

REGRESSION FOR BINARY RESPONSE VARIABLE

Logistic regression is used to find a relationship between a 0-1 dependent variable and a set of predictor variables. The predictor variables can be continuous or discrete (dummy variables are used in discrete case).

Probit Response Function

Let us look at a simple example first.

Example 1: Suppose a researcher in the Registrar's Office of a major university is trying to find a relationship between the results in STATS 101 class and ACT Scores. The data available to the researcher is shown below (PASS = 1 if the student passed in STATS 101, and 0 otherwise):

ACT	PASS	ACT	PASS	ACT	PASS	ACT	PASS
30	1	23	1	20	0	14	0
13	0	34	1	15	0	11	0
27	1	28	1	26	1	27	1
25	1	31	1	16	0	17	0
20	0	35	1	15	0	35	1
31	1	22	0	25	1	14	0
27	1	21	0	11	0	21	1
30	1	11	0	11	0	30	1
21	0	18	0	24	1	26	1
23	0	18	0	34	1	26	1

The result on STATS 101 depends on the semester score of the student which is related to ACT score by a simple linear regression equation:

$$Y_j^C = \beta_0^C + \beta_1^C X_j + e_j^C$$

where

Y_j^C = continuous score on STATS 101, $0 \leq Y_j^C \leq 100$,

β_0^C = intercept for the continuous response

β_1^C = slope for the continuous response

e_j^C = residual, assumed to be $N(0, \sigma^2)$

The response available to the researcher, however, is not Y_j^C but Y which is given by:

$$Y = \begin{cases} 1 & \text{if } Y_j^C \geq 65 \\ 0 & \text{if } Y_j^C < 65 \end{cases}$$

Let π_j = the probability that j-th students passes STATS 101

$$\begin{aligned}
 &= P(Y_j^C \geq 65) \\
 &= P(\beta_0^C + \beta_1^C X_j + e_j^C \geq 65) \\
 &= P(e_j^C \geq 65 - \beta_0^C - \beta_1^C X_j) \\
 &= P\left(\frac{e_j^C}{\sigma} \geq \frac{65 - \beta_0^C}{\sigma} - \frac{\beta_1^C X_j}{\sigma}\right) \\
 &= P\left(-\frac{e_j^C}{\sigma} \leq \frac{\beta_0^C - 65}{\sigma} + \frac{\beta_1^C X_j}{\sigma}\right) \\
 &= P(Z \leq \beta_0^* + \beta_1^* X_j)
 \end{aligned}$$

or

$\pi_j = E(Y_j) = \Phi(\beta_0^* + \beta_1^* X_j)$, $\Phi(\cdot)$ = standard normal cdf

$\Phi^{-1}(\pi_j) = \pi_j' = \beta_0^* + \beta_1^* X_j = \text{probit response function}$

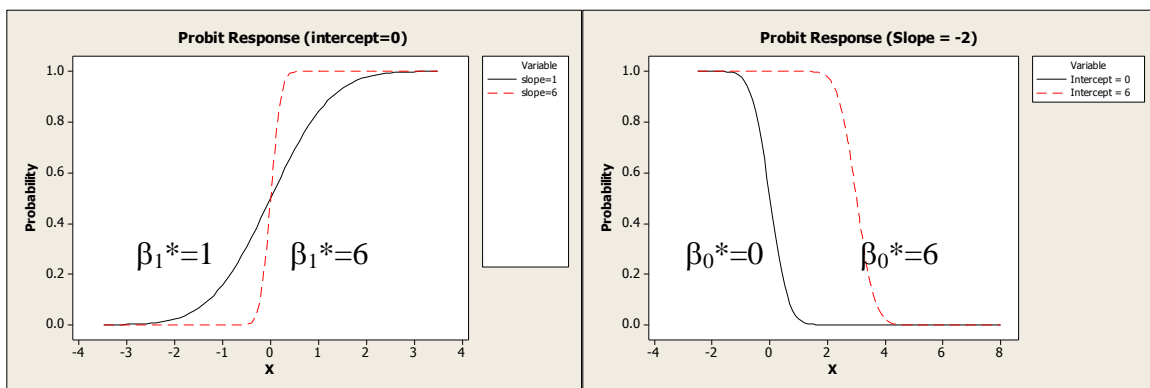


Figure 1(a): Probit for $\beta_0^*=0$

Figure 1(b): Probit for $\beta_1^*=-2$

NOTE: The probit response function is symmetric in the following sense: if the binary response variable Y is recoded as $Y' = 1 - Y$, π' = probability of failure, then the signs of the coefficients are reversed:

$$P(Y'=1) = P(Y = 0) = 1 - \Phi(\beta_0^* + \beta_1^* X) = \Phi(-\beta_0^* - \beta_1^* X)$$

since $\Phi(Z) = 1 - \Phi(-Z)$.

Logistic Response Function

Assuming the error e in the underlying continuous response (Score on STATS 101 in the above example) resulted in the probit response function involving the standard normal cdf $\Phi(\cdot)$ to model π .

If assume the error e in the underlying continuous response to have a logistic distribution, then we get the logistic response function. The logistic distribution with mean 0 and sd $\sigma = \pi/\sqrt{3}$ is given below.

$$X \sim f(x)$$

The pdf $f(x)$ is given by

$$f(x) = \frac{e^{-x}}{[1 + e^{-x}]^2}$$

and the cdf is

$$F(x) = \frac{e^{-x}}{1 + e^{-x}}$$

$$E(X) = 0, \quad \sigma_x = \frac{\pi}{\sqrt{3}}$$

Assuming

$$Y_j^C = \beta_0^C + \beta_1^C X_j + e_j^C$$

where

$$Y_j^C = \text{continuous score on STATS 101, } 0 \leq Y_j^C \leq 100,$$

$$\beta_0^C = \text{intercept for the continuous response}$$

$$\beta_1^C = \text{slope for the continuous response}$$

$$e_j^C = \text{residual, assumed to have a logistic distribution with mean 0 and sd } \sigma_c$$

we get:

$$\begin{aligned} \pi_j &= P(Y_j^C \geq 65) \\ &= P(\beta_0^C + \beta_1^C X_j + e_j^C \geq 65) \\ &= P(e_j^C \geq 65 - \beta_0^C - \beta_1^C X_j) \\ &= P\left(\frac{e_j^C}{\sigma_c} \geq \frac{65 - \beta_0^C}{\sigma_c} - \frac{\beta_1^C X_j}{\sigma_c}\right) \\ &= P\left(-\frac{e_j^C}{\sigma_c} \leq \frac{\beta_0^C - 65}{\sigma_c} + \frac{\beta_1^C X_j}{\sigma_c}\right) \end{aligned}$$

Note that $W = -\frac{e^C}{\sigma_C}$ has a logistic distribution with mean 0 and sd = 1.

Hence

$$\begin{aligned}\pi_j &= P\left(W \leq \frac{\beta_0^C - 65}{\sigma_C} + \frac{\beta_1^C X_j}{\sigma_C}\right) = P\left(\frac{\pi W}{\sqrt{3}} \leq \frac{\pi}{\sqrt{3}} \frac{\beta_0^C - 65}{\sigma_C} + \frac{\pi}{\sqrt{3}} \frac{\beta_1^C X_j}{\sigma_C}\right) \\ &= P(U \leq \beta_0 + \beta_1 X_j), \text{ where } U \sim \text{logistic distribution with mean 0 and sd} = \pi/\sqrt{3}.\end{aligned}$$

Hence

$$\pi_j = \frac{e^{\beta_0 + \beta_1 X_j}}{1 + e^{\beta_0 + \beta_1 X_j}}$$

Thus the logistic mean response function is:

$$\pi_j = E(Y_j) = \frac{e^{\beta_0 + \beta_1 X_j}}{1 + e^{\beta_0 + \beta_1 X_j}}$$

which can be rewritten as

$$\frac{\pi_j}{1 - \pi_j} = e^{\beta_0 + \beta_1 X_j}$$

or

$$\ln\left(\frac{\pi_j}{1 - \pi_j}\right) = \beta_0 + \beta_1 X_j$$

i.e., log-odds is a linear function of the predictor X.

