The Fashion Formula: Understanding What Drives Zara Sales

Abstract

Zara, as the third-largest global fast-fashion retailer, holds significant influence within the fast-fashion apparel market. In this project, we aimed to investigate the key factors that influence sales volume at Zara, as a glimpse into the fashion industry. To this end, we built a multiple linear regression model in R, using the stepwise procedure with the BIC criterion, with a dataset named Zara Sales for EDA on Kaggle. Our final model utilized box-cox transformation and showed that promotion, price, section, and season are significant predictors for sales volumes at Zara.

1. Background and Significance

Zara is a globally recognized fast-fashion retailer founded in 1975 by Amancio Ortega in A Coruña, Spain. Over the past several decades, Zara has become one of the most influential and widely-known fast-fashion brands across the world. As of November 2022, Zara accounted for approximately 13% of the global fast-fashion market, ranking third behind Shein and H&M [1]. As a leading industry player, Zara provides a compelling case through which we can examine the driving force behind the sales volumes in fast fashion. This can help fast-fashion retailers make better strategic marketing and pricing decisions by providing a deeper understanding of consumer behaviors and demands.

Our research question is: *What factors influence the sale volume of clothes at Zara?*

2. Data

2.1 Data Description

For our study, we used the "Zara Sales for EDA" dataset on Kaggle [2]. This dataset has 20252 rows and 17 columns. Table 1 (in appendix) shows the list of variables included in the dataset and their corresponding types. We chose Sales.Volume as the response variable. It is defined as the approximate number of units sold per product.

2.2 Data Preparation

Before starting our model selection process, we removed variables that are either redundant or not meaningful to our study, including the identifiers and the descriptive variables. Specifically, Product Category, Brand, and currency were removed because each contains only a single level. Our trimmed dataset includes 10 columns: Promotion, Product.Position, Seasonal, Season, Sales.Volume, price, terms, section, material, and origin (Table 1). There are no missing values in the trimmed dataset.

In the trimmed dataset, many variables were originally recorded as character strings. To avoid errors during model fitting and comparison, we converted all character variables into factors. This ensures that categorical predictors would be correctly treated as categorical variables, with dummy variables generated automatically during model fitting.

To detect multicollinearity issues, we used the variance inflation test (VIF) with a threshold of 5 to remove highly collinear variables. The results indicated that no variables in our dataset showed severe multicollinearity issues, and all variables had a VIF score less than 2. The cleaned and trimmed dataset has 20252 rows and 10 columns.

3. Methods and Results

3.1 Model Fitting Procedure

After fitting the full model, we first conducted all-subset selection using Mallow's Cp and Adjusted R^2 criterions. The two resulting models were exactly the same. Then, we conducted stepwise regression using AIC and BIC criterions, which resulted in two distinct models different from the previous Mallow's Cp model. In Table 1, we compared the full model, the Mallow's Cp model, the AIC model, and the BIC model based on their Adjusted R^2, prediction error, and the balance between model fitness and model complexity. Their prediction performance was evaluated based on the leave-one-out cross validation score and the 5-fold cross validation score, each square rooted for easier interpretation.

Model	Adjusted R^2	Square-rooted LOOCV	Square-rooted 5-fold CV	AIC	BIC	Num of Preds
Full Model	0.93	78.956	109.01	234431.1	234716.1	9
Mallow's Cp Model	0.93	78.924	**105.78**	234414.7	234596.8	7

AIC Model	0.93	**78.912**	107.97	**234409.3**	234480.5	5
BIC Model	0.93	78.913	106.89	234409.8	**234473.2**	**4**

Table 2

Among the four models, the Mallow's Cp model has the lowest 5-fold CV, the AIC model has the lowest LOOCV and AIC, while the BIC model has the lowest BIC and is the most parsimonious. For these three models, their difference in LOOCV, 5-fold CV, and AIC are negligible (all smaller than 2.5). Therefore, we identified the BIC model as our best first-order model because it finds the best balance between model fitness and model complexity while demonstrating a satisfactory prediction performance.

