CHAPTER 11

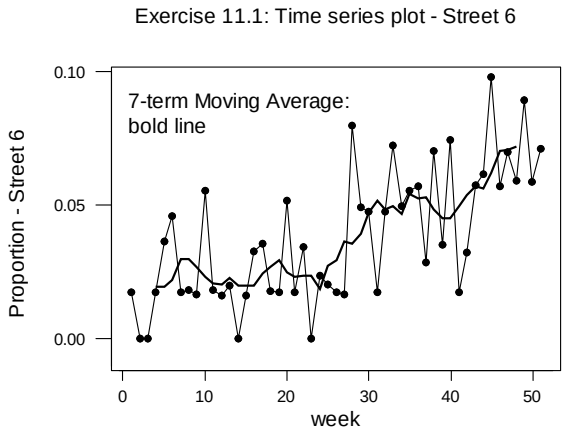
A note on computing with MINITAB (Version 14):

The **Minitab** software is used for fitting the logistic regression models in Chapter 11. Alternatively, one can use the SAS PROC GENMOD procedure; see the explanation in Chapter 12 of this solutions manual.

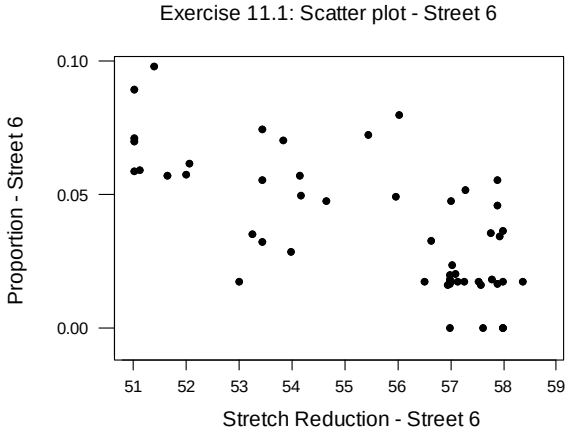
Minitab works like a spreadsheet program. We enter the data into the various columns of the spreadsheet and use the tabs: Stat > Regression > Binary logistic regression. We need to specify the response; either a column of zeros and ones if we work with individual cases, or the number of successes and the number of trials for each constellation if we work with aggregated data. We need to write out the model in model format. We can declare variables as factors – then Minitab will automatically create the needed indicator variables and test for factor effects. We can store the results (fitted values, residuals, …) in unused columns of the worksheet. All diagnostic graphs discussed in Chapter 11 of the book are available in Minitab.

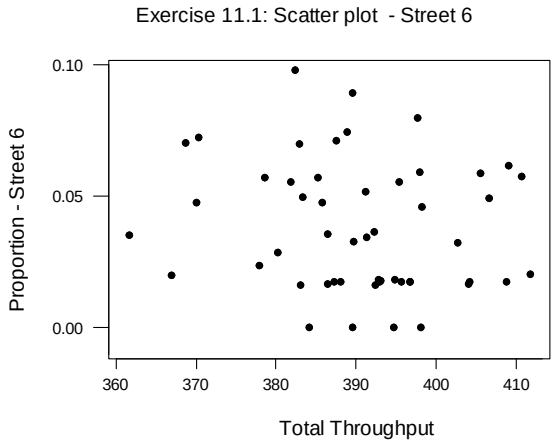
Options for various links (logit, probit, and complementary log-log links), starting values, maximum number of iterations, and number of classes in the Hosmer-Lemeshow test are available. Many other options are available. See the Minitab on-line help for detailed discussion and examples.

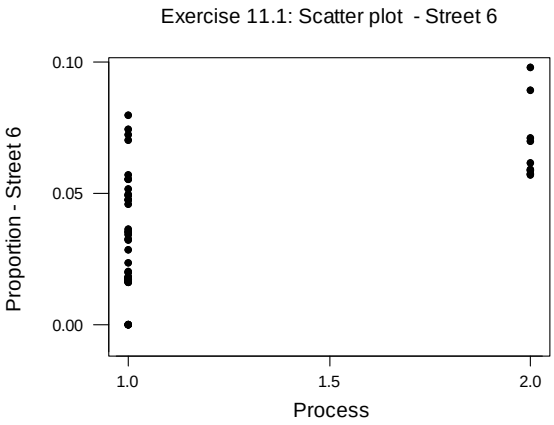
11.1 Time series graphs of weekly proportions of long fibers are given below. 7-term moving averages, $MA_t = (y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2})/7$, are added to these graphs. Moving averages amplify the trend component in a time series graph of noisy observations. The proportions of long fibers increase during the second half of the year.

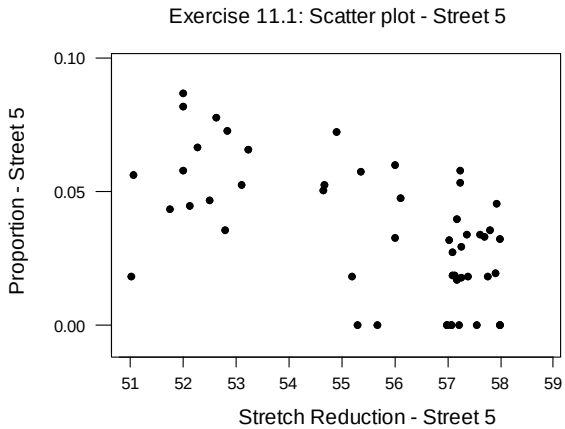


For each street (machine) separately, we construct scatter plots of the proportions of long fibers against stretch reduction, total throughput, and the type of process. The proportions of long fibers decrease with increased stretch reduction. The proportion of long fibers is larger under process 2.




Exercise 11.1: Scatter plot - Street 6
Proportion - Street 6
0.10
0.05
0.00
360
370
380
390
400
410
Total Throughput


Exercise 11.1: Scatter plot - Street 6
Proportion - Street 6
0.10
0.05
0.00
1.0
1.5
2.0
Process


Exercise 11.1: Scatter plot - Street 5
Proportion - Street 5
0.10
0.05
0.00
51
52
53
54
55
56
57
58
59
Stretch Reduction - Street 5

Logistic regression models for machine (street) 6:

Results for the following three logistic regression models are given below:

- model with stretch reduction, throughput, and process
- model with stretch reduction and throughput
- model with stretch reduction only

The total throughput and the type of process are insignificant. Stretch reduction remains as the only significant variable. An increase in the stretch reduction of one unit (percent) changes the odds for long fibers by a (multiplicative) factor of 0.85. That is, an increase in the stretch reduction of one unit (percent) reduces the odds for the occurrence of long fibers by 15 percent. Or, to say this differently: A small stretch reduction increases the odds for quality problems.

