Лекция № 4

Октябрь 2020

1 Различные аспекты множественной регрессии

Рассмотрим некоторые проблемы, которые часто возникают при практическом использовании регрессионных моделей.

На практике исследователю нередко приходится сталкиваться с ситуацией, когда полученная им регрессия является "плохой", т.е. t-статистики большинства оценок малы, что свидетельствует о незначимости соответствующих независимых переменных (регрессоров). В то же время F-статистика может быть достаточно большой, что говорит о значимости регрессии в целом. Одна из возможных причин такого явления носит название *мультиколлинеарности* и возникает при наличии высокой корреляции между регрессорами.

Регрессионные модели являются достаточно гибким инструментом, позволяющим, в частности, оценивать влияние качественных признаков (пол, профессия, наличие детей и т.п.) на изучаемую переменную. Это достигается введением в число регрессоров так называемых *фиктивных переменных*, принимающих, как правило, значения 1 или 0 в зависимости от наличия или отсутствия соответствующего признака в очередном наблюдении. С формальной точки зрения фиктивные переменные ничем не отличаются от других регрессоров. Однако следует обратить особое внимание на правильное их использование и интерпретацию оценок.

1.1 Мультиколлинеарность

Одним из условий классической регрессионной модели является предположеие о линейной независимости объясняющих переменных, что означает *линейную независимость столбцов матрицы регрессоров* X или (эквивалентно) что *матрица* $(X'X)^{-1}$ *имеет полный ранг* k. При нарушении этого условия, т.е. когда один из столбцов матрицы X есть линейная комбинация остальных столбцов, говорят, что имеет место *полная коллинеарность*. В этой ситуации нельзя построить МНК-оценку параметра β, что формально следует из сингулярности матрицы $X'X$ и невозможности решить нормальные уравнения. Нетрудно также понять и содержательный смысл этого

явления. Рассмотрим следующий простой пример регрессии (Greene, 1997):

$$C = \beta_1 + \beta_2 S + \beta_3 N + \beta_4 T + \epsilon$$

, где C - потребление, S - зарплата, N - доход, получаемый вне работы, T - полный доход. Посколько выполнено равенство $T = S + N$, то для произвольного числа h исходную регрессию можно переписать в следующем виде:

$$C = \beta_1 + \beta_2' S + \beta_3' N + \beta_4' T + \epsilon$$

, где $\beta_2' = \beta_2 + h$, $\beta_3' = \beta_3 + h$, $\beta_4' = \beta_4 - h$. Таким образом, одни и те же наблюдения могут быть объяснены различными наборами коэффициентов β. Эта ситуация тесно связана с проблемой *идентифицируемости* системы, о чем более подробно будет говориться позднее. Кроме того, если с учетом равентва $T = S + N$ переписать исходную систему в виде

$$C = \beta_1 + (\beta_2 + \beta_4)S + (\beta_3 + \beta_4)N + \epsilon$$

, то становится ясно, что оценить можно лишь три параметра β_1, $(\beta_2 + \beta_4)$ и $(\beta_3 + \beta_4)$, а не четыре исходных. В общем случае можно показать, что если $rank(X'X) = l < k$, то оценить можно только l линейных комбинаций исходных коэффициентов. Если есть полная коллинеарность, то можно выделить в матрице X максимальную линейно независимую систему столбцов и, удалив остальные столбцы, провести новую регрессию.

На практике полная коллинеарность встречается исключительно редко. Гораздно чаще приходится сталкиваться с ситуацией, когда матрица X имеет полный ранг, но между регрессорами имеется высокая степень корреляции, т.е., говоря нестрого, *когда матрица $X'X$ близка к вырожденной*. Тогда говорят о наличии мультиколлинеарности. В этом случае МНК-оценка формально существует, но обладает "плохими свойствами".

Мультиколлинеарность может возникать в силу разных причин. Например несколько независимых переменных могут иметь общий временной тренд, относительно которого они совершают малые колебания. В частности, так может случиться, когда значения одной независимой переменной являются лагированными значениями другой.

