Лекционный курс по дисциплине: «Статистические методы обработки данных»

Что нужно знать, чтобы получить на экзамене от 4 до 6 баллов.

Шкалы измерений

- Номинальная шкала (шкала наименований). Эта шкала используется только для того, чтобы отнести объект или индивидуум в определенный класс (Распределения учащихся по классам, по половому признаку, по месту жительства, по видам спорта)
- Порядковая шкала. Эта шкала в дополнение к функции отнесения объектов в определенный класс также упорядочивает классы по степени выраженности заданного свойства (учащихся ранжировать по количеству правильно выполненных тестовых заданий)
- Интервальная шкала. Эта шкала позволяет не только классифицировать и упорядочивать объекты и индивидуумы, но и количественно оценивать различие между классами (Шкалы на большинстве физических приборов Шкала коэффициента интеллекта IQ)
- Шкала отношений. Эта шкала отличается от интервальной шкалы лишь тем, что в ней задано абсолютное начало отсчета (отношений являются меры длины (м, см и т. д.) и массы (кг, г и т. д.). Предмет длиной 100 см вдвое длиннее предмета длиной 50 см.)

Математическое ожидание

- Если совокупность случайных величин задана в виде набора дискретных значений, то математическое ожидание случайной величины определяется как среднее значение по выборке:

$$\mu = \sum_{i=1}^{N} p_i x_i$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Дисперсия

- Числовой характеристикой, показывающей степень разброса значений случайной величины относительно математического ожидания, называется дисперсия

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 \, p_i$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Среднеквадратическое отклонение

- Поскольку дисперсия имеет размерность квадрата случайной величины, то для характеристики меры рассеяния значений случайной величины относительно математического ожидания пользуются среднеквадратическим отклонением σ, равным значению квадратного корня из дисперсии:

$$\sigma = \sqrt{\sigma^2}$$

Выборочное среднее, дисперсия и среднеквадратическое отклонение

- Выборочное среднее, представляющее собой оценку математического ожидания генеральной совокупности:

$$\overline{x} = m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Выборочная дисперсия, служащая несмещенной оценкой дисперсии генеральной совокупности:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

- Выборочное среднеквадратическое (стандартное) отклонение:

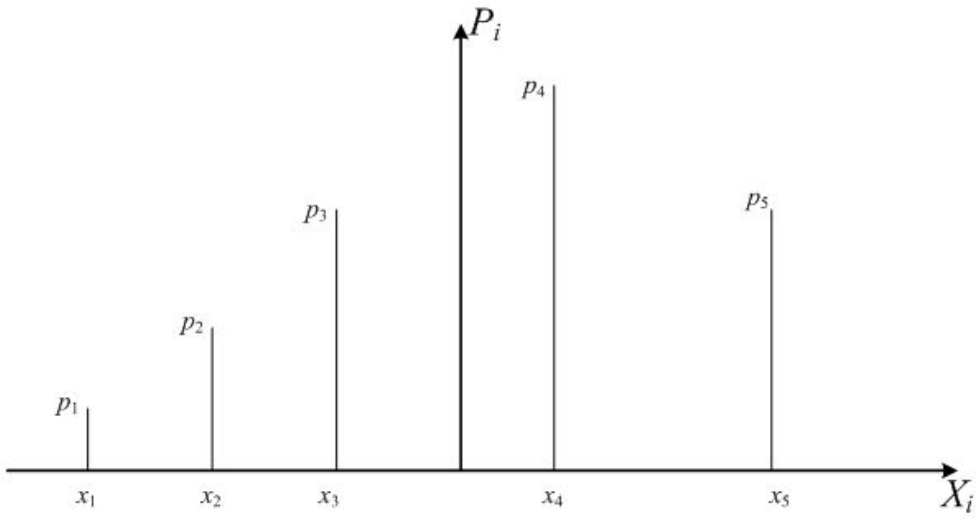
$$s = \sqrt{s^2}$$

Понятие закона распределения

- Полное описание случайной величины дается законом распределения, который устанавливает зависимость между возможными значениями случайной величины и их вероятностями

Задание закона распределения

Закон распределения случайной величины можно задать в виде графика, таблицы или аналитического выражения:



X_i	X_1	X_2	X_3	X_4	X_5
P_i	P_1	P_2	P_3	P_4	P_5

$$P = f(x)$$

Нормальное распределение

Нормальное распределение величины x описывается следующей функцией:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-m)^2}{2\sigma^2} \right]$$

Характеристики распределения Гаусса:

- оно симметрично относительно m
- имеет максимум равный $1/\sqrt{2\pi\sigma^2}$
- монотонно убывает при возрастании $|x-m|$

Нормальное распределение

- Функция распределения, показывающая вероятность случайной величине принять значение меньшее x, определяется выражением

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - m_x)^2}{2\sigma^2} \right) \partial x .$$

Нормальное распределение

При $m_x = 0$ и $\sigma^2 = 1$ имеет место стандартное нормальное распределение, для которого

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \partial x .$$

Нормальное распределение

Функция распределения стандартного нормального распределения носит название интеграла вероятности $\Phi(x)$ и часто используется для определения вероятности нахождения значений случайной величины в заданном интервале (c,d):

$$P(c \le X \le d) = \int_c^d \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \partial x = \Phi(d) - \Phi(c).$$

Нормальное распределение

Числовые характеристики нормального распределения:

– математическое ожидание, мода, медиана:

$$m_x = x_m = x_{med}\,.$$

– дисперсия:

$$D = \sigma^2\,.$$

– коэффициент асимметрии:

$$A = 0\,.$$

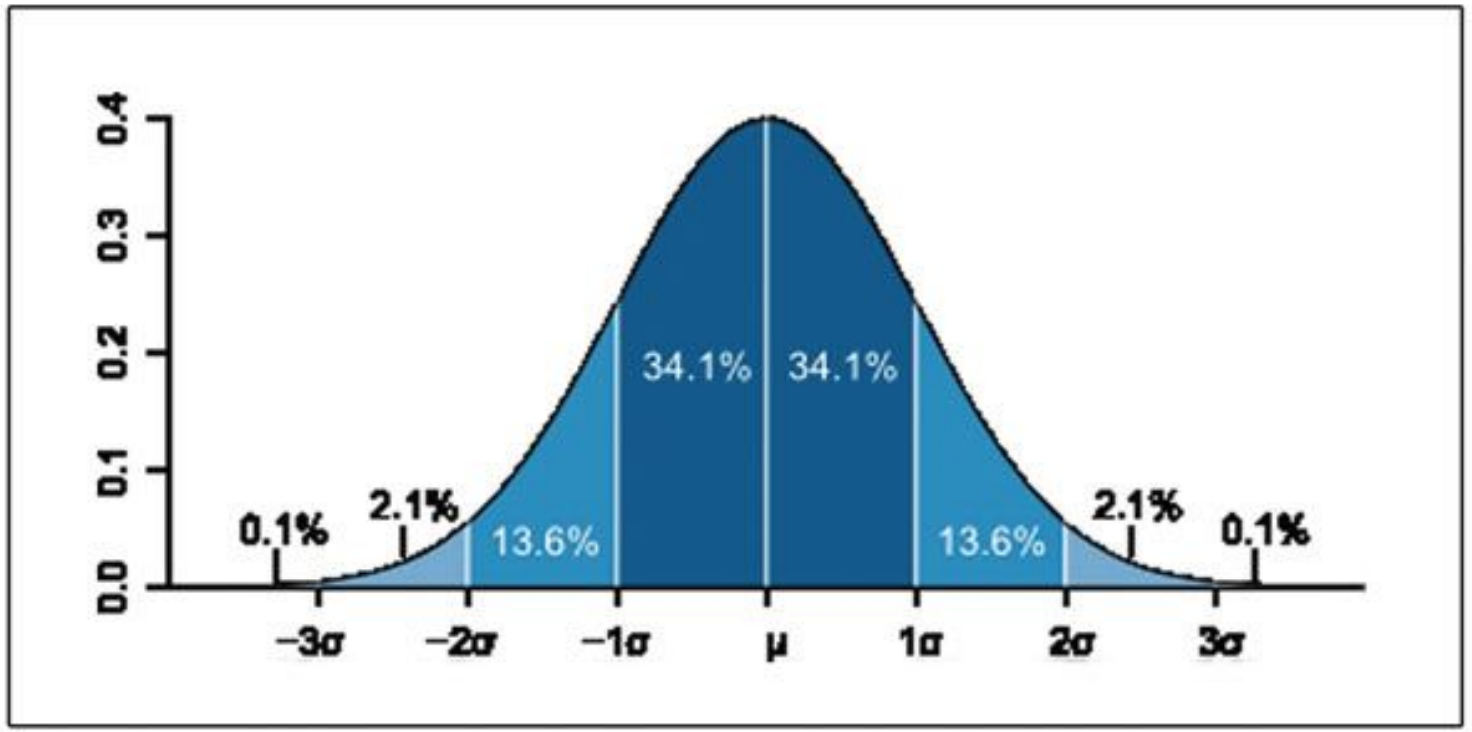
– коэффициент эксцесса:

$$E = 0\,.$$

Доверительная вероятность при нормальном распределении

- Если случайная величина распределена по нормальному закону с математическим ожиданием μ и средним квадратическим отклонением σ, то вероятности ее попадания в интервалы между $(\mu_s + \sigma_s)$ и $(\mu_s - \sigma_s)$; между $(\mu_s + 2\sigma_s)$ и $(\mu_s - 2\sigma_s)$; между $(\mu_s + 3\sigma_s)$ и $(\mu_s - 3\sigma_s)$ равны соответственно: 0,683; 0,955; 0,997

Доверительная вероятность при нормальном распределении



Распределение χ^2

Случайной величиной χ_l^2 называют сумму квадратов l независимых случайных величин, $X_1^2, X_2^2, \ldots, X_l^2$ каждая из которых распределена по нормальному закону с нулевым математическим ожиданием и единичной дисперсией. Число l называют числом степеней свободы случайной величины χ^2.

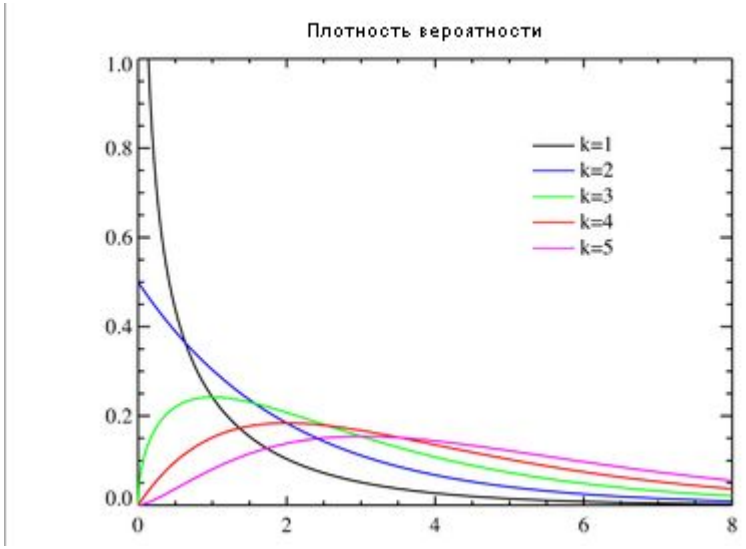
Плотность вероятности распределения χ^2 имеет вид

