Please Click https://canvas.ust.hk and SFQ on the left panel To Fill out End-of-Term Course Survey.

Thanks for your attention!

Chapter 9. Logistic Regression

Outline

9.1. The Logit Model

9.1 The Logit Model

Introduction

In our discussion of regression analysis so far the response variable Y has been regarded as a **continuous quantitative variable**. The predictor variables, however, have been both quantitative, as well as qualitative. Indicator variables, which we have described earlier, fall into the second category. There are situations, however, where the **response variable is qualitative**. In this chapter we present methods for dealing with this situation. The methods presented in this chapter are very different from the method of least squares considered in earlier chapters.

Example

Consider a procedure in which individuals are selected on the basis of their scores in a battery of tests. After five years the candidates are classified as "good" or "poor". We are interested in examining the ability of the tests to predict the job performance of the candidates. Here the **response variable**, performance, is **dichotomous**. We can code "good" as 1 and "poor" as 0, for example. The predictor variables are the scores in the tests.

Example

In a study to determine the risk factors for cancer, health records of several people were studied. Data were collected on several variables, such as age, gender, smoking, diet, and the family's medical history. The response variable was the person had cancer ($Y = 1$) or did not have cancer ($Y = 0$).

Example

In the financial community the "health" of a business is of primary concern. The response variable is solvency of the firm (bankrupt = 0, solvent =1), and the predictor variables are the various financial characteristics associated with the firm. Situations where the response variable is a dichotomous variable are quite common and occur extensively in statistical applications.

9.1 The Logit Model

Modeling Qualitative Data

The qualitative data with which we are dealing, the binary response variable, can always be coded as having two values, 0 or 1. Rather than predicting these two values we try to **model the probabilities that the response takes one of these two values**. The limitation of the previously considered standard linear regression model is obvious.

Example

Given a person's height being 168cm (the value of X), it is not appropriate to say this person must be a man (say, $Y = 1$) or must be a woman (say, $Y = 0$). It is more appropriate to predict the gender by With probability (*say*) 0.6 this person is a man, and with probability 0.4 this person is a woman. For this kind of predictions, we are interested in $P(Y = 1|X = x)$, the conditional probability of $Y = 1$ given the value of $X = x$.

Note that by obtaining $P(Y = 1|X = x)$, you can also make a deterministic prediction. For instance, if by your method, your estimate is $P(Y = 1|X = x) = 0.59$, i.e., it is more likely for $Y = 1$ to happen than $Y = 0$ given the value $X = x$. Then, you can just make the prediction $Y = 1$.

9.1 The Logit Model

Modeling Qualitative Data

Why using a linear function for $P(Y = 1|X = x)$ is a bad idea?

We illustrate this point by considering a simple regression problem in which we have only one predictor. The same considerations hold for the multiple regression case. Let π denote the probability that $Y = 1$ when $X = x$. If we use the standard linear model to describe π, then our model for the probability would be

$$\pi = \Pr(Y = 1|X = x) = \beta_0 + \beta_1 x. \qquad (9.1)$$

Since π is a probability it must lie between 0 and 1. The linear function given in (9.1) is unbounded, and hence cannot be used to model probability. There is another reason why ordinary least squares method is unsuitable. The response variable Y is a binomial random variable, consequently its variance will be a function of π, and depends on X. The assumption of equal variance (**homoscedasticity**) does not hold.

9.1 The Logit Model

Example Data

Table 9.1 Financial Ratios of Solvent and Bankrupt Firms

Row	Y	X_1	X_2	X_3	Row	Y	X_1	X_2	X_3
1	0	−62.8	−89.5	1.7	34	1	43.0	16.4	1.3
2	0	3.3	−3.5	1.1	35	1	47.0	16.0	1.9
3	0	−120.8	−103.2	2.5	36	1	−3.3	4.0	2.7
4	0	−18.1	−28.8	1.1	37	1	35.0	20.8	1.9
5	0	−3.8	−50.6	0.9	38	1	46.7	12.6	0.9
6	0	−61.2	−56.2	1.7	39	1	20.8	12.5	2.4
7	0	−20.3	−17.4	1.0	40	1	33.0	23.6	1.5
8	0	−194.5	−25.8	0.5	41	1	26.1	10.4	2.1
9	0	20.8	−4.3	1.0	42	1	68.6	13.8	1.6
10	0	−106.1	−22.9	1.5	43	1	37.3	33.4	3.5
11	0	−39.4	−35.7	1.2	44	1	59.0	23.1	5.5
12	0	−164.1	−17.7	1.3	45	1	49.6	23.8	1.9
13	0	−308.9	−65.8	0.8	46	1	12.5	7.0	1.8
14	0	7.2	−22.6	2.0	47	1	37.3	34.1	1.5
15	0	−118.3	−34.2	1.5	48	1	35.3	4.2	0.9
16	0	−185.9	−280.0	6.7	49	1	49.5	25.1	2.6
17	0	−34.6	−19.4	3.4	50	1	18.1	13.5	4.0
18	0	−27.9	6.3	1.3	51	1	31.4	15.7	1.9
19	0	−48.2	6.8	1.6	52	1	21.5	−14.4	1.0
20	0	−49.2	−17.2	0.3	53	1	8.5	5.8	1.5
21	0	−19.2	−36.7	0.8	54	1	40.6	5.8	1.8
22	0	−18.1	−6.5	0.9	55	1	34.6	26.4	1.8
23	0	−98.0	−20.8	1.7	56	1	19.9	26.7	2.3
24	0	−129.0	−14.2	1.3	57	1	17.4	12.6	1.3
25	0	−4.0	−15.8	2.1	58	1	54.7	14.6	1.7
26	0	−8.7	−36.3	2.8	59	1	53.5	20.6	1.1
27	0	−59.2	−12.8	2.1	60	1	35.9	26.4	2.0
28	0	−13.1	−17.6	0.9	61	1	39.4	30.5	1.9
29	0	−38.0	1.6	1.2	62	1	53.1	7.1	1.9
30	0	−57.9	0.7	0.8	63	1	39.8	13.8	1.2
31	0	−8.8	−9.1	0.9	64	1	59.5	7.0	2.0
32	0	−64.7	−4.0	0.1	65	1	16.3	20.4	1.0
33	0	−11.4	4.8	0.9	66	1	21.7	−7.8	1.6

9.1 The Logit Model

The relationship between the probability π and X can often be represented by a **logistic response function**. It resembles an S-shaped curve, a sketch of which is given in Figure 9.1. The probability π initially increases slowly with increase in X, then the increase accelerates, finally stabilizes, but does not increase beyond 1. Intuitively this makes sense. Consider the probability of a questionnaire being returned as a function of cash reward, or the probability of passing a test as a function of the time put in studying for it.

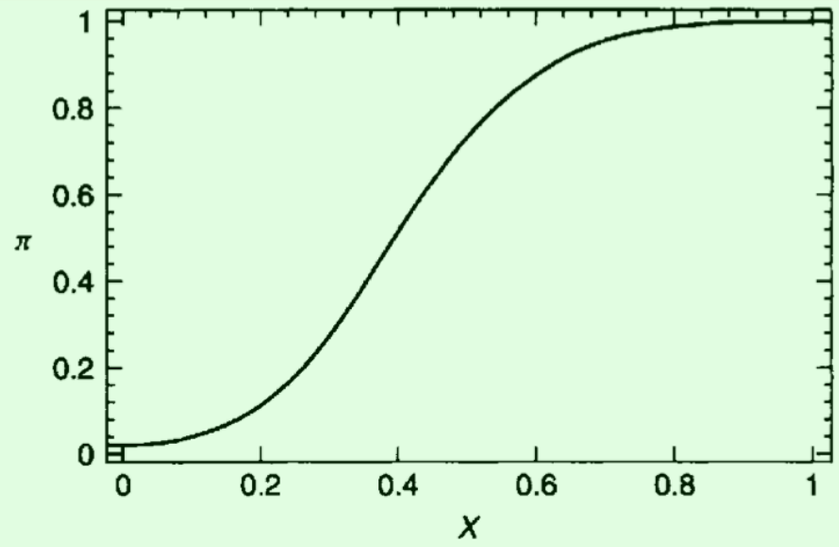


Figure 9.1 Logistic response function.

The shape of the S-curve given in Figure 9.1 can be reproduced if we model the probabilities as follows

$$\pi = \Pr(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}, \qquad (9.2)$$

where e is the base of the natural logarithm.

9.1 The Logit Model

The probabilities here are modeled by the distribution function (**cumulative probability function**) of the **logistic distribution**. There are other ways of modeling the probabilities that would also produce the S-curve. The cumulative distribution of the normal curve has also been used. This gives rise to the **probit model**. We will not discuss the probit model in this course, as we consider the logistic model simpler and superior to the probit model.

The logistic model can be generalized directly to the situation where we have **several predictor variables**. The probability π is modeled as

$$\begin{aligned}\pi &= \Pr(Y=1|X_1=x_1,\cdots,X_p=x_p)\\ &= \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p}}.\end{aligned} \tag{9.3}$$

Equation (9.3) is called the **logistic regression function**. It is nonlinear in the parameters $\beta_0, \beta_1, \cdots, \beta_p$. However, it can be linearized by the **logit transformation**. Instead of working directly with π we work with a transformed value of π. If π is the probability of an event happening, the ratio $\pi/(1-\pi)$ is called the **odds ratio** for the event. Since

$$1-\pi = \Pr(Y=0|X_1=x_1,\cdots,X_p=x_p) = \frac{1}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p}},$$

then

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1x_1+\cdots+\beta_px_p}. \tag{9.4}$$

9.1 The Logit Model

Taking the natural logarithm of both sides of (9.4), we obtain

$$g(x_1, \cdots, x_p) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \qquad (9.5)$$

The logarithm of the odds ratio is called the **logit**. It can be seen from (9.5) that the logit transformation produces a linear function of the parameters $\beta_0, \beta_1, \cdots, \beta_p$. Note also that while the range of values of π in (9.3) is between 0 and 1, the range of values of $\ln[\pi/(1-\pi)]$ is between $-\infty$ and $+\infty$, which makes the logits (the logarithm of the odds ratio) more appropriate for linear regression fitting.

If the **true** parameters are $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)'$, the **probability** of observing the response $y_i \in \{0, 1\}$ given the predictor variables $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{ip})'$ is

$$P(y_i = 0|\mathbf{x}_i) = 1 - P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(\boldsymbol{\beta}'\mathbf{x}_i)}$$

9.1 The Logit Model

Maximum Likelihood Estimator

If the true parameters are $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)'$, the **joint probability function** of observing n observations $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ is

$$\prod_{i=1,\cdots,n: y_i=1} p(y_i = 1|\mathbf{x}_i) \prod_{i=1,\cdots,n: y_i=0} p(y_i = 0|\mathbf{x}_i)$$

which can be conveniently written as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp(y_i \boldsymbol{\beta}' \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}' \mathbf{x}_i)}$$

which is called the **likelihood function**. It represents how likely we can observe the data given that the **true** parameters are $\boldsymbol{\beta}$.

The **maximum likelihood estimator** is obtained by maximizing the likelihood function with respect to $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} \quad \Leftarrow \quad \max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$$

Maximizing $L(\boldsymbol{\beta})$ is equivalent to maximizing $\ell(\boldsymbol{\beta}) = \log\big(L(\boldsymbol{\beta})\big)$, called the **log-likelihood function**. Therefore, the maximum likelihood estimator can also be defined by

$$\hat{\boldsymbol{\beta}} \quad \Leftarrow \quad \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \big(y_i \boldsymbol{\beta}' \mathbf{x}_i - \log(1 + \exp(\boldsymbol{\beta}' \mathbf{x}_i))\big)$$

$\ell(\boldsymbol{\beta})$

9.1 The Logit Model

Maximum Likelihood Estimator

Therefore, the **maximum likelihood estimator** for fitting the model (9.3), called **logistic regression** is defined by

$$\hat{\boldsymbol{\beta}} \quad \Leftarrow \quad \max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i \boldsymbol{\beta}' \mathbf{x}_i - \log(1 + \exp(\boldsymbol{\beta}' \mathbf{x}_i)) \right)$$

Unlike the linear regression where the least squares estimator admits a closed-form solution, there is no closed-form solution for the **maximum likelihood estimator** for logistic regression. The **maximum likelihood estimates** are obtained numerically, using an iterative procedure. We will not go into the computational aspects of the problem.

9.1 The Logit Model

To fit a logistic regression in practice a **computer program** is essential. Most regression packages have a logistic regression option. After the fitting one looks at the same set of questions that are usually considered in linear regression. Questions about the **suitability of the model**, the variables to be **retained**, and **goodness of fit** are all considered. Tools used are not the usual R^2, t-, and F-Tests, the ones employed in least squares regression, but others which provide answers to these same questions. Hypothesis testing is done by different methods, since the method of estimation is **maximum likelihood** as opposed to least squares. Information criteria such as **AIC** and **BIC** can be used for model selection. Instead of SSE, the logarithm of the likelihood for the fitted model is used.

In linear regression, we use residual $e_i = y_i - \hat{y}_i$ or the standardized residual $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}$ and R^2 to quantify the goodness-of-fit of the model to the data. Do we have a counterpart in the context of logistic regression?

The residual e_i represents how the linear model can predict the response y_i.

The counterpart in logistic regression is called the **deviance**.

