ПОДХОДЫ К ПОСТРОЕНИЮ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ЭЛЕКТРОННОЙ ДИАГНОСТИКИ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ[1]

С.И. Сиваков,
ассистент кафедры
информационного менеджмента, НИУ «БелГУ»

При математическом моделировании сердечно-сосудистых заболеваний, причины развития которых не известны, а имеются только предположения влияния тех или иных факторов на появление болезни, используется корреляционный анализ возможных факторов развития. Если выявленные на основе корреляционного анализа связи между изучаемыми факторами окажутся существенными (т.е. достаточно сильными и статистически значимыми), то целесообразно найти их математическое выражение в виде регрессионной модели и оценить ее адекватность. Адекватная регрессионная модель может использоваться для прогнозирования изучаемого явления или показателя.

При большом количестве исследуемых факторов применение обычного регрессионном анализа становится затруднительным. В таких ситуациях применяется пошаговая регрессия [1].

Цель пошаговой регрессии состоит в отборе из большого количества предикатов небольшой подгруппы переменных, которые вносят наибольший вклад в вариацию зависимой переменной. Обычно этот процесс выполняет автоматизированная процедура, которая вводит или выводит предикаты из уравнения регрессии по очереди, основываясь на серии F-тестов, t-тестов или других подходах [2].

Основные подходы:

– прямое включение (прямая пошаговая регрессия). Вначале уравнение регрессии не содержит предикатов. Они вводятся по одному, если удовлетворяют определенному критерию. В основе порядка введения включаемых переменных лежит вклад переменной в объясняемую вариацию;

– исключение переменной (обратная пошаговая регрессия). Вначале все предикаты входят в уравнение регрессии. Затем по очереди выводятся из уравнения исходя из их соответствия критерию;

– пошаговый подход. На каждой стадии прямое включение осуществляют одновременно с исключением переменных, которые больше не удовлетворяют конкретному критерию.

Часто применяют пошаговый подход, когда последовательно включаются факторы в уравнение регрессии и после проверяется их значимость. Факторы поочередно вводятся в уравнение так называемым «прямым методом». При проверке значимости введенного фактора определяется, насколько уменьшается сумма квадратов остатков и увеличивается величина множественного коэффициента корреляции. Одновременно используется и обратный метод, т.е. исключение факторов, ставших незначимыми на основе t-критерия Стьюдента. Фактор является незначимым, если его включение в уравнение регрессии только изменяет значение коэффициентов регрессии, не уменьшая значительно суммы квадратов остатков и не увеличивая их значения. Если при включении в модель соответствующего факторного признака величина множественного коэффициента корреляции увеличивается, а коэффициент регрессии не изменяется (или меняется несущественно), то данный признак существен и его включение в уравнение регрессии необходимо [3].

Анализ качества эмпирического уравнения парной и множественной линейной регрессии начинаются построения эмпирического уравнения регрессии, которое является

[1] Исследование выполнено в рамках Государственного задания Министерства образования и науки РФ на выполнение НИР подведомственным вузам в 2013 году. Проект № 8.8600.2013.

начальным этапом эконометрического анализа. Первое же, построенное по выборке, уравнение регрессии очень редко является удовлетворительным по тем или иным характеристикам. Поэтому следующей важнейшей оценкой является проверка качества уравнения регрессии. В эконометрике принята устоявшаяся схема такой проверки, которая проводится по следующим направлениям:

- проверка статистической значимости коэффициентов уравнения регрессии;
- проверка общего качества уравнения регрессии;
- проверка свойств данных, выполнимость которых предполагалась при оценивании уравнения (проверка выполнимости предпосылок МНК).

Прежде, чем проводить анализ качества уравнения регрессии, необходимо определить дисперсии и стандартные ошибки коэффициентов, а также интервальные оценки коэффициентов. Корреляционный и регрессионный анализ, как правило, проводится для ограниченной по объёму совокупности. Поэтому параметры уравнения регрессии (показатели регрессии и корреляции), коэффициент корреляции и коэффициент детерминации могут быть искажены действием случайных факторов. Чтобы проверить, на сколько эти показатели характерны для всей генеральной совокупности, не являются ли они результатом стечения случайных обстоятельств, необходимо проверить адекватность построенных статистических моделей.

При анализе адекватности уравнения регрессии (модели) исследуемому процессу, возможны следующие варианты:

1. Построенная модель на основе F-критерия Фишера в целом адекватна и все коэффициенты регрессии значимы. Такая модель может быть использована для принятия решений и осуществления прогнозов.

2. Модель по F-критерию Фишера адекватна, но часть коэффициентовне значима. Модель пригодна для принятия некоторых решений, но не для прогнозов.

3. Модель по F-критерию адекватна, но все коэффициенты регрессиине значимы. Модель полностью считается неадекватной. На ее основе не принимаются решения и не осуществляются прогнозы.

