1. Множественная регрессия

Пусть имеются случайные величины $\xi_1, \ldots, \xi_n$ с совместной плотностью $f(x_1, x_2, \ldots, x_n)$. Регрессией или условным математическим ожиданием случайной величины ξ_1 при фиксированных значениях случайных величин $\xi_i,\ i = 2, \ldots, n$, называется функция, зависящая от переменных $x_2, \ldots, x_n$, вида

$$E(\xi_1|\xi_2 = x_2, \ldots, \xi_n = x_n) = m_1(x_2, \ldots, x_n) = \frac{\int\limits_{-\infty}^{\infty} x_1 f(x_1, \ldots, x_n)dx_1}{\int\limits_{-\infty}^{\infty} f(x_1, \ldots, x_n)dx_1}$$

Поверхность регрессии описывается уравнением $x_1 = m_1(x_2, \ldots, x_n)$.

1.1. Линейная среднеквадратичная регрессия

Для упрощения выводов будем считать $E\xi_i = 0$. Формулы для произвольного центра тяжести будут получаться простой заменой ξ_i на $\xi_i - m_i$. Среднеквадратическая регрессия величины ξ_1 относительно $\xi_2, \ldots, \xi_n$ определяется как гиперплоскость

$$\xi_1 = \beta_{12}\xi_2 + \ldots + \beta_{1n}\xi_n, \tag{1}$$

дающая наилучшую аппроксимацию математического ожидания квадрата разности

$$E(\xi_1 - \beta_{12}\xi_2 - \ldots - \beta_{1n}\xi_n)^2 \to min, \tag{2}$$

т.е. правая часть (1) является наилучшей линейной оценкой ξ_1 величинами $\xi_2, \ldots, \xi_n$ в смысле минимума (2). Коэффициенты β_{1i} называются *частными коэффициентами регрессии.*

Для их вычисления продифференцируем (2) по β_{1i}.

$$\begin{cases} \frac{d}{d\beta_{12}}: & -2E\xi_2(\xi_1 - \beta_{12}\xi_2 - \ldots - \beta_{1n}\xi_n) = 0 \\ & \ldots \\ \frac{d}{d\beta_{1n}}: & -2E\xi_n(\xi_1 - \beta_{12}\xi_2 - \ldots - \beta_{1n}\xi_n) = 0 \end{cases}$$

Это равносильно при $\lambda_{ij} = E\xi_i\xi_j$ системе уравнений

$$\begin{cases} \lambda_{22}\beta_{12} + \ldots + \lambda_{2n}\beta_{1n} = \lambda_{21} \\ \ldots \\ \lambda_{n2}\beta_{12} + \ldots + \lambda_{nn}\beta_{1n} = \lambda_{n1} \end{cases} \tag{3}$$

Определитель этой системы равен алгебраическому дополнению Λ_{11} матрицы вторых моментов

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2n} \\ & & \ldots & \\ \lambda_{n1} & \lambda_{n2} & \ldots & \lambda_{nn} \end{bmatrix}$$

По теореме Крамера, β_{12} равно отношению определителя

$$\begin{vmatrix} \lambda_{21} & \lambda_{23} & \ldots & \lambda_{2n} \\ & & \ldots & \\ \lambda_{n1} & \lambda_{n3} & \ldots & \lambda_{nn} \end{vmatrix},$$

который равен $-\Lambda_{12}$, к определителю Λ_{11}, т.е. $\beta_{12} = -\Lambda_{12}/\Lambda_{11}$. Коэффициент β_{13} равен отношению определителя

$$\begin{vmatrix} \lambda_{22} & \lambda_{21} & \lambda_{24} & \ldots & \lambda_{2n} \\ & & & \ldots & \\ \lambda_{n2} & \lambda_{n1} & \lambda_{n4} & \ldots & \lambda_{nn} \end{vmatrix}, \tag{4}$$

полученного в результате замены в матрице Λ_{11} второго столбца на столбец свободных членов системы (3), к определителю Λ_{11}. Поменяем местами

первые два столбца. Определитель (4) равен

$$-\begin{vmatrix} \lambda_{21} & \lambda_{22} & \lambda_{24} & \dots & \lambda_{2n} \\ & & & \dots & \\ \lambda_{n1} & \lambda_{n2} & \lambda_{n4} & \dots & \lambda_{nn} \end{vmatrix} = -\Lambda_{13}, \tag{5}$$

т.е.