The proportion of long fibers π can be obtained from

$$\pi(\mathrm{x}) = \frac{\exp(\beta_0 + \beta_1 \mathrm{x})}{1 + \exp(\beta_0 + \beta_1 \mathrm{x})} = \frac{\exp(5.928 - 0.1662\mathrm{x})}{1 + \exp(5.928 - 0.1662\mathrm{x})}$$

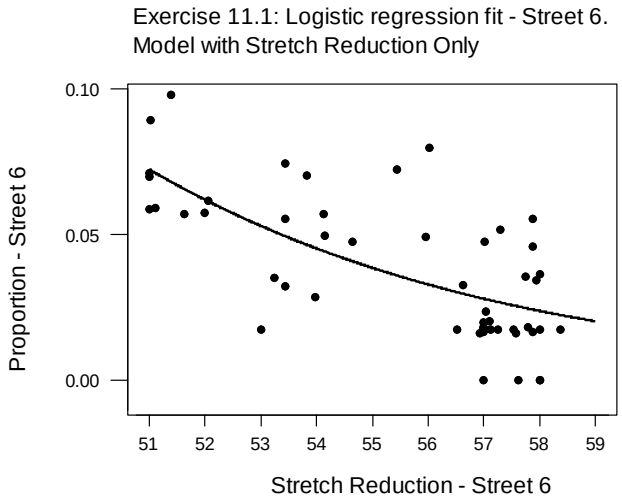
For stretch reduction x = 52, $\pi(\mathrm{x} = 52) = \dfrac{\exp(5.928 - 0.1662(52))}{1 + \exp(5.928 - 0.1662(52))} = 0.062$

For stretch reduction x = 53, $\pi(\mathrm{x} = 53) = \dfrac{\exp(5.928 - 0.1662(53))}{1 + \exp(5.928 - 0.1662(53))} = 0.053$

….

For stretch reduction x = 57, $\pi(\mathrm{x}=57)=\dfrac{\exp(5.928-0.1662(57))}{1+\exp(5.928-0.1662(57))}=0.028$

We have superimposed the fitted values (proportions of long fibers) in the scatter plot of the proportion of long fibers against stretch reduction (street 6). The main features of the scatter plot are well represented by the fitted model.



In this problem there are few exact replicates of the explanatory variable, stretch reduction. Minitab uses the approach by Hosmer and Lemeshow to group the cases on the basis of the estimated probabilities $\hat{\pi}_\mathrm{i}=\hat{\pi}(\mathrm{x}_\mathrm{i})$. It ranks the estimated probabilities from the smallest to the largest, and uses this ranking to break the cases into g = 10 groups of equal size. For each group k , $k = 1, 2, \ldots, g$, it calculates the number of successes o_k and the number of failures $\mathrm{n}_\mathrm{k}-\mathrm{o}_\mathrm{k}$ that are associated with the n_k cases in the group. The observed frequencies are compared with the expected frequencies $\mathrm{n}_\mathrm{k}\bar{\pi}_\mathrm{k}$ and $\mathrm{n}_\mathrm{k}(1-\bar{\pi}_\mathrm{k})$, where $\bar{\pi}_\mathrm{k}=\dfrac{\sum_{\mathrm{i}\in \mathrm{Group}k}\hat{\pi}_\mathrm{i}}{n_\mathrm{k}}$ is the average estimated success probability in the k^{th} group. The Pearson chi-square statistic is calculated from the resulting 2 x g table, and

$$HL=\sum_{\mathrm{k}=1}^{\mathrm{g}}\frac{[\mathrm{o}_\mathrm{k}-\mathrm{n}_\mathrm{k}\bar{\pi}_\mathrm{k}]^2}{\mathrm{n}_\mathrm{k}\bar{\pi}_\mathrm{k}(1-\bar{\pi}_\mathrm{k})}$$

is referred to as the Hosmer-Lemeshow statistic. Hosmer and Lemeshow show that the distribution of HL is well approximated by a chi-square distribution with g - 2 degrees of freedom. Large values of the Hosmer-Lemeshow statistic indicate lack of

fit. In our problem the Hosmer-Lemeshow statistic is *HL* = 6.938. It is quite small when compared to the 95th percentile of chi-square distribution with 10 – 2 = 8 degrees of freedom (15.51). The associated large probability value, 0.435, confirms that the model gives a very adequate representation of the data.

The Pearson residual for each of the 52 weeks is calculated from the equation $r_i = r(y_i, \hat{\pi}_i) = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$, where y_i and n_i are the number of occurrences and the total number of trials in week *i*, and where $\hat{\pi}_i = \hat{\pi}(x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = \frac{\exp(5.928 - 0.1662 x_i)}{1+\exp(5.928 - 0.1662 x_i)}$ is the implied success probability. The autocorrelations for the first six lags are given by

```
0.00, 0.06, -0.22, 0.05, 0.16, -0.02.
```

Comparing these to their approximate standard error, $1/\sqrt{52} = 0.14$, indicates no serial correlation among the residuals.

A note on residuals and fitted values: Minitab stores the residuals and the diagnostic measures for each constellation, and the constellations change with different model specifications. When estimating the logistic regression on stretch reduction alone, there are data for 51 weeks, but there are only 43 different stretch constellations. For three weeks the stretch reduction on street 6 is 51.000, for four weeks it is 57.000, and for four weeks it is 58.000. Minitab aggregates the information and supplies vectors of fitted values and residuals for the 43 constellations. This is fine as far as the usual diagnostic checks are concerned, but it causes difficulties if one wants to calculate the autocorrelations of the residuals where time order is of importance. One cannot compute the autocorrelations of weekly residuals from the vector of the aggregated residuals.
One must first compute the residuals for each week. This can be done by using the weekly frequencies (number of successes and number of trials) and the event probabilities at the constellations (note that these are stored by Minitab).
Alternatively, one can "trick" the program by adding small numbers to the replicates of stretch to make them slightly different (say 51.000, 51.001, and 51.003 for the three weeks with identical stretch reduction 51.000; etc). Then Minitab will treat them as separate constellations and will give you the vector of the 51 weekly residuals automatically.

