Problems to Turn In

Problems 4 and 5 involve variables from the depression data set that I used as an example in class, courtesy of the text *Practical Multivariate Analysis* (formerly *Computer-Aided Multivariate Analysis*) by Afifi, Clark and May. The complete data set is available on line in multiple formats on the UCLA IDRE web site at

https://stats.idre.ucla.edu/other/examples/pma5/

In fact, IDRE has nice worked out examples from the text in several statistical packages (including Chapter 12 on logistic regression) and you can get all the data sets for the book as text files (see the Appendix link at the bottom of the table). I have provided a subset of the data, recoded to make things easier for the models we will be fitting, as part of this week's homework data file. The variables included in this reduced version are **sex** (1 = female, 0 = male), **drink** (1 for a regular drinker and 0 for not), **depressed** (1 = yes and 0 = no), **CESD** which is a continuous measure of depression on a scale of 0-60 with higher being worse, **income** in thousands of dollars, **age** in years and **health status** which is a four level variable with 0 = Excellent, 1 = Good, 2 = Fair and 3 = Poor.

4. Drink is a Downer (ACM 12.9):

For this problem we use drinking status as the outcome and are interested in sex and depression (either categorical or continuous) as predictors.

(a) Fit a simple logistic regression of drink on sex and use it to find the odds of being a regular drinker for women and men and the corresponding odds ratio. Does there appear to be a sex difference?

The odds of being a regular drinker for women is $\exp(1.7813 - 0.6310 * 1) = 3.159$ and the odds of being a regular drinker for men is $\exp(1.7813) = 5.938$. The corresponding odds ratio for being a regular drinker in women vs men is $OR_{\text{F vs M}} = 3.159/5.938 = 0.532$ (95% CI = [0.284, 0.998]). Based on the value of the odds ratio and the 95% confidence interval reported in the SAS output, there does appear to be a sex difference in odds of drinking status (95% CI does not overlap with 1, indicating a true odds ratio of less than 1 with 95% confidence). This suggests that the odds of being a regular drinker in women are lower than the odds of being a regular drinker in men.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.0975	1	0.0429
Score	3.9441	1	0.0470
Wald	3.8676	1	0.0492

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.7813	0.2702	43.4499	<.0001
sex	1	-0.6310	0.3209	3.8676	0.0492

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	0.532	0.284	0.998

Figure 1: Logistic regression of drink on sex

(b) Repeat part (a) but do the calculations separately for people who are depressed and people who are not

depressed and compare the odds ratios for the two groups. (Note: for this part you do not need to give the individual odds by sex; just focus on the odds ratios.)

The odds ratio for being a regular drinker for depressed women vs depressed men is $OR_{\text{F vs M,depressed=1}} = 1.179$ (95% CI = [0.205, 6.789]). The odds ratio for being a regular drinker for non-depressed women vs non-depressed men is $OR_{\text{F vs M,depressed=0}} = 0.461$ (95% CI = [0.234, 0.907]).

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.0332	1	0.8555
Score	0.0339	1	0.8540
Wald	0.0338	1	0.8541

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.3863	0.7906	3.0749	0.0795
sex	1	0.1643	0.8934	0.0338	0.8541

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	1.179	0.205	6.789

Figure 2: Logistic regression of drink on sex in depressed individuals

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5.3642	1	0.0206
Score	5.1666	1	0.0230
Wald	5.0223	1	0.0250

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.8268	0.2880	40.2469	<.0001
sex	1	-0.7743	0.3455	5.0223	0.0250

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	0.461	0.234	0.907

Figure 3: Logistic regression of drink on sex in non-depressed individuals

(c) What do the results from part (b) suggest about whether there is an interaction between sex and depression? Explain carefully what such an interaction would mean. Then fit an appropriate model and test whether the interaction is significant. (The depbysex interaction variable has been provided for your convenience.) Do the results confirm your theory? If not, what do you think might have happened?

