Chapter 4
Transformations of variables

4.1 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 *EAWE* 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.

4.2 Learning outcomes

After working through the corresponding chapter in the text, studying the corresponding slideshows, and doing the starred exercises in the text and the additional exercises in this subject 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.

4.3 Further material

Box–Cox tests of functional specification

The theory behind the procedure for discriminating between a linear and a logarithmic specification of the dependent variable is explained in the Appendix to Chapter 10 of the text. However, the exposition there is fairly brief. An expanded version is offered here. 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) $N(0, \sigma^2)$, the density function for u_i is:

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

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

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

From this we obtain the density function for Y_i:

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

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:

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

and the log-likelihood is:

$$\begin{aligned}\log L(\beta_1,\beta_2,\sigma,\lambda) &= -\frac{n}{2}\log 2\pi\sigma^2 - \sum_{i=1}^{n}\frac{1}{2\sigma^2}(Y_{\lambda i}-\beta_1-\beta_2X_i)^2 + \sum_{i=1}^{n}\log Y_i^{\lambda-1}\\ &= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_{\lambda i}-\beta_1-\beta_2X_i)^2 + (\lambda-1)\sum_{i=1}^{n}\log Y_i.\end{aligned}$$

From the first-order condition $\partial \log L/\partial\sigma = 0$, we have:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(Y_{\lambda i}-\beta_1-\beta_2X_i)^2 = 0$$

giving:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_{\lambda i}-\beta_1-\beta_2X_i)^2.$$

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

$$\log L(\beta_1,\beta_2,\lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\frac{1}{n}\sum_{i=1}^{n}(Y_{\lambda i}-\beta_1-\beta_2X_i)^2 - \frac{n}{2} + (\lambda-1)\sum_{i=1}^{n}\log Y_i.$$

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:

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

With this simplification, the log-likelihood is:

$$\log L(\beta_1,\beta_2,\lambda) = -\frac{n}{2}\left(\log 2\pi + \log\frac{1}{n} + 1\right) - \frac{n}{2}\log\sum_{i=1}^{n}(Y_{\lambda i}^*-\beta_1-\beta_2X_i)^2$$

and it will be maximised when β_1, β_2 and λ are chosen so as to minimise $\sum_{i=1}^{n}(Y_{\lambda i}^*-\beta_1-\beta_2X_i)^2$, 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).

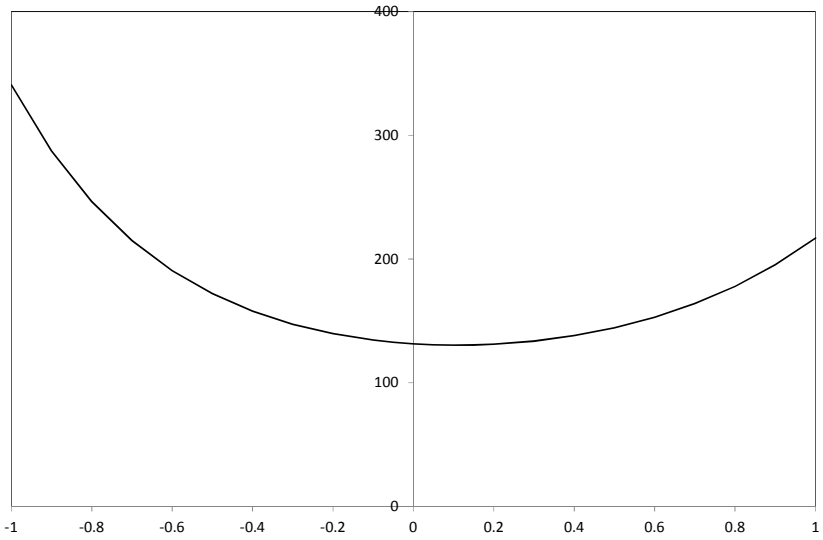
A null hypothesis $\lambda = \lambda_0$ can be tested using a likelihood ratio test in the usual way. Under the null hypothesis, the test statistic $2(\log L_\lambda - \log L_0)$ will have a chi-squared distribution with one degree of freedom, where $\log L_\lambda$ is the unconstrained log-likelihood 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 wage equation example described in Section 4.2 in the text. The maximum likelihood estimate is 0.10, with $RSS = 130.3$. For the linear and logarithmic specifications, RSS was 217.0 and 131.4, respectively, with likelihood ratio statistics $500(\log 217.0 - \log 130.3) = 255.0$ and $500(\log 131.4 - \log 130.3) = 4.20$.The logarithmic specification is clearly much to be preferred, but even it is rejected at the 5 per cent level, with $\chi^2(1) = 3.84$.

