HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 11/25

# Measuring overfitting and mispecification in nonlinear models

Marcel Bilger
Willard G. Manning

August 2011

york.ac.uk/res/herc/hedgwp

# Measuring overfitting and mispecification in nonlinear models*

Marcel Bilger, Willard G. Manning

*The Harris School of Public Policy Studies, University of Chicago, USA*

**Abstract**

We start by proposing a new measure of overfitting expressed on the untransformed scale of the dependent variable, which is generally the scale of interest to the analyst. We then show that with nonlinear models shrinkage due to overfitting gets confounded by shrinkage—or expansion—arising from model misspecification. Out-of-sample predictive calibration can in fact be expressed as in-sample calibration times 1 minus this new measure of overfitting. We finally argue that re-calibration should be performed on the scale of interest and provide both a simulation study and a real-data illustration based on health care expenditure data.

*JEL classification:* C21, C52, C53, I11

*Keywords:* overfitting, shrinkage, misspecification, forecasting, health care expenditure

## 1. Introduction

When fitting a model, it is well-known that we pick up part of the idiosyncratic characteristics of the data as well as the systematic relationship between a dependent and explanatory variables. This phenomenon is known as *overfitting* and generally occurs when a model is excessively complex relative to the amount of data available. Overfitting is a major threat to regression analysis in terms of both inference and prediction. When models greatly over-explain the data at hand they can even show relations which reflect chance only. Consequently, overfitting casts doubt on the true statistical significance of the effects found by the analyst as well as the magnitude of the response. In

* We would like to thank the Swiss National Science Foundation for supporting Marcel Bilger's post doctoral studies at the University of Chicago. The revised paper has benefited from the comments of Randall Ellis, Boston University, when the paper was presented at the International Health Economics Association in Toronto in July 2011.

addition, since the relations found in the estimation sample will in general not replicate, the predictive performance of the models deteriorates when they are applied to new data. An important distinction has thus to be made between *retrospective* prediction, where a model predicts outcomes within the dataset used to estimate it, and *prospective* prediction or forecast, where a previously estimated model forecasts new outcomes in a different data set. We will employ here an alternative way of distinguishing between these two types of prediction by referring to the former as being *in-sample* and the latter *out-of-sample*.

When using parameter estimates from the estimation sample to make predictions in a new sample from the same population, the plot of the actual outcomes against their forecasts should lie on the 45 degree line in the absence of overfitting. This topic has been studied for quite a long time and the deviation from the 45 degree line is referred to as *shrinkage*. An early measure of shrinkage is the adjusted multiple correlation coefficient which was proposed by Wherry (1931). Shrinkage is also often measured by means of cross-validation techniques following the pioneering work by Larson (1931). Our work is based on the seminal work by Copas (1983) who proposed to measure shrinkage by 1 minus the least squares slope of the regression of the observed outcomes on their out-of-sample predictions. Later, Copas (1987) also suggested to estimate this slope by cross-validation and his measure gained further popularity. Many applied researchers use this measure, especially in health sciences (see for instance Harrell et al., 1996; Blough et al., 1999; Harrell, 2001) where it is sometimes referred to as the Copas test of overfitting. The Copas measure of shrinkage is often assessed for various competing models and the results displayed in league tables. The resulting ranking is eventually part of model selection along with other diagnostic tools (e.g. Basu et al., 2006). It should however be stressed that predictive shrinkage is not only relevant in health economics, but in any field of economics where well-calibrated predictions are of primary importance. Government budgeting in public economics and risk adjustment in the economics of risk and uncertainty are two such examples.

In this paper, we revisit the Copas measure of shrinkage in the case of nonlinear models. With nonlinear models, estimation often takes place on a different scale from the one of the dependent variable. The former is sometimes referred to as *scale of estimation*, and the latter as the *scale of interest*, to the extent that the original scale is the scale of scientific or policy interest. Shrinkage is usually measured on the scale of estimation (see for instance Copas, 1983, 1997, in the specific case of the logistic regression). However, a few authors (Veazie et al., 2003; Basu et al., 2006) found it more meaningful to assess the Copas slope on the scale of interest as shrinkage is then measured in the same unit as the dependent variable. We show here that this alternative Copas measure does not constitute a measure of shrinkage arising from only overfitting. This measure also generally picks up the effect of model misspecification on the scale of estimation. We argue that the scale-of-interest Copas measure should instead be viewed as a measure of shrinkage arising from both model misspecification and overfitting. It should be stressed that we still consider this version of the Copas measure to be extremely valuable as it provides a measure of *calibration* of the out-of-sample predictions. Calibration refers here to the extent of bias in the predictions, which is an important component of predictive accuracy (see van Houwelingen and Cessie, 1990, for a discussion). More importantly, while the scale of estimation is often adopted for analytical reasons, the scale of interest, or raw scale, is essential for assessing the implications of policies and behaviors.

We propose to express calibration of the out-of-sample predictions as their in-sample calibration multiplied by 1 minus a new measure of overfitting defined on the scale of interest. This relation makes it possible to measure the respective contributions of model misspecification and overfitting to the predictive bias when applying an estimated model to new data. It notably illustrates the trade-off that the analyst faces when comparing—and selecting—competing models. Should flexibility be increased in order to achieve better in-sample performance? Or should nonlinearity be reduced and some secondary covariates be left out in order to better contain overfitting? Our expression indicates on what side of this trade-off a given model lies, thus providing the analyst with guidance on optimal specification choice.

Given that out-of-sample predictive accuracy depends on both in-sample calibration and overfitting is also crucial when correcting the predictions to account for shrinkage. Copas (1983) proposed a pre-shrunk predictor which is obtained by multiplying the predictions by the estimated Copas measure of shrinkage. However, the problem is that since the Copas measure of shrinkage is defined on the estimation scale, the corresponding pre-shrunk predictor re-calibrates for shrinkage arising from overfitting only. Predictive bias caused by model misspecification will replicate in any new dataset and this is not accounted for. That is why we argue that re-calibration should take place on the scale of interest in order to correct for model misspecification bias as well.

We first illustrate our new measure of overfitting and its relation with in-sample and out-of-sample calibration by means of a simulation study. We also present a model explaining data on individual healthcare expenditures which is a relevant case study as such models are typically highly nonlinear given their strictly positive and right-skewed dependent variable. Our simulations mimic real data and are in particular based on a realistic set of covariates. We employ a baseline GLM model and misspecification is introduced through either an incorrect link function or wrong distributional assumption. This allows us to assess the effect of model misspecification on the various measures of shrinkage presented in this paper. We finally present a real-data illustration by fitting our baseline model on health care expenditure data from a well-known hospitatlist study (Meltzer et al., 2002) in order to show that the distinctions made in this paper matter in practice.

The rest of the paper is organized as follows. Section 2 presents our new measure of overfitting, its relation with out-of-sample and in-sample calibration, and its estimation method. Section 3 presents the simulation design and Section 4 the results from these simulations. Section 5 show our illustration and Section 6 concludes.

