Exercise (2.16). The table below comes from one of the first studies of the link between lung cancer and smoking, by Richard Doll and A. Bradford Hill. In 20 hospitals in London, UK, patients admitted with lung cancer in the previous year were queried about their smoking behavior. For each patient admitted, researchers studied the smoking behavior of a noncancer control patient at the same hospital of the same sex and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least one year:

<table>
<thead>
<tr>
<th>Have Smoked</th>
<th>Lung Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
</tr>
<tr>
<td>Yes</td>
<td>688</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>709</td>
</tr>
</tbody>
</table>

1. Identify the response variable and the explanatory variable.
   (Solution) The response (dependent or Y) variable is the smoking preference of a patient. The explanatory (independent or X) variable is lung cancer.

2. Identify the type of study this was.
   (Solution) This was a case-control study.

3. Can you use the data to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer? Why or why not?
   (Solution) While the patients in the 'control' group do not have lung cancer, they need not be healthy. These other health conditions may have an impact on their lung cancer status. Thus, a comparison of the smokers with nonsmokers in terms of the proportion who suffered lung cancer will not conclusively show that there is a relation between smoking habits and lung cancer.

4. Summarize the association, and explain how to interpret it.
   (Solution) \( \theta = \frac{688 \times 59}{650 \times 21} = 2.9738 > 1 \). The odds of having lung cancer is 2.9738 times higher for a person who smokes than a person who does not smoke.

Exercise (4.7). Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, which is severe forward flexion of the spine following corrective spinal surgery. The age in months at the time of the operation for the 18 subjects for whom kyphosis was present were 12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139, 159, 157 and for the 22 subjects for whom kyphosis was absent were 1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72, 81, 97, 112, 118, 127, 131, 140, 151, 159, 177, 206.

1. Fit a logistic regression model using age as a predictor of whether kyphosis is present. Test whether age has a significant effect.
   (Solution) \( \text{Logit}(\hat{\rho}(x)) = \alpha + \beta x = -0.572693 + 0.004296x \) where \( x \) stands for age. To test the significance of age, let the null hypothesis be \( \beta = 0 \) and the alternate hypothesis be \( \beta \neq 0 \). The \( p \)-value of 0.463 provides strong evidence to not reject the null hypothesis. Thus, age is not significant.

2. Plot the data. Note the difference in dispersion of age at the two levels of kyphosis.
   (Solution) Age seems to have a slightly higher dispersion when kyphosis is absent.
3. Fit the model \[ \logit[\pi(x)] = \alpha + \beta_1 x + \beta_2 x^2. \] Test the significance of the squared age term, plot the fit, and interpret. (The final paragraph of Section 4.1.6 is relevant to these results.)

(Solution) \[ \logit(\hat{\pi}(x)) = -2.0462547 + 0.0600398 x - 0.0003279 x^2. \] The \( p \)-value of 0.0360 indicates to reject the null hypothesis that \( \beta_2 = 0 \) and accept the alternate hypothesis \( \beta_2 \neq 0 \). Hence, the squared age term is significant.

Note that R programming language was used to answer the above questions. My code for this Exercise is given below.

```r
# Exercise 4.7 (Agresti) written by Divya Nair
#=============================================
# Variables
kyphosis<-rep(c(1,0),c(18,22)) # 1=kyphosis present for first 18 age, 0=kyphosis absent for next 22 age
age<-c(12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139, 139, 157,
1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72, 81, 97, 112, 118, 127, 131, 140, 151, 159, 177, 206)
agesq<-age*age

# GLM Estimation
model1<-glm(kyphosis~age,family=binomial(link='logit'))
summary(model1)
model2<-glm(kyphosis~age+agesq,family=binomial(link='logit'))
summary(model2)

# Data plot
df('kyphosis data.pdf')
plot(c(1,210),c(0,1),type='n',xlab='Age',ylab='Kyphosis')
text(30,0.2,"0 = kyphosis absent")
text(30,0.15,"1 = kyphosis present")
points(age,kyphosis,pch=19,col='red')
dev.off()
```

**Exercise (4.10).** An international poll quoted in an Associated Press story (December 14, 2004) reported low approval ratings for President George W. Bush among traditional allies of the United States, such as
32% in Canada, 30% in Britain, 19% in Spain, and 17% in Germany. Let $Y$ indicate approval of Bush’s performance (1 = yes, 0 = no), $\pi = P(Y = 1)$, $c_1 = 1$ for Canada and 0 otherwise, $c_2 = 1$ for Britain and 0 otherwise, and $c_3 = 1$ for Spain and 0 otherwise.

