Istituto Universitario di Architettura di Venezia

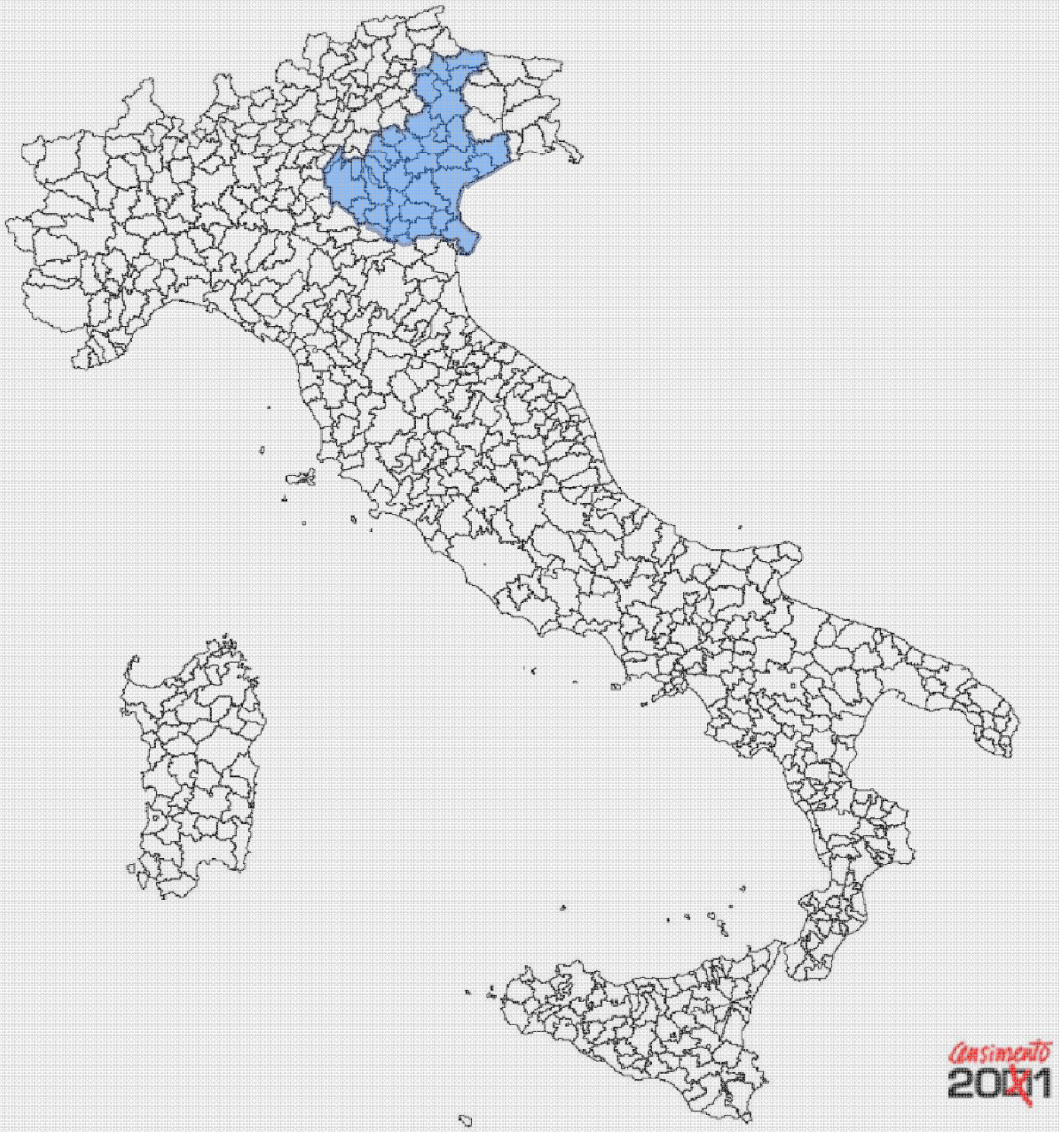
Facoltà di Pianificazione del Territorio

Corso di laurea in Scienze della Pianificazione Urbanistica e Territoriale

Esame di Metodi Quantitativi per l'Analisi Territoriale

Argomento dell'esercitazione: "I Sistemi Locali del Lavoro nella regione Veneto"

Docente: Carlo Grillenzoni



Studente: Chiara Carraretto

Matricola: 265163

Indice:

- Premessa ……..pag. 3
- Tabella dati ……pag. 4
- Analisi di regressione lineare multipla ……pag. 5
- Individuazione residui anomali ……pag. 7
- Ortogonalizzazione delle X ……..pag. 10
- Analisi di regressione lineare multipla tra abitazioni (Y) e i residui delle X …..pag. 15
- Individuazione dei residui anomali ……..……..pag. 16
- *Allegato:* Utilizzo di Statgraphics ……...pag. 19

Si parla di analisi territoriale quando il particolare fenomeno oggetto di studio viene rilevato in un collettivo di unità territoriali.

La mia analisi si riferirà alla distribuzione territoriale dei Sistemi Locali del Lavoro nella regione Veneto.

I Sistemi Locali del Lavoro

I Sistemi Locali del Lavoro (SLL) rappresentano i luoghi della vita quotidiana della popolazione che vi risiede e lavora. Si tratta di unità territoriali identificate da un insieme di comuni contigui legati tra loro dai flussi degli spostamenti quotidiani per motivi di lavoro che vengono rilevati in occasione dei censimenti della popolazione.
I Sistemi Locali del Lavoro sono uno strumento di analisi appropriato per indagare la struttura socio-economica dell'Italia secondo una prospettiva territoriale. La configurazione territoriale dei Sistemi locali del lavoro cambia nel tempo poiché riflette i mutamenti dell'organizzazione territoriale della società e dell'economia del Paese.
L'obiettivo di base è quindi la costruzione di una griglia sul territorio determinata dai movimenti dei soggetti per motivi di lavoro: l'ambito territoriale che ne discende rappresenta l'area geografica in cui maggiormente si addensano quei movimenti.
In questo modo si aggregano unità amministrative elementari (Comuni) individuati sul territorio dalle relazioni socio-economiche.
I criteri adottati per la definizione dei SLL sono l'*autocontenimento*, la *contiguità* e la *relazione spazio-tempo*.
Con il termine autocontenimento si intende un territorio dove si concentrano attività produttive e di servizi in quantità tali da offrire opportunità di lavoro e residenziali alla maggior parte della popolazione che vi è insediata. Il territorio si configura così in un sistema locale, ossia come un'entità socio-economica che fornisce occupazione, acquisti, relazioni e opportunità sociali.
Per contiguità si intende che i comuni contenuti in un SLL devono essere contigui, mentre la relazione spazio-tempo riguarda la distanza e il tempo di percorrenza tra la località di residenza e la località di lavoro, concetto strettamente legato alla presenza di servizi efficienti.
Il SLL presenta inoltre un forte grado di *mutevolezza*, poiché soggetto a cambiamenti che derivano dall'evolversi delle condizioni economiche, sociali, come dimostra ad esempio il forte calo del numero di Sistemi Locali del Lavoro dal censimento del 1981 (955) ai successivi censimenti del 1991 (784) e 2001 (686).

Le variabili che ho scelto sono la superficie totale (kmq), le abitazioni, la popolazione residente e le famiglie. Nella seguente tabella sono riportati i dati relativi ai 34 SLL del Veneto, scaricati dal censimento ISTAT del 2001.

