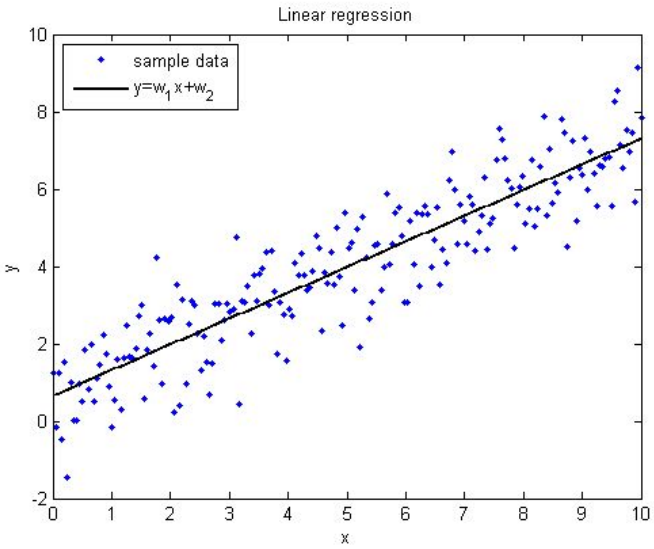
Лекция 6



Корреляционный и регрессионный анализ

Цель лекции: изучить основы корреляционного и регрессионного анализа и их реализацию в решении задач

План лекции:

1. Основы корреляционного и регрессионного анализа.
2. Линейные и нелинейные регрессии.
3. Реализация регрессионного анализа.
4. Реализация корреляционного анализа.

1. Основы корреляционного и регрессионного анализа.

Корреляция — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми).

Корреляционный анализ — метод обработки статистических данных, с помощью которого измеряется теснота связи между двумя или более переменными.

Ограничения корреляционного анализа:

1) Применение возможно при наличии достаточного количества наблюдений для изучения. На практике считается, что число наблюдений должно не менее чем в 56 раз превышать число факторов.

2) Необходимо, чтобы совокупность значений всех факторных и результативного признаков подчинялась многомерному нормальному распределению.

3) Исходная совокупность значений должна быть качественно однородной.

4) Сам по себе факт корреляционной зависимости не даёт основания утверждать, что одна из переменных предшествует или является причиной изменений, или то, что переменные вообще причинно связаны между собой, а не наблюдается действие третьего фактора.

Регрессия – зависимость среднего значения какой-либо случайной величины от некоторой другой величины или нескольких величин.

Регрессионный анализ – раздел математической статистики, объединяющий практические методы исследования регрессионной зависимости между величинами по данным статистических наблюдений.

Задача корреляционного анализа – определение тесноты и направления связи между изучаемыми величинами.

В ходе **регрессионного анализа** определяется аналитическое выражение связи зависимой случайной величины *Y* (результативный признак) с независимыми случайными величинами *X1, X2, …Xm* (факторами).

Практически речь идёт о том, чтобы, анализируя множество точек на графике (т.е. множество статистических данных), найти линию, по возможности точно отражающую заключённую в этом множестве закономерность, тенденцию – ***линию регрессии.***

Уравнение регрессии - это форма связи результативного признака *Y* с факторами *X1, X2, …Xm*. В зависимости от типа выбранного уравнения различают линейную и нелинейную (квадратичную, экспоненциальную, логарифмическую и т.д.) регрессию.

В зависимости от числа взаимосвязанных признаков различают парную и множественную регрессию.

Парная – исследуется связь между двумя признаками (результативным и факторным).

Множественная (многофакторная) – между тремя признаками (результативным и несколькими факторными).

Последовательность этапов регрессионного анализа

1) Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.
2) Определение зависимых и независимых (объясняющих) переменных.
3) Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель.
4) Формулировка гипотезы о форме связи (парная или множественная, линейная или нелинейная).
5) Определение ***функции регрессии*** (заключается в расчете численных значений параметров уравнения регрессии)
6) Оценка точности регрессионного анализа.
7) Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.
8) Предсказание неизвестных значений зависимой переменной.

2. Линейные и нелинейные регрессии.