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

$$\widehat{y} = \widehat{\beta}_1 + \widehat{\beta}_2 x \qquad (3)$$

and:

$$\widehat{z} = \widehat{\alpha}_1 + \widehat{\alpha}_2 x. \qquad (4)$$

- Using the expressions for the OLS regression coefficients, demonstrate that $\widehat{\beta}_2 = \widehat{\alpha}_2 + 1$.
- Similarly, using the expressions for the OLS regression coefficients, demonstrate that $\widehat{\beta}_1 = \widehat{\alpha}_1$.
- Hence demonstrate that the relationship between the fitted values of y, the fitted values of z, and the actual values of x, is $\widehat{y}_i - x_i = \widehat{z}_i$.
- Hence show that the residuals for regression (3) are identical to those for (4).
- Hence show that the standard errors of $\widehat{\beta}_2$ and $\widehat{\alpha}_2$ are the same.
- Determine the relationship between the t statistic for $\widehat{\beta}_2$ and the t statistic for $\widehat{\alpha}_2$, and give an intuitive explanation for the relationship.
- Explain whether R^2 would be the same for the two regressions.

A4.4 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:

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

Comment on this specification.

A4.5 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:

$$\begin{aligned} Y &= \beta_1 + \beta_2 X + u && (1)\\ \log Y &= \beta_1 + \beta_2 X + u && (2)\\ Y &= \beta_1 + \beta_2 \log X + u && (3)\\ \log Y &= \beta_1 + \beta_2 \log X + u. && (4) \end{aligned}$$

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?

A4.6 *Is a logarithmic specification preferable to a linear specification for an expenditure function?*

Use your category of expenditure from the *CES* data set. 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.

A4.7

```
. reg LGEARN S EXP ASVABC SASVABC

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  4,   495) =   22.68
       Model |  23.6368302     4  5.90920754            Prob > F      =  0.0000
    Residual |   128.96239   495   .26053008            R-squared     =  0.1549
-------------+------------------------------           Adj R-squared =  0.1481
       Total |   152.59922   499   .30581006            Root MSE      =  .51042

------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0764243   .0116879     6.54   0.000     .0534603    .0993883
         EXP |   .0400506   .0096479     4.15   0.000     .0210948    .0590065
      ASVABC |  -.2096325   .1406659    -1.49   0.137    -.4860084    .0667434
     SASVABC |   .0188685   .0093393     2.02   0.044     .0005189    .0372181
       _cons |   1.386753   .2109596     6.57   0.000     .9722664     1.80124
------------------------------------------------------------------------------
```

The output above shows the result of regressing the logarithm of hourly earnings on years of schooling, years of work experience, *ASVABC* score, and *SASVABC*, an interactive variable defined as the product of *S* and *ASVABC*, using *EAWE* Data Set 21. The mean values of *S*, *EXP*, and *ASVABC* in the sample were 14.9, 6.4, and 0.27, respectively. Give an interpretation of the regression output.

A4.8 Perform a RESET test of functional misspecification. Using your *EAWE* data set, regress *WEIGHT11* on *HEIGHT*. Save the fitted values as *YHAT* and define *YHATSQ* as its square. Add *YHATSQ* to the regression specification and test its coefficient.

4.5 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:

$$\log \widehat{Y} = \widehat{\beta}_1 + \widehat{\beta}_2 \log X.$$

Suppose $X^* = \mu X$, where μ 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 $\widehat{\beta}_2$ 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:

$$\log \widehat{Y} = \widehat{\beta}_1^* + \widehat{\beta}_2^* \log X^*.$$

Note that:

$$\begin{aligned}
\log X_i^* - \overline{\log X^*} &= \log \mu X_i - \frac{1}{n}\sum_{j=1}^{n} \log X_j^* \\
&= \log \mu X_i - \frac{1}{n}\sum_{j=1}^{n} \log \mu X_j \\
&= \log \mu + \log X_i - \frac{1}{n}\sum_{j=1}^{n} (\log \mu + \log X_j) \\
&= \log X_i - \frac{1}{n}\sum_{j=1}^{n} \log X_j \\
&= \log X_i - \overline{\log X}.
\end{aligned}$$

