Chapter 4: Transformations of variables

Overview

This chapter shows how least squares regression analysis can be extended to fit nonlinear models. Sometimes an apparently nonlinear model can be linearised by taking logarithms. $Y = \beta_1 X^{\beta_2}$ and $Y = \beta_1 e^{\beta_2 X}$ are examples. Because they can be fitted using linear regression analysis, they have proved very popular in the literature, there usually being little to be gained from using more sophisticated specifications. If you plot earnings on schooling, using the *EAEF* data set, or expenditure on a given category of expenditure on total household expenditure, using the *CES* data set, you will see that there is so much randomness in the data that one nonlinear specification is likely to be just as good as another, and indeed a linear specification may not be obviously inferior. Often the real reason for preferring a nonlinear specification to a linear one is that it makes more sense theoretically. The chapter shows how the least squares principle can be applied when the model cannot be linearised.

Learning outcomes

After working through the corresponding chapter in the textbook, studying the corresponding slideshows, and doing the starred exercises in the text and the additional exercises in this guide, you should be able to:

- explain the difference between nonlinearity in parameters and nonlinearity in variables
- explain why nonlinearity in parameters is potentially a problem while nonlinearity in variables is not
- define an elasticity
- explain how to interpret an elasticity in simple terms
- perform basic manipulations with logarithms
- interpret the coefficients of semi-logarithmic and logarithmic regressions
- explain why the coefficients of semi-logarithmic and logarithmic regressions should not be interpreted using the method for regressions in natural units described in Chapter 1
- perform a RESET test of functional misspecification
- explain the role of the disturbance term in a nonlinear model
- explain how in principle a nonlinear model that cannot be linearised may be fitted
- perform a transformation for comparing the fits of models with linear and logarithmic dependent variables.

Further material

Box–Cox tests of functional specification

This section provides the theory behind the procedure for discriminating between a linear and a logarithmic specification of the dependent variable described in Section 4.5 of the textbook. It should be skipped on first reading because it makes use of material on maximum likelihood estimation. To keep the mathematics uncluttered, the theory will be described in the context of the simple regression model, where we are choosing between

$$Y = \beta_1 + \beta_2 X + u$$

and

$$\log Y = \beta_1 + \beta_2 X + u\ .$$

It generalises with no substantive changes to the multiple regression model.

The two models are actually special cases of the more general model

$$Y_\lambda = \frac{Y^\lambda - 1}{\lambda} = \beta_1 + \beta_2 X + u$$

with $\lambda = 1$ yielding the linear model (with an unimportant adjustment to the intercept) and $\lambda = 0$ yielding the logarithmic specification at the limit as λ tends to zero. Assuming that u is iid (independently and identically distributed) N(0, σ^2), the density function for u_i is

$$f(u_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}u_i^2}$$

and hence the density function for $Y_{\lambda i}$ is

$$f(Y_{\lambda i}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2}\ .$$

From this we obtain the density function for Y_i

$$f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2} \left|\frac{\partial Y_{\lambda i}}{\partial Y_i}\right| = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2} Y_i^{\lambda - 1}\ .$$

The factor $\left|\frac{\partial Y_{\lambda i}}{\partial Y_i}\right|$ is the Jacobian for relating the density function of $Y_{\lambda i}$ to that of Y_i. Hence the likelihood function for the parameters is

$$L(\beta_1, \beta_2, \sigma, \lambda) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^n e^{-\frac{1}{2\sigma^2}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2} \prod_{i=1}^n Y_i^{\lambda - 1}$$

and the log-likelihood is

$$\begin{aligned}\log L(\beta_1, \beta_2, \sigma, \lambda) &= -\frac{n}{2}\log 2\pi\sigma^2 - \sum_{i=1}^n \frac{1}{2\sigma^2}(Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2 + \sum_{i=1}^n \log Y_i^{\lambda - 1} \\ &= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2 + (\lambda - 1)\sum_{i=1}^n \log Y_i\end{aligned}$$

From the first order condition $\frac{\partial \log L}{\partial \sigma} = 0$, we have

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n (Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2 = 0$$

giving

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2\ .$$

Substituting into the log-likelihood function, we obtain the concentrated log-likelihood

$$\log L(\beta_1, \beta_2, \lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\frac{1}{n}\sum_{i=1}^n (Y_{\lambda i} - \beta_1 - \beta_2 X_i)^2 - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^n \log Y_i$$

The expression can be simplified (Zarembka, 1968) by working with Y_i^* rather than Y_i, where Y_i^* is Y_i divided by Y_{GM}, the geometric mean of the Y_i in the sample, for

$$\begin{aligned}\sum_{i=1}^{n} \log Y_i^* &= \sum_{i=1}^{n} \log(Y_i / Y_{GM}) = \sum_{i=1}^{n} (\log Y_i - \log Y_{GM}) \\ &= \sum_{i=1}^{n} \log Y_i - n \log Y_{GM} = \sum_{i=1}^{n} \log Y_i - n \log\left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}} \\ &= \sum_{i=1}^{n} \log Y_i - \log\left(\prod_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \log Y_i - \sum_{i=1}^{n} \log Y_i = 0\,.\end{aligned}$$

