Яндекс

Skoltech
Skolkovo Institute of Science and Technology

Линейная регрессия. Выбор модели.

ПМИ ФКН ВШЭ, 23 ноября 2019 г.

Денис Деркач, Алексей Артёмов

ФКН ВШЭ

Оглавление

Регрессия

Стандартная линейная регрессия

Прогнозирование

Оценка качества регрессии

Множественная регрессия

Выбор модели

Регрессия

Регрессия

Регрессия — метод изучения зависимости между откликом Y и регрессором X (признак, независимая переменная).

Один из способов оценить зависимость:

$$r(x) = \mathbb{E}(Y|X = x) = \int y\, f(y|x)dy.$$

Задача состоит в том, чтобы построить оценку $\hat{r}(x)$ функции $r(x)$ по данным

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y},$$

где $F_{X,Y}$ — совместное распределение X и Y.

Стандартная линейная регрессия

Регрессия

Стандартная линейная регрессия

Прогнозирование

Оценка качества регрессии

Множественная регрессия

Выбор модели

Линейная регрессия

Линейная регрессия:

$$r(x) = \beta_0 + \beta_1 x.$$

Определение: простая линейная регрессия

Модель $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, где ε_i — шум с мат. ожиданием $\mathbb{E}(\varepsilon_i | X_i) = 0$ и дисперсией $\mathbb{V}ar(\varepsilon_i | X_i) = \sigma^2$, называется простой линейной регрессией.

Оценивание параметров:

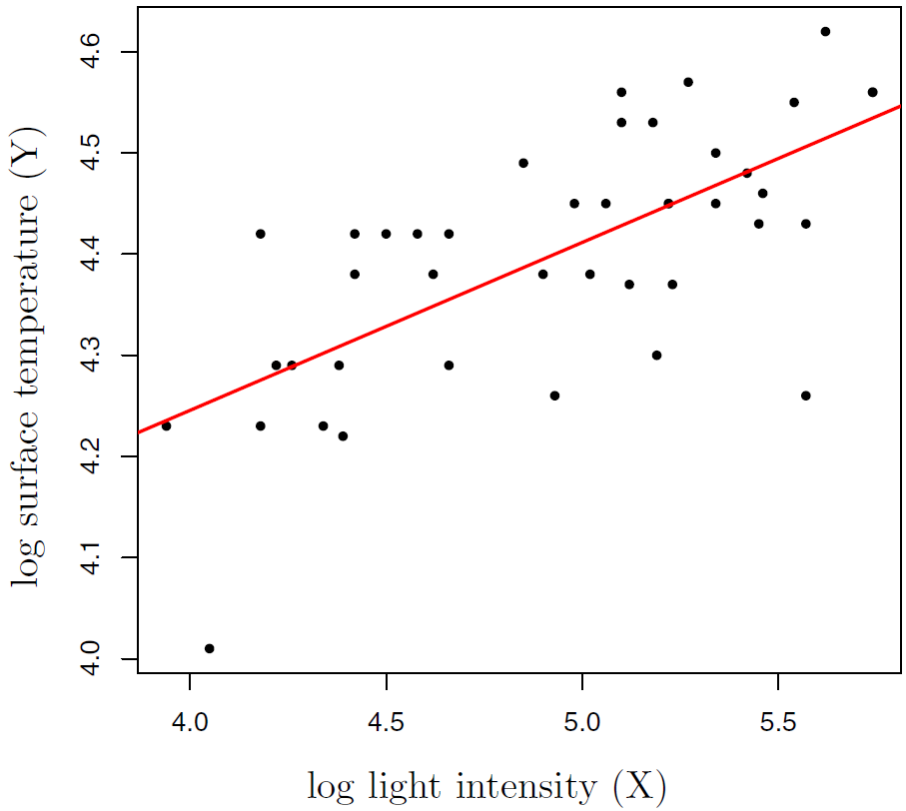
$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Предсказанные значения:

$$\hat{Y}_i = \hat{r}(X_i).$$

Примеры: линейная регрессия

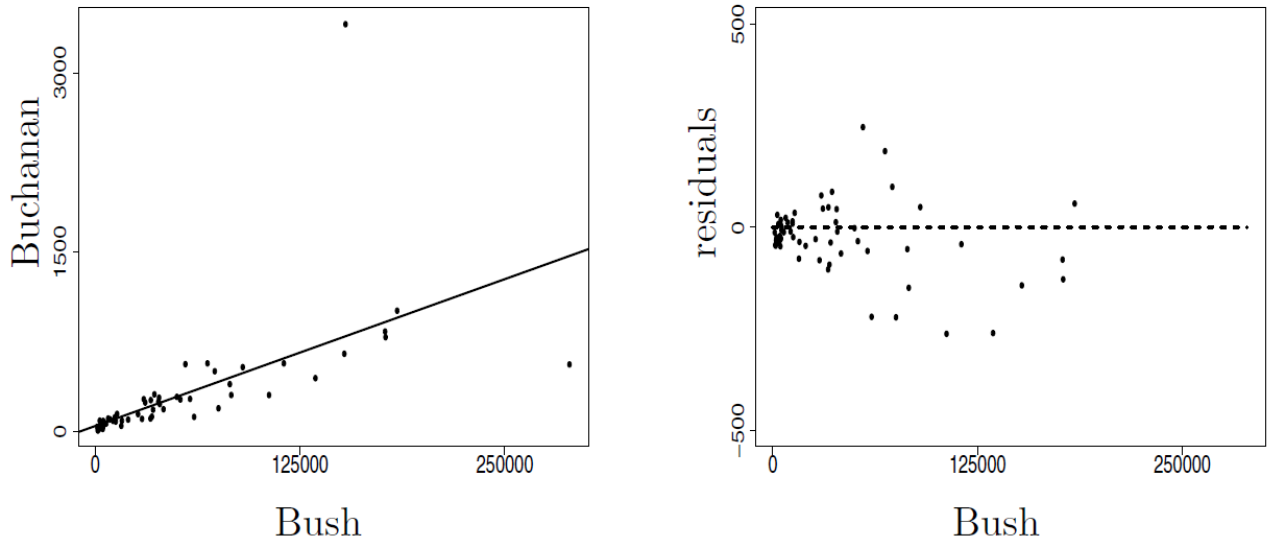
Данные о близлежащих звездах: оценка температуры звезды по её яркости.



Оценки равны: $\hat{\beta}_0 = 3.58$ и $\hat{\beta}_1 = 0.166 \Rightarrow \hat{r}(x) = 3.58 + 0.166x.$

Примеры: стандартная линейная регрессия

Голоса за Buchanan (Y) vs. голоса за Bush (X) во Флориде. Справа на графике указана величина отклонения от прогноза. Гауссовское распределение отклонений будет скорее всего говорить о том, что прогноз выбран правильно.



$$\hat{\beta}_0 = 66.0991, \;\; \hat{se}(\hat{\beta}_0) = 17.2926,$$

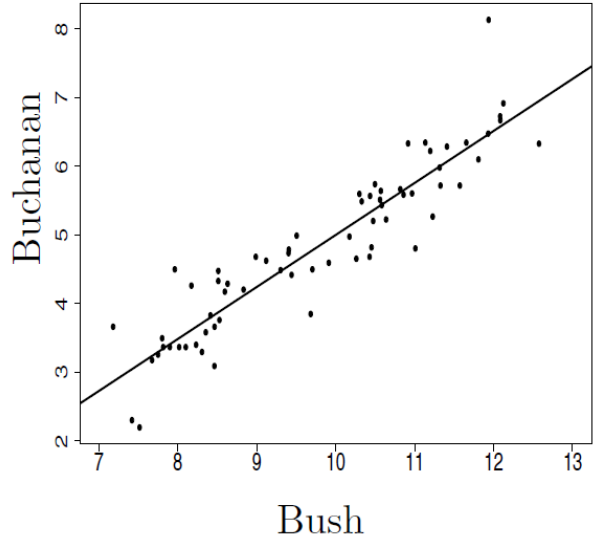
Примеры: стандартная линейная регрессия

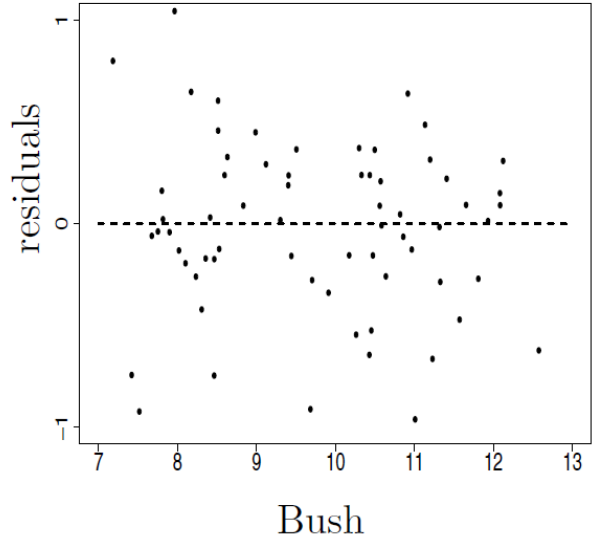
Если прологарифмировать данные, то остатки сильнее будут “напоминать” случайные числа:

$$\hat{\beta}_0 = -2.3298, \;\; \hat{se}(\hat{\beta}_0) = 0.3529,$$

$$\hat{\beta}_1 = 0.7303, \;\; \hat{se}(\hat{\beta}_0) = 0.0358,$$

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush}).$$





Метод наименьших квадратов

Остатки регрессии:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Сумма квадратов остатков (RSS):

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Определение

$\hat{\beta}_0$ и $\hat{\beta}_1$ — оценки неизвестных параметров с помощью метода наименьших квадратов (МНК), если RSS для этих оценок минимальна.

Оценки МНК

Теорема

Оценки параметров β_0 и β_1 с помощью метода наименьших квадратов имеют вид

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}, \text{ где}$$

$$s^2_{xx} = \sum_{i=1}^{n} (X_i - \overline{X})^2 \text{ и } s^2_{xy} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

Остатки МНК

Для неизвестных параметров β_0 и β_1 и $\varepsilon \sim \mathcal{N}(0; \sigma^2)$ несмещённая оценка дисперсии шума σ^2 равна

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

$n - 2$ степени свободы получаются за счёт наличия двух дополнительных связей (β_0, β_1).

Оценки МНК

Пусть $\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_1)^T$ — оценка метода наименьших квадратов. Тогда

$$\mathbb{E}(\hat{\beta}) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},$$

Можно показать, что:

$$\mathbb{V}ar(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{s_{xx}^2} \right), \qquad \mathbb{V}ar(\hat{\beta}_1) = \frac{\sigma^2}{s_{xx}^2}.$$

при $s_{xx}^2 = \sum_{i=1}^n (X_i - \overline{X})^2$.

Оценки дисперсии в МНК

Таким образом,

$$\widehat{\mathbb{V}ar}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{s_{xx}^2} \right), \qquad \widehat{\mathbb{V}ar}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{s_{xx}^2}.$$

Так как ε не зависит от $\hat{\beta}$, то $\widehat{\mathbb{V}ar}(\hat{\beta}_0)$ и $\widehat{\mathbb{V}ar}(\hat{\beta}_0)$ тоже не зависят от $\hat{\beta}$. Мы можем написать для $H_0 : \beta_1 = \beta_1^0$:

$$T = \frac{\hat{\beta}_1 - \beta_1^0}{\sqrt{\widehat{\mathbb{V}ar}\hat{\beta}_1}} \sim t_{n-2}.$$

Пример: критерий Вальда

Замечание

Критерий Вальда для проверки $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ имеет вид $W = \frac{\hat{\beta} - \beta_0}{\hat{se}(\hat{\beta})}$.

Пример

(Выборы) Для регрессии (в логарифмическом масштабе) 95% доверительный интервал имеет вид

$$0.7303 + 2 \times 0.0358 = (0.66, 0.80).$$

Статистика Вальда для проверки $H_0 : \beta_1 = 0$ против альтернативы $H_1 : \beta_1 \neq 0$ равна $|W| = |.7303 - 0|/.0358 = 20.40$. Причем $p - value$ равно $\mathbb{P}(|Z| > 20.40) \approx 0 \Rightarrow$ зависимость действительно существует.

Доверительные интервалы для коэффициента β_1

- Можем написать

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{s_x x^2}}$$

- Тогда:

$$1 - \alpha = \mathbb{P}\left(\left|\frac{\hat{\beta}_1 - \beta_1}{SE(\beta_1}\right| < t_{n-2,1-\alpha/2}\right)$$

- то есть интервал с уровнем доверия $1 - \alpha$ будет записан:

$$\left(\hat{\beta}_1 - t_{n-2,1-\alpha/2} SE(\hat{\beta}_1); \hat{\beta}_1 + t_{n-2,1-\alpha/2} SE(\hat{\beta}_1)\right)$$

- заметим, что при известной σ мы можем заменить t на z.