## 2. Theoretical part

### 2.1. *The Copas measure of shrinkage*

For ease of exposition, we start by restricting the nonlinear models analyzed to the members of the GLM family (Nelder and Wedderburn, 1972). This family is fairly general as it includes many models widely applied in practice, such as the linear model for untransformed continuous variables, Poisson regression for counts, logistic regression for binary variables, and parametric proportional hazard models for durations. An important exception is the Box-Cox models, which do not belong to the GLM family and notably

include the linear regression of the log-transformed dependent variable.

In the GLM family, the distribution of the observed outcomes, $y_{i,i=1,..,n}$, is assumed to be a member of the exponential family where the expectation is related to the linear predictor, $\eta_i = \beta_0 + \boldsymbol{\beta_p'}\mathbf{x_i} \equiv \boldsymbol{\beta'}\mathbf{z_i}$, via the link function $g_T$: $E(y_i|\mathbf{x_i}) = g_T^{-1}(\eta_i) \equiv \mu_i$. In addition, the variance of $y_i$ is supposed to be function of its expectation, i.e. $V(y_i|\mathbf{x_i}) = v_T(\mu_i)$. Note that $\mathbf{x_i}$ refers to a vector of $p$ covariates, $\boldsymbol{\beta_p}$ to the vector of their corresponding parameters, $\beta_0$ to a constant term, and subscript $T$ to the true model followed by $y_i$. Note also that $\mathbf{z_i'} = (1 \quad \mathbf{x_i'})$ and $\boldsymbol{\beta'} = (\beta_0 \quad \boldsymbol{\beta_p'})$.

In his seminal paper, Copas (1983) proposed a very convenient measure of the shrinkage caused by overfitting in the case of the linear regression with multivariate normal covariates. This measure is based on the fact that the conditional expectation of a new outcome, $E(y_i^*|\hat{y}_i^{\text{out}})$, can be expressed as a linear function of its out-of-sample prediction, $\hat{y}_i^{\text{out}}$. Because these two quantities should be equal in the absence of overfitting, 1 minus the least squares slope of this regression provides a measure thereof. Copas also argues that this method can be generalized to the entire GLM family, and illustrates this in the specific case of the logistic regression Copas (1983, 1997). In this paper, we formally show that Copas' intuition is correct by deriving a large sample approximation of the Copas shrinkage factor in the GLM framework. To do so, let us first express the conditional expectation of a new outcome on the scale of estimation, $g_T(E(y_i^*|\hat{\boldsymbol{\beta}}, \mathbf{z_i}))$, as a linear function of its out-of-sample linear predictor, $\hat{\boldsymbol{\beta}}'\mathbf{z_i}$:

$$g_T\left(E(y_i^*|\hat{\boldsymbol{\beta}}, \mathbf{z_i})\right) \simeq \hat{\beta}_0 + \Delta\left(\hat{\boldsymbol{\beta}}'\mathbf{z_i} - \hat{\beta}_0\right), \tag{1}$$

where the covariates $\mathbf{x_i}$ have been appropriately centered. Note that this centering does not lead to any loss of generality as it merely redefines the constant term of the model, $\beta_0$. Equation 1 can be viewed as being the best linear approximation of $g_T(E(y_i^*|\hat{\boldsymbol{\beta}}, \mathbf{z_i}))$ in the maximum likelihood sense. In the absence of overfitting, $\hat{\boldsymbol{\beta}}'\mathbf{z_i}$ is a well-calibrated predictor of $g_T(E(y_i^*|\hat{\boldsymbol{\beta}}, \mathbf{z_i}))$ and $\Delta$ equals 1. On the other hand, the further $\Delta$ is below 1, the greater is the shrinkage resulting from overfitting. The quantity 1-$\Delta$ can thus be interpreted as a measure of the shrinkage caused by overfitting. We show in Appendix A that, asymptotically, $\Delta$ can be expressed as follows:

$$\Delta \simeq 1 - \frac{p}{\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}' I \hat{\boldsymbol{\beta}}_{\boldsymbol{p}}}, \tag{2}$$

where $I$ is the Fisher information matrix. This confirms what Copas (1997) found for the logistic regression, namely that shrinkage increases with the number of covariates $p$, and decreases with the GLM deviance $\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}' I \hat{\boldsymbol{\beta}}_{\boldsymbol{p}}$. So, the better the in-sample fit, the smaller the shrinkage, which completely disappears at the limit. On the other hand, poor in-sample fit results in large out-of-sample shrinkage. Note that when the deviance is less than $p$, $\Delta$ becomes negative. It may finally be stressed that $\Delta$ is defined on the transformed scale of the expectation of the dependent variable. We will thus refer to $1 - \Delta$ as the estimation-scale Copas shrinkage throughout.

Although the vast majority of studies measure shrinkage on the estimation scale (see for instance Blough et al., 1999), a few authors (Veazie et al., 2003; Basu et al., 2006) have found it more meaningful to measure it on the untransformed scale or the scale of interest. This was originally seen as a two degree of freedom measure, but for better comparability with the estimation-scale version, and without any loss of generality, we

here assume that the observations $y_i$ are centered. We formalize this measure as follows:

$$E(y_i^*|\hat{y}_i^{\text{out}}) \simeq \delta \hat{y}_i^{\text{out}}, \tag{3}$$

where $\hat{y}_i^{\text{out}} = g^{-1}(\hat{\boldsymbol{\beta}}'\mathbf{z_i})$ is the out-of-sample prediction of $y_i$, and equation 3 the best linear approximation of $E(y_i^*|\hat{y}_i^{\text{out}})$ in the least squares sense. Absence of overfitting results in $\delta = 1$, whereas $\delta$ is smaller than 1 when shrinkage occurs. Consequently, quantity 1-$\delta$ constitutes an alternative to $1 - \Delta$ for measuring the shrinkage in out-of-sample predictions. The particularity of $\delta$ is that it is expressed on the same scale as the variable of interest $y$. For instance, if $y$ represents individual healthcare expenditure, $\delta = 0.95$ would mean that the deviations above (below) average healthcare expenditure are underestimated (overestimated) by 5%. We refer to $1-\delta$ as the scale-of-interest Copas shrinkage throughout.

### 2.2. *Overfitting and model misspecification*

In practice, models act as approximations of the data generating process of the data at hand. There will always be some degree of departure from these models. To illustrate the consequences of this, let us consider the situation where a wrong link function $g_W$ is used for the model of $y_i$. Note that subscript $W$ refers to a misspecified–or wrong–model throughout.

