BUSINESS DATA ANALYTICS WITH CATEGORICAL AND CENSORED OUTCOMES

Individual Assignment

Analysis of Roulette Casino Game Outcomes Using Econometric Models

Submitted to:

Prof. Gulasekaran Rajaguru

Submitted by:

Thariq Mohammed Z
2301382
PGPM 2023-2025

Submitted on:

14/01/2025

1. INTRODUCTION

The complexity of analyzing gambling data lies in understanding the underlying patterns that drive outcomes and probabilities. Roulette, a classic casino game, provides a fascinating dataset to investigate using statistical methods due to its defined set of rules and outcomes. By examining roulette outcomes, this study aims to delve deeper into the mechanics of limited dependent variable models—an area crucial for fields ranging from behavioural economics to decision sciences.

Limited dependent variable models address situations where the outcomes are inherently restricted, such as binary outcomes, ordinal scales, or count data. They are particularly suited for analyzing scenarios where the dependent variable does not follow the traditional continuous data structure, like determining whether a bet wins or identifying the range of numbers in roulette. These models help reveal relationships between predictors and outcomes in such restricted datasets, providing insights that can influence strategies and decision-making.

This project leverages simulated roulette game data consisting of 100,000 rounds to evaluate several econometric models, including binary choice models (logit and probit), count models (Poisson), ordered models, and multinomial regressions. The primary objective is to identify the factors influencing betting outcomes and evaluate the predictive power of these models. The findings contribute to a better understanding of how such models can be applied to structured, rule-based datasets, offering insights not only into roulette but also into broader applications like risk analysis and behavioural predictions. The analysis, conducted using R, combines theoretical rigor with practical visualization to create a comprehensive exploration of the data.

2. REVIEW OF LITERATURE

2.1 Definition and Significance of Limited Dependent Variable Models

Limited dependent variable models address situations where the outcome variable is constrained in some way, such as being binary, ordinal, or count-based. These models are crucial in fields like finance, healthcare, and behavioural economics, where outcomes often exhibit such constraints.

2.2 Binary Outcomes: Logit and Probit Models

Logit and probit models are the primary tools for binary outcome analysis. They offer flexibility in estimating probabilities and account for non-linear relationships. The logit model uses a logistic distribution, while the probit model assumes a standard normal distribution, making them suitable for different scenarios.

2.3 Count Data Models: Poisson and Negative Binomial

Count data models such as Poisson and Negative Binomial are employed when outcomes are discrete and non-negative. Poisson models assume equal mean and variance, while Negative Binomial models relax this assumption, addressing over-dispersion effectively.

2.4 Ordered Outcomes: Ordered Logit and Probit Models

Ordered regression models are designed for ordinal dependent variables. These models rank outcomes while preserving their inherent order, making them effective for questions involving preferences or classifications.

2.5 Multinomial Models: Logit and Probit Extensions

Multinomial models extend binary choice frameworks to multiple unordered outcomes. These models estimate the probability of each outcome category, facilitating comprehensive analysis of categorical decisions.

2.6 Challenges in Limited Dependent Variable Models

Issues like heteroskedasticity, endogeneity, and multicollinearity can bias estimations. Advances in diagnostics and robust estimation techniques are essential to mitigate these challenges.

3. OBJECTIVE

The primary objective of this study is to explore the factors influencing betting outcomes in roulette using advanced econometric models. Specific questions addressed include:

1. **What factors influence the probability of a bet on red winning?**
   This question examines the predictors for a successful red bet, focusing on the range of winning numbers and their even or odd classification.

2. **Is there a difference in the likelihood of winning a bet on even numbers compared to odd numbers?**

   This aspect evaluates the comparative probabilities of winning based on the parity of numbers.

3. **How does the number of wins vary across different colours or number ranges in a given period?**

   The analysis identifies trends and variations in wins categorized by colour and numerical range.

4. **Does the range of winning numbers (low, medium, high) depend on previous round results?**

   This question explores the influence of past outcomes on the range of current winning numbers.

5. **What factors influence the probability of a specific colour (red, black, or zero) being the winner?**

   The focus here is on understanding the determinants of winning colour probabilities.

6. **Can we predict the winning number given that it falls within a specific range (e.g., greater than 18)?**

   This prediction assesses the feasibility of estimating outcomes within predefined ranges.

7. **How do the models compare in terms of predictive accuracy for different dependent variables?**

   This analysis involves evaluating the performance and accuracy of the applied econometric models.

