48927 13564

Assessment 3 Group Project Report

Turnitin

Document Details

Submission ID

trn:oid:::22828:115584245

Submission Date

Oct 5, 2025, 3:09 PM UTC

Download Date

Oct 5, 2025, 3:10 PM UTC

File Name

unknown_filename

File Size

235.8 KB

16 Pages

3,520 Words

21,785 Characters

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

- **18** Not Cited or Quoted 5%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
2% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

18 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
2% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
	www.conscientiabeam.com	<1%
2	Internet	
	www.coursehero.com	<1%
3	Publication	
	Onyiah, . "Regression Approaches to the Analysis of Responses of Two-Factor Exp...	<1%
4	Student papers	
	Clark University on 2024-04-19	<1%
5	Internet	
	www.mdpi.com	<1%
6	Student papers	
	University of California, Los Angeles on 2025-07-20	<1%
7	Internet	
	blog.genlaw.org	<1%
8	Internet	
	assets-eu.researchsquare.com	<1%
9	Internet	
	ir.juit.ac.in:8080	<1%
10	Publication	
	H. Whitt Kilburn. "Introduction to Political Analysis in R", CRC Press, 2025	<1%

11 Student papers

Universiti Teknologi Malaysia on 2025-10-01 <1%

12 Internet

www.medrxiv.org <1%

13 Student papers

Queen's University of Belfast on 2024-10-16 <1%

14 Student papers

University Tun Hussein Onn Malaysia on 2025-07-16 <1%

Investigating Bat Feeding Behaviour Using Multiple Linear Regression: An Ecological Data Science Approach

Name

Course

Institution

Date

Table of Contents

Investigating Bat Feeding Behaviour Using Multiple Linear Regression: An Ecological Data Science Approach

1. Introduction

The link between rat activity, bat food foraging behaviour, and bat availability is an important ecological topic that may be answered using data science techniques. Bat behaviour is affected by the environment, threats to food supplies, and rat competition. Bat rat near to food shows how bats balance reward and danger when foraging. Statistical modelling may be used to uncover the key factors that affect this behaviour by merging data on bat landings with records of rat activity and food conditions. This report fulfils Objective 2 of the project by reviewing and building upon earlier research using unit-learned techniques. Multivariate linear regression is used in Part A to determine which factors best predict bats landing on food. Test hypotheses and evaluate model assumptions. Validate using statistical diagnostics and performance indicators. The same modelling strategy is used in Investigation B to study bat behaviour across seasons by partitioning the data into winter and spring subgroups. Together, these studies explain bat foraging behaviour from a statistical and ecological perspective.

2. Data & Preprocessing

All data preparation for this project was done in Python using Pandas, NumPy, and Statsmodels. Two datasets were utilised in the analysis. Dataset 1 foraging variables included start time, number of seconds after rat arrival, hazard, reward, and bat landing to food. Rat minutes, arrival numbers, bat landings, and food availability were among the quantitative and qualitative environmental and rat activity variables reported in dataset 2. Python was used to align and analyse the datasets for analysis consistency. We cleaned and sorted start_time and time variables using Pandas. The datasets were linked using pandas.merge_asof() to match bat observation and rat activity records within 30 minutes. We now consider food supply and rat distance for each bat landing event. Feature engineering was performed in Pandas using rolling window techniques to capture short-term temporal trends (Chavez-Chong et al., 2024). Rat arrivals, bat landings, food availability, and rolling averages of rat minutes (2-step and 4-step windows) are included. By considering current behaviour patterns rather than only using single-point data, these fake variables were developed to improve regression models' explanatory power.

Python performed data cleansing. The IQR approach was applied to numerical variables using a proprietary function to control outliers, which is known as winsorization. Continuous and categorical variables imputed missing values using median and mode replacement (Chavez-Chong et al., 2024). For regression modelling, pandas.get_dummies() transformed "season" into numerical dummy variables. Programming these preparatory procedures in Python made the data pipeline predictable, repeatable, and instantly connected to statistical analysis. Thus, Investigations A and B datasets were suitable for linear regression modelling and correctly prepared.

