۱) جدول خلاصه آماری را برای متغیرهای غیر باینری محاسبه کنید و در چند خط نتایج را تفسیر کنید.

Variable	Obs	Mean	Std. dev.	Min	Max
unitvalue	2,433	3672.529	2691.248	200	75000
income	2,433	2.61e+08	1.71e+08	1500000	1.82e+09
weight	2,433	439.3305	280.092	2	2500
age	2,433	44.34525	9.573598	23	86

همانطور که پیداست، دیتاست ۲٤۳۳ مشاهده دارد و میانگین، انحراف معیار، حداکثر و حداقل هر کدام از متغیرهای دیتاست مشخص شده است.

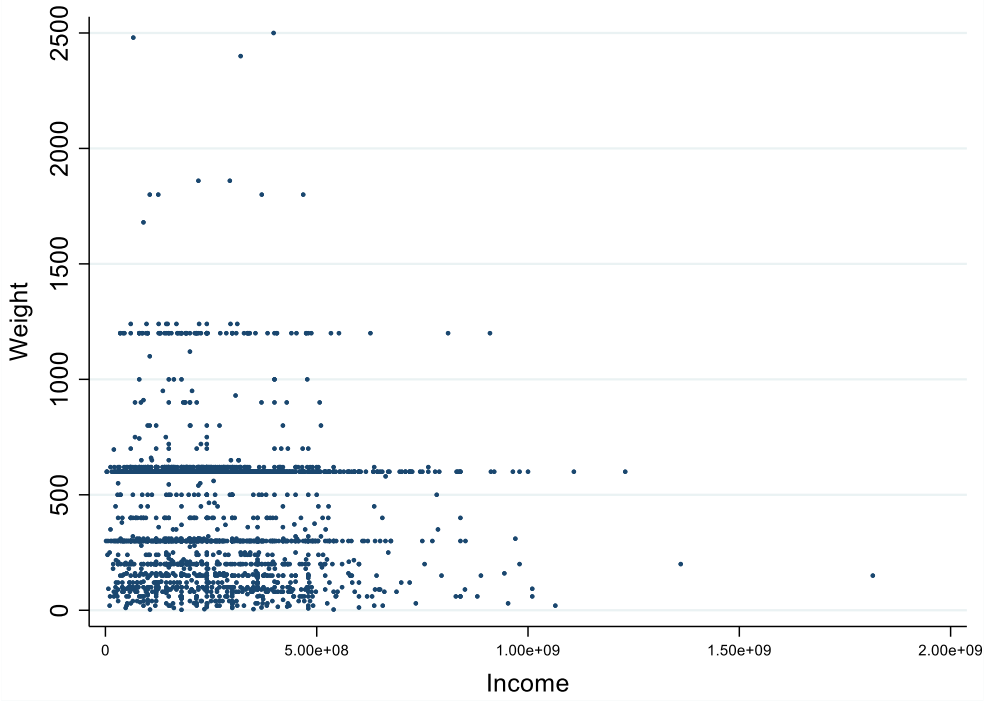
۲) جدول همبستگی متغیرهای غیر باینری را محاسبه نموده و در چند خط تفسیر کنید.

	unitva~e	income	weight	age
unitvalue	1.0000			
income	0.1736	1.0000		
weight	-0.1679	-0.0487	1.0000	
age	-0.0586	-0.0369	0.1179	1.0000

