Recommended Exercise 6 in Statistical Linear Models, Spring 2021

alexaoh

02 april, 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. □

Problem 2 Period of swing of pendulum

```r
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

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

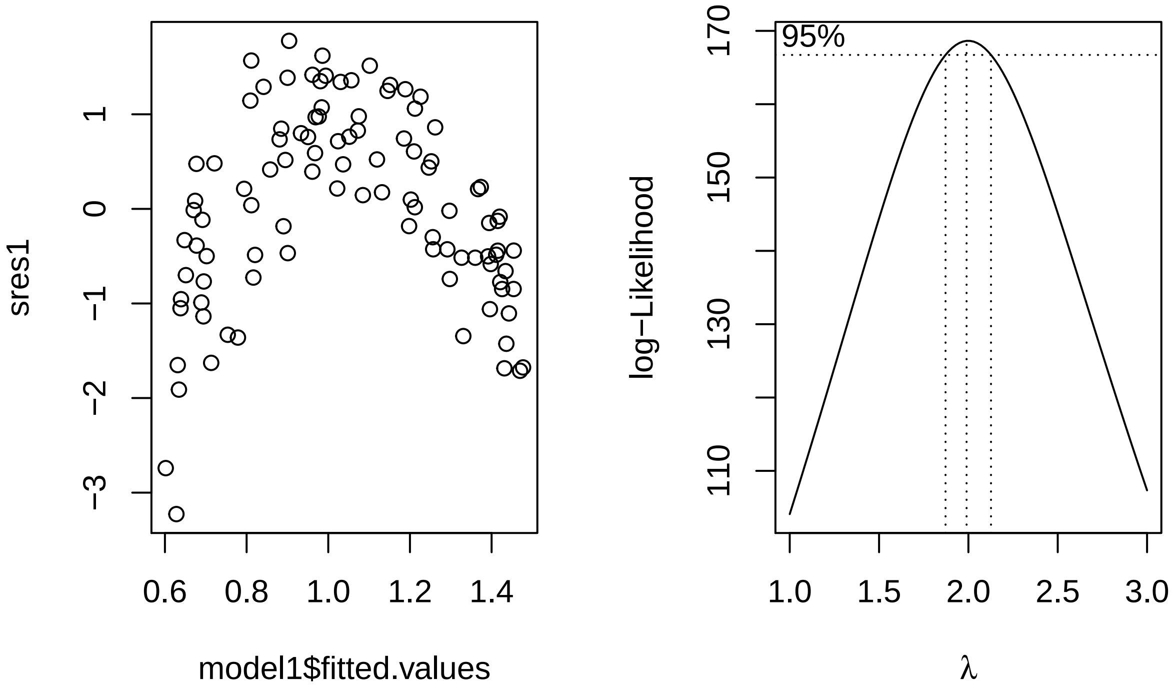
where the x-values are observed values of each of the predictors, and $\hat{\boldsymbol{Y}}$ is a prediction of the response, i.e. the period, based on the model estimates and the values of the predictors. When adding the values from the model, the regression fit takes the form

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

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 intercept 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, since $\lambda = 2$ maximizes the log-likelihood function of the Box-Cox transformed data.

```r
par(mfrow=c(1,2))
sres1 <- rstudent(model1)
plot(model1$fitted.values,sres1)
library(MASS)
boxcox(model1,lambda=seq(1,3,.1))
```



b)

