U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Applied Statistics

Final Project

Miguel Bourgin Gonçalves
nº202107127

Index

Abstract

This project aims to apply statistical methods to analyze specific data and extract relevant information. A quantitative approach will be used, focusing on descriptive and inferential analysis. The study will cover the methodology used for data analysis and model evaluation, followed by the presentation of the results obtained, an interpretation and discussion of the results, and the final conclusions.

1 Introduction

1.1 Linear Regression

Linear regression is a method used to model the relationship between a continuous response variable Y and one or more predictor variables X. It assumes a linear relationship, where the response is predicted as a weighted sum of the predictors, with the goal of minimizing the difference between the predicted and actual values of Y.

Simple Linear Regression

Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y. Mathematically, we can write this linear relationship as

$$Y \approx \beta_0 + \beta_1 X. \tag{1.1}$$

In equation (1.1), β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, β_0 and β_1 are known as the *model coefficients* or *parameters*. Once we have used our training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we can predict future values of Y on the basis of a particular value of X by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{1.2}$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$.

Multiple Linear Regression

Simple linear regression is a useful approach to predicting a response on the basis of a single predictor variable. However, in practice, we often have more than one predictor. Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model.

In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{1.3}$$

where X_j represents the j-th predictor and β_j quantifies the association between that variable and the response. We interpret β_j as the *average effect* on Y of a one-unit increase in X_j, holding all other predictors fixed.

1.2 Logistic Regression

Logistic regression is used to model the probability of a binary outcome Y based on predictor variables X. Unlike linear regression, which can yield probabilities outside the range [0, 1], logistic regression uses the logistic function to constrain the predicted probabilities:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad \text{where} \quad p(X) = \Pr(Y = 1|X)., \tag{1.3}$$

The function (1.3) produces an S-shaped curve, ensuring the probabilities are always between 0 and 1. The odds are defined as:

$$\text{Odds} = \frac{p(X)}{1-p(X)}, \tag{1.4}$$

and taking the logarithm of the odds leads to the log-odds (logit):

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X., \tag{1.5}$$

Here, β_1 represents the change in log-odds for a one-unit increase in X, and e^{β_1} is the factor by which the odds change. The logistic regression model is estimated using maximum likelihood estimation to ensure valid probability predictions.

Ultimately, logistic regression provides a flexible approach for modeling binary outcomes, ensuring valid probabilities through the logistic function and offering interpretability via odds and log-odds transformations.

Multinomial Logistic Regression

For a response variable with more than two categories, we use **multinomial logistic regression**. This model generalizes binary logistic regression by comparing the odds of each category against a baseline category. The probability of each category is modeled as:

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \tag{1.6}$$

For the baseline category $Y = K$, the probability is:

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \tag{1.7}$$

The log-odds of being in class k versus the baseline class K is given by:

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p. \tag{1.8}$$

Note that the choice of baseline class does not affect the fitted values or predictions, but it does affect the interpretation of the coefficients.

Alternatively, the softmax coding can be used, where the probabilities for all K categories are modeled symmetrically:

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

This method estimates coefficients for all categories and calculates the log-odds between any pair of categories.

2 Model Evaluation and Interpretation

Linear Regression

In multiple linear regression, the primary objectives are to assess the relationship between the response variable Y and the predictors $X_1, X_2, \ldots, X_p$, identify important predictors, evaluate model fit, and make predictions with quantifiable uncertainty.

- **Relationship Between Response and Predictors:** We test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ to determine if any predictors are related to Y. This is done using the F-statistic:

  $$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

  A significant F-statistic (small p-value) indicates that at least one predictor is useful.

- **Identifying Important Variables:** Variable selection methods, such as forward, backward, and mixed selection, are used to identify significant predictors, balancing model simplicity and predictive power.
- **Model Fit:** The model fit is evaluated using R^2, which shows the proportion of variance explained, and the Residual Standard Error (RSE), which estimates model accuracy. The Adjusted R^2 accounts for the number of predictors.
- **Predictions and Uncertainty:** Predictions are made based on the fitted model, with uncertainty quantified by confidence intervals for coefficients and predictions. Model bias and irreducible error are considered as sources of prediction uncertainty.

