Chapter 1
Simple regression analysis

1.1 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.

1.2 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 what is meant by:

- dependent variable
- explanatory variable (independent variable, regressor)
- parameter of a regression model
- the nonstochastic component of a true relationship
- the disturbance term
- the least squares criterion of goodness of fit
- ordinary least squares (OLS)
- the regression line
- fitted model
- fitted values (of the dependent variable)
- residuals
- total sum of squares, explained sum of squares, residual sum of squares
- 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.

1.3 Additional exercises

A1.1 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              Number of obs =    6334
-------------+------------------------------           F(  1,  6332) = 3431.01
       Model |   972602566     1   972602566           Prob > F      =  0.0000
    Residual |  1.7950e+09  6332  283474.003           R-squared     =  0.3514
-------------+------------------------------           Adj R-squared =  0.3513
       Total |  2.7676e+09  6333   437006.15           Root MSE      =  532.42

------------------------------------------------------------------------------
        FDHO |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0627099   .0010706    58.57   0.000     .0606112    .0648086
       _cons |   369.4418   10.65718    34.67   0.000     348.5501    390.3334
------------------------------------------------------------------------------
```

A1.2 Download the *CES* data set from the website (see Appendix B of the text), perform a regression parallel to that in Exercise A1.1 for your category of expenditure, and provide an interpretation of the regression coefficients.

A1.3 The output shows the result of regressing the weight of the respondent, in pounds, in 2011 on the weight in 2004, using *EAWE* Data Set 22. Provide an interpretation of the coefficients. Summary statistics for the data are also provided.

```
. reg WEIGHT11 WEIGHT04

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) = 1207.55
       Model |  769248.875     1  769248.875           Prob > F      =  0.0000
    Residual |  317241.693   498  637.031513           R-squared     =  0.7080
-------------+------------------------------           Adj R-squared =  0.7074
       Total |  1086490.57   499  2177.33581           Root MSE      =  25.239
------------------------------------------------------------------------------
    WEIGHT11 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    WEIGHT04 |   .9739736   .0280281    34.75   0.000     .9189056    1.029042
       _cons |   17.42232   4.888091     3.56   0.000     7.818493    27.02614
------------------------------------------------------------------------------
```

```
. sum WEIGHT04 WEIGHT11

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    WEIGHT04 |       500     169.686    40.31215         95        330
    WEIGHT11 |       500     182.692    46.66193         95        370
```

A1.4 The output shows the result of regressing the hourly earnings of the respondent, in dollars, in 2011 on height in 2004, measured in inches, using *EAWE* 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              Number of obs =     500
-------------+------------------------------           F(  1,   498) =    9.23
       Model |  1393.77592     1  1393.77592           Prob > F      =  0.0025
    Residual |  75171.3726   498  150.946531           R-squared     =  0.0182
-------------+------------------------------           Adj R-squared =  0.0162
       Total |  76565.1485   499  153.437171           Root MSE      =  12.286

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HEIGHT |   .4087231   .1345068     3.04   0.003     .1444523    .6729938
       _cons |   -9.26923   9.125089    -1.02   0.310    -27.19765    8.659188
------------------------------------------------------------------------------
```

A1.5 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):

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

The researcher realises 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.

A1.6 A researcher with the same model and data as in Exercise A1.5 believes that GDP in each country has been underestimated by 50 per cent 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.

A1.7 Some practitioners of econometrics advocate 'standardising' each variable in a regression by subtracting its sample mean and dividing by its sample standard deviation. Thus, if the original regression specification is:

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

the revised specification is:

$$Y_i^* = \beta_1^* + \beta_2^* X_i^* + v_i$$

where:

$$Y_i^* = \frac{Y_i - \overline{Y}}{\widehat{\sigma}_Y} \qquad \text{and} \qquad X_i^* = \frac{X_i - \overline{X}}{\widehat{\sigma}_X}$$

$\overline{Y}$ and $\overline{X}$ are the sample means of Y and X, $\widehat{\sigma}_Y$ and $\widehat{\sigma}_X$ are the estimators of the standard deviations of Y and X, defined as the square roots of the estimated variances:

$$\widehat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad \text{and} \qquad \widehat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

and n is the number of observations in the sample. We will write the fitted models for the two specifications as:

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_i$$

and:

$$\widehat{Y}_i^* = \widehat{\beta}_1^* + \widehat{\beta}_2^* X_i^*.$$

Taking account of the definitions of Y^* and X^*, show that $\widehat{\beta}_1^* = 0$ and that $\widehat{\beta}_2^* = \frac{\widehat{\sigma}_X}{\widehat{\sigma}_Y}\widehat{\beta}_2$. Provide an interpretation of $\widehat{\beta}_2^*$.

