МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗЕЙ В БИОМЕДИЦИНСКИХ СИСТЕМАХ С ПОМОЩЬЮ МОДЕЛЕЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Методические указания к лабораторной работе

САМАРА 2012

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗЕЙ В БИОМЕДИЦИНСКИХ СИСТЕМАХ С ПОМОЩЬЮ МОДЕЛЕЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

САМАРА 2012

УДК 621.398

Составитель: В.Н. Конюхов

Исследование взаимосвязей в биомедицинских системах с помощью моделей множественной регрессии. Метод. указания к лабораторной работе/ Самар. гос. аэрокосм. ун-т; Сост. В.Н. Конюхов, Самара, 2012. 25с.

В методических указаниях изложены базовые сведения о регрессионных моделях. Приведены основные области применения и особенности, а также порядок выполнения лабораторной работы.

Методические указания предназначены для магистров, обучающихся по направлению подготовки 201000.68 (Биотехнические системы и технологии) и выполняющих лабораторные работы по дисциплине «Методы математической обработки медико-биологических данных» на кафедре радиотехники и медицинских диагностических систем. Занятия в семестре А.

Печатаются по решению редакционно-издательского совета Самарского государственного аэрокосмического университета им. академика С.П. Королёва

Рецензент: проф. Гречишников В.М.

Цель работы: изучение основных регрессионных моделей и их особенностей, способов оценки параметров множественной линейной регрессии.

1. КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

Регрессионный анализ**-** раздел математической статистики, объединяющий практические методы исследования зависимости между величинами по данным статистических наблюдений. Достижение этой цели оказывается возможным за счет определения вида аналитического выражения, описывающего связь зависимой случайной величины Y (которую в этом случае называют результативным признаком) с независимыми случайными величинами X_1 ,X_2 , ..., X_m (которые называют факторами).

Основной задачей регрессионного анализа является установление формы линии регрессии и изучение зависимости между переменными. Основной задачей корреляционного анализа — выявление связи между случайными переменными и оценка ее тесноты.

Форма связи результативного признака Y с факторами X_1 ,X_2 , ..., X_m называется уравнением регрессии. В зависимости от типа выбранного уравнения различают линейную и нелинейную регрессию (например, квадратичную, логарифмическую, экспоненциальную и т.д.).

Регрессия может быть парная (простая) и множественная, что определяется числом взаимосвязанных признаков. Если исследуется связь между двумя признаками (результативным и факторным), то регрессия называется парной (простой). Если исследуется связь между тремя и более признаками, то регрессия называется множественной (многофакторной).

На этапе регрессионного анализа решаются следующие основные задачи.

1. Выбор общего вида уравнения регрессии и определение параметров регрессии.

2. Определение в регрессии степени взаимосвязи результативного признака и факторов, проверка общего качества уравнения регрессии.

3. Проверка статистической значимости каждого коэффициента уравнения

регрессии и определение их доверительных интервалов.

Самой употребляемой и наиболее простой из моделей множественной регрессии является линейная модель множественной регрессии:

$$y = \alpha' + \beta_1' x_1 + \beta_2' x_2 + \ldots + \beta_p' x_p + \varepsilon \qquad (1)$$

По математическому смыслу коэффициенты β_j' в уравнении (1) равны частным производным результативного признака y по соответствующим факторам:

$$\beta_1' = \frac{\partial y}{\partial x_1}, \beta_2' = \frac{\partial y}{\partial x_2}, \ldots, \beta_p' = \frac{\partial y}{\partial x_p}.$$

Параметр α называется свободным членом и определяет значение y в случае, когда все объясняющие переменные равны нулю. Однако, как и в случае парной регрессии, факторы по своему содержанию часто не могут принимать нулевых значений, и значение свободного члена не имеет смысла. При этом, в отличие от парной регрессии, значение каждого регрессионного коэффициента β_j' равно среднему изменению y при увеличении x_j на одну единицу лишь при условии, что все остальные факторы остались неизменными. Величина ε представляет собой случайную ошибку регрессионной зависимости.

Попутно отметим, что наиболее просто можно определять оценки параметров β_j', изменяя только один фактор x_j, оставляя при этом значения других факторов неизменными. Тогда задача оценки параметров сводилась бы к последовательности задач парного регрессионного анализа по каждому фактору. Однако такой подход, широко используемый в естественнонаучных исследованиях, (физических, химических, биологических), в медицинских обследованиях является неприемлемым. Врач, в отличие от экспериментатора – естественника, лишен возможности регулировать отдельные факторы, поскольку не удаётся обеспечить равенство всех прочих условий для оценки влияния одного исследуемого фактора.

