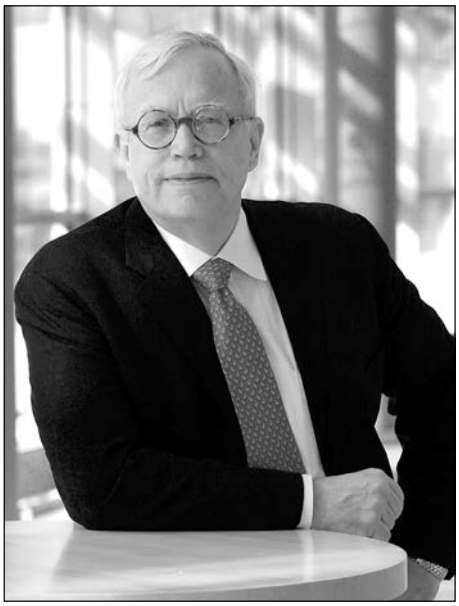
Джеймс Джозеф Хекман



Основное отличие эконометрики от статистических дисциплин, применяемых в других науках, определяется тем, что во многих случаях приходится обрабатывать данные неэкспериментальной природы. При этом большинство задач в эконометрике зачастую состоят в оценивании причинно-следственных связей, в самом определении которых обычно подразумевается некий мысленный эксперимент. Например, вопрос о том, как образование влияет на заработную плату, подразумевает мысленный эксперимент, в котором у случайно выбранного человека меняют образование и измеряют эффект этого изменения на его заработную плату. Проведение такого эксперимента на практике невозможно, в том числе и по этическим причинам, а экономистам приходится довольствоваться выборками из наблюдаемых данных. Джеймс Хекман первый указал на фундаментальную проблему работы с такими данными: наблюдаемые данные часто не являются случайной выборкой.

Выборки наблюдаемых данных являются выборками с искаженным отбором или так называемыми «селективными выборками», в том смысле, что индивиды, оказавшиеся в выборке, зачастую обладают неким общим свойством. Так, например, в случае оценки влияния образования на заработную плату мы часто наблюдаем только индивидов, которые работают. Если не учитывать факт неслучайности выборки, то статистические оценки будут смещенными.

В знаменитой статье Джеймса Хекмана, представленной ниже, разработаны методы статистического оценивания в условиях селективных выборок. Эта статья изменила основы эконометрики и стала классикой эконометрической теории. В 2000 году Нобелевский комитет удостоил Джеймса Хекмана Нобелевской премии по экономике «за разработку теории и методов анализа селективных выборок».

Джеймс Хекман является известным эконометристом, плодотворно работающим как в теории, так и в прикладных задачах. Его исследования посвящены вопросам эффективности государственных программ в области образования, трудоустройства, раннего детского развития. Отвечая на важные экономические вопросы, Хекман разработал множество эконометрических методов работы с селективными выборками, учета гетерогенности и решения других проблем, возникающих при отсутствии экспериментальных данных.

Джеймс Хекман работает профессором экономики в университете Чикаго. Он получил степень PhD в Принстоне в 1971 году.

А. Е. Микушева

Sample selection bias as a specification error

James J. Heckman

Смещение селективной выборки как ошибка спецификации[1,2]

Джеймс Дж. Хекман[3]

В данной работе проблема смещения оценок регрессии, возникающего из-за использования неслучайных выборок, изучается как ошибка спецификации или как смещение, обусловленное «пропущенными переменными». Предлагается простая двухшаговая МНК оценка, которая является состоятельной и позволяет использовать стандартные регрессионные методы. Также выводится асимптотическое распределение оценок коэффициентов.

Ключевые слова: ошибка спецификации; селективность выборки; смещенная выборка.

JEL classification: C01; C13; C24.

(Примечание. JEL classification и ключевые слова добавлены переводчиком).

В данной работе проблема смещения оценок регрессии, возникающего из-за использования селективных выборок, изучается как ошибка спецификации или как смещение, обусловленное «пропущенными переменными». В отличие от обычных смещений оценок, вызванных пропущенными переменными или ошибками спецификации эконометрической модели, в случае использования селективных выборок иногда удается оценить

---

[1] Оригинальная статья: Heckman J. J. Sample Selection Bias as a Specification Error. *Econometrica*, 1979, 47 (1), 153–161. © Econometric Society.

