Recommended Exercise 6 in Statistical Linear Models, Spring 2021

alexaoh

14 mars, 2021

Problem 1 Orthogonally projecting matrices

Show that $R^T(I-R)=O$ iff R is symmetric and idempotent.

Assume that R is symmetric and idempotent. Then, $R^T(I-R)=R^T-R^TR=R-R^2=R-R=O$.

Conversely, assume that $R^T(I-R)=O$ holds. Then, $R^T=R^TR$. Hence, R is symmetric since $R=(R^TR)^T=R^TR=R^T$. Also, R is idempotent, since $R=R^TR=RR=R^2$, from the assumption. $\square$

Problem 2 Period of swing of pendulum

```
pendulum.data <- read.table("https://www.math.ntnu.no/emner/TMA4267/2018v/pendulum.txt")
model1 <- lm(Period~Length+Amplitude+Mass, data = pendulum.data)
summary(model1)
```

```
#>
#> Call:
#> lm(formula = Period ~ Length + Amplitude + Mass, data = pendulum.data)
#>
#> Residuals:
#>       Min        1Q    Median        3Q       Max
#> -0.109411 -0.023820  0.001007  0.027937  0.063272
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 0.4391125  0.0138346  31.740  < 2e-16 ***
#> Length      0.0197488  0.0002723  72.526  < 2e-16 ***
#> Amplitude   0.0448392  0.0296440   1.513  0.13367
#> Mass        0.0232896  0.0070989   3.281  0.00144 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.03644 on 96 degrees of freedom
#> Multiple R-squared:  0.9828, Adjusted R-squared:  0.9823
#> F-statistic:  1827 on 3 and 96 DF,  p-value: < 2.2e-16
```

a)

The fitted regression model has the form

$$\hat{\boldsymbol{Y}}=\hat{\beta}_0+\hat{\beta}_{\text{Length}}x_{\text{Length}}+\hat{\beta}_{\text{Amp}}x_{\text{Amp}}+\hat{\beta}_{\text{Mass}}x_{\text{Mass}},$$

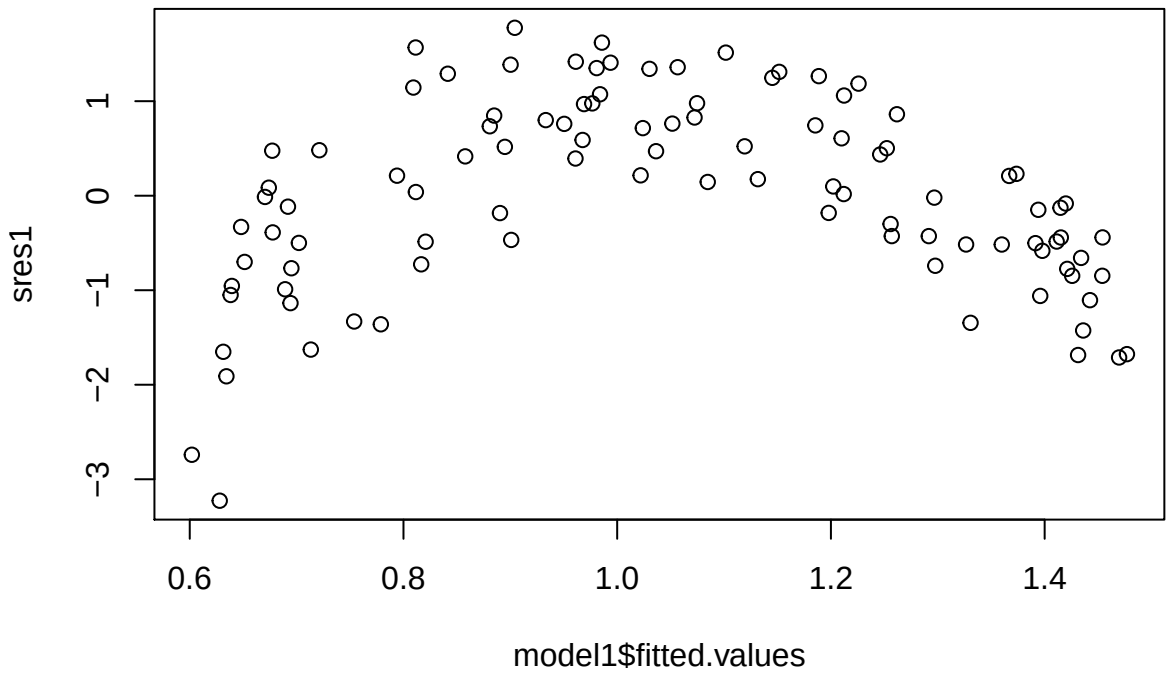
where the x-values are observed values of each of the predictors, and $\hat{\boldsymbol{Y}}$ is a prediction of the response, i.e. the period, based on the model estimates and the values of the predictors. When adding the values from the model, the regression fit takes the form

