УДК.519.257

ИСПОЛЬЗОВАНИЕ БИНАРНЫХ ПЕРЕМЕННЫХ ПРИ РЕГРЕССИОННОМ МОДЕЛИРОВАНИИ СОСТОЯНИЯ ТЕХНИЧЕСКОГО ОБЪЕКТА

© 2014 Д.С. Бубырь[1], В.Н. Клячкин[1], И.Н. Карпунина[2]

[1] Ульяновский государственный технический университет
[2] Ульяновское высшее авиационное училище гражданской авиации

Поступила в редакцию 28.05.2014

Анализируется эффективность применения бинарных переменных при построении регрессионных моделей для оценки состояния технического объекта.

Ключевые слова: *моделирование, технический объект, регрессия, бинарные переменные, коэффициент детерминации*

Для оценки состояния технического объекта могут быть использованы регрессионные модели, отражающие связь параметров функционирования объекта с независимыми факторами, оказывающими влияние на его работоспособность. Такие модели часто строятся по результатам мониторинга системы. Если при этом регрессионная модель предназначена для прогнозирования состояния системы, то одним из важнейших показателей ее качества является коэффициент детерминации R^2 [1-3].

Процедура моделирования и перебора возможных регрессий осуществлялась в интегрированной системе комплексного статистического анализа и обработки данных STATISTICA [7-8]. В первоначальном варианте для поиска регрессий использовалась выборка, состоящая из данных за год (глобальная модель). Проведенные исследования показали, что такие модели обладают недостаточно высокой точностью, что можно объяснить неоднородностью физических свойств системы на области значений регрессоров. Для множественной линейной регрессии определена величина $R^2<0{,}5$. После использования пошаговой регрессии с целью удаления незначимых регрессоров, заметного улучшения значения коэффициента детерминации не наблюдалось: его величина также не превышала 0,5. Перебор различных типов нелинейных моделей (полный/неполный квадрат, куб, сумма всевозможных произведений и др.) привел к улучшению коэффициента детерминации на 10-20% при значительном усложнении структуры (для некоторых моделей количество регрессоров составляло 40 и более).

Бубырь Дмитрий Сергеевич, аспирант
Клячкин Владимир Николаевич, доктор технических наук, профессор кафедры прикладной математики и информатики. E-mail: v_kl@mail.ru.
Карпунина Ирина Николаевна, кандидат технических наук, доцент кафедры общепрофессиональных дисциплин. E-mail: karpunina53@yandex.ru

Значительно увеличить величину коэффициента детерминации получилось за счёт «кусочности» модели, то есть вариации её параметров по области значения регрессоров. Кусочно-линейная зависимость, обладающая достаточно простой структурой, дала улучшение значение коэффициента детерминации по различным выходным параметрам, характеризующим состояние объекта, на 21-56%. Кусочно-линейная регрессия оценивалась в виде:

$$Y_i = (b_{01} + b_{11} \cdot X_1 + \ldots + b_{m1} \cdot X_m) \cdot (Y_i \le c_i) + \\ + (b_{02} + b_{12} \cdot X_1 + \ldots + b_{m2} \cdot X_m) \cdot (Y_i > c_i) \quad , \qquad (1)$$

где m – количество независимых факторов; i – номер выходного параметра; c_i – точка разрыва; $(Y_i \le c_i)$,$(Y_i > c_i)$ – логические выражения, принимающие значения: 1 – если истинно, 0 – если ложно. Разрыв происходит по отклику: точкой разрыва выступает среднее значение отклика Y_i в данной выборке.

Наряду с количественными признаками в моделях иногда бывает необходимо использовать и качественные факторы. Например, это могут быть логические переменные, характеризующие сезонность наблюдения при использовании временных рядов, некие атрибутивные признаки при использовании пространственных данных. Качественные факторы могут быть добавлены в регрессионную модель, если они будут преобразованы в количественные переменные. Такие числовые переменные называются фиктивными или бинарными переменными [9].

С целью повышения значения коэффициента детерминации предпринята попытка введения бинарных переменных в кусочно-линейную регрессию. При этом рассмотрено два случая:

1) Добавление трех бинарных переменных, оценки которых показывают влияние сезонности на значение результирующего признака.

