**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ**

федеральное государственное автономное образовательное учреждение высшего образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ**

**ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

ОТЧЕТ

по лабораторной работе

***Регрессионный анализ парных наблюдений***

по курсу

«Математическая статистика»

Выполнил(а):

Саматов Д. С.

Проверил:

Доцент каф. ВМиМФ: Шинкеев М.Л.

Томск 2022

***Цель работы:***

Подбор и оценка параметров уравнения среднеквадратичной регрессии.

***Задание.***

Дана выборка значений совместно наблюдаемых величин  и  (таблица 1 Приложение 2).

**Требуется:**

1. Отобразить корреляционное поле наблюдаемых значений величин  и  (построить диаграмму рассеяния).
2. Методом наименьших квадратов найти оценки параметров уравнения линейной , квадратичной , логарифмической  и показательной  регрессии  на . Изобразить полученные зависимости на фоне корреляционного поля.
3. Получить оценки остаточной дисперсии и коэффициента детерминации, а также проверить значимость коэффициентов уравнения регрессии и значимость модели в целом (используя достигнутый уровень значимости) для каждой модели, в предположении, что остатки независимые нормальные случайные величины с одинаковой дисперсией.
4. Сравнить модели, используя значения оценок остаточной дисперсии и коэффициента детерминации, а также полученные (достигнутые) уровни значимости коэффициентов моделей и выбрать одну из них (обосновать выбор).
5. Для выбранной модели определить границы доверительных интервалов для значений линии регрессии (отобразить графически) в предположении, что остатки независимые нормальные случайные величины с одинаковой дисперсией.
6. Проверить качественно свойства остатков модели (нормальность, гомоскедастичность), построив гистограмму остатков и диаграмму рассеяния остатков в зависимости от значений переменной .

***Исходные данные: Вариант 18***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 12,98 | 22,16 | 19,95 | 22,57 | 18,82 | 21,97 | 5,87 | 25,08 | 13,58 | 16,45 |
| Y | 9,91 | 22,08 | 8,32 | 23,10 | 19,65 | 30,14 | 10,49 | 21,55 | 1,30 | 12,57 |
| X | 19,60 | 30,75 | 2,48 | 16,07 | 3,14 | 14,08 | 15,63 | 25,50 | 4,89 | 24,59 |
| Y | 11,60 | 20,98 | 0,59 | 11,20 | 19,10 | 28,33 | 3,56 | 12,77 | 0,12 | 4,71 |
| X | 14,89 | 23,27 | 19,21 | 26,41 | 8,78 | 19,87 | 18,79 | 21,73 | 7,68 | 21,94 |
| Y | 9,55 | 19,04 | 10,33 | 26,08 | 6,39 | 17,66 | 18,83 | 29,76 | 1,41 | 11,36 |
| X | 14,74 | 26,70 | 8,48 | 23,43 | 5,75 | 18,52 | 3,33 | 17,34 | 12,55 | 28,00 |
| Y | 13,34 | 20,73 | 12,14 | 22,91 | 4,72 | 21,04 | 0,47 | 3,15 | 7,65 | 20,12 |

***Ход работы:***

Строим диаграммы рассеяния величин  и  с нанесенными на них линиями соответствующих регрессий

Замечаем, что визуально и по значениям  наиболее предпочтительным являтся логарифмическая модель регрессии. Исследуем подробно каждую из моделей.

***Результаты регрессионного анализа для линейной модели :***

Вектор базисных функций:  , вектор коэффициентов уравнения регрессии: .

Оценку вектора коэффициентов уравнения регрессии метода наименьщих квадратов находим по формуле: , где , :

, , , .

Таким образом, оценка функции регрессии: .

Остаточная дисперсия модели регрессии:

.

Коэффициент детерминации модели и скорр. коэффициент детерминации:

, .

Значение статистики Фишера , соответствующий уровень значимости:

( - случайная величина, распределенная по закону Фишера с  и  числом степеней свободы). Следовательно, данное значение статистики является высокозначимым, то есть коэффициент детерминации модели значимо отличается от нуля.

Для линейной модели  значимость коэффициента детерминации равносильна значимости коэффициента уравнения регрессии , поэтому проверка значимости данного коэффициента уравнения регрессии на основе значения статистики Стьюдента избыточна.

***Результаты регрессионного анализа для квадратичной модели*:**

Вектор базисных функций:  , вектор коэффициентов уравнения регрессии: .

Оценку вектора коэффициентов уравнения регрессии метода наименьщих квадратов находим по формуле: , где , :

,, , .

Таким образом, оценка функции регрессии: .

Остаточная дисперсия модели регрессии:

.

Коэффициент детерминации модели и скорр. коэффициент детерминации:

0,692, 0,675.

Значение статистики Фишера 41,52, соответствующий уровень значимости: ( - случайная величина, распределенная по закону Фишера с  и  числом степеней свободы). Следовательно, данное значение статистики является высокозначимым, то есть коэффициент детерминации модели значимо отличается от нуля.

Проверяем значимость коэффициентов регрессионной модели. Для коэффициета значение статистики Стьюдента:

5,416.

Уровень значимости определяем как вероятность события , где  случайная величина, имеющая распределение Стьюдента с  степенями свободы: . Следовательно, можно признать, что значение высоко значимо отличается от нуля.

