Методы статистической обработки информации (практика)

02.08.2019

Метод наименьших квадратов в задаче линейной и нелинейной регрессии

1. Промоделировать нелинейную модель $y = f(x, a, b) + \delta$ с несмещенной нормально распределенной ошибкой, дисперсия которой равна ε, считая x стандартно нормально распределенной случайной величиной. В качестве примера рассматривается последний вариант, поэтому вариантов всего шесть.
   - $f(x, a, b) = e^{ax} + b, a = 0.4, b = 1, \varepsilon = 0.3$
   - $f(x, a, b) = ae^x + b, a = 0.4, b = 1, \varepsilon = 0.8$
   - $f(x, a, b) = a\cos x + b, a = 0.4, b = 1, \varepsilon = 0.8$
   - $f(x, a, b) = a\sin x + bx^2, a = 2, b = 1, \varepsilon = 0.8$
   - $f(x, a, b) = (x + a)^2 + bx^3, a = 1, b = 2, \varepsilon = 4$
   - $f(x, a, b) = (x + a)^3 + bx^2, a = 1, b = 2, \varepsilon = 4$
   - $f(x, a, b) = ax^2 + b, a = 2, b = 3, \varepsilon = 4$
2. Оценить параметры нелинейной модели по методу наименьших квадратов (численно). Применить к модельным данным линейную модель и оценить параметры. Построить на двумерной диаграмме основную и линейную модель. Сравнить невязки для обеих моделей.
3. Для линейной модели выполнить дисперсионный анализ, проверить значимость прогноза и коэффициентов регрессии. Сравнить непосредственные вычисления с результатами встроенной функции.
4. Промоделировать данные для множественной регрессии. Применить функцию lm. Ответить на вопросы о значимости коэффициента детерминации, частных коэффициентов регрессии, о коэффициенте корреляции между остатком и независимыми переменными.

Нелинейная модель

$$y_i = ax^2 + bx_i + \delta_i, \qquad \delta_i \sim \mathcal{N}(0, \sigma)$$

Модель одномерной линейной регрессии

$$y_i = \alpha + \beta x_i + \delta_i, \quad \delta_i \sim \mathcal{N}(0, \sigma)$$

```
# задаем нелинейную и линейную модель с остаточными суммами квадратов
N<-100
f<-function(x,ab) ab[1]*x^2+ab[2]*x
L<-function(X,Y,ab)sum((Y-f(X,ab))^2)

f0<-function(x,AB)AB[1]+AB[2]*x
L0<-function(X,Y,AB)sum((Y-f0(X,AB))^2)
```

```
# задаем параметры нелинейной модели
ab<-c(2,3)
eps<-4


#моделируем данные  нелинейной модели
X<-rnorm(N)
Y<-f(X,ab)+rnorm(N,0,eps)
SLM<-summary(lm(Y~X))
AB<-SLM$coefficients[,1]
Y.<-f0(X,AB)
```

Оценка параметров

$$\hat{\beta} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

```
#  оцениваем параметры линейной модели
EstLM<-function(X,Y)
{
  b.<-(sum(X*Y)-N*mean(X)*mean(Y))/(sum(X^2)-N*mean(X)^2);
  a.<-mean(Y)-AB[2]*mean(X)
c(a.,b.)
}
AB<-EstLM(X,Y);AB
```

```
##        X
## 1.674088 3.767186
```

Наилучший линейный прогноз

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} y_i$$

```
Y.<-f0(X,AB)
```

Источники вариации: общий Q_T, обусловленный регрессией Q_R, невязка Q_E. Коэффициент детерминации R^2.

$$Q_T = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad Q_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \quad Q_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad Q_T = Q_R + Q_E, \quad R^2 = \frac{Q_R}{Q_T}.$$

```
QT<-sum(( Y-mean(Y) )^2);QT
```

```
## [1] 3193.018
```

```
QR<-sum((Y.-mean(Y))^2);QR
```

```
## [1] 1122.749
```

```
QE<-sum((Y-Y.)^2);QE
```

```
## [1] 2070.269
```

```
R2<-QR/QT;R2
```

```
## [1] 0.3516262
```

```
c(QT,QE+QR)
```

```
## [1] 3193.018 3193.018
```

$$S^2 = \frac{Q_E}{n-2}\,, \quad S_\alpha^2 = \frac{S^2}{[x,x]}\cdot\frac{\sum_i x_i^2}{n}\,, \quad S_\beta^2 = \frac{S^2}{[x,x]}\,, \quad [x,x] = \sum_{i=1}^{n}(x_i-\bar{x})^2\,.$$

```
xx<-sum((X-mean(X))^2);xx
```

```
## [1] 79.1131
```

```
S2<-QE/(N-2)
S2a<-S2*sum(X^2)/N/xx
S2b<-S2/xx
```

