УДК 330.43:338

ВЫБОР ЭКЗОГЕННЫХ ФАКТОРОВ В МОДЕЛЬ РЕГРЕССИИ ПРИ МУЛЬТИКОЛЛИНЕАРНОСТИ ДАННЫХ

Орлова И.В., Филонова Е.С.

Финансовый университет при Правительстве Российской Федерации (Финуниверситет), Москва, e-mail: IVOrlova@gmail.com

В работе рассмотрен подход к построению модели множественной регрессии в условиях мультиколлинеарности данных. Рассмотрены типы мультиколлинеарности, причины возникновения и методы уменьшения или устранения ее. Данная статья посвящена рассмотрению и комбинированию различных методик отбора информативных факторов для регрессионного анализа. Предлагаемый подход реализован на основе реальных данных финансовой отчетности предприятий отрасли «Связь». Источник данных http://www.fira.ru/. В ходе решения поставленной задачи подробно показано использование теста Фаррара-Глоубера, пошаговой процедуры отбора факторов в модель, использование теста выбора «короткой» или «длинной» регрессии. Полученная модель была протестирована на мультиколлинеарность с помощью метода дополнительных регрессий (Тест VIF или метод инфляционных факторов). Результат тестирования показал, что полученная модель не содержит коллинеарных факторов и может быть использована для анализа и прогнозирования.

Ключевые слова: моделирование, мультиколлинеарность, кластерный анализ, модель множественной регрессии, тест Фаррара-Глоубера, метод инфляционных факторов

THE CHOICE OF EXOGENOUS FACTORS IN THE REGRESSION MODEL WITH MULTICOLLINEARITY IN THE DATA

Orlova I.V., Filonova E.S.

Financial university under the Government of the Russian Federation, Moscow, e-mail: IVOrlova@gmail.com

The paper considers an approach to the construction of a multiple regression model in terms of multicollinearity data. The types of multicollinearity causes and methods to reduce or eliminate it. This article is devoted to consideration and combining different methods of selection of informative factors for regression analysis. The proposed approach is implemented on the basis of real data of financial reporting industry «Communication». Data source http://www.fira.ru/. In the course of solving this problem is shown in detail the use of test Gloubera Farrar, step by step procedure for the selection factors in the model, the use of the test of choice «short» or «long» regression. The resulting model was tested for multicollinearity using the method of additional regressions (VIF test method or inflation factors). Test results showed that the resulting model contains no collinear factors and can be used for analysis and forecasting.

Keywords: modeling, multicollinearity, cluster analysis, multiple regression model, the test Farrar-Gloubera method inflationary factors

Регрессионный анализ экономических переменных составляет основу эконометрических исследований. Хорошее уравнение регрессии может дать экономисту много важной информации об интересующем его экономическом объекте, а также даёт возможность прогнозирования показателей в зависимости от значений, влияющих на них факторов.

В регрессионном анализе многое зависит от выбора наиболее подходящего вида уравнения: линейное оно или нелинейное; если нелинейное, то какого именно вида. Но все же определяющим этапом исследования, предшествующим регрессионному анализу, являются процедуры, направленные на выбор факторов для уравнения регрессии. Рассмотрению и комбинированию различных методик отбора информативных факторов для регрессионного анализа посвящена данная статья.

При построении модели регрессии к факторам, включаемым в модель, предъявляется ряд требований:

1) каждый фактор должен быть обоснован теоретически;

2) в модель включаются только те факторы, которые могут быть количественно измерены или отождествлены с цифровыми метками (представлены в виде фиктивных переменных);

3) в модель нельзя включать совокупный фактор и факторы, его образующие;

4) факторы должны быть тесно связаны с исследуемой переменной;

5) факторы должны быть линейно независимы друг от друга.

Три последних требования можно коротко сформулировать так: факторы, включаемые в модель регрессии, должны быть тесно связаны с исследуемой переменной и слабо – с другими факторами модели.

При построении множественной линейной регрессии часто сталкиваются с наличием линейной или близкой к ней связи между всеми или некоторыми объясняющими переменными. Это явление называется

мультиколлинеарностью. В математической статистике этот термин используется для обозначения тесной корреляционной взаимосвязи между отбираемыми для анализа факторами, совместно воздействующими на общий результат. Эта связь затрудняет оценивание параметров регрессии. Впервые на проблему мультиколлинеарности обратил внимание норвежский математик Нобелевский лауреат Рагнар Фриш [6].