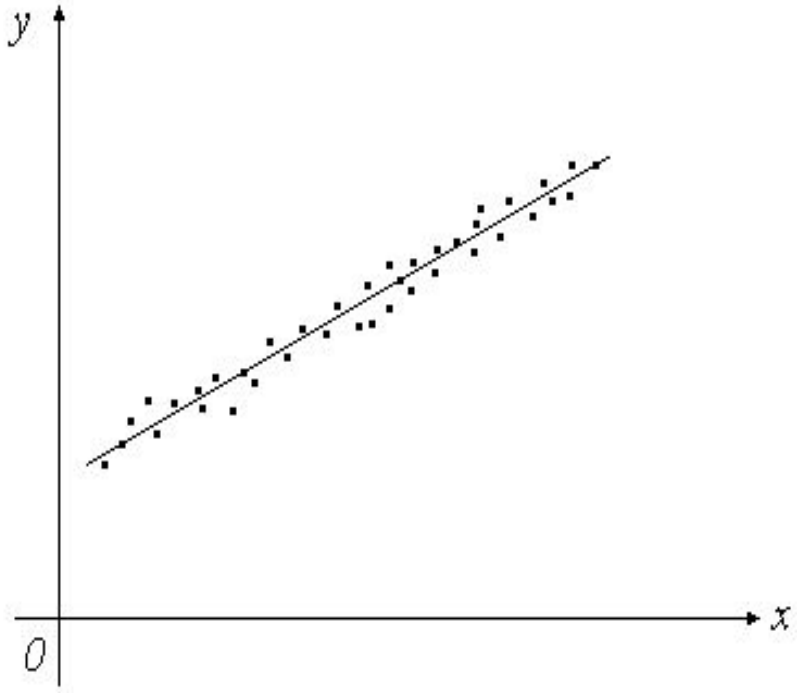


Рисунок 1 – Линейная регрессия

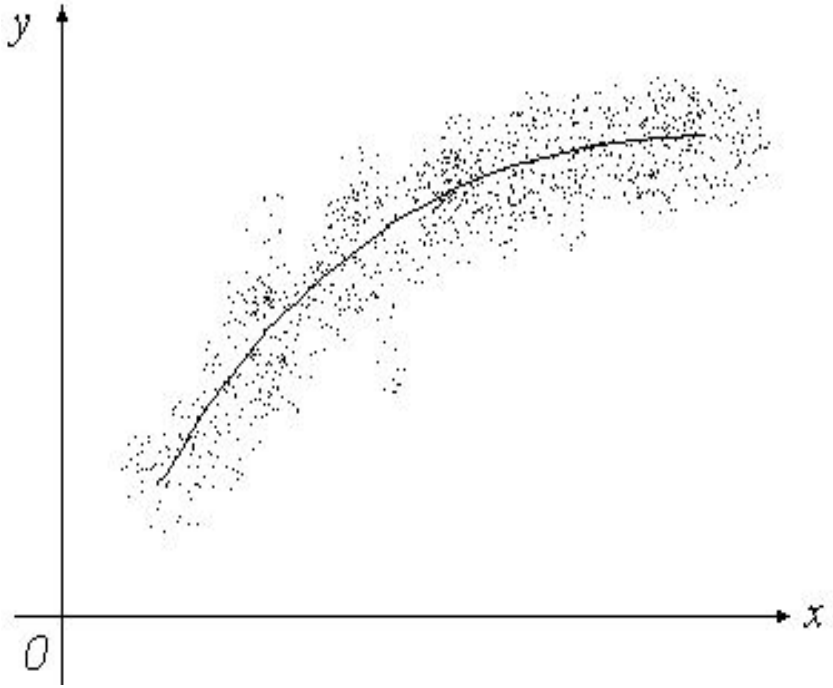


Рисунок 2 – Нелинейная регрессия

Линейная регрессия

При моделировании технологических процессов во многих случаях связь между входными (*x*) и выходными (*y*) параметрами можно аппроксимировать линейным полиномом (зависимостью)