Выделим некоторые наиболее характерные признаки мультиколлинеарности:

1. Небольшое изменение исходных данных (например, добавление новых наблюдений) приводит к существенному изменению оценок коэффициентов модели.

2. Оценки имеют большие стандартные ошибки, малую значимость, в то время как модель в целом является значимой (высокое значение коэффициента детерминации R^2 и соответствующей F-статистики).

3. Оценки коэффициентов имеют неправильные с точки зрения теории знаки или неоправданно большие значения.

Что же делать, если по всем признакам имеется мультиколлинеарность? Однозначного ответа на этот вопрос нет, и среди эконометристов есть разные мнения на этот счет. Существует даже такая школа, представители которой считают, что и не нужно ничего делать, посколько "так устроен мир" (см. Kennedy, 1992). Мы здесь не ставим цель дать достаточно полное описание методов борьбы с мультиколлинеарностью. Более подробно об этом можно прочесть, например, в (Greene, 1997, глава 9).

У неискушенного исследователя при столкновении с проблемой мультиколлинеарности может возникнуть естественное желание отбросить "лишние" независимые переменные, которые, возможно, служат ее причиной. Однако следует помнить, что при этом могут возникнуть новые трудности. Во-первых, далеко не всегда ясно, какие переменные являются лишними в указанном смысле. Мультиколлинеарность означает лишь приблизительную линейную зависимость между столбцами матрицы X, но это не всегда выделяет "лишние" переменные. Во-вторых, во многих ситуациях удаление каких-либо независимых переменных может значительно отразиться на содержательном смысле модели. Наконец, отбрасывание так называемых существенных переменных, т.е. независимых переменных, которые реально влияют на изучаемую зависимую переменную, приводит к смещенности МНК-оценок.

Что можно предпринять в случае мультиколлинеарности?

1. Фактор - σ_E^2.

   Можно попытаться уменьшить σ_ϵ^2. Случайный член отражает воздействие на переменную Y всех влияющих на нее переменных, не включенных в регрессионное уравнение. Таким образом если найти важную переменную,не включенную в модель и, следовательно, вносящую вклад в ϵ, то при ее включении дисперсия случайного члена уменьшится.

2. Фактор - число наблюдений

   Можно увеличить объем выборки

   Можно применить метод группировки. Например, страну можно поделить на области, отдельные города и зоны. Далее используя случайный отбор географических зон и случайной выборки, обеспечивают должное представительство различных регионов. При работе с временными рядами можно увеличить выборку перейдя от длинных к коротким временным интервалам: от годовых к квартальным, месячным и т.д.

3. Увеличение дисперсии объясняющих переменных

   Это можно сделать лишь на стадии планирования проводимого опроса. Например, при планировании опроса семей по доходам и расходам следует включить в опрос относительно богатые и бедные семьи наряду с семьями со средним доходом.

4. На стадии планирования опроса следует получить такую выборку, в которой оюъясняющие переменные были бы как можно меньше связаны между собой.

5. Если коррелированные переменные связаны между собой концептуально, то может быть рационально объединить их в совокупный индекс.

6. Можно исключить некоторые из коррелированных переменных если их коэффициенты незначительны. Это может быть опасно, т.к. некоторые из исключаемы исключаемых коэффициентов могли принадлежать модели и их невключение может служить причиной смещения оценки.

7. Использовать внешнюю информацию относительно коэффициента одной из переменных в уравнении (например, экспертная оценка).

1.2 Фиктивные переменные

Как правило, независимые переменные в регрессионных моделях имеют "непрерывные" области изменения (нац. доход, уровень безработицы, размер зарплаты и т.п.). Однако теория не накладывает никаких ограничений на характер регрессоров, в частности, некоторые переменные могут принимать всего два значения или, в более общей ситуации, дискретное множество значений. Необходимость рассматривать такие переменные возникает довольно часто в тех случаях, когда требуется принимать во внимание какой-либо *качественный признак*. Например, при исследовании зависимости зарплаты от различных факторов может возникнуть вопрос, влияет ли на ее размер, и если да, то в какой степени, наличие у работника высшего образования. Также можно задать вопрос, существует ли дискриминация в оплате труда между мужчинами и женщинами. В принципе можно оценивать соответствующие уравнения внутри каждой категории, а затем изучать различия между ними, но введение дискретных переменных позволяет оценивать одно уравнение сразу по всем категориям.