Model with stretch reduction, throughput, and process:

```
Link Function:  Logit

Response  Information
```

```
Variable  Value          Count
Positive6 Success          139
          Failure         3225
samples6  Total           3364

Logistic Regression Table
                                                    Odds      95% CI
Predictor       Coef    SE Coef        Z     P   Ratio   Lower   Upper
Constant       8.854      4.044     2.19 0.029
stretch6    -0.16900    0.06532    -2.59 0.010    0.84    0.74    0.96
throughput -0.007084   0.008151    -0.87 0.385    0.99    0.98    1.01
process6     -0.0062     0.3606    -0.02 0.986    0.99    0.49    2.02

Log-Likelihood = -566.074
Test that all slopes are zero: G = 25.849, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   29.837    47  0.976
Deviance                  32.323    47  0.949
Hosmer-Lemeshow            2.053     8  0.979

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                                  Group
Value      1     2     3     4     5     6     7     8     9    10   Total
Success
  Obs      8    10     7     8    13    20    17    19    26    11     139
  Exp    7.7   9.3   9.1  10.0  10.5  17.7  17.2  20.0  25.7  11.8
Failure
  Obs    335   378   337   348   331   365   321   323   342   145    3225
  Exp  335.3 378.7 334.9 346.0 333.5 367.3 320.8 322.0 342.3 144.2

  Total  343   388   344   356   344   385   338   342   368   156    3364
```

Model with stretch reduction and throughput:

```
Link Function:  Logit
Response  Information

Variable  Value          Count
Positive6 Success          139
          Failure         3225
samples6  Total           3364

Logistic Regression Table
                                                    Odds      95% CI
Predictor       Coef    SE Coef        Z     P   Ratio   Lower   Upper
Constant       8.818      3.461     2.55 0.011
stretch6    -0.16805    0.03391    -4.96 0.000    0.85    0.79    0.90
throughput -0.007146   0.007285    -0.98 0.327    0.99    0.98    1.01
```

```
Log-Likelihood = -566.075
Test that all slopes are zero: G = 25.849, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   29.840    48  0.982
Deviance                  32.324    48  0.960
Hosmer-Lemeshow            2.063     8  0.979

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                                  Group
Value      1     2     3     4     5     6     7     8     9    10   Total
Success
  Obs      8    10     7     8    13    20    17    19    26    11     139
  Exp    7.7   9.3   9.1  10.0  10.5  17.6  17.2  20.0  25.7  11.8
Failure
  Obs    335   378   337   348   331   365   321   323   342   145    3225
  Exp  335.3 378.7 334.9 346.0 333.5 367.4 320.8 322.0 342.3 144.2

  Total  343   388   344   356   344   385   338   342   368   156    3364
```

Model with stretch reduction only:

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Positive6 Success          139
          Failure         3225
samples6  Total           3364

Logistic Regression Table
                                                    Odds      95% CI
Predictor       Coef    SE Coef        Z     P   Ratio   Lower   Upper
Constant       5.928      1.818     3.26 0.001
stretch6    -0.16619    0.03359    -4.95 0.000    0.85    0.79    0.90

Log-Likelihood = -566.554
Test that all slopes are zero: G = 24.891, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   27.801    41  0.943
Deviance                  26.951    41  0.955
Hosmer-Lemeshow            6.938     7  0.435

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)
```

					Group					
Value	1	2	3	4	5	6	7	8	9	Total
Success										
Obs	9	8	11	6	20	21	17	25	22	139
Exp	9.2	8.8	10.4	11.1	13.1	18.5	19.4	26.0	22.4	
Failure										
Obs	379	345	373	389	367	383	340	363	286	3225
Exp	378.8	344.2	373.6	383.9	373.9	385.5	337.6	362.0	285.6	
Total	388	353	384	395	387	404	357	388	308	3364

Logistic regression models for machine (street) 5:

Results for the following two logistic regressions models are given below:

- model with stretch reduction and throughput
- model with stretch reduction only

Process does not enter here, as machine 5 operates under one production process.

Total throughput is insignificant. Stretch reduction remains as the only significant variable. An increase in the stretch reduction of one unit (percent) changes the odds for long fibers by a (multiplicative) factor of 0.85. That is, an increase in the stretch reduction of one unit (percent) reduces the odds for long fibers by 15 percent. Note that the odds-ratios for stretch reduction are the same on both streets.

The proportion of long fibers π can be obtained from

$$\pi(\mathrm{x}) = \frac{\exp(\beta_0 + \beta_1 \mathrm{x})}{1 + \exp(\beta_0 + \beta_1 \mathrm{x})} = \frac{\exp(6.006 - 0.1681\mathrm{x})}{1 + \exp(6.006 - 0.1681\mathrm{x})}$$

For stretch reduction x = 52, $\pi(\mathrm{x} = 52) = \dfrac{\exp(6.006 - 0.1681(52))}{1 + \exp(6.006 - 0.1681(52))} = 0.061$

For stretch reduction x = 53, $\pi(\mathrm{x} = 53) = \dfrac{\exp(6.006 - 0.1681(53))}{1 + \exp(6.006 - 0.1681(53))} = 0.052$

….

For stretch reduction x = 57, $\pi(\mathrm{x} = 57) = \dfrac{\exp(6.006 - 0.1681(57))}{1 + \exp(6.006 - 0.1681(57))} = 0.027$

We have superimposed the fitted proportions of long fibers in the scatter plot of the proportion of long fibers against stretch reduction (street 5). The main features of the scatter plot are well represented by the fitted model.

The Hosmer-Lemeshow statistic is *HL* = 5.146. It is quite small when compared with the 95th percentile of chi-square distribution with 10 – 2 = 8 degrees of freedom (15.51). The associated large probability value, 0.742, confirms that the model leads to a very adequate representation of the data.

Model with stretch reduction and throughput:

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Positive5 Success          119
          Failure         3014
samples5  Total           3133


Logistic Regression Table
                                                    Odds      95% CI
Predictor       Coef    SE Coef        Z     P   Ratio   Lower   Upper
Constant       9.888      3.934     2.51 0.012
stretch5    -0.17408    0.03979    -4.38 0.000    0.84    0.78    0.91
throughput  -0.009107   0.007761    -1.17 0.241   0.99    0.98    1.01

Log-Likelihood = -496.083
Test that all slopes are zero: G = 19.664, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   36.191    48  0.895
Deviance                  47.990    48  0.473
Hosmer-Lemeshow            5.904     7  0.551
```

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

	Group									
Value	1	2	3	4	5	6	7	8	9	Total
Success										
Obs	7	10	4	11	8	18	21	21	19	119
Exp	7.3	8.2	8.5	9.3	10.4	14.5	18.3	19.5	23.0	
Failure										
Obs	326	336	328	335	341	361	348	316	323	3014
Exp	325.7	337.8	323.5	336.7	338.6	364.5	350.7	317.5	319.0	
Total	333	346	332	346	349	379	369	337	342	3133

Model with stretch reduction only:

Link Function: Logit

Response Information

Variable	Value	Count
Positive5	Success	119
	Failure	3014
samples5	Total	3133

Logistic Regression Table

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	6.006	2.138	2.81	0.005			
stretch5	-0.16807	0.03917	-4.29	0.000	0.85	0.78	0.91

Log-Likelihood = -496.765
Test that all slopes are zero: G = 18.300, DF = 1, P-Value = 0.000