Hence $\widehat{\beta}_2^* = \widehat{\beta}_2$. To compute the standard error of $\widehat{\beta}_2^*$, we will also need $\widehat{\beta}_1^*$.

$$\begin{aligned}
\widehat{\beta}_1^* = \overline{\log Y} - \widehat{\beta}_2^* \overline{\log X^*} &= \overline{\log Y} - \widehat{\beta}_2 \frac{1}{n}\sum_{j=1}^{n} (\log \mu + \log X_j) \\
&= \overline{\log Y} - \widehat{\beta}_2 \log \mu - \widehat{\beta}_2 \overline{\log X} \\
&= \widehat{\beta}_1 - \widehat{\beta}_2 \log \mu.
\end{aligned}$$

Thus the residual $\widehat{u}_i^*$ is given by:

$$\widehat{u}_i^* = \log Y_i - \widehat{\beta}_1^* - \widehat{\beta}_2^* \log X_i^* = \log Y_i - (\widehat{\beta}_1 - \widehat{\beta}_2 \log \mu) - \widehat{\beta}_2 (\log X_i + \log \mu) = \widehat{u}_i.$$

Hence the estimator of the variance of the disturbance term is unchanged and so the standard error of $\widehat{\beta}_2^*$ is the same as that for $\widehat{\beta}_2$. As a consequence, the t statistic must be the same. R^2 must also be the same:

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

4.11 *RSS* was the same in Tables 4.6 and 4.8. Demonstrate that this was not a coincidence.

Answer:

This is a special case of the transformation in Exercise 4.7.

4.14

```
. gen LGHTSQ = ln(HEIGHTSQ)

. reg LGWT04 LGHEIGHT LGHTSQ


      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  211.28
       Model |  7.90843858     1  7.90843858           Prob > F      =  0.0000
    Residual |  18.6403163   498  .037430354           R-squared     =  0.2979
-------------+------------------------------           Adj R-squared =  0.2965
       Total |  26.5487548   499  .053203918           Root MSE      =  .19347

------------------------------------------------------------------------------
      LGWT04 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    LGHEIGHT |  (dropped)
      LGHTSQ |   1.053218   .0724577    14.54   0.000      .9108572   1.195578
       _cons |   -3.78834    .610925    -6.20   0.000     -4.988648  -2.588031
------------------------------------------------------------------------------
```

The output shows the results of regressing, *LGWT04*, the logarithm of *WEIGHT04*, on *LGHEIGHT*, the logarithm of *HEIGHT*, and *LGHTSQ*, the logarithm of the square of *HEIGHT*, using *EAWE* Data Set 21. Explain the regression results, comparing them with those in Exercise 4.2.

Answer:

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

4.18

```
. nl (S = {beta1} + {beta2}/({beta3} + SIBLINGS)) if SIBLINGS>0
(obs = 473)

Iteration 0:  residual SS =  3502.041
Iteration 1:  residual SS =  3500.884
.................................
Iteration 14:  residual SS =  3482.794

      Source |       SS       df       MS
-------------+------------------------------        Number of obs =       473
       Model |  132.339291     2  66.1696453        R-squared     =    0.0366
    Residual |   3482.7939   470  7.41019979        Adj R-squared =    0.0325
-------------+------------------------------        Root MSE      =  2.722168
       Total |  3615.13319   472  7.65918049        Res. dev.     =  2286.658

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /beta1 |   10.45811   5.371492     1.95   0.052    -.0970041    21.01322
      /beta2 |   47.95198   125.3578     0.38   0.702    -198.3791    294.2831
      /beta3 |     8.6994   15.10277     0.58   0.565    -20.97791    38.37671
------------------------------------------------------------------------------
  Parameter beta1 taken as constant term in model & ANOVA table
```

