Journal of Statistical Theory and Applications
https://doi.org/10.1007/s44199-024-00074-y

RESEARCH ARTICLE

Deriving the Distribution and Exploring the Utility of Partial R^2 in the Era of Big Data

Gregory S. Hawk[1] · **Katherine L. Thompson**[1]

Received: 25 October 2023 / Accepted: 31 March 2024

Abstract
A central goal in the world of statistics and data science is the construction of linear regression models for continuous variables of interest. Often, our objective is to examine the impact of one or more explanatory variables, after adjusting for demographic covariates or other known/relevant factor(s). While the traditional approach is to use hypothesis testing to determine statistical significance, the p-values obtained are heavily dependent on sample size. This is particularly problematic for large datasets or "overpowered" studies, where even the tiniest of effects will appear to be highly significant. Computing capabilities and cloud-enhanced data sharing have revolutionized the way we use data worldwide, from healthcare and investments to manufacturing and retail. While machine learning and artificial intelligence are improving predictive analytics, we need better statistical inference to help understand and translate our models into meaningful and actionable insights. The coefficient of partial determination (or **partial R^2**) is widely used in applied science to supplement hypothesis testing, but little work has been done to understand its statistical properties. In this work, we derive the complete distribution of partial R^2 and perform simulated and real-world data analyses to show the advantages of adding it to your next analysis of Big Data.

Keywords Partial R^2 · Big data · Linear regression · Coefficient of partial determination · R^2

1 Introduction

One of the most commonly encountered and fundamentally important tasks facing a data analyst is to evaluate and quantify the relationship between a continuous (quantitative) outcome variable and a set of explanatory variables. In many cases, this task is

✉ Gregory S. Hawk
greg.hawk@uky.edu

[1] Dr. Bing Zhang Department of Statistics, University of Kentucky, Lexington, Kentucky, USA

Published online: 13 May 2024

accomplished using linear regression models, in which this set can be partitioned into two groups: primary explanatory variables and secondary covariates. Primary explanatory variables are of significant interest to the investigators and their potential impact on the designated outcome variable comprises the main focus of the analysis. Often, however, relationships also exist between our outcome variable and non-modifiable demographic variables, like sex, age, and race. While we want to adjust our regression analysis to account for these relationships and model our data more accurately, such variables may not be of primary interest to investigators looking for modifiable, actionable relationships. We may also want to adjust our model for known or relevant factors identified in previous studies, which are not of primary interest to our current study but should be taken into account nevertheless. In each of these cases, we have secondary covariates to include in the model.

The traditional methodology for evaluating and reporting results from such regression models usually centers around hypothesis testing. Individual t-tests or partial F-tests can isolate the effects of one or more primary explanatory variables at a time, after accounting for all other variables in the model. The corresponding p-values can then be calculated and compared to some arbitrary significance level, often defaulting to $\alpha = 0.05$. We know, however, that p-values obtained from these methods are heavily dependent on sample size [1]. While p-values mathematically do not represent the magnitude or strength of a relationship, they are often interpreted as doing so in practice [2]. For under-powered or pilot datasets, statistically significant results are extremely difficult to achieve, even when a true relationship exists. For over-powered or large datasets, on the other hand, even the smallest effects will appear to be highly significant despite not being clinically or scientifically meaningful [1, 3–7].

Though Fisher himself objected to this automatized "accept/reject" process as being philosophically contrary to the principles of sound science, it has nonetheless remained an entrenched practice across the scientific community since its inception [8]. It has been more than a decade since van der Laan and Rose [9] sounded the alarm and issued a challenge to the statistical community with regard to how it analyzes so-called "Big Data". In 2016, the American Statistical Association (ASA) released a statement [2] outlining six principles for improving the use of p-values to show statistical significance, including the assertion that "by itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis". Three years later, the *American Statistician* devoted an entire special edition [10] discussing the limitations of p-values and proposing a myriad of alternative methods. Even journals outside of the field of statistics are becoming more and more aware of the issue, though often proposing less-than-satisfactory solutions to the dilemma, such as simply lowering the p-value threshold to 0.01 or 0.005 [3, 4, 6, 8].

The stark reality, however, is that the overwhelming majority of statistical analyses in 2024, even for Big Data, are still driven by hypothesis testing. Journals are often reluctant (and, at times, completely unwilling) to publish study results without accompanying p-values, though there are encouraging signs that suggest this practice is becoming less frequent [8]. One improvement that is gaining traction in the research community is the reporting of effect sizes along with p-values, to add a dimension of relationship strength. However, many different ways to calculate an effect size for a

given research question exist, many of which are also heavily influenced by sample size.

To address these limitations, this paper focuses on an under-utilized family of alternatives to traditional hypothesis methods known as coefficients of determination. In Section 2, we provide notation for a linear model framework and we define two members of this family --- R^2 and partial R^2 --- which will serve as the central foci of the paper. Following this, in Section 3, we derive the complete distribution of partial R^2. After detailing results from a simulation study and real-world Big Data analysis, we provide our conclusions and future directions in Section 6.

