Diagonistic effectiveness of residuals

Zidong Liu, Xuyan Lu

December 10, 2017

Question of Interest

In the linear regression setting, we can always use residuals to make some diagnostic inference. We may also want to do something to diagonse generalized lienar model. So whether the residuals is effective becomes a question of interest and we put forward the following three questions:

1. Is residuals plot a good diagnostic tool?
2. Is marginal residual plots useful in high dimensional model?
3. Is there any substitution for residuals?

Our discussion will study the residuals of generalized model from following two ways.

1. The shape of fitted residuals against the x value.
2. The empirical distribution of diagnostic statistics.

Part 1: Residual plot analysis

Some concerns for simulation

In this section, we may wish to really look at the shape of residuals in some good fits and bad fits.

So we choose to generate data ourselves.

Before generating the data, there are several problems which we need to consider.

1. The statistical model

There are many possible statistical model we can build a model upon. Here we focus on the logistic model which is binary outcome.

2. The dimension of the model

The dimension are really important in determining the true model. we have to determine the number of predictors involved. To make simulation easier and feasible, in this report, we will only limit the dimension to be less than 3.

3. The functional form of the model

The model's function form is another important aspect. Different forms may lead to different shape of residuals. This is what we want to study. However, we are unable to enumerate all possible caess. So we only consider some basic cases.

4. The distribution of predictors

There are many possible distributions we can generate our data, but here we just focus on normal distribution and uniform distribution.

Simulation process

1. Specify the true model

Here we use a example to illustrate the way we generate the data. $logit(\pi(x)) = \alpha + \beta x$

2. Give the sampling distribution

We simply assume that $X_i \sim N(0, 1)$, and generate 100 i.i.d. x_i from this distribution.

3. Get the probability sequence for binomial trial

Calculate the probability sequence via the following formula:

$\pi_i = \frac{exp(\alpha+\beta x_i)}{1+exp(\alpha+\beta x_i)}$

4. Generate outcome from binomial trial

$Y_i \sim Binomial(1, \pi_i)$

Simulation and analysis

Single predictor simulation

Possible model specification:

1. Linear model: $logit(\pi(x)) = \alpha + \beta x$
2. Quadratic model: $logit(\pi(x)) = \alpha + \beta_1 x + \beta_2 x^2$
3. Exponential increasing model: $logit(\pi(x)) = \alpha + \beta exp(x)$
4. Periodic model: $logit(\pi(x)) = \alpha + \beta sin(ax + b)$

The given model specification is the true settings, however we wish to use lienar model to detect the model misspecification. So in the following simulation, we will compare the effect of a linear model and a perfectly matched model.

Case 1:linear model

Here since linear model has no comparison, we simply give two linear model. One is a strong linear model and the other stands for a weak linear relationship.

Strong linear model

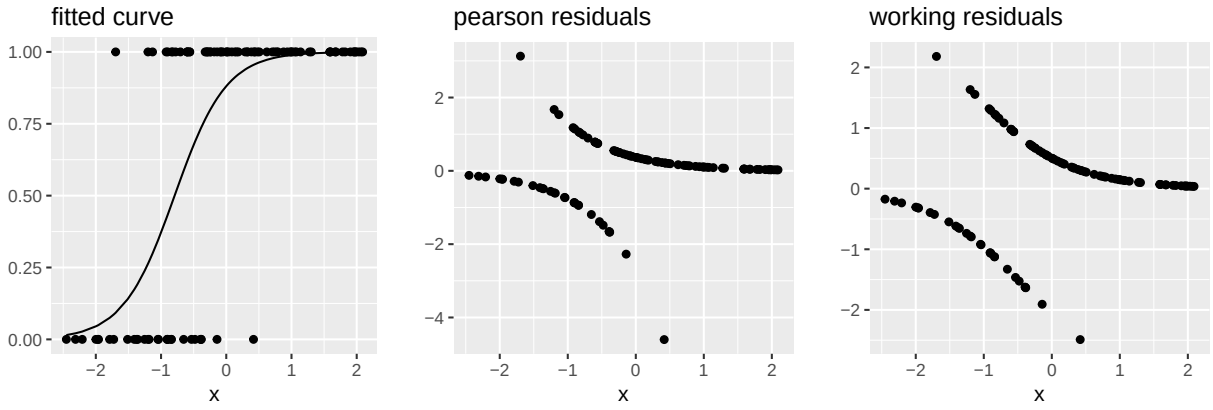
True model: $logit(\pi(x)) = 2 + 3x$

Distribution: $X \sim N(0, 1)$

Number of observations: $N = 100$

Fitted linear model: $logit(\pi(x)) = 1.999 + 2.5x$

Fitted curve and the residuals plots are given below:



Weak linear model

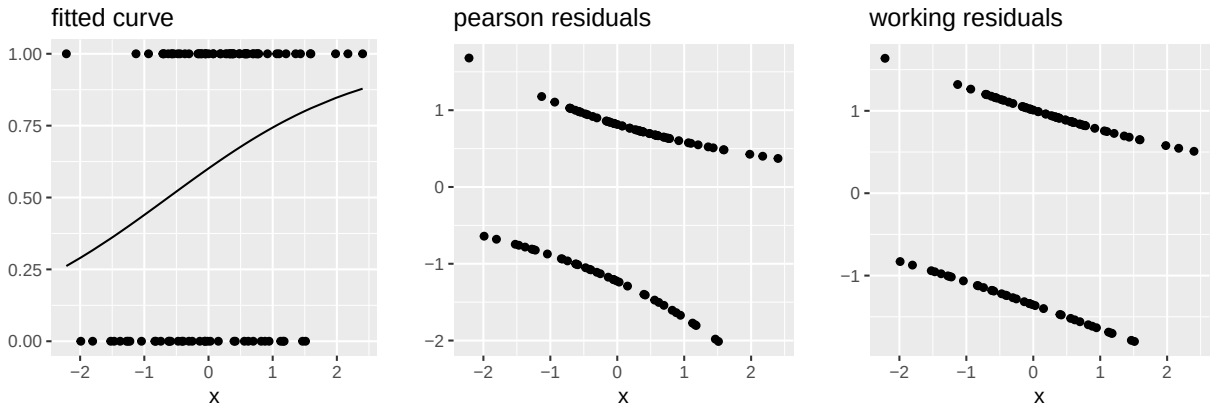
True model: $logit(\pi(x)) = 0.1 + 0.2x$

Distribution: $X \sim N(0, 1)$

Number of observations: $N = 100$

Fitted linear model: $logit(\pi(x)) = 0.4 + 0.65x$

Fitted curve and the residuals plots are given below:



Some discussion for the comparison

First of all, we can look at the bias between the fitted model and true model.

