Процедура використання $\hat{\tau}_{ij}^{(K)}$:

1) якщо $\hat{\tau}_{ij}^{(K)} = 0$, то вважаємо, що то зв'язок між змінними ξ_i та ξ_j відсутній;
2) якщо $\left|\hat{\tau}_{ij}^{(K)}\right| = 1$, то зв'язок між змінними ξ_i та ξ_j функціональний причому:
   - якщо $\hat{\tau}_{ij}^{(K)} = 1$, то ранжування $x^{(i)}$ та $x^{(j)}$ рівні;
   - якщо $\hat{\tau}_{ij}^{(K)} = -1$, то ранжування $x^{(i)}$ та $x^{(j)}$ протилежні;
3) якщо $\left|\hat{\tau}_{ij}^{(K)}\right| \in (0,1)$, то потрібно звернутися до *перевірки на значимість* рангового коефіцієнта кореляції Кендела, тобто перевірити гіпотезу

$$H_0 : \tau_{ij}^{(K)} = 0$$

з деяким рівнем значущості $\alpha > 0$. Зв'язок між ординальними змінними ξ_i та ξ_j зі статистичної точки зору будемо вважати суттєвим при відхиленні гіпотези H_0, а в протилежному разі - не істотним.

При $n = \overline{4,10}$, таку перевірку можна здійснити за допомогою спеціальних таблиць.

А при $n \geq 11$, з'ясувалося, що розподіл статистики

$$3\hat{\tau}_{ij}^{(K)} \sqrt{\frac{n(n-1)}{2(2n+5)}}$$

може бути наближений стандартним нормальним розподілом, тоді враховуючи, що до критичної області H_0 потрібно віднести області екстремальних значень останньої статистики, то область прийняття гіпотези буде мати таке представлення:

$$\left|3\hat{\tau}_{ij}^{(K)} \sqrt{\frac{n(n-1)}{2(2n+5)}}\right| < u_{\frac{\alpha}{2}},$$

де u_β – 100β відсоткова точка нормального розподілу з параметрами 0 і 1.

II випадок. За наявності груп об'єктів з однаковим проявом принаймні за однією зі змінних ξ_i чи ξ_j, ранговий коефіцієнт кореляції Кендела $\hat{\tau}_{ij}^{(K)}$ корегується й використовується у вигляді ***вибіркового модифікованого рангового коефіцієнта кореляції Кендела***:

$$\hat{\hat{\tau}}_{ij}^{(K)} = \frac{\hat{\tau}_{ij}^{(K)} - \left(\delta^{(i)} + \delta^{(j)}\right)}{\sqrt{\left(1-\delta^{(i)}\right)\left(1-\delta^{(j)}\right)}},$$

де $m^{(k)}$ – кількість груп об'єктів з однаковим проявом за змінною ξ_k; $n_l^{(k)}$ – кількість об'єктів, які ввійшли в l-ту групу об'єктів з однаковим проявом за змінною ξ_k;

$$\delta^{(k)} = \frac{\frac{1}{4}\sum_{l=1}^{m^{(k)}} n_l^{(k)}\left(n_l^{(k)}-1\right)}{\frac{1}{4}n(n-1)};\ \ k = i, j.$$

Зауваження 1. Для рангового коефіцієнта кореляції Кендела, коли відсутні групи об'єктів з однаковим проявом за змінними ξ_i та ξ_j, корегуючі величини $\delta^{(i)} = \delta^{(j)} = 0$, оскільки $m^{(i)} = m^{(j)} = n$, а $n_l^{(i)} = n_l^{(j)} = 1,\ l = \overline{1,n}$. У цій ситуації модифікований коефіцієнт $\hat{\hat{\tau}}_{ij}^{(K)} = \hat{\tau}_{ij}^{(K)}$.

У підсумку, для аналізу парного статистичного зв'язку ординальних змінних маємо ще одну характеристику – ранговий коефіцієнт кореляції Кендела.

Зауваження 2. Рангові коефіцієнти кореляції Спірмена та Кендела зв'язані між собою. Так, коли їх модулі не дуже близькі до одиниці та n вже велике, для них має місце таке наближення $\hat{\tau}_{ij}^{(S)} \approx \frac{3}{2}\hat{\tau}_{ij}^{(K)}$.

