Ковариация и корреляция. Линейная регрессия.

Леонид Иосипой

Курс «Вероятностные модели и статистика»
Центр непрерывного образования, ВШЭ

29 апреля 2021

- Ковариация и корреляция
- Многомерное нормальное распределение
- Линейная регрессия

Ковариация и корреляция

Пусть задано распределение случайных величин X и Y. Тогда число $\mathrm{Cov}(X, Y)$ называется их ковариацией определяется по следующей формуле:

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

Интерпретация ковариации: если ковариация положительна, то с ростом значений одной случайной величины, значения второй имеют тенденцию возрастать, а если знак отрицательный — то убывать.

Ковариация и корреляция

Пользуясь свойствами математического ожидания, можно показать, что ковариацию можно вычислять по формуле:

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

Ковариация и корреляция

Важное свойство ковариации: если случайные величины X и Y независимы, то $\mathrm{Cov}(X, Y) = 0$.

Действительно, так как X и Y независимы,

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big] = \mathbb{E}\big[X - \mathbb{E}X\big]\mathbb{E}\big[Y - \mathbb{E}Y\big] = 0.$$

Обратное утверждение не верно: если $\mathrm{Cov}(X, Y) = 0$, то не обязательно X и Y будут независимыми.

Ковариация и корреляция

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

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

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

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

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

Ковариация и корреляция

Более того, можно показать, что:

- $\mathrm{Corr}(X, Y) = 1$ 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 $a < 0$, $b \in \mathbb{R}$.

То есть коэффициент корреляции измеряет наличие прямой линейной зависимости между X и Y.

Ковариация и корреляция

Задача

Пусть $X, Y \sim \mathbf{B}_{1/2}$ независимые. Пусть также $Z = 1 - X$. Найти $\mathrm{Cov}(X, Y)$, $\mathrm{Corr}(X, Y)$ и $\mathrm{Cov}(X, Z)$, $\mathrm{Corr}(X, Z)$.

Ковариация и корреляция

Решение. Так как X и Y независимы, то $\mathrm{Cov}(X, Y) = 0$ и $\mathrm{Corr}(X, Y) = 0$.

Для случайных величин X и Z найдем сначала $\mathbb{E}[XZ]$. Воспользовавшись свойствами математического ожидания:

$$\mathbb{E}[XZ] = \mathbb{E}[X(1 - X)] = \mathbb{E}[X] - \mathbb{E}[X^2] = 0.$$

Чтобы вычислить ковариацию, заметим, что $Z \sim \mathbf{B}_{1/2}$. По определению ковариации:

$$\mathrm{Cov}(X, Z) = \mathbb{E}[XZ] - \mathbb{E}[X] \cdot \mathbb{E}[Z] = 0 - \frac{1}{2} \cdot \frac{1}{2} = -\frac{1}{4}.$$

Вспомним, что $\mathrm{Var}(X) = \mathrm{Var}(Z) = 1/2 \cdot 1/2 = 1/4$. Поэтому

$$\mathrm{Corr}(X, Z) = \frac{\mathrm{Cov}(X, Z)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Z)}} = \frac{-1/4}{1/4} = -1.$$

Ковариация и корреляция

Задача

Пусть $X \sim \mathcal{N}(0, 1)$ и $Y = X^2$. Найти $\mathrm{Cov}(X, Y)$ и $\mathrm{Corr}(X, Y)$.

Ковариация и корреляция

Решение. Найдем сначала $\mathrm{Cov}(X, Y)$. По определению

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X] \cdot \mathbb{E}[X^2].$$

Заметим, что $\mathbb{E}[X] = 0$, поэтому $\mathrm{Cov}(X, Y) = \mathbb{E}[X^3]$. Найдем это значение:

$$\mathbb{E}[X^3] = \int_{-\infty}^{+\infty} u^3 \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = 0,$$

так как подынтегральная функция является нечетной.

Следовательно,

$$\mathrm{Cov}(X, Y) = 0, \qquad \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 0.$$

Ковариация и 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 построена по принципу Монте-Карло с plug-in постановкой оценок для $\mathbb{E}X$ и $\mathbb{E}Y$.

Ковариация и корреляция

Оценка для корреляции выписывается аналогично:

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

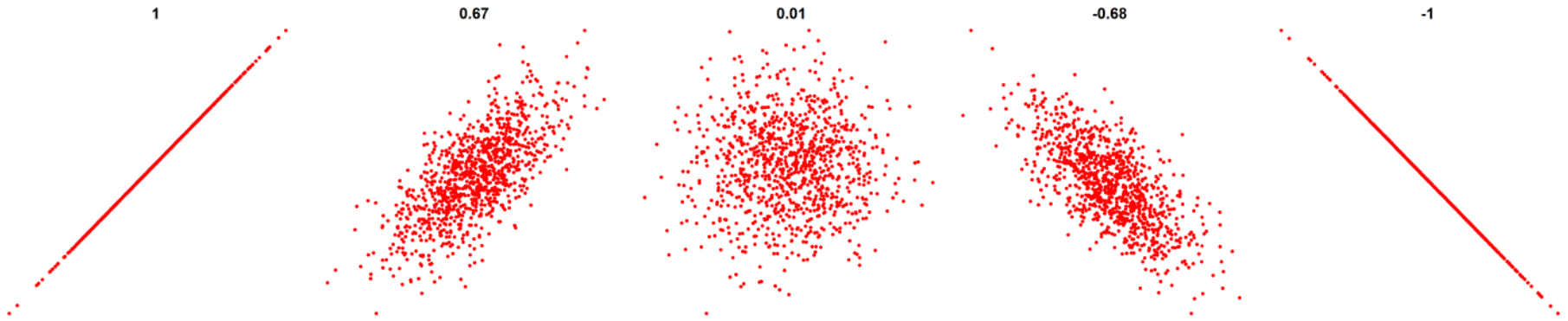
Оценка $\widehat{\rho}_p$ называется коэффициентом корреляции Пирсона.

