Logistic Regression

School of Information Studies
Syracuse University

Learning Topics for This Week

Review of linear prediction

Predicting categorical outcomes with logistic regression

The logistic curve

Generalized linear model and link functions

Simple example of logistic regression with GSR Data

Log odds and odds/overall interpretation

Bayesian estimation of logistic regression

Learning goal: by the conclusion of this class, students should be able to recognize the analytic circumstances where it is useful to predict a dichotomous (binary) outcome using one or more predictor variables; students should be able to explain how the logistic curve is useful for modeling a dichotomous outcome; students should be able to apply traditional and Bayesian inference techniques to examine the usefulness of a set of predictors

Review of Linear Regression

Review of Linear Regression

Last week: a method of predicting a metric variable from a set of one or more other variables (which are generally but not necessarily also metric variables)

Linear multiple regression uses least squares fitting to minimize the sum of squared errors of prediction from the line, plane, hyperplane, or higher dimensional object

The result is a simple, linear equation that multiplies each predictor by its respective "B" weight and adds an intercept to calculate a predicted Y

Using traditional statistical methods an omnibus F-test provides a significance test on R-squared, while t-tests provide tests on Bs

Predicting Categorical Outcomes

This Week

A method of predicting a **categorical** variable from a set of one or more other variables (which are generally but not necessarily also metric variables)

The simplest categorical variable is a binary outcome: on/off, yes/no, success/failure, etc.

The simplest analysis technique for predicting categorical variables is binomial logistic regression:

- Binominal because it predicts a binary outcome
- Logistic because it uses the continuous "logistic" (inverse logit) function to model the binary outcome
- Regression because it is another form of a prediction model, just like linear multiple regression

Examples of Predicting Binary Outcome

Logistic regression facilitates the prediction of a dichotomous (binary) outcome variable (yes/no, true/false, etc.) using a combination of one or more metric predictor variables

- Using a measure of galvanic skin response to predict if someone is telling the truth or lying
- Using a person's income and education level to predict whether they will vote for Candidate A or Candidate B in an election
- Predicting whether a law student will pass the bar exam on the first try based on their LSAT entrance exam score and their first year GPA

There are other kinds of categorical prediction that can predict an outcome with more than two options (multinomial); we will not cover those in this course

The Logistic Curve

Euler and “e”

“e” is a mathematical constant, equal to about 2.718282, but with an infinite number of digits

Considered one of the five most important numbers in mathematics (others include pi and zero), e was discovered by Jakob Bernoulli during his study of compound interest

Important proofs about e were accomplished by Leonhard Euler (1707-1783)

In R, try exp(1)

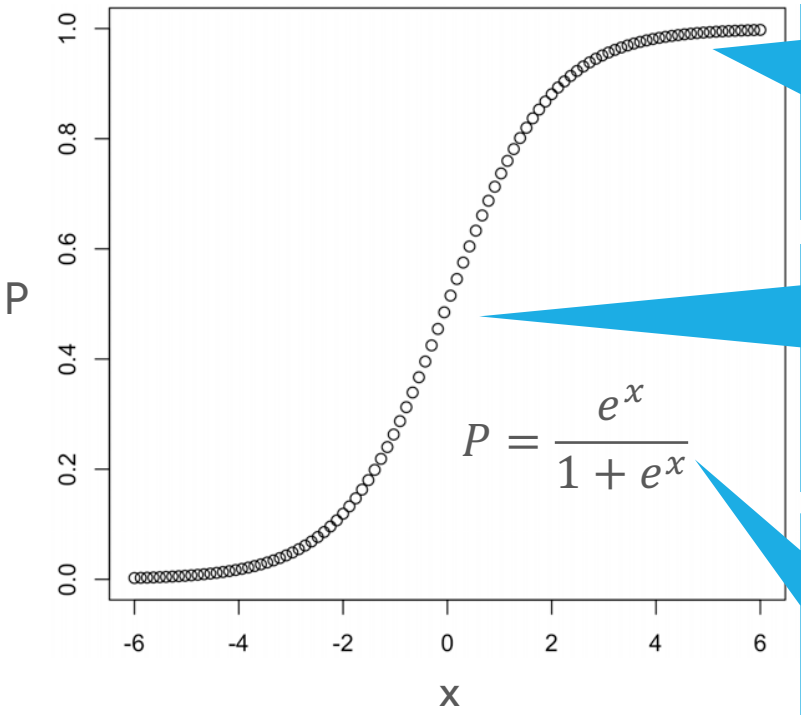
Portrait of Leonhard Euler by J. E., Handmann; image by Kunstmuseum Basel, Public Domain, https://commons.wikimedia.org/w/index.php?curid=893656

