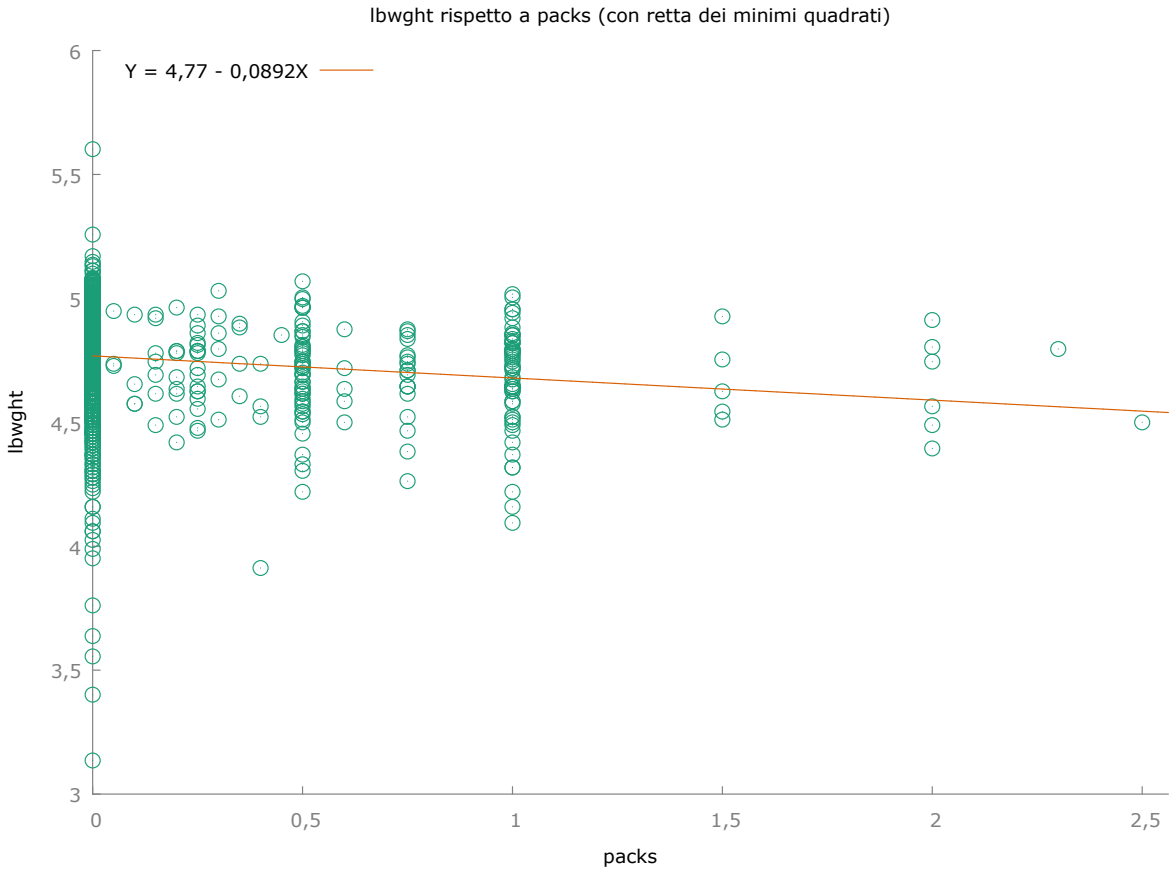
2. I risultati della stima OLS del modello preliminare mostrano che in media c'è una tendenza per la variabile dipendente (peso del neonato) ad assumere valori più bassi all'aumentare del numero dei pacchetti di sigarette fumati. In un modello con una sola esplicativa, se ipotizziamo una situazione in cui non sono presenti altre variabili in grado di spiegare il peso del neonato, la direzione di causalità è univocamente definita: un consumo di sigarette intensivo durante il periodo di gravidanza porta ad un peso minore del neonato. Il parametro β_1 mostra, infatti, che per ogni pacchetto di sigarette fumato al giorno ci si può aspettare in media una riduzione di quasi il 9% rispetto al peso medio dei neonati. Il parametro β_0 rappresenta, invece, il peso medio dei neonati quando l'effetto della variabile indipendente è nullo (condizione in cui non si fumano pacchetti di sigarette durante la gravidanza). Nel campione selezionato il parametro β_0 spiega in maniera molto forte il modello. Questo può essere dovuto alla presenza di pochi fumatori nel campione e, dunque, il peso medio dei bambini associato a tali valori non è molto diverso dal valore medio del peso dei neonati delle donne fumatrici. Sono pertanto necessarie altre variabili per identificare in modo più efficiente l'effetto delle sigarette sul peso dei neonati o individuare con maggiore certezza un nesso di causalità tra le due variabili.



