

Homework #4

1. First use the predictor *flsex* 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. Which gender is more likely to pass the science test? (You can answer this by creating a crosstabulation of *flsex* by *PRE_* values. Use the Analyze, Descriptive Statistics, Crosstabs menu to do this).

COMPOSITE SEX * Predicted probability Crosstabulation

			Predicted probability		Total
			.50685	.61905	
COMPOSITE SEX	MALE	Count	0	63	63
		% within COMPOSITE SEX	.0%	100.0%	100.0%
	FEMALE	Count	73	0	73
		% within COMPOSITE SEX	100.0%	.0%	100.0%
Total		Count	73	63	136
		% within COMPOSITE SEX	53.7%	46.3%	100.0%

Based on the crosstabulation table of *flsex* by predicted *science* scores generated by SPSS, it is more likely for males to pass the science test than females. The predicted probability of passing is 0.61905 for males and 0.50685 for females (as shown at the top of the table), which means approximately 61.9% of males should pass, compared with only 50.7% of females. There are only two predicted probabilities because only *flsex*—a dichotomous predictor with two values—is included in the model. However, both predicted probabilities are more than the cut-off value of 0.50.

- b. Create a crosstabulation of *flsex* 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 science test? How do these proportions relate to the values you saw in answering item a?

COMPOSITE SEX * Science test performance Crosstabulation

			Science test performance		Total
			.00	1.00	
COMPOSITE SEX	MALE	Count	24	39	63
		% within COMPOSITE SEX	38.1%	61.9%	100.0%
	FEMALE	Count	36	37	73
		% within COMPOSITE SEX	49.3%	50.7%	100.0%
Total		Count	60	76	136
		% within COMPOSITE SEX	44.1%	55.9%	100.0%

Based on the crosstabulation table of *flsex* by observed *science* scores generated by SPSS, it is more likely for males to pass the science test than females. The proportions we see in the observed *science* scores are the same as the predicted probabilities from item (a): 61.9% of males pass, compared with only 50.7% of females. This is because the probabilities are based on only one categorical variable, *flsex*; the observed proportions of group membership are the only data used to calculate the predicted probability of students' falling in the "pass" or "fail" groups.

- c. **Is *flsex* 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?**

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1.733	1	.188
	Block	1.733	1	.188
	Model	1.733	1	.188

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	f1sex(1)	.458	.349	1.719	1	.190	1.581
	Constant	.027	.234	.014	1	.907	1.028

a. Variable(s) entered on step 1: f1sex.

flsex is not a significant predictor of *science* scores, as indicated by the significance tests shown in the SPSS output for the logistic regression model (above). For the null hypothesis of $H_0: \beta_{f1sex} = 0$, the p-value calculated based on the Wald test statistic is $p = 0.190$, which is greater than $\alpha = 0.05$. Therefore we cannot reject the null hypothesis and conclude that *flsex* has no predictive power in the model. Our overall model test is also nonsignificant ($p = 0.188$), as it should be since there is only this one predictor in the model. This result tells us that there is no significant difference in the population between how likely males and females are to pass the science test. This can also be concluded from the fact that all the predicted probabilities are

above the 0.50 cutoff, meaning all students are predicted to pass the science test; however, 44.1% of students failed.

The odds ratio is shown by the Exp(B) column in the output. The value shown, 1.581, means that males (the focal group) are 1.581 times more likely to pass the science test than females (the reference group). We can validate this by calculating the odds for males and females as follows:

$$odds_{males} = \frac{p_{males}}{1 - p_{males}} = \frac{0.61905}{1 - 0.61905} = \frac{0.61905}{0.38095} = 1.625$$

$$odds_{females} = \frac{p_{females}}{1 - p_{females}} = \frac{0.50685}{1 - 0.50685} = \frac{0.50685}{0.49315} = 1.028$$

$$odds_{males/females} = \frac{1.625}{1.028} = 1.581$$

This value is the same as calculated by SPSS.

This model does not predict any participants who will fail to pass the test, because the predictor used (*f1sex*) is not significant and thus has no predictive power; in addition as mentioned earlier all of the predicted probabilities calculated are above the cut-off of 0.50.