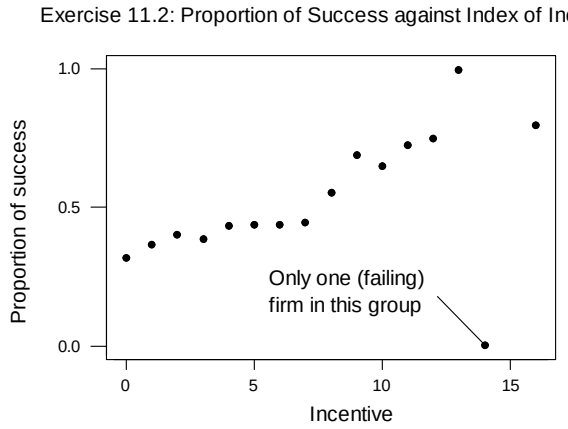
Goodness-of-Fit Tests

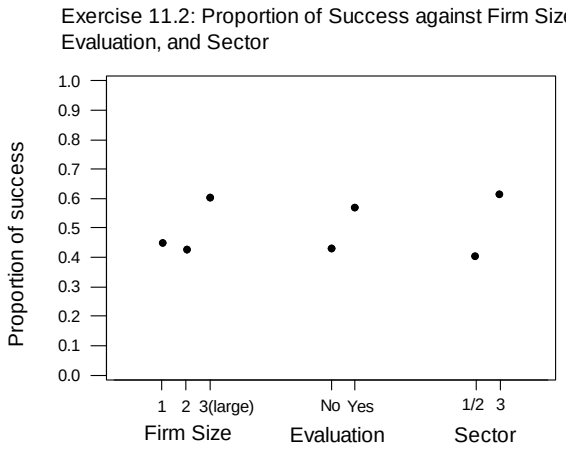
Method	Chi-Square	DF	P
Pearson	33.514	43	0.850
Deviance	44.313	43	0.416
Hosmer-Lemeshow	5.146	8	0.742

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

	Group										
Value	1	2	3	4	5	6	7	8	9	10	Total
Success											
Obs	8	8	10	6	12	11	21	20	22	1	119
Exp	7.9	8.5	8.5	9.2	11.0	11.4	16.6	20.2	21.9	3.9	
Failure											
Obs	329	335	314	336	355	304	330	334	323	54	3014
Exp	329.1	334.5	315.5	332.8	356.0	303.6	334.4	333.8	323.1	51.1	
Total	337	343	324	342	367	315	351	354	345	55	3133

11.2 Scatter plots of the success proportions against the incentive index, the size of the firm, the evaluation indicator, and the sector are given below. We learn that the chance for success increases with the number of offered incentives, and the size of the firm (large firms are usually more successful). Evaluation matters (evaluated firms tend to be more successful), and the sector appears to make a difference (larger success rate in the tertiary sector).





We consider a logistic regression model with the following explanatory variables: incentive index (a linear component), size (a categorical variable with 3 possibilities; we include two parameters for the three groups), evaluation, and sector (since we consider just two sectors - the primary/secondary and the tertiary sectors - we need only one parameter).

Model with incentives, size, evaluation and sector:

```
Link Function:  Logit

Response Information

Variable  Value       Count
profit    1             209  (Event)
          0             220
          Total         429

Factor Information

Factor   Levels Values
size          3 1 2 3

Logistic Regression Table
                                                        Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant     -2.8964     0.5630    -5.15 0.000
incentive    0.13173    0.02945     4.47 0.000     1.14     1.08     1.21
size
 2           -0.0352     0.2519    -0.14 0.889     0.97     0.59     1.58
 3            0.2794     0.2512     1.11 0.266     1.32     0.81     2.16
evaluation    0.4811     0.2096     2.30 0.022     1.62     1.07     2.44
sector2       0.7747     0.2138     3.62 0.000     2.17     1.43     3.30


Log-Likelihood = -271.242
Test that all slopes are zero: G = 51.955, DF = 5, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                  137.900   135  0.415
Deviance                 171.893   135  0.018
Hosmer-Lemeshow            6.499     8  0.592

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                                  Group
Value      1     2     3     4     5     6     7     8     9    10   Total
1
  Obs     10     9    19    18    29    21    28    27    30    18     209
  Exp      9.7  14.3  16.9  18.2  24.4  21.7  25.4  28.8  32.0  17.5
0
  Obs     32    38    29    27    23    21    16    17    14     3     220
  Exp     32.3  32.7  31.1  26.8  27.6  20.3  18.6  15.2  12.0   3.5

  Total   42    47    48    45    52    42    44    44    44    21     429
```

Next, we omit size of the firm (the two size indicators are insignificant), and fit the simpler logistic regression model with the incentive index, evaluation, and sector as explanatory variables.

We can construct a log-likelihood-ratio test to test the statistical significance of the factor "size." We illustrate in detail how one can test whether the size effect is significant. Comparing the log-likelihood = -271.242 of the full model with the log-likelihood of the restricted model (model without size; log-likelihood = -272.029) leads to the log-likelihood ratio test statistic $2(-271.242-(-272.029))=1.57$. Relating this statistic to a chi-square distribution with 2 degrees of freedom leads to the probability value $P(\chi^2(2)\geq 1.57)=0.4561$. Since the probability value is considerably larger than 0.05, we conclude that the factor "size" is not significant. We can work with the simplified model.

All remaining variables are statistically significant. A one unit increase in the incentive index (while keeping the other variables in the model constant) increases the odds for success by 15 percent. Evaluating the firm (and keeping the other variables in the model fixed) increases the odds for success by 64 percent. The odds for success of firms with the same incentive structure and evaluation in the tertiary sector are 127 percent larger than the odds in the primary/secondary sector.

The small Hosmer-Lemeshow statistic (5.905) and its large associated probability value (0.551) indicate that we have found an adequate model.

<u>Model with size omitted from the model:</u>

```
Link Function:  Logit

Response Information

Variable  Value       Count
profit    1             209  (Event)
          0             220
          Total         429

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant     -2.9768     0.5451    -5.46 0.000
incentive    0.13837    0.02893     4.78 0.000     1.15     1.09     1.22
evaluation    0.4926     0.2088     2.36 0.018     1.64     1.09     2.46
sector2       0.8206     0.2102     3.90 0.000     2.27     1.50     3.43

Log-Likelihood = -272.029
Test that all slopes are zero: G = 50.381, DF = 3, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   44.332    54  0.823
Deviance                  49.861    54  0.635
Hosmer-Lemeshow            5.905     7  0.551

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)
```

```
                                Group
Value      1     2     3     4     5     6     7     8     9   Total
1
  Obs     12    10    17    24    29    29    32    27    29     209
  Exp     12.3  13.8  18.0  21.6  26.8  25.7  29.6  28.8  32.4
0
  Obs     39    34    33    27    26    18    16    15    12     220
  Exp     38.7  30.2  32.0  29.4  28.2  21.3  18.4  13.2   8.6

  Total   51    44    50    51    55    47    48    42    41     429
```

11.3 The information can be arranged as a factorial, with the number of affected workers among the total number of workers in each group as the response variable. The 72 groups of the factorial arrangement are formed by all possible level combinations of the five explanatory variables: 3 (Dust) x 2 (Race) x 2 (Sex) x 2 (Smoking) x 3 (Employment). Seven of the 72 categories are empty and are ignored in our analysis. We use the binary logistic regression function in MINITAB, specifying the number of successes and the number of trials, and entering the explanatory variables as (categorical) factors. MINITAB creates the appropriate indicators for the factors automatically.

