Sub-sampling and Weighting Approaches to Model Checking

Li Chenwei

(B.Sc. Nankai University)

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2003

Acknowledgements

I would like to take this opportunity to express my sincere appreciation to my supervisor, Associate Professor Athony Kuk for his guidance and encouragement during my study in National University of Singapore. It is really my privilege to work under his supervision. I am really grateful to him for his valuable comments and suggestions to this thesis.

I wish to express my gratitude to my family for their unfailing and endless support, encouragement and understanding. Special thanks to all the staff and all my friends in our department, who have helped me in one way or another in these two years.

April 2003

Contents

List of Tables

Summary

Motivated by the ideas of sub-sampling and sample splitting, we propose a weighting approach to model checking that exploits the systematic differences under model misspecification between the weighted, the complementarily weighted and the unweighted parameter estimates. Standard error formulae for the differences are derived and the resulting standardised differences can be used to test the adequacy of the postulated model. Unlike many tests in the literature that are designed to test the goodness of fit of a particular class of models such as logistic regression, the proposed approach is very general and can be implemented easily to test the goodness of fit of any parametric model. Marginally specified models, often used in the analysis of clustered or longitudinal data, can also be tested by weighting the quasi-likelihood or the corresponding estimating equation instead of the likelihood. Four examples are given to demonstrate the usefulness of the weighting approach to model checking. Simulation studies and asymptotic power calculation suggest that the proposed tests compare favourably with more sophisticated tests proposed recently in the literature. Keywords: Estimating equation; Lack

of fit test; Maximum likelihood estimation; Quasi likelihood; Sub-sampling with probability proportional to size; Sample splitting; Weighting.

Chapter 1

Introduction

1.1 Review of Approaches to Model Checking

Classical tests for the lack of fit of a regression model typically assume there is more than one observation at each covariate value. When there is no replication, the problem becomes more complicated and one might have to resort to tests based on near replicates (Christensen, 1989; Neill and Johnson, 1989). Alternatively, we can test the postulated model against an extended model that contains one or more extra parameters. A more recent approach to model checking is to compare the parametric fit with a nonparametric fit (Azzalini, Bowman and Hardle, 1989; Azzalini and Bowman, 1993; Firth, Glosup and Hinkley, 1991; le Cessie and van Houwelingen, 1991; Kauermann and Tutz, 1999, 2001). An issue that needs to be addressed when applying nonparametric lack of fit tests is the choice of smoothing parameter. Azzalini and Bowman (1993) and Kauermann and Tutz (2001) suggest

trying a wide range of values for the smoothing parameter h and to plot the p-values against h to get the so-called significance trace. Another difficulty is that it may not be easy to derive the exact or even asymptotic distribution of test statistics based on nonparametric regression. Thus techniques like Johnson curves (Azzalini and Bowman, 1993) or the bootstrap (Azzalini et al., 1989; Kauermann and Tutz, 2001) may have to be used to obtain p-value and to simulate reference band for the estimated curve. For Gaussian data, Fan and Huang (2001) propose to test a parametric regression function by applying the adaptive Neyman test to the residuals but it is not clear how to extend their method to the setting of generalized linear model and discrete data.

1.2 Objects and Questions

In this paper, we propose two alternative yet general approaches to testing the adequacy of any parametric model. The first one is based on the concept of sub-sampling. The basic idea is as follows. If the model is correctly specified, then the maximum likelihood estimator,MLE $\hat{\beta}$, based on the whole sample and the MLE $\hat{\beta}_s$ based on sub-sample, s, should be close as they are both consistent estimators of the true parameter β. If the model is incorrect and simple random sampling is used to select the sub-samples, the difference between $\hat{\beta}_s$ and $\hat{\beta}$ will still cancel out on the average due to the balanced nature of simple random sampling. However, if s is selected in such a way that observations with certain covariate val-

ues or certain combination of covariate values are more likely to be selected, then there will be a systematic difference between $\hat{\beta}_s$ and $\hat{\beta}$ that we can exploit. To avoid over-reliance on the choice of s, we recommend averaging over sub-samples to get $\hat{\beta}_\pi = E_\pi(\hat{\beta}_s)$, where E_π denotes expectation under the sub-sampling scheme. Thus a large difference between $E_\pi(\hat{\beta}_s)$ and $\hat{\beta}$ will imply that the model is mis-specified. Alternatively, we can take a sample splitting approach and look at the difference between $E_\pi(\hat{\beta}_s)$ and $E_\pi(\hat{\beta}_{s'})$, where $s^{'}$ is the complement of s. Further details are given in chapter 2 together with the derivation of variance formulae. The standardized difference is used to test the adequacy of the postulated model. The computational burden is our main issue with the sub-sampling approach. In practice we have to compute sub-sample estimates for a large number of sub-samples and take the average of them. Fortunately, this problem can be overcomed by the second approach, the weighting approach. Indicator variables are used to indicate which terms in the full sample estimating equation (2.1) should be retained in the sub-sample estimating equation. We then bypass sub-sampling by taking expected values of the indicator variables which are just the selection probabilities. As a consequence, we only need to solve a weighted estimating equation once instead of as many times as the numbers of sub-samples. Weighting approach enables us to launch our power study. The power of the proposed weighting approach to test the goodness of fit of a model compares favorably with that of the adaptive Neyman test for the two examples considered by Fan and Huang(2001). This power

study is done in chapter 3. The advantages of weighting approach is its simplicity and generality, which can be applied to test the goodness of fit of any parametric models, such as logistic regression model and Generalized Estimating Equation for discrete data. The lack of fit test for logistic regression data is demonstrated in chapter 4. The weighting approach also compares favorably with Kauermann and Tutz's(1999) varying coefficient approach in a simulation study. The suggested approach can even be applied to over-dispersed data or for testing a model that is specified only up to its marginal distributions by using weighted and unweighted versions of quasi-likelihood or generalized estimating equation. Data from a mouse teratology experiment (Williams, 1988) are used to demonstrate this in chapter 5. In addition, some simulations are performed to verify our results in chapter3 and chapter 4.

Chapter 2

Sub-sampling Approach

2.1 Estimating Equations and Simple Linear Regression

Our basic framework is as follows. Let x be the p-dimensioned covariate vector and y be the response variable with observed values (x_i, y_i) for i=1,...,n. They are independently distributed according to the parametric density $p(y_i; x_i, \beta)$, where β is the unknown parameter of dimension p. A general way to obtain an estimator $\hat{\beta}$ of β is by solving a set of estimating equations of the form

$$\sum_{i=1}^{n} \psi(y_i; x_i, \beta) = 0, \tag{2.1}$$

where $\psi = (\psi_1, \psi_2, \cdots, \psi_p)$ and $\psi_1, \cdots, \psi_p$ are real-valued functions of x, y and β. The estimating equation is said to be unbiased if

$$E_\beta \psi(y; x, \beta) = 0.$$

In this chapter, we focus on linear regression model and least square estimation. Let

$$y_i = x_i\beta + \epsilon_i \qquad i = 1, \cdots, n$$

where x_i is a $1 \times p$ vector of covariates including the intercept and β is a $p \times 1$ vector of regression coefficients and the errors ϵ are independently distributed as $N(0, \sigma^2)$. The least square estimator is obtained by minimizing

$$\sum_{1}^{n}(y_i - x_i\beta)^2$$

In matrix form, the least square estimator β is given by

$$\beta = (X^T X)^{-1} X^T y.$$

To test the goodness of fit of the postulated model, we argue as that if the model is correctly specified, then the estimator based on a sub-sample approach is also consistent. However, if model is misspecified and sub-sample s is obtained by sampling in such a way that observations with certain covariate values or certain combination of covariate values are more likely to be selected, there will be a systematic difference between $\hat{\beta}_s$ and $\hat{\beta}$ that we can exploit. Assume that sub-sample s is obtained by sampling without replacement with unequal selection probabilities

$$\pi_i = P(i \in s)$$

We can get the sub-sample Least Square Estimator $\hat{\beta}_s$ by minimizing

$$\sum_{i \in s}(y_i - x_i\beta)^2 \tag{2.2}$$

resulting in $\hat{\beta}_s = (X_s^T X_s)^{-1} X_s^T y_s$, where X_s is the design matrix based on sub-sample s. Thus a systematic difference between $\hat{\beta}_s$ and $\hat{\beta}$ is an indication that the model is not correct. To remove the dependence on the particular sub-sample selected and to get a more efficient estimator, we consider $\hat{\beta}_\pi = E_\pi(\hat{\beta}_s)$, where E_π denotes expectation under sub-sampling from the full sample which is considered fixed. We can calculate $\hat{\beta}_\pi$ exactly by

$$\hat{\beta}_\pi = E_\pi(\hat{\beta}_s) = \sum_1^M \pi(s_i)\hat{\beta}_{s_i}$$

where M is the number of ways of choosing a sub-sample of size n from the original sample of size N, and $\pi(s_i) \geq 0$ is the probability of selecting the ith sub-sample s_i. Alternatively, we can average $\hat{\beta}_s$ over a large number of randomly drawn sub-samples $s_1, \cdots, s_R$ to yield the approximation

$$\hat{\beta}_\pi \cong \frac{1}{R}\sum_{r=1}^R \hat{\beta}_{s_r}.$$

The standardized difference between $\hat{\beta}_\pi$ and $\hat{\beta}$ is derived by the following procedure under the null hypothesis that the postulated parametric model is correct. We begin with

$$\begin{aligned} V &= var(\hat{\beta} - \hat{\beta}_\pi) \\ &= var(\hat{\beta}) + var(\hat{\beta}_\pi) - 2cov(\hat{\beta}, \hat{\beta}_\pi) \\ &= var(\hat{\beta}) + \sum_{i=1}^M [\pi(s_i)]^2 var(\hat{\beta}_{s_i}) - 2\sum_{i=1}^M \pi(s_i) cov(\hat{\beta}, \hat{\beta}_{s_i}) \\ &\quad + \sum_{i \neq j} \pi(s_i)\pi(s_j) cov(\hat{\beta}_{s_i}, \hat{\beta}_{s_j}) \end{aligned}$$

From

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{\beta}_s = (X_s^T X_s)^{-1} X_s^T y_s$$

$$var(y) = \sigma^2 I_n$$

$$cov(Ay, By_s) = Acov(y, y_s)B^T$$

we have

$$\begin{aligned} cov(\hat{\beta}, \hat{\beta}_{s_i}) &= (X^T X)^{-1} X^T cov(y, y_{s_i}) X_{s_i} (X_{s_i}^T X_{s_i})^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

and

$$cov(\hat{\beta}_{s_i}, \hat{\beta}_{s_j}) = \sigma^2 (X_{s_i}^T X_{s_i})^{-1} \left(X_{(s_i \cap s_j)}^T X_{(s_i \cap s_j)} \right) (X_{s_j}^T X_{s_j})^{-1} \tag{2.3}$$

so V can be estimated by

$$\begin{aligned} \hat{V} &= \hat{\sigma}^2 (X^T X)^{-1} + \hat{\sigma}^2 \sum_{i=1}^{M} [\pi(s_i)]^2 (X_{s_i}^T X_{s_i})^{-1} - 2\hat{\sigma}^2 \sum_{i=1}^{M} \pi(s_i) (X^T X)^{-1} \\ &\quad + \hat{\sigma}^2 \sum_{i \neq j} \pi(s_i)\pi(s_j) (X_{s_i}^T X_{s_i})^{-1} \left(X_{(s_i \cap s_j)}^T X_{(s_i \cap s_j)} \right) (X_{s_j}^T X_{s_j})^{-1}, \end{aligned} \tag{2.4}$$

Where $\hat{\sigma}^2$ is the estimate of σ^2 under the full sample. We are now in a position to test the goodness of fit of the postulated model by looking at the standardized difference

$$\frac{\hat{\beta}_\pi - \hat{\beta}}{\sqrt{\hat{V}}} \sim N(0, 1)$$

for the scalar case. If β is a p-dimensional vector, we can apply the asymptotic normal test to the individual components of β or we can use $(\hat{\beta}_\pi - \hat{\beta})^T \hat{V}^{-1} (\hat{\beta}_\pi - \hat{\beta}) \sim \chi^2_p$. To get a big difference between $\hat{\beta}_\pi = E_\pi(\hat{\beta}_s)$ and $\hat{\beta}$ under model misspecification, we should use a sampling method that selects sub-samples with covariate configuration systematically different from the full sample configuration with large probability. We can then use one of the many existing sampling methods (Brewer and Hanif, 1983) or custom-make a new method to select sub-samples with the desired configuration with large probabilities.

