Лекция 9.

Исследование статистической корреляции

Математическая модель регрессии

Пусть случайная величина X зависит от случайной величины Z (необязательно случайной)

Def. Регрессией X на Z называется функция $f(z) = E(X|Z = z)$. Она показывает зависимость среднего значения X от значения Z

Уравнение $x = f(z)$ называется уравнением регрессии, а график этой функции - линия регрессии

Пусть при n экспериментах при значениях $Z_1, Z_2, \ldots, Z_n$ фактора Z наблюдались значения $X_1, X_2, \ldots, X_n$ случайной величины X

Обозначим через ε_i разницу между экспериментальным и теоретическими значениями случайной величины X, то есть $\varepsilon_i = X_i - f(z_i)$

ε - это случайный член модели или так называемая теоретическая ошибка

Nota. Обычно можно считать, что ε_i независимы друг от друга и имеет нормальное распределение с $a = 0$, так как $E\varepsilon_i = E(X_i - f(Z_i)) = E(X|Z = Z_i) - E(X|Z = Z_i) = 0$

Цель: нам нужно по экспериментальным данным $(z_1, x_1), \ldots, (z_n, x_n)$ как можно лучше оценить функцию $f(z)$

Nota. При этом предполагая (часто из теории), что $f(z)$ - функция определенного вида, но параметры которой неизвестны. Если нет, то начинаем подбирать модели самого простого вида. В противном случае, наилучшим решением была бы кривая, проходящая через все точки

Метод наименьших квадратов

Пусть известен из теории вид функции $f(z)$. Метод наименьших квадратов состоит в выборе параметров $f(z)$ таким образом, чтобы минимизировать сумму квадратов ошибок $\sum\limits_{i=1}^{n} \varepsilon_i^2 = \sum\limits_{i=1}^{n} (X_i - f(Z_i))^2 \to \min$

Def. Пусть θ - набор неизвестных параметров функции $f(z)$. Оценка $\hat{\theta}$ параметра θ, при которой достигается минимум $\sum\limits_{i=1}^{n} \varepsilon_i^2$, называется оценкой метода наименьших квадратов (или ОМНК)

Линейная парная регрессия

Пусть имеется теоретическая модель линейной регрессии

$f(z) = \alpha + \beta z + \varepsilon$ - теоретическая модель, где ε - теоретическая ошибка отражающая влияние невключенных в модель факторов, возможной нелинейности, ошибок измерения и просто случая

Пусть $(z_1, x_1), \ldots, (z_n, x_n)$ - экспериментальные данные. По ним методом наименьших квадратов строим экспериментальную модель линейной регрессии $f(z) = a + bz$, где a и b - ОМНК параметров α и β

$\hat{\varepsilon}_i = X_i - f(Z_i) = X_i - (a + bZ_i)$ - экспериментальная ошибка

Найдем ОМНК параметров α и β

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (X_i - (a + bZ_i))$$

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} -2(X_i - a - bZ_i) = -2\sum_{i=1}^{n} X_i + 2\sum_{i=1}^{n} a + 2b\sum_{i=1}^{n} Z_i = -2(n\overline{x} - na - bn\overline{z})$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} -2Z_i(X_i - a - bZ_i) = -2\sum_{i=1}^{n} X_i Z_i + 2\sum_{i=1}^{n} aZ_i + 2b\sum_{i=1}^{n} Z_i^2 = -2(n\overline{xz} - an\overline{z} - bn\overline{z^2})$$

$$\begin{cases} -2(n\overline{x} - na - nb\overline{z}) = 0 \\ -2(n\overline{zx} - na\overline{z} - nb\overline{z^2}) = 0 \end{cases} \iff \begin{cases} a + b\overline{z} = \overline{x} \\ a\overline{z} + b\overline{z^2} = \overline{zx} \end{cases}$$

Получили систему линейных уравнений. Будем называть ее нормальной системой. При решении получаем:

$$\begin{cases} a = \overline{x} - b\overline{z} \\ (\overline{x} b\overline{z})\overline{z} + b\overline{z^2} = \overline{zx} \end{cases} \iff \begin{cases} a = \overline{x} - b\overline{z} \\ b = \frac{\overline{zx} - \overline{z}\,\overline{x}}{\hat{\sigma}_z^2} \end{cases} \text{ - ОМНК}$$

