TMA4267 Linear Statistical Models
Part 2: Linear regresssion
Recommended exercise 4 - V2017

February 7, 2017

Keywords: multiple linear regression, fitted model, Box-Cox plot, Mallows Cp, model selection, residual plot, R^2, R^2_{adj}, prediction interval.

May 16, 2017: corrected missing transpose in Problem 3.

- Problem 1 and 2 both focuses on different aspects of analysing data with the multiple linear regression model.
- Problem 3 looks at confidence and prediction intervals.

Problem 1: Pendulum

(TMA4267 V2015 Problem 1)

The period of swing of a pendulum was studied as 100 combinations of the pendulum's length (measured in cm), amplitude (the maximum angle that the pendulum swings away from vertical, measured in radians) and mass (kg) were varied. A multiple regression model was fitted. R input and output, a residual plot and a Box–Cox plot are shown in Figure 1.

a) Write down the fitted regression model, and comment briefly on the model fit. What conclusions can you draw from the residual plot? Suggest a transformation based on the Box–Cox plot.

The approximate formula $T \approx 2\pi\sqrt{L/g}$ for the period T of a pendulum, where L is the length and $g \approx 9.8$ m/s^2 is the gravitational acceleration, suggests the use of the square of the period rather than the period as the response in a regression model, and also to drop the intercept. R input and output, a residual plot and a plot of best subset selection based on Mallows' C_P for such models are shown in Figure 2.

b) Would you prefer the original model or the new model just described? Considering submodels of the new model, which would you choose? Briefly justify your answers.

```
> model1<-lm(Period~Length+Amplitude+Mass)
> summary(model1)

Call:
lm(formula = Period ~ Length + Amplitude + Mass)

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

> sres1<-rstudent(model1)
> plot(model1$fitted.values,sres1)
> library(MASS)
> boxcox(model1,lambda=seq(1,3,.1))
```

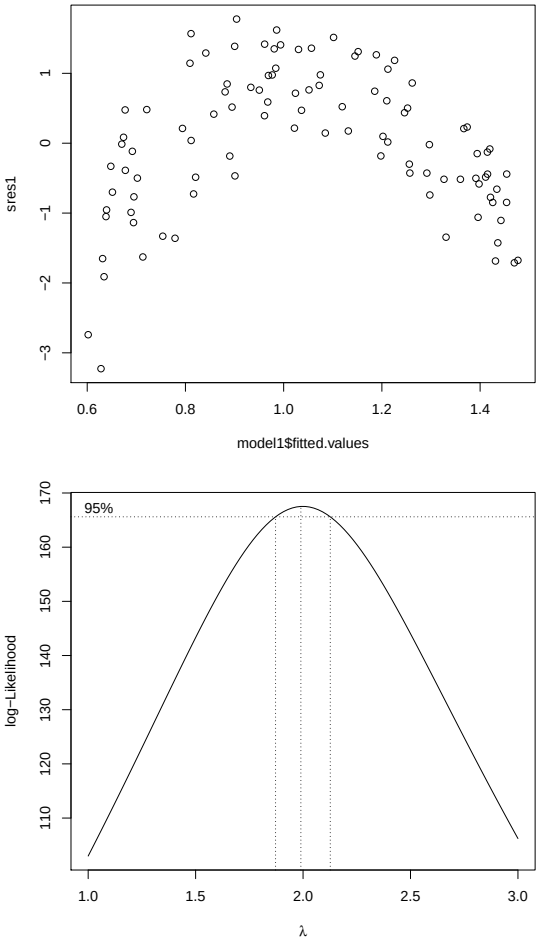

Figure 1: Model from Problem 1a: R input and output (left), residual plot (upper right) and Box–Cox plot (lower right).

```
> model2<-lm(Period^2~Length+Amplitude+Mass-1)
> summary(model2)

Call:
lm(formula = Period^2 ~ Length + Amplitude + Mass - 1)

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$fitted.values,sres2)
> pendulum<-as.data.frame(cbind(Period,Length,Amplitude,Mass))
> library(leaps)
> best<-regsubsets(Period^2~.,data=pendulum,intercept=FALSE)
> summary(best)$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")
```