9.1 The Logit Model

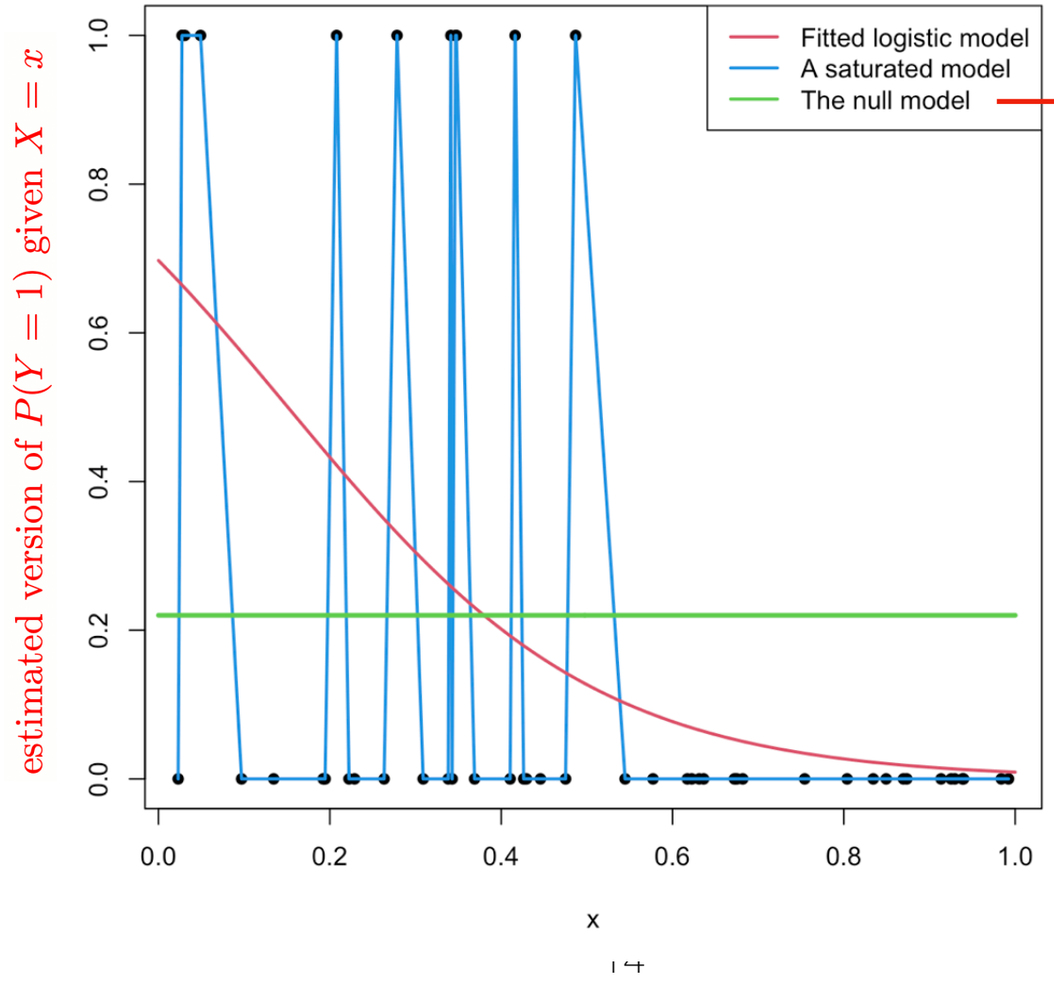
Null deviance and Residual deviance

estimated version of $P(Y = 1|X_1 = x_{i1}, \cdots, X_p = x_{ip})$

The **deviance** is a key concept in logistic regression. Intuitively, it measures the deviance of the fitted logistic model with respect to a perfect model for $P(Y = 1|X_1 = x_1, \cdots, X_p = x_p)$. This perfect model, known as the **saturated model**, denotes an abstract model that fits perfectly the sample, this is, the model such that

$$\hat{P}(Y = 1|X_1 = x_{i1}, \cdots, X_p = x_{ip}) = Y_i, \quad i = 1, \cdots, n$$

This model assigns probability 0 or 1 to Y depending on the actual value of Y_i. To clarify this concept, the following figure shows a saturated model and a fitted logistic regression.



Null model: $P(Y = 1|X_1 = x_1, \cdots, X_p = x_p) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$

black dots denote observations

9.1 The Logit Model

Null deviance and Residual deviance

More precisely, the **residual deviance** is defined as the **difference of likelihood between the fitted model and the saturated model**:

$$D = -2\ell(\hat{\boldsymbol{\beta}})$$

where recall that $\ell(\cdot)$ denotes the **log-likelihood function** and $\hat{\boldsymbol{\beta}}$ denotes the **maximum likelihood estimate**. As a consequence, **the residual deviance is always larger or equal than zero, being zero only if the fit is perfect.**

A benchmark for evaluating the magnitude of the **residual deviance** is the **null deviance**,

$$D_0 = -2\ell(\hat{\beta}_0)$$

which is the **residual deviance of the worst model, the one fitted without any predictor, i.e.** $\beta_1 = \cdots = \beta_p = 0$. The null deviance serves for comparing how much the model has improved by adding the predictors $X_1, \cdots, X_p$. This can be done by means of the R^2 statistic, which is a generalization of the determination coefficient in multiple linear regression:

$$R^2 = 1 - \frac{D}{D_0} = 1 - \frac{\textbf{residual deviance}}{\textbf{null deviance}}$$

In logistic regression, R^2 does not have the same interpretation as in linear regression. It **is not the percentage of variance explained by the logistic model**, but rather a ratio indicating how close is the fit to being perfect or the worst. **It is not related to any correlation coefficient**.

9.1 The Logit Model

G-statistic for model testing

In multiple linear regression, we defined the F-statistic for model testing, i.e., the **full model** versus **reduced model**.

For logistic regression, we have an analogous statistic, called the G-statistic, defined by

$$G = D_0 - D = \textbf{null deviance} - \textbf{residual deviance}$$

for testing the two models:

$$H_0(\textit{null or reduced model}) : P(Y = 1|X_1 = x_1, \cdots, X_p = x_p) = \frac{1}{1 + e^{-\beta_0}} = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

$$H_1(\textit{full model}) : P(Y = 1|X_1 = x_1, \cdots, X_p = x_p) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}}$$

If the **null model** is correct, the G-statistic follows a chi-squared distribution with d.f. $= p$. This is true when the sample size n is large.

9.1 The Logit Model

Distribution of Maximum Likelihood Estimator $\hat{\boldsymbol{\beta}}$

Recall that, in **multiple linear regression**, we have obtained the distribution the **least square estimator (LSE)** $\hat{\boldsymbol{\beta}}$. Basically, the LSE is unbiased meaning that $E[\hat{\beta}_i] = \beta_i$ for all $i = 0, 1, \cdots, p$. Its variance is determined by the matrix $\mathbf{C} = (c_{ij})_{i,j=0}^{p} = (\mathbf{X}'\mathbf{X})^{-1}$, so that, $\text{Var}(\hat{\beta}_i) = \sigma^2 c_{ii}$ and the covariance $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$, where σ^2 is the variance of the random error.

In **logistic regression**, let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimate based on the n observations $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ with binary responses $y_1, \cdots, y_n \in \{0, 1\}$. If the sample size n is large, then

(i) $\hat{\boldsymbol{\beta}}$ is an *approximately* unbiased estimator of $\boldsymbol{\beta}$, i.e., $E(\hat{\beta}_i) \doteq \beta_i$ for $i = 0, 1, \cdots, p$.

(ii) The entries of $\hat{\boldsymbol{\beta}}$ *approximately* follow normal distribution. The **variance** of $\hat{\beta}_i$ and **covariance** between $\hat{\beta}_i$ and $\hat{\beta}_j$ are given by the entries v_{ii} and v_{ij} of a matrix $\mathbf{V} = (v_{ij})_{i,j=0}^{p}$. The matrix $\mathbf{V}$ is given by

$$\mathbf{V} = (v_{ij})_{i,j=0}^{p} = \left(\mathbf{X}' \cdot \begin{bmatrix} \frac{e^{\boldsymbol{\beta}'\mathbf{x}_1}}{(1+e^{\boldsymbol{\beta}'\mathbf{x}_1})^2} & & & \\ & \frac{e^{\boldsymbol{\beta}'\mathbf{x}_2}}{(1+e^{\boldsymbol{\beta}'\mathbf{x}_2})^2} & & \\ & & \ddots & \\ & & & \frac{e^{\boldsymbol{\beta}'\mathbf{x}_n}}{(1+e^{\boldsymbol{\beta}'\mathbf{x}_n})^2} \end{bmatrix} \cdot \mathbf{X} \right)^{-1}$$

where again the $n \times (p+1)$ matrix $\mathbf{X}$ is defined the same as in Chapter 3.

9.1 The Logit Model

Statistical Inference for β

Based on the distribution of $\hat{\boldsymbol{\beta}}$ on last slide, we can conduct statistical inferences for the entries of $\boldsymbol{\beta}$.

After getting the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, plugging it into the formula of $\mathbf{V}$, we can obtain the standard errors of $\hat{\boldsymbol{\beta}}$, i.e.,

$$\text{s.e.}(\hat{\beta}_0),\ \text{s.e.}(\hat{\beta}_1),\ \cdots,\ \text{s.e.}(\hat{\beta}_p)$$

If n is large, for all $i = 0, 1, \cdots, p$, the **Wald statistic**:

$$\frac{\hat{\beta}_i - \beta_i}{\text{s.e.}(\hat{\beta}_i)} \text{ follows approximately the standard normal distribution } N(0,1)$$

To test $\beta_i = 0$, just set β_i to 0 in the Wald statistic

The $100(1-\alpha)\%$ confidence interval for β_i is given by

$$\left[\hat{\beta}_i - z_{\alpha/2} \cdot \text{s.e.}(\hat{\beta}_i),\ \hat{\beta}_i + z_{\alpha/2} \cdot \text{s.e.}(\hat{\beta}_i)\right]$$

9.2. Example: Estimating Probability of Bankruptcies

9.2 Example: Estimating Probability of Bankruptcies

Detecting ailing financial and business establishments is an important function of audit and control. Systematic failure to do audit and control can lead to grave consequences, such as the savings-and-loan fiasco of the 1980s in the United States. Table 9.1 gives some of the operating financial ratios of 33 firms that went bankrupt after 2 years and 33 that remained solvent during the same period.

A multiple logistic regression model is fitted using variables X_1, X_2, and X_3. The output from fitting the model is given in Table 9.2. Three financial ratios were available for each firm:

$$X_1 = \frac{\text{Retained Earnings}}{\text{Total Assets}},$$

$$X_2 = \frac{\text{Earnings Before Interest and Taxes}}{\text{Total Assets}},$$

$$X_3 = \frac{\text{Sales}}{\text{Total Assets}}.$$

The response variable is defined as

$$Y = \begin{cases} 0, & \text{if bankrupt after 2 years,} \\ 1, & \text{if solvent after 2 years.} \end{cases}$$

Table 9.2 Output from the Logistic Regression Using X_1, X_2, and X_3

Variable	Coeff.	s.e.	Z-Test	p-value	Odds Ratio	95% C.I. Lower	95% C.I. Upper
Constant	−10.15	10.84	−0.94	0.35			
X_1	0.33	0.30	1.10	0.27	1.39	0.77	2.51
X_2	0.18	0.11	1.69	0.09	1.20	0.97	1.48
X_3	5.09	5.08	1.00	0.32	161.98	0.01	3.43×10^6
Log-Likelihood = −2.906			$G = 85.683$		df = 3	p-value < 0.000	

Table 9.1

9.2 Example: Estimating Probability of Bankruptcies

Table 9.1 Financial Ratios of Solvent and Bankrupt Firms

Row	Y	X_1	X_2	X_3	Row	Y	X_1	X_2	X_3
1	0	−62.8	−89.5	1.7	34	1	43.0	16.4	1.3
2	0	3.3	−3.5	1.1	35	1	47.0	16.0	1.9
3	0	−120.8	−103.2	2.5	36	1	−3.3	4.0	2.7
4	0	−18.1	−28.8	1.1	37	1	35.0	20.8	1.9
5	0	−3.8	−50.6	0.9	38	1	46.7	12.6	0.9
6	0	−61.2	−56.2	1.7	39	1	20.8	12.5	2.4
7	0	−20.3	−17.4	1.0	40	1	33.0	23.6	1.5
8	0	−194.5	−25.8	0.5	41	1	26.1	10.4	2.1
9	0	20.8	−4.3	1.0	42	1	68.6	13.8	1.6
10	0	−106.1	−22.9	1.5	43	1	37.3	33.4	3.5
11	0	−39.4	−35.7	1.2	44	1	59.0	23.1	5.5
12	0	−164.1	−17.7	1.3	45	1	49.6	23.8	1.9
13	0	−308.9	−65.8	0.8	46	1	12.5	7.0	1.8
14	0	7.2	−22.6	2.0	47	1	37.3	34.1	1.5
15	0	−118.3	−34.2	1.5	48	1	35.3	4.2	0.9
16	0	−185.9	−280.0	6.7	49	1	49.5	25.1	2.6
17	0	−34.6	−19.4	3.4	50	1	18.1	13.5	4.0
18	0	−27.9	6.3	1.3	51	1	31.4	15.7	1.9
19	0	−48.2	6.8	1.6	52	1	21.5	−14.4	1.0
20	0	−49.2	−17.2	0.3	53	1	8.5	5.8	1.5
21	0	−19.2	−36.7	0.8	54	1	40.6	5.8	1.8
22	0	−18.1	−6.5	0.9	55	1	34.6	26.4	1.8
23	0	−98.0	−20.8	1.7	56	1	19.9	26.7	2.3
24	0	−129.0	−14.2	1.3	57	1	17.4	12.6	1.3
25	0	−4.0	−15.8	2.1	58	1	54.7	14.6	1.7
26	0	−8.7	−36.3	2.8	59	1	53.5	20.6	1.1
27	0	−59.2	−12.8	2.1	60	1	35.9	26.4	2.0
28	0	−13.1	−17.6	0.9	61	1	39.4	30.5	1.9
29	0	−38.0	1.6	1.2	62	1	53.1	7.1	1.9
30	0	−57.9	0.7	0.8	63	1	39.8	13.8	1.2
31	0	−8.8	−9.1	0.9	64	1	59.5	7.0	2.0
32	0	−64.7	−4.0	0.1	65	1	16.3	20.4	1.0
33	0	−11.4	4.8	0.9	66	1	21.7	−7.8	1.6

