Prof. Gasparini - Mathematics in Machine Learning
Politecnico di Torino - 20/04/22

Exercises on Linear Models

1. Replicate in Phyton the results of this short version of the R analysis of the insulate data.

```
# reading the data
insulate=read.table("220330insulate.txt", col.names=c("quando","temp","cons"))
attach(insulate) # attach/detach dynamics
# fitting a regression model for consumption with interaction
summary(regr2)
# obtain 90\% confidence intervals for the regression coefficients
confint(regr2, level=0.9)
detach(insulate) # attach/detach dynamics
```

2. With reference to the insulate case study, perform an F-test of the small additive model with respect to the large model with interaction. This is obtained in R using `anova(small,large)`, but also just simply looking at an individual t-test for the interaction term, since that is the only one term regarding interactions.

3. With reference to the insulate case study:

   (a) calculate 99% confidence intervals for the mean response corresponding to quando=prima and temp=3.2 and corresponding to quando=dopo and temp=3.2.

   (b) Calculate 99% prediction intervals for the same two new situations.

4. Generate 100 i.i.d. standard normal observations and call them x, then generate 100 independent observations with mean $1+2x$ and the same standard deviation $\sigma = 0.1$ and call them y. Fit a linear model of y on x and x^2 and compute the 95% confidence intervals for the β coefficients. Comment on the results.

5. The file `220420EXECSAL.txt` is taken from Mendenhall and Sincich (7th ed) *A Second Course in Business Statistics: Regression Analysis* and it contains the salaries of 100 executives together with their years of experience, years of education, gender (male=1), number of supervised people, and a quantification of the assets of the company they work for:

   (a) using residual plots, show that a linear model of log(SALARY) on all other features of the company is better than a linear model of SALARY on all other features;

   (b) interpret the coefficient of EXP in the linear model of log(SALARY) on all other features.

6. The distances obtained having 10 different golf players hit four different brands of golf balls are collected. We are interested in comparing brands, not players. The experiment is balanced and randomized, since each player hits a ball of each of the four different brands in a randomized order. Such an experimental plan is called a *randomized block design*. The example taken from McClave JT., Benson PG. e Sincich T. (2014). *Statistics for Business and Economics.* Pearson Education.

GOLFER	A	B	C	D
1	202.4	203.2	223.7	203.6
2	242	248.7	259.8	240.7
3	220.4	227.3	240	207.4
4	230	243.1	247.7	226.9
5	191.6	211.4	218.7	200.1
6	247.7	253	268.1	195.8
7	214.8	214.8	233.9	227.9
8	245.4	243.6	257.8	227.9
9	224	231.5	238.2	215.7
10	252.2	255.2	265.4	245.2

(a) Read in the data and transform to long format.

(b) fit an additive model containing both BRAND and GOLFER as qualitative predictors (factors) for DISTANCE.

(c) It would not be possible to introduce all brand by player interaction terms with this data. Why?

(d) Test for BRAND differences; are they significant? Are they important? Which is the best brand?

(e) Test for GOLFER differences; are they significant? Are they important?

(f) Calculate a 95% confidence interval for the mean distance difference between BRAND C and BRAND A.

(g) Calculate a 95% confidence interval for the mean distance difference between BRAND C and BRAND A without accounting for GOLFER and comment on the difference with the previous one.

Solutions
4(f): [9.70, 26.86]
4(g): [1.65, 34.91]