

Osnove uporabe programskega paketa Stata

Delavnica SHARE 2017

prof. dr. Miroslav Verbič
www.miroslav-verbic.si

Primer: Datoteka `Podatki_SHARE.dta` vsebuje podatke iz zbirke podatkov easySHARE (različica 6.0.0) v formatu programskega paketa Stata. Gre za vzorec 300 opazovanj, ki predstavljajo posameznike iz Slovenije, zajete v šestem valu raziskave SHARE.

Podatkovna datoteka vsebuje različne spremenljivke, od katerih izpostavljam naslednje:

- ◆ *mergeid*: identifikator posameznika (država-gospodinjstvo-član);
 - ◆ *female*: spol posameznika (1 = ženski, 0 = moški);
 - ◆ *age*: starost posameznika v letih;
 - ◆ *iv009_mod*: območje prebivanja (urbanost, ruralnost);
 - ◆ *eduyears_mod*: število let šolanja posameznika;
 - ◆ *mar_stat*: zakonski stan;
 - ◆ *hhsiz*: število članov gospodinjstva;
 - ◆ *ch001_*: število otrok;
 - ◆ *sphus*: zadovoljstvo z zdravjem;
 - ◆ *casp*: indeks kakovosti življenja in blaginje CASP;
 - ◆ *bmi*: indeks telesne mase;
 - ◆ *ep005_*: kategorija delovne aktivnosti;
 - ◆ *co007_*: shajanje gospodinjstva z dohodkom;
 - ◆ *thinc_m*: skupni neto dohodki gospodinjstva v € na letni ravni;
 - ◆ *ores*: vrednost nepremičnin gospodinjstva v €
 - ◆ *fahc*: izdatki za hrano gospodinjstva (porabljeni doma) v €
 - ◆ *car*: vrednost avtomobilov v €
 - ◆ *hnfass*: neto finančna sredstva gospodinjstva v €
- a) Proučite podatke iz zbirke podatkov easySHARE s pomočjo ustreznih ukazov za pregled in obdelavo podatkov v programskem paketu Stata, še posebej za spremenljivko *fahc*. Pri diskretnih spremenljivkah po potrebi generirajte ustrezne nepravne spremenljivke.
 - b) Prikažite korelacijsko matriko spremenljivk iz zgornjega nabora. Ugotovite tudi statistično značilnost izračunanih Pearsonovih in Spearmanovih korelacijskih koeficientov.
 - c) S pomočjo *t*-preizkusa preverite, ali se aritmetična sredina indeksa kakovosti življenja in blaginje CASP razlikuje glede na spol.
 - d) Oblikujte in ocenite linearno multiplo regresijsko funkcijo, v kateri boste proučevali vpliv različnih (družbenoekonomsko utemeljenih) dejavnikov iz zgornjega nabora na izdatke za hrano gospodinjstva. Odvisnosti med odvisno in pojasnjevalnimi spremenljivkami prikažite najprej grafično. Razložite dobljene rezultate.
 - e) Izračunajte ocenjene vrednosti odvisne spremenljivke in ostanke ter jih izpišite skupaj z opazovanimi vrednostmi odvisne spremenljivke.
 - f) Ocenite še logaritemsko-logaritemsko (linearizirano potenčno) multiplo regresijsko funkcijo, v kateri boste proučevali vpliv ugotovljenih dejavnikov na izdatke za hrano gospodinjstva. Razložite dobljene rezultate. S pomočjo *F*-preizkusa preverite, ali sta učinka neto dohodkov in vrednosti nepremičnin gospodinjstva enaka po jakosti.

Izpis rezultatov obdelav v programskem paketu Stata:

Pregled in ureditev podatkov:

