Прикладная статистика 7. Регрессионный анализ.

11 октября 2013 г.

Постановка задачи линейной регрессии

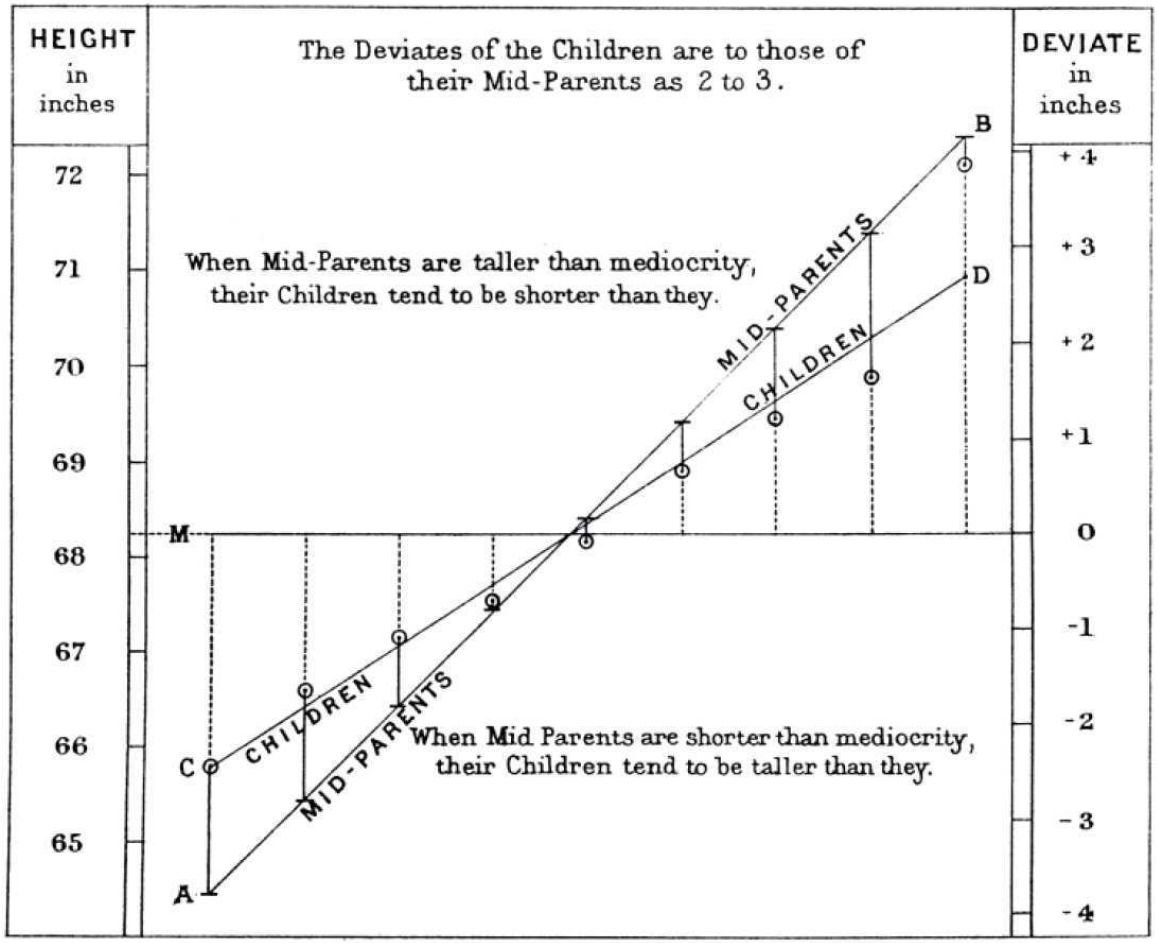
$x_1, \ldots, x_n$ — объекты;
$f_1, \ldots, f_k, y$ — признаки, значения которых измеряются на объектах;
$f_1, \ldots, f_k$ — объясняющие переменные, предикторы, регрессоры, факторы, признаки;
y — зависимая переменная, отклик.

$y \approx F(f_1, \ldots, f_k)$ — модель регрессии;
$y \approx \theta_0 + \sum_{j=1}^{k} \theta_j f_j$ — модель линейной регрессии.

Здесь и далее $n > k$ (или даже $n \gg k$).

