Assessing the usefulness of sea temperature and weather data as predictors of lobster landings.

DS 5020 Introduction to Linear Algebra and Probability for Data Science

Emma Morse

November 2022

1 Introduction

The lobster industry has had a significant effect on the economy, tourism and culture of Maine. Based on a 2021 report by the Maine Department of Marine Resources [1], the estimated value of the fishery landings was almost $900 million, with lobster accounting for 82%. As we face a changing climate and adjustments in marine habitats we may find this effects the expected yield from lobster fisheries.

This paper outlines an analysis using multiple linear regression and dimensionality reduction to determine the strength of climate and sea temperature data in predicting annual lobster landings. Z-scores are calculated for our predictors to aid in determining strength of the overall model.

2 Data

We have collected Lobster landing data from the Department of Marine Resources (DMR) [4]. This is the weight, in pounds (lbs) brought to shore to be sold to another party. In essence the weight of species removed from the ocean per year. The landings could be affected by many external factors, for example, climate changes, lobster demand and value, fishing techniques or changes in landings reporting. However for this analysis, we are most interested in the predicting power of a few climate and sea temperature variables on the landings. We collected sea temperature data from the Boothbay Harbor Environmental Monitoring Program, also reported by DMR [3], and extreme weather index data from the National Oceanic and Atmospheric Administration [2]. These data were wrangled to extract the following predictor variables (p=4):

$\mathbf{x}_1$: extreme weather index for the Northeastern United States
$\mathbf{x}_2$: average of the daily maximum sea surface temperatures
$\mathbf{x}_3$: average of the daily minimum sea surface temperatures
$\mathbf{x}_4$: average daily range in sea surface temperatures

Our target variable $\mathbf{y}$ is reported by year so for this analysis we reduced the data to 1989 - 2019 (n=31).

$\mathbf{y}$: pounds of lobster landings by year

These data are from similar sources so therefore could be correlated. Plotting a scatter matrix visualizes bi-variate relationships by plotting the predictors pairwise. We can visually determine whether the predictors are correlated implying dimensionality reduction with principal component analysis could be useful.

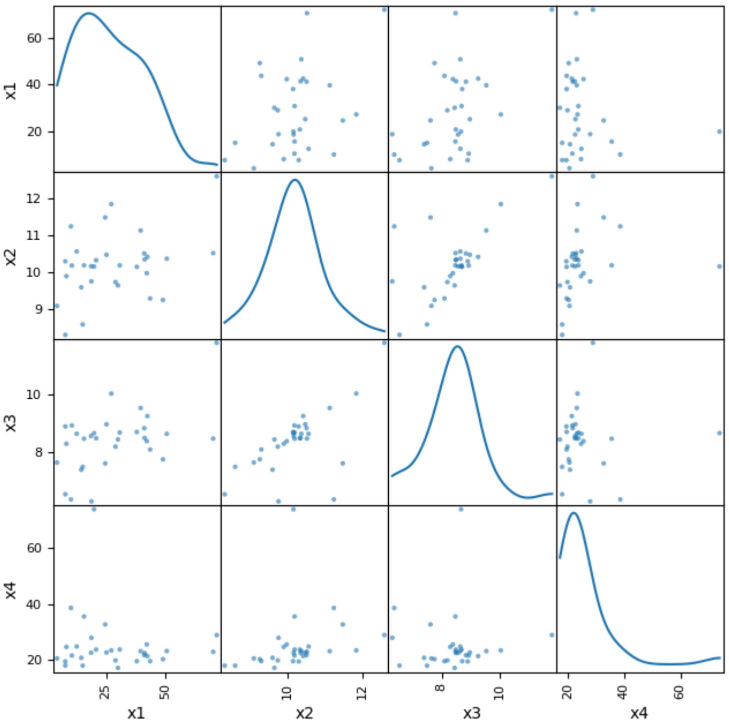

Figure 1: Scatter Matrix for the 4 predictor variables using Kernel density estimation (KDE) for the diagonal plots.