Logistic Regression

- **Relationship Between Response and Predictors:** In logistic regression, the relationship between the response variable Y and the predictor variables is modeled using the logistic function. The significance of the predictors is evaluated using the Wald z-statistic, which tests the null hypothesis

  $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

  If the p-values are small, it indicates that the corresponding predictor variables are significantly related to the outcome.
- **Identifying Important Variables:** In logistic regression, the significance of each predictor is assessed using the p-value associated with the coefficient. A low p-value (typically < 0.05) suggests that the predictor has a significant relationship with the response variable. Using variable selection techniques, such as backward elimination, could help identify the most important predictors.
- **Model Fit:** The model fit is evaluated using the Deviance and AIC (Akaike Information Criterion). The Deviance compares the fitted model with a baseline model (null model). Lower deviance values suggest a better model fit. The AIC helps in comparing different models, where a lower AIC value indicates a better fit while considering model complexity.
- **Predictions and Uncertainty:** Predictions from the logistic regression model are made using the `predict()` function, which provides the probabilities of the event. These probabilities can be converted into class labels by setting a threshold. The uncertainty in predictions is measured through the standard errors of the coefficients. A model with higher accuracy will result in fewer misclassifications and a higher correct prediction rate.

3 Methodology

Linear Regression

To fit a multiple linear regression model using least squares, we use the `lm()` function in R. The model is specified as:

$$y \sim x_1 + x_2 + \cdots + x_n$$

where $x_1, x_2, \ldots, x_n$ are the predictors. The `summary()` function is then used to output the regression coefficients for all the predictors.

We can access individual components of the summary object by name. For instance, the R^2, and `summary(lm.fit)$sigma` are given by `summary(lm.fit)$r.sq` and provides the residual standard error (RSE). Additionally, the `vif()` function from the `car` package is used to compute the variance inflation factors, which help to assess multicollinearity.

Logistic Regression

To fit a logistic regression model, we use the glm() function in R with the binomial family, which specifies that the response variable is binary. The model is expressed as:

$$\text{Response} \sim \text{Predictor}_1 + \text{Predictor}_2 + \cdots + \text{Predictor}_p$$

This model uses maximum likelihood estimation to estimate the coefficients for the predictors. After fitting the model, the summary() function provides the coefficients and their associated p-values, which indicate the significance of each predictor.

We use the predict() function to generate predicted probabilities, which can be converted to class labels by applying a threshold (typically 0.5).

To evaluate the model, we compute a confusion matrix by comparing the predicted labels with the actual labels and calculate the accuracy. The model's performance is evaluated on a test set to ensure it generalizes well and avoids overfitting.

4 Data Pre-Processing

The dataset used in this analysis is the **tips dataset** from the reshape2 package in R. This dataset provides information on restaurant tips, including variables such as the total bill, table size, day and time of day, and customer demographics (e.g., sex, smoking status).

A log transformation was applied to the tip and total_bill variables to address potential skewness and improve the interpretability of the regression model. The transformed variables are now referred to as log_tip and log_total_bill.

Apart from the log transformation, no other modifications were made to the dataset. The categorical variables (e.g., sex, day, smoker, time) were already encoded as factors, and the numerical variables (size, log_tip, log_total_bill) were in the correct format for analysis.

For the Logistic Regression part the response variable (log_tip) is going to be transformed, separating values greater than or equal to its median ('1') from the lower values ('0').

5 Exploratory Data Analysis

Statistical, Numerical, and Graphical Description of the Data

This section provides a statistical summary of the key variables in the dataset. The focus is on the log-transformed tip (log_tip), log-transformed total bill (log_total_bill), and table size (size).

The summary statistics are presented in Table 1.

