

Регрессионный анализ

$$Y = a + b * X; \text{ где:}$$

- Y – зависимая переменная,
- a - константа
- b - угловой коэффициент
- X – независимая переменная

Для многомерной регрессии:

$$Y = a + b_1 * X_1 + b_2 * X_2 + \dots + b_p * X_p$$

Регрессионный анализ

Для аналитического описания связи между признаками могут быть использованы следующие виды уравнений:

- 1) $\bar{y} = a_0 + a_1x$ – прямая, линейная функция;
- 2) $\bar{y} = a_0 + a_1x^2 + a_2x$ – парабола;
- 3) $\bar{y} = a_0 + a_1 \frac{1}{x}$ – гипербола;
- 4) $\bar{y} = a_0x^{a_1}$ – степенная функция;
- 5) $\bar{y} = \exp(a_0 + a_1x)$ – экспонента и др.

Регрессионный анализ

Смысл коэффициента регрессии

В общем случае коэффициент регрессии k показывает, как в среднем изменится *результативный признак* (Y), если *факторный признак* (X) увеличится на единицу .

Свойства коэффициента регрессии

- Коэффициент регрессии принимает любые значения.
- Коэффициент регрессии *не симметричен* , т.е. изменяется, если X и Y поменять местами.
- *Единицей измерения* коэффициента регрессии является отношение единицы измерения Y к единице измерения X ($[Y]/[X]$).
- Коэффициент регрессии *изменяется при изменении единиц измерения* X и Y .

Регрессионный анализ

Единица измерения коэффициента регрессии

В уравнении $Y = 87610 + 2984 X$
коэффициент регрессии равен 2984.

В каких единицах он измеряется?

Если результативный признак Y измеряется в *гривнах*, а факторный признак X в количестве рабочих (чел), то коэффициент регрессии измеряется в *гривнах на человека (грн/чел)*

Сравнение коэффициентов корреляции и регрессии

Коэффициент корреляции

- Принимает значения в диапазоне от -1 до +1
- Безразмерная величина
- Показывает силу связи между признаками
- Знак коэффициента говорит о направлении связи

Коэффициент регрессии

- Может принимать любые значения
- Привязан к единицам измерения обоих признаков
- Показывает структуру связи между признаками
- Знак коэффициента говорит о направлении связи

Регрессионный анализ

Величина *R-квадрат* - мера 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 $[0;1]$.

Регрессионная статистика	
Множественный R	0,99836371
R-квадрат	0,9967301
Нормированный R-квадрат	0,99632137
Стандартная ошибка	0,42404974
Наблюдения	10

	Коэффициенты	Стандартная ошибка	t-статистика
Y-пересечение	2,694545455	0,33176878	8,121757129
x	2,305454545	0,04668634	49,38177965

Наблюдение	Предсказанное y	Остатки
1	9,610909091	-0,610909
2	7,305454545	-0,305454
3	11,91636364	0,083636
4	14,22181818	0,778181
5	16,52727273	0,472727
6	18,83272727	0,167272
7	21,13818182	-0,138181
8	23,44363636	-0,043636



График остатков

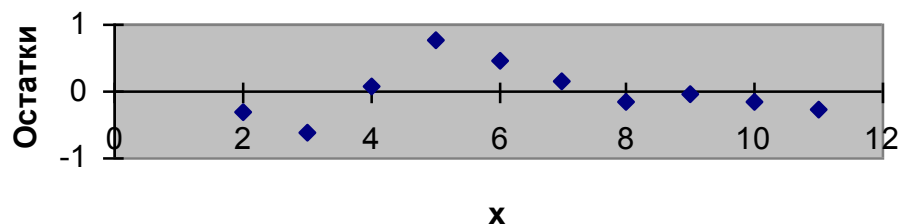
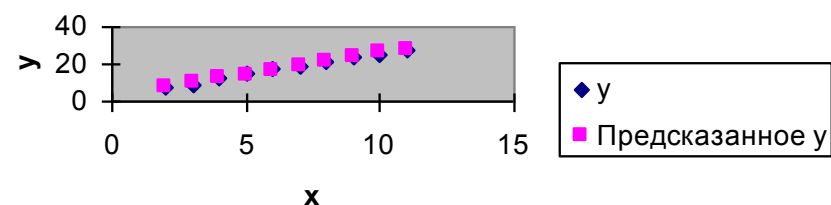


График подбора



F - критерий

Адекватность построенного уравнения данным генеральной совокупности проверяется по 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 $n-m$.

