Многомерный регрессионный анализ

Постановка задачи регрессионного анализа

Ставится задача на основе выборочных данных, представленных в виде вектора $Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$, где y_i – наблюдаемое значение результативного признака Y для i-го объекта и матрицы $X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$, где x_{ij} – наблюдаемое значение j-го объясняющего признака для i-го объекта выборочной совокупности, выявить «зависимость» результативного показателя Y от факторных (объясняющих) признаков $X_1, X_2, \dots, X_k$.

Постановка задачи регрессионного анализа

Ставится задача на основе выборочных данных, представленных в виде вектора $Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$, где y_i – наблюдаемое значение результативного признака Y для i-го объекта и матрицы $X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$, где x_{ij} – наблюдаемое значение j-го объясняющего признака для i-го объекта выборочной совокупности, выявить «зависимость» результативного показателя Y от факторных (объясняющих) признаков $X_1, X_2, ..., X_k$.

Функция $f_Y(X_1, X_2, ..., X_k)$, описывающая зависимость условного среднего значения результативного признака Y от заданных значений факторных признаков $X_1, X_2, ..., X_k$, называется **функцией (уравнением) регрессии**, т.е. $f_Y(X_1, X_2, ..., X_k) = MY / X_1, X_2, ..., X_k$.

Уравнение линейной множественной регрессии

Если зависимость результативного признака (y) и объясняющих переменных ($x_1, x_2, \ldots, x_k$) линейного характера, то линейная функция множественной регрессии имеет вид:

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

или в векторной форме:

$$\tilde{y} = x^T \beta,$$

где $x = (1, x_1, \ldots, x_k)^T$, $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$.

Линейная модель множественной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

y_i — значения зависимой переменной на *i*-ом объекте,

x_{ik} — значения независимых переменных (регрессоров),

k — число коэффициентов (переменных) в модели,

ε_i — случайные ошибки

Линейная модель множественной регрессии

$$f_Y(X_1, X_2, ..., X_k) \approx \tilde{Y} = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k,$$

где $\beta_0, \beta_1, ..., \beta_k$ – коэффициенты линейного уравнения регрессии.

Система линейных уравнений $y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + ... + \beta_k x_{i_k} + \varepsilon_i$, где ε_i – регрессионный остаток, характеризующий расхождение между наблюдаемым значением y_i и «осредненным» значением $\tilde{y}_i$, $i = \overline{1, n}$;

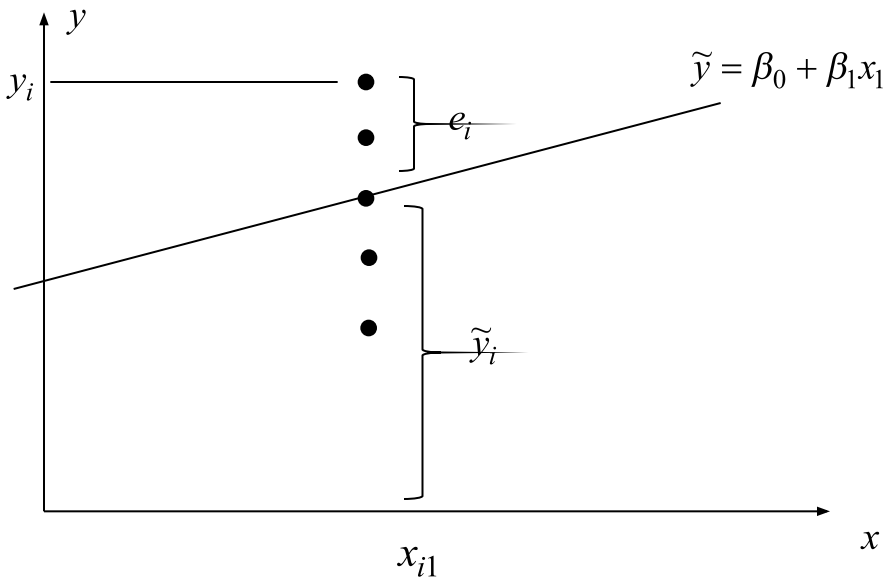
или

$$Y = X\beta + \varepsilon,\ \beta = (\beta_0, ..., \beta_k),\ \varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T,$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \square & x_{1k} \\ 1 & x_{21} & x_{22} & \square & x_{2k} \\ \square & \square & \square & \square & \square \\ 1 & x_{n1} & x_{n2} & \square & x_{nk} \end{pmatrix}$$

называется **линейной моделью множественной регрессии**.

