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
- $R^2$. 
1. Simple regression analysis

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

------------------------------------------------------------------------------
<table>
<thead>
<tr>
<th>FDHO</th>
<th>Coef. Std. Err. t P&gt;</th>
<th>t</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXP</td>
<td>0.0627099   0.0010706 58.57 0.000 0.0606112 0.0648086</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>369.4418 10.65718 34.67 0.000 348.5501 390.3334</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

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
-------------+------------------------------
Model | 769248.875 1 769248.875 F(  1, 498) = 1207.55
Residual | 317241.693 498 637.031513 Prob > F = 0.0000
Total | 1086490.57 499 2177.33581 R-squared = 0.7080
-------------+------------------------------
Adj R-squared = 0.7074
Root MSE = 25.239

------------------------------------------------------------------------------
<table>
<thead>
<tr>
<th>WEIGHT11</th>
<th>Coef. Std. Err. t P&gt;</th>
<th>t</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIGHT04</td>
<td>0.9739736   0.0280281 34.75 0.000 0.9189056 1.029042</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>17.42232    4.888091 3.56 0.000 7.818493 27.02614</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

© Christopher Dougherty, 2016. All rights reserved.
Published on the Online Resource Centre to accompany Dougherty: Introduction to Econometrics, 5th edition, by Oxford University Press.
. sum WEIGHT04 WEIGHT11

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEIGHT04</td>
<td>500</td>
<td>169.686</td>
<td>40.31215</td>
<td>95</td>
<td>330</td>
</tr>
<tr>
<td>WEIGHT11</td>
<td>500</td>
<td>182.692</td>
<td>46.66193</td>
<td>95</td>
<td>370</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs =</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1393.77592</td>
<td>1</td>
<td>1393.77592</td>
<td>F( 1, 498) =</td>
<td>9.23</td>
</tr>
<tr>
<td>Residual</td>
<td>75171.3726</td>
<td>498</td>
<td>150.946531</td>
<td>Prob &gt; F =</td>
<td>0.0025</td>
</tr>
<tr>
<td>Total</td>
<td>76565.1485</td>
<td>499</td>
<td>153.437171</td>
<td>R-squared =</td>
<td>0.0182</td>
</tr>
<tr>
<td></td>
<td>Adj R-squared =</td>
<td>0.0162</td>
<td>Root MSE =</td>
<td>12.286</td>
<td></td>
</tr>
</tbody>
</table>

| 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 = \text{residual sum of squares}$):

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

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
- $RSS$
- $R^2$. 

© Christopher Dougherty, 2016. All rights reserved. Published on the Online Resource Centre to accompany Dougherty: Introduction to Econometrics, 5th edition, by Oxford University Press.
1. Simple regression analysis

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:

\[ Y_i^* = \frac{Y_i - \bar{Y}}{\hat{\sigma}_Y} \quad \text{and} \quad X_i^* = \frac{X_i - \bar{X}}{\hat{\sigma}_X} \]

\( \bar{Y} \) and \( \bar{X} \) are the sample means of \( Y \) and \( X \), \( \hat{\sigma}_Y \) and \( \hat{\sigma}_X \) are the estimators of the standard deviations of \( Y \) and \( X \), defined as the square roots of the estimated variances:

\[ \hat{\sigma}_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \quad \text{and} \quad \hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \]

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

\[ \hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \]

and:

\[ \hat{Y}_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i^* \]

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

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

\[ \hat{Y}_i^* = \hat{\beta}_2^* X_i^* \]

Determine how \( \hat{\beta}_2^* \) is related to \( \hat{\beta}_2 \).

A1.9 A variable \( Y_i \) is generated as:

\[ Y_i = \beta_1 + u_i \quad (1.1) \]

where \( \beta_1 \) is a fixed parameter and \( u_i \) is a disturbance term that is independently and identically distributed with expected value 0 and population variance \( \sigma_u^2 \). The least squares estimator of \( \beta_1 \) is \( \bar{Y} \), 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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>211309 1 211309</td>
<td>F( 1, 498) = 176.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>595389.95 498 1195.56215</td>
<td>Prob &gt; F = 0.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>806698.98 499 1616.63116</td>
<td>Adj R-squared = 0.2605</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 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:

\[ \hat{W}_i = \hat{\alpha}_1 + \hat{\alpha}_2 Y_i \]
\[ \hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 Y_i \]

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

\[ \hat{\alpha}_2 + \hat{\beta}_2 = 1 \]
\[ \hat{\alpha}_1 + \hat{\beta}_1 = 0. \]

Explain intuitively why this should be so.

