STAT0023 Computing for Practical Statistics

In-course assessment 2, take-home component (2023–24 session)

Table of Contents

Rubric

- This assessment is classified as **Coursework**, as defined in the UCL Student Regulations for Exams and Assessments. It contributes 37.5% to the overall mark for this module.
- You may work either alone or in pairs. It is up to you to form your own pairs. **You MUST register your choices via the 'ICA2: with whom will you work?' link on Moodle by 17.00 on Wednesday 20th March 2024, even if you choose to work alone**.
  - If you choose to work in a pair, you will be jointly responsible for the work that is submitted and you will be awarded the same mark.
- The assessment will be released between 13:00 and 14.00 (UK time) on Monday, 18th March 2024.
- The submission deadline is **12:00 (UK time) on Monday, 22 April 2024**.
  - If you are working in a pair, **both members must confirm your submission on Moodle before the deadline**.
- Extensions to the submission deadline can only be granted where a student has been is issued with a SoRA or has made a valid claim for extenuating circumstances. The standard extension length for this assessment type is one week.
  - **If you have a SoRA**, the module organisers will contact you via email to ask whether you wish to make use of your extension. If you choose to do so and you are working in a pair, the extension will also be applied to your partner. **If you have not heard from the organisers by 14.00 on Monday 18th March 2024, please contact them directly.**
  - **Extenuating circumstances** are handled by your parent department and all claims should be submitted via Portico. Depending on the nature and severity of the circumstances, an alternative type of mitigation to a deadline extension may be considered more suitable.

    If you are working in a pair, you should consult with your partner before submitting a claim for extenuating circumstances. When submitting your claim, you should also make clear that the assessment is groupwork.
- Please ensure that you are familiar with the Department of Statistical Science's guidance on academic integrity. When submitting your work, you will be required to make a declaration that you have read and understood this guidance.
  - If you choose to work in a pair, both of you should check what has been submitted before signing the declaration. **If any academic misconduct is identified, you will share responsibility for it** (i.e. the same penalties will be applied to both of you) unless it can be established clearly that one pair member deliberately misled the other.
- Parts of your submission may be scanned using similarity detection software. If any breach of the assessment regulations is suspected, it will be investigated in accordance with UCL's Student Academic Misconduct Procedure.

- To facilitate anonymous marking, you should not write your name anywhere on your work, including in file names or file descriptions requested as part of the submission process.
- You must only submit your work via the designated portal in Moodle. If you try to submit via email or any other channel this will not count as a submission and will not be marked.
- There are strict, non-negotiable penalties for late submission, which for coursework are as follows.
  - Up to 2 working days late: deduction of 10 percentage points, but no lower than the pass mark.
  - 2-5 working days late: capped at the pass mark.
  - More than 5 working days late: mark of 1.00%.
- There are also penalties for over-length submissions. Full details are provided in Section 2.4.
- If the organisers become aware of a significant technical issue or outage affecting Moodle during the assessment, a message will be circulated to explain what has happened and the steps being taken to mitigate the issue. If you do not receive notification of a more widespread issue and you experience technical difficulties, you should refer to the Help & Support resources provided by UCL's central IT service. However, last-minute technical issues will not be considered as valid grounds for missing the deadline, so ensure that you leave plenty of time to prepare, upload and check your submission.
- Non-submission (in the absence of any valid extenuating circumstances) will mean that your mark for this component is recorded as 0.00% and you will be deemed to have made an attempt.
- You should expect to receive feedback on this assessment within one calendar month of the submission deadline. In the event of a delay, the organisers will contact students directly with details of the revised timeline.

---

1 Background and overview

Anemia is a condition in which the oxygen-carrying capacity of blood is reduced relative to the body's requirements. These requirements depend on factors including age, altitude, and pregnancy status. According to a Fact Sheet from the World Health Organisation (WHO), it is a "major public health concern, mainly affecting young children, pregnant and postpartum women, and menstruating adolescent girls and women". Globally it affects 40% of children aged 6–59 months, 30% of women between the ages of 15 and 49 years and 37% of pregnant women.

