統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 15(1): Multiple Regression

上課時間地點: 二 D56, 資訊 140306
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: ____________________ 學號: ____________________ 姓名: ____________________

15.1 Multiple Regression Model

1. (Recall) that the variable being predicted or explained is called the ______________ variable and the variable being used to predict or explain the dependent variable is called the ________________ variable.

2. Multiple regression analysis is the study of how a dependent variable y is related to ____________________________ variables. In the general case, we will use ______ to denote the number of independent variables.

3. The concepts of a regression model and a regression equation introduced in the preceding chapter are ______________ in the multiple regression case.

4. **Multiple regression model**: The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \cdots, x_p$ and an error term is called the multiple regression model.

$$\underline{\hspace{6cm}} \qquad (15.1)$$

5. In the multiple regression model, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are the ________________ and the error term ϵ is a ____________________. y is a linear function of $x_1, x_2, \cdots, x_p$ plus the error term ϵ.

6. The error term accounts for the ______________ in y that _______________________ by the linear effect of the p independent variables.

7. (**Multiple regression equation**):The equation that describes how the mean value of y is related to $x_1, x_2, \cdots, x_p$ is called the multiple regression equation.

   ______________________________ (15.2)

   under the assumption that the mean or expected value of ϵ is zero.

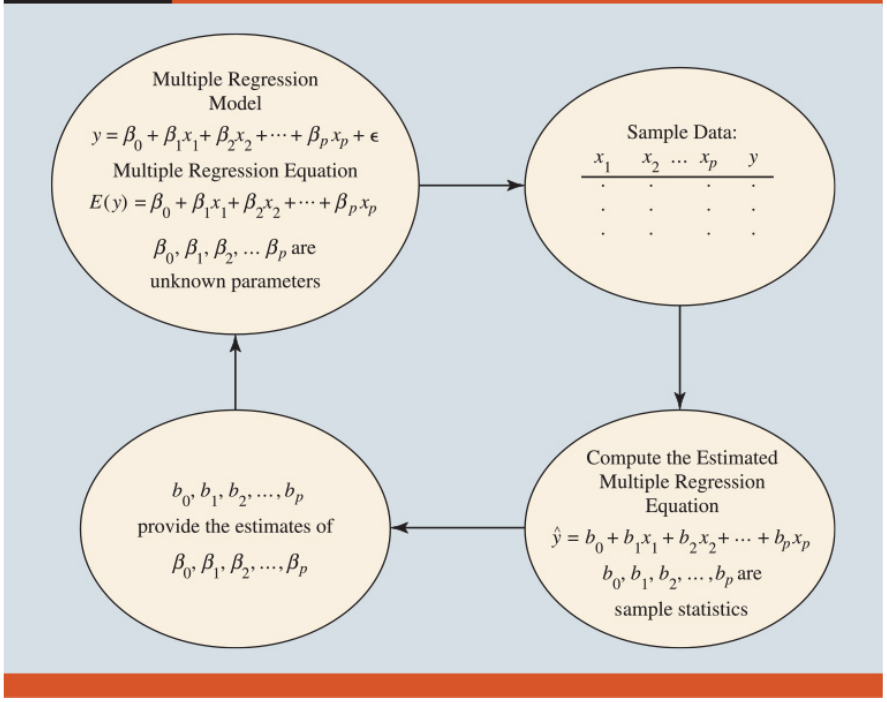
8. **The estimated multiple regression equation**:

   ______________________________ (15.3)

   where $b_0, b_1, b_2, \cdots, b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ and $\hat{y}$ is the predicted value of the dependent variable

9. (Figure 15.1)

FIGURE 15.1 The Estimation Process for Multiple Regression


Multiple Regression Model
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$
Multiple Regression Equation
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
$\beta_0, \beta_1, \beta_2, \cdots \beta_p$ are unknown parameters
Sample Data:
$x_1 \quad x_2 \; \cdots \; x_p \quad y$
Compute the Estimated Multiple Regression Equation
$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$
$b_0, b_1, b_2, \ldots, b_p$ are sample statistics
$b_0, b_1, b_2, \ldots, b_p$
provide the estimates of
$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

