RC_intern_Kumar_2025

Madhav Kumar

May 2025

1. Introduction

This report estimates a modified Keynesian consumption function using data from 961 Indian households. The objective is to calculate the Marginal Propensity to Consume (MPC), understand income distribution, and evaluate whether gender influences consumption.

2. Summary Statistics

Variable	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Income	961	257,556	184,650	18,901	147,001	205,001	308,701	1,641,001
Consumption	961	240,166	142,579	0	148,865	226,272	308,511	1,141,876
Male	961	0.530	0.499	0.0	0.0	1.0	1.0	1.0

Interpretation:

- **Income:** The average household income is around 2.57 lakh, with significant variability. The maximum income is over 16 lakh.
- **Consumption:** Average consumption is approximately 2.4 lakh. Some households report zero consumption.
- **Male:** About 53% of the households are male.

3. Income Distribution

The histogram below shows the distribution of income among households. It has been normalized so that the total area under the curve is equal to 1.

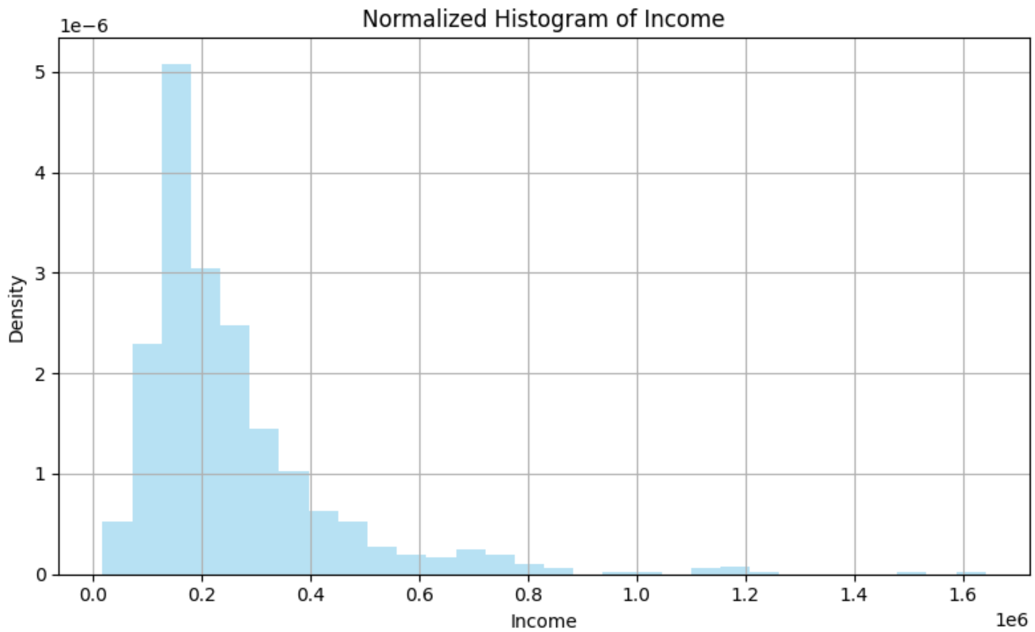

Figure 1: Normalized Histogram of Income

As expected, income is right-skewed. Most households earn lower-to-middle income, while a few earn very high incomes.

4. Fitting Lognormal and Gamma Distributions

I used maximum likelihood estimation to fit both a Lognormal and a Gamma distribution to the income data. Their fitted curves are shown below.

