Using PGA Tour Results to Illustrate the Effects of Selection Bias

Amanda L. Griffith & Todd A. McFall

To cite this article: Amanda L. Griffith & Todd A. McFall (2013) Using PGA Tour Results to Illustrate the Effects of Selection Bias, The Journal of Economic Education, 44:3, 238-248, DOI: 10.1080/00220485.2013.795456

To link to this article: http://dx.doi.org/10.1080/00220485.2013.795456

Published online: 27 Jun 2013.

Submit your article to this journal

Article views: 153

View related articles
ECONOMIC INSTRUCTION

Using PGA Tour Results to Illustrate the Effects of Selection Bias

Amanda L. Griffith and Todd A. McFall

This lesson is designed to give students in a wide variety of undergraduate classes the opportunity to identify the problems inherent in analyzing imperfect data. Students are asked to analyze data, which the authors provide, from the Professional Golfers’ Association Tour to ponder the link between golfers’ skills in different facets of the game and their relative performances in certain tournaments. Selection issues are inherent in the data because of Tour policies regarding golfers’ tournament entry decisions. By the end of the lesson, students should better understand the importance of gaining institutional knowledge to improve the results of a study. Students are asked to mitigate the selection bias with the Heckman two-step procedure, a tool used in studies of labor, insurance, and education markets.

Keywords  econometrics lesson, Heckman two-step, labor economics, selection

JEL codes  A22, A23

Students in economics classes are often asked to ponder the relationship between some variable X and another variable Y. When data are involved in this type of exercise, experienced economists would likely want to know about the nature of the data, as they understand there are a number of issues that if left unaddressed, might create biased results. Students, however, do not tend to have much experience in dealing with data and might need cajoling to believe the econometrician’s war stories of bias.

In this article, we provide instructors of econometrics, labor economics, and sports economics classes the opportunity to allow their students to discern the importance of making investments in learning about processes by which data are collected and how those processes might change the way they perceive the relationship between variables X and Y. Our lesson centers around the
seemingly simple question: How is a professional golfer’s relative performance in a tournament related to the skill he exhibits in different facets of the game? This seemingly simple question regarding the relationship between variables X and Y is made more difficult to answer because of institutional details, which students might not likely know, related to the way golfers on the Professional Golfers’ Association (PGA) Tour enter tournaments. Without investing in learning about some background knowledge regarding Tour policies, students’ answers to the question could be inaccurate due to selection bias issues. So, in the largest sense, at the end of this lesson, students should understand the need to invest in learning about institutional details that affect the data they study because they will have to deal with these dirty data. In a specific sense, because our study’s subjects, professional golfers, are often able to select freely the tournaments in which they would like to compete, students can gain experience in recognizing selection problems and understanding how the Heckman two-step procedure mitigates the issue.

Our lesson is designed to be used whole or in parts, depending upon the needs of the instructor who chooses to use it. We provide links to the data used in the lesson and provide mathematical and intuitive discussions about selection bias in the context of professional golf. Instructors of an upper-level econometrics class could use the entire lesson if they wanted to create an interactive activity in which students could learn about selection issues, while professors of labor economics or sports economics could simply use the results to discuss the effect that controlling for selection bias has on the estimates of the relationships between golfer performance and golfer skill.

Below, we provide a brief discussion of selection bias and the use of the Heckman two-step procedure. We then describe some policies of the PGA Tour and how those policies create selection bias in the golfer performance data that we provide. We illustrate the application of the Heckman two-step procedure to mitigate selection bias in our data and end by discussing the results and suggesting questions that students in various undergraduate economics classes could be asked to further their understanding of various aspects of this lesson.