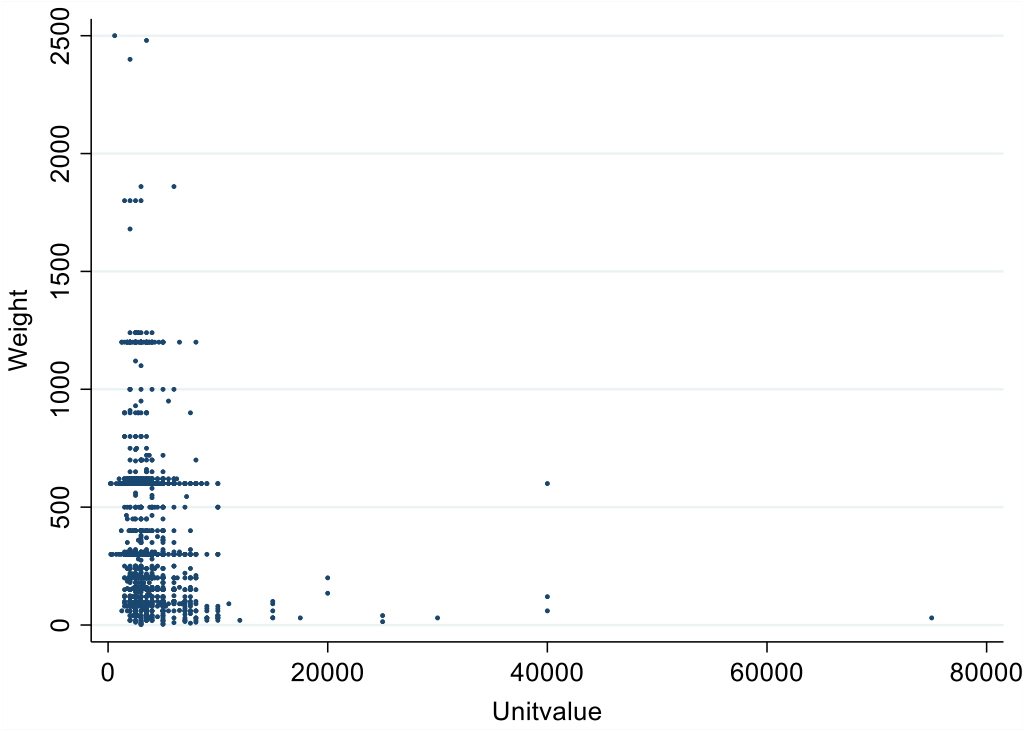
مقدار ضریب همبستگی متغیر قیمت هر عدد سیگار با درآمد فرد مثبت و با متغیر وزن و سن فرد یک عدد منفی است. به این معنا که با افزایش درآمد، قیمت هر واحد سیگار خریداری شده توسط فرد افزایش یافته اما با افزایش سن فرد، او به خرید سیگارهای ارزان‌تر روی می‌آورد. اگر به ضریب همبستگی سن و درآمد نیز توجه شود، مشاهده خواهد شد که با افزایش سن درآمد کاش یافته یا بلعکس. ممکن است که دلیل اینکه با افزایش سن فرد رو به خرید سیگارهای ارزان می‌آورد همین کاهش دستمزد باشد. البته گفتن این نکته ضروریست که جدول همبستگی نشان‌دهندهٔ علیت نیست و جملهٔ بالا یک حدس است که برای نشان دادن درست بودن یا نبود آن به سراغ استنباط آماری رفت.

۳) مقدار مصرف دخانیات و درآمد خانوار را در یک نمودار رسم کرده و رفتار آنها را تفسیر کنید.

با توجه به نمودار به نظر می‌رسد که با افزایش درآمد، وزن مقدار مصرف دخانیات کاهش می‌یابد. اگر به جدول همبستگی رجوع کنیم، این نتیجه تایید می‌شود.



٤) مقدار مصرف دخانیات و قیمت هر واحد را در یک نمودار رسم کرده و رفتار آنها را تفسیر کنید.



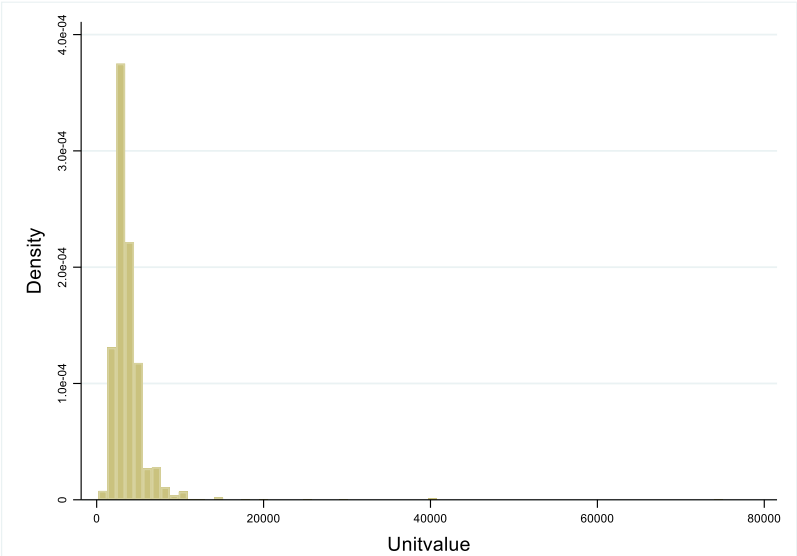
همانند بخش سه، رابطهٔ میان دو متغیر قیمت هر واحد سیگار و مصرف دخانیات یک رابطهٔ معکوس است.

