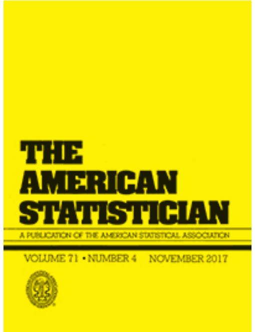

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: www.tandfonline.com/journals/utas20

A Coefficient of Determination for Generalized Linear Models

Dabao Zhang

THE AMERICAN STATISTICIAN
2017, VOL. 71, NO. 4, 310–316: General
https://doi.org/10.1080/00031305.2016.1256839

A Coefficient of Determination for Generalized Linear Models

Dabao Zhang

Department of Statistics, Purdue University, West Lafayette, IN

ABSTRACT

The coefficient of determination, a.k.a. R^2, is well-defined in linear regression models, and measures the proportion of variation in the dependent variable explained by the predictors included in the model. To extend it for generalized linear models, we use the variance function to define the total variation of the dependent variable, as well as the remaining variation of the dependent variable after modeling the predictive effects of the independent variables. Unlike other definitions that demand complete specification of the likelihood function, our definition of R^2 only needs to know the mean and variance functions, so applicable to more general quasi-models. It is consistent with the classical measure of uncertainty using variance, and reduces to the classical definition of the coefficient of determination when linear regression models are considered.

ARTICLE HISTORY
Received December 2015
Revised October 2016

KEYWORDS
Exponential family distribution; Quasi-model; R^2; Variance function

1. Introduction

Consider the linear regression model,

$$y_i = X_i\beta + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad i = 1, 2, \ldots, n,$$

where y_i is the ith component of the n-dimensional column vector Y, X_i is the ith row of the $n \times p$ design matrix $\mathbf{X}$, and β is a p-dimensional column vector of unknown regression coefficients. An estimate $\hat{\beta}$ of β provides fitted values $\hat{y}_i(\mathbf{X}) = X_i\hat{\beta}$, and

$$\text{SSE}(\mathbf{X}) = \sum_{i=1}^{n}\{y_i - \hat{y}_i(\mathbf{X})\}^2$$

accounts for variation in the responses not explained by the available predictors. On the other hand, when $p = 1$ and $X_i \equiv 1$, that is, no predictor is considered, the estimate $\hat{\beta} = \bar{y}$ implies fitted values $\hat{y}_i(\mathbf{1}_n) = \bar{y}$, and

$$\text{SSE}(\mathbf{1}_n) = \sum_{i=1}^{n}\{y_i - \hat{y}_i(\mathbf{1}_n)\}^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \text{SST}$$

accounts for the total variation in the responses. Assuming that every model includes an intercept term, the coefficient of determination

$$R^2 = 1 - \frac{\text{SSE}(\mathbf{X})}{\text{SSE}(\mathbf{1}_n)} \tag{1}$$

measures the proportion of variation in the responses explained by the available predictors.

This coefficient of determination is well-defined for linear regression models, and is popularly used in practice as a measure of goodness of fit of the underlying models. However, its extension to generalized linear models (GLMs) and other more general models is not straightforward. Different perspectives led to several generalizations to the coefficient of determination.

Let $\ell(Y, \mu(\mathbf{X}))$ denote the log-likelihood of model $E[Y|\mathbf{X}] = \mu(\mathbf{X})$ for observed data $(Y, \mathbf{X})$, and $\ell(Y, \mu(\mathbf{1}_n))$ the log-likelihood of model $E[Y] = \mu(\mathbf{1}_n) = \mu\mathbf{1}_n$. Magee (1990) observed the relationship between R^2 and the likelihood ratio statistic in the linear regression models,

$$R^2_{\text{LR}} = 1 - \exp\left\{\frac{2}{n}\ell(Y, \hat{\mu}(\mathbf{1}_n)) - \frac{2}{n}\ell(Y, \hat{\mu}(\mathbf{X}))\right\},$$

where both $\hat{\mu}(\mathbf{1}_n)$ and $\hat{\mu}(\mathbf{X}_n)$ are obtained via maximizing the corresponding likelihood functions. It is therefore proposed to generalize R^2 by using R^2_{LR} for more general models with well-defined likelihood function. This generalization coincides with Maddala (1983) and Cox and Snell (1989).

For a logistic regression model, perfectly fitted values result in $\ell(Y, \hat{\mu}(\mathbf{X})) = 0$, therefore,

$$\max\{R^2_{\text{LR}}\} = 1 - \exp\left\{\frac{2}{n}\ell(Y, \hat{\mu}(\mathbf{1}_n))\right\}.$$

For example, with balanced case-control data, $\max\{R^2_{\text{LR}}\} = 0.75$. R^2_{LR} is bounded from above by $\ell(Y, \hat{\mu}(\mathbf{1}_n))$ and will never attain value one. To address this issue, Nagelkerke (1991) suggested the following correction,

$$R^2_N = \left[1 - \exp\left\{\frac{2}{n}\ell(Y, \hat{\mu}(\mathbf{1}_n)) - \frac{2}{n}\ell(Y, \hat{\mu}(\mathbf{X}))\right\}\right] \Bigg/ \left[1 - \exp\left\{\frac{2}{n}\ell(Y, \hat{\mu}(\mathbf{1}_n))\right\}\right].$$

However, such correction makes R^2_N inconsistent with the classical definition of coefficient of determination.