$$\hat{y} = b_0 + b_1 \cdot x_i$$

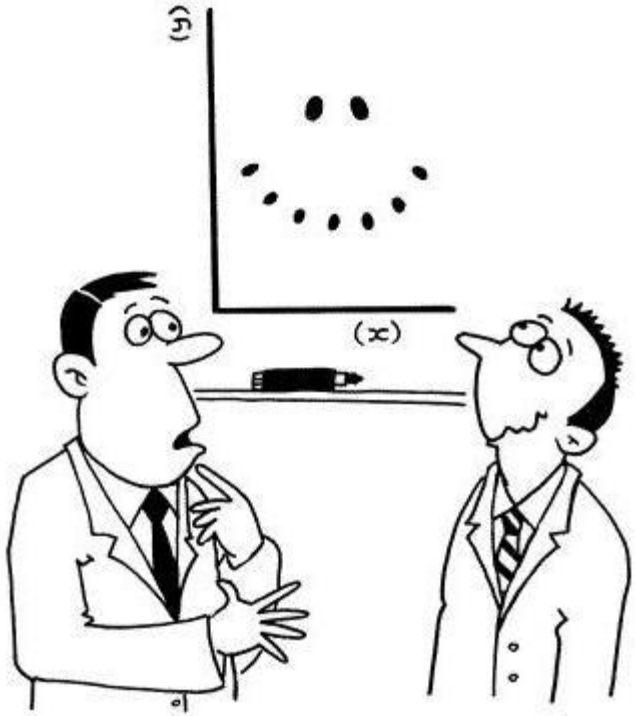
Для получения вида математической модели необходимо определить коэффициенты уравнения регрессии b_0 и b_1. Для этого применяется метод наименьших квадратов.

$$b_0 = \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} (x_i \cdot y_i)}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

$$b_1 = \frac{n \cdot \sum_{i=1}^{n} (x_i \cdot y_i) - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

Пример определения линейной регрессии

	X_i	Y_i	X_i^2	X_iY_i	Y_i-Y_{cp}	$(Y_i-Y_{cp})^2$	Функция
Значе-ния	1	30					
	2	7					
	3	8					
	4	1					
Сумма	**10**	**46**					



Нелинейная регрессия

1) Полиномиальная

$$y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$$

2) Гиперболическая

$$y = a + \frac{b}{x}$$

3) Степенная

$$y = a \cdot x^b$$

4) Показательная

$$y = a \cdot b^x$$

5) Экспоненциальная

$$y = b_0 \cdot e^{b_1 x}$$

3. Реализация регрессионного анализа.

Уравнение множественной линейной регрессии

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m,$$

где $\hat{y}$ – теоретические значения результативного признака, полученные путем подстановки соответствующих значений факторных признаков в уравнение регрессии;

$x_1, x_2, \ldots x_m$ – значения факторных признаков;

$a_0, a_1, \ldots a_m$ – параметры уравнения (коэффициенты регрессии).