A1.8 For the model described in Exercise A1.7, suppose that Y^* is regressed on X^* without an intercept:

$$\widehat{Y}_i^* = \widehat{\beta}_2^{**} X_i^*.$$

Determine how $\widehat{\beta}_2^{**}$ is related to $\widehat{\beta}_2^*$.

A1.9 A variable Y_i is generated as:

$$Y_i = \beta_1 + u_i \tag{1.1}$$

where β_1 is a fixed parameter and u_i is a disturbance term that is independently and identically distributed with expected value 0 and population variance σ_u^2. The least squares estimator of β_1 is $\overline{Y}$, the sample mean of Y. Give a mathematical demonstration that the value of R^2 in such a regression is zero.

1.4 Answers to the starred exercises in the textbook

1.9 The output shows the result of regressing the weight of the respondent in 2004, measured in pounds, on his or her height, measured in inches, using *EAWE* Data Set 21. Provide an interpretation of the coefficients.

```
. reg WEIGHT04 HEIGHT

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  176.74
       Model |      211309     1      211309           Prob > F      =  0.0000
    Residual |   595389.95   498  1195.56215           R-squared     =  0.2619
-------------+------------------------------           Adj R-squared =  0.2605
       Total |   806698.98   499  1616.63116           Root MSE      =  34.577

------------------------------------------------------------------------------
    WEIGHT04 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      HEIGHT |   5.073711    .381639    13.29   0.000      4.32389    5.823532
       _cons |  -177.1703   25.93501    -6.83   0.000    -228.1258   -126.2147
------------------------------------------------------------------------------
```

Answer:

Literally the regression implies that, for every extra inch of height, an individual tends to weigh an extra 5.1 pounds. The intercept, which literally suggests that an individual with no height would weigh -177 pounds, has no meaning.

1.11 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:

$$\begin{aligned} \widehat{W}_i &= \widehat{\alpha}_1 + \widehat{\alpha}_2 Y_i \\ \widehat{P}_i &= \widehat{\beta}_1 + \widehat{\beta}_2 Y_i \end{aligned}$$

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

$$\begin{aligned} \widehat{\alpha}_2 + \widehat{\beta}_2 &= 1 \\ \widehat{\alpha}_1 + \widehat{\beta}_1 &= 0. \end{aligned}$$

Explain intuitively why this should be so.

Answer:

$$\begin{aligned} \widehat{\alpha}_2 + \widehat{\beta}_2 &= \frac{\sum\left(Y_i - \overline{Y}\right)\left(W_i - \overline{W}\right)}{\sum\left(Y_i - \overline{Y}\right)^2} + \frac{\sum\left(Y_i - \overline{Y}\right)\left(P_i - \overline{P}\right)}{\sum\left(Y_i - \overline{Y}\right)^2} \\ &= \frac{\sum\left(Y_i - \overline{Y}\right)\left(W_i + P_i - \overline{W} - \overline{P}\right)}{\sum\left(Y_i - \overline{Y}\right)^2} \\ &= \frac{\sum\left(Y_i - \overline{Y}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(Y_i - \overline{Y}\right)^2} \\ &= 1 \end{aligned}$$

$$\widehat{\alpha}_1 + \widehat{\beta}_1 = \left(\overline{W} - \widehat{\alpha}_2\overline{Y}\right) + \left(\overline{P} - \widehat{\beta}_2\overline{Y}\right) = \left(\overline{W} + \overline{P}\right) - (\widehat{\alpha}_2 + \widehat{\beta}_2)\overline{Y} = \overline{Y} - \overline{Y} = 0.$$

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 $\widehat{W}_i + \widehat{P}_i = (\widehat{\alpha}_1 + \widehat{\alpha}_2 Y_i) + (\widehat{\beta}_1 + \widehat{\beta}_2 Y_i)$, and in general this will be equal to Y_i only if the two conditions are satisfied.