15.2 Least Squares Method

1. The least squares method is used to develop the estimated multiple regression equation:

   ____________________ (15.4)

   where y_i is observed value of the dependent variable for the ith observation, $\hat{y}_i$ is predicted value of the dependent variable for the ith observation

2. In multiple regression, however, the presentation of the formulas for the regression coefficients $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ involves the use of __________________ and is beyond the scope of this text.

3. Therefore, in presenting multiple regression, we focus on how statistical software can be used to obtain the estimated regression equation and other information. The emphasis will be on how to ____________ the computer output rather than on how to make the multiple regression computations.

An Example: Butler Trucking Company

1. The Butler Trucking Company, an independent trucking company in southern California.

2. A major portion of Butler's business involves deliveries throughout its local area. To develop better work schedules, the managers want to predict the total daily travel time for their drivers.

   (a) Initially the managers believed that the total daily travel time would be closely related to the number of miles traveled in making the daily deliveries.

   (b) (Table 15.1)(Figure 15.2) A simple random sample of 10 driving assignments provided the data shown in Table 15.1 and the scatter diagram shown in Figure 15.2.

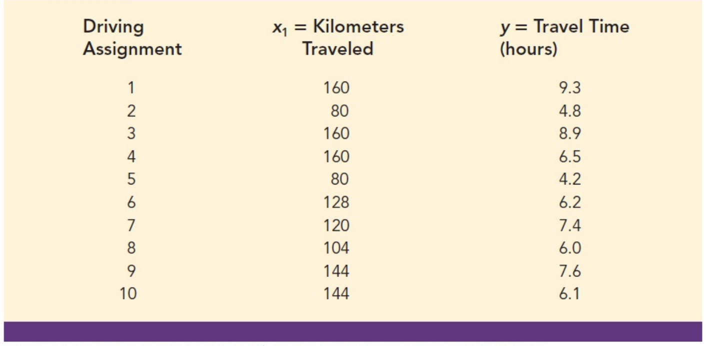


TABLE 15.1 Preliminary Data for Butler Trucking

Driving Assignment	x_1 = Kilometers Traveled	y = Travel Time (hours)
1	160	9.3
2	80	4.8
3	160	8.9
4	160	6.5
5	80	4.2
6	128	6.2
7	120	7.4
8	104	6.0
9	144	7.6
10	144	6.1

Source: PC Magazine website, April, 2015. (https://www.pcmag.com/reviews/monitors)

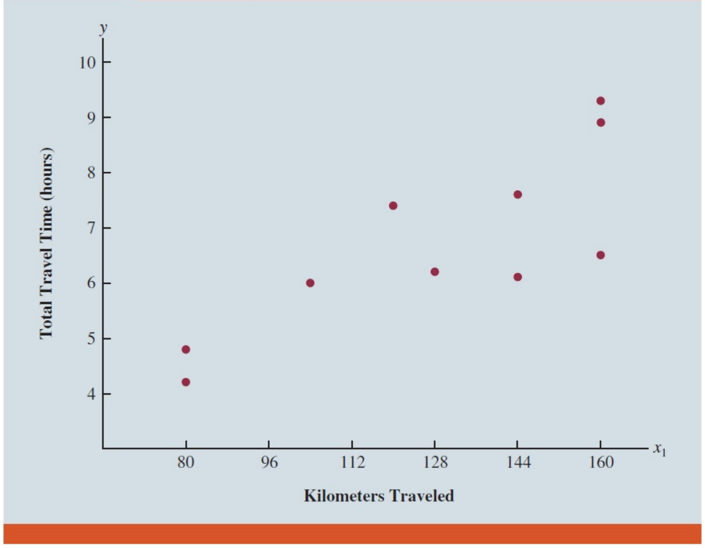


FIGURE 15.2 Scatter Diagram of Preliminary Data for Butler Trucking

(c) After reviewing this scatter diagram, the managers hypothesized that the simple linear regression model $y = \beta_0 + \beta_1 x_1 + \epsilon$ could be used to describe the relationship between the total travel time (y) and the number of miles traveled (x_1).