2.2 Sample Splitting

A natural variation of the proposed sub-sampling test is to compute $\hat{\beta}_s$ as well as the Least Square Estimator based on $s^{'}$, the complement of s. Instead of comparing the sub-sample estimator with the full sample estimator, we can now compare the sub-sample estimator with that of its complement. This leads us to consider

$$\hat{\beta}_\pi - \hat{\beta}'_\pi = E_\pi(\hat{\beta}_{s_i}) - E_\pi(\hat{\beta}_{s_i{}'}) = \sum_{i=1}^{M} \pi(s_i)(\hat{\beta}_{s_i} - \hat{\beta}_{s_i{}'}) \tag{2.5}$$

Similarly we can get the formulae of variance,

$$\begin{aligned}
V^{'} &= var(\hat{\beta}_\pi - \hat{\beta}'_\pi) \\
&= \sum_{i=1}^{M} [\{\pi(s_i)\}^2 \{var(\hat{\beta}_{s_i}) + var(\hat{\beta_{s_i{}'}})\}] - 2 \sum_{j>i} \pi(s_i)\pi(s_j) cov(\hat{\beta}_{s_i} - \hat{\beta}_{s_i{}'}, \hat{\beta}_{s_j} - \hat{\beta}_{s_j{}'}) \\
&= \sum_{i=1}^{M} [\{\pi(s_i)\}^2 \{var(\hat{\beta}_{s_i}) + var(\hat{\beta}_{s_i{}'})\}] - 2 \sum_{j>i} \pi(s_i)\pi(s_j) \{cov(\hat{\beta}_{s_i}, \hat{\beta}_{s_j}) - cov(\hat{\beta}_{s_i}, \hat{\beta}_{s_j{}'}) \\
&\quad - cov(\hat{\beta}_{s_i{}'}, \hat{\beta}_{s_j}) + cov(\hat{\beta}_{s_i}, \hat{\beta}_{s_j{}'})\}
\end{aligned} \tag{2.6}$$

Same as full sample *vs.* sub-sample, the formula of $cov(\hat{\beta}_{s_1}, \hat{\beta}_{s_2})$ can be written specifically. From(2.5) and (2.6) we will get

$$V' \;=\; var(\hat{\beta}_\pi - \hat{\beta}'_\pi).$$

It can be estimated by

$$\begin{aligned}
\hat{V}' \;=\; & \sum_{i=1}^{M} \hat{\sigma}^2\{\pi(s_i)\}^2[(X_{s_i}^T X_{s_i})^{-1} + ({X_{s_i}}'^T X_{s_i'})^{-1}] \\
& - \sum_{i\neq j} \pi(s_i)\pi(s_j)\{(X_{s_i}^T X_{s_i})^{-1}(X_{s_i\cap s_j}^T X_{s_i\cap s_j})({X_{s_j}}^T X_{s_j})^{-1} \\
& -(X_{s_i}^T X_{s_i})^{-1}(X_{s_i\cap s_j'}^T X_{s_i\cap s_j'})({X_{s_j}}'^T X_{s_j'})^{-1} \\
& -(X_{s_i'}^T X_{s_i'})^{-1}(X_{s_i'\cap s_j}^T X_{s_i'\cap s_j})({X_{s_j}}^T X_{s_j})^{-1} \\
& +(X_{s_i'}^T X_{s_i'})^{-1}(X_{s_i'\cap s_j'}^T X_{s_i'\cap s_j'})({X_{s_j}}'^T X_{s_j'})^{-1}\}, \qquad (2.7)
\end{aligned}$$

The statistics $\dfrac{\hat{\beta}_\pi - \hat{\beta}'_\pi}{\sqrt{\hat{V}'}}$ or $(\hat{\beta}_\pi - \hat{\beta}'_\pi)^T {\hat{V}'}^{-1}(\hat{\beta}_\pi - \hat{\beta}'_\pi)$ can be used to test the goodness of fit of a model.

A more naive but closely related method is to simply partition the full sample into two halves s and $s^{'}$ in a deterministic fashion and then calculate the standardised difference between the parameter estimates $\hat{\beta}_s$ and $\hat{\beta}_{s'}$ obtained from the two halves. One drawback of this approach is that its ability to detect model misspecification depends very much on choosing the right partition. The reason we use random sub-sampling rather than a fixed subset is that we do not want to rest all our hope on a particular partition of the sample since we can never be sure that it is the best choice. It is also hoped that the act of taking expectation or averaging over sub-samples to get $\hat{\beta}_\pi = E_\pi(\hat{\beta}_s)$ and $\hat{\beta}'_\pi = E_\pi(\hat{\beta}_{s'})$ will lead to more precise estimators

and subsequently a more powerful test.

2.3 Lack of Fit Test for Normal Regression Model

The following illustrates an example of normal regression model, the dataset is about the mileage y (miles per galloon) and engine size x (displacement in cubic inches) for thirty two 1976-model automobiles, this dataset first appeared in Velleman and Hoalin(1981, p.139). Neilland Johnson first analyzed it in 1989. Firstly we can present a scatter plot(figure 2.1), which suggests that this dataset can't be fitted by a straight line regression model of y on x. But from the scatter plot we can't decide whether a quadratic regression is appropriate. We can estimate the experimental error variance with 5 degrees of freedom due to the existence of the replication of data. The classical lack of fit test of the quadratic model yields an F-value of 3.57 with 24 and 5 degrees of freedom and is not significant at the 0.05 level. In 1989 a new method was proposed by Neill and Johnson, it makes use of the information contained in near replicates. The resulting tests based on different methods of defining near replicates reject the quadratic model at level 0.05. Now we explain our method which we proposed in the previous section. First we try a naive method which divide the sample into two halves according to the criteria $x \leq 167.6$ or $x > 167.6$. The regression parameter estimates based on the two half-samples and their standardized difference are given in table 2.1.

The differences between the two half-sample estimates are not significant at level

Figure 2.1: scatter plot for the mile-age data

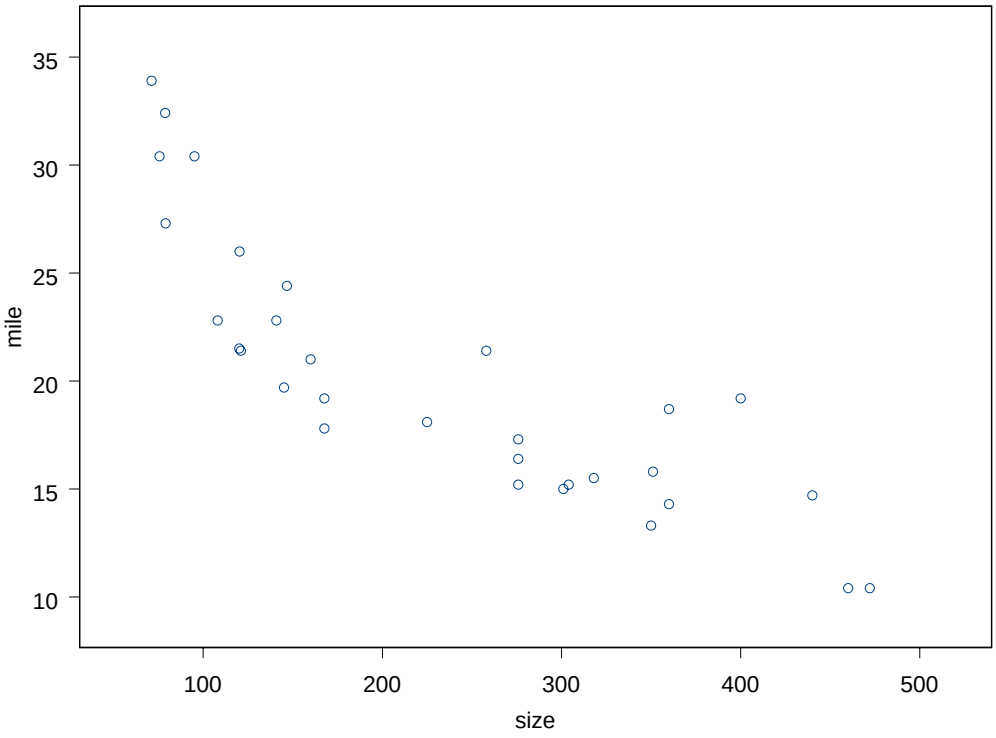


Table 2.1: SD of estimates under deterministic sample splitting

	$\hat{\beta}$	$\hat{\beta}_s$	$\hat{\beta}'_s$	$\frac{\hat{\beta}-\hat{\beta}_s}{\sqrt{\hat{V}}}$	$\frac{\hat{\beta}_s-\hat{\beta}'_s}{\sqrt{\hat{V}'}}$
Intercept	35.83	15.17	53.05	1.21	1.85
Linear	-0.105	0.0277	-0.3617	-1.36	-1.76
Quadratic	0.00126	-0.00007396	0.000975	1.48	1.26

0.05. We have explained that it's not a good idea that the test depends too much on one particular partition of the sample. As mentioned before, the sub-sampling probabilities should be unequal in order for there to be a systematic difference between $\hat{\beta}$ and $E_\pi(\hat{\beta}_s)$. We consider random sub-sampling of the full sample in such a way that observations with large x values are selected with large probability. There are many ways to do this, please see Brewer and Hanif (1983) for a list of 50 methods of sampling with unequal probabilities. The particular method we choose is ordered systematic sampling with probability proportional to size (Madow, 1949) because it can be implemented easily. The size variable used to define the sub-sampling can be the x variable itself but in this example we measure the size of x_i by its rank r_i which is invariant under increasing transformation. The use of ranks as a surrogate of size had also been advocated by Wright (1990). Another advantage of sampling with probability proportional to ranks is that it remains well defined even in cases when the x_i can take on negative values. Thus we use Madow's method to select half-samples with first order inclusion probabilities

$$\pi_i = P(i \in s) \propto r_i,$$

where the proportionality constant is $\frac{n}{\sum_{i=1}^{N} i} = \frac{2n}{N(N+1)} = \frac{1}{N+1}$ for the case of half-sampling. It is well known that there are at most N possible samples under Madow's systematic sampling method. Table 2.2 reports $\hat{\beta}_\pi$ and $\hat{\beta}'_\pi$ together with their standardized difference using the estimate derived from (2.5).

It can be seen that the standardized difference between $\hat{\beta}_\pi$ and $\hat{\beta}'_\pi$ are in the same

Table 2.2: SD of estimates under random sub-sampling

	$\hat{\beta}$	$\hat{\beta}_\pi$	$\hat{\beta}'_\pi$	$\frac{(\hat{\beta}-\hat{\beta}_\pi)}{\sqrt{\hat{V}}}$	$\frac{(\hat{\beta}_\pi-\hat{\beta}'_\pi)}{\sqrt{\hat{V}'}}$
Intercept	35.83	28.56	42.43	2.87	2.01
Linear	-0.105	-0.051	-0.185	-3.05	-2.62
Quadratic	0.000126	0.0000352	0.000314	3.20	3.28

direction but are more significant than the differences reported in Table 2.1 for the case of a fixed partition. We could get a more significant result under the sub-sampling approach than the deterministic sample splitting approach.

Chapter 3

Weighting Approach

3.1 Introduction to the Weighting Approach

3.1.1 Deriving the Estimator of Weighting Approach

Sub-sampling is a general method in model checking, but the computation is time-consuming. Because we need to compute an estimate for every possible sub-sample and average over them. Furthermore, the variance formulae are also tedious to calculate, please refer to (2.6) and (2.7). In addition, power study is also not manageable. To solve these problems, we propose a new approach, called the weighting approach. We can get the methodology of the weighting approach from Sub-sample approach intuitively. Equation (2.2) can be rewritten as

$$\sum_{i=1}^{N} I(i \in s)(y_i - \beta x_i)^2, \tag{3.1}$$

where $I(i \in s)$ is the indicator of whether observation i belongs to the sub-sample s. Instead of solving (3.1) to get $\hat{\beta}_s$ for every sub-sampling s and then take average, we take expectation of (3.1) first to get

$$\sum_{i=1}^{N} w_i (y_i - \beta x_i)^2 \tag{3.2}$$

where

$$w_i = E\{I(i \in s)\} = P(i \in s) = \pi_i.$$

The advantage of doing this is that we only need to minimize (3.2) once. The solution to minimizing (3.2) is of course just the weighted least square estimator

$$\hat{\beta}_w = (X^T W X)^{-1} X^T W y,$$

where W is the diagonal matrix of vector $(w_1, w_2, ..., w_N)$.