9.2 Example: Estimating Probability of Bankruptcies

Use R for last example

```
################################  Example on Bankruptcy data
Bank_dat<-read.table("data/P339.txt",header=TRUE)  ## read data
mod1 <- glm(Y ~ ., data = Bank_dat, family = "binomial")   ## fit a logistic regression model, here glm stands for "generalized linear model" and binomial fits a logistic link function
summary(mod1)
```

```
> summary(mod1)

Call:
glm(formula = Y ~ ., family = "binomial", data = Bank_dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.64148  -0.00008   0.00000   0.00135   1.41755

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.1535    10.8398  -0.937   0.3489
X1            0.3312     0.3007   1.101   0.2707
X2            0.1809     0.1069   1.692   0.0907 .
X3            5.0875     5.0820   1.001   0.3168
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 91.4954  on 65  degrees of freedom
Residual deviance:  5.8129  on 62  degrees of freedom
AIC: 13.813

Number of Fisher Scoring iterations: 12
```

Recall from Table 9.2 that the Log-Likelihood $\ell(\hat{\boldsymbol{\beta}}) = -2.906$. So, the residual deviance $D = -2\ell(\hat{\boldsymbol{\beta}}) = 5.812$

$$G = 91.4954 - 5.8129 = 85.6825$$

9.2 Example: Estimating Probability of Bankruptcies

Table 9.2 has a certain resemblance to the standard regression output. Some of the output serve similar functions. We now describe and interpret the output obtained from fitting a logistic regression. If π denotes the probability of a firm remaining solvent after 2 years, the fitted logit is given

$$\hat{g}(x_1, \cdots, x_p) = -10.15 + 0.33\, x_1 + 0.18\, x_2 + 5.09\, x_3. \qquad (9.6)$$

This corresponds to the fitted regression equation in standard analysis. Here instead of predicting Y we obtain a model to predict the logits, $\log[\pi/(1-\pi)]$. From the logits, after transformation, we can get the predicted probabilities. The constant and the coefficients are read directly from the second column in the table. The standard errors (s.e.) of the coefficients are given in the third column. The fourth column headed by Z is the ratio of the coefficient and the standard deviation. The Z is sometimes referred to as the **Wald Statistic (Test)**. The Z corresponding to the coefficient of X_2 is obtained from dividing 0.181 by 0.107. In the standard regression this would be the t-Test. This ratio for the logistic regression has a **normal distribution** as opposed to a t-distribution that we get in linear regression. The fifth column gives the p-value corresponding to the observed Z value, and should be interpreted like any p-value (see Chapters 2 and 3). These p-values are used to judge the significance of the coefficient. Values smaller than 0.05 would lead us to conclude that the coefficient is significantly different from 0 at the 5% significance level. From the p-values in Table 9.2, we see that none of the variables individually are significant for predicting the logits of the observations.

Table 9.2

9.2 Example: Estimating Probability of Bankruptcies

In the standard regression output the regression coefficients have a simple interpretation. The regression coefficient of the jth predictor variable X_j is the **expected change** in Y for unit change in X_j when other variables are held **fixed**. The coefficient of X_2 in (9.6) is the expected change in the logit for unit change in X_2 when the other variables are held fixed. The coefficients of a logistic regression fit have another interpretation that is of major practical importance. Keeping X_1 and X_3 fixed, for unit increase in X_2 the relative odds of

$$\frac{\Pr(\text{Firm solvent after 2 years})}{\Pr(\text{Firm bankrupt})}$$

is multiplied by $e^{\hat{\beta}_2} = e^{0.181} = 1.198$, that is, there is an increase of 20%. These values for each of the variables is given in the sixth column headed by Odds Ratio. They represent the change in odds ratio for unit change of a particular variable while the others are held constant. The change in odds ratio for unit change in variable X_j, while the other variables are held **fixed**, is $e^{\hat{\beta}_j}$. If X_j was a binary variable, taking values 1 or 0, then $e^{\hat{\beta}_j}$ would be the actual value of the odds ratio rather than the change in the value of the odds ratio.

Table 9.2

9.2 Example: Estimating Probability of Bankruptcies

The 95% confidence intervals of the odds ratios are given in the last two columns of the table. If the confidence interval does not contain the value 1, the variable has a significant effect on the odds ratio. If the interval is below 1, the variable lowers significantly the relative odds. On the other hand, if the interval lies above 1, the relative odds is significantly increased by the variable.

To see whether the variables **collectively contribute** in explaining the logits a test that examines whether the coefficients $\beta_1, \cdots, \beta_p$ are all zero is performed. This corresponds to the case in multiple regression analysis where we test whether all the regression coefficients can be taken to be zero. The statistic G given at the bottom of Table 9.2 performs that task. The statistic G has a chi-square distribution. The p-value is considerably smaller than 0.05, and indicates that the variables collectively influence the logits.

Table 9.2

9.3. Logistic Regression Diagnostic

9.3 Logistic Regression Diagnostic

After fitting a logistic regression model certain diagnostic measures can be examined for the **detection of outliers**, **high-leverage points**, **influential** observations, and other model deficiencies. The diagnostic measures developed in Chapter 4 for the standard linear regression model can be adapted to the logistic regression model. Regression packages with a logistic regression option usually give various diagnostic measures. These include:

$$\pi_i = P(Y = 1 | X_1 = x_{i1}, \cdots, X_p = x_{ip})$$

1. The estimated probabilities $\hat{\pi}_i$, $i = 1, \cdots, n$.
2. One or more types of residuals, for example, the *standardized deviance residuals*, DR_i, and the *standardized Personian residuals*, PR_i, $i = 1, \cdots, n$.
3. The *weighted leverages*, p^*_{ii}, which measure the potential effects of the observations in the predictor variables on the obtained logistic regression results.
4. The scaled difference in the regression coefficients when the ith observation is deleted: DBETA_i, $i = 1, \cdots, n$.
5. The change in the chi-squared statistics G when the ith observation is deleted: DFG_i, $i = 1, \cdots, n$.

9.3 Logistic Regression Diagnostic

The formulas and derivations of these measures are beyond the scope of this course. The above measures, however, can be used in the same way as the corresponding measures obtained from a linear fit (Chapter 4). For example, the following graphical displays can be examined:

1. The scatter plot of DR_i versus $\hat{\pi}_i$
2. The scatter plot of PR_i versus $\hat{\pi}_i$
3. The index plots of DR_i, $DBETA_i$, DG_i, and p^*_{ii}

As an illustrative example using the Bankruptcy data, the index plots of DR_i, $DBETA_i$, and DG_i obtained from the fitted logistic regression model in (9.6) are shown in Figures 9.2, 9.3, and 9.4, respectively. It can easily be seen from these graphs that observations 9, 14,52, and 53 are **unusual** and that they may have **undue influence** on the logistic regression results.

Figure 9.2, 9.3, 9.4

9.3 Logistic Regression Diagnostic

If DR_i or DG_i or $DBETA_i$ is far way from 0, it means the ith observation deserves more inspections.

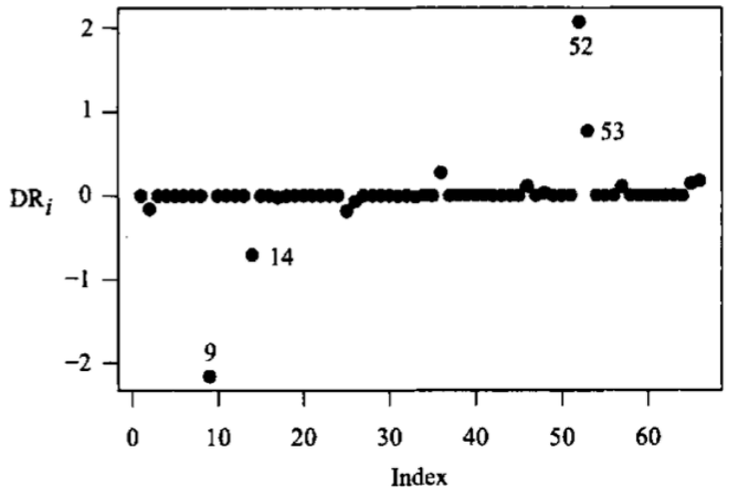


Figure 9.2 Bankruptcy data: Index plot of DR_i, the standardized deviance residuals.

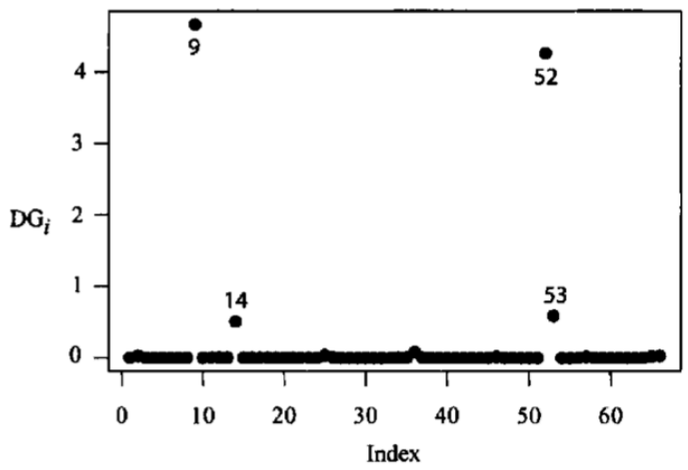


Figure 9.4 Bankruptcy data: Index plot of DG_i, the change in the chi-squared statistics G when the ith observation is deleted.

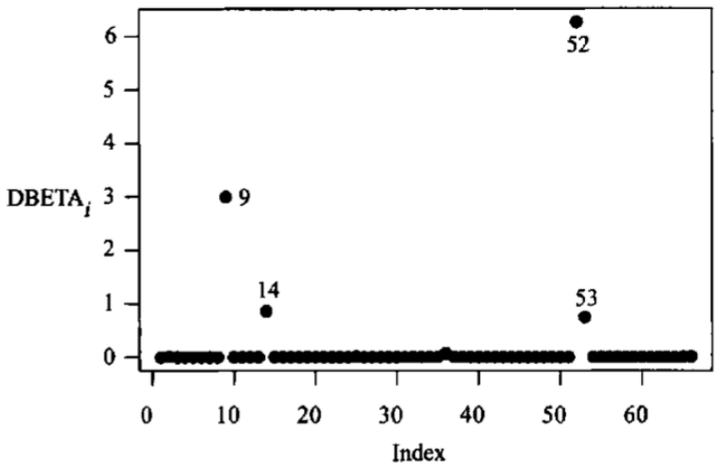


Figure 9.3 Bankruptcy data: Index plot of $DBETA_i$, the scaled difference in the regression coefficients when the ith observation is deleted.

