DATA 605: Computational Mathematics

Gabriel Campos

Last edited December 04, 2024

Contents

Library

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```r
library(readr)
```

Final Examination: Business Analytics and Data Science

Instructions:

You are required to complete this take-home final examination by the end of the last week of class. Your solutions should be uploaded in **pdf** format as a knitted document (with graphs, content, commentary, etc. in the pdf). This project will showcase your ability to apply the concepts learned throughout the course.

The dataset you will use for this examination is provided as retail data.csv, which contains the following variables:

- Product_ID: Unique identifier for each product.
- Sales: Simulated sales numbers (in dollars).
- Inventory_Levels: Inventory levels for each product.
- Lead_Time_Days: The lead time in days for each product.
- Price: The price of each product.
- Seasonality_Index: An index representing seasonality.

Problem 1:

Business Risk and Revenue Modeling

Context: You are a data scientist working for a retail chain that models sales, inventory levels, and the impact of pricing and seasonality on revenue. Your task is to analyze various distributions that can describe sales variability and forecast potential revenue.

```r
retail_df <- read_csv("synthetic_retail_data.csv")
```

Part 1:

Empirical and Theoretical Analysis of Distributions (5 Points)

Task:

1.

Generate and Analyze Distributions:

- **X ~ Sales:** Consider the Sales variable from the dataset. Assume it follows a Gamma distribution and estimate its shape and scale parameters using the fitdistr function from the MASS package.
- **Y ~ Inventory Levels:** Assume that the sum of inventory levels across similar products follows a Lognormal distribution. Estimate the parameters for this distribution.
- **Z ~ Lead Time:** Assume that Lead_Time_Days follows a Normal distribution. Estimate the mean and standard deviation. Calculate Empirical Expected Value and Variance:

Calculate the empirical mean and variance for all three variables. Compare these empirical values with the theoretical values derived from the estimated distribution parameters.