Remarkably, the estimation-scale Copas statistic is not affected by this wrong assumption. This can be shown in equation 1 by replacing $g_T$ with $g_W$, and rearranging:

$$\Delta = \frac{g_W\left(E(y_i^*|\hat{\boldsymbol{\beta}}, \mathbf{z_i})\right) - \hat{\beta}_0}{\hat{\boldsymbol{\beta}}'\mathbf{z_i} - \hat{\beta}_0} \tag{4}$$

This means that, in the absence of overfitting, or in other words when the transformed conditional expectation of $y_i^*$, $g_W\left(E(y_i^*|\hat{\boldsymbol{\beta}}, \mathbf{z_i})\right)$, equals its out-of-sample linear predictor $\hat{\boldsymbol{\beta}}'\mathbf{z_i}$, the estimation-scale Copas statistic still equals 1 despite the misspecification of the model. The reason is that the misspecification has been made twice. It has first been made when computing the linear predictor, $\hat{\boldsymbol{\beta}}'\mathbf{z_i} = g_W(\hat{y}_i^{\text{out}})$, and then when assessing the estimation-scale Copas statistic defined in equation 1.

On the other hand, the scale-of-interest Copas statistic is affected by model misspecification. The reason is that, in the absence of overfitting and when the link function is nonlinear, $\hat{y}_i^{\text{out}}$ is unlikely to equal $E(y_i^*|\hat{y}_i^{\text{out}})$ when the model is misspecified. Indeed, $E(y_i^*|\hat{y}_i^{\text{out}})$ is the expectation of the outcome, $\mu_i$, which will most often not equal $g_W^{-1}(\hat{\boldsymbol{\beta}}'\mathbf{z_i})$ when using the wrong link function $g_W$. So, in equation 3, the slope $\delta$ generally does not equal 1 even in the absence of overfitting. The shrinkage (when $\delta < 1$) or expansion (when $\delta > 1$) results from model misspecification. In practice, some degree of overfitting is always present and this gets confounded with this misspecification effect. The problem lies in the discrepancy between the misspecified model used to predict the outcome and the scale-of-interest Copas statistic defined in equation 3 when this misspecification is ignored.

Note that when a linear model with an untransformed dependent variable is estimated in the first stage, this discrepancy does not exist and $1 - \delta$ can be used as a measure of overfitting. However, in the case of nonlinear models, quantity $1 - \delta$ should not be

viewed as a measure of overfitting alone, but as a measure of shrinkage resulting from both overfitting and model misspecification. We argue that statistic $\delta$ remains extremely meaningful, as it can be interpreted as a broader measure of calibration of the out-of-sample predictions, sensitive to both overfitting and model misspecification.

### 2.3. *A scale-of-interest measure of overfitting*

Our objective here is to propose a new measure of overfitting combining advantages from both the scale-of-interest and estimation-scale Copas statistics. We want this new statistic not only to be expressed on the scale of interest but also to be immune to model misspecification. Let us start by defining statistic $\alpha$ which is the slope of the linear regression of the conditional expectation of the outcome, $E(y_i|\hat{y}_i^{\text{in}})$, on its in-sample prediction, $\hat{y}_i^{\text{in}}$:

$$E(y_i|\hat{y}_i^{\text{in}}) \simeq \alpha \hat{y}_i^{\text{in}}, \tag{5}$$

where, similarly to equation 3, This equation is the best approximation of $E(y_i|\hat{y}_i^{\text{in}})$ in the least squares sense, and the observations $y_i$ are assumed to be centered. When using in-sample predictions, no overfitting can occur since the same vector of covariates is used both when estimating the model and predicting the outcome. When the model is well-specified, $\alpha$ equals 1, whereas we expect that $\alpha$ will most often not equal 1 when the model is misspecified. Note that $\alpha$ can equal 1 even when the model is misspecified, and should thus not be interpreted as a measure of misspecification *per se*. We merely view here $\alpha$ as being a measure of calibration of the in-sample predictions, and quantity $1-\alpha$ as a measure of the shrinkage arising from model misspecification alone. It may be stressed that, unlike $\delta$, $\alpha$ can be greater than one in which case it indicates predictive expansion. that is, over-prediction of large outcomes and under-prediction of small ones.

Interestingly, equations 3 and 5 implicitly define a new measure of overfitting. To see this, it should be stressed that the expectation of $y_i$ is always assumed to be same, both within and outside the sample: $E(y_i^*|\hat{y}_i^{\text{out}}) = E(y_i|\hat{y}_i^{\text{in}}) = \mu_i$. In other words, we assume that the data does not change its probability distribution from one sample to the next. Once $\mu_i$ is substituted for these two expectations in equations 3 and 5, it can easily be shown that:

$$\hat{y}_i^{\text{in}} \simeq \gamma \hat{y}_i^{\text{out}}, \tag{6}$$

where $\gamma = \frac{\delta}{\alpha}$. It is important to note that the same model, possibly misspecified, is used to predict both $\hat{y}_i^{\text{out}}$ and $\hat{y}_i^{\text{in}}$. Any deviation between these 2 quantities can thus only be caused by overfitting. We thus interpret quantity $1-\gamma$ as a measure of the shrinkage caused by overfitting alone. In the absence of overfitting, $\gamma$ equals 1 and no shrinkage arises. When overfitting occurs, the out-of-sample predictions lose their relation with the outcome, $\gamma$ diminishes and the measured shrinkage increases.

Table 1 sums up the main features of the four measures of shrinkage defined above. The first column shows the symbol used to refer to the shrinkage statistic while the second indicates whether the statistic is defined on the estimation scale or on the scale-of-interest. The last two columns respectively indicate whether the statistic captures misspecification, overfitting, or both.

Table 1
Shrinkage statistics

| Symbol | Scale | Effect captured | |
|---|---|---|---|
| | | misspecification | overfitting |
| $1-\Delta$ | estimation | | ✓ |
| $1-\delta$ | interest | ✓ | ✓ |
| $1-\alpha$ | interest | ✓ | |
| $1-\gamma$ | interest | | ✓ |

2.4. *Decomposing the overall shrinkage*

Expression 6 also provides a valuable relation between the 3 scale-of-interest statistics defined above. It shows that out-of-sample calibration $\delta$ equals in-sample calibration $\alpha$ times the overfitting statistic $\gamma$. Further insight can be gained by expressing this relation in terms of shrinkage:

$$1-\delta \simeq (1-\alpha)+\alpha\,(1-\gamma)\,. \tag{7}$$

The overall shrinkage when predicting outcomes in a new sample, $1-\delta$, is thus the sum of shrinkage due to misspecification, $1-\alpha$, shrinkage caused by overfitting, $1-\gamma$, times the in-sample calibration factor, $\alpha$. Note the interaction between shrinkage due to misspecification and overfitting. When the model is perfectly specified (i.e. $\alpha = 1$), our measure of overfitting equals the out-of-sample attenuation $1-\delta$. On the other hand lies the case where the model is so misspecified that $\alpha = 0$, and where overfitting does not play any role as this cannot further deteriorate the fit. It may also be noted that term $1-\alpha$ can be negative in which case it represents expansion caused by model misspecification. Term $1-\delta$ can also be negative when this expansion offsets shrinkage resulting from overfitting.

