Семинар 9

Робастная регрессия. Логит регрессия.

Грауэр Л.В., Архипова О.А.

Санкт-Петербург, 2015

Задание 1. Построение робастной регрессии.

Постройте несколько переменных x_i (N = 30), по ним постройте y с нормально распределенными остатками с м.о. = 0. Добавьте выбросы. Сравните:

- обыкновенную линейную регрессию
- робастную регрессию с M-оценками с функцией Хубера
- регрессию с оценками LMS
- регрессию с оценками LTS

Сравните результаты, для LMS и LTS обратите внимание на разброс коэффициентов регрессии (запустите 1000 раз на тех же данных).

Задание 1. Функции в R

Функции в R:

- library(MASS)
- glm($y \sim x1 + x2 + xleft$) - линейная регрессия
- summary(fit) - вывод результатов
- plot(fit) - диагностические графики
- layout(matrix(c(1,2,3,4),2,2)) и plot(fit) - все 4 графика сразу
- rlm($y \sim x1 + x2 + xleft$, psi = psi.huber) - робастная регрессия с M-оценками. Хубер - по умолчанию
- lqs($y \sim x1 + x2 + xleft$, method = c('lts')) - регрессия с оценками LTS
- или lmsreg($y \sim x1 + x2 + xleft$), ltsreg($y \sim x1 + x2 + xleft$)
- lmsreg_result$coefficients - коэффициенты (получить коэф. для разброса)

Задание 1. Пример. Обыкновенная регрессия.

x1<-rnorm(30,12,4)
x2<-rnorm(30,-3,2)
xleft<–rnorm(30,-3,2)+0.1*rnorm(30)
y<-0.5*x1 + 6*x2+ rnorm(30)
y[1] = -200
y[5] = 150
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.1207	31.5293	0.289	0.7747
x1	-0.2859	2.0046	-0.143	0.8877
x2	8.8265	4.0900	2.158	0.0403 *
xleft	3.1454	4.3691	0.720	0.4780

Задание 1. Пример. Регрессия с M- оценками. Функция Хубера.

Coefficients:

	Value	Std. Error	t value
(Intercept)	-0.1208	0.9453	-0.1278
x1	0.4640	0.0601	7.7205
x2	5.9149	0.1226	48.2370
xleft	0.0891	0.1310	0.6799

Задание 1. Пример. LMS и LTS.

LMS:
Coefficients:
(Intercept) x1 x2 xleft
-1.553558932 0.571310950 5.787719849 0.002555278

LTS:
(Intercept) x1 x2 xleft
-1.56953523 0.55780867 5.76593710 0.02581588

Разброс коэффициентов (5 и 95 percentile):
LMS:

	5%	95%
Intercept	-2.856939901	-1.43115898
x1	0.557884740	0.75303705
x2	5.749951688	6.27122948
xleft	-0.006237836	0.05908828

Задание 2. Логистическая регрессия.

Загрузить данные binary.csv. Переменная rank - должна быть фактором (так как она категориальная).
Построить логистическую регрессию, где зависимая переменная - admit. Посмотреть на результат. Что значат полученные коэффициенты? Получите доверительные интервалы для коэффициентов. Примените критерий Вальда для проверки общей значимости переменной rank, для проверки значимости одной переменной. Перейдите от коэффициентов к odds-ratio (оценки и доверительные интервалы).

Задание 2. Функции в R.

Функции в R:

- library(aod) - для Wald
- mydata = read.csv(file)
- summary(mydata) - описательная статистика для всех переменных датасета
- mydata$rank <- factor(mydata$rank)
- mylogit = glm($admit \sim gre + gpa + rank$, data = mydata, family = "binomial") - логистическая регрессия
- confint(mylogit) - доверительный интервал с помощью функции правдоподобия
- confint.default(mylogit) - дов интервал с помощью standard errors
- wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 4:6)
- wald.test(b = coef(mylogit), Sigma = vcov(mylogit), L = l), l - cbind(0, 0, 0, 1, -1, 0) - проверка значимости коэффициента у 4 переменной по сравнению с пятой

Задание 2. Пример.

```
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = "binomial",
##     data = mydata)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.627  -0.866  -0.639   1.149   2.079
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.98998    1.13995   -3.50  0.00047 ***
## gre          0.00226    0.00109    2.07  0.03847 *
## gpa          0.80404    0.33182    2.42  0.01539 *
## rank2       -0.67544    0.31649   -2.13  0.03283 *
## rank3       -1.34020    0.34531   -3.88  0.00010 ***
## rank4       -1.55146    0.41783   -3.71  0.00020 ***
## ---
## Signif. code    0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion   rameter for binomial family taken to be 1)
##
##     Null dev  nce: 499.98  on 399  degrees of freedom
## Residual dev  nce: 458.52  on 394  degrees of freedom
## AIC: 470.5
##
## Number of Fi  er Scoring iterations: 4
```

Задание 2. Пример. Доверительные интервалы для коэффициентов.

```
##                 2.5 %    97.5 %
## (Intercept) -6.271620 -1.79255
## gre          0.000138  0.00444
## gpa          0.160296  1.46414
## rank2       -1.300889 -0.05675
## rank3       -2.027671 -0.67037
## rank4       -2.400027 -0.75354
```

Wald test:
Chi-squared test:
X2 = 20.9, df = 3, P(> X2) = 0.00011

Задание 2. Пример. Переход от коэффициентов к odds ratio

```
##                    OR   2.5 % 97.5 %
## (Intercept) 0.0185 0.00189  0.167
## gre         1.0023 1.00014  1.004
## gpa         2.2345 1.17386  4.324
## rank2       0.5089 0.27229  0.945
## rank3       0.2618 0.13164  0.512
## rank4       0.2119 0.09072  0.471
```

Задание 3. В дополнение к заданию 2.

Постройте предсказанные вероятности события (admit=1) для каждого уровня категориальной переменной rank, зафиксировав gre и gra как средние значения этих переменных.
Для этого:

1. Создаем наблюдения, для которых хотим получить результат. Каждое наблюдение - это c(rank, mean(gre), mean(gra))
2. Используя predict, получаем вероятности.

Функции в R:

- newdata1 = with(mydata, data.frame(gre = mean(gre), gpa = mean(gpa), rank = factor(1:4)))
- newdata1$rankP = predict(mylogit, newdata = newdata1, type = "response")

Задание 3. Пример.

gre	gpa	rank	rankP
587.7	3.3899	1	0.5166016
587.7	3.3899	2	0.3522846
587.7	3.3899	3	0.2186120
587.7	3.3899	4	0.1846684