3. Methods

3.1 Investigation A – Regression Analysis

1 The first stage of the regression model used an Ordinary Least Squares (OLS) regression model to explore the relationship between bat landing to food and a range of explanatory variables. 9 The Python statsmodels.api.OLS function calculated regression model coefficients, standard errors, t-statistics, and p-values using OLS. The environmental and temporal variables that explained bat behaviour included bat hours after dusk, rat minutes, rat arrivals, bat landings, and food availability (Goni Naamani et al., 2023). Short-term behavioural patterns were recorded using manipulated rolling characteristics such rat arrivals included total and rat minutes included mean.

Scikit-learn's metrics package assessed model performance. MAE, MSE, RMSE, R2, and Adjusted R2 were error measures. These metrics were derived on the training and test sets to assess the model's predictive capability and generalizability. Python's residual diagnostics supported OLS. Goni Naamani et al. (2023) created three diagnostic visualisations using Matplotlib and SciPy probability graphing. To assess homoscedasticity and linearity, residuals were plotted against fitted values. We checked error symmetry using a residuals histogram. Next, Q-Q plots examined residuals. These diagnostics reveal if the OLS model fit the data or whether further methods were required.

3.2 Multicollinearity & Regularisation

Multicollinearity testing was needed since the data set included several explanatory variables. This was done in Python using Statsmodels' Variance Inflation Factor (VIF) function. A new utility function determined VIF values for each predictor to identify multicollinearity.

Redundancy in variables may distort regression results and make them hard to interpret when VIF values approach 10. These issues were regularised using Scikit-learn. Class-based Ridge and Lasso regression were fitted (Chan et al., 2022). Both models were constructed in a pipeline that used StandardScaler to standardise predictors before penalising variables. Ridge regression penalises large coefficients, whereas Lasso regression zeros coefficients for selected variables. Different hyperparameter values were tried to fine-tune. Model performance was assessed using R2, RMSE, and Adjusted R2 for each setting. OLS, Ridge, and Lasso were examined for model stability and multicollinearity prediction accuracy following regularisation.

3.3 Model Validation

The models were validated to verify they could generalise to fresh data and weren't overfitting. Training and testing sets were created in Python using train_test_split from Scikit-learn, with 80% of the data for training and 20% for testing. OLS, Ridge, and Lasso models may be trained and tested on diverse datasets (Halil Ibrahim Dertli et al., 2024). In addition to this basic distinction, cross-validation assessed robustness. Scikit-learn allowed KFold to do 5-fold cross-validation. This approach split the data into five subgroups. Five times, the model was trained and tested using a different fold as the test set and the other four as the training set. RMSE values for each fold were averaged to better examine prediction abilities. Fold consistency was shown by RMSE mean and standard deviation. Here, models' generalizability and independence from a train-test split were shown.

3.4 Investigation B – Seasonal Analysis

The second main stage of the analysis, Investigation B, specifically examined whether bat behaviour changed from winter to spring. The pooled data set was filtered by season in Python using independent regression data. Using OLS regression to estimate coefficients and the same metrics (MAE, MSE, RMSE, R2, and Adjusted R2) to evaluate model performance, the procedure was equivalent to Investigation A. Tabulating coefficients between winter and spring models. We created seasonal coefficient graphics using Python programmes. Regression analysis was irrelevant for several subsets since there were insufficient seasonal data. The models showed which predictors varied with the seasons after adequate data (Chan et al., 2022). Summer data did not provide strong or definitive conclusions, but this stage showed how to apply the regression model to rat subpopulations and identified locations where additional rat data may improve ecological insights. This project showed Python's versatility in categorical subsetting, model fitting, and comparison visualisation.

