4. Линейный парный регрессионный анализ (15 баллов)

Имеются данные о товарообороте и сумме издержек обращения по 10 магазинам города, представленные в таблице

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Товарооборот | 38 | 41 | 43 | 44 | 47 | 49 | 52 | 64 | 65 | 66 |
| Издержки | 3,2 | 2,5 | 3,1 | 2,8 | 2,9 | 3,2 | 3,6 | 3,4 | 3,7 | 4,8 |

Провести линейный регрессионный анализ. Проверить значимость модели с помощью критерия Фишера. Осуществить прогноз с помощью регрессионной модели для товарооборота равного 75

Линейное уравнение регрессии имеет вид y = bx + a

Для оценки параметров a и b - 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 | y | x2 | y2 | x\*y |
| 38 | 3.2 | 1444 | 10.24 | 121.6 |
| 41 | 2.5 | 1681 | 6.25 | 102.5 |
| 43 | 3.1 | 1849 | 9.61 | 133.3 |
| 44 | 2.8 | 1936 | 7.84 | 123.2 |
| 47 | 2.9 | 2209 | 8.41 | 136.3 |
| 49 | 3.2 | 2401 | 10.24 | 156.8 |
| 52 | 3.6 | 2704 | 12.96 | 187.2 |
| 64 | 3.4 | 4096 | 11.56 | 217.6 |
| 65 | 3.7 | 4225 | 13.69 | 240.5 |
| 66 | 4.8 | 4356 | 23.04 | 316.8 |
| 509 | 33.2 | 26901 | 113.84 | 1735.8 |

Для наших данных система уравнений имеет вид

10a + 509\*b = 33.2

509\*a + 26901\*b = 1735.8

Домножим уравнение (1) системы на (-50.9), получим систему, которую решим методом алгебраического сложения.

-509a -25908.1 b = -1689.88

509\*a + 26901\*b = 1735.8

Получаем:

992.9\*b = 45.92

Откуда b = 0.04625

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

10a + 509\*b = 33.2

10a + 509\*0.04625 = 33.2

10a = 9.66

a = 0.966

Получаем эмпирические коэффициенты регрессии: b = 0.04625, a = 0.966

Уравнение регрессии (эмпирическое уравнение регрессии):

y = 0.04625 x + 0.966

**Параметры уравнения регрессии**.

Выборочные средние.

Выборочные дисперсии:

=

=

Среднеквадратическое отклонение

**Коэффициент корреляции**.

Рассчитываем показатель тесноты связи. Таким показателем является выборочный линейный коэффициент корреляции, который рассчитывается по формуле:

Линейный коэффициент корреляции принимает значения от –1 до +1.

Связи между признаками могут быть слабыми и сильными (тесными). Их критерии оцениваются по шкале Чеддока:

0.1 < rxy < 0.3: слабая;

0.3 < rxy < 0.5: умеренная;

0.5 < rxy < 0.7: заметная;

0.7 < rxy < 0.9: высокая;

0.9 < rxy < 1: весьма высокая;

В нашем примере связь между признаком Y и фактором X высокая и прямая.

**Ошибка аппроксимации**.

Для оценки качества параметров регрессии построим расчетную таблицу.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| x | y | y(x) | (yi-ycp)2 | (y-y(x))2 | (xi-xcp)2 | |y - yx|:y |
| 38 | 3.2 | 2.723 | 0.0144 | 0.227 | 166.41 | 0.149 |
| 41 | 2.5 | 2.862 | 0.672 | 0.131 | 98.01 | 0.145 |
| 43 | 3.1 | 2.955 | 0.0484 | 0.0211 | 62.41 | 0.0469 |
| 44 | 2.8 | 3.001 | 0.27 | 0.0404 | 47.61 | 0.0717 |
| 47 | 2.9 | 3.14 | 0.176 | 0.0574 | 15.21 | 0.0826 |
| 49 | 3.2 | 3.232 | 0.0144 | 0.00103 | 3.61 | 0.01 |
| 52 | 3.6 | 3.371 | 0.0784 | 0.0525 | 1.21 | 0.0636 |
| 64 | 3.4 | 3.926 | 0.0064 | 0.277 | 171.61 | 0.155 |
| 65 | 3.7 | 3.972 | 0.144 | 0.074 | 198.81 | 0.0735 |
| 66 | 4.8 | 4.018 | 2.19 | 0.611 | 228.01 | 0.163 |
| 509 | 33.2 | 33.2 | 3.616 | 1.492 | 992.9 | 0.96 |

Оценим качество уравнения регрессии с помощью ошибки абсолютной аппроксимации. Средняя ошибка аппроксимации - среднее отклонение расчетных значений от фактических:

Ошибка аппроксимации в пределах 5%-7% свидетельствует о хорошем подборе уравнения регрессии к исходным данным.

В среднем, расчетные значения отклоняются от фактических на 9.6%. Поскольку ошибка больше 7%, то данное уравнение не желательно использовать в качестве регрессии.

**Коэффициент детерминации**.

Квадрат (множественного) коэффициента корреляции называется коэффициентом детерминации, который показывает долю вариации результативного признака, объясненную вариацией факторного признака.

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

R2= 0.7662 = 0.5873

т.е. в 58.73% случаев изменения х приводят к изменению y. Другими словами - точность подбора уравнения регрессии - средняя. Остальные 41.27% изменения Y объясняются факторами, не учтенными в модели (а также ошибками спецификации).

