Chapter 9 Solutions

9.1 The mean of binary responses is a proportion between 0 and 1 inclusive. However, simple linear regression lines extend below 0 and above 1. So using a line to model mean binary responses would lead to nonsensical predictions such as negative proportions or proportions greater than one. Also, in the case of binomial counts, the standard deviation is not equal at all possible values of p. Specifically, the standard deviation for a given p is $\sqrt{p(1-p)}$.

9.2 In an ordinary regression model, we predict an actual value for an observation with a specific value of the predictor variable. This means we can compute the difference between a real observation at that value of the predictor variable and the predicted value. The difference between these two numbers is the residual and measures how far off the model is for that particular observation. However, in a logistic regression model, we are computing a probability that an observation at a specific value of the predictor variable will be a success. Some observations are likely to be successes and others are not. Unless we have many observations at that particular value of the predictor variable, we have no sense for how far off the predicted probability of success is.

9.3 a. $0.5/0.5 = 1$

b. $0.9/0.1 = 9$

c. $0.1/0.9 = 0.111$

9.4 a. $0.8/0.2 = 4$

b. $0.25/0.75 = 0.333$

c. $0.6/0.4 = 1.5$

9.5 a. $2/(2+1) = 2/3$

b. $10/(10+1) = 0.91$

c. $1/(1+4) = 0.2$

9.6 a. $1/(1+3) = 0.25$

b. $5/(5+2) = 0.71$

c. $1/(1+9) = 0.1$

9.7 $\frac{0.3/0.7}{0.1/0.9} = 3.86$

9.8 $\frac{0.6/0.4}{0.01/0.99} = 148.5$

9.9 a. The curve does not rise so steeply when β_1 decreases from 2 to 1.

b. Increasing the intercept to 8 shifts the curve horizontally to the left.

c. When the sign of the slope changes from positive to negative, the curve falls instead of rises.

9.10 a. The curve does not fall so steeply when β_1 increases from -3 to -1.

b. Increasing the intercept to 5 shifts the curve horizontally to the left.

c. When the sign of the slope changes from negative to positive, the curve rises instead of falls.

9.11 a. This statement might be true. The models give different values for $\log(odds)$ for all x values other than $x = 0$, but it could be that $\log(odds) = 0$ (and $\pi = 1/2$) when $x = 0$.

b. This statement must be true since the models have the same y-intercept.

c. This statement must be true since the asymptotes are at zero and one.

d. This statement might be true. The models give different values for $P(Y = 1)$ for all x values other than $x = 0$, but it could be that $P(Y = 1) = 0.5$ when $x = 0$.

e. This statement cannot be true. The models agree at the point $x = 0$ and slope away from that point at different rates, so they cannot agree when $x = 0.5$.

9.12 a. This statement cannot be true, since parallel lines must have different x-intercepts.

b. This statement cannot be true, since the models have the different y-intercepts.

c. This statement must be true, since the asymptotes are at zero and one.

d. This statement cannot be true, since parallel lines cannot coincide at $y = 0.5$.

e. This statement cannot be true, since parallel lines cannot coincide at $x = 0.5$.

9.13 a. This statement might be true. The models give different values for $\log(odds)$ for most x values, but it could be that the models agree when $\log(odds) = 0$.

b. This statement cannot be true, since the models have different y-intercepts.

c. This statement must be true since the asymptotes are at zero and one.

d. This statement might be true. The models give different values for $P(Y = 1)$ for most x values, but it could be that $P(Y = 1) = 0.5$ at a common value of x.

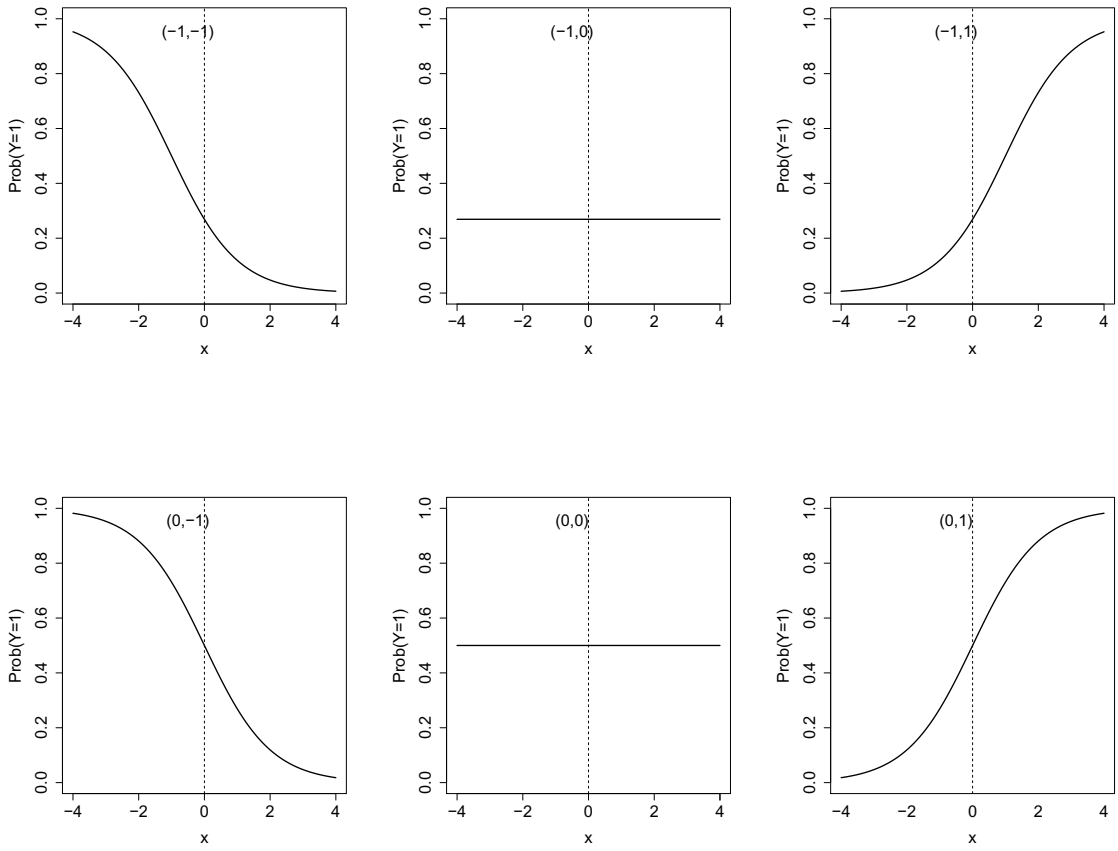
e. This statement might be true. The models give different values for $P(Y = 1)$ for most x values, but it could be that the models agree when $x = 0.5$.