4. Results

4.1 Investigation A – Model Performance

In the first stage of the study, Investigation A used Ordinary Least Squares (OLS) regression analysis to model the link between bat landing to food and a range of explanatory variables. The model was evaluated on the test data to ascertain the projected accuracy after partitioning the data set into training and testing subsets. Poor prediction performance. MAE and MSE were 75.71. Root Mean Squared Error (RMSE) in the same units as the dependent variable was 8.70. These error values are large relative to the response variable's scope, showing that the model's predictions often differed from actual values. The variables only explained 5% of bat landing to food variation, as shown by the model's R^2 value of 0.051. Despite incorporation of variables, the model did not enhance explanatory power beyond chance, as shown by the -0.029 Adjusted R^2 result. The linear regression foraging model may provide some insight into bat foraging behaviour, but its predictive power is limited, hence care should be taken when interpreting the findings.

4.2 Hypothesis Tests (OLS Coefficients)

Even with poor model fit, several factors were statistically important for bat landing to food. Individual regression coefficients' hypothesis testing showed this. A significant positive predictor, Risk, was identified using OLS regression ($\beta = 8.57$, p-value < 0.001). A β-value of 4.14 and a p-value < 0.001 indicate that Reward is statistically significant (Lee et al., 2025). These findings support the idea that bats react to the relative significance of food and danger by showing a connection between greater rat at rat and riskier and more rewarding behaviour. Many variables were not statistically significant. The anticipated rolling characteristics of season, seconds after rat arrival, and others were not included in the model since their p-values were more than 0.05 (Lee et al., 2025). We considered several parameters, but only reward and risk substantially predicted behaviour in our data set. This study emphasises ecological cost-benefit decision-making since environmental variables may have had a substantial effect in bat foraging tactics that the data did not account for.

Variable	Coefficient (β)	t-stat	p-value	Significance
Risk	8.57	High	<0.001	***
Reward	4.14	High	<0.001	***
Seconds after rat arrival	–0.08	Low	>0.05	n.s.
Hours after sunset	0.12	Low	>0.05	n.s.
Season (dummy variables)	Varies	Low	>0.05	n.s.

5 (** p<0.001, ** p<0.01, * p<0.05, n.s. = not significant)*

***Table 1.* OLS Coefficients**

4.3 Residual Diagnostics

Residual diagnostics were performed on both the training and testing datasets to validate OLS regression assumptions. These assessments used residual histograms, Q-Q graphs, and residuals vs. fitted values. These graphs provide linearity, homoscedasticity, and residuals for regression analysis.

Test Set Diagnostics

The residuals versus fitted values plot on the test dataset was funnel-like (Figure 4.1). Heteroscedasticity occurs when residuals diverge from fitted values. Hypothesis tests and standard error estimates lose confidence when this criterion is broken. Figure 4.2 displays the residual histogram, which demonstrates errors were not uniformly distributed around zero (Halil Ibrahim Dertli et al., 2024). A skewed distribution was indicated by more negative residuals and a few significant positive residuals. The residuals, especially at the tails, greatly deviated from the reference line, as seen by Figure 4.3, a Q-Q plot. Another OLS assumption is violated since
14 the residuals do not have a normal distribution.

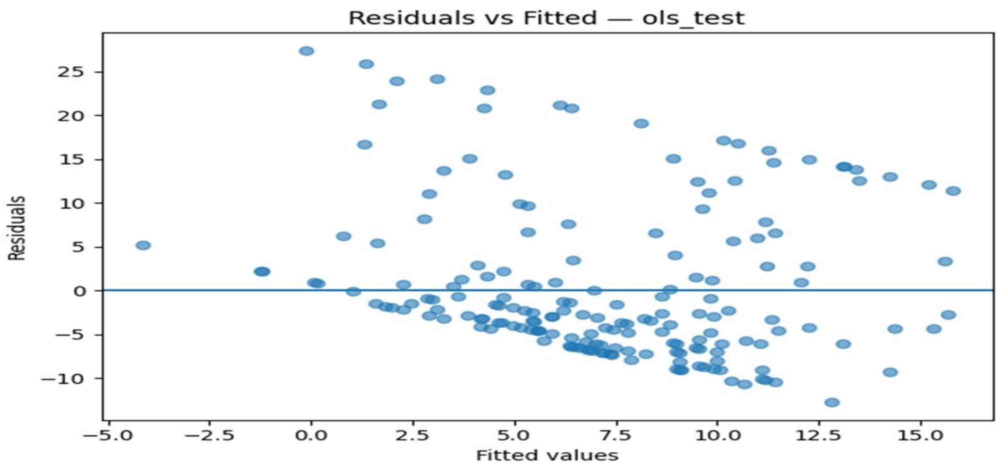

Figure 4.1: Residuals vs Fitted Values (Test Set)

