Linear Models, Problem 4

alexaoh

1 oktober, 2021

```
data <- read.csv2("dcrown.csv")
data$RP <- data$PB/data$PT # Add the given ratio to the data.
head(data)
```

```
#>   DCrown   PB   PT   HT  A       RP
#> 1   5.19 0.92 0.62 2.70 16 1.483871
#> 2   7.03 1.00 0.72 2.54  7 1.388889
#> 3   3.51 0.52 0.36 2.09  7 1.444444
#> 4   5.25 0.72 0.54 2.88 16 1.333333
#> 5   5.33 0.93 0.67 2.90 16 1.388060
#> 6   5.46 0.90 0.65 2.72 16 1.384615
```

```
dim(data)
```

```
#> [1] 311   6
```

```
summary(data)
```

```
#>      DCrown             PB               PT               HT
#>  Min.   : 1.870   Min.   :0.280   Min.   :0.1800   Min.   :1.690
#>  1st Qu.: 3.905   1st Qu.:0.600   1st Qu.:0.4050   1st Qu.:2.325
#>  Median : 4.920   Median :0.760   Median :0.5400   Median :2.590
#>  Mean   : 4.991   Mean   :0.761   Mean   :0.5345   Mean   :2.546
#>  3rd Qu.: 5.930   3rd Qu.:0.905   3rd Qu.:0.6400   3rd Qu.:2.770
#>  Max.   :11.030   Max.   :1.440   Max.   :1.0000   Max.   :3.470
#>        A              RP
#>  Min.   : 7.0   Min.   :1.103
#>  1st Qu.: 7.0   1st Qu.:1.367
#>  Median :16.0   Median :1.418
#>  Mean   :12.1   Mean   :1.433
#>  3rd Qu.:16.0   3rd Qu.:1.487
#>  Max.   :22.0   Max.   :1.750
```

Consider Two Different Linear Models

```
modA <- lm(DCrown~PT+RP+HT+A, data = data)
summary(modA)
```

```
#>
#> Call:
#> lm(formula = DCrown ~ PT + RP + HT + A, data = data)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -2.4770 -0.4398 -0.0358  0.3922  3.4693
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -1.84147    0.73373  -2.510  0.01260 *
#> PT           8.45921    0.34915  24.228  < 2e-16 ***
#> RP           0.77674    0.44621   1.741  0.08273 .
#> HT           0.33440    0.19159   1.745  0.08192 .
#> A            0.02863    0.01026   2.792  0.00558 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.7213 on 306 degrees of freedom
#> Multiple R-squared:  0.8124, Adjusted R-squared:   0.81
#> F-statistic: 331.4 on 4 and 306 DF,  p-value: < 2.2e-16
```

```
modB <- lm(log(DCrown)~log(PT)+log(RP)+log(HT)+log(A), data = data)
summary(modB)
```

```
#>
#> Call:
#> lm(formula = log(DCrown) ~ log(PT) + log(RP) + log(HT) + log(A),
#>     data = data)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -0.35655 -0.09336 -0.00394  0.09086  0.45920
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  1.69533    0.09994  16.964   <2e-16 ***
#> log(PT)      0.88111    0.03435  25.653   <2e-16 ***
#> log(RP)      0.28075    0.12723   2.207   0.0281 *
#> log(HT)      0.23307    0.09806   2.377   0.0181 *
#> log(A)       0.05636    0.02291   2.461   0.0144 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.1422 on 306 degrees of freedom
#> Multiple R-squared:  0.8345, Adjusted R-squared:  0.8323
#> F-statistic: 385.7 on 4 and 306 DF,  p-value: < 2.2e-16
```