3.1.2 Comparing Full Sample and Weighted Estimator

We already know the full sample estimator $\hat{\beta} = (X^T X)^{-1} X^T y$, so

$$\hat{\beta} - \hat{\beta}_w = \left\{ (X^T X)^{-1} X^T - (X^T W X)^{-1} X^T W \right\} y$$

and

$$\begin{aligned} V &= var(\hat{\beta} - \hat{\beta}_w) \\ &= \left\{ (X^T X)^{-1} X^T - (X^T W X)^{-1} X^T W \right\} \sigma^2 I_N \\ &\quad \left\{ (X^T X)^{-1} X^T - (X^T W X)^{-1} X^T W \right\}^T. \end{aligned}$$

V can be estimated by

$$\begin{aligned}\hat{V} &= \left\{(X^TX)^{-1}X^T - (X^TWX)^{-1}X^TW\right\}\hat{\sigma}^2 I_N \\ &\quad \left\{(X^TX)^{-1}X^T - (X^TWX)^{-1}X^TW\right\}^T.\end{aligned}$$

It follows naturally that we can use the component wise standardized difference

$$\frac{\hat{\beta}_{w,k} - \hat{\beta}_k}{\sqrt{\hat{V}_{kk}}} \sim N(0,1), \quad k = 1, \cdots, p,$$

where $\beta_{w,k}$ denotes the kth component of $\hat{\beta}_w$, to test the goodness of fit of the assumed model. A combined test is provided by $(\hat{\beta}_w - \hat{\beta})^T\hat{V}^{-1}(\hat{\beta}_w - \hat{\beta}) \sim \chi^2_p$. It is usually more informative to conduct the componentwise tests than the combined chi-square test. The fact that some component wise tests are significant while others are not may give us some clue on the nature of model misspecification. Conversely, with a specific type of model departure in mind, we will have some idea on which components to look for difference and those component wise tests are likely to be more powerful than the combined test.

3.1.3 Comparing the Weighted and the Complementarily Weighted Estimator

We can also exploit the difference between $\hat{\beta}_w$ and $\hat{\beta}_{w'}$, where $\beta_{w'}$ is the estimator based on weights $w_i^{'} = 1 - w_i$. Note that $1 - w_i = 1 - P(i \in s) = P(i \in s^{'})$ and so $\hat{\beta}_{w'}$ is the analogue of $\hat{\beta}'_\pi = E_\pi(\hat{\beta}_{s'})$ just like $\hat{\beta}_w$ is analogous to $\hat{\beta}_\pi = E_\pi(\hat{\beta}_s)$. We have

$$\hat{\beta}'_w = (X^TW^{'}X)^{-1}X^TW^{'}y,$$

where $W^{'}$ is the diagonal matrix of vector $(w_1^{'}, w_2^{'}, \cdots, w_N^{'})$. Then

$$\hat{\beta}_w - \hat{\beta}_w^{'} = \left\{(X^T W X)^{-1} X^T W - (X^T W^{'} X)^{-1} X^T W^{'}\right\} y.$$

The standardized difference between $\hat{\beta}_w$ and $\hat{\beta}_{w'}$ is

$$\begin{aligned} V^{'} &= var(\hat{\beta}_w - \hat{\beta}_{w'}) \\ &= \sigma^2 \left\{(X^T W X)^{-1} X^T W - (X^T W^{'} X)^{-1} X^T W^{'}\right\} \\ &\quad \left\{(X^T W X)^{-1} X^T W - (X^T W^{'} X)^{-1} X^T W^{'}\right\}^T, \end{aligned}$$

it can be estimated by

$$\begin{aligned} \hat{V}^{'} &= \hat{\sigma}^2 \left\{(X^T W X)^{-1} X^T W - (X^T W^{'} X)^{-1} X^T W^{'}\right\} \\ &\quad \left\{(X^T W X)^{-1} X^T W - (X^T W^{'} X)^{-1} X^T W^{'}\right\}^T, \end{aligned}$$

Thus we can consider the componentwise standardized difference test statistics

$$\frac{\hat{\beta}_{w,k} - \hat{\beta}_{w',k}}{\sqrt{\hat{V}^{'}}}, \quad k = 1, \cdots, p,$$

where $\beta_{w,k}$ and $\beta_{w',k}$ denote the the kth components of $\hat{\beta}_w$ and $\hat{\beta}_{w'}$ respectively. Or the combined test can be provided by $(\hat{\beta}_w - \hat{\beta}_{w'})^T \hat{V^{'}}^{-1} (\hat{\beta}_w - \hat{\beta}_{w'}) \sim \chi_p^2$.

Note that the sample splitting approach considered in section(2.3) can be regarded as a special case of the weighting approach when the weights are either 0 or 1.

3.2 Quadratic Regression Examples Revisited

We have used the sub-sample approach to check the lack of fit of normal regression model with mileage data. In this section we apply the dataset again, but this time

we use weighting approach to check the goodness of fit with weights $w_i = \frac{r_i}{N+1}$. Please refer to table 3.1.

Table 3.1: SD of estimates under weighting with PPR

	$\hat{\beta}$	$\hat{\beta}_w$	$\hat{\beta}_{w'}$	$\frac{(\hat{\beta}-\hat{\beta}_w)}{\sqrt{\hat{V}}}$	$\frac{(\hat{\beta}_w-\hat{\beta}_{w'})}{\sqrt{\hat{V}'}}$
Intercept	35.83	29.49	39.53	2.74	3.43
Linear	-0.105	-0.0567	-0.146	-2.52	-3.40
Quadratic	0.000126	0.000044	0.000214	2.36	3.39

The results are more significant compared with those of the sub-sample approach. This indicates that the weighting approach is at least as effective as the sub-sample approach. Furthermore, weighting approach also avoids the cumbersome computation; it can be done in a few seconds with S-plus. We have also tried weights that are proportional to the value of x_i. The results are listed in table 3.2

Table 3.2: SD of estimates under weighting with PPS

	$\hat{\beta}$	$\hat{\beta}_w$	$\hat{\beta}_{w'}$	$\frac{(\hat{\beta}-\hat{\beta}_w)}{\sqrt{\hat{V}}}$	$\frac{(\hat{\beta}_w-\hat{\beta}_{w'})}{\sqrt{\hat{V}'}}$
Intercept	35.83	31.96	40.29	3.43494	-3.43494
Linear	-0.105	-0.072	-0.158	-3.43494	3.43494
Quadratic	1.25e-4	6.6e-5	2.5e-4	3.43494	-3.43494

An interesting observation under such weights is that the standardized differences for the intercept, linear and quadratic components all equal in magnitude. This is a consequences of the fact that we use $w_i \propto x_i$. A detailed proof of this can be found in Appendix A.

3.3 Asymptotic Power Study

3.3.1 General Theory

In addition to being less computing intensive, another advantage of the weighting approach to model checking over the sub-sampling approach is that it is more amenable to power calculation. We propose a linear regression model as follows,

$$y = X\beta + \epsilon$$

where

$$\epsilon \sim N(0, \sigma^2 I_N).$$

we have

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

and

$$\hat{\beta}_w = (X^T W X)^{-1} X^T W y,$$

where $W = diag(w_1, \cdots, w_N)$. We need a proposition to get our power function.

PROPOSITION

If $E(y) = \mu$ and $var(y) = \Sigma$, then

$$E(y^T A y) = tr(A\Sigma) + \mu^T A \mu$$

Proof:

$$\begin{aligned} y^T A y &= (y - \mu + \mu)^T A (y - \mu + \mu) \\ &= (y - \mu)^T A (y - \mu) + \mu^T A (y - \mu) + (y - \mu)^T A \mu + \mu^T A \mu \end{aligned}$$

because the expectations of the second and third item equal to 0, so

$$E(y^T A y) = E[(y - \mu)^T A (y - \mu)] + \mu^T A \mu,$$

furthermore, we have

$$E[(y - \mu)^T A (y - \mu)] = E[tr\{(y - \mu)^T A (y - \mu)\}] = tr(A\Sigma)$$

and

$$E(y^T A y) = tr(A\Sigma) + \mu^T A \mu. \P$$

Now we can get the formula of the power function. Note that when the assumed

linear model holds,

$$\hat{\beta} - \hat{\beta}_w = Dy \sim N(0, \sigma^2 DD^T),$$

where $D = (X^T X)^{-1} X^T - (X^T W X)^{-1} X^T W$. To facilitate comparison with existing omnibus goodness of fit tests, we consider the chi-square statistic

$$Q = \frac{1}{\hat{\sigma}^2} (\hat{\beta} - \hat{\beta}_w)^T (DD^T)^{-1} (\hat{\beta} - \hat{\beta}_w)$$

where

$$\hat{\sigma}^2 = \frac{(y - X\beta)^T (y - X\beta)}{N - p} = \frac{y^T [I_N - H] y}{N - p}.$$

is the usual unbiased estimator of σ^2 based on the full sample and $H = X(X^T X)^{-1} X^T$ is the hat matrix. In another word, our true model is of the form,

$$y = \mu + \epsilon$$

$$\epsilon \sim N(0, \sigma^2 I_N)$$

$$\mu \neq X\beta.$$

then $\hat{\beta} - \hat{\beta}_w = [(X^T X)^{-1} X^T - (X^T W X)^{-1} X^T W] y = Dy$ has mean $\delta = D\mu$. So $\hat{\beta} - \hat{\beta}_w \sim N(\delta, \sigma^2 DD^T)$. It follows that

$$\frac{1}{\sigma^2} (\hat{\beta} - \hat{\beta}_w)^T (DD^T)^{-1} (\hat{\beta} - \hat{\beta}_w) = \frac{\hat{\sigma}^2}{\sigma^2} Q$$

has a non-central chi-square distribution $\chi^2_p(\lambda)$ with non-centrality parameter

$$\lambda = \frac{1}{\sigma^2} \delta^T (DD^T)^{-1} \delta$$

From proposition

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{\sigma^2 tr[(I_N - H)] + \mu^T (I_N - H)\mu}{N - p} \\ &= \sigma^2 + \frac{\mu^T [I_N - H]\mu}{N - p}. \end{aligned} \tag{3.3}$$

So we get

$$\frac{E(\hat{\sigma}^2)}{\sigma^2} = 1 + \frac{\mu^T (I_N - H)\mu}{\sigma^2 (N - p)} = R$$

It follows that Q is asymptotically distributed like $R^{-1}\chi_p^2(\lambda)$ and so the asymptotic power can be calculated. The standardized differences between the kth component of the weighted and unweighted estimate is given by

$$T_k = \frac{\hat{\beta}_k - \hat{\beta}_{w,k}}{\hat{\sigma}_{c_k}}, \quad (k = 1, \cdots, p),$$

where $\hat{\sigma}_{c_k}$ is the square root of the kth diagonal element of DD^T. Its asymptotic power can be obtained from a normal distribution with shifted mean multiplied by an appropriate factor. The derivation is similar and will be omitted.

3.3.2 Simulation and Comparison

We now perform some power calculation. Consider the normal regression model

$$y_i = \mu(x_i) + e_i, \quad i = 1, \cdots, 64,$$

where the errors e_i are independently distributed as $N(0, 1)$. Suppose we are interested in testing the simple linear regression model

$$H_0 \ : \ \mu(x) = \beta_0 + \beta_1 x$$

versus

$$H_1 \ : \ \mu(x) \neq \beta_0 + \beta_1 x$$

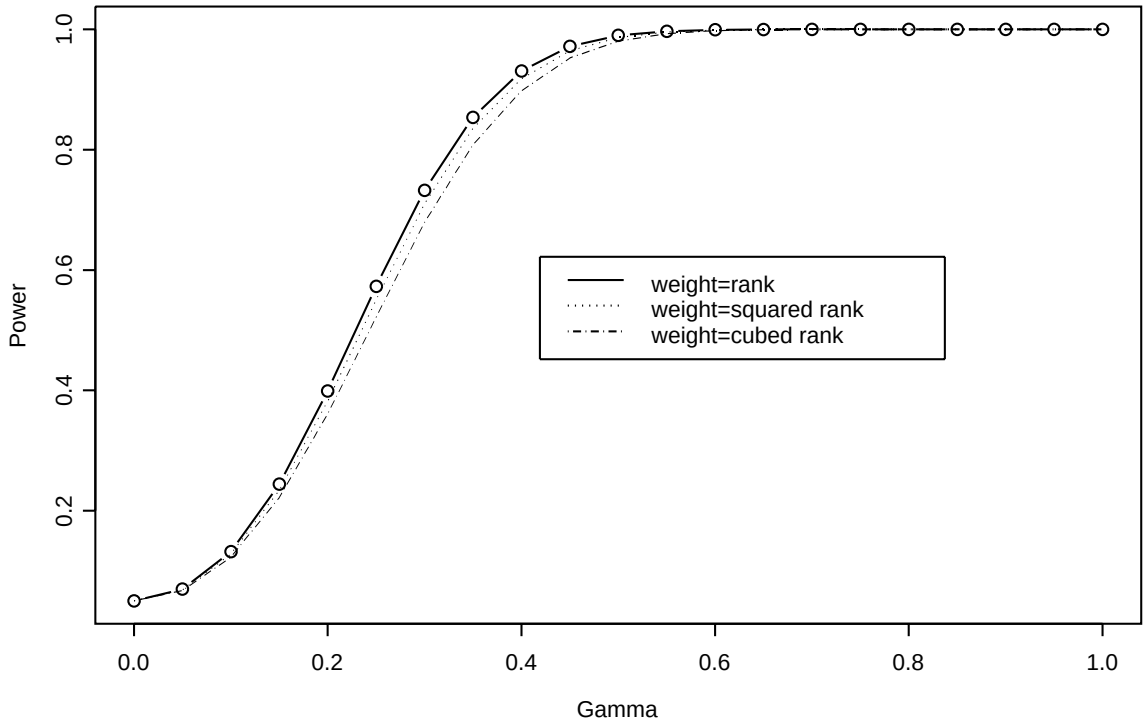
using the omnibus chi-square statistic Q with weights $w_i \propto r_i$. A nominal level of 0.05 is used. We make use of the non-central chi-square distribution to calculate the power of the test when the true mean function is $\mu(x) = 1 + \gamma x^2$ with design points generated from the uniform (-2, 2) distribution as in example 1 of Fan and Huang (2001). Since x_i can take on negative values, we choose the weights to be proportional to the ranks r_i so that observations at the upper end of the x-scale are weighted more heavily. More generally, we can consider the following class of weights

$$w_i = r_i^{\alpha}.$$

By choosing $\alpha > 1$, we will be weighting those observations with large x values more emphatically. In passing, we note that Tamura (1963) had proposed a class of rank tests based on r_i^{α} as well. The power functions of the test Q based on weights $w_i = r_i^{\alpha}$ with $\alpha = 1$,2 and 3 are given in Figure 3.1.