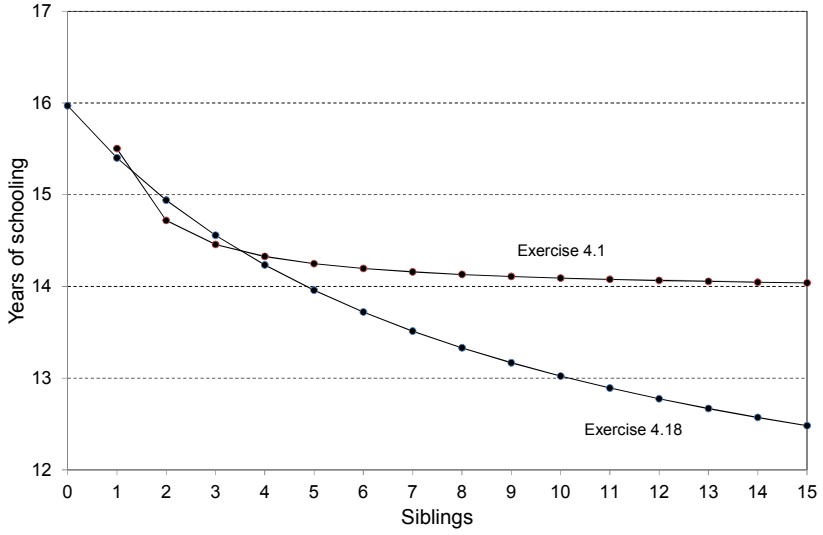
The output uses *EAWE* Data Set 21 to fit the nonlinear model:

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

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, 10.46 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 specification in Exercise 4.1 suggests that the adverse effect is particularly large for the first few siblings, and then attenuates. The revised specification 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.



4.6 Answers to the additional exercises

A4.1

```
. reg LGFDHOPC LGEXPPC

      Source |       SS       df       MS              Number of obs =    6334
-------------+------------------------------           F(  1,  6332) = 4757.00
       Model |  1502.58932     1  1502.58932           Prob > F      =  0.0000
    Residual |  2000.08269  6332  .315869029           R-squared     =  0.4290
-------------+------------------------------           Adj R-squared =  0.4289
       Total |  3502.67201  6333  .553082585           Root MSE      =  .56202
```

```
------------------------------------------------------------------------------
    LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LGEXPPC |   .6092734   .0088338    68.97   0.000     .5919562    .6265905
       _cons |   .8988291   .0703516    12.78   0.000     .7609161    1.036742
------------------------------------------------------------------------------
```

The regression implies that the income elasticity of expenditure on food is 0.61 (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 all the categories, 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	$\widehat{\beta}_2$	s.e.$(\widehat{\beta}_2)$	$t\ (\beta_2 = 0)$	$t\ (\beta_2 = 1)$	R^2	*RSS*
ADM	2,815	1.098	0.030	37.20	3.33	0.330	1,383.9
CLOT	4,500	0.794	0.021	37.34	−9.69	0.237	1,394.0
DOM	1,661	0.812	0.049	16.54	−3.84	0.142	273.5
EDUC	561	1.382	0.090	15.43	4.27	0.299	238.1
ELEC	5,828	0.586	0.011	50.95	−36.05	0.308	2,596.3
FDAW	5,102	0.947	0.015	64.68	−3.59	0.451	4,183.6
FDHO	6,334	0.609	0.009	68.97	−44.23	0.429	4,757.0
FOOT	1,827	0.608	0.027	22.11	−14.26	0.211	488.7
FURN	487	0.912	0.085	10.66	−1.03	0.190	113.7
GASO	5,710	0.677	0.012	56.92	−27.18	0.362	3,240.1
HEAL	4,802	0.868	0.021	40.75	−6.22	0.257	1,660.6
HOUS	6,223	1.033	0.014	73.34	2.34	0.464	5,378.5
LIFE	1,253	0.607	0.047	13.00	−8.40	0.119	169.1
LOCT	692	0.510	0.055	9.29	−8.92	0.111	86.2
MAPP	399	0.817	0.033	9.87	−2.21	0.197	97.5
PERS	3,817	0.891	0.019	48.14	−5.88	0.378	2,317.3
READ	2,287	0.909	0.032	28.46	−2.84	0.262	809.9
SAPP	1,037	0.665	0.045	14.88	−7.49	0.176	221.3
TELE	5,788	0.710	0.012	58.30	−23.82	0.370	3,398.8
TEXT	992	0.629	0.046	13.72	−8.09	0.160	188.2
TOB	1,155	0.721	0.035	20.39	−7.87	0.265	415.8
TOYS	2,504	0.733	0.028	26.22	−9.57	0.216	687.5
TRIP	516	0.723	0.077	9.43	−3.60	0.147	88.9