Геометрическая интерпретация функции регрессии



Классическая линейная модель множественной регрессии (КЛММР)

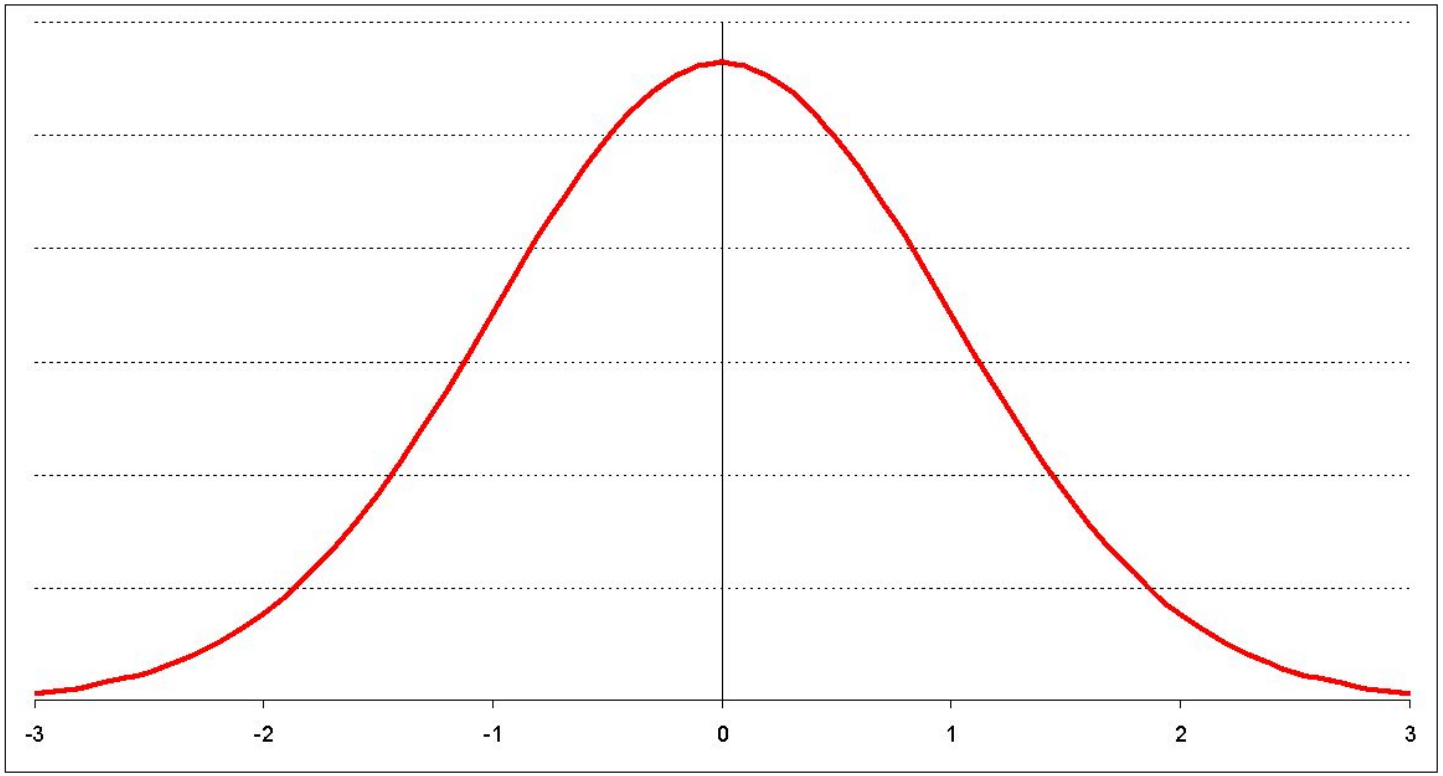
Линейная модель множественной регрессии, удовлетворяющая следующим пяти требованиям, называется **классической линейной моделью множественной регрессии**.

Условия Гаусса–Маркова

1) $X_1, X_2, \ldots, X_k$ – детерминированные переменные;
2) ранг матрицы X равен $k+1$ – среди признаков нет линейно зависимых;
3) $M\varepsilon_i = 0,\ i = \overline{1,n}$ – нет систематических ошибок в измерении Y;
4) $D\varepsilon_i = M\varepsilon_i^2 = \sigma^2,\ i = \overline{1,n}$ – гомоскедастичность регрессионных остатков (равноточные измерения);
5) $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i \cdot \varepsilon_j) = 0$ – условие некоррелированности регрессионных остатков, $i \neq j,\ i = \overline{1,n},\ j = \overline{1,n}$.

Классическая линейная модель множественной регрессии (КЛММР)

(6)* Случайные ошибки ε_i имеют нормальное распределение



Методы оценки коэффициентов КЛММР

Оценку коэффициентов уравнения регрессии можно искать:

- исходя из требований минимума модуля отклонения наблюдаемых значений y_i от "значений" функции регрессии
- исходя из критерия минимума суммы квадратов отклонений наблюдаемых значений y_i от "значений" функции регрессии (метод наименьших квадратов)

