Тема 3. Множественная регрессия и корреляция

Вопрос 1. «Понятие множественной регрессии и корреляции»

Понятие модели множественной регрессии

Модель множественной регрессии – это уравнение, отражающее корреляционную связь между результатом и несколькими факторами.

Цель множественной регрессии

построить модель с несколькими факторами и определить при этом влияние каждого фактора в отдельности, а также их совместное воздействие на результат.

Линейная функция множественной регрессии и корреляции

Линейная функция множественной регрессии и корреляции имеет вид:

$$y = a + b_1 x_1 + b_2 x_2 + b_i x_i + e.$$

Нелинейная функция множественной регрессии и корреляции

В качестве нелинейной функции множественной регрессии и корреляции чаще всего выбирают показательную и степенную.

Показательная функция

Показательная функция имеет вид:

$$y = a + b_1^{x1} + b_2^{x2} + b_i^{xi} + e$$

Степенная функция

Степенная функция имеет вид:

$$y = a + x_1^{b1} + x_2^{b2} + x_i^{bi} + e.$$

Условия для проведения анализа методом множественной регрессии и корреляции

1. При проведении анализа методом множественной регрессии и корреляции предполагается, что наблюдения, на основе которых он проводится, были получены по однородной совокупности единиц.

То есть механизм воздействия факторов на результат должен быть примерно одинаков на разных единицах совокупности.

Условия для проведения анализа методом множественной регрессии и корреляции

2. Результат и факторы – это количественные показатели.

В простейшем случае считают, что для них нет границ изменения, то есть они принадлежат интервалу (- ∞; + ∞) и не являются случайными.

Условия для проведения анализа методом множественной регрессии и корреляции

3. При построении эконометрической модели предполагается, что факторы оказывают влияние на результат, причем влияние отдельного фактора не зависит от влияния других факторов.

В противном случае изменение значения какого-либо фактора окажет на результат, как прямое воздействие, так и косвенное – через другие факторы.

Это может привести к ошибкам в интерпретации результатов исследования.

Интеркорреляция и мультиколлинеарность

Корреляционная связь, которая существует между двумя факторами, называется интеркорреляцией.

Соответственно, корреляционная связь, существующая между несколькими факторами, называет мультиколлинеарностью.

Интеркорреляция и мультиколлинеарность

Существование корреляционной связи между факторами выявляется с помощью коэффициентов корреляции, которые принято записывать в виде матрицы.

Коэффициент корреляции фактора с самим собой равен единице, а коэффициент корреляции первого фактора со вторым фактором равен коэффициенту корреляции второго фактора с первым.

Поэтому матрица является симметричной, в ней указывают только главную диагональ и элементы под ней.

Интеркорреляция и мультиколлинеарность

Наличие мультиколлинеарности подтверждается определителями матрицы.

Если связь между факторами полностью отсутствует, то недиагональные элементы матрицы будут равны нулю, а определители матрицы – единице.

При обнаружении функциональной (очень тесной) связи между факторами определитель матрицы будет близок к нулю.

Вопрос 2. «Правила отбора факторов в модели множественной регрессии и корреляции»

Отбор факторов

Несмотря на то, что теоретически множественная регрессионная модель позволяет учесть любое число факторов, практически в этом нет необходимости.

Отбор факторов производится на основе качественного теоретико-экономического анализа.

Отбор факторов

Однако теоретический анализ часто не позволяет однозначно ответить на вопрос о количественной взаимосвязи рассматриваемых признаков и целесообразности включения фактора в модель.

Поэтому отбор факторов производится в два этапа: сначала отбираются факторы исходя из сути проблемы; затем, на основе матрицы коэффициентов (индексов) корреляции и определения t-статистики Стьюдента для параметров регрессии.

Правила включения в модель факторов

Включаемые в модель множественной регрессии факторы должны объяснять вариацию зависимой переменной.

При построении модели с набором ряда факторов, обязательно следует рассчитать коэффициент (индекс) детерминации (R^2), который зафиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии ряда факторов.

Тогда влияние других, неучтенных в модели факторов, оценивается как (1 - R^2) с соответствующей остаточной дисперсией.

Правила включения в модель факторов

При включении в модель дополнительного фактора коэффициент (индекс) детерминации должен возрастать, а остаточная дисперсия уменьшаться.

Если этого не происходит и коэффициент (индекс) детерминации с остаточной дисперсией до и после включения фактора не отличаются друг от друга, то включаемый в модель дополнительный фактор не улучшает модель и является лишним.

