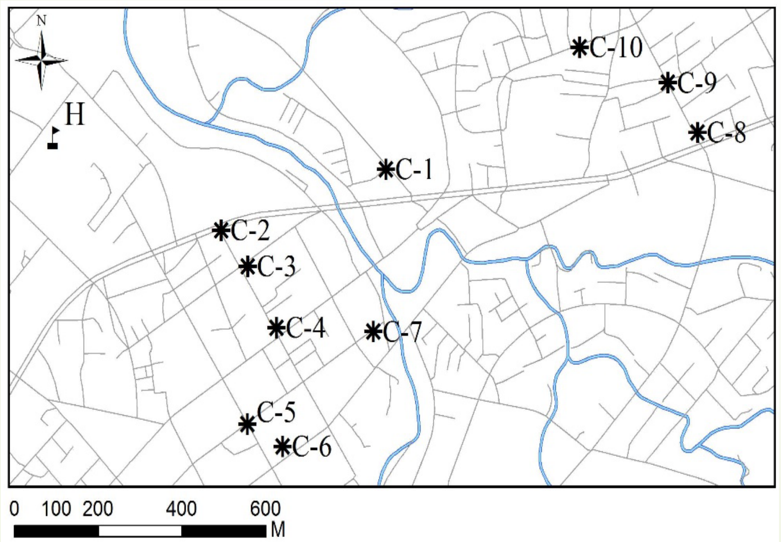
Dataset integration for community, outdoor, indoor, and personal source evaluations (Solutions of Exam 2)

SC Candice Lung & WC Vincent Wang

Advanced Institute on Hi-ASAP (2021)
Academia Sinica, Taiwan

Exam 2 (1/3)

- 1. **Input data:** Hi-ASAP_exam2.csv
- 2. **Ten street-level sites monitor air pollution sources as the table.**
- 3. **Some pollution sources have specific business hours as follows**:
  - Market: 10 am - 8 pm (including 8 pm)
  - Vendor: 1 pm – 9 pm (including 9 pm)
  - Gas station: 8 am – 7 pm (including 7 pm)