Cameron and Windmeijer (1997) proposed to use the Kullback–Leibler divergence to quantify the uncertainty

CONTACT Dabao Zhang zhangdb@stat.purdue.edu Department of Statistics, Purdue University, West Lafayette, IN 47907.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.

remaining in the response after accounting for predictors. That is, variation in the response unexplained by $\mathbf{X}$ is quantified by the estimated Kullback–Leibler divergence between Y and $\hat{\mu}(\mathbf{X})$, with I_n the n-dimensional identity matrix,

$$\widehat{\mathrm{KL}}(Y, \hat{\mu}(\mathbf{X})) = 2\ell(Y, \hat{\mu}(I_n)) - 2\ell(Y, \hat{\mu}(\mathbf{X})).$$

Here, the identity matrix I_n implies that $\mu(I_n)$ is the mean of Y following the saturated model, and $\hat{\mu}(I_n)$ is its maximum likelihood estimator (MLE). Then the coefficient of determination is generalized as

$$R^2_{\mathrm{KL}} = 1 - \widehat{\mathrm{KL}}(Y, \hat{\mu}(\mathbf{X})) \big/ \widehat{\mathrm{KL}}(Y, \hat{\mu}(\mathbf{1}_n)).$$

Since both $\widehat{\mathrm{KL}}(Y, \hat{\mu}(\mathbf{X}))$ and $\widehat{\mathrm{KL}}(Y, \hat{\mu}(\mathbf{1}_n))$ are deviances, R^2_{KL} can be interpreted as the deviance reduction ratio due to the predictors in $\mathbf{X}$.

All the aforementioned generalizations of coefficients of determination are given on the basis of the completely specified likelihood function, but not applicable to more general GLMs, like quasi-models, which specify only the mean and variance functions. The classical R^2 is well defined for general linear models as long as the error terms are homoscedastic. In this work, we intend to define a coefficient of determination for generalized linear models, which acknowledges the relationship between the mean and variance functions.

In the next section, we introduce our definition of coefficient of determination, assuming only mean and variance functions are well-specified. We also propose an adjustment to account for the number of predictors in the model. Advantages and disadvantages of different definitions are investigated via simulation studies in Section 3, which also show the robustness of our proposed definition. We also apply and compare different coefficients of determination to a set of real data in Section 4, and conclude this article with Section 5.

2. Measuring Variation Changes Along the Variance Function

All previous generalizations of coefficient of determination asymptotically approach to measures involving entropy, which measures both uncertainty and information (Shannon 1948). We instead follow popular statistical practice to consider a simpler measure of uncertainty, that is, the variance, which can be specified via a dispersion parameter ϕ and a known variance function $V(\cdot)$ in generalized linear models. That is, with $E[y_i|X_i] = \mu(X_i)$, $i = 1, 2, \ldots, n$,

$$\mathrm{var}(y_i|X_i) = \phi V(\mu(X_i)).$$

In general, as long as the mean $\mu(X_i)$ can be modeled well and linked appropriately to a set of predictors, a generalized linear model with known variance function $V(\cdot)$ can be investigated for the utility of these predictors.

While the variance function describes the effect of the mean on the variation of the response variable besides the dispersion parameter, Jorgensen (1987) showed that the variance function $V(\cdot)$ indeed characterizes the underlying exponential family distribution. For a response variable with its mean changing from a to b, its variation moves accordingly along the variance function from $\phi V(a)$ to $\phi V(b)$. Therefore, the variation change

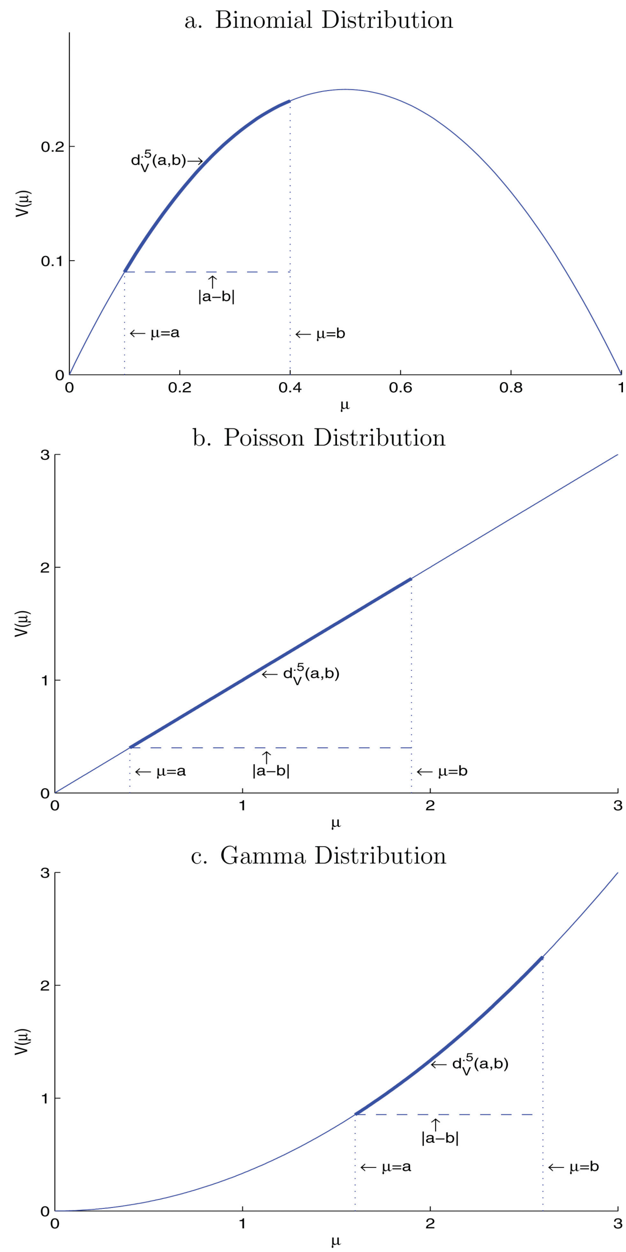

Figure 1. The variance functions of binomial, Poisson and gamma distributions. The length of the thick line is $\sqrt{d_V(a, b)}$, and the length of the dashed line is $|a - b|$. These two lengths may differ dramatically in binomial and gamma distributions.

