Engineering, 2014, 6, 758-772
Published Online November 2014 in SciRes. http://www.scirp.org/journal/eng
http://dx.doi.org/10.4236/eng.2014.612074

Scientific Research

Effect Modeling of Count Data Using Logistic Regression with Qualitative Predictors*

Haeil Ahn

Department of Industrial Engineering, Seokyeong University, Seoul, Republic of Korea
Email: hiahn@skuniv.ac.kr

Received 25 August 2014; revised 24 September 2014; accepted 9 October 2014

Abstract

We modeled binary count data with categorical predictors, using logistic regression to develop a statistical method. We found that ANOVA-type analyses often performed unsatisfactorily, even when using different transformations. The logistic transformation of fraction data could be an alternative, but it is not desirable in the statistical sense. We concluded that such methods are not appropriate, especially in cases where the fractions were close to 0 or 1. The major purpose of this paper is to demonstrate that logistic regression with an ANOVA-model like parameterization aids our understanding and provides a somewhat different, but sound, statistical background. We examined a simple real world example to show that we can efficiently test the significance of regression parameters, look for interactions, estimate related confidence intervals, and calculate the difference between the mean values of the referent and experimental subgroups. This paper demonstrates that precise confidence interval estimates can be obtained using the proposed ANOVA-model like approach. The method discussed here can be extended to any type of experimental fraction data analysis, particularly for experimental design.

Keywords

Logistic Regression, Logit, Logistic Response, Categorical, Binary Count Data

1. Introduction

When manufacturing high-end goods, there is a trade-off between a high yield rate or lower fraction of nonconforming goods. A slight change in the process can drastically affect the yield rate or fraction of defective products, which results in a considerable increase or decrease in product turnover.

To develop a better process, purposeful changes should be made to the input variables of a process or produc-

*Effect modeling means "incremental effect parameterization".

How to cite this paper: Ahn, H. (2014) Effect Modeling of Count Data Using Logistic Regression with Qualitative Predictors. *Engineering*, **6**, 758-772. http://dx.doi.org/10.4236/eng.2014.612074

tion system, so that we can identify the reasons for changes in either the continuous or categorical outcomes and improve the manufacturing conditions. For this reason, it is commonplace in industry to analyze fraction data such as yield rates, percentages, and units of conforming or nonconforming product. When the input variables or regression predictors are all qualitative and the responses are countable, the data are often called categorical outcomes. Analysis of variance (ANOVA) has long been a favorite technique for investigating this type of data, as discussed in Rao [1], Wiener *et al*. [2] and Toutenburg and Shalabh [3].

Unfortunately, however, there are many cases where the fraction of nonconforming units of a product is close to zero or the yield rate of conforming units is close to one. In these cases, conventional analysis techniques often result in yield rate estimates exceeding 100%, or negative defective fraction estimates, as noted by many authors.

The drawbacks of using ANOVA for fraction data were noted by Cochran [4]. According to him, even the square-root or arcsine-square-root transformations of ANOVA-type data do not work properly. As Taguchi noted in Ross [5], the additive property of fraction data does not hold, especially when the fraction is lower than 20% or higher than 80%. He made use of what he called the omega (Ω) transformation for data conversion. Although the omega transformation has its merits, it is not satisfactory in the statistical sense. Jaeger [6] investigated the problem from the point of view of psychological or behavioral sciences, and found that ANOVA can yield spurious results even after applying the arcsine-square-root transformation to proportional data. ANOVAs over proportions can result in hard-to-interpret results, because the confidence intervals can extend beyond the interpretable values (0 and 1). As an alternative, he recommended logistic regression models, which he called the ordinary logit and/or mixed logit models.

In order to avoid above mentioned phenomena, we had better consider the logistic transformation. Dyke and Patterson [7] appear to be the first to use a logistic transformation to analyze ANOVA-type factorial data. Many theoretical backgrounds of logistic regression for categorical data analysis (CDA) are available. Montgomery [8], Kleinbaum and Klein [9], and Agresti [10] discussed the theoretical background in some detail. Some dichotomous response data were touched on in Dobson and Barnett [11] and Sloan *et al*. [12] in relation to contingency table analysis, while some polytomous response data were dealt with in Strokes *et al*. [13] and Dobson and Barnett [11].

In most cases, they dealt with quantitative explanatory variables. However, there are many cases when qualitative predictors are appropriate for modeling and analyses. Even in the comprehensive book by Agresti [10], logistic models with categorical predictors were not fully discussed. He did mention that logistic regression should include qualitative explanatory variables, often called categorical factors. In Agresti [10], the author touched on the ANOVA-type representation of factors and use of the logistic regression model. But the suggested model is quite limited to the case of one factor, and hence is not informative enough for practitioners who want to extend it to models of multiple factors. In Strokes *et al*. [13], the authors briefly introduced model fitting for logistic regression with two quantitative explanatory variables. In our opinion, however, their parameterization is a little confusing, and the ANOVA-model like parameterization is preferable.

Fortunately, modern statistics has presented many ways of extending logistic models. In this study, we consider a binary response variable (*i.e*., countable or categorical) and explanatory variables or predictors with three or more levels that are qualitative or categorical. The response variable may, for example, represent the units of a product manufactured under a certain condition. When trying to determine an appropriate statistical method for analyzing countable data within categorical settings, we have excluded the ANOVA-type analyses. However, we have used an ANOVA-model like parameterization with logistic regression and qualitative predictors. First, we examined the limitations of ANOVA-type analysis in connection to the defective rate or percentage data. Second, we considered logistic regression modeling of two-way binary count data with categorical variables. Then, we examined the behavior of the logistic regression model when fitted to the two-way count data within the logistic regression framework. We investigated this as an alternative to the ANOVA-type model, in an effort to combine logistic regression and qualitative predictors.

When implementing an experiment and analyzing the results, the optimal condition is sought for by testing the significance of regression parameters, evaluating the existence of interactions, estimating related confidence intervals (CIs), assessing the difference of mean values, and so on. The significance of model parameters and fraction estimates are used by the experimenter to identify and interpret the model.

The objectives of this study can be summarized as follows:

- To extend the ANOVA models with qualitative factors to logistic models with qualitative predictors.

- To estimate the main effects and/or interactions of ANOVA-model like parameterization.
- To estimate the confidence intervals (CIs) for model parameters and fractions.
- To ensure that the CIs for fractions are appropriate (between 0 and 1).
- To discuss the interpretation of the analysis results.

We have used a simple, but real, illustrative example to explain how to test the significance of model parameters, ascertain the existence of interactions, estimate the confidence intervals, and find the difference of the mean values. We have used the SAS in Allison [14] and MINITAB [15] logistic regression program to examine the efficiency of models and demonstrate the usefulness of logistic regression with ANOVA-model like parameterization.

2. Logistic Model with Categorical Predictors

2.1. Logistic Transformation

Let π be a fraction representing the probability of the event occurring, then $\pi/(1-\pi)$ the "odds". The naturally logged odds $\ln\left[\pi/(1-\pi)\right]$ are defined as the logistic transformation and also called "logit" link function. The logit link function converts fractions $(\pi's)$ between 0 and 1 to values between minus and plus infinity. For example, if we let $Y \sim Bin(n,\pi)$, then the random variable $\hat{\pi} = Y/n$ has its own probability mass function, which is discrete, non-negative, and asymmetric. The normal approximation of this random variable might cause aforementioned problems, especially when $n\pi < 5$ or $n(1-\pi) < 5$, due to the lack of normality as explained in Montgomery *et al.* [8]. If we take the sample log-odds $\log\left[\hat{\pi}/(1-\hat{\pi})\right]$, then the shape of the distribution becomes the logistic function, which is close to the normal distribution function. The cumulative distribution function is a monotonically increasing function. For another example, Ross [5] introduced Taguchi's omega (Ω) transformation formula in calculating db (decibel) value, which is similar to log-odds. The omega transformation formula is $\Omega(db) = 10\log\left[\pi/(1-\pi)\right]$, $0 < \pi < 1$. In this study, however, only the "logit" conversion is going to be considered.