Equation 7 could be very useful in practice. It illustrates the trade-off that analysts face when comparing—and selecting—competing models, or when judging the adequacy of the model that they have selected. It is well-known that the extent of model flexibility has to balance in-sample quality of fit with containing the amount of overfitting. This is the spirit of Mallows' $C_p$ and the Akaïke information criterion which both measure the goodness of fit of a given model while penalizing for the number of covariates involved. However, model flexibility is not only function of the number of covariates but also of the nonlinearity of its functional form. The more nonlinear it is, the better its in-sample quality of fit, but the greater the potential overfitting as its greater flexibility is more likely to capture non-systematic variability, or noise. Expression 7 measures the extent of bias due to both misspecification and overfitting, thus indicating on what side of this trade-off a given model lies, and providing the analyst with guidance on optimal specification choice. It should be borne in mind, though, that unlike Mallows' $C_p$ and the Akaïke information criterion, quality of fit is not derived from the comprehensive mean square error, but assessed through the more restrictive predictive calibration. In particular, our expression does not account for quality of the predictive discrimination of the model (see for instance van Houwelingen and Cessie, 1990, for a useful discussion

on predictive value assessment). So, expression 7 should not be used as an alternative to—but complementary to—the standard goodness of fit in statistics. From an economics perspective, predictive calibration is often of primary interest per se. This is the case in any area on economics where getting the average prediction of a given population right is of primary importance. Government budgeting in public economics and risk adjustment in the economics of risk and uncertainty are two such examples.

Expression 7 also plays a crucial role when correcting the predictions to account for shrinkage. Copas (1983) proposed a preshrunk predictor which is obtained by multiplying the predictions by the estimated Copas measure of shrinkage. Since the Copas measure of shrinkage is defined on the estimation scale, the corresponding preshrunk predictor re-calibrates for shrinkage arising from overfitting only. However, expression 7 clearly shows that out-of-sample predictive accuracy depends both on in-sample calibration and overfitting. It follows that, when using the Copas preshrunk predictor, the bias caused by model misspecification will replicate in any new dataset since it is not accounted for. That is why we argue that re-calibration should take place on the scale of interest in order to correct for model misspecification bias as well. To be more specific, instead of multiplying the estimation-scale predictor $x_i'\hat{\beta}$ by the corresponding (estimated) shrinkage factor $\hat{\Delta}$, we suggest multiplying the predictions $\hat{y}_i^{\text{out}}$ by the scale-of-interest Copas shrinkage factor $\hat{\delta}$. Doing so yields a well-calibrated predictor of the outcomes when applying an estimated model to new data.

## 3. Simulation design

Our simulation framework aims at being close to real data. We illustrate the measures of overfitting presented in this paper with a model explaining individual healthcare expenditures. Such models are typically highly nonlinear given their strictly positive and right-skewed dependent variable. Our explanatory variables include an evenly split dummy variable, which can be thought of as being a variable indicating gender. We also include a 50%, 35%, and 15% split categorical variable, which approximately corresponds to the adults, children, and elderly age classes found in many countries. In addition, we include both a uniformly and a normally distributed variables to account for the variety of quantitative factors usually included in such models. We choose a sample size of 5,000, which falls in the range of most observational surveys available in practice. An important difference between our simulation framework and real data is that real models usually comprise many more explanatory factors, notably to account for socioeconomic characteristics and of a few rare but important health conditions. Since overfitting tends to increase with the number of covariates, this makes our simulation findings even more relevant to real applications. Finally, another simplification is that all our factors are uncorrelated, except for the indicators for the categorical variable with 50/35/15 split.

The first column of Table 2 shows the different models used to generate the dependent variable, $Y$. The second and third columns give the value of the constant term of the linear predictor, $\beta_0$, and of an ancillary parameter of the distribution considered. Note that these parameters have been computed numerically to ensure that $E(Y) = 1$ and $V(Y) = 2.2$ over all scenarios. Finally, the last two columns of Table 2 present the skewness and kurtosis of $Y$, which have also been determined numerically.

The baseline case is a GLM model with Gamma distribution and logarithmic link,

which is one of the most widely used models for healthcare expenditures (Blough et al., 1999). A whole series of scenarios is then produced with the Extended Estimating Equations (EEE, Basu and Rathouz, 2005), which generalizes the GLM framework, notably through a flexible Box-Cox link function whose parameter $\lambda$ will be estimated along with the parameters of the linear predictor. The EEE model provides us with the opportunity to progressively modify the link function while keeping the distribution of $Y$ unchanged. We explore values of $\lambda$ ranging from -0.75 to 1, where $\lambda = 0$ corresponds to the logarithmic link of the baseline model, and $\lambda = 1$ to the identity link. Another series of scenarios that we explore is produced with the Generalized Gamma model (GENGAM), which was applied by Manning et al. (2005) to account for deviations from the Gamma distribution and improve efficiency of healthcare expenditure models. The GENGAM family makes it possible to progressively modify the distribution of $Y$ while keeping the logarithmic link between the linear predictor and the expectation of $Y$. Note that the GENGAM family does not generalize the GLM family in the sense that most of its models are not member of the GLM family. These additional scenarios are produced by setting the shape parameter $\kappa$ at 0.5, 1, 2, and 3 (in the parametrization used by Manning et al., 2005). Note that the baseline scenario corresponds to a value of approximately 1.5 for $\kappa$.

Each scenario presented in Table 2 is generated 400 times. For each repetition, the explanatory variables are randomly generated first. In order to reduce the Monte Carlo variation in the simulation results, the same explanatory variables are used over all scenarios. At each iteration, we draw the binary and categorical variables so that to ensure exact 50-50 and 50-35-15 splits, which also contributes to containing the Monte Carlo variation. As for the quantitative variables, they are simply drawn from the standard uniform and normal distributions. The covariate matrix is then duplicated 10 times, and the dependent variable $Y$ is randomly drawn from one of the distributions presented in Table 2. Finally, the shrinkage factors are estimated using 10-fold cross-validation (CV, Geisser, 1975) where all groups have the same covariate matrix and only differ with respect to $Y$.

## 4. Simulation results

Table 3 shows the simulation results relative to the specification of the link function. The first column shows the data generating process sorted by increasing value of the Box-Cox parameter $\lambda$. The table is then subdivided into two parts according to which model is estimated: either the GLM with Gamma distribution and logarithmic link or the EEE constrained to the Gamma distribution but with unconstrained $\lambda$. For each estimated model, Table 3 presents the scale-of-interest Copas shrinkage, $1 - \hat{\delta}$, in-sample shrinkage, $1 - \hat{\alpha}$, and shrinkage due to overfitting alone, $1 - \hat{\gamma}$. Note that in this context, the EEE model, because of its ability to estimate the parameter $\lambda$, will thus be unbiased. On the other hand, the GLM model, which is restricted to the special case $\lambda = 0$, will be biased for any other value of $\lambda$.

