di probabilità p oppure in termini odds. Considerando p = 0,5 avremo una probabilità di 0,5 di avere Y = 1, oppure odds = 1 di avere Y = 1, anche se in generale si preferisce utilizzare gli odds. Essendo inoltre derivanti l'uno dall'altro, odds e probabilità risultano avere una corrispondenza biunivoca (es: si usano spesso gli odds all'interno delle scommesse, dove si esprime una relazione x:y per mostrare le probabilità di successo e quelle di insuccesso rapportate, ovvero la formula degli odds). Importante sapere anche che le probabilità variano tra 0-1 mentre gli odds tra $(0 , +\infty)$, ma usiamo log-odds essi varieranno tra $(-\infty , +\infty)$. Devo anche sapere che log-odds saranno corrispondenti al logit, siccome log-odds = $\log(p/(1 - p))$, oltre a sapere eventuali risultati classici degli odds:

- P = 0,5 corrisponde odds = 1
- P < 0,5 corrisponde odds < 1
- P > 0,5 corrisponde odds > 1

Possiamo quindi analizzare le stime dei coefficienti e interpretazioni delle stime degli effetti delle variabili esplicative. In generale, le stime dei B si trovano tramite il metodo generale della massima verosimiglianza, ma non ci interesseremo di vederlo. Otteniamo quindi il modello stimato in cui le stime dei B risultano basate su un campione (tramite la massima verosimiglianza), che poi potremo estendere alla popolazione tramite un test di significatività.

In relazione all'interpretazione delle stime degli effetti eventuali delle variabili esplicative tramite un primo esercizio che consideri fallimento/non fallimento dell'azienda. Nel database troviamo gli indici qr (quick ratio), indipfin (rapporto capitale proprio/capitale investito) e la forma giuridica in grado di distinguere Spa, Srl e Srl ad unico socio. Tramite queste tre variabili cercheremo di trovare l'influenza relativa sulla probabilità di fallire, su un campione di 66 aziende, in modo da definire il modello stimato in cui abbiamo inserito due variabili dummy relative alla forma giuridica. Con un modello ci riferiamo alla stima di un eventuale effetto stimato, analizzando gli effetti degli specifici beta relativi alle variabili.

Per vedere gli effetti della prima variabile IF, aumentiamo in modo unitario il valore di essa così da ottenere gli effetti sul log-odds stimato. Facendo la differenza tra IF+1 e la funzione iniziale F posso ottenere l'effetto di -0,13 che corrisponde alla stima del b_{IF}. Posso anche riscrivere la differenza in base alle seguenti formule, le quali mi dicono come ad un incremento unitario di IF fissate le altre variabili è associato un decremento di 0,13 dei log-odds.

Possiamo anche riscriverlo in termini di odds ratio: ad un incremento unitario di IF (a parità degli altri valori delle rimanenti variabili) è associato un odds ratio pari a $e^{b1} = e^{-0,13}$. Esso ci dice come varino gli odds del risultato 0-1 in base ad una variazione di una delle variabili del modello (es: un odds ratio = 1, che restituisce quindi $e^0 = 1$, ci dirà come un aumento unitario della variabile IF non restituirà alcuna differenza sugli odds di avere Y = 1, mentre un odds ratio > 1 dice che variazioni positive della variabile IF porti ad un aumento degli $odds_{IF+1}$ rispetto agli $odds_{IF}$ per cui non avrò alcun effetto sugli odds di avere Y = 1). Nel nostro caso, avendo un odds ratio = -0,13 (minore di 1) significa che un aumento della variabile IF comporti una riduzione degli odds del fallimento dato che $b_1 = -0,13$. Considerando come $e^{-0,13} = 0,88$ avremo che $odds_{IF+1} = 0,88 \times odds_{IF}$ che riporta come un incremento unitario della variabile IF comporti una riduzione degli odds di fallimento del 12% rispetto al caso senza aumento unitario (attenzione, non parlo di probabilità siccome non conosco di quanto vari la probabilità, per cui posso solo dire che avendo un b_1 negativo un aumento della variabile comporterà sicuramente una riduzione della probabilità di fallimento).

