Семинар 8

Линейная регрессия. МНК, анализ остатков

Грауэр Л.В., Архипова О.А.

Санкт-Петербург, 2015

Задание 1. Построение линейной регрессии. Анализ остатков

Постройте несколько переменных x_i (N = 30), по ним постройте y:

- с нормально распределенными остатками с м.о. = 0
- с нормально распределенными остатками с м.о. = 0, но с outlier
- с любыми другими остатками

Для этих трех случаев постройте линейную регрессию, сравните результаты, постройте диагностические графики для остатков (проверка на наличие выбросов, зависимость от fitted, проверка на нормальность).

Функции в R:

- glm($y \sim x1 + x2 + x1 * x3$) - 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
- residuals(glmobject) - остатки модели. Или residuals(glmobject,type="deviance")
- rstudent(glmmodel) - стьюдентизированные остатки

К заданию 1

$\hat{y} = Hy$, y - observed, $\hat{y}$ - fitted values.
leverage (h_{ii}) - диагональные элементы.
Cook's distanse (D):
$D = \frac{e_i^2 * h_{ii}}{p * MSE * (1 - h_{ii})^2}$,
e_i - residuals, p - число параметров модели.

Задание 1. Пример. Нормальные Остатки.

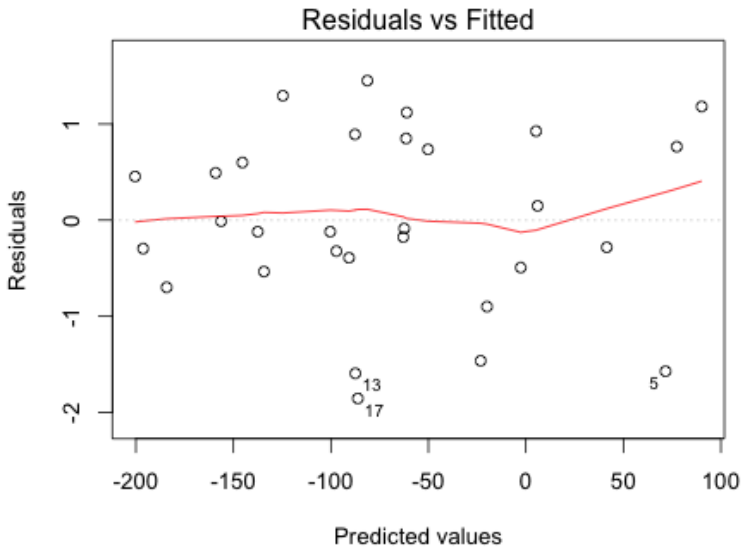
x1=rnorm(30,12,4)
x2=rnorm(30,-3,2)
x3=rnorm(30,-1,2)
y=0.5*x1+2*x2+3*x1*x3 + rnorm(30)
Coefficients:

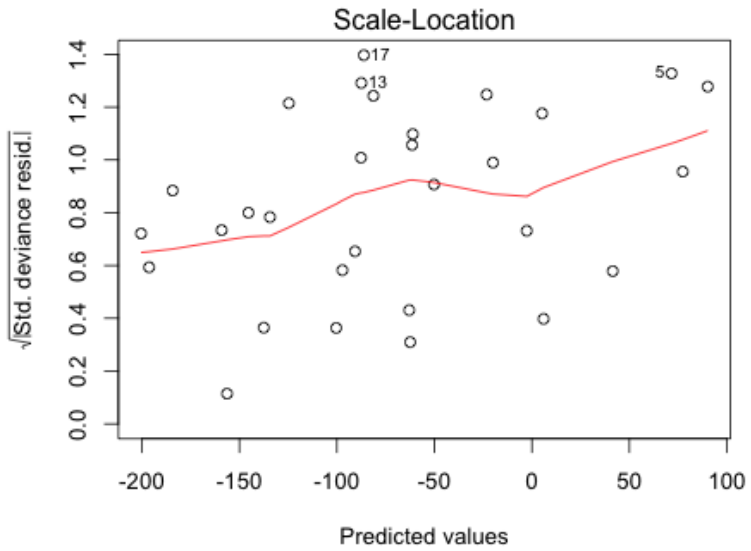
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.07529	0.62571	-1.719	0.0981 .
x1	0.60032	0.04888	12.282	4.36e-12 ***
x2	1.97159	0.07357	26.800	< 2e-16 ***
x3	-0.11306	0.23722	-0.477	0.6378
x1:x3	3.02457	0.02007	150.686	< 2e-16 ***