$$\beta_{13} = -\frac{\Lambda_{13}}{\Lambda_{11}}.$$

Аналогично получаем, что

$$\beta_{1k} = -\frac{\Lambda_{1k}}{\Lambda_{11}}. \tag{6}$$

Простой перестановкой индексов получаем

$$\beta_{ik} = -\frac{\Lambda_{ik}}{\Lambda_{ii}}. \tag{7}$$

1.2. Остатки и остаточная дисперсия

Ограничимся рассмотрением случая, когда $|\Lambda_{11}| \neq 0$ и все $E\xi_i = 0$. Случайная величина вида

$$\eta_{1\cdot 23\dots n} = \xi_1 - \beta_{12}\xi_2 - \dots - \beta_{1n}\xi_n \tag{8}$$

называется остатком величины ξ_1 относительно $\xi_2, \dots, \xi_n$. Покажем, что *остаток не коррелирован ни с одной из "вычитаемых"случайных величин.* Из (6) и из того, что $1 = \Lambda_{11}/\Lambda_{11}$ имеем выражение остатка через линейную комбинацию случайных величин

$$\eta_{1\cdot 23\dots n} = \frac{1}{\Lambda_{11}} \sum_{k=1}^{n} \Lambda_{1k}\xi_k. \tag{9}$$

Отсюда, с учетом того, что $E\xi_i = 0$, получаем

$$E\eta_{1\cdot 23\dots n} = 0. \tag{10}$$

$$E\xi_i\eta_{1\cdot 23\ldots n} = \frac{1}{\Lambda_{11}}\sum_{k=1}^{n} E\xi_i\xi_k\Lambda_{1k} = \frac{1}{\Lambda_{11}}\sum_{k=1}^{n}\lambda_{ik}\Lambda_{1k} = \begin{cases} \frac{\Lambda}{\Lambda_{11}}, & i=1, \\ 0, & i\neq 1, \end{cases} \qquad (11)$$

Так как $\sum\limits_{k=1}^{n}\lambda_{1k}\Lambda_{1k} = \Lambda$, а $\sum\limits_{k=1}^{n}\lambda_{ik}\Lambda_{1k} = 0$ равен определителю матрицы, у которой вместо первой стоит i -ая строчка, а если в матрице две строчки одинаковы, ее определитель равен нулю.

Вычислим дисперсию остатка $\sigma^2_{1\cdot 23\ldots n}$, которую будем называть остаточной. С учетом (10) равенства нулю математических ожиданий компонент

$$\sigma^2_{1\cdot 23\ldots n} = E(\eta^2_{1\cdot 23\ldots n}).$$

Заменим одну из η на (9). В силу (11) получаем

$$\sigma^2_{1\cdot 23\ldots n} = E\left(\frac{1}{\Lambda_{11}}\sum_{i=1}^{n}\Lambda_{1i}\xi_i\eta_{1\cdot 23\ldots n}\right) = E(\xi_1\eta_{1\cdot 23\ldots n}) = \frac{\Lambda}{\Lambda_{11}}. \qquad (12)$$

Остаточная дисперсия используется в качестве меры приближения случайной величины ξ_1 линейной комбинацией $\xi_1^* = \beta_{12}\xi_2 + \ldots + \beta_{1n}\xi_n$.

1.3. Частная корреляция

Частным коэффициентом корреляции между двумя случайными величинами ξ_1 и ξ_2 называется коэффициент корреляции между остатками

$$\varrho_{12\cdot 34\ldots n} = cor(\eta_{1\cdot 34\ldots n}\eta_{2\cdot 34\ldots n}) \qquad (13)$$

имеющий смысл меры линейной зависимости между величинами после устранения изменений, вызванных влиянием $\xi_3, \ldots, \xi_n$. Так как $E\eta_{1\cdot 34\ldots n} = 0$ и $E\eta_{2\cdot 34\ldots n} = 0$, то

$$\varrho_{12\cdot 34\ldots n} = \frac{E(\eta_{1\cdot 34\ldots n}\eta_{2\cdot 34\ldots n})}{\sqrt{E(\eta_{1\cdot 34\ldots n})^2 E(\eta_{2\cdot 34\ldots n})^2}}$$

Итак, в силу (11)

$$E\eta^2_{1\cdot 34\ldots n} = E(\xi_1\eta_{1\cdot 34\ldots n}) = \frac{\Lambda_{22}}{\Lambda_{22\cdot 11}} = \frac{\Lambda_{22}}{\Lambda_{11\cdot 22}},$$
$$E\eta^2_{2\cdot 34\ldots n} = E(\xi_2\eta_{2\cdot 34\ldots n}) = \frac{\Lambda_{11}}{\Lambda_{11\cdot 22}}.$$

Далее

$$E(\eta_{1\cdot 34\ldots n}\eta_{2\cdot 34\ldots n}) = E(\xi_1\eta_{2\cdot 34\ldots n}), \tag{14}$$

