Esercizio 4
(Regressione multipla)

DATI

Il data set employee.sav (o employee.xls; fonte SPSS) contiene 474 dati relativi agli impiegati di un'azienda. Le variabili sono

ID	Employee Code
BDATE	Date of Birth
EDUC	Educational Level (years) (livello di istruzione, in anni di studio)
JOBCAT	Employment Category (1=clerical 2=custodial 3=manager)
SALARY	Current Salary (in $) (stipendio attuale)
SALBEGIN	Beginning Salary (in $) (stipendio iniziale)
JOBTIME	Months since Hire (months) (anzianità, mesi dall'assunzione)
PREVEXP	Previous Experience (months)
MINORITY	Minority Classification (0=no, 1=yes)
GENDER	Sex (1=male, 0=female)

Domanda 1

Vogliamo studiare da quali variabili dipenda lo stipendio (SALARY), utilizzando un modello di regressione lineare.

Analisi

Primi passi.

Il problema: ci chiediamo quali variabili, fra quelle disponibili, possano spiegare, ovvero consentire di prevedere, lo stipendio ("salary").

La variabile dipendente (detta anche *output*) dunque è lo stipendio, ed è quantitativa continua. Fra le variabili indipendenti (dette anche esplicative, o regressori, o *input*), alcune sono quantitative continue (stipendio iniziale), altre quantitative discrete (ad esempio, istruzione intesa come numero di anni di studio), altre sono qualitative (o "fattori"; ad esempio, il genere o la categoria).

Per un primo sguardo sui dati, è utile, se la numerosità non è troppo elevata, disegnare le nuvole di punti (o diagrammi di dispersione) per ciascuna coppia di variabili (quantitative), oppure calcolare la matrice di correlazione.

La tabella **Correlations** contiene i coefficienti di correlazione lineare tra le variabili che compaiono nel modello.

Correlations

		Current Salary	Beginning Salary	Previous Experience (months)	Educational Level (years)	Months since Hire
Pearson Correlation	Current Salary	1,000	,880	-,097	,661	,084
	Beginning Salary	,880	1,000	,045	,633	-,020
	Previous Experience (months)	-,097	,045	1,000	-,252	,003
	Educational Level (years)	,661	,633	-,252	1,000	,047
	Months since Hire	,084	-,020	,003	,047	1,000

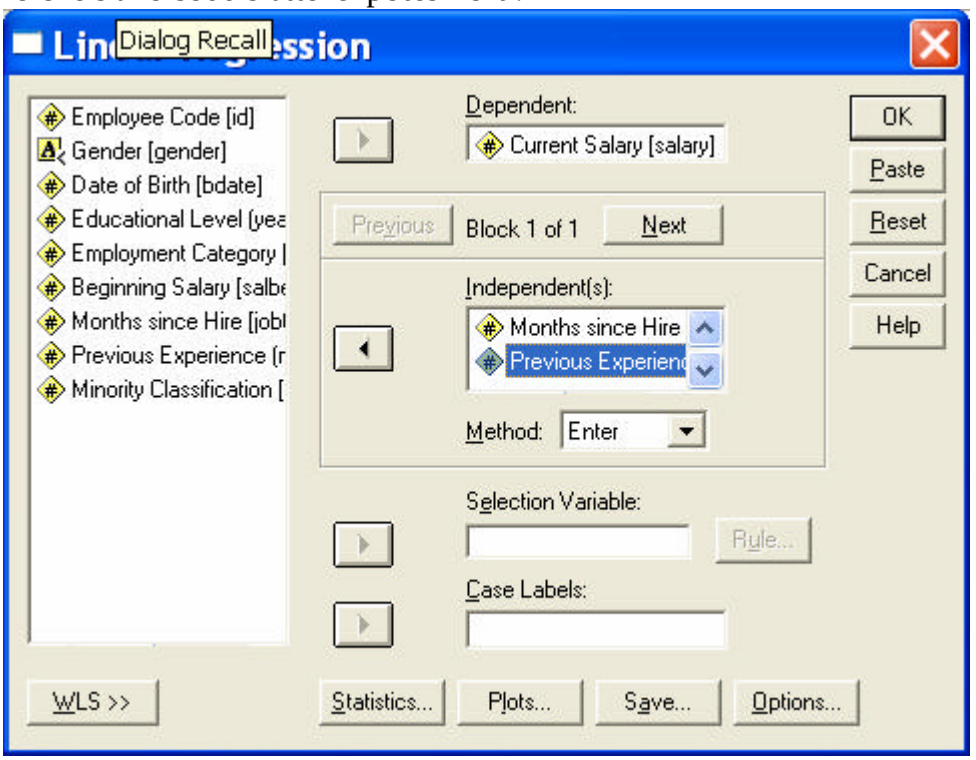
Osserviamo che le variabili SALBEGIN e EDUC risultano correlate positivamente con la variabile SALARY ed anche correlate positivamente tra loro. Potrebbe perciò essere sufficiente inserire una sola delle due (ragionevolmente, SALBEGIN) fra le variabili esplicative del modello.
(Nota: bisogna comunque fare attenzione nell'interpretare la matrice di correlazione. La correlazione fra due variabili può essere influenzata dalla presenza di una terza variabile e dalla sua interazione con le due in esame Per tener conto di questo, può essere preferibile calcolare i coefficienti di correlazione parziale).

1) Come primo esercizio, scegliamo come variabili esplicative tutte le variabili **quantitative** presenti nel data set.

Ipotizziamo, cioè, che valga il seguente modello:

$$SALARY = b_0 + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + b_4 \cdot PREVEXP + e$$

e supponiamo che siano soddisfatte le ipotesi forti.



Dal menu **Analyse**, selezioniamo **Regression** e quindi **Linear**. Selezioniamo come **Dependent variable** SALARY e come **Independent(s) variable** EDUC, JOBTIME, SLBEGIN, PREVEXP. Dalla finestra **Linear Regression** selezioniamo

- **Statistics** => **Descriptives, Collinearity Diagnostics, Part e partial correlations** e dalla finestra **Residuals** => **Casewise Diagnostics**, con **Outliers outside:** 3 **standard deviations.**
- **Save** => dalla finestra **Predicted values** => **Unstandardized**
  dalla finestra **Residuals** => **Standardized** (in questo modo vengono salvate nella **Window SPSS data editor**, contenente la matrice di dati le variabili PRE_1 (che contiene i valori stimati) e ZRE_1 (che contiene i residui standardizzati)
- **Plots** => **Histogram**, **Normal probability plot** e **produce all partial plots**

La tabella **Descriptive Statistics** contiene media e scarto quadratico medio (standard deviation) delle variabili che compaiono nel modello.

