STA302 - Assignment 1

Minh Le Hoang - 999 01 9930

Oct 15 14:05

Part A

1.)

(a) $Y_i^{'} = \frac{(Y_i - a)}{c} \iff Y_i = Y_i^{'} c + a$

$X_i^{'} = \frac{(X_i - d)}{f} \iff X_i = X_i^{'} f + d$

$\overline{Y'} = \frac{\sum_{i=1}^{n} Y_i^{'}}{n} = \frac{\sum_{i=1}^{n} \left[\frac{(Y_i - a)}{c}\right]}{n} = \frac{\sum_{i=1}^{n} Y_i}{c*n} - \frac{a}{c} = \frac{\overline{Y}}{c} - \frac{a}{c} = \frac{\overline{Y} - a}{c}$

$\overline{X'} = \frac{\sum_{i=1}^{n} X_i^{'}}{n} = \frac{\sum_{i=1}^{n} \left[\frac{(X_i - d)}{f}\right]}{n} = \frac{\sum_{i=1}^{n} X_i}{f*n} - \frac{d}{f} = \frac{\overline{X}}{f} - \frac{d}{f} = \frac{\overline{X} - d}{f}$

$b_1^{'} = \frac{\sum_{i=1}^{n} \left[\left(X_i^{'} - \overline{X}\right)\left(Y_i^{'} - \overline{Y}\right)\right]}{\sum_{i=1}^{n} \left(X_i^{'} - \overline{X}\right)^2} = \frac{\sum_{i=1}^{n} \left[\left(X_i^{'} - \overline{X'}\right)\left(Y_i^{'} - \overline{Y'}\right)\right]}{\sum_{i=1}^{n} \left(X_i^{'} - \overline{X'}\right)^2} = \frac{\sum_{i=1}^{n} \left[\left(\frac{(X_i - d)}{f} - \frac{\overline{X}}{f} + \frac{d}{f}\right)\left(\frac{(Y_i - a)}{c} - \frac{\overline{Y}}{c} + \frac{a}{c}\right)\right]}{\sum_{i=1}^{n} \left(\frac{(X_i - d)}{f} - \frac{\overline{X}}{f} + \frac{d}{f}\right)^2}$

$= \frac{\sum_{i=1}^{n} \left[\left(\frac{\left(X_i - d - \overline{X} + d\right)}{f}\right)\left(\frac{\left(Y_i - a - \overline{Y} + a\right)}{c}\right)\right]}{\sum_{i=1}^{n} \left(\frac{\left(X_i - d - \overline{X} + d\right)}{f}\right)^2} = \frac{\sum_{i=1}^{n} \left[\left(\frac{\left(X_i - \overline{X}\right)}{f}\right)\left(\frac{\left(Y_i - \overline{Y}\right)}{c}\right)\right]}{\sum_{i=1}^{n} \left(\frac{\left(X_i - \overline{X}\right)}{f}\right)^2} = \frac{\sum_{i=1}^{n} \left[\frac{1}{fc}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]}{\sum_{i=1}^{n} \frac{1}{f^2}\left(X_i - \overline{X}\right)^2}$

$= \frac{f}{c} \frac{\sum_{i=1}^{n} \left[\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{f}{c} * b_1$

(b) $b_0^{'} = \overline{Y'} - b_1^{'} \overline{X'} = \frac{\overline{Y}}{c} - \frac{a}{c} - \left(\frac{f}{c} * b_1\right) * \left(\frac{\overline{X}}{f} - \frac{d}{f}\right) = \frac{\overline{Y}}{c} - \frac{a}{c} - \left(\frac{b_1}{c}\right) * \left(\overline{X} - d\right) = \frac{\overline{Y}}{c} - \frac{a}{c} - \left(\frac{b_1 \overline{X}}{c}\right) + \left(\frac{b_1 d}{c}\right)$

$= \frac{\overline{Y}}{c} - \frac{a}{c} - \left(\frac{b_1 \overline{X}}{c}\right) + \left(\frac{b_1 d}{c}\right) = \left(\frac{b_1 d}{c}\right) - \frac{a}{c} + \left(\frac{\overline{Y} - b_1 \overline{X}}{c}\right) = \left(\frac{b_1 d}{c}\right) - \frac{a}{c} + \left(\frac{b_0}{c}\right) = \frac{d b_1 - a - + b_0}{c}$

