Using an Informative Missing Data Model to Predict the Ability to Assess Recovery of Balance Control after Spaceflight

Alan Feiveson[1], Scott Wood[2], Varsha Jain[3], William Paloski[1,4]
[1]NASA Johnson Space Center, Houston, TX 77058
[2]Universities Space Research Association, 3600 Bay Area Boulevard, Houston, TX 77058
[3]Imperial College, London, UK
[4]University of Houston, 3855 Holman St. Houston, TX 77204

Abstract
Astronauts show degraded balance control immediately after spaceflight. To assess this change, astronauts' ability to maintain a fixed stance under several challenging stimuli on a movable platform is quantified by "equilibrium" scores (EQs) on a scale of 0 to 100, where 100 represents perfect control (sway angle of 0) and 0 represents data loss where no sway angle is observed because the subject has to be restrained from falling. By comparing post- to pre-flight EQs for actual astronauts vs. controls, we built a classifier for deciding when an astronaut has recovered. For the main EQ of interest, this classifier perfectly separated the groups in our relatively small training data set; hence standard techniques for evaluating ROC area uncertainty were not applicable. The problem of predicting future performance of the classifier was addressed by simulation after modeling P(EQ = 0) in terms of a latent EQ-like beta-distributed random variable.

Key Words: beta distribution, missing data, logistic regression, roc area

1. Introduction

A practical concern in planning long-term space missions is that astronauts show degraded balance control for a period of time after returning to Earth [1]. As a consequence, in early post-flight days, astronauts may be at increased risk of falls during normal activities of daily living. Therefore, for a period of time after flight, they are generally cautioned not to attempt activities requiring good balance control (e.g., contact sports and climbing ladders), and they are restricted from performing more dangerous activities requiring good sensory-motor integration skills (e.g., driving vehicles and operating complex equipment). Although balance control is usually regained within a few days of landing, the recovery trend can vary considerably with mission duration, previous space flight experience, and non-specific individual characteristics.. There is, therefore, a need for diagnostic methodologies that can readily assess whether an astronaut has recovered sufficiently to return to duty.

Computerized dynamic posturography (CDP) [2] has been long used to monitor recovery of balance control in astronauts after space flight. CDP evaluates the ability of a subject to maintain a stable upright stance during 20 sec. trials of several challenging stimuli. Its utility as a diagnostic measure for assessing readiness for return-to-duty is somewhat limited however because of the length of time required to administer the entire test battery. To reduce this time, we sought to determine whether some subset of the CDP stimuli could be sufficient to distinguish between normal and impaired sensory-motor function in returning astronauts. Here we report performance aspects of three methods for classifying a set of longitudinal outcomes of one particular CDP protocol as being representative of either a recovered or a non-recovered subject. Original data and classification results were obtained from an experiment involving astronauts and matched control subjects. However to evaluate future predicted performance and assess the effects of sampling variability, we made use of simulated CDP outcomes based on an informative missing data model.

1.1 Experiment Design

In the classification experiment, one stimulus condition was selected from full CDP test batteries administered to 11 astronauts and 11 matched (by gender and age) non-flying control subjects before and after flight. In CDP, performance is quantified by "equilibrium" scores (EQ), transformed maximum postural sway angles scaled from 0 to 100, where 100 represents perfect control (no sway) and 0 represents a theoretical maximum sway angle of 12.5° that can be achieved without falling or moving one's feet. If a subject moves his/her feet or has to be restrained from falling, no maximum sway angle is actually observed; instead a score of zero is arbitrarily assigned. The CDP condition used for the classification experiment (eyes-closed, unstable foot support surface, dynamic pitch-plane head movements) was repeated twice per session. For future reference, the two pre-flight scores are designated as y_1 and y_2, while the post-flight scores are designated as y_3 and y_4. Time spacing between pre- and post- "flight" sessions for each control subject was matched to that experienced by the corresponding astronaut. All subjects were participating in CDP testing for the first time; hence any learning effects should have been similar for both groups. None reported any history of balance or vestibular abnormalities. Screening was also conducted before trial sessions to ensure that no toxic substances had been taken, there was no evidence of new onset illness and that there had not been any nausea or motion-sickness. All subject selection criteria and experimental procedures were approved by Johnson Space Center Committee for Protecting of Human Subjects, and all subjects provided informed consent prior to inclusion.