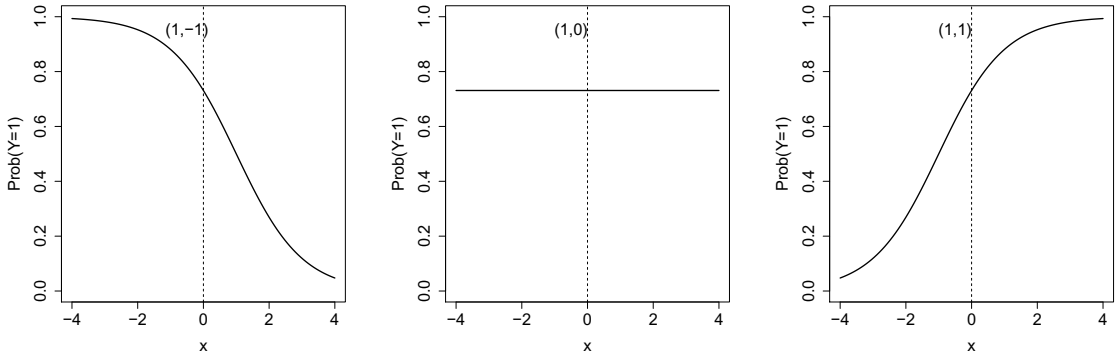
9.14 The two fitted models will have the same intercept—since yards = 0 when feet = 0—but different slopes. The slope for Model B will be three times as large as the slope for Model A since a one-unit change in yards (Model B) corresponds to a three-unit change in feet (Model A).

9.15 Recall that odds ratio $= e^{\text{fitted slope}}$ so fitted slope $= \log(\text{odds ratio})$. Thus to get an odds ratio of 0.01 we need a fitted slope of $\log(0.01)$, or -4.605. To get an odds ratio of 1, we need a fitted slope of $\log(1) = 0$. To get an odds ratio of 2, we need a fitted slope of $\log(2) = 0.693$. To get an odds ratio of e, we need a fitted slope of $\log(e) = 1$.

9.16 Recall that odds ratio $= e^{\text{fitted slope}}$. Thus the odds ratio for a fitted slope of 1 is e^1, or e. The odds ratio for a fitted slope of -0.5 is $e^{-0.5} = 0.607$. The odds ratio for a fitted slope of 5 is $e^5 = 148.413$. The odds ratio for a fitted slope of -4 is $e^{-4} = 0.018$.

9.17 The graphs are given below with their respective values of β_0 and β_1. Note that the equations with a larger y-intercept cross the line $x = 0$ at a higher point. Also, the equations with a negative slope are decreasing, those with 0 slope are flat, and those with positive slope are increasing.





9.18 Independence: Since many pitches will happen in the same game and in fact, one after another, they are likely to be related in some way, so there may be some correlation in the data. Randomness: The pitches are not randomly selected, so the randomness condition does not hold.

9.19 The following output gives us the relevant coefficients needed for both models (see the column marked "Coef").

```
Logistic Regression Table

                                                    Odds      95% CI
Predictor      Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant   -8.71245    3.23653  -2.69  0.007
MCAT       0.245964  0.0893806   2.75  0.006   1.28   1.07   1.52
```

a. Logit form:

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -8.712 + 0.246MCAT$$

Probability form:

$$\hat{\pi} = \frac{e^{-8.712+0.246MCAT}}{1+e^{-8.712+0.246MCAT}}$$

b. The odds ratio is 1.28. This means that the odds of being accepted into medical school increase by a factor of 1.28 for each additional point a student scores on the MCAT.

c. The predicted probability that an applicant with an MCAT of 40 is admitted is estimated to be

$$\hat{\pi} = \frac{e^{-8.712+0.246(40)}}{1+e^{-8.712+0.246(40)}} = 0.755.$$

Alternatively, one can describe this chance in terms of odds,

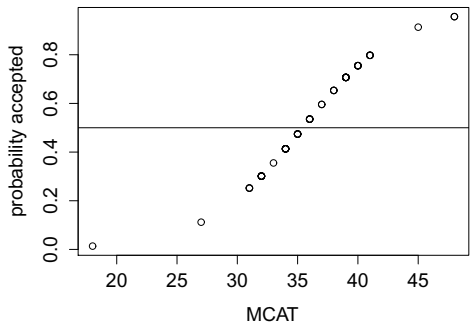
$$\widehat{odds} = e^{-8.712+0.246(40)} = 3.08$$

or approximately 3 to 1.

d. For a 50-50 chance of admission the odds are $50/50 = 1$, so $\log(odds) = \log(1) = 0$. To find the corresponding MCAT score, we solve

$$0 = -8.712 + 0.246MCAT \qquad \text{to get} \qquad MCAT = 8.712/0.246 = 35.4$$

Applicants with an MCAT of approximately 35.4 have a 50-50 chance of admission. This can be seen from the following graph of predicted probability versus MCAT score.



9.20 The following output gives us the relevant coefficients needed for both models (see the column marked "Coef").

```
Logistic Regression Table

Term        Coef  SE Coef       95% CI       Z-Value  P-Value
Constant  -1.883    0.361  (-2.591, -1.176)    -5.22    0.000
Steps     0.1745   0.0525  (0.0716, 0.2775)     3.32    0.001
```

a. Logit form:

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.883 + 0.1745Steps$$

Probability form:

$$\hat{\pi} = \frac{e^{-1.883+0.1745Steps}}{1+e^{-1.883+0.1745Steps}}$$

b. The odds ratio is 1.19. This means that the odds of having walked the dog increase by a factor of 1.19 for every increase of 1000 steps.

c. The predicted probability that a day with 4000 steps will include walking the dogs is estimated to be

$$\hat{\pi} = \frac{e^{-1.883+0.1745(4)}}{1+e^{-1.883+0.1745(4)}} = 0.234.$$

Alternatively, one can describe this chance in terms of odds,

$$\widehat{odds} = e^{-1.883+0.1745(4)} = 0.306$$

or approximately 0.306 to 1.

d. For a 50-50 chance of walking the dogs the odds are $50/50 = 1$, so $\log(odds) = \log(1) = 0$. To find the corresponding value of *Steps*, we solve

$$0 = -1.883 + 0.1745Steps \qquad \text{to get} \qquad Steps = 1.883/0.1745 = 10.79$$

Days where the author walked approximately 10,800 steps have a 50-50 chance of him walking the dogs.