Редакция благодарит Econometric Society за разрешение на публикацию перевода статьи.

Перевод статьи выполнен студентами НИУ ВШЭ И. Станкевичем и Д. Малаховым, под редакцией профессора П. К. Катышева.

[2] Выражение «Смещение селективной выборки» есть один из вариантов перевода оригинального английского термина «sample selection bias». Этот термин применяется в ситуации, когда для оценивания коэффициентов в уравнении регрессии используется неслучайная выборка. В этом случае соответствующие оценки могут оказаться смещенными. — *Прим. переводчика*.

[3] Это исследование выполнено при поддержке гранта HEW Rand Corporation и гранта Департамента труда Национального бюро экономических исследований США. Первоначальная версия этой статьи имела название «Теневые цены, рынок зарплаты и предложение труда: некоторые вычислительные упрощения и обновленные оценки», июнь 1975 г. Неожиданно для меня большое число коллег сделали ценные комментарии к этой статье и ее многочисленным вариантам. Выражаю особую благодарность Takeshi Amemiya, Zvi Griliches, Reuben Gronau, Mark Killingsworth, Ed Leamer, Tom MaCurdy, Bill Rodgers и Paul Schultz. Я несу полную ответственность за любые оставшиеся ошибки.

влияние переменных, которые при исключении из регрессионной модели могут привести к неправильной ее спецификации. Можно использовать оцененные (прогнозные) значения пропущенных переменных в качестве регрессоров в основном уравнении и применять стандартные методы оценивания. Таким образом, в настоящей работе проблема смещения селективной выборки рассматривается как ошибка спецификации модели. Для случая цензурированных выборок предлагается простой состоятельный метод оценивания, устраняющий эту ошибку. Данное исследование проясняет и расширяет анализ, проведенный в (Heckman, 1976): в явном виде получено асимптотическое распределение простой оценки для общего случая, а не частного случая нулевой гипотезы об отсутствии смещения селективной выборки, рассмотренного в (Heckman, 1976). Для улучшения понимания приводится (с исправлениями и упрощениями) вводный материал из статьи (Heckman, 1976).

На практике смещение селективной выборки может возникать по двум причинам. Во-первых, это происходит при самоотборе индивидуумов или наблюдений в выборку, т. е. единицы наблюдения могут попадать в выборку не случайно. Во-вторых, решения исследователей в части построения выборки могут приводить к схожим последствиям.

Можно привести много примеров селективных выборок. Рыночные заработные платы наблюдаются только для тех работающих женщин, чья рыночная зарплата превышает «зарплату» от работы дома. Аналогично, можно наблюдать доходы только тех участников профсоюза, для которых состоять в нем оказалось выгоднее. Также уровни зарплат мигрантов обычно не позволяют получить надежные оценки зарплат, которые получили бы немигранты в случае миграции. Доход индивидов, прошедших обучение, не дает возможность получить оценки заработков необученных индивидов в случае повышения ими квалификации. Во всех вышеперечисленных случаях оценки уравнения зарплаты, полученные по селективной выборке, не дают возможность получить представление об истинных закономерностях в формировании заработных плат. Сравнение зарплат мигрантов и немигрантов (аналогично, сравнение дохода обученных работников и необученных), приводит к смещению в оценках реального влияния миграции, обучения, участия в профсоюзах и т. д.

Выборка может стать неслучайной по причине вмешательства исследователя в структуру данных. В панельных моделях обычно работают лишь со стабильными наблюдениями. Например, постоянство состава домохозяйства часто является необходимым условием добавления его в выборку. В демографических исследованиях и в экспериментах по установлению эффективности обучения обычно рассматриваются лишь те единицы наблюдения, данные по которым есть на протяжении всего периода наблюдения. Эти особенности проведения анализа приводят к результатам, схожим с проблемой самоотбора: при оценивании структурных уравнений «смешиваются» основные факторы, представляющие главный интерес, и факторы, определяющие вероятность попадания в выборку.