8. **Are assumptions like homoscedasticity and independence satisfied?**

   This question addresses the validity of statistical assumptions underlying the chosen models.

4. DATA AND METHODOLOGY

4.1 Dataset Description

The dataset comprises 100,000 observations of roulette game outcomes, with the following variables:

- **Round**: Sequential identifier for each game round.
- **Winning Number**: The number on which the roulette ball landed (1-36, with 0 as a special case).
- **Winning Colour**: The colour associated with the winning number (Red, Black, or Green).
- **Red Bet Win**: Indicator (1/0) for whether a bet on red won.
- **Black Bet Win**: Indicator (1/0) for whether a bet on black won.
- **Even Bet Win**: Indicator (1/0) for whether a bet on even numbers won.
- **Odd Bet Win**: Indicator (1/0) for whether a bet on odd numbers won.
- **Zero Bet Win**: Indicator (1/0) for whether a bet on zero won.

All variables are complete with no missing values. The dataset allows exploration of betting probabilities, winning outcomes, and their determinants.

4.2 Methodology

4.2.1 Binary Choice Models

To examine factors influencing the probability of a bet on red winning, logistic and probit regression models were applied. The dependent variable is Red Bet Win, and independent variables include categorized Winning Number (Low, Medium, High) and Even/Odd classification.

4.2.2 Count Data Models

The variation in wins across colours was analyzed using Poisson regression. Aggregated counts of wins by Winning Colour were modelled to identify trends and associations.

4.2.3 Ordered Regression Models

To determine whether the range of winning numbers depends on previous outcomes, ordered logit and probit models were employed. The independent variable was the lagged range of winning numbers.

4.2.4 Multinomial Regression Models

The probability of a specific colour (Red, Black, or Green) winning was modelled using multinomial logit and probit regressions. Previous round outcomes served as predictors.

4.2.5 Truncated and Censored Regression Models

Truncated regression explored predictions for winning numbers greater than 18. Censored regression (Tobit model) was applied to assess winning numbers within specified ranges.

5. DISCUSSION OF RESULTS

5.1 Binary Choice Models

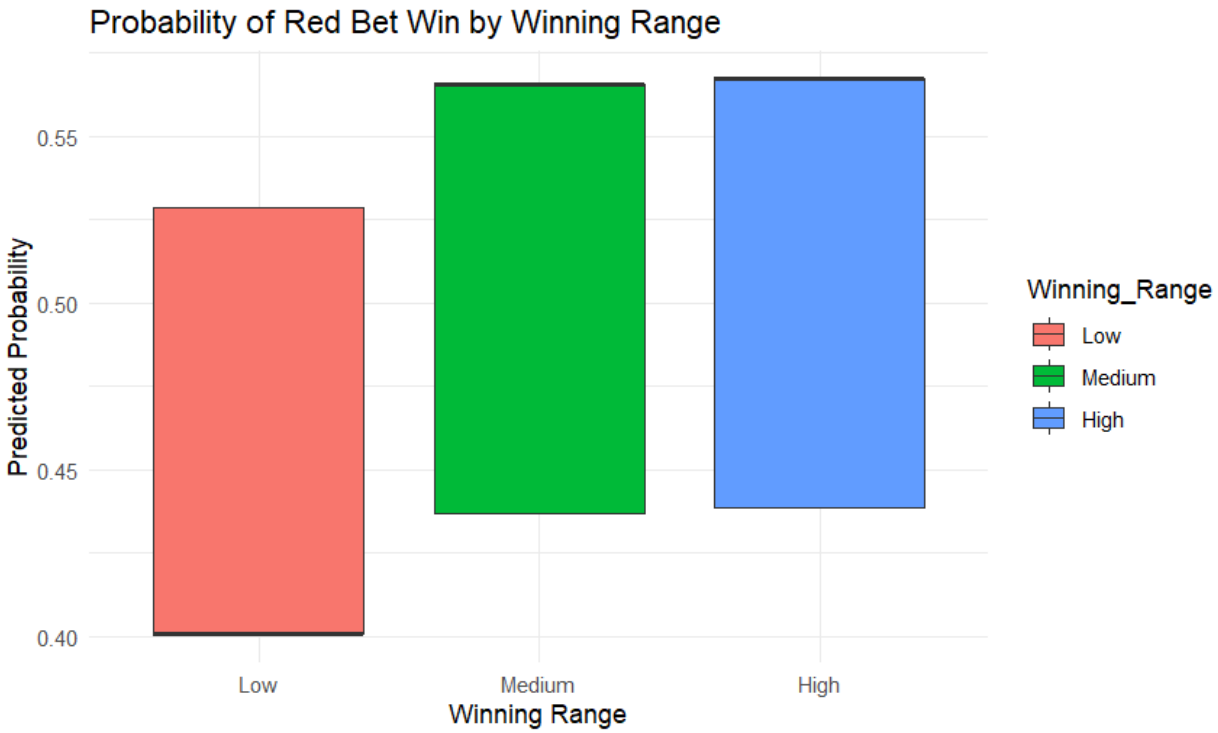
The logit model provided the following key results for predicting the likelihood of a red bet winning:

- The intercept value was -0.4008, with a p-value < 2e-16, confirming its statistical significance.
- Medium and high ranges for the winning number increased the likelihood of a red bet win by 0.1371 and 0.1447 units, respectively. These results were also highly significant (p-value < 2e-16).
- Odd-number outcomes had a strong positive association with red bet wins, with an estimate of 0.5221 and a z-value of 40.865.

The model achieved a residual deviance of 136,731 on 99,996 degrees of freedom, with an AIC of 136,739, indicating good fit.

Marginal effects analysis further clarified the impact of predictors:

- Odd numbers increased the probability of a red bet win by approximately 12.96%.
- Medium and high ranges contributed 3.36% and 3.55%, respectively, to the probability.



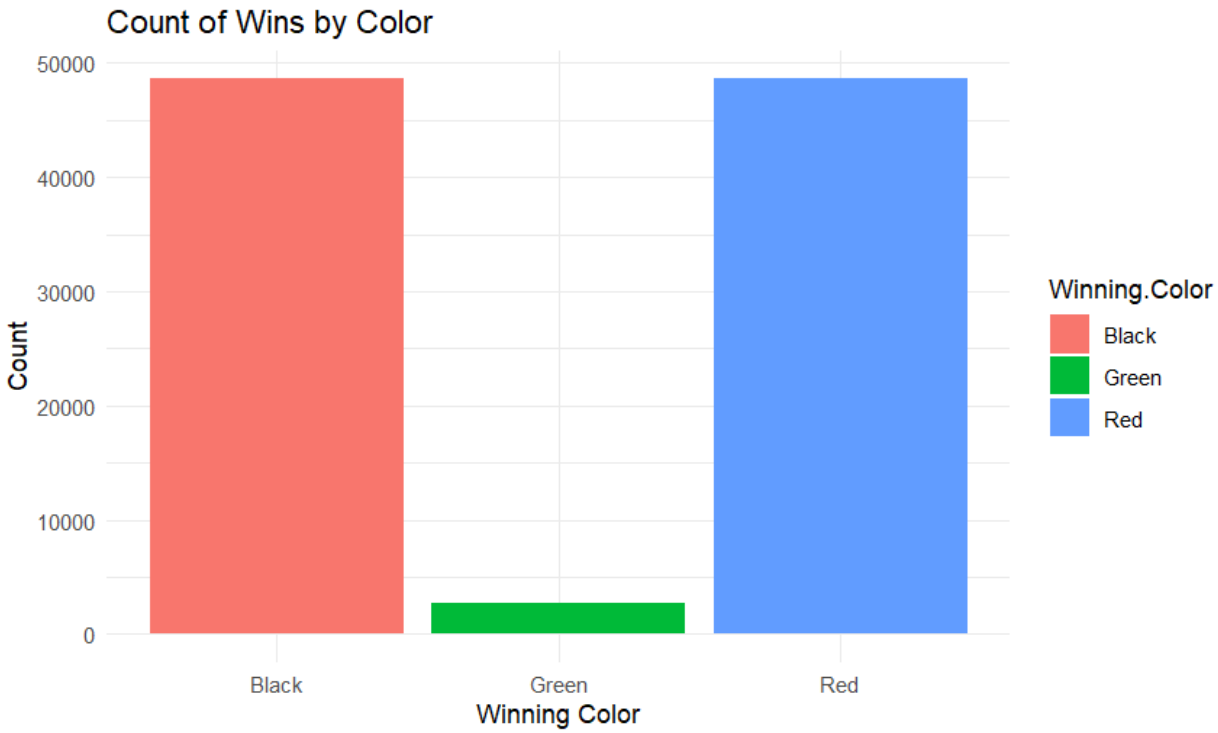
This plot illustrates the predicted probabilities of a red bet win across different winning ranges. Medium and high ranges show slightly higher probabilities (~55%) compared to low ranges (~50%), confirming the positive association observed in the logit model.