<u>Model with all five factors:</u>

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Yes       Success          165
          Failure         5254
Number    Total           5419

   65 cases were used
    7 cases contained missing values

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P     Ratio    Lower    Upper
Constant     -1.9452     0.2334    -8.33 0.000
Dust
 2           -2.5799     0.2921    -8.83 0.000     0.08     0.04     0.13
 3           -2.7306     0.2153   -12.68 0.000     0.07     0.04     0.10
Race
 2            0.1163     0.2072     0.56 0.574     1.12     0.75     1.69
Sex
 2            0.1239     0.2288     0.54 0.588     1.13     0.72     1.77
Smoking
 2           -0.6413     0.1944    -3.30 0.001     0.53     0.36     0.77
Employ
 2            0.5641     0.2617     2.16 0.031     1.76     1.05     2.94
 3            0.7531     0.2161     3.48 0.000     2.12     1.39     3.24

Log-Likelihood = -598.968
```

```
Test that all slopes are zero: G = 279.256, DF = 7, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   37.934    57  0.976
Deviance                  43.271    57  0.910
```

The test statistic for testing the overall significance of the regression (in equation (11.25)) is given by G = 279.256. Its sampling distribution (under the null hypotheses that none of the regressors have an influence on the response) is chi-square with 7 degrees of freedom. The test statistic G = 279.256 is huge compared to the percentiles from that distribution, and its associated probability value is tiny (p value < 0.0001). Hence the regressor variables (all or a subset) have a significant impact on the occurrence of byssinosis.

Race and Sex (both at two levels) have no significant effects. One can see this from the odds-ratios (they are roughly one), their t-ratios (Z-scores) and the associated probability values. The probability values for Race and Sex exceed the usual cutoff 0.05. The insignificance of the effects is also expressed by the confidence intervals of the odds-ratios; the confidence intervals cover one (indicating even odds).

The dustiness of the workplace, the smoking history, and the length of employment matter; the probability values of the estimated coefficients are smaller than 0.05, and the confidence intervals of the resulting odds-ratios do not cover the value one.

The deviance (in equation (11.26)) and the Pearson statistic (in equation (11.31)) compare the fit of the parameterized model (here with 8 = 7 + 1 (for constant) parameters) with the fit of the saturated model where each constellation of the explanatory variables is allowed its own distinct success probability. Here there are 65 = 72 - 7 constellations as seven cells are empty. The deviance is D = 37.9 and the Pearson statistic is $\chi^2 = 43.3$. Large values of these statistics indicate model inadequacy; the appropriate reference distribution is chi-square with 65 – 8 = 57 degrees of freedom. The deviance and the Pearson statistic are smaller than the critical percentile (the 95^{th} percentile is 75.62), implying that the probability values are considerably larger than 0.05. Hence there is no reason to question the adequacy of the model.

Here the deviance and the Pearson chi-square statistics are useful measures of (lack of) fit, as we have replicate observations at each configuration of the explanatory variable(s). In this example there is no reason to consider the Hosmer-Lemeshow statistic which becomes useful if we don't have replicate observations (as is often the case with continuous covariates).

The next steps in the analysis remove the insignificant regressors, sex and race. Because of possible multicollinearity it is always safer to this one step at a time. We first omit race as this variable has the smaller insignificant t-ratio (or, equivalently, the larger probability value). The output of the simplified model is given below:

Model without race:

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Yes       Success          165
          Failure         5254
Number    Total           5419

   65 cases were used
    7 cases contained missing values

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant     -1.8483     0.1549   -11.93 0.000
Dust
 2           -2.6118     0.2864    -9.12 0.000     0.07     0.04     0.13
 3           -2.7623     0.2079   -13.29 0.000     0.06     0.04     0.09
Sex
 2            0.1247     0.2286     0.55 0.586     1.13     0.72     1.77
Smoking
 2           -0.6411     0.1944    -3.30 0.001     0.53     0.36     0.77
Employ
 2            0.5238     0.2512     2.08 0.037     1.69     1.03     2.76
 3            0.6904     0.1844     3.74 0.000     1.99     1.39     2.86

Log-Likelihood = -599.126
Test that all slopes are zero: G = 278.940, DF = 6, P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF      P
Pearson                  28.316    27  0.395
Deviance                 29.716    27  0.327
```

The factor sex is insignificant (t-ratio 0.55, and probability value 0.59), and is omitted in the next model.

Model without race and sex:

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Yes       Success          165
          Failure         5254
Number    Total           5419
```

```
   65 cases were used
    7 cases contained missing values

Logistic Regression Table
                                                        Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant     -1.8336     0.1525   -12.03 0.000
Dust
 2           -2.5493     0.2614    -9.75 0.000     0.08     0.05     0.13
 3           -2.7175     0.1898   -14.31 0.000     0.07     0.05     0.10
Smoking
 2           -0.6210     0.1908    -3.26 0.001     0.54     0.37     0.78
Employ
 2            0.5060     0.2490     2.03 0.042     1.66     1.02     2.70
 3            0.6728     0.1813     3.71 0.000     1.96     1.37     2.80

Log-Likelihood = -599.274
Test that all slopes are zero: G = 278.645, DF = 5, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   13.570    12  0.329
Deviance                  12.094    12  0.438
```

No other variables can be omitted. Smoking is an important contributor to byssinosis. For a non-smoker the odds of contracting byssinosis are 0.54 the odds of a smoker. Everything else equal, not smoking reduces the odds of contracting byssinosis by 46 percent.

The length of employment in the cotton industry matters. The odds that a worker with 10 to 20 years employment contracts byssinosis are 1.66 times the odds of a worker with less than ten years in the industry. The odds for a worker with more than 20 years are twice (1.96) the odds of a worker with less than ten years in the industry.

Dustiness of the workplace clearly matters. The odds of contracting byssinosis at workplaces with medium and low levels of dustiness are considerably smaller than the odds for workplaces with a high level of dustiness (they are 0.08 and 0.07 times the odds of workplaces with high level of dustiness).

Next, we explore whether it is necessary to include interactions. The model with the three factors - smoking, length of employment, and dustiness of the workplace - and all two-factor interactions is given below.

<u>Model with two-factor interactions:</u>

