HW5: soluzione degli esercizi assegnati
Modelli Statistici a.a. 2019/20

November 11, 2019

Esercizi del cap. 8: 3, 4 (dati in salary_alr3) e 5.

Esercizio 3 capitolo 8

Date 40 osservazioni sulle variabili Y , X, Z e W si adatta un modello di regressione lineare multipla normale considerando Y come risposta e le altre variabili come esplicative, ottenendo

$$\begin{array}{rccccc} \hat{Y} = & 199.5 & -9.38x & -0.68z & +9.29w \\ & & (2.24) & (0.56) & (1.94) \end{array}$$

dove i valori tra parentesi sono gli errori standard. La devianza residua è $dev(\hat{\epsilon}_0) = 480.5$. Si adatta poi il sottomodello, prendendo solo X come esplicativa ottenendo

$$\begin{array}{rcc} \hat{Y} = & 204.56 & -9.16x \\ & & (2.83) \end{array}$$

con devianza residua $dev(\hat{\epsilon}_0) = 813.3$.

a) Determinare un intervallo di confidenza al livello del 95% per il coefficiente di regressione di Z.

   L'intervallo di confidenza (IC) al 95% per il coefficiente di Z è dato da:

   $$\hat{\beta}_{YZ.XW} \pm t^*_{0.025,(40-4)} \times SE$$

   Il quantile superiore $t^*_{c/2,n-k}$ per $(1-c) = 0.95$ è $t^*_{0.025,36} = 2.028$ (per calcolare $t^*_{0.025,36}$ si può per esempio utilizzare la funzione di Stata `invt(36,0.975)`). L'IC richiesto è dunque:

   $$-0.68 \pm 2.028 \times 0.56,$$

   cioè $[-1.816, 0.456]$. Si noti che questo intervallo contiene lo zero e quindi coefficiente $\hat{\beta}_{YZ.XW}$ non è significativo.

b) Stimare la varianza degli errori nel primo modello.

La stima della varianza di errore è $S^2 = dev(\hat{\epsilon}_0)/(n-4) = 480.5/(40-4) = 13.347$, da cui $S = 3.653$.

c) Sottoporre a verifica l'ipotesi di uguaglianza a zero dei coefficienti di Z e W.

Il test congiunto dell'ipotesi $H_0 : \beta_{YZ.XW} = \beta_{YW.XZ} = 0$ contro $H_1 : \beta_{YZ.XW} \neq 0 \, o \, \beta_{YW.XZ} \neq 0$ si effettua usando la statistica test

$$F = \frac{[dev(\epsilon_0) - dev(\epsilon)]/(g_0 - g)}{dev(\epsilon)/g}$$

dove $g_0 - g$ è pari al numero di parametri posti a zero in H_0, $g = n - k$ sono i gradi di libertà del modello completo. La statistica F risulta:

$$f_{oss} = \frac{[813.3 - 480.5]/2}{480.5/36} = 12.467$$

Sotto ipotesi nulla il test ha distribuzione $F_{2,36}$, il cui quantile superiore al livello 0.05 è 3.26 (si può calcolare il quantile superiore utilizzando per esempio la funzione di Stata `invF(2,36,0.95)`). Pertanto il test è significativo e concludiamo che c'è evidenza empirica contraria all'ipotesi. Il modello ridotto NON è adeguato.

Esercizio 4, capitolo 8

In uno studio sugli stipendi di un ateneo americano nel 1980, si sono raccolte su un campione di 52 docenti le variabili Salary, lo stipendio annuo in dollari, Year gli anni di anzianità, Sex, il genere (1 = femmina, 0 = maschio), Degree, un indicatore del titolo di dottorato (1 = ha il dottorato, 0 = ha la laurea magistrale). I dati sono contenuti nel data frame salary, nella libreria alr3 di R.

Il file `ex4_ch8.do` contiene i comandi di Stata utilizzati per svolgere questo esercizio.

a) Studiate la dipendenza dello stipendio dal sesso. Qual è l' "effetto" del sesso sul salario? Interpretate con una frase il coefficiente e valutate se è significativo.

La Figura 1 riporta il modello stimato. Si noti che il salario medio atteso per le donne è più basso di quello dei maschi di un importo pari a -3339.647 dollari. Tuttavia il coefficiente non è molto significativo (il p-value associato alla statitsica t è 0.071).

b) Studiate quindi il modello di regressione dello stipendio dal sesso e dall'anzianità.

