Week 9

Logistic Regression: Simple and Multiple Regression Analysis

Dr. Esam Mahdi

Department of Mathematics, Statistics and Physics

Qatar University

Learning Objectives

- Discuss the basic ideas of logistic regression analysis for simple and multiple cases for risk determination.
- Know how to interpret the odds ratio in logistic regression context.
- Test the hypotheses and goodness of fit in multiple logistic regression.
- Specify the strategy for selecting variables included in the multiple logistic regression model.
  - Forward selection procedure.
  - Backward elimination procedure.
  - Stepwise regression procedure.
- Use statistical software such as R and SPSS to fit a logistic regression model to some real data and interpret the output results.

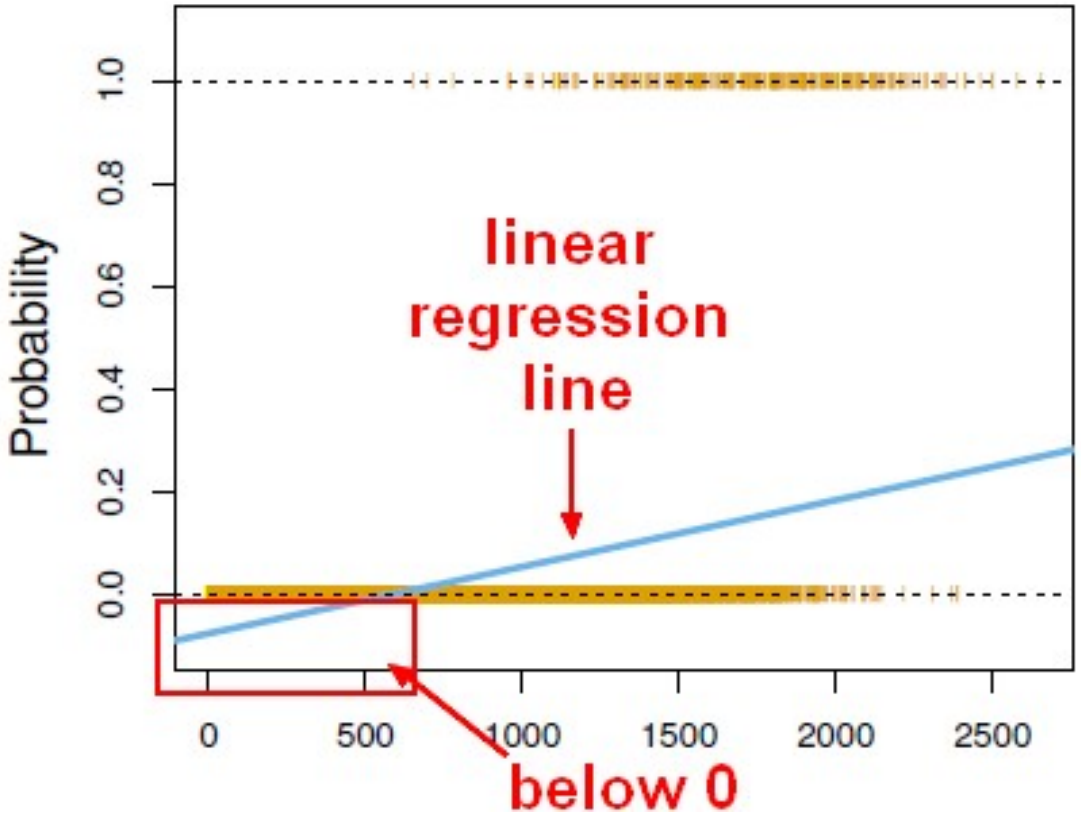
Regression Model

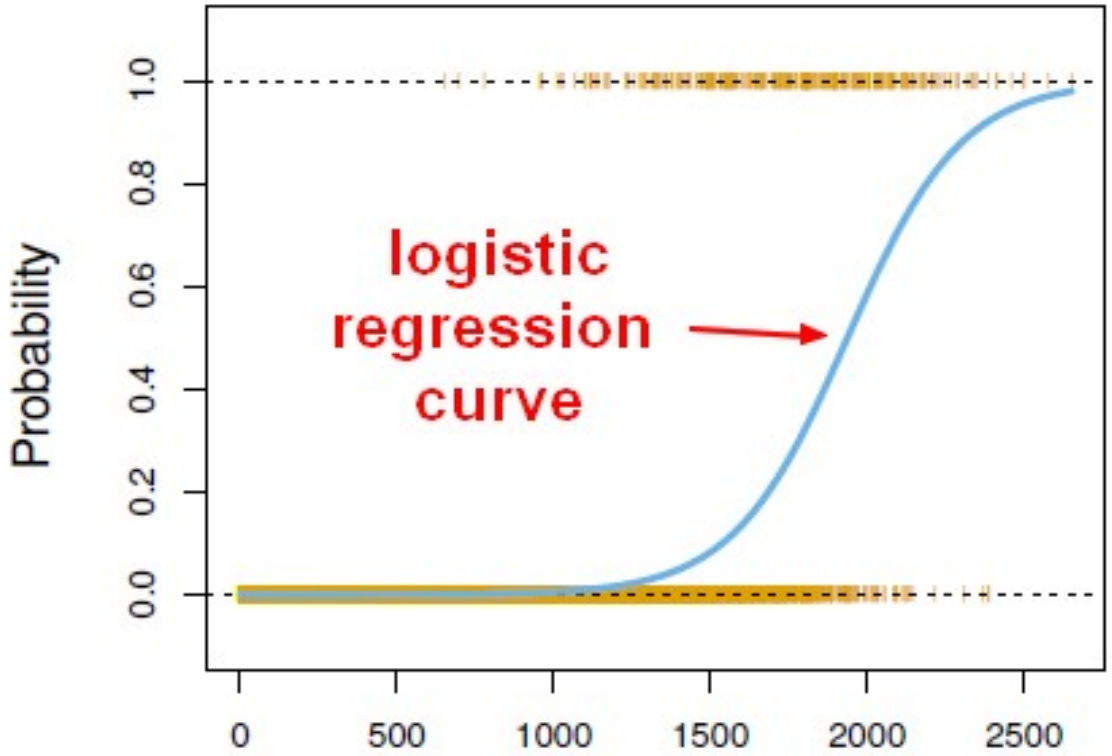
- **Regression** techniques often used as statistical analysis tools to assess the relationships among a set of variables.
- In most cases, one variable is usually taken to be the *response* or *dependent variable*, that is, a variable to be predicted from or explained by other variables.
- The other variables are called *predictors*, *explanatory* variables, or *independent variables*.
- The usual regression analysis goal is to describe the mean of a dependent variable Y as a function of a set of predictor variables.
- Regression analysis serves two major purposes:
  - control or intervention.
  - prediction.

Logistic Regression Model

- In many applications, the dependent variable of interest is not on a continuous scale; it may have only two possible outcomes and therefore can be represented by an indicator variable taking on values 0 and 1.
- In such cases, the dependent variable is **dichotomous (binary)** and hence may be represented by a variable taking the value 1 with probability π and the value 0 with probability $1 - \pi$.
- Such a variable is a point binomial variable, that is, a binomial variable with $n = 1$ trial, and the model often used to express the probability π as a function of potential independent variables under investigation is the **logistic regression model**.

Why Logistic Regression and Not Linear Regression?





Logistic Regression vs Linear Regression

Linear regression	Logistic regression
Response (dependent) variable is continuous	Response (dependent) variable is dichotomous
