Distribution Agreement

____

Svetlana Masalovich Data

Possible benefits of the "logistic flip" in discerning between two logistic regression models

By Svetlana Masalovich
Master of Science
Biostatistics

____

Robert Lyles, PhD, Thesis Advisor

____

Ying Guo, PhD, Committee member

____

Lance Waller, PhD, Department Chair

Accepted:

____

Lisa A.Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

____

Date

Possible benefits of the “logistic flip” in discerning between two logistic regression models

by

Svetlana Masalovich,
M.Sc., Moscow State University, 1990

Advisor:

Robert Lyles, PhD

An abstract of

A thesis submitted to the Faculty of
the James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of Master of Science
in Biostatistics

2010

Abstract

Possible benefits of the “logistic flip” in discerning between two logistic regression models

by Svetlana Masalovich

We investigated the relationship between and properties of odds ratio estimates in two logistic regression models: the model for a dichotomous outcome (*Y*) and dichotomous predictor (*X*) including an auxiliary continuous covariate, and the “flipped” model where *X* and *Y* are interchanged. The odds ratio is invariant to flipping when no additional covariates are considered. However, its estimates yielded by the two models in the presence of covariates are generally different unless some finer adjustments for covariate effects on the outcome in the flipped model are made (e.g., involving polynomial terms). The reason is appearing in the flipped model a non-linear function of the covariates and the model parameters and a function of the conditional probability of the predictor given the covariates are introduced. We demonstrated that the odds ratio estimates from the two models can be similar without adjustment if the function of the covariate in the flipped model is approximately linear and the predictor and covariate in the initial model are independent or related through a logistic regression. When the model for the predictor and covariate is not logistic, nonparametric approaches (such as LOESS) can be employed to estimate the conditional probability of *X* given the covariates.

We found that the extent of the equivalence of odds ratio estimates in initial and flipped models can be useful in data sets with covariates when it is not known which outcome is more appropriate, *Y* or *X* . We hypothesized that the difference between the odds ratio estimates yielded by the initial and “flipped’ models with *Y* as the outcome in the initial model, and that difference when, instead, *X* is the outcome, can be used to discern between the correct and incorrect models. It was expected that the estimates based on the initial and reversed correct model would tend be closer to one another than the corresponding estimates based on the incorrect model. The simulation study

confirmed this hypothesis in general. This approach can be useful in studies where it is not clear which model is more appropriate.

Possible benefits of the “logistic flip” in discerning between two logistic regression models

by

Svetlana Masalovich,
M.Sc., Moscow State University, 1990

Advisor:

Robert Lyles, PhD

A thesis submitted to the Faculty of
the James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of Master of Science
in Biostatistics

2010

Acknowledgments

I would like to acknowledge first of all my thesis advisor, Dr. Robert Lyles for suggesting the interesting idea for the thesis, his encouragement, support and numerous advice. It was great pleasure to work on this topic.

I would like to thank Dr. Waller for thorough reading my thesis and making valuable comments.

I sincerely appreciate Dr. Ying Guo for finding time on the short notice to read my thesis and giving useful suggestions.

I would like to thank Paul Weiss for sharing with me his SAS script that became an important part of the code used for simulation study.

My years of studying at Biostatistics and Bioinformatics department turned to be challenging time in terms of balancing between very intensive academic life, part-time job and family life. I received a lot of understanding, encouragement and advice from many faculty members throughout my studies and difficult times. I am especially grateful for that to Robert Lyles, Ying Guo and Qi Long and my academic advisor DuBios Bowman.

I would like to give my special thank to my husband , Vazha Glonty, for believing in my abilities, encouragement, help and support.

Table of Contents

Tables and Figures

Introduction

Logistic regression is a widely used and intensively studied method for analyzing categorical data. One of the most common measures of effect in the models for categorical variables is the odds ratio (OR). It can be estimated in many study designs, such as cross-sectional, prospective and retrospective. One of the interesting features of the odds ratio is its invariance property. In 2x2 contingency tables, it means that the odds ratio does not change when the orientation of the table reverses [Agresti, 2002, p.45]. An important implication of this property is that the odds ratio can be defined using conditional probabilities in either direction. Accordingly, the sample odds ratio estimates the same parameter in prospective, retrospective or cross-sectional sampling designs.

Similar to the 2x2 table case, the invariance property holds exactly in a logistic regression model with a single dichotomous explanatory variable. If the independent and dependent variables are reversed, both the initial and "flipped" models are of logistic form. By "flipping" we mean that predictor and explanatory variables are reversed so that the predictor is replaced by the original outcome. With a single binary predictor, the OR estimates yielded by the initial model and the reversed model are equivalent [Cornfield,1951]. This is a result of the fact that the estimated regression coefficient of the explanatory variable in the original model is the same as that coefficient in the flipped model. In retrospective studies this symmetry in presentation between Y and X can be used to estimate odds ratios based on a model for disease given the exposure rather than the reverse. Modeling the former, sometimes referred to as prospective probability, is usually more consistent with objectives, but in case-control studies one cannot estimate the prospective probability directly.

The reasonable question arises whether the flip of the outcome and a binary predictor preserves the estimate of the OR for the binary predictor in logistic regression with additional explanatory variables, either continuous or categorical. Anderson [1972] suggested modeling the probability of disease conditional on exposure and categorical explanatory vectors and showed that after maximization over nuisance parameters the maximum likelihood estimate of the logistic regression coefficient of the prospective model is the same that of the retrospective model. Prentice [1976] used the invariance property and explored the effect of a covariate on the odds ratio in retrospective studies that were modeled by a logistic regression model. He illustrated that for the fixed value of a continuous covariate or level of a categorical covariate, the odd ratio estimates are equivalent, if the assumption is made that the influence of the covariate on the probabilities of the predictor conditional on regressors is the same for cases and controls. To relax this restriction, the interaction term between this predictor and the covariate can be added to the model. Then the odds ratio will be a function of the regression coefficients corresponding to the predictor and the interaction term. The author also illustrated that more generally the two models, initial and flipped, yielded quite different parameter estimates.

Pursuing further the results presented by Prentice [1976] and Zelen [1971], Breslow and Powers [1978] compared the estimates of the relative risks from the initial model and the model with flipped predictor and explanatory variables in the context of prospective and retrospective studies. They stated that for analyzing retrospective studies with linear logistic regression there are two valid approaches. In the prospective model the dependent variable is an indicator of case/control status and the independent variables include exposure. In the retrospective model the dependent variable is an exposure and the independent variables include case/control status. Together with predictor, both models may contain confounding factors as well as their interactions and require covariate adjustment to achieve similar estimates of OR. The authors showed that when the covariates were discrete or "nearly" continuous so that the data could be arranged into a series of 2x2 tables, the two models yielded identical estimates of the odds ratios if the covariate effects on the outcome were saturated.

A great body of research conducted to study the invariance property of the OR was motivated by the need to estimate parameters in case-control studies. Prentice and Pyke [1979] extended Anderson's approach, generalized the findings of Breslow and Powers and applied them to various designs in case-control studies. They showed that in case of a very general form of exposure modeling the "flipped" (in our terminology) model is also of logistic form when applied to case-control data. If the sample space for the vector of exposure variable is finite, the model for exposure conditional on disease is an ordinary logistic and the odds ratio parameter estimates as well as their asymptotic variance matrices can be obtained by standard likelihood methods. However, they noted that more generally, due to the unspecified nuisance function in the "flipped" model, nonstandard estimation theory is required. They developed required likelihood equations and asymptotic distribution theory for the case of a very general form of regressor that may be continuous or mixed, forming strata on the basis of the exposure variable. They also noted that if, instead, the auxiliary variable is modeled, the initial and inverse models can be equivalent if the sample space for the auxiliary variable is finite and the functions of the explanatory variables on the right side of the flipped model are unrestricted. If the sample space for auxiliary variable is not finite, these functions usually should be restricted, for example by permitting for them to saturate as it was suggested by Breslow and Powers.

The present work is closely related to Breslow and Powers's paper [1978]. The discussion of Prentice and Pyke [1979] also mentioned some issues considered in this work. We have further explored the relations between the initial and flipped models and the properties of odds ratio estimates when there is one dichotomous explanatory variable and an auxiliary covariate to be considered. However, unlike the Breslow and Powers work, we considered the case where the covariate is continuous. We have studied why in some cases the OR estimates are essentially invariant to "flipping" without the adjustment proposed by Breslow and Powers, and why this adjustment is needed in other instances. Specifically, we have derived that the flipped model contains the sum of two (in general) nonlinear functions: a function of the conditional probability of the predictor on the covariate, and a function of the covariate and the initial model parameters. The former contains the unknown conditional probability of the predictor given the covariate, and it can be linearly or non-linearly associated with the covariate. The latter is generally a nonlinear function of the covariate. In case these terms can be well approximated by a simple linear function of the covariate, the invariance of odds ratio estimates approximately holds. It can be easily achieved when the continuous covariate and binary predictor are also related through a linear logistic regression model (for the first term to be linear), and if either the effect of the covariate or the variance of the covariate is small (for the second term to be linear). We have shown that even in more general cases the invariance of OR estimates can be achieved if the response variable and covariate are related by a linear or polynomial logistic regression model and the non-linear term in the flipped model is well approximated by polynomials of the covariate. This conclusion is consistent with the findings for relative risk presented by Breslow and Powers, but it was made using a different perspective. We did not require the discretization o the explanatory variable as in Breslow and Powers' work, and, unlike Prentice and Pyke, we did not restrict to case-control study designs.

