資訊管理研究所 周秉楠 0653438

統計應用方法:

Homework #3: Analysis of Pizza Data

1. *"Using coal to bake pizzas yields different ratings with those baked by using gas or wood"*. We wish to verify this statement by providing some statistical evidences:
   a. Compute each of the average ratings of the pizzas baked by coal, wood and gas, along with the standard deviations of the ratings. Comment the results. *[hint: you could use codes like* ***pizza[pizza[,"heat"]=="Coal", ratings]*** ***OR*** ***sapply()*** *and a self-defined function to do so]*

Code & Result:

```
pizza <- read.table(file = "pizza2.txt" , header = T)

###### 1.a ########
# read the pizza data and rename to pizza
coal_rating = pizza[pizza[,"heat"]=="Coal","rating"]
# we can also use pizza[pizza[,3]=="Coal","rating"] to get the rating of pizza using Coal
wood_rating = pizza[pizza[,"heat"]=="Wood","rating"]
gas_rating = pizza[pizza[,"heat"]=="Gas","rating"]
#-------------Using by Coal -----------------
mean(coal_rating)
sd(coal_rating)
#-------------Using by Wood -----------------
mean(wood_rating)
sd(wood_rating)
#-------------Using by Gas -----------------
mean(gas_rating)
sd(gas_rating)
```

```
> #-------------Using by Coal -----------------
> mean(coal_rating)
[1] 4.688824
> sd(coal_rating)
[1] 0.4867479
> #-------------Using by Wood -----------------
> mean(wood_rating)
[1] 3.8764
> sd(wood_rating)
[1] 1.537248
> #-------------Using by Gas -----------------
> mean(gas_rating)
[1] 2.961013
> sd(gas_rating)
[1] 1.817251
```

Comment:

以上結果可看出，不同的加熱方式，其 rating 的平均值、標準差皆不同。

另外，由標準差可以看到，用 Coal 加熱標準差最小，表示 Coal 加熱的 Pizza 其 rating 差距不大，亦可推測用 Coal 加熱的 Pizza 品質較其他加熱方法穩定，且 Coal 的 rating 平均數最高，更能表示整體品質應該較高。

b. Perform an ANOVA test to find out if the ratings of the pizzas baked by different heat sources are equal in average. Comment the results.

Code & Result:

```
###### 1.b ########
DataFrame <- data.frame( response_data = c(coal_rating,wood_rating,gas_rating),
Site = factor(rep(c("coal_rating","wood_rating","gas_rating"),times
                  = c(length(coal_rating),length(wood_rating),length(gas_rating))))
)
fm1 <- aov(response_data~Site , data=DataFrame)
anova(fm1)
```

```
> DataFrame <- data.frame( response_data = c(coal_rating,wood_rating,gas_rating),
+ Site = factor(rep(c("coal_rating","wood_rating","gas_rating"),times
+                   = c(length(coal_rating),length(wood_rating),length(gas_rating))))
+ )
> fm1 <- aov(response_data~Site , data=DataFrame)
> anova(fm1)
Analysis of Variance Table

Response: response_data
           Df Sum Sq Mean Sq F value    Pr(>F)
Site        2  58.04  29.022  9.8749 8.184e-05 ***
Residuals 197 578.98   2.939
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comment:

H0: μ1 = μ2 = μ3…(即母體平均數皆相等)
H1: μ1 != μ2 != μ3…(即母體平均數皆不相等)

利用 ANOVA 分析後，由表格可得知 F 查表自由度為(2,197)
F value 為 9.8749(即 29.022/2.939)
Pr(>F)值為 8.184*10^(-5)，為一趨近於 0 的機率，標示***，
表示非常顯著，因此拒絕掉 H0。
接受 H1 結果為母體平均數皆不相等。

c. Fit a simple linear regression by using **rating** as the response variable and **heat** as the predictor variable. Interpret the estimated regression coefficients and the corresponding p-values.

Code & Result:

```
reg <- lm(rating~heat,data=pizza)
summary(reg)
```

```
> reg <- lm(rating~heat,data=pizza)
> summary(reg)

Call:
lm(formula = rating ~ heat, data = pizza)

