# Module 3: Recommended Exercises

TMA4268 Statistical Learning V2020

*Martina Hall, Michail Spitieris, Stefanie Muff, Department of Mathematical Sciences, NTNU*

*January 23, 2020*

Last changes: 22.01.2020

---

**We strongly recommend you to work through the Section 3.6 in the course book (Lab on linear regression)**

---

You need to install the following packages in R to run the code in this file. Again, you only need to install the packages if you have not done it so far (and only before the first time you use them).

```
install.packages("knitr")  #probably already installed
install.packages("rmarkdown")  #probably already installed
install.packages("GLMsData")  #data set for Problem 3
install.packages("ggplot2")  #plotting with ggplot
install.packages("ISLR")  #data problem 1
install.packages("ggfortify")  #diagnostic plots for lm objects
```

## Problem 1 (Book Ex. 9)

This question involves the use of multiple linear regression on the `Auto` data set from `ISLR` package (you may use `?Auto` to see a description of the data). First we exclude from our analysis the variable `name`.

```
library(ISLR)
Auto = subset(Auto, select = -name)
# Auto$origin = factor(Auto$origin)
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
```

**a)**

Use the function `ggpairs()` from `GGally` package to produce a scatterplot matrix which includes all of the variables in the data set.

**b)**

Compute the correlation matrix between the variables.

**c)**

Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables (except `name`) as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

ii. Which predictors show evidence that they are related to the response?

iii. What does the coefficient for the year variable suggest?

**d)**

Use the `autoplot()` function from the `ggfortify` package to produce diagnostic plots of the linear regression fit by setting `smooth.colour = NA`, as sometimes the smoothed line can be misleading. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

**e)**

For beginners, it can be difficult to decide whether a certain QQ plot looks "good" or "bad", because we only look at it and do not test anything. A way to get a feeling for how "bad" a QQ plot may look, even when the normality assumption is perfectly ok, we can use simulations: We can simply draw from the normal distribution and plot the QQ plot. Use the following code to repeat this six times:

```
set.seed(2332)
n = 100

par(mfrow = c(2, 3))
for (i in 1:6) {
    sim = rnorm(n)
    qqnorm(sim, pch = 1, frame = FALSE)
    qqline(sim, col = "blue", lwd = 1)
}
```

**f)**

Let us look at interactions. These can be included via the `*` or `:` symbols in the linear predictor of the regression function (see Section 3.6.4 in the course book).

Fit the same model as before, but now also include an interaction term between `year` and `origin`. Note that `origin` is encoded as 1, 2, 3, but it is actually a qualitative predictor with three levels! To ensure that R treats it correctly, we first need to convert `origin` into a factor variable (a synonymous for "qualitative predictor"):

```
Auto$origin = factor(Auto$origin)
```

Now fit the model. Is there evidence that the interactions term is relevant? Give an interprentation of the result.

```
fit.lm1 = lm(mpg ~ displacement + weight + year * origin, data = Auto)
summary(fit.lm1)
anova(fit.lm1)
```

### g)

Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, $X^2$. See Section 3.6.5 in the course book for how to do this. Comment on your findings.

## Problem 2: Theoretical questions & Simulations

### a)

A core finding for the least-squares estiamtor $\hat{\boldsymbol{\beta}}$ of linear regression models is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \ ,$$

with $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

- Show that $\hat{\boldsymbol{\beta}}$ has this distribution with the given mean and covariance matrix.
- What do you need to assume to get to this result?
- What does this imply for the distribution of the $j$th element of $\hat{\boldsymbol{\beta}}$?
- In particular, how can we calculate the variance of $\hat{\beta}_j$?

### b)

What is the interpretation of a 95% confidence interval? Hint: repeat experiment (on $Y$), on average how many CIs cover the true $\beta_j$? The following code shows an interpentation of a 95% confidence interval. Study and fill in the code where is needed

- Model: $Y = 1 + 3X + \varepsilon$, with $\varepsilon \sim \mathsf{N}(0, 1)$.

```
beta0 = ...
beta1 = ...
true_beta = c(beta0, beta1)  # vector of model coefficients
true_sd = 1  # choosing true sd
X = runif(100, 0, 1)  # simulate the predictor variable X
Xmat = model.matrix(~X, data = data.frame(X))  # create design matrix


ci_int = ci_x = 0  # Counts how many times the true value is within the confidence interval
nsim = 1000
for (i in 1:nsim) {
    y = rnorm(n = 100, mean = Xmat %*% true_beta, sd = rep(true_sd, 100))
    mod = lm(y ~ x, data = data.frame(y = y, x = X))
    ci = confint(mod)
    ci_int[i] = ifelse(..., 1, 0)  # if true value of beta0 is within the CI then 1 else 0
    ci_x[i] = ifelse(..., 1, 0)  # if true value of beta_1 is within the CI then 1 else 0
}

c(mean(ci_int), mean(ci_x))
```

