QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL
Qual. Reliab. Engng. Int. 2006; **22**:971–984
Published online 9 May 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/qre.775

Application

Logistic Regression Analysis of Customer Satisfaction Data

Cathy Lawson[1,*,†] and Douglas C. Montgomery[2]
[1] *General Dynamics C4 Systems, 8220 E. Roosevelt MD R1110, Scottsdale, AZ 85257, U.S.A.*
[2] *Arizona State University, Department of Industrial Engineering, Tempe, AZ 85287-5906, U.S.A.*

Variation exists in all processes. Significant work has been done to identify and remove sources of variation in manufacturing processes resulting in large returns for companies. However, business process optimization is an area that has a large potential return for a company. Business processes can be difficult to optimize due to the nature of the output variables associated with them. Business processes tend to have output variables that are binary, nominal or ordinal. Examples of these types of output include whether a particular event occurred, a customer's color preference for a new product and survey questions that assess the extent of the survey respondent's agreement with a particular statement. Output variables that are binary, nominal or ordinal cannot be modeled using ordinary least-squares regression. Logistic regression is a method used to model data where the output is binary, nominal or ordinal. This article provides a review of logistic regression and demonstrates its use in modeling data from a business process involving customer feedback. Copyright © 2006 John Wiley & Sons, Ltd.

Received 1 April 2005; Revised 10 September 2005

KEY WORDS: logistic regression; business process modeling; categorical data analysis; binary response

INTRODUCTION

Continuous improvement programs have focused a great deal on manufacturing process improvement. Statistical methods such as linear regression are tools that are routinely used to model manufacturing process performance. These tools have not been applied as widely to business processes. An aspect of business processes that makes them somewhat more difficult to characterize is the nature of the output variables of the process and the definition of the process activities. Characterizing complex business processes might be better compared to the process of testing a new drug or studying the effects of a particular treatment on a disease rather than comparing it to a manufacturing process. In the biomedical world, the variables under study are often qualitative, long-term in effect, not easily changed or manipulated and the process can be difficult to isolate. In this scenario, the output variable is whether a particular drug or treatment worked, which is dichotomous. The input variables are highly subjective and may include factors such as determining how the amount of stress an individual experiences affects the outcome of the experiment. Many of the experimental issues facing a biochemist in clinical trials are similar to those facing the company trying to understand and model the behavior

*Correspondence to: Cathy Lawson, General Dynamics C4 Systems, 8220 E. Roosevelt MD R1110, Scottsdale, AZ 85257, U.S.A.
†E-mail: cathy.lawson@gdc4s.com

of their complex business processes. In fact, the approach to studying sources of variation in a business process may resemble qualitative research methods used in the social and anthropological sciences[1].

To illustrate, consider a business process that assesses customer satisfaction. A company wants to know how satisfied their key customers are and whether they would buy from the company again. In addition, the company wants to know customers' preferences for receiving invoices from the company. The company can send invoices electronically, through the mail or they can set up a special Web portal for that customer that allows for automatic billing and payment. The company uses a survey tool to collect these data from their key customer base. In addition to these items, the company is interested in knowing whether certain types of customers will respond differently than others. As part of the survey, they will collect demographic information about the customers to determine whether certain characteristics influence customers' behaviors and buying preferences. Once the data are collected, an analysis technique such as linear regression would be useful to determine whether statistically significant relationships exist between the customer input variables and the output variables describing their attitudes and preferences. The problem with this approach is that linear regression requires the output variable to be continuous in nature and these output variables are not. The choice of buying from the company again is binary. The customer's response concerning their level of satisfaction could be binary or ordinal and the customer's billing preference is nominal. The analysis tool that can model these types of responses is logistic regression.

THE LOGISTIC REGRESSION MODEL

Logistic regression is used primarily when the output variable is binary; however, it can be modified to handle data that are nominal or ordinal in nature. We first consider the analysis of binary data and then discuss the nominal and ordinal cases.

In the case of a binary response variable, the regression model would take the following form

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{1}$$

where $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]$ and the response, y_i, can only take on the values of 0 or 1. The response is assumed to be a Bernoulli random variable with a probability distribution

$$\begin{aligned} y_i &= 1, \quad P(y_i = 1) = \pi_i \\ y_i &= 0, \quad P(y_i = 0) = 1 - \pi_i \end{aligned}$$

As $E(\varepsilon_i) = 0$, the expected value of the response is

$$E(y_i) = \mathbf{x}_i'\boldsymbol{\beta} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \tag{2}$$

The expected value of the response function, $E(y_i) = \mathbf{x}_i'\boldsymbol{\beta}$, is the probability that the response takes the value of 1.

One of the problems with this regression model is that the error term can only take on two values.

$$\varepsilon_i = \begin{cases} 1 - \mathbf{x}_i'\boldsymbol{\beta} & \text{when } y = 1 \\ -\mathbf{x}_i'\boldsymbol{\beta} & \text{when } y = 0 \end{cases}$$

As a result, the errors in this model cannot be normally distributed. Another problem with the model is that the error variance is not constant.

$$\begin{aligned} \sigma^2 &= E\{y_i - E(y_i)\}^2 \\ &= (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i) \\ &= \pi_i(1 - \pi_i) \end{aligned} \tag{3}$$

DOI: 10.1002/qre

which is the same as

$$\sigma^2 = E(y_i)[1 - E(y_i)]$$

This indicates that the variance of the error term is a function of the mean. Finally, there is a constraint in the response function of

$$0 \leq E(y_i) = \pi_i \leq 1$$

Under the assumption of a linear response function, it is possible to fit a model where predicted values of the response could lie outside the 0, 1 range.

When the response is binary, there is evidence to suggest that the shape of the response variable is nonlinear. An s-shaped curve is often used. This function is called the logistic response function or logit and has the form

$$E(y) = \pi = \frac{\mathrm{e}^{g(\mathbf{x})}}{1 + \mathrm{e}^{g(\mathbf{x})}} = \frac{1}{1 + \mathrm{e}^{-g(\mathbf{x})}} \tag{4}$$

where $g(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$. The term $g(\mathbf{x})$ is defined as the linear predictor of the logistic regression model. Through manipulation of Equation (4), we find that

$$g(\mathbf{x}) = \ln \frac{\pi}{1 - \pi} \tag{5}$$

The logistic regression model is fit by obtaining estimates of the parameters (β) using the method of maximum likelihood. There will be $p + 1$ likelihood equations that are obtained by differentiating the log-likelihood function with respect to the $p + 1$ coefficients. The probability distribution of each sample observation is

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \quad \text{for } i = 1, 2, \ldots, n \tag{6}$$

As the observations are independent and take on the values of 0 or 1, the likelihood function is

$$L(y_1, y_2, \ldots, y_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{7}$$

Taking the log of the likelihood function yields

$$\ln L(y_1, y_2, \ldots, y_n, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\pi_i}{1 - \pi_i} \right) \right] + \sum_{i=1}^{n} \ln(1 - \pi_i) \tag{8}$$

If there are multiple observations at each level of the x variables, then Equation (8) becomes

$$\ln L(\mathbf{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \pi_i + \sum_{i=1}^{n} n_i \ln(1 - \pi_i) - \sum_{i=1}^{n} y_i \ln(1 - \pi_i) \tag{9}$$

The log-likelihood can be maximized by using a Newton–Raphson scheme, or equivalently an iteratively re-weighted least-squares (IRLS) procedure can be used to estimate the parameters. Several statistical software packages use the IRLS approach including SAS and Minitab. The estimate of the linear predictor is

$$\hat{g}(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} \tag{10}$$

and the fitted value of the logistic regression model is

$$\begin{aligned} \hat{y} &= \frac{1}{1 + \exp(-\mathbf{x}'\boldsymbol{\beta})} \\ &= \hat{\pi} \end{aligned} \tag{11}$$

For more information on the method of maximum likelihood, see Montgomery *et al.*[2], Hosmer and Lemeshow[3] and Myers *et al.*[4].

DOI: 10.1002/qre

Table I. Summary of customer data

Repeat customer	Account size	Would not buy again	Would buy again	Total
Yes	Small	5	21	27
Yes	Large	4	15	19
No	Small	9	13	10
No	Large	10	11	21
Total		28	57	85

Table II. Summary of estimated probabilities from linear predictor

Repeat customer	Size of account	Probability of buying again
Yes	Small	0.803
Yes	Large	0.795
No	Small	0.532
No	Large	0.518

MODELING CUSTOMER SATISFACTION WITH BINARY LOGISTIC REGRESSION

Data were collected from 85 customer satisfaction surveys. Customers were asked whether they would consider buying from the company again in the future. Customers were described and categorized by two demographic variables, whether the customer was a repeat customer to the company and the size of the account that the customer had with the company. The company was interested in knowing whether repeat customers would continue to buy again from the company. They also wanted to know whether the size of the account the customer had with the company was influential in determining whether the customer would buy again. The response variable is binary in this analysis, whether the customer would buy again or not. The input variable of repeat customer is also binary, depending on whether the customer is new to the company or not. The size of the account is continuous and reflects the amount of dollars spent by each customer. For this analysis this variable was coded categorically into groups of small and large. Small was any account below $100 000 in value. Large accounts were greater than $100 000 in value. A summary of the data is shown in Table I.

Logistic regression is used to analyze the data and determine whether significant relationships exist between the customer demographic variables and the response variable. The estimated linear predictor equation that results is

$$\hat{g}(\mathbf{x}) = 0.075 + 1.281\ (\text{Repeat customer}) + 0.053\ (\text{Account size})$$

The values for the predictor variables used to solve this equation are 0 and 1. If the customer is not a repeat customer, then a value of 0 is used for that x value and the second term of the equation drops out. If the customer is a repeat customer, then $x_1 = 1$ and the second term is added to the final result. If the customer has a small account, the predictor for small is set equal to 1. If the customer has a large account, the x value for account size is set to 0 and the equation is evaluated without the third term. The linear predictor can then be converted back to a probability using Equation (11). The estimated probabilities from the logistic regression equation for each condition in this data set are summarized in Table II.

The results in the table indicate that the probability that a repeat customer will buy again is fairly high. Of those customers who were not repeat customers, approximately half would buy again. The raw data show that 57 of the 85 customers surveyed indicated that they would buy again from the company. It appears that repeat customers are more likely to buy again when compared with new customers. It does not look as if there

DOI: 10.1002/qre

is a difference in response when comparing account size. The next step in the analysis is to determine the significance of these results.

ASSESSING MODEL SIGNIFICANCE

Once a model is fit, the next step is to assess the significance of the model. In logistic regression, one method used to assess goodness of fit involves the model deviance statistic, D. The guiding principle behind this statistic is to compare the log-likelihood of the fitted model to the saturated model. The saturated model is the one that contains as many parameters as there are data points. The value of the log-likelihood function for the fitted model can never exceed the log-likelihood function of the saturated model because the fitted model contains fewer parameters. The model deviance is calculated using the following equation:

$$D(\mathbf{b}) = 2[\ln L(\text{saturated model} - \ln L(\mathbf{b}))] \tag{12}$$

In the saturated model, each observation is replaced by a parameter whose estimate is y_i/n_i:

$$D(\mathbf{b}) = 2\sum_{i=1}^{m}\left[y_i \ln \frac{y_i}{n_i \cdot \hat{\pi}(\mathbf{x}_i)} + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i(1 - \hat{\pi}(\mathbf{x}_i))}\right)\right] \tag{13}$$

If the logistic regression model is correct and the sample size is large, the model deviance has a χ^2 distribution with $n - p$ degrees of freedom. Large values of the model deviance would indicate the model is not correct and small values would indicate that the fitted model fits the data almost as well as the saturated model. The deviance is analogous to the residual sum of squares in linear regression. The test criteria for this model statistic is:

- if $D \leq \chi^2_{\alpha,n-p}$, then conclude that the fitted model is adequate;
- if $D > \chi^2_{\alpha,n-p}$, then conclude that the fitted model is not adequate.

To assess the overall significance of the logistic regression model, the maximum log-likelihood of the full model, $\ln L(\mathbf{b})$, is compared with the maximum log-likelihood of the model with constant success probability. The constant success probability model is

$$\hat{y} = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \tag{14}$$

The maximum likelihood estimator of the constant success probability is y/n where y is the total number of successes and n is the total number of observations. The maximum value of the log-likelihood in this case is

$$\ln L = y\ln(y) + (n - y)\ln(n - y) - n\ln(n) \tag{15}$$

The test statistic for overall significance of regression is

$$G = 2\left[\sum_{i=1}^{m} y_i \ln \hat{\pi}(\mathbf{x}_i) + \sum_{i=1}^{m}(n_i - y_i)\ln(1 - \hat{\pi}(\mathbf{x}_i))\right] - [y\ln(y) + (n - y)\ln(n - y) - n\ln(n)] \tag{16}$$

Large values of this statistic imply that at least one β does not equal 0. Under the hypothesis that all the β_p are equal to zero, the statistic G follows a χ^2 distribution with r degrees of freedom where r is the difference in degrees of freedom between the two models being compared. The test criteria is:

- if $G \geq \chi^2_{\alpha,r}$, then reject the null hypothesis;
- if $G < \chi^2_{\alpha,r}$, then do not reject the null hypothesis.

DOI: 10.1002/qre

Table III. Logistic regression table for binary data

Predictor	Coefficient	Standard error coefficient	Z-value	P-value
Constant	0.075	0.391	0.19	0.85
Repeat	1.281	0.491	2.61	0.009
Account size	0.053	0.485	0.11	0.91

Table IV. Logistic regression table for reduced model

Predictor	Coefficient	Standard error coefficient	Z-value	P-value
Constant	0.1	0.317	0.32	0.75
Repeat customer	1.286	0.489	2.63	0.009

For more information on assessing model fit for logistic regression, see Montgomery *et al.*[2], Collett[5] and Cox and Snell[6].

A method of assessing individual parameter significance is the Wald test. The Wald test compares the maximum likelihood estimate of each β_j with an estimate of its standard error:

$$Z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{17}$$

When the sample size is large, this statistic follows a standard normal distribution. See Daniel[7], Hosmer and Lemeshow[3] and Agresti[8].

The model fit statistics for these data are shown in Table III. The model statistics are

$$G = 7.363, \quad \text{DF} = 2, \quad P\text{-value} = 0.025$$
$$D = 0.011, \quad \text{DF} = 1, \quad P\text{-value} = 0.916$$

The G statistic tests the null hypothesis that the model parameters, β_1 and β_2, are equal to zero. These results show that the null hypothesis should be rejected with the conclusion that at least one parameter is significant in the model. The deviance statistic, D, indicates that the fitted model is close to the saturated model in deviance values and therefore is a good fit. The Wald test indicates that only the repeat customer variable is significant with a Z-value of 2.61 and a P-value of 0.009. This is evident from the estimates of probability made from the linear predictor. There is a change in probability level between repeat and new customers, but there is little change in the probability level between small and large accounts. Knowing that account size makes little difference in determining whether a customer will buy again, the model could be re-fit without the account size variable. The resulting linear predictor is

$$\hat{g}(x) = 0.1 + 1.286 \text{ (Repeat customer)}$$

with estimated probabilities of buying again of 0.52 for new customers and 0.8 for repeat customers. The model statistics for this reduced model are shown in Table IV. The model fit statistics are

$$G = 7.351, \quad \text{DF} = 1, \quad P\text{-value} = 0.007$$
$$D = 0.013, \quad \text{DF} = 1, \quad P\text{-value} = 0.914$$

DOI: 10.1002/qre

ODDS RATIO

The odds ratio is a measure of association that compares the odds of getting one response outcome over the other response outcome[8]. The odds are defined to be

$$\Omega = \frac{\pi}{1-\pi} \tag{18}$$

where π is the probability of occurrence at a particular value of x. In this example, as the binary response refers to buying again from the company, the odds are a measure that estimates the probability of the customer buying again over not buying again. The odds ratio compares the odds from one set of input conditions to another set of inputs. Therefore, the odds ratio is

$$\theta = \frac{\Omega_1}{\Omega_2} \tag{19}$$

The odds ratio can be any non-negative number. The condition where the odds being compared are equal give an odds ratio of 1 and implies independence between the two conditions being compared. When $1 < \theta < \infty$, then the conditions that create Ω_1 are the more likely outcome. When $0 < \theta < 1$, the conditions that create Ω_2 are the more likely outcome.

Using sample data from an experiment, the odds ratio is estimated using the cell counts of the data:

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{20}$$

As odds ratios are calculated for a single variable, the contingency table of data is collapsed on all the variables but one, so that the odds ratio can be calculated for that variable. It can be shown that an equivalent method of estimating the odds ratio is

$$\hat{\theta} = \mathrm{e}^{\hat{\beta}_1} \tag{21}$$

In the model that had both repeat customer and account size, the odds ratio for repeat customer is 3.6 and the odds ratio for account size is 1.05. For the repeat customer variable, this can be interpreted to mean that the odds of a repeat customer buying again from the company are 3.6 times more likely than a new customer. As the odds ratio for account size is nearly 1, this can be interpreted to mean that the likelihood of a customer buying again from the company is independent of the size of the account.

Taking the log of the $\hat{\theta}$ estimator yields a statistic that has been shown to have a normal distribution with mean $\log \hat{\theta}$ and standard error, $\hat{\sigma}$ where

$$\hat{\sigma} = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{1/2} \tag{22}$$

The Wald confidence interval for $\log \hat{\theta}$ is

$$\log \hat{\theta} \pm z_{\alpha/2}\hat{\sigma} \tag{23}$$

To obtain the confidence interval for the actual odds ratio, one takes the anti-log from Equation (21). On conversion from the log scale, the confidence intervals for the odds ratio become highly skewed to the right[9]. In this example, the 95% confidence interval of the odds ratio for the repeat customer variable is 1.37–9.43. The 95% confidence interval for the account size variable is 0.41–2.75. As this interval contains the value of 1, the conclusion is that the change in the response is independent of this variable and this is consistent with the other statistical tests from this model.

DOI: 10.1002/qre

OTHER LINK FUNCTIONS FOR LOGISTIC REGRESSION

Logistic regression is a special case of the generalized linear model (GLM). The GLM is an extension of the standard normal-theory linear regression model to incorporate both linear and nonlinear models and any response distribution that belongs to the exponential family. Prominent members of the exponential family include the normal, inverse normal, binomial, Poisson, exponential and gamma distributions.

Every GLM has three components:

(1) a response distribution that is an exponential family member;
(2) a linear predictor that involves a set of predictor variables or covariates; and
(3) a link function that connects the linear predictor to the mean of the response variable.

In logistic regression, the response distribution is binomial, we have denoted the linear predictor by $\hat{g}(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, and we have used the logistic link function

$$E(y) = \frac{1}{1 + \mathrm{e}^{-g(\mathbf{x})}} \tag{24}$$

Note how the logistic link relates the linear predictor to the expected value of y, the mean response.

Although the logistic link is widely used with logistic regression, there are two other potential choices. The first of these is to use a normal cumulative function as the link,

$$E(y) = \Phi[g(\mathbf{x})] \tag{25}$$

This also provides a symmetric s-shaped curve relating the mean response to the linear predictor. When the normal cumulative $\Phi(\cdot)$ is used as the link, it is customary to refer to the procedure as probit analysis. Another choice of link function is the complimentary log–log function

$$E(y) = 1 - \mathrm{e}^{-\exp[g(\mathbf{x})]} \tag{26}$$

This provides an asymmetric s-shaped function relating the linear predictor to the linear response.

To illustrate these three link functions, suppose that the linear predictor is

$$g(x) = -1 + 2x$$

Figure 1 shows $E(y)$ as a function of x over the range $-3 \leq x \leq +3$. Note that the logit and probit links are very similar except near the extremes when $E(y)$ is very near zero or unity. These links are symmetric and result in $E(y) = 0.5$ at the point where $x = -\beta_0/\beta_1$. This plot clearly shows that the complimentary log–log link is asymmetric. A careful analyst should always consider the selection of the link function. However, it usually requires large sample sizes to see any major differences in the model resulting from the choice of the link function.

NOMINAL LOGISTIC REGRESSION

Logistic regression is used most frequently when the response variable is binary but it is also applicable to multi-level responses. Multi-level responses may be ordinal or nominal. Nominal data are categorical with no implied order. Ordinal data can be ordered in some manner. In the customer survey example, a nominal response variable is the customer's preferred invoicing mechanism, i.e. mail, electronic invoice or Web portal. There is no implied order among these choices. They are simply categories of the various ways the billing process can be handled. When there are more than two nominal response categories, logistic regression fits a model using generalized logits[9]. The generalized logit is defined as

$$h(x)_j = \log\left[\frac{\pi_j}{\pi_r}\right] \tag{27}$$

DOI: 10.1002/qre

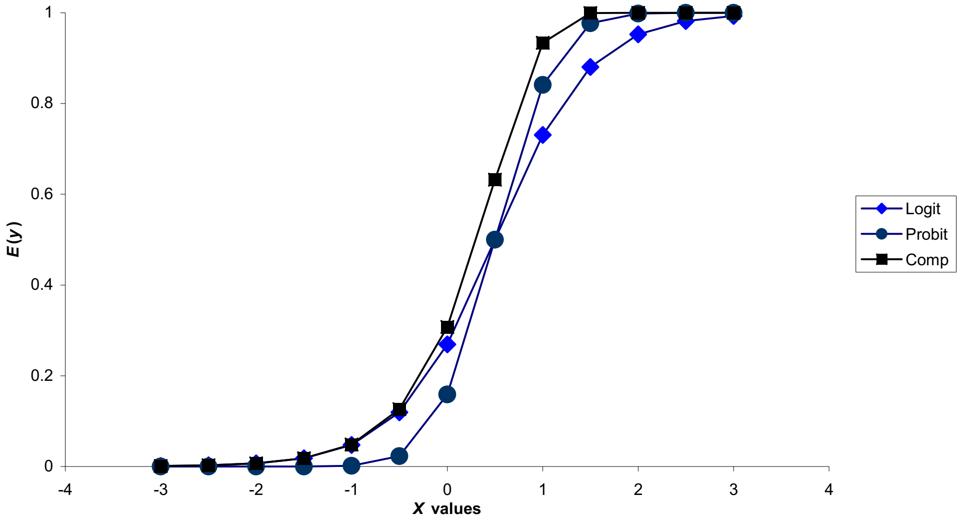

Figure 1. Graph of three link functions

Table V. Customer survey data with nominal response

Repeat customer	Size of account	Invoice mail	Invoice electronically	Invoice Web portal	Total
Yes	Small	12	17	26	55
Yes	Large	7	12	36	55
No	Small	11	15	16	42
No	Large	8	12	30	50
Totals		38	56	108	202

for $j = 1, 2, \ldots, (r - 1)$ nominal response categories. A logit is formed for the probability of each succeeding category over the last response category. The generalized logits for a three-level response such as that in the customer survey example would be

$$h(x)_1 = \log\left[\frac{\pi_1}{\pi_3}\right], \quad h(x)_2 = \log\left[\frac{\pi_2}{\pi_3}\right]$$

These logits will change depending on which nominal response is designated as the last. The last nominal category is often called the reference category. Owing to the way in which the logits are calculated, the reference category becomes the category against which all the other responses are compared. This is especially useful if there is a standard or control group in the experiment and the researcher is interested in knowing the changes in response when compared with the standard. When there are more than two response categories, there will be more than one equation developed to fit the data. The generalized linear predictor model is

$$\hat{g}(x)_k = \beta_{0_k} + \mathbf{x}_i' \boldsymbol{\beta}_k \tag{28}$$

where k is the index of the logits. This implies that there will be separate intercept and regression parameters for each logit.

The data from the customer survey with the nominal response are modeled to illustrate this procedure. The summary data are shown in Table V.

The Web portal is a new mechanism for invoicing and the company wants to know how customers respond to this new billing approach. Therefore, the Web portal response will be used as the reference for the analysis. The logistic model fit for both logits are shown in Tables VI and VII.

DOI: 10.1002/qre

Table VI. Logistic regression model for logit 1 (mail/Web)

Predictor	Coefficient	Standard error coefficient	Z-value	P-value	Odds ratio	95% CI lower	95% CI upper
Intercept	−1.303	0.341	−3.82	0.00			
Repeat	−0.351	0.385	−0.91	0.362	0.7	0.33	1.5
Account	0.899	0.388	2.32	0.02	2.46	1.15	5.26

Table VII. Logistic regression model for logit 2 (electronic/Web)

Predictor	Coefficient	Standard error coefficient	Z-value	P-value	Odds ratio	95% CI lower	95% CI upper
Intercept	−0.871	0.293	−2.98	0.00			
Repeat	−0.271	0.336	−0.81	0.42	0.76	0.39	1.47
Account	0.756	0.336	2.25	0.024	2.13	1.10	4.11

Table VIII. Model summary statistics

Method	Statistic	Degrees of freedom	P-value
G	9.034	4	0.06
D	0.07	2	0.965

Table IX. Predicted probabilities from the first linear predictor

Repeat customer	Size of account	Probability of preferring mail to Web portal
Yes	Small	0.319
Yes	Large	0.160
No	Small	0.400
No	Large	0.214

The goodness of fit and model adequacy tests are computed on the entire data set. Table VIII shows a summary of these statistics. The model statistics indicate that the model is a good fit to the data. The G statistic with a P-value of 0.06 indicates that the null hypothesis that the estimates of the $\boldsymbol{\beta}$ coefficients are equal to zero should be rejected. The model deviance statistic with a P-value of 0.965 indicates the fitted model is adequate. The linear predictor equations indicate that the customers' preferred billing method is related to account size, but not to whether the customer is a repeat customer. Logit 1 compares a mailed form with a Web portal invoicing method. The linear predictor is

$$\hat{g}(\mathbf{x})_1 = -1.303 - 0.351\ (\text{Repeat customer}) + 0.899\ (\text{Account size})$$

The x value for repeat customer is 0 for a new customer and 1 for a repeat customer. The x value for account size is 0 for large accounts and 1 for small accounts. The predictions from this linear predictor are summarized in Table IX.

The analysis shows that customers with a large account have a significantly lower preference for mailing over Web portal invoicing. This relationship holds regardless of whether the customer is a new customer or has done repeated business with the company. The odds ratio for repeat customers is 0.7 with a 95% confidence from 0.33 to 1.5. As this interval includes 1, the preference of mail to Web portal invoicing is independent of the repeat

DOI: 10.1002/qre

Table X. Predicted probabilities from the second linear predictor

Repeat customer	Size of account	Probability of preferring electronic invoice to Web portal
Yes	Small	0.404
Yes	Large	0.242
No	Small	0.471
No	Large	0.295

or new customer. The odds ratio of the account size is 2.46 with a 95% confidence interval from 1.15 to 5.26. There is a significant relationship between account size and invoicing preference. The point estimate of the odds ratio indicates that customers with smaller accounts are 2.46 times more likely to prefer mailing to Web portal invoicing versus customers with larger accounts.

Logit 2 compares electronic invoicing via e-mail to the Web portal invoicing method. The linear predictor for logit 2 is

$$\hat{g}(\mathbf{x})_2 = -0.871 - 0.271\ (\text{Repeat customer}) + 0.756\ (\text{Account size})$$

The predictions for logit 2 are shown in Table X.

As with logit 1, the same types of conclusions can be drawn from this analysis. Customers with larger accounts have a higher preference for the Web portal invoicing method. The reduced models with only the account size in the equation are:

- Mail/Web: $g(x)_1 = -1.481 + 0.879$ (Account size);
- Elec/Web: $g(x)_2 = -1.011 + 0.739$ (Account size).

The nominal logistic regression method allows the researcher to analyze categorical response data and draw conclusions about the effect of input variables on these responses. As the logit function is designed to compare the probabilities of two events, the generalized logit method allows for analysis of multi-level response data through the formation of subset paired groups. The analysis is conducted as if the response is binary and separate logit functions are created for each nominal response compared with the designated reference category. The analysis of the model is performed in the same manner as binary logistic regression. The goodness of fit and model adequacy statistics are also identical. The interpretation of model results is more complex, but draws on the methods used in binary logistic regression.

ORDINAL LOGISTIC REGRESSION

If the response is multi-level but can be ordered from lowest to highest in some manner, ordinal logistic regression can be used. In the customer survey example the ordinal response comes from the question asking the customers how satisfied they are with the company's performance. The satisfaction rating scale goes from 1 to 5, with 5 being the most satisfied. In nominal logistic regression, generalized logits are computed that are based on each response level compared with one designated reference response level. In ordinal regression, cumulative logits are computed that are based on cumulative probabilities of the response levels. This approach, known as the proportional odds model, takes the rank ordering of the response into account. With this model the probability of an equal or smaller response, $Y \leq k$, is compared with the probability of a larger response, $Y > k$,

DOI: 10.1002/qre

Table XI. Ordinal response data from customer surveys

Repeat customer	Size of account	Very dissatisfied 1	Dissatisfied 2	Neutral Neutral 3	Satisfied 4	Very satisfied 5	Total
Yes	Small	0	0	7	10	30	37
Yes	Large	0	0	5	13	25	43
No	Small	0	1	15	33	15	63
No	Large	0	0	16	28	12	56
Total		0	1	43	84	82	210

Table XII. Ordinal logistic regression table

Predictor	Coefficient	Standard error coefficient	Z-value	P-value	Odds ratio	95 % CI lower	95 % CI upper
Intercept (1)	−0.818	0.237	−3.45	0.001			
Intercept (2)	1.163	0.246	4.73	0.000			
Repeat	−1.443	0.281	−5.13	0.000	0.24	0.14	0.41
Account	−0.169	0.265	−0.64	0.522	0.84	0.5	1.42

where k is the rank of the ordinal categories.

$$h_k(x) = \ln\left[\frac{P(Y \le k|\mathbf{x})}{P(Y > k|\mathbf{x})}\right] \tag{29}$$

$$\hat{g}(x)_k = \beta_{0k} + \mathbf{x}'\boldsymbol{\beta} \tag{30}$$

The method used to fit these models is generally the same as for the nominal logistic regression case. There will be a different linear predictor equation for each logit pair. The constraint placed on this model is that the log odds does not depend on the outcome category. Inferences from the proportional odds model look at the direction of the response, but do not focus on specific outcome categories. The results are simpler to describe because the models are fitted with the same set of slope parameters but with different intercepts for each logit. See Agresti[10], Stokes *et al.*[9] and Ananth and Kleinbaum[11]. The data from the customer survey data are summarized in Table XI.

Given that there were no very dissatisfied customers surveyed and only one dissatisfied customer, the analysis will include only the top three categories of customer satisfaction. The proportional odds model will produce one model that compares the satisfied response with the neutral response and then a second model that compares neutral and satisfied with very satisfied. The model output is shown in Table XII. The results indicate that the repeat customer variable is significant, but the account size variable is not. The first linear predictor comparing neutral with satisfied responses is

$$\hat{g}(\mathbf{x})_1 = -0.8182 - 1.4435\ (\text{Repeat customer}) - 0.1698\ (\text{Account size})$$

As in the prior models, the predictor values for repeat customer and account size are 0 and 1, where 0 is assigned to a new customer and those with large accounts; and 1 is assigned to a repeat customer and those with small accounts. The predicted probabilities from this linear predictor are shown in Table XIII.

This linear predictor indicates that repeat customers with large accounts are much less likely to rate their satisfaction as neutral over satisfied than the other groups, although all groups are less likely to rate their satisfaction neutral over satisfied. The linear predictor from logit 2 is

$$\hat{g}(\mathbf{x})_2 = 1.1632 - 1.4435\ (\text{Repeat customer}) - 0.1698\ (\text{Account size})$$

The predicted probabilities from this linear predictor are shown in Table XIV.

DOI: 10.1002/qre

Table XIII. Estimated probabilities for the first linear predictor

Repeat customer	Size of account	Probability of selecting neutral over satisfied
Yes	Small	0.311
Yes	Large	0.094
No	Small	0.271
No	Large	0.306

Table XIV. Estimated probabilities from the second linear predictor

Repeat customer	Size of account	Probability of selecting neutral or satisfied over very satisfied
Yes	Small	0.389
Yes	Large	0.430
No	Small	0.729
No	Large	0.762

Table XV. Model summary statistics for ordinal data

Method	Statistic	Degrees of freedom	P-value
G	27.934	2	0.00
D	5.677	4	0.225

The results from this linear predictor indicate that new customers are more likely to rate either satisfied or neutral over very satisfied. Repeat customers are somewhat more likely to rate very satisfied over satisfied or neutral. There are no differences reflected in the results based on the size of a customer's account. The model adequacy and goodness of fit statistics are shown in Table XV.

The model statistics indicate the model fit is adequate. The G statistic is interpreted the same as for binary logistic regression. It is a test that all slopes are zero. As the P-value is small, the null hypothesis that the slope parameters are all equal to zero is rejected. The deviance statistic is a measure of the difference between the fitted model and the saturated model. As its P-value is >0.05, the null hypothesis is not rejected and the model fit is assumed to be adequate.

CONCLUSIONS

Logistic regression is a flexible statistical modeling technique that provides a powerful mechanism to analyze response data and establish relationships between variables that are not continuous and normally distributed. Response data that are binary or multi-level categorical can be modeled using logistic regression. For industry, this tool provides a mechanism to study and improve processes that have unusual output variables. Business processes are one family of processes that tend to have binary or categorical responses as well as inputs. Business processes account for much of the activity in a company and improvement in the performance of these types of processes will have a significant effect on the bottom line. In this research it was shown how three types

DOI: 10.1002/qre

of logistic regression models could be applied to a business process that is common to many companies and industries, the customer satisfaction assessment process. Customer survey data are typically not continuously distributed and are often categorical. In the example presented here, the researcher looked at two customer demographic variables, repeat customer and size of account to see what relationships existed between that data and the customers' level of satisfaction, billing preferences and attitudes about buying from the company again. It was learned that repeat customers perceive satisfaction differently than new customers and that established customers are more likely to purchase again from the company. It was learned that customers with different account sizes have different preferences in billing methods. This analysis provides the company with a means of taking action to improve their overall customer relationships. Further analysis can be performed using other customer variables to see whether additional insights can be gained to further improve the process.

REFERENCES

1. Bartunek JM, Seo M. Qualitative research can add new meanings to quantitative research. *Journal of Organizational Behavior* 2002; **23**:110–113.
2. Montgomery DC, Peck EA, Vining GG. *Introduction to Linear Regression Analysis*. Wiley: New York, 2001.
3. Hosmer D, Lemeshow S. *Applied Logistic Regression*. Wiley: New York, 2000.
4. Myers RH, Montgomery DC, Vining GG. *General Linear Models with Applications in Engineering and the Sciences*. Wiley: New York, 2002.
5. Collett D. *Modeling Binary Data*. Chapman and Hall: Boca Raton, FL, 2003.
6. Cox DR, Snell EJ. *Analysis of Binary Data*. Chapman and Hall: London, 1989.
7. Daniel WW. *Applied Nonparametric Statistics* (2nd edn). Duxbury: Pacific Grove, CA, 1990.
8. Agresti A. *Categorical Data Analysis*. Wiley: New York, 2002.
9. Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using the SAS System*. SAS Institute: Cary, NC, 1991.
10. Agresti A. *Analysis of Ordinal Categorical Data*. Wiley: New York, 1984.
11. Ananth CV, Kleinbaum DG. Regression models for ordinal data: A review of methods and application. *International Journal of Epidemiology* 1997; **26**:1323–1333.

Authors' biographies

Cathy Lawson is a Division Quality Manager in General Dynamics C4 Systems. She has a BSc in Bio Medical Engineering and a MSc and PhD in Industrial Engineering from Arizona State University. She has published in *Quality and Reliability Engineering International*, the *IEEE Transactions on Reliability* and has co-authored several other publications on Six Sigma and process characterization.

Douglas C. Montgomery is the ASU Foundation Professor of Engineering and Professor of Statistics at Arizona State University. He is an author of 15 books and over 170 technical papers. He is a recipient of the Shewhart Medal, the Brumbaugh Award, the Hunter Award and the Shewell Award (twice) from the American Society for Quality Control. He is also a recipient of the Ellis R. Ott Award. He is the one of the Chief Editors of *Quality and Reliability Engineering International*, a former Editor of the *Journal of Quality Technology* and a member of several other editorial boards. He is a Fellow of the American Statistical Association, a Fellow of the American Society for Quality Control, A Fellow of the Royal Statistical Society, a Fellow of the Institute of Industrial Engineers, and an Elected Member of the International Statistical Institute. He also serves on the Technical Advisory Board of the United States Golf Association.

DOI: 10.1002/qre