Fall 2015 Exam

You may use SAS help at `sas.support.com` to help answer questions. You may also review your PROC IMPORT statements for tab-delimited text files with headers.

A merged data set of Fall 2008, Fall 2009 and Fall 2010 GPA data has been saved in tab-delimited text file `Fall15Q1.txt` on both Blackboard and the course webiste; we will use this data set for the data analysis questions on this exam. Variables and levels are listed below; note that some variables inherit their 2010 value, due to the nature of the merge.

Variable	Notes
Class	Sophomore, Junior, Senior, Graduate, Unknown, Law
GPA2008	2008 Fall GPA
Gender	M, F
Race	White, African American, Other, Unknown
SATV	Verbal SAT
SATM	Math SAT
RegStat	New, Continuing, Other
Enroll	Full-time, Part-time
Housing	1=On-campus, 0=Off-campus
ID	Unique identifier–not to be used in analysis
GPA2009	2009 Fall GPA
GPA2010	2010 Fall GPA

1. Consider a regression of Fall 2008 GPA on Math SAT.

   (a) Plot Fall 2008 GPA against Math SAT and note any possible violations of the usual regression assumptions for a linear model with normal errors.

   (b) Regress Fall 2008 GPA on Math SAT and inspect the residual plot (save the residuals using an OUTPUT statement). Comment on the residual plot and *briefly* discuss outlier diagnostics. Compute absolute residuals and plot the absolute residuals against Math SAT in PROC SGPLOT, overlaying a LOESS curve. Which seems more appropriate here–a WLS regression, or a robust regression using PROC ROBUSTREG?

   (c) Depending on your answer in 1(b), conduct an appropriate analysis of the data. Compare your results for the slope estimate to those in 1(a).

2. We will now consider a set of "kitchen sink" exercises (i.e., throw everything in the model except the kitchen sink–an old American expression). Ideally, these models would be preceded by extensive Exploratory Data Analysis, but we will have to set that aside for this timed exam.

   (a) Find a model to predict Fall 2010 GPA using stepwise selection in PROC GLMSELECT ($\alpha_e = 0.10,\ \alpha_r = 0.15$). Your first model should include all predictors–note any unusual results for slope estimates or effects in the output. For the second model, include all significant predictors found in the first model and their two-way interactions using the `hier=single` option. Using PROC SGPLOT, construct an ANCOVA-style plot of the response for one of the significant interactions between a categorical predictor and a continuous predictor and comment.

(b) Using PROC GAM, construct an additive model (no interactions!) including all predictors. Use `param` for categorical variables and `spline` for continuous variables; comment on the linear terms and spline terms; discuss the spline plot for GPA2008.

3. Consider a regression of Fall 2010 GPA on the continuous covariates Fall 2009 GPA, Fall 2008 GPA, Verbal SAT and Math SAT.

   (a) Compute VIFs for the above regression in PROC REG. Is there strong evidence of collinearity?

   (b) Construct a ridge plot using PROC REG. What seems to be a reasonable value of c to stabilize variance inflation and R^2? What are the slope estimates for this value of c? Compare them to slope estimates from the original regression and comment.

4. A simplifying formula for the deleted residual $d_i = Y_i - \hat{Y}_{i(i)}$ has the form

$$d_i = \frac{e_i}{1 - h_{ii}},$$

where $h_{ii} = x_i'(X'X)^{-1}x_i$. Using the result:

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{e_i}{1 - h_{ii}}(X'X)^{-1}x_i,$$

derive the simplifying formula above for d_i.