A Coefficient of Determination for GLMs

by Dabao Zhang

Rearranged by Jae Ho, Chang
Spring, 2018

OLS

R^2_{OLS} ; A Coefficient of Determination for OLS

Consider a linear model ;

$$\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1} \ , \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathrm{I}_n)$$

Then we have

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

$$R^2_{OLS} \overset{def}{=} 1 - \frac{SSE}{SST}$$

OLS

1. R^2_{OLS} measures **the goodness of fit**
2. However, for GLMs?
3. Some GLMs with well-defined likelihood

Specified Likelihood

$l(\boldsymbol{y}, \mu(\boldsymbol{X}))$; The log-likelihood of model $E(\boldsymbol{y}|\boldsymbol{X}) = \mu(\boldsymbol{X})$ for observed data $(\boldsymbol{y}, \boldsymbol{X})$

$l(\boldsymbol{y}, \mu(\mathbf{1}_n))$; The log-likelihood of model $E(\boldsymbol{y}) = \mu(\mathbf{1}_n) = \mu\mathbf{1}_n$

Magee(1990) ; realtionship between R^2 and the likelihood ratio statistics in the linear regression models

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

where $\hat{\mu}$'s are obtained by MLE for corresponding model.
For proof, refer to "R^2 Measures based on Wald and Likelihood Ratio Joint Significance Tests, Magee, L. ,1990"

Specified Likelihood

For a logistic regression, perfectly fitted values result in $l(\boldsymbol{y}, \hat{\mu}(\boldsymbol{X})) = 0$. That is, samples can be perfectly separated by a linear function.

$$max(R^2_{LR}) = 1 - exp\left\{\frac{2}{n} l(\boldsymbol{y}, \hat{\mu}(\mathbf{1}_n))\right\}$$

For example, with balanced case-control data (Bernoulli-distributed with $\pi = 0.5$), $max(R^2_{LR}) = 0.75$.
That is, R^2_{LR} is bounded from above by $l(\boldsymbol{y}, \hat{\mu}(\mathbf{1}_n))$ and will never attain value 1.
This is incosistent with the existing concept of R^2.

Specified Likelihood

Nagelkerke(1991) ; Suggested the corrected one.

$$R^2_N \overset{def}{=} \frac{R^2_{LR}}{1 - exp\left\{\frac{2}{n} l(\boldsymbol{y}, \hat{\mu}(\mathbf{1}_n))\right\}} = \frac{R^2_{LR}}{max(R^2_{LR})} \in [0, 1]$$

But still inconsistent with the classical definition of cofficient of determination.

Cameron and Windmeijer(1997) ; Use the KL divergence to quantify the uncertainty remaining in the response after accounting for predictors.

Specified Likelihood

$\hat{\theta}_{\boldsymbol{X}} \overset{def}{=} \underset{\theta}{argmax} L(\theta|\boldsymbol{X})$; MLE given data $\boldsymbol{X}$.

$\hat{\theta}_{I_n} \overset{def}{=} \underset{\theta}{argmax} L(\theta|I_n)$; MLE under saturated model.

Saturated model example ; $E(Y) = \boldsymbol{\beta}_{n\times 1}$

$$\hat{KL}(\theta, \boldsymbol{X}) \overset{def}{=} 2\{l(\hat{\theta}_{I_n}|I_n) - l(\hat{\theta}_{\boldsymbol{X}}|\boldsymbol{X})\}$$

$$R^2_{KL} \overset{def}{=} 1 - \frac{\hat{KL}(\theta, \boldsymbol{X})}{\hat{KL}(\theta, \mathbf{1}_n)}$$

Can be interpreted as **the deviance reduction ratio due to the changes of predictors** in $\boldsymbol{X}$.

Limits

All the aforementioned generalized coefficients of determination are given through **completely specified likelihood function**.

However, these are not applicable for more general GLMs like quasi-models which specify only the **mean and variance functions**.

GLM review

Random Component

1. Assume $(y_i|X_i) \overset{iid}{\sim}$ a member of exponential family
2. Model $\mu_i = E(y_i|X_i)$
3. Link function g ; $g(\mu_i) = X_i\boldsymbol{\beta}$ where X_i is an i^{th}row of data matrix.

Exponential family

Any random variable y in the exponential family has a probability density function of the form,

$$f(y, \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} - c(y, \phi) \right\}$$

with loglikelihood,

$$\ell(\theta, y, \phi) = \log\{f(y, \theta, \phi)\} = \frac{y\theta - b(\theta)}{\phi} - c(y, \phi)$$

θ ; The canonical parameter of interest
ϕ : A dispersion parameter which plays a role in the variance
We use the Bartlett's Identities to derive a general expression for the variance function.