. describe

Contains data from Podatki_SHARE.dta

```
obs:      300          easySHARE release 6.0.0
vars:      69
size:     106,500
```

```
-----
variable name    storage   display
                 type      format   variable label
-----
mergeid          str12    %12s     Person identifier (fix across modules and waves)
wave             byte     %9.0g    Wave
country          byte     %37.0g   Country identifier
female          byte     %37.0g   Gender: female=1, male=0
age             float    %37,1f   Age at interview (in years)
iv009_mod       byte     %41.0g   Area of location
isced1997_r     byte     %37.0g   Education of respondent in ISCED-97 Coding
eduyears_mod    float    %37.0g   Years of education
mar_stat        byte     %42.0f   Marital status
hhsz            byte     %37.0g   Household size
partnerinhh     byte     %47.0g   Living with spouse/partner
ch001_          byte     %37.0f   Number of children
ch021_mod       int      %37.0f   Number of grandchildren
sphus           byte     %37.0g   Self-perceived health - us version
chronic_mod     byte     %37.0g   Number of chronic diseases (easySHARE version)
casp            byte     %55.0g   CASP: quality of life and well-being index
eurod           byte     %37.0g   Depression scale EURO-D - high is depressed
hc002_mod       byte     %37.0f   How often seen a medical doctor last 12 months
hc012_         byte     %37.0f   Stayed over night in hospital last 12 months
hc029_         byte     %37.0f   In a nursing home during last 12 months
maxgrip         byte     %24.0g   Max. of grip strength measure
adla            byte     %37.0g   Activities of daily living index (high: diffic.)
bmi             float    %37.0g   Body mass index
bmi2            byte     %37.0g   Body mass index categories
ep005_         byte     %65.0g   Current job situation
ep009_mod       byte     %37.0f   Employee or a self-employed in (main) job
ep011_mod       byte     %41.0f   Term of (main) job
ep013_mod       double   %37.1f   Total hours worked per week (main) job
ep026_mod       byte     %37.0f   Satisfied with (main) job
ep036_mod       byte     %37.0f   Look for early retirement in (main) job
co007_         byte     %37.0f   Is household able to make ends meet
thinc_m        float    %9.0g    Household net income, imputed
inpat6          double   %9.0g    Paid out-of-pocket for inpatient care
outpa6          double   %9.0g    Paid out-of-pocket for outpatient care
drugs6          double   %9.0g    Paid out-of-pocket for prescribed drugs
nurs6           double   %9.0g    Paid out-of-pocket for nursing home/home care
ydip            double   %9.0g    Earnings from employment
yind            double   %9.0g    Earnings from self-employment
ypen1           double   %9.0g    Old age, early retirement, and survivor pensions
ypen2           double   %9.0g    Private and occupational pensions
ypen3           double   %9.0g    Disability pensions/benefits
ypen4           double   %9.0g    Unemployment benefits/insurances
ypen5           double   %9.0g    Social assistance
ypen6           double   %9.0g    Sickness benefits
home            double   %9.0g    Value of main residence
mort            double   %9.0g    Mortgage on main residence
rhre            double   %9.0g    Rent and home-related expenditures
ores            double   %9.0g    Value of real estate - Amount
ysrent          double   %9.0g    Income from rent/sublet
yaohm           double   %9.0g    Income from other household members
fahc            double   %9.0g    Amount spent on food at home
fohc            double   %9.0g    Amount spent on food outside home
hprf            double   %9.0g    Value of home produced food
-----
```

```

hmem          double %9.0g    Paid out-of-pocket for dental care - Amount
bacc          double %9.0g    Bank accounts
bsmf         double %9.0g    Bond, stock and mutual funds
ybbsmf       double %9.0g    Interest/dividend from bank account, bond, funds
slti         double %9.0g    Savings for long-term investments
vbus         double %9.0g    Value of own business
sbus         float  %9.0g    Share of own business
car          double %9.0g    Value of cars
liab         double %9.0g    Financial liabilities
thinc2       double %9.0g    Total household income - Version B
thinc        double %9.0g    Total household income - Version A
hnetw        double %9.0g    Household net worth
yincnrp      double %9.0g    Income from nonresponding partner
hrass        double %9.0g    Household real assets
hgfass       double %9.0g    Household gross financial assets
hnfass       double %9.0g    Household net financial assets

```

Sorted by:

. tab female

Gender: female=1, male=0	Freq.	Percent	Cum.
0. male	124	41.33	41.33
1. female	176	58.67	100.00
Total	300	100.00	

. gen spol=0

. replace spol=1 if female==0

(124 real changes made)

. tab iv009_mod, gen(urb)