The strong linear model tends to have a smaller estimation bias. This is quite natural since the weak linear model has only weak information of the relationship, which would increase the variance of the estimation.

Then, from the residuals plots, we can see the strong linear model tends to have strong curvature while the weak linear model tends to have a straight line relation between residuals and x.

Case 2:quadratic model

True model: $logit(\pi(x)) = 0.1 + 0.2x + 2x^2$

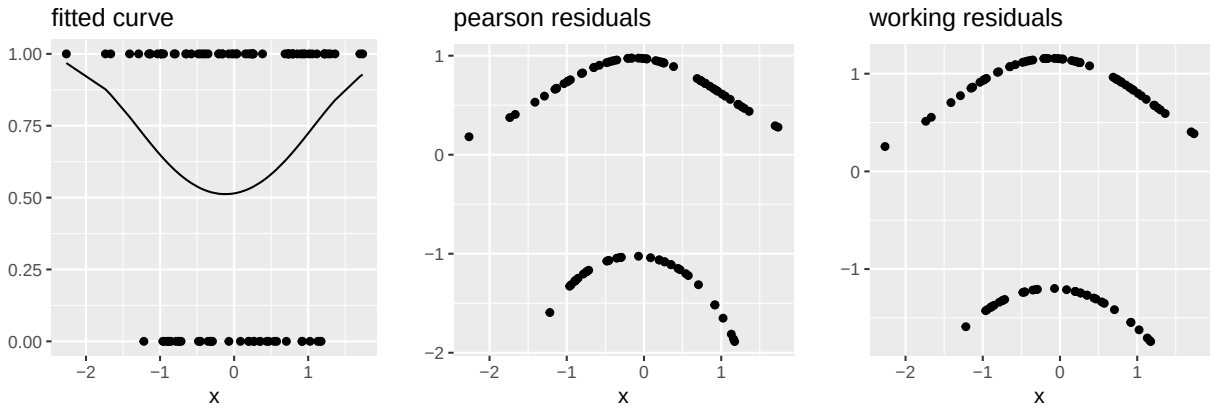
Distribution: $X \sim N(0, 1)$

Number of observations: $N = 100$

Fit a matched model

Fitted quadratic model: $logit(\pi(x)) = 0.05 + 0.17x + 0.7x^2$

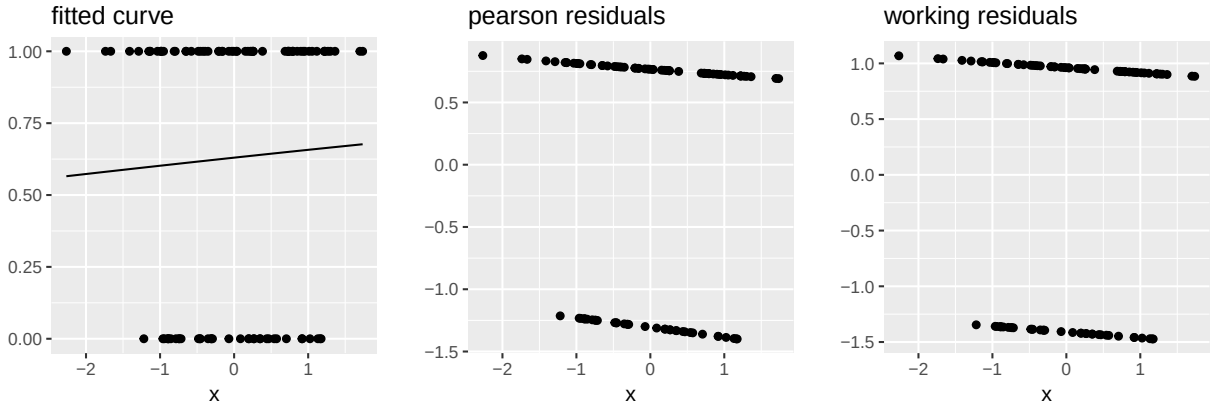
Fitted curve and the residuals plots are given below:



Fit a linear model

Fitted linear model: $logit(\pi(x)) = 0.53 + 0.11x$

Fitted curve and the residuals plots are given below:



Some discussion for the comparison

The matched quadratic model gives a quadratic fitted curve and a quadratic residuals plot while the linear fit residuals plot is a straight line.

From the previous comparison of strong and weak lienar model, we can say the linear fit is weak and it can not give any useful suggestion for the model mis-spacification.