```r
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
```

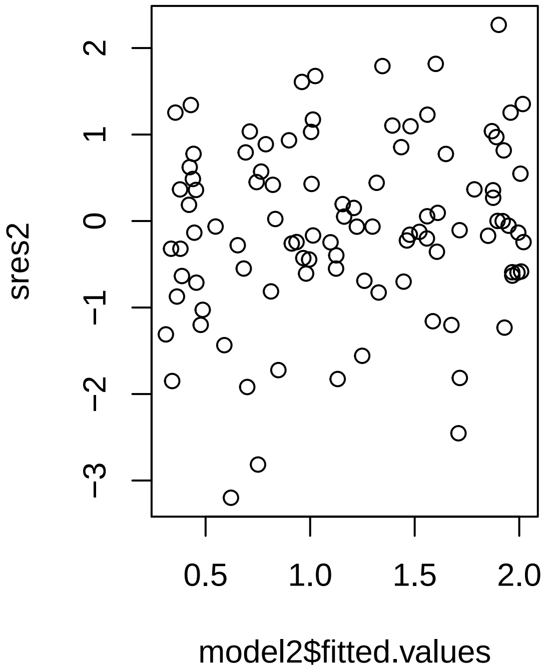
```r
par(mfrow=c(1,2))
sres2 <- rstudent(model2)
plot(model2$fitted.values,sres2)
attach(pendulum.data)
pendulum <- as.data.frame(cbind(Period,Length,Amplitude,Mass))
library(leaps)
best <- regsubsets(Period^2~.,data=pendulum,intercept=FALSE)
summary(best)$which
```

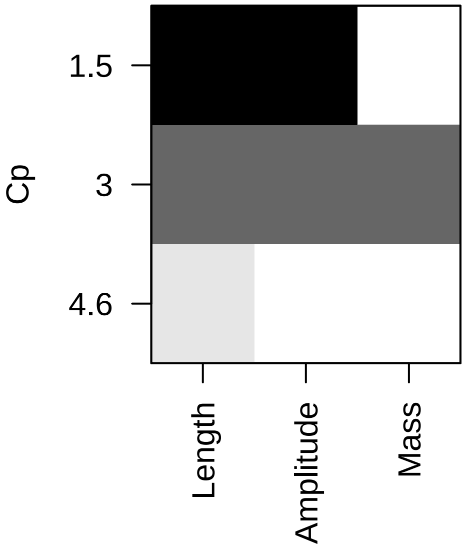
```
#>   Length Amplitude  Mass
#> 1   TRUE     FALSE FALSE
#> 2   TRUE      TRUE FALSE
#> 3   TRUE      TRUE  TRUE
```

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

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

```r
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 $\log(1 + \ldots)$ 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 $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$. For the intercept, i.e. $\hat{\beta}_0$, this means that $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(X^TX)^{-1}_{00})$. We know that $\hat{\sigma}^2 = \frac{\text{SSE}}{n-p}$ is an unbiased estimator of σ^2. Moreover, it can be proved that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. From this, it can be shown that the test statistic

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{00}}} \sim t_{n-p} = t_{97}$$

holds. This leads to the 95% confidence interval $\beta_1 \in \left[\hat{\beta}_1 - t_{0.025}\hat{\sigma}\sqrt{(X^TX)^{-1}_{11}}, \hat{\beta}_1 + t_{0.025}\hat{\sigma}\sqrt{(X^TX)^{-1}_{11}}\right]$. This confidence interval can be found manually or via a command in R. Both are done below.

```r
# Manually.
x <- model.matrix(model3)
n <- 97
p <- 3
s2<-summary(model3)$sigma
c<-diag(solve(t(x)%*%x))
t <- qt(0.975, 97)
coefficients(model3)[1] - t*summary(model3)$coefficients[,"Std. Error"][1] # t*s2*sqrt(c[1])
```

```
#> (Intercept)
#>   -1.649563
```

```r
coefficients(model3)[1] + t*summary(model3)$coefficients[,"Std. Error"][1] # t*s2*sqrt(c[1])
```

```
#> (Intercept)
#>   -1.586134
```

```r
# Command in R.
confint(model3)
```

```
#>                                                      2.5 %     97.5 %
#> (Intercept)                                     -1.6495630 -1.5861342
#> log(Length)                                      0.4928882  0.5119779
#> log(1 + Amplitude^2/16 + 11 * Amplitude^4/3072)  0.1279035  2.3936049
```

This needs to be converted to the correct confidence interval for g, via the same procedure as earlier. This leads to the final confidence interval given by lower bound 9.420163 m/s and upper bound 10.6942901 m/s.

Problem 3 Galápagos species

a)

The fitted regression model has the form

$$\hat{\boldsymbol{Y}} = \hat{\beta}_0 + \hat{\beta}_{\text{Ar}} x_{\text{Ar}} + \hat{\beta}_{\text{El}} x_{\text{El}} + \hat{\beta}_{\text{Near}} x_{\text{near}} + \hat{\beta}_{\text{Scruz}} x_{\text{Scruz}} + \hat{\beta}_{\text{Adj}} x_{\text{Adj}},$$

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 species, based on the model estimates and the values of the predictors. The values for each of the parameters from the summary output can be exchanged with the parameters above for the fitted (estimated) regression model. The p-value of the F-test on the entire model is relatively small, i.e. significant, and the null hypothesis (that all parameters are zero) can be discarded. The model therefore has some merit. The model explains $\approx 77\%$ of the variability in the data. Moreover, t-tests claim that `Elevation` and `Adjacent` are (at least somewhat) significant.

