Computer Modelling & New Technologies, 2002, Volume 6, No.1, 51-61
Transport and Telecommunication Institute, Lomonosov Str.1, Riga, LV-1019, Latvia

СТАТИСТИЧЕСКИЕ МЕТОДЫ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ДАННЫХ

Е. ЗЛОБА, И. ЯЦКИВ

Институт транспорта и связи, Рига, Латвия
Ломоносова 1,Рига, LV 1019, Латвия.
Ph: (+371)-7100594. Fax: (+371)-7100660. E-mail: ivl@tsi.lv

Paper reviews the current state the problem of statistical analysis with missing data and the methods of it decision. The using of *resampling* for this task is offered. The aims of the work are the demonstration of possibilities *resampling* for this task and investigation the effectiveness of *resampling* method and classical methods: Bartlett and means for regression analysis with missing data in dependent variable.
Keywords: missing data, regression analysis

1. Введение

При проведении статистического анализа на практике ограничиваются анализом не всей генеральной совокупности в целом, а лишь некоторого выборочного числа наблюдений. Анализируемая выборка должна отвечать критериям качественности и полноты. В реальности приходится сталкиваться с ситуацией, когда некоторые из свойств одного или нескольких объектов отсутствуют – возникает ситуация данных с пропусками, что значительно осложняет математическую обработку, так как смещение основных статистических характеристик, таких как математическое ожидание или дисперсия, например, возрастает прямо пропорционально числу пропусков. К возникновению пропусков в исходных данных может привести множество причин: например, отсутствие значений вследствие каких-то мелких поломок оборудования, не связанных с экспериментальным процессом, или нежелание респондента при проведении статистического опроса отвечать на вопросы о своих доходах.

На сегодняшний день в математической статистике существует несколько путей решения проблемы неполных данных [1]:

- исключение некомплектных объектов из исходной выборки. Данный подход к проблеме можно охарактеризовать как некорректный, так как неполные данные несут в себе новую информацию, необходимую для исследования, и поэтому их важно включать в анализ;
- применение специально разработанных математических методов анализа неполных данных, таких как метод взвешивания [1] или метод максимального правдоподобия и EM-алгоритм [1] (при этом значительно возрастает сложность проводимого анализа);
- восстановление пропусков (наиболее распространены методы заполнения по среднему и по регрессии). В большинстве случаев именно этот подход считается наиболее эффективным и удобным решением проблемы.

Основным инструментом прикладной статистической обработки данных служат пакеты программ, библиотеки и другие программные продукты. Можно констатировать, что современное статистическое программное обеспечение анализа данных с пропусками в целом находится на начальном уровне. Практически все статистические программные средства, в которых предусмотрена возможность наличия пропусков в данных, содержат лишь простые методы – такие, как, например, исключение некомплектных наблюдений, заполнение пропусков средними, заполнение с помощью регрессии или вычисление ковариационной матрицы и вектора средних парными методами и т.д., то есть методы, которые были реализованы еще в первых версиях пакетов SSP, IMSL, BMDP или Statistica. Однако, как было показано выше, эти методы часто дают неудовлетворительные результаты. В этой связи актуальной является разработка статистического программного обеспечения, основанного на новых подходах.

Одним из перспективных, сравнительно новым в статистическом анализе методом, считается *resampling*-метод, применение которого для задачи заполнения пропусков в неполных

данных и будет подробно исследовано в работе. Целью работы является анализ эффективности метода resampling по сравнению с другими широко применяемыми статистическими методами заполнения пропусков: метод восстановления пропусков Бартлетта; метод восстановления по среднему. Разработано несколько программных модулей с целью дополнить базовый пакет Statistica/Win вышеупомянутыми статистическими методами для восстановления данных с пропущенными значениями.

2. Проблема неполных данных и известные методы ее решения

Пусть исходные данные представлены в виде матрицы $Y_{n\times p}$, строки которой соответствуют n изучаемым объектам, а столбцы представляют собой данные по p переменным, измеряемые для каждого объекта. Элементы матрицы y_{ij}, где $i = 1..n, j = 1..p$ являются действительными числами — значениями непрерывных (например, время или размер дохода), дискретных или категориальных переменных (например, пол человека). В свою очередь категориальные признаки могут быть упорядоченными (например, образование) или неупорядоченными (раса, пол).

Причины пропусков данных могут быть самыми разными, поэтому знание механизма, приводящего к отсутствию значений, является ключевым при выборе методов анализа и интерпретации результатов. Механизм порождения пропусков дает понимание степени важности потерянной информации, ведь неполные данные несут в себе новую информацию, необходимую для исследования, поэтому их важно включать в анализ.

Иногда механизм порождения пропусков управляется статистиком. Например, мы можем считать, что выборочному обследованию пропуски присущи, так как значения части переменных в обследовании присутствуют у всех объектов популяции, а исследуемые переменные «пропущены» у объектов, не включенных в выборку. Здесь механизм порождения пропусков — процесс извлечения выборки. Если объекты извлекаются из популяции случайно, то механизм управляется исследователем и его можно назвать «игнорируемым». Если правило извлечения выборки не соблюдается или для некоторых объектов выборки значения отсутствуют, то механизм порождения пропусков не столь ясен. В этом случае анализ зависит от предположений о механизме образования пропусков, которые следует явно оговаривать.