Descriptive Statistics

	Mean	Std. Deviation	N
Current Salary	$34,419.57	$17,075.66	474
Beginning Salary	$17,016.09	$7,870.64	474
Previous Experience (months)	95,86	104,59	474
Educational Level (years)	13,49	2,88	474
Months since Hire	81,11	10,06	474

La tabella **Correlations** contiene i coefficienti di correlazione lineare tra le variabili che compaiono nel modello.

Correlations

		Current Salary	Beginning Salary	Previous Experience (months)	Educational Level (years)	Months since Hire
Pearson Correlation	Current Salary	1,000	,880	-,097	,661	,084
	Beginning Salary	,880	1,000	,045	,633	-,020
	Previous Experience (months)	-,097	,045	1,000	-,252	,003
	Educational Level (years)	,661	,633	-,252	1,000	,047
	Months since Hire	,084	-,020	,003	,047	1,000

La tabella **ANOVA** contiene la somma dei quadrati del modello di regressione (Regression), la somma dei quadrati dei residui (Residuals) e la somma dei quadrati totali (Total). La statistica F data da

$$F = \frac{\text{Regression}/(n-k-1)}{\text{Residuals}/(n-1)}$$

dove $k = 4$ (numero di variabili esplicative che compaiono nel modello), è altamente significativa (con un p-value prossimo a zero), perciò si rifiuta l'ipotesi nulla del test

$$H_0: b_1 = b_2 = b_3 = b_4 = 0 \quad \text{contro } H_1: b_i \neq 0 \text{ per almeno un } i$$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,12E+11	4	2,79E+10	501,450	,000[a]
	Residual	2,61E+10	469	55728307		
	Total	1,38E+11	473			

a. Predictors: (Constant), Previous Experience (months), Months since Hire, Beginning Salary, Educational Level (years)

b. Dependent Variable: Current Salary

La tabella **Coefficients** contiene le stime dei parametri del modello (B), gli errori standard degli stimatori ottenuti con il metodo dei minimi quadrati (Std.Error) e le statistiche (t) e i p-values (Sig.) dei test di Students che verificano se i parametri siano significativamente diversi da zero.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-16149,7	3255,470		-4,961	,000					
	Educational Level (years)	669,914	165,596	,113	4,045	,000	,661	,184	,081	,516	1,937
	Beginning Salary	1,768	,059	,815	30,111	,000	,880	,812	,605	,551	1,814
	Months since Hire	161,486	34,246	,095	4,715	,000	,084	,213	,095	,992	1,008
	Previous Experience (months)	-17,303	3,528	-,106	-4,904	,000	-,097	-,221	-,099	,865	1,156

a. Dependent Variable: Current Salary

Il *p-value* del test che verifica $H_0: b_0 = 0$ contro $H_1: b_0 \neq 0$ è prossimo a zero, quindi a tutti i livelli di significatività, si rifiuta l'ipotesi che b_0 sia zero.
Analogamente i *p-value* dei test che verificano $H_0: b_i = 0$ contro $H_1: b_i \neq 0$, per $i = 1,\ldots,4$, sono prossimi a zero.

Dalla matrice di correlazione non emerge in maniera evidente un problema di multicollinearità, salvo per le variabili 'stipendio iniziale' (SALBEGIN) e anni di istruzione (EDUC), la cui correlazione è 0.661; le altre variabili esplicative non risultano fortemente correlate tra di loro. Tuttavia, per approfondire questo aspetto, abbiamo, a priori, selezionato Statistics Collinearity Diagnostics, per cui la tabella **Coefficients** contiene i valori delle statistiche: **Tolerance** e **VIF**.
*Per la variabile esplicativa i-esima la statistica **Tolerance** è data da*

$$Tolerance = 1 - R_i^2$$

*dove R_i^2 è il coefficiente di correlazione multipla tra la variabile i-esima e le altre variabili indipendenti. I valori di questa statistica sono compresi tra zero e uno. Quando questa statistica assume valori piccoli, allora la variabile è una combinazione lineare delle altre variabili indipendenti. La statistica **VIF (Variance Inflation Factor)** è il reciproco della statistica **Tolerance.** Un valore di soglia per la statistica VIF è rappresentato da 10 (corrispondente a Tolerance=0.10).*
Nel nostro caso tutti i valori della statistica VIF risultano inferiori al valore di soglia, e ciò non segnala problemi di multicollinearità.

Al fine di verificare se si è in presenza di un problema di multicollinearità, SPSS calcola anche

- Gli autovalori della matrice $X_{(s)}'X_{(s)}$ (Eigenvalue), dove X è la matrice disegno di dimensione $(n\times(p+1))$, con n numero di osservazioni e p numero di regressori, e $X_{(s)} = XD_{(s)}^{(-1)}$, con $D_{(s)} = diag\left(\left\|x_{[0]}\right\|,\ldots,\left\|x_{[k]}\right\|\right)$ e $x_{[j]}$ j-esimo vettore colonna della matrice X.
- i condition indices, dati da

$$\text{condition index}_j = \sqrt{l_{\max}/l_j}$$

dove l_j è il j-esimo autovalore della matrice $X_{(s)}'X_{(s)}$ e $l_{\max} = \max_{0\le j\le p} l_j$

Un condition index molto grande (maggiore di 30) indica un elevato grado di collinearità

- la matrice coefficient variance-decomposition. Ogni riga di tale matrice mostra la proporzione di varianza dello stimatore di ogni coefficiente di regressione attribuibile all'autovalore corrispondente.

L'ultima riga della matrice coefficient variance-decomposition mostra che la proporzione di varianza dello stimatore dell'intercetta e dello stimatore del coefficiente di JOBTIME, attribuibili al quinto autovalore, sono molto elevate, anche se il condition index rimane al di sotto del valore di soglia.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Educational Level (years)	Beginning Salary	Months since Hire	Previous Experience (months)
1	1	4,351	1,000	,00	,00	,00	,00	,01
	2	,500	2,948	,00	,00	,01	,00	,81
	3	,124	5,915	,01	,00	,53	,02	,01
	4	1,754E-02	15,749	,01	,87	,45	,18	,14
	5	6,834E-03	25,232	,97	,12	,02	,79	,03

a. Dependent Variable: Current Salary

Il modello lineare stimato è:

$$SALARY = -16149.7 + 669.914\cdot EDUC + 1.768\cdot SALBEGIN + 161.486\cdot JOBTIME - 17.303\cdot PREVEXP$$

