Data Science.
Lectures. Week 1.
Regression with Multiple Explanatory Variables. Задача регрессии с несколькими независимыми переменными.

Grishina Ekaterina

Contents

1 Основные положения регрессии с несколькими переменными

Множественная регрессия применяется, когда зависимая переменная зависит от нескольких независимых переменных. Общий вид множественной регрессии:

$$Y_i = f(T, X_{ij}, \epsilon_i),$$

где T - набор параметров, X_{ij} - наблюдения независимых переменных и ϵ_i - случайные ошибки. В отличии от простой одномерной регрессии $Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + ... + B_k X_{ki} + \epsilon_i$, где Y_i - i-ое наблюдение независимой переменной, B_0 - свободный член и B_i - коэффициенты наклона, в множественной регрессии X_{ij} - i-ое наблюдение j-ой независимой переменной. Как и в случае одномерной регрессии, f определяет модель, и мы ее задаем сами. X_{ij} может быть как случайной величиной, так и неслучайной. Многомерная линейная регрессия означает, что функция f линейна относительно параметров T и наблюдений независимых переменных X_{ij}.

2 МНК

МНК, как и в одномерии, заключается в минимизации функционала остатков $\sum e_i^2$. В случае с одной независимой переменной мы получали явное выражение для соответсвующих оценок коэффициентов; для многомерния также можно получить подобные оценки, однако они довольно громоздкие, и их вычисляют отдельно в специальных пакетах для статистического анализа.

$$\hat{Y_i} = B_0 + B_1 X_{1i} + B_2 X_{2i} + ... + B_k X_{ki},$$

$$e_i = Y_i - \hat{Y_i}$$

3 Интерпретация коэффициентов

Коэффициент B_0 - значение зависимой переменной, если все независимые переменные равны нулю (как и в одномерном случае).

Каждый из коэффициентов наклона имеет следующий смысл: B_i показывает, насколько изменится зависимая переменная Y_i при единичном изменении независимой переменной при условии, что все остальные независимые переменные являются постоянными. В этой интерпретации предполагается, что все независимые переменные действительно не завяисят друг от друга.

Пример: Пусть регрессионное уравнение имеет вид $Y = 1 + 2X_1 + 3X_2$. Тогда значение зависимой переменной Y равно единице при условии, что все независимые переменные X_1 и X_2 равны нулю. Коэффициент наклона 2 при X_1 означает, что Y изменится на две единицы, если X_1 изменится на одну единицу, а X_2 останется неизменным.

4 Гомоскедастичность и гетероскедастичность

Гомоскедастичность означает, что условная дисперсия ϵ_i при условии, что независимые переменные X_1, ... X_k постоянные равна σ^2: $var(\epsilon_i | X_1, ...X_k) = \sigma^2$ при всех i.

Гетероскедастичность означает, что дисперсия ошибок зависит от выборки. Разделяют на условную и безусловную. Безусловная гетероскедатичность подразумевает, что дисперсия ϵ_i зависит от номера наблюдения, а условная означает, что ошибки ϵ_i зависят не только от номера наблюдений, но и от значений зависимых переменных.

5 Оценка качества регрессии

Точно также, как и в случае простой линейной регрессии определяется сумма квадратов остатков $\sum e_i^2$, назваемая SSR (sum of squared residuals) - это та величина, которая минимизируется в МНК.

Стандартная ошибка регрессии SER строится следущим образом:

$$SER = \sqrt{\frac{SSR}{n-k-1}},$$

где n - количество наблюдений, k - число независимых переменных. Чем меньше SER, тем лучше выбранная модель описывает зависимость.

Величина R^2 (коэффициент детерминации) расчитывается по формуле:

$$R^2 = \frac{\sum(\bar{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

Как известно, значения лежат в диапазоне от 0 до 1, и для одномерной регрессии справделиво, что, чем ближе значение коэффициента к единице, тем лучше точки ложатся на прямой, и тем лучше работает наша модель. Для многомерия имеет место зависимость R^2 от числа независимых переменных: чем их больше, тем ближе величина к единице. Это не есть хорошо: так, в множественной регресии в случае $R^2 = 1$ мы получим, что все точки легли на одну многомерную плоскость, а случайные ошибки равны нулю - мы включили их в нашу модель. Таким образом, вместо отбрасывания случайных ошибок мы их включаем в модель, и прогноз по этой модели будет плохим.

Поэтому вводят другую величину $R_a^2 = 1 - (\frac{n-1}{n-k-1}(1-R^2))$, где n - размер выборки (число наблюдений), k - число независимых переменных. $R_a^2 \leq R^2$; R_a^2 может быть меньше нуля, если R_a^2 достаточно мал. При добавлении переменных R_a^2 может уменьшиться - «штраф» за использование слишком большого числа переменных.

6 Допущения модели множественной линейной регрессии

Предполагается, что имеется линейная зависимость между зависимой и независимыми переменными. В случае, если независимые переменные не являются случайными величинами, никакие из них не должны быть линейно зависимыми.

Условное математическое ожидание случайных ошибок при всех независимых переменных равно нулю. Условная дисперсия ошибки константна для каждого наблюдений при всех независимых переменных. Случайные ошибки должны быть независимыми и одинаково распределенными, обычно полагается, что они имеют нормальное распределение.

7 Проверка гипотез коэффициентов регрессии

Проверить значимость коэффициентов регрессии можно с помощью критерия Стьюдента. Нуль-гипотеза: $B_j = 0$. Альтернативная гипотеза: $B_j \neq 0$. t-статистика, используемая для проверки значимости отдельных коэффициентов в многомерной регрессии рассчитывается по той же формуле, что и одномерная:

$$t = \frac{b_j - B_j}{S_{b_j}},$$

где $b - j$ - оценка из МНК, B_j - предполагаемое значение S_{b_j} - стандартная ошибка b_j.