- d. 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.)**

Predicted group * Predicted probability Crosstabulation

		Predicted probability		Total	
		.50685	.61905		
Predicted group	1.00	Count	73	63	136
		% within Predicted group	53.7%	46.3%	100.0%
Total		Count	73	63	136
		% within Predicted group	53.7%	46.3%	100.0%

Based on the crosstabulation table of predicted group (*PGR*) and predicted probabilities (*PRE*) generated by SPSS (shown above), **all** students are predicted to pass the science exam. This is true because the predicted probability of passing for both groups (females and males) are greater than the cut-off of 0.50 (50%). However, this is not a good prediction because it does not distinguish why a student will pass or fail; as shown in item (c) gender (*f1sex*) is not a significant predictor.

- e. Comment on the quality of this model. How much variation in passing performance does it explain?**

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	184.916 ^a	.013	.017

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Based on the Nagelkerke R^2 value shown in the SPSS output (above) of 0.017, this model only explains 1.7% of the variation in passing performance. This indicates the model is not of high quality. The non-significance of the slope test for *flsex* and non-significance of the likelihood ratio test for the whole model also indicate the model is of low quality.

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

a. Consider the new equation – are both predictors significant? Test the hypotheses that $\beta_{flsex} = 0$ and $\beta_{fltxmstd} = 0$.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	f1sex(1)	.841	.484	3.016	1	.082	2.320
	f1txmstd	.189	.031	36.969	1	.000	1.208
	Constant	-9.862	1.667	35.006	1	.000	.000

a. Variable(s) entered on step 1: f1txmstd.

The logistic regression equation for the new model is

$$\frac{\hat{p}_i}{1 - \hat{p}_i} = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})$$

where

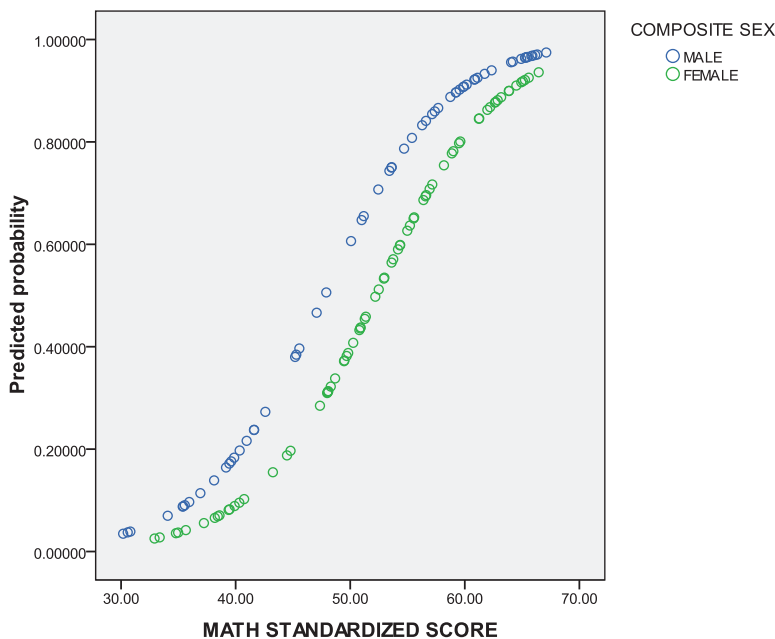
- \hat{p}_i is the predicted probability of passing the science exam (*science*) for student i ;
- β_0 is the intercept, the value of *science* when the predictors are equal to 0;
- β_1 is the amount of change, in the population, in the probability of passing the science exam (*science*) for every one unit change in gender (*flsex*) – i.e. the difference between the two genders in *science* – holding the other predictor (*fltxmstd*) constant (controlling for it);
- X_{1i} is the gender of student i (as measured by *flsex*), coded as 1 for males (the focal group) and 0 for females (the reference group);
- β_2 is the amount of change, in the population, in the probability of passing the science exam (*science*) for every one unit change in math standardized score performance (*fltxmstd*), holding the other predictor (*flsex*) constant (controlling for it); and
- X_{2i} is the math standardized score performance for student i (as measured by *fltxmstd*).