Таблица 1. Бинарные переменные сезонности

Сезон	S_1	S_2	S_3
зима	1	0	0
весна	0	1	0
лето	0	0	1
осень	0	0	0

Кусочно-линейная регрессия оценивалась в виде:

$$Y_i = (b_{01} + b_{11}X_1 + b_{21}X_2 + \ldots + b_{n1}X_n + b_{n+1,1}S_1 + b_{n+2,1}S_2 + b_{n+3,1}S_3)(Y_i \le c_i) + (b_{02} + b_{12}X_1 + b_{22}X_2 + \ldots + b_{n2}X_n + b_{n+1,2}S_1 + b_{n+2,2}S_2 + b_{n+3,3}S_3)(Y_i > c_i) \quad (2)$$

2) Добавление 11 переменных, оценки которых показывают влияние месяца на значение результирующего признака.

Таблица 2. Бинарные переменные месяца

Месяц	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
январь	1	0	0	0	0	0	0	0	0	0	0
февраль	0	1	0	0	0	0	0	0	0	0	0
март	0	0	1	0	0	0	0	0	0	0	0
апрель	0	0	0	1	0	0	0	0	0	0	0
май	0	0	0	0	1	0	0	0	0	0	0
июнь	0	0	0	0	0	1	0	0	0	0	0
июль	0	0	0	0	0	0	1	0	0	0	0
август	0	0	0	0	0	0	0	1	0	0	0
сентябрь	0	0	0	0	0	0	0	0	1	0	0
октябрь	0	0	0	0	0	0	0	0	0	1	0
ноябрь	0	0	0	0	0	0	0	0	0	0	1
декабрь	0	0	0	0	0	0	0	0	0	0	0

Кусочно-линейная регрессия оценивалась в виде:

$$Y_i = (b_{01} + b_{11}X_1 + b_{21}X_2 + \ldots + b_{n1}X_n + b_{n+1,1}M_1 + \ldots + b_{n+11,1}M_{11})(Y_i \le c_i) + (b_{02} + b_{12}X_1 + b_{22}X_2 + \ldots + b_{n2}X_n + b_{n+1,2}M_1 + \ldots + b_{n+11,2}M_{11})(Y_i > c_i) \quad (3)$$

Здесь n – количество регрессоров X, c_i – точка разрыва для показателя Y_i.

Таблица 3. Значения коэффициента детерминации

Показатель качества	Кусочно-линейная регрессия		
	с бинарными переменными		без бинарных переменных
	месяц	сезон	
Y_1	0,64	0,61	0,60
Y_2	0,67	0,65	0,64
Y_3	0,77	0,73	0,72
Y_4	0,71	0,66	0,62
Y_5	0,82	0,80	0,79
Y_6	0,70	0,68	0,68
Y_7	0,77	0,76	0,74

После применения данных регрессий для семи откликов Y (показателей качества функционирования объекта), получены следующие результаты по коэффициенту детерминации.

Из табл. 3 видно, что введение бинарных переменных, учитывающих сезонность, практически не повлияло на качество моделирования по показателю Y_6, максимальное увеличение коэффициента детерминации имеет место для показателя Y_4 (6,2%). Бинарные переменные, учитывающие влияние месяца на функционирование объекта, улучшили значение коэффициента детерминации максимум на 14,7% (по тому же показателю Y_4).

В зависимости от назначения и условий функционирования технического объекта прогнозирование его состояния иногда целесообразно проводить не по данным за год (глобальные модели), а по более коротким промежуткам времени (локальные модели). Исследования эффективности локальных моделей проводились в ситуации, когда для построения регрессионных зависимостей можно использовать от 30 до 40 наблюдений. По сравнению с глобальными моделями коэффициент детерминации значительно повысился. Для дальнейшего увеличения этого коэффициента вновь были введены бинарные переменные. Поскольку в данном случае размер выборки невелик (от одного до полутора месяцев), то добавление бинарных переменных, учитывающих сезон или месяц, не имеет смысла. Были введены переменные, учитывающие день недели (табл. 4).

В результате наблюдалось значительное увеличение коэффициента детерминации для некоторых откликов. Ниже (табл. 5) представлены значения коэффициента детерминации после применения моделей для выборки размера 30 дней. Видно, что использование бинарных переменных и вариация размера выборки позволяет повысить качество

регрессий, моделирующих состояние технического объекта.

