РЕФЕРАТ

по теории экспериментов

Регрессионный анализ.

Оценка параметров линейных регрессионных моделей.

Студент Истомин Н.С.

Группа ЗМ-ЭО-16-1

Руководитель

к. т. н., доцент Шишлин Д. И.

Уравнение регрессии

Уравнение регрессии имеет вид:

$$Y = \varphi(X) + \varepsilon, \text{ где}$$

Y - результирующий признак (отклик, случайная зависимая переменная);

X – фактор (неслучайная независимая переменная);

ε – случайная переменная, характеризующая отклонение фактора X от линии регрессии

Уравнение регрессии записывается в виде:

$$y_x = \varphi(x, b_0, b_1, \ldots, b_p), \text{ где}$$

x – значения величины X;

y_x – значения величины Y;

$b_0, b_1, \ldots, b_p$ – параметры функции регрессии φ.

Задача регрессионного анализа состоит в определении функции и ее параметров и последующего статистического исследования уравнения.

В зависимости от типа выбранной функции

Линейная

Нелинейная
- степенная;
- экспоненциальная;
- логарифмическая;
- и др.

В зависимости от числа взаимосвязанных признаков

Парная (y, x)

Многофакторная (y, x_1, $x_{2,\ldots,}$ x_n)

Для оценки неизвестных параметров b_0, b_1, …, b_p используется метод наименьших квадратов (МНК). Согласно методу неизвестные параметры функции выбираются таким образом, чтобы сумма квадратов отклонений экспериментальных (эмпирических) значений y_i от их расчетных (теоретических) значений была минимальной, т.е.

$$S = \sum_{i=1}^{n} \left(y_{i\,эксп} - y_i^p\right)^2 = \sum_{i=1}^{n} \left(y_{i\,эксп} - \varphi(x_i, b_0, b_1, \ldots b_p)\right)^2 \to \min$$

$y_i - y_i^p = \varepsilon$ – отклонение (ошибка, остаток);

Парная линейная регрессионная модель

$$y_x = ax + b$$

a – коэффициент регрессии (показатель наклона линии линейной регрессии)

Формулы для расчета параметров линейной регрессии

Свободный член b	Коэффициент регрессии a	Коэффициент детерминации
$b = \frac{\bar{y} \cdot \overline{x^2} - \bar{x} \cdot \overline{xy}}{\overline{x^2} - (\bar{x})^2}$	$a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$	$R^2 = \frac{\sum(y_i^p - \bar{y})^2}{\sum(y_i - \bar{y})^2}$
Проверка гипотезы о значимости уравнения регрессии		
$H_0: R^2 = 0$	$H_1: R^2 > 0$	$F_{набл} = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p}$
$F_{кр}(\alpha; k_1; k_2)$, $k_1 = p$, $k_2 = n - p - 1$, (для линейной регрессии $p = 1$)		

Пример: Провести регрессионный анализ данных о поставках

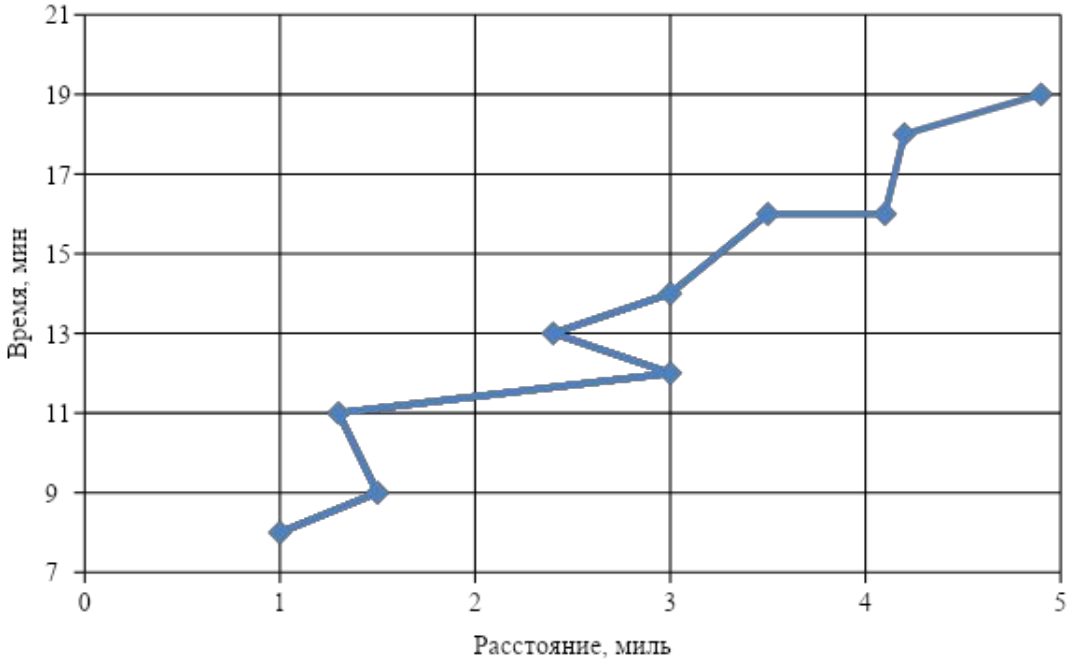
Установить зависимость времени доставки от расстояния