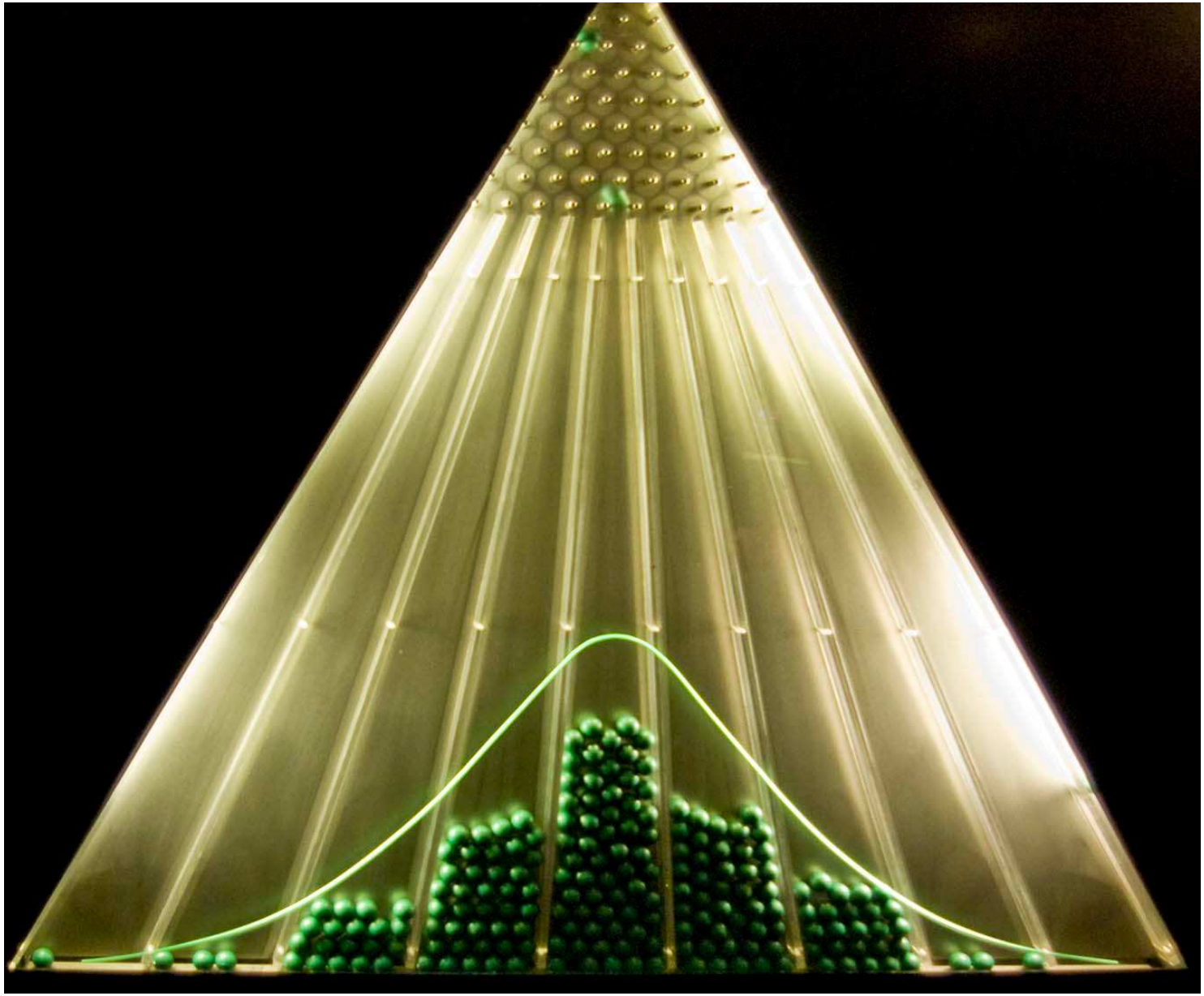
Первое появление

Впервые такая постановка появляется в работе Гальтона 1885 г. «Регрессия к середине в наследственности роста».



$$y - \bar{y} \approx \frac{2}{3}\,(x - \bar{x})\,.$$

Машина Гальтона



МНК

Матричные обозначения:

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}; \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}.$$

Метод наименьших квадратов:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \theta_j x_{ij} \right)^2 \to \min_{\theta};$$

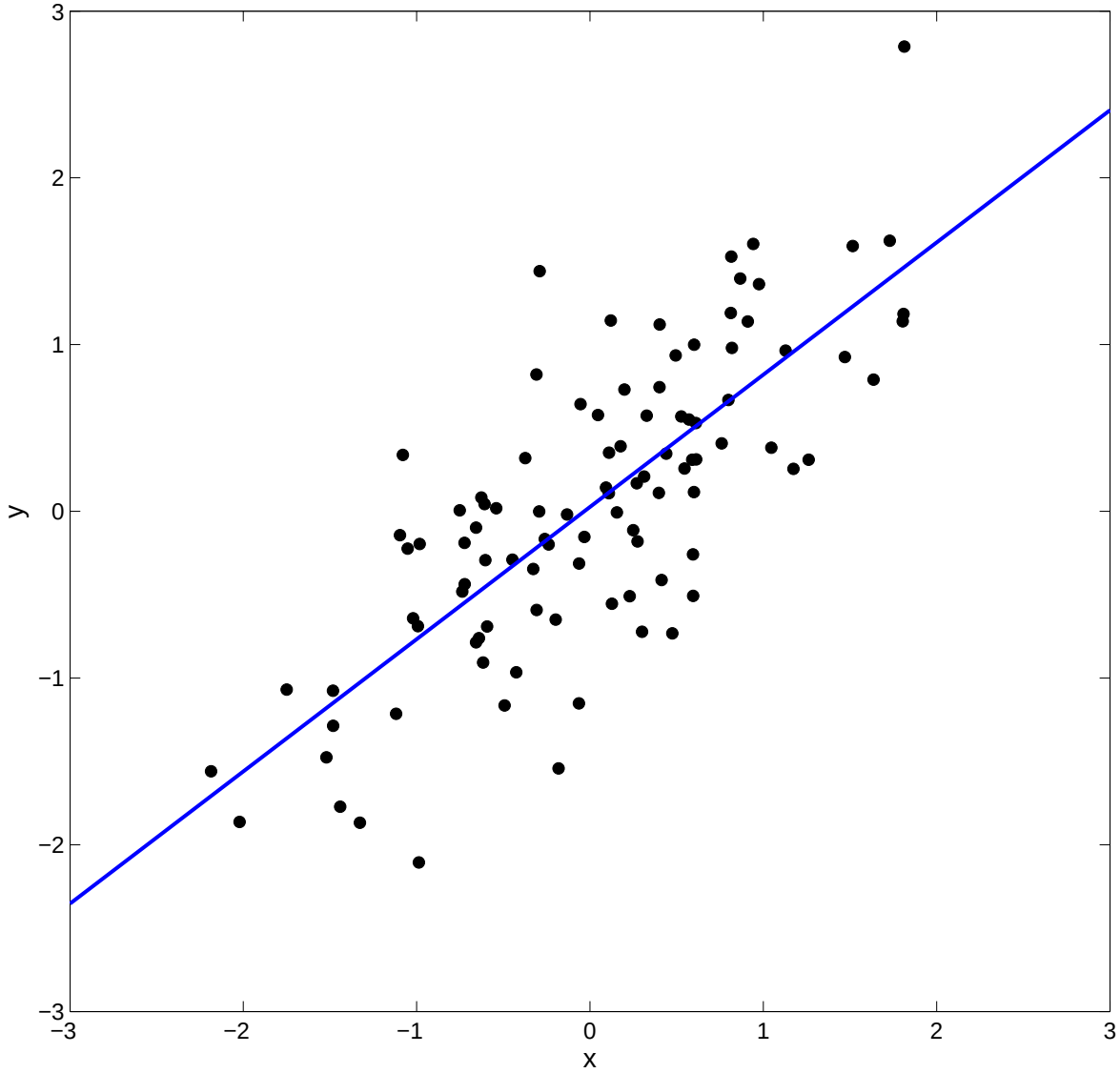
$$\|y - X\theta\|_2^2 \to \min_{\theta};$$

