

```

1      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69
70      *****
71      * Analysis Script
72      * Murphy John
73      * 2025-04-21
74      * This script performs the main analyses of this project.
75      *****;
76
77      title "Setup";
78
79      ** footnote;
80      footnote "Analysis script run on &SYSDATE at &SYSTIME.";
81
82      ** establish library;
83      libname mylib "/home/u63984496/BIOS7400/final-project";
NOTE: Libref MYLIB was successfully assigned as follows:
Engine:          V9
Physical Name:   /home/u63984496/BIOS7400/final-project
84
85      ** set graphics ods;
86      ods graphics on / width=8in height=4in;
87      *****;
88
89      title "Exploratory analysis";
90
91      * plot the outcome of interest, avgprice;
92      ** sort data by type, date, and region;
93      proc sort data=mylib.dat;
94          by type date region;
95      run;

```

NOTE: There were 546 observations read from the data set MYLIB.DAT.

NOTE: The data set MYLIB.DAT has 546 observations and 18 variables.

NOTE: PROCEDURE SORT used (Total process time):

real time	0.01 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	929.31k
OS Memory	29868.00k
Timestamp	04/30/2025 05:45:57 PM
Step Count	318 Switch Count 1
Page Faults	0
Page Reclaims	113
Page Swaps	0
Voluntary Context Switches	37
Involuntary Context Switches	0
Block Input Operations	0
Block Output Operations	264

```

96
97      ** plot price over time by type stratified by region;
98      proc sgpanel data=mylib.dat;
99          panelby type;
100         series x = date y = avgprice / group = region;
101      run;

```

NOTE: PROCEDURE SG PANEL used (Total process time):

real time	0.23 seconds
user cpu time	0.09 seconds
system cpu time	0.01 seconds
memory	12805.35k
OS Memory	39344.00k
Timestamp	04/30/2025 05:45:57 PM
Step Count	319 Switch Count 9
Page Faults	0
Page Reclaims	2608
Page Swaps	0
Voluntary Context Switches	3497
Involuntary Context Switches	7
Block Input Operations	0
Block Output Operations	1656

NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.

NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.

NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.

NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.

NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.  
 NOTE: The column format YYMMDD10 is replaced by an auto-generated format on the axis.  
 NOTE: There were 546 observations read from the data set MYLIB.DAT.

```
102
103      ** plot price by pres_party and type stratified by region;
104      proc sgpanel data=mylib.dat;
105      panelby region type;
106          hbox avgprice / group = pres_party;
107      run;
```

NOTE: PROCEDURE SG PANEL used (Total process time):

real time	0.83 seconds
user cpu time	0.30 seconds
system cpu time	0.10 seconds
memory	3461.84k
OS Memory	38068.00k
Timestamp	04/30/2025 05:45:58 PM
Step Count	320
Switch Count	97
Page Faults	0
Page Reclaims	3353
Page Swaps	0
Voluntary Context Switches	26640
Involuntary Context Switches	39
Block Input Operations	0
Block Output Operations	7048

NOTE: There were 546 observations read from the data set MYLIB.DAT.

```
108
109      ** plot price by temp and type stratified by region;
110      proc sgpanel data=mylib.dat;
111      panelby type;
112          scatter x = temp y = avgprice / group = region;
113      run;
```

NOTE: PROCEDURE SG PANEL used (Total process time):

real time	0.14 seconds
user cpu time	0.05 seconds
system cpu time	0.02 seconds
memory	3134.40k
OS Memory	38068.00k
Timestamp	04/30/2025 05:45:58 PM
Step Count	321
Switch Count	10
Page Faults	0
Page Reclaims	522
Page Swaps	0
Voluntary Context Switches	3476
Involuntary Context Switches	4
Block Input Operations	0
Block Output Operations	1160

NOTE: There were 546 observations read from the data set MYLIB.DAT.

```
114      *****;
115
116      * univariable analysis;
117      ** get means for each type;
118      proc univariate data=mylib.dat plots;
119      var avgprice;
120      class type;
121      run;
```

