Model Control ala Choppin

The exposition that follows is based on the Pair estimation method suggested by Bruce Choppin (1968) when computers were new, slow, and expensive. The method is the basis of the estimation used in our earlier discussions. The Pairwise procedure provides opportunity for interesting investigations for model control, which could be conducted using any of the matrices of the last couple submissions: the 2x2 N_{ij} matrix for each item pair, the *LxL N**-matrix of paired comparisons, or the *R*-matrix of relative difficulties (i.e., log odds.) Since all are based on more or less the same data, the results will be similar. However, since they use the data in different ways, there will be some differences. One uses a rather heuristic standard error, one assumes one count or another is fixed, and the third uses information that wasn't part of the estimation. All are addressing the question, Is the behavior of this item pairing consistent with everything else we know?

The Matrix of Log Odds: ***R***

The *R*-matrix is the log odds comparing pairs of items. Each element $r_{ij} = ln\,(n_{ji} / n_{ij})$ starts with the ratio of the number of examinees for whom item *i* was harder than *j* versus the number for whom *j* was harder than *i*. This is a simple, direct estimate of the difference in logit difficulties $\delta_i - \delta_j$, which is based only on examinees who took that pair of items.

The relative difficulty of the two items is estimated more reliably, given the model, by the marginal estimates d_i and d_j, which use data from every pair involving either of the items. Using the earlier suggestion for standard error of r_{ij} in the denominator gives a familiar looking form:

$$44. \qquad t_{ij} = \frac{r_{ij} - (d_i - d_j)}{\frac{1}{2}\sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{ji}}}}\,.$$

This statistic tests if the direct estimate from the *ij* pair is consistent with all the information we have about the two items from all other pairings. It makes no assumption about the form of the model but does depend on the heuristic standard error estimate and there is a bit of a problem with a lack of independence because r_{ij} contributed to the estimates of the difficulties. Still if this yields a "big" number, there is something funny about the *ij* pair. The *funny* thing may be one item tips the other or it may relate to the specific examinees who took this pair. Diagnosis requires more than statistical rules of thumb about when a number is "big."

The Matrix of Counts: ***N****

The element n_{ij} of N^* is the number of examinees who missed item *i* and passed *j*. The bigger n_{ij} is compared to n_{ji}, the more difficult item *i* is compared to *j*. Each pair of counts in the N^* matrix can be used to compute a statistic that has the form of a χ^2 and addresses the question in a slightly different way. We know from the last chapter that n_{ij} / n_{ji} should equal D_j / D_i, where $D_j = e^{d}{}_{j}$. Given the marginal estimates, d_i and d_j, and the count n_{ji}, we can ask how big n_{ij} should be? One can compute an expected value for either count as:

45. $$E(n_{ij}) = n_{ji}\frac{D_j}{D_i},$$

The calculation of the expected number missing *i* and passing *j* starts with the estimated item difficulties as givens and the number n_{ji} who miss *j* but pass *i* as fixed. The downside is that this calculation will be volatile (partly from rounding error) if n_{ji} is small.

We can take a slightly different tack and assume that $n_{ij} + n_{ji}$ is the fixed bit, rather than n_{ji}, and apportion the total count to the two components based on the odds:

46. $$E(n_{ij}) = (n_{ij} + n_{ji})\frac{D_j}{D_i + D_j} \text{ and } E(n_{ji}) = (n_{ij} + n_{ji})\frac{D_i}{D_i + D_j}.$$

We can put this into a form that looks like a chi-square,

47. $$\chi^2_{ij} = \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})}.$$

All of this is subject to the usual limitations and caveats about minimum counts. But it does not depend on anybody's guess at a standard error.

The Two-by-Two Paired Matrix: ***N***

Finally, the N_{ij}-matrices suggest another possible control. For each pair of items, there are four possible outcomes.