Area of location	Freq.	Percent	Cum.
-15. no information	4	1.33	1.33
-9. filtered: interview not at home	8	2.67	4.00
1. A big city	28	9.33	13.33
2. The suburbs or outskirts of a big ci	18	6.00	19.33
3. A large town	17	5.67	25.00
4. A small town	67	22.33	47.33
5. A rural area or village	158	52.67	100.00
Total	300	100.00	

. tab mar_stat, gen(ms)

Marital status	Freq.	Percent	Cum.
1. Married and living together with spo	217	72.33	72.33
2. Registered partnership	17	5.67	78.00
3. Married, living separated from spous	1	0.33	78.33
4. Never married	11	3.67	82.00
5. Divorced	15	5.00	87.00
6. Widowed	39	13.00	100.00
Total	300	100.00	

. tab sphus, gen(zz)

Self-perceived health - us version	Freq.	Percent	Cum.
1. Excellent	21	7.00	7.00
2. Very good	36	12.00	19.00
3. Good	121	40.33	59.33

4. Fair	82	27.33	86.67
5. Poor	40	13.33	100.00
Total	300	100.00	

. tab ep005_, gen(dakt)

Current job situation	Freq.	Percent	Cum.
-15. no information	1	0.33	0.33
1. retired	224	74.67	75.00
2. employed or self-employed	39	13.00	88.00
3. unemployed	9	3.00	91.00
4. permanently sick or disabled	5	1.67	92.67
5. homemaker	18	6.00	98.67
97. other	4	1.33	100.00
Total	300	100.00	

. tab co007_, gen(sdoh)

Is household able to make ends meet	Freq.	Percent	Cum.
-15. no information	2	0.67	0.67
1. With great difficulty	54	18.00	18.67
2. With some difficulty	115	38.33	57.00
3. Fairly easily	72	24.00	81.00
4. Easily	57	19.00	100.00
Total	300	100.00	

. sum female age edueyears_mod hhszise ch001_ sphus casp bmi thinc_m ores fahc car hnfass

Variable	Obs	Mean	Std. Dev.	Min	Max
female	300	.5866667	.4932544	0	1
age	300	67.944	9.275548	48.1	94.4
edueyears_mod	300	10	5.198405	-15	23
hhszise	300	2.28	.9650758	1	6
ch001_	300	2.136667	1.108491	0	6
sphus	300	3.28	1.063974	1	5
casp	300	34.45667	14.77665	-15	48
bmi	300	26.51387	8.403939	-12	48.88889
thinc_m	300	19096.2	18355.04	669.9236	141360.5
ores	300	20781.04	67807.35	0	500000
fahc	300	3543.656	1641.078	0	8400
car	300	3872.551	4617.122	0	21973.73
hnfass	300	3900.898	9299.317	-16000	63962.95

. tab edueyears_mod

Years of education	Freq.	Percent	Cum.
-15. no information	7	2.33	2.33
0	1	0.33	2.67
2	1	0.33	3.00
3	5	1.67	4.67
4	10	3.33	8.00
5	5	1.67	9.67
6	4	1.33	11.00
7	8	2.67	13.67
8	70	23.33	37.00
9	7	2.33	39.33
10	13	4.33	43.67
11	54	18.00	61.67
12	54	18.00	79.67

13	11	3.67	83.33
14	11	3.67	87.00
15	8	2.67	89.67
16	13	4.33	94.00
17	9	3.00	97.00
18	6	2.00	99.00
19	1	0.33	99.33
20	1	0.33	99.67
23	1	0.33	100.00

Total	300	100.00	

. replace edueyears_mod=. if edueyears_mod==--15
(7 real changes made, 7 to missing)

. replace casp=. if casp==--15
(21 real changes made, 21 to missing)

. replace bmi=. if bmi==--12
(10 real changes made, 10 to missing)

. sum fahc, detail

Amount spent on food at home

Percentiles		Smallest		
1%	720	0		
5%	1200	720		
10%	1800	720	Obs	300
25%	2400	720	Sum of Wgt.	300

50%	3600		Mean	3543.656
		Largest	Std. Dev.	1641.078
75%	4800	8400		
90%	6000	8400	Variance	2693137
95%	6049.852	8400	Skewness	.6922547
99%	8400	8400	Kurtosis	3.122141

