

## Code to implement the GEE approach.

(As described in: Moskowitz and Pepe. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. **Biostatistics** (in press). )

The code given here is for Stata Statistical Software, Release 7.0, but implementation of the GEE approach is relatively simple with any statistical package.

We begin with a demonstration of how to estimate the predictive accuracy of a single prognostic factor,  $X$ , denoted by  $x$  in the code below and then show how to estimate the difference in predictive accuracy between two prognostic factors,  $X_1$  and  $X_2$ , denoted by  $x_1$  and  $x_2$ . The outcome is denoted as  $d$ .

### Steps to Estimate the Predictive Accuracy of a Single Continuous Factor

1. Obtain the proportional ranks ( $F_X(X) = V$ ) for each of the observations:

```
sort x
gen V=_n/_N
```

2. Specify  $S_v = (v_1, \dots, v_J)$ . For this example, we use four equally spaced points at 0, .25, .5, and .75. This code will obviously change somewhat depending on the choice of  $S_v$ .

```
forvalues j = 0 25:75{gen v'j'='j'*.01}
```

3. Rearrange the data so that each individual has  $J$  records corresponding to the  $J$  points in  $S_v$ . In place of the variable  $id$  below, substitute the subject identifier for the data

```
reshape long v, i(id) j(proprank)
```

4. Determine if each individual's proportional rank is greater than the corresponding  $v_j$ .

```
gen r=(V >= v)
```

5. Fit the GEE model

```
xtgee d v if r==1, family(binomial) link(log) i(id) robust corr(indep)
```

## Steps to Estimate the Difference in Predictive Accuracy of Two Continuous Factor

1. Obtain the proportional rankings on the two factors ( $F_X(X) = V$ ) for each of the observations

```
sort x1
gen V1=_n/_N
```

```
sort x2
gen V2=_n/_N
```

2. Rearrange the data so that each individuals two records, one for each test

```
reshape long V, i(id) j(temptest)
gen z=2-temptest
drop temptest
```

3. Specify  $S_v = (v_1, \dots, v_J)$ . For this example, we use four equally spaced points at 0, .25, .5, and .75. This code will obviously change somewhat depending on the choice of  $S_v$ .

```
forvalues j = 0 25:75{gen v'j'='j'*.01}
```

4. Rearrange the data so that each individual has  $2 \times J$  records corresponding to the  $J$  points in  $S_v$ . In place of the variable *id* below, substitute the subject identifier for the data

```
reshape long v, i(id z) j(proprank)
```

5. Determine if each individual's proportional rank is greater than the corresponding  $v_j$ .

```
gen r=(V >= v)
```

6. Obtain the interaction term between the test type and the proportional ranks

```
gen vz=z*v
```

7. Fit the GEE model

```
xtgee d v vz, if r==1, family(binomial) link(log) i(id) robust
corr(indep)
```