Методы оценки коэффициентов КЛММР

Идея метода наименьших квадратов: подбор параметров таким образом чтобы сумма квадратов отклонений

$$e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

была наименьшей

Метод наименьших квадратов

$$F = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{in})^2 =$$

Метод наименьших квадратов

$$F = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{in})^2 =$$

$$= (Y - Xb)^T (Y - Xb) =$$

Метод наименьших квадратов

$$F = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{in})^2 =$$

$$= (Y - Xb)^T (Y - Xb) =$$

$$= Y^T Y - b^T X^T Y - Y^T Xb + b^T X^T Xb =$$

$$= Y^T Y - 2b^T X^T Y + b^T X^T Xb \to \min .$$

Метод наименьших квадратов

$$F = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{in})^2 =$$

$$= (Y - Xb)^T (Y - Xb) =$$

$$= Y^T Y - b^T X^T Y - Y^T Xb + b^T X^T Xb =$$

$$= Y^T Y - 2b^T X^T Y + b^T X^T Xb \to \min .$$

$$2X^T Xb - 2X^T Y = 0$$

Метод наименьших квадратов

$$F = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{in})^2 =$$

$$= (Y - Xb)^T (Y - Xb) =$$

$$= Y^T Y - b^T X^T Y - Y^T Xb + b^T X^T Xb =$$

$$= Y^T Y - 2b^T X^T Y + b^T X^T Xb \to \min .$$

$$2X^T Xb - 2X^T Y = 0$$

$$b_{\text{МНК}} \equiv b = (X^T X)^{-1} X^T Y$$

Метод наименьших квадратов

$$F = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{in})^2 = (Y - X\bar{b})^T (Y - X\bar{b}) =$$

$$= Y^T Y - \bar{b}^T X^T Y - Y^T X\bar{b} + \bar{b}^T X^T X\bar{b} = Y^T Y - 2\bar{b}^T X^T Y + \bar{b}^T X^T X\bar{b} \to \min$$

$$2X^T X\bar{b} - 2X^T Y = 0$$

В силу предположения о справедливости условий Гаусса-Маркова, в частности ($X=k+1$), матрица $X^T X$ – не вырождена и получим МНК - оценки для вектора $\bar{\beta}$:

$$\bar{b}_{\text{МНК}} \equiv \bar{b} = (X^T X)^{-1} X^T Y$$

оценка уравнения регрессии:

$$\hat{y} = \underset{(S_{b0})}{b_0} + \underset{(s_{b1})}{b_1}\ x_1 + \underset{(S_{b2})}{b_2}\ x_2 + \ldots + \underset{(S_{bk})}{b_k}\ x_k$$

Теорема Гаусса — Маркова

Если выполнены условия (1)–(5),
то оценка коэффициентов *модели*,
полученная по методу наименьших квадратов, является:

(а) несмещенной

(б) эффективной

Статистические свойства оценок коэффициентов, полученных МНК

МНК – оценка b является несмещенной оценкой вектора β.

$$b = (X^T X)^{-1} X^T Y =$$

Статистические свойства оценок коэффициентов, полученных МНК

МНК – оценка b является несмещенной оценкой вектора β.

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) =$$

Статистические свойства оценок коэффициентов, полученных МНК

МНК – оценка b является несмещенной оценкой вектора β.

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) = \beta + (X^T X)^{-1}$$

Статистические свойства оценок коэффициентов, полученных МНК

МНК – оценка b является несмещенной оценкой вектора β.

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) = \beta + (X^T X)^{-1} X^T \varepsilon$$

$$M(b) = M(\beta + (X^T X)^{-1} X^T \varepsilon) = \beta + (X^T X)^{-1} X^T M\varepsilon = \beta$$

Статистические свойства оценок коэффициентов, полученных МНК

МНК – оценка b является несмещенной оценкой вектора β.

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) = \beta + (X^T X)^{-1} X^T \varepsilon$$

$$M(b) = M(\beta + (X^T X)^{-1} X^T \varepsilon) = \beta + (X^T X)^{-1} X^T M\varepsilon = \beta$$

МНК - оценки коэффициентов КЛММР являются состоятельными

МНК - оценки коэффициентов КЛММР являются эффективными в классе линейных оценок относительно компонент Y (т.е. имеют наименьшую дисперсию).

Статистические свойства оценок коэффициентов, полученных МНК

$$\bar{b}_{\text{МНК}} \equiv \bar{b} = (X^T X)^{-1} X^T Y$$

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\bar{\beta} + \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon}\,.$$

1) МНК – оценка $\bar{b}$ является несмещенной оценкой вектора $\bar{\beta}$

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\bar{\beta} + \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T \varepsilon$$

$$M\bar{b} = M(\bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T M\varepsilon = \bar{\beta}$$