With this simplification, the log-likelihood is

$$\log L(\beta_1, \beta_2, \lambda) = -\frac{n}{2}\left(\log 2\pi + \log\frac{1}{n} + 1\right) - \frac{n}{2}\log\sum_{i=1}^{n}\left(Y_{\lambda i}^* - \beta_1 - \beta_2 X_i\right)^2$$

and it will be maximised when β_1, β_2 and λ are chosen so as to minimise

$\sum_{i=1}^{n}\left(Y_{\lambda i}^* - \beta_1 - \beta_2 X_i\right)^2$ the residual sum of squares from a least squares regression of the scaled, transformed Y on X. One simple procedure is to perform a grid search, scaling and transforming the data on Y for a range of values of λ and choosing the value that leads to the smallest residual sum of squares (Spitzer, 1982).

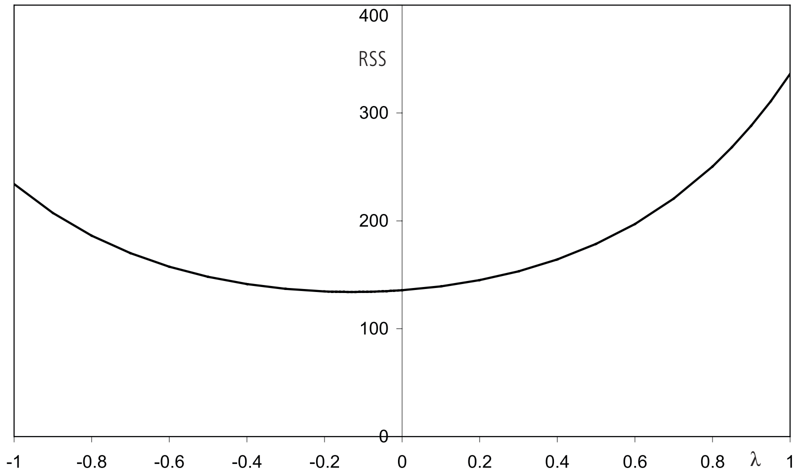
A null hypothesis $\lambda = \lambda_0$ can be tested using a likelihood ratio test in the usual way. Under the null hypothesis, the test statistic $2(\log L_\lambda - \log L_0)$ will have a chi-squared distribution with one degree of freedom, where log L_λ is the unconstrained log-likelihood and L_0 is the constrained one. Note that, in view of the preceding equation,

$$2(\log L_\lambda - \log L_0) = n(\log RSS_0 - \log RSS_\lambda)$$

where RSS_0 and RSS_λ are the residual sums of squares from the constrained and unconstrained regressions with Y^*.

The most obvious tests are $\lambda = 0$ for the logarithmic specification and $\lambda = 1$ for the linear one. Note that it is not possible to test the two hypotheses directly against each other. As with all tests, one can only test whether a hypothesis is incompatible with the sample result. In this case we are testing whether the log-likelihood under the restriction is significantly smaller than the unrestricted log-likelihood. Thus, while it is possible that we may reject the linear but not the logarithmic, or vice versa, it is also possible that we may reject both or fail to reject both.

Example



The figure shows the residual sum of squares for values of λ from –1 to 1 for the earnings function example described in Section 4.5 in the text. The maximum likelihood estimate is –0.13, with RSS = 134.09. For the linear and logarithmic specifications, RSS was 336.29 and 135.72, respectively, with likelihood ratio statistics 540(log 336.29 – log 134.09) = 496.5 and 540(log 135.72 – log 134.09) = 6.52. The logarithmic specification is clearly much to be preferred, but even it is rejected at the 5 per cent level , with $\chi^2(1)$ = 3.84, and nearly at the 1 per cent level.

Additional exercises

A4.1

Is expenditure on your category per capita related to total expenditure per capita? An alternative model specification.

Define a new variable *LGCATPC* as the logarithm of expenditure per capita on your category. Define a new variable *LGEXPPC* as the logarithm of total household expenditure per capita. Regress *LGCATPC* on *LGEXPPC*. Provide an interpretation of the coefficients, and perform appropriate statistical tests.

A4.2

Is expenditure on your category per capita related to household size as well as to total expenditure per capita? An alternative model specification.

Regress *LGCATPC* on *LGEXPPC* and *LGSIZE*. Provide an interpretation of the coefficients, and perform appropriate statistical tests.

A4.3

A researcher is considering two regression specifications:

$$\log Y = \beta_1 + \beta_2 \log X + u \qquad (1)$$

and

$$\log \frac{Y}{X} = \alpha_1 + \alpha_2 \log X + u \qquad (2)$$

where u is a disturbance term.