5.2 Count Data Models

The Poisson regression model examined win counts across colours:

- The intercept was 10.7920, with a z-value of 2,379.874 (p-value < 2e-16), indicating a strong baseline count for black outcomes.
- Green outcomes (representing zero) had a significantly lower count, with a coefficient of -2.8895, translating to a 94.7% reduction in win counts compared to black outcomes.
- Red outcomes did not differ significantly from black (p-value = 0.908).

These results demonstrated the predominance of black wins and the rarity of green outcomes in the dataset.



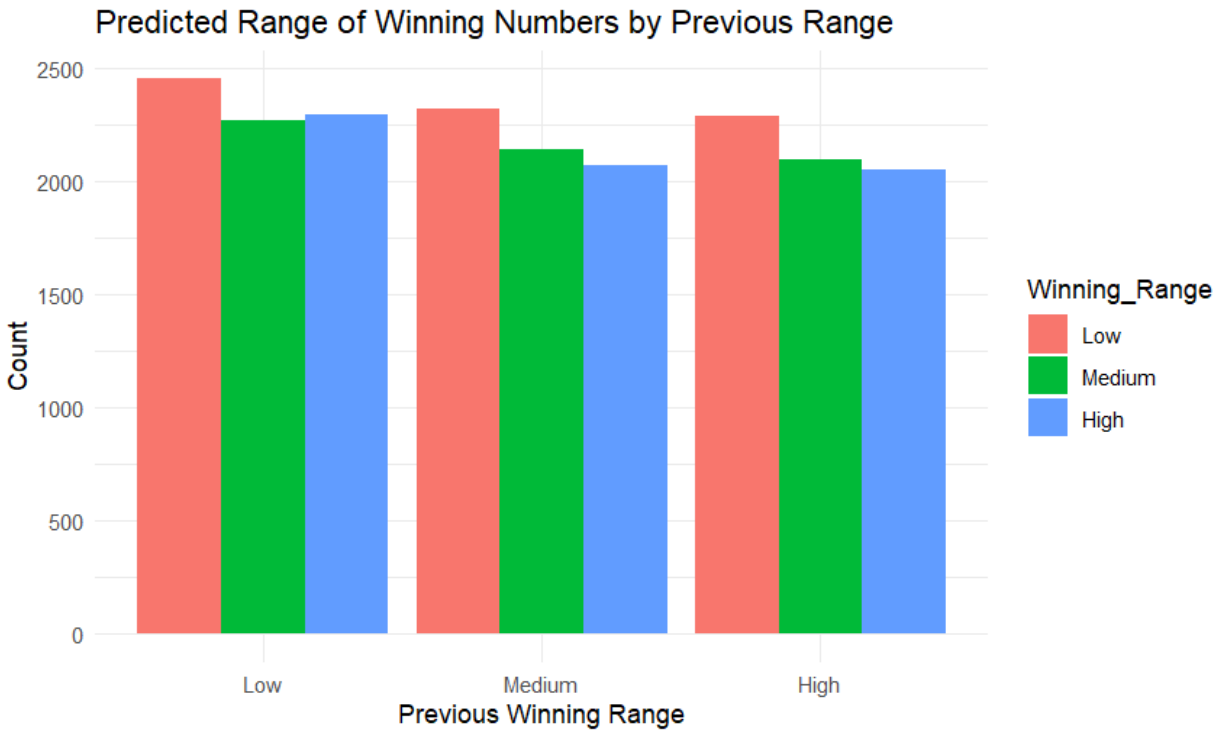
The bar plot shows that black and red outcomes dominate the win counts, with nearly equal distributions. Green outcomes, representing zero, are rare, aligning with the Poisson model's findings.

5.3 Ordered Regression Models

Ordered logit and probit models investigated dependencies between current and previous winning ranges:

- Both models indicated minimal influence of previous outcomes. For instance, the ordered probit model showed a coefficient of 0.0009 for medium ranges in the previous round, with a t-value of 0.1073.
- Intercept values for transitions between low, medium, and high ranges were significant, emphasizing the ordered nature of the dependent variable.

Residual deviance for the ordered probit model was 219,598.80, with an AIC of 219,606.80, suggesting modest predictive power.



The bar chart depicts the predicted distributions of winning ranges based on previous round results. All ranges are evenly distributed, supporting the models' conclusion of minimal dependence.