Bartlett's Identities

The first and second Bartlett results ensures that under suitable conditions (see Leibniz integral rule), for a density function dependent on θ, $f_\theta(\cdot)$,

$$E_\theta\left[\frac{\partial}{\partial\theta}\log(f_\theta(y))\right] = 0$$

$$Var_\theta\left[\frac{\partial}{\partial\theta}\log(f_\theta(y))\right] + \mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log(f_\theta(y))\right] = 0$$

Bartlett's Identities

Expected value of Y : Taking the first derivative with respect to θ of the log of the density in the exponential family form described above, we have

$$\frac{\partial}{\partial\theta}\log(f(y,\theta,\phi)) = \frac{\partial}{\partial\theta}\left[\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right] = \frac{y - b'(\theta)}{\phi}$$

. Then taking the expected value and setting it equal to zero leads to,

$$E_\theta\left[\frac{y - b'(\theta)}{\phi}\right] = \frac{\mathrm{E}_\theta[y] - b'(\theta)}{\phi} = 0$$

$$E_\theta[y] = b'(\theta)$$

Bartlett's Identities

Variance of Y: To compute the variance we use the second Bartlett identity,

$$\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right)\right] + \mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right)\right]$$

is 0. Then,

$$\mathrm{Var}_\theta\left[\frac{y - b'(\theta)}{\phi}\right] + \mathrm{E}_\theta\left[\frac{-b''(\theta)}{\phi}\right] = 0, \quad \mathrm{Var}_\theta[y] = b''(\theta)\phi$$

Bartlett's Identities

We have now a relationship between μ and θ, namely $\mu = b'(\theta)$ and $\theta = b'^{-1}(\mu)$, which allows for a relationship between μ and the variance,

$$V(\theta) = b''(\theta) = \text{the part of the variance that depends on } \theta$$

$$\mathrm{V}(\mu) = b''(b'^{-1}(\mu))$$

or,

$$V(\mu) = (b'' \circ b'^{-1})(\mu)$$

Variance Function - Bernoulli case

For example, let $y \sim Bernoulli(p)$ then we express the density of the Bernoulli distribution in exponential family form,

$$f(y) = \exp\left(y \ln \frac{p}{1-p} + \ln(1-p)\right)$$

$$\theta = \ln \frac{p}{1-p} = logit(p),$$

which gives us $p = \dfrac{e^\theta}{1+e^\theta}$, $\quad b(\theta) = \ln(1+e^\theta)$ and

$$b'(\theta) = \frac{e^\theta}{1+e^\theta} = p = \mu \ , \quad b''(\theta) = \frac{e^\theta}{1+e^\theta} - \left(\frac{e^\theta}{1+e^\theta}\right)^2$$

Variance Function - Bernoulli case

This give us

$$V(\mu) = \mu(1-\mu) = \mu - \mu^2.$$

In this case, dispersion parameter $\phi = 1$.

Measuring Variation Changes Along the Variance Function

ϕ ; dispersion parameter
$V(\cdot)$; known variance function in GLMs
Consider a simpler measure of uncertainty, var, the variation.

$$var(y_i|X_i) \stackrel{def}{=} \phi V\{\mu(X_i)\} = \phi b''\{b'^{-1}(\mu(X_i))\}$$

where $\mu(X_i) \stackrel{def}{=} E(y_i|X_i)$. 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 distributions.