В данной работе мы рассмотрим некоторые подходы к решению проблемы мультиколлинеарности экзогенных (объясняющих) факторов.

Мультиколлинеарность объясняющих переменных вызывает уменьшение точности оценивания или даже невозможность оценки влияния тех или иных переменных на результирующую переменную. Причина заключается в том, что вариации в исходных данных перестают быть независимыми и поэтому невозможно выделить влияние каждой независимой переменной в отдельности на зависимую переменную. Продемонстрируем это на простом примере. Пусть исследуется зависимость себестоимости от объема производства и введенных в действие основных фондов. Следует ожидать, что объем производства зависит также от основных фондов. Если мы обе переменные выберем в качестве объясняющих, то, очевидно, коэффициенты регрессии не будут точно отражать зависимость себестоимости от обоих факторов, так как основные фонды оказывают дополнительное влияние на себестоимость через объем производства.

Виды мультиколлинеарности: строгая (perfect) мультиколлинеарность – наличие линейной функциональной связи между независимыми переменными и нестрогая (imperfect) мультиколлинеарность – наличие сильной линейной корреляционной связи между независимыми переменными. Заметим, что корреляционные связи есть всегда. Проблема мультиколлинеарности – проблема силы проявления корреляционных связей.

Строгая мультиколлинеарность нарушает одно из основных условий теоремы Гаусса-Маркова [6] и делает построение регрессии невозможным. Нестрогая мультиколлинеарность затрудняет работу, но не препятствует получению правильных выводов.

Каковы основные причины возникновения мультиколлинеарности?

1. Ошибочное включение в уравнение двух или более линейно зависимых переменных.

2. В модели использованы факторные признаки, являющиеся составными элементами друг друга.

3. Исходные данные представлены временными рядами, имеющими одинаковые тенденции.

Мультиколлинеарность может проявляться и при отсутствии явных парных корреляционных зависимостей между переменными, так как мультиколлинеарность – ситуация линейной зависимости между объясняющими переменными. Однако вовсе необязательно эта зависимость должна быть парной.

При наличии мультиколлинеарности значимость отдельных коэффициентов регрессии уменьшается, так как стандартные ошибки становятся больше, что приводит к меньшей надежности полученных оценок.

Уменьшение значений t-статистики может выражаться в неверном с содержательной точки зрения знаке коэффициента регрессии. При мультиколлинеарности коэффициенты становятся неустойчивыми, поскольку в этом случае сложно отделить влияние одной переменной от влияния другой. Наличие доминантной переменной (коррелированной с зависимой переменной) делает коэффициенты при остальных объясняющих переменных незначимыми.

Выделим наиболее характерные признаки мультиколлинеарности:

1. Оценки становятся очень чувствительными к незначительному изменению результатов наблюдений и объёма выборки.

2. Оценки имеют большие стандартные ошибки, малую значимость, в то время как модель в целом является значимой и обладает хорошей объясняющей способностью (хорошие значения F-статистики и коэффициента детерминации R^2).

3. Оценки коэффициентов имеют неправильные с точки зрения теории (и логики) знаки или неоправданно большие значения. Коэффициенты, которые по логике должны быть значимы, оказываются незначимыми.

Для устранения или уменьшения мультиколлинеарности используется ряд методов.

Самый простой из них – исключение одной или нескольких переменных. При этом, какую переменную оставить, а какую удалить из анализа, решают в первую очередь на основании содержательных соображений. Если с экономической точки зрения ни одной из переменных нельзя отдать предпочтение, то оставляют ту из двух переменных, которая имеет больший коэффициент корреляции с зависимой переменной.

Другой метод устранения или уменьшения мультиколлинеарности заключается в переходе от несмещённых оценок, определённых по методу наименьших квадратов, к смещённым оценкам, обладающим, однако, меньшим рассеянием относительно оцениваемого параметра. Например,

при использовании «ридж-регрессии» (или «гребневой регрессии») добавление τ (τ – некоторое положительное число, называемое «гребнем» или «хребтом») к диагональным элементам матрицы X`X делает оценки параметров модели смещёнными, но при этом увеличивается определитель матрицы системы нормальных уравнений. Таким образом, становится возможным исключение мультиколлинеарности в случае, когда определитель |X`X| близок к нулю.