The Inverse Logit (Logistic) Curve

The logistic curve is an S-shaped curve with metric values on the X-axis and probabilities on the Y-axis

The "e" in the equation is Euler's number, about 2.718

The curve is useful for modeling a binary outcome, i.e., where the value on the Y-axis is in reality either 0 or 1



The actual Y values we are trying to predict are either zero or one

Steep transition from low probability to high probability when x is near zero

Put a coefficient 'b' in front of each x; logistic regression finds the value of b

Generalized Linear Model

General vs. Generalized

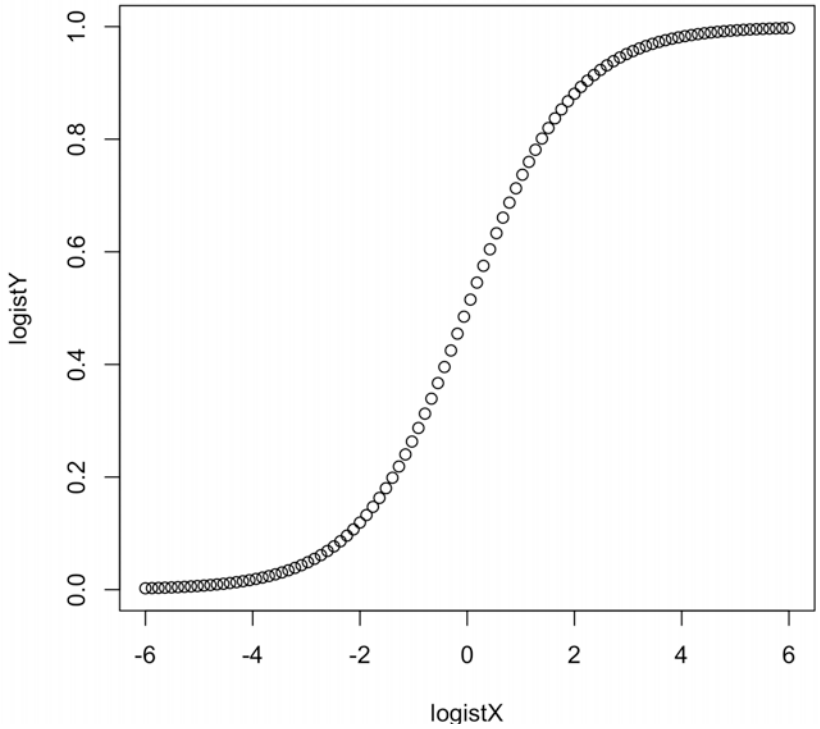
The General Linear Model	The Generalized Linear Model
A special case of the generalized linear model—includes linear multiple regression and ANOVA	A more general method that includes the general linear model
The link function is "identity"	Many possible link functions, including "inverse logit" that we use in this chapter
Outcome variable is metric and normally distributed	Outcome variable may be binomial, Poisson, etc.
Residuals are normally distributed	Residuals need not be normally distributed
Uses ordinary least squares fitting to find solution	Use Maximum Likelihood Estimation to find solution

Link Functions

A “link function” connects a metric predictor to the outcome variable through a mathematical formula

With linear multiple regression from chapter 8, the link function was what mathematicians call “identity” because the linear form of the predictor identically matched the linear form of the outcome

The link function for binomial logistic regression is the inverse logit function previously discussed (and shown again at right)



