Naval Aerospace Medical Research Laboratory

NAMRL-1410

PREDICTING PRIMARY FLIGHT GRADES BY AVERAGING OVER LINEAR REGRESSION MODELS: PART 1

D. J. Blower, H. P. Williams, and A. O. Albert

Naval Aerospace Medical Research Laboratory
51 Hovey Road
Pensacola, Florida 32508-1046

Reviewed and approved 1 JAN 2000

R. R. STANNY, Ph.D.
Technical Director

This research was sponsored by the Office of Naval Research under work unit 62233N.0330.126-7801.

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Volunteer subjects were recruited, evaluated, and employed in accordance with the procedures specified in the Department of Defense Directive 3216.2 and Secretary of the Navy Instruction 3900.39 series. These instructions are based upon voluntary informed consent and meet or exceed the provisions of prevailing national and international guidelines

Trade names of materials and/or products of commercial or nongovernment organizations are cited as needed for precision. These citations do not constitute official endorsement or approval of the use of such commercial materials and/or products.

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
51 HOVEY ROAD, PENSACOLA, FL 32508-1046

NAMRL-1410

PREDICTING PRIMARY FLIGHT GRADES BY AVERAGING OVER LINEAR REGRESSION MODELS: PART 1

D. J. Blower, H. P. Williams, and A. O. Albert

ABSTRACT

This report documents an investigation into two types of variables that might be useful in predicting flight grades in Navy primary flight training. The first set of predictor variables is largely psychomotor in origin and is part of the Computer-Based Performance Test Battery at the Naval Aerospace Medical Research Laboratory. The second set of variables is more cognitive in nature and arises from scores on the Aviation Selection Test Battery (ASTB) and a final grade in Aviation Pre-Flight Indoctrination (API), which is ground school prior to entering primary flight training. The motivation for this research is a joint effort with the Air Force designed to improve selection tests for military aviators. The emphasis in this report is on how to choose good linear regression models which use these variables to predict a criterion variable such as flight grade. In our present case, we have a total of 25 potential predictor variables. As a result, there is a rather large number of possible regression models. Our task is to pick some relatively small number of models that are "best" by some acceptable statistical criterion. The analysis revealed that models with a small number of predictor variables were much superior to models that included a large number of the 25 available variables. The best models consisted of two, three, and four predictor variables and possessed an R^2 of about .35. The single best model contained the final grade from API, a psychomotor tracking variable, and a score from one of the ASTB subtests. A prediction of the flight grade can then be made by averaging over the individual predictions of the single models. Using Bayesian model evaluation techniques, the averaging is carried out by weighting each individual model according to its posterior probability.

Acknowledgments

Our thanks go to Ms. Claire Portman-Tiller and Ms. Kristi Nalley, who were responsible for running the experiments and collecting the ASTB and CBPT data. We also extend our gratitude to Mr. Ed Fisher, data base manager for the primary flight training squadrons at NAS Whiting Field, who supplied the flight grade data.

INTRODUCTION

This report documents an investigation into two types of variables that might be useful in predicting flight grades in Navy primary flight training. The first set of predictor variables is largely psychomotor in origin and is part of the Computer-Based Performance Test Battery at the Naval Aerospace Medical Research Laboratory. The second set of variables is more cognitive in nature and arises from scores on the Aviation Selection Test Battery (ASTB) and a final grade in Aviation Pre-Flight Indoctrination (API), which is ground school prior to entering primary flight training. The motivation for this research stems from a recent collaboration with the Air Force with the intent to improve selection tests for military aviators.

The emphasis in this report is on how to choose a set of good linear regression models which use these variables to predict a criterion variable such as flight grade. In our present case, we have a total of 25 potential predictor variables. As a result, there is a rather large number of possible regression models. Our task is to pick some relatively small number of models that are "best" by some acceptable statistical criterion.

The criterion that we shall be employing is model selection according to Bayesian principles. Generally, this approach finds the posterior probability (i.e., a revised probability after the data have been collected) for any given model. Any two models can be compared by forming the ratio of their posterior probabilities. Such a ratio is known as the posterior odds and reflects the odds in favor of one model over another competing model. Alternatively, one can list the posterior model probabilities of a small subset of good models taken from the larger set of all possible regressions. The main results of the analysis are presented in this fashion. We restrict the phrase "different models" to mean linear regression models with a different number or different composition of predictor variables as available from the entire set of 25 predictor variables.

The technical background for most of what is presented here can be found in a series of articles by Professor Adrian Raftery and his colleagues at the Statistics Department of the University of Washington. Three representative articles concerning Bayesian model selection for linear regression models are listed in the References section [1,2,3]. These articles provided the motivation for the study detailed in this report.

THE EXPERIMENTAL DATA

The data analyzed in this report come from a 3-year study at NAMRL that compared the performance on a traditional paper-and-pencil test (the ASTB) with a computerized version of this same test. As part of this study, the volunteer subjects were also administered NAMRL's Computer Based Performance Test (CBPT) Battery. See Blower and Dolgin [4] for a detailed description of the tests in this battery.

The overall purpose of the CBPT is to sample some fairly basic psychomotor skills such as tracking, dichotic listening, and two-dimensional spatial aptitude. The motivation for the CBPT is that learning about the relative performance of candidates in these areas should improve the prediction of success or failure in the later stages of flight training. Currently, selection into naval aviation training only takes cognitive skills into account and it was thought that tapping into this relatively independent set of skills might improve the selection process. In addition, all the subjects had previously taken the ASTB as part of the routine medical selection process demanded of all candidates for naval aviation training.