Покажем, как это можно сделать в примере с зарплатой. Пусть $x_t = (x_{t1}, ..., x_{tk})'$ - набор объясняющих переменных, т.е. первоначальная модель описывается уравнениями

$$y_t = x_{t1}\beta_1 + .. + x_{tk}\beta_k + \epsilon_t = x_t'\beta + \epsilon_t, t = 1, ..., n, \quad (1)$$

где y_t - размер зарплаты t-го работника. Теперь мы хотим включить в рассмотрение такой фактор, как наличие или отсутствие высшего образования. Введем новую, бинарную переменную d, полагая $d_t = 1$, если в t-м наблюдении индивидуум имеет высшее образование, и $d_t = 0$ в противном случае, и рассмотрим новую систему

$$y_t = x_{t1}\beta_1 + ... + x_{tk}\beta_k + d_t\delta + \epsilon_t = z_t'\gamma + \epsilon_t, t = 1, ..., n, \quad (2)$$

где $z = (x_1, ..., x_k, d)' = (x', d)'$, $\gamma = (\beta_1, ..., \beta_k, \delta)'$. Иными словами, принимая модель 1, мы считаем, что средняя зарплата есть $x'\beta$ при отсутствии высшего образования и $x'\beta + \delta$ - при его наличии. Таким образом, величина δ интерпретируется как среднее изменение зарплаты при переходе из одной категории (без высшего образования) в другую (с высшим образованием) при неизменных значениях остальных параметров. К системе 1 можно применить метод наименьших квадратов и получить оценки соответствующих коэффициентов. Легко понять, что, тестируя гипотезу $\delta = 0$, мы проверяем предположение о несущественном различии в зарплате между категориями.

Замечание. В англоязычной литературе по эконометрике переменные указанного типа называются *dummy variables*, что на русский язык часто переводится как "фиктивные переменные" (см., например, Джонстон, 1980). Следует, однако, ясно понимать, что d такая же "равноправная" переменная, как и любой из регрессоров $x_j, j = 1, ..., k$. Ее "фиктивность" состоит только в том, что она количественным образом описывает качественный признак.

Качественное различие можно формализовать с помощью любой переменной, принимающей два значения, а не обязательно значения 0 или 1. Однако на практике почти всегда используют фиктивные переменные типа "0-1", поскольку в этом случае интерпретация выглядит наиболее просто. Если бы в рассмотренном выше примере переменная d принимала значение, скажем, 5 для индивидуума с высшим образованием и 2 для индивидуума без высшего образования, то коэффициент при этом регрессоре равнялся бы трети среднего изменения зарплаты при получении высшего образования.

Если включаемый в рассмотрение качественный признак имеет не два, а несколько начений, то в принципе можно было бы ввести дискретную переменную, принимающую такое же количество значений. Но этого фактически никогда не делают, так как тогда трудно дать содержательную интерпретацию соответствующему коэффициенту. *В этих случаях целесообразнее использовать несколько бинарных переменных.* Типичным примером подобной ситуации является исследование сезонных колебаний. Пусть, например, y_t - объем потребления некоторого продукта в месяц t, и есть все основания считать, что потребление зависит от времени года. Для выявления влияния сезонности можно ввести три бинарные переменные d_1, d_2, d_3:

1. $d_{t1} = 1$, если месяц t является зимним, $d_{t1} = 0$ в остальных случаях;
2. $d_{t2} = 1$, если месяц t является весенним, $d_{t2} = 0$ в остальных случаях;
3. $d_{t3} = 1$, если месяц t является летним, $d_{t3} = 0$ в остальных случаях,

и оценивать уравнение

$$y_t = \beta_0 + \beta_1 d_{t1} + \beta_2 d_{t2} + \beta_3 d_{t3} + \epsilon_t. \tag{3}$$