1.2 Assessing the Ability to Diagnose Recovery

By assuming that no astronauts had recovered when CDP testing was administered on landing day (2-4 hrs after wheels-stop), and that control subjects were representative of a population of "recovered" astronauts, we built three types of classifiers (linear, conditional linear and marginal likelihood) to separate the groups on the basis of their EQ scores. In an operational setting, one of these classifiers would be used in the days following landing to help decide whether an astronaut has recovered balance control sufficiently to safely return to normal duties and activities.

Predicted diagnostic performance of each classifier depends on both the sampling distribution of parameter estimates obtained in training the classifier as well as the distribution of future input data. With the small amount of data available, asymptotic methods for assessing standard errors of the classifier parameters or of sensitivity and specificity estimates are not reliable. Indeed, some of these methods completely break down when there is perfect separation between the groups in the sample data (as happened in our case). Therefore, to characterize performance, we simulated EQ scores for the current design, trained each classifier on a data set sized similarly to the study, and then applied the estimated classifier to simulated future EQ results. This process was repeated to build an empirical distribution of performance aspects for each of the classification methods.

1.3 Initial Results

The complete experimental data are given in Table 1, where astronauts are coded "1" in the column entitled "*Ast*". Initially, we used logistic regression with a dummy variable indicating control subjects (i.e. the "recovered" group) to construct a linear classifier based on averages $\bar{y}_{pre} = (y_1 + y_2)/2$ and $\bar{y}_{post} = (y_3 + y_4)/2$ of the pre- and post-flight EQ scores, respectively. In a proper logistic regression setting, we would choose the classifier to be of the form: decide "*recovered*" if $c_1 \bar{y}_{pre} + c_2 \bar{y}_{post} > A$; where the estimated probability of a subject belonging to the "recovered" group is $\exp(c_1 \bar{y}_{pre} + c_2 \bar{y}_{post} - A)/[1 + \exp(c_1 \bar{y}_{pre} + c_2 \bar{y}_{post} - A)]$. In other words, the classifier would decide "recovered" if the estimated probability of a subject belonging to the "recovered" group exceeds 0.5. However, because there was complete separation of the groups in $(\bar{y}_{pre}, \bar{y}_{post})$ –space , the likelihood as a function of c_1, c_2, and A was unbounded and estimates of c_1, c_2, and A as implemented in Stata statistical software [3] did not converge. Nevertheless, relative estimates of these parameters were converging as iteration proceeded. Since relative values of c_1, c_2, and A are sufficient to define a classifier, we imposed $c_1 = -1$ and took $c_2 = -c_2^N / c_1^N$ and $A = -A^N / c_1^N$, where c_1^N, $-c_2^N$ and A^N were the N-th-iterate estimates of c_1, c_2, and A. The particular software used terminated estimation after $N = 22$.

Results of the preceding calculations gave $c_1 = -1.0$, $c_2 = 0.929$, and $A = 26.4$. So using this classifier, a NASA flight surgeon would be advised that an astronaut had recovered if 0.929 times the astronaut's post-flight average EQ score exceeded the pre-flight average score by at least 26.4. Figure 1 shows this classifier and the original data in $(\bar{y}_{pre}, \bar{y}_{post})$ –space. But how well can we expect this classifier to perform in future application? With perfect separation, likelihood-based standard errors for the parameter values cannot be evaluated. Furthermore, how can we account for the small size of the training data in assessing future performance? It can be seen from Figure 1 that even a small change in some of the data values would have changed the performance of the classifier to less than perfect. To address these issues, we

modeled the EQ scores and implemented a parametric bootstrap. Section 2 describes the data model and its estimation. The parametric bootstrap simulation and alternatives to the linear classifier are presented in Section 3. Finally bootstrap accuracy assessment results and an overall discussion follow in Sections 4 and 5.

