Module 4

**Explaining Customer Spending for Online Retail:**

**An Analysis of Email Campaign Effectiveness**

James Burnett

School of Information Systems & Management, University of South Florida

ISM 6137: Advanced Statistical Modeling

Dr. Daniel Zantedeschi

March 27, 2025

# Introduction

This report examines whether the mens or womens e‐mail campaign was successful in influencing customer spending. Using data from an online retail campaign, we developed gamma regression models (GLMs) to predict the “recent spend” variable. The analysis includes data cleaning, exploratory analysis, hypothesis formulation, model estimation, and testing of regression assumptions. The aim is to provide clear insights into the factors driving customer spending and inform the retailer’s marketing strategy.

# Data Preparation and Exploratory Analysis

Data were imported from an Excel file and cleaned by standardizing variable names and formats. Key variables such as recency (renamed as “last\_purchased”), past spending, and email campaign type were appropriately recoded. In addition, a new categorical variable, “spending\_category\_yr,” was created to segment customers based on their historical spending. To examine the complete code, include the import and cleaning section, see the appendix.

Exploratory data analysis (EDA) involved visualizing the distribution of the dependent variable, “recent spend.” Initial histograms and Q–Q plots revealed significant right skewness. A log transformation was then applied (via the np.log1p function), resulting in a distribution that approximated normality more closely, thereby justifying the transformation for subsequent modeling.

A graph and a diagram

AI-generated content may be incorrect.

Figure : Evaluation of Data Skeeness

# Comparison of Pre- and Post-Log Transformation Plots

The histogram and Q-Q plot of the original “recent\_spend” variable show a heavily right‐skewed distribution, with many customers spending relatively small amounts and a few outliers spending much larger amounts. This skewness is evident in the Q-Q plot, where the data points deviate substantially from the diagonal at the upper tail. After applying a log transformation, the distribution appears more symmetric, and the Q-Q plot shows points much closer to the diagonal. This indicates that the log transformation reduces the impact of high‐spending outliers and helps meet the assumption of approximate normality, which can improve model performance and interpretability

A graph of a graph of a person

AI-generated content may be incorrect.

Figure : Evaluation of Data Skewness Post-Log Transformation

# Hypothesis Formulation

A hypothesis table was developed to outline the expected relationships between predictors and customer spending. For example, it was hypothesized that:

* Last Purchased: A higher value (indicating a longer time since purchase) is expected to decrease spending.
* Spending Category: Customers with higher past spending are expected to spend more in the future.
* Email Type: Exposure to an email campaign (whether male or female) is anticipated to increase spending relative to receiving no email.
* Interaction Effects: The interaction between email campaign type and new customer status is included to capture potential differential responses.

A complete hypothesis table (included as an Excel output in the analysis) summarizes the predictor, the expected effect (positive, negative, or ambiguous), and the rationale for each.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure : Hypothesis Table

# Model Specification and Estimation

Given the continuous, positive, and skewed nature of the “recent spend” variable, gamma regression with a log link was chosen as the modeling framework. Three alternative GLM models were estimated:

1. GLM Model 1: Baseline model with predictors for recency, email campaign type, and an interaction with new customer status.

A screenshot of a computer

AI-generated content may be incorrect.

Figure : GLM Mode 1 Output

1. GLM Model 2: Extended model that adds purchase channel variables (phone vs. web).

A screenshot of a computer

AI-generated content may be incorrect.

Figure : GLM Model 2 Output

1. GLM Model 3: Full model that further includes the new customer indicator and spending category, as well as a squared term for recency to capture non-linearity.

A screenshot of a computer

AI-generated content may be incorrect.

Figure : GLM Model 3 Output

A screenshot of a computer screen

AI-generated content may be incorrect.Model outputs were aggregated using the Stargazer package to provide a compact summary of coefficient estimates and model fit statistics. Model selection was guided by theoretical considerations and fit criteria, with GLM Model 3 offering the most comprehensive explanation of customer spending.