2 Background and Notation

Consider a linear regression model with p explanatory variables, each with n observations, defined by

$$\vec{\boldsymbol{Y}} = \beta_0 \vec{\mathbf{1}}_{\boldsymbol{n}} + \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}, \tag{1}$$

where $\vec{\boldsymbol{\beta}}$ is the p-dimensional vector of regression coefficients, $\boldsymbol{X}$ is the $n \times p$ design matrix of explanatory variables, and $\vec{\boldsymbol{\varepsilon}} \sim N(\vec{\mathbf{0}}, \sigma^2 \boldsymbol{I_n})$ is the n-dimensional vector of independent error terms.

The most well-known member of the family of coefficients of determination is the **coefficient of multiple determination**, denoted by R^2, which describes the overall strength of a linear regression model. We can interpret the value of R^2 for a given model as the estimated proportion of variability in the response variable that can be collectively explained by the explanatory variables in the model. Mathematically,

$$R^2 = \frac{SSR}{SSTO},$$

where SSR and $SSTO$ are the regression sum of squares and total sum of squares, respectively, for the ordinary least squares (OLS) estimate of model (1) [11]. Because of its construction as a ratio, the usefulness and interpretability of R^2 aren't affected by sample size in the same way that F-statistics and p-values are. Recent advances in methodology have extended its utility as a performance criterion in applications such as machine learning and cluster analysis [12, 13]. Koerts and Abrahamse [14] showed that the coefficient of multiple determination is a consistent estimator for the analogously-interpreted population parameter ϕ, as defined by Barten [15]. Cramer [16] built on this work to show that R^2 follows a non-central beta distribution.

Suppose we can partition the p explanatory variables in the model from (1) into the two subsets described in Section 1. That is, suppose there are q primary variables of interest (call this Subset A, represented by regression coefficient vector $\vec{\boldsymbol{\beta}}_{\boldsymbol{A}}$ and design matrix $\boldsymbol{X_A}$), while the other $p-q$ variables are covariates for which we want to adjust (Subset B, represented by $\vec{\boldsymbol{\beta}}_{\boldsymbol{B}}$ and $\boldsymbol{X_B}$). Then, grouping our explanatory variables by subset, we can rewrite the "full" model from (1) as

$$\vec{\boldsymbol{Y}} = \beta_0 \vec{\mathbf{1}}_{\boldsymbol{n}} + \boldsymbol{X_A}\vec{\boldsymbol{\beta}}_{\boldsymbol{A}} + \boldsymbol{X_B}\vec{\boldsymbol{\beta}}_{\boldsymbol{B}} + \vec{\boldsymbol{\varepsilon}} \tag{2}$$

To study the collective usefulness of the q primary variables of interest in modeling our outcome of interest, we could test the null hypothesis of $\vec{\boldsymbol{\beta}}_{\boldsymbol{A}} = \vec{\mathbf{0}}$. Under this null hypothesis, we can write a "reduced" model as

$$\vec{\boldsymbol{Y}} = \beta_0 \vec{\mathbf{1}}_{\boldsymbol{n}} + \boldsymbol{X_B} \vec{\boldsymbol{\beta}}_{\boldsymbol{B}} + \vec{\boldsymbol{\varepsilon}} \tag{3}$$

Then, we can define the **partial coefficient of determination** for the q primary variables in Subset A, given that the $(p - q)$ covariates in Subset B are already included in the model, as

$$R^2_{Y A|B} = \frac{SSE_{Reduced} - SSE_{Full}}{SSE_{Reduced}}, \tag{4}$$

where the subscripts denote the sum of squares corresponding to either the full or reduced models from (2) and (3). This quantity is commonly referred to as **partial $\boldsymbol{R^2}$** and has an analogous interpretation to that of the coefficient of multiple determination (R^2). Partial R^2 estimates the proportion of remaining variability in the response variable that can be explained collectively by the q primary explanatory variables, after adjusting for the $p - q$ covariates. Unlike R^2, however, the distribution and mathematical properties of partial R^2 have not previously been studied, despite its somewhat frequent use in practice and its inclusion in the default analysis output for many statistical software packages and procedures. The following section begins to fill that gap, providing a derivation of the distribution of partial R^2.

3 Distribution of Partial R^2

Rewriting the denominator of (4) and multiplying by a form of one, we have

$$R^2_{Y A|B} = \frac{\frac{1}{\sigma^2}\left[SSE_{Reduced} - SSE_{Full}\right]}{\frac{1}{\sigma^2}\left[SSE_{Reduced} - SSE_{Full}\right] + \frac{1}{\sigma^2} \cdot SSE_{Full}} \tag{5}$$

We can see that $R^2_{Y A|B}$ can thus be written in the form $\frac{U}{U+V}$. A straightforward two-variable transformation can be used to verify the following property (i.e. from Johnson and Kotz [17]).