Ottenete le stime. Le stime sono riportate nella Figura 2.

Figure 1: Modello di regressione dello stipendio dal sesso.

```
. reg Salary Sex
```

Source	SS	df	MS	Number of obs	=	52
				F(1, 50)	=	3.41
Model	114106220	1	114106220	Prob > F	=	0.0706
Residual	1.6716e+09	50	33432472.8	R-squared	=	0.0639
				Adj R-squared	=	0.0452
Total	1.7857e+09	51	35014310.9	Root MSE	=	5782.1

Salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Sex	-3339.647	1807.716	-1.85	0.071	-6970.55	291.257
_cons	24696.79	937.9776	26.33	0.000	22812.81	26580.77

Figure 2: Modello di regressione dello stipendio dal sesso e dall'anzianità.

```
. reg Salary Sex Year
```

Source	SS	df	MS	Number of obs	=	52
				F(2, 49)	=	23.65
Model	877036388	2	438518194	Prob > F	=	0.0000
Residual	908693470	49	18544764.7	R-squared	=	0.4911
				Adj R-squared	=	0.4704
Total	1.7857e+09	51	35014310.9	Root MSE	=	4306.4

Salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Sex	201.4668	1455.145	0.14	0.890	-2722.757	3125.69
Year	759.0138	118.3363	6.41	0.000	521.2082	996.8195
_cons	18065.41	1247.774	14.48	0.000	15557.91	20572.9

Interpretate i coefficienti. Il coefficiente associato a `Year` è pari a 759.0138 dollari: per ogni anno di anzianità in più il salario atteso aumenta di 759$. Condizionatamente all'anzianità, il coefficiente associato a `Sex` non è più significativo.

Qual è la differenza importante tra i due modelli?

Nel modello di regressione multipla l'interpretazione dei coefficienti cambia. I coefficienti associati alle covariate sono detti coefficienti di regressione parziale, in quanto misurano l'effetto lineare di una covariata X sulla risposta Y *al netto* dell'effetto lineare dell'altra covariata. Quindi, dopo aver depurato le variabili dall'anzianità, la correlazione parziale tra stipendio e sesso non è più significativa.

Spiegate perchè si ottiene questa differenza.

Per capire da cosa dipende questa differenza osserviamo (Figura 3) che le donne hanno un salario in media più basso, ma anche che l'esperienza media

nel gruppo delle donne è la metà di quella degli uomini!

Figure 3: Stipendio medio e esperienza media per sesso.

```
. table Sex, contents(mean Salary mean Year)
```

Sex	mean(Salary)	mean(Year)
0	24696.789	8.7368422
1	21357.143	4.0714288

c) Introducete nel modello anche l'indicatore del dottorato e ripetete l'analisi. Ci sono differenze importanti?

Figure 4: Modello di regressione dello stipendio dal sesso, dall'anzianità e dal dottorato.

```
. reg Salary Sex Year Degree

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(3, 48)        =     15.50
       Model |  878720962          3  292906987    Prob > F        =    0.0000
    Residual |  907008896         48  18896018.7   R-squared       =    0.4921
-------------+----------------------------------   Adj R-squared   =    0.4603
       Total |  1.7857e+09        51  35014310.9   Root MSE        =      4347

------------------------------------------------------------------------------
      Salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Sex |    190.5843   1469.313     0.13   0.897    -2763.668    3144.837
        Year |    763.4602   120.3764     6.34   0.000     521.4272    1005.493
      Degree |    382.3975   1280.723     0.30   0.767    -2192.668    2957.463
       _cons |    17785.04   1571.026    11.32   0.000     14626.28     20943.8
------------------------------------------------------------------------------
```

Dai risultati riportati in Figura 4, si osserva che per coloro che hanno il dottorato ci si attende un valore medio dello stipendio più elevato (+382 \$), ma la differenza rispetto al sotto-gruppo di coloro che non hanno il dottorato non è significativa.

d) Sottoponete a test l'ipotesi che sia l'effetto del sesso che quello del titolo di dottorato siano simultaneamente nulli, usando un test appropriato.