```r
glimpse(retail_df)
```

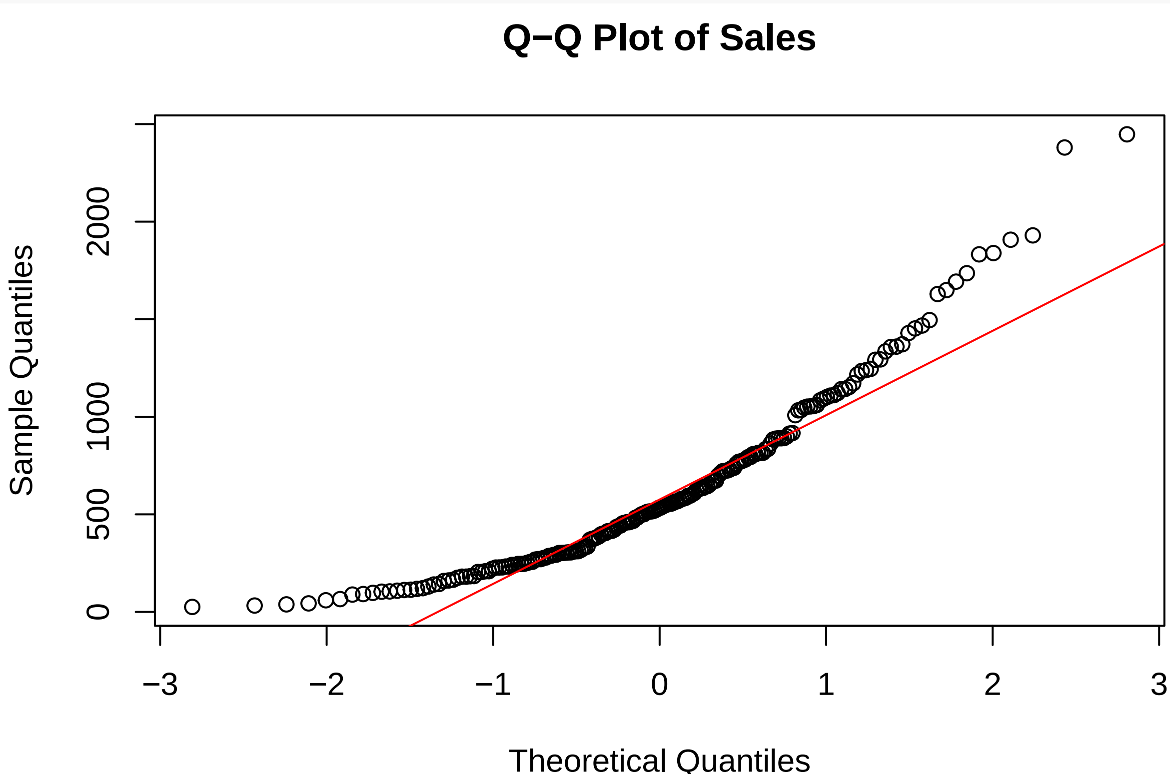
X ~ Sales

```
## Rows: 200
## Columns: 6
## $ Product_ID        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ Sales             <dbl> 158.43952, 278.99020, 698.85868, 1832.39467, 459.703~
## $ Inventory_Levels  <dbl> 367.4421, 426.6512, 407.6394, 392.3912, 448.3120, 54~
## $ Lead_Time_Days    <dbl> 6.314587, 5.800673, 3.071936, 3.534253, 10.802241, 1~
## $ Price             <dbl> 18.795197, 26.089636, 22.399985, 27.092013, 18.30782~
## $ Seasonality_Index <dbl> 1.1839497, 0.8573051, 0.6986774, 0.6975404, 0.840725~
```

```r
summary(retail_df)
```

```
##    Product_ID         Sales         Inventory_Levels Lead_Time_Days
##  Min.   :  1.00   Min.   :  25.57   Min.   : 67.35   Min.   : 0.491
##  1st Qu.: 50.75   1st Qu.: 284.42   1st Qu.:376.51   1st Qu.: 5.291
##  Median :100.50   Median : 533.54   Median :483.72   Median : 6.765
##  Mean   :100.50   Mean   : 636.92   Mean   :488.55   Mean   : 6.834
##  3rd Qu.:150.25   3rd Qu.: 867.58   3rd Qu.:600.42   3rd Qu.: 8.212
##  Max.   :200.00   Max.   :2447.49   Max.   :858.79   Max.   :12.722
##      Price        Seasonality_Index
##  Min.   : 5.053   Min.   :0.3305
##  1st Qu.:16.554   1st Qu.:0.8475
##  Median :19.977   Median :0.9762
##  Mean   :19.560   Mean   :0.9829
##  3rd Qu.:22.924   3rd Qu.:1.1205
##  Max.   :29.404   Max.   :1.5958
```

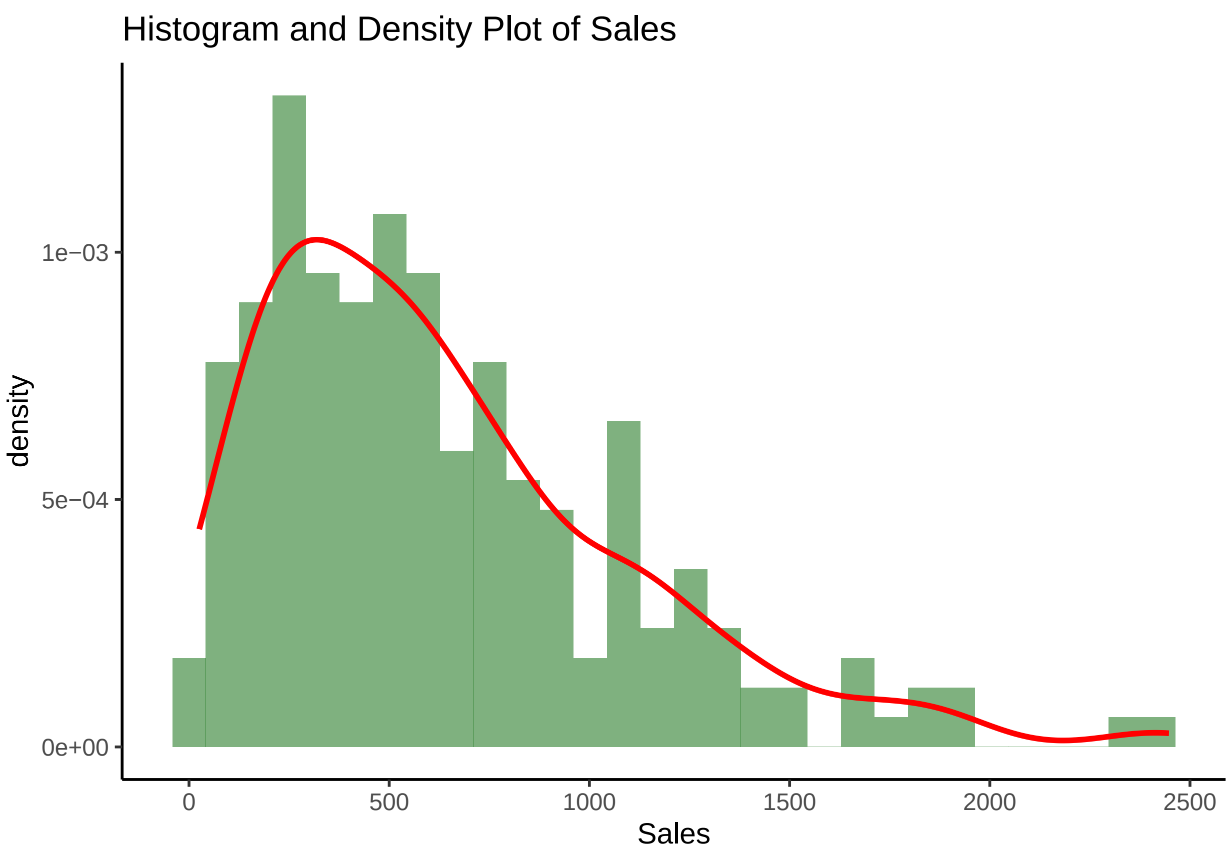
```r
qqnorm(retail_df$Sales, main = "Q-Q Plot of Sales")
qqline(retail_df$Sales, col = "red")
```