Response (dependent) variable usually follows a normal distribution	Response (dependent) variable usually follows a binomial distribution
Estimate the average value of the dependent variable when the covariates (independent or predictors) variables are fixed	Estimate the probability that an event occurs versus the probability that the event does not occur
Predict the effect of a series of values on the mean of a continuous response variable	Predict the effect of a series of values on a binary response variable

- In logistic regression, we ask our self a question like:

  What is the probability that a person having 3 children in age 40 and smoking 10 cigarettes per a day would be died early?

Simple Logistic Regression

- The logistic regression deals with the case where the basic random variable Y of interest is a dichotomous variable taking the value 1 with probability π and the value 0 with probability $1 - \pi$.
- Such a random variable is called a *point-binomial* or *Bernouilli variable*, and it has the simple discrete probability distribution

$$\Pr(Y = y) = \pi^y (1 - \pi)^{1-y} \qquad y = 0, 1$$

- Suppose that for the ith individual of a sample ($i = 1, 2, \ldots, n$), Y_i is a Bernouilli variable with

$$\Pr(Y_i = y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \qquad y_i = 0, 1$$

Simple Logistic Regression

The logistic regression analysis assumes that the relationship between π_i and the covariate value x_i of the same person is described by the logistic function

$$\pi_i = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_i)]} \qquad i = 1, 2, \ldots, n,$$

The basic logistic function is given by

$$\boxed{f(z) = \frac{1}{1 + e^{-z}}}$$

where, as in this simple regression model,

$$z_i = \beta_0 + \beta_1 x_i$$

or, in the multiple regression model of subsequent sections,

$$z_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}$$

representing an index of combined risk factors.

Simple Logistic Regression

There are two important reasons that make logistic regression popular:

1. The range of the logistic function is between 0 and 1; that makes it suitable for use as a probability model, representing individual risk.
2. The logistic curve has an increasing S-shape with a threshold (Figure 9.1); that makes it suitable for use as a biological model, representing risk due to exposure.