Table 1: Summary Statistics of Key Variables

Variable	**Mean**	**Standard Deviation**	**Range (Min–Max)**
Log Tip (log_tip)	1.00	0.44	0.00–2.30
Log Total Bill (log_total_bill)	2.89	0.44	1.12–3.93
Size (size)	2.57	0.95	1.00–6.00

The categorical variables are summarized below:

- **Sex:** 87 Female, 157 Male
- **Smoker:** 151 Non-Smoker, 93 Smoker
- **Day:** 19 Sunday, 87 Thursday, 76 Tuesday, 62 Friday
- **Time:** 176 Dinner, 68 Lunch

The data reveals that log total bill values (mean = 2.89) are generally higher than log tip values (mean = 1.00), with significant variability (SD = 0.44 for both). The wide ranges suggest diverse tipping and bill behaviors. Table size shows moderate variation, while differences in categorical variables like sex and smoker status likely contribute to tipping differences.

Now we will see some graphical visualizations to explore the distribution of key variables and relationships in the dataset.

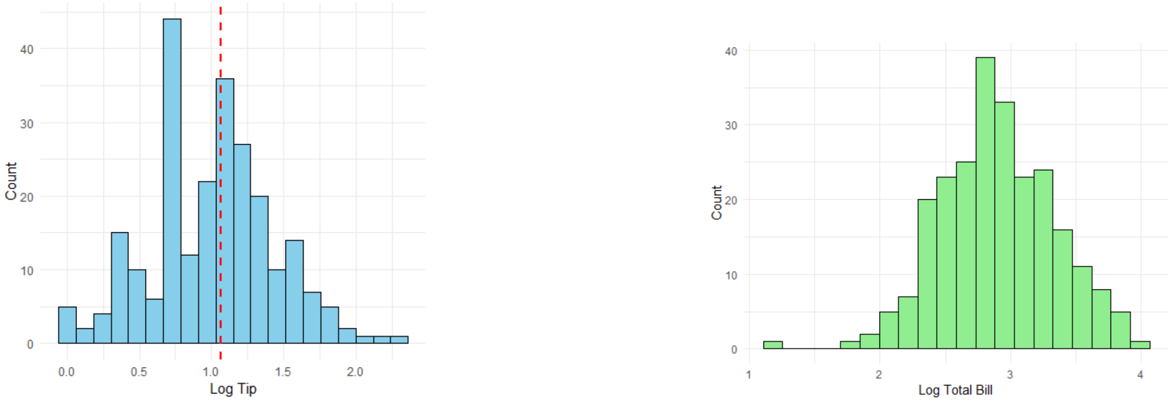

Figure 1: Histograms of Log Tip Amount and Log Total Bill

Figure 1 shows the histograms for the log-transformed tips (`log_tip`) with a vertical line representing the median and the log-transformed total bill (`log_total_bill`).The first histogram of log-transformed tips shows a skewed distribution with a peak around 1, indicating most tips are moderate but there are a few outliers with very low or high tips. The second histogram of log-transformed total bills reveals a more balanced distribution, with values centered around 3, suggesting most customers tend to have higher bills, with fewer extreme low or high values.

