Statistical Linear Models

Homework 5

李乐平 12112627

Dec. 2023

1. Solution:

(a). Rewrite the model in matrix form and get

$$\vec{y} = P\vec{a} + \vec{\varepsilon},$$

where

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, P = \begin{bmatrix} P_0(x_1) & P_1(x_1) & \cdots & P_{p-1}(x_1) \\ P_0(x_2) & P_1(x_2) & \cdots & P_{p-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_0(x_n) & P_1(x_n) & \cdots & P_{p-1}(x_n) \end{bmatrix}, \vec{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{p-1} \end{bmatrix}, \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

and $\vec{\varepsilon} \sim N(\vec{0}, \sigma^2 I)$.

Considering $\sum_{i=1}^{n} P_l(x_i)P_m(x_i) = 0, \forall l, m$, by applying Least Square Method to the model above, we can get

$$\hat{\vec{a}} = (P^T P)^{-1} P^T \vec{y} = \operatorname{diag}\left(\frac{1}{\sum_{i=1}^{n} P_0^2(x_i)}, \cdots, \frac{1}{\sum_{i=1}^{n} P_{p-1}^2(x_i)} \right) P^T \vec{y} = \begin{bmatrix} \frac{\sum_{i=1}^{n} y_i P_0(x_i)}{\sum_{i=1}^{n} P_0^2(x_i)} \\ \frac{\sum_{i=1}^{n} y_i P_1(x_i)}{\sum_{i=1}^{n} P_1^2(x_i)} \\ \vdots \\ \frac{\sum_{i=1}^{n} y_i P_{p-1}(x_i)}{\sum_{i=1}^{n} P_{p-1}^2(x_i)} \end{bmatrix}.$$

Here, we assume (P^TP) is invertible.

Hence the least square estimator of $\hat{a}_j, j = 0, 1, ..., p - 1$ is $\frac{\sum_{i=1}^{n} y_i P_j(x_i)}{\sum_{i=1}^{n} P_j^2(x_i)}$.

Because $\operatorname{Cov}(\hat{\vec{\alpha}}) = \operatorname{Cov}((P^T P)^{-1} P^T Y) = (P^T P)^{-1}\sigma^2 = \sigma^2 \operatorname{diag}(\sum_{i=1}^{n} P_0^2(x_i), \cdots, \sum_{i=1}^{n} P_{p-1}^2(x_i))$ is a diagonal matrix, so that guarantees $\hat{a}_j$ s are uncorrelated for $j = 0, 1, ..., p - 1$.

(b). We want to test H_0: $a_j = 0$, against H_1: $a_j \neq 0$. Here we denote

$\hat{\sigma}^2 = \mathrm{SSE}/(n-p) = (\vec{y} - P\hat{\vec{a}})^T (\vec{y} - P\hat{\vec{a}})/(n-p)$.

Note that $a_j \sim N(a_j,\ \sigma^2 \sum_{i=1}^{n} P_j^2(x_i))$ and $SSE = (\vec{y} - P\hat{\vec{a}})^T (\vec{y} - P\hat{\vec{a}}) = \vec{y}^T (I - P(P^T P)^{-1} P)\vec{y}$.

Set test statistics $t = \frac{a_j - \hat{a}_j}{\hat{\sigma} / \sum_{i=1}^{n} P_j^2(x_i)} \sim t(n-p-1)$, so we have to reject H_0 if

$$\hat{a}_j \notin \left[\frac{-t(\frac{\alpha}{2}, n-p-1)\hat{\sigma}}{\sum_{i=1}^{n} P_j^2(x_i)}, \frac{t(\frac{\alpha}{2}, n-p-1)\hat{\sigma}}{\sum_{i=1}^{n} P_j^2(x_i)} \right]$$ under the significance level of α.