. inspect fahc bmi

fahc: Amount spent on food at home

					Number of Observations				
					Total	Integers	Nonintegers		
		#		Negative	-	-	-		
		#	#	Zero	1	1	-		
		#	#	Positive	299	19	280		
		#	#		-----	-----	-----		
		#	#	Total	300	20	280		
	#	#	#	Missing	-				

+									
0			8400		300				
			(39 unique values)						

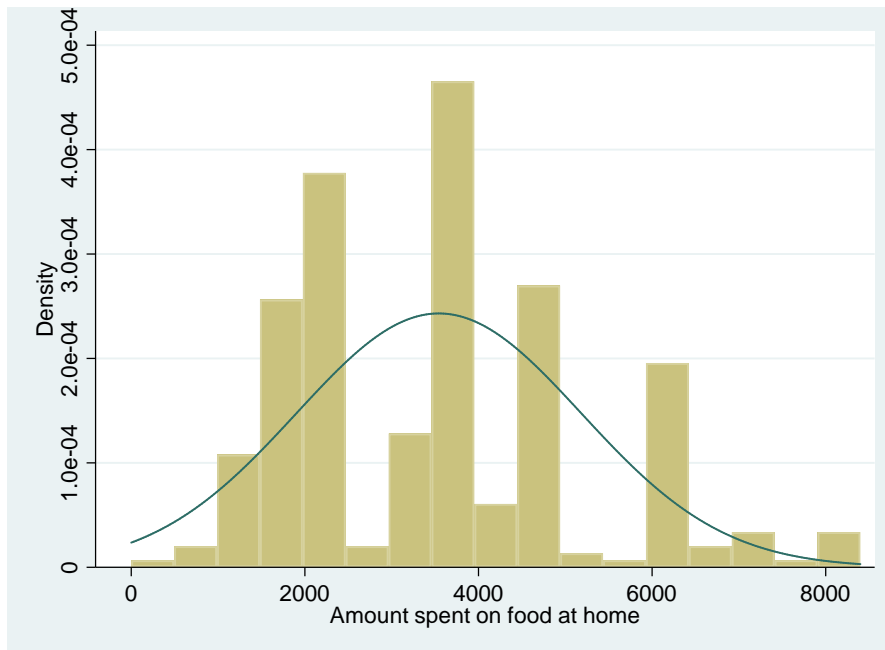
bmi: Body mass index

					Number of Observations				
					Total	Integers	Nonintegers		
		#		Negative	-	-	-		
		#		Zero	-	-	-		
		#		Positive	290	1	289		
		#			-----	-----	-----		
	#	#	#	Total	290	1	289		
	#	#	#	Missing	10				

+									
18.08021			48.88889		300				
			(More than 99 unique values)						

bmi is labeled but 290 values are NOT documented in the label.

```
. hist fahc, normal
(bin=17, start=0, width=494.11765)
```



```
. list mergeid spol age eduyea~d hhszize thinc_m fahc if dakt3==1
```

	mergeid	spol	age	eduyea~d	hhszize	thinc_m	fahc
2.	SI-001361-02	1	56,0	15	2	23781.36	4200
7.	SI-002427-01	0	54,3	12	2	36613.89	4800
13.	SI-005225-01	0	62,0	16	1	15116.28	1800
19.	SI-006643-03	0	56,5	.	5	20348.84	3600
26.	SI-009396-01	1	62,4	11	2	29079.27	4800
28.	SI-010100-01	1	58,0	12	2	27961.89	3600
29.	SI-010100-03	0	53,3	15	2	27961.89	3600
52.	SI-016144-01	1	57,8	8	3	17441.86	4800
68.	SI-018868-01	1	56,8	11	4	21867.71	3600
71.	SI-020327-01	0	56,0	14	1	5813.954	3600
81.	SI-022776-02	0	49,9	11	2	13372.09	2400
87.	SI-023460-01	1	52,9	14	4	1162.791	3600
89.	SI-023530-01	1	61,8	16	1	6193.654	2058.289
92.	SI-023630-01	0	53,3	10	2	8723.188	2400
...							
231.	SI-055711-02	0	67,0	16	3	41860.46	6000
234.	SI-056404-01	0	56,2	13	4	29883.72	3600
235.	SI-056404-02	1	55,2	11	4	29883.72	3600
236.	SI-056467-01	1	54,8	6	2	21162.79	2400
237.	SI-056467-02	0	53,1	11	2	21162.79	2400
263.	SI-062083-01	1	56,8	8	2	18604.65	4800
274.	SI-063660-01	0	58,3	17	4	88997.52	2709.197
277.	SI-066437-01	0	54,3	17	2	2773.216	2400
279.	SI-066644-01	1	60,7	12	2	6408.271	1440
286.	SI-067556-01	0	59,9	23	1	17209.3	4800
287.	SI-067741-01	1	54,6	11	4	21860.64	1904.012
288.	SI-067834-01	0	57,0	17	3	2330.208	6600