1. Характеризация смещения селективной выборки

Для упрощения изложения рассмотрим модель, состоящую из двух регрессионных уравнений. Переход к большему числу уравнений не представляет трудностей.

Пусть есть случайная выборка, состоящая из I наблюдений. Уравнения для i-го объекта выглядят следующим образом:

$$Y_{1i} = X_{1i}\beta_1 + U_{1i}, \quad (1a)$$

$$Y_{2i} = X_{2i}\beta_1 + U_{2i}, \ (i = 1,...,I), \quad (1b)$$

где X_{ji} — вектор экзогенных регрессоров размерности $1 \times K_j$, β_j — вектор коэффициентов размерности $K_j \times 1$ и $E(U_{ji}) = 0, \quad E(U_{ji}, U_{j'i''}) = \begin{cases} \sigma_{jj'}, & i = i'', \\ 0, & i \neq i''. \end{cases}$

Последнее предположение — это следствие того, что наша выборка случайна. Плотность совместного распределения величин U_{1i}, U_{2i} есть $h(U_{1i}, U_{2i})$. Предполагается также, что матрица регрессоров имеет полный ранг, поэтому все параметры модели можно оценить с помощью метода наименьших квадратов (МНК).

Предположим, что мы пытаемся оценить регрессионное уравнение (1a), но обнаруживается, что есть пропущенные значения переменной Y_{1i}. Поэтому закономерно возникает вопрос: «Почему есть пропуски в значениях Y_{1i}?»

Функция регрессии (1a) по всей генеральной совокупности может быть записана так:

$$E(Y_{1i} \mid X_{1i}) = X_{1i}\beta_1, \qquad (i = 1,...,I).$$

Регрессионное уравнение для имеющейся подвыборки выглядит следующим образом:

$$E(Y_{1i} \mid X_{1i}, \textit{правило отбора выборки}) = X_{1i}\beta_1 + E(U_{1i} \mid \textit{правило отбора выборки}), (i = 1, \ldots, I).$$

Для удобства предположим, что для первых $I_1 < I$ наблюдений доступны данные по Y_{1i}.

Если условное математическое ожидание U_{1i} равно нулю, то регрессия по данной подвыборке совпадает с регрессией по всей генеральной совокупности. В этом случае для получения оценки коэффициентов β_1 можно применить метод наименьших квадратов. Таким образом, оценивание регрессии по неполной выборке ведет лишь к потере эффективности.

В общем случае принципы формирования выборки ведут к более серьезным последствиям. Предположим, например, что данные по переменной Y_{1i} есть только в случае, если $Y_{2i} \geq 0$, а для $Y_{2i} < 0$ нет наблюдений. Выбор значения 0 в качестве порога отсечения несущественен — это вопрос нормализации.

В общем случае имеем:

$$E(U_{1i} \mid X_{1i}, \textit{правило отбора выборки}) = E(U_{1i} \mid X_{1i}, Y_{2i} \geq 0) = E(U_{1i} \mid X_{1i}, U_{2i} \geq -X_{1i}\beta_2). \quad (2)$$

В случае независимости U_{1i} и U_{2i}, когда пропуски в Y_{1i} случайны, условное математическое ожидание величины U_{1i} равно нулю. В более общей ситуации условное математическое ожидание U_{1i} не равно нулю и выборочная регрессия выглядит следующим образом:

$$E(Y_{1i} \mid X_{1i}, Y_{2i} \geq 0) = X_{1i}\beta_1 + E(U_{1i} \mid X_{1i}, U_{2i} \geq -X_{2i}\beta_2). \quad (3)$$

Выборочная регрессия зависит от значений X_{1i} и X_{2i}. Оценка регрессии по модели (1a) не учитывает последнее слагаемое уравнения (3), таким образом, смещение, порождаемое неслучайностью выборки, есть следствие обычной проблемы пропущенных переменных.