Figure : Stargazer Output for GML Models

*Note: Significance Levels:*

* \*\*\* (Three stars) → p < 0.01 (Highly significant: strong evidence against the null hypothesis) Interpretation: Very likely to be important predictors.
* \*\*\* (Two stars) → p < 0.05 (Moderately significant: good evidence against the null hypothesis). Interpretation: Have moderate confidence in their impact.
* (One star) → p < 0.1 (Weakly significant: some evidence against the null hypothesis). Interpretation: Suggest a possible effect, but with less certainty
* No stars → p ≥ 0.1 (Not significant: weak or no evidence that the variable has an effect). Interpretation: Variables with no stars likely do not have a strong influence

# Regression Assumptions and Diagnostics

To ensure the validity of our GLM results, the following assumptions were evaluated:

1. *Linearity (on the link scale):*

The log link in the gamma regression implies a linear relationship between the predictors and the log of the expected spending. Scatterplots and the inclusion of a squared term for “last\_purchased” were used to capture and verify potential non-linear effects.

1. *Normality of Residuals:*

Although GLMs do not assume normally distributed errors in the traditional sense, we examined the distribution of the dependent variable before and after log transformation. Q–Q plots showed that the log-transformed “recent spend” more closely approximated a normal distribution, supporting our transformation choice.

1. *Homoscedasticity:*

Gamma regression naturally accommodates heteroscedasticity by modeling the variance as a function of the mean. Residual plots were visually inspected for patterns that might indicate issues. As an area for further improvement, robust standard errors or heteroskedasticity-consistent covariance estimators (e.g., WLS or FGLS) could be implemented to further correct any remaining issues.

1. *Independence:*

The dataset is comprised of independent observations from individual customers. There is no evidence of clustering or serial correlation that would violate the independence assumption.

1. *Multicollinearity:*

Prior to model estimation, correlations among predictors were reviewed. There were no indications of severe multicollinearity. Nonetheless, future work could incorporate formal variance inflation factor (VIF) analysis to ensure predictor independence.

# Results and Discussion

The results from the GLM models suggest that email campaign variables do impact customer spending. For example, receiving an email (whether male or female) is generally associated with higher spending relative to the control group. In addition, historical spending and recency emerged as robust predictors of future spending. The inclusion of purchase channel variables in Model 2 and further demographic segmentation in Model 3 helped improve model fit. Marginal effects analysis (not detailed here) indicated that the economic significance of these predictors is considerable, even if some coefficients are statistically ambiguous.

# Robust Standard Errors for Heteroskedasticity Correction

In addition to fitting the standard Gamma regression models, we applied robust standard errors (sometimes called “White” or “Huber‐White” SEs) to address potential heteroskedasticity in our data. These robust errors do not alter the coefficient estimates themselves but do adjust the standard errors, p -values, and confidence intervals.

After applying robust SEs, the signs and magnitudes of the coefficients in our GLM models remained consistent with the original results, indicating that the overall direction and relative importance of the predictors did not change. However, some coefficients may have slightly different levels of statistical significance (as indicated by their p -values).

Overall, the robust‐SE results reinforce our main findings:

* The email campaign variables (male or female) generally maintain a positive relationship with recent spending, suggesting that targeted emails increase customer spending relative to no email.
* Past spending and recency remain strong predictors of future spending.
* predictors, such as purchase channel, may or may not change in significance, but their coefficient estimates remain in line with prior expectations.

*Note: The outputs for each model after applying robust SEs can be seen in the appendix*

By using robust standard errors, we gain more reliable inferences under the assumption that the variance of errors is not constant across all observations. Thus, we recommend interpreting the robust‐SE results for final conclusions, especially when heteroskedasticity is suspected.

# Conclusion and Future Directions

The analysis indicates that targeted email campaigns have a positive effect on customer spending. The full model (GLM Model 3) provides the best explanation by incorporating a broad range of predictors, including non-linear effects.

Moving forward, implementing heteroskedasticity corrections such as robust standard errors would further strengthen the analysis. Additionally, exploring alternative model specifications (e.g., using different link functions or distribution families) could offer further insights.

Overall, this analysis provides actionable insights for the retailer, emphasizing that a well-constructed email campaign can be an effective tool in driving customer spending.

# Appendix

## Appendix 1: Robust Standard Error Outputs

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

## Appendix 2: Python Code

(or view it on [GitHub here](https://github.com/burnett013/ISM6137_Assignment_M4))

"""

Your job is to tell the retailer if the Mens or Womens e-mail campaign was successful.