Case 3:Exponential model

True model: $logit(\pi(x)) = -0.5 + 2exp(x)$

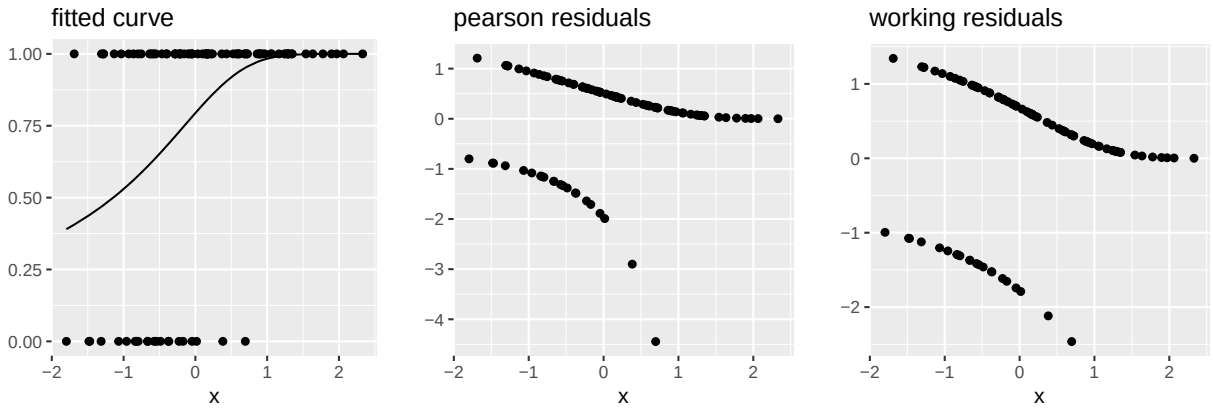
Distribution: $X \sim N(0, 1)$

Number of observations: $N = 100$

Fit a matched model

Fitted exponential model: $logit(\pi(x)) = -0.02 + 0.36x + 1.37exp(x)$

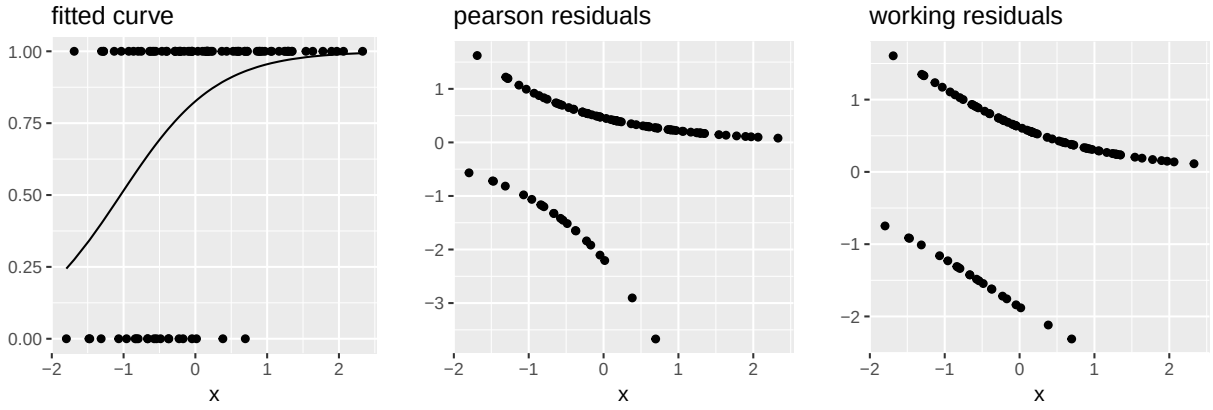
Fitted curve and the residuals plots are given below:



Fit a linear model

Fitted linear model: $logit(\pi(x)) = 1.55 + 1.49x$

Fitted curve and the residuals plots are given below:



Some discussion for the comparison

The exponential model looks interesting, the shape of plots in matched model fit and linear fit looks similar.

This may suggest the similarity between linear setting and exponential setting.

One thing we should note, $exp(\eta) > 0$ while the $\eta \in (-\infty, \infty)$.

The lower bound of the exponential model fitted probability is around 0.5, while that of the linear model could go far below 0.5.

So if we can find some area where the two outcomes are happening with the same frequency, this may suggest a local exponential model form.

But, in terms of the diagonistic effectiveness of linear model, we can still say nothing.

Case 4:Periodic model

True model: $logit(\pi(x)) = 1.5sin(3x)$

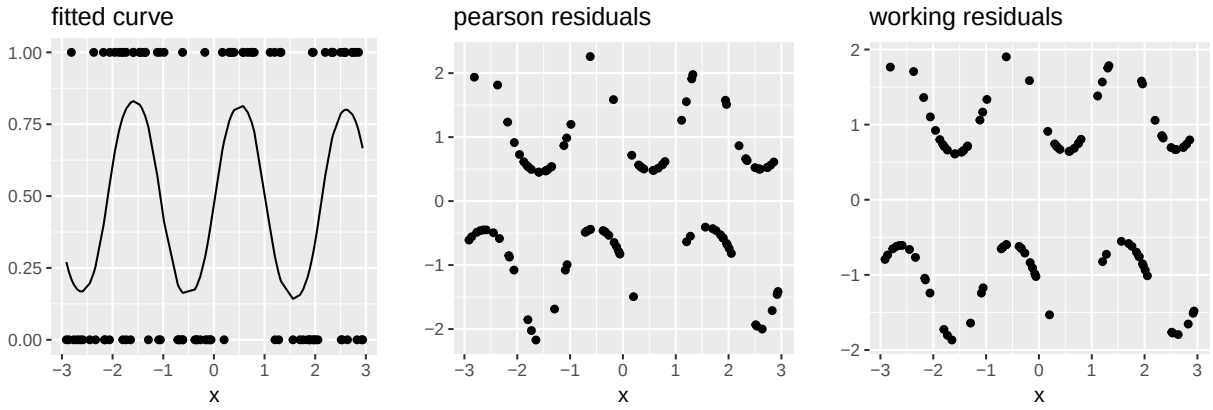
Distribution: $X \sim U(-3, 3)$

Number of observations: $N = 100$

Fit a matched model

Fitted periodic model: $logit(\pi(x)) = 0.09 - 0.385x + 1.03sin(3x)$

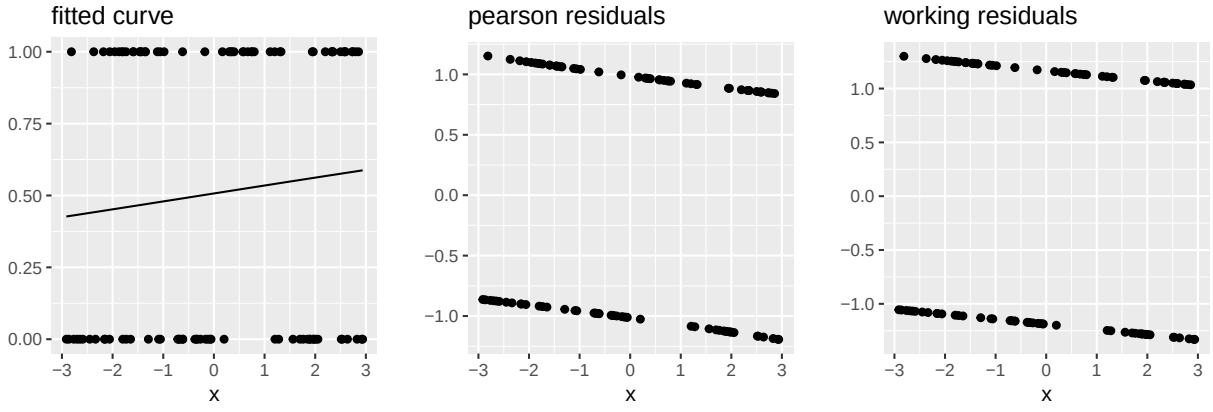
Fitted curve and the residuals plots are given below:



Fit a linear model

Fitted linear model: $logit(\pi(x)) = 0.009 + 0.33x$

Fitted curve and the residuals plots are given below:



Some discussion for the comparison

The linear model has a linear shape which suggests weak linear relationship.

The fitted periodic model has a periodic pattern.

The residuals of the linear model can not give any useful guidance for the model mis-specification.

Some conclusion for single predictor case

1. The shape of the fitted curve looks like the form of the fitted model in general, especially when the true specification is not linear.
2. If the residuals plot of a linear fit has a linear shape, this may suggest the true model is probably none-linear.
3. In general, the residuals plto of linear fit can not give suggestion for the true model when the true model is in other form.

Two predictors simulation

The model specification of two predictors has even more possibilities. Also, we only consider the most common cases.

