Прикладная статистика. Корреляция. Регрессия.

Леонид Иосипой

Программа «Математика для анализа данных»
Центр непрерывного образования, ВШЭ

22 декабря 2020

- Повторение
- Корреляция
- Регрессия

Повторение

Критерии однородности позволяют проверить гипотезу о том, что выборки взяты из одного распределения.

Непараметрические критерии менее чувствительные (потому что более общие), но зато они не требуют «идеальных» условий, таких как, например, нормальность данных.

Часто бывает так, что при совсем небольших отклонениях от «идеальных» условий непараметрические критерии работают значительно лучше параметрических.

Повторение

Мы различали:

1. Случай одной выборки.
2. Случай двух независимых выборок.
3. Случай двух зависимых выборок.

Начнем с проверки однородности двух независимых выборок.
В этой модели выборки могут быть разного размера.

Повторение

1. Критерий Колмогорова-Смирнова. Резюме

выборки: $\mathbf{X} = (X_1, \ldots, X_{n_1})$, $X_i \sim F_X$
$\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$, $Y_i \sim F_Y$

$\mathbf{X}, \mathbf{Y}$ независимые; F_X, F_Y непрерывные

нулевая гипотеза: $H_0 : \ F_X = F_Y$

альтернатива: $H_1 : \ F_X \neq F_Y$

статистика: $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_n$

нулевое распределение: $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_n \sim K$ – распределение Колмогорова

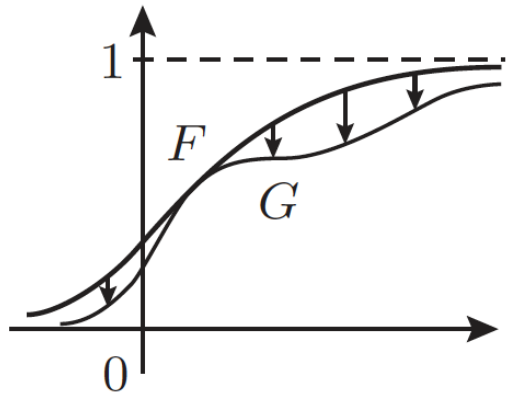
Повторение

1. Критерий Колмогорова-Смирнова

Критерий Колмогорова-Смирнова можно также использовать и при альтернативе доминирования:

$F_X(u) \geq F_Y(u)$ для всех $u \in \mathbb{R}$,
причем хотя бы для одного u
неравенство строгое.

Мы будем обозначать это как $F_X \geq F_Y$.



Данное свойство означает, что случайная величина Y стохастически больше случайной величины X, поскольку

$$\mathbb{P}(X \geq u) \leq \mathbb{P}(Y \geq u) \quad \text{для всех } u \in \mathbb{R}.$$

Повторение

1. Критерий Колмогорова-Смирнова. Резюме

выборки: $\mathbf{X} = (X_1, \ldots, X_{n_1})$, $X_i \sim F_X$
$\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$, $Y_i \sim F_Y$

$\mathbf{X}, \mathbf{Y}$ независимые; F_X, F_Y непрерывные

нулевая гипотеза: $H_0 : \ F_X = F_Y$
альтернатива: $H_1 : \ F_X \geq F_Y$

статистика: $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_n^+$

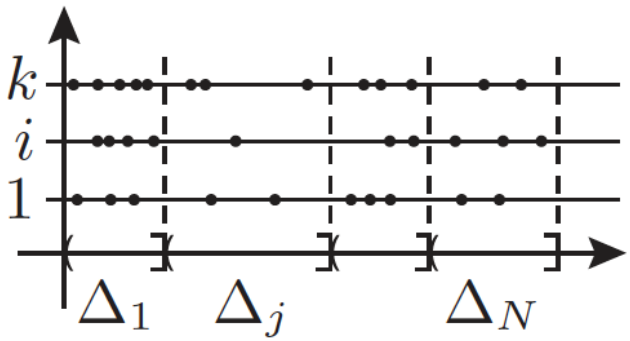
нулевое распределение: $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_n^+ \sim K^+$ – не имеет названия

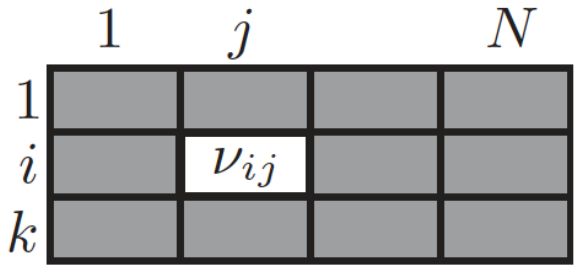
Повторение

2. Критерий хи-квадрат

Как обычно, сгруппируем данные:

- разобьем общую для всех выборок область значений наблюдений на промежутки $\Delta_1, \ldots, \Delta_N$;
- для всех индексов (i, j) подсчитаем ν_{ij} — количество попаданий элементов i-й выборки в j-й промежуток;
- запишем их в таблицу размера $k \times N$, которую и будем анализировать в дальнейшем.





Повторение

2. Критерий хи-квадрат

Если гипотеза однородности верна, то:

- ожидаемое количество наблюдений в ячейке с индексами i и j равно $n_i p_j$, где p_j — вероятность попадания в Δ_j;
- естественной оценкой для p_j служит общая по всем выборкам частота попаданий в Δ_j: $\widehat{p}_j = \frac{\nu_{1,j}+\ldots+\nu_{k,j}}{n}$.
- Тогда статистика

$$T_n = \sum_{i=1}^{k}\sum_{j=1}^{N} \frac{(\nu_{ij} - n_i\widehat{p}_j)^2}{n_i\widehat{p}_j}$$

сходится к распределению хи-квадрат $\chi^2_{(k-1)(N-1)}$.

