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.

   The original estimate of the slope is:

   $$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

   With the rescaling,

   $$b_1' = \frac{\sum_{i=1}^{n}(X_i' - \overline{X'})(Y_i' - \overline{Y'})}{\sum_{i=1}^{n}(X_i' - \overline{X'})^2}$$

   Notice that

   $$\overline{X'} = \frac{\overline{X} - d}{f}$$
   $$\overline{Y'} = \frac{\overline{Y} - a}{c}$$

   Therefore,

$$\begin{aligned} X' - \overline{X'} &= \frac{X_i - d}{f} - \frac{\overline{X} - d}{f} \\ &= \frac{(X_i - d) - (\overline{X} - d)}{f} \\ &= \frac{X_i - \overline{X}}{f} \end{aligned}$$

A similar argument holds to show that

$$Y' - \overline{Y'} = \frac{Y_i - \overline{Y}}{c}$$

Therefore, the new slope is

$$\begin{aligned} b_1' &= \frac{\frac{1}{cf}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{f^2}\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ &= \frac{f}{c}\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ &= \frac{f}{c} b_1 \end{aligned}$$

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

The original estimate for the intercept is

$$b_0 = \overline{Y} - b_1\overline{X}$$

With the transformed data

$$\begin{aligned} b_0' &= \frac{\overline{Y} - a}{c} - \frac{f}{c} b_1 \frac{\overline{X} - d}{f} \\ &= \frac{1}{c}\left((\overline{Y} - a) - b_1(\overline{X} - d)\right) \\ &= \frac{1}{c}\left((\overline{Y} - b_1\overline{X}) + b_1 d - a\right) \\ &= \frac{1}{c}\left(b_0 + b_1 d - a\right) \end{aligned}$$

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

Use the following definition of R^2

$$R^2 = \frac{\left[\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})\right]^2}{SS_x SS_y}$$

Using the previous derivations, we have

$$\begin{aligned}\left[(X_i' - \overline{X'})(Y_i' - \overline{Y'}\right]^2 &= \left[\left(\frac{X_i - \overline{X}}{f}\right)\left(\frac{Y_i - \overline{Y}}{c}\right)\right]^2 \\ &= \frac{1}{(cf)^2} SS_x \bullet SS_y\end{aligned}$$

And also,

$$\begin{aligned}SS_x' &= \sum_{i=1}^{n}(X_i' - \overline{X'})^2 \\ &= \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{f}\right)^2 \\ &= \frac{1}{f^2} SS_x\end{aligned}$$

$$\begin{aligned}SS_y' &= \sum_{i=1}^{n}(Y_i' - \overline{Y'})^2 \\ &= \sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{c}\right)^2 \\ &= \frac{1}{c^2} SS_y\end{aligned}$$

Note the factor of $(cf)^{-1}$ is common to the numerator and denominator, canceling, leaving R^2 unchanged.

(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.

The t-statistic for b_1' is unchanged since

$$\begin{aligned}\frac{b_1'}{\sqrt{\text{Var}(b_1')}} &= \frac{\frac{f}{c}b_1}{\sqrt{\frac{f^2}{c^2}\text{Var}(b1)}} \\ &= \frac{b_1}{\sqrt{\text{Var}(b_1)}}\end{aligned}$$

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.

   Using scalar notation, we have

$$Q = \sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$$

   The first order derivatives are

$$\begin{aligned}\frac{\partial Q}{\partial b_0} &= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) \\ \frac{\partial Q}{\partial b_1} &= -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)X_i\end{aligned}$$

   The second order derivatives and mixed partial are therefore

$$\begin{aligned}\frac{\partial^2 Q}{\partial b_0^2} &= 2n \\ \frac{\partial^2 Q}{\partial b_1^2} &= 2\sum_{i=1}^{n}X_i^2 \\ \frac{\partial^2 Q}{\partial b_0 b_1} &= 2\sum_{i=1}^{n}X_i\end{aligned}$$

