LOGISTIC REGRESSION

"**Regression**" techniques provide **statistical analysis of relationships**. Research designs may be classified as experimental or observational; regression analyses are applicable to both types.

One variable is taken to be the response or dependent variable; a variable to be predicted from or **explained by other variables called predictors**, or explanatory variables, or independent variables, or covariates.

The use of the term “covariate” is not universal. Statisticians may refer to all factors/predictors as covariates; but for investigators/scientists, the term covariates are only used for factors not under investigation – i.e. confounders.

The data are in the form :

$$\{(y_i; x_{1i}, x_{2i}, \cdots, x_{ki})\}_{i=1,\cdots,n}$$

In “regular” Regression Model, we impose the condition that Y is on the continuous scale – We even assume that Y is normally distributed with a constant variance.

COVARIATES

- In biomedical research, independent variables or covariates represent patients' characteristics and, in many cases of clinical research, one represents the treatment – or treatment's characteristics.
- Do we need to impose any assumptions on these predictor variables? Unlike the response variable, **we treat the covariates as "fixed".**
- There are even no restriction on measurement scale; there are **binary** covariates, **categorical** covariates, and **continuous covariates**.

But what's about the Dependent Variable Y?

We impose the condition that Y is on the continuous scale maybe because of the "normal error model" - **not because Y is always on the continuous scale**. In a variety of applications, the **Dependent Variable** of interest may have only two possible outcomes, and therefore can be represented by an **Indicator Variable Y** taking on values 0 and 1. Let:

π = **Pr(Y=1)**

Let Y be the Dependent Variable Y taking on values 0 and 1, and:

π = Pr(Y=1)

Y is said to have the "**Bernouilli distribution**" (Binomial with n = 1).

We have:

$$E(Y) = \pi$$

$$Var(Y) = \pi(1-\pi)$$

Consider, for example, an analysis of **whether or not business firms have a daycare facility**, according to the **number of female employees**, the **size of the firm**, the **type of business**, and the **annual revenue**. The dependent variable Y in this study was defined to have two possible outcomes:

(i) Firm **has a daycare** facility (Y=1), and
(ii) Firm **does not have** a daycare facility (Y=0).

As another example, consider a study of **Drug Use among middle school kids**, as a function of **gender** and **age** of kid, **family structure** (e.g. who is the head of household), and **family income**. In this study, the dependent variable Y was defined to have two possible outcomes:

(i) Kid **uses drug** (Y=1), and
(ii) Kid **does not use** drug (Y=0).

In another example, say, a man has a physical examination; he's concerned: **Does he have prostate cancer?** The "**truth**" would be confirmed by a biopsy. But it's a very painful process (at least, could we say if he needs a biopsy?)
In this study, the dependent variable Y was defined to have two possible outcomes:

(i) Man **has prostate cancer** (Y=1), and
(ii) Man **does not have** prostate cancer (Y=0).

Possible predictors include **PSA level, age, race.**

Suppose Prostate Cancer has been confirmed, the next concern is **whether the cancer has been spread to neighboring lymph nodes**; knowledge would dictate appropriate treatment strategy. The "**truth**" would be confirmed by performing a "laparotomy" (to examine the nodes), but any surgery involves risks; the question is **whether we can** accurately predict **nodal involvement** without a surgery.

In this study, the dependent variable Y was defined to have two possible outcomes:
(i) **With nodal involvement** (Y=1), and
(ii) **Without** nodal involvement (Y=0).

Possible "predictors" include **X-ray reading**, biopsy result pathology reading (**grade**), size and location of the tumor (**stage** - by palpation with the fingers via the rectum), and **"acid phosphatase level"** (in blood serum).

The basic question is: Can we do “regression” when the dependent variable, or “response”, is **binary**?

Recall the "**Normal Error Model**" where we model the "**Mean**" of Y as a function of X's:

$$\mathbf{Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon}$$

$$\boldsymbol{\varepsilon \in N(0, \sigma^2)}$$

For “binary” Dependent Variables, we run into problems with the “**Normal** Error Model” – **The distribution of Y is Bernouilli**. However, the “normal” assumption is not very important; effects of violation is quite minimal!