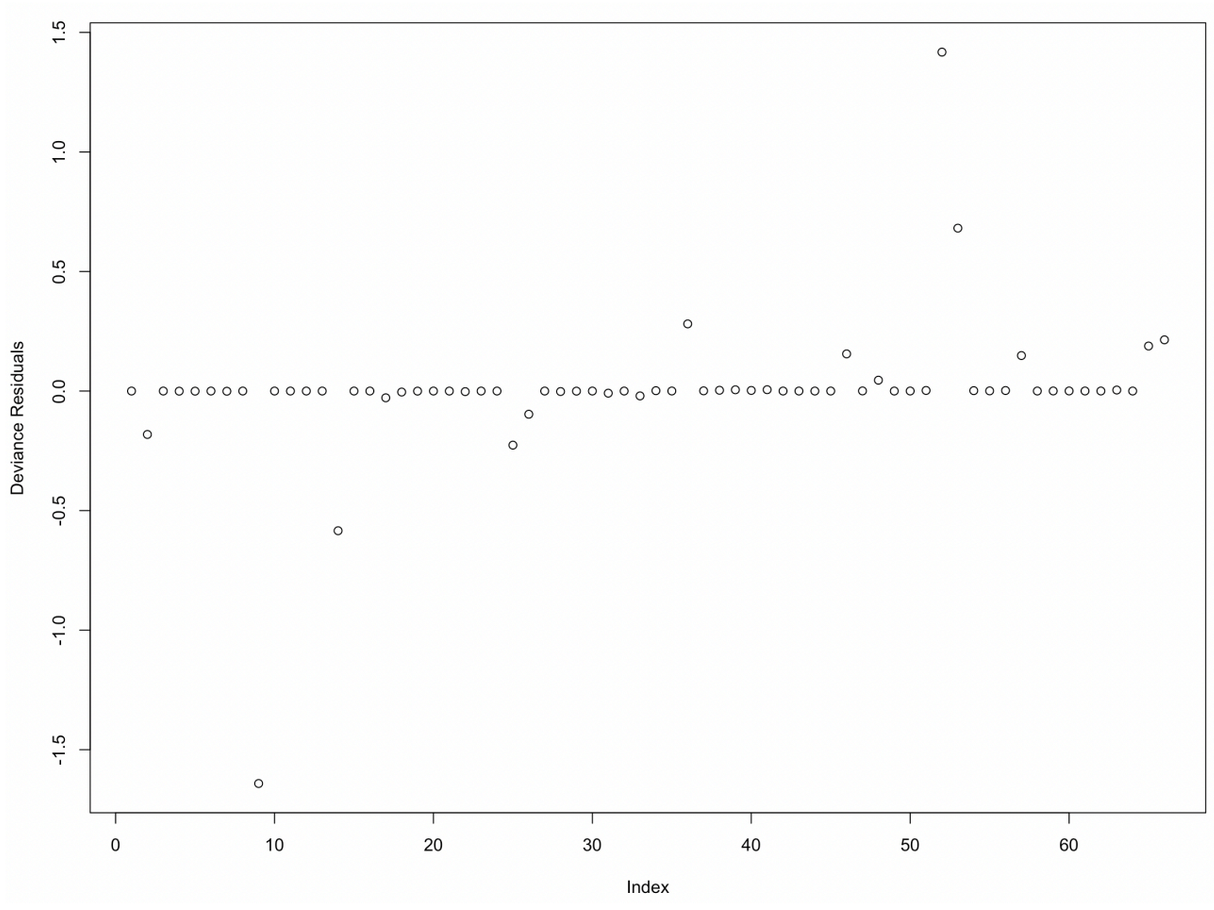
9.3 Logistic Regression Diagnostic

Use R for last example

```
> ############################################## Example on Bankruptcy data
> Bank_dat<-read.table('data/P339.txt',header=TRUE)   ## read data
> mod1<-glm(Y~.,data=Bank_dat,family="binomial")  ## fit a logistic regression model, here glm stands for "generalized linear model" and binomial fits a logistic link function
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> names(summary(mod1))
 [1] "call"           "terms"          "family"         "deviance"       "aic"            "contrasts"      "df.residual"    "null.deviance"  "df.null"        "iter"           "deviance.resid"
[12] "coefficients"   "aliased"        "dispersion"     "df"             "cov.unscaled"   "cov.scaled"
```

It tells what information is stored in the output *mod1*.

```
plot(seq(1,dim(Bank_dat)[1]),summary(mod1)$deviance.resid,xlab="Index",ylab="Deviance Residuals")
```



9.4. Determination of Variables to Retain

9.4 Determination of Variables to Retain

In the analysis of the Bankruptcy data we have determined so far that the variables X_1, X_2, and X_3 **collectively** have explanatory power. Do we need all three variables? This is analogous to the problem of variable selection in multiple regression that was discussed in Chapter 8. Instead of looking at the reduction in the error sum of squares we look at the **change in the likelihood** (more precisely, the logarithm of the likelihood) for the two fitted models. The reason for this is that in logistic regression the fitting criterion is the maximum likelihood, whereas in least squares it is the least sum of squares. Let $L(p)$ denote the logarithm of the likelihood when we have a model with p variables and a constant. Similarly, let $L(p+q)$ be the logarithm of the likelihood for a model in which we have $p+q$ variables and a constant. To see whether the q additional variables contribute significantly we look at $2[L(p+q)-L(p)]$. This quantity is twice the difference between the log-likelihood for the two models. This difference is distributed as a **chi-square variable** with q degrees of freedom.

$$H_0(\textit{reduced model}) : P(Y=1|X_1=x_1,\cdots,X_p=x_p)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\cdots+\beta_px_p)}}$$

Fit the MLE and get the log-likelihood $L(p)$

$$H_1(\textit{full model}) : P(Y=1|X_1=x_1,\cdots,X_{p+q}=x_{p+q})=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\cdots+\beta_{p+q}x_{p+q})}}$$

Fit the MLE and get the log-likelihood $L(p+q)$

9.4 Determination of Variables to Retain

If n is large and the **reduced model** is correct, we have $2[L(p+q) - L(p)]$ follows approximately a chi-squared distribution with a d.f. $= q$.

By definition, we see that $-2L(p)$ is the **residual deviance** of the reduced model, and $-2L(p+q)$ is the **residual deviance** of the full model. Therefore, the test statistic is essentially to calculate the difference of residual deviances.

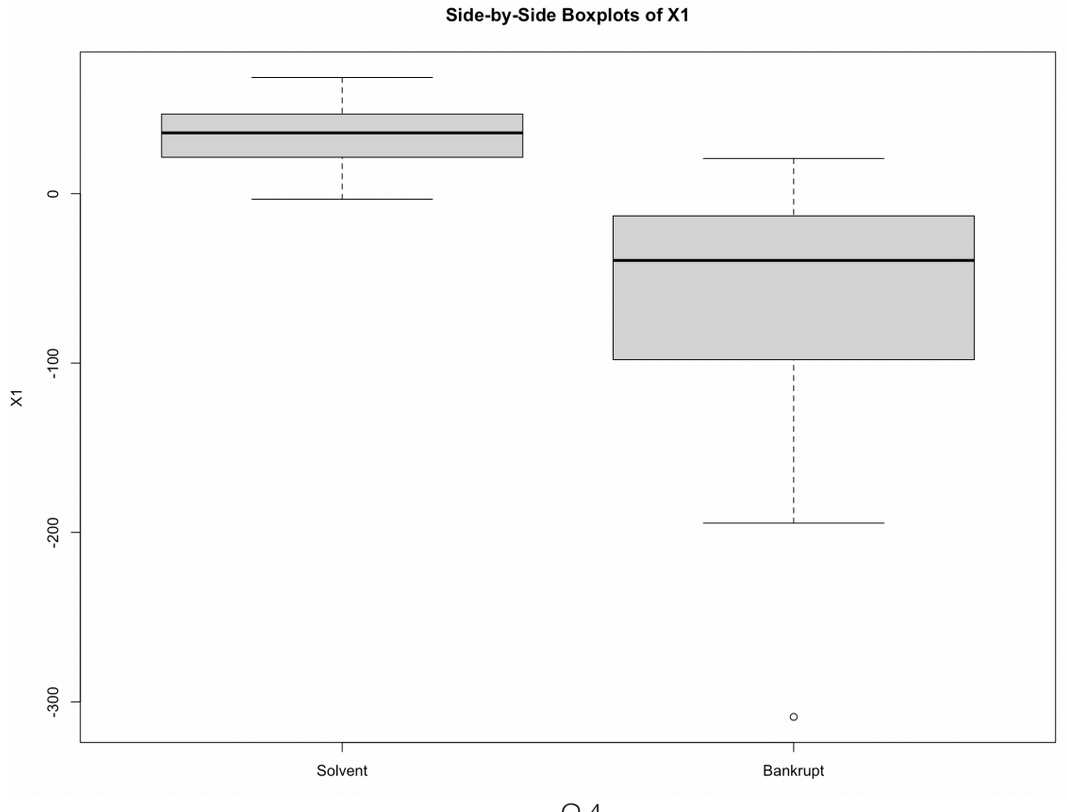
The magnitude of this quantity determines the **significance of the test**. A small value of chi-square would lead to the conclusion that the q variables do not add significantly to the improvement in prediction of the logits, and is therefore not necessary in the model. A large value of chi-square would call for the retention of the q variables in the model. The **critical value** is determined by the significance level of the test. This test procedure is valid when n, the number of observations available for fitting the model, is **large**.

9.4 Determination of Variables to Retain

An idea of the predictive power of a variable for possible inclusion in the logistic model can be obtained from a simple graphical plot. **Side-by-side boxplots are constructed for each of the explanatory variable**. Side-by-side boxplots will indicate the variables that may be useful for this purpose. Variables with boxplots different for the two groups are likely candidates. Note that this does not take into account the correlation between the variables. The formal procedure described above takes into account the correlations. With a large number of explanatory variables the boxplots provide a quick screening procedure.

Example using Bankruptcy data

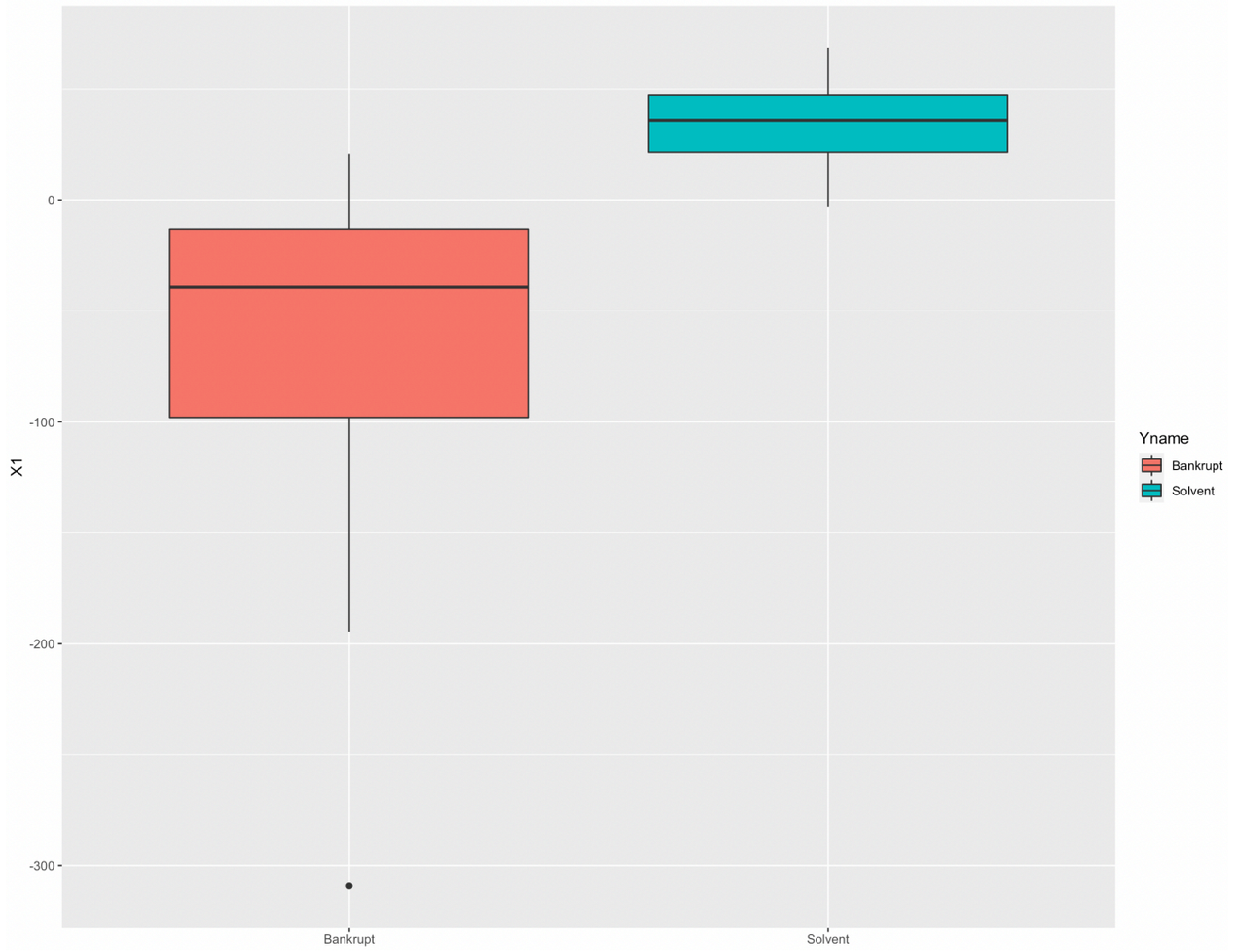
```
################################################## Side-by-side boxplots
Bank_dat<-read.table('data/P339.txt',header=TRUE)   ## read data
## for variable X1
boxplot(Bank_dat$X1[which(Bank_dat$Y==1)],Bank_dat$X1[which(Bank_dat$Y==0)],names=c('Solvent','Bankrupt'),main="Side-by-Side Boxplots of X1",ylab="X1")
```



9.4 Determination of Variables to Retain

Using ggpot2 library to draw fancy Boxplots

```
###### using ggplot 2 to draw more beautiful boxplots
library(ggplot2)
Bank_dat$Yname<-rep('Solvent',66)                    ## assign the names to 0 and 1
Bank_dat$Yname[which(Bank_dat$Y==0)]='Bankrupt'
ggplot(Bank_dat, aes(x=Yname, y=X1,fill=Yname)) + geom_boxplot()
```



9.4 Determination of Variables to Retain

In the Bankruptcy data we are analyzing, let us see if the variable X_3 can be deleted without degrading the model. We want to answer the question: Should the variable X_3 be retained in the model? We fit a logistic regression using X_1 and X_2 . The results are given in Table 9.3. The log-likelihood for the model with X_1, X_2, and X_3 is -2.906, whereas with only X_1 and X_2 it is -4.736. Here $p = 2$ and $q = 1$, and $2[L(3) - L(2)] = 3.66$. This is a chi-square variable with 1 degree of freedom. From the Chi-squared table, we find that the 5% critical value of the chi-square distribution with 1 degree of freedom is 3.84. At the 5% level we can conclude that the variable X_3 can be deleted without affecting the effectiveness of the model.