Lemma 1 *Suppose random variables $U \sim \chi^2_u$ with non-centrality parameter λ and $V \sim \chi^2_v$, such that $U \perp V$. Then, the quantity $W = \frac{U}{U+V} \sim Beta\left(\frac{u}{2}, \frac{v}{2}\right)$ with non-centrality parameter λ.*

From linear model theory, $V = \frac{1}{\sigma^2} \cdot SSE_{Full} \sim \chi^2_{n-p-1}$ [18]. We next seek the distribution of the quantity $U = \frac{1}{\sigma^2}[SSE_{Reduced} - SSE_{Full}]$. Using the matrix representation for each sum of squares quantity [11], we can write

$$\begin{aligned} U &= \frac{1}{\sigma^2}[SSE_{Reduced} - SSE_{Full}] \\ &= \frac{1}{\sigma^2}\left[\vec{\boldsymbol{Y}}^T(\boldsymbol{I_n} - \boldsymbol{H}_{reduced})\vec{\boldsymbol{Y}} - \vec{\boldsymbol{Y}}^T\left(\boldsymbol{I_n} - \boldsymbol{H}_{full}\right)\vec{\boldsymbol{Y}}\right] \\ &= \frac{1}{\sigma^2}\left[\vec{\boldsymbol{Y}}^T\left(\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced}\right)\vec{\boldsymbol{Y}}\right], \end{aligned}$$

where $\boldsymbol{H}$ represents the hat or projection matrix for the indicated model and $\boldsymbol{I_n}$ represents the $n \times n$ identity matrix. Consider the following result (i.e. from Ravishanker and Dey [18]):

Lemma 2 *Let $\vec{\boldsymbol{Y}} \sim MVN_n\left(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma}$ has full rank n. Then, the quadratic form $D = \vec{\boldsymbol{Y}}^T\boldsymbol{A}\vec{\boldsymbol{Y}} \sim \chi^2_r$ with non-centrality parameter $\lambda = \vec{\boldsymbol{\mu}}^T\boldsymbol{A}\vec{\boldsymbol{\mu}}$ if and only if $\boldsymbol{A\Sigma}$ is an idempotent matrix of rank r.*

In our case, we have $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I_n}$ (which clearly has full rank n for $\sigma^2 > 0$), and $\boldsymbol{A} = \frac{1}{\sigma^2}\left(\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced}\right)$. Noting that $\boldsymbol{A\Sigma} = \boldsymbol{H}_{full} - \boldsymbol{H}_{reduced}$ is itself a projection matrix, we know that it is idempotent. Applying properties of rank and trace for idempotent matrices, we have

$$\begin{aligned} \text{rank}\,(\boldsymbol{A\Sigma}) &= \text{trace}\,(\boldsymbol{A\Sigma}) \\ &= \text{trace}\left(\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced}\right) \\ &= \text{trace}(\boldsymbol{H}_{full}) - \text{trace}(\boldsymbol{H}_{reduced}) \\ &= p - (p - q) \\ &= q \end{aligned}$$

So, applying Lemma 2, we have

$$U = \frac{1}{\sigma^2}[SSE_{Reduced} - SSE_{Full}] \sim \chi^2_q,$$

with non-centrality parameter $\lambda = \frac{1}{\sigma^2}\left[\vec{\boldsymbol{\beta}}^T\boldsymbol{X}^T\left(\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced}\right)\boldsymbol{X}\vec{\boldsymbol{\beta}}\right]$.

Finally, we need to establish the independence of U and V, which we can do using Craig's Theorem [18].

Lemma 3 (**Craig's Theorem**). *Let $\vec{\boldsymbol{Y}} \sim MVN_n\left(\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma}$ is positive definite. Then, the quadratic forms $\vec{\boldsymbol{Y}}^T\boldsymbol{A}\vec{\boldsymbol{Y}}$ and $\vec{\boldsymbol{Y}}^T\boldsymbol{B}\vec{\boldsymbol{Y}}$ are independently distributed if and only if $\boldsymbol{A\Sigma B} = \boldsymbol{0}$.*

In our case, we have

$$\begin{aligned}\boldsymbol{A\Sigma B} &= \left(\frac{\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced}}{\sigma^2}\right)\left(\sigma^2 \boldsymbol{I_n}\right)\left(\frac{\boldsymbol{I_n} - \boldsymbol{H}_{full}}{\sigma^2}\right) \\ &= \frac{1}{\sigma^2}\left(\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced}\right)\left(\boldsymbol{I_n} - \boldsymbol{H}_{full}\right) \\ &= \frac{1}{\sigma^2}\left(\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced} - \boldsymbol{H}_{full}^2 + \boldsymbol{H}_{reduced}\boldsymbol{H}_{full}\right)\end{aligned}$$

Since hat matrices are idempotent and the product of nested hat matrices is simply the hat matrix from the reduced model, we have

$$\begin{aligned}\boldsymbol{A\Sigma B} &= \frac{1}{\sigma^2}\left(\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced} - \boldsymbol{H}_{full} + \boldsymbol{H}_{reduced}\right) \\ &= \frac{1}{\sigma^2}\left(\boldsymbol{0}\right) \\ &= \boldsymbol{0},\end{aligned}$$

so we have that U and V are independent.