Ковариация и 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 $a < 0$, $b \in \mathbb{R}$.

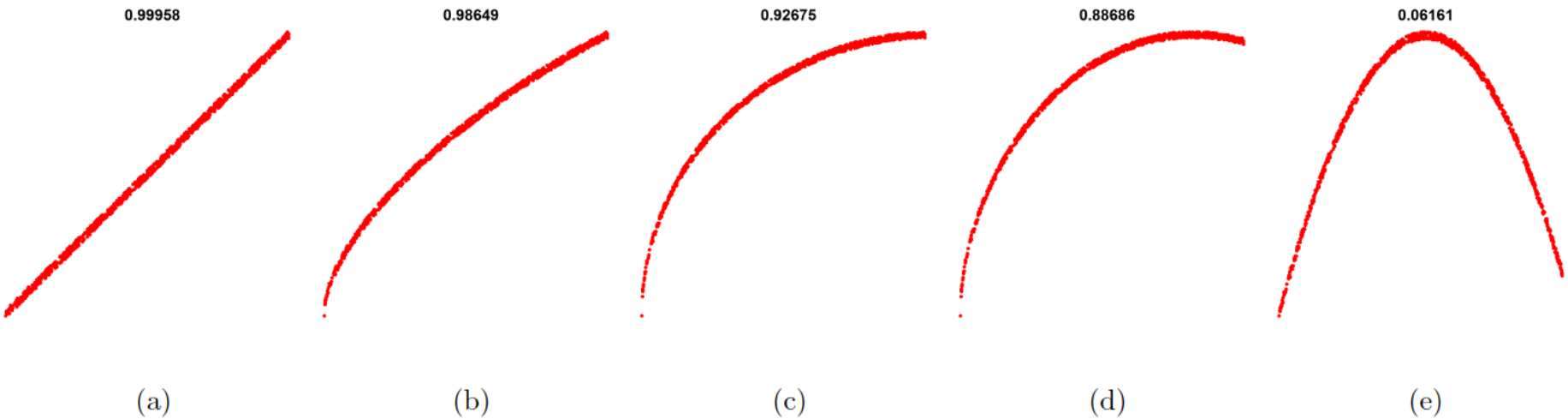
Ковариация и корреляция



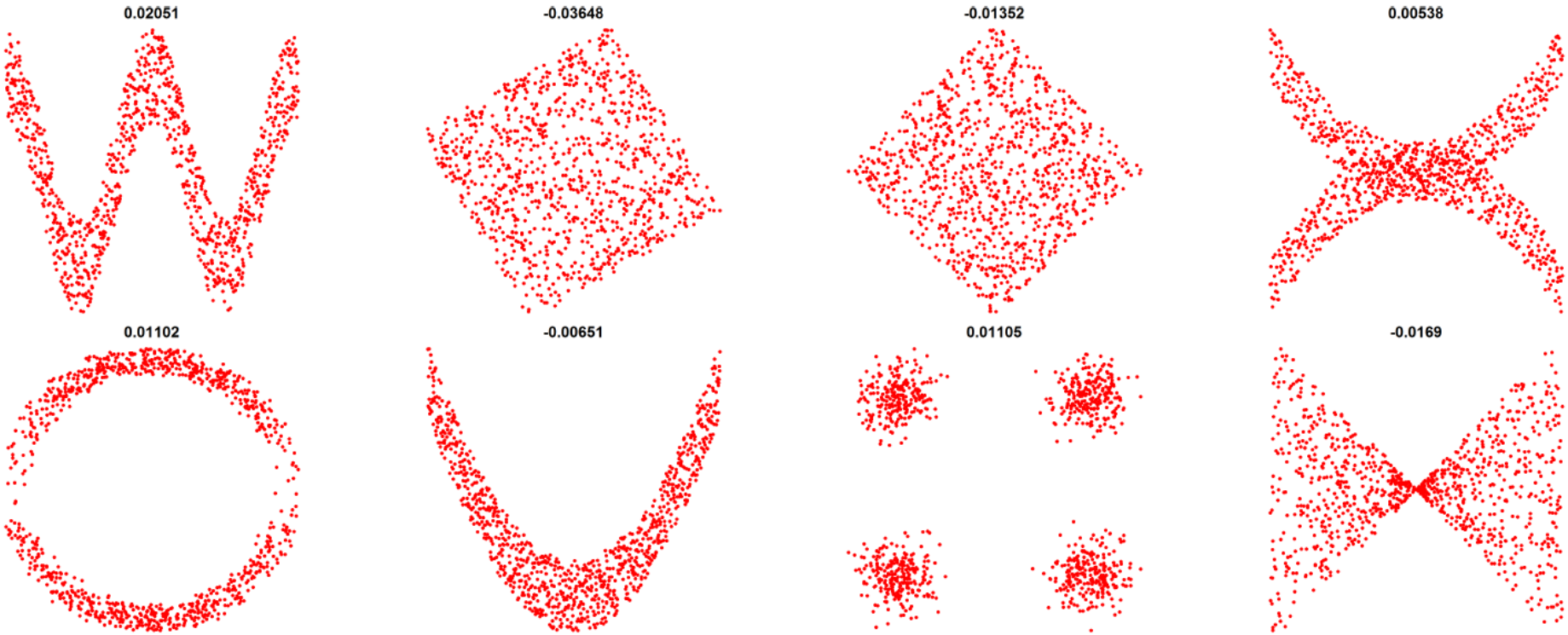
(a) (b) (c) (d) (e)

