Statistical Inference and Prediction Testing to Investigate Relationships in Skincare Data

Lauren Mizner
UCLA - CS249
UCID# 005225768
lmizner@g.ucla.edu

Abstract

Intuition would say that if a product costs more, it should be of higher quality, and vice versa. Through statistical inference models and prediction testing, this study explores that relationship among skincare products and further predicts product price points given the various attributes of said products to find that our intuition may not be correct.

1. Introduction

The purpose of this study is to explore the relationship between price and the various feature attributes of skincare products. This is accomplished through statistical inference by applying multiple linear regression models to better understand the basis of these relationships. From there, the study goes further by developing multiple regression models to predict the price point of a given skincare product based on various characteristics of said product, such as product type, brand, skincare ingredients, skin type, and, most importantly, rating of the product by consumers.

2. Data

For this study, the dataset has been sourced from Kaggle [1]. The dataset consists of 1,472 products featured on Sephora's e-commerce website along with information on the product type, brand, price, rating, skincare ingredients, and skin type that the product is intended for.

Before conducting any analysis or building any models for this study, the data must be cleaned and prepared. First, any missing data is accounted for via mean imputation for numerical variables and mode imputation for categorical variables. Mean imputation is used to replace any missing values with the average value of the observed data for that particular variable. On the other hand, mode imputation is used to replace any missing values with the most frequently occurring value for that particular variable. For this study, no values were missing from the dataset in use.

For the purpose of the linear regression study, outliers were addressed using calculated quartiles and interquartile range (IQR) to identify outliers based on the upper and lower bounds, and then remove them from the dataset. This approach was applied specifically to the price data for each product in the dataset, as there were a few skincare products priced at extreme levels, skewing the data relationships. This process of addressing outliers was also applied to the rating values, particularly to filter out the data on the lower end of the spectrum, as there were some products without any reviews, giving them a default rating of zero. Another option for addressing the outliers in the rating's data could have been to use the mean imputation to replace all zero values with the average rating value, but that was not conducted in this study.

For the purpose of the K-Nearest Neighbors and the Ridge Regression models, all categorical features were converted to 0-1 indicator variables in order to be viable for analysis. This approach was applied specifically to the skincare product type and brand attributes.

In preparing the data for regression models, the data needed to be split into two sets: one for training and one for testing. The dataset was originally sorted by product type, so to ensure that the training is completed on a variety of product types, the data was first shuffled. The data was then split with 80% of the data assigned to a training set and 20% assigned to a testing set.

3. Methods

As previously mentioned, the purpose of this study is to explore the relationships among the various data attributes through statistical inference and prediction testing. For statistical inference, linear regression and hypothesis testing were conducted to determine the relationships between different numerical features. Prediction testing was conducted through a KNN regression model, as well as a Ridge Regression model to predict product price points based on their corresponding features.

3.1. Linear Regression

Linear regression is used to describe the relationship between characteristics of a data source. More specifically, linear regression describes the association between data features and the average response variable in the data. For this study, the interest is in understanding the relationship between the price of a skincare product and the various features of the skincare, such as rating, on a scale of 1 to 5 as reviewed by the consumers.

This can be achieved through plotting the price versus the rating in a scatter plot and then fitting a linear regression model to the data. From the fitted model, useful data results can be determined, such as the R-squared value and coefficients. The R-squared value ranges from 0 to 1 and is used to indicate how well the model fits the data in the question, with a higher value being representative of a better fit. The coefficients are used to explain the relationship between the two data features. For this study, the coefficient will provide the amount the average price is expected to increase or decrease given a 1 unit increase in rating.

It is important to take into account that simple linear regression models do have a sensitivity to outliers in the data, which can affect the overall results. This was addressed in the data cleaning and data exploration process of the study to assist in optimizing the results of the linear regression fitted model and provide accurate relationship coefficients.

3.2. K-Nearest Neighbors

K-Nearest Neighbors (KNN) regression is a method used for predicting a continuous value of an outcome variable based on the values of its nearest neighbors in the feature space. In the case of this study, the regression model is being used to predict the price point of a given skincare product based on the values of its nearest neighbors.