$$\hat{T} = 0.439 + 0.0197 \cdot L + 0.0448 \cdot A + 0.0232 \cdot m,$$

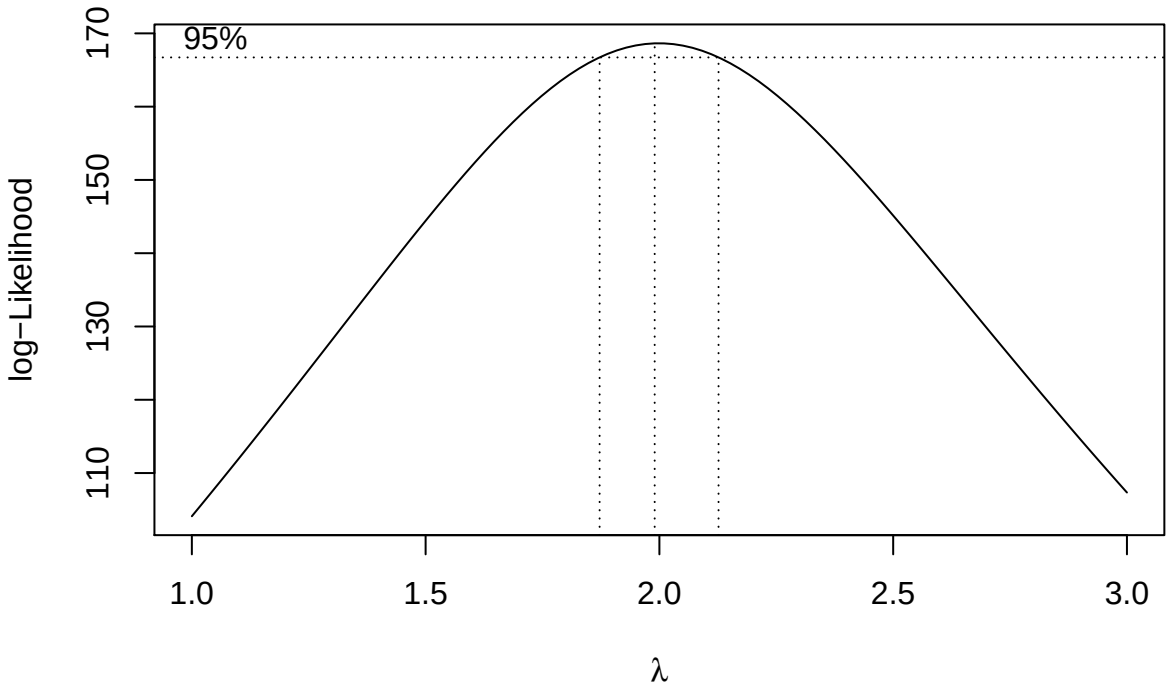
where the response is given the name T and each of the predictors are L, A and m, for length, amplitude and mass respectively. According to the R^2, the model explains 0.983 % of the variance. The p-value of the F-statistic shows that the null-hypothesis in the multiple hypothesis test, i.e. that all coefficients are zero, will be discarded, since it is significant. Moreover, the length, mass and interceipt have significant p-values, which indicates that the null-hypothesis in each of their simple hypothesis tests should be discarded as well.