Codice SLL	Sistema Locale del Lavoro 2001	Numero di comuni	Superficie (kmq)*	Popolazione residente	Famiglie	Abitazioni
132	BOVOLONE	12	398,07	62.712	22.208	23.755
133	GREZZANA	6	214,80	18.962	7.101	13.267
134	LEGNAGO	13	400,44	73.774	27.055	29.430
135	MALCESINE	3	169,63	8.401	3.580	9.367
136	SAN BONIFACIO	23	574,16	117.918	42.180	47.676
137	SAN GIOVANNI ILARIONE	4	88,57	15.083	5.226	6.519
138	VERONA	38	1315,25	540.753	215.633	241.184
139	ARZIGNANO	16	346,92	115.743	43.636	49.217
140	ASIAGO	5	352,53	13.922	5.462	20.828
141	BASSANO DEL GRAPPA	28	576,08	174.859	63.037	75.729
142	SCHIO	8	237,41	78.245	30.287	33.440
143	THIENE	25	450,65	104.094	38.894	46.259
144	VICENZA	30	585,93	278.791	105.816	116.207
145	AGORDO	15	561,40	19.959	8.693	18.988
146	AURONZO DI CADORE	7	564,12	13.143	5.540	12.307
147	BELLUNO	20	1006,38	93.131	38.425	49.460
148	CORTINA D'AMPEZZO	4	389,99	9.513	3.977	10.713
149	FELTRE	14	691,65	58.783	23.798	31.824
150	PIEVE DI CADORE	10	435,54	16.008	6.873	12.365
151	CASTELFRANCO VENETO	24	588,62	207.618	70.785	76.480
152	CONEGLIANO	23	596,07	176.344	67.094	74.643
153	MONTEBELLUNA	17	441,02	119.395	42.435	48.744
154	PIEVE DI SOLIGO	9	191,86	41.816	15.652	18.595
155	TREVISO	25	697,81	298.438	111.942	120.390
156	PORTOGRUARO	19	750,81	115.074	41.564	56.733
157	SAN DONA' DI PIAVE	10	437,59	107.944	39.485	58.940
158	VENEZIA	21	1208,42	600.549	236.247	258.905
159	ESTE	28	554,89	117.712	41.814	46.351
160	MONTAGNANA	21	401,21	64.788	22.733	25.702
161	PADOVA	46	974,94	580.466	219.853	236.159
162	ADRIA	5	362,01	42.859	15.968	18.098
163	BADIA POLESINE	17	362,52	51.005	18.518	20.890
164	ROVIGO	16	416,56	87.266	33.812	36.802
165	PORTO VIRO	7	651,97	50.738	18.745	26.798

È interessante comprendere quale sia la relazione tra queste variabili, ossia se il numero di abitazioni per Sistema Locale del Lavoro sia spiegato dalla quantità di superficie territoriale di ciascun SLL, o dalla quantità di popolazione residente e quindi dalle famiglie presenti per SLL.

Analisi di regressione lineare multipla:

Il metodo che utilizzerò per svolgere questa analisi è la regressione lineare multipla, incrociando tra loro le variabili..

Il modello di regressione serve a stabilire la relazione esistente tra una variabile dipendente, e una o più variabili indipendenti o esplicative. La scelta dell'una o dell'altra variabile come indipendente non è arbitraria ma legata alla natura del fenomeno: si sceglie come variabile indipendente la variabile che sia logicamente antecedente rispetto all'altra.

In un modello di regressione, le variabili indipendenti (dette anche regressori) spiegano, prevedono, simulano, controllano la variabile dipendente e permettono di capire se esiste un trend (positivo o negativo) che consenta di fare previsioni.

La variabile dipendente nell'equazione di regressione è modellata come una funzione delle variabili indipendenti più un termine di errore. Quest'ultimo è una variabile casuale e rappresenta una variazione non controllabile e imprevedibile nella variabile dipendente. I parametri sono stimati in modo da descrivere meglio i dati.

Questo strumento statistico può essere utilizzato solo nel caso di variabili quantitative o qualitative ordinabili. Per questa ragione, tramite questo strumento, andrò a comprendere quale tra le variabili "popolazione residente", "famiglie" e "superficie territoriale" spiega maggiormente la quantità di abitazioni per SLL.
Per fare le mie valutazioni procederò con il valutare la statistica T e l' R^2 delle diverse regressioni multiple.

La statistica T serve a comprendere se il modello funziona: più il valore è alto e più il modello funziona. Se la statistica T cade tra +2 e -2, è paragonabile a 0 quindi il modello non funziona.

L'indice R^2 è chiamato coefficiente di determinazione lineare ed è l'indice di bontà del modello di regressione. Il suo valore è compreso tra 0 e 1. Quando la varianza residua è molto più piccola rispetto alla varianza totale l'indice R2 è vicino ad 1 e il modello di regressione funziona bene. Più il suo valore è alto e più il modello è significativo.

Multiple Regression Analysis

Dependent variable: abitazioni

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	1647,72	1850,68	0,890336	0,3804
famiglie	1,38146	0,2638	5,23675	0,0000
pop_reside	-0,135612	0,100646	-1,34741	0,1879
sup_kmq	10,1931	4,27699	2,38324	0,0237

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	1,40627E11	3	4,68756E10	3078,00	0,0000
Residual	4,56877E8	30	1,52292E7		
Total (Corr.)	1,41084E11	33			

R-squared = 99,6762 percent
R-squared (adjusted for d.f.) = 99,6438 percent
Standard Error of Est. = 3902,46
Mean absolute error = 2545,11
Durbin-Watson statistic = 1,52561

The StatAdvisor

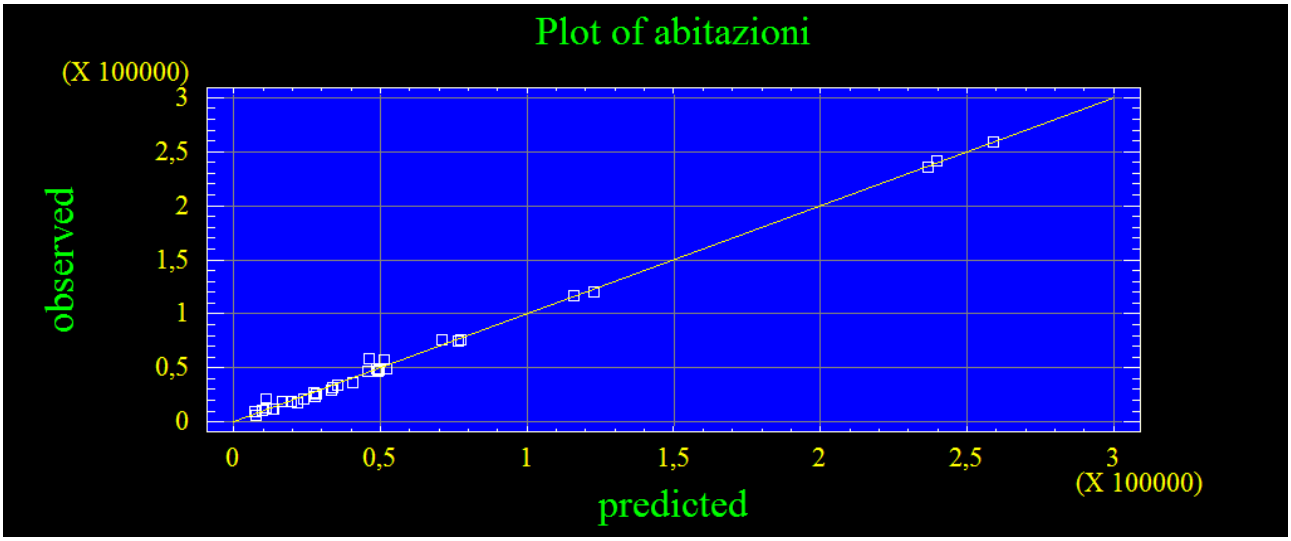
The output shows the results of fitting a multiple linear regression model to describe the relationship between abitazioni and 3 independent variables. The equation of the fitted model is

abitazioni = 1647,72 + 1,38146*famiglie - 0,135612*pop_reside + 10,1931*sup_kmq

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between the variables at the 99% confidence level.

The R-Squared statistic indicates that the model as fitted explains 99,6762% of the variability in abitazioni. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 99,6438%. The standard error of the estimate shows the standard deviation of the residuals to be 3902,46. This value can be used to construct prediction limits for new observations by selecting the Reports option from the text menu. The mean absolute error (MAE) of 2545,11 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the DW value is greater than 1.4, there is probably not any serious autocorrelation in the residuals.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0,1879, belonging to pop_reside. Since the P-value is greater or equal to 0.10, that term is not statistically significant at the 90% or higher confidence level. Consequently, you should consider removing pop_reside from the model.