Il valore del parametro β_0 (4,768) è molto vicino alla media della variabile dipendente (4,758). Dato il suo errore standard molto basso, in un test di verifica di ipotesi sull'intercetta si trova un valore della t di Student superiore a 800. La significatività del parametro è dunque

estremamente alta. Allo stesso modo anche l'effetto dei pacchetti di sigarette è rilevante (t < -5). Viene evidenziata, dunque, una tendenza significativa nel sottogruppo di neonati con donne fumatrici ad assumere un peso sotto la media.

3. Dai risultati del modello OLS si nota che l'R-quadro è vicino a 0. Il modello non riesce a fornire attraverso una relazione di tipo lineare una chiara rappresentazione del fenomeno con un solo regressore. Nonostante la variabile dei pacchetti di sigarette sia fortemente significativa, essendo il numero delle donne fumatrici in gravidanza molto basso rispetto a quello delle donne non fumatrici, packs non riesce a spiegare la parte di varianza relativa a quel sottogruppo (non fumatrici). È proprio questo che spiega l'elevata importanza della costante: in presenza di una donna che non fuma, il peso del neonato non può fare a meno che essere stimato attraverso il valore medio del peso dei neonati del sottogruppo delle donne non fumatrici. Il modello minimizza gli errori approssimando il valore previsto al valore medio del sottogruppo.

Modello 1 prelim:	coefficiente	Errore standard	Statistica t	p-value
Intercetta	4,76800	0,00550367	866,3	0,0000
packs	−0,0892021	0,0169597	−5,26	1,68e-07
Media variabile dipendente	4,758492			
R-quadro	0,019775			

4. La sostituzione della variabile packs con cigs non porta alcun miglioramento, ma neanche un peggioramento al modello preliminare. Questo avviene perché la variabile cigs è una combinazione lineare della variabile packs. Siccome i risultati esplicativi del modello non sono cambiati, nonostante sia variato il parametro β_1 e il suo errore standard, è possibile determinare sulla base del confronto tra i coefficienti β_1 che il numero di sigarette contenuto nei pacchetti è 20.

5. Nel modello:

$$\log(bwght_i) = \beta_0 + \beta_1 packs_i + \beta_2 male_i + \beta_3 parity_i + \beta_4 \log(faminc_i) + u_i$$

Modello 2 (base):	coefficiente	Errore standard	Statistica t	p-value
Intercetta	4,67731	0,0208636	224,2	0,0000
packs	−0,0830647	0,0176624	−4,703	2,83e-06
lfaminc	0,0180313	0,00539497	3,342	0,0009
male	0,0213844	0,0102438	2,088	0,0370
parity	0,0145343	0,00550201	2,642	0,0083
Media variabile dipendente	4,758492			
R-quadro corretto	0,030382			

La costante resta una variabile importante visto il suo valore t > 200. Rispetto al modello precedente il suo coefficiente si è ridotto: ciò implica che parte della sua forza esplicativa è entrata in altri fattori prima non considerati. Il coefficiente della variabile packs, cosi come il

suo standard error, sono invece aumentati leggermente. Ciò può essere dovuto alla presenza di multicollinearità con le variabili appena introdotte. La probabilità di errore rifiutando l'ipotesi nulla resta comunque molto bassa.

