**Роль эффектов взаимодействий в устранении смещений регрессионных коэффициентов, вызванных наличием в данных неслучайных пропусков**

*Жучкова Светлана, Ротмистров Алексей*

*05.11.2019*

Исследование, которое будет обсуждаться в ходе доклада, было посвящено применению метода индикаторной переменной к категориальным данным, содержащим пропущенные значения, в сравнении с анализом полных наблюдений. На практике обычно применяется именно второй из методов – он предполагает, что наблюдения, содержащие пропущенные значения, не участвуют в анализе. Метод индикаторной переменной предполагает создание в пару каждой исходной переменной, в которой встречаются пропуски, новой бинарной, или индикаторной, переменной. Она показывает наличие или отсутствие пропуска в соответствующей ей исходной переменной в соответствующем наблюдении. Этот метод выступает более новой альтернативой, которая позволяет анализировать все наблюдения и при этом, гипотетически, не приводит к искажению результатов анализа. Это последнее свойство было проверено нами посредством статистического эксперимента. В нём контролируются следующие факторы: механизм порождения пропущенных данных, их доля, спецификация регрессионной модели (с взаимодействиями и без). Согласно результатам, оба метода не приводят к смещениям оценок регрессионных коэффициентов; но существенно смещаются оценки их стандартных ошибок, а также величины R2, скорректированного R2 и F-статистики. Причём метод индикаторной переменной сработал заметно хуже. Мы объясняем смещения при использовании этого метода влиянием как раз индикаторной переменной, коэффициенты которой в большинстве моделей оказывались значимыми и довольно большими по модулю.

Эксперимент был проведён на симулированных данных. Симуляция данных для регрессионного моделирования начинается с симуляции предикторов и переменной, отвечающей за случайную ошибку, которые затем подставляются в регрессионное уравнение нужного семейства функций, включающее константу; на основании этого уравнения рассчитываются значения зависимой переменной. Для нашего эксперимента требовались предикторы согласно самой сложной упомянутой выше спецификации – с категориальным и континуальным предикторами и их взаимодействиями. Так, категориальный предиктор ‘A’ генерировался из дискретного равномерного распределения с тремя категориями. Континуальный предиктор ‘B’ и переменная, отвечающая за случайную ошибку, генерировались из нормальных распределений с разными средним и стандартным отклонением . На их основе – с учётом константы и эффектов взаимодействия – рассчитывались значения зависимой переменной ‘C’. Ещё одна переменная – вспомогательная бинарная ‘D’ – была создана для внесения на её основе пропусков по типу СП в категориальный предиктор ‘A’.

Размер совокупности составил 2000 наблюдений, поскольку в социологических исследованиях обычно используются выборки такого размера или несколько меньшего.

Общий дизайн эксперимента выглядел следующим образом:

1. Из исходной совокупности создаётся дочерняя выборка того же размера (2000 наблюдений) с повторениями.
2. На этой выборке из категориального предиктора ‘A’ создаются фиктивные переменные; рассчитываются значения всех параметров регрессионных моделей для каждой из трёх спецификаций параллельно. Итого, три набора значений параметров – для каждой из спецификаций, соответствующие истинными оценкам параметров модели для текущей дочерней выборки.
3. Из этой выборки параллельно формируются девять дублей посредством параллельного отбора 10, 25 и 50% наблюдений для внесения полностью случайных, случайных и неслучайных пропусков (далее – ПСП, СП и ПНП) в категориальный предиктор ‘A’. При этом для ПНП наблюдения отбираются случайным образом из тех наблюдений, которые имеют какое-либо из двух выбранных значений той же переменной ‘A’ (мы выбрали значения ‘1’ и ‘3’); для СП наблюдения отбираются случайным образом из тех наблюдений, которые имеют выбранное значение вспомогательной бинарной переменной ‘D’; для ПСП наблюдения отбираются полностью случайным образом (генератор случайных чисел работает на всей выборке).
4. В каждом из девяти дублей для отобранных наблюдений в недихотомизированный предиктор ‘A’ вносятся пропуски.
5. В каждом из девяти дублей пропуски перекодируются в новое значение ‘4’ (до этого в этой переменной присутствовали только категории ‘1’, ‘2’, ‘3’).
6. В каждом из девяти дублей из недихотомизированного предиктора ‘A’ создаются уже *четыре* бинарные фиктивные переменные (для каждого значения предиктора ‘A’) и рассчитываются значения всех параметров регрессионных моделей для каждой из трёх спецификаций параллельно. Итого, 27 наборов значений параметров (9 дублей \* 3 спецификации), соответствующих *методу индикаторной переменной.*
7. Параллельно в каждом из девяти дублей наблюдения, содержащие значение ‘4’ (бывший «пропуск») удаляются из анализа, а из недихотомизированного предиктора ‘A’ создаются *три* бинарные фиктивные переменные и рассчитываются значения всех параметров регрессионных моделей для каждой из трёх спецификаций параллельно. Итого, 27 наборов значений параметров (9 дублей \* 3 спецификации), соответствующих *анализу полных наблюдений.*
8. Шаги 1-7 повторяются ещё 1999 раз.

