Multiple Regression

Tom Bruning

2018-03-28

Multiple Regression

In this chapter, you learn:

- How to develop a multiple regression model
- How to interpret the regression coefficients
- How to determine which independent variables to include in the regression model
- How to determine which independent variables are most important in * predicting a dependent variable
- How to use categorical independent variables in a regression model
- How to predict a categorical dependent variable using logistic regression
- How to identify individual observations that may be unduly influencing the multiple regression model

The Multiple Regression Model

Idea: Examine the linear relationship between 1 dependent (Y) & 2 or more independent variables (Xi)

Multiple Regression Model with k Independent Variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} \tag{1}$$

Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} \tag{2}$$

Where:

- $\hat{Y}_i$ = Estimated or predicted value of Y
- b_0 = Y intercept
- b_1, b_2, b_k = slope coefficients

Example: 2 Independent Variables

A distributor of frozen dessert pies wants to evaluate factors thought to influence demand

Dependent variable:

- Pie sales (units per week)
  Independent variables:
- Price (in $)
- Advertising ($100's) Data are collected for 15 weeks

Pie Sales	Price	Advertising (000
350	5.5	3.3
460	7.5	3.3
350	8.0	3.0
430	8.0	4.5
350	6.8	3.0
380	7.5	4.0
430	4.5	3.0
470	6.4	3.7
450	7.0	3.5
490	5.0	4.0
340	7.2	3.5
300	7.9	3.2
440	5.9	4.0
450	5.0	3.5
300	7.0	2.7

Excel Multiple Regression Output

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\widehat{Sales} = 306.526 - 24.975(Price) + 74.131(Advertising)$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

The Multiple Regression Equation

$$\hat{Sales} = 306.525 - 24.975(Price) + 74.131(Advertising) \quad (3)$$

where:

- Sales is in number of pies per week
- Price is in $
- Advertising is in $100's.

$b_1 = -24.975$: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising.

$b_2 = 74.131$: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price.

Using The Equation to Make Predictions

Predict sales for a week in which the selling price is $5.50 and advertising is $350:

$$\hat{Sales} = 306.525 - 24.975(5.50) + 74.131(3.5) = 428.62 \quad (4)$$

Note that Advertising is in $100s, so $350 means that $X_2 = 3.5$
Predicted sales is 428.62 pies

The Coefficient of Multiple Determination, r^2

Reports the proportion of total variation in Y explained by all X variables taken together.

$$r^2 = \frac{SSR}{SST} = \frac{regression\ sum\ of\ squares}{total\ sum\ of\ squares} \quad (5)$$

Adjusted r^2

- r^2 never decreases when a new X variable is added to the model
  - This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
  - We lose a degree of freedom when a new X variable is added
  - Did the new X variable add enough explanatory power to offset the loss of one degree of freedom?

r^2 shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used

$$R^2_{adj} = 1 - [(1 - r^2)(\frac{n-1}{n-k-1})] \quad (6)$$

(where n = sample size, k = number of independent variables)
Penalizes excessive use of unimportant independent variables
Smaller than r2
Useful in comparing among models

Using Dummy Variables

- A dummy variable is a categorical independent variable with two levels:
  - yes or no, on or off, male or female
  - coded as 0 or 1
- Assumes the slopes associated with numerical independent variables do not change with the value for the categorical variable
- If more than two levels, the number of dummy variables needed is (number of levels - 1)

Dummy-Variable Example (with 2 Levels)

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 \tag{7}$$

Let:
Y = pie sales
X_1 = price
X_2 = holiday (X_2 = 1 if a holiday occurred during the week)
($X_2 = 0$ if there was no holiday that week)
No Holiday

$$\hat{Y} = b_0 + b_1X_1 + b_2(0) = b_0 + b_1X_1 \tag{8}$$

Holiday

$$\hat{Y} = b_0 + b_1X_1 + b_2(1) = b_0 + b_2 + b_1X_1 \tag{9}$$

Interpreting the Dummy Variable Coefficient (with 2 Levels)

Example:

$$S\hat{a}les = 300 - 30(Price) + 15(Holiday) \tag{10}$$

Sales: number of pies sold per week Price: pie price in $
Holiday:
1 If a holiday occurred during the week
0 If no holiday occurred
$b_2 = 15$
on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price

Dummy-Variable Models (more than 2 Levels)

The number of dummy variables is one less than the number of levels

Example:

- Y = house price ;
- X_1 = square feet
- If style of the house is also thought to matter:
  - Style = ranch, split level, colonial

Three levels, so two dummy variables are needed.

Example: Let “colonial” be the default category, and let X2 and X3 be used for the other two categories:

Y = house price

X_1 = square feet

$X_2 = 1$ if ranch, 0 otherwise

$X_3 = 1$ if split level, 0 otherwise

The multiple regression equation is:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 \tag{11}$$

Interpreting the Dummy Variable Coefficients (with 3 Levels)

Consider the regression equation:

$$\hat{Y} = 20.43 + 0.045(X_1) + 23.53(X_2) + 18.84(X_3) \tag{12}$$

- For a colonial: $X_2 = X_3 = 0$

$$\hat{Y} = 20.43 + 0.045(X_1) \tag{13}$$

- For a ranch: $X_2 = 1; X_3 = 0$

$$\hat{Y} = 20.43 + 0.045(X_1) + 23.53 \tag{14}$$

  With the same square feet, a ranch will have an estimated average price of 23.53 thousand dollars more than a colonial.
- For a split level: $X_2 = 0; X_3 = 1$

$$\hat{Y} = 20.43 + 0.045(X_1) + 18.84 \tag{15}$$

With the same square feet, a split-level will have an estimated average price of 18.84 thousand dollars more than a colonial.