Проверить значимость (качество) уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным, достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной. Чтобы иметь общее суждение о качестве модели, по каждому наблюдению из относительных отклонений определяют среднюю ошибку аппроксимации. Проверка адекватности уравнения регрессии (модели) осуществляется с помощью средней ошибки аппроксимации, величина которой не должна превышать 10-12% (рекомендовано).

$$\bar{\varepsilon} = \frac{1}{n}\sum \frac{|y_i - \tilde{y}_x|}{y_i} * 100\% \qquad (1)$$

Оценка значимости уравнения регрессии в целом производится на основе F-критерия Фишера, которому предшествует дисперсионный анализ. В математической статистике дисперсионный анализ рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели. Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной (y) от среднего значения ($y_{ср.}$) раскладывается на две части – «объясненную» и «необъясненную»:

$$\sum (y - \bar{y})^2 = \sum \left(\hat{y}_x - \bar{y}\right)^2 + \sum \left(y - \hat{y}_x\right)^2 \qquad (2)$$

Схема дисперсионного анализа представлена в таблице.

Таблица

Дисперсионный анализ (n –число наблюдений, m–число параметров при переменной x)

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum(y-\bar{y})^2$	$n-1$	$S^2_{общ}=\frac{\sum(y-\bar{y})^2}{n-1}$
Факторная	$\sum(\hat{y}_x-\bar{y})^2$	m	$S^2_{факт}=\frac{\sum(\hat{y}_x-\bar{y})^2}{m}$
Остаточная	$\sum(y-\hat{y}_x)^2$	$n-m-1$	$S^2_{ост}=\frac{\sum(y-\hat{y}_x)^2}{n-m-1}$

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F-критерия Фишера. Фактическое значение F - критерия Фишера сравнивается с табличным значением $F_{табл.}$ (α, k_1, k_2) при заданном уровне значимости α и степенях свободы k_1= m и k_2=n-m-1. При этом, если фактическое значение F-критерия больше табличного $F_{факт}>F_{теор}$, то признается статистическая значимость уравнения в целом. Для парной линейной регрессии m=1 , поэтому:

$$F=\frac{S^2_{факт}}{S^2_{ост}}=\frac{\sum(\hat{y}_x-\bar{y})^2}{\sum(y-\hat{y}_x)^2}\cdot(n-2) \quad (3)$$

Эта формула в общем виде может выглядеть так:

$$F=\frac{\sigma^2_{факт}(n-m)}{\sigma^2_{ост}(m-1)}\Rightarrow сравнить\ с\ F_{табл} \quad (4)$$

Отношение объясненной части дисперсии переменной (y) к общей дисперсии называют коэффициентом детерминации и используют для характеристики качества уравнения регрессии или соответствующей модели связи. Соотношение между объясненной и необъясненной частями общей дисперсии можно представить в альтернативном варианте:

$$R^2=\frac{\sum(\tilde{y}_x-\bar{y})^2}{\sum(y_i-\bar{y})^2}=1-\frac{\sum(y_i-\tilde{y}_x)^2}{\sum(y_i-\bar{y})^2} \quad (5)$$

Коэффициент детерминации R^2 принимает значения в диапазоне от нуля до единицы $0\leq R^2\leq 1$. Коэффициент детерминации R^2показывает, какая часть дисперсии результативного признака (y) объяснена уравнением регрессии. Чем больше R^2, тем большая часть дисперсии результативного признака (y) объясняется уравнением регрессии и тем лучше уравнение регрессии описывает исходные данные. При отсутствии зависимости между (y) и (x) коэффициент детерминации R^2 будет близок к нулю. Таким образом, коэффициент детерминации R^2 может применяться для оценки качества (точности) уравнения регрессии. Значение R-квадрата является индикатором степени подгонки модели к данным (значение R-квадрата близкое к 1.0 показывает, что модель объясняет почти всю изменчивость соответствующих переменных). Чтобы определить, при каких значениях R^2 уравнение регрессии следует считать статистически не значимым, что, в свою очередь, делает необоснованным его использование в анализе, рассчитывается F-критерий Фишера: $F_{факт}>F_{теор}$– делаем вывод о статистической значимости уравнения регрессии. Величина F-критерия связана с коэффициентом детерминации R^2_{xy} (r^2_{xy}) и ее можно рассчитать по следующей формуле:

$$F=\frac{r^2_{xy}}{1-r^2_{xy}}\cdot(n-2) \quad (6)$$

Либо при оценке значимости индекса детерминации (аналог коэффициента детерминации):

$$F=\frac{i^2(n-m)}{(1-i^2)(m-1)} \quad (7)$$

где: i^2– индекс (коэффициент) детерминации, который рассчитывается:

$$i^2(\eta^2)=1-\frac{\Sigma\left(y_i-\tilde{y}_i\right)^2}{\Sigma\left(y_i-\bar{y}\right)^2} \quad (8)$$

