Влияние включения в модель переменной, которая не должна быть включена

Допустим, что истинная модель представляется в виде:

$$y = \alpha + \beta_1 x_1 + u \qquad (6.15)$$

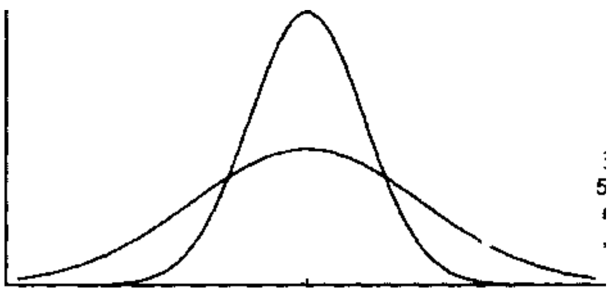
а вы считаете, что ею является

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u \qquad (6.16)$$

и рассчитываете оценку величины b_1, используя формулу (5.12) вместо выражения $Cov\ (x_1,y)/D(x_1)$.

В целом проблемы смещения здесь нет, даже если b_1 будет рассчитана неправильно. Величина $M(b_1)$ (математическое ожидание оценки) остается равной β_1, но в общем оценка будет неэффективной. Она будет более неустойчивой, в смысле наличия большей дисперсии относительно β_1, чем при правильном вычислении. Это проиллюстрировано на рис. 6.2.

(**Эффективная оценка** – это несмещенная оценка, имеющая наименьшую дисперсию среди всех несмещенных оценок. **Несмещенной** называют статистическую оценку, математическое ожидание которой равно оцениваемому параметру. **Смещенной** называют статистическую оценку, математическое ожидание которой не равно оцениваемому параметру.)



b_1

Рис.6.2. Функция плотности вероятности

Это можно легко объяснить интуитивно. Истинная модель может быть записана в виде:

$$y = \alpha + \beta_1 x_1 + 0x_2 + u \qquad (6.17)$$

Таким образом, если вы строите регрессионную зависимость y от x_1 и x_2 , то b_1 будет являться несмещенной оценкой величины β_1 а b_2 будет несмещенной оценкой нуля (при выполнении условий Гаусса-Маркова). Практически вы обнаруживаете для себя, что β_2 равно нулю. Если бы вы заранее поняли, что β_2 равно нулю, то могли бы использовать эту информацию для исключения x_2 и применить парную регрессию, которая в данном случае является более эффективной (***эффективная оценка* — это та, у которой дисперсия минимальна).**

Утрата эффективности в связи со включением x_2 в случае, когда она не должна была быть включена, зависит от корреляции между x_1 и x_2 . Сравните дисперсии величины b_1 при построении парной и множественной регрессии (табл. 6.5).

Дисперсия в общем окажется большей при множественной регрессии, и разница будет тем большей, чем ближе коэффициент корреляции к единице или -1. Единственным исключением в связи с проблемой утраты эффективности является вариант, когда коэффициент корреляции точно равен нулю. В этом случае оценка b_1 для множественной регрессии совпадает с оценкой для парной регрессии. Доказательство этого опустим, поскольку оно довольно простое.

В выводе о несмещенности есть одно исключение, которое необходимо иметь в виду. Если величина ***x2*** коррелирует с ***u*,** то коэффициенты регрессии будут в конечном счете смещенными. Если модель записать как уравнение (6.17), то это будет означать, что четвертое условие Гаусса-Маркова применительно к величине ***x2*** не выполняется.

Иллюстрация, основанная на эксперименте по методу Монте-Карло

В эксперименте по методу Монте-Карло, описанном в разделе 6.2, исследователь переоценил влияние образования на доход из-за того, что он не учел зависимости дохода в данной стране от величины ***IQ*** и того обстоятельства, что величина ***S*** там отчасти играла роль замещающей переменной для ***IQ*** в неправильно специфицированном уравнении парной регрессии. Будем помнить об этом и предположим, что наш исследователь, являющийся уже экспертом в данном вопросе, приглашен в качестве консультанта для проведения аналогичного исследования в соседней стране.