Цензурирование — пример ситуации, когда механизм порождения пропусков может быть неуправляемым, но известным статистику. Данными является время наступления некоторого события (смерть животного в эксперименте, рождение ребенка, перегорание лампочки). Для некоторых объектов выборки время события цензурировано, поскольку событие не успело наступить до окончания эксперимента. Если известна точка (время) цензурирования, то мы имеем частичную информацию о том, что время наступления ненаблюденного события больше времени цензурирования. Такую информацию надо учитывать при анализе, чтобы избежать смещений.

Часто механизм порождения пропусков явно не включают в модель - подразумевается, что этот механизм игнорируется (так например, в пакете Statistica 5.0 предлагается исключить из анализа пропущенные значения, активизировав опцию «Casewise (listwise) deletion of MD»). Однако механизм пропусков можно вводить в статистическую модель, включая в нее распределение индикаторов присутствия – некоторую функцию I={0,1}, принимающую значение 1 при наличии признака y_{ij} и 0 — для пропуска. В общем случае механизмом пропусков нельзя пренебречь.

Методы анализа неполных данных можно условно разбить на следующие группы.

A. Метод исключения некомплектных объектов. При отсутствии у ряда объектов значений каких-либо переменных некомплектные объекты удаляются из анализа. Подход легко реализуется и может быть удовлетворительным при малом числе пропусков. Однако иногда он приводит к серьезным смещениям и обычно не очень эффективен. Главный недостаток такого подхода обусловлен потерей информации при исключении неполных наблюдений.

B. Методы с заполнением. При данном подходе пропущенные значения исходной выборки заполняются и полученные «полные» данные обрабатываются обычными методами. Наиболее часто используются следующие процедуры заполнения пропусков.

Заполнение средними. Подставляются средние присутствующих значений. Метод безусловного среднего - самый простой вид заполнения. Он заключается в оценке отсутствующих значений y_{ij}

средним $\bar{y}_j^{(j)}$ по присутствующим значениям переменной Y_j. Среднее наблюдаемых и подставленных значений равно $\bar{y}_j^{(j)}$ - оценке методом доступных наблюдений.

Заполнение с пристрастным подбором. Пропуски заполняются значениями, полученными для другого сходного объекта выборки. Процедуру можно описать как метод, при котором подстановка выбирается для каждого пропущенного значения по оценке распределения в отличие от заполнения пропусков средними, когда подставляется среднее распределения. В большинстве приложений эмпирическое распределение задается присутствующими значениями, поэтому при заполнении с подбором подставляются различные значения из данных для сходных объектов без пропусков.

Наиболее часто используемые методы: *подстановка с подбором внутри групп* и *подбор ближайшего соседа.* В первом случае формируются группы, и пропуски в каждой группе заполняются присутствующими значениями из нее же. Заполнение с подбором широко распространено. Оно может включать очень сложные схемы отбора объектов. Хотя практика подтвердила достоинства этого метода, литературы, посвященной его теоретическим свойствам, явно недостаточно. Второй подход основан на введении метрики *d* для измерения расстояния между объектами, определенного в пространстве сопутствующих переменных, и выборе подстановки по объекту с присутствующим значением, ближайшему к объекту с пропуском. Например, обозначим $x_{i1},...,x_{iJ}$ - значения J сопеременных, измеренных в нормированных шкалах, у объекта *i* с пропуском y_i. Определим расстояние $d(i,k) = \max_j |x_{ij} - x_{kj}|$ между объектами *i* и *k*. Мы можем выбирать подстановку для y_i из тех *k*-х объектов, у которых:

1) наблюдаются $y_k, x_{k1},..., x_{kJ}$;

2) $d(i,k)$ меньше некоторого порога d_0.

Число «кандидатов» – подходящих *k*-х объектов – можно выбирать, изменяя d_0. Схемы ближайшего соседа требуют значительных вычислительных затрат. Они стали применяться сравнительно недавно.

Заполнение с помощью регрессии. Когда пропущенные значения оцениваются с помощью регрессии на присутствующие для анализируемого объекта переменные. В частности, к этой группе относится метод заполнения условными средними или так называемый метод Бака. Метод является более перспективным способом заполнения пропусков по сравнению с предыдущими методами. Он заключается в подстановке средних, условных по присутствующим в наблюдении переменным и относится к модельным методам.

Если переменные $Y_1,...,Y_k$ распределены по многомерному нормальному закону со средним μ и ковариационной матрицей Σ, то регрессия пропущенных значений в данном наблюдении линейна по присутствующим значениям с коэффициентами, которые являются хорошо известными функциями от μ и Σ. В методе, предложенном Баком, сначала оценивают μ и Σ по полным наблюдениям, а затем используют эти оценки для вычисления линейной регрессии пропущенных переменных по присутствующим для каждого наблюдения. Подставляя значения переменных, присутствующих для данного наблюдения, в регрессионное уравнение, получаем прогноз пропущенных переменных для этого наблюдения. Вычисление регрессионных уравнений для различной структуры пропусков может показаться затруднительным, но на деле оно относительно просто, если использовать оператор свертки. Данные, заполненные по методу Бака, обеспечивают разумные оценки средних, в частности, если приемлемо предположение о нормальности наблюдений. Выборочная ковариационная матрица по заполненным данным занижает величину дисперсии и ковариаций, хотя и не так сильно, как при подстановке безусловных средних. Также среди методов с заполнением можно выделить: *заполнение без подбора, многократного заполнения, составные и другие методы.*