Writing $y = \log Y$, $x = \log X$, and $z = \log \dfrac{Y}{X}$, and using the same sample of n observations, the researcher fits the two specifications using OLS:

$$\hat{y} = b_1 + b_2 x \tag{3}$$

and

$$\hat{z} = a_1 + a_2 x \, . \tag{4}$$

- Using the expressions for the OLS regression coefficients, demonstrate that $b_2 = a_2 + 1$.
- Similarly, using the expressions for the OLS regression coefficients, demonstrate that $b_1 = a_1$.
- Hence demonstrate that the relationship between the fitted values of y, the fitted values of z, and the actual values of x, is $\hat{y}_i - x_i = \hat{z}_i$.
- Hence show that the residuals for regression (3) are identical to those for (4).
- Hence show that the standard errors of b_2 and a_2 are the same.
- Determine the relationship between the t statistic for b_2 and the t statistic for a_2, and give an intuitive explanation for the relationship.
- Explain whether R^2 would be the same for the two regressions.

A4.4

Perform a RESET test of functional misspecification. Using your *EAEF* data set, regress *WEIGHT02* on *HEIGHT*. Save the fitted values as *YHAT* and define *YHATSQ* as its square. Add *YHATSQ* to the regression specification and test its coefficient.

A4.5

Is a logarithmic specification preferable to a linear specification for an expenditure function?

Define *CATPCST* as *CATPC* scaled by its geometric mean and *LGCATST* as the logarithm of *CATPCST*. Regress *CATPCST* on *EXPPC* and *SIZE* and regress *LGCATST* on *LGEXPPC* and *LGSIZE*. Compare the *RSS* for these equations.

A4.6

A researcher hypothesises that a variable Y is determined by a variable X and considers the following four alternative regression specifications, using cross-sectional data:

$$Y = \beta_1 + \beta_2 X + u \tag{1}$$

$$\log Y = \beta_1 + \beta_2 X + u \tag{2}$$

$$Y = \beta_1 + \beta_2 \log X + u \tag{3}$$

$$\log Y = \beta_1 + \beta_2 \log X + u \, . \tag{4}$$

Explain why a direct comparison of R^2, or of *RSS*, in models (1) and (2) is illegitimate. What should be the strategy of the researcher for determining which of the four specifications has the best fit?

A4.7

A researcher has data on a measure of job performance, *SKILL*, and years of work experience, *EXP*, for a sample of individuals in the same occupation. Believing there to be diminishing returns to experience, the researcher proposes the model

$$SKILL = \beta_1 + \beta_2 \log(EXP) + \beta_3 \log(EXP^2) + u.$$

Comment on this specification.

A4.8

```
. reg LGEARN S EXP ASVABC SA

      Source |       SS       df       MS              Number of obs =     270
-------------+------------------------------           F(  4,   265) =   31.72
       Model |  30.0320896     4  7.5080224            Prob > F      =  0.0000
    Residual |  62.7338804   265  .236731624           R-squared     =  0.3237
-------------+------------------------------           Adj R-squared =  0.3135
       Total |    92.76597   269  .344854907           Root MSE      =  .48655

------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |  -.0241627   .0761646    -0.32   0.751    -.1741275    .1258021
         EXP |   .0259103   .0086572     2.99   0.003     .0088646    .0429561
      ASVABC |  -.0095437   .0175083    -0.55   0.586    -.0440169    .0249295
          SA |   .0019856   .0013398     1.48   0.140    -.0006524    .0046237
       _cons |   1.874952   .9344235     2.01   0.046     .0351132    3.714791
------------------------------------------------------------------------------
```

The output above shows the result of regressing the logarithm of hourly earnings on years of schooling, years of work experience, *ASVABC* score, and *SA*, an interactive variable defined as the product of *S* and *ASVABC*, for males in *EAEF* Data Set 21. The mean values of *S*, *EXP*, and *ASVABC* in the sample were 13.7, 17.9, and 52.1, respectively. Give an interpretation of the regression output.

Answers to the starred exercises in the textbook

4.8

Suppose that the logarithm of *Y* is regressed on the logarithm of *X*, the fitted regression being

$$\log \hat{Y} = b_1 + b_2 \log X \cdot$$

Suppose $X^* = \lambda X$, where λ is a constant, and suppose that $\log Y$ is regressed on $\log X^*$. Determine how the regression coefficients are related to those of the original regression. Determine also how the t statistic for b_2 and R^2 for the equation are related to those in the original regression.

Answer:

Nothing of substance is affected since the change amounts only to a fixed constant shift in the measurement of the explanatory variable.