так как $\eta_{1\cdot 34\ldots n} = \frac{1}{\Lambda_{22\cdot 11}} \sum\limits_{k=1,k\neq 2}^{n} \Lambda_{22\cdot 1k}\xi_k$ и

$$E\xi_k\eta_{2\cdot 34\ldots n} \begin{cases} = 0, & k = 3, 4, \ldots, n \\ \neq 0, & k = 2 \\ \neq 0, & k = 1 \end{cases}$$

Так как $\eta_{2\cdot 34\ldots n} = \frac{1}{\Lambda_{11\cdot 22}} \sum\limits_{k=2}^{n} \Lambda_{11\cdot 2k}\xi_k$, то (14) имеет вид

$$E(\xi_1\eta_{2\cdot 34\ldots n}) = \frac{1}{\Lambda_{11\cdot 22}} \sum_{k=2}^{n} \Lambda_{11\cdot 2k}E\xi_1\xi_k = \frac{1}{\Lambda_{11\cdot 22}} \sum_{k=2}^{n} \Lambda_{11\cdot 2k}\lambda_{1k} = -\frac{\Lambda_{12}}{\Lambda_{11\cdot 22}}.$$

Таким образом,

$$\varrho_{12\cdot 34\ldots n} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} \tag{15}$$

Между частной корреляцией и соответствующим частным коэффициентом регрессии имеет место линейная зависимость. Согласно (15) и (6) имеем

$$\varrho_{12\cdot 34\ldots n} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} = -\frac{\Lambda_{12}}{\Lambda_{11}}\sqrt{\frac{\Lambda_{11}}{\Lambda_{22}}} = \beta_{12\cdot 34\ldots n}\sqrt{\frac{\Lambda_{11}}{\Lambda_{22}}}. \tag{16}$$

Для случая $n = 3$ матрица вторых моментов Λ имеет вид

$$\Lambda = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\varrho_{12} & \sigma_1\sigma_3\varrho_{13} \\ \sigma_1\sigma_2\varrho_{12} & \sigma_2^2 & \sigma_2\sigma_3\varrho_{23} \\ \sigma_1\sigma_3\varrho_{13} & \sigma_2\sigma_3\varrho_{23} & \sigma_3^2 \end{bmatrix}$$

$$\varrho_{12\cdot 3} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} = \frac{\sigma_1\sigma_2\varrho_{12}\sigma_3^2 - \sigma_1\sigma_3\varrho_{13}\sigma_2\sigma_3\varrho_{23}}{\sigma_2\sigma_3\sqrt{1-\varrho_{23}^2}\sigma_1\sigma_3\sqrt{1-\varrho_{13}^2}} = \frac{\varrho_{12} - \varrho_{13}\varrho_{23}}{\sqrt{1-\varrho_{23}^2}\sqrt{1-\varrho_{13}^2}}$$

ПРИМЕР. В качестве "зависимой"переменной ξ_1 будем рассматривать среднюю продолжительность в году, а в качестве независимых переменных – национальное богатство ξ_2 и объем ξ_3 продаваемой водки за год, приходящийся на одного человека (в бутылках). В качестве индивидов $N = 7$ будем рассматривать годы: 1970, 1975, 1980, 1985, 1990, 1995 и 1998. Звездочками отмечены артефакты или неполные наблюдения, замененные средними.

j	Годы	Продолж-ть жизни ξ_1	Нац.богатсво (млрд.дол.) ξ_2	Водка (бут.в год) ξ_3
1	1970	68.9	2372*	25.3
2	1975	68.1	2372*	28
3	1980	67.6	2489	30
4	1985	69.2	3379	23.5
5	1990	69.2	4130	18
6	1995	64.6	1171	38.4
7	1998	67	689	29.6

Выборочные средние и стандартные оклонения $m_i \pm S_i$ равны соответственно 67.8 ± 1.6, 2372 ± 1182.7, 27.5 ± 6.3. Матрица вторых моментов (ее детерминант равен $\Lambda = 3882374$) и ее алгебраических дополнений имеют вид:

λ_{ij}	ξ_1	ξ_2	ξ_3
ξ_1	2.31	1291	-8.32
ξ_2		1199027	-5294.52
ξ_3			34.40

Λ_{ij}	ξ_1	ξ_2	ξ_3
ξ_1	13214148	-368	138823
ξ_2		10,1	1468
ξ_3			1097860

Коэффициенты корреляции равны $\varrho_{12} = 0.78$, $\varrho_{13} = -0.93$, $\varrho_{23} = -0.82$. Вычисляем частные коэффициены регрессии частные коэффициенты корреляции и дисперсию остатка:

$$\beta_{12} = -\frac{\Lambda_{12}}{\Lambda_{11}} = -\frac{-368}{13214148} = 2.8 \cdot 10^{-5}, \quad \varrho_{12\cdot 3} = -\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} = 0.032,$$

$$\beta_{13} = -\frac{\Lambda_{13}}{\Lambda_{11}} = -\frac{138823}{13214148} = -0.24, \quad \varrho_{13\cdot 2} = -\frac{\Lambda_{13}}{\sqrt{\Lambda_{11}\Lambda_{33}}} = -0.82,$$

$$\sigma^2_{1\cdot 23} = \frac{\Lambda}{\Lambda_{11}} = \frac{3882374}{13214148} = 0.29.$$