F - критерий

В математической статистике доказывается, что если гипотеза $H_0 : R^2 = 0$

выполняется, то величина *F* имеет *F*-распределение с $k=m$ и $l=n-m-1$ числом степеней свободы, т.е.

$$\frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = F(k=m, l=n-m-1).$$

Гипотеза H_0 о незначимости коэффициента детерминации R^2 отвергается, если

$$F_p > F_{\alpha, m, n-m-1}^{\text{êð}}.$$

При значениях $R^2 > 0,7$ считается, что вариация результативного признака *Y* обусловлена в основном влиянием включенных в регрессионную модель факторов *X*.

Ошибка аппроксимации

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

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

Регрессионный анализ

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.

t-критерий

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

t-критерий

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 $|t_p| > |t_{\hat{e}p}|$.

Границы доверительных интервалов

Зная значение $t_{\hat{e}p}$, можно найти границы доверительных интервалов для коэффициентов регрессии

$$a_i^{\min} = a_i - t_{\hat{e}p} \sigma_{a_i};$$

$$a_i^{\max} = a_i + t_{\hat{e}p} \sigma_{a_i}.$$



	ORP	ZPL	OEX
АРК	22596	2849	849275
Вінницька	22732	2651	622175
Волинська	10187	2580	585473
Дніпропетровська	202318	3335	8967464
Донецька	205594	3755	11335389
Житомирська	15665	2561	550107
Закарпатська	9224	2553	1205576
Запорізька	75835	3142	3320827
Івано-Франк.	20487	2679	435835
Київська	40280	3351	1787605

Кіровоградська	14382	2607
Луганська	67740	3307
Львівська	30845	2707
Миколаївська	21645	3007
Одеська	25016	2907
Полтавська	63646	2907
Рівненська	14474	2807
Сумська	22332	2707
Тернопільська	7609	2307
Харківська	62815	2907
Херсонська	10828	2407
Хмельницька	16524	2607
Черкаська	28249	2607
Чернівецька	4011	2407
Чернігівська	14129	2507
м. Ки	73284	5007
м.Севастополь	3652	3107

The regression equation is

$$\text{ORP} = 152770 - 54.3 \text{ ZPL} + 0.0222 \text{ OEX}$$

Predictor	Coef	StDev	T	P
Constant	152770	37322	4.09	0.000
ZPL	-54.27	14.11	-3.85	0.001
OEX	0.022164	0.002404	9.22	0.000

S = 20737 R-Sq = 85.2% R-Sq(adj) = 83.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	59214393186	29607196593	68.85	0.000
Error	24	10320502145	430020923		
Total	26	69534895331			

