Handout 9

Logistic Regression.

We present a few examples 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.,

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

Note that this ratio of odds does not depend on X, something that is not true if 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. (Population rates of attracting IPO) 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 initially 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 $X = 17$.

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

For the i^{th} subject, let Y_i=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.

The fitted model is (using the R command glm(Y ~X, family = binomial))

$$\begin{aligned} \hat{\pi}'(X) &= -7.0925 + 0.1246X, \\ s(\hat{\beta}_0) &= 3.1709,\ s(\hat{\beta}_1) = 0.0553. \end{aligned}$$

If we test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, then the z-statistic is

$$z = \hat{\beta}_1 / s(\hat{\beta}_1) = 2.254,$$

and the p-value is 0.024. If $\alpha = 0.05$, then we can reject H_0 at level $\alpha = 0.05$.

We can also construct a 95% confidence interval for β_1 as $\hat{\beta}_1 \pm 1.96 s(\hat{\beta}_1)$.

Now let us consider a model of the form

$$\pi' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

where the X_1 is treatment, X_2 is anxiety, and $X_1 X_2$ is the interaction term. Note that when

$$\begin{aligned} X_1 &= 0,\ \pi' = \beta_0 + \beta_2 X_2, \\ X_1 &= 1,\ \pi' = (\beta_0 + \beta_2) + (\beta_2 + \beta_3) X_2. \end{aligned}$$

Thus π' is a linear function of X_2in each case ($X_1 = 0$ and $X_1 = 0$), but with possibly different slopes and different intercepts. This is very much like in the linear regression case, where one independent variable is binary and the other is quantitative.

The fitted logistic regression model is given below [R command: ft=glm($Y\tilde{} X_1 + X_2 + X_1 * X_2, family =' binomial')$)]

$$\hat{\pi}' = -5.7504 - 2.2703 X_1 + 0.1083 X_2 + 0.02181 X_1 X_2.$$

Note that

$$\begin{aligned} X_1 &= 0,\ \hat{\pi}' = -5.7504 + 0.1083 X_2, \\ X_1 &= 1,\ \hat{\pi}' = -8.0207 + 0.1301 X_2. \end{aligned}$$

When $X_1 = 0$ (no anger treatment), the estimated odds of second heart attack at $X_2 = 50$ within a year is

$$\exp(-5.7504 + 0.1083(50)) = 0.7151.$$

When $X_1 = 1$, the estimated odds of second heart attack at $X_2 = 50$ within a year is

$$\exp(-8.0207 + 0.1301(50)) = 0.2197.$$

Clearly the two odds are different.

A more detailed R output shows the following

Coefficients	Estimate	Std.Error (SE)	z value	Pr(>|z|)
Intercept	-5.75043	4.35654	-1.320	0.187
X_1	-2.27025	6.43775	-0.353	0.724
X_2	0.10828	0.07534	1.437	0.151
$X_1 : X_2$	0.02181	0.1106	0.197	0.844

None of the terms in the model are significant, and the best candidate for deletion is the interaction term. One can carry out a backward stepwise method (R Command: step(ft)], and the resulting model is

$$\begin{aligned} \pi' &= -7.0925 + 0.1246X_2, \text{ with} \\ s(\hat{\beta}_0) &= 3.1709,\ s(\hat{\beta}_1) = 0.0553. \end{aligned}$$

Example 4 [Grouped Data].

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. We also write down the observed proportion of killed and the estimated expected proportion of killed under a logistic linear model. The data are in the following table:

Dose	Exposed	Killed	Not-killed	Prop(obs)	Prop(fit)	Expected(killed)
49.1	59	6	53	0.102	0.0707	4.171
53.0	60	13	47	0.217	0.1674	10.044
56.9	62	18	44	0.290	0.3472	21.526
60.8	56	28	28	0.500	0.5845	32.732
64.8	63	52	11	0.825	0.7923	49.915
68.7	59	53	6	0.898	0.9099	53.684
72.6	62	61	1	0.984	0.9639	59.762
76.5	60	60	0	1.000	0.9860	59.160

