# STAT0023 Computing for Practical Statistics

## In-course assessment 2, take-home component (2021–22 session)

## Rubric

- Your solutions should be your own work and are to be submitted electronically to the course Moodle page by **12 noon on MONDAY, 25TH APRIL 2022**.
- You can work in groups of up to 5 for this assessment. It is up to you to form your own groups. *You MUST register your choices on Moodle by 12 noon on MONDAY, 28TH MARCH 2022*.
- All members of a group will be jointly responsible for the work that is submitted and you will be awarded the same mark.
- Ensure that you electronically 'sign' the plagiarism declaration on the Moodle page when submitting your work. All members of a group should check what has been submitted before signing this declaration: if any plagiarism or collusion is identified with anyone outside of your group, you will share responsibility for it.
- Late submission will incur a penalty unless there are extenuating circumstances (e.g. medical) supported by appropriate documentation and notified within one week of the deadline above. Penalties, and the procedure in case of extenuating circumstances, are set out in the latest editions of the Statistical Science Department student handbooks which are available from the departmental web pages.
- Failure to submit this in-course assessment will mean that your overall examination mark is recorded as "non-complete," i.e. you will not obtain a pass for the course.
- Submitted work that exceeds the specified word count will be penalized. The penalties are described in the detailed instructions below.
- Your solutions should be your own work. When uploading your scripts, you will be required to electronically sign a statement confirming this, and that you have read the Statistical Science department's guidelines on plagiarism and collusion (see below).
- Any plagiarism or collusion can lead to serious penalties for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in the departmental student handbooks: the relevant extract is provided on the 'In-course assessment 2' tab on the STAT0023 Moodle page. The Turn-It-In plagiarism detection system may be used to scan your submission for evidence of plagiarism and collusion.
- You will receive feedback on your work via Moodle, and you will receive a provisional grade. *Grades are provisional until confirmed by the Statistics Examiners' Meeting in June 2022*.

## Background and overview

The term "retail industry" refers to sales by retailers directly to end consumers, including spending on goods (in store and online) and services, see the *relevant section of the Office for National Statistics website*. Retail analytics (or retail intelligence) is the collection, management and analysis of data relating to retail sales, with a view to understanding, predicting and optimizing the process on all operational levels, *see the corresponding Wikipedia page*.

Grocery retail primarily refers to the sales of food and household items. The grocery retail market is very competitive and is characterised by very narrow profit margins. Therefore, the use of effective retail analytics tools is crucial to gain efficiencies. One of the primary objectives of retail analytics is the prediction of sales[1]. Demand forecasting can aid pricing, planning, inventory management, optimisation of marketing and promotions, or identification of gaps in the market. Demand forecasting models can use information such as the type of product, the price, the size, the overall value of the brand, the existence of a promotion, as well as characteristics of each particular store, such as local demographics.

In this assessment, we will use data from a leading US supermarket chain about the monthly sales of 15 products in 77 stores from 4 US states over a period of 3 years. The over-arching goal is to create a statistical model to explain and predict sales of each product, in each month, at each store. The data were derived from the *dunnhumby source files* resource; *dunnhumby* is a leading retail analytics company in the UK.

The data are provided as three separate files on the 'In-course assessment 2' tab of the STAT0023 Moodle page. The file `groceryProducts.csv` contains information about each product, `groceryStores.csv` contains information about each store, and `groceryTransactions.csv` contains the number of items sold per product, per store, per month, along with relevant information such as marketing promotions. The first 7,500 rows are complete, i.e., contain all values of sales and covariates. The last 2,500 rows contain all values of the covariates, but `-1` for the number of units sold. More details about the input files can be found in the Appendix.

Your task in this assessment is to carry out some data preprocessing to combine the datasets and then to use the data to build a statistical model that will help you to:

- Understand the role of price, promotion, brand, product and store characteristics and time of the year in product sales;
- Estimate the sales of each of the 2,500 records where you don't have this information.

## Detailed instructions

You may use either R or SAS for this assessment.

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).

[1] Fildes, Robert, Shaohui Ma, and Stephan Kolassa. 2019. "Retail Forecasting: Research and Practice." International Journal of Forecasting. *https://doi.org/10.1016/j.ijforecast.2019.06.004*.