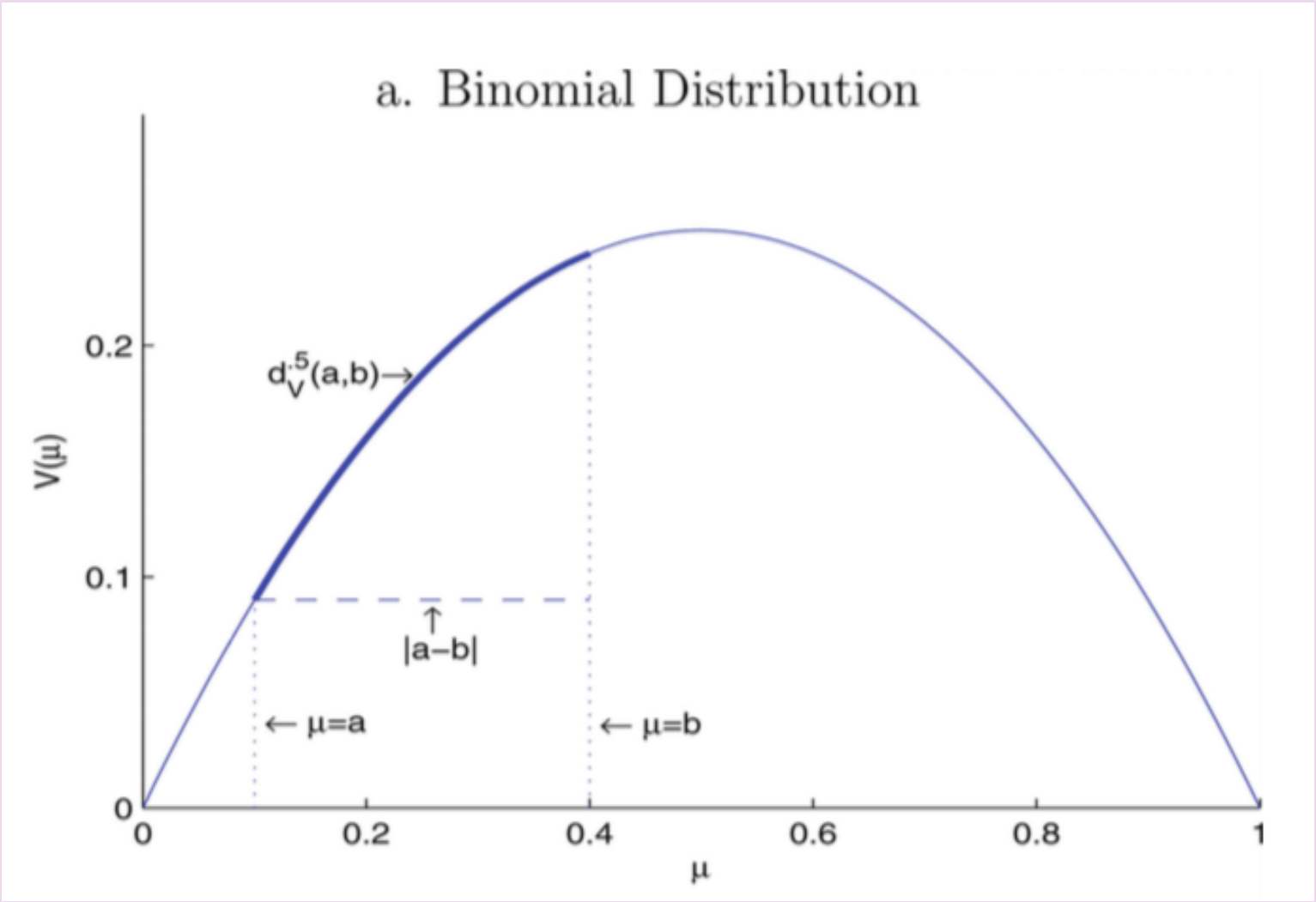
Measuring Variation Changes Along the Variance Function

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 of the response variable should be measured using, instead of $(a-b)^2$,