of the response variable should be measured using, instead of $(a - b)^2$, the squared length of the variance function $V(\cdot)$ between $V(a)$ to $V(b)$, that is,

$$d_V(a, b) = \left\{ \int_a^b \sqrt{1 + [V'(t)]^2 dt} \right\}^2.$$

As shown in Figure 1, $d_V(a, b)$ can differ dramatically from the Euclidean distance $(a - b)^2$ when the underlying variance function is nonlinear.

As shown by Morris (1982, 1983), many popularly considered exponential family distributions, such as binomial, negative binomial, and gamma distributions, have quadratic variance functions. We assume a general case, $v_2 \neq 0$,

$$V(\mu) = v_2\mu^2 + v_1\mu + v_0.$$

Then, $d_V(a, b)$ can be calculated as

$$d_V(a,b) = \frac{1}{16v_2^2}\left\{\log\frac{V'(b)+\sqrt{1+[V'(b)]^2}}{V'(a)+\sqrt{1+[V'(a)]^2}} + V'(b)\sqrt{1+[V'(b)]^2} - V'(a)\sqrt{1+[V'(a)]^2}\right\}^2.$$

When $v_2 = 0$, that is, the variance function is linear or constant as in the case of Poisson distribution or normal distribution, we have $d_V(a, b) = (1 + v_1^2)(b - a)^2$.

While the total variation in Y is $\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{1}_n))$, the model with predictors $\mathbf{X}$ reduces the unexplained variation in Y to $\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{X}))$. Therefore, we define the coefficient of determination as

$$R_V^2 = 1 - \frac{\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{X}))}{\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{1}_n))}. \quad (2)$$

Such a coefficient of determination is well-defined as long as the mean and variance functions are specified, like the quasi-models. Therefore, $\hat{\mu}_i(\mathbf{X})$ and $\hat{\mu}_i(\mathbf{1}_n)$ may be derived from quasi-likelihood estimators, other than MLE.

Because $V'(\cdot)$ is constant for normal and Poisson distributions, R_V^2 is consistent with the classical definition (1) of the coefficient of determination in the case of linear regression models following normal distributions, and log-linear regression models following Poisson distributions, that is, $R_V^2 = R^2$.

Similar to the coefficient of determination, the coefficient of partial determination is well-defined for linear models, measuring the proportion of variation in the response variable not explained by a set of predictors that can be explained by an additional set of predictors. For example, considering two sets of predictors X_1 and X_2 in a linear regression model, we have

$$R^2(X_2|X_1) = 1 - \frac{\mathrm{SSE}(X_1, X_2)}{\mathrm{SSE}(X_1)} = \frac{R^2(X_1, X_2) - R^2(X_1)}{1 - R^2(X_1)},$$

measuring the proportion of remaining variation in the response, when including X_1, explained by X_2. With our definition of R_V^2, we can easily extend it to a coefficient of partial determination for more general models,

$$R_V^2(X_2|X_1) = \frac{R_V^2(X_1, X_2) - R_V^2(X_1)}{1 - R_V^2(X_1)}.$$

Indeed, when X_2 is a univariate variable, we can also define a model-based measure of partial correlation between the response and X_2, given X_1 in the model, as follows:

$$r_V(X_2|X_1) = \mathrm{sign}(\hat{\beta}_2)\sqrt{R_V^2(X_2|X_1)},$$

where $\hat{\beta}_2$ is the regression coefficient of X_2 when regressing against both X_1 and X_2 with a monotonically increasing link function.

While it is well-defined as the mean and variance functions are known, R_V^2 also suffers to increasing numbers of predictors as the classical R^2, and may increase even if irrelevant predictors are added to the underlying model. Therefore, averaged measures of the variation change along the variance function can be used to take consideration of effects caused by different numbers of predictors. That is, we define an adjusted version of R_V^2 as follows:

$$R_{V,\mathrm{adj}}^2 = 1 - \frac{\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{X}))/(n-p)}{\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{1}_n))/(n-1)}. \quad (3)$$

As R_V^2, $R_{V,\mathrm{adj}}^2$ is well-defined as long as the underlying model, like quasi-models, specifies the mean and variance functions.

3. Empirical Studies

To compare the performance of the different definitions of R^2, we consider samples from two populations following the same exponential family distribution but with means μ_1 and μ_2, respectively. A total of 100 datasets are simulated for each of the investigated exponential family distributions, and each dataset has 50 random observations with 25 from each population. Hereafter, we use X_1 to refer to the population of the corresponding observation. We also generate a second variable X_2 from the standard normal distribution, which is independent of X_1 and the response variable.

With a fixed β, we generate 100 datasets from a binomial model by setting $\mu_1 = \frac{e^{-\beta}}{1+e^{-\beta}}$, and $\mu_2 = \frac{e^{\beta}}{1+e^{\beta}}$. That is, we model the mean $\mu(X_1) = \frac{e^{X_1\beta}}{1+e^{X_1\beta}}$, with $X_1 = 1$ or -1 indicating the two different populations. The coefficients of determination are averaged over the 100 datasets for each β ranging from zero to five with step 0.1, and are shown in Figure 2.