From the residual plot, one can conclude that the errors are not homoscedastic. They follow a clear shape, which looks to be an inverted U, perhaps a $-x^2$. This suggests that the linear model is wrong. The Box-Cox plot below suggests that a transformation of the response T of the kind T^2 would be smart.

```
sres1 <- rstudent(model1)
plot(model1$fitted.values,sres1)
```




```
library(MASS)
boxcox(model1,lambda=seq(1,3,.1))
```

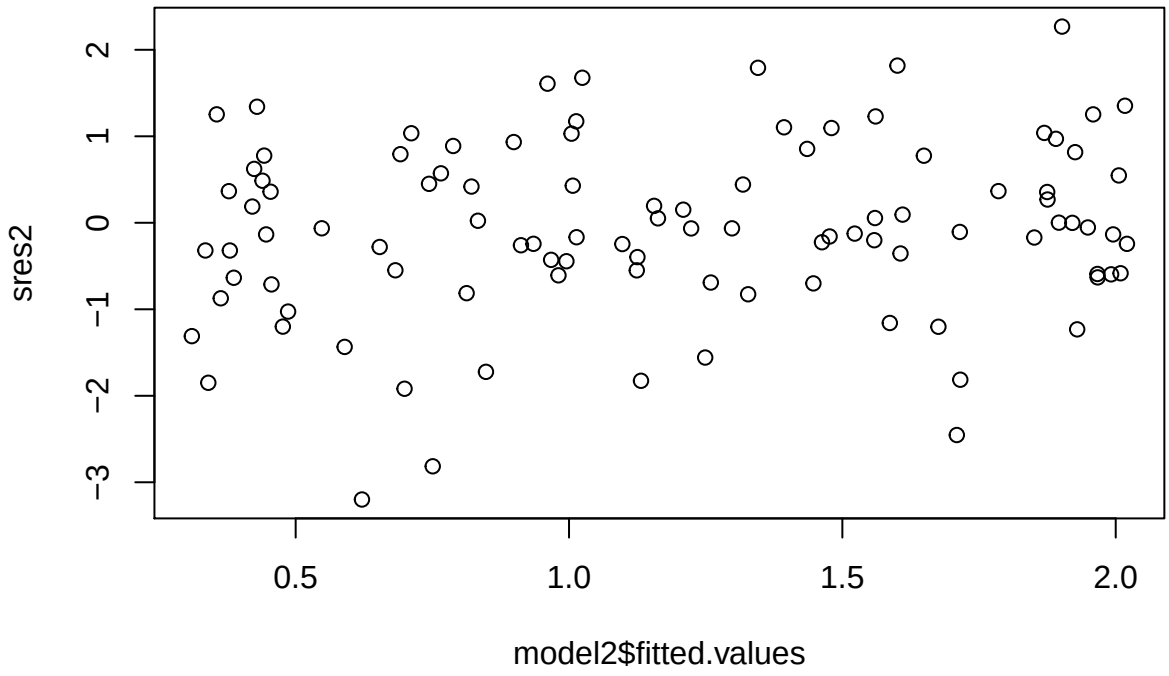


b)

```
model2 <- lm(Period^2~Length+Amplitude+Mass-1, data = pendulum.data)
summary(model2)
```

```
#>
#> Call:
#> lm(formula = Period^2 ~ Length + Amplitude + Mass - 1, data = pendulum.data)
#>
#> Residuals:
#>       Min        1Q    Median        3Q       Max
#> -0.121375 -0.023555 -0.003389  0.023144  0.086937
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> Length     0.0403534  0.0002672 151.008   <2e-16 ***
#> Amplitude  0.0610402  0.0262051   2.329   0.0219 *
#> Mass      -0.0045451  0.0066159  -0.687   0.4937
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.03976 on 97 degrees of freedom
#> Multiple R-squared:  0.9991, Adjusted R-squared:  0.9991
#> F-statistic: 3.566e+04 on 3 and 97 DF,  p-value: < 2.2e-16
```

```
sres2 <- rstudent(model2)
plot(model2$fitted.values,sres2)
```



