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.

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:

- 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.

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 =     868
-------------+------------------------------           F(  1,   866) =  380.37
       Model |   911005795     1   911005795           Prob > F      =  0.0000
    Residual |  2.0741e+09   866  2395045.39           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
------------------------------------------------------------------------------
```

A1.2

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.

A1.3

The output 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.

```
. reg WEIGHT02 WEIGHT85

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

A1.4

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              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   31.72
       Model |  6236.81652     1  6236.81652           Prob > F      =  0.0000
    Residual |  105773.415   538   196.60486           R-squared     =  0.0557
-------------+------------------------------           Adj R-squared =  0.0539
       Total |  112010.231   539  207.811189           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
------------------------------------------------------------------------------
```

A1.5

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)

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

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 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.

A1.7

A variable Y_i is generated as

$$Y_i = \beta_1 + u_i \qquad (1.1)$$

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.

Answers to the starred exercises in the textbook

1.8

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.

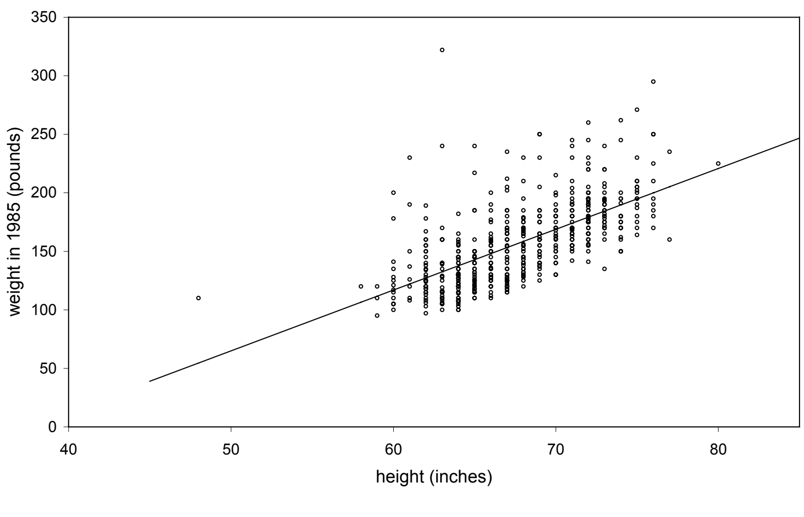
```
. reg WEIGHT85 HEIGHT

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



1.10

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:

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

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

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

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 $[(a_1 + a_2Y) + (b_1 + b_2Y)]$, and in general this will be equal to Y only if the two conditions are satisfied.

1.12

Suppose that the units of measurement of X are changed so that the new measure, X^*, is related to the original one by $X_i^* = \mu_1 + \mu_2 X_i$. Show that the new estimate of the slope coefficient is b_2/μ_2, where b_2 is the slope coefficient in the original regression.

Answer:

$$b_2^* = \frac{\sum_{i=1}^{n} (X_i^* - \overline{X}^*)(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i^* - \overline{X}^*)^2} = \frac{\sum_{i=1}^{n} ([\mu_1 + \mu_2 X_i] - [\mu_1 + \mu_2 \overline{X}])(Y_i - \overline{Y})}{\sum_{i=1}^{n} ([\mu_1 + \mu_2 X_i] - [\mu_1 + \mu_2 \overline{X}])^2}$$

$$= \frac{\sum_{i=1}^{n} (\mu_2 X_i - \mu_2 \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (\mu_2 X_i - \mu_2 \overline{X})^2} = \frac{\mu_2 \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\mu_2^2 \sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{b_2}{\mu_2} .$$

1.13

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 b_1^* and b_2^* be the intercept and slope coefficient in a regression of Y on X^* . Note that $\overline{X}^* = 0$. Then

$$b_1^* = \overline{Y} - b_2^* \overline{X}^* = \overline{Y}.$$

The slope coefficient is not affected by demeaning:

$$b_2^* = \frac{\sum_{i=1}^{n}\left(X_i^* - \overline{X}^*\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i^* - \overline{X}^*\right)^2} = \frac{\sum_{i=1}^{n}\left(\left[X_i - \overline{X}\right] - 0\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(\left[X_i - \overline{X}\right] - 0\right)^2} = b_2 .$$