Model specification under consideration:

1. Linear model: $logit(\pi(x)) = \alpha + \beta_1 x_1 + \beta_2 x_2$
2. Additive quadratic model: $logit(\pi(x)) = \alpha + \beta_1 x_1 + c_1 x_1^2 + \beta_2 x_2 + c_2 x_2^2$
3. Interaction model: $logit(\pi(x)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

Here, we will plot the residuals against each predictors, which is the so-called marginal residual plot.

Case 1:linear model

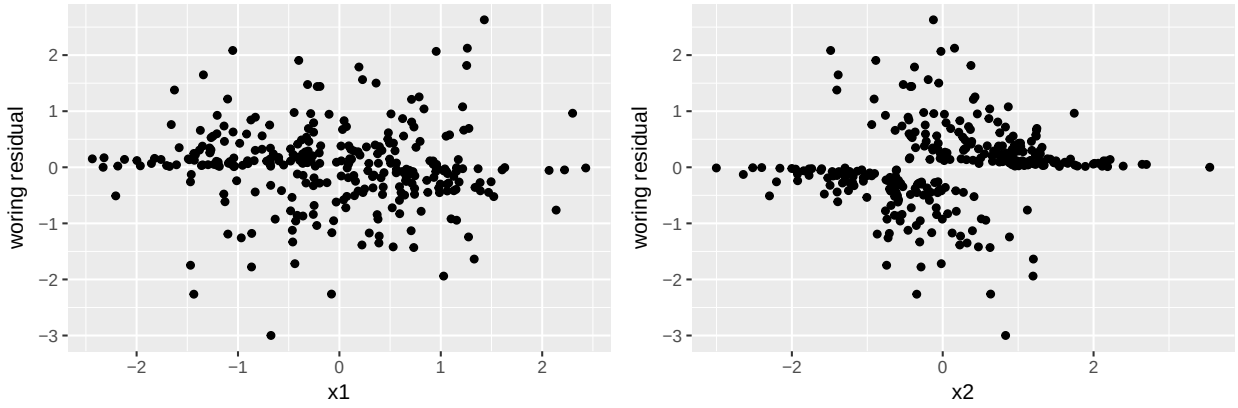
True model: $logit(\pi(x)) = 0.1 - 2x_1 + 3x_2$

Distribution: $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, X_1 independent of X_2

Number of observations: $N = 300$

Fitted linear model: $logit(\pi(x)) = 0.26 + 2.3x_1 + 3.19x_2$

The two marginal residuals plot are given below:



Some discussion for marginal plots

The pattern in the 2-d marginal residual plots are not clear. There is no clear line or curve in the plots.

Case 2:Additive quadratic model

True model: $logit(\pi(x)) = 0.1 + 1.2x_1 - x_1^2 + 2x_2 + x_2^2$

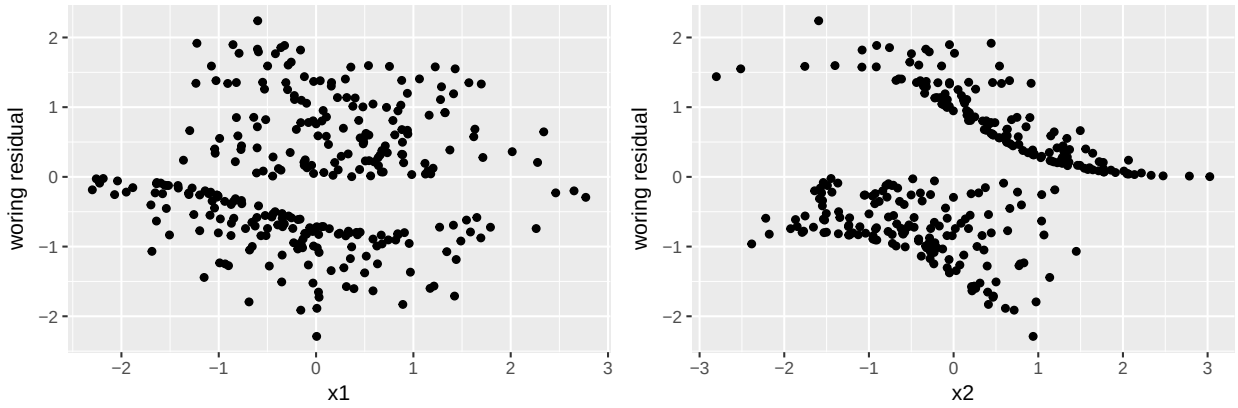
Distribution: $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, X_1 independent of X_2

Number of observations: $N = 300$

Fit a matched model

Fitted quadratic model: $logit(\pi(x)) = 0.15 - 1.23x_1 - 0.87x_1^2 + 1.94x_2 + 0.62x_2^2$

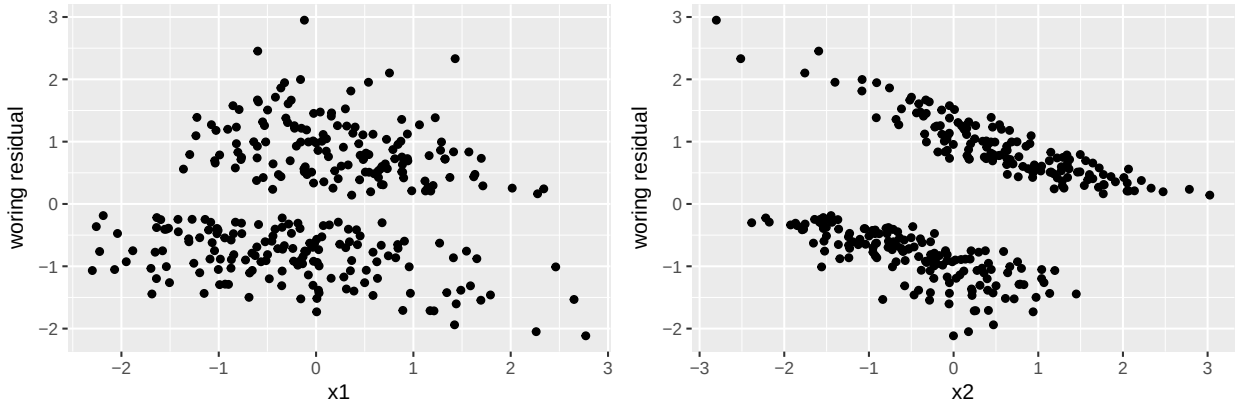
The two marginal residuals plot are given below:



Fit a linear model

Fitted linear model: $logit(\pi(x)) = -0.135 + 0.82x_1 + 1.46x_2$

The two marginal residuals plot are given below:



Some discussion for marginal plots

1. The marginal plots of matched model preserves the quadratic shape in each dimension.
2. The marginal plot of x2 has linear trend while the one for x1 even loses the linear shape.
3. Diagonistic effect of linear model residuals is poor.