When $\beta = 0$, we have $\mu_1 = \mu_2$, and thus corresponding coefficients of determination are supposed to report zero. Though not exactly zero, the calculated values are all close to zero when $\beta = 0$. On the other hand, the difference between the two population means $|\mu_2 - \mu_1|$ reaches the maximum when β goes to infinity. Therefore, it is not surprising to observe that, when including the true predictor X_1, R_N^2, R_{KL}^2, and R_V^2 approach one. However, R_{LR}^2 is bounded under 0.75 as mentioned by Nagelkerke (1991) though it performs similarly as R_V^2 for evaluating the model with irrelevant X_2 only. Overall, R_{KL}^2 reports the smallest values, and R_N^2 instead reports the largest values with R_V^2 in the between.

We also generate datasets from a Poisson model and a gamma model, respectively. We model $\mu(X_1) = e^{X_1\beta}$ for the Poisson model, and $\mu(X_1) = \frac{100}{2+X_1\beta}$ for the gamma model with the shape parameter $\nu = 100$. The coefficients of determination are averaged over 100 datasets for each β ranging from zero to five with step 0.1, and are shown in Figures 3 and 4 for the Poisson model and gamma model, respectively.

For both Poisson model and gamma model, the calculated coefficients of determination are all close to zero when $\beta = 0$, and approach one for large values of β. On the other hand, the difference between the two population means $|\mu_2 - \mu_1|$ reaches the maximum in the Poisson model but is bounded by 50 in the gamma model when β goes to infinity. Therefore, it is not surprising to observe that, when including the true predictor X_1, R_{KL}^2, and R_V^2 approach one in the Poisson model, but are barely bounded away from one in the gamma model. The drawbacks of R_{LR}^2 and R_N^2 are, (i) both R_{LR}^2 and R_N^2 quickly approach one in the Poisson and gamma model as β increases, so they may easily overstate R^2 for both models; (ii) both R_{LR}^2 and R_N^2 can falsely claim high contribution from an irrelevant predictor in the Poisson and gamma model as shown in Figures 3(c) and 4(c), respectively.

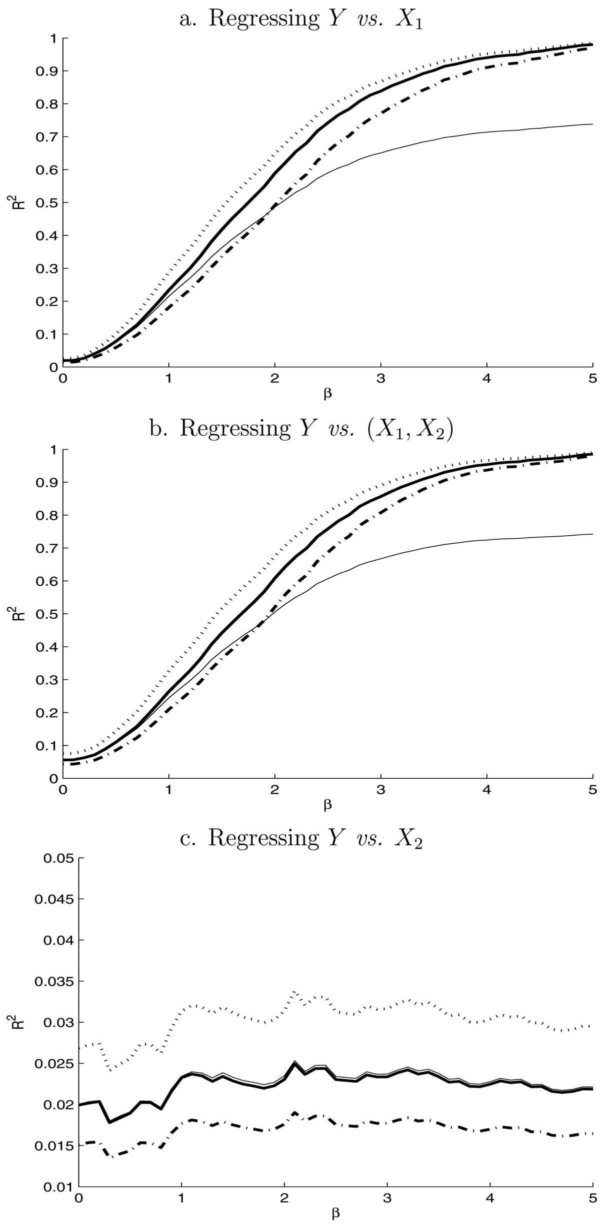

Figure 2. The coefficient of determination in binomial models. Shown in the plot are R^2_{LR} (with thin solid lines), R^2_N (with dotted lines), R^2_{KL} (with dotted-dashed lines), and R^2_V (with thick solid lines).