The results from part (b) suggest that there may be an interaction between sex and depression since the direction of parameter estimates, the odds ratios, and their corresponding 95% CIs are extremely different when running the same analysis stratified by depression status. However, the interaction also may not be significant due to the results from the depressed individuals - we see a lack of significance in the test statistic and the predictors in the output for the logistic regression of drink on sex in depressed individuals (Figure 2). An interaction between **depressed** and **sex** would mean that the log odds/odds for being a regular drinker if the individual is depressed is modified (most likely increased, either additively on the log odds scale, or multiplicatively on the odds ratio scale) if the individual is female instead of male.

The results confirm my theory that the interaction between sex and depression was non-significant. This is possibly because the parameter estimates for that stratified analysis presented in Figure 2 had such wide confidence intervals and low p-values that was then subsequently captured in this model fitting both predictors and the interaction term.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5.6178	3	0.1318
Score	5.5046	3	0.1384
Wald	5.3734	3	0.1464

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.8269	0.2880	40.2469	<.0001
sex	1	-0.7743	0.3455	5.0223	0.0250
depressed	1	-0.4406	0.8414	0.2742	0.6005
depbysex	1	0.9386	0.9579	0.9602	0.3271

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	0.461	0.234	0.907
depressed	0.644	0.124	3.349
depbysex	2.556	0.391	16.711

Figure 4: Logistic regression of drink on sex and depression status, including interaction

(d) Instead of using the case indicator one could instead use the continuous depression rating, CESD. Fit a logistic regression of drink on sex, CESD and their interaction. (the cesdbysex interaction has been provided for your convenience.) Give as precise an interpretation as you can of this interaction. In particular you should carefully explain the meanings of the regression coefficients b_1, b_2 and b_3 and their corresponding odds ratios and create a rough sketch of the log odds as a function of sex and depression score. Based on your model does it seem as if there is a significant interaction? Are the results more or less significant than those in part (c)? Explain what the results tell you in real-world terms and give a possible explanation for why the test has come out the way it has.

β_1 = Regression coefficient for sex. For male individuals, there is a 1.1156 decrease (difference) in log odds of being a regular drinker as compared to females, all else constant. On the odds ratio scale, we say that there is a 67.2% ($\exp(-1.1156) = 0.328$) decrease in odds of being a regular drinker if the individual is male instead of female, all else constant. This regression coefficient was statistically significant, indicating that there is a decrease (difference) in log odds/odds of being depressed if an individual is male instead of female.

β_2 = Regression coefficient for CESD. For a male individual, the log odds of being a regular drinker decrease by 0.0419 for each additional point increase in CESD score, all else constant. On the odds ratio scale, we say that there

is a 4.1% ($\exp(-0.0419) = 0.959$) decrease in odds of being a regular drinker in a male individual, all else constant. This regression coefficient was not statistically significant, indicating that we cannot conclude that there is a decrease in log odds/odds of being depressed when CESD score increases.

β_3 = Regression coefficient for the interaction of CESD with sex. For each additional point increase in CESD score, there is an additional 0.0566 increase in the log odds of being a regular drinker for female individuals compared to male individuals. On the odds ratio scale, we say that there is a 5.8% ($\exp(0.0566) = 1.058$) increase in odds of being a regular drinker for each additional point increase in CESD score for female individuals compared to male individuals. This regression coefficient representing the interaction of CESD with sex was not statistically significant, indicating that we cannot conclude that there is an additional increase in the log odds/odds of being a regular drinker for female individuals when CESD score increases.

Roughly (ignoring significance), the log odds are lower if the individual is male instead of female, holding all else constant, the log odds are lower as CESD score increases, holding all else constant, and as CESD score increases the log odds have an additional linear increase for female individuals compared to male individuals. However, since the regression coefficients for CESD score and the interaction term were not statistically significant, we can only robustly conclude that the log odds are lower if the individual is male instead of female.

Based on my model it does not seem as if there is a significant interaction between CESD and sex (p = 0.135). The results of the interaction term, however, are more significant than those in part (c) (p=0.327). Additionally, the p-values for the parameter estimates were also more significant in this model compared to those in the model from part (c).

For real world explanations, see individual coefficient explanations.