3.2 Model Diagnostics and Refinement

We conducted a residual analysis to evaluate the selected model. A residual plot (Figure 2) showed that the residuals violated the constant variance, linearity, and mean zero assumptions necessary for multiple linear regression. The Q-Q plot (Figure 3) and the time sequence plot (Figure 4) showed that the normality and independence assumptions were met.

First, we considered removing outliers to improve the model. Using studentized deleted residuals, 921 outliers with respect to the response variable were identified; using leverage, 405 outliers with respect to the predictors were identified. To determine if those outliers were influential, we calculated the cook's distance. These values were compared to the 50th percentile of the F distribution (df1=7, df2=20245). Observations with a cook's distance above the threshold (0.907) are considered influential. We found that there were no influential outliers.

To solve the non-constant variance of the residual plot, we conducted a box-cox transformation in the response variable Sales.Volume. The result indicated an optimal transformation of 0.1 (Figure 5), which was easy for interpretation. After fitting the same BIC model with a response variable of $Sales.Volume^{0.1}$, the new residual plot (Figure 6) showed that the constant-variance, linearity, and mean-zero assumptions were all satisfied. The corresponding Q-Q plot (Figure 7) and time sequence plot (Figure 8) indicated that normality and independence assumptions were satisfied. In Table 2, we compared the transformed box-cox model with the original BIC model, which showed significant improvement across all evaluation criterions. Notice that the LOOCV and the 5-fold CV for the box-cox model had been transformed back to the original scale by raising the predicted response values to the power of 10 in the calculation process. Therefore, we could compare the two CV scores for the two models on the same scale. The results showed that the box-cox model was significantly improved in its prediction performance compared to the original BIC model.

Model	Adjusted R^2	Square-rooted LOOCV	Square-rooted 5-fold CV	AIC	BIC	Num of Preds
BIC Model	0.93	78.956	106.89	234409.8	234473.2	4
Box-Cox Model	**0.938**	**17.75**	**41.71**	**-117261**	**-117197.7**	4

Table 3

We also considered interaction terms and higher order terms, neither of which significantly improved the overall model performance.

3. Final Model Interpretation

The box-cox transformed model was chosen as the final best model because it had the highest Adjusted R^2, lowest prediction error, was the most parsimonious, and satisfied all regression assumptions. The final model is shown on the next page. Out of the original 9

predictors, 4 variables were statistically significant, 3 of which were categorical (Promotion, section, season) and 1 was numeric (price). The baselines for the 3 categorical variables were: Promotion=No, section=MAN, and season=Autumn. The interpretation for the coefficients is as follows: for a 1-unit increase in the predictor, $Sales.Volume^{0.1}$ increases by the coefficient value, adjusting for the simultaneous linear change in all other predictors.

```
Call:
lm(formula = I(Sales.Volume^0.1) ~ Promotion + price + section + 
    season, data = zara.anl)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.042902 -0.009345  0.000064  0.009309  0.040165 

Coefficients:
               Estimate Std. Error  t value Pr(>|t|)    
(Intercept)   1.970e+00  2.778e-04 7089.161   <2e-16 ***
PromotionYes  9.344e-02  1.925e-04  485.298   <2e-16 ***
price        -5.869e-04  4.070e-06 -144.204   <2e-16 ***
sectionWOMAN  1.966e-02  1.977e-04   99.422   <2e-16 ***
seasonWinter  2.636e-02  2.415e-04  109.154   <2e-16 ***
seasonSummer  2.637e-02  2.918e-04   90.363   <2e-16 ***
seasonSpring  2.335e-04  2.506e-04    0.932    0.352    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01338 on 20245 degrees of freedom
Multiple R-squared:  0.9383,    Adjusted R-squared:  0.9382 
F-statistic: 5.127e+04 on 6 and 20245 DF,  p-value: < 2.2e-16
```