The oxygen-carrying capacity of blood itself is determined by the concentration of haemoglobin (https://www.webmd.com/a-to-z-guides/understanding-anemia-basics): if haemoglobin levels are too low then the body will not get enough oxygen. There are many forms of anemia, although the most common is caused by iron deficiency. Women of childbearing age are particularly susceptible to anemia, because of the blood loss from menstruation and increased blood supply demands during pregnancy (see the URL just cited). In principle, iron deficiency can be controlled to some extent by diet: red meat is a good source of iron, for example. In developed countries, iron supplements are readily available; in less developed countries however, many people may be unable to afford such supplements or may be unaware of their existence.

In a study published in 2016[1], an attempt was made to identify social, demographic and economic factors associated with anemia among women between 15 and 49 years of age in Afghanistan. The study authors used data from the 2010 UNICEF Multiple Indicator Cluster Survey (MICS) for Afghanistan: see http://mics.unicef.org/ for general information about the MICS surveys, and this link for a report describing the 2010 survey for Afghanistan. The latter report can also be downloaded from the 'In-course assessment 2' section of the STAT0023 Moodle page.

The main points to note about the 2010 MICS survey for Afghanistan are as follows:

- The survey is designed to be nationally representative, with 13,314 households visited over eight regions of the country. Responses were obtained from 98.5% of these households, in which 22,053 women were identified between the ages of 15 and 49. Interviews were carried out with 21,290 of these women.
- Most of the interview questions are related to the social, economic and demographic characteristics of households and individual women: the questions are given in Appendix F of the Afghan survey report (see link above). Additionally, haemoglobin tests were administered to women in half of the households. Haemoglobin levels (in g/Dl) are available for 9,199 women in the survey.

The authors of the 2016 anemia study used the MICS survey data for the 9,199 women with haemoglobin measurements. They discarded cases with haemoglobin levels above 24g/Dl

[1] **Citation**: Flores-Martinez A., G. Zanello, B. Shankar and N. Poole (2016). Reducing Anemia Prevalence in Afghanistan: Socioeconomic Correlates and the Particular Role of Agricultural Assets. *PLoS ONE* **11(6)**: e0156878. A copy of this paper can also be downloaded from the 'In-course assessment 2' section of the STAT0023 Moodle page.

as being unrealistic and therefore probably erroneous: this left 9,174 women in the data set used for further analysis. Each woman was classified as being anemic or not, depending on whether her altitude-adjusted haemoglobin level was below or above a critical threshold: the thresholds were obtained from WHO guidelines (11g/Dl for pregnant women, 12g/Dl for others). The altitude adjustment aimed to provide an 'equivalent haemoglobin level' in terms of blood oxygen capacity, taking into account the fact that oxygen saturation in blood reduces at high altitudes: the adjustment was based on the altitude of the province in which each woman lived. The authors then used logistic regression to model the dependence of anemia status on selected covariates from the survey data. They carried out some further analyses as well, but we don't need to consider these.

From the 2016 anemia study, three features of the authors' analysis are worth noting:

- The altitude-based adjustments to the haemoglobin measurements are imperfect, for two reasons. The first is that it's hard to find any clear justification anywhere for these particular adjustments (the study authors used the adjustments recommended in a 2008 paper,[2] but this doesn't explain how they were derived − nor does it give any indication of how accurate they are). The second is that the precise altitudes at which each woman lived are not known: they have been approximated by the altitude of the provincial capital, which may be very inaccurate in mountainous areas.
- The authors pre-selected a group of potential covariates and included them all in their model, without attempting to further refine the analysis e.g. by dropping non-significant covariates or investigating potential interactions.
- There were many instances of multiple women living in the same household. There is likely to be an association between the anemia status of people in the same household, for example due to shared diets − and it is unlikely that this association can be explained entirely using the covariate information available in the survey. This means that the responses are probably *not* independent given the covariates. In turn, the standard errors, *p*-values and so forth in the study may be incorrect.