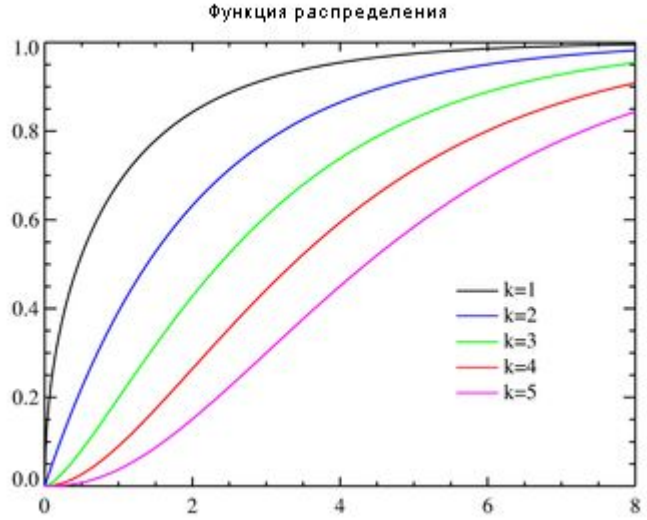
$$P(\chi^2) = \frac{1}{2^{\frac{l}{2}}\Gamma\left(\frac{l}{2}\right)} (\chi^2)^{\frac{l}{2}-1} e^{-\frac{\chi^2}{2}}, \quad 0 < \chi^2 < \infty$$

где $\Gamma\left(\frac{l}{2}\right)$ — гамма-функция.

Распределение χ^2

Распределение χ^2 определяется только одним параметром — числом степеней свободы, $M[\chi^2] = l$, $D[\chi^2] = 2l$.





Распределение Стьюдента

Случайная величина

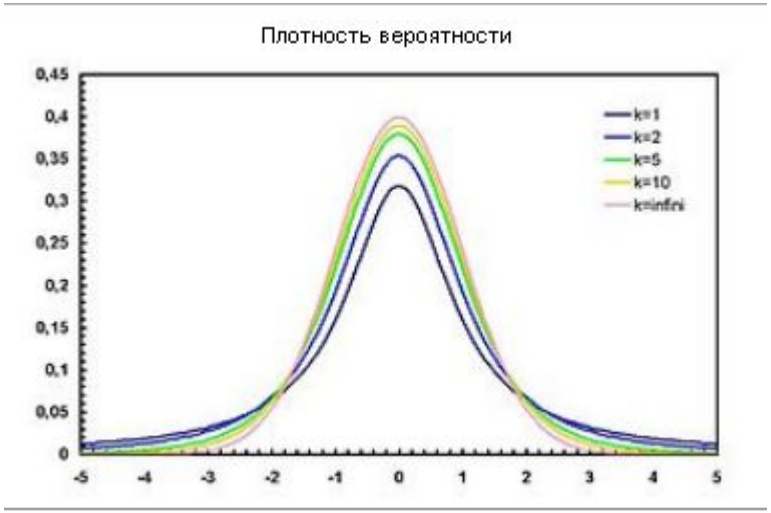
$$T = \frac{Z}{\sqrt{\frac{\chi_l^2}{l}}},$$

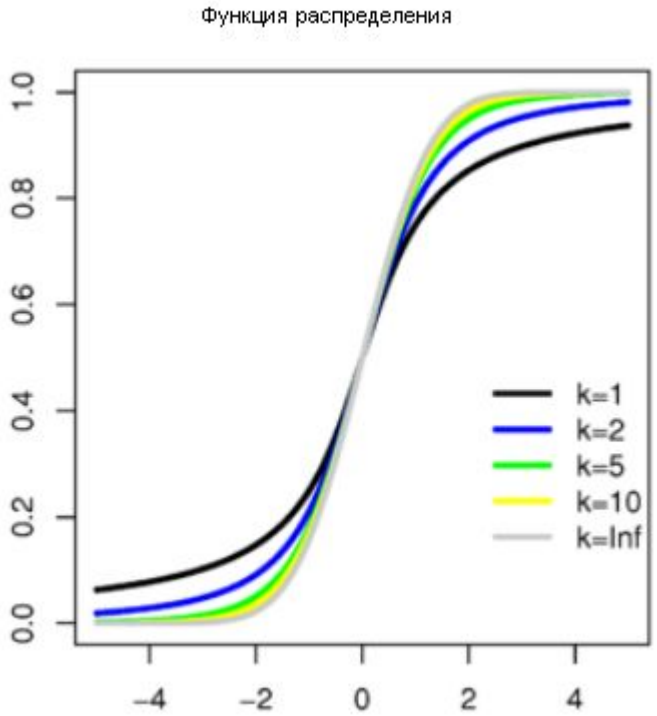
где Z — нормально распределенная случайная величина с $M[Z] = 0$ и $D[Z] = 1$, а χ_l^2 — случайная величина χ^2 с l степенями свободы, подчиняется распределению Стьюдента

$$p(t) = \frac{\Gamma\left(\frac{l+1}{2}\right)}{\sqrt{\pi l}\,\Gamma\left(\frac{1}{2}\right)}\left(1 + \frac{t^2}{l}\right)^{-\frac{l+1}{2}}, \quad -\infty < t < \infty.$$

Распределение Стьюдента

Распределение Стьюдента определяется одним параметром l, $M[T]=0$, $D[T]=\frac{l}{l-2}$ при $l>2$.





Проверка статистических гипотез

- Для того чтобы иметь основания принять или отвергнуть рассматриваемую гипотезу необходимо выработать некоторый критерий, который называют критерием согласия проверяемой гипотезы с результатами эксперимента

Критерий согласия χ^2 (хи-квадрат)

- В качестве меры расхождения между эмпирическим и теоретическим законами распределения Пирсоном была предложена статистика

$$\chi^2 = \sum_{k=1}^{m} \frac{(n_k - np_k)^2}{np_k}$$

Здесь: m — число значений, принятых случайной величиной, n – общее число наблюдений, p_k — вероятность появления k-го значения в теоретическом законе распределения

Непараметрический критерий Вилкоксона для проверки однородности двух независимых выборок

Большинство непараметрических критериев основано на использовании рангов наблюдений.