Правила включения в модель факторов

Насыщение модели лишними факторами не только не снижает величину остаточной дисперсии и не увеличивает коэффициент (индекс) детерминации, но и приводит к статистической незначимости параметров регрессии по t- критерию Стьюдента.

Правила включения в модель факторов

Множественная регрессия характеризуется наличием достаточно большого количества факторов.

При этом отсутствует возможность выделить из них наиболее значимые, подлежащие включению в модель регрессии.

В таких случаях принято рассматривать несколько моделей с разным составом факторов.

Наилучшей выбирается модель, имеющая значимые параметры и максимальный показатель тесноты связи.

Четыре метода подбора факторов при построении модели

- метод последовательного включения факторов
- метод исключения факторов из модели
- шаговый регрессионный анализ
- ступенчатый регрессионный анализ

Метод последовательного включения факторов

При использовании метода последовательного включения факторов сначала должна быть построена модель с фактором, который наиболее тесно связан с результатом.

Затем, поочередно добавляются другие факторы.

После включения каждого фактора обязательно оценивается целесообразность включения нового фактора с точки зрения сокращения остаточной дисперсии.

Метод исключения факторов

Использование метода исключения факторов предполагает, что сначала строится модель с максимально большим количеством факторов, из которой поочередно исключаются незначимые факторы до тех пор, пока модель не будет иметь только значимые параметры при факторах.

Шаговый регрессионный анализ

Шаговый регрессионный анализ является преобразованием метода последовательного включения факторов.

Построение модели начинается с расчета параметров уравнения парной регрессии с фактором, который наиболее тесно связан с результатом.

Добавление каждого нового фактора сопровождается не только оценкой значимости включения данного фактора, но и проверкой значимости влияния на результат факторов, уже включенных в модель.

Выявленные незначимые факторы исключаются из модели.

Завершение процесса происходит тогда, когда добавление нового фактора не приведет к заметному улучшению качества модели.

Ступенчатый регрессионный анализ

Ступенчатый регрессионный анализ начинается с построения уравнения парной регрессии с наиболее значимым по степени влияния на результат фактором.

Затем по полученной модели находят случайные остатки ε.

По причине того, что эти остатки отражают влияние факторов, не включенных в уравнение регрессии, следует построить уравнение зависимости случайного остатка ε от следующего по степени влияния на результат фактора.

Данная процедура повторяется до тех пор, пока вновь полученное уравнение регрессии является значимым.

Этот метод является наиболее простым, но не достаточно точным, так как не учитывает взаимосвязь факторов.

Фиктивные переменные и модель бинарного выбора

Показатели, выбранные в качестве результативного признака и фактора, иногда могут быть неколичественными переменными. В случае если неколичественной переменной является фактор, то она называется фиктивной переменной.

Если неколичественной переменной является результативный признак, то такую модель принято называть моделью бинарного выбора.

Модели с усеченными и цензурированными данными

Чаще всего в моделях результативный признак является количественной переменной, однако его значения могут быть ограничены определенным интервалом.

Для отражения этой особенности существует два типа моделей: модели с усеченными данными и модели с цензурированными данными.

Модель с усеченными данными

При усеченной выборке наблюдения производятся не над всей статистической совокупностью, а над ее частью, для которой свойственно попадание значения результативного признака в определенный числовой интервал.

Модель с цензурированными данными

Цензурированная выборка представляет собой данные наблюдения над всей статистической совокупностью, но в силу каких-либо причин значениям результативного признака, меньшим или большим определенной числовой границы, присваивается значение, равное этой границе.

Частным случаем модели с цензурированными данными является *tobit-модель*.

Вопрос 3. «Показатели тесноты и силы связи между результативным признаком и факторами в уравнении множественной регрессии»

Коэффициенты (индексы) корреляции (детерминации)

В парной линейной регрессии показатели тесноты связи называются коэффициентами корреляции (детерминации), в парной нелинейной регрессии – индексы корреляции (детерминации), а в множественной регрессии – коэффициенты (индексы) корреляции (детерминации).

