Chapter 10 Solutions

10.1 a. The odds a smoker rides with a drinking driver is $(38/62) = 0.613$, whereas the odds a nonsmoker rides with a drinking driver is $(13/87) = 0.149$.

b. The ratio of the smoker odds to the nonsmoker odds gives an odds ratio of $0.613/0.149{=}0.411$.

c. When comparing smokers to nonsmokers the unadjusted OR does not take into account how these groups may differ with respect to SES, substance abuse, and demographics. The adjusted OR makes the comparison between the smokers and nonsmokers given that we are comparing groups with comparable demographics, SES, and identified substance abuse behavior. Thus the comparison of smokers to nonsmokers is not confounded by these other differences, to the extent that is possible.

10.2 a. The odds of back problems for a female is $(44/56) = 0.786$, whereas the odds of back problems for a male is $(18/82) = 0.220$.

b. The ratio of female odds to male odds of back problems gives an odds ratio of $0.786/0.220 = 3.57$.

c. When comparing females to males the unadjusted OR does not take into account how the sexes differ with respect to backpack weight and body weight. The adjusted OR makes the comparison between the females and males given that we are comparing people of comparable backpack weight and body weight. Thus the comparison of females to males is not confounded by differences in backpack weight and body weight.

10.3 The plot suggests the model $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \beta_2 Group1$, where $Group1$ is 1 for $Group = 1$ and 0 for $Group = 2$.

10.4 The plot suggests the model $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \beta_2 Group1 + \beta_3 Group2$, where $Group1$ is 1 for $Group = 1$ and 0 for the other two groups and $Group2$ is 1 for $Group = 2$ and zero for the other two groups.

10.5 The plot suggests the model $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \beta_2 Group1 + \beta_3 X \cdot Group1$, where $Group1$ is 1 for $Group = 1$ and 0 for $Group = 2$.

10.6 The plot suggests the model $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \beta_2 Group1 + \beta_3 X \cdot Group1 + \beta_4 X \cdot Group2$, where $Group1$ is 1 for $Group = 1$ and 0 otherwise and where $Group2$ is 1 for $Group = 2$ and 0 otherwise.

10.7 The curve is concave up, so try squaring.

10.8 The curve is concave down, so a power less than 1 or a logarithm should be tried.

10.9 Because the shift between $Y = 0$ and $Y = 1$ looks the same for both levels of A and because the shift between $A =$ Low and $A =$ High looks the same for $Y = 0$ as it does for $Y = 1$, we can use an additive model: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 Ahigh + \beta_2 X$, where $Ahigh$ is 1 for the high level of A (and 0 for the low level of A).

10.10 Because the shift between $Y = 0$ and $Y = 1$ is different for the two levels of A, we should use an interaction model: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 Ahigh + \beta_2 X + \beta_3 Ahigh \cdot X$, where $Ahigh$ is 1 for the high level of A (and 0 for the low level of A).

10.11 a. Restricted model: $\text{logit}(\pi) = \beta_0 + \beta_1 MCAT + \beta_2 Sex$; Full model: $\text{logit}(\pi) = \beta_0 + \beta_1 MCAT + \beta_2 Sex + \beta_3 GPA$

b. Restricted model: $\text{logit}(\pi) = \beta_0 + \beta_1 MCAT + \beta_2 Sex$; Full model: $\text{logit}(\pi) = \beta_0 + \beta_1 MCAT + \beta_2 Sex + \beta_3 Sex \cdot MCAT$

10.12 a. Restricted model: $\text{logit}(\pi) = \beta_0 + \beta_1 GPA + \beta_2 Sex$; Full model: $\text{logit}(\pi) = \beta_0 + \beta_1 GPA + \beta_2 Sex + \beta_3 MCAT$

b. Restricted model: $\text{logit}(\pi) = \beta_0 + \beta_1 GPA + \beta_2 MCAT + \beta_3 Sex$; Full model: $\text{logit}(\pi) = \beta_0 + \beta_1 GPA + \beta_2 MCAT + \beta_3 Sex + \beta_4 Sex \cdot GPA$

10.13 The coefficient of $LeftHandLow$ is positive, which means that using the “left hand low” grip increases the odds of making a putt. Specifically, having “left hand low” increases the log odds by 0.46. Note that this is almost as much as the effect of a putt being one foot shorter (when *Length* decreases by one foot, the log odds of success increase by 0.56). That is, the fitted model suggests that using the “left hand low” grip is equivalent to having a putt be about 10 inches shorter (since (0.46/0.56)12 is close to 10).

10.14 The coefficient of *Sex_F* is positive, which means that being female increases the odds of being accepted into medical school. Specifically, being female increases the log odds by 1.066.

10.15 a. Using 5 for the value of *Length* and 1 for $LowLeftHand$ and solving gives

$$\hat{\pi} = \frac{e^{3.2262-0.5604(5)+0.4616(1)}}{1+e^{3.2262-0.5604(5)+0.4616(1)}} = 0.7080$$

b. Using 5 for the value of *Length* and 0 for $LowLeftHand$ and solving gives

$$\hat{\pi} = \frac{e^{3.2262-0.5604(5)+0.4616(0)}}{1+e^{3.2262-0.5604(5)+0.4616(0)}} = 0.6045$$

c. Yes. In Exercise **??** we found that the odds of making a putt increased if the “left hand low” grip was used. In this case, if we look at a 5 foot putt, the probability (and therefore the odds) of making the putt are larger when using the “left hand low” grip.

10.16 a. Using 40 for the value of $MCAT$ and 1 for Sex_F and solving gives

$$\hat{\pi} = \frac{e^{-10.02+0.2669(40)+1.066(1)}}{1+e^{-10.02+0.2669(40)+1.066(1)}} = 0.8484$$

b. Using 40 for the value of $MCAT$ and 0 for Sex_F and solving gives

$$\hat{\pi} = \frac{e^{-10.02+0.2669(40)+1.066(0)}}{1+e^{-10.02+0.2669(40)+1.066(0)}} = 0.6584$$

c. Yes. In Exercise **??** we found that the odds of being accepted into medical school were higher for females. In this case, if we look at an MCAT score of 40, the probability (and therefore the odds) of being accepted into medical school are higher for females.

10.17 a. We need a model that includes an interaction between *logContr* and *Dem*. Following is some output:

```
Logistic Regression Table
                                                  Odds        95% CI
Predictor          Coef  SE Coef      Z      P  Ratio  Lower        Upper
Constant       -10.1636  5.40134  -1.88  0.060
LogContr        3.00151  1.35714   2.21  0.027  20.12   1.41       287.57
Dem             2.54364  5.97423   0.43  0.670  12.73   0.00   1548673.91
LogContr*Dem   -1.08829  1.51459  -0.72  0.472   0.34   0.02         6.56

Log-Likelihood = -43.391
Test that all slopes are zero: G = 46.031, DF = 3, P-Value = 0.000
```

The P-value for testing the interaction is 0.472, so the interaction term can be dropped from the model. The effect of *logContr* on $Vote$ appears to be the same for Democrats and Republicans.

b. The output for the reduced model is

```
Logistic Regression Table
                                                Odds     95% CI
Predictor       Coef   SE Coef      Z      P   Ratio  Lower  Upper
Constant    -6.84015   2.49032  -2.75  0.006
LogContr     2.16589  0.613083   3.53  0.000    8.72   2.62  29.01
Dem         -1.73280  0.580447  -2.99  0.003    0.18   0.06   0.55

Log-Likelihood = -43.668
Test that all slopes are zero: G = 45.477, DF = 2, P-Value = 0.000
```

For the full model, $-2\log(L) = -2(-43.391) = 86.782$ and for the restricted model $-2\log(L) = -2(-43.668) = 87.336$. This leads to a difference of 0.554 with 1 degree of freedom. The P-value is 0.4567.

c. The P-values are similar, but they are not the same. This is not surprising: The Wald z-test P-value is based on a normal approximation, and the drop-in-deviance P-value is based on a chi-square approximation. In the limit, these will agree (with $Z^2 = \chi^2$), but for finite samples they will tend to be different, as they are here.

10.18 a. We need a model that includes an interaction between $dilateDiff$ and Sex. Following is some output:

```
Logistic Regression Table
Predictor           Coef    SE Coef        Z         P
Constant         -1.3035     0.4029   -3.235   0.00122
dilateDiff        3.8931     1.3793    2.822   0.00477
SexM              1.3088     0.5049    2.592   0.00953
dilateDiff*SexM  -0.9754     1.7037   -0.573   0.56698
Log-Likelihood = -53.02
Test that all slopes are zero: G = 31.093, DF = 3, P-Value = 0.000
```

The P-value for testing the interaction is 0.567, so the interaction term can be dropped from the model. The effect of $dilateDiff$ on Gay appears to be the same for males and females.

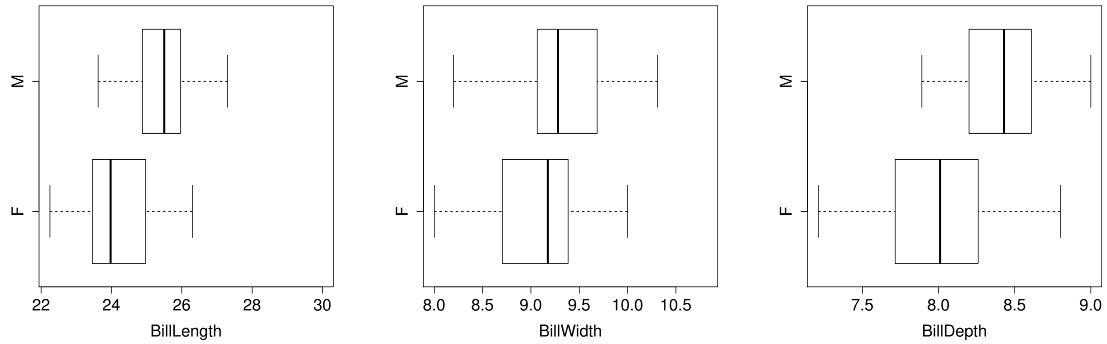
b. The output for the reduced model is

```
Logistic Regression Table
Predictor       Coef    SE Coef        Z          P
Constant     -1.2656     0.3820   -3.313     0.0009
dilateDiff    3.2908     0.8180    4.023   5.75e-05
SexM          1.2897     0.4953    2.604     0.0092
Log-Likelihood = -53.19
Test that all slopes are zero: G = 30.758, DF = 2, P-Value = 0.000
```

For the full model, $-2\log(L) = -2(-53.02) = 106.04$ and for the restricted model $-2\log(L) = -2(-53.19) = 106.38$. This leads to a difference of 0.34 with 1 degree of freedom. The P-value is 0.5625.

c. The P-values are similar, but they are not the same. This is not surprising: The Wald z-test P-value is based on a normal approximation, and the drop in deviance P-value is based on a chi-square approximation. In the limit, these will agree (with $Z^2 = \chi^2$), but for finite samples they will tend to be different, as they are here.

