PROJECT 1: LINEAR REGRESSION
MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION
(WITH DATA GATHERING), 2022
Peer assessment version: **12.30 on Monday 11 April**
Peer assessment comments: **13.00 on Tuesday 12 April**
Final version: **17.00 on Wednesday 13 April**

Determinants of Plasma Beta-Carotene Levels

Introduction

Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. *Determinants of plasma levels of beta-carotene and retinol.* American Journal of Epidemiology 1989;130:511-521.

> Observational studies have suggested that low dietary intake or low plasma concentrations of beta-carotene or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. We designed a cross-sectional study to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of beta-carotene and other carotenoids. Study subjects ($N = 315$) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. We display the data for only one of the analytes.
>
> We conclude that there is wide variability in plasma concentrations of these micronutrients in humans, and that much of this variability is associated with dietary habits and personal characteristics. A better understanding of the physiological relationship between some personal characteristics and plasma concentrations of these micronutrients will require further study.

Data file

The datafile `plasma.txt` contains 315 observations on 13 variables and can be downloaded from the course home page. Save it to your R data directory and then read it into R and put it in a data frame called `plasma` and look at it with

```
plasma <- read.delim("Data/plasma.txt")
head(plasma)
summary(plasma)
```

Variable description

age	Age (years)
sex	Sex (1 = Male, 2 = Female).
smokstat	Smoking status (1 = Never, 2 = Former, 3 = Current Smoker)
quetelet	Quetelet (weight/height^2 kg/m^2) a.k.a. BMI
bmicat	BMI category (1 = Underweight, 2 = Normal, 3 = Overweight, 4 = Obese)
vituse	Vitamin use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
calories	Number of calories consumed per day.
fat	Grams of fat consumed per day.
fiber	Grams of fiber consumed per day.
alcohol	Number of alcoholic drinks consumed per week.
cholesterol	Cholesterol consumed (mg per day).
betadiet	Dietary beta-carotene consumed (μg per day).
betaplasma	Plasma beta-carotene (ng/ml)

Part 1. Plasma beta-carotene and age

(a). Start by modeling how plasma beta-carotene varies with age. We are choosing between a model where `betaplasma` varies as a linear function of `age`, and an alternative model where `log(betaplasma)` varies as a linear function of `age`.

Note: This will cause a problem with one observation. Which one and why? In order of do all the analyses without this problematic observation you should first create a new data frame using `newdata <- olddata[olddata$variable > 0, ]`. This will create a new data frame using only the rows where `variable` is greater than 0.

Use the techniques from Lab 1 to determine which of the two models is best. Your report should include both models, their parameter estimates with confidence intervals, and a residual analysis.

In the rest of the project, use the transformation from the preferred model.

(b). Plot plasma beta-carotene against age together with the estimated relationship, a confidence interval for the fitted line and a prediction interval for new observations.

Describe what happens, on average, to the plasma beta-carotene level if we increase the age by 1 year. Include a 95 % confidence interval for this change rate.

Does the size of the change (measured in ng/ml) depend on age? Explain why and illustrate by calculating the additive difference in expected plasma beta-carotene levels between a 25-year old person and a 26-uear old, comparing with the difference between a 75-year old and a 76-year old. Relate the size of the differences to the yearly change.

(c). Calculate a 95 % prediction interval for the observed plasma beta-carotene of a 25 year old person, as well as for a 75 year old person. Are there any substantial differences in the widths of the two intervals? Why or why not?

Part 2. Plasma Beta-carotene and the background variables

We now want to model how (the suitable transformation of) Plasma Beta-carotene varies as a function of all the background variables, i.e, age, sex, smoking and BMI, using the techniques from Lab 2. We have two variables for BMI, the continous version, `quetelet`, and the categorical version, `bmicat`, where the categories are defined as Underweight = BMI < 18.5, Normal = 18.5–25, Overweight = 25–30, and Obese = BMI > 30.