Residuals:
   Min     1Q Median     3Q    Max
-3.506 -1.715  0.379  1.562  2.039

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6888     0.4158  11.277  < 2e-16 ***
heatGas      -1.7278     0.4376  -3.948 0.000109 ***
heatWood     -0.8124     0.5389  -1.507 0.133289
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.714 on 197 degrees of freedom
Multiple R-squared:  0.09112,   Adjusted R-squared:  0.08189
F-statistic: 9.875 on 2 and 197 DF,  p-value: 8.184e-05
```

d. Compare and contrast the results in 2a., 2b. and 2c.. In other words, what information are shown from both analyses, *OR* from one analysis, but not from the others?

在 1a.時，蒐集出來的樣本只是母體的一小部分，我們可以先以抽出的樣本來算出標準差、平均數。如 1a.可看出其「樣本平均數」其實差異很大，就可以事先了解到以 1b.方式的 ANOVA 分析會拒絕 H0 假設。(之所以利用 ANOVA 分析，就是想要從樣本資料去判斷兩個母體以上的母體平均數是否有差異)

2. Fit two multiple linear regression by using **rating** as the response variable, and
   a. **heat**, **area** and **cost** as the predictor variables.
   b. **heat_re**, **area** and **cost** as the predictor variables.

Code & Result:

```
37 ###### 2.a ########
38 multi_reg_a <- lm(rating~heat+area+cost,data=pizza)
39 summary(multi_reg_a)
40
41 ###### 2.b ########
42 multi_reg_b <- lm(rating~heat_re+area+cost,data=pizza)
43 summary(multi_reg_b)
```

```
> ###### 2.a ########
> multi_reg_a <- lm(rating~heat+area+cost,data=pizza)
> summary(multi_reg_a)

Call:
lm(formula = rating ~ heat + area + cost, data = pizza)

Residuals:
     Min       1Q   Median       3Q      Max
-1.98864 -0.52516  0.00599  0.51428  1.92332

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.72260    0.34461   2.097  0.03731 *
heatGas        -1.59555    0.20526  -7.773 4.52e-13 ***
heatWood       -0.45753    0.26056  -1.756  0.08069 .
areaEVillage    4.17970    0.24628  16.971  < 2e-16 ***
areaLES         2.37294    0.26106   9.089  < 2e-16 ***
areaLittleItaly 0.78700    0.25268   3.115  0.00212 **
areaSoHo        3.65362    0.24498  14.914  < 2e-16 ***
cost            0.43865    0.06613   6.633 3.26e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7957 on 192 degrees of freedom
Multiple R-squared:  0.8092,    Adjusted R-squared:  0.8022
F-statistic: 116.3 on 7 and 192 DF,  p-value: < 2.2e-16
```

```
> ###### 2.b ########
> multi_reg_b <- lm(rating~heat_re+area+cost,data=pizza)
> summary(multi_reg_b)

Call:
lm(formula = rating ~ heat_re + area + cost, data = pizza)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97759 -0.51011 -0.02969  0.52497  2.15583

