Лекция 10.

Свойство ковариации

Мет. Ковариацией случайных величин X и Y называется величина $\operatorname{cov}(X,Y) = E((X-EX)(Y-EY)) = E(XY) - EX \cdot EY$

Ковариация является индикатором наличия направления связи между двумя случайными величинами

Пусть имеется $(X_1, Y_1), \dots, (X_n, Y_n)$ случайных величин X и Y

Def. Выборочной ковариацией называется величина $\widehat{\operatorname{cov}}(X,Y) = \overline{xy} - \overline{x} \cdot \overline{y}$

По Закону Больших Чисел ясно, что $\widehat{\operatorname{cov}}(X,Y) \longrightarrow \operatorname{cov}(X,Y)$, поэтому выборочная ковариация является оценкой

Th. Выборочная ковариация является точечной состоятельной, но смещенной оценкой ковариации. Несмещенной оценкой будет $\dfrac{n}{n-1}\widehat{\operatorname{cov}}(X,Y)$

Ковариация и выборочная ковариация обладают свойствами

1. $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
2. $\operatorname{cov}(X,a) = 0$, где $a = \text{const}$
3. $\operatorname{cov}(X,bY) = b\operatorname{cov}(X,Y)$
4. $\operatorname{cov}(X+Y,Z) = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z)$
5. $\operatorname{cov}(X,X) = D(X), \widehat{\operatorname{cov}}(X,X) = D^*(X)$
6. $D(X+Y) = DX + DY + 2\operatorname{cov}(X,Y)$

Nota. В дальнейшем под $\operatorname{cov}(X,Y)$ будет пониматься выборочная ковариация

Анализ модели линейной парной регрессии

Пусть при n экспериментах получены значения случайных величин X и Z: $(X_1, Z_1), \dots, (X_n, Z_n)$

Пусть $X = \alpha + \beta Z + \varepsilon$ - теоретическая модель линейной регрессии, где ε - случайная величина, отражающая влияние невключенных факторов, нелинейность модели, ошибок измерений и просто случая.

Пусть построили с помощью метода наименьших квадратов выборочное уравнение линейной регрессии $\hat{X} = a + bZ$

Обозначим $\hat{\varepsilon}_i = X_i - \hat{X}_i$ - экспериментальная ошибка, разница между наблюдаемыми значениями и вычисляемыми по модели

Тогда $X_i = \hat{X}_i + \hat{\varepsilon}_i$ или $X_i = a + bZ_i + \hat{\varepsilon}_i$, где a и b - точечные оценки параметров α и β

Свойства $\hat{\varepsilon}_i$:

1. $\overline{\hat{\varepsilon}_i} = 0$

$$a = \overline{X} - b\overline{Z} \Longrightarrow a + b\overline{Z} = \overline{X} \Longrightarrow \overline{\hat{\varepsilon}_i} = \overline{X_i - (a + bZ_i)} = \overline{X} - \overline{a + bZ_i} = \overline{X} - \overline{X} = 0$$

2. $\operatorname{cov}(\hat{X}, \hat{\varepsilon}) = 0$

$$b = \overline{\overline{xz} - \overline{x} \cdot \overline{z}}\hat{\sigma}_z^2 = \overline{\operatorname{cov}(X, Z)}D(Z) \Longrightarrow \operatorname{cov}(X, Z) - bD(Z) = 0$$
$$\operatorname{cov}(\hat{X}, \hat{\varepsilon}) = \operatorname{cov}(a + bZ, X - a - bZ) = \operatorname{cov}(bZ, X - bZ) = \operatorname{cov}(bZ, x) - \operatorname{cov}(bZ, bZ) = b \operatorname{cov}(Z, X) - b^2 D(Z) = b(\operatorname{cov}(Z, X) - bD(Z)) = 0$$

Анализ дисперсии результата

Def. $D(X) = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ - дисперсия наблюдаемых значений

Def. $\hat{D}(X) = \dfrac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - \overline{X})^2$ - дисперсия расчетных значений

Def. $D(\hat{\varepsilon}) = \dfrac{1}{n}\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2$ - дисперсия остатков

Так как $X = \hat{X} + \hat{\varepsilon}$, $\operatorname{cov}(\hat{X}, \hat{\varepsilon}) = 0$, то $D(X) = D(\hat{X}) + D(\hat{\varepsilon}) + 2\operatorname{cov}(\hat{X}, \hat{\varepsilon}) = D(\hat{X}) + D(\hat{\varepsilon})$

Th. $D(X) = D(\hat{X}) + D(\hat{\varepsilon})$

Очевидно, что качество модели будет тем лучше, чем меньше будет дисперсия остатков

