Mini-Project: Scatterplots

"Sixty minutes of thinking of any kind is bound to lead to confusion and unhappiness."
—JAMES THURBER

Overview
The purpose of this project is to practice making scatter plots and regression equations with data software. You will use R to make a total of three scatterplots, each with a graph of a regression equation.

Plot 1: from the `hanes` data, a plot of weight, based on height, and a regression line with an indicator variable for male or female.
Plot 2: from the `hanes` data, a plot of weight, based on height, and a regression line with an indicator variable for male or female, plus one interaction term for gender:height.
Plot 3: on your own, find population data for at least ten years for a city, region, state, or country of your choice at statistista.com. Find two different regression polynomials and plot graphs of both.

For each plot, you will include the regression equation(s) and the coefficient(s) of determination.

Your primary references are Dr. Whalen's notes *Lesson 05* and *Lesson 06*. You are also expected to find population data from statista.com. Any other or additional references (websites, forums, books, etc.) must be listed in your submission (including what information you got from the reference). Also, the name of anyone that you worked with on this project must be listed in the references.

Project Requirements

1. The regression line plus indicator scatter plot.
    a. From the `hanes` data, make a scatter plot of weight (y) based on height (x).
    b. Distinguish the men and the women in the scatter plot with two different symbols and colors. Please use entries of the form `pch=ifelse(…)` and `col=ifelse(…)` in the plot function to make this distinction. This requires you to figure out how to use the `ifelse` function combined with data sub-setting code that you already know how to do.
    c. Find the regression equation of weight based on height plus a gender indicator. Write down (type out) the equation.
    d. Add a graph of the equation to the scatter plot (it will be two parallel lines). Hint: find the equations of the two lines by substituting 1 or 0 for the indicator. Then use `abline` twice, same slope, but different intercept.
    e. Paste the plot into your document editor (Word). With math typesetting, provide the regression equation and the coefficient of determination below the plot.

2. The regression line plus indicator plus interaction.
    a. Repeat part 1, but this time add the interaction term gender:height to the regression equation. Write down (type out) the equation.
    b. Add a graph of the equation to the scatter plot (it will be two lines, but not parallel). Hint: find the equations of the two lines by substituting 1 or 0 for the indicator and collecting like terms. Then use `abline` twice, different slopes, different intercepts.
    c. Which line is for which gender? Distinguish between them using `lty` and/or `col` within the `abline` command.
    d. Paste the plot into your document editor (Word). With math typesetting, provide the regression equation and the coefficient of determination below the plot.

3. Population polynomial
    a. At statista.com, find population data by year for a city, region, state, or country of your choice. Ideally, find one with an interesting trend, where a polynomial rather than a line would be helpful. There must be at least 10 years recorded, up to a recent year (say, 2021).
    b. Store the data in R. The simplest way is to store two vectors (lists of numbers combined with `c(…)`), one for the years, and another for the population values. The years can be the absolute year numbers or a counting number starting at 0. Alternatively, you could look into storing it as a `data.frame`. Or, you could type (or download) the data into an Excel file and import the Excel file.
    c. Make a plot of the population based on the year.
    d. You **must** include an x-axis label, something like "Year" or "Years since _____" with the blank filled in with the starting year.
    e. You **must** include a y-axis label. This may be "Population" or "Population (thousands)" or "Population (millions)" and so on.
    f. You must include a `main` title that says what place in the world the data represents.
    g. Following the `Galileo` example in Lesson 06, find the least-squares polynomial of degree two (or higher). Type out the equation.
    h. Use `curve` to add the polynomial to your population scatter plot.
    i. Find the least-squares polynomial with one more degree (so, if you found a degree 2 polynomial, now find a degree 3 polynomial). Type out the equation.
    j. Add this polynomial to the scatter plot using a different `col` and `lty`.
    k. Paste the plot (with the population dots and graphs of two polynomials) into your document editor (Word). With math typesetting, provide the equations of the two polynomials below the plot, along with the coefficients of determination for each.

Project Submission

1. Page 1—put the two `hanes` plots here, with the equations and coefficients of determination below each plot. Recall that images copied from R may not display with the default paste—in Microsoft Word, choose Paste > Picture (from the top-left menu area under the Home tab).
2. Page 2—put the population plot here, with the equations and coefficients of determination of the two polynomials below each plot.
3. Page 3—list the names (full elearning name) of *anyone* that you worked with and what sources, if any, you used (except for statista.com, Dr. Whalen's notes, and the textbooks mentioned in the syllabus).
4. Pages 4 and beyond—copy and paste your final R script. Make sure that the grader can copy your script, run it line by line, and recreate your work.
5. Save the document as a PDF.
6. Upload to the Mini-Project Scatterplot Submission assignment in elearning.

Warnings

- You are expected to submit your own individual work.
- Working together is great, but you must list anyone you worked with (by full elearning name) and be careful not simply to copy someone else's code and work.
- Beware of using online sources. If you do use any, include them by name and web address in your source list and mention what you got from it.
- Please do not use any R graphics packages (such as `ggplot`) for this project. Same comment as before: Plots in R itself have a higher "data-to-ink" ratio. Plots from `ggplot` or other packages can get too "busy." Of course, you're welcome to experiment with packages on your own.)