It is not an unusual situation in biomedical studies when is not clear which variable would be best modeled better as a dependent and which as an independent variable. For example, in cross-sectional studies the exposure and disease are measured at the same point in time, and it may not be possible to distinguish whether the exposure preceded or followed the disease. In our terminology, the choice of dependent variable becomes a decision regarding which model, the initial or flipped one, is more appropriate.

Breslow and Powers [1978] discussed "flipping", but in a slightly different context. Both the prospective and retrospective models considered by the authors included a vector of explanatory (nuisance) variables in the form of functions of covariates and the interaction terms between this function and the predictor. The two models provided similar estimates of the relative risk and OR with an increasing degree of covariate adjustment that was achieved by adding polynomial terms to both models. The authors noted that one of the models can be preferred to another depending on the degree of covariate adjustment, the type of risk factors, the simplicity of the modeling, or how strong the association between the covariate and disease or risk factors is.

We were interested in applying the invariance property of the OR to a somewhat different situation than studied by Breslow and Powers and Prentice and Pyke, and not necessarily assuming a retrospective design.

Objectives

The purpose of this work is to investigate the properties of OR estimates in the initial and flipped models and find the conditions when the estimate of OR is invariant. We aimed to find out how this invariance can be achieved if it is not invariant in the reversed model with the original covariate. We were interested also in a possible application of this invariance property of OR estimates. We found that it can be used to inform the choice of which variable (*X* or *Y*) is best treated as the dependent variable in a logistic regression when the goal is to estimate their adjusted odds ratio. We were also interested in deriving a measure or criteria for how to choose between the models. This is motivated by the following reasoning:

Let us consider the case of a data set with two categorical variables (*Y* and *X*) and one continuous covariate (*Z*), when it is not known whether Y or X is best treated as the true outcome and therefore, whether the logistic model $Y|X,Z$ or $X|Y,Z$ is the correct one. We demonstrate that very similar estimates of OR in the two models can be achieved by adding polynomial terms of the covariate to the flipped model if *Y* is the true outcome. However, if the wrong model ($X|Y,Z$) is fit as the initial one, it would be harder to achieve the invariance of OR estimates by "flipping" because this invariance is predicated on the assumption that the initial model is correct. We expect this will most often result in the finding that the initial and flipped OR estimates will be more different than the estimates were in the first case, when the correct model was fit as the initial one. So the closeness of the two pairs of OR estimates can be used as a criteria to determine the model that describes the data the best.

Methodology

I. Invariance of Odds Ratio

Let *Y* and *X* be correspondingly response and explanatory binary variables taking values 0 and 1, and let Z be a continuous covariate.

Denote by $\underline{Z}$ a vector of covariates of length *m*, and by α, β and γ the regression coefficients in the model for *Y* conditional on *X* and $\underline{Z}$.

Using Bayes' rule it easy to show that the OR is invariant in the following sense:

$$\begin{aligned} OR &= \frac{\Pr(Y=1 \mid X=1, Z=z)/\Pr(Y=0 \mid X=1, Z=z)}{\Pr(Y=1 \mid X=0, Z=z)/\Pr(Y=0 \mid X=0, Z=z)} \\ &= \frac{\Pr(X=1 \mid Y=1, Z=z)/\Pr(X=0 \mid Y=1, Z=z)}{\Pr(X=1 \mid Y=0, Z=z)/\Pr(X=0 \mid X=0, Z=z)} \end{aligned}$$

Now let us consider the logistic regression model with one predictor:

$$\operatorname{logit}\left[\Pr(Y=1 \mid X=x)\right] = \alpha^* + \beta y \,.$$

This equation implies that

$$\Pr(Y=1 \mid X=x) = \pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} = 1/(1+\exp(-\alpha-\beta x)) \,.$$

It can be shown (Cornfield, 1951) that the above model also implies a logistic model of the binary variable X=1 conditional on y, i.e.,

$$\text{logit}\left[\Pr(X=1\,|\,Y=y)\right]=\alpha^{*}+\beta y\,,$$

with the same coefficient β as the model above.

Indeed, since $Y\sim Bin(1,\pi(x))$,

$$\Pr(Y=y\mid X=x)=\pi(x)^{y}(1-\pi(x))^{1-y}=\frac{\exp(\alpha+\beta x)y}{1+\exp(\alpha+\beta x)}$$

Applying Bayes' theorem we can derive

$$\Pr(X=1\,|\,Y=y)=\frac{\Pr(Y=y\mid X=1)\Pr(X=1)}{\Pr(Y=y\mid X=1)\Pr(X=1)+\Pr(Y=y\mid X=0)\Pr(X=0)}$$

$$=\left[1+\left[\frac{1-p_x}{p_x}\right]\left[\frac{1+\exp(\alpha+\beta)}{1+\exp(\alpha)}\exp(-\beta y)\right]\right]^{-1}=1/(1+\exp(-\alpha^{*}-\beta y)),$$

where $p_x=\Pr(X=1)$ and $\alpha^{*}=\ln\left[\dfrac{p_x}{1-p_x}\right]+\ln\left[\dfrac{1+\exp(\alpha)}{1+\exp(\alpha+\beta)}\right].$

Hence, $OR=\exp(\beta)$ is the same as in the previous model and the logistic link is preserved.

Now consider the logistic regression models with one predictor and covariates:

$$\text{logit}[\Pr(Y=1\mid X=x,\underline{Z}=\underline{z})=\text{logit}[\pi(x,\underline{z})]=\alpha+\beta x+\underline{\gamma}'\underline{z}\,, \qquad (1)$$

We can show that logistic model for probability *X*=1 conditional on *y* and *z* is given by equation:

$$\text{logit}[\Pr(X=1\,|\,Y=y,\underline{Z}=\underline{z})\approx\alpha^{*}+\beta y+\underline{\gamma}^{*\prime}\,\underline{z}$$

in special cases (details to follow), or by the equation

$$\text{logit}[\Pr(X=1\,|\,Y=y,\underline{Z}=\underline{z})=\alpha^{*}+\beta y+\underline{\gamma}^{*\prime}u(\underline{z}) \qquad (2)$$

in the general case. Since $Y \sim Bin(1, \pi(x, \underline{z}))$,

$$\Pr(Y = y \mid X = x, \underline{Z} = \underline{z}) = \pi(x, \underline{z})^{y}(1 - \pi(x, \underline{z}))^{1-y} = \frac{\exp(\alpha + \beta x + \underline{\gamma}'\underline{z})y}{1 + \exp(\alpha + \beta x + \underline{\gamma}'\underline{z})}.$$

Then we derive

$$\Pr(X = 1 \mid Y = y, \underline{Z} = \underline{z}) =$$

$$= \left[1 + \left[\frac{1 - p_{x|z}}{p_{x|z}}\right]\left[\frac{1 + \exp(\alpha + \beta + \underline{\gamma}'\underline{z})}{1 + \exp(\alpha + \underline{\gamma}'\underline{z})}\right]\exp(-\beta y)\right]^{-1} = \frac{1}{1 + \exp(-\alpha_z^* - \beta y)},$$

where $p_{x|z} = \Pr(X = 1 \mid \underline{Z} = \underline{z})$, and

$$\alpha_z^* = \ln\left[\frac{p_{x|z}}{1 - p_{x|z}}\right] + \ln\left[\frac{1 + \exp(\alpha + \underline{\gamma}'\underline{z})}{1 + \exp(\alpha + \beta + \underline{\gamma}'\underline{z})}\right]. \qquad (3)$$

Throughout the paper we will use the following notation:

$$g_1(\underline{z}) = \ln\left[\frac{p_{x|z}}{1 - p_{x|z}}\right]$$

and

$$g_2(\underline{z}) = \ln\left[\frac{1 + \exp(\alpha + \underline{\gamma}'\underline{z})}{1 + \exp(\alpha + \beta + \underline{\gamma}'\underline{z})}\right].$$

If we rewrite the logistic equation for X in the form

$$\Pr(X = 1 \mid Y = y, \underline{Z} = \underline{z})$$

$$= \beta y + \ln\left[\frac{p_{x|z}}{1 - p_{x|z}}\right] + \ln\left[\frac{1 + \exp(\alpha + \underline{\gamma}'\underline{z})}{1 + \exp(\alpha + \beta + \underline{\gamma}'\underline{z})}\right] = \beta y + \alpha_z^* \qquad (4)$$

it is easy to notice that (4) is a linear -logistic regression equation if $g_1(\underline{z})$ and $g_2(\underline{z})$ can be well approximated by a simple linear function of $\underline{z}$. Then the OR invariance property will hold directly in the sense that a logistic regression of X on Y and the elements of $\underline{z}$ allows valid estimates of β. In general, however, neither $g_1(\underline{z})$ nor $g_2(\underline{z})$ is a linear function of $\underline{z}$. Note that $g_1(\underline{z})$ contains the unknown probability of X given $\underline{z}$, whereas $g_2(\underline{z})$ is a known and generally non-linear function of unknown parameters α, β, γ, and $\underline{z}$.

In fact, for the model (4) to be logistic we do not need to restrict $g_1(\underline{z})$ and $g_2(\underline{z})$ to only simple linear functions of $\underline{z}$. If there is any function, linear with respect to regression coefficients, that closely approximates α_z^*, then (4) still will approximate a linear logistic model. This will be the case, for example, if α_z^* is well approximated by a linear predictor involving higher order terms in $\underline{z}$.