The increased significance in this model as compared to the previous model suggests that CESD score provides us more information/evidence for determining the outcome of drinking status as compared to the dichotomous variable of depression status. While the previous results in earlier parts of this question suggest there should be an interaction, we do not see a significant interaction possibly because of skew in how many individuals are depressed or not depressed (there are 50 individuals categorized as depressed, 244 individuals categorized as not depressed, and CESD score is skewed right (skewness = 1.69), distribution has a long right tail). We would need a larger sample or a more balanced sample to be able to discern the presence of this interaction better.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.2504	3	0.1000
Score	5.9273	3	0.1152
Wald	5.7354	3	0.1252

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.1279	0.4010	28.1534	<.0001
sex	1	-1.1156	0.4704	5.6249	0.0177
cesd	1	-0.0419	0.0325	1.6590	0.1977
cesdbysex	1	0.0566	0.0378	2.2393	0.1345

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	0.328	0.130	0.824
cesd	0.959	0.900	1.022
cesdbysex	1.058	0.983	1.140

Figure 5: Logistic regression of drink on sex and CES-D score, including interaction

5. Homework Is Depressing (ACM):

For this problem we will model depression status (yes or no) as a function of sex, age, income and general health (but not how much statistics homework one has!)

(a) Fit the logistic regression model with all 4 predictors, treating health as a continuous variable. Which of these variables appear to be associated with depression status? Are all the relationships in the expected direction? Discuss briefly.

Variables sex and health appear to have a positive association with depression status, and age and income appear to have a negative association with depression status. The parameter estimates for all four predictors appear to be significant at the $\alpha = 0.05$ level. The relationships to appear to be in the expected direction: women tend to have higher odds of depression diagnosis, poorer health (higher value of **health** variable) is typically associated with higher odds of depression, depressive symptoms are more commonly observed in younger individuals, and lower household income is typically associated with increased odds of depression and mental disorders.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	270.125	249.693
SC	273.808	268.111
-2 Log L	268.125	239.693

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.4310	4	<.0001
Score	26.4779	4	<.0001
Wald	23.3210	4	0.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8516	0.5864	2.1092	0.1464
sex	1	0.8938	0.3903	5.2432	0.0220
age	1	-0.0290	0.00978	8.7773	0.0030
income	1	-0.0333	0.0140	5.6749	0.0172
health	1	0.5491	0.1947	7.9535	0.0048

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	2.444	1.137	5.253
age	0.971	0.953	0.990
income	0.967	0.941	0.994
health	1.732	1.182	2.537

Figure 6: Logistic regression of depression status on sex, age, income, and general health, treating health as a continuous variable

(b) Perform an appropriate graphical check to determine whether a linear term adequately describes the relationship between health status and depression. Note: Because the health status variable only has four

values you do not need to go through the tedious process of calculating quartiles–you simply need to rerun the model using an appropriate set of indicator variables (which I have provided for you!) In addition to creating the appropriate plot, discuss whether the model using the indicators appears superior to the original model with the linear term. Is it legitimate to formally test this? If the relationship appears non-linear suggest and test an appropriate transformation of the health variable.

The $\hat{\beta}$s for health status are linear for the bins representing levels 1, 2, and 3 for health. The slope between the $\hat{\beta}$ values representing health = 0 and health = 1 have a different slope. From my observations it seems that the graphical check shows linearity (we can ignore the referent group since it is not something we are focusing on in our model) and we can graphically conclude that a linear term adequately describes the relationship between health status and depression.

The two models are nested (health as a categorical variable estimates 4 coefficients, whereas health as a continuous variable estimates 1 coefficient, so the continuous health model is nested within the categorical health model), so it is legitimate to formally test which of the two models is superior. To do this, we would look at the log likelihoods between the two models to conduct an LRT. The test gives us the test statistic $\chi^2_{LR} = 0.888$ with a p-value of $p = 0.172$. We fail to reject the null hypothesis H_0 and say that the more complex model with indicator variables is no better than the reduced model with just a continuous variable for health.

$$\begin{aligned}
\text{LRT} : \chi^2 &= D_{reduced} - D_{full} \\
&= -2 \log L_{reduced} - 2 \log L_{full} \\
&= 239.693 - 238.805 \\
&= 0.888 \\
\text{p} - \text{value} &= P(\chi^2_{df=3} \geq 0.888) = 0.172
\end{aligned}$$