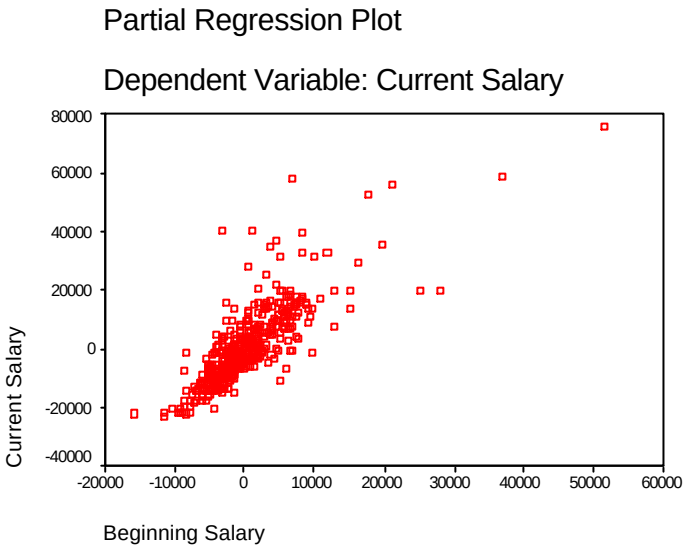
$\hat{b}_1 = 669.914\$$ rappresenta l'incremento che subisce la variabile dipendente SALARY, allorché il il numero di anni di studio (EDUC) aumenta di uno, fermi restando i valori assunti dalle altre variabili.

$\hat{b}_2 = 1.768\$$ rappresenta l'incremento che subisce la variabile dipendente SALARY, allorché il salario iniziale (SALBEGIN) subisce un incremento di 1$, fermi restando i valori assunti dalle altre variabili.

$\hat{b}_3 = 161.486$ 1$ rappresenta l'incremento che subisce la variabile dipendente SALARY, allorché il numero di mesi lavorativi nell'azienda (JOBTIME) subisce un incremento di uno, fermi restando i valori assunti dalle altre variabili.

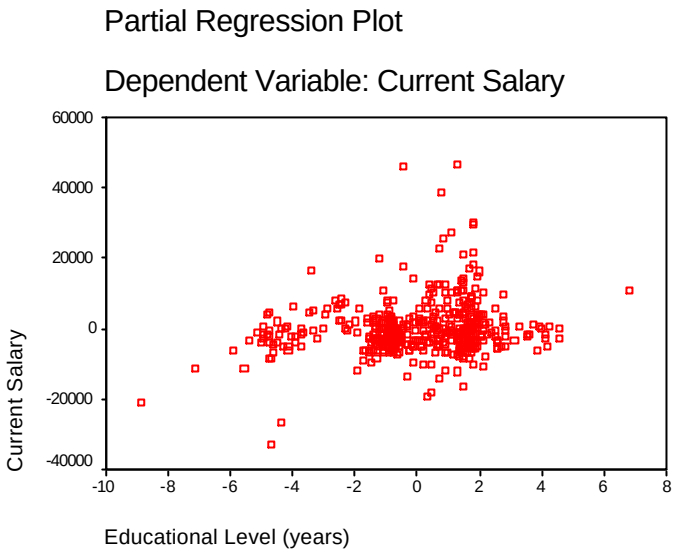
$\hat{b}_4$ = -17.303$ rappresenta il decremento che subisce la variabile dipendente SALARY, allorché il numero di anni di esperienza lavorativa aumenta di uno, fermi restando i valori assunti dalle altre variabili.

I **partial residuals plots** forniscono una versione grafica della correlazione parziale di ogni variabile esplicativa con la variabile dipendente, dopo che è stata rimossa l'influenza delle altre variabili.
Nel grafico relativo a SALBEGIN, riportato sotto, sono rappresentati sull'asse delle ascisse i residui della regressione di SALBEGIN sulle altre variabili esplicative (EDUC, JOBTIME, PREVEXP), sull'asse delle ordinate i residui della regressione di SALARY su SALBEGIN. La correlazione tra i due insiemi di residui rappresenta la correlazione parziale tra SALARY e SALBEGIN, al netto dell'influenza delle altre variabili esplicative. Vi sono, in particolare, due osservazioni che si allontanano dalla nuvola di punti sia rispetto all'asse delle ascisse che rispetto all'asse delle ordinate.



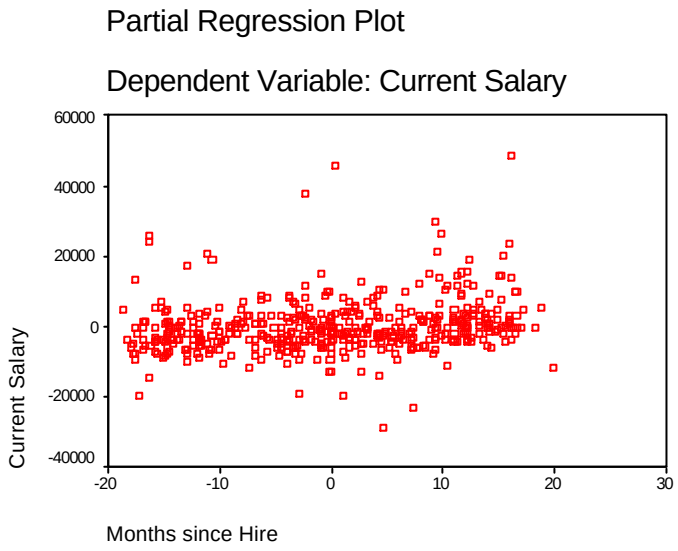
La correlazione semplice tra SALARY e SALBEGIN è 0.880. Dopo aver rimosso l'influenza delle altre variabili, questo legame diminuisce, ma rimane comunque molto forte. Il coefficiente di correlazione parziale tra SALARY e SALBEGIN, al netto dell'influenza delle altre variabili è pari a 0.812 (Tabella **Coefficients)**

Nel grafico relativo a EDUC, riportato sotto, sono rappresentati sull'asse delle ascisse i residui della regressione di EDUC sulle altre variabili esplicative (SALBEGIN, JOBTIME, PREVEXP) e sull'asse delle ordinate i residui della regressione di SALARY su EDUC. La correlazione tra i due insiemi di residui rappresenta la correlazione parziale tra SALARY e EDUC, al netto dell'influenza delle altre variabili.