This script:

1. Imports and cleans the data.

2. Explores the skewness of the 'recent\_spend' variable.

3. Applies a log transformation.

4. Constructs a hypothesis table.

5. Encodes categorical variables and engineers new features.

6. Fits three GLM (Gamma) models.

7. Applies robust standard errors to correct for heteroskedasticity.

8. Uses Stargazer to summarize results.

"""

# --- 1. Import Toolbox ---

import pandas as pd

import numpy as np

import statsmodels.api as sm

import statsmodels.formula.api as smf

import statsmodels.stats.sandwich\_covariance as sw

import seaborn as sns

import matplotlib.pyplot as plt

import scipy.stats as stats

from stargazer.stargazer import Stargazer

from sklearn.preprocessing import StandardScaler

# --- 2. Import and Clean the Data ---

df = pd.read\_excel("OnlineRetailCampaign-2.xlsx", sheet\_name="Data")

df.columns = df.columns.str.lower()

df = df.apply(lambda x: x.str.lower() if x.dtype == "object" else x)

# Rename columns

df.rename(columns={

'recency': 'last\_purchased',

'history': 'past\_spending',

'mens': 'purchased\_male',

'womens': 'purchased\_female',

'zipcode': 'zipcode\_cat',

'channel': 'purchase\_channel',

'campaign': 'email\_type',

'visit': 'recent\_web\_visit',

'conversion': 'was\_converted',

'spend': 'recent\_spend'

}, inplace=True)

# Change data types

df = df.astype({

"last\_purchased": int,

"past\_spending": float,

"purchased\_male": int,

"purchased\_female": int,

"newcustomer": int,

"recent\_web\_visit": int,

"was\_converted": int,

"recent\_spend": float

})

df["zipcode\_cat"] = df["zipcode\_cat"].astype("category")

df["purchase\_channel"] = df["purchase\_channel"].astype("category")

df["email\_type"] = df["email\_type"].astype("category")

# Standardize email campaigns

df['email\_type'] = df['email\_type'].replace({

'womens e-mail': 'female',

'mens e-mail': 'male',

'no e-mail': 'none'

})

# Create spending categories

bins = [0, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1600, 2000, 3000, 4000]

labels = list(range(1, len(bins)))

df['spending\_category\_yr'] = pd.cut(df['past\_spending'], bins=bins, labels=labels, right=True, include\_lowest=True)

df['spending\_category\_yr'] = df['spending\_category\_yr'].astype(int)

# Create a subset of converted customers

df\_converted = df[df["was\_converted"] == 1].copy()

df\_converted.drop(columns=["was\_converted"], inplace=True)

df\_converted.to\_csv("converted\_only.csv", index=False)

df\_converted = pd.read\_csv("converted\_only.csv")

# --- 3. Evaluate Skewness of Recent Spending ---

skew\_value = df\_converted['recent\_spend'].skew()

plt.figure(figsize=(10, 6))

sns.histplot(df\_converted['recent\_spend'], bins=30, kde=True)

plt.title("Histogram of Recent Spending Among Converted Customers")

plt.show()

plt.figure(figsize=(8, 6))

stats.probplot(df\_converted['recent\_spend'], dist="norm", plot=plt)

plt.title("Q-Q Plot for Recent Spend")

plt.show()

# --- 4. Apply Log Transformation ---

df\_converted['recent\_spend\_log'] = np.log1p(df\_converted['recent\_spend'])

skew\_value\_log = df\_converted['recent\_spend'].skew()

plt.figure(figsize=(10, 6))

sns.histplot(df\_converted['recent\_spend\_log'], bins=30, kde=True)

plt.title("Histogram of Recent Spending Among Converted Customers After Log Transformation")

plt.show()

plt.figure(figsize=(8, 6))

stats.probplot(df\_converted['recent\_spend\_log'], dist="norm", plot=plt)

plt.title("Q-Q Plot for Recent Spend After Log Transformation")

plt.show()

# --- 5. Create a Hypothesis Table ---

hypothesis\_data = {

"Predictor": [

"Last Purchased",

"Spending Category (Year)",

"Purchase Channel",

"Email Type",

"New Customer",

"Average Spending",

"Last Purchased^2",

"Email \* New Customer Interaction"

],

"Effect (+/-)": [

"-",

"+",

"?",

"+",

"+/-",

"+",

"?",

"?"