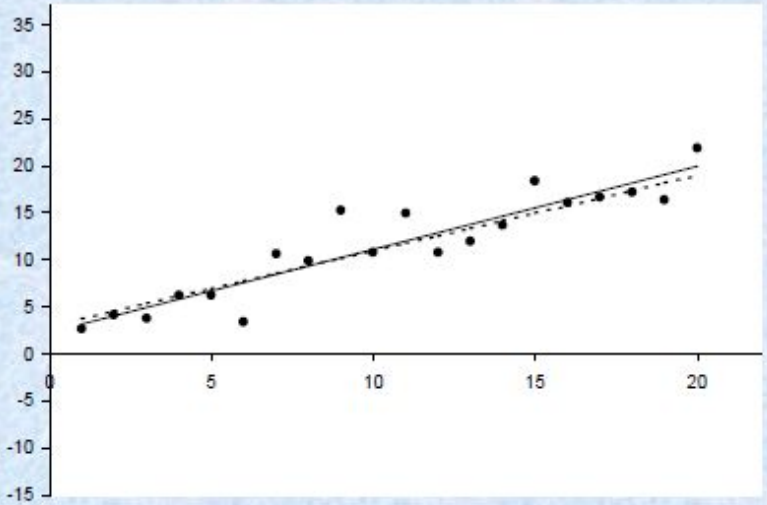
2) Свойство состоятельности в данном случае определяется структурой матрицы X: наименьшее собственное число матрицы $X^T X$ стремится к ∞, при $n \to \infty$.

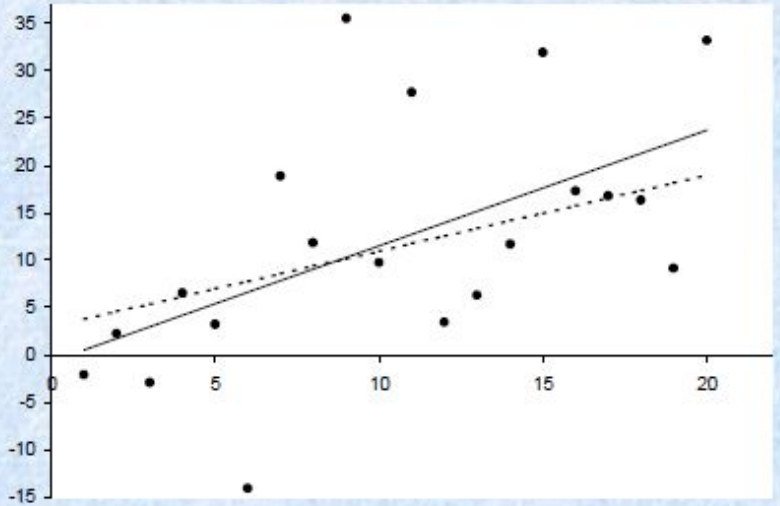
3) Оценки считаются эффективными, если они характеризуются наименьшей дисперсией. Эффективность МНК-оценок доказывается в предположении о нормальности регрессионных остатков ММП.

СВОЙСТВА ОЦЕНОК ПАРАМЕТРОВ

влияние параметров выборки на дисперсию оценок (на примере парной регрессии)

