In Pursuit of Higher Power Through Integrated Multivariate Regression

Ryan Shahbaba[†]

Project advisor: Professor Annie Qu[‡]

Abstract. Univariate regression models are commonly used in statistics and machine learning to examine the relationship between an outcome variable and a set of explanatory variables, and possibly use this relationship to predict the unknown values of the outcome variable. However, when dealing with multiple outcome variables that are interrelated, multivariate regression models are preferred. These models simultaneously capture the dependencies between outcome variables and their collective relationships with explanatory variables. While multivariate regression models provide a rigorous and comprehensive understanding of factors associated with outcomes of interest, they have several limitations including: increased model complexity, larger sample size requirements, and lack of interpretability. To address these issues, we propose an alternative approach, called Integrated Multivariate Regression (IMR) that reduces the dimensionality of the outcome variables by transforming them into one or more derived outcome variables that retain important information. Using simulated and real data, we demonstrate that IMR simplifies the analysis and increases statistical power by reducing the number of parameters, while simultaneously maintaining interpretability and accounting for interdependencies among the outcome variables.

1. Introduction.

In statistics and machine learning, univariate regression models are commonly used to capture the relationship between an outcome of interest (also known as response variable or dependent variable) and one or more explanatory variables (also known as covariates or independent variables). However, many scientific problems involve the analysis of multiple outcome variables measured simultaneously. For example, we may measure multiple cognitive tests for a group of subjects and examine the associations between these measurements and specific demographic and clinical factors. Other instances include evaluating variations in expression values of numerous genes in response to different experimental conditions, assessing improvements across various health outcomes for different treatments, and analyzing changes in various crime rates in response to a new policy. Real-life applications of such nature can be challenging to analyze, particularly when dealing with small sample sizes, which is often the case. Hence, it is crucial to develop new methods that are flexible and powerful enough to handle the complexity of regression problems with multiple outcomes.

For such problems, we could simply use *individual regression (IR)* models, i.e., one regression model for each outcome. As we show later, such an approach is not efficient – it lacks statistical power since it ignores the underlying relationship among the outcome variables. Alternatively, we can use *multivariate regression (MR)* models [13], which are particularly useful when the outcome variables are interrelated and may exhibit dependencies that need to be accounted for in the analysis. Multivariate regression allows for modeling the interdependencies between the outcome variables and capturing their collective relationships with the explanatory variables. By considering multiple outcome variables simultaneously, researchers can gain a more comprehensive understanding of the corresponding scientific problem. However,

[†]Sage Hill School, Newport Beach, CA, USA (ryanshahbaba@gmail.com)
[‡]Department of Statistics, UC Irvine, CA, USA (aqu2@uci.edu)

in practice, multivariate regression models have some limitations. 1) With multiple outcomes, the complexity of the regression model increases. As the number of outcome variables grows, the model's parameters, including the regression coefficients, also increase. This complexity can lead to challenges in fitting the model (see Section 4 for real examples). 2) Interpreting the results of multivariate regression models with multiple outcomes can be more challenging compared to models with a single outcome. In these models, the coefficients capture the relationship between the explanatory variables and each outcome, considering the other outcomes' influence. Interpreting the individual coefficients requires careful consideration and understanding of the underlying relationships and dependencies between the outcomes. Additionally, presenting and communicating the results of multivariate regression models in a meaningful and concise manner can be complex. 3) Multivariate regression models generally require larger sample sizes compared to models with a single outcome because of their relatively large number of parameters. Insufficient sample sizes may result in unstable or unreliable models, limiting their generalizability and usefulness.

To address these issues, one possible solution is to use canonical correlation analysis (CCA) [12]. CCA reduces the number of parameters while exploring and quantifying the relationship between two sets of variables, namely explanatory and outcome variables. Its objective is to identify the linear combinations of variables from each set that exhibit the highest correlation with each other. By doing so, CCA aims to provide insights into the underlying associations and dependencies between two multivariate datasets. However, interpreting the results can still be challenging due to the complex nature of the analysis. Additionally, CCA treats outcome and explanatory variables equally, disregarding the central role of outcome variables as the primary focus of the analysis.

In this paper, we propose a simpler and more interpretable approach called *integrated multivariate regression (IMR)*. Similar to standard multivariate regression, IMR accounts for the interdependencies among multiple outcomes. However, in contrast, IMR achieves this while utilizing a substantially smaller number of parameters to increase statistical power. More specifically, IMR relies on Principal Component Analysis (PCA), which is a statistical technique used for dimensionality reduction. While PCA is commonly used to reduce the dimensionality of explanatory variables, in this study, we employ it to reduce the dimensionality of the outcome variables. Our objective is to transform the multiple outcome variables into a single derived outcome variable that preserves as much information as possible. Specifically, we define our single outcome variable as the first principal component, which captures the maximum variance among the multiple outcome measurements. If necessary, we can utilize two or more principal components to retain additional information, while still maintaining a significantly lower number of parameters compared to standard multivariate regression.

This paper is organized as follows. In Section 2, we provide a brief review of individual regression (IR), multivariate regression (MR), and principal component analysis (PCA). We then explain our proposed integrated multivariate regression (IMR) method in detail. In Section 3, we evaluate the performance of our method using a set of simulation studies. Section 4 focuses on the application of IMR to two real problems. The first problem involves examining the association between multiple cognitive tests and a set of demographic and clinical variables. The second problem pertains to evaluating the effect of a specific genotype on 120 genes potentially involved in nutritional problems. Finally, in Section 5, we discuss

the advantages and limitations of our method and briefly explore possible future directions.