Le variabili esplicative prima omesse sono risultate significative nello spiegare la variabile dipendente. Di particolare importanza è l'esplicativa che esprime il logaritmo del reddito: l'elasticità del peso del neonato rispetto al reddito è positiva. Viene individuata, dunque, una relazione secondo cui al crescere del reddito ci si aspetta in media un peso del neonato superiore mantenendo fissi gli altri fattori. Altresì importante è l'effetto generato dalla variabile parity: la stima evidenzia che in media ciascun figlio aggiuntivo al primo presenta un peso maggiore (1.5%). La variabile male permette di distinguere per genere i neonati: questa suddivisione in gruppi è da ritenersi importante perché sussistono differenze oggettive tra il peso alla nascita dei maschi e delle femmine. Il coefficiente della variabile male è 0.02: ciò significa che in media i neonati maschi (individuati dal valore 1 della variabile dicotomica male) pesano il 2% in più rispetto alle femmine.

6. Il secondo modello non vincolato presenta un R-quadro corretto maggiore del modello con una esplicativa. L'aggiunta di ulteriori variabili ha diminuito l'errore standard della regressione perché alcune variabili prima omesse erano in grado di spiegare la variabile dipendente. La stima del modello preliminare è dunque affetta da distorsione da variabili omesse. L'importanza di queste variabili è confermata attraverso il test F. La F individua un valore pari a 7,81 sotto l'ipotesi nulla che le variabili prima omesse siano uguali a 0. La probabilità di errore rifiutando l'ipotesi nulla è molto bassa (pressoché nulla). Questo ci indica che le nuove variabili esplicative introdotte nel modello sono significative a livello statistico, il che conferma l'importanza di includerle nel modello.

```
Ipotesi nulla: i parametri della regressione valgono zero per le variabili
  lfaminc, male, parity
Statistica test: F robusta(3, 1334) = 7,81845, p-value 3,56404e-005
L'omissione delle variabili ha migliorato 1 dei 3 criteri di informazione.
```

7.

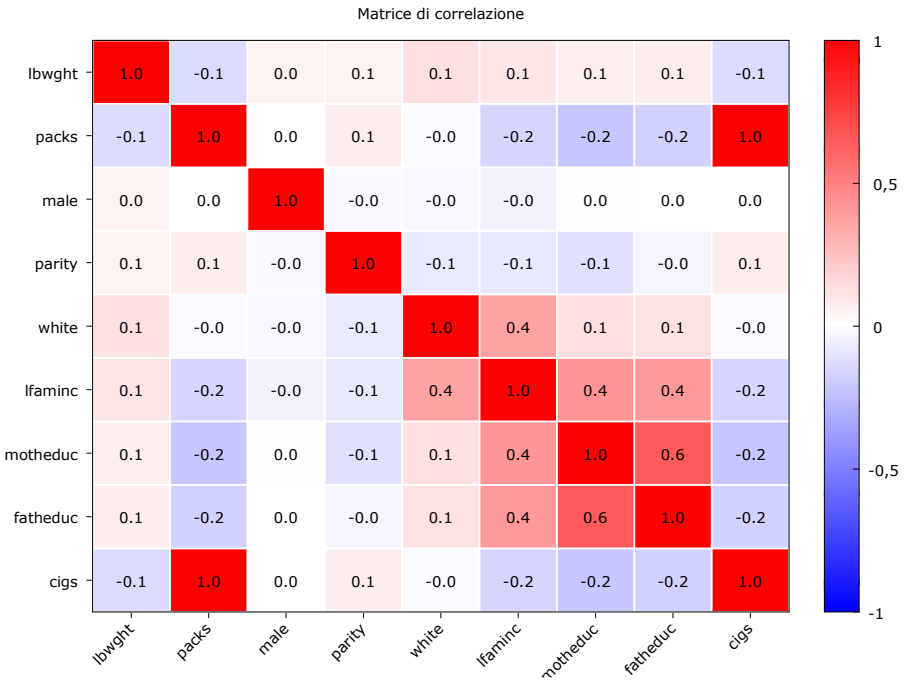
Modello 1:	coefficiente	Errore standard	Statistica t	p-value
intercetta	4,76800	0,00550367	866,3	0,0000
packs	−0,0892021	0,0169597	−5,26	1,68e-07
R-quadro	0,019775			
Modello 2:				
Intercetta	4,67731	0,0208636	224,2	0,0000
packs	−0,0830647	0,0176624	−4,703	2,83e-06
lfaminc	0,0180313	0,00539497	3,342	0,0009
male	0,0213844	0,0102438	2,088	0,0370
parity	0,0145343	0,00550201	2,642	0,0083
R-quadro corretto	0,030382			

