5. Линейный множественный регрессионный анализ

(15 баллов)

Имеются следующие данные о курсе доллара , фондовом индексе  и котировке акций за 10 дней.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 40,7 | 38,7 | 39,5 | 41,9 | 42,8 | 39,3 | 37,9 | 37,1 | 39,0 | 39,9 |
|  | 4,3 | 4,5 | 4,2 | 5,2 | 4,8 | 4,2 | 4,1 | 4,1 | 4,6 | 4,24 |
|  | 111 | 115 | 109 | 106 | 109 | 106 | 115 | 103 | 109 | 106 |

Провести линейный множественный регрессионный анализ. Проверить значимость модели. Проверить модель на мультиколлинеарность. Спрогнозируйте котировку акций, если курс доллара составит 33,5 руб., а значение фондового индекса равно 3.

Эмпирическое уравнение множественной регрессии представим в виде:

Y = b0 + b1X1 + b1X1 + ... + bmXm + e

Здесь b0, b1, ..., bm - оценки теоретических значений β0, β1, β2, ..., βm коэффициентов регрессии (эмпирические коэффициенты регрессии); e - оценка отклонения ε.

При выполнении предпосылок МНК относительно ошибок εi, оценки b0, b1, ..., bm параметров β0, β1, β2, ..., βm множественной линейной регрессии по МНК являются несмещенными, эффективными и состоятельными (т.е. BLUE-оценками).

Для оценки параметров уравнения множественной регрессии применяют МНК.

**1. Оценка уравнения регрессии**.

Определим вектор оценок коэффициентов регрессии. Согласно методу наименьших квадратов, вектор *s* получается из выражения: s = (XTX)-1XTY

К матрице с переменными Xj добавляем единичный столбец:

|  |  |  |
| --- | --- | --- |
| 1 | 40.7 | 4.3 |
| 1 | 38.7 | 4.5 |
| 1 | 39.5 | 4.2 |
| 1 | 41.9 | 5.2 |
| 1 | 42.8 | 4.8 |
| 1 | 39.3 | 4.2 |
| 1 | 37.9 | 4.1 |
| 1 | 37.1 | 4.1 |
| 1 | 39 | 4.6 |
| 1 | 39.9 | 4.24 |

Матрица Y

|  |
| --- |
| 111 |
| 115 |
| 109 |
| 106 |
| 109 |
| 106 |
| 115 |
| 103 |
| 109 |
| 106 |

Матрица XT

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 40.7 | 38.7 | 39.5 | 41.9 | 42.8 | 39.3 | 37.9 | 37.1 | 39 | 39.9 |
| 4.3 | 4.5 | 4.2 | 5.2 | 4.8 | 4.2 | 4.1 | 4.1 | 4.6 | 4.24 |

Умножаем матрицы, (XTX)

В матрице, (XTX) число 10, лежащее на пересечении 1-й строки и 1-го столбца, получено как сумма произведений элементов 1-й строки матрицы XT и 1-го столбца матрицы X

Умножаем матрицы, (XTY)

Находим обратную матрицу (XTX)-1

=

Вектор оценок коэффициентов регрессии равен

= =

Уравнение регрессии (оценка уравнения регрессии)

Y = 116.1948-0.1601X1-0.2144X2

**Матрица парных коэффициентов корреляции R**.

Число наблюдений n = 10. Число независимых переменных в модели равно 2, а число регрессоров с учетом единичного вектора равно числу неизвестных коэффициентов. С учетом признака Y, размерность матрицы становится равным 4. Матрица, независимых переменных Х имеет размерность (10 х 4).