**LESSON—EXPLAINING PERFORMANCE OF EXEMPT GOLFERS ON PGA TOUR**

Selection bias is a specification error that arises when the dependent variable is limited, either because the data are incomplete, or the values are censored. The corruption of the data occurs because certain observations are not recorded for one of two reasons. Either the manner in which data were collected is such that certain observations are ignored, or the individuals recorded in the data do not participate in the activity of interest during a given time period. No matter the reason for the existence of selection bias, when the problem is not mitigated, the result is an inaccurate analysis that may lead to faulty conclusions in studies of labor markets, goods markets, and insurance markets.¹

One popular method of mitigating the effects of selection bias is the Heckman two-step procedure, outlined originally by the eponymous author in his seminal article on the subject (Heckman 1979). When one uses least-squares to analyze data plagued by selection bias, the missing observations may cause biased estimates because the error term is correlated with the probability of an individual either participating in the market or being recorded as having participated in the market. Heckman’s procedure calls for one to first estimate with a selection equation the effect that individual characteristics have on the probability of the individual participating in a market...
(and therefore being recorded in the data). This part of the procedure requires the proper identification of an exclusion restriction, a variable that is an economically and statistically important variable in the selection equation but does not belong in the least-squares estimate. The second step then involves the inclusion of the inverse Mills ratio, which is calculated using the results of a probit analysis, in the least-squares model. Done properly, the result mitigates the bias, and under certain conditions, can produce consistent estimates.

The procedures on the PGA Tour allow for some golfers to choose the tournaments that they would like to enter. If they do not wish to play, Tour rules do not prohibit them from making this choice. So, not all golfers play in all Tour events; consequently, the data recording golfer performance are riddled with selection issues. This fact makes Tour data a great place for students to learn to recognize why selection issues exist in certain settings and how to control for the issues in order to obtain unbiased results.

With these institutional details in mind, we begin the lesson by asking students to contemplate the following question: What is the relationship between a golfer’s relative performance in a given tournament and skills levels he exhibited in various facets of the game during the tournament? To answer this question, we want to build upon previously published studies of the relative performance and earnings of golfers on the PGA Tour. These studies generally attempt to explain golfers’ performances or earnings by using as independent variables measurements from the different skills that a golfer must master in order to be successful in professional golf (driving, approach shots, chipping, and putting). Alexander and Kern (2005) focused on determining the extent to which the importance of driving and putting has changed in determining a golfer’s earnings. Moy and Liaw (1998) studied the determinants of golfer earnings. Finally, Shmanske (1992) studied the human capital formation of the PGA Tour golfers.

We expand the general empirical models developed in these articles to account directly for sample selection, which exists due to Tour policies regarding tournament entry procedures. We think that students will be quick to use regression techniques to answer the question we pose but will be slow to recognize the need to control for selection. The results we discuss below show that controlling for selection issues in this setting alters our estimates and are interesting in light of the questions we ask students to ponder. Unfortunately, we cannot control for selection issues without investing some time in understanding why these issues exist in the first place, an important point for students to grasp. To ease preparation costs for instructors, we provide data that can be used in class as well as a primer for the uninitiated on the way Tour events are conducted and its policy of rewarding players who meet certain performance criteria.

We collected data from the 2006–8 PGA Tour seasons from the exceptional ShotLink database, which is a record of the characteristics of every shot attempted in Tour events. In each of these years, the Tour’s season consisted of about 45 events per year, roughly one per week, starting in early January and ending in November. There are two types of tournaments on the schedule: tournaments that are owned and operated by the PGA Tour and those that are owned by other entities, yet are still recognized as part of the Tour’s schedule. The large majority of these 45 tournaments are organized and owned by the PGA Tour, which operates the tournaments in a highly uniform fashion. These tournaments are contested in places like San Diego (Farmers Insurance Open), Phoenix (Waste Management Phoenix Open), Charlotte (Wells Fargo Championship), and San Antonio (Valero Texas Open). At the start of each tournament, between 125 and 180 golfers compete in two 18-hole rounds. After two rounds, the field is cut to the lowest scoring 70 golfers, who then play two more rounds to determine relative finish within the field, which is
TABLE 1
Descriptive Statistics for 2008 PGA Tour Tournaments in Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purse (Millions $2012)</td>
<td>5.54</td>
<td>0.935</td>
<td>3.19</td>
<td>6.8</td>
</tr>
<tr>
<td>Number of players</td>
<td>143.71</td>
<td>12.72</td>
<td>128</td>
<td>180</td>
</tr>
<tr>
<td>N</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based upon the golfers’ four-round aggregate score. Golfers are rewarded prize money from the tournament’s purse—the pool of money that a tournament collects from corporate sponsorships and ticket sales. Although the sizes of the purses vary widely across tournaments, the method for disbursing prize money does not, as the winner always receives 18 percent of the tournament’s total purse, the runner-up receiving 10.8 percent, third receiving 6.8 percent, and so on, to the 70th-place golfer, who receives 0.2 percent of the purse.3