1. Explain why these results suggest that for the identity link function, $\hat{\pi} = 0.17 + 0.15c_1 + 0.13c_2 + 0.02c_3$.
   (Solution) The probability of approval for Canada is $P(Y = 1) = \hat{\pi} = 0.17 + 0.15(1) = 0.32 = 32\%$. The probability of approval for Britain is $P(Y = 1) = \hat{\pi} = 0.17 + 0.13(1) = 0.30 = 30\%$. The probability of approval for Spain is $P(Y = 1) = \hat{\pi} = 0.17 + 0.02(1) = 0.19 = 19\%$. The probability of approval for Germany is $P(Y = 1) = \hat{\pi} = 0.17 = 17\%$.

2. Show that the prediction equation for the logit link function is $\logit(\hat{\pi}) = -1.59 + 0.83c_1 + 0.74c_2 + 0.14c_3$.
   (Solution) $\logit(\hat{\pi}) = \ln \frac{\pi}{1-\pi}$ can be found for each country with the given probabilities. This value will be the same as the value obtained from the given logit link function for each corresponding $c_i = 1$. $\logit(\hat{\pi})$ for Germany is $\ln \frac{0.17}{1 - 0.17} = -1.59$, $\logit(\hat{\pi})$ for Canada is $\ln \frac{0.32}{1 - 0.32} = -0.75 = -1.59 + 0.83(1)$, $\logit(\hat{\pi})$ for Britain is $\ln \frac{0.30}{1 - 0.30} = -0.85 = -1.59 + 0.74(1)$, and $\logit(\hat{\pi})$ for Spain is $\ln \frac{0.19}{1 - 0.19} = -1.45 = -1.59 + 0.14(1)$.

**Exercise.** Derive the non-linear GLM $\logit(P(x)) = \alpha + \beta x + \gamma x^2$.

**Proof.** Let $Y$ be a binary response variable so that $Y \sim \text{Bernoulli}(\pi)$. Let $X$ be the explanatory variable such that it has a conditional normal distribution, that is,

$$
\begin{align*}
  f(X|Y=1) &= N(\mu_1, \sigma^2_1) = \frac{\exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2_1}\right)}{\sigma_1 \sqrt{2\pi}}, \\
  f(X|Y=0) &= N(\mu_0, \sigma^2_0) = \frac{\exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2_0}\right)}{\sigma_0 \sqrt{2\pi}}.
\end{align*}
$$

By Bayes’ Theorem,

$$
P(Y = 1|X = x) = \frac{f(X = x|Y = 1) \cdot P(Y = 1)}{P(X = x)}.
$$

By law of total probability, the denominator above can be written as

$$
P(X = x) = f(X = x|Y = 1) \cdot P(Y = 1) + f(X = x|Y = 0) \cdot P(Y = 0).
$$

Denote $P(Y = 1) = p$ and $P(Y = 1|X = x) = P(x)$. Then,

$$
P(x) = \frac{f(X = x|Y = 1) \cdot P(Y = 1)}{f(X = x|Y = 1) \cdot P(Y = 1) + f(X = x|Y = 0) \cdot P(Y = 0)} = \frac{\exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2_1}\right) \cdot p}{\exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2_1}\right) \cdot p + \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2_0}\right) \cdot (1 - p)}.
$$

Divide the right-hand side by $\exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2_1}\right) \cdot p$ to obtain

$$
P(x) = \frac{1}{1 + \frac{1-p}{p} \cdot \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2_0} + \frac{(x-\mu_1)^2}{2\sigma^2_1}\right)}.
$$

The exponent expands to

$$
\frac{(x - \mu_1)^2}{2\sigma^2_1} - \frac{(x - \mu_0)^2}{2\sigma^2_0} = x^2(2\sigma^2_0 - 2\sigma^2_1) + \frac{x(4\sigma^2_1\mu_0 - 4\sigma^2_1\mu_1)}{4\sigma^2_1\sigma^2_0} + \frac{2\sigma^2_0 - 2\sigma^2_1\mu_0^2}{4\sigma^2_1\sigma^2_0}.
$$
Hence,

\[ P(x) = \frac{1}{1 + \exp(\alpha + \beta x + \gamma x^2)} \]

where \( \alpha = \frac{2\sigma_0^2 - 2\sigma_1^2 \mu_0^2}{4\sigma_1^2\sigma_0^2} + \ln \left( \frac{1-p}{p} \right) \), \( \beta = \frac{4\sigma_1^2 \mu_0 - 4\sigma_0^2 \mu_1}{4\sigma_1^2\sigma_0^2} + \ln \left( \frac{1-p}{p} \right) \), and \( \gamma = \frac{2\sigma_0^2 - 2\sigma_1^2}{4\sigma_1^2\sigma_0^2} + \ln \left( \frac{1-p}{p} \right) \). Then

\[ \frac{P(x)}{1 - P(x)} = \exp(\alpha + \beta x + \gamma x^2). \]

Therefore, \( \log \left( \frac{P(x)}{1 - P(x)} \right) = \logit(P(x)) = \alpha + \beta x + \gamma x^2. \)