Получение оценок параметров $\alpha', \beta_1', \beta_2' \ldots, \beta_p'$ уравнения регрессии (1) – одна из важнейших задач множественного регрессионного анализа. Самым распространенным методом решения этой задачи является метод наименьших квадратов (МНК). Его суть состоит в минимизации суммы квадратов отклонений наблюдаемых значений зависимой переменной у от её значений, получаемых по уравнению регрессии. Поскольку параметры $\alpha', \beta_1', \beta_2'..., \beta_p'$ являются случайными величинами, определить их истинные значения по выборке невозможно. Поэтому вместо теоретического уравнения регрессии (1) оценивается так называемое эмпирическое уравнение регрессии, которое можно представить в виде:

$$y = a + b_1x_1 + b_2x_2 + ... + b_px_p + e \tag{2}$$

Здесь $a, b_1, b_2, ..., b_p$ - оценки теоретических значений $\alpha', \beta_1', \beta_2'..., \beta_p'$, или эмпирические коэффициенты регрессии, е – оценка отклонения ε.

Оценим параметры с помощью метода наименьших квадратов. Для этого построим систему нормальных уравнений, решение которой позволяет получить оценки параметров регрессии:

$$\begin{cases} \sum y = na + b_1\sum x_1 + b_2\sum x_2 + ... + b_p\sum x_p, \\ \sum yx_1 = a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_2x_1 + ... + b_p\sum x_px_1, \\ ... \\ \sum yx_p = a\sum x_p + b_1\sum x_1x_p + b_2\sum x_2x_p + ... + b_p\sum x_p^2. \end{cases} \tag{3}$$

Другой вид уравнения множественной регрессии – уравнение регрессии в стандартизированном масштабе:

$$t_y = \beta_1t_{x_1} + \beta_2t_{x_2} + ... + \beta_pt_{x_p}, \tag{4}$$

где $t_y = \dfrac{y - \bar{y}}{\sigma_y}$, $t_{x_i} = \dfrac{x_i - \overline{x_i}}{\sigma_{x_i}}$ – стандартизированные переменные;

β_i – стандартизированные коэффициенты регрессии.

К уравнению множественной регрессии в стандартизированном масштабе применим МНК, что приводит к решению системы уравнений:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_2x_1} + \beta_3 r_{x_3x_1} + \ldots + \beta_p r_{x_px_1}, \\ r_{yx_2} = \beta_1 r_{x_1x_2} + \beta_2 + \beta_3 r_{x_3x_2} + \ldots + \beta_p r_{x_px_2}, \\ \ldots \\ r_{yx_p} = \beta_1 r_{x_1x_p} + \beta_2 r_{x_2x_p} + \beta_3 r_{x_3x_p} + \ldots + \beta_p. \end{cases} \tag{5}$$

Для двухфакторной модели линейной регрессии $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2}$ расчет β-коэффициентов можно выполнить по формулам (следуют из решения системы (2.4)):

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r_{x_1x_2}^2}, \ \beta_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r_{x_1x_2}^2} \tag{6}$$

Связь коэффициентов множественной регрессии b_i со стандартизированными коэффициентами β_i описывается соотношением:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}, \ \beta_i = b_i \frac{\sigma_{x_i}}{\sigma_y}. \tag{7}$$

При этом: $a = \bar{y} - b_1 \overline{x_1} - b_2 \overline{x_2}$.

Тесноту совместного влияния факторов на результат оценивает *коэффициент множественной корреляции*, который можно определить по формуле:

$$R_{yx_1x_2\ldots x_p} = \sqrt{\sum \beta_i r_{yx_i}}, \tag{8}$$

где β_i – стандартизированные коэффициенты регрессии,

r_{yx_i} – парные коэффициенты корреляции между переменными y и x_i.

Качество построенной модели в целом оценивает коэффициент (индекс) детерминации. *Коэффициент множественной детерминации* рассчитывается как квадрат индекса множественной корреляции:

$$R_{yx_1x_2\ldots x_p}^2. \tag{9}$$

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при устранении влияния (при закреплении их влияния на постоянном уровне) других факторов, включенных в уравнение регрессии. Для двухфакторной модели их можно определить по формулам:

$$r_{yx_1\cdot x_2}=\frac{r_{yx_1}-r_{yx_2}\cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)(1-r_{x_1x_2}^2)}};\ r_{yx_2\cdot x_1}=\frac{r_{yx_2}-r_{yx_1}\cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}};$$
$$r_{x_1x_2\cdot y}=\frac{r_{x_1x_2}-r_{yx_1}\cdot r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{yx_2}^2)}}. \qquad (10)$$

При построении уравнения множественной регрессии может возникнуть проблема мультиколлениарности факторов (тесная линейная зависимость более двух факторов). Считается, что две переменные явно *коллинеарны*, если $r_{x_ix_j}>0{,}7$.