],

"Rationale": [

"Customers who purchased more recently are more likely to spend; a larger value indicates a longer gap, which may reduce spending.",

"Customers with higher historical spending are likely to continue spending more.",

"Effect depends on user experience and channel preference (phone vs. web).",

"Email campaigns (male or female) are hypothesized to increase spending vs. no email.",

"New customers may spend more (enthusiasm) or less (hesitation).",

"Higher average spending in the past signals a greater propensity to spend now.",

"A squared term can capture potential non-linear effects of recency.",

"Interaction may reveal different responses to email among new customers."

]

}

hypothesis\_df = pd.DataFrame(hypothesis\_data)

hypothesis\_df.to\_excel("hypothesis\_table\_v5.xlsx", index=False)

display(hypothesis\_df)

# --- 6. One-Hot Encode Categorical Variables ---

categorical\_columns = ['zipcode\_cat', 'purchase\_channel', 'email\_type']

df\_converted\_encoded = pd.get\_dummies(df\_converted, columns=categorical\_columns, drop\_first=False)

# --- 7. Feature Engineering ---

df\_converted\_encoded["email\_newcustomer"] = df\_converted\_encoded["email\_type\_male"] \* df\_converted\_encoded["newcustomer"]

df\_converted\_encoded["avg\_spending"] = df\_converted\_encoded["past\_spending"] / df\_converted\_encoded["last\_purchased"]

# --- 8. Scale Numerical Features ---

scaler = StandardScaler()

df\_converted\_encoded[['last\_purchased', 'past\_spending']] = scaler.fit\_transform(

df\_converted\_encoded[['last\_purchased', 'past\_spending']]

)

# Polynomial term for non-linear relationships

df\_converted\_encoded["last\_purchased\_sq"] = df\_converted\_encoded["last\_purchased"] \*\* 2

# --- 9. GLM Models (Gamma) with Different Predictor Combinations ---

glm\_model\_1 = smf.glm(

formula="recent\_spend ~ last\_purchased + email\_type\_female + email\_type\_male + email\_type\_none + email\_newcustomer + avg\_spending",

data=df\_converted\_encoded,

family=sm.families.Gamma(link=sm.families.links.log())

).fit()

glm\_model\_2 = smf.glm(

formula="recent\_spend ~ last\_purchased + email\_type\_female + email\_type\_male + email\_type\_none + purchase\_channel\_phone + purchase\_channel\_web + email\_newcustomer + avg\_spending",

data=df\_converted\_encoded,

family=sm.families.Gamma(link=sm.families.links.log())

).fit()

glm\_model\_3 = smf.glm(

formula="recent\_spend ~ last\_purchased + email\_type\_female + email\_type\_male + email\_type\_none + purchase\_channel\_phone + purchase\_channel\_web + newcustomer + spending\_category\_yr + email\_newcustomer + avg\_spending",

data=df\_converted\_encoded,

family=sm.families.Gamma(link=sm.families.links.log())

).fit()

# Display GLM model summaries

print("==== GLM Model 1 ====")

print(glm\_model\_1.summary())

print("\n==== GLM Model 2 ====")

print(glm\_model\_2.summary())

print("\n==== GLM Model 3 ====")

print(glm\_model\_3.summary())

# --- 10. Heteroskedasticity Corrections (Robust SE) ---

robust\_cov\_1 = sw.cov\_white\_simple(glm\_model\_1)

glm\_model\_1.normalized\_cov\_params = None

glm\_model\_1.cov\_params\_default = robust\_cov\_1

robust\_cov\_2 = sw.cov\_white\_simple(glm\_model\_2)

glm\_model\_2.normalized\_cov\_params = None

glm\_model\_2.cov\_params\_default = robust\_cov\_2

robust\_cov\_3 = sw.cov\_white\_simple(glm\_model\_3)

glm\_model\_3.normalized\_cov\_params = None

glm\_model\_3.cov\_params\_default = robust\_cov\_3

print("\n==== GLM Model 1 (Robust SE) ====")

print(glm\_model\_1.summary())

print("==== GLM Model 2 (Robust SE) ====")

print(glm\_model\_2.summary())

print("==== GLM Model 3 (Robust SE) ====")

print(glm\_model\_3.summary())

# --- 11. Aggregate and Evaluate Models Using Stargazer ---

stargazer\_glm = Stargazer([glm\_model\_1, glm\_model\_2, glm\_model\_3])

with open("glm\_model\_results.html", "w") as f:

f.write(stargazer\_glm.render\_html())

stargazer\_glm