

EDF 5401 HW 4

Due 08/ 02/ 10

Please type your answers as much as possible.

Total 100 points

For this assignment you will first need to create the outcome score we will use in our analyses. We'll use the **hw3.sav** data set and the outcome of interest is science test performance. We'll call the outcome "science" and it will be defined using the science standardized score *f1txsstd*. Follow these steps to create *science*.

Using the **Transform** menu, select **Recode, Into different variables**. Put the test score *f1txsstd* into the window labeled "numeric variable ->output variable" and type the name *science* in the box for the output variable. Hit the button that says "change" underneath of the name.

Next click the "old and new values" button. You'll need to move three sets of values into the "Old -> New" box.

- First select "system- or user-missing" in the old value side and "system missing" on the new value side. Click "add" to add this to the list.
- Next select Range: Lowest through ___ in the old value side and type 50 into the box. Select Value and type in 0 into the blank box on the new value side. Click "add" to add this to the list.
- Last select "All other values" in the old value side and choose value and type in 1 for the value on the new value side. Click "add" to add this to the list.

Click continue, then back on the main recode window click OK. This should create your new variable.

Before running any analyses,

- First, reset the cases. Do this by using the menu Data, Select cases, and choose All cases and Then choose OK. You won't need to do this if you have not omitted cases using a filter variable as described above (e.g., this won't be needed if this is the first analysis you do in an SPSS session).
- Second, filter out cases that do not have scores on *f1txsstd* (science standardized score), which we will use in our analyses. Do this by using the menu Data, Select cases, and choose "Use filter variable". Move the variable name *f1txsstd* into the box. Then choose OK.

Check to see that you've gotten the correct values by running a **Frequencies** analysis on the **Analyze, Descriptive Statistics** menu. You should get the output below.

science

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	60	37.7	44.1	44.1
	1.00	76	47.8	55.9	100.0
	Total	136	85.5	100.0	
Missing	System	23	14.5		
	Total	159	100.0		

1. First use the predictor *f1sex* in a logistic regression to predict *science*. Save the predicted probabilities and the predicted group membership values. Answer the following questions about this model.

a. (7 points) Which gender is more likely to pass the science test? (You can answer this by creating a crosstabulation of *f1sex* by *PRE_values*. Use the **Analyze, Descriptive Statistics, Crosstabs** menu to do this).

b. (7 points) Create a crosstabulation of *f1sex* by *science*. Compute the proportions of males who pass and of females who pass (Use the **Analyze, Descriptive Statistics, Crosstabs**, select **Cells**, choose **Row in Percentage** box). Which group is more likely to pass the math test? How do these proportions relate to the values you saw in answering item a?

c. (9 points) Is *f1sex* a significant predictor of science? Interpret the meaning of the significance test – what does it tell you about the likelihoods of males and females passing the test? (Also, report the odds ratio, and validate the odds ratio by computing odds for males and females and odds ratio of that - use the probabilities of males and females) Does this model predict any participants who will fail to pass the test?

d. (5 points) How many students in the entire sample are predicted to pass the exam, based on this analysis? Here you can examine the saved values of PGR – predicted group membership. How many students have PGR = 1? (To do this, create a crosstabulation of *PGR* and *PRE_values*.)

e. (7 points) Comment on the quality of this model. How much variation in passing performance does it explain?

2. Next run a model with *f1sex* and math standardized score performance *f1txmstd* as predictors. You may want to run the analysis using blocks – with *f1sex* in the first block and *f1txmstd* in the second block.

a. (7 points) Consider the new equation – are both predictors significant? Test the hypotheses that $\beta_{f1sex} = 0$ and $\beta_{f1txmstd} = 0$.

b. (7 points) Use the saved predicted probabilities (the *PRE_* values) to make a scatterplot. Put math score, *f1txmstd*, on the X axis and *PRE_* for this model as Y. Use “Set markers by” with *f1sex* to obtain different markers for males and females. Explain the plot. What is this plot showing you?

c. (7 points) Now consider the predicted group membership values (the *PGR_* values). Make a scatterplot with math score, *f1txmstd*, on the X axis and *PGR_* for this model as Y. Also use “Set markers by” with *f1sex* to obtain different markers for boys and girls. (I suggest using different colors or shapes for the markers in this plot, because the differences do not show up so well using symbols). Explain the plot. What is this plot showing you?

d. (7 points) Does adding the math score to the model change the model significance tests? Remark on the change in the chi-square due to the addition of *fltxmstd*.

e. (7 points) Now that you have added the math score to the model, how has the predictive power of the model changed? Comment on both the R^2 -like measures as well as the classification table.

f. (7 points) Finally consider model fit as represented by the Hosmer–Lemeshow test. Is this test significant? Are there any values in the H-L contingency table that appear problematic to you?

3. Now add one other predictor to the model. Add the variable *fltxsstd*. Note this is the score we used to create the outcome *science*. You would NEVER do this in practice, in general because you would not have both the continuous and categorical versions of an outcome. Also we will see the results are quite unusual.

a. (6 points) Examine the model test and the individual slope tests. What do you see? Are *flsex* and *fltxmstd* still important? Explain what you see.

b. (6 points) Examine the Hosmer–Lemeshow test and Hosmer–Lemeshow contingency table. What do you notice about the values?

c. (6 points) What about the R^2 values for this model. Comment in particular about the difference between the Cox and Snell and Nagelkerke values.

d. (5 points) Comment on the values of the predicted probabilities PRE_{\cdot} for this model. How do these values differ from the values from your earlier models? Why do you get only two values of PRE from this model?