Def. Коэффициентом детерминации R^2 называется величина $R^2 = \dfrac{D(\hat{X})}{D(X)}$ или $R^2 = 1 - \dfrac{D(\hat{\varepsilon})}{D(X)}$

Nota. Смысл R^2 - доля объясненной дисперсии, а $1 - R^2$ - доля необъясненной дисперсии

Свойства:

1. $0 \le R^2 \le 1$
2. Если $R^2 = 1$, то $D(\hat{\varepsilon}) = 0 \Longrightarrow \hat{\varepsilon}_i = \overline{\hat{\varepsilon}_i} = 0$, то есть точки лежат строго на линии регрессии, модель идеальна
3. Если $R^2 = 0$, то $D(\hat{X}) = 0 \Longrightarrow \hat{X} = \overline{x}$, то есть получаем примитивную, ничего не объясняющую модель

Чем больше R^2, тем лучше качество модели

Проверка гипотезы о значимости уравнения регрессии

Проверяется $H_0 : R^2_{\text{теор}} = 0$ (уравнение регрессии статистически не значимо) против $H_1 : R^2_{\text{теор}} \neq 0$

Th. Если H_0 верна, то $F = \dfrac{R^2(n-2)}{1-R^2} \in F(1, n-2)$

Пусть t_α - квантиль $F(1, n-2)$ уровня α, тогда:

$$\begin{cases} H_0 : R^2_{\text{теор}} = 0 & \text{если } F < t_\alpha \\ H_0 : R^2_{\text{теор}} \neq 0 & \text{если } F \geq t_\alpha \end{cases}$$

Nota. Если $H_0 : R^2_{\text{теор}} = 0$, то $H_0 : \beta = 0$

Связь между коэффициентом детерминации и коэффициентом линейной корреляции

1. $\sqrt{R^2} = r_{\hat{X},X}$ - коэффициент корреляции между $\hat{X}$ и X

$$r_{\hat{X},X} = \frac{\operatorname{cov}(\hat{X}, X)}{\sqrt{D(\hat{X})D(X)}} = \frac{\operatorname{cov}(\hat{X}, \hat{X} + \hat{\varepsilon})}{\sqrt{D(\hat{X})D(X)}} = \frac{D(\hat{X}) + \cancelto{0}{\operatorname{cov}(\hat{X}, \hat{\varepsilon})}}{\sqrt{D(\hat{X})D(X)}} = \sqrt{\frac{D(\hat{X})}{D(X)}} = R$$

2. $r_{\hat{X},X} = |r_{X,Z}|$

$$\operatorname{cov}(\hat{X}, X) = \operatorname{cov}(a + bZ, X) = b\operatorname{cov}(Z, X)$$
$$D(\hat{X}) = D(a + bZ) = b^2 D(Z)$$
$$r_{\hat{X},X} = \frac{\operatorname{cov}(\hat{X}, X)}{\sqrt{D(\hat{X})D(X)}} = \frac{b\operatorname{cov}(Z, X)}{\sqrt{b^2 D(Z)D(X)}} = \left|\frac{\operatorname{cov}(X, Z)}{\sqrt{D(Z)D(X)}}\right| = |r_{X,Z}|$$

Следствие 1: в случае линейной парной регрессии коэффициент детерминации равен квадрату коэффициенту корреляции

Следствие 2: в случае линейной парной регрессии совпадают результаты проверки гипотез $H_0 : R^2_{\text{теор}} = 0 \iff H_0 : r = 0 \iff H_0 : \beta = 0$

Теорема Гаусса-Маркова

Th. Пусть $X_i = \alpha + \beta Z_i + \varepsilon_i$ - теоретическая модель регрессии

$X = a + bZ$ - модель, полученная по методу наименьших квадратов

Если выполнено условия:

а) Случайные члены ε_i независимые случайные величины, имеющие одинаковое нормальное распределение $\varepsilon_i \in N(0, \sigma^2)$

б) Случайные величины ε_i и Z_i - независимы

Тогда a и b - состоятельные, несмещенные, эффективные оценки параметров α и β, то есть

1. Состоятельность: $a \xrightarrow[n\to\infty]{p} \alpha, b \xrightarrow[n\to\infty]{p} \beta$
2. Несмещенность: $Ea = \alpha, Eb = \beta$
3. Наименьшая дисперсия, равная:
   $Da = \dfrac{\overline{z^2}\sigma^2}{nD(Z)}, Db = \dfrac{\sigma^2}{nD(Z)}$

Nota. Если не выполняется условие а), то есть ошибки зависимы или имеют разную дисперсию, то оценки становятся неэффективными. Если не выполнено условие б), то оценки становятся смещенными и несостоятельными