Отметим, что мы не вводим четвертую бинарную переменную d_4, относящуюся к осени, иначе тогда для любого месяца t выполнялось бы тождество $d_{t1} + d_{t2} + d_{t3} + d_{t4} = 1$, что означало бы линейную зависимость регрессоров

в 3 и, как следствие, невозможность получения МНК-оценок. Такая ситуация, когда сумма фиктивных переменных тождественно равна константе, также включенной в регрессию, называется *"dummy trap"*. Иначе говоря, среднемесячный объем потребления есть β_0 для осенних месяцев, $\beta_0 + \beta_1$ - для зимних, $\beta_0 + \beta_2$ - для весенних и $\beta_0 + \beta_3$ - для летних. Таким образом, оценки коэффициентов $\beta_i, i = 1, 2, 3$, показывают средние сезонные отклонения в объеме потребления по отношению к осенним месяцам. Тестируя, например, стандартную гипотезу $\beta_3 = 0$, мы проверяем предположение о несущественном различии в объеме потребления между летним и осенним сезоном, гипотеза $\beta_1 = \beta_2$ эквивалентна предположению об отсутствии различия в потреблении между зимой и весной и т.д: $H_0 : \beta_3 = 0, H_0 : \beta_1 = \beta_2$.

Фиктивные переменные, несмотря на свою внешнюю простоту, являются весьма гибким инструментом при исследовании влияния качественных признаков. Рассмотрим еще один пример. В предыдущей модели мы интересовались сезонными различиями лишь для среднемесячного объема потребления. Модифицируем ее, введя новую независимую переменную i - доход, используемый на потребление. Как известно, в регрессии

$$y_t = \beta_0 + \beta_1 i_t + \epsilon_t \tag{4}$$

коэффициент β_1 носит название "склонность к потреблению". Поэтому естественно поставить задачу исследовать влияние сезона на склонность к потреблению. Для этого можно рассмотреть модель

$$y_t = \beta_0 + \beta_1 d_{t1} + \beta_2 d_{t2} + \beta_3 d_{t3} + \beta_4 d_{t1} i_t + \beta_5 d_{t2} i_t + \beta_6 d_{t3} i_t + \beta_7 i_t + \epsilon_t, \tag{5}$$

согласно которой склонность к потреблению зимой, весной, летом и осенью есть $\beta_4 + \beta_7, \beta_5 + \beta_7, \beta_6 + \beta_7$, и β_7 соответственно. Как и в предыдущей модели, можно тестировать гипотезы об отсутствии сезонных влияний на склонность к потреблению.

Фиктивные переменные позволяют строить и оценивать так называемые кусочно-линейные модели, которые можно применять для исследования структурных изменений. Как и раньше, проще всего это продемонстрировать на примере.

Пусть y - зависимая переменная и пусть для простоты есть только две независимые переменные: x и постоянный член. Предположим, что x и y представлены в виде временных рядов $\{(x_t, y_t), t = 1, ..., n\}$ (например, x_t - размер основного фонда некоторого предприятия в период t, y_t - объем продукции, выпущенной в этот же период). Из некоторых априорных соображений исследователь считает, что в момент времени t_0 произошла структурная перестройка и линия регрессии будет отличаться от той, что была до момента t_0, но общая линия остается непрерывной (рис. 1).

Чтобы оценить такую модель, введем бинарную переменную $r_t = 0$, если $t \leq t_0$ и $r_t = 1$, если $t > t_0$ и запишем следующее регрессионное уравнение:

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 (x_t - x_{t_0}) r_t + \epsilon_t. \tag{6}$$