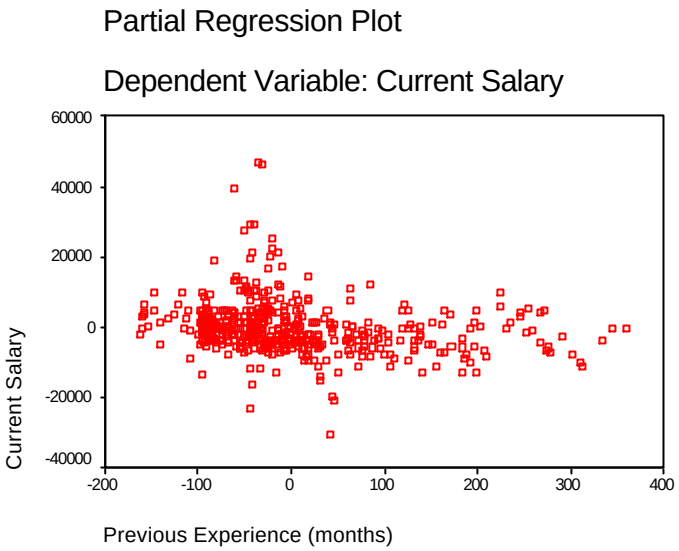
La correlazione semplice tra SALARY e EDUC è 0.661. Dopo aver rimosso l'influenza delle altre variabili esplicative, questo legame diminuisce notevolmente. Il coefficiente di correlazione parziale tra SALARY e EDUC, al netto dell'influenza delle altre variabili esplicative è pari a 0.184 (Tabella **Coefficients**).

Analogamente è costruito il grafico relativo a JOBTIME.



La correlazione semplice tra SALARY e JOBTIME è 0.084. Dopo aver rimosso l'influenza delle altre variabili esplicative, questo legame aumenta.. Il coefficiente di correlazione parziale tra SALARY e JOBTIME, al netto dell'influenza delle altre variabili esplicative, è pari a 0.213 (Tabella **Coefficients**).

Analogamente, il grafico relativo a PREVEXP illustra la correlazione parziale tra SALARY e PREVEXP, al netto dell'influenza delle altre variabili.



La correlazione semplice tra SALARY e PREVEXP è -0.097. Dopo aver rimosso l'influenza delle altre variabili esplicative, questo legame aumenta.. Il coefficiente di correlazione parziale tra SALARY e PREVEXP, al netto dell'influenza delle altre variabili esplicative è pari a -0.221. (Tabella **Coefficients**)

Il coefficiente di determinazione multipla R^2, che possiamo leggere sull'output riportato sotto, è pari a 0.810. Le variabili SALBEGIN, EDUC, JOBTIME, PREVEXP spiegano circa il 81% della variabilità di SALARY.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,900[a]	,810	,809	$7,465.14

a. Predictors: (Constant), Previous Experience (months), Months since Hire, Beginning Salary, Educational Level (years)

b. Dependent Variable: Current Salary

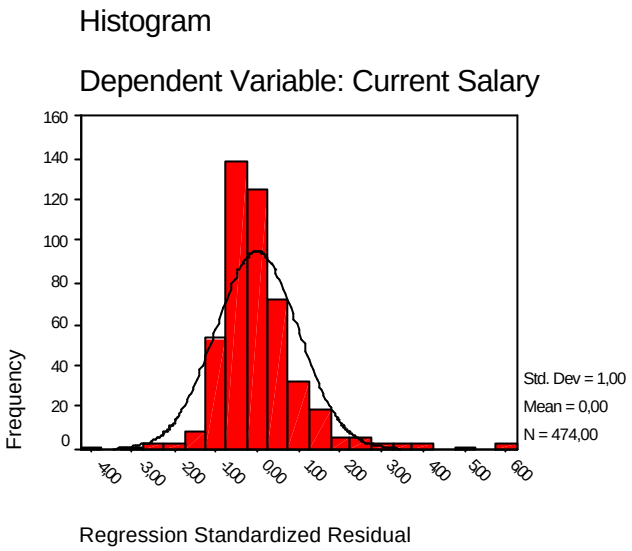
Analisi dei residui.

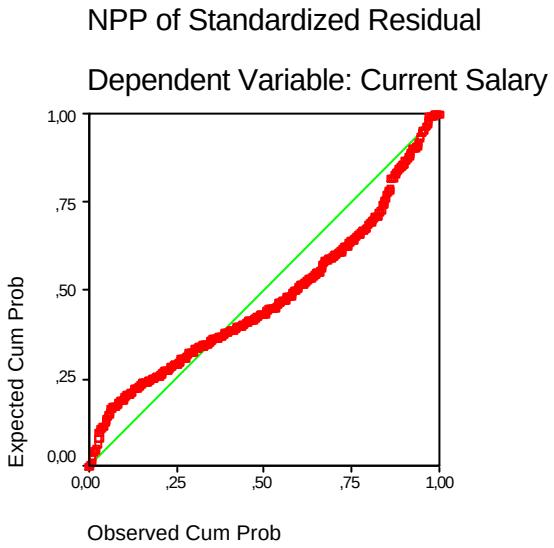
Un ulteriore strumento per controllare la bontà di un modello di regressione è dato dall'analisi dei residui. Se sono verificate le ipotesi forti

- i residui sono normali di media zero e varianza costante;
- i residui sono indipendenti
- i residui e i valori stimati sono indipendenti

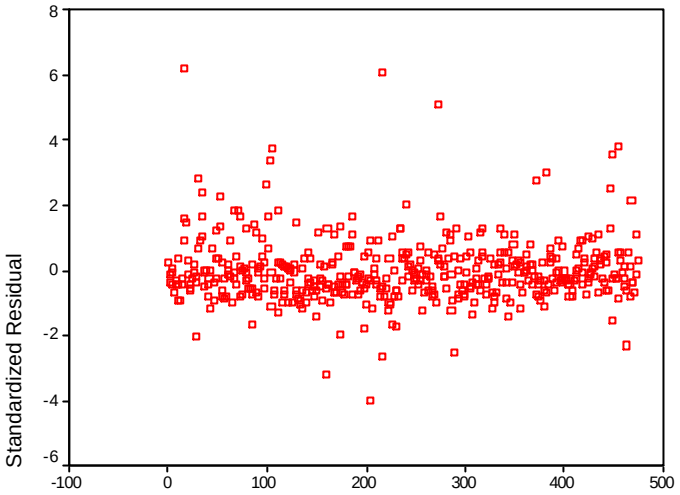
I due grafici successivi, un istogramma e un *normal probability plot* (NPP) dei residui standardizzati, sono utilizzati per verificare se sia plausibile l'assunzione di normalità dei residui. Come possiamo osservare i residui seguono approssimativamente una distribuzione normale, sebbene sia riscontrabile una certa asimmetria nei dati, e l'istogramma risulti più appuntito. Nel NPP, i punti tendono a disporsi, anche se con approssimazione non del tutto soddisfacente, lungo una retta. Si può concludere tuttavia che non c'è grossa evidenza di una forte violazione dell'ipotesi di normalità.

.





Il plot dei residui standardizzati rispetto agli indici del data set, riportato sotto, mostra che vi sono alcune osservazioni con residui standardizzati superiori a 3 in valore assoluto (si veda anche la tabella **Casewise Diagnostics**).