Let the fitted regression be

$$\log \hat{Y} = b_1^* + b_2^* \log X^* \quad .$$

Note that

$$\log X_i^* - \overline{\log X^*} = \log \lambda X_i - \frac{1}{n}\sum_{j=1}^{n} \log X_j^* = \log \lambda X_i - \frac{1}{n}\sum_{j=1}^{n} \log \lambda X_j$$

$$= \log \lambda + \log X_i - \frac{1}{n}\sum_{j=1}^{n}\left(\log \lambda + \log X_j\right) = \log X_i - \frac{1}{n}\sum_{j=1}^{n} \log X_j$$

$$= \log X_i - \overline{\log X}.$$

Hence $b_2^* = b_2$. To compute the standard error of b_2^*, we will also need b_1^*.

$$b_1^* = \overline{\log Y} - b_2^* \overline{\log X^*} = \overline{\log Y} - b_2 \frac{1}{n}\sum_{j=1}^{n}\left(\log\lambda + \log X_j\right)$$

$$= \overline{\log Y} - b_2 \log\lambda - b_2 \overline{\log X} = b_1 - b_2 \log\lambda.$$

Thus the residual e_i^* is given by

$$e_i^* = \log Y_i - b_1^* - b_2^* \log X_i^* = \log Y_i - \left(b_1 - b_2 \log\lambda\right) - b_2\left(\log X_i + \log\lambda\right) = e_i.$$

Hence the estimator of the variance of the disturbance term is unchanged and so the standard error of b_2^* is the same as that for b_2. As a consequence, the t statistic must be the same. R^2 must also be the same:

$$R^{2^*} = 1 - \frac{\sum e_i^{*2}}{\sum\left(\log Y_i - \overline{\log Y}\right)} = 1 - \frac{\sum e_i^2}{\sum\left(\log Y_i - \overline{\log Y}\right)} = R^2.$$

4.14

```
. reg LGS LGSM LGSMSQ

      Source |       SS       df       MS              Number of obs =     536
-------------+------------------------------           F(  1,   534) =   56.99
       Model |  1.62650898     1  1.62650898           Prob > F      =  0.0000
    Residual |  15.2402109   534  .028539721           R-squared     =  0.0964
-------------+------------------------------           Adj R-squared =  0.0947
       Total |  16.8667198   535  .031526579           Root MSE      =  .16894

------------------------------------------------------------------------------
         LGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        LGSM |  (omitted)
      LGSMSQ |    .100341   .0132915     7.55   0.000     .0742309    .1264511
       _cons |    2.11373   .0648636    32.59   0.000     1.986311    2.241149
------------------------------------------------------------------------------
```

The output shows the results of regressing, *LGS*, the logarithm of *S*, on *LGSM*, the logarithm of *SM*, and *LGSMSQ*, the logarithm of *SMSQ*. Explain the regression results.

Answer:

LGSMSQ = 2*LGSM*, so the specification is subject to exact multicollinearity. In such a situation, Stata drops one of the variables responsible.

4.16

```
. nl (S = {beta1} + {beta2}/({beta3} + SIBLINGS)) if SIBLINGS>0
(obs = 529)

Iteration 0:  residual SS =  2962.929
Iteration 1:  residual SS =  2951.616
....................................
Iteration 13:  residual SS =  2926.201

      Source |       SS       df       MS
-------------+------------------------------         Number of obs =       529
       Model |  206.566702     2  103.283351         R-squared     =    0.0659
    Residual |  2926.20078   526  5.56311936         Adj R-squared =    0.0624
-------------+------------------------------         Root MSE      =  2.358627
       Total |  3132.76749   528  5.93327175         Res. dev.     =  2406.077

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /beta1 |   11.09973   1.363292     8.14   0.000     8.421565     13.7779
      /beta2 |   17.09479   18.78227     0.91   0.363    -19.80268    53.99227
      /beta3 |   3.794949    3.66492     1.04   0.301    -3.404729    10.99463
------------------------------------------------------------------------------
  Parameter beta1 taken as constant term in model & ANOVA table
```

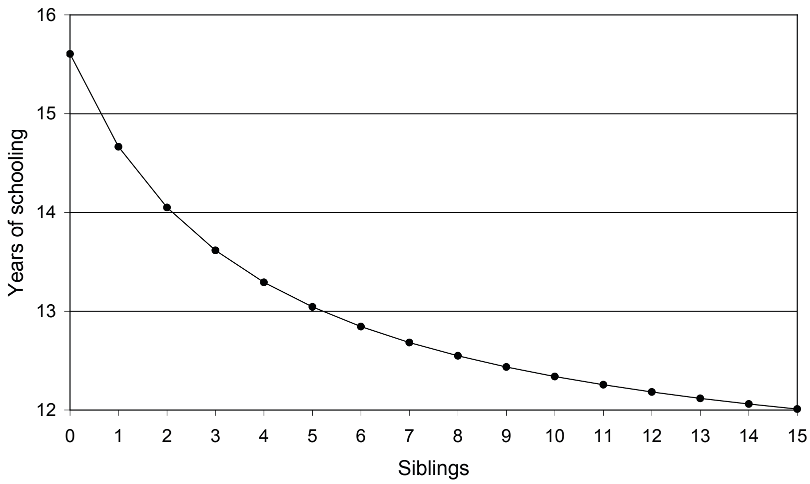
The output uses *EAEF* Data Set 21 to fit the nonlinear model

$$S = \beta_1 + \frac{\beta_2}{\beta_3 + SIBLINGS} + u$$

where S is the years of schooling of the respondent and *SIBLINGS* is the number of brothers and sisters. The specification is an extension of that for Exercise 4.1, with the addition of the parameter β_3. Provide an interpretation of the regression results and compare it with that for Exercise 4.1.

