Прикладная статистика. Корреляция. Регрессия I.

Леонид Иосипой

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

4 марта 2021

- Корреляция
- Многомерное нормальное распределение
- Регрессия

Корреляция

Перейдем теперь к вопросам зависимости и независимости случайных величин и ее оценке 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:

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

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

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

Такой величиной удобно пользоваться на практике, так как ее значения будут уже лежать в отрезке $[-1, 1]$.

Корреляция

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

Корреляция

Оценка для корреляции выписывается аналогично:

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

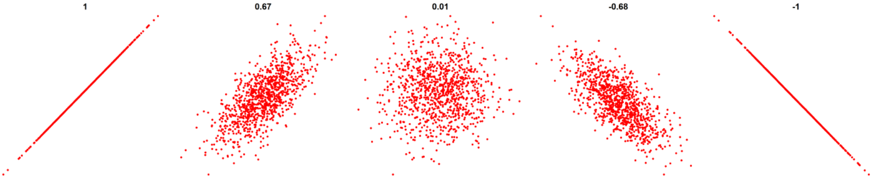
Оценка $\widehat{\rho}_p$ называется коэффициентом корреляции Пирсона.

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



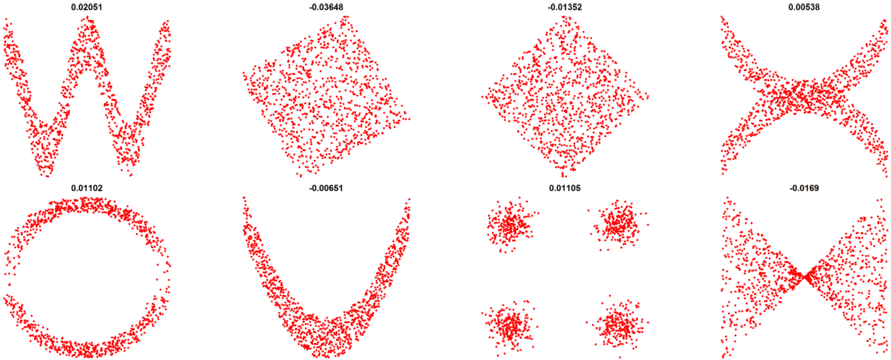
(a) (b) (c) (d) (e)

Корреляция

Коэффициент корреляции Пирсона может быть нечувствительным к другим видам зависимостей.

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

Корреляция

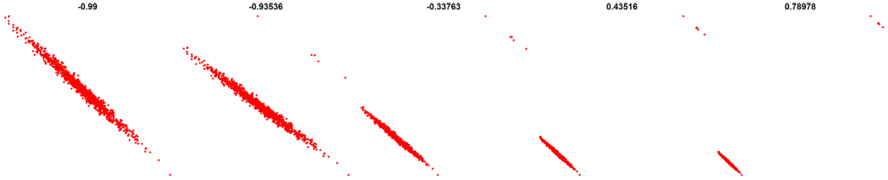


Корреляция

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

Корреляция



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

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

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

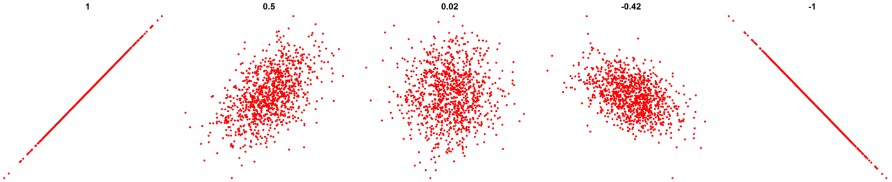
Корреляция

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.

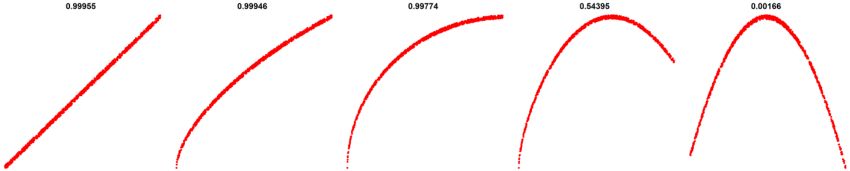
Давайте посмотрим на наши старые эксперименты, но уже для коэффициента Спирмена.

