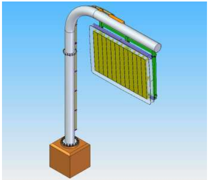
Прикладная статистика 8. Регрессионный анализ, часть вторая.

8 апреля 2013 г.

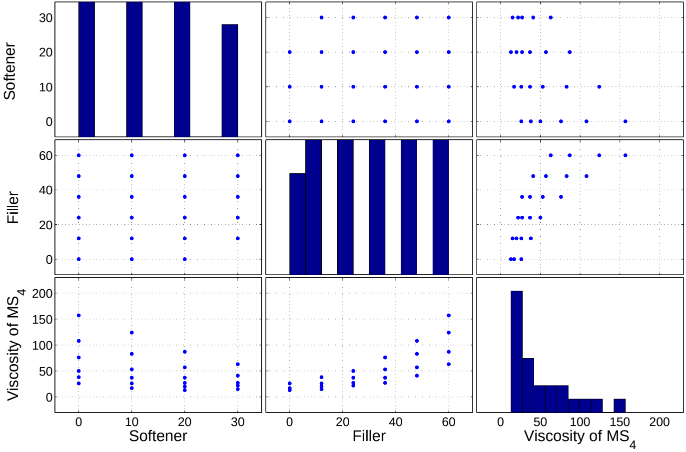
Вязкость MS_4

Derrigher GC, An empirical model for viscosity of filled and plasticized elastomer products (1974): исследовалась вязкость MS_4 при $100\,^\circ C$ при разных уровнях наполнителя и пластификатора.

Найти преобразование отклика, обеспечивающее хороший подбор модели первого порядка.



Вязкость MS_4



$$\max y / \min y = 12.0769.$$

Преобразования Бокса-Кокса

Пусть имеются положительные значения отклика $y_1, y_2, \ldots, y_n$.
Если отношение наибольшего наблюдаемого y к наименьшему превосходит 10, стоит рассмотреть возможность преобразования y.
В каком виде искать преобразование?

Часто полезно рассмотреть преобразования вида y^λ, но оно не имеет смысла при $\lambda = 0$.
Вместо него можно рассмотреть преобразование вида

$$W = \begin{cases} \left(y^\lambda - 1\right)/\lambda, & \lambda \neq 0; \\ \ln y, & \lambda = 0, \end{cases}$$

но оно сильно варьируется по λ.
Вместо него можно рассмотреть преобразование вида

$$V = \begin{cases} \left(y^\lambda - 1\right)/\left(\lambda \dot{y}^{\lambda-1}\right), & \lambda \neq 0; \\ \dot{y} \ln y, & \lambda = 0, \end{cases}$$

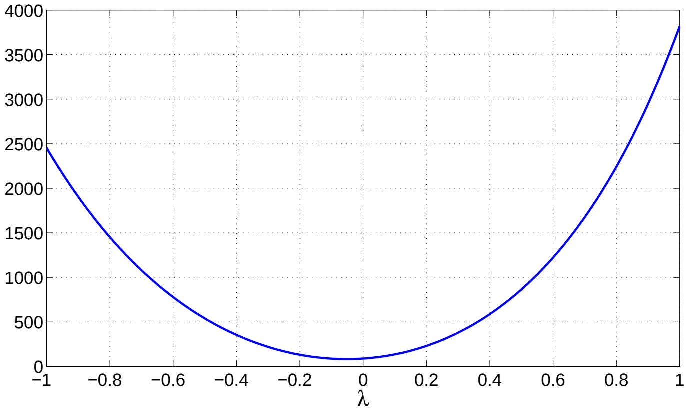
где $\dot{y} = (y_1 y_2 \ldots y_n)^{1/n}$ — среднее геометрическое наблюдений.

Метод Бокса-Кокса

Процесс подбора λ:

1. выбирается набор значений λ в некотором интервале, например, $(-2; 2)$;
2. для каждого значения λ выполняется преобразование отклика V, строится регрессия, вычисляется остаточная сумма квадратов $RSS\,(\lambda, V)$;
3. строится график зависимости $RSS\,(\lambda, V)$ от λ, по нему определяется оптимальное значение λ;
4. выбирается ближайшее к оптимальному удобное значение λ (например, полуцелое);
5. строится окончательная модель регрессии с откликом y^{λ} или $\ln y$.

Вязкость MS_4



Выбираем $\lambda = 0$, т. е., $y = \ln y$.

Вязкость MS_4

Доверительный интервал для λ выбирается из уравнения

$$L\left(\hat{\lambda}\right) - L\left(\lambda\right) \leq \frac{1}{2}\chi_1^2\left(1-\alpha\right),$$

где $L\left(\lambda\right) = -\frac{1}{2}n\ln\left(\frac{RSS(\lambda)}{n}\right)$.

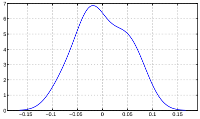
Если он содержит единицу, возможно, не стоит выполнять преобразование. Если он содержит несколько удобных значений λ, то всё равно, какое из них выбирать.