Table 1: Training Data

		Pre-flight		*Post-flight*	
Subject	*Ast*	y_1	y_2	y_3	y_4
1	0	45	43	42	41
2	0	33	65	68	80
3	0	28	73	56	73
4	0	63	69	72	66
5	0	41	25	29	0
6	0	35	56	69	59
7	0	48	37	41	51
8	0	74	71	51	66
9	0	44	58	27	38
10	0	84	86	70	73
11	0	71	71	77	76
12	1	56	71	26	39
13	1	57	49	0	0
14	1	63	78	0	29
15	1	73	77	0	0
16	1	83	75	0	0
17	1	68	65	0	0
18	1	57	61	0	0
19	1	60	66	36	0
20	1	72	64	0	0
21	1	67	40	0	32
22	1	72	55	68	0

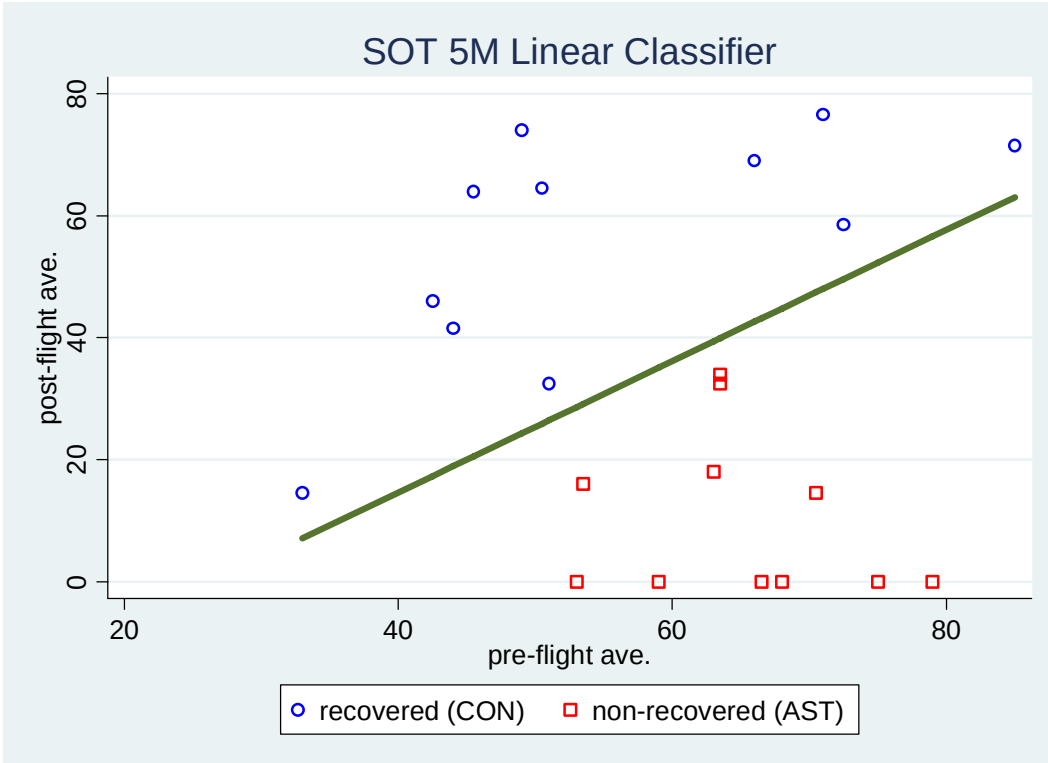

Figure 1: Linear classifier estimated from original experiment data.