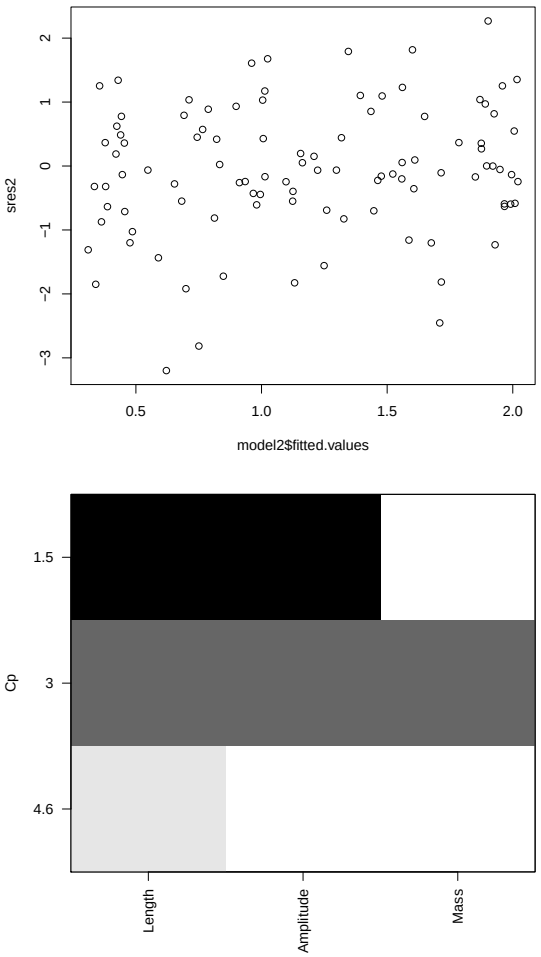

Figure 2: Model from Problem 1b: R input and output (left), residual plot (upper right) and a graphical table of best subsets using Mallows' C_P as the statistic for ordering models (lower right). Note that the information of the graphical table is also included in the R output.

```
> model3<-lm(log(Period)~log(Length)+log(1+Amplitude^2/16+11*Amplitude^4/3072))
> summary(model3)

Call:
lm(formula = log(Period) ~ log(Length) + log(1 + Amplitude^2/16 +
    11 * Amplitude^4/3072))

Residuals:
     Min       1Q   Median       3Q      Max
-0.09906 -0.01002  0.00126  0.01266  0.08019

Coefficients:
                                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                -1.617849   0.015979 -101.247   <2e-16 ***
log(Length)                                 0.502433   0.004809  104.474   <2e-16 ***
log(1 + Amplitude^2/16 + 11 * Amplitude^4/3072)  1.260754   0.570785    2.209   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
```

Figure 3: Model from Problem 1c: R input and output.

Finally, a more exact formula, $T = 2\pi\sqrt{\frac{L}{g}}(1 + \frac{1}{16}\theta^2 + \frac{11}{3072}\theta^4 + \cdots)$, or $\ln T = \ln(2\pi/\sqrt{g}) + \frac{1}{2}\ln L + \ln(1 + \frac{1}{16}\theta^2 + \frac{11}{3072}\theta^4 + \cdots)$, where θ is amplitude, suggests a third model in which both the response variable and the covariates are transformed. R input and output are shown in Figure 3.

c) How do the estimates of the coefficients agree with the physical model given above? Find an estimate of g, the gravitational acceleration, and a 95% confidence interval for g.

Problem 2: Galapagos species

(TMA4267 V2014 Problem 2)

This data set concerns the number of species of tortoise on the various Galapagos Islands, and is taken from the book "Practical Regression and Anova using R" by Julian J. Faraway.

The data set contains measurements on 30 islands, and we study the following 6 variables:

- `Species`: The number of species of tortoise found on the island.
- `Area`: The area of the island (km^2).
- `Elevation`: The highest elevation of the island (m).
- `Nearest`: The distance from the nearest island (km).
- `Scruz`: The distance from Santa Cruz island (km).
- `Adjacent`: The area of the adjacent island (km^2).