Линейная комбинация β_1 и β_0

Может быть также показано:

$$SE(a_0\hat{\beta}_0 + a_1\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{a_0^2}{n} + \frac{(a_0\bar{X} - a_1)^2}{s_{xx}^2}}$$

То есть, границы интервалов будут заданы приблизительно также:

$$a_0\hat{\beta}_0 + a_1\hat{\beta}_1 \pm t_{n-2,1-\alpha/2}SE(a_0\hat{\beta}_0 + a_1\hat{\beta}_1).$$

NB: приблизительно, так как нам необходимо точно подсчитать степени свободы.

Прогнозирование

Регрессия

Стандартная линейная регрессия

Прогнозирование

Оценка качества регрессии

Множественная регрессия

Выбор модели

Прогнозирование

Модель — $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, построенная по выборке данных $(X_1, Y_1), \ldots, (X_n, Y_n)$.
Необходимо предсказать значение отклика Y_* при $X = x_*$:

$$\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*.$$

При этом модель выглядит:

$$Y_* = \beta_0 + \beta_1 x_* + \varepsilon_*$$

Прогнозирование: доверительные интервалы

Мы можем использовать результаты предыдущей секции для $a_0 = 1$ и $a_1 = x^*$:

$$SE(\hat{Y}_*) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\bar{X} - x_*)^2}{s_{xx}^2}}$$

Отсюда получается ошибка прогноза:

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t_{n-2,1-\alpha/2} SE(SE(\hat{Y}_*))$$

Пример: прогнозирование

Пример

1. Выборы

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush}).$$

2. В Palm Beach за Bush отдали 152 954 голосов, а за Buchanan — 3 476.
3. В логарифмической шкале это составляет 11.93789 и 8.151045 соответственно.
4. Насколько вероятен этот исход в предположении, что модель верна?
   - Предсказание для Buchanan равно -2.3298 + 0.7303 * 11.93789 = 6.388441.
5. Существенно ли это меньше, чем мы наблюдаем на практике?
   - $\hat{\xi}_n = 0.093775$ и 95% доверительный интервал имеет вид

Оценка качества регрессии

Адекватность модели

Как и в случае дисперсионного анализа, мы можем написать некоторое количество сумм квадратов:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2;$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2;$$

$$SSR = \sum_{i=1}^{n} (\bar{Y} - \hat{Y})^2;$$

Как и раньше: $SST = SSE + SSR$.