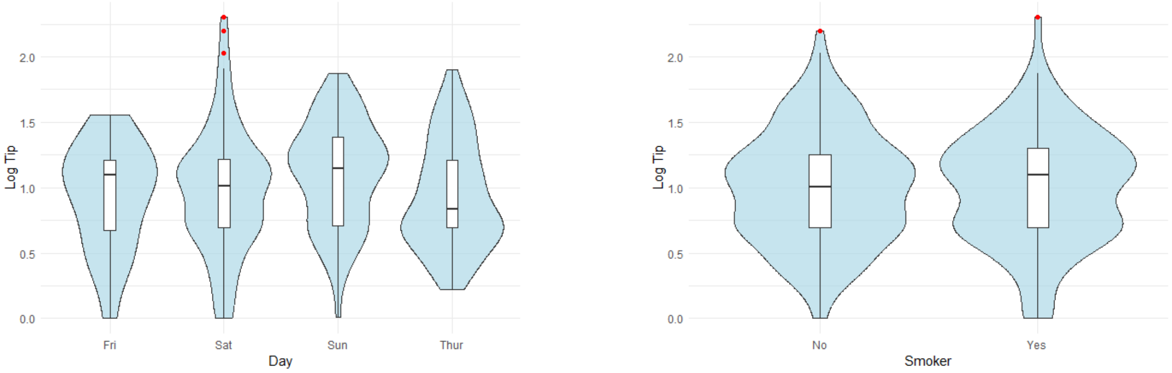

Figure 2: Violin Plots of Log Tip Amount by Day and Smoker

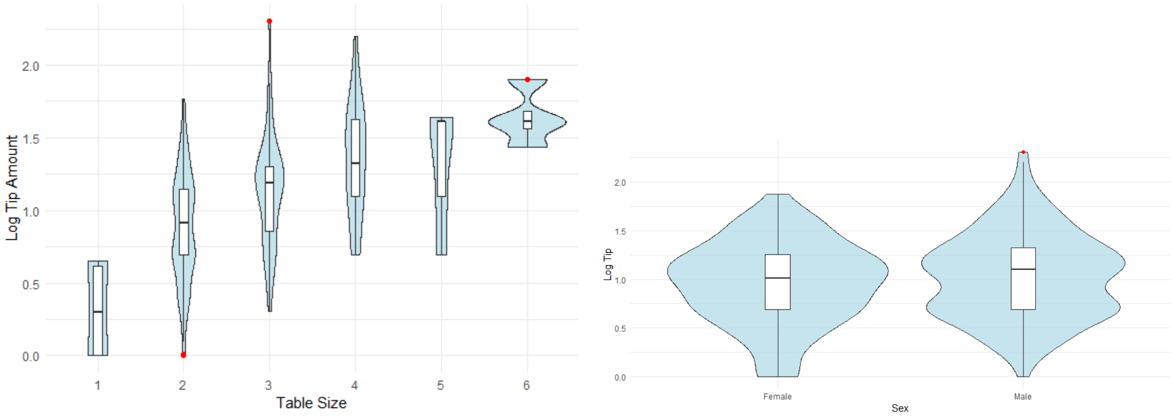

Figure 3: Violin Plots of Log Tip Amount by Size and Sex

Violin plots were used to explore the distribution of tips across different categorical variables. Figure 2 and figure 3 shows the distributions of log_tip by day, smoker, and table size (size).Larger table sizes generally show bigger in log tips. Both smoking and sex status have minimal impact on tipping, though non-smokers and male exhibit slightly higher median tips. Across days, tipping is most variable on Saturdays, with more high-value outliers, while Thursdays have the most consistent behavior. The log transformation effectively normalizes tip amounts, highlighting trends across these features.