2. Method.

Individual Regression Model. Standard univariate regression with a single outcome variable and multiple explanatory variables can be presented as follows:

$$E(Y|\mathbf{X}) = \beta_0 + \mathbf{X}\beta, \tag{2.1}$$

which models the expectation (mean) of the outcome variable, Y, for a given set of explanatory variables, $\mathbf{X}$, as a linear combination of them. Here, Y is an $n \times 1$ vector of the outcome values for n data points, $\mathbf{X}$ is a matrix of $n \times p$ for p explanatory variables, β_0 is the intercept, and β is a $p \times 1$ vector of regression coefficients. Alternatively, we can write the model as follows:

$$E(Y|\mathbf{X}) = \mathbf{X}\beta. \tag{2.2}$$

In this case, $\mathbf{X}$ is an $n \times (p+1)$ matrix, where the first column is one, and β is a $(p+1) \times 1$ vector, where the first element is the intercept.

The regression parameters are obtained using the least squares estimates. The significance of each explanatory variable, X_j, can be evaluated based on its corresponding regression coefficient, β_j, by testing the null hypothesis $\beta_j = 0$. For this, we can use Analysis of Variance (ANOVA). When there are multiple outcome variables, we can simply use a separate univariate regression model for each outcome. We refer to this approach as individual regression (IR) model.

Multivariate Regression Model. As discussed in the Introduction section, when there are multiple, m, outcome variables, it is more appropriate to use a multivariate regression (MR) model. In this case, we rewrite Equation 2.1 as follows:

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}. \tag{2.3}$$

Here, $\mathbf{Y}$ is an $n \times m$ matrix of the outcome values, $\mathbf{X}$ is an $n \times (p+1)$ matrix, and $\boldsymbol{\beta}$ is a matrix of $(p+1) \times m$ regression parameters. As before, the regression parameters β_0 and β can be obtained using the least squares estimates. Then, multivariate analysis of variance (MANOVA) can be used to evaluate the overall and collective association of a specific explanatory variable with the set of outcome variables.

Principal Component Analysis (PCA). As mentioned above, we propose to use PCA in order to reduce the dimensionality of the outcome variables. The concept of PCA was first introduced by Pearson [18] to study the lines and planes of closest fit for high-dimensional data by finding the direction that maximizes the variance in multivariate data. Later, Hotelling extended this concept, established its statistical and mathematical foundation, and showed that PCA is equivalent to finding the eigenvectors and eigenvalues of the covariance matrix of multivariate data [11].

Consider a set of outcome variables, $Y_1, \ldots, Y_m$, which are standardized to have zero means and variances of one. We denote the centered matrix of observed data as $y_1, \ldots, y_m$. The principal components are a set of orthonormal basis, $v_1, v_2, \ldots, v_m$, such that v_1 is the

basis with the largest sample variance, v_2 is the basis with the second largest sample variance that it is orthogonal to v_1, v_3 is the basis with the third largest sample variance that it is orthogonal to v_1, v_2, and so forth.

To find these principal components, we first calculate their covariance matrix $\boldsymbol{S}$ and its corresponding eigenvectors: $v_1, v_2, \ldots, v_m$. Then, we order the eigenvectors based on the descending order of their corresponding eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_m$,

$$\boldsymbol{S}v_j = \lambda_j v_j, \qquad j = 1, \ldots, m, \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m.$$

The eigenvectors represent the directions in the original feature space, and the corresponding eigenvalues indicate the amount of variance explained by each component. The eigenvectors are then used to project the data onto the new coordinate system defined by the principal components. By selecting a subset, $q \ll m$, of the principal components, one can effectively reduce the dimensionality of the data while preserving the most informative aspects. Here, we set $q = 1$ to select the first principal component, denoted as Z, which can be written as a weighted sum of the original variables:

$$Z = v_{11}Y_1 + v_{12}Y_2 + \ldots + v_{1m}Y_m, \tag{2.4}$$

where $v_{11}, \ldots, v_{1m}$ are the elements of the first eigenvector (i.e., the eigenvector with the highest eigenvalue, λ_1).

Integrated Multivariate Regression. Given the first principal component of the outcome variables, we reduce Equation 2.3 to the simple form presented in Equation 2.2:

$$E(Z|\mathbf{X}) = \mathbf{X}\beta. \tag{2.5}$$

From Equation 2.4, we have

$$E[Z|\mathbf{X}] = v_{11}E[Y_1|\mathbf{X}] + v_{12}E[Y_2|\mathbf{X}] + \ldots + v_{1m}E[Y_m|\mathbf{X}]. \tag{2.6}$$

Therefore, we can rewrite Equation 2.2 as

$$v_{11}E[Y_1|\mathbf{X}] + v_{12}E[Y_2|\mathbf{X}] + \ldots + v_{1m}E[Y_m|\mathbf{X}] = \mathbf{X}\beta. \tag{2.7}$$