**Оценка параметров уравнения регрессии**.

Несмещенной оценкой дисперсии возмущений является величина:

S2 = 0.187 - необъясненная дисперсия или дисперсия ошибки регрессии (мера разброса зависимой переменной вокруг линии регрессии).

S = 0.43 - 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.

**Доверительные интервалы для зависимой переменной**.

Экономическое прогнозирование на основе построенной модели предполагает, что сохраняются ранее существовавшие взаимосвязи переменных и на период упреждения. Для прогнозирования зависимой переменной результативного признака необходимо знать прогнозные значения всех входящих в модель факторов.

Прогнозные значения факторов подставляют в модель и получают точечные прогнозные оценки изучаемого показателя.

(a + bxp ± ε)

Рассчитаем границы интервала, в котором будет сосредоточено 95% возможных значений Y при неограниченно большом числе наблюдений и Xp = 75

tкрит (n-m-1;α/2) = (8;0.025) = 2.306

y(75) = 0.0462\*75 + 0.966 = 4.435

Вычислим ошибку прогноза для уравнения y = bx + a

=

4.435 ± 0.824

(3.61;5.26)

С вероятностью 95% можно гарантировать, что значения Y при неограниченно большом числе наблюдений не выйдет за пределы найденных интервалов.

**F-статистика. Критерий Фишера.**

Коэффициент детерминации R2 используется для проверки существенности уравнения линейной 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.

=

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:

=

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.

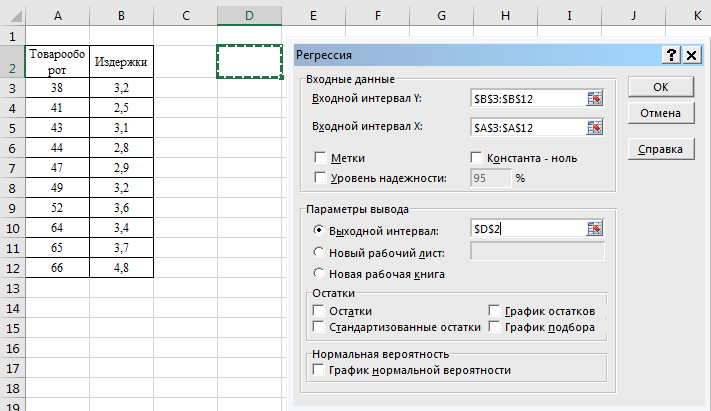
В противном случае, нулевая гипотеза отклоняется и с вероятностью (1-α) принимается альтернативная гипотеза о статистической значимости уравнения в целом.

Табличное значение критерия со степенями свободы k1=1 и k2=8, Fтабл = 5.32

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

Такое же решение можно быстро получить в Excel, воспользовавшись инструментом Анализ данных – Регрессия

Заполним окно инструмента



Решение

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ВЫВОД ИТОГОВ |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Регрессионная статистика* |  |  |  |  |  |  |
| Множественный R | 0,766 |  |  |  |  |  |
| R-квадрат | 0,587 |  |  |  |  |  |
| Нормированный R-квадрат | 0,536 |  |  |  |  |  |
| Стандартная ошибка | 0,432 |  |  |  |  |  |
| Наблюдения | 10 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Дисперсионный анализ |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Значимость F* |  |
| Регрессия | 1 | 2,124 | 2,124 | 11,385 | 0,010 |  |
| Остаток | 8 | 1,492 | 0,187 |  |  |  |
| Итого | 9 | 3,616 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Коэффициенты* | *Стандартная ошибка* | *t-статистика* | *P-Значение* | *Нижние 95%* | *Верхние 95%* |
| Y-пересечение | 0,966 | 0,711 | 1,359 | 0,211 | -0,673 | 2,605 |
| Переменная X 1 | 0,046 | 0,014 | 3,374 | 0,010 | 0,015 | 0,078 |

Коэффициенты в нижней таблице дают коэффициенты уравнения

y = 0,966+0,046x

R-квадрат в верхней таблице есть коэффициент детерминации = 0,587

Значимость уравнения можно сразу определить исходя из величины Значимость F = 0,01 второй таблицы и поскольку она меньше принятого уровня значимости = 0,05, то уравнение в целом значимо

Значимость коэффициентов можно определить по величине P-значение, если оно для данного коэффициент меньше 0,05, то коэффициент значим. В нашем случае свободный член незначим, коэффициент при х – значим.

**Выводы**.

Изучена зависимость Y (издержки) от X (товарооборот). На этапе спецификации была выбрана парная линейная регрессия. Оценены её параметры методом наименьших квадратов. Статистическая значимость уравнения проверена с помощью коэффициента детерминации и критерия Фишера. Установлено, что в исследуемой ситуации 58.73% общей вариабельности Y объясняется изменением X. Возможна экономическая интерпретация параметров модели - увеличение X (товарооборота) на 1 ед. изм. приводит к увеличению Y (издержек) в среднем на 0.0462 ед.изм. Полученные оценки уравнения регрессии позволяют использовать его для прогноза. При x=75, Y (издержки) будет находиться в пределах от 3.14 до 5.73 ед.изм. и с вероятностью 95% не выйдет за эти пределы.