Note that $g_1(\underline{z})$ can be approximated by a function of $\underline{z}$ linear w.r.t regression coefficients, if a linear- logistic regression is an appropriate model for $x|\underline{z}$. In other words, if

$$\operatorname{logit}\Pr(X = 1 \mid \underline{z}) = \ln\left[\frac{p_{x|z}}{1 - p_{x|z}}\right] = a_1 + \underline{b}_1' v(\underline{z}) \ , \qquad (5)$$

where $v(\underline{z})$ is any function of $\underline{z}$ linear w.r.t. coefficients of $\underline{z}$, including a function with higher power terms in $\underline{z}$. Note that in case of perfect independency between X and $\underline{Z}$, $p_{x|z} = p_x$ and $g_1(\underline{z})$ is a constant that does not depend on $\underline{z}$. However, this rarely occurs in practice.

The third term in (4), $g_2(\underline{z})$, is generally an S-shaped (sigmoidal) function and it is not a linear function of $\underline{z}$, although it is approximately linear for some range of values of $\underline{z}$ or if $|\underline{\gamma}'\underline{z}|$ is small enough as it will be shown in the appendix. Note that at very large values of $|\underline{\gamma}'\underline{z}|$ the solution to equation (1) may not exist, since in this situation $\Pr(Y = y \mid X = x, \underline{Z} = \underline{z})$ is close to 1 and therefore, Y takes values 1 (or 0) at nearly all values of $\underline{z}$ (see Appendix for details). Therefore, we expect that $g_2(\underline{z})$ will be a linear or almost linear function of $\underline{z}$ in many data sets. In the general case, $g_2(\underline{z})$ can be approximated by a linear combination of polynomials of $\underline{z}$:

$$g_2(\underline{z}) = \ln\frac{1+\exp(\alpha+\underline{\gamma}'\underline{z})}{1+\exp(\alpha+\beta+\underline{\gamma}'\underline{z})} \approx a_2 + \underline{b}_2' w(\underline{z}), \tag{6}$$

where $w(\underline{z})$ can be any function of $\underline{z}$, linear w.r.t. coefficients of $\underline{z}$ including a function with higher power terms in $\underline{z}$. Note that any smooth function can be approximated by polynomials to some degree of precision.

Finally, (3) can be written as

$$\alpha_z^* = \alpha^* + \underline{\gamma}^{*\prime} u(\underline{z}) \tag{7}$$

Here $u(\underline{z}) = v(\underline{z}) + w(\underline{z})$ and $\alpha^* \approx a_1 + a_2$. Note that first order term $\underline{\gamma}_1{}^* \approx \underline{b}_1 + \underline{b}_2$.

Consequently, (2) is a logistic model for *X* conditional on *y* and $\underline{z}$, and it can be written as:

$$\operatorname{logit}(X = 1 \mid y, \underline{z}) = \alpha^* + \beta x + \underline{\gamma}^{*\prime} u(\underline{z}) \tag{8}$$

Therefore, we can expect that with an appropriate choice of $u(\underline{z})$ the estimate of β in (8) would equal or closely approximate the estimate of β in the initial model (1). This implies an OR estimate invariance property in the model with a continuous covariate.

How can we choose $u(\underline{z})$? First we need to estimate the expressions for $g_1(\underline{z})$ and $g_2(\underline{z})$. In case the association between x and $\underline{z}$ is adequately described by the linear logistic regression for X conditional on $\underline{z}$, the linear coefficients for $g_1(\underline{z})$ can be estimated by fitting this model with first or higher order terms of $\underline{z}$.

We are also interested in very realistic practical situations when the association between X and $\underline{Z}$ is not well described by a logistic model and an underlying parametric form for $g_1(\underline{z})$ is unknown. To obtain the coefficients for linear combinations approximating $g_1(\underline{z})$ in this case, we need to estimate $\Pr(X \mid \underline{z})$. It turns out that $\Pr(X \mid \underline{z})$ can be well approximated by predicted values of X obtained by fitting the model for X on z by nonparametric approaches. Among those are, for example, the generalized additive model (GAM) [Hastie and Tibshirani, 1990] and the local regression method (LOESS) [Cleveland, 1988]. We used the latter because it yielded better results in our experimental cases. In local regression, the relationship between the dependent and independent variables is modeled locally by weighted regression. One of the attractive features of local regression for estimating $\Pr(X \mid \underline{z})$ is that the fitting is performed in a moving fashion, similar to what could be done if we wanted to roughly estimate $\Pr(X \mid \underline{z})$ without any regression techniques [Copas,1983]. In the LOESS method, the regression surface is estimated by fitting locally linear or quadratic functions of predictors in some parametric class using weighted least squares. The smoothness of the estimated surface is controlled by the fraction of the data in each local neighborhood, called the smoothing parameter. The smoothing parameter can be chosen automatically, for example, by a method that minimizes a criterion that incorporates both the tightness of the fit and the model complexity. One such method, AIC_{C1} , is based on bias- corrected Akaike information criteria [Cohen,1999]. AIC_{C1} was shown to avoid the tendency of uncorrected AIC to undersmooth and it seems to be the most appropriate for our goal .

After the estimates for $\Pr(X \mid \underline{z})$ are obtained, the function $g_1(\underline{z})$ can be easily calculated. To find the approximation to $g_2(\underline{z})$, the logistic model (1) is fitted and the estimates of α, β and γ are used to compute the $g_2(\underline{z})$ for each z. Next, using the estimated $g_1(\underline{z})$ and $g_2(\underline{z})$, α_z^* is calculated and a multiple regression model for α_z^* with linear and higher degree terms in z is fitted. We use a model selection procedure to select a parsimonious set of terms that provides a desirable level of R^2 . If it is reached, it will suggest that the corresponding multiple linear regression model is adequate for approximating α_z^*, and the selected polynomial terms in z will comprise $u(\underline{z})$ in equation (8). Note, for example, the regression model could be fitted for $g_2(\underline{z})$ only, if $g_1(\underline{z})$ is well described by a logistic model and/or if the range of $g_1(\underline{z})$ is very small compared to that of $g_2(\underline{z})$. However, if neither is the case, fitting the regression model for the sum of these functions may allow one to select fewer polynomial terms. So $\alpha_z^*(\underline{z})$ was fitted in all cases in what follows.

II. **Logistic flip in discerning between two logistic models**

Now consider a data set with two binary variables (*Y* and *X*) and one continuous covariate (*Z*) where it is not known which of the binary variables is best treated as the outcome to obtain a valid estimate of the adjusted OR (β). Suppose the true outcome is *Y* and the data are well described by model (1). Introducing additional indices with the aim to distinguish between the models for the correct outcome and the model for the wrong one (*c* and *w*, respectively) as well as to indicate the initial and flipped models (*i* and *f*, “initial” and “flipped” ones), the initial model (1) for the true outcome can be written as

$$\text{logit}[\Pr(Y=1\,|\,X=x,Z=z)=\alpha_{ci}+\beta_{ci}x+\gamma_{ci}z \ . \qquad (9a)$$

The reversed model for the true outcome is described by the equation:

$$\text{logit}[\Pr(X=1\,|\,Y=y,Z=z)=\alpha_{cf}+\beta_{cf}y+\gamma_{cf}u_c(z) \ . \qquad (9b)$$

If the function $\alpha_{cf}+\gamma_{cf}u(z)$ is a good linear approximation of α^*_{cz}, the estimates of β_{ci} and β_{cf} should be similar ($\hat{\beta}_{ci}\approx\hat{\beta}_{cf}\approx\hat{\beta}_c$).

The model for the “wrong” outcome, *X,* is given by

$$\text{logit}[\Pr(X=1\,|\,Y=y,Z=z)=\alpha_{wi}+\beta_{wi}y+\gamma_{wi}z \ . \qquad (10a)$$

and the corresponding flipped model can be written as

$$\text{logit}[\Pr(Y=1\,|\,X=x,Z=z)=\alpha_{wf}+\beta_{wf}x+\gamma_{wf}u_w(z) \ . \qquad (10b)$$

Here α^*_{wz} is the sum of $g_{w1}(z)$ and $g_{w2}(z)$, which are now the functions of *Pr(Y|z)* and α_{wi}, β_{wi} , and γ_{wi}, respectively .

If model (10a) is not correct and hence, does not fit well, the estimate of β will be biased. Assuming the polynomial terms are needed in (10b), this model will generally not fit as well as (9a) does, since the latter is assumed to be the correct model. Hence, the estimate of β will be not the same as in (9a) and it will be biased. So we expect greater departure of the estimates of OR yielded by the models (10a) and (10b) from the true OR and between themselves, compared with the OR estimates and their difference given by the models (9a) and (9b). We conducted a simulation study to illustrate this point and to assess the extent of equivalence of parameter estimates from the two models. Note that in case both α^*_{wz} and α^*_{cz} are close to linear in the first order term *z,* $u_c(z) = u_w(z) = z$ and all four models yield similar estimates of true OR parameter. We will discuss later that for α^*_{cz}, it occurs when b_1 in the equation analogous to (5) with Y as an outcome is small enough.

Simulation study

The simulation study was conducted to illustrate the behavior of β (OR) estimates from the initial and flipped logistic models, and possible benefits of the interchanging of *Y* with *X* . In particular, our goal was to compare the difference between the two estimates obtained under assumptions that the initial model *Y|x,z* is correct on one side, and that the flipped model *X|y,z* is correct, on another. For this, the true model parameters have to be known. For simplicity we included only one continuous covariate $Z \sim N(0,\sigma)$. To simulate the data for the logistic model, the outcome for the correct model was modeled as $Y \sim Bin(1,\pi_y(x))$ with probability

$$\pi_y(x) = 1/(1+\exp(-\alpha-\beta x-\gamma z)),$$

X was simulated as a binomial variable $X \sim Bin(1,\pi_x(z))$, where $\pi_x(z)$ referred before as $\Pr(x \mid z)$, does not have to be a function of z. Parameters σ and γ were chosen to obtain linear and nonlinear functions for α_z^*, depending on the example.