Повторение

2. Критерий хи-квадрат. Резюме

выборки: $\mathbf{X_1} = (X_{1,1}, \ldots, X_{1,n_1})$, $X_1 \sim F_1$

$\ldots$

$\mathbf{X_k} = (X_{k,1}, \ldots, X_{k,n_k})$, $X_k \sim F_k$

$\mathbf{X_1}, \ldots, \mathbf{X_k}$ независимые

нулевая гипотеза: $H_0 : \; F_1 = F_2 = \ldots = F_k$

альтернатива: $H_1 : \;$ есть различия

статистика: T_n

нулевое распределение: $T_n \sim \chi^2_{(k-1)(N-1)}$ – хи-квадрат с $(k-1)(N-1)$ степенями свободы

Повторение

3. Критерий Манна-Уитни

Напомним, что по любой выборке $X_1, \ldots, X_n$ всегда можно сопоставить вариационный ряд, то есть упорядочить её по неубыванию:

$$X_{(1)} \leq \ldots < \underbrace{X_{(j_1)} = \ldots = X_{(j_2)}}_{\text{связка размера } j_2 - j_1 + 1} < \ldots \leq X_{(n)}.$$

Рангом наблюдения X_i называется:

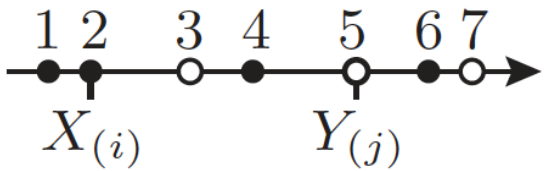
- если X_i не попадает в связку, то его позиция в вариационном ряду;
- если X_i попадает в связку, то $(j_1 + j_2)/2$; то есть в связке все объекты получают одинаковый средний ранг.

Повторение

3. Критерий Манна-Уитни

Вычислим статистику V_n критерия Манна-Уитни:

1. Обозначим через R_j ранг порядковой статистики $Y_{(j)}$, $j = 1, \ldots, m$, в вариационном ряду, построенном по объединенной выборке $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$.
2. Положим $V_n = R_1 + \ldots + R_{n_2}$.



Повторение

3. Критерий Манна-Уитни. Резюме

выборки: $\mathbf{X} = (X_1, \ldots, X_{n_1})$, $X_i \sim F_X$
$\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$, $Y_i \sim F_Y$

$\underline{\mathbf{X}, \mathbf{Y} \text{ независимые}}$

нулевая гипотеза: $H_0 : \ F_X = F_Y$

альтернатива: $H_1 : \ F_X \geq F_Y$ или $F_X \leq F_Y$ или $F_X \neq F_Y$

статистика: V_n

нулевое распределение: табличное для малых выборок
нормальное для больших выборок

Повторение

Перейдем теперь к проверке однородности для случая двух зависимых выборок. В этой модели выборки должны быть одинакового размера.

Обычно проверка однородности двух зависимых выборок заключается в проверке каких-то свойств приращений

$$Z_i = Y_i - X_i, \quad i = 1, \ldots, n.$$

Повторение

Разложим каждое приращение на две части:

$$Z_i = \theta + \varepsilon_i, \quad i = 1, \ldots, n,$$

где θ — интересующий нас эффект воздействия (систематический сдвиг), а ε_i — случайные ошибки, включающие в себя влияние неучтенных факторов на Z_i.

Мы будем дополнительно далее предполагать, что $\varepsilon_1, \ldots, \varepsilon_n$ независимы и имеют непрерывные (вообще говоря, разные) распределения с равной нулю медианой.

Повторение

4. Критерий знаков

Статистикой критерия знаков является величина

$$S_n = \sum_{i=1}^{n} \mathbf{I}_{\{Z_i>0\}}.$$

Тогда при верной гипотезе однородности S_n будет иметь биномиальное распределение $\mathbf{B}_{n,1/2}$, а при альтернативе — биномиальное распределение $\mathbf{B}_{n,p}$ с $p \neq 1/2$. Для больших n можно использовать сходимость к нормальному закону.

Повторение

4. Критерий знаков. Резюме

выборки: $\mathbf{X} = (X_1, \ldots, X_n),\ X_i \sim F_X$
$\mathbf{Y} = (Y_1, \ldots, Y_n),\ Y_i \sim F_Y$

$\underline{\mathbf{X}, \mathbf{Y} \text{ зависимые}}$

нулевая гипотеза: $H_0 :\ \theta = 0$

альтернатива: $H_1 :\ \theta > 0$ или $\theta < 0$ или $\theta \neq 0$

статистика: S_n

нулевое распределение: $S_n \sim \mathbf{B}_{n,1/2}$ для малых выборок
нормальное для больших выборок

Повторение

5. Критерий знаковых рангов Уилкоксона

Сделаем дополнительное предположение, что случайные величины $\varepsilon_1, \ldots, \varepsilon_n$ имеют одинаковое распределение, симметричное относительно нуля (то есть медианы).

Условие строгой симметрии относительно медианы является почти столь же нереалистичным, как и предположение, что распределение величин Z_i в точности нормально.