Case 3:Interaction model

True model: $logit(\pi(x)) = 0.6 + 0.2x_1 - 0.5x_2 + 3x_1x_2$

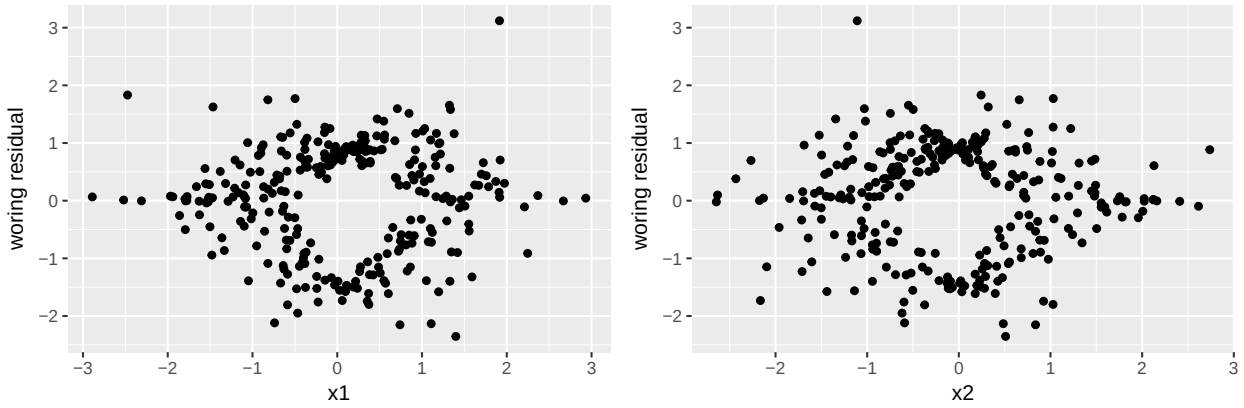
Distribution: $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, X_1 independent of X_2

Number of observations: $N = 300$

Fit a matched model

Fitted interaction model: $logit(\pi(x)) = 0.675 + 0.11x_1 - 0.43x_2 + 2.9x_1x_2$

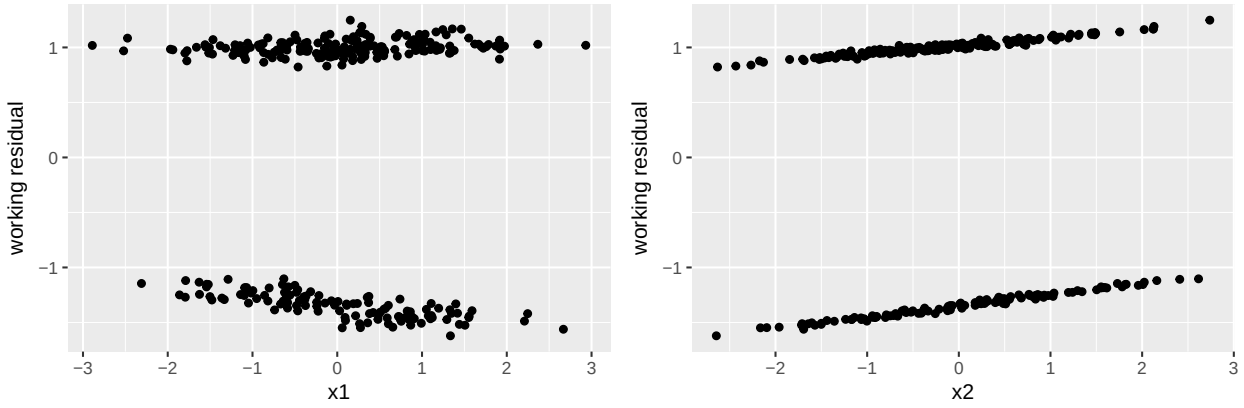
The two marginal residuals plot are given below:



Fit a linear model

Fitted linear model: $logit(\pi(x)) = 0.39 + 0.048x_1 - 0.21x_2$

The two marginal residuals plot are given below:



Some discussion for the marginal plots

1. The weird shape of the matched model is hard to explain. Actually, the 2-d shape of the fitted curve is a saddle shape which makes the marginal shape quite strange.
2. The linear model fit shows strong linear trend in each dimension. However, it still can not give any useful information about the model misspecification.

Conclusion from the Simulation and Analysis

1. Residual plots will preserve the shape of the fitted curve which is specified by the fitted model.
2. Some matched model will also have some strange marginal residuals plot.
3. The residuals of logistic model has no power in detecting the model mis-specification.

Part 2: The Distribution of Diagnostic Statistics

When we have a logistic regression model:

$$logit(\pi(x)) = \alpha + \beta^T x, \text{ where: } x = (x_1, \cdots, x_p)^T, \beta = (\beta_1, \cdots, \beta_p)^T$$

we always want to test the goodness of fit of a it, and we can construct some test statistic to do hypothesis test. Here we consider two statistics that we used in our text book (*Categorical Data Analysis*): Pearson statistic (X^2) and likelohood-ratio ststistic (G^2), the formula is below:

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \to X^2 = \sum_{i=1}^{n} r_i^2$$

$$G^2 = D(y; \hat{\pi}) = -2[L(\hat{\pi}; y) - L(y; y)]$$

where $L(\hat{\pi}; y)$ is the maximum likelihood for model and $L(y; y)$ is the maximum likelihood for saturated model.

When we performing the good of fit test, we always assume the test statistic (Pearson and Likelihood-ratio) following a χ^2_{n-p} distribution, where p is the number of parameters we need to estimate in our model (if we set p as the number of variables, the degree of freedom should be $n - (p + 1)$). But we are curious whether this assumption is reasonable, in another words, what is the true distribution of Pearson and likelihood-ration statistic. Our research is based on simulation (we generate data by ourself) so we know the true model, and we are going to see the behaviour of test ststistic under different model specification.

Comparison of True Empirical Distribution and Chi-squared Distribution

We generated our 2 dimentional input variable X from a multinomial normal distribution:

$$X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix})$$

Then we generate each $logit(\pi_i) = f(X_i)$ by a specific formula, and after we got the sequence $\{\pi_i\}$, we can generate each Y_i by sample once from binomial distribution with $\mu = \pi_i$, so finally we can get our $\{Y_i\}$ sequence. So now we have our two dimentional input variable sequence: $\{X_i\}$ and our response variable sequence: $\{Y_i\}$, and we also know the true relationship between π_i and X_i: $logit(\pi_i) = f(X_i)$.