4. Conclusion

4.1. Discussion

The model indicates four significant predictors that influence the products' sale volumes at Zara, including Promotion, price, section, and season. We found that adjusting for the linear change in the rest of other predictors, sale volumes are positively associated with the presence of promotion (Figure 9), and negatively associated with the price of the products (Figure 10). Additionally, sales volumes of products in the Woman section tend to be larger than products in the Man section (Figure 11). We also discovered an important seasonal trend: sales volumes of winter and summer products tend to be larger than spring and autumn products (Figure 12).

These findings are relevant for Zara when formulating marketing strategies. To increase product sales volumes, the company could consider conducting promotional activities more frequently, keeping overall prices as low as possible, allocating more resources to women's apparel and fewer to men's apparel, and offering a greater variety of winter and summer products compared to those for spring and autumn. However, since higher sales volumes do not necessarily translate into higher net profits, Zara must also account for their costs in this analysis. For example, the frequency of promotional activities should not exceed the point at which net profits begin to decline due to increased costs.

4.2. Limitations

Our model-fitting process did not assess all possible factors that may influence the sales volumes of Zara products, largely because the predictors available in the dataset were limited. In particular, we were unable to account for potentially important but difficult-to-measure factors, such as design aesthetics and product trendiness. As a result, our analysis of sales volume focuses primarily on external factors and does not sufficiently capture internal product characteristics (the only such predictor was material, which was determined to be insignificant).

Another limitation is evident in the residual plot (Figure 6). Although the residuals appear randomly scattered overall, the constant-variance assumption holds only within two distinct clusters, which correspond to the two promotion levels (Figure 9). By modeling promoted and non-promoted products jointly, we may have overlooked differences in the factors influencing sales, given the distinction in their price and sales patterns.

4.3. Future Considerations

To gain a deeper understanding of the factors influencing sales volumes in fast-fashion products, future analyses could incorporate additional predictors such as design aesthetics and product trendiness. We could also fit separate models for promoted and non-promoted products to examine whether the determinants of sales differ across promotion levels. Finally, to assess the generalizability of our findings, similar analyses could be conducted using data from other major fast-fashion retailers, such as Shein and H&M.

Reference

[1] Cardona, N. *Fast Fashion Statistics 2025*. Uniform Market, 2025. Available at: www.uniformmarket.com/statistics/fast-fashion-statistics (accessed Dec. 16, 2025).'

[2] Idrissi, M. *Zara Sales for EDA* [Dataset]. Kaggle, 2025. https://doi.org/10.34740/KAGGLE/DSV/13507590

Appendix

Table 1: Column Names and Data Types

The columns names and their corresponding data types in the dataset Zara Sales for EDA

	Variable	Type
1	Product ID	Integer
2	Product Position	Categorical (Aisle/End-cap/Front of Store)
3	Promotion	Categorical (Yes/No)
4	Product Category	Categorical (clothing)
5	Seasonal	Categorical (Yes/No)
6	Season	Categorical (Spring/Summer/Autumn/Winter)
7	Sales Volume	Integer
8	Brand	Categorical (Zara)
9	url	Text
10	name	Text
11	description	Text
12	price	Float
13	currency	Text (USD)
14	terms	Categorical (jackets/jeans/shoes/sweaters t-shirts)
15	section	Categorical (WOMAN/MAN)
16	material	Categorical (Acrylic/Cotton/Denim/…11 categories in total)
17	origin	Categorical (Argentina/Bangladesh/Brazil/… 12 categories in total)