Свободный член вычисляется непосредственно $\beta_{11} = m_1 - \beta_{12}m_2 - \beta_{13}m_3 = 74.28$. Наблюдаемая положительная корреляция между продолжительностью жизни и ростом национального богатсва объясняется тем, что при снижении его уровня увеличивается продажа водки, которая и приводит к снижению продолжительности жизни.

1.4. Множественный коэффициент корреляции

Множественным коэффициентом корреляции $\varrho_{1(23\ldots n)}$ между случайной величиной ξ_1 и множеством случайных величин $\xi_2, \ldots, \xi_n$ называется коэффициент корреляции ξ_1 и ее наилучшей среднеквадратической оценкой, построенной по набору случайных величин $\xi_2, \ldots, \xi_n$. В случае $E\xi_i = 0$ множественный коэффициент корреляции имеет вид

$$\varrho_{1(23\ldots n)} = \frac{E(\xi_1 \xi_1^*)}{\sqrt{E\xi_1^2 E(\xi_1^*)^2}}, \tag{17}$$

где $\xi_1^* = \beta_{12}\xi_2 + \ldots + \beta_{1n}\xi_n$ – наилучшая оценка ξ_1 по $\xi_2, \ldots, \xi_n$. По (8) $\xi_1^* = \xi_1 - \eta_{1\cdot 23\ldots n}$, поэтому с учетом (11) и (12)

$$E(\xi_1\xi_1^*) = E(\xi_1(\xi_1 - \eta_{1\cdot 23\ldots n})) = E\xi_1^2 - E\xi_1\eta_{1\cdot 23\ldots n} = \lambda_{11} - \frac{\Lambda}{\Lambda_{11}}$$

$$E(\xi_1^*)^2 = E(\xi_1^2 - 2\xi_1\eta_{1\cdot 23\ldots n} + \eta_{1\cdot 23\ldots n}^2) = \lambda_{11} - \frac{\Lambda}{\Lambda_{11}}.$$

Следовательно,

$$\varrho_{1(23\ldots n)} = \frac{\sqrt{\lambda_{11} - \frac{\Lambda}{\Lambda_{11}}}}{\sqrt{\lambda_{11}}} = \sqrt{1 - \frac{\Lambda}{\Lambda_{11}\lambda_{11}}}. \tag{18}$$

Для численных вычислений оказывается полезной формула

$$\varrho_{1(23\ldots n)}^2 = 1 - \frac{\sigma_{1\cdot 23\ldots n}^2}{\sigma_1^2} = \frac{\sigma_1^2 - \sigma_{1\cdot 23\ldots n}^2}{\sigma_1^2}, \tag{19}$$

так как по (12) $\sigma_{1\cdot 23\ldots n}^2 = \frac{\Lambda}{\Lambda_{11}}$, а $\lambda_{11} = \sigma_1^2$.

Квадрат множественного коэффициента корреляции или *коэффициент детерминации* равен доле дисперсии, объясняемой регрессией, то есть влиянием случайных величин $\xi_2, \ldots, \xi_n$ на ξ_1.

ПРИМЕР (продолжение). Коэффициент детерминации равен

$$R^2 = \varrho_{1(23)}^2 = \frac{\sigma_1^2 - \sigma_{1\cdot 23}^2}{\sigma_1^2} = \frac{2.31 - 0.29}{2.31} = 0.87,$$

множественный коэффициент корреляции равен $\varrho_{1(23)} = 0.93$.