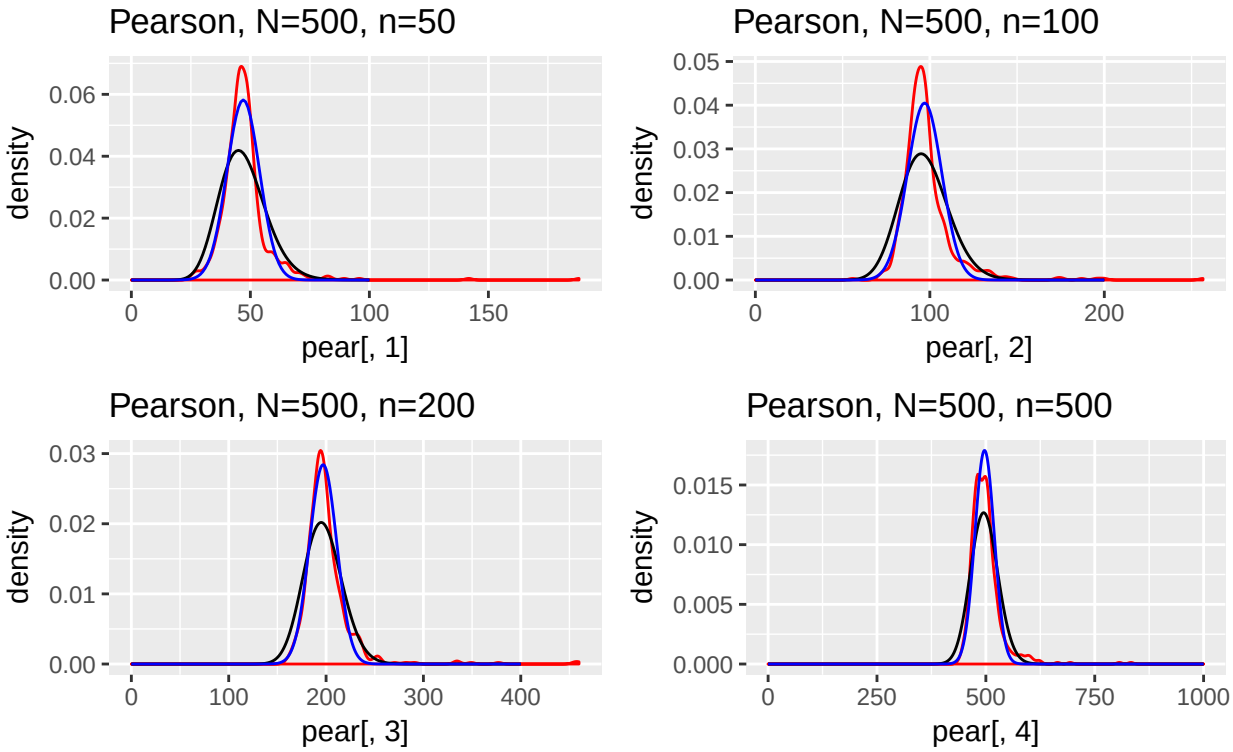
We set the number of obervations n (length of $\{(X_i, Y_i)\}_n$ sequence) to be 50, 100, 200 and 500 to see whether sample size have a effect on the distribution of test statistics. We generated our $\{(X_i, Y_i)\}_n$ sequence 500 times, each time we fit a logistic model with right model specification: $Y \sim X_1 + X_2$ (by R function *glm()*) and calculate the value of the test statistic, our true expirical distribution is given by these 500 values of test statistic.

In the first case we set the relationship between π_i and X_i to be:

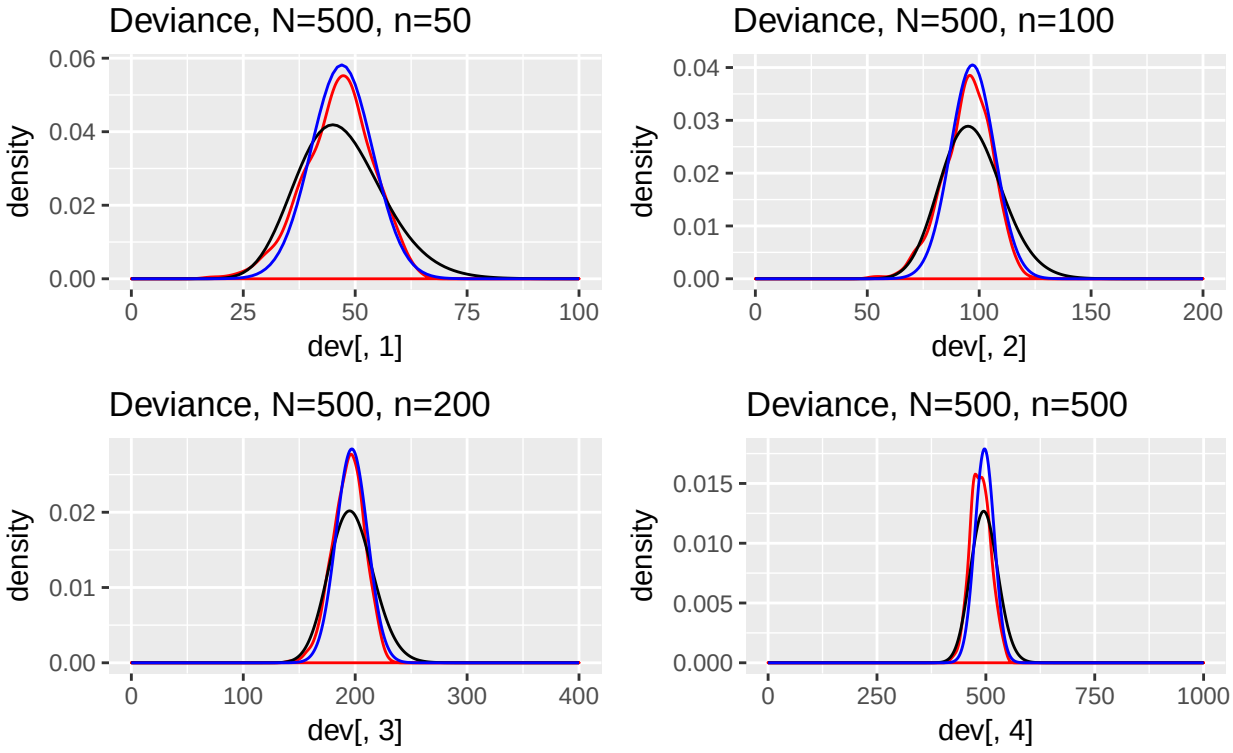
$$logit(\pi) = 1 + x_1 + x_2$$

and we compared the ture distribution of test statistic with the χ^2_{n-p} distribution, where $p = 3$ because we have 3 parameters to be estimated: x_1, x_2 and interception, $n = 50, 100, 200, 500$. The density plot is shown in the plot below:

Pearson Statistic:



Likelihood ratio Statistic:



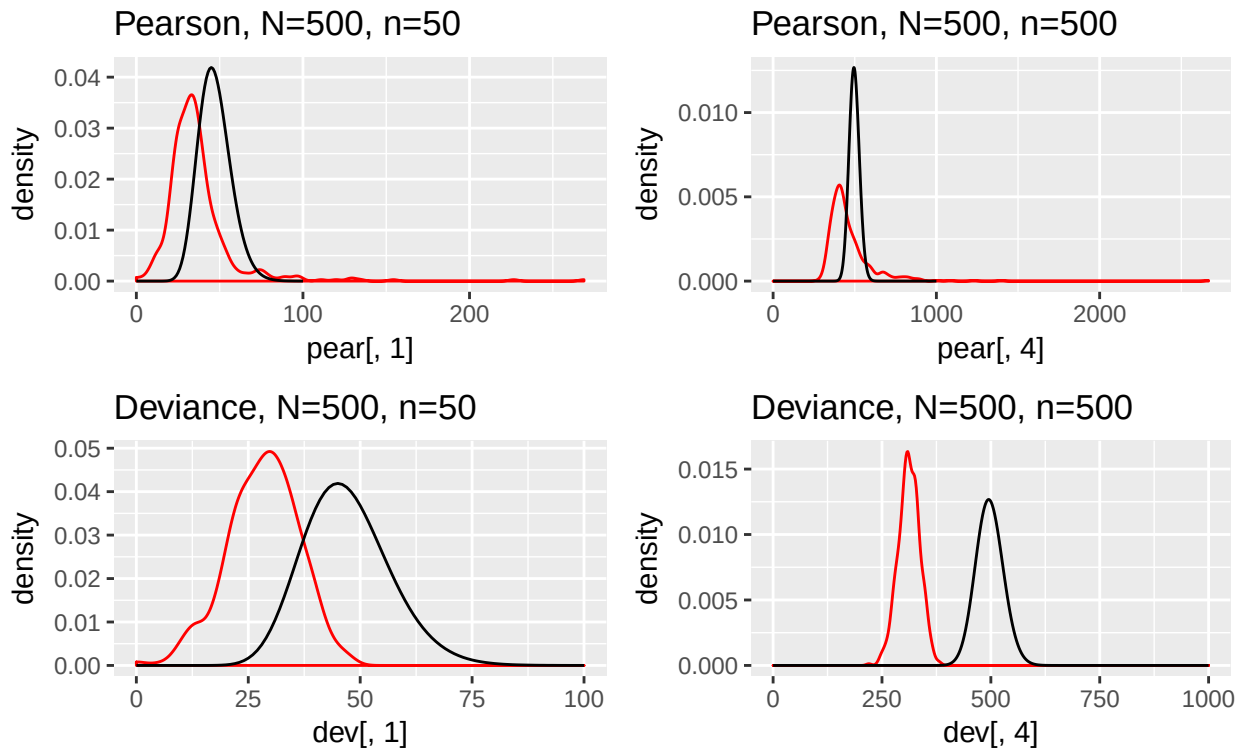
Note: the red line is the true empirical distribution; the black line is the Chi-squared distribution with df=n-p, the blue line is the normal distribution with $\mu = n - p$, $\sigma^2 = n - p$.

From the plots we can see that the mean of true empirical distribution (both Pearson and likelihood-ratio) is similar to the mean of χ^2_{n-p}, but the shape is not that similar, and we find the shape is more similar to $N(n-p, n-p)$. However in general, the these 3 distribution are not so different with each other, especially when n becomes larger, so in this case it seems the chi-squared assumption of test statistic is reasonable, and normal assumption looks better. But is this case representative? Let's look at another model specification.

In the second case we set the relationship between π_i and X_i to be:

$$logit(\pi) = 2x_1 + 3x_2$$

and we also compared the ture distribution of test statistic with the χ^2_{n-p} distribution, where $p = 3$, $n = 50, 100, 200, 500$. To save space we only shown the plot of $n = 50$ and $n = 500$ below:



Note: the red line is the true empirical distribution; the black line is the Chi-squared distribution with df=n-p

From the density plots we can see that the true empiracal distribution is very different with the χ^2_{n-p} distribution, and the mean of the true empiracal distribution is obviously less than the mean of χ^2_{n-p} distribution, especially when n becomes larger. So in this case, if we assume our test statistic follows a Chi-squared distributon and do a hypothesis test, the result will be very inaccurate, and we easily get a conclusion that the model fits the data bad, even under the right model specification. And we also find out that there Pearson statistic have some very large value, but likelihood-ratio statistic doesn't have, which means likelihood-ratio statistic is more stable.

So the assumption that Pearson statistic and likelihood-ratio statistic follows χ^2_{n-p} distribution is not appropriate. It may work in some cases but may also fail, so we need a better way to get the distribution of the test statistic.