The data used in the 2016 study have kindly been provided to us by one of its authors: Prof Bhavani Shankar (now at the University of Sheffield). Therefore: on the 'In-course assessment 2' tab of the STAT0023 Moodle page, you will find a `CSV` file called `AnemiaData.csv` which contains a modified version of the data used by the authors of the 2016 study.

In this dataset, one woman has been randomly sampled from each household so that there is no remaining within-household dependence. The file contains 5,421 records (i.e. rows of data): each record represents one woman. Haemoglobin measurements (in g/Dl) are provided for the first 4,382 women, along with other information about the women and their households: full details can be found in the Appendix to these instructions. For the remaining 1,039 women, however, the haemoglobin measurements are not provided: they are given as `-1`.

---

[2] **Citation**: Sullivan, K. M., Mei, Z., Grummer-Strawn, L. and Parvanta, I. (2008). Haemoglobin adjustments to define anemia. *Tropical Medicine & International Health*, **13**: 1267-1271. As usual, a copy is available from the Moodle page.

Your task in this assessment is to carry out some data preprocessing and then to use the data from the first 4,382 records, to build a statistical model that will help you to:

- Understand the social, demographic and economic factors associated with variation in haemoglobin levels between Afghan women in the 15-49 age range; and
- Estimate the haemoglobin levels for each of the 1,039 records where you don't have this information.

2 Detailed instructions

You may use either `R` or `SAS` for this assessment.

2.1 Your tasks

1. Read the data into your chosen software package, and carry out any necessary recoding (e.g. to deal with the fact that `-1` represents a missing value).
2. Carry out an exploratory analysis that will help you to start building a sensible statistical model to understand and predict each woman's haemoglobin level. This analysis should aim to identify an appropriate set of candidate variables to take into the subsequent modelling exercise, as well as to identify any important features of the data that may have some implications for the modelling. You will need to consider the context of the problem to guide your choice of exploratory analysis. See the 'Hints' below for some ideas.
3. Using your exploratory analysis as a starting point, develop a statistical model that enables you to *predict* each woman's haemoglobin level based on (a subset of) the other variables in the dataset; and also to *understand* the variation of haemoglobin levels between women. To be convincing, you will need to consider a range of models and to use an appropriate suite of diagnostics to assess them. Ultimately however, you are required to recommend a single model that is suitable for interpretation, and to justify your recommendation. Your chosen model should be either a linear model, a generalized linear model or a generalized additive model.
4. Use your chosen model to predict the haemoglobin levels for each of the women where this information is missing, and also to estimate the standard deviation of your prediction errors.

2.2 Submission requirements

Submission for this assessment is electronic, via the STAT0023 `Moodle` page. You are required to submit three files, as follows:

1. A PDF file containing a report on your analysis.
2. An R script or SAS program corresponding to your analysis and predictions.
3. An ASCII file containing your predictions for the 1,039 women with missing haemoglobin measurements.

More details on these files are as follows.

2.2.1 PDF file: report on your analysis

Your report should be submitted as a `PDF` file named as `########_rpt.pdf`, where `########` is your group ID. For example, if your group ID is 'ICA2_Group789' then your report should be named `ICA2_Group789_rpt.pdf`. **Your file name must be EXACTLY correct**, to ensure that we can identify it.

Your report must not exceed 2,500 words of text plus two pages of graphs and / or tables. Section 2.4 describes what is included in the word count.

Your report should be in three sections, as follows:

I. Introduce the problem context and describe briefly what aspects you considered at the outset, how you used these to start your exploratory analysis, and what were the important points to emerge from this exploratory analysis.
II. Describe briefly (without too many technical details) what models you considered in step (3) above, and why you chose the model that you did.
III. State your final model clearly, summarise what your model tells you about the factors associated with variation of haemoglobin levels in Afghan women in the 15–49 age range, and discuss any potential limitations of the model.

