

Chapter 1

Simple regression analysis

Overview

This chapter introduces the least squares criterion of goodness of fit and demonstrates, first through examples and then in the general case, how it may be used to develop expressions for the coefficients that quantify the relationship when a dependent variable is assumed to be determined by one explanatory variable. The chapter continues by showing how the coefficients should be interpreted when the variables are measured in natural units, and it concludes by introducing R^2 , a second criterion of goodness of fit, and showing how it is related to the least squares criterion and the correlation between the fitted and actual values of the dependent variable.

Further material

Definition of cross-sectional, time series, and panel data

Data sets may be of three types: cross-sectional data, time series data, and panel data. Cross-sectional data sets are generated at one moment in time and the observations generally relate to households, individuals, enterprises, or geographical areas. It is usually very desirable that the sample should be drawn from a well-defined population using a statistically respectable sampling scheme, so that one may generalize from the results. For example, the Consumer Expenditure Survey data set, whose observations relate to households, employs a sampling scheme that makes it nationally representative of the United States.

Time series data sets consist of repeated observations on a set of variables over an interval of time. Generally the interval between the observations is fixed, often being a year, a quarter, or a month, but in some cases, such as analysis using stock market prices, the frequency may be much greater. An example is the Demand Functions data set, used for examples and exercises in Chapters 11–13 of the text. It consists of 45 annual observations over the period 1959–2003 on aggregate expenditure on different categories of consumer expenditure, the prices of the categories, and aggregate disposable income.

Panel data sets have both cross-sectional and time series dimensions, being repeated observations over an interval of time on the same cross-section sample. The National Longitudinal Survey of Youth is an example. The 6,000 males and females in the core sample were first interviewed in 1979. From then until 1994 they were reinterviewed every year, and since 1994 every two years. The *EAEF* data sets derived from it and used as examples and exercises in the first part of the text are effectively cross-sectional data sets because the earnings data relate to only one year, 2002, and the other characteristics, such as educational attainment, are mostly fixed.

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 guide, you should be able to explain what is meant by:

- the least squares criterion of goodness of fit
- the nonstochastic component of a true relationship
- the disturbance term
- the regression line
- the residuals

- R^2 .

In addition, you should be able to explain the difference between

- the nonstochastic component of a true relationship and a fitted regression line, and
- the values of the disturbance term and the residuals.

Additional exercises

A1.1 The output below shows the result of regressing the number of children in the family on the years of schooling of the mother, using *EAEF* Data Set 22. Provide an interpretation of the coefficients.

```
. g CHILDREN = SIBLINGS + 1
. reg CHILDREN SM
```

Source	SS	df	MS			
Model	171.813775	1	171.813775	Number of obs =	540	
Residual	2205.43622	538	4.09932384	F(1, 538) =	41.91	
				Prob > F =	0.0000	
				R-squared =	0.0723	
				Adj R-squared =	0.0705	
Total	2377.25	539	4.41048237	Root MSE =	2.0247	

CHILDREN	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
SM	-.215508	.0332882	-6.47	0.000	-.2808988	-.1501172
_cons	6.750292	.3959112	17.05	0.000	5.972571	7.528013

A1.2 The output below gives the result of regressing *FDHO*, annual household expenditure on food consumed at home, on *EXP*, total annual household expenditure, both measured in dollars, using the Consumer Expenditure Survey data set. Give an interpretation of the coefficients.

```
. reg FDHO EXP if FDHO>0
```

Source	SS	df	MS			
Model	911005795	1	911005795	Number of obs =	868	
Residual	2.0741e+09	866	2395045.39	F(1, 866) =	380.37	
				Prob > F =	0.0000	
				R-squared =	0.3052	
				Adj R-squared =	0.3044	
Total	2.9851e+09	867	3443039.33	Root MSE =	1547.6	

FDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
EXP	.0527204	.0027032	19.50	0.000	.0474149	.058026
_cons	1922.939	96.50688	19.93	0.000	1733.525	2112.354

A1.3 Download the *CES* data set from the website (see Appendix B of the text), perform a regression parallel to that in Exercise A1.2 for your category of expenditure, and provide an interpretation of the regression coefficients.

A1.4 The output below shows the result of regressing the weight of the respondent, in pounds, in 2002 on the weight in 1985, using *EAEF* Data Set 22. Provide an interpretation of the coefficients. Summary statistics for the data are also provided.