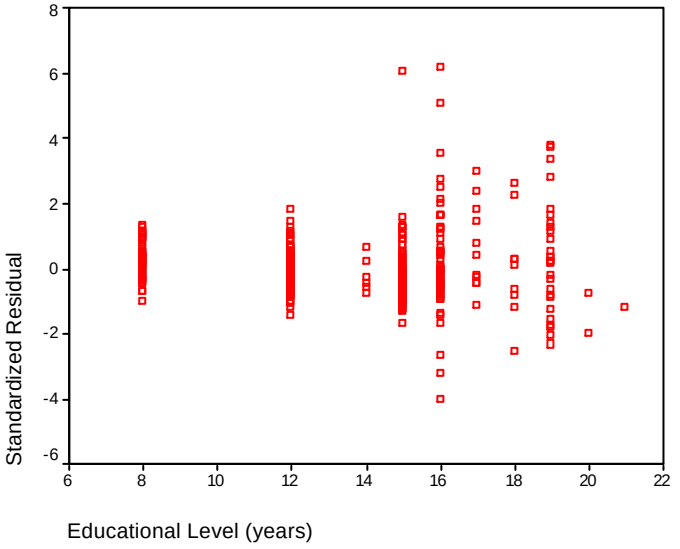


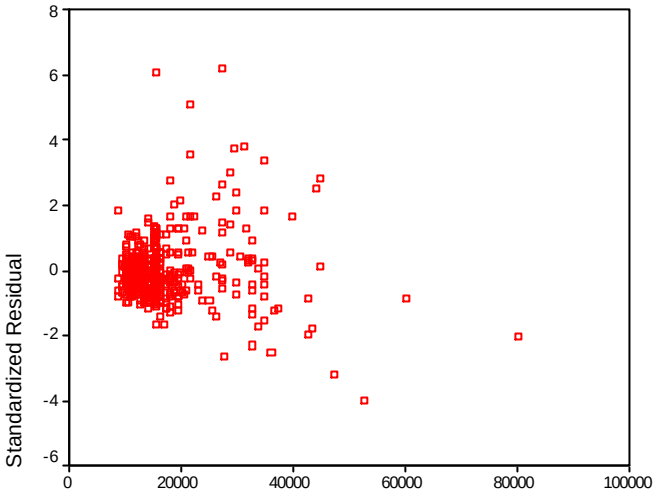
Casewise Diagnostics[a]

Case Number	Std. Residual	Current Salary
18	6,173	$103,750
103	3,348	$97,000
106	3,781	$91,250
160	-3,194	$66,000
205	-3,965	$66,750
218	6,108	$80,000
274	5,113	$83,750
449	3,590	$70,000
454	3,831	$90,625

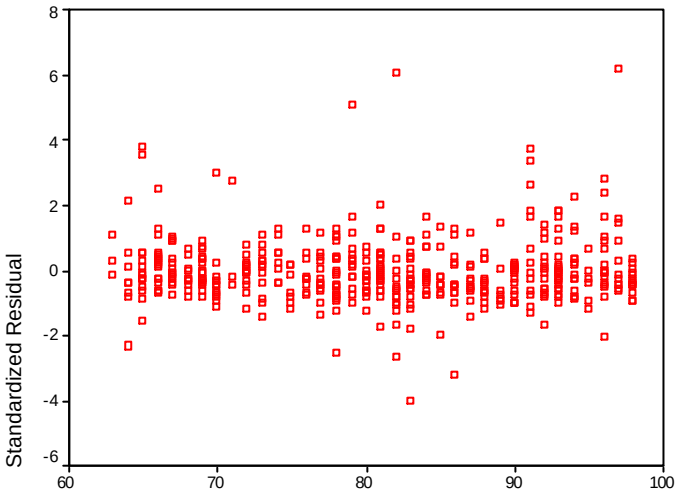
a. Dependent Variable: Current Salary

Sono riportati di seguito i plots dei residui standardizzati rispetto alle variabili esplicative.

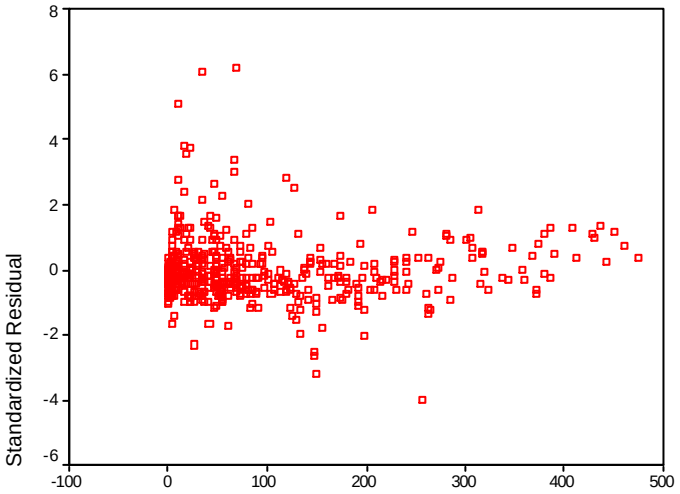




Beginning Salary

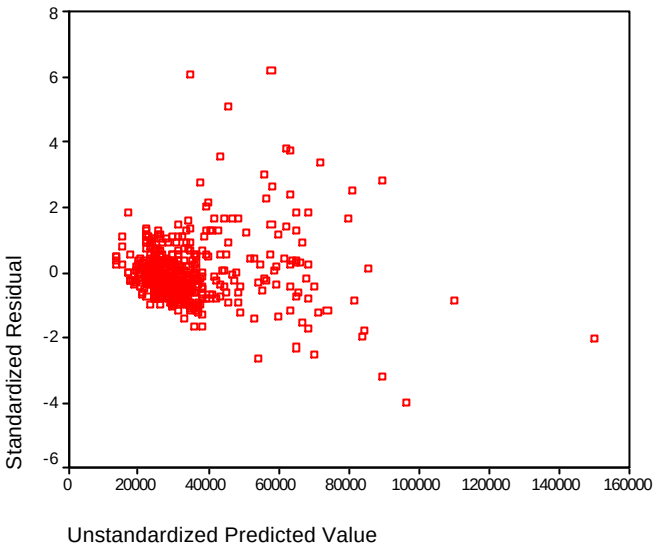


Months since Hire



Previous Experience (months)

Se sono verificate le ipotesi forti, residui standardizzati e valori stimati sono indipendenti, quindi la nuvola di punti nel grafico riportato sotto deve disporsi in modo casuale.



