Мультиколлинеарность (Multicollinearity)

Мультиколлинеарность — это явление, при котором одна из входных переменных статистической модели (например, множественной линейной регрессии) линейно зависит от других входных переменных, т.е. между ними наблюдается сильная корреляция. В этой ситуации оценки коэффициентов (параметров) модели могут случайно и значительно изменяться даже при небольших изменениях в исходных данных, т.е. решение становится неустойчивым.

При этом возможны два случая:

1. **Полная коллинеарность** — имеет место, если между входными переменными присутствует функциональная зависимость (например если одна переменная — зарплата сотрудника в рублях, а другая — в долларах). Если модель содержит две входных переменных x_1 и x_2, то линейная функциональная зависимость между ними может иметь вид $x_2 = b \cdot x_1$, где b — константа. В этой ситуации оказывается, что в двумерном пространстве признаков вектор решения оказывается не единственным, а решение образует целую прямую, каждая точка которой представляет собой истинный вектор параметров модели. Такая модель принципиально неидентифицируема. Проблема полной коллинеарности может быть решена только путем соответствующей организации формирования выборки и отбора переменных.
2. **Мультиколлениарность** — возникает когда зависимость между входными переменными не функциональная, а статистическая, т.е. имеет место сильная корреляция. Если полная коллинеарность вызывает неопределенность значений параметров модели, то мультиколлинеарность приводит к неустойчивости их оценок, которая выражается в увеличении статистической неопределенности и росту их дисперсии. На практике, это приводит к тому что оценки могут сильно изменяться даже при незначительных изменениях в исходных данных.

Для пояснения сказанного рассмотрим модель множественной линейной регрессии с двумя переменными:

$y = a_0 + a_1 x_1 + a_2 x_2$.

Из теории метода наименьших квадратов известно, что

$\sigma^2_{a_1} \simeq \frac{1}{(1-r^2)}$,

т.е. дисперсия оценки параметра a_1 растет при увеличении коэффициента корреляции r между переменными. И когда $r \to 1$ дисперсия оценки стремится к бесконечности. Когда $r = 1$ между переменными x_1 и x_2 возникает функциональная зависимость и модель

становится неопределенной (имеет место полная коллинеараность).

Чтобы избежать проблем, связанных с мультиколлинеарностью при построении регрессионных моделей, ее наличие необходимо сначала обнаружить. Признаками мультколлинеарности могут быть:

1. Высокие стандартные ошибки оценок параметров модели.
2. Низкая значимость оценок параметров модели при том, что вся модель признается статистически значимой.
3. Значительные изменения оценок параметров модели при изменении в выборке.
4. В корреляционной матрице входных переменных присутствуют большие значения коэффициентов парной корреляции (0.7 и более).
5. Знаки коэффициентов регрессии противоречат бизнес-логике задачи.

В простейшем случае для решения проблемы мультиколлинерности можно попытаться исключить попадание в выборку зависимых признаков. Но этот метод не всегда приводит к желаемым результатам, поэтому на практике чаще используются различные методы декорреляции переменных, например, метод главных компонент. В результате вместо исходного набора признаков получается набор ортогональных, т.е. статистически независимых факторов. Недостатком здесь является проблема их интерпретации.

В Loginom существует инструменты для выявления мультиколлинеарности и борьбы с ней. Так, специализированный обработчик Корреляционный анализ позволяет производить расчет коэффициентов корреляции между признаками набора данных. В обработчике Факторный анализ можно производить декорреляцию признаков с помощью метода главных компонент.