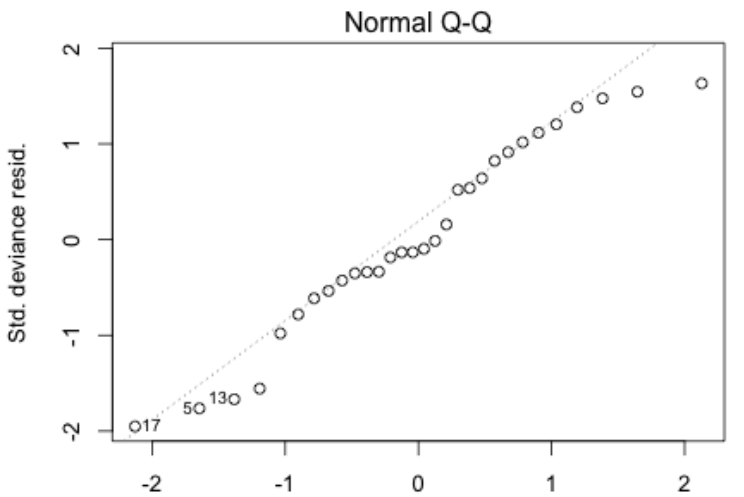
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.9468812)
Null deviance: 183017.092 on 29 degrees of freedom
Residual deviance: 23.672 on 25 degrees of freedom
AIC: 90.029

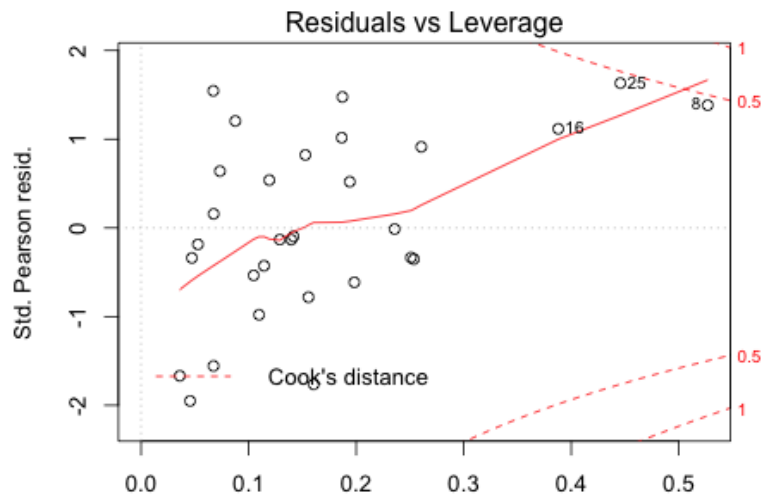
Задание 1. Пример. Нормальные Остатки.

Модель с нормальными остатками



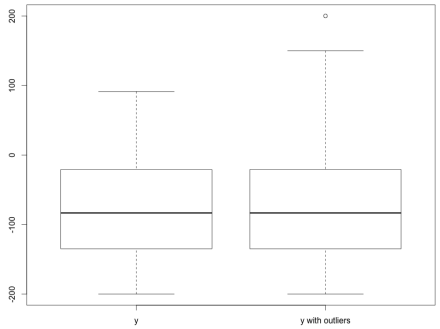






Задание 1. Пример. Выбросы.