Чем больше дисперсия оценок- тем менее точны оценки коэффициентов

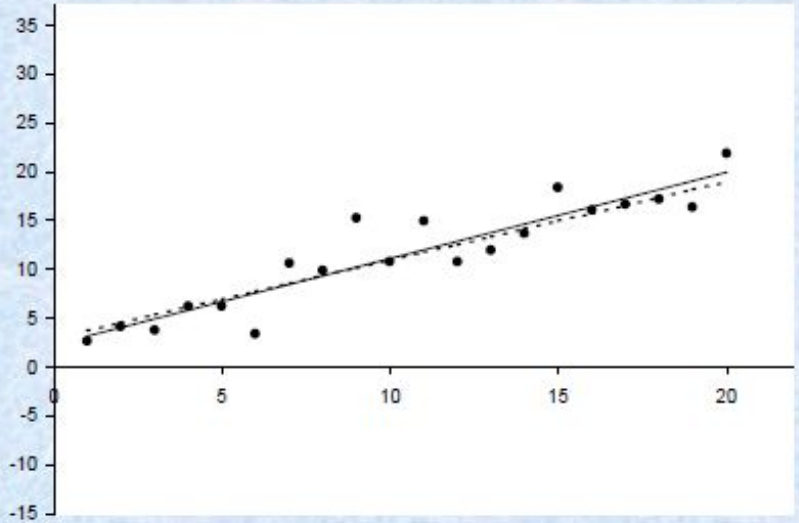


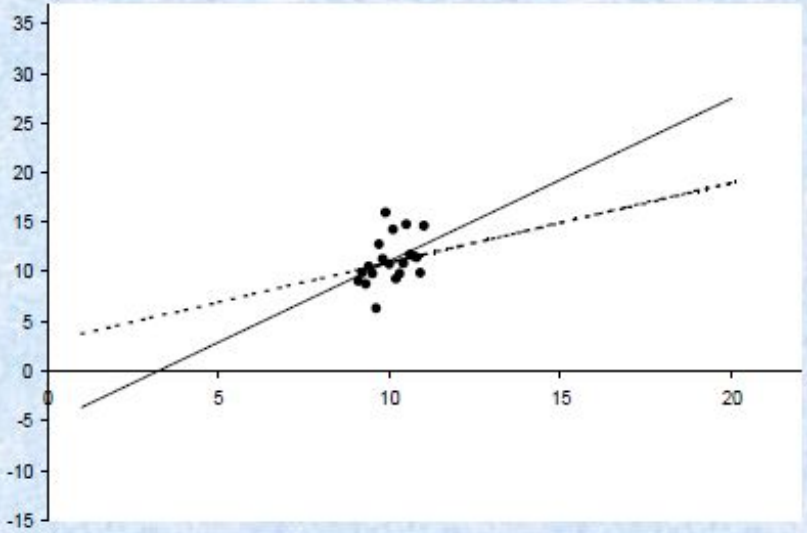


СВОЙСТВА ОЦЕНОК ПАРАМЕТРОВ

влияние параметров выборки на дисперсию оценок (на примере парной регрессии)

Оценки тем точнее, чем разнообразнее выборка по значениям регрессоров





Ковариационная матрица вектора оценок коэффициентов

Найдем ковариационную матрицу случайного вектора b, воспользовавшись условиями Гаусса-Маркова

$$\Sigma_{\bar{b}} = M[(b - Mb)(b - Mb)^T] =$$

Ковариационная матрица вектора оценок коэффициентов

Найдем ковариационную матрицу случайного вектора b, воспользовавшись условиями Гаусса-Маркова

$$\Sigma_{\overline{b}} = M[(b - Mb)(b - Mb)^T] = M[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T)] =$$

Ковариационная матрица вектора оценок коэффициентов

Найдем ковариационную матрицу случайного вектора b, воспользовавшись условиями Гаусса-Маркова

$$\Sigma_{\bar{b}} = M[(b - Mb)(b - Mb)^T] = M[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T)] =$$

$$= M[(X^T X)^{-1} X^T \varepsilon\varepsilon^T (X^T X)^{-1}] =$$

Ковариационная матрица вектора оценок коэффициентов

Найдем ковариационную матрицу случайного вектора b, воспользовавшись условиями Гаусса-Маркова

$$\Sigma_{\bar{b}} = M[(b - Mb)(b - Mb)^T] = M[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T)] =$$

$$= M[(X^T X)^{-1} X^T \varepsilon\varepsilon^T (X^T X)^{-1}] = (X^T X)^{-1} X^T M(\varepsilon\varepsilon^T) X (X^T X)^{-1} =$$

Ковариационная матрица вектора оценок коэффициентов

Найдем ковариационную матрицу случайного вектора b, воспользовавшись условиями Гаусса-Маркова

$$\Sigma_{\bar{b}} = M[(b - Mb)(b - Mb)^T] = M[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T)] =$$

$$= M[(X^T X)^{-1} X^T \varepsilon\varepsilon^T (X^T X)^{-1}] = (X^T X)^{-1} X^T M(\varepsilon\varepsilon^T) X (X^T X)^{-1} =$$

$$= (X^T X)^{-1} X^T \sigma^2 \varepsilon_n X (X^T X)^{-1} =$$

Ковариационная матрица вектора оценок коэффициентов

Найдем ковариационную матрицу случайного вектора b, воспользовавшись условиями Гаусса-Маркова

$$\Sigma_{\bar{b}} = M[(b - Mb)(b - Mb)^T] = M[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T)] =$$

$$= M[(X^T X)^{-1} X^T \varepsilon\varepsilon^T (X^T X)^{-1}] = (X^T X)^{-1} X^T M(\varepsilon\varepsilon^T) X (X^T X)^{-1} =$$

$$= (X^T X)^{-1} X^T \sigma^2 \varepsilon_n X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Ковариационная матрица вектора оценок коэффициентов

Найдем ковариационную матрицу случайного вектора b, воспользовавшись условиями Гаусса-Маркова

$$\Sigma_{\bar{b}} = M[(b - Mb)(b - Mb)^T] = M[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T)] =$$

$$= M[(X^T X)^{-1} X^T \varepsilon\varepsilon^T (X^T X)^{-1}] = (X^T X)^{-1} X^T M(\varepsilon\varepsilon^T) X (X^T X)^{-1} =$$

$$= (X^T X)^{-1} X^T \sigma^2 \varepsilon_n X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Откуда, в частности,

$$D_{bj} = \sigma^2 [(X^T X)^{-1}]_{jj}, \quad j = \overline{0,k}$$

Ковариационная матрица вектора оценок коэффициентов

Найдем ковариационную матрицу случайного вектора b, воспользовавшись условиями Гаусса-Маркова

$$\Sigma_{\bar{b}} = M[(b - Mb)(b - Mb)^T] = M[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T)] =$$
$$= M[(X^T X)^{-1} X^T \varepsilon\varepsilon^T (X^T X)^{-1}] = (X^T X)^{-1} X^T M(\varepsilon\varepsilon^T) X (X^T X)^{-1} =$$
$$= (X^T X)^{-1} X^T \sigma^2 \varepsilon_n X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Откуда, в частности,