1.5. Отношение дисперсий и частный коэффициент корреляции

$$\frac{\beta_{12}^2\sigma_{2\cdot34\ldots n}^2}{\sigma_{1\cdot234\ldots n}^2} = \frac{\varrho_{12\cdot34\ldots n}^2}{1-\varrho_{12\cdot34\ldots n}^2} \tag{20}$$

Доказательство этого утверждения опирается на равенство Якоби:

$$\Lambda\cdot\Lambda_{11\cdot22} = \Lambda_{11}\cdot\Lambda_{22} - \Lambda_{12}\cdot\Lambda_{21} \tag{21}$$

Используя из (6) и (12) выражения для

$$\beta_{12}^2 = \frac{\Lambda_{12}^2}{\Lambda_{11}^2},$$
$$\sigma_{1\cdot23\ldots n}^2 = \frac{\Lambda}{\Lambda_{11}}, \qquad \sigma_{2\cdot34\ldots n}^2 = \frac{\Lambda_{11}}{\Lambda_{11\cdot22}},$$

получаем левую часть в виде:

$$\frac{\beta_{12}^2\sigma_{2\cdot34\ldots n}^2}{\sigma_{1\cdot234\ldots n}^2} = \frac{\Lambda_{12}^2}{\Lambda_{11}^2}\cdot\frac{\Lambda_{11}}{\Lambda_{11\cdot22}}\cdot\frac{\Lambda_{11}}{\Lambda} = \frac{\Lambda_{12}^2}{\Lambda\cdot\Lambda_{11\cdot22}} = \frac{\Lambda_{12}^2}{\Lambda_{11}\cdot\Lambda_{22}-\Lambda_{12}\cdot\Lambda_{21}}.$$

С другой стороны, из (15) $\varrho_{12\cdot34\ldots n}^2 = \frac{\Lambda_{12}^2}{\Lambda_{11}\Lambda_{22}}$, следовательно,

$$\frac{\varrho_{12\cdot34\ldots n}^2}{1-\varrho_{12\cdot34\ldots n}^2} = \frac{\Lambda_{12}^2}{\Lambda_{11}\Lambda_{22}\left(1-\frac{\Lambda_{12}^2}{\Lambda_{11}\Lambda_{22}}\right)}.$$

Равенство Якоби получается из произведения определителей $(\Lambda_{11}\Lambda_{22} - \Lambda_{12}\Lambda_{21})|\Lambda| = |\Lambda|^2\Lambda_{11\cdot22}$ соответствующих матриц

$$\left[\begin{array}{cc|ccc} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} & \ldots & \Lambda_{1n} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} & \ldots & \Lambda_{2n} \\ \hline \multicolumn{2}{c|}{\mathbb{O}} & \multicolumn{3}{c}{\mathbb{I}} \end{array}\right] \left[\begin{array}{cccc} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \lambda_{n1} & \lambda_{n2} & \ldots & \lambda_{nn} \end{array}\right] = \left[\begin{array}{cc|ccc} |\Lambda| & 0 & 0 & \ldots & 0 \\ 0 & |\Lambda| & 0 & \ldots & 0 \\ \hline \lambda_{31} & \lambda_{32} & \lambda_{33} & \ldots & \lambda_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \lambda_{n1} & \lambda_{n2} & \lambda_{3n} & \ldots & \lambda_{nn} \end{array}\right]$$

1.6. Задача регрессии в матричном виде

Уравнения линейной регрессии представим в виде $Y = X\beta + e$, где зависимая переменная задается вектором $Y = (y_1, \ldots, y_n)^T$, матрица

$$X^T = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_{11} & x_{12} & \ldots & x_{1n} \\ & & \ldots & \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{bmatrix} \tag{22}$$

называется матрицей плана, $\beta = (\beta_0, \ldots, \beta_p)^T$ – вектор параметров, e – нормально распределенный вектор с нулевым средним и ковариационной матрицей вида $\sigma^2\mathbf{I}$.

При минимизации квадратичной формы $S = (Y - X\beta)^T(Y - X\beta)$ вектор оценок $b = (b_0, \ldots, b_p)^T$ является решением системы нормальных уравнений

$$(X^TX)\beta = X^TY \tag{23}$$

Действительно, дифференцирование по вектору параметров приводит к

$$2X^T(Y - X\beta) = 0 \Longleftrightarrow X^TY = X^TX\beta.$$

Следовательно,

$$b = (X^TX)^{-1}X^TY. \tag{24}$$

Ковариация оценок

Ковариационная матрица оценок частных коэффициентов регрессии можно представить в виде:

$$\begin{aligned} Cov(b) = \mathbf{E}(b - \mathbf{E}b)(b - \mathbf{E}b)^T = \\ = \mathbf{E}((X^TX)^{-1}X^T(Y - \mathbf{E}Y))((X^TX)^{-1}X^T(Y - \mathbf{E}Y))^T = \\ = (X^TX)^{-1}X^TCov(Y)((X^TX)^{-1}X^T)^T = \\ = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}. \end{aligned}$$