Добавление выбросов. (у меня выбросы 1ое и 5ое наблюдения)



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.4255	15.0462	1.690	0.1035
x1	0.2070	1.1754	0.176	0.8617
x2	3.9597	1.7690	2.238	0.0343 *
x3	7.8461	5.7044	1.375	0.1812
x1:x3	2.8669	0.4827	5.940	3.36e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 547.5324)

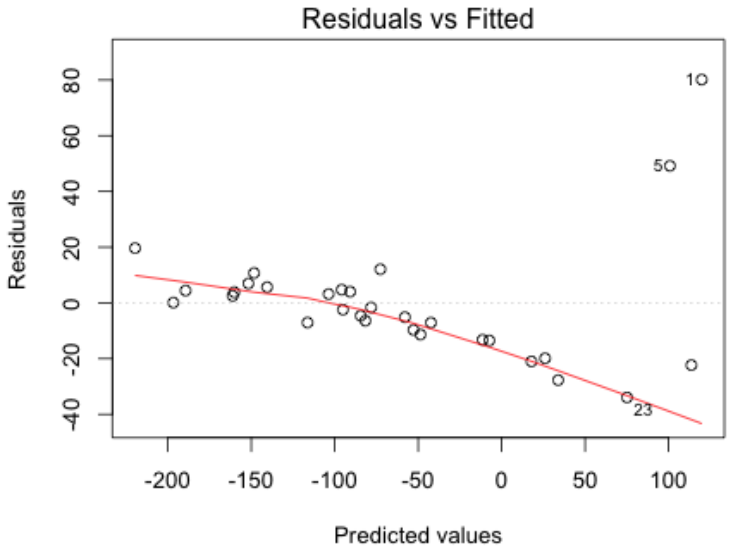
Null deviance: 261692 on 29 degrees of freedom

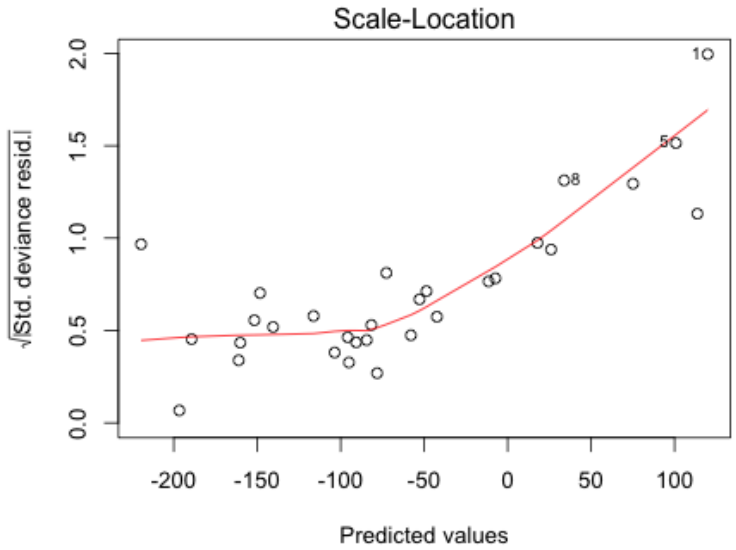
Residual deviance: 13688 on 25 degrees of freedom