Для каждого из 54 наборов (27 после шага 6 и столько же после шага 7) оценённых параметров регрессионной модели на основании 2000 дочерних выборок рассчитываются средние арифметические значения каждого параметра в качестве его точечной оценки, а также интервальные оценки (верхняя и нижняя границы для 95-процентного доверительного интервала – это 2,5 и 97,5 перцентили соответственно).

В ходе этого исследования был обнаружен один любопытный вывод: выбор определённых правил для внесения неслучайных пропусков в совокупности с параметрами распределения имеющихся переменных могут также определять результаты применения выбранных методов обработки пропусков. Так, например, при выборе другого правила для внесения неслучайных пропусков в нашем эксперименте была обнаружена следующая закономерность: если пропуски вносятся в те категории переменной, среднее значение зависимой переменной в которых по своему знаку отличается от среднего по прочим категориям, то это приводит к значимому смещению коэффициента при континуальном предикторе (во второй спецификации). Однако если в модели присутствуют эффекты взаимодействия, то никакого смещения не происходит. Мы задались вопросом – почему? Каким образом эффекты взаимодействия нивелируют смещения, вызыванные неслучайными пропусками?

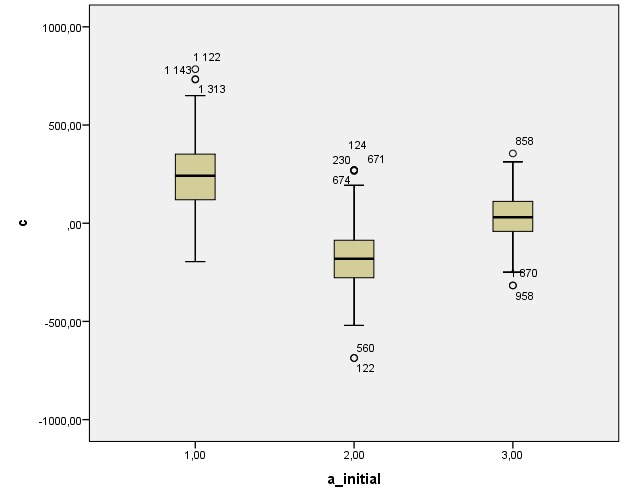


Рис. 1. Условное распределение зависимой переменной (С) по категориям исходной переменной А (на примере одной бутстрепированной выборки)

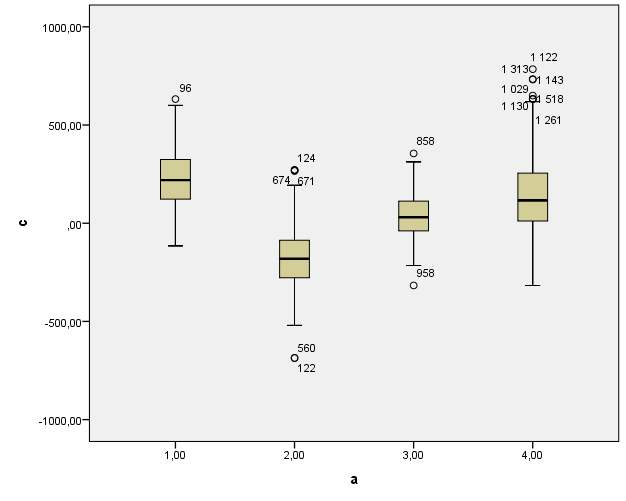


Рис. 2. Условное распределение зависимой переменной (С) по категориям итоговой переменной А (на примере одной бутстрепированной выборки)