```
Link Function:  Logit
Response  Information

Variable  Value          Count
Yes       Success          165
          Failure         5254
Number    Total           5419
```

```
Factor Information

Factor   Levels Values
Dust          3 1 2 3
Smoking       2 1 2
Employ L      3 1 2 3

   65 cases were used
    7 cases contained missing values

Logistic Regression Table
                                                          Odds        95% CI
Predictor             Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant           -1.9545     0.1922   -10.17 0.000
Dust
 2                 -2.7064     0.4775    -5.67 0.000     0.07     0.03     0.17
 3                 -2.4646     0.3274    -7.53 0.000     0.09     0.04     0.16
Smoking
 2                 -0.7242     0.3516    -2.06 0.039     0.48     0.24     0.97
Employ
 2                  0.8287     0.3324     2.49 0.013     2.29     1.19     4.39
 3                  0.9904     0.2551     3.88 0.000     2.69     1.63     4.44
Dust*Smoking
 2*2                1.1956     0.5501     2.17 0.030     3.31     1.12     9.72
 3*2                0.4546     0.4375     1.04 0.299     1.58     0.67     3.71
Dust*Employ
 2*2               -0.1908     0.7751    -0.25 0.806     0.83     0.18     3.78
 2*3               -0.5094     0.5881    -0.87 0.386     0.60     0.19     1.90
 3*2               -1.0915     0.6432    -1.70 0.090     0.34     0.10     1.18
 3*3               -0.4572     0.4103    -1.11 0.265     0.63     0.28     1.41
Smoking*Employ
 2*2               -0.0556     0.6162    -0.09 0.928     0.95     0.28     3.16
 2*3               -0.4911     0.4183    -1.17 0.240     0.61     0.27     1.39

Tests for terms with more than 1 degree of freedom

Term              Chi-Square    DF      P
Dust                  73.005     2  0.000
Employ                16.025     2  0.000
Dust*Smoking           4.863     2  0.088
Dust*Employ            3.712     4  0.446
Smoking*Employ         1.473     2  0.479

Log-Likelihood = -593.735
Test that all slopes are zero: G = 289.723, DF = 13, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                    1.005     4  0.909
Deviance                   1.016     4  0.907
```

The interactions between dust and employment length and between smoking history and employment length matter little, and are omitted from the model at the next step. The chi-square tests for the Dust*EmployLength interaction is 3.712 with probability value 0.446, and the Smoking*EmployLength interaction is 1.473 with probability value 0.479. These chi-square tests compare the full model with the model that restricts the interactions under consideration to zero.

Fitting the simpler model with the three factors smoking, length of employment, and dustiness of the workplace and the remaining 2-factor interaction between dust and smoking is shown below.

Model with the dustiness by smoking interaction:

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Yes       Success          165
          Failure         5254
Number    Total           5419

Factor Information

Factor   Levels Values
Dust          3 1 2 3
Smoking       2 1 2
Employ L      3 1 2 3

   65 cases were used
    7 cases contained missing values


Logistic Regression Table
                                                        Odds        95% CI
Predictor           Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant         -1.7573     0.1555   -11.30 0.000
Dust
 2               -2.9576     0.3565    -8.30 0.000     0.05     0.03     0.10
 3               -2.8325     0.2230   -12.70 0.000     0.06     0.04     0.09
Smoking
 2               -0.9573     0.2751    -3.48 0.001     0.38     0.22     0.66
Employ
 2                0.4990     0.2499     2.00 0.046     1.65     1.01     2.69
 3                0.6638     0.1819     3.65 0.000     1.94     1.36     2.77
Dust*Smoking
 2*2              1.1807     0.5490     2.15 0.031     3.26     1.11     9.55
 3*2              0.4864     0.4338     1.12 0.262     1.63     0.69     3.81

Tests for terms with more than 1 degree of freedom

Term          Chi-Square    DF      P
Dust             198.232     2  0.000
Employ            13.717     2  0.001
Dust*Smoking       4.840     2  0.089

Log-Likelihood = -596.848
Test that all slopes are zero: G = 283.496, DF = 7, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                    7.289    10  0.698
Deviance                   7.243    10  0.702
```

We illustrate in detail how one can test whether the interaction is significant. Comparing the log-likelihood = -596.848 of the full model with the log-likelihood of the restricted model (model without the interaction; log-likelihood = -599.274) leads to the log-likelihood ratio test statistic $2(-596.848-(-599.274)) = 4.84$. Relating this statistic to a chi-square distribution with 2 degrees of freedom leads to the probability value $P(\chi^2(2) \geq 4.84) = 0.089$. Note that the test-statistic (4.84) and the probability value (0.089) are given in the previous computer output. Since the probability value is larger than 0.05, we conclude that the interaction is not significant. Of course, at the ten percent significance level one would conclude that there is a smoking by dustiness interaction effect on the odds of contracting byssinosis. While there is some evidence of an interaction, the evidence is certainly not very strong.

How would one interpret the coefficients and the odds-ratios in the interaction component? One can write out the logistic regression model with the interaction terms and look at the odds for fixed levels of dustiness of the workplace.

(i) Comparing the odds for a non-smoker at a high-level dusty workplace (dust level 1), exp(constant - 0.9573), to those of a smoker at a high-level dusty workplace, exp(constant), leads to the odds-ratio exp(-0.9573) = 0.38. At a dusty workplace, nonsmoking reduces the odds of contracting byssinosis by 62 percent.

(ii) The odds-ratio for a non-smoker at a medium-level dusty workplace (dust level 2) is 0.38exp(1.1807) = (0.38)(3.26) = 1.25. At a medium-level dusty workplace the odds of contracting byssinosis for smokers and non-smokers are about the same. At medium-level dusty workplaces the smoking history has little influence on the odds of contracting the disease.

(iii) The odds-ratio for a non-smoker at a low-level dusty workplace (dust level 3) is 0.38exp(0.4864) = (0.38)(1.63) = 0.62. However, note the confidence interval for the interaction effect for (non)smoking and low dustiness (level 3) is quite wide (extending from 0.69 to 3.81) making the interpretation for low-level dustiness quite uncertain. The odds of contracting byssinosis for smokers and non-smokers may in fact be the same.

In summary, nonsmoking reduces the odds of contracting byssinosis, and the reduction is largest in very dusty workplaces.