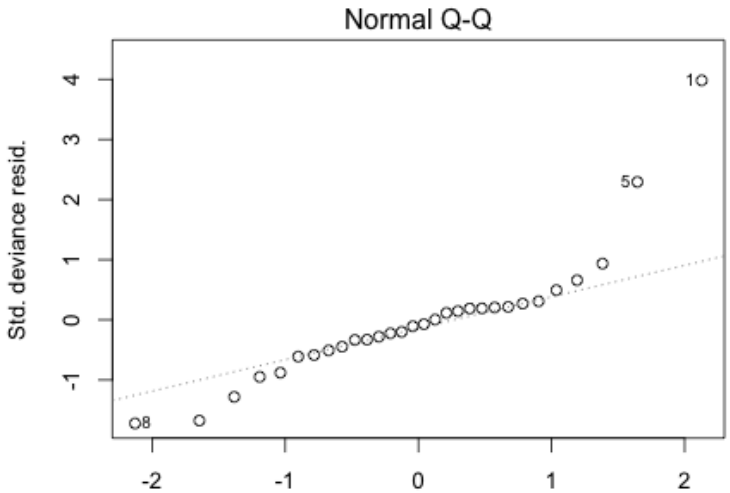
AIC: 280.83

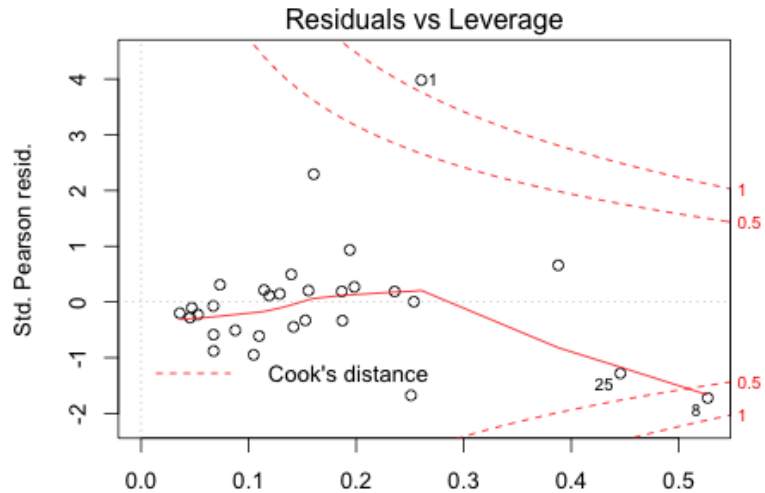
Задание 1. Пример. Выбросы.

Модель с выбросами









Задание 1. Пример. Другие остатки.

y_nnr=$0.5 * x1 + 2 * x2 + 3 * x1 * x3 + rnorm(30)^6$

Coefficients:

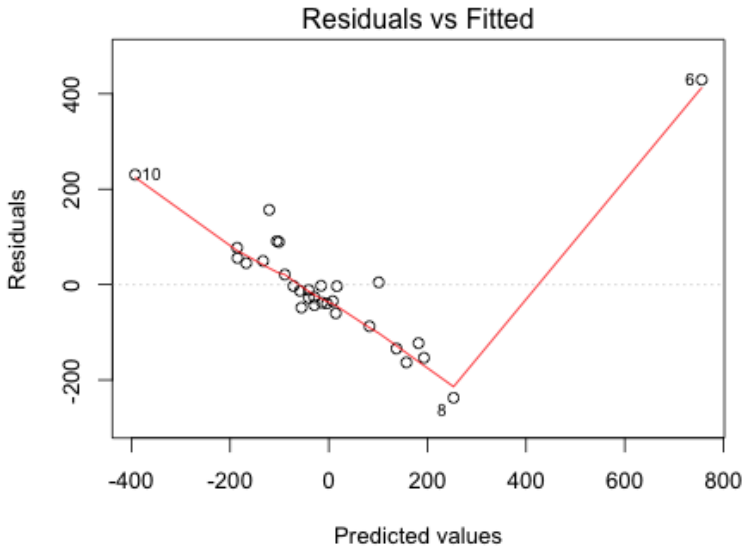
(Intercept)	x1	x2	x3	x1:x3
-326.975	32.946	-3.207	-210.707	23.862

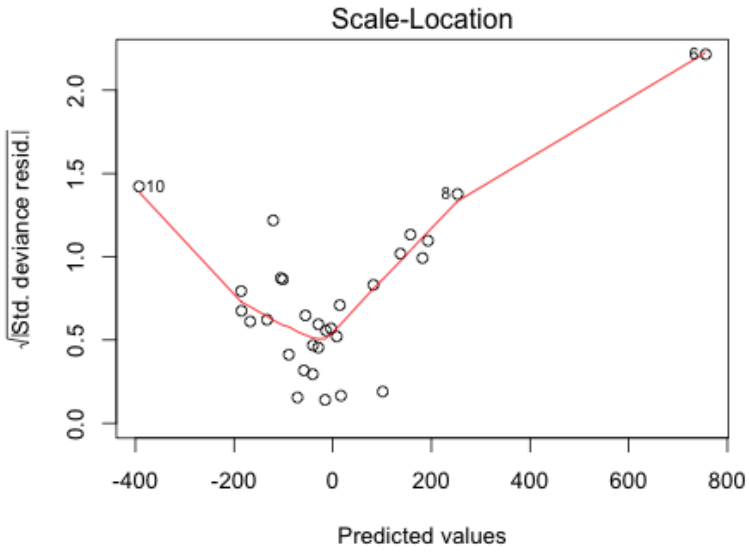
Degrees of Freedom: 29 Total (i.e. Null); 25 Residual