Для устранения мультиколлинеарности может быть использован переход от исходных объясняющих переменных $X_1, X_2, \ldots, X_n$, связанных между собой достаточно тесной корреляционной зависимостью, к новым переменным, представляющим линейные комбинации исходных. В качестве таких переменных берут, например, так называемые главные компоненты вектора исходных объясняющих переменных и рассматривают регрессию на главных компонентах, в которой последние выступают в качестве обобщённых объясняющих переменных, подлежащих в дальнейшем содержательной (экономической) интерпретации.

Можно применить преобразование мультиколлинеарных переменных:

– использовать нелинейные формы;

– использовать агрегаты (линейные комбинации нескольких переменных);

– использовать первые разности вместо самих переменных.

Рассмотрим конкретный набор данных и поставим задачу определения совокупности экзогенных переменных, наиболее подходящих для регрессионного анализа.

В качестве исходных данных будем использовать основные показатели баланса компаний-эмитентов, относящихся к сфере деятельности «Связь»[1] (всего 122 компании). Анализу показателей этой отрасли посвящены работы [8], [10].

В качестве результирующей (эндогенной) переменной выбираем показатель *чистая прибыль (убыток)* компаний.

В силу неоднородности объектов (компаний) была проведена их классификация на однородные группы методами кластерного анализа [2], [5]. В результате получены три кластера со следующими профилями (Кластерные профили – это, по сути, средние значения группирующих переменных в кластере):

1. Первый кластер самый большой – 109 компаний. Среди них много убыточных. Все показатели баланса предприятий этой группы на несколько порядков ниже, чем у других.

2. Второй кластер – 10 крупных прибыльных компаний.

3. Третий кластер – наименьший по количеству компаний, их всего три – Вымпел-Коммуникации (Билайн), Мобильные ТелеСистемы (МТС), Мегафон, но это самые крупные и самые прибыльные компании.

Объективную картину количественных взаимосвязей финансовых показателей предприятий данной отрасли можно получить, используя данные только о компаниях самого крупного первого кластера.

Краткая экономическая характеристика этих показателей:

– валюта баланса (ВБ) – это итог по всем счетам бухгалтерского баланса, сумма всех активов или всех пассивов компании;

– выручка (нетто) от продаж (ВП), дебиторская задолженность (краткосрочная) (ДЗ), запасы готовой продукции и товаров для перепродажи (ЗП), оборотные активы (ОА), основные средства (ОС), прибыль (убыток) от продаж (ПП), чистая прибыль (убыток) (ЧП) – это представители группы активов компании (здесь присутствуют как оборотные, так и внеоборотные активы);

– долгосрочные обязательства (ДО), краткосрочные обязательства (КО) – представители группы пассивов.

Схема отбора экзогенных факторов при решении данной задачи будет выглядеть следующим образом:

I. Корреляционный анализ данных, включая проверку теста Фаррара-Глоубера на мультиколлинеарность факторов;

II. Пошаговый отбор факторов методом исключения из модели статистически незначимых переменных;

III. При несоответствии результатов, полученных в пунктах 1) и 2), проверка теста на «длинную» и «короткую» регрессию.

I. Корреляционный анализ данных, включая проверку теста Фаррара-Глоубера [11] на мультиколлинеарность факторов

В табл. 1 представлена матрица коэффициентов парной корреляции для всех переменных, участвующих в рассмотрении.

Визуальный анализ матрицы позволяет установить:

- ЧП имеет довольно высокие парные корреляции со всеми переменными, кроме переменной ЗП (запасы готовой продукции и товаров для перепродажи) далее ее не будем рассматривать, что вполне объяснимо, так как предприятия отрасли «Связь» имеют специфическую продукцию;
- большинство переменных анализа демонстрируют довольно высокие парные корреляции, что обуславливает необходимость проверки факторов на наличие между ними мультиколлинеарности.

[1] Источник данных – http://www.fira.ru/.