10.19 a. All three pairs of parallel boxplots show larger measurements for males than for females. The biggest shifts are for *BillLength* and *BillDepth*, whereas *BillWidth* shows the most overlap between males and females. This suggests that *BillWidth* is the weakest predictor of sex. *BillLength* shows the least overlap between the male and female plots so it is the strongest predictor.



b. Here is the output from fitting the model:

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) -48.1113     8.6998  -5.530  3.2e-08 ***
BillDepth     2.9367     0.8311   3.533 0.000410 ***
BillWidth    -0.2405     0.4698  -0.512 0.608756
BillLength    1.0537     0.2879   3.660 0.000253 ***
```

The P-value for *BillWidth* is the largest. The other two predictors have roughly the same predictive ability.

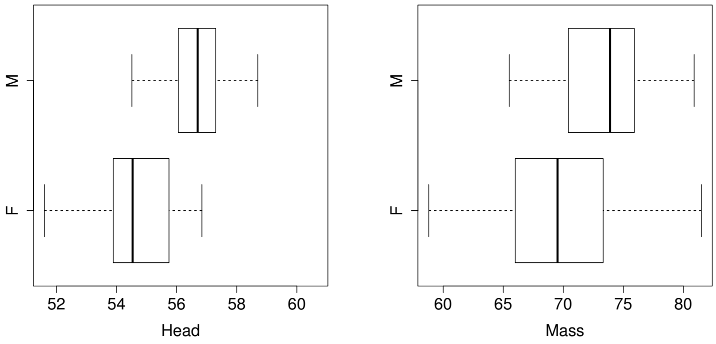
c. Here is the output from fitting the model:

```
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.5393     3.3045  -1.979   0.0478 *
BillWidth     0.7149     0.3581   1.996   0.0459 *
```

The P-value for *BillWidth* is just below 0.05, showing that *BillWidth* has a clear relationship with sex.

d. *BillWidth* does not emerge as a strong predictor in the multiple logistic model because of collinearity between the three predictors. Much of the information in *BillWidth* about males and females is already accounted for in *BillLength* and *BillDepth*.

10.20 a. Both pairs of parallel boxplots show larger measurements for males than for females. The bigger shift is for *Head*, which suggests that *Head* is the stronger predictor of *Sex*.



b. Here is the output from fitting the model:

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.17171    3.24990  -3.745 0.000180 ***
Mass          0.17073    0.04528   3.770 0.000163 ***
```

The P-value for $Mass$ is well below any reasonable significance level (e.g., 0.01).

c. Here is the output from fitting the model:

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -89.04780   15.87609  -5.609 2.04e-08 ***
Mass         -0.07609    0.06815  -1.116    0.264
Head          1.69650    0.32271   5.257 1.46e-07 ***
```

Now the P-value for $Mass$ is fairly large (0.264), showing that the added predictive power of $Mass$ is modest if $Head$ is already included in the model.

10.21 a. We need a model that includes an interaction between GPA and Sex. Here is some output:

```
Logistic Regression Table
                                                          95% CI
Predictor        Coef  SE Coef      Z      P  Odds Ratio  Lower        Upper
Constant     -24.3852  10.0818  -2.42  0.016
SexCode       4.90102  13.1550   0.37  0.709      134.43   0.00  2.11875E+13
GPA           7.08340  2.89044   2.45  0.014     1192.01   4.13    344052.31
GPA*SexCode  -1.70876  3.70906  -0.46  0.645        0.18   0.00       260.05

Log-Likelihood = -26.864
Test that all slopes are zero: G = 22.062, DF = 3, P-Value = 0.000
```

The P-value for testing the interaction is 0.645, so the interaction term can be dropped from the model. The effect of GPA on *Acceptance* appears to be the same for men and women.

b. The output for the restricted model is given below.

```
Logistic Regression Table
                                                            95% CI
Predictor       Coef   SE Coef      Z      P  Odds Ratio  Lower     Upper
Constant    -21.0680   6.40252  -3.29  0.001
SexCode     -1.16970  0.717834  -1.63  0.103        0.31   0.08      1.27
GPA          6.13242   1.82832   3.35  0.001      460.55  12.79  16578.45

Log-Likelihood = -26.972
Test that all slopes are zero: G = 21.846, DF = 2, P-Value = 0.000
```

For the full model, $-2\log(L) = -2(-26.864) = 53.728$, and for the restricted model, $-2\log(L) = -2(-26.972) = 53.944$. This leads to a difference of 0.216 with 1 degree of freedom. The P-value is 0.642.

c. The P-values are similar, but they are not the same. This is not surprising: The Wald z-test P-value is based on a normal approximation, and the drop-in-deviance P-value is based on a chi-square approximation. In the limit, these will agree (with $Z^2 = \chi^2$), but for finite samples they will tend to be different, as they are here.

10.22 a. We need a model that includes an interaction between $MCAT$ and Sex. Here is some output:

```
Logistic Regression Table

Predictor       Coef   SE Coef       Z      P
Constant     -6.1804    4.3247  -1.429  0.153
SexCode      -7.2122    7.1083  -1.015  0.310
MCAT          0.1887    0.1212   1.557  0.119
MCAT*SexCode 0.1697     0.1946   0.872  0.383

Log-Likelihood = -30.462
Test that all slopes are zero: G = 14.867, DF = 3, P-Value = 0.0019
```

The P-value for testing the interaction is 0.383, so the interaction term can be dropped from the model. The effect of $MCAT$ on *Acceptance* appears to be the same for men and women.

b. The output for the restricted model is given below.

```
Logistic Regression Table

Predictor    Coef   SE Coef       Z        P
Constant -8.95142   3.35391  -2.669  0.00761
SexCode  -1.06624   0.63335  -1.684  0.09228
MCAT      0.26689   0.09407   2.837  0.00455

Log-Likelihood = -30.8555
Test that all slopes are zero: G = 14.08, DF = 2, P-Value = 0.0009
```

For the full model, $-2\log(L) = -2(-30.462) = 60.924$, and for the restricted model, $-2\log(L) = -2(-30.8555) = 61.711$. This leads to a difference of 0.787 with 1 degree of freedom. The P-value is 0.375.

c. The P-values are similar, but they are not the same. This is not surprising: The Wald z-test P-value is based on a normal approximation, and the drop in deviance P-value is based on a chi-square approximation. In the limit, these will agree (with $Z^2 = \chi^2$), but for finite samples they will tend to be different, as they are here.

10.23 a. Here is some output from fitting a two-predictor logistic regression model to predict the probability of Titanic survival based on *Age* and *SexCode*.

```
Variable  Value  Count
Survived  1        313  (Event)
          0        443
          Total    756
                                                   Odds     95% CI
Predictor         Coef   SE Coef       Z      P  Ratio  Lower  Upper
Constant      -1.15984  0.219651  -5.28  0.000
Age         -0.0063520  0.0061869  -1.03  0.305   0.99   0.98   1.01
SexCode        2.46600  0.178455  13.82  0.000  11.78   8.30  16.71

Log-Likelihood = -397.793
Test that all slopes are zero: G = 229.987, DF = 2, P-Value = 0.000
```

The logit form of the model is

$$log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.160 - 0.00635Age + 2.466SexCode$$

The probability form of the model is

$$\hat{\pi} = \frac{e^{-1.160-0.00635Age+2.466SexCode}}{1+e^{-1.160-0.00635Age+2.466SexCode}}$$

b. *Age* is not a effective predictor in this model (P-value $= 0.305$), but *SexCode* is very effective (P-value ≈ 0).

c. We use the logit form of the model to estimate the log(odds) when $Age = 18$ and $SexCode = 0$

$$log(\widehat{odds}) = -1.160 - 0.00635(18) + 2.466(0) = -1.2743$$

The predicted odds are

$$\widehat{odds} = e^{-1.2743} = 0.2796$$

The probability that an 18 year-old-male would survive is

$$\hat{\pi} = \frac{e^{-1.160-0.00635(18)+2.466(0)}}{1 + e^{-1.160-0.00635(18)+2.466(0)}} = \frac{0.2796}{1 + 0.2796} = 0.219$$

d. To estimate the odds for an 18-year-old female, we use $Age = 18$ and $SexCode = 1$:

$$log(\widehat{odds}) = -1.160 - 0.00635(18) + 2.466(1) = 1.1917$$

The predicted odds are

$$\widehat{odds} = e^{1.1917} = 3.293$$

The probability that an 18-year-old female would survive is

$$\hat{\pi} = \frac{e^{-1.160-0.00635(18)+2.466(1)}}{1 + e^{-1.160-0.00635(18)+2.466(1)}} = \frac{3.293}{1 + 3.293} = 0.767$$

The odds ratio for a female compared to a male is

$$OR = \frac{odds_f}{odds_m} = \frac{3.293}{0.2796} = 11.78$$

The odds that an 18-year-old female survived are 11.78 times greater than the survival odds for an 18-year-old male. Note that this matches the odds ratio for *SexCode* in the output.

e. For a 50-year-old male

$$\begin{aligned} \widehat{odds} &= e^{-1.160-0.00635(50)+2.466(0)} = e^{-1.4775} = 0.2282 \\ \hat{\pi} &= \frac{0.2282}{1 + 0.2282} = 0.186 \end{aligned}$$

For a 50-year-old female

$$\begin{aligned} \widehat{odds} &= e^{-1.160-0.00635(50)+2.466(1)} = e^{0.9885} = 2.687 \\ \hat{\pi} &= \frac{2.8687}{1 + 2.687} = 0.729 \end{aligned}$$

For 50-year-olds, the ratio of survival odds is

$$OR = \frac{odds_f}{odds_m} = \frac{2.687}{0.2282} = 11.77$$

f. Up to round-off differences, the ratio of survival odds for females to males is the same (11.78) for every age.

10.24 a. We fit the model $logit(\pi) = \beta_0 + \beta_1 Age + \beta_2 SexCode + \beta_3 Age \cdot SexCode$ to produce the output below.

```
Variable  Value  Count
Survived  1        313  (Event)
          0        443
          Total    756
                                                      Odds     95% CI
Predictor         Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant    -0.298750   0.277699  -1.08  0.282
Age        -0.0363669  0.0092629  -3.93  0.000   0.96   0.95   0.98
SexCode      0.599858   0.408050   1.47  0.142   1.82   0.82   4.05
AgeSex      0.0657179  0.0136862   4.80  0.000   1.07   1.04   1.10

Log-Likelihood = -385.278
Test that all slopes are zero: G = 255.017, DF = 3, P-Value = 0.000
```

The constant term and coefficient of *Age* are the intercept and slope for the linear model for log(odds) for males (when $Sex = 0$).

$$\text{Males: } logit(\hat{\pi}) = -0.2988 - 0.0364 Age$$

The coefficients for *SexCode* and the interaction $Age \cdot SexCode$ indicate how the intercept and slope, respectively, change for the linear model for females.

$$\text{Females: } logit(\hat{\pi}) = -0.2988 - 0.0364 Age + 0.600(1) + 0.0657 Age(1) = 0.3012 + 0.0293 Age$$

b. For the model in (a), we test $H_0 : \beta_1 = \beta_3 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_3 \neq 0$. The reduced model uses just the *SexCode* predictor to give the output below. Note: There are quite a few missing values for the *Age* variable. To make a proper comparison of nested models, those cases are also deleted when running the model for *SexCode* alone.

```
Variable  Value  Count
Survived  1        313  (Event)
          0        443
          Total    756
                                                  Odds     95% CI
Predictor       Coef   SE Coef       Z      P  Ratio  Lower  Upper
Constant    -1.35455  0.114476  -11.83  0.000
SexCode      2.47176  0.178319   13.86  0.000  11.84   8.35  16.80

Log-Likelihood = -398.322
Test that all slopes are zero: G = 228.929, DF = 1, P-Value = 0.000
```

We compute a chi-square statistic by comparing G values between the full and reduced models.

$$\chi^2 = G_{Full} - G_{Reduced} = 255.017 - 228.929 = 26.088$$

We find a P-value using the upper tail of a chi-square distribution with 2 degrees of freedom, P-value $= P(\chi^2_2 > 26.088) = 0.000002$. This is a very small P-value and provides strong evidence that at least one of the terms involving *Age* is important in this model.

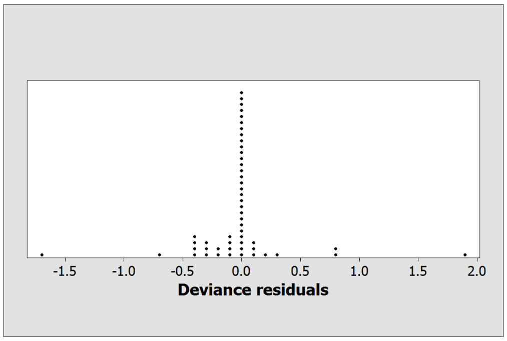
10.25 a. Here is some ouput for fitting a multiple regression model to predict *ObamaWin* based on *Dem.Rep*, *HS*, *BA*, and *Income*.

