2024 年 7 月 15 日
情報システム工学実験レポート

球体-木製ブロック衝突実験の重回帰分析を用いた分析

情報経営システム工学分野 B3
学籍番号 ： 学籍番号
氏名 ： 本間三暉
グループ名 ： 10

1　はじめに

1.1　前置き

今回の実験で物理実験を行うに当たり，16 班に分かれ，各班ごとに実験を行い実験結果を Excel にまとめた．その実験結果を一つにまとめる際ラベル等の指定がなかったため，おおよそデータの分析を行うとは思えないような形式でまとめている班が散見された．

そのため，各班に呼びかけ形式の統一を行うよう呼びかけた．しかし，レポート執筆時点 (2024 年 6 月 27 日) で，形式の統一が行われておらずどの数字が実験データかわからない班が 3 班，単位が正しく直されていなさそうな班が 1 班いた．

そのような数字はデータの前処理の段階で排除したため，16 班分のデータを用いた想定通りのデータ分析が行えないことを断っておきたい．また，形式を統一した際に，平均値のみを提出した班と生データを提出した班があるため，データの数に偏りがあることも断っておきたい．

1.2　目的

現代社会において，データサイエンスの重要性はますます高まっている．データサイエンスとは，与えられたデータに基づいて知見を見出し，その知見を次の行動に活かすことを目的としている．本レポートでは，データ分析の基本的な手法とその応用について実験を通じて探求する．

本実験は，データ分析の 5 つの手順 (PPDAC サイクル) に基づいて進められる．PPDAC サイクルは，問題の把握 (Problem)，調査の計画 (Plan)，データの収集 (Data)，データの分析 (Analysis)，および結論の考 (Conclusion) の 5 つのステップから成り立つ．これらのステップを踏むことで，データから有用な情報を引き出し，具体的な課題解決に繋げることを目指す．

本レポートの目的は，実験を通じてデータ分析の基本的な手法を理解し，実際のデータを用いた分析のプロセスを経験することである．さらに，得られた知見をもとに具体的な解決策を提案し，実務に応用できるスキルを身につけることを目指す．

2　データ (実験概要)

今回の実験では，図 1 のような装置を用いて行う．

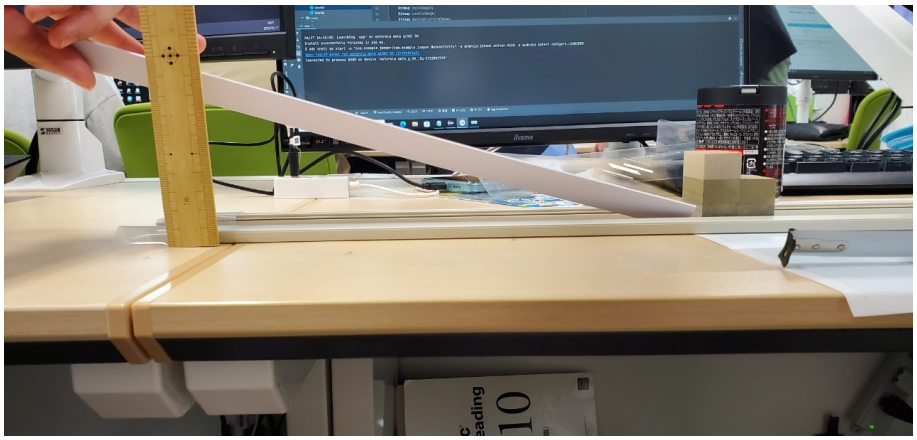

図 1　実験装置

今回，私の班はボールが転がり始める高さ (以下高さ)，ボールが転がった距離の内地面と平行な向きに移動した距離 (以下底辺)，物体の移動距離 (以下移動距離) を測定し，ボールが転がる距離 (以下斜辺)，斜辺と底辺のなす角 (角度) などは計算で導出することにした．斜辺の導出方法は，斜辺を C，高さを A，底辺を B とした時，三平方の定理から式 (2) で導出できる．

$$C^2 = A^2 + B^2 \tag{1}$$

$$C = \sqrt{A^2 + B^2} \tag{2}$$

また，角度は測定値である高さと底辺から導出する．高さを A，底辺を B，角度を θ とした時，タンジェントの定義式から，アークタンジェントを用いた式 (4) から導出できる．