Plugging in the values generated for these parameters by SPSS (shown above) gives a regression equation of

$$\frac{\hat{p}_i}{1 - \hat{p}_i} = \exp(-9.862 + 0.841X_{1i} + 0.189X_{2i})$$

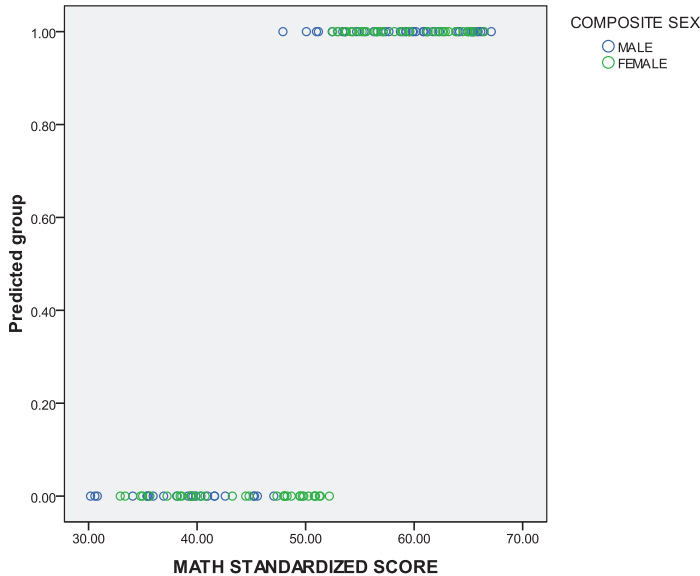
Only *fltxmstd* is significant, as shown by the p-values in the SPSS output. The null hypothesis of $H_0: \beta_1 = 0$ (for the slope of *flsex*) cannot be rejected because the Wald test statistic is not significant, as shown by the p-value of $p = 0.082$ which is not less than $\alpha = 0.05$. However, the null hypothesis of $H_0: \beta_2 = 0$ (for the slope of *fltxmstd*) is rejected because the Wald test statistic is significant, as shown by the p-value of $p < 0.001$ (less than $\alpha = 0.05$).

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



The scatterplot shows us the relationship between math standardized score performance and the predicted probability of passing the science exam. We can see that, as a student’s math score increases, their predicted probability of passing the science exam also increases; the correlation is positive. However, this is not a linear relationship as shown by the curved shape on the scatterplot. The scatterplot also shows that males are predicted to be slightly more likely to pass the science exam, but as shown in question 1 this is not a significant difference between the genders.

- c. Now consider the predicted group membership values (the PGR_ values). Make a scatterplot with math score, *fltxmstd*, on the X axis and PGR_ for this model as Y. Also use “Set markers by” with *flsex* 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?



The scatterplot shows the relationship between math standardized score performance and the predicted group (passing or failing the science exam). We can see that students with math scores less than about 45 are predicted to fail the science exam, while those with math scores greater than about 55 are predicted to pass the science exam. For those scoring between 45 and 55 on math, there is some overlap but in general males are predicted to pass the science exam, while females are predicted to fail it. This can be seen by the preponderance of green (female) dots at the bottom of the graph near *fltxmstd* = 50, compared with the blue (male) dots at the top of the graph near *fltxmstd* = 50.

- d. 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*.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	66.784	1	.000
	Block	66.784	1	.000
	Model	68.517	2	.000

As shown in the “Block 2” results from SPSS (shown above), adding *fltxmstd* (math score) to the logistic regression model changes the model significance tests dramatically. The model is now significant as shown by the chi-square value of $\chi^2 = 68.517$ and associated p-value of

$p < 0.001$. Adding *fltxmstd* contributes 66.784 to the chi-square test statistic, a dramatic increase over the value for “Block 1” (the model with only *flsex*) of 1.733 (also shown below).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1.733	1	.188
	Block	1.733	1	.188
	Model	1.733	1	.188

- e. 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.

The model now has predictive power, unlike the first model with only *flsex*, as shown by the chi-square test and p-value discussed in item (d).

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	118.133 ^a	.396	.530

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The Nagelkerke pseudo- R^2 value has greatly increased, from 0.017 to 0.530. Adding *fltxmstd* to the model increases the amount of variation explained from 1.7% to 53%, an enormous change. The Cox and Snell pseudo- R^2 value also greatly increased, from 0.013 to 0.396; as expected its value is less than the Nagelkerke value.