```
Variable   Value  Count
ObamaWin   1         29  (Event)
           0         22
           Total     51
                                                       Odds      95% CI
Predictor        Coef    SE Coef      Z      P   Ratio   Lower  Upper
Constant     -53.6135    37.7332  -1.42  0.155
Dem.Rep      0.635334   0.272602   2.33  0.020    1.89    1.11   3.22
HS           0.151444   0.389400   0.39  0.697    1.16    0.54   2.50
BA           0.521433   0.394681   1.32  0.186    1.68    0.78   3.65
Income      0.0006445  0.0004828   1.33  0.182    1.00    1.00   1.00

Log-Likelihood = -4.863
Test that all slopes are zero: G = 60.012, DF = 4, P-Value = 0.000
```

The strongest predictor in this model is *Dem.Rep* with the only small P-value (0.020) for testing a coefficient.

b. Each of the other predictors, *HS* (P-value $= 0.697$), *BA* (P-value $= 0.186$), and *Income* (P-value $= 0.182$), are not strongly related to *ObamaWin* in this model.

c. Here is a dotplot of the deviance residuals.



The state with a large positive deviance residual is Indiana (1.938). The state with a large negative deviance residual is Missouri (−1.746).

d. At the first step, we eliminate HS (P-value = 0.697) from the model to give the output below.

```
                                                  Odds     95% CI
Predictor        Coef    SE Coef       Z       P  Ratio  Lower  Upper
Constant    -41.7249    18.5953   -2.24  0.025
Dem.Rep     0.615488   0.253717    2.43  0.015   1.85   1.13   3.04
BA          0.573496   0.373035    1.54  0.124   1.77   0.85   3.69
Income     0.0006371  0.0004605    1.38  0.167   1.00   1.00   1.00

Log-Likelihood = -4.944
Test that all slopes are zero: G = 59.848, DF = 3, P-Value = 0.000
```

Now the weakest predictor is $Income$ (P-value = 0.167), so we eliminate it to give

```
                                              Odds     95% CI
Predictor       Coef   SE Coef       Z       P  Ratio  Lower  Upper
Constant   -21.9388   8.94576   -2.45  0.014
Dem.Rep    0.511617  0.198333    2.58  0.010   1.67   1.13   2.46
BA         0.698647  0.313005    2.23  0.026   2.01   1.09   3.71

Log-Likelihood = -7.195
Test that all slopes are zero: G = 55.348, DF = 2, P-Value = 0.000
```

In this two-predictor model, the P-values for both coefficients, *Dem.Rep* (0.010) and BA (0.026), are less than 0.10, so we stop and call this our final model.

$$logit(\hat{\pi}) = -21.94 + 0.5116 Dem.Rep + 0.6986 BA$$

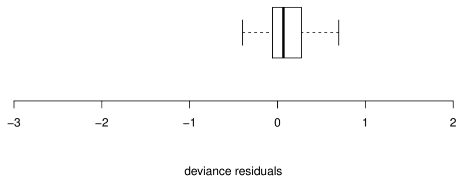
10.26 a. Here is some ouput for fitting a multiple regression model to predict $TrumpWin$ based on *Dem.Rep*, HS, BA, and *Income*.

```
Predictor   Coef   SE Coef       Z      P
Constant  11.879    19.324    0.61  0.539
Dem.Rep   -0.252     0.113   -2.24  0.025
HS         0.066     0.251    0.26  0.792
BA        -0.275     0.273   -1.01  0.314
Income -0.000178  0.000144   -1.24  0.214
Log-Likelihood = -9.778
Test that all slopes are zero: G = 47.746, DF = 4, P-Value = 0.000
```

The strongest predictor in this model is *Dem.Rep* with the only small P-value (0.025) for testing a coefficient.

b. Each of the other predictors, HS (P-value = 0.792), BA (P-value = 0.314), and $Income$ (P-value = 0.214), are not strongly related to $TrumpWin$ in this model.

c. Here is a dotplot of the deviance residuals.



The state with the largest positive deviance residual is Pennsylvania (1.35). The state with the largest negative deviance residual is Nevada (−2.45).

d. At the first step, we eliminate HS (P-value = 0.792) from the model to give the output below.

```
Predictor    Coef  SE Coef       Z       P
Constant   16.765    6.319    2.65  0.0079
Dem.Rep    -0.254   0.1099   -2.31   0.021
BA         -0.264    0.267   -0.99   0.325
Income  -0.000166 0.000134   -1.24   0.213
Log-Likelihood = -9.812
Test that all slopes are zero: G = 47.678, DF = 3, P-Value = 0.000
```

Now the weakest predictor is BA (P-value = 0.325), so we eliminate it to give

```
Predictor   Coef   SE Coef      Z     P
Constant  14.584     5.308  2.75 0.006
Dem.Rep   -0.340     0.113 -2.73 0.006
Income  -0.00027  0.000099 -2.71 0.0067
Log-Likelihood = -10.397
Test that all slopes are zero: G = 46.507, DF = 2, P-Value = 0.000
```

In this two-predictor model, the P-values for both coefficients, $Dem.Rep$ (0.006) and $Income$ (0.0067), are less than 0.10, so we stop and call this our final model.

$$logit(\hat{\pi}) = 14.58 - 0.340 Dem.Rep - 0.00027 Income$$

10.27 a. Here is some output for fitting $logit(\pi) = \beta_0 + \beta_1 DJIAch + \beta_2 lagNik$.

```
Variable  Value  Count
Up        1         29  (Event)
          0         27
          Total     56
                                                      Odds     95% CI
Predictor         Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant   -0.0272749   0.316434  -0.09  0.931
DJIAch      0.0134848  0.0042225   3.19  0.001   1.01   1.01   1.02
lagNik     -0.0038203  0.0022833  -1.67  0.094   1.00   0.99   1.00

Log-Likelihood = -29.987
Test that all slopes are zero: G = 17.587, DF = 2, P-Value = 0.000
```

The Dow Jones change predictor is important in this model (P-value = 0.001), while the previous day's Nikkei change is not so important (P-value = 0.094) and would only be considered significant at a 10% level.

b. When $DJIAch = 0$ and $lagNik = 0$, the predicted log(odds) for the Nikkei going up is

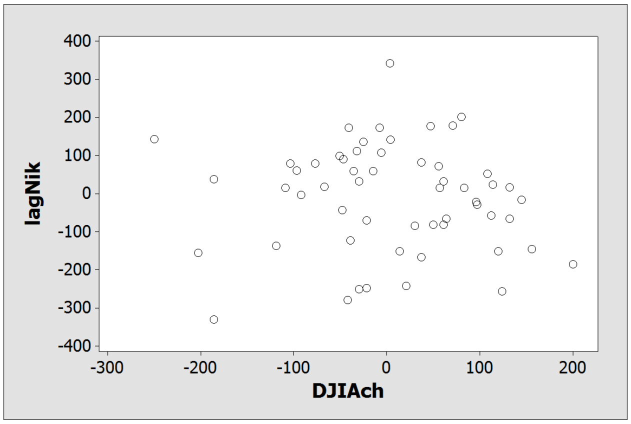
$$log(\widehat{odds}) = -0.0273 + 0.0135(0) - 0.0038(0) = -0.0273$$

The odds of going up are $e^{-0.0273} = 0.973$, which converts to a probability of

$$\hat{\pi} = \frac{0.973}{1 + 0.973} = 0.493$$

If neither market changed the previous day, we would expect about a 49.3% chance that the Nikkei 225 would go up.

c. The correlation between $DJIAch$ and $lagNik$ for the 56 sample points is -0.068, which is not very far from zero (P-value is 0.620 when testing if $\rho = 0$). The scatterplot below of $lagNik$ versus $DJIAch$ shows no consistent relationship, so we should not be concerned about multicollinearity for these two predictors.



10.28 a. Here is some output for using the six predictors to model the probability a patient responds to treatment.

```
Variable  Value  Count
Resp      1         24  (Event)
          0         27
          Total     51
                                                         Odds     95% CI
Predictor        Coef    SE Coef      Z       P  Ratio  Lower  Upper
Constant      108.331    41.8431   2.59  0.010
Age        -0.0623121  0.0274613  -2.27  0.023   0.94   0.89   0.99
Smear      -0.0046898  0.0400465  -0.12  0.907   1.00   0.92   1.08
Infil       0.0310448  0.0378937   0.82  0.413   1.03   0.96   1.11
Index        0.372812   0.132467   2.81  0.005   1.45   1.12   1.88
Blasts      0.0326739  0.0460503   0.71  0.478   1.03   0.94   1.13
Temp        -0.111622  0.0426300  -2.62  0.009   0.89   0.82   0.97

Log-Likelihood = -19.638
Test that all slopes are zero: G = 31.249, DF = 6, P-Value = 0.000
```

Three of the predictors, *Age* (P-value = 0.023), *Index* (P-value = 0.005), and *Temp* (P-value = 0.009), have small P-values and should be considered valuable in this model to predict response to treatment. The other three variables, *Smear* (P-value = 0.907), *Infil* (P-value = 0.413), and *Blasts* (P-value = 0.478), have larger P-values and do not appear to be very effective in this model.

b. *Age* is an effective predictor in this model (P-value = 0.023) and has a negative coefficient, so the probability of responding to treatment appears to decrease for older patients. According to the odds ratio, the odds of responding go down by a factor of about 0.94 for every extra year of age, after accounting for the other variables in the model. *Temp* is also an effective term in the model (P-value = 0.009) with a negative coefficient. Higher temperatures appear to be associated with lower probabilities of responding, with the odds going down by about a factor of 0.89 for every extra tenth of a degree, after accounting for the other variables in the model.

c. Yes, a predictor that is "insignificant" in one model might actually be important to include in a final model. For example, we might have two predictors that are strongly related to each other and the response. When both are in the model, their individual tests may show large P-values since each is not needed if the other is in the model. However, dropping one might cause the P-value for the other to decrease dramatically when the similar predictor is no longer in the model, making it important to keep at least one of the two predictors in the final model.

d. We fit the reduced model $logit(Resp) = \beta_0 + \beta_1 Age + \beta_4 Index + \beta_6 Temp$ to test $H_0 : \beta_2 = \beta_3 = \beta_5 = 0$ versus H_a : Either $\beta_2 \neq 0$ or $\beta_3 \neq 0$ or $\beta_5 \neq 0$. Here is some output for the reduced model.