С. Методы взвешивания. Рандомизированные выводы по данным выборочных обследований с пропусками построены на весах плана, обратно пропорциональных вероятности выбора. Пусть y_i — значение переменной Y i-го объекта популяции. Тогда среднее популяции часто оценивают величиной

$$\sum \pi_i^{1} y_i \Big/ \sum \pi_i^{-1}, \qquad (1)$$

где суммы берутся по извлеченным объектам, π_i — вероятность извлечения *i*-го объекта, π_i^{-1} - вес плана *i*-го элемента. Методы взвешивания изменяют веса, чтобы учесть отсутствие значений. Оценка (1) заменяется оценкой

$$\sum(\pi_i \hat{p}_i)^{-1} y_i \Big/ \sum(\pi_i \hat{p}_i)^{-1}, \qquad (2)$$

где суммы берутся по извлеченным объектам, в которых нет пропусков, $\hat{p}_i$ — оценка вероятности присутствия значения для *i*-го объекта (обычно доля объектов выборки с присутствующим значением). Взвешивание связано с заполнением средними. Например, если веса плана постоянны в подгруппах выборки, то заполнение пропусков в каждой подгруппе средними подгруппы и взвешивание присутствующих значений с помощью их доли в каждой подгруппе ведут к одинаковым оценкам среднего популяции, хотя оценки выборочной дисперсии различны, если только не используются поправки на заполнение средними.

D. Методы, основанные на моделировании. Широкий класс методов основывается на построении модели порождения пропусков. Выводы получают с помощью функции правдоподобия, построенной при условии справедливости этой модели, с оцениванием параметров методами типа максимального правдоподобия.

В методах, использующих функцию правдоподобия, реализована относительно старая идея обработки неполных данных:

- заполнение пропусков оценками пропущенных значений;
- оценивание параметров;
- повторное оценивание пропущенных значений (оценки параметров считаются точными);
- повторное оценивание параметров и так далее до сходимости процесса.

Преимущества такого подхода состоят в том, что он гибок; позволяет отказаться от методов, разработанных для частных случаев; позволяет оценивать в приближении большой выборки дисперсии оценок с помощью матрицы вторых производных функций правдоподобия для неполных данных; обеспечивает надежную сходимость, т.е. в определенных нестрогих условиях каждая итерация увеличивает логарифм правдоподобия и последовательность сходится к некоторому стационарному значению. Недостаток алгоритма заключается в том, что скорость сходимости может быть очень низкой, если пропущено много данных.

3. Исследуемые методы заполнения пропущенных значений

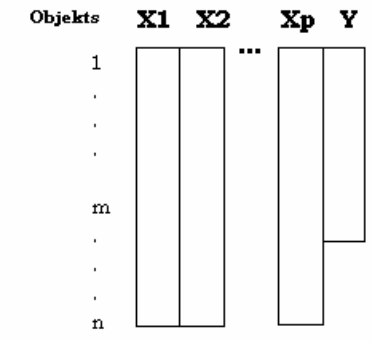


Рисунок 1. Монотонная структура с пропусками в одной переменной Y

Рассмотрим задачу с пропусками в зависимой переменной. Во многих задачах полностью присутствуют значения $p>1$ переменных для всех n объектов. Такие данные можно представить рисунок 1, где X представляет собой матрицу nxp, а пропущенные значения представляют $(n-m)$ объектов Y. При обычном анализе неполных данных используют предположение, что данные отсутствуют случайно, т.е. предполагают, что вероятность пропуска y_i может меняться в зависимости от переменных плана, но при данном значении x_i, *i-й* строки X, вероятность отсутствия y_i не зависит от y_i. В практических приложениях следует проверять допустимость такого предположения. Анализ строится так, чтобы использовать «почти сбалансированность» получаемого множества данных для упрощения вычислений. При подстановке оценок пропущенных значений вместо пропусков следует уделить внимание таким вопросам, как выбор значений для подстановки и модификации методов с целью учета этих подстановок.

Рассмотрим ситуацию, когда X – некоторые задаваемые исследователем факторы, а Y – зависимая от этих факторов переменная. Поскольку в эксперименте значения факторов задаются статистиком, то пропуски, если они есть, содержатся в выходной переменной Y намного чаще, чем в значениях факторов X. Поэтому мы ограничимся ситуацией, когда пропуски только в Y.

3.1. МЕТОД БАРТЛЕТТА ДЛЯ ЗАПОЛНЕНИЯ ПРОПУСКОВ

Метод, предложенный Бартлеттом для решения данной проблемы (1937), заключается в подстановке начальных значений вместо пропусков и проведении ковариационного анализа с сопутствующей переменной пропусков для каждого пропущенного значения.

Допустим, что каждый пропуск y_i заполняется начальным значением, чтобы вектор значений **Y** был полон. Обозначим начальные значения $\tilde{y}_i, i = 1,..., m_0$. Пусть Z – *n* x *m*-матрица m_0 сопутствующих переменных пропусков. По определению *i–я* сопутствующая переменная пропусков – это индикатор *i–го* пропущенного значения, т.е. всегда 0, за исключением случая, когда пропущено *i–е* значение, тогда она равна 1. Первая строка Z, z_1, равна (1,0,…,0), …, строка m_0 равна (0,…,0,1), а все z_i при $i > m_0$ равны (0,…,0), так как они соответствуют присутствующим y_i. При ковариационном анализе используется и X, и Z для предсказания Y.