All the subjects (Ss) were commissioned officers in either the Navy or Marine Corps. They were awaiting entry into API at Naval Aviation Schools Command, NAS Pensacola, the academic ground school portion of training before the candidates actually began the flight curriculum. There were 265 Ss in the original study of whom 260 successfully graduated from API. We wanted to concentrate solely on pilot flight grades so we eliminated another 26 Ss who chose the Naval Flight Officer (NFO) pipeline. Some Ss were eliminated because of missing data in the set of predictor variables and other Ss could not be used because they had not yet completed training. After these Ss were eliminated, there remained a total of 210 Ss for whom the subsequent analysis described in this report was carried out.

Of these 210 Ss, 200 subjects were male and 10 were female. Of these, 148 were in the Marine Corps and 62 were in the Navy while 185 were right-handed, 18 were left-handed, and 7 were ambidextrous. The mean age of the subjects was 24.10 with a standard deviation of 1.72 years. The youngest subject tested was 21 years old and the oldest was 31.

Table 1 presents a description of each of the 25 predictor variables in this study. The abbreviation for each predictor variable is given in column one (PMT stands for Psychomotor Test) and the origin of each variable is given in column two. DLT stands for Dichotic Listening Test, SR for Stick and Rudder, SRT for Stick, Rudder, and Throttle, and HTAD for Horizontal Tracking with Absolute Difference. The formula for deriving the predictor variable from the raw data is given in the final column.

These formulas were designed to (1) put all of the predictor variables onto roughly the same scale, (2) let higher scores represent better performance, and (3) for those variables that reflected multi-tasking, assure that the score did not allow Ss to concentrate on one task to the exclusion of the other task.

For example, PMT4, a multi-tasking test consisting of dichotic listening combined with tracking by stick and rudder input controls, divides a DLT score by an average of log error scores on the two tracking tasks. Raw DLT scores are in the range of about 100, and log tracking scores are in the range of about 4 so this puts PMT4 into a scale centered at about 25. Should DLT performance deteriorate to a raw score of 80 and average tracking error remain constant at 4, then PMT4 decreases to 20. Suppose DLT performance remains constant at 100, and average tracking error increases to 5, then PMT4 again decreases to 20. If a subject were to devote processing resources exclusively to, say, tracking with the stick to maximize performance on that task, but divert attention away from the tracking with the rudder, then his raw stick tracking scores might improve to 3.5 with his rudder tracking score ballooning to 6.5. When these two tracking scores are averaged to a 5, a suitably lower quantitative measure for PMT4 results.

Some of the predictor variables represent repeated sessions on the same task. For example, PMT8 through PMT13 represent six sessions on the horizontal tracking task. Likewise, PMT14 through PMT16 stand for three sessions on the multi-tasking test of horizontal tracking with absolute difference. Finally, PMT17 through PMT20 represent four sessions on the Manikin task, a test of two-dimensional spatial aptitude.

The single dependent variable was a standardized score of flight grades from primary flight training. Primary flight training is conducted at two locations, NAS Whiting Field, Milton, Florida, and NAS Corpus Christi, Corpus Christi, Texas. The flight grade data were obtained from three squadrons, VT-2, VT-3, and VT-6, at Whiting Field and two squadrons at Corpus Christi, VT-27 and VT-28. In addition, seven students received primary flight training at training squadron HT-8 and were destined for helicopter pilot training. These students were included in the analysis as well. By construction, these standardized scores have a mean of 50 and a standard deviation of 10 with a possible range extending from 20 to 80. The standardization process is supposed to even out differences among the various training squadrons in how the grades are assigned.

Sometimes a flight grade was given to a student who attrited from primary flight training. These flight grades were used in the analysis. Other students who attrited were not given a flight grade. Seven of these students who were not assigned a flight grade were coded as DOR (Drop on Request - Assigned to non-aviation role in Navy). It was decided to assign these students low flight grades, i.e., standardized scores below 30, on the supposition that they were failing but given the option of DOR. Other subjects who attrited for non-DOR reasons, such as medical reasons, were not assigned any flight grade and thus were not included in the analysis.

Table 2 is a table of descriptive statistics for all 25 predictor variables and the one dependent variable. It shows the minimum score, maximum score, mean, standard deviation, and sample size.

Table 1: The 25 predictor variables used in the linear regression models to predict primary flight grade.