		Item *i*	
		Wrong	Right
Item *j*	Wrong	n_{00}	n_{10}
	Right	n_{01}	n_{11}

Each cell in each matrix contains the count of the examinees who had the appropriate pair of scores on items *i* and *j*. An expected value can be computed for each score pattern using the model and the best ability estimate we have for each person.

48. $$E^{xy}_{ij} = \sum_{v=1}^{N} P_{vi}(x)P_{vj}(y),\ x = 0 \text{ or } 1 \text{ on item } i;\ y = 0 \text{ or } 1 \text{ on item } j,$$

where *N* is the total number of examinees who attempted the item pair, and $P_{vi}(x)$ is the probability that person *v* will score *x* on item *i*.

49. $$P_{vi}(x) = \frac{B_v}{B_v + D_i} \quad \text{if } x = 1; \text{ or} \quad P_{vi}(x) = \frac{D_i}{B_v + D_i} \quad \text{if } x = 0.$$

50. $$P_{vi}(x) = \frac{xB_v + (1-x)D_i}{B_v + D_i} \quad x = 0 \text{ or } 1.$$

With these expected and observed counts, another standard χ^2 *goodness-of-fit* statistic can be computed for each N_{ij}-matrix; identical in form to expression 47 and summed over the four cells.

This computation differs from the first two controls in at least one significant way. The upper left cell (both items incorrect) and the lower right cell (both items correct) are included, although neither cell is part of the estimation.

Because the people can take different item sets (which is one of the strengths of the Pair algorithm) rather than a fixed form, the summation in expression 48 and subscript for ability in expression 49 are for the people rather than the scores. If a fixed form is used, we can continue to index ability by *r*, the number correct score, and the computer's job is a little easier.

While the form of each of these statistics suggests a possible distribution, I will leave them as suggestions and avoid definitive statements. The statistics are included here as a starting point in that discussion and to re-emphasize the importance of model control to the application of Rasch models.

A Little Arithmetic: t*-test Example*

To illustrate the calculations for the control process, we will use the same data for a five-item test and 500 simulated examinees that were used earlier to illustrate the difficulty estimation process. The matrix *R* of log odds, with the difficulty estimates in the last column, is again:

Table IV.1: Matrix of Log Odds, Five items, 500 Examinees

R-Matrix of Log Odds					Recovered Difficulties
	-1.764	-2.471	-3.850		-2.795
1.764		-0.869	-1.908	-4.373	-1.077
2.471	0.869		-1.007	-3.034	-0.140
3.850	1.908	1.007		-2.180	0.917
	4.373	3.034	2.180		3.095

Cells *(1,5)* and *(5,1)* are empty because there was no data to provide a direct estimate of the two difficulties. The expected values for each entry in the *R*-matrix are based on the difficulty estimates for the appropriate pair of items:

Table 4.3: Expected Value for Each Entry from the Margin

Expected Log Odds for Entry				
0.000	-1.718	-2.655	-3.712	-5.891
1.718	0.000	-0.937	-1.994	-4.173
2.655	0.937	0.000	-1.057	-3.236
3.712	1.994	1.057	0.000	-2.178
5.891	4.173	3.236	2.178	0.000

For example, the expected value for row 3, column 5 is the estimated difficulty (from Table 4.1) for item 3 minus the estimated difficulty for item 5, or:

50. $$\hat{R}_{35} = -0.140 - 3.095 = -3.235.$$

The standard errors for each entry in the *R*-matrix are based on the number of useful observations:

Table 4.2: Standard Errors for Log Odds

Standard Errors for Entry					Standard Error for Margin
	0.221	0.213	0.292		0.244
0.221		0.117	0.123	0.252	0.188
0.213	0.117		0.090	0.137	0.146
0.292	0.123	0.090		0.103	0.173
	0.252	0.137	0.103		0.176

The t-statistic for cell (3,5) is:

51. $$t = \frac{-3.034-(-3.235)}{0.137} = 1.47,$$

which is within rounding error of the value in the table below[1]. This happens to be the largest value in the table and this is probably the only time I've ever said that about a *t = 1.47*.