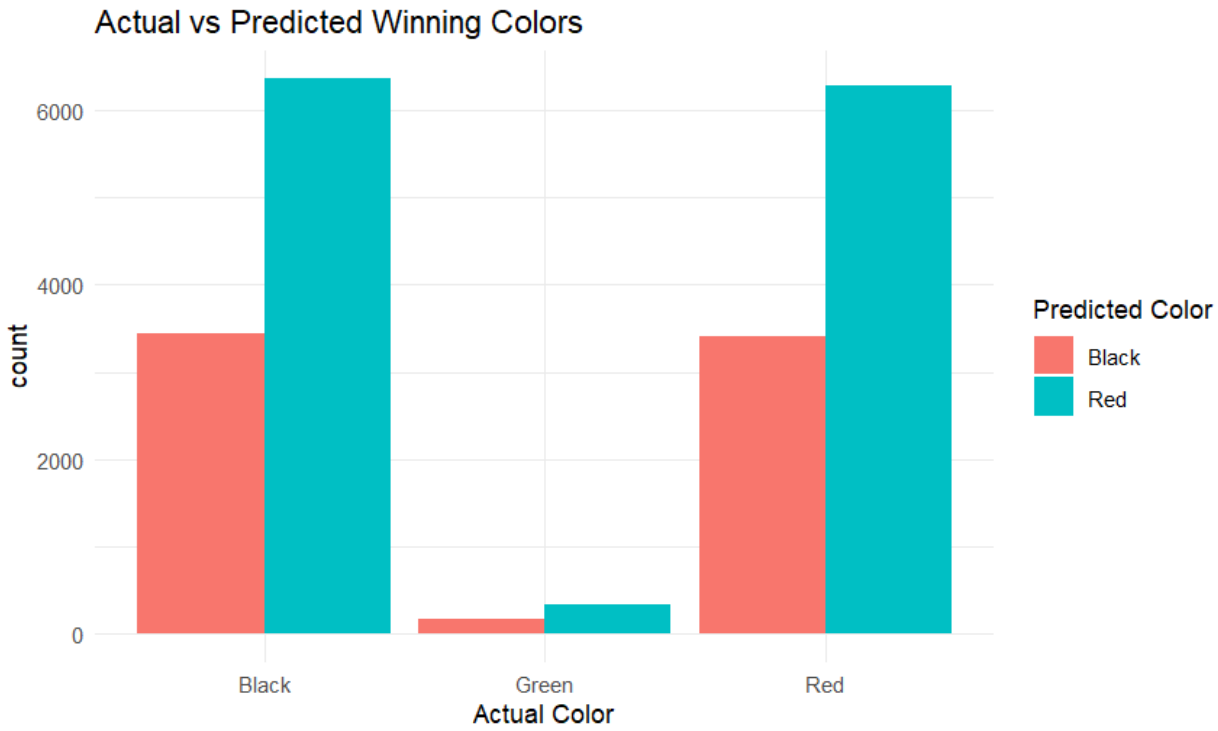
5.4 Multinomial Regression Models

The multinomial logit model explored the probabilities of specific colours winning:

- Black outcomes served as the baseline category.
- Green outcomes had an intercept of -2.9343, with medium and high ranges in the previous round contributing 0.0256 and 0.1093, respectively.
- Red outcomes exhibited negligible influence from previous ranges, with coefficients near zero.

Residual deviance for the model was 159,732.9, with an AIC of 159,744.9. Predictive performance was moderate, as indicated by confusion matrix results:

- Predicted black outcomes matched 3,441 actual black outcomes, but 6,364 black outcomes were misclassified as red.
- Green outcomes were consistently misclassified, reflecting their rarity.