One example aimed to confirm the invariance of OR estimates in the models with almost linear α_z^*. We also wanted to investigate the effect of various functional forms of association between the binary predictor X and covariate z, such as independence, linear-logistic and non-logistic association. Of special interest was to assess the performance of local regression in estimating Pr(*X*|*z*).

The software SAS 9.2 (SAS Institute, North Carolina,US) was used for simulating observations and estimating the parameters in the simulation study. In particular, the LOGISTIC procedure was used for analyses of logistic models, the REG was employed for assessing linearity of α_z^*, and the LOESS procedure was used to estimate Pr(*X*|*z*) when needed.

The sample size contains 500 observations, since the OR and β estimates tend to be biased for sample size less than 500 [Nemes, 2009]. The number of simulations is 1000 in each case.

The covariate was centered in an effort to avoid potential multicollinearity problems caused by correlations among the polynomial terms. It was not required in our example since the mean of *z* was zero, but it makes the code applicable to any distribution of covariates.

If the association between *X* and *z* can be well described by logistic model (with or without higher order terms), fitting the logistic regression (5) is expected to provide a good approximation. As a criteria of good fit of the model (5), the Hosmer-Lemeshow test was used, with a p-value < 0.05 taken to indicate that a better model is desirable.

To find the approximation to $g_1(z)$ when the logistic model does not fit well, we have to estimate Pr(*X*|*z*). As aforementioned, we used loess regression as it is implemented in SAS by PROC LOESS, with an automated procedure of choosing the smoothing parameter based on AIC_{C1}. In cases when the true model for *X* conditional on *z* was logistic, possibly with higher order terms, we found that $g_1(z)$ was as well approximated by loess regression as by logistic regression with one exception. Namely, in some models the estimated values of Pr(*X*|*z*) for small percentage of z values close to the end of its range were out of range [0,1]. We choose to replace predicted probabilities less than 1E-6 by 1E-6 and predicted probabilities greater then 0.999999 by 0.999999 similar to the approach used by LOGISTIC procedure in SAS. Even if it caused jumps in α_z^* values at the end of the *z* range and consequently, small changes in the coefficients estimates in the linear regression for α_z^*, we found that the main conclusion about model choice remained the same as could be made with the alternative method, where these values were replaced with predicted probabilities obtained with logistic regression. Figure 1 shows plots of Pr(*X*|*z*) obtained by fitting both logistic and LOESS regression overlaid on the true curve, for the artificial case when Pr(*X*|*z*) = *abs(sin(z))*. Obviously, polynomial terms were required in the logistic model to provide a good fit to the non-linear curve. They were added automatically using a model selection procedure based on Shtatland's paper [2001]. Similar results were observed when the original logistic regression had higher order terms. The approximation seems to be adequate, so LOESS was the method of choice for all general forms or unknown Pr(X|z) considered. In particular, in our simulation examples involving non-logistic regression, Pr(X|z) was estimated using LOESS regression.

After the estimates of Pr(X|z) and model (1) parameters have been obtained, α_z^* is calculated and the polynomial approximation to α_z^* is found by fitting the α_z^* values as a dependent variable and the set of polynomials of z as independent variables in a linear regression model. The model selection procedure is performed to select the set of terms that provides desirable R^2 of the model equal to or exceeding 0.975.

The parameters common to for all simulation examples were $\alpha = 0.1$, β=2, and a_1 = 0.5, if applicable . We have also investigated the performance of our approach in the case of more extreme parameter values in the models.

We simulated four situations:

Simulation 1. To confirm the equivalence of β (OR) estimates in the models with linear α_z^*, $g_1(z)$ and $g_2(z)$ were modeled as almost perfectly linear functions of z. For that, X was simulated as a binomial variable with probability $\pi_z(x) = 1/(1+\exp(-a_1 - b_1 z))$, so that $g_1(z)$ would be a linear function of z . The value of b_1 = 0.5 implies rather weak dependency of X on z. We will discuss below that at larger values of b_1 simulation results can be different. The parameters γ and σ were chosen to be small enough, so that $g_2(z)$ would be almost a linear function of z (see appendix): $\gamma = 0.1$ and $\sigma = 1$. In rare instances of simulated data set with non-linear α_z^* (corresponding to R^2 <0.975) were excluded from the further analysis.

We observed that the β estimates obtained by fitting the initial and flipped models without polynomial adjustment agree up to the third decimal point for the assumed correct model (1.9989 (SE=0.2406) and 1.9991 (SE=0.2407)). For the "wrong" model they were respectively, 1.9991 (SE=0.2406) and 2.0 (SE=0.2408). The mean difference between $\hat{\beta}_{ci}$ and $\hat{\beta}_{cf}$ over 1000 simulations for correct model, *Y|X,z,* was 0.003, while that difference under the assumption that, instead, the model *X|Y,z* is correct was 0.004. This demonstrates very small difference between the estimates of β (OR) yielded by the initial and reversed model under both assumptions, correct and incorrect models, and implies that the equivalence of OR estimates holds directly. In terms of our method for discerning the correct model from the incorrect one, even in this example it provided some clues on which model is better: the model *Y|X,z* was chosen 38.5% of times, while the model *X|Y,z* was chosen 29% of times. In the other cases, neither the models was preferred to the other. The means of estimates, $\hat{\beta}_c$ (1.9994) and $\hat{\beta}_w$ (1.9991), were very close to the true value of β, with a slight edge in favor of the correct model.

The results of this and other simulations are summarized in Table 1 in the Appendix.

Simulation 2. For the logistic regression for *X* conditional on *z* and nonlinear $\alpha*_z$, we set a=0.5, b$_1$ = 1, $\gamma = 2$ and $\sigma = 1$.

We observed that β estimates obtained by fitting the initial and flipped models without polynomial adjustment are rather different: $\hat{\beta}_c$ =2.0228 (SE=0.2672) and 1.9512 (SE=0.2752), respectively. After adding polynomials the estimates have agreed up to the third decimal point for the model assumed correct ($\hat{\beta}_{cf}$ = 2.096) and they have been still different for the incorrect model ($\hat{\beta}_{wi}$ = 1.9512 and $\hat{\beta}_{wf}$ =2.0228). Note that no polynomial terms were needed in the reversed "wrong" model. The mean difference between $\hat{\beta}_i$ and $\hat{\beta}_f$ over 1000 simulations under the assumption that the initial model is correct was 0.021. That difference under the assumption that the flipped model is correct was slightly larger (0.074). The correct model was chosen 89.5% of times, and the mean $\hat{\beta}$ = 2.0228 , while the incorrect one was chosen only 9.7% of times, and the mean $\hat{\beta}$ = 1.9512 was farther from the true value.

To illustrate the performance of our method, the Appendix provides an excerpt of SAS output for a single data set in simulation 2. As we see, $\hat{\beta}_{ci}$ = 2.213. To adjust for non-linear functions in the flipped models, four polynomial terms were selected to meet the criterion that R^2 >0.975. This is the most parsimonious model, as can be seen from the output on p 36. After this adjustment, the estimates for the reversed model became $\hat{\beta}_{cf}$ = 2.206, which is very close to $\hat{\beta}_{ci}$. The estimates provided by fitting the "wrong" initial model differed to a greater extent: $\hat{\beta}_{wi}$ = 2.244 and $\hat{\beta}_{wf}$ = 2.213. It is noteworthy that based on the Hosmer-Lemeshow test the correct model is not always preferred to the wrong one. We observed in several simulated data sets that even when the correct model was chosen by our approach, the Hosmer-Lemeshow test sometimes indicated worse fit for the initial model than for the wrong model. In 1000 simulated data sets for simulation example 2, the Hosmer-Lemeshow test indicated better fit for the correct model in only about 43% of times, while this model was chosen 89.5% of times.

We also simulated the logistic model for *X* conditional on z with a large coefficient for *z*, such as 2 and larger. Although at b > 2.5 the correct model was chosen fewer times then the incorrect one, both models were chosen only a very few times (<1%). In other cases, the mean difference between the initial and flipped estimates was the same regardless of which model was taken to be the initial one . We also observed that at b>2.5 the range of $g_1(z)$, which is a linear function in this example, becomes much larger than the range of $g_2(z)$. This resulted in the observation that α_z^* is almost a linear function in both, "correct" and "wrong" models. Hence, no polynomials were required in either of the flipped models, so that the difference between the estimates was nearly or completely identical, as in the case when b=3 and β=1. This resulted in a situation when it was impossible to discern between the correct and the wrong models with these parameters.

<u>Simulation 3</u>. The independency of X and z can be simulated assuming a uniform distribution: $X \sim Uniform(0,1)$. α_z^* was chosen to be nonlinear, so the parameters for model (1) were the same as in the example 2. The loess regression was used to estimate Pr(*X*|*z*) for the correct model and Pr(*Y*|*z*) for the wrong model.

The mean difference between $\hat{\beta}_{ci}$ and $\hat{\beta}_{cf}$ when the initial model is assumed to be the correct one was 0.04, while that difference under the assumption that the reversed model is correct was 0.169. The correct model was chosen 94% of times and the mean $\hat{\beta}$ = 2.03 was close enough to the true value, while the "wrong" one was chosen only 6% of times and the mean $\hat{\beta}$ = 1.88 was substantially farther from the true value.

If X and $\underline{Z}$ are independent, $\hat{p}_{x|z} = \hat{p}_x$ can also be obtained directly from the data as a ratio of counts of ones and zeros.

