Data Science.
Lectures. Week 1.

Quantitative Analysis Regression Diagnostics.

Мария Зыкова

28 сентября 2022 г.

Contents

1 Смещение вследствие пропущенных переменных(Omitted variable bias)

Пропущенная переменная — это переменная, которая имеет ненулевой коэффициент и не включена в модель.
Когда в модели линейной регрессии отсутствуют правильные переменные, результаты, скорее всего, приведут к неверным выводам, поскольку оценки, полученные методом наименьших квадратов(МНК), могут неточно отражать фактические данные.

Варианты влияния пропущенной переменной на оценки коэффициента наклона и остатки:

- Если пропущенная переменная не зависит от выбранных переменных, это приводит к тому, что члены ошибки больше истинных, однако ее влияние на оценки коэффициенты наклона незначительно.
- Если пропущенная переменная каким-либо образом коррелирует с другими переменными, учтенными в модели, в этом случае предположения МНК нарушены, поскольку она фактически включена в термин ошибки, таким образом, делая независимую переменную и термин ошибки коррелированными.
  Оценки МНК тогда уже не будут обладать правильными и полезными свойствами.

Проблема смещения пропущенной переменной возникает независимо от размера выборки и приводит к непоследовательности оценок МНК.

Корреляция между пропущенной переменной и независимой переменной определяет размер смещения.

Множественный регрессионный анализ используется для устранения смещения пропущенной переменной, поскольку он может оценить влияние одной независимой переменной на зависимую переменную, сохраняя все остальные переменные постоянными.

2 Компромисс между смещением и дисперсией

Если добавить в модель переменную, которая объясняет не существенную часть изменения зависимой переменной или совсем не объясняет, то оценки коэффициента МНК будут менее точными.

Есть двойственность выбора между пропуском подходящей переменной и ее включением.

– С одной стороны, оценки коэффициента наклона показывают смещение, когда используется меньше переменных (из-за пропущенных), но являются более точными.

– С другой, включение нерелевантных переменных могло бы уменьшить смещение, но оценки становятся менее точными из-за увеличения числа переменных.

Такой выбор в конечном счете является компромиссом между смещением и дисперсией.

Чтобы достигнуть точности, необходимо увеличить размер выборки.

Есть два подхода к тому, как определить кол-во переменных в модели.

1. **от общего к частному.**
   Начинается с указания большой модели, включающей все соответствующие переменные, и удаления статистически незначимых переменных(т.е. тех, что имеют коэффициент наклона равным 0), начиная с той, которая имеет наименьшее абсолютное значение статистики.

2. **подход кросс валидации**(перекрестная проверка)

   Осуществляется в следующие этапы:

   - разбиение данных на две части. На одной оцениваются параметры регрессии, вторая используется для проверки, где считаются сумма квадратов остатков
   - перебор подмножеств переменных таким образом, чтобы сумма остаточных квадратов для оставшихся данных была наименьшей

   Т.е. на одной части данных определяется модель, а на другой – точность прогноза.
   Тем самым выбираются переменные, которые показывают лучшие прогнозы.

3 Гомоскедастичность и гетероскедастичность в множественной регрессии.

Гомоскедастичность относится к условию, при котором дисперсия члена ошибки постоянна для всех независимых переменных, $Var(\epsilon_i|X_1, ..., X_k) = \sigma^2$
Дисперсия членов ошибки варьируется в зависимости от выборки, т.е. непостоянна.

Как правило, гетероскедастичность не влияет на состоятельность оценок полученных по методу наименьших квадратов, но изменяет их распределение и соответствующие оценки стандартной ошибки.
Требуются специальные оценки для случая гетероскедастичности.

Как проверить данные?

1. график остатков(графический анализ)
2. Тест Эйкера-Уайта:

   - построение регрессии квадратов остатков по членам квадратичной формы(комбинации независимых переменных степени не больше двух, т.е. $X_1, X_2, X_1^2, X_2^2, X_1 * X_2$
   - применение критерия - все ли коэффициенты наклона равны нулю. Если нулевая гипотеза отвергается, мы имеем гетероскедастичность, иначе гомоскедастичность.

Несмотря на то, что свойство гетероскедастичности часто встречается, обычно оно не влияет на состоятельность оценок, но оно влияет на их распределение и на применимость соответствующих критериев доверительных интервалов. В этом случае необходимо использовать специализированные оценки для случая гетероскедастичности и соответствующие специализированные критерии доверительных интервалов.

4 Мультиколлинеарность

Мультиколлинеарность - свойство, когда две или более независимых переменных или линейные комбинации независимых переменных в множественной регрессии сильно коррелируют друг с другом.

- Если независимые переменные – не случайные величины, мультиколлинеарность означает, что какая-то из этих переменных может быть выражена в виде линейной комбинации других.
- Если независимые переменные – случайные величины, мультиколлинеарность означает, что какие-то из переменных оказались сильно коррелированными.

Если одна из независимых переменных является линейной комбинацией других независимых переменных (или корреляцию $\pm$ 1), то модель демонстрирует **полную мультиколлинеарность.**

- При полной мультиколлинеарности оценки метода наименьших квадратов(МНК) нестабильны или не могут быть получены(уравнение не будет иметь решения и таким образом ничего оценить не получится)

Например, если независимые переменные являются фиктивными переменными и каждое наблюдение связано только с одним классом, регрессия будет демонстрировать полную мультиколлинеарность; в этом случае следует исключить одну их фиктивных переменных (т.е. использовать n-1 фиктивных переменных).

Неполная мультиколлинеарность возникает, когда две или более независимых переменных имеют высокую корреляцию, но менее чем полную корреляцию.

- В результате неполной мультиколлинеарности существует большая вероятность неправильного заключения, что переменная не является статистически значимой.

Обычно мультиколлинеарность обнаруживается, когда оценки по критерию Стьюдента показывают, что ни один из вычисленных коэффициентов существенно не отличается от нуля, в то время как R^2 высок(т.е. модель хорошо описывает изменения зависимой переменной)

Высокая корреляция между независимыми переменными иногда рассматривается как признак мультиколлинеарности, однако низкая корреляция между независимыми переменными не обязательно указывает на отсутствие мультиколлинеарности.

Наиболее распространенным методом исправления мультиколлинеарности является исключение одной из коррелированных переменных; недостатком здесь является то, что нелегко найти, какие переменные необходимо исключить.

5 Выбросы.

Выбросы(Outliers) — это значения, которые при удалении из выборки приводят к значительным изменениям в оценочных коэффициентах.

Для определения выбросов можно использовать расстояние Кука, которое выражается следующей формулой:

$$d_j = \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \widehat{Y}_i^0\right)^2}{ks^2} \tag{1}$$

где s^2 оценка отклонения ошибки от модели, которая использует все наблюдения,
$\widehat{Y}_i$ - значение регрессии в X_i;
$\widehat{Y}_i^0$ значение регрессии в X_i, когда значение X_j отбрасывается.
k – кол-во независимых переменных.

Если значение расстояния больше 1, значение Y_j считается выбросом.

Выбросы представляют собой проблему, поскольку они значительно изменяют оценочные значения и ухудшают их свойства.

Способы устранения проблемы выбросов:

– увеличение количества наблюдений. Если число выбросов не возрастает, то постепенно точность оценок будет улучшаться, а эффект от выбросов уменьшатся

– использование робастных оценок.

– удаление всех выбросов из наблюдений.

Обратите внимание, что из-за рисков выбросы могут представлять наибольший интерес для аналитики.