From the plots, one can draw some conclusions. Firstly, the residuals do not look homoscedastic, since they have a higher density for lower fitted values. Furthermore, the quantiles do not seem to match the theoretical quantiles in the Q-Q-plot. Also, the Anderson-Darling normality test suggests that the errors are not normal, since the p-value is low, which means that the null-hypothesis that the errors are normal can be rejected. Hence, some of the assumptions of the linear model seem to fail. The Box-Cox-plot suggests that a transform $\text{Species}^{\frac{1}{3}}$, of the response, could be clever.

b)

The cube root transformation of the response is used in a new fitted model, using the same covariates as in the first model. The four missing numerical values from the new fit are

- Estimate for the intercept. This can be calculated from the t-value and the standard error: $7.365 \cdot 0.3052013 \approx 2.2478$. This is the first element of the vector of regression coefficient estimates $\hat{\boldsymbol{\beta}}$.
- P-value for the t-test of `Area`. This can be approximately found using a table. Since the quantile for $\alpha = 0.025$ for a t-distribution with 24 degrees of freedom is 2.064, which is the closest quantile to the given t-value from the summary of the model, the p-value is approximately 0.05. This is a test of the null hypothesis $H_0 : \beta_{\text{Areav}} = 0$ vs the alternative hypothesis $H_a : \beta_{\text{Areav}} \neq 0$, when all the other predictors are kept in the model.
- Std. error of `Nearest`. This can be found from the t-value, used in the t-test of the covariate: $0.0118152/0.703 \approx 0.0168$. This is the estimated standard deviation of the coefficient estimate of `Nearest.`
- Adjusted R-squared. This can be found by inserting into the formula that converts regular R-squared to the adjusted R-squared. This can be used for model selection or assessment when comparing models with different amounts of covariates.

I would prefer this model, with the cube root transformation, since the assumptions of a linear model seem to hold more closely in this case, based on the plots and the Anderson-Darling test, compared to the first model. The Anderson-Darling test does not reject the null-hypothesis of normal data. Also, the residual plot has no clear structure and the Q-Q-plot looks better for the new model compared to the old one.

c)

Results from performing best subset selection are shown.

The definitions of R^2 and R^2_{adj} are $R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\boldsymbol{Y}^T(I-H)\boldsymbol{Y}}{\boldsymbol{Y}^T(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Y}}$ and $R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} = 1 - \frac{(\boldsymbol{Y}^T(I-H)\boldsymbol{Y})/(n-p)}{(\boldsymbol{Y}^T(I-\frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Y})/(n-1)}$, where it is assumed that there are $p = k + 1$ covariates (including intercept). When comparing models with the same amount of covariates, R^2 is sufficient for model selection. Simply select the model with the highest R^2 for highest predictive power. However, when comparing models with different amounts of covariates, the use of R^2_{adj} is necessary, since the R^2 always increases with increasing amounts of covariates in the model. Therefore, a penalization for larger models with "noisy"-predictors is introduced in R^2_{adj}, in order to do proper model selection.

Based on R^2_{adj}, the best model is the model with 4 covariates, i.e. the model with the largest value. This model uses the covariates `Area`, `Elevation`, `Scruz` and `Adjacent` (+ intercept).

The Lasso performs regularization or variable selection, since it penalizes larger models in the ℓ^1-norm, i.e. the term $\lambda \sum_{j=1}^{p} |\beta_j|$ is added to the usual term $(\boldsymbol{Y} - X\boldsymbol{\beta})^(\boldsymbol{Y} - X\boldsymbol{\beta})T$, when optimizing, where $\lambda \geq 0$ is a penalty parameter. When $\lambda = 0$ estimates become the regular least squares estimates, but when $\lambda \to \infty$ all predictors are set to zero. The R^2_{adj} can be used to select the penalty parameter in the Lasso, because it does not depend on λ. The number of covariates that should enter into the R^2_{adj}-formula is not defined, since the Lasso shrinks the coefficients by different amounts, and there is no way to insert λ into the formula for the adjuster R^2.

The fitted regression model from the Lasso, found by CV, is

$$\widehat{\text{Species}}^{1/3} = 3.5388701794 + 0.0002804519 \cdot \text{Elevation}.$$