Answer:

As in Exercise 4.1, the estimate of β_1 provides an estimate of the lower bound of schooling, 11.10 years, when the number of siblings is large. The other parameters do not have straightforward interpretations. The figure below represents the relationship. Comparing this figure with that for Exercise 4.1, it can be seen that it gives a very different picture of the adverse effect of additional siblings. The figure in Exercise 4.1, reproduced after it, suggests that the adverse effect is particularly large for the first few siblings, and then attenuates. This figure indicates that the adverse effect is more evenly spread and is more enduring. However, the relationship has been fitted with imprecision since the estimates of β_2 and β_3 are not significant.



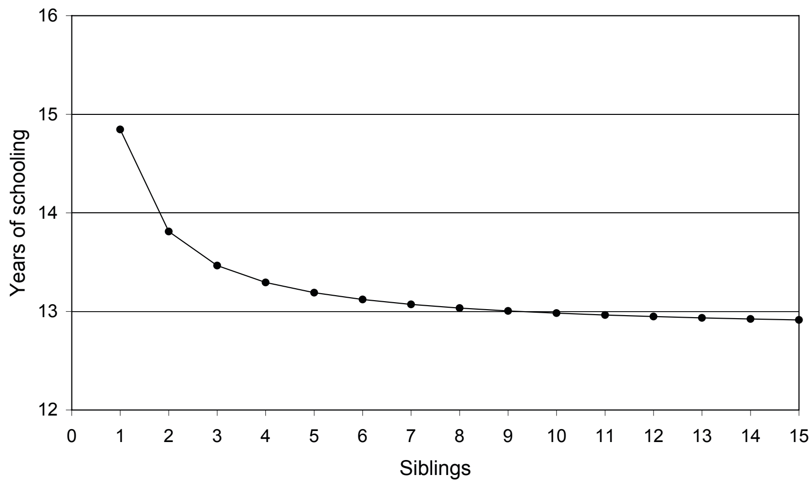


Figure for Exercise 4.1

Answers to the additional exercises

A4.1

```
. g LGEXPPC =LGEXP -LGSIZE
. g LGFDHOPC=LGFDHO-LGSIZE
(1 missing value generated)

. reg LGFDHOPC LGEXPPC

      Source |       SS       df       MS              Number of obs =     868
-------------+------------------------------           F(  1,   866) =  313.04
       Model |  51.4364294     1  51.4364294           Prob > F      =  0.0000
    Residual |  142.293979   866  .164311754           R-squared     =  0.2655
-------------+------------------------------           Adj R-squared =  0.2647
       Total |  193.730408   867  .223449145           Root MSE      =  .40535

------------------------------------------------------------------------------
    LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LGEXPPC |    .376283   .0212674    17.69   0.000     .3345414    .4180246
       _cons |   3.700667   .1978924    18.70   0.000     3.312263    4.089072
------------------------------------------------------------------------------
```

The regression implies that the income elasticity of expenditure on food is 0.38 (supposing that total household expenditure can be taken as a proxy for permanent income). In addition to testing the null hypothesis that the elasticity is equal to zero, which is rejected at a very high significance level for this and all the other categories except *LOCT*, one might test whether it is different from 1, as a means of classifying the categories of expenditure as luxuries (elasticity > 1) and necessities (elasticity < 1).

The table gives the results for all the categories of expenditure.

Regression of *LGCATPC* on *EXPPC*							
	n	b_2	$s.e.(b_2)$	$t\ (\beta_2 = 0)$	$t\ (\beta_2 = 1)$	R^2	RSS
FDHO	868	0.3763	0.0213	17.67	–29.28	0.2655	142.29
FDAW	827	1.3203	0.0469	28.15	6.83	0.4903	608.05
HOUS	867	1.1006	0.0401	27.45	2.51	0.4653	502.08
TELE	858	0.6312	0.0353	17.88	–10.45	0.2717	380.59
DOM	454	0.7977	0.1348	5.92	–1.50	0.0719	1325.21
TEXT	482	1.0196	0.0813	12.54	0.24	0.2469	560.37
FURN	329	0.8560	0.1335	6.41	–1.08	0.1117	697.33
MAPP	244	0.7572	0.1161	6.52	–2.09	0.1496	291.76
SAPP	467	0.9481	0.0810	11.70	–0.64	0.2275	522.31
CLOT	847	0.9669	0.0487	19.85	–0.68	0.3184	686.45
FOOT	686	0.7339	0.0561	13.08	–4.74	0.1999	589.34
GASO	797	0.7107	0.0379	18.75	–7.63	0.3062	366.92
TRIP	309	1.2434	0.1305	9.53	1.87	0.2283	527.42
LOCT	172	0.1993	0.1808	1.10	–4.43	0.0071	450.92
HEAL	821	0.8629	0.0716	12.05	–1.91	0.1505	1351.63
ENT	824	1.3069	0.0521	25.08	5.89	0.4336	754.86
FEES	676	1.5884	0.0811	19.59	7.26	0.3629	1145.09
TOYS	592	0.9497	0.0771	12.32	–0.65	0.2045	809.01
READ	764	1.1532	0.0641	17.99	2.39	0.2982	897.63
EDUC	288	1.2953	0.1600	8.10	1.85	0.1865	828.35
TOB	368	0.6646	0.0817	8.13	–4.11	0.1530	385.63