Предположим, что для выходной переменной $Y = (y_1,..., y_n)^T$ верна линейная модель:

$$Y = X_\beta + Z_\gamma + e, \qquad (3)$$

где γ – вектор-столбец из m_0 коэффициентов регрессии для сопутствующих переменных пропусков, $e = (e_1,..., e_n)^T$, e_i - независимо и одинаково распределены с нулевым средним и одинаковой дисперсией σ^2, β - оцениваемый параметр – вектор длины p.

Классическая оценка наименьших квадратов β равна:

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \qquad (4)$$

если $(X^T X)$ имеет полный ранг. Если $(X^T X)$ невырождена, то $\hat{\beta}$ – несмещенная оценка β с минимальной дисперсией. Если e_i распределены нормально, то $\hat{\beta}$ - оценка максимального правдоподобия, распределенная нормально со средним β и дисперсией $\sigma^2 (X^T X)^{-1}$.

Рассмотрим метод для нашей задачи. Остаточная сумма квадратов, минимизируемая по (β, γ), равна: $SS(\beta,\gamma) = \sum_{i=1}^{m_0} (\tilde{y}_i - x_i\beta - z_i\gamma)^2 + \sum_{i=m_0+1}^{n} (y_i - x_i\beta - z_i\gamma)^2$.

Упростим, используя определение матрицы *Z*

$$SS(\beta,\gamma) = \sum_{i=1}^{m_0} (\tilde{y}_i - x_i\beta - \gamma_i)^2 + \sum_{i=m_0+1}^{n} (y_i - x_i\beta)^2 . \qquad (5)$$

Назовем $\hat{\beta}_*$ – правильная оценка наименьших квадратов β, полученная по формуле (4) по присутствующим значениям, т.е. по последним $m = n\text{-}m_0$ строкам (Y,X). Она минимизирует вторую сумму в выражении (5). Но если при $\beta = \hat{\beta}$ положить $\gamma = (\hat{\gamma}_1,..., \hat{\gamma}_{m_0})^T$, где

$$\hat{\gamma}_i = \tilde{y}_i - x_i\hat{\beta}, \quad i = 1,..., m_0 ; \qquad (6)$$

m_0 – число отсутствующих значений, которыми для простоты мы будем считать первые m_0 наблюдений, то будет минимизирована и обратится в нуль первая сумма в (5), так что

$$SS(\hat{\beta}_*, \hat{\gamma}) = \sum_{i=m_0+1}^{n} (y_i - x_i\hat{\beta}_*)^2 . \qquad (7)$$

Значит, $(\hat{\beta}_*, \hat{\gamma})$ минимизирует $SS(\beta,\gamma)$ и является оценкой наименьших квадратов (β,γ), получаемой из модели (3). Уравнение (5) означает также, что точная оценка наименьших квадратов отсутствующего значения y_i, т.е. $\hat{y}_i = x_i\hat{\beta}_*$, есть $\tilde{y}_i - \hat{\gamma}_i$ или в словесной формулировке: *прогноз i–го пропущенного значения методом наименьших квадратов есть начальное значение для i–го пропуска минус коэффициент для сопутствующей переменной i–го пропуска.*

В работе Бартлетта все $\tilde{y}_i$ приравниваются по этому методу к нулю, но с вычислительной точки зрения использование в качестве $\tilde{y}_i$ общего среднего более привлекательно и дает точную сумму квадратов отклонений от среднего (доказательство в книге [1]).

Метод имеет следующие преимущества.

- Он неитеративный, и, следовательно, снимает вопрос о сходимости.

- Если структура пропусков обладает вырожденностью (например, в том случае, когда нельзя оценить некоторые параметры, как при отсутствии всех значений для какой-то обработки), этот метод «предупреждает» исследователя, тогда как итеративные методы приводят к ответу, возможно, недопустимому.
- Метод дает правильные оценки и остаточные суммы квадратов, а также верные стандартные ошибки, суммы квадратов и F-критерии.

Хотя этот метод привлекателен в определенных отношениях, его часто нельзя реализовать непосредственно, потому что специализированные программы дисперсионного анализа могут не обладать возможностью вести обработку при многих сопутствующих переменных.

3.2. МЕТОД ЗАПОЛНЕНИЯ СРЕДНИМ

В пакете *Statistica* для заполнения пропусков в данных предусмотрена возможность замены по среднему значению. Это можно сделать в специализированном модуле по работе с данными Data Management при помощи команды Replace Missing Data by Means - подставляются средние присутствующих значений. Поэтому метод среднего включен в исследование как метод, наиболее часто используемый в статистических пакетах.

Метод безусловного среднего - самый простой вид заполнения. Он заключается в оценке отсутствующих значений y_{ij} средним $\bar{y}_j^{(j)}$ по присутствующим значениям переменной Y_j. Среднее наблюдаемых и подставленных значений равно $\bar{y}_j^{(j)}$ - оценке методом доступных наблюдений. Для равновероятного плана среднее популяции $\bar{Y}$ можно оценить средним присутствующих и подставленных значений $\bar{y}_j^{(j)}$, а именно $\sum_j n_j \bar{y}_j^{(j)} / \sum_j n_j$. Дисперсия наблюденных и подставленных значений равна $\left[(n^{(j)}-1)/(n-1)\right] s_{jj}^{(j)}$, где $s_{jj}^{(j)}$ - оценка дисперсии методом доступных наблюдений. При условии случайного отсутствия данных $s_{jj}^{(j)}$ - состоятельная оценка истинной дисперсии, так что выборочная дисперсия для данных после заполнения – заниженная в $(n^{(j)}-1)/(n-1)$ раз оценка дисперсии. Это занижение – естественное следствие заполнения пропусков средним (значением в центре распределения).