Ковариация и корреляция

Коэффициент корреляции Пирсона может быть нечувствительным к другим видам зависимостей.



Ковариация и корреляция

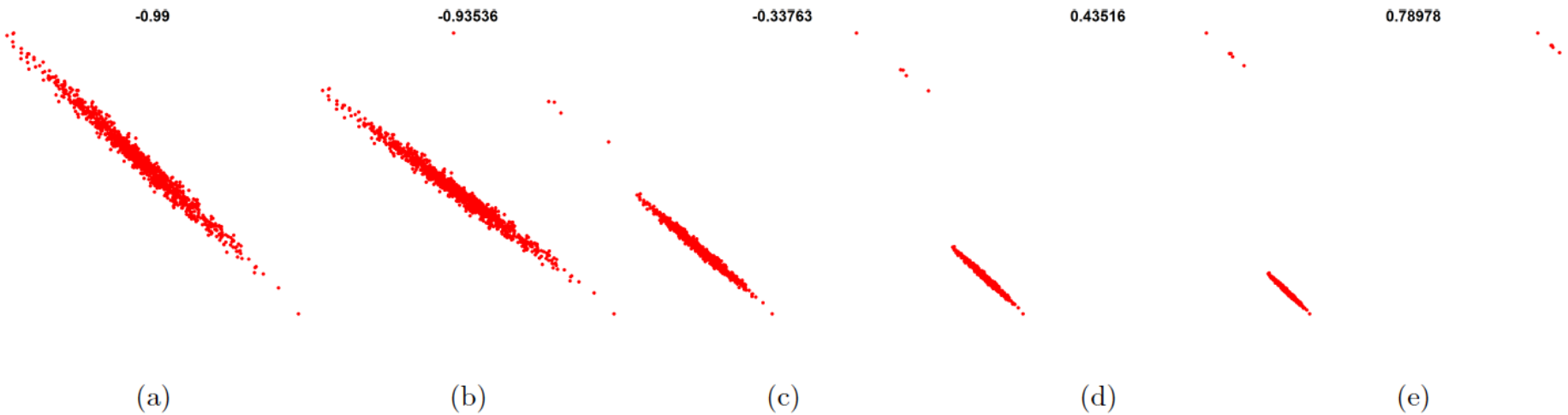


Ковариация и корреляция

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

Ковариация и 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.

Ковариация и 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

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

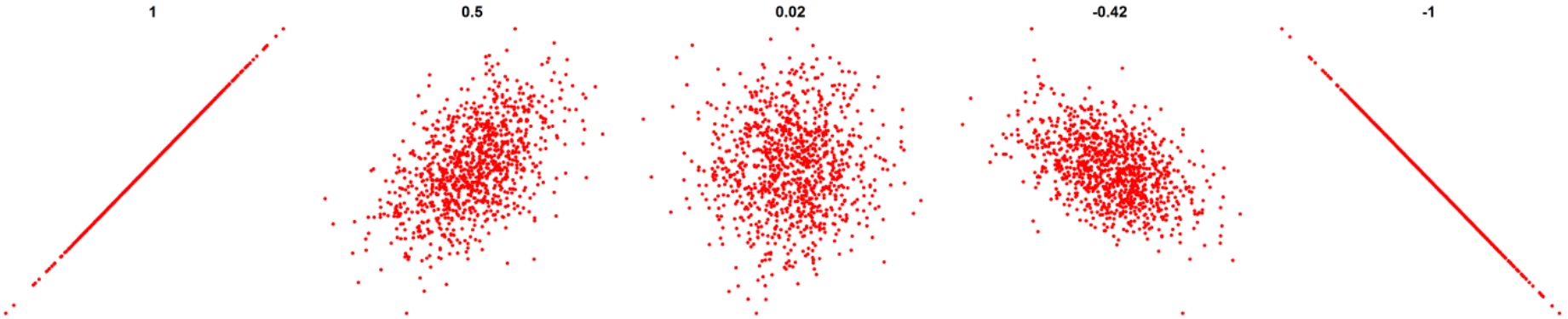
Ковариация и корреляция

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.