Расстояние, миль	3,5	2,4	4,9	4,2	3,0	1,3	1,0	3,0	1,5	4,1
Время, мин	16	13	19	18	12	11	8	14	9	16

Для проведения регрессионного анализа:

1. Построить график исходных данных, приближенно определить характер зависимости;
2. Выбрать вид функции регрессии и определить численные коэффициенты модели методом наименьших квадратов и направление связи;
3. Оценить силу регрессионной зависимости с помощью коэффициента детерминации;
4. Оценить значимость уравнения регрессии;

1. Построим график исходных данных



Построенные точки не находятся точно на линии: помимо расстояния на время поставки влияют пробки на дорогах, время суток, дорожные работы, погода, квалификация водителя, вид транспорта. Но эти точки собраны вдоль прямой линии, поэтому можно предположить линейную положительную связь между параметрами.

2. Вычислим коэффициенты модели с помощью МНК

№	x_i	y_i	x_i^2	$x_i y_i$	y_i^p	$(y_i^p - \bar{y})^2$	$(y_i - \bar{y})^2$
1	3,5	16	12,25	56,00	15,22	2,63	5,76
2	2,4	13	5,76	31,20	12,30	1,70	0,36
3	4,9	19	24,01	93,10	18,95	28,59	29,16
4	4,2	18	17,64	75,60	17,09	12,15	19,36
5	3,0	12	9,00	36,00	13,89	0,08	2,56
6	1,3	11	1,69	14,30	9,37	17,88	6,76
7	1,0	8	1,00	8,00	8,57	25,27	31,36
8	3,0	14	9,00	42,00	13,89	0,09	0,16
9	1,5	9	2,25	13,50	9,90	13,67	21,16
10	4,1	16	16,81	65,60	16,82	10,36	5,76
Σ	28,9	136	99,41	435,30	–	112,42	122,40

$$\bar{x} = \frac{\sum x_i}{n} = 2{,}89; \qquad \bar{y} = \frac{\sum y_i}{n} = 13{,}6;$$

$$b = \frac{13{,}6 \cdot 9{,}941 - 2{,}89 \cdot 43{,}53}{9{,}941 - 2{,}89^2} = 5{,}91; \qquad a = \frac{43{,}53 - 2{,}89 \cdot 13{,}6}{9{,}941 - 2{,}89^2} = 2{,}66.$$

Искомая регрессионная зависимость имеет вид: $y^p = 2{,}66x + 5{,}91$

3. Оценим силу регрессионной зависимости

$$R^2 = \frac{112,42}{122,40} = 0,92$$

Таким образом, линейная модель объясняет 92% вариации времени поставки, что означает правильность выбора фактора (расстояния). Не объясняется 8% вариации времени, которые обусловлены остальными факторами, влияющими на время поставки, но не включенными в линейную модель регрессии.