Applying Lemma 1 to (5), we have our final result:

Theorem 4 (**Distribution of Partial R^2**).

$$R^2_{YA|B} \sim Beta\left(\frac{q}{2}, \frac{n-p-1}{2}\right),$$

with non-centrality parameter $\lambda = \frac{1}{\sigma^2}\left[\vec{\boldsymbol{\beta}}^T \boldsymbol{X}^T \left(\boldsymbol{H}_{full} - \boldsymbol{H}_{reduced}\right) \boldsymbol{X}\vec{\boldsymbol{\beta}}\right]$.

4 Simulation Study

4.1 Design

To confirm the work above, we simulated datasets of size $n = 100$ for a variety of parameter settings. For each combination of settings, a fixed predictor matrix $\boldsymbol{X}$ was used, containing three primary variables of interest ($q = 3$) and two adjustment variables ($p - q = 2$). The five explanatory variables included a mix of binary and continuous variables in each subset and were simulated independently, using the distributions shown in Table 1.

Regression coefficients ($\beta_{A1}, \ldots, \beta_{B2}$) were systematically varied from 0 to 10 to produce a wide variety of effect sizes for simulation testing. For each unique combination of regression coefficients, $B = 1{,}000{,}000$ response vectors ($\vec{\boldsymbol{Y}}$) were simulated

Table 1 Distributions for the Five Simulated Explanatory Variables

Primary Variables	Adjustment Variables
$X_{A1} \sim Bernoulli(0.6)$	$X_{B1} \sim Uniform(-1, 1)$
$X_{A2} \sim Uniform(-1, 1)$	$X_{B2} \sim Bernoulli(0.8)$
$X_{A3} \sim Uniform(-1, 1)$	

as the sum of $\boldsymbol{X_A}\vec{\boldsymbol{\beta}}_{\boldsymbol{A}}$, $\boldsymbol{X_B}\vec{\boldsymbol{\beta}}_{\boldsymbol{B}}$, and independent random errors following a standard normal distribution.

4.2 Results

After fitting full and reduced models for each of the response vectors generated above, the value of partial R^2 was calculated for each of the B models.

For the null case of $\vec{\boldsymbol{\beta}}_{\boldsymbol{A}} = \vec{\mathbf{0}}$, the non-centrality parameter from our proposed Beta distribution reduces to zero. Thus, in this case, the values of $\vec{\boldsymbol{\beta}}_{\boldsymbol{B}}$ shouldn't affect the distribution of partial R^2. We can indeed see this from Fig. 1, which shows histograms of the B values of partial R^2 for two selected coefficient combinations ($\vec{\boldsymbol{\beta}}_{\boldsymbol{B}} = \vec{\mathbf{0}}$ and $\vec{\boldsymbol{\beta}}_{\boldsymbol{B}} = \vec{\mathbf{1}}$), with the corresponding density curves overlaid from our proposed distribution given by Theorem 4 above. We can see that the histograms appear to be virtually identical, as expected, and the overlaid distribution curves provide an excellent fit.

For the alternative case of $\vec{\boldsymbol{\beta}}_{\boldsymbol{A}} \neq \vec{\mathbf{0}}$, the non-centrality parameter from our proposed Beta distribution depends on the projection matrices from the full and reduced models. In this case, the values of all regression parameters have a marked effect on the distribution of partial R^2, as we can see in Fig. 2. For two selected coefficient combinations ($\vec{\boldsymbol{\beta}}_{\boldsymbol{A}} = \vec{\mathbf{1}}$ and $\vec{\boldsymbol{\beta}}_{\boldsymbol{A}} = 10 \cdot \vec{\mathbf{1}}$; $\vec{\boldsymbol{\beta}}_{\boldsymbol{B}} = \vec{\mathbf{1}}$), histograms of the corresponding 1,000,000 values of partial R^2 are shown, with the density curves for our proposed distribution overlaid.