(c) $R'^2 = \frac{(SS_{X'Y'})^2}{SS_{X'} SS_{Y'}} = \frac{\left[\sum_{i=1}^{n} \left(X_i' - \overline{X'}\right)\left(Y_i' - \overline{Y'}\right)\right]^2}{\left[\sum_{i=1}^{n} \left(X_i' - \overline{X'}\right)^2\right]\left[\sum_{i=1}^{n} \left(Y_i' - \overline{Y'}\right)^2\right]} = \frac{\left[\sum_{i=1}^{n} \left(\frac{(X_i - d)}{f} - \frac{\overline{X} - d}{f}\right)\left(\frac{(Y_i - a)}{c} - \frac{\overline{Y} - a}{c}\right)\right]^2}{\left[\sum_{i=1}^{n} \left(\frac{(X_i - d)}{f} - \frac{\overline{X} - d}{f}\right)^2\right]\left[\sum_{i=1}^{n} \left(\frac{(Y_i - a)}{c} - \frac{\overline{Y} - a}{c}\right)^2\right]}$

$\frac{\left[\sum_{i=1}^{n} \left(\frac{(X_i - d)}{f} - \frac{\overline{X} - d}{f}\right)\left(\frac{(Y_i - a)}{c} - \frac{\overline{Y} - a}{c}\right)\right]^2}{\left[\sum_{i=1}^{n} \left(\frac{(X_i - d)}{f} - \frac{\overline{X} - d}{f}\right)^2\right]\left[\sum_{i=1}^{n} \left(\frac{(Y_i - a)}{c} - \frac{\overline{Y} - a}{c}\right)^2\right]} = \frac{\left[\sum_{i=1}^{n} \left(\frac{\left(X_i - \overline{X}\right)}{f}\right)\left(\frac{\left(Y_i - \overline{Y}\right)}{c}\right)\right]^2}{\left[\sum_{i=1}^{n} \left(\frac{\left(X_i - \overline{X}\right)}{f}\right)^2\right]\left[\sum_{i=1}^{n} \left(\frac{\left(Y_i - \overline{Y}\right)}{c}\right)^2\right]}$

$= \frac{\left[\sum_{i=1}^{n} \frac{1}{fc}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]^2}{\left[\sum_{i=1}^{n} \frac{1}{f^2}\left(X_i - \overline{X}\right)^2\right]\left[\sum_{i=1}^{n} \frac{1}{c^2}\left(Y_i - \overline{Y}\right)^2\right]} = \frac{\left(\frac{1}{fc}\right)^2 \left[\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]^2}{\frac{1}{f^2}\left[\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right]\frac{1}{c^2}\left[\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2\right]} = \frac{\left[\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]^2}{\left[\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right]\left[\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2\right]}.$

Then $R'^2 = R^2$

(d) $\widehat{Y_i'} = b_0' + b_1' X_i' = \frac{d b_1 - a - b_0}{c} + \frac{f}{c} * b_1 * \frac{(X_i - d)}{f} = \frac{d b_1 - a - b_0}{c} + \frac{b_1 (X_i - d)}{c} = \frac{d b_1 - a - b_0}{c} + \frac{b_1 X_i - b_1 d}{c}$

$\widehat{Y_i'} = \frac{d b_1 - a - b_0 + b_1 X_i - b_1 d}{c} = \frac{b_1 X_i + b_0 - a}{c} = \frac{\widehat{Y_i} - a}{c}.$

$s^2\left\{b_1'\right\} = \sum_{i=1}^{n} \left(Y_i' - \widehat{Y_i'}\right)^2 = \sum_{i=1}^{n} \left(\frac{(Y_i - a)}{c} - \frac{\widehat{Y_i} - a}{c}\right)^2 = \sum_{i=1}^{n} \left(\frac{\left(Y_i - \widehat{Y_i}\right)}{c}\right)^2 = \frac{1}{c^2} \sum_{i=1}^{n} \left(Y_i - \widehat{Y_i}\right)$

Then $s^2\{b_1'\} = \frac{1}{c^2} s^2\{b_1\}$.