Dougherty: Introduction to Econometrics 3e
Study Guide

```
. reg WEIGHT02 WEIGHT85
```

Source	SS	df	MS			
Model	620662.43	1	620662.43	Number of obs =	540	
Residual	290406.035	538	539.788169	F(1, 538) =	1149.83	
Total	911068.465	539	1690.294	Prob > F =	0.0000	
				R-squared =	0.6812	
				Adj R-squared =	0.6807	
				Root MSE =	23.233	

WEIGHT02	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
WEIGHT85	1.013353	.0298844	33.91	0.000	.9546483	1.072057
_cons	23.61869	4.760179	4.96	0.000	14.26788	32.96951

```
. sum WEIGHT85 WEIGHT02
```

Variable	Obs	Mean	Std. Dev.	Min	Max
WEIGHT85	540	155.7333	33.48673	89	300
WEIGHT02	540	181.4315	41.11319	103	400

A1.5 The output shows the result of regressing the hourly earnings of the respondent, in dollars, in 2002 on height in 1985, measured in inches, using *EAEF* Data Set 22. Provide an interpretation of the coefficients, comment on the plausibility of the interpretation, and attempt to give an explanation.

```
. reg EARNINGS HEIGHT
```

Source	SS	df	MS			
Model	6236.81652	1	6236.81652	Number of obs =	540	
Residual	105773.415	538	196.60486	F(1, 538) =	31.72	
Total	112010.231	539	207.811189	Prob > F =	0.0000	
				R-squared =	0.0557	
				Adj R-squared =	0.0539	
				Root MSE =	14.022	

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
HEIGHT	.8025732	.1424952	5.63	0.000	.522658	1.082488
_cons	-34.67718	9.662091	-3.59	0.000	-53.65723	-15.69713

A1.6 A researcher has data for 50 countries on N , the average number of newspapers purchased per adult in one year, and G , GDP per capita, measured in US \$, and fits the following regression (RSS = residual sum of squares)

$$\hat{N} = 25.0 + 0.020 G \quad R^2 = 0.06, RSS = 4,000.0$$

The researcher realizes that GDP has been underestimated by \$100 in every country and that N should have been regressed on G^* , where $G^* = G + 100$. Explain, with mathematical proofs, how the following components of the output would have differed:

- the coefficient of GDP
- the intercept
- RSS
- R^2

A1.7 A researcher with the same model and data as in Exercise A1.6 believes that GDP in each country has been underestimated by 50 percent and that N should have been regressed on G^* , where $G^* = 2G$. Explain, with mathematical proofs, how the following components of the output would have differed:

- the coefficient of GDP
- the intercept
- RSS
- R^2

Answers to the starred exercises in the text

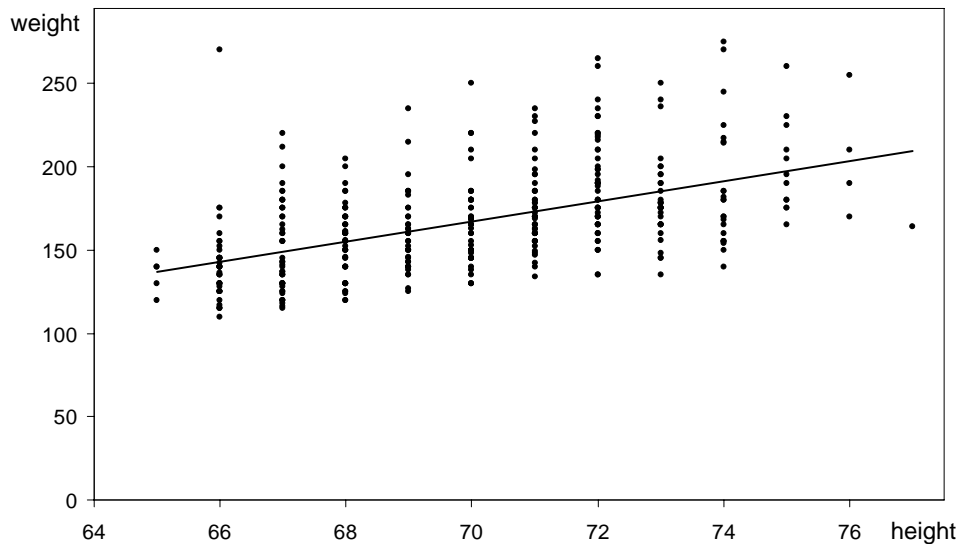
1.5 The output below shows the result of regressing the weight of the respondent in 1985, measured in pounds, on his or her height, measured in inches, using *EAEF* Data Set 21. Provide an interpretation of the coefficients.

```
. reg WEIGHT85 HEIGHT
```