When building a KNN regression model, there are a few techniques that can be completed in order to ensure optimization of the model. These include optimizing the k value or number of nearest neighbors and utilizing k-folds cross validation.

First, the optimal number of k nearest neighbors is calculated before conducting the KNN regression model. The value of k will inform the model of how many neighbors it needs to consider the target values of when predicting the value of the new data point, typically through calculating a weighted average of the neighbor values. The choice of the value of k is critical to the model, as too small of a k value will lead to more noisy predictions, but too large of a k value will lead to smoother, but possibly biased predictions. This parameter is key in optimizing the regression model.

Next, k-fold cross validation is utilized to randomly split the data into training and test sets for the purpose of fitting the model to the training data. The data is randomly split into k parts, in this case 10. From there, the model will be trained on the k parts and the error will be computed for each part. The prediction error is then estimated by averaging the error over all k parts.

After the KNN regression model has been optimized, trained, and tested, the results are evaluated using the Root Mean Squared Error (RMSE). The RMSE provides a measure of the average deviation between the predicted values and the actual observed values in the dataset. In general, a lower RMSE value indicates that the model's predictions are closer to the actual values in the dataset, therefore giving a higher accuracy prediction. In the case of comparing the training and testing results via RMSE, the dataset with the lower RMSE value is considered to be more accurate in its predictions.

The RMSE evaluation metric is sensitive to outliers, and large errors for an individual data point can significantly inflate the RMSE value for the model. Any potential outlier values were addressed in the data cleaning and data exploration process of the study to assist in optimizing the results of the KNN regression model and provide accurate RMSE values.

3.3. Ridge Regression

Ridge regression is another method used for predicting a continuous outcome variable based on the features of the input data. Ridge regression utilizes a regularization technique that adds a penalty term to the ordinary least squares (OLS) function to control the complexity of the model and mitigate overfitting. In this case, the model is designed to predict the price associated with skincare products based on their input features, such as brand, consumer rating, ingredients, and skin type the product is intended for.

The regularization parameter, lambda, must be optimized for the Ridge Regression model to produce the best results. The lambda value controls the strength of the penalty term on coefficient values, to shrink the coefficients. The higher the lambda values, the more regularization and, therefore, the more shrinkage of coefficients in the model. By finding the optimal lambda value, we can balance the trade-off between bias and variance to produce a more accurate prediction model.

The optimal regularization parameter, or lambda, can be chosen based on the performance obtained during cross validation. In this case, k-fold cross validation is utilized in the Ridge regression model to randomly split the data into training and test sets for the purpose of fitting the model to the training data. The data is randomly split into k parts, in this case 10. From there, the model will be trained on the k parts and the error will be computed for each part. The prediction error is then estimated by averaging the error over all k parts. Cross-validation is a crucial step within ridge regression as it assists in ensuring the model generalizes well to unseen data, as well as mitigating potential overfitting.

4. Results

Based on the methods previously described, the following gives details of the results found through the implementations of Linear Regression, K-Nearest Neighbors Regression, and Ridge Regression, along with the evaluation metrics used to measure the success or failure of each of their performance.

4.1. Linear Regression

The key relationship this study is investigating is the correlation between product price and the various features of the skincare products. Linear regression allows us to take a look at the direct correlation between numerical data in the dataset, specifically between product price and overall customer satisfaction rating. In conducting a simple linear regression analysis, the relationship between these variables has been identified for both the full data set, as well as the data sets established for each product type.

A summary of the coefficients for each dataset can be seen in Table 1, and the figures associated with each dataset can be found in the Appendix in Figures 1 through 7. The full dataset showed that, if all other features remain the same, there is not much of a correlation or relationship established across all skincare products. When breaking down the data set by skincare product type, the various relationships can be seen to develop.

Table 1: Summary of Linear Regression Coefficients

Dataset	Coefficient
Full Dataset	0.59
Cleanser	-12.92
Moisturizer	-14.78
Treatment	19.12
Face Mask	9.95
Eye Cream	-12.33
Sun Protection	-5.05

Depending on the product type, a significant increase or decrease in cost can be seen for each unit of consumer rating. The largest increase in cost is seen with skincare treatments, where there is an increase of $19.12 for every 1 unit increase in consumer rating, implying that to obtain a better quality skincare treatment product, one would need to pay the extra money. On the other hand, the largest decrease in cost is seen with moisturizers, where there is a decrease of $14.78 for every 1 unit increase in consumer rating, implying that to obtain a better quality skincare moisturizer, it's not necessary to put forth the extra money.