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 $n_2 = 60$ independent observations $Y_{2j}, j = 1, \ldots, n_2 = 60$. Each Y_{2j} is 0-1 valued, i.e., Y_{2j} is Bernoulli with probability of getting killed is π_2. So $Y_{2\cdot} = \sum_{j=1}^{n_2} Y_{2j}$ has a binomial(n_2, π_2) distribution. In general we have c distinct doses $X_1, ..., X_c$; n_i beetles were exposed to dose X_i and $Y_{i\cdot}$ were killed out of n_i flies. So $Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$ has a binomial(n_i, π_i) distribution. We now want to model π_i', the logit of π_i, as a linear function of X_i. Thus we mode $\pi_i' = \beta_0 + \beta_1 X_i, i = 1, \ldots, c = 8$. 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(cbind(killed,not killed)~X,family=binomial)". The fitted model is

$$\begin{aligned}\hat{\pi}' &= -14.82300 + 0.24942X,\\ s(\hat{\beta}_0) &= 1.28959,\ s(\hat{\beta}_1) = 0.02139,\\ z &= \hat{\beta}_1/s(\hat{\beta}_1) = 11.66.\end{aligned}$$

If we test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, then the z-statistic is $z = \hat{\beta}_1/s(\hat{\beta}_1) = 11.66$ and the p-value≈ 0. Thus we can reject H_0 quite safely.

Poisson Regression

Example 5 (Ampule breakage data)

A pharmaceutical company ships a certain medicine to wholesalers in cartons of 100 ampules. Some ampules are broken during shipping. The following data are obtained from a random sample 20 cartons.

X = number of transfers during shipment

Y = number of broken ampules in a carton 1000.

Y	11	9	17	55	12	22	13	40	8	25	8	47	19	16	34	64	19	11	28	44
X	1	0	2	5	0	3	1	4	0	3	0	5	2	1	14	5	2	0	3	4
$\hat{\mu}$	13.15	9.20	18.79	54.85	9.20	26.85	13.15	38.38	9.20	26.85	9.20	54.85	18.79	13.15	38.38	54.85	18.79	9.20	26.85	38.38

Note that one observation corresponding to $X = 14$ is an outlier and it has been deleted for further analysis.

Here we are assuming that a Poisson distribution is appropriate for Y each level of X, and the mean is modeled as $\log(\mu) = \beta_0 + \beta_1 X$. Using R [command: glm(Y~X,family='poisson')], we get the following

$$\begin{aligned}\log(\hat{\mu}) &= 2.21895 + 0.35711X, \text{ with}\\ s(\hat{\beta}_0) &= 0.1015,\ s(\hat{\beta}_1) = 0.0273.\end{aligned}$$

An approximate 95% confidence interval for β_1 is

$$\begin{aligned}&\hat{\beta}_1 \pm 1.96 s(\hat{\beta}_1),\ \text{ie, } 0.3571 \pm (1.96)(0.0273),\ \text{ie,}\\ &0.3571 \pm 0.0535,\ \text{ie, } (0.3037, 0.4106).\end{aligned}$$

Clearly, this confidence interval does not include zero. We can easily reject the null hypothesis $H_0 : \beta_1 = 0$ at level $\alpha = 0.05$ (the alternative is $H_1 : \beta_1 \neq 0$).

Example 6. **(Non-melanoma skin cancer among women)** [**Modeling rate**].

The following are the number of cases of nonmelanoma skin cancer among women reported in 1974 in two cities Minneapolis-St.Paul(Minn) and Dallas-Ft. Worth(Dallas). Also reported are the age groups and the number of individuals in each group. For the purpose of analysis we will take the midpoint of each age group as the representative age of that group and we will take 90 to be the representative age of the 85+ group (not the best thing to do and there are other methods for handling this).