Может оказаться, что в новой стране подход более формален, чем в первой, и доход здесь определяется только образованием (и удачей) без учета способностей как таковых. Пусть базовый доход здесь снова равен 10 ООО, с добавлением 2000 за каждый год учебы сверх минимальных 10 лет, плюс (или минус) некоторая величина, зависящая от фактора удачи. Истинным соотношением поэтому будет:

$$\boldsymbol{y} = 10\,000 + 2000\,(S - 10) + \boldsymbol{u} = -10\,000 + 2000S + \boldsymbol{u} \qquad (6.18)$$

Исследователь снова делает выборку из 20 человек, и по удивительному совпадению все они имеют одинаковые характеристики, показанные в первой части табл. 5.2. В этом случае имеются также данные о величине ***IQ.*** Считая, что включение величины ***IQ*** в уравнение регрессии не причинит вреда, исследователь проводит эту операцию и получает следующее соотношение (стандартные ошибки указаны в скобках):

$$\hat{y} = -13\ 336 + 2140S + 18IQ. \qquad (6.19)$$
$$(4155) \quad (151) \quad (43)$$

Результат действительно неплохой. 95-процентный доверительный интервал для константы включает в себя ее истинное значение —10 000, и аналогичный интервал для ***S*** включает значение 2000. Таким образом, полученные оценки незначимо отличаются от истинных величин с 5-процентным уровнем значимости. Точно так же коэффициент ***IQ*** незначимо отличается от нуля.

Если бы при этом была использована правильная спецификация, то результатом было бы:

$$\hat{y} = -11\ 782 + 2163S. \qquad (6.20)$$
$$(с.о.)\ (1851) \quad (137)$$

Оценка константы здесь лучше, однако оценка коэффициента при переменной ***S*** недостаточно хороша (влияние фактора удачи оказалось относительно незначительным).

И вновь здесь нельзя слишком полагаться на результаты одного эксперимента.

Экс пери мент	*Правильная спецификация*					*Спецификация исследователя*				
	Конс танта	*с.о.*	S	*с.о.*	*Конс танта*	*с. о.*	S	*с.о.*	*IQ*	*с. о.*
1	**-11781**	**185**	**2163**	**137**	**-13336**	**4155**	**2140**	**151**	**18**	**43**
2	**-11940**	**249**	**215**	**184**	**-3019**	**5067**	**2290**	**184**	**-10**	**52**
3	**-7092**	**234**	**182**	**173**	**-11463**	**5150**	**1755**	**187**	**51**	**53**
4	**-7152**	**213**	**172**	**158**	**-15371**	**4273**	**1597**	**155**	**95**	**44**
5	**-9116**	**204**	**195**	**151**	**-14535**	**4371**	**1872**	**158**	**63**	**45**
6	**-12446**	**157**	**2230**	**116**	**-16742**	**3352**	**2167**	**121**	**50**	**35**
7	**-12510**	**246**	**2177**	**182**	**-6727**	**5329**	**2263**	**193**	**-67**	**55**
8	**-11487**	**236**	**2164**	**175**	**-18187**	**5005**	**2065**	**181**	**77**	**52**
9	**-4733**	**232**	**1644**	**172**	**-5384**	**5354**	**1634**	**190**	**8**	**54**
10	**-13742**	**194**	**2290**	**144**	**-13839**	**4386**	**2289**	**159**	**1**	**45**

В табл. 6.6 сведены вместе результаты повторения еще девяти таких же экспериментов с изменением в каждой выборке только значений случайной составляющей. Из табл. 6.6 можно сделать следующие выводы.

1. Результаты исследователя не выглядят смещенными, даже если спецификация является неправильной. Оценка константы колеблется около —10 000, а оценка коэффициента величины ***S***

— около 2000. (Естественно, что результаты оценивания правильной спецификации тоже будут несмещенными.)

2. Результаты оценивания правильной спецификации в целом более точны, поскольку эта спецификация оказывается более эффективной. Но данное утверждение не всегда верно, и в ряде случаев неправильная спецификация дает результат ближе к истине. Причиной этого является то, что относительная неэффективность спецификации исследователя зависит от корреляции между ***S*** и ***IQ***, а корреляция оказывается не достаточно тесной (в выборке из табл. 6.1), чтобы вызвать большие расхождения с истинными значениями.
3. Более высокая эффективность правильной спецификации должна отражаться в меньших стандартных ошибках, и это в целом действительно подтверждается.
4. Оценки коэффициентов при ***IQ*** в спецификации исследователя в целом незначительно отличны от нуля. В эксперименте ***4*** имеется единственное отклонение, когда истинная гипотеза о нулевом значении отвергается при 5-процентном уровне значимости. Это является хорошим примером ошибки I рода (см. Обзор).