Since

$$\begin{aligned}
\frac{\partial^2 Q}{\partial b_0^2} &> 0 \\
\frac{\partial^2 Q}{\partial b_1^2} &> 0 \\
&\text{and} \\
\frac{\partial^2 Q}{\partial b_0^2}\frac{\partial^2 Q}{\partial b_1^2} - \left(\frac{\partial^2 Q}{\partial b_0 b_1}\right)^2 &= 4n\sum_{i=1}^{n} X_i^2 - \left(2\sum_{i=1}^{n} X_i\right)^2 \\
&= 4n\sum_{i=1}^{n} X_i^2 - \left(2n\overline{X}\right)^2 \\
&= 4n\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \\
&> 0
\end{aligned}$$

We have this point being a minimum. ■

Alternatively, using matrix notation. Define,

$$\begin{aligned}
\mathbf{y} &= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \\
\beta &= \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad (1) \\
\mathbf{X} &= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}
\end{aligned}$$

The error function can be expressed as

$$Q = \sum_{i=1}^{n} e_i^2 = (\mathbf{y} - \beta\mathbf{X})^\top (\mathbf{y} - \beta\mathbf{X}) \tag{2}$$

The first derivative recovers the Normal Equations

$$\begin{aligned} \frac{\partial Q}{\partial \beta} &= -2\mathbf{X}^\top (\mathbf{y} - \beta\mathbf{X}) \\ &\overset{\text{set}}{=} 0 \\ \implies \mathbf{X}^\top \mathbf{X}\beta &= \mathbf{X}^\top \mathbf{y} \end{aligned} \tag{3}$$

With solution

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{4}$$

The Hessian matrix is

$$\frac{\partial^2 Q}{\partial \beta^2} = 2\mathbf{X}^\top \mathbf{X} \tag{5}$$

Which is a positive definite matrix, which is a sufficient condition to prove it is a minimum (recall from calculus). To prove it is positive definite, recall from linear algebra that for a matrix $\mathbf{X}$ and vector $\mathbf{v}$, $\mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} = (\mathbf{X}\mathbf{v})^\top (\mathbf{X}\mathbf{v}) \geq \mathbf{0}$, with this being strictly greater as the design matrix $\mathbf{X}$ is nonsingular.

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.

   The relevant code is given in the appendix.

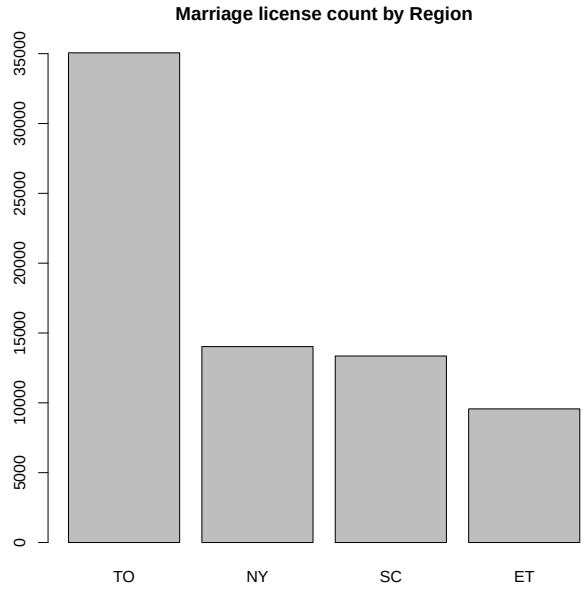
2. Let's summarize the Marriage data with tables and graphs.

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

	Total Number of Marriage Licenses
ET	9568
NY	14028
SC	13354
TO	35060

   (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.