1.13 Suppose that the units of measurement of X are changed so that the new measure, X^*, is related to the original one by $X_i^* = \mu_2 X_i$. Show that the new estimate of the slope coefficient is $\widehat{\beta}_2/\mu_2$, where $\widehat{\beta}_2$ is the slope coefficient in the original regression.

Answer:

$$\begin{aligned}
\widehat{\beta}_2^* &= \frac{\sum\left(X_i^* - \overline{X}^*\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i^* - \overline{X}^*\right)^2} \\
&= \frac{\sum\left(\mu_2 X_i - \mu_2\overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(\mu_2 X_i - \mu_2\overline{X}\right)^2} \\
&= \frac{\mu_2\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\mu_2^2\sum\left(X_i - \overline{X}\right)^2} \\
&= \frac{\widehat{\beta}_2}{\mu_2}.
\end{aligned}$$

1.14 Demonstrate that if X is demeaned but Y is left in its original units, the intercept in a regression of Y on demeaned X will be equal to $\overline{Y}$.

Answer:

Let $X_i^* = X_i - \overline{X}$ and $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ be the intercept and slope coefficient in a regression of Y on X^*. Note that $\overline{X}^* = 0$. Then:

$$\widehat{\beta}_1^* = \overline{Y} - \widehat{\beta}_2^*\overline{X}^* = \overline{Y}.$$

The slope coefficient is not affected by demeaning:

$$\widehat{\beta}_2^* = \frac{\sum\left(X_i^* - \overline{X}^*\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i^* - \overline{X}^*\right)^2} = \frac{\sum\left([X_i - \overline{X}] - 0\right)\left(Y_i - \overline{Y}\right)}{\sum\left([X_i - \overline{X}] - 0\right)^2} = \widehat{\beta}_2.$$

1.15 The regression output shows the result of regressing weight on height using the same sample as in Exercise 1.9, but with weight and height measured in kilos and centimetres: $WMETRIC = 0.454 * WEIGHT04$ and $HMETRIC = 2.54 * HEIGHT$. Confirm that the estimates of the intercept and slope coefficient are as should be expected from the changes in the units of measurement.

```
. gen WTMETRIC = 0.454*WEIGHT04
. gen HMETRIC = 2.54*HEIGHT

. reg WTMETRIC HMETRIC

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  176.74
       Model |  43554.1641     1  43554.1641           Prob > F      =  0.0000
    Residual |  122719.394   498  246.424486           R-squared     =  0.2619
-------------+------------------------------           Adj R-squared =  0.2605
       Total |  166273.558   499  333.213544           Root MSE      =  15.698

------------------------------------------------------------------------------
     WMETRIC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HMETRIC |   .9068758   .0682142    13.29   0.000     .7728527    1.040899
       _cons |  -80.43529   11.77449    -6.83   0.000    -103.5691   -57.30148
------------------------------------------------------------------------------
```

Answer:

Abbreviate *WEIGHT04* to W, *HEIGHT* to H, *WMETRIC* to WM, and *HMETRIC* to HM. $WM = 0.454W$ and $HM = 2.54H$. The slope coefficient and intercept for the regression in metric units, $\widehat{\beta}_2^M$ and $\widehat{\beta}_1^M$, are then given by:

$$
\begin{aligned}
\widehat{\beta}_2^{\mathrm{M}} &= \frac{\sum\left(HM_i - \overline{HM}\right)\left(WM_i - \overline{WM}\right)}{\sum\left(HM_i - \overline{HM}\right)^2} \\
&= \frac{\sum 2.54\left(H_i - \overline{H}\right)0.454\left(W_i - \overline{W}\right)}{\sum 2.54^2\left(H_i - \overline{H}\right)^2} \\
&= 0.179\frac{\sum\left(H_i - \overline{H}\right)\left(W_i - \overline{W}\right)}{\sum\left(H_i - \overline{H}\right)^2} \\
&= 0.179\widehat{\beta}_2 \\
&= 0.179 \times 5.074 \\
&= 0.908 \\
\widehat{\beta}_1^M &= \overline{WM} - \widehat{\beta}_2^M\overline{HM} \\
&= 0.454\overline{W} - \left(\frac{0.454}{2.54}\widehat{\beta}_2\right)(2.54\overline{H}) \\
&= 0.454(\overline{W} - \widehat{\beta}_2\overline{H}) \\
&= 0.454\widehat{\beta}_1 \\
&= 0.454 \times -177.2 \\
&= -80.4.
\end{aligned}
$$