Повторение

5. Критерий знаковых рангов Уилкоксона

Критерий знаковых рангов Уилкоксона основан на статистике

$$W_n = U_1 R_1 + \ldots + U_n R_n,$$

где

- $U_i = \mathbf{I}_{\{Z_i > 0\}}$;
- R_i — ранги величин $|Z_i|$ в ряду $|Z_1|, \ldots, |Z_n|$.

Повторение

5. Критерий знаковых рангов Уилкоксона. Резюме

выборки:	$\mathbf{X} = (X_1, \ldots, X_n)$, $X_i \sim F_X$
	$\mathbf{Y} = (Y_1, \ldots, Y_n)$, $Y_i \sim F_Y$
	$\underline{\mathbf{X}, \mathbf{Y} \text{ зависимые}}$
нулевая гипотеза:	$H_0 : \ \theta = 0$
альтернатива:	$H_1 : \ \theta > 0$ или $\theta < 0$ или $\theta \neq 0$
статистика:	W_n
нулевое распределение:	табличное для малых выборок нормальное для больших выборок

Повторение

Оценка параметра сдвига

Если гипотеза однородности H_0 отвергнута в пользу альтернативы доминирования, можно дополнительно оценить параметр «сдвига».

В качестве оценки сдвига можно взять:

- в случае независимых выборок — выборочную медиану попарных приращений

$$\text{MED}\{X_i - Y_j,\ \ i = 1, \ldots, n_1, j = 1, \ldots, n_1\}.$$

- в случае зависимых выборок — выборочную медиану приращений

$$\text{MED}\{Z_i,\ \ i, j = 1, \ldots, n\}.$$

Повторение

Резюме

- При проверке гипотезы однородности крайне важно определиться, с каким из двух случаев мы имеем дело: двумя реализациями независимых между собой наблюдений или парными повторными наблюдениями.
- Критерии, применимые для зависимых выборок, можно использовать и для независимых. Однако при этом игнорируется важная информация о совместной независимости, что снижает чувствительность методов.
- В свою очередь применение критериев, предполагающих независимость выборок, в зависимом случае представляет собой грубую ошибку.

Повторение

- Параметрические критерии однородности Стьюдента, опирающиеся на предположение о нормальности данных, рекомендуется использовать только как вспомогательный инструмент. Он неустойчив к выбросам и быстро теряет эффективность даже при небольшом утяжелении хвостов наблюдений.

Корреляция

Перейдем теперь к вопросам зависимости и независимости случайных величин и оценки этой характеристики по реализации выборки.

Напомним, что ковариация случайных величин X и Y определяется следующим образом

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Корреляция

Известно, что значение ковариации двух случайных величин не превышает корня из произведения их дисперсий:

$$\mathrm{Cov}(X, Y) \leq \sqrt{\mathrm{Var}\, X \cdot \mathrm{Var}\, Y}.$$

Если разделить ковариацию на эту оценку сверху, мы получим корреляцию случайных величин

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}\, X \cdot \mathrm{Var}\, Y}},$$

которая так же будет характеризовать зависимость случайных величин, но ее значения будут уже лежать в отрезке $[-1, 1]$.

Корреляция

Естественно, на практике удобнее пользоваться корреляцией, так как близкие по модулю значения к 1 будут нам указывать, на возможную зависимость X и Y.

Теоретический коэффициент корреляции измеряет наличие прямой линейной зависимости. Причем

- $\mathrm{Corr}(X, Y) = 1$ тогда и только тогда, когда $Y = aX + b$ для некоторых $a > 0$, $b \in \mathbb{R}$;
- $\mathrm{Corr}(X, Y) = -1$ тогда и только тогда, когда $Y = aX + b$ для некоторых $a < 0$, $b \in \mathbb{R}$.

Корреляция

Пусть теперь у нас есть реализации $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ из законов X и Y соответсвенно.

Чтобы оценить $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$ можно воспользоваться следующей формулой:

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y}),$$

где $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ — средние значения выборок.

Эта оценка является оценкой Монте-Карло с plug-in постановкой оценок для $\mathbb{E}X$ и $\mathbb{E}Y$.

Корреляция

Оценка для корреляции выписывается аналогично:

$$\widehat{\rho}_p = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}.$$

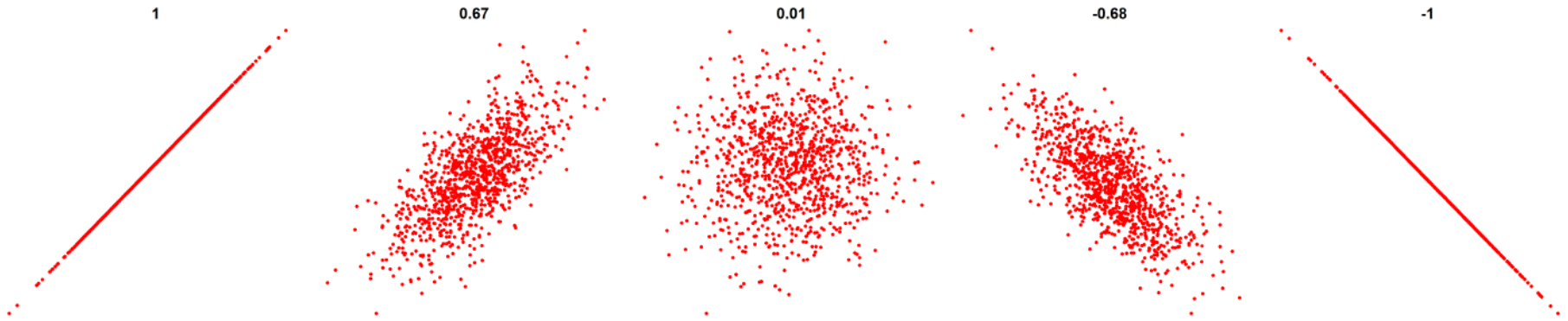
Оценка $\widehat{\rho}_p$ называется коэффициентом корреляции Пирсона.

