# Assignment

For questions 1 to 4, short and concise, but correct, answers is required (preferably in red font). Statistical software is needed to solve question number 5 and 6. The answer should be complemented with a commented syntax script which can be run from the first to the last line without halting, including code to read the data from the file system of your computer, code to create new variables, and code for the relevant statistical analyses.

The datasets for the last two questions is attached (CTC.sav, fracture\_bmi.sav)

Reference to the syntax or the output produced by running the code is not sufficient. Copy the relevant output from the statistics package and comment the results in this document.

**The syntax files should not be converted to PDF/Word.**

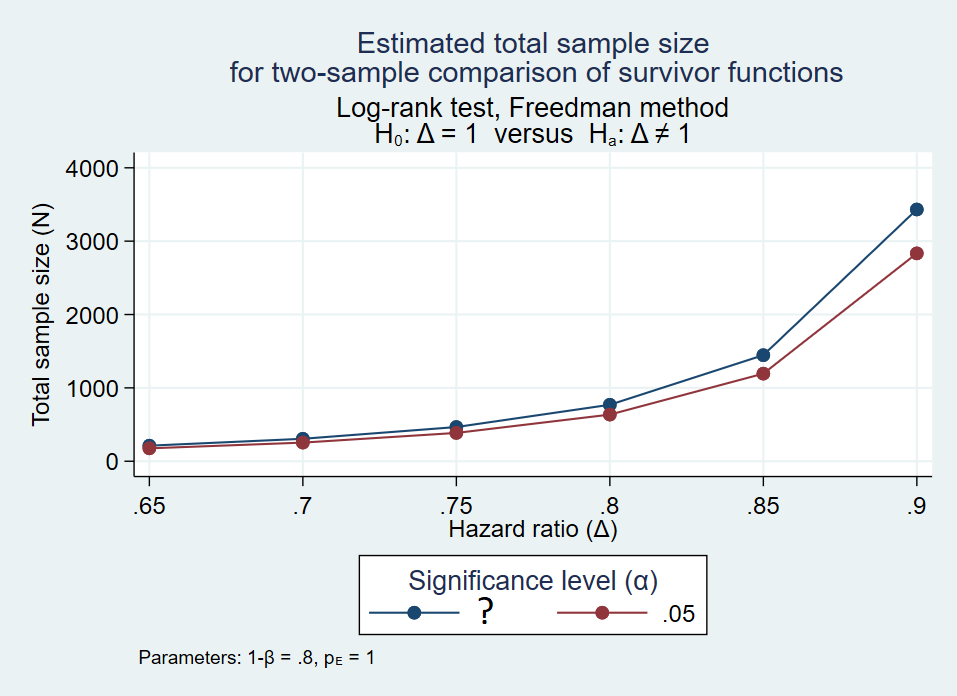
Write your full name both inside the document and in the document name. Use “Exam Nov 2022 + your name” as subject of the e-mail.

Ask for a delivery confirmation in your e-mail program if you want to make sure that your   
e-mail has arrived. The final results will be reported, together with individual feedback, no later than three weeks after the deadline.

1. **Study design and statistical power (2.5p)**

In the planning of a phase III randomized clinical trial for comparing progression-free survival after a new treatment A compared to the conventional treatment B, the investigators made a simple power calculation (see figure below) to guide the decision on how many patients to include in the study to have 80% power.

1. The red line shows the number of patients needed to detect different hazard ratios (HR) with a significance level (alpha) of 0.05. The label for the blue line is missing. What can you say about the significance level? Motivate briefly. (1p)
2. According to the figure, how many patients do we need to include if we want to be able to detect a 25% lower progression rate among patients treated with the new treatment compared to the old treatment? Assume no drop out and a significance level of 0.05. (0.5p)
3. When analyzing the data from the trial, the p-value for the estimated HR was 0.0003 which is much lower than the significance level used when calculating the required sample size for this trial (α = 0.05). Name at least one possible explanation to this low P-value? (1p)



1. **Simple logistic regression (3.5p)**

Age at cardiac arrest is a risk factor for death within 30 days after successful cardiopulmonary rescue. A researcher investigates if this hypothesis holds also in her own cohort study with patients alive and with preserved cognitive function 24 hours after the arrest. Since the outcome is binary she chooses logistic regression analysis.

1. Would it have been wrong to use Cox regression if both the event (dead within 30 days, yes/no) and the time to event (or last follow-up for event-free) had been recorded in the database? Motivate your answer briefly. (0.5p)

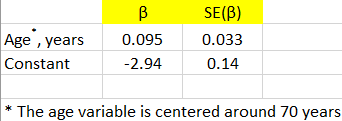
She defined the logistic regression model this way:

log(p/(1-p)) = β0 + β1\*Agec

where p = probability of death within 30 days

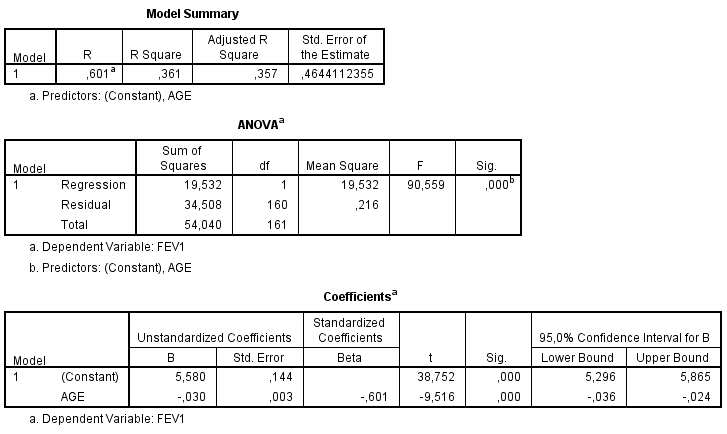
and Agec = Age in years minus 70 years

The results of her analysis is presented in the table below:



1. Explain why it might be a good idea to center the age variable. (0.5p)
2. Interpret the results of the logistic regression analysis. (1.5p)
3. Is the effect of age on the outcome (death within 30 days) significant at the 5% level? Why/why not? (1p)
4. **Correlation and linear regression (3p)**

The relation between age and lung function (FEV1) was studied in 160 male non-smokers, and analyzed using simple linear regression:



a) Write down the equation of the fitted regression line and interpret the coefficients of the regression model. (1p)

b) What is the interpretation of R-squared=0.36? (0.5p)

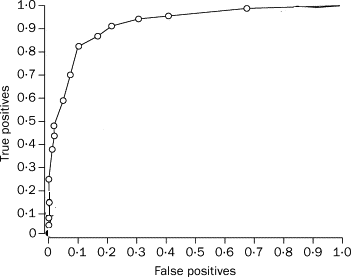
c) What is the estimated lung function for a non-smoking man who is 55 years old? (1p)

d) What is the Pearson correlation for age and lung function? (0.5p)

1. **Diagnostic tests (4p)**

Acute upper-gastrointestinal hemorrhage is a common reason for emergency hospital admission and it is important to be able to predict which patients will require clinical intervention. A research group therefore developed a risk score from a logistic regression model to assess whether patients presenting with acute upper-gastrointestinal bleeding will require admission for treatment to manage their bleeding.

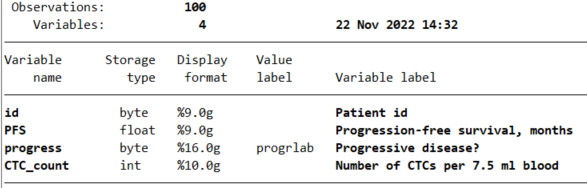
The validity of the scoring system was assessed by plotting a receiver operating characteristic (ROC) curve in a validation set of 197 patients.



| **Predicted need of treatment** | **Eventual need of treatment** | | |
| --- | --- | --- | --- |
|  | **No** | **Yes** | **Total** |
| No | 43 | 1 | 44 |
| Yes | 65 | 88 | 153 |
| Total | 108 | 89 | 197 |

1. Argue, from the ROC-curve above, if the overall ability of the score to predict need of treatment is better than random? (0.5p)
2. Calculate the sensitivity and specificity from the table, and identify the chosen cut-off in the ROC-graph above. (1p)
3. Calculate the positive predictive value for the same cutoff and interpret it (in the context of predicting need of treatment). (1p)
4. What would be the effect on PPV if the same score was used in a population with lower prevalence of severe acute upper-gastrointestinal bleeding? Higher PPV, lower PPV or no difference (on average)? A brief motivation is required. (1.5p)
5. **Survival analysis (5.5p)**

Circulating tumor cells in blood is a relatively new biomarker which can be used to monitor metastatic disease in for example breast cancer. In the data file CTC you find information on number of CTCs per 7.5 ml blood for 100 patients diagnosed with metastatic breast cancer. These 100 patients are a random sample from a larger prospective cohort study. The counts and the progression status (yes=1/no=0) are real data, but a little random noise have been added to the follow-up in months. The data set has the following variables:



1. Describe the distribution of CTC-count separately for patients with and without progress. (1p)
2. Is the evidence for a difference between the two CTC-distributions strong? (0.5p)
3. Categorize CTC\_count into three groups and draw a Kaplan-Meier graph of progression-free survival, one curve per CTC-group. (1p)
4. Is the evidence for a dose-response relationship strong? (1p)
5. Finally, fit one Cox-regression with CTC\_count as a covariate (no transformation) and one with CTC-count as a factor on three levels and discuss the estimates. Which model do you prefer and why? (2p)
6. **Multiple logistic regression (8.5p)**

Low BMI has been shown to increase fracture risk, possibly because low BMI is associated with low bone mineral density (BMD), less soft tissue, and muscle weakness. To investigate this further you have here received a dataset *bmi\_fracture* that includes 1690 individuals and the following variables:

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Values |
| id | Unique identifier | 1-1690 |
| fracture | Fracture | 0 = No  1 = Yes |
| bmi | Body mass index, BMI (kg/m2) | 14-38 |
| bmi\_cat | BMI, categorized | 0 = underweight (BMI<20)  1 = normal weight (BMI 20-24)  2 = overweight (BMI≥25) |
| sex | Gender | 1 = male  2 = female |
| age | Age in years | 33-90 |

1. Use summary measures and/or graphics to explore the data. Comment in particular on the relation between BMI, sex and fracture. (1p)
2. The focus here is on the effect of BMI and the risk of fracture. Fit a logistic regression model with fracture as the outcome and BMI as a linear predictor. Adjust the model for age and sex. (1p)
3. Report the adjusted odds ratio for BMI with a 95% CI and interpret it in your own words. (2p)
4. To account for a possible non-linear effect of BMI on fracture, include BMI as a factor on 3 levels (underweight, normal weight and overweight) instead of as a linear covariate in the model. Chose an appropriate reference level and present and interpret the result. (1p)
5. Which of the two models do you prefer? Motivate briefly. (1p)

In a previously published meta-analysis, higher BMIs lowered the fracture risk in women but not in men which suggests that BMI affects fracture risk differently in men and women.

1. Create a new variable “underweight” that is 1 if BMI<20 and 0 otherwise. Report the proportion of men that are underweight (0.5p)
2. Now, fit a logistic regression model to test the hypothesis that the effect of being underweight on the odds of fracture is the same in men and woman. What do you conclude? (2p)