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	270.125	252.805
SC	273.808	278.590
-2 Log L	268.125	238.805

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	29.3195	6	<.0001
Score	27.6677	6	0.0001
Wald	24.0569	6	0.0005

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7146	0.6006	1.4155	0.2341
sex	1	0.9162	0.3923	5.4542	0.0195
age	1	-0.0293	0.00995	8.6800	0.0032
income	1	-0.0341	0.0140	5.9086	0.0151
health3	1	1.7984	0.6765	7.0671	0.0079
health2	1	0.9851	0.4981	3.9116	0.0480
health1	1	0.2461	0.3859	0.4067	0.5236

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	2.500	1.159	5.393
age	0.971	0.952	0.990
income	0.966	0.940	0.993
health3	6.040	1.604	22.743
health2	2.678	1.009	7.108
health1	1.279	0.600	2.725

Figure 7: Logistic regression of depression status on sex, age, income, and general health, using indicator variables for health

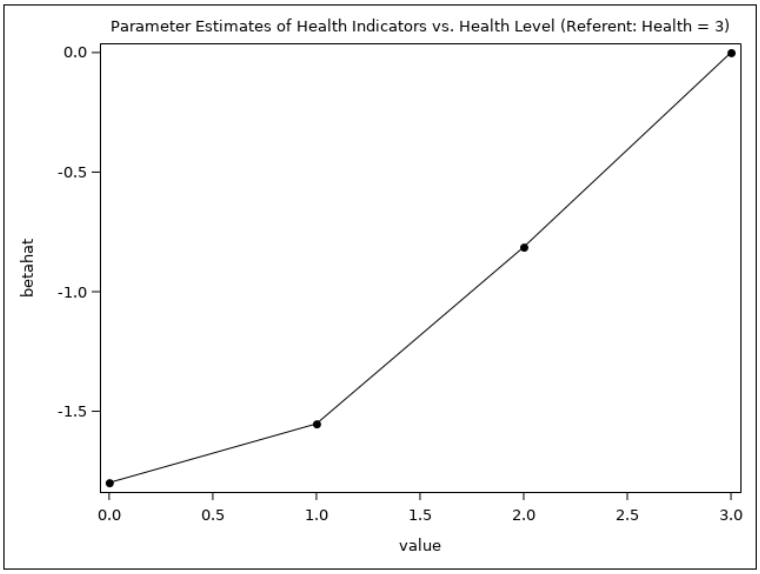

Figure 8: Plot of parameter estimates $\hat{\beta}_{0,1,2,3}$ of health indicator levels vs. health level values (0, 1, 2, 3) using referent: **health** = 0 on the log odds scale

Note: For the rest of the problem, use the model from part (a).

(c) Briefly describe the Hosmer-Lemeshow test for overall goodness of fit and use it to evaluate this model. Does the model appear to fit adequately?

The Hosmer-Lemeshow test for overall goodness of fit tells us how well-calibrated a model is. Predicted probabilities are split into k equal groups (in this case, SAS splits our data into 10 equal groups), and the average predicted probability is calculated for each group. We compare the average predicted probability for each group to the average observed outcome (proportion) for each group, and the model is more well calibrated if the predicted probability for each group comes close to the true proportion for each group. A p-value below a certain significance level indicates that the model is not well calibrated and is not an adequate fit. From our output, we see that the test statistic for the goodness of fit test is $\chi^2 = 11.359$ with $p = 0.182$, indicating that the model does appear to fit adequately according to this goodness of fit test.

Partition for the Hosmer and Lemeshow Test					
		depressed = 1		depressed = 0	
Group	Total	Observed	Expected	Observed	Expected
1	29	0	0.85	29	28.15
2	29	3	1.53	26	27.47
3	29	2	2.27	27	26.73
4	29	1	2.92	28	26.08
5	29	4	3.69	25	25.31
6	29	6	4.36	23	24.64
7	29	3	5.48	26	23.52
8	29	8	6.91	21	22.09
9	29	13	8.39	16	20.61
10	33	10	13.61	23	19.39

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
11.3589	8	0.1822

Figure 9: Hosmer-Lemeshow goodness of fit test for logistic regression of depression status on sex, age, income, and general health, treating health as a continuous variable

(d) Obtain one or more pseudo-R2 values for this data set and explain as carefully as you can what they tell you about the model performance.

