機械学習シリーズ：線形回帰

原文はこちら

このブログ記事は、Vertica の機械学習アルゴリズムに関する一連のブログ記事の 1 つです。今後のアップデートをご期待ください。

線形回帰とは？

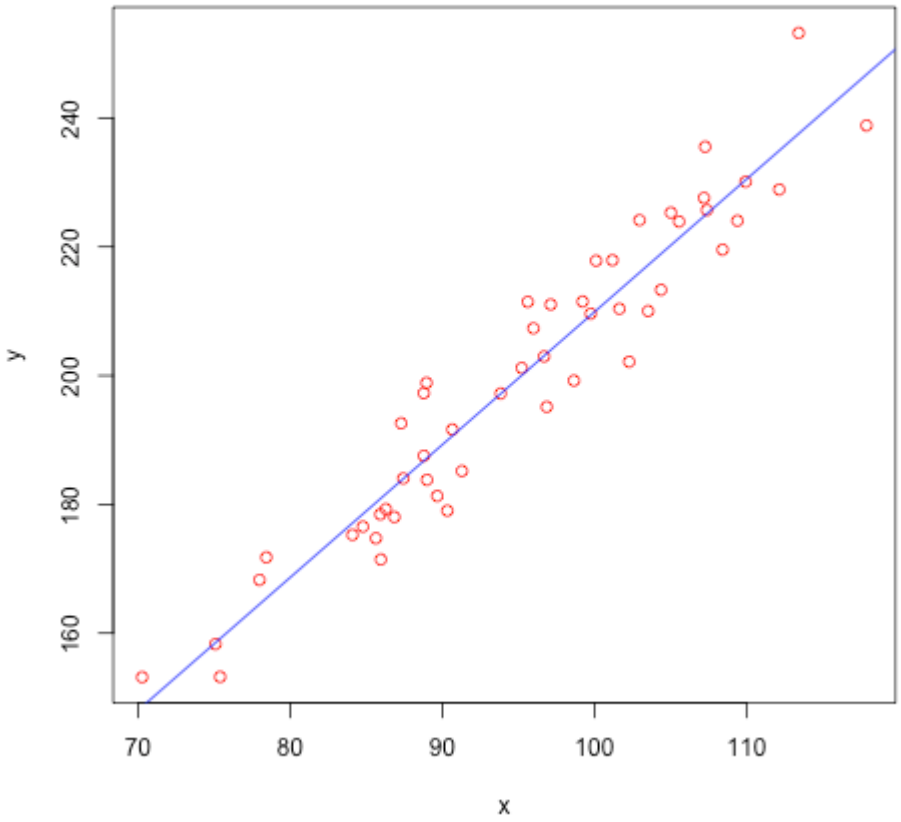
基礎から始めましょう。線形回帰は、最も古くから最も広く使われている予測モデルの 1 つです。線形回帰は、問題の 1 つ以上の予測因子（独立変数）と応答（従属変数）の間の関係を推定する統計モデルです。一般に、独立変数は連続的、離散的、またはカテゴリー的でありえます。Vertica はカテゴリー予測変数をサポートしていないことに注意してください。従属変数は、回帰アプリケーションでは常に連続ですが、分類アプリケーションでは連続していません。予測因子とデータポイントを通る直線との相関を表すことができるはずです。
たとえば、ボストンで家を購入しようとしているとします。あなたはすでに市の住宅のコストに関するデータを面積に基づいて持っています。このデータを使用して、家を購入するのに最も経済的な市内の場所はどこかを決定したいと考えています。線形回帰を使用してこの問題を評価し、購入する場所を予測することができます。このシナリオでは、面積という予測因子に基づいて、価格という応答を得ます。データポイント間の関係を記述する関数は次のとおりです。

$$y_i = \alpha + \beta x_i + \epsilon_i \qquad (1)$$

このシナリオでは、データポイントに最も適した直線方程式の係数を求めたいとします。

$$y = \alpha + \beta x \qquad (2)$$

線形回帰の方程式のために見出される最適な係数のセットは、モデルと呼ばれます。この例では、線形回帰方程式には価格と面積という 2 つの変数しかありません。モデルとデータポイントを次のように視覚化することができます。ここでは、面積が x 軸に、価格は y 軸上にあるとします。

