重回帰

概要

アシスタントの重回帰手順は、最小二乗推定を使用して、最大 5 つまでの予測変数（X）と 1 つの連続応答変数を含む線形および 2 次モデルを適合します。ユーザーがモデルタイプを選択し、アシスタントがモデル項を選択します。本書では、アシスタントが回帰モデルの選択に使用する基準について説明します。

さらに、有効な回帰モデルを取得するために重要ないくつかの因子を調査します。まず、検定に十分な検出力、および X と Y の関係の強度を推定するために十分な精度を提供できる大きさのサンプルが必要です。次に、分析の結果に影響を及ぼす可能性がある異常なデータを特定することが重要です。また、誤差項は正規分布に従っているという仮定を考慮し、全体のモデルの仮説検定に対する非正規性の影響も評価します。

アシスタントはこれらの因子に基づき、データで次のチェックを行い、レポートカードに結果を表示します。

- データ量
- 異常なデータ
- 正規性

本書では、これらの因子が実際にどのように回帰分析に関連するかを調査し、アシスタントでこれらの因子をチェックするためのガイドラインをどのように定めたかについて説明します。

回帰法

モデル選択

アシスタントの回帰分析は、1 つの連続応答変数と 2～5 つの予測変数を含むモデルを適合します。予測変数の 1 つはカテゴリ変数の場合があります。次の 2 つのタイプのモデルがあります。

- 線形：$F(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
- 2 次：$F(x) = \beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i<j} \beta_{ij} X_i X_j$

アシスタントは、完全な線形モデルまたは 2 次モデルからモデル項を選択します。

目的

アシスタントでどちらのモデルを使用するかを決定するため、モデル選択に使用できる異なる方法を調査します。

方法

後方、前方、ステップワイズの 3 つの異なるモデル選択タイプを調査しました。これらのモデル選択タイプには次のようないくつかのオプションが含まれており、それらのオプションについても調査しました。

- モデルへの項の入力または削除を行うために使用する基準。
- モデルに特定の項を強制的に含めるかどうか、または最初のモデルに特定の項を含めるかどうか。
- モデルの階層。
- モデルの X 変数の標準化。

これらのオプションを確認し、手順の結果でその効果を調べ、どの方法が専門家に好まれるかを検討しました。

結果

アシスタントでのモデル項の選択に使用した手順は次のとおりです。

- ステップワイズモデル選択を使用します。多くの場合、潜在的な X 変数のセットは相関しており、1 つの項の効果はモデル内に含まれる他の項によって異なります。ステップワイズ選択は、1 つのステップで項を入力し、モデル内に含まれる他の項に応じて後からその項を削除できるため、この状況に最適な手法と言えます。
- モデルの階層は各ステップで維持され、同一ステップ内で複数の項をモデルに入力できます。たとえば、最も有意な項がX_1^2の場合、この項はX_1が有意であるかどうかに関わらずX_1とともに入力されます。階層は標準化された単位から非標準化単位にモデルを変換できるため、好ましいと考えられます。また、任意のステップで複数の項をモデルに入力できるため、関連する線形項が応答に強く関係していない場合でも、重要な二乗項または交互作用項を特定できます。

- 項は$\alpha = 0.10$に基づいてモデルに入力されたりモデルから削除されたりします。$\alpha = 0.10$を使用することで、$\alpha = 0.15$を使用するコア Minitab のステップワイズ手順よりも選択的になります。
- モデル項を選択するために、平均を引いて標準偏差で割ることで予測変数が標準化されます。最終モデルは、非標準化 X の単位で表示されます。X の標準化によって、線形項と二乗項間のほとんどの相関が削除され、高次の項が不必要に追加される可能性が低くなります。

データチェック

データ量

検出力は、仮説検定が偽の場合にどの程度の確率で帰無仮説を棄却するかによります。回帰の場合、帰無仮説は X と Y の間に関係はないことを示します。データセットが小さすぎる場合、検定の検出力は実際に存在する X と Y の関係を適切に検出できない可能性があります。そのため、高確率で実質的に重要な関係を検出するのに十分な大きさのデータセットが必要です。

