

Solution to Selected Problems from HW1

October 4, 2009

1 Exercise 3.3

Here is a general approach to attacking Exercise 3.3.

The setup goes as follows: we have $N = 100$ observations, some values of $\{y_i\}_{i=1}^N$, and up to three regressors $\{x_{1i}, x_{2i}, x_{3i}\}_{i=1}^N$, which are dummy variables that indicate that observations fall within a certain range in $[1, N]$.

In case indicator variables divide the range $[1, N]$ into disjoint intervals (like in parts (a) and (c) of the exercise), it is straightforward to find the OLS estimate of the corresponding values of β . Consider (c), where:

$$\begin{cases} x_{1i} = 1 \{i < 34\} \\ x_{2i} = 1 \{i \in [34, 66]\} \\ x_{3i} = 1 \{i > 66\} \end{cases} .$$

The OLS problem looks as follows:

$$\min_{\{\beta_j\}_{j=1}^3} \sum_{i=1}^N \left(y_i - \sum_{j=1}^3 \beta_j x_{ji} \right)^2, \quad (1)$$

and given the definitions of x , it can be rewritten as:

$$\min_{\{\beta_j\}_{j=1}^3} \sum_{i=1}^{33} (y_i - \beta_1)^2 + \sum_{i=34}^{66} (y_i - \beta_2)^2 + \sum_{i=67}^{100} (y_i - \beta_3)^2,$$

and it should be immediate that:

$$\begin{cases} \hat{\beta}_1 = \frac{1}{33} \sum_{i=1}^{33} y_i \\ \hat{\beta}_2 = \frac{1}{33} \sum_{i=34}^{66} y_i \\ \hat{\beta}_3 = \frac{1}{34} \sum_{i=67}^{100} y_i \end{cases} .$$

which means that OLS estimates are just the average values of y over the subsamples where the corresponding indicator variables are equal to 1.

Now, things get somewhat more tricky in case indicators do not break the $[1, N]$ range into disjoint subgroups. Consider this example, where $1 < a < b < N$:

$$\begin{cases} x_{1i} = 1 \ \forall i \in [1, N] \\ x_{2i} = 1 \ \{i > a\} \\ x_{3i} = 1 \ \{i < b\} \end{cases} \quad . \quad (2)$$

Here x_{1i} is a constant, x_{2i} is an indicator for observations that follow a and x_{3i} is the indicator for observations that precede b . Assuming that we still want to solve (1) for $\{\beta_j\}_{j=1}^3$, it can be done in terms of subsample averages of y , but we need to be a bit more creative.

Partition the interval $[1, N]$ as $[1, a] \cup (a, b) \cup [b, N]$ and denote $\bar{y}_a \equiv \frac{1}{a} \sum_{i=1}^a y_i$, $\bar{y}_{ab} \equiv \frac{1}{b-a} \sum_{i=a+1}^b y_i$, and $\bar{y}_b \equiv \frac{1}{N-b} \sum_{i=b+1}^N y_i$. I then suggest to write the following table:

	$i \in [1, a]$	$i \in (a, b)$	$i \in [b, N]$
\tilde{x}_{1i}	1	0	0
\tilde{x}_{2i}	0	1	0
\tilde{x}_{3i}	0	0	1
x_{1i}	1	1	1
x_{2i}	0	1	1
x_{3i}	1	1	0

Here is how you read the table. We use $\{\tilde{x}_j\}_{j=1}^3$ to denote the ‘‘artificial’’ indicators, which equal 1 if observations are within one of the three intervals denoted in the column titles of the table. We use $\{x_j\}_{j=1}^3$ to denote the real indicators that we face. The zeros and ones in the cells show the value of a corresponding variable on a given interval. So, for example, when $i \in (a, b)$, $\tilde{x}_{2i} = 1$ because we define \tilde{x}_{2i} to be the indicator for observations for which $i \in (a, b)$; and all $\{x_j\}_{j=1}^3 = 1$, since this is how these regressors are defined in (2).

Denote $\{\tilde{\beta}_j\}_{j=1}^3$ – the OLS estimates if we use $\{\tilde{x}_j\}_{j=1}^3$ as regressors. Obviously, $\tilde{\beta}_1 = \bar{y}_a$, $\tilde{\beta}_2 = \bar{y}_{ab}$, and $\tilde{\beta}_3 = \bar{y}_b$ (it follows from our previous discussion on indicators for disjoint subsets of $[1, N]$). The table is useful since it implies the following three equations:

$$\begin{cases} \bar{y}_a = \hat{\beta}_1 + \hat{\beta}_3 \\ \bar{y}_{ab} = \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 \\ \bar{y}_b = \hat{\beta}_1 + \hat{\beta}_2 \end{cases} \quad ,$$

and treating the RHS variables as known, it is easy to solve for the LHS variables.

In this particular case, we get:

$$\begin{cases} \hat{\beta}_1 = \bar{y}_a - \bar{y}_{ab} + \bar{y}_b \\ \hat{\beta}_2 = \bar{y}_{ab} - \bar{y}_a \\ \hat{\beta}_3 = \bar{y}_{ab} - \bar{y}_b \end{cases} \quad .$$

This method is quite general, and can be easily applied if you have more than 3 regressors. Moreover, it helps to understand what happens when you have multiple indicator variables that may partially overlap.