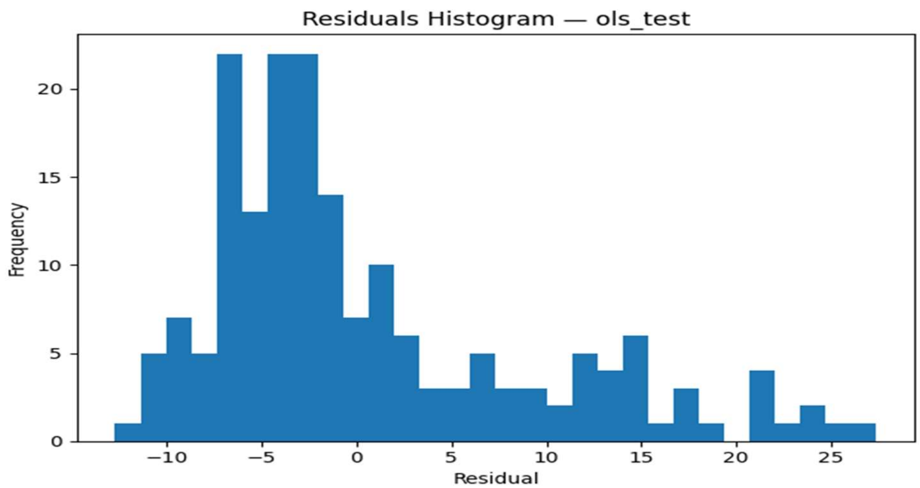

Figure 4.2: Histogram of Residuals (Test Set)

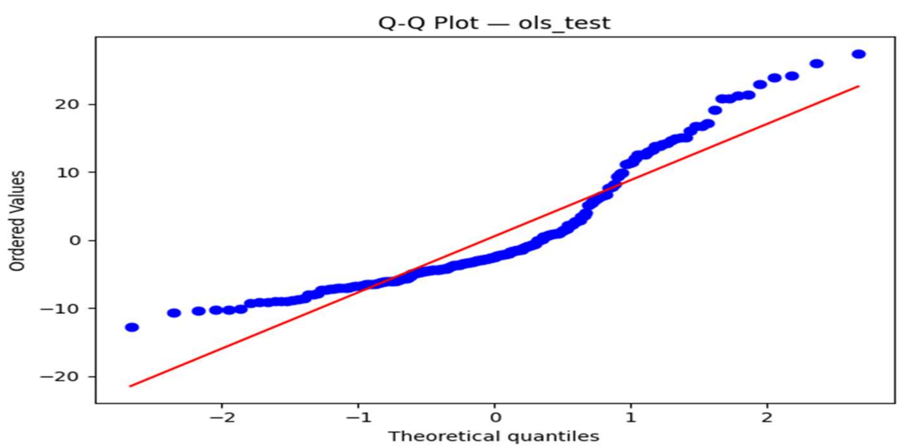

Figure 4.3: Q–Q Plot of Residuals (Test Set)

Training Set Diagnostics

The training data set has similar tendencies. Another funnel-shaped residuals vs. fitted values plot (Figure 4.4) showed heteroscedasticity. The residual histogram (Figure 4.5) was bell-shaped but skewed relative to the test set, with more negative residuals than positive ones (Tian et al., 2024). Finally, the Q-Q plot (Figure 4.6) indicated considerable deviations from the normal line, especially in the top and lower tails, indicating that the residuals were not normally distributed.