Simulation 4. For the case of non-logistic regression between X and z and nonlinear α^*_z, we choose the hypothetical from of probability Pr(X|z) = *abs(sin(z)),* since this is a smooth function that takes values in [0,1]. The mean difference between $\hat{\beta}_{ci}$ (2.053) and $\hat{\beta}_{cf}$ (2.086) substantially differs from that between $\hat{\beta}_{wi}$ (1.355) and $\hat{\beta}_{wf}$ (2.057): 0.169 (0.138) and 0.716 (0.274), respectively. This indicates that the estimates yielded by the incorrect model in case of arbitrary relationship between X and Z are very different from each other and from the true value. "Flipping" again allowed us to discern between the "correct" and "wrong" models: the former was chosen about 97% of times, while the latter about 3 % of times.

Discussion

The simulation studies demonstrated that the initial and flipped models often yield estimates of the parameter β that are extremely close to each other and close to the true value if α^*_z is nearly a simple linear function of z. In particular, if α^*_{cz} and α^*_{wz} are both approximately linear, all four models yield similar estimates of the true OR. It always occurs when X and Z are independent or weakly dependent compared to the strength of the association between Y and Z , while γ_{ci} is small enough. This corresponds to the case when Z is not very strongly associated with either Y or X, so that each model can almost be reduced to the one involving only Y and X. In this case it does not matter which model to fit, because in the absence of a covariate we cannot distinguish between dependent and independent variables in the framework of the logistic model. The identity of the estimates of β (or OR) reflects this notion. As expected, if α_z^* is not linear in z, the estimate of β (OR) yielded by the correct initial and flipped models are similar if the appropriate polynomial adjustment is done.

Our simulation study confirmed that the estimates of ln (OR) yielded by the initial model for $Y|X,Z$ under the assumption that Y is the correct outcome is closer to the β estimate obtained by fitting "flipped" model than those estimates yielded by the initial and flipped models when X is erroneously assumed to be a correct outcome. The size of the difference between the estimates yielded by of the initial and flipped models depends on how much of a polynomial adjustment is required for α_z^* to be well approximated by a linear function of z. If more polynomial terms are needed for the flipped correct model than flipped wrong one, the β estimates will be generally closer for the correct model.

If this adjustment can be easily achieved when the "wrong" model is fitted as an initial one, and the flipped correct model does not require any polynomial terms to provide the equivalent estimates of OR by both initial and flipped models, then the incorrect model can be chosen by our method as a "better" model. It happens when X and Z are associated more strongly than Y and Z , so that γ_{ci} is larger than γ_{wi} and, consequently, $g_{w2}(z)$ (and, hence, α^*_{wz}) requires more polynomial terms to be well-approximated by a linear function of z than $g_{w2}(z)$ (and, hence, α^*_{wz}). As a result, in these cases there is a tendency for the model with *X* as the dependent and *Y* and *Z* as independent variables (flipped model) rather than the model having *Y* as the outcome to be selected by our method as yielding better OR estimates.

Note that the OR estimates yielded by the initial "correct" model in all simulation examples are closer to the true values than the estimates obtained by fitting the initial "wrong" model. It is important to notice that although the Hosmer-Lemeshow test results for the correct and wrong models are often consistent with the results of our method, we observed on several occasions that the Hosmer-Lemeshow test indicated worse fit for the initial model than for the wrong one even when the correct model was chosen by our approach. Overall, the correct models were chosen more often based on our approach than based on theHosmer-Lemeshow test.

It is worthwhile to note that when one wishes to choose which model, prospective or retrospective, to fit, that Breslow and Powers recommended fitting the model that requires less covariate adjustment. This leads to choosing the model that may have a very strong association between the covariate and the independent variable and a weak association between the covariate and the dependent variable as the preferred one. This does not invalidate our findings however, because the model in Breslow and Powers' paper includes an interaction term and more generally, the procedure proposed by these authors is only aimed to obtain the RR estimate.

Our simulation study demonstrated that LOESS regression and logistic regression performed equally well when there were higher order terms in the logistic model. In more general, non-logistic cases, LOESS regression outperformed logistic regression in estimating the predicted probabilities.

Unlike Breslow and Powers, we did not require the presence of polynomial terms in both models to achieve similar OR estimates. We also showed that the invariance can be achieved without any adjustment for covariate effect in some circumstances (including when *X* and *Y* on one side, and *Z* on another, are weakly associated). We did not necessarily assume a retrospective design. Our suggested method is aimed primary to cross-sectional studies. In the future its application to case-control studies can be considered.

The Breslow and Powers approach required the sum of the covariate values to be equal to zero. The reason for this restriction can be found in Zelen [1971]. It does not cause loss of generality; however, often it requires covariate transformations that may not be feasible in practice. Our approach avoids this inconvenience, at least in the case of the simple models considered in our work. However, there are several limitations in our approach. First, we assume that one of the initial models is perfectly correct, which is an idealization of situations commonly seen in practice. Also, little theoretical background was developed so far to support the method of discerning between the models, although some efforts have been made. The theoretical justification of the "flipped" model can be easily done if the covariate is categorical or can be categorized. With an essentially continuous covariate, theoretical development is somewhat impeded by that fact that the maximum likelihood approach used to estimate the parameters of logistic model does not produce closed-form solutions so developing new approaches or substantial modification of existing ones is required. In other words, the background for our criteria of choosing the best model is rather data driven. We also restrict ourselves (for simplicity) to the case where the covariate is a scalar. Although it is still possible to extend our approach to handle a few more covariates, managing a large number of covariates can be difficult.

Another (unavoidable in our approach) drawback is that the polynomials were chosen automatically (Royston and Altman [1997]). Multicollinearity was not a serious problem in our simulation, but in other general data sets the inclusion of polynomials in the model results in multicollinearity problems. When the values of the covariate are equally spaced, orthogonal polynomials would be preferred to centered ones. In the case when the covariate takes only positive values, fitting fractional polynomials can give more flexibility and requires fewer terms (Royston and Altman [1997]) .

Accordingly, for future improvements we would suggest further developing the theoretical background for the criteria of choosing the best model. Orthogonal or possibly fractional polynomials can be included as an option in the code. Some other suggestions for the interaction terms and form of covariates appropriate for logistic regression such as by Kay and Little [1987] can be considered. The described approach can be easily extended for the case when the initial model is a polytomous regression model. Alternatively, one could attempt to work directly with the model for $X|Y,Z$ that is implied by model (1), rather than with the appropriate logistic regression model in (2).

The idea of "flipping" can be useful in other contexts, such as when the predictor (X) is continuous (e.g., Lyles, Guo and Hill [2009]). It may also have potential benefits when the predictor (X) and/or the outcome (Y) are measured with error. Another possible application is a nonlinear model with a function of the form of α_z^* on the right side. We have not seen the problems with the nonlinear relationship described by the form exactly as α_z^*, but similar forms are common in pharmaceutical and growth data.

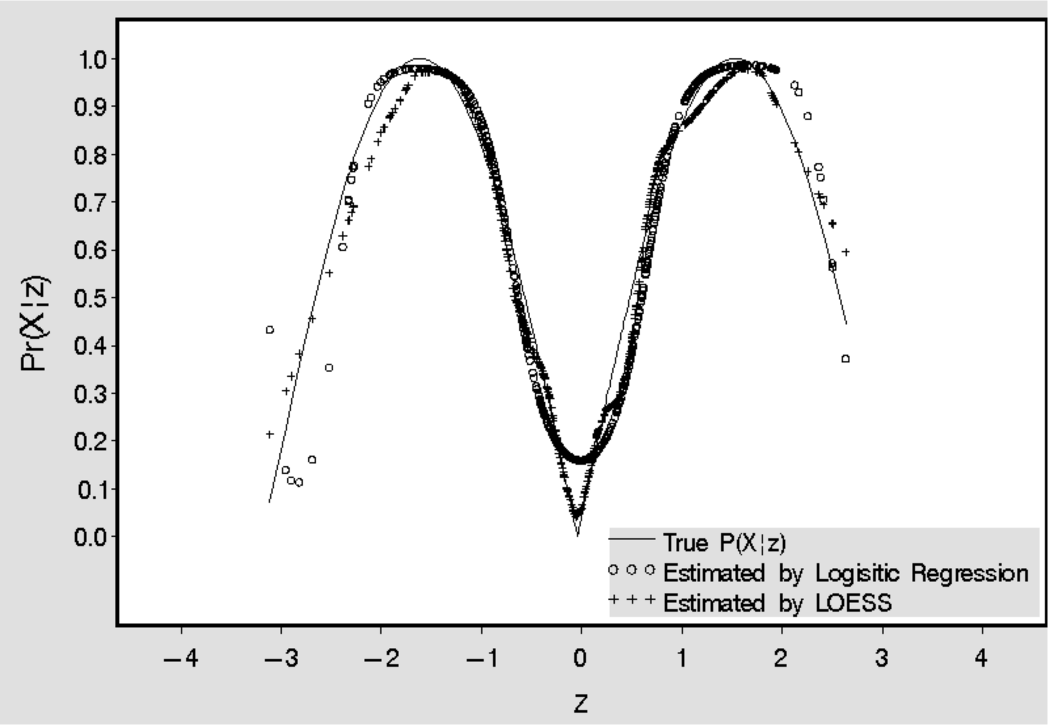
To summarize, exchanging the dependent variable with the independent one in a logistic model can be helpful in discerning between the correct and incorrect models, when one assumes that either logistic regression of *Y* on (*X*,*Z*) or a logistic regression of *X* on (*Y*,*Z*) is the true model that generates the data in most data sets. This approach is potentially useful for ensuing a valid estimate of the adjusted OR that characterizes the association between *X* and *Y*.