(a). Preliminaries: Start by turning the categorical variables into factors, e.g.,

```
plasma$sex <- factor(plasma$sex,
   levels = c(1, 2),
   labels = c("Male", "Female"))
```

Also change `smokstat` and `bmicat` in the same way.

Present a frequency table for each of these variables and reflect on which category would be the best to use as reference category.

(b). Reference choice: Fit a model where plasma beta-carotene depends only on the categorical variable `bmicat`, using the default "Underweight" as reference category. Present the β-estimates and their standard errors.

Then change the reference category into a more suitable category (using `relevel`) and present the new β-estimates with standard errors.

Compare the sizes of the standard errors for these two models. Explain why they are so different.

In future models, use the suitable reference category. Also relevel the other categorical variables, if necessary.

(c). Are the variables significant: Fit a model where plasma beta-carotene depends on the continous variable `age` and the categorical variables `sex`, `smokstat` and `bmicat` and present a table with the β-estimates as well as the e^{β}-estimates together with confidence intervals for the e^{β}.

Perform the following tests

- Is this model significantly better than a model with only an intercept (null model)?
- Is this model significantly better than the model in Part 1, using only `age`?
- For each of the variables (`age`, `sex`, `smokstat` and `bmicat`), is a model with that variable significantly better than a model without that variable, keeping all the other variables?
- Is the "Underweight" BMI category significantly different from the reference category, given all the other variables?

Present the results in a table containing the type of test you use, the null hypothesis H_0, the test statistic, the distribution of the test statistic when H_0 is true, the P-value and the conclusion (yes/no). Explain why you choose the different types of tests and comment on the result.

(d). Plotting: Calculate the fitted values, confidence intervals, and prediction intervals for the data. Plot log plasma beta-carotene against age, add the fitted line and the intervals. Do not set colors for the different lines. Instead, use different colours for men and women, and separate subplots for the different combinations of smoke and bmi categories (The relevel command ensures that the BMI categories are plotted in the logical order):

```
ggplot(data = ..., aes(x = age, y = log(betaplasma), color = sex)) +
   geom_points() + [etc] +
   facet_grid(smokstat ~ relevel(bmicat, "Underweight"))
```

Do we have any underweight men in the data? Any underweight former smokers of either sex? Use the model to estimate the plasma beta-carotene level, with confidence and prediction intervals, for an underweight former smoking male of average age. Should we rely on this estimate?

What happens to the (log) plasma beta-carotene levels when the BMI increases? Relate this to the corresponding β-estimates.

(e). Continous BMI: It might be better to use the continous BMI-variable, `quetelet`, instead of the categorical version. Fit a new model with `age`, `sex`, `smokstat` and `quetelet` and present the β- and e^{β}-estimates, with confidence intervals for e^{beta}. Did the parameters for the other variables change in any substantial way?

Use both models to estimate, with confidence intervals, the average (median) plasma beta-carotene level in a man and in a woman, both 40-year old former smokers with a normal BMI of 22.

Explain why the confidence intervals for the man are wider than the corresponding intervals for the woman.

How would the expected plasma beta-carotene level change if we used a BMI of 33 (obese) instead of 22 (normal)? For both models, express the relative difference as a function of the relevant β-parameters and estimate it, with confidence intervals.

(f). Investigate what happens if we fit a model using *both* `quetelet` and `bmicat`, in addition to the other variables. Is `quetelet` significant in this extended model? Is `bmicat` significant? Explain why it is problematic to use both variables in the model.

Part 3. Model validation and selection

. . . in which we will use the dietary variables as well to build an even better model, compare the different models, even when they are not nested, and investigate any problematic observations.

NOTE: In order to preserve space in the report, resize single plots, i.e., when not using `+facet_wrap()`, so that you can fit at least two plots side by side on a page. But make sure the labels are still readable with `+theme(text = element_text(size = 18))` experimenting with the size.

(a). Compare the two models from Part 2(c) and 2(e) and determine which version of BMI is best to have in the model, the continous `quetelet` or the categorical `bmicat`, taking into account that the two models are not nested, and also that they have different number of parameters.