Статистики для проверки значимости прогноза и коэффициентов регрессии

$$F = \frac{Q_R}{Q_E}(n-2) \;\sim\; \mathbf{F}(1,n-2), \quad T_\alpha = \frac{\hat{\alpha}-\alpha}{S_\alpha} \;\sim\; \mathbf{T}(n-2), \qquad T_\beta = \frac{\hat{\beta}-\beta}{S_\beta} \;\sim\; \mathbf{T}(n-2)\,.$$

```
F.<-QR/QT*(N-2);F.
```

```
## [1] 34.45937
```

```
Pf<-1-pf(F.,2,N-2)
Ta<-AB[1]/sqrt(S2a);Ta
```

```
##        X
## 3.642027
```

```
Tb<-AB[2]/sqrt(S2b);Tb
```

```
##
## 7.290225
```

```
Pa<-2*(1-pt(abs(Ta),N-2));Pa
```

```
##             X
## 0.0004345074
```

```
Pb<-2*(1-pt(abs(Tb),N-2));Pb
```

```
##
## 8.015855e-11
```

```
# проверяем при помощи встроенной функции
LM<-lm(Y~X)
SLM<-summary(LM);
cbind(AB,SLM$coefficients[,1])
```

```
##          AB
## X 1.674088 1.674088
##   3.767186 3.767186
```

```
c(R2=R2,SLM$r.squared)
```

```
##        R2
## 0.3516262 0.3516262
```

```
df<-SLM$df[seq(2)];df
```

```
## [1]  2 98
```

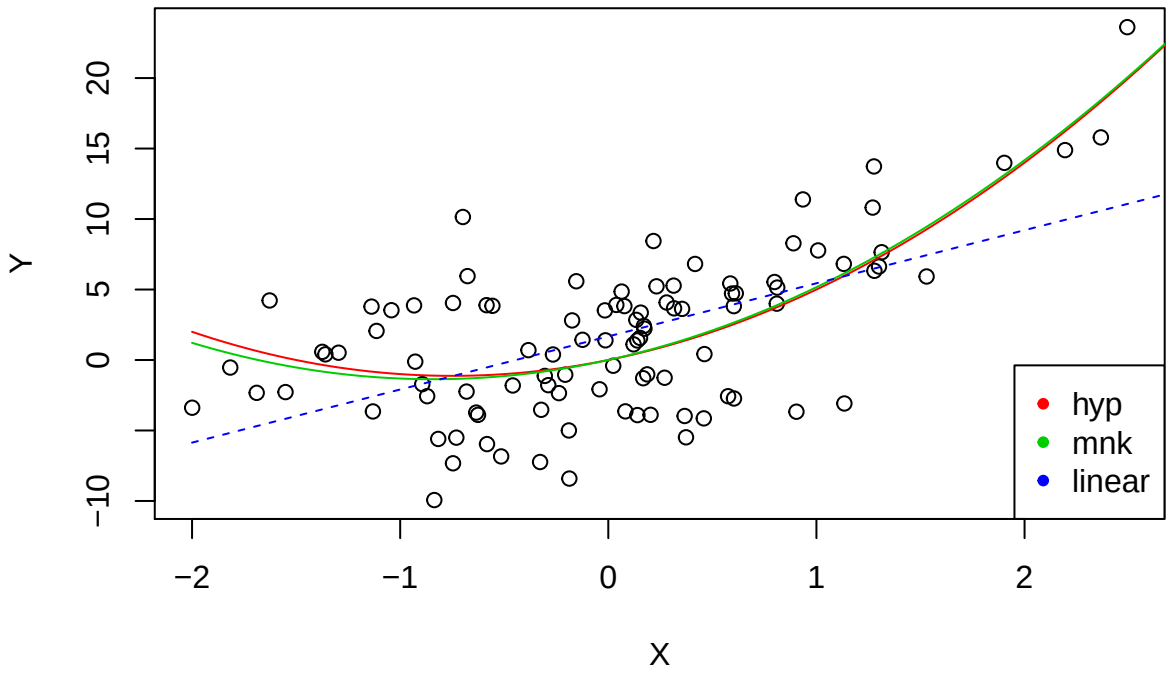
```
Pf.lm<-1-pf(SLM$fstatistic[1],df[1],df[2])
cbind(c(Pf=Pf,Pa=Pa,Pb=Pb),c(Pf=Pf.lm,SLM$coefficients[,4]))
```

```
##                [,1]          [,2]
## Pf    4.648726e-12 2.220446e-16
## Pa.X 4.345074e-04 4.345074e-04
## Pb    8.015855e-11 8.015840e-11
```

```
#  оцениваем параметры нелинейной модели
NLM<-nlm(function(ab)L(X,Y,ab),c(1,1))
ab.<-NLM$estimate
cbind(ab.=ab.,ab=ab)
```

```
##           ab. ab
## [1,] 1.924097  2
## [2,] 3.239133  3
```

```
plot(X,Y)
f_<-function(x)f(x,ab); curve(f_,-2,3,add=TRUE,col=2)
f_<-function(x)f(x,ab.); curve(f_,-2,3,add=TRUE,col=3)
f_<-function(x)f0(x,AB); curve(f_,-2,3,add=TRUE,col=4,lty=2)
legend('bottomright',c('hyp','mnk','linear'),pch=20,col=c(2,3,4))
```




```
# errors
c(Q.linear=L0(X,Y,AB),Q.model=L(X,Y,ab),Q.model.hat=L(X,Y,ab.))
```

```
##    Q.linear     Q.model Q.model.hat
##    2070.269    1624.648    1619.773
```