- Рангом наблюдения называют тот номер, который получит это наблюдение в упорядоченной совокупности всех данных после их упорядочения по определенному правилу, например от меньших значений к большим или наоборот.

Ранги и ранжирование

Трудности в назначении рангов возникают, если среди элементов выборки встречаются совпадающие. В этом случае обычно используют средние ранги.

Непараметрический критерий Вилкоксона

В критерии Вилкоксона в качестве в качестве статистики используется случайная величина

$$W = R_1 + R_2 + \ldots + R_n$$

Здесь R_j – ранги наблюдений второй выборки в общей объединенной выборке.

Непараметрический критерий Вилкоксона

Для проверки с уровнем значимости α гипотезы H_0 об однородности выборок при альтернативной гипотезе H_1: $F_x(x) > F_y(y)$ по имеющимся таблицам находят верхнее критическое значение $w_в(\alpha, m, n)$ статистики W, т. е. такое значение, для которого

$$P\left(W \geq w_в\left(\alpha, m, n\right)\right) = \alpha$$

Гипотезу об однородности выборок следует отвергнуть с уровнем значимости α, если рассчитанное значение статистики W больше критического значения.

Критерий Вилкоксона для проверки однородности двух зависимых выборок

Порядок применения критерия следующий:

1. Вычисляются абсолютные разности наблюдений в паре:

$$|z_i| = |x_{i2} - x_{i1}|, \qquad i = 1,\ldots,n$$

2. Осуществляется ранжирование этих разностей в порядке возрастания и каждому значению ранга присваивается знак его разности.

Критерий Вилкоксона для проверки однородности двух зависимых выборок

3. Вычисляется сумма значений рангов, которая образует статистику T.

4. Проверяется, принадлежит ли вычисленное значение T критической области, границы которой находятся по таблицам процентных точек распределения Вилкоксона для парных выборок.

Критерий Вилкоксона для проверки однородности двух зависимых выборок

Если вычисленное значение статистики T

$$T \geq t\left(\frac{\alpha}{2}, n\right) \quad \text{или} \quad T \leq \frac{n(n+1)}{2} - t\left(\frac{\alpha}{2}, n\right)$$

то гипотеза об однородности двух выборок отклоняется при уровне значимости α в пользу альтернативной гипотезы H_1: выборки неоднородны.

При альтернативной гипотезе H_1: распределение разности смещено вправо относительно нуля, гипотеза об однородности отклоняется, если вычисленное значение статистики T превышает критическое значение

$$T \geq t(\alpha, n)$$

Однофакторный дисперсионный анализ. Проверка гипотезы о влиянии фактора на исследуемую величину

Рассмотрим простейший случай дисперсионного анализа, когда изучается влияние на исследуемую величину какого-либо одного фактора *A*. Будем считать, что фактор *A* изучается на *k* уровнях A_1, A_2, ..., A_k. Пусть для простоты рассмотрения на каждом уровне производится одинаковое число *n* наблюдений исследуемой величины.

Проверка гипотезы о влиянии фактора на исследуемую величину

Оценка генерального среднего

$$\mu = \bar{x}.. = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}$$

Несмещенная оценка дисперсии генеральной совокупности

$$\sigma^2 = s^2 = \frac{1}{nk-1}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij} - \bar{x}..\right)^2$$

Проверка гипотезы о влиянии фактора на исследуемую величину

При справедливости нулевой гипотезы любая из выборочных дисперсий дает одинаково хорошую оценку. Поэтому в качестве оценки дисперсии генеральной совокупности возьмем среднее выборочных дисперсий. Эта оценка называется внутри групповой дисперсией:

$$s_0^2 = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_{\cdot j}\right)^2\right)$$

Проверка гипотезы о влиянии фактора на исследуемую величину

Оценим теперь дисперсию совокупности по выборочным средним. Поскольку мы предположили, что все выборки извлечены из одной совокупности, то стандартное отклонение выборочных средних будет служить оценкой ошибки среднего:

$$s_x = \frac{s}{\sqrt{n}}$$

Отсюда находим межгрупповую оценку дисперсии

$$s_A^2 = ns_x^2 = n\sum_{j=1}^{k} \frac{\left(\bar{x}_{.j} - \bar{x}_{..}\right)^2}{k-1}$$

Проверка гипотезы о влиянии фактора на исследуемую величину

В результате задача проверки гипотезы H_0 сводится к проверке гипотезы о равенстве дисперсий s_A^2 и s_0^2. При справедливости допущения о нормальном распределении случайных величин ε_{ij} отношение

$$F = \frac{s_A^2}{s_0^2}$$

в случае справедливости нулевой гипотезы подчиняется F-распределению с $l_1 = k\text{-}1$ и $l_2 = k(n\text{-}1)$ числом степеней свободы.

Проверка гипотезы о влиянии фактора на исследуемую величину

Влияние фактора *A* на исследуемый признак считается значимым с уровнем значимости *α*, если

$$\frac{s_A^2}{s_0^2} > f_{k-1;k(n-1);\alpha}$$

т. е. когда расчетное значение статистики *F* превышает значение α-процентной точки распределения Фишера.

Проверка гипотезы о влиянии фактора на исследуемую величину

Результаты дисперсионного анализа в общем случае обычно представляют в виде следующей таблицы

Источник дисперсии	Сумма квадратов	Степени свободы	Дисперсии	*F* отношение
Между группами	$CK_A = \sum_{i=1}^{n} n_j \left(\bar{x}_{.j} - \bar{x}_{..}\right)^2$	$k-1$	$s_A^2 = \frac{CK_A}{k-1}$	$F = \frac{s_A^2}{s_0^2}$
Внутри групп	$CK_0 = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_{.j}\right)^2$	$\sum_{j=1}^{k} n_j - k$	$s_0^2 = \frac{CK_0}{\sum_{j=1}^{k} n_j - k}$	
Полная	$CK_n = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_{..}\right)^2$	$\sum_{j=1}^{k} n_j - 1$		

Двухфакторный дисперсионный анализ. Виды взаимосвязи между двумя факторами

Пусть на исследуемую величину могут оказывать влияние два фактора *A* и *B*, каждый из которых имеет конечное число уровней. При этом ставится вопрос, как влияют и влияют ли вообще эти факторы на исследуемую величину. Здесь уже необходимо уделить внимание способу взаимосвязи факторов. Для большинства практических задач достаточно ограничиться двумя способами: пересечением и группировкой.