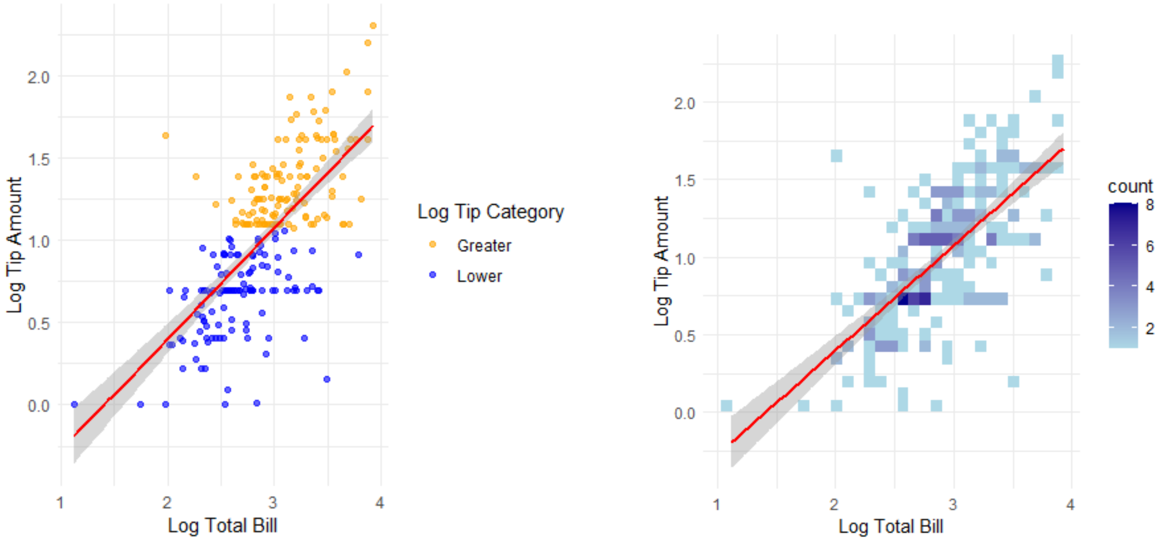

Figure 4: Visualizations of Log Tip Amount by Log Total Bill

The first plot in Figure 4 visualizes the log tip amount categorized into lower and greater than the median, with the points colored accordingly and the regression line illustrating the relationship for both categories. The second plot shows a heatmap of the log-transformed total bill and tip amount, with a linear regression line (in red) highlighting the trend. Both plots imply that there is a correlation between this two variables.

6 Model Selection Process

The model selection process was carried out iteratively to identify the most suitable predictors for explaining the variation in the response variable (tip). Several models were evaluated using statistical metrics, as discussed in Section 2. . The following is the step-by-step process:

Linear Model

Model	Formula	R^2	Adj. R^2	F-statistic	Residuals (Range)	AIC
Initial	tip ∼ total_bill + sex + day + smoker + size + time	0.4701	0.4520	26.06	$[-2.85, 4.11]$	715.0
Model 1	tip ∼ total_bill + size + sex + time + total_bill*size + total_bill*time	0.4710	0.4576	35.17	$[-2.85, 4.14]$	710.5
Model 2	tip ∼ total_bill + size + time + total_bill*size + total_bill*time	0.4710	0.4598	42.37	$[-2.86, 4.14]$	708.6
Model 3	tip ∼ total_bill + time + size + total_bill*time	0.4703	0.4614	53.05	$[-2.90, 4.14]$	706.8
Model 4	tip ∼ total_bill + size	0.4679	0.4635	105.9	$[-2.93, 4.04]$	704.0
Model 5	log_tip ∼ log_total_bill + size	0.4717	0.4674	107.6	$[-1.18, 1.22]$	138.8

Table 2: Summary of Models with Key Metrics and AIC

We began with an initial model including all predictors, assessing significance and fit. Interaction terms were added in Model 1, but non-significant variables like sex were removed in Model 2. In Model 3, the interaction size*total_bill was excluded due to irrelevance. Model 4 simplified further by including only total_bill and size. Finally, Model 5 applied log-transformation to stabilize variance and improve interpretability. Each step refined the model by removing insignificant terms or enhancing interpretability while maintaining statistical rigor, leading to a final, simpler model that effectively explained tip variations with only two variables: log_total_bill and size. Addicionaly, we saw that AIC for Model 5 was the lowest.

Logistic Regression Models

Model	Formula	Pseudo R^2	Residual Deviance	Null Deviance	Residuals (Range)	AIC
Initial	log_tip_category ∼ log_total_bill + size + sex + smoker + time	0.10	240.50	338.26	$[-2.85, 2.14]$	252.5
Model 1	log_tip_category ∼ log_total_bill + size + sex + time	0.11	240.83	338.26	$[-2.85, 2.14]$	250.83
Model 2	log_tip_category ∼ log_total_bill + size + time	0.11	241.37	338.26	$[-2.86, 2.14]$	249.37
Model 3	log_tip_category ∼ log_total_bill + time	0.12	241.96	338.26	$[-2.90, 2.14]$	247.96
Model 4	log_tip_category ∼ log_total_bill + log_total_bill*time	0.13	237.04	338.26	$[-2.93, 2.14]$	245.04
Model 5	log_tip_category ∼ log_total_bill + factor(day) + log_total_bill * time	0.14	231.63	338.26	$[-2.85, 2.12]$	245.63
Model 6	log_tip_category ∼ log_total_bill * factor(day) + time	0.15	226.19	338.26	$[-2.85, 2.13]$	244.19
Model 7	log_tip_category ∼ log_total_bill * factor(day)	0.13	227.14	338.26	$[-2.86, 2.13]$	243.14

Table 3: Summary of Logistic Regression Models with Key Metrics and AIC

We began with an initial logistic regression model that included all predictors: log_total_bill, size, sex, smoker, and time. In **Model 1**, we removed smoker, as it was not statistically significant. In **Model 2**, the sex variable was excluded for similar reasons, leaving log_total_bill, size, and time. **Model 3** further simplified the model by removing size, leaving only log_total_bill and time. Finally, in **Model 4**, we introduced an interaction term between log_total_bill and time.