Your report should not include any computer code. It should include some graphs and / or tables, but only those that support your main points. **Graphs and tables must appear on separate pages** (up to two pages in total, as described above), or they will be not be marked and will contribute to your word count.

As well as describing your analysis, **if you are working as a pair then you must include an additional page at the end of your report, describing how each pair member contributed to the project**. If both pair members agree that they contributed equally then it is sufficient to write a single sentence saying so. Alternatively you may describe your own personal contributions to the project. Note that this page will not be marked and does not contribute to the word count; nor will it be used to allocate different marks to the pair members. The purpose is to encourage you to be mindful about contributing to this piece of group-work.

2.2.2 R script or SAS program

Your script / program should be named `########.r` or `########.sas` as appropriate, where `########` is your group ID with underscores instead of spaces. For example, if your group ID is 'ICA2_Group789' and you use `R`, your script should be named `ICA2_Group789.r`.

Your script / program should run *without user intervention* on any computer with R or SAS installed, providing the file `AnemiaData.csv` is present in the current working directory / current folder. When run, it should produce any results that are mentioned in your report, together with:

- all graphs that are included in your report, saving them automatically to the current folder / working directory;
- a file containing your predictions and the associated standard deviations, named correctly (see below).

You may not create any additional input files that can be referenced by your script; nor should you write code that requires access to the internet in order to run it. If you use R however, you may use the following additional packages if you wish (together with other libraries that are loaded automatically by these): `mgcv`, `ggplot2`, `grDevices`, `RColorbrewer`, `lattice` and `MASS`. You may not use any other add-on packages: for present purposes, an "add-on package" is one that requires a `library()` or `require()` command or equivalent (e.g. the `package::command` syntax) before it can be used, if your `R` system is installed using default settings.

2.2.3 ASCII file containing your predictions

This file should be named `########_pred.dat`, where `########` is your group ID. The file should contain three columns, separated by spaces and with *no header*. The first column should be the record identifier (corresponding to variable ID in file `AnemiaData.csv`); the second should be the corresponding haemoglobin prediction, and the third should be the standard deviation of your prediction error.

2.3 Marking criteria

There are **75 marks** for this exercise. These are broken down as follows:

- **Report: 40 marks.** The marks here are for: displaying awareness of the context for the problem and using this to inform the statistical analysis; good judgement in the choice of exploratory analysis and in the model-building process; a clear and well-justified argument; clear conclusions that are supported by the analysis; and appropriate choice and presentation of graphs and / or tables. The mark breakdown is as follows:
  - **Awareness of context: 5 marks.**
  - **Exploratory analysis: 10 marks.** These marks are for (a) tackling the problem in a sensible way that is justified by the context (b) carrying out analyses that are designed to inform the subsequent modelling.
  - **Model-building: 10 marks.** The marks are for (a) starting in a sensible place that is justified from the exploratory analysis (b) appropriate use of model output and diagnostics to identify potential areas for improvement (c) awareness of different modelling options and their advantages and disadvantages (d) consideration of the social, economic and demographic context during the model-building process.
  - **Quality of argument: 5 marks.** The marks are for assembling a coherent 'narrative', for example by drawing together the results of the exploratory analysis so as to provide a clear starting point for model development, presenting the model-building exercise in a structured and systematic way and, at each stage, linking the development to what has gone before.
  - **Clarity and validity of conclusions: 5 marks.** These marks are for stating clearly what you have learned about how and why haemoglobin levels vary between women, and for ensuring that this is supported by your analysis and modelling.
  - **Graphs and / or tables: 5 marks.** Graphs and / or tables need to be relevant, clear and well presented (for example, with appropriate choices of symbols, line types,

captions, axis labels and so forth). There is a one-slide guide to 'Using graphics effectively' in the Week 1 slides for the course.