Запишем уравнение линейной регрессии в удобном виде: $\overline{x}_z = f(z) = E(X|Z = z)$

$\overline{x}_z = a + bz$

$\overline{x}_z = \overline{x} - b\overline{z} + bz$

$$\overline{x}_z - \overline{x} = \frac{\overline{zx} - \overline{x}\,\overline{z}}{\hat{\sigma}_z^2}(z - \overline{z})$$

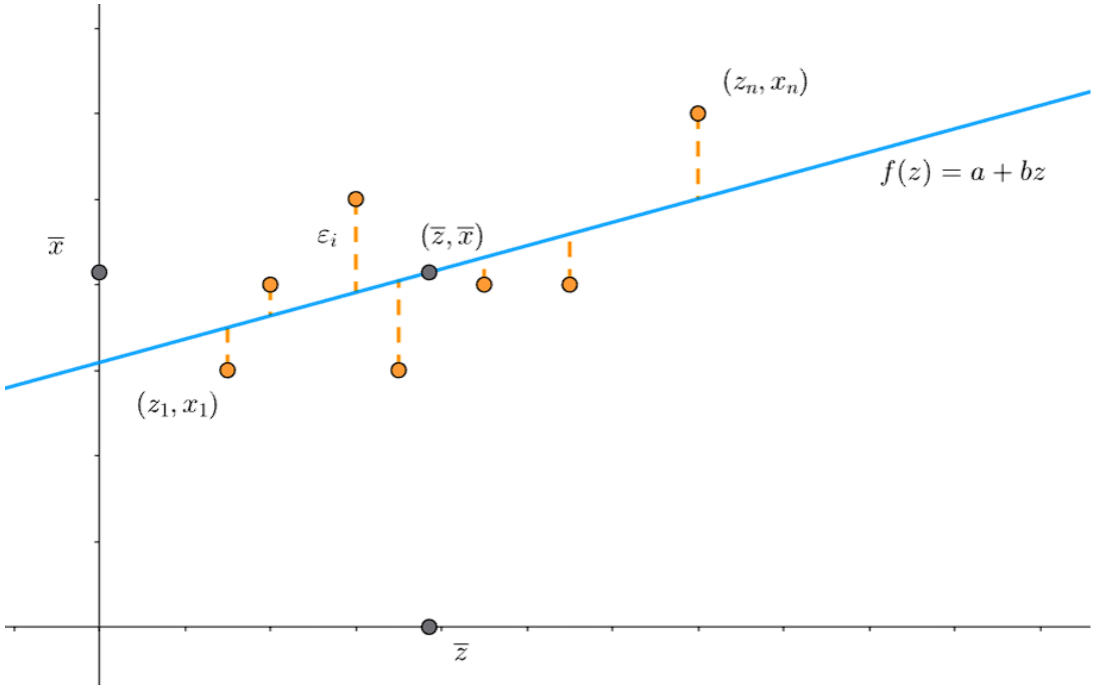
$\overline{x}_z - \overline{x} = \dfrac{\hat{\sigma}_x}{\hat{\sigma}_z}\dfrac{\overline{zx} - \overline{x}\,\overline{z}}{\hat{\sigma}_z\hat{\sigma}_x}(z - \overline{z}) = \dfrac{\hat{\sigma}_x}{\hat{\sigma}_z}\hat{r}(z - \overline{z})$, где $\hat{r}$ - выборочный коэффициент линейной корреляции

Или $\dfrac{\overline{x}_z - \overline{x}}{\hat{\sigma}_x} = \hat{r}\dfrac{z - \overline{z}}{\hat{\sigma}_z}$ - выборочное уравнение линейной регрессии

Nota. Прямая регрессии проходит через точку из выборочных средних

Nota. При $n \to \infty$ $\overline{x} \longrightarrow EX, \overline{z} \longrightarrow EZ, \hat{\sigma}_x \longrightarrow \sigma_x, \hat{\sigma}_z \longrightarrow \sigma_z, \overline{x}_z \longrightarrow E(X|Z = z), \hat{r} \longrightarrow r$, получаем $\dfrac{E(X|Z = z) - EX}{\sigma_x} = r\dfrac{z - EZ}{\sigma_z}$ - теоретическое уравнение линейной регрессии

