TMA4267 Linear Statistical Models
Part 2: Linear regresssion
Solutions to recommended exercise 4 - V2017

February 7, 2017

May 16, 2017: corrected transposes in Problem 3.

Problem 1: Pendulum

(Exam TMA4267 V2015, Problem 1)

a) The fitted regression model is $\widehat{T} = 0.44 + 0.0197L + 0.045\,\theta + 0.023\,m$, where T is period (in s), L length (cm), θ amplitude (radians) and m mass (kg).

The model explains 98% of the variation of the data. The hypothesis that all regression coefficients are zero is rejected. The intercept and the coefficients of length and mass are significantly different from zero at the 1% level.

The residual plot shows low residuals for small and large values of fitted periods and high residuals for medium values of fitted residuals, suggesting that the model is wrong. Many wrote in their papers that the appearance of the residual plot were due to residuals not being independent, but in fact they may be independent but the linear model wrong.

The Box–Cox plot suggests a square transform of the response variable T.

b) I would prefer the new model, since the residual plot shows no clear structure.

The best subset selection suggests that L should always be present as a covariate. With two covariates, also θ should be present. Based on Mallows' C_P, the overall "best" model is the one including L and θ only.

c) The estimate of the coefficient of $\ln L$ agrees well with the theory, which states that it should be $\frac{1}{2}$. The estimate of the coefficient of the term involving θ agrees less well – it should be 1; however, the standard error of the estimate is large.

The intercept β_0 of the regression model corresponds to $\ln(2\pi/\sqrt{g})$ of the physical model, $\beta_0 = \ln(2\pi/\sqrt{g})$. Thus, $\hat{\beta}_0 = \ln(2\pi/\sqrt{\hat{g}})$ defines an estimate of g. Solving, we get $\hat{g} = 4\pi^2 e^{-2\hat{\beta}_0} = 4\pi^2 e^{-2(-1.62)} = 1.0 \cdot 10^3$, that is, $1.0 \cdot 10^3$ cm/s^2 = 10 m/s^2 (the units of the data were s and cm).

$(\beta_0 - \hat{\beta}_0)/\operatorname{se}\hat{\beta}_0$ has the t distribution with $100 - 3 = 97$ degrees of freedom, where $\operatorname{se}\hat{\beta}_0$ denotes the standard error of $\hat{\beta}_0$ (the denominator in Corollary 3.33 of Bingham and Fry). Let t denote the upper 0.025 critical value of this distribution. Then

$$\begin{aligned}
0.95 &= P\bigg(-t < \frac{\beta_0 - \hat{\beta}_0}{\operatorname{se}\hat{\beta}_0} < t\bigg) = P\bigg(-t < \frac{\ln(2\pi/\sqrt{g}) - \hat{\beta}_0}{\operatorname{se}\hat{\beta}_0} < t\bigg) \\
&= P\big(\hat{\beta}_0 - t\operatorname{se}\hat{\beta}_0 < \ln(2\pi/\sqrt{g}) < \hat{\beta}_0 + t\operatorname{se}\hat{\beta}_0\big) = P\big(e^{\hat{\beta}_0 - t\operatorname{se}\hat{\beta}_0} < 2\pi/\sqrt{g} < e^{\hat{\beta}_0 + t\operatorname{se}\hat{\beta}_0}\big) \\
&= P\big(4\pi^2 e^{-2(\hat{\beta}_0 + t\operatorname{se}\hat{\beta}_0)} < g < 4\pi^2 e^{-2(\hat{\beta}_0 - t\operatorname{se}\hat{\beta}_0)}\big).
\end{aligned}$$

The statistical tables give $t = 1.98$, and from the R output we have $\hat{\beta}_0 = -1.62$ and $\operatorname{se}\hat{\beta}_0 = 0.0160$ for our data. Inserting these values in the inequalities above, we get $4\pi^2 e^{-2(-1.618+1.98\cdot 0.0160)} = 0.94 \cdot 10^3$ and $4\pi^2 e^{-2(-1.618-1.98\cdot 0.0160)} = 1.07 \cdot 10^3$ as bounds of the confidence interval, that is, the interval is $(9.4, 10.7)$ with unit $\mathrm{m/s^2}$.

Problem 2: Galapagos

(Exam TMA4267 V2014, Problem 2)

a) The fitted regression model is:

$$\widehat{\texttt{Species}} = 7.07 - 0.02 \cdot \texttt{Area} + 0.32 \cdot \texttt{Elevation} + 0.009 \cdot \texttt{Nearest} - 0.24 \cdot \texttt{Scruz} - 0.75 \cdot \texttt{Adjacent}$$

This model explains 77% of the variability in the data. The regression is significant (the hypothesis that all regression coefficients are zero is rejected) and t-tests claim that `Elevation` and `Adjacent` are significant covariates.

