

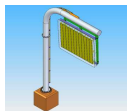
Прикладная статистика. Занятие 8. Регрессионный анализ, часть вторая.

10 апреля 2012 г.

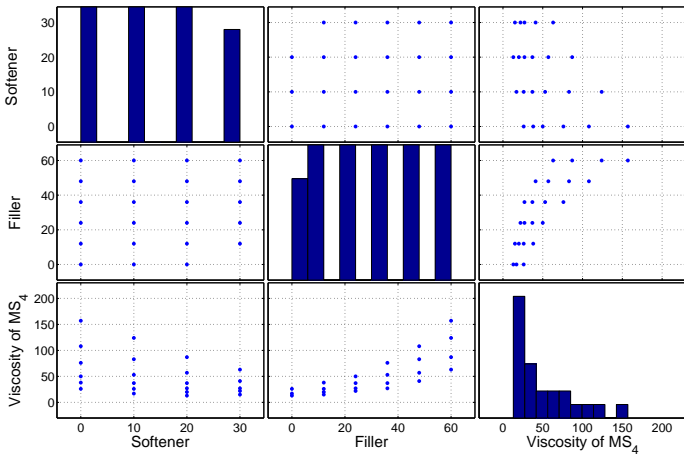
Вязкость MS_4

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



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

Вязкость MS_4

Пусть имеются положительные значения отклика Y_1, Y_2, \dots, 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

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

но оно сильно варьируется по λ .

Вместо него можно рассмотреть преобразование вида

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

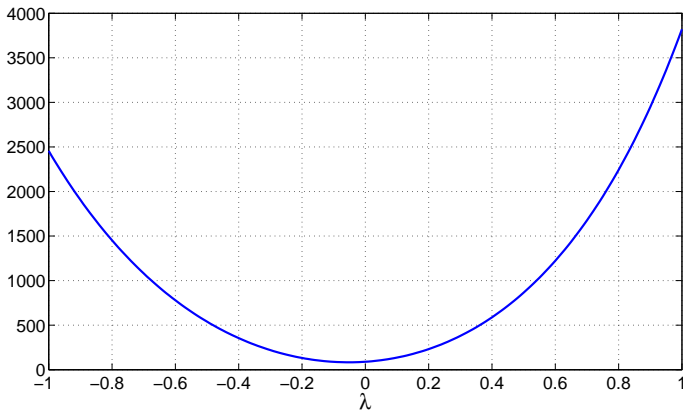
где $\dot{Y} = (Y_1 Y_2 \dots Y_n)^{1/n}$ — среднее геометрическое наблюдений.

Вязкость MS_4

Процесс подбора λ :

- 1 выбирается набор значений λ в некотором интервале, например, $(-2; 2)$;
- 2 для каждого значения λ выполняется преобразование отклика, строится регрессия, вычисляется остаточная сумма квадратов SSE ;
- 3 строится график $RSS(\lambda)$, 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

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

где $L(\lambda) = n \ln(RSS(\lambda)/n)$.

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:

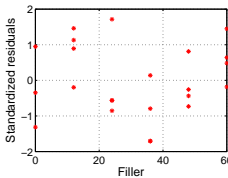
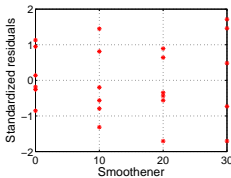
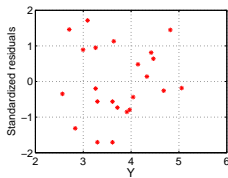
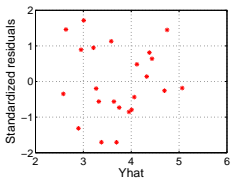
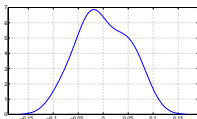
$\ln Y = 3.212 + 0.03088f - 0.03152p$, $F = 2045$, $p \approx 0$, $R^2 = 0.9951$
(модель объясняет $100R^2 = 99.51\%$ отклонения от среднего значения).

Без преобразования:

$Y = 28.184 + 1.55f - 1.717p$, $F = 72.9$, $p \approx 0$, $R^2 = 0.8793$ (модель объясняет $100R^2 = 87.93\%$ отклонения от среднего значения).

Вязкость MS_4

Особенно важно исследовать остатки.



Содержание свободного хлора

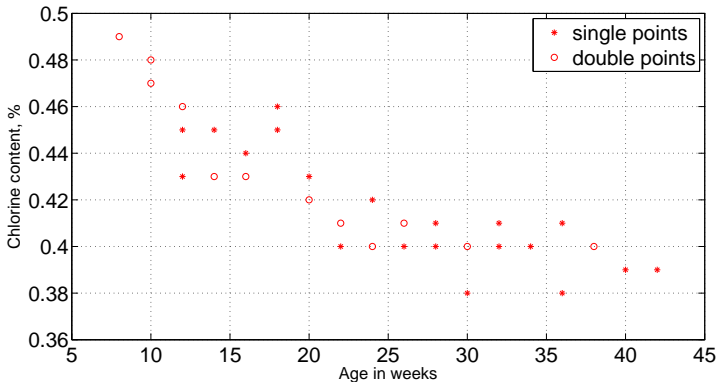
Smith H, Dubey SD, Some reliability problems in the chemical industry (1964): исследование корпорации Procter & Gamble. Исследуется продукт А, в момент производства доля свободного хлора в нём должна составлять 0.5. Известно, что со временем содержание хлора в продукте снижается. За первые 8 недель содержание 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