9.21 a. The odds of metastasis is $e^{-2.086+0.5117(6)} = e^{0.9842} = 2.676$.

b. The probability of metastasis is

$$\frac{e^{-2.086+0.5117*6}}{1+e^{-2.086+0.5117(6)}} = \frac{2.676}{1+2.676} = 0.728$$

c. The odds of metastasis for a 7-cm tumor is

$$\frac{e^{-2.086+0.5117(7)}}{e^{-2.086+0.5117(6)}} = 1.67$$

times that of the odds for a 6-cm tumor. Note that this is equal to the exponentiated coefficient for size, $e^{0.5117} = 1.67$.

d. The probability of metastasis for a 7-cm tumor is $\frac{e^{-2.086+0.5117(7)}}{1+e^{-2.086+0.5117(7)}} = 0.817$ in contrast to the probability for a 6-cm tumor was found in part (b) to be 0.728.

9.22 a. When $MMSE = -4$, the fitted model gives log(odds) of $-0.742 - 0.294(-4) = 0.434$. Thus the odds are $e^{0.434} = 1.54$.

b. From part (a) we know that the odds are 1.54. Thus the probability is $1.54/2.54 = 0.606$.

c. When $MMSE = -3$, the fitted model gives log(odds) of $-0.742 - 0.294(-3) = 0.14$. Thus the odds are $e^{0.14} = 1.15$.

d. From part (c) we know that the odds are 1.15. Thus the probability is $1.15/2.15 = 0.535$.

9.23 When $\pi = 0.8$, the odds are $\pi/(1-\pi)$ or $0.8/0.2 = 4$. The model provides an estimate of the log(odds) or logit to be $-2.086 + 0.5117Size$. Thus when $\pi = 0.8$

$$\log(4) = -2.086 + 0.5117Size$$

Solving for $Size$ we get

$$Size = (\log(4) + 2.086)/0.5117 = 6.79 \text{ cm}$$

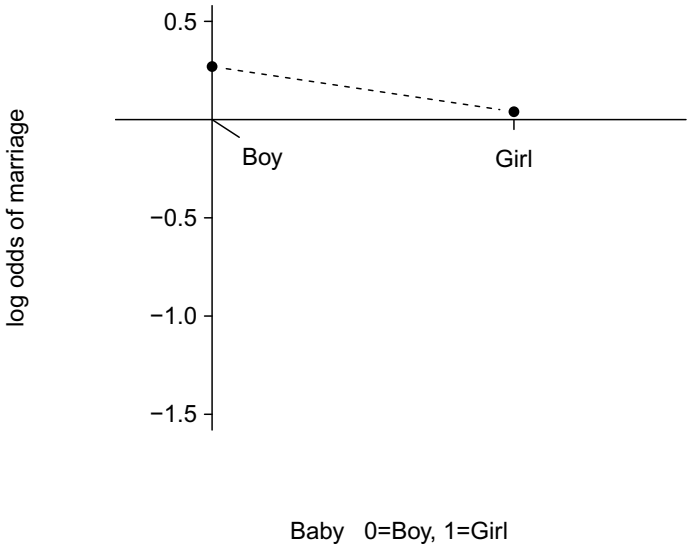
9.24 When $\pi = 0.75$, the odds are $\pi/(1-\pi)$ or $0.75/0.25 = 3$. The model provides an estimate of the log(odds) or logit to be $-0.742 - 0.294MMSE$. Thus when $\pi = 0.75$

$$\log(3) = -0.742 - 0.294MMSE$$

Solving for $MMSE$ we get

$$MMSE = (\log(3) + 0.742)/(-0.294) = -6.26 \text{ cm}$$

9.25 The odds of eventual marriage for mothers with boy babies is $176/134 = 1.313$ and for girl babies it's $148/142 = 1.042$. The change in log($odds$) from boys (0) to girls (1) is $log(1.042) - log(1.313) = -0.23$. That is the slope of the line connecting the two log(odds) values in the plot below.



9.26 Starting with the logit form

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Exponentiate both sides (to lose the log)

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

Multiply both sides by $1-\pi$

$$\pi = (1-\pi)e^{\beta_0+\beta_1 X} = e^{\beta_0+\beta_1 X} - \pi e^{\beta_0+\beta_1 X}$$

Put the terms involving π on the same side of the equation

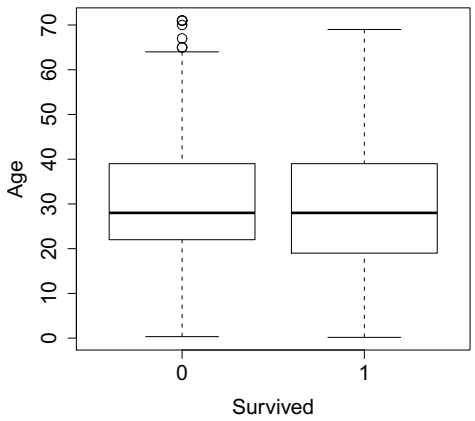
$$\pi + \pi e^{\beta_0+\beta_1 X} = \pi(1 + e^{\beta_0+\beta_1 X}) = e^{\beta_0+\beta_1 X}$$

Divide both sides by $1 + e^{\beta_0+\beta_1 X}$

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

If you would rather start with the probability form of the model, just read the sequences of steps above from bottom to top!

9.27 a. Here is a boxplot comparing the age distribution for survivors and nonsurvivors. The two age distributions look similar.



b. Here is some output from a logistic regression model to predict *Survived* based on *Age*.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.081428   0.173862  -0.468   0.6395
Age         -0.008795   0.005232  -1.681   0.0928
```

The estimated logistic model is $logit(\hat{\pi}) = -0.0814 - 0.008795Age$. The coefficient of Age is negative, indicating some evidence that older passengers were less likely to survive, but the P-value $= 0.0928$ would only be considered significant at a 10% level.

9.28 a. Here is a two-way table of Sex by $Survived$.