It appears that the choice of α does not make a wholelot difference in this example with the test based on $\alpha = 1$ performing slightly better. Comparing with Figure 1 of Fan and Huang (2001), it can be seen that the proposed test based on the weighting approach appears to be more powerful than Fan and Huang's adaptive Neyman test when the true mean function is quadratic. For example, when $\gamma = 0.2$, the proposed test statistic Q based on $\alpha = 1$ has power close to 0.4 whereas the

Figure 3.1: Power of the quadratic form goodness of fit test for simple linear regression based on the standardised difference between weighted and unweighted estimates when the true mean function is quadratic and n = 64



adaptive Neyman test only has power close to 0.2. At $\gamma = 0.4$, the proposed Q has power 0.93 compared with 0.8 for the adaptive Neyman test. In fact, the proposed test has power close to that of the F test for testing linear versus quadratic regression, which is the optimal test in this setting.

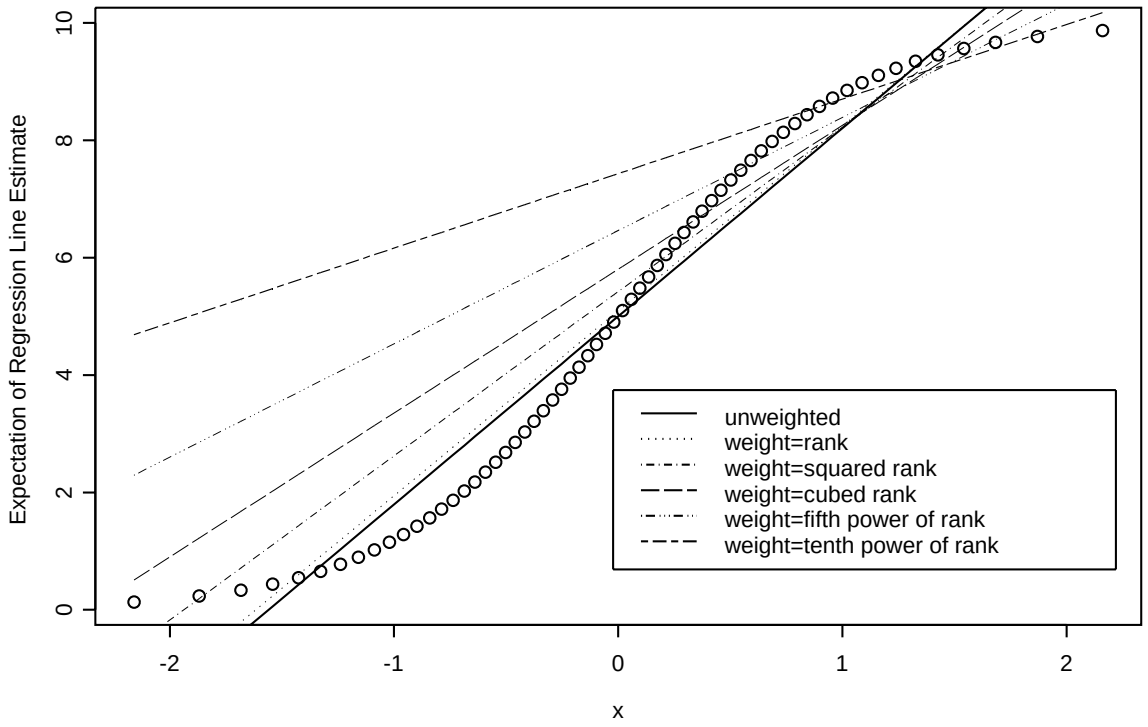
When the true mean function is logistic,

$$\mu(x) = \frac{10}{1 + \gamma exp(-2x)},$$

with design points $x_i = \Phi^{-1}(\frac{i}{65})$, $i = 1, \cdots, 64$, test Q based on $w_i = r_i^{\alpha} (\alpha = 1, 2, 3)$ is able to maintain its power above 0.98 whereas the power of the adaptive Neyman test drops to around 0.9 (Fan and Huang, 2001) when γ is near 1. To gain more insights

into how the choice of weights affects the test, let us suppose the true mean function is given by the above logistic function with $\gamma = 1$. The true means at the 64 design points are depicted by the dots in Figure 3.2.

Figure 3.2: Power of the quadratic form goodness of fit test for simple linear regression based on the standardised difference between weighted and unweighted estimates when the true mean function is logistic and n = 64



The expectation of the unweighted least squares line, $E(\hat{\beta}_0 + \hat{\beta}_1 x)$, is given by the solid line in Figure 3.2. Also shown in Figure 3.2 are the expectations of the weighted least squares line, $E(\hat{\beta}_{w,0} + \hat{\beta}_{w,1} x)$, with weights $w_i = r_i^\alpha$ and α =1,2,3,5 and 10. It can be seen that the logistic means follow more or less a straight line in the middle section but level out at both ends. In view of this, it is not surprising that as increases, so that the observations with large x values get weighted more and more heavily, the expected weighted least squares line gets rotated in the counter-clockwise direction and becomes increasily different from the unweighted line on the average. The most extreme difference occurs at $\alpha = 10$. However, the big difference between $\hat{\beta}$ and $\hat{\beta}_w$ when extreme weights such as $w_i = r_i^{10}$ are used is offset by the large variability of $\hat{\beta}_w$ and hence $\hat{\beta} - \hat{\beta}_w$ so that the test $Q = (\hat{\beta} - \hat{\beta}_w)^T \hat{V}^{-1} (\hat{\beta} - \hat{\beta}_w)$ is still most powerful if the usual weights $w_i = r_i$ are used. When the true mean function is logistic, the power of Q at $\lambda = 1$ is 0.995 if $w_i = r_i$, 0.993 if $w_i = r_i^2$, 0.984 if $w_i = r_i^3$ and drops to 0.658 if $w_i = r_i^{10}$. Our experience indicates that our original suggestion of weighting according to the ranks of x, i.e., $w_i = r_i$, often works well in practice and is a good default choice in the absence of other knowledge.

Not only does the proposed weighting approach acquitted itself well in power calculation, it has the additional advantage that it can be applied in principle to test the goodness of fit of any parametric model. In comparison, it is not clear how to extend the adaptive Neyman procedure to test, say, a logistic regression model due

to the discrete nature of the residuals.

Chapter 4

Weighted Likelihood Approach

4.1 The Likelihood Score Equation

Let the observed data be independently distributed according to the parametric density $p(y; x, \beta)$, where x_i is a covariate value and β is the unknown parameter. Consistent estimator of β can be obtained by solving estimating equation (2.1). In this section, we will concentrate on maximum likelihood estimator so that $\phi(\beta; x, y) = l^{'}(\beta; x_i, y_i)$ is the derivative of $l(\beta; y, x) = logp(\beta; x_i, y_i)$ with respect to the parameter, so (2.1) becomes

$$\sum_{i=1}^{N} l^{'}(\beta; x_i, y_i) = 0. \tag{4.1}$$

But the theory we develop can be applied to other choices of ϕ, such as those used in robust estimation or generalized estimating equations with minor modification, see chapter 5. To test the goodness of fit of the postulated model, we argue as follows.

If the model is correctly specified, then the MLE based on a random sub-sample s of size n of the original sample is also consistent. If the model is misspecified and s is obtained by sampling without replacement with unequal selection probabilities

$$\pi_i = P(i \in s),$$

then it is well known in the survey sampling literature (Pfeffermann, 1993) that the subsample MLE $\hat{\beta}_s$ obtained by solving the equation

$$\sum_{i \in s} l^{'}(\beta; x_i, y_i) = 0 \tag{4.2}$$

is in general a biased estimator of the full sample MLE $\hat{\beta}$. This has led Pfeffermann (1993) and Binder (1992) to advocate the use of pseudo score function

$$\sum_{i \in s} \frac{l^{'}(\beta; x_i, y_i)}{\pi_i} = 0,$$

where weights inversely proportional to the selection probabilities π_i are applied. Since we are interested in model checking rather than finite population inference, pseudo score function is not relevant to us.

Recall (3.1) and (3.2), we can re-write (4.2) as

$$\sum_{i=1}^{n} I(i \in s) l^{'}(\beta; x_i, y_i) = 0. \tag{4.3}$$

Same idea as chapter 3, instead of solving (4.3) to get $\hat{\beta}_s$ for every sub-sampling s and then take average, we take expectation of (4.3) first to get

$$\sum_{i=1}^{n} w_i l^{'}(\beta; x_i, y_i) = 0. \tag{4.4}$$

where

$$w_i = E\{I(i \in s)\} = P(i \in s) = \pi_i.$$

Using standard Taylor Series expansions

$$\hat{\beta} - \beta \cong -\left\{\sum_{i=1}^{n} l^{''}(\beta; x_i, y_i)\right\}^{-1} \left\{\sum_{i=1}^{n} l^{'}(\beta; x_i, y_i)\right\} \tag{4.5}$$

$$\hat{\beta}_w - \beta \cong -\left\{\sum_{i=1}^{n} w_i l^{''}(\beta; x_i, y_i)\right\}^{-1} \left\{\sum_{i=1}^{n} w_i l^{'}(\beta; x_i, y_i)\right\} \tag{4.6}$$

According

$$var(AX) = A \cdot var(X) \cdot A^T$$

$$cov(AX, BY) = A \cdot cov(X, Y) \cdot B^T,$$

we can also obtain the following asymptotic variance-covariance matrices from (4.5) and (4.6)

$$var(\hat{\beta}) = -\left\{\sum_{i=1}^{n} l^{''}(\beta; x_i, y_i)\right\}^{-1}$$

$$var(\hat{\beta}_w) = -\left\{\sum_{i=1}^{n} w_i l^{''}(x_i, \beta)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} {w_i}^2 l^{''}(\beta; x_i, y_i)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} w_i l^{''}(\beta; x_i, y_i)\right\}^{-1}$$

$$cov(\hat{\beta}, \hat{\beta}_w) = -\left\{\sum_{i=1}^{n} l^{''}(\beta; x_i, y_i)\right\}^{-1}.$$

Here we make use of the following equation

$$var(l^{'}(\beta; x_i, y_i)) = -E_\beta l^{''}(\beta; x_i, y_i) \simeq -l^{''}(\beta; x_i, y_i). \tag{4.7}$$

This is a standard well known result and its proof is follows. PROOF

Firstly from the definition of information we know

$$I(\beta) = E_\beta \left(\frac{\partial}{\partial \beta} log p(\beta; x, y)\right)^2 = \int \frac{[p^{'}(\beta; x, y)]^2}{p(\beta; x, y)} dx. \tag{4.8}$$

Secondly, since $-\frac{\partial^2}{\partial\beta^2}\int p(\beta;x,y)dx = 0$ we have

$$\begin{aligned} -E_\beta \frac{\partial^2}{\partial\beta^2} logp(\beta;x,y) &= -E_\beta\left(\frac{(p''(\beta;x,y)p(x,\beta) - [p(\beta;x,y)]^2}{p(\beta;x,y)^2}\right) \\ &= -E_\beta\left(\frac{p''(\beta;x,y)}{p(\beta;x,y)}\right) + I(\beta) \\ &= -\int p''(\beta;x,y)dx + I(\beta) \\ &= -\frac{\partial^2}{\partial\beta^2}\int p(\beta;x,y)dx + I(\beta) \\ &= I(\beta). \end{aligned} \tag{4.9}$$

Thirdly, from $-\frac{\partial}{\partial\beta}\int p(\beta;x,y)dx = 0$ we can get

$$\begin{aligned} I(\beta) &= E_\beta\left(\frac{\partial}{\partial\beta} logp(\beta;x,y)\right)^2 \\ &= E_\beta\left(\frac{\partial}{\partial\beta} logp(\beta;x,y) - E_\beta\left(\frac{\partial}{\partial\beta} logp(\beta;x,y)\right)\right)^2 \\ &= var\left(\frac{\partial}{\partial\beta} logp(\beta;x,y)\right). \end{aligned} \tag{4.10}$$

Equation(4.7) comes naturally from(4.8), (4.9) and (4.10)¶

Now we explore the difference between $\hat{\beta}$, $\hat{\beta}_w$. If the model is correct, they are both consistent estimators of β and therefore the difference is asymptotically normal with mean zero and variance-covariance matrices given by

$$\begin{aligned} V &= var(\hat{\beta} - \hat{\beta}_w) \\ &= var(\hat{\beta}_w) + var(\hat{\beta}) - cov(\hat{\beta}, \hat{\beta}_w) - cov(\hat{\beta}_w, \hat{\beta}) \end{aligned} \tag{4.11}$$

It can be estimated by

$$\hat{V} = -\left\{\sum_{i=1}^n w_i l''(\beta;x_i,y_i)\right\}^{-1} \cdot \left\{\sum_{i=1}^n {w_i}^2 l''(\beta;x_i,y_i)\right\} \cdot \left\{\sum_{i=1}^n w_i l''(\beta;x_i,y_i)\right\}^{-1}$$

$$+\left\{\sum_{i=1}^{n} l''(\beta; x_i, y_i)\right\}^{-1}. \tag{4.12}$$

Then we can get our componentweise test statistics $\frac{\hat{\beta}_k - \hat{\beta}_{w,k}}{\hat{V}}$ $k = 1, \cdots, p$ or quadratic form $(\hat{\beta} - \hat{\beta}_w)^T \hat{V}^{-1} (\hat{\beta} - \hat{\beta}_w)$ which can both be used to test the goodness of fit of our model.