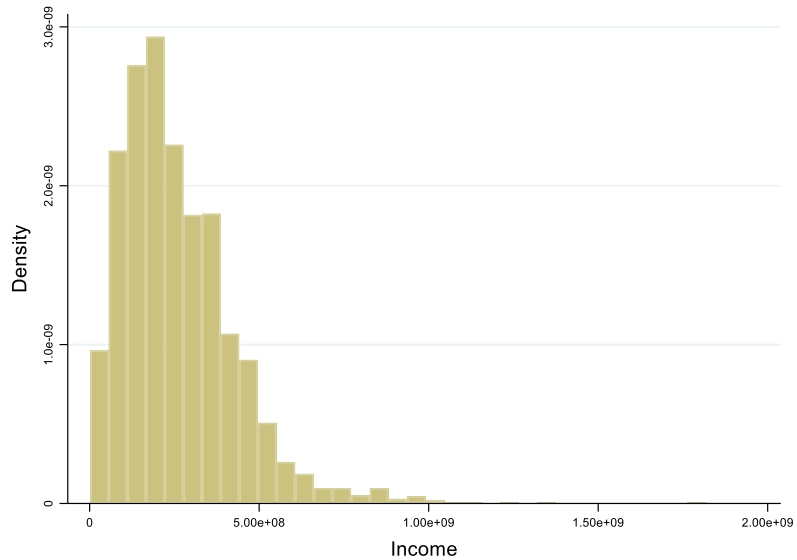
۵) نرمال بودن متغیرهای درآمد خانوار و قیمت هر نخ سیگار را تست کنید و همچنین با رسم نمودار هیستوگرام برای دو متغیر آن را بررسی کنید.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
income	2,433	0.90106	140.201	12.663	0.00000
unitvalue	2,433	0.45146	777.327	17.051	0.00000

فرضیهٔ صفر تست شپیرو–ویلک نرمال بودن توزیع متغیر است و با توجه به آمارهٔ به دست آمده و پی‌ولیو در سطح یک درصد، پنج درصد و ده درصد نرمال بودن توزیع رد می‌شود.





هیستوگرام این دو متغیر شاهدی بر درستی ادعای تست شپیرو–ویلک است.

رگرسیون زیر را برآورد کنید:

$$\text{weight} = \beta_0 + \beta_1 \text{wsregular\_y} + \beta_2 \text{unitvalue} + \beta_3 \text{age} + \beta_4 \text{female} + \beta_5 \text{ceduc} + \beta_6 \text{own} + \beta_7 \text{child\_less\_14} + \varepsilon$$

٦) ضرایب را تفسیر کنید و تست معناداری آن‌ها و کل رگرسیون را انجام دهید.

Source	SS	df	MS		
Model	9012579.94	7	1287511.42	Number of obs	= 2,433
Residual	181781534	2,425	74961.4575	F(7, 2425)	= 17.18
				Prob > F	= 0.0000
Total	190794114	2,432	78451.5273	R-squared	= 0.0472
				Adj R-squared	= 0.0445
				Root MSE	= 273.79

weight	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
income	2.66e-08	3.57e-08	0.75	0.456	-4.34e-08	9.67e-08
unitvalue	-.0161684	.0021001	-7.70	0.000	-.0202866	-.0120502
age	2.751557	.7155617	3.85	0.000	1.348382	4.154733
female	-77.29919	73.74761	-1.05	0.295	-221.914	67.31565
ceduc	-97.33213	24.04575	-4.05	0.000	-144.4845	-50.1798
own	3.847417	12.43939	0.31	0.757	-20.54552	28.24035
child_less_14	-9.434709	13.84527	-0.68	0.496	-36.58449	17.71507
_cons	380.121	38.42759	9.89	0.000	304.7667	455.4753

