

NOTES ON COMPARING REGRESSIONS

Suppose that you are interested in the effect of a number of *housing attributes*, (x_1, \dots, 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, $(y_i^1, x_{i1}^1, \dots, x_{ki}^1)$, $i = 1, \dots, n_1$, from Philadelphia and comparable housing data, $(y_i^2, x_{i1}^2, \dots, x_{ki}^2)$, $i = 1, \dots, n_2$, 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 β_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

$$(1) \quad c = \begin{cases} 1, & \text{city} = \text{Philadelphia} \\ 0, & \text{city} = \text{Chicago} \end{cases}$$

and consider the multiple regression model

$$(2) \quad Y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \alpha_0 c_i + \sum_{j=1}^k \alpha_j (c_i \cdot x_{ij}) + \varepsilon_i, \quad i = 1, \dots, n_1 + n_2$$

applied to the combined data set $(y_i, c_i, x_{i1}, \dots, x_{ki})$, $i = 1, \dots, n_1 + n_2$, 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:

Y	X1	X2	C	(C)X1	(C)X2

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

$$(3) \quad Y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \alpha_0 + \sum_{j=1}^k \alpha_j x_{ij} + \varepsilon_i = (\beta_0 + \alpha_0) + \sum_{j=1}^k (\beta_j + \alpha_j) x_{ij} + \varepsilon_i$$

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

$$(4) \quad Y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij}$$

Hence it is clear that for each $j = 1, \dots, k$, the parameter α_j in fact represents the *difference* between the slopes $\left[(\beta_j + \alpha_j) - \beta_j \right]$ for attribute x_j in the Philadelphia model (3) and the Chicago (4). If we now let $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k, \hat{\alpha}_0, \hat{\alpha}_1, \dots, \hat{\alpha}_k)$ 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.