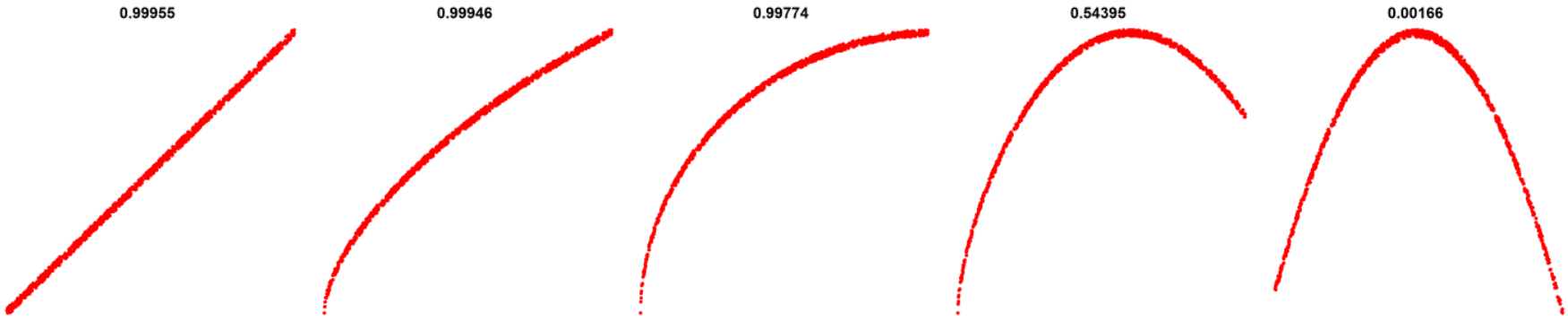
Давайте посмотрим на наши старые эксперименты, но уже для коэффициента Спирмена.

Ковариация и корреляция



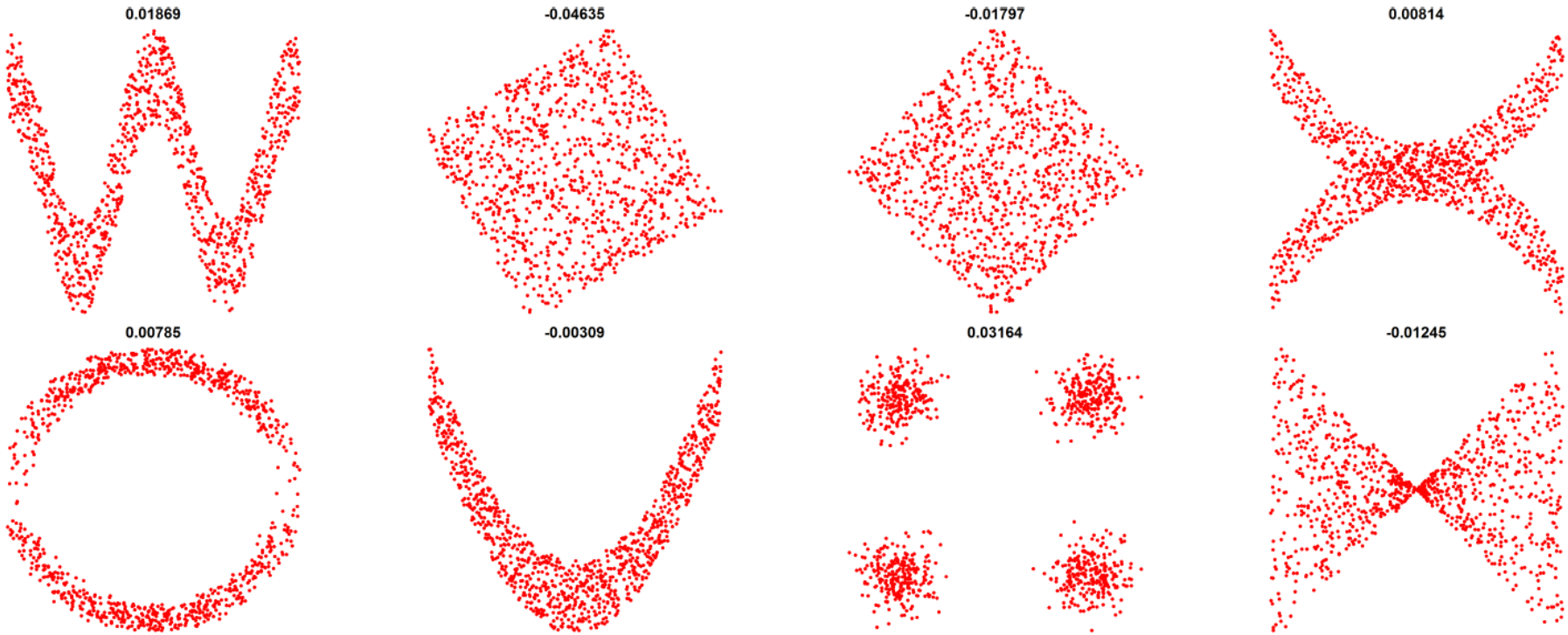
(a) (b) (c) (d) (e)

