**5.1.1 – SAS: Logistic Regression**

Example: (Text Table 14.3) Individuals were randomly sampled within two sectors of a city, and checked for presence of disease (here, spread by mosquitoes). Subjects’ age (in years), socioeconomic status (low, medium, high), and city sector are to be used to predict the probability of contracting the disease.

**/\* Input data -- see Table 14.3 in text**

**Case = subject ID**

**Age = years**

**SES\_mid = indicator of middle socioeconomic status**

**SES\_low = indicator of low socioeconomic status**

**(upper is reference level for socioeconomic status)**

**Sector = indicator of sector 2 in city**

**(sector 1 is reference level)**

**Disease = indicator of disease presence**

**\*/**

**filename myurl url "http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%2014%20Data%20Sets/CH14TA03.txt";**

**data outbreak;**

**infile myurl delimiter = '09'x;**

**input Case Age SES\_mid SES\_low Sector Disease;**

**Observation = \_n\_;**

**run;**

**/\* Run logistic regression, checking for lack of fit \*/**

**proc logistic data=outbreak plots=(roc effect);**

**model Disease(event = '1') = Age SES\_mid SES\_low Sector /**

**clparm=wald alpha=.05 lackfit;**

**SES: test SES\_mid=SES\_low=0;**

**output out=a1out prob=phat;**

**title1 'Logistic Regression';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Logistic Regression*** |     Probability modeled is Disease=1.     | **Model Convergence Status** | | --- | | Convergence criterion (GCONV=1E-8) satisfied. |      | **Model Fit Statistics** | | | | --- | --- | --- | | **Criterion** | **Intercept Only** | **Intercept and Covariates** | | **AIC** | 124.318 | 111.054 | | **SC** | 126.903 | 123.979 | | **-2 Log L** | 122.318 | 101.054 |      | **Testing Global Null Hypothesis: BETA=0** | | | | | --- | --- | --- | --- | | **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** | | **Likelihood Ratio** | 21.2635 | 4 | 0.0003 | | **Score** | 20.4067 | 4 | 0.0004 | | **Wald** | 16.6437 | 4 | 0.0023 |      | **Analysis of Maximum Likelihood Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** | | **Intercept** | 1 | -2.3127 | 0.6426 | 12.9545 | 0.0003 | | **Age** | 1 | 0.0297 | 0.0135 | 4.8535 | 0.0276 | | **SES\_mid** | 1 | 0.4088 | 0.5990 | 0.4657 | 0.4950 | | **SES\_low** | 1 | -0.3051 | 0.6041 | 0.2551 | 0.6135 | | **Sector** | 1 | 1.5746 | 0.5016 | 9.8543 | 0.0017 |      | **Odds Ratio Estimates** | | | | | --- | --- | --- | --- | | **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | | | **Age** | 1.030 | 1.003 | 1.058 | | **SES\_mid** | 1.505 | 0.465 | 4.868 | | **SES\_low** | 0.737 | 0.226 | 2.408 | | **Sector** | 4.829 | 1.807 | 12.907 |      | **Association of Predicted Probabilities and Observed Responses** | | | | | --- | --- | --- | --- | | **Percent Concordant** | 77.5 | **Somers' D** | 0.554 | | **Percent Discordant** | 22.1 | **Gamma** | 0.556 | | **Percent Tied** | 0.3 | **Tau-a** | 0.242 | | **Pairs** | 2077 | **c** | 0.777 |      | **Parameter Estimates and Wald Confidence Intervals** | | | | | --- | --- | --- | --- | | **Parameter** | **Estimate** | **95% Confidence Limits** | | | **Intercept** | -2.3127 | -3.5721 | -1.0533 | | **Age** | 0.0297 | 0.00328 | 0.0562 | | **SES\_mid** | 0.4088 | -0.7653 | 1.5828 | | **SES\_low** | -0.3051 | -1.4891 | 0.8789 | | **Sector** | 1.5746 | 0.5915 | 2.5578 |      | **Linear Hypotheses Testing Results** | | | | | | --- | --- | --- | --- | --- | |  | **Label** | **Wald Chi-Square** | **DF** | **Pr > ChiSq** | |  | **SES** | 1.2053 | 2 | 0.5474 |      | **Partition for the Hosmer and Lemeshow Test** | | | | | | | --- | --- | --- | --- | --- | --- | | **Group** | **Total** | **Disease = 1** | | **Disease = 0** | | | **Observed** | **Expected** | **Observed** | **Expected** | | **1** | 10 | 0 | 0.79 | 10 | 9.21 | | **2** | 10 | 1 | 1.02 | 9 | 8.98 | | **3** | 11 | 2 | 1.51 | 9 | 9.49 | | **4** | 10 | 1 | 1.78 | 9 | 8.22 | | **5** | 10 | 3 | 2.34 | 7 | 7.66 | | **6** | 10 | 4 | 3.09 | 6 | 6.91 | | **7** | 10 | 7 | 3.91 | 3 | 6.09 | | **8** | 11 | 3 | 5.51 | 8 | 5.49 | | **9** | 10 | 5 | 6.32 | 5 | 3.68 | | **10** | 6 | 5 | 4.75 | 1 | 1.25 |      | **Hosmer and Lemeshow Goodness-of-Fit Test** | | | | --- | --- | --- | | **Chi-Square** | **DF** | **Pr > ChiSq** | | 9.1871 | 8 | 0.3268 | |