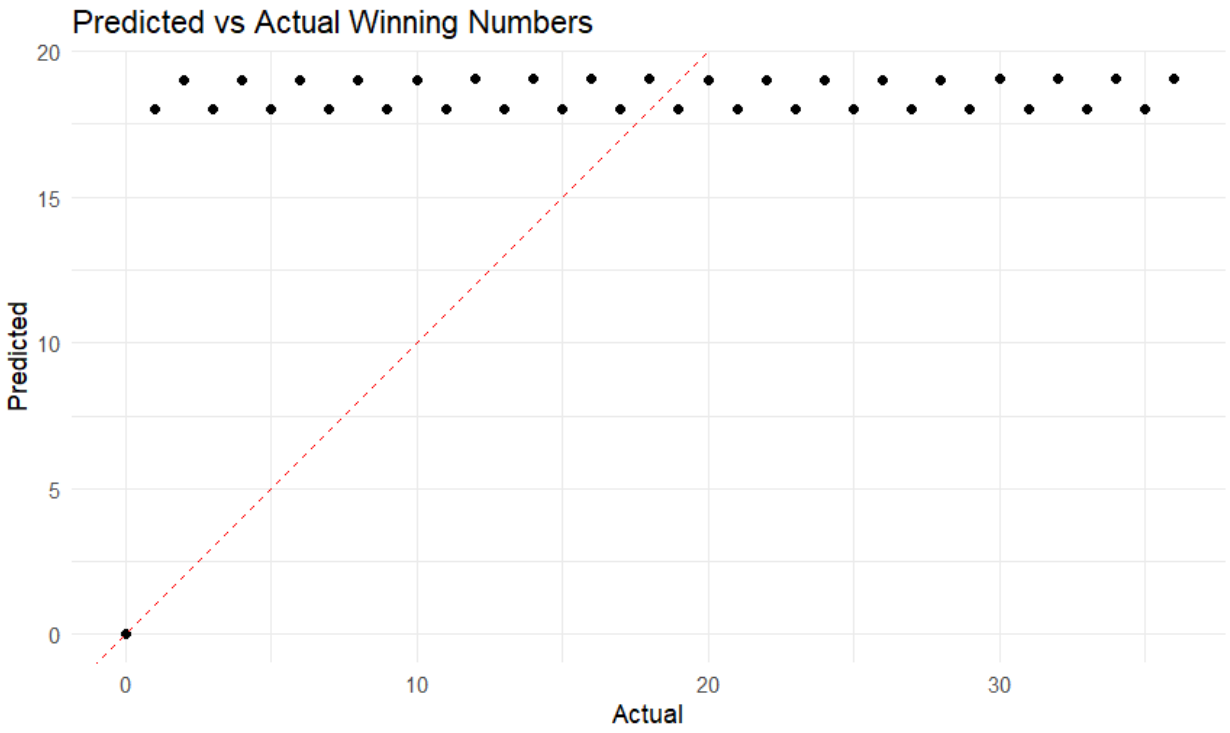
This plot highlights discrepancies in predicted outcomes, particularly the misclassification of black as red and the rarity of green outcomes.

5.5 Truncated and Censored Regression Models

Truncated regression analysis of outcomes greater than 18 showed:

- An intercept of 26.9627, highlighting the concentration of high numbers in the dataset.
- Even numbers had a stronger association with high outcomes (coefficient = 1.0772, p-value < 2e-16).

Tobit regression results were less conclusive due to large standard errors for predictors. While the intercept was significant, coefficients for predictors like even and odd outcomes were not statistically robust.



This scatter plot compares predicted and actual winning numbers. The clustering around higher values reflects the model's bias towards predicting high outcomes, consistent with the truncated regression results.

6. CONCLUSION

This study demonstrates the effectiveness of limited dependent variable models in analyzing structured datasets, using roulette as a practical example. The analysis highlights significant predictors such as the range of winning numbers and the parity of outcomes, with the logit model effectively showcasing how medium and high winning ranges positively influence the likelihood of red bet wins. Additionally, Poisson regression results confirm the predominance of black and red outcomes, while green (zero) outcomes remain rare. These findings align with theoretical expectations and validate the dataset's structure and reliability.

Based on the results, the following recommendations are proposed:

1. **Strategic Betting on Red**: Bettors should consider medium and high winning ranges as favorable conditions for red bets, as these significantly increase the probability of success, as shown by the logit model.
2. **Focus on Dominant Colours**: Given the high counts of black and red outcomes observed in the Poisson regression, betting strategies could focus on these colours while minimizing stakes on green (zero) outcomes.
3. **Incorporate Range-Specific Predictions**: Future betting strategies can leverage truncated regression insights, which highlight the importance of focusing on outcomes within specific ranges (e.g., numbers greater than 18).

Beyond identifying predictors, this project bridges theoretical econometric modelling with practical visualization, offering accessible interpretations of complex statistical outputs. The methodologies and insights derived from this analysis extend beyond roulette, providing a robust framework for examining restricted outcomes in various domains such as behavioural science and financial decision-making. Future research can build on these findings by exploring additional predictors or testing these models on more complex datasets, enhancing their applicability and impact.