As expected, larger effect sizes for the primary variables of interest result in much larger values of partial R^2 and reduced variability in the distribution. Additionally,

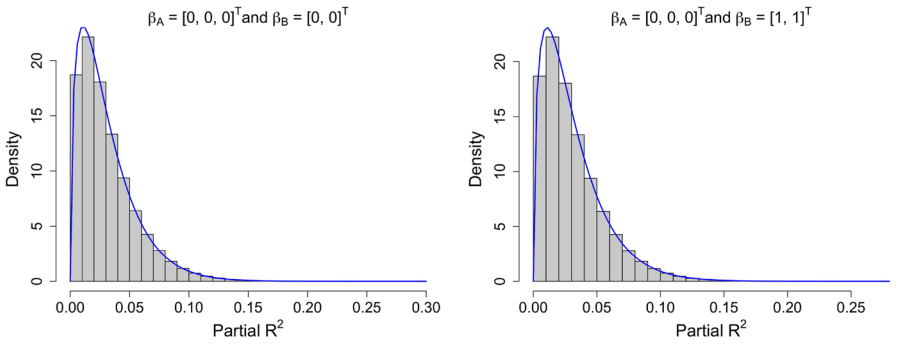

Fig. 1 For two different combinations of parameter settings when $\vec{\boldsymbol{\beta}}_{\boldsymbol{A}} = \vec{\mathbf{0}}$, histograms of the $B = 1{,}000{,}000$ values of partial R^2 are shown, with the density curve from Theorem 4 overlaid in blue

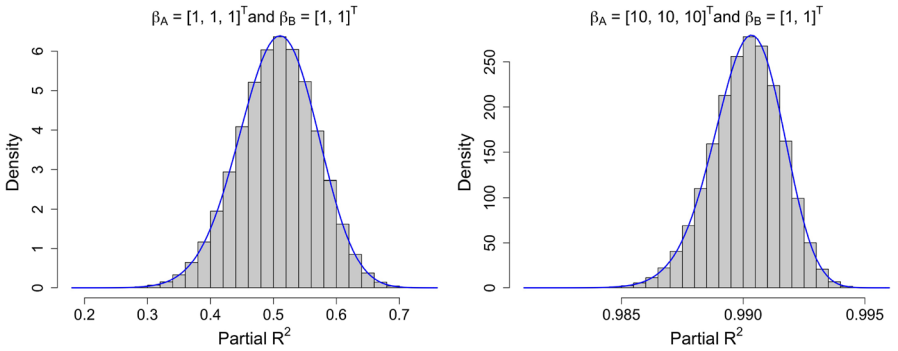

Fig. 2 For two different combinations of parameter settings when $\vec{\boldsymbol{\beta}}_A \neq \vec{\mathbf{0}}$, histograms of the $B = 1{,}000{,}000$ values of partial R^2 are shown, with the density curve from Theorem 4 overlaid in blue

the overlaid distribution curves once again provide an excellent fit, lending additional credence to our proposed distribution in Theorem 4.

5 Application

For decades, it has been widely known in the medical community that increased levels of high-density lipoprotein-cholesterol (HDL-C) are associated with decreased risks of cardiovascular disease, from atherosclerosis to coronary artery disease [19]. According to the Centers for Disease Control and Prevention (CDC) and the American Heart Association (AHA), cardiovascular disease continues to be the leading cause of death in the United States, resulting in over 650,000 deaths per year and costing over $200 billion in healthcare utilization and lost productivity. Coronary artery disease alone is responsible for more than half of these deaths and is present in nearly 7% of adults age 20 and older [20–23].

Every two years since 1999, the National Center for Health Statistics (NCHS) has published an extensive set of data collectively known as the National Health and Nutrition Examination Survey (NHANES), available at https://www.cdc.gov/nchs/nhanes/index.htm. Combining laboratory results, medical examinations, and interviews from approximately 5,000 American children and adults each year, the ongoing goal of NHANES is to provide a cross-sectional snapshot of the health and nutritional state of the country. To increase reliability, the survey over-samples from several minority racial groups and from those aged 60 and older, demographic features we should account for in the covariate portion of our regression models.

For this analysis, we combined twenty years of NHANES data from 1999 through 2018, all of which is publicly available on the NCHS website (see the link above). A total of $n = 101{,}316$ subjects were available for inclusion in the analysis. As with most surveys, however, there was a sizeable amount of missingness present and changes were frequent in the survey variables selected by the NCHS for inclusion over time. Thus, for modeling purposes, potential explanatory variables were only considered if they were collected during all twenty years. Since the primary purpose

Table 2 For each covariate in the reduced HDL-C model, individual Type III test statistics and their corresponding p-values are given

Variable	F-statistic	p-value
Age	43.2	5.1×10^{-11}
Race	122.0	$< 2.2 \times 10^{-16}$
Gender	2597.5	$< 2.2 \times 10^{-16}$
Education Level	47.3	$< 2.2 \times 10^{-16}$
Year	30.4	$< 2.2 \times 10^{-16}$

of this analysis is not to suggest new scientific relationships nor to illustrate model selection techniques, two ordinary linear models were fit for log-transformed HDL-C levels, somewhat naïvely assuming independence between subjects but adjusting for collection year. All analyses were completed in R, version 3.6.3 (R Foundation for Statistical Computing; Vienna, Austria).