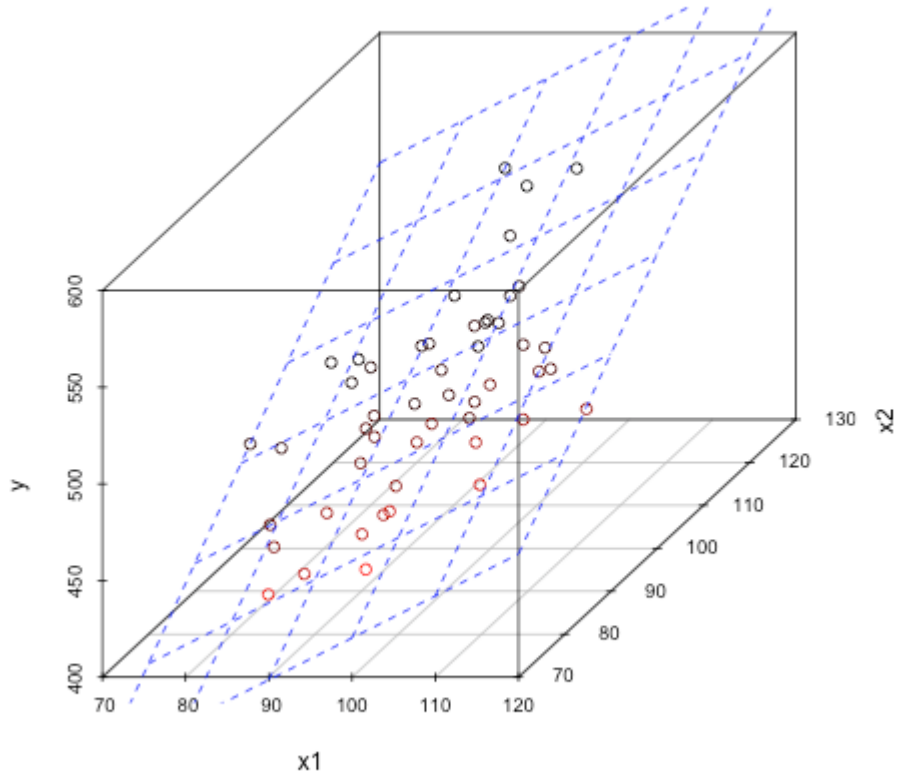


この予測は意味をなします。価格が高いほど、家の面積が多くなります。もちろん、これは比較的簡単な例です。複数の予測変数を調べることで、より複雑な方程式を使って家の価格を予測することができます。

たとえば、面積を考慮することに加えて、家の築年数も考慮したいと考えているとします。同様に、目的は、次の式の係数のセットを見つけることです。

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \qquad (3)$$

この例では複数の予測因子が存在するため、モデルを視覚化するために複数の次元が必要です。2 つの予測因子と 1 つの応答を使用すると、次のようにモデルを視覚化できます。



Vertica では、線形回帰を使用するには、以下を実行する必要があります。

1. トレーニングデータを作成します。
2. モデルを構築します。
3. 新しいデータにモデルを使用して予測を行います。

Prestige データセット

Prestige データセットを使用した例を見てみましょう。このデータセットを使用して、所得が他の変数とどのように関連しているかを調べることができます。このデータセットはオンラインで入手可能です[2]。

この例では、データセットをロードし、Vertica で LINEAR_REG 関数を使用する方法を示します。予測因子がどのように応答に影響を与えるかを理解するには、各自の Vertica データベースでこの例を使用する必要があります。どの予測因子が収入に最も大きな影響を及ぼしているか把握できますか？

データセットには、次の情報が含まれています。

職業名
教育（年）
所得 - 1971 年の現職者の平均収入（ドル）
女性 - 女性である現職者の割合
プレステージ - 1960 年代半ばに実施された社会調査から、職業の Pineo-Porter プレステージスコア。
国勢調査 - カナダの国勢調査の職業コード

タイプ - 職業タイプ。bc はブルーカラーを示し、wc はホワイトカラーを示し、prof は専門職、管理職または技術職

ゴールは、データセットの他の値に基づいて収入を予測するこのデータセットを使用して線形回帰モデルを構築することです。次に、モデルの適合度を評価する必要があります。

では、このモデルでどの変数を選択するかはどのように選択するのでしょうか？ Vertica は、現在、カテゴリー予測変数をサポートしていないため、タイプの列を削除できます。職業名および国勢調査の列には、多くのユニークな値が含まれています。これらの列は、おそらくユースケースの収入を予測するのに最も適していません。そのため、教育、プレステージ、女性を選んでみましょう。

