

Computer Exercise 4 (Answers)

Read in the data and generate the log of household expenditure

```
g logexpeq=log(expeq)
```

1. The simple regression OLS estimate of the Engel curve gives

```
reg foodsh2 logexpeq
```

Source	SS	df	MS			
Model	116951.426	1	116951.426	Number of obs =	2804	
Residual	237437.174	2802	84.7384631	F(1, 2802) =	1380.15	
Total	354388.6	2803	126.431894	Prob > F =	0.0000	
				R-squared =	0.3300	
				Adj R-squared =	0.3298	
				Root MSE =	9.2053	

foodsh2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
logexpeq	-9.208081	.2478601	-37.15	0.000	-9.694088	-8.722074
_cons	71.3734	1.315884	54.24	0.000	68.7932	73.9536

The coefficient on log expenditure is a semi-elasticity and gives the percentage point change in the budget share of each item following a 1% change in total household expenditure, multiplied by 100, (since $dw_i/d\text{Log}(x) = b_i = dw_i/(dx/x) = \text{unit change in } w \text{ with respect to a 1 percentage change in } x * 100$)

So a 1% increase in expenditure is associated with a 0.09 percentage point fall (-9.2/100) in the share of the household budget spent on food. The negative sign confirms that food is a necessity (expenditure share falls as income rises)

At face value, the coefficients do not look particularly amiss

2. It is important however to look at the data **before** doing any regression

```
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
edage	2804	17.08631	3.202251	13	99
age	2804	49.73787	15.12236	18	79
foodsh2	2804	22.91637	11.24419	0	73.35155
london	2804	.0937946	.2915947	0	1
female	2804	.3598431	.4800398	0	1
employed	2804	.6073466	.4884279	0	1
expeq	2804	426.5682	5779.861	25.65605	180068.7

The mean age is 49.7, mean weekly household expenditure is £426 and mean food budget share is 22.9 (ie 22.9% of total expenditure goes on food)

36% of the sample of household heads are female, 9.4% live in London and 60.7% are employed (the mean of a dummy variable gives the sample proportion. If there are N_1

coded 1 and N_0 coded zero then $\frac{\sum_{i=1}^{N_1} 1 + \sum_{i=N_1+1}^{N_0} 0}{N} = \frac{N_1}{N}$)

Should be able to spot that two variables give cause for concern. The maximum of edage is 99. A cross-tabulation of the variable suggest that this observation is wrongly coded (particularly as the maximum age of the sample is 79)

. tab edage

age completed continuous full time education	Freq.	Percent	Cum.
13	16	0.57	0.57
14	224	7.99	8.56
15	627	22.36	30.92
16	868	30.96	61.88
17	240	8.56	70.44
18	258	9.20	79.64
19	83	2.96	82.60
20	60	2.14	84.74
21	147	5.24	89.98
22	112	3.99	93.97
23	58	2.07	96.04
24	52	1.85	97.90
25	22	0.78	98.68
26	15	0.53	99.22
27	4	0.14	99.36
28	9	0.32	99.68
29	4	0.14	99.82
30	1	0.04	99.86
31	2	0.07	99.93
32	1	0.04	99.96
99	1	0.04	100.00
Total	2,804	100.00	

Similarly the mean of household expenditure is quite high and the maximum value is very high

```
. su expeq, detail
      total household expenditure (equivalised) ]
-----
```

Percentiles		Smallest		
1%	36.76984	25.65605		
5%	60.01889	25.88354		
10%	78.46063	26.20582	Obs	2804
25%	124.901	27.18727	Sum of Wgt.	2804
50%	197.763		Mean	426.5682
		Largest	Std. Dev.	5779.861
75%	298.614	1672.697		
90%	432.3807	173720.4	Variance	3.34e+07
95%	548.7453	176933.1	Skewness	30.49271
99%	827.2877	180068.7	Kurtosis	931.8264

In fact the 3 highest values lie a long way from the main body of data (by a factor of 100). This suggests that they have been wrongly coded (probably using pence instead of pounds)

Also the food share value of 0 is rather implausible, since it means the household spent nothing on food

So you should drop the following

```
. drop if edage==99
(1 observation deleted)

. drop if expeq>173720
(3 observation deleted)

. drop if foodsh==0
(4 observations deleted)
```

This leaves a working sample of 2796

```
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
edage	2796	17.05222	2.803548	13	32
age	2796	49.73247	15.1205	18	79
foodsh2	2796	22.9717	11.20698	.0392941	73.35155
london	2796	.0926323	.2899684	0	1
female	2796	.3594421	.4799227	0	1
employed	2796	.6076538	.4883605	0	1
expeq	2796	237.2969	170.8488	25.65605	1672.697