Table 9.3 Output From the Logistic Regression Using X_1 and X_2

Variable	Coefficient	s.e.	Z-Test	p-value	Odds Ratio	95% C.I. Lower	95% C.I. Upper
Constant	−0.550	0.951	−0.58	0.563			
X_1	0.157	0.075	2.10	0.036	1.17	1.01	1.36
X_2	0.195	0.122	1.59	0.112	1.21	0.96	1.54
Log-Likelihood = −4.736		G = 82.024		df = 2	p-value < 0.000		

9.4 Determination of Variables to Retain

Let us now see if we can delete X_2. The result of regressing Y on X_1 is given in Table 9.4. The resulting log-likelihood is -7.902. The test statistic, which we have described earlier, has a value of 6.332. This is distributed as a chi-square random variable with 1 degree of freedom. The 5% value, as we saw earlier, was 3.84. The analysis indicates that we should not delete X_2 from our model. The p-value for this test, as can be verified, is 0.019. To predict probabilities of bankruptcies of firms in our data we should include both X_1 and X_2 in our model.

Table 9.4 Output from the Logistic Regression Using X_1

Variable	Coefficient	s.e.	Z-Test	p-value	Odds Ratio	95% C.I. Lower	95% C.I. Upper
Constant	−1.167	0.816	−1.43	0.153			
X_1	0.177	0.057	3.09	0.002	1.19	1.07	1.33
Log-Likelihood = −7.902		$G = 75.692$		df = 1	p-value < 0.000		

The procedure that we have outlined above enables us to test any **nested model**. A set of models are said to be nested if they can be obtained from a larger model as special cases. The methodology is similar to that used in analyzing nested models in multiple regression. The only difference is that here our test statistic is based on the **log of the likelihood** instead of **sum of squares**.

9.4 Determination of Variables to Retain

The AIC and BIC criteria discussed in Chapter 8 can be used to **judge the suitability of various logistic models**, and thereby the desirability of retaining a variable in the model. In the context of p-term logistic regression, AIC and BIC are

$$\text{AIC} = -2(\text{Log-Likelihood of the Fitted Model}) + 2p, \tag{9.7}$$

$$\text{BIC} = -2(\text{Log-Likelihood of the Fitted Model}) + p\log n, \tag{9.8}$$

where p denotes the number of variables in the model. Table 9.5 shows AIC and BIC for all possible models. The best AIC model is the one that includes all three variables (lowest AIC). While BIC picks X_1, X_2 as the best model, but the one containing all three variables is equally adequate. The BIC for the two top models differ by less than 2.

Table 9.5 The AIC and BIC Criteria for Various Logistic Regression models

Variables	AIC	BIC
$X_1X_2X_3$	13.81	22.57
X_1X_2	15.47	22.04
X_1X_3	18.12	24.69
X_2X_3	33.40	39.97
X_1	19.80	24.18
X_2	34.50	38.88
X_3	92.46	96.84
None	93.50	95.69

9.5. Judging the Fit of a Logistic Regression

9.5 Judging the Fit of a Logistic Regression

The overall fit of a multiple regression model is judged, for example, by the value of R^2 from the fitted model. **No such simple satisfactory measure exists** for logistic regression. Some ad hoc measures have been proposed which are based on the ratio of likelihoods. Most of these are functions of the ratio of the likelihood for the model and the likelihood of the data under a binomial model. These measures are not particularly informative and we will consider a different approach.

The logistic regression equation attempts to model probabilities for the two values of Y (0 or 1). To judge how well the model is doing we will determine the number of observations in the sample that the model is **classifying correctly**. Our approach will be to fit the logistic model to the data, and calculate the fitted logits. From the fitted logits we will calculate the fitted probabilities for each observation. If the fitted probability for an observation is greater than 0.5, we will assign it to Group 1 ($Y = 1$), and if less than 0.5 we will classify it in Group 0 ($Y = 0$). We will then determine what proportion of the data is classified correctly. **A high proportion of correct classification will indicate to us that the logistic model is working well**. A low proportion of correct classification will indicate poor performance.

Different cutoff values, other than 0.5, have been suggested in the literature. In most practical situations, without any auxiliary information, such as the relative cost of misclassification or the relative frequency of the two categories in the population, 0.5 is recommended as a cutoff value.

9.5 Judging the Fit of a Logistic Regression

A slightly more problematical question is how high the correct classification probability has to be before logistic regression is thought to be effective. Suppose that in a sample of size n there are n_1 observations from Group 1, and n_2 from Group 2. I f we classify all the observations into one group or the other, then we will get either n_1/n or n_2/n proportions of observations classified correctly. **As a base level for correct classification we can take the** $\max(n_1/n, n_2/n)$. The proportion of observation classified correctly by the logistic regression should be much higher than the base level for the logistic model to be deemed useful.

For the Bankruptcy data that we have been analyzing logistic regression performs very well. Using variables X_1 and X_2, we find that the model misclassifies one observation from the solvent group (observation number 36) and one observation from the bankruptcy group (observation number 9). The overall correct classification rate $(64/66) = 0.97$. This is considerably higher than the base level rate of 0.5.

9.5 Judging the Fit of a Logistic Regression

The concept of overall correct classification for the observed sample to judge the adequacy of the logistic model that we have discussed has been generalized. This generalization is used to produce a statistic to judge the fit of the logistic model. It is sometimes called the **Concordance Index** and is denoted by C. This statistic is calculated by considering all possible pairs formed by taking one observation from each group. **Each of the pairs** is then classified by using the fitted model. The Concordance Index is the percent of all possible pairs that is classified correctly. Thus, C lies between 0.5 and 1. Values of C close to 0.5 shows the logistic model performing poorly (no better than guessing). The value of C for the logistic model with X_1, X_2, X_3 is 0.99. Several currently available software computes the value of C.

n_1 observations with $Y_i = 0$
n_2 observations with $Y_i = 1$ → In total, there are $n_1 n_2$ pairs.

For every pair, say $(\mathbf{x}_1, y_1 = 0)$ and $(\mathbf{x}_2, y_2 = 1)$, let $\hat{\pi}_1 = P(Y = 1|\mathbf{X} = \mathbf{x}_1)$ and $\hat{\pi}_2 = P(Y = 1|\mathbf{X} = \mathbf{x}_2)$ be the **estimated probability by the logistic regression output**. If $\hat{\pi}_2 > \hat{\pi}_1$, we say this pair is **correctly classified**.

$$\textbf{Concordance index} = \frac{\textbf{\# of pairs correctly classified}}{n_1 n_2}$$

9.5 Judging the Fit of a Logistic Regression

The observed correct classification rate should be treated with caution. In practice, if this logistic regression was applied to a new set of observations from this population, it would be very unlikely to do as well. The classification probability has an **upward** . The bias arises due to the fact that the same data that were used to fit the model was used to judge the performance of the model. The model fitted to a given body of data is expected to perform well on the same body of data. The true measure of the performance of the logistic regression model for classification is the probability of classifying a **future observation** correctly and not a sample observation. This upward bias in the estimate of correct classification probability can be reduced by using resampling methods, such as **jack-knife** or **bootstrap**. These will not be discussed in this course.

9.6. The Multinomial Logit Model

9.6 The Multinomial Logit Model

Multinomial Logistic Regression

In our discussion of logistic regression we have so far assumed that the qualitative response variable assumes only **two** values, generically, 1 for success and 0 for failure. The logistic regression model can be extended to situations where the response variable assumes more than two values. In a study of the choice of mode of transportation to work, the response variable may be private automobile, car pool, public transport, bicycle, or walking. The response falls into **five categories**. There is no natural ordering of the categories. We might want to analyze how the choice is related to factors such as age, gender, income, distance traveled, and so forth. The resulting model can be analyzed by using slightly modified methods that were used in analyzing the dichotomous outcomes. This method is called the **multinomial (polytomous) logistic regression**.

We have n independent observations with p explanatory variables. The qualitative response variable has k categories. To construct the logits in the multinomial case one of the categories is considered the base level and all the logits are constructed relative to it. Any category can be taken as the **base level**. We will take category k as the base level in our description of the method. Since there is no ordering, it is apparent that any category may be labeled k. Let π_j denote the multinomial probability of an observation falling in the jth category. We want to find the rela- tionship between this probability and the p explanatory variables, $X_1, X_2, \cdots, X_p$.

The multiple logistic regression model then is

$$\ln\left(\frac{\pi_j(x_i)}{\pi_k(x_i)}\right) = \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi}; \qquad \begin{matrix} j = 1, 2, \cdots, (k-1), \\ i = 1, 2, \cdots, n. \end{matrix}$$

9.6 The Multinomial Logit Model
Multinomial Logistic Regression

For Example

If $k = 4$, i.e, there are four categories, we let the category 4 be the **base level**.

Given an observation $\mathbf{x}_i = (x_{1i}, \cdots, x_{pi})'$, we denote

$$\pi_1(\mathbf{x}_i) = P(Y = 1|\mathbf{X} = \mathbf{x}_i), \quad \pi_2(\mathbf{x}_i) = P(Y = 2|\mathbf{X} = \mathbf{x}_i)$$
$$\pi_3(\mathbf{x}_i) = P(Y = 3|\mathbf{X} = \mathbf{x}_i), \quad \pi_4(\mathbf{x}_i) = P(Y = 4|\mathbf{X} = \mathbf{x}_i)$$

The multiple logistic regression model assumes

$$\ln\left(\frac{\pi_1(\mathbf{x}_i)}{\pi_4(\mathbf{x}_i)}\right) = \beta_{01} + \beta_{11}x_{1i} + \beta_{21}x_{2i} + \cdots + \beta_{p1}x_{pi}$$
$$\ln\left(\frac{\pi_2(\mathbf{x}_i)}{\pi_4(\mathbf{x}_i)}\right) = \beta_{02} + \beta_{12}x_{1i} + \beta_{22}x_{2i} + \cdots + \beta_{p2}x_{pi}$$
$$\ln\left(\frac{\pi_3(\mathbf{x}_i)}{\pi_4(\mathbf{x}_i)}\right) = \beta_{03} + \beta_{13}x_{1i} + \beta_{23}x_{2i} + \cdots + \beta_{p3}x_{pi}$$

9.6 The Multinomial Logit Model
Multinomial Logistic Regression

Since all the π's add to unity, this reduces to

$$\ln\left(\pi_j(x_i)\right) = \frac{\exp\left(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi}\right)}{1 + \sum_{j=1}^{k-1} \exp\left(\beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi}\right)},$$

for $j = 1, 2, \cdots, (k-1)$. The model parameters are estimated by the method of maximum likelihood. Statistical software is available to do this fitting. We illustrate the method by an example.

Example

9.6 The Multinomial Logit Model

Example: Determining Chemical Diabetes

To determine the treatment and management of diabetes it is necessary to determine whether the patient has chemical diabetes or overt diabetes. The data presented in Tables 9.6 and 9.7 is from a study conducted to determine the nature of chemical diabetes. The measurements were taken on 145 nonobese volunteers who were subjected to the same regimen. Many variables were measured, but we consider only three of them. These are, insulin response (IR), the steady-state plasma glucose (SSPG), which measures insulin resistance, and relative weight (RW). The diabetic status of each subject was recorded. The clinical classification (eC) categories were overt diabetes (1), chemical diabetes (2), and normal (3).

Side-by-side boxplots of the explanatory variables indicate that the distribution of IR and SSPG differ for the three categories. The distribution of RW on the other hand does not differ substantially for the three categories. The boxplots are shown in Figure 9.5.