SAS gives the likelihood-based pseudo R-square measure and its rescaled measure (which allows the maximum value to reach 1) when specifying the `lackfit rsquare` option in the `proc logistic` procedure. For the model in part (a), SAS tells us that the likelihood-based pseudo R-square is $R^2 = 0.0922$ and the max-rescaled R-square is $R^2_{adj} = 0.1541$. The two R-square measures have quite low values, indicating that our model is quite far from the best possible fit and that the model has a relatively poor fit. This may indicate that sex, age, income, and general health still do not capture a lot of the variability in what is predictive of depression status and that there are many other important factors.

R-Square	0.0922	Max-rescaled R-Square	0.1541

Figure 10: Likelihood-based pseudo R-square measure and corresponding rescaled measure for logistic regression of depression status on sex, age, income, and general health, treating health as a continuous variable

(e) Compute (i) the true positive rate (sensitivity) and (ii) the true negative rate (specificity)–or if you prefer the false negative rate for this data set for a range of threshold values and plot them to identify a good cutoff for predicting that a person will be depressed. What are your sensitivity, specificity and overall error rate at this cutoff? Do you think the model performs well in this sense?

Sensitivity and specificity for this data set were computed at a range of values from 0.1 to 0.9 with step size of 0.05 to yield more values of sensitivity and specificity for a more accurate plot. The plot below shows the two curves superimposed on one plot to show how sensitivity and specificity vary with different probability cutoffs p_c. To keep the overall error rate consistent in both sensitivity and specificity, I identified that a good probability cutoff value would be where the sensitivity and specificity curves intersect, yielding a threshold value of $p_c = 0.17$ (marked by the dashed vertical line on the plot as well). At this threshold, sensitivity and specificity are around 66% each, and the overall error rate is thus equal to 100% - 66% = around 34%. The model seems to not perform that well in this sense,

as the overall error rate is still quite high for a prediction model and the sensitivity and specificity are quite that high and indicate that the true positive rate and true negative rate are only slightly better than random guessing.

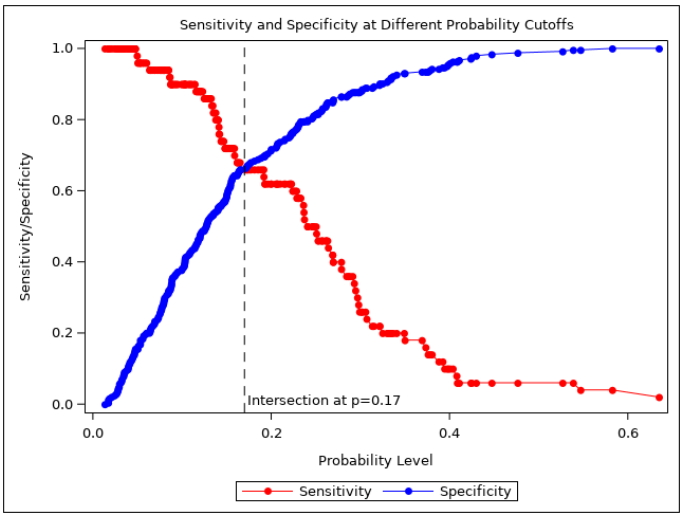

Figure 11: Plot of sensitivity and specificity over probability values from 0 to 1 for logistic regression of depression status on sex, age, income, and general health, treating health as a continuous variable

(f) Obtain the ROC curve to go with your calculations from part (e) and the corresponding AUC value. Does the model perform well by this standard?

The ROC curve yields a curve that rises upwards away from the diagonal line (see Figure 9) with an Area Under the Curve (AUC) value of 0.7314. The AUC value is above the general rule of wanting AUC values ≥ 0.7, indicating that the model performs fairly well and indicates that the model and included variables provide predicted information that is better than random guessing.