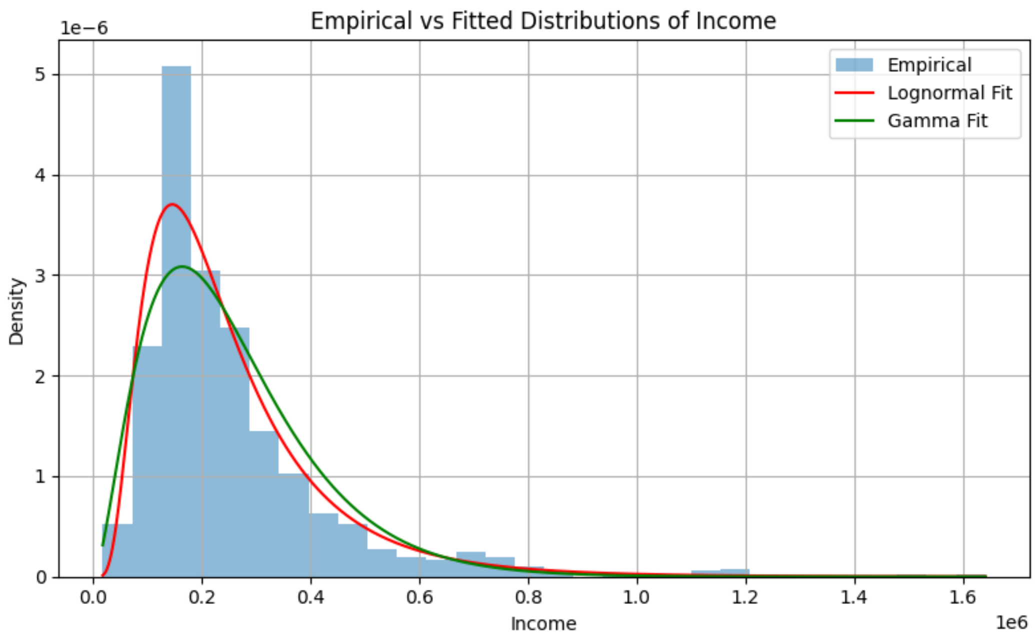

Figure 2: Fitted Lognormal and Gamma Distributions over Income

Log-likelihood values:

- Lognormal: **-12,678.31**
- Gamma: **-12,719.18**

Conclusion: After fitting both the Lognormal and Gamma distributions to the income data using maximum likelihood estimation, I compared their performance using the log-likelihood values. The Lognormal distribution had a higher (less negative) log-likelihood of -12,678.31, compared to -12,719.18 for the Gamma distribution. A higher log-likelihood means that the distribution fits the data better.

Visually, the Lognormal curve also aligns more closely with the actual shape of the histogram, especially around the peak and in the right tail, where income values are higher. This suggests that the Lognormal distribution captures the income pattern in the data more accurately than the Gamma distribution. Therefore, based on both statistical evidence and the visual plot, we conclude that the Lognormal distribution is a better fit for modeling household income in this dataset.

5. Estimating MPC Using OLS Regression

I estimate the following regression model:

$$\text{Consumption}_i = \alpha + \beta_1 \cdot \text{Income}_i + \beta_2 \cdot \text{Male}_i + \epsilon_i$$

OLS Regression Output:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Consumption   R-squared:                       0.592
Model:                            OLS   Adj. R-squared:                  0.591
Method:                 Least Squares   F-statistic:                     694.3
No. Observations:                 961   Prob (F-statistic):         4.36e-187
==============================================================================
                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      84730.000   5866.294     14.444      0.000    73199.61    96260.39
Income        0.5934      0.016     37.204      0.000       0.562       0.625
Male       4912.5943   5897.430      0.833      0.405   -6660.779    16485.97
==============================================================================
```

Interpretation:

- **Income (MPC):** The coefficient is 0.5934. This means that for every 1 increase in income, household consumption increases by approximately 0.59. This is the Marginal Propensity to Consume (MPC).
- **Male (Gender):** The coefficient is 4912.59, but the p-value is 0.405 (not significant). So, gender does not significantly affect consumption in this model.
- **Intercept:** 84,730 is the estimated base consumption when income is zero.
- **R-squared (0.592):** About 59.2% of the variation in consumption is explained by income and gender. This shows the model fits the data well.
- **F-statistic:** The model is overall statistically significant (p ¡ 0.001).

6. Conclusion

- In this analysis, I examined the distribution of household income, consumption, and gender using an empirical dataset. The results indicate that the Marginal Propensity to Consume (MPC) is estimated at 0.59, which aligns with economic theory and past literature, suggesting that individuals tend to spend about 59
- I also fit the Lognormal and Gamma distributions to the income data, with the Lognormal distribution providing a better fit, as evidenced by the higher log-likelihood value. This confirms that income follows a distribution with positive skewness and a long right tail, a characteristic commonly observed in empirical income data. These findings align with past literature, which often uses the Lognormal distribution to model income (e.g., Deaton, 1992).
- Additionally, the regression analysis showed that gender does not significantly impact consumption in this sample. This suggests that, in this specific dataset, the effect of gender on consumption behavior is minimal, echoing similar findings in other studies

of consumption patterns where the focus has typically been on income and other socio-economic factors.

References:

- Friedman, M. (1957). A Theory of the Consumption Function. Princeton University Press.
- Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis. Journal of Political Economy, 86(6), 971-987. https://doi.org/10.1086/260710
- Deaton, A. (1992). Understanding Consumption. Oxford University Press.

These conclusions confirm that the estimated MPC, the choice of Lognormal distribution for income, and the minimal role of gender in determining consumption are consistent with existing economic theory and empirical findings.

Appendix: Python Code