Il modello di regressione multipla:

“Abitazioni (Y) = 1647,72 + 1,38146*famiglie – 0,135612*pop_reside + 10,1931*sup_kmq” è molto significativo, come si può notare dal valore di R^2 (99,7 %).

Dai valori riportati nella tabella precedente si può dedurre che le famiglie e la superficie totale hanno un forte legame dipendente con le abitazioni, mentre la popolazione residente non influenza il numero di abitazioni presenti per Sistema Locale del Lavoro.

Individuazione dei residui anomali:

Il modello di regressione si può considerare come una retta in un piano di n-dimensioni (rappresentate dal numero di variabili) ed è possibile calcolare di quanto le unità statistiche si scostano dalla retta (osservazioni anomale).

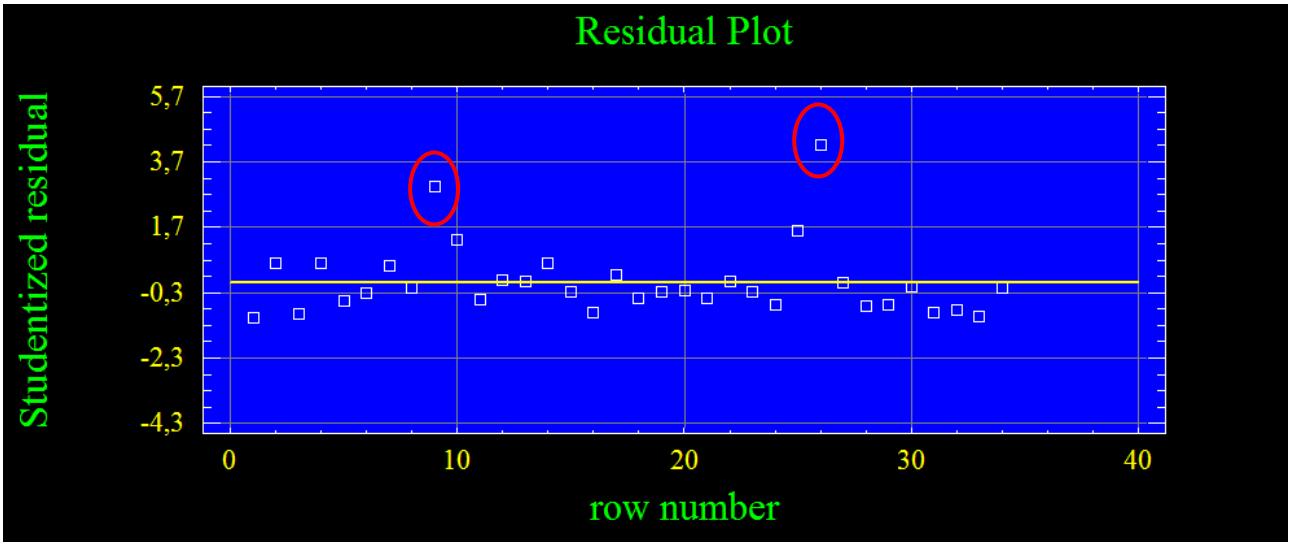
Un’osservazione è anomala non perché sta fuori dalla nuvola di punti, ma perché il livello della sua X non è associato alla sua Y come negli altri dati.

I residui sono quindi le distanze (positive o negative) dalla retta e si distribuiscono come una Normale (con media = 0 e varianza = $\sigma 2e$).

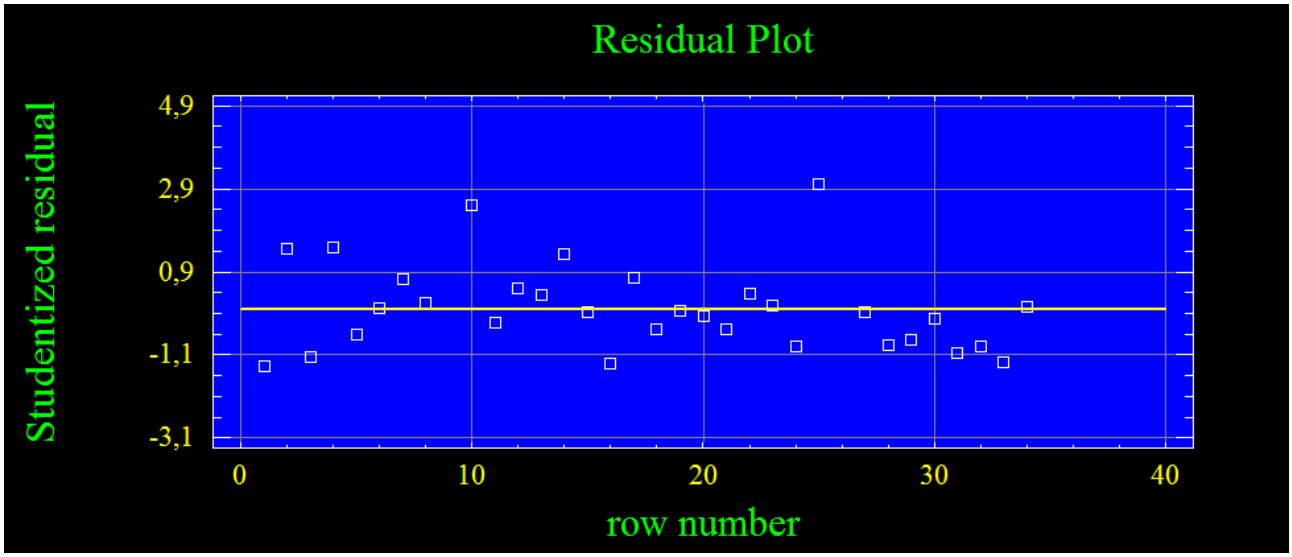
Dal grafico seguente si possono individuare i valori anomali nel campione di dati elaborati per costruire il modello di regressione.

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
9	20828,0	10898,6	9929,38	2,93
26	58940,0	46016,5	12923,5	4,21



Di seguito è riportato il grafico depurato dai residui anomali n.9 e n.26.



Togliendo le unità statistiche anomale Statgraphics ricalcola il modello, generando di conseguenza altri residui anomali.

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
10	75729,0	70301,8	5427,17	2,51
25	56733,0	50665,1	6067,91	3,01

Lascio inserite le unità statistiche n.10 e n.25 in quanto risultano di non molto fuori dal range.

Multiple Regression Analysis

Dependent variable: abitazioni

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	247,023	1191,66	0,207292	0,8373
famiglie	1,36274	0,16798	8,11248	0,0000
pop_reside	-0,127921	0,0641253	-1,99486	0,0559
sup_kmq	11,3179	2,71189	4,17345	0,0003

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	1,39488E11	3	4,64958E10	7623,18	0,0000
Residual	1,7078E8	28	6,09928E6		
Total (Corr.)	1,39658E11	31			

R-squared = 99,8777 percent
R-squared (adjusted for d.f.) = 99,8646 percent
Standard Error of Est. = 2469,67
Mean absolute error = 1768,15
Durbin-Watson statistic = 2,37557

The StatAdvisor

The output shows the results of fitting a multiple linear regression model to describe the relationship between abitazioni and 3 independent variables. The equation of the fitted model is

abitazioni = 247,023 + 1,36274*famiglie - 0,127921*pop_reside + 11,3179*sup_kmq

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between the variables at the 99% confidence level.

The R-Squared statistic indicates that the model as fitted explains 99,8777% of the variability in abitazioni. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 99,8646%. The standard error of the estimate shows the standard deviation of the residuals to be 2469,67. This value can be used to construct prediction limits for new observations by selecting the Reports option from the text menu. The mean absolute error (MAE) of 1768,15 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the DW value is greater than 1.4, there is probably not any serious autocorrelation in the residuals.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0,0559, belonging to pop_reside. Since the P-value is less than 0.10, that term is statistically significant at the 90% confidence level. Depending on the confidence level at which you wish to work, you may or may not decide to remove pop_reside from the model.