Виды взаимосвязи между двумя факторами

Два фактора *A* и *B* называются пересекающимися, если в плане эксперимента предусмотрены все возможные сочетания факторов.

	A_1	...	A_i	...	A_k
$B1$	x_{111}	...	x_{i11}	...	x_{k11}
	...	...	...	...	...
	x_{11t}	...	x_{i1t}	...	x_{k1t}
	...	...	...	...	...
	x_{11m}	...	x_{i1m}	...	x_{k1m}
...	...	...	...	...	...
Bj	x_{1j1}	...	x_{ij1}	...	x_{kj1}
	...	...	...	...	...
	x_{1jt}	...	x_{ijt}	...	x_{kjt}
	...	...	...	...	...
	x_{1jm}	...	x_{ijm}	...	x_{kjm}
...	...	...	...	...	...
Bn	x_{1n1}	...	x_{in1}	...	x_{kn1}
	...	...	...	...	...
	x_{1nt}	...	x_{int}	...	x_{knt}
	...	...	...	...	...
	x_{1nm}	...	x_{inm}	...	x_{knm}

Виды взаимосвязи между двумя факторами

Фактор *B* группируется фактором *A*, если каждый уровень фактора *B* сочетается не более, чем с одним уровнем фактора *A*.

A_1		A_2		A_3	
B_1	B_2	B_3	B_4	B_5	B_6
x_{111}	x_{121}	x_{231}	x_{241}	x_{351}	x_{361}
…	…	…	…	…	…
x_{11t}	x_{12t}	x_{23t}	x_{24t}	x_{35t}	x_{36t}
…	…	…	…	…	…
x_{11m}	x_{12m}	x_{23m}	x_{24m}	x_{35m}	x_{36m}

Двухфакторный дисперсионный анализ с пересечением уровней

Рассматривая совокупность данных как одну выборку из генеральной совокупности, получим оценку генерального среднего в виде

$$\mu = \overline{x}... = \frac{1}{knm}\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{t=1}^{m} x_{ijt}$$

и несмещенную оценку дисперсии генеральной совокупности

$$\sigma^2 = s^2 = \frac{1}{knm-1}\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{t=1}^{m}\left(x_{ijt} - \overline{x}...\right)^2$$

Двухфакторный дисперсионный анализ с пересечением уровней

Входящую в оценку дисперсии генеральной совокупности сумму квадратов можно представить в виде суммы четырех отдельных сумм квадратов $СК_A$, $СК_B$, $СК_{AB}$, $СК_0$:

характеризует разброс наблюдаемых значений между столбцами (уровнями фактора А) таблицы данных

$$СК_A = nm\sum_{i=1}^{k}\left(x_{i}.. - \overline{x}...\right)^2$$

характеризует разброс наблюдаемых значений между строками (уровнями фактора В) таблицы

$$СК_B = mk\sum_{j=1}^{n}\left(x_{.j}. - \overline{x}...\right)^2$$

Двухфакторный дисперсионный анализ с пересечением уровней

характеризует эффект взаимодействия факторов

$$CK_{AB} = m\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\overline{x}_{ij}. - \overline{x}_{i}.. - \overline{x}_{.j}. - \overline{x}...\right)^2$$

остаточная сумма квадратов

$$CK_{0} = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{t=1}^{m}\left(x_{ijt} - \overline{x}_{ij}.\right)^2$$

Двухфакторный дисперсионный анализ с пересечением уровней

С учетом числа степеней свободы каждой суммы квадратов, получим следующие выражения для оценок дисперсий:

$$\sigma_A^2 = \frac{1}{k-1}$$

$$\sigma_{AB}^2 = \frac{1}{(k-1)(n-1)}$$

$$\sigma_B^2 = \frac{1}{n-1}$$

B

$$\sigma_0^2 = \frac{1}{kn(m-1)}$$

0

Двухфакторный дисперсионный анализ с пересечением уровней

Гипотеза $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$ проверяется с помощью отношения

$$F = \frac{s_A^2}{s_0^2}$$

Гипотеза $H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0$ проверяется с помощью отношения

$$F = \frac{s_B^2}{s_0^2}$$

Двухфакторный дисперсионный анализ с пересечением уровней

Гипотеза об отсутствии взаимодействия между факторами (гипотеза об аддитивности) проверяется с помощью отношения

$$F = \frac{s_{AB}^2}{s_0^2}$$

Двухфакторный дисперсионный анализ с пересечением уровней

Результаты дисперсионного анализа представляют следующей таблицей

Источник дисперсии	Сумма квадратов	Степени свободы	Дисперсии
Фактор A	$СК_A$	$k-1$	$s_A^2 = \frac{СК_A}{k-1}$
Фактор B	$СК_B$	$n-1$	$s_B^2 = \frac{СК_B}{n-1}$
Взаимодействие	$СК_{AB}$	$(k-1)(n-1)$	$s_{AB}^2 = \frac{СК_{AB}}{(k-1)(n-1)}$
Остаток (ошибка)	$СК_0$	$kn(m-1)$	$s_0^2 = \frac{СК_0}{kn(m-1)}$
Полная	$СК_п = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{t=1}^{m}(x_{ijt} - \bar{x}...)^2$	$knm-1$	

Двухфакторный дисперсионный анализ с группировкой уровней

Фактор *B* группируется фактором *A*, если каждый уровень фактора *B* сочетается не более, чем с одним уровнем фактора *A*.

