Review of Logistic Regression

CRP 245 Tutorial

Duke University Clinical Research Training Program

2026-01-11

Table of contents

Introduction

Learning Objectives

After completing this tutorial, you will be able to:

- **Understand** the application of logistic regression for binary clinical outcomes.
- **Visualize** the relationship between a continuous predictor and a binary outcome.
- **Interpret** logistic regression coefficients on the log-odds scale.
- **Calculate** and **Interpret** Odds Ratios (OR) and their 95% Confidence Intervals.
- **Predict** the probability of a clinical outcome for an individual patient.

0.1 Clinical Context

When a patient is diagnosed with prostate cancer, an important question in deciding on a treatment strategy is whether or not the cancer has spread to neighboring lymph nodes. This relationship is critical because lymph node status significantly influences both treatment decisions and long-term prognosis.

Why Study This?

As physician-scientists, we frequently encounter biomarkers (like serum acid phosphatase) that might signal disease severity. However, raw clinical data can be messy. Logistic regression provides a rigorous mathematical framework to translate these biomarkers into actionable probabilities of risk, helping us move from "intuition" to "evidence-based" clinical decision-making.

0.2 Study Overview

In this analysis, we examine data from 53 prostate cancer patients. Several possible predictor variables were measured before surgery, and the patients then underwent surgery to confirm the presence or absence of nodal involvement.

0.3 Data Dictionary

Table 1: Prostate Cancer Dataset Variables

Variable	Description
X	Patient Identification Number
Xray	Measure of cancer seriousness from X-ray (1 = more severe; 0 = less severe)
Grade	Dichotomized tumor measure (1 = more serious; 0 = less serious)
Stage	Dichotomized tumor measure (1 = more serious; 0 = less serious)
Age	Age at diagnosis (years)
Acid	Serum acid phosphatase level (multiplied by 100)
Nodes	Lymph node involvement found at surgery (1 = present; 0 = absent)

---

1 Setup and Data Loading

1.1 Setup: Environment Configuration

Following our standard workflow, we first ensure the necessary statistical packages are installed and loaded.

```
# Check for and install 'stats' (built-in)
if (!requireNamespace("stats", quietly = TRUE)) {
  install.packages("stats")
}
library(stats)

# Check for and install 'graphics' for plotting
if (!requireNamespace("graphics", quietly = TRUE)) {
  install.packages("graphics")
}
library(graphics)
```

1.2 Loading the Dataset

We load the `prostate_node` dataset directly from the course website and examine our study population.

```r
# Load data
load(url("https://www.duke.edu/~sgrambow/crp241data/prostate_node.RData"))

# Examine basic characteristics
summary(prostate_node)
```

```
       X              Xray            Grade            Stage             Age       
 Min.   : 1.00   Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :45.00  
 1st Qu.:14.00   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:56.00  
 Median :27.00   Median :0.000   Median :0.0000   Median :1.0000   Median :60.00  
 Mean   :27.00   Mean   :0.283   Mean   :0.3774   Mean   :0.5094   Mean   :59.38  
 3rd Qu.:40.00   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:65.00  
 Max.   :53.00   Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :68.00  
      Acid            Nodes       
 Min.   : 40.00   Min.   :0.0000  
 1st Qu.: 50.00   1st Qu.:0.0000  
 Median : 65.00   Median :0.0000  
 Mean   : 69.42   Mean   :0.3774  
 3rd Qu.: 78.00   3rd Qu.:1.0000  
 Max.   :187.00   Max.   :1.0000  
```

Missing Values Note

Always check for missing values (NAs) during data loading. In this specific trial dataset, there are **0 missing values**. However, consistently checking for NAs is critical in clinical research, as missing data can bias your model results if not handled correctly.

Statistical Interpretation

- **Sample Size:** $n = 53$ patients.
- **Biomarker:** Acid phosphatase (`Acid`) ranges from 40 to 187 with a median of 65.
- **Outcome:** The mean of `Nodes` (0.3774) indicates that approximately 38% (20 out of 53) of patients in this sample have nodal involvement.

Clinical Interpretation

The dataset represents a group of prostate cancer patients with a moderate prevalence of lymph node spread. The biomarker of interest, acid phosphatase, shows a right-skewed distribution, meaning while most patients have lower levels, a few have very high levels.