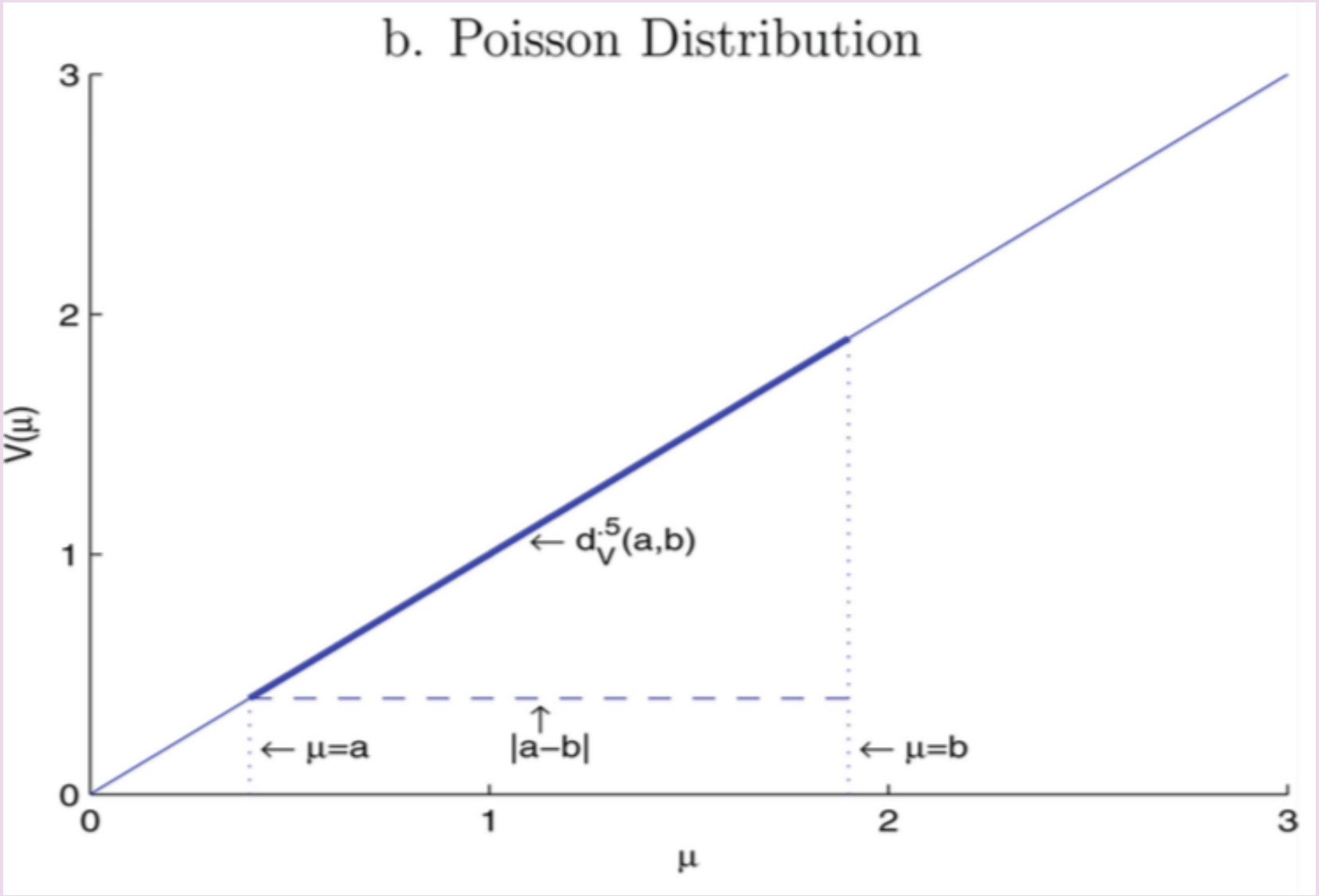
$$d_V(a,b) = \left[\int_a^b \sqrt{1+\{V'(t)\}^2}dt\right]^2$$

which is an squared length of $V(\cdot)$ between $V(a)$ and $V(b)$.

