21. MULTIPLE REGRESSION

Multiple regression provides a method of predicting a response variable $y$ from two or more explanatory $x$ variables.

Multiple regression is potentially much more useful than simple regression, because there are often several important explanatory factors, instead of just one.

The purposes of multiple regression are the same as in simple regression:

(1) Describing the relationship.

- Salary ($y$) of company employees may depend on several factors, such as years of experience ($x_1$), years of education ($x_2$), and gender ($x_3$, represented as 0 or 1 to distinguish male and female).

Describing and understanding how these $x$ factors influence $y$ would provide important evidence in a gender discrimination lawsuit.

The regression coefficient for gender would give an estimate of how large the salary gap is for men and women after adjustment for age and experience.

(2) Predicting a new observation.

- Beer drinkers may want to understand how calories per 12 oz serving ($y$) depends on % Alcohol ($x_1$) and Carbohydrates ($x_2$).

<table>
<thead>
<tr>
<th>Brand</th>
<th>% Alcohol</th>
<th>Calories/12 oz</th>
<th>Carbohydrates (g.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor Porter</td>
<td>5.60</td>
<td>209</td>
<td>0.00</td>
</tr>
<tr>
<td>Anchor Steam</td>
<td>4.90</td>
<td>153</td>
<td>16.00</td>
</tr>
<tr>
<td>A. Busch Nat. Light</td>
<td>4.20</td>
<td>95</td>
<td>3.20</td>
</tr>
<tr>
<td>A. Busch Nat. Ice</td>
<td>5.90</td>
<td>157</td>
<td>8.90</td>
</tr>
<tr>
<td>Aspen Edge</td>
<td>4.10</td>
<td>94</td>
<td>2.60</td>
</tr>
</tbody>
</table>

- Suppose we study the finishing time ($y$) for horses racing at Belmont Park. If we know how $y$ depends on the length of the race ($x_1$), the post position ($x_2$), the horse's finishing time in his previous race ($x_3$), the length of his previous race ($x_4$), his total life time winnings ($x_5$) and other factors, we could try to forecast the winning times of all horses in tomorrow's race, and thereby forecast the winner.

- Predict unemployment ($y$) using current GDP ($x_1$) as well as current unemployment ($x_2$). We may want a point prediction (the fitted value) or an interval prediction (PI).
(3) Adjusting and controlling a process.

• If we know how ice cream sales \( (y) \) depend on fat content \( (x_1) \), price \( (x_2) \), sugar content \( (x_3) \), and other factors, we can reformulate our product and set its price to (hopefully) increase profits.

Suppose we have \( k \) explanatory variables \( x_1, \ldots, x_k \) which we feel may be useful in predicting \( y \).

The multiple linear regression model is

\[
y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon
\]

where \( \alpha \) is the intercept, or constant term, \( \beta_1, \ldots, \beta_k \) are the true regression coefficients, and \( \varepsilon \) is a random error variable.

The error term \( \varepsilon \) is included to take account of the randomness in our response \( y \). The values of \( \varepsilon \) are assumed to be independent and normally distributed with mean zero and variance \( \sigma^2 \) which does not change with the explanatory variables.

The response surface (or true regression function)

\[
\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k
\]

accounts for the part of \( y \) which depends systematically on the explanatory variables. Since there are several explanatory variables, the response surface is not a line, but a flat surface called a hyperplane.

So the observed response \( y_i \) is the sum of a systematic part (the response surface) and a random part (the error). We seek to estimate the response surface based on our data.

The least squares estimators of the parameters \( \alpha, \beta_1, \ldots, \beta_k \) are the values \( \hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k \) which minimize the residual sum of squares,

\[
SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \text{where} \quad \hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k.
\]

The assumptions of the model may be checked by plotting the residuals \( e_i = y_i - \hat{y}_i \) against the fitted values \( \hat{y}_i \).

The plot should show no structure, no strong outliers, and no tendency for the variability to change with \( \hat{y} \).