(c) Under the condition of $x = x^*$, we have

$$E(\vec{y}^*) - \hat{\vec{y}}^* \sim N(0, t(\frac{\alpha}{2}, n-p-1)\hat{\sigma}\sqrt{\vec{p}^*(P'P)^{-1}\vec{p}^*}),$$

where

$$\vec{p}^{*T} = \begin{bmatrix} P_0(x^*) & P_1(x^*) & \cdots & P_{p-1}(x^*) \end{bmatrix}.$$

So a 100(1-α)% confidence interval for the mean of y^* at $x = x^*$ is

$$[\vec{p}^{*T}\hat{\vec{a}} - t(\frac{\alpha}{2}, n-p-1)\hat{\sigma}\sqrt{\vec{p}^*(P'P)^{-1}\vec{p}^*}, \vec{p}^{*T}\hat{\vec{a}} + t(\frac{\alpha}{2}, n-p-1)\hat{\sigma}\sqrt{\vec{p}^*(P'P)^{-1}\vec{p}^*}].$$

2. Solution:

(a).

```
data = read.csv("./6data.csv")
summary(data)
##          Y                X1               X2               X3
##  Min.   :10.50   Min.   : 1.00   Min.   : 3.000   Min.   :0.0000
##  1st Qu.:13.75   1st Qu.: 2.00   1st Qu.: 8.143   1st Qu.:0.0000
##  Median :14.81   Median : 3.50   Median : 9.550   Median :0.0400
##  Mean   :14.69   Mean   : 7.55   Mean   : 9.505   Mean   :0.0870
##  3rd Qu.:16.31   3rd Qu.:14.00   3rd Qu.:11.235   3rd Qu.:0.1225
##  Max.   :17.50   Max.   :16.00   Max.   :14.620   Max.   :0.6000
##          X4
##  Min.   : 3.001
##  1st Qu.: 8.181
##  Median :12.650
##  Mean   :15.780
##  3rd Qu.:23.899
##  Max.   :36.601
fit = lm(Y ~ X1 + X2 + X3 + X4, data = data)
summary(fit)
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.17355 -0.55425 -0.00316  0.61569  2.02727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.22211    0.71119  17.185  < 2e-16 ***
## X1          -0.18698    0.02497  -7.489 9.04e-09 ***
## X2           0.29510    0.07349   4.016 0.000298 ***
## X3          -1.21196    1.40668  -0.862 0.394786
## X4           0.07479    0.01637   4.569 5.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.9353 on 35 degrees of freedom
## Multiple R-squared:  0.7541, Adjusted R-squared:  0.726
## F-statistic: 26.84 on 4 and 35 DF,  p-value: 3.088e-10
```

A multiple linear model of $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ has been established now. And the test H0: $\beta=0$ against H1: at least some $\beta_i \neq 0$ can be tested through a F-statistic. Since the p-value = 3.088e-10 << 0.05, we should reject H0 and believe that the model is effective.