Note that **you will only receive credit for the graphs in your report if your submitted script / program generates and automatically saves *all* of these graphs, without user intervention, when it is run (e.g. using `source()` in R)**.

- **Coding: 15 marks.** There are 3 marks here for reading the data and handling missing values correctly; 7 marks for effective use of your chosen software (e.g. programming efficiently and correctly); and 5 marks for clarity of your code − commenting, layout, choice of variable / object names and so forth.
- **Prediction quality: 20 marks.** The remaining 20 marks are for the quality of your predictions. Note, however, that **you will only receive credit for your predictions if your submitted script generates an identical copy of your submitted prediction file, without user intervention, when it is run.** If this is not the case, your predictions will earn zero marks.

  For the prediction quality marks, *you are competing against each other*. Your predictions will be assessed using the following score:

$$S = \sum_{i=1}^{1\,039} \left[\log\hat{\sigma}_i + \frac{(Y_i - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2} \, , \right].$$

  where:

  - Y_i is the actual actual haemoglobin measurement (which the examiners know) for the ith prediction;
  - $\hat{\mu}_i = \widehat{\mathbb{E}}(Y_i)$ is your corresponding prediction;
  - $\hat{\sigma}_i$ is your quoted standard deviation for the prediction error.

  The score S is an approximate version of a *proper scoring rule*, which is designed to reward predictions that are close to the actual observation and are also accompanied by an accurate assessment of uncertainty (this was discussed during the Week 10 lecture, along with the rationale for using this score for the assessment). Low values are better. The scores of all of the students in the class (and the lecturer) will be compared: students with the lowest scores will receive all 20 marks, whereas those with the highest scores will receive fewer marks. The precise allocation of marks will depend on the distribution of scores in the class.

  If you don't supply standard deviations for your prediction errors, the values of the $\{\hat{\sigma}_i\}$ will be taken as zero: this means that your score will be $-\infty$ if you predict every value perfectly (this is the smallest possible score, so you'll get 20 marks in this case), and $+\infty$ otherwise (this will earn you zero marks).

2.4 Word count

As noted above, your report must not exceed 2,500 words. This word count includes titles, footnotes, appendices, references etc. − in fact it includes everything except the two pages of graphs / tables and, if present, the separate page describing the contribution of each pair member.

You will be penalised if your report exceeds EITHER the specified 2,500-word limit or the number of pages of graphs and / or tables. Following UCL guidelines, the maximum penalty is 7 marks, and no penalty will be imposed that takes the final mark below 30/75 if it was originally higher.

Subject to these conditions, penalties are as follows:

- *More than two pages of graphs and / or tables:* zero marks for graphs and / or tables, in the marking scheme given above.
- *Exceeding the word count by 10% or less:* mark reduced by 4.
- *Exceeding the word count by more than 10%:* mark reduced by 7.

In the event of disagreement between reported word counts on different software systems, the count used will be that from the examiner's system. The examiners will use an R function called `PDFcount` to obtain the word count in your PDF report: this function is available from the Moodle page in file `PDFcount.r`.

There are no word or page limits for your `R` script / `SAS` program, or for your file of predictions.

2.5 Hints on tackling the assessment

1. There is no single 'right' answer to this assignment. There is a huge range of options available to you, and many of them will be sensible.
2. You are being assessed not only on your computing skills, but also on your ability to carry out an informed statistical analysis: material from other statistics courses (in particular STAT0006, for students who have taken it) will be relevant. To earn high marks, you need to take a structured and critical approach to the analysis and to demonstrate appropriate judgement in your choice of material to present.
3. At first sight, the task may appear challenging. However, there is a lot of information that can guide you: look at some of the web links earlier in these instructions, and at other commentaries on anemia as a disease, to gain some understanding of what kinds of relationships you might look for in the data.
4. When building your model, you have two main decisions to make. The first is: should it be a linear, generalized linear or generalized additive model? The second is: which covariates should you include? You might consider the following points:
    - **Linear, generalized linear or generalized additive?** This is best broken down into two further questions, as follows:
        - *Conditional on the covariates, can the response variable be assumed to follow a normal distribution with constant variance?* In this assignment, the response variable cannot be negative; nor can it exceed 24g/Dl (see above). Therefore, it cannot have exactly a normal distribution. However, you may find that the residuals from a linear regression model are *approximately* normal − and you may judge that the approximation is adequate for your purposes.