2. Combine the data from `groceryTransactions.csv`, `groceryProducts.csv` and `groceryStores.csv` into a single dataset (a data frame if you're using R, a dataset in SAS), so that each row corresponds to the monthly sales of a single product at a single store and includes the information about the product and the store. You may notice that the files `groceryProducts.csv` and/or `groceryStores.csv` may contain information about products/stores which are not included in the transaction dataset and so are not relevant here; you will have to make sure you deal with these accordingly.
3. Carry out an exploratory analysis that will help you to start building a sensible statistical model to understand what drives the numbers of sales for each product, in each store, at each time point (the number of sales is called `UNITS` in the dataset). 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.
4. Using your exploratory analysis as a starting point, develop a statistical model that enables you to *predict* the number of sales on (a subset of) the other variables in the dataset, and also to *understand* the variation in sales. 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.
5. Use your chosen model to predict the number of sales of each product at each store and time point where this information is missing, and also to estimate the standard deviation of your prediction errors.

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

- A report on your analysis, not exceeding 2 500 words of text plus two pages of graphs and / or tables. The word count includes titles, footnotes, appendices, references etc. – in fact it includes everything except the two pages of graphs / tables and the separate page describing the contribution of each group member (see below). Your report should be in three sections, as follows:

  **Section I**: Describe briefly what aspects of the problem context 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.
  **Section 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.
  **Section III**: State your final model clearly, summarise what your model tells you about the factors associated with product sales, 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**, or they will be included in the word count.

In addition to your data analysis, **you must include an additional page at the end of the report where each member of the group briefly describes their contribution to the project**. You will need to agree this in your groups before submitting the report. If all members agree that they contributed equally then it is sufficient to write a single sentence to that effect, or alternatively you are very welcome to describe your own personal contribution to the project. Note that this page will not be marked and does not contribute to the word count; nor will different marks be allocated to different members based on this. The purpose is to encourage you all to be mindful about contributing to this piece of group-work.

**Your report should be submitted as a `PDF` file named as `########_rpt.pdf`, where `########` is your group ID, with any spaces replaced by underscores (IMPORTANT!!!)**. For example, if your group ID is 'ICA2Group C,' your report should be named `ICA2Group_C_rpt.pdf`.

- An R script or SAS program corresponding to your analysis and predictions. Your script/program should run *without user intervention* on any computer with R or SAS installed, providing the files `groceryProducts.csv`, `groceryStores.csv` and `groceryTransactions.csv` are present in the current working directory / current folder. When run, it should produce any results that are mentioned in your report, together with the predictions and the associated standard deviations. **The 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 'ICA2Group C' and you use R, your should be named `ICA2Group_C.r`.

  You may not create any additional input files that can be referenced by your script; nor should you write any code that requires access to the internet in order to run it. If you use R however, you may use the following additional libraries 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 libraries: for present purposes, an "add-on library" 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.
- A text file containing your predictions for the 2 500 observations with missing counts. **This file should be named `########_pred.dat`, where `########` is your group ID with underscores instead of spaces**. 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 `groceryTransactions.csv`); the second should be the corresponding count prediction, and the third should be the standard deviation of your prediction error.
- **NOTE: all members of a group must confirm their submission on Moodle before the submission deadline**.

## Hints on tackling the assessment

1. There is not a single 'right' answer to this assignment. There is a huge range of options available to you, and many of them will be sensible.

2. 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 and it is an integer. 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 several covariates, some of which are more important than others. 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 important factors for grocery retail sales*. What are considered to be the most important factors for predicting product sales? 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.
     - *Which factor variables?* Each observation has the store number, its corresponding city as well as the state it belongs to. Choosing whether to include these (or which one), depends on whether your data contains enough resolution to warrant it. You may also want to think about how one might cluster cities/stores into groups that share similar characteristics.
     - *How to incorporate time?* In the STAT0023 module we have not studied models for time series. However, you can think about including time as an additional covariate: make sure you clearly think about and explain what the implications of this are.

   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.

3. Don't forget to look for interactions! For example, the effect of price reductions may be different in different cities.
4. 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 just don't provide very much useful information. You should focus on finding the best model that you can, therefore − and acknowledge any deficiencies in your discussion.
5. 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 $i$th predicted death count 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 $5\,401$ 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 the materials from Weeks 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 linear model then this is just the error variance estimate, whereas if you're using a Poisson distribution (which is a possibility when the response variable is a count) then $\widehat{\text{Var}}(Y_i) = \hat{\mu}_i$.
   vii. Hence you can estimate the standard deviation of your prediction error as $\hat{\sigma}_i = \sqrt{\widehat{\text{Var}}(Y_i) + \text{Var}(\hat{\mu}_i)}$. In fact, for the case of linear models this is exactly the calculation that is used in the construction of prediction intervals (see your STAT0006 notes or equivalent).
6. Larger stores will tend to have more sales simply because they serve more customers, so you may want to think carefully about how to incorporate the average number of baskets per store into your model. Some of you may choose to model the effects of covariates on approximate sales *rates* (i.e. the numbers of sales as a proportion of the average total baskets per store) instead of the actual counts. However, the assignment instructions tell you to model and predict the *numbers* of sales, so you need to make sure you estimate the correct quantity at the end. One option here would be to fit models to the rates and then to derive the corresponding expressions for the counts (since the count is equal to the rate times the average total baskets per store).

