SAS Statistical Business Analysis Using SAS 9: Regression and Modeling Exam

ANOVA - 10%

Verify the assumptions of ANOVA

- Explain the central limit theorem and when it must be applied
- Examine the distribution of continuous variables (histogram, box-whisker, Q-Q plots)
- Describe the effect of skewness on the normal distribution
- Define H0, H1, Type I/II error, statistical power, p-value
- Describe the effect of sample size on p-value and power
- Interpret the results of hypothesis testing
- Interpret histograms and normal probability charts
- Draw conclusions about your data from histogram, box-whisker, and Q-Q plots
- Identify the kinds of problems may be present in the data: (biased sample, outliers, extreme values)
- For a given experiment, verify that the observations are independent
- For a given experiment, verify the errors are normally distributed
- Use the UNIVARIATE procedure to examine residuals
- For a given experiment, verify all groups have equal response variance
- Use the HOVTEST option of MEANS statement in PROC GLM to asses response variance

Analyze differences between population means using the GLM and TTEST procedures

- Use the GLM Procedure to perform ANOVA
  - CLASS statement
  - MODEL statement
  - MEANS statement
  - OUTPUT statement
- Evaluate the null hypothesis using the output of the GLM procedure
- Interpret the statistical output of the GLM procedure (variance derived from MSE, F value, p-value R**2, Levene's test)
- Interpret the graphical output of the GLM procedure
- Use the TTEST Procedure to compare means

Perform ANOVA post hoc test to evaluate treatment effect

- Use the LSMEANS statement in the GLM or PLM procedure to perform pairwise comparisons
- Use PDIFF option of LSMEANS statement
- Use ADJUST option of the LSMEANS statement (TUKEY and DUNNETT)
- Interpret diffograms to evaluate pairwise comparisons
- Interpret control plots to evaluate pairwise comparisons
- Compare/Contrast use of pairwise T-Tests, Tukey and Dunnett comparison methods

Detect and analyze interactions between factors

- Use the GLM procedure to produce reports that will help determine the significance of the interaction between factors. MODEL statement
- LSMEANS with SLICE=option (Also using PROC PLM)
- ODS SELECT
- Interpret the output of the GLM procedure to identify interaction between factors: p-value
- F Value
- R Squared
- TYPE I SS
- TYPE III SS

Linear Regression - 20%

Fit a multiple linear regression model using the REG and GLM procedures

- Use the REG procedure to fit a multiple linear regression model
- Use the GLM procedure to fit a multiple linear regression model

Analyze the output of the REG, PLM, and GLM procedures for multiple linear regression models

- Interpret REG or GLM procedure output for a multiple linear regression model: convert models to algebraic expressions
- Convert models to algebraic expressions
- Identify missing degrees of freedom
- Identify variance due to model/error, and total variance
- Calculate a missing F value
- Identify variable with largest impact to model
- For output from two models, identify which model is better
- Identify how much of the variation in the dependent variable is explained by the model
- Conclusions that can be drawn from REG, GLM, or PLM output: (about H0, model quality, graphics)

Use the REG or GLMSELECT procedure to perform model selection

- Use the SELECTION option of the model statement in the GLMSELECT procedure
- Compare the different model selection methods (STEPWISE, FORWARD, BACKWARD)
- Enable ODS graphics to display graphs from the REG or GLMSELECT procedure
- Identify best models by examining the graphical output (fit criterion from the REG or GLMSELECT procedure)
- Assign names to models in the REG procedure (multiple model statements)

Assess the validity of a given regression model through the use of diagnostic and residual analysis

- Explain the assumptions for linear regression
- From a set of residuals plots, asses which assumption about the error terms has been violated
- Use REG procedure MODEL statement options to identify influential observations (Student Residuals, Cook's D, DFFITS, DFBETAS)
- Explain options for handling influential observations
- Identify collinearity problems by examining REG procedure output
- Use MODEL statement options to diagnose collinearity problems (VIF, COLLIN, COLLINOINT)

Logistic Regression - 25%

Perform logistic regression with the LOGISTIC procedure

- Identify experiments that require analysis via logistic regression
- Identify logistic regression assumptions
- logistic regression concepts (log odds, logit transformation, sigmoidal relationship between p and X)
- Use the LOGISTIC procedure to fit a binary logistic regression model (MODEL and CLASS statements)

Optimize model performance through input selection

- Use the LOGISTIC procedure to fit a multiple logistic regression model
- LOGISTIC procedure SELECTION=SCORE option
- Perform Model Selection (STEPWISE, FORWARD, BACKWARD) within the LOGISTIC procedure