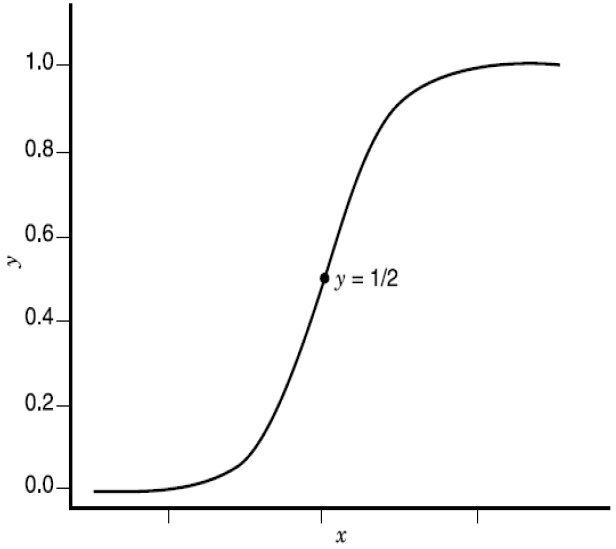


Figure 9.1 General form of a logistic curve.

Simple Logistic Regression

Under the simple logistic regression model, the likelihood function is given by

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

$$= \prod_{i=1}^{n} \frac{[\exp(\beta_0 + \beta_1 x_i)]^{y_i}}{1 + \exp(\beta_0 + \beta_1 x_i)} \qquad y_i = 0, 1$$

from which we can obtain maximum likelihood estimates of the parameters β_0 and β_1. As mentioned previously, the logistic model has been used both extensively and successfully to describe the probability of developing $(Y = 1)$ some disease over a specified time period as a function of a risk factor X.

Measure of Association

- In epidemiological studies, the strength of a statistical relationship between the binary dependent variable and each independent variable or covariate measured are usually measured by the **relative risk** or **odds ratio.**
- When the logistic model is used, such effects are usually measured by the odds ratio.
- The logistic function for the probability π_i can also be expressed as a linear model in the log scale (of the odds):

$$\ln \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_i$$

Logistic Model and Odds Ratio for Binary Covariate

- Consider the case of a binary covariate with the conventional coding:

$$X_i = \begin{cases} 0 & \text{if the patient is not exposed} \\ 1 & \text{if the patient is exposed} \end{cases}$$

- It can be seen that from the log-linear form of the logistic regression model,

$$\ln(\text{odds; nonexposed}) = \beta_0$$

$$\ln(\text{odds; exposed}) = \beta_0 + \beta_1$$

- After exponentiating, the difference leads to

$$e^{\beta_1} = \frac{(\text{odds; exposed})}{(\text{odds; nonexposed})}$$

- This represents the odds ratio (OR) associated with the exposure, exposed versus nonexposed. *In other words, the primary regression coefficient β_1 is the value of the odds ratio on the log scale.*

Logistic Model and Odds Ratio for continuous covariate

- Consider a continuous covariate X and any value x of X,

$$\ln(\text{odds};\, X = x) = \beta_0 + \beta_1(x)$$
$$\ln(\text{odds};\, X = x+1) = \beta_0 + \beta_1(x+1)$$

- After exponentiating, the difference leads to

$$e^{\beta_1} = \frac{(\text{odds};\, X = x+1)}{(\text{odds};\, X = x)}$$

- This represents the odds ratio (OR) associated with a 1-unit increase in the value of X, $X = x + 1$ versus $X = x$.
- For example, a systolic blood pressure of 114 mmHg versus 113 mmHg. For an m-unit increase in the value of X, say $X = x + m$ versus $X = x$, the corresponding odds ratio is $e^{m\beta_1}$.

The primary regression coefficient β_1 (and β_0, which is often not needed) can be estimated iteratively using a computer-packaged program such as SAS, SPSS, R, etc.

- The point estimate $\widehat{\text{OR}} = e^{\hat{\beta}_1}$ and its 95% confidence interval $\exp[\hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1)]$

Effect of Measurement Scale

- The odds ratio, used as a measure of association between the binary dependent variable and a covariate, depends on the coding scheme for a binary covariate and for a continuous covariate X, the scale with which to measure X.
  - For example, if we use the following coding for a factor,

$$X_i = \begin{cases} -1 & \text{if the subject is not exposed} \\ 1 & \text{if the subject is exposed} \end{cases}$$

then

$$\ln(\text{odds; nonexposed}) = \beta_0 - \beta_1$$
$$\ln(\text{odds; exposed}) = \beta_0 + \beta_1$$

so that

$$\begin{aligned} \text{OR} &= \exp[\ln(\text{odds; exposed}) - \ln(\text{odds; nonexposed})] \\ &= e^{2\beta_1} \end{aligned}$$

and its 95% confidence interval,

$$\exp[2(\hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1))]$$

Example 9.1

- When a patient is diagnosed as having cancer of the prostate, an important question in deciding on treatment strategy for the patient is whether or not the cancer has spread to neighboring lymph nodes.