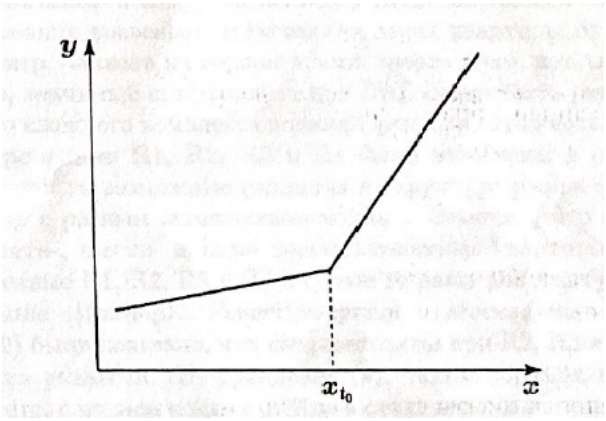

Рис. 1: Рис. 1

Нетрудно проверить, что регрессионная линия, соответствующая 6, имеет коэффициент наклона β_2 для $t \leq t_0$ и $\beta_2 + \beta_3$ для $t > t_0$, и разрыва в точке x_{t_0} не происходит. Таким образом, тестируя гипотезу $\beta_3 = 0$, мы проверяем предположение о том, что *фактически структурного изменения не произошло*, $H_0 : \beta_3 = 0$.

Выводы:

- для исследования влияния качественных признаков в модель можно вводить бинарные (фиктивные) переменные, которые, как правило, принимают значение 1, если данный качественный признак присутствует в наблюдении, и значение 0 при его отсутствии;
- способ включения фиктивных переменных зависит от априорной информации относительно влияния соответствующих качественных признаков на зависимую переменную и от гипотез, которые проверяются с помощью модели;
- от способа включения фиктивной переменной зависит и интерпретации оценки коэффициента при ней.

1.3 Частная корреляция

В том случае, когда имеются одна независимая и одна зависимая переменные, естественной мерой зависимости (в рамках линейного подхода) является (выборочный) коэффициент корреляции между ними. Использование множественной регрессии позволяет обобщить это понятие на случай, когда имеется несколько независимых переменных. Корректировка здесь необходима по следующим очевидным соображениям. Высокое значение коэффициента может означать высокую степень зависимости или то, что существует третья переменная, от которой зависят и первая и вторая. Поэтому

возникает естественная задача найти "чистую" корреляцию между двумя переменными, исключив (линейное) влияние других факторов. Это можно сделать с помощью коэффициента частной корреляции. Для простоты предположим, что имеется регрессионная модель

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \epsilon,$$

где, как обычно, $y - n \times 1$ вектор наблюдений зависимой переменной, $x_1, x_2 - n \times 1$ векторы независимых переменных, $\beta_0, \beta_1, \beta_2$ – скалярные параметры, $\epsilon - n \times 1$ вектор ошибок. Наша цель - определить корреляцию между y и, например, первым регрессором x_1 после исключения влияния x_2.

Соответствующая процедура устроена следующим образом:

1. Осуществим регрессию y на x_2 и константу и получим прогнозные значения $\hat{y} = \hat{\alpha}_1 + \hat{\alpha}_2 x_2$.

2. Осуществим регрессию x_1 на x_2 и константу и получим прогнозные значения $\hat{x}_1 = \hat{\gamma}_1 + \hat{\gamma}_2 x_2$.

3. Удалим влияние x_2, взяв остатки $e_y = y - \hat{y}$ и $e_{x_1} = x_1 - \hat{x}_1$.

4. Определим (выборочный) *коэффициент частной корреляции* между y и x_1 при исключении влияния x_2 как (выборочный) коэффициент корреляции между e_y и e_{x_1}:

$$r(y, x_1|x_2) = r(e_y, e_{x_1}). \tag{7}$$

   Напомним, что из свойств метода наименьших квадратов следует, что e_y и e_{x_1} не коррелированы с x_2: $x_1^T e_{x_1} = 0$, $x_2^T e_y = 0$. Именно в этом смысле указанная процедура соответствует интуитивному представлению об "исключении (линейного) влияния переменной x_2"

Прямыми вычислениями можно показать (см. упражнение 2), что справедлива следующая формула, связывающая коэффициенты частной и обычной корреляции:

$$r(y, x_1|x_2) = \frac{r(y, x_1) - r(y, x_2)r(x_1, x_2)}{\sqrt{1 - r^2(x_1, x_2)}\sqrt{1 - r^2(y, x_2)}}. \tag{8}$$

Значения $r(y, x_1|x_2)$ лежат в интервале $[-1, 1]$, как у обычного коэффициента корреляции. Равенство коэффициента $r(y, x_1|x_2)$ нулю означает, говоря нестрого, отсутствие *прямого линейного влияния переменной x_1 на y*.