```
                                                   Odds     95% CI
Predictor         Coef     SE Coef       Z      P  Ratio  Lower  Upper
Constant       87.3880     35.4581    2.46  0.014
Age         -0.0585016   0.0255764   -2.29  0.022   0.94   0.90   0.99
Index         0.384926    0.121518    3.17  0.002   1.47   1.16   1.86
Temp        -0.0889732   0.0360684   -2.47  0.014   0.91   0.85   0.98

Log-Likelihood = -21.633
Test that all slopes are zero: G = 27.259, DF = 3, P-Value = 0.000
```

The improvement in the G-statistic is $\chi^2 = G_{Full} - G_{Reduced} = 31.249 - 27.259 = 3.99$, which we compare to a chi-square distribution with 3 degrees of freedom (the number of terms we are testing to leave out). The P-value is $P(\chi^2_3 \geq 3.99) = 0.2625$, which is not small. Thus we lack evidence to show that any of the three predictors, $Smear$, $Infil$, or $Blasts$, are useful in helping to predict whether or not leukemia patients respond to the treatment.

e. Age is quite consistent, with the estimated coefficient at approximately 0.06 and a P-value of approximately 0.22. The $Temp$ coefficient changes somewhat (-0.089 in the reduced model and -0.112 in the full model) and becomes somewhat less significant in the reduced model (P-value of 0.014 in the reduced model and P-value of 0.009 in the full model).

10.29 a. Here is some output for a logistic regression model for $Survive$ based on Age, $SysBP$, and $Pulse$.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.0612654  1.2280930   0.864  0.38750
Age         -0.0283710  0.0107816  -2.631  0.00850 **
SysBP        0.0167644  0.0058740   2.854  0.00432 **
Pulse       -0.0009298  0.0067020  -0.139  0.88966
```

After Age and $SysBP$ explain variability in $Survive$, $Pulse$ is not helpful for explaining any remaining unexplained variation in survival status at discharge (P-value = 0.89). Here is some output when $Pulse$ is dropped from the model.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.962471   1.000272   0.962  0.33594
Age         -0.028407   0.010774  -2.637  0.00838 **
SysBP        0.016831   0.005859   2.873  0.00407 **
```

b. The predicted log(odds) of survival for an 87-year-old patient with a systolic blood pressure of 80 is

$$log(\widehat{odds}) = 0.962471 - 0.028407(87) + 0.016831(80) = -0.162$$

The estimated odds of survival are $e^{-0.162} = 0.85$, and the estimated probability that this person survives his visit is

$$\hat{\pi} = \frac{e^{-0.162}}{1 + e^{-0.162}} = \frac{0.85}{1 + 0.85} = 0.46$$

c. This question is similar to asking whether you are surprised to observe "heads" on a flip of a fair coin. It is certainly not "very surprising" and not necessarily "very likely" but rather "reasonably likely."

10.30 a. The estimated coefficients and corresponding P-values for the models using the predictors individually appear in the table below.

Single variable in model	Estimated coefficient	P-value
Sex	−0.1054	0.771
Infection	−0.9163	0.011
Emergency	−2.1849	0.003

Individually in simple logistic regressions, an emergency admission or a suspected infection are significant predictors of survival with statistically significant OR of 0.112 and 0.400 (P-value = 0.004 and P-value = 0.011, respectively), suggesting that either event is associated with the lower odds for survival when present. Those admitted as emergencies have only approximately one-tenth the odds of survival compared to those who are not emergency admissions. The odds of survival for those with suspected infections is 60% less than for those without. Marginally, the odds of survival do not appear to be associated with sex (P-value = 0.771).

b. From the table below, we can see that *Sex* remains insignificant, while the magnitude of the effect of both $Infection$ and $Emergency$ has been reduced, probably indicating that there is a common portion of the variation in survival explained by both.

Multiple logistic model	Estimated coefficient	P-value
Intercept	3.496	0.000
Sex	0.05408	0.886
Infection	−0.7576	0.042
Emergency	−2.0749	0.006

c. The estimated probability of survival for an 87-year-old man who was admitted as an emergency with an infection is

$$\hat{\pi} = \frac{e^{3.496+0.05408(0)-0.7576(1)-2.0749(1)}}{1 + e^{3.496+0.05408(0)-0.7576(1)-2.0749(1)}} = \frac{1.94}{1 + 1.94} = 0.66$$

d. This event is not really surprising given that one would expect a man admitted under these circumstances to survive about 66% of the time.

10.31 a. Table of counts of *Survive* by *PClass*.

Survived	1st	2nd	3rd	Total
No	129	160	573	863
Yes	193	119	138	450
Total	322	279	711	1313

The proportions surviving in each class are

$$\hat{p}_1 = 193/322 = 0.599 \qquad \hat{p}_2 = 119/279 = 0.427 \qquad \hat{p}_3 = 138/711 = 0.194$$

The proportions surviving appear to decrease as class increases. First-class passengers may have had an easier time making it off the *Titanic* than those in third-class areas.

b. Here is some output for a chi-square analysis showing observed an expected counts based on the 2×3 table, assuming a null hypothesis of no relationship.

```
Rows: Survived   Columns: PClass
       1st    2nd    3rd

0      129    160    573
     211.6  183.3  467.1

1      193    119    138
     110.4   95.7  243.9

Pearson Chi-Square = 172.519, DF = 2, P-Value = 0.000
Likelihood Ratio Chi-Square = 173.144, DF = 2, P-Value = 0.000
```

The P-value is small, so we reject the null hypothesis and find that there is some association between survival and the travel class on the *Titanic*. (Note that there is one observation for which the class was not known. This observation has been eliminated for this analysis.)

c. Here is some output using indicators of first and second classes in a logistic model for survival.

```
                                                 Odds     95% CI
Predictor        Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant    -1.42363  0.0948238  -15.01  0.000
PClass_1st   1.82651   0.148070   12.34  0.000   6.21   4.65   8.30
PClass_2nd   1.12758   0.153769    7.33  0.000   3.09   2.28   4.17

Log-Likelihood = -757.041
Test that all slopes are zero: G = 173.144, DF = 2, P-Value = 0.000
```

The log(odds) of survival are 1.827 higher for passengers in first class (compared to third-class passengers) and 1.128 higher for second-class passengers. If we exponentiate each of these, we get the odds ratios shown in the output. The odds of survival were 6.21 times greater for first-class passengers and 3.09 times greater for second-class passengers, both as compared to third-class passengers.

d. Here are computations of the predicted survival proportion for each class, based on the model in (c).

$$\begin{aligned}
\text{1st class: } \hat{\pi}_1 &= \frac{e^{-1.42363+1.82651}}{1+e^{-1.42363+1.82651}} = \frac{1.496}{1+1.496} = 0.599 \\
\text{2nd class: } \hat{\pi}_2 &= \frac{e^{-1.42363+1.12758}}{1+e^{-1.42363+1.12758}} = \frac{0.7438}{1+0.7438} = 0.427 \\
\text{3rd class: } \hat{\pi}_3 &= \frac{e^{-1.42363}}{1+e^{-1.42363}} = \frac{0.2408}{1+0.2408} = 0.194
\end{aligned}$$

These predicted proportions match the proportions surviving in each class from part (a).

e. To test $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 = 0$ or $\beta_2 = 0$ for the model in part (c), we use the G-statistic (173.144), which gives a P-value ≈ 0 based on a chi-square distribution with 2 degrees of freedom. This is similar to the chi-square statistic (labeled as Pearson) in the output of part(b) and exactly matches the likelihood ratio chi-square statistic in that output. The conclusion that at least one of the class indicators is important for explaining survival is equivalent to the conclusion that there is some association between class and survival.

10.32 a. We fit the model $logit(\pi) = \beta_0 + \beta_1 Altitude + \beta_2 Lateral$, where π is the probability at least 10% of the birds flying away. Here is some output for that model.

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.39541    0.30551   7.841 4.48e-15
Altitude     0.19652    0.06745   2.914  0.00357
Lateral     -0.23883    0.02248 -10.625  < 2e-16
```

Both predictor variables have strong relationships with the response, *Flight*, in the presence of each other (P-values = 0.00357 and ≈ 0, respectively). In the presence of *Lateral*, a unit change in *Altitude* is associated with a 0.197 increase in the log odds of flight, which is to say that the odds of flight increase by a factor of $e^{0.197} = 1.217$. In the presence of *Altitude*, a unit change in *Lateral* is associated with a 0.239 decrease in the log odds of flight, which is to say that the odds of flight go down by a factor of $e^{-0.239} = 0.787$.

b. Following is some output for logistic regression models to predict *Flight* based on *Lateral* for each of the three altitude groups.

```
LOW altitude:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.64997    0.50189   7.272 3.53e-13
Lateral     -0.35995    0.05105  -7.051 1.78e-12

MID altitude:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.15153    0.42494   7.416 1.20e-13
Lateral     -0.24931    0.03714  -6.712 1.92e-11

HIGH altitude:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.41157    0.50297   4.795 1.63e-06
Lateral     -0.09981    0.03248  -3.073  0.00212
```

We see that the three slopes are -0.36, -0.25, and -0.10 (rounding to two decimal places). The standard errors of these slopes are 0.05, 0.04, and 0.03, so clearly these three slopes differ from one another by statistically significant amounts. The effect of *Lateral* on *Flight* is greatest at low altitudes, with the effect becoming weaker as altitude increases.

c. We fit the model $logit(\pi) = \beta_0 + \beta_1 Altitude + \beta_2 Lateral + \beta_3 Altitude \cdot Lateral$ to get

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.825039   0.499216   7.662 1.83e-14
Altitude         -0.203013   0.105580  -1.923   0.0545
Lateral          -0.390944   0.048391  -8.079 6.54e-16
Altitude:Lateral  0.040137   0.009487   4.231 2.33e-05
```

The interaction between *Lateral* and *Altitude* is highly significant (P-value = 0.0000233). We note that the interaction coefficient is positive, meaning that as *Altitude* increases the effect of *Lateral* moves closer to zero (as we saw in part (b)).

d. When *Altitude* and *Lateral* are both large, the interaction effect in the modeling will be pronounced. For example, case 4 in **FlightResponse** has *Altitude* = 9.14 and *Lateral* = 21.08. For this case, the model from part (a) gives a prediction of -0.84 for the log odds of flight, which means a predicted probability of 0.30, while the model from part (c) gives a prediction of 1.46 for the log odds, which means a predicted probability of 0.81.

10.33 a. We fit the model $logit(\pi) = \beta_0 + \beta_1 Seed$, where π is the probability of making it to the Final Four, to produce the output below.

```
            Estimate   Std. Error   z value   Pr(>|z|)
(Intercept) -0.2227     0.15287     -1.457      0.145
Seed        -0.4848     0.03975    -12.198    < 2e-16
```

Seed has a strong relationship with *Final4* (P-value ≈ 0) with a negative coefficient: The higher the seed, the lower the probability that the team makes it into the Final Four.

b. If we add a term for $Year$ to the model, we get the output below.

```
              Estimate   Std. Error    z value   Pr(>|z|)
(Intercept) -2.227e-01   1.595+01      -0.014     0.989
Seed        -4.484e-01  3.975e-02     -12.198   < 2e-16
Year         6.961e-15  7.983e-03       0.000     1.000
```

In the presence of *Seed*, $Year$ has no effect (P-value = 1.00). (This is not surprising since exactly four teams make the Final Four every year, so the effect of $Year$ should be zero.)

10.34 a. We fit the model $logit(\pi) = \beta_0 + \beta_1 Seed + \beta_2 Izzo$, where π is the probability of making it to the Final Four, to obtain the output below.

```
             Estimate   Std. Error   z value     Pr(>|z|)
(Intercept) -0.22597      0.16930    -1.335     0.18196
Seed        -0.49708      0.04458   -11.152     < 2e-16
Izzo         1.52755      0.56001     2.728     0.00638
```

This gives a fitted model $logit(\pi) = -0.22597 - 0.49708 Seed + 152755 Izzo$.

b. In the presence of *Seed*, $Izzo$ has a strong positive effect. The small P-value (0.00638) for the positive $Izzo$ coefficient tells us that a team coached by Tom Izzo has a better chance of making the Final Four than the team's *Seed* would predict and that this difference is greater than chance variation would suggest.

10.35 a. There is a strong positive relationship between the two variables. The following output shows that as *StateOpinion* increases by one point, the log odds of a "yes" vote increase by 0.254.