- We are interested in the predictive value of the level of acid phosphatase in blood serum for 53 prostate cancer patients receiving surgery.
- Table 9.1 presents the complete data set.
  - For each of the 53 patients, there are two continuous independent variables:
    - age at diagnosis and level of serum acid phosphatase (×100; called ‘‘acid’’),
  - Three binary variables:
    - x-ray reading, pathology reading (grade) of a biopsy of the tumor obtained by needle before surgery, and a rough measure of the size and location of the tumor (stage) obtained by palpation with the fingers via the rectum.
    - For these three binary independent variables a value of 1 signifies a positive or more serious state and a 0 denotes a negative or less serious finding.
  - The sixth column presents the finding at surgery—the primary binary response or dependent variable Y, a value of 1 denoting nodal involvement, and a value of 0 denoting no nodal involvement found at surgery.

TABLE 9.1

X-ray	Grade	Stage	Age	Acid	Nodes	X-ray	Grade	Stage	Age	Acid	Nodes
0	1	1	64	40	0	0	0	0	60	78	0
0	0	1	63	40	0	0	0	0	52	83	0
1	0	0	65	46	0	0	0	1	67	95	0
0	1	0	67	47	0	0	0	0	56	98	0
0	0	0	66	48	0	0	0	1	61	102	0
0	1	1	65	48	0	0	0	0	64	187	0
0	0	0	60	49	0	1	0	1	58	48	1
0	0	0	51	49	0	0	0	1	65	49	1
0	0	0	66	50	0	1	1	1	57	51	1
0	0	0	58	50	0	0	1	0	50	56	1
0	1	0	56	50	0	1	1	0	67	67	1
0	0	1	61	50	0	0	0	1	67	67	1
0	1	1	64	50	0	0	1	1	57	67	1
0	0	0	56	52	0	0	1	1	45	70	1
0	0	0	67	52	0	0	0	1	46	70	1
1	0	0	49	55	0	1	0	1	51	72	1
0	1	1	52	55	0	1	1	1	60	76	1
0	0	0	68	56	0	1	1	1	56	78	1
0	1	1	66	59	0	1	1	1	50	81	1
1	0	0	60	62	0	0	0	0	56	82	1
0	0	0	61	62	0	0	0	1	63	82	1
1	1	1	59	63	0	1	1	1	65	84	1
0	0	0	51	65	0	1	0	1	64	89	1
0	1	1	53	66	0	0	1	0	59	99	1
0	0	0	58	71	0	1	1	1	68	126	1
0	0	0	63	75	0	1	0	0	61	136	1
0	0	1	53	76	0						

Example 9.2

- Refer to the data for patients diagnosed as having cancer of the prostate in Example 9.1 (Table 9.1)
- Investigate the relationship between nodal involvement found at surgery and the level of acid phosphatase in blood serum in two different ways using either
  a) X = acid.
  b) X = log10 (acid).

Solution:

(a) For $X = \text{acid}$, we find that

$$\hat{\beta}_1 = 0.0204$$

from which the odds ratio for $(\text{acid} = 100)$ versus $(\text{acid} = 50)$ would be

$$\begin{aligned} \text{OR} &= \exp[(100 - 50)(0.0204)] \\ &= 2.77 \end{aligned}$$

(b) For $X = \log_{10}(\text{acid})$, we find that

$$\hat{\beta}_1 = 5.1683$$

from which the odds ratio for $(\text{acid} = 100)$ versus $(\text{acid} = 50)$ would be

$$\begin{aligned} \text{OR} &= \exp\{[\log_{10}(100) - \log_{10}(50)](5.1683)\} \\ &= 4.74 \end{aligned}$$

Overdispersion

- **Overdispersion** is a common phenomenon in practice and it could cause problems.

Measuring and Monitoring Dispersion

- **Dispersion** is measured by the scaled deviance or scaled Peason chi-square.
- scaled deviance and scaled Peason chi-square are respectively the deviance and Pearson chi-square divided by the degrees of freedom.
- The **deviance** is a goodness-of-fit statistic defined as two times the log-likelihood ratio of the full model compared to the reduced model of the regression parameters.
- Suppose that data are with replications consisting of m subgroups; then
  - the Pearson chi-square is $X_P^2 = \sum_i \sum_j \frac{(r_{ij} - n_i p_{ij})^2}{n_i p_{ij}}$
  - the deviance is $X_D^2 = \sum_i \sum_j r_{ij} log \frac{r_{ij}}{n_i p_{ij}}$

- Each of these goodness-of-fit statistics divided by the appropriate degrees of freedom, called the scaled Pearson chi-square and scaled deviance, respectively, can be used as a measure for overdispersion:
- When their values are much larger than 1, the assumption of binomial variability may not be valid and the data are said to exhibit overdispersion.