Predictor Variable	*Test*	*Formula*
PMT 1	DLT	DLT/4
PMT 2	Stick	$100/(\log_{10}(\text{STICK error}))$
PMT 3	DLT+Stick	$\text{DLT}/(\log_{10}(\text{STICK error}))$
PMT 4	DLT+SR	$\text{DLT}/.5 \times (\log_{10}(\text{STICK}) + \log_{10}(\text{RUDDER}))$
PMT 5	Stick+Rudder	$100/.5 \times (\log_{10}(\text{STICK}) + \log_{10}(\text{RUDDER}))$
PMT 6	SRT	$100/.33 \times (\log_{10}(\text{STICK}) + \log_{10}(\text{RUDDER}) + \log_{10}(\text{THROTTLE}))$
PMT 7	Absolute Difference	Number Correct − Number Incorrect
PMT 8	Horizontal Tracking 1	$100/\log_{10}(\text{HT error 1})$
PMT 9	Horizontal Tracking 2	$100/\log_{10}(\text{HT error 2})$
PMT 10	Horizontal Tracking 3	$100/\log_{10}(\text{HT error 3})$
PMT 11	Horizontal Tracking 4	$100/\log_{10}(\text{HT error 4})$
PMT 12	Horizontal Tracking 5	$100/\log_{10}(\text{HT error 5})$
PMT 13	Horizontal Tracking 6	$100/\log_{10}(\text{HT error 6})$
PMT 14	HTAD 1	$(\text{Correct 1} - \text{Incorrect 1})/\log_{10}(\text{HT error 1})$
PMT 15	HTAD 2	$(\text{Correct 2} - \text{Incorrect 2})/\log_{10}(\text{HT error 2})$
PMT 16	HTAD 3	$(\text{Correct 3} - \text{Incorrect 3})/\log_{10}(\text{HT error 3})$
PMT 17	Manikin 1	(Correct 1 − Incorrect 1)
PMT 18	Manikin 2	(Correct 2 − Incorrect 2)
PMT 19	Manikin 3	(Correct 3 − Incorrect 3)
PMT 20	Manikin 4	(Correct 4 − Incorrect 4)
MVT	Math/Verbal	Raw MVT subtest score
MCT	Mechanical Comprehension	Raw MCT subtest score
SAT	Spatial Apperception	Raw SAT subtest score
ANI	Aviation/Nautical Interest	Raw ANI subtest score
API	API NSS	Navy Standard Score at completion of API

Table 2: Descriptive statistics for the 25 predictor variables and the dependent variable of flight grade.

Predictor Variable	*Minimum*	*Maximum*	*Mean*	*SD*	*N*
PMT1	21.00	27.00	25.81	1.14	210
PMT2	20.50	27.76	24.50	1.50	210
PMT3	8.96	26.91	22.44	2.51	210
PMT4	11.60	24.64	20.70	2.33	210
PMT5	20.95	25.70	23.49	0.98	210
PMT6	20.55	25.27	22.97	0.95	210
PMT7	20.00	94.00	53.85	13.68	210
PMT8	20.91	27.59	22.87	1.09	210
PMT9	20.56	30.81	23.66	1.48	210
PMT10	20.28	31.25	23.71	1.48	210
PMT11	20.43	30.50	23.98	1.69	210
PMT12	20.25	37.11	24.12	1.80	210
PMT13	20.20	43.30	24.28	2.11	210
PMT14	0.60	22.07	10.77	3.42	210
PMT15	−10.30	22.10	12.07	3.78	210
PMT16	1.58	24.61	12.53	4.15	210
PMT17	0.00	126.00	68.65	23.65	210
PMT18	15.00	138.00	80.96	21.41	210
PMT19	18.00	146.00	84.27	21.35	210
PMT20	28.00	147.00	87.51	21.72	210
MVT	14.00	36.00	26.73	5.06	210
MCT	14.00	30.00	22.49	3.15	210
ANI	11.00	30.00	19.31	3.18	210
SAT	12.00	35.00	28.90	4.29	210
API	34.00	67.00	52.14	6.42	210
Flight Grade	20.00	80.00	50.29	10.82	210

THE BAYESIAN FORMALISM FOR MODEL EVALUATION

In this section, we develop the equation for the ratio of the posterior probability for any two given models. By Bayes's Theorem, the posterior probability of Model A, $\mathcal{M}_A$, as conditioned upon the observation of the data, D, is

$$P(\mathcal{M}_A|D) = \frac{P(D|\mathcal{M}_A)\ P(\mathcal{M}_A)}{\sum_{i=1}^{K} P(D|\mathcal{M}_i)\ P(\mathcal{M}_i)} \tag{1}$$

where $P(D|\mathcal{M}_A)$ is the probability of the data conditioned upon Model A, otherwise known as the likelihood. $P(\mathcal{M}_A)$ is the prior probability assigned to Model A. The denominator in Bayes's Theorem is the sum of the expression in the numerator over all K models that are being considered.

The posterior probability of a second competing model, Model B, is expressed in exactly the same way.

$$P(\mathcal{M}_B|D) = \frac{P(D|\mathcal{M}_B)\ P(\mathcal{M}_B)}{\sum_{i=1}^{K} P(D|\mathcal{M}_i)\ P(\mathcal{M}_i)} \tag{2}$$

Now, we form the ratio of Equations (1) and (2), i.e., the ratio of the posterior probabilities for Models A and B, to eliminate the denominator in each equation.

$$\frac{P(\mathcal{M}_A|D)}{P(\mathcal{M}_B|D)} = \frac{P(D|\mathcal{M}_A) \times P(\mathcal{M}_A)}{P(D|\mathcal{M}_B) \times P(\mathcal{M}_B)} \tag{3}$$

Equation (3) results in a number that shows the odds in favor of Model A relative to Model B. As a simple numerical example, suppose that the prior probability of the two models are equal so that

$$P(\mathcal{M}_A) = P(\mathcal{M}_B) = 1/2.$$

However, the likelihood of the data under Model A is $P(D|\mathcal{M}_A) = .20$ and the likelihood of the data under Model B is $P(D|\mathcal{M}_A) = .02$. Then we say that the odds are 10:1 in favor of Model A over Model B. The first term on the right-hand side of Equation (3)

$$\frac{P(D|\mathcal{M}_A)}{P(D|\mathcal{M}_B)}$$

is called the *Bayes Factor* and given the notation B_{AB}. The second term in Equation (3)

$$\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)}$$

is called the *prior odds.*

THE BAYESIAN APPROXIMATION TO MODEL LIKELIHOOD

From the above discussion we see that, in order to compute the posterior probability of any model of interest, it is necessary to find the likelihood of the data as conditioned on some given model, $P(D|\mathcal{M}_k)$, as well as the prior probability of that model, $P(\mathcal{M}_k)$. $P(D|\mathcal{M}_k)$ is found by another application of Bayes's Theorem, this time at a lower level where the explicit parameters of the model ($\boldsymbol{\theta}$) are taken into account.