$$2X^T (y - X\theta) = 0,$$

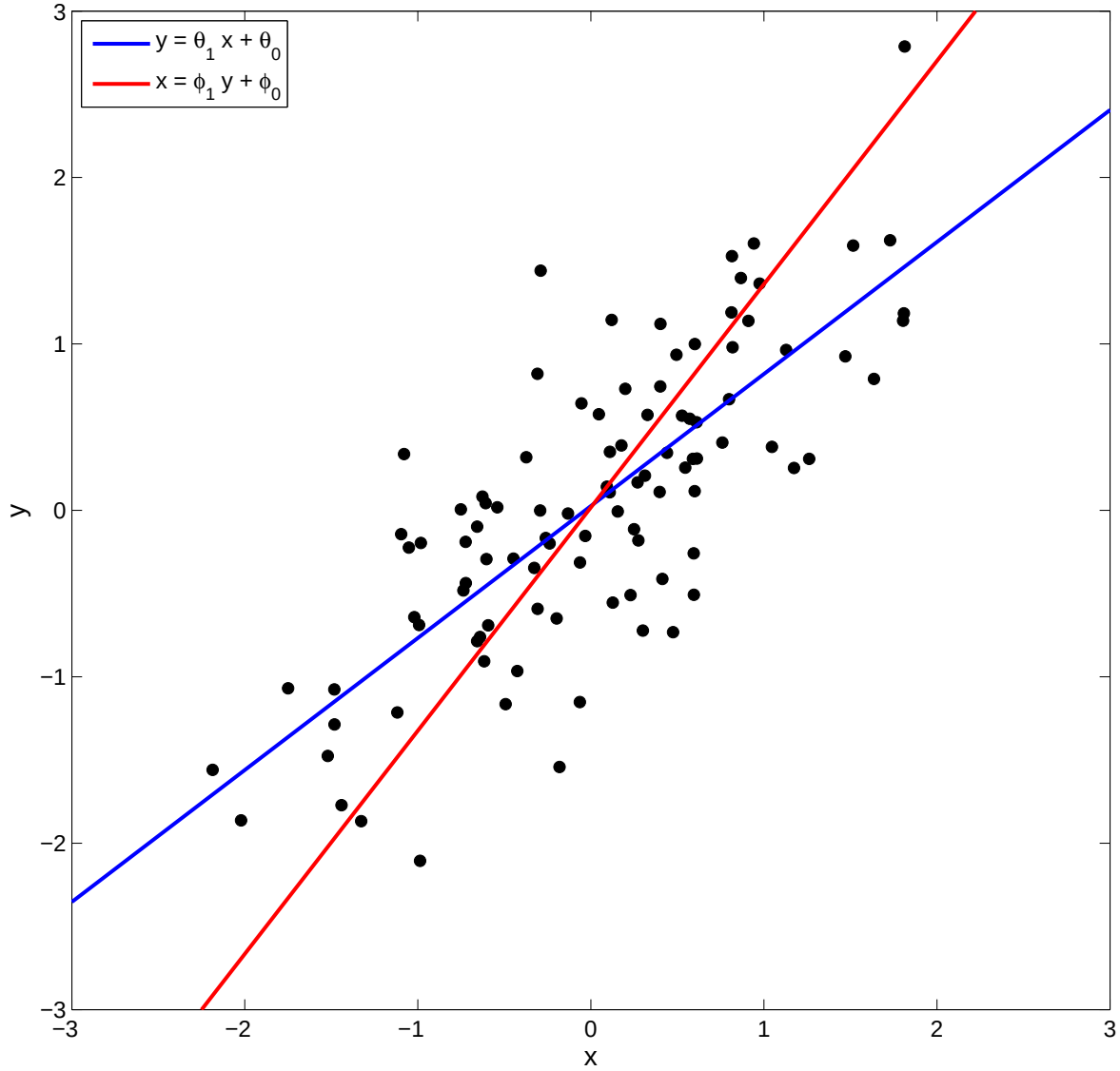
$$\hat{\theta} = \left(X^T X \right)^{-1} X^T y,$$

