STAT 210
Applied Statistics and Data Analysis:
Homework 6

Due on November 09/2025

You cannot use artificial intelligence tools to solve this homework.

Show complete solutions to get full credit. Writing code is not enough to answer a question. Your comments are more important than the code. Do not write comments in chunks. Label your graphs appropriately

For all tests in this HW use a significance level of $\alpha = 0.02$.

Question 1 (50 pts)

The data for this question is stored in the file `CHFLS` in the library `HSAUR3` and comes from a survey of 60 villages and urban neighborhoods in China published in 2003. It has 1534 observations of 10 variables, but we will focus on `R_age` (age), `R_happy` (self-reportet happiness), and `R_region` (region).

(a) Creat a new data frame called `df1` that only includes `R_age`, `R_happy` and `R_region`. Check whether the new data frame has missing data. Explore the distribution of `R_age` for the different regions. Do boxplots of age as a function of region and comment on what you observe. Calculate mean, standard deviation, median and interquartile range for `R_age` for each of the six regions and comment.

(b) Create a new ordered factor in `df1` named `R_a` by dividing `R_age` into five groups having approximately the same number of subjects. Name the levels `a1, a2, a3, a4,` and `a5`.

(c) Produce a table of `R_region` against `R_a` (age should be in the columns of the table). Graph a mosaic plot of this table using different colors for the rectangles. Comment on what you observe. Produce a second table with proportions relative to the regions. Comment.

(d) You want to determine whether the age groups have a homogeneous distribution across regions. Which test (or tests) do you know that can be used for this? What are the underlying assumptions? Are they satisfied in this case? Carry out all the tests that apply and discuss the results. What are your conclusions?

(e) To explore the relation between age and happiness, build a contingency table for `R_happy` against `R_a`. Use the Chi-square test on this table. Are the conditions for the test satisfied? Why or why not?

(f) Create a new ordered factor called `R_h` in `dt1` by joining the two lower levels of `R_happy`, i.e., `R_h` will have three level named `unhappy`, `Somewhat happy`, and `very happy`. The values for `unhappy` come from re-naming the levels `Very unhappy` and `Not too happy` as `unhappy`. One easy way to do this is to use the `labels` argument. Look at the help of the `factor` function to see how this is done.

(g) Build a contingency table for `R_h` against `R_a`. Graph a mosaic plot of this table using different colors for the rectangles. Comment on what you observe. Use the Chi-square test on this table. Are the conditions for the test satisfied? Why or why not? What is your conclusion after using the test?

Question 2 (50 pts)

For this problem use the data set `women` which is available in `R`, and has 'average heights and weights for American women aged 30-39', according to the help file for the data set. Height is measured in inches while weight is measured in pounds.

(a) Fit a simple linear regression model for `weight` as a function of `height`. Produce a scatterplot and add the regression line.

(b) Print a summary table for the model and interpret the results. Write an equation for the model and interpret the coefficients. What is the R^2 for this model?

(c) Predict the weight for a woman of 65 inches including a confidence interval.

(d) State explicitly the assumptions on which the model is based and using plots and tests verify whether they are satisfied.

(e) Use the function `residualPlots` in the `car` library. The argument of the function is the name of your model and the output is a couple of residual plots plus some summary information about two tests. The first plot is residuals vs. the regressor (`height` in this case) and the second is residuals against fitted values. In both cases, the blue line shown in the plot is not a local smoother but a quadratic term added to the model. If the line is flat, it indicates that the term will not improve the model. The results of tests of curvature are also shown. We will consider only the first one, corresponding to `height`. This is a test on the coefficient for a quadratic term in `height` added to the model. The null hypothesis is that coefficient corresponding to the quadratic term is zero. Interpret the output you get when using this function on your model.

(f) Add a quadratic term in `height` to your model (you have to use the expression `I(height^2)` in the equation to do this). Print the summary table and interpret the results. What is the R^2 for this model and how does it compare with the previous model?

(g) Plot the data and add the lines corresponding to the two model you fitted.

(h) Write down an equation for the final model. Predict the weight for a woman of 65 inches including a confidence interval using the quadratic model and compare with your previous prediction.