Source	SS	df	MS			
Model	261111.383	1	261111.383	Number of obs =	540	
Residual	394632.365	538	733.517407	F(1, 538) =	355.97	
Total	655743.748	539	1216.59322	Prob > F =	0.0000	
				R-squared =	0.3982	
				Adj R-squared =	0.3971	
				Root MSE =	27.084	

WEIGHT85	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
HEIGHT	5.192973	.275238	18.87	0.000	4.6523	5.733646
_cons	-194.6815	18.6629	-10.43	0.000	-231.3426	-158.0204

Answer: Literally the regression implies that, for every extra inch of height, an individual tends to weigh an extra 5.2 pounds. The intercept, which literally suggests that an individual with no height would weigh -195 pounds, has no meaning. The figure shows the observations and the fitted regression line.



1.7 Derive, with a proof, the slope coefficient that would have been obtained in Exercise 1.5 if weight and height had been measured in metric units. (*Note:* one pound is 454 grams and one inch is 2.54 cm.)

Answer: Let the weight and height be W and H in imperial units (pounds and inches) and WM and HM in metric units (kilos and centimetres). Then $WM = 0.454W$ and $HM = 2.54H$. The slope coefficient for the regression in metric units, b_2^M , is given by

$$b_2^M = \frac{\sum (HM_i - \overline{HM})(WM_i - \overline{WM})}{\sum (HM_i - \overline{HM})^2} = \frac{\sum 2.54(H_i - \overline{H})0.454(W_i - \overline{W})}{\sum 2.54^2(H_i - \overline{H})^2}$$

$$= 0.179 \frac{\sum (H_i - \overline{H})(W_i - \overline{W})}{\sum (H_i - \overline{H})^2} = 0.179b_2 = 0.929$$

In other words, weight increases at the rate of almost one kilo per centimetre. The regression output below confirms that the calculations are correct (subject to rounding error in the last digit).

```
. g WM = 0.454*WEIGHT85
. g HM = 2.54*HEIGHT
. reg WM HM
```

	Source	SS	df	MS		
	Model	53819.2324	1	53819.2324	Number of obs =	540
	Residual	81340.044	538	151.189673	F(1, 538) =	355.97
	Total	135159.276	539	250.759325	Prob > F =	0.0000
					R-squared =	0.3982
					Adj R-squared =	0.3971
					Root MSE =	12.296

	WM	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
	HM	.9281928	.0491961	18.87	0.000	.8315528 1.024833
	_cons	-88.38539	8.472958	-10.43	0.000	-105.0295 -71.74125

- 1.8 A researcher has data on the aggregate expenditure on services, Y , and aggregate disposable personal income, X , both measured in \$ billion at constant prices, for each of the US states and fits the equation

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$

The researcher initially fits the equation using OLS regression analysis. However, suspecting that tax evasion causes both Y and X to be substantially underestimated, the researcher adopts two alternative methods of compensating for the under-reporting:

1. The researcher adds \$90 billion to the data for Y in each state and \$200 billion to the data for X .
2. The researcher increases the figures for both Y and X in each state by 10 percent.

Evaluate the impact of the adjustments on the regression results.

Answer:

First adjustment: Let the adjusted data be Y^* and X^* where $Y^* = Y + 90$ and $X^* = X + 200$. Then

$$b_2^* = \frac{\sum (X_i^* - \overline{X}^*)(Y_i^* - \overline{Y}^*)}{\sum (X_i^* - \overline{X}^*)^2} = \frac{\sum ([X_i + 200] - [\overline{X} + 200])([Y_i + 90] - [\overline{Y} + 90])}{\sum ([X_i + 200] - [\overline{X} + 200])^2}$$

$$= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = b_2$$

Hence the new slope coefficient will be identical to the original one. The constant will be affected since

$$b_1^* = \overline{Y}^* - b_2^* \overline{X}^* = (\overline{Y} + 90) - b_2(\overline{X} + 200) = \overline{Y} - b_2 \overline{X} + 90 - 200b_2 = b_1 + 90 - 200b_2$$