The regression output confirms that the calculations are correct (subject to rounding error in the last digit).

1.16 Consider the regression model:

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

It implies:

$$\overline{Y} = \beta_1 + \beta_2 \overline{X} + \bar{u}$$

and hence that:

$$Y_i^* = \beta_2 X_i^* + v_i$$

where $Y_i^* = Y_i - \overline{Y}$, $X_i^* = X_i - \overline{X}$ and $v_i = u_i - \bar{u}$.

Demonstrate that a regression of Y^* on X^* using (1.49) will yield the same estimate of the slope coefficient as a regression of Y on X. Note: (1.49) should be used instead of (1.35) because there is no intercept in this model.

Evaluate the outcome if the slope coefficient were estimated using (1.35), despite the fact that there is no intercept in the model.

Determine the estimate of the intercept if Y^* were regressed on X^* with an intercept included in the regression specification.

Answer:

Let $\widehat{\beta}_2^*$ be the slope coefficient in a regression of Y^* on X^* using (1.49). Then:

$$\widehat{\beta}_2^* = \frac{\sum X_i^* Y_i^*}{\sum X_i^{*2}} = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} = \widehat{\beta}_2.$$

Let $\widehat{\beta}_2^{**}$ be the slope coefficient in a regression of Y^* on X^* using (1.35). Note that $\overline{Y}^*$ and $\overline{X}^*$ are both zero. Then:

$$\widehat{\beta}_2^{**} = \frac{\sum \left(X_i^* - \overline{X}^*\right)\left(Y_i^* - \overline{Y}^*\right)}{\sum \left(X_i^* - \overline{X}^*\right)^2} = \frac{\sum X_i^* Y_i^*}{\sum X_i^{*2}} = \widehat{\beta}_2.$$

Let $\widehat{\beta}_1^{**}$ be the intercept in a regression of Y^* on X^* using (1.35). Then:

$$\widehat{\beta}_1^{**} = \overline{Y}^* - \widehat{\beta}_2^{**}\overline{X}^* = 0.$$

1.18 Demonstrate that the fitted values of the dependent variable are uncorrelated with the residuals in a simple regression model. (This result generalises to the multiple regression case.)

Answer:

The numerator of the sample correlation coefficient for $\widehat{Y}$ and $\widehat{u}$ can be decomposed as follows, using the fact that $\overline{\widehat{u}} = 0$:

$$\begin{aligned}
\frac{1}{n}\sum \left(\widehat{Y}_i - \overline{\widehat{Y}}\right)\left(\widehat{u}_i - \overline{\widehat{u}}\right) &= \frac{1}{n}\sum \left([\widehat{\beta}_1 + \widehat{\beta}_2 X_i] - [\widehat{\beta}_1 + \widehat{\beta}_2 \overline{X}]\right)\widehat{u}_i \\
&= \frac{1}{n}\widehat{\beta}_2 \sum \left(X_i - \overline{X}\right)\widehat{u}_i \\
&= 0
\end{aligned}$$

by (1.65). Hence the correlation is zero.

1.23 Demonstrate that, in a regression with an intercept, a regression of Y on X^* must have the same R^2 as a regression of Y on X, where $X^* = \mu_2 X$.