The results may be summarised as follows:

- Significantly greater than 1, at the 1 per cent level: *FDAW*, *ENT*, *FEES*.
- Significantly greater than 1, at the 5 per cent level: *HOUS*, *READ*.
- Not significantly different from 1 *DOM*, *TEXT*, *FURN*, *SAPP*, *CLOT*, *TRIP*, *HEAL*, *TOYS*, *EDUC*.
- Significantly less than 1, at the 1 per cent level: *FDHO*, *TELE*, *FOOT*, *GASO*, *LOCT*, *TOB*.
- Significantly less than 1, at the 5 per cent level: *MAPP*.

A4.2

```
. reg LGFDHOPC LGEXPPC LGSIZE

      Source |       SS       df       MS              Number of obs =     868
-------------+------------------------------           F(  2,   865) =  210.94
       Model |  63.5111789     2  31.7555894           Prob > F      =  0.0000
    Residual |  130.219229   865  .150542462           R-squared     =  0.3278
-------------+------------------------------           Adj R-squared =  0.3263
       Total |  193.730408   867  .223449145           Root MSE      =    .388

------------------------------------------------------------------------------
    LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LGEXPPC |   .2866812   .0226824    12.64   0.000     .2421622    .3312003
      LGSIZE |  -.2278489   .0254412    -8.96   0.000    -.2777826   -.1779152
       _cons |   4.720269   .2209996    21.36   0.000     4.286511    5.154028
------------------------------------------------------------------------------
```

The income elasticity, 0.29, is now a little lower than before. The size elasticity is significantly negative, suggesting economies of scale and indicating that the model in the previous exercise was misspecified. *t* tests of the hypothesis that the income elasticity is equal to 1 produce the following results:

- Significantly greater than 1, at the 1 per cent level: *FDAW*, *ENT*, *FEES*.
- Significantly greater than 1, at the 5 per cent level: *CLOT* .
- Not significantly different from 1: *HOUS*, *DOM*, *TEXT*, *TRIP*, *TOYS*, *READ*, *EDUC*.
- Significantly less than 1, at the 1 per cent level: *FDHO*, *TELE*, *FURN*, *MAPP*, *SAPP*, *FOOT*, *GASO*, *LOCT*, *HEAL*, *TOB*.
- Significantly less than 1, at the 5 per cent level: none.

Dependent variable *LGCATPC*								
		LGEXPPC		LGSIZE				
	n	b_2	$s.e.(b_2)$	b_3	$s.e.(b_3)$	R^2	F	RSS
FDHO	868	0.2867	0.0227	–0.2278	0.0254	0.3278	210.9	130.22
FDAW	827	1.4164	0.0529	0.2230	0.0588	0.4990	410.4	597.61
HOUS	867	1.0384	0.0446	–0.1566	0.0498	0.4714	385.2	496.41
TELE	858	0.4923	0.0378	–0.3537	0.0423	0.3268	207.5	351.81
DOM	454	0.8786	0.1470	0.2084	0.1520	0.0758	18.5	1319.71
TEXT	482	0.9543	0.0913	–0.1565	0.1005	0.2507	80.1	557.55
FURN	329	0.6539	0.1511	–0.4622	0.1677	0.1319	24.8	681.45
MAPP	244	0.5136	0.1381	–0.4789	0.1533	0.1827	26.9	280.41
SAPP	467	0.7223	0.0899	–0.5076	0.0973	0.2703	85.9	493.39
CLOT	847	1.1138	0.0539	0.3502	0.0597	0.3451	222.4	659.59
FOOT	686	0.6992	0.0638	–0.0813	0.0711	0.2015	86.2	588.21
GASO	797	0.6770	0.0433	–0.0785	0.0490	0.3084	177.0	365.73
TRIP	309	1.0563	0.1518	–0.3570	0.1510	0.2421	48.9	517.96
LOCT	172	–0.0141	0.1958	–0.5429	0.2084	0.0454	4.0	433.51
HEAL	821	0.6612	0.0777	–0.5121	0.0849	0.1868	93.9	1294.03
ENT	824	1.4679	0.0583	0.3771	0.0658	0.4554	343.2	725.85
FEES	676	1.7907	0.0940	0.4286	0.1042	0.3786	205.0	1117.00
TOYS	592	0.9522	0.0905	0.0054	0.1011	0.2045	75.7	809.01
READ	764	0.9652	0.0712	–0.4313	0.0768	0.3262	184.2	861.92
EDUC	288	1.2243	0.1882	–0.1707	0.2378	0.1879	33.0	826.85
TOB	368	0.4329	0.0915	–0.5379	0.1068	0.2080	47.9	360.58