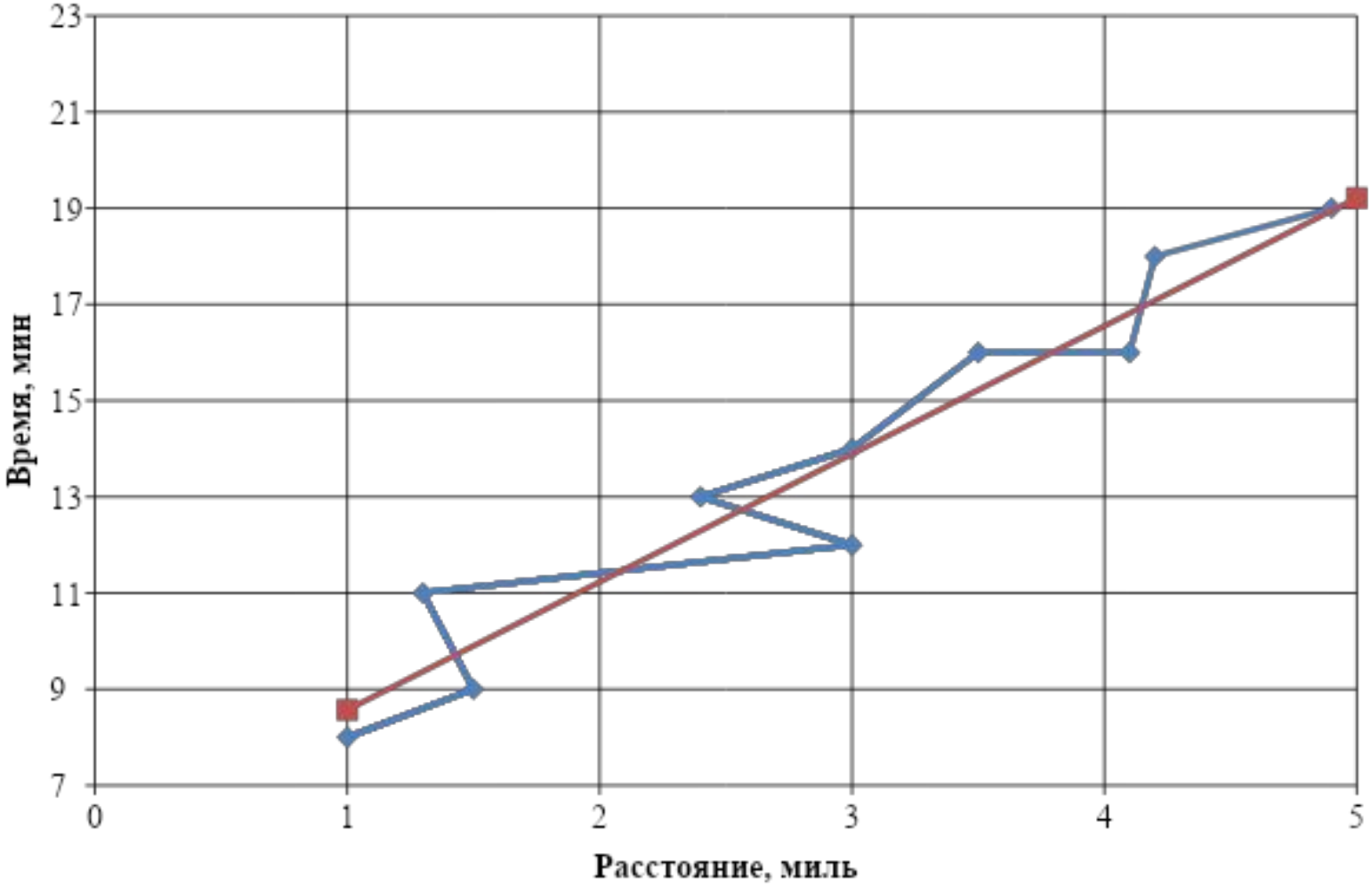
4. Оценим значимость уравнения регрессии