Importante sapere come si parli sempre di odds stimati, dato che potremo estenderlo a tutta al mia popolazione solo a seguito di un test d'inferenza (come per la regressione lineare normale). Per un veloce ripasso del modello logit devo sapere che stiamo cercando di spiegare gli effetti delle nostre variabili x_i con effetti B sulla funzione logit, la quale coinvolge la probabilità $p = p(Y = 1)$ tramite il modello stimato logit con le varie x_i e relativi b_i. In particolare, nel modello logit non possiamo fare una espressione diretta della Y tramite le variabili x_i ma solo in termini di E(Y). Inoltre, qualunque interpretazione degli effetti delle variabili del modello non posso essere riferite alla probabilità, ma solo in termini di odds di avere Y=1 tramite la formula seguente. Una volta aver stimato gli effetti possiamo riferirli all'intera popolazione tramite dei test d'inferenza, riferiti alle singole variabili, ai gruppi di variabili (confronto tra modelli annidati) o all'intero modello.

Iniziando con la verifica della significatività di singole variabili si utilizza un test d'ipotesi con una $H_0 : B_1 = 0$ ed una $H_1 : B_1 \neq 0$, come sempre riferiti ai B non stimati ma della popolazione. In tal caso il test ha una regione di rifiuto riferita al solito rapporto b_1/se_b che deve essere maggiore del solito quantile, non della t di Student, ma della normale standard (quantile $z_{1-a/2}$). Passando quindi al test per la verifica della significatività di un gruppo di variabili, utilizziamo due modelli, uno ridotto in cui inseriamo q variabili e uno completo con p variabili (dove p > q). Come prima, il test si riferisce alla ipotesi $H_0 : B_{q+1} = B_{q+2} = \ldots = B_p = 0$ ed una H_1 : almeno uno dei coefficienti $\neq 0$, dove ognuna di queste variabili si riferisce alle extra-variabili del modello completo rispetto al ridotto. Tale test non usa la statistica F: per fare il test utilizzeremo quindi la statistica X^2, la quale deriva dalla differenza tra le due devianze residue dei modelli.

Essendo tale differenza necessariamente ≥ 0, come nella distribuzione X^2 con p - q gradi di libertà, per cui dovremo rifiutare H_0 nel caso in cui la statistica $X^2 >$ quantile di ordine 1-a (non utilizzo un quantile di ordine 1-a/2 dato che si usa un test unilaterale). In particolare, se tra le extra-variabili

del modello completo troviamo tre dummy (che identificano quattro modalità possibili) stiamo in realtà verificando la significatività della singola variabile qualitativa. Tanto maggiore la differenza tra le devianze residue dei modelli ridotto-completo e maggiore sarà la statistica test che porterà a rifiutare l'ipotesi nulla (per cui è meglio il modello completo). Possiamo quindi passare al test di significatività globale, ovvero mirato a verificare la significatività rispetto al modello nullo senza le variabili scelte (quindi risulta un caso particolare del test precedente con $q = 0$). Le ipotesi sono le stesse di prima, con una ipotesi nulla che richieda che ogni singola variabile sia pari a zero. Posso quindi rifiutare l'ipotesi nulla solo se il modello nel suo complesso sia significativo, ovvero se una statistica $X^2 \geq$ quantile della distribuzione di ordine 1-a con p gradi di libertà (dato che $q = 0$).

Considerando il data frame Election, possiamo vedere una serie di variabili che influenzino la mia variabile dipendente Y relativa al fatto di essere eletti. Per utilizzare un GLM al posto di modelli di natura lineare possiamo usare la formula *mod1 <- glm(win ~ incumbent + money + attitude, family = binomial(logit))*. Facendo un summary troviamo la significatività delle variabili in base alla statistica test z. Osservando la variabile *incumbentyes*, il segno negativo significa che il fatto di esserlo riduce gli odds e quindi anche la probabilità di vincere, oltre al fatto che essendo pari a -1,6 (minore di 1) significa che la riduzione degli odds sarà pari all'80% (dato che $e^b = 0{,}2$). Osservando invece come la dummy moderate risulta significativa, mentre quella aggressive non lo è, capiamo che essere dei moderati aiuta a vincere mentre essere troppo difensivi/aggressivi non aiuta (per cui potremmo inserire solo la variabile relativi al moderato, secondo questo modello almeno). Le ultime due righe del summary mostrano la *residual deviance* (devianza del modello ristretto) e *null deviance* (devianza del modello nullo), le quali possono essere utilizzati nei test detti in precedenza. Possiamo quindi provare a togliere la variabile dummy *attitude*, per poi fare *anova(mod2, mod1, test = "Chisq")* così da ricevere la soluzione del test X^2 che paragoni modello ristretto-completo, che restituisce la grande significatività del modello completo (per cui è meglio mettere la variabile tolta). Sempre in questa tabella, la voce deviance rappresenta la devianza rimasta (differenza tra le due devianze). Per fare invece un test globale dobbiamo fare un modello senza variabili, per poi fare la stessa *anova* così da ottenere il paragone del modello completo-nullo.