          The 'constant variance' assumption may also be suspect: for positive-valued quantities, it is common for the variability to increase with the mean. If this is the

case here, you need to decide whether it varies enough to matter: you need to think about whether the effect is big enough that you can improve your predictions (and hence your score!) by accounting for it e.g. using a GLM. You might consider using your exploratory analysis to gain some preliminary insights into this point.

- *Are the covariate effects best represented parametrically or nonparametrically?* Again, your exploratory analysis can be used to gain some preliminary insights into this. You may want to look at the material from week 6, for examples of situations where a nonparametric approach is needed.

– **Which covariates?** The data file contains a lot of potential covariates, some of them factors with several levels. You have many choices here, and you will need to take a structured approach to the problem in order to avoid running into difficulties. The following are some potentially useful ideas:

- *Look at other literature on anemia, and on the structure of Afghan society at the time when the data were gathered*. What factors are considered to be the most important characteristics controlling haemoglobin levels? Were there known health inequalities within Afghanistan at the time? Can these be linked to covariates for which you have information? Obviously, if you do this then you will need to acknowledge your sources in your report.
- *Define useful summary measures on contextual grounds, and work with these*. For example, many of the potential covariates are binary factors indicating ownership of different types of animals: you might decide to combine these by summing them. Another covariate is 'age': you might decide to divide this into three or so groups.
- *Define new variables based on the correlations between the existing variables, and work with these*. If several continuous variables are highly correlated, then it is difficult to disentangle their effects and it may be preferable to work with a single 'index' that combines all of them. This is the basis of techniques such as Principal Components Analysis, that were discussed during the Week 10 lecture (along with how to implement them in `R` and `SAS`).

You should not start to build any models until you have formed a fairly clear strategy for how to proceed. Your decisions should be guided by your exploratory analysis, as well as your understanding of the context.

5. Don't forget to look for interactions! For example, one of the variables in the data set is `Sheep`, which is a factor (i.e. categorical covariate) indicating whether or not the woman's household owns sheep: the authors of the 2016 study concluded that this variable was significantly associated with a woman's anemia status. Another variable is `WealthScore`, which is an aggregate index of household wealth. It is conceivable that sheep ownership is important for lower-income families where home-produced food may contribute a substantial proportion to the diet, but that it is less important for wealthier families who can afford to buy food from elsewhere. Look at the analysis of the iris data from Workshop 2, for a similar kind of situation.

Sometimes people get confused about the difference between interactions and collinearity. **Reminder:**

- An *interaction* describes the way in which covariates must be considered in combination to characterise their relationship with the response variable.
- By contrast, *collinearity* is about correlations between the covariates: this has no reference to the response variable. Collinearity just makes it harder to identify which covariates are genuinely associated with the response (recall the "sheep energy" example from Week 9).