Of the 45 events that the PGA Tour incorporates into its season, there are about 14 tournaments that are owned by other entities or are owned by the PGA Tour and operated in a fashion different than the “typical” tournament we described above. Even though the Tour recognizes golfer earnings and performances in the events, we chose not to include these events in the dataset because the way the tournament fields are populated deviates significantly from the typical Tour event. Thus, we excluded from our dataset the four so-called “major” championships that are conducted each season—the Masters (owned by the Augusta National Golf Club), the U.S. Open (owned by the U.S. Golf Association), the British Open (owned by the Royal & Ancient Society), and the PGA Championship (owned by the PGA of America), the four World Golf Association-owned events, the Players Championship (owned by the PGA Tour), and five “invitational” tournaments that have unique entry procedures.

This variation in purse size can be seen in table 1, where we show descriptive statistics for purse size, expressed in 2012 terms with Consumer Price Index data,4 and the number of players per tournament for Tour events that took place in 2008.

In our dataset, we observe, by round, the score and measurements of golfer performance in various skill categories, but only for golfers who chose to enter a given tournament. This seemingly obvious detail is important, because for any given tournament in our dataset, there are two types of golfers, so-called exempt golfers and nonexempt golfers. Exempt golfers have met one or more performance criteria and are described as such because they are exempt from qualifying for tournaments, meaning they can enter any of the tournaments in our dataset that they would like. Nonexempt golfers do not have this luxury. Because Tour regulations create a priority list for tournament entries, nonexempt players must wait in line for an opportunity to play in any one of these tournaments. Without knowing that exempt golfers can select into the tournaments of their choice, students probably will not recognize the need to control for selection bias. Investing in learning about the institutional details of the Tour gives students a chance to understand how to control for selection bias in this particular instance and to reap unbiased results, which are the benefits of more careful analysis.

How does a golfer reach different priorities of exemption status? We describe several ways below, all of which we controlled for in our dataset. As in many other worlds, the world of professional golf is characterized by the saying “to the victor goes the spoils.” The exemption
system reflects this belief because the criteria create a hierarchy that is a monotonic function of the difficulty and prestige associated with meeting a criterion. The golfers who have met the most difficult challenges are granted priority for entering the tournaments that we analyze. The golfers who have met the next most difficult challenges are next in line and so forth. Because of this hierarchy, not all golfers who have an exemption status can enter all the tournaments in which they would like to compete because golfers with higher priority have filled up the field.

While constructing the data, we attempted to reflect this hierarchy because we identified golfers who earned a level of priority such that they could enter—without reservation—any of the tournaments we included in our dataset. To accomplish this, we identified golfers who met one or more of four exemption criteria. First, the highest exemption status is granted to the last five winners of the U.S. Open, the PGA Championship, The Masters, the British Open, or the Players Championship. Next in line are last three winners of a World Golf Championship tournament or the Tour Championship. Third on the list is any golfer who won a “typical” Tour-owned event within the last two years. Last on our list are the players who ranked among the top 125 money winners on the Tour. (The money rankings are an aggregate total of money won on Tour-recognized events over an entire season.) So, all golfers who were ranked among the top 125 money winners in 2005, for instance, could play in all the Tour-owned tournaments in 2006. Any golfers who earned exemption status beneath these golfers were not guaranteed spots in the tournaments in our dataset; therefore, we excluded them from our list of exempt members.