Существует тесная связь между коэффициентом частной корреляции $r(y, x_1|x_2)$ и коэффициентом детерминации R^2, а именно

$$r^2(y, x_1|x_2) = \frac{R^2 - r^2(y, x_2)}{1 - r^2(y, x_2)}, \tag{9}$$

или

$$1 - R^2 = (1 - r^2(y, x_2))(1 - r^2(y, x_1|x_2)).$$

Описанная выше процедура очевидным образом обобщается на случай, когда исключается влияние не одной, а нескольких переменных: достаточно переменную x_2 заменить на набор переменных X_2, сохраняя определение 7. Формула 8, естественно, усложнится. Подробнее об этом можно прочесть в книге (Айвазян и др., 1986).

Проиллюстрируем приведенное выше понятие частных коэффициентов корреляции и их отличие от обычных коэффициентов корреляции на следующем примере.

Пример. Рынки валютных фьючерсов. Рассмотрим вопрос о связи российского и западных рынков валютных фьючерсов.

В настоящее время несколько российских бирж ведут торговлю срочными контрактами на поставку доллара США: МТБ, МЦФБ, РТСБ и др. Однако (см. Яковлев, Бессонов, 1995а, 1995б) в течение периода наблюдений (ноябрь 1992 г. - сентябрь 1995 г.) на МТБ приходилось от 75 до 85% общего объема торговли. Поэтому в качестве цен фьючерсных контрактов на поставку доллара США мы выбрали котировки контрактов на МТБ.

Динамика цен валютных фьючерсов на Западе не сильно зависит от биржи. Для анализа мы взяли биржу с наибольшим объемом торговли - IMM (International Monetary Market, Chicago).

Мы используем ежедневные данные - цена закрытия для IMM и котировочная цена для МТБ - показатели, которые используют торговые палаты этих бирж для ежедневного перерасчета позиций инвесторов (вариационной маржи).

В качестве параметров для сравнения мы взяли не сами цены контрактов, а "доходности", приведенные к годичному базису, определяемые как

$$y_{t,T} = (\ln F_t^T - \ln S_t)/(T-t) \cdot 365,$$

где F_t^T - цена контракта в момент времени t на поставку 1 доллара в момент времени T (т.е. со сроком до поставки $T-t$); S_t - спот-курс доллара в момент t. (Для рубля - данные ММВБ, для немецкой марки DM, британского фунта BP, японской иены JY - данные IMM.) $y_{t,T}^{RU}$,$y_{t,T}^{DM}$,$y_{t,T}^{BP}$,$y_{t,T}^{JY}$ обозначают доходности контрактов на поставку 1 доллара в рублях, DM, BP, JY. На наш взгляд, этот показатель в меньшей мере зависит от темпа инфляции, чем сама цена контракта. Время t изменяется в днях.

Рассмотрим *таблицу коэффициентов корреляции доходностей* $y_{t,T}^{RU}$,$y_{t,T}^{DM}$, $y_{t,T}^{BP}$,$y_{t,T}^{JY}$:

	RU	DM	BP	JY
RU	1			
DM	0.626	1		
BP	0.380	0.775	1	
JY	0.615	0.919	0.602	1

Из таблицы видны высокие (0.602, 0.775, 0.919) значения коэффициентов корреляции показателей для западных валют, что неудивительно ввиду высокой степени интегрированности западных финансовых рынков. Удивле-

ние вызывают высокие 0.615 (0.626) значения коэффициентов корреляции показателей для рубля и японской иены (немецкой марки).

Рассмотрим теперь таблицу коэффициентов частной корреляции между доходностями $y_{t,T}^{XX}$ для XX = RU,DM,BP,JY (устранено влияние временного тренда t).

	RU	DM	BP	JY
RU	1			
DM	0.024	1		
BP	0.008	0.807	1	
JY	-0.003	0.488	0.276	1

Теперь мы видим картину более реалистичную! Наиболее тесно связаны между собой европейские валюты (BP,DM), слабее связь европейских валют и японской иены и практически отсутствует связь российской валюты с западными.