3. A repeat of the original regression on the new sample gives

```
reg foodsh2 logexpeq
```

Source	SS	df	MS			
Model	121797.548	1	121797.548	Number of obs =	2796	
Residual	229244.327	2794	82.0487928	F(1, 2794) =	1484.45	
				Prob > F =	0.0000	
				R-squared =	0.3470	
				Adj R-squared =	0.3467	
				Root MSE =	9.0581	
foodsh2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
logexpeq	-9.93004	.2577317	-38.53	0.000	-10.4354	-9.424676
_cons	75.15202	1.365118	55.05	0.000	72.47528	77.82876

So that the estimate budget share effect has become somewhat larger. Now a 1% increase in expenditure is associated with a 0.1 percentage point fall (-9.9/100)

4. Adding age to the model gives

```
reg foodsh2 logexpeq age
```

Source	SS	df	MS			
Model	123104.572	2	61552.2858	Number of obs =	2796	
Residual	227937.304	2793	81.6102054	F(2, 2793) =	754.22	
				Prob > F =	0.0000	
				R-squared =	0.3507	
				Adj R-squared =	0.3502	
				Root MSE =	9.0338	
foodsh2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
logexpeq	-9.779796	.2597691	-37.65	0.000	-10.28916	-9.270437
age	.0457055	.0114209	4.00	0.000	.0233113	.0680996
_cons	72.08948	1.5618	46.16	0.000	69.02708	75.15188

This reduces the estimated expenditure effect a little

This is not a coincidence. Older households tend to spend a greater proportion of their incomes on food but total expenditure (income) and age are negatively correlated – pensioners incomes are lower So part of the expenditure effect observed earlier is an age-related effect.

$$\hat{\beta}_{\log \exp}^{2 \text{ var}} = \hat{\beta}_{\log \exp}^{3 \text{ var}} + \beta_{age} \frac{\text{Cov}(\text{Log exp}, \text{Age})}{\text{Var}(\text{Log exp})}$$

Since the age coefficient is positive and the variance of log expenditure (or any variable) is always positive, then you can conclude that the estimated coefficient on expenditure in the

2 variable model will be larger (in absolute value) than the “true” coefficient in the 3 variable model if $\text{Cov}(\text{Log exp}, \text{Age}) < 0$

You can check the covariance value using the command

```
. corr foodsh2 logexpeq age
(obs=2796)
```

	foodsh2	logexpeq	age
foodsh2	1.0000		
logexpeq	-0.5890	1.0000	
age	0.1455	-0.1445	1.0000

which is indeed negative

Note that the standard error on female is slightly higher in the multiple compared to the simple regression. Specification bias says that the standard error will be biased in omitted model (biased downward **only** if σ^2 used to calculate the true variance is known and the variables are not orthogonal, as shown by the non-zero covariance term above – see lecture notes -) Since σ^2 is unknown then the upward bias to the standard errors cause by omitting variables is greater than offset the bias induced by an incorrect estimate of σ^2)

5. Adding all the other variables to the model gives

```
reg foodsh2 logexpeq age edage female london employed
```

Source	SS	df	MS			
Model	124378.757	6	20729.7929	Number of obs =	2796	
Residual	226663.118	2789	81.2703901	F(6, 2789) =	255.07	
Total	351041.875	2795	125.596378	Prob > F =	0.0000	
				R-squared =	0.3543	
				Adj R-squared =	0.3529	
				Root MSE =	9.015	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
logexpeq	-9.898546	.2881178	-34.36	0.000	-10.46349	-9.333601
age	.0586426	.0136046	4.31	0.000	.0319665	.0853186
edage	.062799	.0672754	0.93	0.351	-.0691157	.1947136
female	.3343683	.3651811	0.92	0.360	-.3816842	1.050421
london	2.028052	.5972053	3.40	0.001	.857043	3.199061
employed	.4925856	.440643	1.12	0.264	-.3714337	1.356605
_cons	70.39186	1.888329	37.28	0.000	66.68919	74.09452

To test the hypothesis that *all* the right hand side variables are jointly significant, use the variant of the F test which reduces to

$$\text{Using } F = (R^2 / k - 1) / ((1 - R^2) / (N - k)) \sim F(k - 1, N - k) \quad k \text{ includes the constant}$$

$$= (0.3543 / 7 - 1) / ((1 - 0.3543) / (2796 - 7)) \sim F(7 - 1, 2796 - 7)$$

which can be done automatically in stata using the command

```
test logexpeq age edage female london employed
```

```
( 1) logexpeq = 0
( 2) age = 0
( 3) edage = 0
( 4) female = 0
( 5) london = 0
( 6) employed = 0
```

```
F( 6, 2789) = 255.07
Prob > F = 0.0000
```

and this is indeed the value given in the top right hand corner of the stata output

Since the estimated F is greater than the critical value (not given in the output but the p value is). From Tables the 5% critical value is 2.09. Conclude model does have significant joint explanatory power.