$$P(\boldsymbol{\theta}|D, \mathcal{M}_k) = \frac{P(D|\mathcal{M}_k, \boldsymbol{\theta})\ P(\boldsymbol{\theta}|\mathcal{M}_k)}{\int d\boldsymbol{\theta}\ P(D|\mathcal{M}_k, \boldsymbol{\theta})\ P(\boldsymbol{\theta}|\mathcal{M}_k)}. \tag{4}$$

Because the denominator of Bayes's Theorem in Equation (4) represents the marginalization over all parameter values, it is the value we seek.

$$P(D|\mathcal{M}_k) = \int d\boldsymbol{\theta}\ P(D|\mathcal{M}_k, \boldsymbol{\theta})\ P(\boldsymbol{\theta}|\mathcal{M}_k) \tag{5}$$

The vector of parameters for a given model is indicated by $\boldsymbol{\theta}$. For the linear regression models of interest in this analysis, the vector of parameters is

$$\boldsymbol{\theta} = \{\beta_0, \beta_i, \sigma\} \tag{6}$$

the intercept parameter, the regression coefficients, and the standard deviation of the error. The usual notation for the linear regression model is used where

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{7}$$

$\boldsymbol{Y}$ is the vector containing the dependent variables, here the set of flight grades. $\boldsymbol{X}$ is the matrix of the known predictor variables, here the values of the 20 variables from the CBPT and the 5 cognitive variables from the ASTB and API. The first column of $\boldsymbol{X}$ is filled with 1s for the intercept term. $\boldsymbol{\beta}$ is the vector of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is the error vector assumed to be independently normally distributed with a mean of 0 and a standard deviation of σ.

The Bayes Factor is given the generic notation B_{jk} to indicate the ratio of the likelihood of Model j to the likelihood of Model k. A special model called the null model that uses none of the predictor variables is labeled as Model 0. Therefore, for two models, say Models A and B, the Bayes Factor in favor of the null model is

$$B_{0A} \equiv \frac{P(D|\mathcal{M}_0)}{P(D|\mathcal{M}_A)} \tag{8}$$

and

$$B_{0B} \equiv \frac{P(D|\mathcal{M}_0)}{P(D|\mathcal{M}_B)}. \tag{9}$$

Now if we form the ratio of these two Bayes Factors, the null model cancels out and we are left with the Bayes Factor of Model B compared to Model A. Notice the inversion of which model is compared to the other.

$$\frac{B_{0A}}{B_{0B}} = \frac{\frac{P(D|\mathcal{M}_0)}{P(D|\mathcal{M}_A)}}{\frac{P(D|\mathcal{M}_0)}{P(D|\mathcal{M}_B)}} \tag{10}$$

$$= \frac{P(D|\mathcal{M}_B)}{P(D|\mathcal{M}_A)} \tag{11}$$

$$= B_{BA}. \tag{12}$$

It happens that a logarithmic transform of the left-hand side of Equation (10) is useful for further mathematical manipulations:

$$2\ln\left[\frac{B_{0A}}{B_{0B}}\right] = 2\left[\ln B_{0A} - \ln B_{0B}\right] \tag{13}$$

$$= 2\ln B_{0A} - 2\ln B_{0B} \tag{14}$$

$$2\ln B_{0A} = BIC_A \tag{15}$$

$$2\ln B_{0B} = BIC_B \tag{16}$$

$$2\ln\left[\frac{B_{0A}}{B_{0B}}\right] = BIC_A - BIC_B \tag{17}$$

$$\frac{P(D|\mathcal{M}_B)}{P(D|\mathcal{M}_A)} = \exp\left[1/2\left(BIC_A - BIC_B\right)\right]. \tag{18}$$

Twice the logarithmic transform of the Bayes Factor comparing the null model with some other model is known as the *Bayesian Information Criterion*, here shortened to BIC_A for Model A, BIC_B for Model B, and so on.

For a linear regression model of the kind represented by Equation (7), Raftery [3] has shown that the posterior odds can be approximated by

$$2\ln B_{0A} \equiv BIC_A \tag{19}$$

$$\equiv 2\ln\left[\frac{P(D|\mathcal{M}_0)}{P(D|\mathcal{M}_A)}\right] \tag{20}$$

$$= n\ln(1-R_A^2) + p_A \ln n. \tag{21}$$

Here $\mathcal{M}_0$ is the notation for the null model with no predictor variables. n is the number of subjects in the analysis, R_A^2 is the squared sample multiple correlation coefficient for model A, and p_A is the number of predictor variables (not including the intercept) in model A. p_A will range from 1 to 25 depending on how many predictor variables are included in any particular model.

If we take as a simplifying assumption that the prior probabilities of all models are equal, then the posterior probability of the kth model can then be written as

$$P(\mathcal{M}_k|D) \approx \frac{\exp(-1/2\ BIC_k)}{\sum_{i=1}^{K}\exp(-1/2\ BIC_i)}. \tag{22}$$

In the denominator, we are summing over K models, which may not necessarily be all possible models, but instead those for which some appreciable probability exists.

