

HW 9

Logistic Regression Models

December 10, 2025

💡 Assignment Goals

1. Conduct inference for simple and multiple logistic regression models
2. Appropriately interpret the coefficients of a logistic regression model with interactions
3. Build and interpret confidence intervals for a linear combination of logistic regression coefficients
4. Perform nested likelihood ratio tests by hand

Q1 (Connecting Logistic Regression and Contingency Table Analyses)

In class, we analyzed survival data from 91 individuals on the [Donner Party](#) expedition to answer the question: “are there sex-specific differences in one’s ability to withstand and survive harsh conditions?” Our analysis was based on the 2×2 contingency table below:

sex	died	survived	Total
Female	10	25	35
Male	32	24	56
Total	42	49	91

In particular, we estimated that the probability of survival for men was $\hat{\pi}(1) = 0.43$, that the probability of survival for women was $\hat{\pi}(0) = 0.71$, and that the corresponding odds ratio comparing survival among men to survival among women was $\widehat{OR} = 0.30$ (90% CI: 0.14 to 0.64). What conclusions would we have reached if we analyzed the data using logistic regression instead?

The data for this analysis is available in the `Donner` object from the `alr4` package. To read this data into R, run the following line of code:

```
donner <- alr4::Donner
```

(a) Fit the logistic regression model of Donner Party survival as a function of sex (with female sex taken to be the reference level) and print out the corresponding model summary table in R.

(b) Use the fitted logistic regression model from (a) to predict the probability of survival among male members of the expedition and among female members of the expedition. How do your results compare to those from the contingency table?

(c) Use the fitted logistic regression model from (a) to calculate the odds ratio for the association between survival and sex as well as the corresponding 90% Wald confidence interval. How do your results compare to those from the contingency table?

(d) In a first course in statistics, we might have tested the hypothesis that “there are sex-specific differences in one’s ability to withstand and survive harsh conditions” by looking at the difference in survival proportions:

$$H_0 : \pi(1) - \pi(0) = 0.$$

In a logistic regression analysis, we conduct hypothesis testing for this same question using the odds ratio instead:

$$H_0 : OR = \theta(1)/\theta(0) = 1.$$

Show that these two null hypotheses are, in fact, mathematically equivalent, i.e., that if the difference in proportions is zero, the odds ratio must be 1.

(e) Use the fitted logistic regression model from (a) to test the hypotheses in (d) in two ways: first using a Wald test, and then using a likelihood ratio test. (Take the significance level to be $\alpha = 0.05$.) Interpret the conclusions of your hypothesis tests in context.

(f) The correctness of both the contingency table analysis and the logistic regression analysis requires that the observations in the data are independent of one another. Do you think this condition is plausible for the Donner Party data? Why or why not?

Q2 (Intensive Care Unit)

The ICU data set in the `Stat2Data` package contains information on a random sample of 200 patients who were part of a larger study conducted in a hospital’s Intensive Care Unit (ICU). Since an ICU often deals with serious, life-threatening cases, a key variable to study is patient survival (`Survive`), which is coded as a 1 if a patient lived to be discharged and a 0 if the patient died.

```
library(Stat2Data)
data(ICU)
head(ICU)
```

	ID	Survive	Age	AgeGroup	Sex	Infection	SysBP	Pulse	Emergency
1	4	0	87	3	1	1	80	96	1
2	8	1	27	1	1	1	142	88	1
3	12	1	59	2	0	0	112	80	1
4	14	1	77	3	0	0	100	70	0
5	27	0	76	3	1	1	128	90	1
6	28	1	54	2	0	1	142	103	1

(a) Fit a logistic regression model for the probability of surviving until discharge from the ICU, modeled as a function of a patient's systolic blood pressure (SysBP, measured in mmHg); sex, dichotomized as male and female with male sex as the reference level (**Sex**); and their interaction. Write down the logit (log-odds) form of the fitted model, defining any notation that you use, and then visualize it by adding it to a scatter plot of the ICU data.

(b) Based on this fitted regression model, what is the estimated change in the odds of survival for a one mmHg increase in systolic blood pressure among male patients? Compute and interpret a 95% confidence interval for this association. Given this interval, is there a statistically significant association between systolic blood pressure and survival for men?

(c) What is the estimated change in the odds of survival for a one mmHg increase in systolic blood pressure among female patients? Compute and interpret a 95% confidence interval for this association. Given this interval, is there a statistically significant association between systolic blood pressure and survival for women? *Hint: the estimated association involves a linear combination of coefficients, which means that you will need to use the `vcov()` function to help you calculate your standard error.*

(d) Compute and interpret the odds ratio comparing survival among female patients to survival among male patients who are all admitted to the ICU with a systolic blood pressure of 90 mmHg. Repeat this calculation and interpretation for female patients and male patients who are admitted with a systolic blood pressure of 150 mmHg.