Ковариация и корреляция

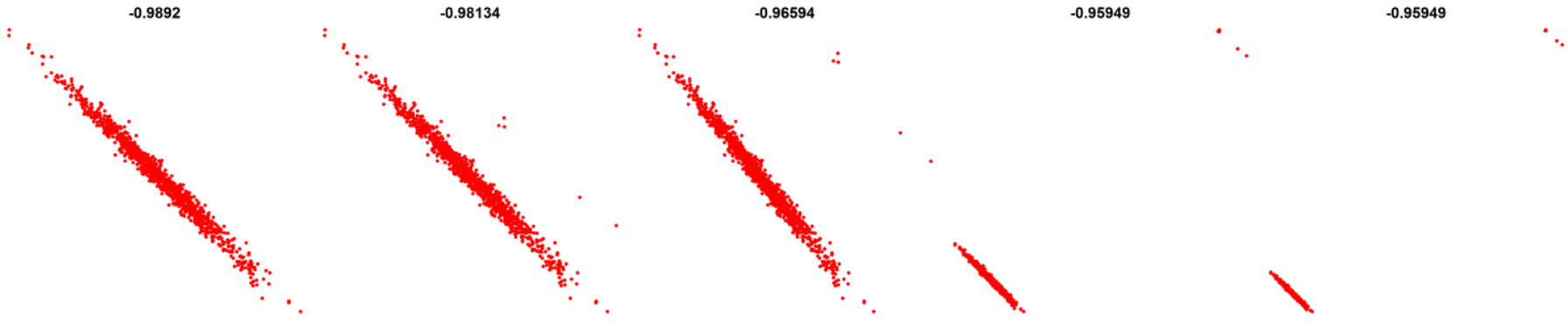


(a) (b) (c) (d) (e)

Ковариация и корреляция



Ковариация и корреляция



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

Ковариация и 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 $\widehat{\rho}_s$.

Ковариация и корреляция

После того, как был посчитан коэффициент корреляции, можно проверить его значимость 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 $Z \sim \mathcal{N}(0,1)$

Ковариация и 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

Ковариация и 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

Ковариация и 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.

Ковариация и 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?

Ковариация и 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.

Многомерное нормальное распределение

Пусть $Z_1, \ldots, Z_n$ независимые и $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

Составим из них вектор $Z = (Z_1, \ldots, Z_n)^\top$. Он будет иметь многомерное нормальное распределение

$$Z \sim \mathcal{N}(\mu, \Sigma), \quad \text{где} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}.$$

Плотность этого вектора Z будет равна

$$f_Z(u_1, \ldots, u_n) = f_{Z_1}(u_1) \cdot \ldots \cdot f_{Z_n}(u_n) = \frac{1}{(2\pi)^{n/2}\sigma_1 \cdot \ldots \cdot \sigma_n} e^{-\frac{1}{2}\sum_{i=1}^{n} \frac{(u_i - \mu_i)^2}{2\sigma_i^2}}.$$

Многомерное нормальное распределение

В записи $Z \sim \mathcal{N}(\mu, \Sigma)$ параметры играют следующую роль:

- μ — вектор средних;
- Σ — ковариационная матрица.

Ковариационной матрицей (произвольного) случайного вектора $Z = (Z_1, \ldots, Z_n)^\top$ называется матрица с элементами

$$\Sigma_{ij} = \mathrm{Cov}(Z_i, Z_j) = \mathbb{E}\big[(Z_i - \mu_i)(Z_j - \mu_j)\big].$$

В матричном виде это можно записать так:

$$\Sigma = \mathbb{E}\big[(Z - \mu)(Z - \mu)^\top\big].$$

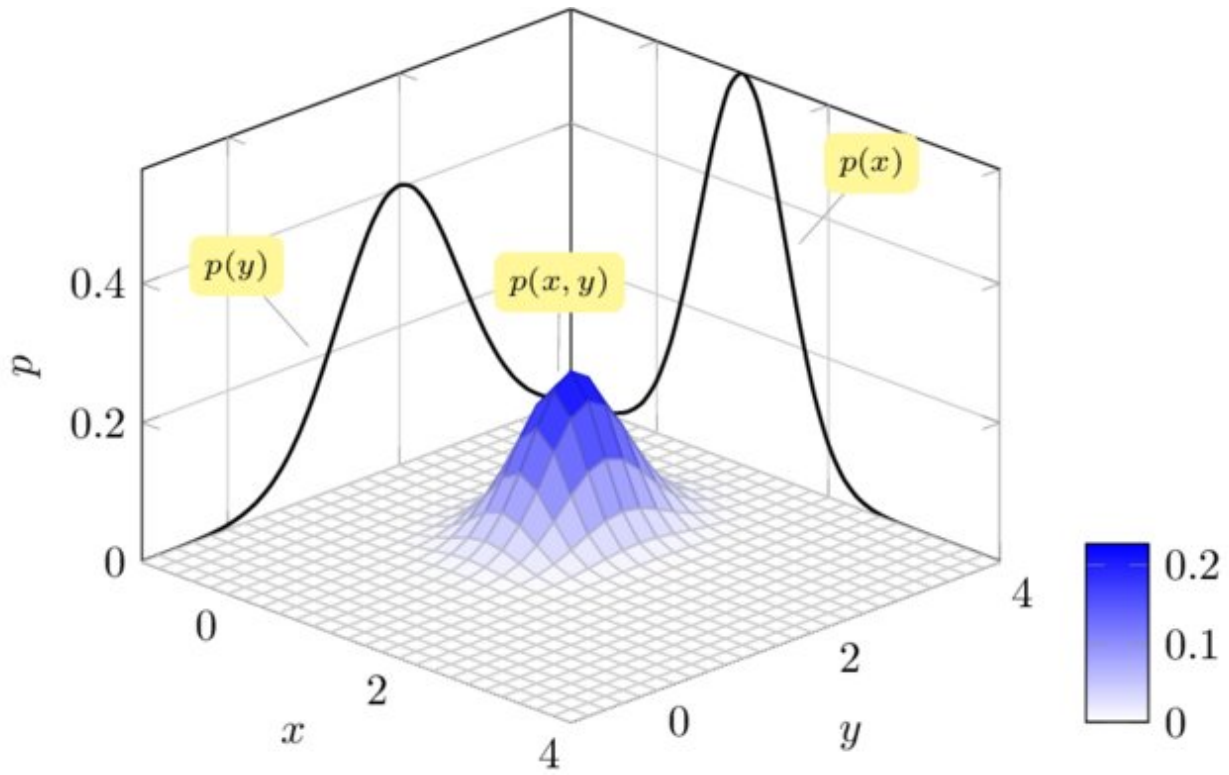
Все ковариационные матрицы должны быть симметричными и неотрицательно определенными.

