Регрессия

Стандартная линейная регрессия. НМК. Прогнозирование.

ПМИ ФКН ВШЭ, 03 ноября 2018 г.

Денис Деркач[1]

[1] ФКН ВШЭ

Оглавление

Введение

Регрессия

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

Один из способов оценить зависимость:

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

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

Линейная регрессия

Линейная регрессия:

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

Определение: простая линейная регрессия

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

Оценивание параметров:

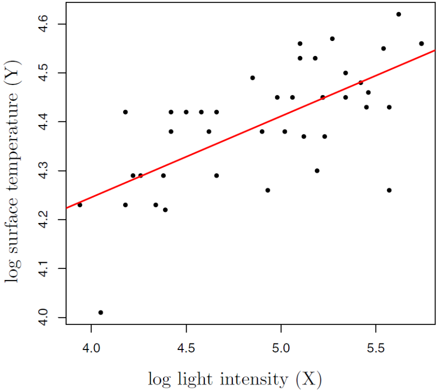
$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Предсказанные значения:

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

Примеры: линейная регрессия

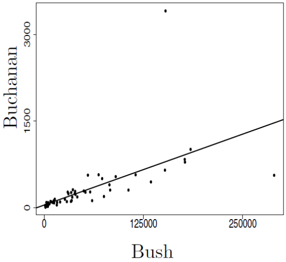
Данные о близлежащих звездах: оценка температуры звезды по её яркости.

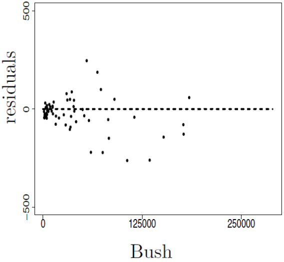


Оценки равны: $\hat{\beta}_0 = 3.58$ и $\hat{\beta}_1 = 0.166 \Rightarrow \hat{r}(x) = 3.58 + 0.166x.$

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.





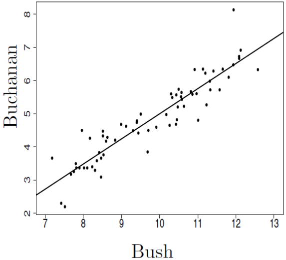
$$\hat{\beta}_0 = 66.0991,\ \hat{se}(\hat{\beta}_0) = 17.2926,$$

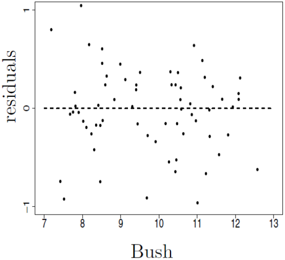
Примеры: стандартная линейная регрессия

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):

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

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

Метод наименьших квадратов

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2},$$

$$\hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{X}_n.$$

При этом несмещенная оценка дисперсии шума σ^2 равна

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

Свойства оценок МНК

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

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

$$\mathbb{V}ar(\hat{\beta}|X^n) = \frac{\sigma^2}{ns_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix}$$

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

Таким образом,

$$\hat{se}(\hat{\beta}_0) = \frac{\hat{\sigma}}{s_X\sqrt{n}} \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}, \qquad \hat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{s_X\sqrt{n}}.$$

Свойства оценок МНК

1. $\hat{\beta}_0 \xrightarrow{P} \beta_0,\ \hat{\beta}_1 \xrightarrow{P} \beta_1$.
2. $\frac{\hat{\beta}_0 - \beta_0}{\hat{se}(\hat{\beta}_0)} \rightsquigarrow \mathcal{N}(0,1),\ \frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\hat{\beta}_1)} \rightsquigarrow \mathcal{N}(0,1)$.
3. Приближенные доверительные интервалы размера $1-\alpha$ для параметров:
$$\hat{\beta}_0 \pm z_{\alpha/2}\hat{se}(\hat{\beta}_0) \text{ и } \hat{\beta}_1 \pm z_{\alpha/2}\hat{se}(\hat{\beta}_1).$$
4. Тест Вальда для проверки $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ имеет вид: H_0 отклоняется, если $|W| > z_{\alpha/2}$, где $W = \hat{\beta}_1 / \hat{se}(\hat{\beta}_1)$.

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

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

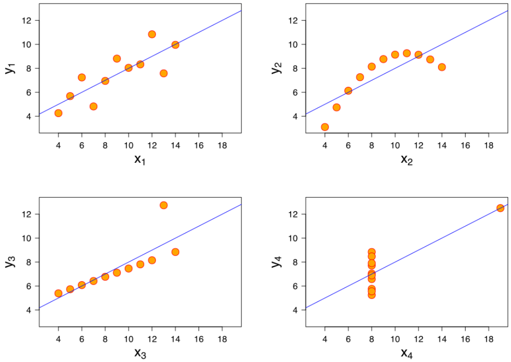
Статистика Вальда для проверки $H_0 : \beta_1 = 0$ против альтернативы $H_1 : \beta_1 \neq 0$ равна $|W| = |.7303 - 0|/.0358 = 20.40$. Причем $p - value$ равно $\mathrm{P}(|Z| > 20.40) \approx 0 \Rightarrow$ зависимость действительно существует.