Il grafico sembra mostrare presenza di eteroschedasticità con varianza crescente. Ricordiamo però che i dati sono riferiti a impiegati uomini e donne; l'eteroschedasticità potrebbe essere dovuta ad una diversa variabilità all'interno dei due gruppi.
L'analisi successiva inserisce l'informazione GENDER (uomo/donna) nel modello.

2) Nell'analisi precedente abbiamo considerato come regressori le sole variabili quantitative. Inseriamo ora tra le variabili esplicative la variabile qualitativa (*dummy*) GENDER, a valori 1 (maschio) e 0 (femmina). Ipotizziamo, cioè, che valga il seguente modello:

$$SALARY = b_0 + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + b_4 \cdot PREVEXP + g_1 \cdot GENDER + e$$

e supponiamo che siano soddisfatte le ipotesi forti.

Dal menu **Analyse**, selezioniamo **Regression** e quindi **Linear**. Selezioniamo come **Dependent variable** SALARY e come **Independent(s) variable** EDUC, JOBTIME, SLBEGIN, PREVEXP, GENDER. Dalla finestra **Linear Regression** selezioniamo

- **Statistics** => dalla finestra **Residuals** => **Casewise Diagnostics**, con **Outliers outside:** 3 **standard deviations.**
- **Save** => dalla finestra **Predicted values** => **Unstandardized**
  dalla finestra **Residuals** => **Standardized** (in questo modo vengono salvate nella **Window SPSS data editor**, contenente la matrice di dati, le variabili PRE_1 (che contiene i valori stimati) e la variabile ZRE_1 (che contiene i residui standardizzati)
- **Plots** => **Histogram**, **Normal probability plot**

La statistica F contenuta nella tabella **ANOVA**

$$F = \frac{\text{Regression}/(n-k-1)}{\text{Residuals}/(n-1)}$$

dove $k = 5$ (numero di variabili esplicative che compaiono nel modello), è altamente significativa (con un p-value prossimo a zero), perciò si rifiuta l'ipotesi nulla del test che i coefficienti siano tutti nulli.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,12E+11	5	2,24E+10	408,692	,000[a]
	Residual	2,57E+10	468	54914875		
	Total	1,38E+11	473			

a. Predictors: (Constant), Sex, Months since Hire, Previous Experience (months), Beginning Salary, Educational Level (years)

b. Dependent Variable: Current Salary

I coefficienti delle variabili esplicative, contenuti nella tabella **Coefficients**, risultano tutti significativamente diversi da zero.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-14782,9	3267,788		-4,524	,000
	Beginning Salary	1,723	,061	,794	28,472	,000
	Months since Hire	154,536	34,085	,091	4,534	,000
	Previous Experience (months)	-19,436	3,583	-,119	-5,424	,000
	Educational Level (years)	593,031	166,630	,100	3,559	,000
	Gender	2232,917	792,078	,065	2,819	,005

a. Dependent Variable: Current Salary

Il modello lineare stimato è dato da

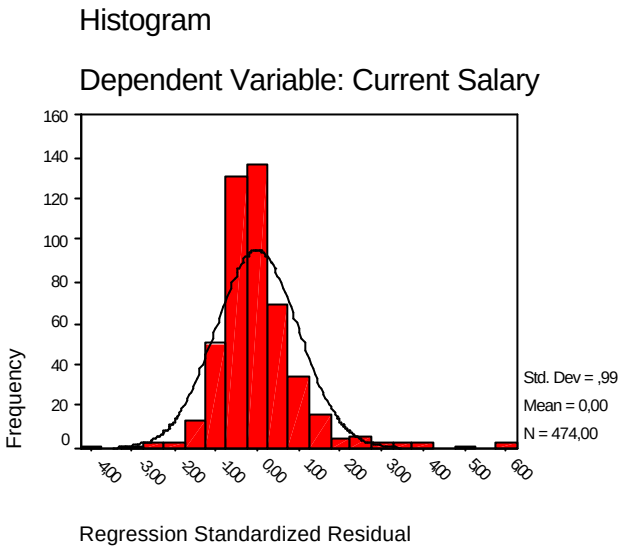
$$SALARY = \text{-}12549.983 + 93.031 \cdot EDUC + 1{,}723 \cdot SALBEGIN +$$
$$+154.536 \cdot JOBTIME - 19.436 \cdot PREVEXP \quad \text{se} \quad GENDER = 1 \quad \text{(male)}$$
$$SALARY = -14782.9 + 593.031 \cdot EDUC + 1.723 \cdot SALBEGIN +$$
$$+154.536 \cdot JOBTIME - 19.436 \cdot PREVEXP \quad \text{se} \quad GENDER = 0 \quad \text{(female)}$$

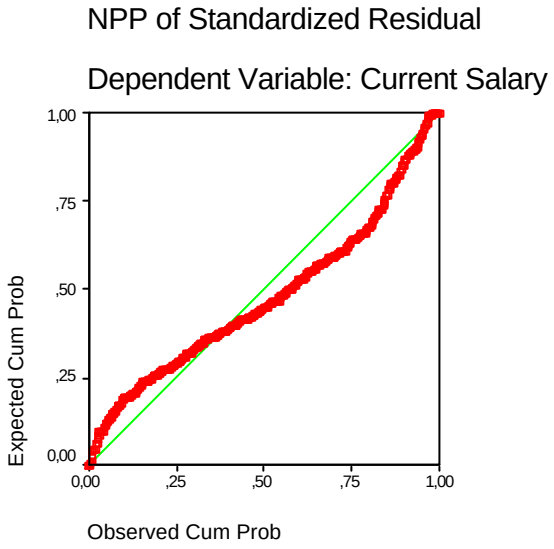
Il coefficiente di determinazione R^2, nella tabella **Model Summary**, è pari a 0.814 (leggermente superiore rispetto al coefficiente R^2 del modello con quattro regressori). Tuttavia $R^2 = \dfrac{\text{Regression}}{\text{Total}} = 1 - \dfrac{\text{Residual}}{\text{Total}}$ cresce all'aumentare del numero di regressori, per cui è utile, per confrontare due modelli, considerare

$$R^2_{ad} = 1 - \frac{\text{Residual}/(n\text{-}k\text{-}1)}{\text{Total}/(n\text{-}1)} \quad \text{(Adjusted R square)}$$

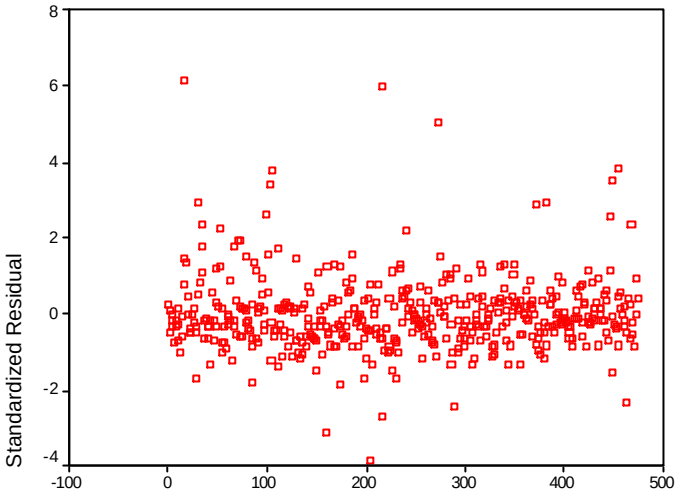
che tiene conto del numero k di regressori. Si ha $R^2_{ad} = 0.809$ per il modello con quattro regressori stimato al punto (1), e $R^2_{ad} = 0.812$ per il modello con cinque regressori, per cui la variabile GENDER si dimostra utile nello spiegare la variabile dipendente SALARY (c'è differenza fra uomini e donne..).