**Answer:**

\[
\hat{\alpha}_2 + \hat{\beta}_2 = \frac{\sum (Y_i - \bar{Y})(W_i - \bar{W})}{\sum (Y_i - \bar{Y})^2} + \frac{\sum (Y_i - \bar{Y})(P_i - \bar{P})}{\sum (Y_i - \bar{Y})^2} = 1
\]

"EC2020 Elements of Econometrics" offered as part of the University of London International Programmes in Economics, Management, Finance, and the Social Sciences.

© Christopher Dougherty, 2016. All rights reserved.
Published on the Online Resource Centre to accompany Dougherty: Introduction to Econometrics, 5th edition, by Oxford University Press.
1. Simple regression analysis

\[ \hat{\alpha}_1 + \hat{\beta}_1 = (\bar{W} - \hat{\alpha}_2\bar{Y}) + (\bar{P} - \hat{\beta}_2\bar{Y}) = (\bar{W} + \bar{P}) - (\hat{\alpha}_2 + \hat{\beta}_2)\bar{Y} = \bar{Y} - \bar{Y} = 0. \]

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 \( \bar{W}_i + \bar{P}_i = (\hat{\alpha}_1 + \hat{\alpha}_2 Y_i) + (\hat{\beta}_1 + \hat{\beta}_2 Y_i) \), 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 \( \hat{\beta}_2^*/\mu_2 \), where \( \hat{\beta}_2 \) is the slope coefficient in the original regression.

**Answer:**

\[
\hat{\beta}_2^* = \frac{\sum (X^*_i - \bar{X}^*) (Y_i - \bar{Y})}{\sum (X^*_i - \bar{X}^*)^2} = \frac{\sum (\mu_2 X_i - \mu_2 \bar{X}) (Y_i - \bar{Y})}{\sum (\mu_2 X_i - \mu_2 \bar{X})^2} = \frac{\mu_2 \sum (X_i - \bar{X}) (Y_i - \bar{Y})}{\mu_2^2 \sum (X_i - \bar{X})^2} = \frac{\hat{\beta}_2}{\mu_2}.
\]

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 \( \bar{Y} \).

**Answer:**

Let \( X^*_i = X_i - \bar{X} \) and \( \hat{\beta}_0^* \) and \( \hat{\beta}_2^* \) be the intercept and slope coefficient in a regression of \( Y \) on \( X^* \). Note that \( \bar{X}^* = 0 \). Then:

\[ \hat{\beta}_0^* = \bar{Y} - \hat{\beta}_2^*\bar{X}^* = \bar{Y}. \]

The slope coefficient is not affected by demeaning:

\[ \hat{\beta}_2^* = \frac{\sum (X^*_i - \bar{X}^*) (Y_i - \bar{Y})}{\sum (X^*_i - \bar{X}^*)^2} = \frac{\sum (X_i - \bar{X}) (Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \hat{\beta}_2. \]

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: \( W\text{METRIC} = 0.454 \times \text{WEIGHT04} \) and \( H\text{METRIC} = 2.54 \times \text{HEIGHT} \). 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
----------|--------|-------
Model | 43554.1641 1 43554.1641
Residual | 122719.394 498 246.424486
----------|--------|-------
Total | 166273.558 499 333.213544

F( 1,  498) = 176.74
Prob > F = 0.0000

R-squared = 0.2619
Adj R-squared = 0.2605

ROOT MSE = 15.698

| Coef. Std. Err. t P>|t| [95% Conf. Interval] |
|------------------|-------------|--------|--------|
| WMETRIC | .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, \( \hat{\beta}_2^M \) and \( \hat{\beta}_1^M \), are then given by:

\[
\hat{\beta}_2^M = \frac{\sum (H_i - \bar{H})(W_i - \bar{W})}{\sum (H_i - \bar{H})^2} = \frac{\sum 2.54(H_i - \bar{H})0.454(W_i - \bar{W})}{\sum 2.54^2(H_i - \bar{H})^2} = 0.179 \frac{\sum (H_i - \bar{H})(W_i - \bar{W})}{\sum (H_i - \bar{H})^2} = 0.179\hat{\beta}_2
\]

\[
= 0.179 \times 5.074 = 0.908
\]

\[
\hat{\beta}_1^M = \bar{W}M - \hat{\beta}_2^MHM = 0.454\bar{W} - (\frac{0.454}{2.54}\hat{\beta}_2) (2.54\bar{H}) = 0.454(\bar{W} - \hat{\beta}_2\bar{H}) = 0.454\hat{\beta}_1 = 0.454 \times -177.2 = -80.4.
\]
1. Simple regression analysis