5.1 Reduced Model

As is often the case with healthcare modeling, we want to adjust our models for age, race, gender, and level of education. Each of these demographic variables is potentially important to the model as a whole, but none of them are modifiable risk factors for further scientific study or patient intervention. As stated above, we also want to adjust for the year of data collection to account for potential changes in the population over time. Thus, we will have $p - q = 5$ covariates in our reduced model.

Fitting the reduced model and performing individual Type III F-tests for each adjustment variable leads to the results given in Table 2.

All five covariates appear to be extremely significant from a hypothesis testing standpoint. In fact, four of the five p-values from the individual Type III F-tests are below the floating-point limit of the analysis software used. However, the calculated R^2 for this reduced model is only 0.111, meaning that these five variables collectively explain only about one-ninth of the variability in HDL-C levels from our data.

5.2 Full Model #1

After employing a model selection procedure known as the Feasible Solutions Algorithm [24], we arrived at the following $q = 3$ primary explanatory variables of interest: body mass index (BMI), triglyceride level, and mean blood cell volume (BCV). Fitting the full model and again performing individual Type III F-tests for each variable leads to the results given in Table 3.

It appears that the relationships between HDL-C levels and each of the eight variables in our full model are highly significant, as all eight p-values are now below the floating point limit of our analysis software. Thus, it seems that the addition of our three primary variables of interest improved the model, which is supported by a partial-F test statistic of 2823.2 and corresponding p-value $< 2.2 \times 10^{-16}$.

Table 3 For each explanatory variable in the first full HDL-C model, individual Type III test statistics and their corresponding p-values are given

Variable	F-statistic	p-value
Age	593.6	$< 2.2 \times 10^{-16}$
Race	93.7	$< 2.2 \times 10^{-16}$
Gender	3237.4	$< 2.2 \times 10^{-16}$
Education Level	113.7	$< 2.2 \times 10^{-16}$
Year	45.0	$< 2.2 \times 10^{-16}$
BMI	2208.1	$< 2.2 \times 10^{-16}$
Triglycerides	4055.6	$< 2.2 \times 10^{-16}$
Mean BCV	251.0	$< 2.2 \times 10^{-16}$

As we saw with the reduced model, small p-values do not necessarily guarantee a strong model. Looking at the coefficient of multiple determination for this full model, however, we get a calculated R^2 of 0.313. This means that we are now explaining almost a third of the variability in HDL-C levels from our data, a marked improvement from the reduced model. This improvement is quantified by a calculated partial R^2 of 0.227, meaning that the three primary variables of interest are collectively able to account for almost a quarter of the remaining variability in HDL-C levels, after adjusting for our five covariates.

5.3 Full Model #2

An alternative model selection procedure resulted in a different set of $q = 3$ primary explanatory variables of interest: diastolic blood pressure (BP), mean platelet volume (MPV), and monocyte percentage. Fitting the full model and performing individual Type III F-tests for each variable leads to the results given in Table 4.

Once again, it appears that the relationships between HDL-C levels and each of the eight variables in this second full model are highly significant from a hypothesis testing standpoint, as all eight p-values are far below any reasonable significance level, including six that are below the floating point limit of our analysis software. It

Table 4 For each explanatory variable in the second full HDL-C model, individual Type III test statistics and their corresponding p-values are given

Variable	F-statistic	p-value
Age	58.9	1.7×10^{-14}
Race	115.5	$< 2.2 \times 10^{-16}$
Gender	2721.7	$< 2.2 \times 10^{-16}$
Education Level	58.3	$< 2.2 \times 10^{-16}$
Year	28.6	$< 2.2 \times 10^{-16}$
Diastolic BP	89.7	$< 2.2 \times 10^{-16}$
MPV	64.9	8.3×10^{-16}
Monocyte %	181.0	$< 2.2 \times 10^{-16}$

would appear that the addition of this new set of three primary variables of interest improved the model as well, which is supported by a partial-F test statistic of 118.3 and corresponding p-value $< 2.2 \times 10^{-16}$.

Looking at the coefficient of multiple determination for this model, however, we get a calculated R^2 of just 0.122, which doesn't suggest much of an improvement from the reduced model R^2 of 0.111. This lack of meaningful improvement is quantified by a calculated partial R^2 of just 0.012. That is, this set of three primary variables of interest are only able to collectively account for just over 1% of the remaining variability in HDL-C levels, after adjusting for our five covariates.

5.4 Discussion

From a hypothesis testing perspective, both full models in our example analysis appear to contain a set of explanatory variables with strong evidence supporting their inclusion in any regression model for HDL-C levels, even after adjusting for the five covariates detailed in the Reduced Model section. Even the most reasonably conservative cutoff value of α or the most conservative multiple-testing correction procedures wouldn't come close to changing the high level of significance indicated by the p-values associated with each of the effects from these models. And for most peer reviewers, citing such significant p-values as evidence of markedly strong association would be sufficiently conclusive. In fact, it might be difficult for the data scientist to choose a "best" model between the two, leading to the temptation to simply aggregate them into a single, larger model containing all of these seemingly-important predictor variables.