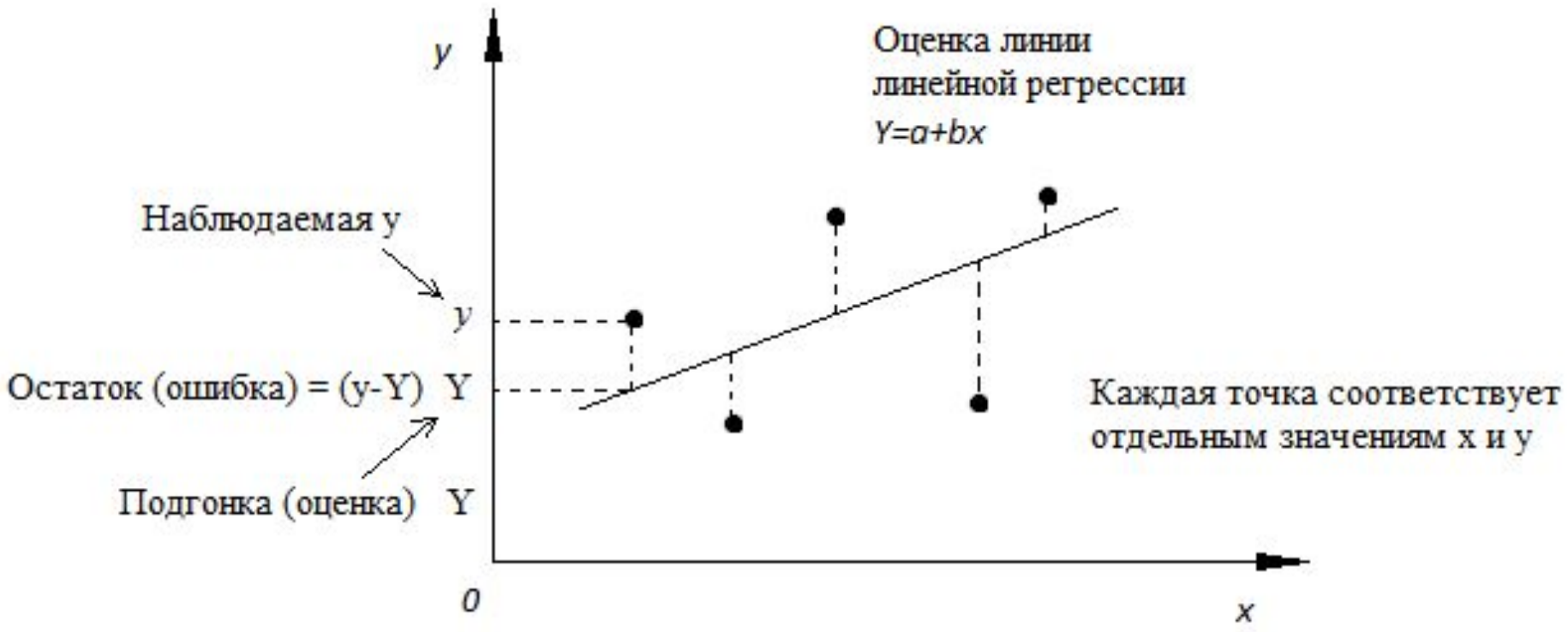


Рисунок - Линия линейной регрессии с изображенными остатками (вертикальные пунктирные линии) для каждой точки.

Метод наименьших квадратов

Параметры уравнения регрессии могут быть определены с помощью метода наименьших квадратов, который используется в пакете анализа данных «Регрессия» (MS Excel):

находятся параметры модели, при которых минимизируется сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по выбранному уравнению регрессии, т.е.

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2_i} - \ldots - a_m x_{mi})^2 \to \min.$$

Рассматривая S в качестве функции параметров a_i и проводя математические преобразования (дифференцирование), получаем систему нормальных уравнений с *m* неизвестными (по числу параметров a_i).

$$\begin{cases} \sum y = na_0 + a_1 \sum x_1 + a_2 \sum x_2 + ... + a_m \sum x_m, \\ \sum yx_1 = a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_2 x_1 + ... + a_m \sum x_m x_1, \\ ... \\ \sum yx_m = a_0 \sum x_m + a_1 \sum x_1 x_m + a_2 \sum x_2 x_m + ... + a_m \sum x_m^2. \end{cases}$$

Здесь *n* – число наблюдений, *m* – число факторов в уравнении регрессии.

Решение системы позволяет получить значения параметров регрессии a_i .

Для определения величины степени стохастической взаимосвязи результативного признака *Y* и факторов *X* необходимо знать следующие дисперсии:

общую дисперсию результативного признака *Y*, отображающую влияние как основных, так и остаточных факторов:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n},$$

где $\bar{y}$ - среднее значение результативного признака *Y*.

- ***факторную дисперсию*** результативного признака *Y*, отображающую влияние только основных факторов

$$\sigma_F^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n};$$

- ***остаточную дисперсию*** результативного признака *Y*, отображающую влияние только остаточных факторов

$$\sigma_O^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - (m+1)}.$$

При корреляционной связи результативного признака и факторов выполняется соотношение

$$\sigma_F^2 < \sigma_Y^2,$$

при этом

$$\sigma_Y^2 = \sigma_F^2 + \sigma_O^2.$$

Определение коэффициента детерминации R^2

Для анализа общего качества уравнения линейной многофакторной регрессии используют множественный коэффициент детерминации R^2 , называемый также квадратом коэффициента множественной корреляции *R*

$$R^2 = \frac{\sigma_F^2}{\sigma_y^2}$$

и определяет долю вариации результативного признака, обусловленную изменением факторных признаков, входящих в многофакторную регрессионную модель.

