3.4.1: Model Validation

Dr. Bean - Stat 5100

```
/* Project 2 is focused on using information regarding Tinder profiles
to predict the genuineness of the user. Information regarding the total
set of variables are included in the project 2 description. For purposes
of illustration, only a subset of variables are considered here. */

/* This first line of code will need to be changed */
FILENAME REFFILE '/home/u41171697/data/project2/tinder.csv';

PROC IMPORT DATAFILE=REFFILE replace
        DBMS=CSV
        OUT=WORK.tinder;
        GETNAMES=YES;
RUN;


/* Separate Into Training and Test Sets.
Only Fit Models to the Training Set. The variable
"Selected" separates training (0) from test (1)
seed - sets a random seed that allows your code to be reproduced
out - the name of the output dataset that includes the selected variable
rate - the percentage of points (between 0 and 1) that will be "selected"
       for validation */
proc surveyselect data=tinder seed=12345 out=tinder2
     rate=0.2 outall; /* Withold 20% for validation */
run;

proc print data=tinder2;
run;
```

Obs	Selected	ID	Genuine	SocPrivConc	InstPrivConc	Narcissism	SelfEsteem	Loneliness	Hookup	Friends
1	0	Subj57	-0.5	1	1	1.75	3.4	2.71	1	4
2	0	Subj310	-0.25	1.5	2.75	2.5	4	3.52	4	1.5
3	0	Subj303	1.5	3.75	4	1	3.4	3.27	3.25	4.25
4	0	Subj309	4	5	5	1.5	4.2	2.94	1	4.5
5	0	Subj426	2	2.25	3	2	2.2	4.19	5	3.5
6	0	Subj316	1.75	3.5	2.75	2.75	2.2	1.98	4.5	2
7	1	Subj5	2.5	2	2	3	3.2	2.98	3.75	1.75
8	0	Subj115	2	2.25	3	3	4	1	3.5	4
9	0	Subj327	0.5	1	1	2.25	3.8	1.1	2.75	2.5
10	0	Subj252	-2	1	3.75	3	2	2.79	4	3
11	0	Subj339	-1.5	4.25	4.5	2	3.8	2.57	2	4.25

```
data train; set tinder2;
if Selected = 0;
run;

data test; set tinder2;
```

```
if Selected = 1;
run;

proc print data = train;
run;

/* Fit one model with 4 variables.  */
proc reg data=train noprint;
 model genuine = socprivconc instprivconc narcissism selfesteem;
store regModel;
run;

/* Fit another model with more variables.  */
proc reg data=train noprint;
 model genuine = socprivconc instprivconc narcissism selfesteem loneliness
                 hookup friends partner travel selfvalidation entertainment;
store regModel2;
run;

/* Fit a third model with NO variables  */
proc reg data=train noprint;
 model genuine = ;
store regModel3;
run;

/* Calculate MSPR for each model by first making predictions
(via proc plm), then estimating errors (via a data step) and
calculating the means (via proc means). */
proc plm restore=regModel;
 score data=test out=newTest predicted;
 run;
 proc plm restore=regModel2;
 score data=test out=newTest2 predicted;
 run;

proc plm restore=regModel3;
 score data=test out=newTest3 predicted;
 run;


data newTest; set newTest;
ASE = (Genuine - Predicted)**2;
run;
data newTest2; set newTest2;
ASE = (Genuine - Predicted)**2;
run;
data newTest3; set newTest3;
ASE = (Genuine - Predicted)**2;
run;
```

```
proc means data = newTest;
var ASE;
run;
proc means data = newTest2;
var ASE;
run;
proc means data = newTest3;
var ASE;
run;
```

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	TINDER
Random Number Seed	12345
Sampling Rate	0.2
Sample Size	100
Selection Probability	0.200803
Sampling Weight	4.98
Output Data Set	TINDER2

The MEANS Procedure

Analysis Variable : ASE				
N	Mean	Std Dev	Minimum	Maximum
99	3.4138416	4.4208025	8.8289731E-6	18.9501242

The MEANS Procedure

Analysis Variable : ASE				
N	Mean	Std Dev	Minimum	Maximum
99	2.7276352	3.5324712	0.000862991	17.8117101

The MEANS Procedure

Analysis Variable : ASE				
N	Mean	Std Dev	Minimum	Maximum
99	3.5206134	3.8275996	0.0065659	21.7992795