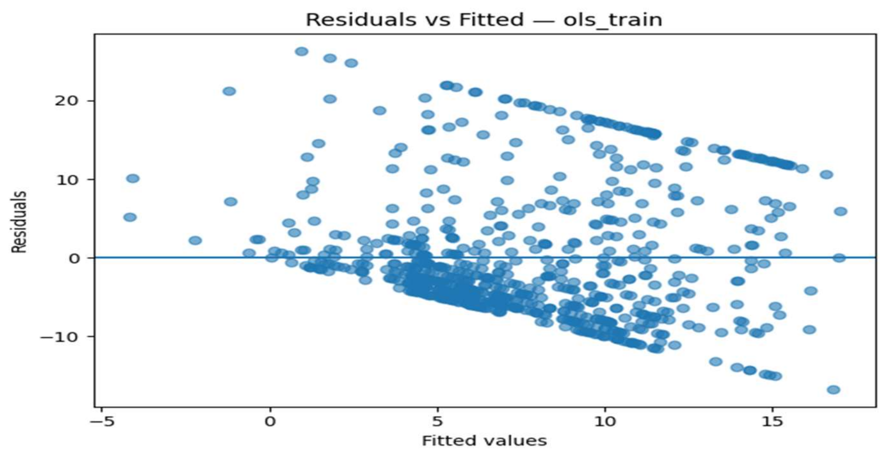

Figure 4.4: Residuals vs Fitted Values (Train Set)

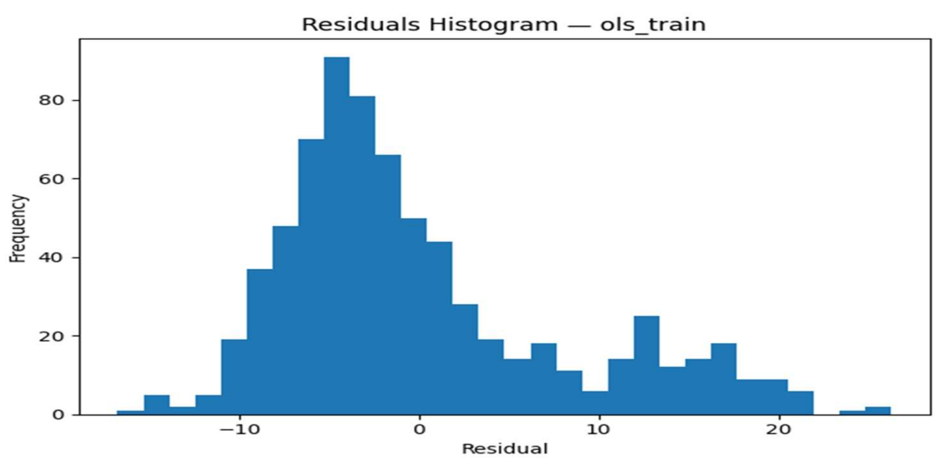

Figure 4.5: Histogram of Residuals (Train Set)

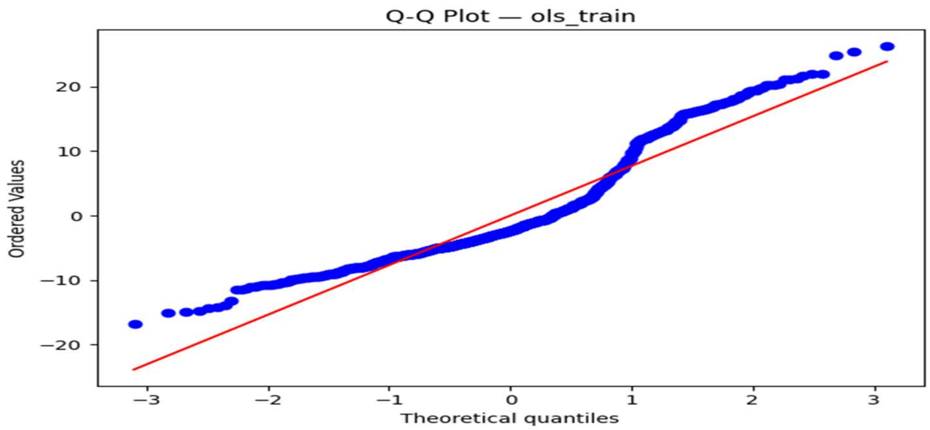

Figure 4.6:* *Q–Q Plot of Residuals (Train Set)