Answer:

Let the fitted regression of Y on X^* be written $\widehat{Y}_i^* = \widehat{\beta}_1^* + \widehat{\beta}_2^* X_i^*$. $\widehat{\beta}_2^* = \widehat{\beta}_2/\mu_2$ (Exercise 1.13).

$$\widehat{\beta}_1^* = \overline{Y} - \widehat{\beta}_2^* \overline{X}^* = \overline{Y} - \frac{\widehat{\beta}_2}{\mu_2}\mu_2 \overline{X} = \widehat{\beta}_1.$$

Hence:

$$\widehat{Y}_i^* = \widehat{\beta}_1 + \frac{\widehat{\beta}_2}{\mu_2}\mu_2 X_i = \widehat{Y}_i.$$

The fitted and actual values of Y are not affected by the transformation and so R^2 is unaffected.

1.25 The output shows the result of regressing weight in 2011 on height, using *EAWE* Data Set 21. In 2011 the respondents were aged 27–31. Explain why R^2 is lower than in the regression reported in Exercise 1.9.

```
. 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
------------------------------------------------------------------------------
```

Answer:

The explained sum of squares is actually higher than that in Exercise 1.9. 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 variations in eating habits.

1.5 Answers to the additional exercises

A1.1 Expenditure on food consumed at home increases by 6.3 cents for each dollar of total household expenditure. Literally the intercept implies that \$369 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.

A1.2 For each category, the regression sample has been restricted to households with nonzero expenditure. All the slope coefficients are highly significant. Housing has the largest coefficient, as one should expect. Surprisingly, it is followed by education. However, most households spent nothing at all on this category. For those that did, it was important.

	EXP		
	n	$\widehat{\beta}_2$	R^2
ADM	2,815	0.0235	0.228
CLOT	4,500	0.0316	0.176
DOM	1,661	0.0409	0.134
EDUC	561	0.1202	0.241
ELEC	5,828	0.0131	0.180
FDAW	5,102	0.0527	0.354
FDHO	6,334	0.0627	0.351
FOOT	1,827	0.0058	0.082
FURN	487	0.0522	0.102
GASO	5,710	0.0373	0.278
HEAL	4,802	0.0574	0.174
HOUS	6,223	0.1976	0.469
LIFE	1,253	0.0193	0.101
LOCT	692	0.0068	0.059
MAPP	399	0.0329	0.102
PERS	3,817	0.0069	0.213
READ	2,287	0.0048	0.104
SAPP	1,037	0.0045	0.034
TELE	5,788	0.0160	0.268
TEXT	992	0.0040	0.051
TOB	1,155	0.0165	0.088
TOYS	2,504	0.0145	0.076
TRIP	516	0.0466	0.186

A1.3 The summary data indicate that, on average, the respondents put on 13 pounds over the period 2004–2011. Was this due to the relatively heavy becoming even heavier, or to a general increase in weight? The regression output indicates that weight in 2011 was approximately equal to weight in 2004 plus 17 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 2004 weight accounts for 71 per cent of the variance in 2011 weight, so other factors are important.

A1.4 The slope coefficient indicates that hourly earnings increase by 41 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.

A1.5 *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, $\overline{G}^* = \overline{G} + 100$ and so $G_i^* - \overline{G}^* = G_i - \overline{G}$ for all i. Hence the new slope coefficient is:

$$\widehat{\beta}_2^* = \frac{\sum \left(G_i^* - \overline{G}^*\right)\left(N_i - \overline{N}\right)}{\sum \left(G_i^* - \overline{G}^*\right)^2} = \frac{\sum \left(G_i - \overline{G}\right)\left(N_i - \overline{N}\right)}{\sum \left(G_i - \overline{G}\right)^2} = \widehat{\beta}_2.$$

The coefficient is unchanged.

The intercept: The new intercept is:

$$\widehat{\beta}_1^* = \overline{N} - \widehat{\beta}_2^*\overline{G}^* = \overline{N} - \widehat{\beta}_2\left(\overline{G} + 100\right) = \widehat{\beta}_1 - 100\widehat{\beta}_2 = 23.0.$$

RSS: The residual in observation i in the new regression, $\widehat{u}_i^*$, is given by:

$$\widehat{u}_i^* = N_i - \widehat{\beta}_1^* - \widehat{\beta}_2^* G_i^* = N_i - (\widehat{\beta}_1 - 100\widehat{\beta}_2) - \widehat{\beta}_2(G_i + 100) = \widehat{u}_i$$

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

R^2:

$$R^2 = 1 - \frac{RSS}{\sum \left(N_i - \overline{N}\right)^2}$$

and is unchanged since RSS and $\sum \left(N_i - \overline{N}\right)^2$ 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.

