Heckman selection models

We are running two models with Heckman selection regression. The first model is a choice model – are you in the group or not? For example, are you in the labor market (and have a wage) or not. We use predictors to determine this. The second stage then examines the effects of the independent variables on the outcome (wages, for example). Each stage has a residual for each observation, or a set of unknowns for each observation. To test for bias, we examine the relationship between the residuals for the two stages (stage 1 and stage 2). If the unobservables in the selection model are correlated with the unobservables in the stage 2 model, we have biased estimates without correction (or in an OLS model). This is basically saying that unobservables in the selection (or choice) of entering the labor market are also affecting the stage 2 model. If the unobservables in stage 1 are unrelated to the unobservables in stage 2, then we are saying that stage 1 does not affect stage 2 results. This is another way of saying that selection into the sample of stage 2 is a random process, unaffected by different unobservables. If we can pick all the right variables for our models, and leave few unobservable variables that affect our outcome, then chances are good that we will not have selection bias.

When rho is positive, this indicates that unobservables are positively correlated with one another. Thus, in a wage model, if ability is an unobservable, and it is positively related to working (stage 1), and positively related to wages (stage 2), we will find a positive rho coefficient.

When rho is negative, this indicates that unobservables are negatively correlated with one another. Thus, in a wage model, having long hair is unobserved, and is negatively related to the choice of work but is positively related to dollar wages, then rho will be negative.

At the very bottom of the regression output for a Heckman selection model examining wages we will have estimates for a number for the following:

\[
\begin{align*}
\text{lambda} & | \quad .141005 \\
\text{rho} & | \quad .26729 \\
\text{sigma} & | \quad .52753667 \\
\text{lambda} & | \quad .14100497
\end{align*}
\]

I give you some such estimates above.

The adjusted standard error for the wage equation regression is given by sigma=0.527 and the correlation coefficient between the unobservables that determine selection into waged employment and the unobservables that determine the wage is given by rho=0.267. The estimated selection coefficient lambda = sigma × rho = 0.527 × 0.267 = 0.141.

The next thing to determine is how to interpret the estimated selection effect itself. In order to do this we need to compute the average selection or truncation effect. We first need to get the average value for the selection term for the sample of women that are in
waged employment. The mills ratio is determined by a process we went over last week. The summary statistics on the inverse Mills ratio are given by:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>706</td>
<td>1.1165</td>
<td>.5217</td>
<td>.2645</td>
<td>2.863</td>
</tr>
</tbody>
</table>

The average truncation effect is computed as $\lambda \times [\text{average mills value}] = 0.141 \times 1.117 = 0.157$. This gives by how much the conditional wages are shifted up (or down) due to the selection or truncation effect. The interpretation of this is that a woman with sample average characteristics who selects (or is selected) into waged employment secures $[\exp(0.157) - 1] \times 100 = 17.0\%$ higher hourly wages than a woman drawn at random from the population with the average set of characteristics. Thus, the numerical values suggest there is positive selection or truncation effects in these data and those who select into waged employment get higher wages than a random drawing from the population of women with a comparable set of characteristics would get. However, this value is dependent on whether or not there is a statistically significant effect of selection, or is the chi-square value for rho statistically significant. If it is not, we would find that there no effects of selection (those who select into the wage sample have no higher wages relative to those with average characteristics drawn at random from the population).

Parts of this were taken from:
Source: www.sussex.ac.uk/Units/economics/qm2/stata_4s.doc