Series 2

1. The following R-code genereates an artificial dataset with predictors `x1`, `x2` and response `y`.

```
> set.seed(0)
> n<-100
> z1<-rnorm(n)
> z2<-rnorm(n)
> M=matrix(c(1,1,0.1,-0.1),2,2)
> X=t(M%*%rbind(z1,z2))
> beta<-c(0.5,-1.0)
> x1=X[,1]
> x2=X[,2]
> y=5+beta[1]*x1+beta[2]*x2 +rnorm(n)
```

a) Create a plot of the observations of the two predictor variables `x1` and `x2`.

b) Fit a linear model `fit1<-lm(y~x1+x2)` and print the summary using `summary(fit1)`.

c) Recompute the t-value corresponding to $\hat{\beta}_1$ by hand using the estimate $\hat{\beta}_1$ and its estimated standard error $\mathrm{se}(\hat{\beta}_1)$.

d) Give the definition of a p-value. Then compute the p-value corresponding to $\hat{\beta}_1$ using the t-value from part c) and the quantile function of the `t`-distribution `pt()`.
Note: You need to provide the correct number of degrees of freedom.

e) Report the p-value of the overall F-test and reproduce it using `anova()`.

f) The overall F-test is significant. However, the p-values for `x1` and `x2` are not significant. Explain how this can be true.

g) Report the residual standard error, interpret it, and recompute it based on `residuals(fit1)`.

h) Report the R^2 value, interpret it, and recompute it using `residuals(fit1)`.

i) Assume now that we only observed the values for `x1` and `y` whereas `x2` is a hidden predictor that we do not observe. Fit the model `fit3<-lm(y~x1)` and print the summary `summary(fit3)`. Compare the estimated coefficient corresponding to `x1` to the one in part b). Interpret the coefficient of `x1` in both models.

2. In this exercise, we will code a categorical variable by hand. The dataset `Carseats` contains the number of child car seat sales and several predictors in 400 locations. We will only use the quantitative predictor `advertising` (local advertising budget for company at each location in thousands of dollars) and the qualitative predictor `shelveloc` (a factor with levels 'Bad', 'Good' and 'Medium' indicating the quality of the shelving location for the car seats at each site). Consider the following R code:

```
> # prepare data
> library(ISLR)
> data(Carseats) #use ?Carseats for an explaination of the dataset
> shelveloc=Carseats$ShelveLoc
> sales=Carseats$Sales
> advertising=Carseats$Advertising
> # fit using automatic coding
> fit<-lm(sales~shelveloc+advertising)
> summary(fit)
```

a) Encode the factor variable `shelveloc` in the same way as done automatically by R by constructing appropriate predictors `a1` and `a2`. `a1` shall be 1 when the level of `shelveloc` is `medium` and `a2` shall be 1 if its level is `good`. The so-called *contrast coding* in this case can be seen in Table 1. Fit the model `fit_a<-lm(sales~a1+a2+advertising)`. Verify that `fit` and `fit_a` are indeed equal and give an interpretation of the coefficients corresponding to `a1` and `a2`.
R-hint:

```
> # boolean vectors for easy construction of a1, a2, b1,...
> bad<- levels(shelveloc)[1]==shelveloc
> medium<- levels(shelveloc)[3]==shelveloc
> good<- levels(shelveloc)[2]==shelveloc
> a1<-medium*1
```

Table 1: Contrast codings in a), b), c)

shelveloc	a1	a2	shelveloc	b1	b2	shelveloc	c1	c2	c3
bad	0	0	bad	1	0	bad	1	0	0
medium	1	0	medium	0	0	medium	0	1	0
good	0	1	good	0	1	good	0	0	1

b) Construct predictor variables `b1` and `b2` according to the contrast coding in Table 1 and fit the model `fit_b<-lm(sales~b1+b2+advertising)`. Give an interpretation of the coefficients of `b1` and `b2`.

c) Construct predictor variables `c1`, `c2` and `c3` according to Table 1.
Then fit the model `fit_c<-lm(sales~c1+c2+c3+advertising)`. This causes a problem. Why?

d) Remove the intercept by using `fit_c<-lm(-1+...)`. Interpret the coefficients corresponding to `c1`, `c2` and `c3`.

e) Show that the fitted values are the same for `fit_a`, `fit_b` and `fit_c`.
Note: Due to rounding errors the values are not *exactly* the same. Show that they are very close.
R-hint: `max(abs(fitted(fit_a)-fitted(fit_b)))`

f) We now want to know if distinguishing between all three categories is significantly better than distinguishing only between "bad" (level `bad`) and "not bad" (level `medium` or `good`) each time also accounting for advertising. In which of the summaries of the fits `fit_a`, `fit_b`, `fit_c` can we see this directly? Explain.

g) Suppose we used the coding from `fit_a`. Conduct a partial F-test to check if we need to distinguish between `medium` and `good` by fitting a model `fit_d` with a new dummy variable.

3. The dataset `airline` contains the monthly number of flight passengers in the USA in the years 1949-1960 ranging from January 1949 to December 1960. Read the data with the command:

```
airline <- scan("http://stat.ethz.ch/Teaching/Datasets/airline.dat")
```

a) Plot the data against time and describe what you observe.

b) Compute the logarithm of the data and plot against time. Comment on the difference.

c) Define a linear model of the form

$$\log(y_t) = \beta t + \sum_{j=1}^{12} \gamma_j x_{tj} + \epsilon_t$$

where the month is coded in the predictors $x_{\cdot,1}, \ldots x_{\cdot,12}$, i.e. for $j \in \{1, \cdots, 12\}$

$$x_{tj} = \begin{cases} 1 & \text{if } t \text{ corresponds to the } j\text{-th month in a year} \\ 0 & \text{otherwise.} \end{cases}$$

and $t \in \{1, \cdots, 144\}$ is the month index starting with 1 for January 1949. Construct appropriate predictors `t`, `x1`,...,`x12` and fit this model in R.
R-hint: You should not use an intercept parameter (see 2 c)). Use `-1` in the model formula of `lm()` to exclude the intercept.
R-hint: `x1<-rep(c(1,rep(0,11)),12)` and `t<-1:144`.

d) Plot the fitted values and residuals against time. Do you think that the model assumptions hold?

e) Give an interpretation of the parameter β in the above model if we consider the original scale.
Hint: How does a model prediction $\widehat{y}_t := \exp(\widehat{\log(y_t)})$ change if we increase t by 12?

f) Conduct a partial F-test to check whether we can use four predictors indicating the seasons $s_1, \cdots, s_4$ (s_1 for spring (month 3,4,5),..., s_4 for winter (month 12,1,2)) instead of twelve indicators $x_1, \cdots, x_{12}$ encoding the month.
R-hint: Construct appropriate predictors `s1`,...,`s4` for the seasons.

Preliminary discussion: Friday, March 08.

Deadline: Friday, March 15.