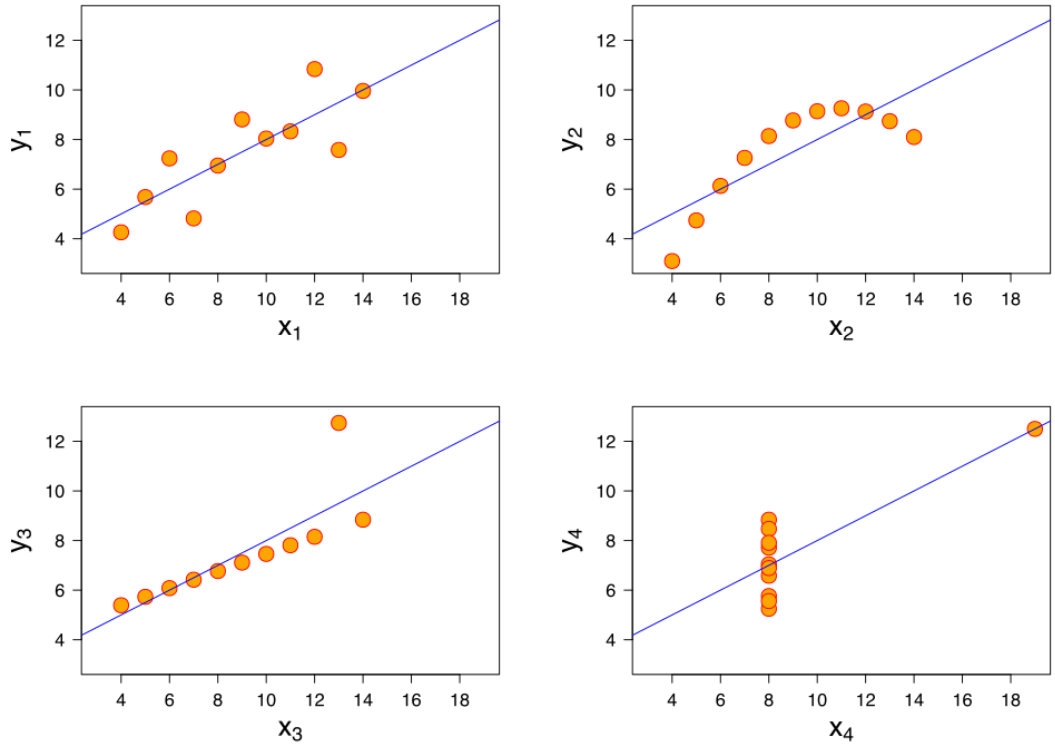
Коэффициент детерминации

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \widehat{\rho}^2(X, Y),$$

где ρ — коэффициент корреляции Пирсона.

- R^2 указывает на размер объяснённых изменений Y;
- $R^2 \leq 1$ и обычно больше 0;
- связан с MSE: $1 - R^2 = MSE / \mathbb{V}ar(Y)$;
- растёт при добавлении новых переменных, потому лучше использовать коррекции: $R^2_{\text{adj}} = 1 - \frac{SSE/dfe}{SST/dft}$, где $dfe = n - p - 1$, $dft = n - 1$, количество степеней свободы для выборки размером n и количеством объяснённых параметров p.

Квартет Энскомба



Все 4 семейства имеют одинаковое средние, дисперсии, уравнения регрессии, R^2.

Datasaurus:https://bit.ly/2wtDgyFl

Способы оценки качества регрессии

- Mean square (L2) loss (MSE): $\mathrm{MSE}(h, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - h(x_i))^2$
- Root MSE: $\mathrm{RMSE}(h, X^\ell) = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - h(x_i))^2}$
- Coefficient of determination: $R^2(h, X^\ell) = 1 - \frac{\sum_{i=1}^{\ell}(y_i - h(x_i))^2}{\sum_{i=1}^{\ell}(y_i - \mu_y)^2}$

  with $\mu_y = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i$
- Mean absolute error: $\mathrm{MAE}(h, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}|y_i - h(x_i)|$

NB: это способы оценки качества конкретной регрессии, а не способ сравнения регрессий.

Множественная регрессия

Регрессия

Стандартная линейная регрессия

Прогнозирование

Оценка качества регрессии

Множественная регрессия

Выбор модели

Множественная регрессия

В этом случае данные имеют вид

$$(X_1, Y_1), \ldots, (X_i, Y_i), \ldots, (X_n, Y_n),$$

$$X_i = (X_{i1}, \ldots, X_{ik}) \in \mathbb{R}^k.$$

Модель имеет вид $(i = 1, \ldots, n)$

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i,$$

$$\mathbb{E}(\varepsilon_i | X_{1i}, \ldots, X_{ki}) = 0.$$

Чтобы включить нулевой коэффициент, обычно полагают $X_{i1} = 1$ при $i = 1, \ldots, n$.

Множественная регрессия

Модель может быть выписана:

$$
\begin{aligned}
y_1 &= \beta_1 x_{11} + \dots \beta_d x_{d1} + \varepsilon_1, \\
y_2 &= \beta_1 x_{12} + \dots \beta_d x_{d2} + \varepsilon_2, \\
&\dots \\
y_\ell &= \beta_1 x_{1\ell} + \dots \beta_d x_{d\ell} + \varepsilon_\ell,
\end{aligned}
$$

или в матричной форме:

$$
\begin{bmatrix} y_1 \\ \vdots \\ y_\ell \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{d1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1\ell} & x_{2\ell} & \dots & x_{d\ell} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_\ell \end{bmatrix} \quad \longleftrightarrow \quad \boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}
$$

Множественная регрессия

Теорема

Для невыроженной матрицы $X^T X$ размера $k \times k$

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$
$$\mathbb{V}ar(\hat{\beta}|X^n) = \sigma^2 (X^T X)^{-1},$$
$$\hat{\beta} \approx \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}).$$

Оценка функции регрессии имеет вид

$$\hat{r}(x) = \sum_{j=1}^{k} \hat{\beta}_j x_j,$$

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{\varepsilon}_i^2, \qquad \hat{\varepsilon} = X\hat{\beta} - Y \ \text{— вектор остатков.}$$

Поведение оценок

- Оценки параметров:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$$

- Оценки дисперсии шума:

$$\sigma^2 = \frac{SSE}{n-p} \sim \chi^2_{n-p}$$

- По теореме Гаусса-Маркова в наших условиях оценки метода наименьших квадратов оптимальны в классе линейных несмещённых оценок.