3.3. *RESAMPLING* МЕТОД

В 1977 году американским статистиком Бредли Эфроном был предложен метод «bootstrap» [2]. Первоначально этот метод возник как средство преодоления смещения, обусловленного выборкой, затем, начал широко применяться для работы с любыми статистическими задачами: проверка гипотезы о законах распределения случайных величин, регрессия, дисперсионный анализ или многомерная классификация. Основным преимуществом бутстреп-подхода является то, что он не нуждается в априорном знании закона вероятностного распределения исходных данных, а значит подходит для работы с любыми данными. Отличие бутстрепа от традиционных методов заключается в том, что он предполагает *многократную обработку* различных частей одних и тех же данных, как бы поворот их разными гранями, и сопоставление полученных таким образом результатов.

Разновидностью бутсреп-метода является сравнительно новый метод обработки статистических данных, называемый resampling. В данной работе *resampling* метод применяется для решения задачи заполнения пропусков в неполных данных, когда значения для заполнения пропущенных элементов выбираются случайным образом из исходного множества данных X_i. Значение для замены пропуска можно выбрать двумя способами: с возвращением и без возвращения. Будем использовать способ с возвращениями, когда раннее выбранное значение может участвовать в замене еще раз.

Рассмотрим применение *resampling* метода в ситуации, когда данные представлены множеством значений НОР СВ X и зависимым от них откликом Y, причем некоторые значения отклика отсутствуют. Предположим, что имеется некоторая случайная выборка, состоящая из независимых непрерывных СВ $X = \{X_1, X_2, ..., X_m\}$ и значения отклика *Y*. Причем выборка такая,

что некоторые значения отклика пропущены. Расположим данные в монотонную структуру. Тогда присутствующие значения будут Y_i , $i = 1..k$, пропуски окажутся в Y_i , $i = (k+1)..n$.

Для каждой из независимых величин $X_1, X_2, ..., X_m$ и отклика Y возможно получить некоторое выборочное множество полных наблюдений $H_i = \{X_{i1}, X_{i2}, ..., X_{in_i}, Y_i\}$. Предположим, что распределение СВ $\{X_i\}$ и Y неизвестно. Применение resampling-метода для задачи замены пропусков в данном случае может быть осуществлено двумя способами.

1 способ (resampling 1).

- Строится матрица полных наблюдений $H_{(m+1)xk} = \{X_1, X_2, ..., X_m, Y\}$, где k – число присутствующих наблюдений.
- Для каждого пропуска выбирают из выборки *H* случайным образом наблюдение, которым замещают пропущенное значение *Y* и соответствующие ему X_1, X_2, …, X_m. Для этого генерируется случайное число j = Rnd(), производится замена наблюдения с пропуском $\{X_{i1}, X_{i2}, ..., X_{im}, Y_i\}$, *i=(k+1)..n* на $H_j = \{X_{j1}, X_{j2}, ..., X_{jm}, Y_j\}$
- По данным, полученным при заполнении *resampling 1* методом, строится регрессионная модель, и находятся оценки $\hat{\beta}_i$ коэффициентов, i=1,..,m и свободного члена $\hat{\beta}_0$.

2 способ (resampling 2).

- По присутствующим наблюдениям строится регрессионная модель, и находятся оценки $\hat{\beta}_i$ коэффициентов, *i=1,..,k.*
- Находится оценка $\hat{Y}_i$ по регрессионной модели для *i=1,..,k.*
- Находится ошибка $\varepsilon_i = Y_i - \hat{Y}_i, i = 1..k$
- Для каждого пропуска, подставляя значения сопутствующих переменных X_1, X_2, …, X_m в полученное регрессионное уравнение, находим оценку $\hat{Y}_i$, *i=(k+1),..,n.*
- Значение, которым замещают пропуск, получают по формуле: $Y_i = \hat{Y}_i + \varepsilon_j, i = (k+1)..n$, где ε_j выбирается случайно из ранее рассчитанных ошибок (генерируется случайное число j = Rnd()).
- По данным, полученным после заполнения, строится регрессионная модель, и находятся оценки $\hat{\beta}_i$ коэффициентов, i=1,..,m и свободного члена $\hat{\beta}_0$.

Данные алгоритмы повторяют r раз и после этого находятся средние значения коэффициентов регрессионной модели, которые рассчитываются как:

$$\overline{\beta}_0 = \frac{\sum \hat{\beta}_0}{r}, \qquad \overline{\beta}_i = \frac{\sum \hat{\beta}_i}{r}, \; i=1,..,m.$$

Найденные значения $\overline{\beta}_0$ и $\overline{\beta}_i$, i=1,..,m, являются результатом применения *resampling*-методов.

Положительным фактором в пользу *resampling*-метода является повторное использование исходных данных, ведь увеличение числа подвыборок позволяет наиболее полно и информативно использовать исходную информацию. С другой стороны, число новой информации уменьшается для каждой новой подвыборки, так как увеличивается вероятность того, что данные элементы выборки были уже выбраны раньше – это основной недостаток метода.