Same as chapter 3 we can compare the weighted and the complementarily weighted estimator too. That is to say we can exploit the difference between $\hat{\beta}_w$ and $\hat{\beta}_{w'}$, where $w^{'}$ is the estimator based on weights $w_i^{'} = 1 - w_i$. We also have

$$1 - w_i = 1 - P(i \in s) = P(i \in s^{'}).$$

Using Standard Taylor series expansion,

$$\hat{\beta}_{w'} - \beta \cong -\left\{\sum_{i=1}^{n} w_i' l''(\beta; x_i, y_i)\right\}^{-1} \left\{\sum_{i=1}^{n} w_i' l'(\beta; x_i, y_i)\right\}.$$

Make use of (4.7) again

$$var(\hat{\beta}_{w'}) = -\left\{\sum_{i=1}^{n} w_i' l''(x_i, \beta)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} w_i^{'2} l''(\beta; x_i, y_i)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} w_i' l''(\beta; x_i, y_i)\right\}^{-1}$$

$$cov(\hat{\beta}_w, \hat{\beta}_{w'}) = -\left\{\sum_{i=1}^{n} w_i l''(x_i, \beta)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} w_i w_i' l''(\beta; x_i, y_i)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} w_i' l''(\beta; x_i, y_i)\right\}^{-1}$$

So

$$\begin{aligned} V^{'} &= var(\hat{\beta}_w - \hat{\beta}_w^{'}) \\ &= var(\hat{\beta}_w) + var(\hat{\beta}_w^{'}) - 2cov(\hat{\beta}_w, \hat{\beta}_w^{'}). \end{aligned}$$

It can be estimated by

$$\begin{aligned}\hat{V}' &= -\left\{\sum_{i=1}^{n} w_i l''(\beta; x_i, y_i)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} {w_i}^2 l''(\beta; x_i, y_i)\right\} \cdot \left\{\sum_{i=1}^{n} w_i l''(\beta; x_i, y_i)\right\}^{-1} \\ &\quad -\left\{\sum_{i=1}^{n} (1-w_i) l''(\beta; x_i, y_i)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} (1-w_i)^2 l''(\beta; x_i, y_i)\right\} \cdot \left\{\sum_{i=1}^{n} (1-w_i) l''(\beta; x_i, y_i)\right\}^{-1} \\ &\quad +2\left\{\sum_{i=1}^{n} w_i l''(\beta; x_i, y_i)\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} w_i(1-w_i) l''(\beta; x_i, y_i)\right\} \\ &\quad \cdot \left\{\sum_{i=1}^{n} (1-w_i) l''(\beta; x_i, y_i)\right\}^{-1}.\end{aligned}$$

By doing so, we get statistics $\dfrac{\hat{\beta}_{w,k} - \hat{\beta}_{w',k}}{\hat{V}'_{kk}}$ $k = 1, \cdots, p$ or $(\hat{\beta}_w - \hat{\beta}_{w'})^T \hat{V}'^{-1}_{kk} (\hat{\beta}_w - \hat{\beta}_{w'})$ which can also both be used to test the goodness of fit of our model.

4.2 Logistic Regression with Menarche Data

Next we consider a well known data set first reported by Milicer and Szczotka(1966) which is about the age of menarche in a sample of 3918 Warsaw girls who are grouped into $N = 25$ age groups. Let m_i, $i = 1, ..., 25$, denote the group size, y_i denote the number of girls in group i who have reached menarche and x_i is the mid-point of the class interval for age. It is assumed that $y_i \sim$ Binomial(m_i, p_i) independently and we would like to test the goodness of fit of the liner logistic model

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \tag{4.13}$$

The usual Pearson's chi-square test for this model has a statistic value of 26.7 with 23 degree of freedom and a p value of around 0.25. However, an examination of

the residuals shows that the fit could be improved at the lower tail. Stukel(1988) reports the results of 13 score tests of the logistic family and 5 of those tests are significant at level 0.05. More recently, Fygenson(1997) fit a decreasing odds ratio model to this data.

To see whether the same logistic model holds for different age groups, we divide the 25 binomial observations into two subsamples $s(x \geq 13.33)$ and $s^{'}(x < 13.33)$. The results of fitting logistic regression to the two sub-samples are given in table 4.1.

Table 4.1: SD of estimates of logistic regression under sample splitting

	$\hat{\beta}$	$\hat{\beta}_s$	$\hat{\beta}_{s'}$	$\frac{\hat{\beta}-\hat{\beta}_s}{\sqrt{\hat{V}}}$	$\frac{\hat{\beta}_s-\hat{\beta}_{s'}}{\sqrt{\hat{V}'}}$
Intercept	-21.23	-19.73	-23.62	-0.86	-1.56
Linear	1.632	1.522	1.828	0.90	1.63

From the table we can see the differences between the two sets of parameter estimates are not significant at level 0.05.

Instead of using a fixed partition of the data set, we consider the weighting approach with weights w_i proportional to age ordering and requires $\Sigma w_i = 12$ like in half samping. The unweighted MLE $\hat{\beta}$ as well as the weighted $\hat{\beta}_w$ and $\hat{\beta}_{w'}$ are given in the table 4.2. The variance estimates can be obtained from (4.12) with

observation information $-l''(\hat{\beta}; x_i, y_i)$ with the expected information

$$m_i\hat{p}_i(1-\hat{p}_i)\begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}.$$

where

$$\hat{p}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}.$$

in the variance formulae. It can be seen that component-wise standardized differences between $\hat{\beta}$ and $\hat{\beta}_w$, as well as between $\hat{\beta}_w$ and $\hat{\beta}_{w'}$ are statistically significant at level 0.05. Therefore the departure away from the logistic model is detected by the weighting approach.

Table 4.2: SD of estimates of logistic regression under weighting

	$\hat{\beta}$	$\hat{\beta}_w$	$\hat{\beta}_{w'}$	$\frac{\hat{\beta}-\hat{\beta}_w}{\sqrt{V}}$	$\frac{\hat{\beta}_w-\hat{\beta}_{w'}}{\sqrt{V'}}$
Intercept	-21.23	-20.31	-22.08	-2.06	2.19
Linear	1.632	1.563	1.700	2.07	-2.20

4.3 Simulation Results

Our first set of simulations is about logistic regression models. The same m_i and x_i as in the menarche data set are adopted. Firstly, we simulate 1000 datasets according to the linear logistic model (4.13) with parameters set to the MLE computed from the menarche data (i.e., $\beta_0 = -21.23$, $\beta_1 = 1.632$). The actual levels of

Table 4.3: Simulation result under logistic model

Componense-wise difference	$\frac{\hat{\beta}-\hat{\beta}_w}{\sqrt{V}}$	$\frac{\hat{\beta}_w-\hat{\beta}_{w'}}{\sqrt{V'}}$
Intercept	0.065	0.064
Linear	0.059	0.063

the tests are slightly above 0.05. Please refer to table 4.3.

Secondly we simulate 1000 data sets $y_i \sim Binomial(m_i, p_i)$, where p_i is the raw proportion from age group i in the menarche data set. We call this the separate proportion model. It appears that all powers exceed 0.7 when data are simulated from the separate proportion model. Please refer to table 4.4.

Table 4.4: Simulation result under separate proportion model

Componense-wise difference	$\frac{\hat{\beta}-\hat{\beta}_w}{\sqrt{V}}$	$\frac{\hat{\beta}_w-\hat{\beta}_{w'}}{\sqrt{V'}}$
Intercept	0.727	0.729
Linear	0.763	0.765

Kauermann and Tutz (1999) motivated the use of varying coefficient models for diagnostics in regression models with continuous and factorial covariates. We adopt their method to simulate our binary observations from the following logistic model

$$\eta \;=\; log\frac{p}{1-p} = 0.5 - 4(\mu - 0.5)^2 + x(1.5\mu - 0.5),$$

where x is the factorial regressor in $\{0, 1\}$ and μ is continuous, uniformly distributed on 20 equidistant design points from [0, 1]. At each design point, four observations

are simulated, two with $x = 0$ and two with $x = 1$. Therefore there is a total of 80 observations in each simulated sample. The same as the previous example we simulate 1000 samples. Following Kauermann and Tutz (1999), we are interested in testing the goodness of fit of the following three models,

$$M_0 : \eta = \beta_0 + \beta_\mu \mu + \beta_{\mu\mu} \mu^2 + \beta_x x + \beta_{\mu x} \mu x$$

$$M_1 : \eta = \beta_0 + \beta_\mu \mu + \beta_x x + \beta_{\mu x} \mu x$$

$$M_2 : \eta = \beta_0 + \beta_\mu \mu + \beta_x x.$$

Recall that we simulate our sample from the full model, which includes both the quadratic term and the interaction term. So M_0 is the true model. We get model M_1 by subtracting the quadratic term from M_0. Similarly, M_2 is derived by taking the interaction term away from M_1. Detailed simulation results are shown in table 4.5.

The good performance of the β_μ component test is to be expected because the erroneous omission of a quadratic term during fitting will obviously lead to a systematic difference in the estimates of the linear term depending on whether we weight the small $\mu^{'}s$ or the large $\mu^{'}s$ heavily. Tests based on the standardized differences in the β_0 component also do fairly well and reject M_1 and M_2 around 19% and 15% of the times respectively. The interaction term μx is omitted from M_2 only but not from M_1, this explains why the tests based on the standardized differences in the β_x component reject M_2 (19.5%) more often than M_1 (7.5%). The above results illustrate the comment we made towards the end of section 3.1 that with a specific

Table 4.5: Simulation result of varing coefficient model

Model	Componense-wise difference	β_0	β_μ	β_x	$\beta_{\mu x}$	$\beta_{\mu\mu}$
M_0	$\frac{\hat{\beta}-\hat{\beta}_w}{\sqrt{V}}$	0.028	0.032	0.040	0.038	0.028
M_0	$\frac{\hat{\beta}_w-\hat{\beta}_{w'}}{\sqrt{V'}}$	0.028	0.031	0.050	0.033	0.031
M_1	$\frac{\hat{\beta}-\hat{\beta}_w}{\sqrt{V}}$	0.191	0.188	0.075	0.073	
M_1	$\frac{\hat{\beta}_w-\hat{\beta}'_w}{\sqrt{V'}}$	0.185	0.184	0.074	0.076	
M_2	$\frac{\hat{\beta}-\hat{\beta}_w}{\sqrt{V}}$	0.168	0.250	0.195		
M_2	$\frac{\hat{\beta}_w-\hat{\beta}_{w'}}{\sqrt{V'}}$	0.136	0.263	0.194		

type of model departure in mind, we will have some idea on which components to look for difference. Conversely, results of the componentwise tests can give us some clue on the nature of model misspecification. Similar remarks were made by Kauermann and Tutz (1999) for their class of tests.