Корреляция

Коэффициент корреляции Пирсона тоже будет лежать в диапазоне $[-1, 1]$ и будет измерять наличие прямой линейной зависимости. Аналогично,

- $\widehat{\rho}_p = 1$ тогда и только тогда, когда $y_i = ax_i + b$ для некоторых $a > 0$, $b \in \mathbb{R}$;
- $\widehat{\rho}_p = -1$ тогда и только тогда, когда $y_i = ax_i + b$ для некоторых $a < 0$, $b \in \mathbb{R}$.

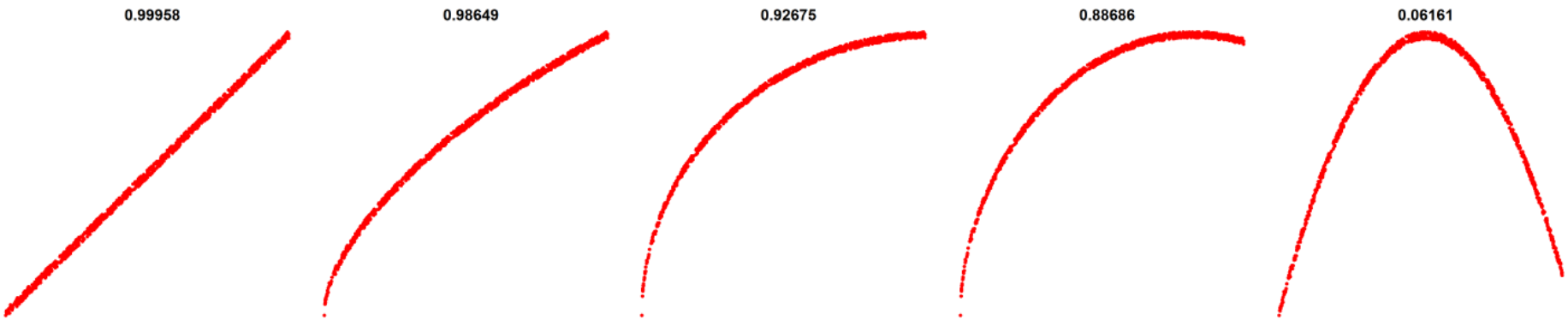
Корреляция



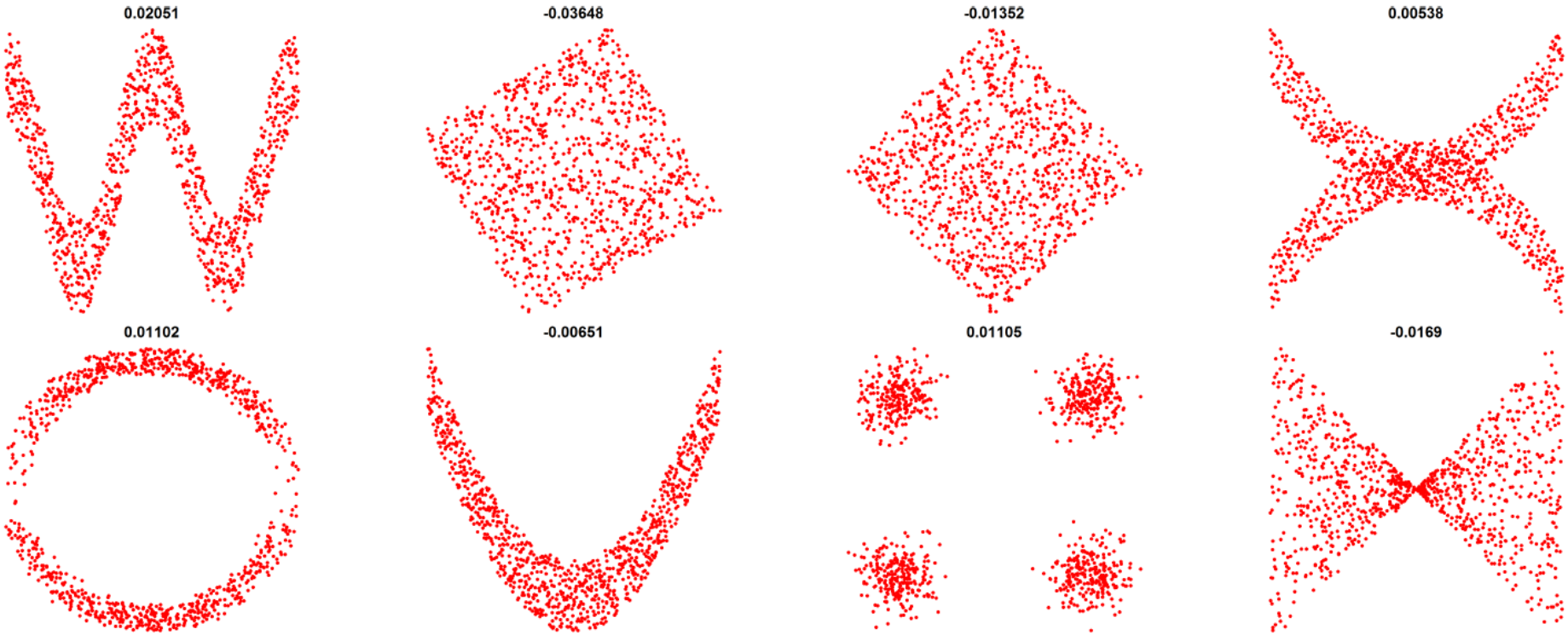
(a) (b) (c) (d) (e)

Корреляция

Коэффициент корреляции Пирсона может быть нечувствительным к другим видам зависимостей.