It can first be seen that the scale-of-interest Copas shrinkage $(1 - \hat{\delta})$ obtained with the GLM and EEE models can be substantially different. For instance, for $\lambda = -0.75$, this measure shows 2.99% expansion for the GLM and 10.04% shrinkage for the EEE model. Because overfitting *alone* cannot lead to expansion, this clearly indicates that another

Table 2
Data generating processes

| Data Generating Process | Parameters[a] | | Higher Moments | |
|---|---|---|---|---|
| | $\beta_0$ | $\nu$ / $\sigma$[b] | skewness | kurtosis |
| **Generalized Linear Model** | | | | |
| Gamma distribution, log. link | -0.312 | 0.5 | 3.26 | 20.4 |
| **Extended Estimating Equations** (Gamma distribution) | | | | |
| $\lambda$ = -0.75 | -0.339 | 0.515 | 3.50 | 29.5 |
| $\lambda$ = -0.5 | -0.329 | 0.509 | 3.35 | 22.3 |
| $\lambda$ = -0.25 | -0.320 | 0.505 | 3.29 | 21.0 |
| $\lambda$ = 0.25 | -0.306 | 0.494 | 3.25 | 20.0 |
| $\lambda$ = 0.5 | -0.297 | 0.494 | 3.22 | 19.6 |
| $\lambda$ = 0.75 | -0.291 | 0.494 | 3.21 | 19.4 |
| $\lambda$ = 1 | -0.283 | 0.494 | 3.20 | 19.2 |
| **Generalized Gamma** | | | | |
| $\kappa$ = 0.5 | -0.723 | 1.26 | 4.86 | 53.6 |
| $\kappa$ = 1 | -0.516 | 1.38 | 3.78 | 28.6 |
| $\kappa$ = 2 | -0.020 | 1.38 | 2.79 | 14.6 |
| $\kappa$ = 3 | 0.398 | 1.21 | 2.32 | 9.9 |

[a] Parameters computed numerically so that $E(Y) = 1$ and $V(Y) = 2.2$.
[b] Either parameter $\nu$ for the Extended Estimating Equations or $\sigma$ for the Generalized Gamma.

factor is in play. This is confirmed by the measure of in-sample shrinkage which shows near-perfect calibration in the case of the EEE model (i.e. $1 - \hat{\alpha} \approx 0$) and substantial deviations for the GLM that span from 7.70% expansion for $\lambda = -0.75$ to 6.76% shrinkage for $\lambda = 1$. The simulation results thus clearly illustrate that the Copas measure of shrinkage, when applied to scale-of-interest predictions, cannot be considered to be a measure of overfitting. As discussed in Section 2, we propose measuring shrinkage due to overfitting alone by $1 - \hat{\gamma}$. As expected, the simpler GLM exhibits lower overfitting than the EEE model which requires the estimation of nonlinear parameter $\lambda$. For the highly nonlinear data where $\lambda = -0.75$, shrinkage due to overfitting in the EEE (9.75%) is more than twice as large as what is observed for the GLM (4.37%). When nonlinearity decreases, so does the difference in overfitting between the GLM and EEE models.

The simulation results illustrate well our decomposition of the out-of-sample shrinkage given by Equation 7. For the EEE model, or when the link function is well-specified, out-of-sample shrinkage equals shrinkage due to overfitting. On the other hand, for the mis-specified GLM model, in-sample shrinkage can add up to the shrinkage due to overfitting and further deteriorate out-of-sample calibration, which can be seen in our simulations

Table 3
Measure of shrinkage (percentage) and specification of the link function[a]

| **Data Generating Process** | **GLM**[b] | | | **EEE**[c] | | |
|---|---|---|---|---|---|---|
| | 1-$\hat{\delta}$ | 1-$\hat{\alpha}$ | 1-$\hat{\gamma}$ | 1-$\hat{\delta}$ | 1-$\hat{\alpha}$ | 1-$\hat{\gamma}$ |
| EEE, $\lambda = -0.75$ | -2.99 | -7.70 | 4.37 | 10.04 | 0.35 | 9.75 |
| | (0.26) | (0.26) | (0.06) | (0.67) | (0.13) | (0.66) |
| EEE, $\lambda = -0.5$ | 0.35 | -4.57 | 4.71 | 9.22 | 0.28 | 8.99 |
| | (0.26) | (0.26) | (0.07) | (0.49) | (0.12) | (0.48) |
| EEE, $\lambda = -0.25$ | 3.31 | -1.78 | 5.01 | 7.62 | 0.22 | 7.41 |
| | (0.23) | (0.22) | (0.07) | (0.33) | (0.10) | (0.32) |
| GLM, log. Link | 5.70 | 0.36 | 5.37 | 7.07 | 0.26 | 6.83 |
| | (0.21) | (0.21) | (0.07) | (0.15) | (0.08) | (0.13) |
| EEE, $\lambda = 0.25$ | 7.31 | 1.79 | 5.63 | 7.57 | 0.05 | 7.53 |
| | (0.24) | (0.23) | (0.08) | (0.37) | (0.09) | (0.36) |
| EEE, $\lambda = 0.5$ | 8.84 | 3.35 | 5.69 | 6.82 | -0.07 | 6.89 |
| | (0.23) | (0.22) | (0.08) | (0.21) | (0.10) | (0.19) |
| EEE, $\lambda = 0.75$ | 10.52 | 4.86 | 5.96 | 6.74 | -0.09 | 6.83 |
| | (0.23) | (0.22) | (0.08) | (0.21) | (0.10) | (0.19) |
| EEE, $\lambda = 1$ | 12.86 | 6.77 | 6.54 | 6.98 | 0.20 | 6.80 |
| | (0.30) | (0.27) | (0.13) | (0.27) | (0.18) | (0.20) |

[a] Average Monte Carlo shrinkage with standard error in parentheses.
[b] Generalized Linear Model with a Gamma distribution and a logarithmic link function.
[c] Extended Estimating Equations constrained to the Gamma distribution, whereas $\lambda$ is estimated.

when $\lambda > 0$. However, it can also be the case that model misspecification and overfitting work in the opposite direction, as shown in our simulations when $\lambda < 0$. That is why we argue that the scale-of-interest Copas shrinkage, $1 - \hat{\delta}$, even though it does not measure overfitting per se, remains a valuable measure when assessing out-of-sample predictive performance. For instance, for $\lambda = -0.5$, the out-of-sample predictions obtained with the GLM are well-calibrated ($1 - \hat{\delta} = 0.35$) as in-sample expansion and overfitting cancel each other. On the other hand, the well-specified EEE model shows considerable out-of-sample shrinkage ($1 - \hat{\delta} = 9.22$), which is driven by overfitting alone. The analysts who are primarily interested in forecasting are likely to prefer the latter model.