Korelacijska matrika:

```
. correlate age eduyea~d hhsizе thinc_m fahc
(obs=293)
```

	age	eduyea~d	hhsizе	thinc_m	fahc
age	1.0000				
eduyea~d	-0.1733	1.0000			
hhsizе	-0.2881	-0.0010	1.0000		
thinc_m	-0.1162	0.2272	0.0758	1.0000	
fahc	-0.0622	0.2753	0.2722	0.2845	1.0000

```
. pcorr age eduyea~d hhsizе thinc_m fahc, sig
```

	age	eduyea~d	hhsizе	thinc_m	fahc
age	1.0000				
eduyea~d	-0.1733	1.0000			
	0.0029				
hhsizе	-0.2993	-0.0010	1.0000		
	0.0000	0.9865			
thinc_m	-0.1220	0.2272	0.0650	1.0000	
	0.0346	0.0001	0.2616		
fahc	-0.0566	0.2753	0.2638	0.2662	1.0000
	0.3289	0.0000	0.0000	0.0000	

```
. spearman hhsizе ch001_ sphus co007_
(obs=300)
```

	hhsizе	ch001_	sphus	co007_
hhsizе	1.0000			
ch001_	0.1934	1.0000		
sphus	-0.0379	0.1155	1.0000	
co007_	-0.0192	-0.0975	-0.3090	1.0000

```
. spearman hhsizе ch001_ sphus co007_, stats(rho p)
(obs=300)
```

Key
rho
Sig. level

	hhsizе	ch001_	sphus	co007_
hhsizе	1.0000			
ch001_	0.1934	1.0000		
	0.0008			
sphus	-0.0379	0.1155	1.0000	
	0.5137	0.0457		
co007_	-0.0192	-0.0975	-0.3090	1.0000
	0.7401	0.0918	0.0000	

Preverjanje enakosti aritmetičnih sredin na dveh pod vzorcih:

`. ttest casp, by(spol)`

Two-sample t test with equal variances

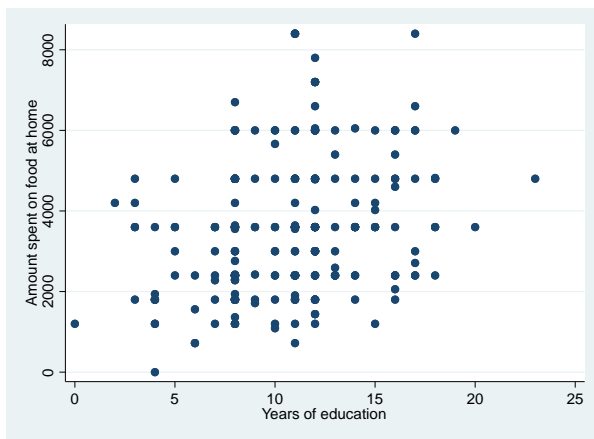
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	167	37.98204	.4881152	6.307839	37.01832	38.94575
1	112	38.47321	.5255627	5.562033	37.43178	39.51465
combined	279	38.17921	.3600538	6.014084	37.47043	38.88799
diff		-.4911784	.7352538		-1.938573	.9562166

diff = mean(0) - mean(1) t = -0.6680
 Ho: diff = 0 degrees of freedom = 277

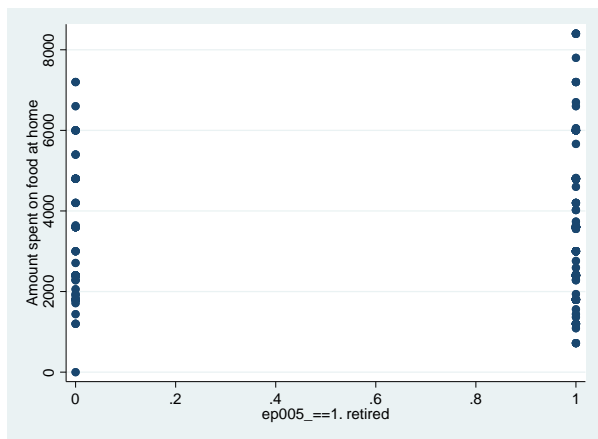
Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
 Pr(T < t) = 0.2523 Pr(|T| > |t|) = 0.5047 Pr(T > t) = 0.7477