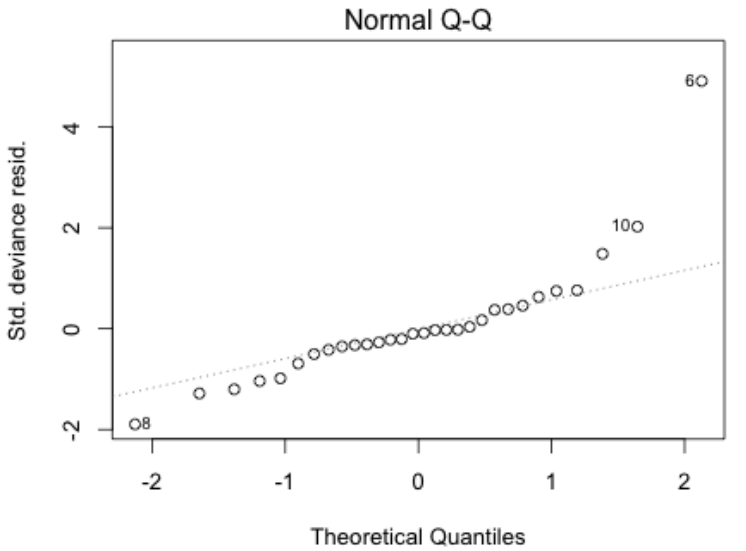
Null Deviance: 1550000

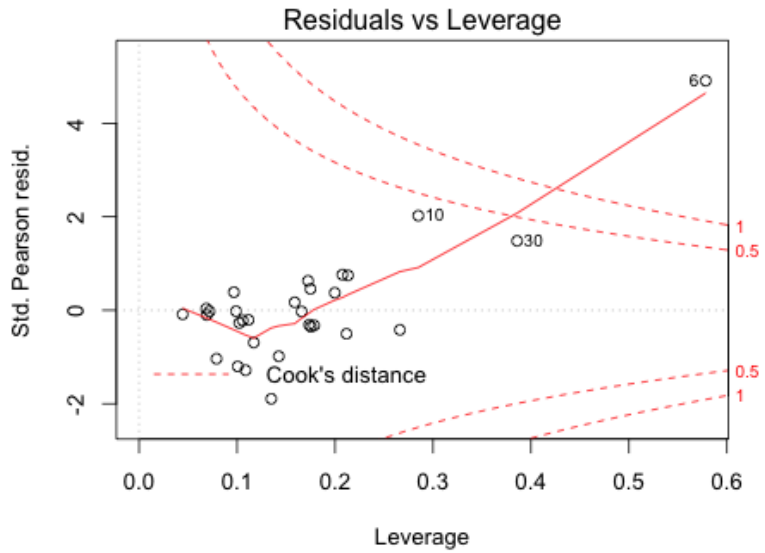
Residual Deviance: 453400 AIC: 385.8

Задание 1. Другие остатки.









