19. Свойства коэффициентов уравнения парной линейной регрессии.

Определение 7.13. *Регрессионный анализ* — статистический метод, предназначенный для определения **вида** (аналитического выражения) **связи** зависимой случайной величины y (называемой результативным признаком) от независимых случайных величин $X_1, X_2, \ldots, X_p$ (называемых факторами или объясняющими переменными).

Форма связи результативного признака Y с факторами $X = (X_1, \ldots, X_p)^T$ получила название **уравнения регрессии**:

$$y = f(X, \varepsilon),$$

где случайная составляющая ε характеризует, с одной стороны, влияние на y не входящих в X факторов, а с другой - погрешность в измерении показателя y.

Модель парной линейной регрессии имеет вид:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Индексы i показывают, что модель записана для **каждого наблюдения** в выборке.

где:

- y_i — **зависимая переменная** (отклик) для i-го наблюдения,
- x_i — **независимая переменная** (фактор, предиктор) для i-го наблюдения,
- ε_i — **случайная ошибка** для i-го наблюдения.
- β_0 — **свободный член** (константа, intercept),
- β_1 — **коэффициент наклона** (slope),

Смысл коэффициентов:

- β_0 (**intercept**) — это **ожидаемое значение** y, когда $x = 0$ (если такое значение имеет смысл в контексте задачи).
- β_1 (**slope**) — показывает, **насколько в среднем изменяется** y при увеличении x на одну единицу.

Так как мы оцениваем $\beta_0 = b_0$ и $\beta_1 = b_1$ стоит пояснить что за коэффициенты b_0 и b_1:

Коэффициент b_1 — угловой коэффициент регрессии, он показывает, на сколько единиц в среднем меняется переменная y при увеличении независимой переменной x на единицу.

Постоянная b_0 дает прогнозируемое значение зависимой переменной при $x = 0$. Это может иметь смысл в зависимости от того, как далеко находится $x = 0$ от выборочных значений x.

Пример:
Если y — зарплата (в тыс. руб.), x — стаж (в годах), и

$$y = 30 + 5x + \varepsilon,$$

то:

- $\beta_0 = 30$ — ожидаемая зарплата при нулевом стаже,
- $\beta_1 = 5$ — при увеличении стажа на 1 год зарплата в среднем растёт на 5 тыс. руб.

Свойства коэффициентов уравнения парной линейной регрессии

Рассматриваемая нами модель регрессии (7.14)

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i. \tag{1}$$

Её оценкой по имеющимся данным является выборочное уравнение регрессии (7.15)

$$\widehat{y_i} = b_0 + b_1 \cdot x_i. \tag{2}$$

МНК-оценки коэффициентов β_0 и β_1, найденные нами ранее:

$$b_0 = \overline{y} - b_1\overline{x}, \tag{3.a}$$

$$b_1 = \frac{\operatorname{cov}(x,y)}{D(x)}. \tag{3.b}$$

Рассмотрим выражение (3.b). Подставим в него вместо y его выражение из (1):

$$b_1 = \frac{\operatorname{cov}(x,y)}{D(x)} = \frac{\operatorname{cov}(x, \beta_0 + \beta_1 \cdot x + \varepsilon)}{D(x)} = \frac{\operatorname{cov}(x,\beta_0) + \beta_1\operatorname{cov}(x,x) + \operatorname{cov}(x,\varepsilon)}{D(x)} =$$
$$= \frac{0 + \beta_1 D(x) + \operatorname{cov}(x,\varepsilon)}{D(x)} = \beta_1 + \frac{\operatorname{cov}(x,\varepsilon)}{D(x)}.$$

Здесь учтено, что $\operatorname{cov}(x,\beta_0) = 0$, т.к. $\beta_0 = \text{const}$, и то, что $\operatorname{cov}(x,x) = D(x)$.