7. REFERENCES

- Data source is Kaggle (https://www.kaggle.com/datasets/flynn28/simulated-roulette-data)
- Greene, W. H. (2018). *Econometric Analysis*. Pearson Education.
- Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Cambridge University Press.
- Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*. SAGE Publications.
- R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Effective use of GAI for rephrasing the report contents

8. APPENDIX

R Code Used for the Project:

```
# Loading Required Libraries
# install.packages(c("dplyr", "MASS", "nnet", "car", "AER", "margins", "ggplot2"))
library(dplyr)
library(MASS)
library(nnet)
library(AER)
library(margins)
library(ggplot2)

# Loading the Dataset
data <- read.csv("../../Project/Individual/roulette_100000_rounds.csv")
str(data)
head(data)

# ------------------------------------------------------------------------------
# Binary Choice Models (Logit and Probit)
# ------------------------------------------------------------------------------
# Question: What factors influence the probability of a bet on red winning?

# Preparing data to create independent variables
data$Winning_Range <- cut(data$Winning.Number, breaks = c(-1, 12, 24, 36), labels = c("Low",
"Medium", "High"))
data$Even_Odd <- ifelse(data$Winning.Number %% 2 == 0, "Even", "Odd")

# Logit Model
logit_model <- glm(Red.Bet.Win ~ Winning_Range + Even_Odd, family = binomial(link =
"logit"), data = data)
summary(logit_model)

# Probit Model
probit_model <- glm(Red.Bet.Win ~ Winning_Range + Even_Odd, family = binomial(link =
"probit"), data = data)
summary(probit_model)

# Marginal Effects for Logit
logit_marginal <- margins(logit_model)
summary(logit_marginal)

# Marginal Effects for Probit
probit_marginal <- margins(probit_model)
summary(probit_marginal)

# ------------------------------------------------------------------------------
```

```
# Count Data Models
# ------------------------------------------------------------------------------
# Question: How does the number of wins vary across colours in a given period?

# Aggregating the data for count analysis
data_summary <- data %>%
  group_by(Winning.Colour) %>%
  summarise(Count = n())

# Poisson Model
poisson_model <- glm(Count ~ Winning.Colour, family = poisson, data = data_summary)
summary(poisson_model)

# ------------------------------------------------------------------------------
# Ordered Regression Models (Ordered Probit and Ordered Logit)
# ------------------------------------------------------------------------------
# Question: Does the range of winning numbers (low, medium, high) depend on previous round
results?

# Preparing the data for Lag in winning range as independent variable
data <- data %>%
  mutate(Prev_Winning_Range = lag(Winning_Range))

# Ordered Probit Model
ordered_probit <- polr(Winning_Range ~ Prev_Winning_Range, data = data, method = "probit")
summary(ordered_probit)

# Ordered Logit Model
ordered_logit <- polr(Winning_Range ~ Prev_Winning_Range, data = data, method = "logistic")
summary(ordered_logit)

# Brant Test for Ordered Logit
# install.packages("brant")
library(brant)
brant(ordered_logit)

# ------------------------------------------------------------------------------
# Multinomial Regression Models (Logit and Probit)
# ------------------------------------------------------------------------------
# Question: What factors influence the probability of a specific colour (red, black, or zero) being
the winner?

# Multinomial Logit Model
multinomial_logit <- multinom(Winning.Colour ~ Prev_Winning_Range, data = data)
summary(multinomial_logit)
```

```
# Multinomial Probit Model
multinomial_probit <- multinom(Winning.Colour ~ Prev_Winning_Range, data = data, method =
"probit")
summary(multinomial_probit)

# ------------------------------------------------------------------------------
# Truncated/Censored Regression Models
# ------------------------------------------------------------------------------
# Question: Can we predict the winning number given that it falls within a specific range (e.g., >
18)?

# Using Truncated Data by filering rows with Winning Number > 18
truncated_data <- subset(data, Winning.Number > 18)

# Least Squares on Full Sample
lm_full <- lm(Winning.Number ~ Red.Bet.Win + Black.Bet.Win + Even.Bet.Win + Odd.Bet.Win,
data = data)
summary(lm_full)

# Least Squares on Truncated Sample
lm_truncated <- lm(Winning.Number ~ Red.Bet.Win + Black.Bet.Win + Even.Bet.Win +