```
Coefficients
Term            Coef   SE Coef        95% CI      Z-Value  P-Value
Constant      -16.48      3.60  (-23.53, -9.43)    -4.58    0.000
StateOpinion  0.2538    0.0547  (0.1467, 0.3609)    4.64    0.000
```

b. Both *StateOpinion* and *Party* are useful predictors of *Vote*. Of the two, *Party* has the stronger effect. See the following output.

```
Coefficients
Term           Coef   SE Coef        95% CI    Z-Value  P-Value
Constant      -46.9      16.9  (-80.0, -13.8)    -2.78    0.005
StateOpinion  0.627     0.228  (0.181, 1.074)     2.75    0.006
Party_R       10.30      3.02  (4.39, 16.22)      3.41    0.001
```

10.36 a. There is a positive relationship between the variables. The output below shows that as *StateOpinion* increases by one point, the log odds of a “yes” vote increase by 0.115.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.78210    3.95962  -1.965   0.0494 *
StateOpinion  0.11499    0.05779   1.990   0.0466 *
```

b. The output below shows that both *StateOpinion* and *Party* are useful predictors of *Vote*. Of the two, *Party* has the stronger effect.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -27.2119     9.7521  -2.790  0.00526 **
StateOpinion   0.3715     0.1390   2.674  0.00750 **
PartyR         5.7888     1.2601   4.594 4.35e-06 ***
```

c. The output below shows that there is very little evidence of an interaction between *StateOpinion* and *Party* in their effects on *Vote*. The P-value for testing the interaction term is 0.431.

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -34.2211    14.2054  -2.409   0.0160 *
StateOpinion          0.4713     0.2019   2.334   0.0196 *
PartyR               20.3686    18.7708   1.085   0.2779
StateOpinion:PartyR  -0.2179     0.2766  -0.788   0.4308
```

10.37 a. Here is some output from fitting the model $logit(\pi) = \beta_0 + \beta_1 Uninsured$ where π is the probability of voting “yes” on insurance reform.

```
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept)   -0.1440     0.2582  -0.558    0.577
Uninsured      0.9875     1.4176   0.697    0.486
```

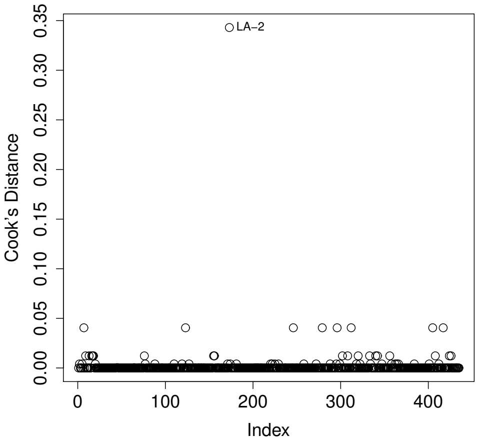
The coefficient of *Uninsured* ($\hat{\beta}_1 = 0.9875$) is positive, indicating a higher predicted probability of support as the percentage of uninsured in a district goes up. However, the P-value for testing this coefficient (0.486) is not small, which means the relationship is not strong and could be due to just random chance.

b. Following is some output using two predictors, *Dem* and *Obama*, to model the proportion supporting the insurance reform bill.

```
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -7.2976     1.0958  -6.659  2.75e-11
Dem            6.7453     1.0644   6.337  2.34e-10
Obama          3.7891     0.4641   8.165  3.22e-16
```

Both of the predictors have strong relationships with *InsVote*, but the z-value for the *Obama* coefficient is larger than that for *Dem*, so *Obama* has a slightly stronger impact on this model.

c. Here is a plot of Cook's distance for each congressional district. The extreme value is for the 2^{nd} congressional district in Louisiana. This is the only district for which a Republican voted in favor of the bill.



10.38 a. Here is some output from fitting the model $logit(\pi) = \beta_0 + \beta_1 uniChange$ where π is the probability of voting "yes" on the AHCA.

```
           Estimate Std. Error  z value   Pr(>|z|)
(Intercept)  1.0518     0.2444    4.304   1.68e-05
uniChange  -20.3054     4.4453   -4.568   4.93e-06
```

The coefficient of *uniChange* ($\hat{\beta}_1 = -20.305$) is negative, indicating a lower predicted probability of voting yes as the change in uninsured rate in a district goes up. The P-value for testing this coefficient is very small, which means the relationship is strong and is not attributable to just random chance.

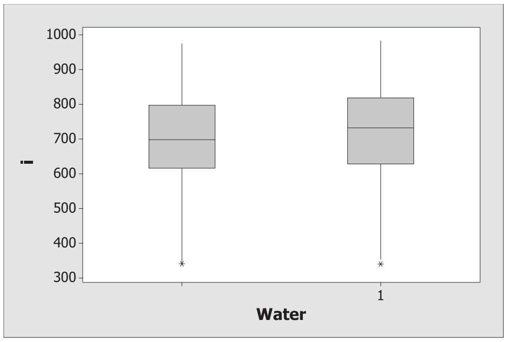
b. Following is some output using two predictors, *uniChange* and *Trump*, to model the proportion supporting the bill.

```
           Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -2.9808     0.5645   -5.281  1.29e-07
uniChange    6.1426     7.8405    0.783    0.433
Trump        4.8903     0.3865   12.652   <2e-16
```

Trump has a strong relationship with *AHCAvote*, but with Trump in the model the z-value for the *uniChange* coefficient is quite modest and this predictor is not significant.

c. The largest deviance residual is 2.41 for IL-06, a district that Trump lost but where the representative voted Yes on the AHCA. The smallest deviance residual is -2.21 for WA-03, a district that Trump won but where the representative voted No on the AHCA.

10.39 a. Side-by-side boxplots show that the mean time difference between the presence and absence of water is relatively small for the amount of interval variation in the samples. The mean for absence and presence are, respectively, 695 to 715, which (relative to the sample standard deviations of 132 and 139) seems rather small. (Note that this difference is statistically significant, which is not so very relevant given the large sample sizes.)



b. The output, which includes the coefficients, follows:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.252693   0.681876  -0.371 0.710946
Time        -0.006106   0.001055  -5.786  7.2e-09 ***
Water        1.209619   0.341653   3.540 0.000399 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 383.76  on 1589  degrees of freedom
AIC: 389.76
```

Summary: Both *Time* and *Water* are statistically significant predictors for the presence of a gunnel. The relationship between presence of a gunnel and *Time* is negative: The longer one waits to look after midnight, the less likely one is to find a gunnel. This is "controlling for" *Water*. The relationship between presence of gunnels and *Water* is positive: Controlling for *TIME*, one is more likely to find gunnels in standing water.

c. To get the estimated logit compute

$$-0.252693 + (-0.006106)(600) + 1.209619(1) = -2.706674$$

To get the estimated odds, exponentiate this number, that is $e^{(-2.706674)} = 0.0667584$, which is approximately an odds of 1:15 for finding a gunnel under those conditions.

d. To get the estimated logit compute

$$-0.252693 + (-0.006106)(600) + 1.209619(0) = -3.916293$$

To get the estimated odds, exponentiate this number, that is $e^{(-3.916293)} = 0.0199147$, which is approximately an odds of 1:50 for finding a gunnel under those conditions.

e. The odds ratio would be 0.0667584/0.0199147 = 3.352. At 10 a.m., the odds increase by a factor of about 3.35, which is a pretty substantial increase.

10.40 a. The relevant output is below:

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    6.449e+00  5.072e+00   1.272   0.2035
Time          -2.565e-02  1.900e-02  -1.350   0.1769
Water         -1.003e+01  6.020e+00  -1.666   0.0957 .
Time.SQ        1.235e-05  1.684e-05   0.733   0.4636
Time:Water     3.293e-02  2.172e-02   1.516   0.1295
Water:Time.SQ -2.166e-05  1.879e-05  -1.153   0.2490
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 373.65  on 1586  degrees of freedom
```

None of the variables are statistically significant, given the inclusion of the others in the model. This indicates much redundancy within this set of five predictors.

b. We compute a drop-in-deviance test using the output from the full model from (b) and the reduced model from the previous exercise: $383.76 - 373.65 = 10.11 =$ drop in deviance, with d.f. $= 1589 - 1586 = 3$. So using a chi-square distribution with 3 degrees of freedom, the P-value (the area to the right of 10.11) is 0.01765387. This is below the 0.05 threshold and suggests that one might want to entertain a model with more than just the two predictors *Time* and *Water*. The simpler model may, indeed, be inadequate.

10.41 a. Here is the output from the logistic model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.136977   0.781146  -1.456  0.14552
Time        -0.004276   0.001450  -2.949  0.00318 **
Fromlow     -0.030865   0.005422  -5.693 1.25e-08 ***
Water        0.460082   0.443433   1.038  0.29948
Slope        0.025813   0.013936   1.852  0.06398 .
Rw           1.431063   0.592338   2.416  0.01569 *
Pool         0.443370   0.410614   1.080  0.28024
Cobble       2.682660   0.394588   6.799 1.06e-11 ***
---
    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 247.88  on 1584  degrees of freedom
```

The variables deemed significant in this model are *Time*, *Fromlow*, *Rw*, and *Cobble*; the other three—*Water*, *Slope*, and *Pool*—are nonsignificant (although *Slope* is close to significant).

b. The P-values tell whether a single predictor adds to the model, beyond the model fitting all other predictors. So if a variable is not statistically significant, you know you can remove that one variable and not substantially lessen the quality of the model, and that is true for each of the three variables we mentioned in (a), but that is a one-at-a-time thing. You cannot tell whether the entire group of three can be removed or not until you do the nested LRT-test.

c. The model fit in (a) becomes the full model. The reduced model is given here:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.631075   0.692790  -0.911  0.36234
Time        -0.003983   0.001380  -2.886  0.00391 **
Fromlow     -0.031769   0.005441  -5.838 5.28e-09 ***
Rw           1.551527   0.576740   2.690  0.00714 **
Cobble       2.593051   0.371157   6.986 2.82e-12 ***
---
    Null deviance: 430.69  on 1591  degrees of freedom
Residual deviance: 253.70  on 1587  degrees of freedom
```

Here is the calculation for the drop-in-deviance test:

Test statistic $= 253.70 - 247.88 = 5.82$.
Degrees of freedom $= 1587 - 1584 = 3$.