Prikaz odvisnosti v razsevnem diagramu in ocenjevanje linearnega modela:

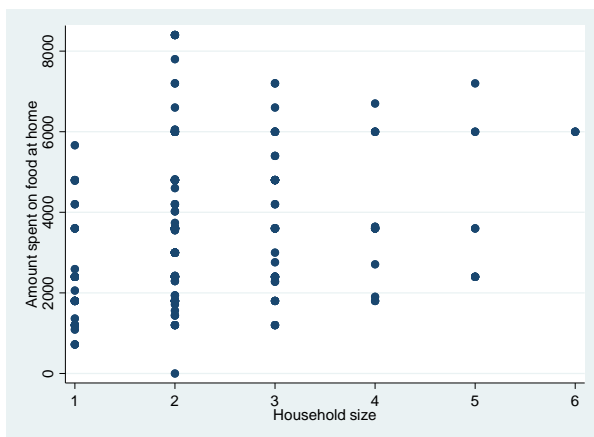
`. scatter fahc eduyears_mod`



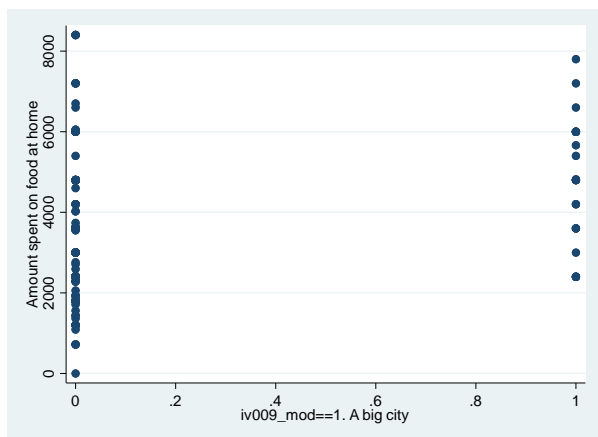
`. scatter fahc dakt2`



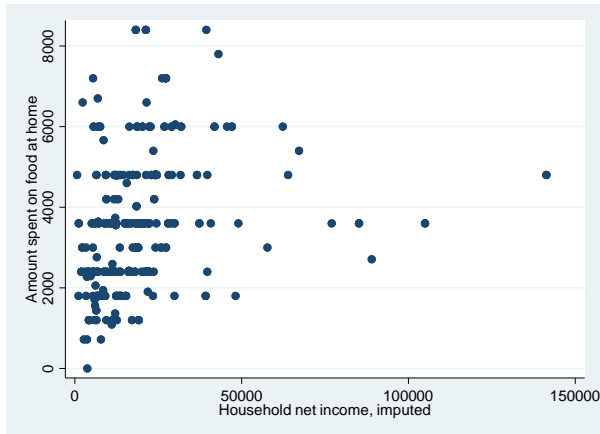
`. scatter fahc hhszise`



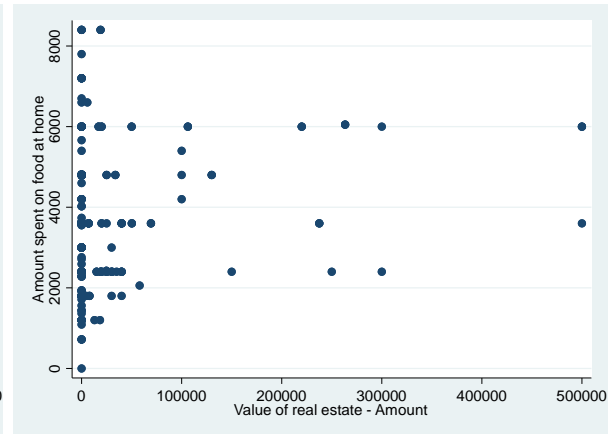
`. scatter fahc urb3`



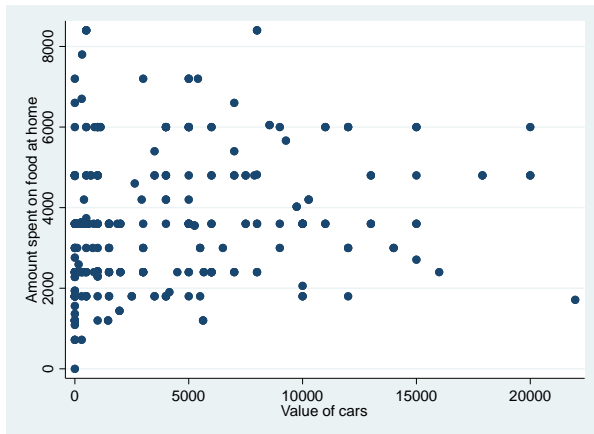

```
. scatter fahc thinc_m
```



```
. scatter fahc ores
```



```
. scatter fahc car
```



```
. regress fahc edueyears_mod dakt2 hhsiz e urb3 thinc_m ores car
```

Source	SS	df	MS	Number of obs	=	293
Model	216705116	7	30957873.7	F(7, 285)	=	15.09
Residual	584865698	285	2052160.34	Prob > F	=	0.0000
Total	801570814	292	2745105.53	R-squared	=	0.2704
				Adj R-squared	=	0.2524
				Root MSE	=	1432.5

	fahc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
edueyears_mod		68.38859	25.80981	2.65	0.009	17.58655 119.1906
dakt2		394.9844	198.8579	1.99	0.048	3.567916 786.4009
hhsiz e		514.3805	90.23081	5.70	0.000	336.7772 691.9839
urb3		1203.913	292.3475	4.12	0.000	628.4791 1779.348