Стандартные ошибки коэффициентов регрессии

Из теоремы видим, что Da и Db зависят от дисперсии σ^2 случайного члена. По экспериментальным ошибкам получаем оценку данной дисперсии:

$D(\hat{\varepsilon}) = \dfrac{1}{n}\sum\limits_{i=1}^{n} \hat{\varepsilon}_i^2 \xrightarrow[n\to\infty]{p} \sigma^2$

Однако эта оценка является смещенной:

$E(D(\hat{\varepsilon})) = \dfrac{n-2}{n}\sigma^2$

Поэтому несмещенной оценкой дисперсии σ^2 является величина $S^2 = \dfrac{1}{n-2}\sum\limits_{i=1}^{n} \hat{\varepsilon}_i^2$

Def. Величина S называется стандартной ошибкой регрессии

Смысл: характеризует разброс наблюдаемых значений вокруг линии регрессии

Nota. Заменим в теореме Гаусса-Маркова σ^2 на S^2, получаем оценки дисперсий Da и Db:

$S_a^2 = \dfrac{\overline{z^2}S^2}{nD(z)}, S_b^2 = \dfrac{S^2}{nD(Z)}$

Def. S_a и S_b называются стандартным ошибками коэффициентов регрессии

Прогнозирование регрессионных моделей

Пусть $X = \alpha + \beta Z + \varepsilon$ - теоретическая модель

$\hat{X} = a + bZ$ - модель МНК, построенная по выборке объема n

С помощью данной модели надо дать прогноз значения X_p при заданном значении Z_p и оценить качество прогноза

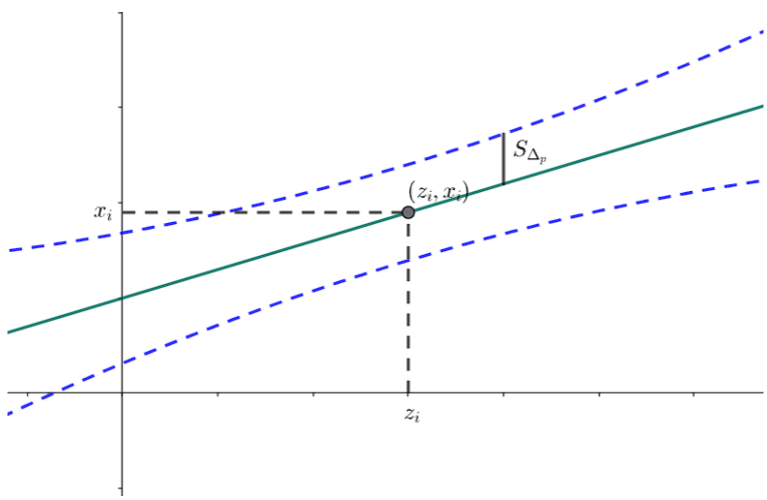
Теоретическое значение - $X_p = \alpha + \beta Z_p + \varepsilon$, а точечный прогноз $\hat{X}_p = a + bZ_p$

Разность между ними $\Delta_p = \hat{X}_p - X_p$ называется ошибкой предсказания

Свойства Δ_p:

1. $E\Delta_p = 0$
2. $D(\Delta_p) = \left(1 + \frac{1}{n} + \frac{(Z_p - \overline{z})^2}{nDZ}\right)\sigma^2$

   Заменив σ^2 на S^2, получим стандартную ошибку прогноза: $S_{\Delta_p} = S\sqrt{1 + \frac{1}{n} + \frac{(Z_p - \overline{z})^2}{nDZ}}$
3. $D(\Delta_p) > \sigma^2$ - то есть точность прогноза ограничена случайным членом ε
4. При $n \to \infty$ $D(\Delta_p) \xrightarrow{p} \sigma^2$ - качество модели тем лучше, чем больше объем выборки
5. Чем больше Z_p отклоняется от $\overline{z}$, тем хуже качество прогноза. Наилучшее качество в точке $Z_p = \overline{z}$: $D(\Delta_p) = \left(1 + \frac{1}{n}\right)\sigma^2$



Доверительные интервалы прогноза и коэффициентов уравнения линейной регрессии

Пусть t_γ - квантиль $|T_{n-2}|$ уровня γ

Тогда доверительные интервалы надежности γ для параметров α и β:

$\alpha : (a - t_\gamma S_a; a + t_\gamma S_a)$

$\beta : (b - t_\gamma S_b; b + t_\gamma S_b)$

Доверительный интервал для прогноза $X_p : (\hat{X}_p - t_\gamma S_{\Delta_p}; \hat{X}_p + t_\gamma S_{\Delta_p})$