(d) (Figure 15.3) we show statistical software output from applying simple linear regression to the data in Table 15.1. The estimated regression equation is ____________________

i. At the 0.05 level of significance, the F value of __________ and its corresponding p-value of __________ indicate that the relationship is significant; that is, we can reject $H_0 : \beta_1 = 0$ because the p-value is less than $\alpha = 0.05$.

ii. Note that the same conclusion is obtained from the t value of __________ and its associated p-value of __________.

iii. Thus, we can conclude that the relationship between the total travel time and the number of miles traveled is ______________; longer travel times are associated with more miles traveled.

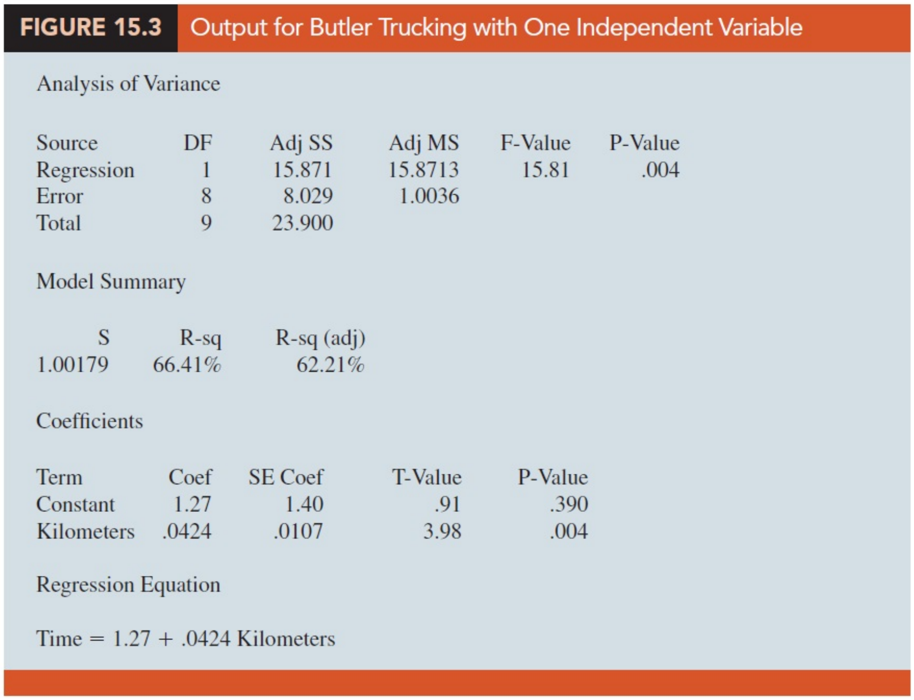


FIGURE 15.3 Output for Butler Trucking with One Independent Variable

```
Analysis of Variance

Source       DF     Adj SS     Adj MS    F-Value    P-Value
Regression    1     15.871    15.8713      15.81       .004
Error         8      8.029     1.0036
Total         9     23.900

Model Summary

      S      R-sq     R-sq (adj)
1.00179    66.41%        62.21%

Coefficients

Term          Coef    SE Coef     T-Value     P-Value
Constant      1.27       1.40         .91        .390
Kilometers   .0424      .0107        3.98        .004

Regression Equation

Time = 1.27 + .0424 Kilometers
```

iv. With a coefficient of determination (expressed as a percentage) of __________________, we see that ___________________________ in travel time can be explained by the linear effect of the number of miles traveled.

3. (Table 15.2) The managers might want to consider adding a second independent variable (number of deliveries) to explain some of the remaining variability in the dependent variable.