Second adjustment: Let the adjusted data again be Y^* and X^* where $Y^* = 1.1Y$ and $X^* = 1.1X$. Then

$$b_2^* = \frac{\sum (X_i^* - \bar{X}^*)(Y_i^* - \bar{Y}^*)}{\sum (X_i^* - \bar{X}^*)^2} = \frac{\sum (1.1X_i - 1.1\bar{X})(1.1Y_i - 1.1\bar{Y})}{\sum (1.1X_i - 1.1\bar{X})^2}$$

$$= \frac{1.21 \sum (X_i - \bar{X})(Y_i - \bar{Y})}{1.21 \sum (X_i - \bar{X})^2} = b_2$$

Hence the new slope coefficient will again be identical to the old one. The constant is affected since

$$b_1^* = \bar{Y}^* - b_2^* \bar{X}^* = 1.1\bar{Y} - b_2(1.1\bar{X}) = 1.1b_1$$

and so it is increased by 10 percent.

- 1.9 A researcher has international cross-sectional data on aggregate wages, W , aggregate profits, P , and aggregate income, Y , for a sample of n countries. By definition,

$$Y_i = W_i + P_i.$$

The regressions

$$\hat{W}_i = a_1 + a_2 Y_i$$

$$\hat{P}_i = b_1 + b_2 Y_i$$

are fitted using OLS regression analysis. Show that the regression coefficients will automatically satisfy the following equations:

$$a_2 + b_2 = 1$$

$$a_1 + b_1 = 0.$$

Explain intuitively why this should be so.

Answer:

$$a_2 + b_2 = \frac{\sum (Y_i - \bar{Y})(W_i - \bar{W})}{\sum (Y_i - \bar{Y})^2} + \frac{\sum (Y_i - \bar{Y})(P_i - \bar{P})}{\sum (Y_i - \bar{Y})^2}$$

$$= \frac{\sum (Y_i - \bar{Y})(W_i + P_i - \bar{W} - \bar{P})}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (Y_i - \bar{Y})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} = 1$$

$$a_1 + b_1 = (\bar{W} - a_2 \bar{Y}) + (\bar{P} - b_2 \bar{Y}) = (\bar{W} + \bar{P}) - (a_2 + b_2) \bar{Y} = \bar{Y} - \bar{Y} = 0$$

The intuitive explanation is that the regressions break down income into predicted wages and profits and one would expect the sum of the predicted components of income to be equal to its actual level. The sum of the predicted components is $[(a_1 + a_2 Y) + (b_1 + b_2 Y)]$, and in general this will be equal to Y only if the two conditions are satisfied.

- 1.10 Derive from first principles the least squares estimator of β_2 in the model

$$Y_i = \beta_2 X_i + u_i$$

First define RSS , the sum of the squares of the residuals, and then differentiate.

Answer: Let the fitted line be $\hat{Y}_i = b_2 X_i$. Then $e_i = Y_i - b_2 X_i$ and

$$\begin{aligned}
 RSS &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - b_2 X_i)^2 = \sum_{i=1}^n (Y_i^2 - 2b_2 X_i Y_i + b_2^2 X_i^2) \\
 &= \sum_{i=1}^n Y_i^2 - 2b_2 \sum_{i=1}^n X_i Y_i + b_2^2 \sum_{i=1}^n X_i^2
 \end{aligned}$$

The first-order condition for a minimum is

$$\frac{dRSS}{db_2} = -2 \sum_{i=1}^n X_i Y_i + 2b_2 \sum_{i=1}^n X_i^2 = 0$$

and so

$$b_2 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

The second derivative is positive, confirming that we have found a minimum.

- 1.16 The output shows the result of regressing weight in 2002 on height, using *EAEF* Data Set 21. In 2002 the respondents were aged 37–44. Explain why R^2 is lower than in the regression reported in Exercise 1.5.

```
. reg WEIGHT02 HEIGHT
```

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

Answer: The explained sum of squares (described as the model sum of squares in the Stata output) is actually higher than that in Exercise 1.5. The reason for the fall in R^2 is the huge increase in the total sum of squares, no doubt caused by the cumulative effect of diversity in eating habits.