Таблица 1

Матрица коэффициентов парной корреляции

	ВП	ДЗ	ДО	ЗП	КО	ОА	ОС	ПП	ЧП
ВП	1								
ДЗ	0,703	1							
ДО	0,619	0,711	1						
ЗП	0,207	0,214	0,191	1					
КО	0,872	0,766	0,761	0,261	1				
ОА	0,628	0,909	0,687	0,216	0,687	1			
ОС	0,885	0,658	0,632	0,113	0,760	0,560	1		
ПП	0,937	0,625	0,626	0,114	0,796	0,538	0,845	1	
ЧП	0,848	0,567	0,642	0,126	0,776	0,529	0,723	0,896	1

Таблица 2

Матрица межфакторных корреляций R

Переменная	ВП	ДЗ	ДО	КО	ОА	ОС	ПП
ВП	1,00	0,70	0,62	0,87	0,63	0,89	0,94
ДЗ	0,70	1,00	0,71	0,77	0,91	0,66	0,62
ДО	0,62	0,71	1,00	0,76	0,69	0,63	0,63
КО	0,87	0,77	0,76	1,00	0,69	0,76	0,80
ОА	0,63	0,91	0,69	0,69	1,00	0,56	0,54
ОС	0,89	0,66	0,63	0,76	0,56	1,00	0,85
ПП	0,94	0,62	0,63	0,80	0,54	0,85	1,00

Таблица 3

Обратная матрица R^{-1}

21,37	1,10	4,33	– 7,73	– 3,10	– 5,52	– 10,95
1,10	7,78	0,14	– 1,86	– 5,67	– 1,30	– 0,35
4,33	0,14	3,61	– 2,85	– 1,40	– 1,39	– 2,21
– 7,73	– 1,86	– 2,85	7,49	1,28	1,63	2,16
– 3,10	– 5,67	– 1,40	1,28	6,55	1,29	1,68
– 5,52	– 1,30	– 1,39	1,63	1,29	5,42	0,28
– 10,95	– 0,35	– 2,21	2,16	1,68	0,28	10,00

Таблица 4

Значения F-критериев

F1 (ВП)	F2 (ДЗ)	F3 (ДО)	F4 (КО)	F5 (ОА)	F6 (ОС)	F7 (ПП)
293,966	97,812	37,709	93,621	80,047	63,808	129,863

Таблица 5

Матрица коэффициентов частных корреляций $R_{\text{частные}}$

Переменная	ВП	ДЗ	ДО	КО	ОА	ОС	ПП
ВП							
ДЗ	– 0,09						
ДО	– 0,49	– 0,03					
КО	0,61	0,24	0,55				
ОА	0,26	0,79	0,29	– 0,18			
ОС	0,51	0,20	0,31	– 0,26	– 0,22		
ПП	0,75	0,04	0,37	– 0,25	– 0,21	– 0,04	

Для выявления мультиколлинеарности факторов выполним тест Фаррара-Глоубера по факторам ВП, ДЗ, ДО, КО, ОА, ОС, ПП, используя межфакторные корреляционные связи.

Проверка теста Фаррара-Глоубера на мультиколлинеарность факторов включает несколько этапов [6], [9] реализация которых представлена ниже.

1. *Проверка наличия мультиколлинеарности всего массива переменных*

● Найдем определитель матрицы межфакторных корреляций (табл. 2) R (det[R] = 0,0001) с помощью функции *МОПРЕД*. Определитель матрицы R близок к нулю, что позволяет сделать предположение об общей мультиколлинеарности факторов. Подтвердим это предположение оценкой статистики Фаррара-Глоубера.

● Вычислим наблюдаемое значение статистики Фаррара – Глоубера по формуле:

$$FG = -\left[n - 1 - \frac{1}{6}\cdot\left(2\cdot k + 5\right)\right]\cdot \ln\left(\det\left[R\right]\right),$$

где n = 109 – количество наблюдений (компаний); k = 7 – количество факторов (переменных анализа).

$$FG = -\left[109 - 1 - \frac{1}{6}\cdot\left(2\cdot 7 + 5\right)\right]\times$$

$$\times \ln\left(0,0001\right) = 953,87.$$

Фактическое значение этого критерия FG сравниваем с табличным значением критерия χ^2 с $\frac{1}{2}\cdot k\cdot\left(k-1\right) = \frac{1}{2}\cdot 7\cdot\left(7-1\right) = 21$ степенью свободы и уровне значимости α = 0,05. Табличное значение χ^2 = 32,64 можно найти с помощью функции *ХИ2ОБР*.

Так как $FG > \chi^2$ (953,87 > 32,67), то в массиве объясняющих переменных существует мультиколлинеарность.