$$Y = \alpha + (0.49 - \alpha) e^{-\beta(X-8)} + \epsilon$$

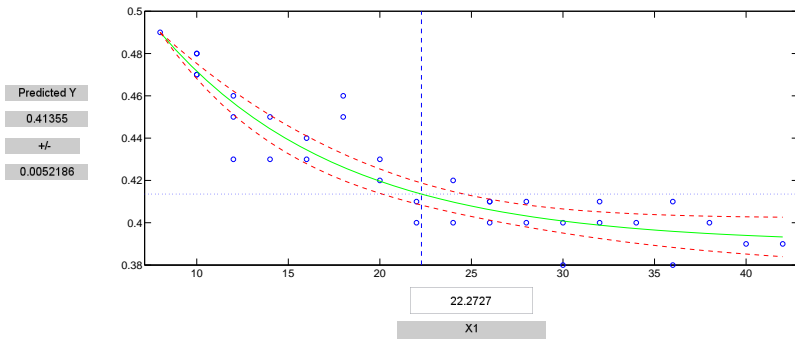
описывает содержание хлора в продукте при $X \geq 8$. Требуется оценить параметры α и β по данным.



Содержание свободного хлора



Содержание свободного хлора



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

Что дальше?

Содержание свободного хлора

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

$Y_{11}, Y_{12}, \dots, Y_{1n_1}$ — n_1 повторных наблюдений при X_1 ;

$Y_{21}, Y_{22}, \dots, Y_{2n_2}$ — n_2 повторных наблюдений при X_2 ;

...

$Y_{m1}, Y_{m2}, \dots, Y_{mn_m}$ — n_m повторных наблюдений при X_m .

$$\hat{\sigma}^2 = \frac{S_{pe}}{n_e} = \frac{\sum_{j=1}^m \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2}{\sum_{j=1}^m n_j - m}.$$

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

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

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

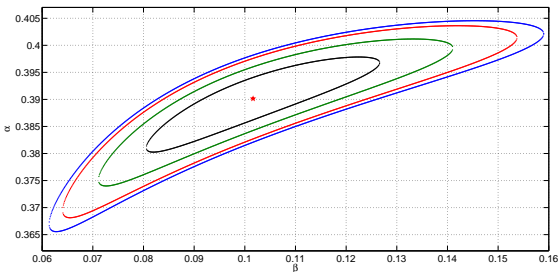
Формально F -критерий неприменим, но можно на него ориентироваться:
 $F(16; 26; 0.95) = 2.08$, $\frac{0.00016}{0.00009} = 1.8$ — можно надеяться, что модель подобрана хорошо.

Содержание свободного хлора

Можно построить приближительные $100(1 - q)$ -процентные доверительные области для значений параметров α и β , подбирая их из уравнения

$$\sum_{u=1}^n (Y_u - \hat{Y}_u)^2 = RSS \times \left(1 + \frac{p}{n-p} F(p; n-p; 1-q) \right),$$

где p — число оцениваемых параметров, n — число точек в массивах X и Y , F — распределение Фишера.



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

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.

Была подобрана полная линейная модель вида

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4.$$

Необходимо проверить гипотезу $H_0: \beta_1 = \beta_3, \beta_2 = \beta_4$.

Химический состав цемента

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

Примеры:

- модель $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, гипотеза $H_0: \beta_1 = 0, \beta_2 = 0$ — две линейно независимые функции;
- модель $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$, гипотеза $H_0: \beta_1 = \beta_2 = \dots = \beta_k = \beta \Leftrightarrow H_0: \beta_1 - \beta_2 = 0, \beta_2 - \beta_3 = 0, \dots, \beta_{k-1} - \beta_k = 0$ — $k - 1$ линейно независимых функций;
- общий случай: модель $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$, гипотеза $H_0: \mathbf{C}\beta = 0$ — q линейно независимых функций, $m - q$ являются линейными комбинациями.

Химический состав цемента

Проверка общей линейной гипотезы:

RSS_{full} — остаточная сумма квадратов исходной модели, $n - p$ степеней свободы;

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

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

Химический состав цемента

Матрица корреляций исходных признаков (Пирсона):

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

Полная модель:

$$Y = 62.4 + 1.55X_1 + 0.51X_2 + 0.102X_3 - 0.144X_4, \quad RSS = 47.8636.$$

Сокращённая модель:

$$Y = 241.1 - 1.106(X_1 + X_3) - 2.104(X_2 + X_4), \quad RSS = 2510.1032.$$

$$F = \left(\frac{2510.1032 - 47.8636}{2} \right) / \left(\frac{47.8636}{8} \right) = 205.7714, \quad p \approx 0.$$

Гипотеза H_0 отвергается.

Прикладная статистика
Семинар 8. Регрессионный анализ, часть вторая.

Рябенко Евгений
riabenko.e@gmail.com