- 1.17 The useful results in Box 1.2 are in general no longer valid if the model does not contain an intercept. Demonstrate, in particular, that \bar{e} will not in general be equal to zero.

Answer:

$$e_i = Y_i - \frac{\sum X_i Y_i}{\sum X_i^2} X_i$$

Hence

$$\bar{e} = \bar{Y} - \frac{\sum X_i Y_i}{\sum X_i^2} \bar{X}$$

There is no reason why the right side of the equation should be zero and one can provide an empirical counterexample by running a regression. The following output shows the result of regressing earnings on schooling, using *EAEF* Data Set 21, with no intercept. The residuals are saved as *E*. The mean value of *E* is -0.43 .

```
. reg EARNINGS S, nocon
```

Source	SS	df	MS	Number of obs =	540
Model	224309.221	1	224309.221	F(1, 539) =	1260.52
Residual	95914.8706	539	177.949667	Prob > F =	0.0000
Total	320224.092	540	593.007577	R-squared =	0.7005
				Adj R-squared =	0.6999
				Root MSE =	13.34

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
S	1.467575	.0413357	35.50	0.000	1.386376 1.548773

```
. predict E, resid
. sum E
```

Variable	Obs	Mean	Std. Dev.	Min	Max
E	540	-.4287836	13.33287	-20.26634	96.70881

Answers to the additional exercises

- A1.1 The slope coefficient indicates that every extra year of schooling reduces the number of children by 0.22. Literally, the intercept indicates that a mother with no schooling would have 6.75 children, but it is dangerous to extrapolate this far since few mothers had no schooling. The results are consistent with the literature.
- A1.2 Expenditure on food consumed at home increases by 5.3 cents for each dollar of total household expenditure. Literally the intercept implies that \$1,923 would be spent on food consumed at home if total household expenditure were zero. Obviously, such an interpretation does not make sense. If the explanatory variable were income, and household income were zero, positive expenditure on food at home would still be possible if the household received food stamps or other transfers. But here the explanatory variable is total household expenditure.
- A1.3 Housing has the largest coefficient, followed perhaps surprisingly by food consumed away from home, and then clothing. All the slope coefficients are highly significant, with the exception of local public transportation. Its slope coefficient is 0.0008, with *t* statistic 0.40, indicating that this category of expenditure is on the verge of being an inferior good.

	<i>EXP</i>				
	<i>n</i>	<i>b</i> ₂	<i>s.e.</i> (<i>b</i> ₂)	<i>R</i> ²	<i>F</i>
<i>FDHO</i>	868	0.0527	0.0027	0.3052	380.4
<i>FDAW</i>	827	0.0440	0.0021	0.3530	450.0
<i>HOUS</i>	867	0.1935	0.0063	0.5239	951.9
<i>TELE</i>	858	0.0101	0.0009	0.1270	124.6
<i>DOM</i>	454	0.0225	0.0043	0.0581	27.9
<i>TEXT</i>	482	0.0049	0.0006	0.1119	60.5
<i>FURN</i>	329	0.0128	0.0023	0.0844	30.1
<i>MAPP</i>	244	0.0089	0.0018	0.0914	24.3
<i>SAPP</i>	467	0.0013	0.0003	0.0493	24.1
<i>CLOT</i>	847	0.0395	0.0018	0.3523	459.5
<i>FOOT</i>	686	0.0034	0.0003	0.1575	127.9
<i>GASO</i>	797	0.0230	0.0014	0.2528	269.0
<i>TRIP</i>	309	0.0240	0.0038	0.1128	39.0
<i>LOCT</i>	172	0.0008	0.0019	0.0009	0.2
<i>HEAL</i>	821	0.0226	0.0029	0.0672	59.0
<i>ENT</i>	824	0.0700	0.0040	0.2742	310.6
<i>FEES</i>	676	0.0306	0.0026	0.1667	134.8
<i>TOYS</i>	592	0.0090	0.0010	0.1143	76.1
<i>READ</i>	764	0.0039	0.0003	0.1799	167.2
<i>EDUC</i>	288	0.0265	0.0054	0.0776	24.1
<i>TOB</i>	368	0.0071	0.0014	0.0706	27.8