As discussed in the methods section, the R-squared value is used as an evaluation metric to indicate how well the linear regression model fits the data. The R-squared value associated with each model can be seen summarized in Table 2.

Table 2: Summary of R-Squared Values

Dataset	R-Squared
Full Dataset	0.000
Cleanser	0.015
Moisturizer	0.006
Treatment	0.015
Face Mask	0.009
Eye Cream	0.015
Sun Protection	0.004

The R-squared value ranges from 0 to 1 and is used to indicate how well the model fits the data in the question, with a higher value being representative of a better fit. As seen in Table 2, the best performing models are for cleanser, treatment, and eye cream data and only obtain an R-Square value of 0.015. The worst performing model is for the full dataset and obtains an R-Squared value of 0. Unfortunately, these R-Squared values indicate that fitted models do not explain the data variance very well and the overall model fit is poor.

4.2. K-Nearest Neighbors

For the KNN regression model, the model was trained for predicting the price of a given skincare product based on the following data features: product type, consumer ranking, and the skin type that the product is intended for.

First, the optimal number of k nearest neighbors is calculated before conducting the KNN regression model. The value of k will inform the model of how many neighbors it needs to consider the target values of when predicting the value of the new data point. The optimal k value is found by comparing the cross validation scores across k values from 1 to 50. The k value associated with the smallest RMSE score is the optimal k value for the model.

For this study the optimal k value was found to be 19 nearest neighbors. The model was then trained and tested utilizing k-fold cross validation with 10 folds and k set to the optimal value of 19. The final RMSE value for the model was found by averaging the data found over all 10 folds. The model was then applied to the test data to retrieve that RMSE value.

The RMSE value is the evaluation metric used for the performance of the KNN regression model. A lower RMSE value indicates a better regression model. The results for each k-fold, as well as the final average calculated RMSE values are summarized in Table 3.

Table 3: KNN Regression Model Train and Test RMSE

K-Fold	Train RMSE	Test RMSE
1	42.32	34.88
2	39.66	59.12
3	41.81	42.33
4	40.50	52.43
5	41.45	45.21
6	41.54	43.59
7	42.23	35.75
8	42.33	33.68
9	40.83	47.52
10	42.24	35.35
Average	41.70	46.38

If the testing RMSE is significantly higher than the training RMSE, then it would suggest that the model may be overfitting to the training data. This means that the model tends to capture noise and patterns specific to the training data, leading to poor generalization when performing on unseen data.

Based on the values in Table 3, the model performed better on the training data, as the training RMSE is 41.70, which is closer to zero than the testing RMSE of 46.38. The testing RMSE is a little higher than the training RMSE, suggesting there may be some slight overfitting which could be better fine tuned through hyperparameter selection. Overall, the training and testing RMSE values are relatively close in value indicating that the test model performed fairly well given the training.

4.3. Ridge Regression

For the Ridge regression model, the model was trained for predicting the price of a given skincare product based on the following data features: product type, consumer ranking, and the skin type that the product is intended for.

First, the optimal lambda value must be calculated before conducting the Ridge regression model. The value of lambda will control the strength of the penalty term on the coefficient values. The optimal lambda value is found by comparing the cross validation scores across 1000 different lambda values between 0 and 10. The lambda value associated with the smallest RMSE score is the optimal lambda value for this model.

For this study, the optimal lambda value was found to be 2.70. The model was then trained and tested utilizing k-fold cross validation with 10 folds and lambda set to 2.70. The final RMSE value for the model was found by averaging the data found over all 10 folds. The model was then applied to the test data to retrieve that RMSE value.

The RMSE value is the evaluation metric used for the performance of the Ridge Regression model. A lower RMSE value indicates a better regression model. The result for each k-fold, as well as the final average calculated RMSE values are summarized in Table 4.

