**Корреляционный анализ.**

**Уравнение парной регрессии**.

**Использование графического метода**.

Этот метод применяют для наглядного изображения формы связи между изучаемыми экономическими показателями. Для этого в прямоугольной системе координат строят график, по оси ординат откладывают индивидуальные значения результативного признака Y, а по оси абсцисс - индивидуальные значения факторного признака X.

Совокупность точек результативного и факторного признаков называется **полем корреляции**.

На основании поля корреляции можно выдвинуть гипотезу (для генеральной совокупности) о том, что связь между всеми возможными значениями X и Y носит экспоненциальный характер.

Экспоненциальное уравнение регрессии имеет вид y = a∙ebx

Оценочное уравнение регрессии (построенное по выборочным данным) будет иметь вид y = a∙ebx + ε, где ei – наблюдаемые значения (оценки) ошибок εi, a и b соответственно оценки параметров α и β регрессионной модели, которые следует найти.

Здесь ε - случайная ошибка (отклонение, возмущение).

Причины существования случайной ошибки:

1. Невключение в регрессионную модель значимых объясняющих переменных;

2. Агрегирование переменных. Например, функция суммарного потребления – это попытка общего выражения совокупности решений отдельных индивидов о расходах. Это лишь аппроксимация отдельных соотношений, которые имеют разные параметры.

3. Неправильное описание структуры модели;

4. Неправильная функциональная спецификация;

5. Ошибки измерения.

Так как отклонения εi для каждого конкретного наблюдения i – случайны и их значения в выборке неизвестны, то:

1) по наблюдениям xi и yi можно получить только оценки параметров α и β

2) Оценками параметров α и β регрессионной модели являются соответственно величины а и b, которые носят случайный характер, т.к. соответствуют случайной выборке;

После линеаризации получим: ln(y) = ln(a) + bx

Для оценки параметров α и β - используют МНК (метод наименьших квадратов).

Метод наименьших квадратов дает наилучшие (состоятельные, эффективные и несмещенные) оценки параметров уравнения регрессии. Но только в том случае, если выполняются определенные предпосылки относительно случайного члена (ε) и независимой переменной (x).

Формально критерий МНК можно записать так:

S = ∑(yi - y∙i)2 → min

Система нормальных уравнений.

a·n + b·∑x = ∑y

a·∑x + b·∑x2 = ∑y·x

Для расчета параметров регрессии построим расчетную таблицу (табл. 1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | ln(y) | x2 | ln(y)2 | x∙ln(y) |
| 109 | 5.5759 | 11881 | 31.0912 | 607.7785 |
| 188 | 5.5013 | 35344 | 30.2638 | 1034.2365 |
| 266 | 3.9512 | 70756 | 15.6123 | 1051.0308 |
| 563 | 15.0285 | 117981 | 76.9674 | 2693.0458 |

Для наших данных система уравнений имеет вид

3a + 563·b = 15.028

563·a + 117981·b = 2693.046

Домножим уравнение (1) системы на (-187.667), получим систему, которую решим методом алгебраического сложения.

-563a -105656.521 b = -2820.344

563∙a + 117981∙b = 2693.046

Получаем:

12324.479∙b = -127.298

Откуда b = -0.01033

Теперь найдем коэффициент «a» из уравнения (1):

3a + 563∙b = 15.028

3a + 563∙(-0.01033) = 15.028

3a = 20.843

a = 6.9478

Получаем эмпирические коэффициенты регрессии: b = -0.01033, a = 6.9478

Уравнение регрессии (эмпирическое уравнение регрессии):

y = e6.9477709141977e-0.01033x = 1040.82705e-0.01033x

Эмпирические коэффициенты регрессии a и b являются лишь оценками теоретических коэффициентов βi, а само уравнение отражает лишь общую тенденцию в поведении рассматриваемых переменных.

**1. Параметры уравнения регрессии**.

Выборочные средние.

Выборочные дисперсии:

Среднеквадратическое отклонение

Коэффициент корреляции b можно находить по формуле, не решая систему непосредственно:

**1.6. Индекс детерминации**.

Величину R2 (равную отношению объясненной уравнением регрессии дисперсии результата *у* к общей дисперсии *у*) для нелинейных связей называют **индексом детерминации**.

Чаще всего, давая интерпретацию индекса детерминации, его выражают в процентах.

т.е. в 46.23% случаев изменения х приводят к изменению y. Другими словами - точность подбора уравнения регрессии - средняя. Остальные 53.77% изменения Y объясняются факторами, не учтенными в модели (а также ошибками спецификации).

