Лабораторная работа №3
«Построение моделей регрессии в R»

Цель работы

Изучить средства языка R для проведения регрессионного анализа: построения моделей регрессии, проверки на адекватность полученной модели, оценки качества модели и ее способности давать точный прогноз.

Методические указания

Регрессионный анализ – статистический метод исследования влияния одной или нескольких независимых переменных $x_i, i=\overline{1,q}$ на зависимую переменную y. Регрессионный анализ служит для определения вида этой связи и предоставляет возможность прогнозирования значений зависимой переменной по значениям независимых. Различают простую и множественную регрессию: различие состоит в том, что в первом случае рассматривают одну независимую переменную, а во втором – несколько.

Запишем линейную (по параметрам) регрессионную модель в общем виде:

$$y = b_0 + b_1 f_1\left(x_1,...,x_q\right) + b_2 f_2\left(x_1,...,x_q\right) + ... + b_m f_m\left(x_1,...,x_q\right),$$

где q – число независимых переменных, m – число регрессоров, включенных в модель. Регрессоры $f_j\left(x_1,...,x_q\right)$ могут представлять собой как степенные функции от отдельных переменных (и в целом нелинейные функции от них), так и взаимодействия нескольких переменных.

В регрессионной модели зависимости могут выражаться линейными или нелинейными функциями. Многие связи по своей природе, то есть в реальной жизни, либо являются собственно линейными, либо их можно привести к линейному виду, используя стандартные преобразования (возведение в степень, логарифмирование и пр.). Причем эти процедуры часто применяют не только к исходным независимым переменным, но к предсказываемой зависимой переменной: так, логарифмирование отклика –

это часто используемый в эконометрике прием, позволяющий одновременно и линеаризировать связи с независимыми переменными, и приблизить распределение наблюдений к нормальному закону.

Существует два принципиально различных подхода к процедуре формирования модели по множеству рассматриваемых регрессоров. Первый, традиционный, заключается в том, что все регрессоры одновременно включаются в модель, и определяется их значимость их, в зависимости от этого делаются выводы о качестве построенной модели. При этом подходе на первоначальных этапах анализа необходимо определить, есть ли среди рассматриваемых переменных x_i взаимозависимые . При обнаружений взаимокоррелированных x_i (ситуация мультиколлинеарности), необходимо решить вопрос о процедуре их включения в модель (возможные варианты – выделение групп независимых переменных, проведение факторного анализа).

Другим вариантом отбора переменных при построении модели является метод пошагового включения переменных в регрессионную модель (*Forward stepwise*) или пошагового исключения переменных из модели (*Backward stepwise*).

Определив вид регрессионной модели, и получив оценки параметров при соответствующих регрессорах, необходимо проверить качество и адекватность построениенной модели. В данной работе для контроля качества модели предлагается воспользоваться методом *скользящего контроля* или *кросс-валидации*. Метод заключается в следующем: на первом шаге исходная выборка разделяется на несколько подвыборок, далее одна из них рассматривается в качестве контрольной выборки, по которой вычисляется прогноз, а остальные – в качестве единой обучающей выборки. Такая процедура продолжается до тех пор, пока каждая из подвыборок по очереди не выступит в качестве контрольной. В результате такой серии проверок можно сделать выводы о качестве исследуемой, построенной до проведения кросс-валидации, модели на основании средней ошибки прогноза

и усредненных характеристик качества модели, в частности, значения коэффициента детерминации. В случае, если модель качественна, ее характеристики не будут существенно расходиться с усредненными, полученными по итогам проведения кросс-валидации.

В качестве проверки на адекватность построенной регрессионной модели предлагается анализировать выборку остатков полученной модели на близость к нормальному закону, поскольку именно выполнение предположений нормальности обеспечивает корректность выводов, связанных с гипотезами, рассматриваемыми в регрессионном анализе.

В случае, если проверка на качество и/или адекватность модели показала отрицательный результат, необходимо вернуться к началу и пересмотреть вид модели регрессии.

Рассмотрим примерную схему действий при проведении регрессионного анализа.

1. По рассматриваемым данным определить вид регрессионной модели: переменные, включаемые в модель в качестве регрессоров (отбор «вручную» с учетом корреляции x_i, либо с использованием пошаговых процедур типа *stepwise* или процедур выбора регрессоров по всем подмножествам), их взаимодействие и характер влияния на значения отклика. При необходимости, преобразовать переменные в фиктивные (т.н. *dummy* переменные).
2. Построить регрессионную модель по выбранным независимым переменным, получить оценки параметров модели, значение коэффициента детерминации R^2, проверить значимость регрессоров.
3. По значению R^2 сделать вывод о значимости построенной модели: в случае, когда $R^2 > 0.75$, т.е. включенные регрессоры объясняют более 75% изменчивости отклика, можно считать, что модель в достаточной степени отражает взаимосвязи между зависимой и независимыми переменными, и переходить к пункту 4. Иначе необходимо изменить структуру модели, в т.ч. включив в

рассмотрение новые, не используемые ранее регрессоры от исходных переменных, и вернуться к пункту 1 .