4. Технология проведения эксперимента

Для оценки эффективности заполнения пропусков *resampling*-методом по сравнению с методами Бартлетта и средних, в ходе работы проводился следующий эксперимент: одни и те же данные с пропусками поочередно обрабатываются вышеупомянутыми методами – заполняются все пропуски. Затем анализируется регрессионная модель, находятся коэффициенты модели для данных, восстановленных различными методами. Полученные результаты сравниваются с «истинными» значениями коэффициентов (которыми мы считаем коэффициенты, полученные по полной модели без пропусков).

Следует выделить несколько основных этапов эксперимента.

- Для полных исходных данных оценим β-коэффициенты, стандартные ошибки оценок

коэффициентов, получим значения критерия Фишера, коэффициента множественной детерминации и скорректированного коэффициента множественной детерминации, проведем анализ остатков модели.

- В исходные полные данные случайным образом внесем пропуски.
- Пропущенные значения восстановим методами Бартлетта, методом средних и *resampling*-методом.
- Полученные при заполнении разными методами «полные» выборки вновь исследуем - построим регрессионную модель, оценим β-коэффициенты модели, стандартные ошибки оценок, значения критериев качества.
- На основании полученных результатов оценим эффективность заполнения пропусков *resampling*- методом.

5. Описание программного обеспечения

Разработано программное обеспечение, необходимое для проведения эксперимента и расчета результатов. Сам эксперимент и все необходимые расчеты в работе проводились в интегрированной среде статистического анализа и обработки данных Statistica, а также с помощью таблиц MS Excel. В качестве языков программирования были выбраны встроенные в пакет Statistica язык Statistica Basic и командный язык SCL (Statistical Command Language), которые позволяют пользователю расширить стандартные возможности системы Statistica, а также язык Visual Basic для написания макросов в пакете MS Excel.

Программное обеспечение реализует:

- генерацию случайных пропусков (misgener.stb);
- методы заполнения пропусков (resampling.stb bartlett.stb);
- расчет регрессионной модели и нахождение коэффициентов регрессионной модели (my_regressn.stb);
- анализ эффективности методов.

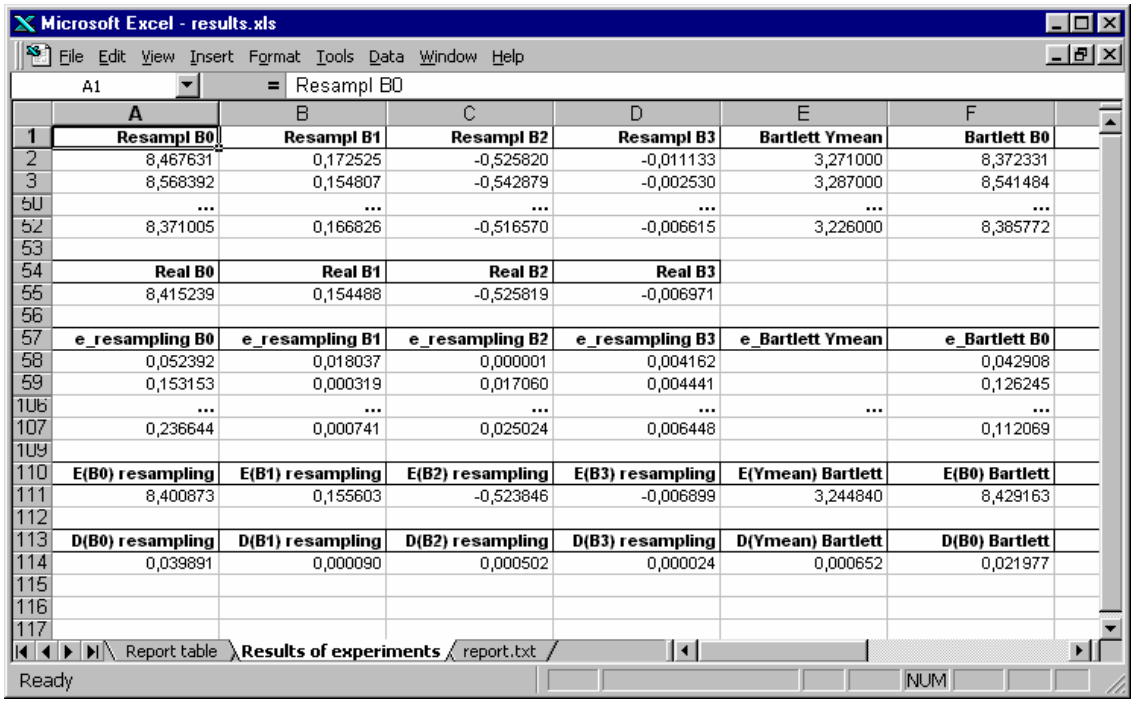
Программное обеспечение состоит из независимых между собой программных модулей, каждый из которых может быть запущен в системе Statistica как самостоятельная программа. Кроме того, написан макрос в MS Excel для представления результатов экспериментирования в удобном виде, расчета и формирования таблицы результатов.

6. Исходные данные для экспериментирования

Для задачи с пропусками в зависимой переменной в качестве исходных данных была взята выборка, предложенная для анализа Дрейпером и Смитом в [3]. В выборке представлены данные по трем независимым переменным и одному отклику. Данные представляют собой значения экспериментально управляемых переменных и среднего размера частиц: X_1 – скорость на входе на единицу длины (см/с/см); X_2 – окружная скорость ротора (см/с); X_3 – вязкость на входе (пуазы); Y – средний размер частиц μ.