```
        Survived
Sex        0   1
  female 154 308
  male   709 142
```

The proportion surviving for each gender are

$$\hat{p}_f = \frac{308}{154 + 308} = 0.667 \qquad \hat{p}_m = \frac{142}{709 + 142} = 0.167$$

The proportion of females who survived (0.667) is much higher than the proportion of males who survived (0.167).

b. Here is some output from fitting a logistic regression model to predict $Survived$ based on sex, where $SexCode = 1$ for females and 0 for males.

```
Logistic Regression Table

                                                  Odds     95% CI
Predictor      Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant   -1.60803  0.0919376  -17.49  0.000
SexCode     2.30118   0.134881   17.06  0.000   9.99   7.67  13.01
```

The small P-value ≈ 0 gives strong evidence that there is some relationship between sex and survival. The positive slope indicates that the probability of survival is lower for males than it is for females. This is consistent with the proportions found in part (c).

9.29 a. From the coefficient of $SexCode$ we estimate an odds ratio to be $OR = e^{2.30118} = 9.986$. This means that the odds for a female surviving are about 10 times the odds for a male surviving.

b. Using the output from the previous answer, we find a 95% confidence interval for the odds ratio to be (7.67, 13.01). We are 95% confidence that the odds for females surviving are between 7.7 and 13.0 times the odds for males surviving.

c. Based on the two-way table, we can find the odds ratio for survival of females compared to males.

$$OR = \frac{308/154}{142/709} = \frac{709(308)}{142(154)} = 9.986$$

Finding the log of the odds ratio gives the estimated coefficient for $SexCode$ in the model, $log(9.986) = 2.301$.

d. $\hat{p}_f = \dfrac{e^{-1.60803+2.301(1)}}{1+e^{-1.60803+2.301(1)}} = \dfrac{2.00}{3.00} = 0.667$

e. We should worry about independence since passengers on the *Titanic* are likely to be related to each other. The bigger concern is with randomness since these data represent *all* passengers on the ship, so we really have the entire population, not a sample of a larger population (unless we are willing to assume the *Titanic* passengers are a sample of all ocean liner passengers in that era).

9.30 a. From the coefficient of *Age* we estimate an odds ratio to be $OR = e^{-0.00879} = 0.9912$. This means that the odds for surviving decrease by a factor of 0.9912 for every year older a person is.

b. Using the output from the previous answer, we find a 95% confidence interval for the odds ratio to be (0.9811, 1.0015). We are 95% confident that the odds for surviving are between 0.9811 and 1.9915 greater for each year older a person is.

c. $\hat{p}_f = \dfrac{e^{-0.081-0.00879(40)}}{1+e^{-0.081-0.00879(40)}} = \dfrac{0.6488}{1.6488} = 0.394$

9.31 There is convincing evidence about a relationship between *Sex* and *Survived*. The proportion of males who survived (0.167) is much smaller than the proportion of females who survived (0.667). This large a difference would be almost impossible to see by random chance alone.

9.32 The relationship between survival and age is not very strong. There is some suggestion that older passengers were less likely to survive, but the P-value (0.0928) is not small enough to make a strong conclusion about that relationship.

9.33 a. Odds of flight for different altitudes:

	Flight	
AltCat	no	yes
high	17	59
mid	77	121
low	85	105

Here are the odds of flight for each altitude

$$\text{high: } \frac{59}{17} = 3.47 \qquad \text{mid: } \frac{121}{77} = 1.57 \qquad \text{low: } \frac{105}{85} = 1.24$$

The log(odds) are log(3.47) = 1.244 (high), log(1.57) = 0.451 (mid), and log(1.24) = 0.215 (low).

The odds of flight with high altitudes are well over twice that of low altitudes, so there does appear to be an association between altitude and propensity for flight.

b. Logistic regression results:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.09600    0.17201   0.558   0.5768
Altitude     0.11503    0.04563   2.521   0.0117 *
```

These results suggest that higher altitudes are associated with greater probability of flight. Specifically, the odds of flight increase by a factor of $e^{0.11503} = 1.12$, or 12% ($p = 0.01$), for each additional 100 m of altitude.

9.34 a. Odds of flight for different lateral distances:

```
          Flight
LatCat    no  yes
     1    37  243
     2    68   37
     3    44    4
     4    30    1
```

Here are the odds of flight for each lateral distance

$$1: \quad \frac{243}{37} = 6.57 \qquad 2: \quad \frac{37}{68} = 0.54 \qquad 3: \quad \frac{4}{44} = 0.091 \qquad 4: \quad \frac{1}{30} = 0.033$$

The log(odds) are $1 : \log(6.57) = 1.88$, $2 : \log(0.54) = -0.61$, $3 : \log(0.091) = -2.40$, and $4 : \log(0.033) = -3.41$.

The odds of flight is dramatically higher at lateral level 1, much less at level 2, and less still at levels 3 and 4. There appears to be an association, but it does not look linear.

b. Logistic regression results:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.96206    0.25604   11.57   <2e-16 ***
Lateral     -0.23325    0.02165  -10.77   <2e-16 ***
```

There is a statistically significant relationship between the lateral distance and the propensity for flight ($p \approx 0$). For each additional 100 m, the odds of flight decrease by a factor of $e^{-0.2335} = 0.792$ (or about 20.8%).

9.35 a. Here is some output from fitting a logistic regression for Up based on $DJIAch$:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.021359   0.308101  -0.069  0.94473
DJIAch       0.013215   0.004151   3.183  0.00146 **
```

The test for the coefficient of $DJIAch$ indicates that it is a statistically significant predictor of whether the Nikkei 225 went up ($p = 0.00146$).

b. Logistic regression of Up on $lagNik$:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.040834   0.275280   0.148   0.8821
lagNik      -0.003525   0.002082  -1.694   0.0903
```

At a 5% significance level, $lagNik$ is not a significant predictor of Up ($p = 0.09$).

c. $DJIAch$ is a better predictor of Up than $lagNik$ based on their relative P-values.

9.36

TrumpWin as a function of income

```
Coefficients
Term          Coef   SE Coef   Z-Value    P-Value
Constant     11.18      3.08      3.64      0.000
Income  -0.000197  0.000056     -3.52      0.000
```

TrumpWin as a function of high school education

```
Coefficients
Term         Coef  SE Coef    Z-Value   P-Value
Constant     9.07     8.64       1.05     0.294
HS        -0.0981   0.0977      -1.00     0.315
```

TrumpWin as a function of college education

```
Coefficients
Term        Coef   SE Coef    Z-Value   P-Value
Constant   18.00      5.11       3.52     0.000
BA        -0.599     0.174      -3.45     0.001
```

TrumpWin as a function of % Dem − % Rep

```
Coefficients
Term        Coef   SE Coef    Z-Value   P-Value
Constant   0.323     0.435       0.74     0.458
Dem.Rep -0.2503    0.0745      -3.36     0.001
```

Income, *BA*, and *Dem.Rep* have similar *P*-values, suggesting that they would be similarly effective in explaining variation in Trump wins. Differing levels of having at least a high school education is not as effective ($p = 0.315$). Note that the magnitude of the estimated coefficients is dependent upon the metric for each variable, so comparing the magnitudes of β_1's is less useful for assessing effectiveness in prediction.

9.37 a. Here is some output from a logistic model to predict $TrumpWin$ based on $Income$.