The Mean of Y is in well-defined but it has **limited range**:

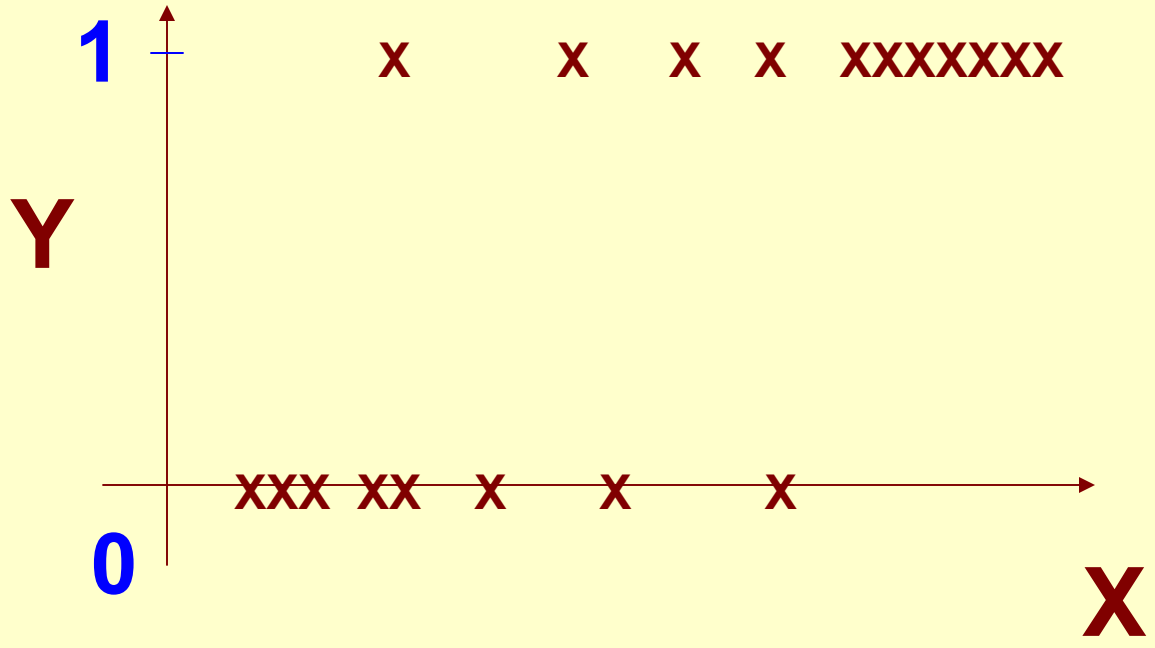
Mean of Y = Pr(Y=1) = π
$0 \leq \pi \leq 1$,

and **fitted values may fall outside of (0,1).** However, that's a minor problem.

The Variance (around the regression line) **is not constant** (a model violation that we learn in diagnostics); variance is function of the Mean π of Y (which is **a function of predictors**):

$$\sigma^2 = \pi(1- \pi)$$

More important, **the relationship is not linear**. For example, with one predictor X, we usually have:



In other words, We still can focus on "**modeling the mean**", in this case it is a Probability, π **= Pr(Y=1)**, but the usual linear regression with the "normal error regression model" is definitely not applicable – **all assumptions are violated, some may carry severe consequences.**

EXAMPLE: Dose-Response

Data in the table show the effect of different concentrations of (nicotine sulphate in a 1% saponin solution) on fruit flies; here X = log(100xDose), just making the numbers easier to read.

Dose(gm/100cc)	# of insects, n	# killed, r	x	p (%)
0.1	47	8	1.000	17.0
0.15	53	14	1.176	26.4
0.2	55	24	1.301	43.6
0.3	52	32	1.477	61.5
0.5	46	38	1.699	82.6
0.7	54	50	1.845	92.6
0.95	52	50	1.978	96.2

Proportion p is an estimate of Probability π

UNDERLYING ASSUMPTION

It is assumed that **each subject/fly has its own tolerance** to the drug. The amount of the chemical needed to kill an individual fruit fly, called "**individual lethal dose**" (ILD), cannot be measured - because **only one fixed dose is given to a group of n flies** (indirect assay)