Togliendo i residui anomali aumenta l'indice R^2 (99,9 %). e quindi la bontà del modello. La variabile "popolazione residente" spiega il numero di abitazioni, perciò il segno della statistica T dev'essere positivo. In questo caso però risulta essere negativo. Questo problema viene risolto tramite l'ortogonalizzazione dei regressori.

Ortogonalizzazione delle variabili indipendenti (X)

Come ho detto l'analisi della regressione lineare multipla risponde all'obiettivo di studiare la dipendenza di una variabile quantitativa Y da un insieme di n variabili esplicative quantitative X1, …, Xn, dette regressori, mediante un modello lineare.

In molte situazioni, le variabili esplicative possono essere tra loro molto correlate e in questo caso ci troviamo in una situazione di multicollinearità.
L'effetto principale dovuto alla multicollinearità è quello di aumentare considerevolmente la varianza degli stimatori dei minimi quadrati dei coefficienti di regressione. Questo aumento ha degli effetti negativi sull'inferenza dei coefficienti di regressione. In particolare porta all'aumento dell'ampiezza dell'intervello di confidenza. Inoltre l'aumento dell'errore standard fa diminuire il valore assoluto della T_statistica portando più facilmente ad accettare l'ipotesi nulla, anche se questa non è vera.
La presenza di elevata multicollinearità comporta il cambiamento nei valori delle stime dei coefficienti di regressione in conseguenza a lievi modificazioni dei valori osservati, a eliminazione o aggiunta di qualche variabile esplicativa, all'aggiunta di nuove osservazioni. Tuttavia, la multicollinearità, non altera la bontà di adattamento del modello e la sua capacità previsiva rispetto alla variabile risposta.

Trasformo il modello attraverso l'ortogonalizzazione dei residui (cioè sono stocasticamente incorrelati). In questo modo l'impulso unitario di una variabile su di un'altra può essere analizzato senza l'interferenza causata dai legami istantanei tra le variabili.

Procederò inizialmente analizzando la regressione multipla tra i regressori utilizzando ciascuna variabile X come variabile dipendente nelle tre regressioni .

- famiglie X1 → pop.residente X2, sup.kmq X3
- pop.residente X2 → famiglie X1, sup.kmq X3
- sup.kmq X3 → famiglie X1, po.residente X2

Dalle tre regressioni ricavo i rispettivi residui. Infine la regressione multipla verrà eseguita tra le abitazioni ed i residui delle tre precedenti regressioni (X1_Rs, X2_Rs, X3_Rs).

Regressione tra la variabile famiglie (X1) e le variabili: pop. residente (X2) e sup. kmq (X3)

```
Multiple Regression Analysis
----------------------------------------------------------------------------
Dependent variable: famiglie
----------------------------------------------------------------------------
                                     Standard          T
Parameter                Estimate        Error      Statistic        P-Value
----------------------------------------------------------------------------
CONSTANT                -3171,06       1123,95       -2,82136         0,0083
pop_reside              0,380585    0,00480311        79,2373         0,0000
sup_kmq                  5,47118       2,74113        1,99596         0,0548
----------------------------------------------------------------------------

                          Analysis of Variance
----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
----------------------------------------------------------------------------
Model                 1,22594E11      2   6,12969E10     8683,05       0,0000
Residual               2,18841E8     31    7,05937E6
----------------------------------------------------------------------------
Total (Corr.)         1,22813E11     33

R-squared = 99,8218 percent
R-squared (adjusted for d.f.) = 99,8103 percent
Standard Error of Est. = 2656,95
Mean absolute error = 1888,63
Durbin-Watson statistic = 2,25281


The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear
regression model to describe the relationship between famiglie and 2
independent variables.  The equation of the fitted model is

famiglie = -3171,06 + 0,380585*pop_reside + 5,47118*sup_kmq

Since the P-value in the ANOVA table is less than 0.01, there is a
statistically significant relationship between the variables at the
99% confidence level.

     The R-Squared statistic indicates that the model as fitted
explains 99,8218% of the variability in famiglie.  The adjusted
R-squared statistic, which is more suitable for comparing models with
different numbers of independent variables, is 99,8103%.  The standard
error of the estimate shows the standard deviation of the residuals to
be 2656,95.  This value can be used to construct prediction limits for
new observations by selecting the Reports option from the text menu.
The mean absolute error (MAE) of 1888,63 is the average value of the
residuals.  The Durbin-Watson (DW) statistic tests the residuals to
determine if there is any significant correlation based on the order
in which they occur in your data file.  Since the DW value is greater
than 1.4, there is probably not any serious autocorrelation in the
residuals.

   In determining whether the model can be simplified, notice that the
highest P-value on the independent variables is 0,0548, belonging to
sup_kmq.  Since the P-value is less than 0.10, that term is
statistically significant at the 90% confidence level.  Depending on
the confidence level at which you wish to work, you may or may not
decide to remove sup_kmq from the model.
```

C’è un forte legame tra le famiglie e la popolazione residente mentre il legame è minimo tra le famiglie e la superficie territoriale.

Regressione tra la variabile pop. residente (X2) e le variabili: famiglie (X1) e sup. kmq (X3)

```
Multiple Regression Analysis
----------------------------------------------------------------------------
Dependent variable: pop_reside
----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
----------------------------------------------------------------------------
CONSTANT                  7765,7        2993,6         2,5941          0,0144
famiglie                2,61462     0,0329974        79,2373          0,0000
sup_kmq                -12,0904        7,31697       -1,65238          0,1086
----------------------------------------------------------------------------

                          Analysis of Variance
----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square     F-Ratio       P-Value
----------------------------------------------------------------------------
Model                  8,11989E11      2   4,05994E11     8371,38        0,0000
Residual                1,50344E9     31    4,84979E7
----------------------------------------------------------------------------
Total (Corr.)          8,13492E11     33

R-squared = 99,8152 percent
R-squared (adjusted for d.f.) = 99,8033 percent
Standard Error of Est. = 6964,04
Mean absolute error = 5015,72
Durbin-Watson statistic = 2,23483

The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear
regression model to describe the relationship between pop_reside and 2
independent variables.  The equation of the fitted model is

pop_reside = 7765,7 + 2,61462*famiglie - 12,0904*sup_kmq

Since the P-value in the ANOVA table is less than 0.01, there is a
statistically significant relationship between the variables at the
99% confidence level.

     The R-Squared statistic indicates that the model as fitted
explains 99,8152% of the variability in pop_reside.  The adjusted
R-squared statistic, which is more suitable for comparing models with
different numbers of independent variables, is 99,8033%.  The standard
error of the estimate shows the standard deviation of the residuals to
be 6964,04.  This value can be used to construct prediction limits for
new observations by selecting the Reports option from the text menu.
The mean absolute error (MAE) of 5015,72 is the average value of the
residuals.  The Durbin-Watson (DW) statistic tests the residuals to
determine if there is any significant correlation based on the order
in which they occur in your data file.  Since the DW value is greater
than 1.4, there is probably not any serious autocorrelation in the
residuals.

   In determining whether the model can be simplified, notice that the
highest P-value on the independent variables is 0,1086, belonging to
sup_kmq.  Since the P-value is greater or equal to 0.10, that term is
not statistically significant at the 90% or higher confidence level.
Consequently, you should consider removing sup_kmq from the model.
```

Tra la popolazione residente e le famiglie esiste una forte dipendenza, mentre la superficie territoriale non influenza la popolazione residente.