The residual plots: The plot of studentized residuals vs. fitted values hints to heteroscedasticity in the errors (differing variances), and the qq-plot shows deviance from the normal distribution in the tails. The latter is also observed by looking at the Anderson-Darling normality test, which gives a p-value of 0.0002 (reject the null hypothesis that the errors are normal). The Box-Cox plot doesn't include 1 in the 95% confidence interval (dotted lines in the plot), and suggests that the cube root transform ($\lambda = 1/3$) may be suitable as a variance stabilizating transform.

b) Let us assume that we have p covariates – where the intercept is included. (Different sources include and not the intercept.)

Estimate (Intercept): t value*Std. Error=7.365· 0.305=2.25
Meaning: estimate for the regression coefficient. Intercept associated with first column of design matrix (first column of ones) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

p-value of `Area`:two tails of t-distribution with 24 degrees of freedom. Can't find precise value, but from table 4 of Tabeller og formler i statistikk we see that the critical value in the t-distribution with 24 degrees of freedom is 2.064 for $\alpha = 0.025$. This means that the p-value will approximately be 0.05.
Meaning: Test the null hypothesis that $\beta_{Area} = 0$ vs. $\beta_{Area} \neq 0$, with the other four covariates present in the model, and produce a p-value of the test.

Std.Error of `Nearest`:estimate/tobs=0.012/0.7=0.017
Meaning: the estimated standard deviation for the regression estimate. Mathematically the

corresponding (`Nearest`) diagonal element of the square root of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}s^2$, where s^2 is the estimate for the regression variance σ^2.

Adjusted R-squared: $1-(1-R^2)(n-1)/(n-p) = 1-(1-0.7543)\cdot 29/24 = 0.7032$, or in a two stage process by first observing SSE=$s^2\cdot(n-p) = 0.9716^2\cdot 24 = 22.65$ and then finding SST from R^2=1-SSE/SST, SST=SSE/(1-R^2), and finally using $R^2_{\text{adj}} = 1-\frac{\frac{\text{SSE}}{n-p-1}}{\frac{\text{SST}}{n-1}}$.

Yes, I would prefer model B to A. The plot of standardized residuals vs fitted values shows no clear structure, and the qq-plot looks much better for B and A. The Anderson-Darling normality test doesn't reject the null hypothesis of normal data.

c) Let SSE be the sum-of-squares of error, SSR be the regression sum-of-squares, and SST be the total som of squares. Then R^2: coefficient of multiple determination is defined as

$$1-\text{SSE/SST} = \text{SSR/SST}$$

and is interpreted as the amount of variability in the data that is accounted for by the regression. R^2 will increase when regressors are added to the model, even if the new regressors are independent of the response. Why? The least squares estimator will minimize SSE and if the regression coefficient for the new regressor is estimated to be a value different from zero, this means that the SSE of this larger model will be smaller than the SSE of the smaller model.

The R^2_{adj} is constructed to also include information about the number of parameters estimated and the number of observations in the data set. Assume we have p regression parameters, then

$$R^2_{\text{adj}} = 1-\frac{\frac{\text{SSE}}{n-p}}{\frac{\text{SST}}{n-1}}$$

R^2 will always increase when new covariates are added to the model, so R^2 can only be used to select the best model among models with the same number of covariates. This is done when in best subset selection one model is reported for each total number of covariates. To choose between these models a criterion taking into account the number of covariates in the model need to be used, and one such criterion is R^2_{adj}. We therefore use R^2_{adj} to choose between the best models of each size.

In our example the best model is according to this strategy the model with four covariates. These are all covariates except `Nearest`. To write down the estimated regression equation we need to refit this model.

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

a) $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$, so that the linear transformation $\boldsymbol{X}_0^T\hat{\boldsymbol{\beta}}$ of $\hat{\boldsymbol{\beta}}$ has the distribution $N(\boldsymbol{X}_0^T\boldsymbol{\beta}, \sigma^2\boldsymbol{X}_0^T(X^TX)^{-1}\boldsymbol{X}_0)$, which is univariate. So its expected value is the EY_0 specified by the model, and is thus unbiased.

b) $\text{SSE}/\sigma^2 \sim \chi^2(n-p)$, where n is the length of $\boldsymbol{Y}$ and p is the length of $\boldsymbol{\beta}$, and it is independent of $\hat{\boldsymbol{\beta}}$ and thus of $\boldsymbol{X}_0^T\hat{\boldsymbol{\beta}}$. Standardize $\boldsymbol{X}_0\hat{\boldsymbol{\beta}}$ and divide it by $\sqrt{\text{SSE}/((n-p)\sigma^2)}$ to get

$$T = \frac{\boldsymbol{X}_0^T\hat{\boldsymbol{\beta}} - EY_0}{\sqrt{\boldsymbol{X}_0^T(X^TX)^{-1}\boldsymbol{X}_0\,\text{SSE}/(n-p)}} \sim t(n-p).$$

