Лекция 7

Спецификация множественной регрессии

1. Результаты ошибочного отбора факторов.

2. Отбор факторов в уравнение множественной регрессии.

3. Выбор формы связи.

1. Результаты ошибочного отбора факторов.

Все предыдущие рассуждения и выводы, касающиеся классической или обобщенной модели множественной регрессии, основывались на предположении, что выполнена *правильная спецификация* модели.

Как уже отмечалось, под спецификацией множественной модели понимается *отбор* объясняющих переменных в модель и установление *формы связи* между этими переменными и зависимой переменной.

Факторы, включаемые в модель, должны удовлетворять следующим требованиям:

1. Факторы должны быть количественно измеримы. Если необходимо включить качественный фактор, то ему нужно придать количественную определенность, т.е. ввести в рассмотрение фиктивные переменные.

2. Между факторами не должно быть высокой корреляции ($r_{x_i x_j} \geq 0{,}8$), тем более линейной функциональной зависимости. Иначе нельзя допускать мультиколлинеарности объясняющих переменных.

3. Каждый отбираемый фактор должен быть достаточно тесно связан с зависимой переменной y . Если фактор мало влияет на y , то его не следует включать в модель.

При отборе факторов возможны ошибки **двух** типов. Можно ошибочно включить в уравнение переменные, которых там не должно быть или ошибиться и не включить фактор, который там должен присутствовать.

Какие последствия этих ошибок? Оказывается, свойства оценок коэффициентов регрессии в значительной мере зависят от правильной спецификации модели. Рассмотрим эти вопросы подробнее.

Рассмотрим вначале случай, когда в модели отсутствует *существенная* переменная.

Пусть переменная y зависит от двух факторов x_1 и x_2 в соответствии с соотношением:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Однако мы не уверены в значимости фактора x_2 и считаем, что модель должна выглядеть так:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

Получим оценку парной регрессии

$$\tilde{y} = b_0 + b_1 x_1,$$

вычислив параметр b_1 по формуле:

$$b_1 = \frac{\overline{x_1 y} - \bar{x}_1 \cdot \bar{y}}{\overline{x_1^2} - (x_1)^2} = \frac{\operatorname{cov}(x_1, y)}{\sigma_{x_1}^2}. \quad (1)$$

Убедимся, что оценка (1) будет смещенной, если $\beta_2 \neq 0$.

Для этого выполним следующие преобразования

$$b_1 = \frac{\operatorname{cov}(x_1, y)}{\sigma_{x_1}^2} = \frac{\operatorname{cov}(x_1, \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon)}{\sigma_{x_1}^2} =$$

$$= \frac{1}{\sigma_{x_1}^2}\left(\operatorname{cov}(x_1, \beta_0) + \operatorname{cov}(x_1, \beta_1 x_1) + \operatorname{cov}(x_1, \beta_2 x_2) + \operatorname{cov}(x_1, \varepsilon)\right)$$

$$= \frac{1}{\sigma_{x_1}^2}\left(0 + \beta_1 \operatorname{cov}(x_1, x_1) + \beta_2 \operatorname{cov}(x_1, x_2) + \operatorname{cov}(x_1, \varepsilon)\right) =$$

$$= \beta_1 \frac{\operatorname{cov}(x_1, x_1)}{\sigma_{x_1}^2} + \beta_2 \frac{\operatorname{cov}(x_1, x_2)}{\sigma_{x_1}^2} + \frac{\operatorname{cov}(x_1, \varepsilon)}{\sigma_{x_1}^2}.$$

Поскольку $\operatorname{cov}(x_1, x_1) = \sigma^2_{x_1}$ и в силу того, что x_1 не является случайной величиной (именно в этом предпосылка 1°), то имеем

$$\operatorname{cov}(x_1, \varepsilon) = 0$$

Отсюда окончательно получаем:

$$b_1 = \beta_1 + \beta_2 \frac{\operatorname{cov}(x_1, x_2)}{\sigma^2_{x_1}}.$$

Находим математическое ожидание от обеих частей последнего выражения

$$M(b_1) = \beta_1 + \beta_2 \frac{\operatorname{cov}(x_1, x_2)}{\sigma_{x_1}^2},$$

так как слагаемые в правой части остаются неизменными.

Таким образом, при неравенстве $\beta_2 \neq 0$ и $\operatorname{cov}(x_1, x_2) \neq 0$ т.е. оценка $M(b_1) \neq \beta_1$, является смещенной на величину b_1

$$\Delta = \beta_2 \frac{\operatorname{cov}(x_1, x_2)}{\sigma_{x_1}^2}.$$

Направление смещения будет зависеть от знаков величин β_2 и $\mathrm{cov}(x_1, x_2)$. Если они будут одного знака (например, больше нуля), то b_1 будет давать в среднем завышенные оценки коэффициента β_1.

