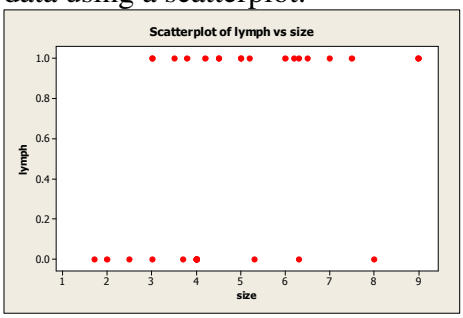
Homework 11 due in lab Dec. 8 ---- **Solutions**

(assignment for these solutions can be found on the last page)

1. *Redo the analysis of Example 12.44 p.582 of SW using Minitab. You do not need to replicate the plots presented there, but you do need to confirm the estimates of coefficients are the same as in SW. Interpret the tests of significance and the goodness of fit tests. Explain what the odds ratio means.*

First, while not necessary, we confirm that we're looking at the right data by plotting the data using a scatterplot.



The estimates of the coefficients in the logistic regression model match those in the equations at the bottom of p. 584, below in **bold**. The constant -2.08579 has a non-significant p-value (0.089), but size 0.511654 is significant at the 0.05 level (0.046). The goodness-of-fit test has a significant Deviance p-value (0.047) at the 0.05 level, thus the model does not fit well.

Therefore, though size is significant in this logistic model, the model does not adequately describe the relationship between size and lymph, so we should abandon this model.

With everything else held constant, the model predicts that a unit increase in size results in a 1.67 increase in the odds of lymph being a 1 to being a 0. That is, when size is increased one unit, lymph is 1.67 times more likely to be a 1 than to be a 0.

```
Logistic Regression Table
                                                  Odds     95% CI
Predictor      Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant   -2.08579    1.22564   -1.70  0.089
size        0.511654  0.256117    2.00  0.046   1.67   1.01   2.76

Goodness-of-Fit Tests
Method           Chi-Square  DF      P
Pearson             26.3404  19  0.121
Deviance            30.4102  19  0.047
Hosmer-Lemeshow     13.1280   8  0.108
```

2. *Analyze the HPV data on page 102 of the notes using logistic regression. You want to compare odds of HPV+ for Cases to that for controls. Compare this to the naive estimate of OR obtained using methods as on pages 106-7 of the notes.*

Data:

HPV Outcome	Cases	Controls
Positive	164	130
Negative	11	178
Sample size	175	308

Minitab:

HPV	group	freq
1	1	164
0	1	11
1	0	130
0	0	178

Below, both of the constant and group are significant in the logistic model with HPV. With everything else held constant, the model predicts that a unit increase in group (from control to case) results in a 20.41 increase in the odds of HPV being positive (1) to being negative (0). That is, when group is increased one unit, HPV is 20.41 times more likely to be positive.

Binary Logistic Regression: HPV versus group

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-0.314249	0.115370	-2.72	0.006			
group	3.01622	0.332109	9.08	0.000	**20.41**	10.65	39.14

Naïve estimate of the odds ratio is the odds of positive to negative for cases over the odds for controls. This matches the logistic analysis above.

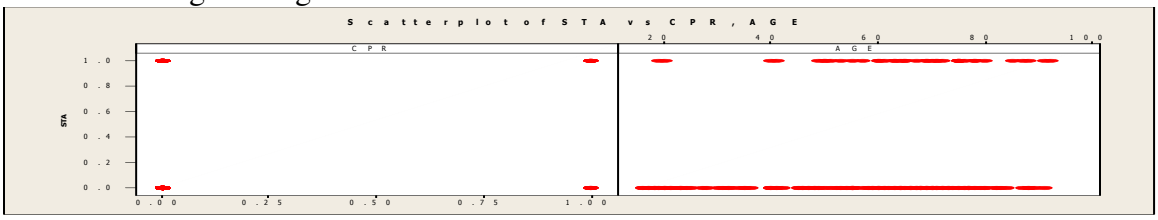
OR = ((164/175)/(11/175)) / ((130/308)/(178/308)) = 20.4140.

3. *The ICU data description and actual data set are on the web page. The major goal of the study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. There are lots of variables in the data set, but just use two - CPR status and AGE - for predicting STA. Evaluate the model (within the limits of what we have covered), and explain what your analysis says is the relationship between CPR status and AGE with probability of survival.*

The ICU data set consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients and to study the risk factors associated with ICU mortality.

STA	vital STAtus	0 = Lived	1 = Died
AGE	AGE	Years	
CPR	CPR Prior to ICU Admission	0 = No	1 = Yes

The scatterplots below do not give a great deal of information about how AGE and CPR are related to STA, except that AGE appears to have a 0 STA everywhere, but a 1 STA more on the higher range of AGE.



Below we see that both CPR (1.78409) and AGE (0.0296074) have a positive relationship to STA, and are both highly significant (p-values 0.003 and 0.008).
The Goodness-of-Fit deviance test gives a large p-value (0.656), indicating there is not significant evidence the model does not fit.

With everything else held constant, for a one unit increase in CPR (having CPR versus not) the model predicts an increase of 5.95 in the odds of death. Thus, having to receive CPR increases the chances of death by 6 fold.
With everything else held constant, for a one unit increase in AGE (one year) the model predicts an increase of 1.03 in the odds of death. While this odds ratio is significant, it is not very different from 1. However, an increase of 30 or 50 years surely has a drastic effect on the probability (or odds) of death.

Binary Logistic Regression: STA versus CPR, AGE

```
Link Function: Logit

Logistic Regression Table
                                                        Odds     95% CI
Predictor       Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant    -3.35196   0.745500   -4.50  0.000
CPR          1.78409   0.607297    2.94  0.003   5.95   1.81  19.58
AGE        0.0296074  0.0111489    2.66  0.008   1.03   1.01   1.05

Goodness-of-Fit Tests
Method           Chi-Square  DF      P
Pearson             57.9912  73  0.900
Deviance            67.6202  73  0.656
Hosmer-Lemeshow      2.7620   8  0.948
```

Stat 538 Fall 2005

Homework 11 due in lab Dec. 8

1. Redo the analysis of Example 12.44 p.582 of SW using Minitab. You do not need to replicate the plots presented there, but you do need to confirm the estimates of coefficients are the same as in SW. Interpret the tests of significance and the goodness of fit tests. Explain what the odds ratio means.

2. Analyze the HPV data on page 102 of the notes using logistic regression. You want to compare odds of HPV+ for Cases to that for controls. Compare this to the naive estimate of OR obtained using methods as on pages 106-7 of the notes.

3. The ICU data description and actual data set are on the web page. The major goal of the study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. There are lots of variables in the data set, but just use two - CPR status and AGE - for predicting STA. Evaluate the model (within the limits of what we have covered), and explain what your analysis says is the relationship between CPR status and AGE with probability of survival.