However, when we move beyond hypothesis testing and consider other measures of model quality – in this case, coefficients of multiple and partial determination – we see that there are stark contrasts between each full model's ability to represent the NHANES data. Table 5 summarizes these measures for each model discussed above.

Full Model #1 explains nearly 20% more overall variability in HDL-C levels than its counterpart. Even more strikingly, it explains nearly 20 times more of the post-adjustment variability in HDL-C levels than Full Model #2. In large datasets like NHANES, the usefulness and informativeness of hypothesis testing is reduced, demanding a deeper and more nuanced approach to regression modeling. Measures like R^2 and partial R^2, whose construction as ratios makes them more interpretable and more robust to extreme sample sizes, can help us better understand and identify the real relationships that exist (or fail to exist) in our data.

A natural follow-up question we might ask is whether a partial R^2 of 0.227 is large enough to be useful or meaningful to doctors helping patients manage their cholesterol levels in the real world, given that nearly three-quarters of the post-adjustment

Table 5 For each of the three HDL-C models fit, the values of the coefficients of multiple determination and partial determination are given

Model	R^2	Partial R^2
Reduced Model	0.111	–
Full Model #1	0.313	0.227
Full Model #2	0.122	0.012

variability (and two-thirds of the overall variability) is still unaccounted for in our best model. And might there exist a Full Model #3 whose $q = 3$ explanatory variables explain even more of the variability in HDL-C levels than Full Model #1 does? While the focus of this paper is not on model-building, nor was our motivation for this example to discover new scientific relationships regarding a person's HDL-C levels, these questions are important to carefully consider in practice.

Unlike hypothesis testing, in which the significance level $\alpha = 0.05$ has been the gatekeeper of statistical significance for nearly a century, there is no universally-accepted "cutoff" for statistical (or practical) significance of a regression model or subset of variables based on R^2 or partial R^2. This provides statisticians and their collaborators with both substantial freedom and a substantial challenge. On the one hand, researchers are given the flexibility to decide what a meaningful value of R^2 or partial R^2 might be in their particular context and field of study. Additionally, they are given a statistic whose interpretability greatly improves their ability to describe the linear relationships being reported from their data. On the other hand, making an intelligent choice requires careful consideration and collaborative thinking, in context, for each specific discipline of study. Additionally, justifying any such decision to academic journal reviewers and communicating the implications to the reader becomes more challenging. But as our NHANES data analysis demonstrates, the dangers of relying solely on p-values to assess relationships in a regression context strikingly jeopardize the quality and effectiveness of our modeling efforts and the overarching scientific research they represent.

6 Conclusion

In this paper, we derived the complete distribution of the partial coefficient of determination in the context of linear regression modeling, with supporting evidence from a simulation study. We showed that partial R^2 follows a non-central beta distribution similar in structure to that of the coefficient of multiple determination, though the dependence of the non-centrality parameter on the nested projection matrices makes our continued study of its statistical properties more nuanced. From our analyses of the aggregated NHANES dataset, we demonstrated the urgent need to move beyond the reporting of p-values in isolation, particularly for linear regression models involving large datasets. R^2 and partial R^2, as the estimated proportions of response variability explained by the model or by a model subset, add a richer and more informative element to regression analysis.

Although statistical inference methods would not provide additional information about partial R^2 for large data sets, future directions of this work include using the distributional results derived here to develop methodology for performing confidence intervals and hypothesis testing for researchers to use when analyzing small- or medium-sized data sets. Future extensions of this research also include pseudo-R^2 measures, which are often used to describe regression models built under generalized linear model frameworks like logistic and Poisson regression. We also see extensions of this theory to repeated-measures and mixed-modeling applications as high-leverage opportunities for improved practice and theoretical understanding. Clearly, within the

context of each researcher's field of study, work remains to determine what values of partial R^2 are noteworthy and represent appreciable contextual value. But as we move beyond a world where $p < 0.05$ is the blind gatekeeper to statistical significance and scientific importance, measures like the coefficients of determination will become an increasingly invaluable tool for analyzing Big Data.

Acknowledgements The authors would like to thank Dr. Arnold Stromberg, Dr. Constance Wood, and Dr. Cale Jacobs at the University of Kentucky for their expertise and guidance throughout the writing of this manuscript.

Data Availibility Statement The data that support the findings of this study are openly available on the NHANES website, at https://www.cdc.gov/nchs/nhanes/index.htm.

