Lab 05: Logistic Regression

**Due date:** Thursday, March 21, 2025 submitted as Word document to Canvas ***Lab05*** link

This lab counts 9 % toward your total grade.

**Objectives:** In this lab, you will practice your skills in

1. Explore logistic regression
2. All effects plot
3. Conditional effects plot

**Format of answer:** Submit your answers as a **Word document** with graphs and verbal descriptions, properly labeled in the task sequence, with answers in red text and only relevant content included

**Accident\_data.shp** show the spatial distribution of traffic accidents in the city of Dallas. The accident is represented as 1 (Yes, a traffic accident occurred at the location) and 0 (No, no traffic accident occurred at location).

A map with red and blue dots

AI-generated content may be incorrect.

# Task 1: Load data (2 pts)

We will use the **Accident\_data.shp** to practice the skills in logistic regression

1. Load the **Accident\_data.shp** data using **sf::st\_read()** (1 pt)

A computer code with text

AI-generated content may be incorrect.

1. Check the geometry column and explain what information are stored in geometry column using **data$geometry**. (0.5 pt)
2. Each feature is a point with location coordinates
3. Coordinate reference system (CRS) is Lambert Conformal Conic. It’s a metric coordinate system.
4. Dimension is xy. Which means no Z (elevation) or M(measure) value.
5. Bounding box provides spatial extent of shapefile.

A screenshot of a computer code

AI-generated content may be incorrect.

1. Convert column ‘**accident**’ to factor. (0.5 pt)

df$accident = as.factor(df$accident)

Why as.factor is used?

 **Converts** a numeric or character column into a factor (categorical variable).

 **Enables grouping in visualizations**, e.g., different colors/shapes for each accident type in ggplot2.

 **Tells models** (like glm, lm, etc.) to treat accident as a **categorical predictor**, not a continuous one.

The dependent variable in the data set is whether there is a traffic accident at the location.

|  |  |
| --- | --- |
| **Dependent variable** | **Description** |
| **accident** | a factor with levels: ‘0’; ‘1’ |

The independent variable describes the geographical environments around the location.

|  |  |
| --- | --- |
| **Independent variable** | **Description** |
| **Hour** | Hour of the day, from 0 - 23 |
| **F\_SYSTEM** | Road type at different levels: 0 -7 |
| **intersect** | Whether the accident occurred at a major intersection |
| **NUM\_LANES** | The number of lanes |
| **lane\_width** | Lane width |
| **BelowFreez** | Whether the temperature is below freezing, a factor with levels: ‘FALSE’; ‘TRUE’ |
| **Fog** | Whether foggy conditions are present: ‘FALSE’; ‘TRUE’ |
| **Thunder** | Whether thunderstorms are present: ‘FALSE’; ‘TRUE’ |
| **FrozenPrec** | Whether there is frozen precipitation: ‘FALSE’; ‘TRUE’ |

# Task 2: Build logistic regression model (3 pts)

1. Build a logistic regression for the probability of accident with independent variable list above and provide the 95% confidence intervals around the logistic regression parameters.

logistic\_acc =glm(accident~hour+F\_SYSTEM+intersect+NUM\_LANES+lane\_width+BelowFreez+Fog+Thunder+FrozenPrec,

family = binomial(logit),data = df)

summary(logistic\_acc)

confint(logistic\_acc, level = 0.95)

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. Discuss the model output from logistic regression (significant of the variable) and analyze whether the results from confidence intervals align with the model results.

Yes, the results of confidence interval are aligned with the output of the logistic regression. For all variables where the p-value < 0.05, the 95% CI does not contain 0 (e.g., hour, F\_SYSTEM, intersect, NUM\_LANES). This suggested that hour of day, road system type, intersection presence, and number of lanes are significant.

1. Using all effects plot to interpret the calibrated logistic regression model in terms of probabilities.

library(effects)

plot(allEffects(logistic\_acc), type="response", ylim=c(0,1), ask=FALSE)

A group of graphs with text

AI-generated content may be incorrect.