Доверительные интервалы: множественная регрессия

Приближенный доверительный интервал размера $1 - \alpha$ для β_j равен

$$\hat{\beta}_j \pm z_{\alpha/2} \hat{se}(\hat{\beta}_j),$$

где $\hat{se}^2(\hat{\beta}_j)$ — j-ый диагональный элемент матрицы $\hat{\sigma}^2(X^T X)^{-1}$. Более точно использовать t статистику.

Пример: множественная регрессия

Пример

Данные о преступлениях по 47 штатам США в 1960г.
http://lib.stat.cmu.edu/DASL/Stories/USCrime.html

Регрессор	$\hat{\beta}_j$	$\hat{se}(\hat{\beta}_j)$	t-value	p-value
Нулевой коэффициент	-589.39	167.59	-3.51	0.001
Возраст	1.04	0.45	2.33	0.025
Южный штат(да/нет)	11.29	13.24	0.85	0.399
Образование	1.18	0.68	1.7	0.093
Расходы	0.96	0.25	3.86	0.000
Труд	0.11	0.15	0.69	0.493
Количество мужчин	0.30	0.22	1.36	0.181
Численность населения	0.09	0.14	0.65	0.518
Безработные (14-24)	-0.68	0.48	-1.4	0.165
Безработные (25-39)	2.15	0.95	2.26	0.030
Доход	-0.08	0.09	-0.91	0.367

Метод оценивания на основе максимизации правдоподобия

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

$$Y_i | X_i \sim \mathcal{N}(\mu_i, \sigma^2), \text{ где } \mu_i = \beta_0 + \beta_1 X_i.$$

Правдоподобие имеет вид

$$\prod_{i=1}^{n} f(X_i, Y_i) = \prod_{i=1}^{n} f_X(X_i) f_{Y|X}(Y_i|X_i) =$$

$$= \prod_{i=1}^{n} f_X(X_i) \times \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) = \mathcal{L}_1 \times \mathcal{L}_2,$$

$$\mathcal{L}_1 = \prod_{i=1}^{n} f_X(X_i), \qquad \mathcal{L}_2 = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i)$$

Метод оценивания на основе максимизации правдоподобия

Функция $\mathcal{L}_1$ не содержит параметры β_0 и β_1.

Рассмотрим $\mathcal{L}_2$ — условную функцию правдоподобия:

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) \propto \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (Y_i - \mu_i)^2 \right.$$

$$\ell(\beta_0, \beta_1, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2. \quad (1)$$

ОМП $(\beta_0, \beta_1) \Leftrightarrow$ максимизация (1)$\Leftrightarrow$ минимизация RSS,

$$RSS = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2.$$

Метод оценивания на основе максимизации правдоподобия

Теорема

В предположении нормальности ОМП оценка совпадает с оценкой метода наименьших квадратов.

Максимизируя $\ell(\beta_0, \beta_1, \sigma)$ по σ, получаем ОМП оценку

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i \hat{\varepsilon}_i^2.$$

Некоторые наблюдения

- Регрессия может сильно ошибаться при неправильном выборе предикторов.
- Зависимые переменные можно линеаризовать (например, взять логарифм, как в примере), но надо следить за погрешностями (кроме того, критерии качества относятся непосредственно к регрессии, а не к изначальным переменным.
- Негауссов шум или гетероскедастичный шум нарушают предположения регрессии, можно взвесить, но тогда все приведённые тесты нужно будет изменять.
- Выбросы могут сильно влиять на качество регрессии — при этом искать их надо с учётом того, что выбросы могут быть случайны.
- Некоторые точки дают больший вклад, чем другие — можно сделать тест, чтобы найти эти точки (не забывайте про множественное тестирование).

Выбор модели

Регрессия

Стандартная линейная регрессия

Прогнозирование

Оценка качества регрессии

Множественная регрессия

Выбор модели

Выбор модели

All models are wrong, but some are useful

George Box

Выбор модели

Бритва Оккама — не надо “плодить” сущности. Много переменных приводят к большой дисперсии прогноза, но маленькому смещению, и наоборот.