NOTE: PROCEDURE UNIVARIATE used (Total process time):

real time	0.20 seconds
user cpu time	0.13 seconds
system cpu time	0.00 seconds
memory	4176.59k
OS Memory	37900.00k
Timestamp	04/30/2025 05:45:58 PM
Step Count	322
Switch Count	0
Page Faults	0
Page Reclaims	614
Page Swaps	0
Voluntary Context Switches	395
Involuntary Context Switches	6
Block Input Operations	0
Block Output Operations	648

```

123      * bivariable analysis;
124      ** check correlations of numerical variables;
125      proc corr pearson data=mylib.dat;
126      var temp avgprice;
127      by type;
128      run;

```

NOTE: PROCEDURE CORR used (Total process time):

```

real time      0.04 seconds
user cpu time   0.04 seconds
system cpu time 0.00 seconds
memory         1304.62k
OS Memory      37288.00k
Timestamp      04/30/2025 05:45:58 PM
Step Count     323  Switch Count  0
Page Faults    0
Page Reclaims  102
Page Swaps     0
Voluntary Context Switches 7
Involuntary Context Switches 2
Block Input Operations 0
Block Output Operations 16

```

```

129
130      ** create macro to check means of categorical variables;
131      %macro means(var);
132      proc means data = mylib.dat;
133      var avgprice;
134      class &var;
135      by type;
136      run;
137      %mend means;
138
139      * run;
140      %means(month_num);

```

NOTE: There were 546 observations read from the data set MYLIB.DAT.

NOTE: PROCEDURE MEANS used (Total process time):

```

real time      0.03 seconds
user cpu time   0.03 seconds
system cpu time 0.00 seconds
memory         2088.90k
OS Memory      38316.00k
Timestamp      04/30/2025 05:45:58 PM
Step Count     324  Switch Count  2
Page Faults    0
Page Reclaims  208
Page Swaps     0
Voluntary Context Switches 17
Involuntary Context Switches 2
Block Input Operations 0
Block Output Operations 0

```

```

141      %means(year);

```

NOTE: There were 546 observations read from the data set MYLIB.DAT.

NOTE: PROCEDURE MEANS used (Total process time):

```

real time      0.02 seconds
user cpu time   0.03 seconds
system cpu time 0.00 seconds
memory         2171.03k
OS Memory      38316.00k
Timestamp      04/30/2025 05:45:58 PM
Step Count     325  Switch Count  2
Page Faults    0
Page Reclaims  198
Page Swaps     0
Voluntary Context Switches 15
Involuntary Context Switches 1
Block Input Operations 0
Block Output Operations 16

```

```

142      %means(region);

```

NOTE: There were 546 observations read from the data set MYLIB.DAT.

NOTE: PROCEDURE MEANS used (Total process time):

```

real time      0.03 seconds
user cpu time   0.03 seconds

```

```

system cpu time    0.00 seconds
memory            2060.21k
OS Memory         38316.00k
Timestamp         04/30/2025 05:45:58 PM
Step Count        326  Switch Count  2
Page Faults       0
Page Reclaims     180
Page Swaps        0
Voluntary Context Switches 15
Involuntary Context Switches 3
Block Input Operations 0
Block Output Operations 16

```

```
143      %means(pres_party);
```

NOTE: There were 546 observations read from the data set MYLIB.DAT.

NOTE: PROCEDURE MEANS used (Total process time):

```

real time         0.02 seconds
user cpu time     0.02 seconds
system cpu time   0.00 seconds
memory           2033.75k
OS Memory         38316.00k
Timestamp         04/30/2025 05:45:58 PM
Step Count        327  Switch Count  2
Page Faults       0
Page Reclaims     180
Page Swaps        0
Voluntary Context Switches 14
Involuntary Context Switches 1
Block Input Operations 0
Block Output Operations 0

```

```
144      *****;
```

```
145
146      title "Regression Model fits";
```

```

147
148      * create train and test data;
149      ** randomly select 80 percent of the data for the training and reserve the remainder for testing;
150      proc surveyselect data=mylib.dat
151          out=dat_select
152          samprate=0.8
153          outall
154          seed=333;
155      run;

```

NOTE: The data set WORK.DAT\_SELECT has 546 observations and 19 variables.