Summary statistics are given below for the Galapagos data set.

```
Summary statistics for the Galapagos data set
        Species      Area Elevation Nearest  Scruz Adjacent
Min.       2.00    0.0100     25.00    0.20   0.00     0.03
1st Qu.   13.00    0.2575     97.75    0.80  11.02     0.52
Median    42.00    2.5900    192.00    3.05  46.65     2.59
Mean      85.23  261.7000    368.00   10.06  56.98   261.10
3rd Qu.   96.00   59.2400    435.20   10.02  81.08    59.24
Max.     444.00 4669.0000   1707.00   47.40 290.20  4669.00
```

A multiple linear regression model was fitted to the Galapagos data set, with `Species` as response and the remaining five variables as covariates. Call this Model A. Code and printout from R is found in Figure 4 and accompanying plots in Figures 5 and 6.

a) Write down the fitted regression model, and comment *briefly* on the model fit. What conclusions can you draw from the residual plots and the Box–Cox transformation plot?

The cube root transformation of `Species` will from now on be used as response in a new multiple linear regression model, with the same five covariates as for Model A. Call this Model B. Code and printout from R is found in Figure 7 and accompanying plots in Figure 8.

b) In the printout in Figure 7 from fitting Model B four numerical values are substituted by question marks. Calculate numerical values for each of these, and explain what each of the numbers means.

Would you prefer Model B to Model A? Justify *briefly* your answer.

c) The results from performing best subset selection is reported in Figure 9, where also R^2 and R^2_{adj} is listed numerically for the five models reported.

Write down the definition for R^2 and R^2_{adj} and explain how you can use these to compare the different models.

Choose the "best" out of these five models. Justify your choice.

```
> fit1 <- lm(Species~Area+Elevation+Nearest+Scruz+Adjacent,
data=gala)
> summary(fit1)

Call:
lm(formula = Species ~ Area + Elevation +
    Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07

> plot(fit1$fitted,rstudent(fit1))
> qqnorm(rstudent(fit1))
> qqline(rstudent(fit1))
> ad.test(rstudent(fit1))
Anderson-Darling normality test
data:  rstudent(fit1)
A = 1.7071, p-value = 0.0001729
> boxcox(fit1)
> abline(v=1/3,lty=1)
```

Figure 4: Printout from statistical analyses for Model A for the Galapagos data set.

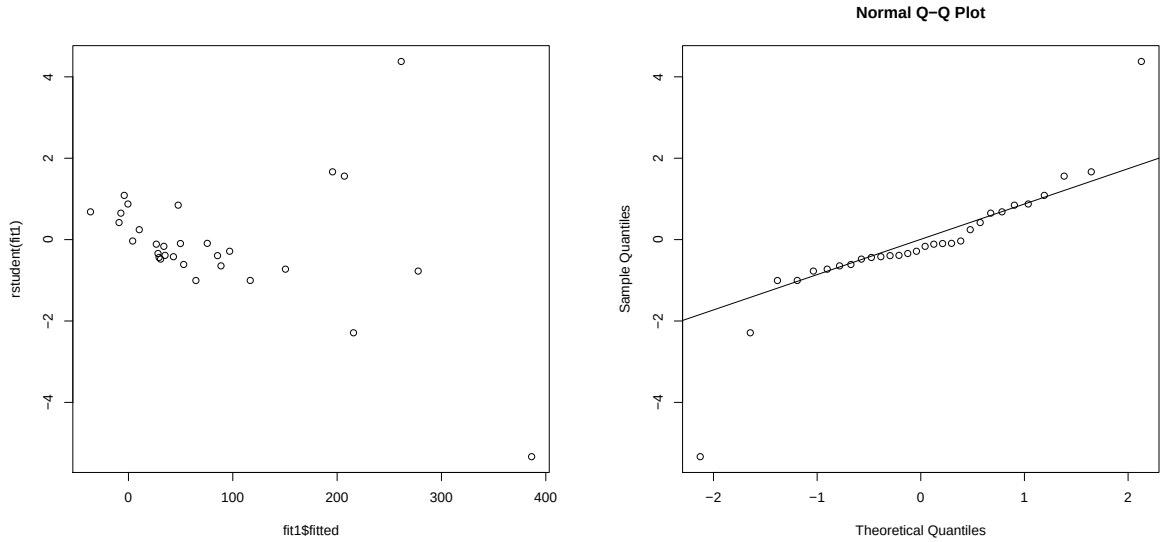

Figure 5: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for Model A for the Galapagos data set.