Table 4.4: t-Statistics for Each Entry Based on Log Odds

t-Statistic for Entry				
	-0.21	0.87	-0.47	
0.21		0.58	0.70	-0.80
-0.87	-0.58		0.56	1.47
0.47	-0.70	-0.56		-0.02
	0.80	-1.47	0.02	

A Little Less Arithmetic; More Interesting Example

Simulated data has its uses but diagnosing sources of anomalies isn't one. For a more interesting example, I return to the football field. In an earlier section, we arrived at logit measures of proficiency for the 32 professional teams in the National Football League based on 2012 results. The table below starts with the logits for the twelve teams that qualified for post-season play (third column or third row). The entries in the body of the table, also logits, describe an encounter between the row team and the column team and are simply the row logit minus the column logit. For example, the Minnesota Vikings have a logit of 0.25 and the Green Bay Packers a logit of 0.30, hence the logit for the Vikings against the Packers is *0.25 – 0.30 = -0.05*. In words, the Packers are slightly favored.

A positive value in a cell implies the row team is favored; a negative value, the column team is favored. The entire row for the San Francisco 49ers is positive so it should beat any of the other eleven teams. By contrast, the Vikings should lose to any other NFC team except Washington (d_{97} = 0.16) and beat every AFC team except Denver (d_{96} = –0.27).

The first table shows our expectation, in logits, for any of the 66 games that might be played among the twelve playoff teams. Eleven games were actually played. The eleven

[1] The tables shown here used the full precision of the spread sheet program; the arithmetic done here used the three digits shown.

are highlighted in yellow for the favored team; there is an equal but opposite value for the unfavored team.

First Table: Expected Logits for 66 Possible Games for the 12 Playoff Teams

		Team	NE	Cin	Bal	Hou	Ind	Den	Was	GB	Mn	Atl	SF	Sea
	Team	Logit	0.53	0.22	0.11	0.23	-0.21	0.52	0.09	0.3	0.25	0.52	0.61	0.58
AFC	Patriots	0.53	0.00	0.31	0.42	0.30	0.74	0.01	0.44	0.23	0.28	0.01	-0.08	-0.05
AFC	Bengals	0.22	-0.31	0.00	0.11	-0.01	0.43	-0.30	0.13	-0.08	-0.03	-0.30	-0.39	-0.36
AFC	Ravens	0.11	-0.42	-0.11	0.00	-0.12	0.32	-0.41	0.02	-0.19	-0.14	-0.41	-0.50	-0.47
AFC	Texans	0.23	-0.30	0.01	0.12	0.00	0.44	-0.29	0.14	-0.07	-0.02	-0.29	-0.38	-0.35
AFC	Colts	-0.21	-0.74	-0.43	-0.32	-0.44	0.00	-0.73	-0.30	-0.51	-0.46	-0.73	-0.82	-0.79
AFC	Broncos	0.52	-0.01	0.30	0.41	0.29	0.73	0.00	0.43	0.22	0.27	0.00	-0.09	-0.06
NFC	Washington	0.09	-0.44	-0.13	-0.02	-0.14	0.30	-0.43	0.00	-0.21	-0.16	-0.43	-0.52	-0.49
NFC	Packers	0.30	-0.23	0.08	0.19	0.07	0.51	-0.22	0.21	0.00	0.05	-0.22	-0.31	-0.28
NFC	Vikings	0.25	-0.28	0.03	0.14	0.02	0.46	-0.27	0.16	-0.05	0.00	-0.27	-0.36	-0.33
NFC	Falcons	0.52	-0.01	0.30	0.41	0.29	0.73	0.00	0.43	0.22	0.27	0.00	-0.09	-0.06
NFC	49ers	0.61	0.08	0.39	0.50	0.38	0.82	0.09	0.52	0.31	0.36	0.09	0.00	0.03
NFC	Seahawks	0.58	0.05	0.36	0.47	0.35	0.79	0.06	0.49	0.28	0.33	0.06	-0.03	0.00

The second table shows the points scored by each team in the eleven playoff games, with the winning score highlighted. If every game had followed our expectation, the same cells would be highlighted as in the first table; not everything went quite as we expected.