Then

$s\{b_1'\} = \sqrt{\frac{s^2\{b_1'\}}{\sum_{i=1}^n \frac{1}{f}(X_i' - \overline{X'})^2}} = \sqrt{\frac{\frac{1}{c^2} s^2\{b_1\}}{\sum_{i=1}^n \frac{1}{f^2}(X_i' - \overline{X'})^2}} = \frac{f}{c} * s\{b_1\}$

With $E[b_1'] = E\left[\frac{f}{c} * b_1\right] = \frac{f}{c} * E[b_1] = \frac{f}{c} * \beta_1$ Then $\beta'_1 = \frac{f}{c} * \beta_1$

$\frac{b_1' - \beta_1'}{s\{b_1'\}} = \frac{\frac{f}{c} * b_1 - \frac{f}{c} * \beta_1}{\frac{f}{c} * s\{b_1\}} = \frac{b_1 - \beta_1}{s\{b_1\}}$

Then the transformations above do not affect inference for β_1, the slope parameter

2.)

We have $SSE_{b_0;b_1} = \sum_{i=1}^n \left(Y_i - \widehat{Y_{(b_0;b_1)i}}\right)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$

Let $b_1' = b_1 + \Delta_1$ and $b_0' = b_0 + \Delta_0$ with $\Delta_1, \Delta_0 \neq 0$

Let $SSE' = \sum_{i=1}^n \left(Y_i - \widehat{Y_i'}\right)^2 = \sum_{i=1}^n (Y_i - b_0' - b_1' X_i)^2$

$= \sum_{i=1}^n (Y_i - (b_0 + \Delta_0) - (b_1 + \Delta_1) X_i)^2$

$= \sum_{i=1}^n (Y_i - b_0 - \Delta_0 - b_1 X_i - \Delta_1 X_i)^2$

$= \sum_{i=1}^n (Y_i - b_0 - b_1 X_i - \Delta_1 X_i - \Delta_0)^2$

$= \sum_{i=1}^n (Y_i - b_0 - b_1 X_i - (\Delta_1 X_i + \Delta_0))^2$

$= \sum_{i=1}^n \left((Y_i - b_0 - b_1 X_i)^2 + (\Delta_1 X_i + \Delta_0)^2 - 2(Y_i - b_0 - b_1 X_i)(\Delta_1 X_i + \Delta_0)\right)$

$= SSE_{b_0;b_1} + \sum_{i=1}^n (\Delta_1 X_i + \Delta_0)^2 - 2\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)(\Delta_1 X_i + \Delta_0)$

$= SSE_{b_0;b_1} + \sum_{i=1}^n (\Delta_1 X_i + \Delta_0)^2 - 2\sum_{i=1}^n \left(e_{(b_0;b_1)i}\right)(\Delta_1 X_i + \Delta_0).$

$= SSE_{b_0;b_1} + \sum_{i=1}^n (\Delta_1 X_i + \Delta_0)^2 - 2\sum_{i=1}^n \Delta_1 e_{(b_0;b_1)i} X_i - 2\sum_{i=1}^n e_{(b_0;b_1)i}\Delta_0$

$= SSE_{b_0;b_1} + \sum_{i=1}^n (\Delta_1 X_i + \Delta_0)^2 - 2\Delta_0 \sum_{i=1}^n e_{(b_0;b_1)i} X_i - 2\Delta_0 \sum_{i=1}^n e_{(b_0;b_1)i}$

$= SSE_{b_0;b_1} + \sum_{i=1}^n (\Delta_1 X_i + \Delta_0)^2.$

Because $\sum_{i=1}^n (\Delta_1 X_i + \Delta_0)^2 \geq 0$

Then $SSE_{b_0;b_1} + \sum_{i=1}^n \left(\Delta\left[X_i - \overline{X}\right]\right)^2 \geq SSE_{b_0;b_1} \iff SSE' \geq SSE_{b_0;b_1}$

Then $SSE_{b_0;b_1}$ is the minimum

Then SSE is minimized with b_1

Part B

1.)