Многомерное нормальное распределение

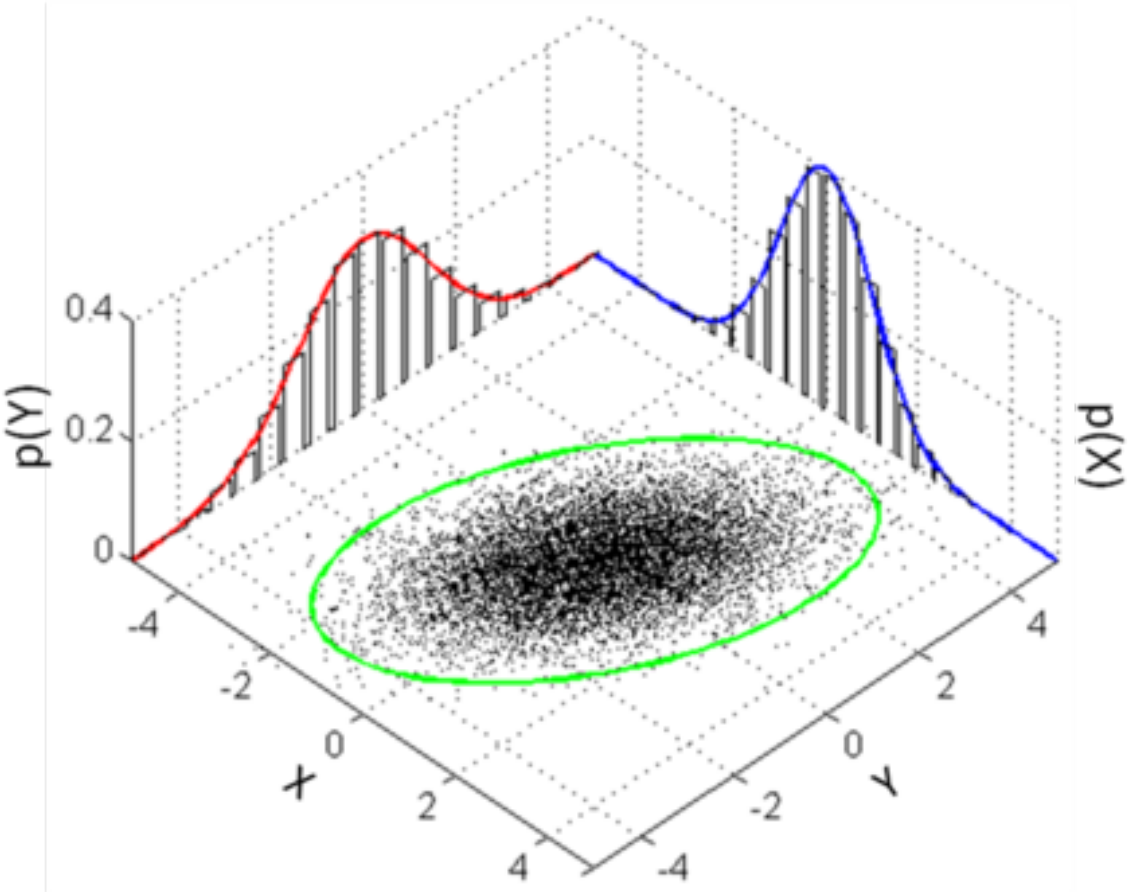
В общем случае, многомерным нормальным вектором $Z \sim \mathcal{N}(\mu, \Sigma)$ со средним μ и ковариационной матрицей Σ называется вектор с плотностью

$$f_Z(u) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(u-\mu)^\top \Sigma^{-1}(u-\mu)}.$$

Многомерное нормальное распределение



Многомерное нормальное распределение



Многомерное нормальное распределение

Как меняются параметры многомерного нормального распределения при линейных преобразованиях?