Геометрический смысл линии регрессии



Суть МНК: находим такую прямую, чтобы сумма квадратов длин этих отрезков (по сути отклонений) была минимальна (или дисперсия экспериментальных данных относительно прямой была минимальна)

Выборочный коэффициент линейной корреляции

Def. $\hat{r} = \dfrac{\overline{zx} - \overline{x}\,\overline{z}}{\hat{\sigma}_z \hat{\sigma}_x}$ называется выборочным коэффициентов линейной корреляции. Ясно, что она будет точечной оценкой теоретического коэффициента линейной корреляции. Также $\hat{r}$ является несмещенной оценкой

Поэтому выборочный коэффициент корреляции характеризует силу линейной связи. Знак коэффициента показывает направления корреляции (прямая или обратная)

Силу связи можно примерно оценить от шкале Чеддока:

Количественная мера $\hat{r}$	Качественная мера
0.1 – 0.3	Слабая
0.3 – 0.5	Умеренная
0.5 – 0.7	Заметная
0.7 – 0.9	Высокая
> 0.9	Весьма высокая

Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть (Z, X) распределена нормально. По выборке объема n вычислен выборочный коэффициент корреляции $\hat{r}$, а r - теоретический коэффициент корреляции

Проверяется $H_0 : r = 0$ (выборочный коэффициент корреляции статистически незначим) против $H_1 : r \neq 0$ (коэффициент статистически значим)

Если H_0 верна, то $K = \dfrac{\hat{r}\sqrt{n-2}}{\sqrt{n-\hat{r}^2}} \in T_{n-2}$ - распределение Стьюдента с степенью $n-2$

Получаем критерий. Пусть t_α - квантиль $|T_{n-2}|$ (двухстороннее распределение Стьюдента) уровня α

$$\begin{cases} H_0 : r = 0, & \text{если } |K| < t_\alpha \\ H_1 : r \neq 0, & \text{если } |K| \geq t_\alpha \end{cases}$$

Надо понимать, что корреляция - более тонкое понятие, чем зависимость

А термин *регрессия* получил свое название чисто исторически: статистик Гальтон в 1886 году исследовал зависимость роста детей от роста родителей

$E(\text{Р}_{\text{сына}} | Z_{\text{отца}} = Z_1, Z_{\text{матери}} = Z_1) = 0.27Z_1 + 0.2Z_2 + \text{const}$

$E(\text{Р}_{\text{дочери}} | Z_{\text{отца}} = Z_1, Z_{\text{матери}} = Z_1) = \dfrac{1}{1.08}\text{Р}_{\text{сына}}$

Дальше он заметил, что при у самых высоких родителей рост детей был меньше относительно них (скатывался к среднему, происходил регресс)

Позже исследовали экономические результаты фирм, показатели спортсменов, которые после успешного сезона уменьшались, после чего появлялось куча теорий. Сейчас все это объясняется простым случаем

Выборочное корреляционное отношение

Выборочный коэффициент корреляции характеризует только силу линейной связи. Следующий подход основан на однофакторном дисперсионном анализе

Пусть есть k выборок случайной величины X при k различных уровнях фактора Z. Вычислены общая, внутригрупповая и межгрупповая дисперсии. По теореме $D_{\text{O}} = D_{\text{M}} + D_{\text{B}}$

Def. Выборочным корреляционным отношением X на Z называется величина $\eta_{X,Z} = \sqrt{\dfrac{D_{\text{M}}}{D_{\text{O}}}}$

Свойства:

1. $0 \leq \eta_{X,Z} \leq 1$ $(D_{\text{M}}, D_{\text{O}} \geq 0)$
2. Если $\eta = 1$, то $D_{\text{M}} = D_{\text{O}} \Longrightarrow D_{\text{B}} = 0$, имеем функциональную зависимость X от Z
3. Если $\eta = 0$, то $D_{\text{M}} = 0 \Longrightarrow$ корреляция отсутствует

4. $\eta \geq |\hat{r}|$
5. Если $\eta = |\hat{r}|$, то все точки экспериментальных данных лежат на прямой линейной регрессии (то есть данная линейная модель является идеальной)