بنابر رگرسیون تخمین‌زده‌شده، با افزایش یک واحد درآمد، قیمت سیگار و سن، میزان مصرف دخانیات به ترتیب به مقدار ۰,۰۰۰۰۰۰۰۲٦٦ افزایش، ۰,۰۱٦۱٦ کاهش و ۲,۷۵ واحد افزایش می‌یابد. همچنین زنان به طور میانگین حدودا هفتاد هفت کیلو، دارندگان تحصیلات دانشگاهی حدودا ۹۷ کیلو و افرادی که کودک زیر چهارده سال دارند حدودا نه و نیم کیلو کمتر دخانیات مصرف می‌کنند. صاحبان خانه نیز به طور میانگین سه کیلو و هشتصد گرم بیشتر از افرادی که خانه ندارند سیگار می‌کشند. البته گفتن این نکته ضروریست که ضرایب درآمد، جنسیت، صاحب خانه بودن و داشتن فرزند زیر چهارده سال به طور آماری معنادار نیست. و آزمون اف نشان می‌دهد که ضرایب رگرسیون توامان معنی‌دارند.

۷) رگرسیون اصلی را این بار با مقادیر لگاریتمی متغیرهای مستقل غیر باینری (درآمد و قیمت واحد) انجام داده و در مدل جدید ضرایب را تفسیر کنید. کدام مدل بهتر هست؟

ضرایب سن، جنسیت، تحصیلات دانشگاهی، صاحب‌خانه بودن و داشتن فرزند کمتر از چهارده سال را به همان شیوهٔ بالا می‌توان تفسیر کرد. اما تفسیر دو متغیر لگاریتمیک اندکی متفاوت است. با افزایش یک درصد درآمد، میزان مصرف دخانیات به مقدار ۰,۲۳ واحد افزایش می‌یابد و افزایش یک درصدی قیمت سیگار موجب کاهش مصرف به میزان ۱,۲٦ واحد می‌شود. با توجه به افزایش ضریب تعیین تعدیل‌شده و آمارهٔ اف و کاهش خطای جذر میانگین مربعات، مدل دوم مدل بهتری است.

Source	SS	df	MS		
				Number of obs	= 2,433
				F(7, 2425)	= 21.76
Model	11278124.8	7	1611160.69	Prob > F	= 0.0000
Residual	179515989	2,425	74027.2122	R-squared	= 0.0591
				Adj R-squared	= 0.0564
Total	190794114	2,432	78451.5273	Root MSE	= 272.08

weight	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ln_income	23.21075	7.817107	2.97	0.003	7.881849	38.53965
ln_unitvalue	-126.5787	13.36357	-9.47	0.000	-152.7839	-100.3735
age	2.588639	.7112215	3.64	0.000	1.193975	3.983304
female	-66.78519	73.35465	-0.91	0.363	-210.6294	77.05907
ceduc	-94.66647	22.88841	-4.14	0.000	-139.5493	-49.7836
own	.8648095	12.36269	0.07	0.944	-23.37773	25.10734
child_less_14	-13.86073	13.77732	-1.01	0.314	-40.87726	13.15581
_cons	920.5201	169.3002	5.44	0.000	588.5321	1252.508

۸) استدلال خود از خطای تصریح مدل در خصوص مدل جدید را بیان کنید. آیا هیچ متغیر توضیحی مشاهده‌نشده‌ای (غیرقابل اندازه‌گیری) در مدل وجود دارد؟ توضیح دهید.

بعد از تبدیل لگاریتمی متغیر درآمد، مشاهده می‌شود که ضریب درآمد در رگرسیون معنادار می‌شود. بنابراین یکی از استدلال‌هایی که می‌توان در این مورد کرد این است که مدل خطی اول یک مدل اشتباه است. البته از طرفی این پرسش پیش می‌آید که مگر سیگار و به طور کلی دخانیات، مشروبات الکلی و مواد مخدر با توجه به طبیعت اعتیادآور آن‌ها نباید یک کالای کم‌کشش باشد؟ بنابراین شاید بهتر باشد که حتی متغیر وابسته‌مان در این مدل حالت لگاریتمی به خود بگیرد تا بینش بهتری نسبت به ارتباط این دو متغیر داشته باشیم. در رابطه با متغیر توضیحی مشاهده‌نشده یکی از اولین مسائلی که به ذهن انسان می‌رسد وضعیت روحی و روانی مصرف‌کننده سیگار است. شاید این متغیر بر روی میزان مصرف تاثیرگذار باشد اما پرداختن به اینکه وضعیت روانی باید چگونه اندازه‌گیری و وارد مدل شود خود یک مشکل دیگر این مدل است. برای چک کردن اینکه آیا مدل دچار مشکل OVB است یا نه می‌توان تست رمزی ریست را انجام داد.

```
Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of weight

H0: Model has no omitted variables

F(3, 1783) =   0.30
  Prob > F = 0.8248
```

آمارهٔ اف و پی–ولیوی آزمون بیانگر وجود متغیر محذوف است.