Classification Table^a

Observed		Predicted		
		Science test performance		Percentage Correct
		.00	1.00	
Step 1	Science test performance	.00	1.00	73.3
	Overall Percentage			77.9

a. The cut value is .500

The classification table (shown above) indicates that most of the predictions are correct; the overall percentage correct is 77.9%. Unfortunately, 16 students were predicted to pass the science test when, in actuality, they failed it; 14 students were predicted to fail but actually passed. These values are circled in blue above.

- f. 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?

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.645	8	.687

Contingency Table for Hosmer and Lemeshow Test

		Science test performance = .00		Science test performance = 1.00		Total
		Observed	Expected	Observed	Expected	
Step 1	1	13	13.309	1	.691	14
	2	12	12.348	2	1.652	14
	3	9	10.418	5	3.582	14
	4	9	8.360	5	5.640	14
	5	7	6.159	7	7.841	14
	6	7	4.188	7	9.812	14
	7	2	2.440	12	11.560	14
	8	1	1.500	13	12.500	14
	9	0	.954	14	13.046	14
	10	0	.323	10	9.677	10

The Hosmer-Lemeshow test, as shown in the SPSS output above, tests whether the observed group membership and predicted group membership match. In this case, the p-value of $p = 0.687$ means that this test is not significant and we conclude that, in this model, the observed and predicted group membership are a good match; the model fits well in this respect. There are no values in the contingency table that appear problematic to us; while the last values (for step 10) for those passing the science test ($science = 1$) are less than the values before them (for step 9), this is only a small decrease and the overall pattern is, as expected, decreasing values for $science = 0$ and increasing values for $science = 1$. We do not expect the expected (predicted) values to perfectly match the observed values, and this is true in this case.

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

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	118.133	1	.000
	Block	118.133	1	.000
	Model	186.649	3	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	f1sex(1)	2.195	1539.332	.000	1	.999	8.976	.000	.
	f1txmstd	-.660	110.068	.000	1	.995	.517	.000	2.529E93
	f1txsstd	20.681	701.596	.001	1	.976	9.590E8	.000	.
	Constant	-990.507	33568.669	.001	1	.976	.000		

a. Variable(s) entered on step 1: f1txsstd.

As shown in the model test above (output by SPSS), the model as a whole is significant ($p < 0.001$). However, as shown by the “Variables in the Equation” table (above) none of the predictors in the model are significant; *f1sex* and *f1txmstd* are no longer important or significant predictors in the model with the addition of *f1txsstd*. We believe this has occurred because there will naturally be a strong correlation between *f1txsstd* (a continuous variable measuring science standardized scores) and *science* (a categorical variable measuring whether someone passed the science test, based in turn on science standardized scores). This correlation has caused the model as a whole to be highly significant and explain all of the variation in *science*, as shown by the Nagelkerke pseudo- R^2 value. *f1sex* and *f1txmstd* are, in relation to *f1txsstd*, not as significant and as such the tests for these predictors have become non-significant in the SPSS output. We are not sure why *f1txsstd* is not shown as significant in this model, but assume it is because *science* is derived directly from *f1txsstd*, causing this anomaly.

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

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	5	1.000

Contingency Table for Hosmer and Lemeshow Test

		Science test performance = .00		Science test performance = 1.00		Total
		Observed	Expected	Observed	Expected	
Step 1	1	14	14.000	0	.000	14
	2	14	14.000	0	.000	14
	3	14	14.000	0	.000	14
	4	14	14.000	0	.000	14
	5	4	4.000	10	10.000	14
	6	0	.000	1	1.000	1
	7	0	.000	65	65.000	65

The Hosmer-Lemeshow test, as shown in the SPSS output above, shows a p-value of $p = 1.000$. While this means that this test is not significant and would imply a good fit between observed group membership and predicted group membership, the contingency table does not fit the expected pattern. The transition towards lower values (for *science* = 0) and higher values (for *science* = 1) is far from smooth, the changes coming in large jumps at steps 5, 6, and 7. Step 6 also breaks the pattern of increasing values for *science* = 1. The biggest problem is that the observed and expected values are identical; this is because *flxsstd* allows us to predict *science* with perfect accuracy, since the latter is derived directly from *flxsstd*.