```
Coefficients
Term          Coef    SE Coef    Z-Value    P-Value
Constant     11.18       3.08       3.64      0.000
Income  -0.000197   0.000056      -3.52      0.000
```

For each \$1 increase in income, the odds of a Trump win increase by a factor of $e^{-0.000197} = 0.9998$.

b. The standard error for the coefficient of $Income$ is 0.000056, so we find a 95% confidence interval for the coefficient with

$$-0.000197 \pm 1.96(0.000056) = -0.000197 \pm 0.00010976 = (-0.00030676, -0.00008724)$$

We exponentiate these values to find a 95% confidence interval for the odds ratio.

$$(e^{-0.00030676}, e^{-0.00008724}) = (0.9997, 0.9999)$$

95% CI for OR: odds of a Trump win with \$1 increase in income is (0.9997, 0.9999).

9.38 a. When using $IncomeTh$ (Income in \$1000's) as the predictor, the logistic regression output is shown below.

```
Coefficients
Term          Coef    SE Coef      Z-Value    P-Value
Constant     11.18       3.08         3.64      0.000
IncomeTh  -0.1967     0.0558        -3.52      0.000
```

The fitted logit equation is

$$\log\left(\frac{\hat{\pi}}{1+\hat{\pi}}\right) = 11.18 - 0.1967 IncomeTh$$

The predicted probability when income is measured in dollars is

$$\hat{\pi} = \frac{e^{11.18 - 0.000197 Income}}{1 + e^{11.18 - 0.000197 Income}}$$

The predicted probability when income is measured in thousands of dollars is

$$p = \frac{e^{11.18 - 0.1967 IncomeTh}}{1 + e^{11.18 - 0.1967 IncomeTh}}$$

Since $-0.000197 Income = -0.1967 Income/1000$, the two probabilities are equal.

b. *Income* was recorded in dollars; however, this makes interpreting the odds ratio cumbersome. Easier and more meaningful is to account for income in \$1000s of dollars: For each additional \$1000 dollars in income, the odds of a Trump win increases by a factor of $e^{-0.1967} = 0.8215$, or a decrease of about 22%.

c. The 95% CI for the OR using this scale ranges from 0.7363 to 0.9164.

$$e^{-0.1967 \pm 1.96(0.0558)} = (e^{-0.3061}, e^{-0.0873}) = (0.7363, 0.9164)$$

9.39 a. H_0 : The slope in the logistic model is 0.

b. The probability decreases as time goes forward; the negative estimated slope (-0.005899) tells us this.

c. $-0.005899 \pm 1.96(0.001049)$, that is, $(-0.007955, -0.003843)$

d. the predicted log(odds) of the presence of a gunnel at time 600, which is 10 a.m.

e. To get the estimated odds ratio, we compute $e^{-0.005899}$, or 0.9941. The odds of a find go down by a factor of 0.9941 for each additional minute of time.

f. To get the confidence interval, we exponentiate the CI from part (c): $(0.9921, 0.9962)$.

g. $e^{(60\hat{beta}_1)} = e^{60(-0.005899)} = 0.7019$. Thus the odds of a gunnel being present goes down by a factor of 0.7019 for each passage of an hour.

9.40 a. H_0 : The slope in the logistic model is 0.

b. The probability decreases as time goes forward; the negative estimated slope (-0.691) tells us this.

c. $-0.691 \pm 1.96(0.346)$, that is, $(-1.369, -0.013)$

9.41 a. Age is significantly associated with response to treatment. The odds ratio is $OR = e^{-0.04676} = 0.95$ and the P-value is 0.017. For each additional year older, the odds of responding to treatment decrease by a factor of 0.95 (or about 5%).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.19678    1.00548   2.185   0.0289 *
Age         -0.04676    0.01952  -2.395   0.0166 *
```

b. To get a two-way table we need to divide the ages into two groups. We could do this in many different ways. For example, 50 is the median age in this study so we can make a new indicator of whether a subject is at least 50 years old.

```
               Respond
               no  yes
less than 50    9   15
50 or more     18    9
```

For the younger patients, $\frac{15}{9+15} = 0.625$, or 62.5%, responded. For the older patients, $\frac{9}{18+9} = 0.333$, or 33%, responded. This corresponds to what we found in part (a). Younger patients seem to respond better than older patients.

c. Temperature is only moderately associated with response to treatment. The odds ratio is $OR = e^{-0.03884} = 0.96$ and the P-value is 0.069. For each additional degree in temp, the odds of response to treatment decreases by a factor of 0.96 (or about 4%).

```
 Coefficients:
             Estimate Std. Error z value Pr($>$|z|)
(Intercept) 38.55067   21.23644   1.815   0.0695 .
Temp        -0.03884    0.02135  -1.819   0.0689 .
```

Again, we have some freedom in choosing a temperature cutoff to make a two-way table. Since 990 is the median temp in this study, a new indicator of whether a subject's temperature is greater than 990 or not was used to construct a 2×2 table.

```
               Response
               no    yes
less than 991  14     15
991 or more    13      9
```

For the patients with lower temps, $\frac{15}{14+15} = 0.517$, or 51.7%, responded. For the patients with higher temps, $\frac{9}{13+9} = 0.409$, or 40.9%, responded. This corresponds to what we found in above. Patients with lower temps are more likely to patients that have higher temps.