LEZIONE 7 - PREVISIONI e REGRESSIONE LOGISTICA MULTIPLA

Abbiamo visto che nel modello di regressione lineare multipla per effettuare una previsione devo semplicemente sostituire i valori delle variabili nel modello, ottenendo la previsione della media o la previsione delle singole osservazioni Y. Abbiamo anche utilizzato la devianza residua come una misura dell'errore di previsione, siccome confronta ciò che abbiamo trovato nel campione con ciò che avremmo previsto per quella specifica osservazione ($\hat{y}_i$). Nella funzione logistica la previsione avviene in modo diverso siccome dovrò sostituire valori delle variabili dentro la stima della solita funzione logistica, dove $E(Y) = p$. Dobbiamo cercare di prevedere la probabilità $p(Y = 1)$, cosa che possiamo fare sostituendo valori delle x_i dentro la funzione logistica, così da ottenere il valore del logit e quindi anche la probabilità stimata p.

Per fare invece la valutazione della qualità della previsione possiamo procedere nello stesso modo della regressione lineare. Prevediamo quindi p_i in corrispondenza di ogni unità del campione, le quali saranno ovviamente caratterizzate da diversi valori delle x_i e quindi anche del logit con cui potremo derivare p_i. Vogliamo quindi fare la stessa cosa della previsione relativa al modello lineare normale, in cui siamo interessati a prevedere le varie $\hat{Y}_i$ in modo da fare il paragone con il valore reale e trovare lo scostamento. Il problema risulta quindi passare dalla previsione di p_i al $\hat{Y}_i$, cosa che possiamo fare mettendo:

Maggiore il numero di valori che troviamo uguali tra vari $\hat{Y}_i$ e Y_i migliore sarà la valutazione della qualità della previsione. Per fare ciò possiamo fare la cosiddetta matrice di confusione, che cerca di calcolare il numero di volte in cui abbiamo fatto giusta la previsione. Per valutare la qualità posso utilizzare un indicatore basato sul rapporto $(n_{11} + n_{00})/n$ che mostra il numero di osservazioni per cui abbiamo fatto previsioni corrette sul totale delle previsioni nel campione. In casi reali, ottenere

previsioni con $\hat{Y}_i = 1$ mentre $Y_i = 0$ potrebbe avere effetti decisamente differenti da un errore avuto al contrario, siccome spesso risulta molto peggio prevedere un successo rispetto all'insuccesso. Tali pesi diversi dei due errori portano a cambiare il peso del passaggio da p stimato a $\hat{Y}_i$ in modo da evitare il più possibile di avere previsioni dannose nel mio ambito.

Introduciamo adesso l'AIC (Akaike information criterion), indicatore descrittivo di qualità del mio modello che tiene conto tanto della capacità esplicativa quanto della semplicità del modello in base al numero di variabili (mentre la capacità esplicativa si riferisce alla devianza residua). Tale AIC si definisce come devianza residua + 2(p + 1) = -2 log L(M). Il modello è tanto migliore quanto più ne deriva un valore inferiore, valore decisamente opposto alla lettura data dal $R^2_{adjusted}$ che rispecchia un valore migliore più è alto. Possiamo fare una previsione in R tramite la funzione predict in cui si devono inserire le variabili da prevedere: nel nostro caso, relativamente al data frame delle elezioni possiamo inserire *predict(mod1, data.frame(money=2, incumbent="YES"), type="response")* che trova la previsione della probabilità di essere eletto per gli specifici valori delle variabili inserite. Per fare la valutazione della previsione possiamo inserire la la previsione della probabilità tramite la funzione *election$p_hat <- predict(mod1, type="response")* che inserisce un'altra colonna nel data frame che mi mostra la probabilità per ogni osservazione di essere rieletti per ogni osservazione (posso guardare singolarmente se ogni probabilità di essere rieletti coincide con l'effettivo risultato in base al mio criterio probabilità-Y). Avendo trovato la probabilità posso ora trovare $\hat{Y}_i$ tramite la funzione ifelse, con la funzione *election$y_hat <- ifelse(election$p_hat > 0.5, "WIN", "LOSE")* con cui abbiamo messo un'altra colonna contenente la previsione $\hat{Y}_i$ che potremo comparare con il risultato effettivo.