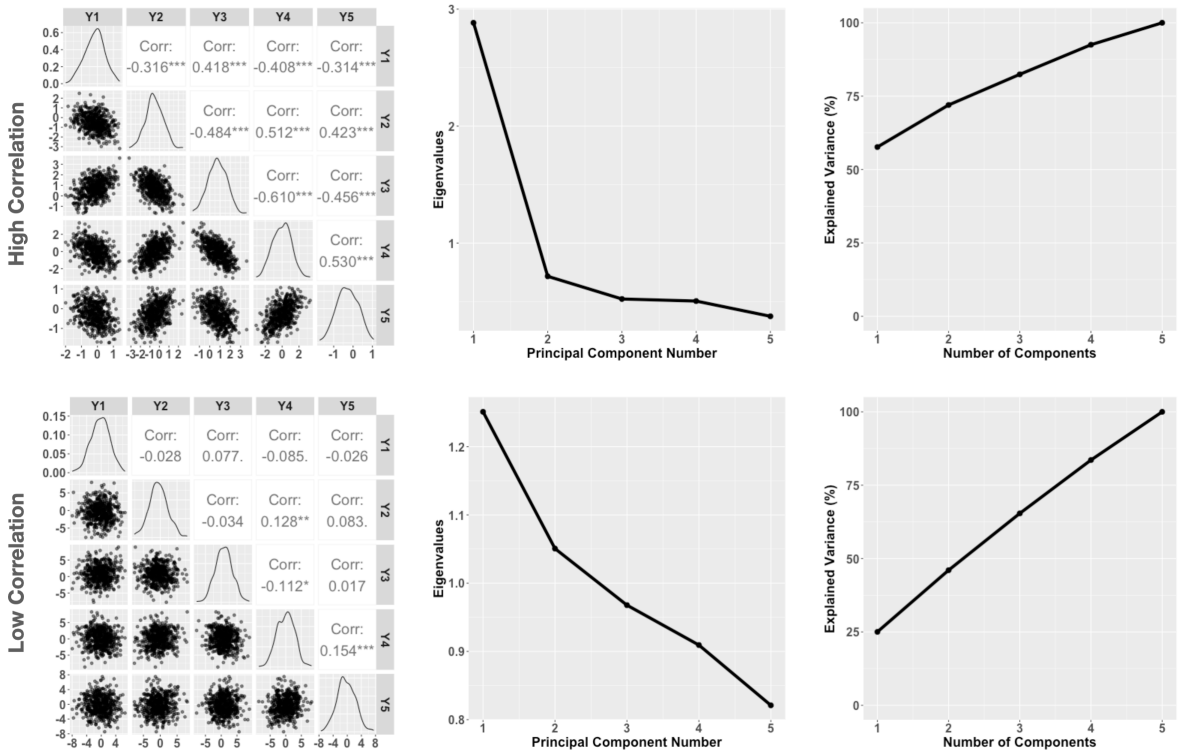
Then, for Y_1, we have

$$\begin{aligned} E[Y_1|\mathbf{X}] &= \frac{1}{v_{11}}\{\mathbf{X}\beta - (v_{12}E[Y_2|\mathbf{X}] + \ldots + v_{1m}E[Y_m|\mathbf{X}])\} \\ &= \mathbf{X}\beta^*. \end{aligned}$$

Similar equations can be written for $Y_2, \ldots, Y_m$. Note that the resulting model for each outcome variable itself is in the form of a univariate regression (Equation 2.2), but its parameters are not the same as what we would obtain by fitting individual regression (IR) models.

We can extend this approach to cases where the first principal component alone does not capture a significant portion of the variance in the data. In such situations, we have the option to include multiple principal components in the analysis. By doing so, we can still effectively reduce the dimensionality of the outcome variables and minimize the number of required parameters. However, it is important to note that the resulting model may not be as straightforward compared to using only the first principal component.

Figure 1. *Correlation matrix (left), scree plot (middle), and cumulative proportion of variance explained by the principal components for simulated data with high (top row) and low (bottom row) correlations among the outcome variables.*



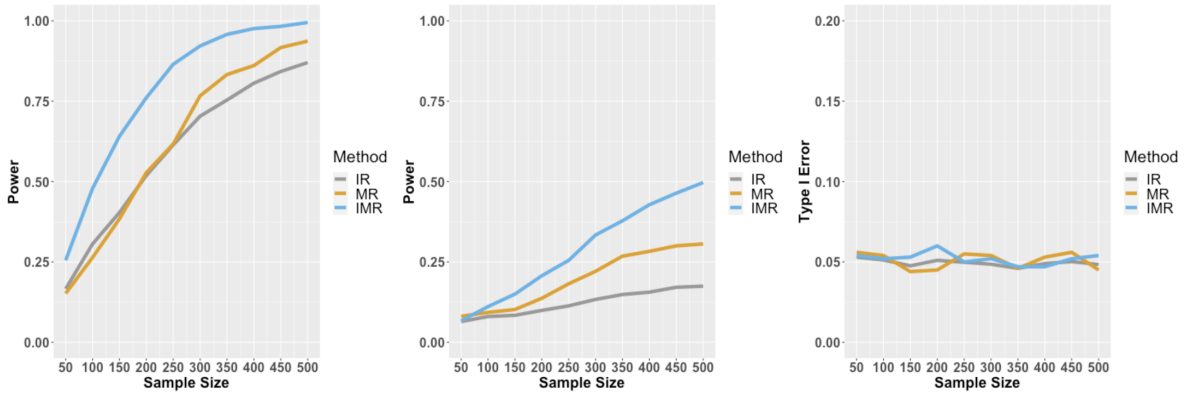
3. Simulation Results. In this section, we evaluate the performance of our proposed method using simulation studies. We assume that we are interested in the association between a genetic factor, denoted as a binary variable X (e.g., 1 for allele A and 0 for allele a), and the overall health, denoted as Z, in a population. While we cannot directly observe Z, we can use five different medical tests, $Y_1, \ldots, Y_5$ to measure it indirectly. We expect these measurements to be correlated since they all represent the overall health of a subject. More specifically, we simulate data as follows:

$$\begin{aligned} X &\sim \text{Bernoulli}(0.25), \\ Z &\sim N(1 + \beta X, 1), \\ Y_j &\sim N(a_j + b_j Z, s_j^2), \quad j = 1, \ldots, 5. \end{aligned}$$

Here, we assume that 25% of the population have allele A and their overall measurement of health, Z, increases by β on average. We assume that Z and the medical tests are normally distributed. For each medical test, the average measurement is $a_j + b_j Z$ (i.e., a function of Z, where b_j captures the strength of relationship) and the measurement noise is controlled by s_j.