9.42 a. Two-sample t-test results:
There is a statistically significant difference in mean GPA between those students who are first generation and those who are not (2.89 for first generation, 3.12 for those not first generation, $p = 0.02$).

```
Welch Two Sample t-test

data:  GPA by FirstGen
t = 2.4474, df = 31.406, p-value = 0.02017
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 0.03820979 0.41912630
sample estimates:
mean in group 0 mean in group 1
       3.122268        2.893600
```

b. Simple linear regression results:
The coefficient for the indicator FirstGen is significantly different from 0 ($\hat{\beta}_1 = -0.229$, $p = 0.02$).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.12227    0.03308  94.377   <2e-16 ***
FirstGen    -0.22867    0.09792  -2.335   0.0204 *
```

c. Logistic regression results:
The coefficient for GPA is significantly different from 0 ($\hat{\beta}_1 = -1.0381$, $p = 0.02$).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0751     1.3527   0.795    0.427
GPA          -1.0381     0.4567  -2.273    0.023 *
```

The P-value is identical for all three models; however, the point estimate differs because each model measures the relationship in different ways.

9.43 a. The proportion effective in each group are

$$\hat{p}_{THC} = \frac{36}{79} = 0.456 \qquad \hat{p}Pro = \frac{16}{78} = 0.205$$

b. Here is some output for predicting *Effective* using an indicator for THC.

```
Variable   Value      Count
Effective  Event         52
           Non-event    105
Patients   Total        157

                                                Odds     95% CI
Predictor      Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant   -1.35455  0.280409  -4.83  0.000
THC         1.17686  0.360087   3.27  0.001   3.24   1.60   6.57

Log-Likelihood = -94.028
Test that all slopes are zero: G = 11.345, DF = 1, P-Value = 0.001
```

The fitted logit equation is $log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.35455 + 1.17686THC$.

c. For the THC group, log(odds) $= -1.35455 + 1.1786(1) = -0.176$,
odds $= e^{-0.176} = 0.84$
probability $= e^{-0.176}/(1 + e^{-0.176}) = \frac{0.84}{1+0.84} = 0.456$

For the Prochlorperazine group, log(odds) $= -1.35455$
odds $= e^{-1.35} = 0.259$
probability $= e^{-1.35}/(1 + e^{-1.35}) = \frac{0.259}{1+0.259} = 0.206$

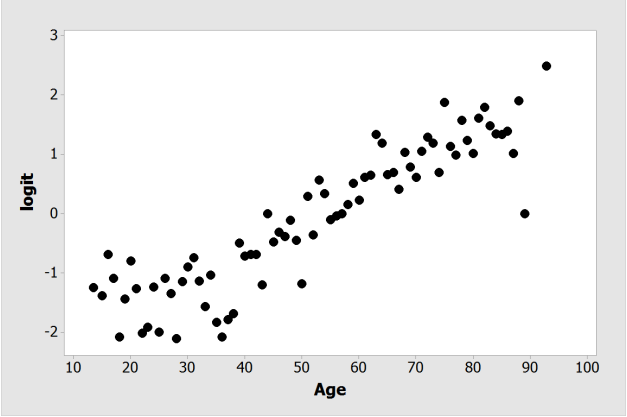
The probabilities match those calculated from the frequencies in the table.

d. The odds ratio is $e^{1.17686} = 3.24$. The estimated odds of having effective treatment for nausea are about 3.24 times higher when THC is used compared to prochlorperazine. A 95% confidence interval for the odds ratio is given in the computer output as (1.60, 6.57). We could compute this directly using the standard error in the output (SE = 0.36) with

$$e^{1.17686 \pm 1.96(0.36)} = 1.60 \text{ to } 6.57$$

e. To see if THC is more effective, we test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 > 0$. We can use either the P-value for the individual z-test for the coefficient of THC or the G-statistic (both P-values = 0.001). However, those are two-tailed tests, so the P-value for one tail is about 0.0005. This is a very small P-value, providing strong evidence that THC is more effective than prochlorperazine at preventing nausea with chemotherapy.

9.44 a. The empirical logit plot is reasonably linear. This suggests that the logistic model is appropriate.



b. The logistic equation is $logit = -2.779 + 0.05172Age$ and P-value is ≈ 0 so the relationship is statistically significant.

```
 Coefficients

Term          Coef  SE Coef    Z-Value  P-Value
Constant    -2.779    0.140     -19.92    0.000
Age        0.05172  0.00250      20.71    0.000
```

c. The probability of death increases as age increases. We see this from the positive slope 0.05172.