Table 4: Ridge Regression Model Train and Test RMSE

K-Fold	Train RMSE	Test RMSE
1	43.34	38.29
2	40.69	59.15
3	43.25	39.64
4	41.80	52.04
5	42.56	45.56
6	42.95	42.17
7	43.56	36.01
8	43.65	35.00
9	42.48	46.43
10	43.57	35.66
Average	42.86	45.45

Again, if the testing RMSE is significantly higher than the training RMSE, then it would suggest that the model may be overfitting to the training data. This means that the model tends to capture noise and patterns specific to the training data, leading to poor generalization when performing on unseen data.

Based on the values in Table 4, the model performed better on the training data, as the training RMSE is 42.86, which is closer to zero than the testing RMSE of 45.45. The testing RMSE is a little higher than the training RMSE, suggesting there may be some slight overfitting which could be better fine tuned through hyperparameter selection. Overall, the training and testing RMSE values are relatively close in value indicating that the test model performed fairly well given the training.

5. Conclusion

Based on the results of each model and the evaluation metrics associated with them, the following conclusions on this study have been made.

5.1. Linear Regression

The results from the simple linear regression model left much to be desired. Although the fitted models did produce some interesting relationship information between product price and consumer rating when the data was broken down by product type, the R-Squared values indicated that the overall model fit was poor and not an accurate representation of the variance in the models.

One aspect that may improve this study, would be pulling data on the number of reviews associated with each product rating. This way, a weighted value can be used to better validate the consumer ratings. Currently, a 4.5 rating could be based on 5 reviews of 5000 reviews. It would make sense to assign a rating of 4.5 based on 5000 reviews a greater weight than that of 5 reviews.

It would also be interesting to see if more data, or better fine tuned data, would show potential for a stronger performing linear regression model and, therefore, a more accurate relationship between the product price and consumer rating. Another idea would be to explore other data attributes to see which features have the strongest relationships. For example, it may be worthwhile to break down products by ingredients and see if certain ingredients are associated with a higher price point for skincare.

5.2. KNN Regression and Ridge Regression

The results from the KNN Regression model and the Ridge Regression model are almost identical, as seen in their summary tables, Table 3 and Table 4, respectively. Based on the results, the Ridge Regression model performs slightly better as the RMSE of the testing data came in at 45.45, while the RMSE of the KNN regression testing data came in at 46.38.

Based on the results, both models perform fairly well as the RMSE test values are only slightly higher than their equivalent training RMSE values. It would be interesting to see if this performance could be further improved upon given more skincare attributes to base the testing on.

5.3. Future Studies

Although the main interest in this study was to investigate the relationships between different skincare attributes and the price point of said products, more questions have been raised than answered. While intuition showed that the price may be correlated with the overall quality of the product, gathered through consumer ratings, that may not be the case. Additional studies would need to be conducted to further explore other relationships in the data, along with their effectiveness in prediction testing.

In terms of linear regression models, it may be worthwhile to consider other relationships outside of product price and consumer ratings. Other options would be to explore the relationships between skincare ingredients and consumer ratings or skincare ingredients and product price.

Another consideration would be to shift the focus of the study from product price to product quality. Instead of investigating the relationships and prediction testing tied to the price of a product, it could be tied to consumer rating and what makes a product a quality product in the eyes of the consumer.

By converting the problem from one of regression prediction to one of classification prediction, the technique of feature selection could be utilized to better understand which attributes contribute the most in our model. This could be done through converting the consumer ratings into a categorical variable. For example, a rating in the range of 0 to 1 is bad, a rating of 1 to 2 is poort, a rating of 2 to 3 is fair, a rating of 3 to 4 is good, and a rating of 4 to 5 is excellent. From there we could use feature selection to analyze which features produce the highest marginal gain and see which features are the most prominent in determining the quality or consumer rating level of a product.

Through a categorical model, such as this, it would be interesting to see if the features with the highest gain differ depending on the rating level that's being predicted and classified with the model. There may be a different set of features that have a higher marginal gain when classifying a product that is of poor quality versus a product that is of excellent quality. This would show which attributes potentially contribute more to an excellent product versus which attributes contribute more to a bad product, which, in turn, could lead to further exploration of the relationships between those attributes and the consumer rating.

References

[1] Awan, Abid Ali. “Cosmetics Datasets.” *Kaggle*, 16 Dec. 2020, www.kaggle.com/datasets/kingabzpro/cosmetics-datasets.