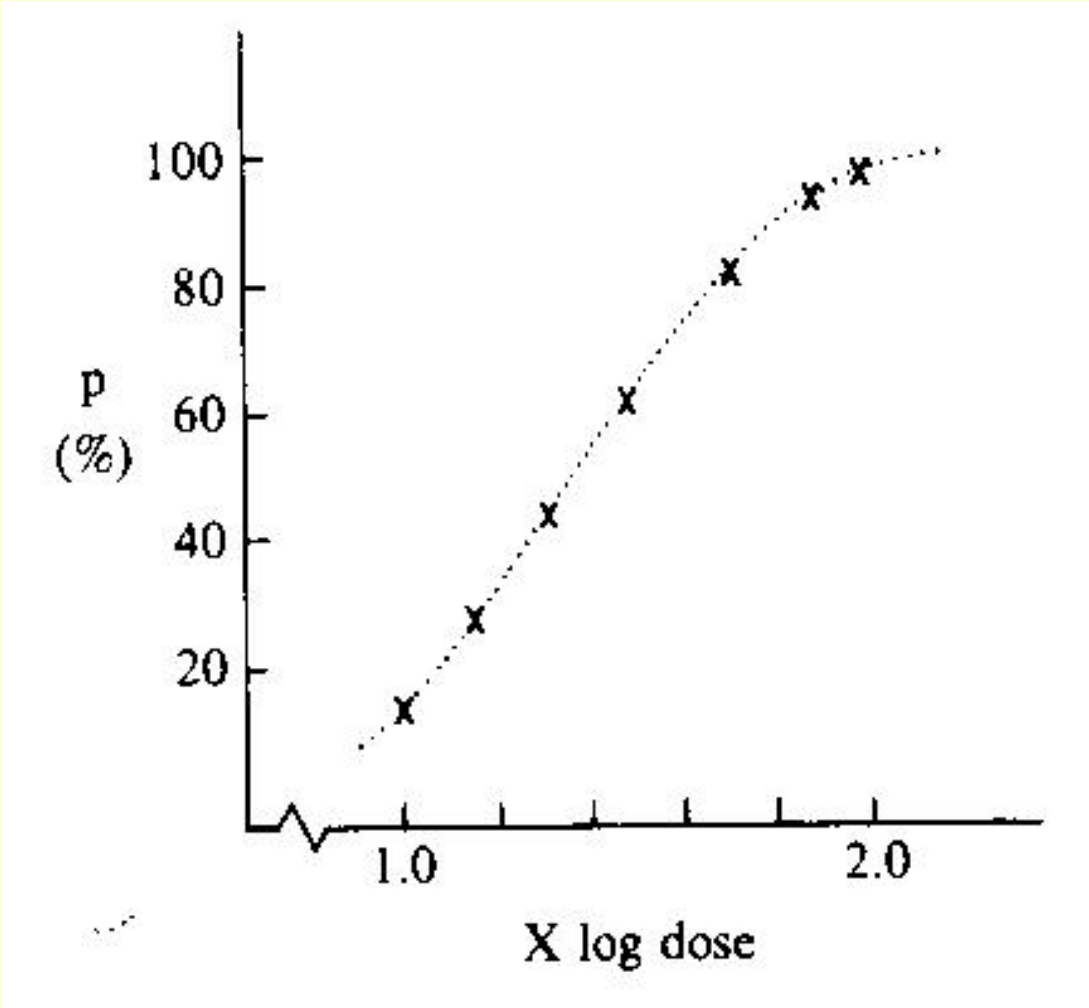
(1) If that dose is below some particular fly's ILD, the insect survived.

(2) Flies who **died** are those with **ILDs below the given fixed dose.**

INTERPRETATION OF DATA

Dose	# n	# killed	X	p(%)
0.1	47	8	1.000	17.0
0.15	53	14	1.176	26.4
0.2	55	24	1.301	43.6
0.3	52	32	1.477	61.5
0.5	46	38	1.699	82.6
0.7	54	50	1.845	92.6
0.95	52	50	1.978	96.2

- 17% (8 out of n=47) of the first group respond to dose of .1gm/100cc (x=1.0); that **means 17% of subjects have their ILDs less than .1**
- 26.4% (14 out of n=53) of the 2nd group respond to dose of .15gm/100cc (X=1.176); that means **26.4% of subjects have their ILDs less than .15**
- we view each dose D (with X = logD) as **upper endpoint** of an interval and **p the cumulative relative frequency.**



A symmetric **sigmoid dose-response curve** suggests that it be seen as some cumulative distribution function (**cdf**).