---

2 Exploratory Visualization

2.1 Visualization of Nodal Involvement

Question: Is there visual evidence of an association between nodal involvement and serum acid phosphatase level?

```
# We plot Acid (predictor) on x-axis and Nodes (outcome) on y-axis
plot(prostate_node$Acid, prostate_node$Nodes,
     main = "Nodal Involvement vs. Acid Phosphatase",
     xlab = "Serum Acid Phosphatase (x 100)",
     ylab = "Nodal Involvement (1=Present, 0=Absent)",
     cex = 1.5, pch = 1, col = "blue")
```

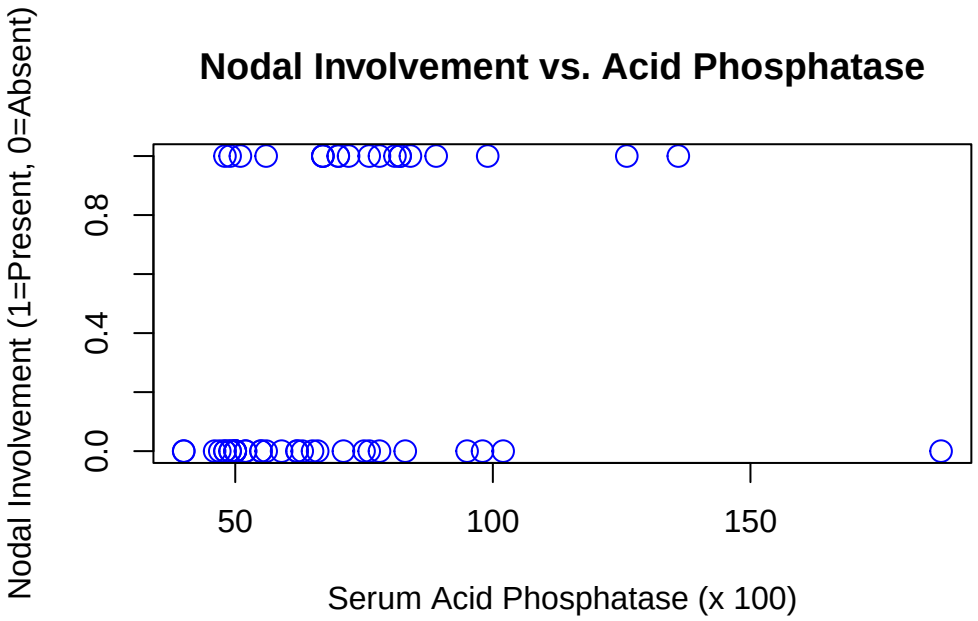

Figure 1: Scatterplot of Nodal Involvement by Acid Phosphatase Level

Statistical Interpretation

The scatterplot displays binary outcomes (0 and 1) on the y-axis. While it is difficult to see a clear linear trend with discrete data, patients with nodal involvement (1) appear somewhat more concentrated at higher acid phosphatase levels than those without (0).

Clinical Interpretation

Visual inspection suggests a possible trend, but the significant overlap between groups makes it impossible to draw a definitive conclusion without formal modeling. Logistic regression is needed to quantify the strength of this relationship.

---

3 Logistic Regression Modeling

3.1 Testing the Association

Question: Is there statistically significant evidence of an association between serum acid phosphatase and nodal involvement at the $\alpha = 0.05$ level?

```
# Fit logistic regression model
# NOTE: family='binomial' is essential for binary outcomes
acid.fit <- glm(Nodes ~ Acid, data=prostate_node, family='binomial')

# Display model results
summary(acid.fit)
```

```
Call:
glm(formula = Nodes ~ Acid, family = "binomial", data = prostate_node)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.92703    0.92104  -2.092   0.0364 *
Acid         0.02040    0.01257   1.624   0.1045
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 67.116  on 51  degrees of freedom
AIC: 71.116

Number of Fisher Scoring iterations: 4
```

! Programming Note: The `family` Argument

In R, the `glm()` function defaults to linear regression (Gaussian) if the `family` argument is omitted. For binary clinical outcomes (Yes/No), you **must** specify `family='binomial'` to perform logistic regression.

i Statistical Interpretation

- **Coefficient (Slope) for Acid:** 0.02040 (representing the change in log-odds).
- **Standard Error:** 0.01257.
- **p-value:** 0.1045.
- **Conclusion:** At $\alpha = 0.05$, we fail to reject the null hypothesis because the p-value is greater than 0.05.