$$\hat{y} = X \left(X^T X \right)^{-1} X^T y.$$

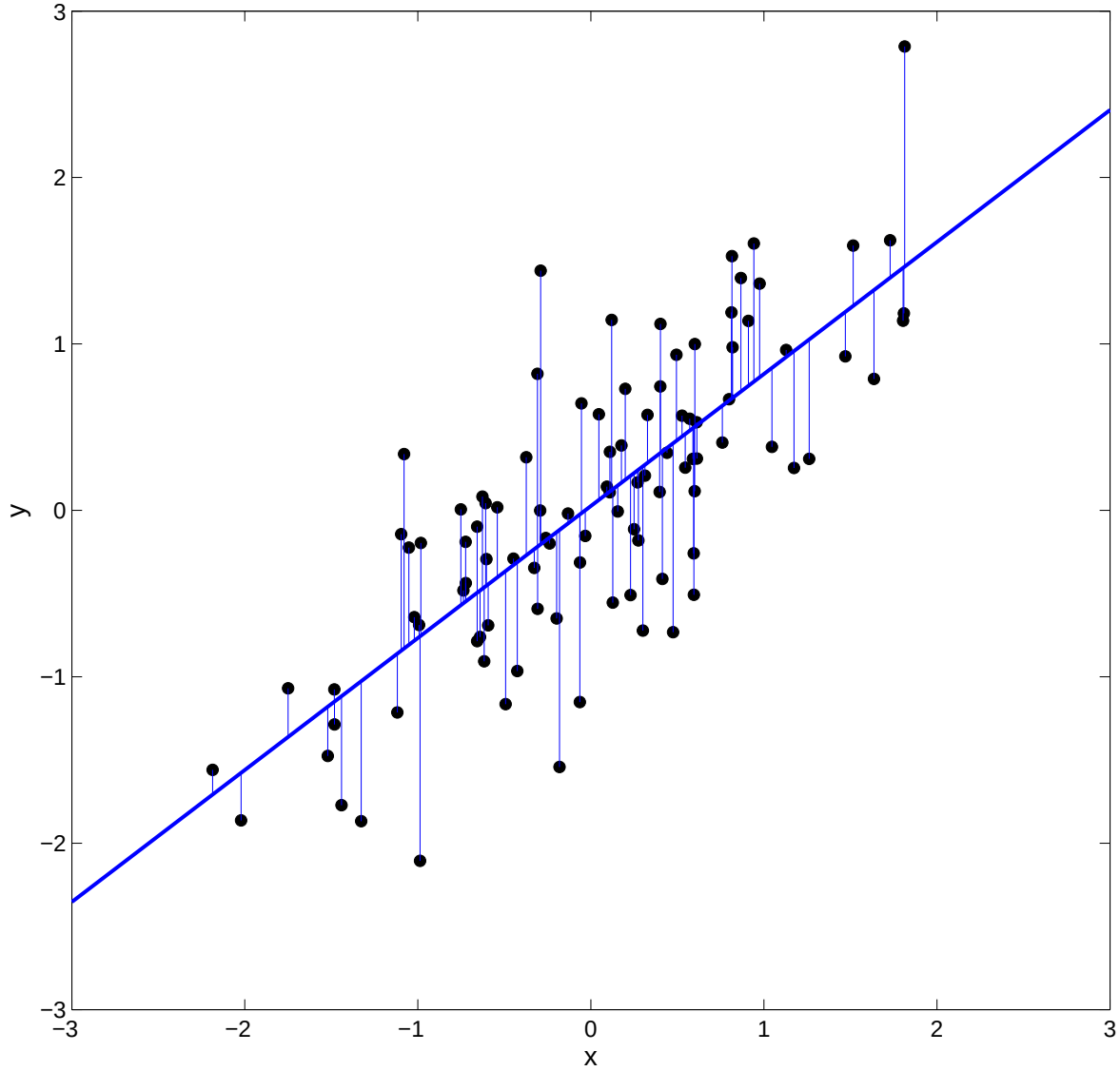
Инверсия задачи регрессии



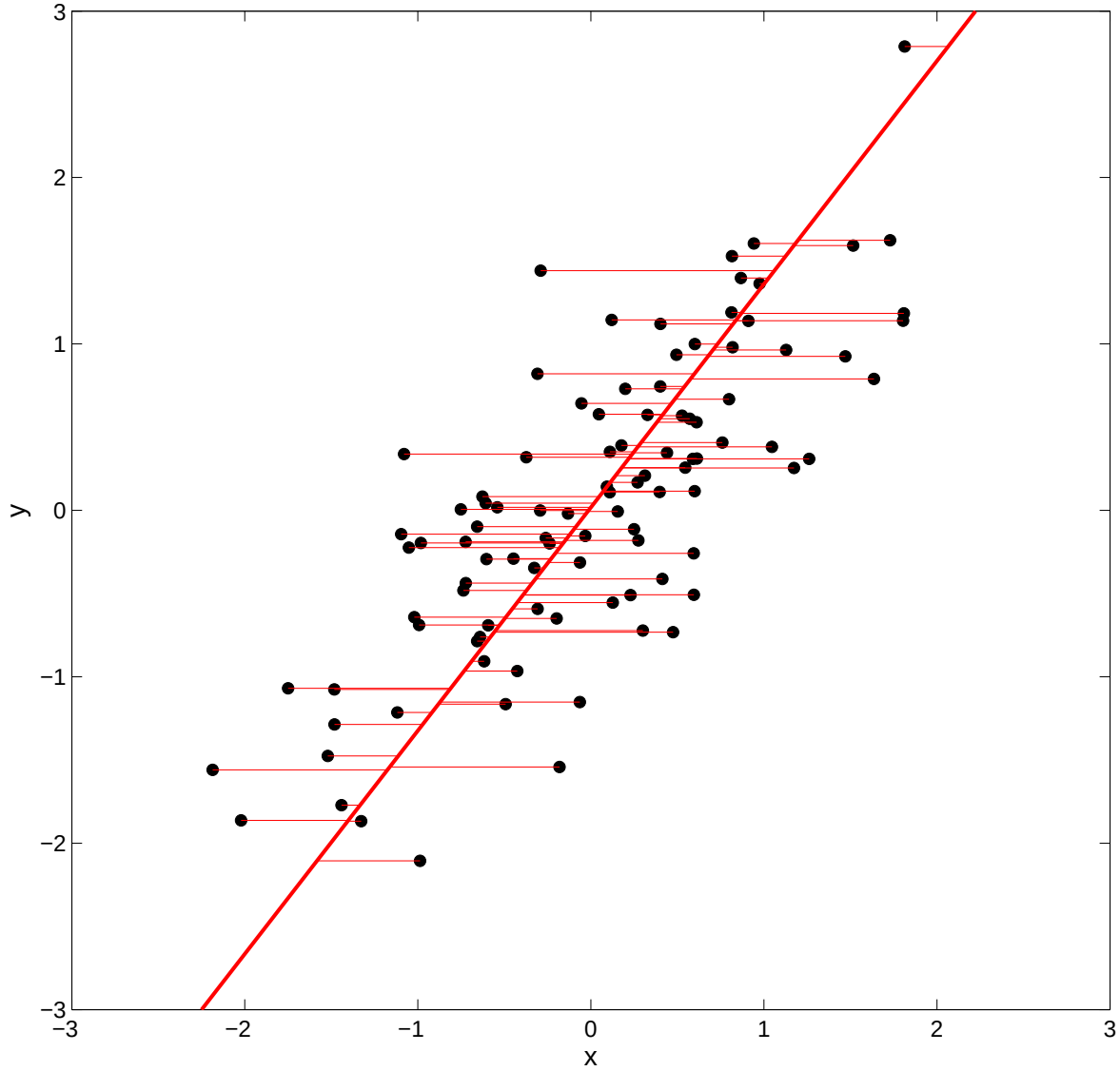
Инверсия задачи регрессии



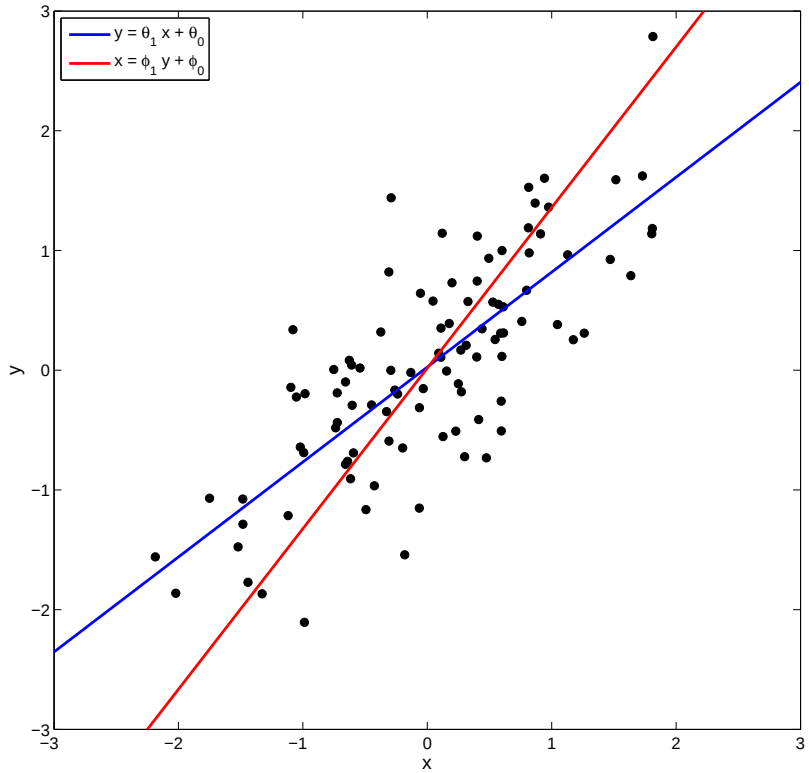
Инверсия задачи регрессии



Инверсия задачи регрессии



Инверсия задачи регрессии



- Две прямые пересекаются в точке $(\bar{x}, \bar{y})$.
- Косинус угла между прямыми, осуществляющими линейную МНК-регрессию x на y и y на x, равен значению выборочного коэффициента корреляции между x и y.

Goodness-of-fit

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(Total Sum of Squares);}$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{(Explained Sum of Squares);}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{(Residual Sum of Squares);}$$

$$TSS = ESS + RSS.$$

Коэффициент детерминации:

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

$R^2 = corr^2(y, \hat{y})$ — квадрат коэффициента множественной корреляции y с X.

Предположения модели

1. Линейность: $y = X\theta + \varepsilon$.
2. Случайность выборки: имеется независимая выборка наблюдений $(x_i, y_i), i = 1, \ldots, n$.
3. Полнота ранга: ни в популяции, ни в выборке ни один из признаков не является линейной комбинацией других $(\operatorname{rank} X = k)$.
4. Случайность ошибок: $\mathbb{E}(\varepsilon | X) = 0$.

В предположениях (1-4) МНК-оценки коэффициентов θ являются несмещёнными:

$$\mathbb{E}\hat{\theta}_j = \theta_j, \ j = 0, \ldots, k,$$

и состоятельными:

$$\forall \gamma > 0 \ \lim_{n \to \infty} P\left(\left|\theta_j - \hat{\theta}_j\right| < \gamma\right) = 1, \ j = 0, \ldots, k.$$

Предположения модели

1. Линейность: $y = X\theta + \varepsilon$.
2. Случайность выборки: имеется независимая выборка наблюдений $(x_i, y_i), i = 1, \ldots, n$.
3. Полнота ранга: ни в популяции, ни в выборке ни один из признаков не является константой; кроме того, ни один из признаков не является линейной комбинацией других ($\operatorname{rank} X = k$).
4. Случайность ошибок: $\mathbb{E}(\varepsilon | X) = 0$.
5. Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков: $\mathbb{D}(\varepsilon | X) = \sigma^2$.

(предположения Гаусса-Маркова).

Теорема Гаусса-Маркова: в предположениях (1-5) МНК-оценки имеют наименьшую дисперсию в классе оценок θ, линейных по y.

Дисперсия $\hat{\theta}_j$

В предположениях (1-5) дисперсии МНК-оценок коэффициентов θ задаются следующим образом:

$$\mathbb{D}\hat{\theta}_j = \frac{\sigma^2}{TSS_j \left(1 - R_j^2\right)},$$

где $TSS_j = \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2$, R_j^2 — коэффициент детерминации при регрессии x_j на все остальные признаки из X.