Chapter 5

Generalized Estimating Equation

5.1 Extension to GEE

Generally, the MLE method should only be used if the complete distribution is correctly specified. If this is not the case, misspecification may yield inconsistent estimators of the parameters. The over-dispersion often occurs due to the positive correlation within clusters in Longitudinal data and Clustered data. There is a scarcity of models for correlated discrete data. It is also deemed desirable not to make too many assumptions about higher order moments as they are hard to verify. A popular approach is to model only the marginal mean of the response together with a working specification of the variance function or correlation structure and then make use of quasi-likelihood or generalized estimating equations to estimate model parameters. These methods can be made robust to misspecification of the correlation structure in the sense that the marginal parameters can still be

estimated consistently with valid sandwich type variance estimates. Our approach to testing goodness of fit can be applied equally well to this setting. All that is required is to come up with a weighted version of the estimating equation and to derive the standard error of the difference between the weighted and the unweighted estimates. The extension will be discussed concretely. We analyze data collected from teratology experiments which are typically recorded as $m_i, y_i, x_i, i = 1, ..., N$, where m_i is the size of litter, y_i is the number of dead or malformed foetuses within litter i and x_i is a covariate vector consisting of dose level and possibly litter size as well. In a typical generalized linear model(GLM), one may assume

$$g(p_i) = x_i^T \beta,$$

where p_i is the marginal probability that a foetus in litter i is dead or malformed so that $\mu_i = E(y_i) = m_i p_i$, $g(.)$ is the link function, and x_i, β are covariate and parameter vectors of matching dimension respectively. A working specification of the variance function could be

$$var(y_i) = \phi m_i p_i (1 - p_i) = V_i,$$

where ϕ is a multiplicative over-dispersion factor. The parameter β can be estimated by solving the quasi-score equation (Wedderburn, 1974)

$$U(\beta, \phi) = \sum_{i=1}^{N} \left(\frac{\partial \mu}{\partial \beta} \right)' V_i^{-1}(\mu_i; \phi)(y_i - \mu_i(\beta)) = 0. \tag{5.1}$$

Here ϕ is estimated through methods of moments,

$$\hat{\phi} = \frac{1}{N - p} \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\mu_i(m_i - \mu_i)/m_i}.$$

Another variance function given by (Liang and Henfelt,1994),

$$V_i = \mu_i(m_i - \mu_i)(1 + (m_i - 1)\phi)/m_i,$$

can also be used for our estimation.

Given a initial estimate of β, we can estimate ϕ by solving the moment equation

$$\sum_{1}^{N}[(y_i - \mu_i)^2/\{\mu_i(m_i - \mu_i)(1 + (m_i - 1)\phi)/m_i\}] - (N - p) = 0. \tag{5.2}$$

After getting $\hat{\phi}$, solve (5.1) to get our new β. We then iterate between (5.1) and (5.2) until the estimation converge.

Two variance estimates for β are available, respectively ${\Sigma_1}^{-1}$ or ${\Sigma_1}^{-1}\Sigma_2{\Sigma_1}^{-1}$, where

$$\Sigma_1 = \sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)' V_i^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right)$$
$$\Sigma_2 = \sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)' V_i^{-1}(y_i - \mu_i)^2 V_i^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right),$$

where ${\Sigma_1}^{-1}\Sigma_2{\Sigma_1}^{-1}$ is called the robust variance estimate for $\hat{\beta}$ (Liang and Helflt, 1994). We make use of it in our weighted estimating equation. Now we explore our weighted estimator β_w. It can be solved from the weighed estimating equation

$$U_w(\beta, \phi) = \sum_{i=1}^{N} w_i\left(\frac{\partial\mu_i}{\partial\beta}\right) V_i^{-1}(y_i - \mu_i) = 0.$$

Using the standard Taylor Series expansions,

$$\hat{\beta}_w - \beta = \left\{\sum_{I=1}^{N} w_i(\frac{\partial\mu_i}{\partial\beta})' V_i^{-1}\frac{\partial\mu_i}{\partial\beta}\right\}^{-1}\left\{\sum_{i=1}^{N} w_i\frac{\partial\mu_i}{\partial\beta}V_i^{-1}(y_i - \mu_i)\right\}$$
$$\hat{\beta}_{w'} - \beta = \left\{\sum_{I=1}^{N} w_i'(\frac{\partial\mu_i}{\partial\beta})' V_i^{-1}\frac{\partial\mu_i}{\partial\beta}\right\}^{-1}\left\{\sum_{i=1}^{N} w_i'\frac{\partial\mu_i}{\partial\beta}V_i^{-1}(y_i - \mu_i)\right\}$$

We can obtain the following estimate of $var(\hat{\beta}_w - \hat{\beta})$.

$$\hat{V} = {\Sigma_1}^{-1}\Sigma_2{\Sigma_1}^{-1} + {\Sigma_{w,1}}^{-1}\Sigma_{ww,2}{\Sigma_{w,1}}^{-1} - {\Sigma_{w,1}}^{-1}\Sigma_{w,2}{\Sigma_1}^{-1} - {\Sigma_1}^{-1}\Sigma_{w,2}{\Sigma_{w,1}}^{-1},$$

where

$$\Sigma_{w,1} = \sum_{i=1}^{N} w_i \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right),$$
$$\Sigma_{w,2} = \sum_{i=1}^{N} w_i \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}(y_i - \mu_i)^2 V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right),$$

and

$$\Sigma_{ww,2} = \sum_{i=1}^{N} {w_i}^2 \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}(y_i - \mu_i)^2 V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right).$$

We also can get the following estimate of $var(\hat{\beta}_w - \hat{\beta}_{w'})$,

$$\begin{aligned}\hat{V}' = {} & {\Sigma_{w,1}}^{-1}\Sigma_{ww',2}{\Sigma_{w,1}}^{-1} + {\Sigma_{w',1}}^{-1}\Sigma_{w'w',2}{\Sigma_{w,1}}^{-1} \\ & -{\Sigma_{w,1}}^{-1}\Sigma_{ww',2}{\Sigma_{w',1}}^{-1} - {\Sigma_{w',1}}^{-1}\Sigma_{ww',2}{\Sigma_{w,1}}^{-1}\end{aligned}$$

where

$$\Sigma_{w',1} = \sum_{i=1}^{N} {w_i}' \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right),$$
$$\Sigma_{ww',2} = \sum_{i=1}^{N} w_i {w_i}' \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}(y_i - \mu_i)^2 V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right),$$

and

$$\Sigma_{w'w',2} = \sum_{i=1}^{N} {w_i}'^2 \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}(y_i - \mu_i)^2 V_i^{-1} \left(\frac{\partial \mu_i}{\partial \beta}\right).$$

And all the matrices are evaluated at $\hat{\beta}$. If the form of the marginal probability is misspecified, then $\hat{\beta}$, $\hat{\beta}_w$ and $\hat{\beta}_{w'}$ may not be estimating the same thing and we can use the standardized difference between $\hat{\beta}$ and $\hat{\beta}_{w'}$ or the standardized difference between $\hat{\beta}_w$ and $\hat{\beta}_{w'}$ to test the correctness of the link function.

5.2 Example

As an illustration, we consider the data given in Table 6.17 of Morgan (1992). The data were first presented by Williams (1988) and consist of the numbers of deaths per litter from a mouse teratology experiment with 4 dose groups. Williams (1982) fitted the model

$$log(\frac{p}{1-p}) = \beta_{d1}I_1 + \beta_{d2}I_2 + \beta_{d3}I_3 + \beta_{d4}I_4 + \beta_{m1}m + \beta_{m2}m^2.$$

m is the size of the litter, p is the probability of death given litter and dose and I_1, I_2, I_3, I_4 are indicators for the 4 dose groups. In this example we use weights proportional to litter sizes. To test the appropriateness of this quadratic model we compare the weighted against unweighted estimates of β. In addition, we also compare the weighted against the complementarily weighted estimators of β. The results are given in the table 5.1.

It is shown that none of the componentwise standardized differences are significant at level 0.05, which seem to support Williams's model.

Table 5.1:

	$\hat{\beta}$	$\hat{\beta}_w$	$\hat{\beta}_{w'}$	$\hat{V}$	$\hat{V}'$	$\frac{\hat{\beta}-\hat{\beta}_w}{\sqrt{V}}$	$\frac{\hat{\beta}_w-\hat{\beta}_{w'}}{\sqrt{V'}}$
β_{d1}	6.6016	5.429	6.402	0.561	0.948	1.047	-1.026
β_{d2}	5.917	5.377	6.261	0.563	0.970	0.96	0.911
β_{d3}	6.154	5.537	6.562	0.576	0.973	1.071	-1.053
β_{d4}	6.78	6.207	7.154	0.556	0.947	1.03	-1.00
β_{m1}	-1.288	-1.195	-1.355	0.088	0.158	-1.044	1.012
β_{m2}	0.049	0.0459	0.0523	0.033	0.0063	1.058	-1.02

If we drop the quadratic term and re-fit the model, then the model becomes

$$log(\frac{p}{1-p}) = \beta_{d1}I_1 + \beta_{d2}I_2 + \beta_{d3}I_3 + \beta_{d2}I_2 + \beta_{m1}m,$$

And there are significant differences between the weighted and unweighted estimates for every component. It appears that the proposed weighting scheme has good power in detecting the omission of the quadratic term.

The results are listed in table 5.2

There are significant differences between the weighted and unweighted estimates for every component. It appears that the proposed weighting scheme has good power in detecting the omission of the quadratic term.

Table 5.2:

	$\hat{\beta}$	$\hat{\beta}_w$	$\hat{\beta}_{w'}$	$\hat{V}$	$\hat{V}'$	$\frac{\hat{\beta}-\hat{\beta}_w}{\sqrt{V}}$	$\frac{\hat{\beta}_w-\hat{\beta}_{w'}}{\sqrt{V'}}$
β_{d1}	-0.9999	-2.181	-0.185	0.461	0.802	2.56	-2.49
β_{d2}	-1.0121	-2.11	-0.294	0.426	0.728	2.58	-2.49
β_{d3}	-1.7661	-2.05	0.1223	0.493	0.856	2.61	-2.54
β_{d4}	-3485	-1.44	0.4176	0.420	0.730	2.61	-2.55
β_{m1}	-0821	0.0048	-0.1475	0.034	0.060	-2.57	-2.53

Chapter 6

Conclusion

The main appeal of the proposed weighting approach to model checking lies in its generality and simplicity. In principle, the method can be applied to test the goodness of fit of any parametric model and its implementation requires only minor modifications to the existing codes for finding MLE and its standard error. The method can even be applied in situations where the models are only marginally specified that necessitate the use of quasi-likelihood or generalised estimating equations. Unlike goodness of fit tests based on nonparametric regression, there is no smoothing parameter to be selected and there is no need to perform bootstrap to simulate the null distribution of the test statistic. The usefulness of the proposed approach is demonstrated in three examples involving normal regression, logistic regression and marginal regression model for litter data. Simulation studies and asymptotic power calculation suggest that our simple tests compare favourably with more sophisticated tests proposed recently by Fan and Huang (2001) and

Kauermann and Tutz (1999).

Bibliography

[1] Azzalini, A. and Bowman, A. W. (1993) *On the use of nonparametric regression for checking linear relationships. J. R.* Statist. Assoc. B, 55, 549-57.

[2] Azzalini, A., Bowman, A. W. and Hardle, W. (1989) *On the use of nonparametric regression for model checking.* Biometrika, 76, 1-11.

[3] Binder, D. A. (1992) *Fitting Cox's proportional hazards models from survey data.* Biometrika, 79, 139-47.

[4] Brewer, K. R. W. and Hanif, M. (1983) *Sampling With Unequal Probabilities.* New York: Springer-Verlag.

[5] Christensen, R. R. (1989) *Lack of fit tests based on near or exact replicates.* Ann. Statist., 17, 673-83.

[6] Fan, J. and Huang, L. (2001) *Goodness-of-fit tests for parametric regression models.* J. Am. Statist. Assoc., 96, 640-51.

[7] Firth, D., Glosup, J. and Hinkley, D. V. (1991) *Model checking with nonparametric curves.* Biometrika, 78, 245-52.

[8] Fygenson, M. (1997) *A new approach in modelling a categorical response. I. Binary case.* J. Am. Statist. Assoc., 92, 322-32.

[9] Kauermann, G. and Tutz, G. (1999) *On model diagnostics using varying coefficient models.* Biometrika, 86, 119-28.

[10] Kauermann, G. and Tutz, G. (2001) *Testing generalized linear and semi-parametric models against smooth alternatives.* J. R. Statist. Assoc. B, 63, 147-66.

[11] le Cessie, S. and van Houwelingen, J. C. (1991) *A goodness-of-fit test for binary regression models, based on smoothing methods.* Biometrics, 47, 1267-82.

[12] Liang, K.Y and Hanfelt, J. (1994) *On the use of quasi-likelihood method in teratological experiments.* Biometrics, 50, 872-880.

[13] Madow, W. G. (1949) *On the theory of systrmatic sampling, II.* Ann. Math. Statist., 20, 333-54.

[14] Milicer, H. and Szczokta, F. (1966) *Age at menarche in Warsaw girls in 1965.* Human Biology, 38, 199-203.

[15] Morgan, B.J.T. (1992) *Analysis of Quantal Response Data.* London: Chapman and Hall.

[16] Neill, J. W. and Johnson, D. E. (1989) *A comparison of some lack of fit tests based on near replicates.* Commun. Statist. B, 18, 3533-70.