**/\* Make better 'Conditional Effect' plot, compare predicted**

**disease probabilities for sector 1 (Sector=0) vs**

**sector 2 (Sector=1) at low socioeconomic status**

**(SES\_mid=0, SES\_low=1), as a function of Age \*/**

**data new; set outbreak;**

**p1 = 1/(1+exp(-(-2.3127+0.0297\*Age+0.4088\*0**

**-0.3051\*1+1.5746\*0)));**

**p2 = 1/(1+exp(-(-2.3127+0.0297\*Age+0.4088\*0**

**-0.3051\*1+1.5746\*1)));**

**label p1 = 'Sector 1'**

**p2 = 'Sector 2';**

**proc sort data=new; by Age;**

**proc sgplot data=new;**

**series y=p1 x=Age / lineattrs=(pattern=solid);**

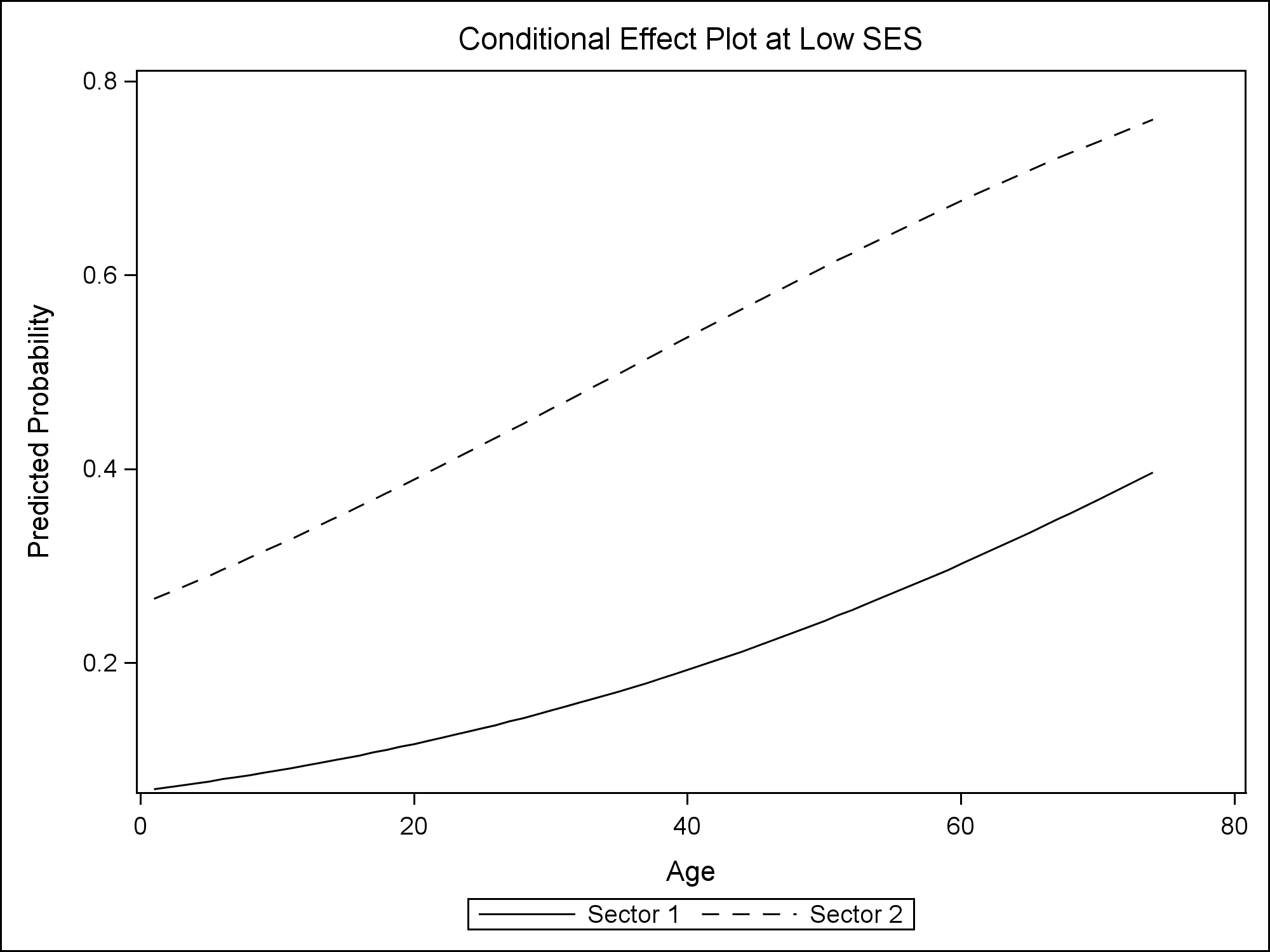
**series y=p2 x=Age / lineattrs=(pattern=dash);**

**xaxis label='Age';**

**yaxis label='Predicted Probability';**

**title1 'Conditional Effect Plot at Low SES';**

**run;**



**/\* Check for multicollinearity \*/**

**proc reg data=outbreak;**

**model Disease = Age SES\_mid SES\_low Sector / vif collin;**

**title1 'Collinearity Check';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Collinearity Check*** |      | **Parameter