$$D_{bj} = \sigma^2 [(X^T X)^{-1}]_{jj}, \quad j = \overline{0,k}$$

Несмещенная оценка для σ^2 определяется по формуле:

$$\hat{S}^2 = \frac{1}{n-k-1}(Y - Xb)^T (Y - Xb)$$

$$\hat{\Sigma}_b = \hat{S}^2 (X^T X)^{-1}$$

Качество подгонки модели

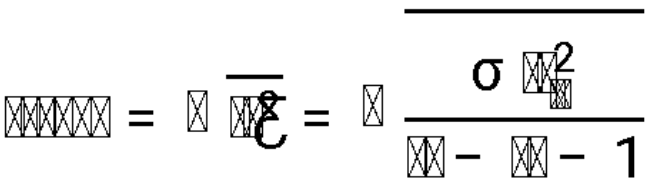
1. Стандартная ошибка регрессии
2. Коэффициент детерминации R^2
3. Скорректированный (нормированный) коэффициент детерминации R^2

Стандартная ошибка регрессии

Несмещенная оценка для σ^2 определяется по формуле:

$$\hat{S}^2 = \frac{1}{n-k-1}(Y - Xb)^T (Y - Xb)$$

$\hat{S}^2$ используется для расчета стандартных ошибок коэффициентов и стандартной ошибки регрессии
(Standard Error of Estimate — не путать с ESS)



измеряет среднюю величину ошибки модели

Используется для оценки качества подгонки модели: чем меньше SEE, тем точнее модель

Можно использовать для сравнения нескольких однотипных уравнений регрессии

Анализ вариации результативного признака Коэффициент детерминации R^2

В качестве характеристики степени рассеивания случайной величины Y относительно функции регрессии используется в случае нелинейной связи корреляционное отношение:

$$\rho^2_{Y/X_1,\ldots,X_k} = 1 - \frac{M(Y - f_Y(X_1, X_2, \ldots, X_k))^2}{\sigma_Y^2} = \frac{M(f_Y(X_1, X_2, \ldots, X_k) - MY)^2}{\sigma_Y^2},$$

которое характеризует качество подгонки функции регрессии под выборочные данные. В случае линейной регрессии $\rho^2_{Y/X_1,\ldots,X_k}$ называется коэффициентом детерминации $R^2_{Y/X_1,\ldots,X_k} \equiv R^2$.

Анализ вариации результативного признака Коэффициент детерминации R^2

Коэффициент детерминации строится из тех соображений, что общая дисперсия результативного признака складывается из факторной и остаточной дисперсий:

$$\sigma_Y^2 = \sigma_{факт}^2 + \sigma_{ост}^2,$$

где σ_Y^2 – дисперсия результативного признака;

$\sigma_{факт}^2 = M(f_Y(X_1,\ldots,X_k) - MY)^2$ – факторная дисперсия;

$\sigma_{ост}^2 = M(Y - f_Y(X_1,\ldots,X_k))^2$ – остаточная дисперсия.

Анализ вариации результативного признака Коэффициент детерминации R^2

Можно показать, что общая вариация (дисперсия) результативного признака складывается из вариации функции регрессии, обусловленной варьированием значений объясняющих переменных $x_1,..,x_k$, (факторной дисперсии) и из вариации случайной величины относительно функции регрессии (остаточной дисперсии):

$$Q_{общ} = (Y - \bar{Y})^T (Y - \bar{Y}) = \sum_{i=1}^{n} (y_i - \bar{y})^2 = Q_{ост} + Q_{факт}$$

Определим выборочную вариацию результативного признака y.

$$Вар(y) = Q_{общ} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (Y - \bar{Y})^T (Y - \bar{Y}),$$

где $\bar{y} = \sum_{i=1}^{n} y_i / n$ - выборочное среднее;

$Y = (y_1,..., y_n)^T$;

$\bar{Y}_{n*1} = (\bar{y}, \bar{y},..., \bar{y})^T$.

Анализ вариации результативного признака Коэффициент детерминации R^2

Значение оцененной функции регрессии в точке $X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_k x_{ik},\ i = 1,2,\ldots,n$$

Выражение $Q_{общ} = Q_{ост} + Q_{факт}$, разделим на $Q_{общ}$,

тогда $$1 = \frac{Q_{ост}}{Q_{общ}} + \frac{Q_{факт}}{Q_{общ}},$$

где $$Q_{ост} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \equiv \sum_{i=1}^{n} e_i^2, \qquad e_i = y_i - \hat{y}_i;$$

$$Q_{факт} = Bap(\hat{y}) = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

Анализ вариации результативного признака
Выборочный коэффициент детерминации

Возьмем в качестве выборочной оценки коэффициента детерминации:

$$\hat{R}^2_{y/x_1,\dots,x_n} \equiv \frac{Q_{факт}}{Q_{общ}} \equiv \frac{(\hat{Y}-\bar{Y})^T(\hat{Y}-\bar{Y})}{(Y-\bar{Y})^T(Y-\bar{Y})} \equiv 1-\frac{Q_{ост}}{Q_{общ}} \equiv \frac{(Y-\hat{Y})^T(Y-\hat{Y})}{(Y-\bar{Y})^T(Y-\bar{Y})} \equiv$$