Корреляция

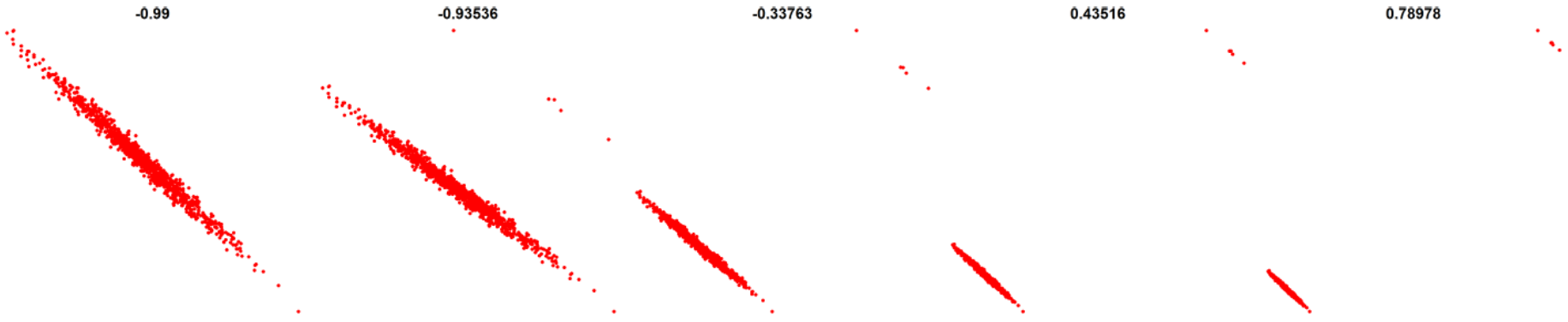


Корреляция

Более того, коэффициент корреляции Пирсона неустойчив к выбросам: небольшое количество точек могут оказывать на него существенное влияние, если они находятся достаточно далеко от основного облака.

В следующем примере мы возьмем 5 из 1000 точек из облака с сильной отрицательной корреляцией и начнем их постепенно отодвигать в верхний правый угол.

Корреляция



Мы видим, что с какого-то момента коэффициент Пирсона становится больше 0. Достаточно сильно отодвинув всего 5 точек из 1000, можно получить большой положительный коэффициент корреляции.

Корреляция

Рассмотрим теперь еще один коэффициент корреляции — ранговый коэффициент корреляции Спирмена $\widehat{\rho}_s$.

Заменим x_i на их ранги R_i в ряду $x_1, \ldots, x_n$, а y_i — на их ранги S_i в ряду $y_1, \ldots, y_n$. Тогда коэффициентом корреляции Спирмена называется величина

$$\widehat{\rho}_s = \frac{\sum_{i=1}^{n} \left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}}.$$

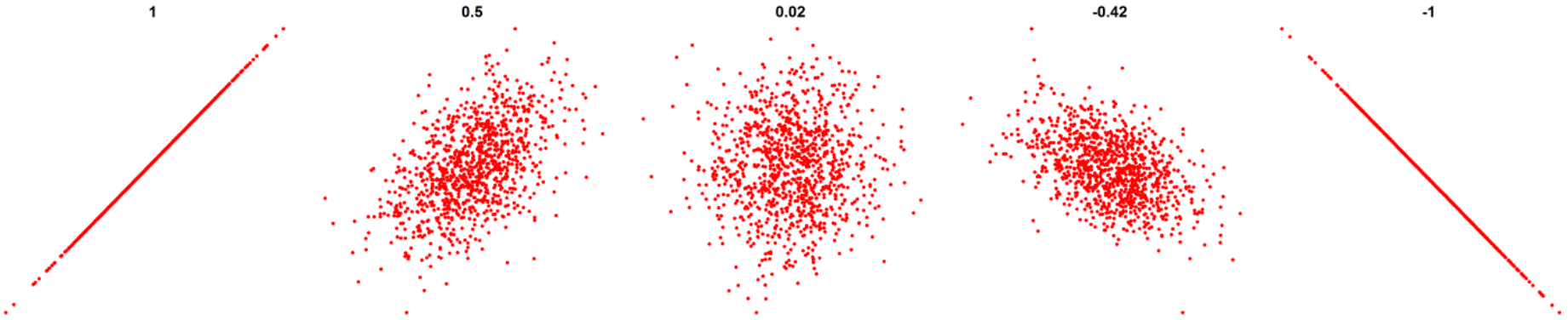
Корреляция

Если $\widehat{\rho}_s$ близок по абсолютному значению к 1, то это означает, что R_i почти линейно зависят от S_i, то есть зависимость X_i от Y_i монотонна.

Более того, из-за того, что мы перешли от наблюдений к их рангам, коэффициент корреляции Спирмена стал более устойчив к выбросам.