[17] Peter J. bickle and Kjell A. Docsum (2001) Mathemetical Statistics New Jersey: Prentice Hall

[18] Pfeffermann, D. (1993) *The role of sampling weights when modelling survey data.* Int. Statist. Rev., 61, 317-37.

[19] Stukel, T. A. (1989) *Generalized logistic models.* J. Am. Statist. Assoc., 83, 426-31.

[20] Tamura, R. (1963) *On a modification of certain rank tests.* Ann. Math. Statist., 34, 1101-1103.

[21] Velleman, P. F. and Hoaglin, D. C. (1981) *Applications, Basics and Computing of Exploratory Data Analysis.* Boston: Duxbury.

[22] Williams, D. A. (1988) *Extra-binomial variation in toxicology.* In Proceedings of the XIV th International Biometric Conference, pp.301-13.

[23] Wright, T. (1990) *Probability proportional to size (PPS) sampling using ranks.* Commun. Statist.-Theory Meth., 19, 347-362.

Appendix

Appendix A: Proof for the fact that the componentwise standardized differences $\frac{\hat{\beta}_{w,k} - \hat{\beta}_k}{\sqrt{\hat{V}}}$. have equal magnitude when the weights w_i are proportional to x_i

Recall

$$\hat{\beta}_w = (X^T W X)^{-1} X^T W y$$

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

where

$$X = \begin{pmatrix} 1 & x_1 & {x_1}^2 \\ 1 & x_2 & {x_2}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & {x_n}^2 \end{pmatrix}$$

and

$$W = \begin{pmatrix} x_1 & & 0 \\ & x_2 & \\ & & \ddots \\ 0 & & x_n \end{pmatrix}.$$

If we let

$$A = [(X^T W X)^{-1} X^T W - (X^T X)^{-1} X^T] = \begin{pmatrix} a_1^T \\ a_2^T \\ a_3^T \end{pmatrix},$$

then, we have

$$(\hat{\beta}_w - \hat{\beta})^2 = A y y^T A^T = \begin{pmatrix} a_1^T y y^T a_1 & a_1^T y y^T a_2 & a_1^T y y^T a_3 \\ a_2^T y y^T a_1 & a_2^T y y^T a_2 & a_2^T y y^T a_3 \\ a_3^T y y^T a_1 & a_3^T y y^T a_2 & a_3^T y y^T a_3 \end{pmatrix}$$

and

$$V = \sigma^2 \begin{pmatrix} a_1^T a_1 & a_1^T a_2 & a_1^T a_3 \\ a_2^T a_1 & a_2^T a_2 & a_2^T a_3 \\ a_3^T a_1 & a_3^T a_2 & a_3^T a_3 \end{pmatrix}.$$

So our issue becomes to prove

$$\frac{a_1^T y y^T a_1}{a_1^T a_1} = \frac{a_2^T y y^T a_2}{a_2^T a_2} = \frac{a_3^T y y^T a_3}{a_3^T a_3}$$

for all y. By letting y take on $(1, 0, \cdots, 0)^T$, $(0, 1, 0, \cdots, 0)$,$\cdots$,$0, \cdots, 0, 1$, we obtain the necessary condition.

$$\frac{a_{1i}^2}{\| a_1 \|} = \frac{a_{2i}^2}{\| a_2 \|} = \frac{a_{3i}^2}{\| a_3 \|},$$

for $i = 1, \cdots, n$. This means that our task becomes that of proving

$$a_1 \propto a_2 \propto a_3,$$

Which of course is also a sufficient condition for $(A.1)$ to hold. We can get

$$X^T W X = \begin{pmatrix} \sum X_i & \sum X_i^2 & \sum X_i^3 \\ \sum X_i^2 & \sum X_i^3 & \sum X_i^4 \\ \sum X_i^3 & \sum X_i^4 & \sum X_i^5 \end{pmatrix}.$$

Let $\sum X_i = a$, $\sum X_i^2 = b$, $\sum X_i^3 = c$, $\sum X_i^4 = d$ and $\sum X_i^5 = e$,

we have

$$(X^T W X)^{-1} X^T W = \frac{1}{F_1} \begin{pmatrix} x_1(ce - d^2) + x_1^2(cd - be) + x_1^3(bd - c^2), & \cdots \\ x_1(cd - be) + x_1^2(ae - c^2) + x_1^3(bc - ad), & \cdots \\ x_1(bd - c^2) + x_1^2(bc - ad) + x_1^3(ac - b^2), & \cdots \end{pmatrix}$$

and

$$(X^T X)^{-1} X^T = \frac{1}{F_2} \begin{pmatrix} (bd - c^2) + x_1(bc - ad) + x_1^2(ac - b^2), & \cdots \\ (bc - ad) + x_1(nd - b^2) + x_1^2(ab - nc), & \cdots \\ (ac - b^2) + x_1(ad - nc) + x_1^2(nb - a^2), & \cdots \end{pmatrix},$$

where

$$F_1 = -c^3 + 2bcd - ad^2 - b^2 e + ace,$$

and

$$F_2 = -b^3 + 2abe - nc^2 - a^2 d + nbd.$$

They are the determinants of X^TWX snf X^TX respectively.

So $A = (X^TWX)^{-1}X^TW - (X^TX)^{-1}X^T$ has the following form.

$$\frac{1}{F_1F_2}\begin{pmatrix} -F_1(bd-c^2)+\{F_2(ce-d^2)-F_1(bc-ad)\}x_1+\{F_2(cd-be) & \\ -F_1(ac-b^2)\}x_1^2+F_2(bd-c^2)x_1^3 & \cdots \\ -F_1(bc-ad)+\{F_2(cd-be)-F_1(nd-b^2)\}x_1+\{F_2(ae-c^2) & \\ -F_1(ab-nc)\}x_1^2+F_2(bc-ad)x_1^3 & \cdots \\ -F_1(ac-b^2)+\{F_2(bd-c^2)-F_1(ab-nc)\}x_1+\{F_2(bc-ad) & \\ -F_1(nb-a^2)\}x_1^2+F_2(ac-b^2)x_1^3 & \cdots \end{pmatrix}.$$

We want to prove that the row vectors of (A.2) are proportional to one another.

In the first row, the coefficient of x_1 is

$$\begin{aligned} & \frac{1}{F_1\cdot F_2}[(-b^3+2abc-nc^2-a^2d+nbd)(ce-d^2) \\ & \qquad -(-c^3+2bcd-ad^2-b^2e+ace)(bc-ad)] \\ = \quad & \frac{1}{F_1\cdot F_2}(bd-c^2)(b^2d+nce-nd^2-bc^2+acd-abc). \end{aligned}$$

In the second row the coefficient of x_1 is

$$\begin{aligned} & \frac{1}{F_1\cdot F_2}[(-b^3+2abc-nc^2-a^2d+nbd)(cd-bc) \\ & \qquad -(-c^3+2bcd-ad^2-b^2e+ace)(nd-b^2)] \\ = \quad & \frac{1}{F_1\cdot F_2}(bc-ad)(b^2d+nce-nd^2-bc62+acd-abc). \end{aligned}$$

The coefficient of x_1^2 in the first row is

$$\frac{1}{F_1\cdot F_2}[(-b^3+2abc-nc^2-a^2d+nbd)(cd-be)$$

$$-(-c^3 + 2bcd - ad^2 - b^2e + ace)(ac - b^2)]$$

$$= \frac{1}{F_1 \cdot F_2}(bd - c^2)(b^2c + a^2e + ncd - nbe - abd - ac^2).$$

The coefficient of x_1^2 in the second row is

$$\frac{1}{F_1 \cdot F_2}[(-b^3 + 2abc - nc^2 - a^2d + nbd)(ae - c^2)$$

$$-(-c^3 + 2bcd - ad^2 - b^2e + ace)(ab - nc)]$$

$$= \frac{1}{F_1 \cdot F_2}(bc - ad)(b^2c + a^2encd - nbe - abd - ac^2).$$

Similarly we can get the coefficients of the constant items and the cubic items. We can observe that the two vectors a_1^T, a_2^T of A only differ in the multipliers, which are $bd - c^2$ for row 1 and $ac - b^2$ for row 2. From the above discussion we conclude that

$$\frac{a_1^T}{bc - ad} = \frac{a_2^T}{bd - c^2}$$

or

$$a_2^T = \frac{bd - c^2}{bc - ad}a_1^T$$

Similiarly, we can show that the multipier for the third row is $ac - b^2$. So

$$\frac{a_3^T}{=}\frac{bc - ad}{ac - b^2}bc - ada_2^T$$

Appendix B: Selective Code in S-Plus2000

Mile Age Example

```
#####################################################################
```

```
## First we input tmpsize and mile from dataset or other files ###
########      The dataset's name is mileagedata      #############
#######################################################################
attach(mileagedata)
qfit <- lm(formula = mile ~ size + size^2,na.action = na.exclude)
qfit.sigma <- summary(qfit)$sigma
qfit.coef <- summary(qfit)$coef
w1 <-   15*size/sum(size)
w2 <- 1-w1
x<- cbind(1,size,size^2)
WXTX1 <- t(x)%*%diag(w)%*%x
w <- w1^2
WXTX2 <- t(x)%*%diag(w)%*%x
w <- w2
WXTX3 <- t(x)%*%diag(w)%*%x
w <- w2^2
WXTX4 <- t(x)%*%diag(w)%*%x
w <- w1*w2
WXTX5 <- t(x)%*%diag(w2)%*%diag(w1)%*%x
var1 <- solve(WXTX1)%*%WXTX2%*%solve(WXTX1)
var2 <- solve(WXTX3)%*%WXTX4%*%solve(WXTX3)
var3 <- solve(WXTX1)%*%WXTX5%*%solve(WXTX3)
qfit1 <- lm(formula = mile ~size+size^2,data = mileagedata, weight = w1)
qfit1.coef1 <- summary(qfit1)$coef

qfit2 <- lm(formula = mile ~ size + size^2,data = mileagedata, weight=w_2)
qfit2.coef2 <- summary(qfit2)$coef
var <- t(x)%*%x
var <- solve(var)

final <- qfit.sigma^2*(var1+var2-var3-t(var3))
a <- qfit1.coef1[,1]
b <- qfit2.coef2[,1]
(b-a)/sqrt(diag(final))
```

Marnache Example

```
Mens.girl <- c(0,0,0,2,2,5,10,17,16,29,39,51,47,67,81,88,79,90,113,95,117,107,92,112,1049)
All.girl <- c(376,200,93,120,90,88,105,111,100,93,100,108,99,106,105,117,98,97,120,102,
            122,111,94,114,1049)
age <- c(9.21,10.21,10.58,10.83,11.08,11.33,11.58,11.83,12.08,12.33,12.58,12.83,13.08,13.33,
13.58,13.83,14.08,14.33,14.58,14.83,15.08,15.33,15.58,15.83,17.58)

Com.Mens.girl <- All.girl - Mens.girl
SF <- cbind(Mens.girl,Com.Mens.girl)
options(contrats=c("contr.treatment","contr.poly"))
qfit <- glm(SF~age,family=binomial)
Beta <- summary(qfit)$coef[,1]
w1 <- 12*(c(1:25))/sum(c(1:25))
qfit1 <- glm(SF~age,family=binomial,w=w1)
```

```
w2 <- 1-w1
qfit2 <- glm(SF~age,family=binomial,w=w2)

temp <- exp(Beta[1]+Beta[2]*age)
x <- rbind(1,age)
P <- temp/(1+temp)

var1 <- 0
var1 <- var1 + x%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
var.pop <- solve(var1)

v<- summary(qfit)$cov.unscale
var1 <- 0
var2 <- 0
var1 <- var1 + x%*%diag(w1)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
var2 <- var2 + x%*%diag(w1)%*%diag(w1)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
var.pop1 <- solve(var1)%*%var2%*%solve(var1)

var1 <- 0
var2 <- 0
var1 <- var1 + x%*%diag(w2)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
var2 <- var2 + x%*%diag(w2)%*%diag(w2)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
var.pop2 <- solve(var1)%*%var2%*%solve(var1)

var1 <- 0
var2 <- 0
var3 <- 0
var1 <- var1 + x%*%diag(w1)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
var2 <- var2 + x%*%diag(w2)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
var3 <- var3 + x%*%diag(w1)%*%diag(w2)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
var.pop3 <- solve(var1)%*%var3%*%solve(var2)

var <- var.pop1 + var.pop2 - var.pop3 - t(var.pop3)
var <- var.pop1-var.pop
expect <- summary(qfit)$coef
expect1 <- summary(qfit1)$coef
(expect[, 1] - expect1[, 1])/diag(sqrt(var))
```