(a) .

```
'data.frame': 224 obs. of 3 variables:
$ TIME_PERIOD : Factor w/ 56 levels "2011-01","2011-02",..: 1 1 1 1 2 2 2 2 3 3 ...
$ CIVIC_CENTRE : Factor w/ 4 levels "ET","NY","SC",..: 1 2 3 4 1 2 3 4 1 2 ...
$ MARRIAGE_LICENSES: int 80 136 159 367 109 150 154 383 177 231 ...
NULL
```

(b) .

(c) .

(d) .

```
_ TIME_PERIOD CIVIC_CENTRE MARRIAGE_LICENSES
1 2011-01 ET 80
2 2011-01 NY 136
3 2011-01 SC 159
4 2011-01 TO 367
5 2011-02 ET 109
6 2011-02 NY 150
......
```

(e) .

```
_ TIME_PERIOD Month Year MARRIAGE_LICENSES
1 2011-04 Apr 2011 1376
2 2011-08 Aug 2011 1933
3 2011-12 Dec 2011 785
4 2011-02 Feb 2011 796
5 2011-01 Jan 2011 742
6 2011-07 Jul 2011 1943
......
```

2.)

(a) .

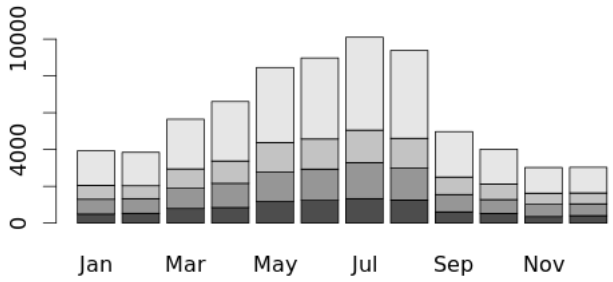
```
ET NY SC TO
9568 14028 13354 35060
```

(b) .

(c) .

```
_____ 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 NA NA NA NA
```

(d) .



3.)

(a) .

```
'data.frame': 143618 obs. of 16 variables:
$ Category : chr "TAXICAB BROKER" "TAXICAB BROKER" "TAXICAB BROKER" "TAXICAB BRO-
KER" ...
$ Licence.No. : chr "B032020976" "B032025475" "B032022604" "B030000236" ...
$ Operating.Name : chr "EMERALD TAXI" "PREMIER TAXI" "A TORONTO TAXI" "KIPLING KAB"
...
$ Issued : chr "24/03/95" "24/06/96" "17/08/95" "03/03/88" ...
$ Client.Name : chr "NAEGELI,LEO/PATEL,CHHOTUBHAI" "PATEL, CHHOTUBHAI" ...
$ Business.Phone : num NA NA NA NA NA NA NA NA NA NA ...
$ Business.Phone.Ext. : chr "" "" "" "" ...
$ Licence.Address.Line.1 : chr "420 ORMONT DR" "131 HAWKSHEAD CRES" "99 MORBANK DR" ...
$ Licence.Address.Line.2 : chr "TORONTO, ON" "TORONTO, ON" "TORONTO, ON" "TORONTO,
ON" ...
$ Licence.Address.Line.3 : chr "M9L 1N9" "M1W 2Z4" "M1V 2M1" "M9W 2Z2" ...
$ Conditions : chr "" "" "" "" ...
$ Free.Form.Conditions.Line.1: chr "" "" "" "" ...
$ Free.Form.Conditions.Line.2: chr "" "" "" "" ...
$ Plate.No. : chr "" "" "" "" ...
$ Endorsements : chr "TAXICAB BROKER;" "TAXICAB BROKER;" "TAXICAB BROKER;" ...
$ Cancel.Date : chr "01/05/97" "20/05/97" "31/12/97" "08/06/98" ...
NULL
```