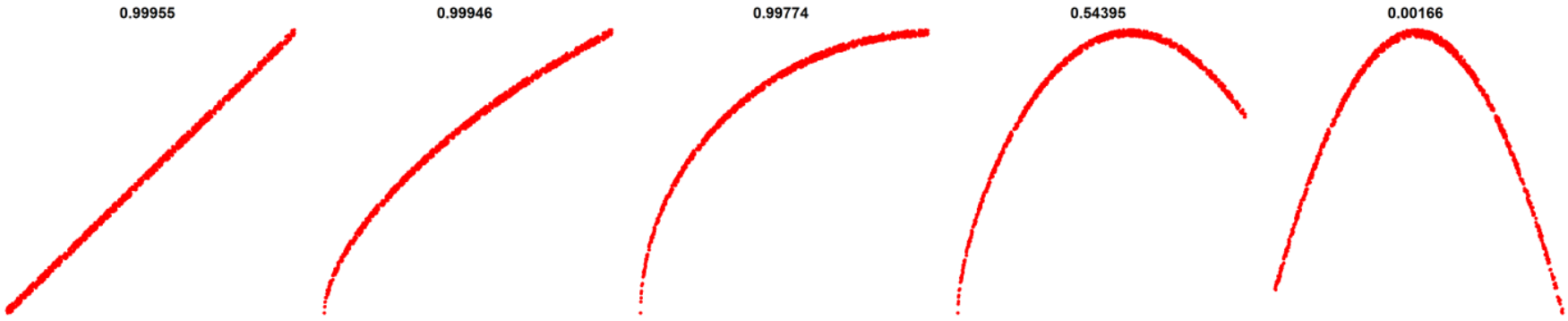
Давайте посмотрим на наши старые эксперименты, но уже для коэффициента Спирмена.

Корреляция

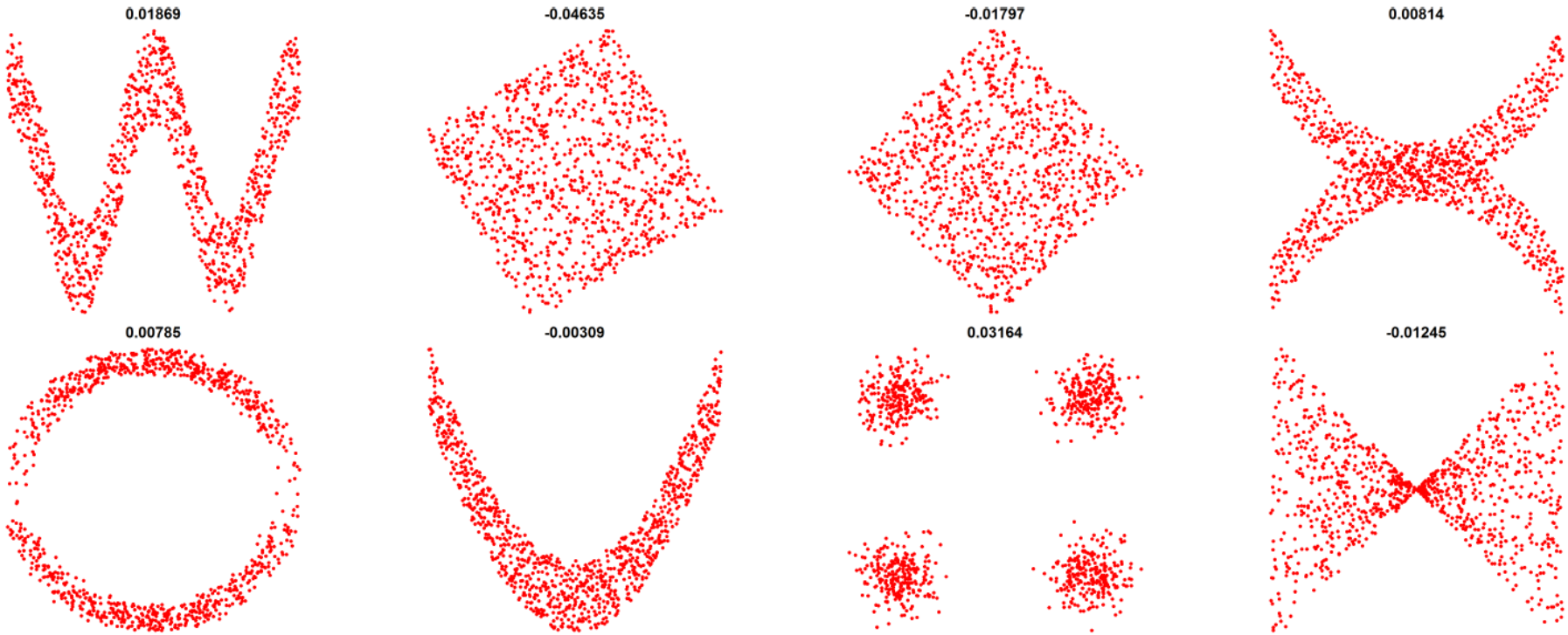


(a) (b) (c) (d) (e)