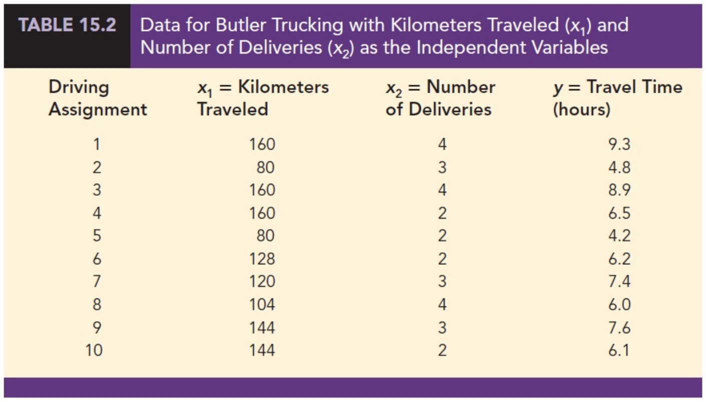


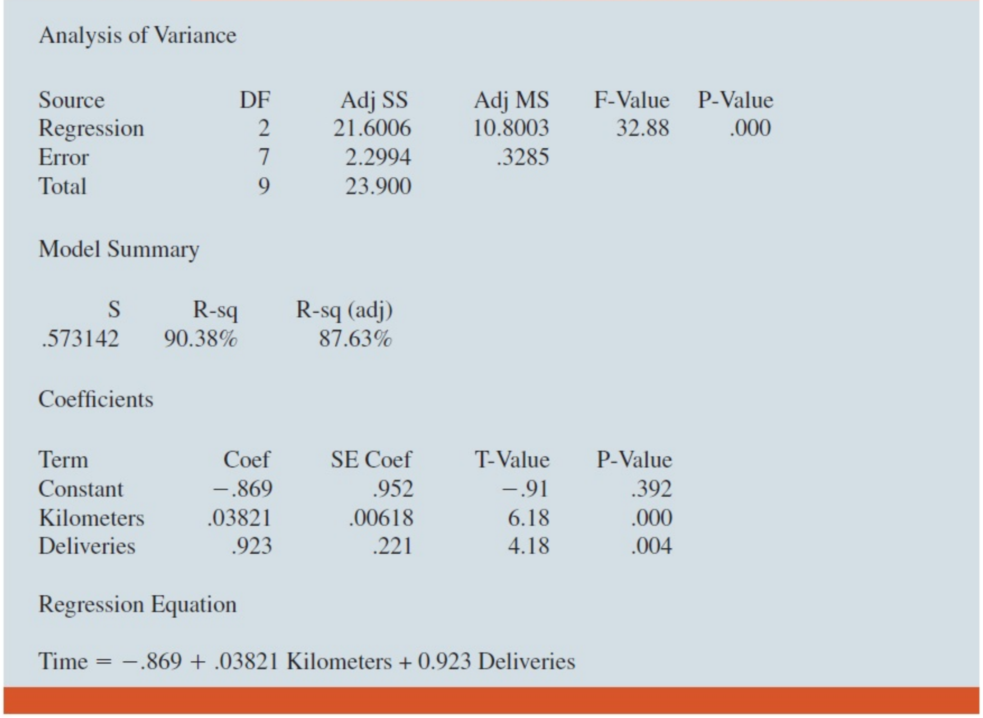
TABLE 15.2 Data for Butler Trucking with Kilometers Traveled (x_1) and Number of Deliveries (x_2) as the Independent Variables

Driving Assignment	x_1 = Kilometers Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
1	160	4	9.3
2	80	3	4.8
3	160	4	8.9
4	160	2	6.5
5	80	2	4.2
6	128	2	6.2
7	120	3	7.4
8	104	4	6.0
9	144	3	7.6
10	144	2	6.1