目的

X と Y の関係に対する全体の F 検定の検出力、および X と Y の関係の強度の推定値である R 二乗（調整済み）の精度に対して、データ量がどのように影響するかを判断します。この情報は、データで観測される関係の強度が、真の関係の基本強度の確実な指標であると信頼するために十分な大きさのデータセットがあるかどうかを判断するために不可欠です。R 二乗（調整済み）の詳細は、「付録 A」を参照してください。

方法

単回帰で使用した推奨サンプルサイズの判断方法と同じ方法を使用しました。R 二乗（調整済み）値の変動性を調べ、R 二乗（調整済み）が ρ 二乗（調整済み）に近くなるために必要なサンプルの大きさを判断しました。また、Y 変数と X 変数の関係の強度がやや弱い場合でも、推奨サンプルサイズで妥当な検出力を得られることも確認しました。この計算の詳細は、「付録 B」を参照してください。

結果

単回帰と同様、R 二乗（調整済み）の観測値が ρ 二乗（調整済み）の 0.20 内になることを 90%信頼できるのに十分な大きさのサンプルサイズが推奨されます。モデルに項を追加するたびに必要なサンプルサイズが増えることが判明したため、各モデルサイズで必要なサンプルサイズを計算しました。推奨サイズは 5 の倍数に最も近い値に切り上げられます。たとえば、定数に加えて 8 つの係数（4 つの線形項、3 つの交互作用項、および 1 つの二乗項など）がモデルに含まれている場合、基準を満たすために必要な最小サンプルサイズは n = 49 です。アシスタントはこれを推奨サンプルサイズの n = 50 に切り上げます。項の数に基づく推奨サンプルサイズの詳細は、「付録 B」を参照してください。

また、推奨サンプルサイズで十分な検出力を得られることも確認しました。ρ 二乗（調整済み）= 0.25 のやや弱い関係の場合、検出力は通常約 80%以上であることがわかりました。したがって、アシスタントの推奨サンプルサイズに従うことで、関係の強度の推定で妥当な検出力と精度を得ることができます。

これらの結果に基づき、データ量をチェックするときに、アシスタントレポートカードに次の情報が表示されます。

ステータス	状態
	**サンプルサイズ<推奨サイズ** サンプルサイズは、関係の強度を非常に正確に推定するために十分な大きさではありません。R 二乗や R 二乗（調整済み）などの関係の強度の測定値は大きく異なる場合があります。正確に推定するには、より大きなサンプルをこのサイズのモデルに使用する必要があります。 **サンプルサイズ>=推奨サイズ** サンプルは、関係の強度を非常に正確に推定するのに十分な大きさです。

異常なデータ

アシスタントの回帰手順では、異常なデータを標準化残差またはてこ比値の大きい観測値と定義します。通常、これらの測定値は回帰分析の異常なデータを特定するために使用されます（Neter et al、1996）。異常なデータは結果に大きく影響する可能性があるため、分析を有効にするにはデータの修正が必要な場合があります。ただし、異常なデータは工程の自然変動により発生する可能性もあります。そのため、異常な振る舞いの原因を特定し、そのようなデータ点の処理方法を決定することが重要です。

目的

データ点が異常であるという信号を出すために必要な標準化残差およびてこ比値の大きさを判断します。

方法

Minitab 標準の回帰手順（[統計] > [回帰] > [回帰]）に基づき、異常な観測値を特定するためのガイドラインを作成しました。

結果

標準化残差

標準化残差は、残差（e_i）をその標準偏差の推定値で割ったものです。一般に、標準化残差が 2 より大きい場合に観測値が異常であると見なされますが、このガイドラインはやや保守的です。すべての観測値のうち約 5%が偶然の所産としてこの基準を満たすことが予測されます（誤差が正規分布に従う場合）。そのため、異常な振る舞いの原因を調査し、観測値が実際に異常であるかどうかを判断することが重要です。

てこ比値

てこ比値は観測の X 値にのみ関連し、Y 値には依存しません。てこ比値がモデル係数（p）を観測値数（n）で割った数の 3 倍よりも大きい場合、観測値が異常と見なされます。一部の教科書では$\frac{2 \times p}{n}$を使用していますが、これが一般的に使用される基準値です（Neter et al、1996）。