In the case of Multiple logistic Regression Model, we have p predictors X_1, X_2, \dots, X_p some of which could be continuous and some binary or dummy variables. The odds ratio is related to the predictors X_1, X_2, \dots, X_p as follows:

$$\frac{\pi_j}{1 - \pi_j} = e^{\beta_0 + \beta_1 X_{1j} + \dots + \beta_p X_{pj}}$$

or

$$\ln\left(\frac{\pi_j}{1 - \pi_j}\right) = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \dots + \beta_p X_{pj}, \quad j = 1, 2, \dots, n$$

i.e., log-odds is a linear function of the predictors X_1, X_2, \dots, X_p .

We will focus on the logistic regression model for a binary response or dependent variable (whether it arises from a continuous response or not) as this is the most commonly used model for a binary response. The method of MAXIMUM LIKELIHOOD is used for estimation of $\beta_0, \beta_1, \dots, \beta_p$. The Probit model is also fitted using the method of maximum likelihood and computer search maximization.

Likelihood Estimation:

The sample in the case of a multiple logistic model is: $\{(Y_1, X_{11}, X_{21}, \dots, X_{p1}), \dots, (Y_n, X_{1n}, X_{2n}, \dots, X_{pn})\}$, where Y_j is a Bernoulli random variable with

$$P(Y_j = 1) = \pi_j$$

$$P(Y_j = 0) = 1 - \pi_j$$

The likelihood of the independent observations Y_1, Y_2, \dots, Y_n is

$$L(\pi_1, \pi_2, \dots, \pi_j | Y_1, Y_2, \dots, Y_n) = \prod_{j=1}^n \pi_j^{Y_j} (1 - \pi_j)^{1 - Y_j}$$

$$\ln(L) = \sum_{j=1}^n [Y_j \ln(\pi_j) + (1 - Y_j) \ln(1 - \pi_j)]$$

$$= \sum_{j=1}^n [Y_j \ln\left(\frac{\pi_j}{1 - \pi_j}\right)] + \sum_{j=1}^n \ln(1 - \pi_j)$$

$$= \sum_{j=1}^n Y_j (\beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \dots + \beta_p X_{pj}) - \sum_{j=1}^n \ln(1 + e^{\beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \dots + \beta_p X_{pj}})$$

A numerical search procedure for maximization is used to find the β 's that maximize the above likelihood function, and then

$$\hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p}}$$

Goodness of Fit Tests

Several measures of testing the significance of the fitted logistic model (or testing goodness of fit) exist. It is easier to use matrix notation to define various statistics. The multiple logistic regression model is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}} = [1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}]^{-1}$$

In matrix notation,

$$\pi = [1 + e^{-X'\beta}]^{-1}$$

$$X_{(p+1) \times 1} = \begin{bmatrix} 1 \\ X_1 \\ \dots \\ X_p \end{bmatrix}, \quad X_i = \begin{bmatrix} 1 \\ X_{1i} \\ \dots \\ X_{pi} \end{bmatrix}, \quad \beta_{(p+1) \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{bmatrix}$$

$$X' \beta = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

$$X_i' \beta = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}$$

Some of the goodness of fit tests for the logistic model (just as in the case of multiple linear regression model) need repeated X-values, as shown below:

Y	X ₁	...	X _p		
Y _{ij}		...		Group 1 replicates has n ₁ observations, X = X ₍₁₀₎	O ₁₁ = # of Y equal to 1 in Group 1, O ₁₀ = # of Y equal to 0 in Group 1, O ₁₀ = n ₁ - O ₁₁
Y _{ij}		...		Group 2 replicates has n ₂ observations, X = X ₍₂₀₎	O ₂₁ = # of Y equal to 1 in Group 2, O ₂₀ = # of Y equal to 0 in Group 2, O ₂₀ = n ₂ - O ₂₁
		...			
Y _{ij}		...		Group K replicates has n _K observations, X = X _(K0)	O _{K1} = # of Y equal to 1 in Group 1, O _{K0} = # of Y equal to 0 in Group K, O _{K0} = n _K - O _{K1}

K = total number of groups of distinct X-values

Y_{ij} = i-th binary response at predictor combination X_j , $j = 1, 2, \dots, K$
 n_j = number of observations in j-th group, $j = 1, 2, \dots, K$

Pearson χ^2 - Statistic

The null hypothesis is:

$$H_0 : \pi_j = [1 + e^{-X_j' \beta}]^{-1}$$

Using the logistic model, π_j is estimated by

$$\hat{\pi}_j = [1 + e^{-X_j' \hat{\beta}}]^{-1}$$

The expected number of cases with $Y_{ij} = 1$ is $E_{j1} = n_j \hat{\pi}_j$,

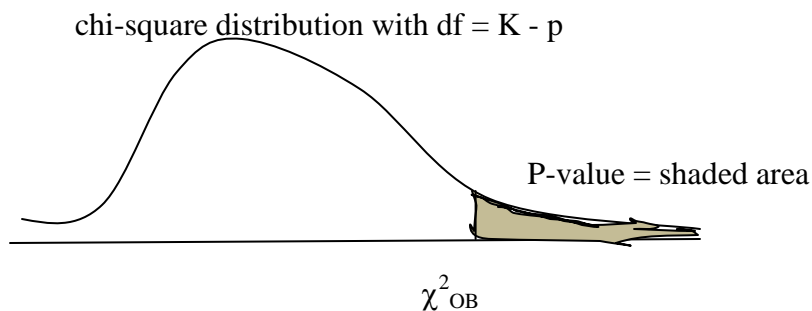
and the expected number of cases with $Y_{ij} = 0$ is $E_{j0} = n_j(1 - \hat{\pi}_j) = n_j - E_{j1}$

$$\chi_{OBS}^2 = \sum_{j=1}^K \sum_{i=0}^1 \frac{(O_{ji} - E_{ji})^2}{E_{ji}}$$

The null distribution of the chi-square statistic is approximately $\chi_{df=K-p}^2$ provided $K > p$.

For the chi-square approximation to be valid, ideally $E_{ji} \geq 5$.

The null hypothesis is rejected if $\chi_{OBS}^2 = \sum_{j=1}^K \sum_{i=0}^1 \frac{(O_{ji} - E_{ji})^2}{E_{ji}} > \chi_{df=K-p, 1-\alpha}^2$



G²-Statistics based on Deviance

In multiple linear regression, lack of fit test is based on testing the reduced model $E(Y_{ij}) = X' \beta$ vs. the full model $E(Y_{ij}) = \mu_{ij}$. Similarly, for the logistic regression, we test $E(Y_{ij}) = \pi_j = [1 + e^{-X_j' \beta}]^{-1}$ (reduced model) vs.

$E(Y_{ij}) = \pi_j, j = 1, 2, \dots, K$ (unknown parameters, not functions of X), the full model

L_R = maximum value of likelihood under the reduced model, $\hat{\pi}_j = [1 + e^{-X_j' \hat{\beta}}]^{-1}$

L_F = maximum value of likelihood under the full model, $\hat{\pi}_j = p_j = \frac{O_{j1}}{n_j}$

The G^2 – statistic is

$$G^2 = -2[\ln(L_R) - \ln(L_F)] = -2\left[O_{j1} \ln\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - O_{j1}) \ln\left(\frac{1 - \hat{\pi}_j}{1 - p_j}\right)\right] = DEV(1, X_1, X_2, \dots, X_p)$$

The null distribution of the G^2 - statistic is approximately $\chi^2_{df=K-p}$ provided $K > p$.

Hosmer-Lemeshow Goodness of Fit Test

For data sets with no replicates ($n_j = 1$ for all j) or very few replicates, Hosmer-Lemeshow test can be used. The data is grouped into classes with similar fitted values $\hat{\pi}_j$ with equal number of cases in each group (approximately), or similar fitted log-odd values. The observed and expected frequencies for each class are calculated, and the standard chi-square goodness of fit test statistic is calculated. The df for the Hosmer-Lemeshow chi-square test is $K - 2$, where K is the number of classes formed.

Measures of Association in Logistic Regression

$$\text{Goodman - Kruskal's Gamma} = \frac{\# \text{ of concordant pairs} - \# \text{ of discordant pairs}}{\# \text{ of concordant pairs} + \# \text{ of discordant pairs}}$$

$$\text{Somers's d Statistic} = \frac{\# \text{ of concordant pairs} - \# \text{ of discordant pairs}}{\# \text{ of concordant pairs} + \# \text{ of discordant pairs} + \# \text{ of pairs tied on Dependent Variable}}$$

$$\text{Kendall's } \tau_a = \frac{\# \text{ of concordant pairs} - \# \text{ of discordant pairs}}{\frac{n(n-1)}{2}}$$

Concordant and discordant pairs are calculated from cross-tabulated data as follows:

VARIABLE 1 (X)	VARIABLE 2 (Y)	
	LOW	HIGH
LOW	a	b
MIDDLE	c	d
HIGH	e	f

A pair $(X_1, Y_1), (X_2, Y_2)$ is concordant if $\text{rank}(X_1) < \text{rank}(X_2)$, $\text{rank}(Y_1) < \text{rank}(Y_2)$

A pair $(X_1, Y_1), (X_2, Y_2)$ is discordant if $\text{rank}(X_1) > \text{rank}(X_2)$, $\text{rank}(Y_1) > \text{rank}(Y_2)$

A pair $(X_1, Y_1), (X_2, Y_2)$ is tied if $\text{rank}(X_1) = \text{rank}(X_2)$, or $\text{rank}(Y_1) = \text{rank}(Y_2)$ or both.

Hence

- any (X_1, Y_1) from cell (LOW, LOW) and any (X_2, Y_2) from cell (MIDDLE, HIGH) is concordant since $\text{LOW} < \text{MIDDLE}$, $\text{LOW} < \text{HIGH}$,
- any (X_1, Y_1) from cell (LOW, HIGH) and any (X_2, Y_2) from cell (MIDDLE, LOW) is discordant since $\text{LOW} < \text{MIDDLE}$, $\text{HIGH} < \text{LOW}$, and
- any pair $(X_1, Y_1), (X_2, Y_2)$ from *same cell*, say both points from (MIDDLE, HIGH), will be tied on both variables.

The total number of concordant pairs = $ad + af + cf$,

total number of discordant pairs = $bc + be + de$,

If the two observations $(X_1, Y_1), (X_2, Y_2)$ come from the same cell, the pair will be tied on both variables, and hence the total number of pairs tied on both variables equals

$$\binom{a}{2} + \binom{b}{2} + \binom{c}{2} + \binom{d}{2} + \binom{e}{2} + \binom{f}{2}$$

Total number of pairs tied on $X_1 = ad + cd + ef$ (both observations come from same row, different columns)

Total number of pairs tied on $X_2 = a(c+e) + ce + b(d+f) + df$ (both observations come from same column, 2nd observation from a row below)

Hence total number of tied pairs =

$$\binom{a}{2} + \binom{b}{2} + \binom{c}{2} + \binom{d}{2} + \binom{e}{2} + \binom{f}{2} + ab + cd + ef + a(c+e) + ce + b(d+f) + df$$

Note that the total number of pairs = $\binom{n}{2} = \frac{n(n-1)}{2}$, $n = a + b + c + d + e + f$

Example (from Statistical Analysis for Public Administration, by Lawrence Giventer, Jones & Bartlett Learning, 2008, page 202). In May 1983, there were an estimated 865 cases of gastrointestinal disease occurrences in Greenville, Florida. Following table summarizes the result of a survey of 187 randomly selected residents of Greenville conducted to evaluate a possible association between water consumption and gastrointestinal disease:

Average # of 8-oz glasses of water drunk per day	Illness		Total
	Not Ill (0)	Ill	
0	72	4	76
1	12	1	13
2	8	9	17
3	4	13	17
≥4	15	49	64
Total	111	76	187

Number of concordant pairs = $72(1+9+13+49) + 12(9+13+49) + 8(13+49) + 4(49) = 6728$

Number of discordant pairs = $4(12+8+4+15) + 1(8+4+15) + 9(4+15) + 13(15) = 549$

Number of pairs tied on Illness (dependent variable) =
 $72(12+8+4+15) + 12(8+4+15) + 8(4+15) + 4(15) +$
 $4(1+9+13+49) + 1(9+13+49) + 9(13+49) + 13(49) = 4898$

Number of pairs tied on Water Consumption = $72(4) + 12(1) + (9) + 4(13) + 15(49) = 1159$

Number of pairs tied on both variables =
 $\binom{72}{2} + \binom{4}{2} + \binom{12}{2} + \binom{1}{2} + \binom{8}{2} + \binom{9}{2} + \binom{4}{2} + \binom{13}{2} + \binom{15}{2} + \binom{49}{2} = 4057$

Check on computations: $6728 + 549 + 4898 + 1159 + 4057 = 17391 = 187(187-1)/2$

Goodman-Kruskal's $\gamma = (6728 - 549)/(6728 + 549) = 0.849$ (strong association)

Somer's $d = (6728 - 549)/(6728 + 549 + 4898) = 0.508$ (strong association)

Kendall's $\tau_a = (6728 - 549)/17391 = 0.36$ (moderate to strong)

EXAMPLES:

Example 1) The fitted simple logistic model for the (PASS, SAT) data of Example 1, obtained from MINITAB, is given below:

Binary Logistic Regression: PASS versus ACT

Link Function: Logit

Response Information

Variable	Value	Count	
PASS	1	21	(Event)
	0	19	
Total		40	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-27.9154	12.6416	-2.21	0.027			
ACT	1.24467	0.564159	2.21	0.027*	3.47	1.15	10.49

*ACT is significant at 5%

Log-Likelihood = -4.582

Test that all slopes are zero: G = 46.189, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	2.09143	18	1.000
Deviance	2.57151	18	1.000
Hosmer-Lemeshow	0.99010	7	0.995

Since P-values for each of the goodness of fit tests > .05, we cannot reject the logistic regression model, i.e., there is no lack of fit.

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Value	Group									Total	
	1	2	3	4	5	6	7	8	9		
1											
Obs	0	0	0	1	2	5	4	5	4	21	
Exp	0.0	0.0	0.0	0.5	2.6	4.9	4.0	5.0	4.0		
0											
Obs	4	5	4	4	2	0	0	0	0	19	
Exp	4.0	5.0	4.0	4.5	1.4	0.1	0.0	0.0	0.0		
Total	4	5	4	5	4	5	4	5	4	40	

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	394	98.7	Somers' D	0.98 all 3 measures
Discordant	2	0.5	Goodman-Kruskal Gamma	0.99 show good prediction
Ties	3	0.8	Kendall's Tau-a	0.50 from logistic
Total	399	100.0		regression

Interpretation of the Fitted Logistic Regression Model

The fitted logistic model is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 = -27.9154 + 1.24467 \times \text{ACT}$$

$$\pi = \frac{e^{-27.9154 + 1.24467 \times \text{ACT}}}{1 + e^{-27.9154 + 1.24467 \times \text{ACT}}}$$

$$\text{Odds}(\text{Passing STATS 101} | \text{ACT} = a + 1) = e^{-27.9154 + 1.24467 \times (a+1)}$$

$$\text{Odds}(\text{Passing STATS 101} | \text{ACT} = a) = e^{-27.9154 + 1.24467 \times a}$$

$$\frac{\text{Odds}(\text{Passing STATS 101} | \text{ACT} = a + 1)}{\text{Odds}(\text{Passing STATS 101} | \text{ACT} = a)} = \frac{e^{-27.9154 + 1.24467 \times (a+1)}}{e^{-27.9154 + 1.24467 \times a}} = e^{1.24467} = 3.47$$

The rate of increase in OR with respect to ACT score = $\exp(1.24467) = 3.47$

i.e., the odds of passing STATS 101 increases by $100 \times (3.47 - 1) / 1 = 247\%$ with unit increase in ACT score.

Figure 2 shows the fitted logistic regression model for data of Example 1.

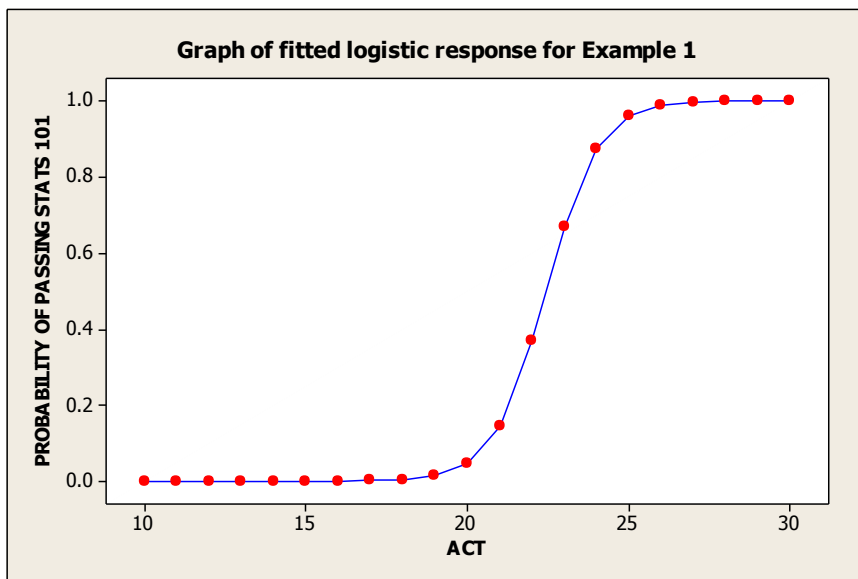


Figure 2: Fitted logistic response for Example 1

Example 2 (Titanic Data Set): The `titanic3` dataset gives the survival status of individual passengers on the Titanic. The principal source for data about Titanic passengers is the [Encyclopedia Titanica](http://www.encyclopedia-titanica.org/). Thomas Cason of UVa filled in missing ages for many passengers, dropped duplicate passengers, and corrected several other errors in creating the datafile `titanic3`. For more information on this dataset, please see <http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic.html>.

The file `titanic3.xlsx` has 14 columns. We are using the first eight columns in this example:

pclass **survived** **name** **sex** **age** **sibsp** **parch** **Gender**

pclass = passenger class (1, 2, 3)

sibsp = Number of Siblings/Spouses Aboard

parch = Number of Parents/Children Aboard

To run logistic regression, we created two dummy variable columns for the variable pclass:

D1 = 1 if pclass = 1 else 0

D2 = 1 if pclass = 2 else 0

The output from MINITAB is given below;

Binary Logistic Regression for `titanic3`: survived versus age, sibsp, ...

Link Function: Logit

Response Information

Variable	Value	Count	
survived	1	427	(Event)
	0	619	
	Total	1046	

* NOTE * 1046 cases were used

* NOTE * 263 cases contained missing values

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI		
						Lower	Upper	
Constant	-1.00209	0.221298	-4.53	0.000				
age	-0.0394887	0.0066346	-5.95	0.000	0.96	0.95	0.97	SIG*
sibsp	-0.352915	0.105356	-3.35	0.001	0.70	0.57	0.86	SIG
parch	0.0743609	0.0999108	0.74	0.457	1.08	0.89	1.31	NOT SIG
Gender	2.55686	0.173281	14.76	0.000	12.90	9.18	18.11	SIG
D1	2.35202	0.228819	10.28	0.000	10.51	6.71	16.45	SIG
D2	0.985266	0.199394	4.94	0.000	2.68	1.81	3.96	SIG

* Term is SIGNIFICANT at 5% test size.

Log-Likelihood = -485.060

Test that all slopes are zero: G = 444.501, DF = 6, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	625.938	608	0.299*
Deviance	617.171	608	0.389*
Hosmer-Lemeshow	26.930	8	0.001

Since P-values for each of the goodness of fit tests > .05, we cannot reject the logistic regression model, i.e., there is no lack of fit.

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Value	Group										Total	
	1	2	3	4	5	6	7	8	9	10		
1												
Obs	7	19	17	14	23	37	43	74	92	101	427	
Exp	6.7	10.3	13.3	16.8	26.7	40.2	58.9	72.8	85.2	96.2		
0												
Obs	99	85	87	90	82	67	63	31	12	3	619	
Exp	99.3	93.7	90.7	87.2	78.3	63.8	47.1	32.2	18.8	7.8		
Total	106	104	104	104	105	104	106	105	104	104	1046	

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	223445	84.5	Somers' D	0.69
Discordant	39975	15.1	Goodman-Kruskal Gamma	0.70
Ties	893	0.3	Kendall's Tau-a	0.34
Total	264313	100.0		

Since the term parch is not significant ($P = .457$), we run the logistic regression procedure after dropping this term, and get:

Binary Logistic Regression for titanic3: survived versus age, sibsp, Gender, D1, D2 (D1 = 1, pclass = 1, 0 o.w., D2 = 1, pclass = 2, 0 o.w.)

Link Function: Logit

Response Information

Variable	Value	Count	
survived	1	427	(Event)
	0	619	
	Total	1046	

* NOTE * 1046 cases were used

* NOTE * 263 cases contained missing values

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-0.982783	0.219635	-4.47	0.000			
age	-0.0396824	0.0066274	-5.99	0.000	0.96	0.95	0.97
sibsp	-0.329198	0.100126	-3.29	0.001	0.72	0.59	0.88
Gender	2.58022	0.170786	15.11	0.000	13.20	9.44	18.45
D1	2.35008	0.228533	10.28	0.000	10.49	6.70	16.41
D2	0.981398	0.199093	4.93	0.000	2.67	1.81	3.94

All terms are highly significant.

Log-Likelihood = -485.336

Test that all slopes are zero: G = 443.949, DF = 5, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	558.184	520	0.120
Deviance	553.748	520	0.148
Hosmer-Lemeshow	26.229	8	0.001

Since P-values for each of the goodness of fit tests > .05, we cannot reject the logistic regression model, i.e., there is no lack of fit.

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Value	Group										Total	
	1	2	3	4	5	6	7	8	9	10		
1												
Obs	8	18	18	14	23	38	43	75	94	96	427	
Exp	6.8	10.8	13.5	17.2	26.7	42.5	58.5	73.6	85.7	91.8		
0												
Obs	99	89	86	90	81	70	61	30	10	3	619	
Exp	100.2	96.2	90.5	86.8	77.3	65.5	45.5	31.4	18.3	7.2		
Total	107	107	104	104	104	108	104	105	104	99	1046	

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	223472	84.5	Somers' D	0.69
Discordant	40013	15.1	Goodman-Kruskal Gamma	0.70
Ties	828	0.3	Kendall's Tau-a	0.34
Total	264313	100.0		

all 3 measures show good prediction from logistic regression.

Interpretation of the Fitted Logistic Regression Model

OR = P(survial)/P(not surviving)

$$\ln\left(\frac{\pi}{1-\pi}\right) = -0.98 - 0.04Age - 0.33Sibsp + 2.58Gender + 2.35D1 + .98D2$$

$$\pi = \frac{e^{-0.98 - 0.04Age - 0.33Sibsp + 2.58Gender + 2.35D1 + .98D2}}{1 + e^{-0.98 - 0.04Age - 0.33Sibsp + 2.58Gender + 2.35D1 + .98D2}}$$

where Age and Sibsp are continuous or quantitative variables, and the other three are binary variables.

Gender = 1 for Female passenger, 0 for male passenger

D1 = 1 if Passenger Class (pclass) = 1, 0 otherwise

D2 = 1 if Passenger Class (pclass) = 2, 0 otherwise

From the above fitted regression model, we can see that

Odds Ratio (ln(OR)) goes down with Age at the rate of $\exp(-.04) = 0.96$, which equals a $100(.96-1)/1 = 4\%$ decrease.

Odds Ratio (ln(OR)) goes down with Age at the rate of $\exp(-.33) = 0.72$, which equals a $100(.72-1)/1 = 28\%$ decrease.

The three remaining variables are binary.

$$\frac{\text{OR}(\text{Gender} = 1, \text{ i.e., passenger is Female})}{\text{OR}(\text{Gender} = 0, \text{ i.e., passenger is Male})} = e^{2.58} = 13.2$$

This means the female survival odds are 13.2 times that for the males.

$$\frac{\text{OR}(D1 = 1, \text{ i.e., passenger is in 1}^{\text{st}} \text{ class})}{\text{OR}(D2 = 1, \text{ i.e., passenger is in 2}^{\text{nd}} \text{ class})} = e^{2.35-.98} = e^{1.37} = 3.94$$

which implies that a 1st class passenger has survival odds 3.94 times the survival odds of a 2nd class passenger.