Корреляция



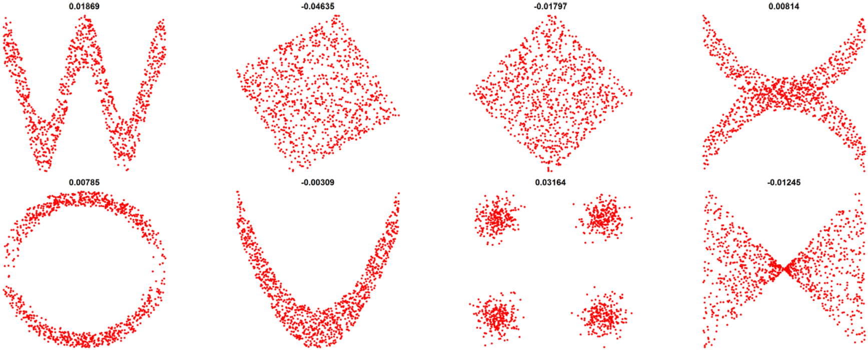
(a) (b) (c) (d) (e)

Корреляция

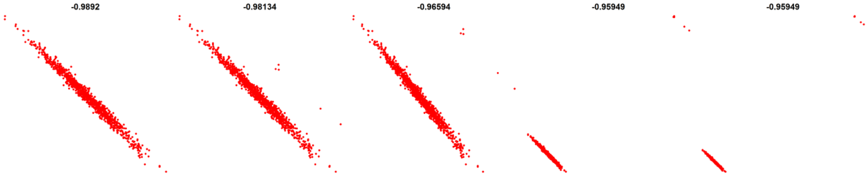


(a) (b) (c) (d) (e)

Корреляция



Корреляция



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

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

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

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.

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:

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

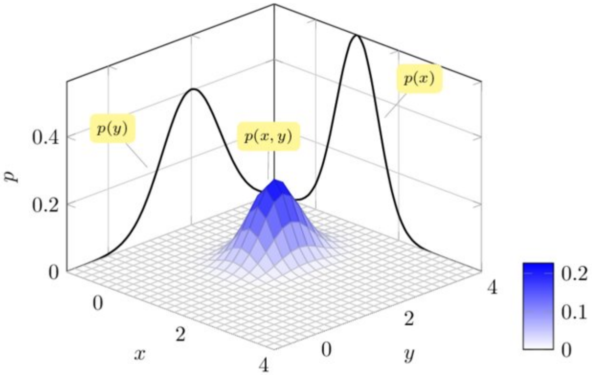
Все ковариационные матрицы должны быть симметричными и неотрицательно определенными.

Многомерное нормальное распределение

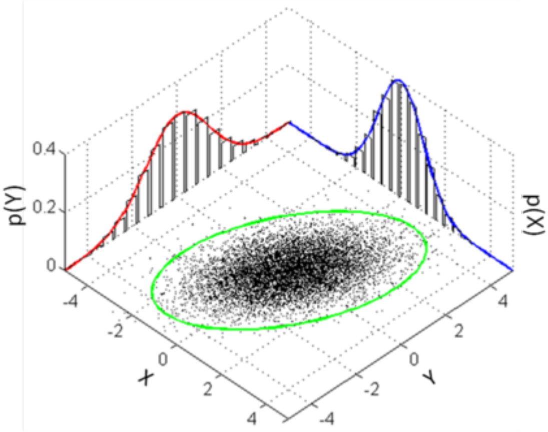
В общем случае, многомерным нормальным вектором $Z \sim \mathcal{N}(\mu, \Sigma)$ со средним μ и ковариационной матрицей Σ называется вектор с плотностью

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

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

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

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

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

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

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

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

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 $n \times n$.

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.

Регрессия

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

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

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

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

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:

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

- 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:

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

Регрессия

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

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

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

(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):

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

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 $\varepsilon_1, \dots, \varepsilon_n$ 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.

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

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

Регрессия

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

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

И, наконец,

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

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

Регрессия

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

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

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

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

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

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

Регрессия

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

- для неизвестной дисперсии шума σ^2:

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

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

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

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

  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}.

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

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

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!

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)

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?

Регрессия

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

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.

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).

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?

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.

Спасибо за внимание!