МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ  
«Харківський Політехнічний Інститут»  
  
Кафедра Стратегічного Управління

ЗВІТ

з лабораторної роботи № 11

з дисципліни Математична статистика

на тему

«Регресія»

Перевірила: старший викладач  
Мошко Є. О.  
Виконав: ст. гр. КН-27

Харків, 2019

Цель: Провести двухфакторный дисперсионный анализ без повторений.

Задание: Можно ли утверждать, что фильмы снятые одним режиссером более успешны, чем фильмы других режиссеров. Существует ли разница между успехом фильмов, снятых в разных жанрах разными режиссерами? Определить какой режиссер и какой жанр в комбинации дают самые успешные фильмы.

Ход работы.

Теоретическая часть.

Линейная регрессия

Представим зависимость y от x в виде линейной модели первого порядка:

Будем считать, что значения x определяются без ошибки, β0 и β1 – параметры модели, а ε – ошибка, распределение которой подчиняется нормальному закону с нулевым средним значением и постоянным отклонением σ2. Значения параметров β заранее не известны и их нужно определить из набора экспериментальных значений (xi, yi), i = 1,n. Таким образом мы можем записать:

где означает предсказанное моделью значение y при данном x, b0 и b1 – выборочные оценки параметров модели. Определим также – значение ошибки аппроксимации для i-го наблюдения.

Метод наименьших квадратов даёт следующие формулы для вычисления параметров данной модели и их отклонений:

Обычно истинные величины коэффициентов регрессии β0 и β1 не известны. Известны только их оценки b0 и b1. Иначе говоря, истинная прямая регрессии может пройти иначе, чем построенная по выборочным данным. Можно вычислить доверительную область для линии регрессии. При любом значении соответствующие значения распределены нормально. Средним является значение уравнения регрессии . Неопределённость его оценки характеризуется стандартной ошибкой регрессии:

Теперь можно вычислить -процентный доверительный интервал для значения уравнения регрессии в точке x:

где t(1−α/2, n−2) – t-значение распределения Стьюдента.

Практическая часть

Имеется фактор Х – года и зависимая переменная У – уровень продаж предприятия (млн. руб) – рисунок 1. Предполагается что У зависит от Х линейно. Необходимо провести регрессионный анализ, вычислить параметры наклона и смещения и спрогнозировать доход для последующих лет.

Воспользуемся пакетом анализа для проведения регрессионного анализа – рисунок 2.



Рисунок 1 – Исходные данные



Рисунок 2 – Регрессионная статистика

Здесь:

1. Множественный R - множественный коэффициент детерминации, характеризует, на сколько процентов построенная модель регрессии объясняет вариацию значений результативной переменной относительно своего среднего уровня, т. е. показывает долю общей дисперсии результативной переменной, объяснённой вариацией факторных переменных, включённых в модель регрессии;
2. R-квадрат - коэффициент детерминации R2, показывает, насколько полезна построенная нами линейная регрессионная модель;
3. Нормированный R-квадрат - чтобы была возможность сравнивать модели с разным числом факторов так, чтобы число регрессоров (факторов) не влияло на статистику R2 обычно используется скорректированный коэффициент детерминации, в котором используются несмещённые оценки дисперсий, который даёт штраф за дополнительно включённые факторы;
4. Стандартная ошибка - показывает, насколько велика ошибка предсказания значений переменной Y на основании значений Х;
5. Наблюдения - объемы выборок.

Кроме этого, для расчета регрессии автоматически проводится дисперсионный анализ:



Рисунок 3 – Дисперсионный анализ

Здесь:

1. df - число степеней свободы, использующееся для расчета квантилей;
2. SS - сумма квадратов;
3. MS - средний квадрат;
4. F – F-статистика;
5. Значимость F - верхний -квантиль;
6. Коэффициенты - коэффициентов регрессии β0 и β1;
7. Стандартная ошибка - показывает, насколько велика ошибка предсказания значений переменной Y на основании значений Х;
8. t-статистка – статистика Стьюдента;
9. Р-значение - вероятность, что F-статистика примет вычисленное значение;
10. Нижние 95% и Верхние 95% - границы критических значений, за которыми у нас будет обоснование отвергнуть нулевую теорию.

Используя исходные данные можно построить линию тренда. Используя полученные данные (стандартную ошибку) можно построить линию регрессии и рассчитать верхнюю и нижнюю границы области рассеивания эмпирических данных. Эта границы определяются как +- 2\*SEy. 95% точек располагается в пределах этой области.

Рисунок 4 – График для набора данных

Вывод

Рассчитаем двусторонний квантиль для случайной величины a – коэффициент наклона. Его можно найти по формуле СТЬЮДРАСПОБР(0.05,49)= 2.001. Так как t-статистика для переменной X1 (рисунок 3), что представляет собой уровень наклона, больше чем рассчитанный квантиль, то делаем вывод о том, что переменная Y зависит от X. При малых значения начального наклона часто бывает, что t-статистика будет меньше двустороннего квантиля и в этом случае нулевая гипотеза об отсутствии зависимости между Х и У не отклоняется.