При выборе подходящей модели возникают две задачи:

1. выбор целевой функции для характеризации качества используемой модели;
2. поиск оптимальной модели согласно выбранному критерию качества.

Критерии выбора модели

В принципе, в статистике нет ответа, как ”правильно” выбирать модель.

- Зачем нужна модель?
- Что она должна делать?
- С какой скоростью она должна работать?

Вложенные и невложенные модели

Выбор модели сильно зависит от того, какие модели мы сравниваем.

Определение

Если модель A является частным случаем модели B, то есть, A можно получить из B, наложив на параметры некоторые ограничения (или сделав замену переменных), то говорят, что модель A вложена в модель B.

Вложенные модели

- Вложенные модели могут сравниваться по критериям качества, которые мы обсуждали ранее.
- Мы можем использовать индексы, основанные на тестах согласия (например, χ^2 или $\chi^2/ndof$, что является эквивалентом тесту отношений правдоподобий).
- Предпочтительны индексы, которые штрафуют за сложность и чувствительны к малому n, например R_{adj}.

Статистика C_p Mallow

$$C_p(M) = \frac{SSE(M)}{\hat{\sigma}^2} - n + 2p(M).$$

- $\hat{\sigma}^2 = \frac{SSE(F)}{df_F}$ - лучшая оценка $\hat{\sigma}^2$, которая у нас есть (из самой сложной модели с максимальным набором регрессоров);
- $SSE(M) = ||Y - \hat{Y}_M||^2$ — сумма квадратов остатков в модели M.
- $p(M)$ - количество предикторов модели (или количество степеней свободы, которое забираем модель).
- Мотивирована разложением смещения-вариации.

AIC

AIC (Akaike information criterion):

$$AIC(S) = -2\ell(M) + 2p(M)$$

где $\ell(M) = \ell_M(\hat{\beta})$ — логарифм правдоподобия модели, где в качестве неизвестных параметров были подставлены их оценки, полученные с помощью максимизации $\ell_M(\beta)$.

В линейной регрессии в случае нормальных ошибок (шум берется равным оценке, полученной по полной модели) минимизация AIC эквивалента минимизации C_p.

BIC

BIC (Bayesian information criterion):

$$BIC(S) = -2\ell(M) + 2p(M)\log(n).$$

Этот функционал имеет байесовскую интерпретацию.

- Пусть $\mathcal{S} = \{S_1, \ldots, S_m\}$ — множество возможных моделей.
- Допустим, что априорное распределение имеет вид $\mathbb{P}(S_j) = 1/m$.
- Также предположим, что параметры внутри каждой модели имеют некоторое “гладкое” априорное распределение.
- Можно показать, что апостериорная вероятность модели примерно равна

$$\mathbb{P}(S_j|\text{выборка}) \approx \frac{\exp\Big(BIC(S_j)\Big)}{\sum_{r=1}^{m}\exp\Big(BIC(S_r)\Big)}.$$

BIC

Таким образом, выбор модели с наибольшим BIC эквивалентен выбору модели с наибольшей апостериорной вероятностью.

BIC также можно интерпретировать с точки зрения теории минимальной длины описания информации: BIC обычно “выбирает” модели с меньшим числом параметров.

Кросс-проверка

Оценка риска с помощью кросс-проверки (cross-validation; leave-one-out):

$$\hat{R}_{CV}(M) = \sum_{i=1}^{n} (\hat{Y}_{(i)} - Y_i)^2,$$

где $\hat{Y}_{(i)}$ — предсказание значения Y_i, полученное по модели, параметры которой оценены на обучающей выборке без i входа.

$$\hat{R}_{CV}(M) = \sum_{i=1}^{n} \frac{(\hat{Y}_i - Y_i)^2}{1 - U_{ii}(M)},$$

$$U(M) = X_S (X_S^T X_S)^{-1} X_S^T.$$

K-кратная кросс-проверка

1. Данные случайным образом делятся на k непересекающихся подвыборок (часто берут $k = 10$).
2. По одной подвыборке за раз удаляется (с возвращением), по остальным происходит оценка параметров.
3. Риск полагается равным $\sum_i (\hat{Y}_i - Y_i)^2$ (сумма берется по наблюдениям из удаленной подвыборки, данные оцениваются с помощью полученной модели).
4. Процесс повторяется для остальных подвыборок, после чего полученная оценка риска усредняется.

Для линейной регрессии оценка на основе коэффициента C_p Mallow и оценка на основе K-кратной кросс-проверки зачастую совпадают.