L'istogramma e il *normal probability plot* dei residui standardizzati, riportati sotto, mostrano una maggiore simmetria rispetto agli stessi grafici relativi al modello con quattro regressori (soprattutto tra 0 e 0.25). E' ancora presente tuttavia uno scostamento dalla normalità.





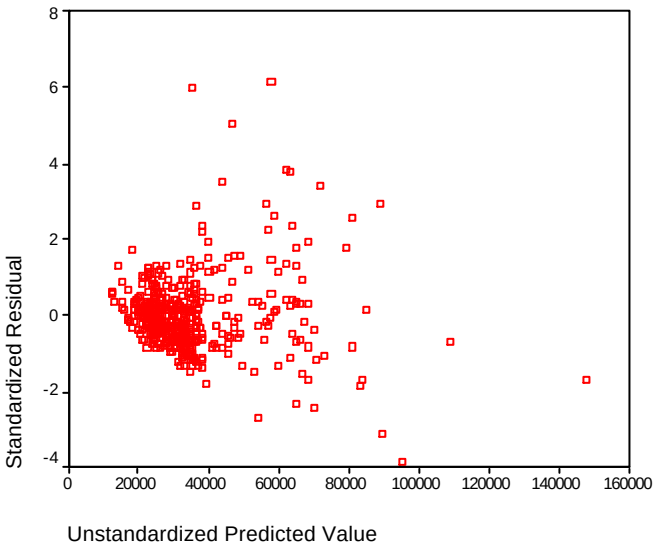
Sono riportati sotto i grafici dei residui standardizzati (ZRE_1), rispetto agli indici del data set (ID) e rispetto ai valori stimati (PRE_1) rispettivamente. Non c'è evidenza di un marcato miglioramento rispetto agli stessi grafici relativi al modello con quattro regressori.



Casewise Diagnostics[a]

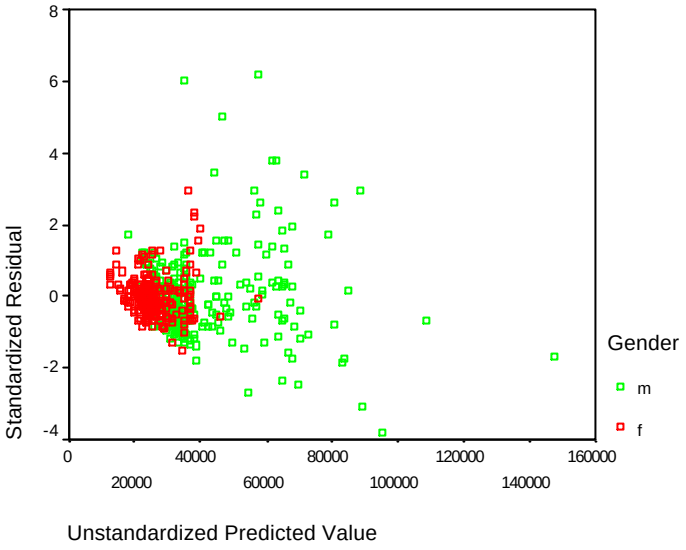
Case Number	Std. Residual	Current Salary
18	6,179	$103,750
103	3,404	$97,000
106	3,794	$91,250
160	-3,121	$66,000
205	-3,839	$66,750
218	6,006	$80,000
274	5,042	$83,750
449	3,497	$70,000
454	3,829	$90,625

a. Dependent Variable: Current Salary



Riportiamo ora il grafico che ha sull' asse delle ordinate i residui standardizzati e sull'asse delle ascisse i valori stimati, contrassegnando i residui in base alla variabile GENDER.

Osserviamo che i residui tendono ad avere un comportamento differente all'interno delle due categorie. Il modello costruito sembra più adeguato a spiegare il salario delle donne che non quello degli uomini.



3) Inseriamo ora tra le variabili esplicative due nuove variabili *dummy* che rappresentino le tre categorie lavorative descritte dalla variabile JOBCAT.

Definiamo le variabili

$$JB1 = \begin{cases} 1 & \text{se} \quad JOBCAT = 1 \\ 0 & \text{altrimenti} \end{cases}$$

$$JB2 = \begin{cases} 1 & \text{se} \quad JOBCAT = 2 \\ 0 & \text{altrimenti} \end{cases}$$

Quindi

per JOBCAT=1 (clerical) $\Rightarrow$ JB1=1, JB2=0

per JOBCAT=2 (custodial) $\Rightarrow$ JB1=0, JB2=1

per JOBCAT=3 (manager) $\Rightarrow$ JB1=0, JB2=0

Il modello di regressione diventa:

$$SALARY = b_0 + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + b_4 \cdot PREVEXP + g_1 \cdot GENDER + \\ + d_1 \cdot JB1 + d_2 \cdot JB2 + e$$

Valutando l'equazione riportata sopra, per i differenti valori delle variabili *dummy*, otteniamo sei differenti equazioni:

GENDER	JOBCAT	Equazione di regressione
male	clerical	$SALARY = (b_0 + g_1 + d_1) + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + + b_4 \cdot PREVEXP + e$
male	custodial	$SALARY = (b_0 + g_1 + d_2) + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + + b_4 \cdot PREVEXP + e$
male	manager	$SALARY = (b_0 + g_1) + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + + b_4 \cdot PREVEXP + e$
female	clerical	$SALARY = (b_0 + d_1) + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + + b_4 \cdot PREVEXP + e$
female	custodial	$SALARY = (b_0 + d_2) + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + + b_4 \cdot PREVEXP + e$
female	manager	$SALARY = b_0 + b_1 \cdot EDUC + b_2 \cdot SALBEGIN + b_3 \cdot JOBTIME + + b_4 \cdot PREVEXP + e$