Статистическая значимость уравнения множественной регрессии в целом оценивается с помощью общего F-критерия Фишера:

$$F=\frac{R^2_{yx_1x_2\ldots x_p}}{1-R^2_{yx_1x_2\ldots x_p}}\cdot\frac{n-m-1}{m}, \qquad (11)$$

где m – число факторов в линейном уравнении регрессии;
n – число наблюдений.

Вывод о статистической значимости уравнения множественной регрессии в целом и коэффициента множественной детерминации можно сделать, если наблюдаемое значение критерия больше табличного, найденного для заданного уровня значимости (например, $\alpha=0{,}05$) и степенях свободы $k_1=m$, $k_2=n-m-1$.

Частный F-критерий оценивает статистическую значимость присутствия каждого из факторов в уравнении множественной регрессии. Для двухфакторной модели F_{x_1} оценивает целесообразность включения в уравнение фактора x_1 после того, как в него был включен фактор x_2; F_{x_2} оценивает целесообразность включения в уравнение фактора x_2 после того, как в него был включен фактор x_1:

$$F_{x_1}=\frac{R^2_{yx_1x_2}-r^2_{yx_2}}{1-R^2_{yx_1x_2}}\cdot\frac{n-m-1}{1},\ F_{x_2}=\frac{R^2_{yx_1x_2}-r^2_{yx_1}}{1-R^2_{yx_1x_2}}\cdot\frac{n-m-1}{1}, \qquad (12)$$

где m – число факторов в линейном уравнении регрессии;
n – число наблюдений.

Фактическое значение частного F-критерия сравнивается с табличным при 5%-ном или 1%-ном уровне значимости и числе степеней свободы: $k_1 = 1$, $k_2 = n - m - 1$. Если фактическое значение превышает табличное, то дополнительное включение соответствующего фактора в модель статистически оправдано, в противном случае фактор в модель включать нецелесообразно.

2.ФУНКЦИИ ПАКЕТА MATLAB ДЛЯ ПРОВЕДЕНИЯ МНОЖЕСТВЕННОГО РЕГРЕССИОННОГО АНАЛИЗА.

В пакете Matlab предусмотрено множество функций для проведения как линейного, так и нелинейного регрессионного анализа. К функциям, для проведения линейного анализа, относятся, например, такие как:

glmfit - Определение параметров обобщенной линейной модели

glmval - Прогнозирование с использованием обобщенной линейной модели

leverage - Оценка степени влияния отдельных наблюдений в исходном многомерном множестве данных на значения параметров линии регрессии.

lscov - Линейная регрессия (метод наименьших квадратов) при заданной матрице ковариаций (встроенная функция MATLAB)

polyconf - Определение доверительных интервалов для линии регрессии

polyfit - Полиномиальная регрессия (встроенная функция MATLAB)

polyval - Прогноз с использованием полиномиальной регрессии (встроенная функция MATLAB)

rcoplot - График остатков

regress - Множественная линейная регрессия

regstats - Функция диагностирования линейной множественной модели. Графический интерфейс.

robustfit - Робастная оценка параметров регрессионной модели

stepwise - Пошаговая регрессия (графический интерфейс пользователя).

К функциям нелинейного-

lsqnonneg - Функция реализует метод наименьших квадратов и возвращает только неотрицательные значения параметров модели (встроенная функция MATLAB)

nlinfit - Нелинейный метод наименьших квадратов (метод Гаусса-Ньютона)

nnls - Решение системы линейных уравнений методом наименьших квадратов для неотрицательных значений аргумента

nlintool - График прогнозируемых значений

nlparci - Вектор доверительных интервалов для параметров модели

nlpredci - Прогнозируемые значения и их доверительные интервалы.

Для проведения множественного линейного регрессионного анализа в данной лабораторной работе используется функция b = regress(y,X), которая предназначена для расчета точечных оценок коэффициентов линейного уравнения регрессии b. Расчет точечных оценок коэффициентов выполняется методом наименьших квадратов из следующего уравнения линейной модели:

$$y = X\beta + \varepsilon,$$

где y - вектор значений зависимой переменной; β - вектор коэффициентов линейной модели; X - матрица значений независимых переменных; ε - вектор случайных возмущающих факторов, распределенных по нормальному закону с нулевым математическим ожиданием и дисперсией σ^2, $\varepsilon \sim N(0,\sigma^2 I)$.