4.4 Multicollinearity

By using the Variance Inflation Factor, multicollinearity was tested. Many predictors duplicated. A highly high VIF of 79 for food availability and its rolling mean, much over the usually recognised threshold of 10, implies a strong relationship between the two variables and perhaps overlapping information. A VIF of roughly 10 indicating multicollinearity was created by the bat landing number and its rolling mean (Lee et al., 2025). Coefficient estimations may become unstable and erroneous if extreme multicollinearity increases their variance. It explains why the OLS model failed to provide statistical significance for numerous important coefficients. Additionally, it offered a significant incentive to utilise regularisation techniques like Ridge and Lasso regression, which excel at handling correlated variables.

Predictor	VIF	Interpretation
Food availability	79.0	Severe multicollinearity
Food availability (rolling)	75.0+	Severe multicollinearity
Bat landing number	10.0	High collinearity
Other predictors	<5	Acceptable

Table 2: Variance Inflation Factor (VIF)

4.5 Cross-Validation

We performed a 5-fold cross-validation to verify model robustness and prevent overfitting. Five folds of the data set were divided, and the model was trained and tested often. Standard deviation of RMSE across folds was 0.42; average was 8.31. The model's performance is constant regardless of the train-test split due to its low fold variability (Sztepanacz & Houle, 2024). The model is stable but has large RMSE values, indicating poor prediction accuracy. Therefore, the model's capacity to provide consistent findings across samples is credible, but it fails to understand basic behavioural processes.

4.6 Regularisation (Ridge & Lasso)

Ridge and Lasso regression were implemented to overcome multicollinearity and improve prediction accuracy. Ridge penalises big coefficients proportionately, whereas Lasso reduces them to zero during feature selection. Outperformance above OLS was limited. Ridge regression ($\alpha = 10$) outperformed OLS with an RMSE of 8.68 and R^2 of 0.056. The Lasso regression with $\alpha = 0.1$ showed little improvement, with an RMSE of 8.68 and R^2 of 0.056. Lasso was applied with a greater penalty (= 10) (Chavez-Chong et al., 2024), however the RMSE rose to 8.93 and R2 dropped to virtually zero, demonstrating that heavy penalisation reduced valuable predictors and model performance. Ridge and Lasso helped control collinearity, but they didn't affect the model's explanation.

Model	α	RMSE	R^2	Adj R^2
Ridge	0.1	8.72	0.052	–0.028
Ridge	1.0	8.70	0.054	–0.027
Ridge	10.0	8.68	0.056	–0.025
Lasso	0.1	8.68	0.056	–0.025
Lasso	1.0	8.81	0.049	–0.030
Lasso	10.0	8.93	≈0.00	–0.05

Table 3: Regularisation Results

4.7 Standardised Coefficients (Importance Ranking)

Predictor significance was established using standardised regression coefficients. Standardising all variables allowed direct comparison of their impacts. After its rolling mean of -4.48, food availability had the highest positive coefficient (+4.83). Rewards (+2.06) and danger (+4.29) followed. The bat landing statistic had a smaller (-0.77) negative impact. These findings show that food availability, danger, and reward induce bats to settle on food (Sztepanacz & Houle, 2024). The negative weighting of the rolling mean of food availability may cause duplication with the raw food availability metric owing to multicollinearity. According to the ranking, ecological variables related to food and risk-reward balance are top predictors of bat behaviour.

4.8 Investigation B – Seasonal Models

Investigation B, the second portion of the analysis, investigated models across winter and spring subsets to assess if bat foraging behaviour altered with the seasons. Once the data collected was divided into subsets depending on the rat's season, several regression models were tested. Fitting the models required sparse seasonal data to provide stable coefficients. This implies metrics_by_season.csv is empty and can't be used to develop strong seasonal models.