A4.2

```
. reg LGFDHOPC LGEXPPC LGSIZE

------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    6334
-----------+------------------------------           F(  2,  6331) = 2410.79
     Model |  1514.30728     2   757.15364           Prob> F       =  0.0000
  Residual |  1988.36473  6331  .314068035           R-squared     =  0.4323
-----------+------------------------------           Adj R-squared =  0.4321
     Total |  3502.67201  6333  .553082585           Root MSE      =  .56042
------------------------------------------------------------------------------
  LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
   LGEXPPC |   .5842097   .0097174    60.12   0.000     .5651604    .6032591
    LGSIZE |  -.0814427   .0133333    -6.11   0.000    -.1075806   -.0553049
     _cons |   1.158326   .0820119    14.12   0.000     .9975545    1.319097
------------------------------------------------------------------------------
```

The income elasticity, 0.58, 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.

The specification is equivalent to that in Exercise 4.5 in the text. Writing the latter again as:

$$LGCAT = \beta_1 + \beta_2 LGEXP + \beta_3 LGSIZE + u$$

we have:

$$LGCAT - LGSIZE = \beta_1 + \beta_2(LGEXP - LGSIZE) + (\beta_3 + \beta_2 - 1)LGSIZE + u$$

and so:

$$LGCATPC = \beta_1 + \beta_2 LGEXPPC + (\beta_3 + \beta_2 - 1)LGSIZE + u.$$

Note that the estimates of the income elasticity are identical to those in Exercise 4.5 in the text. This follows from the fact that the theoretical coefficient, β_2, has not been affected by the manipulation. The specification differs from that in Exercise A4.1 in that we have not dropped the $LGSIZE$ term and so we are not imposing the restriction $\beta_3 + \beta_2 - 1 = 0$.

Dependent variable *LGCATPC*								
		LGEXPPC		*LGSIZE*				
	n	$\widehat{\beta}_2$	s.e.$(\widehat{\beta}_2)$	$\widehat{\beta}_3$	s.e.$(\widehat{\beta}_3)$	R^2	F	*RSS*
ADM	2,815	1.080	0.033	−0.055	0.043	0.330	692.9	3,945.2
CLOT	4,500	0.842	0.024	0.146	0.032	0.240	710.1	5,766.1
DOM	1,661	0.941	0.054	0.415	0.075	0.157	154.6	4,062.5
EDUC	561	1.229	0.101	−0.437	0.139	0.311	125.9	1,380.1
ELEC	5,828	0.372	0.012	−0.362	0.017	0.359	1,627.8	2,636.3
FDAW	5,102	0.879	0.016	−0.213	0.022	0.461	2,176.6	3,369.1
FDHO	6,334	0.584	0.010	−0.081	0.013	0.432	2,410.8	1,988.4
FOOT	1,827	0.396	0.031	−0.560	0.042	0.281	356.1	1,373.5
FURN	487	0.807	0.103	−0.246	0.137	0.195	58.7	913.9
GASO	5,710	0.676	0.013	−0.004	0.018	0.362	1,691.8	2,879.3
HEAL	4,802	0.779	0.023	−0.306	0.031	0.272	894.6	6,062.5
HOUS	6,223	0.989	0.016	−0.140	0.021	0.467	2,729.5	4,825.6
LIFE	1,253	0.464	0.050	−0.461	0.065	0.154	113.4	1,559.2
LOCT	692	0.389	0.060	−0.396	0.086	0.138	54.9	1,075.1
MAPP	399	0.721	0.094	−0.264	0.123	0.206	51.5	576.8
PERS	3,817	0.824	0.020	−0.217	0.028	0.388	1,206.3	3,002.2
READ	2,287	0.764	0.034	−0.503	0.047	0.297	482.8	2,892.1
SAPP	1,037	0.467	0.048	−0.592	0.066	0.236	160.1	1,148.9
TELE	5,788	0.640	0.013	−0.222	0.018	0.386	1,816.3	3,055.1
TEXT	992	0.388	0.049	−0.713	0.067	0.246	161.0	1,032.9
TOB	1,155	0.563	0.037	−0.515	0.049	0.329	282.1	873.4
TOYS	2,504	0.638	0.031	−0.304	0.043	0.231	375.8	2,828.3
TRIP	516	0.681	0.083	−0.142	0.109	0.150	45.3	792.8

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.