```r
ggplot(retail_df, aes(x = Sales)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30, fill = "darkgreen",
                 alpha = 0.5) +
  geom_density(color = "red", size = 1) +
  ggtitle("Histogram and Density Plot of Sales") +
  theme_classic()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```r
shapiro.test(retail_df$Sales)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  retail_df$Sales
## W = 0.90377, p-value = 4.397e-10
```

2.

Part 2:

Probability Analysis and Independence Testing (5 Points)

Task:

1.

Empirical Probabilities: For the Lead_Time_Days variable (assumed to be normally distributed), calculate the following empirical probabilities:

- $P(Z > \mu | Z > \mu - \sigma)$
- $P(Z > \mu + \sigma > Z > \mu)$
- $P(Z > \mu + 2\sigma > Z > \mu)$

2.

Correlation and Independence:

- Investigate the correlation between Sales and Price. Create a contingency table using quartiles of Sales and Price, and then evaluate the marginal and joint probabilities.
- Use Fisher's Exact Test and the Chi-Square Test to check for independence between Sales and Price. Discuss which test is most appropriate and why.

Problem 2

Advanced Forecasting and Optimization (Calculus) in Retail

Context: You are working for a large retail chain that wants to optimize pricing, inventory management, and sales forecasting using data-driven strategies. Your task is to use regression, statistical modeling, and calculus-based methods to make informed decisions.

Part 1

Descriptive and Inferential Statistics for Inventory Data (5 Points)

Task:

1.

Inventory Data Analysis:

- Generate univariate descriptive statistics for the Inventory_Levels and Sales variables.
- Create appropriate visualizations such as histograms and scatterplots for Inventory_Levels, Sales, and Price.
- Compute a correlation matrix for Sales, Price, and Inventory_Levels.
- Test the hypotheses that the correlations between the variables are zero and provide a 95% confidence interval.

2.

Discussion:

- Explain the meaning of your findings and discuss the implications of the correlations for inventory management. Would you be concerned about multicollinearity in a potential regression model? Why or why not?

Part 2

Linear Algebra and Pricing Strategy (5 Points)

Task:

1.

Price Elasticity of Demand:

- Use linear regression to model the relationship between Sales and Price (assuming Sales as the dependent variable).
- Invert the correlation matrix from your model, and calculate the precision matrix.
- Discuss the implications of the diagonal elements of the precision matrix (which are variance inflation factors).
- Perform LU decomposition on the correlation matrix and interpret the results in the context of price elasticity.

Part 3:

Calculus-Based Probability & Statistics for Sales Forecasting (5 Points)

Task:

1.

Sales Forecasting Using Exponential Distribution:

- Identify a variable in the dataset that is skewed to the right (e.g., Sales or Price) and fit an exponential distribution to this data using the fitdistr function.
- Generate 1,000 samples from the fitted exponential distribution and compare a histogram of these samples with the original data's histogram.
- Calculate the 5th and 95th percentiles using the cumulative distribution function (CDF) of the exponential distribution.
- Compute a 95% confidence interval for the original data assuming normality and compare it with the empirical percentiles.

2.

Discussion:

- Discuss how well the exponential distribution models the data and what this implies for forecasting future sales or pricing. Consider whether a different distribution might be more appropriate.

Part 4

Regression Modeling for Inventory Optimization (10 Points)

Task:

1.

Multiple Regression Model:

- Build a multiple regression model to predict Inventory_Levels based on Sales, Lead_Time_Days, and Price.
- Provide a full summary of your model, including coefficients, R-squared value, and residual analysis.

2.

Optimization:

- Use your model to optimize inventory levels for a peak sales season, balancing minimizing stockouts with minimizing overstock.