We can see from Figure 1 that our predictors are correlated, for example, x2 and x3. Therefore dimensionality reduction will be necessary to determine the usefulness of the predictors.

3 Methods

The first step is to perform a multiple linear regression using the 4 predictor variables. This follows:

$$\mathbf{y} = X\beta + \epsilon \tag{1}$$

Equation 1 for Multiple Linear Regression.

Where $\mathbf{y}$ is the target vector, X is the design matrix consisting of the predictor vectors and a vector of ones, β is a vector of model parameters and ϵ is a vector of residuals:

$$\mathbf{y} = \begin{bmatrix} 23.368719 \\ 28.068238 \\ \vdots \\ 102.220639 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 14.60 & \cdots & 20.800 \\ 1 & 43.68 & \cdots & 19.600 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 41.36 & \cdots & 22.499 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_4 \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_{30} \end{bmatrix}$$

In order to make predictions using our design matrix, we will need to find parameter estimates that best combine our predictors by minimizing the residuals:

$$\epsilon = \mathbf{y} - X\beta = \mathbf{y} - (\beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4) \tag{2}$$

Equation 2. the residual vector in terms of the target and our predictor variables.

3.1 Least Squares Optimization for finding parameter estimates

The least squares method finds parameter estimates $\hat{\beta}$ that minimize the residuals squared $||\epsilon||^2$. We can use Equation 2 to find $||\epsilon||^2$.

$$\begin{aligned}
\epsilon &= \mathbf{y} - X\beta \\
||\epsilon||^2 &= (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) \\
||\epsilon||^2 &= (\mathbf{y}^T - \beta^T X^T)(\mathbf{y} - X\beta) \\
||\epsilon||^2 &= \mathbf{y}^T\mathbf{y} - \mathbf{y}\beta^T X^T - \mathbf{y}^T X\beta + \beta^T X^T X\beta \\
||\epsilon||^2 &= \mathbf{y}^T\mathbf{y} - 2X^T\beta^T\mathbf{y} + \beta^T X^T X\beta
\end{aligned} \tag{3}$$

To minimize Equation 3 we can find $\nabla||\epsilon||^2$ in terms of β and set it to zero. This allows us to solve for the parameter estimates $\hat{\beta}$:

$$\nabla||\epsilon||^2 = -2X^T\mathbf{y} + 2X^T X\beta \tag{4}$$

$$0 = -2X^T\mathbf{y} + 2X^T X\beta \tag{5}$$

We can confirm that this $\hat{\beta}$ minimizes $||\epsilon||^2$ by checking if the Hessian is positive definite. $X^T X$ is always positive semi definite, and verifying from the design matrix, $X^T X$ has only positive non zero eigenvalues. From Equation 4 we get:

$$\nabla^2||\epsilon||^2 = 2X^T X$$

The $\hat{\beta}$ that solves Equation 5 can be found by re-arranging the equation:

$$\hat{\beta} = (X^T X)^{-1}X^T\mathbf{y} \approx \begin{bmatrix} -51.69530843 \\ 0.62361141 \\ 15.08815022 \\ -6.38130378 \\ 0.21430962 \end{bmatrix} \tag{6}$$

Now that we have determined the parameter estimates that minimizes the residuals, we can make a preliminary prediction of lobster landings following:

$$\hat{y} = X\hat{\beta} = -51.69530843 + 0.62361141 \cdot \mathbf{x}_1 + 15.08815022 \cdot \mathbf{x}_2 - 6.38130378 \cdot \mathbf{x}_3 + 0.21430962 \cdot \mathbf{x}_4$$

3.2 More on the Residuals Vector

Plotting the residuals against years provides an interesting insight, see Figure 2. We can visually see that after around 2004 the residuals from the model abruptly increase. Interestingly in 2004 the Department of Marine Resources made lobster landing reporting mandatory for all Maine dealers buying directly from harvesters [4]. The 2 lines of best fit have been added for before and after this change was made. What this can tell us is that the regression model is not capturing this change in the target sufficiently.