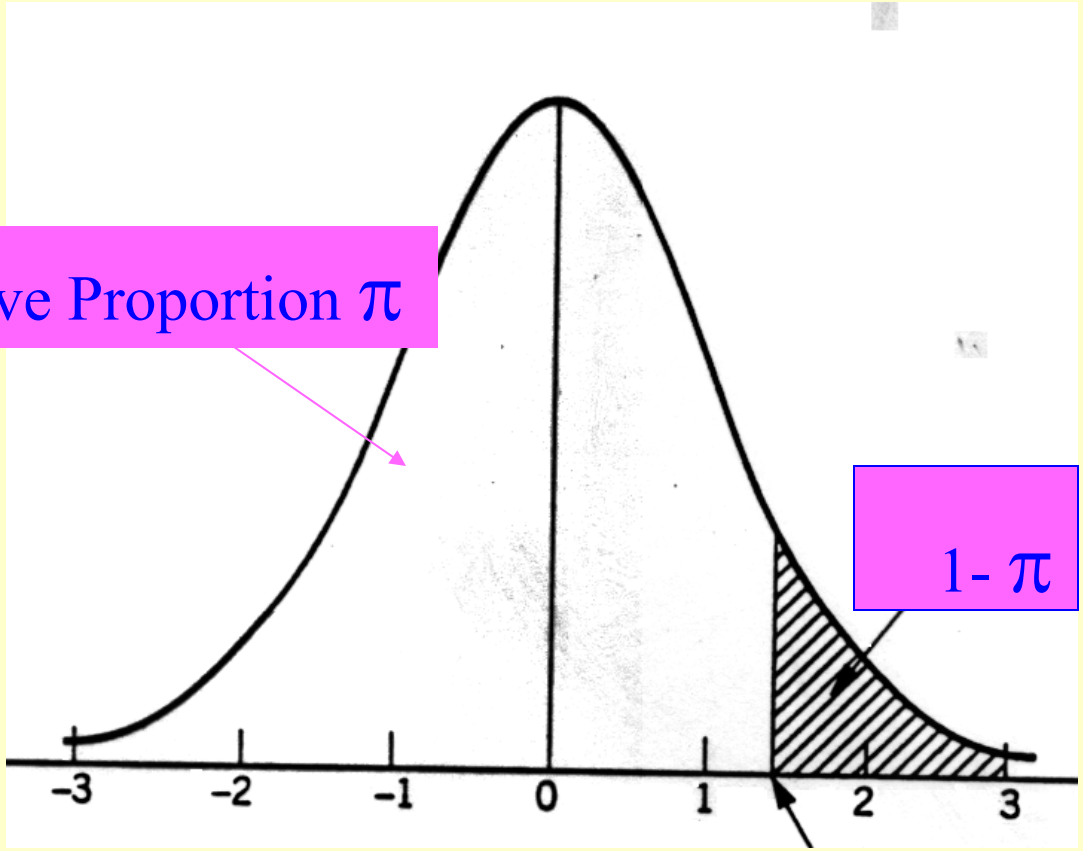
"**Empirical evidence**", i.e. data, suggest that we view **p the cumulative relative frequency**. This leads to a **"transformation" from "π" to an "upper endpoint", say Y*** (which is on the continuous scale) corresponding to that cumulative frequency of some cdf. After this transformation, **the regression model is then imposed on Y*,** the transformed value of π.

MODELING A PROBABILITY

Let π be the probability "to be modeled" and X a covariate (let consider only one X for simplicity). The first step in the regression modeling process is to obtain "**the equivalent deviate Y* of π**" using the following **transformation:**

$$\pi = \int_{-\infty}^{Y^*} f(z)dz \quad \text{or} \quad \int_{Y^*}^{\infty} f(z)dz$$

f(z) is some probability density function.