Numerical examples

In this section, we present some simple numerical examples to illustrate the use of Equations (21) and (22) and to prepare for their use on the actual data. First, look at Table 3 which contains some numerical values for computing Equation (21).

Table 3: A numerical example for computing Equation (21).

Model	p_k	R_k^2	$1-R_k^2$	$n\ln(1-R_k^2)$	$p_k \ln n$	BIC_k
$\mathcal{M}_A$	5	.25	.75	−42.86	25	−17.86
$\mathcal{M}_B$	10	.40	.60	−76.11	50	−26.11
$\mathcal{M}_C$	15	.45	.55	−89.08	75	−14.08
$\mathcal{M}_D$	20	.48	.52	−97.44	100	+2.56

We are comparing only four models in this exercise, Models A, B, C and D, as shown in the first column. The second column gives the number of predictor variables in the regression equation. The third column shows the resulting squared multiple correlation coefficient. As usual, R^2 will increase with the addition of extra predictor variables. The question is, however, *"Are we overfitting the model by including extra predictor variables, and if so, how can we penalize for the inclusion of the extra variables that are not really making a genuine contribution?"* The sixth column shows the effect of the penalty term in the approximation to the likelihood of the data given the model. We chose $n = 149$ in this numerical example simply because $\ln(149) \approx 5$. So we multiply the number of predictor variables in the regression equation by 5 to yield the penalty term. When we add the fifth and sixth columns, we get in the final column the BIC value for each of the four models.

The BIC value for the null model with no predictor variables is equal to 0. Therefore, any model with a negative value in the last column is better than the null model. Models A, B, and C are observed to be better models than the null model. On the other hand, any model with a positive value is worse than the null model. Model D with 20 predictor variables, therefore, is not preferred even to the model with no predictor variables. This kind of behavior exhibits the penalty imposed for a large number predictor variables without a sufficiently large compensating increase in R^2.

Among the good models, $\mathcal{M}_A$, $\mathcal{M}_B$, and $\mathcal{M}_C$, the more negative the BIC value, the better the model. Therefore, Model A is better than Model C, but Model B is better than both Model A and Model C. How much better is Model B than Model A?

$$\begin{aligned}
2\ln\left[\frac{P(D|\mathcal{M}_B)}{P(D|\mathcal{M}_A)}\right] &= BIC_A - BIC_B \\
&= -17.86 - (-26.11) \\
&= 8.25 \\
\frac{P(D|\mathcal{M}_B)}{P(D|\mathcal{M}_A)} &= e^{4.125} \\
&= 61.87 \\
\frac{P(\mathcal{M}_B|D)}{P(\mathcal{M}_A|D)} &= \frac{P(D|\mathcal{M}_B) \times P(\mathcal{M}_B)}{P(D|\mathcal{M}_A) \times P(\mathcal{M}_A)} \\
&= 61.87 \times \frac{1/4}{1/4} \\
&= 61.87.
\end{aligned}$$

It turns out that the odds in favor of Model B with 10 predictor variables over Model A with 5 predictor variables is approximately 62:1. By doing a similar calculation, the posterior odds in favor of Model A over Model C is only about 7:1.

$$\begin{aligned}
2\ln\left[\frac{P(D|\mathcal{M}_A)}{P(D|\mathcal{M}_C)}\right] &= BIC_C - BIC_A \\
&= -14.08 - (-17.86) \\
&= 3.78 \\
\frac{P(D|\mathcal{M}_A)}{P(D|\mathcal{M}_C)} &= e^{1.89} \\
&= 6.62 \\
\frac{P(\mathcal{M}_A|D)}{P(\mathcal{M}_C|D)} &= \frac{P(D|\mathcal{M}_A) \times P(\mathcal{M}_A)}{P(D|\mathcal{M}_C) \times P(\mathcal{M}_C)} \\
&= 6.62 \times \frac{1/4}{1/4} \\
&= 6.62.
\end{aligned}$$

If one prefers, the posterior probability for each model can be reported instead of the ratio of any two models. To find, for example, the posterior probability of Model B we use Equation (22).

$$
\begin{aligned}
P(\mathcal{M}_B|D) &= \frac{e^{-1/2\,BIC_B}}{\sum_{i=i}^{K} e^{-1/2\,BIC_i}} \\
&= \frac{e^{-1/2\,(-26.11)}}{e^{-1/2\,(-26.11)} + e^{-1/2\,(-17.86)} + e^{-1/2\,(-14.08)} + e^{-1/2\,(+2.56)}} \\
&= \frac{4.67 \times 10^5}{(4.67 \times 10^5) + (7.56 \times 10^3) + (1.14 \times 10^3) + .278} \\
&= \frac{4.67 \times 10^5}{4.76 \times 10^5} \\
&= .9811.
\end{aligned}
$$

In like manner, the posterior probabilities of Models A and C can be found as

$$P(\mathcal{M}_A|D) = .0159$$

and

$$P(\mathcal{M}_C|D) = .0030.$$

The posterior probability of Model D is negligible, so the posterior probabilities of Models A, B, and C should together add up to 1. In this example, it is seen that Model B is by far the most likely model, after the data have been gathered, of the four models under consideration.

PROBABILITY OF SOME SELECTED MODELS FOR PREDICTING FLIGHT GRADES

In this section, we turn to the analysis of the data described in the second section. There is a grand total of $2^{25} = 33,554,432$ possible linear regression models for this set of predictor variables. It is obvious that we cannot examine them all. However, we can obtain a fairly good idea of where the high probability models are located and concentrate our search effort in that vicinity.