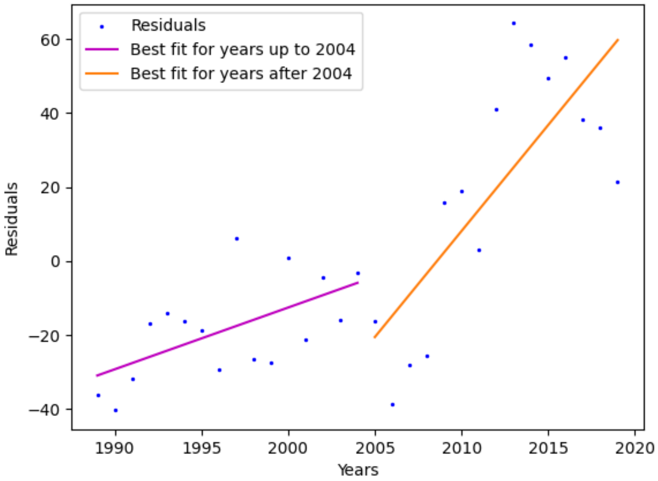

Figure 2: Multiple Linear Regression Model residuals plotted over time

Note that the residuals vector $\epsilon_{n\times 1}$ can be considered a random vector in R^n following a multivariate Normal Distribution $\sim$ Normal(μ, $\sigma^2 I$). Where $\mu = 0$ and $\sigma^2 > 0$. We can estimate the σ^2 as $\hat{\sigma^2}$ using the residuals as a population calculated from the prediction:

$$\epsilon = y - \hat{y}$$

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}(\epsilon - \mu)^2$$

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}(\epsilon \cdot \epsilon) \approx 1193.9167$$

$\hat{\sigma}^2$ is the estimated **unexplained variability** in the target as this is the estimated variability of the residuals from the model.

3.3 Predictor Strength

Since our predictor vectors were not all on the same scale we will need to run additional statistical tests to help determine usefulness and strength. We cannot simply compare the values of $\hat{\beta}_1$,...,$\hat{\beta}_p$. We can calculate z-scores which will standardize the $\hat{\beta}_j$ for $j = 1, 2, 3, 4$ following:

$$z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{V[\hat{\beta}_j]}}$$

Noting that this is comparing the $\hat{\beta}_j$ to the β_j from the null model. The null model does not contain any of the predictors, so there $\beta_j = 0$. $V[\hat{\beta}_j]$ is still unknown but we can estimate it using conditional expectation and variance:

Consider:

$$V[\hat{\beta}|X] = V[(X^TX)^{-1}X^T\mathbf{y}|X]$$
$$V[\hat{\beta}|X] = ((X^TX)^{-1}X^T)V[\mathbf{y}|X]((X^TX)^{-1}X^T)^T$$
$$\text{Where } V[\mathbf{y}|X] = \sigma^2 I \text{ as the variance in } \mathbf{y} \text{ is inherited from the variance in } \epsilon$$
$$\text{Therefore } V[\hat{\beta}|X] = ((X^TX)^{-1}X^T)\sigma^2 I((X^TX)^{-1}X^T)^T$$
$$= \sigma^2((X^TX)^{-1}X^T)(X(X^TX)^{-1})$$
$$= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}$$
$$= \sigma^2 I(X^TX)^{-1}$$
$$= \sigma^2(X^TX)^{-1} \approx \hat{\sigma}^2(X^TX)^{-1}$$