The Econometrics of Selection Bias on the PGA Tour

Once students have gained some institutional knowledge about the PGA Tour, they could easily hypothesize that a golfer’s score might be a function of his ability to hit tee shots, approach shots, recovery shots, and putts. Anticipating this hypothesis, we constructed the data for any particular tournament as follows:

\[ \text{Rank}_{i, t, y} \text{ Skills}_{i, t, y} \]

\[ \text{Rank}_{2, t, y} \text{ Skills}_{2, t, y} \]

\[ \ldots \]

\[ \text{Rank}_{N, t, y} \text{ Skills}_{N, t, y}, \]

where \( \text{Rank}_{i, t, y} \) refers to the relative rank of golfer \( i \) after two rounds of tournament \( t \) in year \( y \) and \( \text{Skills}_{i, t, y} \) refers to observations on golfer skills through two rounds of a tournament.

After the data are organized and explained, students should be asked to develop a least-squares regression model that allows them to measure the impact that certain skills have on a golfer’s relative ranking. We recommend that they follow the methods employed in Alexander and Kern (2005). If they do, then they will decide upon a model similar to equation (1) below:

\[ R_{i, t, y} = X_{1i, t, y} \beta_1 + u_{1i, t, y} \] (1)

where \( R_{i, t, y} \) is the relative ranking of golfer \( i \) after two rounds of tournament \( t \) in season \( y \). The vector \( X_{1i, t, y} \) is a six-column vector of golfer performance characteristics, and \( u_{1i, t, y} \) is the error term. The performance variables contained in \( X_{1i, t, y} \) are defined as
After students use OLS to estimate equation (1), they should check their results against the results we present in column 1 of table 2. Students can be directed to note that of the six golfer characteristics included in the regression, all but one have significance levels of 1 percent or better, and the signs of the coefficients are sensible. (Remember, a golfer would want to be in first place, not 70th place.) The coefficient for experience is significant and positive, which might

```
<table>
<thead>
<tr>
<th>Variable</th>
<th>1–Rank (OLS) (N = 5,584)</th>
<th>2–Rank (OLS) w/purse (N = 5,584)</th>
<th>3–Rank (Selection) (N = 7,037)</th>
<th>4–Selection equation (N = 7,037)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>0.170*** (0.056)</td>
<td>0.124** (0.055)</td>
<td>0.094 (0.074)</td>
<td>−0.010*** (0.003)</td>
</tr>
<tr>
<td>Purse</td>
<td></td>
<td>−7.223*** (0.417)</td>
<td></td>
<td>0.289*** (0.017)</td>
</tr>
<tr>
<td>Drive distance</td>
<td>−0.036*** (0.012)</td>
<td>−0.079*** (0.012)</td>
<td>−0.035** (0.014)</td>
<td></td>
</tr>
<tr>
<td>Drive accuracy</td>
<td>−0.034 (0.028)</td>
<td>−0.124*** (0.028)</td>
<td>−0.048 (0.036)</td>
<td></td>
</tr>
<tr>
<td>GIR</td>
<td>−1.524*** (0.035)</td>
<td>−1.617*** (0.035)</td>
<td>−1.502*** (0.044)</td>
<td></td>
</tr>
<tr>
<td>Avg putts GIR</td>
<td>1.711*** (0.030)</td>
<td>1.756*** (0.029)</td>
<td>1.718*** (0.038)</td>
<td></td>
</tr>
<tr>
<td>Sand saves</td>
<td>−0.041*** (0.010)</td>
<td>−0.042*** (0.010)</td>
<td>−0.034*** (0.013)</td>
<td></td>
</tr>
<tr>
<td>Scrambling</td>
<td>−0.932*** (0.022)</td>
<td>−0.9418*** (0.021)</td>
<td>−0.989*** (0.028)</td>
<td></td>
</tr>
<tr>
<td>Lag drive distance</td>
<td></td>
<td></td>
<td></td>
<td>0.015*** (0.004)</td>
</tr>
<tr>
<td>Lag drive accuracy</td>
<td></td>
<td></td>
<td></td>
<td>0.004*** (0.001)</td>
</tr>
<tr>
<td>Lag GIR</td>
<td></td>
<td></td>
<td></td>
<td>−0.006 (0.007)</td>
</tr>
<tr>
<td>Lag GIR puts</td>
<td></td>
<td></td>
<td></td>
<td>−0.007 (0.006)</td>
</tr>
<tr>
<td>Lag sand saves</td>
<td></td>
<td></td>
<td></td>
<td>−0.004 (0.002)</td>
</tr>
<tr>
<td>Lag scrambling</td>
<td></td>
<td></td>
<td></td>
<td>0.015*** (0.004)</td>
</tr>
<tr>
<td>Lag money</td>
<td></td>
<td></td>
<td></td>
<td>−0.265*** (0.018)</td>
</tr>
<tr>
<td>Lag wins</td>
<td></td>
<td></td>
<td></td>
<td>−0.041 (0.027)</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td>9.870*** (2.103)</td>
</tr>
<tr>
<td>Constant</td>
<td>−66.17*** (9.379)</td>
<td>5.522 (10.03)</td>
<td>−73.05*** (11.72)</td>
<td>−3.529*** (1.297)</td>
</tr>
</tbody>
</table>
```