Regressione tra la variabile sup. kmq (X3) e le variabili: famiglie (X1) e pop.residente (X2)

```
Multiple Regression Analysis
----------------------------------------------------------------------------
Dependent variable: sup_kmq
----------------------------------------------------------------------------
                                     Standard          T
Parameter                Estimate        Error      Statistic        P-Value
----------------------------------------------------------------------------
CONSTANT                  373,575      39,2165        9,52598         0,0000
famiglie                0,0208139    0,0104281        1,99596         0,0548
pop_reside             -0,0066951    0,0040518       -1,65238         0,1086
----------------------------------------------------------------------------

                          Analysis of Variance
----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
----------------------------------------------------------------------------
Model                   1,66516E6      2     832582,0      31,00        0,0000
Residual                 832533,0     31      26855,9
----------------------------------------------------------------------------
Total (Corr.)           2,4977E6     33

R-squared = 66,668 percent
R-squared (adjusted for d.f.) = 64,5175 percent
Standard Error of Est. = 163,878
Mean absolute error = 117,52
Durbin-Watson statistic = 1,92832


The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear
regression model to describe the relationship between sup_kmq and 2
independent variables.  The equation of the fitted model is

sup_kmq = 373,575 + 0,0208139*famiglie - 0,0066951*pop_reside

Since the P-value in the ANOVA table is less than 0.01, there is a
statistically significant relationship between the variables at the
99% confidence level.

     The R-Squared statistic indicates that the model as fitted
explains 66,668% of the variability in sup_kmq.  The adjusted
R-squared statistic, which is more suitable for comparing models with
different numbers of independent variables, is 64,5175%.  The standard
error of the estimate shows the standard deviation of the residuals to
be 163,878.  This value can be used to construct prediction limits for
new observations by selecting the Reports option from the text menu.
The mean absolute error (MAE) of 117,52 is the average value of the
residuals.  The Durbin-Watson (DW) statistic tests the residuals to
determine if there is any significant correlation based on the order
in which they occur in your data file.  Since the DW value is greater
than 1.4, there is probably not any serious autocorrelation in the
residuals.

   In determining whether the model can be simplified, notice that the
highest P-value on the independent variables is 0,1086, belonging to
pop_reside.  Since the P-value is greater or equal to 0.10, that term
is not statistically significant at the 90% or higher confidence
level.  Consequently, you should consider removing pop_reside from the
model.
```

La superficie totale non dipende dalle famiglie né dalla popolazione residente.

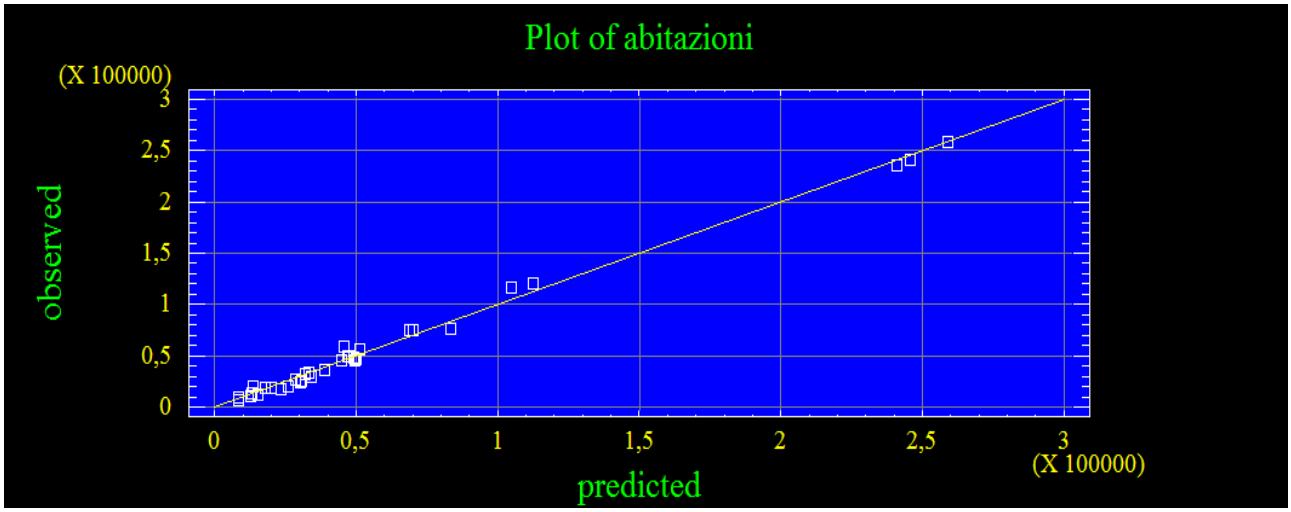
Successivamente ad ognuna delle tre regressioni appena svolte, ho salvato i residui che mi serviranno per ricalcolare il modello di regressione lineare multipla tra la variabile Y=abitazioni e i residui di X1, X2, X3.

	x1_Rs	x2_Rs	x3_Rs
1	-0,251538	0,243996	-0,109455
2	0,728928	-0,706002	-1,12903
3	-0,015882	0,0159931	-0,258824
4	1,03495	-1,00367	-1,42006
5	-1,02247	0,995563	0,696644
6	0,872698	-0,80872	-1,90727
7	2,87384	-2,81589	0,593078
8	0,328091	-0,279807	-0,99077
9	0,535624	-0,561478	-0,256492
10	-1,35379	1,36653	0,38433
11	0,933843	-0,874058	-1,54376
12	-0,00646048	0,0120175	-0,216868
13	-0,12442	0,212091	-0,773085
14	0,463959	-0,554522	0,883448
15	0,241729	-0,338799	1,02745
16	0,292628	-0,507317	3,35462
17	0,531602	-0,572716	-0,0165194
18	0,319549	-0,426367	1,37843
19	0,599641	-0,650741	0,161664
20	-3,80894	3,84202	0,982561
21	-0,0414201	0,0516173	0,0405411
22	-0,857149	0,874736	-0,101643
23	0,726852	-0,679418	-1,44251
24	-0,888008	0,951969	-0,0481665
25	-1,26435	1,17415	1,82818
26	-0,30935	0,321061	-0,214866
27	2,00174	-1,85298	-0,471179
28	-1,09368	1,07268	0,616474
29	-0,35845	0,351107	-0,0721014
30	-1,49039	1,7014	-0,661721
31	0,320794	-0,329505	-0,349895
32	0,11112	-0,114719	-0,337174
33	0,56516	-0,559198	-0,469999
34	-0,375164	0,274582	1,4425

Regressione tra Y=abitazioni e i residui di x1 (fam.), x2 (pop.) e x3 (sup.)

```
Multiple Regression Analysis
----------------------------------------------------------------------------
Dependent variable: abitazioni
----------------------------------------------------------------------------
                                     Standard          T
Parameter                Estimate        Error      Statistic        P-Value
----------------------------------------------------------------------------
CONSTANT                  54689,5          902,0        60,6313         0,0000
x1_R                    1,32252E6        18589,9        71,1414         0,0000
x2_R                    1,29851E6        18326,8        70,8532         0,0000
x3_R                      78728,9        1395,37        56,4214         0,0000
----------------------------------------------------------------------------

                          Analysis of Variance
----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
----------------------------------------------------------------------------
Model                   1,40256E11      3   4,67521E10    1695,12       0,0000
Residual                 8,27414E8     30    2,75805E7
----------------------------------------------------------------------------
Total (Corr.)           1,41084E11     33

R-squared = 99,4135 percent
R-squared (adjusted for d.f.) = 99,3549 percent
Standard Error of Est. = 5251,71
Mean absolute error = 3839,31
Durbin-Watson statistic = 1,25821

The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear
regression model to describe the relationship between abitazioni and 3
independent variables.  The equation of the fitted model is

abitazioni = 54689,5 + 1,32252E6*x1_R + 1,29851E6*x2_R + 78728,9*x3_R

Since the P-value in the ANOVA table is less than 0.01, there is a
statistically significant relationship between the variables at the
99% confidence level.

     The R-Squared statistic indicates that the model as fitted
explains 99,4135% of the variability in abitazioni.  The adjusted
R-squared statistic, which is more suitable for comparing models with
different numbers of independent variables, is 99,3549%.  The standard
error of the estimate shows the standard deviation of the residuals to
be 5251,71.  This value can be used to construct prediction limits for
new observations by selecting the Reports option from the text menu.
The mean absolute error (MAE) of 3839,31 is the average value of the
residuals.  The Durbin-Watson (DW) statistic tests the residuals to
determine if there is any significant correlation based on the order
in which they occur in your data file.  Since the DW value is less
than 1.4, there may be some indication of serial correlation.  Plot
the residuals versus row order to see if there is any pattern which
can be seen.

   In determining whether the model can be simplified, notice that the
highest P-value on the independent variables is 0,0000, belonging to
x1_R.  Since the P-value is less than 0.01, the highest order term is
statistically significant at the 99% confidence level.  Consequently,
you probably don't want to remove any variables from the model.
```