The statistical software package *SPSS Version 9.0* was used to perform linear regression on specified subsets of the 25 predictor variables. When referring to some given subset, we call it the kth model. The value of the sample squared multiple correlation coefficient for the kth model, R_k^2, as produced by SPSS was used as input to Equation (21). An initial effort running models with different number of predictor variables showed that the good models were concentrated on models with three or four predictor variables.

Table 4 presents a listing of 18 models culled from this kind of nonexhaustive search. The model number is listed in the first column. The second column contains the names of the variables in the specified model. API is the standardized score at the completion of ground school. PMTn refers to the nth psychomotor variable from the CBPT. A brief description of each of the 20 psychomotor variables was given in Table 1. MVT, SAT, and ANI refer to subtests of the ASTB. MVT is the Math/Verbal subtest, SAT is the Spatial Apperception subtest, and ANI is the Aviation/Nautical Interest subtest. The third column shows the number of predictor variables in the kth model, while the fourth column shows the squared sample multiple correlation coefficient. The fifth column lists the BIC value as computed from Equation (21). The more negative a BIC value, the better the model. A positive BIC value indicates a model that is worse than the null model. Finally, in the last column is the posterior probability for the kth model, $P(\mathcal{M}_k|D)$, as computed by Equation (22). Here, the set of models summed over in the denominator of Equation (22) is $K = 18$ in number. The sum of the posterior probabilities of these models considered in Table 4 must equal 1. Many models were examined with low posterior probabilities (BIC values less negative than, say, -60) and are not shown in this table. The only exception was made for Model 1.

Table 4: The posterior probabilities of some linear regression models which will be used for predicting primary flight grade.

k	Model	p_k	R_k^2	BIC_k	$P(\mathcal{M}_k \mid D)$
1	API	1	.239	−52.01	.0000
2	API,PMT2	2	.320	−70.29	.1613
3	API,PMT6	2	.303	−65.11	.0121
4	API,PMT2,MVT	3	.343	−72.17	.4126
5	API,PMT6,MVT	3	.327	−67.12	.0330
6	API,PMT2,PMT6	3	.327	−67.12	.0330
7	API,PMT6,SAT	3	.308	−61.27	.0018
8	API,PMT2,PMT20	3	.320	−64.95	.0111
9	API,PMT2,PMT14	3	.321	−65.26	.0130
10	API,PMT2,MVT,ANI	4	.351	−69.40	.1031
11	API,PMT2,MVT,PMT6	4	.349	−68.75	.0746
12	API,PMT2,MVT,SAT	4	.346	−67.79	.0460
13	API,PMT2,MVT,PMT14	4	.345	−67.47	.0392
14	API,PMT2,MVT,PMT20	4	.343	−66.83	.0285
15	API,PMT6,MVT,ANI	4	.339	−65.55	.0151
16	API,PMT2,MVT,ANI,PMT14	5	.355	−65.35	.0136
17	API,PMT2,MVT,ANI,PMT6,PMT14	6	.360	−61.64	.0021
18	All 25 Variables	25	.412	+22.16	.0000
				Sum	1.0000

API, the final standardized grade from ground school, was the single best predictor and appears in all the models. Perhaps the importance of API is due to the fact that it is the variable closest in time to actual flight training. However, when used in a regression by itself, its posterior probability is negligible as shown in line 1. $R_1^2 = .239$ is not sufficiently large compared to the increase that occurs when extra variables are added to the regression equation. In line 2, with the addition of PMT2, the psychomotor tracking variable using only the joystick to center the cursor, R_2^2 jumps to .320. The posterior probability of this model is the second best at $P(\mathcal{M}_2|D) = .1613$. If one were to consider substituting PMT6, the psychomotor tracking variable using the stick, throttle, and rudder to center three cursors, for PMT2 as in $\mathcal{M}_3$, BIC_3 increases to -65.11. Thus, with this increase in the BIC, the posterior probability for this third model drops to .0121.

Next, we look at some good models with three predictor variables. $\mathcal{M}_4$ with API, PMT2 and MVT is the best model of all because it has the lowest BIC value of $BIC_4 = -72.17$. Consequently, it also possesses the highest posterior probability by far of .4126. This fact, of course, was not immediately evident until many other regressions with more than three predictor variables were examined. Other various models with three predictors were analyzed, and a sampling of the ones with any appreciable posterior probability are shown as the next five models.

As the analysis proceeded, we bumped up to four variables in the regression equation. As can be observed, R_{10}^2 through R_{14}^2 are all greater or equal to $R_4^2 = .343$ of Model 4. The Bayesian modeling approach penalizes these models for including an additional variable. Since the increase in R_k^2 is not sufficient to overcome this penalty, their posterior probabilities are not as high as $\mathcal{M}_4$. Nonetheless, these models would contribute to any final prediction with a weight proportional to their respective posterior probabilities. $\mathcal{M}_{10}$ is the third best model and achieves this by including the ANI subtest to the three variables already in $\mathcal{M}_4$.

Attempts to find good models with five, six, or any greater number of predictor variables were not successful. $\mathcal{M}_{16}$ and $\mathcal{M}_{17}$ are examples where trying to add additional variables to increase R_k^2 did not pan out. For example,

adding PMT14 to the best four predictor model resulted in a negligible increase to R^2_{16}. Such a model would have only a small impact in the final prediction. As the ultimate example of trying to add predictor variables, we present Model 18 in the final line. This model includes all 25 predictor variables and suffers a severe penalty for doing so. BIC_{25} is positive, indicating that such a model is even worse than the null model, which does not include any predictor variables.