Overdispersion

- Several factors can cause overdispersion; among these are such problems as outliers in the data, omitting important covariates in the model, and the need to transform some explanatory factors.
- It is preferable to form subgroup when you fit a logistic regression.
- Without such a grouping, data may be too sparse, the Pearson chi-square and deviance do not have a chi-square distribution, and the scaled Pearson chi-square and scaled deviance cannot be used as indicators of overdispersion.
- A large difference between the scaled Pearson chi-square and scaled deviance provides evidence of this situation.

Fitting an Overdispersed Logistic Model

- One way of correcting overdispersion is to multiply the covariance matrix by the value of the overdispersion parameter Ø, scaled Pearson chi-square, or scaled deviance

$$E(p_i) = \pi_i$$

$$\text{Var}(p_i) = \phi\pi_i(1 - pi_i)$$

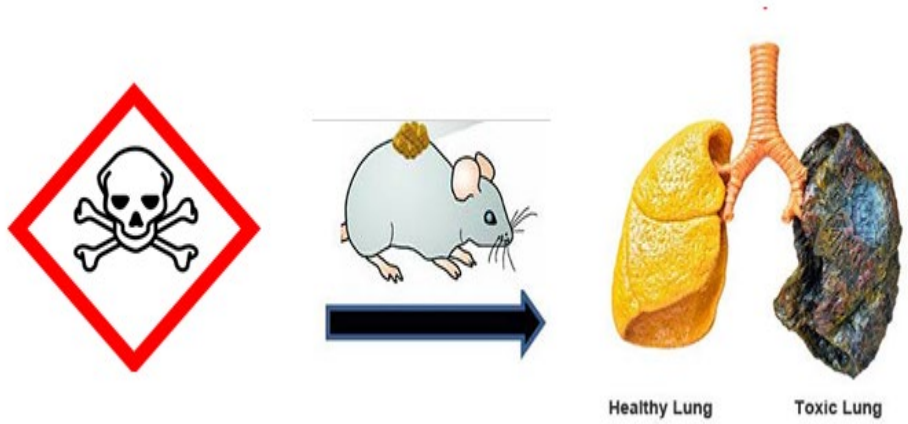
- In this correction process, the parameter estimates are not changed. However, their standard errors are adjusted (increased), affecting their significant levels (reduced).

Example 9.3

- In a study of the toxicity of certain chemical compound, five groups of 20 rats each were fed for four weeks by a diet mixed with that compound at five different doses.
- At the end of the study, their lungs were harvested and subjected to histopathological examinations to observe for sign(s) of toxicity (yes = 1, no = 0).
- The results are shown in Table 9.2.

TABLE 9.2

Group	Dose (mg)	Number of Rats	Number of Rats with Toxicity
1	5	20	1
2	10	20	3
3	15	20	7
4	20	20	14
5	30	20	10



Example 9.3 (Cont.)

- Fit of the simple logistic regression model yields Table 9.3.

TABLE 9.3

Variable	Coefficient	Standard Error	*z* Statistic	*p* Value
Intercept	−2.3407	0.5380	−4.3507	0.0001
Dose	0.1017	0.0277	3.6715	0.0002

- The results in Table 9.4 for the monitoring of overdispersion

TABLE 9.4

Parameter	Chi-Square	Degrees of Freedom	Scaled Parameter
Pearson	10.9919	3	3.664
Deviance	10.7863	3	3.595

- The results above indicate an obvious sign of overdispersion.
- By fitting an overdispersed model, controlling for the scaled deviance, we have Table 9.5.

TABLE 9.5

Variable	Coefficient	Standard Error	*z* Statistic	*p* Value
Intercept	−2.3407	1.0297	−2.2732	0.0230
Dose	0.1017	0.0530	1.9189	0.0548

- Compared to the previous results, the point estimates remain the same but the standard errors are larger.
- The effect of dose is no longer significant at the 5% level.

Multiple Regression Analysis

- The effect of some factor on a dependent or response variable may be influenced by the presence of other factors through effect modifications (i.e., interactions) .
- It is desirable to consider a large number of factors and sort out which ones are most closely related to the dependent variable.
- **Multiple logistic regression analysis**, involves a linear combination of the explanatory or independent variables
  - The variables must be quantitative with particular numerical values for each patient.
  - A covariate or independent variable, such as a patient characteristic, may be dichotomous, polytomous, or continuous (categorical factors will be represented by dummy variables).
    - Examples of dichotomous covariates are gender and presence/absence of certain comorbidity.
    - Examples of polytomous covariates include race and different grades of symptoms; these can be covered by the use of dummy variables.
    - Examples of continuous covariates include patient age and blood pressure.
- In many cases, data transformations (e.g., taking the logarithm) may be desirable to satisfy the linearity assumption.