Modello 3:				
Intercetta	4,66128	0,0217819	214,0	0,0000
packs	−0,0861549	0,0177503	−4,854	1,35e-06
lfaminc	0,00978990	0,00597743	1,638	0,1017
male	0,0219495	0,0101822	2,156	0,0313
parity	0,0157566	0,00553206	2,848	0,0045
white	0,0502679	0,0154560	3,252	0,0012
R-quadro corretto	0,039814			
Modello 4:				
Intercetta	4,65739	0,0398853	116,8	0,0000
packs	−0,103730	0,0208428	−4,977	7,46e-07
lfaminc	0,0109039	0,00809160	1,348	0,1781
male	0,0299584	0,0109341	2,740	0,0062
parity	0,0168272	0,00602560	2,793	0,0053
white	0,0417112	0,0183108	2,278	0,0229
fatheduc	0,00337324	0,00237371	1,421	0,1556
motheduc	−0,00303210	0,00280314	−1,082	0,2796
R-quadro corretto	0,039630			
Modello 5:				
const	4,67574	0,0209904	222,8	0,0000
packs	−4,2552e+06	7.26E+11	−0,5857	0,5582
cigs	0,0188085	0,00556726	3,378	0,0008
male	0,0206776	0,0103350	2,001	0,0456
parity	0,0147158	0,00549723	2,677	0,0075
lfaminc	212758	370728	0,5739	0,5661
R-quadro corretto	0,032112			

8. Nei primi modelli considerati si è ipotizzato che non esistessero altre variabili in grado di spiegare la variabile dipendente, il modello risultante era una forte semplificazione del fenomeno che non prendeva in considerazione altri fattori esterni. L'aggiunta di altre variabili statisticamente significative è stata importante nel ridurre l'incertezza sul valore teorico della regressione e migliorare il fit. Nonostante l'aggiunta di alcune variabili abbia fatto aumentare l'errore standard della variabile packs, i regressori introdotti sono significativi e devono essere inclusi. Il terzo modello presenta l'R-quadro più elevato: particolare attenzione è da porre a ciò che è successo alla variabile del reddito familiare. L'errore, rifiutando l'ipotesi che il coefficiente di regressione sia nullo, non cade più dentro l'intervallo di confidenza del 95%, questo aumento può essere dovuto all'introduzione della variabile white. Analizzando le correlazioni tra le due variabili (0.36) si nota che in media la componente relativa al colore del neonato è moderatamente correlata con il reddito familiare. L'esplicativa del reddito potrebbe

essere comunque correlata con fattori esterni (endogena). Questo può portare a distorsioni nella stima del suo effetto sulla variabile dipendente.