THE POSTERIOR ODDS FOR SELECTED MODELS

Presenting the results of the analysis in terms of the posterior probability of some selected models is preferable, but one can also calculate the posterior odds for any two models if so desired. For example, the posterior odds of the best model, $\mathcal{M}_4$, which includes API, PMT2, and MVT can be compared to model $\mathcal{M}_1$ which consists of just API.

$$\begin{aligned} 2\ln\left[\frac{P(\mathcal{M}_4|D)}{P(\mathcal{M}_1|D)}\right] &= BIC_1 - BIC_4 \\ &= -52.01 - (-72.17) \\ &= 20.16 \\ \frac{P(\mathcal{M}_4|D)}{P(\mathcal{M}_1|D)} &= e^{10.08} \\ &= 23,861. \end{aligned}$$

The odds are therefore overwhelmingly in favor of a model that includes these three predictor variables when compared to the model that has just one of them.

As a less extreme example, consider comparing the relative merits of two of the four predictor variable models, $\mathcal{M}_{13}$ and $\mathcal{M}_{14}$. $\mathcal{M}_{14}$ substitutes PMT20, two- dimensional spatial aptitude taken during the fourth session, for PMT14, horizontal tracking with absolute difference on the first session, into the set of variables for the best model.

$$\begin{aligned} 2\ln\left[\frac{P(\mathcal{M}_{13}|D)}{P(\mathcal{M}_{14}|D)}\right] &= BIC_{14} - BIC_{13} \\ &= -66.83 - (-67.47) \\ &= .64 \\ \frac{P(\mathcal{M}_{13}|D)}{P(\mathcal{M}_{14}|D)} &= e^{.32} \\ &= 1.38. \end{aligned}$$

In this case, there is really nothing to distinguish between these two models. One is as good as the other, or stated in different terms, the predictions from both models would be weighted almost evenly in any kind of averaging over models.

It is perhaps worthwhile to mention the distinction between classical hypothesis testing and the kind of Bayesian model evaluation we have done here. Classical hypothesis testing sets up two models, one the *null hypothesis* and the other the *alternative hypothesis*. Then, if a statistic with the appropriate properties can be found, the null hypothesis may be rejected. The null hypothesis is never "accepted" in the classical approach, and it is hard to understand how one might find support for the null hypothesis within this traditional framework.

In contrast, the Bayesian approach simply tabulates the quantitative evidence in favor of any one hypothesis over another competing hypothesis. This is certainly a more compelling intuitive rationale than the classical

methods. The state of knowledge available about the hypotheses before the data were gathered is updated from the prior probability to reflect the new information contained in the data. Hypotheses are neither rejected nor accepted; the odds in favor of any one hypothesis are simply adjusted up or down as the data dictate. Of course, the odds in favor of any one hypothesis over a competing hypothesis may become so large that, from a pragmatic point of view, the hypothesis is effectively rejected. We saw this in the example of $\mathcal{M}_4$ compared to $\mathcal{M}_1$. The posterior probability of $\mathcal{M}_1$ is so low that its contribution to any averaging process will be inconsequential.

PREDICTION BY AVERAGING OVER MODELS

This report concludes with an example of forming a prediction about a candidate's flight grade by averaging over a set of good linear regression models. The averaging is accomplished by weighting the selected models according to their posterior probabilities. Table 5 shows 7 out of the 18 models of Table 4 with the highest posterior probabilities. Just 7 models are selected to make the numerical example easy to follow. Normally, many more models than this would enter into the averaging procedure.

Table 5: The constant intercept term, β_0, and the regression coefficients, β_i, for seven models with the highest posterior probability.

$\mathcal{M}_k$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Model	Constant	API	PMT2	MVT	ANI	PMT6	SAT	PMT14	PMT20
$\mathcal{M}_4$	−34.271	.886	1.944	-.346	*	*	*	*	*
$\mathcal{M}_2$	−41.027	.781	2.065	*	*	*	*	*	*
$\mathcal{M}_{10}$	−38.126	.857	1.899	-.331	.314	*	*	*	*
$\mathcal{M}_{11}$	−48.532	.861	1.432	-.341	*	1.217	*	*	*
$\mathcal{M}_{12}$	−38.553	.879	1.939	-.331	*	*	.151	*	*
$\mathcal{M}_{13}$	−32.634	.880	1.842	-.376	*	*	*	.182	*
$\mathcal{M}_{14}$	−34.165	.884	1.930	-.349	*	*	*	*	.004

The estimate of the intercept term, β_0, and the estimates of the regression parameters, β_i, for each of the seven models are presented. These are the intercept term and the unstandardized regression coefficients as reported by SPSS. It can be seen that each model uses a different number or different set of predictor variables.

Suppose, for example, that we want to predict the standardized flight grade for a candidate who has taken the CBPT and has just completed API. We know his/her scores on all the predictor variables that play a role in the seven models. For the sake of a numerical computation, let's say that these values are as follows:

Predictor Variable	Score
API	45
PMT2	25
MVT	35
ANI	20
PMT6	22
SAT	31
PMT14	10
PMT20	100

Each of the seven models makes a prediction of primary flight grade given these scores. Each of these individual predictions is then averaged to form a final prediction. The weight used in this average is the model's posterior probability. Since we are picking out only the top seven models, we need to renormalize their posterior probabilities so that together they sum to 1.