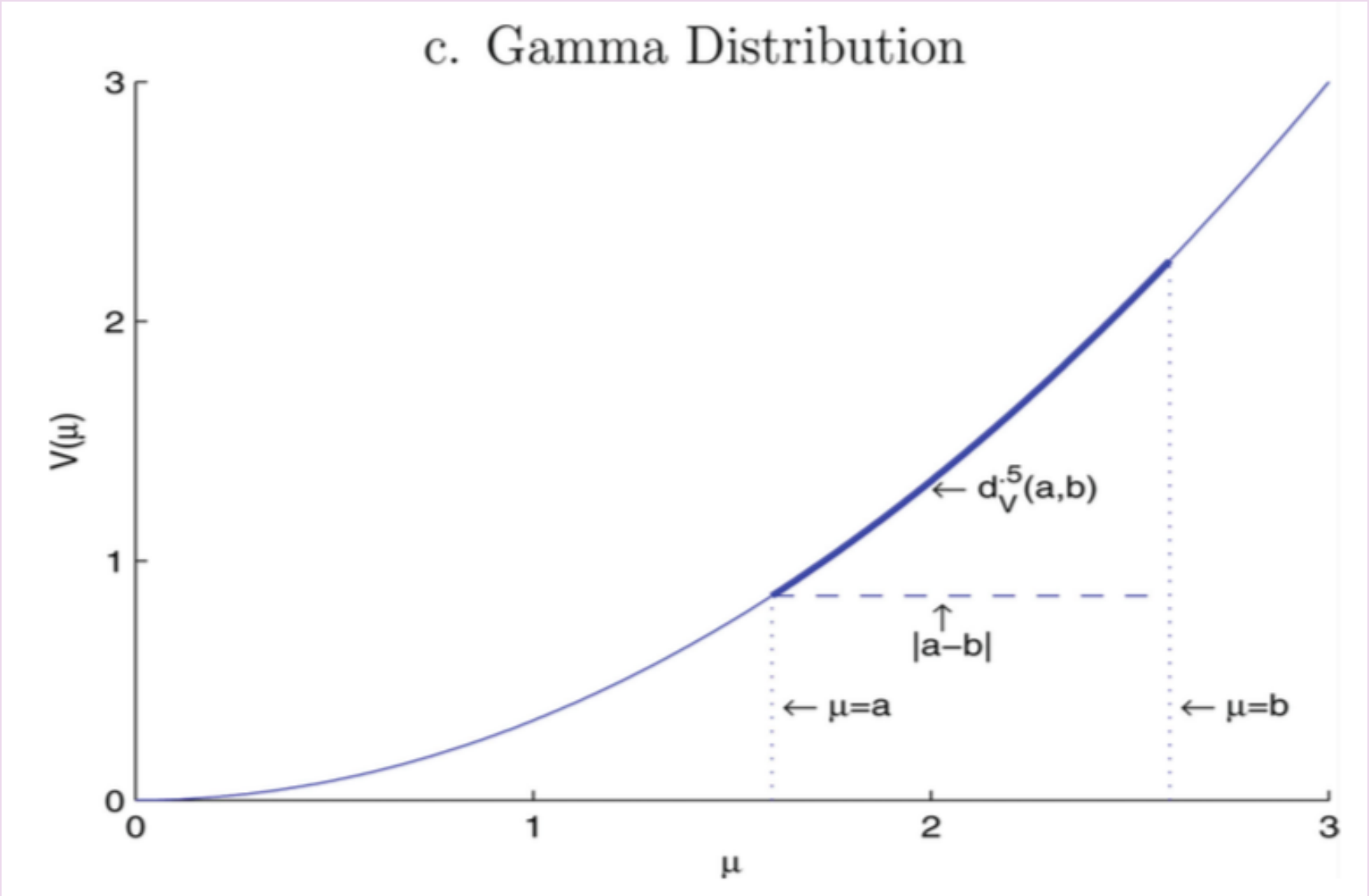
Variation change along the variance function



Variation change along the variance function



Variation change along the variance function



Variation

$d_V(a, b)$ 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 ;

$$V(\mu) = v_2\mu^2 + v_1\mu + v_0 \ , \ v_2 \neq 0$$

Variation

$$\int_a^b \sqrt{1+V'(t)^2}dt = \int_a^b \sqrt{1+(2v_2t+v_1)^2}dt$$

$$= \frac{V'(t)\sqrt{1+V'(t)^2}+\sinh^{-1}(V'(t))}{4v_2}\Big|_a^b = \sqrt{d_V(a,b)}$$

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 $d_V(a,b) = (1+v_1^2)(b-a)^2$.

R_V^2

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

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

Appropriate when only mean and variance functions can be specified(like the quasi-models).
Therefore, $\hat{\mu}(\boldsymbol{X})$ and $\hat{\mu}(\mathbf{1}_n)$ may be derived from quasi-likelihood estimators, other than MLE.

Consistency of R_V^2 and its Extension

$V'(\cdot)$ is **constant** for normal and Poisson distributions.
That is, $R_V^2 = R^2$ for OLS with normal distribution and log-linear model with Poisson distribtution.

Similar to the coefficient of determination, the coefficient of **partial determination** is well-defined for linear models.

Measures 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.

Partial Determination

For example, considering two sets of predictors X_1 and X_2 in a linear regression model, we have

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

($\because$) Recall that $1 - R^2(X_1) = \frac{SSE(X_1)}{SST}$.

measuring the proportion of remaining variation in the response, when including X_1, explained by X_2 .

Partial Determination

With our definition of R^2_V , we can easily extend it to a coefficient of partial determination for more general models.

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

Adjusted R^2_V

R^2_V also suffers to increasing numbers of predictors as the classical R^2.
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.

$$R^2_{V,adj} \stackrel{def}{=} 1 - \frac{\sum_{i=1}^{n} d_V\{y_i, \hat{\mu}_i(\boldsymbol{X})\}/(n-p)}{\sum_{i=1}^{n} d_V\{y_i, \hat{\mu}_i(\mathbf{1}_n)\}/(n-1)}.$$

Empirical Comparison

1. Random Sampling for 100 times from 2 Populations
2. from a member of Exponential family $f(\cdot)$ with means $\mu_1,\ \mu_2$ respectively
3. 50 random samples for each sampling ;
   $x_{1,1}, \cdots, x_{1,25} \overset{iid}{\sim} f(\ \cdot \mid \mu_1), \quad x_{2,1}, \cdots, x_{2,25} \overset{iid}{\sim} f(\ \cdot \mid \mu_2)$

X_1 ; The population of the corresponding observation
X_2 ; From the standard normal distribution, independent of X_1 and the response variable

I'm gonna draw a sketch for this data structure.