Коэффициенты (индексы) корреляции

Формулы для расчета коэффициентов (индексов) корреляции, при наличии двух факторов имеют вид:

$$r_{yx_1} = \frac{\overline{X_1 * Y} - \bar{X}_1 * \bar{Y}}{\sqrt{\left(\overline{X_1^2} - (\bar{X}_1^2)\right)} * \sqrt{\left(\overline{Y^2} - (\bar{Y}^2)\right)}}$$

$$r_{yx_2} = \frac{\overline{X_2 * Y} - \bar{Y} * \overline{X_2}}{\sqrt{\left(\overline{X_2^2} - (\bar{X}_2^2)\right)} * \sqrt{\left(\overline{Y^2} - (\bar{Y}^2)\right)}}$$

$$r_{x_1x_2} = \frac{\overline{X_1 * X_2} - \overline{X_1} * \overline{X_2}}{\sqrt{\left(\overline{X_1^2} - (\bar{X}_1^2)\right)} * \sqrt{\left(\overline{X_2^2} - (\bar{X}_2^2)\right)}}$$

Первая формула (первый фактор)

$$r_{yx_1} = \frac{\overline{X_1 * Y} - \bar{X}_1 * \bar{Y}}{\sqrt{\left(\overline{X_1^2} - (\bar{X}_1^2)\right)} * \sqrt{\left(\overline{Y^2} - (\bar{Y}^2)\right)}}$$

Вторая формула (второй фактор)

$$r_{yx_2} = \frac{\overline{X_2 * Y} - \bar{Y} * \overline{X_2}}{\sqrt{\left(\overline{X_2^2} - (\bar{X}_2^2)\right)} * \sqrt{\left(\overline{Y^2} - (\bar{Y}^2)\right)}}$$

Третья формула (два фактора)

$$r_{x_1 x_2} = \frac{\overline{X_1 * X_2} - \overline{X_1} * \overline{X_2}}{\sqrt{\left(\overline{X_1^2} - (\overline{X_1^2})\right)} * \sqrt{\left(\overline{X_2^2} - (\overline{X_2^2})\right)}}$$

Совокупный коэффициент (индекс) множественной корреляции

$$R_{YX_1X_2} = \sqrt{\frac{r_{YX_1}^2 + r_{YX_2}^2 - 2 * r_{YX_1} * r_{YX_2} * r_{X_1X_2}}{1 - r_{X_1X_2}^2}}$$

Совокупный коэффициент (индекс) множественной корреляции

$$R_{YX_1X_2} = \sqrt{\frac{r_{YX_1}^2 + r_{YX_2}^2 - 2 * r_{YX_1} * r_{YX_2} * r_{X_1X_2}}{1 - r_{X_1X_2}^2}}$$

Интерпретация значений коэффициентов (индексов) корреляции

0,1- 0,3- слабая связь
0,3-0,5 – умеренная связь
0,5-0,7- заметная связь
0,7-0,9- тесная связь
0,9-0,99- весьма тесная

Коэффициент (индекс) детерминации

Коэффициент (индекс) детерминации определяется возведением в квадрат коэффициента (индекса) корреляции.

Коэффициент эластичности

Ввиду того, что величины абсолютных показателей силы связи определяются единицами измерения факторов, они не являются сравнимыми между собой.