Table 2: Housing Data

Style	House Price in $1000s (Y)	Square Feet (X)	Col	SL
Col	245	1400	1	0
Col	312	1600	1	0
Col	279	1700	1	0
Col	308	1875	1	0
Col	199	1100	1	0
Col	219	1550	1	0
Col	405	2350	1	0
Col	324	2450	1	0
Col	319	1425	1	0
Col	255	1700	1	0
SL	345	1400	0	1
SL	412	1600	0	1
SL	379	1700	0	1
SL	408	1875	0	1
SL	299	1100	0	1
SL	319	1550	0	1
SL	505	2350	0	1
SL	424	2450	0	1
SL	419	1425	0	1
SL	355	1700	0	1
Ran	295	1400	0	0
Ran	362	1600	0	0
Ran	329	1700	0	0
Ran	358	1875	0	0
Ran	249	1100	0	0
Ran	269	1550	0	0
Ran	455	2350	0	0
Ran	374	2450	0	0
Ran	369	1425	0	0
Ran	305	1700	0	0

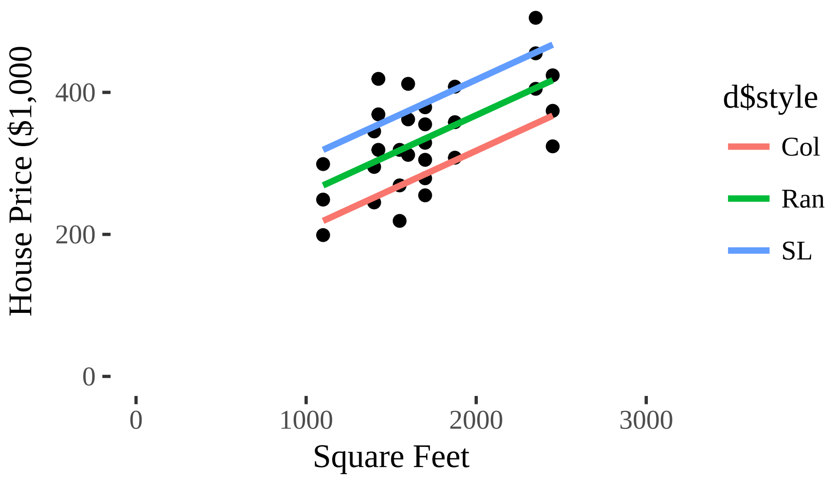
```
##
## Call:
## lm(formula = d$price ~ d$sqft)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.388 -45.342  -0.295  41.838 114.333
##
## Coefficients:
```

```
##             Estimate Std. Error t value
## (Intercept) 148.24833   46.21500   3.208
## d$sqft        0.10977    0.02626   4.181
##            Pr(>|t|)
## (Intercept) 0.003339 **
## d$sqft      0.000258 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 57.01 on 28 degrees of freedom
## Multiple R-squared:  0.3843, Adjusted R-squared:  0.3623
## F-statistic: 17.48 on 1 and 28 DF,  p-value: 0.0002583
```

```
##
## Call:
## lm(formula = d$price ~ d$sqft + d$style)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.388 -29.853  -6.388  38.123  64.333
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  98.24833   33.78449   2.908
## d$sqft        0.10977    0.01829   6.002
## d$styleRan   50.00000   17.75836   2.816
## d$styleSL   100.00000   17.75836   5.631
##            Pr(>|t|)
## (Intercept)  0.00735 **
## d$sqft      2.45e-06 ***
## d$styleRan   0.00916 **
## d$styleSL   6.41e-06 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 39.71 on 26 degrees of freedom
## Multiple R-squared:  0.7226, Adjusted R-squared:  0.6906
## F-statistic: 22.58 on 3 and 26 DF,  p-value: 2.074e-07
```

```
##
## Call:
## lm(formula = d$price ~ d$sqft + d$Col + d$SL)
##
```

```
## Residuals:
##      Min      1Q  Median      3Q     Max
## -49.388 -29.853  -6.388  38.123  64.333
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept) 148.24833   33.78449   4.388
## d$sqft        0.10977    0.01829   6.002
## d$Col1      -50.00000   17.75836  -2.816
## d$SL1        50.00000   17.75836   2.816
##             Pr(>|t|)
## (Intercept) 0.000169 ***
## d$sqft      2.45e-06 ***
## d$Col1      0.009164 **
## d$SL1       0.009164 **
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 39.71 on 26 degrees of freedom
## Multiple R-squared:  0.7226, Adjusted R-squared:  0.6906
## F-statistic: 22.58 on 3 and 26 DF,  p-value: 2.074e-07
```



Logistic Regression

- Used when the dependent variable Y is binary (i.e., Y takes on only two values)
- Examples
  - Customer prefers Brand A or Brand B
  - Employee chooses to work full-time or part-time

  - Loan is delinquent or is not delinquent
  - Person voted in last election or did not

- Logistic regression allows you to predict the probability of a particular categorical response

- Logistic regression is based on the odds ratio, which represents the probability of an event of interest compared with the probability of not an event of interest

$$Odds\ Ratio = \frac{probability\ of\ an\ event\ of\ interest}{1 - probability\ of\ an\ event\ of\ interest} \tag{16}$$

- The logistic regression model is based on the natural log of this odds ratio

- Logistic Regression Model

$$ln(odds\ ratio) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + e_i \tag{17}$$

- Logistic Regression Equation

$$ln(odds\ ratio) = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} \tag{18}$$