11.4
Occurrence of proteinurea only:

Model with Smoking and Class: The test statistic for testing the overall significance of the logistic regression (in equation (11.25)) is G = 83.82. The sampling distribution (under the null hypotheses that none of the regressors have an influence on the

response) is chi-square with 6 degrees of freedom. The test statistic is large compared to the percentiles from that distribution and the probability value is small (p value < 0.001). Hence the regressor variables (all or some) have a significant impact on the presence of proteinurea.

The deviance (in equation (11.26)) and the Pearson statistic (in equation (11.31)) compare the fit of the parameterized model (here with 7 = 6 + 1 (for constant) parameters) with the fit of the saturated model where each constellation of the explanatory variables is allowed its own distinct success probability. Here there are 15 = (5)(3) constellations. The deviance is D = 15.35 and the Pearson statistic is $\chi^2 = 16.08$. Large values of these statistics indicate model inadequacy; the appropriate reference distribution is chi-square with 15 – 7 = 8 degrees of freedom. The deviance and the Pearson statistic are roughly the same size as the critical percentile (the 95th percentile is 15.51), implying probability values that are about 0.05. This leaves some doubt whether the model is adequate.

Individually, the coefficients for the four classes (class 2 through 5) are insignificant. These four coefficients express the incremental effect of class 2 through class 5, with class 1 acting as the standard. One can see the insignificance from the odds-ratios (they are roughly one, hence not changing the odds of class 1), their t-ratios (Z-scores), and the associated probability values. The probability values exceed the usual cutoff 0.05, with the one for the second class coming closest to 0.05 (it is 0.087). All four confidence intervals of their odds-ratios cover the value one (even odds).

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Proteinu  Success         2715
          Failure        10669
Total     Total          13384

Factor Information

Factor   Levels Values
Smoking       3 1 2 3
Class         5 1 2 3 4 5

Logistic Regression Table
                                                      Odds        95% CI
Predictor       Coef    SE Coef        Z     P      Ratio    Lower    Upper
Constant     -1.2964     0.1078   -12.03 0.000
Smoking
 2          -0.38319    0.04770    -8.03 0.000     0.68     0.62     0.75
 3          -0.26838    0.09115    -2.94 0.003     0.76     0.64     0.91
Class
 2            0.2102     0.1227     1.71 0.087     1.23     0.97     1.57
 3            0.0802     0.1112     0.72 0.471     1.08     0.87     1.35
 4           -0.0088     0.1222    -0.07 0.943     0.99     0.78     1.26
 5            0.0071     0.1386     0.05 0.959     1.01     0.77     1.32
```

```
Tests for terms with more than 1 degree of freedom

Term      Chi-Square    DF      P
Smoking       66.685     2  0.000
Class          8.385     4  0.078

Log-Likelihood = -6708.093
Test that all slopes are zero: G = 83.819, DF = 6, P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF      P
Pearson                  16.077     8  0.041
Deviance                 15.351     8  0.053
```

Model with Smoking Only: The next step in the analysis is to remove the factor "class" from the model (that is, omitting all four class indicators). We can test whether the factor class (with its five categories) is significant.

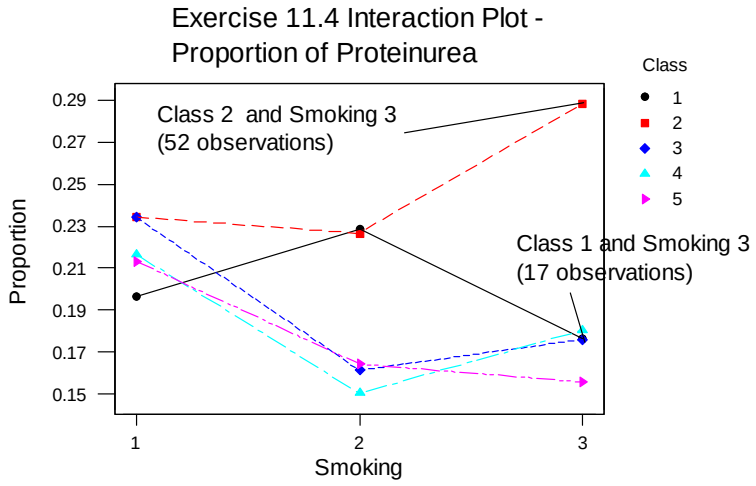
Comparing the log-likelihood -6,708.093 of the full model with the log-likelihood of the restricted model (model without class; log-likelihood = -6,712.254) leads to the log-likelihood ratio test statistic $2(-6{,}708.093-(-6{,}712.254)) = 8.38$. Relating the test statistic to a chi-square distribution with 4 degrees of freedom leads to the probability value $P(\chi^2(4) \geq 8.38) = 0.078$. Since this probability value is larger than 0.05, we conclude that the factor "class" is insignificant. "Class" can be omitted from the model. Note that the test statistic and its probability value are part of the earlier output for the model with both smoking and class.

The odds-ratios for smoking (0.67 and 0.75) imply that smoking is beneficial in reducing the onset of proteinurea. It seems beneficial for mothers to smoke!! Other studies also found that toxemia is less frequent in smokers than in non-smokers. The medical explanation for this is unclear. Brown et al quote evidence that nicotine dilates the muscle capillaries. Furthermore, research suggests that the cyanide in tobacco is detoxicated in the body to thiocyanate which has a known effect on hypertension and may be the active agent in reducing toxaemia.

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Proteinu  Success         2715
          Failure        10669
Total     Total          13384

Factor Information

Factor   Levels Values
Smoking       3 1 2 3
```

```
Logistic Regression Table

                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant    -1.21512    0.02730   -44.51 0.000
Smoking
 2          -0.39654    0.04716    -8.41 0.000     0.67     0.61     0.74
 3          -0.29167    0.09052    -3.22 0.001     0.75     0.63     0.89

Tests for terms with more than 1 degree of freedom

Term      Chi-Square    DF      P
Smoking       73.867     2  0.000

Log-Likelihood = -6712.254
Test that all slopes are zero: G = 75.498, DF = 2, P-Value = 0.000

* NOTE * No goodness of fit tests performed.
       * The model uses all degrees of freedom.
```

Comment: The model with smoking and class considered above is barely adequate, with goodness-of-fit statistics right at the critical 95th percentile. This fact may be the result of an interaction effect. The following interaction plot shows that this lack of fit may originate from the data for class 1 and 2 at smoking level 3. Unfortunately these cells are the ones with the smallest numbers of trials, and the somewhat unusual proportions at these cells may be an artifact of the small sample size.



Occurrence of hypertension only:

Model with Smoking and Class: The test statistic for testing the overall significance of the regression (in equation (11.25)) is G = 29.27 (with probability value = 0.000). Hence the regressor variables (all or some) have a significant impact on the presence of hypertension. The deviance D = 8.1 and the Pearson statistic $\chi^2 = 6.9$, and their respective probability values 0.42 and 0.55, give us no reason to question the adequacy of the model.

