PROJECT 1: LINEAR REGRESSION
MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION (WITH DATA GATHERING), 2025

Peer assessment version: **12.30 on Monday 14 April**
Peer assessment comments: **13.00 on Tuesday 15 April**
Final version: **17.00 on Wednesday 16 April**

Introduction — Determinants of plasma β-carotene levels

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.

The datafile `carotene.xlsx` contains 315 observations on the following 12 variables.

Variable namn	Description
age	Age (years)
sex	Sex (1 = Male, 2 = Female)
smokstat	Smoking status (1 = Never, 2 = Former, 3 = Current Smoker)
bmi	Body mass index, BMI = weight/height^2 (kg/m^2)
vituse	Vitamin use (1 = Yes, fairly often, 2 = Yes, not often, 3 = No)
calories	Calories consumed per day (MJ, 2500 kcal $\approx$ 10 MJ)
fat	Fat consumed per day (g)
fiber	Fiber consumed per day (g)
alcohol	Number of alcoholic drinks consumed per week
cholesterol	Cholesterol consumed per day (mg)
betadiet	Dietary β-carotene consumed per day (mg)
betaplasma	Plasma β-carotene (ng/ml)

Our goal is to model how the plasma β-carotene level, `betaplasma`, varies as a function of one or several of the other variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. We will use a linear regression model $Y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$ where the random errors ε_i are assumed to be pairwise independent and $N(0,\ \sigma^2)$. In order to fulfill these model assumptions we will have to use suitable transformations.

1 Plasma β-carotene and body mass index

We start by modeling how plasma β-carotene depends on body mass index, `bmi`.

Lec.2 1(a). Make a graphical comparison of the residuals for model *Lin* and *Log*, and motivate why version *Log* is more suitable than version *Lin*.

Lin: $\texttt{betaplasma}_i = \beta_0 + \beta_1 \cdot \texttt{bmi}_i + \varepsilon_i$ where $\varepsilon_i \in N(0,\ \sigma^2)$ for $i = 1, \ldots, n$

Log: $\ln(\texttt{betaplasma}_i) = \beta_0 + \beta_1 \cdot \texttt{bmi}_i + \varepsilon_i$ where $\varepsilon_i \in N(0,\ \sigma^2)$ for $i = 1, \ldots, n$

Lec.2 1(b). Use version *Log* (*Model.1(b)*) and present a table with the β-estimates and their 95 % confidence intervals.

Plot log plasma β-carotene against BMI together with this estimated linear relationship, its 95 % confidence interval and a 95 % prediction interval for future observations.

Transform the relationship back to $\texttt{betaplasma} = \ldots$, and plot plasma β-carotene (ng/ml) against BMI together with the estimated relationship, its 95 % confidence interval and a 95 % prediction interval.

Lec.2 1(c). Express how, according to *Model.1(b)*, the expected plasma β-carotene (ng/ml) changes when BMI is

(i) *increased* by 1 unit,

(ii) *decreased* by 1 unit,

(iii) *decreased* by 10 units.

Also calculate 95 % confidence intervals for these change rates.

Lec.4 1(d). Test whether there is a significant ($\alpha = 5\,\%$) linear relationship between log plasma β-carotene and BMI, according to *Model.1(b)*. Report the name of the test you use, the null and alternative hypotheses expressed using the model parameters, the observed value of the test statistic and its distribution when the null hypothesis is true (including the degrees of freedom), the P-value of the test and the conclusion.

2 Plasma β-carotene and smoking habits

We now turn to smoking habits.

Lec.3 2(a). Turn the categorical variable `smokstat` into a factor variable, with suitable category labels, and present a frequency table for it. Add the mean and standard deviation of plasma β-carotene and of log plasma β-carotene for each of the three categories to the table. Also indicate which category would be most suited to use as reference category. Motivate your choice.

Present boxplots, and/or violin-plots, of both plasma β-carotene and log plasma β-carotene against `smokstat`, and motivate whether we should still use log plasma β-carotene as dependant variable.

Lec.3 2(b). Fit two versions of a model where log plasma β-carotene depends on the categorical variable `smokstat`. Use category ”Never” as reference in one version, and ”Current smoker” in the other version. State how you define the different dummy variables and which β-parameter they each are connected to.

Present parameter estimates, with standard errors, for both versions and explain why the standard errors are larger in one version. The version with the more reasonable reference category will be referred to as *Model.2(b)*.

Lec.3 2(c). Use both model versions and calculate the expected log plasma β-carotene with 95 % confidence intervals, in a current smoker, a former smoker, and someone who has never smoked. Also calculate the expected plasma β-carotene levels (ng/ml) with 95 % confidence intervals.

Relate these expected values to the corresponding means in 2(a). Also explain why the predictions and their confidence intervals are the same regardless of which model version you used.

Lec.4 2(d). Perform a suitable test for whether there are significant ($\alpha = 5\,\%$) differences in log plasma β-carotene between any of the `smokstat` categories. Report the name of the test you use, the null and alternative hypothesis, expressed using the model parameters, the observed value of the test statistic and its distribution when the null hypothesis is true (including the degrees of freedom), the P-value of the test and the conclusion.