Clinical Interpretation

The p-value of 0.10 indicates that we do not have enough evidence in this specific sample of 53 patients to claim a statistically significant association. However, the positive coefficient (0.02) suggests a clinical trend where higher acid phosphatase levels are associated with a greater risk of nodal involvement.

---

4 Interpretation of Results

4.1 Odds Ratios and the Nature of Association

Question: How do we interpret the risk on a scale that is clinically intuitive?

In logistic regression, we exponentiate the coefficients to convert them from the log-odds scale to **Odds Ratios (OR)**.

```
# Calculate odds ratios
exp(acid.fit$coefficients)
```

```
(Intercept)        Acid
  0.1455796   1.0206103
```

Statistical Interpretation

- **Intercept ($e^{-1.927} = 0.145$):** The odds of nodal involvement when `Acid` is zero (not clinically useful as the range starts at 40).
- **Slope ($e^{0.0204} = 1.0206$):** The multiplicative change in odds for every 1-unit increase in acid phosphatase.

Clinical Interpretation

An Odds Ratio (OR) of **1.021** means that for every **1-unit** increase in serum acid phosphatase, the odds of nodal involvement increase by approximately **2.1%**.

4.2 Clinically Meaningful Change

Question: What is the risk associated with a **10-unit** increase in acid phosphatase?

Individual unit changes are often too small to be clinically relevant. We can calculate the OR for a 10-unit increase and its 95% Confidence Interval.

```
# OR for 10-unit increase
exp(10 * acid.fit$coefficients)
```

```
 (Intercept)         Acid
4.275675e-09 1.226308e+00
```

```
# 95% Confidence Interval for 10-unit increase
exp(10 * confint(acid.fit))
```

```
                  2.5 %     97.5 %
(Intercept) 7.917783e-18 0.07548575
Acid        9.794927e-01 1.62198170
```

Statistical Interpretation

- **OR for 10-unit increase:** 1.226.
- **95% Confidence Interval:** [0.978, 1.616].

Clinical Interpretation

A 10-unit increase in acid phosphatase is associated with a **23% increase** in the odds of nodal involvement. However, because the 95% Confidence Interval includes **1.0** (the value of "no effect"), this increase is not statistically significant at the 5% level. This highlights how clinical importance (a 23% increase) and statistical significance can sometimes diverge in small studies.

---

5 Individual Prediction and Visualization

5.1 Predicted Probability for a Patient

Question: What is the predicted probability of nodal involvement for a patient with an acid phosphatase level of **78**?

```
# Calculation using the logistic equation
xb <- -1.92703 + (0.02040 * 78)
predicted_prob <- exp(xb) / (1 + exp(xb))
print(predicted_prob)
```

```
[1] 0.4168228
```

```
# Alternate calculation using predict function
predict(acid.fit, data.frame(Acid=78), type="response")
```

```
        1
0.4168367
```

Clinical Interpretation

For a patient with a serum acid phosphatase level of 78, the model predicts a **41.7% probability** of lymph node involvement. In a clinical context, such a patient might be considered at high risk (nearly a 1 in 2 chance), suggesting that further diagnostic investigation or a more aggressive treatment strategy might be warranted.

5.2 Visualizing the Logistic Curve

Finally, we visualize how the probability of the outcome changes across the entire range of the biomarker.

```
# Sequence of Acid values
xAcid <- seq(40, 187, length=100)
# Predicted probabilities
yxAcid <- predict(acid.fit, list(Acid=xAcid), type="response")

# Plot
plot(prostate_node$Acid, jitter(prostate_node$Nodes, amount=0.05),
     pch=1, col="blue",
     main = "Logistic Regression: Predicted Probability Curve",
     xlab = "Serum Acid Phosphatase",
     ylab = "Probability of Nodal Involvement")
lines(xAcid, yxAcid, col="darkred", lwd=2)
```