Determine whether (2) is a reparameterised or a restricted version of (1).

(2) may be rewritten:

$$\log Y = \alpha_1 + (\alpha_2 + 1) \log X + u$$

so it is a reparameterised version of (1) with $\beta_1 = \alpha_1$ and $\beta_2 = \alpha_2 + 1$.

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

$$\widehat{y} = \widehat{\beta}_1 + \widehat{\beta}_2 x \qquad (3)$$

and:

$$\widehat{z} = \widehat{\alpha}_1 + \widehat{\alpha}_2 x. \qquad (4)$$

Using the expressions for the OLS regression coefficients, demonstrate that $\widehat{\beta}_2 = \widehat{\alpha}_2 + 1$.

$$\begin{aligned}\widehat{\alpha}_2 &= \frac{\sum(x_i-\bar{x})(z_i-\bar{z})}{\sum(x_i-\bar{x})^2} = \frac{\sum(x_i-\bar{x})([y_i-x_i]-[\bar{y}-\bar{x}])}{\sum(x_i-\bar{x})^2} \\ &= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} - \frac{\sum(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2} = \widehat{\beta}_2 - 1.\end{aligned}$$

Similarly, using the expressions for the OLS regression coefficients, demonstrate that $\widehat{\beta}_1 = \widehat{\alpha}_1$.

$$\widehat{\alpha}_1 = \bar{z} - \widehat{\alpha}_2\bar{x} = (\bar{y}-\bar{x}) - \widehat{\alpha}_2\bar{x} = \bar{y} - (\widehat{\alpha}_2+1)\bar{x} = \bar{y} - \widehat{\beta}_2\bar{x} = \widehat{\beta}_1.$$

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

$$\widehat{z}_i = \widehat{\alpha}_1 + \widehat{\alpha}_2 x_i = \widehat{\beta}_1 + (\widehat{\beta}_2 - 1)x_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i - x_i = \widehat{y}_i - x_i.$$

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

Let $\widehat{u}_i$ be the residual in (3) and $\widehat{v}_i$ the residual in (4). Then:

$$\widehat{v}_i = z_i - \widehat{z}_i = y_i - x_i - (\widehat{y}_i - x_i) = y_i - \widehat{y}_i = \widehat{u}_i.$$

Hence show that the standard errors of $\widehat{\beta}_2$ *and* $\widehat{\alpha}_2$ *are the same.*

The standard error of $\widehat{\beta}_2$ is:

$$\text{s.e.}(\widehat{\beta}_2) = \sqrt{\frac{\sum \widehat{u}_i^2/(n-2)}{\sum(x_i-\bar{x})^2}} = \sqrt{\frac{\sum \widehat{v}_i^2/(n-2)}{\sum(x_i-\bar{x})^2}} = \text{s.e.}(\widehat{\alpha}_2).$$

Determine the relationship between the t *statistic for* $\widehat{\beta}_2$ *and the* t *statistic for* $\widehat{\alpha}_2$*, and give an intuitive explanation for the relationship.*

$$t_{\widehat{\beta}_2} = \frac{\widehat{\beta}_2}{\text{s.e.}(\widehat{\beta}_2)} = \frac{\widehat{\alpha}_2 + 1}{\text{s.e.}(\widehat{\alpha}_2)}.$$

The t statistic for $\widehat{\beta}_2$ 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 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.

A4.5 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.

A4.6 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.)

The following Stata output gives the results of fitting the model for *FDHO*, assuming that both the dependent variable and the explanatory variables are subject to the Box–Cox transformation with the same value of λ. Iteration messages have been deleted. The maximum likelihood estimate of λ is 0.10, so the logarithmic specification is a better approximation than the linear specification. The latter is very soundly rejected by the likelihood-ratio test.

```
. boxcox FDHOPC EXPPC SIZE if FDHO>0, model(lambda)
                                                Number of obs   =       6334
                                                LR chi2(2)      =    3592.55
Log likelihood = -41551.328                     Prob > chi2     =      0.000

------------------------------------------------------------------------------
      FDHOPC |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     /lambda |   .1019402   .0117364     8.69   0.000     .0789372    .1249432
------------------------------------------------------------------------------
Estimates of scale-variant parameters
-----------------------------
             |      Coef.
-------------+---------------
Notrans      |
       _cons |   2.292828
-------------+---------------
Trans        |
       EXPPC |   .4608736
        SIZE |  -.1486856
-------------+---------------
      /sigma |   .9983288
-----------------------------
```

```
-------------------------------------------------------
   Test         Restricted     LR statistic      P-value
    H0:       log likelihood       chi2       Prob > chi2
-------------------------------------------------------
lambda = -1     -50942.835     18783.01           0.000
lambda =  0     -41590.144        77.63           0.000
lambda =  1     -44053.749      5004.84           0.000
-------------------------------------------------------
```