Il quarto modello aggiunge al terzo modello variabili relative al livello di istruzione dei genitori. Non è rilevante determinare a priori il peso del neonato con tali variabili: ipotizzare che lo siano significa porre una relazione di causalità tra il numero di anni di istruzione dei genitori ed il peso del neonato. Tuttavia, il contributo dato dal numero di anni di istruzione non influisce in alcun modo sul peso del neonato. Se il modello evidenzia significatività in tali parametri, significa che sussistono fattori esterni non inclusi nel modello per spiegare la variabile dipendente. Le due variabili potrebbero, per esempio, esercitare un effetto sull'esplicativa del reddito. Oltretutto, dallo studio delle correlazioni si nota che tra tutte le variabili analizzate, i livelli di istruzione dei genitori sono i più correlati tra loro: questo rende complesso isolare gli effetti di una variabile rispetto all'altra e porta ad una maggiore incertezza nella stima. Nel modello in cui sono presenti in contemporanea sia la variabile cigs che packs, per la condizione di ceteris paribus, Gretl non è in grado di determinare la stima attraverso il modello OLS perché la correlazione $\rho = 1$ (è presente una perfetta combinazione lineare). Se varia la variabile pack tenendo fisso cigs, devono necessariamente variare anche le sigarette.



Idealmente, per spiegare in modo più accurato la variabile relativa al peso del neonato, si potrebbe analizzare attraverso l'uso di variabili dicotomiche la presenza o meno di problemi di salute genetico-ereditari (per esempio casi di obesità riconducibili a fattori genetici): tale variabile potrebbe spiegare i valori estremi anomali (outliers).

Il modello che ci risultava essere migliore per analizzare il fenomeno è quello in cui viene esclusa la variabile relativa al reddito. L'obiettivo dell'analisi è individuare l'effetto dei pacchi di sigarette fumati sul peso dei neonati. L'esclusione della variabile, nonostante riduca leggermente il valore dell'R-quadro rispetto al terzo modello, porta ad una riduzione

dell'incertezza relativa alla variabile dei pacchetti di sigarette, restringendo l'intervallo di confidenza del rispettivo coefficiente di regressione. Depura inoltre il modello dagli effetti dovuti alla correlazione tra le esplicative. Il reddito familiare può, inoltre, essere correlato con il termine di errore, ciò porta a distorsioni nella stima del modello (causalità simultanea). L'analisi presuppone uguaglianza tra tutti i pacchi di sigarette, tuttavia in pratica i pacchetti possono differire tra loro. Una misura che depura dal potenziale errore di misura è cigs, tuttavia nei sondaggi è difficile, specialmente per donne fumatrici, ricordarsi l'esatto numero di sigarette fumate (servirebbe un contatore) per cui nel modello, nonostante sia presente un margine di errore, viene mantenuta la variabile packs come esplicativa. L'errore di misura di cigs potrebbe essere molto più elevato di packs.

Il modello presenta i seguenti risultati di stima:

Modello finale:	coefficiente	Errore standard	Statistica t	p-value
Intercetta	4,68686	0,0176880	265,0	0,0000
packs	−0,0907940	0,0174384	−5,207	2,23e-07
lfaminc	0,0212300	0,0102245	2,076	0,0380
male	0,0152350	0,00556168	2,739	0,0062
parity	0,0580805	0,0141004	4,119	4,04e-05
white				
R-quadro corretto	0,030382			

9. È plausibile che la variabile packs sia correlata con il termine di errore. Il consumo dei pacchetti di sigarette può essere, infatti, ricondotto a dipendenze o ad abitudini che possono però essere influenzate da fattori esterni. Per esempio, coloro che fumano per moda, o coloro che desiderano uscire dalla loro dipendenza potrebbero passare a beni sostituti (sigarette elettroniche), ciò altera il loro comportamento in modo significativo sul numero di pacchetti di sigarette fumate. Le campagne di sensibilizzazione contro il fumo, l'assenza di politiche di welfare per i fumatori di certe aree e la preoccupazione di dover pagare cifre molto elevate per la cura di malattie legate al fumo possono portare altresì ad un disincentivo al consumo di sigarette. Effetti più diretti possono essere legati, invece, al prezzo delle sigarette. La variabile packs è anche sensibile all'errore di misura, i pacchetti di sigarette, infatti, possono contenerne un numero diverso, viene preso un valore medio per le sigarette contenute.