Резюме индексов

TABLE 13.2. Model Selection Indices for Non-Nested Models

Equation No.	Index	Rationale	Measure of model complexity	Pros	Cons
T15	$AIC = f + 2k$	Selects the model that had least expected discrepancy from the true model.	Number of parameters.	Easy to calculate. Performs well at large sample sizes.	Tends to select too complex models. Bad in recovering true model at small sample *N*.
T16	$CAIC = f + [1 + \ln(N)]k$	Selects the model that had least expected discrepancy from the true model.	Number of parameters and sample size.	Easy to calculate. Performs better than AIC at small sample size and large number of parameters.	
T17	$BIC = f + k\ln(N)$	Selects the model that is most likely to have generated the data in the Bayesian sense.	Number of parameters and sample size.	Easy to calculate. Performs well under large sample size.	Tends to select a model with too few parameters. Bad in recovering true model at small *N*.
T18	$CV = -\ln\int(y_{val} \mid \hat{\theta}_{cal})$	Selects the model that has more generalizability to the sample from the population.	Complexity penalty is implicit.	Easy to calculate. More consistent with the implication of generalizability.	Requires sample split. Estimates are often unreliable, especially for small sample size.
T19	$ECVI = f + \frac{2k}{N}$	Expected value of CV.	Number of parameters and sample size.	Can be calculated using one sample. More consistent with the implication of generalizability.	Assumes multivariate normality.
T20	$BMS = \ln\int_{\Theta} f(y \mid \theta)\pi(\theta)d\theta$	Selects the model with the highest mean likelihood of the data over the parameter space.	Number of parameters, sample size, and functional form of a model.	Includes more accurate measure on model complexity. Leads to more accurate model selection.	Computational burden. Hard to specify and calculate for most SEM models.

Note. AIC, Akaike information criterion; BIC, Bayesian information criterion; CV, cross-validation index; BMS, Bayesian model selection; *f*, minimized discrepancy function; *k*, the number of free parameters of the model; *N*, sample size; ln, natural logarithm; y_{val}, the validation sample; y_{cal}, the calibration sample; $\hat{\theta}_{cal}$, the parameter values estimated by the calibration sample; $\pi(\theta)$, the prior density of the parameters; Θ, the parameter space.

from: Handbook of Structural Equation Modeling (ed. by Rick Hoyle)

Перебор моделей

- Если в модели максимальное количество регрессоров равно k, то существует 2^k всевозможных моделей.
- В идеале необходимо “просмотреть” все модели, для каждой найти значение критерия качества и выбрать наилучшую согласно этому критерию.
- При большом количестве регрессоров для уменьшения трудоемкости используют регрессию методом включений, исключений или включений-исключений.

Метод включений / метод исключений

- Включения:
  - на первом шаге регрессоров нет вообще;
  - далее добавляется регрессор, для которого критерий качества максимальный и т.д.
- Исключения:
  - на первом шаге количество регрессоров максимальное;
  - на каждом шаге удаляется регрессор, исключение которого приводит к максимальному значению критерия качества.

Пример: метод исключений

Пример

Данные о преступлениях. Используем критерий AIC, что эквивалентно минимизации C_p Mallow.

В модели с полным набором регрессоров AIC = -310.37. В порядке убывания AIC при удалении каждой из переменных равен:

Численность населения (AIC = -308), Труд (AIC = -309), Южный штат (AIC = -309), Доход (AIC = -309), Количество мужчин (AIC = -310), Безработные I (AIC = -310), Образование (AIC = -312), Безработные II (AIC = -314), Возраст (AIC = -315), Расходы (AIC = -324).

Таким образом, имеет смысл удалить переменную “Население”.

Пример: метод исключений

Южный штат (AIC = -308), Труд (AIC = -308), Доход (AIC = -308), Количество мужчин (AIC = -309), Безработные I (AIC = -39), Образование (AIC = -310), Безработные II (AIC = -313), Возраст (AIC = -313), Расходы (AIC = -329).
Удаляем переменные до тех пор, пока не удастся больше получить увеличения AIC.

Уровень преступности = 1.2 Возраст + 0.75 Образование + 0.87 Расходы + 0.34 Количество мужчин − 0.86 Безработные I + 2.31 Безработные II.

Замечание: не дан ответ на то, какие переменные вызывают рост уровня преступности!