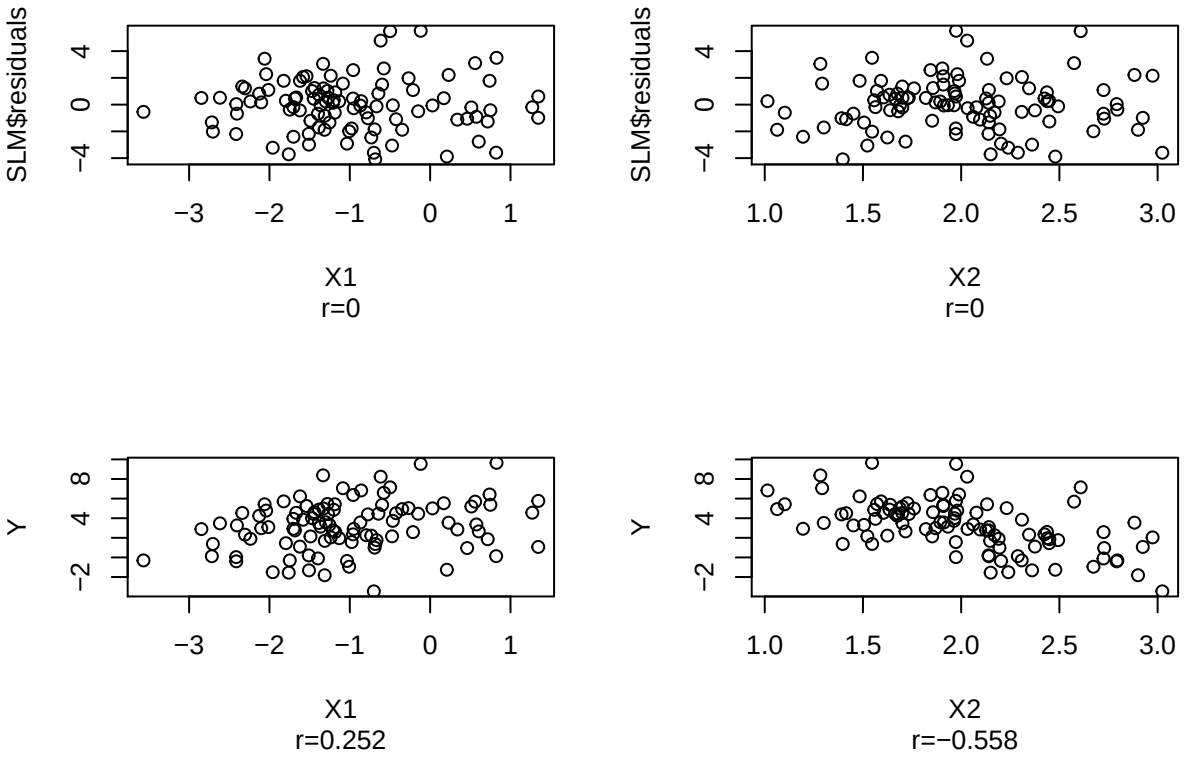
Множественная регрессия

```
N<-100
a<-c(1,-3,10)
eps<-2
X1<-rnorm(N,-1,1);X2<-rnorm(N,2,0.5)
Y<-a[1]*X1+a[2]*X2+a[3]+rnorm(N,0,eps)

LM<-lm(Y~X1+X2)
SLM<-summary(LM)
SLM
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0899 -1.1452 -0.0038  1.0900  5.5196
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.5478     0.9314  11.325  < 2e-16 ***
## X1            0.7790     0.1971   3.952 0.000147 ***
## X2           -3.2618     0.4353  -7.494 3.14e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.973 on 97 degrees of freedom
## Multiple R-squared:  0.4069, Adjusted R-squared:  0.3947
## F-statistic: 33.28 on 2 and 97 DF,  p-value: 9.897e-12
```