A4.3

- *Using the expressions for the OLS regression coefficients, demonstrate that* $b_2 = a_2 + 1$.

$$a_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})([y_i - x_i] - [\bar{y} - \bar{x}])}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = b_2 - 1.$$

- *Similarly, using the expressions for the OLS regression coefficients, demonstrate that* $b_1 = a_1$.

$$a_1 = \bar{z} - a_2\bar{x} = (\bar{y} - \bar{x}) - a_2\bar{x} = \bar{y} - (a_2 + 1)\bar{x} = \bar{y} - b_2\bar{x} = b_1 .$$

- *Hence demonstrate that the relationship between the fitted values of y, the fitted values of z, and the actual values of x, is* $\hat{y}_i - x_i = \hat{z}_i$.

$$\hat{z}_i = a_1 + a_2 x_i = b_1 + (b_2 - 1)x_i = b_1 + b_2 x_i - x_i = \hat{y}_i - x_i .$$

- *Hence show that the residuals for regression (3) are identical to those for (4).*

Let e_i be the residual in (3) and f_i the residual in (4). Then

$$f_i = z_i - \hat{z}_i = y_i - x_i - (\hat{y}_i - x_i) = y_i - \hat{y}_i = e_i .$$

- *Hence show that the standard errors of b_2 and a_2 are the same.*

The standard error of b_2 is

$$\text{s.e.}(b_2) = \sqrt{\frac{\sum e_i^2 /(n-2)}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{\sum f_i^2 /(n-2)}{\sum (x_i - \bar{x})^2}} = \text{s.e.}(a_2).$$

- *Determine the relationship between the t statistic for b_2 and the t statistic for a_2, and give an intuitive explanation for the relationship.*

$$t_{b_2} = \frac{b_2}{\text{s.e.}(b_2)} = \frac{a_2 + 1}{\text{s.e.}(a_2)} .$$

The t statistic for b_2 is for the test of H_0: $\beta_2 = 0$. Given the relationship, it is also for the test of H_0: $\alpha_2 = -1$. The tests are equivalent since both of them reduce the model to log Y depending only on an intercept and the disturbance term.

- *Explain whether R^2 would be the same for the two regressions.*

R^2 will be different because it measures the proportion of the variance of the dependent variable explained by the regression, and the dependent variables are different.

A4.4

In the first part of the output, *WEIGHT02* is regressed on *HEIGHT*, using *EAEF* Data Set 21. The predict command saves the fitted values from the most recent regression, assigning them the variable name that follows the command., in this case *YHAT*. *YHATSQ* is defined as the square of *YHAT*, and this is added to the regression specification. Its coefficient is significant at the 1 per cent level, indicating, as one would expect, that the relationship between weight and height is nonlinear.

```
. reg WEIGHT02 HEIGHT

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  216.95
       Model |  311260.383     1  311260.383           Prob > F      =  0.0000
    Residual |  771880.527   538  1434.72217           R-squared     =  0.2874
-------------+------------------------------           Adj R-squared =  0.2860
       Total |  1083140.91   539  2009.53787           Root MSE      =  37.878

------------------------------------------------------------------------------
    WEIGHT02 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HEIGHT |   5.669766   .3849347    14.73   0.000     4.913606    6.425925
       _cons |  -199.6832   26.10105    -7.65   0.000    -250.9556   -148.4107
------------------------------------------------------------------------------

. predict YHAT
(option xb assumed; fitted values)

. g YHATSQ = YHAT*YHAT

. reg WEIGHT02 HEIGHT YHATSQ

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  114.87
       Model |  324546.101     2   162273.05           Prob > F      =  0.0000
    Residual |  758594.809   537  1412.65328           R-squared     =  0.2996
-------------+------------------------------           Adj R-squared =  0.2970
       Total |  1083140.91   539  2009.53787           Root MSE      =  37.585

------------------------------------------------------------------------------
    WEIGHT02 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HEIGHT |  -7.240152    4.22697    -1.71   0.087    -15.54358    1.063271
      YHATSQ |   .0062029   .0020226     3.07   0.002     .0022296    .0101761
       _cons |   460.3737   216.7846     2.12   0.034     34.52394    886.2234
------------------------------------------------------------------------------
```

A4.5

The *RSS* comparisons for all the categories of expenditure indicate that the logarithmic specification is overwhelmingly superior to the linear one. The differences are actually surprisingly large and suggest that some other factor may also be at work. One possibility is that the data contain many outliers, and these do more damage to the fit in linear than in logarithmic specifications. To see this, plot *CATPC* and *EXPPC* and compare with a plot of *LGCATPC* and *LGEXPPC*. (Strictly speaking, you should control for *SIZE* and *LGSIZE* using the Frisch–Waugh–Lovell method described in Chapter 3.)