Сделаем несколько замечаний. Во-первых, если вектор X_{2i}, который определяет выбор конкретной подвыборки, состоит лишь из «1» (т. е. есть свободного члена, *прим. перевод*.) (поэтому вероятность включения в выборку для всех наблюдений одинакова), то условное математическое ожидание U_{1i} — константа, и смещение значений β_1 заключается лишь в сме-

щении значения константы. Также легко показать, что МНК-оценка дисперсии σ_{11} смещена вниз. Во-вторых, хорошим индикатором селективного смещения служит следующий факт: незначимые в популяционной модели регрессоры, включенные в X_{2i}, но не включенные в X_{1i}, становятся значимыми в модели, оцененной по имеющейся выборке. В-третьих, данная модель — обобщение некоторых уже существующих моделей. Например, если $h(U_{1i}, U_{2i})$ — сингулярное одномерное нормальное распределение $(U_{1i} \equiv U_{2i})$ и $X_{2i} = X_{1i}$, $\beta_2 \equiv \beta_1$, рассматриваемая модель превращается в Тобит-модель. Более подробно взаимосвязь с уже созданными моделями рассмотрена в (Heckman, 1976). В-четвертых, модели с большим числом уравнений, будучи простым математическим обобщением приведенной выше модели, содержательно могут представлять значительный интерес. Рассмотрим такой пример. Предположим, что мигранты выбирают один из K регионов, куда они могут мигрировать. Если наиболее предпочтителен регион с наивысшим потенциальным доходом, модель, описывающая поведение мигрантов, является расширением двухуровневой модели.

2. Оценка для случая нормальных ошибок и ее свойства[4]

Пусть $h(U_{1i}, U_{2i})$ — двумерная нормальная функция плотности. Используя известные результаты (см. (Johnston, Kotz, 1972, p. 112–113)), имеем:

$$E(U_{1i} \mid U_{2i} \geq -X_{2i}\beta_2) = \frac{\sigma_{12}}{(\sigma_{22})^{1/2}}\lambda_i,$$

$$E(U_{2i} \mid U_{2i} \geq -X_{2i}\beta_2) = \frac{\sigma_{22}}{(\sigma_{22})^{1/2}}\lambda_i,$$

где

$$\lambda_i = \frac{\varphi(Z_i)}{1-\Phi(Z_i)} = \frac{\varphi(Z_i)}{\Phi(-Z_i)}.$$

Здесь φ и Φ — функция плотности и функция распределения стандартной нормальной величины соответственно, и

$$Z_i = -\frac{X_{2i}\beta_2}{(\sigma_{22})^{1/2}}.$$

Величина λ_i — обратное отношение Миллса. Это монотонно убывающая функция вероятности попадания наблюдения в выборку, $\Phi(-Z_i)$ $(=1-\Phi(Z_i))$. В частности, $\lim\limits_{\Phi(-Z_i)\to 1}\lambda_i = 0$, $\lim\limits_{\Phi(-Z_i)\to 0}\lambda_i = \infty$, и $\dfrac{\partial\lambda_i}{\partial\Phi(-Z_i)} < 0.$

Теперь можно привести полную статистическую модель для случая нормальных ошибок. Условная функция регрессии для имеющихся наблюдений может быть записана как:

[4] Версия метода оценивания для сгруппированных данных, рассматриваемая здесь, была предложена в (Gronau, 1974) и (Lewis, 1974). Однако они не исследуют статистические свойства метода и не рассматривают микроверсию оценки, представленную здесь.

$$E(Y_{1i} \mid X_{1i}, Y_{2i} \geq 0) = X_{1i}\beta_1 + \frac{\sigma_{12}}{(\sigma_{22})^{1/2}}\lambda_i,$$

$$E(Y_{2i} \mid X_{2i}, Y_{2i} \geq 0) = X_{2i}\beta_2 + \frac{\sigma_{22}}{(\sigma_{22})^{1/2}}\lambda_i,$$

$$Y_{1i} = E(Y_{1i} \mid X_{1i}, Y_{2i} \geq 0) + V_{1i}, \quad (4a)$$

$$Y_{2i} = E(Y_{2i} \mid X_{2i}, Y_{2i} \geq 0) + V_{2i}, \quad (4b)$$

где

$$E(V_{1i} \mid X_{1i}, \lambda_i, U_{2i} \geq -X_{2i}\beta_2) = 0, \quad (4c)$$

$$E(V_{2i} \mid X_{2i}, \lambda_i, U_{2i} \geq -X_{2i}\beta_2) = 0, \quad (4d)$$

$$E(V_{ji}V_{j''i'} \mid X_{1i}, X_{2i}, \lambda_i, U_{2i} \geq -X_{2i}\beta_2) = 0, \quad \text{для } i \neq i'. \quad (4e)$$