Interpret the output of the LOGISTIC procedure

- Interpret the output from the LOGISTIC procedure for binary logistic regression models: Model Convergence section
- Testing Global Null Hypothesis table
- Type 3 Analysis of Effects table
- Analysis of Maximum Likelihood Estimates table

- Association of Predicted Probabilities and Observed Responses

Score new data sets using the LOGISTIC and PLM procedures

- Use the SCORE statement in the PLM procedure to score new cases
- Use the CODE statement in PROC LOGISTIC to score new data
- Describe when you would use the SCORE statement vs the CODE statement in PROC LOGISTIC
- Use the INMODEL/OUTMODEL options in PROC LOGISTIC
- Explain how to score new data when you have developed a model from a biased sample

Prepare Inputs for Predictive Model Performance - 20%

Identify the potential challenges when preparing input data for a model

- Identify problems that missing values can cause in creating predictive models and scoring new data sets
- Identify limitations of Complete Case Analysis
- Explain problems caused by categorical variables with numerous levels
- Discuss the problem of redundant variables
- Discuss the problem of irrelevant and redundant variables
- Discuss the non-linearities and the problems they create in predictive models
- Discuss outliers and the problems they create in predictive models
- Describe quasi-complete separation
- Discuss the effect of interactions
- Determine when it is necessary to oversample data

Use the DATA step to manipulate data with loops, arrays, conditional statements and functions

- Use ARRAYs to create missing indicators
- Use ARRAYS, LOOP, IF, and explicit OUTPUT statements

Improve the predictive power of categorical inputs

- Reduce the number of levels of a categorical variable
- Explain thresholding
- Explain Greenacre's method
- Cluster the levels of a categorical variable via Greenacre's method using the CLUSTER procedure
    - METHOD=WARD option
    - FREQ, VAR, ID statement

    - Use of ODS output to create an output data set
- Convert categorical variables to continuous using smooth weight of evidence

Screen variables for irrelevance and non-linear association using the CORR procedure

- Explain how Hoeffding's D and Spearman statistics can be used to find irrelevant variables and non-linear associations
- Produce Spearman and Hoeffding's D statistic using the CORR procedure (VAR, WITH statement)
- Interpret a scatter plot of Hoeffding's D and Spearman statistic to identify irrelevant variables and non-linear associations

Screen variables for non-linearity using empirical logit plots

- Use the RANK procedure to bin continuous input variables (GROUPS=, OUT= option; VAR, RANK statements)
- Interpret RANK procedure output
- Use the MEANS procedure to calculate the sum and means for the target cases and total events (NWAY option; CLASS, VAR, OUTPUT statements)
- Create empirical logit plots with the SGPLOT procedure
- Interpret empirical logit plots

Measure Model Performance - 25%

Apply the principles of honest assessment to model performance measurement

- Explain techniques to honestly assess classifier performance
- Explain overfitting
- Explain differences between validation and test data
- Identify the impact of performing data preparation before data is split

Assess classifier performance using the confusion matrix

- Explain the confusion matrix
- Define: Accuracy, Error Rate, Sensitivity, Specificity, PV+, PV-
- Explain the effect of oversampling on the confusion matrix
- Adjust the confusion matrix for oversampling

Model selection and validation using training and validation data

- Divide data into training and validation data sets using the SURVEYSELECT procedure
- Discuss the subset selection methods available in PROC LOGISTIC
- Discuss methods to determine interactions (forward selection, with bar and @ notation)

- Create interaction plot with the results from PROC LOGISTIC
- Select the model with fit statistics (BIC, AIC, KS, Brier score)

Create and interpret graphs (ROC, lift, and gains charts) for model comparison and selection

- Explain and interpret charts (ROC, Lift, Gains)
- Create a ROC curve (OUTROC option of the SCORE statement in the LOGISTIC procedure)
- Use the ROC and ROCCONTRAST statements to create an overlay plot of ROC curves for two or more models
- Explain the concept of depth as it relates to the gains chart

Establish effective decision cut-off values for scoring

- Illustrate a decision rule that maximizes the expected profit
- Explain the profit matrix and how to use it to estimate the profit per scored customer
- Calculate decision cutoffs using Bayes rule, given a profit matrix
- Determine optimum cutoff values from profit plots
- Given a profit matrix, and model results, determine the model with the highest average profit

---

Note: All 22 main objectives will be tested on every exam. The 126 expanded objectives are provided for additional explanation and define the entire domain that could be tested.