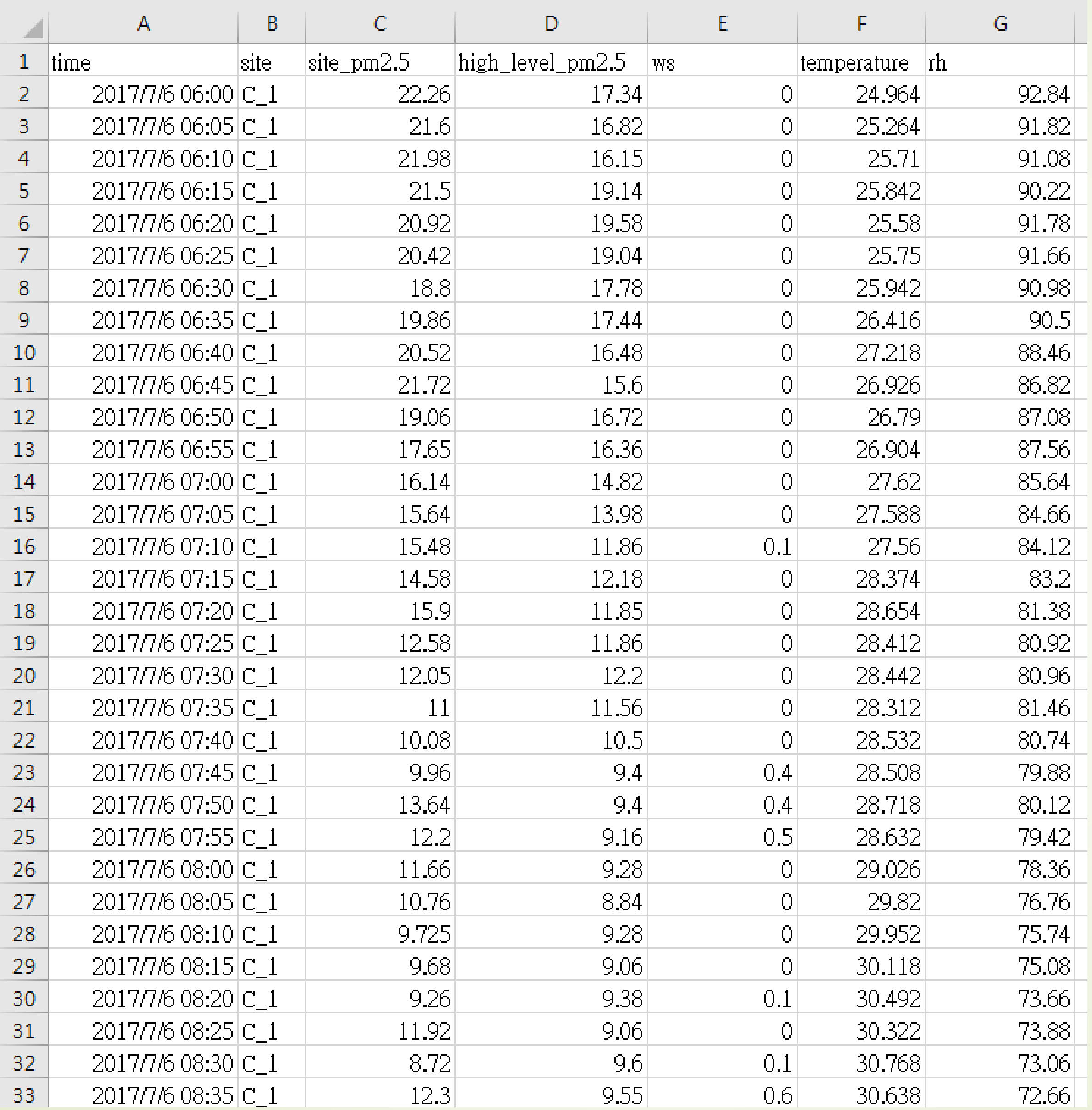


Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles

Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Exam 2 (2/3): Hi-ASAP_exam2.csv



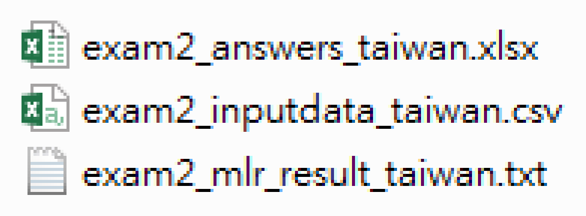
	A	B	C	D	E	F	G
1	time	site	site_pm2.5	high_level_pm2.5	ws	temperature	rh
2	2017/7/6 06:00	C_1	22.26	17.34	0	24.964	92.84
3	2017/7/6 06:05	C_1	21.6	16.82	0	25.264	91.82
4	2017/7/6 06:10	C_1	21.98	16.15	0	25.71	91.08
5	2017/7/6 06:15	C_1	21.5	19.14	0	25.842	90.22
6	2017/7/6 06:20	C_1	20.92	19.58	0	25.58	91.78
7	2017/7/6 06:25	C_1	20.42	19.04	0	25.75	91.66
8	2017/7/6 06:30	C_1	18.8	17.78	0	25.942	90.98
9	2017/7/6 06:35	C_1	19.86	17.44	0	26.416	90.5
10	2017/7/6 06:40	C_1	20.52	16.48	0	27.218	88.46
11	2017/7/6 06:45	C_1	21.72	15.6	0	26.926	86.82
12	2017/7/6 06:50	C_1	19.06	16.72	0	26.79	87.08
13	2017/7/6 06:55	C_1	17.65	16.36	0	26.904	87.56
14	2017/7/6 07:00	C_1	16.14	14.82	0	27.62	85.64
15	2017/7/6 07:05	C_1	15.64	13.98	0	27.588	84.66
16	2017/7/6 07:10	C_1	15.48	11.86	0.1	27.56	84.12
17	2017/7/6 07:15	C_1	14.58	12.18	0	28.374	83.2
18	2017/7/6 07:20	C_1	15.9	11.85	0	28.654	81.38
19	2017/7/6 07:25	C_1	12.58	11.86	0	28.412	80.92
20	2017/7/6 07:30	C_1	12.05	12.2	0	28.442	80.96
21	2017/7/6 07:35	C_1	11	11.56	0	28.312	81.46
22	2017/7/6 07:40	C_1	10.08	10.5	0	28.532	80.74
23	2017/7/6 07:45	C_1	9.96	9.4	0.4	28.508	79.88
24	2017/7/6 07:50	C_1	13.64	9.4	0.4	28.718	80.12
25	2017/7/6 07:55	C_1	12.2	9.16	0.5	28.632	79.42
26	2017/7/6 08:00	C_1	11.66	9.28	0	29.026	78.36
27	2017/7/6 08:05	C_1	10.76	8.84	0	29.82	76.76
28	2017/7/6 08:10	C_1	9.725	9.28	0	29.952	75.74
29	2017/7/6 08:15	C_1	9.68	9.06	0	30.118	75.08
30	2017/7/6 08:20	C_1	9.26	9.38	0.1	30.492	73.66
31	2017/7/6 08:25	C_1	11.92	9.06	0	30.322	73.88
32	2017/7/6 08:30	C_1	8.72	9.6	0.1	30.768	73.06
33	2017/7/6 08:35	C_1	12.3	9.55	0.6	30.638	72.66