The computer gives a P-value of 0.1207037

This P-value is not small enough to reject the null hypothesis and the null represents the reduced model here. Thus we prefer the reduced model with just the four predictors: *Time*, *Fromlow*, *Rw*, and *Cobble*. This says that the presence of gunnels depends upon how long after midnight and the last low tide the observation occurs at, and also depends the percentage of plants and weeds in the quadrat (*Rw*) and the presence or not of rocky cobbles.

d. The reduced model was adequate, so there is no further work to do here and our final model is that given in (c).

e. Based upon the summary table, we see that the probability of finding a gunnel:

- goes down the further from midnight we observe;
- goes down the further from the last low tide;
- goes up as the amount of plants and weeds increases; and
- goes up as the amount of cobbles increases.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.631075   0.692790  -0.911  0.36234
Time        -0.003983   0.001380  -2.886  0.00391 **
Fromlow     -0.031769   0.005441  -5.838 5.28e-09 ***
Rw           1.551527   0.576740   2.690  0.00714 **
Cobble       2.593051   0.371157   6.986 2.82e-12 ***
```

10.42 a. Here is the output from the logistic model:

```
Coefficients:
           Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   3.364       3.276     1.03    0.3045
Trust         0.845       0.471     1.79    0.0730 .
Faithful     -1.173       0.451    -2.60    0.0092 **
Attract      -0.103       0.416    -0.25    0.8044
SexDimorph   -0.441       0.445    -0.99    0.3216
---
Null deviance: 108.509 on 87 degrees of freedom
Residual deviance: 99.747 on 83 degrees of freedom
```

The variables deemed significant in this model are *Trust* and *Faithful*; the other two—*Attract* and *SexDimorph*—are non-significant.

b. The model fit in (a) becomes the full model. The reduced model is given here:

```
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -0.0633      1.4050    -0.05     0.964
Trust        0.7166      0.3645     1.97     0.049 *
Faithful    -0.8102      0.3346    -2.42     0.015 *
---
Null deviance: 108.51 on 87 degrees of freedom
Residual deviance: 101.40 on 85 degrees of freedom
```

Here is the calculation for the drop-in-deviance test:
Test statistic $= 101.4 - 99.7 = 1.65$.
Degrees of freedom $= 85 - 83 = 2$.
The computer gives a P-value of 0.44
This P-value is not small enough to reject the null hypothesis and the null represents the reduced model here. Thus we prefer the reduced model with just the two predictors: *Trust* and *Faithful*. This says that the likelihood that a man has cheated depends on his trustworthiness rating and on his faithfulness rating.

c. Based upon the summary table, we see that the probability that a man has cheated goes up the higher his trustworthiness rating is (which is the opposite of what we might expect) and goes down the higher his faithfulness rating is (which is consistent with what we might expect).

10.43 We first note that the age and grade predictors are strongly correlated ($r = 0.868$), so we are likely to have multicollinearity problems if both are in the same model. This also applies if we use the *DriverLicense* variable, which is just an indicator for an age of at least 16 years old. In any model with two (or all three) of these predictors, at most one is considered significant by the z-tests for individual coefficients. For example, here is some output with all five potential predictors.

```
                                                  Odds     95% CI
Predictor            Coef    SE Coef       Z       P  Ratio  Lower  Upper
Constant         -1.51203   0.442989   -3.41  0.001
female           0.697824  0.0455928   15.31  0.000   2.01   1.84   2.20
grade          -0.0765862  0.0367130   -2.09  0.037   0.93   0.86   1.00
age4            0.0376973  0.0403516    0.93  0.350   1.04   0.96   1.12
smoke            0.956550  0.0449624   21.27  0.000   2.60   2.38   2.84
DriverLicense  -0.0621409  0.0775284   -0.80  0.423   0.94   0.81   1.09

Log-Likelihood = -7089.683
Test that all slopes are zero: G = 1094.103, DF = 5, P-Value = 0.000
```

We drop the *DriverLicense* predictor since it is really just an indicator for two age groups and now decide between including *age4* or *grade* as a third predictor along with *female* and *smoke*. Here is some output for each of those models.

```
Variable         Value  Count
ride.alc.driver  1       3858  (Event)
                 0       8445
                 Total  12303

* NOTE * 12303 cases were used
* NOTE * 1084 cases contained missing values
                                                      Odds     95% CI
Predictor        Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant     -1.06648   0.279917  -3.81  0.000
female       0.698163  0.0455136  15.34  0.000   2.01   1.84   2.20
age4       -0.0424819  0.0176205  -2.41  0.016   0.96   0.93   0.99
smoke        0.960549  0.0448814  21.40  0.000   2.61   2.39   2.85

Log-Likelihood = -7104.825
Test that all slopes are zero: G = 1093.727, DF = 3, P-Value = 0.000

*******************************************************************
Variable         Value  Count
ride.alc.driver  1       3853  (Event)
                 0       8443
                 Total  12296

* NOTE * 12296 cases were used
* NOTE * 1091 cases contained missing values
                                                      Odds     95% CI
Predictor        Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant     -1.09719   0.195591  -5.61  0.000
female       0.701228  0.0451904  15.52  0.000   2.02   1.85   2.20
grade      -0.0624426  0.0188040  -3.32  0.001   0.94   0.91   0.97
smoke        0.958337  0.0448908  21.35  0.000   2.61   2.39   2.85

Log-Likelihood = -7097.510
Test that all slopes are zero: G = 1095.252, DF = 3, P-Value = 0.000
```

As anticipated, these models are very similar and would probably work equally well in practice. In both cases, the individual z-tests for all of the coefficients have small P-values (slightly smaller for *grade* compared to *age*4). The G-statistics are both huge and very similar in magnitude, with a slightly larger value for the model with *grade*. However, due to missing values, the two models are fit to slightly different sets of data, with a few more cases in the fit for the model using *age*4. In choosing between these similar models, we might take into account what specific research questions are of interest. For example, in applying the model to other youths, it may be easier to obtain information on age rather than grade level.

The interpretations of the two models are also very similar. The odds ratio for the $female$ term indicates that females have about twice as large odds of riding with a drinking driver than males, after accounting for the other two variables. The effect of smoking is even stronger with those who have smoked having odds about 2.6 times higher than nonsmokers. For most youths, grade level increases at about the same rate as age. We see that the odds of riding with a drinking driver actually decrease, by about a factor of 0.94 or 0.96 for each additional year, after also accounting for *smoke* and $female$.

It is interesting to note that if we use *age4* or *grade* alone in a logistic model to predict *ride.alc.drink* their coefficients are positive and significant.

```
                                                Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant    -2.18115   0.251822  -8.66  0.000
age4       0.0859281  0.0154970   5.54  0.000   1.09   1.06   1.12
********************************************************************
                                                Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant    -1.31729   0.177593  -7.42  0.000
grade      0.0498955  0.0167492   2.98  0.003   1.05   1.02   1.09
```

However, we also see in the data that the probability of smoking increases with age and, curiously, the older subjects in this study are more likely to be females. Thus we have some additional multicollinearity issues that make the interpretation of the individual coefficients problematic.

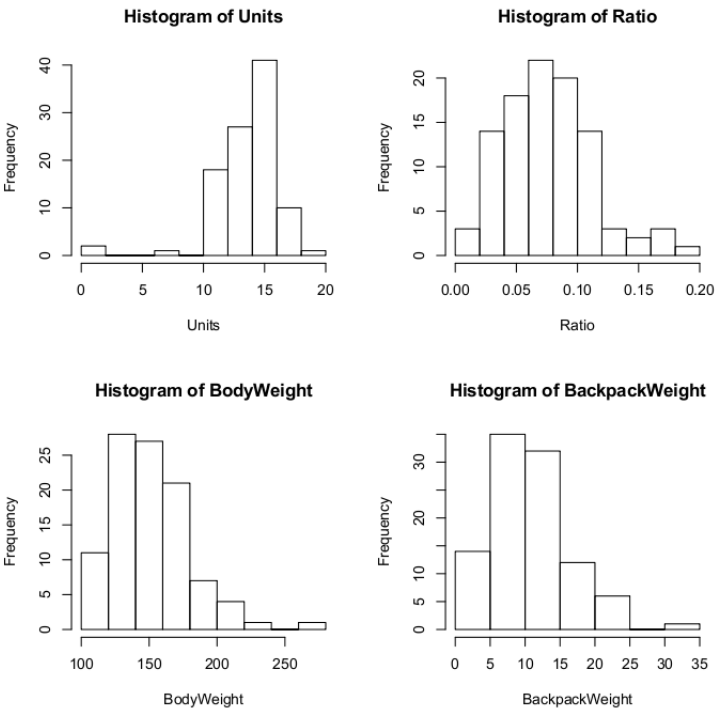
10.44 a. **Exploratory Data Analysis** First, we will tabulate and cross-tabulate variables that are **categorical** explanatory variables.

There are $n = 100$ observations, of which only 3 are graduate students, while 55 are females and 45 males. Thirty-two of these 100 students report back problems.

The potential explanatory variable $Major$ has a large number of categories (over 40!), so categories will need to be collapsed in order for this variable to be useful. We will keep in mind that our concern is how often students are likely to be carrying heavy loads and collapse accordingly.

The $Year$ is fairly evenly distributed with the exception of one student with a $Year$ of 0 and five students with $Year$s of six. In order to avoid removing observations unnecessarily, we will include the zero in $Year1$ and the five 6's with the 14 $Year5$s. A new variable $Year5$ was created collapsing categories in this manner.

Continuing the univariate EDA with the **quantitative** variables.
$BackpackWeight$ and $BodyWeight$ are positively skewed. The distribution of $Units$ or what is often referred to credits is negatively skewed with one very low value of 0. *Ratio* is slightly positively skewed.

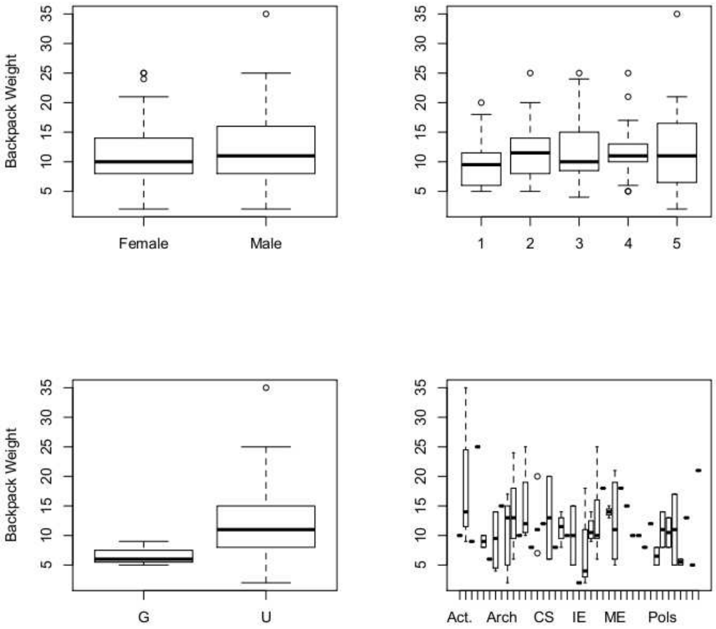


Examining **pairwise associations**, first with the outcome variable of interest $BackProblems$ Women report significantly more back problems than men (44% versus 18%, P-value = 0.01).

```
                  Sex
BackProblems Female Male
           0   0.56 0.82
           1   0.44 0.18

Pearson's Chi-squared test
X-squared = 6.4635, df = 1, p-value = 0.01101
```

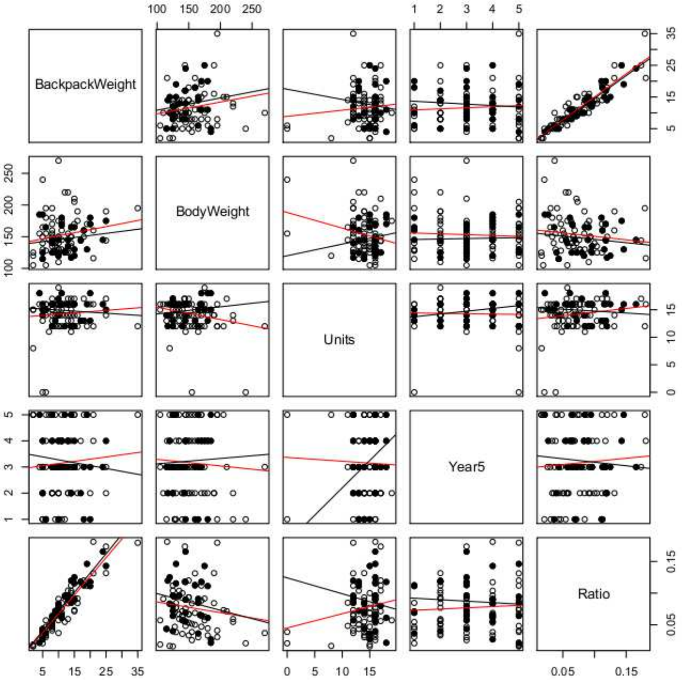
After collapsing $Year$ to five categories, the P-value for the χ^2 test is simulated because the small expected values make the χ^2 approximation questionable. We do not find any significant association of $Year$ with back problems.