[2] Shirani-Mehr, Houshmand. “Lecture 6: Linear Regression.” CS249: Data Science Fundamentals, University of California, Los Angeles.

[3] Shirani-Mehr, Houshmand. “Lecture 7: Prediction I.” CS249: Data Science Fundamentals, University of California, Los Angeles.

[4] Shirani-Mehr, Houshmand. “Lecture 8: Prediction II.” CS249: Data Science Fundamentals, University of California, Los Angeles.

Appendix

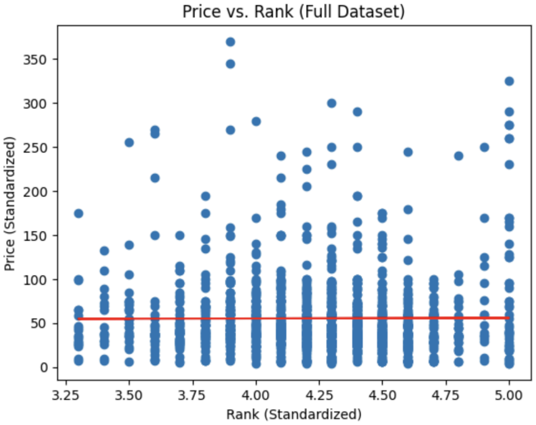

Figure 1: Linear Regression on Full Dataset

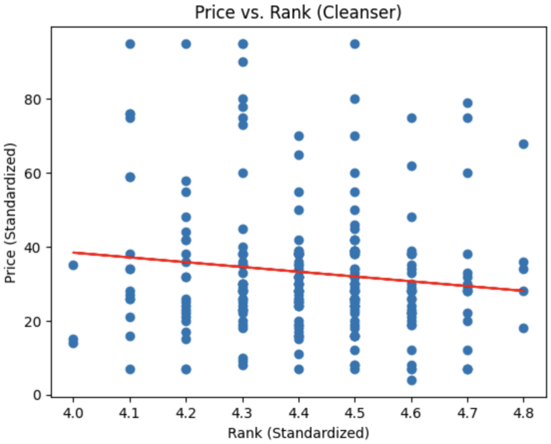

Figure 2: Linear Regression on Cleanser Data