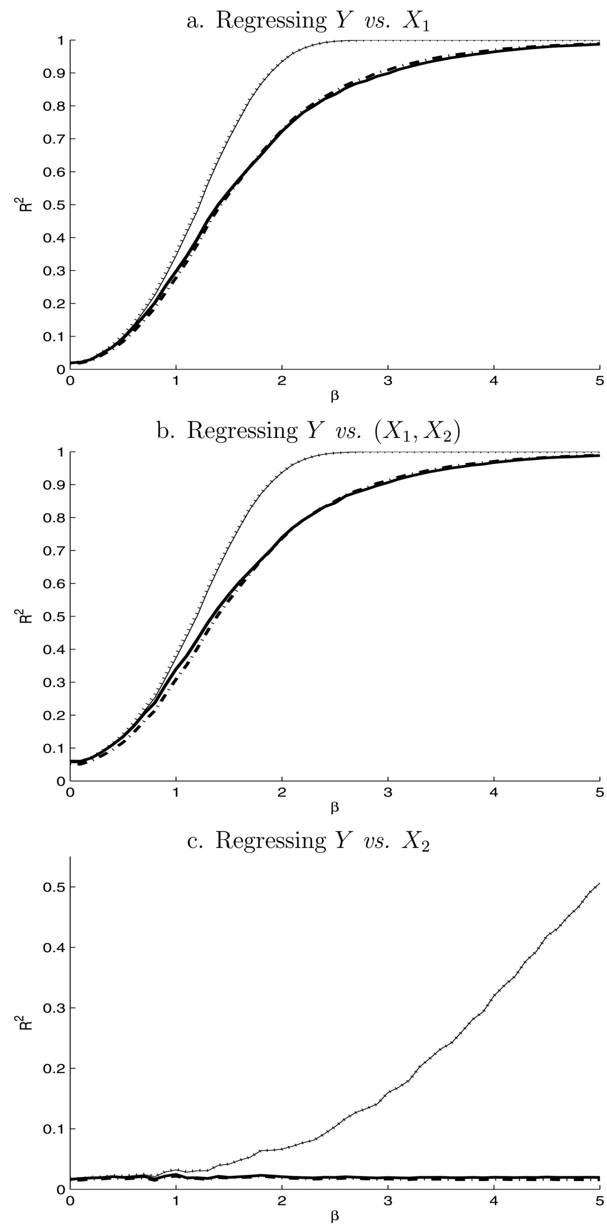

Figure 3. The coefficient of determination in Poisson models. Shown in the plot are R^2_{LR} (with thin solid lines), R^2_N (with dotted lines), R^2_{KL} (with dotted-dashed lines), and R^2_V (with thick solid lines).

To understand the false claim of high level of R^2 by R^2_{LR} and R^2_N, we consider the case that $\beta = 5$, and split each dataset into two subsets, one with $E[X_1] = \mu_1$ and another with $E[X_1] = \mu_2$. We further calculate the sample mean of X_2 for each subset, denoted as $\bar{X}_{21}$ and $\bar{X}_{22}$, respectively. With the total sample size at 50, we have $\bar{X}_{21} - \bar{X}_{22} \sim N(0, 0.08)$. A large value of $\bar{X}_{21} - \bar{X}_{22}$ implies falsely correlated X_1 and X_2 although X_1 and X_2 are truly independent. We plot calculated R^2 versus $\bar{X}_{21} - \bar{X}_{22}$ in Figure 5 to show the robustness of different coefficients of determination when only a falsely correlated X_2 is included. While all versions of coefficients of determination reported low values in Binomial models, R^2_{LR} and R^2_N are vulnerable to the falsely correlated X_2 in both Poisson and gamma models as they can be over 0.8 when $|\bar{X}_{21} - \bar{X}_{22}|$ is greater than 0.4 (note that its probability is more than 0.15). Therefore, the false claim of large contribution from an irrelevant predictor by R^2_{LR} and R^2_N is due to their tendency to severely overstate the variation proportion explained by the Poisson or gamma model. On the other hand, both R^2_{KL} and R^2_V are more robust to such false correlation.

Including the true predictor X_1, the models with X_1 and X_2 reported all coefficients of determination similarly to the models with X_1 only, though the former models always have slightly higher values. We define $R^2_{V,\mathrm{adj}}$ to adjust for increasing numbers of predictors. Since the Kullback–Leibler divergences used to define R^2_{KL} are indeed deviances, we can also address the issue of different degrees of freedom in the deviances by defining an adjusted version of R^2_{KL} as follows:

$$R^2_{\mathrm{KL,adj}} = 1 - \frac{\widehat{\mathrm{KL}}(Y, \hat{\mu}(\mathbf{X}))/(n-p)}{\widehat{\mathrm{KL}}(Y, \hat{\mu}(\mathbf{1}_n))/(n-1)}.$$

As shown in Figure 6, both $R^2_{V,\mathrm{adj}}$ and $R^2_{\mathrm{KL,adj}}$ have lowered values when irrelevant predictors X_2 and X_3 are added to the model, with X_3 independently simulated from a standard normal distribution.