Пусть $Z \sim \mathcal{N}(\mu, \Sigma)$. Пусть также $a \in \mathbb{R}^n$ — произвольный вектор и X — произвольная матрица размера $k \times n$. Тогда

- $Z + a \sim \mathcal{N}(\mu + a, \Sigma)$

  $\mathbb{E}[Z + a] = \mu + a;$

  $\mathbb{E}\big[(Z + a - (\mu + a))(Z + a - (\mu + a))^\top\big] = \mathbb{E}\big[(Z - \mu)(Z - \mu)^\top\big] = \Sigma.$

- $XZ \sim \mathcal{N}(X\mu, X\Sigma X^\top)$

  $\mathbb{E}[XZ] = X\mathbb{E}[Z] = X\mu;$

  $\mathbb{E}\big[(XZ - X\mu)(XZ - X\mu)^\top\big] = \mathbb{E}\big[X(Z - \mu)(Z - \mu)^\top X^\top\big] = X\Sigma X^\top.$

Многомерное нормальное распределение

Из этих свойств, например, видно, что $Z \sim \mathcal{N}(\mu, \Sigma)$ можно представить как

$$Z = \mu + \Sigma^{1/2} Z, \quad Z \sim \mathcal{N}(0, I_n),$$

где I_n — единичная матрица размера $n \times n$.

Линейная регрессия

Регрессионный анализ решает задачу выявления искаженной случайным «шумом» зависимости некоторого 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.

Линейная регрессия

Основной целью обычно является как можно более точный прогноз Y по новым измеряемым переменным $X_1, \ldots, X_k$.

Кроме этого, с помощью регрессии можно: измерить влияние факторов на отклик, исключить ненужные/неудобные факторы, найти выбросы.

Линейная 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 $\beta_0, \beta_1, \ldots, \beta_k$.

Линейная регрессия

Регрессионное равенство можно переписать в матричном виде как

$$y \approx X\beta,$$

где

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

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

Линейная 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:

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

Линейная 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.

Линейная регрессия

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

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

Линейная регрессия

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

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

- RSS (Residual Sum of Squares) — это сумма квадратов отклонений предсказанных y от их истинных значений:

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

- ESS (Explained Sum of Squares) — это сумма квадратов отклонений среднего y от предсказанных y:

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

Линейная 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.

Линейная 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 $\mathbb{E}[\varepsilon_i] = 0$.

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

Линейная 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.

Линейная регрессия

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

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

Линейная регрессия

Более того, предположим еще что:

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

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

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.

Линейная регрессия

Рассмотрим еще одно предположение:

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

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

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.

Линейная регрессия

Более того, при выполнении (П1)-(П5) мы можем посчитать распределения всех случайных объектов в модели.

Действительно, мы имеем

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

где I_n — единичная матрица размера $n \times n$.

Так как X и β не случайны, то

$$y \sim \mathcal{N}\Big(X\beta, \sigma^2 I_n\Big).$$

Линейная регрессия

Далее, так как $\widehat{\beta} = (X^\top X)^{-1} X^\top y$ и $y \sim \mathcal{N}\Big(X\beta, \sigma^2 I_n\Big)$, то

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

И, наконец,

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

Зная распределения, мы можем строить доверительные интервалы, проверять гипотезы о значимости и т.д.

Линейная регрессия

В прошлых формулах присутствовала дисперсия шума σ^2, которую мы не знаем. Ее можно оценить с помощью RSS:

$$\widehat{\sigma}^2 = \frac{\mathrm{RSS}}{n-k-1}.$$

Кроме того, отношение RSS к истинной дисперсии σ^2 будет иметь распределение хи-квадрат:

$$\frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{n-k-1}.$$

Более того, для любого вещественного вектора $c \in \mathbb{R}^{k+1}$ справедливо следующее утверждение:

$$\frac{c^\top(\beta - \widehat{\beta})}{\widehat{\sigma}^2\sqrt{c^\top(X^\top X)^{-1}c}} \sim T_{n-k-1}.$$

Линейная регрессия

Эти факты позволяют нам построить следующие доверительные интервалы уровня доверия $1-\alpha$, $\alpha \in (0,1)$:

- для неизвестной дисперсии шума σ^2:

$$\mathbb{P}\left(\frac{\mathrm{RSS}}{c_{1-\alpha/2}} \leq \sigma^2 \leq \frac{\mathrm{RSS}}{c_{\alpha/2}}\right) = 1-\alpha,$$

  где c_α — квантиль уровня α распределения χ^2_{n-k-1}.

- для регрессионных коэффициентов $\beta_0, \ldots, \beta_k$:

$$\mathbb{P}\left(\widehat{\beta}_j - c_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}} \leq \beta_j \leq \widehat{\beta}_j + c_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}\right) = 1-\alpha,$$

  где $(X^\top X)^{-1}_{jj}$ — j,j элемент матрицы $(X^\top X)^{-1}$ и c_α — квантиль уровня α распределения T_{n-k-1}.

Линейная регрессия

Аналогично построению доверительных интервалов, можно проверить гипотезу о том, что признак j незначим, то есть что $\beta_j = 0$, $j = 0, \ldots, k$.

Критерий Стьюдента

нулевая гипотеза: $H_0 : \beta_j = 0$

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

статистика: $T = \dfrac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}}$

нулевое распределение: $T \sim T_{n-k-1}$

Линейная регрессия

Можно также проверить гипотезу о том, что сразу несколько коэффициентов β_j равны 0.