Матрица A, составленная из Y и X

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 111 | 40.7 | 4.3 |
| 1 | 115 | 38.7 | 4.5 |
| 1 | 109 | 39.5 | 4.2 |
| 1 | 106 | 41.9 | 5.2 |
| 1 | 109 | 42.8 | 4.8 |
| 1 | 106 | 39.3 | 4.2 |
| 1 | 115 | 37.9 | 4.1 |
| 1 | 103 | 37.1 | 4.1 |
| 1 | 109 | 39 | 4.6 |
| 1 | 106 | 39.9 | 4.24 |

Транспонированная матрица.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 111 | 115 | 109 | 106 | 109 | 106 | 115 | 103 | 109 | 106 |
| 40.7 | 38.7 | 39.5 | 41.9 | 42.8 | 39.3 | 37.9 | 37.1 | 39 | 39.9 |
| 4.3 | 4.5 | 4.2 | 5.2 | 4.8 | 4.2 | 4.1 | 4.1 | 4.6 | 4.24 |

Матрица XTX.

|  |  |  |  |
| --- | --- | --- | --- |
| 10 | 1089 | 396.8 | 44.24 |
| 1089 | 118731 | 43206.3 | 4816.84 |
| 396.8 | 43206.3 | 15772.2 | 1759.516 |
| 44.24 | 4816.84 | 1759.516 | 196.858 |

Полученная матрица имеет следующее соответствие:

|  |  |  |  |
| --- | --- | --- | --- |
| ∑n | ∑y | ∑x1 | ∑x2 |
| ∑y | ∑y2 | ∑x1 y | ∑x2 y |
| ∑x1 | ∑yx1 | ∑x1 2 | ∑x2 x1 |
| ∑x2 | ∑yx2 | ∑x1 x2 | ∑x2 2 |

Найдем парные коэффициенты корреляции.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Признаки x и y | ∑xi |  | ∑yi |  | ∑xiyi |  |
| Для y и x1 | 396.8 | 39.68 | 1089 | 108.9 | 43206.3 | 4320.63 |
| Для y и x2 | 44.24 | 4.424 | 1089 | 108.9 | 4816.84 | 481.684 |
| Для x1 и x2 | 44.24 | 4.424 | 396.8 | 39.68 | 1759.516 | 175.952 |

Дисперсии и среднеквадратические отклонения.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Признаки x и y |  |  |  |  |
| Для y и x1 | 2.718 | 13.89 | 1.649 | 3.727 |
| Для y и x2 | 0.114 | 13.89 | 0.338 | 3.727 |
| Для x1 и x2 | 0.114 | 2.718 | 0.338 | 1.649 |

Матрица парных коэффициентов корреляции R:

|  |  |  |  |
| --- | --- | --- | --- |
| - | y | x1 | x2 |
| y | 1 | -0.08496 | -0.07121 |
| x1 | -0.08496 | 1 | 0.7318 |
| x2 | -0.07121 | 0.7318 | 1 |

**Анализ мультиколлинеарности**.

Если факторные переменные связаны строгой функциональной зависимостью, то говорят о полной мультиколлинеарности. В этом случае среди столбцов матрицы факторных переменных Х имеются линейно зависимые столбцы, и, по свойству определителей матрицы, det(XTX = 0).

Вид мультиколлинеарности, при котором факторные переменные связаны некоторой стохастической зависимостью, называется частичной. Если между факторными переменными имеется высокая степень корреляции, то матрица (XTX) близка к вырожденной, т. е. det(XTX ≧ 0) (чем ближе к 0 определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии).

1. Анализ мультиколлинеарности на основе матрицы коэффициентов корреляции.

Если в матрице есть межфакторный коэффициент корреляции rxjxi > 0.7, то в данной модели множественной регрессии существует мультиколлинеарность.

**В нашем случае rx1x2 имеют |r|>0.7, что говорит о мультиколлинеарности факторов и о необходимости исключения одного из них из дальнейшего анализа.**

**Анализ параметров уравнения регрессии**.

Перейдем к статистическому анализу полученного уравнения регрессии: проверке значимости уравнения и его коэффициентов, исследованию абсолютных и относительных ошибок аппроксимации

Для несмещенной оценки дисперсии проделаем следующие вычисления:

Несмещенная ошибка ε = Y - Y(x) = Y - X\*s (абсолютная ошибка аппроксимации)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Y | Y(x) | ε = Y - Y(x) | ε2 | (Y-Yср)2 | |ε : Y| |
| 111 | 108.758 | 2.242 | 5.027 | 4.41 | 0.0202 |
| 115 | 109.035 | 5.965 | 35.578 | 37.21 | 0.0519 |
| 109 | 108.972 | 0.0285 | 0.000812 | 0.01 | 0.000261 |
| 106 | 108.373 | -2.373 | 5.631 | 8.41 | 0.0224 |
| 109 | 108.315 | 0.685 | 0.47 | 0.01 | 0.00629 |
| 106 | 109.004 | -3.004 | 9.021 | 8.41 | 0.0283 |
| 115 | 109.249 | 5.751 | 33.073 | 37.21 | 0.05 |
| 103 | 109.377 | -6.377 | 40.668 | 34.81 | 0.0619 |
| 109 | 108.966 | 0.0342 | 0.00117 | 0.01 | 0.000314 |
| 106 | 108.899 | -2.899 | 8.404 | 8.41 | 0.0273 |
|  |  |  | 137.873 | 138.9 | 0.269 |

Средняя ошибка аппроксимации

Оценка дисперсии равна:

se2=(Y-Y(X))T(Y-Y(X))=137.873

Несмещенная оценка дисперсии равна:

Оценка среднеквадратичного отклонения (стандартная ошибка для оценки Y):

Найдем оценку ковариационной матрицы вектора k = S2 • (XTX)-1

=

Дисперсии параметров модели определяются соотношением S2i = Kii, т.е. это элементы, лежащие на главной диагонали

**Множественный коэффициент корреляции (Индекс множественной корреляции).**

Тесноту совместного влияния факторов на результат оценивает индекс множественной корреляции.

В отличии от парного коэффициента корреляции, который может принимать отрицательные значения, он принимает значения от 0 до 1.

Поэтому R не может быть использован для интерпретации направления связи. Чем плотнее фактические значения yi располагаются относительно линии регрессии, тем меньше остаточная дисперсия и, следовательно, больше величина Ry(x1,...,xm).

Таким образом, при значении R близком к 1, уравнение регрессии лучше описывает фактические данные и факторы сильнее влияют на результат. При значении R близком к 0 уравнение регрессии плохо описывает фактические данные и факторы оказывают слабое воздействие на результат.

=

Коэффициент множественной корреляции можно определить через матрицу парных коэффициентов корреляции:

где Δr - определитель матрицы парных коэффициентов корреляции; Δr11 - определитель матрицы межфакторной корреляции.

Коэффициент множественной корреляции

Аналогичный результат получим при использовании других формул:

Связь между признаком Y и факторами Xi низкая.

**Коэффициент детерминации**.

R2= 0.085982 = 0.00739

Более объективной оценкой является скорректированный коэффициент детерминации:

Чем ближе этот коэффициент к единице, тем больше уравнение регрессии объясняет поведение Y.

Добавление в модель новых объясняющих переменных осуществляется до тех пор, пока растет скорректированный коэффициент детерминации.

**Оценка значения результативного признака при заданных значениях факторов**.

Y(33.5,3) = 116.19-0.16\*33.5-0.214\*3 = 110.189

**Проверка общего качества уравнения множественной регрессии**.

Оценка значимости уравнения множественной регрессии осуществляется путем проверки гипотезы о равенстве нулю коэффициент детерминации рассчитанного по данным генеральной совокупности: R2 или b1 = b2 =... = bm = 0 (гипотеза о незначимости уравнения регрессии, рассчитанного по данным генеральной совокупности).

Для ее проверки используют F-критерий Фишера.

При этом вычисляют фактическое (наблюдаемое) значение F-критерия, через коэффициент детерминации R2, рассчитанный по данным конкретного наблюдения.

По таблицам распределения Фишера-Снедоккора находят критическое значение F-критерия (Fкр). Для этого задаются уровнем значимости α (обычно его берут равным 0,05) и двумя числами степеней свободы k1=m и k2=n-m-1.

F-статистика. Критерий Фишера.

=

Проверим гипотезу об общей значимости - гипотезу об одновременном равенстве нулю всех коэффициентов регрессии при объясняющих переменных:

H0: R2 = 0; β1 = β2 = ... = βm = 0.