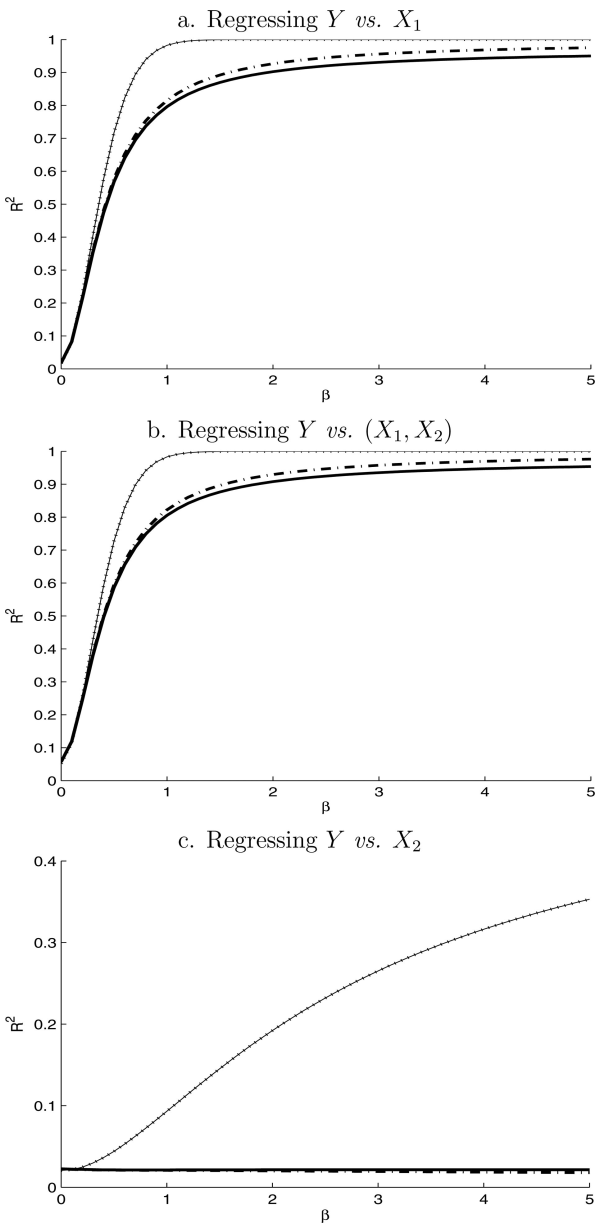

Figure 4. The coefficient of determination in gamma models. Shown in the plot are R^2_{LR} (with thin solid lines), R^2_N (with dotted lines), R^2_{KL} (with dotted-dashed lines), and R^2_V (with thick solid lines).

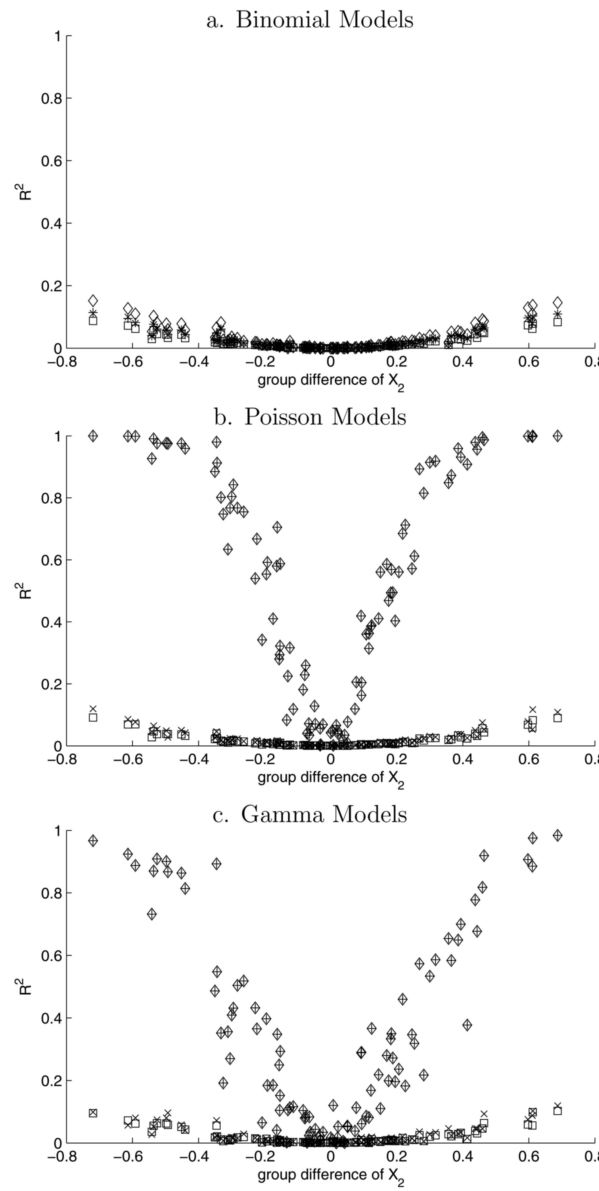

Figure 5. Effects of falsely correlated X_2 on R^2 when regressing Y versus X_2. The group difference of X_2 is calculated as the difference of averaged X_2 between the two groups defined by X_1, that is, $\bar{X}_{21} - \bar{X}_{22}$. Shown in the plot are R^2_{LR} (with $+$), R^2_N (with $\Diamond$), R^2_{KL} (with $\Box$), and R^2_V (with $\times$) for the corresponding models with $\beta = 5$.

4. Real Data Analysis

We illustrate the different definitions of R^2 by applying them to the data from a study of nesting horseshoe crabs included in Agresti (1996). This study collected colors (C), spine conditions (SC), carapace widths (CW), and weights (W) of 173 female crabs, each with a male crab attached to her in her nest. This study intended to investigate whether these factors affect the number of satellites, that is, any other males riding near a female crab.

We consider binomial models to investigate factors affecting whether a female crab had any satellites, and Poisson models to investigate factors affecting how many satellites a female crab had. Three dummy variables are used to code the four different colors, and two dummy variables are used to code the three spine conditions. For carapace width and weight, we also include their quadratic terms in the corresponding models. The different coefficients of determination are calculated for each model, shown in Table 1 as well as the Akaike information criterion (AIC) by Akaike (1974) and Bayesian information criterion (BIC) by Schwarz (1978). As demonstrated in the previous section, both R^2_{LR} and R^2_N are larger than the others, and indeed severely overstate the variation proportion explained by the Poisson models.