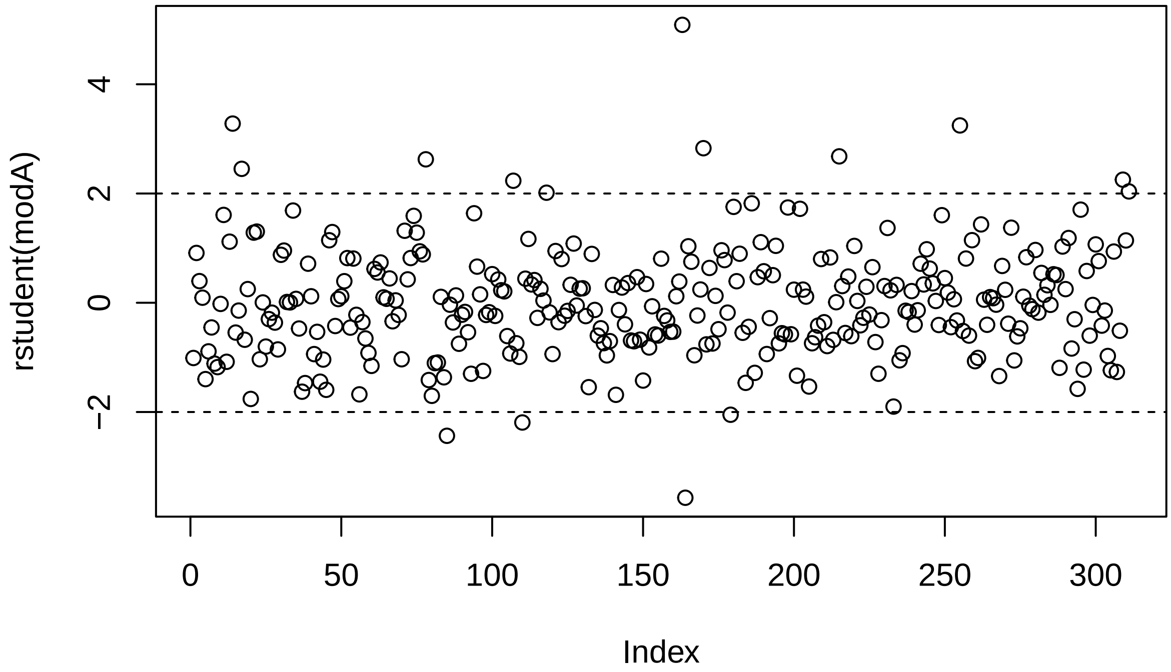
a) Questions about `modA`

(1) All the coefficient estimates are not significantly different from zero, with a significance level of 5%. Only `PT` and `A` (plus the intercept) are significant to this level.

(2) The estimation of the residual variance is (Residual standard error)^2 = 0.5202814.

(3) The studentized residuals are plotted below.

```
plot(rstudent(modA))
abline(h = c(-2, 2), lty = 2)
```



As is apparent, 15 points are outside the interval (-2,2). This represents a percentage of approximately 4.82 % of the points, which is reasonable, considering 1.96 ≈ 2, where 1.96 is the 0.975 quantile of the standard normal distribution. Hence, (-2,2) is an approximation of a 5% confidence interval.

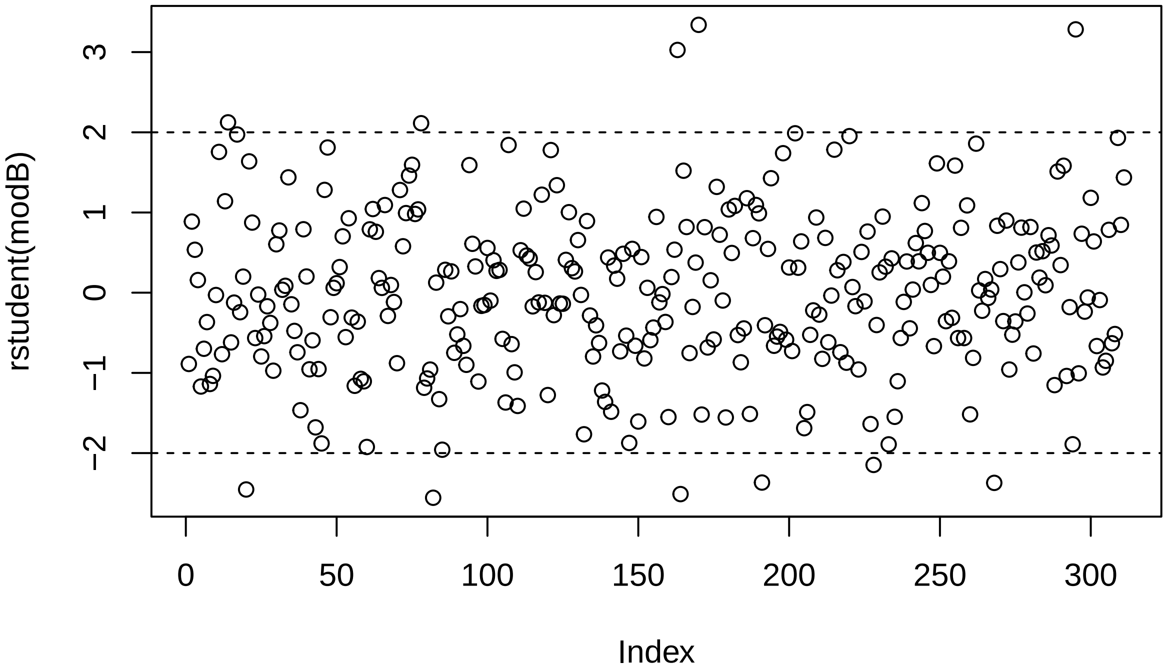
b) Questions about `modB`

(1) In the second model all the coefficient estimates are significantly different from zero, with a significance level of 5%.

(2) The estimation of the residual variance is $(\text{Residual standard error})^2 = 0.0202191$.

(3) The studentized residuals are plotted below.

```
plot(rstudent(modB))
abline(h = c(-2, 2), lty = 2)
```



As is apparent, 11 points are outside the interval (-2,2). This represents a percentage of approximately 3.54% of the points, which is reasonable, considering 1.96 ≈ 2, where 1.96 is the 0.975 quantile of the standard normal distribution. Hence, (-2,2) is an approximation of a 5% confidence interval.

c)

I would choose `modB` since, compared to `modB`, R^2 is larger, the F-statistic is larger (even though both have small p-values for the F-test) and all the coefficients are significant (in contrast to the first model, which only has a few estimates that are significantly non-zero, to a 5% level). Note that I have assumed that the assumptions of the linear model are verified for both models (as it says in the problem), which is why these assumptions are not discussed.

d)

```
new.data <- data.frame(PT = 0.4, PB = 0.6, HT = 2.3, A = 10, RP = 0.6/0.4)
predict(modB, new.data, interval = "confidence", level = 0.95)
```

```
#>        fit      lwr      upr
#> 1 1.325713 1.302734 1.348691
```