Dal menu **Analyse**, selezioniamo **Regression** e quindi **Linear**. Selezioniamo come **Dependent variable** SALARY e come **Independent(s) variable** EDUC, JOBTIME, SLBEGIN, PREVEXP, GENDER, JB1, JB2. Dalla finestra **Linear Regression** selezioniamo

- **Statistics** => dalla finestra **Residuals** => **Casewise Diagnostics**, con **Outliers outside:** 3 **standard deviations.**
- **Save** => dalla finestra **Predicted values** => **Unstandardized**
  dalla finestra **Residuals** => **Standardized** (in questo modo vengono salvate nella **Window SPSS data editor**, contenente la matrice di dati le variabili PRE_1 (che contiene i valori stimati) e ZRE_1 (che contiene i residui standardizzati)

Otteniamo il seguente output:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,918[a]	,843	,841	$6,813.57

a. Predictors: (Constant), JB2, Months since Hire, Beginning Salary, Previous Experience (months), Gender, Educational Level (years), JB1

b. Dependent Variable: Current Salary

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,16E+11	7	1,66E+10	357,821	,000[a]
	Residual	2,16E+10	466	46424804		
	Total	1,38E+11	473			

a. Predictors: (Constant), JB2, Months since Hire, Beginning Salary, Previous Experience (months), Gender, Educational Level (years), JB1

b. Dependent Variable: Current Salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2799,937	3807,502		,735	,462
	Educational Level (years)	499,308	159,988	,084	3,121	,002
	Beginning Salary	1,341	,073	,618	18,397	,000
	Months since Hire	148,615	31,346	,088	4,741	,000
	Previous Experience (months)	-22,283	3,567	-,136	-6,248	,000
	Gender	1870,063	760,481	,055	2,459	,014
	JB1	-11255,5	1364,208	-,279	-8,251	,000
	JB2	-4518,651	2160,506	-,061	-2,091	,037

a. Dependent Variable: Current Salary

Casewise Diagnostics[a]

Case Number	Std. Residual	Current Salary	Predicted Value	Residual
18	6,067	$103,750	$62,415.08	$41,334.92
32	3,598	$110,625	$86,110.32	$24,514.68
103	3,504	$97,000	$73,125.96	$23,874.04
106	3,600	$91,250	$66,724.48	$24,525.52
205	-3,363	$66,750	$89,665.41	-$22,915.41
218	6,831	$80,000	$33,459.03	$46,540.97
274	4,468	$83,750	$53,306.30	$30,443.70
446	3,116	$100,000	$78,768.48	$21,231.52
454	3,712	$90,625	$65,332.68	$25,292.32

a. Dependent Variable: Current Salary

Osserviamo che

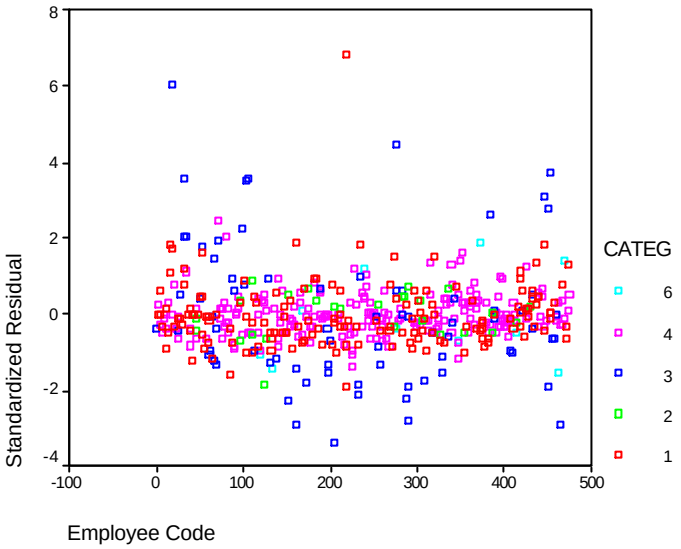
- il p-value del test che verifica l'ipotesi nulla che tutti i coefficienti siano nulli (tabella **ANOVA**) è prossimo a zero, quindi rifiutiamo l'ipotesi nulla
- La tabella **Coefficients** mostra che il p-value che verifica l'ipotesi nulla

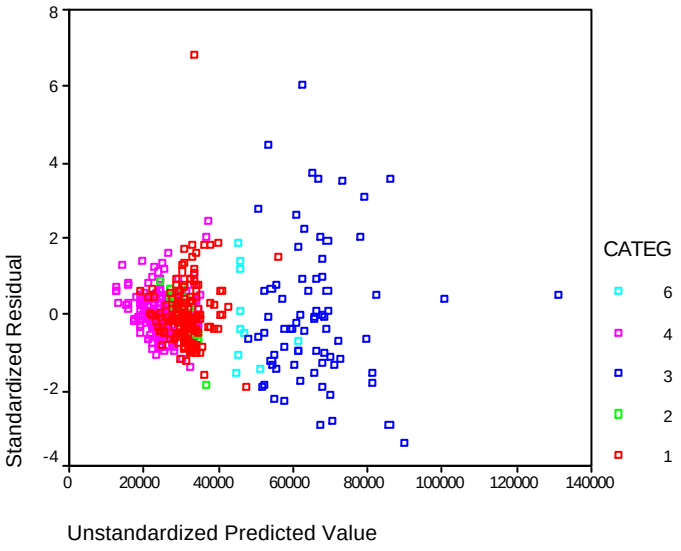
  $$H_0 : b_0 = 0 \text{ contro } H_1 : b_0 \neq 0$$

  è molto elevato, per cui la costante non è statisticamente significativa.

- Il coefficiente R^2_{ad} (pari a 0.841) è maggiore rispetto a quello dei modelli considerati precedentemente.

Sono riportati sotto i grafici che riportano i residui standardizzati (rispetto agli indice del data set e ai valori stimati rispettivamente) contrassegnati in base alle sei categorie descritte dal modello lineare (1=male clerical, 2=male custodial, 3=male manager, 4=female clerical, 5=female custodial, 6= female manager). Il secondo grafico, in particolare, mostra che i residui relativi alla categoria 3 tendono ad avere un comportamento diverso rispetto a quelli relativi alle altre categorie.





E' possibile che gli effetti di GENDER e JOBCAT sulla determinazione dello stipendio non siano additivi. Effetti non additivi di queste variabili possono essere valutati costruendo delle variabili aggiuntive che misurino gli effetti moltiplicativi o di interazione. Tuttavia l'inserimento

di tali variabili tra le variabile esplicative non ha condotto alla costruzione di un modello migliore. Perciò un modello lineare risulta inadeguato a descrivere il salario degli impiegati appartenenti alla terza categoria.