Практика №4 Метод наименьших квадратов в задаче линейной и нелинейной регрессии

Долаева А. Р., г .20.М04-мм

13/05/2021

Вариант №5.

$f(x, a, b) = (x + a)^2 + bx^3, a = 1, b = 2, \varepsilon = 4$

Задача - промоделировать заданную нелинейную модель и проверить зависимость между переменными x и y.

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.

Задаем нелинейную (f) по варианту и линейную модели (f0) для вычисления отдаленности отклонения наблюдений от прямой, а также их парметры.

```
N<-100;
ab<-c(1,2);
eps<-4;

f<-function(x,ab) (ab[1]+x)^2+ab[2]*x^3;
f0<-function(x,AB)AB[1]+AB[2]*x
```

Моделируем данные нелинейной модели (Y),
где вектор X задается в соответствии с стандартным нормальным распределением,
Y - в соответствии с нормальным распределением, дисперсия которого равна ε.

Моделируем данные линейной модели (Y0), используя коэффициенты (AB), вычисленные регрессионным анализом (lm).

```
X<-rnorm(N);
Y<-f(X,ab)+rnorm(N,0,eps);

AB<-summary(lm(Y~X))$coefficients[,1];AB;
```

```
## (Intercept)           X
##    2.272380    7.734528
```

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

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.

Функции вычисления ошибок (остаточными суммами квадратов). $\sum_{i=1}^{N}(y_i - f(x_i, a, b))^2$
L для нелинейной модели, L0 - линейной.

```
L<-function(X,Y,ab)sum((Y-f(X,ab))^2);
L0<-function(X,Y,AB)sum((Y-f0(X,AB))^2)
```

Оцениваем параметры нелинейной модели:
минимизируем (nlm) остаточные суммы квадратов (L); начальные параметры минимизации c(1,1);
выводим вычисленные (ab.) и начальные коэффициенты (ab).

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

```
##            ab. ab
## [1,] 1.008168  1
## [2,] 2.072893  2
```

Оценки близки к начальным коэффициентам.

Оцениваем параметры линейной модели $\hat{\beta} = \frac{\sum_i x_i y_i - n\overline{x}\overline{y}}{\sum_i x_i^2 - n\overline{x}^2}$, $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$

Из формул вычисляем коэффициенты (a, b):

```
EstLM<-function(X,Y)
{
   mmx<-mean(X);
   mmy<-mean(Y);
   b.<-(sum(X*Y)-N*mmx*mmy)/(sum(X^2)-N*mmx^2);
   a.<-mmy-AB[2]*mmx;
   c(a.,b.)
}

AB<-EstLM(X,Y);
c(a=AB[1], b=AB[2])
```

```
##      a.X        b
## 2.272380 7.734528
```

Наилучший линейный прогноз: $\hat{y}_i = \hat{\alpha} + \hat{\beta} y_i$

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

Источники вариации:

общий: $Q_T = \sum_{i=1}^{N}(y_i - \overline{y})^2$
обусловленный регрессией: $Q_R = \sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2$
невязка: $Q_E = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

y_i - значения наблюдаемой переменной (Y);
$\overline{y}$ - среднее значение по наблюдаемым данным;
$\hat{y}_i$ - модельные значения,построенные по оцененным параметрам (Y.).

```r
QT<-sum((Y-mean(Y) )^2);
QR<-sum((Y.-mean(Y))^2);
QE<-sum((Y-Y.)^2);
c(QT=QT, QR=QR, QE=QE);
```

```
##       QT       QR       QE
## 8768.212 5752.673 3015.539
```

```r
paste('equality check');
```

```
## [1] "equality check"
```

```r
c(QT=QT,'QE+QR'=QE+QR)
```

```
##       QT    QE+QR
## 8768.212 8768.212
```

Коэффициент детерминации: $R^2 = \frac{Q_R}{Q_T}$

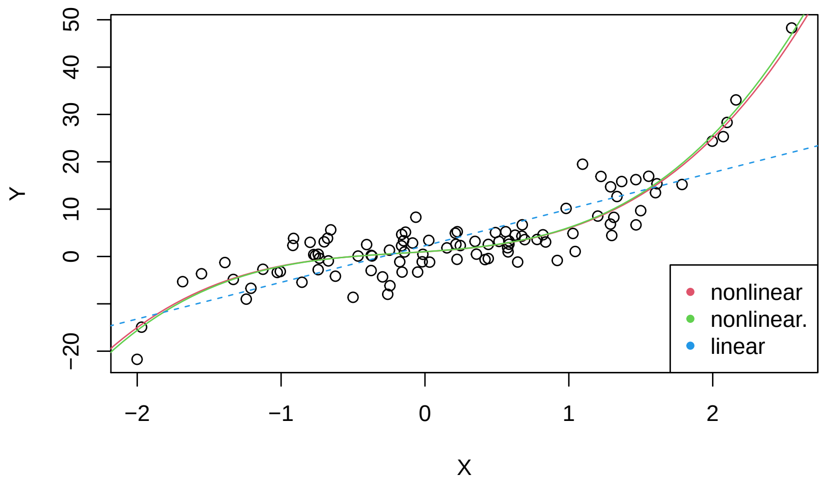
```r
R2<-QR/QT;R2
```

```
## [1] 0.6560828
```

Коэффициент детерминации показывает, какая доля вариации объясняемой переменной Y обусловлена влиянием на нее фактора X.
Хорошим показателем R^2 является значение выше 0.8,
если больше 0.5, то расчетные параметры модели объясняют зависимость и изменения изучаемого параметра Y от исследуемого фактора X,
а если меньше 0.5, то смысл такой модели можно смело ставить под большой вопрос, и зависимость параметра Y от исследуемых фактора X скорее всего отсутствует.