Transformation: π to Y*
which is on a linear scale

As a result, the proportion π has been transformed into a variable Y* on the "linear" or continuous scale with unbounded range. We can use Y* as the dependent variable in a regression model. (We now should only worry about "normality" which is not very important)

The relationship between covariate X (in the example, log of the dose) or covariates X's and Probability π (through Y) is then stipulated by the usual simple linear regression:

$$Y^* = \beta_0 + \beta_1 x$$

or multiple regression :

$$\mathrm{Y}^* = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

All we need is a "**probability density function**" f(.) in order to translate π to Y^* through :

$$\boldsymbol{\pi} = \int_{-\infty}^{\mathbf{Y}^*} \mathbf{f(z)dz} \text{ or } \int_{Y^*}^{\infty} f(z)dz$$

In theory, any probability density function can be used. We can choose one either by its **simplicity** and/or its **extensive scientific supports**. And we can check to see **if the data fit** the model (however, it's practically hard because we need lots of data to tell).

A VERY SIMPLE CHOICE

A possibility is "Unit Exponential Distribution"
with density :

$$f(z) = e^{-z};\ z \geq 0$$

Result (for one covariate X) is:

$$\pi = \int_{-\beta_0-\beta_1 x}^{\infty} e^{-z} dz$$

$$= e^{\beta_0+\beta_1 x}; \text{ or}$$

$$\ln\pi = \beta_0 + \beta_1 x$$

That is to **model the "log" of the probability as a "linear function" of covariates.**

Of course, you could use

"**multiple regression**" too :

$$\ln\pi = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

The advantage of the approach of modeling the "log" of the probability as a "linear function" of covariates, is **easy interpretation** of model parameters, the **probability is changed by a multiple constant** (i.e."**multiplicative model**" which is usually plausible)

Example : Say X_1 is binary

$$\ln \pi = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$x_1 = 0 \text{ (unexposed)} : \ln \pi_{unexposed} = \beta_0 + \beta_2 x_2$$

$$x_1 = 1 \text{ (exposed)} : \ln \pi_{exposed} = \beta_0 + \beta_1 + \beta_2 x_2$$

$$\beta_1 = \ln \pi_{exposed} - \ln \pi_{unexposed}$$

$$= \ln \frac{\pi_{exposed}}{\pi_{unexposed}}$$

$$= \ln(Odds)$$

The model is plausible; calculations could be simple too; after the log transformation of “p”, proceeding with usual steps in regression analysis.

this approach has a small problem: the exponential distribution is defined only on the whole positive range **and** certain choice of “x” could make the **fitted probabilities exceeding 1.0**

$$\ln\pi = \beta_0 + \beta_1 x$$

A HISTORICAL CHOICE

Besides the Unit Exponential probability density, one can also use of the **Standard Normal** density in the transformation of π:

$$\pi = \int_0^{y^*} f(\theta) d\theta$$

"f" is the Standard Normal density :

$$f(\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\theta^2}{2})$$

This “**Probit Transformation**” leads to the “**Probit Model**”; Y* is called the “probit” of π. The word “probit” is a shorten form of the phrase “PROBability unIT” (but it is not a probability), it is a standard normal variate.

The Probit Model was popular in years past and had been **used almost exclusively to analyze "bioassays" for many decades**. However, there is **no closed-form formula for Y*** (it's not possible to derive an equation relating π to x without using an integral sign):

$$\pi = \int_{0}^{\beta_0+\beta_1 x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{\theta^2}{2}) d\theta$$

Since it's not possible to derive an equation relating π to x without using an integral sign, the computation is much more complicated.

There is a SAS program (It's **PROC PROBIT**) but the use of the Probit Model has been faded.