2. *Проверка наличия мультиколлинеарности каждой переменной с другими переменными*.

Вычислим обратную матрицу R^{-1} с помощью функции Excel *МОБР* (табл. 3), диагональные элементы которой назовем c_{jj}.

Вычисленные F-критерии

$$F_j = \left(c_{jj}\right)\frac{n-k-1}{k},$$

где c_{jj} – диагональные элементы матрицы R^{-1} приведены в (табл. 4).

Фактические значения F-критериев сравниваются с табличным значением $F_{\text{табл}} = 2,1$ при $v_1 = 7$ и $v_2 = n - k - 1 = 109 - 7 - 1 = 101$ степенях свободы и уровне значимости α = 0,05, где k – количество факторов. Так как все значения F-критериев больше табличного, то все исследуемые независимые переменные мультиколлинеарны с другими. Больше других влияет на общую мультиколлинеарность факторов фактор ВП – выручка (нетто) от продаж, меньше – фактор ДО – долгосрочные обязательства .

3. *Проверка наличия мультиколлинеарности каждой пары переменных*

Вычислим частные коэффициенты корреляции по формуле

$$r_{ij(\ )} = \frac{-c_{ij}}{\sqrt{c_{ii}\cdot c_{jj}}},$$

где c_{jj} – элементы матрицы R^{-1}. Матрицу коэффициентов частной корреляции $R_{\text{частные}}$, можно получить с помощью программ *VSTAT, SPSS* (табл. 5).

В табл. 6 серым цветом выделены значения t-критерия, которые меньше табличного значения, а в табл. 5 –соответствующие им статистически незначимые коэффициенты частной корреляции. Жирным шрифтом цветом выделены значения t-критерия, которые больше табличного значения, и соответствующие им статистически значимые коэффициенты частной корреляции.

Вычисление t-критериев по формуле $t_{ij} = \frac{r_{ij(\ )}\sqrt{n-k-1}}{\sqrt{1-r^2_{ij(\ )}}}$ (табл. 6).

Таблица 6

t-критерии для коэффициентов частной корреляции

Переменная	ВП	ДЗ	ДО	КО	ОА	ОС	ПП
ВП							
ДЗ	– 0,86						
ДО	– 5,69	– 0,26					
КО	7,75	2,52	6,59				
ОА	2,73	13,12	3,02	– 1,87			
ОС	6,01	2,05	3,32	– 2,66	– 2,24		
ПП	11,35	0,40	3,97	– 2,60	– 2,14	– 0,38	

Фактические значения t-критериев сравниваются с табличным значением $t_{табл} = 1{,}98$ при степенях свободы $(n - k - 1) = 109 - 7 - 1 = 101$ и уровне значимости $\alpha = 0{,}05$.

Из табл. 6 и 7 видно, что две пары факторов *ОА* и *ДЗ*, *ПП* и *ВП* имеют высокую статистически значимую частную корреляцию, то есть являются мультиколлинеарными. Для того, чтобы избавиться от мультиколлинеарности, можно исключить одну из переменных коллинеарной пары. В паре *ПП* и *ВП* оставляем *ПП,* так как у нее меньше связи с другими факторами; в паре *ОА* и *ДЗ* оставим *ОА,* во-первых, с экономической точки зрения, а, во-вторых, так как у нее меньше значение F-критерия и, значит, она меньше влияет на общую мультиколлинеарность факторов.

Таким образом, в результате проверки теста Фаррара-Глоубера остается пять факторов: *ДО, КО, ОА, ОС, ПП.*

Завершая процедуры корреляционного анализа, целесообразно посмотреть частные корреляции выбранных факторов с результатом *ЧП. В* последнем столбце табл. 7 представлены значения t-критерия для столбца *ЧП.*

Из табл. 7 видно, что межфакторные частные корреляции слабые, а переменная *ЧП* имеет высокую и одновременно статистически значимую частную корреляцию только с фактором *ПП* (соответствующие значения в табл. 7 выделены жирным шрифтом).

Уточнение набора факторов, наиболее подходящих для регрессионного анализа, осуществим другими методами отбора.