Logistic Regression Model with Several Covariates

Suppose that we want to consider k covariates simultaneously; the simple logistic model of Section 9.1 can easily be generalized and expressed as

$$\pi_i = \frac{1}{1 + \exp[-(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji})]} \qquad i = 1, 2, \ldots, n$$

or, equivalently,

$$\ln \frac{\pi_i}{1 - \pi_i} = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}$$

This leads to the likelihood function

$$L = \prod_{i=1}^{n} \frac{[\exp(\beta_0 + \sum_{i=1}^{k} \beta_j x_{ji})]^{y_i}}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji})} \qquad y_i = 0, 1$$

from which parameters can be estimated iteratively using a computer-packaged

Logistic Regression Model with Several Covariates

- Similar to the univariate case, $\exp(\beta_i)$ represents one of the following:

1. The odds ratio associated with an exposure if X_i is binary (exposed $X_i = 1$ versus unexposed $X_i = 0$); or
2. The odds ratio due to a 1-unit increase if X_i is continuous ($X_i = x + 1$ versus $X_i = x$).

- After $\hat{\beta}_i$ and its standard error have been obtained, a 95% confidence interval for the odds ratio above is given by

$$\exp[\hat{\beta}_i \pm 1.96\ \mathrm{SE}(\hat{\beta}_i)]$$

- The use of products such as X_1X_2 and higher power terms such as X_1^2 may be necessary and can improve the goodness of fit in some cases.
- The cross-product term X_1X_2 is called an **interaction term**.

Effect Modifications

Consider the model

$$\pi_i = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i})]} \qquad i = 1, 2, \ldots, n$$

The meaning of β_1 and β_2 here is not the same as that given earlier because of the cross-product term $\beta_3 x_1 x_2$. Suppose that both X_1 and X_2 are binary.

1. For $X_2 = 1$, or exposed, we have

$$(\text{odds ratio; not exposed to } X_1) = e^{\beta_0 + \beta_2}$$

$$(\text{odds ratio; exposed to } X_1) = e^{\beta_0 + \beta_1 + \beta_2 + \beta_3}$$

   so that the ratio of these ratios, $e^{\beta_1 + \beta_3}$, represents the odds ratio associated with X_1, exposed versus nonexposed, in the presence of X_2, whereas

2. For $X_2 = 0$, or not exposed, we have

$$(\text{odds ratio; not exposed to } X_1) = e^{\beta_0} \text{ (i.e., baseline)}$$

$$(\text{odds ratio; exposed to } X_1) = e^{\beta_0 + \beta_1}$$

   so that the ratio of these ratios, e^{β_1}, represents the odds ratio associated with X_1, exposed versus nonexposed, in the absence of X_2. In other words, the effect of X_1 depends on the level (presence or absence) of X_2, and vice versa.

Effect Modification

- The use of the cross-product term X_1X_2 will help in the investigation of possible effect modifications.
- If $\beta_3 = 0$, the effect of two factors acting together, as measured by the odds ratio, is equal to the combined effects of two factors acting separately, as measured by the product of two odds ratios:

$$e^{\beta_1+\beta_2} = e^{\beta_1}e^{\beta_2}$$

- This fits the classic definition of no interaction on a multiplicative scale.

Testing Hypotheses in Multiple Logistic Regression

- Once we have fit a multiple logistic regression model and obtained estimates for the various parameters of interest, we want to answer questions about the contributions of various factors to the prediction of the binary response variable.
- **There are three types of such questions:**

1) **Overall test:** *Does the entire set of explanatory or independent variables contribute significantly to the prediction of response?*
2) **Test for the value of a single factor:** Does the addition of one particular variable of interest add significantly to the prediction of response over and above that achieved by other independent variables?
3) **Test for contribution of a group of variables:** Does the addition of a group of variables add significantly to the prediction of response over and above that achieved by other independent variables?

Overall Regression Tests

- Consider the first question stated above concerning an overall test for a model containing k factors.

The null hypothesis for this test may be stated as: "All k independent variables *considered together* do not explain the variation in the responses." In other words,

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Two likelihood-based statistics can be used to test this *global* null hypothesis; each has a symptotic chi-square distribution with k degrees of freedom under H_0.