Binomial Model

Let

$$\mu_1 = \frac{e^{-\beta}}{1+e^{-\beta}}, \quad \mu_2 = \frac{e^{\beta}}{1+e^{\beta}} \ .$$

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. Here, $\mu_1 + \mu_2 = 1$.

The coefficients of determination are averaged over the 100 datasets for each β ranging from 0 to 5 with step=0.1.

Binomial Model

When $\beta = 0$, we have $\mu_1 = \mu_2$.

Thus corresponding $R^2_\cdot$'s are supposed to report 0 (or very close to 0).

On the other hand, $\mu_1 \to 0,\ \mu_2 \to 1$ as $\beta \to \infty$. Which leads to the single-population based samples.
Therefore, it is not surprising to observe that, when including the true predictor X_1, R^2_N, R^2_{KL}, and R^2_V approach 1.

See Fig.2.

Candidates

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

$$R^2_N = \frac{R^2_{LR}}{max(R^2_{LR})}$$

$$R^2_{KL} = 1 - \frac{\hat{KL}(\beta, \boldsymbol{X})}{\hat{KL}(\beta, \mathbf{1}_n)}$$

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

Poisson & Gamma Model

Poisson Model ; $\mu(X_1) = e^{X_1\beta}$

Gamma Model ; $\mu(X_1) = \frac{100}{2+X_1\beta}$, shape par.=100

The coefficients of determination are averaged over the 100 datasets for each β ranging from 0 to 5 with step=0.1.

See Fig.3, 4.

Poisson & Gamma Model

When $\beta \to \infty$, $|\mu_1 - \mu_2|$ reaches the maximum in poisson model but is bounded by 50 in the gamma model.

Therefore, it is not surprising to observe that, when including the true predictor X_1,R^2_{KL} and R^2_V approach one in the Poisson model, but are barely bounded away from one in the gamma model.

Why R_N^2, R_{LR}^2 falsely claim high R^2?

We consider the case $\beta = 5$.
Denote X_{21} as the subset of X_2 with $E(X_1) = \mu_1$ and
X_{22} as the subset of X_2 with $E(X_1) = \mu_2$.

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.
Plot of calculated R^2 versus $\bar{X}_{21} - \bar{X}_{22}$ in fig.5.
It shows the robustness of different coefficients of determination when only a falsely correlated X_2 is included.

Why R^2_N, R^2_{LR} falsely claim high R^2?

R^2_{LR} and R^2_N ; **Have 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.

X_1 vs (X_1, X_2)

$R^2(X_1, X_2)$ is similar to $R^2(X_1)$
. Though the former models always have slightly higher values.

We defined $R^2_{V,adj}$ to adjust for increasing numbers of predictors. Since the K-L 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;

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

X_1 vs (X_1, X_2)

As shown in Figure 6, both $R^2_{V,adj}$ and $R^2_{KL,adj}$ 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.

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).

(C ; colors), (SC ; spine conditions), (CW ; carapace widths), (W ; weights) 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.

Real Data Analysis

We consider binomial models ;

$$\boldsymbol{y} = Whether\ a\ female\ crab\ had\ any\ satellites$$

and Poisson models ;

$$\boldsymbol{y} = \#\ of\ satellites\ a\ female\ crab\ had$$

As demonstrated in the previous section, both R^2_{LR} and R^2_N indeed severely overstate the variation proportion explained by the Poisson models.

Real Data Analysis

Quasi-binomial models and quasi-Poisson models can be fitted for the above cases, with R_V^2 and partial R_V^2 calculated.

$\hat{\phi} = 1.0266$ for the binomial full model, and $\hat{\phi} = 3.2354$ for the Poisson full model.
Use quasi-Poisson models to allow overdispersion(that is, use of likelihood-based R^2 is inappropriate)!

For Poisson or quasi-Poisson models we have same regression coefficients. This leads to the same R_V^2 and partial R_V^2.

Conclusions

R^2 ; Measures goodness of fit , also provides a measure of predictability.
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 *adj.* versions 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 $R^2_{V,adj}$ makes all these possible when any statistical model with a well-defined variance function, such as quasi model.