Lab 2 - Statistics and Data Analysis II

November 11, 2024

To be handed in (Lisam) no later than November 22. The submission should include code, relevant output, as well as answers to questions. Preferably, all combined in a ".pdf" file. We recommend the use of RMarkdown to create the report.

You have access to a data set containing survey data from Swedish schools during 2014 (`hsbc_lab2.csv`). The data extract you will use for this assignment includes children's body mass index (BMI=mass (kg)/height2 (meters)), their life satisfaction and their gender.

1) How is children's BMI (`BMI`) associated with life satisfaction (`lifesat`)? There is not an obvious direction of this correlation. For the sake of this exercise, think of $X =$ `BMI` and $Y =$ `lifesat`. Fit a regression $Y = \beta X + \epsilon$ (that is, do not include any control variables) and interpret the result. In particular interpret the p-value of the estimated β-coefficient.

2) Include a squared term of BMI. What does the result suggest about the relationship between BMI and life satisfaction?

3) Drop the squared term from the model and instead add gender (`sex`), as well as an interaction between gender and BMI. So the model now reads $Y = \beta_1 X + \beta_2 Z + \beta_3 (XZ) + \epsilon$, where Z represents the gender variable. How do you interpret the results? Which gender's life satisfaction is more affected by BMI? Boys or girls?

4) Based on your model from #3, predict the life satisfaction for two hypothetical girls with a BMI of 20 and 30, respectively. Store your predictions in a `data.frame`. Hint: to do this, you may follow these three steps:

   i. Create a `data.frame` containing two observations (and two columns) as specified. See table below.
   ii. Use the `predict()` function with two inputs: `object=` your model from #3, and `newdata=` the data.frame you created in step i.
   iii. Store the predictions as a third column in the data.frame you created in step i.

sex	BMI
Girl	20
Girl	30

5) Using the same steps as in #4, predict the life satisfaction for two hypothetical boys with a BMI of 20 and 30, respectively.

sex	BMI
Boy	20
Boy	30

6) Create a plot of the predictions made in 4) and 5). Interpret this plot: what does it reveal about how BMI affects life satisfaction for Boys and for Girls? To create the plot, you may follow these two steps:

   i. Create a new `data.frame` which combines the data.frames from 4) and 5) in a row-wise manner, i.e., by using the `rbind()` function.
   ii. Create a line-plot using `ggplot` (hint: `geom_line()` for a line-plot), with `BMI` on the x-axis, the *predictions* on the y-axis, and *colored* by `sex`.

7) To assess how well the assumptions of the OLS model are met, please do the following:

   a. Assess the assumption of *homoscedasticity* (constant variance of residuals).
   b. Assess the assumption of *normality of errors*.
   c. Discuss the assumption of *no omitted variables* which are correlated with both X and Y.
   d. Consider the information provided in the beginning of the document, and discuss the *appropriateness of the data collection*. Which population can we make inference about from these analyses?