Let $\tilde{y}_k$ be the notation for the prediction based on the linear regression of the kth model. $P(\mathcal{M}_k|D)$ is, as before, the posterior probability of the kth model. The average of the individual model predictions is then

$$\tilde{y} = \sum_{k=1}^{7} \tilde{y}_k \, P(\mathcal{M}_k|D). \tag{23}$$

Table 6 presents the individual predictions for each of the seven models. The next to last column shows the posterior probability for that model as renormalized. This value serves then as the weighting value in the averaging process indicated by Equation (23). The global prediction for this student's standardized flight grade based on Bayesian model averaging is shown in the final row as 42.73.

Table 6: Forming a global prediction by Bayesian model averaging from the individual predictions of seven models.

$\mathcal{M}_k$	$\tilde{y}_k$	$P(\mathcal{M}_k \mid D)$	Eq. (23)
$\mathcal{M}_4$	42.09	.4768	20.0685
$\mathcal{M}_2$	45.74	.1864	8.5259
$\mathcal{M}_{10}$	42.61	.1191	5.0749
$\mathcal{M}_{11}$	40.85	.0862	3.5213
$\mathcal{M}_{12}$	42.57	.0532	2.2647
$\mathcal{M}_{13}$	41.68	.0453	1.8881
$\mathcal{M}_{14}$	42.05	.0329	1.3834
Sums		1.0000	42.7268

DISCUSSION

A legitimate question could be raised about whether the very best models have, in fact, been found. The search through the space of all models was nonexhaustive, and it was only because a few models with a large number of predictor variables were found to possess low probabilities that the inference was drawn that all such models suffered from this same defect.

We address this concern in a subsequent report [5]. The 25 predictor variables examined in this study were collapsed down into 8 predictor variables. This reduced set of 8 predictor variables did, however, sample the same skills as the larger set. We then calculated the posterior probability of all 256 possible linear regression models based on this reduced set of 8 variables. As a result of this exhaustive search through the space of models, only a few of the 256 models possessed any significant probability.

These models with significant probability were essentially the same ones identified in this report. That is, models consisting of the API score, the single psychomotor tracking variable, and composites of the ASTB that included the MVT and ANI subtests were the preferred models. This result lends support to the conjecture that, from the over 33 million possible models with 25 predictor variables, the few models that have been highlighted here are really the significant models one ought to be concerned about.

REFERENCES

1. Raftery, A.E., Madigan, D., and Hoeting, J.A. Model selection and accounting for model uncertainty in linear regression models. *J. of the Am. Stat. Assoc.,* 92:179-191,1997.
2. Kass, R.E., and Raftery, A.E. Bayes Factors. *J. of the Am. Stat. Assoc.,* 90:773-795, 1995.
3. Raftery, A.E. Bayesian model selection in social research (with Discussion). In *Sociological Methodology 1995,* ed. by P.V. Marsden, pp. 111–163, Blackwell Publishers, Cambridge, MA, 1995.
4. Blower, D. J. and Dolgin, D.L. An Evaluation of Performance-Based Tests Designed to Improve Naval Aviation Selection. *NAMRL-1363,* Naval Aerospace Medical Research Laboratory, Pensacola, FL, August 1991.
5. Blower, D. J., Albert, A. O., and Williams, H. P. Predicting Primary Flight Grades by Averaging Over Linear Regression Models: Part 2. NAMRL Technical Report in review, Naval Aerospace Medical Research Laboratory, Pensacola, FL, March 2000.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1 January 2000	

4. TITLE AND SUBTITLE
Predicting Primary Flight Grades by Averaging Over Linear Regression Models: Part I

5. FUNDING NUMBERS
62233N.0330.126-7801

6. AUTHOR(S)
D. J. Blower, H. P. Williams and A. O. Albert

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Aerospace Medical Research Laboratory
51 Hovey Road
Pensacola Fl 32508-1046

8. PERFORMING ORGANIZATION REPORT NUMBER
NAMRL-1410

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217-5660

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Approved for public release; distribution unlimited.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

12b. DISTRIBUTION CODE

12. ABSTRACT *(Maximum 200 words)*

This report documents an investigation into two types of variables that might be useful in predicting flight grades in Navy primary flight training. The first set of predictor variables is largely psychomotor in origin and is part of the Computer-Based Performance Test Battery at the Naval Aerospace Medical Research Laboratory. The second set of variables is more cognitive in nature and arises from scores on the Aviation Selection Test Battery (ASTB) and a final grade in Aviation Pre-Flight Indoctrination (API), which is ground school prior to entering primary flight training. The motivation for this research is a joint effort with the Air Force designed to improve selection tests for military aviators. The emphasis in this report is on how to choose good linear regression models which use these variables to predict a criterion variable such as flight grade. In our present case, we have a total of 25 potential predictor variables. As a result, there is a rather large number of possible regression models. Our task is to pick some relatively small number of models that are "best" by some acceptable statistical criterion. The analysis revealed that models with a small number of predictor variables were much superior to models that included a large number of the 25 available variables. The best models consisted of two, three, and four predictor variables and possessed an R^2 of about .35. The single best model contained the final grade from API, a psychomotor tracking variable, and a score from one of the ASTB subtests. A prediction of the flight grade can then be made by averaging over the individual predictions of the single models. Using Bayesian model evaluation techniques, the averaging is carried out by weighting each individual model according to its posterior probability.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Personnel selection, Linear models, Bayesian model evaluation, prediction	21
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR