TABLE 5.18 Data for Problem 5.17

Patient	D	T	Y	Patient	D	T	Y	Patient	D	T	Y
1	45	0	0	13	50	1	0	25	20	1	0
2	15	0	0	14	75	1	1	26	45	0	1
3	40	0	1	15	30	0	0	27	15	1	0
4	83	1	1	16	25	0	1	28	25	0	1
5	90	1	1	17	20	1	0	29	15	1	0
6	25	1	1	18	60	1	1	30	30	0	1
7	35	0	1	19	70	1	1	31	40	0	1
8	65	0	1	20	30	0	1	32	15	1	0
9	95	0	1	21	60	0	1	33	135	1	1
10	35	0	1	22	61	0	0	34	20	1	0
11	75	0	1	23	65	0	1	35	40	1	0
12	45	1	1	24	15	1	0				

Source: Data from D. Collett, in *Encyclopedia of Biostatistics* (New York: Wiley: 1998), pp. 350–358.

1 Logistic regression

Q1.1: See Section 12.2.4 in Weisberg (2005).

The Titanic was a luxury passenger ship that sunk when it struck an iceberg on its maiden voyage to New York City from Southampton, England. There were 2201 passengers and crew members on board, out of them 711 people were reported to have survived. The data in the file *TitanicSurvival* contains a list of passengers with the following information about them: Sex, Age (years), Class (first, second, third), the survival status (survived or not).

(a) Fit a logistic regression model with Sex, Age, and Class as explanatory variables.

(b) Compare the odds of survival of the two genders. Which gender has a higher chance of survival? How many times higher?

(c) Compute the odds of survival of each passenger class.

(d) Which class has the highest survival rate?

(e) Fit a logistic regression model with the same explanatory variables and two-way interaction between Class and Sex.

(f) Use the analysis of deviance to see that the second model fits better than the first one. Is the difference in fit significant?

(g) Based on the second model,

- Compute the difference in survival rates of two sexes.
- Using an appropriate odds ratio, compare the odds of survival of males in the first class with those of males in the third class.

Q1.2: Problem 5.17 in Agresti (p. 204) describes data on 35 patients who received general anesthesia for surgery. The outcome of interest is whether a patient experienced a sore throat upon awakening: "0" - no, "1" - yes. For each patient, the duration of surgery in minutes and the type of device used to secure the airway ("0" = laryngeal mask airway, "1" = tracheal tube) were also recorded. The observed data are summarized in the following table:

(a) Disregarding the duration information, compute the observed joint distribution of sore throat incidence and type of anesthesia and carry out a chi-squared test to see whether these two variables are independent from each other.

(b) Fit a logistic regression model to investigate whether the surgery duration and anesthesia type have a significant effect on the occurrence of sore throat.

(c) Which statistical test can be used to assess whether the effect of duration differs for the two types of anesthesia?

2 Multiple linear regression

Q2.1: Hastie, Tibshirani, and Friedman (2009). "Elements of statistical learning".

The data come from a study by Stamey et al. (1989), aiming to examine the correlation between the level of prostate-specific antigen and several clinical measures in men who were about to receive a radical prostatectomy.

(a) Download the data file *prostate.dat*.

(b) Compute the correlation matrix of the data and a scatterplot matrix of pairwise plots (omit the dummy variable called *train*).

(c) The response variable is *lpsa*. Which predictor variables are highly correlated with the response?

(d) Which predictor variables are highly correlated with each other?

(e) Standardize the predictors to have unit variance.

(f) Choose the subset of the data for which *train* is TRUE.

(g) To these data, fit a linear regression model including all predictor variables.

(h) Does the model summary support the information about correlation which you could take from the correlation plot?

(i) Which variables have the strongest effect?

(j) Which effects are not significant?

(k) Drop out the non-significant terms and fit a linear regression model without them.

(l) Use F-test to check whether the simpler model could also be appropriate.