Таблица 4. Бинарные переменные, учитывающие день недели

День	D1	D2	D3	D4	D5	D6
понедельник	1	0	0	0	0	0
вторник	0	1	0	0	0	0
среда	0	0	1	0	0	0
четверг	0	0	0	1	0	0
пятница	0	0	0	0	1	0
суббота	0	0	0	0	0	1
воскресение	0	0	0	0	0	0

Таблица 5. Значение коэффициента детерминации (выборки объемом 30 наблюдений)

Показатель качества	Кусочно-линейная регрессия	
	с бинарными переменными	без бинарных переменных
Y_1	0,92	0,86
Y_2	0,99	0,89
Y_3	0,98	0,83
Y_4	0,99	0,97
Y_5	0,99	0,98
Y_6	0,97	0,76
Y_7	0,97	0,86

Работа выполнена в рамках задания Минобрнауки России №2014/232.

СПИСОК ЛИТЕРАТУРЫ:

1. *Айвазян, С.А.* Прикладная статистика и основы эконометрики / *С.А. Айвазян, В.С. Мхитарян.* – М.: ЮНИТИ, 1998. 1022 с.
2. *Валеев, С.Г.* Регрессионное моделирование при обработке наблюдений. – М.: Наука, 1991. 272 с.
3. *Валеев, С.Г.* Особенности построения регрессионных моделей при многомерном контроле технологического процесса / *С.Г. Валеев, В.Н. Клячкин* // Радиоэлектроника. Информатика. Управление. 2002. №1. С.48-51.
4. *Валеев, С.Г.* Критерии выбора многоотключковых регрессий при контроле технологического процесса / *С.Г. Валеев, В.Н. Клячкин* // Проектирование и технология электронных средств. 2003. №2. С. 34-39.
5. *Клячкин, В.Н.* Статистические методы в управлении качеством: компьютерные технологии. – М.: Финансы и статистика, ИНФРА-М, 2009. 304 с.
6. *Клячкин, В.Н.* Идентификация режима статистического контроля многопараметрического технологического процесса / *В.Н. Клячкин, А.Ю. Михеев* // Автоматизация и современные технологии. 2011. №12. С. 27-31.
7. *Халафян, А.А.* STATISTICA 6. Статистический анализ данных. 3-е изд. – М.: ООО «Бином-Пресс», 2007. 512 с.
8. Statistica documentation [Электронный ресурс] // URL: http://documentation.statsoft.com (дата обращения: 31.03.2014)
9. *Каракозов, С.Г.* Основы эконометрики: учебное пособие. – Ульяновск: УлГУ, 2008. 127 с.
10. *Крашенинников, В.Р.* Кусочно-квадратичное моделирование регрессионных зависимостей при оценке качества / *Крашенинников В.Р., Бубырь Д.С.* // Междисциплинарные исследования в области математического моделирования и информатики. Мат-лы 3-й науч.-практ. internet-конференции. 20-21 февраля 2014 г. – Ульяновск: SIMJET, 2014. С. 233-236.
11. *Васильев, К.К.* Статистический анализ многомерных изображений / *К.К. Васильев, В.Р. Крашенинников.* – Ульяновск: УлГТУ, 2007. 170 с.
12. *Клячкин, В.Н.* Информационно-математическая система раннего предупреждения об аварийной ситуации / *В.Н. Клячкин, Ю.Е. Кувайскова, А.А. Алёшина, Ю.А. Кравцов* // Известия Самарского научного центра РАН. 2013. №4(4). С. 919-923.
13. *Кувайскова, Ю.Е.* Прогнозирование состояния технического объекта на основе мониторинга его параметров / *Ю.Е. Кувайскова, В.Н. Клячкин, Д.С. Бубырь* // XII Всероссийское совещание по проблемам управления. Институт проблем управления им. Трапезникова РАН [Электронный ресурс] URL: http://vspu2014.ipu.ru/node/2940 (дата обращения: 16.05.2014)

USE OF BINARY VARIABLES IN THE REGRESSION MODELING OF THE TECHNICAL OBJECT STATE

© 2014 D.S. Bubyr[1], V.N. Klyachkin[1], I.N. Karpunina[2]

[1] Ulyanovsk State Technical University
[2] Ulyanovsk Higher Civil Aviation School

In this article the effectiveness of binary variables use in the construction of regression models for technical object state estimation is analyzed. Quality models estimated by using the coefficient of determination. As the sample observations are considered for the year and in a shorter period of time.

Key words: *modeling, technical object, regression, binary variables, coefficient of determination*

Dmitriy Bubyr, Post-graduate Student; Vladimir Klyachkin, Doctor of Technical Sciences, Professor at the Department of Applied Mathematics and Computing Science. E-mail: v_kl@mail.ru; Irina Karpunina, Candidate of Technical Sciences, Associate Professor at the Department of Professional Disciplines. E-mail: karpunina53@yandex.ru