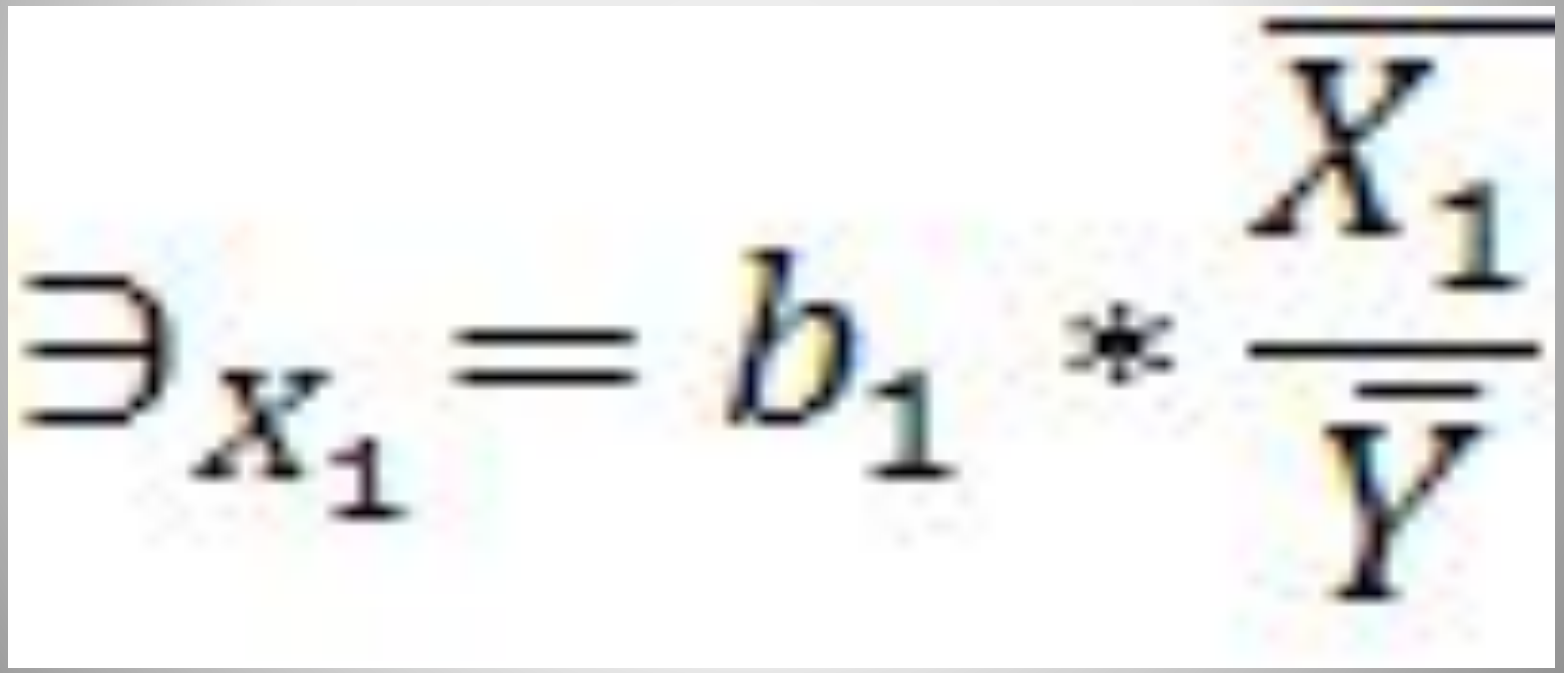
Для сопоставления факторов по силе влияния используют относительные показатели силы связи – коэффициенты эластичности.

Коэффициент эластичности

$$\exists_{x_1} = b_1 * \frac{\overline{X_1}}{\overline{Y}}$$

Коэффициенты эластичности показывают, на сколько процентов в среднем изменится результат при изменении фактора на 1% и значениях других факторов, фиксированных на средних уровнях.

Коэффициент эластичности



Стандартизированные коэффициенты регрессии

Во множественной регрессии и корреляции относительным показателем силы связи также являются стандартизированные коэффициенты регрессии.

Как и коэффициенты эластичности, они сопоставимы между собой по силе влияния факторов на результат.

Стандартизированные коэффициенты регрессии показывают, на сколько своих среднеквадратических отклонений в среднем изменится результат при изменении любого конкретного фактора на одно свое среднеквадратическое отклонение при фиксированном уровне других факторов, включенных в модель множественной регрессии.

Вопрос 4. «Оценка параметров модели множественной регрессии и корреляции»

Несмещенность, эффективность и состоятельность оценок параметров

Параметры уравнения множественной регрессии являются выборочными оценками неизвестных параметров по генеральной совокупности, поэтому следует проверить их качество.

В модели множественной регрессии принято использовать оценки параметров, которые являются несмещенными, эффективными и состоятельными.

Несмещенность, эффективность и состоятельность оценок параметров

Оценка параметра является несмещенной, если ее математическое ожидание равно оцениваемому параметру.

Несмещенность, эффективность и состоятельность оценок параметров

Оценка параметра является эффективной, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок данного параметра по выборкам одного и того же объема.

Несмещенность, эффективность и состоятельность оценок параметров

Оценка параметра является состоятельной, если с увеличением числа наблюдений оценка параметра стремится к его значению в генеральной совокупности.

Метод наименьших квадратов

Наиболее простым методом оценки параметров уравнения множественной регрессии является метод наименьших квадратов

Он применяется в случае соблюдения определенных предпосылок:

1. факторы являются неслучайными величинами, не связанными между собой;

2. результат является случайной величиной, не ограниченной сверху или снизу;

3. для каждого конкретного значения фактора (или факторов) результат рассматривается как отдельная случайная величина результативного признака;

4. различные случайные величины независимы друг от друга.

Метод максимального правдоподобия

Если значения факторов и результативного признака не удовлетворяют перечисленным предпосылкам, то для нахождения параметров модели регрессии можно использовать метод максимального правдоподобия.

Для его применения необходимо знать закон распределения результативного признака.

Благодарю за внимание