Некоррелированность остатков

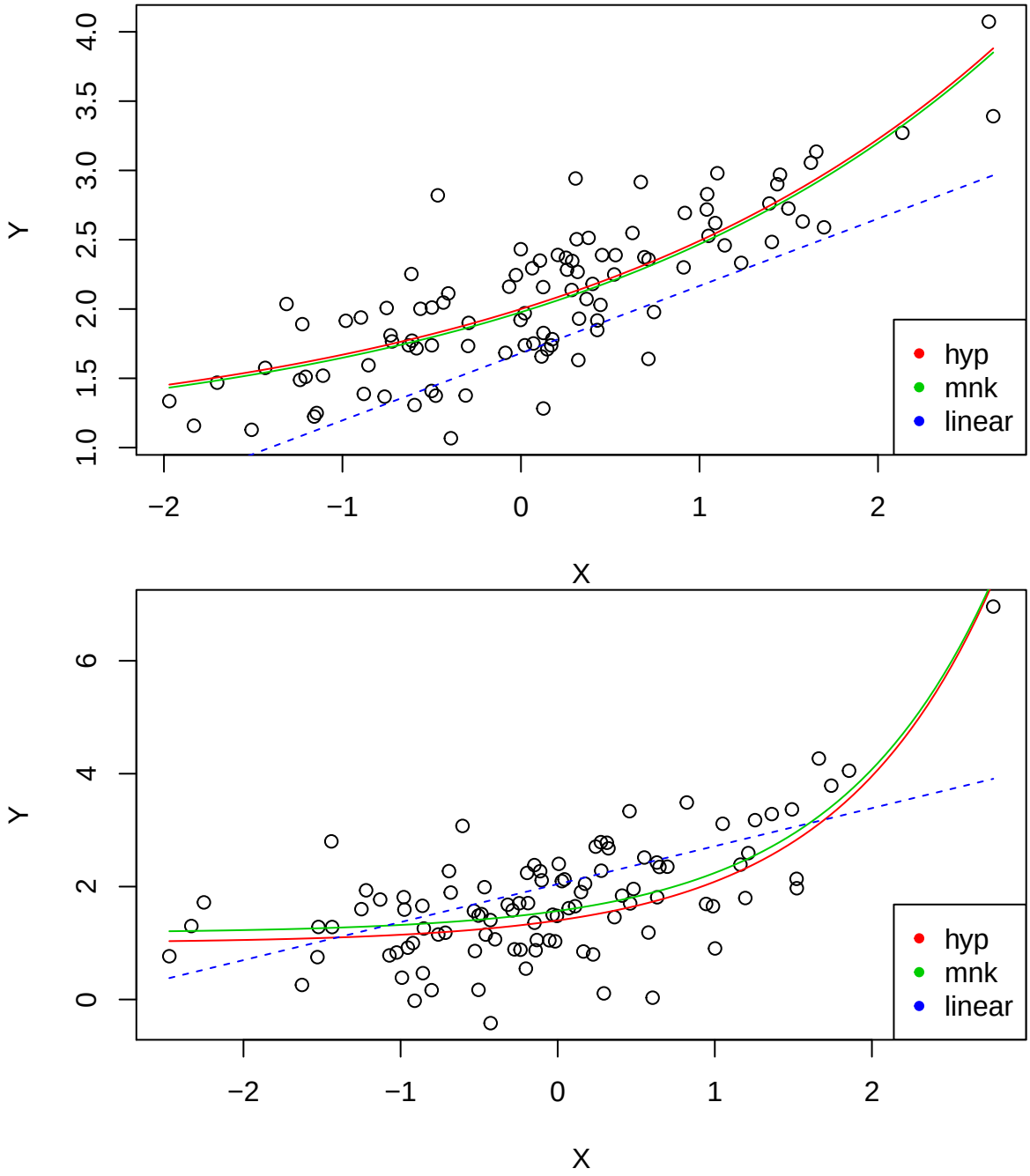
```
op <- par(mfrow = c(2, 2))
plot(X1,SLM$residuals)
title(sub=paste("r",round(cor(SLM$residuals,X1),3), sep="="))
plot(X2,SLM$residuals)
title(sub=paste("r",round(cor(SLM$residuals,X2),3), sep="="))

plot(X1,Y)
title(sub=paste("r",round(cor(X1,Y),3), sep="="))
plot(X2,Y)
title(sub=paste("r",round(cor(X2,Y),3), sep="="))
```

