



第6章 モデル診断とモデル 選択

model diagnosis and selection

HDS 中村 知繁

はじめに：線形回帰モデルの復習

このチュートリアルへようこそ！

個人の医療費を予測する線形回帰モデルを題材に、データ分析の核心的なプロセスを学びます。

ゴール：単にモデルを作って終わりではなく、データと対話しながらモデルを診断し、より良いモデルへと「育てていく」一連の思考プロセスを体験すること。

本日の学習ステップ

モデル改善のための強力な武器を身につけていきましょう。

1. ベースラインモデル構築

全ての分析の出発点となる最初のモデルを立てます。

2. モデルの診断と改善①

目的変数の「対数変換」を行い、モデルの基本的な仮定を満たすようにします。

3. モデルの改善②

「交互作用項」を追加し、より現実に即した複雑な関係性を捉えます。

4. 総合的なモデル選択

複数の評価指標を駆使して、最終的なモデルを決定します。

第1章

すべての基本、線形回帰モデル

1.1 線形回帰モデルとは？

予測したい変数（**目的変数**）を、予測に使う変数（**説明変数**）の足し算で説明するシンプルなアイデアです。

ベースラインモデル

$$\begin{aligned}\text{charges} = & \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{bmi} + \beta_3 \times \text{children} \\ & + \beta_4 \times \text{sex}_{\text{male}} + \beta_5 \times \text{smoker}_{\text{yes}} \\ & + \beta_6 \times \text{region}_{\text{northwest}} + \beta_7 \times \text{region}_{\text{southeast}} \\ & + \beta_8 \times \text{region}_{\text{southwest}} + \epsilon\end{aligned}$$

1.1 線形回帰モデルとは？(続き)

- Y (**charges**): 予測したい**目的変数**（年間医療費）。
- X (**age, bmi**など): 予測の材料となる**説明変数**。カテゴリ変数は0/1の**ダミー変数**に変換します。
- $\beta_1, \beta_2, \dots$ (**回帰係数**): 各説明変数が目的変数に与える影響の大きさ（重み）。
 - 例：年齢が1歳上がると、医療費は平均で β_1 ドル増える。
- β_0 (**切片**): 全ての説明変数が0のときのベースとなる医療費。
- ϵ (**誤差項**): モデルでは説明しきれない変動のすべて。

1.2 係数の推定：最小二乗法

モデルの係数(β)は、**最小二乗法 (Ordinary Least Squares, OLS)** で決定されます。

アイデア：「予測値と実際の値の差（**残差**）の二乗和」が最小になるように係数を決める。

$$\text{残差平方和(RSS)} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \rightarrow \quad \text{これを最小化する}$$

- y_i : i 番目の人の**実際の**医療費
- $\hat{y}_i$: モデルが予測した i 番目の人の医療費

これにより、全体として最も「当てはまりの良い」モデルを作成します。

第2章

モデルの診断と改善① - 対数変換による安定化

2.1 なぜ変数変換が必要か？

ベースラインモデルの「健康診断」をします。特に、目的変数 `charges` の分布に課題がありました。

これには2つの問題があります。

1. **正規性の仮定違反:** 誤差項 ϵ が正規分布に従う、という線形回帰の仮定が満たされにくい。
2. **不等分散性:** 予測値が大きいほど、誤差のばらつきも大きくなる。残差プロットがラッパ状に広がり、高額な医療費での予測が不安定になります。

2.2 対数変換という「治療法」

この問題を解決する強力な手法が、目的変数の**対数変換**です。

$$\log(\text{charges}) = \beta'_0 + \beta'_1 \times \text{age} + \dots + \epsilon'$$

対数をとることで、大きな値は小さく圧縮され、分布が正規分布に近づきます。
結果として、不等分散性が改善され、より安定したモデルになることが期待できます。

【解釈のポイント】

係数の解釈が変わります。

- **変更前:** 年齢が1歳上がると、医療費は β_1 **ドル** 増加する。
- **変更後:** 年齢が1歳上がると、医療費は **約** $100 \times \beta'_1\%$ 増加する。
(変化が「絶対額」から「変化率」になります)