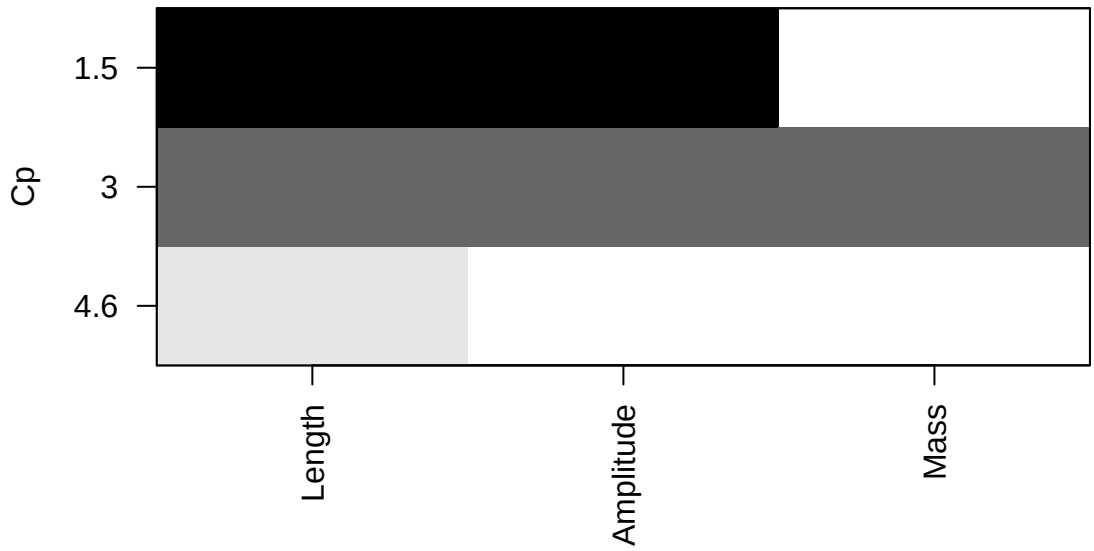

```
attach(pendulum.data)
pendulum <- as.data.frame(cbind(Period,Length,Amplitude,Mass))
library(leaps)
best <- regsubsets(Period^2~.,data=pendulum,intercept=FALSE)
summary(best)$which
```

```
#>   Length Amplitude  Mass
#> 1   TRUE     FALSE FALSE
#> 2   TRUE      TRUE FALSE
#> 3   TRUE      TRUE  TRUE
```

```
summary(best)$cp
```

```
#> [1] 4.569336 1.471964 3.000000
```

```
plot(best,scale="Cp")
```



Based on the residual plot alone, I would much prefer this model over the former `model1`. This has no clear pattern in the residuals, which does not indicate any grave mistakes in the model assumptions. Considering the fact that Mallows' C_p should be as small as possible, the best of the submodels includes the length and the amplitude, since this gives the lowest value for C_p according to the given data.

c)

```
model3 <- lm(log(Period)~log(Length)+log(1+Amplitude^2/16+11*Amplitude^4/3072), data = pendulum.data)
summary(model3)
```

```
#>
#> Call:
#> lm(formula = log(Period) ~ log(Length) + log(1 + Amplitude^2/16 +
#>     11 * Amplitude^4/3072), data = pendulum.data)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -0.09906 -0.01002  0.00126  0.01266  0.08019
#>
#> Coefficients:
#>                                                 Estimate Std. Error  t value
#> (Intercept)                                    -1.617849   0.015979 -101.247
#> log(Length)                                     0.502433   0.004809  104.474
#> log(1 + Amplitude^2/16 + 11 * Amplitude^4/3072)  1.260754   0.570785    2.209
#>                                                Pr(>|t|)
#> (Intercept)                                      <2e-16 ***
```

```
#> log(Length)                                          <2e-16 ***
#> log(1 + Amplitude^2/16 + 11 * Amplitude^4/3072)   0.0295 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.02705 on 97 degrees of freedom
#> Multiple R-squared:  0.9912, Adjusted R-squared:  0.9911
#> F-statistic:  5491 on 2 and 97 DF,  p-value: < 2.2e-16
```

The coefficient for log(Length) agrees with the theory, which states that it should be $\frac{1}{2}$. Moreover, the coefficient for $log($1 + \dots$) should be 1 according to theory, and is estimated to being ≈ 1.261, which is an error of ≈ 20.7 %. It is worth to note that the standard error of this estimate is large. An estimate of g is 1003.703 cm/s, which is ≈ 10.037 m/s.

To calculate a 95% confidence interval we use the fact that

\textcolor{red}{Gjør denne skikkelig senere! Hvordan kan de si det de sier om fordelingen til beta? Hva skjer med c_i i nevneren, dvs diagonal i ($X^{TX)}${-1}? (se notater fra forelesning). Se mer på denne}

Problem 3 Galápagos species

a)