Age	City	Pop	Cases
$[15, 25)$	Minn	172,675	1
$[25, 35)$	Minn	123,065	16
$[35, 45)$	Minn	96,216	30
$[45, 55)$	Minn	92,051	71
$[55, 65)$	Minn	72,159	102
$[65, 75)$	Minn	54,722	130
$[75, 85)$	Minn	32,185	133
85+	Minn	8,328	40
$[15, 25)$	Dallas	181,343	4
$[25, 35)$	Dallas	146,207	38
$[35, 45)$	Dallas	121,374	119
$[45, 55)$	Dallas	111,353	221
$[55, 65)$	Dallas	83,004	259
$[65, 75)$	Dallas	55,932	310
$[75, 85)$	Dallas	29,007	226
85+	Dallas	7,538	65

Note that the number of cases should be proportion to the population size in the group. If μ_i is the expected number of cases in the i^{th} cell, then we should be modeling the rate, i.e., the number of cases per 1000 women. Let w_i be the population (in thousands) for the i^{th} cell. Thus $w_1 = 172.675, w_2 = 123.065$ etc. Note that w_i is called the "offset". We may model $\log(\mu_i/w_i)$ as a linear function of city and age. Let X_{i1} be the age and X_{i2} be the city (Dallas=0 and Minnesota=1). So our model is of the form

$$\log(\mu_i/w_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}, \text{ i.e., } \log(\mu_i) = \log(w_i) + \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}.$$