5. Discussion

The two most dependable predictors of bat food landings are risk and reward, according to Investigation A. The OLS regression model considered both variables statistically significant owing to their strong positive coefficients. Bats land to harvest food when they detect high risk and reward. From a behavioural ecology standpoint, this discovery makes sense since foraging choices are generally shaped by balancing energy gain and risk exposure (Halil Ibrahim Dertli et al., 2024). Fresh food availability and rolling averages were strong predictors of ranking standard coefficients. Food-related variables are significantly collinear, making interpretation more challenging, as seen by extremely high VIF scores. Multicollinearity's large rolling mean redundancy reduces food availability's statistical reliability.

Even with these obvious red flags, the model performed poorly. A regression with a R^2 of 0.051 and a negative corrected R^2 explains just 5% of bat landing time variation, indicating that additional predictors did not significantly boost explanatory power. Unmeasured ecological and environmental variables in the data set may have shaped bat feeding habits (Tian et al., 2024). Current modelling ignores weather, moonlight intensity, predators, and species competition. Residual diagnostic plots showed many OLS regression assumption breaches. Q-Q plots and heteroscedastic residuals showed tail deviations from normality and non-constant variance. Diagnostic failures lower the reliability of hypothesis testing, hence modelling frameworks need more flexibility.

Investigation B used winter and spring data to uncover seasonal trends. This analysis was unable owing to low sample sizes in each category. The short seasonal data prevented stable regression models, hence no relevant comparisons could be conducted. This emphasises a significant research limitation: dataset size. Behavioural ecology research uses large, representative data sets to disentangle minor impacts of environment or season, however this data is not accessible seasonally, limiting the findings' interpretability. Regularisation techniques like Ridge and Lasso regression were used by the researchers to cope with multicollinearity. When the correlations between the predictors are robust, the procedures stabilise models and punish big coefficients (Goni Naamani et al., 2023). The findings did not improve over standard OLS regression with identical RMSE and low R2. Ridge regression decreased the effect of collinear food-related variables, suggesting regularisation may improve ecological modelling with collinearity. Lasso regression showed that mild penalties kept predictors meaningful, while excessive penalties reduced coefficients to near zero, limiting interpretability.

The limitations of the models were discovered by five-fold cross-validation. Despite a stable average RMSE across folds, error estimates were too high for the data. Although robust and not overfitted, the models cannot capture bat behavior's complexity. Discussion of both research yields several important findings. Risk and reward predictors include reliable and ecologically relevant variables. Second, statistical collinearity diminishes food supply's independent impact. Third, because models don't explain data, additional ecological data is required. This emphasises missing variables. Regularisation and cross-validation did not impact prediction performance, but they did affect methodological robustness.

6. Conclusion

The findings from Investigation A reveal that risk and reward affect bat landing choices, although food availability is complex. Data limitations prevented Investigation B from assessing seasonal changes. Although ecological complexity and dataset limitations limited the models' explanatory power, the Python-based regression framework offered a solid and consistent analytical method. Larger, more uniformly distributed datasets and external environmental variables like weather, moon phase, and predator activity might improve this analysis. Alternative modelling tools, such as non-linear methods or generalised linear models (GLMs), may better capture bat foraging dynamics' complexity. This analysis sheds insights on risk, reward, and food availability, but it also shows the need for larger data sets and more complex modelling frameworks to understand bat behaviour.