Per fare questo test usiamo il test F, confrontando la devianza di errore del modello con solo `Year` con la devianza di errore del modello con `Year`, `Sex`

e `Degree`. Il test risulta non significativo (la statistica test e il p -value si possono ottenere con il comando di Stata `test Sex Degree`):

$$F(2, 48) = 0.05 \ , \ Prob > F = 0.9475.$$

Esercizio 5, capitolo 8

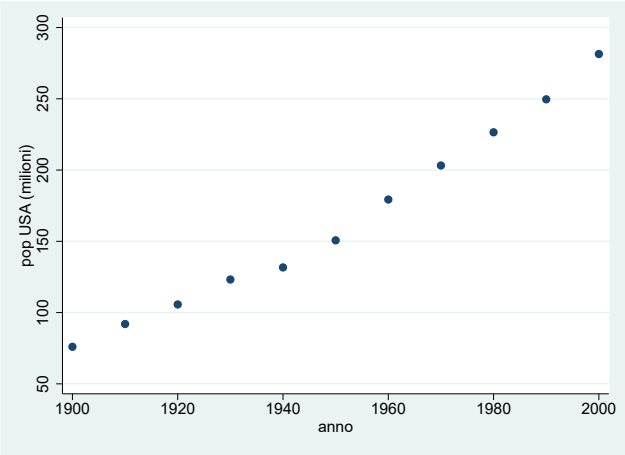
Nella tabella seguente è riportata la popolazione degli USA (in milioni di abitanti) dal 1900 al 2000, rilevata ogni 10 anni.

1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
75.995	91.972	105.711	123.203	131.669	150.697	179.323	203.212	226.505	249.633	281.422

Le istruzioni di Stata per svolgere l'esercizio sono nel file `ese5_ch8.do`.

a) *Fate un grafico che riporti in ascissa il tempo x e in ordinata la popolazione y.* La Figura 5 riporta il grafico ottenuto con il comando `twoway (scatter y x, sort)`.

Figure 5: Popolazione degli USA (in milioni di abitanti) dal 1900 al 2000.



b) *Adattate un modello lineare e un modello quadratico. Riportate sul grafico il modello stimato con i minimi quadrati.* La Figura 6 riporta le stime ottenute con i due modelli. Si noti che il termine di secondo grado nel modello quadratico è significativo. Inoltre per questo modello R^2 è più elevato e l'errore quadratico medio (`rmse`) sensibilmente più piccolo. Il grafico con le due funzioni di regressione stimate è riportato in Figura 7: si noti il miglior adattamento del modello quadratico.

c) *Valutate guardando i residui quale dei due modelli sembra più appropriato.* La Figura 8 riporta il grafico dei residui rispetto all'anno per i due modelli. Si osservi che mentre i residui dal modello lineare mostrano una relazione di tipo quadratico tra residuo e anno, i residui dal modello quadratico mostrano un andamento casuale.

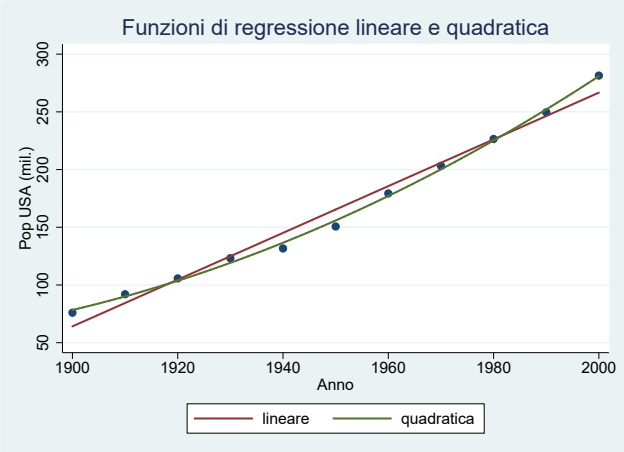
Figure 6: Modelli di regressione lineare e quadratico per la popolazione degli USA (in milioni di abitanti) dal 1900 al 2000.

```
. estimate table M1 M2, stats(r2 rmse) b(%7.3f) star
```

Variable	M1	M2
x	2.025***	-34.987***
c.x#c.x		0.009***
_cons	-3.8e+03***	3.2e+04***
r2	0.981	0.998
rmse	9.864	3.585

legend: * p<0.05; ** p<0.01; *** p<0.001

Figure 7: Popolazione degli USA (in milioni di abitanti) dal 1900 al 2000: dati osservati e funzioni di regressione adattate.



d) *Adattate un polinomio di grado 8 e riportate il polinomio stimato sul grafico.* Per verificare l'assunzione di linearità si può guardare lo scatterplot dei residui rispetto all'anno di Figura 8: l'andamento osservato nel grafico suggerisce una relazione non monotona.