We conduct three simulation studies. In each simulation, we use the above model to generate 1000 simulated datasets. For each simulated data, we compare our method (IMR) to five individual regression (IR) models, and a multivariate regression (MR) model that captures

Figure 2. *Left panel: Statistical power comparison among the three alternative methods using simulated data with highly correlated outcome variables. Middle panel: Statistical power comparison when the outcome variables have low correlations. Right panel: Comparing the Type I Error rates among the three methods.*



the effect of the gene on the five outcome variables simultaneously using the Pillai test statistic [19] implemented in the `car` package [8] in R. For all models, we set the significance level at 0.05 and reject the null hypothesis if p-value < 0.05. We compare the three models in terms of their *statistical power*, which refers to the percentage of times (across 1000 simulated datasets) that a model correctly rejects the null hypothesis. Additionally, we examine their *Type I error*, which represents the percentage of times (across 1000 simulated datasets) that a model incorrectly rejects the null hypothesis.

For Simulation I, we assume $\beta = 0.5$ so the genetic factor is indeed associated with the overall health. Therefore, under this scenario, the models should reject the null hypothesis and conclude that the gene is in fact associated with the outcome variables. For this simulation, we compare the models in terms of their statistical power. To generate data, we randomly sample a_j and b_j from the $N(0, 1)$ distribution and sample s_j from the Uniform(0, 1) distribution. For a sample dataset under Simulation I, Figure 1 (top row) shows the pairwise scatter plots and correlation coefficients (left panel), the scree plot (i.e., the eigenvalues of the principal components; middle panel), and the cumulative percentage of the overall variance explained by the principal components (right panel). As we can see, $Y_1, \ldots, Y_5$ are significantly correlated, and the first principal component is a good representative of the data, which captures around 60% of the overall variation.

For Simulation II, we follow the same process to generate data, but this time we sample s_j from the Uniform(2, 3) distribution. By doing so, we increase the noise in the observed outcome variables, and as a result, weaken the correlations among them. The bottom row in Figure 1 shows the corresponding plots. As we can see, the correlation coefficients are small, and the first principal component only captures about 25% of the overall variation in the data. Nevertheless, as before, we expect the models to reject the null hypothesis and conclude that the gene is significantly associated with the overall health since we kept $\beta = 0.5$. Similar to Simulation I, under this scenario, we compare the three models in terms of their statistical power.

For Simulation III, we use a similar setup as Simulation II, but this time we set $\beta = 0$

Table 1

Definition and summary statistics for the NACC data. The outcome variables include MMSE, ANIMALS, and TMTB, and explanatory variables include disease status (diagnosis), gender (female), age, years of education, and systolic blood pressure (bpsys). For categorical variables, the frequency and percentage (in parenthesis) of each category are presented. For numerical variables, we provide the median and interquartile range (IQR, in parenthesis).

Variable	N = 2,700	Descriptoin
diagnosis		Alzheimer's disease diagnosis
0	1,534 (57%)	Healthy controls (HC)
1	613 (23%)	Mild cognitive impairment (MCI)
2	553 (20%)	Alzheimer's disease (AD)
female		Binary variable indicating whether the subject is female
0	1,151 (43%)	No
1	1,549 (57%)	Yes
age	72 (64, 78)	Age of subject when tests were taken
education	16.0 (13.0, 18.0)	Years of education
bpsys	130 (120, 143)	Systolic blood pressure
mmse	28.0 (26.0, 30.0)	Mini Mental State Examination
animals	18 (13, 22)	Cognitive impairment test - recite animals
trail	90 (62, 140)	Cognitive impairment test - trail letters and numbers

to remove the effect of the gene on the overall health. In this setting, the models should not reject the null hypothesis, which indicates that the gene is not associated with the outcome variables. Failing to do so will contribute to the Type I error. Therefore, under this scenario, we compare the three models in terms of their Type I error rate. Because we have set the significance level at 0.05, we expect the Type I error rate for the models to be close to 0.05.

Figure 2 compares the three alternative models (IR, MR, and IMR) in terms of their statistical power under Simulations I & II, and their Type I error rate under Simulation III. For Simulation I, the left panel shows that our proposed approach has substantially higher power compared to the two alternatives. Additionally, while IR and MR perform similarly for small sample sizes, MR outperforms IR as the sample size increases. For Simulation II (middle panel), the overall power decreases for all three methods due to the high level of noise. Nevertheless, despite the absence of substantial correlation between the outcome variables, IMR continues to exhibit superior performance compared to the other two models. For Simulation III (right panel), where the null hypothesis is true, the observed trend suggests that all three models adhere to their expected Type I error rate of 0.05.

4. Real Data Analysis.

Demographic and Clinical Factors Associated with Cognitive Performance. The number of patients suffering from Alzheimer's disease (AD) is projected to reach 13 million by 2050 [1]. Therefore, there is a great need to 1) understand the underlying factors contributing to this disease, 2) identify it as early as possible, and 3) improve patients' quality of life by reducing and controlling its symptoms. Here, we focus on a battery of cognitive tests commonly used to examine the severity of the disease and aim to identify different demographic and

clinical factors associated with these tests. To achieve this goal, we will use the data obtained from the National Alzheimer's Coordinating Center (NACC) [2, 23], which standardizes data collected from more than 30 Alzheimer's Disease Research Centers (ADRC) across the United States. The data contain information from multiple time points and across different data modalities, including clinical information from the Uniform Data Set (UDS) [16, 4]. Here, we have selected all baseline visits and applied the Crosswalk Study [15] to convert UDS 3.0 neuropsychological testing features to 2.0 features to preserve more subjects.

The final dataset includes 2700 subjects classified into one of three categories: healthy controls (HC), mild cognitive impairment (MCI), and Alzheimer's disease (AD). Each subject was examined using several cognitive tests, including the Mini Mental State Examination (MMSE) test [7], ANIMALS [9], and Trail Making Test Part B (TMTB) [20, 6]. MMSE, which is by far one of the most commonly used screening method at primary care visits, is a brief questionnaire-based test that helps evaluate an individual's cognitive functioning across several areas, including orientation, memory, attention, language, and visuospatial skills. The total score ranges from 0 to 30, with higher scores indicating better cognitive function. The ANIMALS test involves reciting the name of animals in one minute. Similar to MMSE, higher scores are better. For TMTB, participants are given 25 circles on a piece of paper, with numbers (1 to 12) and letters (A through L). They are asked to connect the circles in ascending order as quickly as possible, alternating back and forth from numbers to letters. Unlike MMSE and ANIMALS, where higher scores are better, lower scores are preferred for TMTB. We expect MMSE and ANIMALS to be positively correlated with each other, and both negatively correlated with TMTB. The three test indirectly capture the overall cognitive ability of the participants.

