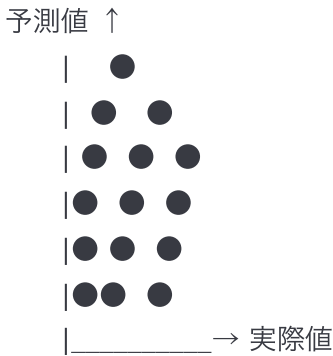


残差プロット完全解説ガイド

1. 実際値 vs 予測値プロット (Actual vs Predicted)

基本的な見方



✓ 理想的なパターン

特徴:

- すべての点が赤い対角線 ($y=x$) の近くに密集している
- 点に対角線からの距離が均等に散らばっている
- 外れ値がほとんどない

意味:

- モデルの予測が実際の値と非常に近い
- 過小評価・過大評価のバイアスがない

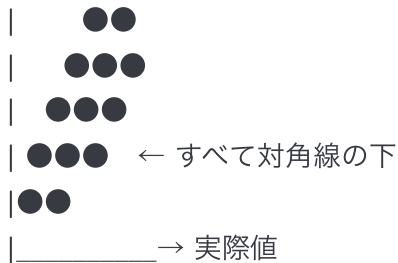
例:

実際値: 0.05 → 予測値: 0.051 ← 良い！
実際値: 0.10 → 予測値: 0.098 ← 良い！
実際値: 0.15 → 予測値: 0.152 ← 良い！

⚠ 問題があるパターン

パターンA: 系統的な過小評価

予測値 ↑



問題:

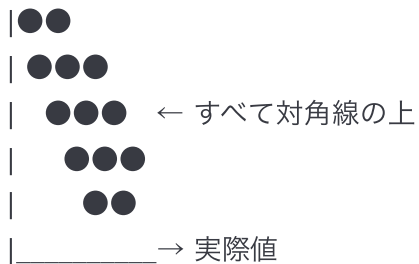
- 高い値を予測するときに常に低めに予測している
- モデルが保守的すぎる

対策:

- 高リスク領域のデータを増やす
- 特徴量を追加して高値を説明できるようにする

パターンB: 系統的な過大評価

予測値 ↑



問題:

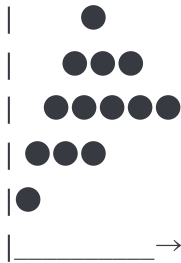
- 低い値を予測するときに常に高めに予測している
- モデルが過敏すぎる

対策:

- 正則化を強くする
- ノイズデータを除去する

パターンC: S字型（非線形性）

予測値 ↑



問題:

- 中間の値では良いが、極端な値で予測が悪い
- モデルが非線形関係を捉えきれていない

対策:

- 多項式特徴量を追加 (curvature², log(speed)など)
- より複雑なモデルを使用

統計指標の解釈

RMSE (Root Mean Squared Error):

$$\text{RMSE} = \sqrt{(\sum (\text{実際値} - \text{予測値})^2 / n)}$$

例: RMSE = 0.05569

→ 平均的に予測値は実際値から±0.0557ずれている

解釈:

- 小さいほど良い
- 目的変数のスケールで評価する
 - 事故リスク (0-1範囲) でRMSE=0.055 → **約5.5%の誤差** (良好)
 - 住宅価格 (数千万円) でRMSE=0.055 → **55円の誤差** (完璧)

MAE (Mean Absolute Error):

$$\text{MAE} = \sum |\text{実際値} - \text{予測値}| / n$$

例: MAE = 0.045

→ 平均的に予測値は実際値から0.045ずれている

解釈:

- RMSEより外れ値の影響を受けにくい

- 実務的な誤差の目安になる

R² (決定係数):

$$R^2 = 1 - (\text{予測の誤差} / \text{平均からの誤差})$$

例: $R^2 = 0.85$

→ モデルがデータの分散の85%を説明している

解釈:

- 1に近いほど良い (1 = 完璧)
- 0以下 = モデルが無意味
- 目安:
 - $R^2 > 0.9$: 非常に良い
 - $R^2 > 0.7$: 良い
 - $R^2 > 0.5$: まあまあ
 - $R^2 < 0.5$: 改善が必要