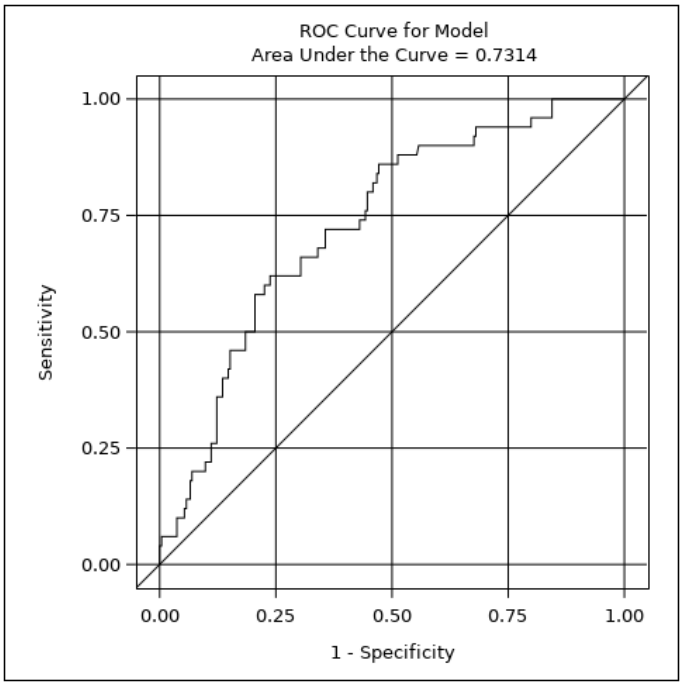

Figure 12: ROC curve and corresponding AUC for logistic regression of depression status on sex, age, income, and general health, treating health as a continuous variable

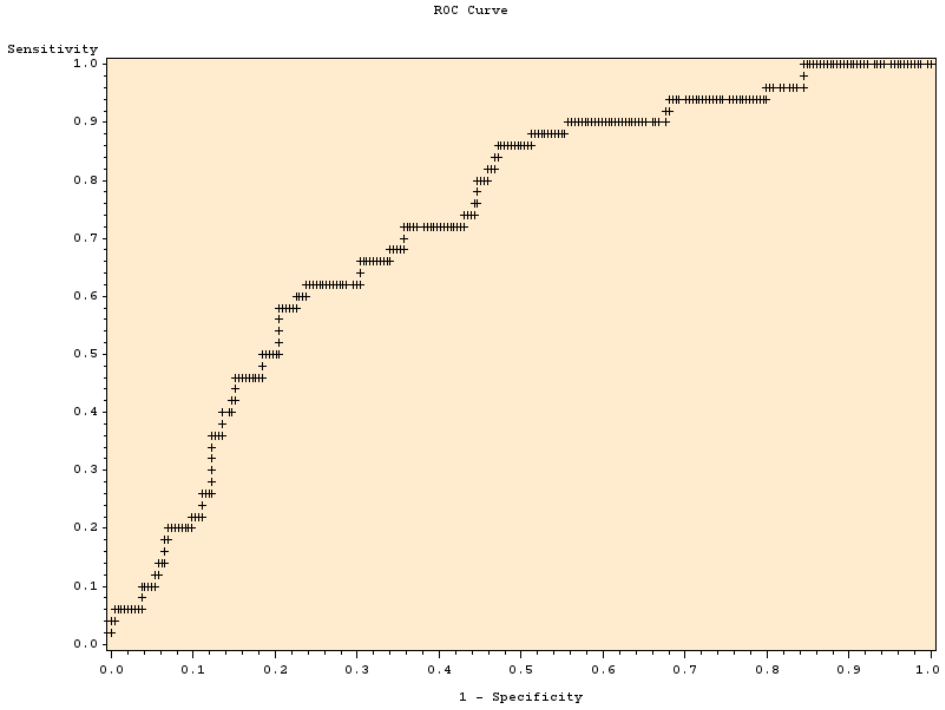

Figure 13: Alternate visualization of ROC curve for logistic regression of depression status on sex, age, income, and general health, treating health as a continuous variable