In this paper, our goal is to investigate whether these test scores are significantly associated with the diagnosis, as well as several demographic and clinical variables including age, gender, education, and blood pressure. Table 1 provides a summary of all explanatory and outcome variables along with a brief description.

We start our analysis by fitting three individual regression (IR) models, one for each

Table 2

Individual regression models for the NACC Data with three outcome variables.

MMSE

Characteristic	Beta	95% CI[1]	p-value
diagnosis			
0	—	—	
1	-2.068	-2.357, -1.779	<0.001
2	-7.083	-7.382, -6.785	<0.001
age	-0.019	-0.029, -0.008	<0.001
female			
0	—	—	
1	0.067	-0.166, 0.300	0.6
education	0.236	0.202, 0.270	<0.001
bpsys	-0.006	-0.012, 0.001	0.077

[1] CI = Confidence Interval

ANIMALS

Characteristic	Beta	95% CI[1]	p-value
diagnosis			
0	—	—	
1	-4.356	-4.855, -3.857	<0.001
2	-9.370	-9.886, -8.854	<0.001
age	-0.125	-0.143, -0.107	<0.001
female			
0	—	—	
1	-0.185	-0.587, 0.218	0.4
education	0.375	0.316, 0.433	<0.001
bpsys	-0.003	-0.014, 0.008	0.6

[1] CI = Confidence Interval

TMTB

Characteristic	Beta	95% CI[1]	p-value
diagnosis			
0	—	—	
1	46.013	40.030, 51.996	<0.001
2	89.071	82.886, 95.255	<0.001
age	1.384	1.169, 1.599	<0.001
female			
0	—	—	
1	0.259	-4.566, 5.084	>0.9
education	-4.748	-5.454, -4.043	<0.001
bpsys	0.110	-0.021, 0.241	0.10

[1] CI = Confidence Interval

Table 3
Multivariate Analysis of Variance (MANOVA) of the NACC data using Pillai test statistic [19].

Variables	Df	Test Stat	Approx F	P-value
diagnosis	2	0.581	367.39	<0.001
age	1	0.098	97.72	<0.001
female	1	0.001	0.46	0.711
education	1	0.134	139	<0.001
bpsys	1	0.002	1.96	0.118

outcome variable. The results are presented in Table 2 (which are prepared using the R package `gtsummary` [22]). As we can see, diagnosis, age, and education show strong statistical significance (all p-values < 0.001), whereas the associations with gender and blood pressure are not significant (at the 0.05 level) according to these models.

Using a multivariate regression (MR) model to examine the associations with the three outcome variables simultaneously, our analysis further confirms that diagnosis, age, and education are indeed statistically significant factors (Table 3). As before, this model shows that gender and blood pressure lack statistical significance when examined in relation to the set of outcome variables. These findings shed light on the multivariate relationships between the three cognitive tests and the set of demographic (age, gender, education) and clinical (diagnosis, blood pressure) variables under investigation.

Finally, we apply our proposed IMR method to the NACC data. As we can see in Figure 3 (left panel), the three outcome variables, MMSE, ANIMALS, and TMTB, are highly correlated – MMSE and ANIMALS are positively correlated, with the correlation coefficient equal to

Figure 3. *Correlation matrix (left), scree plot (middle), and cumulative proportion of variance explained by principal components (right) for the three outcome variables from the NACC data. As we can see, the three variables are highly correlated (both positively and negatively) and the first principal component captures almost 2/3 of the overall variance.*

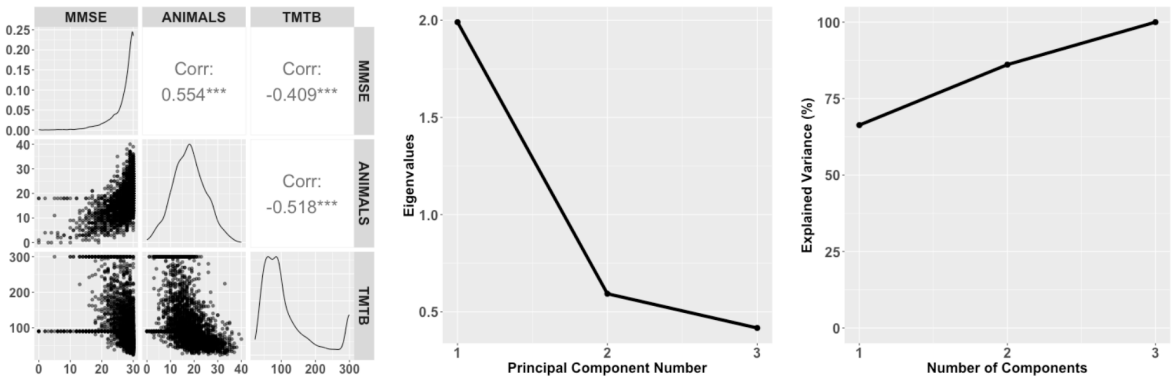


Table 4
Regression analysis results for the NACC data using our proposed IMR method.

Characteristic	Beta	95% CI[1]	p-value
diagnosis			
0	—	—	
1	0.989	0.908, 1.071	<0.001
2	2.412	2.328, 2.496	<0.001
age	0.023	0.020, 0.026	<0.001
female			
0	—	—	
1	0.009	-0.056, 0.075	0.8
education	-0.099	-0.108, -0.089	<0.001
bpsys	0.002	0.000, 0.004	0.049

[1] CI = Confidence Interval

0.554, and they are both negatively correlated with TMTB, with the correlation coefficients equal to -0.409 and -0.518 respectively. The scree plot (middle panel) and the proportion of variance explained (right panel) justify using the first principal component as a singular combined variable in place of the the original three outcome variables. In this case, the first principle component captures almost 2/3 of the overall variance.