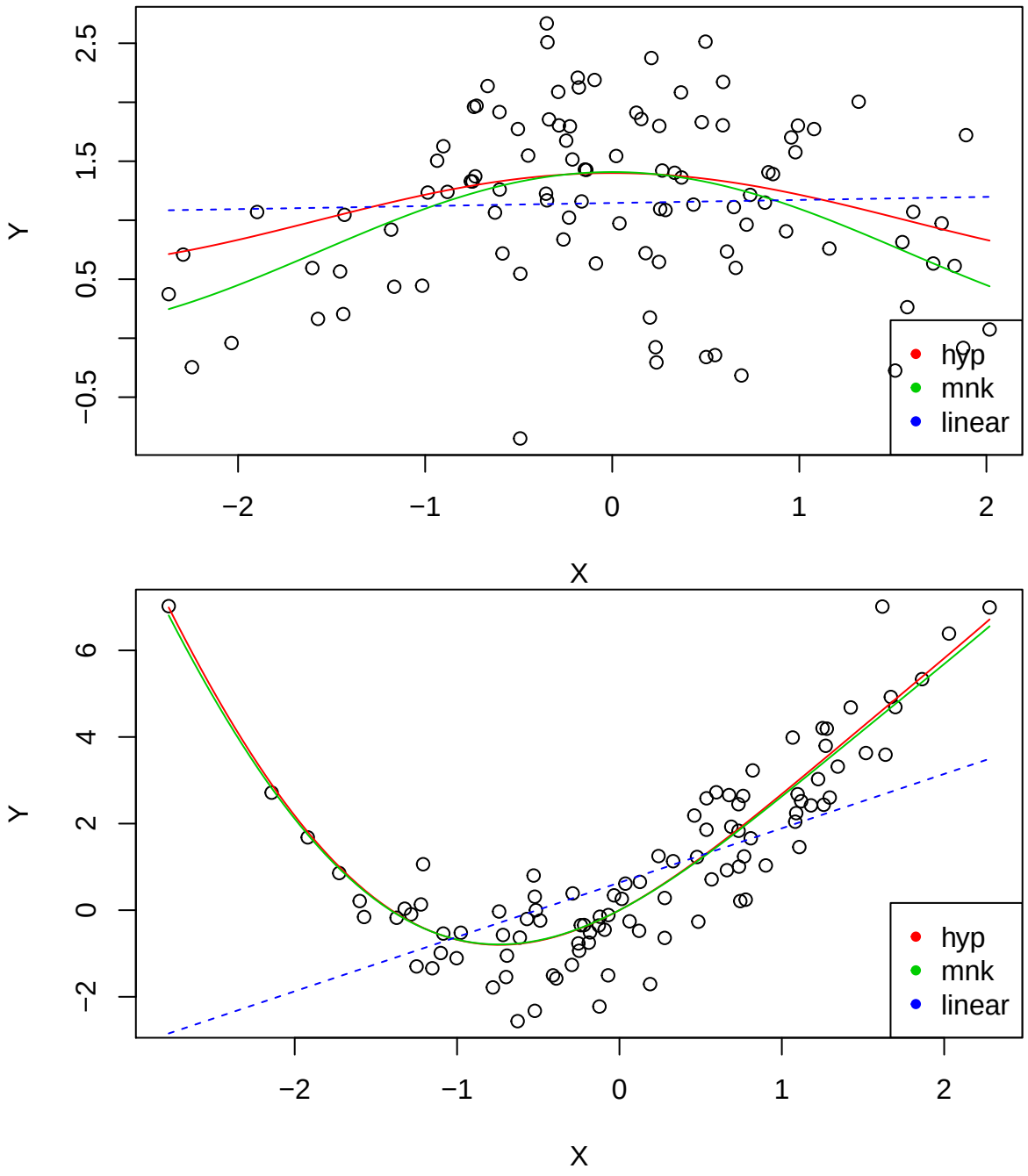



```
par(op)

cor(X1,SLM$residuals)
```

```
## [1] -4.516073e-18
```

Варианты




Y
2.5
1.5
0.5
−0.5
−2
−1
0
1
2
X
hyp
mnk
linear

Y
6
4
2
0
−2
−2
−1
0
1
2
X
hyp
mnk
linear