۹) از داده‌های سلامت بودجهٔ خانوار استفاده کرده و یک پراکسی مناسب برای آثار مخرب مصرف دخانیات انتخاب و به صورت لگاریتمی وارد مدل کنید. مدل نهایی را با وجود پراکسی برآورد و ضرایب و معناداری کل رگرسیون را تفسیر کنید.

تفسیر ضرایب رگرسیون سوم را با توجه به مطالب گفته شده در بخش شش و هفت می‌توان انجام داد. تنها تفاوت این مدل با مدل بخش هفت اضافه شدن پراکسی مخارج درمانی به رگرسیون است. با افزایش یک درصدی مخارج درمانی، میزان مصرف دخانیات به میزان ۰,۰۱۷٦ واحد کاش می‌یابد. البته ضریب این متغیر به لحاظ آماری معنادار نیست و همچنین آمارهٔ اف مدل نسبت به مدل بخش هفت کاهش یافته.

Source	SS	df	MS		
				Number of obs	= 1,794
				F(8, 1785)	= 16.29
Model	9569371.13	8	1196171.39	Prob > F	= 0.0000
Residual	131090175	1,785	73439.8741	R-squared	= 0.0680
				Adj R-squared	= 0.0639
Total	140659546	1,793	78449.2729	Root MSE	= 271

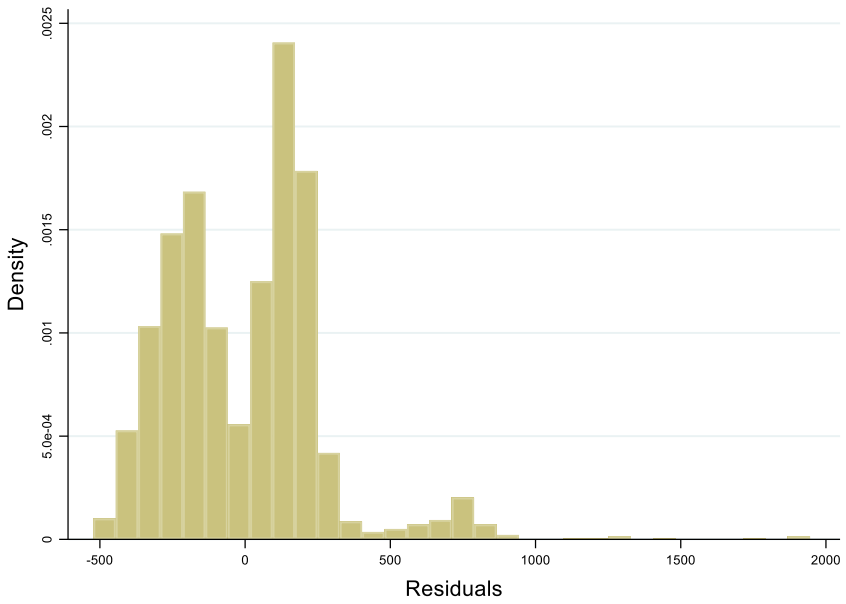
weight	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ln_income	28.90134	9.344691	3.09	0.002	10.57366	47.22903
ln_unitvalue	-130.6947	15.82341	-8.26	0.000	-161.7291	-99.66035
ln_sum_of_value	-1.766046	4.928618	-0.36	0.720	-11.43251	7.900423
age	2.295019	.8488905	2.70	0.007	.6300953	3.959943
female	-105.3917	79.11909	-1.33	0.183	-260.5675	49.78414
leduc	-18.81171	4.345237	-4.33	0.000	-27.334	-10.28942
own	4.536739	14.59804	0.31	0.756	-24.0943	33.16778
child_less_14	-12.35954	16.25296	-0.76	0.447	-44.23637	19.51729
_cons	912.7151	200.8774	4.54	0.000	518.7355	1306.695

با توجه به مدل نهایی:

۱۰) آزمون نرمال بودن جمله اخلال را انجام دهید.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
resid	1,794	0.90449	102.602	11.733	0.00000



همانطور که از نتیجهٔ تست و هیستوگرام مشخص است، توزیع جملهٔ اخلال نرمال نیست.