On the other hand, the logistic model is set up to ensure that whatever estimate for success or failure we have, it will always be some number between 0 and 1. Thus, we can never get a success or failure estimate either above 1 or below 0. For the variable x, the standard logistic response function, called $\pi = \pi(x) = F(x)$, is given by e to the x over 1 plus e to the x or, alternatively, 1 over 1 plus e to minus x.

$$\pi = \pi(x) = \frac{e^{x}}{1+e^{x}} = \frac{1}{1+e^{-x}} \quad -\infty < x < \infty \quad 0 < \pi < 1 \tag{1}$$

For linear variety, $\beta_0 + \beta_1 x$, the logistic response function is:

$$\pi = \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad -\infty < x < \infty \quad 0 < \pi < 1 \tag{2}$$

Let us think of a simple regression analysis where there exist several types of responses such as observations, regression line, confidence interval, and prediction interval. The logistic response function transforms the responses into some number between 0 and 1, which results in S-shaped curves.

Generally, for a linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$, the logistic response function $\pi = \pi(\boldsymbol{x})$ is given as:

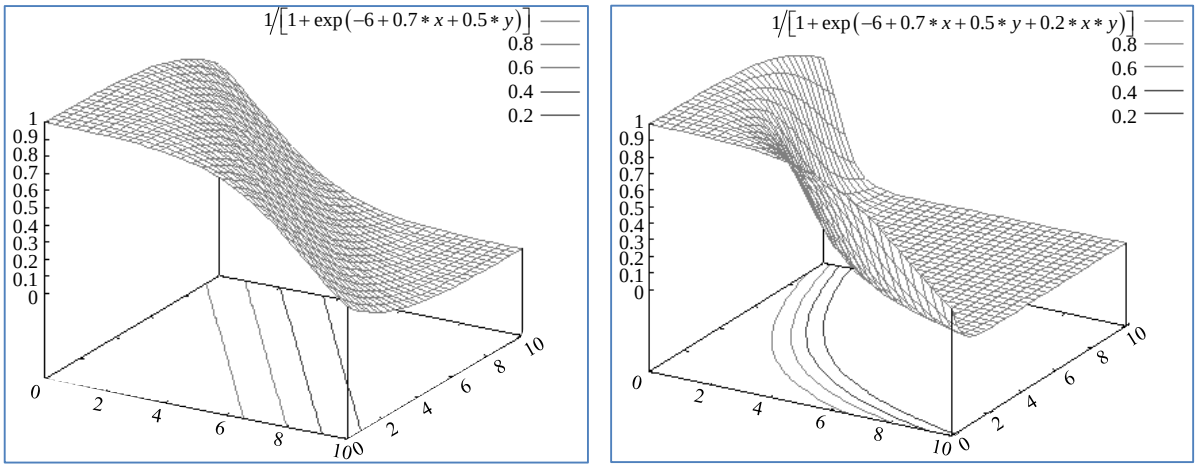
$$\pi = \pi(\boldsymbol{x}) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}} = \frac{1}{1+e^{-\boldsymbol{x}'\boldsymbol{\beta}}} \qquad -\infty < \boldsymbol{x} < \infty \qquad 0 < \pi < 1 \tag{3}$$

Typical response functions with and without interaction term can be depicted as in (b) and (a) of **Figure 1**. The logit link function called "log-odds" and logistic response function are reciprocal to each other. The logistic model is widely used for binomial data and is implemented in many statistical programs, such as SAS and MINITAB.

2.2. Models for Two-Way Responses

Let us consider two-way $l \times m$ layout data with n_{ij} observations within each subgroup. A typical ANOVA model looks like what follows.

$$\text{logit}(\pi_{ij}) = \alpha + \beta_i + \gamma_j + \delta_{ij} \qquad i = 1,2,\cdots,l \quad j = 1,2,\cdots,m \tag{4}$$



(a) (b)

Figure 1. Response of (a) $\pi = 1/\left[1+\exp\left(-6+0.7x+0.5v\right)\right]$ and (b) $\pi = 1/\left[1+\exp\left(-6+0.7x+0.5v+0.2xv\right)\right]$.

where β_i is the effect of the ith level of the row factor and γ_j the effect of the jth level of the column factor, respectively. The term δ_{ij} represents the effect of the interaction between β_i and γ_j. Normally, this model is subject to the following constraints.

$$\sum_{i=1}^{l}\beta_i = \sum_{j=1}^{m}\gamma_j = \sum_{i=1}^{l}\delta_{ij} = \sum_{j=1}^{m}\delta_{ij} = 0 \tag{5}$$

Such an ANOVA model can be transformed into a regression model. One way of defining the regression model corresponding to this model is as follows:

$$\operatorname{logit}\left(\pi_{ij}\right) = \alpha + \sum_{i=1}^{l}\beta_i x_i + \sum_{j=1}^{m}\gamma_j v_j + \sum_{i=1}^{l}\sum_{j=1}^{m}\delta_{ij}x_i v_j = \boldsymbol{x}'\boldsymbol{\beta} \tag{6}$$

$$x_s = \begin{cases} 1, & s = i \\ 0, & s \neq i, \quad s = 1,2,\cdots,l \end{cases} \qquad v_t = \begin{cases} 1, & t = j \\ 0, & t \neq j, \; t = 1,2,\cdots,m \end{cases}$$

This model is also subject to the constraints in Equation (5). This type of modeling is often called “effect modeling” or “incremental effect parameterization”.

2.3. Odds Ratio Assessment

The odds ratio (OR) is defined as the ratio of any odds of experimental subgroup to that of the referent one.

$$OR_{ij} = OR\left(A_i B_j\right) = \frac{\pi_{ij}/\left(1-\pi_{ij}\right)}{\pi_{11}/\left(1-\pi_{11}\right)} \tag{7}$$

In this study, we found that if the odds ratio is less than or equal to one; *i.e.*, $OR_{ij} \le 1$, then the following holds:

$$\begin{aligned} &\pi_{ij}/\left(1-\pi_{ij}\right) \le \pi_{11}/\left(1-\pi_{11}\right) \\ &\pi_{ij} - \pi_{ij}\pi_{11} \le \pi_{11} - \pi_{ij}\pi_{11} \\ &\pi_{ij} \le \pi_{11} \end{aligned} \tag{8}$$

If we are sure that the upper and lower limits of $OR_{ij} \le 1$ with $100\left(1-\alpha\right)\%$ confidence, then the upper limit of two-sided $100\left(1-\alpha\right)\%$ CI corresponds to that of one-sided $100\left(1-\alpha/2\right)\%$ confidence interval.

2.4. Interaction Assessment

As defined in Kleinbaum and Klein [9], an equation for assessing interaction can be identified as follows. We

begin with the null hypothesis that:

$$H_0 : \delta_{ij} = \ln\left[OR_{ij} / OR_{i1} \times OR_{1j} \right] = 0 \tag{9}$$

It is interesting to note that the interaction effects can be expressed to be multiplicative. One way to state this null hypothesis, in terms of odds ratios, is that OR_{ij} equals the product of OR_{i1} and OR_{1j}.

$$H_0 : OR_{ij} = OR_{i1} \times OR_{1j} \tag{10}$$

If the equation of this null hypothesis is satisfied, we say that there is "no interaction on a multiplicative scale." In contrast, if this expression does not hold, we say that there is "evidence of interaction on a multiplicative scale."

We can make use of this formula to test the null hypothesis (H_0) of no interaction on a multiplicative scale. If null hypothesis is true, then we can interpret the hypothesis in either way.

$$H_0 : OR_{ij} = OR_{i1} \times OR_{1j} \Leftrightarrow H_0 : \delta_{ij} = 0 \tag{11}$$

2.5. Estimation of Regression Parameters

We consider generalized linear models in which the outcome variables are measured on a binary scale, as explained in Dobson and Barnett [11]. For example, the responses may be "success" or "failure" or non-conforming or conforming. "S" and "F" denoted by 1 or 0 are used as generic terms of the two categories. First, the binary random variable is defined.

$$U = \begin{cases} 1 & \text{if the outcome is a success} \\ 0 & \text{if the outcome is a failure} \end{cases} \tag{12}$$

with probabilities $\Pr(U=1) = \pi$ and $\Pr(U=0) = 1-\pi$, which is the Bernoulli distribution. Suppose there are n such random variables $U_1, \cdots, U_n$ which are independent of each other with $\Pr(U_j = 1) = \pi_j$. In the case where the π_j's are all equal, we can define a new random variable $Y = \sum_{j=1}^{n} U_j$ so that Y is the number of successes in n trials. The random variable Y has the binomial distribution:

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0,1,2,\cdots,n \tag{13}$$

In Dobson and Barnett [11], the one factor case of N independent random variables $Y_1, Y_2, \cdots, Y_N$ corresponding to the number of successes in N different subgroups. If $Y_i \sim Bin(n_i, \pi_i)$, the log-likelihood function is

$$l(\pi_1, \cdots, \pi_N; y_1, \cdots, y_N) = \sum_{i=1}^{N} \left[y_i \ln\left[\pi_i / (1-\pi_i) \right] + n_i \ln(1-\pi_i) + \ln\binom{n_i}{y_i} \right] \tag{14}$$

In this study, we intend to extend this one-way single factor case to two-way two factor case.

We consider the case of $N = l \times m$ independent random variables $Y_{11}, Y_{12}, \cdots, Y_{lm}$ corresponding to the numbers of successes in $N = l \times m$ two-way subgroups as in **Table 1**.

If we define $\Pr(U_{ij} = 1) = \pi_{ij}$, then the likelihood function can be given by:

$$\prod_{i=1}^{l} \prod_{j=1}^{m} \pi_{ij}^{u_{ij}} (1-\pi_{ij})^{1-u_{ij}} = \exp\left\{ \sum_{i=1}^{l} \sum_{j=1}^{m} u_{ij} \ln\left[\pi_{ij} / (1-\pi_{ij}) \right] + \sum_{i=1}^{l} \sum_{j=1}^{m} \ln(1-\pi_{ij}) \right\} \tag{15}$$

Since $Y_{ij} \sim Bin(n_{ij}, \pi_{ij})$, the log-likelihood function becomes:

$$l(\pi_{11}, \cdots, \pi_{lm}; y_{11}, \cdots, y_{lm}) = \sum_{i=1}^{l} \sum_{j=1}^{m} \left[y_{ij} \ln\left[\pi_{ij} / (1-\pi_{ij}) \right] + n_{ij} \ln(1-\pi_{ij}) + \ln\binom{n_{ij}}{y_{ij}} \right] \tag{16}$$

Table 1. Frequencies of $l \times m$ binomial distributions.

Factors			B			
			1	2	⋯	m
A	1	Successes	Y_{11}	Y_{12}	⋯	Y_{1m}
		Number of Trials	n_{11}	n_{12}	⋯	n_{1m}
	2	Successes	Y_{21}	Y_{22}	⋯	Y_{2m}
		Number of Trials	n_{21}	n_{22}	⋯	n_{2m}
	⋮	⋮	⋮	⋮	⋱	⋮
	l	Successes	Y_{l1}	Y_{l2}	⋯	Y_{lm}
		Number of Trials	n_{l1}	n_{l2}	⋯	n_{lm}

As defined in Equation (7), $\pi_{ij} = \exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}\right) \Big/ \left[1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}\right)\right]$. It follows that $1-\pi_{ij} = 1 \Big/ \left[1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}\right)\right]$ and $\ln\left(1-\pi_{ij}\right) = -\ln\left[1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}\right)\right]$. The log likelihood function takes the form of:

$$l\left(\boldsymbol{\beta}; y_{11}, \cdots, y_{lm}\right) = \sum_{i=1}^{l}\sum_{j=1}^{m}\left\{ y_{ij}\boldsymbol{x}_{ij}'\boldsymbol{\beta} - n_{ij}\ln\left[1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}\right)\right] + \ln\binom{n_{ij}}{y_{ij}}\right\} \tag{17}$$

The partial derivative and the Hessian of this likelihood function with respect to $\boldsymbol{\beta}$ give the score vector and the information matrix $\boldsymbol{I}(\boldsymbol{\beta})$, respectively. As explained in Montgomery *et al.* [8], numerical search methods could be used to compute the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$. Alternatively, one can use iteratively reweighted least squares (IRLS) to actually find the MLEs. The MLE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ can be obtained by the following recursive equation.

$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k-1)} - \left[\boldsymbol{I}\left(\boldsymbol{\beta}\right)^{(k-1)}\right]^{-1}\boldsymbol{u}^{(k-1)}, \quad k = 1, 2, 3, \cdots \tag{18}$$

This is usually called generalized estimating equation (GEE). For more details, refer to Montgomery *et al.* [8] and Dobson and Barnett [11]. The information matrix corresponds to $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$. The inverse of information matrix $\left[\boldsymbol{I}(\boldsymbol{\beta})\right]^{-1}$ is the estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}$. If we let $\hat{\boldsymbol{\beta}}$ be the final estimate vector, then

$$E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}, \quad Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\right)^{-1} = \left[\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \tag{19}$$

where the $\boldsymbol{V}$ is an $N \times N$ diagonal matrix containing the estimated variance on the main diagonal; that is, the ith diagonal element of $\boldsymbol{V}$ is $n_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)$. If the model is a good fit of the data, the deviance D should approximately have the distribution $\chi^2\left(N-r\right)$, where N is the different values of x and r is the number of parameters in the model.

2.6. Interval Estimation of Fractions

The odds can be estimated by $\exp\left(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}\right)$ and the odds ratio can be expressed by the antilog of the corresponding parameter. The fitted fraction estimate for the logistic regression model is written by:

$$\hat{\pi}_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}\right)} = \frac{1}{1+\exp\left(-\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}\right)} \tag{20}$$

This is the point estimate for the fraction of each subgroup of outcomes. To obtain the interval estimation, prediction vectors are needed. Let π_0 be the fraction to be estimated and $\boldsymbol{x}_0'$ be a row vector of prediction.

$$\hat{\pi}_0 = \frac{\exp\left(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right)} = \frac{1}{1+\exp\left(-\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right)} \tag{21}$$

The expectation and variance can be given by:

$$E\left(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right)=\boldsymbol{x}_0'\boldsymbol{\beta},\quad \mathrm{Var}\left(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right)=\boldsymbol{x}_0'\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{x}_0=\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0 \tag{22}$$

The $100(1-\alpha)\%$ lower and upper limit of $\mathrm{logit}(\pi_0)$ at $\boldsymbol{x}_0$ can be calculated as follows:

$$L(\boldsymbol{x}_0)=\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}-z_{1-\alpha/2}\sqrt{\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0},\ U(\boldsymbol{x}_0)=\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}+z_{1-\alpha/2}\sqrt{\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0} \tag{23}$$

By taking the reciprocals, $100(1-\alpha)\%$ confidence interval for π_0 can be calculated by:

$$\left(\frac{\exp\left[L(\boldsymbol{x}_0)\right]}{1+\exp\left[L(\boldsymbol{x}_0)\right]},\frac{\exp\left[U(\boldsymbol{x}_0)\right]}{1+\exp\left[U(\boldsymbol{x}_0)\right]}\right)\ \text{or}\ \left(\frac{1}{1+\exp\left[-L(\boldsymbol{x}_0)\right]},\frac{1}{1+\exp\left[-U(\boldsymbol{x}_0)\right]}\right) \tag{24}$$

In the estimation process, the precision varies depending on the sample size of each subgroup. The bigger the sample size is, the more accurate the fraction estimates are.

There exist excellent computer programs that implement maximum-likelihood estimation for logistic regression, such as SAS PROC LOGISTIC in Allison [14] and MINITAB [15]. We have only to apply ourselves to modeling and parameterization.

3. Illustrative Example of Qualitative Predictors

3.1. Illustrative Example

A manufacturing company produces touch-screen panels on a massive scale to deliver to a customer company, who produces smart phones and tablet PCs. Currently, resistive touch-screen panels (TSP) are being widely used. The company plans to produce capacitive TSP (CTSP) to minimize the thickness. However, the following problems may be caused during the fabrication of CTSP. For example, after performing the screen printing process, when an Ag paste is cured at a high temperature, cracks may occur in a fine indium tin oxide (ITO) line. Moreover, many defects such as air bubbles, alien substances and scratches, may take places during the interlayer lamination process. The defective items are the major source of failure cost of the product.

An experimenter is seeking a method of fabricating the CTSP that can efficiently reduce the cost of CTSP fabrication, which can be assessed in terms of yield rate or fraction non-conforming. There are four patented methods of fabrication and four types of facilities available for the process operation. The experimenter is concerned with handling with explanatory variables that are qualitative and contain three or more levels.

3.2. Units of Nonconforming

Since the example lends itself to the problem of two-way binary data, let us consider two qualitative factor experiments with $l=4$ levels of factor A (facility), $m=3$ levels of factor B (manufacturing method), and $k=100$ replicates except the current manufacturing condition with $k=500$ replicates. The units of product data either conforming or nonconforming are given in **Table 2**. The data are grouped as frequencies for each combination of factor levels. There are sixteen cells within the table, each of which corresponds to a certain manufacturing condition. A fixed effect model seems to be appropriate for the data, since the levels of factors are not randomly chosen.

Expressed in terms of the combination of factor levels, the subgroup A_1B_1 is the referent cell representing current manufacturing condition. The referent cell, also called the control subgroup (A_1B_1) of this experiment, consists of the individual Bernoulli trials. The trials belonging to the cells other than the referent cell are members of the corresponding experiment subgroup or experiment cell. Let π_{ij} denote the fraction of each subgroup where units of product will not conform to specification. The quantity $(1-\pi_{ij})$ is so called yield rate of each process corresponding to the combination of factor levels.

The experimenter is concerned about the optimal subgroup and the significance of the fractions of these experimental subgroups, eventually to find out the improved experimental subgroup, if any, which gives the lowest π_{ij} in the statistical sense.

3.3. ANOVA with Logistic Transformation

The logistically transformed log-odds data shown in **Table 3** can be regarded as a two-way layout without

Table 2. Units of product nonconforming.

Mfg Method Facility	B_1		B_2		B_3	
	Number Of Units	Units Non-Conforming	Number Of Units	Units Non-Conforming	Number Of Units	Units Non-Conforming
A_1	500	25	100	10	100	37
A_2	100	6	100	20	100	53
A_3	100	8	100	3	100	21
A_4	100	20	100	22	100	38

Table 3. Logistically transformed data.

Work Method Machine	B_1	B_2	B_3	Mean
A_1	−2.94444	−2.19722	−0.53222	−1.89129
A_2	−2.75154	−1.38629	0.12014	−1.33923
A_3	**−2.44235**	−3.47610	−1.32493	**−2.41446**
A_4	−1.38629	−1.26567	−0.48955	−1.04717
Mean	**−2.38115**	−2.08132	−0.55664	−1.67304

replications.

The following model seems to be relevant as far as the ANOVA is concerned.

$$w_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \quad \varepsilon_{ij} \sim NI\left(0, \sigma_e^2\right) \quad i = 1,2,3,4 \quad j = 1,2,3 \tag{25}$$

Since this is a two-way layout without replication, the interactions are not considered in the model. Based on the model, the ANOVA can be conducted as in **Table 4**.

Factor B looks significant at 95% significance level, factor A does not. The smaller the nonconforming fraction is, the better its performance is. The minimum point estimate for nonconforming fraction can be calculated by:

$$\hat{\mu}_{31} = \hat{\mu}\left(A_3B_1\right) = \bar{w}_{3\bullet} + \bar{w}_{\bullet 1} - \bar{\bar{w}} = -2.38115 - 2.41446 + 1.67304 = -3.12257 \tag{26}$$

Since the number of effective replication is given by:

$$n_e = lm/\left(l + m - 1\right) = 4 \times 3/\left(4 + 3 - 1\right) = 2 \tag{27}$$

The 95% confidence interval for $\mu\left(A_3B_1\right)$ can be calculated as:

$$-3.12258 \pm t_{0.975}\left(6\right)\sqrt{MS_E/2} = -3.12258 \pm 2.447 \times 0.43850 = \left(-4.19557, -2.04957\right) \tag{28}$$

The result of Minitab processing can be obtained as in **Figure 2**.

The point estimate for $\pi\left(A_3B_1\right)$ can be given by:

$$\tilde{\pi}_{31} = \frac{\mathrm{e}^{\hat{\mu}_{31}}}{1 + \mathrm{e}^{\hat{\mu}_{31}}} = \frac{1}{1 + \mathrm{e}^{-\hat{\mu}_{31}}} = \frac{1}{1 + \mathrm{e}^{3.12258}} = 0.0424185 = 4.24185\left(\%\right) \tag{30}$$

Likewise, the 95% confidence interval for $\pi\left(A_3B_1\right)$ can be calculated as:

$$\left(0.014840, 0.114086\right) = \left(1.4840\%, 11.4086\%\right) \tag{30}$$

The result is quite plausible in that the interval is narrower than before and the lower limit can never be a negative. However, the optimal manufacturing condition of ANOVA is A_3B_1, not A_3B_2. Moreover, we have some misgivings about the fact that since the sample size is ignored. Some of the sample information must be lost during the course of the analyzing process. For this reason, we are encouraged to give a try to logistic regression.

3.4. Logistic Regression for the Illustrative Example

The data structure for the illustrative example shown in **Table 2** is identified as follows:

$$\text{logit}\left(\pi_{ij}\right)=\ln\left[\pi_{ij}\big/\left(1-\pi_{ij}\right)\right]=\alpha+\beta_i+\gamma_j+\delta_{ij}\quad i=1,2,3,4\quad j=1,2,3 \tag{31}$$

where β_i and γ_j denotes the main effects of factor A and B, respectively. As a preliminary model, δ_{ij} is needed to represent potential interaction effect. In a logistic regression, if we adopt the logit link function, then the model for the data in **Table 2** can be stated as in the following equation. This is an extension of one predictor model in Agresti [10] and a leverage of two predictor model in Strokes *et al*. [13]. This is what we call ANOVA-model like parameterization of logistic regression.

$$\text{logit}\left(\pi_{ij}\right)=\ln\left[\pi_{ij}\big/\left(1-\pi_{ij}\right)\right]=\boldsymbol{x}_{ij}'\boldsymbol{\beta}=\alpha+\sum_{i=1}^{4}\beta_i x_i+\sum_{j=1}^{3}\gamma_j v_j+\sum_{i=1}^{4}\sum_{j=1}^{3}\delta_{ij}x_i v_j\quad(\text{model 1}) \tag{32}$$

where $\boldsymbol{\beta}$ is the column vector of model parameters and $\boldsymbol{x}_{ij}'$ is the row vector of corresponding indicator values. This model is also subject to the following constraints.

$$\sum_{i=1}^{4}\beta_i=0\quad\sum_{j=1}^{4}\gamma_j=0\quad\sum_{i=1}^{4}\delta_{ij}=0,\ j=1,2,3\quad\sum_{j=1}^{3}\delta_{ij}=0,\ i=1,2,3,4 \tag{33}$$

There are several ways of handling these constraints as in Dobson and Barnett [11]. One of those methods is to set to zero the first term of each constraint. That is,

$$\beta_1=\gamma_1=0,\quad\delta_{11}=\delta_{12}=\delta_{13}=0,\quad\delta_{21}=\delta_{31}=\delta_{41}=0 \tag{34}$$

If α in the model is so determined, then α in represents the mean of the referent subgroup A_1B_1. The parameters β_2, β_3 and β_4 are the incremental main effects for method 2, 3 and 4, respectively as compared to referent subgroup. Likewise, γ_2 and γ_3 are the incremental main effects for facility 2 and 3, respectively. The interactions are also incremental. Expressed in terms of matrix and vector notation, the incremental effect parametric logistic regression model is analogous to the conventional model of $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$.

Let $\boldsymbol{x}_{ij}'$ denote the row vector of the design matrix $\boldsymbol{X}$ corresponding to the combination of factor levels A_iB_j. The sample logit vector $\boldsymbol{y}$, the design matrix $\boldsymbol{X}$, and the parameter vector $\boldsymbol{\beta}$ are as follows.

$$\boldsymbol{y}=\begin{bmatrix}\text{logit}\left(\hat{\pi}_{11}\right)\\ \text{logit}\left(\hat{\pi}_{12}\right)\\ \text{logit}\left(\hat{\pi}_{13}\right)\\ \text{logit}\left(\hat{\pi}_{21}\right)\\ \text{logit}\left(\hat{\pi}_{22}\right)\\ \text{logit}\left(\hat{\pi}_{23}\right)\\ \text{logit}\left(\hat{\pi}_{31}\right)\\ \text{logit}\left(\hat{\pi}_{32}\right)\\ \text{logit}\left(\hat{\pi}_{33}\right)\\ \text{logit}\left(\hat{\pi}_{41}\right)\\ \text{logit}\left(\hat{\pi}_{42}\right)\\ \text{logit}\left(\hat{\pi}_{43}\right)\end{bmatrix}\quad\boldsymbol{X}=\begin{array}{c}\begin{matrix}\alpha&\beta_2&\beta_3&\beta_4&\gamma_2&\gamma_3&\delta_{22}&\delta_{23}&\delta_{32}&\delta_{33}&\delta_{42}&\delta_{43}\end{matrix}\\ \begin{bmatrix}1&0&0&0&0&0&0&0&0&0&0&0\\1&0&0&0&1&0&0&0&0&0&0&0\\1&0&0&0&0&1&0&0&0&0&0&0\\1&1&0&0&0&0&0&0&0&0&0&0\\1&1&0&0&1&0&1&0&0&0&0&0\\1&1&0&0&0&1&0&1&0&0&0&0\\1&0&1&0&0&0&0&0&0&0&0&0\\1&0&1&0&1&0&0&0&1&0&0&0\\1&0&1&0&0&1&0&0&0&1&0&0\\1&0&0&1&0&0&0&0&0&0&0&0\\1&0&0&1&1&0&0&0&0&0&1&0\\1&0&0&1&0&1&0&0&0&0&0&1\end{bmatrix}\end{array}\quad\boldsymbol{\beta}=\begin{bmatrix}\alpha\\ \beta_2\\ \beta_3\\ \beta_4\\ \gamma_2\\ \gamma_3\\ \delta_{22}\\ \delta_{23}\\ \delta_{32}\\ \delta_{33}\\ \delta_{42}\\ \delta_{43}\end{bmatrix} \tag{35}$$

The explanatory variables are all indicators. Notice that the columns corresponding to β_1, γ_2, δ_{11}, δ_{12} and δ_{21} are eliminated to avoid the redundancy of column vectors. Some columns of the matrix are orthogonal to each other and some columns are interrelated. It is interesting to note that the design matrix enjoys a special structure. The adoption of some parameters is a matter of choice. For instance, we can eliminate β_4, δ_{42} and δ_{43} columns and the corresponding rows $\hat{\pi}_{41}$, $\hat{\pi}_{42}$ and $\hat{\pi}_{43}$ from the design matrix $\boldsymbol{X}$ simultaneously without affecting the estimates of the rest of the parameters.

3.5. Estimation of Parameters

The full pattern of model 1 is fitted into the data in **Table 2**. Minitab logistic regression output is displayed in **Figure 3**.

Table 4. ANOVA table.

Source	SS	dof	MS	F_0	p-value
A	3.3014	3	1.10048	2.86	0.126
B	7.6579	2	3.82894	9.96*	0.012
E	2.3071	6	0.38456		
T	13.2664	11			

S = 0.620098 R-square = 82.61% R-square(adj) = 68.12%

*: significant at 5% **: significant at 1%

Fitted Value	SE	95% CI	95% PI
−3.12258	0.438475	(−4.19549, −2.04967)	(−4.98091, −1.26424)

Figure 2. 95% confidence and prediction interval (Minitab).

Prediction						90% CI	
Variables	Coeff	SE	Z	P	OR	Lower	Upper
Const	−2.94444	0.205196	−14.35	0.000			
x2	0.192904	0.468412	0.41	0.680	1.21	0.56	2.62
x3	0.502092	0.421871	1.19	0.234	1.65	0.83	3.31
x4	1.55814	0.323427	4.82	0.000	4.75	2.79	8.09
v2	0.747214	0.391429	1.91	0.056	2.11	1.11	4.02
v3	2.41222	0.291557	8.27	0.000	11.16	6.91	18.03
x2v2	0.618027	0.626914	0.99	0.324	1.86	0.66	5.20
x2v3	0.459457	0.549959	0.84	0.403	1.58	0.64	3.91
x3v2	−1.78097	0.795442	−2.24	0.025	0.17	0.05	0.62
x3v3	−1.29480	0.530238	−2.44	0.015	0.27	0.11	0.66
x4v2	−0.626586	0.523442	−1.20	0.231	0.53	0.23	1.26
x4v3	−1.51548	0.435833	−3.48	0.001	0.22	0.11	0.45

Figure 3. Parameter estimates for model 1.

The p-values and the CI's of odds ratios can be regarded as measures of the significance tests of regression parameters. Some parameters look significant, but others do not. As a matter of fact, regardless of whether parameters are significant or not, we can eliminate any rows or columns from the table on purpose without affecting the estimation of other parameters, owing to the incremental effect parameterization.

For example, we are interested in comparison of reference subgroup and strong candidate subgroup for optimality. Since the combination A_3B_2 is the strong candidate for optimality, we can eliminate row 2 and 4 and also columns 3 from **Table 2**. By doing so, we can drastically reduce the data table until it becomes as small as 2×2, 2×1 or 1×2. If we had known the fact, we could have started with the data table of size 2×2, 2×1 or 1×2. In this case, the resultant table is 2×2 as shown in **Table 5**. Note that the table includes both referent and candidate subgroups.

Likewise, the logistic regression model reduces to the following.

$$\text{logit}\left(\pi_{ij}\right)=\ln\left[\pi_{ij}\Big/\left(1-\pi_{ij}\right)\right]=\boldsymbol{x}_{ij}'\boldsymbol{\beta}=\alpha+\beta_3x_3+\gamma_2v_2+\delta_{32}x_3v_2\ \ (\text{model 2}) \tag{36}$$

The design matrix $\boldsymbol{X}$ and the parameter vector $\boldsymbol{\beta}$ are identified as follows.

$$\begin{matrix}\alpha & \beta_3 & \gamma_2 & \delta_{32}\end{matrix}\qquad \boldsymbol{X}=\begin{bmatrix}1&0&0&0\\1&0&1&0\\1&1&0&0\\1&1&1&1\end{bmatrix},\quad \boldsymbol{\beta}=\begin{bmatrix}\alpha\\\beta_3\\\gamma_2\\\delta_{32}\end{bmatrix} \tag{37}$$

Table 5. First reduced table.

Working Method Machine	B_1		B_2	
	Number of Units	Units Non-Conforming	Number of Units	Units Non-Conforming
A_1	500	25	100	10
A_3	100	8	100	3

The parameter estimates are shown in **Figure 4**. It is worthy to note that the parameter estimates remain the same. We can ensure that the elimination of rows and columns does not affect the parameter estimates. The phenomenon that makes matters simple is the major difference between ANOVA-type and incremental effect modeling.

On the one hand, $\hat{\beta}_3$ seems insignificant because the corresponding p-value 0.234 is greater than 0.10 and the 90% confidence interval for odds ratio (0.83, 3.31) contains one. On the other hand, $\hat{\delta}_{32}$ is affirmatively significant in that the upper and lower limits of 90% CI are smaller than one. We decide to eliminate the β_3 column and the corresponding row from the design matrix for the parsimony of the model. The model becomes

$$\text{logit}\left(\pi_{ij}\right) = \ln\left[\pi_{ij}/\left(1-\pi_{ij}\right)\right] = \boldsymbol{x}_{ij}'\boldsymbol{\beta} = \alpha + \gamma_2 v_2 + \delta_{32} x_3 v_2 \quad \left(\text{model } 3\right) \tag{38}$$

The design matrix $\boldsymbol{X}$ and the parameter vector $\boldsymbol{\beta}$ are reduced and identified as follows:

$$\begin{matrix} & \alpha & \gamma_2 & \delta_{32} \end{matrix}$$
$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \gamma_2 \\ \delta_{32} \end{bmatrix} \tag{39}$$

The parameter estimates are shown in **Figure 5**.

The estimate for $\hat{\delta}_{32}$ changes, but still $\hat{\gamma}_2$ does not. The estimates for γ_2 and δ_{32} seem to be significant at 10%. Note that the equation of model 3 could be the final one, because all parameter estimates are significant at 10%. We can draw a conclusion that the model is appropriate.

3.6. Existence of Interactions

Notice that the estimate for the interaction δ_{32} is significant at 10%. The last line of **Figure 5** gives the information on the point estimate and CI for δ_{32}. The estimate for $\mathrm{e}^{\delta_{32}}$ is 0.28, which is the quantity of $OR_{32}/\left(OR_{31}\cdot OR_{12}\right)$. The 90% CI for $\mathrm{e}^{\delta_{32}}$ is (0.09, 0.84) as in **Figure 5**, the upper and lower limits of which are smaller than 1 and hence affirmatively significant. Therefore, it is the evidence that there exists interaction. In this case, $\hat{\beta}_3 = 0$, thus we can make sure that

$$OR_{32} = \frac{\pi_{32}/\left(1-\pi_{32}\right)}{\pi_{11}/\left(1-\pi_{11}\right)} = \frac{\mathrm{e}^{\alpha+\beta_3+\gamma_2+\delta_{32}}}{\mathrm{e}^{\alpha}} = \mathrm{e}^{\beta_3+\gamma_2+\delta_{32}} \neq \mathrm{e}^{\beta_3+\gamma_2} = \frac{\mathrm{e}^{\alpha+\beta_3}}{\mathrm{e}^{\alpha}}\cdot\frac{\mathrm{e}^{\alpha+\gamma_2}}{\mathrm{e}^{\alpha}} = OR_{31}\cdot OR_{12} \tag{40}$$

Usually, equality does not hold, unless $\delta_{32} = 0$. In other words, there is the evidence that interaction exists if and only if $\delta_{32} \neq 0$ or the antilog of δ_{32} is other than 1. If the interaction does not exist, then

$$H_0 : \frac{OR_{32}}{OR_{31}\cdot OR_{12}} = \frac{\mathrm{e}^{\beta_3+\gamma_2+\delta_{32}}}{\mathrm{e}^{\beta_3}\cdot\mathrm{e}^{\gamma_2}} = \mathrm{e}^{\delta_{32}} = \mathrm{e}^{0} = 1 \tag{41}$$

3.7. Estimation of Confidence Intervals

We can ensure that the point estimates for each fraction can be obtained as follows:

$$\text{logit}\left(\hat{\pi}_{11}\right) = \ln\left[\hat{\pi}_{11}/\left(1-\hat{\pi}_{11}\right)\right] = \hat{\alpha} = -2.94444$$
$$\Rightarrow \hat{\pi}_{11} = 1/\left[1+\exp\left(-\hat{\alpha}\right)\right] = 1/\left[1+\exp\left(2.94444\right)\right] = 0.05$$

Prediction						90% CI	
Variables	Coeff	SE	Z	P	OR	Lower	Upper
Const	−2.94444	0.205196	−14.35	0.000			
x3	0.502092	0.421870	1.19	0.234	1.65	0.83	3.31
v2	0.747214	0.391421	1.91	0.056	2.11	1.11	4.02
x3v2	−1.78097	0.795423	−2.24	0.025	0.17	0.05	0.62

Figure 4. Parameter estimates for model 2.

Prediction						90% CI	
Variables	Coeff	SE	Z	P	OR	Lower	Upper
Const	−2.94444	0.205196	−14.35	0.000			
v2	0.747214	0.391429	1.91	0.056	2.11	1.11	4.02
x3v2	−1.27887	0.674354	−1.90	0.058	0.28	0.09	0.84

Figure 5. Parameter estimates for model 3.

$$\text{logit}\left(\hat{\pi}_{12}\right)=\ln\left[\hat{\pi}_{12}/\left(1-\hat{\pi}_{12}\right)\right]=\hat{\alpha}+\hat{\gamma}_2=-2.94444+0.747214=-2.197226$$

$$\Rightarrow \hat{\pi}_{12}=1/\left[1+\exp\left(-\hat{\alpha}-\hat{\gamma}_2\right)\right]=1/\left[1+\exp\left(2.197226\right)\right]=0.10$$

$$\text{logit}\left(\hat{\pi}_{32}\right)=\ln\left[\hat{\pi}_{32}/\left(1-\hat{\pi}_{32}\right)\right]=\hat{\alpha}+\hat{\gamma}_2+\hat{\delta}_{32}=-2.94444+0.747214-1.27887=-3.476096$$

$$\Rightarrow \hat{\pi}_{32}=1/\left[1+\exp\left(-\hat{\alpha}-\hat{\gamma}_2-\hat{\delta}_{32}\right)\right]=1/\left[1+\exp\left(3.476096\right)\right]=0.03$$

From a conventional statistical view point, we might like to calculate the confidence intervals. To find confidence intervals, we have to know the standard errors $SE(\cdot)$ of the model parameter estimates. The parameter estimates reported in **Figure 5** are just the square roots of the main diagonal elements of the variance-covariance matrix $\text{Var}\left(\hat{\boldsymbol{\beta}}\right)=\left(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\right)^{-1}$. The confidence interval for π_{11} can be calculated as follows:

90% CI for α: $\hat{\alpha}\pm z_{0.95}\times SE\left(\hat{\alpha}\right)=-2.94444\pm 1.645\times 0.205196=\left(-3.281947,-2.60685\right)$

90% CI for π_{11}: $\left[1/\left(1+\exp\left(3.281947\right)\right),1/\left(1+\exp\left(2.60685\right)\right)\right]=\left(0.036194,0.068694\right)$

In the same way, we can calculate the confidence intervals for π_{12} and π_{32}. By the way, we need to know the information on the variance-covariance of parameter estimates, which can be obtained in the form of variance-covariance matrix. But the calculations are not that simple due to the fact that

$$\text{Var}\left(\hat{\alpha}+\hat{\gamma}_2\right)=\text{Var}\left(\hat{\alpha}\right)+\text{Var}\left(\hat{\gamma}_2\right)+2\text{Cov}\left(\hat{\alpha},\hat{\gamma}_2\right) \tag{42}$$

$$\text{Var}\left(\hat{\alpha}+\hat{\gamma}_2+\hat{\delta}_{32}\right)=\text{Var}\left(\hat{\alpha}\right)+\text{Var}\left(\hat{\gamma}_2\right)+\text{Var}\left(\hat{\delta}_{32}\right)+2\text{Cov}\left(\hat{\alpha},\hat{\gamma}_2\right)+2\text{Cov}\left(\hat{\alpha},\hat{\delta}_{32}\right)+2\text{Cov}\left(\hat{\gamma}_2,\hat{\delta}_{32}\right) \tag{43}$$

In general, the standard errors corresponding to π_{11}, π_{12} and π_{32} can be calculated by

$$\sqrt{\text{Var}\left(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right)}=\sqrt{\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0} \tag{44}$$

where $\boldsymbol{x}_0'$ is a row vector of $\boldsymbol{X}$ corresponding to π_{11}, π_{12} or π_{32}. There exists commercialized program such as SAS PROC LOGISTIC and/or MINITAB, which performs those cumbersome calculations for us. Once we have those computer programs at our finger tips and the knack of modeling, it is no big deal to calculate the confidence intervals. The point and interval estimates for fractions can be reported as in **Table 6**, which is a translation of MINITAB output.

The confidence intervals can provide us with the information on whether the sample size is large enough or not. For example, the interval estimates for π_{11} and π_{32} in **Table 6** overlap considerably and hence the interval estimates are not discriminative, seemingly because of the fact that the number of replication of subgroup A_3B_2 is not large enough. But we have to bear in mind that this is the conventional way of interpreting the result. We do not have to see the problem in this manner.

Table 6. Point and interval estimates for fractions of model 3.

Factor Levels		1	v_2	x_3v_2	Observed Fractions	Estimated Fractions	90% CI	
i	j	α	γ_2	δ_{32}	y_{ij}/n_{ij}	$\hat{\pi}_{ij}$	Lower Limit	Upper Limit
1	1	1	0	0	25/500	0.05	0.0361594	0.068694
1	2	1	1	0	10/100	0.10	0.0603408	0.161252
3	2	1	1	1	3/100	0.03	0.0116546	0.075030

3.8. Collection of More Data

Sometimes we might have to see the problem in another way. In order to make the two subgroups A_1B_1 and A_3B_2 contrasting, we construct a 2×1 data table as in **Table 7**.

The logistic regression model becomes as simple as the following.

$$\text{logit}\left(\pi_{ij}\right)=\ln\left[\pi_{ij}/\left(1-\pi_{ij}\right)\right]=\boldsymbol{x}_{ij}'\boldsymbol{\beta}=\alpha+\eta x \qquad \left(\text{model } 4\right) \tag{45}$$

The parameter η is adopted as a combination of γ_2 and δ_{32}.The following is the corresponding design matrix and parameter vector.

$$\begin{matrix}\alpha & \eta\end{matrix}$$
$$\boldsymbol{X}=\begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix},\quad \boldsymbol{\beta}=\begin{bmatrix}\alpha\\ \eta\end{bmatrix} \tag{46}$$

The parameter estimates are shown in **Figure 6**.

The interval estimate (0.21, 1.63) for η contains 1, which means $\hat{\eta}$ is insignificant. We need more replications especially with the A_3B_2 subgroup. There exist related sample size formulae for comparing proportions in order to calculate the required sample size for a simple logistic regression model in Hsieh *et al*. [16].

A simulated data are given in **Table 8** where the numbers of replications are increased up to 1000 with the referent subgroup A_1B_1 and up to 500 with the experimental subgroup A_3B_2, while the fractions remain the same.

Model 4 is fitted into the data. The design matrix $\boldsymbol{X}$ and the parameter vector $\boldsymbol{\beta}$ are the same with model 4. The parameter estimates are shown in **Figure 7**.

The estimate $\hat{\eta}$ is significant at 10% and both the lower and upper limits are smaller than 1. We can say with 95% confidence that $\pi_{32}<\pi_{11}$ as explained before. Seen from the last line of **Figure 7**, we can ascertain that the odds ratio is less than one, since we know that if $OR_{32}<1$, then

$$\begin{aligned}&\pi_{32}/\left(1-\pi_{32}\right)<\pi_{11}/\left(1-\pi_{11}\right)\\ &\pi_{32}<\pi_{11}\end{aligned} \tag{47}$$

If we are sure that the upper and lower limits of OR_{32} is smaller than one with 90% confidence, then we can say that the upper limit of two-sided 90% confidence interval corresponds to that of one-sided 95% confidence interval.

Seen from the conventional view point of statistics, the point and interval estimates for fractions can be given as in **Table 9**. Since the confidence intervals overlap, the intervals are not discriminative enough.

In this manner, an experimenter can decide on whether parameter estimates are significant, whether the model is appropriate, whether sample size is large enough, and whether the fraction of candidate subgroup is smaller, until he or she is convinced that the candidate subgroup is superior to the current one.

4. Conclusion and Further Study

In reality, there are many cases where an experimenter has to analyze fraction data, usually provided in the form of percentages or yield rates as the outcomes of an experiment. The input variables are quantitative, qualitative, or both. In this study, the case that the two input variables are all qualitative and the responses are countable is considered for study in order to extend the model in Agresti [10] and leverage the logistic model in Strokes *et al*. [13]. That is to say, an attempt is given to the problem of binary outcomes with two categorical predictors by

Table 7. Second reduced table.

Manufacturing Condition	Number of Units	Units Non-Conforming
A_1B_1	500	25
A_3B_2	100	3

Table 8. Simulated data table.

Manufacturing Condition	Number of Units	Units Non-Conforming
A_1B_1	1000	50
A_3B_2	500	15

Table 9. Point and interval estimates for fractions of model 3.

Factor Levels		1	x	Observed Fractions	Estimated Fractions	90% CI	
i	j	α	η	y_{ij}/n_{ij}	$\hat{\pi}_{ij}$	Lower Limit	Upper Limit
1	1	1	0	50/1000	0.05	0.0398067	0.0626332
3	2	1	1	15/500	0.03	0.0196985	0.0454389

Prediction						90% CI	
Variables	Coeff	SE	Z	P	OR	Lower	Upper
Const	−2.94444	0.205196	−14.35	0.000			
x	−0.531660	0.621085	−0.86	0.392	0.59	0.21	1.63

Figure 6. Parameter estimates for model 4.

Prediction						90% CI	
Variables	Coeff	SE	Z	P	OR	Lower	Upper
Const	−2.94444	0.145095	−20.29	0.000			
x	−0.531660	0.299635	−1.77	0.076	0.59	0.36	0.96

Figure 7. Parameter estimates for model 4.

utilizing logistic regression. In this study, we excluded ANOVA-type analyses, but we adopted ANOVA-model like parameterization, that is, incremental effect modeling.

The optimal manufacturing condition can be ensured, mainly by testing the significance of regression parameters, testing the existence of interactions, estimating related confidence intervals, testing the difference of mean values, and so on. The conventional ANOVA-type analyses are based on the assumption of normality, independence, and equality of variances of experimental observations. For this reason, the ANOVA-type model entails much detrimental to the goodness-of-fit test and the efficient and precise estimation of regression parameters, mainly because the additive property of fraction data is no longer valid, especially when the fractions are close to zero or is near one, as discussed by Jaeger [6].

As it is always the case with logistic regression, the point estimates are more accurate than those of ANOVA-type modeling. Not only is the lower limit always positive, but also the upper limit is always less than one. The significance test of a parameter can be performed by checking whether the confidence interval of the corresponding odds ratio contains one or not, based on the assumption that the null hypothesis (H_0) is true. The interpretation from the viewpoint of logistic regression is not only different from, but also superior to that of ANOVA-type analysis in the statistical sense, as far as the fraction data are concerned. We have to see the model and interpret the result as it is. The model may not be seen from conventional statistical view point.

When dealing with logistic regression with categorical predictors, the generalized estimating equations (GEE) must be utilized to estimate the parameters. These demerits, nevertheless, can be easily overcome by making use of commercialized computer programs such as SAS PROC LOGISTIC and MINITAB. The analyzing process is somewhat different from the conventional statistical analysis method. We might have to abandon our conven-

tional ANOVA-type of way to interpret the analysis result.

The use of logistic regression has its merits: 1) the analyzer can never get a yield rate or defective rate estimate either above 1 or below 0, 2) the estimates for parameters are more efficient and accurate compared to those of the ANOVA-type model since the logistic regression model describes more accurately the intrinsic nature of the count data, and 3) the significance test of regression parameters is easily performed by checking the interval estimates for odds ratios.

There exist other types of transformations, not mentioned in this study, such as probit and complementary log-log transformations, which seems to be worthy of trying. The logistic regression model is sometimes called ordinary logit model to distinguish it from what they call mixed logit model. The mixed logit model could be the next topic of this study.

The analyses method discussed throughout this study can be extended to the case of multiple qualitative predictors for count data, just as there are a variety of models available in the literature, especially in the area of experimental design and regression analysis.

Acknowledgements

This research was supported by Seokyeong University in 2013.

References

[1] Rao, M.M. (1960) Some Asymptotic Results on Transformations in the Analysis of Variance. ARL Technical Note, Aerospace Research Laboratory, Wright-Patterson Air Force Base, Dayton, 60-126.

[2] Wiener, B.J., Brown, D.R. and Michels, K.M. (1971) Statistical Principles in Experimental Design. McGraw Hill, New York.

[3] Toutenburg, H. and Shalabh (2009) Statistical Analysis of Designed Experiments. 3rd Edition, Springer Texts in Statistics.

[4] Cochran, W.G. (1940) The Analysis of Variances When Experimental Errors Follow the Poisson or Binomial Laws. *The Annals of Mathematical Statistics*, **11**, 335-347. http://dx.doi.org/10.1214/aoms/1177731871

[5] Ross, P.J. (1989) Taguchi Techniques for Quality Engineering. McGraw Hill, Singapore.

[6] Jaeger, T.F. (2008) Categorical Data Analysis: Away from ANOVAs (Transformation or Not) and towards Logit Mixed Models. *Journal of Memory and Language*, **59**, 434-446. http://dx.doi.org/10.1016/j.jml.2007.11.007

[7] Dyke, G.V. and Patterson, H.D. (1952) Analysis of Factorial Arrangements When the Data Are Proportions. *Biometrics*, **8**, 1-12. http://dx.doi.org/10.2307/3001521

[8] Montgomery, D.C., Peck, E.A., and Vining, G.G. (2006) Introduction to Linear Regression Analysis. 4th Edition, John Wiley & Sons, Inc., Hoboken.

[9] Kleinbaum, D.G. and Klein, M. (2010) Logistic Regression: A Self Learning Text. 3rd Edition, Springer, New York. http://dx.doi.org/10.1007/978-1-4419-1742-3

[10] Agresti, A. (2013) Categorical Data Analysis. 3rd Edition, John Wiley & Sons Inc., Hoboken.

[11] Dobson, A.J. and Barnett, A.G. (2008) An Introduction to Generalized Linear Models. 3rd Edition, CRC Press, Chapman & Hall, Boca Raton.

[12] Sloan, D. and Morgan, S.P. (1996) An Introduction to Categorical Data Analysis. *Annual Review of Sociology*, **22**, 351-375. http://dx.doi.org/10.1146/annurev.soc.22.1.351

[13] Strokes, M.E., Davis, C.S. and Koch, G.G. (2000) Categorical Data Analysis Using the SAS System. 2nd Edition, SAS Institute Inc., Cary, NC.

[14] Allison, P.D. (1999) Logistic Regression Using the SAS System—Theory and App. SAS Institute Inc., Cary, NC.

[15] Minitab (2011) Minitab Manual. Minitab Inc. http://www.minitab.com/en-us/

[16] Hsieh, F.Y., Bloch, D.L. and Larsen, M.D. (1998) A Simple Method of Sample Size Calculation for Linear and Logistic Regression. *Statistics in Medicine*, **17**, 1623-1634. http://dx.doi.org/10.1002/(SICI)1097-0258(19980730)17:14<1623::AID-SIM871>3.0.CO;2-S