A1.6 *The coefficient of GDP:* Let the revised measure of GDP be denoted G^*, where $G^* = 2G$. Since $G_i^* = 2G_i$ for all i, $\overline{G}^* = 2\overline{G}$ and so $G_i^* - \overline{G}^* = 2\left(G_i - \overline{G}\right)$ for all i. Hence the new slope coefficient is:

$$\begin{aligned}
\widehat{\beta}_2^* &= \frac{\sum \left(G_i^* - \overline{G}^*\right)\left(N_i - \overline{N}\right)}{\sum \left(G_i^* - \overline{G}^*\right)^2} \\
&= \frac{\sum 2\left(G_i - \overline{G}\right)\left(N_i - \overline{N}\right)}{\sum 4\left(G_i - \overline{G}\right)^2} \\
&= \frac{2\sum \left(G_i - \overline{G}\right)\left(N_i - \overline{N}\right)}{4\sum \left(G_i - \overline{G}\right)^2} \\
&= \frac{\widehat{\beta}_2}{2} \\
&= 0.010
\end{aligned}$$

where $\widehat{\beta}_2 = 0.020$ is the slope coefficient in the original regression.

The intercept: The new intercept is:

$$\widehat{\beta}_1^* = \overline{N} - \widehat{\beta}_2^* \overline{G}^* = \overline{N} - \frac{\widehat{\beta}_2}{2} 2\overline{G} = \overline{N} - \widehat{\beta}_2 \overline{G} = \widehat{\beta}_1 = 25.0$$

the original intercept.

RSS: The residual in observation i in the new regression, $\widehat{u}_i^*$, is given by:

$$\widehat{u}_i^* = N_i - \widehat{\beta}_1^* - \widehat{\beta}_2^* G_i^* = N_i - \widehat{\beta}_1 - \frac{\widehat{\beta}_2}{2} 2G_i = \widehat{u}_i$$

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

R^2:

$$R^2 = 1 - \frac{RSS}{\sum \left(N_i - \overline{N}\right)^2}$$

and is unchanged since RSS and $\sum \left(N_i - \overline{N}\right)^2$ are unchanged. As in Exercise A1.6, this makes sense intuitively.

A1.7 By construction, $\overline{Y}^* = \overline{X}^* = 0$. So $\widehat{\beta}_1^* = \overline{Y}^* - \widehat{\beta}_2^* \overline{X}^* = 0$.

$$\begin{aligned}
\widehat{\beta}_2^* &= \frac{\sum \left(X_i^* - \overline{X}^*\right)\left(Y_i^* - \overline{Y}^*\right)}{\sum \left(X_i^* - \overline{X}^*\right)^2} \\
&= \frac{\sum X_i^* Y_i^*}{\sum X_i^{*2}} \\
&= \frac{\sum \left(\frac{X_i - \bar{X}}{\widehat{\sigma}_X}\right)\left(\frac{Y_i - \bar{Y}}{\widehat{\sigma}_Y}\right)}{\sum \left(\frac{X_i - \bar{X}}{\widehat{\sigma}_X}\right)^2} \\
&= \frac{\widehat{\sigma}_X}{\widehat{\sigma}_Y} \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} \\
&= \frac{\widehat{\sigma}_X}{\widehat{\sigma}_Y} \widehat{\beta}_2.
\end{aligned}$$

$\widehat{\beta}_2^*$ provides an estimate of the effect on Y, in terms of standard deviations of Y, of a one-standard deviation change in X.

A1.8 We have:

$$\widehat{\beta}_2^{**} = \frac{\sum X_i^* Y_i^*}{\sum X_i^{*2}} = \frac{\sum \left(X_i^* - \overline{X}^*\right)\left(Y_i^* - \overline{Y}^*\right)}{\sum \left(X_i^* - \overline{X}^*\right)^2} = \widehat{\beta}_2^*.$$

A1.9 We have:

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

and $\widehat{Y}_i = \overline{Y}$ for all i.