Second Table: Scores of Eleven Games Actually Played

		Team	NE	Cin	Bal	Hou	Ind	Den	Was	GB	Mn	Atl	SF	Sea
	Team	Logit	0.53	0.22	0.11	0.23	-0.21	0.52	0.09	0.3	0.25	0.52	0.61	0.58
AFC	Patriots	0.53			13	41								
AFC	Bengals	0.22				13								
AFC	Ravens	0.11	28				24	38					34	
AFC	Texans	0.23	28	19										
AFC	Colts	-0.21			9									
AFC	Broncos	0.52			35									
NFC	Washington	0.09												14
NFC	Packers	0.30									24		31	
NFC	Vikings	0.25								10				
NFC	Falcons	0.52											24	30
NFC	49ers	0.61			31					45		28		
NFC	Seahawks	0.58							24			28		

The entries in each row are the points scored by the team; the values in the column are the points allowed. The Vikings scored 10 but allowed 24; the Packers scored 24 and allowed 10. These scores can be used to compute an *observed* logit; i.e., Vikings versus Packers = *ln(10/24) = -0.88*. Comparing this to the expected value of -0.05 in the first value says that the Vikings lost as expected (because both values are negative) but lost more handily than expected (because -0.88 is more negative than -0.05.) Complementary comments can be made about the Packers, but they may be less than complimentary outside of Wisconsin.

The third table has the observed logits for the eleven games. A positive value in a row means the team won; if it is larger than the expected value in first table, the team won by more than expected.

Third Table: Observed Logits for Eleven Games Actually Played

		Team	NE	Cin	Bal	Hou	Ind	Den	Was	GB	Mn	Atl	SF	Sea
	Team	Logit	0.53	0.22	0.11	0.23	-0.21	0.52	0.09	0.3	0.25	0.52	0.61	0.58
AFC	Patriots	0.53			-0.77	0.38								
AFC	Bengals	0.22				-0.38								
AFC	Ravens	0.11	0.77				0.98	0.08					0.09	
AFC	Texans	0.23	-0.38	0.38										
AFC	Colts	-0.21			-0.98									
AFC	Broncos	0.52			-0.08									
NFC	Washington	0.09												-0.54
NFC	Packers	0.30									0.88		-0.37	
NFC	Vikings	0.25								-0.88				
NFC	Falcons	0.52											-0.15	0.07
NFC	49ers	0.61			-0.09					0.37		0.15		
NFC	Seahawks	0.58							0.54			-0.07		

The fourth table makes the comparisons easier by subtracting the expected from the observed and *standardizing* by dividing by something that might be a standard error (or proportional to it.) We have been playing rather fast and loose with the model by treating points scored like counts but *bigger numbers* in the table will still indicate bigger surprises than *small numbers*. It probably comes as no surprise to any follower of American football in 2012 that four of the five biggest playoff surprises are associated with the Baltimore Ravens. The other surprise, perhaps not in Minnesota, was how convincingly the Vikings lost to the Packers.

Fourth Table: "Standardized" Values of Observed Logits for Games Actually Played

		Team	NE	Cin	Bal	Hou	Ind	Den	Was	GB	Mn	Atl	SF	Sea
	Team	Logit	0.53	0.22	0.11	0.23	-0.21	0.52	0.09	0.3	0.25	0.52	0.61	0.58
AFC	Patriots	0.53			-7.1	0.7								
AFC	Bengals	0.22				-2.1								
AFC	Ravens	0.11	7.1				3.4	4.2					4.8	
AFC	Texans	0.23	-0.7	2.1										
AFC	Colts	-0.21			-3.4									
AFC	Broncos	0.52			-4.2									
NFC	Washington	0.09												-0.3
NFC	Packers	0.30									4.4		-0.5	
NFC	Vikings	0.25								-4.4				
NFC	Falcons	0.52											-0.5	1.0
NFC	49ers	0.61			-4.8					0.5		0.5		
NFC	Seahawks	0.58							0.3			-1.0		