The elements of the first eigenvector (i.e., with the highest eigenvalue) are $v_{11} = -0.57$, $v_{12} = -0.61$, $v_{13} = 0.55$. The first principal component uses these elements to create a combined score, henceforth called the Combined Cognitive Score (CCS), as follows:

$$\text{CCS} = -0.57 \times \text{MMSE}^* - 0.61 \times \text{ANIMALS}^* + 0.55 \times \text{TMTB}^*, \tag{4.1}$$

where $\text{MMSE}^*, \text{ANIMALS}^*$, and TMTB^* are the standardized (mean zero and variance 1) versions of the original outcome variables. This way, CCS has the same direction as TMTB and opposite direction as MMSE and ANIMALS. Therefore, for CCS lower scores are preferred.

Using CCS as the outcome variable, we fit a linear regression model with diagnosis, gender, age, education, and blood pressure as the explanatory variables. The results are presented in

Figure 4. *Left panel: The distribution of p-values for the genotype (wild-type and PPARα-deficient) effect on 120 genes potentially involved in nutritional problems, while controlling for the type of diet. Middle panel: The scree plot from applying PCA to the gene expression data. Note that there are only 40 principal components since the sample size is 40 even though there are 120 variables. Right panel: The cumulative proportion of variance explained by principal components for the gene expression data. As we can see, the scree plot levels off after the first three principal components, which together capture about 65% of the overall variance.*

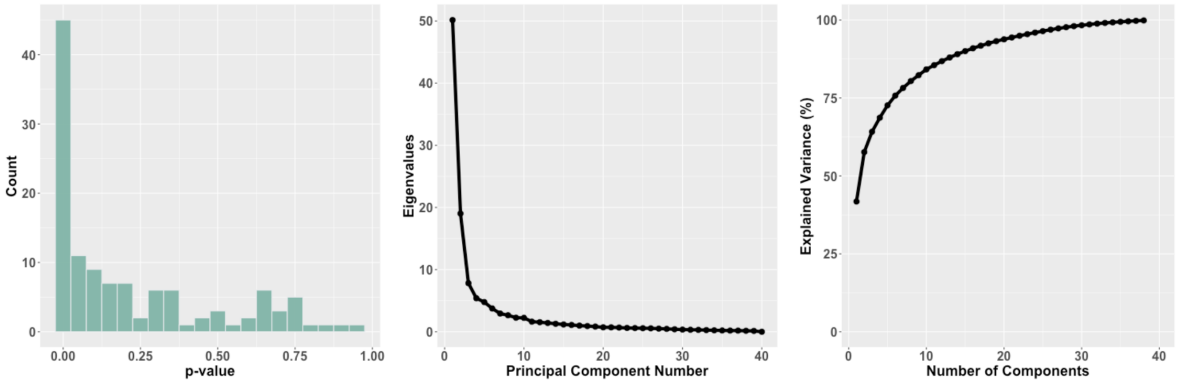


Table 4. While our findings are similar to those provided by IR and MR, our method shows that blood pressure is marginally significant (p-value = 0.049). Our model indicates that compared to the healthy controls, on average CCS increases by 0.989 for patients with MCI and by 2.412 for patients with AD. As age increases by one year, the average CCS increases by 0.023. Education on the other hand improves the score: for each extra year of education, CCS decreases (improves) by -0.099. Finally, for each 10 units increase in systolic blood pressure, CCS increases (worsens) by 0.02 on average. This finding is consistent with some recent studies [21, 17] that have shown there is a slight correlation between the cognitive status of people and their blood pressure level. More specifically, these studies confirm an underlying association between hypertension and poor performances in executive function and attention tests, as well as possible dementia later in life. Our IMR approach confirms these findings. As illustrated by simulation studies earlier, we believe this is due to our model's ability to identify weak signals because of its improved power.

Effect of Diet and Genotype on Lipid Metabolism. Next, we re-examine the study conducted by Martin et al. [14] that highlights the importance of proliferator-activated receptor-α (PPARα) in regulating lipid and xenobiotic metabolism and provides novel insights into its regulatory mechanisms through a nutrigenomic approach. The study examines the effects of different fatty acids (FAs) compositions on liver lipids and gene expression in wild-type and PPARα-deficient mice under low-fat intake conditions. More specifically, the study includes 20 wild-type and 20 PPARα-deficient mice under five different low fat diets. For each mouse, they measured the expression level of 120 genes related to class II nuclear receptor (NR) signaling. The study suggests that even under conditions of low-fat intake, dietary FAs can reduce hepatic steatosis. Further, it highlights the role of PPARα in regulating hepatic FA content and composition.

Using individual regression (IR) models and controlling for the type of diet, we find that

the genotype (i.e, wild-type vs. PPARα-deficient) is significantly associated (p-value < 0.05) with 52 out of 120 genes. The distribution of p-values are presented in Figure 4 (left panel). Before making any conclusion, we need to account for the fact that multiple (120) hypotheses are tested simultaneously. However, different methods lead to substantially different results. Using the Holm-Bonferroni [10] method, only 26 genes remain significant; whereas, using the False Discovery Rate (FDR) approach [3] to adjust the p-values leads to 38 significant genes at the 0.05 level.

As discussed above, alternatively we could evaluate the association of the 120 gene expression values simultaneously with the genotype using a multivariate regression (MR) model to account for their underling correlation structure. In this case, however, we only have 40 subjects for the 120 outcome variables. Therefore, the standard multivariate regression model (like the one we used earlier) is infeasible for this problem.