Аналіз множинних зв'язків для ординальних змінних. Коефіцієнт конкордації.

Нехай серед усіх ординальних змінних $\eta,\ \xi_1,\ \xi_2,\ \ldots,\xi_q\ (\xi_0 \equiv \eta)$ відібрано m змінних: $\xi_{i_1},\xi_{i_2},\ \ldots,\xi_{i_m},\ 2 \le m \le q+1$. Перейдемо до аналізу наявності статистичного зв'язку між цими вибраними ординальними змінними $\xi_{i_1},\xi_{i_2},\ldots,\xi_{i_m}$, причому нехай для них доступні відповідні ранжування

$$x^{(i_j)} = \left(x_1^{(i_j)},x_2^{(i_j)},\ldots,x_n^{(i_j)}\right)^T,\ j=\overline{1,m},\ 2 \le m \le q+1,$$

де n – кількість об'єктів, що досліджуються.

Для розв'язання цієї проблеми М. Кендел запропонував використовувати спеціальний показник.

I випадок. Розглянемо спочатку випадок, коли відсутні групи об'єктів з однаковим проявом у змінних $\xi_{i_1},\xi_{i_2},\ldots,\xi_{i_m},\ 2 \le m \le q+1$.

Введемо позначення $\vec{\zeta} = \left(\xi_{i_1},\xi_{i_2},\ldots,\xi_{i_m}\right)^T$.

Означення. ***Вибірковим коефіцієнтом конкордації ординальних змінних $\xi_{i_1},\xi_{i_2},\ldots,\xi_{i_m}$ з ранжуваннями $x^{(i_1)},x^{(i_2)},\ldots,x^{(i_m)}$*** називається числова характеристика, що задається таким чином:

$$\hat{W}_{\vec{\zeta}} = \frac{\sum_{k=1}^{n}\left(\sum_{j=1}^{m}\left(x_k^{(i_j)}\right)-\frac{m(n+1)}{2}\right)^2}{\frac{m^2\left(n^3-n\right)}{12}},$$

де $x^{(i_j)} = \left(x_1^{(i_j)},x_2^{(i_j)},\ldots,x_n^{(i_j)}\right)^T,\ j=\overline{1,m},\ 2 \le m \le q+1$.

Для коефіцієнта конкордації мають місце такі **властивості**:

1) $0 \le \hat{W}_{\bar{\xi}} \le 1$;

2) якщо $\hat{W}_{\bar{\xi}} = 0$, то зв'язок між змінними $\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_m}$ відсутній;

3) якщо $\hat{W}_{\bar{\xi}} = 1$, то зв'язок функціональний, а саме ранжування $x^{(i_1)}, x^{(i_2)}, \dots, x^{(i_m)}$ будуть рівні;

4) зміна порядку розташування змінних у послідовності $\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_m}$ не змінює значення коефіцієнта конкордації.

У загальному випадку, коли кількість об'єктів n не є постійною, для вибіркового значення коефіцієнта конкордації, яке вже може залежати від n, будемо використовувати позначення $\hat{W}_{\bar{\xi}}(n)$.

В свою чергу введемо поняття.

Означення. ***Теоретичним коефіцієнтом конкордації ординальних змінних*** $\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_m}$ називається числова характеристика $W_{\bar{\xi}}$, яка визначається таким чином:

$$W_{\bar{\xi}} = \lim_{n \to \infty} \hat{W}_{\bar{\xi}}(n).$$

Процедура використання $W_{\bar{\xi}}$:

1) якщо $\hat{W}_{\bar{\xi}} = 0$, то зв'язок між змінними $\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_m}$ відсутній;

2) якщо $\hat{W}_{\bar{\xi}} = 1$, то зв'язок функціональний, а саме ранжування $x^{(i_1)}, x^{(i_2)}, \dots, x^{(i_m)}$ будуть рівні;

3) якщо $\hat{W}_{\bar{\xi}} \in (0,1)$, то потрібно з'ясувати, чи суттєво відхиляється від нуля коефіцієнт конкордації зі статистичної точки зору, тобто здійснити *перевірку* його *на значимість*, а саме перевірити гіпотезу

$$H_0 : W_{\bar{\xi}} = 0$$

з деяким рівнем значущості $\alpha > 0$. Якщо гіпотеза буде відхилена, то вважається, що зв'язок між ординальними змінними $\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_m}$ зі статистичної точки зору є істотним, у протилежному випадку – не суттєвим.

При $n=\overline{3,7}$, $m=\overline{2,20}$, перевірка гіпотези H_0 здійснюється за допомогою спеціальних таблиць.

А при $n \geq 8$, можна скористатися тим фактом, що розподіл статистики $m(n-1)\hat{W}_{\bar{\xi}}$ можна наблизити χ^2-розподілом з $(n-1)$ ступенями свободи за умови справедливості гіпотези H_0. У результаті область прийняття гіпотези набуває вигляду:

$$m(n-1)\hat{W}_{\bar{\xi}} < \chi^2_\alpha(n-1),$$

де $\chi^2_\alpha(n)$ – 100α %-ва точка χ^2-розподілу з n ступенями свободи.

II випадок. Додаткової уваги потребує випадок наявності груп об'єктів з однаковим проявом принаймні за однією зі змінних $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$, $2 \leq m \leq q+1$. Після відповідної корекції, коефіцієнт конкордації $\hat{W}_{\bar{\xi}}$, можна використовувати у вигляді такого ***вибіркового модифікованого коефіцієнта конкордації***:

$$\hat{\hat{W}}_{\bar{\xi}} = \frac{\sum\limits_{k=1}^{n}\left(\sum\limits_{j=1}^{m}\left(x_k^{(i_j)}\right) - \frac{m(n+1)}{2}\right)^2}{\frac{m^2\left(n^3-n\right)}{12} - \frac{m}{2}\sum\limits_{j=1}^{m}\Delta^{(i_j)}},$$

де $m^{(i_j)}$ – кількість груп об'єктів з однаковим проявом за змінною ξ_{i_j};

$n_l^{(i_j)}$ – кількість об'єктів, які ввійшли в l-ту групу об'єктів з однаковим проявом за змінною ξ_{i_j};

$$\Delta^{(i_j)} = \frac{1}{6}\sum_{l=1}^{m^{(i_j)}}\left(\left(n_l^{(i_j)}\right)^3 - n_l^{(i_j)}\right);\ j=\overline{1,m};\ 2 \leq m \leq q+1.$$

Кореляційний аналіз номінальних змінних

...

Самостійна робота №5. З навчального посібника «Слабоспицький О.С. Основи кореляційного аналізу даних, 2006». Пропрацювати матеріал наведений у **Розділі 3.**: «Дослідження наявності зв'язку між номінальними змінними.»

(пропустити мінімум 4 стор.)

■ №7(ЗПМ)(2020.10.08). буде перед знаходженням оцінки ЗМНК у Р.А.

Регресійний аналіз

Регресійний аналіз один з розділів аналізу даних, який займається побудовою математичних моделей істотних зв'язків між кількісними змінними.

Нехай аналізується зв'язок між ***залежною*** скалярною змінною η та вектором ***незалежних*** змінних $\vec{\xi}$:

$$\eta \in \mathbb{R}, \quad \vec{\xi} \in \mathbb{R}^q.$$

Якщо кореляційний аналіз встановив, що статистичний зв'язок між цими змінними є істотний, то математичну модель залежності η від $\vec{\xi}$ можна шукати у вигляді ***регресійної моделі*** η ***щодо*** $\vec{\xi}$:

$$\eta = f\left(\vec{\xi}\right) + \varepsilon,$$

де $f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x})$ – функція регресії η щодо $\vec{\xi}$, ε – залишкова похибка апроксимації, причому $f(\vec{\xi})$ буде найкращою у середньоквадратичному розумінні апроксимацією η на класі борелівських функцій на $\mathbb{R}^q$.