- Чем больше дисперсия ошибки σ^2, тем больше дисперсия оценки $\hat{\theta}_j$.
- Чем больше вариация значений признака x_j в выборке, тем меньше дисперсия оценки $\hat{\theta}_j$.
- Чем лучше признак x_j объясняется линейной комбинацией оставшихся признаков, тем больше дисперсия оценки $\hat{\theta}_j$.

$R_j^2 < 1$ по предположению (3); тем не менее, может быть $R_j^2 \approx 1$.

Дисперсия $\hat{\theta}_j$

В матричном виде:

$$\operatorname{cov} \hat{\theta} = \sigma^2 \left(X^T X \right)^{-1}.$$

Если столбцы X почти линейно зависимы, то матрица $X^T X$ плохо обусловлена, и дисперсия оценок $\hat{\theta}_j$ велика.

Близкая к линейной зависимость между двумя или более признаками x_j называется **мультиколлинеарностью**.

Неправильное определение модели

Недоопределение: если зависимая переменная определяется моделью

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_{j-1} x_{j-1} + \theta_j x_j + \theta_{j+1} x_{j+1} + \cdots + \theta_k x_k,$$

а вместо этого используется модель

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_{j-1} x_{j-1} + \theta_{j+1} x_{j+1} + \cdots + \theta_k x_k,$$

то МНК-оценки $\hat{\theta}_0, \ldots, \hat{\theta}_{j-1}, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_k$ являются смещёнными и несостоятельными оценками $\theta_0, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k$.

Переопределение: если признак x_j не влияет на y, т. е. $\theta_j = 0$, то МНК-оценка $\hat{\theta}$ остаётся несмещённой состоятельной оценкой θ, но дисперсия её возрастает.

Бинарные признаки

Если x_j принимает только два значения, то они кодируются нулём и единицей. Например, если x_j — пол испытуемого, то можно задать $x_j = [\text{пол} = \text{мужской}]$.

Механизм построения регрессии не меняется.

Категориальные признаки

Как кодировать дискретные признаки x_j, принимающие более двух значений?

Пусть y — средний уровень заработной платы, x — тип должности (рабочий / инженер / управляющий). Допустим, мы закодировали эти должности следующим образом:

Тип должности	x
рабочий	1
инженер	2
управляющий	3

и построили регрессию $y = \theta_0 + \theta_1 x$. Тогда для рабочего, инженера и управляющего ожидаемые средние уровни заработной платы определяются следующим образом:

$$y_{bc} = \theta_0 + \theta_1,$$
$$y_{pr} = \theta_0 + 2\theta_1,$$
$$y_{wc} = \theta_0 + 3\theta_1.$$

Согласно построенной модели, разница в средних уровнях заработной платы рабочего и инженера в точности равна разнице между зарплатами инженера и управляющего.

