*定量的ソフトウェア開発管理　第三回レポート*

*43M22403　古谷康平*

課題説明①開発工数を予測する線形重回帰モデルを構築する

②構築したモデルの予測精度を評価する

扱うデータ：**FakedataMonden.csv**

1. に関しては、FakedataMonden.csvの2005-2011年のデータを使った
2. に関しては、FakedataMonden.csvの2012-2013年のデータを使った
3. **データクリーニング**：不要データ削除

下記の112と114のような開発規模と開発工数が欠損しているデータは削除した。

また、124のような開発規模が多いにもかかわらず開発工数が0のデータは削除した。

画像には示していないが、開発期間が欠損しているデータも削除した。

カレンダー

自動的に生成された説明

1. **データクリーニング**：外れ値の特定と削除

生産性（計算：開発規模＿FP/開発工数＿人月）の外れ値を削除する。log(生産性)の値を参照し、下記のように極端に次のデータと離れているデータは削除した。

テーブル

自動的に生成された説明　（右記：log(生産性)の値、１、２番が外れ値）

1. **モデル化：変数選択**

PM経験年数と生産性の相関図を下記のように作成したところ、相関が見られなかったので変数「PM経験年数」は不採用とした。

（相関係数：-0.01785）

グラフ, 散布図

自動的に生成された説明

また、カテゴリ変数である「開発種別」、「アーキテクチャ」は下図のようにそれぞれの項目にほとんど差がないため、不採用とした。

グラフ, 箱ひげ図

自動的に生成された説明（開発種別）

グラフ, 箱ひげ図

自動的に生成された説明

また、下図を見ると要求仕様が明確であるかどうかと生産性には特に目立った関係が見られないため、「要求仕様＿明確度合」変数は不採用とした。グラフ, 散布図

自動的に生成された説明

以上で変数選択は終了である。

1. **エクセルを使ったモデル構築：*回帰分析***

カテゴリ変数「業種」と「主開発言語」は下図のようになる。

グラフ, 箱ひげ図

自動的に生成された説明

グラフ, 箱ひげ図

自動的に生成された説明

ここで「業種」では、生産性の値の高い（製造・建設）と、値の低い（金融・研究）に着目する。

同じ理由で「主開発言語」では、（ASP）と（COBOL）に着目する。

開発工数は、業種、主開発言語、開発規模、開発期間を説明変数として予測する。

2005-2011年のデータを回帰分析すると結果は以下のようになった。

テーブル

自動的に生成された説明

*構築されたモデル式*：log(開発工数)= -1.3666+0.001053\*製造・建設

+0.314033\*金融・研究-0.34828\*ASP+0.155882\*COBOL

+0.412319\*log(開発期間) +0.898564\*log(開発規模)

開発工数の予測値（上記のモデル式）と実測値の関係は以下のようになる。

モデル式の予測精度は高いと考えられる。

グラフ, 散布図

自動的に生成された説明

1. **予測精度算出**

ここで、2012-2013年のデータを使ってモデルの予測精度を評価する。

エクセルでABS関数を使って予測値と実測値の差（残差）を求め、SUM関数で合計を出して絶対誤差平均を算出した。

また、求めた残差を実測値でわって、SUM関数で合計を出して相対誤差平均を算出した。

結果は以下のようになる。

・絶対誤差平均＝0.2749349

・相対誤差平均＝0.2034390

1. 得られた感想

エクセルの計算を久々にしたので時間はかかったが、実際の開発のデータ分析をしているようで将来のためになるなと感じた。