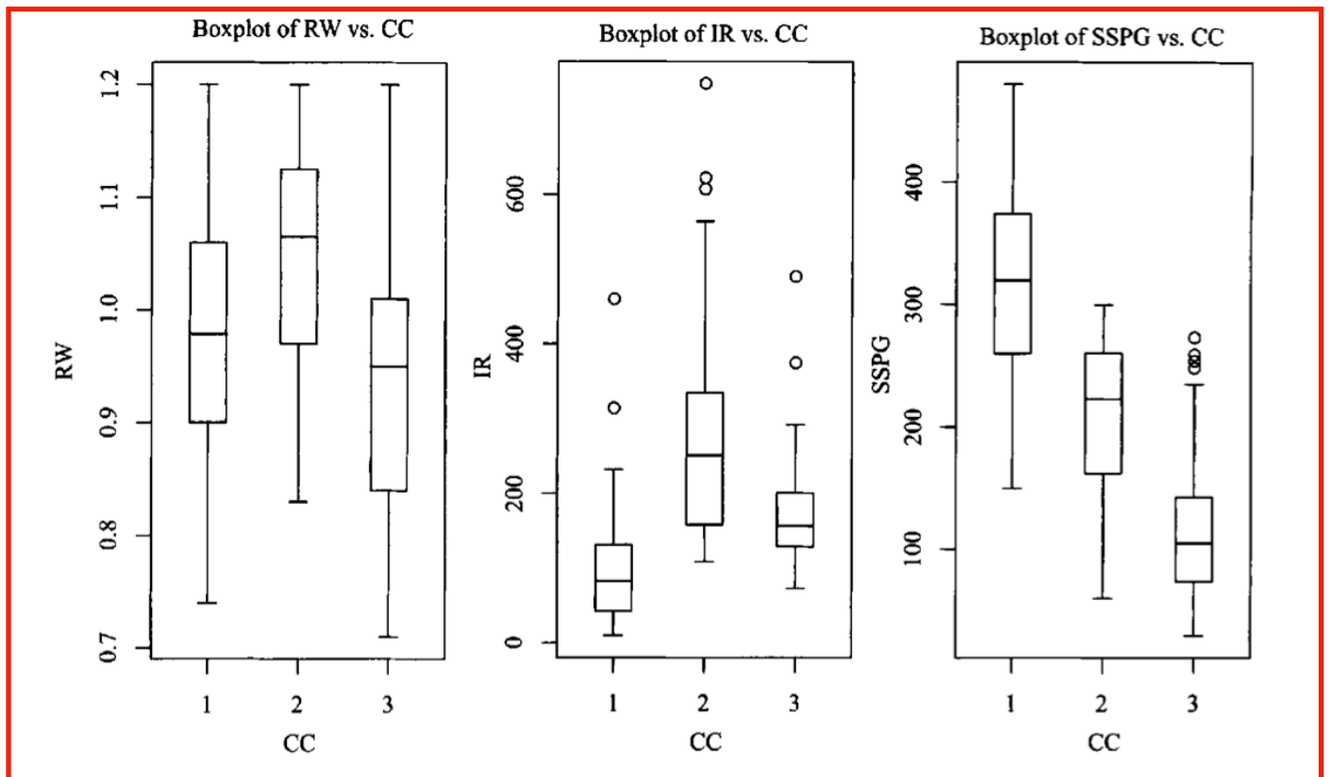


Figure 9.5 Side-by-side boxplots for the Diabetes data.

Table 9.6, 9.7

9.6 The Multinomial Logit Model

Example: Determining Chemical Diabetes

Table 9.6 Diabetes Data: Blood Glucose, Insulin Levels, Relative Weight, Clinical Classification (Patients 1 to 90)

Patient	RW	IR	SSPG	CC	Patient	RW	IR	SSPG	CC
1	0.81	124	55	3	46	0.91	106	56	3
2	0.95	117	76	3	47	0.95	118	122	3
3	0.94	143	105	3	48	0.95	112	73	3
4	1.04	199	108	3	49	1.03	157	122	3
5	1.00	240	143	3	50	0.87	292	128	3
6	0.76	157	165	3	51	0.87	200	233	3
7	0.91	221	119	3	52	1.17	220	132	3
8	1.10	186	105	3	53	0.83	144	138	3
9	0.99	142	98	3	54	0.82	109	83	3
10	0.78	131	94	3	55	0.86	151	109	3
11	0.90	221	53	3	56	1.01	158	96	3
12	0.73	178	66	3	57	0.88	73	52	3
13	0.96	136	142	3	58	0.75	81	42	3
14	0.84	200	93	3	59	0.99	151	122	2
15	0.74	208	68	3	60	1.12	122	176	3
16	0.98	202	102	3	61	1.09	117	118	3
17	1.10	152	76	3	62	1.02	208	244	2
18	0.85	185	37	3	63	1.19	201	194	2
19	0.83	116	60	3	64	1.06	131	136	3
20	0.93	123	50	3	65	1.20	162	257	2
21	0.95	136	47	3	66	1.05	148	167	2
22	0.74	134	50	3	67	1.18	130	153	3
23	0.95	184	91	3	68	1.01	137	248	3
24	0.97	192	124	3	69	0.91	375	273	3
25	0.72	279	74	3	70	0.81	146	80	3
26	1.11	228	235	3	71	1.10	344	270	2
27	1.20	145	158	3	72	1.03	192	180	3
28	1.13	172	140	3	73	0.97	115	85	3
29	1.00	179	145	3	74	0.96	195	106	3
30	0.78	222	99	3	75	1.10	267	254	3
31	1.00	134	90	3	76	1.07	281	119	3
32	1.00	143	105	3	77	1.08	213	177	2
33	0.71	169	32	3	78	0.95	156	159	3
34	0.76	263	165	3	79	0.74	221	103	3
35	0.89	174	78	3	80	0.84	199	59	3
36	0.88	134	80	3	81	0.89	76	108	3
37	1.17	182	54	3	82	1.11	490	259	3
38	0.85	241	175	3	83	1.19	143	204	2
39	0.97	128	80	3	84	1.18	73	220	3
40	1.00	222	186	3	85	1.06	237	111	2
41	1.00	165	117	3	86	0.95	748	122	2
42	0.89	282	160	3	87	1.06	320	253	2
43	0.98	94	71	3	88	0.98	188	211	2
44	0.78	121	29	3	89	1.16	607	271	2
45	0.74	73	42	3	90	1.18	297	220	2

Table 9.7 Diabetes Data: Blood Glucose, Insulin Levels, Relative Weight, Clinical Classification (Patients 91 to 145)

Patient	RW	IR	SSPG	CC	Patient	RW	IR	SSPG	CC
91	1.20	232	276	2	119	1.06	76	260	1
92	1.08	480	233	2	120	0.92	42	346	1
93	0.91	622	264	2	121	1.20	102	319	1
94	1.03	287	231	2	122	1.04	138	351	1
95	1.09	266	268	2	123	1.16	160	357	1
96	1.05	124	60	2	124	1.08	131	248	1
97	1.20	297	272	2	125	0.95	145	324	1
98	1.05	326	235	2	126	0.86	45	300	1
99	1.10	564	206	2	127	0.90	118	300	1
100	1.12	408	300	2	128	0.97	159	310	1
101	0.96	325	286	2	129	1.16	73	458	1
102	1.13	433	226	2	130	1.12	103	339	1
103	1.07	180	239	2	131	1.07	460	320	1
104	1.10	392	242	2	132	0.93	42	297	1
105	0.94	109	157	2	133	0.85	13	303	1
106	1.12	313	267	2	134	0.81	130	152	1
107	0.88	132	155	2	135	0.98	44	167	1
108	0.93	285	194	2	136	1.01	314	220	1
109	1.16	139	198	2	137	1.19	219	209	1
110	0.94	212	156	2	138	1.04	100	351	1
111	0.91	155	100	2	139	1.06	10	450	1
112	0.83	120	135	2	140	1.03	83	413	1
113	0.92	28	455	1	141	1.05	41	480	1
114	0.86	23	327	1	142	0.91	77	150	1
115	0.85	232	279	1	143	0.90	29	209	1
116	0.83	54	382	1	144	1.11	124	442	1
117	0.85	81	378	1	145	0.74	15	253	1
118	1.06	87	374	1					

9.6 The Multinomial Logit Model

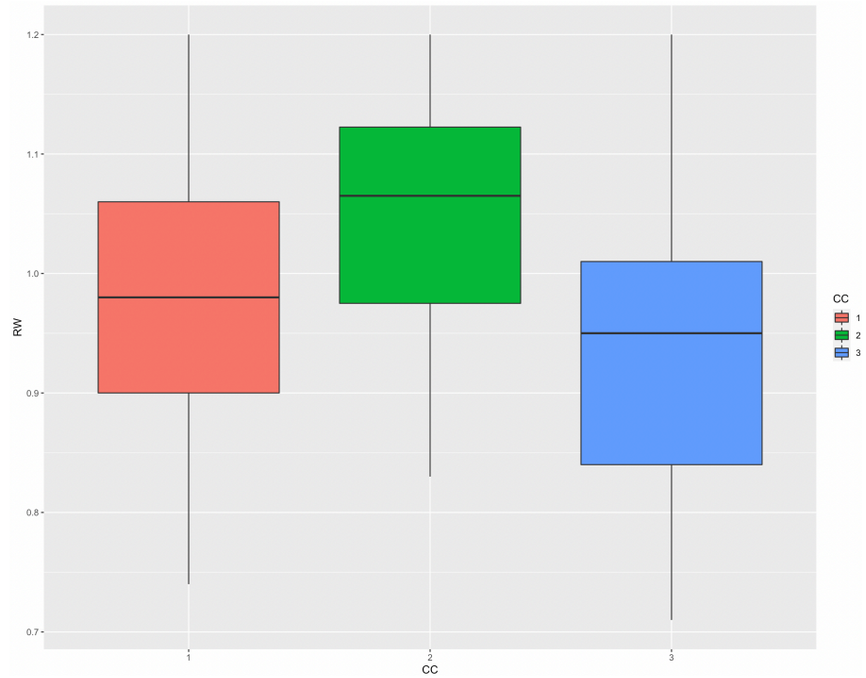
Example: Determining Chemical Diabetes

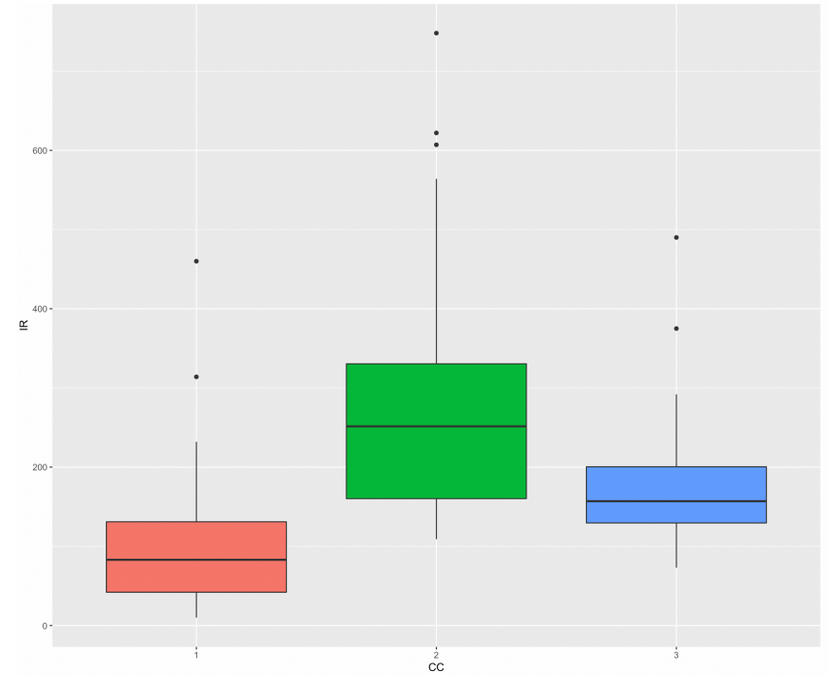
Using ggpot2 library to draw fancy Boxplots

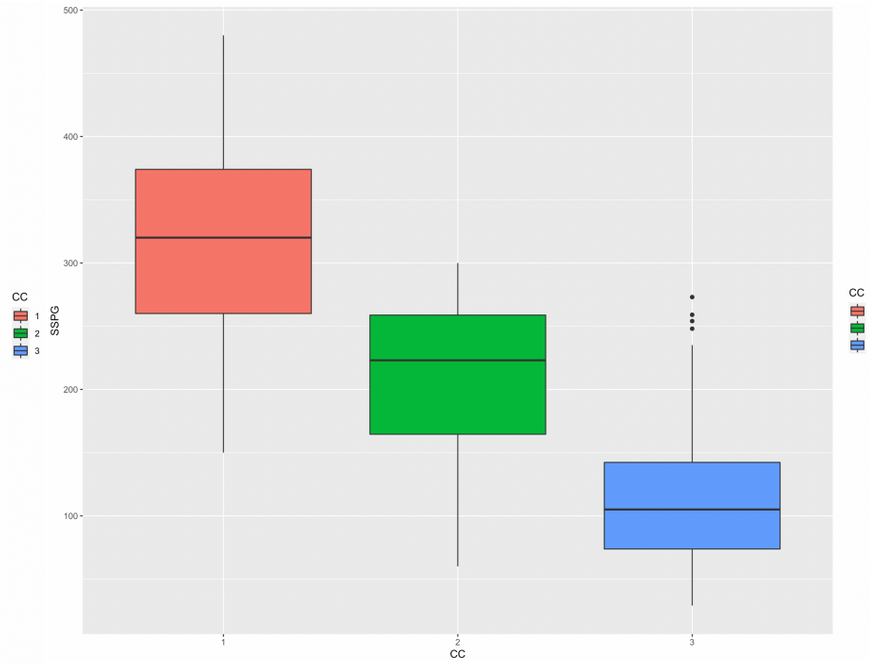
```
################################################   Examples using Diabetes data
diabetes<-read.table('data/P349-50.txt',header=TRUE)   ## read data
diabetes$CC<-as.factor(diabetes$CC)   ## assign the variable to a factor class so that R treats it as categorical
ggplot(diabetes,aes(x=CC,y=RW,fill=CC))+geom_boxplot()    ## side-by-side boxplot for RW

ggplot(diabetes,aes(x=CC,y=IR,fill=CC))+geom_boxplot()    ## side-by-side boxplot for IR

ggplot(diabetes,aes(x=CC,y=SSPG,fill=CC))+geom_boxplot()    ## side-by-side boxplot for SSPG
```







9.6 The Multinomial Logit Model

Example: Determining Chemical Diabetes

The results of fitting a multinomial logistic model using the variables IR, SSPG, and RW are given in Table 9.8. Each of the logistic models is given relative to normal patients. We see that RW has insignificant values in each of the logit models. This is consistent with what we observed in the side-by-side boxplots. We now fit the multinomiallogistic model with two variables, SSPG and IR. The results are given in Table 9.9.