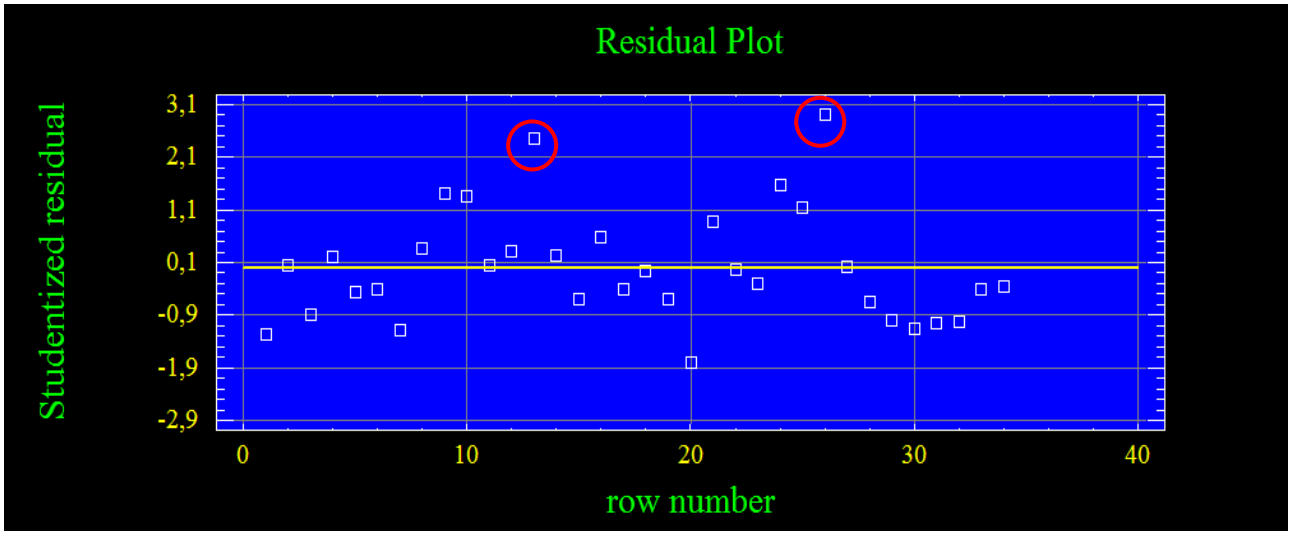
Il modello di regressione multipla calcolato tra le abitazioni ed i residui delle variabili X1,X2 e X3 è comunque molto buono, nonostante l'indice R^2 (99,4 %) è diminuito rispetto al modello calcolato tra le variabili senza l'ortogonalizzazione dei regressori. Esiste una relazione forte di dipendenza tra Y e le variabili X1, X2, X3.



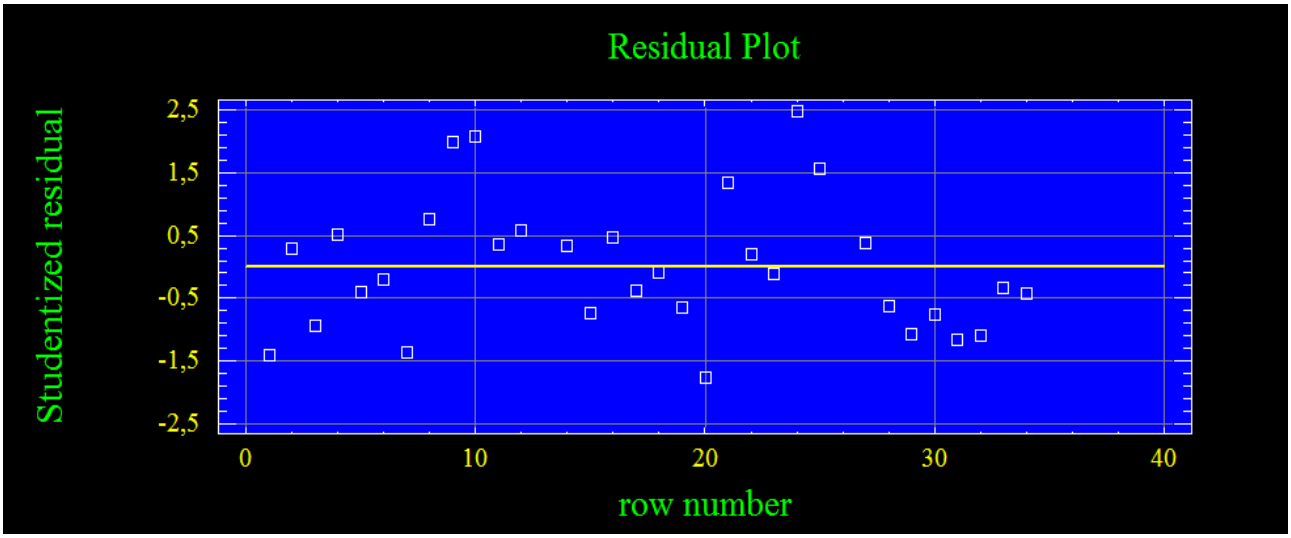
Individuazione dei residui anomali:

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
13	116207,0	104681,0	11526,1	2,46
26	58940,0	45555,1	13384,9	2,90



Di seguito è riportato il grafico depurato dai residui anomali (n.13 e n.26).



Togliendo le unità statistiche anomale Statgraphics ricalcola il modello, generando di conseguenza altri residui anomali.

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
10	75729,0	67923,7	7805,26	2,06
24	120390,0	111342,0	9048,04	2,46

Lascio inserite le unità statistiche n.10 e n.24 in quanto non escono di molto rispetto all'intervallo di accettabilità.

Multiple Regression Analysis

Dependent variable: abitazioni

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	53890,5	738,58	72,9649	0,0000
x1_Rs	1,31795E6	14900,2	88,4518	0,0000
x2_Rs	1,29374E6	14694,3	88,0439	0,0000
x3_Rs	78900,2	1109,61	71,1064	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	1,37102E11	3	4,57005E10	2628,82	0,0000
Residual	4,86764E8	28	1,73844E7		
Total (Corr.)	1,37588E11	31			

R-squared = 99,6462 percent
R-squared (adjusted for d.f.) = 99,6083 percent
Standard Error of Est. = 4169,46
Mean absolute error = 3180,55
Durbin-Watson statistic = 1,24082