Далее,

$$E(V_{1i}^2 \mid X_{1i}, \lambda_i, U_{2i} \geq -X_{2i}\beta_2) = \sigma_{11}\left((1-\rho^2) + \rho^2(1 + Z_i\lambda_i - \lambda_i^2)\right), \quad (4f)$$

$$E(V_{1i}V_{2i} \mid X_{1i}, X_{2i}\lambda_i, U_{2i} \geq -X_{2i}\beta_2) = \sigma_{12}(1 + Z_i\lambda_i - \lambda_i^2), \quad (4g)$$

$$E(V_{2i}^2 \mid X_{2i}, \lambda_i, U_{2i} \geq -X_{2i}\beta_2) = \sigma_{22}(1 + Z_i\lambda_i - \lambda_i^2), \quad (4h)$$

где

$$\rho^2 = \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}$$

и

$$0 \leq 1 + Z_i\lambda_i - \lambda_i^2 \leq 1. \quad (5)$$

Зная Z_i и, следовательно, λ_i, можно добавить λ_i как регрессор в уравнение (4a) и оценить его методом наименьших квадратов. МНК оценки параметров β_1 и $\sigma_{11}/(\sigma_{22})^{1/2}$ являются несмещенными, но не эффективными. Неэффективность — следствие гетероскедастичности, вытекающей из уравнения (4f), когда X_{2i} (а следовательно, и Z_i) содержит нетривиальные регрессоры (т. е. регрессоры, отличные от константы, *прим. перевод*.) Вследствие неравенства (5) стандартная МНК оценка дисперсии σ_{11} смещена вниз. Из уравнения (4g) и неравенства (5) следует, что обычная оценка ковариации между уравнениями смещена вниз. Для получения правильных стандартных отклонений оценок коэффициентов первого уравнения может быть использована обычная ОМНК процедура (подробнее см. Heckman (1976)).

На практике величина λ_i неизвестна. Но в случае цензурированной выборки, когда нет информации о величинах Y_{1i} при $Y_{2i} \leq 0$, но известны X_{2i} для наблюдений с $Y_{2i} \leq 0$, можно оценить λ_i при помощи следующей процедуры.

(1) Оценить параметры вероятности того, что $Y_{2i} \geq 0$ (т. е. $\beta_2/(\sigma_{22})^{1/2}$) при помощи пробит-модели по всей выборке[5].

(2) Из оценки параметра $\beta_2/(\sigma_{22})^{1/2}$ $(=\beta_2^*)$ получить оценку величины Z_i и, следовательно, величины λ_i. Все эти оценки являются состоятельными.

[5] В случаях, где Y_{2i} наблюдается, можно оценить β_2, σ_{22}, и, следовательно, $\beta_2/(\sigma_{22})^{1/2}$ при помощи МНК.

(3) Оцененная переменная λ_i может быть использована как регрессор в уравнении (4a), оцениваемом на подвыборке. Оценки параметров β_1 и $\sigma_{12}/(\sigma_{22})^{1/2}$ (коэффициентов при X_{1i} и λ_i соответственно) являются состоятельными[6].

(4) Состоятельную оценку параметра σ_{11} можно получить следующим образом. На шаге 3 можно получить состоятельную оценку параметра $C = \rho(\sigma_{11})^{1/2} = \sigma_{12}/(\sigma_{22})^{1/2}$. Обозначим через $\hat{V}_{1i}$ остаток для i-го наблюдения на шаге 3, а оценку параметра C — через $\hat{C}$. Тогда оценку величины σ_{11} можно получить так:

$$\hat{\sigma}_{11} = \frac{\sum_{i=1}^{I_1} \hat{V}_{1i}^2}{I_1} - \frac{\hat{C}}{I_1} \sum_{i=1}^{I_1} (\hat{\lambda}_i \hat{Z}_i - \hat{\lambda}_i^2),$$

где $\hat{\lambda}_i$ и $\hat{Z}_i$ — оценки параметров Z_i и λ_i, полученные на шаге 2. Эта оценка состоятельна и положительна, т. к. элементы второй суммы отрицательны (см. неравенство (5)).