$$F_{набл} = \frac{0,92^2}{1 - 0,92^2} \cdot \frac{10 - 1 - 1}{1} = 44,1$$

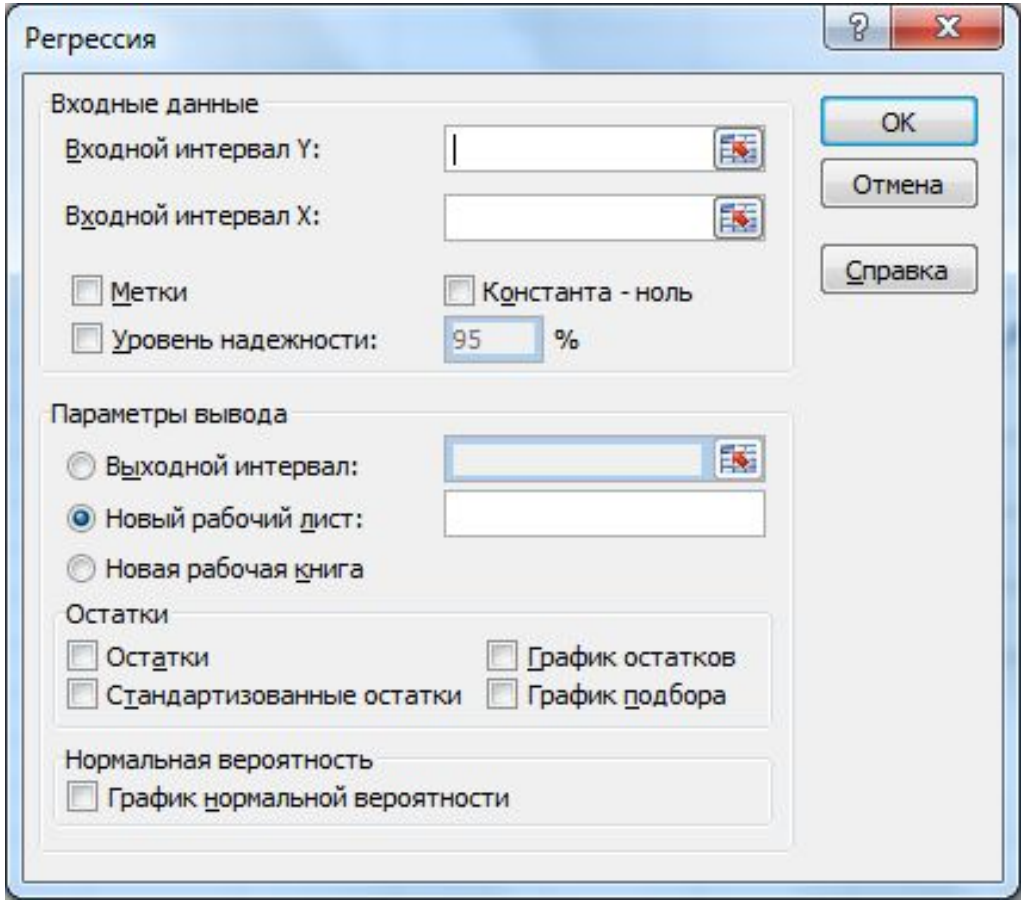
$$F_{набл} = 44,1 > F_{кр}(0,05; 1; 10 - 1 - 1) = 5,32$$

– уравнение регрессии (линейной модели) статистически значимо.

График полученного уравнения



Регрессионный анализ с использованием возможностей MS Office Excel



Окно надстройки анализа данных «Регрессия»

Результаты регрессионного анализа в MS Office Excel

ВЫВОД ИТОГОВ				
Регрессионная статистика				
Множественный R	0,958275757			
R-квадрат	0,918292427			
Нормированный R-квадрат	0,90807898			
Стандартная ошибка	1,11809028			
Наблюдения	10			
	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересечение	5,913462144	0,884389599	6,6864899	0,000155
Переменная X 1	2,65970168	0,280497238	9,4820958	1,26E-05

R-квадрат – соответствует коэффициенту детерминации R^2

Множественный R – корень из Коэффициента детерминации

Y-пересечение – коэффициент b

Переменная X1 – коэффициент a

Сравнивая попарно значения столбцов Коэффициенты и Стандартная ошибка в таблице, видим, что абсолютные значения коэффициентов больше, чем их стандартные ошибки. К тому же эти коэффициенты являются значимыми, о чем можно судить по значениям показателя P-значение, которые меньше заданного уровня значимости $\alpha=0,05$.

Вывод остатков

Наблюдение	Предсказанное Y	Остатки	Стандартные остатки
1	15,22241803	0,777581975	0,737641894
2	12,29674618	0,703253823	0,667131568
3	18,94600038	0,053999622	0,051225961
4	17,0842092	0,915790799	0,868751695
5	13,89256718	-1,892567185	-1,795356486
6	9,371074328	1,628925672	1,545256778
7	8,573163824	-0,573163824	-0,543723571
8	13,89256718	0,107432815	0,101914586
9	9,903014664	-0,903014664	-0,8566318
10	16,81823903	-0,818239033	-0,776210624

При помощи этой части отчета мы можем видеть отклонения каждой точки от построенной линии регрессии.

Для лучшей интерпретации этих данных строят график исходных данных и построенной линией регрессии.

СПАСИБО ЗА ВНИМАНИЕ!