2. Modeling EQ Scores

2.1 Missing Data Model

In order to accomplish the bootstrap analysis, it was necessary to formulate and estimate a complete structural model for EQ scores that includes allowance for no actual maximum sway angle being observed when a subject falls, as well as repeated observations pertaining to the same subject. Following [4], we modeled the marginal distribution of y = EQ/100 as equal to y^*, a Beta-distributed normalized "latent" score when a fall does not occur and equal to zero when a fall does occur. The probability of a fall; *i. e.* P($y = 0$) is further modeled as conditional on y^*. In particular we used the fall model

$$P(\text{fall} \mid y^*) = (1 - y^*)^{\theta} \text{ for some } \theta > 0. \tag{1}$$

In other words, the lower the latent score, the more likely it is that a fall would occur. As y^* approaches zero, a fall becomes almost a certainty; conversely as y^* approaches one (perfect control), the probability of a fall becomes negligible.

2.2 Longitudinal Model

To account for repeated observations from the same subject, the model for y^* incorporates random effects as well as fixed effects explaining the effect of spaceflight. More specifically, assume $y^* \sim \beta(p, q)$; *i.e.* y^* has density

$$f_{\beta}(y^*) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}(y^*)^{p-1}(1-y^*)^{q-1}, \tag{2}$$

where the log mean, $\zeta = \log\ p/(p+q)$ is modeled by

$$\zeta = \zeta_0 + \zeta_1 I_{post}(A) + u^{(1)} \tag{3}$$

where ζ_0 and ζ_1 are fixed parameters, $u^{(1)}$ is a random subject effect distributed $N(0, \sigma_1^2)$ and $I_{post}(A)$ indicates whether the observation was taken post-flight for an astronaut ($I_{post}(A) = 1$); otherwise $I_{post}(A) = 0$. In addition, $\varphi = \log (p + q)$ also incorporates random effects:

$$\varphi = \varphi_0 + u^{(2)} \tag{4}$$

where $u^{(2)}$ is another random subject effect distributed $N(0, \sigma_2^2)$. Finally, θ, the parameter affecting the conditional probability of a fall given y*, follows a log-linear model but without random effects:

$$\log \theta = \tau_0 + \tau_1 I_{post}(A) \tag{5}$$

With this model, the distribution of y^* varies over subjects; being the same pre- and post-flight for a given control subject, but changing after flight for a given astronaut subject.

2.3 Estimation

From (1) and (2), it can be seen that the marginal likelihood of y for a subject with random effects $\boldsymbol{u} = (u^{(1)}, u^{(2)})'$ is

$$f(y \mid \boldsymbol{u}) = \begin{cases} [1-(1-y)^{\theta}] f_{\beta}(y \mid \boldsymbol{u}) & y > 0 \\ P(fall \mid \boldsymbol{u}) & y = 0 \end{cases} \tag{6}$$

where $f_\beta(y \mid \boldsymbol{u})$ is the same as (2), except that the parameters p and q are altered by (4) and (5); and where the unconditional probability of a fall for that subject is

$$P(fall \mid \boldsymbol{u}) = \int_0^1 (1-y^*)^{\theta} f_{\beta}(y^* \mid \boldsymbol{u}) dy^* \tag{7}$$

Preliminary estimates of ζ_0, ζ_1, τ_0, and τ_1 were obtained by maximizing the "pseudo"-likelihood

$$\tilde{L}(\underline{y}) = \prod_{y>0} f(y \mid \boldsymbol{u} = 0) \prod_{y=0} P(fall \mid \boldsymbol{u} = 0) \tag{8}$$

which would be the actual likelihood if all random effects were zero and all observations were independent. These estimates were then used to form prior distributions for subsequent Bayesian estimation using WINBUGS software [5]. Final estimates of all parameters, including σ_1^2 and σ_2^2 were taken as posterior medians.