LOGISTIC TRANSFORMATION

(Standard) **Logistic Distribution** with density :

$$f(\theta) = \frac{\exp(\theta)}{[1+\exp(\theta)]^2}$$

Result is:

$$\pi = \int_{-\infty}^{Y^* = \beta_0 + \beta x} \frac{e^{\theta}}{[1 + e^{\theta}]^2} d\theta$$

$$= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$= \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$

$$\boldsymbol{\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}$$

$$1 - \pi = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\frac{\pi}{1 - \pi} = e^{\beta_0 + \beta_1 x}$$

$$\boldsymbol{\log \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 x}$$

We refer to this as "**Logistic Regression**"

Exponential transformation leads to a linear model of "**Log of Probability**": $\ln(\pi)$;
Logistic transformation leads to a linear model of "**Log of Odds**": $\ln[\pi/(1-\pi)]$

When π is small (rare disease/event), the probability and the odds are approximately equal.

$$Odds = \frac{\pi}{1-\pi}$$

$$\pi = \frac{Odds}{1+Odds}$$

Advantages:

(1) Also very simple data transformation:

$Y = \log\{p/(1-p)\}$

(2) The **logistic density**, with **thicker tails** as compared to normal curve, may be a **better representation of real-life processes (compared to Probit Model which is based on the normal density).**

A POPULAR MODEL

- Although one can use the Standard Normal density in the regression modeling process (or any density function for that purpose),
- The Logistic Regression, as a result of choosing Logistic Density remains the **most popular choice** for a number of reasons: closed form formula for π, easy computing (Proc **LOGISTIC**)
- The most important reasons: **interpretation of model parameter** and **empirical supports**!

REGRESSION COEFFICIENTS

$$\pi = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\alpha + \beta x)}$$

$$\ln \frac{P}{1-P} = \beta_0 + \beta_1 x$$

β_1 represents the **log of the odds ratio** associated with X, if X is binary, or with "an unit increase" in X if X is on continuous scale; β_0 only depends on "event prevalence"- just like any **intercept**.

Example: Say X_1 is binary

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$x_1 = 0 \text{ (unexposed)} : \ln Odds_{unexposed} = \beta_0 + \beta_2 x_2$$

$$x_1 = 1 \text{ (exposed)} : \ln Odds_{exposed} = \beta_0 + \beta_1 + \beta_2 x_2$$

$$\ln Odds_{exposed} - \ln Odds_{unexposed} = \beta_1$$

$$\frac{Odds_{exposed}}{Odds_{unexposed}} = \text{Odds Ratio} = (e^{\beta_1})$$

β_1 is the odds ratio on the log scale if X is binary

Logistic Regression applies in both prospective and retrospective (case-control) designs. In **prospective design**, we can calculate/estimate the **probability of an event** (for specific values of covariates). In **retrospective design**, we cannot calculate/estimate the probability of events because the "intercept" is meaningless but **relationship between event and covariates are valid**.

SUPPORTS FOR LOGISTIC MODEL

The fit and the origin of the linear logistic model could be easily traced as follows. When a dose D of an agent is applied to a pharmacological system, the fractions f_a and f_u of the system affected and unaffected satisfy the so-called "median effect principle" (Chou, 1976):

$$\frac{f_a}{f_u} = \left\{ \frac{d}{ED_{50}} \right\}^m$$

where ED_{50} is the "median effective dose" and "m" is a Hill-type coefficient; m = 1 for first-degree or Michaelis-Menten system. The median effect principle has been investigated much very thoroughly in pharmacology.

If we set " $\pi = f_a$", the **median effect principle and the logistic regression model are completely identical** with a slope $\beta_1 = m$.

Besides the Model, the other aspect where Logistic Regression, both simple and multiple, is very different from our usual approach is **the way we estimate the parameters or regression coefficients.** The obstacle is the **lack of homoscedasticity**: we cannot assume a constant variance after the logistic transformation.

$$Y^* = \log \frac{P}{1-P}$$

$$Var(Y^*) = [\frac{dY}{dp}]^2 Var(p)$$

$$= \frac{1}{[p(1-p)]^2} Var(p)$$

$$= \frac{1}{[p(1-p)]^2} \frac{p(1-p)}{n}$$

$$= \frac{1}{np(1-p)}$$

Not constant

SOLUTION #1: WEIGHTED LS

Instead of minimizing the "sum of squares", we minimize the "weighted sum of squares"

$$\Sigma w[y^* - (\alpha + \beta x)]^2$$

where the weight for the value Y* is 1/Var(Y*). This **can be done but much more complicated**.

