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:
    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]*

heat <fctr>	rating_mean <dbl>	rating_sd <dbl>
Coal	4.688824	0.4867479
Gas	2.961013	1.8172511
Wood	3.876400	1.5372483

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。

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.

```
Analysis of Variance Table

Response: rating
           Df Sum Sq Mean Sq F value    Pr(>F)
heat        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
```

H_0：$\mu_{Coal} = \mu_{Gas} = \mu_{Wood}$ vs. H_1：至少有一烘烤方式平均 rating μ 與其他不相等

TS：F*=9.8749

R..R.：Reject H_0 if P(F*<F) < 0.05

p-value P(F>F*)=8.184e-05 < 0.05 → reject H_0

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。

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.

```
lm(formula = rating ~ heat, data = pizza2)

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
```

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：$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$，其中假設 $\varepsilon \sim^{iid} N(0, \sigma^2)$

heat	(X_1, X_2)
Coal (baseline)	(0,0)
Gas	(1,0)
Wood	(0,1)

$E(Y|(X_1, X_2)=(0,0))=\beta_0$ (用 Coal 烤的披薩平均 rating)

$E(Y|(X_1, X_2)=(1,0))=\beta_0+\beta_1$ (用 Gas 烤的披薩平均 rating)

$E(Y|(X_1, X_2)=(0,1))=\beta_0+\beta_2$ (用 Wood 烤的披薩平均 rating)

	estimate	假設檢定	T.S	p-value	結論
β_0	4.6888	$H_0: \beta_0 = 0$	t*=11.277	2e-16<0.05	Reject H_0 β_0 *具顯著性*
β_1	-1.7278	$H_0: \beta_1 = 0$	t*=-3.948	0.000109<0.05	Reject H_0 β_1 *具顯著性*
β_2	-0.8124	$H_0: \beta_2 = 0$	t*=-1.507	0.133289>0.05	Accept H_0 β_2 *不具顯著性*

d. 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。

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

```
lm(formula = rating ~ heat + area + cost, data = pizza2)

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
```

    b. **heat_re**, **area** and **cost** as the predictor variables.

```
lm(formula = rating ~ heat_re + area + cost, data = pizza2)

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**.

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 −0.87601，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：

```
lm(formula = rating ~ heat_re_modified + area + cost, data = pizza2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2887 -0.6213 -0.0169  0.4364  3.1674

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.32407    0.39796  -0.814    0.416
heat_re_modified -0.05123    0.15689  -0.327    0.744
areaEVillage      3.93336    0.29796  13.201  < 2e-16 ***
areaLES           2.10901    0.31581   6.678 2.51e-10 ***
areaLittleItaly   0.49712    0.30518   1.629    0.105
areaSoHo          3.44958    0.29688  11.620  < 2e-16 ***
cost              0.44295    0.08044   5.507 1.16e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9679 on 193 degrees of freedom
Multiple R-squared:  0.7162,    Adjusted R-squared:  0.7074
F-statistic: 81.17 on 6 and 193 DF,  p-value: < 2.2e-16
```

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

```
       fit       lwr      upr
1 2.606232 0.9747882 4.237676
```

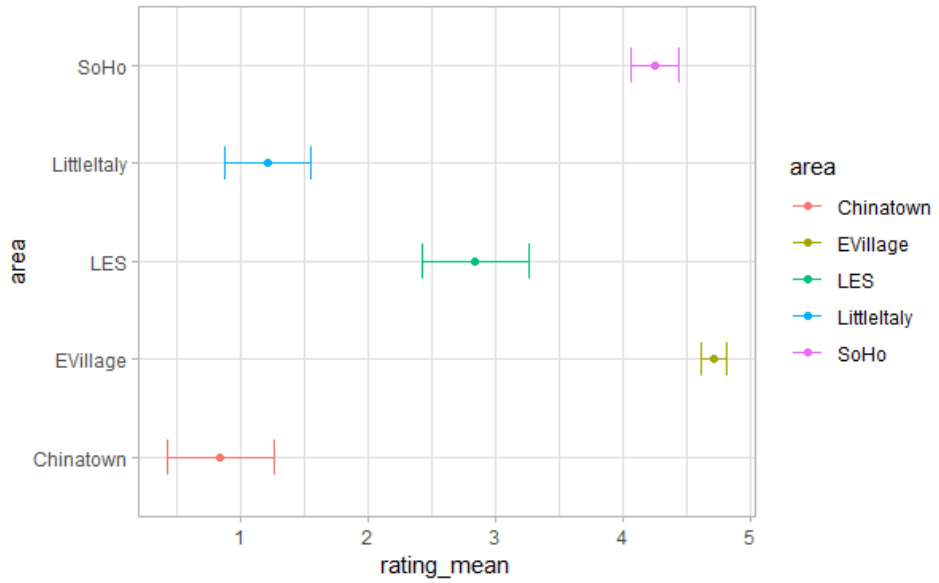
在 95%信心水準下，new_data 的 rating 預測區間介於 0.9747882 和 4.237676 之間

Model_2b

```
       fit      lwr      upr
1 2.776521 1.14876 4.404281
```

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

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.*)



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。

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

詳細比對如下：

（其中 t_left 和 t_right 為區間下界和上界）

area <fctr>	rating_mean <dbl>	rating_sd <dbl>	count <int>	t_bias <dbl>	t_left <dbl>	t_right <dbl>
Chinatown	0.8414286	0.8822598	14	0.4175752	0.4238534	1.259004
EVillage	4.7097917	0.4190338	48	0.1014849	4.6083068	4.811277
LES	2.8408571	1.4628445	35	0.4181078	2.4227493	3.258965
LittleItaly	1.2125581	1.3183072	43	0.3381396	0.8744185	1.550698
SoHo	4.2506667	0.8744526	60	0.1886519	4.0620147	4.439319

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