Notes: Standard errors in parentheses. Purse and lagged money in millions of 2012$.  
***p < 0.01; **p < 0.05; *p < 0.10.

- Average Driving Distance (Drive Distance)—the average length of each tee shot for a golfer in the first two rounds.
- Average Driving Accuracy (Drive Accuracy)—the percentage of occasions in which a golfer’s tee shot finished in a fairway.
- Greens in Regulation (GIR)—the percentage of occasions in which a golfer was able to place his ball on the green in a number of attempts less than or equal to two less than the par value of a hole. (Example: If a golfer takes two shots to get his ball on the green of a hole with a par value of four, then GIR = 1 for that golfer on that hole; GIR = 0 if it takes a golfer three or more shots to hit his ball on the green.)
- Average Putts per GIR (Avg Putts GIR)—the average number of putts a golfer takes on holes in which GIR = 1.
- Average Sand Saves (Sand Saves)—the percentage of occasions in which a golfer takes two or fewer shots to complete a hole from a greenside bunker.
- Average Scrambling (Scrambling)—the percentage of occasions in which a golfer takes two or fewer shots to complete a hole when GIR = 0.
seem a little strange, as one could hypothesize that increased experience, as in any labor supply or production function, should result in increased output or at the very least, after having controlled for skills and performance, have no impact on output. Instead, one might interpret these results as saying that holding skill level constant, more experienced, or possibly older, golfers perform worse on average. However, students understandably might be satisfied that they have unlocked the secrets to performing well as a professional golfer—drive the ball a long way, hit the ball on the green from the fairway, putt well once the ball is on the green—and accept that output goes down the longer one is in the business.

Of course, students would be wrong about this, as they must be directed to think about the nature of the data in equation (1) and why the selection of golfers into particular tournaments might cause the error term to be correlated with the vector of characteristics in \( X_{1t} \). In any given event, there are golfers who select not to play in the tournament, which means these golfers’ scores are zero and they have no relative ranking, an omitted variables issue that is analogous to the problem labor economists must solve when studying worker wages. After all, only workers who are employed earn a nonzero wage, just like golfers who participate are the only golfers who have a nonzero number of strokes.

