Множественная линейная регрессия

Грауэр Л.В.

Регрессионный анализ

Y — зависимая переменная / отклик

$X_1, \ldots, X_k$ — независимые переменные / факторы / предикторы

$$y = f(x_1, \ldots, x_k) + \varepsilon$$

$(x_{i1}, \ldots, x_{ik}, y_i): \quad y_i = f(x_{i1}, \ldots, x_{ik}) + \varepsilon_i, \quad i = 1, \ldots, n$

$E\varepsilon_i = 0,\ D\varepsilon_i = \sigma^2,\ K(\varepsilon_i \varepsilon_j) = 0, i \neq j$

$\hat{y} = \hat{f}(x_1, \ldots, x_k)$

$$y_i = f(x_{i1}, \ldots, x_{ik}; \beta) + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\hat{y} = f(x_1, \ldots, x_k; \hat{\beta})$$

$$E_x Y = E_x[f(x_1, \ldots, x_k) + \varepsilon] = f(x_1, \ldots, x_k)$$

Модель простой линейной регрессии

$Y, \quad X$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$

$$E\varepsilon_i = 0,\ D\varepsilon_i = \sigma^2,\ K(\varepsilon_i \varepsilon_j) = 0, i \neq j$$

$$\varepsilon_i \sim N(0, \sigma)$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to \min_{\beta_0, \beta_1}$$

МНК-оценки

$$\frac{\partial Q}{\partial \beta_0} =$$

$$\frac{\partial Q}{\partial \beta_1} =$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Остаточная сумма квадратов

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Остатки $\quad e_i = y_i - \hat{y}(x_i), \quad i = 1, \dots, n$

$$RSS = \sum_{i=1}^{n} e_i^2$$

$$S^2 = \frac{RSS}{n-2}$$

$$ES^2 = \sigma^2$$

$$Z_S = \frac{S^2(n-2)}{\sigma^2}$$

Свойства МНК-оценок β_0, β_1

$$E\hat{\beta}_0 = \beta_0, \quad E\hat{\beta}_1 = \beta_1,$$

$$D\hat{\beta}_0 = \frac{\sigma^2 \sum_{i=1}^n x_i^2 / n}{nD_x^*}, \qquad D\hat{\beta}_1 = \frac{\sigma^2}{nD_x^*}$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sqrt{D\hat{\beta}_0}\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \sqrt{D\hat{\beta}_1}\right)$$

$$Z_\beta = \frac{\sigma}{S} \frac{\hat{\beta}_i - \beta_i}{\sqrt{D\hat{\beta}_i}}$$

Свойства МНК-оценок линейной регрессии

$$E\hat{y}(x) = \beta_0 + \beta_1 x$$

$$D\hat{y}(x) = \frac{\sigma^2}{n}\left(1 + \frac{(x - \bar{x})^2}{D_x^*}\right)$$

$$\hat{y}(x) \sim N\left(\beta_0 + \beta_1 x, \sqrt{D\hat{y}(x)}\right)$$

$$Z_y = \frac{\sigma}{S}\frac{\hat{y}(x) - (\beta_0 + \beta_1 x)}{\sqrt{D\hat{y}(x)}}$$

Интервальные оценки

$$\beta_0 \quad \hat\beta_0 \pm t_{1-\alpha/2}(n-2) S \sqrt{\frac{\sum_{i=1}^{n} x_i^2 / n}{n D_x^*}}$$

$$\beta_1 \quad \hat\beta_1 \pm t_{1-\alpha/2}(n-2) S \sqrt{\frac{1}{n D_x^*}}$$

$$y(x) \quad \hat y(x) \pm t_{1-\alpha/2}(n-2) S \sqrt{\frac{1 + \frac{(x-\bar x)^2}{D_x^*}}{n}}$$

$$\sigma^2 \quad \left(\frac{S^2(n-2)}{\chi^2_{1-\alpha/2}(n-2)}, \ \frac{S^2(n-2)}{\chi^2_{\alpha/2}(n-2)} \right)$$

Пример

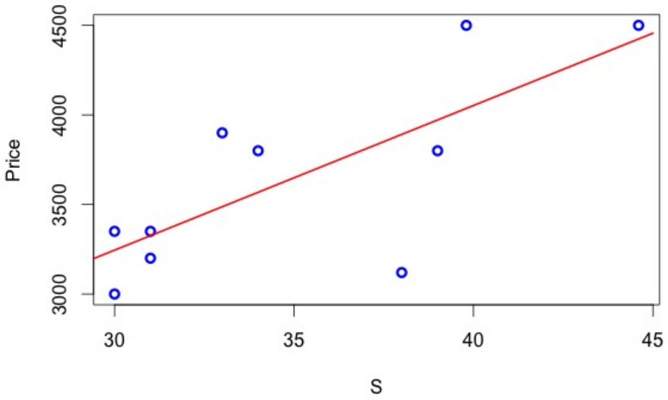
S, м^2	39.8	38.0	44.6	31.0	31.0	30.0	33.0	39.0	34.0	30.0
P, тыс.руб	4500	3120	4500	3350	3200	3350	3900	3800	3800	3000