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Nickerson, R.S.: Null-hypothesis Significance Testing: Misconceptions. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science., pp. 1006–1009. Springer, Berlin, Heidelberg (2011)
2. Wasserstein, R.L., Lazar, N.A.: The ASA Statement on p-values: Context, Process, and Purpose. American Stat. **70**(2), 129–133 (2016). https://doi.org/10.1080/00031305.2016.1154108
3. Khalilzadeh, J., Tasci, A.D.: Large Sample size, Significance Level, and the Effect Size: Solutions to Perils of Using Big Data for Academic Research. Tourism Manag. **62**, 89–96 (2017). https://doi.org/10.1016/j.tourman.2017.03.026
4. Lin, M., Jr., H.C.L., Shmueli, G.: Too Big to Fail: Large Samples and the p-value Problem. Inf. Syst. Res. **24**(4), 906–917 (2013). https://doi.org/10.1287/isre.2013.0480
5. Xiang, Z., Schwartz, Z., Gerdes, J.H., Uysal, M.: What Can Big Data and Text Analytics Tell Us About Hotel Guest Experience and Satisfaction? Int. J. Hospitality Manag. **44**, 120–130 (2015). https://doi.org/10.1016/j.ijhm.2014.10.013
6. Lantz, B.: The Large Sample Size Fallacy. Scandinavian J. Caring Sci. **27**(2), 487–492 (2013). https://doi.org/10.1111/j.1471-6712.2012.01052.x
7. Lee, C.H., Yoon, H.J.: Medical Big Data: Promise and Challenges. Kidney Res. Clinical Pract. **36**(1), 3 (2017). https://doi.org/10.23876/j.krcp.2017.36.1.3
8. Fraser, D.A.S.: On Evolution of Statistical Inference. J. Stat. Theory Appl. **17**, 193–205 (2018). https://doi.org/10.2991/jsta.2018.17.2.1
9. Laan, M., Rose, S.: Statistics Ready For a Revolution: Next Generation of Statisticians Must Build Tools for Massive Data Sets. Amstat News **399**, 38–39 (2010)
10. Wasserstein, R.L., Schirm, A., Lazar, N.: Moving to a World Beyond 'p<0.05'. American Stat. **73**(sup1), 1–19 (2019). https://doi.org/10.1080/00031305.2019.1583913
11. Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models, vol. 5. McGraw-Hill Irwin, Boston (2005)978-0-07-310875-2

12. Kaur, K.: Artificial Neural Network Model to Forecast Energy Consumption in Wheat Production in India. J. Stat. Theory Appl. **22**, 19–37 (2023). https://doi.org/10.1007/s44199-023-00052-w
13. Kim, D.Y., Tsokos, C.P.: A Stochastic Approach in Modeling of Regional Atmospheric CO_2 in the United States. J. Stat. Theory Appl. **19**, 10–20 (2020). https://doi.org/10.2991/jsta.d.200224.002
14. Koerts, J., Abrahamse, A.P.J.: The Correlation Coefficient in the General Linear Model. European Econ. Rev. **1**(3), 401–427 (1970)
15. Barten, A.P.: Note on Unbiased Estimation of the Squared Multiple Correlation Coefficient. Stat. Neerlandica **16**(2), 151–164 (1962). https://doi.org/10.1111/j.1467-9574.1962.tb01062.x
16. Cramer, J.S.: Mean and Variance of R^2 in Small and Moderate Samples. J. Economet. **35**(2–3), 253–266 (1987). https://doi.org/10.1016/0304-4076(87)90027-3
17. Johnson, N.L., Kotz, S.: Distributions in Statistics: Continuous Univariate Distributions, vol. 2. Wiley, New York (1970)978-0-4714-4626-2
18. Ravishanker, N., Dey, D.K.: A First Course in Linear Model Theory. Chapman & Hall/CRC, New York (2002)1-58488-247-6
19. Ali, K.M., Wonnerth, A., Huber, K., Wojta, J.: Cardiovascular Disease Risk Reduction by Raising HDL Cholesterol - Current Therapies and Future Opportunities. British J. Pharmacol. **167**(6), 1177–1194 (2012). https://doi.org/10.1111/j.1476-5381.2012.02081.x
20. Centers for Disease Control and Prevention: Underlying Cause of Death, 1999-2018. CDC Wonder Online Database (2018). Accessed August 10, 2021
21. Virani, S.S., Alonso, A., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P.: Heart Disease and Stroke Statistics - 2020 Update: a Report from the American Heart Association. Circulation **141**(9), 139–596 (2020). https://doi.org/10.1161/CIR.0000000000000757
22. Fryar, C.D., Chen, T., Li, X.: Prevalence of Uncontrolled Risk Factors for Cardiovascular Disease: United States, 1999-2010. NCHS Data Brief (103) (2012)
23. Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P.: Heart Disease and Stroke Statistics - 2019 Update: a Report from the American Heart Association. Circulation **139**(10), 56–528 (2019). https://doi.org/10.1161/CIR.0000000000000659
24. Lambert, J., Gong, L., Elliott, C.F., Thompson, K., Stromberg, A.: rFSA: An R Package for Finding Best Subsets and Interactions. R J. **10**(2) (2018). https://doi.org/10.32614/RJ-2018-059