Tables and Figures

Table 1. The results of "flipping" simulation study for different distributions and models for X conditional on z and for various α_z^* degrees of linearity. True $\beta = 2$.

Distribution		Parameter estimates			% of correct	"Chosen"
of X on z and α_z^* as a		$\hat{\beta}$ (SE)		mean$\left\|\hat{\beta}_i - \hat{\beta}_f\right\|$ (SD)	model choice	$\hat{\beta}$ (SE)
function of z	Model	Initial	Flipped			
Logit(X) ~ z,	Correct	1.9989(0.2406)	1.9991(0.2407)	0.003(0.003)	38.5	1.9989(0.2406)
α_z^* - linear	Wrong	1.9991(0.2406)	2.0395(0.2408)	0.004(0.004)	29	
Logit(X) ~ z,	Correct	2.0228(0.2752)	2.0195(0.2757)	0.0218(0.019)	89.5	2.0228(0.2752)
α_z^* - non-linear	Wrong	1.9512(0.2672)	2.0229(0.2752)	0.0742(0.051)	9.7	
X ~ Uniform(0,1),	Correct	2.030(0.280)	2.0185(0.277)	0.040(0.033)	93.9	2.025(0.278)
α_z^* - non-linear	Wrong	1.881(0.258)	2.049(0.285)	0.169(0.100)	6.1	
Logit(X) ~ abs(sin(z)),	Correct	2.053(0.302)	2.086(0.366)	0.169(0.138)	96.6	2.053(0.302)
α_z^* - non-linear	Wrong	1.355(0.233)	2.057(0.362)	0.716(0.274)	3.4	

Figure 1. The performance of proc loess and proc log with polynomial terms in estimating Pr(X|z).



References

Agresti, A. Categorical data analyses. 2002. John Wiley & Sons, Inc.

Anderson, J. (1972). Separate sample logistic discrimination. Biometika Vol 59, pp19-35. (Invariance OR inn case-control studies with only variables for disease and exposure)

Royston, P. and Altman, D. (1997). Approximate statistical functions by using fractional polynomial regression. Statistician, 46, No.3 pp.411-422.

Cleveland, W., Devlin, S. and Grosse, E. (1988). Regression by local fitting. Journal of Econometrics 37, 87-114.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast and cervix. Journal of the Nat. Cancer Institute. 11, 1269-75 (about Invariance OR)

Copas, J. (1983). Plotting p against x. Applied Statistics, Vol.32, pp.23-31. (modeling P(x|z)

Breslow, N. (1976) Regression analysis of the Odds Ratio: A Method for retrospective studies. Biometrics, Vol 32, No. 2, pp.409-416.

Breslow N. and Powers.W. (1978). Are there two logistic regressions for retrospective studies? Biometrics, Vol 34, No. 1, pp.100-105. (the main paper)

Cohen, A. (1999) An introduction to PROC LOESS for local regression. SUGI 2001 proceedings, paper 273. Cary, North Carolina. SAS Institute Inc.

Hastie, T. and Tibshirani, R., (1990). Generalized Additive Model. London: Chapman & Hall.

Kay, R. and Little, S. (1987) Transformation of the explanator variables in the logistic model for binary data. Biometrika, Vol. 74, No.3, pp.495-501.

Nemes, S. at el. (2009) Bias in odds ratios by logistic regression modeling and sample size. BMC Medical Research Methodology. Vol 9, No 56. (sample size)

Lyles, R., Guo, Y. and Hill, A. (2009). A Fresh Look at the Discriminant Function Approach for Estimating Crude or Adjusted Odds Ratios The American Statistician. Vol. 63(4), pp. 320-327.

Prentice, R. (1976) Use of the Logistic Model in retrospective studies. Biometrics, Vol 32, No. 3, pp.599-606. (about Invariance OR)

Prentice, R. and Pyke, R. Biometrika. (1979) Logistic disease incidence models and case-control studies. Vol 66, No.3, pp.403-411. (another important paper).

Shtatland, E. Cain, E. and Barton, M. (2001) The perils of stepwise logistic regression and how to escape them using information criteria and the output delivery system. SUGI 2001 proceedings, paper 222. Cary, North Carolina. SAS Institute Inc.

Zelen M. (1971). The analysis of several 2x2 contingency tables. Biometrika, Vol.58, pp. 129-137.

Appendix

I. The condition for linearity of α_z^*

We are interested in the cases where α_z^* is a linear function of $\underline{z}$, and hence, where the invariance of OR estimates holds directly .

First, note that if $\underline{\gamma} > 0$ $g_2(\underline{z}) -> \ln\left(\dfrac{1+e^{\alpha}}{1+e^{\alpha+\beta}}\right)$ as $\underline{z} -> 0$, and $g_2(\underline{z}) -> -\beta$ as $\underline{z} -> \infty$ and $g_2(\underline{z}) -> 0$ as $\underline{z} -> -\infty$.

If $\underline{\gamma} < 0$ $g_2(\underline{z}) -> \ln\left(\dfrac{1+e^{\alpha}}{1+e^{\alpha+\beta}}\right)$ as $\underline{z} -> 0$ and $g_2(\underline{z}) -> 0$ as $\underline{z} -> \infty$ $\underline{z} -> \infty$ and $g_2(\underline{z}) -> -\beta$ as $\underline{z} -> -\infty$.

Hence, g(z) takes values in the interval $(-\beta, 0)$ if $\beta > 0$ and in the interval $(0, -\beta)$ if $\beta < 0$ for any $\underline{\gamma}$.

The expression for $g_2(\underline{z})$ as a function of power terms of $\underline{z}$ can be obtained by expanding $g_2(\underline{z})$ by a Taylor series at $\underline{z}$=0:

$$g_2(\underline{z}) \approx \ln\frac{1+\exp(\alpha)}{1+\exp(\alpha+\beta)} + \frac{\exp(\alpha)[1-\exp(\beta)]}{[1+\exp(\alpha)][1+\exp(\alpha+\beta)]}\underline{\gamma}'\underline{z}$$

$$+\frac{\exp(\alpha)[1-\exp(\beta)][1-\exp(2\alpha+\beta)]}{2[1+\exp(\alpha)]^2[1+\exp(\alpha+\beta)]^2}(\underline{\gamma}^{2\prime}\underline{z}^2 + \sum_{i>j}^{n}\gamma_i z_i \gamma_j z_j)$$

$$+\exp(\alpha)[1-\exp(\beta)][1-\exp(\alpha)-\exp(\alpha+\beta)-6\exp(2\alpha+\beta)$$

$$-\exp(2\alpha+\beta)(\exp(\alpha)+\exp(\alpha+\beta))+\exp(4\alpha+2\beta)]$$

$$\times(1/3!)[1+\exp(\alpha)]^{-2}[1+\exp(\alpha+\beta)]^{-2}\underline{\gamma}^{3\prime}\underline{z}^{3}...,$$

where $\underline{\gamma}^2$ is a vector with elements γ_i^2 and $\underline{z}^2$ is a vector with elements z_i^2, i=1, …,m .

since $g_2'(z_i) \approx \dfrac{\exp(\alpha+\underline{\gamma}'\underline{z})-\exp(\alpha+\beta+\underline{\gamma}'\underline{z}))}{(1+\exp(\alpha+\underline{\gamma}'\underline{z}))(1+\exp(\alpha+\beta\underline{\gamma}'\underline{z}))}\gamma_i$

and

$$g_2''(z_i)\approx \frac{[\exp(\alpha+\underline{\gamma}'\underline{z})-\exp(\alpha+\beta+\underline{\gamma}'\underline{z}))][1-\exp(2\alpha+\beta+2\underline{\gamma}'\underline{z})]}{2[1+\exp(\alpha+\underline{\gamma}'\underline{z})]^2[1+\exp(\alpha+\beta+\underline{\gamma}'\underline{z})]^2}\sum_{i,j}^{n}\gamma_i\gamma_j),$$

In case of scalar z

$$g'''(z_i)\approx\exp(\underline{\gamma}'\underline{z})[\exp(\alpha)-\exp(\alpha+\beta)][1-\exp(\alpha+\underline{\gamma}'\underline{z})-\exp(\alpha+\beta+\underline{\gamma}'\underline{z})$$

$$-6\exp(2\alpha+\beta+2\underline{\gamma}'\underline{z})-\exp(2\alpha+\beta+2\underline{\gamma}'\underline{z})[\exp(\alpha+\underline{\gamma}'\underline{z})+$$

$$\exp(\alpha+\beta+\underline{\gamma}'\underline{z})]+\exp(4\alpha+2\beta+4\underline{\gamma}'\underline{z})]$$

$$\times[1+\exp(\alpha+\underline{\gamma}'\underline{z})]^{-2}[1+\exp(\alpha+\beta+\underline{\gamma}'\underline{z})]^{-2}\gamma_i^{3},$$

It can be seen that $g_2(\underline{z})$ is an approximately linear function of $\underline{z}$ if the quadratic and higher power terms in the expression for $g_2(\underline{z})$ are negligibly small comparing with the linear term. Generally it is true when $|\underline{\gamma}'\underline{z}| \square\ 1$. It is easy to see that first and second coefficients are always less than 1, so for the quadratic term to be small the requirement that

$$|\underline{\gamma}'\underline{z}| \square \left|\frac{2[1+\exp(\alpha)][1+\exp(\alpha+\beta)]}{1-\exp(2\alpha+\beta)}\right| = c*$$

is enough.