$P = a + bS + \varepsilon$

$\hat{a} = 821.89, \quad \hat{b} = 80.77$

$\hat{P}(S) = 821.89 + 80.77S$

$\hat{P}(S_1) = 4036.54$, $\hat{P}(S_2) = 3891.15$, $\hat{P}(S_3) = 4424.23$,
$\hat{P}(S_4) = 3325.76$, $\hat{P}(S_5) = 3325.76$, $\hat{P}(S_6) = 3244.99$,
$\hat{P}(S_7) = 3487.30$, $\hat{P}(S_8) = 3971.92$, $\hat{P}(S_9) = 3568.07$,
$\hat{P}(S_{10}) = 3244.99$

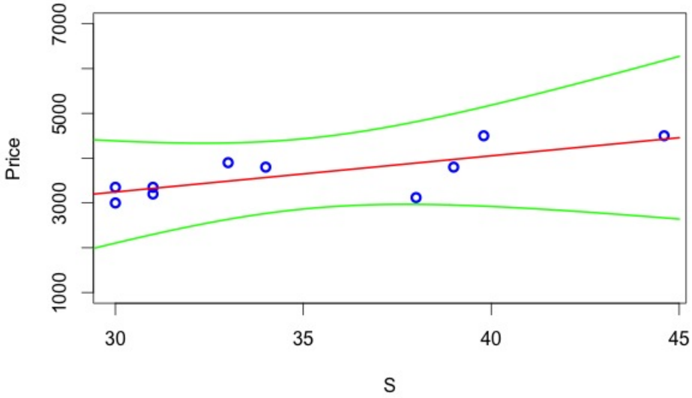


$$RSS = 1156332$$

$\gamma = 1 - \alpha = 0.95$

a : $(-926.35,\ 2570.13)$

b : $(31.33,\ 130.21)$



$\hat{P}(40) = 4052.69$

$P\{P(40) > 5000\}$

Нелинейные модели, сводящиеся к линейным

Обратное преобразование: $Y = \beta_0 + \beta_1(1/X) + \varepsilon$.
Замена $Z = 1/X$.

Логарифмическое преобразование: $Y = \beta_0 + \beta_1 \ln X + \varepsilon$.
Замена $Z = \ln X$.

Мультипликативная модель: $Y = \alpha X^{\beta}\varepsilon$.

$$\ln Y = \ln\alpha + \beta\ln X + \ln\varepsilon.$$

Обратная экспоненциальная модель: $Y = \dfrac{1}{1 + \alpha e^{\beta_1 X + \varepsilon}}$.

$$\ln(1/Y - 1) = \ln\alpha + \beta_1 X + \varepsilon.$$

Фиктивные переменные

X — категориальная переменная: $X_1, \ldots, X_p$

Фиктивные переменные

$Z_1, \ldots, Z_{p-1}$

$$
\begin{array}{c|l}
X = X_1 & Z_1 = Z_2 = \ldots = Z_{p-1} = 0 \\
X = X_2 & Z_1 = 1, Z_2 = \ldots = Z_{p-1} = 0 \\
X = X_3 & Z_1 = 0, Z_2 = 1, Z_3 = \ldots = Z_{p-1} = 0 \\
\ldots & \ldots \\
X = X_p & Z_1 = \ldots = Z_{p-2} = 0, Z_{p-1} = 1
\end{array}
$$

Множественная линейная регрессия

$$Y,\ X = (X_1, \dots, X_k):$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \dots, n$$

$$E\varepsilon_i = 0,\ i = 1, \dots, n$$

$$K(\varepsilon_l \varepsilon_u) = 0 \text{ при } l \neq u$$

$$D\varepsilon_i = \sigma_i^2,\ i = 1, \dots n$$

Матричное представление

$$Y = A\beta + \varepsilon,$$

где $Y = (y_1, \dots, y_n)^T$, $\beta = (\beta_0, \beta_1, \dots, \beta_k)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$,

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

— матрица порядка $n \times (k+1)$.
$A = (X_0, X_1, \dots, X_K)$, где $X_0 = (1, 1, \dots, 1)^T$.