7 columns in the input file:

1. time
2. site: 10 stations; C_1~C_10
3. site_pm2.5
4. high_level_pm2.5
5. ws: wind speed
6. temperature
7. rh: relative humidity

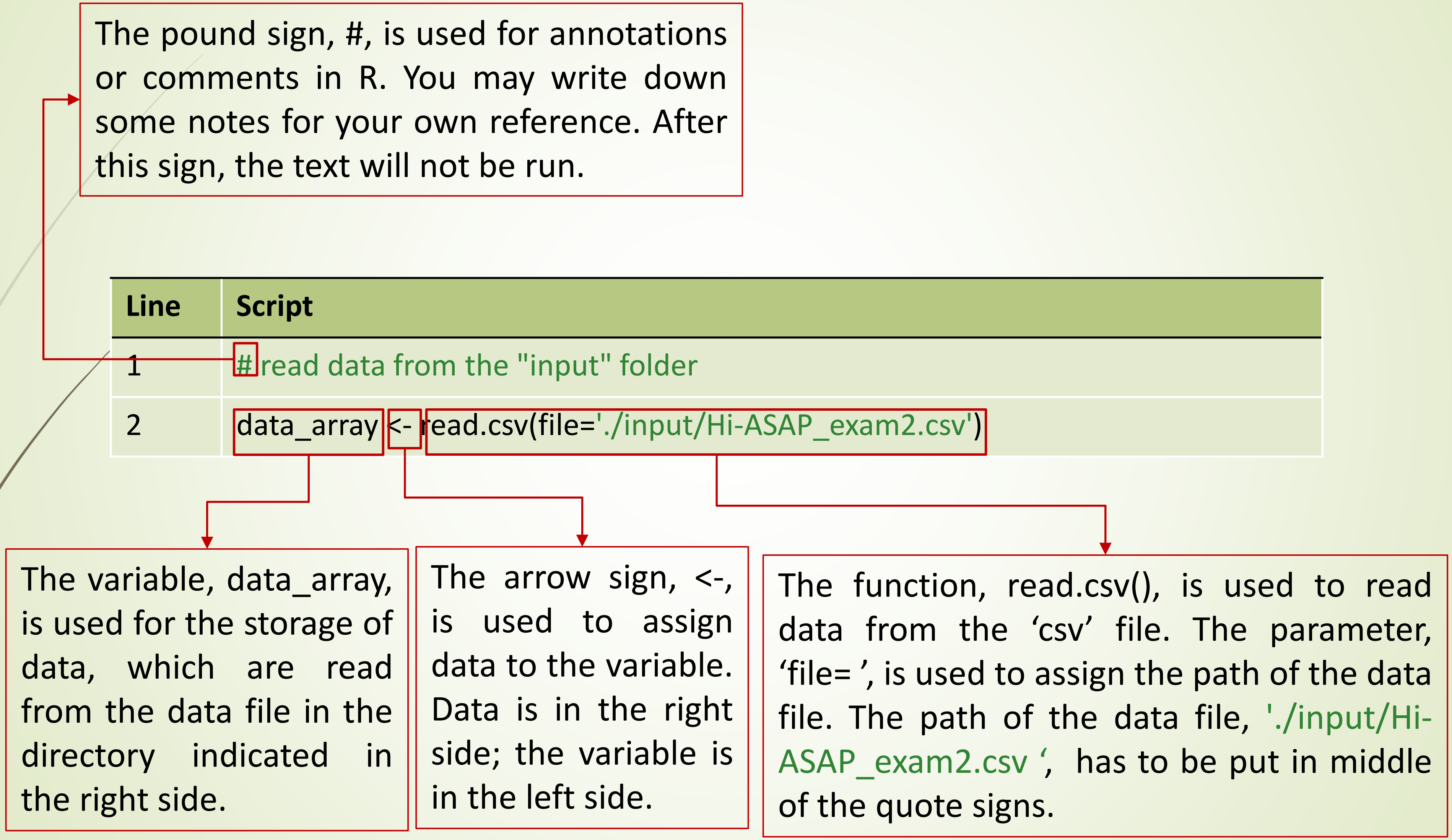
Exam 2 (3/3)