Предлагаемая модель, основанная на теоретических представлениях, имеет вид: $Y = \beta_0 X_1^{\beta 1} X_2^{\beta 2} X_3^{\beta 3} \varepsilon$. Логарифмированием по основанию e. преобразуем модель в линейную форму. Проведем регрессионный анализ для исходных данных и найдем значения оценок β - коэффициентов, которые будем считать «истинными» (с помощью модуля *Multiple Regression* ***Statistica***). По критерию Стьюдента, третья сопутствующая переменная оказалась незначимой (t(28)=-0.8797), и ее можно исключить из модели. После исключения получили результаты: критерий Фишера увеличился по сравнению с предыдущим шагом (286<432), уменьшились ошибки оценок β_i-коэффициентов, увеличилось значение скорректированного квадрата коэффициента детерминации (adjusted RI = 0,9653097). На этом шаге мы получили значимые оценки β_i-коэффициентов, построенная регрессионная модель объясняет 96.5% разброса значений Y относительно среднего и имеет большой коэффициент Фишера. Таким образом, регрессионная модель, построенная по исходным данным, имеет вид:

$$\ln \hat{Y} = 8.448016 + 0.153934 \ln X_1 - 0.528304 \ln X_2 .$$



Microsoft Excel - results.xls

File Edit View Insert Format Tools Data Window Help

A1 = Resampl B0

	A	B	C	D	E	F
1	**Resampl B0**	**Resampl B1**	**Resampl B2**	**Resampl B3**	**Bartlett Ymean**	**Bartlett B0**
2	8,467631	0,172525	-0,525820	-0,011133	3,271000	8,372331
3	8,568392	0,154807	-0,542879	-0,002530	3,287000	8,541484
50	...	...	...	...	...	...
52	8,371005	0,166826	-0,516570	-0,006615	3,226000	8,385772
53						
54	**Real B0**	**Real B1**	**Real B2**	**Real B3**		
55	8,415239	0,154488	-0,525819	-0,006971		
56						
57	**e_resampling B0**	**e_resampling B1**	**e_resampling B2**	**e_resampling B3**	**e_Bartlett Ymean**	**e_Bartlett B0**
58	0,052392	0,018037	0,000001	0,004162		0,042908
59	0,153153	0,000319	0,017060	0,004441		0,126245
106	...	...	...	...	...	...
107	0,236644	0,000741	0,025024	0,006448		0,112069
109						
110	**E(B0) resampling**	**E(B1) resampling**	**E(B2) resampling**	**E(B3) resampling**	**E(Ymean) Bartlett**	**E(B0) Bartlett**
111	8,400873	0,155603	-0,523846	-0,006899	3,244840	8,429163
112						
113	**D(B0) resampling**	**D(B1) resampling**	**D(B2) resampling**	**D(B3) resampling**	**D(Ymean) Bartlett**	**D(B0) Bartlett**
114	0,039891	0,000090	0,000502	0,000024	0,000652	0,021977
115						
116						
117						

Report table | Results of experiments | report.txt

Ready NUM

Рисунок 2. Окно "Результаты экспериментов" содержит:

- оценки свободного члена $\hat{\beta}_0$ и коэффициентов регрессионной модели $\hat{\beta}_i, i := 1..m$ для методов *resampling*, Бартлетта и замены по среднему;
- истинные значения свободного члена $\hat{\beta}_0$ и коэффициентов модели $\hat{\beta}_i, i := 1..m$;
- расчетные (с помощью макроса) значения абсолютной ошибки $\varepsilon = |\hat{\beta}_i - \beta_i|$ для методов *resampling*, Барлетта и среднего;
- расчетные значения математического ожидания оценки коэффициентов $E(\hat{\beta}_i) = \overline{\beta}_i = \frac{1}{1} \cdot \sum_{j=1}^{1} \hat{\beta}_{ij}$ для методов *resampling,* Барлетта и среднего;
- расчетные значения дисперсии оценки коэффициентов $D(\hat{\beta}_i) = \sqrt{\frac{1}{1-1} \cdot \sum_{j=1}^{1} (\hat{\beta}_{ij} - \overline{\beta}_i)^2}$ для методов *resampling,* Барлетта и среднего.

7. Результаты экспериментирования

Для оценки эффективности применения метода *resampling* 2 находится регрессионная модель на основании данных, в которых пропуски заполнены по *resampling*-методу. Затем производится анализ полученных оценок свободного члена $\hat{\beta}_0$ и коэффициентов модели $\hat{\beta}_i, i := 1..m$ по сравнению с «истинными» значениями, найденными по исходной статистике, а также с оценками $\hat{\beta}_i, i := 1..m$, полученными методами Бартлетта и средних.

На рисунке 3 представлены результаты оценки коэффициентов по этим трем методам. Видно, что методы Бартлетта и *resampling* 2 дают примерно одинаковые результаты, намного превосходящие по качеству метод заполнения средними.

Для анализа сходимости результатов, полученных на основе *resampling* метода, повторим множество раз эксперимент по заполнению данных с одними и теми же пропусками и найдем величину абсолютной ошибки. Зададим следующие параметры экспериментов: процентное содержание пропусков в экспериментальных данных misPercent = 10 %, число *resampling* реализаций R = 100, 300, 500, 1000.