We could use our proposed integrated multivariate regression (IMR) model. We start by finding the principal components. The scree plot (Figure 4, middle panel) and the cumulative proportion of variance explained by the principal components (Figure 4, right panel) indicate that using only the first principal component is not sufficient. Therefore, we use the first three principal components instead, since that is the point the scree plot levels off, and we can capture 65% of overall variance. Table 5 provides the list of top five genes with the highest absolute weights for the first three principal components (PC1, PC2, and PC3).

Using the top three principal components leads to a slightly more complex model compared to the ones we used earlier. However, by using the first three principal components as the derived outcome variables (instead of the original 120 gene expressions), we are able to fit a multivariate regression model with substantially smaller number of parameters. Our results (Table 6) shows that the overall gene expression values (captured by their first three principal components) are significantly associated with both diet and genotype (p-values < 0.001). Our results confirm the finding of Martin et al. [14].

Table 5

Top five genes with the highest absolute weights for each of the first three principal components.

Nutrimouse Data	Principal Components		
	PC1	PC2	PC3
Top Genes	c.fos	PMDCI	S14
	CYP26	THIOL	ACC2
	PPARg	L.FABP	HMGCoAred
	MDR1	ACBP	cHMGCoAS
	FXR	CYP3A11	CYP4A14

Table 6

Evaluating the effect of diet and genotype on overall gene expression using IMR with three principal components.

Variables	Df	Test Stat	Approx F	P-value
diet	4	0.823	3.22	<0.001
genotype	1	0.916	115.86	<0.001

5. Discussion. In this paper, we have proposed a simple, yet powerful model for regression problems with multiple outcome variables. Our simulation results show that this approach outperforms some commonly-used alternative models, including individual univariate regression, which ignore the correlation structure in the outcome variables, and multivariate regression, which require estimating a large number of parameters. We have also applied our method to two scientific problems involving multiple cognitive tests and genomics data.

One of the main drawbacks of our method is that by reducing the dimensionality of the outcome variables, it could lose substantial amount of information, especially if the outcome variables are not highly correlated. To avoid this issue, as we discussed above, one could use multiple principal components in the analysis. This, however, requires finding the right trade-off between reducing the number of parameters and information loss. This can be the focus of future research.

When finding the principal components (and eigenvectors), it is important to recognize that what matters is the line along which the variance is maximized. The direction along that line is arbitrary and can change depending on how the eigenvectors are computed. Therefore, The utility of the principal components (and eigenvectors) remains unchanged if we reverse their direction by multiplying them by -1. As an example, consider the definition of the Combined Cognitive Score (CCS) for Alzheimer's disease diagnosis in Section 4 based on the first principal component. In this case, CCS is calculated by multiplying MMSE and ANIMALS by negative numbers (-0.57 and -0.61 respectively), and multiplying TMTB by a positive number (0.55). This way, PCA aligns MMSE and ANIMALS with TMTB, so the overall score, CCS, is also aligned with TMTB (i.e., lower scores are better). The overall results would remain unchanged if we reverse the signs of the three multipliers, resulting in an overall score aligned with MMSE and ANIMALS, where higher scores are indicative of better cognitive performance. While this changes the interpretation of the score, our results and findings remain the same.

For the IR model discussed in this paper, one needs to account for multiple hypothesis testing, since the error rate of individual tests no longer accurately represents the overall error rate. While this is outside of the scope of this current paper, it is common practice to address this issue by employing methodologies that adjust the p-values. Popular methods include: Bonferroni [5] Holm-Bonferroni [10], and False Discovery Rate (FDR) [3]. Future research could involve comparing our proposed approach to various methods designed for taking multiple hypothesis testing into account.

Acknowledgments. I would like to express my gratitude to my advisor and mentor, Professor Qu, for her invaluable support and guidance in refining my idea and patiently guiding me throughout this project. Additionally, I wish to express my thanks to Dr. Sam Behseta and Dr. Jessica Jaynes for teaching me statistical analysis and R programming. I am also grateful to Dr. Steven Cramer, who provided me with the opportunity to work on his research project, which led to the idea presented in this paper. Finally, I would like to thank the National Alzheimer's Coordinating Center (NACC) for generously sharing their data with me, and Yueqi Ren for introducing me to the NACC dataset and providing guidance on how to prepare it for analysis. The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

REFERENCES

[1] *2023 alzheimer's disease facts and figures*, Alzheimer's & Dementia, 19 (2023), pp. 1598–1695, https://doi.org/https://doi.org/10.1002/alz.13016.

[2] D. Beekly, E. Ramos, W. Lee, W. Deitrich, M. Jacka, J. Wu, J. Hubbard, T. Koepsell, J. Morris, W. Kukull, E. Reiman, N. Kowall, G. Landreth, M. Shelanski, K. Welsh-Bohmer, A. Levey, H. Potter, B. Ghetti, D. Price, B. Hyman, R. Petersen, M. Sano, S. Ferris, M. Mesulam, J. Kaye, D. Bennett, J. Yesavage, D. Marson, C. Beck, C. DeCarli, C. Cotman, J. Cummings, L. Thal, W. Markesbery, S. Gilman, J. Trojanowski, S. DeKosky, H. Chui, R. Rosenberg, and M. Raskind, *The national alzheimer's coordinating center (nacc) database: The uniform data set*, Alzheimer Disease and Associated Disorders, 21 (2007), pp. 249–258, https://doi.org/10.1097/WAD.0b013e318142774e.

[3] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society: Series B (Methodological), 57 (1995), pp. 289–300.

[4] L. Besser, W. Kukull, D. S. Knopman, H. Chui, D. Galasko, S. Weintraub, G. Jicha, C. Carlsson, J. Burns, J. Quinn, R. A. Sweet, K. Rascovsky, M. Teylan, D. Beekly, G. Thomas,

M. Bollenbeck, S. Monsell, C. Mock, X. H. Zhou, N. Thomas, E. Robichaud, M. Dean, J. Hubbard, M. Jacka, K. Schwabe-Fry, J. Wu, C. Phelps, and J. C. Morris, *Version 3 of the national alzheimer's coordinating center's uniform data set*, Alzheimer Disease & Associated Disorders, 32 (2018), pp. 351–358.

