Handout 16

Logistic Regression.

So far we have only dealt withy the cases where the response variable is quantitative and the mathematical assumption is that the response is a continuous variable. Let us look at two data sets. The first data has a binary response in which the response takes values either 0 or 1. In the second case the response is a count. In none of theses cases, can we realistically model the response as a continuous variable. In this handout we will deal with data sets where the response is a 0-1 variable.

Example 1. (CHD percentages in the population) On the basis of last 5 years' records in a city, we have the following table which gives the percentages of people with CHD (coronary heart disease) for different age groups.

Age group	Percent with CHD
[20,30)	0.9
[30,40)	5.1
[40,50)	21.8
[50,60)	55.9
[60,70)	84.3
[70,80]	96.1

Let $\pi(X)$ be the proportion of people at age X. Here we take the midpoints of the intervals to be the representative age of the group. When $\pi(X)$ is plotted against X, it shows a sigmoidal (or S-shaped). Moreover, the plot of the logit of $\pi(X)$ [logit of p is defined to be $\log(p/(1-p))$] against X is almost linear. Thus, it may be reasonable to model logit of $\pi(X)$ by $\beta_0 + \beta_1 X$, which leads to $\pi(X)$ being approximately equal to $\exp(\beta_0 + \beta_1 X)[1 + \exp(\beta_0 + \beta_1 X)]$. This is called a "logistic linear model", i.e., the logit transform of $\pi(X)$ is linear in X. There are cases where the logit transform of $\pi(X)$ may be modeled by a quadratic form as Example 2 shows.

For this example we may fit a straight line to logit $\pi(X)$. It turns out that $\beta_0 \approx -8.467$ and $\beta_1 \approx 0.157$. At $X = 45$, using this line, we have

$$\pi(45) = \exp(\beta_0 + \beta_1(45))[1 + \exp(\beta_0 + \beta_1(45))] = 0.1957.$$

This value is reasonably close to the population value for the age group 40-50.. Now the odds of CHD at $X = 45$ is defined to be $\pi(45)/(1 - \pi(45)) = 0.2432$. In general, the odds of CHD at age X is

$$\pi(X)/(1 - \pi(X)) = \exp(\beta_0 + \beta_1 X)$$

. Now consider the ratio of odds at age $X+1$ and X; i.e.,

$$\frac{\text{odds of CHD at age } X+1}{\text{odd of CHD at age } X} = \frac{\pi(X+1)/(1-\pi(X+1))}{\pi(X)/(1-\pi(X))} = \frac{\exp(\beta_0+\beta_1(X+1))}{\exp(\beta_0+\beta_1 X)} = e^{\beta_1}.$$

Note that this ratio of odds does not depend on X, something that is not true of the logit of $\pi(X)$ is not linear in X. Here $e^{\beta_1} = e^{0.157} = 1.170$. So the odds of CHD at age $X+1$ is about 1.17 times the odds of CHD at age X.

Now if we select a random sample from the population and record, for each person, the age X and whether or not the person has CHD ($Y=1$ if the person has CHD, 0 otherwise), then we can say $\pi_i = P(Y_i = 1)$. Denoting π_i', the logit transform of π_i, we may try to model π_i' as a linear function of X_i, i.e., $\pi_i' = \beta_0 + \beta_1 X_i$ and estimate the parameters β_0 and β_1 on the basis of the data. In some cases (as in the next example), it may turn out that π_i' is not linear in X_i and in that case we may try a quadratic function of X_i. Usually, there are more background information such as gender, educational level, socio- economic back grounds. In such a case π_i' may be modeled as a linear function of other variables the same manner as in the case of linear models. Thus if we have the data $(Y_i, X_{i1}, X_{i2}, X_{i3}), i = 1, \ldots, n$, then we may model π_i' as $\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}$ and use the data to estimate the unknown beta parameters.

Example 2. In a certain state, we have the percentage (over a period of 10 years) of companies whose IPO (initial public) offerings) were financed by venture capital funds. Also known were the face values of the companies. Since face value distribution is highly skewed we have taken X to be the logarithm of the face value

X	Percent
12	1
13	8.2
14	26.5
15	56.1
16	59.5
17	43.7
18	15.3
19	1.5

For this case, note that the proportion of companies attracting venture capital inutility increases with face value then decreases. A plot of logit of the proportion against X shows that it initially decreases and then decreases. In this case, one needs to use a quadratic in X, i.e., logit of $\pi(X)$ needs to be modeled as a quadratic function of X. Also note that in this case, the a reasonable model for logit of $\pi(X)$ is $0.2983 + 0.2597(X-\bar{X}) - 0.3864(X-\bar{X})^2$, where $\bar{X} = 15.3$.

Since the logit of $\pi(X)$ is being modeled as a quadratic, it is fairly easy to see that the odds ratio at

$X+1$ to X depends on X. First a formula for the odds at X - it is equal to

$$\exp[\beta_0 + \beta_1(X - \bar{X}) + \beta_{11}(X - \bar{X})^2],$$

where the values of $\beta_0, \beta_1, \beta_{11}$ and $\bar{X}$ are given for the quadratic model above.

The odds of getting venture capital at $X = 13, 14, 17, 18$ are 0.096031253, 0.500402521, 0.686039117 and 0.162478130 respectively. Thus

$$\frac{\text{odds at } 14}{\text{odd at } 13} = \frac{0.500402521}{0.096031253} = 5.21,$$
$$\frac{\text{odds at } 18}{\text{odd at } 17} = \frac{0.162478130}{0.686039117} = 0.237.$$

Unlike in the linear case, the odds ratio is not constant at different values of X. The odds at $X = 14$ is 5.21 times larger than the odds at $X = 13$. Whereas the odds at $X = 18$ is 0.237 times the odds at $X = 17$. Here, the odds of attracting venture capital increases at $X = 13$, whereas the odds decreases at $X = 17$.

Logistic regression: heart attack data (Example 3):

Suppose we have observations in heart attack data in which, for the i^{th} subject, Y_i is 1 if there is a second heart attack within a year of the first attack and 0 otherwise. We look at an independent variable X, which is anxiety. At this moment we will ignore the other variable if the patient has gone through treatment for anger management. Thus we have observations, $(Y_i, X_i), i = 1, \ldots, n$, and if we write $\pi_i = P(Y_i = 1)$, then let π' be the vector of logit transformations of π_i, i.e., $\pi_i' = \log[\pi_i/(1 - \pi_i)]$, then we we are modeling $\pi_i' = \beta_0 + \beta_1 X_i, i = 1, ..., n$, or $\pi' = X\beta$ in the matrix notations.

We now fit a logistic model for this data using the R command "attack=glm(y~X,family=binomial)". The command "summary(attack)" will give us the following output

Call:

glm(formula = Y ~X, family = binomial)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.62461	-0.83983	-0.01232	0.64540	2.10801

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.0925	3.1709	-2.237	0.0253 *
X	0.1246	0.0553	2.254	0.0242 *

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 27.726 on 19 degrees of freedom

Residual deviance: 19.601 on 18 degrees of freedom

AIC: 23.601

Number of Fisher Scoring iterations: 4

Inference. Unlike in the linear regression or ANOVA cases, all the inference in logistic regression model are approximate since it is not possible to find the exact distribution of the parameter estimates b_i's. For instance, in order construct confidence intervals for β_1 or carrying out tests for it, we will use the normal table. Mathematically, the normal approximation is justified if the sample size n is large.

An approximate 95% confidence interval for β_1 is $b_1 \pm 1.96 s(b_1)$, i.e., $0.1246 \pm (1.96)(0.0553)$ i.e., 0.1246 ± 0.1084, i.e., $(0.016, 0.233)$. Similarly we can carry out a test for the hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, by using the z-statistic $z^* = b_1/s(b_1) = 2.254$.The p-value is 0.0242 as given above.

Estimating the probability of a second heart attack at $X = 71$. This is simply done by plugging in the value of of $X = 71$ in the fitted logistic regression. However, a package like R will also do this.

Plots: The plot of the fitted $\hat{\pi}_i$ against X_i is plotted. When it comes to residuals, there are multiple definitions of residuals for the logistic regression. Two most common ones are:

Pearson Residual: $r_{Pi} = (Y_i - \hat{\pi}_i)/\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}$

Deviance Residual: $dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2[Y_i \log \hat{\pi}_i + (1-Y_i)\log(1-\hat{\pi}_i)]}$.

[R commands: resid(attack), resid(attack,type="pear"), will yield deviance residuals and Pearson residuals, respectively].

Here are two measure that are somewhat similar to SSE in the regression models discussed in STA 206.

(a) Pearson: $X^2 = \sum r_{Pi}^2$, and (ii) Deviance: $G^2 = \sum dev_i^2$.

Degrees of freedom for each of them is n-(# of beta parameters estimated). Hence, the df=n-2=18 here. Even though, neither X^2 nor G^2 has exactly a χ^2 distribution with $n-2$ df, we may still use the χ^2 table to see if the calculated values of X^2 or G^2 are too high. High values indicate inadequacy of the model. For instance, $\chi^2(0.95; 18) = 28.87$. The value of $G^2 = 19.601$ which is clearly smaller than 28.87. Thus we may conclude that there does not seem to be a departure from the linear model fitted here.

We may plot the residuals against the fitted $\hat{\pi}_i$'s, and usually there are parallel curves. It is not very clear how much information these plots convey. To extract more information, we may plot the loess of the residuals, if the loess seem to be quite close to the horizontal line at 0, it would indicate the model may be adequate in estimating π, the probability of a second hear attack. Another plot known as the half normal plot allows us to check adequacy of the model and identify outliers.[R command for half-normal plot is "halfnorm" (package "faraway"). Half normal-plot here does not seems to show any serious outliers, nor any inadequacy of the logistic linear model.

An important technical note: Following the notations in the text, the logistic linear regression model is called the reduced model, and the full model is the one where no specific model assumption is made. If we

denote the maximum likelihoods of the reduced and the full model by $L(R)$ and $L(F)$ respectively, then (see below)

$$G^2 = -2[\log L(R) - \log L(F)] = -2\log L(R).$$

Analysis of the data in Example 4.

Note that for this data, we have the number exposed and killed at $c = 8$ different doses. For instance, $n_2 = 60$ were exposed to dose $X_2 = 53$ out of which 13 were killed. This is an example of Binomial regression which is not really much different from the previous example. Here, the total number of observations is $n = 420$. One may think at dose $X_2 = 53$, we have $n_2 = 60$ observations $Y_{i2}, i = 1, \ldots, n_2 = 60$. Each Y_{ij} is 0-1 valued, i.e., Y_{ij} is Bernoulli with probability of getting killed is π_j. So $Y_{\cdot 2} = \sum_i Y_{i2}$ has a Binomial(n_2, π_2) distribution. In general we have c distinct doses $X_1, ..., X_c$; n_j beetles were exposed to dose X_j and $Y_{\cdot j}$ were killed out of n_j flies. So $Y_{\cdot j}$ has a Binomial(n_j, π_j) distribution. We now want to model π'_j, the logit of π_j, as a linear function of X_j. Thus we mode $\pi'_j = \beta_0 + \beta_1 X_j, j = 1, \ldots, c = 8$. If this is not adequate we may also try a polynomial model such as quadratic.

We will describe the linear modeling here. The R command is "glm(c(killed,not killed)~X,family=binomial)". Here is a summary of the output.

Call:

glm(formula = cbind(kill, surv) ~dose, family = binomial)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2746	-0.4668	0.7688	0.9544	1.2990

Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-14.82300	1.28959	-11.49	<2e-16
dose	0.24942	0.02139	11.66	<2e-16

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.2024 on 7 degrees of freedom

Residual deviance: 7.3849 on 6 degrees of freedom

AIC: 37.583

Number of Fisher Scoring iterations: 4

This output clearly shows that the dose is a significant variable, i.e., if we test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, then the z-statistic is $z = b_1/s(b_1) = 11.66$ and the p-value$\approx = 0$. Thus we can reject H_0 quite safely.

Goodness-of-fit tests.

In Example 4, we may wish to investigate if fitting a logistic linear model is appropriate, i.e., test $H_0 : \pi'_j = \beta_0 + \beta_1 X_j$ vs. $H_1 : \{\pi'_j\}$ do not line on straight line.

There are two popular tests for this:

(a) Pearson Chi-square test,

(b) Likelihood ratio test (Deviance goodness-of-fit test).[Note at the end of this handout.]

(b) Let us discuss (b) first. This test statistic G^2 is given by R. Here $G^2 = 7.3849$ and the df$= c-p = 8-2 = 6$, where p is the number of beta parameters estimated. So we reject H_0 if $G^2 > \chi^2(0.95; 6) = 12.59$. Clearly, $G^2 < 12.59$ and hence we cannot reject H_0. Conclusion: a linear logistic fit seems to be reasonable.

What is the structure of this test? The full model is the saturated model, whereby the $\{\pi_j\}$ are allowed to be arbitrary, and the reduced model is $\pi'_j = \beta_0 + \beta_1 X_j$. So we need to obtain the likelihood $L(F)$ and $L(R)$ under the full and the reduced models. The test statistic is

$$\begin{aligned} G^2 &= -2[\log(L(R)) - \log(L(F))] \\ &= -2\sum_{j=1}^{c}\left[Y_{\cdot j}\log\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - Y_{\cdot j})\log\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right], \end{aligned}$$

where $\hat{\pi}_j$ is the estimated value of π_j under the reduced model and $p_j = Y_{\cdot j}/n_j$. Note that p_j is the MLE for π_j under the full (here saturated) model. Under H_0, $G^2 \sim \chi^2_{c-p}$.

(a) Pearson Chi-square goodness-of-fit test.

Let $\hat{\pi}_j$ be the estimated value of π_j under the reduced model. Then the expected number of killed at X_j is $E_{j1} = n_j\hat{\pi}_j$ and expected number of not-killed is $E_{j0} = n_j(1-\hat{\pi}_j)$. Denote O_{j1} is the observed number of killed (which is $Y_{\cdot j}$ in our notation) and the observed number of not-killed is $O_{j0} = n_j - Y_{\cdot j}$. The Pearson statistics is

$$X^2 = \sum_{j=1}^{c}\sum_{l=0}^{1}\frac{(O_{jl} - E_{jl})^2}{E_{jl}}.$$

Under H_0, $X^2 \sim \chi^2_{c-p}$, under the assumption that all the expected frequencies $\{E_{jl}\}$ are large. A rule of thumb for this "largeness" is to have all (or almost all) the expected frequencies 5 or larger. Otherwise, frequencies for neighboring X_j are merged together. In our example, we have the following table

Dose	Exposed	Killed		Not-killed	
X_j	n_j	Observed O_{j1}	Expected E_{j1}	Observed O_{j0}	Expected E_{j0}
49.1	59	6	4.171	53	54.829
53.0	60	13	10.044	47	49.956
56.9	62	18	21.526	44	40.474
60.8	56	28	32.732	28	23.268
64.8	63	52	49.915	11	13.085
68.7	59	53	53.684	6	5.316
72.6	62	61	59.762	1	2.238
76.5	60	60	59.160	0	0.840

Note that the expected number of not-killed are rather small for dose X_7 and X_8. It may be worthwhile to merge the doses X_6, X_7 and X_8 together. Here is the table after merging these.

Dose	Exposed	Killed		Not-killed	
X_j	n_j	Observed O_{j1}	Expected E_{j1}	Observed O_{j0}	Expected E_{j0}
49.1	59	6	4.171	53	54.829
53.0	60	13	10.044	47	49.956
56.9	62	18	21.526	44	40.474
60.8	56	28	32.732	28	23.268
64.8	63	52	49.915	11	13.085
$\geq$68.7	181	174	172.606	7	8.438

Now note that we have $c-2=6$ dose categories, so the Pearson statistic X^2 will have $df = c-2-p = 2$. The value of the statistics is $X^2 = 5.101$. Since $\chi^2(0.92; 2) = 5.991$, we cannot reject H_0. Thus a linear logistic model is reasonable here.

Remark: We should note in passing the requirement that all the expected frequencies (i.e., $E_{j1} = n_j\hat{\pi}_j$ and $E_{j0} = n_j(1-\hat{\pi}_j)$) is also valid for the Deviance goodness-of-fit test. Even though, we have not done it here, same merging of categories are also needed for this example.

Multiple Logistic regression.

Let us look at the "Arthritis" data set. Arthritis patients were asked if the conditions were improved (none, some, marked). Some of the patients were given the treatment, others none (placebo). Also recorded were gender and age of patients. Suppose we merge the groups: some and marked improvement. So Y will take values 0 (no improvement) and 1 (improvement). The other variables are treatment (X_1, categorical: 0-1 valued), gender (X_2, categorical: 0-1 valued), age (X_3, quantitative). So we the data $(Y_i, X_{i1}, X_{i2}, X_{i3})$. Let π_i be the probability that $Y_i = 1$. Thus a multiple logistic model can be written as $\pi_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$, where π_i' is the logit transform of π_i. In vector-matrix notations we have $\pi' = X\beta$, where X is $n \times 4$. If we decide the add the interactions: gender*treatment, gender*age and

treatment*age, then we will have three more predictor variables and in matrix notations we have $\pi' = X\beta$ with X as an $n \times 7$ matrix.

The methods for multiple regression have their analogues for linear logistic regression such as simultaneous inference, estimation of mean response (i.e., estimating π at given X values), nonlinearity, variable deletion, stepwise regression etc.

Estimation of parameters (Maximum likelihood).

For logistic regression, one may use the maximum likelihood (ML) method. The likelihood is

$$L = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}, \text{ and}$$
$$\log L = \log L = \sum [Y_i \log \pi_i + (1-Y_i)\log(1-\pi_i)].$$

If π_i involves the parameters β_0 and β_1, i.e. $\pi_i = \exp(\beta_0 + \beta_1 X_i)/[1+\exp(\beta_0+\beta_1 X_i)]$. Then

$$\begin{aligned}
\log L(\beta_0, \beta_1) &= \sum_{i=1}^{n} [Y_i \log \pi_i + (1-Y_i)\log(1-\pi_i)] \\
&= \sum_{i=1}^{n} [Y_i \log\{\pi_i/(1-\pi_i)\} + \log(1-\pi_i)] \\
&= \sum_{i=1}^{n} [Y_i(\beta_0+\beta_1 X_i) + \log(1-\pi_i)\} \\
&= \sum_{i=1}^{n} [Y_i(\beta_0+\beta_1 X_i) - \sum_{i=1}^{n} \log[1+\exp(\beta_0+\beta_1 X_i)].
\end{aligned}$$

In order to maximize $\log L(\beta_0, \beta_1)$ one differentiates it with respect to β_0 and β_1 and equates the derivatives to zero and that leads to the equations

$$\sum \pi_i = \sum Y_i, \quad \sum \pi_i X_i = \sum \pi_i Y_i.$$

A solution of these equations lead to estimates b_0 and b_1 of β_0 and β_1. No explicit forms of b_0 and b_1 are available. They are usually solved via iterative methods. Thus the estimated probabilities are $\hat{\pi}_i = \exp(b_0 + b_1 X_i)/[1+\exp(b_0+b_1 X_i)]$.

If we write the model $\pi' = X\beta$, where X is $n \times 2$, then the likelihood equations can be written as

$$X^T(Y - \hat{\pi}) = 0.$$

Incidentally, this form for the likelihood equation holds even when X is a $n \times p$ matrix (i.e., there are multiple predictors). How good are the estimators? Here the results are only approximate unlike in the usual linear regression case. It can be shown that

$$E(b) \approx \beta, \ \sigma^2(b) = Cov(b) \approx (X^T W X)^{-1},$$

where W is a $n \times n$ diagonal matrix whose diagonal entries are $\pi_1(1-\pi_1), \ldots, \pi_n(1-\pi_n)$. Since W can be estimated by a diagonal matrix $\hat{W}$ whose diagonal entries are $\hat{\pi}_1(1-\hat{\pi}_1), \ldots, \hat{\pi}_n(1-\hat{\pi}_n)$, we can therefore estimate $Cov(b)$ by

$$s^2(b) = (X^T \hat{W} X)^{-1}.$$

Unlike in the linear regression case, we do not know the exact distribution of b_j. It turns however out that for n large, the distribution of $(b_j - \beta_j)/s(b_j)$ is approximate $N(0,1)$. Thus an approximate $(1-\alpha)100$ confidence interval for β_j is $b_j \pm z(1-\alpha/2)s(b_j)$, where $z(1-\alpha/2)$ is obtained from the normal table and $s^2(b_j)$ is the j^{th} diagonal entry of $s^2(b)$. We can similarly carry out a hypothesis test for β_j.

Deviance G^2: We know that the log-likelihood is

$$\log L = \sum [Y_i \log \pi_i + (1-Y_i)\log(1-\pi_i)].$$

Suppose we have two models: a reduced model and a full model. Let $\hat{\pi}_i^{(R)}$ and $\hat{\pi}_i^{(F)}$ be the estimate of π_i under the reduced and the full models respectively. The deviance is defined to be

$$\begin{aligned} G^2 &= -2[\log L(R) - \log L(F)] \\ &= -2 \sum [Y_i \log\{\hat{\pi}_i^{(R)}/\hat{\pi}_i^{(F)}\} + (1-Y_i)\log\{(1-\hat{\pi}_i^{(R)})/(1-\hat{\pi}_i^{(F)})\}]. \end{aligned}$$

Special case: If the full model is the saturated case in which no assumption is being made on π_i, i.e., they are allowed to be arbitrary. Then for the full model, $\hat{\pi}_i = Y_i$ and

$$\begin{aligned} \log L(F) &= \sum [Y_i \log \hat{\pi}_i + (1-Y_i)\log(1-\hat{\pi}_i)] \\ &= \sum [Y_i \log Y_i + (1-Y_i)\log(1-Y_i)] = 0. \end{aligned}$$

For this case, then

$$\begin{aligned} G^2 &= -2[\log L(R) - \log L(F)] = -2 \log L(R) \\ &= -2 \sum [Y_i \log\{\hat{\pi}_i^{(R)}\} + (1-Y_i)\log\{(1-\hat{\pi}_i^{(R)})\}. \end{aligned}$$

Example 3. Suppose that we are working with some doctors on heart attack patients. The dependent variable is whether the patient has had a second heart attack within 1 year (yes = 1). We have two independent variables, one is whether the patient completed a treatment consistent of anger control practices (yes=1). The other IV is a score on a trait anxiety scale (a higher score means more anxious).

Person	2^{nd} heart attack	Treatment of anger	Trait Anxiety
1	1	1	70
2	1	1	80
3	1	1	50
4	1	0	60
5	1	0	40
6	1	0	65
7	1	0	75
8	1	0	80
9	1	0	70
10	1	0	60
11	0	1	65
12	0	1	50
13	0	1	45
14	0	1	35
15	0	1	40
16	0	1	50
17	0	0	55
18	0	0	45
19	0	0	50
20	0	0	60

Example 4.

Beetles were exposed to gaseous carbon disulphide at various concentrations (in mf/L) for five hours (Bliss, 1935) and the number of beetles killed were noted. The data are in the following table:

Dose	Exposed	Killed	Prop(obs)	Prop(fit)	Expected(killed)
49.1	59	6	0.102	0.0707	4.171
53.0	60	13	0.217	0.1674	10.044
56.9	62	18	0.290	0.3472	21.526
60.8	56	28	0.500	0.5845	32.732
64.8	63	52	0.825	0.7923	49.915
68.7	59	53	0.898	0.9099	53.684
72.6	62	61	0.984	0.9639	59.762
76.5	60	60	1.000	0.9860	59.160

The fitted proportion and the expected number of killed are according to the fitted linear logistic model.

Other important models for modeling binary response

There are two other popular models for modeling binary response, i.e., when Y is 0-1 valued. We will describe then for the case when there is a single predictor, but the same ideas continue to be true when there are multiple predictors.

(a) Probit regression: $\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 X_i$, where Φ is the cdf of the standard normal distribution

(b) Complimentary log-log: $\log(-\log \pi_i) = \beta_0 + \beta_1 X_i$.

All packages including R have the option of data analysis using probit and complimentary log-log models.

In actual data analysis, logistic and probit models tend yield similar results.

Motivation behind the different models for binary response.

Let us think of the CHD example where Y is 0-1 valued with $Y = 1$ denotes the presence of CHD. Suppose that there is some (unknown) quantitative measure Y^c of chemical balance in the body and when it is below a threshold y^* it leads to CHD. Thus $Y = 1$ is the same as the event $Y^c \leq y^*$. Let us also assume that there is a linear relation between Y^c and X, age, i.e.., $Y^c = \beta_0^c + \beta_1^c X + \varepsilon$. Then

$$\begin{aligned}\pi = P(Y = 1) &= P(Y^c \leq y^*) = P(\beta_0^c + \beta_1^c X + \varepsilon \leq y^*) \\ &= P(\varepsilon \leq y^* - \beta_0^c - \beta_1^c X).\end{aligned}$$

It turns out that different models such as logistic, probit, complimentary log-log etc. consequences on what type of distributional assumptions are made of the distribution of ε. Please note that $E(\varepsilon) = 0$.

(a) Logistic distribution of ε: A random variable Z is said to have a standard logistic distribution if its pdf f and cdf F are

$$f(z) = e^z/(1 + e^z)^2,\ F(z) = e^z/(1 + e^z), -\infty < z < \infty.$$

This pdf is symmetric about zero and is bell shaped. If we assume that $\varepsilon = \sigma Z$, where Z has a standard logistic distribution and $\sigma > 0$ is a scale parameter, then

$$\begin{aligned}\pi &= P(\varepsilon \leq y^* - \beta_0^c - \beta_1^c X) = P(Z \leq (y^* - \beta_0^c - \beta_1^c X)/\sigma) \\ &= P(Z \leq \beta_0 + \beta_1 X)\,, \text{ with } \beta_0 = (y^* - \beta_0^c)/\sigma, \beta_1 = -\beta_1^c/\sigma \\ &= e^{\beta_0 + \beta_1 X}/(1 + e^{\beta_0 + \beta_1 X}).\end{aligned}$$

Clearly this leads to the logistic model since

$$\pi' = \log(\pi/(1 - \pi)) = \beta_0 + \beta_1 X.$$

(b) Probit model: If we assume that $\varepsilon = \sigma Z$, where $Z \sim N(0, 1)$, we have

$$\begin{aligned}\pi &= P(\varepsilon \leq y^* - \beta_0^c - \beta_1^c X) = P(Z \leq (y^* - \beta_0^c - \beta_1^c X)/\sigma) \\ &= P(Z \leq \beta_0 + \beta_1 X)\,, \text{ with } \beta_0 = (y^* - \beta_0^c)/\sigma, \beta_1 = -\beta_1^c/\sigma \\ &= \Phi(\beta_0 + \beta_1 X),\end{aligned}$$

where Φ is the cdf of the standard normal distribution. Thus we are lead to the probit model since

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 X.$$

(c) Complimentary log-log: A random variable Z has a standard Gumbel distribution if its pdf f and cdf F are

$$f(z) = \exp(-z - e^z),\ F(z) = \exp(-e^z),\ \ -\infty < z < \infty.$$

Gumbel distribution is not symmetric about 0, it is skewed to the right. If we assume that $\varepsilon = \sigma Z$, where $\sigma > 0$ is a scale parameter and Z has a standard Gumbel distribution, then

$$\begin{aligned}\pi &= P(\varepsilon \le y^* - \beta_0^c - \beta_1^c X) = P(Z \le (y^* - \beta_0^c - \beta_1^c X)/\sigma) \\ &= P(Z \le \beta_0 + \beta_1 X)\,, \text{ with } \beta_0 = (y^* - \beta_0^c)/\sigma, \beta_1 = -\beta_1^c/\sigma \\ &= \exp(-e^{\beta_0 + \beta_1 X}).\end{aligned}$$

This leads to the complimentary log-log model since

$$\log - \log \pi) = \beta_0 + \beta_1 X.$$

Remark: It is interesting to note that we get the three models (logistic, probit and complimentary log-log) by considering three distribution of ε, namely normal, logistic and Gumbel. However, it is clearly possible to get other models by considering other distributional forms for ε.

Deviance Goodness-of-fit test

Note that $Y_{\cdot j} \sim$Binomial(n_j, π_j) and $Y_{\cdot j}$'s are independent. Thus the likelihood is

$$\prod_{j=1}^{c} \binom{n_j}{Y_{\cdot j}} \pi_j^{Y_{\cdot j}} (1-\pi_j)^{n_j - Y_{\cdot j}}.$$

Under the saturated model, i.e., when π_j 's are not restricted, then the MLE for π_j is $p_j = Y_{\cdot j}/n_j$. Under the reduced model (i.e., under logistic linear model), estimate of π_j is $\hat{\pi}_j = \exp(b_0 + b_1 X_j)/[1 + \exp(b_0 + b_1 X_j)]$. Thus we have

$$L(F) = \prod_{j=1}^{c} \binom{n_j}{Y_{\cdot j}} p_j^{Y_{\cdot j}} (1-p_j)^{n_j - Y_{\cdot j}},\ L(R) = \prod_{j=1}^{c} \binom{n_j}{Y_{\cdot j}} \hat{\pi}_j^{Y_{\cdot j}} (1-\hat{\pi}_j)^{n_j - Y_{\cdot j}}.$$

Thus

$$\begin{aligned}G^2 &= -2[\log(L(R)) - \log(L(F))] \\ &= -2\sum_{j=1}^{c}[\{Y_{\cdot j}\log\hat{\pi}_j + (n_j - Y_{\cdot j})\log(1-\hat{\pi}_j)\} - \{Y_{\cdot j}\log p_j + (n_j - Y_{\cdot j})\log(1-p_j)\}] \\ &= -2\sum_{j=1}^{c}\left[Y_{\cdot j}\log\left(\frac{\hat{\pi}_j}{p_j}\right) + (n_j - Y_{\cdot j})\log\left(\frac{1-\hat{\pi}_j}{1-p_j}\right)\right].\end{aligned}$$

Figure 1: Logistic model: CHD and IPO

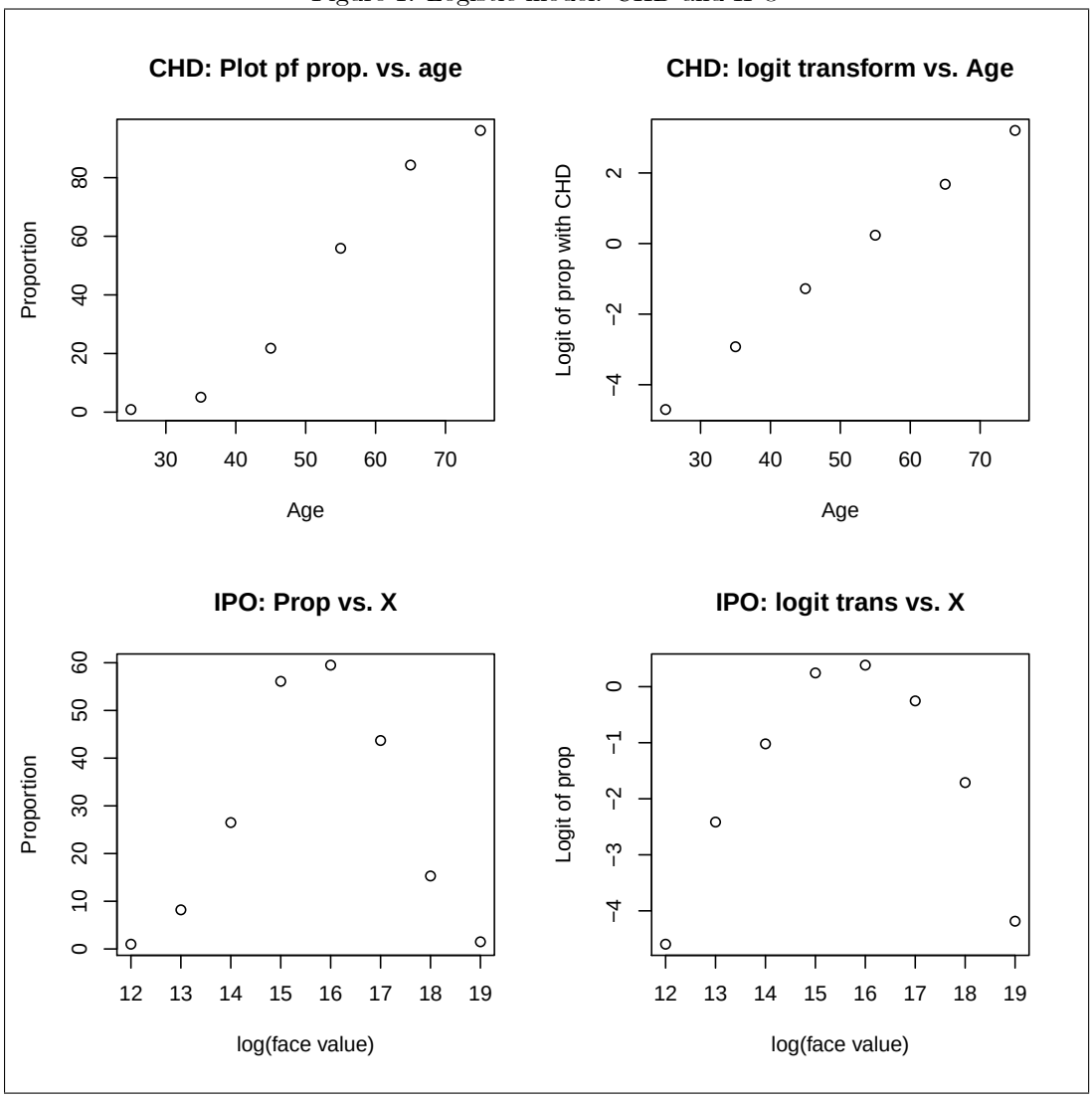

Figure 2: Heart-attack data: heart attack vs. anxiety