Двумерная диаграмма
нелинейной модели с заданными параметрами (ab), выделенной красным;
нелинейной модели с оцененными параметрами (ab.), выделенной зеленым;
линейной модели с оцененными параметрами (AB), выделенной синим.

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



Сраваем невязки для обеих моделей:

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

```
##        Q.linear     Q.nonlinear Q.nonlinear.hat
##        3015.539        1300.022        1294.150
```

Обычно наименьшая сумма квадратов отклонений у модели с оцененными параметрами (Q.nonlinear.hat), как и в нашем случае.

3. Для линейной модели выполнить дисперсионный анализ, проверить значимость прогноза и коэффициентов регрессии. Сравнить непосредственные вычисления с результатами встроенной функции.

Дисперсионный анализ: $[x,x] = \sum_{i=1}^{N}(x_i - \overline{x})^2$ $S^2 = \frac{Q_E}{n-2}$, $S^2_\alpha = \frac{S^2}{[x,x]}\frac{\sum_i x_i^2}{n}$, $S^2_\beta = \frac{S^2}{[x,x]}$

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

c(xx=xx, S2=S2, S2a=S2a, S2b=S2b)
```

```
##         xx         S2        S2a        S2b
## 96.1616968 30.7708038  0.3195956  0.3199902
```

Статистики для проверки значимости прогноза: $F = \frac{Q_R}{Q_E}(n-2) \sim F(1, n-2)$ $T_\alpha = \frac{\hat{a}-a}{S_\alpha} \sim T(n-2)$ $T_\beta = \frac{\hat{b}-b}{S_\beta} \sim T(n-2)$

```
F.<-QR/QT*(N-2);
Ta<-AB[1]/sqrt(S2a);
Tb<-AB[2]/sqrt(S2b);
```

```
c(F.=F., Ta=Ta, Tb=Tb)
```

```
##        F.      Ta.X        Tb
## 64.296117  4.019579 13.673051
```

Высокая значимость прогноза по Фишеру (F.), так как высокий доверительный уровень.

Провериям значимость коэффициентов регрессии:

```
Pf<-1-pf(F.,1,N-2);
Pa<-2*(1-pt(abs(Ta),N-2));
Pb<-2*(1-pt(abs(Tb),N-2));

c(Pf=Pf, Pa=Pa, Pb=Pb)
```

```
##           Pf         Pa.X           Pb
## 2.316702e-12 1.145481e-04 0.000000e+00
```

Высокая значимость коэффициентов регрессии pvalue<0.05. Значимые отконения от нуля.

Проверяем при помощи встроенной функции: коэффициенты, вычисленные вручную (ab) и при помощи встроенной функции (ab.)

```
LM<-lm(Y~X)
SLM<-summary(LM);
cbind("ab"=AB, "ab."=SLM$coefficients[,1])
```

```
##         ab      ab.
## X 2.272380 2.272380
##   7.734528 7.734528
```

Коэффициенты детерминации: вычисленные для параметров вручной (R2) и для встроенной функции (R2.)

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

```
##        R2       R2.
## 0.6560828 0.6560828
```

Проверяем значимость прогноза и коэффициентов регрессии:

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

```
##        [,1]  [,2]
## Pf    0e+00 0e+00
## Pa.X 1e-04 1e-04
## Pb    0e+00 0e+00
```

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.

Параметры подбирались самостоятельно, так как в задании они не были указаны.
$f(x1, x2, a, b) = ax + bx + c, a = 1, b = 2, c = 7, \varepsilon = 4$
Для X1 $\mu = -1$ $\varepsilon = 1$, для X2 $\mu = 2$ $\varepsilon = 0.5$