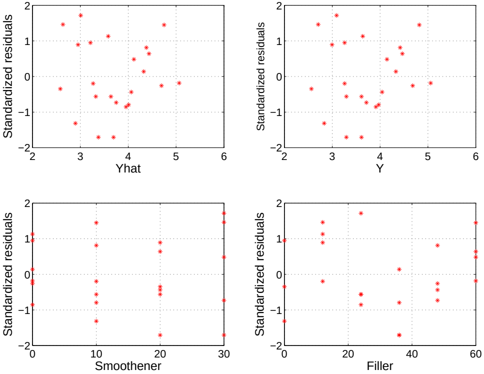
Для нашей задачи 95% доверительный интервал — $-0.13 \leq \lambda \leq 0.03$.

Итоговое уравнение:
$\ln y = 3.212 + 0.03088f - 0.03152p, \quad F = 2045, \quad p \approx 0, \quad R^2 = 0.9951$ (модель объясняет $100R^2 = 99.51\%$ отклонения от среднего значения).
Без преобразования:
$y = 28.184 + 1.55f - 1.717p, \quad F = 72.9, \quad p \approx 0, \quad R^2 = 0.8793$ (модель объясняет $100R^2 = 87.93\%$ отклонения от среднего значения).

Вязкость MS_4

Особенно важно исследовать остатки.





Химический состав цемента

Woods H, Steinour HH, Starke HR, Effect of composition of Portland cement on heat involved during hardening (1932): измерено тепло, выделенное цементом при отвердевании (калорий на грамм цемента), а также количество в составе цемента трикальциум аллюмината, трикальциум силиката, тетракальциум аллюминиоферрита и дикальциум силиката. Матрица корреляций Пирсона признаков:

r	X_1	X_2	X_3	X_4
X_1	1.0000	0.2286	**-0.8241**	-0.2454
X_2	0.2286	1.0000	-0.1392	**-0.9730**
X_3	**-0.8241**	-0.1392	1.0000	0.0295
X_4	-0.2454	**-0.9730**	0.0295	1.0000

Была подобрана линейная модель вида

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4.$$

Необходимо проверить гипотезы $H_0\colon \theta_1 = -\theta_3, \theta_2 = -\theta_4$ и $H_0\colon \theta_3 = \theta_2 - \theta_1$.

Общая линейная гипотеза

Общая линейная гипотеза — гипотеза, содержащая одно или несколько утверждений о линейных комбинациях коэффициентов регрессии.

Примеры:

- модель $E\left(y|\,X\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$,
  гипотеза:
  $$H_0\colon \theta_1 = 0, \theta_2 = 0$$
  две линейно независимые функции;
- модель $E\left(y|\,X\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k$,
  гипотеза:
  $$H_0\colon \theta_1 = \theta_2 = \ldots = \theta_k = \theta \Leftrightarrow H_0\colon \theta_1 - \theta_2 = 0, \theta_2 - \theta_3 = 0, \ldots, \theta_{k-1} - \theta_k = 0$$
  $k-1$ линейно независимых функций;
- общий случай: модель $E\left(y|\,X\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_k x_k$,
  гипотеза:
  $$H_0\colon \mathbf{C}\theta = 0$$
  $C \in \mathbb{R}^{m \times k}$, q линейно независимых строк, $m-q$ являются линейными комбинациями.

Проверка общей линейной гипотезы

RSS_{full} — остаточная сумма квадратов исходной модели, $n-k$ степеней свободы;
RSS_{short} — остаточная сумма квадратов модели при справедливости общей линейной гипотезы, $n-k+q$ степеней свободы.

$$\left(\frac{RSS_{short}-RSS_{full}}{q}\right) \Big/ \left(\frac{RSS_{full}}{n-k}\right) \sim F\left(q;n-k\right).$$

Химический состав цемента

Полная модель:
$y = 62.4 + 1.55x_1 + 0.51x_2 + 0.102x_3 - 0.144x_4, \quad RSS_{full} = 47.8636.$

Гипотеза: $H_0\colon \theta_1 - \theta_3 = 0, \theta_2 - \theta_4 = 0 \Rightarrow$ сокращённая модель
$y = \theta_0 + \theta_1 (x_1 - x_3) + \theta_2 (x_2 - x_4), \quad RSS_{short} = 109.0523.$
$F = \left(\frac{109.0523 - 47.8636}{2}\right) \Big/ \left(\frac{47.8636}{8}\right) = 5.1136, \quad p = 0.0371.$
Гипотеза H_0 отвергается.

Гипотеза: $H_0\colon \theta_3 = \theta_2 - \theta_1 \Rightarrow$ сокращённая модель
$y = \theta_0 + \theta_1 (x_1 + x_3) + \theta_2 (x_2 - x_3) + \theta_4 x_4, \quad RSS_{short} = 57.1888.$
$F = \left(\frac{57.1888 - 47.8636}{1}\right) \Big/ \left(\frac{47.8636}{8}\right) = 1.5586, \quad p = 0.2472.$
Гипотеза H_0 не отвергается.