A_1		A_2		A_3	
B_1	B_2	B_3	B_4	B_5	B_6
x_{111}	x_{121}	x_{231}	x_{241}	x_{351}	x_{361}
…	…	…	…	…	…
x_{11t}	x_{12t}	x_{23t}	x_{24t}	x_{35t}	x_{36t}
…	…	…	…	…	…
x_{11m}	x_{12m}	x_{23m}	x_{24m}	x_{35m}	x_{36m}

Двухфакторный дисперсионный анализ с группировкой уровней

Результаты дисперсионного анализа оформляются в виде следующей таблицы

Источник дисперсии	Сумма квадратов	Степени свободы	Дисперсии
Фактор A	$\text{СК}_A = mn\sum_{i=1}^{k}(\bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot\cdot\cdot})^2$	$k-1$	$s_A^2 = \frac{\text{СК}_A}{k-1}$
Фактор B (внутри A)	$\text{СК}_{B(A)} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot})^2$	$k(n-1)$	$s_{B(A)}^2 = \frac{\text{СК}_B}{k(n-1)}$
Остаток (ошибка)	$\text{СК}_0 = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{t=1}^{m}(x_{ijt} - \bar{x}_{ij\cdot})^2$	$kn(m-1)$	$s_0^2 = \frac{\text{СК}_0}{kn(m-1)}$
Полная	$\text{СК}_\text{п} = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{t=1}^{m}(x_{ijt} - \bar{x}_{\cdot\cdot\cdot})^2$	$knm-1$	

Двухфакторный дисперсионный анализ с группировкой уровней

Статистики для проверки гипотез имеют вид:

для гипотезы H_0: все $\alpha_i = 0$

$$F = \frac{s_A^2}{s_0^2}$$

для гипотезы H_0: $\sigma_{b(a)} = 0$

$$F = \frac{s_{B(A)}^2}{s_0^2}$$

Задачи корреляционного анализа

В математическом анализе зависимость между величинами *x* и *y* выражается функцией *y* = *f*(*x*), где каждому значению *x* соответствует одно и только одно значение *y*. Такая связь называется функциональной.

Для случайных величин *X* и *Y* такую зависимость можно установить не всегда. Связь между случайными величинами является не функциональной, а случайной (стохастической), при которой изменение переменной *X* влияет на значения переменной *Y* через изменение закона распределения случайной величины *Y*.

Задачи корреляционного анализа

Таким образом задача корреляционного анализа исследование наличия взаимосвязей между отдельными группами переменных и установление тесноты (силы) связи между ними.

Измерители парной статистической связи. Корреляционное отношение

Очевидно, что $0 \le \rho^2_{yx} \le 1$. Стремление ρ^2_{yx} к нулю означает, что доля дисперсии, обусловленная функциональной связью, очень мала. Наоборот, стремление ρ^2_{yx} к единице показывает, что случайными изменениями Y можно пренебречь и вся дисперсия обусловлена функциональной зависимостью $Y = \phi(X)$.

Аналогично определяется квадрат корреляционного отношения ρ^2_{xy} переменной X по Y. Однако между ρ^2_{yx} и ρ^2_{xy} нет какой-либо простой зависимости.

Измерители парной статистической связи. Корреляционное отношение

Положительный корень из ρ^2_{yx} носит название корреляционного отношения, которое является показателем статистической связи между двумя случайными величинами X и Y для самой общей ситуации, когда закон распределения системы (X,Y) является произвольным.

Измерители парной статистической связи

В общем случае показатели ρ^2_{xy} и r^2 связаны неравенствами $0 \le r^2 \le \rho^2_{xy} \le 1$

При этом возможны следующие варианты:

- $r^2 = \rho^2_{yx} = 1$ **только тогда, когда имеется строгая линейная функциональная зависимость** ***Y*** **от** ***X***
- $r^2 < \rho^2_{yx} = 1$ **только тогда, когда имеется строгая нелинейная функциональная зависимость** ***Y*** **от** ***X***
- $r^2 = \rho^2_{yx} < 1$ **только тогда, когда зависимость** ***Y*** **от** ***X*** **строго линейна, но нет функциональной зависимости**
- $r^2 < \rho^2_{yx} < 1$ **указывает на то, что не существует функциональной зависимости, а некоторая нелинейная кривая "подходит" лучше, чем "наилучшая" прямая линия.**

Измерители парной статистической связи

Таким образом, в качестве показателя статистической связи между двумя случайными количественными переменными X и Y следует выбрать корреляционное отношение ρ_{yx} (или ρ_{xy}) , если закон распределения системы (X,Y) вызывает сомнение. Если же можно с большой степенью уверенности считать закон распределения системы (X,Y) нормальным, то вместо корреляционного отношения следует использовать коэффициент корреляции r.

Регрессионный анализ

Основные понятия регрессионного анализа

Для математического описания статистических связей между изучаемыми переменными величинами следует решить следующие задачи:

- **подобрать класс функций, в котором целесообразно искать наилучшую (в определенном смысле) аппроксимацию интересующей зависимости;**
- **найти оценки неизвестных значений параметров, входящих в уравнения искомой зависимости;**
- **установить адекватность полученного уравнения искомой зависимости;**
- **выявить наиболее информативные входные переменные.**

Простая линейная регрессия

Простейшей моделью регрессии является простая (одномерная, однофакторная, парная) линейная модель, имеющая следующий вид:

$$y_i = a + bx_i + \varepsilon_i \qquad i = 1,\ldots,n$$

где ε_i – некоррелированные между собой случайные величины (ошибки), имеющие нулевые математические ожидания и одинаковые дисперсии σ^2, a и b – постоянные коэффициенты (параметры), которые необходимо оценить по измеренным значениям отклика y_i.

Простая линейная регрессия

Для нахождения оценок параметров *a* и *b* линейной регрессии, определяющих наиболее удовлетворяющую экспериментальным данным прямую линию:

$$f_a(x) = a + bx$$

применяется метод наименьших квадратов.

