Assignment 1 - Statistics for molecular biomedicine

March 25, 2021

- Please read the assignment instructions in the assignment module on Absalon.
- In this assignment only use base R - dont use other packages i.e. dont use library().
- In the following questions use a significance threshold of 0.05 unless otherwise stated.
- This is an individual assignment. You cannot share code and answers with someone else.
- Use the discussion board for question about the assignment - however DO NOT post your suggested solution (see pinned post in the discussion board)

Part 1

In a large population based Danish cohort 5764 individuals were diagnosis tested for type 2 diabetes (T2D) in a lab. For some of the individuals the lab did this using a method based on a so-called oral glucose tolerance test (OGTT) in which fasting plasma glucose (pglu0) or 2-hour glucose after sugar intake is measured. For the rest of the individuals the lab did it using a method based on so-called glycated hemoglobin (HbA1c), which is a measure for long term glucose levels. The lab used the WHO criteria for T2D (definition for the curious). In addition to performing the T2D diagnosis test the lab also measured the individuals' BMI, age and sex. The data from the lab is found in the file T2D.txt. For T2D the data is coded as 1 for T2D (is diagnosed to have T2D) and 0 for 'NOT T2D' (is diagnosed NOT to have T2D), while sex is coded as 1 for males and 2 for females. The OGGT column indicates which diagnosis method was used with '1' for OGTT and '0' for Hb1Ac.

We are interesting in knowing which factors affect developing T2D and also whether the choice of T2D diagnosis method (OGTT or HbA1c) has an impact on the diagnosis. Below T2D refers to being diagnosed with type 2 diabetes.

1A First we want to test whether the two diagnosis methods are equally likely to give the T2D (diagnosis of type 2 diabetes). Perform an appropriate test that tests if there is a difference in diagnosis outcome from the two different diagnosis methods (OGTT or HbA1c). Report the test, the p-value and the conclusion.

1B Report the odds ratio (OR) and the relative risk of getting T2D from an OGTT compared to HbA1c.

1C We want to explore which factors affect T2D. Fit an appropriate model that, in addition to the type of diagnosis method, includes age, sex and BMI. Based on this model, which includes the four factors, what is the odds ratio of getting a T2D from an OGTT compared to HbA1c?

1D Using the fitted model perform tests that determine which factors affect T2D. Report the p-values for each test and a conclusion.

1E Provide the effect sizes in OR with 95%confidence intervals for each of the four factors

1F Based on the model, what is that probability that a 36 old male individual with a BMI of 24.5 is diagnosed with T2D using a OGTT?

1G A 36 year old female was tested using a OGTT and was predicted to have the same probability. What was her BMI?

1H BMI is often grouped into normal weight (< 25), overweight ($20 - 30$) and obese (> 30). Instead of BMI as a quantitative trait use BMI groups in the model. Perform a test to determine whether the BMI groups have an effect on T2D. Report the p-value and the conclusion

1I It is known that being obese increases the risk of T2D (compared to normal weight) much more than being overweight. It has been stated that the effect of being obese is 4 times bigger than the effect of being overweight. Here effect refers to the regression coefficients (β) and not the OR. Perform a test to see if you can reject this hypothesis based on this data. Report the p-value and your conclusion.

Part 2

COVID-19 has affected most countries in the world. Mexico is a country who has registered and published detailed data for each case of a positive COVID-19 test. You can find the data here `https://www.gob.mx/salud/documentos/datos-abiertos-152127`. I downloaded the data from June first 2020, I randomly sampled 1000 individual from each of 8 age groups. I cleaned aand recoded the data for easy use. The data is found in the file `corona.txt` which contains the columns

firstSymptomDay the day where the first symptom and/or positive test. The day refers to the day of 2020 e.g. day 1 is january 1st 2020.

eventDay the day where the person died from corona or the data where the data was collected (June 1st)

event Status of the individual on June 1st

- **deadCorona** died from corona
- **deadUnrelated** died from unrelated causes
- **lost** lost to followup (individual not be tracked after positive test)
- **alive** alive june 1st

ageGroup the age group for each individual in 8 different age bins

DIABETES Diabetes status for each individual (0 no diabetes, 1 diabetes)

In this assignment we are interested in the survival of individuals who have tested postive. We will use a significant threshold of $\alpha = 0.01$.

2A How many individuals in the study were censored?

2B Make a plot the cumulative survival rates

2C Since we only follow the individuals for a finite amount of time we can only calculate the mean survival in a restricted time period. Calculate the mean restricted survival and the mean survival of the non-censored individuals. Restrict the period of time to the longest survival time for the non-censored individuals.

2D Estimate the probability of surviving the first 14 days after first symptoms. Include the 99% CI in the answer.

2E If you survived the first 14 days what is the probability of surviving the next 14 days?

2F Does having diabetes affect the survival? Include the test name, a p-value and a conclusion in the answer.

2G Age has a big effect on survival rates with the older being at most risk. Which ages group have lower risk compared to the group aged>80.

Part 3

Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. A womanś behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight. The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. Four variables which were thought to be of importance were age, race, smoking and history of hypertension. The data set 'birthWeight.txt' contains the following columns

low indicator of birth weight less than 2.5 kg.

age mothers age in years.

race mother's race ('1' = white, '2' = African American, '3' = other).

smoke smoking status during pregnancy. (0 no, 1 = yes)

ptl number of previous premature labours.

ht history of hypertension (0 no, 1 yes)

3A First we want to test whether it is true that smoking affects the risk of low birth weight. Perform an appropriate test and report choice of test, the p-value and the conclusion.

3B Report the odds ratio (OR) and the relative risk for low birth weight in smokers.

3C We want to explore which factors affect low birth weight. Fit an appropriate model that, in addition to smoking, includes age, race and hypertension status. Report the odds ratio (OR) for low birth weight in smokers with 95% CI

3D Based on the fitted model, what is that probability that a white 36 old smoker with no history of hypertension gives birth to an underweight individual?

3E (Based on same model) if a smoking white individual with a history of hypertension was predicted to have a risk of low birth weight of 35%. What would be her age?

3F Using the fitted model perform tests that determine which of the risk factors; age, hypertension, race and smoking affects low birth weight. Report the p-values for each test and a conclusion.

3G It is known that the risk of low birth is different between individuals who identify as African Americans and individuals who identify as white. Test if there is a difference in low birth risk between African American individuals and individuals in the category 'other'. Answer the question while still taking the other risk factors into account. Report the p-value and the conclusion