Обычные формулы для стандартных ошибок коэффициентов, оцененных МНК, *не* работают, кроме одного важного случая — нулевой гипотезы об отсутствии селективного смещения ($C = \sigma_{12}/(\sigma_{22})^{1/2} = 0$). В этой ситуации можно использовать обычные стандартные ошибки и проводить тестирование гипотезы $C = 0$, используя t-распределения. Если же $C \neq 0$, стандартная процедура занижает оценки стандартных ошибок и завышает значимость переменных.

Вывод корректного асимптотического распределения для этой оценки в общем случае требует определенных усилий[7]. Заметим, что уравнение (4a) с оцененной величиной λ_i, использованной вместо истинной λ_i, может быть записано так:

$$Y_{1i} = X_{1i}\beta_1 + C\hat{\lambda}_i + C(\lambda_i - \hat{\lambda}_i) + V_{1i}. \qquad (4a')$$

Остаток состоит из последних двух членов уравнения.

Величина λ_i рассчитывается при помощи параметра $\beta_2/(\sigma_{22})^{1/2}$ $(= \beta_2^*)$, который, в свою очередь, оценивается с помощью пробит-модели на полной выборке из I наблюдений методом максимального правдоподобия[8]. Поэтому в силу того, что λ_i — дважды непрерывно дифференцируемая функция от β_2^*, величина $\sqrt{I}(\hat{\lambda}_i - \lambda_i)$ имеет асимптотически нормальное распределение

$$\sqrt{I}(\hat{\lambda}_i - \lambda_i) \sim N(0, \Sigma_i),$$

где Σ_i — асимптотическая ковариационная матрица, полученная из ковариационной матрицы β_2^* следующим образом:

---

[6] Предполагается, что вектор X_2 содержит нетривиальные регрессоры, или что β_1 не содержит константу, или и то и другое одновременно.

[7] Эта часть работы была вдохновлена комментариями T. Amemiya. Разумеется, он не несет ответственности за любые ошибки в рассуждениях.

[8] Дальнейший анализ может быть очевидным образом модифицирован, если Y_{2i} наблюдается и β_2^* оценивается методом наименьших квадратов.

$$\Sigma_i = \left(\frac{\partial \lambda_i}{\partial Z_i}\right)^2 X_{2i} \Sigma X_{2i}',$$

где $\frac{\partial \lambda_i}{\partial Z_i}$ — производная λ_i по Z_i, и Σ — асимптотическая ковариационная матрица величины $\sqrt{I}(\hat{\beta}_2^* - \beta_2^*)$.

Мы ищем асимптотическое распределение вектора

$$\sqrt{I_1}\begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{C} - C \end{bmatrix} = I_1 \begin{bmatrix} \Sigma X_{1i}' X_{1i} & \Sigma X_{1i}' \hat{\lambda}_i \\ \Sigma X_{1i} \hat{\lambda}_i & \Sigma \hat{\lambda}_i^2 \end{bmatrix}^{-1} \frac{1}{\sqrt{I_1}} \begin{bmatrix} \Sigma X_{1i}' (C(\lambda_i - \hat{\lambda}_i) + V_{1i}) \\ \Sigma \hat{\lambda}_i (C(\lambda_i - \hat{\lambda}_i) + V_{1i}) \end{bmatrix}.$$

В дальнейшем важно помнить, что пробит-модель оценена на полной выборке из I наблюдений, тогда как основная модель оценивается на подвыборке из I_1 $(< I)$ наблюдений, в которых наблюдается Y_{1i}. Далее необходимо отметить, что, в отличие от двухшагового метода наименьших квадратов, часть остатка, которая возникает из-за использования оцененной величины λ_i вместо истинной λ_i, *не* ортогональна вектору X_1.