(b) .

```
'data.frame': 143618 obs. of 18 variables:
$ Category : chr "TAXICAB BROKER" "TAXICAB BROKER" "TAXICAB BROKER" "TAXICAB BRO-
KER" ...
$ Licence.No. : chr "B032020976" "B032025475" "B032022604" "B030000236" ...
$ Operating.Name : chr "EMERALD TAXI" "PREMIER TAXI" "A TORONTO TAXI" "KIPLING KAB"
...
$ Issued : Date, format: "1995-03-24" "1996-06-24" "1995-08-17" ...
$ Client.Name : chr "NAEGELI,LEO/PATEL,CHHOTUBHAI" "PATEL, CHHOTUBHAI" ...
$ Business.Phone : num NA NA NA NA NA NA NA NA NA NA ...
$ Business.Phone.Ext. : chr "" "" "" "" ...
$ Licence.Address.Line.1 : chr "420 ORMONT DR" "131 HAWKSHEAD CRES" "99 MORBANK DR" ...
$ Licence.Address.Line.2 : chr "TORONTO, ON" "TORONTO, ON" "TORONTO, ON" "TORONTO,
ON" ...
$ Licence.Address.Line.3 : chr "M9L 1N9" "M1W 2Z4" "M1V 2M1" "M9W 2Z2" ...
$ Conditions : chr "" "" "" "" ...
$ Free.Form.Conditions.Line.1: chr "" "" "" "" ... $ Free.Form.Conditions.Line.2: chr "" "" "" "" ...
$ Plate.No. : chr "" "" "" "" ...
$ Endorsements : chr "TAXICAB BROKER;" "TAXICAB BROKER;" "TAXICAB BROKER;" ...
$ Cancel.Date : chr "01/05/97" "20/05/97" "31/12/97" "08/06/98" ...
$ Month : chr "Mar" "Jun" "Aug" "Mar" ...
$ Year : num 1995 1996 1995 1988 1994 ...
NULL
```

(c) .

```
__ Month Year YYYYMM
80 Jan 2013 201301
81 May 2012 201205
82 Aug 2012 201208
84 Sep 2012 201209
85 Oct 2012 201210
86 Feb 2011 201102
......
```

(d) .

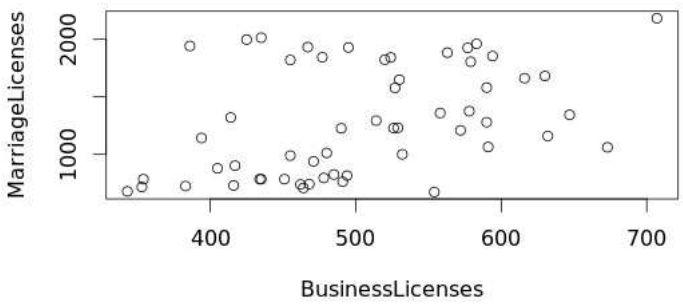
```
_____ Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2011 468 478 572 578 563 520 386 467 414 480 494 451
2012 417 405 490 529 530 524 435 495 394 591 485 354
2013 491 383 455 558 590 527 425 455 526 471 464 343
2014 435 353 673 514 630 579 583 477 590 532 554 434
2015 462 416 632 647 616 577 707 594 NA NA NA NA
```

Part C

1.)

```
____ BusinessLicenses MarriageLicenses
[1,] 468 742
[2,] 417 902
[3,] 491 763
[4,] 435 785
[5,] 462 739
[6,] 478 796
....
```

2.)



We can still do regression but there will be a lot of errors

3.)

   (a) .
   Marriage Licenses =126.250
   BusinessLicenses = 2.302

   (b) .
   Our interception presents total number of new Marriage Licenses within a month of there is no new Business Licenses within a month
   Our slope present number of new Marriage Licenses within a month for every new Marriage Licenses within a month

   (c) .
   88% Confidence Interval for:
   - Intercept: $[-432.902886; 685.40321]$
   - Slope: $[1.206791; 3.39676]$

   (d) .
   The predicted number of marriage licenses issued if there will be 550 business licenses issued: 1277.138, with $[402.8037; 2151.472]$ is 95% prediction interval

   (e) Cannot do this without first regressing Business Licenses (X) onto Marriage Licenses (Y)

4.)

Because data points are a month of a certain year, months within a year might have correlation with each other. This makes our data point not independent of each other