For factors affecting whether a female crab had any satellites, since SC itself contributed little, we may consider either (C,CW,W) or (C,CW) the best model, on the basis of different unadjusted versions of R^2. Unsurprisingly, all these R^2 reported the full model with the highest R^2 values. On the other hand, both $R^2_{\mathrm{KL,adj}}$ and $R^2_{V,\mathrm{adj}}$ reported the model (C,CW) with the highest R^2 values, and this model also has the smallest AIC value but the full model has the smallest BIC value. As shown in Table 2, given C and CW in the model, the other predictors make little contribution to explain the variation in the response variable, which is confirmed by the likelihood ratio tests

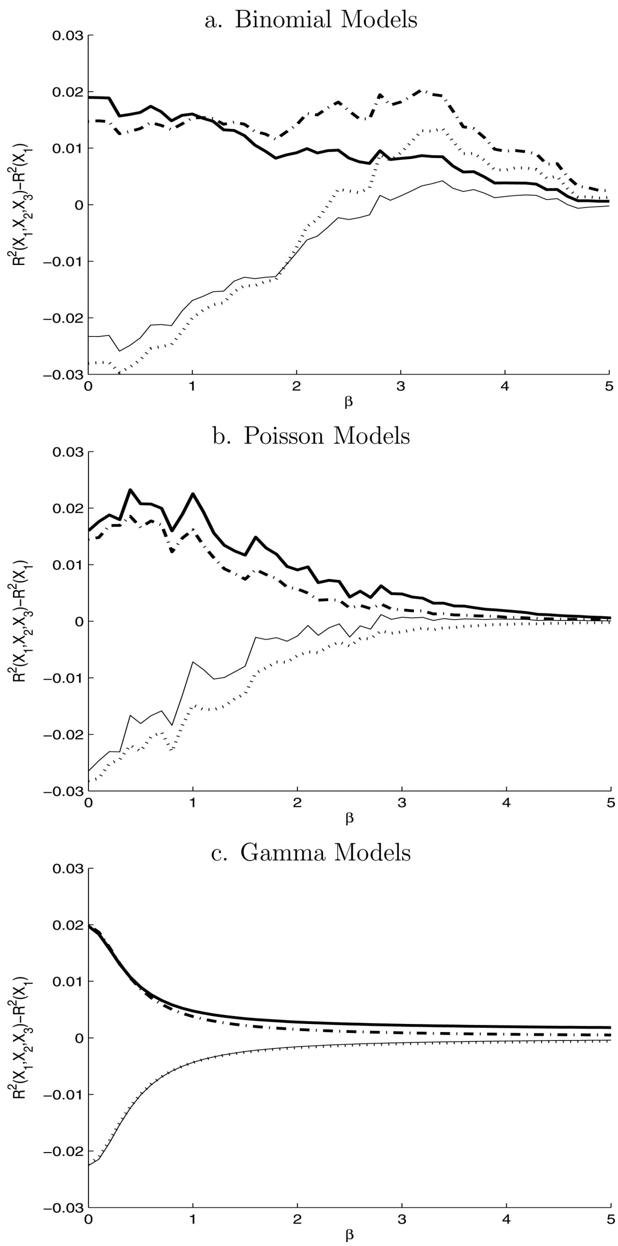

Figure 6. The difference of coefficients of determination between regressing Y versus (X_1, X_2, X_3) and regressing Y versus X_1. Shown in the plot are R^2_{KL} (with dotted-dashed lines), $R^2_{\mathrm{KL,adj}}$ (with dotted lines), R^2_V (with thick solid lines), and $R^2_{V,\mathrm{adj}}$ (with thin solid lines).

(with statistics shown under χ^2 in Table 2). $R^2_{\mathrm{KL,adj}}$ and $R^2_{V,\mathrm{adj}}$ further confirm that SC made no contribution to whether a female crab had any satellites.

For the number of satellites, we fitted Poisson models. Based on $R^2_{\mathrm{KL,adj}}$ and $R^2_{V,\mathrm{adj}}$, SC had almost no impact on the number of satellites a female crab might have, and C had very little impact, though both R^2_{LR} and R^2_N are more than 12% for C only, and more than 6% for SC only. Overall, we may take either the full model or (C,CW,W) based on different unadjusted versions of R^2. Indeed, the full model has the smallest BIC value and the model (C,CW,W) has the smallest AIC value. However, both $R^2_{\mathrm{KL,adj}}$ and $R^2_{V,\mathrm{adj}}$ would select W as the best model, which is further supported by the likelihood ratio tests and coefficients of partial determination in Table 2. Note that negative R^2_V(SC|C,CW,W) is because R^2_V of the full model is less

Table 1. Analysis of the horseshoe crabs data.