(b).

```
r1 = rstandard(fit)
r2 = rstudent(fit)
leverage = hatvalues(fit)
dffits = dffits(fit)
cooks = cooks.distance(fit)

result = data.frame(r1, r2, leverage, dffits, cooks)
names(result) = c(
  "SR",              # Studentized Residual
  "SDR",             # Studentized Deleted Residual",
  "LV",              # Leverage Value",
  "DFFITS",          # DFFITS
  "CD"               # Cook's Distance
)

result
##                   SR             SDR            LV           DFFITS                CD
## 1   -0.881313929 -0.878434210 0.13469748 -0.3465811939 2.418147e-02
## 2   -0.547027035 -0.541475420 0.14719530 -0.2249577359 1.032980e-02
## 3    0.114526120  0.112899333 0.36914451  0.0863623876 1.534990e-03
## 4    0.001844146  0.001817610 0.08380724  0.0005497282 6.221786e-08
## 5    0.908126276  0.905794106 0.05964149  0.2281166588 1.046110e-02
## 6   -2.520284183 -2.745620746 0.14979177 -1.1524494177 2.238163e-01
## 7   -0.239868320 -0.236611361 0.10028484 -0.0789952047 1.282644e-03
## 8    2.137663634  2.259566038 0.52194545  2.3610156478 9.978296e-01
## 9    2.429263793  2.625895934 0.20392176  1.3290197065 3.023341e-01
## 10   0.503208583  0.497771708 0.05951632  0.1252196844 3.204873e-03
## 11   0.730853920  0.725897877 0.16180864  0.3189369120 2.062290e-02
## 12  -0.303250984 -0.299280866 0.07711837 -0.0865136968 1.536902e-03
## 13   0.775534861  0.771029052 0.07742031  0.2233553000 1.009447e-02
## 14   0.271122006  0.267501818 0.15000359  0.1123748224 2.594443e-03
## 15  -0.131615597 -0.129753863 0.10232066 -0.0438067804 3.948997e-04
## 16  -0.941069237 -0.939490166 0.14157544 -0.3815356611 2.921184e-02
## 17   0.400112460  0.395260142 0.06830628  0.1070229134 2.347370e-03
## 18  -1.151851331 -1.157426559 0.14920288 -0.4846955654 4.653439e-02
```

```
## 19 -0.586051309 -0.580473602 0.18019839 -0.2721469992 1.509883e-02
## 20 -0.608227275 -0.602668830 0.05929007 -0.1513010173 4.663243e-03
## 21  0.067067672  0.066106867 0.09300209  0.0211684851 9.224501e-05
## 22  0.338608468  0.334284136 0.09243832  0.1066850849 2.335616e-03
## 23 -0.008970208 -0.008841144 0.09775913 -0.0029102204 1.743692e-06
## 24  0.270208563  0.266598685 0.08846972  0.0830557620 1.417267e-03
## 25 -0.431894646 -0.426818898 0.12784686 -0.1634151456 5.468686e-03
## 26 -1.776371827 -1.835507218 0.17601682 -0.8483479432 1.348136e-01
## 27  1.247561875  1.257897088 0.08369103  0.3801575295 2.843094e-02
## 28  0.174568468  0.172131513 0.08252251  0.0516236723 5.481995e-04
## 29 -0.733720368 -0.728789274 0.07953229 -0.2142246442 9.303065e-03
## 30  0.730635239  0.725677315 0.10296404  0.2458563427 1.225482e-02
## 31 -1.249953962 -1.260421576 0.08660188 -0.3881051666 2.962683e-02
## 32 -0.108416507 -0.106874423 0.07103079 -0.0295526507 1.797489e-04
## 33 -1.043192193 -1.044548682 0.07718264 -0.3020860161 1.820382e-02
## 34  0.688281672  0.683015946 0.09531188  0.2216944949 9.981838e-03
## 35  1.002134178  1.002197155 0.08919865  0.3136320696 1.967054e-02
## 36  1.600626622  1.638711484 0.05806995  0.4068823945 3.158951e-02
## 37  1.721521774  1.773496669 0.13203018  0.6916950670 9.016202e-02
## 38 -1.693003760 -1.741473039 0.17171092 -0.7929115605 1.188398e-01
## 39 -0.109289486 -0.107735278 0.07536270 -0.0307574597 1.947026e-04
## 40 -0.640939528 -0.635457160 0.12206680 -0.2369487002 1.142353e-02
```

(c).

```
abs_sr = abs(r1)
which(abs_sr > 2)
## 6 8 9
## 6 8 9
```

The 6th, 8th and 9th observations of Y are outlying observations.

(d).

```
p = 4
n = 40
which(leverage > 2 * (p + 1) / n)
## 3 8
## 3 8
```

We set criterion of leverage > 2(p+1) / n and find out that the 3rd and 8th observations of X are outlying observations.

(e).

```
abs_dffits = abs(dffits)
which(abs_dffits > 2 * sqrt((p + 1) / n))
##  6  8  9 26 38
```

```
##  6  8  9 26 38
```

Under the criterion of |DFFITS| > $2\sqrt{((p+1) / n)}$, we can say that the 6th, 8th, 9th, 26th and 38th observations may be influential.

```
which(cooks > qf(0.5, p, n - p))
## 8
## 8
```

While under the criterion of Cook's Distance > F(0.5, p, n - p), the 8th observation may be influential.