Размерности векторов значений зависимой переменной y и случайных возмущающих факторов ε - $n\times1$, где n - количество наблюдений. Размерность матрицы X равна $n\times p$, где p - количество независимых переменных. Столбцы матрицы X соответствуют независимым переменных, строки - наблюдениям. Размерность вектора коэффициентов линейной регрессионной модели равна $p\times1$. Коэффициенты множественной линейной регрессионной модели в векторе b располагаются по возрастанию степени независимых переменных.

[b,bint,r,rint,stats] = regress(y,X) функция возвращает: b - вектор точечных оценок коэффициентов линейного уравнения регрессии, bint - матрицу интервальных оценок параметров линейной регрессии, r - вектор остатков, rint - матрицу 95% доверительных интервалов остатков, stats - структуру, содержащую значения статистики R^2 с соответствующими ей F статистикой и уровнем значимости p для регрессионной модели.

Размерность матрицы bint составляет p×2, где первый столбец матрицы задает нижнюю границу 95% доверительного интервала, второй - верхнюю границу 95% доверительного интервала. Количество элементов вектора r равно n. Размерность матрицы rint равна n×2, где первый и второй столбцы используются для задания нижней и верхней границ 95% доверительного интервала по каждому из n наблюдений.

[b,bint,r,rint,stats] = regress(y,X,alpha) входной параметр alpha позволяет задать величину уровня значимости. Уровень значимости используется для расчета границ доверительных интервалов bint и rint с доверительной вероятностью определяемой как 100(1-alpha)%. Значение alpha=0.2 будет соответствовать 80% границам доверительных интервалов bint и rint.

3. ОПИСАНИЕ ДАННЫХ И ЭКСПОРТ ДАННЫХ В ПАКЕТ MATLAB

Исходные данные для выполнения лабораторной работы приведены в файле rus_dasl13. В таблице Excel содержаться четыре переменные- пол, возраст, общий холестерин, индекс массы тела. Экспорт данных в matlab можно провести с помощью функции xlsread. Описание использования функции xlsread можно получить с помощью команды help.

Цель обработки исходных данных заключается в построении зависимости переменной VAR3 от переменных VAR2 и VAR4.

4. ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

1. Считая зависимой переменной общее содержание холестерина в крови построить парную линейную регрессию, считая независимыми переменными возраст и индекс массы тела не учитывая пол.
2. Считая зависимой переменной общее содержание холестерина в крови построить множественную линейную регрессию, считая независимыми переменными возраст и индекс массы тела не учитывая пол.
3. Повторить п.1 и 2 с учетом пола пациента.
4. Построить для полученных в п.п 1-3 регрессионных моделей вектор остатков и их границы доверительных интервалов.
5. Оценить для моделей п. 2 и 3 целесообразность включения второго фактора.

5. СОДЕРЖАНИЕ ОТЧЕТА

1. Наименование и цель работы.
2. Графики зависимостей, полученные в п.1-4.
3. Расчеты оценок целесообразности включения в регрессионную модель дополнительного фактора.
4. Выводы.

6. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое регрессионный анализ?
2. Чем отличается парная регрессия от множественной?
3. Какие задачи решает регрессионный анализ?
4. Назовите основные этапы регрессионного анализа.
5. Приведите теоретическое уравнение множественной линейной регрессии. Поясните смысл коэффициентов.
6. Как определяются коэффициенты множественной линейной регрессии на практике?
7. Что такое стандартизированные коэффициенты регрессии?

8. Как связаны коэффициенты множественной регрессии со стандартизированными коэффициентами?

9. Что такое коэффициент множественной корреляции?

10. Что такое коэффициент множественной детерминации?

11. Частные коэффициенты корреляции?

12. В каком случае переменные коллинеарны?

13. Как оценить статистическую значимость уравнения множественной регрессии?

14. Как оценить статистическую значимость присутствия каждого из факторов в уравнении множественной регрессии?

15. Поясните структуру, возвращаемую функцией regress.

СПИСОК ЛИТЕРАТУРЫ

1. Орлов А.И. Прикладная статистика. М: Издательство «Экзамен», 2004.
2. http://matlab.exponenta.ru/statist/book2/11/regress.php
3. http://statsoft.ru/home/textbook/modules/stmulreg.html
4. https://function-x.ru/statistics_regression1.html#paragraph2

Учебное издание

ИССЛЕДОВАНИЕ ВЗАИМОСВЯЗЕЙ В БИОМЕДИЦИНСКИХ СИСТЕМАХ С ПОМОЩЬЮ МОДЕЛЕЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Методические указания

Составитель: Конюхов Вадим Николаевич

Самарский государственный аэрокосмический университет
имени академика С.П. Королёва.
443086 Самара, Московское шоссе, 34