Для коэффициета  значение статистики Стьюдента:

-3,296

.

Соответствующий уровень значимости: 0,002. Следовательно, можно признать, что значение сильно значимо отличается от нуля.

Таким образом, мы получили для квадратичнй модели, что коэффициент высоко значим, а коэффициент сильно значим.

***Результаты регрессионного анализа для логарифмической модели*** :

Вектор базисных функций:  , вектор коэффициентов уравнения регрессии: .

Оценку вектора коэффициентов уравнения регрессии метода наименьщих квадратов находим по формуле , где , :

, , , .

Таким образом, оценка функции регрессии: .

Остаточная дисперсия модели регрессии:

.

Коэффициент детерминации модели и скорр. коэффициент детерминации:

0,751, 0,745.

Значение статистики Фишера 114,92, соответствующий уровень значимости: 5,97 ( - случайная величина, распределенная по закону Фишера с  и  числом степеней свободы). Следовательно, данное значение статистики является высокозначимым, то есть коэффициент детерминации модели значимо отличается от нуля.

Для модели  значимость коэффициента детерминации равносильна значимости коэффициента уравнения регрессии , поэтому проверка значимости данного коэффициента уравнения регрессии на основе значения статистики Стьюдента избыточна.

***Результаты регрессионного анализа для показательной модели*** :

Показательная модель  не является линейной, поэтому предварительно преобразуем модель прологарифмировав обе части уравнения регрессии:

 или , где: , , .

Результаты регрессионного анализа для модели :

Вектор базисных функций:  , вектор коэффициентов уравнения регрессии: .

Оценку вектора коэффициентов уравнения регрессии метода наименьщих квадратов находим по формуле , где , :

, , ,

.

Таким образом, оценка функции регрессии: .

Остаточную дисперсию для данной модели не оцениваем (имеет смысл оценивать данную величину для непреобразованной модели).

Коэффициент детерминации модели и скорр. коэффициент детерминации:

0,472 , 0,458.

Значение статистики Фишера 33,97, соответствующий уровень значимости: 2,64 ( - случайная величина, распределенная по закону Фишера с  и  числом степеней свободы). Следовательно, данное значение статистики является высокозначимым, то есть коэффициент детерминации модели значимо отличается от нуля.

Для модели  значимость коэффициента детерминации равносильна значимости коэффициента уравнения регрессии , поэтому проверка значимости данного коэффициента уравнения регрессии на основе значения статистики Стьюдента избыточна.

Оценки коэффициентов исходной модели :

11,66, .

Оценка функции регрессии исходной модели:

Остаточная дисперсия: .

***Сравним полученные модели:***

|  |  |  |  |
| --- | --- | --- | --- |
| Линейная модель | | | |
| R-квадрат (испр.) | Остат. дисперсия | Статистика F | Уровень знач-ти |
|  |  |  |  |
| Квадратичная модель | | | |
| R-квадрат (испр.) | Остат. дисперсия | Статистика F | P-значение |
| 0,675 |  | 41,52 |  |
| Логарифмическая модель | | | |
| R-квадрат (испр.) | Остат. дисперсия | Статистика F | P-значение |
| 0,745 |  | 114,92 | 5,97 |
| Показательная модель | | | |
| R-квадрат (испр.) | Остат. дисперсия | Статистика F | P-значение |
| 0,458 |  | 33,97 | 2,64 |

Все модели высоко значимы (), по величине коэффициента детерминации и значению стандартной ошибки наилучшей является логарифмическая модель, которую и выбираем в качестве регрессионной модели для наших данных.

***Находим границы доверительных интервалов*** уровня 0,95 для линии квадратичной регрессии (в предположении, что остатки независимые нормальные случайные величины с одинаковой дисперсией). Для этого вычисляем оценку дисперсии оценки значения функции регрессии в произвольной точке :

10,19074.

Определяем величину доверительного отклонения для вероятности  в точке : , где  – квантиль распределения Стьюдента уровня  ( с числом степеней свободы . Соответственно, получаем границы доверительного интервала в точке :

,

.

График линии регрессии с планками погрешностей в указанных точках на фоне корреляционного поля приведен на рисунке.

***Исследуем свойства остатков модели.***

Строим гистограмму остатков с наложенной на нее кривой плотности нормального распределения:

Видим, что гистограмма остатков не согласуется с кривой плотности нормального распределения.

Для исследования однородности остатков (гомоскедастичности) построим диаграмму рассеяния остатков от значений переменной :

Из приведенной диаграммы можно сделать вывод о достаточной однородности разброса остатков относительно значений переменной , то есть можно считать, что остатки гомоскедастичны.

**Выводы:**

В ходе проделанной работы была исследована выборочная совокупность данных на линейную, квадратичную, логарифмическую и показательную регрессию. В результате наилучшей была признана модель логарифмической регрессии:. Уровень значимости данной модели по Фишеру 5,97, коэффициент детерминации , то есть модель объясняет 74,5% разброса исходных данных. Для данной модели были определены границы 95% доверительного интервала для значений регрессии (на рассматриваемом интервале значений независимой переменной). По результатам исследования свойств остатков модели, можно говорить о высокой достоверности полученных результатов.