Possiamo anche vedere un caso di logit con interazioni, ad esempio nel caso in cui volessi valutare l'influenza derivante sia da incumbent che money, come sempre tramite la funzione glm in cui ne inserisco l'interazione tramite l'asterisco. In sostanza stiamo cercando di descrivere l'effetto della variabile money in base al fatto di essere incumbent o no. Facendo il summary avremo una riga in più rispetto a prima che riguarda l'interazione, la quale riguarda l'interazione: nel caso incumbent = 1 avremo un effetto di money dato dalla somma dei b_i relativi al money (0,69) ed all'interazione (3,38) per un totale di 4,07, mentre nel caso di nessun incumbent avremo un effetto pari solo a 0,69. Potremo quindi calcolare gli effetti sugli odds in base a tali valori di b_i nei due casi.

Possiamo quindi introdurre il modello logistico multinomiale, in cui la variabile dipendente risulta una variabile qualitativa (nominale nel nostro caso) almeno $r \geq 3$ modalità. Per costruire il modello devo considerare r-1 delle r modalità totali, dove la modalità esclusa costituirà il riferimento nella valutazione finale. Possiamo poi scrivere r-1 equazioni di regressione logistica che confortano prob. delle r-1 modalità scelte con la modalità di riferimento. Se consideriamo modalità A, B e C dove ne tolgo quella C, avremo due modelli formati da variabili con effetti indicizzati che mostrano tutti gli effetti della variabile di x_i in relazione al relativo rapporto di A o B con la variabile C. Siccome tali coefficienti sono diversi in base al rapporto, anche il relativo effetto sugli odds risulta diverso (per cui i due rapporti tra le probabilità p vengono chiamati odds condizionali per singola coppia, dato che supponendo esistano solo A e C oppure B e C entrami i rapporti mostreranno il rapporto tra le probabilità di successo e di insuccesso 1-p per ognuna delle due coppie, ovvero odds = p/(1-p)).

Considerando che $p_A + p_B + p_C = 1$, il modello realizzato sarà mirato a capire come sempre tutti gli effetti delle variabili esplicative del modello, però in relazione agli odds condizionali (motivo per cui ogni coefficiente B dei due modelli risultano indicizzati e per questo diversi tra le due formule). Il problema è quello di passare delle equazioni teoriche a quelle stimate, dove il metodo di stima è lo stesso del modello logit normale, ovvero la massima verosimiglianza che stima congiuntamente tutte le formule del modello siccome formate dalle stesse probabilità. Le equazioni stimate sono:

Si vuole studiare la dipendenza delle scelte di modalità di pagamento (Y), dove A rappresenta dei contanti, B la carta di debito e C quella di credito, mentre A rappresenta la modalità di riferimento. Il modello si riferisce a due variabili esplicative, una quantitativa e una qualitativa inserita in base a due variabili dummy per le tre modalità del reddito. Possiamo quindi iniziare a studiare tutti gli

effetti delle variabili esplicative, sia in relazione a B che C. Passando da AGE a AGE+1 per fare una differenza avremo:

Gli stessi risultati li potremo ottenere tramite il rapporto invece che differenza, così da avere i soliti log-odds ratio che potranno essere trasformati negli odds ratio togliendo il logaritmo. Avremo per cui che l'effetto di un aumento unitario di AGE per la prima formula avremo un aumento di odds pari al 0,007 = 0,7%, mentre per la seconda formula avremo una riduzione di 0,002 = 0,2% (per cui un aumento di AGE porta ad un aumento dell'utilizzo della carta di debito rispetto ai contanti con odds pari al 0,7%, ma anche contemporaneamente ad una riduzione pari al 0,2% degli odds di un utilizzo della carta di credito). In sostanza, l'aumento dell'età aumenta la probabilità di utilizzare la carta di debito rispetto ai contanti ma riduce la probabilità di usare la carta di credito.