```
            Year5
BackProblems    1    2    3    4    5
           0 0.67 0.85 0.56 0.64 0.74
           1 0.33 0.15 0.44 0.36 0.26

Pearson’s Chi-squared test with simulated p-value
(based on 2000 replicates)

X-squared = 5.0623, p-value = 0.2974
```



There do not appear to be exceptionally high correlations between the quantitative variables with the exception of $BackpackWeight$ and $Ratio$, which is to be expected given the definitions. It does suggest that we do not want to have both of these variables in the model.

```
               BackpackWeight BodyWeight Units Year5 Ratio
BackpackWeight           1.00       0.19  0.09  0.08  0.93
BodyWeight               0.19       1.00 -0.24 -0.06 -0.14
Units                    0.09      -0.24  1.00 -0.03  0.17
Year5                    0.08      -0.06 -0.03  1.00  0.07
Ratio                    0.93      -0.14  0.17  0.07  1.00
```



b. The 15% threshold is attained so infrequently that it is probably best to stick to the 10%. We can do some simple two-way tables: Sex is not related with the 10% threshold, but back problems is. People with back problems are more likely to have exceeded the 10% threshold (34% vs. 18%). A logistic model with the single binary predictor *BackProblems*, shows a P-value of 0.068. Comparative boxplots suggest that exceeding the 10% threshold is not related to body weight.

One could build some more logistic models, trying to add predictors to the model with back problems, but the goal needs to be clarified before going too far into this.

10.45 Initially, we consider recommendation taken, $RecTaken$, as the only predictor. The original structure of the $RecTaken$ variable is complicated. It is based on a variable *Recommends*, which has values such as R0, R1, R2,. . . R8. These recommendation values imply a course or set of courses based on the placement exam that are appropriate for that student. The $RecTaken$ variable equals 1 if the student took a recommended course and 0 otherwise. In addition, for later work or a project, you could also investigate the role of the variables $TooLow$ and $TooHigh$, defined for students taking a class at a level lower than recommended, and some students taking courses higher than recommended, $TooHigh$.

We need to define the response variable for this analysis. The exercise defines “success” as a grade of B or higher, so we create a variable called *CourseSuccess* as an indicator for whether students achieved a B or better grade.

EDA Findings Grade is needed to define our response of *CourseSuccess*, and we see from the summary command that 567 out of 2696 students are missing a grade. Out of the 2134 students with a grade, nearly 68% were successful in their chosen course. The average *ACTM* score is 27.0 (sd = 3.81). The mean *GPAadj* is 35.7 (sd = 4.6). Males make up only 44% of this dataset.

Missing Data From the summary command, we also note that there are quite a few missing values. As in past problems, we need to be careful that models we wish to compare contain the same observations. We examine the missing data for each variable. There is a way to use *PSATM* or *SATM* to estimate missing *ACTM* (it is based on a regression!), but we won’t use it here to fill in missing *ACTM*. *PSATM* and *SATM* will not be used. The binary response variable for this investigation is based on grade, so we will have at least 567 missing out of 2696 students in the data.

Here is a table showing the proportions successful and not successful for students who did and did not take a recommended course.

	Course Success	
Recommendation taken	Not successful	Successful
No	0.38	0.62
Yes	0.30	0.70

A chi-square test based on the original counts for this table chi-square test gives $\chi^2 = 15.3$ and P-value < 0.0001. There is undoubtedly an association between the recommendation and course success. We can reproduce this result in logistic regression:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.47203    0.08108   5.822 5.82e-09 ***
RecTaken       0.39070    0.09899   3.947 7.91e-05 ***
```

The odds of success in the course for those who took the recommendation is nearly 50% greater ($e^{0.3907} = 1.478$) than it was for those not taking the recommendation. This OR is unadjusted. It doesn’t take into account differences in *ACTM* between the groups or *GPAadj*. And possibly gender differs between the two groups. In fact, a look at some t-tests reveal that the two groups differed significantly with respect to higher high school adjusted GPA (P-value < 0.0001), while they do not differ with respect to *ACTM* (P-value $= 0.85$). A χ^2 test finds that the gender composition differs for the two groups (P-value < 0.00001).

There are actually a number of interesting questions you could pursue with this data, but we will focus on whether the response to the recommendation is associated with course success. The OR of 1.478 found above is not adjusted for other variables that may be responsible for this result.

Model 1: *ACTM* and *RecTaken* as predictors

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.64966    0.39890 -11.656  < 2e-16 ***
RecTaken       0.34757    0.11429   3.041  0.00236 **
ACTM           0.19342    0.01482  13.053  < 2e-16 ***
```

Both predictors remain highly significant and the adjusted OR for *rec.taken* is now at $e^{0.1934} = 1.42$.

Let's also see if *Gender* and *GPAadj* contribute further to the model. When gender is added to the model with *RecTaken* and *ACTM*, all predictors are significant. However, when *GPAadj* is added, all of the predictors are significant EXCEPT *RecTaken* (P-value = 0.60). Possibly the department could do away with the placement test altogether and use some combination of *ACTM*, *Gender*, and *GPAadj*.

```
                                                  Odds     95% CI
Predictor        Coef     SE Coef       Z       P  Ratio  Lower  Upper
Constant     -9.44573     1.22588   -7.71   0.000
RecTaken    0.0900720    0.242834    0.37   0.711   1.09   0.68   1.76
ACTM         0.128191   0.0318922    4.02   0.000   1.14   1.07   1.21
Gender      -0.365124    0.228519   -1.60   0.110   0.69   0.44   1.09
GPAadj       0.196362   0.0322788    6.08   0.000   1.22   1.14   1.30
```

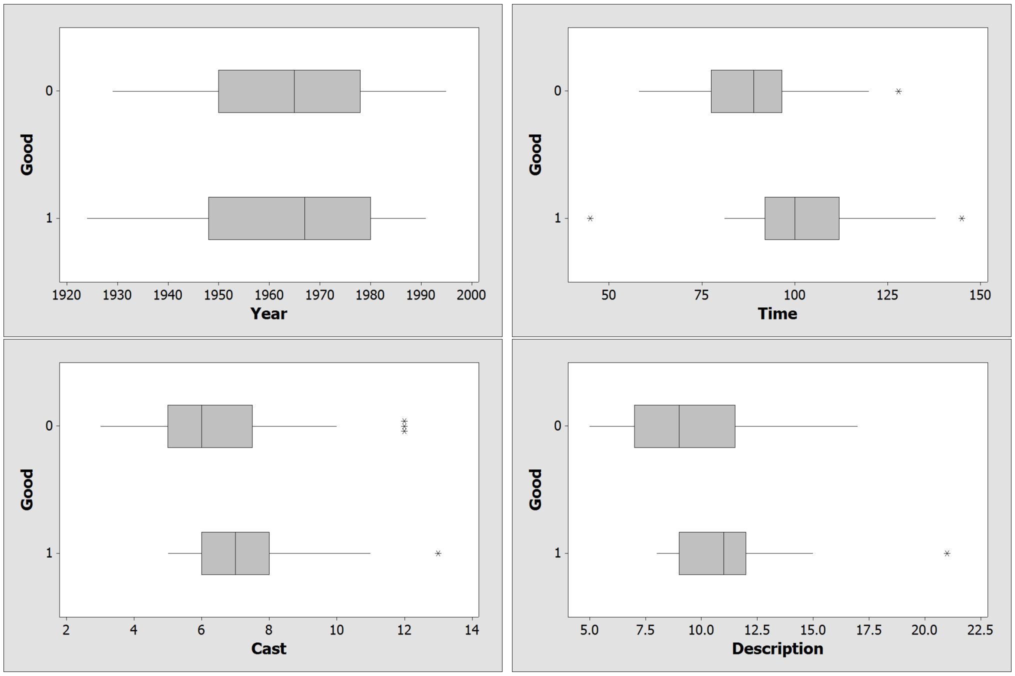
Strictly speaking these model statistics are not comparable because, with different variables in the model, different observations will be missing. One way around this is to do a complete case analysis, that is, we only keep and analyze observations that have all the values for our covariates. There are 505 complete cases.

```
                                               Odds     95% CI
Predictor       Coef   SE Coef      Z       P  Ratio  Lower  Upper
Constant    0.229574  0.170146   1.35   0.177
RecTaken    0.731254  0.206491   3.54   0.000   2.08   1.39   3.11
```

We see that the unadjusted OR (2.08) for *RecTaken* is greater than the previous one. The full model has different results with an odds ratio of 1.09.

```
                                                  Odds     95% CI
Predictor        Coef     SE Coef       Z       P  Ratio  Lower  Upper
Constant     -9.44573     1.22588   -7.71   0.000
RecTaken    0.0900720    0.242834    0.37   0.711   1.09   0.68   1.76
Gender      -0.365124    0.228519   -1.60   0.110   0.69   0.44   1.09
GPAadj       0.196362   0.0322788    6.08   0.000   1.22   1.14   1.30
ACTM         0.128191   0.0318922    4.02   0.000   1.14   1.07   1.21
```

10.46 To get a feel for relationships between the four quantitative predictors and the *Good* indicator, we look at side-by-side boxplots for each variable between good and poorer rated movies.



From the plots, it looks like the $Year$ distribution is the similar for good and lower-rated movies, while the good movies tend to have somewhat larger casts, run longer times, and have longer descriptions.

We'll first try a logistic regression model using all four predictors. Here is some output for fitting that model.

```
Variable  Value  Count
Good      1         31  (Event)
          0         69
          Total    100
                                                      Odds     95% CI
Predictor          Coef    SE Coef       Z      P    Ratio  Lower  Upper
Constant        73.1403    31.6717    2.31  0.021
Year         -0.0415906  0.0166677   -2.50  0.013     0.96   0.93   0.99
Time          0.0545986  0.0167714    3.26  0.001     1.06   1.02   1.09
Cast          0.0847997   0.133973    0.63  0.527     1.09   0.84   1.42
Description    0.191787   0.108009    1.78  0.076     1.21   0.98   1.50

Log-Likelihood = -50.927
Test that all slopes are zero: G = 21.966, DF = 4, P-Value = 0.000
```

Overall, the model is effective (P-value ≈ 0 for the test of all slopes based on the G-statistic), but the $Cast$ variable looks like a fairly weak predictor in this model (P-value $= 0.527$), so we'll try a reduced model with this predictor omitted.