Для оценки качества параметров регрессии построим расчетную таблицу (табл. 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | y | y(x) |  |  |
| 109 | 264 | 337.64 | 5929 | 5422.869 |
| 188 | 245 | 149.312 | 3364 | 9156.113 |
| 266 | 52 | 66.715 | 18225 | 216.531 |
| 563 | 561 | 553.668 | 27518 | 14795.513 |

**2. Оценка параметров уравнения регрессии**.

**2.3. Анализ точности определения оценок коэффициентов регрессии**.

Несмещенной оценкой дисперсии возмущений является величина:

S2 = 14795.513 - необъясненная дисперсия или дисперсия ошибки регрессии (мера разброса зависимой переменной вокруг линии регрессии).

S = 121.64 - стандартная ошибка оценки.

Стандартная ошибка регрессии рассматривается в качестве меры разброса данных наблюдений от смоделированных значений. Чем меньше значение стандартной ошибки регрессии, тем качество модели выше.

Sa - стандартное отклонение случайной величины a.

Sb - стандартное отклонение случайной величины b.

**2.5. Проверка гипотез относительно коэффициентов линейного уравнения регрессии**.

1) t-статистика. Критерий Стьюдента.

С помощью МНК мы получили лишь оценки параметров уравнения регрессии, которые характерны для конкретного статистического наблюдения (конкретного набора значений x и y).

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитываются t-критерий Стьюдента и доверительные интервалы каждого из показателей. Выдвигается гипотеза Н0 о случайной природе показателей, т.е. о незначимом их отличии от нуля.

Чтобы проверить, значимы ли параметры, т.е. значимо ли они отличаются от нуля для генеральной совокупности используют статистические методы проверки гипотез.

В качестве основной (нулевой) гипотезы выдвигают гипотезу о незначимом отличии от нуля параметра или статистической характеристики в генеральной совокупности. Наряду с основной (проверяемой) гипотезой выдвигают альтернативную (конкурирующую) гипотезу о неравенстве нулю параметра или статистической характеристики в генеральной совокупности.

Проверим гипотезу H0 о равенстве отдельных коэффициентов регрессии нулю (при альтернативе H1 не равно) на уровне значимости α=0.05.

H0: b = 0, то есть между переменными x и y отсутствует линейная взаимосвязь в генеральной совокупности;

H1: b ≠ 0, то есть между переменными x и y есть линейная взаимосвязь в генеральной совокупности.

В случае если основная гипотеза окажется неверной, мы принимаем альтернативную. Для проверки этой гипотезы используется *t-критерий Стьюдента*.

Найденное по данным наблюдений значение t-критерия (его еще называют наблюдаемым или фактическим) сравнивается с табличным (критическим) значением, определяемым по таблицам распределения Стьюдента (которые обычно приводятся в конце учебников и практикумов по статистике или эконометрике).

Табличное значение определяется в зависимости от уровня значимости (α) и числа степеней свободы, которое в случае линейной парной регрессии равно (n-2), n-число наблюдений.

Если фактическое значение t-критерия больше табличного (по модулю), то основную гипотезу отвергают и считают, что с вероятностью (1-α) параметр или статистическая характеристика в генеральной совокупности значимо отличается от нуля.

Если фактическое значение t-критерия меньше табличного (по модулю), то нет оснований отвергать основную гипотезу, т.е. параметр или статистическая характеристика в генеральной совокупности незначимо отличается от нуля при уровне значимости α.

tкрит(n-m-1;α/2) = tкрит(1;0.025) = 25.452

Поскольку 0.00943 < 25.452, то статистическая значимость коэффициента регрессии b не подтверждается (принимаем гипотезу о равенстве нулю этого коэффициента). Это означает, что в данном случае коэффициентом *b* можно пренебречь.

Поскольку 0.032 < 25.452, то статистическая значимость коэффициента регрессии a не подтверждается (принимаем гипотезу о равенстве нулю этого коэффициента). Это означает, что в данном случае коэффициентом *a* можно пренебречь.

**Доверительный интервал для коэффициентов уравнения регрессии**.

Определим доверительные интервалы коэффициентов регрессии, которые с надежность 95% будут следующими:

(b - tкрит Sb; b + tкрит Sb)

(-0.0103 - 25.452∙1.096; -0.0103 + 25.452∙1.096)

(-27.897;27.877)

С вероятностью 95% можно утверждать, что значение данного параметра будут лежать в найденном интервале.