Efficiency has been shown to be an important concern in healthcare expenditure models (see for instance Manning and Mullahy, 2001). We take advantage of the fact that, with GLM models, efficiency is conditioned by the choice of the distribution. The simulation results presented in Table 4 illustrate the relationship between overfitting and the specification of the distribution. The first column shows the GENGAM models used to generate the data sorted by increasing value of the shape parameter $\kappa$. Similarly to Table

3, this table is then subdivided into two parts according to which model is estimated: either the GLM with Gamma distribution and logarithmic link or the GENGAM. For each estimated model, the table presents the scale-of-interest Copas shrinkage, $1-\hat{\delta}$, in-sample shrinkage, $1-\hat{\alpha}$, and shrinkage due to overfitting alone, $1-\hat{\gamma}$. Note that in this context, the GLM model, which is restricted to the special case $\kappa \simeq 1.5$ corresponding to the Gamma distribution in our example, will thus be misspecified for any other value of the shape parameter.

Table 4
Measure of shrinkage (percentage) with misspecified distribution[a]

| **Data Generating Process** | **GLM**[b] | | | **GENGAM**[c] | | |
|---|---|---|---|---|---|---|
| | 1-$\hat{\delta}$ | 1-$\hat{\alpha}$ | 1-$\hat{\gamma}$ | 1-$\hat{\delta}$ | 1-$\hat{\alpha}$ | 1-$\hat{\gamma}$ |
| GENGAM, $\kappa = 0.5$ | 5.50 | 0.28 | 5.23 | 4.60 | 0.20 | 4.41 |
| | (0.23) | (0.23) | (0.08) | (0.30) | (0.31) | (0.06) |
| GENGAM, $\kappa = 1$ | 5.56 | 0.28 | 5.29 | 5.36 | 0.24 | 5.13 |
| | (0.23) | (0.23) | (0.08) | (0.24) | (0.25) | (0.08) |
| GLM, Gamma distribution | 5.70 | 0.36 | 5.37 | 5.70 | 0.36 | 5.37 |
| | (0.21) | (0.21) | (0.07) | (0.21) | (0.21) | (0.07) |
| GENGAM, $\kappa = 2$ | 5.66 | 0.29 | 5.39 | 5.50 | 0.38 | 5.15 |
| | (0.23) | (0.23) | (0.08) | (0.26) | (0.25) | (0.08) |
| GENGAM, $\kappa = 3$ | 5.67 | 0.30 | 5.39 | 4.64 | 0.51 | 4.16 |
| | (0.23) | (0.23) | (0.08) | (0.34) | (0.34) | (0.06) |

[a] Average Monte Carlo shrinkage factor with standard error in parentheses.
[b] Generalized Gamma model, unconstrained.
[c] Generalized Linear Model with a Gamma distribution and a logarithmic link function.

As expected, the misspecification of the GLM model does not lead to in-sample bias as the measured in-sample shrinkage, $1-\hat{\alpha}$, is never significantly different from zero. Consequently, in the absence of in-sample misspecification bias, $1-\hat{\delta}$ equals $1-\hat{\gamma}$. The scale-of-interest Copas shrinkage, $1-\hat{\delta}$, is thus an adequate measure of the overfitting that occurs on that scale. This measure shows that the efficiency loss of the misspecified GLM model, even though it does not adversely affect in-sample calibration, leads to greater out-of-sample overfitting. For instance, for $\kappa = 3$, the misspecification of the distribution results in an increase of out-of-sample shrinkage by 1.23% compared to the GENGAM model. This is due to the double burden of inefficiency. Not only inefficient models will be less precise in-sample, but they will also have reduced out-of-sample predictive performance as this precision loss leads to greater overfitting. Again, measuring overfitting is very useful as this reveals here the out-of-sample shortcomings of the simpler GLM specification which appears to be unbiased when judged on in-sample grounds only. Conversely, this also shows that the efficiency gain of the GENGAM more than cancels out the greater overfitting induced by its greater complexity.

## 5. Illustration

To illustrate our measures of shrinkage in a real context, we use data from an hospitalist study which took place at the University of Chicago hospital (Meltzer et al., 2002) and which provides us with a fairly large data set of 6,500 observations. The outcome variable is patient-level health care expenditure excluding physician fees, and the key covariates relate to physician characteristics: whether the physician is an hospitalist or not, as well as disease-specific experience. Many control variables are also present such as patient commorbidities, relative utilization weight of diagnosis, admission month dummy variables, and an indicator for transfer from another institution. Note that in addition to being used to explain the lower cost of hospitalist care, this data set has also been used to illustrate the EEE (Basu and Rathouz, 2005) and GENGAM (Manning et al., 2005) models. Indeed, the marked right skewness of patient-level health expenditure makes a fruitful ground to illustrate the use of nonlinear models. In our example, we fit a log-Gamma GLM which has been widely applied in practice.

Table 5 shows our measures of shrinkage for this model. The first 4 columns show the shrinkage estimates obtained by repeated 10-fold CV, while the estimates presented in the last 4 columns have been obtained by repeated 2-fold CV (also known as the 2-way Copas test). The first two lines of the table present the estimates and standard errors [1] for the full sample (n = 6,500). Because issues of overfitting and misspecification may depend on sample size, the last two lines display the same information for a quarter of the sample (n = 1,625). To obtain a representative subsample, we have first randomly drawn 101 subsamples, then computed the scale-of-interest Copas shrinkage for each one of them, and finally picked the subsample with median value. [2] In what follows, we refer to the full and reduced samples as the large and small hospital ones respectively.

Let us start by interpreting the large hospital results obtained by repeated 10-fold CV. There is a striking difference between the raw scale Copas shrinkage, $1 - \hat{\delta}$, and our new measure of shrinkage arising from overfitting alone, $1 - \hat{\gamma}$: while the former shows significant shrinkage (16.11%), the latter reveals that overfitting plays a secondary role only (2.10%). The most important problem, by far, is the lack of fit within the sample, as shown by the in-sample measure of shrinkage ($1 - \hat{\alpha} = 14.32\%$). In the small hospital sample, shrinkage caused by overfitting considerably increases to 6.19%, but in-sample misspecficiation still remains the main issue. Note also that our raw scale measure of shrinkage is roughly fifty percent larger than with the estimation scale Copas, $1 - \hat{\Delta}$, for both hospital sizes. This clearly illustrates that the scale of analysis matters when assessing overfitting.

Table 5 also shows the relative gain from using 10-fold CV over 2-fold CV which is widely used in practice when measuring shrinkage. Since CV methods consist in holding out part of the data for validation when estimating the model, they yield an overly pessimistic estimate of its predictive accuracy. By holding out 50 percent of the sample,

[1] An unbiased measure of the standard error of CV estimates has yet to be found (Arlot and Celisse, 2009). What we report is the standard deviation of 400 repetitions of the k-fold CV estimates. It can easily be shown that this measure is an upper bound for the estimate averaged over all 400 repetitions that we report.

[2] We have performed our selection on the basis of the raw scale Copas measure of shrinkage in order to balance in- and out-of sample predictive performance.