Estimates** | | | | | | | | --- | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** | | **Intercept** | **1** | 0.04699 | 0.10470 | 0.45 | 0.6546 | 0 | | **Age** | **1** | 0.00555 | 0.00238 | 2.33 | 0.0218 | 1.05242 | | **SES\_mid** | **1** | 0.07595 | 0.11139 | 0.68 | 0.4970 | 1.24616 | | **SES\_low** | **1** | -0.04150 | 0.10323 | -0.40 | 0.6886 | 1.34514 | | **Sector** | **1** | 0.31702 | 0.09180 | 3.45 | 0.0008 | 1.09668 |      | **Collinearity Diagnostics** | | | | | | | | | --- | --- | --- | --- | --- | --- | --- | --- | | **Number** | **Eigenvalue** | **Condition Index** | **Proportion of Variation** | | | | | | **Intercept** | **Age** | **SES\_mid** | **SES\_low** | **Sector** | | **1** | 2.91249 | 1.00000 | 0.01791 | 0.02993 | 0.02390 | 0.01836 | 0.03753 | | **2** | 1.03987 | 1.67357 | 0.00177 | 0.00043477 | 0.23886 | 0.23672 | 0.02873 | | **3** | 0.56543 | 2.26957 | 0.00369 | 0.01213 | 0.41619 | 0.08529 | 0.46221 | | **4** | 0.36812 | 2.81280 | 0.00172 | 0.50905 | 0.06301 | 0.18584 | 0.37255 | | **5** | 0.11410 | 5.05233 | 0.97491 | 0.44845 | 0.25805 | 0.47378 | 0.09897 | |