Comparison of True and Bootstrap Empirical Distribution

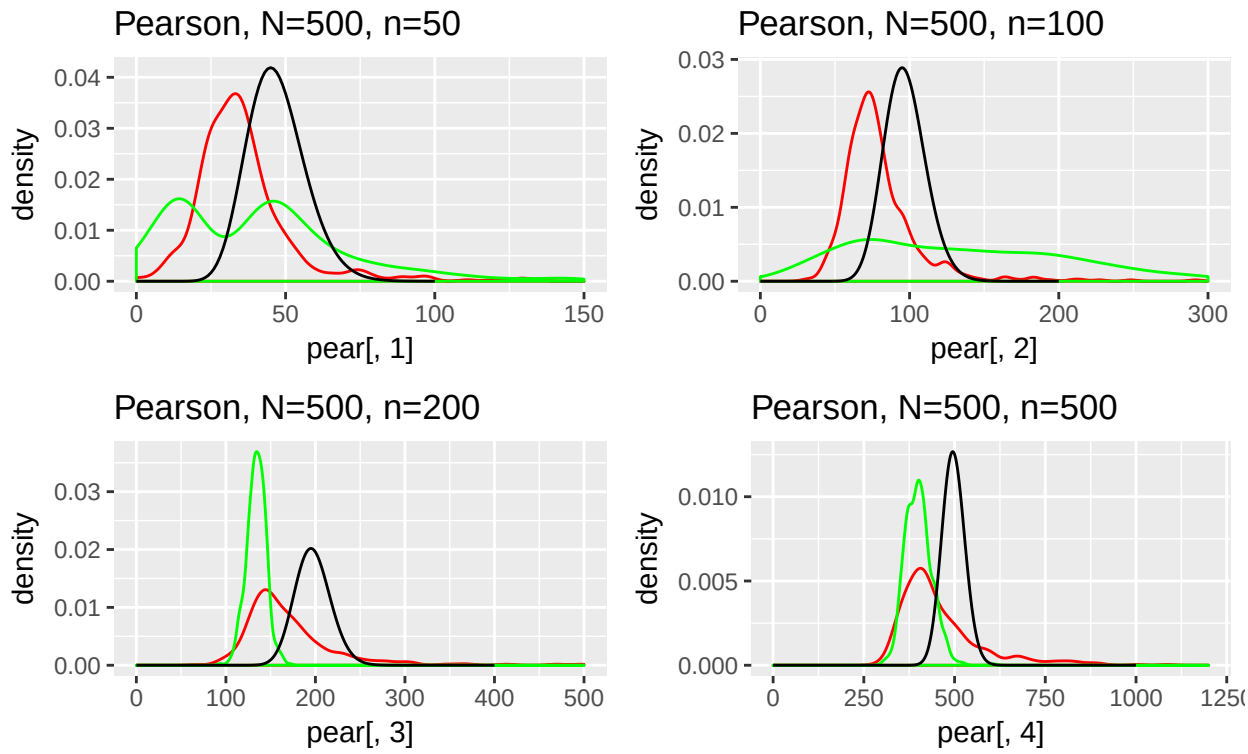
In the previous part, we generated 500 sequences of $\{(X_i, Y_i)\}_n$ (fixed n) and then get the true empirical distribution of the test statistic. But in most of the situations, we only have one sequence of $\{X_i, Y_i\}$, so we need to use bootstrap to get the empirical distribution of our test statistic. The algorithm is:

1. Sample (X_i, Y_i) n times from our sequence of $\{(X_i, Y_i)\}_n$ with replacement, and get a new sequence $\{(X_i^*, Y_i^*)\}_n$
2. Fit a logistic regression model (under right model specification) using the new sequence $\{(X_i^*, Y_i^*)\}_n$, and calculate the value of test statistic.
3. Repeat step 1 and step 2 B times, and get B values of test statistic. So now we have a bootsrap empirical distribution of test statistic.

We still consider the second case in the previous part where the χ^2 assumption is inappropriate. We use only one sequence of $\{(X_i, Y_i)\}_n$, and we do 500 times bootstrap ($B = 500$). The result is shown in the plot below:

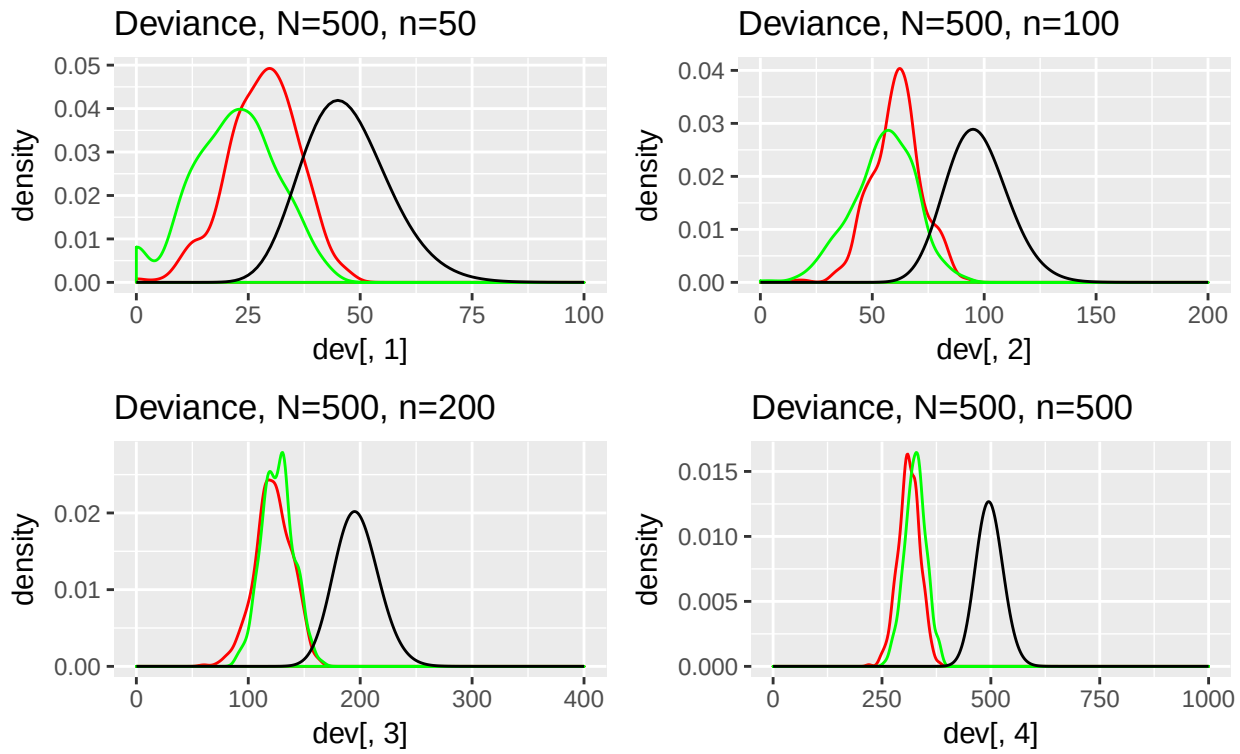
Note: the red line is the true empirical distribution; the green line is the bootstrap empiracal distribution; the black line is the Chi-squared distribution with df=n-p

Pearson statistic:



From the plot we can see that even under the right model specification, the bootstrap distribution of Pearson statistic is not similar to the true distribution of Pearson statistic.

likelihood-ratio statistic:



From the plots we can see that the bootstrap distribution of likelihood-ratio statistic is similar to the true distribution of Pearson statistic, especially when sample size n large.

Conclusion

When we want to test the good of fit of a logistic regression model, the χ^2 assumption of Pearson and likelihood-ratio statistic may be inappropriate, and likelihood-ratio statistic tend to be more stable compare to Pearson statistic. The most secure way is to use likelihood-ratio statistic and compare it to the empirical distribution getting by bootstrap.