第3章

モデルの改善② - 交互作用による表現力向上

3.1 交互作用とは？

“ 「ある変数の効果が、別の変数の水準によって変化すること」

”

EDAから、年齢と医療費の関係が**喫煙の有無**で大きく異なることが示唆されました。

- **非喫煙者**: 年齢とともに医療費は**緩やかに**増加する。
- **喫煙者**: 年齢とともに医療費は**急激に**増加する。

これは、**年齢の効果が喫煙ステータスによって変わる**典型的な交互作用の例です。
この関係性をモデルに組み込むのが**交互作用項**です。

3.2 交互作用項の導入と解釈

モデル式に **年齢** × **喫煙ダミー** という新しい項を追加します。

$$\log(\text{charges}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{smoker}_{\text{yes}} + \beta_3 (\text{age} \times \text{smoker}_{\text{yes}}) + \dots$$

- 非喫煙者 (**smoker_yes = 0**) の場合：

$$\log(\text{charges}) = \beta_0 + \beta_1 \text{age} + \dots$$

年齢が1歳上がると、 $\log(\text{charges})$ は β_1 増加します。

- 喫煙者 (**smoker_yes = 1**) の場合：

$$\log(\text{charges}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{age} + \dots$$

年齢が1歳上がると、 $\log(\text{charges})$ は $(\beta_1 + \beta_3)$ 増加します。

β_3 は「喫煙による、年齢効果への**追加的な上乘せ分**」を意味します。

第4章

総合的なモデル選択と最終評価

4.1 モデル選択の4つの視点

複数のモデル候補から、どれを採用すべきか？異なる指標で多角的に評価します。

評価指標	目的	特徴と計算式
決定係数 (R^2)	モデルの説明力 (基本)	目的変数の変動のうち、モデルで説明できた割合。変数を増やすだけで上昇する欠点あり。 $R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$
自由度調整済み決定係数 (Adj. R^2)	モデルの説明力 (改善版)	変数の数を考慮して R^2 を調整。異なる変数を持つモデル間の比較に適している。 $\text{Adj. } R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$
赤池情報量規準 (AIC)	当てはまりと簡潔さのバランス	当てはまりが良く、シンプルなモデルほど値が小さい。 AICが最も小さいモデルが良い。 $\text{AIC} = n \log(\text{RSS}/n) + 2(p + 1)$
RMSE (二乗平均平方根誤差)	モデルの予測精度	予測誤差の平均。小さいほど良い。テストデータで計算し、汎化性能を測る。 $\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}$

4.2 最終モデルの決定プロセス

1. 訓練データでの比較

- 複数のモデル候補の **Adj. R^2** と **AIC** を比較します。
- Adj. R^2 が高く、AICが低いモデルほど、訓練データへの当てはまりが良いと言えます。

2. テストデータでの最終評価

- 最も性能が良さそうなモデルを使い、未知のデータ（**テストデータ**）に対する予測精度を **RMSE** で評価します。
- これが、そのモデルの真の実力（**汎化性能**）を示す最終スコアとなります。

“ **注意点:** 対数変換モデルのRMSEを計算する際は、予測値 `log(charges)` を `np.expm1()` などでも元のドル単位に戻してから計算するのを忘れないようにしましょう。

”

結論

データ分析は対話のプロセス

まとめ

このチュートリアルを通じて、私たちはモデルを継続的に改善するプロセスを学びました。

- **診断:** 残差分析でモデルの健康状態をチェック
- **治療:** 変数変換でモデルの仮定を満たす
- **機能拡張:** 交互作用項で複雑な関係性を捉える

Key Takeaways

- **Adjusted R^2** や **AIC** は、手元のデータでモデルを比較検討するための羅針盤です。
- 最終的なモデルの価値は、未知のデータに対する予測精度、すなわち**テストデータでのRMSE**によって決まります。

線形回帰で学んだこの思考法は、より高度な機械学習モデルにも通じる普遍的なスキルです。ぜひこの経験を活かして、さらなるデータ分析の世界を探求してください。