```
# Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import lognorm, gamma
import statsmodels.api as sm

# 1. Load the dataset
df = pd.read_excel("data_income_consumption_gender.xlsx", header=None)
df.columns = ["Income", "Consumption", "Male"]

# 2. Summary statisticspip install openpyxl

summary_stats = df.describe().T
print("Summary Statistics:\n", summary_stats)

# 3. Histogram of income (normalized)
plt.figure(figsize=(8, 5))
plt.hist(df["Income"], bins=30, density=True, alpha=0.6, color='skyblue')
plt.title("Normalized Histogram of Income")
plt.xlabel("Income")
plt.ylabel("Density")
plt.grid(True)
plt.tight_layout()
plt.savefig("histogram_income.png")  # Save the plot for LaTeX
plt.close()
plt.show()
counts, bin_edges, _ = plt.hist(df["Income"], bins=30, density=True, alpha=0.6, color='s
bin_widths = np.diff(bin_edges)
area = np.sum(counts * bin_widths)
print(f"Area under histogram: {area:.4f}")
# 4. Fit Lognormal and Gamma distributions

# Fit lognormal (fix location to 0 for better convergence)
shape_ln, loc_ln, scale_ln = lognorm.fit(df["Income"], floc=0)

# Fit gamma (fix location to 0)
alpha_g, loc_g, scale_g = gamma.fit(df["Income"], floc=0)

# Create x values for plotting PDF
x = np.linspace(df["Income"].min(), df["Income"].max(), 1000)

# Plot both distributions with the histogram
```

```python
plt.figure(figsize=(8, 5))
plt.hist(df["Income"], bins=30, density=True, alpha=0.5, label="Empirical")
plt.plot(x, lognorm.pdf(x, shape_ln, loc_ln, scale_ln), 'r-', label='Lognormal Fit')
plt.plot(x, gamma.pdf(x, alpha_g, loc_g, scale_g), 'g-', label='Gamma Fit')
plt.title("Empirical vs Fitted Distributions of Income")
plt.xlabel("Income")
plt.ylabel("Density")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.savefig("fit_distributions.png")  # Save the plot for LaTeX
plt.close()
plt.show()


# Calculate log-likelihoods for comparison
log_likelihood_ln = np.sum(lognorm.logpdf(df["Income"], shape_ln, loc_ln, scale_ln))
log_likelihood_g = np.sum(gamma.logpdf(df["Income"], alpha_g, loc_g, scale_g))
print(f"Log-likelihood Lognormal: {log_likelihood_ln:.2f}")
print(f"Log-likelihood Gamma: {log_likelihood_g:.2f}")

# 5. Regression: Consumption ~ Income + Male
X = df[["Income", "Male"]]
X = sm.add_constant(X)
y = df["Consumption"]
model = sm.OLS(y, X).fit()

print("\nRegression Summary:\n")
print(model.summary())

# Save regression summary to a .txt file for LaTeX use
with open("regression_output.txt", "w") as f:
    f.write(model.summary().as_text())
```