Maximum Likelihood Fitting

The goal of the fitting process is to locate values for the coefficient(s), i.e., the weights on the predictors and the value of the intercept, that would maximize the probability of observing all data points in the sample

This is somewhat similar in thinking to a Bayesian approach although in this case there is no concern for prior probabilities

This equation represents the function whose value we are trying to optimize by adjusting the beta weight (for a one predictor model):

$$-LL(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i \ln\left(1 + e^{-x_i\boldsymbol{\beta}}\right)\right) - (1\text{-}\, y_i)(\ln\left(1 + e^{x_i\boldsymbol{\beta}}\right))$$

The function must be tested repeatedly while varying beta; this can be computationally expensive

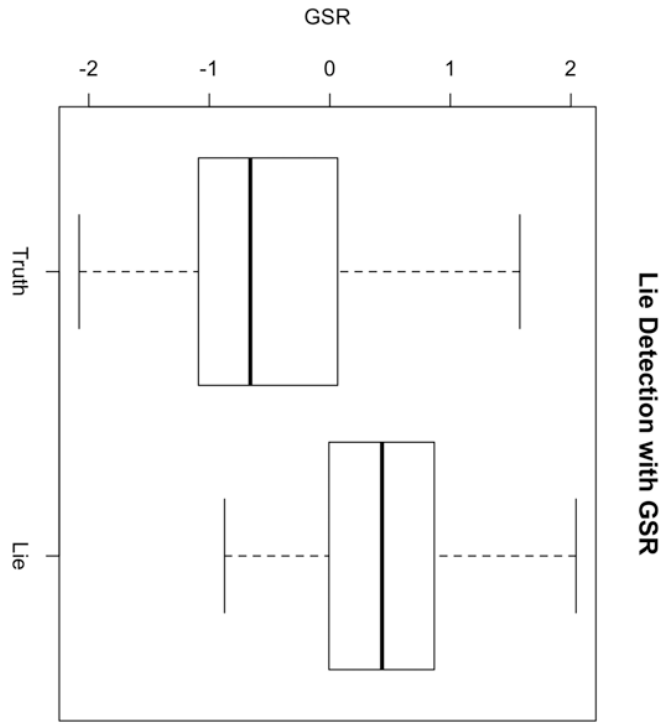
Simple Example

Fake GSR Example

```
set.seed(123)
logistX <- rnorm(n=100,mean=0,sd=1)
logistY <- exp(logistX)/(exp(logistX)+1)
binomY <- round(logistY)
logistX <- logistX/1.41 + rnorm(n=100,mean=0,sd=1)/1.41
binomY <- factor(round(logistY), labels=c('Truth','Lie'))
logistDF <- data.frame(logistX, logistY, binomY)
boxplot(formula=logistX ~ binomY, data=logistDF, ylab="GSR",
main=NULL)
```

Using a Boxplot to Examine GSR Data

Turning the diagram 90 degrees puts the X-axis where it belongs: we are showing the distributions of the X variable for each of the two values of Y



Interpreting the Logistic Regression

```
glmOut <- glm(binomY ~ logistX, data=logistDF, family=binomial())
summary(glmOut)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3216  -0.7982   0.3050   0.8616   1.7414

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1199     0.2389   0.502    0.616
x             1.5892     0.3403   4.671 3e-06 ***

Null deviance: 138.47  on 99  degrees of freedom
Residual deviance: 105.19  on 98  degrees of freedom
AIC: 109.19

Number of Fisher Scoring iterations: 4
```

The difference between the null model and the residual model is distributed as chi-square and can be used as an omnibus test

The Wald z-test on the predictor is significant, but what is the meaning of the estimate?

Omnibus Test and Interpretation

> anova(glmOut, test="Chisq") # Compare null model to one predictor model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			99	138.47	
x	1	33.279	98	105.19	7.984e-09 ***

The chi-square test shows a difference of 33.3 on one degree of freedom, which is statistically significant. Adding the predictor, which cost one degree of freedom, was "worth it" because it significantly reduced the residual deviance (the amount of variation among residuals).