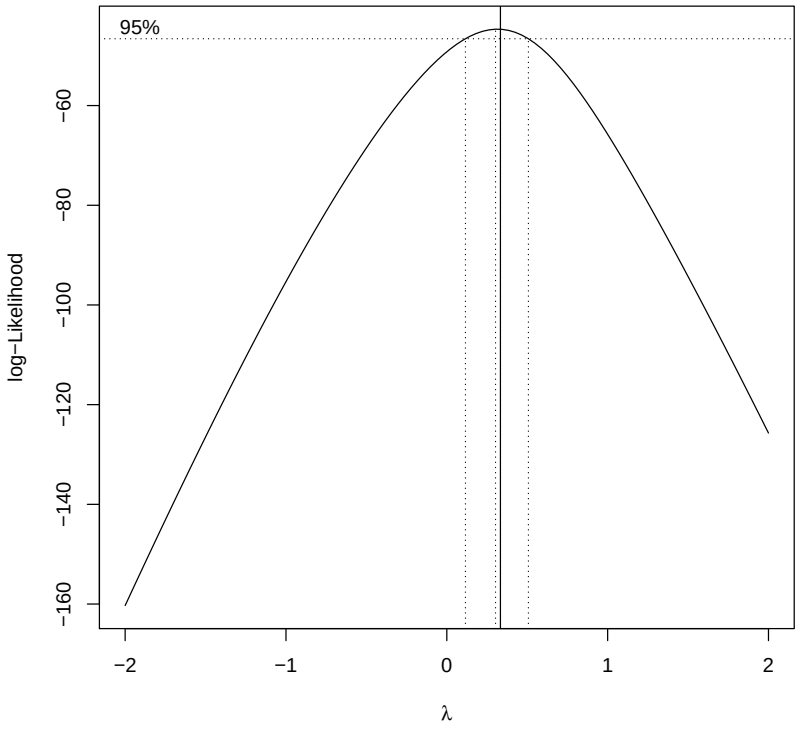

Figure 6: Box–Cox transformation plot based on Model A for the Galapagos data set.

```
> fit2 <- lm(Species^(1/3)~Area+Elevation+Nearest+Scruz+Adjacent,
x=TRUE,data=gala)
> summary(fit2)

Call:
lm.default(formula = Species^(1/3) ~ Area + Elevation + Nearest +
    Scruz + Adjacent, data = gala, x = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54306 -0.47863 -0.08499  0.56349  1.83283

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)         ?  0.3052013   7.365 1.32e-07
Area        -0.0007349  0.0003573  -2.057        ?
Elevation    0.0054510  0.0008551   6.375 1.37e-06
Nearest      0.0118152          ?   0.703  0.48855
Scruz       -0.0045951  0.0034322  -1.339  0.19317
Adjacent    -0.0010597  0.0002820  -3.757  0.00097
---
Residual standard error: 0.9716 on 24 degrees of freedom
Multiple R-squared:  0.7543,Adjusted R-squared:  ?
F-statistic: 14.74 on 5 and 24 DF,  p-value: 1.192e-06

> plot(fit2$fitted,rstudent(fit2))
> qqnorm(rstudent(fit2))
> qqline(rstudent(fit2))
> ad.test(rstudent(fit2))
Anderson-Darling normality test
data:  rstudent(fit2)
A = 0.2639, p-value = 0.6738
```

Figure 7: Printout from R of fitting the multiple linear regression model B for the Galapagos island data set. Response is cube root of `Species`.