thinc_m		.0174705	.0050296	3.47	0.001	.0075706 .0273704
ores		.002435	.0012669	1.92	0.056	-.0000586 .0049287
car		.0403618	.019394	2.08	0.038	.0021882 .0785354
_cons		689.7797	380.3994	1.81	0.071	-58.96911 1438.528

Prikaz ocenjenih vrednosti in ostankov:

```
. qui regress fahc edueyears_mod dakt2 hhsiz e urb3 thinc_m ores car
```

```
. predict e, resid
```

(7 missing values generated)

```

. predict fahc_hat, xb
(7 missing values generated)

. list fahc fahc_hat e if fahc_hat!=. & e!=., mean

```

	fahc	fahc_hat	e
1.	4200	3900.802	299.1984
2.	4200	3574.206	625.7942
3.	1800	2548.558	-748.5578
4.	1089.749	2476.018	-1386.269
5.	1800	3297.425	-1497.425
6.	1800	3092.259	-1292.259
7.	4800	3983.448	816.5518
8.	4800	4378.433	421.5674
9.	2400	3619.107	-1219.108
10.	2400	3550.719	-1150.719
11.	0	2057.34	-2057.34
12.	1712.354	3324.83	-1612.475
13.	1800	2784.456	-984.4559
...			
295.	1200	2584.355	-1384.355
296.	2400	4888.494	-2488.494
297.	4800	3819.498	980.5018
298.	4800	3365.784	1434.216
299.	4800	3365.784	1434.216
300.	3000	3269.574	-269.5744
Mean	3531.604	3531.604	-1.92e-06