GEE Example

```
all.num <- c(10,11,12,4,10,11,9,11,10,10,12,10,8,11,6,9,
            14,12,11,13,14,10,12,13,10,14,13,4,8,13,12,10,3,13,12,14,9,
            13,16,11,4,1,12,8,11,14,14,11,3,13,9,17,15,2,14,8,6,17)
r.num   <- c(1,4,9,4,10,9,9,11,10,7,12,9,8,9,4,7,14,7,9,8,5,10,10,8,10,
3,13,3,8,5,12,1,1,1,0,4,2,2,1,0,0,0,0,0,1,0,1,0,0,0,2,2,0,0,1,0,0,0)
hb <- c(4.1,3.2,4.7,3.5,3.2,5.9,4.7,4.7,3.5,4.8,4.3,4.1,3.2,6.3,4.3,
        3.1,3.6,4.1,4.8,4.7,4.8,6.7,5.2,4.3,3.9,6.3,4.4,5.2,3.9,7.7,
        5.0,8.6,11.1,7.2,8.8,9.3,9.3,8.5,9.4,6.9,8.9,11.1,9.0,11.2,11.5,
        12.6,9.5,9.8,16.6,14.5,15.4,14.5,14.6,16.5,14.8,13.6,14.5,12.4)
sf <- cbind(r.num,all.num-r.num)
qfit <- glm(sf~hb,family=binomial)
```

```
##################
fai <- 0
stopsign <- 10
beta <- c(0,0)
beta1 <- c(0,0)
while(stopsign > 1e-5){
    beta1 <- beta
    qfit <- glm(sf~hb,family=binomial,weight=(1+(all.num-1)*fai))
    beta <- qfit$coef
    temp <- exp(beta[1]+beta[2]*hb)
    pi<-temp/(1+temp)

    cubic <- function(z,pi,m,y){
        a <- (y-m*pi)^2
        b <- m*pi*(1-pi)*((1+(m-1)*z))
        c <- 56
        sum(a/b)-c
    }
    cubic(0,pi,all.num,r.num)
    tmp<- uniroot(cubic,lower=0, upper=10,pi=pi,m=all.num,y=r.num)
    fai <- tmp$root
    stopsign <- sum(abs(beta1-beta))
}

w1 <- 29*(hb)/sum(hb)
c <- ifelse(w1 >1,1,w1)
w1 <- c
w2 <- 1 - w1

qfit1 <- glm(sf~hb,family=binomial,weight=w1*(1+(all.num-1)*fai))
qfit2 <- glm(sf~hb,family=binomial,weight=w2*(1+(all.num-1)*fai))
beta <- qfit$coef
beta1 <- qfit1$coef
beta2 <- qfit2$coef

temp <- exp(beta[1]+beta[2]*hb)
pi<-temp/(1+temp)

beta
fai

######### variance #########
var1 <- rbind(1,hb)%*%diag(all.num*pi*(1-pi)/(1+(all.num-1)*fai))%*%cbind(1,hb)
var2 <- rbind(1,hb)%*%diag(((r.num-all.num*pi)/(1+(all.num-1)*fai))^2)%*%cbind(1,hb)
var.beta <- solve(var1)%*%var2%*%solve(var1)

var1 <- rbind(1,hb)%*%diag(w1)%*%diag(all.num*pi*(1-pi))%*%cbind(1,hb)
var2 <- rbind(1,hb)%*%diag(w1^2)%*%diag(((r.num-all.num*pi))^2)%*%cbind(1,hb)
var.beta1 <- solve(var1)%*%var2%*%solve(var1)
```

```
var1 <- rbind(1,hb)%*%diag(w2)%*%diag(all.num*pi*(1-pi))%*%cbind(1,hb)
var2 <- rbind(1,hb)%*%diag(w2^2)%*%diag(((r.num-all.num*pi))^2)%*%cbind(1,hb)
var.beta2 <- solve(var1)%*%var2%*%solve(var1)

var1 <- rbind(1,hb)%*%diag(w1)%*%diag(all.num*pi*(1-pi))%*%cbind(1,hb)
var2 <- rbind(1,hb)%*%diag(w1*w2)%*%diag(((r.num-all.num*pi))^2)%*%cbind(1,hb)
var3 <- rbind(1,hb)%*%diag(w2)%*%diag(all.num*pi*(1-pi))%*%cbind(1,hb)

var.inter.12 <- solve(var1)%*%var2%*%solve(var3)

var1 <- rbind(1,hb)%*%diag(w1)%*%diag(all.num*pi*(1-pi))%*%cbind(1,hb)
var2 <- rbind(1,hb)%*%diag(w1)%*%diag(((r.num-all.num*pi))^2)%*%cbind(1,hb)
var3 <- rbind(1,hb)%*%diag(all.num*pi*(1-pi))%*%cbind(1,hb)
var.inter.1.all <- solve(var1)%*%var2%*%solve(var3)

var1 <- rbind(1,hb)%*%diag(w2)%*%diag(all.num*pi*(1-pi))%*%cbind(1,hb)
var2 <- rbind(1,hb)%*%diag(w2)%*%diag(((r.num-all.num*pi))^2)%*%cbind(1,hb)
var3 <- rbind(1,hb)%*%diag(all.num*pi*(1-pi))%*%cbind(1,hb)
var.inter.2.all <- solve(var1)%*%var2%*%solve(var3)

ret1 <-(beta1-beta2)/diag(sqrt(var.beta1+var.beta2-var.inter.12-t(var.inter.12)))
ret2 <-(beta-beta1)/diag(sqrt(var.beta1+var.beta-var.inter.1.all-t(var.inter.1.all)))
ret3 <-(beta-beta2)/diag(sqrt(var.beta1+var.beta-var.inter.2.all-t(var.inter.2.all)))
```

Simulation about Menarche girl data

```
orig.Mens.girl <- c(0,0,0,2,2,5,10,17,16,29,39,51,47,67,81,88,79,90,113,95,117,
                    107,92,112,1049)
All.girl <- c(376,200,93,120,90,88,105,111,100,93,100,108,99,106,105,117,98,97,
            120,102,122,111,94,114,1049)
age <- c(9.21,10.21,10.58,10.83,11.08,11.33,11.58,11.83,12.08,12.33,12.58,12.83,
13.08,13.33,13.58,13.83,14.08,14.33,14.58,14.83,15.08,15.33,15.58,15.83,17.58)
orig.Com.Mens.girl <- All.girl - orig.Mens.girl
SF <- cbind(orig.Mens.girl,orig.Com.Mens.girl)
options(contrats=c("contr.treatment","contr.poly"))
qfit <- glm(SF~age,family=binomial)
summary(qfit)
Beta <- summary(qfit)$coef[,1]

main <-function(Mens.girl){
    Com.Mens.girl <- All.girl - Mens.girl
    SF <- cbind(Mens.girl,Com.Mens.girl)
    options(contrats=c("contr.treatment","contr.poly"))
    qfit <- glm(SF~age,family=binomial)
    summary(qfit)
    Beta <- summary(qfit)$coef[,1]
    w1 <- 11*(age-9.21)/sum((age-9.21))
    w1 <- 5*(c(25:1)^4)/sum(c(25:1)^4)
    w1 <- 12*(c(1:25))/sum(c(1:25))
    qfit1 <- glm(SF~age,family=binomial,w=w1)
    w2 <- 1-w1
```

```
    qfit2 <- glm(SF~age,family=binomial,w=w2)
    temp <- exp(Beta[1]+Beta[2]*age)
    x <- rbind(1,age)
    P <- temp/(1+temp)

    var1 <- 0
    var1 <- x%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var.pop <- solve(var1)

    var1 <- 0
    var2 <- 0
    var1 <- x%*%diag(w1)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var2 <-  x%*%diag(w1)%*%diag(w1)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var.pop1 <- solve(var1)%*%var2%*%solve(var1)
    var1 <- 0
    var2 <- 0
    var1 <- x%*%diag(w2)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var2 <- x%*%diag(w2)%*%diag(w2)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var.pop2 <- solve(var1)%*%var2%*%solve(var1)
    var1 <- 0
    var2 <- 0
    var3 <- 0
    var1 <- x%*%diag(w1)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var2 <- x%*%diag(w2)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var3 <- x%*%diag(w1)%*%diag(w2)%*%diag(All.girl)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var.pop3 <- solve(var1)%*%var3%*%solve(var2)
    #var <- var.pop2 - var.pop
    var <- var.pop1 + var.pop2 - var.pop3 - t(var.pop3)
    expect1 <- summary(qfit1)$coef
    expect2 <- summary(qfit2)$coef
    ret.value <- (expect1[, 1] - expect2[, 1])/diag(sqrt(var))
}

Pi.vec <- exp(Beta%*%rbind(1,age))/(1+exp(Beta%*%rbind(1,age)))

j <- 0
k <- 0
set.seed(1)
for (i in 1:500){
    ret.value <- main(rbinom(rep(1,25),All.girl,Pi.vec) )
    print(ret.value)
    if(abs(ret.value[1]) > 5.991) j<-j+1
    if(abs(ret.value[1]) > 1.96) j<-j+1
    if(abs(ret.value[2]) > 1.96) k<-k+1
}
j
k
```

Simulation about varying coefficient data

```
equi.point <- rep(seq(1:20)/20,rep(4,20))
```

```
factor <- rep(c(0,0,1,1),20)
temp <- exp(0.5 - 4*(equi.point-0.5)^2 + factor*(-0.5+1.5*equi.point))
Pi.vec <- temp/(1+temp)
k <- rep(0,5)
l <- rep(0,5)
m <- rep(0,5)
random.num <- rep(0,80)
for(i in 1:100){
    random.num <- rbinom(rep(1,80),rep(1,80),Pi.vec)
    fit <- glm(random.num ~ equi.point+equi.point^2+factor+factor*equi.point,family=binomial)
    fit.coef <-summary(fit)$coef
    expect <- fit.coef
    Beta <- fit.coef[,1]
    w1 <- rep(10*seq(1:20)/sum(c(1:20)),rep(4,20))
    w2 <- 1-w1

    fit1 <- glm(random.num ~ equi.point+equi.point^2+factor+factor*equi.point,
    family=binomial,weight=w1)
    fit1.coef <-summary(fit1)$coef
    expect1 <- fit1.coef

    fit2 <- glm(random.num ~ equi.point+equi.point^2+factor+factor*equi.point,
    family=binomial,weight=w2)
    fit2.coef <-summary(fit2)$coef
    expect2 <- fit2.coef

    temp <- exp(Beta[1]+Beta[2]*equi.point+Beta[3]*equi.point^2+Beta[4]*factor+Beta[5]
    *factor*equi.point)

    x <- rbind(1,equi.point,equi.point^2,factor,factor*equi.point)


    P <- temp/(1+temp)
    var1 <- x%*%diag(P)%*%diag(1-P)%*%(t(x))
    var.pop <- solve(var1)
    var1 <- 0
    var2 <- 0
    var1 <- var1 + x%*%diag(w1)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var2 <- var2 + x%*%diag(w1)%*%diag(w1)%*%diag(P)%*%diag(1-P)%*%t(x)
    var.pop1 <- solve(var1)%*%var2%*%solve(var1)

    var1 <- 0
    var2 <- 0
    var1 <- var1 + x%*%diag(w2)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var2 <- var2 + x%*%diag(w2)%*%diag(w2)%*%diag(P)%*%diag(1-P)%*%t(x)
    var.pop2 <- solve(var1)%*%var2%*%solve(var1)
    var1 <- 0
    var2 <- 0
    var3 <- 0
    var1 <- var1 + x%*%diag(w1)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var2 <- var2 + x%*%diag(w2)%*%diag(P)%*%diag(1-P)%*%(t(x))
```

```
    var3 <- var3 + x%*%diag(w1)%*%diag(w2)%*%diag(P)%*%diag(1-P)%*%(t(x))
    var.pop3 <- solve(var1)%*%var3%*%solve(var2)
    var <- var.pop1 + var.pop2 - var.pop3 - t(var.pop3)
    ret.value <- (expect1[, 1] - expect2    [, 1])/diag(sqrt(var))
    if(ret.value[1]>1.96) k[1]<-k[1]+1
    if(ret.value[2]>1.96) k[2]<-k[2]+1
    if(ret.value[3]>1.96) k[3]<-k[3]+1
    if(ret.value[4]>1.96) k[4]<-k[4]+1
    if(ret.value[5]>1.96) k[5]<-k[5]+1
    var <- var.pop1 - var.pop
    ret.value <- (expect1[, 1] - expect[, 1])/diag(sqrt(var))

    if(ret.value[1]>1.96) l[1]<-l[1]+1
    if(ret.value[2]>1.96) l[2]<-l[2]+1
    if(ret.value[3]>1.96) l[3]<-l[3]+1
    if(ret.value[4]>1.96) l[4]<-l[4]+1
    if(ret.value[5]>1.96) l[5]<-l[5]+1

    var <- var.pop2 - var.pop
    ret.value <- (expect2[, 1] - expect[, 1])/diag(sqrt(var))

    if(ret.value[1]>1.96) m[1]<-m[1]+1
    if(ret.value[2]>1.96) m[2]<-m[2]+1
    if(ret.value[3]>1.96) m[3]<-m[3]+1
    if(ret.value[4]>1.96) m[4]<-m[4]+1
    if(ret.value[5]>1.96) m[5]<-m[5]+1

}
```