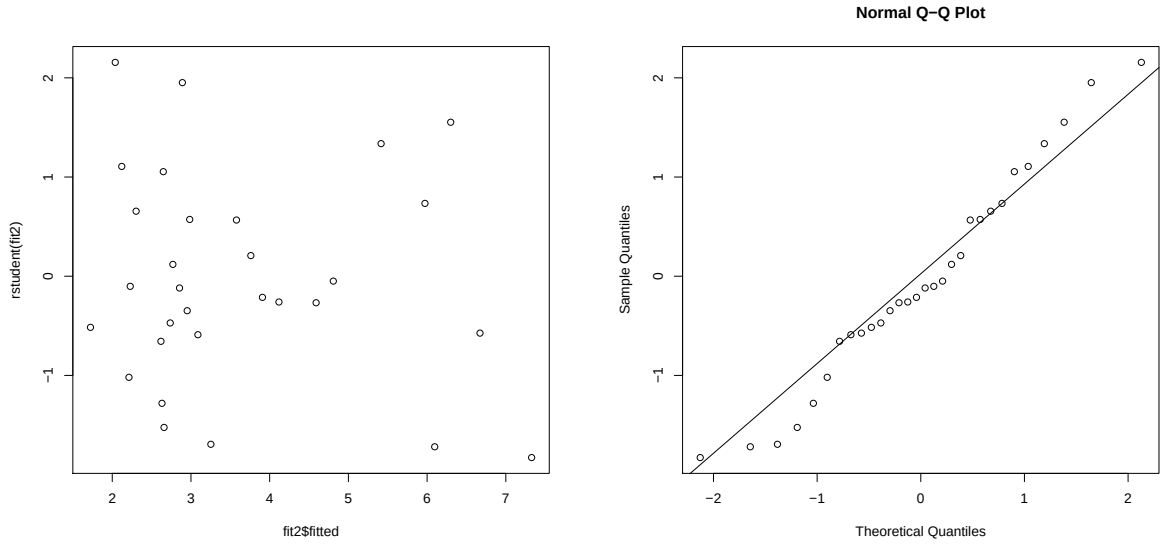

Figure 8: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for Model B (cube root of `Species`) for the Galapagos data set.

```
> x <- fit2$x[,-1]
> y <- gala$Species^(1/3)
> library(leaps)
> bests <- regsubsets(x,y)
> sumbests <- summary(bests)
> sumbests
Subset selection object
5 Variables  (and intercept)
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         Area Elevation Nearest Scruz Adjacent
1  ( 1 ) " "  "*"       " "     " "   " "
2  ( 1 ) " "  "*"       " "     " "   "*"
3  ( 1 ) "*"  "*"       " "     " "   "*"
4  ( 1 ) "*"  "*"       " "     "*"   "*"
5  ( 1 ) "*"  "*"       "*"     "*"   "*"
> plot(1:5, sumbests$rsq,type="l") #solid line
> lines(1:5, sumbests$adjr2,lty=2) #dashed line
> sumbests$rsq # R^2
[1] 0.5570784 0.6893784 0.7356845 0.7492704 0.7543353
> sumbests$adjr2 # R^2_adjusted
[1] 0.5412597 0.6663694 0.7051866 0.7091536 0.7031552
```

Figure 9: Printout from R of fitting best subset selection to the Galapagos island data set. Response is cube root of `Species`.

Problem 3: Inference about a new observation in multiple linear regression

Consider again the multiple linear regression model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$. Assume that $Y_0 = \boldsymbol{X}_0^T\boldsymbol{\beta} + \epsilon_0$ is a new observation, with ϵ_0 independent of $\boldsymbol{\epsilon}$.

a) Show that $\boldsymbol{X}_0^T\hat{\boldsymbol{\beta}}$ is an unbiased estimator of EY_0, where $\hat{\boldsymbol{\beta}}$ is the least-square estimator of β. Find the distribution of the estimator.

b) Find a $100(1-\alpha)\%$ confidence interval for EY_0.

c) Find a $100(1-\alpha)\%$ prediction interval for Y_0, that is, an interval that will contain Y_0 with probability $1-\alpha$.

d) Use the acid rain data (see course www page exercise tab) to find a confidence interval for the expected value of a new observation having covariates $(\texttt{x1}, \ldots, \texttt{x7}) = (1, 3, 50, 1, 50, 2, 1, 0)$. Also find a prediction interval for such a new observation.

e) From the theory of simple linear regression, you know that the bounds of the confidence interval are

$$\hat{y}_0 \pm t_{\alpha/2}\sqrt{\frac{\text{SSE}}{n-2}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2)}{\sum_i (x_i - \bar{x})^2}\right)},$$

where $\hat{y}_0$ is the estimator of EY_0, $-t_{\alpha/2}$ the $\alpha/2$-quantile of a $t(n-2)$ variable, n the number of observations, x_i the covariates and x_0 the new covariate. Show that this is the same confidence interval as found above.