---

# The dataset

## Data source and pre-processing

The file `groceryTransactions.csv` contains a sample of monthly sales and promotional information of 15 products over a period of 3 years in individual stores in 4 US states. The file `groceryStores.csv` contains information about each of these stores, and the file `groceryProducts.csv` contains information about each product. All 15 products belong to the "cold cereal" product category. The data were derived from the *dunnhumby source files database*, "Breakfast at the Frat" section and have been preprocessed to make them suitable for this assessment in the following ways.

- The period of January 2009 to December 2011 was used.
- All products in the "cold cereal" category were used.
- Some variables (for example the address of each store, the product sizes and the product descriptions) were removed from the data.
- Weekly data were aggregated into monthly data by mapping each week to the month corresponding to the first day of the week.
- The names and codes of the stores and cities were shuffled to make it impossible for you to figure out the missing records from the original data.
- A random sample of 10,000 rows was taken from these records, to make computational feasible.
- A sample of 75% of the records was selected for use in the "model building" part of the assessment (this will be referred to as the "training set" below), with the remaining 25% used for 'prediction' (referred to as the "held-out set"). This was done in such a way that the two samples were non-overlapping but had similar characteristics. Specifically, the subsampling was done such that any month, year, store and product type appearing in the held-out set also appeared in the training set (but not necessarily vice-versa).

## The `groceryProducts.csv` dataset

This file contains information about each product. Specifically, each row contains

- `UPC`, the unique product number.
- `MANUFACTURER`, the brand.
- `CATEGORY`, the product category (in the case of this dataset "cold cereal").
- `SUB_CATEGORY`, the product sub-category.

## The `groceryStores.csv` dataset

This file contains information about each store in the dataset. Specifically

- `STORE_NUM`, the unique store number.
- `CITY`, the city of the store.
- `STATE`, the US state.
- `AREA_CODE`, the region of the store.
- `STORE_TYPE`, the type of the store (`VALUE`, `UPSCALE` or `MAINSTREAM`).

## The `groceryTransactions.csv` dataset

This file contains the information about monthly sales of the products at each store. Specifically

- `ID`, the record ID, from 1 to 10,000.
- `STORE_NUM`, the store number.

- `UPC`, the unique product code.
- `MONTH`, the month.
- `YEAR`, the year.
- `NWEEKS`, the number of weeks in the data whose first day falls in `MONTH` (the maximum being 5). Some entries have `NWEEKS` equal to a smaller number (eg 1 or 2) because some products may have not been available at that store throughout the month.
- `FEATURE`, the proportion of weeks in `MONTH` that the product was promoted through a marketing circular.
- `DISPLAY`, the proportion of weeks in `MONTH` that the product was on a special display in store.
- `TPR_ONLY`, the proportion of weeks in `MONTH` that the product was on temporary price reduction only.
- `PRICE`, the average price in that month.
- `BASE_PRICE`, the average ”regular price” (i.e., without any promotions) in that month.
- `UNITS`, the number of items of each product sold in any week whose first day of `MONTH`, i.e., your response variable that you’re trying to predict.

---

## 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 retail and marketing 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 the numbers of sales vary, 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 materials of Week 1 of 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 when it is run.

  ***Note** that 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`*.*

- **Coding: 15 marks.** There are 4 marks here for reading the data, preprocessing and setting up variable names correctly and efficiently; 7 marks for effective use of your chosen software in the exploratory analysis and modelling (e.g. programming efficiently and correctly); and 4 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* `########_pred.dat` *file is identical to that produced by your script / program when it is run: if this is not the case, your predictions will earn zero marks.*

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

$$S = \sum_{i=1}^{2500} \left[ \log \sigma_i + \frac{(Y_i - \hat{\mu}_i)^2}{2\sigma_i^2} \, , \right] .$$

  where:

  - $Y_i$ is the actual number of deaths (which the examiners know) for the $i$th prediction;
  - $\hat{\mu}_i = \widehat{\mathbb{E}}(Y_i)$ is your corresponding prediction;
  - $\sigma_i$ 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 $\{\sigma_i\}$ 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).