STA302/1001: Assignment 1

Craig Burkett

Oct 15, 2015

Due at the beginning of lecture on Thursday, Oct 15^{th}*. Please hand it in on 8.5 x 11 inch paper, stapled in the upper left, with no other packaging and no title page. Please try to make this assignment look like something you might hand in to your boss at a job. In particular, it is inappropriate to hand in pages of R output without explanation or interpretation. Quote relevant numbers from your R output as part of your solutions. The only direct R output you should submit with the assignment are relevant plots.* ***You must append your R program file to the end of the assignment, formatted nicely with a fixed-width font.*** *No assignment will be marked without a program file, and marks will be deducted if the instructions above are not followed.*

In this assignment, you will examine some datasets from the Toronto Open Data Portal, containing statistical information on marriage and business licenses. Any time that I use the words {Present, State, Give, Show, Predict, Display}, you must supply that plot/table/output/prediction in your submission. If I say {Produce, Make}, you do not need to show what you produced or made, but you still need to do it.

The data dictionary and a link to the datasets are available on Portal. To answer most of these questions you can use the sample code from lecture, but you will also have to search for some functions online. This is how 99%* of all useRs learn R. You can always use the forum on Portal if you get frustrated, so start early!

*Note: Made up statistic

A Pen and Paper (15 marks)

Please solve the following questions by hand, or typeset using something appropriate like LaTeX, and show all of your work.

1. Suppose you observe n pairs of data (X_i, Y_i) and fit the Simple Linear Regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with the usual Gauss-Markov assumptions. Let b_0, b_1 be the LS estimates of the regression coefficients for these data. Consider a linear transformation to the X and Y variables of the form:
   $Y_i' = (Y_i - a)/c$
   $X_i' = (X_i - d)/f$

   (a) Compute the new estimate of the slope b_1' in terms of the original slope.

   (b) Compute the new estimate of the intercept b_0' in terms of the original intercept and slope.

   (c) Compute the new coefficient of determination $R^{2\prime}$ in terms of the original R^2.

   (d) Show that the transformations above do not affect inference for β_1, the slope parameter. It is sufficient to show that the t-statistics are the same.

   NB: In terms of inference for linear regression, linear transformations to either variable give equivalent results. This is not the case for non-linear transformations, like $Y' = exp(Y)$

2. In class we derived the least squares solutions to the Normal Equations, (b_0, b_1), and showed that they resulted in an extreme value of SSE. Show that these solutions actually *minimize* SSE.

B Initial Data Analysis (10 marks)

Before we get to some regression models, let's get comfortable using R.

1. To begin with, let's format the Marriage data.

   (a) Read in the data file, and check that each column was stored correctly in R using str().

   (b) Make a new factor called *Year*, which is just the four-digit year code, in numbers.

   (c) Make a new factor called *Month*, which is just the two-digit month code, in numbers.

   (d) Months are really much better with a 3-letter code (like "Jan" instead of "01"). Switch those digits to the letter code, or combine with previous step.

   There are about a half-dozen ways of accomplishing this. Some suggestions are to search: `gsub()`, `substr()`, `month.abb`, `paste()`, `as.Date()`, and the {`lubridate`} package.

   (e) Produce a second data frame, aggregating the marriage license counts over all civic centres, and save it for later.

2. Let's summarize the Marriage data with tables and graphs.

   (a) Present a 1D table showing the total number of marriage licenses issued by Civic Centre.

   (b) Display this same information visually using a bar plot. These plots look nice in *Pareto* style, that is, sorted from largest count to smallest, so please display it that way.

   (c) Present a 2D table showing the total number of marriage licenses issued by month (across columns) and by year (across rows).

   (d) Display a stacked bar plot showing the counts of marriage licenses issued in each civic centre, grouped by month, in month order. To be clear, your plot should show January (all 4 civic centres stacked), followed by February (all 4), ... until you get to December. The stacking order doesn't matter. Do you notice a 'high season' for marriage licenses?

3. Let's now clean and format the Business data.

   (a) Read in the data file, and check that each column was stored correctly in R using str(). In order to store objects as strings rather than factors, use a line like this:

   ```
   df <- read.csv(filename, head=T, stringsAsFactors = F)
   ```

   It looks like the *Issued* column has the issue date, formatted as dd/mm/yy for some dates and as dd/mm/yyyy for others. (*I know this from inspecting the data file after reading it into R – Didn't they learn anything from Y2K?*) We should clean these dates, but actually their 'dirtiness' won't affect this assignment. And there are only 30 dates with a 4-digit year.

   You can cast this character string as a date object using:

   ```
   as.Date(Issued, format = "%d/%m/%y")
   ```

   This will change all two-digit years from 69-99 into 1969-1999, which is good, and from 00-68 into 2000-2068, which is bad, because you can't get issued a business licence from the future. But it will correctly convert the dates we are going to use, so let's just sweep this problem under the rug for now.

   (b) Make two new factors called *Year* and *Month*, to match the factors created in the Marriage data.

   (c) Subset the data by the dates that are present in the Marriage data, and keep only the following columns:

   ```
   (Month, Year)
   ```

   (d) Now we have essentially the same data as the Marriage dataset, except in 'long' format. Convert this dataset to the same format as the (aggregated) Marriage data by aggregating over the unique Year/Month combinations.

C Our first regression model (15 marks)

Let's see if we can predict the number of Marriage Licenses (Y) issued by the city using the number of Business Licenses (X).

1. You'll have to join these two datasets together. You can join using the combined key of Month and Year.

2. Before fitting any kind of model, we should always visualize our data. Present a bivariate plot of these two variables, and assess whether they are suitable for linear regression.

3. Fit a Simple Linear Regression model and answer the following:

   (a) State the estimated regression equation using variable names (not X and Y)

   (b) Give an interpretation of the slope and intercept parameters in the context of this question, specifically

   (c) Give an 88% Confidence Interval for the slope and the intercept

   (d) A new month is upon us! Predict the number of marriage licenses issued if we know that there will be 550 business licenses issued, and supply an appropriate interval.

   (e) Yet another month has arrived. Predict the number of business licenses issued if we know that there will be 2500 marriage licenses issued.

4. Do you have any problems with what we just did? If so, state what problem(s) you have in a clear English sentence. You do not need to fix any problems you identify, at this point.

D Format **(10 marks)**

Please make your submission look nice. This means:

- Proper paragraph structure free of typing errors (2 mks)
- Graphs and tables should be in the **body of the report**, not thrown in at the end (2 mks)
- R Code should be appended to the assignment (3 mks), in a small fixed-width font like `courier new` (1 mks)
- Somewhere on the front page of the assignment, write the exact date and time when you finished your final copy, in 24h format (ie. if you finished Oct 13 at 8:39pm write 'Oct 13 20:39'). (2 mks)

  You are not being marked on when you finish, but if you don't write the time you won't get these 2 marks. We will use this data, anonymously, for the next assignment, so please be honest otherwise it won't work out and your assignment will be harder. And also God will kill a kitten.

  NB: These format marks will become a malus on the next assignment, instead of a bonus.