Здається, що залишається тільки знайти вигляд функції $f(\vec{x})$, але для цього потрібно знати відповідний розподіл для η та $\vec{\xi}$. На практиці цей розподіл фактично є невідомим. І, як правило, єдиною доступною інформацією є тільки спостереження над цими змінними:

$$\eta : y(1), y(2), \ldots, y(N),$$
$$\vec{\xi} : \vec{x}'(1), \vec{x}'(2), \ldots, \vec{x}'(n).$$

Тобто потрібно спираючись тільки на ці спостереження знайти апроксимацію для $f(\vec{x})$.

Основні етапи розв'язання задачі **регресійного аналізу**:

1. вибір класу апроксимуючих функцій $\tilde{F}$ для $f(\vec{x})$, тобто для функції регресії η щодо $\vec{\xi}$:
$$\tilde{f}(\vec{x}, \alpha) \in \tilde{F}, \quad \vec{x} \in \mathbb{R}^q, \alpha \in \mathbb{R}^p,$$
де α - вектор невідомих параметрів;
2. визначення
оптимальної, згідно деякого критерію якості, точкової оцінки $\hat{\alpha}$ для α та її характеристики розсіювання $M(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)^T$, або множинної оцінки для α, тобто довірчої області для α;
3. якщо модель лінійна по α, то здійснюємо перевірку на значимість параметри моделі, а саме перевіряємо гіпотези:
$$H_0 : \alpha = \theta, \gamma > 0, \quad \text{або}$$
$$H_0 : \alpha_i = 0, \gamma > 0.$$
4. перевірка на адекватність отриманої моделі.

Зауваження. В якості критерію якості на другому етапі при обчисленні точкової оцінки $\hat{\alpha}$ для α найчастіше використовують такий функціонал:

$$M\left(\eta - \tilde{f}\left(\vec{\xi}, \alpha\right)\right)^2.$$

Вибіркове представлення його має такий вигляд:

$$\frac{1}{N}\sum_{k=1}^{n}\left(y(k) - \tilde{f}\left(\vec{x}'(k), \alpha\right)\right)^2. \qquad (1.1)$$

Саме його і буде використано у подальшому.

Класичний регресійний аналіз

Нехай потрібно побудувати математичну модель зв'язку залежної скалярної змінної η від вектора незалежних змінних $\vec{\xi} = \left(\xi_1, \xi_2, \ldots, \xi_q\right)^T$:

$$\eta \in \mathbb{R}, \quad \vec{\xi} \in \mathbb{R}^q.$$

У класичному регресійному аналізі в якості класу апроксимуючих функцій для $f(\vec{x})$ беруть клас функцій лінійних по вектору невідомих параметрів $\alpha = \left(\alpha_1, \alpha_2, \ldots, \alpha_p\right)^T$. Тобто математичну модель шукають у вигляді:

$$\eta = \sum_{i=1}^{p} \alpha_i \varphi_i\left(\vec{\xi}\right) + \varepsilon, \qquad (1.2)$$

де $\left\{\varphi_i(\bullet)\right\}_{i=1}^{p}$ - обраний (відомий) набір функцій, ε - похибка моделі.

Як правило, у якості $\left\{\varphi_i(\bullet)\right\}_{i=1}^{p}$ обирають набір незалежних (ще краще ортогональних) функцій.

У подальшому, $\varphi_i\left(\vec{\xi}\right)$ будемо називати i-**им регресором**, $i=\overline{1,p}$. А відповідно ξ_i – i-**ою незалежною змінною**, $i=\overline{1,q}$.

У k-ий момент модель (1.2) набуває вигляду:

$$y(k)=\sum_{i=1}^{p}\alpha_i\varphi_i\left(\vec{x}'(k)\right)+e(k), k=\overline{1,N} \tag{1.3}$$

Введемо позначення

$$x(k)=\begin{pmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_p(k) \end{pmatrix}=\begin{pmatrix} \varphi_1\left(\vec{x}'(k)\right) \\ \varphi_2\left(\vec{x}'(k)\right) \\ \vdots \\ \varphi_p\left(\vec{x}'(k)\right) \end{pmatrix}.$$