$\hat{\sigma}^2(X^TX)^{-1}$ is the covariance matrix for the parameter estimates $\hat{\beta}$. We only need $\hat{\beta}_j$ variance for calculating each individual z-score. The diagonal elements of the covariance matrix contain these variances, σ_j^2, as a covariance matrix is as follows:

$$\sigma^2(X^TX)^{-1} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2p}\sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1}\sigma_p\sigma_1 & \rho_{p2}\sigma_p\sigma_1 & \cdots & \sigma_p^2 \end{bmatrix}$$

We can simplify our z-score calculation to:

$$z_j = \frac{\hat{\beta}_j}{\sigma_j}$$

Results from computing z-scores for our predictors are as follows:

z_1	z_2	z_3	z_4
1.50577554	1.52128926	-0.76093487	0.33168843

From the z-scores we can say that $\mathbf{x}_1$, the extreme weather index for the Northeastern United States, and $\mathbf{x}_2$, the average of the daily maximum sea surface temperatures, are the most useful predictors in the model (given the presence of the other 2 predictors). We can find an estimate of the explained variance in the target variable by finding the ratio of the estimated variance of the prediction and the actual targets variance:

$$\sigma_y^2 = \frac{1}{n-1}(\mathbf{y} - \overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y}) \approx 1226.9104$$
$$\text{Remember: } \hat{\sigma^2} \approx 1193.9167$$
$$R^2 = 1 - \frac{(n-(p-1))\hat{\sigma^2}}{(n-1)\sigma_y^2} \approx 0.189$$

This tells us that our predictor variables explained 18.9% of the variability in the target variable.

3.4 Principal Component Analysis

Performing Principal Component Analysis using singular value decomposition (SVD) will identify a reduced number of predictor vectors that contain most of the information from our design matrix X. We must start with centered data and find the SVD, see Equation 7:

$$\tilde{X} = [\mathbf{x}_1 - \overline{x}_1 \cdots \mathbf{x}_p - \overline{x}_p]$$
$$\tilde{\mathbf{y}} = \mathbf{y} - \overline{y}$$
$$\tilde{X} = U\Sigma V^T \tag{7}$$

where $U =$ is a matrix containing the eigenvectors of $\tilde{X}\tilde{X}^T$ accounting for the column space of $\tilde{X}$

where $V =$ is a matrix containing the eigenvectors of $\tilde{X}^T\tilde{X}$ accounting for the row space of $\tilde{X}$

where $\Sigma =$ diagonal matrix containing the singular values σ of $\tilde{X}\tilde{X}^T$and $\tilde{X}^T\tilde{X}$

$$\text{So then: } \tilde{X} = U\Sigma V^T = \sum_{i=1}^{p} \mathbf{u}_i \sigma_i \mathbf{v}_i^T$$

From this decomposition we can find our principal components P: containing principal component vectors ordered from the direction of most to least variance in the column space. Meaning that the first principal component tells us the combination of predictor variables that contribute to the most variance. Plotting the singular values in Σ as a Scree plot will help in deciding how many principal components to keep, see Figure 3.

$$P = \tilde{X}V = U\Sigma(V^T V) = U\Sigma$$

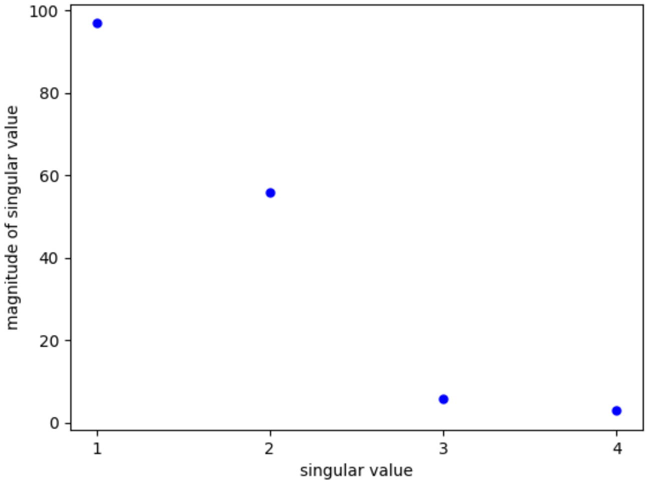

Figure 3: Scree plot of Singular values from Singular value decomposition of the centered design matrix