1. Likelihood ratio test:

$$\chi^2_{\mathrm{LR}} = 2[\ln L(\hat{\boldsymbol{\beta}}) - \ln L(\mathbf{0})]$$

2. Score test:

$$\chi^2_S = \left[\frac{\delta \ln L(\mathbf{0})}{\delta \beta}\right]^T \left[-\frac{\delta^2 \ln L(\mathbf{0})}{\delta \beta^2}\right]^{-1} \left[\frac{\delta \ln L(\mathbf{0})}{\delta \beta}\right]$$

- Both statistics are provided by most standard computer programs

Example 9.4

- Refer to the data set on prostate cancer of Example 9.1 (Table 9.1).
- With all five covariates, we have the following test statistics for the global null hypothesis:

1. Likelihood test:

$$\chi^2_{\mathrm{LR}} = 22.126 \text{ with } 5 \text{ df}; \quad p = 0.0005$$

2. Score test:

$$\chi^2_S = 19.451 \text{ with } 5 \text{ df}; \quad p = 0.0016$$

- Both tests yield the same result that the null hypothesis should be rejected. Thus the entire set of covariates contribute significantly to the prediction of response.

Tests for a Single Variable

- Assume that we wish to test whether the addition of one particular independent variable of interest adds significantly to the prediction of the response over and above that achieved by other factors already present in the model.
- The null hypothesis for this test may stated as:

"Factor X_i does not have any value added to the prediction of the response given that other factors are already included in the model." In other words H_0: $\beta_i = 0$

- To test the null hypothesis, one can perform a likelihood ratio chi-squared test, with 1 df,

$$\chi^2_{\mathrm{LR}} = 2[\ln L(\hat{\boldsymbol{\beta}};\ \text{all } X\text{'s}) - \ln L(\hat{\boldsymbol{\beta}};\ \text{all other } X\text{'s with } X_i \text{ deleted})]$$

- A much easier alternative method is using the z-statistic,

$$z_i = \frac{\hat{\beta}_i}{\mathrm{SE}(\hat{\beta}_i)}$$

where $\hat{\beta}_i$ is the corresponding estimated regression coefficient and SE($\hat{\beta}_i$) is the estimate of the standard error of ^ bi

Example 9.5

- Refer to the data set on prostate cancer of Example 9.1 (Table 9.1).
- With all five covariates, we have the results shown in Table 9.6.

TABLE 9.6

Variable	Coefficient	Standard Error	z Statistic	p Value
Intercept	0.0618	3.4599	0.018	0.9857
X-ray	2.0453	0.8072	2.534	0.0113
Stage	1.5641	0.7740	2.021	0.0433
Grade	0.7614	0.7708	0.988	0.3232
Age	−0.0693	0.0579	−1.197	0.2314
Acid	0.0243	0.0132	1.850	0.0643

- The effects of x-ray and stage are significant at the 5% level, whereas the effect of acid is marginally significant ($p = 0.0643$).

Contribution of a Group of Variables

- This testing procedure addresses the more general problem of assessing the additional contribution of two or more factors to the prediction of the response over and above that made by other variables already in the regression model.
- In other words, the null hypothesis is of the form

$$\mathrm{H}_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

- To test such a null hypothesis, one can perform a likelihood ratio chi-square test, with m df,

$$\chi^2_{\mathrm{LR}} = 2[\ln L(\hat{\beta};\ \text{all } X\text{'s}) - \ln L(\hat{\beta};\ \text{all other } X\text{'s with } X\text{'s under investigation deleted})]$$

- As with the z test, this multiple contribution procedure is useful for assessing the importance of potential explanatory variables.
- In particular it is often used to test whether a similar group of variables, such as demographic characteristics, is important for the prediction of the response; these variables have some trait in common.

Example 9.7 Refer to the data set on prostate cancer of Example 9.1 (Table 9.1) with all five covariates. We consider, collectively, these four interaction terms: acid × x-ray, acid × stage, acid × grade, and acid × age. The basic idea is to see if *any* of the other variable would modify the effect of the level of acid phosphatase on the response.

1. With the original five variables, we obtained $\ln L = -24.063$.
2. With all nine variables, five original plus four products, we obtained $\ln L = -20.378$.

Therefore,

$$\begin{aligned}\chi^2_{\text{LR}} &= 2[\ln L(\hat{\boldsymbol{\beta}};\ \text{nine variables}) - \ln L(\hat{\boldsymbol{\beta}};\ \text{five original variables})] \\ &= 7.371; 4\ \text{df}, 0.05 \leq p\text{-value} \leq 0.10\end{aligned}$$

In other words, all four interaction terms, *considered together*, are marginally significant ($0.05 \leq p$-value ≤ 0.10); there may be some weak effect modification and that the effect of acid phosphatase on the response may be somewhat stronger for a certain combination of levels of the other four variables.

Stepwise Regression

- In many applications, we wish to avoid type I error.
- In regression analysis, a type I error corresponds to including a predictor that has no real relationship to the outcome.
- Thus, we need to identify from many available factors a small subset of factors that relate significantly to the outcome (e.g., the disease under investigation).
- In a multiple regression analysis, this goal can be achieved by using a strategy that adds into or removes from a regression model one factor at a time according to a certain order of relative importance.
- **The two important steps are as follows:**
- Specify a criterion or criteria for selecting a model.
- Specify a strategy for applying the criterion or criteria chosen.