6. The Ramsey RESET test is done in stata using the command

```
predict yhat
(option xb assumed; fitted values)
```

```
g yhat2=yhat^2
g yhat3=yhat^3
```

```
reg foodsh2 logexpeq age edage female london employed yhat2 yhat3
```

Source	SS	df	MS	Number of obs = 2796		
Model	125961.667	8	15745.2084	F(8, 2787) = 194.96		
Residual	225080.208	2787	80.7607493	Prob > F = 0.0000		
-----				R-squared = 0.3588		
Total	351041.875	2795	125.596378	Adj R-squared = 0.3570		
-----				Root MSE = 8.9867		
foodsh2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
logexpeq	4.34954	3.842793	1.13	0.258	-3.185469	11.88455
age	-.0249538	.0265719	-0.94	0.348	-.0770565	.0271488
edage	-.0143773	.0710222	-0.20	0.840	-.1536388	.1248842
female	-.1646363	.3885975	-0.42	0.672	-.9266043	.5973317
london	-.8299194	.9919055	-0.84	0.403	-2.774863	1.115024
employed	-.1142501	.4644314	-0.25	0.806	-1.024914	.7964142
yhat2	.070773	.0173355	4.08	0.000	.0367813	.1047647
yhat3	-.0010501	.0002442	-4.30	0.000	-.001529	-.0005712
_cons	-22.67602	24.71877	-0.92	0.359	-71.14498	25.79293

```
. test yhat2 yhat3
```

```
( 1) yhat2 = 0
( 2) yhat3 = 0
F( 2, 2787) = 9.80
Prob > F = 0.0001
```

The RESET test is done in stata using the command

```
. ovtest
```

```
Ramsey RESET test using powers of the fitted values of foodsh2  
Ho: model has no omitted variables  
F(3, 2784) = 3.95  
Prob > F = 0.0080
```

In both cases the estimated F value is greater than the relevant critical values at the 5% level, (2.60), so reject null of no omitted (higher order) variables.

Remember using 3 powers (implicit in the degrees of freedom of the numerator) is arbitrary. The test result may vary depending on the number of powers included.

7. The test of joint significance of these extra variables is given by

```
. test edage female employed  
  
( 1) edage = 0  
( 2) female = 0  
( 3) employed = 0  
  
F( 3, 2789) = 0.90  
Prob > F = 0.4408
```

So the estimated F is below the critical value so accept null that these variables do not contribute to the model – so might consider dropping them.

The test of equality of age and education effects is given by

```
test age=edage  
  
( 1) age - edage = 0  
  
F( 1, 2789) = 0.00  
Prob > F = 0.9490
```

Note that this is a form of the F test given by

$$(\beta_i - \beta_j)^2 / \text{Var}(\hat{\beta}_i - \hat{\beta}_j) \sim F(1, N-k)$$

which can check from the variance/covariance matrix of the OLS estimates given by

```

. matrix list e(V)

matrix list e(V)

symmetric e(V)[7,7]
      logexpeq      age      edage      female      london      employed
_cons
logexpeq  .08301184
      age  -.00042045  .00018508
      edage -.00415854  .00024098  .00452598
      female .00907849  .0005872  .00013008  .13335724
      london .0080197  6.233e-06  -.00660455  -.00570793  .35665418
employed -.04398337  .00282222  -.00049277  .02623468  .00494676  .19416623
      _cons -.32166729  -.01303118  -.06644586  -.14247377  .03617833  -.02870365
3.5657861

```

Can check this is correct since the square root of the i^{th} element on the main diagonal should equal the standard error on the i^{th} variable in the regression and the relevant covariance green is highlighted (in green)

$$\text{Var}(\hat{\beta}_i - \hat{\beta}_j) = .00018 + .00456 - 2(.0024)$$

To test for effect of log expenditure being - 10

```

test logexp=-10

( 1)  logexpeq = -10

      F( 1, 2789) =      0.12
      Prob > F =      0.7248

```

Estimated F below critical value so can't reject the null

8. London is a dummy variable living in London raises the budget share on food by 2.02 percentage points, other things equal. Food is either more expensive in London or eating out is.