```
                                                     Odds     95% CI
Predictor          Coef    SE Coef       Z       P  Ratio  Lower  Upper
Constant        75.1126    31.6490    2.37   0.018
Year         -0.0424456  0.0166751   -2.55   0.011   0.96   0.93   0.99
Time          0.0550073  0.0166451    3.30   0.001   1.06   1.02   1.09
Description    0.216623   0.101725    2.13   0.033   1.24   1.02   1.52

Log-Likelihood = -51.127
Test that all slopes are zero: G = 21.567, DF = 3, P-Value = 0.000
```

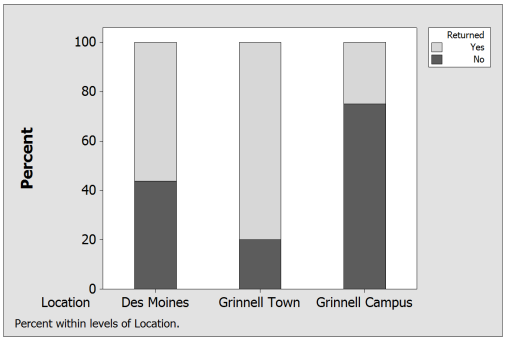
In this model, each of the three terms have significant coefficients (at a 5% level), and the G-statistic is virtually the same as the four-predictor model (21.567 compared to 21.966). This looks like a reasonable model to choose if we are limited to these four potential predictors. The fitted logistic regression model is

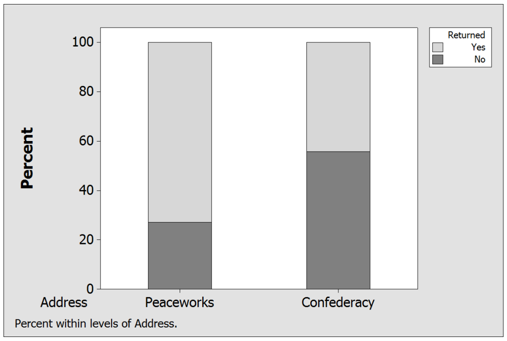
$$logit(\pi) = 75.1126 - 0.0424Year + 0.0550Time + 0.2166Description$$

Interpreting the odds ratios for each term we estimate that, after adjusting for the other predictors, the odds of being a good movie:

- go down by a factor of 0.96 for each extra year (so newer movies are less likely to be "good"),
- go up by a factor of 1.06 for each additional minute in length,
- go up by a factor of 1.24 for each additional line in the description.

10.47 We start with plots showing the percentage of letters returned or not from each of the three locations and two addresses. It looks like letters are more likely to be returned from Grinnell town and less likely to be returned from the Grinnell campus, with the Des Moines rate somewhere in between. Also, the letters addressed to Iowa Peaceworks are somewhat more likely to be returned than those addressed to Friends of the Confederacy.





To see which of these relationships are stronger than might be expected by random chance alone, we fit a multiple logistic regression model with indicators for one of the addresses (*Peaceworks*) and two of the three locations (*GrinnellTown* and *GrinnellCampus*).

$$logit(\pi) = \beta_0 + \beta_1 Peaceworks + \beta_2 GrinnellTown + \beta_3 GrinnellCampus$$

Here is some output for fitting this model.

```
Variable  Value  Count
Returned  1         82  (Event)
          0         58
          Total    140
                                                     Odds     95% CI
Predictor           Coef   SE Coef       Z       P  Ratio  Lower  Upper
Constant       -0.424347  0.299532  -1.42  0.157
Peaceworks       1.41655  0.398286   3.56  0.000   4.12   1.89   9.00
GrinnellTown     1.25197  0.478586   2.62  0.009   3.50   1.37   8.94
GrinellCampus   -1.50895  0.596497  -2.53  0.011   0.22   0.07   0.71

Log-Likelihood = -79.202
Test that all slopes are zero: G = 31.542, DF = 3, P-Value = 0.000
```

We see that overall, the model as a whole is effective (P-value ≈ 0), and each of the individual coefficients are significant at a 5% level (*Peaceworks* P-value = 0.000, *GrinnellTown* P-value = 0.009, and *GrinnellCampus* P-value = 0.011). Furthermore, we can interpret the individual coefficients and odds ratios to confirm and quantify the relationships observed in the plots.

The coefficient of *Peaceworks* is positive, indicating that those letters are more likely to be returned than those addressed to the Confederacy. How much more likely? From the odds ratio we estimate the odds of return are about 4.12 times higher. The letters lost in Des Moines serve as the reference group for the three locations. Letters left in the Grinnell town have about 3.5 times higher odds of being returned (compared to Des Moines), while those left on the Grinnell campus have return odds that are less than Des Moines by a factor of 0.22.

10.48 We first need to compute some new indicator variables to assess the coach's keys:

Key #1: There is no variable that measures trips up and down the court. Therefore, to assess the keys using logistic regression either we can use variables in the dataset to define a surrogate variable to stand in for trips or we can ignore trips. Let's try the first option.

Define a new variable: $Trips = (GrAtt - GrOR + GrTO) + (OppAtt + OppTO)$

The rationale is this: Each shot Grinnell attempts comes from a trip down the court; likewise for the opponent. Thus we add the terms *GrAtt* and *OppAtt* into the formula. Each Grinnell offensive rebound means a Grinnell shot attempt that was unsuccessful and yet did not lead to a separate trip down the court, as an offensive rebound gives Grinnell another attempt, without taking a trip up and down the court. We would likewise wish to have a term “$-OppOR$” in the formula, but opponent offensive rebounds was not in the dataset. The terms $GrTO$ and $OppTO$ are added in because each turnover committed by a team means a trip down the court in which a shot attempt did not occur, and we want to include these trips in our total. This surrogate $Trips$ variable is deficient by not including an $OppOR$ term, and there probably should be some term that takes into account free throw attempts because a certain percentage of these correspond to trips down the court, where no field goal was attempted. But since it is not clear how to incorporate free throws, we leave it out.

To use the coach’s key, we can also define an indicator variable ($Key1$) to track when $Trips \geq 150$.

Key #2: To find the percent of Grinnell attempts that are three-point shots, we create a variable $ThreePer = (Gr3Att/GrAtt)100$. An indicator for $Key2$ should equal to one when $GrAtt \geq 94$ and $ThreePer \geq 50$.

Key #3: To find the percentage of Grinnell shots for which they successfully get an offensive rebound, we define a variable $ORPer = (GrOR/(GrOR + OppDR))100$ and then find an indicator for $Key3$ when $ORPer \geq 33$.

Key #4: The indicator for $Key4$ tracks when $OppTO \geq 32$.

To check the coach’s keys, we run a multiple logistic regression model to predict $WinLoss$ using the four indicators for the keys.

```
Variable  Value  Count
WinLoss    1        77  (Event)
          -1        70
          Total    147
                                                    Odds     95% CI
Predictor        Coef   SE Coef       Z      P  Ratio  Lower  Upper
Constant   -0.634523  0.610367  -1.04  0.299
Key1        0.184094  0.611439   0.30  0.763   1.20   0.36   3.98
Key2       -0.128976  0.367696  -0.35  0.726   0.88   0.43   1.81
Key3        0.675581  0.369305   1.83  0.067   1.97   0.95   4.05
Key4        0.762509  0.438875   1.74  0.082   2.14   0.91   5.07

Log-Likelihood = -98.088
Test that all slopes are zero: G = 7.275, DF = 4, P-Value = 0.122
```

We see that none of the four keys have a significant coefficient (at a 5% level) although $Key3$ and $Key4$ are significant at 10%. Furthermore, the P-value for the G-statistic to test the overall effectiveness of the model is 0.122, which is not small enough to reject a null hypothesis that all of the coefficients are equal to zero. It looks like the coach's four keys are actually not very good indicators for when games are won.

We can also use the actual values of the relevant variables for these keys (rather than the indicators for thresholds being met) to ascertain whether the coach has landed on the variables that are likely to predict win-loss success. Here is some output for a logistic regression model using the five variables that help make the keys.

```
                                                    Odds     95% CI
Predictor         Coef    SE Coef       Z       P  Ratio  Lower  Upper
Constant      -3.15333    2.06326   -1.53   0.126
Trips        0.0122654  0.0153109    0.80   0.423   1.01   0.98   1.04
GrAtt       -0.0214052  0.0225615   -0.95   0.343   0.98   0.94   1.02
ThreePer    -0.0131069  0.0185877   -0.71   0.481   0.99   0.95   1.02
ORPer        0.0542074  0.0264334    2.05   0.040   1.06   1.00   1.11
OppTO        0.0686978  0.0313065    2.19   0.028   1.07   1.01   1.14

Log-Likelihood = -95.790
Test that all slopes are zero: G = 11.873, DF = 5, P-Value = 0.037
```

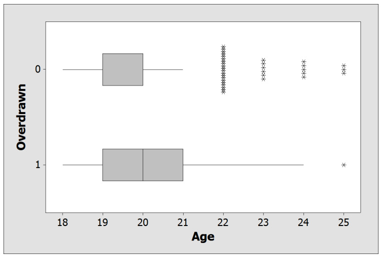
We see that only *ORPer* and *OppTO* have coefficients with P-values less than 5%. Let's try a model with just those two predictors,

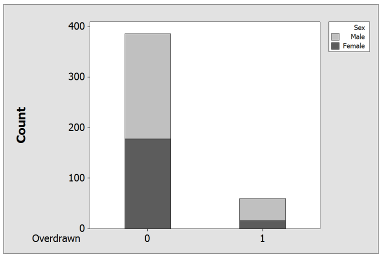
```
                                                  Odds     95% CI
Predictor        Coef    SE Coef       Z       P  Ratio  Lower  Upper
Constant     -3.45577    1.20683   -2.86   0.004
ORPer       0.0435866  0.0236368    1.84   0.065   1.04   1.00   1.09
OppTO       0.0739600  0.0279823    2.64   0.008   1.08   1.02   1.14

Log-Likelihood = -96.476
Test that all slopes are zero: G = 10.500, DF = 2, P-Value = 0.005
```

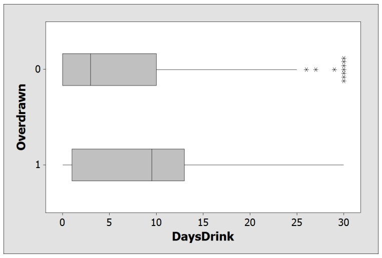
Both coefficients are positive, indicating that higher values of those variables are associated with a better chance of winning a game. Moreover, at a 5% level, *OppTO* is highly significant (P-value = 0.008) and *ORPer* is nearly significant (P-value = 0.065). Thus the coach could probably base his keys on just two of the given variables. Note that the G-statistic (10.5) and P-value (0.005) indicate that something is worthwhile in this model to predict wins.

One could also explore different predictors than the five mentioned by the coach or try different thresholds than those used in his keys to success. There are lots of possible avenues for investigating what relates to winning in the **Hoops** data.

10.49 We first look at some plots of each of the potential predictors, *Age*, *Sex*, and *DaysDrink*, versus *Overdrawn*.







The plots suggest that the overdrawn cases tend to be a bit older, more likely male, and drink more often.

We fit a logistic regression model based on all three predictors, $logit(\pi) = \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 DaysDrink$ to model the proportion overdrawn. Some output from this model is shown below.

```
Variable   Value  Count
Overdrawn  1         56  (Event)
           0        381
           Total    437


                                                  Odds     95% CI
Predictor       Coef    SE Coef       Z       P  Ratio  Lower  Upper
Constant    -8.19320    1.99002   -4.12  0.000
Age         0.249493  0.0957992    2.60  0.009   1.28   1.06   1.55
Sex          1.25477   0.347885    3.61  0.000   3.51   1.77   6.94
DaysDrink  0.0706817  0.0182138    3.88  0.000   1.07   1.04   1.11

Log-Likelihood = -153.031
Test that all slopes are zero: G = 28.548, DF = 3, P-Value = 0.000
```

We see that each of the predictors has a very small P-value, so all three are important contributors to this model. The P-value for testing the model as a whole is approximately zero, so this would appear to be an effective model. Interpreting the odds ratios, we see the estimated odds of overdrawing increasing by a factor of 1.28 for each additional year in age, being about 3.51 times higher for males compared to females, and going up by a factor of 1.07 for each additional day of drinking.

The fitted model would suggest predicting the log(odds) of overdrawing with

$$log(\widehat{odds}) = -8.193 + 0.2495 Age + 1.2548 Sex + 0.0707 DaysDrink$$