Рассмотрим выражение (3.b), подставив в него вместо b_1 полученное только что выражение, а вместо $\overline{y}$ — выражение $\frac{1}{n}\sum y_i$, где y_i определяется в (1):

$$b_0 = \overline{y} - b_1\overline{x} = \frac{1}{n}\sum(\beta_0 + \beta_1 \cdot x_i + \varepsilon_i) - \overline{x} \cdot \left[\beta_1 + \frac{\operatorname{cov}(x,\varepsilon)}{D(x)}\right] =$$
$$= \frac{1}{n}\sum \beta_0 + \beta_1\frac{1}{n}\sum x_i + \frac{1}{n}\sum \varepsilon_i - \overline{x}\beta_1 - \frac{\overline{x} \cdot \operatorname{cov}(x,\varepsilon)}{D(x)} =$$
$$= \frac{1}{n}n \cdot \beta_0 + \beta_1\overline{x} + \frac{1}{n}\sum \varepsilon_i - \overline{x}\beta_1 - \frac{\overline{x} \cdot \operatorname{cov}(x,\varepsilon)}{D(x)} = \beta_0 + \left[\frac{1}{n}\sum \varepsilon_i - \frac{\overline{x} \cdot \operatorname{cov}(x,\varepsilon)}{D(x)}\right].$$

Получили, что коэффициенты b_0 и b_1 можно выразить через «истинные» коэффициенты β_0, β_1 и значения независимой переменной x:

$$b_1 = \beta_1 + \frac{\operatorname{cov}(x,\varepsilon)}{D(x)}, \tag{4}$$

$$b_0 = \beta_0 + \left[\frac{1}{n}\sum \varepsilon_i - \frac{\overline{x} \cdot \operatorname{cov}(x,\varepsilon)}{D(x)}\right]. \tag{5}$$

Несмещённость и дисперсии оценок

Покажем, что оценки b_0 и b_1 коэффициентов β_0 и β_1 являются несмещенными и дисперсии их равны следующим величинам:

$$D(b_0) = \sigma_{b_0}^2 = \frac{\overline{x^2} \cdot \sigma_\varepsilon^2}{n \cdot D(x)}; \qquad D(b_1) = \sigma_{b_1}^2 = \frac{\sigma_\varepsilon^2}{n \cdot D(x)}$$

Предварительно оценим (учитывая, что $\overline{\varepsilon} = 0$)

$$\operatorname{cov}(x, \varepsilon) = \frac{1}{n}\sum_i (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}) = \frac{1}{n}\sum_i (x_i - \overline{x})\varepsilon_i - \frac{\overline{\varepsilon}}{n}\sum_i (x_i - \overline{x}) = \frac{1}{n}\sum_i (x_i - \overline{x})\varepsilon_i$$

Принимая предположение теоремы Гаусса-Маркова о том, что x - неслучайная (детерминированная) величина, оценим математическое ожидание и дисперсию коэффициентов, используя их представление (4) и (5). Из (4) оценим:

$$M[\operatorname{cov}(x, \varepsilon)] = M\left[\frac{1}{n}\sum_i (x_i - \overline{x})\varepsilon_i\right] = \frac{1}{n}\sum_i (x_i - \overline{x}) M[\varepsilon_i] = 0, \tag{6}$$

$$D[\operatorname{cov}(x, \varepsilon)] = D\left[\frac{1}{n}\sum_i (x_i - \overline{x})\varepsilon_i\right] = \frac{1}{n^2}\sum (x_i - \overline{x})^2 D(\varepsilon_i) = \frac{D(x)}{n}\sigma_\varepsilon^2. \tag{7}$$

Из (5) оценим:

$$M\left[\frac{1}{n}\sum \varepsilon_i\right] = \frac{\sum M(\varepsilon_i)}{n} = 0; \tag{8}$$

$$D\left[\frac{1}{n}\sum \varepsilon_i\right] = \frac{\sum D(\varepsilon_i)}{n^2} = \frac{\sigma_\varepsilon^2 \cdot n}{n^2} = \frac{\sigma_\varepsilon^2}{n}; \tag{9}$$

$$M\left[\frac{\overline{x}\ \operatorname{cov}(x, \varepsilon)}{D(x)}\right] = \frac{\overline{x}}{D(x)} M[\operatorname{cov}(x, \varepsilon)]_{(6)} = 0; \tag{10}$$

$$D\left[\frac{\overline{x}\ \operatorname{cov}(x, \varepsilon)}{D(x)}\right] = \frac{(\overline{x})^2}{D^2(x)} D[\operatorname{cov}(x, \varepsilon)]_{(7)} = \frac{(\overline{x})^2}{D^2(x)} \frac{D(x)}{n}\sigma_\varepsilon^2 = \frac{(\overline{x})^2 \sigma_\varepsilon^2}{nD(x)}. \tag{11}$$