SOLUTION #2: MLE

Model :

$$\pi = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\alpha + \beta x)}$$

Likelihood Function :

$$L = \prod_{i=1}^{n} \Pr(Y_i = y_i)$$

$$= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

$$= \prod_{i=1}^{n} \frac{\{\exp[\beta_0 + \beta_1 x_i]\}^{y_i}}{1 + \exp[\beta_0 + \beta_1 x_i]}; y_i = 0/1$$

Maximum Likelihood Estimation (MLE) process gives us estimates of all regression coefficients and their standard errors, b_i (estimate of β_i) and $SE(b_i)$

TEST FOR SINGLE FACTOR

- The **question** is: "Does the addition of one particular factor of interest add significantly to the prediction of Pr(Y=1) **over and above that achieved by other factors**?".
- The **Null Hypothesis** for this test may stated as: "Factor X_i does not have any value added to the prediction of the probability of response over and above that achieved by other factors ". In other words,

$$H_0 : \beta_i = 0$$

TEST FOR SINGLE FACTOR

- The Null Hypothesis is $H_0 : \beta_i = 0$
- Regardless of the number of variables in the model, one simple approach is using $$\mathbf{z = \frac{b_i}{SE(b_i)}}$$
- Refer it to the percentiles of the **standard normal distribution**, where b_i is the corresponding estimated regression coefficient and $SE(b_i)$ is the standard error of β_i , both of which are provided by any computer package**.**

ESTIMATING ODDS RATIO

- General form of 95% CI for β_i: $b_i \pm 1.96*SE(b_i)$; b_i is point estimate of β_i, provided by SAS, and $SE(b_i)$ from Information matrix, also by SAS
- Transforming the 95% confidence interval for the parameter estimates to 95% C.I. for Odds Ratios:

$$\exp\left[b_i \pm 1.96 SE(b_i)\right]$$

Logistic Model For Interaction

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

"Usual approach": use the product of individual terms to represent "interaction" – also called "effect modification"

Example:

X1 = 1 for treatment and 0 for placebo

X2 = 1 for age $\geq$ 55 and 0 for age < 55

X3 = X1 * X2

So, X3 = 1 for treatment and age $\geq$ 55

X3 = all other combinations.

Logistic Model For Interaction

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

X1 = 1 treatment and 0 for placebo

X2 = 1 for age ≥ 55 and 0 for age < 55

X3 = X1 * X2

Log Odds (placebo, young) = β_0

Log Odds (active, young) = $\beta_0 + \beta_1$

Dif = β_1; exp(β_1) is odds (A v P) for young

Log Odds (placebo, old) = $\beta_0 + \beta_2$

Log Odds (active, old) = $\beta_0 + \beta_1 + \beta_2 + \beta_3$

Dif = $\beta_1 + \beta_3$; exp($\beta_1 + \beta_3$) is odds (A v P) for old

What does b_3 Mean?

Log Odds (placebo, young) $= \beta_0$

Log Odds (active, young) $= \beta_0 + \beta_1$

} $\exp(\beta_1)$ is odds (A v P) for young

Log Odds (placebo, old) $= \beta_0 + \beta_2$

Log Odds (active, old) $= \beta_0 + \beta_1 + \beta_2 + \beta_3$

} $\exp(\beta_1 + \beta_3)$ is odds (A v P) for old

$$\frac{\text{Odds (A v P) for Old}}{\text{Odds (A v P) for Young}} = \frac{\exp(\beta_1 + \beta_3)}{\exp(\beta_1)} = \exp(\beta_3)$$

A ratio of odds ratios!! Same multiplicative model

Interaction Hypothesis

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Q: Does the effect of active treatment on CVD differ for young versus older persons?

H_0: $\beta_3 = 0$

H_A: $\beta_3 \neq 0$

(Interaction = Effect Modification)

Because it is difficult to judge the lack of linearity, quadratic and polynomial models are rarely used in Logistic Regression. But it can be done

Simply assuming a Quadratic Model, then check for Quadratic Effect: $H_0: \beta_2 = 0$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$