Использование коэффициента множественной детерминации R^2 для оценки качества модели, обладает тем недостатком, что включение в модель нового фактора (даже несущественного) автоматически увеличивает величину R^2. Поэтому, при большом количестве факторов, предпочтительнее использовать, так называемый, улучшенный, скорректированный коэффициент множественной детерминации R^2, определяемый соотношением:

$$\overline{R}^2=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2:(n-p-1)}{\sum_{i=1}^{n}(y_i-\bar{y})^2:(n-1)}=1-\frac{n-1}{n-p-1}(1-R^2) \quad (9)$$

где p – число факторов в уравнении регрессии, n – число наблюдений. Чем больше величина p, тем сильнее различия между множественным коэффициентом детерминации R^2и скорректированным R^2. При использовании скорректированного R^2, для оценки целесообразности включения фактора в уравнение регрессии, следует учитывать, что увеличение его величины (значения), при включении нового фактора, не обязательно свидетельствует о его значимости, так как значение увеличивается всегда, когда t-статистика больше единицы (|t|>1). При заданном объеме наблюдений и при прочих равных условиях, с увеличением числа независимых переменных (параметров), скорректированный коэффициент множественной детерминации убывает. При небольшом числе наблюдений, скорректированная величина коэффициента множественной детерминации R^2 имеет тенденцию переоценивать долю вариации результативного признака, связанную с влиянием факторов, включенных в регрессионную модель. Низкое значение коэффициента множественной корреляции и коэффициента множественной детерминации R^2 может быть обусловлено следующими причинами:

- в регрессионную модель не включены существенные факторы;
- неверно выбрана форма аналитической зависимости, которая нереально отражает соотношения между переменными, включенными в модель.

Следует также обратить внимание на важность анализа остатков (остаточной, «необъясненной» дисперсии). Остаток представляет собой отклонение фактического значения зависимой переменной от значения, полученного расчетным путем. При построении уравнения регрессии, мы можем разбить значение (y) в каждом наблюдении на 2 составляющие:

$$y_i=\tilde{y}_i+\varepsilon_i$$

Отсюда:

$$\varepsilon_i=y_i-\tilde{y}_i$$

Если $\varepsilon_i=0$, то для всех наблюдений фактические значения зависимой переменной совпадают с расчетными (теоретическими) значениями. Графически это означает, что теоретическая линия регрессии (линия, построенная по функции $y=a_0+a_1x$) проходит через все точки корреляционного поля, что возможно только при строго функциональной связи. Следовательно, результативный признак (*y)* полностью обусловлен влиянием фактора (*x).*На практике, как правило, имеет место некоторое рассеивание точек корреляционного поля относительно теоретической линии регрессии, т.е. отклонения эмпирических данных от теоретических $\varepsilon_i\neq0$. Величина этих отклонений и лежит в основе расчета показателей качества (адекватности) уравнения.

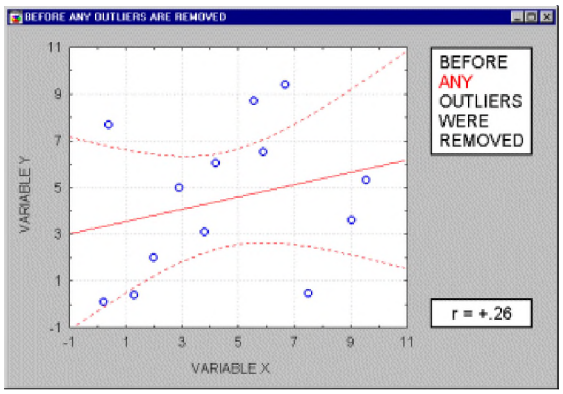

Рис. 1. Рассеивание точек корреляционного поля относительно теоретической линии регрессии

Большинство предположений множественной регрессии нельзя в точности проверить, однако можно обнаружить отклонения от этих предположений. В частности, выбросы (экстремальные наблюдения) могут вызвать серьезное смещение оценок, сдвигая линию регрессии в определенном направлении и, тем самым, вызывая смещение коэффициентов регрессии. Часто исключение всего одного экстремального наблюдения приводит к совершенно другому результату. Выбросы оказывают существенное влияние на угол наклона регрессионной линии и, соответственно, на коэффициент корреляции. Всего один выброс может полностью изменить наклон регрессионной линии и, следовательно, вид зависимости между переменными. Одна точка выброса обуславливает высокое значение коэффициента корреляции, в то время, как в отсутствие выброса, он практически равен нулю.