NOTE: PROCEDURE SURVEYSELECT used (Total process time):

```

real time         0.01 seconds
user cpu time     0.01 seconds
system cpu time   0.00 seconds
memory           935.90k
OS Memory         37548.00k
Timestamp         04/30/2025 05:45:58 PM
Step Count        328  Switch Count  2
Page Faults       0
Page Reclaims     91
Page Swaps        0
Voluntary Context Switches 21
Involuntary Context Switches 1
Block Input Operations 0
Block Output Operations 264

```

```

156
157      ** create sepearte data sets;
158      data dat_train dat_test;
159          set dat_select;
160          if selected then output dat_train;
161          else output dat_test;
162      run;

```

NOTE: There were 546 observations read from the data set WORK.DAT\_SELECT.

NOTE: The data set WORK.DAT\_TRAIN has 437 observations and 19 variables.

NOTE: The data set WORK.DAT\_TEST has 109 observations and 19 variables.

NOTE: DATA statement used (Total process time):

```

real time         0.00 seconds
user cpu time     0.00 seconds
system cpu time   0.00 seconds
memory           1342.18k

```

```

OS Memory          37808.00k
Timestamp          04/30/2025 05:45:58 PM
Step Count         329  Switch Count  4
Page Faults        0
Page Reclaims      157
Page Swaps         0
Voluntary Context Switches  28
Involuntary Context Switches 0
Block Input Operations  0
Block Output Operations 528

```

```

163
164
165 * simple linear regression;
166 ** avgprice by type, only;
167 proc glm data=dat_train plots=all;
168 class type (ref='conventional');
169 model avgprice = type;
170 run;

171
172 *** rmse = 0.251153, rsq = 0.485679;
173
174 * bivariable regression with interaction;
175 ** model avgprice by type, covariate, and interaction;
176 ** covariates are month_num, year, temp, region, pres_party;
177
178 * write macro to fit the numerical covariate model;
179 %macro bivariable_num(covariate);
180 proc glm data=dat_train plots=all;
181 class type (ref='conventional');
182 model avgprice = type &covariate type*&covariate;
183 run;
184
185 %mend bivariable_num;
186
187 * write macro to fit the categorical covariate models;
188 %macro bivariable_cat(covariate);
189 proc glm data=dat_train plots=all;
190 class type (ref='conventional') &covariate;
191 model avgprice = type &covariate type*&covariate;
192 run;
193
194 %mend bivariable_cat;
195
196 * run;
197
198 %bivariable_num(temp);

```

NOTE: PROCEDURE GLM used (Total process time):

```

real time          0.27 seconds
user cpu time      0.13 seconds
system cpu time    0.03 seconds
memory            10373.12k
OS Memory          44764.00k
Timestamp          04/30/2025 05:45:58 PM
Step Count         330  Switch Count  23
Page Faults        0
Page Reclaims      11934
Page Swaps         0
Voluntary Context Switches  765
Involuntary Context Switches 12
Block Input Operations  0
Block Output Operations 1488

```

```

199 *** rmse = 0.234790, rsq = 0.552579;
200
201 %bivariable_cat(region);

```

NOTE: PROCEDURE GLM used (Total process time):

```

real time          0.34 seconds
user cpu time      0.17 seconds
system cpu time    0.03 seconds
memory            10281.09k
OS Memory          44764.00k
Timestamp          04/30/2025 05:45:59 PM
Step Count         331  Switch Count  23
Page Faults        0

```

Page Reclaims	11801
Page Swaps	0
Voluntary Context Switches	2138
Involuntary Context Switches	8
Block Input Operations	0
Block Output Operations	2480

```

202      *** rmse = 0.202991, rsq = 0.673291;
203
204      %bivariable_cat(pres_party);