With these data for these eleven games, it happened that every winning score (second table) is associated with a positive residual (third table) but that does not need to be the case. In fact, I would be happier if half had been negative. A positive residual means that the winner won by more than we expected. For example, the Vikings against the Packers had an expected logit of –0.05, implying the Vikings should lose by a little. They actually lost by a lot (10 to 24,) giving an observed logit of $ln(10/24) = -0.88$. Had the Vikings managed to score 23 points, the logit then would have been $ln(23/24) = -0.04$ and the logit residual would have been positive, $(-0.04 - (-0.05)) = 0.01$. They still would have lost but had the moral victory of achieving a positive residual. A *moral victory* means they exceeded expectations, but positive residuals don't win contract extensions for coaches.

We haven't tried to predict the actual scores, just the log of the ratios of the points. There are many scores that would give the same ratio; we have just shown one. The assumption of the last paragraph was that the Packers would score exactly 24. We could just as well assume that there would be a total of 10 + 24 = 34 points scored. Then the Vikings should, in theory, score 16.6 and the Packers 17.4.

Any other positive integer would be just as defensible[2] but these two approaches give the appearance of having some basis in reality. To get the expected Viking score in the first case, we need to solve $\boldsymbol{exp(-0.05) = X/24)}$; and in the second, we solved $\boldsymbol{exp(-0.05) = X/(34-X)}$.

All these fun and games with logits have a point, which is to suggest another approach to doing the arithmetic for spotting funny outcomes; i.e., discrepancies between the observed and expected points. The table below scores the results for the eleven play-off games. For these calculations, we fixed the total points scored as the sum of the points scored by the team and its opponent and used the logits to apportion the points between the two teams.

Fifth Table: "Chi-squared" Values for Points Scored in Games Actually Played

Team	Logit	Obs	Exp	χ^2	Opp	Logit	Obs	Exp	χ^2	Sum	"t"	T*
Ravens	0.11	28	16.3	8	NE	0.53	13	24.7	6	14	7	14.4
Ravens	0.11	34	24.5	4	SF	0.61	31	40.5	2	6	5	6.1
Vikings	0.25	10	16.6	3	GB	0.30	24	17.4	2	5	4	5.0
Ravens	0.11	38	29.1	3	Den	0.52	35	43.9	2	5	4	4.5
Ravens	0.11	24	19.1	1	Ind	-0.21	9	13.9	2	3	3	2.7
Bengals	0.22	13	15.9	1	Hou	0.23	19	16.1	1	1	0	-0.1
Falcons	0.52	30	28.1	0	Sea	0.58	28	29.9	0	0	0	0.0
Patriots	0.53	41	39.6	0	Hou	0.23	28	29.4	0	0	0	-0.1
Packers	0.30	31	32.2	0	SF	0.61	45	43.8	0	0	0	-0.1
Falcons	0.52	24	24.8	0	SF	0.61	28	27.2	0	0	0	0.0
Washington	0.09	14	14.4	0	Sea	0.58	24	23.6	0	0	1	-0.1

The basic calculation for "expected" points is based on the expression we just used, $exp(-0.05) = X/(34-X)$, rearranged as:

$$52.\quad E_{ij} = \frac{O_i + O_j}{1 + e^{d_j - d_i}}.$$

That's not as mysterious as it looks: add the points scored by both teams together; divide by *e* raised to the power of the difference in logits plus one. For the Vikings and Packers,

[2] Not quite any positive integer. In American football, you can't score just one. But if the Packers could score one, the Vikings would be expected to score $\boldsymbol{exp(-0.05) = 0.95}$. That's what odds (0.95 to 1) and log odds, i.e., logits, (-0.05) mean in this world.

it's *10+24 = 34* over *exp(0.3 – 0.25) +1 = 2.05* for the Vikings and *exp(0.25 -0.3) + 1 = 1.95* for the Packers. The Vikings get *34 / 2.05 = 16.6* of the *34* points scored and the Packers get *34 / 1.95 = 17.4*. Or that's the way it was supposed to work.