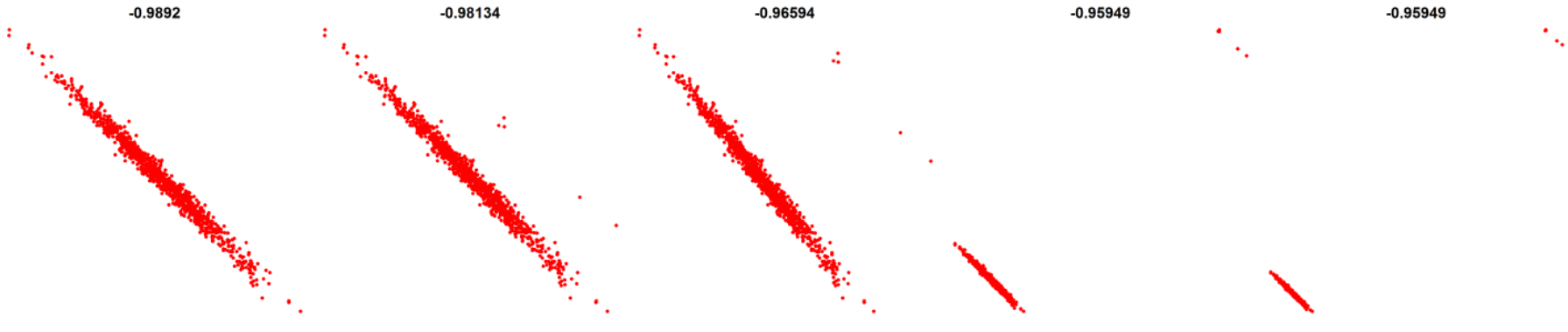
Корреляция



Корреляция



Корреляция



Видим, что коэффициент корреляции Спирмена гораздо более устойчив к выбросам.

Корреляция

Иногда еще используют коэффициент корреляции Кенделла.

Назовем две пары значений x_i, y_i и x_j, y_j согласованными, если $x_i - x_j$ и $y_i - y_j$ — одного знака. Пусть C — количество согласованных пар, а D — количество несогласованных пар.

Коэффициентом корреляции Кенделла называется величина

$$\widehat{\rho}_k = \frac{C - D}{C + D} = \frac{2(C - D)}{n(n-1)}.$$

Коэффициент $\widehat{\rho}_k$ сильно коррелирован с коэффициентом $\widehat{\rho}_s$.

Корреляция

После того, как был посчитан какой-то коэффициент корреляции, можно проверить значимость этого коэффициента с помощью критерия.

Опять более удобными оказываются ранговые критерии:

- у них однозначно определено нулевое распределение при достаточно общих предположениях;
- при больших n можно воспользоваться сходимостью к нормальному закону

$$\frac{\widehat{\rho}_s}{\sqrt{\operatorname{Var}\widehat{\rho}_s}} = \widehat{\rho}_s\sqrt{n-1} \to Z, \qquad \frac{\widehat{\rho}_k}{\sqrt{\operatorname{Var}\widehat{\rho}_k}} = \widehat{\rho}_k\sqrt{\frac{9n(n-1)}{2(2n+5)}} \to Z,$$

где $Z \sim \mathcal{N}(0,1)$

Корреляция

Критерий Спирмена

выборки: $\mathbf{X} = (X_1, \ldots, X_n),\ X_i \sim F_X$
$\mathbf{Y} = (Y_1, \ldots, Y_n),\ Y_i \sim F_Y$

нулевая гипотеза: $H_0 :\ \widehat{\rho}_s = 0$

альтернатива: $H_1 :\ \widehat{\rho}_s \neq 0$ или $\widehat{\rho}_s < 0$ или $\widehat{\rho}_s > 0$

статистика: $\widehat{\rho}_s$

нулевое распределение: известное для малых выборок
нормальное для больших выборок

Корреляция

Критерий Кенделла

выборки: $\mathbf{X} = (X_1, \ldots, X_n),\ X_i \sim F_X$
$\mathbf{Y} = (Y_1, \ldots, Y_n),\ Y_i \sim F_Y$

нулевая гипотеза: $H_0 :\ \widehat{\rho}_k = 0$

альтернатива: $H_1 :\ \widehat{\rho}_k \neq 0$ или $\widehat{\rho}_k < 0$ или $\widehat{\rho}_k > 0$

статистика: $\widehat{\rho}_k$

нулевое распределение: известное для малых выборок
нормальное для больших выборок

Корреляция

Даже если вы обнаружили корреляцию между двумя признаками, и она оказалась значимой, то это не значит, что между этими признаками есть какая-либо причинно-следственная связь.

Корреляция

Пример

Представим, что дети пишут языковой тест, X — их оценка, Y — вес ребенка. Пусть мы обнаружили, что x_i в целом больше, когда больше y_i. Можно ли говорить, что больший вес детей влечет лучшую успеваемость?

Корреляция

- А что, если дети разных возрастов от 5 до 15 лет?
- А что, если среди детей есть дети из двух разных стран, причем в одной стране дети в целом крупнее, чем в другой?
- А что, если родители кормят детей конфетами, если те хорошо учатся?

Таким образом, исследование причинности достаточно сложно. Причинно-следственная связь может идти от чего-то третьего, она может быть «дискретной» (как в примере с двумя странами) или может быть и вовсе быть обратной.

Регрессия

Регрессионный анализ решает задачу выявления искаженной случайным «шумом» функциональной зависимости интересующего исследователя показателя Y от измеряемых переменных $X_1, \ldots, X_k$.

Обычно:

- Y называют откликом, зависимой или критериальной переменной;
- $X_1, \ldots, X_k$ называют факторами, предикторами или регрессорами.

Регрессия

Основной целью обычно является как можно более точный прогноз (предсказание) Y на основе измеряемых переменных $X_1, \ldots, X_k$ (например, на новых данных).