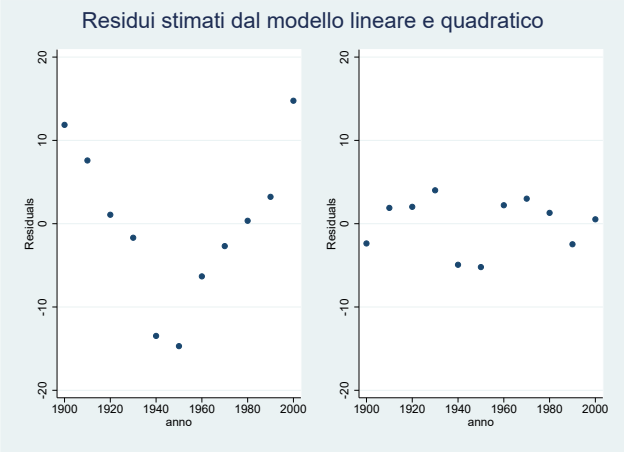
Per tenere conto di questo andamento si può usare un modello di regressione polinomiale del tipo:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_h X^k$$

dove k è il grado del polinomio. Anche se questo modello consente di specificare una relazione non lineare tra Y e X è ancora un modello di regressione lineare, perchè è lineare rispetto ai coefficienti di regressione $\beta_1, \ldots \beta_k$.

Un problema che si incontra nella stima di un modello polinomiale è che il cal-

Figure 8: Residui dei modelli di regressione lineare e quadratico rispetto all'anno.



colo dell'inversa della matrice $\mathbf{X}'\mathbf{X}$ diventa inaccurato al crescere dell'ordine del polinomio, introducendo così errori nella stima dei parametri. In particolare, se i valori di X variano poco il problema della multicollinearità è sicuramete presente.

Nel nostro caso i valori di X vanno da 1900 a 2000, quindi un il campo di variazione in termini percentuali del 5% $(100 \times (2000 - 1900)/1900)$. Per ovviare al problema, possiamo creare una nuova variabile scartando i valori $\widetilde{x} = x - 1900$. I calcoli e la stima dei vari modelli polinomiali sono riporati nel file `ese5_ch8.do`. Osserviamo che si riesce a stimare un polinomio fino al grado $k = 7$. quando si arriva a $k = 8$ Stata omette automaticamente il termine di ottavo grado e restituisce un messaggio di errore, per via della multicollinearità.

In conclusione, un polinomio di secondo grado non è ottimale per descrivere l'andamento della popolazione nel tempo, infatti abbiamo notato che permangono delle non-linearità. Tuttavia un polinamio di grado superiore al secondo non si adatta bene ai dati: i termini di grado superiore a 2 sono tutti non significativi, già a partire da un polinomio di terzo grado. Questo succede perchè l'andamento della popolazione è diverso nei diversi periodi considerati. Un apprccio che si usa spesso in questi casi consiste nel suddividere l'intervallo in sotto-intervalli per poi adattare una funzione di regressione diversa per ciascun sotto-intervallo. Questo adattamento polinamiale a tratti si può fare introducendo degli indicatori di ciascun sotto-periodo, oppure per via non parametrica (per esempio con stimatori kernel, utilizzando il comando `lpoly` di Stata come riporato nel file `ese5_ch8.do`).

e) *Determinate la popolazione stimata degli USA nel 2020 usando il modello quadratico e l'ultimo modello polinomiale. Cosa si nota?* Il modello di ottavo grado non si può stimare. Per fare il confronto con il modello quadratico, consideriamo un modello polinomiale di grado $k = 5$. La popolazione stimata

per il 2020 con il modello quadratico è pari a 344.1787 milioni di abitanti: un valore molto elevato! Poichè il modello polinomiale di 5^o grado prevede una diminuzione, la previsione che si ottiene è sensibilmente più piccola e più verosimile: 105.9479 milioni di abitanti.