```

```

NOTE: PROCEDURE GLM used (Total process time):
real time          0.32 seconds
user cpu time      0.15 seconds
system cpu time    0.04 seconds
memory            10232.43k
OS Memory          44764.00k
Timestamp          04/30/2025 05:45:59 PM
Step Count         332  Switch Count  23
Page Faults        0
Page Reclaims      11590
Page Swaps         0
Voluntary Context Switches  3427
Involuntary Context Switches  9
Block Input Operations  0
Block Output Operations 1704

```

```

205      *** rmse = 0.244407, rsq = 0.515176;
206
207      %bivariable_cat(year);

```

```

NOTE: PROCEDURE GLM used (Total process time):
real time          0.31 seconds
user cpu time      0.14 seconds
system cpu time    0.04 seconds
memory            10270.43k
OS Memory          44764.00k
Timestamp          04/30/2025 05:45:59 PM
Step Count         333  Switch Count  23
Page Faults        0
Page Reclaims      11602
Page Swaps         0
Voluntary Context Switches  3336
Involuntary Context Switches  9
Block Input Operations  0
Block Output Operations 1536

```

```

208      *** rmse = 0.239643, rsq = 0.538200;
209
210      %bivariable_cat(month_num);

```

```

NOTE: PROCEDURE GLM used (Total process time):
real time          0.31 seconds
user cpu time      0.14 seconds
system cpu time    0.04 seconds
memory            10145.96k
OS Memory          44252.00k
Timestamp          04/30/2025 05:46:00 PM
Step Count         334  Switch Count  23
Page Faults        0
Page Reclaims      11642
Page Swaps         0
Voluntary Context Switches  3375
Involuntary Context Switches  6
Block Input Operations  0
Block Output Operations 1568

```

```

211      *** rmse = 0.231129, rsq = 0.586450;
212
213      * create table of metrics by hand;

```

```

NOTE: PROCEDURE GLM used (Total process time):
real time          0.32 seconds

```

```

user cpu time      0.14 seconds
system cpu time    0.04 seconds
memory            10289.00k
OS Memory         44252.00k
Timestamp         04/30/2025 05:46:00 PM
Step Count                335  Switch Count  23
Page Faults                0
Page Reclaims            11730
Page Swaps                0
Voluntary Context Switches 3512
Involuntary Context Switches 11
Block Input Operations     0
Block Output Operations    1768

```

```

214      data stats;
215          length Covariate $30;
216          infile datalines dsd trunccover;
217          input Covariate :$30. R_Square RMSE;
218          datalines;

```

NOTE: The data set WORK.STATS has 6 observations and 3 variables.

NOTE: DATA statement used (Total process time):

```

real time      0.00 seconds
user cpu time   0.00 seconds
system cpu time 0.00 seconds
memory         784.78k
OS Memory      38056.00k
Timestamp      04/30/2025 05:46:00 PM
Step Count                336  Switch Count  2
Page Faults                0
Page Reclaims            85
Page Swaps                0
Voluntary Context Switches 11
Involuntary Context Switches 1
Block Input Operations     0
Block Output Operations    264

```

```

225      ;
226      run;
227
228      proc print data=stats noobs;
229          var Covariate RMSE R_Square;
230      run;

```

NOTE: There were 6 observations read from the data set WORK.STATS.

NOTE: PROCEDURE PRINT used (Total process time):

```

real time      0.00 seconds
user cpu time   0.01 seconds
system cpu time 0.00 seconds
memory         610.90k
OS Memory      38056.00k
Timestamp      04/30/2025 05:46:00 PM
Step Count                337  Switch Count  0
Page Faults                0
Page Reclaims            62
Page Swaps                0
Voluntary Context Switches 0
Involuntary Context Switches 0
Block Input Operations     0
Block Output Operations    0

```

```

231
232      *** All three bivariable models improve the fit above the univariable model as measured by the rmse and r-squared.
233      *** The bivariable model with region has the lowest rmse and highest r-squared.;
234
235      * full model;
236      ** avgprice by type, temp, region, pres_party and interactions;
237      proc glm data=dat_train plots=all;
238          class pres_party type (ref='conventional') region month_num year;
239          model avgprice = type temp region pres_party month_num year
240              type*temp type*region type*pres_party type*month_num type*year;
241      run;
242
243      *** rmse = 0.145137, rsq = 0.844828;
244
245      * reduced model;
246      ** the interaction between type and temp, month_num, and pres_party have small Type III SS and large p-values.
247      ** Reduce the model by removing these terms.;