Table 9.8 Multinomial Logistic Regression Output with RW, SSPG, and IR (Base Level =3)

Variable	Coefficient	s.e.	Z-Test	p-value	Odds Ratio	95% C.I. Lower	95% C.I. Upper
Logit 1: (2/3)							
Constant	–7.615	2.336	–3.26	0.001			
RW	3.473	2.446	1.42	0.156	32.23	0.27	3894.2
SSPG	0.016	0.005	3.29	0.001	1.02	1.01	1.03
IR	0.004	0.002	1.53	0.127	1.00	1.00	1.01
Logit 2: (1/3)							
Constant	–1.845	3.463	–0.53	0.594			
RW	–5.868	3.867	–1.52	0.129	0.00	0.00	5.53
SSPG	0.046	0.009	4.92	0.000	1.05	1.03	1.07
IR	–0.0134	0.005	–2.66	0.008	0.99	0.98	1.00
Log-Likelihood = –68.415		G = 159.369		df = 6	p-value < 0.000		

Table 9.9

9.6 The Multinomial Logit Model

Example: Determining Chemical Diabetes

Using R

```
################################################   Multinomial logistic regression on Diabetes data
diabetes<-read.table('data/P349-50.txt',header=TRUE)   ## read data
diabetes$CC<-as.factor(diabetes$CC)   ## assign the variable to a factor class so that R treats it as categorical
diabetes$CC<-relevel(diabetes$CC,ref=3)  ## set the class 3 as the base class
library(nnet)      ## requires the library nnet for the multinom function
mod1<-multinom(CC~.-Patient, data=diabetes)      ## note that the patient column is the patient ID which should be deleted from the model
summary(mod1)
```

```
> mod1<-multinom(CC~.-Patient, data=diabetes)      ## note that the patient column is the patient ID which should be deleted from the model
# weights:  15 (8 variable)
initial  value 159.298782
iter  10 value 69.027793
iter  20 value 68.418245
iter  30 value 68.414665
final  value 68.414644
converged
> summary(mod1)
Call:
multinom(formula = CC ~ . - Patient, data = diabetes)

Coefficients:
  (Intercept)        RW           IR       SSPG
1   -1.845230 -5.867196 -0.013353688 0.04550552
2   -7.615261  3.472572  0.003586749 0.01641449

Std. Errors:
  (Intercept)       RW          IR        SSPG
1    3.463507 3.866580 0.005019289 0.009241721
2    2.335615 2.446151 0.002349168 0.004981886

Residual Deviance: 136.8293
AIC: 152.8293
```

9.6 The Multinomial Logit Model

Example: Determining Chemical Diabetes

Table 9.9 Multinomial Logistic Regression Output with SSPG and IR (Base Level = 3)

Variable	Coefficient	s.e.	Z-Test	p-value	Odds Ratio	95% C.I. Lower	95% C.I. Upper
Logit 1: (2/3)							
Constant	–4.549	0.771	–5.90	0.000			
SSPG	0.020	0.004	4.38	0.000	1.02	1.01	1.03
IR	0.003	0.002	1.42	0.155	1.00	1.00	1.01
Logit 2: (1/3)							
Constant	–7.111	1.688	–4.21	0.000			
SSPG	0.043	0.008	5.34	0.000	1.04	1.03	1.06
IR	–0.013	0.005	–2.89	0.004	0.99	0.98	1.00
Log-Likelihood = −72.029		$G = 152.141$		df = 4	p-value < 0.000		

Looking at Logit(1/3), we see that higher values of SSPG increases the odds of overt diabetes, while a decrease in IR reduces the same odds when compared to normal subjects. Looking at Logit(2/3), we see that the higher values SSPG increases the odds of chemical diabetes when compared to the normal subjects. The IR value does not significantly affect the odds. This indicates the difference between chemical and overt diabetes and has implications for the treatment of the two conditions.

9.6 The Multinomial Logit Model

Example: Determining Chemical Diabetes

Using R

```
################################################   Multinomial logistic regression on Diabetes data
diabetes<-read.table('data/P349-50.txt',header=TRUE)   ## read data
diabetes$CC<-as.factor(diabetes$CC)   ## assign the variable to a factor class so that R treats it as categorical
diabetes$CC<-relevel(diabetes$CC,ref=3)  ## set the class 3 as the base class
library(nnet)      ## requires the library nnet for the multinom function
mod1<-multinom(CC~IR+SSPG, data=diabetes)      ## note that the patient column is the patient ID which should be deleted from the model
summary(mod1)
```

```
> mod1<-multinom(CC~IR+SSPG, data=diabetes)      ## note that the patient column is the patient ID which should be deleted from the model
# weights:  12 (6 variable)
initial  value 159.298782
iter  10 value 72.172679
iter  20 value 72.028901
final  value 72.028883
converged
> summary(mod1)
Call:
multinom(formula = CC ~ IR + SSPG, data = diabetes)

Coefficients:
  (Intercept)           IR       SSPG
1   -7.110590 -0.013427199 0.04259435
2   -4.548408  0.003257602 0.01951007

Std. Errors:
  (Intercept)          IR        SSPG
1   1.6882103 0.004651300 0.007973417
2   0.7714595 0.002292307 0.004451874

Residual Deviance: 144.0578
AIC: 156.0578
```

9.6 The Multinomial Logit Model

Example: Determining Chemical Diabetes

Although we have taken 3 as the base level, from our computation we can derive other comparisons. We can get Logit (1/2) from the relation

$$\text{Logit}(1/2) = \text{Logit}(1/3) - \text{Logit}(2/3). \tag{9.9}$$

We can judge how well the multinomial logistic regression classifies the observations into different categories. The methodology is similar to binary logistic regression. An observation is classified to that category for which it has the highest estimated probability. The classification table for the multinomial logistic regression is given in Table 9.10.

Table 9.10 Classification Table of Diabetes Data Using Multinomial Logistic Regression

	Predict			
CC	1	2	3	All
1	27	3	3	33
2	1	22	13	36
3	2	5	69	76
All	30	30	85	145

One can see that 118 out of 145 subjects studied are classified correctly by this procedure. Thus, 81% of the observations are correctly classified. This is considerably higher than the maximum correct rate 59% (85/145), which would have been obtained if all the observations were put in one category. Multinomial logistic regression has performed well on this data. It is a powerful technique that should be used more extensively.

9.6 The Multinomial Logit Model

Ordinal Logistic Regression (optional)

The response variable in many studies, as has been pointed out earlier, can be qualitative and fall in more than two categories. The categories may sometimes be **ordered**. In a consumer satisfaction study, the responses might be highly satisfied, satisfied, dissatisfied, and highly dissatisfied. An analyst may want to study the socioeconomic and demographic factors that influence the response. The logistic model, slightly modified, can be used for this analysis. The logits here are based on **the cumulative probabilities**. Several logistic models can be based on the cumulative logits. We describe one of these, the **proportional odds model**.

Again, we have n independent observations with p predictors. The response variable falls into k categories $(1, 2, \cdots, k)$. The k categories are ordered. Let Y denote the response variable. The cumulative distribution for Y is

$$F_j(x_i) = \Pr(Y \le j | X_i = x_{i1}, \cdots, X_p = x_{ip},); \quad j = 1, 2, \cdots, (k-1).$$

The proportional odds model is given by,

$$L_j(x_i) = \ln\left(\frac{F_j(x_i)}{1 - F_j(x_i)}\right) = \beta_{0j} + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

for $j = 1, 2, \cdots, (k-1)$. The comulative logit has a simple interpretation. It can be interpreted as the logit for a binary response in which the categories from 1 to j is one category, and the remaining categories from $j+1$ to k is the second category. Note that the coefficients for $X_1, \cdots, X_p$ are the same for different log-odds.

9.6 The Multinomial Logit Model

Ordinal Logistic Regression (optional)

The model is fitted by the maximum likelihood method. Several statistical software packages will carry out this procedure. Increase in the value of a response variable with a positive β will increase the probability of being in a lower numbered category, all other variables remaining the same. The number of parameters estimated to describe the data is fewer in the ordinal than in the nominal model.

Example: Determining Chemical Diabetes Revisited

We will use the data on chemical diabetes to illustrate ordinal logistic regression. The clinical classifications in the previous categories are ordered but we did not take it into consideration in our analysis. The progression of diabetes goes from normal (3), chemical (2) , to overt diabetes (1). The classification states have a natural order and we will use them in our analysis. We will fit the proportional odds logit model. The results of the fit are given in Table 9.11.

Table 9.11 Ordinal Logistic Regression Model (Proportional Odds) Using SSPG and IR

					Odds	95% C.I.	
Variable	Coefficient	s.e.	Z-Test	p-value	Ratio	Lower	Upper
Constant 1	–6.794	0.872	–7.79	0.000			
Constant 2	–4.189	0.665	–6.30	0.000			
IR	–0.004	0.002	–2.30	0.021	1.00	0.99	1.00
SSPG	0.028	0.004	7.73	0.000	1.03	1.02	1.04
Log-Likelihood = –81.749		G = 132.700		df = 2	p-value < 0.000		

9.6 The Multinomial Logit Model

Ordinal Logistic Regression (optional)

The fit for the model is good. Both variables have a significant relationship to the group membership. The coefficient of SSPG is positive. This indicates that higher values of SSPG increase the probability of being in a lower numbered category, other factors being the same. The coefficient of IR is negative, indicating that higher values of this variable increase the probability of being in a higher numbered category, other factors remaining the same. The coefficient of concordance is high (0.90) showing the ability of the model to classify the group membership is high. In Table 9.12, we give the classification table for the ordinal logistic regression.

Of the 145 subjects ordinallogit regression classifies 114 subjects to their correct group. This gives the correct classification rate as 79%. This is comparable to the rate achieved by the multinomial logit model. It is generally expected that the ordinal model will do better than the multinomial model because of the additional information provided by the ordering of the categories. It should be also noted the ordinal logit model uses fewer parameters than the multinomial model. In our example the ordinal model uses 4 parameters, while the nominal version uses 6.

Table 9.12 Classification Table of Diabetes Data Using Multinomial Logistic Regression

	Predict			
CC	1	2	3	All
1	26	5	2	33
2	3	20	13	36
3	0	8	68	76
All	29	33	83	145

9.7. Working with R

9.7 Working with R

On Simulated Data

```r
sim_logistic_data = function(sample_size = 25, beta_0 = -2, beta_1 = 3) {
  x = rnorm(n = sample_size)
  eta = beta_0 + beta_1 * x
  p = 1 / (1 + exp(-eta))
  y = rbinom(n = sample_size, size = 1, prob = p)
  data.frame(y, x)
}
```

Generate simulated data

You might think, why not simply use ordinary linear regression? Even with a binary response, our goal is still to model (some function of) $\mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}]$. However, with a binary response coded as 0 and 1, $\mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}] = P[Y = 1 \mid \mathbf{X} = \mathbf{x}]$ since

$$\begin{aligned}\mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}] &= 1 \cdot P[Y = 1 \mid \mathbf{X} = \mathbf{x}] + 0 \cdot P[Y = 0 \mid \mathbf{X} = \mathbf{x}] \\ &= P[Y = 1 \mid \mathbf{X} = \mathbf{x}]\end{aligned}$$

Then why can't we just use ordinary linear regression to estimate $\mathrm{E}[Y \mid \mathbf{X} = \mathbf{x}]$, and thus $P[Y = 1 \mid \mathbf{X} = \mathbf{x}]$?

To investigate, let's simulate data from the following model:

$$\log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = -2 + 3x$$

Another way to write this, which better matches the function we're using to simulate the data:

$$\begin{aligned}Y_i \mid \mathbf{X_i} = \mathbf{x_i} &\sim \mathrm{Bern}(p_i) \\ p_i = p(\mathbf{x_i}) &= \frac{1}{1 + e^{-\eta(\mathbf{x_i})}} \\ \eta(\mathbf{x_i}) &= -2 + 3x_i\end{aligned}$$

ogistic Regression

9.7 Working with R

On Simulated Data

After simulating a dataset, we'll then fit both ordinary linear regression and logistic regression. Notice that currently the responses variable `y` is a numeric variable that only takes values `0` and `1`. Later we'll see that we can also fit logistic regression when the response is a factor variable with only two levels. (Generally, having a factor response is preferred, but having a dummy response allows use to make the comparison to using ordinary linear regression.)

```
# ordinary linear regression
fit_lm  = lm(y ~ x, data = example_data)
# logistic regression
fit_glm = glm(y ~ x, data = example_data, family = binomial)
```

Notice that the syntax is extremely similar. What's changed?

- `lm()` has become `glm()`
- We've added `family = binomial` argument

In a lot of ways, `lm()` is just a more specific version of `glm()`. For example

```
glm(y ~ x, data = example_data)
```

would actually fit the ordinary linear regression that we have seen in the past. By default, `glm()` uses `family = gaussian` argument. That is, we're fitting a GLM with a normally distributed response and the identity function as the link.

