NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4267 Linear Statistical Models**

Academic contact during examination: Øyvind Bakke

Phone: 73 59 81 26, 990 41 673

Examination date: 22 May 2015

Examination time (from–to): 9:00–13:00

Permitted examination support material: Yellow stamped A5 sheet with your own handwritten notes, specific basic calculator, *Tabeller og formler i statistikk* (Tapir forlag), *Matematisk formelsamling* (K. Rottmann)

Other information:
In the grading, each of the eight points counts equally.

Language: English

Number of pages: 4

Number pages enclosed: 0

Checked by:

Date Signature

```
> 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))
```

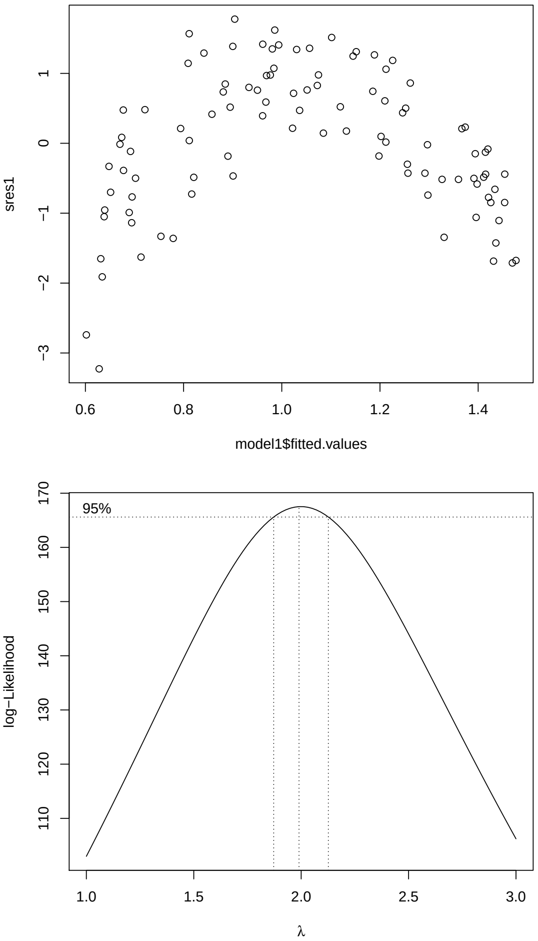

Figure 1: Model from Problem 1a: R input and output (left), residual plot (upper right) and Box–Cox plot (lower right).

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.

```
> 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")
```

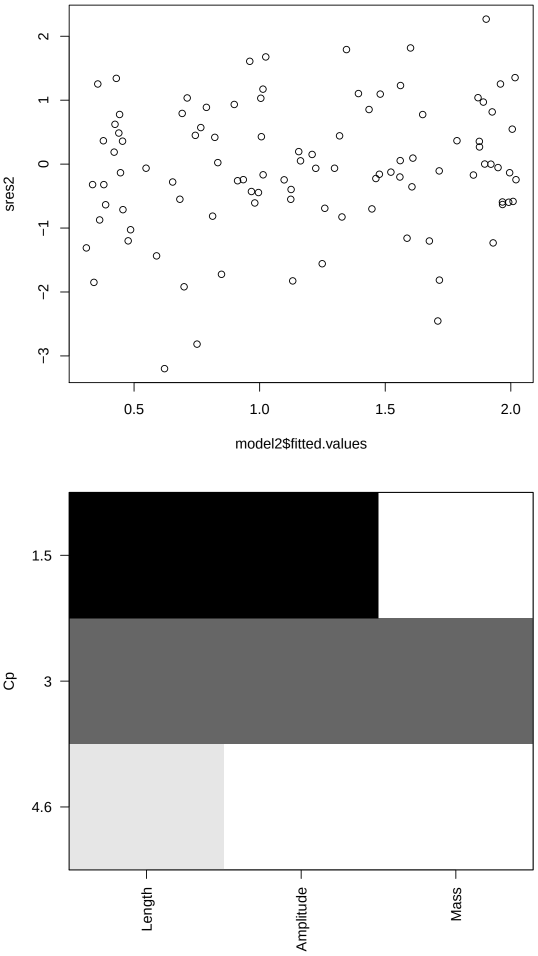

Figure 2: Model from Problem 1b: R input and output (left), residual plot (upper right) and a graphical table of best subsets using Mallows' C_P as the statistic for ordering models (lower right). Note that the information of the graphical table is also included in the R output.