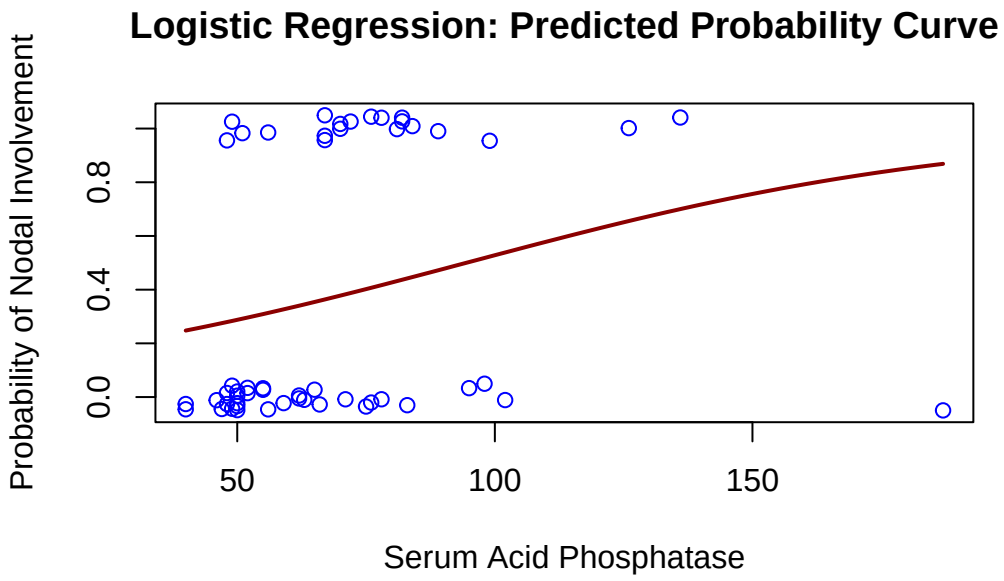

Figure 2: Logistic Regression Curve showing Probability of Nodal Involvement

Clinical Summary

The "S-shaped" logistic curve illustrates the continuous increase in risk as acid phosphatase levels rise. While the trend is clinically meaningful, physician-scientists should look for larger validation studies before relying solely on this single biomarker for surgical planning.

---

6 Generalizability and Limitations

6.1 External Validity

Clinical Question: Can we apply these findings to patients in a different clinical setting?

Generalizability Note

This model was developed using a sample of 53 patients from a specific cohort. External validity (how well it works in the "real world") should be confirmed with larger, multi-center trials before these exact thresholds are used for surgical decision-making. Factors such as race, comorbidities, and different laboratory standards for measuring acid phosphatase can all influence the results.

6.2 Avoiding Extrapolation

Clinical Question: Can we use this model for a patient with an acid phosphatase level of 250?

Do Not Extrapolate

Statistical models are valid only within the range of the observed data. In our study, acid phosphatase levels ranged from **40 to 187**. Applying the model to values outside this range (extrapolation) is statistically risky and could lead to highly inaccurate predictions of risk.

---

7 Summary and Conclusions

i Key Takeaways

1. **Binary Outcomes:** Logistic regression is the standard tool for modeling binary (Yes/No) clinical events.
2. **Clinical Interpretation:** Odds Ratios (OR) provide the most intuitive way for clinicians to understand how risk changes per unit of a biomarker.
3. **Group vs. Individual:** While a model may not reach statistical significance in a small group ($p > 0.05$), the predicted probabilities for an *individual* patient (like our 41.7% example) may still be clinically concerning.
4. **Significance vs. Effect Size:** A "non-significant" result doesn't always mean there is "no effect"—it sometimes means the study was too small to be certain of the effect we observed.

! Key Teaching Checklist

- ☐ **Logistic Regression** is required for modeling binary clinical outcomes.
- ☐ **Odds Ratios (OR)** are obtained by exponentiating (e^x) the model coefficients.
- ☐ **Confidence Intervals** that include 1.0 signify a lack of statistical significance at the chosen α level.
- ☐ **S-Curves** visualize how probability changes non-linearly across the biomarker range.

---

8 Practice Exercises

1. **Manual Calculation:** Use the model intercept (-1.927) and slope (0.0204) to manually calculate the predicted probability for a patient with an acid level of 100. (Hint: Calculate xb first, then use $e^{xb}/(1+e^{xb})$).
2. **Context Clues:** Why is the p-value for `Acid` (0.1045) considered non-significant at the 0.05 level? What might happen to this p-value if the sample size was 500 patients instead of 53?
3. **Visualization:** Generate the KM curve plot for this data if it were survival data (status = Nodes, time = futime). How does the visual representation differ from the logistic S-curve?

This tutorial was developed for CRP 245 at Duke University.