Ocenjevanje logaritemsko-logaritemskega modela:

```

. gen lfahc=log(fahc)
(1 missing value generated)

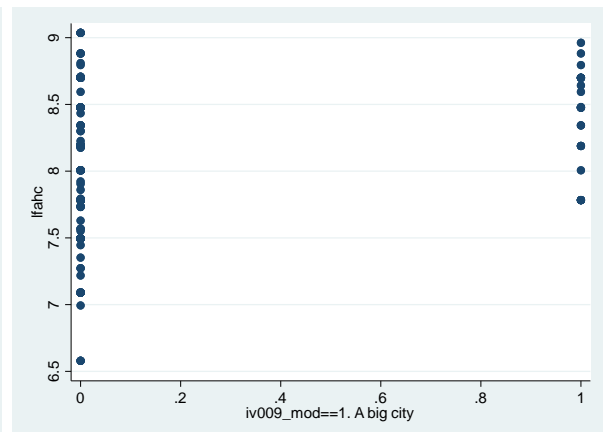
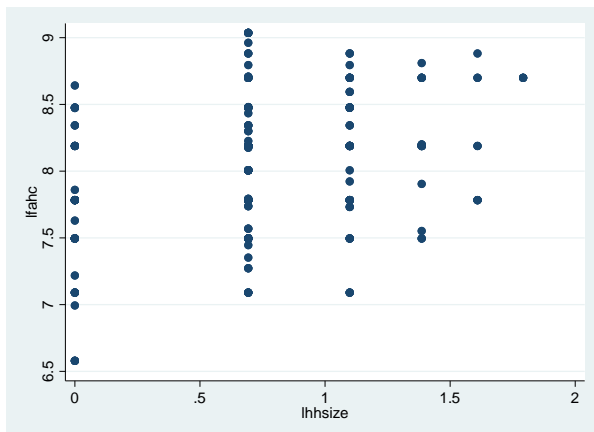
. gen lhhsz=log(hhsz)

. gen lthinc_m=log(thinc_m)

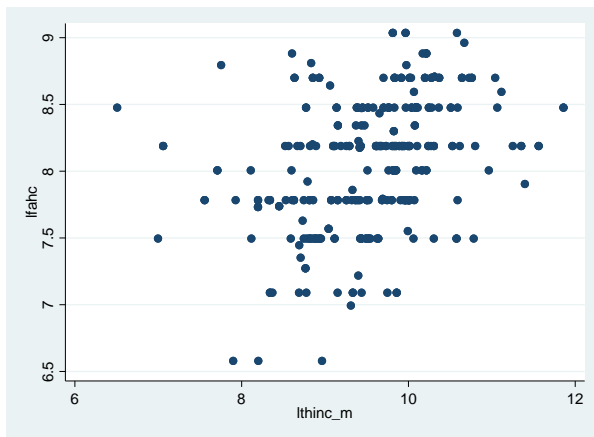
. gen lores=log(ores)
(223 missing values generated)

. scatter lfahc lhhsz
. scatter lfahc lurb3

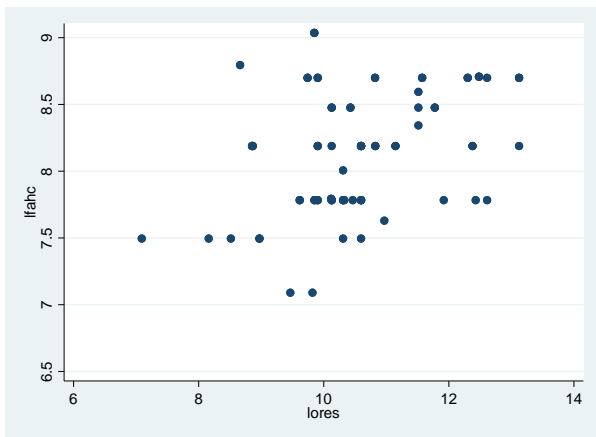
```



```
. scatter lfahc lthinc_m
```



```
. scatter lfahc lores
```



```
. regress lfahc lhsize urb3 lthinc_m lores
```

Source	SS	df	MS	Number of obs	=	77
Model	7.10437973	4	1.77609493	F(4, 72)	=	14.57
Residual	8.7766655	72	.121898132	Prob > F	=	0.0000
Total	15.8810452	76	.208961121	R-squared	=	0.4473
				Adj R-squared	=	0.4166
				Root MSE	=	.34914

lfahc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lhsize	.4640447	.0950227	4.88	0.000	.2746205 .653469
urb3	.2566703	.1441502	1.78	0.079	-.0306878 .5440283
lthinc_m	.1209255	.0492399	2.46	0.016	.0227675 .2190834
lores	.1099059	.033519	3.28	0.002	.043087 .1767248
_cons	5.419108	.5336027	10.16	0.000	4.35539 6.482825

```
. test lthinc_m=lores
```

```
( 1) lthinc_m - lores = 0
```

```
F( 1, 72) = 0.03
Prob > F = 0.8645
```