Таким образом, высокие коэффициенты корреляции в первой таблице, например 0.626 для RU-DM, были лишь следствием того, что на интервале наблюдений (ноябрь 1992 г. - сентябрь 1995 г.) отмечалось падение курса рубля по отношению к доллару и падение курса доллара по отношению к немецкой марке, т.е. эта корреляция является следствием наличия временного тренда в $y_{t,T}^{RU}$ и $y_{t,T}^{DM}$. Наш вывод подтверждается также тем, что коэффициенты корреляции $y_{t,T}^{RU}$ и $y_{t,T}^{DM}$ с t достаточно высоки (-0.673, 0.920).

1.4 Процедура пошагового отбора переменных

Коэффициент частной корреляции часто используется при решении проблемы *спецификации модели*.

Иногда исследователь заранее знает характер зависимости исследуемых величин, опираясь, например, на экономическую теорию, предыдущие результаты, априорные знания и т.п., и задача состоит лишь в оценивании неизвестных параметров. (По существу, во всех наших предыдущих рассуждениях мы неявно предполагали, что имеется именно такая ситуация.) Классический пример - оценивание параметров производственной функции Кобба-Дугласа $Y = AK^{\alpha}L^{\beta}$, где Y - совокупный выпуск, K - капиталовложения и L - трудозатраты. Логарифмируя это выражение получаем линейное относительно $\ln A, \alpha, \beta$ уравнение, из которого, например, с помощью метода наименьших квадратов можно получить оценки этих параметров, проверять те или иные гипотезы и т.д.

Однако на практике довольно часто приходится сталкиваться с ситуацией, когда имеется большое число наблюдений различных параметров (независимых переменных), но нет априорной модели изучаемого явления. Возникает естественная проблема, какие переменные включить в регрессионную схему. Теоретические вопросы, связанные с этой проблемой, будут изложены далее.

В компьютерные пакеты включены различные эвристические процедуры *пошагового отбора регрессоров*. Основными пошаговыми процедурами

являются *1) процедура последовательного присоединения, 2) процедура присоединения-удаления и 3) процедура последовательного удаления.* Опишем кратко одну из таких процедур, использующую понятие коэффициента частной корреляции.

Процедура присоединения-удаления

На первом шаге из исходного набора объясняющих переменных выбирается (включается в число регрессоров) переменная, имеющая наибольший по модулю коэффициент корреляции с зависимой переменной y.

$$x_i : |r(y, x_i)| = \max_j |r(y, x_j)|.$$

Второй шаг состоит из двух подшагов. На первом из них, который выполняется, если число регрессоров уже больше двух, делается попытка исключить один из регрессоров. Ищется тот регрессор x_s, удаление которого приводит к наименьшему уменьшению коэффициента детерминации. Затем сравнивается значение F-статистики для проверки гипотезы $H_0 : \beta_s = 0$ о незначимости этого регрессора с некоторым заранее заданным пороговым значением $F_{\text{искл}}$. Если $F < F_{\text{искл}}$, то x_s удаляется из списка регрессоров. Второй подшаг состоит в попытке включения нового регрессора из исходного набора предсказывающих переменных. Ищем переменную x_p с наибольшим по модулю частным коэффициентом корреляции (исключается влияние ранее включенных в уравнение регрессоров) и сравниваем значение F-статистики для проверки гипотезы $H_0 : \beta_p = 0$ о незначимости этого регрессора с некоторым заранее заданным пороговым значением $F_{\text{вкл}}$.

$$x_p : |r(y, x_p|x_i)| = max_j r(y, x_j|x_i)|.$$

Если $F > F_{\text{вкл}}$, то x_p включается в список регрессоров. Обычно выбирают $F_{\text{искл}} < F_{\text{вкл}}$. Второй шаг повторяется до тех пор, пока происходит изменение списка регрессоров. Конечно, ни одна из пошаговых процедур не гарантирует получение оптимального по какому-либо критерию набора регрессоров.

Подробное описание пошаговых процедур содержится в книге (Айвазян и др., 1985).