At this point in our lesson, we offer two paths for instructors, as we now discuss the selection issue in the context of our data by following closely the notation and discussion given in Heckman (1979). Either path allows students the opportunity to think very carefully about why this specification error can result in bias. Instructors of upper-level econometrics courses might want to follow our mathematical discussion that illustrates selection issues, while instructors in lower-level econometrics courses or labor or sports economics courses might only follow our intuitive discussion and the results we provide, which illustrate well the intuition behind selection bias in these golf data.

To represent mathematically the problem of selection bias, students must think about when we observe values for equation (1) and when we do not. Because golfers are deciding which tournaments to enter, presumably based on whether they have a positive expected payoff from entering, we can develop a selection rule as follows—we only observe a golfer’s relative ranking in a particular tournament \( R_{it} \) if their expected payoff from entering is greater than or equal to zero. Let expected payoff (\( EP_{it} \)) be:

\[
EP_{it} = X_{2it} \beta_2 + u_{2it}
\]

where the vector \( X_{2it} \) contains measures of golfer skill and tournament specific variables such as total purse. To understand how this selection rule impacts our estimates for relative rank, we must first think about how we would discuss this relationship if there were no selection rule. For unbiased estimates, we expect that the conditional expectation of relative ranking looks like (3):

\[
E(R_{it} | X_{1it}) = X_{1it} \beta_1
\]

However, given our sample selection rule it actually is as in (4):

\[
E(R_{it} | X_{1it}, EP \geq 0) = X_{1it} \beta_1 + E(u_{1it} | EP \geq 0)
\]

Because of potential omitted variable bias, it is this second term that gives us trouble if the conditional expectation of \( u_{1it} \) is not zero.

So, when is selection a problem? Students can be shown the setting for this problem in the context of the question we are trying to answer by substituting the expected payoff
equation (2) into the conditional expectation for $u_{1ity}$, and after a slight rearrangement of terms we get equation (5):

$$E\left(u_{1ity} \mid EP \geq 0\right) = E\left(u_{1ity} \mid u_{2ity} \geq -X_{2ity} \beta_2\right)$$

(5)

which leads us to the complete conditional expectation for relative ranking in equation (6):

$$E\left(R_{ity} \mid X_{ity}, EP \geq 0\right) = X_{ity} \beta_1 + E\left(u_{ity} \mid u_{2ity} \geq -X_{2ity} \beta_2\right)$$

(6)

Given that selection into tournaments is nonrandom, students should see that the conditional expectation for $u_{1ity}$ is nonzero, which implies that the coefficients in $\beta_1$ will be biased if any correlation exists between $X_{1ity}$ and $X_{2ity}$. In context of our lesson, if any of the characteristics that affect a golfer’s performance also affect his decision to enter a given tournament, then OLS estimates using equation (1) will be biased.

The intuition behind the math we have presented is pretty straightforward and should be grasped by students at various levels, even when the math is not at the forefront of a class presentation. Students should understand that not only golfers’ various skills are correlated with their relative performances, but also their decisions to enter tournaments. Not controlling for the correlation between these variables confounds the results of our regression of skill measures on performance.

As an example, we could link golf to the consulting profession. In both vocations, there are rare superstars like Tiger Woods who, because of their superior skills, have a reasonable expectation that they will always perform relatively well and earn high wages for doing so. It probably would not be surprising to find that these superstars do not play every event or accept every consulting matter that avails itself to them. On the other hand, less-talented golfers and consultants have lower expectations about their ability to finish relatively high in a tournament or be chosen to represent a client. These golfers’ lower expected wages in all tournaments make it more likely that they would be willing to enter more tournaments, even those with relatively small purses, just like a mediocre consultant will likely find it not profitable to turn down work that darkens his or her door.

