

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

## 分散分析 - 10%

### 分散分析の仮定の検証

- 中心極限定理と、いつ中心極限定理を適用する必要があるかについて
- 連続変数の分布の調査（ヒストグラム、箱ひげ図、Q-Qプロット）
- 正規分布に歪度が与える影響の説明
- $H_0$ 、 $H_1$ 、第1種 / 第2種の過誤、統計的な検出力、p値の定義
- サンプルサイズがp値と検出力に及ぼす影響の説明
- 仮説検定の結果の解釈
- ヒストグラムと正規確率プロットの解釈
- ヒストグラム、箱ひげ図、Q-Qプロットからデータに関する結論を導き出す
- データに存在する問題の種類の特定：（偏った標本、外れ値、極値）
- 所与の実験における、オブザベーションの独立性の検証
- 所与の実験において、誤差が正規分布に従っているかの検証
- UNIVARIATEプロシジャを使用した残差の調査
- 所与の実験における、全ての群についての応答変数の等分散性の検証
- PROC GLMのMEANSステートメントのHOVTESTオプションを使用した、応答変数の分散の評価

### GLMおよびTTESTプロシジャを使用した母平均の差の分析

- GLMプロシジャを使用した分散分析の実施
  - CLASSステートメント
  - MODELステートメント
  - MEANSステートメント
  - OUTPUTステートメント
- GLMプロシジャの出力を使用した帰無仮説の評価
- GLMプロシジャの統計的な出力の解釈（MSE、F値、p値、R2乗、Levene検定から導かれる分散）
- GLMプロシジャのグラフィカルな出力の解釈
- TTESTプロシジャを使用した平均の比較

### 分散分析におけるポストホック検定の実施による処理効果の評価

- GLMまたはPLMプロシジャのLSMEANSステートメントを使用したペアごとの比較の実施
- LSMEANSステートメントのPDIFオプションの使用
- LSMEANSステートメントのADJUSTオプションの使用（TUKEYとDUNNETT）
- ペアごとの比較を評価するための、ディフォグラムの解釈
- ペアごとの比較を評価するための、コントロール・プロットの解釈
- 対ごとのt検定、TukeyとDunnett比較法の使用の対比較

## 因子間の交互作用の検出と分析

- 因子間の交互作用の重要性を判断するのに役立つレポートの作成のための、GLMプロシジャの使用MODELステートメント
- SLICE =オプションを伴うLSMEANS（PROC PLMも使用）
- ODS SELECT
- 因子間の交互作用を識別するための、GLMプロシジャ出力の解釈：p値
- F値
- R<sup>2</sup>乗
- タイプIの平方和
- タイプIIIの平方和

## 線形回帰 - 20%

### REGおよびGLMプロシジャを使用した線形重回帰モデルの当てはめ

- 線形重回帰モデルの当てはめのための、REGプロシジャの使用
- 線形重回帰モデルの当てはめのための、GLMプロシジャの使用

### 線形重回帰モデルのREG、PLM、およびGLMプロシジャ出力の分析

- 線形重回帰モデルのREGまたはGLMプロシジャ出力の解釈：モデルの代数式への変換
- モデルの代数式への変換
- 欠損の自由度の識別
- モデル/誤差、および全分散による分散の識別
- 欠損したF値の計算
- モデルへの影響が最も大きい変数の特定
- 2モデルからの出力において、どちらのモデルが良いかの識別
- 従属変数の変動がモデルによってどの程度説明されているかの特定
- REG、GLM、またはPLMの出力から導くことができる結論：（H0、モデルの品質、グラフィックスについて）

### REGプロシジャやGLMSELECTプロシジャを使用したモデル選択の実施

- GLMSELECTプロシジャにおけるmodelステートメントのSELECTIONオプションの使用
- さまざまなモデルの選択方法の比較（STEPWISE、FORWARD、BACKWARD）
- REGまたはGLMSELECTプロシジャのグラフを表示するための、ODSグラフィックスの有効化
- グラフィカルな出力を調べることによる最良のモデルの識別（REGまたはGLMSELECTプロシジャの当てはまりの基準）
- REGプロシジャでのモデルへの名前の割り当て（複数のモデルステートメント）

### 診断分析と残差分析の使用による回帰モデルの妥当性の評価

- 線形回帰の仮定について
- 一連の残差プロットから、どの誤差項の仮定が違反しているかの評価
- 影響力のあるオブザベーションの特定のための、REGプロシジャのMODELステートメントオプションの使用（RStudent 残差、Cookの D、DFFITS、DFBETAS）
- 影響力のあるオブザベーションをハンドリングするためのオプションについて
- REGプロシジャ出力の調査による共線性の問題の識別
- 共線性の問題を診断するための、MODELステートメントオプションの使用（VIF、COLLIN、COLLINOINT）

## ロジスティック回帰 - 25%

### LOGISTICプロシジャによるロジスティック回帰の実施

- ロジスティック回帰による分析が必要な実験の識別
- ロジスティック回帰の仮定の識別
- ロジスティック回帰の概念（対数オッズ、ロジット変換、pとX間のシグモイドの関係）
- バイナリロジスティック回帰モデルの当てはめのための、LOGISTICプロシジャの使用（MODELおよびCLASSステートメント）

### 入力選択によるモデル性能の最適化

- 複数ロジスティック回帰モデルの当てはめのための、LOGISTICプロシジャの使用
- LOGISTICプロシジャのSELECTION=SCOREオプション
- LOGISTICプロシジャ内でモデル選択の実行（STEPWISE、FORWARD、BACKWARD）