Кроме того, с помощью регрессии можно измерить влияние факторов на отклик, найти выбросы, исключать ненужные/неудобные факторы.

Регрессия

Мы будем изучать линейную регрессию.

В линейной регрессии мы делаем предположение, что

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik}, \quad i = 1, \ldots, n,$$

где

- $y_i, x_{i1}, \ldots, x_{ik}$ — отклик и значения k признаков для этого отклика (нам известные);
- $\beta_0, \beta_1, \ldots, \beta_k$ — константы, которые не зависят от номера отклика (нам неизвестные).

Задача состоит в том, чтобы оценить $\beta_0, \beta_1, \ldots, \beta_k$.

Регрессия

Регрессионное равенство можно переписать в матричном виде как

$$y \approx X\beta,$$

где

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}.$$

Здесь мы добавили в матрицу X единичный столбец, чтобы больше не думать про коэффициент β_0.

Регрессия

Мы будем изучать свойства метода наименьших квадратов без использования каких-либо регуляризаторов:

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

Точное решение $\widehat{\beta}$ этой задачи известно и равно

$$\widehat{\beta} = (X^\top X)^{-1} X^\top y.$$

Можно посчитать и предсказание модели $\widehat{y}$ на объектах, на которых она обучается:

$$\widehat{y} = X(X^\top X)^{-1} X^\top y.$$

Регрессия

Итак, строить обычную линейную регрессию очень просто.

Однако если по построенной модели хочется делать какие-то выводы с использованием статистических методов, необходимо приложить дополнительные усилия.

Именно этим мы и займемся.

Регрессия

Чтобы исследовать качество решения метода наименьших квадратов, определим величину TSS (Total Sum of Squares) — разброс y относительно своего среднего:

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Регрессия

Оказывается, что (если в модель включен коэффициент β_0) TSS можно представить в виде суммы:

$$\text{TSS} = \text{RSS} + \text{ESS},$$

- RSS — это сумма квадратов отклонений предсказанных y от их истинных значений:

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

- ESS — это сумма квадратов отклонений среднего y от предсказанных y:

$$\text{ESS} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2.$$

Регрессия

По величинам RSS и ESS можно составить меру R^2, которая называется коэффициентом детерминации:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

По сути, это доля объясненной дисперсии отклика во всей дисперсии отклика.

Регрессия

Сделаем следующие предположения:

(П1) Истинная модель действительно является «зашумленной» линейной:

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

для некоторых (неизвестных) коэффициентов $\beta_0, \ldots, \beta_k \in \mathbb{R}$ и некоторой случайной ошибки ε_i с $\mathbb{E}[\varepsilon_i | X] = 0$.

(П2) Наблюдения действительно случайны, то есть $(y_i, x_{i1}, \ldots, x_{ik})$ для $i = 1, \ldots, n$ образуют независимую выборку.

Регрессия

(П3) Матрица X является матрицей полного (столбцового) ранга:

$$\operatorname{rank} X = k + 1.$$

То есть ни один из признаков не должен являться линейной комбинацией других. Поскольку среди столбцов есть константа, никакой из признаков в выборке не должен быть константой.

Регрессия

Уже из этих трех предположений можно вывести, что оценки, получаемые методом наименьших квадратов, являются несмещенными и состоятельными:

$$\mathbb{E}\left[\widehat{\beta}_j\right] = \beta_j \quad \text{и} \quad \widehat{\beta}_j \xrightarrow{\mathbb{P}} \beta_j, \quad j = 0, \ldots, k.$$

Регрессия

Более того, предположим еще что:

(П4) Ошибки $\varepsilon_1, \ldots, \varepsilon_n$ имеют одинаковую дисперсию, которая не зависит от значений признаков (гомоскедастичность ошибок):

$$\mathrm{Var}(\varepsilon_i | X) = \sigma^2, \quad i = 1, \ldots, n,$$

где $\sigma^2 > 0$ — неизвестный параметр.

Тогда можно показать, что дисперсия оценок, получаемых методом наименьших квадратов, является наименьшей в классе всех оценок, линейных по y (теорема Гаусса-Маркова).

То есть оценки метода наименьших квадратов (П1)-(П4) являются в некотором смысле оптимальными.

Регрессия

Рассмотрим еще одно предположение:

(П5) Ошибки $\varepsilon_1, \ldots, \varepsilon_n$ имеют нормальное распределение

$$\varepsilon_i | X \sim \mathcal{N}(0, \sigma^2).$$

Если выполняются (П1)-(П5), то оценки метода наименьших квадратов совпадают с оценками максимального правдоподобия.

Это означает, что оценки метода наименьших квадратов обладают всеми свойствами, которыми обладают оценки максимального правдоподобия.

Регрессия

При выполнении (П1)-(П5) можно показать, что

$$y|X \sim \mathcal{N}\Big(X\beta, \sigma^2 \mathbf{I}_n\Big), \quad \mathbf{I}_n \text{ — единичная матрица размера } n.$$

Оценки $\widehat{\beta}$ тоже будут иметь нормальное распределение

$$\widehat{\beta}|X \sim \mathcal{N}\Big(\beta, \sigma^2 (X^\top X)^{-1}\Big).$$

Регрессия

Если выполняются описанные предположения, то можно строить доверительный интервалы для коэффициентов, доверительные интервалы для отклика и проверять гипотезы о значимости каких-то коэффициентов.

Мы этим займемся в следующий раз.

Спасибо за внимание!