RSS from Zarembka transformations			
	n	*RSS linear*	*RSS logarithmic*
FDHO	868	197.58	130.22
FDAW	827	2993.63	597.61
HOUS	867	888.75	496.41
TELE	858	1448.27	351.81
DOM	454	61271.17	1319.71
TEXT	482	20655.14	557.55
FURN	329	6040.07	681.45
MAPP	244	1350.83	280.41
SAPP	467	3216.40	493.39
CLOT	847	1919.32	659.59
FOOT	686	1599.01	588.21
GASO	797	597.57	365.73
TRIP	309	3828.14	517.96
LOCT	172	2793.50	433.51
HEAL	821	2295.19	1294.03
ENT	824	6267.20	725.85
FEES	676	33224.88	1117.00
TOYS	592	4522.51	809.01
READ	764	2066.83	861.92
EDUC	288	44012.28	826.85
TOB	368	617.45	360.58

A4.6

In (1) R^2 is the proportion of the variance of Y explained by the regression. In (2) it is the proportion of the variance of log Y explained by the regression. Thus, although related, they are not directly comparable. In (1) *RSS* has dimension the squared units of Y. In (2) it has dimension the squared units of log Y. Typically it will be much lower in (2) because the logarithm of Y tends to be much smaller than Y.

The specifications with the same dependent variable may be compared directly in terms of *RSS* (or R^2) and hence two of the specifications may be eliminated immediately. The remaining two specifications should be compared after scaling, with Y replaced by Y^* where Y^* is defined as Y divided by the geometric mean of Y in the sample. *RSS* for the scaled regressions will then be comparable.

A4.7

The proposed model

$$SKILL = \beta_1 + \beta_2 \log(EXP) + \beta_3 \log(EXP^2) + u$$

cannot be fitted since

$$\log(EXP^2) = 2\log(EXP)$$

and the specification is therefore subject to exact multicollinearity.

A4.8

Let the theoretical model for the regression be written

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 ASVABC + \beta_5 SA + u.$$

The estimates of β_2 and β_4 are negative, at first sight suggesting that schooling and cognitive ability have adverse effects on earnings, contrary to common sense and previous results with wage equations of this kind. However, rewriting the model as

$$LGEARN = \beta_1 + (\beta_2 + \beta_5 ASVABC)S + \beta_3 EXP + \beta_4 ASVABC + u$$

it can be seen that, as a consequence of the inclusion of the interactive term, β_2 represents the effect of a marginal year of schooling for an individual with an *ASVABC* score of zero. Since no individual in the sample had a score less than 25, the perverse sign of the estimate illustrates only the danger of extrapolating outside the data range. It makes better sense to evaluate the implicit coefficient for an individual with the mean *ASVABC* score of 52.1. This is (–0.024163 + 0.001986*52.1) = 0.079, implying a much more plausible 7.9 per cent increase in earnings for each year of schooling. The positive sign of the coefficient of *SASVABC* implies that the coefficient is somewhat higher for those with above-average *ASVABC* scores and somewhat lower for those with below average scores. For those with the highest score, 66, it would be 10.7, and for those with the lowest score, 25, it would be 2.5.

Similar considerations apply to the interpretation of the estimate of β_4 , the coefficient of *ASVABC*. Rewriting the model as

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + (\beta_4 + \beta_5 S)ASVABC + u$$

it can be seen that β_4 relates to the effect on hourly earnings of a one-unit increase in *ASVABC* for an individual with no schooling. As with β_2, this is outside the data range in the sample, no individual having fewer than 8 years of schooling. If one calculates the implicit coefficient for an individual with the sample mean of 13.7 years of schooling, it comes to (–0.009544 + 0.001986*13.7) = 0.018.

As shown in the exercise, one way of avoiding nonsense parameter estimates is to measure the variables in question from their sample means. This has been done in the regression output below, where *S1* and *ASVABC1* are schooling and *ASVABC* measured from their sample means and *SASVABC1* is their interaction. The only differences in the output are the lines relating to the coefficients of schooling, *ASVABC*, and the intercept, the point estimates of the coefficients of *S* and *ASVABC* being as calculated above.

```
. reg LGEARN S1 EXP ASVABC1 SASVABC1

      Source |       SS       df       MS              Number of obs =     270
-------------+------------------------------           F(  4,   265) =   31.72
       Model |  30.0320902     4  7.50802256           Prob > F      =  0.0000
    Residual |  62.7338798   265  .236731622           R-squared     =  0.3237
-------------+------------------------------           Adj R-squared =  0.3135
       Total |    92.76597   269  .344854907           Root MSE      =  .48655

------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          S1 |   .0793138   .0171164     4.63   0.000     .0456124    .1130153
         EXP |   .0259103   .0086572     2.99   0.003     .0088646    .0429561
     ASVABC1 |   .0177037   .0040138     4.41   0.000     .0098007    .0256067
    SASVABC1 |   .0019856   .0013398     1.48   0.140    -.0006524    .0046237
       _cons |   2.465968    .163862    15.05   0.000     2.143331    2.788605
------------------------------------------------------------------------------
```