4. Проверить адекватность полученной модели путем проверки остатков на нормальность. В случае, когда модель не проходит эту проверку, изменить модель: например, провести логарифмирование отклика y (и пересчитать параметры модели уже для преобразованного y-ка), расширить множество регрессоров, включенных в модель, в т.ч. и за счет незначимых.
5. Проверить качество построенной регрессионной модели с использованием метода кросс-валидации: сделать вывод о степени изменения значения коэффициента детерминации и показателя качества прогноза. Если такие изменения существенны, то модель нельзя считать качественной, и необходимо изменение ее структуры, либо анализ данных на предмет наличия выбросов в выборке.

При проведении регрессионного анализа в системе R, предлагается воспользоваться операциями и функциями, представленными в таблицах 1, 2.

Таблица 1 – Символы, использующиеся при построении моделей регрессии в системе R

Символ	**Назначение**
~	Отделяет зависимые переменные от независимых. Например, предсказание значений y по значениям x, z и w будет закодировано так: $y \sim x+z+w$.
+	Разделяет независимые переменные
:	Обозначает взаимодействие между независимыми переменными. Предсказание значений y по значениям x, z будет закодировано как $y \sim x+z+x:z$
*	Краткое обозначение для всех возможных взаимодействий. Код $y \sim x*z*w$ в полном виде означает $y \sim x+z+w+x:z+x:w+z:w+$ $+x:z:w$
^	Обозначает взаимодействие до определенного порядка. Код $y \sim (x+z+w)\wedge 2$ в полном виде будет записан как $y \sim x+z+w+$ $+x:z+x:w+z:w$
.	Символ заполнитель для всех переменных в таблице данных, кроме зависимой. Например, если таблица данных содержит переменные x, y, z и w, то код $y \sim .$ будет означать $y \sim x+z+w$
-	Знак минуса удаляет переменную из уравнения. Например,

	$y \sim (x+z+w)\wedge 2 - x:w$ соответствует $y \sim x+z+w++x:z+z:w$
-1	Подавляет свободный член уравнения. Например, формула $y \sim x-1$ позволяет подогнать такую регрессионную модель для предсказания значений y по x, чтобы ее график проходил через начало координат
$I(\ )$	Элемент в скобках интерпретируется как арифметическое выражение. Например, $y \sim x+(z+w)\wedge 2$ означает $y \sim x+z+w+.$ $+z:w$. Для сравнения $y \sim x+I((z+w)\wedge 2)$ означает $y \sim x+h$, где h – это новая переменная, полученная при возведении в квадрат суммы z и w
function	В формулах можно использовать математические функции. Например, $\log(y) \sim x+z+w$ будет предсказывать значения $\log(y)$ по значениям x, z и w

Таблица 2 – Функции, использующиеся при построении моделей регрессии в системе R

Функция	**Действие**
summary()	Показывает детальную информацию о подогнанной модели
coefficients()	Перечисляет параметры модели (свободный член и регрессионные коэффициенты)
confint()	Вычисляет доверительные интервалы для параметров модели (по умолчанию 95%)
fitted()	Выводит на экран предсказанные значения согласно подогнанной модели
residuals()	Показывает остатки для подогнанной модели
anova()	Создает таблицу ANOVA (дисперсионного анализа) для подогнанной модели или таблицу ANOVA, сравнивающую две или более моделей
vcov()	Выводит ковариационную матрицу для параметров модели
AIC()	Вычисляет информационный критерий Акаике (Akaike's Information Criterion)
stepAIC()	Проведение пошаговой регрессии с использованием точного критерия AIC в качестве критерия включения или удаления переменных
crossval()	Функция для k-кратной кросс-валидации
predict()	Использует подогнанную модель для предсказания зависимости переменной для нового набора данных

Задание

1. По данным из лабораторной работы №1 построить регрессионную модель.

2. Проверить качество и адекватность построенной регрессионной модели, сделать соответствующие выводы, при необходимости вернуться на этап определения вида модели.
3. Для наиболее оптимальной построенной модели записать прогностическую модель, описать в терминах предметной области вид и характер полученных зависимостей, сделать сравнительный анализ степени влияния независимых переменных на отклик, характеристики качества получаемых прогнозов.

Список литературы

1. Статистические методы анализа данных : учеб.-метод. пособие / А. А. Попов. - : НГТУ, 2004. - 31 с.
2. R в действии. Анализ и визуализация данных в программе R. / пер. с англ. Полины А. Волковой. Роберт И. Кабаков – М.: ДМК Пресс, 2014. – 588 с. http://www.ievbras.ru/ecostat/Kiril/R/Biblio/R_rus/Kabacoff2014ru.pdf
3. Статистический анализ и визуализация данных с помощью R. С.Э. Мастицкий, В.К. Шитиков – Хайдельберг – Лондон – Тольятти, 2014. – 401с. http://www.ievbras.ru/ecostat/Kiril/R/Mastitsky%20and%20Shitikov%202014.pdf