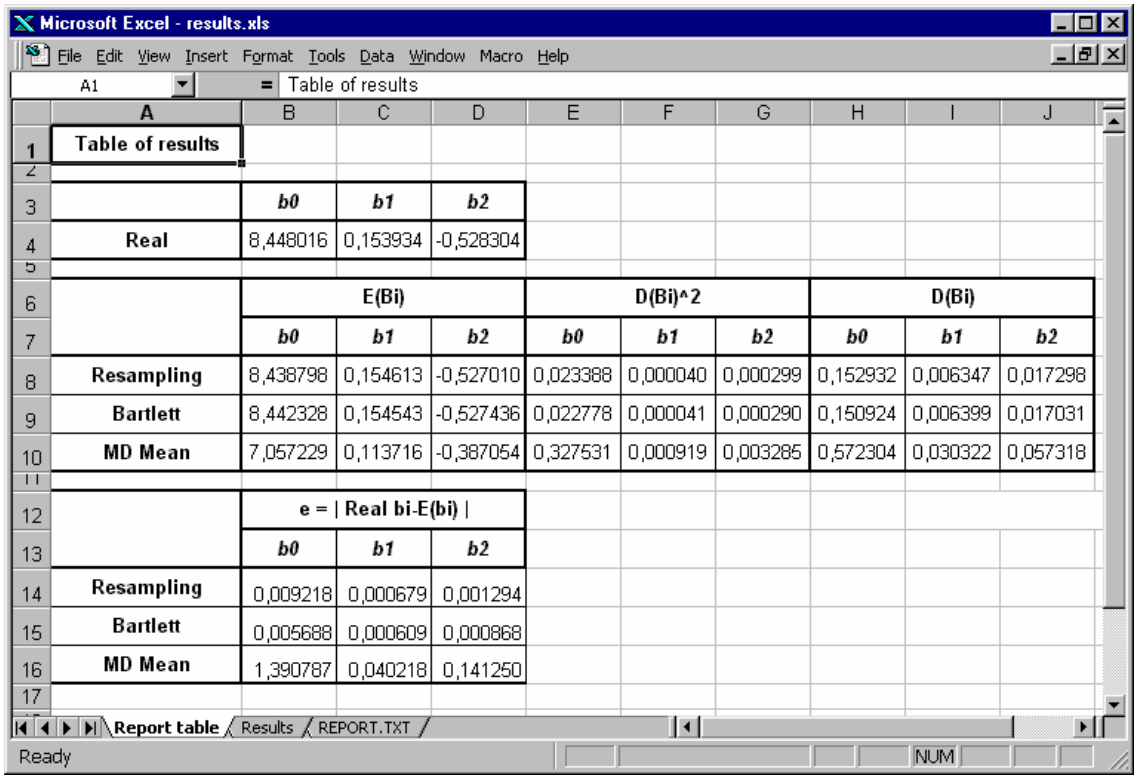


Рисунок 3. Окно работы макроса (содержит сводную таблицу результатов для оценки эффективности заполнения пропусков методами resampling, Bartlett и по среднему)

ТАБЛИЦА.

		b0	b1	b2
	Истинные значения	8,448016	0,153934	-0,528304
	Дисперсия	**0,042463**	**0,0009581**	**0,005423**
	$E(b_i)$ – м.Барлетта	8,453265	0,153120	-0,529207
	$D(b_i)$ – м. Барлетта	**0,098490**	**0,003891**	**0,010912**
	$E(b_i)$ - м. средним	7,682187	0,128941	-0,451341
	$D(b_i)$ - м. средним	**0,396834**	**0,021398**	**0,042271**
100	***$E(b_i)$ Resampling***	8,453057	0,153140	-0,529186
	$D(b_i)$ Resampling	**0,098703**	**0,003837**	**0,010917**
300	***$E(b_i)$ Resampling***	8,445874	0,154514	-0,527885
	$D(b_i)$ Resampling	**0,086763**	**0,003382**	**0,009498**
500	***$E(b_i)$ Resampling***	8,428843	0,154523	-0,525955
	$D(b_i)$ Resampling	**0,074847**	**0,003373**	**0,008206**
1000	***$E(b_i)$ Resampling***	8,437481	0,154979	-0,526734
	$D(b_i)$ Resampling	**0,070166**	**0,003620**	**0,008230**

Из результатов в таблице видно, что чем больше число реализаций, тем дисперсия коэффициентов, т.е. при числе реализаций $\rightarrow \infty$, значения оценок β -коэффициентов, полученные методом *resampling*, стремятся к «истинным». С небольшим преимуществом *resampling*-метод дает более точные оценки коэффициентов регрессии, чем метод Бартлетта, следовательно, его можно признать альтернативным методом заполнения пропусков.

8. Выводы

- Проведено исследование методов для заполнения пропусков в неполных данных.
- Рассмотрено использование *resampling* процедур для этой проблемы.
- Разработаны и реализованы программно алгоритмы методов Resampling, Бартлетта и замены по среднему.
- Оценена эффективность вышеупомянутых методов применительно к проблеме пропусков в переменной-отклике в регрессионных моделях
- Метод *resampling* можно считать альтернативным методу Бартлетта, являющимся более простым алгоритмически и дающим результаты того же качества.

Литература

[1] Литтл Р.Дж.А., Рубин Д.Б. (1991) *Статистический анализ данных с пропусками.* Финансы и статистика, Москва

[2] Эфрон Б. (1988) *Нетрадиционные методы многомерного статистического анализа.* Финансы и статистика, Москва

[3] Дрейпер Н., Смит Г. (1988) *Прикладной регрессионый анализ. Т.1,2.* Машиностроение, Москва

Received on the 1st of July 2002