We ran a Poisson regression with offset w_i=(pop. size)/1000. Here is the output.

```
Call:
glm(formula = y ~age + age2 + city + age * city + offset(log(w)),
family = poisson)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	$-5.478e+00$	$2.982e-01$	-18.373	$< 2e-16***$
age	$1.678e-01$	$9.940e-03$	16.883	$< 2e-16***$
age2	$-9.262e-04$	$8.165e-05$	-11.343	$< 2e-16***$
city	$-1.351e+00$	$2.380e-01$	-5.677	$1.37e-08***$
age:city	$8.395e-03$	$3.551e-03$	2.364	$0.0181*$

Note that the all the parameters are highly significant. In this case the fitted model not being adequate. Clearly, better modeling may be possible in this case. As a matter of fact, plot of $\log(y_i/w_i)$ against X_{i1} will show two parallel nonlinear curves each of which is may be cubic.

Generalized linear model.

Let us first look at the Logistic and Poisson regression cases. In the logistic case Y is Binomial$(1, \pi)$, and thus $\mu = E(Y) = \pi$ and logit of π is being modeled as a linear function of the independent variables. In the Poisson regression case Y is Poisson(μ) and $\log(\mu)$ is being modeled as linear function of the independent variables. The logit function in the Binomial case and logarithm in the Poisson case are called the link functions. Thus in both cases, we have

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where $X_1, \ldots, X_p$ are independent or explanatory variables. There is a wide class of such cases and they can be described nicely in the framework of exponential family of distributions (even though more general formulations is possible).We have