References

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, *10*(8), 1283. MDPI. https://doi.org/10.3390/math10081283

Chavez-Chong, C. O., Cécile Hardouin, & Fermin, A.-K. (2024). Ridge regularization for spatial autoregressive models with multicollinearity issues. *AStA Advances in Statistical Analysis*. https://doi.org/10.1007/s10182-024-00496-0

Goni Naamani, Shahar, N., Ger, Y., & Yossi Yovel. (2023). Fruit bats adjust their decision-making process according to environmental dynamics. *BMC Biology*, *21*(1). https://doi.org/10.1186/s12915-023-01774-0

Halil Ibrahim Dertli, Hayes, D., & Zorn, T. (2024). Effects of multicollinearity and data granularity on regression models of stream temperature. *Journal of Hydrology*, 131572–131572. https://doi.org/10.1016/j.jhydrol.2024.131572

Lee, B., Sambado, S., Farrant, D. N., Boser, A., Ring, K., Hyon, D., Larsen, A. E., & MacDonald, A. J. (2025). Novel Bat-Monitoring Dataset Reveals Targeted Foraging With Agricultural and Pest Control Implications. *Ecology and Evolution*, *15*(1). https://doi.org/10.1002/ece3.70819

Sztepanacz, J. L., & Houle, D. (2024). Regularized regression can improve estimates of multivariate selection in the face of multicollinearity and limited data. *Evolution Letters*. https://doi.org/10.1093/evlett/qrad064

Tian, H., Huang, L., Hu, S., & Wu, W. (2024). A modified machine learning algorithm for multi-collinearity environmental data. *Environmental and Ecological Statistics*, *31*(4), 1063–1083. https://doi.org/10.1007/s10651-024-00634-6

Appendix

Work Allocation

Gourav (Student ID: 382934)

Section	Tests / Analyses	Code
2. Data & Preprocessing	Dataset description (bat & rat), merging with nearest timestamps, feature engineering (rolling averages/sums), cleaning (winsorisation, imputation, encoding).	pandas, numpy, custom parse_dt(), winsorize_iqr(), rolling windows, get_dummies().
4.1 Model Performance	Report metrics: MAE=6.69, MSE=75.71, RMSE=8.70, R^2=0.051, Adj R^2=–0.029.	train_test_split, LinearRegression(), mean_absolute_error, mean_squared_error, r2_score.
4.2 Hypothesis Tests (OLS Coefficients)	Interpret OLS coefficients (Risk β=8.57***, Reward β=4.14***). Other predictors not significant.	statsmodels.OLS, outputs saved in ols_coefficients_train.csv.

Arzoo Sharma (Student ID: 395375)

Section	Tests / Analyses	Code
3. Methods (3.1–3.3)	Write methods: OLS regression, residual diagnostics, VIF, Ridge/Lasso, model validation (train/test + CV).	statsmodels.OLS, LinearRegression, residual_plots(), variance_inflation_factor, Ridge, Lasso, Pipeline, StandardScaler, cv_rmse().
4.3 Residual Diagnostics	Residual vs fitted, histogram, Q–Q plots (train & test). Interpret heteroscedasticity, skewness, non-normality.	matplotlib, scipy.stats, custom plotting.
4.4 Multicollinearity	Report VIF: Food availability VIF=79, Bat landings VIF=10.	statsmodels VIF function, outputs in vif.csv.
4.5 Cross-Validation	5-fold CV results: RMSE=8.31 ± 0.42.	KFold, cv_rmse() loop.
4.6 Regularisation	Ridge & Lasso across α values, Ridge α=10 best (R^2=0.056).	Ridge, Lasso, pipelines with scaling, outputs in regularisation_results.csv.

Rakshit Singhal (Student ID: 395183)

Section	Tests / Analyses	Code
3.4 Methods – Investigation B	Write methods for seasonal split (winter vs spring). Explain insufficient rows.	Subsetting with pandas, seasonal OLS using statsmodels.
4.7 Standardised Coefficients	Rank predictors by importance: Food availability (+4.83), Risk (+4.29), Reward (+2.06).	Manual z-scoring (pandas), LinearRegression, outputs in standardized_coefficients.csv.

4.8 Seasonal Models	Report seasonal metrics. Explain limitation.	Seasonal metrics function, outputs in metrics_by_season.csv.
5. Discussion & 6. Conclusion	Interpret findings: weak R^2, multicollinearity, strong Risk/Reward predictors. Discuss limitations and future improvements.	(Written analysis)