**/\* Variable selection \*/**

**/\* - here, backward elimination;**

**may also consider selection=stepwise \*/**

**proc logistic data=outbreak;**

**model Disease(event = '1') = Age SES\_mid SES\_low Sector /**

**selection=backward slstay=0.10;**

**title1 'Backward Elimination';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Backward Elimination*** |   Probability modeled is Disease=1.  Backward Elimination Procedure  ***Step 0. The following effects were entered:***  ***Intercept Age SES\_mid SES\_low Sector***  ***Step 1. Effect SES\_low is removed:***  ***Step 2. Effect SES\_mid is removed:***   |  |  | | --- | --- | | ***Note:*** | ***No (additional) effects met the 0.1 significance level for removal from the model.*** |      | **Summary of Backward Elimination** | | | | | | | --- | --- | --- | --- | --- | --- | | **Step** | **Effect Removed** | **DF** | **Number In** | **Wald Chi-Square** | **Pr > ChiSq** | | **1** | **SES\_low** | 1 | 3 | 0.2551 | 0.6135 | | **2** | **SES\_mid** | 1 | 2 | 0.9590 | 0.3274 |      | **Analysis of Maximum Likelihood Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** | | **Intercept** | 1 | -2.3350 | 0.5111 | 20.8713 | <.0001 | | **Age** | 1 | 0.0293 | 0.0132 | 4.9455 | 0.0262 | | **Sector** | 1 | 1.6734 | 0.4873 | 11.7906 | 0.0006 | |

**/\* Variable selection \*/**

**/\* - here, display the 'best' two models containing**

**between 1 and 2 predictors \*/**

**proc logistic data=outbreak;**

**model Disease(event = '1') = Age SES\_mid SES\_low Sector /**

**selection=score best=2 start=1 stop=2;**

**title1 'Variable Selection: best by score';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Variable Selection: best by score*** |   Probability modeled is Disease=1.     | **Regression Models Selected by Score Criterion** | | | | --- | --- | --- | | **Number of Variables** | **Score Chi-Square** | **Variables Included in Model** | | **1** | **14.7805** | Sector | | **1** | **7.5802** | Age | | **2** | **19.5250** | Age Sector | | **2** | **15.7058** | SES\_low Sector | |