Figure 1:



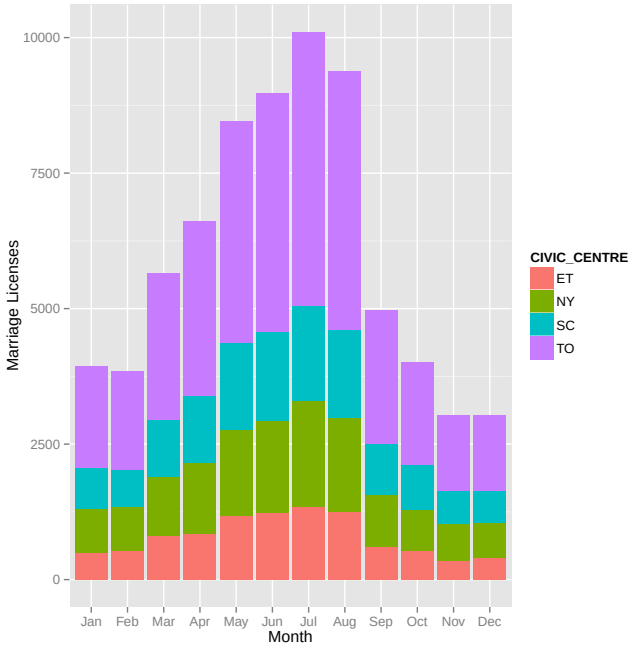
(c) Present a 2D table showing the total number of marriage licenses issued by month (across columns) and by year (across rows).

Table 1: Total number of marriage licenses issued for each month (columns) and year (rows)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2011	742	796	1210	1376	1885	1824	1943	1933	1321	1013	816	785
2012	902	879	1227	1232	1650	1843	2015	1930	1143	1065	826	785
2013	763	725	990	1360	1581	1579	1999	1821	1229	940	709	679
2014	785	717	1062	1294	1682	1806	1962	1845	1280	1001	673	785
2015	739	730	1160	1344	1663	1927	2184	1855				

(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

Figure 2: fig2d



season' for marriage licenses?

As indicated in Figure 2, there is a clear 'high season' in the summer months.

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.

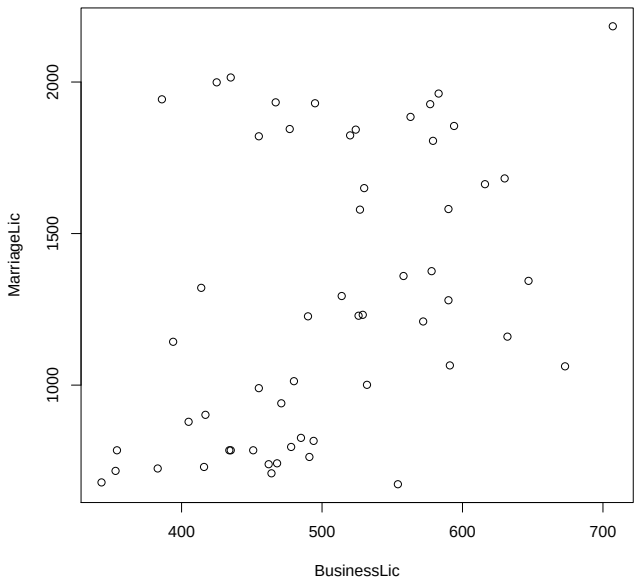
C Our first regression model (15 marks)

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.

Figure 3:



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

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

The estimated number of marriages licenses is expected to be 126.250, plus an additional 2.302 for each business licence.

$$\text{MarriageLic} = 126.250 + 2.302 * \text{BusinessLic} \tag{6}$$

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

Each month, we expect there will be, on average, 126 marriage licenses (intercept), with an additional 2.3, on average, for each business license issued (slope).

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

	6 %	94 %
(Intercept)	-432.90	685.40
BusinessLic	1.21	3.40

(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.

Table 2: 95% confidence interval

Predicted	Lower Bound	Upper Bound
1392.23	515.55	2268.90

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

You cannot do this without first regressing X onto Y

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.

There are a few problems here

(a) Linear regression requires the observations to be independent. This is time series data (the months are the time periods), and are not independent. The earlier work showing a cyclical rise and fall in the marriage license data illustrated this.