The columns I have labeled χ^2 are labeled that because the algebra looks like the standard form for a *goodness-of-fit* χ^2:

53. $\chi^2 = \frac{(O-E)^2}{E} = \frac{(10-16.6)^2}{16.6} = 2.6$ for the Vikings against the Packers and

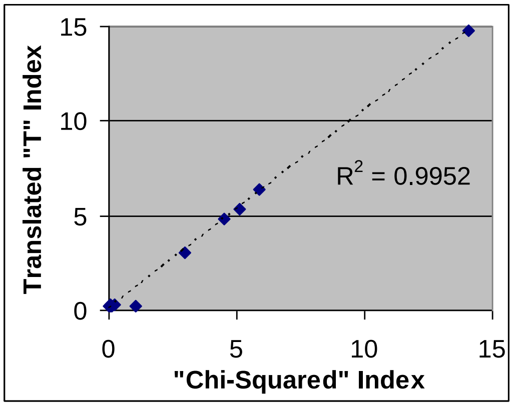
54. $\chi^2 = \frac{(O-E)^2}{E} = \frac{(24-17.4)^2}{17.4} = 2.5$ for the Packers against the Vikings.

The column labeled *Sum* is the sum of the two χ^2. Everything in these columns has been rounded to integers because any decimals would imply more confidence in the numbers than I feel or professional sports warrants.

After all this, it didn't matter if we used the *pseudo-t* (penultimate column of the fifth table, copied from the fourth table) or the *pseudo-*χ^2 (antepenultimate column); we would reach the same conclusions. Where it matters (i.e., with the large numbers), results are indistinguishable given appropriate choices for what defines a large number. Four of the five largest numbers involve the Ravens and the third largest in either metric is the Packers-Viking score. I can make the numbers, not just the conclusions, match in this example by translating the *t*-values with:

55. $T^* = t\,(t-1)/3.$

This is the basis for the figure below. The small values don't matter much because they don't imply any surprise and any fluctuations are due to being restricted to points coming in clusters, or at least quanta.



First Figure: "Chi-squared" Values and Translated "t" Values

Slicing and Dicing the Pair Matrix

Any of the approaches to model control we have been describing yield a statistic that quantifies our degree of surprise for each item pair (expressions 44, 47, and 48). We have at least three *LxL* matrices of surprise. We can then add up the rows or columns to see if our surprise attaches to one or more items across the board. We can in fact take slices bigger than a single row across the matrices to ask if the surprise attaches to items of a particular difficulty range, type, content, format, type-face, source, age, etc.

The n_{ij} can be diced on the person dimension, within the cells. When building the matrices containing the pairwise counts of examinees that passed one item and failed the other, we paid no attention to anything about the examinees, looking simply at pairs of items, one examinee's response string at a time. When Specific Objectivity holds, all attributes of the examinees (including their abilities) are irrelevant. We mustn't take this on faith and should verify that it is in fact the case. Then a Pair *R-matrix* using, say, just the boys should look just like the total matrix, which should look just like the expected matrix that we constructed from the difficulty estimates based on the row averages. The counts of course will be lower for the partial group, more cells may be empty, the degrees of freedom may be lower, and everything subject to statistical variability, but the logic is exactly the same as the contingency table analyses just described for the total matrix. We now have multiple versions of the *LxL* matrices of counts and log odds, the versions defined by a level of some factor for the people, e.g., gender, ethnicity, ability, SES, LEP, age, grade, year, *ad nauseum*. All the same analyses can be done.

Depending on how much data and time we have, and how much money the client has, we should continue slicing across facets of the items and dicing among factors for the people that might be dangerous, problematic, or just interesting. The questions are only limited by our imagination, paranoia, and information available.