Остаточная сумма квадратов

Выясним, насколько согласовано с выражением для остаточной дисперсии $\sigma^2_{1\cdot 23\ldots n} = \frac{\Lambda}{\Lambda_{11}}$ выражение для суммы квадратов остатков

$$\begin{aligned}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) = Y^TY - Y^TX\hat{\beta} = \\ = Y^TY - Y^TX(X^TX)^{-1}X^TY = \\ = \begin{vmatrix} Y^TY & Y^TX \\ X^TY & X^TX \end{vmatrix} / |X^TX| = \frac{1}{G_{11}},\end{aligned} \tag{25}$$

где $G = S^{-1}$, $S = Z^TZ$ матрица вторых выборочных моментов с точностью до множителя n, $Z = [Y, X]$ – матрица размерности n на $p+1$, X – матрица плана размерности n на p. Заметим, что сумма квадратов остатков не зависит от смещений, поэтому можно говорить о центральных моментах. Здесь мы воспользовались разложением определителя матрицы блочной структуры

$$|X| = |D|(a - b^TD^{-1}c) = a|D| - \sum_i b_i \sum_k D_{ki}c_k,$$

$$\text{где} \quad X = \begin{bmatrix} a & b^T \\ c & D \end{bmatrix}.$$

Таблица анализа дисперсий

Обозначим через $\hat{\mathbf{Y}} = Xb$ вектор наилучшего в среднеквадратичном смысле линейного предсказания переменной Y по переменным $X_1, \ldots, X_p$.

Из таблицы 1 дисперсионного анализа для модели множественной регрессии может быть получена оценка коэффициента детерминации $R^2 = \frac{SS_R}{SS_T}$ и выражение для статистики $F = \frac{MS_R}{MS_E}$, которая при справедливости нулевой гипотезы о равенстве нулю частных коэффициентов регрессии имеет распределение Фишера с p и $n - p - 1$ степенями свободы.

2. Проверка нормальности остатков

Рассмотрим подробнее процедуру построения графика соотношения эмпирических и теоретических квантилей $QQ - plot$. Пусть имеется выборка $x_1, \ldots, x_n$, распределение которой нас интересует. Для примера, рассмотрим данные, полученные в результате моделирования в R при $n = 100$

$$x < -rnorm(n, mean = 1, sd = 2)$$

Построим ее эмпирическую функцию распределения, при помощи функции $plot(ecdf(x))$, график функции распределения, например, $curve(pnorm, -3, 3)$ и отметим эмпирические Qe и теоретические квантили Qt для всех значений $F^{-1}\left(\frac{i-0.5}{n}\right)$, где $F(x)$ функция распределения. Половинка вычитается для того чтобы не получить $F^{-1}(1) = +\infty$.

Источник дисперсии	Сумма квадратов	степени свободы	средний квадрат
Регрессия	$SS_R = \sum_{i=1}^{n} (\hat{\mathbf{y}}_i - \bar{y})^2$	$\nu_R = p$	$MS_R = SS_R/\nu_R$
Откл.от регр.	$SS_E = \sum_{i=1}^{n} (y_i - \hat{\mathbf{y}}_i)^2$	$\nu_E = n - p - 1$	$MS_E = SS_E/\nu_E$
полная	$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$\nu_T = n - 1$	$MS_T = SS_T/\nu_T$

Таблица 1. Дисперсионный анализ для модели множественной регрессии ($SS_E = SS_T - SS_R$)