4. (Figure 15.4) Computer output with both miles traveled (x_1) and number of deliveries (x_2) as independent variables is shown in Figure 15.4. The estimated regression equation is

$$\hat{y} = \underline{\hspace{6cm}} \qquad (15.6)$$

FIGURE 15.4 Output for Butler Trucking with Two Independent Variables



```
Analysis of Variance

Source         DF     Adj SS     Adj MS    F-Value   P-Value
Regression      2    21.6006    10.8003      32.88      .000
Error           7     2.2994      .3285
Total           9     23.900

Model Summary

       S      R-sq    R-sq (adj)
 .573142    90.38%        87.63%

Coefficients

Term           Coef    SE Coef    T-Value   P-Value
Constant      −.869       .952       −.91      .392
Kilometers   .03821     .00618       6.18      .000
Deliveries     .923       .221       4.18      .004

Regression Equation

Time = −.869 + .03821 Kilometers + 0.923 Deliveries
```

Note on Interpretation of Coefficients

1. One observation can be made at this point about the relationship between the estimated regression equation with only the miles traveled as an independent variable and the equation that includes the ____________________ as a second independent variable.

2. The value of ______ is not the same in both cases. In simple linear regression, we interpret β_1 as an estimate of the change in y for a ________________ in the independent variable.

3. In multiple regression analysis, we interpret each regression coefficient as follows: b_i represents an estimate of the _____________ corresponding to a ________________ ________ when all other independent variables are ______________.

4. Butler Trucking example

   (a) $\beta_1 = 0.06113$, an estimate of the expected increase in travel time corresponding to an increase of one mile in the distance traveled when the number of deliveries is held constant is 0.06113 hours.

   (b) $\beta_2 = 0.923$, an estimate of the expected increase in travel time corresponding to an increase of one delivery when the number of miles traveled is held constant is 0.923 hours.

😍 **EXERCISES 15.2**: 1, 5, 6

15.3 Multiple Coefficient of Determination

1. In simple linear regression, we showed that the total sum of squares can be partitioned into two components: the sum of squares due to regression and the sum of squares due to error. The same procedure applies to the sum of squares in multiple regression.

   ______________________ (15.7)

   where

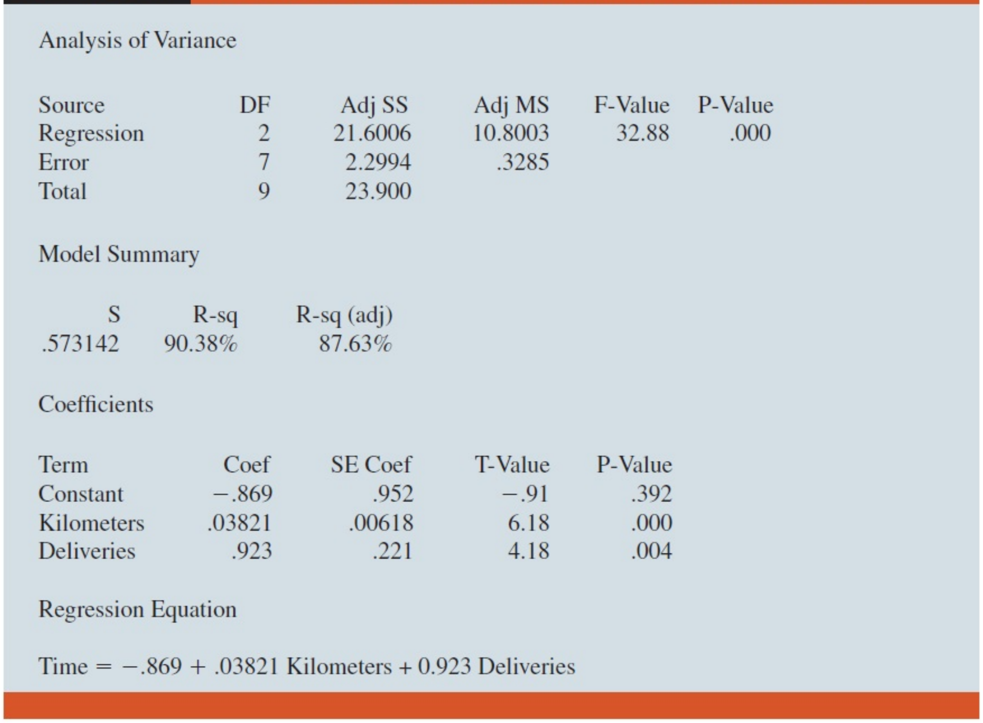
   SST: total sum of squares = ______________.