При выполнении общих требований к регрессорам, рассматриваемых в (Amemiya, 1973) и (Jennrich, 1969), получаем

$$\underset{I_1 \to \infty}{p\lim} I_1 \begin{bmatrix} \Sigma X_{1i}' X_{1i} & \Sigma X_{1i}' \hat{\lambda}_i \\ \Sigma X_{1i} \hat{\lambda}_i & \Sigma \hat{\lambda}_i^2 \end{bmatrix}^{-1} = \underset{I_1 \to \infty}{p\lim} I_1 \begin{bmatrix} \Sigma X_{1i}' X_{1i} & \Sigma X_{1i}' \lambda_i \\ \Sigma X_{1i} \lambda_i & \Sigma \lambda_i^2 \end{bmatrix}^{-1} = B,$$

где B — конечная положительно определённая матрица[9]. При выполнении этих требований

$$\sqrt{I_1}\begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{C} - C \end{bmatrix} \sim N(0, B\psi B'),$$

где $$\psi = \underset{\substack{I_1 \to \infty \\ I \to \infty}}{p\lim} \left\{ \sigma_{11} \begin{bmatrix} \frac{\Sigma X_{1i}' X_{1i} \eta_i}{I_1} & \frac{\Sigma X_{1i}' \lambda_i \eta_i}{I_1} \\ \frac{\Sigma \lambda_i X_{1i} \eta_i}{I_1} & \frac{\Sigma \lambda_i^2 \eta_i}{I_1} \end{bmatrix} + C^2 \left(\frac{I_1}{I}\right) \begin{bmatrix} \sum_{i=1}^{I_1}\sum_{i'=1}^{I_1} \frac{X_{1i}' X_{1i'} \theta_{ii'}}{I_1^2} & \sum_{i=1}^{I_1}\sum_{i'=1}^{I_1} \frac{X_{1i}' \pi_{ii'}}{I_1^2} \\ \sum_{i=1}^{I_1}\sum_{i'=1}^{I_1} \frac{X_{1i} \pi_{ii'}}{I_1^2} & \sum_{i=1}^{I_1}\sum_{i'=1}^{I_1} \frac{\Omega_{ii'}}{I_1^2} \end{bmatrix} \right\},$$

$$\underset{\substack{I_1 \to \infty \\ I \to \infty}}{p\lim} \frac{I_1}{I} = k, \quad 0 < k < 1,$$

где

$$C = \sigma_{12} / (\sigma_{22})^{1/2},$$

$$\eta_i = (1 + C^2 (Z_i \lambda_i - \lambda_i^2) / \sigma_{11}),$$

$$\pi_{ii'} = \left(\frac{\partial \lambda_i}{\partial Z_i}\right)\left(\frac{\partial \lambda_{i'}}{\partial Z_{i'}}\right) \lambda_i X_{2i} \Sigma X_{2i'}',$$

[9] Для этого необходимо, чтобы X_2 содержал нетривиальные регрессоры, или чтобы в уравнении не было константы, или выполнения обоих условий одновременно.

$$\theta_{ii'} = \left(\frac{\partial \lambda_i}{\partial Z_i}\right)\left(\frac{\partial \lambda_{i'}}{\partial Z_{i'}}\right) X_{2i} \Sigma X_{2i'}',$$

$$\Omega_{ii'} = (\lambda_i \lambda_{i'})\left(\frac{\partial \lambda_i}{\partial Z_i}\right)\left(\frac{\partial \lambda_{i'}}{\partial Z_{i'}}\right) X_{2i} \Sigma X_{2i'}',$$

где $\frac{\partial \lambda_i}{\partial Z_i}$ — производная λ_i по Z_i,

$$\frac{\partial \lambda_i}{\partial Z_i} = \lambda_i^2 - Z_i \lambda_i .$$

Можно заметить, что если $C = 0$, матрица $B\psi B'$ сводится к стандартной ковариационной матрице для оценки метода наименьших квадратов. Отметим также, что из-за того, что вторая матрица в ψ положительно определена, при $C \neq 0$ корректная асимптотическая ковариационная матрица ($B\psi B'$) дает более высокие оценки стандартных ошибок коэффициентов регрессии, чем некорректная «стандартная» ковариационная матрица $\sigma_{11}B$. Таким образом, стандартная процедура оценивания, корректно работающая при известных λ_i, приводит к недооценке истинных стандартных отклонений и переоценке значимости коэффициентов, когда используется оцененное значение λ_i и $C \neq 0$.