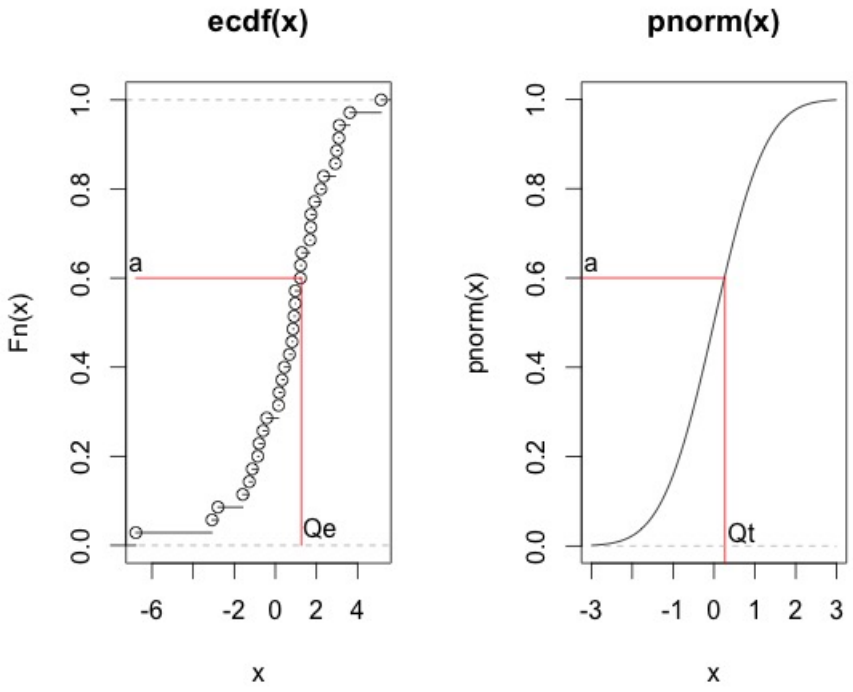
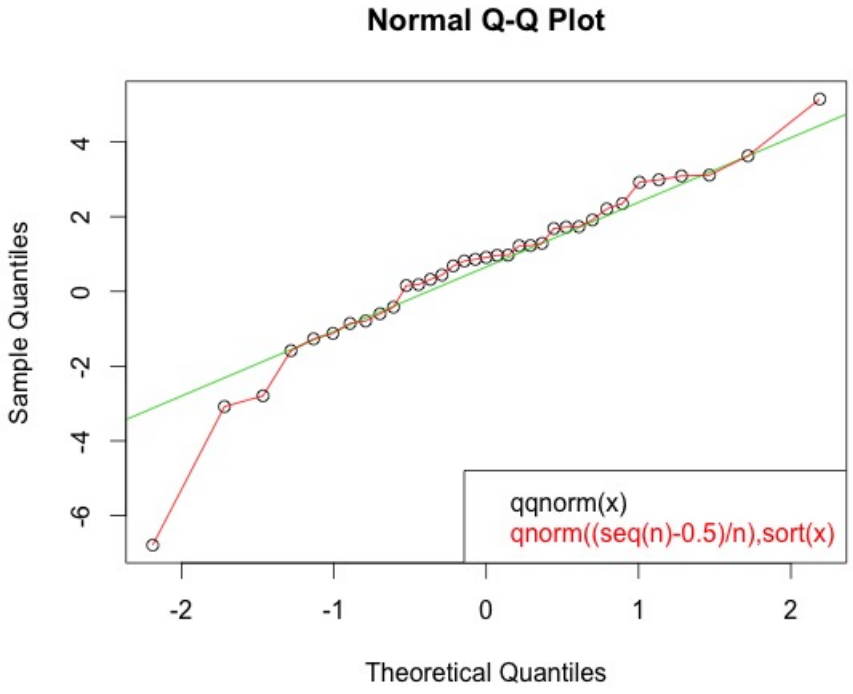
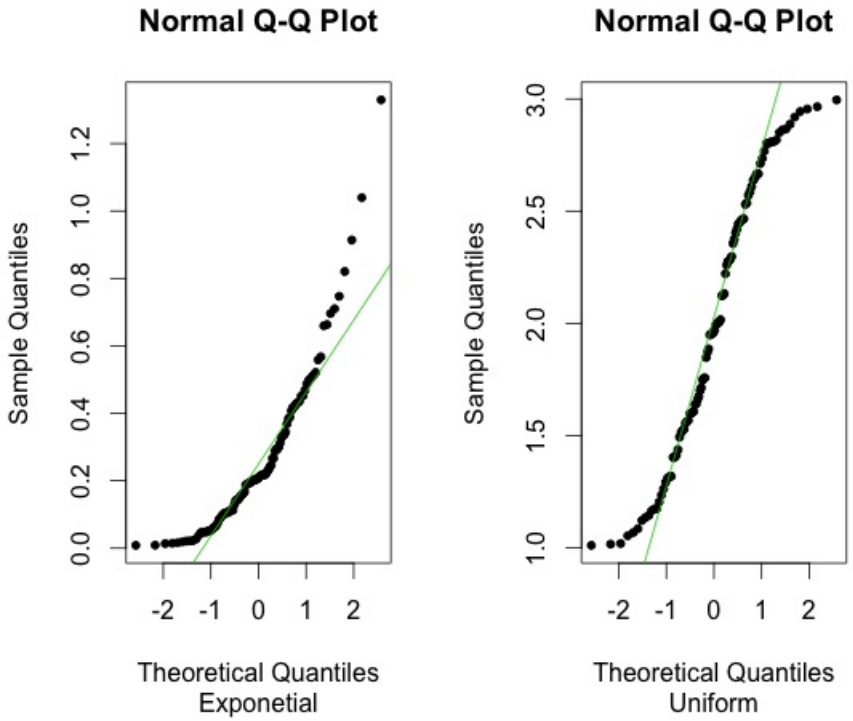


График квантиль-квантиль строится при помощи функции $qqnorm(x)$. По вертикальной оси откладываются упорядоченные значения $x_1, \ldots, x_n$, по горизонтальной $F^{-1}\left(\frac{i-0.5}{n}\right)$, в случае нормального распределения

$$qnorm((seq(n) - 0.5)/n).$$





Примеры графиков квантиль-квантиль с теоретическими нормальными квантилями и эмпирическими квантилями равномерного и экспоненциального распределений представлены на рисунках.

