**Defining Dummies**

The following will create a dataset called Train2 that has the dummies defined.

**Data save.train2;**

**set save.train;**

if vjobmos <= **6** then vjobmos1 = **1**; else vjobmos1 = **0**;

if vjobmos > **6** and vjobmos <= **11** then vjobmos2 = **1**; else vjobmos2 = **0**;

if vjobmos > **60** then vjobmos3 = **1**; else vjobmos3 = **0**;

if mileag <= **50000** then mileag1 = **1**; else mileag1 = **0**;

if mileag > **85000** and mileag <= **95000** then mileag2 = **1**; else mileag2 = **0**;

if mileag > **95000** and mileag <= **105000** then mileag3 = **1**; else mileag3 = **0**;

if mileag > **105000** and mileag <= **116000** then mileag4 = **1**; else mileag4 = **0**;

if mileag > **116000** then mileag5 = **1**; else mileag5 = **0**;

if hst03x >= **0** and hst03x <= **1** then hst03x1=**1**; else hst03x1 = **0**;

if hst03x = **3** then hst03x2=**1**; else hst03x2 = **0**;

if hst03x = **4** then hst03x3=**1**; else hst03x3 = **0**;

if hst03x >= **5** and hst03x <= **7** then hst03x4=**1**; else hst03x4 = **0**;

if hst03x >= **8** then hst03x5=**1**; else hst03x5 = **0**;

if ageotd > **0** and ageotd <= **42** then ageotd1=**1**; else ageotd1 = **0**;

if ageotd > **84** and ageotd <= **115** then ageotd2=**1**; else ageotd2 = **0**;

if ageotd > **115** and ageotd <= **160** then ageotd3=**1**; else ageotd3 = **0**;

if ageotd > **160** then ageotd4=**1**; else ageotd4 = **0**;

if vage > **27** and vage <=**32** then vage1=**1**; else vage1=**0**;

if vage > **32** and vage <=**50** then vage2=**1**; else vage2=**0**;

if vage > **50** then vage3=**1**; else vage3=**0**;

**run**;

**quit**;

The above is just an example. Use your crosstabs to determine how to define the dummies. Check that the dummies look correct in save.train2 by running the following:

**proc** **print** data=save.train2 (obs=**10**);

var

vjobmos vjobmos1--vjobmos3

mileag mileag1--mileag5

hst03x hst03x1--hst03x5

ageotd ageotd1--ageotd4

vage vage1--vage3;

format \_all\_;

**run**;

**Running Regression Analysis**

The following program runs the regression and stores the estimates in a file called *estfile*. The dummy variable list should, of course, be your own. Please edit. Assume that train2 is the dataset that has the dummy variables defined in it.

**proc** **reg** data=save.train2 outest=estfile;

**bgscore**: model Good =

vjobmos1--vjobmos3

mileag1--mileag5

hst03x1--hst03x5

ageotd1--ageotd4

vage1--vage3

;

**run**;

Note that you run the regression with all the dummies for all the variables together. Do not run separate regressions for each variable!

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Scoring (predicting) in SAS**

The following will score the and save the output in save.scrtrain. The following **score program** **must be repeated for the validation dataset**. Be careful, however, to run the **regression on the training set only**! The model is built on the training set, and tested on both training and validation sets.

**proc** **score** data=save.train2 score=estfile type=parms out=save.scrtrain;

var

vjobmos1--vjobmos3

mileag1--mileag5

hst03x1--hst03x5

ageotd1--ageotd4

vage1--vage3;

**run**;

**quit**;

Add this format to the format.sas program and run it.

VALUE bgscore

**0**-**.05**='0 to 50' **.05**<-**.10**='51 to 100'

**.10**<-**.15**='101 to 150' **.15**<-**.20**='151 to 200'

**.20**<-**.25**='201 to 250' **.25**<-**.30**='251 to 300'

**.30**<-**.35**='301 TO 350' **.35**<-**.40**='351 TO 400'

**.40**<-**.45**='401 TO 450' **.45**<-**.50**='451 TO 500'

**.50**<-**.55**='501 TO 550' **.55**<-**.60**='551 TO 600'

**.60**<-**.65**='601 TO 650' **.65**<-**.70**='651 TO 700'

**.70**<-**.75**='701 TO 750' **.75**<-**.80**='751 TO 800'

**.80**<-**.85**='801 TO 850' **.85**<-**.90**='851 TO 900'

**.90**<-**.95**='901 TO 950' **.95**<-**1.00**='951 TO 1000'

**1.00**<-HIGH='OVER 1000'

;

Now you are ready to print a crosstab of the scores (predictions) against the real Good/Bad values from the sample. Use ***Proc Freq*** to do so. This will help you assess how well the predictions are able to match reality.

ODS html …*fill this in to create html output file;*

**proc** **freq** data=save.scrtrain;

tables bgscore\*good;

format bgscore bgscore. ;

**run**;

ODS html close;

Read into Excel. Complete the KS spreadsheet.