Методы статистической обработки информации. Задание 3

Ершов А. С., гр. 22.М04-мм

Вариант (b)

$$f(x, a, b) = ae^x + b, \quad a = 0.4, b = 1, \varepsilon = 0.8$$

Задание:

1. Промоделировать нелинейную модель $y = f(x, a, b) + \delta$ с несмещенной нормально распределенной ошибкой, дисперсия которой равна ε, считая хстандартно нормально распределенной случайной величиной.
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.

Нелинейная модель

$$y_i = ax_i^2 + bx_i + \delta_i, \quad \delta_i\, N(0, \sigma)$$

Модель одномерной линейной регрессии

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

```
# Задаём линейную и нелинейную модель с остаточными суммами квадратов
set.seed(19)
N <- 100
f <- function(x, ab)
  ab[1] * exp(x) + ab[2]
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(0.4, 1)
eps <- 0.8
```

```
# 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 $\hat{\alpha}$, $\hat{\beta}$

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

```
# 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
## 1.7930499 0.8131022
```

Наилучший линейный прогноз

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

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

Вычислим источники вариации - общий Q_T, обусловленный регрессией Q_R, невязка Q_E и коэффициент детерминации R^2.

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

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

```
## [1] 174.5807
```

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

```
## [1] 69.03135
```

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

```
## [1] 105.5493
```

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

```
## [1] 0.3954123
```

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

```
## [1] 174.5807 174.5807
```

Равенство $Q_T = Q_R + Q_E$ выполняется.

Вычислим

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

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

```
## [1] 104.4134
```

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

Посчитаем статистики для проверки значимости прогноза и коэффициентов регрессии.

$$F = \frac{Q_R}{Q_e}(n-2) \sim \mathbf{F}(1, n-2)$$

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

```
## [1] 38.7504
```

Рассчитаем функцию распределения **F**:

```
Pf <- 1 - pf(F., 1, N - 2)
Pf
```

```
## [1] 1.197418e-08
```

Вычислим

$$T_\alpha = \frac{\hat{\alpha} - \alpha}{S_\alpha} \sim \mathbf{T}(n-2).$$

```
Ta <- AB[1] / sqrt(S2a)
Ta
```

```
##        X
## 17.27606
```

Рассчитаем

$$T_\beta = \frac{\hat{\beta} - \beta}{S_\beta} \sim \mathbf{T}(n-2).$$

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

```
##
## 8.005869
```

Посчитаем функции распределения $\mathbf{T}_a$ и $\mathbf{T}_b$:

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

```
## X
## 0
```

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

```
##
## 2.464695e-12
```

Для проверки воспользуемся встроенной функцией и сравним результаты оценок параметров линейной модели:

```
LM <- lm(Y ~ X)
SLM <- summary(LM)

cbind(AB, SLM$coefficients[, 1])
```

```
##          AB
## X 1.7930499 1.7930499
##   0.8131022 0.8131022
```

Вычисленные значения оценок и значения, полученные с использованием встроенной функции, совпали.

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

```
##        R2
## 0.3954123 0.3954123
```

То же верно для коэффициента детерминации R^2, значения совпали.

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

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

Посмотрим на результаты расчетов значений P_f, P_a, P_x.

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

```
##              [,1]         [,2]
## Pf   1.197418e-08 2.464917e-12
## Pa.X 0.000000e+00 1.673583e-31
## Pb   2.464695e-12 2.464856e-12
```

Видно, что рассчитанные значения P_a почти совпали, значения P_f, P_a не совпали, но близки.

Для нелинейной модели результаты следующие:

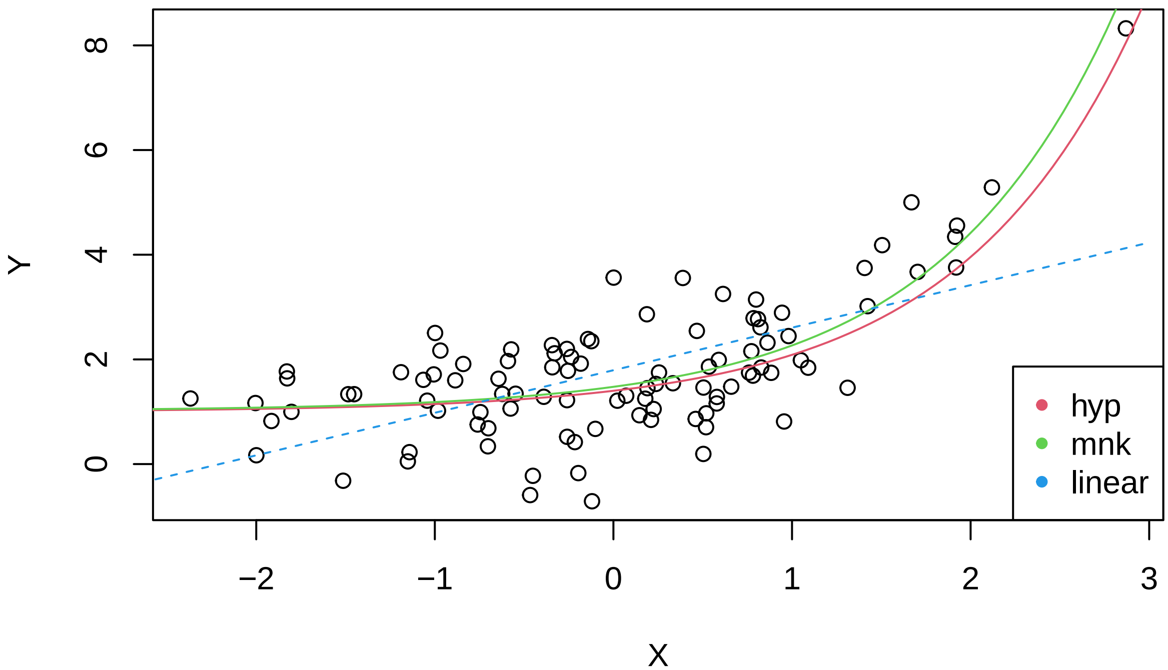
```r
NLM <- nlm(function(ab)
  L(X, Y, ab), c(1, 1))
ab. <- NLM$estimate
cbind(ab. = ab., ab = ab)
```

```
##            ab.  ab
## [1,] 0.4601712 0.4
## [2,] 1.0156014 1.0
```

Значения параметров нелинейной модели почти совпали.

Визуализируем модели, построив диаграмму. Красным отмечена основная модель, зеленым - нелинейная с параметрами полученными по результатам оценки, синим - линейная модель.

```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('hyp', 'mnk', 'linear'),
       pch = 20,
       col = c(2, 3, 4))
```



Вычислим невязки моделей:

```
# Ошибки
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
##   105.54934    69.48832    66.22923
```

Значения невязок исходной модели и модели с параметрами полученными по результатам оценки близки, в то же время невязки нелинейной и линейной модели отлчиаются в полтора раза.