II. Пошаговый отбор факторов методом исключения из модели статистически незначимых переменных

В соответствии с общим подходом, пошаговый отбор следует начинать с включения в модель всех имеющихся факторов, то есть в нашем случае с восьмифакторной регрессии. Но мы не будем включать в модель факторы из заранее известных коллинеарных пар (в связи с наличием коллинеарности ранее были исключены из рассмотрения *ВП* и *ДЗ*), а также фактор *ЗП*, имеющий слабую связь с *ЧП.* Таким образом, пошаговый отбор факторов начнем с пятифакторного уравнения. Фрагмент протокола пятифакторного регрессионного анализа представлен в табл. 8.

Статистически незначимыми ($t_{табл} < \left|t_{a_j}\right|$) оказались три фактора (в табл. 8 они выделены жирным шрифтом). На следующем этапе пошагового отбора удаляем статистически незначимый фактор с наименьшим значением t-критерия, то есть фактор *ОА* (в табл. 8 выделен цветом).

Аналогично поступаем до тех пор, пока не получим уравнение, в котором все факторы окажутся статистически значимыми. Этапы получения такого уравнения, то есть фрагменты протоколов соответствующих регрессионных анализов, представлены в табл. 9 и 10.

Таблица 7

Матрица коэффициентов частной корреляции с результатом *ЧП*

Переменная	ДО	КО	ОА	ОС	ПП	ЧП	*t*-критерий $t_{табл} = (0{,}05;\ 102) = 1{,}98$
ДО	1,00	0,34	0,34	0,12	– 0,12	0,16	1,63
КО	0,34	1,00	0,28	0,17	0,15	0,17	1,75
ОА	0,34	0,28	1,00	0,07	– 0,04	– 0,02	– 0,24
ОС	0,12	0,17	0,07	1,00	0,59	– 0,24	– 2,49
ПП	– 0,12	0,15	– 0,04	0,59	1,00	0,71	10,27
ЧП	0,16	0,17	– 0,02	– 0,24	0,71	1,00	

Таблица 8

Фрагмент протокола пятифакторного регрессионного анализа

		$t_{табл}(0{,}05;\ 109 - 5 - 1 = 103) =$ **1,98326409**	
	Коэффициенты	Стандартная ошибка	*t*-статистика
Y– пересечение	– 2067,779334	16246,6282	– 0,127274368
ОС	– 0,040553788	0,016198212	– 2,503596652
ПП	0,649466697	0,062951463	10,31694366
ДО	0,033862469	0,02067002	1,638240731
КО	0,049965808	0,028431981	1,75738047
ОА	– 0,006074787	0,025402164	– 0,239144461

Таблица 9

Фрагмент протокола четырехфакторного регрессионного анализа

		$t_{табл}$(0,05; 109 – 4 – 1 = 104) = **1,983037471**	
	Коэффициенты	Стандартная ошибка	*t*-статистика
Y-пересечение	– 3255,832024	15398,16512	– 0,211442857
ОС	– 0,040859333	0,016074384	– 2,541891019
ПП	0,650673211	0,062463899	10,41678825
ДО	0,032173752	0,019338145	1,663745481
КО	0,048029464	0,027130844	1,770290058

Таблица 10

Фрагмент протокола трехфакторного регрессионного анализа

		$t_{табл}$(0,05; 109 – 3 – 1 = 105) = **1.982815217**	
	Коэффициенты	Стандартная ошибка	*t*-статистика
Y-пересечение	– 4456,711199	15510,19708	– 0,28734072
ОС	– 0,037629315	0,016090498	– 2,338604743
ПП	0,647303561	0,062954486	10,28208794
КО	0,071691944	0,023297943	3,07717916

Из табл. 10 видно, что уравнение с тремя факторами *ОС, ПП* и *КО* обладает статистически значимыми коэффициентами перед факторами (в нем незначим только свободный член), а, значит, и сами эти факторы статистически значимы.

Таким образом, в результате пошагового отбора получено трехфакторное уравнение регрессии, все коэффициенты которого (кроме свободного члена) значимы при 5 %-ном уровне значимости, вида

$$Y = –4456{,}7 – 0{,}038 \cdot X_1 + \\ + 0{,}647 \cdot X_2 + 0{,}072 \cdot X_3,$$

где *Y – ЧП, – ОС, – ПП, – КО.*

III. Проверка теста на «длинную» и «короткую» регрессии

По результатам пунктов 1) и 2) возникает необходимость выбора из двух регрессий: «длинной» – с тремя факторами (*ОС, ПП* и *КО*) и «короткой» – с одним фактором (*ПП*).