A4.7 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 estimate of β_4 is negative, at first sight suggesting that cognitive ability has an adverse effect on earnings, contrary to common sense and previous results with wage equations of this kind. However, rewriting the model as:

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

it can be seen that, as a consequence of the inclusion of the interactive term, β_4 represents the effect of a marginal year of schooling for an individual with no schooling. Since no individual in the sample had fewer than 8 years of schooling, 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 years of schooling, 14.9. This is $(-0.2096 + 0.0189 \times 14.9) = 0.072$, implying a much more plausible 7.2 per cent increase in earnings for each standard deviation increase in cognitive ability. The positive sign of the coefficient of SA suggests that schooling and cognitive ability have mutually reinforcing effects on earnings.

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 coefficients of S and $ASVABC$ now provide estimates of their effects when the other variable is equal to its sample mean.

```
. reg LGEARN S1 EXP ASVABC1 SASVABC1
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  4,   495) =   22.68
     Model |  23.6368304     4  5.90920759           Prob > F      =  0.0000
  Residual |  128.962389   495  .260530079           R-squared     =  0.1549
-----------+------------------------------           Adj R-squared =  0.1481
     Total |   152.59922   499   .30581006           Root MSE      =  .51042
------------------------------------------------------------------------------
    LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
        S1 |   .0815188   .0116521     7.00   0.000     .0586252    .1044125
       EXP |   .0400506   .0096479     4.15   0.000     .0210948    .0590065
   ASVABC1 |   .0715084   .0298278     2.40   0.017     .0129036    .1301132
  SASVABC1 |   .0188685   .0093393     2.02   0.044     .0005189    .0372181
     _cons |   2.544783   .0675566    37.67   0.000      2.41205    2.677516
------------------------------------------------------------------------------
```

A4.8 In the first part of the output, *WEIGHT11* is regressed on *HEIGHT*, using *EAWE* 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. Somewhat surprisingly, its coefficient is not significant. A logarithmic regression of *WEIGHT11* on *HEIGHT* yields an estimated elasticity of 2.05, significantly different from 1 at a high significance level. Multicollinearity is responsible for the failure to detect nonlinearity hear. *YHAT* is very highly correlated with *HEIGHT*.

```
. reg WEIGHT11 HEIGHT
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  1,   498) =  139.97
     Model |  236642.736     1  236642.736           Prob > F      =  0.0000
  Residual |  841926.912   498  1690.61629           R-squared     =  0.2194
-----------+------------------------------           Adj R-squared =  0.2178
     Total |  1078569.65   499  2161.46222           Root MSE      =  41.117
------------------------------------------------------------------------------
  WEIGHT11 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
    HEIGHT |   5.369246   .4538259    11.83   0.000     4.477597    6.260895
     _cons |  -184.7802    30.8406    -5.99   0.000    -245.3739   -124.1865
------------------------------------------------------------------------------


. predict YHAT
. gen YHATSQ = YHAT*YHAT

. reg WEIGHT11 HEIGHT YHATSQ
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =     500
-----------+------------------------------           F(  2,   497) =   70.33
     Model |  237931.888     2  118965.944           Prob > F      =  0.0000
  Residual |   840637.76   497  1691.42407           R-squared     =  0.2206
-----------+------------------------------           Adj R-squared =  0.2175
     Total |  1078569.65   499  2161.46222           Root MSE      =  41.127
------------------------------------------------------------------------------
  WEIGHT11 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
    HEIGHT |  -.4995924   6.737741    -0.07   0.941    -13.73756    12.73837
    YHATSQ |   .0030233    .003463     0.87   0.383    -.0037807    .0098273
     _cons |   114.5523   344.2538     0.33   0.739    -561.8199    790.9244
------------------------------------------------------------------------------
```