- c. **What about the R^2 values for this model. Comment in particular about the difference between the Cox and Snell and Nagelkerke values.**

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000 ^a	.747	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

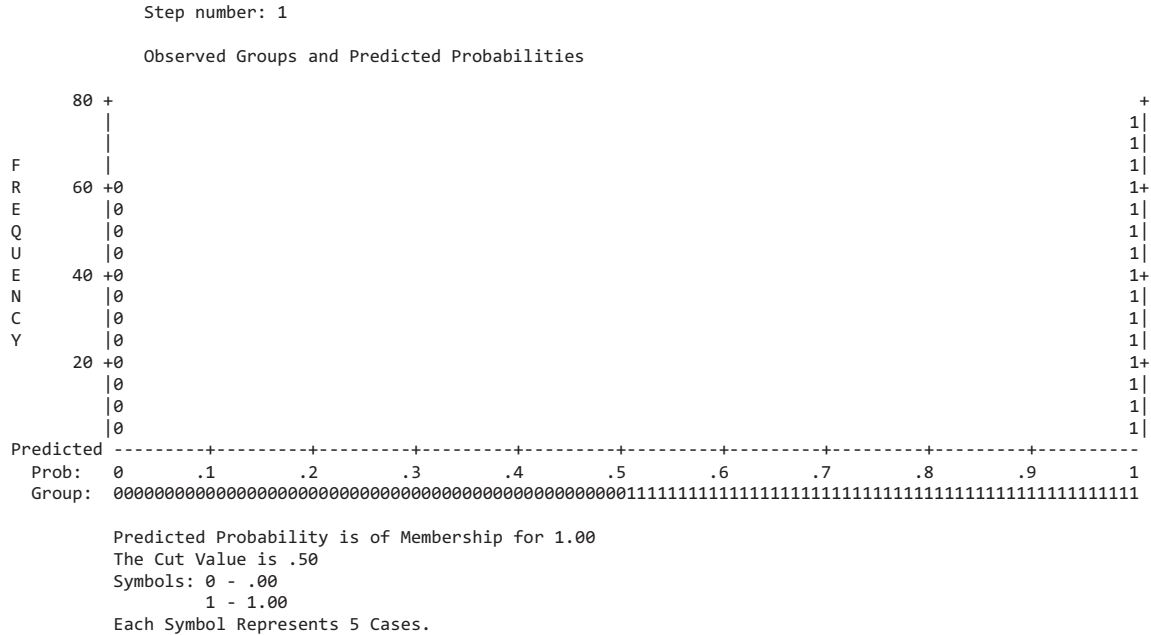
Science test performance

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	60	44.1	44.1	44.1
	1.00	76	55.9	55.9	100.0
Total		136	100.0	100.0	

The Cox and Snell and Nagelkerke pseudo- R^2 values (in the SPSS output shown above) are very large. The Cox and Snell value cannot be compared directly with a value of 1 because its maximum value depends on the proportion of the sample in the focal group. In this case, as shown in the frequency table above 55.9% of the sample is in the focal group (passing the science test, that is, *science* = 1), and the maximum value for the Cox and Snell value is also 0.747. Therefore it is equal to its maximum value. The Nagelkerke value, on the other hand, can

be compared against 1 because it modifies the Cox and Snell value by dividing it by its maximum value. $0.747/0.747 = 1$, as shown in the SPSS output above.

- d. Comment on the values of the predicted probabilities PRE_ 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?**



As shown in the classification plot above, the only predicted probabilities for this model are 0.00 and 1.00; there are approximately 60 of the former and 80 of the latter. The predicted probabilities in the SPSS data file (PRE_3 in our case) also agree with the classification plot; there are no values other than 0.00 and 1.00 listed. These values are different than the earlier models because they are at the extremes, rather than being spread out across the possible probabilities. There are also only two values included here. The two extreme predicted values can be explained by the fact that this model perfectly predicts *science* scores based on *fltxsstd*, since the former is derived directly from the latter. While the first model also only had two predicted values, this was because it only included *flsex*, a dichotomous predictor with two values, and thus had limited predictive power. Those predicted values were also not at the extremes shown in this model.