Воспользуемся тестом на «длинную» и «короткую» регрессии [6]. Этот тест используется для отбора наиболее существенных объясняющих переменных. Иногда переход от большего числа исходных показателей анализируемой системы к меньшему числу наиболее информативных факторов может быть объяснен дублированием информации, из-за сильно взаимосвязанных факторов. Стремление к построению более простой модели приводит к идее уменьшения размерности модели без потери её качества. Для этого используют тест проверки «длинной» и «короткой» регрессий.

Рассмотрим две модели регрессии:

$yi = \beta 0 + \beta 1\ xi1 + \ldots + \beta k\ xik + \varepsilon\ i$ (длинную),

$yi = \beta 0 + \beta 1\ xi1 + \ldots + \beta k\ xik\text{-}q + \varepsilon i$ (короткую).

Предположим, что модель не зависит от последних q объясняющих переменных и их можно исключить из модели. Это соответствует гипотезе

$$H0\text{: } \beta k\text{-}q+1 = \beta k\text{-}q + 2 \ldots = \beta k = 0,$$

т.е. последние q коэффициентов β_i равны нулю.

Алгоритм проверки следующий:

– Построить по МНК длинную регрессию по всем факторам $X_1, \ldots, X_k$ и найти для неё сумму квадратов остатков – $ESS_{длин}$.

– Построить по МНК короткую регрессию по первым $(k – q)$ факторам $X_1, \ldots, X_{k–q}$ и найти для неё сумму квадратов остатков – $ESS_{кор}$.

– Вычислить *F*-статистику

$$F_{набл} = \frac{\left(ESS_{кор} - ESS_{длин}\right)/q}{ESS_{длин}/\left(n-k-1\right)}.$$

– Если $F_{набл} > F_{табл}$ $(\alpha,\ v1 = q,\ v2 = n – k – 1)$, гипотеза отвергается (выбираем длинную регрессию), в противном случае – выбираем короткую регрессию.

На основании данных нашего примера сравним две модели: «длинную» (с факторами X_1, X_2, X_3) и «короткую» (только с фактором X_2).

1. Построим длинную регрессию по трем факторам X_1, X_2, X_3 и найдем для неё сумму квадратов остатков – $ESS_{длин}$ (табл. 11).

Таблица 11

Фрагмент протокола регрессионного анализа для длинной (трехфакторной) регрессии

Дисперсионный анализ			
	df	SS	MS
Регрессия	3	1,04794E+13	3,49313E+12
Остаток	105	2,25564E+12	21482327265
Итого	108	1,2735E+13	
	Коэффициенты	Стандартная ошибка	*t*-статистика
Y-пересечение	– 4456,711199	15510,19708	– 0,28734072
ОС	– 0,037629315	0,016090498	– 2,338604743
ПП	0,647303561	0,062954486	10,28208794
КО	0,071691944	0,023297943	3,07717916

2. Построим короткую регрессию по одному фактору X_2 и найдем для неё сумму квадратов остатков – $ESS_{\text{кор}}$ (табл. 12).

Таблица 12

Фрагмент протокола регрессионного анализа для короткой (однофакторной) регрессии

Дисперсионный анализ			
	df	SS	MS
Регрессия	1	1,02234E+13	1,02234E+13
Остаток	107	2,51168E+12	23473610976
Итого	108	1,2735E+13	
	Коэффициенты	Стандартная ошибка	*t*-статистика
Y-пересечение	1286,42961	15643,62168	0,08223349
ПП	0,658080318	0,031533476	20,86925995

3. Вычислим F-статистику

$$F = \frac{\left(ESS_{\text{кор}} - ESS_{\text{длин}}\right)/q}{ESS_{\text{длин}}/\left(n-k-1\right)} = \frac{\left(2{,}51168\,E+12-2{,}25564\,E+12\right)/2}{2{,}25564E+12/\left(109-3-1\right)} = 5{,}959,$$

$F_{\text{табл}}(0{,}05;\ 109-3-1=105) = 3{,}083$.

4. Так как $F > F_{\text{табл}}$, отдаем предпочтение длинной регрессии

$$Y = -4456{,}7 - 0{,}038\cdot X_1 + \\ + 0{,}647\cdot X_2 + 0{,}072\cdot X_3.$$

Набор факторов в этой модели вполне соответствует результатам проверки теста Фаррара-Глоубера.