	R^2_{LR}	R^2_N	R^2_{KL}	$R^2_{\mathrm{KL,adj}}$	R^2_V	$R^2_{V,\mathrm{adj}}$	AIC	BIC
Binomial model								
(C,SC,CW,W)	0.2132	0.2925	0.1837	0.1387	0.2178	0.1746	204.27	194.59
(C,CW,W)	0.2080	0.2854	0.1787	0.1439	0.2130	0.1796	201.41	195.72
(C,CW)	0.2025	0.2779	0.1734	0.1487	0.2052	0.1814	198.61	196.91
(C,W)	0.1945	0.2669	0.1657	0.1408	0.2006	0.1767	200.34	198.65
(CW,W)	0.1760	0.2415	0.1483	0.1281	0.1727	0.1530	202.27	202.58
C	0.0761	0.1044	0.0607	0.0440	0.0807	0.0644	220.06	222.37
SC	0.0145	0.0199	0.0112	-.0004	0.0149	0.0033	229.23	233.54
CW	0.1695	0.2326	0.1423	0.1322	0.1636	0.1537	199.63	203.93
W	0.1607	0.2204	0.1342	0.1240	0.1582	0.1483	201.46	205.77
Poisson model								
(C,SC,CW,W)	0.4309	0.4324	0.1541	0.1074	0.1515	0.1046	910.56	900.87
(C,CW,W)	0.4298	0.4313	0.1536	0.1177	0.1518	0.1158	906.89	901.20
(C,CW)	0.3786	0.3798	0.1301	0.1040	0.1264	0.1002	917.79	916.10
(C,W)	0.4153	0.4167	0.1467	0.1212	0.1417	0.1160	907.25	905.55
(CW,W)	0.4057	0.4070	0.1423	0.1218	0.1422	0.1218	908.07	908.37
C	0.1287	0.1282	0.0374	0.0203	0.0396	0.0226	972.44	974.74
SC	0.0650	0.0652	0.0184	0.0068	0.0198	0.0083	982.46	986.76
CW	0.3501	0.3513	0.1178	0.1074	0.1171	0.1067	919.53	923.84
W	0.3909	0.3922	0.1356	0.1254	0.1341	0.1239	908.31	912.62

Table 2. Comparing candidate models for the horseshoe crabs data.

	Partial R^2_V	χ^2_{df}	df	p-Value
Binomial model				
(SC|C,CW,W)	0.0062	1.1313	2	0.5680
(SC,W|C,CW)	0.0159	2.3265	4	0.6759
(W|C,CW)	0.0098	1.1952	2	0.5501
Poisson model				
(SC|C,CW,W)	− 0.0003	0.3338	2	0.8462
(C,SC,CW|W)	0.0201	11.7500	7	0.1091
(C,CW|W)	0.0205	11.4160	5	0.0438

than that of model (C,CW,W), which further implies that SC does not affect the number of satellites.

Quasi-binomial models and quasi-Poisson models can be fitted for the above cases, with R^2_V and partial R^2_V calculated. Indeed, the estimated dispersion parameter is 1.0266 for the binomial full model, and 3.2354 for the Poisson full model. Therefore, it seems more appropriate to take quasi-Poisson models to investigate the risk factors for the number of satellites, and invalidates the use of likelihood-based coefficients of determination. Nonetheless, assuming Poisson or quasi-Poisson models, we will have identical estimators of the regression coefficients, which will lead to the same values of R^2_V and partial R^2_V, and hence the same conclusion as above.

5. Conclusion

The coefficient of determination, a.k.a. R^2, is a key statistic indicating how well a model including a set of predictors accounts for the variation in the response variable. While it shows the utility of these predictors in fitting the model, it also provides a measure of predictability of the response variable using the set of predictors. R^2 can be used to choose the optimal set of predictors when the model size, that is, the number of predictors, is fixed. The adjusted version of R^2, that is, R^2_{adj}, can be used to compare models including different numbers of predictors. For this reason, R^2_{adj} can be also used to help model selection, tuning parameter selection, etc. Our extension $R^2_{V,\mathrm{adj}}$ makes all these possible when any statistical model with a well-defined variance function, such as generalized linear models or even quasi-models, is considered.

As shown in the empirical studies and real data analysis, both R^2_{LR} and R^2_N tend to overstate the variation proportion explained by the model. Indeed, they may severely overstate the variation proportion explained by a Poisson model or gamma model so as to falsely claim a large contribution to an irrelevant predictor. R^2_V and R^2_{KL} avoid such overstatement and still range from zero to one. Furthermore, as it is well defined for quasi-models, R^2_V may be used to evaluate predictability of models built in machine learning or deep learning (Krizhevsky, Sutskever, and Hinton 2012). For example, when identifying risk factors for personalized medicines or cancer studies (Sirinukunwattana, Raza, and Tsang 2016), we may use R^2_V to measure the predictability with available risk factors, and investigate the importance of a set of potential risk factors using the corresponding coefficient of partial determination.

Acknowledgments

The author gratefully thanks Bruce A. Craig, George P. McCabe, the associate editor, and two anonymous reviewers for helpful comments that improved this article.

ORCID

Dabao Zhang http://orcid.org/0000-0003-0629-8672

References

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley. [314]

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723. [314]

Cameron, A. C., and Windmeijer, A. G. (1997), "An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models," *Journal of Econometrics*, 77, 329–342. [310]

Cox, D. R., and Snell, E. J. (1989), *The Analysis of Binary Data* (2nd ed.), London: Chapman and Hall. [310]

Jorgensen, B. (1987), "Exponential Dispersion Models," *Journal of the Royal Statistical Society*, Series B, 49, 127–162. [311]

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012), "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, 25. [316]

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge, UK: Cambridge University. [310]

Magee, L. (1990), "R^2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests," *The American Statistician*, 44, 250–253. [310]

Morris, C. N. (1982), "Natural Exponential Families with Quadratic Variance Functions," *The Annals of Statistics*, 10, 65–80. [311]

——— (1983), "Natural Exponential Families with Quadratic Variance Functions: Statistical Theory," *The Annals of Statistics*, 11, 515–529. [311]

Nagelkerke, N. J. D. (1991), "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78, 691–692. [310,312]

Schwarz, G. E. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [314]

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *The Bell System Technical Journal*, 27, 379–423, 623–656. [311]

Sirinukunwattana, K., Raza, S., Tsang, Y. W., Snead, D., Cree, I., and Rajpoot, N. (2016), "Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images," *IEEE Transactions on Medical Imaging*, 35, 1196–1206. [316]