We can also obtain a "pseudo-R-squared" using the PseudoR2() command in the "BaylorEdPsych" package. There is no one version of pseudo-R-squared accepted by all statisticians.

Log Odds and Odds

Probabilities and Odds

The term "odds" is used in everyday life to express probability. 5:1 odds on winning a race is pretty good.

Odds and probability are related: Odds = P/(1-P) so P = Odds/(1 + Odds)

So odds of 5:1 is the same as saying P=0.83. Raising the odds from 5:1 up to 10:1 increases P from 0.83 to 0.91.

The output of logistic regression calculates the prediction coefficients as the **natural logarithm of the odds ratio**. We need to convert this to something more interpretable.

Try this code:

```
logOdds <- seq(from=-2, to=2, length.out=100)

plot(logOdds, exp(logOdds))
```

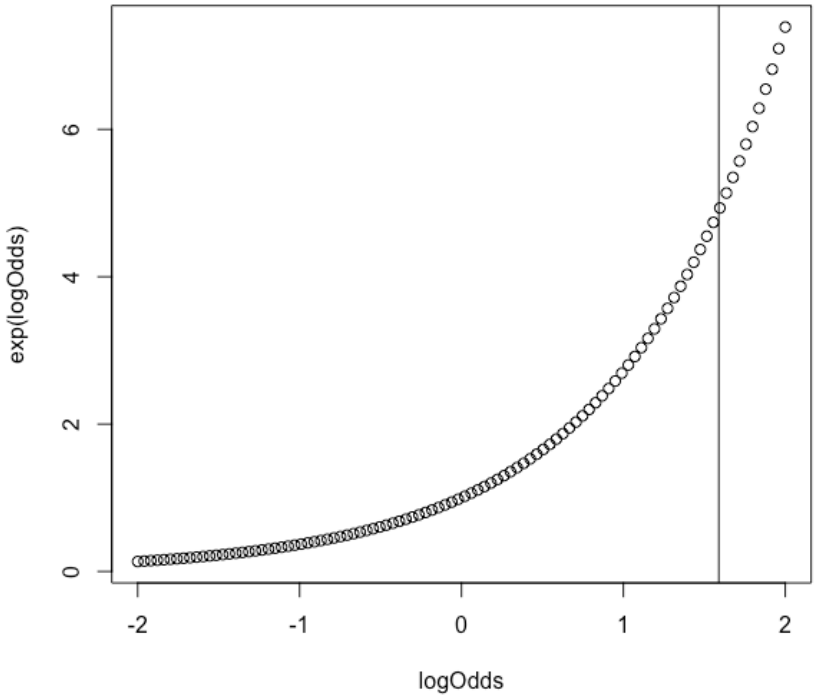
Converting Log Odds to Odds

Use a simple command to convert from log-odds to odds:

```
> exp(coef(glmOut)) # Convert log odds to odds
  (Intercept)      logistX
  1.127432   4.900041
```

So a log odds of 1.5892 converts to 4.9 in plain odds. Try: abline(v=1.5892)

For each unit change in the value of X, odds that Y=1 is the correct prediction increase by 4.9:1.



Confidence Intervals Around Plain Odds

```
> exp(confint(glmOut))
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) 0.7057482  1.812335
logistX     2.6547544 10.190530
```

Overlaps with 1:1 (a "null" result)

Interpretation: the 95% confidence interval for GSR ranged from 2.65:1 up to 10.2:1, expressed in plain odds; if the study was repeated 100 times, 95% of similarly constructed intervals would contain the true population value

Putting It All Together: Part 1

We tested a measure of galvanic skin response (GSR) that varied from about -3 (dry) to +3 (wet) to see if it could predict the truthfulness of statements spoken by a research participant. A chi-square omnibus test on the results of logistic regression was significant, chisq(1) = 32.3, $p<.001$. A Wald's z-test on the GSR coefficient was also significant, $z=4.67$, $p<.001$.

When converted to odds, the coefficient on GSR was 4.9, suggesting that for each unit increase in GSR, the odds of the statement being a lie increased by 4.9:1. This was strong evidence suggesting that GSR could serve as a useful lie detector test.