1. List the p-value of the overall regression model.
2. List the adjusted R^2 of the overall regression model.
3. List the contribution of the market.
4. List the contribution of the gas station.
5. Deliver three result files, which are the answers to exam 2, the regression result, and input data including the established dummy variables.
6. Pleas follow the file naming rules:
    - exam2_answers_[team name].xlsx
    - exam2_inputdata_[team name].csv
    - exam2_mlr_result_[team name].txt



exam2_answers_taiwan.xlsx
exam2_inputdata_taiwan.csv
exam2_mlr_result_taiwan.txt

Source code of R (1/13): Read the data file



The pound sign, #, is used for annotations or comments in R. You may write down some notes for your own reference. After this sign, the text will not be run.

Line	Script
1	#read data from the "input" folder
2	data_array <- read.csv(file='./input/Hi-ASAP_exam2.csv')

The variable, data_array, is used for the storage of data, which are read from the data file in the directory indicated in the right side.

The arrow sign, <-, is used to assign data to the variable. Data is in the right side; the variable is in the left side.

The function, read.csv(), is used to read data from the 'csv' file. The parameter, 'file= ', is used to assign the path of the data file. The path of the data file, './input/Hi-ASAP_exam2.csv ', has to be put in middle of the quote signs.

Source code of R (2/13): Conversion of data time

strptime: convert a string type of data time to a date object.

Line	Script
4	# convert a character string type to "Date Time" type
5	data_time <- strptime(data_array$time, "%Y/%m/%d %H:%M")
6	data_array$month <- as.integer(strftime(data_time,"%m"))
7	data_array$hour <- as.integer(strftime(data_time,"%H"))

as.integer: convert a string type of time to an integer.

strftime(data_time, "%H"): extract the “hour” component from the date object.

strftime(data_time,"%m"): extract the “month” component from the date object.