Это значение нужно сравнить с верхним и нижним критическими значениями распределения Стьюдента с n-k-1 степенями свободы, где n - количество наблюдений, k - число независимых переменных. Если значение t-статистики оказалось больше верхнего критического или меньше нижнего критического значения, гипотеза отвергается. Иначе отвергается нуль-гипотеза.

Тот же самый вывод может быть сделан по p-значению - наименьшему уровню значимости, при котором нулевая гипотеза может быть отвергнута. Сравниваем p-значение с уровнем значимости, и если p-значение оказалось меньше уровня значимости нулевую гипотезу отвергают, иначе - не отвергают.

Пример: Рассчитанное p-значение для коэффициента y-пересечения равно 0.031. Определите, является ли оно значимым на уровне значимости 2%?

а) Коэффициент сильно отличается от нуля

б) Коэффициент незначительно отличается от нуля

p-значение больше двух процентов ($0.02 < 0.031$), значит не отвергаем нуль-гипотезу о том, что коэффициент равен нулю. Ответ б).

8 Построение доверительных интервалов для коэффициентов регрессии

Доверительный интервал для коэффициента b_j: $b_j \pm t_{critical} s_{b_j}$, где $t_{critical}$ - критическое значение, полученное из распределения Стьюдента с числом степеней свободы n-k-1, s_{b_j} - стандартная ошибка коэффициента b_j.

9 Прогноз

Прогнозируемое значение вычисляется следующим образом: $\hat{Y_i} = b_0 + b_1 \hat{X_{1i}} + b_2 \hat{X_{2i}} + ... + b_k \hat{X_{ki}}$, где $\hat{Y_i}$ - прогнозируемое значение зависимой переменной, b_j - оценки коэффициентов регрессии, X_{ij} - прогноз j-ой независимой переменной. Для пронозирования используют все оценки коэффициентов регрессии независимо от их статистической значимости.

10 Совместная проверка гипотез

Совместная гипотеза проверяет два и более коэффициентов одновременно. Так, нулевой гипотезой может быть предположение, что $b_1 = b_2 = 0$, а альтернативной предположение о том, что хотя бы одно из b_1 и b_2 отлично от нуля. Использование совместной проверки предпочтительно в определенных сценариях, поскольку тестирование коэффициентов по отдельности приводит к большей вероятности отклонения нулевой гипотезы. F-статистика - надежный метод для совместной проверки гипотез, особенно, если независимые переменные коррелируют.

11 F-статистика

F-статистика рассчитывается по формуле:

$$F = \frac{\frac{ESS}{k}}{\frac{SSR}{n-k-1}},$$

где $ESS = \sum(\bar{Y_i} - \bar{Y})^2$ (explained sum of squares), $SSR = \sum(Y_i - \bar{Y})^2$ (sum of squared residuals), n - количество наблюдений, k - число независимых переменных.

Вычисленное значение статистики сравнивается с критическим значением распределения Фишера (F-распределение) с n-k-1 степенями свободы. Если полученное значение больше F-критического значения, нулевую гипотезу отвергают.

Заметим, что данный критерий является односторонним. Он проверяет, равны ли нулю все коэффициенты наклона одновременно.

12 Specification bias

Величина specification bias показывает, насколько мы бы ошиблись, если вместо множественной регрессии применили бы одномерную регрессию. Так, для одного и того же коэффициента мы получим различные значения в многомерном и одномерном случае. Разность между этими значениями называется specification bias. Если эта величина относительно мала, то это означает, что зависимыя переменная сильно зависит от выбранной независимой переменной и практически не зависит от остальных независимых переменных.

13 Ограничения на коэффициенты

Часто при построении регрессионных моделей приходится накладывать ограничения на то, какими могут быть коэффициенты. Например, если у нас есть уравнение множественной двумерной регрессии $Y_i = B_0 + B_1X_{1i} + B_2X_{2i}$, мы ее можем превратить в одномерную, положив $B_2 = 0$.

Если мы накладываем ограничения, то для нахождения оценок коэффициентов применяется МНК с ограничениями. Величина R^2 рассчитывается по тем же самым формулам, нок ней приписывается индекс r: R^2_r. Для R^2, рассчитанного без ограничений, используем индекс ur (unrestricted): R^2_{ur}.

Возникает вопрос: является ли существенным наложенное ограничение? Для проверки может использоваться F-статистика:

$$F = \frac{\frac{R^2_{ur} - R^2_r}{m}}{\frac{1-R^2_{ur}}{n-k_{ur}-1}},$$

где m - число ограничений.

Пусть есть ограничение на коэффициенты: $B_1 = B_2$. Можно воспользоваться критерием Фишера и рассчитать F-статистику. Можно действовать по-другому: возьмем уравнение регрессии $Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + \epsilon_i$, добавим и вычтем слагаемое B_2X_{1i}, получим $Y_i = B_0 + (B_1 - B_2)X_{1i} + B_2(X_{1i} + X_{2i}) + \epsilon_i$. Так, если $B_1 = B_2$, получим уравнение $Y_i = B_0 + B_2(X_{1i} + X_{2i}) + \epsilon_i$, коэффициент при X_{1i} должен быть статистически незначимым. Перейдем к другим независимым переменным - X_{1i} и $(X_{1i} + X_{2i})$, построим множественную линейную регрессию и проверим значимость коэффициента при X_{1i}. Если этот коэффициент окажется статистичеки незначимым, то есть гипотеза о том, что $B_1 = B_2$ не отвергается, то в исходном уравнении регрессии не отвергается та же самая гипотеза. Иначе используем МНК с учетом данного ограничения.