Part 1: Analysis with Galton's original data set

Galton's work on children and parents' height was published in: Galton, F. (1886): "Regression towards mediocrity in hereditary stature", *Journal of the Anthropological Institute*, 15: 246-63. In this first part of the project you are asked to reconstruct the original data from this original article and replicate his analysis.

- **Question 1.1.** Find Galton's original article (you can use www.jstor.org). You can also find it on LEARN. On Table I of his article, the data used is summarized. You need to create a STATA data set that contains the 928 observations that Galton collected. It is recommended that you first type the data in an excel file and then have STATA read that file. Some versions of the Galton data set are available online. You are advised NOT to use them. It is part of this project that you show that you understand how to make a data set from such a table. There are important conceptual issues that you will miss if you borrow the data from somewhere else.

  (i) For those observations reported in Table I of Galton's article as "below" or "above" the minimum and maximum height values, you need to assume some particular values. Please state these explicitly in a table (Table 1.1.a.) and provide a justification with one sentence.
  (ii) Given your assumptions, what is the sample mean height and standard deviation for adult children and for parents, respectively? Report this in a table (Table 1.1.b.).

- **Question 1.2.** For the rest of part 1, assume that there are 928 parents in the sample rather than 205. Define "tall parents" and "short parents". Then divide your sample into two corresponding groups.

  (i) Are children of "tall parents" as tall as their parents? And similarly, are children of "short parents" as short as their parents? Report your results in a table.
  (ii) Does the assumption of having 928 parents rather than 205 matter for this exercise?

- **Question 1.3.** Galton was the first to describe and explain the phenomenon of "regression towards the mean". Being concerned about the height of the English aristocracy, he interpreted his results as "regression to mediocrity" (hence the name "regression").

  (i) Regress the height of adult children against the height of parents. Report your results in a table and interpret the estimated coefficients.
  (ii) What can you say about the relationship between the height of parents and their children? How does it relate to the findings in question 1.2.? You can answer these questions with a short paragraph and a graph.

- **Question 1.4.** Now regress the height of parents against the height of adult children. Report your results in a table. Explain in a short paragraph whether this regression is equivalent to the one in question 1.3.

- **Question 1.5.** Taking your regression results from question 1.3., and using your definition of "tall parents" and "short parents" from question 1.2:

  (i) Calculate the predicted adult children's height whose parents are "tall" after 1, 2, 3, ..., Z generations? And similarly, what is your prediction for adult children's height whose parents are "short" after 1, 2, 3, ..., Z generations? Report your results in a table. Is there convergence in heights? If so, how many generations does it take?

  (ii) How do you interpret the results? Did Galton do something wrong in his regression? You can answer this question with a short paragraph.

Template Answer Sheet

GROUP PROJECT, ESSENTIALS OF ECONOMETRICS

Group number: XX
Group members (student numbers only): s123, s345, etc.

Declaration:

Group XX, composed by students s123, s345, etc., confirms that the data collection has been conducted under the ethical guidelines of the School of Economics (www.ed.ac.uk/economics/research/ethics). The data collection has only involved individuals 18 years old and over. The information collected is not sensitive in any way that can harm the well-being and dignity of the subjects interviewed. To maintain confidentiality, no names or any contact details have been collected. All interviewed subjects have been told that this is part of the EofE course project.

Recall: each question has to be answered in one page maximum, font 12, double spaced.

- **Question 1.1.** Table 1.1.a. summarizes the height values assumed for cases below/above the minimum/maximum height. We have assumed these values because ....

Table 1.1.a. Heights below/above the minimum/maximum height

	assumed height
Adult children	
Heights below the minimum	xyz
Heights above the maximum	xyz
Mid-parents	
Heights below the minimum	xyz
Heights above the maximum	xyz

Reference: Table I in Galton (1886).

Table 1.1.b reports the mean and standard deviation for adult children and parents.

Table 1.1.b. Summary Statistics

	Mean	Standard Deviation	Number obs.
Adult Children	xyz	xyz	xyz
Mid parents	xyz	xyz	xyz

Source: Galton's data with authors assumptions.

- **Question 1.2.** We have defined tall and short as... (i) Table 1.2. reports .... (ii) The assumption of having 928 parents rather than 205...
- **Question 1.3.** Table 1.3 reports...
- **Question 1.4.** Table 1.4 reports...
- **Question 1.5.** Table 1.5 reports...

- **Question 2.1.** Description of own data set: population of interest, data collection process, survey ... Table 2.1 reports ....
- **Question 2.2.** Table 2.2. reports... Figure 2.2. shows...
- **Question 2.3.** Table 2.3. reports...
- **Question 2.4.** Table 2.4. reports...
- **Question 2.5.** Another literature in Economics that has analysed regression towards the mean is....

Template Appendix: log file resulting running from EEproject.do

---

```
clear
capture log close
set more 1
log using project.log, replace

/*question 1.1*/
use namedatagalton.dta
.  sum X
Variable | Obs    Mean    Std.  Dev.      Min    Max
---------+--------------------------------------------

       X |    n    xyz   ....

/*question 1.2*/
etc..

log close
```