$$\tan\theta = \frac{A}{B} \tag{3}$$

$$\theta = \arctan\left(\frac{A}{B}\right) \tag{4}$$

Excel の関数 ATAN() は戻り値の単位がラジアンなので，これを度数法に直す必要がある．弧度法は，扇形を考えた時，中心角 θ は円弧の長さ l に比例する．円弧の長さ l と扇型の半径 r の比をとると，同じ角度 θ に対して扇型の大きさにかかわらずこの比は一定である．このような性質を利用して，角度の大きさは式 (5) を用いて定義されている．

$$\theta = \frac{l}{r} \tag{5}$$

よって弧度法から度数法に変換するためには，弧度法の角度を θ，度数法の角度を θ' とした時，式 (6) を用いれば良い．

$$\theta' = \frac{360}{2\pi}\theta \tag{6}$$

また，今回は情報経営システム工学分野 B3 の学生を 4 人で 1 つの班に分け，全部で 16 班作成する．各班ごとに使用するボールの重さや材質，物体の重さを変え実験を行う．ボールの種類は以下の 7 種類である．

- 鉄球 1
- 鉄球 2
- 鉄球 3
- ビー玉 1
- ビー玉 2
- スーパーボール
- ゴムボール

また，物体の重さは 5 段階である．

私の班は重さが 66.9$[g]$ の鉄球をボールに用いて，重さが 25.7$[g]$ で地面との接地面積が 1350$[mm^2]$ の物体を用いて実験を行った．

3　データの前処理

3.1　データの前処理について

データの前処理は，データ分析工程の 8 割を占めるとも言われるほど重要な工程である．必要な情報が完璧にそろっているデータは珍しく，多くのデータは必要な数値が抜けている場合がほとんどであ

る．さらに，データごとに形式が違うこともあり，数値に変換されていないテキストデータ（文字データ）しかない場合も多々ある．このような状態のデータではエラーが発生しやすく，不十分な結果しか得られないため，データの前処理を行って事前にデータを整理する必要がある．データの形式や質・量が予測精度を決定するため，前処理はスムーズに分析を行うために不可欠であり，予測精度を担保するための重要な開発工程である．

データの前処理には大きく分けて 5 つの方法がある．

3.1.1 数値型への変換

本来数値型が入力されるべき場所に文字列や記号が入力されている場合がある．例えば，Excel のデータで他の欄には数値データが入っているにもかかわらず，一箇所に「測定不可」のようなメモが残されている場合がある．そのような文字列データを除去し，全てのデータを数値型に置き換える必要がある．

3.1.2 欠損値処理

欠損値とは，何らかの理由により記載されなかったり欠落したりした値のことである．欠損値の対処法としては以下の方法がある．

- 平均値や最頻値で補完する：欠損値を含むデータが分析に必要な場合は，平均値や最頻値で補完する．欠損値を含むデータを全て除外するとデータ数が不足する場合などは，できる限り補完を検討する．
- 行または列ごと除外する：欠損値の割合が高い行や列がある場合は，そのデータを分析対象から除外する．

3.1.3 外れ値処理

ヒストグラムや散布図を用いてデータの分布を確認し，外れ値の有無を確認する．外れ値の対処法としては以下のようなものがある．

- 正しい値に修正する：外れ値がデータの入力ミスやシステムのエラーによる場合は，正しい値に修正する．
- 行ごと除外する：外れ値が分析結果に大きな影響を与える場合は，行ごと除外する．
- そのまま使用する：外れ値が分析結果に影響しない場合は，そのまま使用することもある．

3.1.4 スケーリング

データの値を特定の範囲に変換する前処理方法である．例えば，異なる単位のデータを比較する際に単位をそろえたり，データの値を同じスケールにそろえたりするためによく使われるスケーリングの手法には，「Min-Max スケーリング」や「Z スコアスケーリング」などがある．

- Min-Max スケーリングはデータを最小値 0，最大値 1 にスケーリングする方法である．
- Z スコアスケーリングは平均 0，分散 1 にスケーリングする方法で，大きな外れ値がある場合に適している．
- Robust スケーリングは，データの中央値を 0，四分位範囲を 1 として変換する手法で，外れ値の影響を受けにくい．