9.45 a. Here is some output for fitting a logistic regression model to predict *Result* based on *StartSpeed*.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.409761   0.473634  -2.976  0.00292 **
StartSpeed   0.021235   0.005367   3.957  7.6e-05 ***
---
    Null deviance: 4540.4  on 3401  degrees of freedom
Residual deviance: 4524.8  on 3400  degrees of freedom
```

The estimated intercept is $\widehat{\beta_0} = -1.4098$ and slope is $\widehat{\beta_1} = 0.021235$, so the logit form of the fitted model is

$$log(\widehat{odds}) = -1.4098 + 0.021235 StartSpeed$$

Converting this to the probability form gives

$$\widehat{\pi} = \frac{e^{-1.4098+0.021235 StartSpeed}}{1 + e^{-1.4098+0.021235 StartSpeed}}$$

b. Since the slope of the logit form of the model is positive ($\widehat{\beta_1} = 0.021235$), faster pitches appear to have a higher probability of giving positive results.

c. For $StartSpeed = 95$ we have

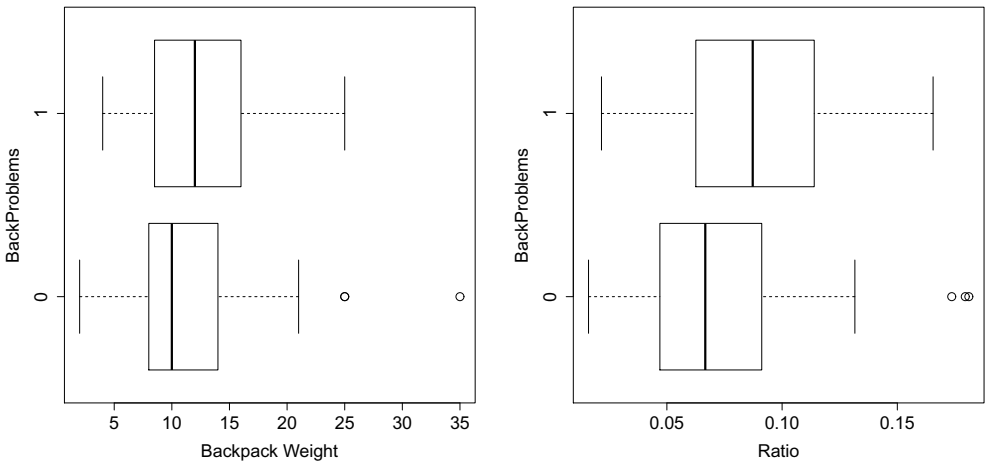
$$\widehat{\pi} = \frac{e^{-1.4098+0.021235(95)}}{1 + e^{-1.4098+0.021235(95)}} = \frac{e^{0.6075}}{1 + e^{0.6075}} = \frac{1.8359}{2.8359} = 0.647$$

For $StartSpeed = 75$ we have

$$\widehat{\pi} = \frac{e^{-1.4098+0.021235(75)}}{1 + e^{-1.4098+0.021235(75)}} = \frac{e^{0.6075}}{1 + e^{0.6075}} = \frac{1.2006}{2.2006} = 0.546$$

d. If we test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, the output shows $z = 3.957$ and P-value$= 0.000076$. This gives strong evidence that the slope is positive and that the speed of a pitch has some value in predicting the result.

9.46 In this analysis, we will explore the possibility of backpacks being responsible for back problems. The response variable will be whether or not a student has back problems. This is a binary variable, where 0 indicates that the person does not report back problems and 1 indicates that the person reported back problems. Of the 100 students, 32 reported having back problems, 68 did not. Two potential explanatory variables were considered: backpack weight and the ratio of backpack weight to body weight. Backpack weight is a quantitative variable, ranging from 2 to 35 pounds and averaging 11.7 pounds for students in the dataset. The following boxplot shows two outlier values with backpack weights much larger than expected. Both of these outlier values are in the group not reporting back problems. The median weight for the back-problem-free group appears to be lower than the median weight for those with back problems. The ratio of backpack weight to body weight ranges from 0.016 to 0.181, with a mean of 0.07713. It too displays outliers in the back-problem-free group. There is an even greater disparity in the median ratios than the median backpack weights.



In order to determine if backpacks are responsible for back problems, we will build a binary logistic regression model using *BackpackWeight* as an explanatory variable. Here is some output for that model.

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	-1.27118	0.496480	-2.56	0.010			
BackpackWeight	0.0435057	0.0369607	1.18	0.239	1.04	0.97	1.12

This gives a fitted equation for estimating the probability of back pain.

$$\hat{\pi} = \frac{e^{-1.27+0.0435BackpackWeight}}{1+e^{-1.27+0.0435BackpackWeight}}$$

According to this model, for every one pound increase in backpack weight, the odds that a person will have back problems increases by a factor of 1.04, or about 4% ($p = 0.239$, 95% CI 0.97, 1.12).

One explanation for why there is not a significant relationship between backpack weight and back pain is that the students in the sample vary greatly in weight, and a person's weight will also play a role in whether a heavy bag hurts a person's back. The smallest body weight is 105 pounds, and the largest body weight is 270 pounds, with a mean body weight of 153 pounds.

In order to incorporate a student's body weight with respect to the backpack weight, a binary logistic regression model using the *Ratio* of $BackpackWeight$ to $BodyWeight$ as an explanatory variable. Here is some output for that model.

						95% CI	
Predictor	Coef	SE Coef	Z	P	Odds Ratio	Lower	Upper
Constant	-1.60090	0.531022	-3.01	0.003			
Ratio	10.6681	5.95113	1.79	0.073	42964.54	0.37	4.99713E+09

Note that the coefficient is large so that the odds ratio is huge. This happens since the actual changes in *Ratio* are very small (in fact, the maximum ratio is 0.18, so it's hard to interpret what is meant by a change in the ratio as large as one). For this reason, we create a new variable ($RatioPct$), by multiplying each *Ratio* by 100. Thus we can think of $RatioPct$ as backpack weight as a percent of body weight.

Fitting the model with $RatioPct$ gives

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	-1.60090	0.531022	-3.01	0.003			
RatioPct	0.106681	0.0595113	1.79	0.073	1.11	0.99	1.25

This gives a fitted equation for estimating the probability of

$$\hat{\pi} = \frac{e^{-1.60+0.1067RatioPct}}{e^{-1.60+0.1067RatioPct}}$$

This model suggests that for every increase of 1% in the percentage ratio of backpack weight to body weight, the odds that a person will have back problems increases by a factor of 1.11 (or about 11%). However, we do not have statistically significant evidence ($p = 0.07303$, $z = 1.79$) that there is a strong relationship between the ratio of the backpack weight as a percentage of body weight and back pain.

Neither model provides statistically significant coefficients for their respective explanatory variables. In the next chapter, we will see more tools for constructing and comparing models like this including multiple logistic regression. Some statistics, such as AICs and deviances, will also be seen to be useful for comparing models.

To answer the research question, $BackpackWeight$ is not an especially strong predictor of back pain, however, our observations are in a direction that supports the initial hypothesis that heavy backpacks are associated with back problems.

9.47 a. We can do this directly using the 2×2 table:

```
      ride.alc.driver
female   no   yes
    no  4742 1224
   yes  3921 2745
```

The odds ratio for females RDD compared to males is

$$OR = \frac{2745/3921}{1224/4742} = \frac{2745(4742)}{1224(3921)} = 2.71$$

Or this can done using logistic regression with $female$ as a predictor for *ride.alc.driver*.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.35433    0.03206  -42.24   <2e-16
female       0.99777    0.04059   24.58   <2e-16
```

We get the odds ratio by exponentiating the coefficient of $female$, $OR = e^{0.99777} = 2.71$. This matches the odds ratio from the two-way table.

The odds that a female is riding with a drinker are about a 2.71 times greater than the odds for a male. The z-statistic for the female coefficient is very large ($z = 24.58$) and P-value ≈ 0, so the YRBS data provide strong support for the claim that young women are more likely to be riding with a drinking driver.

b. We created a *DriverLicense* variable by assuming that all 16-year-old students get their driver's license. (Is this a reasonable assumption?)