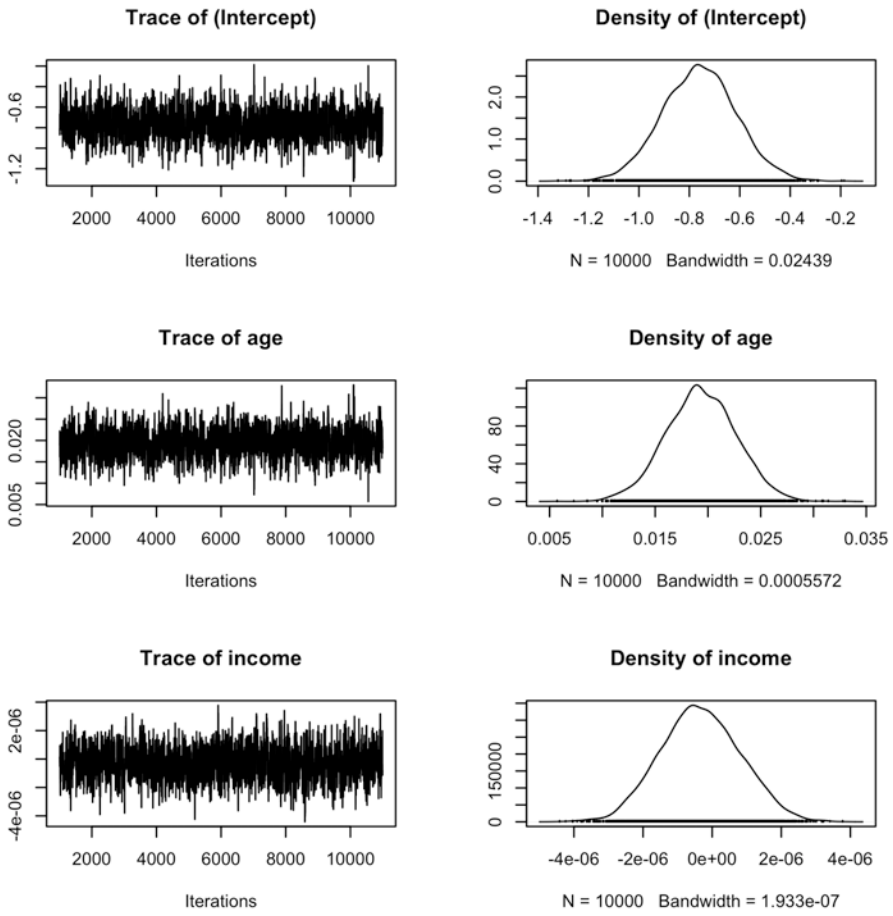
Bayesian Example With Real Data

Bayesian Approach With MCMClogit()

Using the Chile vote data described in the book, the research question is whether age and income predicted Chileans' votes on a plebiscite in 1988. Voting "Yes" was a vote to keep then-president Augusto Pinochet in office.

```
> bayesLogitOut <-
MCMClogit(formula = vote ~ age + income, data = ChileYN)
> summary(bayesLogitOut)
> plot(bayesLogitOut)
```

Bayesian Approach With MCMClogit()



Trace plots show the progress of the MCMC estimation process.

Density plots show the posterior distribution of each coefficient. Income is centered near zero. What does that mean?

HDIs Provide the Details

The mean value of each coefficient is the "point estimate" at the center of the density distribution. These are fairly close to the output of glm().