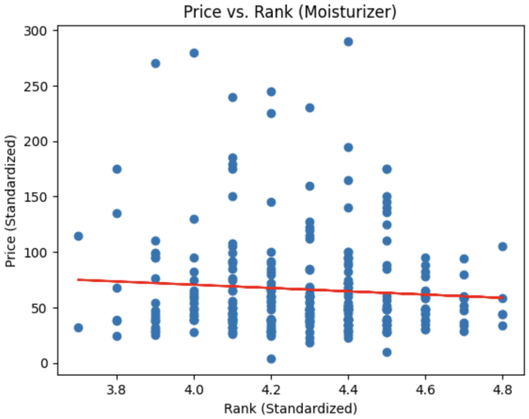

Figure 3: Linear Regression on Moisturizer Data

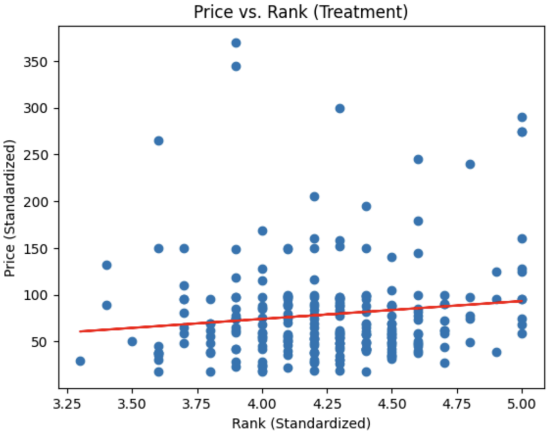

Figure 4: Linear Regression on Treatment Data

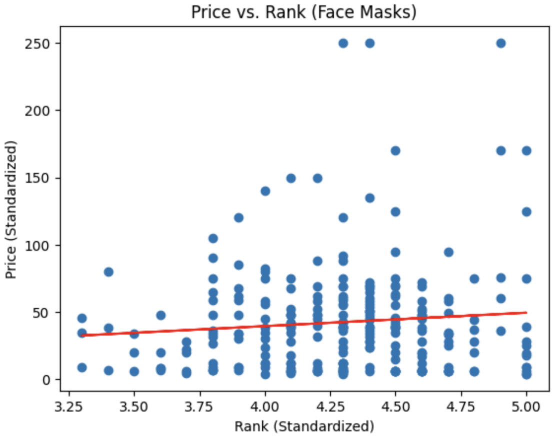

Figure 5: Linear Regression on Face Mask Data

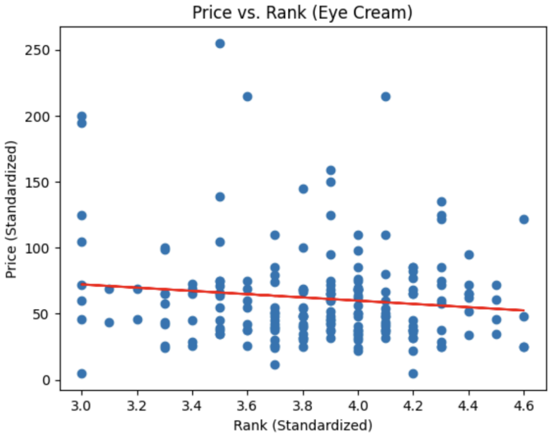

Figure 6: Linear Regression on Eye Cream

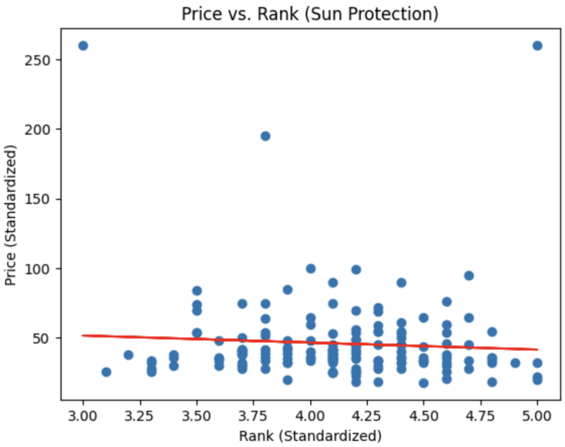

Figure 7: Linear Regression on Sun Protection

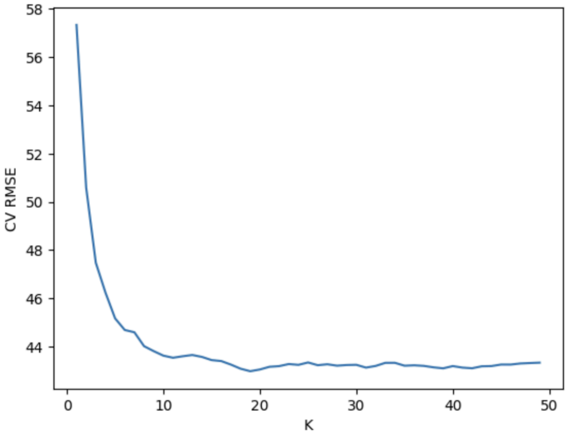

Figure 8: Cross Validation RMSE Scores over K

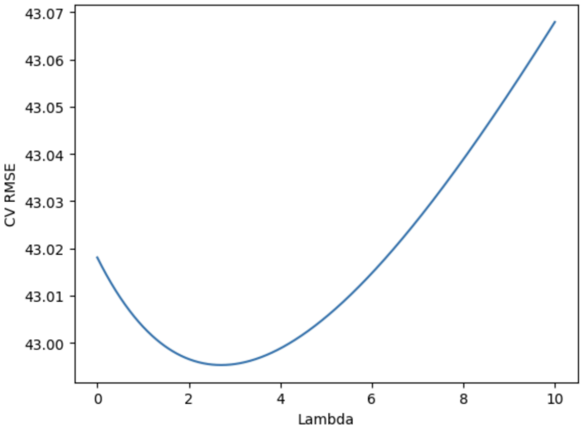

Figure 9: Cross Validation RMSE Scores over Lambda