In **Model 5**, we included factor(day) to account for day-of-week variations and added an interaction term with log_total_bill. **Model 6** further extended the interaction model by combining log_total_bill, factor(day), and time. **Model 7** focused solely on the interaction between log_total_bill and factor(day).

Each step in model refinement aimed to reduce complexity while maintaining significant predictors. **Model 6** showed the best overall performance, balancing model fit and complexity, with the lowest AIC and a higher pseudo R^2. However, despite the improvements, the models still perform poorly, likely due to the simplification of the response variable into a binary category, which may not fully capture the variability in the data.

7 Analysis of Variables

Linear Regression: Continuous Variable (log_total_bill) and Categorical Variable (factor(size)

Raw and Adjusted Effects of log_total_bill and size

The raw effect of log_total_bill on the response variable is significant, with an estimated coefficient of 0.675 with a very high significance level, indicating that as log_total_bill increases by one unit, the response variable increases by approximately 67.5%. While the adjusted effect of log_total_bill, controlling for other variables, is 0.602 and remains significant.

Category	Raw Estimate	Raw p-value	Adjusted Estimate	Adjusted p-value
2	0.5640	0.00386	0.05346	0.754
3	0.8271	5.19e-05	0.11170	0.542
4	1.0315	5.88e-07	0.17584	0.353
5	1.0154	9.70e-05	0.11860	0.611
6	1.3249	1.70e-06	0.34020	0.170

Table 4: Raw and Adjusted Effects for size

As we can see in Table 4 the raw effects of size categories on the response variable are significant, with increasing magnitudes across categories, indicating meaningful associations. However, after adjusting for other variables, none of the adjusted effects are statistically significant, suggesting that the raw associations may be influenced by confounding factors. While Category 6 shows the largest raw effect (1.3249), the adjusted effect is much smaller and non-significant ($p = 0.170$). These findings imply that size categories are not independent predictors of the response variable after accounting for other factors in the model. Additionally, this suggests that a linear model may not be the best approach for this dataset.

Effect of Changing Categories in Size

The effect of changing from the third category of size to the second category is estimated to be -0.058, indicating a small decrease in the response variable. However, the 95% confidence interval for this effect is $[-0.548, 0.431]$, and the 90% confidence interval is $[-0.469, 0.353]$. Since both confidence intervals include zero, we conclude that the change from the third to the second size category is not statistically significant.

Investigation of Interaction Between Log Total Bill and Size

The interaction terms between `log_total_bill` and `size` were found to be statistically insignificant (all $p > 0.05$), suggesting that the relationship between `log_total_bill` and `log_tip` does not substantially vary across different size categories. This is further confirmed by an ANOVA comparison, which shows no significant improvement in model fit when including interaction terms ($p = 0.551$). Additionally, the adjusted R^2 (0.4577) and the residual standard error (0.3212) indicate minimal differences between models with and without the interaction terms. Thus, the simpler additive model, without interactions, is sufficient to describe the data and provides a more straightforward interpretation without sacrificing explanatory power.

Logistic Regression: Continuous Variable (log_total_bill), Categorical Variable (factor(day))

Raw and Adjusted Effects of Log Total Bill and Day of the Week

The raw effect of log total bill on the likelihood of receiving a tip is highly significant, with an estimated coefficient of 3.9430 ($p < 0.001$). This result suggests that for every one-unit increase in the log total bill, the odds of receiving a tip increase substantially. The adjusted effect of log total bill remains equally strong and significant, with a similar coefficient of 3.3376 ($p < 0.001$). This confirms a robust positive relationship between the total bill and the likelihood of tipping, even after accounting for the day of the week.