Рассмотрим теперь последствия того случая, когда в модель включена *несущественная* переменная.

Допустим, что истинная модель имеет вид

$$y = \beta_0 + \beta_1 x_1 + \varepsilon,$$

а мы считаем, что ею является уравнение

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

и рассчитываем оценку b_1 по формуле (для двухфакторной регрессии):

$$\hat{b}_1 = \frac{\operatorname{cov}(x_1, y) \cdot \sigma_{x_1}^2 - \operatorname{cov}(x_2, y) \cdot \operatorname{cov}(x_1, x_2)}{\sigma_{x_1}^2 \cdot \sigma_{x_2}^2 - \left(\operatorname{cov}(x_1, x_2)\right)^2} \qquad (2)$$

вместо выражения (1).

Оценка $\hat{b}_1$ будет несмещенной ($M(\hat{b}_1) = \beta_1$), но в общем случае она будет неэффективной.

Действительно, можно показать, что дисперсии параметров b_1 и $\hat{b}_1$ вычисляются по формулам:

$$D(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3)$$

$$D(\hat{b}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \frac{1}{1 - r_{x_1 x_2}^2}. \qquad (4)$$

Из формулы (4) видно, что $D(\hat{b}_1)$ зависит от коэффициента корреляции $r_{x_1 x_2}$. Если $r_{x_1 x_2} = 0$, то дисперсии $D(b_1) = D(\hat{b}_1)$ совпадают , а в противном случае $D(b_1) < D(\hat{b}_1)$, т.е. оценка $\hat{b}_1$ не является эффективной.

В итоге можно сделать следующие выводы.

1. Если опущена переменная, которая должна быть включена в модель, то оценки регрессии, вообще говоря, оказываются *смещенными*, их значения могут существенно отличаться от оцениваемых значений коэффициентов .

2. Если в модель включена переменная, которой там не должно быть, то оценки регрессии становятся *неэффективными*. Стандартные ошибки оценок будут большими, и результаты тестирования будут неверными.

2. Отбор факторов в уравнение множественной регрессии.

При отборе факторов для множественной регрессии часто используются *частные* коэффициенты корреляции. С их помощью можно ранжировать факторы по степени влияния на результирующий признак и затем исключить маловлияющие факторы.

Другой подход основан на анализе матрицы коэффициентов корреляции.

На первом этапе в модель отбираются *потенциальные факторы* исходя из представления исследователя о природе взаимосвязи моделируемого показателя с другими экономическими переменными, т.е. исходя из сущности проблемы. Пусть их число рав-но n.

На втором этапе из числа потенциальных факторов выбираются такие объясняющие переменные, которые сильно коррелируют с объясняемой переменной и, одновременно, слабо коррелируют между собой.

В качестве исходной информации служит матрица коэффициентов корреляции, найденная для всех потенциальных переменных модели.

Задаётся уровень значимости α и для числа степеней свободы $k = n - 2$ рассчитывается критическое значение коэффициента корреляции

$$r^* = \sqrt{\frac{t_{кр}^2}{t_{кр}^2 + n - 2}},$$

где $t_{кр}$ –критическое значение распределения Стьюдента для двусторонней критической области.

Далее процедура отбора факторов состоит из следующих шагов.

1. Из множества потенциальных объясняющих переменных исключаются все факторы, для которых справедливо неравенство

$$\left| r_{yx_j} \right| \le r^*,$$

так как они несущественно коррелируют с переменной y .

2. Из множества оставшихся факторов x_h выбирается тот фактор , для которого выполняется равенство

$$\left|r_{yx_h}\right| = \max_i \left|r_{yx_i}\right|,$$

поскольку он является носителем наибольшего количества информации о переменной y.

3. Из оставшегося множества потенциальных переменных исключаются все факторы, для которых выполняется неравенство

$$\left| r_{x_i x_h} \right| > r^*,$$

поскольку эти факторы слишком сильно коррелируют с x_h и, следовательно, только воспроизводят представленную ею информацию.

Указанные шаги 1-3 повторяются вплоть до опустошения множества потенциальных объясняющих переменных до определенного числа p .

Для отбора факторов также используют так называемые процедуры *пошагового отбора переменных*. В компьютерные пакеты включены различные эвристические процедуры отбора факторов:

- процедура последовательного присоединения;
- процедура последовательного присоединения-удаления;
- процедура последовательного удаления.

К первому типу относят процедуру "всех возможных регрессий", которая заключается в следующем.

Для заданного значения $k\,(k=1,2,\ldots,n-1)$ путем полного перебора возможных комбинаций из k объясняющих переменных, отобранных из исходного потенциального набора факторов $x_1, x_2, \ldots, x_n$, определяют такие переменные $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$, для которых коэффициент детерминации с y был бы максимальным.