1.14

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

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

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

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              Number of obs =     540
-------------+------------------------------           F(  1,    538) =  355.97
       Model |  53819.2324     1  53819.2324           Prob > F      =  0.0000
    Residual |   81340.044   538  151.189673           R-squared     =  0.3982
-------------+------------------------------           Adj R-squared =  0.3971
       Total |  135159.276   539  250.759325           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
------------------------------------------------------------------------------
```

1.15

Consider the regression model

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

It implies

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

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 $v_i = u_i - \overline{u}$.

Demonstrate that a regression of Y^* on X^* using (1.40) will yield the same estimate of the slope coefficient as a regression of *Y* on *X*. *Note*: (1.40) should be used instead of (1.28) because there is no intercept in this model.

Evaluate the outcome if the slope coefficient were estimated using (1.28), 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 b_2^* be the slope coefficient in a regression of Y^* on X^* using (1.40). Then

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

Let b_2^{**} be the slope coefficient in a regression of Y^* on X^* using (1.28). Note that $\overline{Y}^*$ and $\overline{X}^*$ are both zero. Then

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

Let b_1^{**} be the intercept in a regression of Y^* on X^* using (1.28). Then

$$b_1^{**} = \overline{Y}^* - b_2^{**} \overline{X}^* = 0 .$$

1.17

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

Answer:

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

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)\left(e_i - \overline{e}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\left[b_1 + b_2 X_i\right] - \left[b_1 + b_2 \overline{X}\right]\right)e_i \\ &= \frac{1}{n} b_2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)e_i \\ &= 0\end{aligned}$$

by (1.53). Hence the correlation is zero.

1.22

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 $X^* = \mu_1 + \mu_2 X$.

Answer:

Let the fitted regression of Y on X^* be written $\hat{Y}_i^* = b_1^* + b_2^* X_i^*$. $b_2^* = b_2 / \mu_2$ (Exercise 1.12).

$$b_1^* = \overline{Y} - b_2^* \overline{X}^* = \overline{Y} - b_2 \overline{X} - \frac{\mu_1 b_2}{\mu_2} = b_1 - \frac{\mu_1 b_2}{\mu_2}.$$

Hence

$$\hat{Y}_i^* = b_1 - \frac{\mu_1 b_2}{\mu_2} + \frac{b_2}{\mu_2}\left(\mu_1 + \mu_2 X_i\right) = \hat{Y}_i.$$

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

1.24

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              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
------------------------------------------------------------------------------
```

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.

Answers to the additional exercises

A1.1

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.

A1.2

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.

	EXP				
	n	b_2	$s.e.(b_2)$	R^2	F
FDHO	868	0.0527	0.0027	0.3052	380.4
FDAW	827	0.0440	0.0021	0.3530	450.0
HOUS	867	0.1935	0.0063	0.5239	951.9
TELE	858	0.0101	0.0009	0.1270	124.6
DOM	454	0.0225	0.0043	0.0581	27.9
TEXT	482	0.0049	0.0006	0.1119	60.5
FURN	329	0.0128	0.0023	0.0844	30.1
MAPP	244	0.0089	0.0018	0.0914	24.3
SAPP	467	0.0013	0.0003	0.0493	24.1
CLOT	847	0.0395	0.0018	0.3523	459.5
FOOT	686	0.0034	0.0003	0.1575	127.9
GASO	797	0.0230	0.0014	0.2528	269.0
TRIP	309	0.0240	0.0038	0.1128	39.0
LOCT	172	0.0008	0.0019	0.0009	0.2
HEAL	821	0.0226	0.0029	0.0672	59.0
ENT	824	0.0700	0.0040	0.2742	310.6
FEES	676	0.0306	0.0026	0.1667	134.8
TOYS	592	0.0090	0.0010	0.1143	76.1
READ	764	0.0039	0.0003	0.1799	167.2
EDUC	288	0.0265	0.0054	0.0776	24.1
TOB	368	0.0071	0.0014	0.0706	27.8

A1.3

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.

A1.4

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.

A1.5

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

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

The coefficient is unchanged.

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

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

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

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

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

$$b_2^* = \frac{\sum\left(G_i^* - \overline{G}^*\right)\left(N_i - \overline{N}\right)}{\sum\left(G_i^* - \overline{G}^*\right)^2} = \frac{\sum 2\left(G_i - \overline{G}\right)\left(N_i - \overline{N}\right)}{\sum 4\left(G_i - \overline{G}\right)^2}$$

$$= \frac{2\sum\left(G_i - \overline{G}\right)\left(N_i - \overline{N}\right)}{4\sum\left(G_i - \overline{G}\right)^2} = \frac{b_2}{2} = 0.010$$

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

The intercept: The new intercept is $b_1^* = \overline{N} - b_2^*\overline{G}^* = \overline{N} - \dfrac{b_2}{2}2\overline{G} = \overline{N} - b_2\overline{G} = b_1 = 25.0$,

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

$R^2 = \frac{\sum_i \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_i \left(Y_i - \overline{Y}\right)^2}$ and $\hat{Y}_i = \overline{Y}$ for all *i*.

Notes