**c)**

What is the interpretation of a 95% prediction interval? Hint: repeat experiment (on $Y$) for a given $\boldsymbol{x}_0$. Write R code that shows the interprentation of a 95% PI. Hint: In order to produce the PIs use the data point $x_0 = 0.4$. Furthermore you may use a similar code structure as in b).

**d)**

Construct a 95% CI for $\boldsymbol{x}_0^T\beta$. Explain what is the connections between a CI for $\beta_j$, a CI for $\boldsymbol{x}_0^T\beta$ and a PI for $Y$ at $\boldsymbol{x}_0$.

**e)**

Explain the difference between *error* and *residual*. What are the properties of the raw residuals? Why don't we want to use the raw residuals for model check? What is our solution to this?

# Problem 3 (Compulsory 1, 2019)

The lung capacity data `lungcap` (from the `GLMsData` R package) gives information on health and on smoking habits of a sample of 654 youths, aged 3 to 19, in the area of East Boston during middle to late 1970s.

We will focus on modelling forced expiratory volume `FEV`, a measure of lung capacity. For each person in the data set we have measurements of the following 5 variables:

- `FEV` the forced expiratory volume in litres, a measure of lung capacity; a numeric vector,
- `Age` the age of the subject in completed years; a numeric vector,
- `Ht` the height in inches; a numeric vector,
- `Gender` the gender of the subjects: a numeric vector with females coded as 0 and males as 1,
- `Smoke` the smoking status of the subject: a numeric vector with non-smokers coded as 0 and smokers as 1

First we transform the height from inches to cm. Then a multiple normal linear regression model is fitted to the data set with `log(FEV)` as response and the other variables as covariates. The following R code may be used.

```
library(GLMsData)
data("lungcap")
lungcap$Htcm = lungcap$Ht * 2.54
modelA = lm(log(FEV) ~ Age + Htcm + Gender + Smoke, data = lungcap)
summary(modelA)
```

```
##
## Call:
## lm(formula = log(FEV) ~ Age + Htcm + Gender + Smoke, data = lungcap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63278 -0.08657  0.01146  0.09540  0.40701
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.943998   0.078639 -24.721  < 2e-16 ***
## Age          0.023387   0.003348   6.984  7.1e-12 ***
```

```
## Htcm         0.016849   0.000661  25.489  < 2e-16 ***
## GenderM      0.029319   0.011719   2.502   0.0126 *
## Smoke       -0.046067   0.020910  -2.203   0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1455 on 649 degrees of freedom
## Multiple R-squared:  0.8106, Adjusted R-squared:  0.8095
## F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16
```

We call the model fitted above `modelA`.

### a)

Write down the equation for the fitted `modelA`.

### b)

Explain (with words and formulas) what the following in the `summary`-output means, use the `Age` and/or the `Smoke` covariate for numerical examples.

- `Estimate` - in particular interpretation of `Intercept`
- `Std.Error`
- `Residual standard error`
- `F-statistic`

### c)

What is the proportion of variability explained by the fitted `modelA`? Comment.

### d)

Produce residual plots using the `autoplot()` function, and comment on what you see.

### e)

Now fit a model, call this `modelB`, with `FEV` as response, and the same covariates as for `modelA`. Would you prefer to use `modelA` or `modelB` when the aim is to make inference about `FEV`? Explain what you base your conclusion on.

```
modelB = lm(FEV ~ Age + Htcm + Gender + Smoke, data = lungcap)
autoplot(modelB)
```

### f)

Construct a 95% and a 99% confidence interval for $\beta_{\text{Age}}$ (write out the formula and calculate the interval numerically). Explain what these intervals tell you.

### g)

Consider a 16 year old male. He is 170 cm tall and not smoking.

```
new = data.frame(Age = 16, Htcm = 170, Gender = "M", Smoke = 0)
```

What is your best guess for his `log(FEV)`? Construct a 95% prediction interval for his forced expiratory volume `FEV`. Comment. Hint: first contruct values on the scale of the response `log(FEV)` and then transform the upper and lower limits of the prediction interval. Do you find this prediction interval useful? Comment.