6. You probably won't find a perfect model in which all the assumptions are satisfied: models are just models. Moreover, you should not necessarily *expect* that your model will have much predictive power: maybe the covariates in the data set don't provide much useful information about a woman's haemoglobin levels. You should focus on finding the best model that you can, therefore − and acknowledge any deficiencies in your discussion.
7. To obtain the standard deviations of your prediction errors, you need to do some calculations. Specifically:
   i. Suppose $\hat{\mu}_i = \widehat{\mathbb{E}}(Y_i)$ is your ith predicted haemoglobin level, and that Y_i is the corresponding actual value.
   ii. Then your prediction error will be $Y_i - \hat{\mu}_i$.
   iii. Y_i and $\hat{\mu}_i$ are independent, because $\hat{\mu}_i$ is computed using only information from the first 4,382 records and Y_i relates to one of the 'new' records.
   iv. The *variance* of your prediction error is thus equal to $\text{Var}(Y_i) + \text{Var}(\hat{\mu}_i)$.
   v. You can calculate the standard error of $\hat{\mu}_i$ in both `R` and `SAS`, when making predictions for new observations − see Workshops 6 and 9. Squaring this standard error gives you $\text{Var}(\hat{\mu}_i)$.
   vi. You can estimate $\text{Var}(Y_i)$ by plugging in the appropriate formula for your chosen distribution − for example, if you're using a gamma distribution (which is a possibility when using GLMs for non-negative response variables) then $\widehat{Var}(Y_i) = \hat{\phi}\hat{\mu}_i^2$, where $\hat{\phi}$ is the estimated dispersion parameter for your model (see Section 2.1 of the Week 6 self-study materials).
   vii. Hence you can estimate the standard deviation of your prediction error as $\hat{\sigma}_i = \sqrt{\widehat{\text{Var}}(Y_i) + \text{Var}(\hat{\mu}_i)}$.[3]

---

[3] In fact, for the case of linear models this is esssentially the calculation that is used in the construction of prediction intervals (see your STAT0006 notes or equivalent).

Appendix: the `AnemiaData.csv` data set

Data sources and processing

The data provided in `AnemiaData.csv` are ultimately derived from the full 2010 Afghanistan MICS dataset, available from http://mics.unicef.org/surveys. The authors of the 2016 study selected a subset of the variables from this survey as described in their supporting information (click the blue text to follow the link). These authors' data have subsequently been processed in the following way to create `AnemiaData.csv`:

1. The variable names were modified for ease of interpretation.
2. One women was randomly sampled from each household, so that the resulting data set does not contain any within-household dependence.
3. The original dataset contained many dummy variables representing different levels of the same factor: for example, there were binary variables representing each of the eight regions of Afghanistan. Each group of dummy variables has been aggregated into a single factor variable with multiple levels: for example, the eight binary regional variables have been aggregated into a single factor `Region` with eight levels.
4. Some less relevant variables, variables with large quantities of missing data, and variables that could be calculated from other information in the data set, were removed. An example of a 'less relevant' variable is the survey weight given to a particular woman: this would be useful if we wanted to estimate (say) the mean haemoglobin level for all women in Afghanistan, but it is not needed here. A variable with large quantities of missing data is the mean upper-arm circumference (MUAC), which was not measured for any pregnant woman. Variables that could be calculated from other information include 'wealth quintiles': these can be calculated from the `WealthScore` variable.
5. The `Province` variable, originally provided as a numeric code, was relabelled with the actual province names.
6. The rows of the dataset were randomly shuffled: this is just to make it harder to identify the rows on the basis of information that may be available on the internet.
7. A sample of roughly 80% of the records was selected for use in the 'model building' part of the assessment (this will be referred to as 'Group 1' below), with the remaining 20% used for 'prediction' ('Group 2'). This was done in such a way that the two samples were non-overlapping but had very similar distributions of all potential covariates. Specifically:
   a. For each combination of the `Province` and `Pregnant` variables (see below), 80% of the women were sampled at random, without replacement, as candidates to use in Group 1; and the remaining 20% were allocated to Group 2.
   b. For each of the numeric covariates in the data set, a Kolmogorov-Smirnov test was performed to test the null hypothesis that the underlying distributions in Groups 1 and 2 are the same.
   c. For each of the categorical covariates in the data set, a chi-squared test was performed to test the null hypothesis that the category proportions in Groups 1 and 2 are the same.
   d. The samples were accepted only if the p-values for *all* of the Kolmogorov-Smirnov and chi-squared tests were greater than 0.01. Otherwise, a new candidate sample was drawn in step (a) and the procedure was repeated.