For the third term in the case of scalar z

$$|\gamma \mathrm{z}| \square \ \left[6[1+\exp(\alpha)][1+\exp(\alpha+\beta)]\right]^{1/2}$$
$$\times\left[1-\exp(\alpha)-\exp(\alpha+\beta)-6\exp(2\alpha+\beta)-\exp(2\alpha+\beta)(\exp(\alpha)+\exp(\alpha+\beta))+\exp(4\alpha+2\beta)\right]^{-1/2}$$
$$= c**$$

Hence, in case of one covariate, for linearity of α_z^* we need that $|\gamma \mathrm{z}| \square \ \min\{c*, c**\}$.

When α_z^* is linear function of $\underline{z}$, using (5) it easy to derive an approximation for the parameters in flipped model

$$\alpha* \approx a_1 + \ln\frac{1+\exp(\alpha)}{1+\exp(\alpha+\beta)}$$

and

$$\underline{\gamma}*' \approx \underline{b}_1' + \frac{\exp(\alpha)[1-\exp(\beta)]}{[1+\exp(\alpha)][1+\exp(\alpha+\beta)]}\underline{\gamma}'.$$

If , additionally , X and $\underline{Z}$ are independent, then

$$\alpha* \approx \ln\left[\frac{p_x}{1-p_x}\right] + \ln\frac{1+\exp(\alpha)}{1+\exp(\alpha+\beta)} \quad \text{and} \quad \underline{\gamma}*' \approx \frac{\exp(\alpha)[1-\exp(\beta)]}{[1+\exp(\alpha)][1+\exp(\alpha+\beta)]}\underline{\gamma}'.$$

The derivations above demonstrate that if X and $\underline{Z}$ are related through logistic model, so that $g_1(\underline{z})$ is linear in $\underline{z}$, OR estimates in the initial and reversed models will be equivalent when $|\underline{\gamma}'\underline{z}|$ is small enough.

II. Example of SAS output

Here we presented trimmed SAS output for a single data set in the simulation example 2.

INITIAL CORRECT MODEL

The Logistic procedure

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.1416	0.1961	0.5213	0.4703
x	1	2.2126	0.2823	61.4123	<.0001
z1	1	2.0209	0.2233	81.8849	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
x	9.140	5.255	15.896
z1	7.545	4.870	11.688

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
3.5459	8	0.8956

LOGISTIC model for Px|z

The Logistic procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4998	0.0994	25.2850	<.0001
z1	1	0.8510	0.1133	56.3901	<.0001

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
13.5799	8	**0.0934**

MODEL SELECTION for alpha star z

The REG regression procedure

Obs	z1	z2	z3	z4	z5	z6	z7	z8	z9	z10	_RSQ_
1	0.26856	.	.	.	.	.	.	.	.	.	0.64831
2	0.25483	0.11639	.	.	.	.	.	.	.	.	0.86086
3	0.12988	0.11241	0.045851	.	.	.	.	.	.	.	0.94210
4	**0.11959**	**0.22222**	**0.048984**	**-0.02119**	**.**	**.**	**.**	**.**	**.**	**.**	**0.99131**
5	0.12398	0.27577	0.047361	-0.04600	.	0.002390	.	.	.	.	0.99503

FINAL MODEL for alpha star z adj

The REG regression procedure

Root MSE	0.03092	**R-Square**	**0.9913**
Dependent Mean	-0.98239	**Adj R-Sq**	0.9912
Coeff Var	-3.14759		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.15459	0.00198	-584.29	<.0001
z1	1	0.11959	0.00232	51.55	<.0001
z2	1	0.22222	0.00233	95.39	<.0001
z3	1	0.04898	0.00067680	72.38	<.0001
z4	1	-0.02119	0.00040032	-52.93	<.0001

FLIPPED Correct MODEL

The Logistic procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0200	0.2162	22.2629	<.0001
y	1	2.2439	0.2770	65.6088	<.0001
z1	1	0.2061	0.1375	2.2452	0.1340

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
8.8011	8	0.3593

FLIPPED adjusted Correct MODEL

The Logistic procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9490	0.2581	13.5219	0.0002
y	1	2.2057	0.2803	61.9056	<.0001
z1	1	0.4610	0.2079	4.9170	0.0266
z2	1	-0.1540	0.1934	0.6342	0.4258
z3	1	-0.0870	0.0560	2.4131	0.1203
z4	1	0.0409	0.0343	1.4228	0.2330

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
14.5336	8	0.0689

INITIAL Wrong MODEL

The Logistic procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0200	0.2162	22.2629	<.0001
y	1	2.2439	0.2770	65.6088	<.0001
z1	1	0.2061	0.1375	2.2452	0.1340

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
y	9.430	5.479	16.230
z1	1.229	0.938	1.609

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
8.8011	8	0.3593

MODEL SELECTION for alpha star z

The REG regression procedure

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2719	0.1457	76.2489	<.0001
z1	1	2.1600	0.2042	111.9399	<.0001

FINAL MODEL for alph_strz adj

The REG regression procedure

z1	z2	z3	z4	z5	z6	z7	z8	z9	z10	_RSQ_
2.05619	.	.	.	.	.	.	.	.	.	1.00000

FLIPPED adjusted Wrong MODEL

The Logistic procedure

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.1416	0.1961	0.5213	0.4703
x	1	2.2126	0.2823	61.4123	<.0001
z1	1	2.0209	0.2233	81.8849	<.0001

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
3.5459	8	0.8956

Summary table

Variable	Label	Mean
bet_yCorrect	beta from initial model y~x+z	2.2126466
se_bet_yCorrect	Standard Error	0.2823479
bet_xCorrect	beta from flipped model x~y+z	2.2439072
se_bet_xCorrect	Standard Error	0.2770282
bet_adj_xCorrect	beta from flipped model x~y+z(+?)	2.2056726
se_bet_adj_xCorrect	Standard Error	0.2803342
RSQ_alphCorrect	alp~z1+	0.9913051
HLPiCorrect	H-L p-val for y~x,z	0.8956017
true_bet		2.0000000
bet_xWrong	beta from initial x~y+z	2.2439072
se_bet_xWrong	Standard Error	0.2770282
bet_yWrong	beta from flipped y~x+z	2.2126466
se_bet_yWrong	Standard Error	0.2823479
bet_adj_yWrong	beta from flipped y~x+z(+?)	2.2126466
se_bet_adj_yWrong	Standard Error	0.2823479
RSQ_alphWrong	alp~z1+	0.9999999
HLPiWrong	Pr > Chi-Square	0.3593480
trueOR		7.3890561
dif_betc	abs(bet_ycorrect-bet_adj_xcorrect)	0.0069740
dif_betw	abs(bet_xwrong-bet_adj_ywrong)	0.0312606
HLPdiff_count	% times when HLPiCorrect>HLPiWrong	1.0000000
HLPdiff_cw		0.5362537
choose_corY		1.0000000
bestbeta		2.2126466
choose_corX		0