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:

\[ \bar{Y} = \beta_1 + \beta_2 \bar{X} + \bar{u} \]

and hence that:

\[ Y_i^* = \beta_2 X_i^* + v_i \]

where \( Y_i^* = Y_i - \bar{Y} \), \( X_i^* = X_i - \bar{X} \) 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 \( \hat{\beta}_2^* \) be the slope coefficient in a regression of \( Y^* \) on \( X^* \) using (1.49). Then:

\[ \hat{\beta}_2^* = \frac{\sum X_i^* Y_i^*}{\sum X_i^{*2}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \hat{\beta}_2. \]

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

\[ \hat{\beta}_2^{**} = \frac{\sum (X_i^* - \bar{X}^*) (Y_i^* - \bar{Y}^*)}{\sum (X_i^* - \bar{X}^*)^2} = \frac{\sum X_i^* Y_i^*}{\sum X_i^{*2}} = \hat{\beta}_2. \]

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

\[ \hat{\beta}_1^{**} = \bar{Y}^* - \hat{\beta}_2^{**} \bar{X}^* = 0. \]

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 \( \hat{Y} \) and \( \hat{u} \) can be decomposed as follows, using the fact that \( \bar{\hat{u}} = 0 \):

\[ \frac{1}{n} \sum (\hat{Y}_i - \bar{Y})(\hat{u}_i - \bar{\hat{u}}) = \frac{1}{n} \sum \left( [\hat{\beta}_1 + \hat{\beta}_2 X_i] - [\hat{\beta}_1 + \hat{\beta}_2 \bar{X}] \right) \hat{u}_i \]

\[ = \frac{1}{n} \hat{\beta}_2 \sum (X_i - \bar{X}) \hat{u}_i \]

\[ = 0. \]
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

$$\hat{Y}_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i^*, \quad \hat{\beta}_2^* = \frac{\hat{\beta}_2}{\mu_2}$$

(Exercise 1.13).

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

Hence:

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

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
-------------+------------------------------ 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.
1. Simple regression analysis

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.

<table>
<thead>
<tr>
<th>EXP</th>
<th>( n )</th>
<th>( \hat{\beta}_2 )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADM</td>
<td>2,815</td>
<td>0.0235</td>
<td>0.228</td>
</tr>
<tr>
<td>CLOT</td>
<td>4,500</td>
<td>0.0316</td>
<td>0.176</td>
</tr>
<tr>
<td>DOM</td>
<td>1,661</td>
<td>0.0409</td>
<td>0.134</td>
</tr>
<tr>
<td>EDUC</td>
<td>561</td>
<td>0.1202</td>
<td>0.241</td>
</tr>
<tr>
<td>ELEC</td>
<td>5,828</td>
<td>0.0527</td>
<td>0.180</td>
</tr>
<tr>
<td>FDAW</td>
<td>6,334</td>
<td>0.0527</td>
<td>0.134</td>
</tr>
<tr>
<td>FDHO</td>
<td>1,827</td>
<td>0.0058</td>
<td>0.351</td>
</tr>
<tr>
<td>FOOT</td>
<td>487</td>
<td>0.0522</td>
<td>0.102</td>
</tr>
<tr>
<td>GASO</td>
<td>5,710</td>
<td>0.0069</td>
<td>0.101</td>
</tr>
<tr>
<td>HEAL</td>
<td>4,802</td>
<td>0.0373</td>
<td>0.174</td>
</tr>
<tr>
<td>HOUS</td>
<td>6,223</td>
<td>0.1076</td>
<td>0.469</td>
</tr>
<tr>
<td>LIFE</td>
<td>1,253</td>
<td>0.0193</td>
<td>0.101</td>
</tr>
<tr>
<td>LOCT</td>
<td>692</td>
<td>0.0068</td>
<td>0.059</td>
</tr>
<tr>
<td>MAPP</td>
<td>399</td>
<td>0.0329</td>
<td>0.102</td>
</tr>
<tr>
<td>PERS</td>
<td>3,817</td>
<td>0.0069</td>
<td>0.213</td>
</tr>
<tr>
<td>READ</td>
<td>2,287</td>
<td>0.0048</td>
<td>0.104</td>
</tr>
<tr>
<td>SAPP</td>
<td>1,037</td>
<td>0.0045</td>
<td>0.034</td>
</tr>
<tr>
<td>TELE</td>
<td>5,788</td>
<td>0.0466</td>
<td>0.268</td>
</tr>
<tr>
<td>TEXT</td>
<td>992</td>
<td>0.0040</td>
<td>0.051</td>
</tr>
<tr>
<td>TOB</td>
<td>1,155</td>
<td>0.1065</td>
<td>0.088</td>
</tr>
<tr>
<td>TOYS</td>
<td>2,504</td>
<td>0.0145</td>
<td>0.076</td>
</tr>
<tr>
<td>TRIP</td>
<td>516</td>
<td>0.0466</td>
<td>0.186</td>
</tr>
</tbody>
</table>