На первом шаге процедуры, полагая $k = 1$, находят одну объясняющую переменную из всего набора, которая является наиболее информированным фактором, т.е. для неё

$$R^2 \to \max.$$

На втором шаге процедуры ($k = 2$) определяется уже наиболее информативная пара переменных, которая имеет наиболее тесную связь с результатом y . В эту пару может и не войти тот фактор, который был отобран на первом шаге. Такой процесс продолжается до значения $k = n - 1$

В качестве критерия остановки процесса, т.е. выбора оптимального числа факторов модели предлагается следующее.

На каждом шаге вычисляется нижняя доверительная граница коэффициента детерминации

$$R^2_{\min}(k) = \hat{R}^2(k) - 2\sqrt{\frac{2k(n-k-1)}{(n-1)(n^2-1)}} \cdot \left(1 - R^2(k)\right), \quad (5)$$

где $\hat{R}^2(k)$ – скорректированный коэффициент детерминации для наиболее информативных факторов, $R^2(k)$ – обычный коэффициент детерминации.

В соответствии с критерием останова следует выбрать такое k_0 , при котором величина (5) достигает своего максимума.

3. Выбор формы связи.

При выборе формы связи между результирующим признаком и факторами начинают с наиболее простой зависимости – линейной.

Если линейная модель множественной регрессии неадекватно отражает исследуемое явление или процесс, то лучшее приближение могут дать нелинейные уравнения. Они, в свою очередь, могут быть нелинейными только по факторам, но линейными по параметрам, либо нелинейными как по параметрам, так и по факторам.

Например, уравнение, нелинейное только по факторам, имеет вид

$$\tilde{y} = b_0 + b_1 x_1^2 + b_2 x_2^{0,5} + b_3 x_3^{-2}.$$

Если ввести в рассмотрение новые переменные

$$x_1' = x_1^2,\ x_2' = x_2^{0,5},\ x_3' = x_3^{-2},$$

то получим уже линейное уравнение с новыми переменными

$$\widetilde{y} = b_0 + b_1 x_1' + b_2 x_2' + b_3 x_3',$$

для получения оценок которого в случае выполнения всех предпосылок используется обычный МНК.

Модели второго типа, например, функцию спроса

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \cdot \varepsilon$$

после предварительного логарифмирования

$$\ln y = \ln \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \ln \varepsilon$$

можно линеаризовать

$$y' = \beta_0' + \beta_1 x_1' + \beta_2 x_2' + \varepsilon',$$

путём введения соответствующих новых переменных.

Измерение тесноты связи переменной y с факторами для моделей первого типа осуществляется с помощью индекса корреляции

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

или индекса детерминации $R^2 = (R)^2$.

При рассмотрении альтернативных нелинейных моделей такого типа с одной функциональной формой зависимой переменной y процедура выбора модели проста: по значениям R^2.

Если же модели второго типа, то такой подход неправомерен. В этом случае используют остаточную сумму $\sum_{i=1}^{n} e_i^2$: чем она меньше, тем точнее модель.

Однако если модели используют разные функциональные формы y , то проблема выбора формы усложняется: нельзя непосредственно сравнивать коэффициенты детерминации R^2 или остаточные суммы $\sum_{i=1}^{n} e_i^2$.

Например, если в одной модели в левой части уравнения стоит y , а в другой модели - $\log y$, то такое сравнение бессмысленно по суммам $\sum_{i=1}^{n} e_i^2$.

В этом случае можно использовать стандартную процедуру, известную под названием *теста Бокса-Кокса* (или его частный случай тест Зарембки), который заключается в следующем.

1.Вычисляется среднее геометрическое $y_г$ в выборке.

2. Пересчитываются наблюдения зависимой переменной по формуле:

$$y_i^* = \frac{y_i}{y_г}.$$

3. Оценивается регрессия для линейной модели с использованием y_i^* вместо y_i и в логарифмической модели с заменой $\log(y_i)$ на $\log(y_i^*)$. Теперь остаточные суммы $\sum_{i=1}^{n} e_i^2$ сравнимы и модель с меньшей суммой является более точной.

4. Для того чтобы проверить не является ли одна из моделей значимо лучше, используют статистику

$$\chi^2 = \frac{n}{2}\log z,$$

где n – число наблюдений, z – отношение остаточных сумм $\sum_{i=1}^{n} e_i^2$ в пересчитанных регрессиях.

Эта статистика имеет χ^2-распределение с числом степеней свободы $k = n - 1$. Если

$$\chi^2 > \chi^2_{кр},$$

то имеется значимая разница в качестве моделей.