H1: R2 ≠ 0.

Проверка этой гипотезы осуществляется с помощью F-статистики распределения Фишера (правосторонняя проверка).

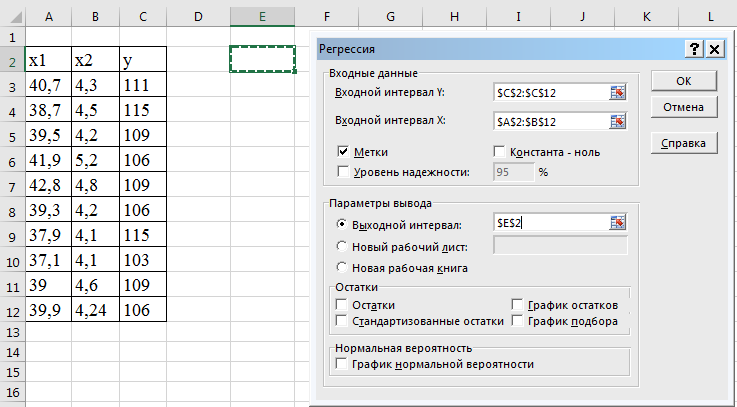
Если F < Fkp = Fα ; n-m-1, то нет оснований для отклонения гипотезы H0.

Табличное значение при степенях свободы k1 = 2 и k2 = n-m-1 = 10 - 2 - 1 = 7, Fkp(2;7) = 4.74

Поскольку фактическое значение F < Fkp, то коэффициент детерминации статистически не значим и уравнение регрессии статистически ненадежно (совместная незначимость коэффициентов при факторах xi подтверждается).

Решим задачу в Excel

Заполним окно инструмента Регрессия



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ВЫВОД ИТОГОВ |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Регрессионная статистика* | |  |  |  |  |  |
| Множественный R | 0,0860 |  |  |  |  |  |
| R-квадрат | 0,0074 |  |  |  |  |  |
| Нормированный R-квадрат | -0,2762 |  |  |  |  |  |
| Стандартная ошибка | 4,4380 |  |  |  |  |  |
| Наблюдения | 10 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Дисперсионный анализ |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Значимость F* |  |
| Регрессия | 2 | 1,0271 | 0,5135 | 0,0261 | 0,9744 |  |
| Остаток | 7 | 137,8729 | 19,6961 |  |  |  |
| Итого | 9 | 138,9000 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Коэффициенты* | *Стандартная ошибка* | *t-статистика* | *P-Значение* | *Нижние 95%* | *Верхние 95%* |
| Y-пересечение | 116,1948 | 35,0623 | 3,3139 | 0,0129 | 33,2855 | 199,10 |
| x1 | -0,1599 | 1,2491 | -0,1280 | 0,9017 | -3,1136 | 2,799 |
| x2 | -0,2147 | 6,0992 | -0,0352 | 0,9729 | -14,6371 | 14,207 |

Коэффициенты в нижней таблице дают коэффициенты уравнения

Y = 116.1948-0.1599X1-0.2147X2

R-квадрат в верхней таблице есть коэффициент детерминации = 0,0074

Значимость уравнения можно сразу определить исходя из величины Значимость F = 0,9744 второй таблицы и поскольку она больше принятого уровня значимости = 0,05, то уравнение в целом незначимо.

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

**Выводы**.

В результате расчетов было получено уравнение множественной регрессии:

Y = 116.1948-0.1601X1-0.2144X2.

Возможна экономическая интерпретация параметров модели: увеличение X1 на 1 ед.изм. приводит к уменьшению Y в среднем на 0.16 ед.изм.; увеличение X2 на 1 ед.изм. приводит к уменьшению Y в среднем на 0.214 ед.изм. Статистическая значимость уравнения проверена с помощью коэффициента детерминации и критерия Фишера. Установлено, что в исследуемой ситуации 0.74% общей вариабельности Y объясняется изменением факторов Xj.

Полученное уравнение в целом незначимо.