STAT 659 — Exam 2 **Name:** _____
Summer 2014

INSTRUCTIONS FOR THE STUDENT:

1. You have exactly 70 minutes to complete the exam.
2. There are 12 pages including this cover sheet and 5 pages of SAS output.
3. Each lettered part of a question is worth 8 points unless otherwise marked.
4. Please answer all questions.
5. Show all your work on the test booklet.
6. Do not discuss or provide any information to any one concerning any of the questions on this exam or your solutions until I post the solutions.
7. You may use a calculator that does not have the capability of phoning, texting, or accessing the internet and two $8\frac{1}{2} \times 11$ formula sheets (you may use both sides). Do not use the textbook or class notes.
8. Carry out tests at level 0.05 unless otherwise stated.
9. Be sure to clearly state the hypotheses, the test statistic and its value, and conclusion for all tests.

I attest that I spent no more than 70 minutes to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature_____

INSTRUCTIONS FOR PROCTOR:

(1) Record the time at which the student starts the exam: _____

(2) Record the time at which the student ends the exam: _____

(3) Immediately after the student completes the exam, please scan the exam to a .pdf file and have the student upload it to webassign.

(4) Collect all portions of this exam at its conclusion. Do not allow them to take any portion with them.

(5) Please keep these materials until July 30, at which time you may either dispose of them or return them to the student.

I attest that the student has followed all the INSTRUCTIONS FOR THE STUDENT listed above and that the exam was scanned into a pdf and uploaded to webassign in my presence:

Proctor's Signature_____

Some Chi-Squared Percentiles

	Right-Tail Probability			
df	0.100	0.050	0.025	0.010
1	2.71	3.84	5.02	6.63
2	4.61	5.99	7.38	9.21
3	6.25	7.81	9.35	11.34
4	7.78	9.49	11.14	13.28
5	9.24	11.07	12.83	15.09
6	10.64	12.59	14.45	16.81
7	12.02	14.07	16.01	18.48
8	13.36	15.51	17.53	20.09
9	14.68	16.92	19.02	21.67
10	15.99	18.31	20.48	23.21

Some Normal Percentiles

Right-Tail Probability			
0.100	0.050	0.025	0.010
1.282	1.645	1.960	2.326

1. We revisit the authorship example that we briefly discussed in lecture. Authors have distinctive literary styles that serve as a way to identify them from their writing. One approach to identification uses frequencies of usage of words with little contextual meaning. We will analyze a set of data for four authors, Jane Austen, Jack London, John Milton, and William Shakespeare, where word counts were made from blocks of text by these authors, each containing 1700 words. Baseline-category logistic regression models were fit with Shakespeare as the baseline category using frequencies of various words as predictors. Use the accompanying SAS output to help answer this problem.

   (a) A model using the frequencies of 11 words (`be, been, had, it, may, not, on, the, upon, was, which`) was fit to the data. Since the effects due to `may` and `had` appear to be not statistically significant, we also fit a nine-variable model omitting these two variables. Are we justified in omitting both the variables, `may` and `had`? Explain your answer.

   A baseline-category logistic regression model with Shakespeare as the baseline category was fit using only the frequency of the word "`the`" as a predictor.

   (b) Write out the estimated logit for comparing the probability that the author is Jane Austen relative to the probability that Jack London is the author. Then obtain the range of frequencies for the word "`the`" where the estimated probability that the author is Jane Austen is larger than the estimated probability that the author is Jack London.

(c) Estimate the odds ratio for the author being William Shakespeare rather than Jack London if the frequency of the word "`the`" increases by 10.

(d) Estimate the probability that the author is John Milton when the frequency of the word "`the`" equals 100.

2. Researchers are interested in the relationship between gender (`gender`= `male or female`) and depression (`depress`=`yes or no`). Since level of education is thought to be related to depression, the level of education in years of schooling (`educ`=`low` for less than 12 years or =`high` for 12 years or more) was recorded for each subject in the study. A random sample of subjects from a population known to be prone to depression was taken. Logistic regression models relating `depress` to `gender` and `educ` were fit to the data. Use the accompanying SAS output to help you answer this problem.

   (a) Carry out a test of equal odds ratios between `gender` and `depress` for the two levels of education.

   (b) Carry out a test of partial association of `gender` and `depress`, controlling for level of education.