При выполнении упомянутых выше условий Amemiya–Jennrich, ψ является ограниченной положительно определенной матрицей. Нетрудно оценить параметры матриц ψ и B. Оцененные λ_i, C и σ_{11} могут быть использованы вместо истинных величин для получения состоятельной оценки $B\psi B'$. Оценка ковариационной матрицы требует обращения матрицы размера $(K_1 + 1) \times (K_1 + 1)$ и проста с вычислительной точки зрения. Копия программы для оценки коэффициентов β_2^* пробит-модели и коэффициентов $\hat{\beta}_1$ и $\hat{C}$ регрессии, вычисляющая корректные асимптотические оценки стандартных ошибок в общем случае, доступна по запросу у автора[10].

Для оценки таких моделей возможно разработать ОМНК процедуру (Heckman, 1977), но она будет требовать более сложных вычислений, и при этом ОМНК-оценки не являются асимптотически эффективными, поэтому использовать ее не рекомендуется.

Обсуждаемый в настоящей работе метод оценивания уже был применен на практике. Появляется все больше свидетельств (Griliches et al., 1977; Heckman, 1976) того, что получаемые с его помощью оценки могут быть использованы в качестве хороших начальных значений для получения оценок методом максимального правдоподобия, т. к. они оказываются близки к ММП-оценкам. Учитывая простоту и гибкость процедуры, можно рекомендовать ее к использованию в эмпирических работах.

3. Заключение

В данной работе смещение, возникающее из-за использования неслучайных выборок при оценке моделей, рассматривается как ошибка спецификации модели в духе работ (Griliches, 1957) и (Theil, 1957). Предлагается простая с вычислительной точки зрения техника, позволяющая использовать обычные регрессионные методы для оценивания моделей и не при-

[10] Это предложение действительно в течение двух лет после публикации статьи. После этого программа будет предоставляться за плату (статья опубликована в 1979 г. — *Прим. редакции*).

водящая к смещению в случае использования цензурированной выборки. Рассматриваются асимптотические свойства оценок.

Другая простая оценка, применимая для усеченных выборок, рассматривается в (Amemiya, 1973). Сравнение результатов (Amemiya, 1973) с оценкой, предлагаемой в данной работе, было бы очень полезно, но не рассматривается в рамках данной работы. Обобщение анализа, проведенного в моей работе 1976 года, на многомерный случай дано в работе (Hanoch, 1976). Предлагаемая здесь простая процедура может быть использована для исследования моделей с усеченными выборками, селективными выборками и с ограниченными зависимыми переменными, равно как и для систем одновременных уравнений с эндогенными дамми-переменными (Heckman, 1976, 1978).

University of Chicago

Manuscript received March, 1977; final revision received July, 1978.

Список литературы

Amemiya T. (1973), Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41 (6), 997–1016.

Griliches Z. (1977). Specification bias in estimates of production functions. *Journal of Farm Economics*, 39, 8–20.

Griliches Z., Hall B., Hausman J. (1977). Missing data and self selection in large panels. Harvard University.

Gronau R. (1974). Wage comparisons — A selectivity bias. *Journal of Political Economy*, 82 (6), 1119–1143.

Hanoch G. (1976). A multivariate model of labor supply: Methodology for estimation. *Rand Corporation Paper* R-1980.

Heckman J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5 (4), 475–492.

Heckman J. (1977). Sample selection bias as a specification error with an application to the estimation of labor supply functions. *NBER Working Paper* #172 (revised).

Heckman J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46 (4), 931–959.

Jennrich R. (1969). Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics*, 40 (2), 633–643.

Johnson N., Kotz S. (1972), *Distribution in statistics: Continuous multivariate distributions*. New York: John Wiley & Sons.

Lewis H. (1974). Comments on selectivity biases in wage comparisons. *Journal of Political Economy*, 82 (6), 1145–1155.

Theil H. (1957). Specification errors and the estimation of economic relationships. *Revue de l'Institut International de Statistique*, 25 (1–3), 41–51.