Способы оценки качества регрессии

- Mean square (L2) loss (MSE): $\mathrm{MSE}(h, X^\ell) = \frac{1}{\ell}\sum\limits_{i=1}^{\ell}(y_i - h(x_i))^2$
- Root MSE: $\mathrm{RMSE}(h, X^\ell) = \sqrt{\frac{1}{\ell}\sum\limits_{i=1}^{\ell}(y_i - h(x_i))^2}$
- 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: $\mathrm{MAE}(h, X^\ell) = \frac{1}{\ell}\sum\limits_{i=1}^{\ell}|y_i - h(x_i)|$

Квартет Энскомба



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

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

Множественная регрессия

Введение

Стандартная линейная регрессия

Множественная регрессия

Прогнозирование

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 $(i = 1, \ldots, n)$

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i,$$
$$\mathrm{E}\left(\varepsilon_i | X_{1i}, \ldots, X_{ki}\right) = 0.$$

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

Множественная регрессия

Модель может быть выписана:

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

или в матричной форме:

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

Множественная регрессия

Предположим, что матрица $X^T X$ размера $k \times k$ невырожденная, тогда

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

Оценка функции регрессии имеет вид

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

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

$\hat{\varepsilon} = X\hat{\beta} - Y$ — вектор остатков.

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 $\hat{\sigma}^2(X^T X)^{-1}$.

Пример: множественная регрессия

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:

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

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

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

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

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

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

Прогнозирование

Введение

Стандартная линейная регрессия

Множественная регрессия

Прогнозирование

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 $X = x_*$:

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

$$\mathbb{V}ar(\hat{Y}_*) = \mathbb{V}ar(\hat{\beta}_0 + \hat{\beta}_1 x_*) = \mathbb{V}ar(\hat{\beta}_0) + x_*^2 \mathbb{V}ar(\hat{\beta}_1) + 2x_* Cov(\hat{\beta}_0, \hat{\beta}_1)$$

$\Rightarrow$ можно подсчитать $\hat{se}(\hat{Y}_*)$, используя в качестве оценки σ^2 величину $\hat{\sigma}^2$.

Прогнозирование

Пусть

$$\hat{\xi}_n^2 = \hat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_{i=1}^n (X_i - \overline{X})^2} + 1 \right).$$

Приблизительный prediction interval для Y_* размера $1 - \alpha$ имеет вид

$$\hat{Y}_* \pm z_{\alpha/2} \hat{\xi}_n.$$

Пример: прогнозирование

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 $(6.2, 6.578)$, или, в исходных единицах — $(493, 717)$, что мало в сравнении с 3476.

Выбор модели

Введение

Стандартная линейная регрессия

Множественная регрессия

Прогнозирование

Выбор модели

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.

Обозначения

Пусть $S \subset \{1, \ldots, k\}$ — подмножество регрессоров. Тогда

- β_S — коэффициенты при соответствующих регрессорах, $\hat{\beta}_S$ — их оценки;
- X_S — подматрица матрицы типа X в соответствии с данным подмножеством регрессоров;
- $\hat{r}_S(x)$ — оцененная функция регрессии, $\hat{Y}_i(S) = \hat{r}_S(X_i)$ — предсказанные значения.

Риск прогноза

Риск прогноза:

$$R(S) = \sum_{i=1}^{n} \mathrm{E}\,(\hat{Y}_i(S) - Y_i^*)^2,$$

где $Y_i^* = X_i\beta$ — истинное значения выхода для X_i.

Задача состоит в выборе подмножества S, которое минимизирует $R(S)$.

Оценка риска прогноза

Оценка риска прогноза (ошибка на обучающей выборке):

$$\hat{R}_{tr}(S) = \sum_{i=1}^{n} (\hat{Y}_i(S) - Y_i)^2.$$

Оценка риска прогноза смещена по сравнению с реальным значением риска прогноза:

$$\left(\hat{R}_{tr}(S)\right) = \mathbb{E}\hat{R}_{tr}(S) - R(S) = \sum_{i=1}^{n} \left(\sigma_i^2 - 2\,\mathrm{cov}\,(\hat{Y}_i(S), Y_i)\right).$$

Оценка риска прогноза

- Причина в том, что данные использовались дважды — для оценки параметров и для оценки риска прогноза.
- Если параметров много, то $\mathrm{cov}\,(\hat{Y}_i(S), Y_i)$ принимает большое значение.
- При этом прогноз на данных, отличных от данных в обучающей выборке, может оказаться существенно хуже!

Статистика C_p Mallow

Статистика C_p Mallow:

$$\hat{R}(S) = \hat{R}_{tr}(S) + 2|S|\hat{\sigma}^2,$$

где $|S|$ — число регрессоров, $\hat{\sigma}^2$ — оценка дисперсии шума σ^2, полученная по полной модели (т.е. с включением всех регрессоров).

Критерий включает оценку риска прогноза на обучающей выборке и “сложность” модели (регуляризация).

AIC

AIC (Akaike information criterion):

$$AIC(S) = |S| - \ell_S,$$

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

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

Кросс-проверка

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

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

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

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

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

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. В более сложных случаях кросс-проверка работает лучше.

BIC

BIC (Bayesian information criterion):

$$BIC(S) = \frac{|S|}{2}\log n \ - \ell_S.$$

Этот функционал имеет байесовскую интерпретацию.

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

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

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.

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

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 = 309), 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.