Category	Raw Estimate	Raw p-value	Adjusted Estimate	Adjusted p-value
Saturday	-0.5351	0.123	-1.4708	0.0598
Sunday	-0.3245	0.215	-0.5021	0.311
Thursday	-0.2504	0.315	-0.1320	0.511

Table 5: Raw and Adjusted Effects for Day of the Week

As shown in Table 5, the raw effects for day of the week are not significant. After adjusting for the other variables on the model, only the effect for Saturday approaches significance. This suggests that tipping behavior on Saturday may slightly differ from Friday, though the evidence is marginal. The effects of Sunday and Thursday remain non-significant in the adjusted model, indicating no substantial differences in tipping likelihood on these days compared to Friday.

Effect of Changing Categories in Day of the Week

The effect of changing from the third category of `day` (Sunday) to the second category (Saturday) is estimated to be 0.0864, indicating a small increase in the likelihood of receiving a tip. However, the 95% confidence interval for this effect is $[-2.4247, 2.5974]$, and the 90% confidence interval is $[-2.0211, 2.1939]$. Since both confidence intervals include zero, we conclude that the change from Sunday to Saturday is not statistically significant.

Investigation of Interaction Between Log Total Bill and Day of the Week

The interaction terms between log total bill and day of the week were found to be statistically significant for the model comparison ($p = 0.01063$), suggesting that the relationship between log total bill and the likelihood of receiving a tip does indeed vary across different days of the week.

The coefficients for Saturday, Sunday, and Thursday are -10.190, -9.642, and -5.468, respectively, with p-values of 0.1304, 0.1534, and 0.4349. These results indicate no significant effects for these days compared to Friday.

However, the ANOVA comparison between the models with and without the interaction terms yields a p-value of 0.01063, which is less than 0.05. This indicates that the inclusion of interaction terms significantly improves the model fit, suggesting that the effect of log total bill on the likelihood of receiving a tip does vary depending on the day of the week, justifying the inclusion of the interaction terms in the model. Additionally, the Akaike Information Criterion (AIC) for the model with interaction is 243.14, compared to 248.35 for the simpler additive model.

The inclusion of interaction terms between log total bill and day of the week improves model fit, as indicated by the ANOVA ($p = 0.01063$) and the lower residual deviance and AIC. Although the coefficients for individual days were not significant, the interaction model offers a better fit overall and is preferred over the simpler model.

8 Conclusion

The linear regression analysis demonstrated that the log-transformed total bill and table size are the most important predictors of tipping behavior. The model refinement process revealed that including interaction terms did not significantly improve the model's performance. The final model, which used only log-transformed total bill and size as predictors, had the lowest AIC and provided a straightforward explanation of the data. However, adjusted effects for table size categories were not significant, suggesting confounding factors influence the raw associations. Interaction terms between log total bill and size were also insignificant, confirming that a simpler additive model was sufficient. The linear model highlights a strong positive association between log total bill and tip amounts but suggests limited predictive power for other variables like sex, smoker status, and day of the week.

The logistic regression analysis explored the relationship between log-transformed total bill and tipping likelihood, categorized into high or low tips. The final logistic model, which incorporated interaction terms between log total bill and day of the week, showed the best fit with the lowest AIC and highest pseudo R^2. While individual effects for day categories were not significant, the interaction terms significantly improved model fit, indicating that tipping likelihood varies across days depending on the total bill. The binary classification approach, however, may have oversimplified the variability in tipping behavior, potentially limiting model performance. Nonetheless, the results underscore the robust positive influence of total bill on tipping likelihood, with nuanced variations by day.

Summary

The linear model effectively captures the continuous relationship between tipping amounts and predictors, while the logistic model provides insights into tipping likelihood and its interactions with categorical variables. The linear model is preferred for detailed predictions of tip amounts, whereas the logistic model highlights trends in tipping behavior. Both models underscore the importance of log-transformed total bill as the dominant predictor but suggest limited utility for other factors like table size, sex, and smoker status.

References

[1] An Introduction to Statistical Learning. Available: https://www.statlearning.com/

[2] Classification Metrics Walkthrough. [Online]. Available: https://www.kdnuggets.com/2022/10/classification-metrics-walkthrough-logistic-regression-accuracy-precision-recall-roc.html

[3] Linear and Logistic Regression Slides from Moodle.