۱۲) آزمون معناداری مشترک را برای موارد زیر انجام داده و نتایج را تفسیر کنید.

آ. پروکسی آثار مخرب و داشتن فرزند زیر ١٤ سال

ب. درآمد و own

ج. درآمد و تحصیلات

فرضیهٔ صفر تست آ رد نمی‌شود اما فرضهٔ صفر تست ب و ج رد خواهد شد و ضرایب لگاریتم درآمد با صاحب خانه بودن و لگاریتم درآمد و داشتن تحصیلات دانشگاهی در سطح پنج درصد با هم معنادارند.

```
. test child_less_14 ln_sum_of_value

 ( 1)  child_less_14 = 0
 ( 2)  ln_sum_of_value = 0

       F(  2,  1785) =    0.45
            Prob > F =    0.6400

. test ln_income own

 ( 1)  ln_income = 0
 ( 2)  own = 0

       F(  2,  1785) =    3.84
            Prob > F =    0.0218

. test ln_income ceduc

 ( 1)  ln_income = 0
 ( 2)  ceduc = 0

       F(  2,  1785) =    8.35
            Prob > F =    0.0002
```

```
Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of weight

H0: Constant variance

    chi2(1) =  51.55
Prob > chi2 = 0.0000

. estat imtest

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	86.89	40	0.0000
Skewness	36.01	8	0.0000
Kurtosis	6.20	1	0.0128
Total	129.10	49	0.0000

۱۳) ناهمسانی واریانس رگرسیون را با ۲ روش دلخواه آزمون کنید و تفسیری مختصر از نتایج ارائه کنید.

نتیجهٔ هر دو تست انجام‌شده رد شدن فرضیهٔ صفر آزمون یا وجود واریانس ناهمسانی در رگرسیون است.

۱٤) تفاوت میزان مصرف دخانیات در گروه‌های مردان دارای تحصیلات دانشگاهی و مردان بدون تحصیلات دانشگاهی چقدر است؟ (راهنمایی: از متغیرهای موهومی کنشی استفاده کنید.)

Source	SS	df	MS		
Model	1718370.94	3	572790.313	Number of obs	= 1,794
Residual	138941175	1,790	77620.7684	F(3, 1790)	= 7.38
				Prob > F	= 0.0001
Total	140659546	1,793	78449.2729	R-squared	= 0.0122
				Adj R-squared	= 0.0106
				Root MSE	= 278.61

weight	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female						
Male	95.07927	84.28111	1.13	0.259	-70.22045	260.379
1.ceduc	-55.45455	290.9935	-0.19	0.849	-626.1772	515.2681
female#ceduc						
Male#1	-62.02372	292.1303	-0.21	0.832	-634.976	510.9286
_cons	355.4545	84.00258	4.23	0.000	190.7011	520.208

رگرسیون بالا به ما می‌گوید که مردان دارای تحصیلات دانشگاهی به طور میانگین ٦٢ کیلوگرم کمتر دخانیات مصرف می‌کنند. که البته ضریب آن به لحاظ آماری معنادار نیست.

۱۵) استدلال شما از گنجانده شدن متغیر child_less_14 در مدل چیست؟

شاید بتوان گفت افراد با توجه به آسیب‌پذیر بودن کودکان، سعی می‌کنند که اگر کودکی عضو خانواده‌شان باشد، به میزان کمتری سیگار استعمال کنند یا حتی اصلا سیگار نکشند. همهٔ ما در اطرافیانمان کسی را سراغ داریم که بعد از فرزنددار شدن ترک کرده است پس حضور این متغیر در مدل به نظر منطقی‌ست. با این حال هر سه مدل پروژه معنادار بودن این متغیر را رد می‌کنند که این مسئله از دو حالت خارج نیست، یا تئوری پشت حضور این متغیر در مدل یک تئوری نادرست است و یا مدل دارای مشکلاتی از قبیل خطای تصریح است.