Number of excluded rows: 2

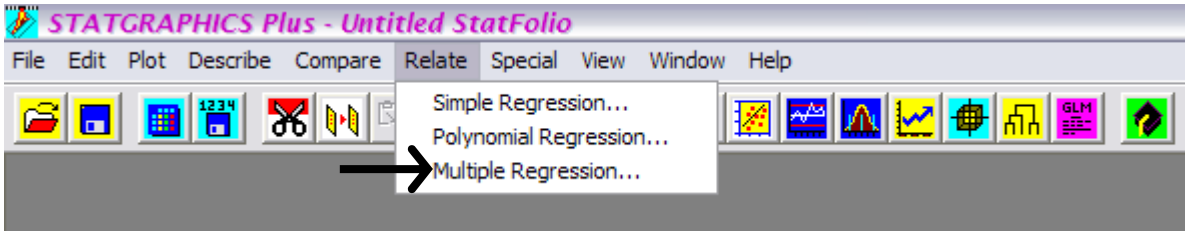
```
The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear
regression model to describe the relationship between abitazioni and 3
independent variables.  The equation of the fitted model is

abitazioni = 53890,5 + 1,31795E6*x1_Rs + 1,29374E6*x2_Rs +
78900,2*x3_Rs

Since the P-value in the ANOVA table is less than 0.01, there is a
statistically significant relationship between the variables at the
99% confidence level.

     The R-Squared statistic indicates that the model as fitted
explains 99,6462% of the variability in abitazioni.  The adjusted
R-squared statistic, which is more suitable for comparing models with
different numbers of independent variables, is 99,6083%.  The standard
error of the estimate shows the standard deviation of the residuals to
be 4169,46.  This value can be used to construct prediction limits for
new observations by selecting the Reports option from the text menu.
The mean absolute error (MAE) of 3180,55 is the average value of the
residuals.  The Durbin-Watson (DW) statistic tests the residuals to
determine if there is any significant correlation based on the order
in which they occur in your data file.  Since the DW value is less
than 1.4, there may be some indication of serial correlation.  Plot
the residuals versus row order to see if there is any pattern which
can be seen.

   In determining whether the model can be simplified, notice that the
highest P-value on the independent variables is 0,0000, belonging to
x1_Rs.  Since the P-value is less than 0.01, the highest order term is
statistically significant at the 99% confidence level.  Consequently,
you probably don't want to remove any variables from the model.
```

Il modello è buono (R^2=99,6% →vicino a 1).

Il numero di abitazioni per Sistema Locale del Lavoro nella regione Veneto dipende fortemente dal numero di famiglie, dalla popolazione residente e dalla superficie totale.

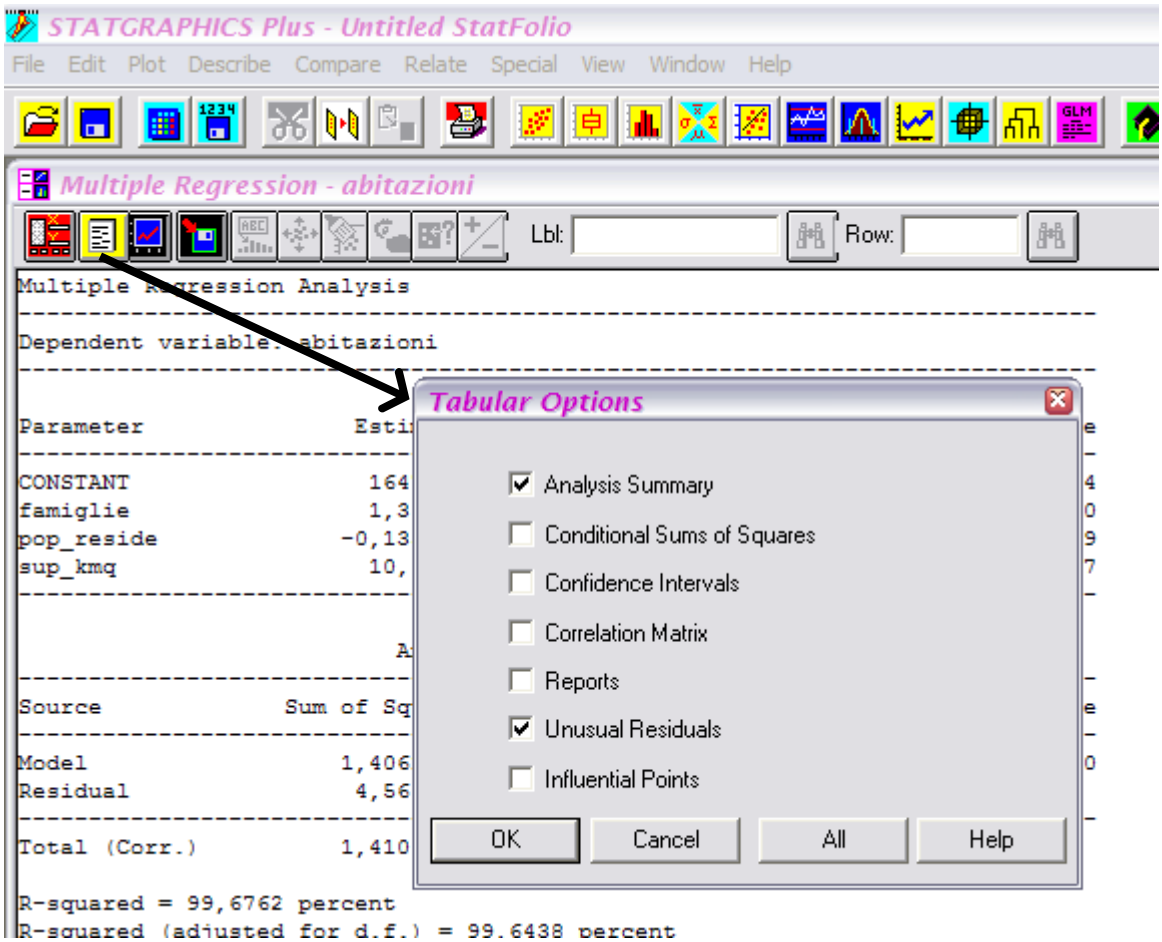
Allegato: Utilizzo di Statgraphics

Per analizzare la regressione lineare multipla cliccare **Relate → Multiple Regression**



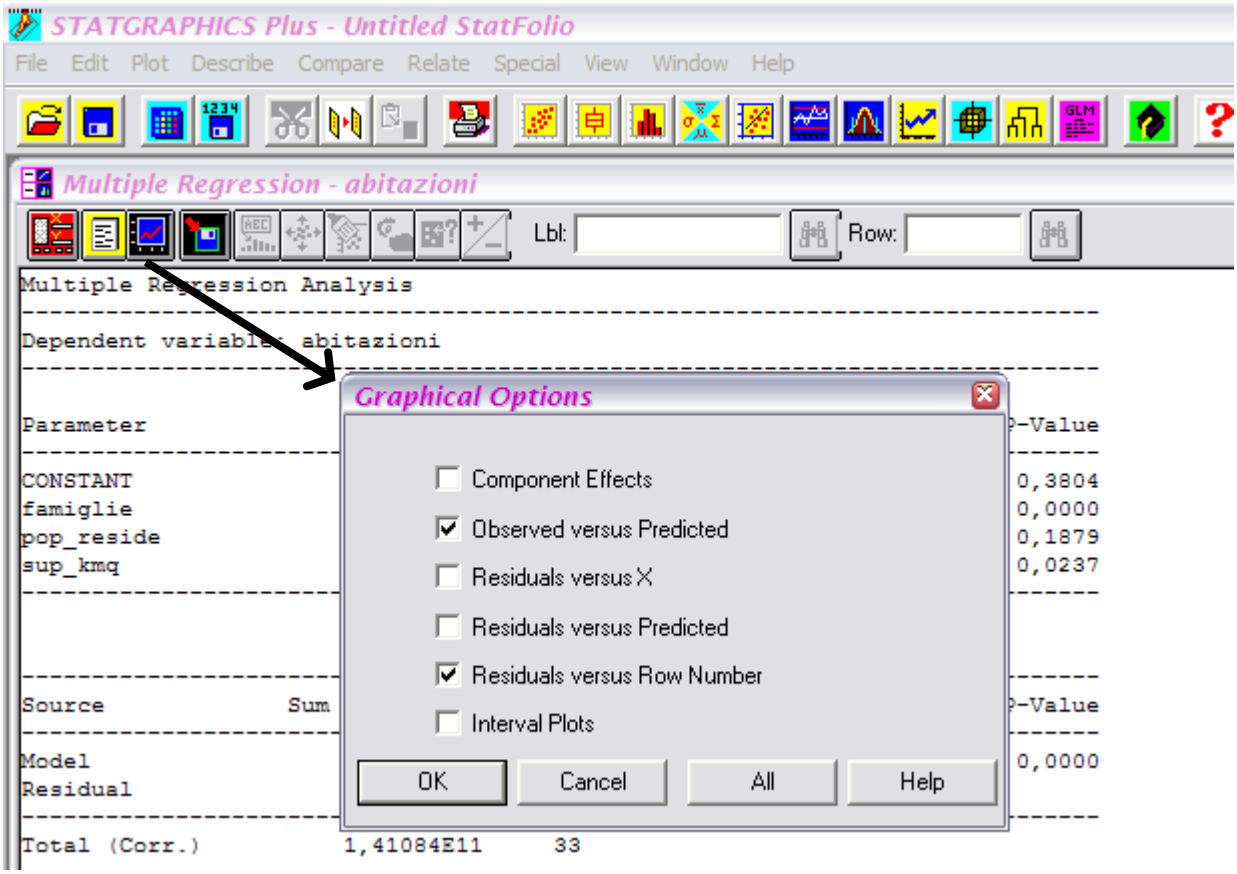
Una volta scelta le variabili indipendenti (X) e le rispettiva variabile dipendente (Y) si aprirà una schermata dove possono essere prodotti grafici e tabelle per approfondire l'analisi.