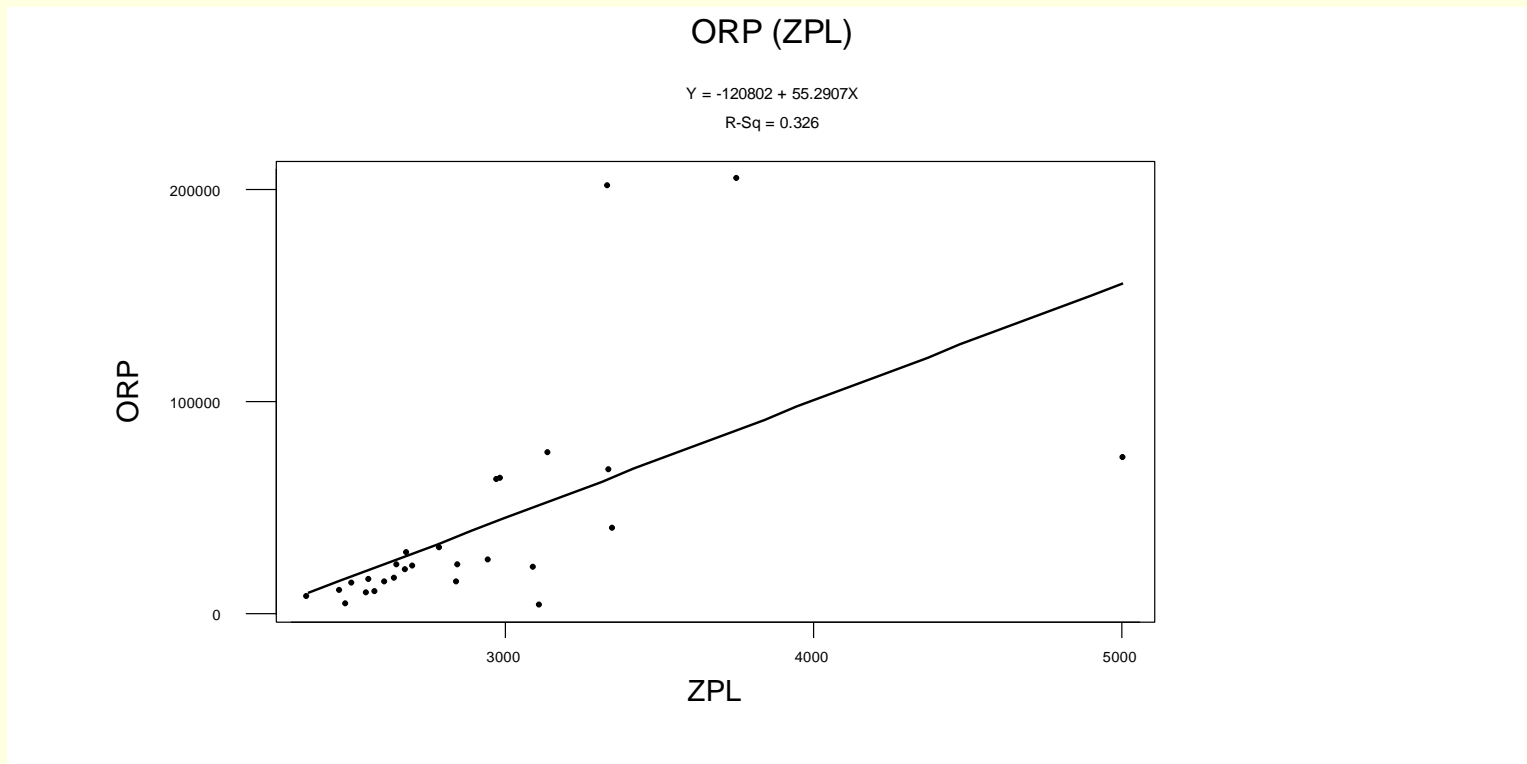
Source	DF	Seq SS
ZPL	1	22657300530
OEX	1	36557092656

Unusual Observations

Obs	ZPL	ORP	Fit	StDev Fit	Residual	St Resid
4	3335	202318	170527	12655	31791	1.94 X
5	3755	205594	200216	14391	5378	0.36 X
26	5007	73284	124629	16699	-51346	-4.18RX

R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

1. Методы первичной обработки данных



1. Методы первичной обработки данных


The regression equation is
 $y = -120802 + 55.3 x$

Predictor	Coef	StDev	T	P
Constant	-120802	47278	-2.56	0.017
x	55.29	15.91	3.48	0.002

S = 43302 R-Sq = 32.6% R-Sq(adj) = 29.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	22657300530	22657300530	12.08	0.002
Error	25	46877594802	1875103792		
Total	26	69534895331			

 (Ctrl) ▾

$$MS = \frac{SS}{df}$$

$$F_p = \frac{MS(\text{Регрессия})}{MS(\text{Остатки})}$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

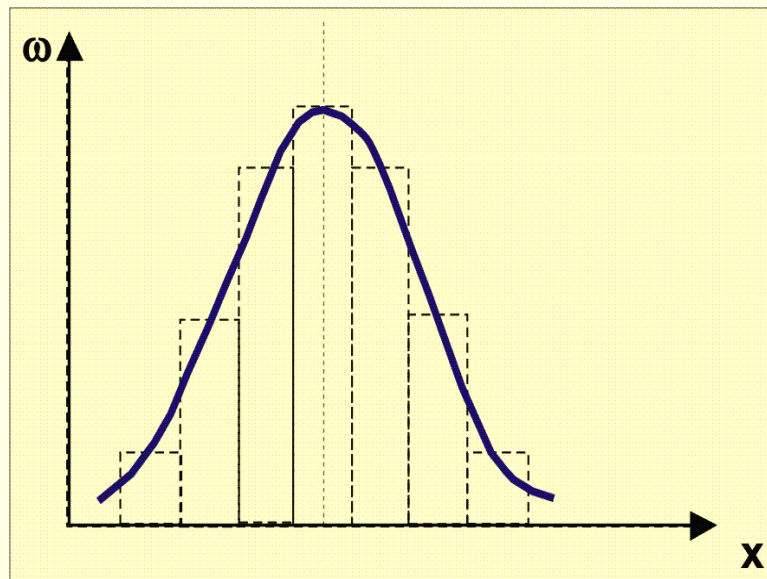
Общая сумма квадратов
отклонений

= Сумма квадратов
отклонений, объясненная
регрессией + Остаточная сумма
квадратов
отклонений