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 $i$, $G^* = G + 100$ and so $G_i^* - G^* = G_i - G$ for all $i$. Hence the new slope coefficient is:

$$\hat{\beta}_2^* = \frac{\sum (G_i^* - G^*) (N_i - \bar{N})}{\sum (G_i^* - G^*)^2} = \frac{\sum (G_i - G) (N_i - \bar{N})}{\sum (G_i - G)^2} = \hat{\beta}_2.$$ 

The coefficient is unchanged.

The intercept: The new intercept is:

$$\hat{\beta}_1^* = \bar{N} - \hat{\beta}_2^* G^* = \bar{N} - \hat{\beta}_2 (G + 100) = \hat{\beta}_1 - 100 \hat{\beta}_2 = 23.0.$$ 

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

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

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.

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

$$\hat{\beta}_2^* = \frac{\sum (G_i^* - G^*) (N_i - \bar{N})}{\sum (G_i^* - G^*)^2}$$

$$= \frac{\sum 2 (G_i - G) (N_i - \bar{N})}{\sum 4 (G_i - G)^2}$$

$$= \frac{2 \sum (G_i - G) (N_i - \bar{N})}{4 \sum (G_i - G)^2}$$

$$= \frac{\hat{\beta}_2}{2}$$

$$= 0.010$$
1. Simple regression analysis

where \( \hat{\beta}_2 = 0.020 \) is the slope coefficient in the original regression.

The intercept: The new intercept is:

\[
\hat{\beta}_1^* = \overline{N} - \hat{\beta}_2^* \overline{G} = \overline{N} - \frac{\hat{\beta}_2 G}{2} = \overline{N} - \hat{\beta}_2 G = \hat{\beta}_1 = 25.0
\]

the original intercept.

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

\[
\hat{u}_i^* = N_i - \hat{\beta}_1^* - \hat{\beta}_2^* G_i^* = N_i - \hat{\beta}_1 - \frac{\hat{\beta}_2 G}{2} = \hat{u}_i
\]

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

\[
R^2 = 1 - \frac{RSS}{\sum (N_i - \overline{N})^2}
\]

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

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

\[
\hat{\beta}_2^* = \frac{\sum (X_i - \overline{X}^*) (Y_i - \overline{Y}^*)}{\sum (X_i - \overline{X}^*)^2} = \frac{\sum X_i Y_i^*}{\sum X_i^2} = \frac{\sum \left( \frac{X_i - \overline{X}}{\overline{X}} \right) \left( \frac{Y_i - \overline{Y}}{\overline{Y}} \right)}{\sum \left( \frac{X_i - \overline{X}}{\overline{X}} \right)^2} = \frac{\hat{\sigma}_X \sum (X_i - \overline{X}) (Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\hat{\sigma}_X}{\hat{\sigma}_Y} \hat{\beta}_2.
\]

\( \hat{\beta}_2^* \) 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:

\[
\hat{\beta}_2^{**} = \frac{\sum X_i Y_i^*}{\sum X_i^2} = \frac{\sum (X_i - \overline{X}^*) (Y_i - \overline{Y}^*)}{\sum (X_i^* - \overline{X}^*)^2} = \hat{\beta}_2^*.
\]
A1.9 We have:

\[ R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \]

and \( \hat{Y}_i = \bar{Y} \) for all \( i \).