```
N<-100
a<-c(1,2,7)
eps<-4
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
## -11.3557  -2.6235   0.2195   2.4958   8.9720
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.5471     1.6953   3.862 0.000203 ***
## X1            0.8567     0.4039   2.121 0.036461 *
## X2            2.0938     0.8144   2.571 0.011662 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.11 on 97 degrees of freedom
## Multiple R-squared:  0.08898,    Adjusted R-squared:  0.07019
## F-statistic: 4.737 on 2 and 97 DF,  p-value: 0.01089
```

коэффициенты (Estimate);
Высокая значимость коэффициентов регрессии pvalue<0.05. Pr(>|t|) по Стьюденту;
Степени свободы 2 и 97;
Multiple R-squared меньше 50%, зависимость параметра Y от исследуемых факторов X1 и X2 скорее всего отсутствует.

Проверим согласованность остатков нормальному распределению:

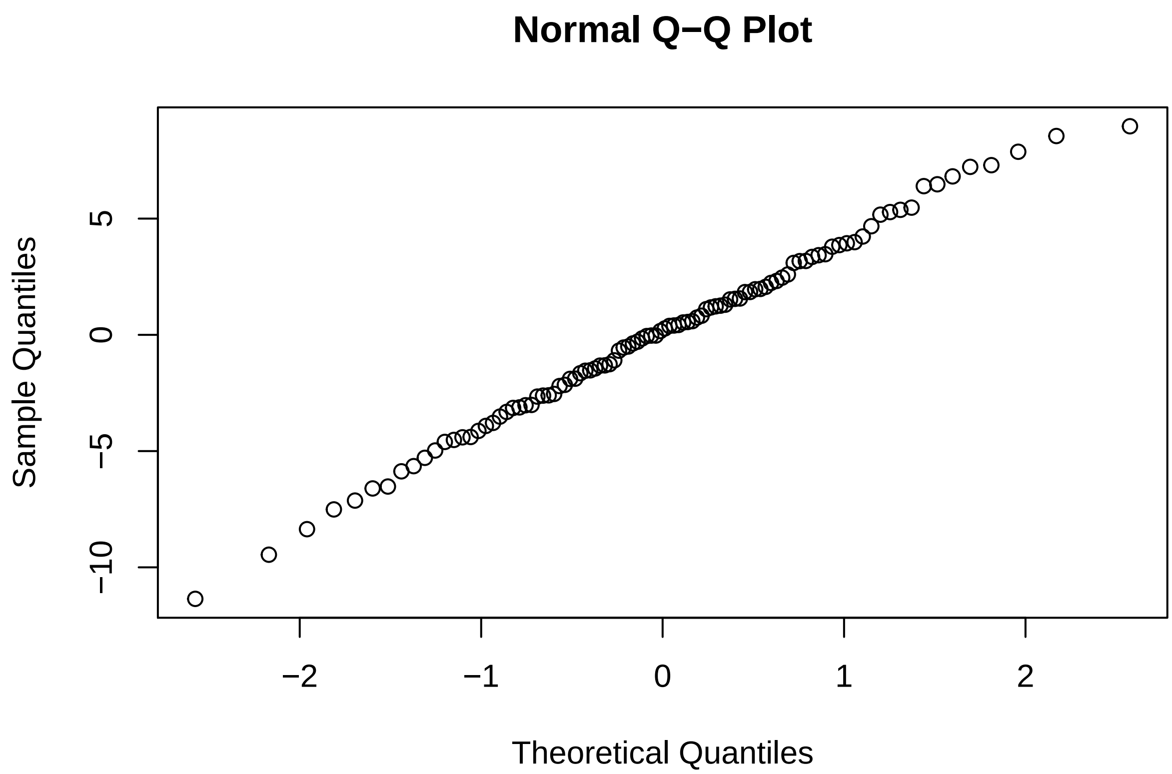
```
shapiro.test(SLM$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SLM$residuals
## W = 0.995, p-value = 0.9751
```

Высокий доверительный уровень, остатки согласованы с нормальным распределением.

Построение QQPlot:

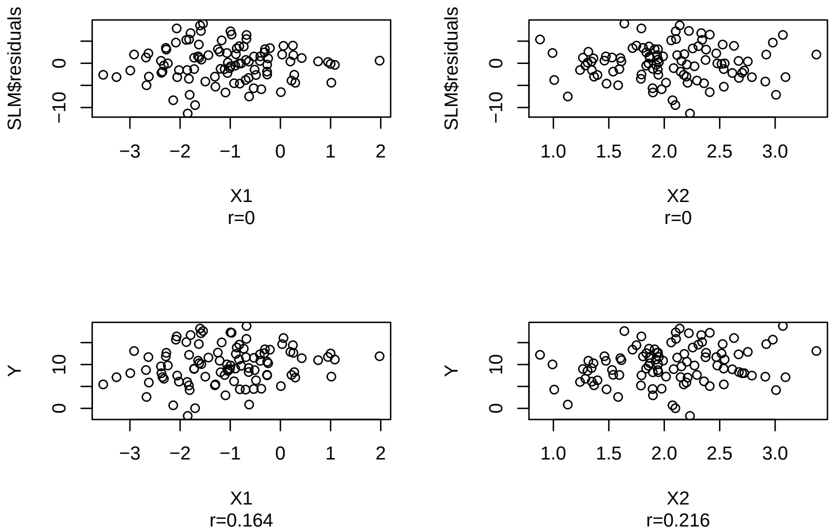
```
qqnorm(SLM$residuals)
```



Значения остатков близки линейной модели, что подтверждает согласованность с нормальным распределением.

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="="))
```



Отсутствует коррелированность остатков с параметрами X1 и X2. А также слабая отрицательная корреляция между У факторами X1 и X2.