P can be considered the new predictor matrix, however the lower order principal components are removed as they do not contribute much additional information. The Scree plot from Figure 3 shows that the first 2 principal components contain most of the information. We can reduce P to just the first 2 principal components and perform a new multiple linear regression model using the least squared optimization:

$$\mathbf{y} = P\gamma + \epsilon$$
$$\mathbf{y} = p_1\gamma_1 + p_2\gamma_2 + \epsilon$$

$$\hat{\gamma} = (P^T P)^{-1} P^T \mathbf{y} \approx \begin{bmatrix} 0.67602481 \\ -0.62588626 \end{bmatrix} \tag{8}$$

Interestingly, the reduced model explains 11.7% of the variability in the target. This reduction in the explained variability is surprisingly large, with the reduced predictor matrix accounting for ≈ 62% of the full predictor matrix's result. However, there is an improvement in z-scores:

z_1	z_2
1.92293987	-1.02672395

4 Conclusion

The set of predictor variables do not sufficiently predict annual lobster landings. The overall model was able to explain 18.9% of the variability in lobster landings. The extreme weather index for the Northeastern United States, and the average of the daily maximum sea surface temperatures where the strongest of our original predictor set. Performing a dimensionality reduction did not capture most of the original predictor's information as it significantly decreased the explained variability in the target, however it did provide improved z-scores. There is some promise in the results however, better capturing the change in lobster landing reporting could provide improved results from the model.

5 Appendix

Please see https://github.com/emgrotto/maine-lobster-mlr for the source code for this project. This project was inspired by an analysis of the climate effects on rice yield in Nigeria [5].

6 References

[1] Department of Marine Resources, "Most Recent Maine Commercial Landings". [Online]. Available: https://www.maine.gov/dmr/fisheries/commercial/landings-data

[2] National Oceanic and Atmospheric Administration, National Centers for Environmental Information, "U.S. Climate Extremes Index (CEI)". [Online]. Available: https://www.ncei.noaa.gov/access/monitoring/cei/

[3] Department of Marine Resources, "DMR Boothbay Harbor Environmental Monitoringhttps://www.overleaf.com/project/6 Program". [Online]. Available: https://www.maine.gov/dmr/science/weather-tides/boothbay-harbor-environmental-data

[4] Department of Marine Resources, "Historical Maine Fisheries Landings Data". [Online]. Available: https://www.maine.gov/dmr/fisheries/commercial/landings-data/historical-data

[5] P. G. Oguntunde, G. Lischeid, and O. Dietrich, "Relationship between rice yield and climate variables in southwest Nigeria using multiple linear regression and support vector machine analysis," International journal of biometeorology, vol. 62, no. 3, pp. 459–469, 2018, doi: 10.1007/s00484-017-1454-6

7 Questions

How do you get the symbol for Real numbers?
Is there a use in calculating z scores for PCA?

What is happening in PCA:

with 1 component
explained variability using pstar principal comps
0.08333040318542806

with 2 components
explained variability using pstar principal comps
0.11658960789781525
diff = 0.03325920471

with 3 components
explained variability using pstar principal comps
0.13041885232462413
diff = 0.01382924442

with 4 components
explained variability using pstar principal comps
0.18907638111384661
diff = 0.05865752878

What does it mean if removing one of the predictors significantly improves the results?

$$
\begin{gathered}
E[\mathbf{y}|X] = E[(X\beta + \epsilon)|X] \\
E[\mathbf{y}|X] = E[X\beta|X] + E[\epsilon|X] \\
E[\mathbf{y}|X] = E[X\beta|X] + 0 \\
E[\mathbf{y}|X] = X\beta
\end{gathered}
$$