Фиктивные переменные

Верный способ использования категориальных признаков в регрессии — введение бинарных фиктивных переменных (dummy variables).
Пусть признак x_j принимает m различных значений, тогда для его кодирования необходима $m-1$ фиктивная переменная.

Способы кодирования:

	Dummy		Deviation	
Тип должности	x_1	x_2	x_1	x_2
рабочий	0	0	1	0
инженер	1	0	0	1
управляющий	0	1	-1	-1

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

- При dummy-кодировании коэффициенты θ_1, θ_2 оценивают среднюю разницу в уровнях зарплат инженера и управляющего с рабочим.
- При deviation-кодировании коэффициенты θ_1, θ_2 оценивают среднюю разницу в уровнях зарплат инженера и управляющего со средним по всем должностям.

Вопросы

1. Как найти доверительные интервалы для θ_j и проверить гипотезу $H_0 \colon \theta_j = 0$?
2. Как найти доверительный интервал для значений отклика на новом объекте $y(x_0)$?
3. Как проверить адекватность построенной модели?

Предположение о нормальности ошибок

6. Нормальность ошибок: $\varepsilon \sim N\left(0, \sigma^2\right)$.

- В предположениях (1-6) МНК-оценки совпадают с оценками максимального правдоподобия.

ММП:

$$p\left(\varepsilon_i\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma}\varepsilon_i^2},$$

$$\ln \prod_{i=1}^{n} p\left(\varepsilon_i\right) \to \max_{\theta},$$

$$\sum_{i=1}^{n} \left(-\frac{1}{2} \ln\left(2\pi\sigma\right) - \frac{1}{2\sigma}\varepsilon_i^2 \right) \to \max_{\theta},$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \theta_j x_{ij} \right)^2 \to \min_{\theta}.$$

Предположение о нормальности ошибок

- Эквивалентная запись предположения (6):

$$y\,|X \sim N\left(X\theta, \sigma^2\right).$$

- МНК-оценки $\hat{\theta}$ имеют наименьшую дисперсию среди всех несмещённых оценок θ.
- МНК-оценки $\hat{\theta}$ имеют нормальное распределение $N\left(\theta, \sigma^2\left(X^TX\right)^{-1}\right)$.
- Несмещённой оценкой σ^2 является

$$\hat{\sigma}^2 = \frac{1}{n-k-1}RSS;$$

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

- $\forall c \in \mathbb{R}^{k+1}$

$$\frac{c^T\left(\theta - \hat{\theta}\right)}{\hat{\sigma}\sqrt{c^T\left(X^TX\right)^{-1}c}} \sim St(n-k-1).$$

Доверительные интервалы

$100(1-\alpha)\%$ доверительный интервал для σ:

$$\sqrt{\frac{RSS}{\chi^2_{n-k-1,1-\alpha/2}}} \leq \sigma \leq \sqrt{\frac{RSS}{\chi^2_{n-k-1,\alpha/2}}}.$$

Возьмём $c = \left(0 \dots 0 \underset{j}{1} 0 \dots 0\right)$; $100(1-\alpha)\%$ доверительный интервал для θ_j:

$$\hat{\theta}_j \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}.$$

Для нового объекта x_0 возьмём $c = x_0$; $100(1-\alpha)\%$ доверительный интервал для $y(x_0)$:

$$x_0^T\hat{\theta} \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{x_0^T\left(X^TX\right)^{-1}x_0}.$$

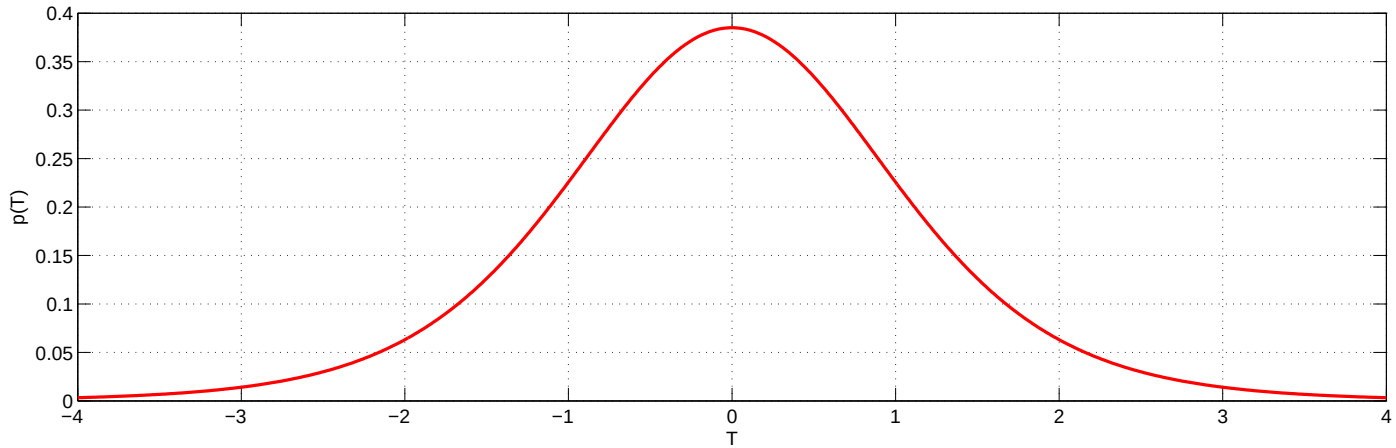
t-критерий Стьюдента

нулевая гипотеза: $H_0 : \theta_j = 0;$

альтернатива: $H_1 : \theta_j <\neq> 0;$

статистика: $T = \dfrac{\hat{\theta}_j}{\frac{RSS}{n-k-1}\sqrt{\left(X^T X\right)^{-1}_{jj}}};$

$T \sim St(n-k-1)$ при $H_0;$



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - tcdf(t, n-k-1), & H_1 : \theta_j > 0, \\ tcdf(t, n-k-1), & H_1 : \theta_j < 0, \\ 2\left(1 - tcdf(|t|, n-k-1)\right), & H_1 : \theta_j \neq 0. \end{cases}$$

t-критерий Стьюдента

Пример: имеется 12 испытуемых, x — результат прохождения испытуемым составного теста скорости реакции, y — результат его теста на симулятора транспортного средства. Проведение составного теста значительно проще и требует меньших затрат, поэтому ставится задача предсказания y по x, для чего строится линейная регрессия согласно модели

$$y = \theta_0 + \theta_1 x + \varepsilon.$$

Значима ли переменная x для предсказания y?