Once students have been introduced to the source of the bias from the omitted variables, they can start to use the Heckman two-step procedure in order to mitigate these selection issues. The first step in the procedure involves identifying the equation that determines selection into tournaments. To do this, students must not only define the variables that affect whether a golfer enters a particular tournament but also perform the sometimes difficult task of finding a variable that satisfies the exclusion restriction. Because selecting to enter a tournament is a binomial decision, students should recognize that they could use a probit model with golfer characteristics as independent variables to measure the conditional likelihood that a golfer enters a tournament. Exempt golfers who can enter tournaments freely will be apt to enter tournaments that offer them the largest expected returns on their labor, which they will do either by entering tournaments with the biggest purses, which should be obvious to students, or choosing tournaments that are contested on courses that match well with their playing style, a reason that might be less obvious to students. (The idea that players match themselves to certain courses is referred to by some as “courses for horses”). Some course designs reward golfers who can drive the ball with accuracy inordinately well, while other courses reward players who can hit the ball a long way. (Jim Furyk is an example of a golfer who is very accurate; Bubba Watson is a golfer who is very long.) To address these two concerns, we collected data on the six skills mentioned above for the season
The variable purse is crucial to our procedure because it is used in the selection equation and because the variable satisfies well the exclusion restriction, a necessity that must be met when using the Heckman procedure. A variable that satisfies the exclusion restriction must be economically and statistically significant in the selection equation (2) while also not being of use in the least-squares equation (1). Purse does this because golfers will find strong incentives to enter tournaments with relatively big purses, as this will likely increase their expected earnings, holding finish constant. However, including it in the structural equation for relative rank makes little sense for most golfers, as the size of the purse should not change everyone’s rankings in one direction or the other on average; as one person improves, another must fall in the rankings. This fact is something about which students should be asked to think, as finding an exclusion restriction can often be a difficult thought exercise in questions of selection bias. In this setting, the variable purse provides a fairly straightforward example of a proper exclusion restriction that we believe illuminates the concept very well.

With regards to our regression results, the use of purse as an exclusion restriction should really drive the concept home for students. In addition to biased coefficients, models estimated on selected data also can show significance for irrelevant variables that play a role in the selection process, but not in the equation of interest. To illustrate the danger of including a nonrelevant variable in the original OLS equation (1) in selected samples, we ask students to add purse to their original OLS regression. The results of this exercise are shown in column 2 of table 2. The coefficient on purse is in fact highly significant, but its significance belies its lack of economic significance, as not all golfers can move down in rankings when purse increases. An intuitive explanation of this significant coefficient for purse can be found in our previous example of golfers like Tiger Woods versus less-talented golfers. The significance of this coefficient is in fact driven by the selection equation (2) and the correlation between purse and other factors that both influence a golfer’s decision to enter a tournament, as well as their overall performance in the tournament. This provides a clear example of how one could improperly interpret regression results from a selected sample if one ignores the selection problem. This also underscores the econometrician’s mantra: The best models are defined on good theory, not on significant coefficients.

Now that students have become aware of the exclusion restriction, they can be asked to estimate the first step of the Heckman procedure. We do not observe expected payoff measures that golfers are using to select into tournaments, but rather a censored version where, if expected payoff is greater than or equal to zero, we observe golfers entering a tournament. Therefore, we can model this entry decision using equation (7):

\[
P(\text{Enter}_{ity}) = \beta X_{2ity} + u_{2ity}
\]

Let Enter_{ity} = 1 if golfer \(i\) enters tournament \(t\) in year \(y\) and 0, otherwise. The vector \(X_{2ity}\) contains golfers’ tenure as professional golfers, purse of the event, and averages of the six golfer statistics used in equation (1), calculated for the prior season. Assuming that the error term is distributed normally, the students can use a probit to estimate equation (4), the results of which can be used to calculate the inverse Mills ratio. Depending on the level or need of the class, at this point students can be asked to calculate the inverse Mills ratio themselves and enter it into the second-stage estimation of equation (2), or a software package such as Stata can do this for them directly.
After estimating equation (7), students should find their results matching those in column 4 of table 2. Note the large and statistically significant coefficient on purse, indicating the strength of the exclusion restriction. Not surprisingly, golfers’ respond strongly to the potential for large winnings.