Figure 2: Residual plot of the BIC model

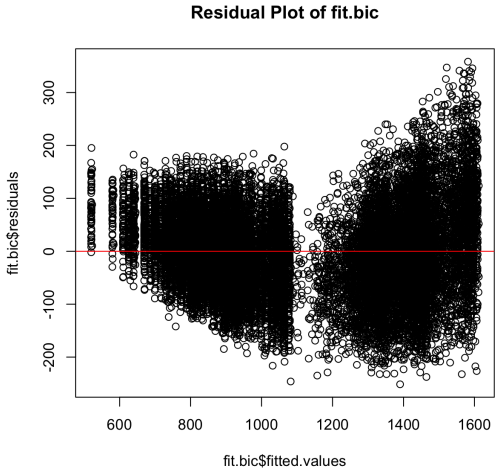

Figure 3: Normal Q-Q Plot of the BIC model

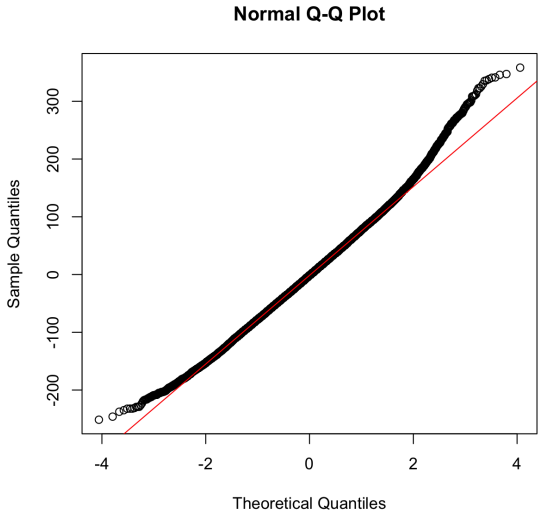

Figure 4: Time sequence plot of the BIC model

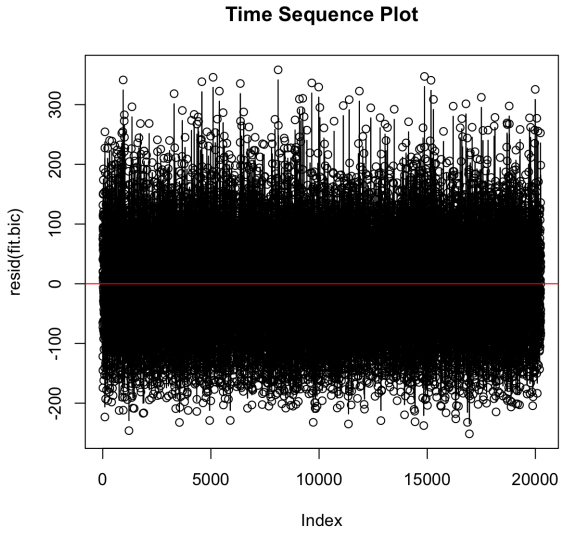

Figure 5: Result of box-cox transformation (λ=0.1)