3.1.5 ダミーデータ

ダミー変数の作成がある．質的データやカテゴリカルデータを数字に変換する手法であり，例えばカテゴリを「0」と「1」に変換する方法がある．例えば，「はい」を 1，「いいえ」を 0 にしたり，「男」を 1，「女」を 0 に変換する．3 つ以上のカテゴリがある場合は，各カテゴリを 1 とし，その他を 0 とした数列に変換する．

3.2 グループデータ

実験時にデータを取得した際，高さ 200[mm]，底辺 100[mm] のときボールが物体を弾いてしまい上手くデータが取れなかったため，”N/A”と記述した．しかし，データ分析の際にはそのようなデータは不要なため行ごと削除した．

3.3 統合データ

実験時各班でデータを作成したため，フォーマットや基準にする単位がバラバラになってしまっていた．それだけならいいが，表を横に作成する班，シートを増やしすぎている上に名前を変えずそもそもデータがどこにあるか分かりにくい班，単位を明記していない班など様々であった．

そのため，各班ごとにフォーマットとデータの単位を表 1 のような形に整形するよう呼びかけた．また，表が横に長いため途中で折り返して示す．

表 1　整形後フォーマット

移動距離 [mm]	底辺 [mm]	高さ [mm]	斜辺 [mm]	理論値角度 [度]
20	430	100	441.475	13.092
理論値角度 [ラジアン]	球体の質量 [g]	物体の質量 [g]	球体の種類	物体の底面積 [mmˆ2]
0.228	5.5	10.5	鉄球 1	1305

呼びかけの結果，多くの班でデータのフォーマットが統一された．フォーマットを統一してくれず基準にした単位も記述されていないような修正のしようがない班のデータは除外した．また，ボールの素材によって別の分析をする必要があったのでダミー変数を追加した．ダミー変数を追加する際，式 (7) に示す Excel 関数を用いて作成した．

$$= IF(MID(\$I2, 1, 1) = \text{”鉄”}, 1, 0) \tag{7}$$

ここで，セル $I2$ は球体の種類の列の値である．この式は球体の種類の列を参照し，1 文字目が”鉄”であったら 1，そうでなかったら 0 を入力する関数である．各グループの実験で用いたボールがどの素材であるかは 1 文字目を参照し，ダミー変数を 1 にしたい材質の 1 文字目と比べれば良い．

このようにして表 1 の右に追加したダミー変数を表 2 に示す．

表 2　ダミー変数追加

鉄球	ビー玉	ゴムボール	スーパーボール
1	0	0	0

このように処理したデータを用いて分析を行う．

4　相関分析・単回帰分析

私のグループで取ったデータに関して相関分析と単回帰分析を行う．

4.1　相関分析

相関分析とは，2 つの変数間の関係性を定量的に評価する統計手法である．この分析手法は，様々な分野で広く活用されており，データに隠れた関係性を発見し，因果関係の手がかりを得るために重要な役割を果たしている．

相関分析では，2 つの変数間の関係の強さと方向を相関係数によって表現する．相関係数は-1 から 1 の間の値を取り，1 に近いほど正の相関が強く，-1 に近いほど負の相関が強いことを示す．0 に近い値は，相関がほとんどない，または全くないことを意味する．

相関分析を行う際には，まず散布図を作成して変数間の関係を視覚的に確認することが重要である．散布図から関係のパターン（線形，非線形など）を把握した上で，相関係数を計算し，その値に基づいて相関の強さを解釈する．

ただし，相関分析の結果を解釈する際には，いくつかの注意点がある．相関があることは因果関係を意味するわけではなく，第三の要因が両方の変数に影響を与えている可能性もある．また，相関係数は外れ値に敏感であるため，データの前処理で外れ値の処理が重要となる．さらに，Pearson の相関係数は線形関係を前提としているため，非線形の関係がある場合には他の方法を検討する必要がある．

具体的なデータセットに対して相関分析を適用する際は，データの特性に応じた前処理が必要となる．例えば，カテゴリカルデータを数値に変換するためにダミー変数を使用したり，変数のスケールを揃えるために標準化を行ったりする．このように，相関分析は変数間の関係性を理解するための強力なツールであり，適切に使用することで，データから有益な知見を得ることができる．しかし，分析結果の解釈には注意が必要であり，常に因果関係の可能性を念頭に置きながら，多角的な視点でデータを見ることが重要である．