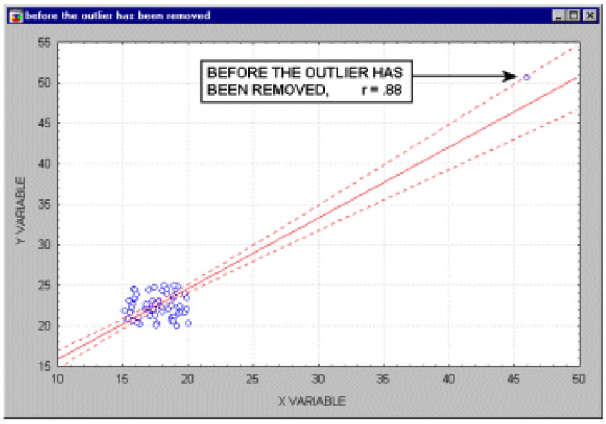

Рис. 2. Смещение оценок экстремальных наблюдений

При численности объектов анализа до 30 единиц возникает необходимость проверки значимости (существенности) каждого коэффициента регрессии. При этом выясняют, насколько вычисленные параметры характерны для отображения комплекса условий: не являются ли полученные значения параметров результатами действия случайных причин. Значимость коэффициентов простой линейной регрессии (применительно к совокупностям, у которых n<30) осуществляют с помощью t-критерия Стьюдента. При этом вычисляют расчетные (фактические) значения t-критерия для параметров a_0 a_1:

$$t_{a_0} = \frac{a_0\sqrt{n-2}}{\sigma_\varepsilon} \qquad t_{a_1} = \frac{a_1\sqrt{n-2}}{\sigma_\varepsilon}\sigma_x$$

$$\sigma_\varepsilon = \sqrt{\frac{\Sigma(Y_i - \tilde{Y})^2}{n-m}} \qquad \sigma_x = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-m}} \qquad (10)$$

– n – число наблюдений, m-число параметров уравнения регрессии,

– σ_ε – (остаточное) среднее квадратическое отклонение результативного признака от выровненных значений $\hat{y}$,

– σ_x– среднее квадратическое отклонение факторного признака от общей средней.

Вычисленные, по вышеприведенным формулам, значения сравнивают с критическими t, которые определяют по таблице значений Стьюдента с учетом принятого уровня значимости (α)и числа степеней свободы вариации k (v)=n-2. В социально-экономических исследованиях уровень значимости α обычно принимают равным 0,05. Параметр признается значимым (существенным) при условии, если $t_{расч.} > t_{табл.}$ В этом случае, практически невероятно, что найденные значения параметров обусловлены только случайными совпадениями.

Для оценки значимости парного коэффициента корреляции (корень квадратный из коэффициента детерминации), при условии линейной формы связи между факторами, можно использовать t-критерий Стьюдента:

$$t_r = \frac{|r| \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = |r| \cdot \sqrt{\frac{n-2}{1-r^2}} \quad (11)$$

Анализ качества эмпирического уравнения множественной линейной регрессии предусматривает оценку мультиколлинеарности факторов. При оценке мультиколлинеарности факторов следует учитывать, что чем ближе к нулю определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. Для отбора наиболее значимых факторов X_i должны быть учтены следующие условия:

– связь между результативным признаком и факторным должна быть выше межфакторной связи

– связь между факторами должна быть не более 0.7

– при высокой межфакторной связи признака отбираются факторы с меньшим коэффициентом корреляции между ними

Более объективную характеристику тесноты связи дают частные коэффициенты корреляции, измеряющие влияние на результативный фактор $У_i$ фактора X_i при неизменном уровне других факторов. Коэффициент частной корреляции отличается от простого коэффициента линейной парной корреляции тем, что он измеряет парную корреляцию соответствующих признаков (У и X_i) при условии, что влияние на них остальных факторов (X_j) устранено.

$$r_{yx_1}/x_2 = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1-r^2_{yx_2})(1-r^2_{x_1x_2})}} \quad (12)$$

$$r_{yx_2}/x_1 = \frac{r_{yx_2} - r_{yx_1} r_{x_2x_1}}{\sqrt{(1-r^2_{yx_1})(1-r^2_{x_2x_1})}} \quad (13)$$

$$r_{x_1x_2}/y = \frac{r_{x_1x_2} - r_{x_1y} r_{x_2y}}{\sqrt{(1-r^2_{x_1y})(1-r^2_{x_2y})}} \quad (14)$$

Литература

1. НорманДрейпер, Гарри Смит Прикладной регрессионный анализ. Множественная регрессия = AppliedRegressionAnalysis. – 3-е изд. – М.: «Диалектика», 2007. – С. 912.

2. Радченко Станислав Григорьевич, Устойчивые методы оценивания статистических моделей: Монография. – К.: ПП «Санспарель», 2005. – С. 504.

3. Радченко Станислав Григорьевич, Методология регрессионного анализа: Монография. – К.: "Корнийчук", 2011. – С. 376.