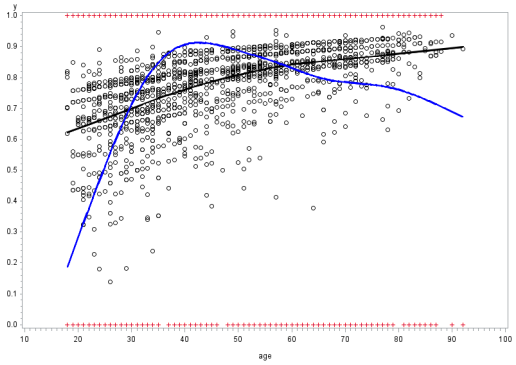
(c) Estimate the odds ratio between `gender` and `depress` for a highly educated person (`educ=high`) using: (i) the homogeneous association model and (ii) the model with interaction. Comment on which estimate is more appropriate for this data set.

3. The German Social Survey provides data on the number of children (`child`) for women in addition to their age (`age`) in years, nationality (`nation`= 1, if German, = 0, otherwise), duration of school education (`dur` ranging from 5 to 23 years), university degree (`univ`= 1 if yes, = 0 if no), and religion (`relig`= $1, \ldots, 6$ representing answers to the question "God is the most important in man." with 1 = strongly agree and 6 = strongly disagree). The response variable was defined to be `y` = 1 if `child` ≥ 1 and = 0 if `child` = 0.

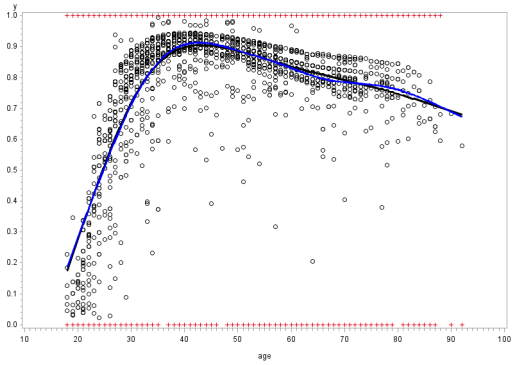
   Use the accompanying SAS output to help you answer the following questions.

   (a) A model (Model A) with the predictors `age,nation,dur,univ,relig` was fit to the data. The researchers felt that the effects of age and duration of education might be nonlinear, so they fit a second model (Model B) including the predictors in the first model as well as three additional polynomial terms in age (`age2,age3,age4`) and a quadratic term in duration (`dur2`). Marginal model plots for the variable `age` were made for each of the two models. For Model B the two plotted curves are nearly the same. Interpret these plots.

Model A

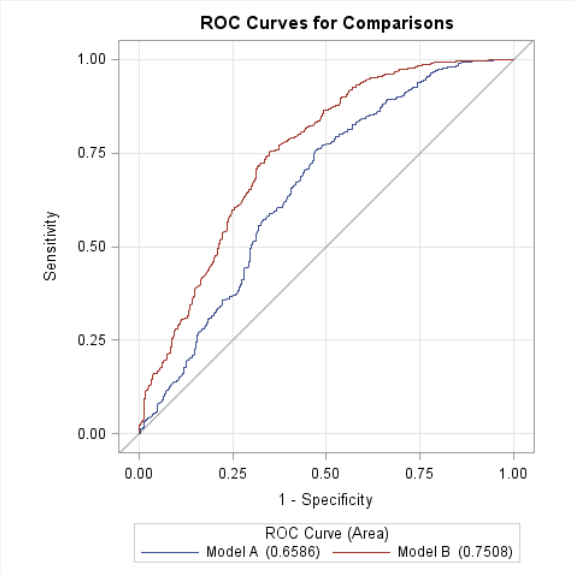


Model B



(b) Carry out a likelihood ratio test to determine whether the additional polynomial terms included in Model B described in part (a) significantly improve the fit relative to Model A (also described in part (a)).

(c) The ROC curves for Models A and B appear in the plot below. The lower curve corresponds to Model A, and the upper curve corresponds to Model B. Based on these ROC curves, which model has better predictive power? Explain why.



(d) Use the logistic regression model (Model A) without the additional polynomial terms in `age` and `dur` to answer this part of the problem. What is the effect of having a university degree on the estimated probability of a woman's having at least one child, keeping all other variables constant?

(e) Use the logistic regression model (Model B) with the additional polynomial terms in `age` and `dur` to answer this part of the problem. Is there any evidence of lack of fit for this model? Explain your answer.