**統計應用方法:**

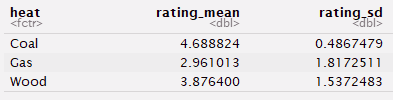
Homework #3: Analysis of Pizza Data 數據所 賴品儒 0656704

Pizza was original invented in Naples, Italy in the early 19th century. It is a kind of flat bread baked by oven and is usually topped with cheese, tomato sauce, meat and vegetables. Pizza has become a common delicacy around the world. Suppose the dataset pizza2.txt contains the data of the pizzas in some US pizzeria, which could furnish some clues for us to make inferences about their ratings. The table below shows some brief information about the data.

|  |  |  |
| --- | --- | --- |
| **Data** | pizza2.txt | |
| **Description** | Data about pizza | |
| **Variables descriptions** | rating | Rating for the pizza |
| cost | Cost per slice |
| heat | Heat source used (Gas/Coal/Wood) |
| brick | The use of brick oven (TRUE/FALSE) |
| area | The location of pizzeria |
| heat\_re | Same as the **heat** variable but it is just numerically coded instead of using strings.  0 – Coal  1 – Wood  2 – Gas |

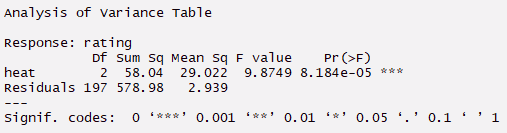
**Tasks:**

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:
   1. 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]*



可以發現用Coal烤出的披薩，平均rating相對其他兩者高，rating的變異程度也比另外兩者小。而用Gas考出之披薩平均rating較低，rating變異程度也較大。

* 1. 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.



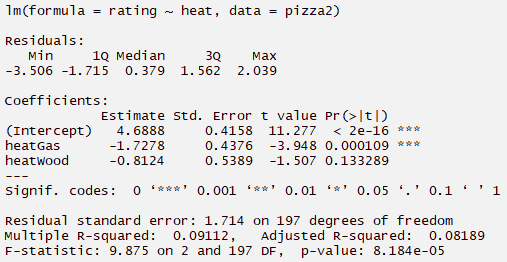
TS：F\*=9.8749

R..R.：Reject H0 if P(F\*<F) < 0.05

p-value P(F>F\*)=8.184e-05 < 0.05 → reject H0

結論：我有足夠證據推論不同烘烤方式披薩的平均rating並非全部都相等。

* 1. 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.



heat為披薩烤的方式，共有三種，分別為：Coal、Gas、Wood，我們需要製造2個dummy variables來表示。

Model：

|  |  |
| --- | --- |
| heat | () |
| Coal (baseline) | (0,0) |
| Gas | (1,0) |
| Wood | (0,1) |

E(Y|()=(0,0))=

E(Y|()=(1,0))=

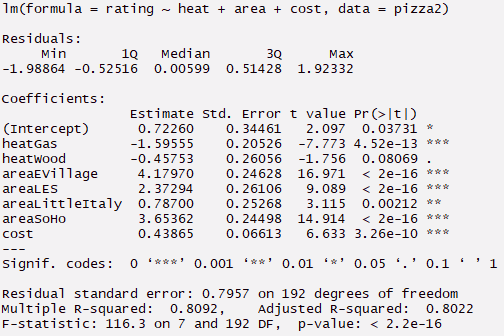
E(Y|()=(0,1))=

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | estimate | 假設檢定 | T.S | p-value | 結論 |
|  | 4.6888 |  | t\*=11.277 | 2e-16<0.05 | Reject H0 |
|  | -1.7278 |  | t\*=-3.948 | 0.000109<0.05 | Reject H0 |
|  | -0.8124 |  | t\*=-1.507 | 0.133289>0.05 | Accept H0 |

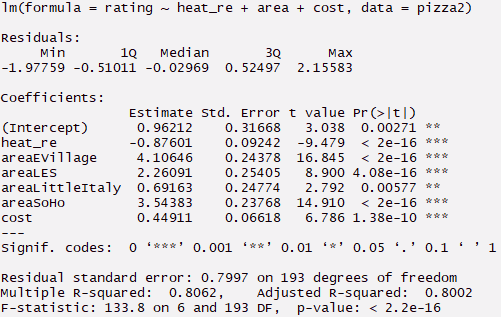
* 1. Compare and contrast the results in 1a., 1b. and 1c.. In other words, what information are shown from both analyses, *OR* from one analysis, but not from the others?

|  |  |
| --- | --- |
| result | compare |
| 1a | 列出不同烤披薩方式 rating的平均和各別rating的變異程度。 |
| 1b | 檢測整個model是否為適當的。  檢測解釋變數是否無法解釋反應變數。 |
| 1c | 檢定各烘烤方式迴歸係數組之顯著性。 |

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



* 1. **heat\_re**, **area** and **cost** as the predictor variables.