Величина *R-квадрат*, называемая также мерой определенности, характеризует качество полученной регрессионной прямой. Это качество выражается степенью соответствия между исходными данными и регрессионной моделью (расчетными данными). Мера определенности всегда находится в пределах интервала [0;1].

В большинстве случаев значение *R-квадрат* находится между этими значениями, называемыми экстремальными, т.е. между нулем и единицей.

Если значение *R-квадрата* близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных. И наоборот, значение *R-квадрата*, близкое к нулю, означает плохое качество построенной модели.

Множественный R - коэффициент множественной корреляции R - выражает степень зависимости независимых переменных (*X*) и зависимой переменной (*Y*).

Множественный R равен квадратному корню из коэффициента детерминации, эта величина принимает значения в интервале от нуля до единицы.

В простом линейном регрессионном анализе *множественный R* равен коэффициенту корреляции Пирсона.

Определение F критерия Фишера

Так как в большинстве случаев уравнение регрессии приходится строить на основе выборочных данных, то возникает вопрос об адекватности построенного уравнения данным генеральной совокупности. Для этого проводится проверка статистической значимости коэффициента детерминации R^2 на основе F-критерия Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m},$$

где n – число наблюдений;

m – число факторов в уравнении регрессии.

Если в уравнении регрессии свободный член $a_0 = 0$, то числитель n-m-1 следует увеличить на 1, т.е. он будет равен n-m.

Определение ошибки аппроксимации

Для оценки адекватности уравнения регрессии часто также используют показатель средней ошибки аппроксимации

$$\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}\right|}{y_i}\cdot 100\%.$$

Возможна ситуация, когда часть вычисленных коэффициентов регрессии не обладает необходимой степенью значимости, т.е. *значения данных коэффициентов будут меньше их стандартной ошибки*. В этом случае такие коэффициенты должны быть *исключены* из уравнения регрессии.

Поэтому проверка адекватности построенного уравнения регрессии наряду с проверкой значимости коэффициента детерминации R^2 включает также и проверку значимости каждого коэффициента регрессии.

Определение t-критерия

Для оценки адекватности уравнения регрессии часто также используют показатель средней ошибки аппроксимации

$$t = \frac{a_i}{\sigma_{a_i}},$$

где σ_{a_i} - стандартное значение ошибки для коэффициента регрессии a_i

В математической статистике доказывается, что если гипотеза $H_0 : a_i = 0$ выполняется, то величина t имеет распределение Стьюдента с *k*=*n*-*m*-1 числом степеней свободы, т.е.

$$\frac{a_i}{\sigma_{a_i}} = t(k = n - m - 1).$$

Гипотеза $H_0 : a_i = 0$ о незначимости коэффициента регрессии отвергается, если $|t_p| > |t_{kp}|$.

Определение границ доверительных интервалов

Зная значение t_{kp}, можно найти границы доверительных интервалов для коэффициентов регрессии

$$a_i^{\min} = a_i - t_{kp}\sigma_{a_i};$$

$$a_i^{\max} = a_i + t_{kp}\sigma_{a_i}.$$

Результаты регрессионного анализа, полученные с помощью MS Excel

Показывает, что 91,5% общей вариации результативного признака объясняется вариацией факторных признаков Xi.