データに高いてこ比値点が含まれる場合、それらの点がデータに適合するために選択したモデルに不適切に影響するかどうかを調査します。たとえば、1 つの極端な X 値によって、線形モデルの代わりに 2 次モデルが選択される可能性があります。2 次モデルで観測された曲面性が、工程に関する理解と一致するかどうかを確認する必要があります。一致しない場合、より単純なモデルをデータに適合するか、追加データを収集してさらに細かく工程を調査します。

異常なデータをチェックするときに、アシスタントレポートカードに次のステータスインジケータが表示されます。

ステータス	状態
	異常なデータ点はありません。
	1 つ以上の大きな標準化残差または 1 つ以上の高いてこ比点があります。

正規性

一般に、回帰はランダム誤差（ε）が正規分布に従っているという仮定に基づきます。係数（β）の推定の仮説検定を実施する場合、正規性の仮定が重要になります。ただし、ランダム誤差が正規分布に従っていない場合でも、サンプルが十分に大きければ検定結果は多くの場合信頼できます。

目的

正規分布に基づいて信頼できる結果を得るために必要なサンプルのサイズを判断します。実際の検定結果がその検定の目標有意水準（α、またはタイプ I 過誤率）とどの程度近く一致するか、つまり、異なる非正規分布で期待されるより高い頻度または低い頻度で検定が帰無仮説を誤って棄却するかどうかを判断します。

方法

タイプ I 過誤率を推定するため、正規分布から大きく逸脱する歪んだ分布、裾の重い分布、および裾の軽い分布で複数のシミュレーションを実行しました。これらのシミュレーションは 15 のサンプルサイズを使用して行い、いくつかのモデルで全体の F 検定を調べました。

各条件に対して 10,000 回の検定を行い、各検定で帰無仮説が真になるように、ランダムデータを生成しました。次に、目標有意水準に 0.10 を使用した検定を実行し、10,000 回のうち実際に帰無仮説を棄却した回数を数え、この比率を目標有意水準と比較しました。検定が適切に機能する場合、タイプ I 過誤率は目標有意水準に非常に近くなります。このシミュレーションの詳細は、「付録 C」を参照してください。

結果

両方の全体の F 検定で、統計的に有意な結果を得る確率はどの非正規分布でも大きな差はありません。タイプ I 過誤率はすべて 0.08820～0.11850 で、目標有意水準の 0.10 に適度に近くなっています。

検定は比較的小さいサンプルで適切に機能するため、アシスタントはデータの正規性を検定しません。代わりに、サンプルのサイズをチェックし、サンプルが 15 未満の場合に通知します。アシスタントの異常なデータをチェックするときに、アシスタントの回帰レポートカードに次のステータスインジケータが表示されます。

ステータス	状態
	サンプルサイズが 15 以上であるため、正規性は問題になりません。
	サンプルサイズが 15 未満であるため、正規性が問題になる可能性があります。p 値を解釈するときに注意が必要です。小さなサンプルを使用する場合、p 値の正確性は非正規残差誤差の影響を受けやすくなります。

参考文献

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (1996). *Applied linear statistical models*. Chicago: Irwin.

付録 A: モデルと統計量

応答 Y に対する予測変数 X に関する回帰モデルは次の形式です。

$Y = f(X) + \varepsilon$

ここで、関数 f(X)は与えられた X に対する Y の期待値（平均）を表します。

アシスタントでは、関数 f(X)の形式として次の 2 つの選択肢があります。

モデルタイプ	f(X)
線形	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
2 次	$\beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i<j} \beta_{ij} X_i X_j$

係数βの値は未知で、データから推定する必要があります。推定方法は、サンプルの残差の平方和を最小化する最小二乗です。

$$\min \sum_{i=1}^{n} \left(Y_i - \hat{f}(X_i)\right)^2.$$

残差は、推定された係数に基づく観測応答Y_iと適合値$\hat{f}(X_i)$間の差です。この平方和の最小化された値は、指定されたモデルの SSE（誤差の平方和）です。