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

Подправленная на несмещенность оценка $\hat{R}^{*2}_{y/x_1,\dots,x_n}$ коэффициента детерминации $R^2_{y/x_1,\dots,x_n}$ имеет вид:

$$\hat{R}^{*2}_{y/x_1,\dots,x_n} \approx 1-(1-\hat{R}^2_{y/x_1,\dots,x_n})\frac{n-1}{n-k-1}.$$

О чем говорит и о чем не говорит R^2

- Выборочный коэффициент детерминации находится в промежутке [0;1]
- Характеризует долю общей вариации результативного признака *y*, объясняемую вариацией выборочной функции регрессии .
- Чем меньше разброс статистических данных относительно уравнения регрессии, тем меньше остаточная дисперсия, и тем ближе R^2 к единице.

О чем говорит и о чем не говорит R^2

- Высокий R^2 сам по себе не гарантирует, что модель является хорошей.
  Остается риск *ложной регрессии*
- Низкий R^2 говорит о том, что существуют важные факторы, которые мы не учли в нашей модели

О чем говорит и о чем не говорит R^2

Высокий R^2 сам по себе не гарантирует, что модель является хорошей.
Остается риск *ложной регрессии*

Низкий R^2 говорит о том, что существуют важные факторы, которые мы не учли в нашей модели

Использование R^2 и R^2_c

Есть соблазн свести выбор уравнения к задаче максимизации R^2 или R^2_c. Не стоит этого делать.

1. Высокий R^2 (или R^2_c) **говорит** о том, что регрессоры предсказывают большую долю изменений ***y***.
2. Высокий R^2 (или R^2_c) **не говорит** о том, что вы верно выявили причинно-следственную связь между переменными
3. Высокий R^2 (или R^2_c) **не гарантирует** отсутствия смещения оценок из-за некорректной спецификации

45

Тестирование гипотез

1. Тестирование значимости уравнения
2. Тестирование значимости коэффициента
3. Доверительный интервал для коэффициента
4. Построение доверительного интервала для среднего значения предсказания и для конкретного предсказанного значения

46

Проверка гипотезы об адекватности линейной модели выборочным данным

Исследование свойств оценок классической линейной модели множественной регрессии проводится при дополнительном предположении и нормальном характере распределения регрессионных остатков:

$$\varepsilon_i \in N(0,\sigma^2), \quad i = \overline{1...n}.$$

H_0: $\beta_1=\beta_2=...=\beta_k=0$ (линейная модель множественной регрессии неадекватна выборочным данным);

H_1: $\exists j \in [\overline{1,n}]: \quad \beta_j \neq 0$ (линейная модель множественной регрессии адекватна выборочным данным).

Для проверки гипотезы H_0 используется статистика:

$$F = \frac{Q_{факт}/k}{Q_{ост}/(n-k-1)} = \frac{\hat{R}^2_{Y/X_1,...,X_k}/k}{(1-\hat{R}^2_{Y/X_1,...,X_k})/(n-k-1)},$$

которая в случае справедливости H_0 имеет распределение Фишера с числом степеней свободы $\nu_1 = k$ и $\nu_2 = n-k-1$.

Проверка гипотезы об адекватности линейной модели выборочным данным

Рассматриваемая модель

$$y_i = \beta_1 + \beta_2 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

Тестируемая гипотеза

H0: $\beta_2 = \ldots = \beta_k = 0$ «Все переменные $x_1 \ldots x_k$ не оказывают значимого влияния на переменную y»

Альтернативная гипотеза

H1: «Хотя бы одна из переменных $x_1 \ldots x_k$ оказывает значимое влияние на переменную y»

48

Проверка гипотезы об адекватности линейной модели выборочным данным

Алгоритм проведения теста

Шаг 1

Вычисляем расчетное значение F-статистики

$$F = \frac{Q_{факт} / k}{Q_{ост} /(n-k-1)} = \frac{\hat{R}^2_{y/x_1,\ldots,x_n} / k}{(1-\hat{R}^2_{y/x_1,\ldots,x_n}) /(n-k-1)}$$

Шаг 2

Выбираем **уровень значимости**

49

Проверка гипотезы об адекватности линейной модели выборочным данным

Шаг 3

Из таблиц F-распределения Фишера находим критическое значение F-статистики $F_{кр}$

Оно зависит от уровня значимости α и от числа степеней свободы, которые равны

$$\nu_1 = k \; u \; \nu_2 = n - k - 1$$

50

Проверка гипотезы об адекватности линейной модели выборочным данным

Шаг 4

Сравниваем расчетное и критическое значение F-статистик

Если $F_{расч} < F_{кр}$,

то гипотеза H_0 не отклоняется (принимается),

то есть мы делаем вывод о том, что все переменные $x_1 \ldots x_k$ не оказывает значимого влияния на переменную y

В этом случае уравнение называют незначимым.