Согласно методу наименьших квадратов оценки параметров *a* и *b* находят из условия минимизации суммы квадратов отклонений значений y_i по вертикали от "истинной" линии регрессии:

$$D = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

Простая линейная регрессия

Для минимизации *D* приравняем к нулю частные производные по *a* и *b*:

$$\frac{\partial D}{\partial a} = 2\sum_{i=1}^{n}\left(y_i - a - bx_i\right)\left(-1\right) = 0$$

$$\frac{\partial D}{\partial b} = 2\sum_{i=1}^{n}\left(y_i - a - bx_i\right)\left(-x_i\right) = 0$$

В результате получим следующую систему уравнений для нахождения оценок *a* и *b*:

$$\sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right)x_i = 0$$

Простая линейная регрессия

Решение этих двух уравнений дает:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2}$$

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} x_i$$

Простая линейная регрессия

Выражения для оценок параметров *a* и *b* можно представить также в виде:

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

где $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – среднее значение случайной величины Y;

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – среднее значение величин x_i.

Простая линейная регрессия

Тогда эмпирическое уравнение регрессионной прямой *Y* на *X* можно записать в виде:

$$\hat{y} = \hat{a} + \hat{b}x = \bar{y} - \hat{b}\left(x - \bar{x}\right)$$

Заметим, что если $x = \bar{x}$, то $\hat{y} = \bar{y}$, т.е. точка лежит на подобранной регрессионной прямой.

Простая линейная регрессия

Несмещенная оценка дисперсии σ^2 отклонений значений y_i от подобранной прямой линии регрессии дается выражением (остаточная дисперсия)

$$s_0^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right)^2$$

Проверка значимости линии регрессии

Найденная оценка $b \neq 0$ может быть реализацией случайной величины, математическое ожидание которой равно нулю, т. е. может оказаться, что никакой регрессионной зависимости на самом деле нет.

Чтобы разобраться с этой ситуацией, следует проверить гипотезу H_0: $b = 0$ при конкурирующей гипотезе H_1: $b \neq 0$.

Проверку значимости линии регрессии можно провести с помощью дисперсионного анализа.

Проверка значимости линии регрессии

Вычисления по проверки значимости регрессии проводят в следующей таблице дисперсионного анализа

Источник дисперсии	Суммы квадратов	Степени свободы	Дисперсии (средние квадраты)	F-отношение
регрессия	$\text{СК}_\text{р} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$s_\text{р}^2 = \frac{\text{СК}_\text{р}}{1}$	$F = \frac{s_\text{р}^2}{s_0^2}$
остаточная	$\text{СК}_0 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$s_0^2 = \frac{\text{СК}_0}{n-2}$	
полная (общая)	$\text{СК}_\text{п} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

Проверка адекватности линейной модели регрессии

Под адекватностью построенной регрессионной модели понимается то, что никакая другая модель не дает значимого улучшения в предсказании отклика.

Если все значения откликов получены при разных значениях x, т. е. нет нескольких значений отклика, полученных при одинаковых x_i, то можно провести лишь ограниченную проверку адекватности линейной модели. Основой для такой проверки являются остатки:

$d_i = y_i - \hat{y}_i$ **- отклонения от установленной закономерности:** $\hat{y}_i = \hat{a} + \hat{b}x_i$

Коэффициент детерминации

Иногда для характеристики качества линии регрессии используют выборочный коэффициент детерминации R^2, показывающий, какую часть (долю) сумма квадратов, обусловленная регрессией $СК_р$, составляет в полной сумме квадратов $СК_п$:

$$R^2 = \frac{СК_Р}{СК_П} = 1 - \frac{СК_0}{СК_П}$$

Чем ближе R^2 к единице, тем лучше регрессия аппроксимирует экспериментальные данные, тем теснее наблюдения примыкают к линии регрессии. Если $R^2 = 0$, то изменения отклика полностью обусловлены воздействием неучтенных факторов, и линия регрессии параллельна оси x-ов. В случае простой линейной регрессии коэффициент детерминации R^2 равен квадрату коэффициента корреляции r^2 .

Коэффициент детерминации

Максимальное значение $R^2 = 1$ может быть достигнуто только в случае, когда наблюдения проводились при различных значениях *x*-ов. Если же в данных имеются повторяющиеся опыты, то величина R^2 не может достичь единицы, как бы ни была хороша модель.

Вместо коэффициента детерминации R^2 можно использовать статистику - нормированная (приведенная) R^2– статистика. Она имеет следующий вид:

$$\hat{R}^2_{\text{н}} = 1 - \frac{\text{СК}_0/(n-p)}{\text{СК}_\text{р}/(n-1)},$$

где p – число параметров линейной модели регрессии.

Коэффициент детерминации

Применительно к простой линейной регрессии

$$\hat{R}^2_{\text{н}} = 1 - \left(1 - \hat{R}^2\right)\frac{n-1}{n-2}$$

Отметим, что коэффициент R^2 имеет смысл рассматривать только при наличии в уравнении регрессии свободного члена a, так как лишь в этом случае верно равенство

$$СК_{п} = СК_{р} + СК_{0}$$

Сравнение двух линий регрессии

Часто требуется сравнить линии регрессии, рассчитанные по двум выборкам. Это можно сделать тремя способами:

- **Сравнить коэффициенты наклона *b***
- **Сравнить коэффициенты сдвига *a***
- **Сравнить линии в целом**

Сравнение двух линий регрессии

Если нужно проверить, значимо ли различие в наклоне двух прямых регрессии, критерий Стьюдента *t* вычисляется по формуле:

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}}$$

где b_1–b_2 — разность коэффициентов наклона, а s_{b1-b2} — ее стандартная ошибка.

Затем вычисленное значение *t* сравнивают, с критическим значением, имеющим n_1+n_2-4 степени свободы.