Here is a two-way table for RDD status by *DriverLicense*.

```
             ride.alc.driver
DriverLicense    no  yes
          no   3077 1231
         yes 1 6096 2929
```

The odds ratio for RDD comparing those who smoke to those who don't is

$$OR = \frac{2929/6096}{1231/3077} = \frac{2929(3077)}{6096(1231)} = 1.20$$

Following is some output using a logistic regression model to predict *ride.alc.driver* using *DriverLicense*.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.91613    0.03372 -27.165  < 2e-16
DriverLicense  0.18316    0.04053   4.519 6.22e-06
```

These results support the hypothesis that kids get more reckless after they get their driver's license (or at least after they turn 16). The odds of a respondent who has their driver's license riding with a drinking driver are nearly 20% higher than the odds for a respondent under 16. The P-value ≈ 0 for the test of the *DriverLicense* coefficient demonstrates that this large a difference in the odds would be very unlikely to happen by random chance.

c. Here is a two-way table for RDD status by *smoke*.

```
     ride.alc.driver
smoke    no  yes
   no  4940 1113
  yes  4027 2919
```

The odds ratio for RDD comparing those with to those without a license is

$$OR = \frac{2919/4027}{1113/4940} = \frac{2919(4940)}{4027(1113)} = 3.22$$

Here is some output using a logistic regression model to predict *ride.alc.driver* using *smoke*.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.49031    0.03318  -44.92   <2e-16
smoke        1.16853    0.04113   28.41   <2e-16
```

The logistic regression results support the claim that smokers are more likely to ride with a drinking driver. Specifically, a smoker is over 3 times as likely to RDD(OR = 3.22, $p < 0.0001$).

9.48 The response variable is whether a movie received 3 or more stars. This variable is binary: 0 indicates that the movie is bad (below 3 stars) and 1 indicates that the movie is good (3 stars or above). There are 69 bad movies and 31 good movies in the dataset.

Logistic regressions were performed with a single covariate from each of the covariates: $Year, Time, Cast.$ and *Description*. Here is a summary of key quantities for fitting those models:

Variable	coeff	OR	P-value	95% CI lower	95% CI upper
Year	−0.0009821	0.999	0.935	0.976	1.023
Time	0.04518	1.046	0.001	1.019	1.078
Cast	0.2106	1.234	0.057	0.997	1.53
Description	0.20070	1.222	0.013	1.043	1.433

The running *Time* of the movie was most predictive of the *Good* or *Bad* rating ($p = 0.001$) followed by the number of lines in the *Description* ($p = 0.013$). Here is a short summary for each of the predictors individually:

- For each additional minute, the odds of a good rating for a movie increases by approximately 5%.
- For each additional line of description, the odds of a good rating increase by a factor of 1.22.
- The number of *Cast* members mentioned exhibited a moderate, yet not statistically significant, association with the *Good* ($p = 0.057$).
- There seemed to be little or no association with the *Year* the movie was made and the *Good* rating ($p = 0.935$).

9.49 We fit the model $logit(\pi) = \beta_0 + \beta_1 Peaceworks$, where π is the proportion of letters returned and $Peaceworks = 1$ for letters addressed to Iowa Peaceworks and $Peaceworks = 0$ for letters addressed to the Friends of the Confederacy. We fit this model separately for each location to obtain the three sets of output below.

```
DES MOINES:
Variable  Value  Count
Returned  1         45  (Event)
          0         35
          Total     80
                                                 Odds     95% CI
Predictor        Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant    -0.302281  0.319847  -0.95  0.345
Peaceworks    1.14958  0.470478   2.44  0.015   3.16   1.26   7.94

Log-Likelihood = -51.709
Test that all slopes are zero: G = 6.233, DF = 1, P-Value = 0.013

GRINNELL TOWN:
Variable  Value  Count
Returned  1         32  (Event)
          0          8
          Total     40
                                               Odds     95% CI
Predictor       Coef   SE Coef     Z      P  Ratio  Lower  Upper
Constant    0.619039  0.468807  1.32  0.187
Peaceworks   2.32540   1.12800  2.06  0.039  10.23   1.12  93.34

Log-Likelihood = -16.919
```

```
Test that all slopes are zero: G = 6.194, DF = 1, P-Value = 0.013

GRINNELL CAMPUS:
Variable  Value  Count
Returned  1          5  (Event)
          0         15
          Total     20
                                                Odds     95% CI
Predictor       Coef  SE Coef      Z      P  Ratio  Lower  Upper
Constant    -2.19722  1.05409  -2.08  0.037
Peaceworks   1.79176  1.23603   1.45  0.147   6.00   0.53  67.65

Log-Likelihood = -9.981
Test that all slopes are zero: G = 2.532, DF = 1, P-Value = 0.112
```

In Des Moines, we have fairly strong evidence (P-value $= 0.015$) that the coefficient of the *Peaceworks* indicator is different from zero. The fact that the estimated coefficient is positive suggests that letters addressed to Peaceworks are more likely to be returned (when "lost" in Des Moines). From the estimated odds ratio, we see that the odds of a Peaceworks letter being returned are about 3 times greater than for Confederacy.

In the town of Grinnell, we see similar strength results (P-value $= 0.039$) for the effectiveness of the Peaceworks indicator, although the sample size is smaller. The return rate is higher in the town of Grinnell (32 out of 40), and the estimated odds ratio is over 10.

The sample size is even smaller at the Grinnell campus—only 5 of the 20 "lost" letters were returned. Although the estimated coefficient of *Peaceworks* is also positive, the P-value (0.147) is not small enough for this evidence to be considered strong. The sample size and number of letters returned are both too small for the Grinnell campus to be able to tell much about the comparison for that population.

9.50 The question suggests running the logistic regression model $\log(\frac{\pi}{1+\pi}) = \beta_0 + \beta_1 Inform + \epsilon$, where π is the proportion of students who have voted (coded as $Participate = 1$). Here is some output for fitting this model to the data in **Political**.

```
Variable     Value  Count
Participate  1         33  (Event)
             0         22
             Total     55
                                                Odds     95% CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant   -0.817108   1.01454  -0.81  0.421
Inform      0.411878  0.332044   1.24  0.215   1.51   0.79   2.89
```

```
Log-Likelihood = -36.206
Test that all slopes are zero: G = 1.618, DF = 1, P-Value = 0.203
```

To see if there is evidence that "better informed citizens tend to be the ones who vote," we test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 > 0$. The P-value from the output is 0.215, but that is for a two-tailed test. Since the coefficient is in the direction of the alternative, the one-tail P-value is half that amount, so P-value = 0.1075. This is not small enough to reject H_0, even at a 10% significance level, so we don't have sufficient evidence in this sample to conclude that better-informed students are more likely to vote.