Дело в том, что механизм генерирования миссингов MNAR (missing not at random) предполагал вероятностный «забор» наблюдений из 1 и 3 категорий категориального предиктора ‘А’. Эти категории имеют средний отклик положительный. А оставшаяся нетронутой категория 2 имеет средний отклик отрицательный. Поэтому средний отклик для немиссинговых значений сместился в отрицательную область – см. рис. 1-2.

Учитывая, что групповые средние отклики до и после «забора» миссингов не изменились значимо («забор» же был вероятностным), регрессионные коэффициенты фиктивных переменных, а также эффектов взаимодействия не сместились значимо – см. Таблицы 1-2, в которых представлены коэффициенты корреляции исходных и итоговых категорий переменной А (дамми-переменных) с зависимой и независимой переменными (коэффициенты корреляции являются прокси для регрессионных коэффициентов).

*Таблица 1*

**Коэффициенты корреляции категорий исходной переменной А**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlations** | | | | |
| a\_initial | | | b | c |
| 1,00 | b | Pearson Correlation | 1 | ,802\*\* |
| Sig. (2-tailed) |  | ,000 |
| N | 639 | 639 |
| c | Pearson Correlation | ,802\*\* | 1 |
| Sig. (2-tailed) | ,000 |  |
| N | 639 | 639 |
| 2,00 | b | Pearson Correlation | 1 | -,712\*\* |
| Sig. (2-tailed) |  | ,000 |
| N | 735 | 735 |
| c | Pearson Correlation | -,712\*\* | 1 |
| Sig. (2-tailed) | ,000 |  |
| N | 735 | 735 |
| 3,00 | b | Pearson Correlation | 1 | ,342\*\* |
| Sig. (2-tailed) |  | ,000 |
| N | 626 | 626 |
| c | Pearson Correlation | ,342\*\* | 1 |
| Sig. (2-tailed) | ,000 |  |
| N | 626 | 626 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | |

*Таблица 2*

**Коэффициенты корреляции категорий итоговой переменной А**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlations** | | | | |
| a | | | b | c |
| 1,00 | b | Pearson Correlation | 1 | ,754\*\* |
| Sig. (2-tailed) |  | ,000 |
| N | 121 | 121 |
| c | Pearson Correlation | ,754\*\* | 1 |
| Sig. (2-tailed) | ,000 |  |
| N | 121 | 121 |
| 2,00 | b | Pearson Correlation | 1 | -,712\*\* |
| Sig. (2-tailed) |  | ,000 |
| N | 735 | 735 |
| c | Pearson Correlation | -,712\*\* | 1 |
| Sig. (2-tailed) | ,000 |  |
| N | 735 | 735 |
| 3,00 | b | Pearson Correlation | 1 | ,373\*\* |
| Sig. (2-tailed) |  | ,000 |
| N | 144 | 144 |
| c | Pearson Correlation | ,373\*\* | 1 |
| Sig. (2-tailed) | ,000 |  |
| N | 144 | 144 |
| 4,00 | b | Pearson Correlation | 1 | ,518\*\* |
| Sig. (2-tailed) |  | ,000 |
| N | 1000 | 1000 |
| c | Pearson Correlation | ,518\*\* | 1 |
| Sig. (2-tailed) | ,000 |  |
| N | 1000 | 1000 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | |

Однако в спецификациях, в которых континуальный предиктор ‘B’ есть и при этом не встроен в группы (посредством эффектов взаимодействия), его коэффициенты смещаются значимо из-за того, что после «забора» миссингов оставшиеся значения безусловного отклика смещены в отрицательную область. Эффекты взаимодействия же фактически встраивают эту переменную обратно в группы другой переменной, выравнивая таким образом различия между исходной и итоговой ситуациями.