[5] C. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilita*, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8 (1936), pp. 3–62.

[6] S. Correia, D. C. Ahern, A. R. Rabinowitz, T. J. Farrer, A. K. Smith Watts, S. Salloway, P. F. Malloy, and S. C. Deoni, *Lowering the Floor on Trail Making Test Part B: Psychometric Evidence for a New Scoring Metric*, Archives of Clinical Neuropsychology, 30 (2015), pp. 643–656, https://doi.org/10.1093/arclin/acv040.

[7] S. Creavin, S. Wisniewski, A. Noel-Storr, C. Trevelyan, T. Hampton, D. Rayment, V. Thom, K. Nash, H. Elhamoui, R. Milligan, A. Patel, D. Tsivos, T. Wing, E. Phillips, S. Kellman, H. Shackleton, G. Singleton, B. Neale, M. Watton, and S. Cullum, *Mini-mental state examination (mmse) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations*, Cochrane Database of Systematic Reviews, (2016), https://doi.org/10.1002/14651858.CD011145.pub2, https://doi.org//10.1002/14651858.CD011145.pub2.

[8] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, Sage, Thousand Oaks CA, third ed., 2019, https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

[9] H. Hanyu, K. Kazumasa Kume, Y. Takada, T. Onuma, and T. Iwamoto, *The 1-minute mental status examination in the memory clinic*, Journal of the American Geriatrics Society, 57 (2009), pp. 1130–1131, https://doi.org/https://doi.org/10.1111/j.1532-5415.2009.02287.x, https://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2009.02287.x, https://arxiv.org/abs/https://agsjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1532-5415.2009.02287.x.

[10] S. Holm, *A simple sequentially rejective multiple test procedure*, Scandinavian Journal of Statistics, 6 (1979), pp. 65–70.

[11] H. Hotelling, *Analysis of a complex of statistical variables into principal components.*, Journal of Educational Psychology, 24 (1933), pp. 498–520.

[12] H. Hotelling, *Relation between two sets of variables*, Biometrika, 28 (1936), pp. 321–377, https://doi.org/10.1093/biomet/28.3-4.321.

[13] R. Johnson and D. Wichern, *Applied multivariate statistical analysis*, Prentice Hall, Upper Saddle River, NJ, 6. ed ed., 2002.

[14] P. G. P. Martin, H. Guillou, F. Lasserre, S. DÃ©jean, A. Lan, J.-M. Pascussi, M. SanCristobal, P. Legrand, P. Besse, and T. Pineau, *Novel aspects of pparÎ±-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study*, Hepatology, 45 (2007), pp. 767–777, https://doi.org/https://doi.org/10.1002/hep.21510.

[15] S. E. Monsell, H. H. Dodge, X.-H. Zhou, Y. Bu, L. M. Besser, C. Mock, S. E. Hawes, W. A. Kukull, and S. Weintraub, *Results from the NACC uniform data set neuropsychological battery crosswalk study*, Alzheimer Disease & Associated Disorders, 30 (2016), pp. 134–139, https://doi.org/10.1097/wad.0000000000000111, https://doi.org/10.1097/wad.0000000000000111.

[16] J. C. Morris, S. Weintraub, H. C. Chui, J. Cummings, C. DeCarli, S. Ferris, N. L. Foster, D. Galasko, N. Graff-Radford, E. R. Peskind, D. Beekly, E. M. Ramos, and W. A. Kukull, *The uniform data set (UDS): Clinical and cognitive variables and descriptive data from alzheimer disease centers*, Alzheimer Disease & Associated Disorders, 20 (2006), pp. 210–216, https://doi.org/10.1097/01.wad.0000213865.09806.92, https://doi.org/10.1097/01.wad.0000213865.09806.92.

[17] V. Novak, *The relationship between blood pressure and cognitive function*, Nature Reviews Cardiology, 7 (2010), https://doi.org/10.1038/nrcardio.2010.161.

[18] K. Pearson, *On lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2 (1901), pp. 559–572, https://doi.org/10.1080/14786440109462720.

[19] K. C. S. Pillai, *Some New Test Criteria in Multivariate Analysis*, The Annals of Mathematical Statistics, 26 (1955), pp. 117 – 121, https://doi.org/10.1214/aoms/1177728599, https://doi.org/10.1214/aoms/1177728599.

[20] R. M. Reitan, *Investigation of the validity of halstead's measures of biological intelligence*, AMA Archives of Neurology & Psychiatry, 73 (1955), pp. 28–35.

[21] W. S. Shaw, *Accelerated risk of hypertensive blood pressure recordings among alzheimer caregivers*, ScienceDirect, 46 (1999), pp. 215–227, https://doi.org/10.1038/nrcardio.2010.161.

[22] D. D. Sjoberg, K. Whiting, M. Curry, J. A. Lavery, and J. Larmarange, *Reproducible summary tables with the gtsummary package*, The R Journal, 13 (2021), pp. 570–580, https://doi.org/10.32614/RJ-2021-053, https://doi.org/10.32614/RJ-2021-053.

[23] S. Weintraub, D. P. Salmon, N. Mercaldo, S. Ferris, N. R. Graff-Radford, H. Chui, J. L. Cummings, Charles, Decarli, N. L. Foster, D. R. Galasko, E. R. Peskind, W. Dietrich, D. L. Beekly, W. A. Kukull, C. John, and Morris, *The alzheimer's disease centers' uniform data set (UDS): The neuropsychological test battery*, Alzheimer Disease and Associated Disorders, 23 (2009), pp. 91–101.