$$\begin{array}{rcl} \text{logistic} & : & g(\mu) = \log\left(\dfrac{\mu}{1-\mu)}\right), 0 < \mu < 1, \\ \text{Poisson} & : & g(\mu) = \log(\mu),\ \mu > 0, \\ \text{Negative binomial} & : & g(\mu) = \log(\mu), \mu > 0, \\ \text{Normal} & : & g(\mu) = \mu. \end{array}$$

Clearly, there are other link functions and packages allow user defined specifications.

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

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

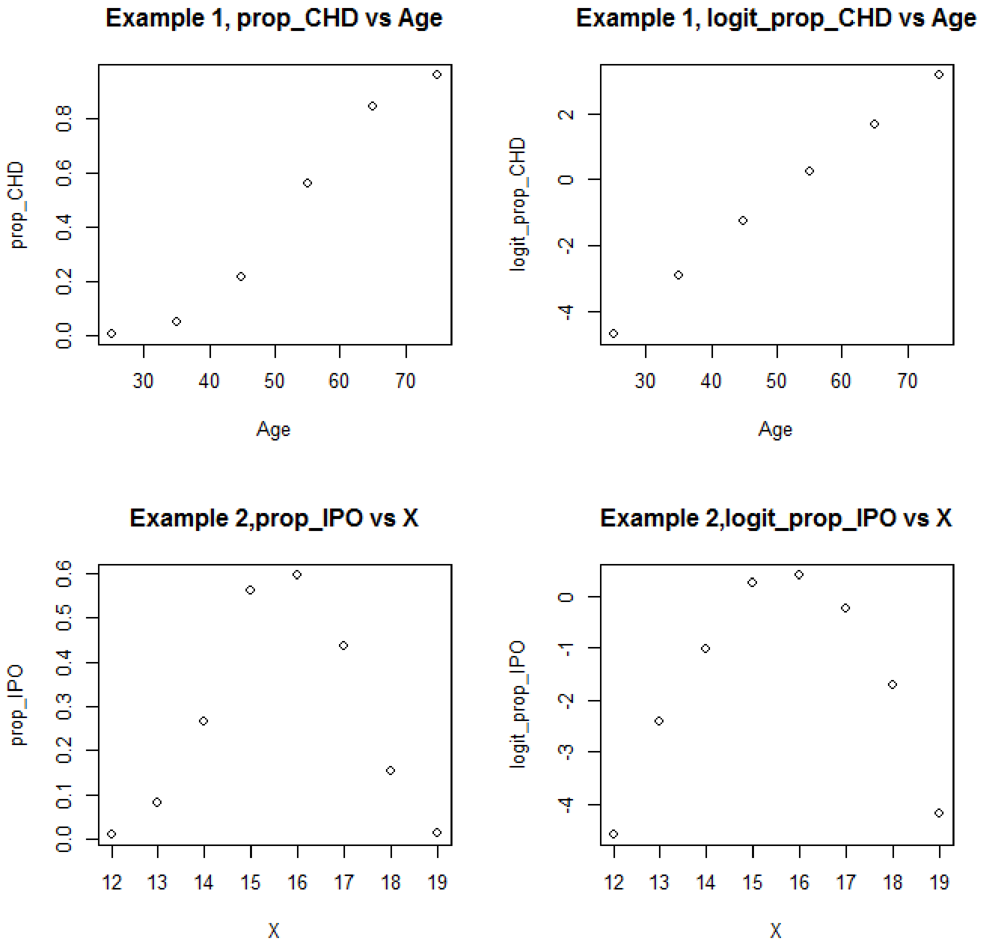
Gumbel distribution is not symmetric about 0, it is skewed to the right. If we assume that $\varepsilon = \sigma Z$, where $\sigma > 0$ is a scale parameter and Z has a standard Gumbel distribution, then

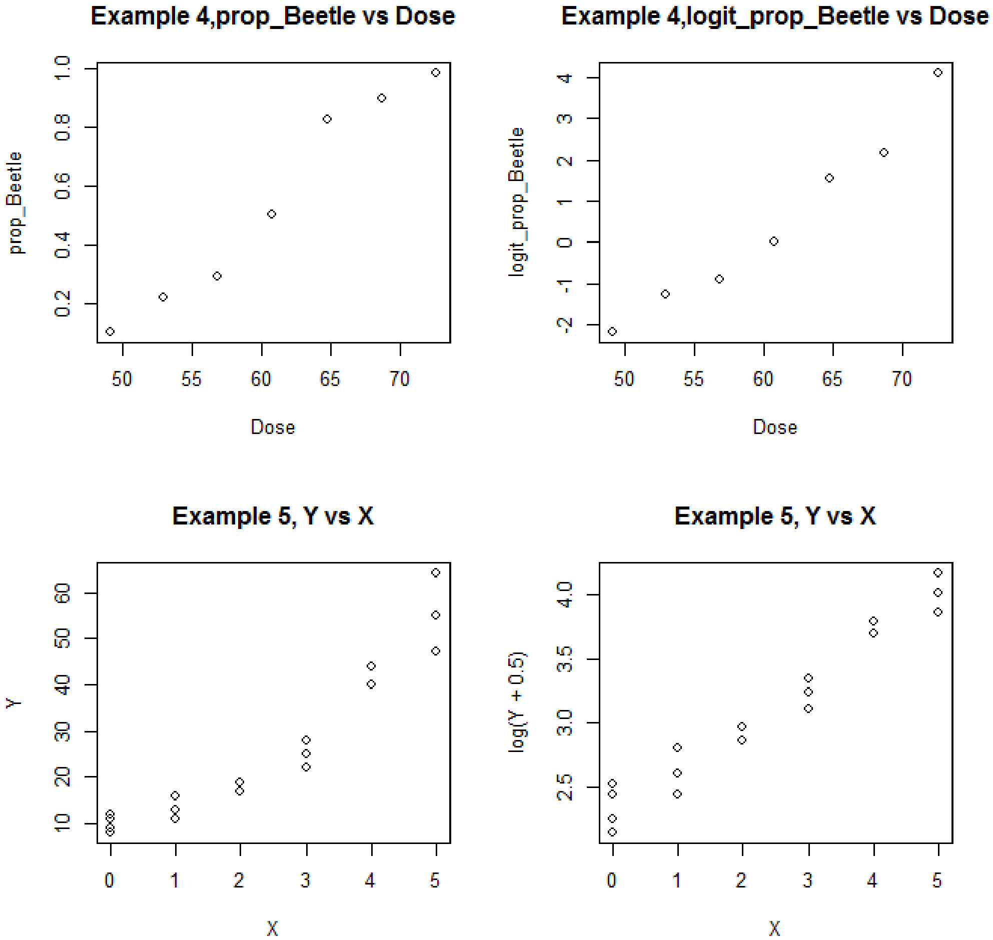
$$\begin{aligned} \pi &= P(\varepsilon \leq y^* - \beta_0^c - \beta_1^c X) = P(Z \leq (y^* - \beta_0^c - \beta_1^c X)/\sigma) \\ &= P(Z \leq \beta_0 + \beta_1 X)\,, \text{ with } \beta_0 = (y^* - \beta_0^c)/\sigma, \beta_1 = -\beta_1^c/\sigma \\ &= \exp(-e^{\beta_0 + \beta_1 X}). \end{aligned}$$

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.


Example 1, prop_CHD vs Age
prop_CHD
0.0
0.2
0.4
0.6
0.8
30
40
50
60
70
Age
Example 1, logit_prop_CHD vs Age
logit_prop_CHD
-4
-2
0
2
30
40
50
60
70
Age
Example 2,prop_IPO vs X
prop_IPO
0.0
0.1
0.2
0.3
0.4
0.5
0.6
12
13
14
15
16
17
18
19
X
Example 2,logit_prop_IPO vs X
logit_prop_IPO
-4
-3
-2
-1
0
12
13
14
15
16
17
18
19
X

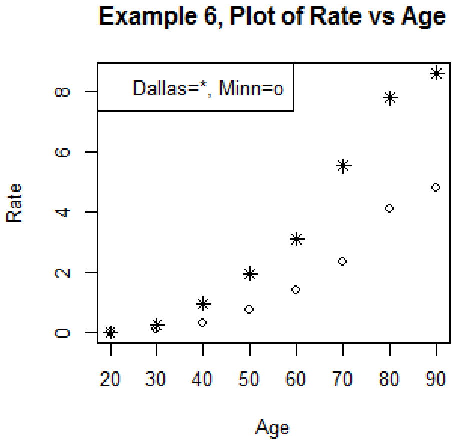


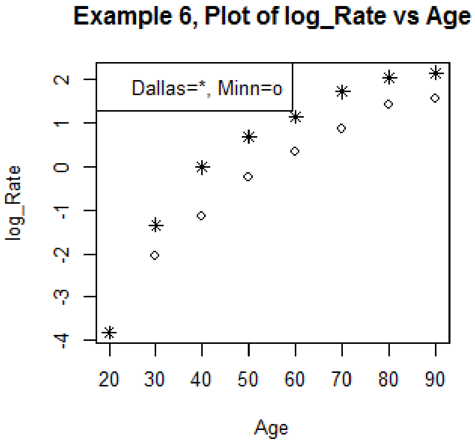
Example 4,prop_Beetle vs Dose

Example 4,logit_prop_Beetle vs Dose

Example 5, Y vs X

Example 5, Y vs X


Example 6, Plot of Rate vs Age
Dallas=*, Minn=o
Rate
0
2
4
6
8
20
30
40
50
60
70
80
90
Age


Example 6, Plot of log_Rate vs Age
Dallas=*, Minn=o
log_Rate
-4
-3
-2
-1
0
1
2
20
30
40
50
60
70
80
90
Age