Table 5
Shrinkage (percentage) in the log-Gamma GLM when applied to the Hospitalist data

| | **10-fold CV** | | | | **2-fold CV**[†] | | | |
|---|---|---|---|---|---|---|---|---|
| | *raw scale* | | | *est. scale* | *raw scale* | | | *est. scale* |
| *sample size n* | 1-$\hat{\delta}$ | 1-$\hat{\alpha}$ | 1-$\hat{\gamma}$ | 1-$\hat{\Delta}$ | 1-$\hat{\delta}$ | 1-$\hat{\alpha}$ | 1-$\hat{\gamma}$ | 1-$\hat{\Delta}$ |
| **6500** | 16.11 | 14.32 | 2.10 | 1.33 | 17.32 | 14.18 | 3.62 | 2.37 |
| | (0.65) | (0.05) | (0.79) | (0.14) | (2.77) | (1.30) | (4.00) | (0.69) |
| **1625**[‡] | 19.79 | 14.49 | 6.19 | 4.89 | 23.72 | 13.96 | 11.15 | 8.34 |
| | (1.92) | (0.14) | (2.34) | (0.51) | (7.39) | (3.34) | (9.76) | (2.21) |

Standard errors are displayed in brackets.
[†]Both 10- and 2-fold CV estimates have been averaged over 400 repetitions.
[‡]Subsample with median raw scale Copas shrinkage (among 101 random draws from the full sample).

2-fold CV is more prone to this bias than 10-fold CV which makes use of 90% of the data when estimating the model. What Table 5 clearly shows is that this bias also depends on sample size. In the large hospital sample, our measure of shrinkage caused by overfitting alone is inflated from 2.10% to 3.62%, whereas this quantity jumps from 6.19% to 11.15% in the small hospital case. Not only does the bias of the $k$-fold CV increase with $k$, but also does its efficiency. We can see that the standard errors reported for the out-of-sample shrinkage estimates are approximatively 4 times greater when using the 2-fold CV. The accuracy of the in-sample measures of shrinkage gets hit even harder with standard errors more that 25 times larger. 2-fold CV should thus be avoided, unless computational cost is an issue and sample size is large enough in the sense that holding out half the data does not excessively impact the estimation of the model.

## 6. Conclusion

In this paper, we propose a new measure of overfitting for nonlinear models, which is expressed on the untransformed scale of the dependent variable. This is typically the scale of interest in terms of assessing behaviours or policy analysis. We then show that, in the case of nonlinear models, shrinkage due to overfitting gets confounded by shrinkage—or expansion—arising from model misspecification. We also show that out-of-sample predictive calibration can be expressed as in-sample calibration times 1 minus this new measure of overfitting. We then illustrate our new measure of overfitting and its relation with out-of-sample and in-sample calibration by means of a simulation study. Our simulations are based on a model intended to emulate individual healthcare expenditures as such models are typically highly nonlinear given their strictly positive and right-skewed dependent variable. The baseline model is a GLM model and misspecification is introduced either through an incorrect link function or wrong distributional assumption, making it possible to assess the effect of model misspecification on the various measures of shrinkage presented in this paper. We finally present a real-data illustration by fitting our baseline model on health care expenditure data from a well-known hospitatlist study (Meltzer et al., 2002) in order to show that the distinctions we make matter in practice.

An important result from our simulations is that in-sample model misspecification easily outweighs overfitting in the narrow sense. Thus, when evaluating the out-of-sample predictive accuracy of their model, the analysts should take into account both overfitting and in-sample model specification. This is especially relevant for model selection. Large in-sample misspecification bias calls for actions such as changing the functional form and adding complexity to the model in order to more accurately capture the variations of the dependent variable. On the other hand, large overfitting calls for actions such as reducing model complexity and increasing the efficiency of the estimation method. That last point is well-illustrated by our simulations where we show that an inefficient GLM can lead to considerable out-of-sample bias, and this despite of its in-sample robustness to model misspecification within the exponential family. More generally, our results highlight the fact that indiscriminately preferring an unbiased estimator over an efficient one is by no means a safe strategy, as the inefficiency of the former does not only weakens inference, but also ultimately results in biased out-of-sample predictions.

Our real-data illustration shows that the scale on which overfitting is measured matters a lot, as the estimated shrinkage can substantially differ between the original and estimation scales of the dependent variable. The scale-of-interest measure of overfitting we propose might thus be relevant to those primarily interested in the original scale, which might have relevant scientific or policy interpretations to them. The illustration also confirms that in-sample misspecification matters a lot as the resulting shrinkage dominates the one due to overfitting in all our examples. Finally, the illustration shows that the role played by sample size can be considerable by comparing the results obtained with the full sample to those obtained with only part of the data.

As for further directions of research, the small sample properties of our shrinkage statistics deserve some attention given the numerous types or research based on limited datasets, especially those from randomized medical trials. A related point would be to improve the efficiency of our measures of shrinkage as efficiency is critical in resampling estimation procedures. A solution could be to use the variance function estimated in the first stage of the estimation procedure as weights in the second stage of the procedure. It is also important to stress that we have used the GLM family because it is a very convenient way to introduce in-sample biases by means of inadequate link functions and inefficiency through wrong distributional assumptions. However, when measuring shrinkage on the scale of interest, the GLM framework is no longer needed.

A related point is that when using a log-GLM as a baseline scenario, the shape of the distribution needs to be monotonically decreasing in order to get the type of over-dispersed data that we used, which is a common feature of health care expenditures. We are currently working on the issue that more often than not the distribution is a skewed bell-shaped. Thus, the updated simulations will have more relevance to health data that have the other shape to the pdf. The hospitalist data application clearly indicate that the qualitative story should be similar and that the magnitudes are of interesting size.

Finally, as noted by Blough et al. (1999), a practical advantage of the Copas (1983) preshrunk estimator is that the shrinkage parameter can be nonparametrically estimated from the data, for instance by cross-validation. Thus, our suggestion to correct the Copas (1983) preshrunk predictor by calibrating on the scale of interest instead of the scale of estimation might be very helpful in practice. Indeed, in addition to correcting for in-sample miss-calibration, our suggested pre-shrunk predictor also would have the valuable advantage to dissociate the estimation method from recalibration. Given the

numerous challenges raised by most data, the analysts might appreciate to address these challenges with what they consider to be the most appropriate estimation method, and later recalibrate their predictions by using our scale-of-interest pre-shrunk predictor.