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.

```
> 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

Assume a linear regression model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{Y}$ is an n-dimensional random vector, X an $n \times p$ design matrix, $\boldsymbol{\beta}$ a p-dimensional parameter (coefficient) vector and $\boldsymbol{\epsilon}$ n-dimensional multinormal with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$, where I is an $n \times n$ identity matrix.

Assume further that the columns of X are orthogonal.

a) Show that the least squares estimator of β_j, the jth entry of $\boldsymbol{\beta}$, depends only on the jth column of X (i.e., the jth covariate vector) and $\boldsymbol{Y}$.

Consider a two-level 2^2 factorial unreplicated experiment in which the levels are coded -1 and 1. The response vector is $(6 \;\; 4 \;\; 10 \;\; 7)^{\mathrm{T}}$, corresponding to levels $(-1 \;\; 1 \;\; -1 \;\; 1)^{\mathrm{T}}$ of the first factor and $(-1 \;\; -1 \;\; 1 \;\; 1)^{\mathrm{T}}$ of the second.

b) Estimate the interaction effect (twice the coefficient) of the two factors.

Problem 3

Assume a linear regression model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{Y}$ is an n-dimensional random vector, X an $n \times p$ design matrix, $\boldsymbol{\beta}$ a p-dimensional parameter vector and $\boldsymbol{\epsilon}$ n-dimensional multinormal with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$, where I is an $n \times n$ identity matrix.

We consider a reduced model involving only the r first covariates, with $r < p$. Let X_0 denote the design matrix consisting only of the first r columns of X. Let $\hat{\boldsymbol{\beta}}_{(0)} = (X_0^{\mathrm{T}} X_0)^{-1} X_0^{\mathrm{T}} \boldsymbol{Y}$ denote the least squares estimator of the parameters of the submodel, and let $\hat{\boldsymbol{\beta}}_0$ be $\hat{\boldsymbol{\beta}}_{(0)}$ extended by zeros so that $\hat{\boldsymbol{\beta}}_0$ has length p, that is $\hat{\boldsymbol{\beta}}_0^{\mathrm{T}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{(0)}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \end{pmatrix}$, where $\mathbf{0}$ is a zero vector of length $p - r$.

We would like to measure the adequacy of the submodel by

$$J_0 = \frac{1}{\sigma^2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^{\mathrm{T}} X^{\mathrm{T}} X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0),$$

called "the scaled sum of squared errors" by Mallows. A problem is of course that the parameters $\boldsymbol{\beta}$ (and σ^2) are unknown. We consider the original model to be "true", so that $E\boldsymbol{Y} = X\boldsymbol{\beta}$.

a) What is the covariance matrix of $\hat{\boldsymbol{\beta}}_{(0)}$? Find the covariance matrix $\mathrm{Cov}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)$ of $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0$. Show that the trace of $\frac{1}{\sigma^2} X^{\mathrm{T}} X \, \mathrm{Cov}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)$ is r.

Let $H_0 = X_0 (X_0^{\mathrm{T}} X_0)^{-1} X_0^{\mathrm{T}}$ be the projection matrix for projecting onto the column space of X_0 (the "hat matrix" of the submodel).

b) Show that $E(X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)) = (I - H_0) X \boldsymbol{\beta}$. Find EJ_0. (Hint: Use the trace formula, $E(\boldsymbol{Z}^{\mathrm{T}} A \boldsymbol{Z}) = \mathrm{tr}(A \,\mathrm{Cov}\, \boldsymbol{Z}) + (E\boldsymbol{Z}^{\mathrm{T}}) A (E\boldsymbol{Z})$.)

Let $\mathrm{SSE}_0 = \boldsymbol{Y}^{\mathrm{T}} (I - H_0) \boldsymbol{Y}$ be the error (residual) sum of squares of the submodel.

c) Show that its expected value is $E\,\mathrm{SSE}_0 = (n - r)\sigma^2 + \boldsymbol{\beta}^{\mathrm{T}} X^{\mathrm{T}} (I - H_0) X \boldsymbol{\beta}$. Combine the expressions for EJ_0 and $E\,\mathrm{SSE}_0$ to show that $EJ_0 = \frac{1}{\sigma^2} E\,\mathrm{SSE}_0 - n + 2r$. Discuss briefly how this motivates the use of Mallows' C_P statistic in submodel selection.