The `family` argument to `glm()` actually specifies both the distribution and the link function. If not made explicit, the link function is chosen to be the **canonical link function**, which is essentially the most mathematical convenient link function. See `?glm` and `?family` for details. For example, the following code explicitly specifies the link function which was previously used by default.

```
# more detailed call to glm for logistic regression
fit_glm = glm(y ~ x, data = example_data, family = binomial(link = "logit"))
```

On Simulated Data

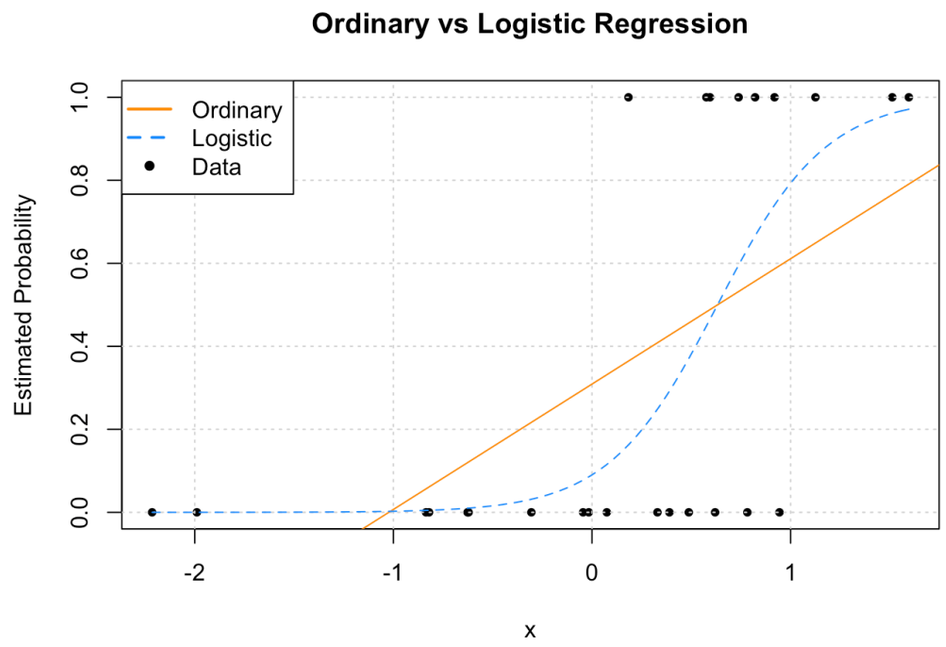
9.7 Working with R

Making predictions with an object of type `glm` is slightly different than making predictions after fitting with `lm()`. In the case of logistic regression, with `family = binomial`, we have:

`type`	Returned
`"link"` [default]	$\hat{\eta}(\mathbf{x}) = \log\left(\frac{\hat{p}(\mathbf{x})}{1-\hat{p}(\mathbf{x})}\right)$
`"response"`	$\hat{p}(\mathbf{x}) = \frac{e^{\hat{\eta}(\mathbf{x})}}{1+e^{\hat{\eta}(\mathbf{x})}} = \frac{1}{1+e^{-\hat{\eta}(\mathbf{x})}}$

That is, `type = "link"` will get you the log odds, while `type = "response"` will return the estimated mean, in this case, $P[Y = 1 \mid \mathbf{X} = \mathbf{x}]$ for each observation.

```
plot(y ~ x, data = example_data,
     pch = 20, ylab = "Estimated Probability",
     main = "Ordinary vs Logistic Regression")
grid()
abline(fit_lm, col = "darkorange")
curve(predict(fit_glm, data.frame(x), type = "response"),
      add = TRUE, col = "dodgerblue", lty = 2)
legend("topleft", c("Ordinary", "Logistic", "Data"), lty = c(1, 2, 0),
       pch = c(NA, NA, 20), lwd = 2, col = c("darkorange", "dodgerblue", "black"))
```



Since we only have a single predictor variable, we are able to graphically show this situation. First, note that the data, is plotted using black dots. The response `y` only takes values `0` and `1`.

On Simulated Data

9.7 Working with R

Next, we need to discuss the two added lines to the plot. The first, the solid orange line, is the fitted ordinary linear regression.

The dashed blue curve is the estimated logistic regression. It is helpful to realize that we are not plotting an estimate of Y for either. (Sometimes it might seem that way with ordinary linear regression, but that isn't what is happening.) For both, we are plotting $\hat{\mathrm{E}}[Y \mid \mathbf{X} = \mathbf{x}]$, the estimated mean, which for a binary response happens to be an estimate of $P[Y = 1 \mid \mathbf{X} = \mathbf{x}]$.

We immediately see why ordinary linear regression is not a good idea. While it is estimating the mean, we see that it produces estimates that are less than 0! (And in other situations could produce estimates greater than 1!) If the mean is a probability, we don't want probabilities less than 0 or greater than 1.

Enter logistic regression. Since the output of the inverse logit function is restricted to be between 0 and 1, our estimates make much more sense as probabilities. Let's look at our estimated coefficients. (With a lot of rounding, for simplicity.)

```
round(coef(fit_glm), 1)
```

```
## (Intercept)           x
##        -2.3         3.7
```

Our estimated model is then:

$$\log\left(\frac{\hat{p}(\mathbf{x})}{1 - \hat{p}(\mathbf{x})}\right) = -2.3 + 3.7x$$

Because we're not directly estimating the mean, but instead a function of the mean, we need to be careful with our interpretation of $\hat{\beta}_1 = 3.7$. This means that, for a one unit increase in x, the log odds change (in this case increase) by 3.7. Also, since $\hat{\beta}_1$ is positive, as we increase x we also increase $\hat{p}(\mathbf{x})$. To see how much, we have to consider the inverse logistic function.

For example, we have:

$$\hat{P}[Y = 1 \mid X = -0.5] = \frac{e^{-2.3+3.7\cdot(-0.5)}}{1 + e^{-2.3+3.7\cdot(-0.5)}} \approx 0.016$$

$$\hat{P}[Y = 1 \mid X = 0] = \frac{e^{-2.3+3.7\cdot(0)}}{1 + e^{-2.3+3.7\cdot(0)}} \approx 0.09112296$$

$$\hat{P}[Y = 1 \mid X = 1] = \frac{e^{-2.3+3.7\cdot(1)}}{1 + e^{-2.3+3.7\cdot(1)}} \approx 0.8021839$$

On SAheart Data

9.7 Working with R

To illustrate the use of logistic regression, we will use the `SAheart` dataset from the `ElemStatLearn` package.

```
# install.packages("bestglm")
library(bestglm)
```

```
## Loading required package: leaps
```

```
data("SAheart")
```

sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1
132	6.20	6.47	36.21	Present	62	30.77	14.14	45	0

This data comes from a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. The `chd` variable, which we will use as a response, indicates whether or not coronary heart disease is present in an individual. Note that this is coded as a numeric `0` / `1` variable. Using this as a response with `glm()` it is important to indicate `family = binomial`, otherwise ordinary linear regression will be fit. Later, we will see the use of a factor variable response, which is actually preferred, as you cannot accidentally fit ordinary linear regression.

The predictors are various measurements for each individual, many related to heart health. For example `sbp`, systolic blood pressure, and `ldl`, low density lipoprotein cholesterol. For full details, use `?SAheart`.

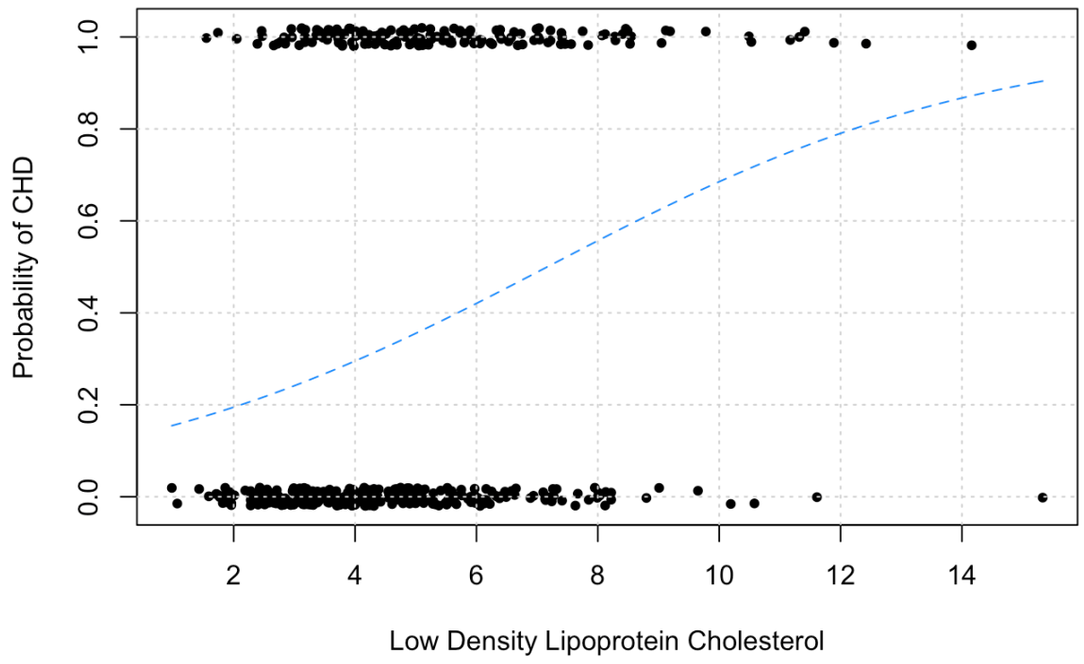
We'll begin by attempting to model the probability of coronary heart disease based on low density lipoprotein cholesterol. That is, we will fit the model

$$\log\left(\frac{P[\texttt{chd} = 1]}{1 - P[\texttt{chd} = 1]}\right) = \beta_0 + \beta_{\texttt{ldl}} x_{\texttt{ldl}}$$

On SAheart Data

9.7 Working with R

```
chd_mod_ldl = glm(chd ~ ldl, data = SAheart, family = binomial)
plot(jitter(chd, factor = 0.1) ~ ldl, data = SAheart, pch = 20,
     ylab = "Probability of CHD", xlab = "Low Density Lipoprotein Cholesterol")
grid()
curve(predict(chd_mod_ldl, data.frame(ldl = x), type = "response"),
      add = TRUE, col = "dodgerblue", lty = 2)
```



As before, we plot the data in addition to the estimated probabilities. Note that we have "jittered" the data to make it easier to visualize, but the data do only take values `0` and `1`.

As we would expect, this plot indicates that as `ldl` increases, so does the probability of `chd`.

```
coef(summary(chd_mod_ldl))
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -1.9686681 0.27307908 -7.209150 5.630207e-13
## ldl          0.2746613 0.05163983  5.318787 1.044615e-07
```

On SAheart Data

9.7 Working with R

To perform the test

$$H_0 : \beta_{\texttt{ldl}} = 0$$

we use the `summary()` function as we have done so many times before. Like the t-test for ordinary linear regression, this returns the estimate of the parameter, its standard error, the relevant test statistic (z), and its p-value. Here we have an incredibly low p-value, so we reject the null hypothesis. The `ldl` variable appears to be a significant predictor.

When fitting logistic regression, we can use the same formula syntax as ordinary linear regression. So, to fit an additive model using all available predictors, we use:

```
chd_mod_additive = glm(chd ~ ., data = SAheart, family = binomial)
```

We can then use the likelihood-ratio test to compare the two models. Specifically, we are testing

$$H_0 : \beta_{\texttt{sbp}} = \beta_{\texttt{tobacco}} = \beta_{\texttt{adiposity}} = \beta_{\texttt{famhist}} = \beta_{\texttt{typea}} = \beta_{\texttt{obesity}} = \beta_{\texttt{alcohol}} = \beta_{\texttt{age}} = 0$$

We could manually calculate the test statistic,

```
-2 * as.numeric(logLik(chd_mod_ldl) - logLik(chd_mod_additive))
```

```
## [1] 92.13879
```

Or we could utilize the `anova()` function. By specifying `test = "LRT"`, `R` will use the likelihood-ratio test to compare the two models.

```
anova(chd_mod_ldl, chd_mod_additive, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: chd ~ ldl
## Model 2: chd ~ sbp + tobacco + ldl + adiposity + famhist + typea + obesity +
##     alcohol + age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       460     564.28
## 2       452     472.14  8   92.139 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the test statistic that we had just calculated appears in the output. The very small p-value suggests that we prefer the larger model.

9.7 Working with R

On SAheart Data

We can create confidence intervals for the β parameters using the `confint()` function as we did with ordinary linear regression.

```
confint(chd_mod_selected, level = 0.99)
```

```
## Waiting for profiling to be done...
```

```
##                       0.5 %      99.5 %
## (Intercept)    -8.941825274 -4.18278990
## tobacco         0.015704975  0.14986616
## ldl             0.022923610  0.30784590
## famhistPresent  0.330033483  1.49603366
## typea           0.006408724  0.06932612
## age             0.024847330  0.07764277
```

Note that we could create intervals by rearranging the results of the Wald test to obtain the Wald confidence interval. This would be given by

$$\hat{\beta}_j \pm z_{\alpha/2} \cdot \text{SE}[\hat{\beta}_j].$$

However, `R` is using a slightly different approach based on a concept called the profile likelihood. (The details of which we will omit.) Ultimately the intervals reported will be similar, but the method used by `R` is more common in practice, probably at least partially because it is the default approach in `R`. Check to see how intervals using the formula above compare to those from the output of `confint()`. (Or, note that using `confint.default()` will return the results of calculating the Wald confidence interval.)