10. Le variabili cigtax e cigprice sono rilevanti a priori per la variabile packs. Infatti, un aumento del prezzo dei pacchi di sigarette, cosi come un aumento nella tassazione, porta secondo la teoria microeconomica ad una riduzione del consumo dei pacchi fumati in caso di assenza di una funzione di domanda perfettamente inelastica. L'analisi non tiene però in considerazione il fatto che certe società, per mantenere competitività sul lato dei prezzi, possano rifiutare di internalizzare la tassazione e sopportare in modo completo l'onore dell'imposta. Di norma la correlazione di cigtax e cigprice rispetto a packs è dunque negativa. La variabile cigtax è esogena, non sono presenti fattori esterni correlati con il livello di tassazione nel termine di errore poiché il valore viene imposto dal legislatore. Cigprice, invece, è una variabile endogena, è influenzabile sia da fattori microeconomici (tassazione, mode, concorrenza) sia macroeconomici (prezzo dei fattori produttivi, forza sindacati, inflazione).

Modello TSLS 1 (Inst. Cigtax)	coefficiente	Errore standard	Statistica t	p-value
Intercetta	4,39152	0,330325	13,29	5,77e-038
packs	1,10873	1,41985	0,7809	0,4350
male	0,0261144	0,0218165	1,197	0,2315
parity	−0,0070832	0,0333827	−0,2122	0,8320
lfaminc	0,0806133	0,0737974	1,092	0,2749
R-quadro corretto	0,011191			
Modello TSLS 2 (Inst. cPrice)	coefficiente	Errore standard	Statistica t	p-value
Intercetta	4,42949	0,318384	13,91	3,46e-041
packs	0,950419	1,36113	0,6983	0,4851
male	0,0254861	0,0197690	1,289	0,1976
parity	−0,0042116	0,0307479	−0,1370	0,8911
lfaminc	0,0723001	0,0706996	1,023	0,3067
R-quadro corretto	0,010782			

Modello TSLS 3 (Sovraident.)	coefficiente	Errore standard	Statistica t	p-value
Intercetta	4,39571	0,325404	13,51	4,55e-039
packs	1,09126	1,39877	0,7802	0,4354
male	0,0260451	0,0215717	1,207	0,2275
parity	−0,0067663	0,0328793	−0,2058	0,8370
lfaminc	0,0796959	0,0726767	1,097	0,2730
R-quadro corretto	0,011152			

Modello OLS 1 (dip. packs). Primo stadio	coefficiente	Errore standard	Statistica t	p-value
Intercetta	0,210414	0,181863	1,157	0,2475
cigtax	0,00094549	0,00231675	0,4081	0,6833
cigprice	9,58113e-05	0,00169608	0,05649	0,9550
Male	−0,0047749	0,0162828	−0,2933	0,7694
parity	0,0180292	0,0118560	1,521	0,1286
lfaminc	−0,0529557	0,0113798	−4,653	3,59e-06
R-quadro corretto	0,027379			

Modello OLS 2 (dip. bwght)	coefficiente	Errore standard	Statistica t	p-value

Secondo stadio				
Intercetta	4,39571	0,149011	29,50	1,17e-147
Fit salvato	1,09126	0,622142	1,754	0,0797
male	0,0260451	0,0103775	2,510	0,0122
parity	−0,00676636	0,0129293	−0,5233	0,6008
lfaminc	0,0796959	0,0329050	2,422	0,0156
R-quadro corretto	0,015875			

11. L'uso delle variabili strumentali ha portato ad un peggioramento del fit. Ciò può essere dovuto alla poca rilevanza degli strumenti inclusi nel modello. In particolare, analizzando le correlazioni tra le variabili strumentali e packs, si nota che tale valore è molto basso, vicino allo 0. Ciò evidenzia una debole relazione tra le variabili che rende lo strumento inadatto, sebbene a livello teorico ci aspettassimo un impatto maggiore nello spiegare la variabile packs. La variabile cigprice è endogena. È influenzata in maniera molto forte dalla variabile cigtax, la correlazione tra le due variabili è vicino a 0.9. Questo può portare a collinearità specialmente al primo stadio nel modello OLS. Siccome le due variabili non sono rilevanti nello spiegare il numero dei pacchetti di sigarette fumate, sia il modello esattamente identificato sia quello sovra-identificato peggiorano il modello di base.

