**Quiz 3**

*Content: Linear model, adding categorical X into model*

*Functions in R: lm(), predict().*

Consider data set <crab.csv> where we are interested in satell - the number of male crabs grouped around a female horse shoe crab during the mating season (called satellite). The possible regressors are the **weight** (kg) and the **width** (cm) of the female crab; the female crab's **color** (categorical, 2 = light; 3 = medium, 4 = dark, and 5 = darker).

Data set was imported into R and a model with 3 variables weight, width and color was built.

**Extra question.** How many coefficients that the correct model with 3 variables (weight, width and color) should have (including the intercept)?

4, 5, 6, 7

ANS: 6

Intercept + 1 for weight + 1 for width + 3 for color (because color is categorical)🡪6

**Q1.** R code is given in the photo below. How many coefficients does model M1 created from this code have?

4, 5, 6, 7

ANS: 4.   
Because column “color” inside data frame “crab” is still recognized by R as a numeric column by the code at line 8. In line 10, *as.factor()* is used, but it is to create a vector called “color” – be a categorical, but this vector is NOT inside data frame “crab”.

Model M1 is fitted in line 12 uses column “color” inside data frame “crab” which is still numeric.

**Q2.** In the R code given, how many coefficients does model M2 have?

4, 5, 6, 7

ANS: 6  
R code for model M2 in line 12 doesn’t specify the data frame where the variables come from. Hence, R will use satell, width and weight from data frame “crab” since the command attach(crab) was run, hence R knows what satell, width and weight are. For color, after *attach(crab)*, R updates that “color” is categorical from the code in line 10.

A screenshot of a computer

Description automatically generated

**Q3.** The coefficient table from the summary output of model M2 and model M3 will be the same. True or False?

ANS: TRUE

The R code in line 16 has declared with R that column “color” inside data frame “crab” is categorical, hence, when fitting M3 with “*data = crab*” in line 18, R will have 3 coefficients for color.

**It’s recommended to fit a linear model as M3.**

**Q4.** What is the output of the code at line 20, <*predict(M3)*>? Multiple Answer

The fitted values of model M3.

The predicted values of satell from model M3, for all the crabs in the data set.

The real values of response for the crabs in the data set.

ANS:

The fitted values of model M3.

The predicted values of satell from model M3, for all the crabs in the data set.

**Q5.** To get the predicted response values using model M3, for all the crabs given in the data set, we use R code in line 22, <*M3$fitted.values*>. True or False?

ANS: TRUE

**Q6.** Using model M3 to predict the number of satell for a new crab with color = 4, width = 26 cm, weight = 2.6 kg. The code is given in lines 24 and 26. The code is correct. True or False?

ANS: FALSE.

Model M3 has color as categorical, hence when creating a new crab, we need to specify color as categorical, by putting the number inside a quote sign, such as:

*new.crab = data.frame(weight = 2.6, width = 26, color = “4”)*

*predict(M3, newdata = new.crab)*

should predict the number of satellites for the new crab.

**Q7.** A simple model M4 is fitted where the number of satellites only depends on the weight.

After that, using model M4 to predict the response for the new crab with weight = 2.6 kg. R code is given below. The code is correct. True or False?

A screenshot of a computer

Description automatically generated

Ans: FALSE.

From the code in line 28, we told R that model M4 is built with the values for “weight” that must be from a column inside a data frame called “crab”, specified by crab$weight.

Hence, when we use model M4 to predict for a new data point where this new one has “weight” from a data frame with the name “new” (in the line 30), R will not produce the prediction we want.

R will just produce the prediction for the data that were used to fit the model, “crab”.

How to fix the issue above?

Either we change the way to fit M4 as below, then the code will work:

M4 = lm(satell ~ weight, data = crab) *# note: do not use the form of data$y ~ data$x*

new = data.frame(weight = 2.6)

predict(M4, newdata = new)

OR we must name the new data frame as the same as the original data frame as below.

M4 = lm(crab$satell ~ crab$weight, data = crab)

crab = data.frame(weight = 2.6)

predict(M4, newdata = crab)

Note: this latter method is not recommended, since if we do this, we will not have the original data frame “crab” in R anymore, since it is now replaced by the new one.