Задание 2. Мультиколлинеарность, выбор модели. Forward , Backward selection

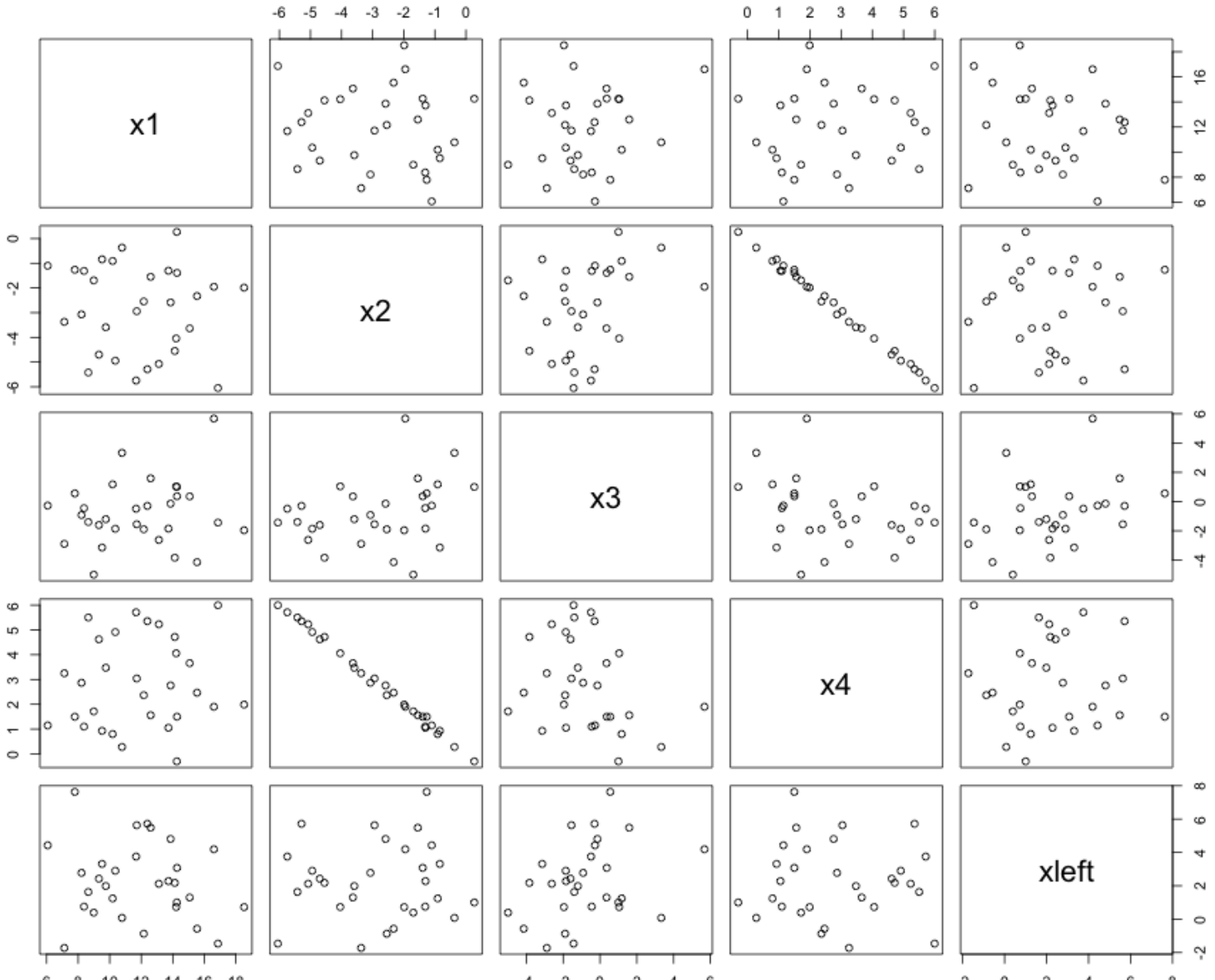
- Смоделируйте x_i и y. Смоделируйте ‘лишний‘ x: 1) x не зависит от существубщих переменных. 2) x сильно скоррелирован с какой-либо переменной. Сравните результат.
- Смоделируйте x_i и y. Сравните неполную модель и полную (ту, по которой построен y). Сравните полную модель и модель с ‘лишней‘ переменной
- Сделайте forward, backward selection для вашей модели по всем данным (y и все x_i)

Функции в R:

- anova(glm1, glm2, test="F")
- pairs(x1 + x2 + x3) - попарные графики зависимости переменных модели
- library(MASS)
- step = stepAIC(fit, direction="both");
- step($glm(y \sim 1, data = mydata), direction = "forward" \ scope = list(lower = \sim 1, upper = \sim x1 + x2... + x3 * x4)$)

Задание 2. Пример. Мультиколлинеарность.

Пример результата paires.