3. Simulation

Using the model described by (1) - (5) with the above parameter estimates, we generated 500 pairs of training and validation data sets, each with the same design (11 astronauts, 11 controls, pre-flight/post-flight, 2 replicates) as the actual experiment. For each pair, we trained each of three classifiers on the first ("training") set and then applied the estimated classifier to the second ("validation") set. The three classifiers consisted of the linear classifier (L), estimated by logistic regression as described in Sec. 1.3, and two alternatives – denoted "conditional linear" (CL) and "pseudo-likelihood ratio" (PLR). The CL-classifier was the same as the L-classifier except for the added rule that any post-flight fall automatically classified the subject as "non-recovered". This rule puts more penalty on a post-flight fall than would averaging it in as a value of zero. Unlike L and CL, the PLR-classifier utilizes the 4 distinct observations $y_1 - y_4$ for an individual by constructing the pseudo-likelihood (8), first assuming the individual was "recovered" ($I_{post}(A) = 0$), and then assuming "non-recovered" ($I_{post}(A) = 1$). The classifier would then decide "recovered" if the former pseudo-likelihood exceeded the latter. Classification results were then aggregated over the 500 validation data sets and various aspects of performance were compared.

4. Results

4.1 Parameter Estimates

Table 2 shows estimates(posterior medians) and the 95% Bayes credible interval for the EQ model parameters ζ_0, ζ_1, φ_0, τ_1, τ_2, σ_1, and σ_2 as described in (1) – (5). Note that for astronauts post-flight, the mean latent EQ score is estimated to be considerably lower (ζ_1 = -1.24). For example, using the estimates in Table 2, and applying (3) with $u^{(1)} = 0$ (a "typical" astronaut), the mean latent EQ score is $p/(p + q) = 0.31$ post-flight as opposed to 0.61 prior to flight. Also, from (2) and (7), the probability of a fall for a typical astronaut post-flight is estimated at 0.73, as compared with 0.015, pre-flight.

Table 2: EQ Model Parameter Estimation Results

Parameter	*Estimate*[1]	*95% low*[2]	*95% upper*[2]
ζ_0	0.433	0.154	0.707
ζ_1	-1.238	-1.692	-0.734
ϕ_0	3.101	2.517	3.695
τ_0	2.182	1.481	3.377
τ_1	-2.517	-4.101	-1.362
σ_1	0.530	0.387	0.736
σ_2	0.690	0.399	1.135

[1]Posterior median (10,000 MCMC draws)
[2]Bayes credible interval

4.2 Model Check

Before proceeding with the classification simulation described in Sec. 3, we first checked the model by using the above point estimates to simulate 500 data sets similar to the actual one and then compared various statistics between the actual data and averaged over the 500 simulated sets. Table 3 contains shows these comparisons for two types of statistics: averages and ANOVA mean squares. Values in the column labelled "Actual" are calculated averaged from the actual data. Values in the column labelled "Simulated" are averages over the 500 simulated data sets of quantities calculated for each set. The ANOVA means squares were calculated for comparison purposes only and were not meant for statistical inference on group, phase, effects, etc. Here "phase" means pre-flight vs. post-flight. As can be seen from Table 3, there is good agreement between the actual and simulated data except for the group mean square statistic. This quantity reflects the difference in EQ scores between astronauts and controls, averaged over both pre- and in-flight. Low values of this mean square tend to occur if the average pre-flight score for the astronaut group happens to be considerably higher than it is for the control group. Then the negative effect of flight is partially cancelled and the overall means don't differ by too much. The group mean square was less than 0.702 in about 13% of the simulated training sets; thus an observed value that low is not overly inconsistent with the model.

Table 3: Comparison between actual and simulated data

Criterion	*Actual*	*Simulated (500 sets)*
Averages		
EQ/100: ast(post)	0.105	0.096
EQ/100: con & ast(pre)	0.587	0.602
P(fall): ast(post)	0.727	0.759
P(fall): con & ast(pre)	0.0152	0.0091
ANOVA Mean Squares		
group (ast or con)	0.702	1.48
sub/group	0.0624	0.063
phase	1.62	1.43
phase*group	1.65	1.43
phase*sub/group	0.0261	0.02
residual	0.0178	0.0175

4.3 Classification Assessment

The classification performances of each classifier on the original data and on the simulated verification data are summarized in Tables 4 and 5, respectively. As reported above, there was perfect separation of groups in the original data for the linear classifier (L), but one control subject had a fall in one of the post-flight trials, hence that subject was classified as not recovered by the conditional linear classifier (CL). The pseudo-likelihood ratio classifier (PLR) was slightly more conservative, missing missed two control subjects (Table 4).