全体の F 検定

この方法は、全体のモデル（線形または 2 次）の検定です。選択された形式の回帰関数 f(X)で次を検定します。

$H_0: f(X)$ は定数です

$H_1: f(X)$ は定数ではありません

調整済みR^2

調整済みR^2（R^2_{adj}）は、モデルによって応答の変動性がどの程度 X に起因するかを測定します。観測された X と Y の関係の強度を測定する一般的な方法は、次の 2 つです。

$$R^2 = 1 - \frac{SSE}{SSTO}$$

および

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

ここで

$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

SSTO は平方総和で、全体の平均$\bar{Y}$での応答の変動を測定します。SSE は、回帰関数 f(X)での変動を測定します。R 二乗（調整済み）の調整は完全モデルでの係数の数（p）用で、n - p の自由度によってεの分散を推定します。より多くの係数がモデルに追加された場合、R^2が減少することはありません。ただし、R 二乗（調整済み）は調整されるため、追加係数によってモデルが改善されない場合は減少する可能性があります。そのため、モデルへの別の項の追加によって応答の追加分散は説明されず、R 二乗（調整済み）が減少し、追加項は役に立たないことが示されます。したがって、異なるサイズのモデルの比較には調整された測定を使用する必要があります。

F 検定と R 二乗（調整済み）の関係

全体のモデル検定の F 統計量は、R 二乗（調整済み）の計算でも使用される SSE と SSTO で表すことができます。

$$F = \frac{(SSTO - SSE)/(p-1)}{SSE/(n-p)}$$

$$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R^2_{adj}}{1-R^2_{adj}}$$.上記の計算式は、F 統計量が R 二乗（調整済み）の増加関数であることを示しています。そのため、R 二乗（調整済み）が検定の有意水準（α）で決定された特定の値を超えた場合にのみ、検定で H0 が棄却されます。

付録 B: データ量

このセクションでは、全体のモデル検定の検出力、およびモデルの強度の推定値である R 二乗（調整済み）の精度に観測値数の n がどのような影響を及ぼすかを調査します。

関係の強度を数量化するため、新しい ρ 二乗（調整済み）を前述のサンプル統計量 R 二乗（調整済み）の母集団に対応する数量として導入します。

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

これに対して、次のように定義します。

$$\rho^2_{adj} = 1 - \frac{E(SSE|X)/(n-p)}{E(SSTO|X)/(n-1)}$$

演算子 E(·|X) は、期待値、または X の値が与えられた場合の確率変数の平均を表します。正しいモデルが独立同一分布に従う ε を含む$Y = f(X) + \varepsilon$と仮定すると、次のようになります。

$$\frac{E(SSE|X)}{n-p} = \sigma^2 = Var(\varepsilon)$$

$$\frac{E(SSTO|X)}{n-1} = \sum_{i=1}^{n} \frac{(f(X_i) - \bar{f})^2}{(n-1)} + \sigma^2$$

ここで、$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.

したがって、

$$\rho^2_{adj} = \frac{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1)}{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1) + \sigma^2}$$

全体のモデルの有意性

全体のモデルの統計的有意性を検定する場合、ランダム誤差 ε が独立正規分布に従っていると仮定します。Y の平均が定数$(f(X) = \beta_0)$であるという帰無仮説では、F 検定統計量に$F(p-1, n-p)$ 分布があります。対立仮説では、F 統計量には非心パラメータを使用した非心$F(p-1, n-p, \theta)$分布があります。

$$\theta = \sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/\sigma^2$$

$$= \frac{(n-1)\rho^2_{adj}}{1 - \rho^2_{adj}}$$

H_0 を棄却する確率は、n と ρ 二乗（調整済み）の両方で増加している非心パラメータによって高くなります。

関係の強度

単回帰で示したように、データの統計的に有意な関係は X と Y の強い基本関係を示すとは限りません。そのため、多くのユーザーは関係の実際の強度を R 二乗（調整済み）などの指標で確認します。R 二乗（調整済み）を ρ 二乗（調整済み）の推定値とする場合、その推定値が真の ρ 二乗（調整済み）値に適度に近いことを信頼できる必要があります。