   SSR: sum of squares due to regression = ______________.

   SSE: sum of squares due to error = ______________.

2. Example Butler Trucking problem (Figure 15.4)

FIGURE 15.4 Output for Butler Trucking with Two Independent Variables



```
Analysis of Variance

Source        DF    Adj SS     Adj MS    F-Value   P-Value
Regression     2   21.6006    10.8003      32.88      .000
Error          7    2.2994      .3285
Total          9    23.900

Model Summary

     S       R-sq     R-sq (adj)
.573142    90.38%       87.63%

Coefficients

Term           Coef     SE Coef    T-Value    P-Value
Constant      −.869        .952       −.91       .392
Kilometers   .03821      .00618       6.18       .000
Deliveries     .923        .221       4.18       .004

Regression Equation

Time = −.869 + .03821 Kilometers + 0.923 Deliveries
```

$SST = 23.900$, $SSR = 21.6006$, and $SSE = 2.2994$.

3. With only one independent variable (number of miles traveled), the output in Figure 15.3 shows that $SST = 23.900$, $SSR = 15.871$, and $SSE = 8.029$. The value of SST is the same in both cases because it does not depend on $\hat{y}$, but SSR increases and SSE decreases when a second independent variable (number of deliveries) is added.

4. The multiple coefficient of determination, denoted R^2, measures the goodness of fit for the estimated multiple regression equation.

$$\underline{\hspace{4cm}} \qquad (15.8)$$

5. The multiple coefficient of determination can be interpreted as the ______________________ ______________ in the dependent variable that can be explained by the estimated multiple regression equation.

6. Hence, when multiplied by 100, it can be interpreted as the percentage of the variability in y that can be explained ______________________________________.

7. Example In the two-independent-variable Butler Trucking example, with $SSR = 21.6006$ and $SST = 23.900$, we have $R^2 = 21.6006/23.900 = 0.9038$.

8. Therefore, 90.38% of the variability in travel time y is explained by the estimated multiple regression equation with miles traveled and number of deliveries as the independent variables.

9. (Figure 15.3) the R-sq value for the estimated regression equation with only one independent variable, number of miles traveled (x_1), is 66.41%. Thus, the percentage of the variability in travel times that is explained by the estimated regression equation increases from ____________________ when number of deliveries is added as a second independent variable.

10. In general, R^2 always increases as independent variables are added to the model.

11. Many analysts prefer adjusting R^2 for the number of independent variables to avoid __________________ the impact of adding an independent variable on the amount of variability exlained by the estimated regression equation.

12. With n denoting the number of observations and p denoting the number of independent variables, the adjusted multiple coefficient of determination is computed as follows:

    ______________________________ (15.9)

13. Example With $n = 10$ and $p = 2$, we have

$$R^2 = 1 - (1 - 0.9038)\frac{10-1}{10-2-1}$$

14. Thus, after adjusting for the two independent variables, we have an adjusted multiple coefficient of determination of 0.8763. This value (expressed as a percentage) is provided in the output in Figure 15.4 as ______________________.

15. If the value of R^2 is small and the model contains a large number of independent variables, the adjusted coefficient of determination can take a _________________; in such cases, statistical software usually sets the adjusted coefficient of determination to ________.

☺ **EXERCISES 15.3**: 11, 14, 15