(a - tкрит Sa; a + tкрит Sa)

(6.948 - 25.452∙217.282; 6.948 + 25.452∙217.282)

(-5523.302;5537.198)

С вероятностью 95% можно утверждать, что значение данного параметра будут лежать в найденном интервале.

Так как точка 0 (ноль) лежит внутри доверительного интервала, то интервальная оценка коэффициента *a* статистически незначима.

2) F-статистика. Критерий Фишера.

Индекс детерминации R2 используется для проверки существенности уравнения нелинейной регрессии в целом.

Проверка значимости модели регрессии проводится с использованием F-критерия Фишера, расчетное значение которого находится как отношение дисперсии исходного ряда наблюдений изучаемого показателя и несмещенной оценки дисперсии остаточной последовательности для данной модели.

Если расчетное значение с k1=(m) и k2=(n-m-1) степенями свободы больше табличного при заданном уровне значимости, то модель считается значимой.

 = 1 - = 0.4623

где m – число факторов в модели.

Оценка статистической значимости парной линейной регрессии производится по следующему алгоритму:

1. Выдвигается нулевая гипотеза о том, что уравнение в целом статистически незначимо: H0: R2=0 на уровне значимости α.

2. Далее определяют фактическое значение F-критерия:

или по формуле:

F = = ∙ = 0.86

где

где m=1 для парной регрессии.

3. Табличное значение определяется по таблицам распределения Фишера для заданного уровня значимости, принимая во внимание, что число степеней свободы для общей суммы квадратов (большей дисперсии) равно 1 и число степеней свободы остаточной суммы квадратов (меньшей дисперсии) при линейной регрессии равно n-2.

Fтабл - это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α. Уровень значимости α - вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно α принимается равной 0,05 или 0,01.

4. Если фактическое значение F-критерия меньше табличного, то говорят, что нет основания отклонять нулевую гипотезу.

В противном случае, нулевая гипотеза отклоняется и с вероятностью (1-α) принимается альтернативная гипотеза о статистической значимости уравнения в целом.

Табличное значение критерия со степенями свободы k1=1 и k2=1, Fтабл = 161.4476

Поскольку фактическое значение F < Fтабл, то коэффициент детерминации статистически не значим (найденная оценка уравнения регрессии статистически не надежна).

Связь между F-критерием Фишера и t-статистикой Стьюдента выражается равенством:

**Дисперсионный анализ**.

При анализе качества модели регрессии используется теорема о разложении дисперсии, согласно которой общая дисперсия результативного признака может быть разложена на две составляющие – объясненную и необъясненную уравнением регрессии дисперсии.

Задача дисперсионного анализа состоит в анализе дисперсии зависимой переменной:

∑(yi - ycp)2 = ∑(y(x) - ycp)2 + ∑(y - y(x))2

где

∑(yi - ycp)2 - общая сумма квадратов отклонений;

∑(y(x) - ycp)2 - сумма квадратов отклонений, обусловленная регрессией («объясненная» или «факторная»);

∑(y - y(x))2 - остаточная сумма квадратов отклонений.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Источник вариации | Сумма квадратов | Число степеней свободы | Дисперсия на 1 степень свободы | F-критерий |
| Модель (объясненная) | 12722.487 | 1 | 12722.487 | 0.86 |
| Остаточная | 14795.51 | 1 | 14795.51 | 1 |
| Общая | 27518 | 3-1 |   |   |

**Показатели качества уравнения регрессии**.

|  |  |
| --- | --- |
| Показатель | Значение |
| Коэффициент детерминации | 0.4623 |
| Средний коэффициент эластичности | не был рассчитан |
| Средняя ошибка аппроксимации | не был рассчитан |

**Выводы**.

Изучена зависимость Y от X. На этапе спецификации была выбрана парная экспоненциальная регрессия. Оценены её параметры методом наименьших квадратов: y = 1040.827 e-0.0103 x Статистическая значимость уравнения проверена с помощью коэффициента детерминации и критерия Фишера. Установлено, что в исследуемой ситуации 46.23% общей вариабельности Y объясняется изменением X. Установлено также, что параметры модели статистически не значимы.

Решение было получено и оформлено с помощью сервиса:

[Уравнение нелинейной регрессии](https://math.semestr.ru/corel/noncorel.php)

Вместе с этой задачей решают также:

[Уравнение множественной регрессии](https://math.semestr.ru/regress/corel.php)

[Выявление тренда методом аналитического выравнивания](https://math.semestr.ru/trend/analis.php)

[Показатели вариации](https://math.semestr.ru/group/group.php)

[Показатели динамики](https://axd.semestr.ru/dinam/group.php)