```

NOTE: PROCEDURE GLM used (Total process time):

real time	0.28 seconds
user cpu time	0.14 seconds
system cpu time	0.04 seconds
memory	10195.18k
OS Memory	45020.00k
Timestamp	04/30/2025 05:46:00 PM
Step Count	338 Switch Count 23
Page Faults	0
Page Reclaims	11651
Page Swaps	0
Voluntary Context Switches	601
Involuntary Context Switches	7
Block Input Operations	0
Block Output Operations	1784

```
247      proc glm data=dat_train plots=all alpha=0.05;
248          class type (ref='conventional') region pres_party month_num year;
249          model avgprice = type temp region pres_party month_num year
250              type*region;
251          store out=final_model;
252      run;

253      quit;
```

NOTE: The GLM procedure generated the model item store WORK.FINAL\_MODEL.

NOTE: PROCEDURE GLM used (Total process time):

real time	0.28 seconds
user cpu time	0.14 seconds
system cpu time	0.03 seconds
memory	10612.03k
OS Memory	45276.00k
Timestamp	04/30/2025 05:46:01 PM
Step Count	339 Switch Count 23
Page Faults	0
Page Reclaims	11631
Page Swaps	0
Voluntary Context Switches	607
Involuntary Context Switches	7
Block Input Operations	0
Block Output Operations	2032

```
254      *** rmse = 0.152099, rsq = 0.823078;
255
256      *** The reduced model is very similar in terms of fit to the full model but does not include unnessecary interaction
257      ! terms.
258      *** Select this as the final model.;
259      *****,
260
261      title 'Model Evaluation';
262      * use the final model to generate predictions on the test data;
263      proc plm restore=final_model;
264          score data=dat_test out=predictions predicted;
265      run;
```

NOTE: The data set WORK.PREDICTIONS has 109 observations and 20 variables.

NOTE: PROCEDURE PLM used (Total process time):

real time	0.01 seconds
user cpu time	0.02 seconds
system cpu time	0.00 seconds
memory	1117.46k
OS Memory	38316.00k
Timestamp	04/30/2025 05:46:01 PM
Step Count	340 Switch Count 2
Page Faults	0
Page Reclaims	143
Page Swaps	0
Voluntary Context Switches	20
Involuntary Context Switches	1
Block Input Operations	0
Block Output Operations	272

```
265
266      * compute rmse and r-squared;
267      ** get residuals;
268      data eval;
269          set predictions;
```



```

270         resid = avgprice - predicted;
271         sq_resid = resid**2;
272     run;

```

NOTE: There were 109 observations read from the data set WORK.PREDICTIONS.

NOTE: The data set WORK.EVAL has 109 observations and 22 variables.

NOTE: DATA statement used (Total process time):

```

real time          0.00 seconds
user cpu time      0.00 seconds
system cpu time    0.00 seconds
memory            1088.12k
OS Memory          38316.00k
Timestamp          04/30/2025 05:46:01 PM
Step Count                341  Switch Count  2
Page Faults              0
Page Reclaims           112
Page Swaps              0
Voluntary Context Switches 19
Involuntary Context Switches 0
Block Input Operations   0
Block Output Operations 264

```

```

273
274     * compute mean;
275     proc means data=eval noprint;
276         var avgprice predicted sq_resid;
277         output out=metrics
278             mean(avgprice)=mean_y
279             sum(sq_resid)=ss_res
280             n=samples;
281     run;

```

NOTE: There were 109 observations read from the data set WORK.EVAL.

NOTE: The data set WORK.METRICS has 1 observations and 5 variables.