Задание 2. Пример. Мультиколлинеарность.

xleft=-rnorm(30,-3,2)+0.1*rnorm(30) - распределена как $-x_2$, но не скоррелирована с ней.

(Intercept)	x1	x2	x3	xleft	x1:x3
-1.37	0.55 ***	2.01***	-0.2	0.14	3.0 ***

x4=(-1)*x2 + 0.1*rnorm(30)

(Intercept)	x1	x2	x3	x4	x1:x3
-1.07	0.56 ***	2.61	-0.11	0.61	3.0 ***

Задание 2. Пример. Сравнение моделей.

Сравнение $glm(y \sim x1 + x2)$ и $glm(y \sim x1 + x2 + x1 * x3)$: pval = 2.2e-16

Сравнение $glm(y \sim x1 + x2 + x1 * x3)$ и $glm(y \sim x1 + x2 + x1 * x3 + x4)$: pval = 0.7208

Задание 2. Пример. Алгоритмы выбора модели (fb selection)

Справка: $AIC = 2k - 2ln(L)$, L - likelihood of the model, k - number of params

Коэффициент детерминации R^2: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$

fit = $glm(y \sim x1 + x2 + x3 + xleft + x1 * x3, data = mydata)$
step = step(fit, direction="both")
Step: AIC=88.37
y x1 + x2 + x3 + x1:x3

step =step(glm(y 1,data = mydata),
direction="forward $scope = list(lower =\sim 1, upper =\sim$
$x1 + x2 + x1 * x2 + x1 * x3 + x2 * x3 + x3 + xleft + xleft * x2))$
Step: AIC=88.37
y x3 + x1 + x2 + x3:x1

step <- step(fit, direction="backward")

Задание 3. Факторы и замена переменных.

- Смоделируйте x_i и y, такой что какой-то из x_i входит в него нелинейно (например + $ln(x_i)$). Постройте регрессию без замены переменных и с заменой, сравните результаты.
- Добавьте фактор в модель, посмотрите на результаты. Что значат коэффициенты у фактора?

Функции в R:

- xf=factor(x1) - создание категориальной переменной
- $glm(y \sim x1 + x2 + x3 + xleft + x4 + relevel(xf, 2))$ - указание уровня для сравнения (по умолчанию сравнивают с первым)

Задание 3. Пример. Без замены.

$y = 10 + 2 * x1 - 0.7 * x2 + 5 * (x3)^3 + rnorm(30)$

$glm(y \sim x1 + x2 + x3 + xleft)$

Значим только коэффициент у x3.

Coefficients:

(Intercept)	x1	x2	x3	xleft
27.6031	0.6472	-1.0485	32.0049	-3.3755

Degrees of Freedom: 29 Total (i.e. Null); 25 Residual

Null Deviance: 92400

Residual Deviance: 23250 AIC: 296.7

Задание 3. Пример. С заменой.

$x_zp = x3^3$

$y = 10 + 2 * x1 - 0.7 * x2 + 5 * x_zp + rnorm(30)$

Все коэффициенты, кроме xleft значимы.

Coefficients:

(Intercept)	x1	x2	x_zp	xleft
9.66778	1.98629	-0.74642	5.00393	0.07649

Degrees of Freedom: 29 Total (i.e. Null); 25 Residual

Null Deviance: 92400

Residual Deviance: 29.53 AIC: 96.66

Задание 3. Фактор

xfactor = 1,2,3

$y = 10 + 2 * x1 - 0.7 * x2 + 3 * x4 + 5 * (xfactor)^2 + rnorm(30)$

xf=factor(xfactor)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.41211	0.87291	19.947	1.41e-15 ***
x1	1.85203	0.04626	40.035	< 2e-16 ***
x2	-0.46727	0.10963	-4.262	0.000318 ***
x3	0.72581	0.30278	2.397	0.025453 *
xleft	0.01932	0.08481	0.228	0.821885
x4	2.94970	0.02280	129.386	< 2e-16 ***
xf2	15.02477	0.40679	36.935	< 2e-16 ***
xf3	39.53555	0.43493	90.901	< 2e-16 ***