Число

Определяется числ
количеством переме
k_o=n-(m+1)

СКО
эмпирическ
данных от
теоретических

Диспе
-факто
-остат

Расчетное значение критерия Фишера: долж
интервале($F_{кр}$
определяется
=FРАСПОБР(0

Уровень значимости: должен быть меньше 0,05

	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	*Регрессионная с*						
4	Множественный R	0,956697117					
5	R-квадрат	0,915269373					
6	Нормированный R-квадрат						
7	Стандартная ошибка						
8	Наблюдения						
9							
10	Дисперсионный анализ						
11						*F*	*Значимость F*
12	Регрессия	3	487339909	162446636,3	14,40281045	0,013074704	
13	Остаток	4	45115260,5	11278815,13			
14	Итого	7	532455169,5				
15							
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
17	Y-пересечение	-4326,547909	3824,113175	-1,131385948	0,321126121	-14943,9882	6290,8924
18	Переменная X 1	750,7161125	121,4702651	6,180245936	0,003482475	413,4605895	1087,97164
19	Переменная X 2	14,88213576	6,744514864	2,206553927	0,091978433	-3,84363952	33,607911
20	Переменная X 3	15,1512605	39,91165855	0,379619917	0,723521764	-95,6612685	125,96379

Оценка коэффициентов регрессии

	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	*Регрессионная статистика*						
4	Множественный R	0,956697117					
5	R-квадрат	0,915269373					
6	Нормированный R-квадрат	0,851721402					
7	Стандартная ошибка	3358,394724					
8	Наблюдения	8					
9							
10	Дисперсионный анализ						
11							
12	Регрессия						
13	Остаток						
14	Итого						
15							
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
17	Y-пересечение	-4326,547909	3824,113175	-1,131385948	0,321126121	-14943,9882	6290,8924
18	Переменная X 1	750,7161125	121,4702651	6,180245936	0,003482475	413,4605895	1087,97164
19	Переменная X 2	14,88213576	6,744514864	2,206553927	0,091978433	-3,84363952	33,607911
20	Переменная X 3	15,1512605	39,91165855	0,379619917	0,723521764	-95,6612685	125,96379

Значения
используемы
построени
регресси

Должн
меньш
знач
коэффи

Должен попа
область:
Опр
=СТЬЮДРА

Значение до
меньше
значимо

Показывает нижние и верхние границы доверительных интервалов. Не должен проходить через 0.

Таким образом, регрессионная модель будет иметь вид:

$$\hat{y} = 750 \cdot x_1$$

4. Реализация корреляционного анализа.

Определение коэффициента корреляции

Пусть *r* обозначает выборочный коэффициент корреляции, полученный по извлеченным из двумерного нормального распределения пар наблюдений $(x_1, y_1),\ldots,(x_n, y_n)$.

Коэффициент корреляции неизвестен, но может быть оценен по выборке с помощью выборочного коэффициента корреляции *r*:

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}}.$$

$Если\ r_{xy} \le 0,3 - корелляция\ слабая;$

$0,3 \le r_{xy} \le 0,7 - корелляция\ средняя;$

$r_{xy} > 0,7 - корелляция\ сильная.$

Проверка значимости коэффициента корреляции.

Нулевая гипотеза состоит в том, что коэффициент корреляции равен нулю, альтернативная - не равен нулю:

$$H_0: \ \rho = 0$$

$$H_1: \ \rho \neq 0$$

Очевидно, достаточно большое по *абсолютной* величине значение величины *r* будет стремиться опровергнуть нулевую гипотезу.

Возникает вопрос.

Насколько большое должно быть абсолютное значение величины *r*?

Для того чтобы проверить гипотезу, мы должны знать распределение величины *r*.

Собственное распределение величины *r* довольно сложное, поэтому мы применим преобразование:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

Итак, выборочное распределение этой статистики есть распределение Стьюдента с *n-2* степенями свободы.

При заданном уровне значимости (α) определяем критическое значение $t_{кр}$.

Принимаем решение об отклонении или не отклонении нулевой гипотезы:

- $|t| > t_{кр}$ яем H_0
- $|t| < t_{кр}$ оняем H_0

Вычисление уровня значимости коэффициента корреляции

Для определения фактического уровня значимости коэффициента корреляции запишем:

$$SL = P(T \geq |t|) + P(T \leq -|t|) = 2P(T \geq |t|)$$

Где T подчиняется распределению Стьюдента с $n-2$ степенями свободы, а значение величины t вычисляется в соответствии с формулой

Вычисление уровня значимости эквивалентно определению площади под правым и левым хвостами функции, ограниченной значениями $-t$ и t.