NOTE: PROCEDURE MEANS used (Total process time):

```

real time          0.00 seconds
user cpu time      0.00 seconds
system cpu time    0.01 seconds
memory            6640.31k
OS Memory          43712.00k
Timestamp          04/30/2025 05:46:01 PM
Step Count                342  Switch Count  3
Page Faults              0
Page Reclaims           1454
Page Swaps              0
Voluntary Context Switches 34
Involuntary Context Switches 0
Block Input Operations   0
Block Output Operations 272

```

```

282
283     * compute metrics;
284     data results;
285         set metrics;
286         ss_total = 0;
287         do i = 1 to samples;
288             set eval point=i nobs=n;
289             ss_total + (avgprice - mean_y)**2;
290         end;
291         rmse = sqrt(ss_res / samples);
292         rsq = 1 - (ss_res / ss_total);
293         keep rmse rsq;
294     run;

```

NOTE: There were 1 observations read from the data set WORK.METRICS.

NOTE: The data set WORK.RESULTS has 1 observations and 2 variables.

NOTE: DATA statement used (Total process time):

```

real time          0.00 seconds
user cpu time      0.00 seconds
system cpu time    0.00 seconds
memory            1438.50k
OS Memory          38320.00k
Timestamp          04/30/2025 05:46:01 PM
Step Count                343  Switch Count  3
Page Faults              0
Page Reclaims           153
Page Swaps              0
Voluntary Context Switches 19
Involuntary Context Switches 1

```

```
Block Input Operations      0
Block Output Operations     264
```

```
295
296     proc print data=results;
297         title "RMSE and R-squared on Test Data";
298     run;
```

NOTE: There were 1 observations read from the data set WORK.RESULTS.

NOTE: PROCEDURE PRINT used (Total process time):

```
real time      0.00 seconds
user cpu time   0.01 seconds
system cpu time 0.00 seconds
memory         719.40k
OS Memory      37800.00k
Timestamp      04/30/2025 05:46:01 PM
Step Count          344  Switch Count  1
Page Faults         0
Page Reclaims       62
Page Swaps          0
Voluntary Context Switches  11
Involuntary Context Switches 0
Block Input Operations  0
Block Output Operations  0
```

```
299
300     * residuals vs fitted plot;
301     proc sgplot data=eval;
302         scatter x=predicted y=resid / markerattrs=(symbol=circlefilled color=black);
303         refline 0 / axis=y lineattrs=(color=red pattern=shortdash);
304         xaxis label="Fitted Values";
305         yaxis label="Residuals";
306         title "Residuals vs. Fitted Values";
307     run;
```

NOTE: PROCEDURE SGPLOT used (Total process time):

```
real time      0.07 seconds
user cpu time   0.03 seconds
system cpu time 0.00 seconds
memory         1781.59k
OS Memory      38668.00k
Timestamp      04/30/2025 05:46:01 PM
Step Count          345  Switch Count  2
Page Faults         0
Page Reclaims       279
Page Swaps          0
Voluntary Context Switches  145
Involuntary Context Switches  3
Block Input Operations  0
Block Output Operations 384
```

NOTE: There were 109 observations read from the data set WORK.EVAL.

```
308
309     * predicted vs observed plot;
310     proc sgplot data=predictions;
311         scatter x=predicted y=avgprice / markerattrs=(symbol=circlefilled color=black);
312         lineparm x=0 y=0 slope=1 / lineattrs=(color=red pattern=shortdash);
313         xaxis label="Predicted Values";
314         yaxis label="Observed Values";
315         title "Predicted vs. Observed Values";
316     run;
```

NOTE: PROCEDURE SGPLOT used (Total process time):

```
real time      0.07 seconds
user cpu time   0.02 seconds
system cpu time 0.00 seconds
memory         2169.40k
OS Memory      38668.00k
Timestamp      04/30/2025 05:46:01 PM
Step Count          346  Switch Count  2
Page Faults         0
Page Reclaims       280
Page Swaps          0
Voluntary Context Switches  143
Involuntary Context Switches  1
Block Input Operations  0
Block Output Operations 416
```

NOTE: There were 109 observations read from the data set WORK.PREDICTIONS.

```
317
318      * END OF SCRIPT;
319
320      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
330
```