Критерий Фишера

нулевая гипотеза:	$H_0 : \beta_{j_1} = \ldots = \beta_{j_m} = 0$ для некоторых $0 \le j_1 < \ldots < j_m \le k$
альтернатива:	$H_1 : \beta_{j_1}, \ldots, \beta_{j_m} \neq 0$ одновременно
статистика:	$F = \ldots$
нулевое распределение:	$F \sim F_{m,n-k-1}$ — распределение Фишера

Линейная регрессия

Обратите внимание, что доверительные интервалы и критерии строятся в предположениях (П1)-(П5).

Если ошибки имеют разную дисперсию и/или распределены не нормально, то доверительные интервалы будут неверными!

Линейная регрессия

Есть несколько типичных ошибок, которые следует иметь в виду, применяя регрессионный анализ. Сами по себе они достаточно очевидны. Тем не менее, о них часто забывают при работе с реальными данными и в результате приходят к неверным выводам.

Существуют три вида лжи: ложь, наглая ложь и статистика.
(Марк Твен)

Линейная регрессия

Пример

При исследовании зависимости веса Z студентов двух групп от их роста X и размера обуви Y в первой группе было получено регрессионное уравнение

$$Z = 0.9X + 0.1Y,$$

а для второй группы:

$$Z = 0.2X + 0.8Y.$$

Как объяснить существенное различие коэффициентов этих двух моделей?

Линейная регрессия

Ответ: дело здесь в том, что X и Y сильно зависимы, поэтому «весовые» коэффициенты при X и Y случайным образом распределяются между слагаемыми.

Линейная регрессия

Пример

Во время второй мировой войны англичане исследовали зависимость точности бомбометания Z от ряда факторов, в число которых входили высота бомбардировщика H, скорость ветра V, количество истребителей противника X.

Как и ожидалось, Z увеличивалась при уменьшении H и V. Однако (что поначалу представлялось необъяснимым), точность бомбометания Z возрастала также и при увеличении X.

Линейная регрессия

Ответ: дальнейший анализ позволил понять причину этого парадокса. Дело оказалось в том, что первоначально в модель не был включен такой важный фактор, как Y — облачность. Он сильно влияет и на Z (уменьшая точность), и на X (бессмысленно высылать истребители, если ничего не видно).

Линейная регрессия

Пример

Если найти зависимость между ежегодным количеством родившихся в Голландии детей Z и количеством прилетевших аистов X, то она окажется довольно значительной. Можно ли на основе этого статистического результата заключить, что детей приносят аисты?

Линейная регрессия

Ответ: рассмотрим проблему на содержательном уровне. Аисты появляются там, где им удобно вить гнезда; излюбленным же местом их гнездовья являются высокие дымовые трубы, какие строят в голландских сельских домах.

По традиции новая семья строит себе новый дом — появляются новые трубы и, естественно, рождаются дети. Таким образом, и увеличение числа гнезд аистов, и увеличение числа детей являются следствиями одной причины Y — образования новых семей.

Линейная регрессия

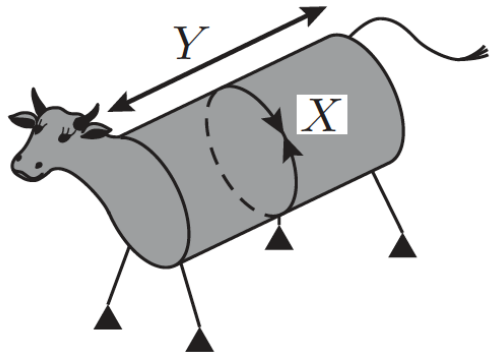
Пример

Рассмотрим в качестве отклика Z вес коровы, а в качестве предикторов — окружность ее туловища X и расстояние от хвоста до холки Y. Сравнительному анализу были подвергнуты три регрессионные модели:

(1) линейная: $Z = \theta_1 + \theta_2 X + \theta_3 Y$;

(2) степенная: $Z = \theta_1' X^{\theta_2'} Y^{\theta_3'}$;

(3) учитывающая содержательный смысл задачи $Z = \theta_0 X^2 Y$.



Регрессия

Модель	По всем наблюдениям		По части наблюдений	
	$\widehat{\boldsymbol{\theta}}$	$\widehat{\sigma}$	$\widehat{\boldsymbol{\theta}}_{\text{тяж}}$	$\widehat{\sigma}_{\text{лег}}$
1	$\widehat{\theta}_1 = -984{,}7$ $\widehat{\theta}_2 = 4{,}73$ $\widehat{\theta}_3 = 4{,}70$	25,9	$\widehat{\theta}_1 = 453{,}2$ $\widehat{\theta}_2 = 0{,}62$ $\widehat{\theta}_3 = -0{,}22$	81
2	$\widehat{\theta}'_1 = 0{,}0011$ $\widehat{\theta}'_2 = 1{,}556$ $\widehat{\theta}'_3 = 1{,}018$	24,5	$\widehat{\theta}'_1 = 266{,}4$ $\widehat{\theta}'_2 = 0{,}203$ $\widehat{\theta}'_3 = -0{,}072$	79
3	$\widehat{\theta}_0 = 1{,}13 \cdot 10^{-4}$	26,6	$\widehat{\theta}_0 = 1{,}11 \cdot 10^{-4}$	28

Подробнее про пример можно посмотреть в книге Лагутина.

Спасибо за внимание!