Chapter 11. Linear Regression
Multiple Linear Regression

J.C. Wang

Outline

Multiple Linear Regression
  Multiple Linear Regression Modeling
  Issues for MLR

Analysis of Real Estate Data
  Summary Statistics
  Statistical Inferences
  Residual Analysis and Model Goodness
Multiple Linear Regression

The multiple linear regression analysis concentrates on a model that has more than one independent (explanatory) variable. The independent variables are used to predict the dependent variable.

Multiple Linear Regression Modeling

- Purpose of multiple regression analysis is prediction
- Model: \( y = b_0 + b_1x_1 + \ldots + b_n x_n \); where \( b_i \) are the slopes, \( y \) is a dependent variable and \( x_i \) is an independent variable.
- Correlation coefficient, \( r_{ij} \).
- Coefficient of determination, \( R^2 \) (or multiple \( R^2 \)).
▶ Standard error of the estimated regr. line, $s$.
▶ Test hypothesis of slopes, $p$-value.
▶ Slope confidence intervals.
▶ Residual calculation.

MLR
Real Estate Example

A realtor in a suburban town would like to study the relationship between the size of a single-family house (as measured by the number of rooms) and the selling price of the house. The study is to be carried out in two different neighborhood, one on the east side (code=0) of the town and the other on the west side (code=1). A random sample of 8 houses was selected with the following results:
MLR: Real Estate Example

continued

<table>
<thead>
<tr>
<th>Selling Price</th>
<th># of Rooms</th>
<th>Neighborhood</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.2</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>109.6</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>119.3</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>135.3</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>108.5</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>126.7</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>138.8</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>143.8</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

Things to Consider in MLR

- Scatter plots
- Correlation
- MLR equation, $R^2$, CIs of slopes, and residuals
Correlations

<table>
<thead>
<tr>
<th>Correlation Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>price</td>
</tr>
<tr>
<td>price</td>
</tr>
</tbody>
</table>

- correlation coefficient of price and rooms is 0.9682 and sign is positive.
- correlation coefficient of price and hood is 0.4537 and sign is positive.
- correlation coefficient of rooms and hood is 0.2750 and sign is positive.
- There is no multi-collinearity.

Regression Equation

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>coeff</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>rooms</td>
</tr>
<tr>
<td>hood</td>
</tr>
</tbody>
</table>

Using the information in the above table, write the estimated regression equation for this problem.

price = 38.99 + 10.21 × rooms + 6.19 × hood
Using the estimated regression equation, predict the average selling \textit{price} for a house with nine (9) \textit{rooms} that is located on the east side (\textit{code}=0) of the town.

\[
\text{price} = 38.99 + 10.21 \times 9 + 6.19 \times 0 = 130.88
\]

Is the previous prediction extrapolation? Why or why not? No, because number of \textit{rooms} = 9 is between 6 and 10, and \textit{hood} = 0 is between 0 and 1.
Interpretation of Slopes

▶ Holding \textbf{hood} constant, as \textbf{rooms} increases by one, \textbf{price} increases by 10.21
▶ Holding \textbf{rooms} constant, as \textbf{hood} increases by one, \textbf{price} increases by 6.19

Multiple R-Squared

\textbf{Excel Output}

\begin{tabular}{|l|c|}
\hline
Regression Statistics & \textbf{R-square} \\
\hline
\textbf{R-square} & 0.9755 \\
Adjusted R-sq & 0.9657 \\
Standard Error & 3.0238 \\
Observations & 8 \\
\hline
\end{tabular}

Interpret the meaning of the coefficient of multiple determination in this problem. What measure did you use to answer this questions?

\[ R^2 = 0.9755. \]

98\% of the variation in \textbf{price} is explained by this model.

\[ R^2 = \frac{SSR}{SST} = \frac{1818.28}{1863.99} = 0.9755 \]
Standard Error of Estimated Equation

**Excel Output**

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R-square</td>
<td>0.9755</td>
</tr>
<tr>
<td>Adjusted R-sq</td>
<td>0.9657</td>
</tr>
<tr>
<td><strong>Standard Error</strong></td>
<td><strong>3.0238</strong></td>
</tr>
<tr>
<td>Observations</td>
<td>8</td>
</tr>
</tbody>
</table>

What is the standard error of the estimated regression equation? Include the unit of measurement in your answer. $s = 3.0238$ thousands of dollars

\[
s = \sqrt{\frac{SSE}{n-3}} = \sqrt{\frac{45.72}{5}} = 3.0238 \ (= \sqrt{MSE} = \sqrt{9.14})
\]

Significance of Regression Equation

Determine whether there is a significant relationship between selling price and the two explanatory variables at the 5% level of significance.


**ANOVA**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>1818.28</td>
<td>909.14</td>
<td>99.43</td>
<td>9e−5</td>
</tr>
<tr>
<td>Residual</td>
<td>5</td>
<td>45.72</td>
<td>9.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>1864.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$H_0 : \beta_1 = \beta_2 = 0$ vs. $H_a :$ at least one slope is not zero

$F = 99.43, p$-value = 9e−5, Reject $H_0$. 
Significance of Slopes

Determine whether each explanatory variable makes a significant contribution to the regression model. On the basis of these results, indicate the regression model that should be used in this problem at a 5% level of significance. Be sure to state hypotheses, test statistic, p-value, and conclusion.

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>coeff</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>rooms</td>
</tr>
<tr>
<td>hood</td>
</tr>
</tbody>
</table>

Significance of Slopes

continued

- **rooms**: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, $t = 12.5$, $p$-value = $6e−5$, Reject $H_0$.
- **hood**: $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$, $t = 2.8$, $p$-value = 0.039, Reject $H_0$.
- Since both slopes are significantly different from zero, both slopes are making a significant contribution to this model.
Confidence Intervals of Slopes

Set up 95% confidence interval estimates of the population slope for the relationship between selling price and number of rooms and neighborhood.

<table>
<thead>
<tr>
<th>Regression Coefficients</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>22.80</td>
<td>55.19</td>
</tr>
<tr>
<td>rooms</td>
<td>8.12</td>
<td>12.31</td>
</tr>
<tr>
<td>hood</td>
<td>0.47</td>
<td>11.91</td>
</tr>
</tbody>
</table>

- 95% CI for rooms: (8.12, 12.31)
- 95% CI for hood: (0.47, 11.91)

Residual Plot

Describe what you see on the residual plot.

There is a random distribution of points.
Goodness of Fitted Model

Is the linear model a good fit for this data? Consider the following:

- $R^2 = 0.9755$, $s = 3.0238$
- Both $\text{rooms}$ and $\text{hood}$ are making a significant contribution to the linear model.
- The residual plot shows a random distribution of points.

Yes, this model appears to be appropriate for this data.

Calculation of Predicted Value and Residual

For this data set, the first set of observations is $\text{price} = 98.2$, $\text{rooms} = 6$, $\text{hood} = 0$. Calculate the residual, i.e., compute the difference between the observed $\text{price}$ and the predicted $\text{price}$. Show all the formulas and calculations.

\[
\hat{y}_1 = 38.99 + 10.21 \times 6 + 6.19 \times 0 = 100.25
\]
\[
e_1 = y_1 - \hat{y}_1 = 98.2 - 100.25 = -2.05
\]
iClicker question 11.5