Протестируем полученную модель на мультиколлинеарность с помощью метода дополнительных регрессий (Тест VIF или метод инфляционных факторов).

Для измерения эффекта мультиколлинеарности используется показатель VIF – «фактор инфляции вариации»:

$$VIF_{x_j} = \frac{1}{\left(1 - R^2_{x_j x_1 \ldots x_{j-1} x_{j+1} \ldots x_p}\right)},$$

где $R^2_{X_j X_1 \ldots X_{j-1} X_{j+1} \ldots X_k}$ – это значение коэффициента множественной детерминации, полученное для регрессора X_j как зависимой переменной и остальных переменных. При этом степень мультиколлинеарности, представляемая в регрессии переменной X_j, когда все переменные X включены в регрессию, есть функция множественной корреляции между X_j и другими переменными X.

Если $VIF_{x_j} > 10$, то считается, что данный регрессор приводит к мультиколлинеарности.

Расчеты были выполнены с помощью программы Gretl, хотя сложность расчетов невелика и это вполне можно было вычислить в Excel. В нашем случае мы получили следующие VIF значения:

$$VIF_{x_1} = 3{,}783,$$

$$VIF_{x_2} = 4,355,$$

$$VIF_{x_3} = 2,944.$$

Так как все VIF значения меньше 10, то можно сделать вывод, что построенная модель $Y = -4456,7 - 0,038 \cdot X_1 + 0,647 \cdot X_2 + 0,072 \cdot X_3$ не содержит коллинеарных факторов и может быть использована для анализа и прогнозирования. Тем самым, поставленная задача выбора факторов в модель регрессии решена.

Список литературы

1. Гусарова О.М. Моделирование результатов бизнеса в менеджменте организации // В сборнике: Перспективы развития науки и образования сборник научных трудов по материалам Международной научно-практической конференции. – Тамбов, 2014. – С. 42–43.

2. Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: Учебное пособие / Под ред. И.В. Орловой / Орлова И.В., Концевая Н.В., Турундаевский В.Б., Уродовских В.Н., Филонова Е.С. – М.: Вузовский учебник, 2009.

3. Орлова И.В. Линейная алгебра и аналитическая геометрия для экономистов: учебник и практикум для прикладного бакалавриата / И.В. Орлова, В.В. Угрозов, Е.С. Филонова. – М.: Издательство Юрайт, 2014 – 370 с. – Серия: Бакалавр. Прикладной курс.

4. Орлова И.В. Экономико-математическое моделирование: Практическое пособие по решению задач. – 2-е издание, испр. и доп. – М.: Вузовский учебник: ИНФРА-М, 2012.

5. Орлова И.В., Концевая Н.В., Турундаевский В.Б., Уродовских В.Н., Филонова Е.С. Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS (учебное пособие) // Международный журнал прикладных и фундаментальных исследований. – 2014. – № 3–2. – С. 248–250.

6. Орлова И.В., Половников В.А. Экономико-математические методы и модели: компьютерное моделирование / учебное пособие для студентов высших учебных заведений, обучающихся по специальности «Статистика» и другим экономическим специальностям. – М., 2011. Сер. Вузовский учебник (3-е издание, переработанное и дополненное).

7. Орлова И.В., Половников В.А., Филонова Е.С., Гусарова О.М., Малашенко В.М., Дайитбегов Д.М. Эконометрика. Учебно-методическое пособие. – М., 2010.

8. Орлова И.В., Филонова Е.С. Эконометрическое моделирование финансовой эффективности предприятий, относящихся к виду экономической деятельности «Связь» // Международный бухгалтерский учет. – 2012. – № 43. – С. 22–24.

9. Орлова И.В., Филонова Е.С., Агеев А.В. Эконометрика. Компьютерный практикум для студентов третьего курса, обучающихся по специальностям 080105.65 «Финансы и кредит», 080109.65 «Бухгалтерский учет, анализ и аудит». – М., 2011.

10. Филонова Е.С. Состояние и тенденции финансовой эффективности предприятий отрасли (на примере предприятий отрасли «СВЯЗЬ»). В сборнике: Анализ современных проблем развития регионов Российской Федерации. – Орел, 2011. – С. 154–157.