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
(Intercept)	-7.602e-01	1.452e-01	1.452e-03	4.849e-03
age	1.935e-02	3.317e-03	3.317e-05	1.089e-04
income	-3.333e-07	1.151e-06	1.151e-08	3.723e-08

2. Quantiles for each variable:

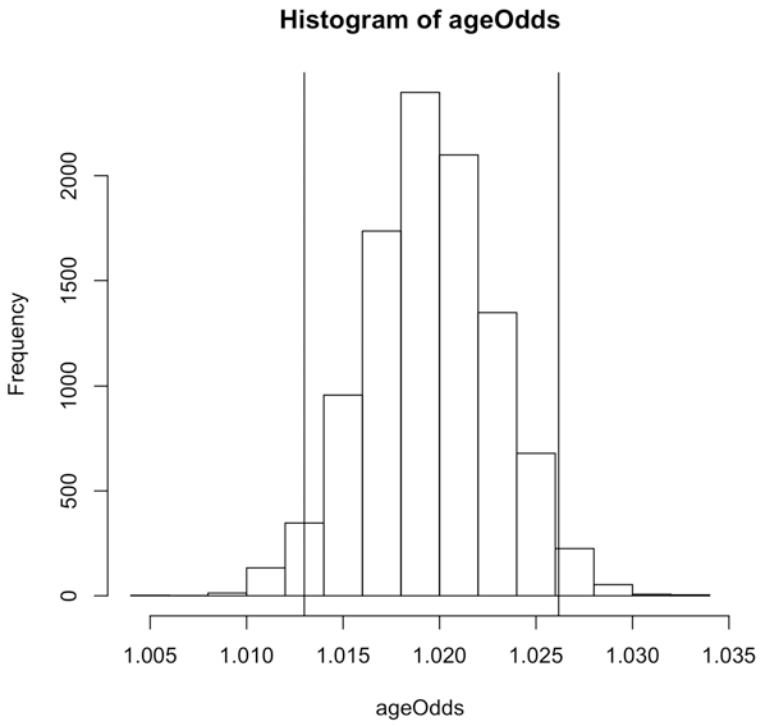
	2.5%	25%	50%	75%	97.5%
(Intercept)	-1.044e+00	-8.588e-01	-7.609e-01	-6.639e-01	-4.710e-01
age	1.278e-02	1.711e-02	1.930e-02	2.157e-02	2.584e-02
income	-2.549e-06	-1.113e-06	-3.662e-07	4.437e-07	1.926e-06

The 2.5% and 97.5% values mark the extent of the 95% HDI for each coefficient. Does the HDI for "age" overlap with zero?

Convert MCMC Output to Odds

```
> ageLogOdds <- as.matrix(bayesLogitOut[,"age"])
> ageOdds <- apply(ageLogOdds,1,exp)
> hist(ageOdds)
> abline(v=quantile(ageOdds,c(0.025)), col="black")
> abline(v=quantile(ageOdds,c(0.975)), col="black")
```

Odds of 1.02:1 for the coefficient on the age predictor are tricky to interpret because age is calibrated in years and a single year obviously does not make much difference in odds. What if you divided the age variable by 10 and ran the analysis again?



Confusion Matrix: How Accurate Are Predictions?

```
chOut <- glm(formula = vote ~ age + income, family = binomial(), data = ChileYN)

actualVote <- ChileYN$vote

predictedVote <- round(predict(chOut, type="response")) # round() splits probabilities at 0.5

table(predictedVote, actualVote)
```

	N	Y
0	565	449
1	302	387

The actual vote is shown in the columns and the predictions (probabilities rounded to 0 or 1) are represented in the rows.

Incorrect predictions are on the off-diagonal. So one measure of error rate would be (449+302)/(565+387+449+302) = 44%

Putting It All Together: Part 2

We examined data from the 1988 Chilean plebiscite, to see if the age and income of a voter could predict whether an individual would vote in favor of keeping Augusto Pinochet in office. We conducted a Bayesian logistic analysis, using age and income to predict votes. The posterior distribution of the coefficient for income (calibrated as log odds) overlapped squarely with zero, suggesting that income was not a meaningful predictor of votes. In contrast, the Highest Density Interval of age did not overlap with zero. When converted to regular odds, the mean value of the posterior distribution for age was 1.02 to 1, suggesting that for every additional year of age, an individual was about 2% more likely to vote to keep Pinochet. However, a confusion matrix showed that the overall error rate was 44% indicating that the logistic model was not particularly good at predicting votes.

Independent Work: Redo Chilean Votes Analysis With GLM