### LOGISTICプロシジャの出力の解釈

- バイナリロジスティック回帰モデルについてのLOGISTICプロシジャの出力の解釈モデル収束セクション
- 包括的帰無仮説の検定表
- 効果に対するType 3分析
- 最尤推定値の分析表

- 予測確率と観測データの応答との関連性

## LOGISTICプロシジャやPLMプロシジャを使用した新規データセットのスコアリング

- 新しいケースをスコアリングするための、PLMプロシジャのSCOREステートメントの使用
- 新しいデータをスコアリングするための、PROC LOGISTICのCODEステートメントの使用
- PROC LOGISTICでのSCOREステートメントとCODEステートメントの使用について
- PROC LOGISTICのINMODEL / OUTMODELオプションの使用
- 偏った標本からモデルを作成したときの、新しいデータをスコアリングする方法の説明

## 予測モデルの性能に向けた入力の準備 - 20%

### 入力データにおける潜在的な問題の同定

- 予測モデルを作成して新しいデータセットをスコアリングする際に欠損値が原因で発生する可能性がある問題の識別
- 完全データとしての解析の限界の識別
- 多数のレベルを持つカテゴリ変数によって引き起こされる問題の説明
- 冗長な変数の問題について
- 無関係で冗長な変数の問題について
- 非線形性と予測モデルで非線形性が引き起こす問題について
- 外れ値と予測モデルで外れ値が引き起こす問題について話し合う
- 準完全分離について
- 交互作用の影響について
- データをオーバーサンプリングする必要がある場面

### DATAステップの利用によるループや配列、条件文、関数を用いたデータ・ハンドリング

- 欠損値の指標を作成するための、配列の使用
- 配列、ループ、IF、および明示的なOUTPUTステートメントの使用

### カテゴリ変数入力における予測力の向上

- カテゴリ変数の水準数の削減
- 閾値について
- Greenacreの方法について
- CLUSTERプロシジャでGreenacreの方法を使用したカテゴリ変数の水準のクラスタ化
  - METHOD = WARDオプション
  - FREQ、VAR、IDステートメント

- 出力データセットを作成するためのODS出力の使用
- 平滑化された証拠の重みを使用してカテゴリ変数を連続値に変換する

## CORRプロシジャを使用した不要な変数のスクリーニング

- HoeffdingのD統計量とSpearmanの相関統計量を使用して、無関係な変数や非線形関係を見つける方法の説明
- CORRプロシジャ（VAR、WITHステートメント）を使用したSpearmanとHoeffdingのD統計量の生成
- 無関係な変数や非線形関係を識別するための、HoeffdingのD統計量とSpearmanの相関統計量の散布図の解釈

## 経験ロジットのプロットを使用した非線形性を持つ変数のスクリーニング

- 連続入力変数をビン化するための、RANKプロシジャの使用（GROUPS =、OUT = オプション、VAR、RANKステートメント）
- RANKプロシジャの出力の解釈
- ターゲット・ケースおよびイベントの総数の合計と平均を計算するためのMEANSプロシジャの使用（NWAYオプション、CLASS、VAR、OUTPUTステートメント）
- SGPLOTプロシジャでの経験ロジットのプロットの作成
- 経験ロジットのプロットの解釈

## モデル性能の計測 - 25%

### モデル性能の計測 「正直な評価」 原理を適用したモデル性能の計測

- 分類器の性能を正當に評価するためのテクニックについて
- オーバーフィッティングについて
- 検証データとテストデータの違いについて
- データを分割する前にデータの準備を実施することの影響の識別

### 混同行列を使用した分類性能の評価

- 混同行列について
- 定義：正確度、誤分類率、感度、特異度、PV +、PV -
- 混同行列に対するオーバーサンプリングの影響について
- オーバーサンプリングのための混同行列の調整

### 学習および評価データを使用したモデル選択と評価

- SURVEYSELECTプロシジャを使用した、学習および検証データセットへのデータ分割
- PROC LOGISTICでの利用可能なサブセットの選択方法について
- 交互作用を定義するについて（前進選択、バー表記と@表記）

- PROC LOGISTICの結果と交互作用プロットの作成
- 当てはめ統計量（BIC、AIC、KS、Brier Score）でのモデルの選択

## モデルの比較や選択を行うためのグラフ（ROC、リフト、およびゲインチャート）の作成と解釈

- チャート（ROC、リフト、ゲイン）の説明と解釈
- ROC曲線の作成（LOGISTICプロシジャでのSCOREステートメントのOUTROCオプション）
- 複数モデルのROC曲線の重ね合わせプロットを作成するための、ROCおよびROCCONTRASTステートメントの使用
- ゲインチャートに関連する深さの概念について

## スコアリングを行う上での効果的なカットオフ値の確立

- 期待利益を最大化する決定ルールについて
- 利益行列と、それを使ってスコアリングされた顧客1人あたりの利益を推定する方法について
- 所与の利益行列における、Bayesルールを使ったカットオフ値の計算
- 利益プロットからの最適なカットオフ値の決定
- 所与の利益行列とモデル結果における、最も高い平均利益を持つモデルの決定

---

**注意：** 全22の主な目的は、すべての試験でテストされます。126の拡張された目的は、追加の説明のために提供され、テストされる可能性のある領域全体を定義します。