В противном случае гипотеза H_0 не принимается (отклоняется).

51

Проверка гипотезы о значимости коэффициентов КЛММР

H_0: $\beta_j = 0$ (коэффициент β_j незначимо отличен от нуля);
H_1: $\beta_j \neq 0$ (коэффициент β_j – значимо отличен от нуля).
Для проверки таких гипотез H_0 строятся статистики

$$t = \frac{b_j}{S_{b_j}}, \quad j = 1,2,\ldots,k, \quad S_{b_j} = \hat{S}\sqrt{[(X^T X)^{-1}]_{jj}},$$

которые в случае справедливости H_0, имеют распределение Стьюдента с $\nu = n - k - 1$ степенями свободы.

Для коэффициентов уравнения регрессии значимо отличных от нуля находим доверительные интервалы, используя статистику

$$t = \frac{b_j - \beta_j}{S_{b_j}}$$

имеющую распределение Стьюдента с $\nu = n - k - 1$ степенями свободы.

Проверка гипотезы о значимости коэффициентов КЛММР

Рассматриваемая модель

$$y_i = \beta_1 + \beta_2 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

Тестируемая гипотеза

H0: $\beta_j = 0$ «Переменная x_j не оказывает значимого влияния на переменную y»

Альтернативная гипотеза

H1: $\beta_j \neq 0$ «Переменная x_j оказывает значимое влияние на переменную y»

53

Проверка гипотезы о значимости коэффициентов КЛММР

Шаг 1

Вычисляем расчетное значение t-статистики

$$t_j = \frac{b_j}{S_{b_j}}, \quad j = 1,2,\ldots,k, \quad S_{b_j} = \hat{S}\sqrt{[(X^T X)^{-1}]_{jj}}$$

Шаг 2

Выбираем **уровень значимости**

Уровень значимости — **вероятность ошибки** первого рода, то есть вероятность отклонить гипотезу H0, если на самом деле гипотеза H0 верна.

В эконометрике обычно используется уровень значимости $\alpha = 0,01 = 1\%$ или $\alpha = 0,05 = 5\%$.

54

Проверка гипотезы о значимости коэффициентов КЛММР

Шаг 3

Из таблиц t-распределения Стьюдента находим критическое значение t-статистики $t_{кр}$

Оно зависит от уровня значимости α и так называемого числа степеней свободы, которое в случае нашего теста равно $\nu = n - k - 1$

55

Проверка гипотезы о значимости коэффициентов КЛММР

Шаг 4

Сравниваем расчетное и критическое значение t-статистик

Если $\left|t_{расч}\right| < t_{кр}$,

то гипотеза H0 не отклоняется (принимается),

то есть мы делаем вывод о том, что переменная x_j не оказывает значимого влияния на переменную y. В этом случае коэффициент при переменной x_j называют незначимым.

В противном случае гипотеза H0 не принимается (отклоняется).

56

Доверительные интервалы для коэффициентов

Для коэффициентов уравнения регрессии, значимо отличных от нуля, находим доверительные интервалы, используя статистику

$$tj = \frac{b_j - \beta_j}{S_{b_j}},$$

имеющую распределение Стьюдента с $\nu = n - k - 1$ степенями свободы

Построение доверительного интервала для $\tilde{y}(X_{n+1})$ и $y(X_{n+1})$

$$\hat{y}_{n+1} = X_{n+1}^T \bar{b} \text{ при } X_{n+1} = (1, x_{n+1,1}, \ldots, x_{n+1,k})^T$$

Так как:

$$M\hat{y}_{n+1} = MX_{n+1}^T \bar{b} = X_{n+1}^T \bar{\beta} = \tilde{y}_{n+1}$$

$$D\hat{y}_{n+1} = D(X_{n+1}^T \bar{b}) = X_{n+1}^T \Sigma_b X_{n+1} = \sigma^2 X_{n+1}^T (X^T X)^{-1} X_{n+1},$$

$$S_{\hat{y}_{n+1}}^2 = \hat{S}^2 X_{n+1}^T (X^T X)^{-1} X_{n+1},$$

то для построения доверительного интервала для $\tilde{y}_{n+1}$ построим статистику

$$t = \frac{\hat{y}_{n+1} - \tilde{y}_{n+1}}{\hat{S}_{\hat{y}_{n+1}}},$$

имеющую распределение Стьюдента с $\nu = n - k - 1$ степенями свободы.

Построение доверительного интервала для $\tilde{y}(X_{n+1})$ и $y(X_{n+1})$

При построении доверительного интервала для $y(X_{n+1}) = y_{n+1}$ воспользуемся статистикой

$$t = \frac{\hat{y}_{n+1} - y_{n+1}}{\hat{S}\sqrt{X_{n+1}^T (X^T X)^{-1} X_{n+1} + 1}},$$

имеющую распределение Стьюдента с степенями свободы $\nu = n - k - 1$.