Forward Selection Procedure

Strategies: This is concerned with specifying whether and which particular variable should be added to a model or whether any variable should be deleted from a model at a particular stage of the process.

Forward Selection Procedure

1. Fit a simple logistic regression model to each factor, one at a time.
2. Select the most important factor according to certain predetermined criterion.
3. Test for the significance of the factor selected in step 2 and determine, according to a certain predetermined criterion, whether or not to add this factor to the model.
4. Repeat steps 2 and 3 for those variables not yet in the model. At any subsequent step, if none meets the criterion in step 3, no more variables are included in the model and the process is terminated.

Backward Elimination Procedure

1. Fit the multiple logistic regression model containing all available independent variables:
2. Select the least important factor according to a certain predetermined criterion; this is done by considering one factor at a time and treat it as though it were the last variable to enter.
3. Test for the significance of the factor selected in step 2 and determine, according to a certain predetermined criterion, whether or not to delete this factor from the model.
4. Repeat steps 2 and 3 for those variables still in the model. At any subsequent step, if none meets the criterion in step 3, no more variables are removed in the model and the process is terminated.

Stepwise Regression Procedure

- Stepwise regression is a modified version of forward regression that permits reexamination, at every step, of the variables incorporated in the model in previous steps.
- A variable entered at an early stage may become superfluous at a later stage because of its relationship with other variables now in the model; the information it provides becomes redundant.
- That variable may be removed, if meeting the elimination criterion, and the model is refitted with the remaining variables, and the forward process goes on.
- The entire process, one step forward followed by one step backward, continues until no more variables can be added or removed.

Criteria

Criteria For the first step of the forward selection procedure, decisions are based on individual score test results (chi-square, 1 df). In subsequent steps, both forward and backward, the ordering of levels of importance (step 2) and the selection (test in step 3) are based on the likelihood ratio chi-square statistic:

$$\chi^2_{\mathrm{LR}} = 2[\ln L(\hat{\beta};\ \text{all other } X\text{'s}) - \ln L(\hat{\beta};\ \text{all other } X\text{'s with one } X \text{ deleted})]$$

Example 9.8

- Refer to the data set on prostate cancer of Example 9.1 (Table 9.1)
- With all five covariates: x-ray, stage, grade, age, and acid.
- This time we perform a stepwise regression analysis in which we specify that a variable has to be significant at the 0.10 level before it can enter into the model and that a variable in the model has to be significant at the 0.15 for it to remain in the model (most standard computer programs allow users to make these selections; default values are available).

Example 9.8 (Cont.)

- **First**, we get these individual score test results for all variables (Table 9.8).
- These indicate that x-ray is the most significant variable.
- **Step 1:** Variable ‘‘x-ray’’ is entered. Analysis of variables not in the model is shown in Table 9.9.
- **Step 2:** Variable ‘‘stage’’ is entered. Analysis of variables in the model (Table 9.10) shows that neither variable is removed. Analysis of variables not in the model is shown in Table 9.11.
- **Step 3:** Variable ‘‘acid’’ is entered. Analysis of variables in the model is shown in Table 9.12. None of the variables are removed. Analysis of variables not in the model is shown in Table 9.13. No (additional) variables meet the 0.1 level for entry into the model.

TABLE 9.8

Variable	Score χ^2	p Value
X-ray	11.2831	0.0008
Stage	7.4383	0.0064
Grade	4.0746	0.0435
Age	1.0936	0.2957
Acid	3.1172	0.0775

TABLE 9.9

Variable	Score χ^2	p Value
Stage	5.6394	0.0176
Grade	2.3710	0.1236
Age	1.3523	0.2449
Acid	2.0733	0.1499

TABLE 9.10

Factor	Coefficient	Standard Error	z Statistic	p Value
Intercept	−2.0446	0.6100	−3.352	0.0008
X-ray	2.1194	0.7468	2.838	0.0045
Stage	1.5883	0.7000	2.269	0.0233

TABLE 9.11

Variable	Score χ^2	p value
Grade	0.5839	0.4448
Age	1.2678	0.2602
Acid	3.0917	0.0787

TABLE 9.12

Factor	Coefficient	Standard Error	z Statistic	p Value
Intercept	−3.5756	1.1812	−3.027	0.0025
X-ray	2.0618	0.7777	2.651	0.0080
Stage	1.7556	0.7391	2.375	0.0175
Acid	0.0206	0.0126	1.631	0.1029

TABLE 9.13

Variable	Score χ^2	p Value
Grade	1.065	0.3020
Age	1.5549	0.2124