• You can obtain the least squares estimates from the Minitab output.
Interpreting The Coefficients

Suppose we want to predict the selling price (\( y \); Thousands of Dollars) of a single family home based on the house size (\( x_1 \); Hundreds of Square Feet), the age (\( x_2 \); Years), and the lot size (\( x_3 \); Thousands of Square Feet). The data are in House.MTP.

The estimated regression coefficients are

\[
\hat{\beta}_1 = 41.46, \hat{\beta}_2 = -2.36, \hat{\beta}_3 = 48.31.
\]

The fitted response surface is therefore

\[
y = -161 + 41.46x_1 - 2.36x_2 + 48.31x_3.
\]

Regression Analysis: Price versus Size, Age, Lot Size

Model Summary

<table>
<thead>
<tr>
<th>S</th>
<th>R-sq</th>
<th>R-sq(adj)</th>
<th>R-sq(pred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.9399</td>
<td>91.61%</td>
<td>89.32%</td>
<td>87.66%</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-161</td>
<td>191</td>
<td>-0.84</td>
<td>0.418</td>
</tr>
<tr>
<td>Size</td>
<td>41.46</td>
<td>7.51</td>
<td>5.52</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>-2.36</td>
<td>8.81</td>
<td>-0.27</td>
<td>0.794</td>
</tr>
<tr>
<td>Lot Size</td>
<td>48.31</td>
<td>9.01</td>
<td>5.36</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Regression Equation

\[
\text{Price} = -161 + 41.46 \text{ Size} - 2.36 \text{ Age} + 48.31 \text{ Lot Size}
\]

Interpretation of the Regression Coefficients:

The estimated regression coefficient for a given \( x \)-variable estimates the effect of that \( x \) variable on \( E[y|x] \) after an adjustment has been made for the other \( x \)-variables.

In other words, the coefficient \( \hat{\beta}_j \) (for the \( j \)-th \( x \)-variable, \( x_j \)) estimates how much larger you expect \( y \) to be if \( x_j \) is increased by one unit and all other \( x \)-variables are held fixed.

Note that the meaning and interpretation of a given regression coefficient depends on what other variables are included in the model.
Eg: In the above example, it is not wise to try to interpret the term $\hat{\beta}_2$ since we have no data for $x_1, x_2, x_3$ near zero. Nevertheless, it is almost always a good idea to include a constant term ($\alpha$) in the model, as this will usually produce better predictions.

Since we obtained $\hat{\beta}_2 = -2.36$, we can say that for a given house size and a given lot size, we estimate that each year of age reduces the mean selling price by $2,360$.

If we run a *simple* regression of $y$ using the single explanatory variable $x_2$ (Age), we find that

$$\hat{y} = 1125 - 27.3x_2.$$  

Thus, *without* controlling for house size and lot size, we estimate that each additional year of age reduces the mean selling price by an average of $27,300$.

This is very different from the $2,360$ figure obtained above, which adjusts for house size and lot size.

As we see, the interpretation of a regression coefficient, as well as its numerical value, depends on what other variables are used in the model. This is one reason why we need to think carefully about which variables to use.

\[
\text{Regression Analysis: Price versus Age} \\
\text{Model Summary} \\
\begin{array}{cccc}
S & R$^2$ & R-sq(adj) & R-sq(pred) \\
186.875 & 27.13\% & 21.53\% & 5.87\% \\
\end{array}
\]

\[
\text{Coefficients} \\
\begin{array}{ccccc}
\text{Term} & \text{Coef} & \text{SE Coef} & \text{T-Value} & \text{P-Value} \\
\text{Constant} & 1125 & 118 & 9.55 & 0.000 \\
\text{Age} & -27.3 & 12.4 & -2.20 & 0.046 \\
\end{array}
\]

\[
\text{Regression Equation} \\
\text{Price} = 1125 - 27.3 \text{ Age}
\]