この実験では Excel の機能を用いて相関分析を行う．相関分析を行った結果について表 3 に示す．

表 3　相関分析

	移動距離 [mm]	底辺 [mm]	高さ [mm]	斜辺 [mm]	理論値角度 [度]
移動距離 [mm]	1				
底辺 [mm]	0.456861417	1			
高さ [mm]	0.898531378	0.136363636	1		
斜辺 [mm]	0.616331666	0.972160283	0.352605016	1	
理論値角度 [度]	0.25097754	-0.66280947	0.589354047	-0.476740815	1

表 3 から移動距離と強い正の相関があるのは高さであることがわかる．よって単回帰分析は移動距離と高さで行う．

4.2　単回帰分析

単回帰分析は，一つの説明変数と目的変数の関係を調べるための統計的手法である．この分析手法は，説明変数の値から目的変数の値を予測するためのモデルを構築することを目的としている．

単回帰分析を実施するには，まず説明変数と目的変数のデータを収集し，それらの関係を散布図で視覚化する．散布図から，両変数の間に直線的な関係が存在するかどうかを判断する．次に，最小二乗法を用いて回帰式の係数を推定する．この回帰式は，説明変数の値から目的変数の値を予測するために使用される．

推定された回帰式の精度を評価するために，決定係数 (R^2) を確認する．R^2 は，回帰式がデータの変動をどの程度説明できるかを示す指標である．R^2 の値が高いほど，回帰式の当てはまりが良いことを意味する．また，回帰式の統計的有意性を確認するために，F 検定や t 検定を行う．

回帰式の係数が統計的に有意であれば，説明変数が目的変数に対して有意な影響を与えていると判断できる．この場合，説明変数の値を変化させることで，目的変数の値を制御できる可能性がある．

強い正の相関がある移動距離と高さについて単回帰分析を行った結果を表 4，5 に示す．

表 4　単回帰分析

回帰統計	
重相関 R	0.89853138
重決定 R2	0.80735864
補正 R2	0.80471971
標準誤差	91.1275435
観測数	75

表 5　分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	2540616.66	2540616.66	305.942502	8.0769E-28
残差	73	606208.73	8304.22918		
合計	74	3146825.39			

	係数	標準誤差	t	P-値
切片	23.0606061	25.5787529	0.90155318	0.37025921
高さ [mm]	3.39827273	0.19428458	17.4912121	8.0769E-28

表 4 を見ると，決定係数 R^2 の値が大きいのでこのモデルにはかなりの説得力がある．表 5 の有意 F は 0.01 以下の値となっていることから統計的に有意である，この回帰モデルには十分な意味があると解釈できる。また，P-値を確認すると 0.01 以下の値となっていることから，高さと移動距離には関係があると考えられる．同様に，t 値の絶対値が 17 を超えることから，高さは移動距離にとても影響する事がわかる．

単回帰分析の結果は目的変数を y，切片を β_0，回帰係数を β_1，説明変数を x としたとき，回帰式は式 (9) のようになる．

$$y = \beta_0 + \beta_1 \times x \tag{8}$$

$$\text{移動距離} = 23.0660 + 3.3983 \times \text{高さ} \tag{9}$$

5　重回帰分析

私のグループで行った実験のデータを用いて重回帰分析した結果を 5.1 節に，全グループのデータを統合したデータに対して住家行く分析した結果を 5.2 節に示す．

5.1 グループデータ

5.2 全グループデータ

6 おわりに

6.1 結論

6.2 考察

6.3 反省点

6.4 要望

- 4 節の解説がなかったので，”スライドを参照すること”などでいいので多少書いてもらえると嬉しかった．
- データを入力する時に班によってフォーマットや基準にする単位が違い，全グループの統合データを作成するのがとても大変だったので，先生の方から表 1 に示すようなテンプレートファイルを配布し，フォーマットや単位の統一を促してほしかった．
- ilias にレポートの提出場所を作るのが提出期限の直前になってしまっていたので実験が終わり次第すぐ上げてもらえると嬉しかった．また，そこに word のテンプレートをつけてほしい．
- あまり期待はしていないが，TeX で書きたい学生もいるので TeX でも対応してほしい．
-