Table 4: Classification Results on Original Data

		L-Classifier	
		ast	con
Actual	AST	11	0
	CON	0	11

		CL-Classifier	
		ast	con
Actual	AST	11	0
	CON	1	10

		PLR-Classifier	
		ast	con
Actual	AST	11	0
	CON	2	9

Table 5 shows corresponding results from the simulation. Entries in this table are average frequencies over the 500 classified data sets, each the same size as the original. Note that for the linear classifier an average of almost 2.5 of 11 astronauts were erroneously classified as “recovered” – a serious error. These results suggest that the perfect separation on the original data was somewhat fortuitous – in fact this happened in 98 of the 500 simulated studies. By contrast, adding the simple rule of deciding “not recovered” if there is a post-flight fall considerably reduced the error rate. The pseudo-likelihood classifier performed about as well as the conditional linear – so based on these results, the conditional linear classifier, being much easier to implement, would be the rule of choice.

Table 5: Average Classification Results on Simulated Data

		L-Classifier	
		ast	con
Actual	AST	10.73	0.27
	CON	2.47	8.53

		CL-Classifier	
		ast	con
Actual	AST	10.64	0.36
	CON	0.51	10.49

		PLR-Classifier	
		ast	con
Actual	AST	10.43	0.57
	CON	0.51	10.49

We also looked at average areas under ROC curves for each classifier, obtained by calculating sensitivity and specificity as functions of the threshold *A* for each of the 500 simulated validation data sets. Results (Table 6) again suggest that the conditional linear classifier is a good choice.

Table 6: ROC areas

Classifier	*ROC area*
Linear	0.921
Conditional Linear	0.966
Pseudo-Likelihood Ratio	0.976

5.0 Discussion/Conclusions

An important conclusion reached from the simulation is that the perfect classification performance of the linear classifier on the original data was somewhat fortuitous. As can be seen in Fig. 1, slight "migration" of some of the data points in $(\bar{y}_{pre}, \bar{y}_{post})$ –space could have resulted in 2 or 3 errors. In fact, perfect separation in $(\bar{y}_{pre}, \bar{y}_{post})$ –space occurred in 98 of the 500 simulated training sets. By contrast, adding the simple rule of deciding "not recovered" if there is a post-flight fall considerably reduced the error rate. The pseudo-likelihood classifier performed about as well as the conditional linear – so based on these results, the conditional linear classifier, being much easier to implement, would be the rule of choice. This finding is also supported by the ROC areas in Table 6. Because of the complexity of the EQ data model, with log-linear random subject effects, Bayesian methods were used to estimate the model parameters which in turn enabled us to simulate realistic EQ data. Because of the relatively small sample size (22 subjects), the resulting parameter estimates were somewhat dependent on prior distributions. To control this dependence we constructed realistic centers for priors by maximizing the "pseudo"-likelihood (8) and using robust standard errors to define spread. A final sanity check was made by comparing statistics calculated form the actual and simulated data.

References

[1] Paloski, W.H., M.F. Reschke, and F.O. Black, Recovery of postural equilibrium control following space flight (DSO 605), in Extended Duration Orbiter Medical Project. Final Report, C.F. Sawin, G.R. Taylor, and W.L. Smith, Editors. 1999, National Aeronautics and Space Administration: Houston. p. 1-11.

[2] Nashner, L.M., *Computerized dynamic posturography: clinical applications*, in *Handbook of Balance Function Testing*, G.P. Jacobson, C.W. Newman, and J.M. Kartush, Editors. 1993, Mosby-Year Book, Inc.: Chicago, IL. p. 308-334.

[3] StataCorp. 2007. Stata Statistical Software: Release 10. College Station, TX: Stata Corp LP.

[4] Feiveson, A.H., E.J. Metter, and W.H. Paloski, *A statistical model for interpreting computerized dynamic posturography data*. IEEE Trans Biomed Eng, 2002. **49**(4): p. 300-9.

[5] Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**:325--337.