Сравнение двух линий регрессии

Если обе регрессии оценены по одинаковому числу наблюдений, то стандартная ошибка разности

$$s_{b_1-b_2} = \sqrt{s_{b_1}^2 + s_{b_2}^2}$$

Если же объемы выборок различны, следует воспользоваться объединенной оценкой остаточной дисперсии

$$s_{0_{общ}}^2 = \frac{(n_1-2)s_{0_1}^2 + (n_2-2)s_{0_2}^2}{n_1+n_2-4}$$

Тогда стандартная ошибка разности

$$s_{b_1-b_2} = \sqrt{\frac{s_{0_{общ}}^2}{(n_1-2)s_{x_1}^2} + \frac{s_{0_{общ}}^2}{(n_2-2)s_{x_2}^2}}$$

Сравнение двух линий регрессии

Аналогично сравниваются и коэффициенты сдвига a_1 и a_2. В этом случае

$$t = \frac{a_1 - a_2}{s_{a_1 - a_2}}$$

где a_1–a_2 — разность коэффициентов сдвига, а $s_{a1–a2}$ — стандартная ошибка разности коэффициентов сдвига

Затем вычисленное значение t сравнивают, с критическим значением, имеющим $n_1+n_2–4$ степени свободы.

Сравнение двух линий регрессии

Таким образом алгоритм сравнения двух линии регрессии следующий:

- **Построить прямую регрессии для каждой из выборок.**
- **По остаточным дисперсиям $s^2_{0_1}$ и $s^2_{0_2}$ каждой из регрессий вычислить объединенную оценку остаточной дисперсии $s^2_{0_{общ}}$**
- **Объединить обе выборки. Построить прямую регрессии для получившейся выборки и вычислить остаточную дисперсию s^2_0**

Множественная линейная регрессия

Модель множественной линейной регрессии имеет следующий вид:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + \varepsilon_i$$

Предположения относительно множественной линейной регрессии аналогичны тем, которые применялись для простой линейной регрессии. В частности, что все x_i считаются фиксированными и для любого набора x_i значения y_i распределены по нормальному закону с постоянной дисперсией.

Множественная линейная регрессия

Для получения оценок параметров b_0, b_1, ...,b_k методом наименьших квадратов нужно минимизировать по этим параметрам выражение

$$D = \sum_{i=1}^{n} \left(y_i - b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} \right)^2$$

Множественная линейная регрессия

Приравняв нулю частные производные

$$\frac{\partial D}{\partial b_0}, \frac{\partial D}{\partial b_1}, \frac{\partial D}{\partial b_2}, \ldots, \frac{\partial D}{\partial b_k}$$

после упрощений получается следующая система нормальных уравнений для нахождения оценок параметров:

$$n\hat{b}_0 + \hat{b}_1 \sum_{i=1}^{n} x_{1i} + \hat{b}_2 \sum_{i=1}^{n} x_{2i} + \ldots + \hat{b}_k \sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i ,$$

$$\hat{b}_0 \sum_{i=1}^{n} x_{1i} + \hat{b}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{b}_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \ldots + \hat{b}_k \sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i ,$$

$$\hat{b}_0 \sum_{i=1}^{n} x_{2i} + \hat{b}_1 \sum_{i=1}^{n} x_{1i}x_{2i} + \hat{b}_2 \sum_{i=1}^{n} x_{2i}^2 + \ldots + \hat{b}_k \sum_{i=1}^{n} x_{2i}x_{ki} = \sum_{i=1}^{n} x_{2i}y_i ,$$

$$\hat{b}_0 \sum_{i=1}^{n} x_{ki} + \hat{b}_1 \sum_{i=1}^{n} x_{1i}x_{ki} + \hat{b}_2 \sum_{i=1}^{n} x_{2i}x_{ki} + \ldots + \hat{b}_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} x_{ki}y_i .$$

Множественная линейная регрессия

Пусть b – вектор-столбец размера $(k+1)$, состоящий из коэффициентов b_0 , b_1, …, b_k , y – вектор-столбец из n наблюдений, ε – вектор-столбец из n ошибок и X – матрица наблюдений размером $n(k+1)$:

$$\mathbf{b} = \begin{vmatrix} b_0 \\ b_1 \\ \dots \\ b_k \end{vmatrix} \qquad \mathbf{y} = \begin{vmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{vmatrix} \qquad \boldsymbol{\varepsilon} = \begin{vmatrix} \varepsilon_1 \\ \varepsilon_1 \\ \dots \\ \varepsilon_n \end{vmatrix} \qquad \mathbf{X} = \begin{vmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{kn} \end{vmatrix}$$

Множественная линейная регрессия

Тогда уравнение модели регрессии можно записать в виде:

$$\mathbf{y} = \mathbf{X}^T\mathbf{b} + \varepsilon$$

Выражение для *D* можно представить в матричном виде:

$$\mathbf{D} = \left(\mathbf{y} - \mathbf{Xb}\right)^T \left(\mathbf{y} - \mathbf{Xb}\right)$$

тогда вектор оценок *b* получается из решения системы уравнений:

$$\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{b} = \mathbf{X}^T\mathbf{y}$$

решение которой имеет вид:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Множественная линейная регрессия

Несмещенной оценкой дисперсии является:

$$s_0^2 = \frac{1}{n-k-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^T \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)$$

Дисперсионный анализ множественной линейной регрессии проводится в следующей таблице:

Источник дисперсии	Суммы квадратов	Степени свободы	Дисперсии (средние квадраты)	F-отношение
регрессия	$СК_р = b^T X^T y - n\bar{y}^2$	k	$s_р^2 = \frac{СК_р}{k}$	$F = \frac{s_р^2}{s_0^2}$
остаточная	$СК_0 = СК_п - СК_р$	$n-k-1$	$s_0^2 = \frac{СК_0}{n-k-1}$	
полная (общая)	$СК_п = y^T y - n\bar{y}^2$	$n-1$		