Source code of R (3/13): create the dummy variable array

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles
Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Pollution source: traffic type 1

Line	Script
10	## for traffic type 1_traffic_passing_by
11	data_array$traffic_passing_by <- 0
12	data_array$traffic_passing_by[(data_array$site %in% c('C_1','C_2','C_3','C_4','C_5'))] <- 1

Create a dummy variable column which was named “traffic with passing-by vehicles” and set to be zero first.

Then, set the sites with the emission source “traffic with passing-by vehicles” to be 1

Source code of R (4/13): create the dummy variable array

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles
Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Create a dummy variable column which was named “stop-and-go traffic” and set to be zero first.

Pollution source: traffic type 2

Line	Script
14	## for traffic type 2_traffic_stop_n_go
15	data_array$traffic_stop_n_go <- 0
16	data_array$traffic_stop_n_go [(data_array$site %in% c('C_6','C_8','C_9','C_10'))] <- 1

Then, set the sites with the emission source “stop-and-go traffic” to be 1

Source code of R (5/13): create the dummy variable array

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles
Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Pollution source: temple

Line	Script
18	## for temple
19	data_array$temple <- 0
20	data_array$temple [(data_array$site %in% c('C_6','C_9','C_10'))] <- 1

Create a dummy variable column which was named “temple” and set to be zero first.

Then, set the sites with the emission source “temple” to be 1

Source code of R (6/13): create the dummy variable array

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles

Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Some pollution sources have specific business hours as follows:

- ➢Market: 10 am - 8 pm (including 8 pm)
- ➢Vendor: 1 pm – 9 pm (including 9 pm)
- ➢Gas station: 8 am – 7 pm (including 7 pm)

Pollution source: market

Line	Script
22	## for market, time for 10-20
23	data_array$market <- 0
24	data_array$market [(data_array$site %in% c('C_3')) & (data_array$hour>=10) & (data_array$hour<=20)] <- 1

Create a dummy variable column which was named “market” and set to be zero first.

Then, set the sites with the emission source “market” and specific business hours to be 1

Source code of R (7/13): create the dummy variable array

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles
Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Some pollution sources have specific business hours as follows:

- Market: 10 am - 8 pm (including 8 pm)
- Vendor: 1 pm – 9 pm (including 9 pm)
- Gas station: 8 am – 7 pm (including 7 pm)

Pollution source: vendor

Line	Script
26	## for vendor, time for 13-21
27	data_array$vendor <- 0
28	data_array$vendor [(data_array$site %in% c('C_5')) & (data_array$hour>=13) & (data_array$hour<=21)<- 1

Create a dummy variable column which was named “vendor” and set to be zero first.

Then, set the sites with the emission source “vendor” and specific business hours to be 1

Source code of R (8/13): create the dummy variable array

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles
Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Some pollution sources have specific business hours as follows:

- ➢Market: 10 am - 8 pm (including 8 pm)
- ➢Vendor: 1 pm – 9 pm (including 9 pm)
- ➢Gas station: 8 am – 7 pm (including 7 pm)

Pollution source: gas station

Line	Script
30	## for gas station, time for 8-19
31	data_array$gas_stat <- 0
32	data_array$gas_stat [(data_array$site %in% c('C_4')) & (data_array$hour>=8) & (data_array$hour<=19)<- 1

Create a dummy variable column which was named “gas station” and set to be zero first.

Then, set the sites with the emission source “gas station” and specific business hours to be 1

Source code of R (9/13): create the dummy variable array

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles
Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Pollution source: school

Line	Script
34	## for school
35	data_array$school <- 0
36	data_array$school [(data_array$site %in% c('C_1'))] <- 1

Create a dummy variable column which was named “school” and set to be zero first.

Then, set the sites with the emission source “school” to be 1

Source code of R (10/13): create the dummy variable array

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles
Traffic type 2: Stop-and-go traffic (stop near the traffic light)

Dummy variable: season

Line	Script
38	## for season
39	data_array$season[data_array$month==7] <- 0
40	data_array$season[data_array$month==12] <- 1

Create a dummy variable column which was named “season” and set to be zero when the "month" variable is 7.