Используя полученные оценки, найдем метематическое ожидание и дисперсию коэффициентов b_0 и b_1.

$$M(b_1) = M\left[\beta_1 + \frac{\operatorname{cov}(x, \varepsilon)}{D(x)}\right] = \beta_1 + \underbrace{\frac{M[\operatorname{cov}(x, \varepsilon)]}{D(x)}}_{(6)} = \beta_1$$

$$D(b_1) = M\left[\beta_1 + \frac{\operatorname{cov}(x, \varepsilon)}{D(x)}\right] = \underbrace{\frac{D[\operatorname{cov}(x, \varepsilon)]}{D^2(x)}}_{(7)} = \frac{D(x)}{n}\sigma_\varepsilon^2 \frac{1}{D^2(x)} = \frac{\sigma_\varepsilon^2}{n \cdot D(x)}$$

$$M(b_0) = M[\beta_0] + \underbrace{M\left[\frac{1}{n}\sum \varepsilon_i\right]}_{(8)} - \underbrace{M\left[\frac{\overline{x}\ \operatorname{cov}(x, \varepsilon)}{D(x)}\right]}_{(10)} = \beta_0;$$

$$D(b_0) = \underbrace{D\left[\frac{1}{n}\sum \varepsilon_i\right]}_{(9)} + \underbrace{D\left[\frac{\overline{x}\ \operatorname{cov}(x, \varepsilon)}{D(x)}\right]}_{(11)} = \frac{\sigma_\varepsilon^2}{n} + \frac{(\overline{x})^2 \sigma_\varepsilon^2}{nD(x)} = \frac{\sigma_\varepsilon^2}{n}\left[1 + \frac{(\overline{x})^2}{D(x)}\right] =$$

$$= \frac{\sigma_\varepsilon^2 \overline{x^2}}{nD(x)} \qquad \text{так как} \quad 1 + \frac{(\overline{x})^2}{D(x)} = \frac{D(x) + (\overline{x})^2}{D(x)} = \frac{(\overline{x^2} - (\overline{x})^2) + \overline{x})^2}{D(x)} = \frac{\overline{x^2}}{D(x)}$$

Оценкой дисперсии ошибки σ_ε^2 по выборке служит величина

$$S_e^2 = \frac{Q_e}{n-2} = \frac{\sum (y_i - \widehat{y}_i)^2}{n-2}.$$

Тогда для коэффициентов b_0 и b_1 выборочного уравнения регрессии имеем:

$$M(b_0) = \beta_0, \qquad M(b_1) = \beta_1.$$

Это означает, что найденные оценки являются несмещенными.

Дисперсии и средние квадратические отклонения коэффициентов b_0 и b_1:

$$D(b_0) = \frac{S_e^2 \overline{x^2}}{nD(x)}, \qquad S_{b_0} = \frac{S_e \sqrt{\overline{x^2}}}{\sqrt{n}\sigma_x},$$
$$D(b_1) = \frac{S_e^2}{nD(x)}, \qquad S_{b_1} = \frac{S_e}{\sqrt{n}\sigma_x}.$$

Статистические свойства (Что мы рассмотрели выше)

При использовании метода наименьших квадратов (МНК) для оценки коэффициентов, они обладают набором оптимальных свойств,:

- **Несмещенность**: Математическое ожидание оценок b_0 и b_1 совпадает с истинными значениями параметров в генеральной совокупности, что гарантирует **отсутствие систематической ошибки** при оценивании.
- **Эффективность**: Оценки МНК имеют **минимально возможную дисперсию** (разброс) среди всех линейных несмещенных оценок, что делает их наиболее надежными.
- **Состоятельность**: С увеличением объема выборки значения коэффициентов стремятся (по вероятности) к своим теоретическим значениям.

Важное замечание: Эти свойства выполняются только при соблюдении предпосылок регрессионного анализа, таких как **гомоскедастичность** (равенство дисперсий ошибок) и отсутствие их автокорреляции.