In particolare cliccando il tasto  è possibile creare tabelle d'analisi.



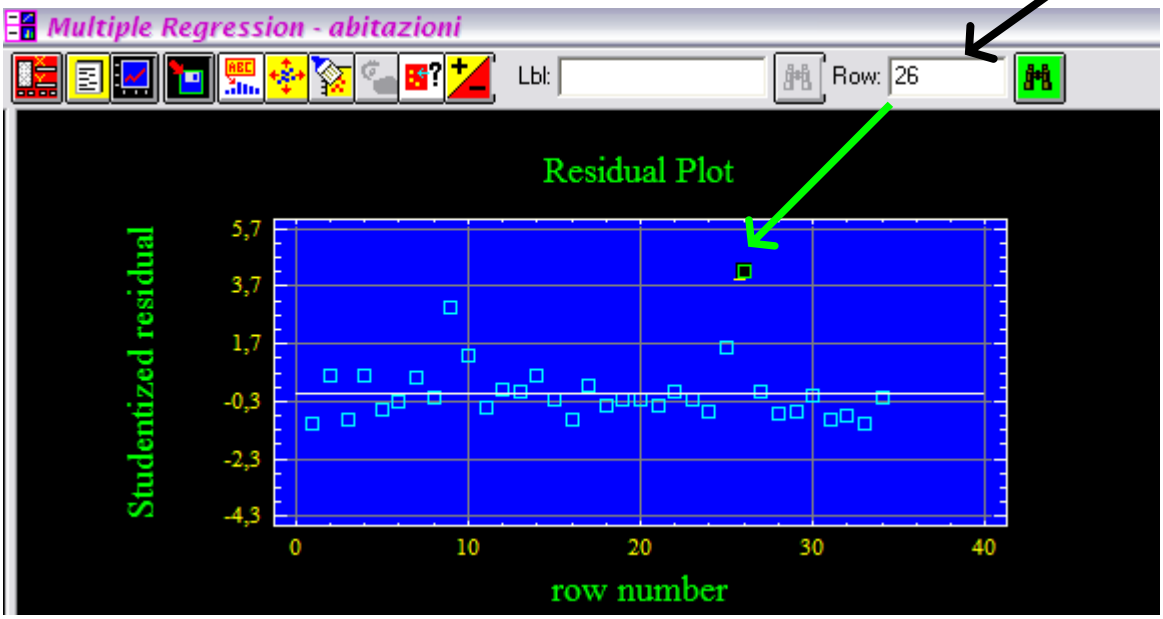
Spuntando **Analysis Summary** si apre la tabella generale di analisi della regressione multipla mentre **Unusual Residuals** crea la tabella di analisi dei residui standardizzati.

Con il tasto  si possono creare i grafici.

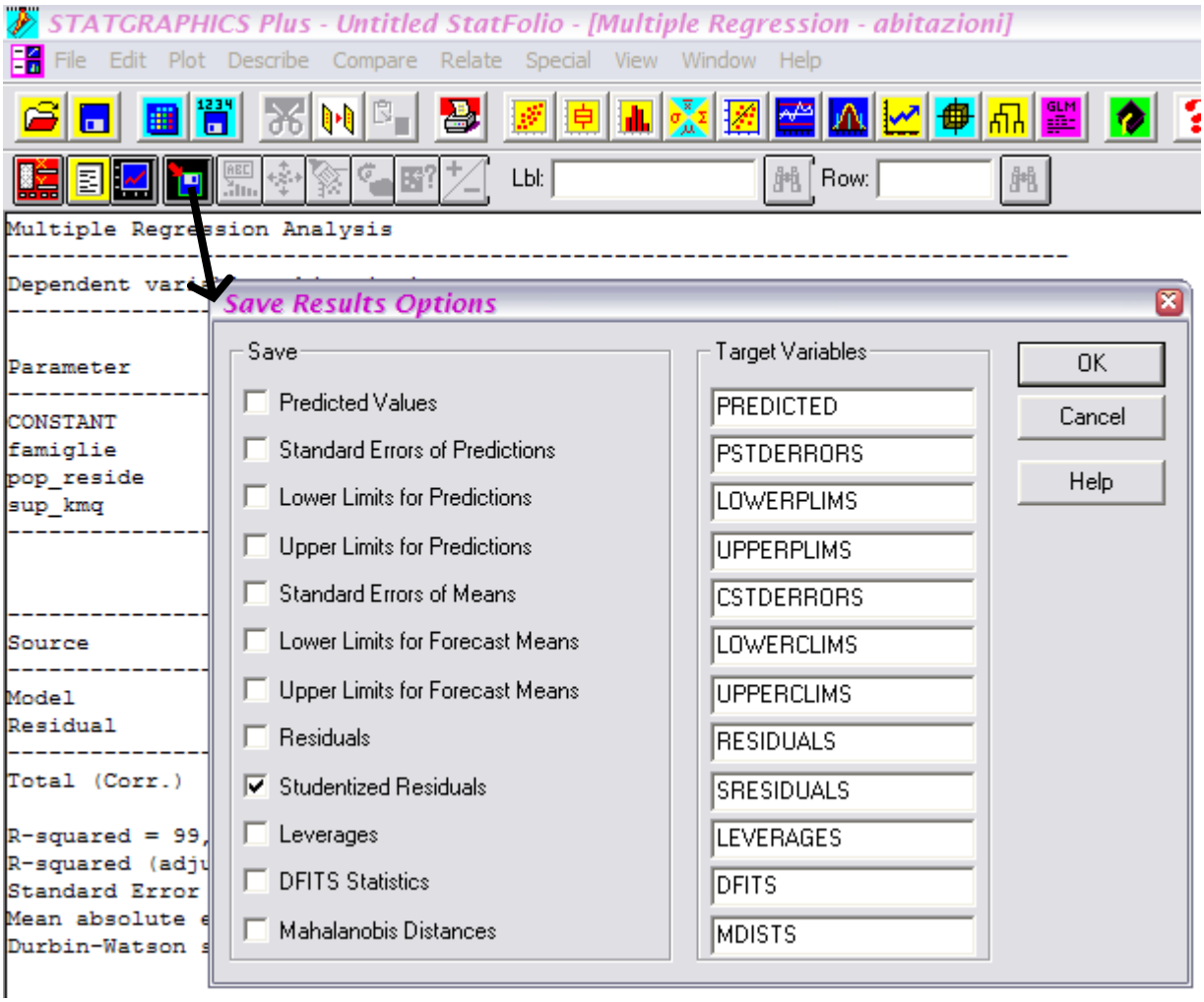


Cliccando **Observed versus Predicted** si aprirà il grafico del trend di regressione, mentre cliccando **Residuals versus Row Number** si aprirà il grafico dei residui.

Per depurare il grafico dai residui anomali si deve cliccare due volte sull'immagine del grafico, poi si clicca il residuo che si vuole eliminare (il numero del residuo compare nella barra Row) e tramite questo tasto  lo si toglie.



Per salvare i residui della regressione multipla si clicca sul tasto  .



Spuntando **Studentized Residuals** Statgraphics salva i residui standardizzati della regressione nella schermata dove abbiamo inserito la tabella dei nostri dati.