Set the variable of the season to be 1 when the "month" variable is 12.

Source code of R (11/13): build the multiple regression model

lm(formula=) is the function to establish the multiple regression model.

Line	Script
42	## the multiple regression model
43	mlr<-lm(formula=site_pm2.5 ~ traffic_passing_by + traffic_stop_n_go + temple
44	+ market + gas_stat + vendor + school + season + high_level_pm2.5 + ws
	+ temperature + rh, data=data_array)

Input the data with the dummy variables which are created by the above steps to the multiple regression model.

Source code of R (12/13): save the result of the regression model

Line	Script
50	# save the result of the multiple regression model in the "txt" file
51	sink("./output/exam2_mlr_result_taiwan.txt")
52	summary(mlr)
53	sink() # returns to the console

Use sink() to output the result in the ‘txt’ file.

summary() is to present the results of the multiple regression model.

The second sink() is to declare the end of the sink function.

Source code of R (13/13):
save the input data with dummy variables

Line	Script
55	# save the data which is used in the multiple regression model in the "csv" file
56	write.csv(data_array,file="./output/exam2_inputdata_taiwan.csv",row.names = FALSE)

write.csv() is to output the data in the ‘csv’ file.

The storage pathway and the file name of the ‘csv’.

‘row.names = FALSE’ means not to output the row number in the file.

Results: exam2_mlr_result_[team name].txt

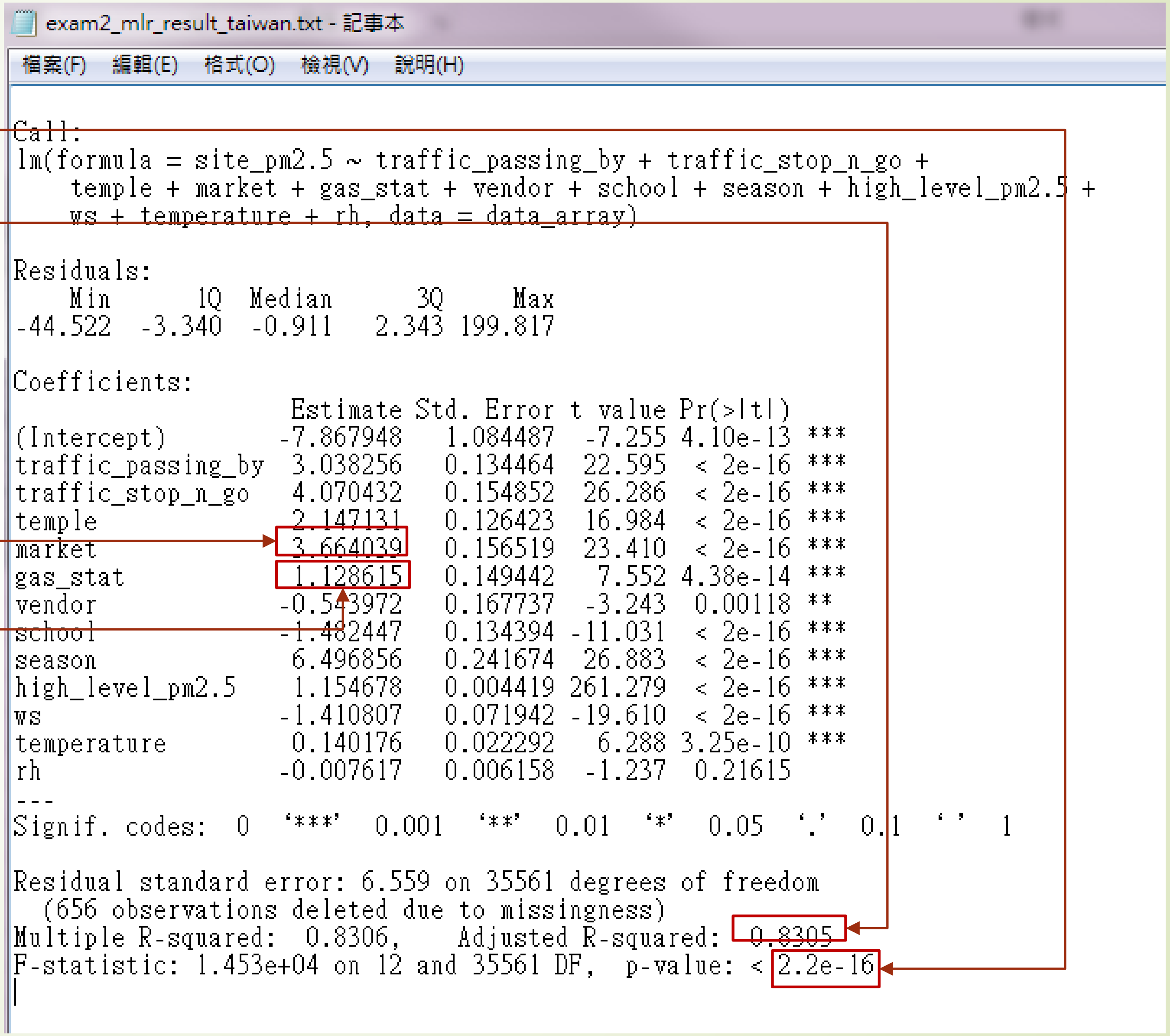
Exam 2:

1.List the p-value of the overall regression model.

2.List the adjusted R^2 of the overall regression model.

3.List the contribution of the market.

4.List the contribution of the gas station.


exam2_mlr_result_taiwan.txt - 記事本
檔案(F) 編輯(E) 格式(O) 檢視(V) 說明(H)

```
Call:
lm(formula = site_pm2.5 ~ traffic_passing_by + traffic_stop_n_go +
    temple + market + gas_stat + vendor + school + season + high_level_pm2.5 +
    ws + temperature + rh, data = data_array)

Residuals:
    Min      1Q  Median      3Q     Max
-44.522  -3.340  -0.911   2.343 199.817

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -7.867948   1.084487  -7.255 4.10e-13 ***
traffic_passing_by  3.038256   0.134464  22.595  < 2e-16 ***
traffic_stop_n_go   4.070432   0.154852  26.286  < 2e-16 ***
temple              2.147131   0.126423  16.984  < 2e-16 ***
market              3.664039   0.156519  23.410  < 2e-16 ***
gas_stat            1.128615   0.149442   7.552 4.38e-14 ***
vendor             -0.543972   0.167737  -3.243  0.00118 **
school             -1.482447   0.134394 -11.031  < 2e-16 ***
season              6.496856   0.241674  26.883  < 2e-16 ***
high_level_pm2.5    1.154678   0.004419 261.279  < 2e-16 ***
ws                 -1.410807   0.071942 -19.610  < 2e-16 ***
temperature         0.140176   0.022292   6.288 3.25e-10 ***
rh                 -0.007617   0.006158  -1.237  0.21615
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.559 on 35561 degrees of freedom
  (656 observations deleted due to missingness)
Multiple R-squared:  0.8306,    Adjusted R-squared:  0.8305
F-statistic: 1.453e+04 on 12 and 35561 DF,  p-value: < 2.2e-16
```

Results: exam2_answers_[team name].xlsx

Exam 2:

1.List the p-value of the overall regression model.

2.List the adjusted R^2 of the overall regression model.

3.List the contribution of the market.

4.List the contribution of the gas station.

Exam 2	Answer
1. List the p-value of the overall regression model.	2.20E-16
2. List the adjusted R^2 of the overall regression model.	0.8305
3. List the contribution of the market.	3.664039
4. List the contribution of the gas station.	1.128615

- Thank you for your participation!