ПРИМЕР (продолжение).y - продолжительность жизни, x_1 - объем национального богатства, x_2 - количество продаваемых за год бутылок водки, приходящееся на одного человека, уравнение регрессии:

$$y = 74.27634 + 0.00003x_1 - 0.23754x_2,$$

множественный коэффициент корреляции $R = 0.9341$, коэффициент детерминации $R^2 = 0.8726$, $p < 0.01624$ значимость статистики Фишера $F(2,4) = 13.696$, частные коэффициенты корреляции:

$$r_{12\cdot 3} = 0.0318,\ p = 0.9523, r_{13\cdot 2} = -0.824,\ p = 0.0437$$

.

2.1. Задание по теме „Множественная регрессия“

Для того чтобы сделать обработку по множественной регрессии в программе $StatSoft$, нужно запустить программу *Statistics - Multiple Regression*. Далее нужно выбрать переменные ($Variables$). При необходимости привлечения пошаговой процедуры ставится галочка в *Advanced options* (*stepwise or ridge regression*). Далее OK. В таком случае нужно будет выбрать метод: $Forward$ или $Backward$ в зависимости от того, происходит ли отбор информативных или исключение малоинформативных признаков. Коэффициенты регрессии, их значимость можно получить из $Summary$: *Regression results*, таблица дисперсионного анализа из $ANOVA$ ($Overal\ goodnessof fit$), частные коэффициенты регрессии соответственно из *Partial correlations*.

1. ASIAN% – процент лиц азиатского происхождения;
2. BLACK% – процент афроамериканцев;
3. TEMPER – среднеиюльская температура;
4. GROSS – средеяя арендная плата;
5. PRECEP – годовое количество осадков;
6. TRANSP – процент лиц, использующих в городе общественный транспорт;
7. DEGREE – процент людей с образованием бакалавр и выше.

Построить уравнение регрессии по данным из табл.2, взяв в качестве зависимой переменной один признак (номер варианта), в качестве независимых – все остальные. Проверить значимость регрессии и частных коэффициентов, построить частные коэффициенты корреляции. Для проверки вычислить коэффициенты корреляции между остатком и вычитаемыми переменными.

Вариант		1	2	3	4	5	6	7
city	st	asian%	black%	temper	cross	precep	transp	degree
COLUMBUS	1	2,4	22,6	73,2	422	38,1	4,7	24,6
CLEVELAN	1	1	46,6	71,9	322	36,6	14,1	8,1
CINCINNA	1	1,1	37,9	76,4	329	40,7	11,2	22,2
TOLEDO	1	1	19,7	72,1	378	33	3	14,1
AKRON	1	1,2	24,5	71,9	360	36,8	3,3	14,9
HOUSTON	2	4,1	28,1	83,5	390	50,8	6,5	25,1
DALLAS	2	2,2	29,5	85,9	426	36,1	6,7	27,1
SAN ANTO	2	1,1	7	85	369	31	4,9	17,8
EL PASO	2	1,2	3,4	82,3	349	8,8	3,1	16,2
AUSTIN	2	3	12,4	84,5	410	31,9	5,1	34,4
FORT WOR	2	2	22	85,3	403	33,7	1,7	21,5
ARLINGTO	2	3,9	8,4	85,3	444	33,7	0,2	30
CORPUS C	2	0,9	4,8	84,1	373	30,1	1,9	17,8
LOS ANGE	3	9,8	14	74,3	600	14,8	10,5	23
SAN DIEG	3	11,8	9,4	71	602	9,9	4,2	29,8
SAN JOSE	3	19,5	4,7	69,5	755	14,4	3,5	25,3
SAN FRAN	3	29,1	10,9	59,1	653	19,7	33,5	35
LONG BEA	3	13,6	13,7	73,1	605	11,8	6	23,2
SACRAMEN	3	15	15,3	75,7	495	17,5	4	23,5
FRESNO	3	12,5	8,3	81,9	441	10,6	2,1	19,1
OAKLAND	3	14,8	43,9	62,1	538	24,3	17,9	27,2
SANTA AN	3	9,7	2,6	72,6	736	12,3	8,4	10,6
ANAHEIM	3	9,4	2,5	72,6	712	12,3	3,8	18,8
RIVERSID	3	5,2	7,4	77,9	575	9,6	1,6	19,3
STOCKTON	3	22,8	9,6	77,7	476	14	2	15

Таблица 2. Статистические данные об американских городах.