Solve the double inequality $-t_{n-p} \leq T \leq t_{n-p}$, where $-t_{\alpha/2}$ is the $\alpha/2$-quantile of a $t(n-p)$ variable, to get a confidence interval having bounds

$$\boldsymbol{X}_0^T\hat{\boldsymbol{\beta}} \pm \sqrt{\boldsymbol{X}_0^T(X^TX)^{-1}\boldsymbol{X}_0\,\mathrm{SSE}/(n-p)}.$$

c) Since Y_0 is independent of $\boldsymbol{\epsilon}$ and thus of $\boldsymbol{X}_0^T\hat{\boldsymbol{\beta}}$, $Y_0 - \boldsymbol{X}_0^T\hat{\boldsymbol{\beta}} \sim N(0, \sigma^2(1+\boldsymbol{X}_0^T(X^TX)^{-1}\boldsymbol{X}_0))$. Standardize $Y_0 - \boldsymbol{X}_0^T\hat{\boldsymbol{\beta}}$ and proceed as in (b) to get a prediction interval having bounds

$$\boldsymbol{X}_0^T\hat{\boldsymbol{\beta}} \pm \sqrt{\left(1+\boldsymbol{X}_0^T(X^TX)^{-1}\boldsymbol{X}_0\right)\mathrm{SSE}/(n-p)}.$$

d)

```
# How does pH in Nowegian lakes depend on sulfate, nitrate, calcium, aluminium
# and organic content (x1, ... x5), area of lake (x6) and location (x7 = 0,
# Telemark, or x7 = 1, Tr?ndelag)? Data from Statens forurensningstilsyn
# (1986). Here 26 random lakes from Telemark and Tr?ndelag out of 1005 lakes
# have been drawn .

acidrain <-
  read.table("https://www.math.ntnu.no/emner/TMA4267/2017v/acidrain.txt",
  header=TRUE)

attach(acidrain)

n <- length(y)

x <- cbind(rep(1,n),acidrain[,2:8]) # we want intercept in model
names(x)[1] <- 1
x <- as.matrix(x)
p <- dim(x)[2]

i <- diag(n)
h <- x%*%solve(t(x)%*%x)%*%t(x)

# Test of submodel where coefficients of x2, x4, x5, x6, x7 are zero:
r <- 3
x0 <- x[,c(1,2,4)]
h0 <- x0%*%solve(t(x0)%*%x0)%*%t(x0)
f <- t(y)%*%(h-h0)%*%y/(p-r)/(t(y)%*%(i-h)%*%y/(n-p))
f
pf(f,p-r,n-p,lower.tail=FALSE) # p-value - cannot reject null hypothesis
# or by R functions (1 is added by R as a covariate):
fit <- lm(y~x1+x2+x3+x4+x5+x6+x7)
fit0 <- lm(y~x1+x3)
anova(fit0,fit)

# Test of whether all coefficients except the intercept are zero
r <- 1
x0 <- x[,1]
h0 <- x0%*%solve(t(x0)%*%x0)%*%t(x0)
f <- t(y)%*%(h-h0)%*%y/(p-r)/(t(y)%*%(i-h)%*%y/(n-p))
f
pf(f,p-r,n-p,lower.tail=FALSE) # p-value - cannot reject null hypothesis
# or by R functions:
fit0 <- lm(y~1)
anova(fit0,fit)
# or:
fit2 <- lm(y~x)
anova(fit2)

# Confidence interval for EY0 and prediction interval for Y0 with new
# covariate vector
x0 <- c(1,3,50,1,50,2,1,0) # remember intercept (the first 1)
```

```
sse <- y%*%(i-h)%*%y # R automatically transposes vector when necessary
betahat <- solve(t(x)%*%x)%*%t(x)%*%y
# Confidence interval:
halflength <-
  qt(.025,n-p,lower.tail=FALSE)*sqrt(x0%*%solve(t(x)%*%x)%*%x0*sse/(n-p))
sum(x0*betahat)-halflength
sum(x0*betahat)+halflength
# Prediction interval:
halflength2 <-
  qt(.025,n-p,lower.tail=FALSE)*sqrt((1+x0%*%solve(t(x)%*%x)%*%x0)*sse/(n-p))
sum(x0*betahat)-halflength2
sum(x0*betahat)+halflength2

# By R functions instead:
newdata <- data.frame(x1=3,x2=50,x3=1,x4=50,x5=2,x6=1,x7=0)
# or the following two lines
newdata <- data.frame(3,50,1,50,2,1,0)
names(newdata) <- names(coefficients(fit)[-1])
# In either case:
predict(fit,newdata,level=.95,interval="confidence")
predict(fit,newdata,level=.95,interval="prediction")
```

e) The design matrix in this case is

$$X = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}^T,$$

and you can verify that

$$\boldsymbol{X}_0^T (X^T X)^{-1} \boldsymbol{X}_0 = \frac{\sum_i (x_i - x_0)^2}{n \sum_i (x_i - \bar{x})^2}.$$

Next, write $x_i - x_0 = (x_i - \bar{x}) + (\bar{x} - x_0)$ to get $\sum_i (x_i - x_0)^2 = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - x_0)^2$ (crossterms vanish). The bounds given follow.