## Appendix A. Large sample approximation of the estimation scale Copas measure of shrinkage

In this Appendix, we generalize the demonstration by Copas (1997) for the logistic regression in the wider framework of the GLM family. Note that we do not make use of the more general setting presented in Copas (1997) as, unlike with the least squares slope, there is no closed form solution for the shrinkage factor in the GLM case. Let us start with the first order condition corresponding to the shrinkage factor $\Delta$ in equation 1:

$$\sum_{i=1}^{n} \frac{y_i^* - \tilde{\mu}_i}{v(\tilde{\mu}_i)} \left( \frac{\partial g(\tilde{\mu}_i)}{\partial \mu} \right)^{-1} \hat{\boldsymbol{\beta}}'_{\boldsymbol{p}} \mathbf{x_i} = 0, \tag{A.1}$$

where $\tilde{\mu}_i = g^{-1}(\hat{\beta}_0 + \Delta \hat{\boldsymbol{\beta}}'_{\boldsymbol{p}} \mathbf{x_i})$. The problem is that the new outcomes $y^*_{i,i=1,..,n}$ are not observed but, under the assumption the that these follow the same distribution as the observed ones, $y_{i,i=1,..,n}$, we do know that $E(y_i^*) = \mu_i$. We also know that shrinkage vanishes asymptotically: plim $\Delta = 1$, which also implies that plim $\tilde{\mu}_i = \hat{\mu}_i$. In large samples, we can also replace $\tilde{\mu}_i$ by its first order Taylor approximation around $\Delta = 1$:

$$\tilde{\mu}_i \simeq \hat{\mu}_i + \left( \frac{\partial g(\hat{\mu}_i)}{\partial \mu} \right)^{-1} \hat{\boldsymbol{\beta}}'_{\boldsymbol{p}} \mathbf{x_i} (\Delta - 1) \tag{A.2}$$

So, by taking the expectation of equation A.1, replacing $\tilde{\mu}_i$ by its approximation, and the functions of $\tilde{\mu}_i$ by their limits in probability, the first order condition corresponding to $\Delta$ can be rewritten as follows:

$$\sum_{i=1}^{n} \frac{\hat{\mu}_i - \mu_i}{v(\hat{\mu}_i)} \left( \frac{\partial g(\hat{\mu}_i)}{\partial \mu} \right)^{-1} \hat{\boldsymbol{\beta}}'_{\boldsymbol{p}} \mathbf{x_i} = (1 - \Delta) \hat{\boldsymbol{\beta}}'_{\boldsymbol{p}} I \hat{\boldsymbol{\beta}}_{\boldsymbol{p}}, \tag{A.3}$$

where $I = \sum_{i=1}^{n} (\frac{\partial g(\hat{\mu}_i)}{\partial \mu})^{-2} \frac{\mathbf{x_i} \mathbf{x_i}'}{v(\hat{\mu}_i)}$ is the Fisher information matrix corresponding to $\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}$. We can now replace $\mu_i$, which is a function of parameters $\beta_0$ and $\boldsymbol{\beta_p}$, by its first order Taylor approximation around the estimated parameters $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{p}}$, to obtain:

$$(\hat{\beta}_0 - \beta_0) \hat{\boldsymbol{\beta}}'_{\boldsymbol{p}} \sum_{i=1}^{n} \left( \frac{\partial g(\hat{\mu}_i)}{\partial \mu} \right)^{-2} \frac{\mathbf{x_i}}{v(\hat{\mu}_i)} + (\hat{\boldsymbol{\beta}}_{\boldsymbol{p}} - \boldsymbol{\beta_p})' \sum_{i=1}^{n} \left( \frac{\partial g(\hat{\mu}_i)}{\partial \mu} \right)^{-2} \frac{\mathbf{x_i} \mathbf{x_i}'}{v(\hat{\mu}_i)} \hat{\boldsymbol{\beta}}_{\boldsymbol{p}} = (1 - \Delta) \hat{\boldsymbol{\beta}}'_{\boldsymbol{p}} I \hat{\boldsymbol{\beta}}_{\boldsymbol{p}}, \tag{A.4}$$

Since the $\mathbf{x_i}$ are assumed to be centered, the first sum equals 0. It follows:

$$\hat{\boldsymbol{\beta}}' I (\hat{\boldsymbol{\beta}}_{\boldsymbol{p}} - \boldsymbol{\beta_p}) = (1 - \Delta) \hat{\boldsymbol{\beta}}'_{\boldsymbol{p}} I \hat{\boldsymbol{\beta}}_{\boldsymbol{p}}, \tag{A.5}$$

Finally, we obtain equation 2 by taking the expectation of the LHS:

$$E(\hat{\boldsymbol{\beta}}' I (\hat{\boldsymbol{\beta}}_{\boldsymbol{p}} - \boldsymbol{\beta_p})) = \text{trace}(I I^{-1}) = p. \tag{A.6}$$

**References**


Arlot, S., Celisse, A., Jul. 2009. A survey of cross-validation procedures for model selection.

Basu, A., Arondekar, B. V., Rathouz, P. J., 2006. Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. Health Economics 15 (10), 1091–1107.

Basu, A., Rathouz, P. J., 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostatistics 6 (1), 93–109.

Blough, D. K., Madden, C. W., Hornbrook, M. C., 1999. Modeling risk using generalized linear models. Journal of Health Economics 18 (2), 153–171.

Copas, J. B., 1983. Regression, prediction and shrinkage. Journal of the Royal Statistical Society. Series B (Methodological) 45 (3), 311–354.

Copas, J. B., 1987. Cross-validation shrinkage of regression predictors. Journal of the Royal Statistical Society. Series B (Methodological) 49 (2), 175–183.

Copas, J. B., 1997. Using regression models for prediction: shrinkage and regression to the mean. Statistical Methods in Medical Research 6 (2), 167–183.

Geisser, S., 1975. The predictive sample reuse method with applications. Journal of the American Statistical Association 70, 320–328.

Harrell, F. E., 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer.

Harrell, F. E., Lee, K. L., Mark, D. B., 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15 (4), 361–387.

Larson, S. C., 1931. The shrinkage of the coefficient of multiple correlation. [article]. Journal of Educational Psychology 22(1), 45–55.

Manning, W., Mullahy, J., 2001. Estimating log models: To transform or not to transform? Journal of Health Economics 20 (4), 461–494.

Manning, W. G., Basu, A., Mullahy, J., 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. Journal of Health Economics 24 (3), 465–488.

Meltzer, D., Manning, W. G., Morrison, J., Shah, M. N., Jin, L., Guth, T., Levinson, W., Dec 2002. Effects of physician experience on costs and outcomes on an academic general medicine service: results of a trial of hospitalists. Ann Intern Med 137 (11), 866–874.

Nelder, J. A., Wedderburn, R. W. M., 1972. Generalized linear models. Journal of the Royal Statistical Society. Series A (General) 135 (3), 370–384.

van Houwelingen, J. C. V., Cessie, S. L., 1990. Predictive value of statistical models. Statistics in Medicine 9 (11), 1303–1325.

Veazie, P. J., Manning, W. G., Kane, R. L., 2003. Improving risk adjustment for medicare capitated reimbursement using nonlinear models. Medical Care 41 (6), 741–752.

Wherry, R. J., 1931. A new formula for predicting the shrinkage of the coefficient of multiple correlation. The Annals of Mathematical Statistics 2 (4), 440–457.