The Kolmogorov-Smirnov and chi-squared tests are used here as a convenient way to measure whether two distributions are roughly similar. Note, however, that the haemoglobin levels were *not* included in this balancing exercise: this is because the performance of predictions would be artificially enhanced if they were included (for example, we would know that the mean haemoglobin level for Group 2 is similar to that for Group 1). Note also that no attempt has been made to balance the groups in terms of *combinations* of the covariates.

8. The 'Group 2' records were placed at the end of the data table, with their haemoglobin levels set to `-1`; and a new `ID` variable was created so that each record has an ID number between 1 and 5,421.

Description of variables

This section gives a brief description of each of the variables in `AnemiaData.csv`.

Variable name	Description
`ID`	Record ID, from 1 to 5,421
`Haemoglobin`	Individual's haemoglobin level (g/Dl)
`Age`	Individual's age (years)
`RecentBirth`	Has the individual given birth in the last two years? This takes one of two values: `Yes` and `No`.
`HHSize`	Number of household members.
`HHUnder5s`	Number of children under the age of 5 in the household.
`CleanWater`	Does the household have access either to water from a protected source (including a borehole), or to treated drinking water? (`Yes` / `No`)
`TreatedWater`	Is the household's water treated for drinking? (`Yes` / `No`)
`Electricity`	Does the household have electricity? (`Yes` / `No`)
`Toilet`	Does the household have toilet facilities (flushing toilet, pit latrine, composting toilet, bucket, vault or sanitation)? (`Yes` / `No`)
`BikeScootCar`	What proportion of the following does the household own: (a) bike (b) scooter / motorcycle (c) car / truck (recorded as a value of 0, $1/3$, $2/3$ or 1).
`AnimCart`	Does any household member own an animal-drawn cart? (`Yes` / `No`)
`AgricLandOwn`	Does any household member own agricultural land? (`Yes` / `No`)
`Cows`	Does the household own any cattle? (`Yes` / `No`)
`Horses`	Does the household own any horses, donkeys or mules? (`Yes` / `No`)
`Goats`	Does the household own any goats? (`Yes` / `No`)
`Sheep`	Does the household own any sheep? (`Yes` / `No`)

Variable name	Description
`Chickens`	Does the household own any chickens? (`Yes` / `No`)
`Rural`	Is the household in a rural area? (`Yes` / `No`)
`Province`	Which province is the household in? (this is a factor with 34 levels, corresponding to the provinces in Afghanistan − see Wikipedia page for maps).
`TotalChildren`	Total number of children ever born to the individual
`WealthScore`	Index of household wealth, provided as part of the MICS dataset and created using a principal components analysis incorporating information on water source, sanitation facility, house construction characteristics, ratio of people to rooms, cooking fuel type, and ownership of appliances such as a refrigerator, TV and radio. Negative values indicate households that are less wealthy than average, positive values are more wealthy.
`AgricArea`	How much agricultural land does the household own (hectares). A value of 0 when `AgricLandOwn` is `Yes` means that the household owns less than 1ha of land. Values up to 94 are rounded down; a value of 95 means 'at least 95ha'.
`Pregnant`	Is the individual currently pregnant? (`Yes` / `No`)
`Education`	To what level is the individual educated? This is a factor with three levels: `None` (no education), `Primary` (educated to primary level) and `Secondary+` (educated to secondary level or above).
`HHEducation`	To what level is the head of the household educated? This is a factor with three levels, coded as for `Education`.
`Region`	Name of the region in which the household is situated. There are several provinces in a region. The values are `central`, `central_highlands`, `east`, `northeast`, `south`, `southeast`, `west` and `north`.
`Ethnicity`	Head of the household's ethnic group or primary language. Possible values are `Dari`, `Pashto`, `Uzbek`, `Turkmen` and `Other/missing`. The MICS survey data do not distinguish between `Other` groups and missing values.