12.

	Modello OLS	Modello cigtax	Modello cigprice	TSLS overidentified
Intercetta	4,67731	4,39152	4,42949	4,39571
packs	−0,0830647	1,10873	0,950419	1,09126
male	0,0180313	0,0261144	0,0254861	0,0260451
parity	0,0213844	−0,0070832	−0,0042116	−0,0067663
lfaminc	0,0145343	0,0806133	0,0723001	0,0796959
R-quadro corretto	0,030382	0,011191	0,010782	0,011152

13. Il modello esattamente identificato che ha come strumento la variabile cigtax genera risultati inattesi e contrastanti: se in precedenza avevamo individuato una relazione negativa tra il peso e il numero dei pacchetti di sigarette fumate, adesso il modello individua un valore positivo del coefficiente. Tale valore a noi sembra distorto ed è dovuto alla poca rilevanza dello strumento selezionato. Il fit del modello è inferiore al modello di base. Analogamente, il modello che pone come variabile strumentale cigprice, presenta un valore positivo per il coefficiente di packs ed errori standard molto elevati. L'incertezza della sua stima è aumentata in maniera considerevole ed il fit del modello si è ridotto rispetto al modello con variabile strumentale cigtax. Il test di Hausman in entrambi i casi mostra valori che tenderebbero a considerare i risultati di stima inconsistenti.

Nel modello sovraidentificato, il test di Hausman mostra che le stime OLS sono inconsistenti se rigettiamo un p-value superiore a 0.05. Visti i coefficienti ottenuti siamo propensi a dire che i risultati di stima sono però poco affidabili. L'output del test di Sargan mostra che tra gli

strumenti utilizzati alcuni possono essere poco validi (p-value >0.8). Il test sugli strumenti deboli evidenzia, inoltre, che le variabili strumentali sono inadeguate nello spiegare la variabile endogena. Il fit di tale modello è decisamente inferiore al modello di base. L'incertezza del parametro della variabile packs è aumentata in modo molto forte rispetto al modello di base.

Nella stima OLS a due stadi si ottiene un fit migliore del modello TSLS sia per quanto riguarda quello esattamente identificato sia quello sovraidentificato. Il coefficiente relativo al fit del primo stadio è vicino all'essere significativo, il suo valore continua ad essere positivo anche analizzando l'estremo inferiore nel suo intervallo di confidenza del 95%. La collinearià del fit salvato è molto forte (> 40) con la variabile relativa al reddito familiare: se viene omessa, il coefficiente relativo al fit diventa negativo e il suo errore standard si riduce.

14. Le variabili strumentali disponibili per lo studio del fenomeno non sono rilevanti nello spiegare i pacchi di sigarette fumati. Questo porta, nella parte relativa allo studio dei modelli IV, a delle stime distorte e inconsistenti. L'endogena nel modello (packs) è inoltre affetta da possibili errori di misura. Sulla base del dataset disponibile, troviamo stime più accurate utilizzando gli OLS senza valutare gli effetti generati dalla tassazione e dal livello dei prezzi sulla variabile packs.

Le variabili che possono, dunque, esercitare un effetto diretto sul peso del neonato sono: il numero di figli avuti dalla madre, il sesso del neonato, il reddito e i pacchi fumati. Tutti i modelli OLS considerati mostrano una relazione negativa tra il peso del neonato e il numero dei pacchetti di sigarette fumati dalla madre. I risultati osservati sono in linea con la teoria secondo la quale il fumo danneggia la salute del neonato perché lo priva di ossigeno e del nutrimento per crescere. Il peso medio del neonato di una madre fumatrice si riduce, infatti, di circa l'8% in base al numero di pacchetti di sigarette fumate al giorno.