# Each panel shows the marginal effect of a single predictor on the predicted probability of an accident keeping rest of the predictors constant. We can see that how the predicted probability of an accident changes when just one variable changes, keeping all others constant.

# Task 3: Perform one likelihood ratio test (2 pts)

1. Refine the model from Task 2 by dropping all variables that are not significant important for dependent variable. Test whether the removed variables had a significant influence on the dependent variable using ANOVA-based LRT or Manual LRT (logLik()). (2 pts)

logistic\_acc02 = glm(accident~hour+F\_SYSTEM+intersect+NUM\_LANES+Fog,

family = binomial(logit),data = df)

summary(logistic\_acc02)

anova(logistic\_acc02,logistic\_acc)

## Likelihood ratio test

lkh = logLik(logistic\_acc02)

lh = logLik(logistic\_acc)

LR = -2 \* (lkh - lh)

pchisq(LR[1], df = 4, lower.tail = F)

A screen shot of a computer code

AI-generated content may be incorrect.

The 4 added variables do not significantly improve the model.

Bonus (+1 point): Try both approaches and compare results. Explain if they have the same result or not.

The p-value of the Chi-squared test from ANOVA-based LRT and Manual LRT (logLik()) are both equal to 0.9309.

# Task 4: Conditional effects plots (2 pts)

1. Generate conditional effects for the NUM\_LANES variable based on the refined model in Task 2.a to estimate the probability of traffic accident. (1.5 pts)

Assume two scenarios with the following values for the additional independent variable in the logistic regression model:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Low probability** | **High probability** |
| **hour** | 3 | 19 |
| **F\_SYSTEM** | 6 | 2 |
| **Intersect** | ‘no’ | ‘yes’ |

par(mfrow = c(1, 1))

library(effects)

# Compute effect for low probability scenario

low\_prb <- effect("NUM\_LANES", logistic\_acc02, given.values = c(hour = 3, F\_SYSTEM = 6, intersectyes = 0))

# Compute effect for high probability scenario

high\_prb <- effect("NUM\_LANES", logistic\_acc02, given.values = c(hour = 19, F\_SYSTEM = 2, intersectyes = 1))

1. Discuss the conditional effects plots for the two scenarios. (0.5 pt)

# Extract data

low\_data <- as.data.frame(low\_prb)

high\_data <- as.data.frame(high\_prb)

# Plot the low probability case first

plot(low\_data$NUM\_LANES, low\_data$fit, type = "l", col = "blue", lwd = 2, ylim = c(0, 1),

ylab = expression(Pr(Y[i] == "Accident probability")), xlab = "Number of Lanes",

main = "Accident Probability vs. Number of Lanes")

# Add shaded confidence intervals for low probability

polygon(c(low\_data$NUM\_LANES, rev(low\_data$NUM\_LANES)),

c(low\_data$lower, rev(low\_data$upper)), col = rgb(0, 0, 1, 0.2), border = NA)

# Add high probability scenario

lines(high\_data$NUM\_LANES, high\_data$fit, col = "red", lwd = 2)

# Add shaded confidence intervals for high probability

polygon(c(high\_data$NUM\_LANES, rev(high\_data$NUM\_LANES)),

c(high\_data$lower, rev(high\_data$upper)), col = rgb(1, 0, 0, 0.2), border = NA)

# Add legend

legend("bottomright", legend = c("Low Probability", "High Probability"), col = c("blue", "red"), lwd = 2)

A graph of accident probability

AI-generated content may be incorrect.

|  |  |  |
| --- | --- | --- |
| Aspect | Description | Interpretation |
| Both curves rise | Accident probability increases with number of lanes | Suggests more lanes - higher complexity, traffic volume, decision-making demands |
| Blue vs. Red curve | Red curve is always above blue | Accident probability depends on context - number of lanes is important, but its impact is conditioned by other variables |
| Shaded area | Confidence intervals | Wider at extremes (e.g., 10 lanes) in low probability scenario but narrow in high probability scenario- more uncertainty in low probability scenario due to more variability |