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.96212    0.31668   3.038  0.00271 **
heat_re        -0.87601    0.09242  -9.479  < 2e-16 ***
areaEVillage    4.10646    0.24378  16.845  < 2e-16 ***
areaLES         2.26091    0.25405   8.900 4.08e-16 ***
areaLittleItaly 0.69163    0.24774   2.792  0.00577 **
areaSoHo        3.54383    0.23768  14.910  < 2e-16 ***
cost            0.44911    0.06618   6.786 1.38e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7997 on 193 degrees of freedom
Multiple R-squared:  0.8062,    Adjusted R-squared:  0.8002
F-statistic: 133.8 on 6 and 193 DF,  p-value: < 2.2e-16
```

Assume that coal-baked pizzas produce the highest ratings, followed by using wood, and then gas, compare the two models. It is not reasonable to not use dummy(indicator) variables in model fitting (as in 2b.), why? Justify your answer by comparing the interpretations of the regression coefficients of **heat** and **heat_re**.

因為 heat_re 直接對應到數字(Coal=0 , Wood=1, Gas=2)，但數字的數值會影響整體的分析數據，即三種加熱法不是以公平的方式去分析，因此 2a.的 heat 利用 dummy 即可避免這種不公平的方式來分析數據，更能準確的預測整體的 rating。

Then, predict the rating for a coal baked pizza that costs $2.50 per slice in LittleItaly and find the corresponding prediction interval using both of the models built in 3a. and 3b.. *[hint: use **predict()**]*

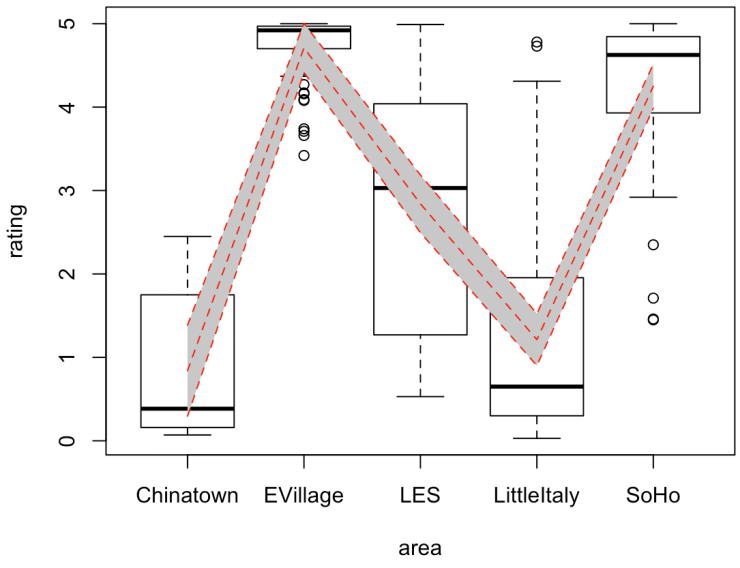
```
> predict_a = predict(multi_reg_a,data.frame("heat"="Coal","area"="LittleItaly","cost"=2.50))
> predict_b = predict(multi_reg_b,data.frame("heat_re"=0,"area"="LittleItaly","cost"=2.50))
> predict_a
       1
2.606232
> predict_b
       1
2.776521
```

3. Construct the 95% t-based confidence intervals for the mean rating for each pizzeria location (**area**). Plot **all** of the intervals in a single plot and briefly comment the results. (*Hint: you could make use of* ***plot(), lines()*** *and* ***points() OR*** *search online*[1] *for some ways to plot confidence intervals.*)

Code & Result:

```
LittleItaly=pizza[pizza[,"area"]=="LittleItaly","rating"]
SoHo=pizza[pizza[,"area"]=="SoHo","rating"]
Chinatown=pizza[pizza[,"area"]=="Chinatown","rating"]
LES=pizza[pizza[,"area"]=="LES","rating"]
EVillage=pizza[pizza[,"area"]=="EVillage","rating"]

#Data2 = data.frame(
#  Y=c(LittleItaly, SoHo, Chinatown,LES,EVillage),
#  Site =factor(rep(c("LittleItaly", "SoHo", "Chinatown", "LES", "EVillage"), times=c(length(Litt
#)

#boxplot(Y ~ Site, data = Data2)

# predicts + interval
newx <- factor(c("Chinatown","EVillage","LES","LittleItaly","SoHo"))
fittttt = lm(rating ~ area, data=pizza)

preds <- predict(fittttt, newdata = data.frame(area=newx),interval = 'confidence', level = 0.95)

#-------

# plot
plot(rating ~ area, data = pizza)
# add fill
polygon(c(rev(newx), newx), c(rev(preds[ ,3]), preds[ ,2]), col = 'grey80', border = NA)
# intervals
lines(newx, preds[ ,1], lty = 'dashed', col = 'red')
lines(newx, preds[ ,2], lty = 'dashed', col = 'red')
lines(newx, preds[ ,3], lty = 'dashed', col = 'red')
```

Plot:



(含盒狀圖，可看出各 Area 的 rating 分佈狀況)

[1] *http://stackoverflow.com/questions/14069629/plotting-confidence-intervals*

Comment:

根據此圖可看出，rating 與 area 之間的關係，圖中以灰色區域來表示以 95%信賴區間。

因為母體參數的信賴區間=點估計量+-抽樣誤差，對應灰色區域中間的虛線極為「點估計量」所連接的線，其他以虛線擴展的上下兩個灰色區塊則為「抽樣誤差」。95%的信賴區間即表示有 95%的機率上面的公式會成立。