$H_0 : \theta_1 = 0.$
$H_1 : \theta_1 \neq 0 \Rightarrow p = 2.2021 \times 10^{-5}.$

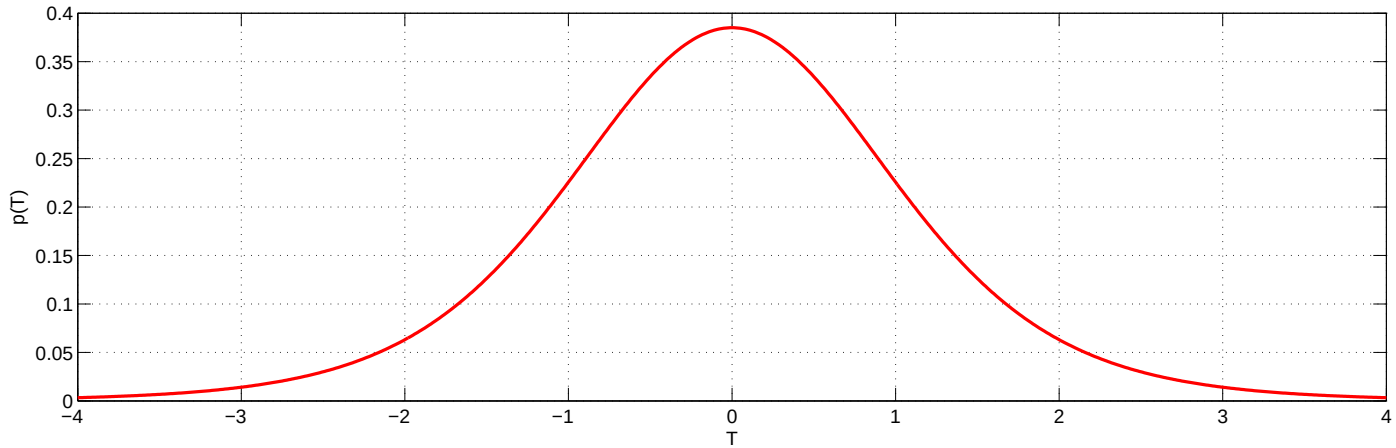
t-критерий Стьюдента

нулевая гипотеза: $H_0 : \theta_j = a;$

альтернатива: $H_1 : \theta_j <\neq> a;$

статистика: $T = \dfrac{\hat{\theta}_j - a}{\frac{RSS}{n-k-1}\sqrt{\left(X^T X\right)^{-1}_{jj}}};$

$T \sim St(n-k-1)$ при $H_0;$



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - tcdf(t, n-k-1), & H_1 : \theta_j > a, \\ tcdf(t, n-k-1), & H_1 : \theta_j < a, \\ 2\left(1 - tcdf(|t|, n-k-1)\right), & H_1 : \theta_j \neq a. \end{cases}$$

t-критерий Стьюдента

Пример: по выборке из 506 жилых районов, расположенных в пригородах Бостона, строится модель средней цены на жильё следующего вида:

$$\ln price = \theta_0 + \theta_1 \ln nox + \theta_2 \ln dist + \theta_3 rooms + \theta_4 stratio + \varepsilon,$$

где nox — содержание в воздухе двуокиси азота, dis — взвешенное среднее расстояние от жилого района до пяти основных мест трудоустройства, $rooms$ — среднее число комнат в доме жилого района, $stratio$ — среднее отношения числа студентов к числу учителей в школах района.
Коэициент θ_1 имеет смысл эластичности цены по признаку nox. По экономическим соображениям интерес представляет гипотеза о том, что эластичность равна -1.

$H_0 : \theta_1 = -1.$
$H_1 : \theta_1 \neq -1 \Rightarrow p = 0.6945.$

Критерий Фишера

$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2}\right); \quad \underset{(k+1)\times 1}{\theta^T} = \left(\underset{(k+1-k_1)\times 1}{\theta_1^T}, \underset{k_1\times 1}{\theta_2^T}\right)^T;$$

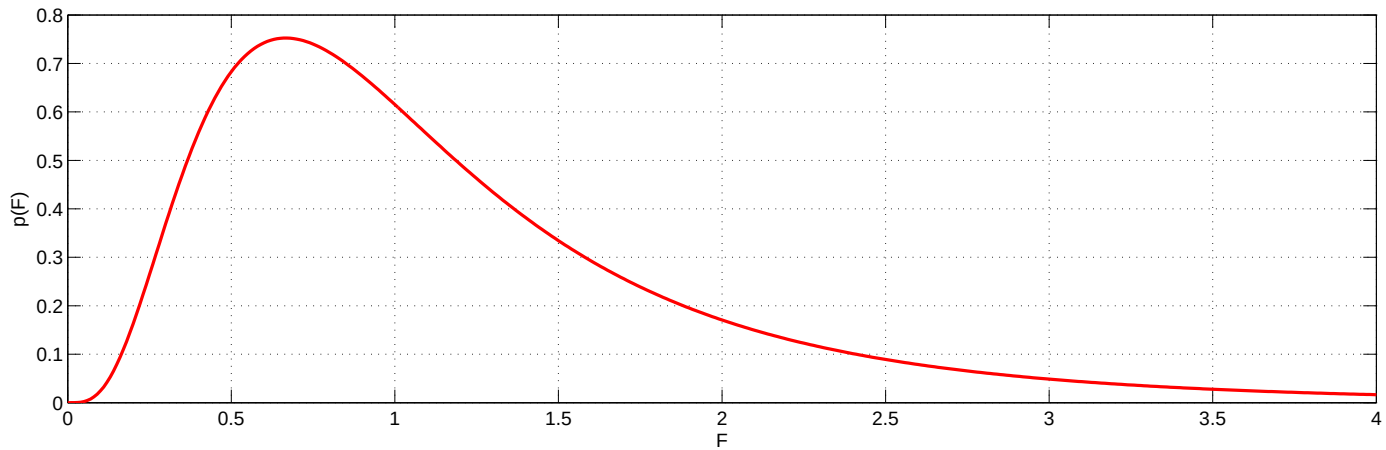
нулевая гипотеза: $H_0 : \theta_2 = 0;$

альтернатива: $H_1 : H_0$ неверна;