Odd.Bet.Win, data = truncated_data)
summary(lm_truncated)

# Tobit Model
tobit_model <- tobit(Winning.Number ~ Red.Bet.Win + Black.Bet.Win + Even.Bet.Win +
Odd.Bet.Win, left = 18, data = data)
summary(tobit_model)

# ------------------------------------------------------------------------------
# Diagnostics for Checking the Assumptions
# ------------------------------------------------------------------------------
# Heteroskedasticity: Breusch-Pagan Test
# install.packages("lmtest")
library(lmtest)
bptest(logit_model)

# Multicollinearity: Variance Inflation Factor (VIF)
vif(logit_model)

# ------------------------------------------------------------------------------
# Model Performance Evaluation
# ------------------------------------------------------------------------------
# Splitting the data into training (80%) and testing (20%)
set.seed(123)  # For reproducibility
train_indices <- sample(1:nrow(data), size = 0.8 * nrow(data))
```

```
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

# Predictive performance for Logit Model
logit_model_train <- glm(Red.Bet.Win ~ Winning_Range + Even_Odd, family = binomial(link =
"logit"), data = train_data)
logit_predictions <- predict(logit_model_train, newdata = test_data, type = "response")
logit_class <- ifelse(logit_predictions > 0.5, 1, 0)
confusion_matrix_logit <- table(Predicted = logit_class, Actual = test_data$Red.Bet.Win)
print("Confusion Matrix for Logit Model:")
print(confusion_matrix_logit)

# Predictive performance for Probit Model
probit_model_train <- glm(Red.Bet.Win ~ Winning_Range + Even_Odd, family = binomial(link
= "probit"), data = train_data)
probit_predictions <- predict(probit_model_train, newdata = test_data, type = "response")
probit_class <- ifelse(probit_predictions > 0.5, 1, 0)
confusion_matrix_probit <- table(Predicted = probit_class, Actual = test_data$Red.Bet.Win)
print("Confusion Matrix for Probit Model:")
print(confusion_matrix_probit)

# Predictive performance for Multinomial Logit Model
multinomial_logit_train <- multinom(Winning.Colour ~ Prev_Winning_Range, data = train_data)
multinomial_predictions <- predict(multinomial_logit_train, newdata = test_data)
confusion_matrix_multinomial <- table(Predicted = multinomial_predictions, Actual =
test_data$Winning.Colour)
print("Confusion Matrix for Multinomial Logit Model:")
print(confusion_matrix_multinomial)

# Predictive performance for Truncated Model (using RMSE)
predicted_lm <- predict(lm_full, newdata = test_data)
rmse_lm <- sqrt(mean((test_data$Winning.Number - predicted_lm)^2))
print("RMSE for Truncated Model:")
print(rmse_lm)

# ------------------------------------------------------------------------------
# Visualizations for Business Questions
# ------------------------------------------------------------------------------

# Binary Choice Models: Logit predictions for "Red Bet Win"
ggplot(data = test_data, aes(x = Winning_Range, y = logit_predictions)) +
  geom_boxplot(aes(fill = Winning_Range)) +
  labs(title = "Probability of Red Bet Win by Winning Range", y = "Predicted Probability", x =
"Winning Range") +
  theme_minimal()
```

```
# Count Data Models: Wins by Colour
colour_counts <- data %>% group_by(Winning.Colour) %>% summarise(Count = n())
ggplot(colour_counts, aes(x = Winning.Colour, y = Count, fill = Winning.Colour)) +
  geom_bar(stat = "identity") +
  labs(title = "Count of Wins by Colour", y = "Count", x = "Winning Colour") +
  theme_minimal()

# Ordered Regression Models: Predictions by Previous Winning Range
ggplot(data = test_data, aes(x = Prev_Winning_Range, fill = Winning_Range)) +
  geom_bar(position = "dodge") +
  labs(title = "Predicted Range of Winning Numbers by Previous Range", x = "Previous Winning
Range", y = "Count") +
  theme_minimal()

# Multinomial Regression Models: Actual vs Predicted Winning Colours
ggplot(test_data, aes(x = Winning.Colour, fill = factor(multinomial_predictions))) +
  geom_bar(position = "dodge") +
  labs(title = "Actual vs Predicted Winning Colours", x = "Actual Colour", fill = "Predicted
Colour") +
  theme_minimal()

# Truncated Regression Models: Predicted vs Actual Winning Numbers
rmse_plot <- ggplot(test_data, aes(x = Winning.Number, y = predicted_lm)) +
  geom_point(alpha = 0.5) +
  geom_abline(slope = 1, intercept = 0, colour = "red", linetype = "dashed") +
  labs(title = "Predicted vs Actual Winning Numbers", x = "Actual", y = "Predicted") +
  theme_minimal()
print(rmse_plot)
```