Possiamo quindi provare a fare previsioni delle varie probabilità relative ai tre metodi, tramite un sistema di tre equazioni, dove le prime due sono relative ai log-odds stimati condizionati, e l'altra si riferisce all'equazione $p_A + p_B + p_C = 1$. Considerando un valore di AGE = 40 ed INC = average (per cui $INC_{low} = 0$) avremo un totale pari alle equazioni:

Togliendo il logaritmo e facendo il sistema avremo delle stime di $p_A = 0,177$ mentre per $p_B = 0,468$ e per $p_C = 0,355$. Sulle slide si trovano altre cose, ma all'esame saranno richieste solo le cose trattate a lezione (per cui non guardarle).

ESERCITAZIONE 4 - RIEPILOGO

Utilizzando l'esercitazione su BB chiamata Esame online 31 Agosto 2020 bisogna definire la prima domanda dove chiedere di dare l'assunzione di normalità, ovvero che i termini di errore abbiano la distribuzione di normalità (e_i). Per identificarlo possiamo usare il QQplot, in modo da vedere tutti i residui standardizzati/studentizzati. La seconda domanda dice di definire l'effetto di *shipping costs* sulla media delle vendite, la quale risulta maggiore nel caso di interazione con la variabile *product container box* (dato che se box = 1, ovvero se non è large ma standard, il coefficiente dell'interazione risulta positivo per cui aumenterà l'effetto di shipping costs sulle vendite) ed essendo significativo possiamo dire che l'impatto di shipping costs sulle vendite è significativamente maggiore per ogni livello di significatività. La terza domanda chiede di fare una previsione in relazione alla large box (standard box = 0), per cui avremo y = 3713,3311 + 33,688 x 10 = 4050,21. La quarta domanda porta a definire il primo grafico come caratterizzato da variabilità costante (omoschedasticità), mentre la quinta domanda rappresenta il primo grafico siccome si riferisce all'unico in cui cambia solamente l'intercetta della retta. La sesta domanda ci permette immediatamente di vedere come la dummy abbia come variabile di riferimento east, siccome le due variabili presenti si riferiscono al west e al south, mentre la risposta della domanda risulta quella in cui dice di dover fare un test d'ipotesi per verificare la significatività del modello (le altre risposte sono sbagliate). La settima domanda dice di stimare i log-odds, cosa che possiamo fare semplicemente inserendo valori nella regressione con i dati ottenuti da R per i coefficienti; avremo quindi un log-odds = 0,2293 - 0,001714 - 0,48 = -0,2524.

La domanda otto chiede di definire la frazione di errori sul totale, ovvero (403 + 702) / totale delle osservazioni di tutta la tabella. La domanda nove ci presenta dei log-odds condizionali relativi alle tre variabili; siccome abbiamo un cliente privato dovremo mettere customer segment = 1, cosa che porterà ad un effetto di *sales* su *region* pari a 0,04 (in base ai log-odds), per cui ad un incremento di sales unitario avremo un incremento degli odds di *east* rispetto a *west* del 4%. Avremo la stessa cosa per l'altra funzione che mette in relazione *south* e *west*. La decima domanda chiede semplicemente di fare la solita previsione tramite la sostituzione degli elementi (tramite il sistema troveremo totale della probabilità in % pari a 56,44).

Passando quindi all'Esame online 31 Agosto parte 2 possiamo vedere il *dataset* relativo all'esercizio precedente. Per la prima domanda facciamo la funzione lm e con summary troviamo la risposta, il cui effetto risulterà di 427,512 = 8 x 53 (possiamo moltiplicare per 8 sol perché stiamo usando un modelli lineare). Per la seconda domanda facciamo sempre un modello e guardiamo il R^2 che mette la spiegazione della variabilità pari a 0,19%. Per la terza domanda guardiamo come l'effetto della variabile *order priority standard* = 1 risulta pari a -27,27 (ovvero un valore negativo, per cui profitto medio è più basso rispetto al caso con variabile = 0, anche se il coefficiente non è significativo). Tale domanda ha come risposta il fatto che il profitto è maggiore per ordini con priorità alta al livello di significatività al 10%. Per la quarta domanda facciamo un modello con interazione, in cui dovremo inserire i valori della predizione con la funzione *predict(mod, data.frame(Product-container = "large box", Shipping.cost = 51)* che riporta un valore finale di 5431,42. La domanda cinque chiede di dare l'estremo inferiore per un intervallo di previsione che faremo sempre con la funzione predict in cui metteremo internal = "confidence" e level = 0,99 che restituirà 1163,05.