статистика: $RSS_r = \|y - X_1\theta_1\|_2^2, \quad RSS_{ur} = \|y - X\theta\|_2^2,$

$$F = \frac{(RSS_r - RSS_{ur})/k_1}{RSS_{ur}/(n-k-1)};$$

$F \sim F(k_1, n-k-1)$ при H_0;



достигаемый уровень значимости:

$$p(f) = fcdf(1/f, n-k-1, k_1).$$

Критерий Фишера

Пример: для веса ребёнка при рождении имеется следующая модель:

$$weight = \theta_0 + \theta_1 cigs + \theta_2 parity + \theta_3 inc + \theta_4 med + \theta_5 fed + \varepsilon,$$

где $cigs$ — среднее число сигарет, выкуривавшихся матерью за один день беременности, $parity$ — номер ребёнка у матери, inc — среднемесячный доход семьи, med — длительность в годах получения образования матерью, fed — отцом. Данные имеются для 1191 детей.
Зависит ли вес ребёнка при рождении от уровня образования родителей?

$H_0 : \theta_4 = \theta_5 = 0.$
$H_1 : H_0$ неверна $\Rightarrow p = 0.2421.$

Связь между критериями Фишера и Стьюдента

Если $k_1 = 1$, критерий Фишера эквивалентен критерию Стьюдента для двусторонней альтернативы.

Иногда критерий Фишера отвергает гипотезу о незначимости признаков X_2, а критерий Стьюдента не признаёт значимым ни один из них. Возможные объяснения:

- отдельные признаки из X_2 недостаточно хорошо объясняют y, но совокупный эффект значим;
- признаки в X_2 мультиколлинеарны.

Иногда критерия Фишера не отвергает гипотезу о незначимости признаков X_2, а критерий Стьюдента признаёт значимыми некоторые из них. Возможные объяснения:

- незначимые признаки в X_2 маскируют влияние значимых;
- значимость отдельных признаков в X_2 — результат множественной проверки гипотез.

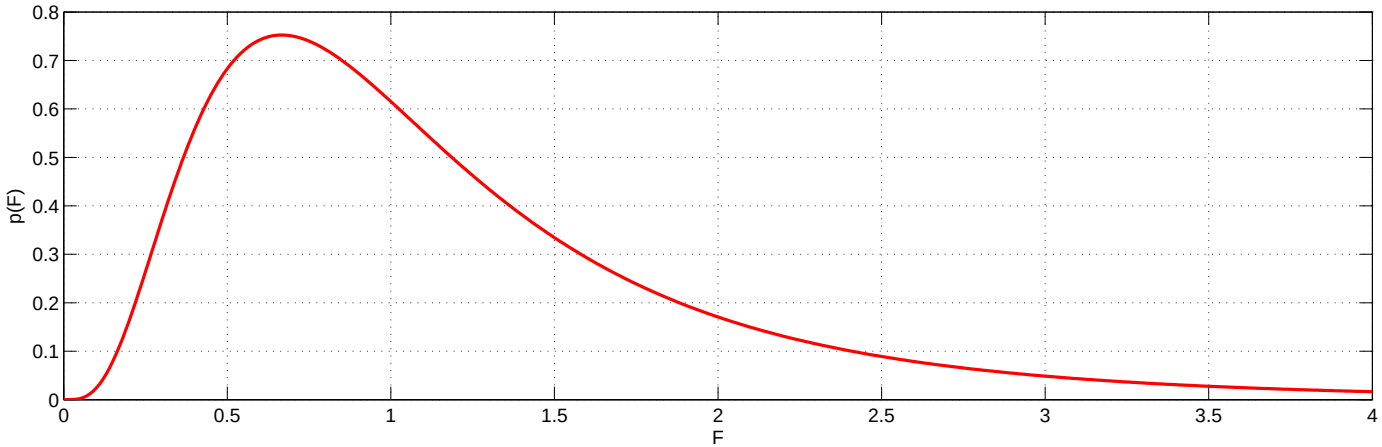
Критерий Фишера

нулевая гипотеза: $H_0 : \theta_1 = \cdots = \theta_k = 0;$

альтернатива: $H_1 : H_0$ неверна;

статистика: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)};$

$F \sim F(k, n-k-1)$ при $H_0;$



достигаемый уровень значимости:

$$p(f) = fcdf(1/f, n-k-1, k).$$

Критерий Фишера

Пример: имеет ли вообще смысл модель веса ребёнка при рождении, рассмотренная выше?

$H_0 : \theta_1 = \cdots = \theta_5 = 0.$
$H_1 : H_0$ неверна $\Rightarrow p = 6.0331 \times 10^{-9}$.

Приведённый коэффициент детерминации

Стандартный коэффициент детерминации всегда увеличивается при добавлении регрессоров в модель, поэтому для отбора признаков его использовать нельзя.

Для сравнения моделей, содержащих разное число признаков, можно использовать приведённый коэффициент детерминации:

$$R_a^2 = \frac{ESS/(n-k-1)}{TSS/(n-1)} = 1 - \left(1 - R^2\right) \frac{n-1}{n-k-1}.$$

Пошаговая регрессия

- **Шаг 0**. Настраивается модель с одной только константой, а также все модели с одной переменной. Рассчитывается F-статистика каждой модели и достигаемый уровень значимости. Выбирается модель с наименьшим достигаемым уровнем значимости. Соответствующая переменная X_{e1} включается в модель, если этот достигаемый уровень значимости меньше порогового значения $p_E = 0.05$.
- **Шаг 1**. Рассчитывается F-статистика и достигаемый уровень значимости для всех моделей, содержащих две переменные, одна из которых X_{e1}. Аналогично принимается решение о включении X_{e2}.
- **Шаг 2**. Если была добавлена переменная X_{e2}, возможно, X_{e1} уже не нужна. В общем случае просчитываются все возможные варианты исключения одной переменной, рассматривается вариант с наибольшим достигаемым уровнем значимости, соответствующая переменная исключается, если он превосходит пороговое значение $p_R = 0.1$.
- ...

Эксперимент Фридмана

David A. Freedman, A Note on Screening Regression Equations. The American Statistician, Vol. 37, No. 2 (May, 1983), pp. 152-155.

Отбор признаков с учётом эффекта множественной проверки гипотез

Прикладная статистика
7. Регрессионный анализ.

Рябенко Евгений
riabenko.e@gmail.com