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Hyperten  Success          589
          Failure        12795
Total     Total          13384

Factor Information
Factor   Levels Values
Smoking       3 1 2 3
Class         5 1 2 3 4 5

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant     -3.0430     0.2024   -15.04 0.000
Smoking
 2           0.23179    0.08989     2.58 0.010     1.26     1.06     1.50
 3            0.3339     0.1575     2.12 0.034     1.40     1.03     1.90
Class
 2           -0.3829     0.2442    -1.57 0.117     0.68     0.42     1.10
 3           -0.2277     0.2095    -1.09 0.277     0.80     0.53     1.20
 4            0.0255     0.2254     0.11 0.910     1.03     0.66     1.60
 5            0.2582     0.2431     1.06 0.288     1.29     0.80     2.08

Log-Likelihood = -2400.886
Test that all slopes are zero: G = 29.270, DF = 6, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                    6.904     8  0.547
Deviance                   8.122     8  0.422
```

Model with Smoking only: We assess whether it is possible to omit the factor "class" from the model. Comparing the log-likelihood -2,400.886 of the full model with the log-likelihood of the restricted model (model without class; log-likelihood = -2,409.267) leads to the log-likelihood ratio test statistic $2(-2{,}400.886 - (-2{,}409.267)) = 16.76$. Relating it to a chi-square distribution with 4 degrees of freedom leads to the probability value $P(\chi^2(4) \geq 16.76) = 0.0022$. Since this probability value is small, we conclude that the factor "class" is significant. It cannot be omitted from the model.

Smoking increases the odds for hypertension (odds-ratios of 1.26 and 1.40).

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Hyperten  Success          589
          Failure        12795
Total     Total          13384
```

```
Logistic Regression Table
                                                        Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant    -3.21356    0.05948   -54.03 0.000
Smoking
 2           0.27020    0.08861     3.05 0.002     1.31     1.10     1.56
 3            0.3966     0.1559     2.54 0.011     1.49     1.10     2.02

Log-Likelihood = -2409.276
Test that all slopes are zero: G = 12.492, DF = 2, P-Value = 0.002

* NOTE * No goodness of fit tests performed.
       * The model uses all degrees of freedom.
```

Occurrence of both hypertension and proteinurea:

Model with smoking and class:

```
Link Function:  Logit

Response  Information


Variable  Value          Count

Both hyp  Success          665
          Failure        12719
Total     Total          13384

Factor Information

Factor   Levels Values
Smoking       3 1 2 3
Class         5 1 2 3 4 5

Logistic Regression Table
                                                        Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant     -2.6142     0.1779   -14.70 0.000
Smoking
 2          -0.40768    0.08910    -4.58 0.000     0.67     0.56     0.79
 3           -0.5793     0.1903    -3.04 0.002     0.56     0.39     0.81
Class
 2           -0.4695     0.2191    -2.14 0.032     0.63     0.41     0.96
 3           -0.1641     0.1849    -0.89 0.375     0.85     0.59     1.22
 4           -0.1036     0.2049    -0.51 0.613     0.90     0.60     1.35
 5           -0.0101     0.2321    -0.04 0.965     0.99     0.63     1.56

Tests for terms with more than 1 degree of freedom

Term      Chi-Square    DF      P
Smoking       26.488     2  0.000
Class          7.884     4  0.096

Log-Likelihood = -2627.725
Test that all slopes are zero: G = 33.644, DF = 6, P-Value = 0.000
```

```
Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                    6.673     8  0.572
Deviance                   7.240     8  0.511
```

Model with Smoking Only:

```
Link Function:  Logit

Response  Information

Variable  Value          Count
Both hyp  Success          665
          Failure        12719
Total     Total          13384

Factor Information

Factor   Levels Values
Smoking       3 1 2 3



Logistic Regression Table
                                                        Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant    -2.79260    0.04917   -56.80 0.000
Smoking
 2          -0.38545    0.08809    -4.38 0.000     0.68     0.57     0.81
 3           -0.5453     0.1893    -2.88 0.004     0.58     0.40     0.84


Log-Likelihood = -2631.881
Test that all slopes are zero: G = 25.332, DF = 2, P-Value = 0.000

* NOTE * No goodness of fit tests performed.
       * The model uses all degrees of freedom.
```

The factor “class” can be omitted from the model. Smoking decreases the odds of developing both hypertension and proteinurea.

Occurrence of either hypertension or proteinurea (or both):

Model with Smoking and Class:

```
Link Function:  Logit

Response  Information

Variable  Value          Count
EitherOr  Success         3969
          Failure         9415
Total     Total          13384
```

```
Factor Information

Factor   Levels Values
Smoking       3 1 2 3
Class         5 1 2 3 4 5

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant    -0.71346    0.09336    -7.64 0.000
Smoking
 2          -0.33729    0.04142    -8.14 0.000     0.71     0.66     0.77
 3          -0.25676    0.07919    -3.24 0.001     0.77     0.66     0.90
Class
 2           -0.0080     0.1078    -0.07 0.941     0.99     0.80     1.23
 3          -0.02587    0.09641    -0.27 0.788     0.97     0.81     1.18
 4           -0.0239     0.1056    -0.23 0.821     0.98     0.79     1.20
 5            0.0748     0.1187     0.63 0.529     1.08     0.85     1.36

Tests for terms with more than 1 degree of freedom

Term      Chi-Square    DF      P
Smoking       69.021     2  0.000
Class          1.811     4  0.770

Log-Likelihood = -8100.026
Test that all slopes are zero: G = 72.512, DF = 6, P-Value = 0.000

Goodness-of-Fit Tests

Method                Chi-Square    DF      P
Pearson                   13.141     8  0.107
Deviance                  12.867     8  0.117
```

Model with Smoking Only:

```
Link Function:  Logit

Response  Information

Variable  Value          Count
EitherOr  Success         3969
          Failure         9415
Total     Total          13384

Factor Information

Factor   Levels Values
Smoking       3 1 2 3

Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant    -0.73006    0.02448   -29.82 0.000
Smoking
 2          -0.33465    0.04093    -8.18 0.000     0.72     0.66     0.78
 3          -0.25381    0.07864    -3.23 0.001     0.78     0.67     0.91
```

```
Tests for terms with more than 1 degree of freedom

Term      Chi-Square    DF      P
Smoking       69.868     2  0.000

Log-Likelihood = -8100.923
Test that all slopes are zero: G = 70.719, DF = 2, P-Value = 0.000

* NOTE * No goodness of fit tests performed.
       * The model uses all degrees of freedom.
```

The factor “class” has no influence on the odds of developing either hypertension or proteinurea. Smoking decreases the odds of developing either one of these conditions.