III. SAS partial code

```
/***************************************************************

                  SAS code for flipping logistic regression

***************************************************************/

%global varlist varlist1 nsim  betT  n;
%let  n=500 ;     *sample size;
%let  betT=2;     *true beta;
%let nsim=1000; *number of simulations;

/********      MAIN  macro  ***************/

%macro sim;

%do q=1 %to &nsim;
dm 'clear log';      **Try to keep SAS log from getting too large**;
dm 'clear output';   **Try to keep SAS output from getting too large**;

      %data;*simulating data set;
      *fitting initial correct model;
      %init_model (y= y,x=x,m=C);
      %Pxz_log (x=x);

      /*if need loess;
      %Pxz_loess (x=x); */

      %alph_strz(m=C);
      *fitting flipped correct model ith polyn terms selected in macro
alph_strz;
      %flip_model (x=x ,y=y, m=C);

      ***NOW REPEAT CODE, ASSUMING THE WRONG MODEL IN THE ORIGINAL
FIT***;

      %init_model (y=x,x=y,m=W);
      %Pxz_log (x=y);
      *%Pxz_loess (x=y);
      %alph_strz(m=W);
      %flip_model (x=y ,y=x, m=W);

%end;

%compar; ***print out summary output;

%mend sim;

/*******************************
                  Macros
********************************/

%macro data;
```

```
data data;
      a=0.5; b=1.; alphT=0.1; gammT=2; sigma=1.;
      do i=1 to  &n;
            z1=0 +  sigma*rannor(0);
      px=exp(a + b*z1)/(1 + exp(a + b*z1));
            *px=uniform(0);
            *px= abs(sin(z1));
            *px=exp(a + b*z1+ z2)/(1 + exp(a + b*z1+ z2));
            x=ranbin(0,1,px);
            py=exp(alphT+&betT*x+gammT*z1)/(1 +
exp(alphT+&betT*x+gammT*z1));
            logity= -log((1-py)/py);  logitx= -log((1-px)/px);
            y=ranbin(0,1,py);
      output;
     end;
 drop i  a b  alphT gammT sigma;

 proc sort data=data;
 by z1;

 **centering;
 proc means data=data  noprint;
 output out=meanout  mean=meanz;
 var  z1;
 run;

 data  data;
 IF _N_ = 1 then set  meanout (keep= meanz);
      set  data ;
      if meanz LE 0.01 then  z1=z1-meanz; *0.001;
      z2=z1**2; z3=z1**3; z4=z1**4; z5=z1**5; z6=z1**6; z7=z1**7;
z8=z1**8; z9=z1**9; z10=z1**10;
  run;

 %mend data;

 /****  fitting   initial  model     ****/
 *m= model - correct or wrong;

%macro init_model (y= ,x=, m=);

title "run#&q. INITIAL &m MODEL ";
proc logistic descending data=data;
 model &y=&x z1;
    ods output ParameterEstimates = outPE;
 run; title;


…
%mend init_model;

/***** Logistic model for Px|z (defined logit(Px) as linear function of
z1) ****/

%macro Pxz_log (x=);
```

```
title "LOGISTIC  model  for P&x|z1";
proc logistic descending data=data;
 model &x=z1 /lackfit;
  ods output  LackFitChiSq = outfit ;
  output Out=outlogP PREDICTED=pred_log;
run; title;

***Based on Hosmer-Lemeshow test P-value, decide whether the logistic
model is appropriate;
data  outfit_p(keep=HLP_p); set outfit;
      if probChiSq<0.05 then Call symput ("CallLogModSelect", "yes");
      else  Call symput ("CallLogModSelect", "no");
HLP_p =probChiSq;
run;

/***  Macro to choose polyniomial terms if the model is logistic with
higher order terms  ***/
*will run only if CallLogModSelect=yes;
%LogModSelect(x=&x);

data merge1_log ;
 merge tran1  outlogP(drop = _LEVEL_);
 g1= -log((1-Pred_log)/Pred_log);
 g2= -log((1+exp(alph_h+bet_h+gamm_h*z1))/(1+exp(alph_h+gamm_h*z1)));
 alph_strz= g1 + g2;
run;

….
%mend Pxz_log;

***if for logit(x) ~z1  HLP (lackfit)  <.05 then select polynimial for
it from z1--z10 ;

/****  Model selection (based on the Shtatland paper,but use only SC as
more restrictive criteria) *********/

%macro LogModSelect(x=);

 %if &CallLogModSelect = yes %then %do;
   title "Log Model selection for Logistic  model  for Px|z1";

 proc logistic descending data=data;
      model &x=z1 z2--z10/selection=STEPWISE  slentry=1 slstay=1
      include=1;
      ods output FitStatistics=FIT;
  run; title;

      data fit1;set fit;
       where  Criterion='SC';
      run;

      proc means data= fit1 min noprint;
      output out=minout min=minSC;
      var  InterceptAndCovariates;
      run;
      data  min_all (drop=_type_ _freq_);  *combine min and orig data
set;
            IF _N_ = 1 then set; *  (keep =minSC);
      set  fit1 (keep=step InterceptAndCovariates);
      bestset=0;
      if    InterceptAndCovariates= minSC then bestset=step+1;
      run;

      ***convert bestset to number to use for best in score selection
procedure;
      data min_allbest;
      set min_all (keep= bestset);
            if bestset NE 0 then call symput ("bestnum",bestset);
      run;

      proc logistic descending data=data; * noprint;
            model &x=z1 z2--z10/selection=SCORE  best=&bestnum
include=1;
      ods output BestSubsets = best_subsets;
    run; title;

      data best_subsets1; set best_subsets (keep= VariablesinModel
NumberOfVariables);
      by  NumberOfVariables;
    if last.NumberOfVariables; *save the last row for each unique
NumberOfVariables value;
      where NumberOfVariables=&bestnum ;
       call symput ("varlist1", VariablesinModel);
      run;

      title "New logist model";
      proc logistic descending data=data;
            model &x= z1 &varlist1/lackfit;
      run; title;
%end;
title;

%mend LogModSelect;


/***********  Estimating Px|z1 (Py|z1) using Loess regression ********/

***obtain predicted values for Px|z1;
%macro Pxz_loess (x=);

title "Loess model  for Px|z1";
proc loess data=data;
 model &x=z1/select= aicc;
ods output OutputStatistics=outloessP;
run;

proc sort data=outloessP; by z1;

*****replace outsiders with .0001 and .9999;
data out_fin1;
```

```
merge  outloessP(rename =(Pred=pred_loess) keep= Pred)
outlogP(keep=Pred_log z1);
if   pred_loess < 1E-6 then  pred_loess =1E-6;
if    pred_loess > 0.999999   then pred_loess=0.999999 ;
run;

options ps=1100 ls=100; *calculate alph_str for loess;
data merge1_loess1;
merge tran1  out_fin1 (keep=Pred_loess  Pred_log z1);
 g1= -log((1-Pred_loess)/Pred_loess);
 g2= -log((1+exp(alph_h+bet_h+gamm_h*z1))/(1+exp(alph_h+gamm_h*z1)));
 alph_strz= g1 + g2;
run;

…
%mend Pxz_loess;

/***Macro below supplied by Paul Weiss, Nov 3, 2009***/
*keep polynomials as a macro var;

%macro keeper;

data polynom; set  finalmodel;
   var="z1 "; z=z1; power=1; output;
      var="z2 "; z=z2; power=2; output;
   var="z3 "; z=z3; power=3; output;
      var="z4 "; z=z4; power=4; output;
   var="z5 "; z=z5; power=5; output;
      var="z6 "; z=z6; power=6; output;
   var="z7 "; z=z7; power=7; output;
      var="z8 "; z=z8; power=8; output;
   var="z9 "; z=z9; power=9; output;
      var="z10"; z=z10; power=10; output;
  keep var z power; run;

data variables;set polynom;
if    z ne . then call symput (var, var);
 else call symput (var, " "); proc print; run;

%let varlist = &z1 &z2 &z3 &z4 &z5 &z6 &z7  &z8 &z9 &z10 ;
%mend;


/**** linear regression model for g2/alph_strz: model selection *****/

%macro      alph_strz(m=);

 proc reg data=all outest=estA  noprint;
  title "Model selection for alph_strz";
      model alph_strz=z1 z2--z10 / selection = rsquare singular=.01
include=1;  *include - to keep linear term;
 run; quit;

**Choosing the most parsimonious model that gives an Rsquare > 0.975
**;
data estA1; set estA;
```

```
  by _IN_; if first._IN_;  *save the first row for each _IN_= # of var
in the model (w/highest R-sq);
  keep z1--z10  _IN_  _RSQ_;
run;

***find max  _RSQ_;
proc means data= estA1 noprint;
output out=maxout max=maxRSQ; var  _RSQ_;
run;

data  estA1_mean;  *combine max of RSQ from proc means and data-set;
IF _N_ = 1 then set  maxout (keep= maxRSQ);
set  estA1;
run;

data estA1_mean_fin; set estA1_mean;
if maxRSQ > .975 AND  _RSQ_ < .975 then delete;   *?? is .975 not too
low?;
else if maxRSQ <= .975 AND _IN_ NE 10  then delete;
keep z1--z10 _IN_  _RSQ_;
run;

*Select the model with smallest # of variables.
*but it has the smallest R-sq;
data finalmodel; set estA1_mean_fin; if _n_=1;
if   abs(z2) < 1.e-5 then z2='.';
if   abs(z3) < 1.e-5 then z3='.';
if   abs(z4) < 1.e-5 then z4='.';
if   abs(z5) < 1.e-5 then z5='.';
if   abs(z6) < 1.e-5 then z6='.';
if   abs(z7) < 1.e-5 then z7='.';
if   abs(z8) < 1.e-5 then z8='.';
if   abs(z9) < 1.e-5 then z9='.';
if   abs(z10) < 1.e-5 then z10='.';
drop _IN_;
proc print data=finalmodel; run;

%keeper;

data finalmodel; set finalmodel (keep= _RSQ_);
rename _RSQ_ = RSQ_alph&m; run;

****test alph_str;
proc reg data=all outest=est_alp0;
title "Final model for alph_strz vs z";
model alph_strz=z1/ rsquare;  run ;quit;

….
%mend alph_strz;


/*****  Fit the flipped model with chosen polynom covariates  ***/

%macro flip_model (x= ,y=, m=);

 title  "FLIPPED adjusted &m MODEL ";
```

```
proc logistic descending data=data;
 model &x=&y &varlist /lackfit;
  ods output OddsRatios = outOR_adj;
  ods output ParameterEstimates = outPE_adj; run;
title;

…
***save final results;
title "data log for  model &m";
data log&m&q;
merge outPE_i outPE_f   outPE_f_adj finalmodel est_alp;
bet_diff&m=bet_y&m-bet_x&m;
run; title;

%mend flip_model;


/***** Results comparison and output ****/

%macro compar;

options ls=100;

/******Combining the simulation results for correct model*****/
title1 'Results where model with Y is correct and model with Y is the
unflipped model';
data biglogC;
  set logC1-logC&nsim;
…
run;

/******Combining the simulation results for wrong model*****/
title1 'Results where model with Y is correct and model with X is the
unflipped model';
data biglogW;
  set logW1-logW&nsim ;
trueOR=exp(&betT);
…
run;

/**********************************************************************
Criteria for choosing between the unflipped estimate with Y as the
outcome  vs. the unflipped estimate with X as the outcome;
*********************************************************************/
data  log_all;
merge  biglogC (keep= bet_xC  bet_adj_xC bet_yC   SE_BET_YC
SE_BET_ADJ_XC )
biglogW (keep= bet_xW bet_yW bet_adj_yW  SE_BET_xW SE_BET_ADJ_yW );
*dif_betw  dif_betc;
label   bestbeta = chosen beta
            choose_corY= % of times correct model was chosen
            choose_corX= % of times wrong  model was chosen;

dif_betc = abs(bet_yc-bet_adj_xc);
dif_betw= abs(bet_xw-bet_adj_yw) ;
choose_corY=0;   bestbeta=bet_xw;
```

```
   if dif_betw  > dif_betc then do;
    choose_corY=1; bestbeta=bet_yc;
      end;
choose_corX=0;   bestbeta=bet_yc;
   if dif_betw  < dif_betc then do;
    choose_corX=1; bestbeta=bet_yw;
   end;
run;

title "Diff btw bet in correct model for Y/X and wrong for Y/X";
proc means data=log_all mean; * stdDev;
….
run;

%mend;
```