3 Multiple linear regression

Lec.3 3(a). Turn `sex` and `vituse` into factor variables as well, with suitable category labels, and present frequency tables for them. For each variable, also indicate which category would be most suited to use as reference category. Motivate your choices.

Lec.4 3(b). Calculate pairwise correlations between the continuous x-variables `bmi`, `age`, `calories`, `fat`, `cholesterol`, `fiber`, `alcohol`, and `betadiet`, and present, and plot, the pairs where the correlations are stronger than ± 0.6. Comment on any other potential problems you find with any of the variables, illustrated with a suitable plot.

Is the person consuming the equivalent 200 alcoholic drinks, corresponding to 12 bottles of Absolut Vodka, each week, extreme in any other variables as well? *Hint*: consuming 12 bottles of Vodka per week adds ca 12 MJ per day to the calorie consumption.

Lec.4 3(c). Ignore any potential problems and fit a model where log plasma β-carotene depends on all the other variables, `bmi`, `age`, `calories`, `fat`, `cholesterol`, `fiber`, `alcohol`, `betadiet`, `smokstat`, `sex`, and `vituse`.

Present the VIF/GVIF-values for the variables, and indicate any variables where more than 80 % of the variablility can be explained using the other x-variables.

Remove the most problematic x-variable and refit the model without it (*Model.3(c)*). Present, and comment on, the new VIF/GVIF-values.

Lec.4 3(d). Use *Model 3(c)* and state how you define the different variables and which β-parameter they each are connected to. Present a table with the β-estimates as well as the e^{β}-estimates together with 95 % confidence intervals for e^{β}.

Perform the following tests and present the results in a table containing the name of test you use, the null hypothesis H_0 expressed using the β-parameters, the test statistic, the distribution of the test statistic when H_0 is true, the P-value and the conclusion of the test. Explain why you choose the different types of tests and comment on the result.

(i) Is there a significant relationship between log plasma β-carotene and BMI, given the other variables in the model?

(ii) Is this model significantly better than *Model 1(b)*,which used only `bmi`?

(iii) Is this model significantly better than *Model 2(b)*,which used only `smokstat`?

Lec.5 3(e). Make a visual inspection of the studentized residuals for *Model.3(c)*, looking for outliers, non-constant variance, and non-normality, using suitable plots with suitable reference lines. Comment on the results.

Lec.5 3(f). Calculate the leverage for *Model.3(c)* and plot them against the linear predictor, with suitable reference lines. Identify the observation with the largest leverage and determine why it has such a large leverage, using suitable plots.

Lec.5 3(g). Calculate Cook's distance for *Model.3(c)* and plot against the linear predictor, using suitable reference lines. Identify the observation with the largest Cook's distance, calculate the DFBETAS and identify the β-parameters that have been affected. Illustrate by plotting the relevant DFBETAS against the corresponding x-variable, with suitable reference lines.

Explain why the observation has had a large influence on the estimate, with the help of suitable plots.

Investigate the observation with the largest leverage, identified in 3(f), and determine whether it has had any alarming influence on the estimates.

4 Removing the influential observation

Lec.5 4(a). Create a new dataset where the influential observation identified in 3(g) has been removed and estimate a new version of *Model.3(c)* on this reduced data set.
Hint: `newmodel <- update(oldmodel, data = newdata)`

Calculate and plot Cook's distance against the linear predictor, with suitable reference lines, for *Model.3(c)* on the reduced data set and compare with the corresponding plot for the full data set in 3(g). Comment on any interesting differences.

Recalculate the pairwise correlations between the x-variables as well as the VIF/GVIF values for *Model.3(c)* on the reduced data set and comment on any interesting differences or similarities with the full data set in 3(b) and 3(c).

Lec.6 4(b). For both data sets, full and reduced, perform a stepwise variable selection, starting with the null model, using the null model as the smallest model allowed and the full model as the largest model allowed, with BIC as criterion. Report the order in which the variables are included/excluded in both data sets, and present the final β-estimates, with 95 % confidence intervals. The models will be referred to as *Model.4(b)*.

Reflect on any interesting differences in which variables where included, and why the difference might be reasonable, considering your earlier finds.

5 Fine-tuning the model

Lec.6 5(a). For the reduced data set, determine whether it is necessary to have three categories in the categorical x-variable(s) in *Model.4(b)*. Perform a suitable test and update the variable(s) and model as necessary. Present the new β-estimates, with 95 % confidence intervals. Call the new model version *Model.5(a)*.

Lec.6 5(b). Refit *Model.1(b)* and *Model.2(b)* on the reduced data set as well.

For the reduced data set, present a table with the number of β-parameters, the residual standard deviation, the R^2, adjusted R^2, AIC and BIC for the all five models (*1(b)*, *2(b)*, *3(c)*, *4(b)*, and *5(a)*). State which model you find best, and the reasons for your choice.

End of Project 1