Q3 (Selection, Naturally)

The year: 1898. The place: a biology lecture in Woods Hole, Massachusetts. The scientific theory under discussion: natural selection.¹

Professor Herman Bumpus, a biologist at Brown University, is giving a lecture on data he collected in order to test this new theory. He collected data on 87 house sparrows that were found on the ground following a particularly severe winter storm in 1898. Some of these birds had died in the storm; others had survived. Professor Bumpus theorized that the birds that survived the storm did so because they had superior physical traits, i.e., that the winter storm had led to *selective* elimination (like that predicted by the theory of natural selection).

Traits that Professor Bumpus measured on the birds included:

- Total weight (WT, measured in grams)
- The length of their beak and head (BH, measured in mm)
- The length of their humerus (HL, measured in mm)
- Their alar extent, which is the length from tip to tip of the extended wings (AE, measured in mm)
- Their age (AG, classified as adult and juvenile)

¹Natural selection was first proposed as a mechanism for evolution in 1858, when Charles Darwin theorized that some organisms possess heritable traits (like larger brains or longer limbs) that make them better suited to their environments and thus more likely to survive. The surviving organisms then pass these beneficial traits on to their offspring. This process of selection allows species to change and evolve over time: beneficial traits that increase the likelihood of survival become more common while harmful traits die out.

Each of the three logistic regression models below (shown on the next two pages) models the probability that a bird survived the storm as a function of one or more of these traits.

Model A:

Call:

```
glm(formula = Survived ~ WT + BH + HL + AE, family = "binomial",
     data = bumpus)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.22732	18.33765	-0.285	0.775598
WT	-0.85487	0.25374	-3.369	0.000754 ***
BH	0.44844	0.46460	0.965	0.334437
HL	2.17097	0.74456	2.916	0.003548 **
AE	-0.10921	0.08756	-1.247	0.212333

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom
Residual deviance: 94.943 on 82 degrees of freedom
AIC: 104.94

Number of Fisher Scoring iterations: 4

Model B:

Call:

```
glm(formula = Survived ~ AG, family = "binomial", data = bumpus)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.37729	0.26502	1.424	0.155
AGjuvenile	-0.08961	0.46483	-0.193	0.847

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.01 on 86 degrees of freedom
Residual deviance: 117.97 on 85 degrees of freedom
AIC: 121.97

Number of Fisher Scoring iterations: 4

Model C:

Call:

```
glm(formula = Survived ~ WT + BH + HL + AE + AG, family = "binomial",  
     data = bumpus)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.62095	18.49013	-0.250	0.80265
WT	-0.87329	0.25937	-3.367	0.00076 ***
BH	0.43656	0.46733	0.934	0.35022
HL	2.30616	0.77053	2.993	0.00276 **
AE	-0.11780	0.08848	-1.331	0.18310
AGjuvenile	-0.47007	0.55237	-0.851	0.39477

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom
Residual deviance: 94.213 on 81 degrees of freedom
AIC: 106.21

Number of Fisher Scoring iterations: 4

Please use this output to assist you in answering the following questions, showing all of your work:

(a) Using Model A, conduct a likelihood ratio test (by hand) to answer the question: is the probability of survival associated with one or more of the physical characteristics that Bumpus collected? *If so, this would be consistent with the theory of natural selection!* In your response, please:

- Provide the null and alternative hypotheses for this test
- Report the deviances for the full model and the reduced model that you are comparing
- Calculate the observed value of the test statistic
- Find the corresponding p -value using a χ^2 distribution with the appropriate degrees of freedom
- Report the conclusions of your hypothesis test in a sentence, taking $\alpha = 0.05$.

(b) An inquisitive member of the audience is concerned that any survival differences detected in part (a) might be attributable to confounding: because the sample of 87 birds includes both adult and juvenile sparrows, it's possible that these *age differences* (and not the physical traits themselves) explain the different probabilities of survival. Conduct a second likelihood ratio test that answers the question: among sparrows of the same age (either adult or juvenile), is the probability of survival associated with one or more of the physical characteristics that Bumpus collected? *Once again, please include the same five bulleted items in your response.*

(c) Using Model C, which physical trait(s) are associated with significantly *improved* survival (at the $\alpha = 0.05$ level), after accounting for age and the other traits in the model? Please interpret the corresponding coefficient(s) in context.

(d) Using Model C, which physical trait(s) are associated with significantly *decreased* survival (at the $\alpha = 0.05$ level), after accounting for age and the other traits in the model? Please interpret the corresponding coefficient(s) in context.