Once students have collected information on the selectivity regressor from the probit equation, they can proceed to the second step of the Heckman procedure by adding the inverse Mills ratio to equation (1). Although it may be obvious to the experienced economist what is happening when this substitution occurs, it is very important for students to understand that the inverse Mills ratio is representing the variable or variables omitted in the original estimate of equation (1) by controlling for the probability that a given observation would be observed. The results of this estimation can be found in column 3 of table 2. Students should be directed to take note of three points of interest. First, note that in column 3 the number of observations regressed is larger than the number regressed in column 1 because all observations, those with seen and unseen relative ranking, are used to calculate the inverse Mills ratio. Second, the coefficient estimate on the inverse Mills ratio is large and significant, suggesting strongly that selection bias existed in the original estimation of equation (1). Finally, the coefficient on experience is smaller in magnitude and no longer statistically significant. These results mesh more easily with economic theory than the results in column 1, suggesting now that once measures of skill and performance are controlled for, experience does not have a significant impact on relative ranking. Clearly, correcting for selection bias matters, as this correction changes our conclusion about the effect of worker experience on output in this setting.

To end this lesson, students can be directed to discuss the differences in results from column 1 to column 3, and how they reflect the selection problem. They can also be asked to discuss how any conclusions they would reach would have been wrong if they had not corrected for selection, and how this correction changes the overall interpretation of the results. Finally, students can use this knowledge to think about how the results of studies in other areas, such as labor supply studies, may be affected by selection bias problems.

CONCLUSION

To conclude, we have presented a project that an econometrics class at the undergraduate or graduate level could be asked to perform. Alternatively, the main points of the lesson can be used to illustrate the importance of institutional knowledge and selection bias in labor economics or sports economics courses, as selection issues often must be addressed in studies of a number of markets. The activity asks students to analyze data from PGA Tour golf tournaments. Along the way, students should be asked to ponder the role that PGA Tour procedures have in creating selection issues in the data and to use the Heckman two-step procedure to mitigate the selection bias in their analysis of golfers’ relative performances. At the end of the lesson, we are confident that students will have a healthier respect for the returns they can earn from learning about the effect that institutional operating procedures might have on the data they collect, while also improving their econometric modeling skills via the use of tools like the Heckman two-step procedure.
1. Students can be directed to important labor economics studies such as Heckman’s 1993 survey of the labor supply literature to further illustrate the importance of this issue.

2. The ShotLink data can be accessed freely by students and teachers at http://www.pgatour.com/stats/shotlinkintelligence/. We have placed the data at https://sites.google.com/site/mcfallecon/home/sports-datasets.

3. The Tour’s Web site, www.pgatour.com, provides more institutional details. Note the nonlinear structure of prize money disbursements, which exists to induce golfers to take risks, which can be exciting, in order to leap competitors on the last day of competition.


5. Exemption rules can be found at http://www.pgatour.com/news/2012/12/golfer_exemptions.html. We control for four of the criteria because only these golfers would have been guaranteed a spot in any of the tournaments in the dataset. Those lower on the exemption priority rankings had to wait to see if there was a spot open for them because golfers with higher priority chose not to play in a particular event.

6. We use only the first two rounds because tournaments cut fields at this point in the tournament. Extending our analysis to four rounds would have required another layer of selection bias.

7. Tournament entry decisions can lead to an interesting discussion, because under some conditions, the expected earnings of a golfer might actually increase as purse decreases. This fact rests upon the elasticity of tournament field strength with respect to purse. On the other end, sometimes lesser-talented golfers will want to fake their talent and hope to get lucky in tournaments with big purses. See Lazear and Rosen (1981) for discussion of this incentive.

8. Heckman (1979) never discussed specifically the importance of finding a variable that satisfies the exclusion restriction. Puhani (2000) presented a fine discussion on this topic.

9. Stata has a command called “heckman” which will do this nicely, and if students employ the “twostep” option, they can observe both steps. This procedure is also available in EViews and SAS with slightly more manipulation.

REFERENCES