**/\* Check for outliers using the half-normal probability**

**plot with simulated envelope**

**-- note that this macro can be slow for large**

**sample sizes**

**\*/**

**/\* Alternative way to access simulated envelope macro:**

**filename macrourl "<add SAS studio file path here>";**

**%include macrourl;**

**OR Just load the one line version of the macro provided on canvas into your SAS session**

**\*/**

**%macro simEnv(dataset, response, predictors, N); proc ...**

**/\* Call simEnv macro; arguments:**

**dataset = name of dataset containing data**

**response = name of response variable in dataset,**

**coded 0/1**

**predictors = name(s) of predictor variable(s)**

**(if multiple, separated by spaces)**

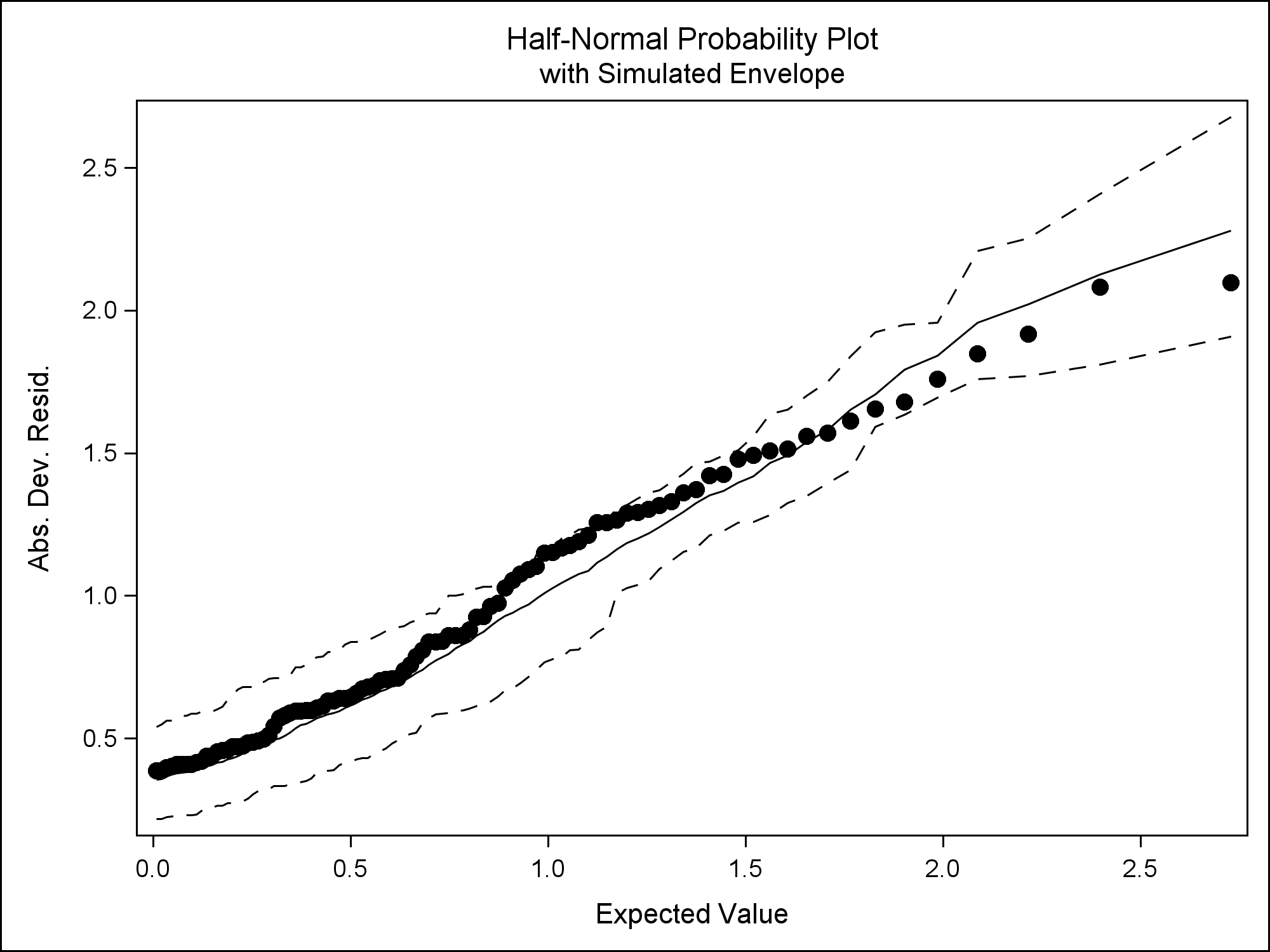
**N = number of observations (sample size)**

**\*/**

**%*simEnv*(dataset = outbreak, response = disease,**

**predictors = Age SES\_mid SES\_low Sector, N=98);**

**run;**



**/\* Check for influential observations \*/**

**proc logistic data=outbreak**

**plots(only label)=(phat influence dpc roc);**

**ID Case;**

**model Disease(event = '1') = Age SES\_mid SES\_low Sector;**

**run;**

|  |
| --- |
|  |

**/\* Look at suspect observation \*/**

**proc print data=outbreak;**

**where Case = 48;**

**var Case Age SES\_mid SES\_low Sector Disease;**

**title1 'Suspect point';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Suspect point*** |  | **Obs** | **Case** | **Age** | **SES\_mid** | **SES\_low** | **Sector** | **Disease** | | --- | --- | --- | --- | --- | --- | --- | | **48** | 48 | 65 | 0 | 1 | 1 | 0 | |