注：実際には、Vertica の線形回帰モデルをトレーニングするためのカテゴリー予測変数を検討する必要がある場合は、事前に数値に変換してください。カテゴリー変数を数値の変数に変換するには、いくつかの手法があります。たとえば、ワンホットエンコードを使用できます。

データをロードする

次に、プレステージのデータセットを格納するテーブル定義を示します。

```
=> DROP TABLE IF EXISTS public.prestige CASCADE;
=> CREATE TABLE public.prestige (
    occupation VARCHAR(25),
    education NUMERIC(5,2), -- avg years of education
    income INTEGER, -- avg income
    women NUMERIC(5,2), -- % of woman
    prestige NUMERIC(5,2), -- avg prestige rating
    census INTEGER, -- occupation code
    type CHAR(4) -- Professional & managerial (prof)
                 -- White collar (wc)
                 -- Blue collar (bc)
                 -- Not Available (na)
   );
```

プレステージのデータセットから Vertica のテーブルにデータをロードするには、次の SQL 文を使用します。

```
=> COPY public.prestige
   FROM stdin
    DELIMITER ','
    SKIP 1
    ABORT ON ERROR
    DIRECT ;
```

線形回帰モデルを作成する

ここで、Vertica の機械学習関数 LINEAR_REG を使って線形回帰モデルを作成しましょう。

モデルを作成するには、次のように public.prestige テーブルに対して LINEAR_REG 関数を実行します。このステートメントでは、収入は応答であり、予測変数は教育、女性、プレステージです。

```
=> SELECT LINEAR_REG(
            'prestige',
            'public.prestige',
            'income',
            'education,women,prestige');
```

このステートメントは、次の式の係数を特定しようとしています。

$$income = \alpha + \beta_1 education + \beta_2 women + \beta_3 prestige \tag{4}$$

モデルを作成後、SUMMARIZE_MODEL 関数を使用してモデルのプロパティを観察します。

```
=> SELECT SUMMARIZE_MODEL('prestige');
```

SUMMARIZE_MODEL は次の情報を返します。

```
SUMMARIZE_MODEL| coeff names : {Intercept, education, women, prestige}
coefficients: {-253.8390442, 177.1907572, -50.95063456, 141.463157}
p_value: {0.83275, 0.37062, 4.1569e-08, 8.84315e-06}
```

これらの係数を使用して、方程式(4)を書き直すと次のようになります。

$$\begin{aligned} income = & -253.8390442 \\ & +177.1907572 * education \\ & -50.950663456 * women \\ & +141.463157 * prestige \end{aligned} \tag{5}$$

最後に、線形回帰モデルがデータにどの程度適合しているかを測定する方法を探っていきましょう。Vertica では、PREDICT_LINEAR_REG 関数は入力テーブルに線形回帰モデルを適用します。この関数の詳細については、Vertica documentation を参照してください。

適合度

線形回帰モデルが観測データにどの程度適合しているかをテストするために使用される一般的な方法は、決定係数です。係数は次の式で定義されます。

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{6}$$

決定係数 R2 は、0(非適合)と 1(完全適合)の間の範囲となります。決定係数を計算するには、Vertica の RSQUARED 関数を使用します。

```
=> SELECT RSQUARED(income, predicted) OVER()
   FROM ( SELECT
```

```
          income,
          PREDICT_LINEAR_REG (
            prestige, women
            USING PARAMETERS OWNER='dbadmin',
            MODEL_NAME='prestige')
          AS predicted
          FROM public.prestige
     ) x ;
               rsq   | comment
-------------------+-------------------------------
  0.63995924449805 | Of 102 rows, 102 were used ...
```

注：OWNER パラメーターは、Vertica 8.1 で非推奨となる予定です。

決定係数の評価は、あなたが調査している分野によって決まることがよくあります。社会科学では、係数 0.6 はかなり良いと考えられています。[3]

モデルを評価する際には、複数の指標を考慮する必要があります。メトリックが 1 つのメトリックであれば、良い値が得られるかもしれませんが、モデルそのものは必要なほど役に立たない可能性があります。適合度を評価するには、R 二乗値と他のメトリックを理解することが重要です。

このシリーズでは、次の機械学習のブログに注目してください！

参考文献

[1] Prestige Data Set
[2] Canada (1971) Census of Canada. Vol. 3, Part 6. Statistics Canada.
[3] Julian J. Faraway. Linear Models of R, second edition. CRC Press, 2014
[4] Machine Learning for Predictive Analytics, Vertica documentation.
[5] Machine Learning Functions, Vertica documentation.
[6] Machine Learning for Predictive Analytics, Hands On Vertica.