2. 残差プロット (Residual Plot)

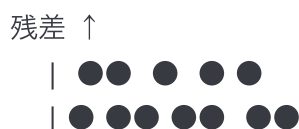
基本的な見方



残差 = 実際値 - 予測値

- 正の残差: モデルが**過小評価** (実際より低く予測)
- 負の残差: モデルが**過大評価** (実際より高く予測)
- 残差=0: **完璧な予測**

理想的なパターン (ランダム)





特徴:

- 残差が0を中心にランダムに散らばる
- パターンや傾向が見えない
- 分散が一定（等分散性）

意味:

- モデルが適切に学習できている
- バイアスがない
- これ以上改善の余地が少ない

⚠ 問題があるパターン

パターンA: 右上がりの傾向



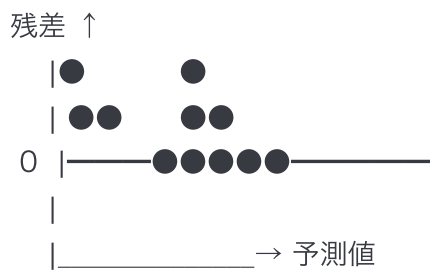
問題:

- 予測値が大きくなるほど過小評価している
- モデルに系統的なバイアスがある

対策:

- 特徴量の変換（対数変換など）
- 新しい特徴量を追加
- モデルの複雑度を上げる

パターンB: U字型（非線形パターン）



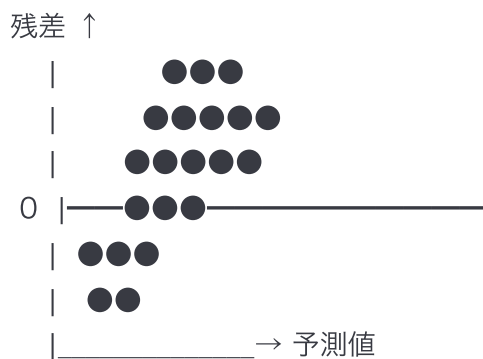
問題:

- 中間の予測値で過大評価、極端な値で過小評価
- モデルが非線形関係を捉えていない

対策:

- 多項式特徴量を追加
- 特徴量同士の交互作用項を追加
- より柔軟なモデルを使用

パターンC: 扇形（不均一分散）



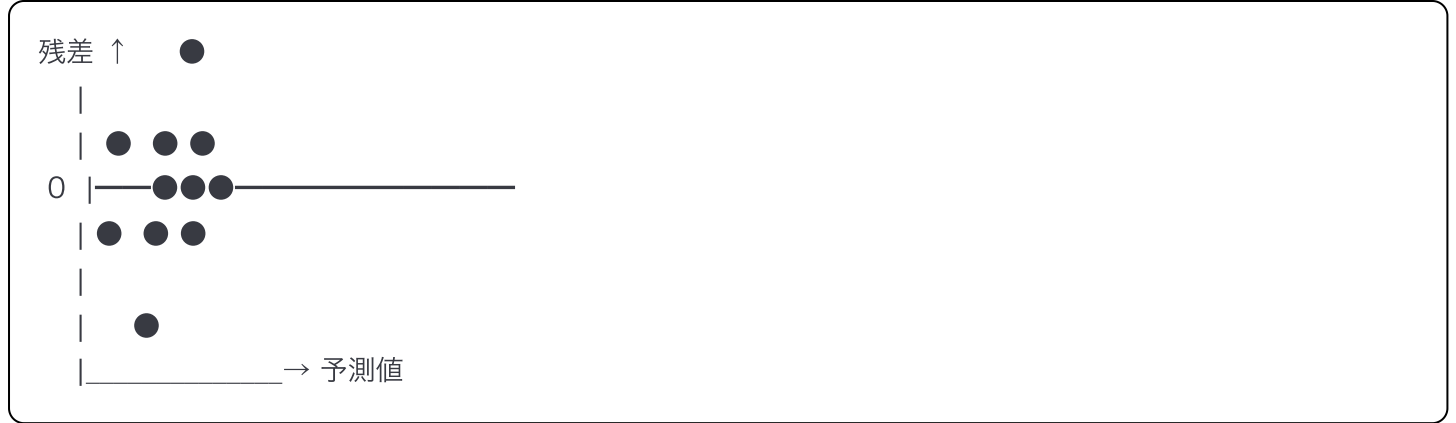
問題:

- 予測値が大きくなるほど誤差が大きくなる
- 等分散性の仮定が崩れている

対策:

- 目的変数を対数変換 ($\log(y)$)
- 重み付き回帰を使用
- 外れ値を調査・除去

パターンD: 明確な外れ値



問題:

- 特定のデータポイントで大きな誤差
- データエラーまたは特殊なケース

対策:

- $\pm 2\sigma$ または $\pm 3\sigma$ を超える点を調査
- データの入力ミスをチェック
- 特殊ケースなら除外を検討

±2σ ラインの意味

正規分布の場合:

- $\pm 1\sigma$ 内: 約68%のデータ
- $\pm 2\sigma$ 内: 約95%のデータ ← これを超えるのは異常
- $\pm 3\sigma$ 内: 約99.7%のデータ

判断基準:

- $\pm 2\sigma$ 外が5%以下: 正常
- $\pm 2\sigma$ 外が10%以上: モデルに問題あり

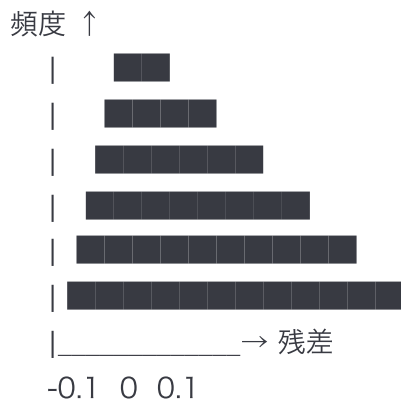
3. 残差のヒストグラム (Distribution of Residuals)

基本的な見方



|_____→ 残差
0

✔ 理想的なパターン（正規分布）



特徴:

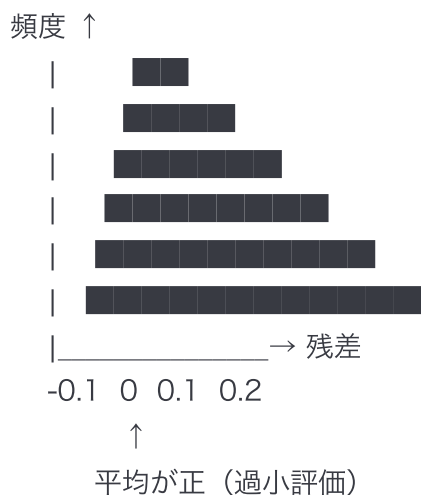
- 平均が0に近い
- 左右対称の釣鐘型
- 正規分布に近い形

意味:

- 誤差がランダムに分布
- モデルの仮定が適切
- 過小評価・過大評価が均等

⚠ 問題があるパターン

パターンA: 平均がゼロでない（バイアス）



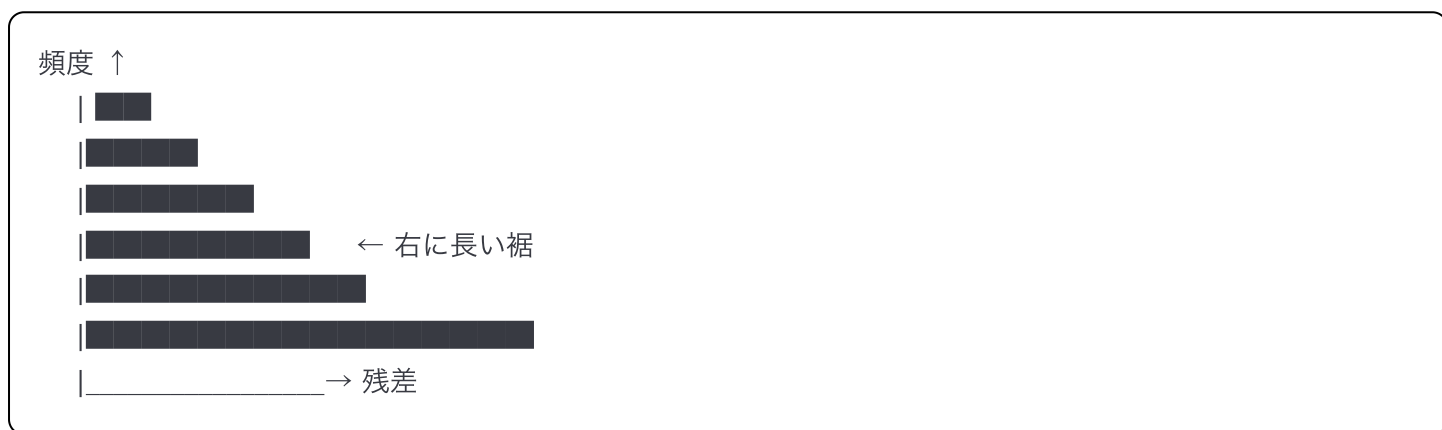
問題:

- 系統的に予測が低い（または高い）
- モデルにバイアスがある

対策:

- 切片項の調整
- 特徴量の標準化
- 新しい特徴量の追加

パターンB: 左右非対称（歪度がある）



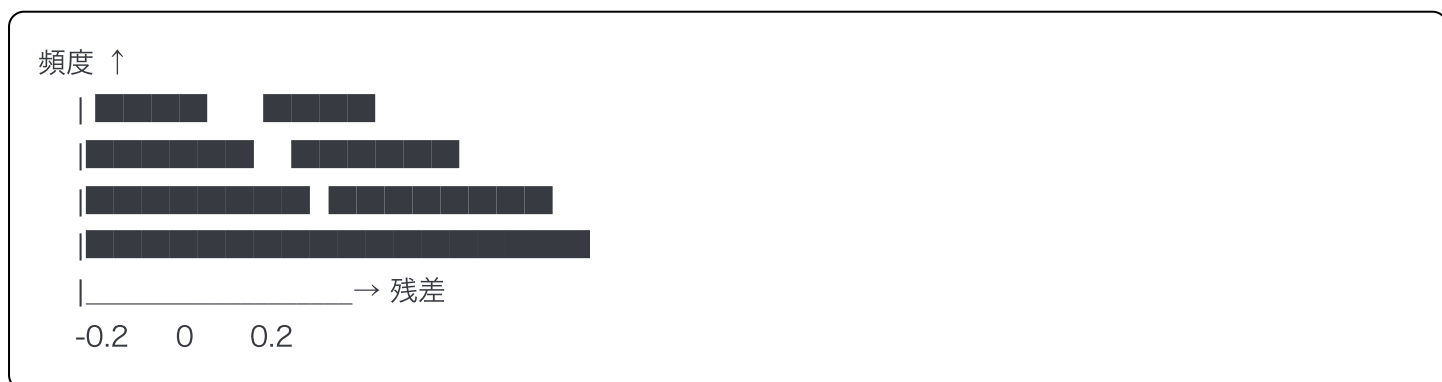
問題:

- 大きな正の残差（または負の残差）が多い
- 分布が偏っている

対策:

- 目的変数を変換（log, sqrt）
- 外れ値を調査
- ロバストな損失関数を使用

パターンC: 二峰性（複数のピーク）



問題:

- データに2つ以上のグループが混在
- 異なる挙動のデータが混ざっている

対策:

- データを層別化して分析
- カテゴリ変数を追加
- 異なるモデルを使い分ける

パターンD: 裾が重い（外れ値が多い）



問題:

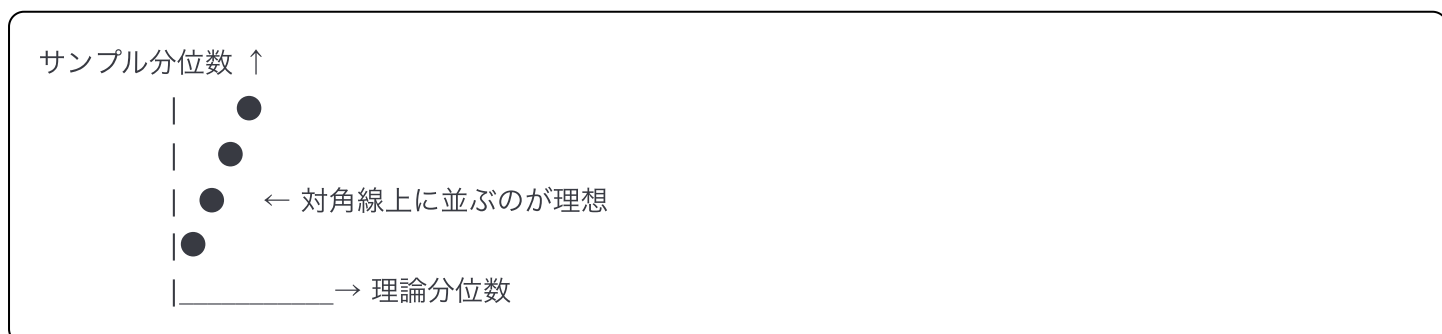
- 正規分布より極端な値が多い
- ロバスト性が不足

対策:

- 外れ値の除去または別処理
- ロバストな回帰手法（Huber損失など）
- アンサンブル学習

4. Q-Qプロット（Quantile-Quantile Plot）

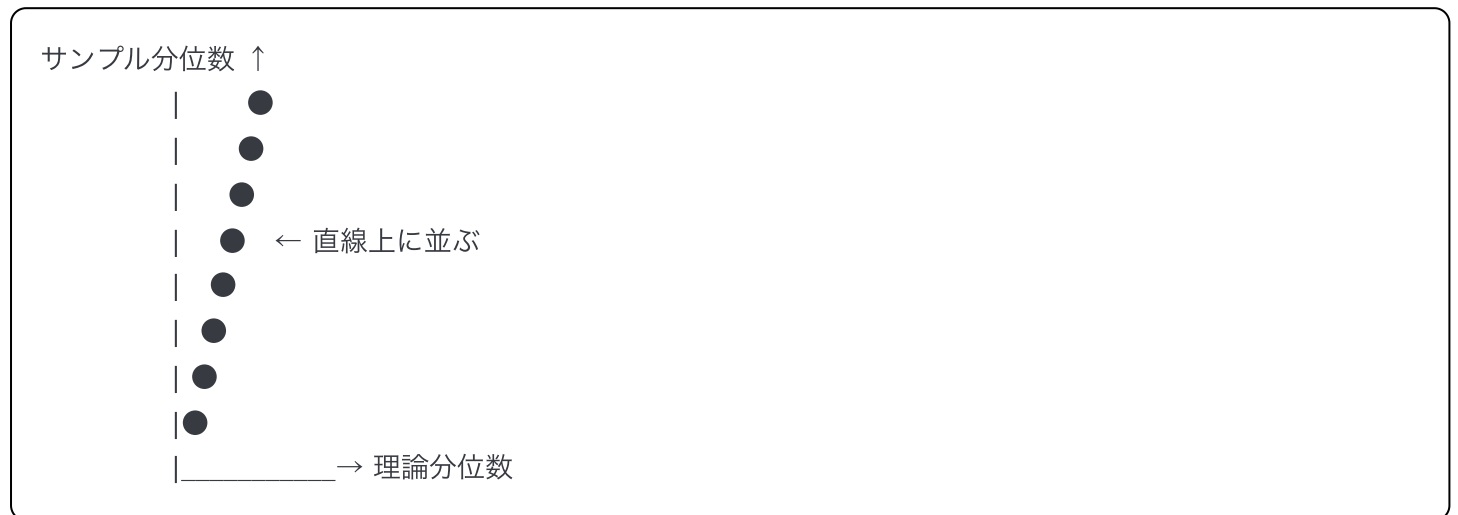
基本的な見方



Q-Qプロットとは:

- 残差の分布が正規分布に従うかを確認
- X軸: 理論的な正規分布の分位数
- Y軸: 実際のデータの分位数

✅ 理想的なパターン（正規性あり）



特徴:

- すべての点が赤い対角線上に並ぶ
- ズレが小さい

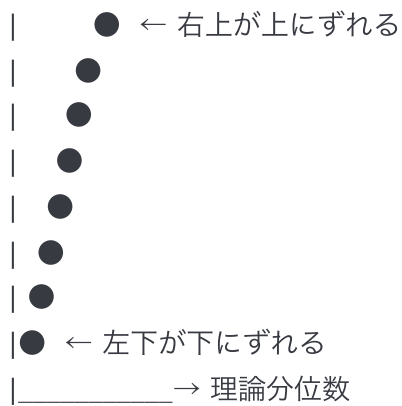
意味:

- 残差が正規分布に従う
- 統計的仮定が満たされている
- 信頼区間や検定が有効

⚠ 問題があるパターン

パターンA: 裾が重い (Heavy Tails)

サンプル分位数 ↑



問題:

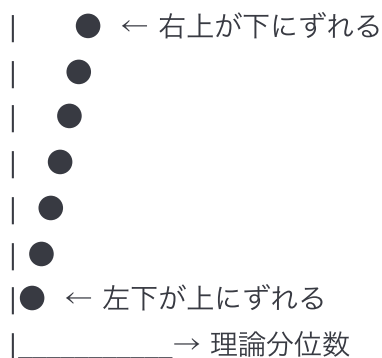
- 極端な値が正規分布より多い
- 外れ値が多い

対策:

- 外れ値を調査・除去
- ロバストな手法を使用
- データ変換（Box-Cox変換など）

パターンB: 裾が軽い（Light Tails）

サンプル分位数 ↑



問題:

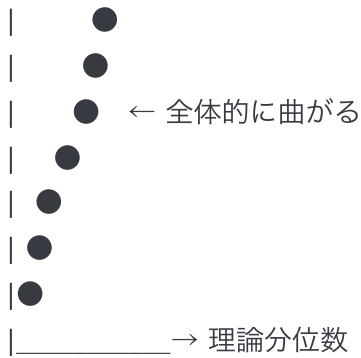
- 極端な値が正規分布より少ない
- 分布が正規分布より尖っている

対策:

- データの生成過程を確認
- モデルが単純すぎる可能性
- より複雑なモデルを試す

パターンC: 歪度がある (Skewness)

サンプル分位数 ↑



問題:

- 分布が左右非対称
- 正規分布の仮定が崩れている

対策:

- 目的変数を変換 (log, sqrt, Box-Cox)
- 歪度を持つ分布を仮定したモデル

🎯 総合的な診断フロー

Step 1: 実際値 vs 予測値プロット

↓

対角線に沿っている？

YES → 良好！

NO → Step 2へ

Step 2: 残差プロット

↓

ランダムに散らばっている？

YES → Step 3へ

NO → パターンあり

→ 特徴量追加・変換が必要

Step 3: 残差のヒストグラム

↓

平均 \approx 0、左右対称？

YES → Step 4へ

NO → バイアスあり

→ モデルの再調整

Step 4: Q-Qプロット



直線上に並んでいる？

YES →  モデル完璧！

NO → 正規性の問題

→ データ変換検討

チェックリスト

良いモデルの条件

- ☐ 実際値 vs 予測値が対角線に沿っている
- ☐ RMSE < 目的変数の標準偏差の30%
- ☐ $R^2 > 0.7$
- ☐ 残差プロットにパターンがない
- ☐ 残差の平均 ≈ 0
- ☐ 残差が正規分布に近い
- ☐ $\pm 2\sigma$ 外の外れ値 < 5%
- ☐ Q-Qプロットが直線上

改善が必要なサイン

- ☐ 残差に明確なパターンがある
- ☐ 残差の分散が一定でない（扇形）
- ☐ 残差の平均が0から大きくずれている
- ☐ 外れ値が10%以上
- ☐ $R^2 < 0.5$
- ☐ Q-Qプロットが大きくずれている

実践的なアドバイス

1. まず実際値 vs 予測値を見る

- 全体的な予測精度の把握
- 系統的なバイアスの確認

2. 残差プロットで詳細分析

- パターンがあれば特徴量追加
- 扇形なら変数変換

3. ヒストグラムで分布確認

- 平均がゼロでなければバイアス修正
- 歪んでいればデータ変換

4. Q-Qプロットは最後の確認

- 統計的仮定のチェック
- 大きな問題がなければOK

重要: 完璧なモデルは存在しない！少しのズレは許容範囲です。