- A1.4 The summary data indicate that, on average, the respondents put on 25.7 pounds over the period 1985–2002. Was this due to the relatively heavy becoming even heavier, or to a general increase in weight? The regression output indicates that weight in 2002 was approximately equal to weight in 1985 plus 23.6 pounds, so the second explanation appears to be the correct one. Note that this is an instance where the constant term can be given a meaningful interpretation and where it is as of much interest as the slope coefficient. The R^2 indicates that 1985 weight accounts for 68 percent of the variance in 2002 weight, so other factors are important.
- A1.5 The slope coefficient indicates that hourly earnings increase by 80 cents for every extra inch of height. The negative intercept has no possible interpretation. The interpretation of the slope coefficient is obviously highly implausible, so we know that something must be wrong with the model. The explanation is that this is a very poorly specified earnings function and that, in particular, we are failing to control for the sex of the respondent. Later on, in Chapter 5, we will find that males earn more than females, controlling for observable characteristics. Males also tend to be taller. Hence we find an apparent positive association between earnings and height in a simple regression. Note that R^2 is very low.
- A1.6 *The coefficient of GDP:* Let the revised measure of GDP be denoted G^* , where $G^* = G + 100$. Since $G_i^* = G_i + 100$ for all i , $\bar{G}^* = \bar{G} + 100$ and so $G_i^* - \bar{G}^* = G_i - \bar{G}$ for all i . Hence the new slope coefficient is

$$b_2^* = \frac{\sum(G_i^* - \bar{G}^*)(N_i - \bar{N})}{\sum(G_i^* - \bar{G}^*)^2} = \frac{\sum(G_i - \bar{G})(N_i - \bar{N})}{\sum(G_i - \bar{G})^2} = b_2.$$

The coefficient is unchanged.

The intercept: The new intercept is $b_1^* = \bar{N} - b_2^* \bar{G}^* = \bar{N} - b_2(\bar{G} + 100) = b_1 - 100b_2 = 23.0$.

RSS: The residual in observation i in the new regression, e_i^* , is given by

$$e_i^* = N_i - b_1^* - b_2^* G_i^* = N_i - (b_1 - 100b_2) - b_2(G_i + 100) = e_i,$$

the residual in the original regression. Hence *RSS* is unchanged.

R^2 : $R^2 = 1 - \frac{RSS}{\sum(N_i - \bar{N})^2}$ and is unchanged since *RSS* and $\sum(N_i - \bar{N})^2$ are unchanged. Note that this

makes sense intuitively. R^2 is unit-free and so it is not possible for the overall fit of a relationship to be affected by the units of measurement.

A1.7 *The coefficient of GDP:* Let the revised measure of GDP be denoted G^* , where $G^* = 2G$. Since $G_i^* = 2G_i$ for all i , $\bar{G}^* = 2\bar{G}$ and so $G_i^* - \bar{G}^* = 2(G_i - \bar{G})$ for all i . Hence the new slope coefficient is

$$\begin{aligned} b_2^* &= \frac{\sum(G_i^* - \bar{G}^*)(N_i - \bar{N})}{\sum(G_i^* - \bar{G}^*)^2} = \frac{\sum 2(G_i - \bar{G})(N_i - \bar{N})}{\sum 4(G_i - \bar{G})^2} \\ &= \frac{2\sum(G_i - \bar{G})(N_i - \bar{N})}{4\sum(G_i - \bar{G})^2} = \frac{b_2}{2} = 0.010 \end{aligned}$$

where $b_2 = 0.020$ is the slope coefficient in the original regression.

The intercept: The new intercept is $b_1^* = \bar{N} - b_2^* \bar{G}^* = \bar{N} - \frac{b_2}{2} 2\bar{G} = \bar{N} - b_2 \bar{G} = b_1 = 25.0$, the original intercept.

RSS: The residual in observation i in the new regression, e_i^* , is given by

$$e_i^* = N_i - b_1^* - b_2^* G_i^* = N_i - b_1 - \frac{b_2}{2} 2G_i = e_i$$

the residual in the original regression. Hence *RSS* is unchanged.

R^2 : $R^2 = 1 - \frac{RSS}{\sum(N_i - \bar{N})^2}$ and is unchanged since *RSS* and $\sum(N_i - \bar{N})^2$ are unchanged. As in Exercise

A1.6, this makes sense intuitively.