Тоді (1.3) можна переписати таким чином

$$y(k)=\sum_{i=1}^{p}\alpha_i x_i(k)+e(k), k=\overline{1,N}.$$

Або у такому представленні

$$y(k)=x^T(k)\alpha+e(k), k=\overline{1,N}. \tag{1.4}$$

Останню систему рівнянь можна записати у матричному вигляді:

$$y=X\alpha+e \tag{1.5}$$

де

$$y=\begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{pmatrix}, X=\begin{pmatrix} x^T(1) \\ x^T(2) \\ \vdots \\ x^T(N) \end{pmatrix}\in M_{N,p}(\mathbb{R}), e=\begin{pmatrix} e(1) \\ e(2) \\ \vdots \\ e(N) \end{pmatrix}.$$

Припущення класичного регресійного аналізу:

I. $e \sim \mathcal{N}\left(\theta_N, \sigma^2 E_N\right), \sigma^2 \in \mathbb{R}_+$,

II. $rank\left(X\right) = p$,

III. немає ніяких обмежень на α, тобто $\alpha \in \mathbb{R}^p$.

Зауваження. З цих припущень випливає:

1. з першого, що $e\left(k\right) \sim \mathcal{N}\left(0, \sigma^2\right)$ та є незалежними, $k = \overline{1, N}$.
2. з другого, що матриця X має повний ранг по стовпчикам, а $N \geq p$.
3. з третього, що α може набувати довільного значення з простору $\mathbb{R}^p$, бо на α не накладено ніяких обмежень.

Переходимо до 2-го етапу розв'язання задачі регресійного аналізу, а саме до пошуку оптимальної, згідно критерію якості (1.1), точкової оцінки $\hat{\alpha}$ для α та її характеристики розсіювання $M(\hat{\alpha}-\alpha)(\hat{\alpha}-\alpha)^T$.

Тобто потрібну оптимальну оцінку $\hat{\alpha}$ шукаємо таким чином:

$$\hat{\alpha} = \arg\min_{\alpha}\left\{\frac{1}{N}\sum_{k=1}^{N}\left(y(k) - \tilde{f}\left(\vec{x}'(k), \alpha\right)\right)^2\right\} =$$

$$= \arg\min_{\alpha}\left\{\sum_{k=1}^{N}\left(y(k) - x^T(k)\alpha\right)^2\right\} =$$

$$= \arg\min_{\alpha}\left\{\sum_{k=1}^{N} e^2(k)\right\} = \arg\min_{\alpha}\|e\|^2,$$

де $\|e\|$ - евклідова норма вектора e.

T

Означення. Оцінка $\hat{\alpha}$ для вектора невідомих параметрів α моделі (1.5), яка розв'язком задачі

$$\hat{\alpha} = \arg\min_{\alpha} \|e\|^2,$$

називається **оцінкою методу найменших квадратів** (МНК).

Ця оцінка була запропонована у 1795 році (тобто у 18 столітті) **Карлом Фрідріхом Гауссом**, коли йому було тільки 18 років.

Потрібну оцінку $\hat{\alpha}$ знайдемо як частинний випадок більш загальної оцінки.

Означення. Оцінка $\hat{\alpha}$ для вектора невідомих параметрів α моделі (1.5), яка розв'язком задачі

$$\hat{\alpha}_W = \arg\min_{\alpha} \|e\|_W^2, \; W > 0,$$

називається **оцінкою зваженого методу найменших квадратів** (ЗМНК), де $\|e\|_W^2 = e^T W e$.

Спочатку знайдемо оцінку $\hat{\alpha}_W$, а оцінку МНК $\hat{\alpha}$ отримаємо, як частинний випадок оцінки ЗМНК, бо $\hat{\alpha} = \hat{\alpha}_E$.

Зауваження. Відомо, що

$$grad_{\alpha}\left(\alpha^T \beta\right) = grad_{\alpha}\left(\beta^T \alpha\right) = \beta,$$

$$grad_{\alpha}\left(\alpha^T A \alpha\right) = \left(A + A^T\right)\alpha,$$

де α, β - вектори, а A - матриця відповідних розмірностей.

Переходимо до знаходження оцінки $\hat{\alpha}_W$. Розпишемо спочатку вираз критерію