(b) Figure 2 is a stacked bar plot, aggregating the data from Table 1 across years into a single month. In other words, the plot illustrates the column sums of this table. For the months of Jan - Aug, there are five numbers contributing to the sum. For Sep - Dec, there are four numbers contributing to the sum (there is no Sep - Dec 2015). Hence, the figure is misleading.

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.

```
#### A1 ####
### Marriage Licenses ###
require(plyr)
wed <- read.csv("marriage.csv", head=T)
str(wed)

wed <- within(wed, {
  Year <- factor(gsub("-.*", "", TIME_PERIOD))
  Month <- factor(month.abb[as.numeric(gsub(".*-", "",
    ↪ TIME_PERIOD))], levels = month.abb)
})
wed.agg <- ddply(wed, .(Year, Month), summarize,
  ↪ MarriageLic = sum(MARRIAGE_LICENSES))

# 1D table
```

```
wed.tab1 <- with(wed, tapply(MARRIAGE_LICENSES, CIVIC_
    ↪ CENTRE, sum)); wed.tab1

# Barplot
wed.bar <- arrange(ddply(wed, "CIVIC_CENTRE", summarize
    ↪ , MarriageLic = sum(MARRIAGE_LICENSES)), -
    ↪ MarriageLic)
with(wed.bar, barplot(MarriageLic, names.arg=CIVIC_
    ↪ CENTRE, main="Marriage license count by Region"))

# 2D table
wed.tab2 <- with(wed, tapply(MARRIAGE_LICENSES, list(
    ↪ Year, Month), sum)); wed.tab2

# Stacked bar
wed.bar2 <- ddply(wed, c("CIVIC_CENTRE", "Month"),
    ↪ summarize, MarriageLic = sum(MARRIAGE_LICENSES))
require(ggplot2) # I like these plots better than base
    ↪ R graphics. Also easier to code
ggplot(wed.bar2, aes(x= Month, y= MarriageLic, fill=
    ↪ CIVIC_CENTRE)) + geom_bar(stat="identity") + labs
    ↪ (y= "Marriage Licenses")
# barplot(xtabs(MarriageLic ~ CIVIC_CENTRE + Month,
    ↪ data=wed.bar))
# barplot(with(wed, tapply(MARRIAGE_LICENSES, list(
    ↪ CIVIC_CENTRE, Month), sum)), col=1:4)
# legend("topright", legend= levels(wed$CIVIC_CENTRE),
    ↪ fill= 1:4, title= "Civic Centres")

### Business Licenses ###
biz <- read.csv("businessLicences.csv", head=T,
    ↪ stringsAsFactors = F)
str(biz)

require(lubridate) # Nice package for dealing with
    ↪ dates, although you don't need it
biz <- within(biz, {
```

```
#    Date <- as.Date(Issued, format = ifelse(nchar(
    ↪ Issued) > 8, "%d/%m/%Y", "%d/%m/%y"))
  Date <- as.Date(Issued, format = "%d/%m/%y")
  Year <- factor(year(Date))
  Month <- factor(month.abb[month(Date)], levels =
     ↪ month.abb)
})
biz <- subset(biz, Year %in% unique(wed$Year), select=c
   ↪ (Month, Year)) # Doing it here to save RAM, but
   ↪ could be done in join instead
biz <- droplevels(biz) # Gets rid of unused factor
   ↪ levels
biz.agg <- ddply(biz, .(Year, Month), summarize,
   ↪ BusinessLic = length(Month))

### Regression with combined data ###
licenses <- join(wed.agg, biz.agg, by = c("Year", "
   ↪ Month"))
rm(wed, wed.agg, biz, biz.agg, wed.bar, wed.bar2, wed.
   ↪ tab1, wed.tab2)

with(licenses, plot(BusinessLic, MarriageLic))
fit <- lm(MarriageLic ~ BusinessLic, data=licenses)
summary(fit)
confint(fit, level = 0.88)
newData <- data.frame(BusinessLic = 550)
predict(fit, newData, interval="prediction")
```