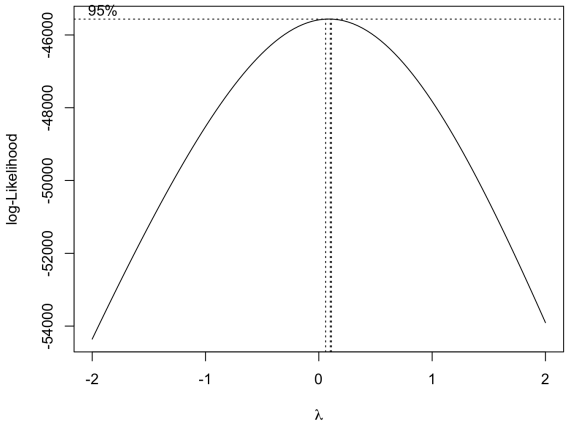

Figure 6: Residual plot of the transformed box-cox model

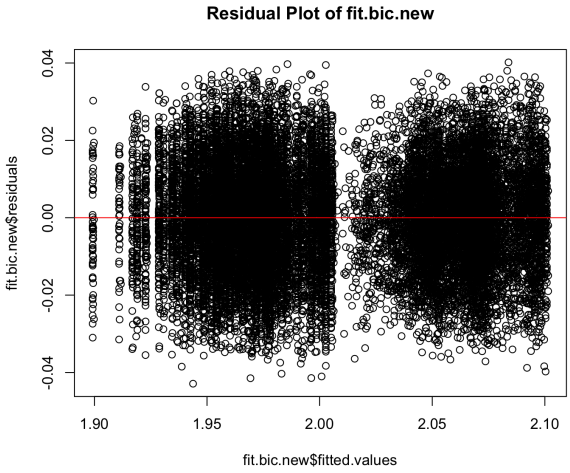

Figure 7: Q-Q plot of the transformed box-cox model

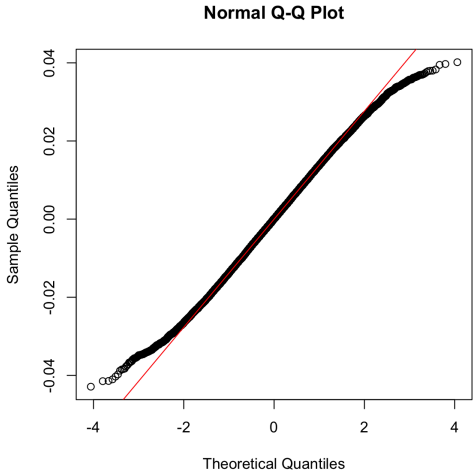

Figure 8: Time sequence plot of the transformed box-cox model

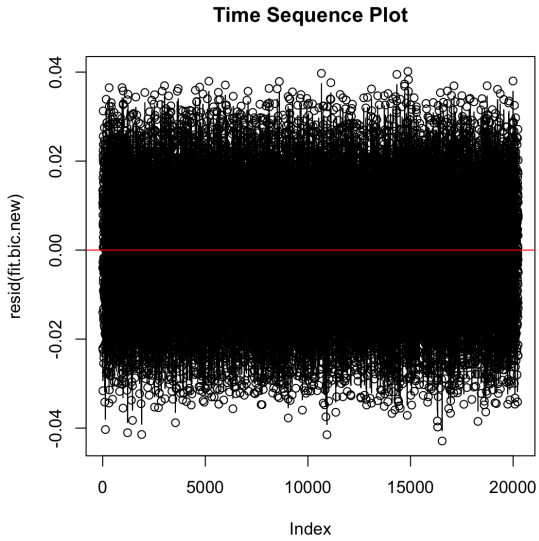

Figure 9: Effect of Promotion on sales volumes

Figure 10: Scatter plot between sales volume and price with a fitted linear regression line

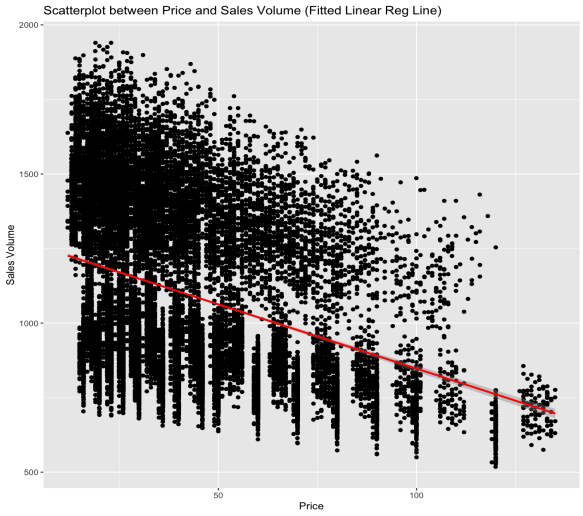

Figure 11: Effect of section on sales volumes

Figure 12: Effect of season on sales volumes

AI Statement: We used AI in helping us debug the code and doing the fast fashion industry research. AI is not used in the writing process.