The better version of the model will be referred to as the *Background* model. The model from Part 1 with only age as covariate will be referred to as the *Age* model.

(b). Before we start adding the dietary x-variables to the model we much check that they are not highly correlated to each other. Create a data frame with all the continous x-variables

```
contx <- plasma[, c("age", "quetelet", "calories", "fat", "fiber",
                    "alcohol", "cholesterol", "betadiet")]
```

Then calculate all pairwise correlations between them with `cor(contx)` and plot them against each other with `pairs(contx)`, or, if you install the package `GGally` and load it with `library(GGally)`, you can get both the correlations and the plots at the same time with `ggpairs(contx)`.

Report all pairs where the correlation is stronger than ± 0.7 and present the plots for these pairs. Comment on any other potential troubles you find with any of the variables. Illustrate with a suitable plot.

Report a frequency table for the categorical variable `vituse` and select a suitable reference category.

(c). Ignore any potential problems and fit a model using all the dietary variables, `vituse`, `calories`, `fat`, `fiber`, `alcohol`, `cholesterol`, and `betadiet`.

Investigate the leverage for this set of covariates identifying any that are alarmingly high and identify the variable that is causing the problem, and why.

Reflect on whether it would be a good idea to use the logarithm of the alcohol consumption instead. Why not? *Hint*: how many persons do not consume any alcohol at all? Is the extreme person (drinking the equivalent of 1 liter of vodka a day!) extreme in any other variables as well? Investigate by plotting alchohol against each of the other dietary variables, and illustrate any interesting finds.

Note: Use the alcohol variable as it is in the models but keep track of the extreme individual and invesigate what problems it causes, if any.

(d). Make a visual inspection of the studentized residuals, looking for trends, outliers, and non-constant variance. using suitable plots, with reference lines. Highlight the observation with the problematic alcohol consumption. Comment on the results and identify any observations with an unusually large ($\pm$) residual for future reference.

(e). Perform a visual inspection of Cook's distance in a plot with suitable reference line(s). Highlight both the high leverage observation and the observation with the largest residual and comment on their influence on the estimates, if any. Also identify the observation with the largest Cook's D.

Investigate which of the parameters the observation with the largest residual and the observation with the largest Cook's D actually influenced by plotting DFBETAS for the different β-parameters against their corresponding x-variable, with suitable reference lines, and highlighting the three problematic observations (leverage, residual, Cook's D). Present the plots where the influence is detected and comment on the results.

(f). Use backwards elimination to reduce the mode, using AIC as criterion, and present the final model and a table with the resulting β-estimates together with the e^{β}-estimates and their 95 % confidence intervals.

Also comment on what happened to the variable pairs with strong correlation found in (b).

This reduced model will be referred to as the *Dietary* model.

(g). Combine the *Backgound* and *Dietary* models using a stepwise procedure, starting with the *Dietary* model. The largest model allowed should be the full model containing all variables present in either the *Background* or the *Dietary* model. The smallest model allowed should be the null model (`lm(log(betaplasma) ~ 1, ...)`).

Perform one version using AIC as criterion and another using BIC instead. Present the final models and a table with the resulting e^{β}-estimates and their 95 % confidence intervals. Use one row for each variable that is present in either model and separate columns for the two models, leaving empty places if the variable is not present in that model. Comment on any interesting differences between the two models.

These models will be referred to as the *Step AIC* and *Step BIC* model, respectively.

(h). Compare the five models, *Age*, *Background*, *Dietary*, *Step AIC*, and *Step BIC* in a table presenting their R^2 and R^2_{adj}.

How much of the variability of log plasma beta-carotene can be explained using only the background variables? Using only the dietary variables? Which model is best according to R^2_{adj}. This is the *Final* model.

At the beginning of your project report, include a very short Abstract or Executive summary (no formulas!), describing the *Final* model. Describe the reason for the log-transformation, which variables are left in the model, whether they each are associated with an increase or a decrease in plasma beta-carotene levels, and how much of the variability the model can explain.

End of Project 1