МНК-оценки

$$(Y - A\beta)^T(Y - A\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2 \to \min_{\beta}$$

$$(A^T A)\hat{\beta} = A^T Y \quad \Rightarrow \quad \hat{\beta} = (A^T A)^{-1} A^T Y$$

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

$$\hat{\varepsilon} = (e_1, \ldots, e_n): \quad e_i = y_i - \hat{y}(x_i), \quad i = 1, \ldots, n$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$S^2 = \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - k - 1} = \frac{1}{n - k - 1}\hat{\varepsilon}^T\hat{\varepsilon}$$

Свойства МНК-оценок

$$E\hat{\beta} = (A^T A)^{-1} A^T EY = (A^T A)^{-1} A^T A\beta = \beta.$$

Теорема Гаусса-Маркова
Оценки метода наименьших квадратов $\hat{\beta}$ являются наилучшими линейными несмещенными оценками, т.е.

$$D\tilde{\beta}_i \geqslant D\hat{\beta}_i, \quad i = 0, 1, 2, \dots, k,$$

для любых несмещенных оценок $\tilde{\beta} = CY$.

$$D\hat{\beta} = E\{(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T\} = \sigma^2 (A^T A)^{-1}.$$

Распределения оценок

Пусть $\quad \varepsilon^T = (\varepsilon_1, \dots, \varepsilon_n) \sim N(0, \sigma^2 E_n)$

- $\dfrac{(n-k-1)S^2}{\sigma^2} \sim \chi^2(n-k-1)$
- $\hat{\beta} \sim N(\beta, \sigma^2 (A^T A)^{-1})$
- $(n-k-1)S^2/\sigma^2$ взаимно независима с вектором оценок $\hat{\beta}$
- $\dfrac{\widehat{\beta}_j - \beta_j}{S\sqrt{[(A^T A)^{-1}]_{(j+1)(j+1)}}} \sim T_{n-k-1}, \;\; j = 0, \dots, k$
- $\dfrac{\widehat{y}(x) - y(x)}{S\sqrt{x^T (A^T A)^{-1} x}} \sim T_{n-k-1}, \;\; \mathrm{x}{=}(1{,}\mathrm{x}_1, \dots, x_k)$

Доверительные интервалы

$\beta_j, \quad j = 0, 1, \dots, k$

$$\hat{\beta}_j \pm t_{1-\frac{\alpha}{2}}(n-k-1)S\sqrt{[(A^TA)^{-1}]_{(j+1)(j+1)}}$$

$y(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

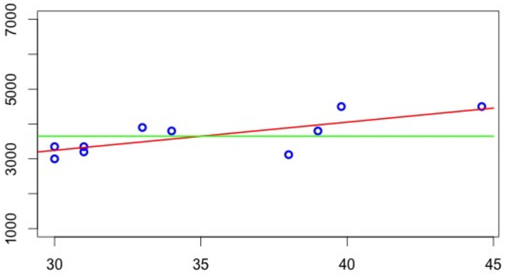
$$\hat{y}(x) \pm t_{1-\alpha/2}(n-k-1)S\sqrt{x^T(A^TA)^{-1}x}$$

σ^2

$$\left(\frac{S^2(n-k-1)}{\chi^2_{1-\alpha/2}(n-k-1)}, \quad \frac{S^2(n-k-1)}{\chi^2_{\alpha/2}(n-k-1)}\right)$$

Коэффициент детерминации

$$R^2 = \frac{\left(\sum\limits_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{y})\right)^2}{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2 \sum\limits_{i=1}^{n}(\hat{y}_i - \overline{y})^2} = 1 - \frac{RSS}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



Значимость модели

Пусть $\varepsilon^T = (\varepsilon_1, \ldots, \varepsilon_n) \sim N(0, \sigma^2 E_n)$

$$H_0\colon \beta_1 = \ldots = \beta_k = 0$$

$$H_1\colon \exists \beta_r \neq 0$$

$$F = \frac{R^2}{1 - R^2} \frac{n - k - 1}{k} \sim \mathcal{F}_{k, n-k-1}$$

H_0: $\beta_{k_1} = \ldots = \beta_{k_q} = 0,\ k_i \neq 0$

H_1: $\exists \beta_{k_i} \neq 0$

$$F = \frac{(RSS_{H_0} - RSS)/q}{RSS/(n-k-1)} \sim F(q, n-k-1)$$

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

$$t_{\beta_j} = \frac{\hat{\beta}_j}{S\sqrt{[(A^T A)^{-1}]_{(j+1)(j+1)}}} \sim T(n-k-1)$$

Информационные критерии Акаике и Шварца

Статистика критерия Акаике

$$AIC = 2k + n\left[\ln\frac{RSS}{n} + 1\right]$$

Статистика критерия Шварца

$$BIC = k\ln n + n\left[\ln\frac{RSS}{n} + 1\right]$$

Из двух моделей предпочтительно выбрать модель с меньшим значением статистики критерия Акаике или статистики критерия Шварца.