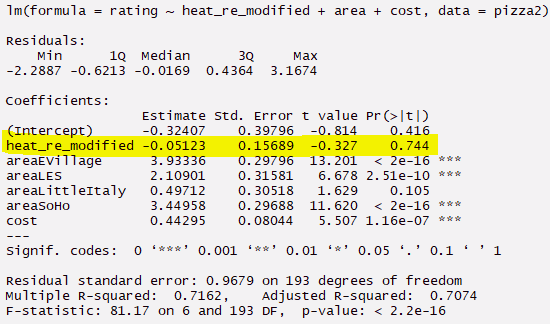
Assume that coal-baked pizzas produce the highest ratings, followed by using wood, and then gas, compare the two models. It is not reasonable to not use dummy(indicator) variables in model fitting (as in 2b.), why? Justify your answer by comparing the interpretations of the regression coefficients of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

heat是指不同烘烤披薩的方式，他代表一種特質，而非量，故使用numerical編碼是不合理的。

2a 使用dummy variables，我們可以看出不同披薩烘烤方式迴歸係數的顯著性，如下：

|  |  |  |  |
| --- | --- | --- | --- |
| heat | Estimate (迴歸係數) | p-value | 結論 |
| Coal | 0.72260 | 0.03731<0.05 \* | 迴歸係數具顯著性 |
| Gas | -1.59555 | 4.52e-13<0.05 \*\*\* | 迴歸係數具顯著性 |
| Wood | -0.45753 | 0.08069>0.05 | 迴歸係數不具顯著性 |

在2b，heat\_re是用numerical方式編碼 (Coal,Wood,Gas)=(0,1,2)，其迴歸係數為，這代表在其他解釋變數固定下，heat\_re增加1，平均rating就會減少0.87601，也就是Coal變成Wood平均rating會降0.87601，Wood變Gas也是，已知三種rating由高到低排序為Coal>Wood>Gas，可看出heat\_re的迴歸係數具顯著性，但若此時編碼改成(Coal,Gas,Word) = (0,1,2) 就會發生問題，heat\_re\_modified的迴歸係數不具顯著性，如下：



由此可知使用dummy variables較合理，他代表特質而非數值多寡。

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 2a. and 2b.. *[hint: use* ***predict()****]*

(default 0.95 confidence level)

new\_data：(heat=”Coal”, area=” LittleItaly”, cost=2.50, heat\_re=0)

**Model\_2a**



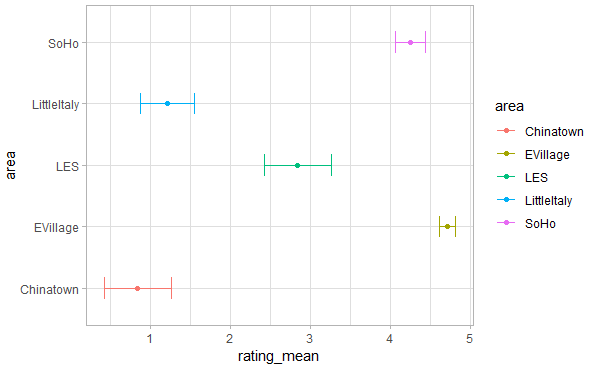
在95%信心水準下，new\_data的rating預測區間介於0.9747882和4.237676之間

**Model\_2b**



在95%信心水準下，new\_data的rating預測區間介於1.14876和4.404281之間

1. 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]](#footnote-1) for some ways to plot confidence intervals.*)

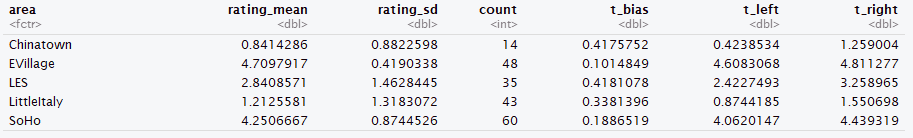


在EVillage的披薩有最高rating平均，在95%信心水準下，大約落在4.6和4.8之間。

其次為SoHo、LES、LittleItaly，而Chinatown的披薩rating平均最低，在95%信心水準下，大約落在0.42和1.26之間。

詳細比對如下：

(其中t\_left和t\_right為區間下界和上界)



Code：<https://github.com/ayamisea/Applied-Methods-in-Statistics/blob/master/homework/HW03/HW03.Rmd>

1. [*http://stackoverflow.com/questions/14069629/plotting-confidence-intervals*](http://stackoverflow.com/questions/14069629/plotting-confidence-intervals) [↑](#footnote-ref-1)