Содержание свободного хлора

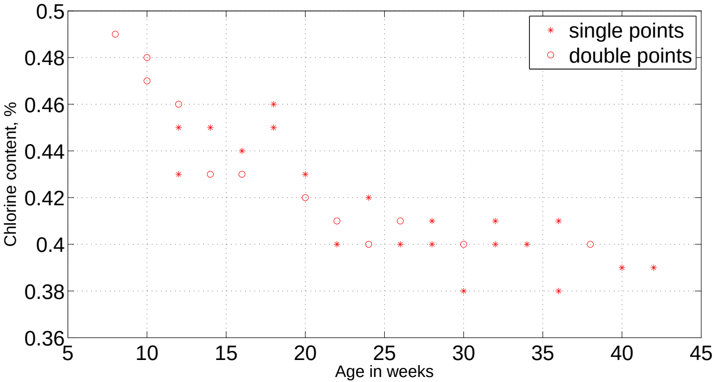
Smith H, Dubey SD, Some reliability problems in the chemical industry (1964): исследование корпорации Procter & Gamble. Исследуется продукт A, в момент производства доля свободного хлора в нём должна составлять 0.5. Известно, что со временем содержание хлора в продукте снижается. За первые 8 недель содержание хлора снизится до 0.49, но в более поздние сроки из-за влияния большого количества неконтролируемых факторов теоретические расчёты не могут достаточно надёжно предсказать содержание свободного хлора. Для определения закона убывания концентрации свободного хлора она была измерена в 44 образцах на разных сроках хранения.

Была выдвинута гипотеза, что модель вида

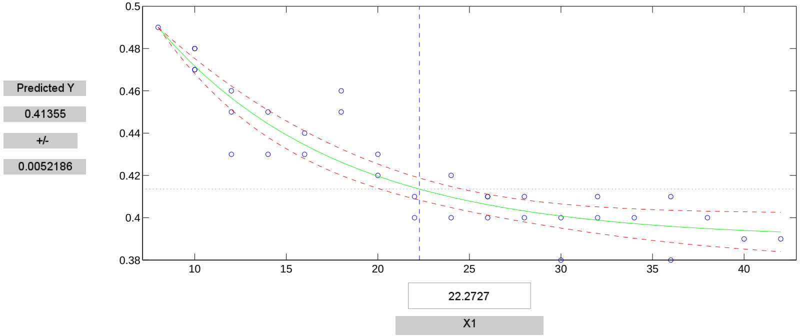
$$y = \alpha + (0.49 - \alpha)\, e^{-\beta(x-8)} + \varepsilon$$

описывает содержание хлора в продукте при $x \geq 8$.
Требуется оценить параметры α и β по данным.

Содержание свободного хлора



Содержание свободного хлора



$$\hat{\alpha} = 0.3901, \quad \hat{\beta} = 0.1016, \quad RSS = 0.00500168.$$

Что дальше?

Сравнение RSS с чистой ошибкой

Чистая ошибка σ^2 — дисперсия ϵ, может быть оценена по повторяющимся наблюдениям.

$y_{11}, y_{12}, \ldots, y_{1n_1}$ — n_1 повторных наблюдений при x_1;
$y_{21}, y_{22}, \ldots, y_{2n_2}$ — n_2 повторных наблюдений при x_2;
...
$y_{m1}, y_{m2}, \ldots, y_{mn_m}$ — n_m повторных наблюдений при x_m.

$$\hat{\sigma}^2 = \frac{S_{pe}}{n_e} = \frac{\sum\limits_{j=1}^{m} \sum\limits_{u=1}^{n} (y_{ju} - \bar{y}_j)^2}{\sum\limits_{j=1}^{m} n_j - m}.$$

В нашем случае $S_{pe} = 0.0024, \quad n_e = 26;$

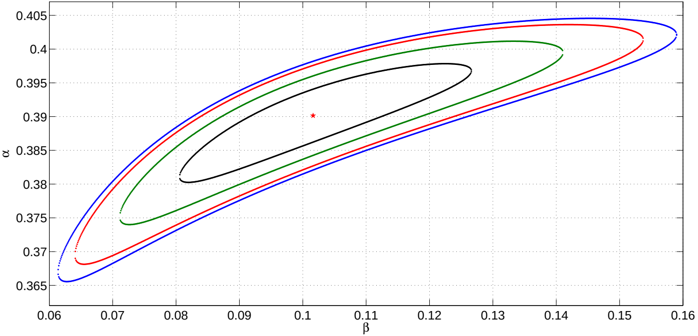
$$\frac{RSS - S_{pe}}{44 - 2 - n_e} = 0.00016,$$

$$\frac{S_{pe}}{26} = 0.00009.$$

Формально F-критерий неприменим, но можно на него ориентироваться: $F(16; 26; 0.95) = 2.08, \quad \frac{0.00016}{0.00009} = 1.8$ — можно надеяться, что модель подобрана хорошо.

Доверительные области

Приблизительные $100\,(1-q)$-процентные доверительные области для значений параметров α и β:



Синий контур — $q = 0.005$, красный — $q = 0.01$, зелёный — $q = 0.05$, чёрный — $q = 0.25$.

Прикладная статистика
8. Регрессионный анализ, часть вторая.

Рябенко Евгений
riabenko.e@gmail.com