NOTES ON COMPARING REGRESSIONS

Suppose that you are interested in the effect of a number of housing attributes, \((x_1, \ldots, x_k)\), on housing prices, \(Y\), and wish to compare the relative significance of these attributes for two different cities, say Philadelphia and Chicago. To do so, you have collected housing data, \((y_i^1, x_{i1}^1, \ldots, x_{ik}^1), i = 1, \ldots, n_1\), from Philadelphia and comparable housing data, \((y_i^2, x_{i1}^2, \ldots, x_{ik}^2), i = 1, \ldots, n_2\), from Chicago. Each of these regressions can be run separately, but it is difficult to compare them directly. For example, while one can determine whether the coefficient \(\beta_j\) for attribute \(x_j\) is significant in Philadelphia and/or Chicago, one cannot determine whether they are significantly different. To do so, they must be part of the same regression.

This can be accomplished in exactly the same way as in the class example where wages were compared for male and female employees within a single large firm. In the present case, the attribute “sex” can be replaced by “city”. More formally, let the “city” indicator variable, \(c\), be defined by

\[ c = \begin{cases} 1, & \text{city Philadelphia} \\ 0, & \text{city Chicago} \end{cases} \]

and consider the multiple regression model

\[ Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \alpha_0 c_i + \sum_{j=1}^{k} \alpha_j (c_i x_{ij}) + \epsilon_i, \quad i = 1, \ldots, n_1 + n_2 \]

applied to the combined data set \((y_i, c_i, x_{i1}, \ldots, x_{ik}), i = 1, \ldots, n_1 + n_2\), where \(c_i\) denotes the city location of each house \(i\) [say with the first \(n_1\) rows corresponding to the Philadelphia houses \((c_i = 1)\) and the last \(n_2\) rows corresponding to the Chicago houses \((c_i = 0)\)]. If \(k = 2\) then the column headings for this regression in JMPIN might look something like the following:

<table>
<thead>
<tr>
<th>Y</th>
<th>X1</th>
<th>X2</th>
<th>C</th>
<th>(C)X1</th>
<th>(C)X2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To analyze the results of this regression, observe first that for each Philadelphia house $i$ ($c_i = 1$) the model in (2) has the form

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ij} + \epsilon_i = (\beta_0 + \alpha_0) + \sum_{j=1}^{k} (\beta_j + \alpha_j) x_{ij} + \epsilon_i$$

and for each Chicago house $i$ ($c_i = 0$) it has the form

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Hence it is clear that for each $j = 1, \ldots, k$, the parameter $\alpha_j$ in fact represents the difference between the slopes $[\left(\beta_j + \alpha_j\right) - \beta_j]$ for attribute $x_j$ in the Philadelphia model (3) and the Chicago (4). If we now let $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_k\right)$ denote the estimated coefficients from this combined regression model, then a low $P$-value for $\hat{\alpha}_j$ now indicates that there is a significant difference between the effects of attribute $x_j$ on housing prices in Philadelphia versus Chicago. For example, if $x_j$ denotes “number of bedrooms” and if $\hat{\alpha}_j$ is significantly positive, then this would indicate that the addition of one bedroom has a greater effect on expected housing prices in Philadelphia than in Chicago.