$$= \text{FРАСП}(F_p; df(\text{регрессия}); df(\text{остаток})).$$

Нормальный закон распределения

- 1) Количество вариантов (значений СВ), превышающих среднее значение, равно количеству вариантов, которые меньше его (примерная симметричность диаграммы).
- 2) Частота вариантов тем больше, чем ближе к среднему значению они расположены (гистограмма имеет наибольшие ординаты в центре и наименьшие – у краев).

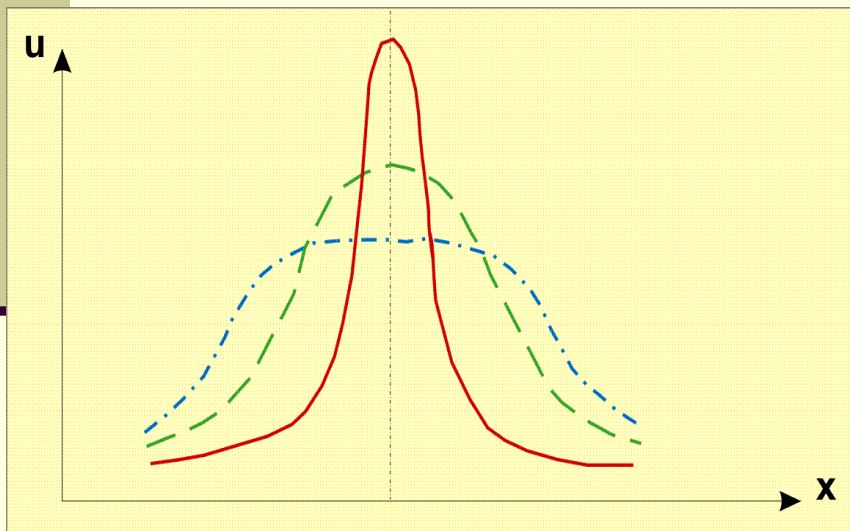


1. Хорошо изучен, методика проста и отработана
2. При увеличении объёма выборки

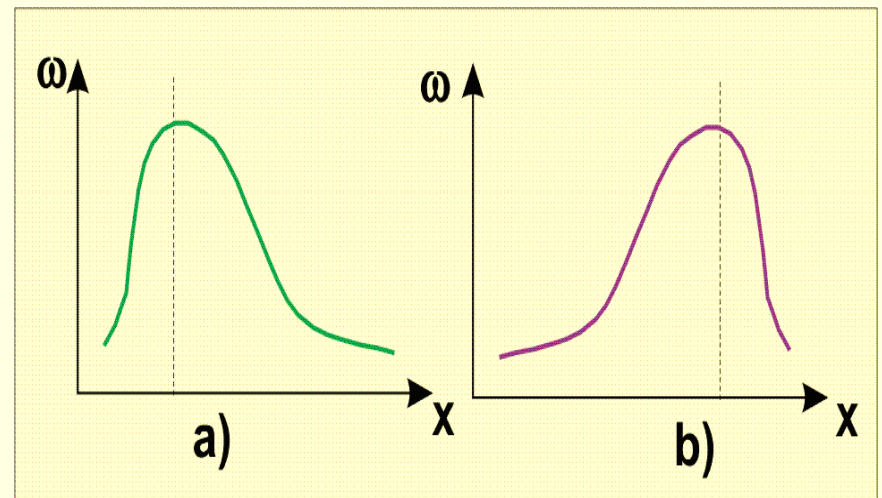
Если результаты измерений вызывают сомнение в применимости НЗ, необходимо увеличить объём выборки

По своему виду кривые нормального распределения могут быть:

- нормальновыпуклыми;
- туповыпуклыми;
- островершинными
- иметь положительную асимметрию;
- иметь отрицательную асимметрию.



Кривые нормального распределения



Кривые нормального распределения с положительной и отрицательной асимметрией