考えられるモデルサイズごとに、0.20 より大きい差分絶対値| R 二乗（調整済み）- ρ 二乗（調整済み）|が 10%以下の確率で発生する n の最小値を特定することで、適切な許容サンプルサイズのしきい値を判断します。これは ρ 二乗（調整済み）の真の値には依存しません。次の表に、推奨サンプルサイズ n(T)を要約します。ここで、T は定数係数以外のモデル内の係数の数です。

T	n(T)
1～3	40
4～6	45
7～8	50
9～11	55
12～14	60
15～18	65
19～21	70
22～24	75
25～27	80
28～31	85
32～34	90
35～38	95
39～41	100
42～45	105
46～48	110
49～52	115
53～56	120
57～59	125
60～63	130
64～67	135

T	n(T)
68～70	140
71～73	145

$\rho^2_{adj} = 0.25$のやや弱い値で、モデル全体の F 検定の検出力を評価し、推奨サンプルサイズで十分な検出力を得られることを確認しました。次の表のモデルサイズは、n(T)の各値での最悪の場合を表しています。同じ n(T)でモデルサイズを小さくすると、検出力は高くなります。

T	n(T)	検出力 $\rho^2_{adj} = 0.25$
3	40	0.902791
6	45	0.854611
8	50	0.850675
11	55	0.831818
14	60	0.820592
18	65	0.798003
21	70	0.796425
24	75	0.796911
27	80	0.798856
31	85	0.789861
34	90	0.794367
38	95	0.788625
41	100	0.794511
45	105	0.790864
48	110	0.797487
52	115	0.79525
56	120	0.793698
59	125	0.800982
63	130	0.800230
67	135	0.799906

T	n(T)	検出力 $\rho_{adj}^2 = 0.25$
69	140	0.814664

付録 C: 正規性

アシスタントで使用される回帰モデルは、すべて次の形式です。

$Y = f(X) + \varepsilon$

ランダム項εは、平均ゼロおよび一般分散σ^2による独立同一分布に従う正規確率変数であると一般的に仮定されます。βパラメータの最小二乗推定は、εが正規分布に従うと仮定しない場合でも、最良の線形不偏推定です。正規性の仮定は、f(X)の仮説検定と同様にこれらの推定に確率を追加する場合にのみ重要になります。

正規性の仮定に基づく回帰分析の結果を信頼できるのに十分なnの大きさを判断するため、さまざまな非正規誤差分布で仮説検定のタイプⅠ過誤率を調べるシミュレーションを実行しました。

次の表1は、3つの異なるモデルのさまざまなεの分布において、全体のF検定が$\alpha = 0.10$で有意であった10,000回のシミュレーションの比率を示しています。これらのシミュレーションでは、XとYの間に関係はないという帰無仮説は真でした。MinitabのRANDOMコマンドで、X値が多変量正規変数として生成されました。すべての検定で使用したサンプルサイズはn = 15です。すべてのモデルには5つの連続予測変数が含まれます。最初のモデルは、5つすべてのX変数が含まれる線形モデルです。2番目のモデルは、すべての線形項と二乗項が含まれます。3番目のモデルには、すべての線形項と7つの双方向の交互作用が含まれます。

表1 非正規分布でのn = 15を使用した全体のF検定のタイプⅠ過誤率

分布	線形	線形+ 二乗	線形+ 7つの交互作用
正規	0.09910	0.10270	0.10060
t(3)	0.09840	0.1185	0.118
t(5)	0.09980	0.10010	0.10430
ラプラス	0.09260	0.09400	0.09650
一様	0.10630	0.10080	0.09480
Beta(3, 3)	0.09980	0.10120	0.10020
指数	0.08820	0.09500	0.09960
Chi(3)	0.09890	0.11400	0.10970
Chi(5)	0.09730	0.10590	0.10330
Chi(10)	0.10150	0.09930	0.10360
Beta(8, 1)	0.09870	0.10230	0.10490

シミュレーション結果では、統計的に有意な結果を得る確率はすべての誤差分布で 0.10 の望目特性と大きな差はないことが示されています。観測されたタイプ I 過誤率はすべて 0.08820～0.11850 です。

© 2015, 2017 Minitab Inc. All rights reserved.
Minitab®, Quality. Analysis. Results.® and the Minitab® logo are all registered trademarks of Minitab, Inc., in the United States and other countries. See minitab.com/legal/trademarks for more information.