ISE 5103 Intelligent Data Analytics

Logistic Regression

Charles Nicholson, Ph.D.
cnicholson@ou.edu

University of Oklahoma
Gallogly College of Engineering
School of Industrial and Systems Engineering

Outline

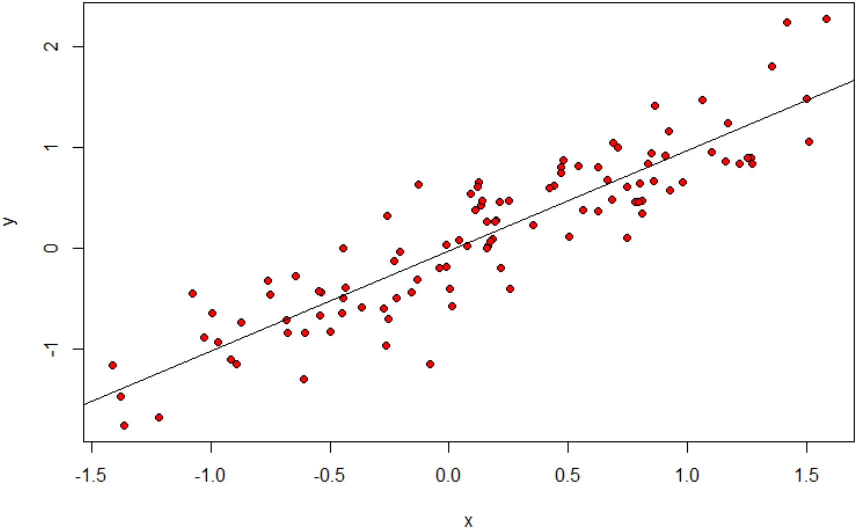
1. Logistic Regression Theory
2. Logistic Regression Example
3. Diagnostics

MLR: goals and terminology

Goal: Find linear function relating $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ji} + \epsilon_i$$

- $i = 1, \ldots, n$ rows of data
- p predictors (which may have been engineered)
- $\epsilon \sim N(0, \sigma^2)$
- $\hat{\beta}$ are found using a closed-form solution
- Fitted value: $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ji}$
- $e_i = y_i - \hat{y}_i$


y
-1
0
1
2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
x

logistic regression

Logistic regression is a classification technique; specifically, we will consider dichotomous targets.

Example 1. Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (target) variable is binary (0/1); win or lose.

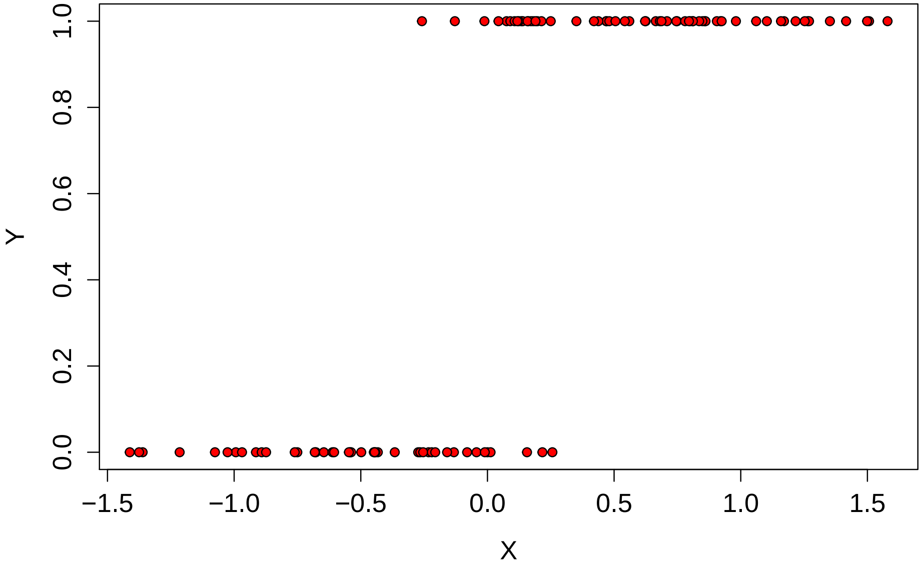
Example 2. A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

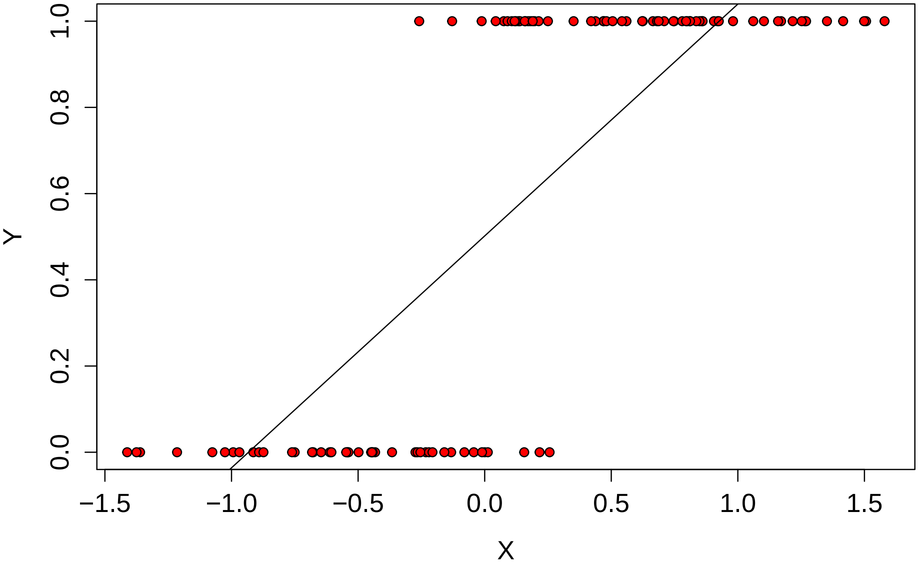
Examples... Health: benign, malignant; Customer: buy/no buy; Finance: pay back loan, don't pay back; fraud: fraudulent or not

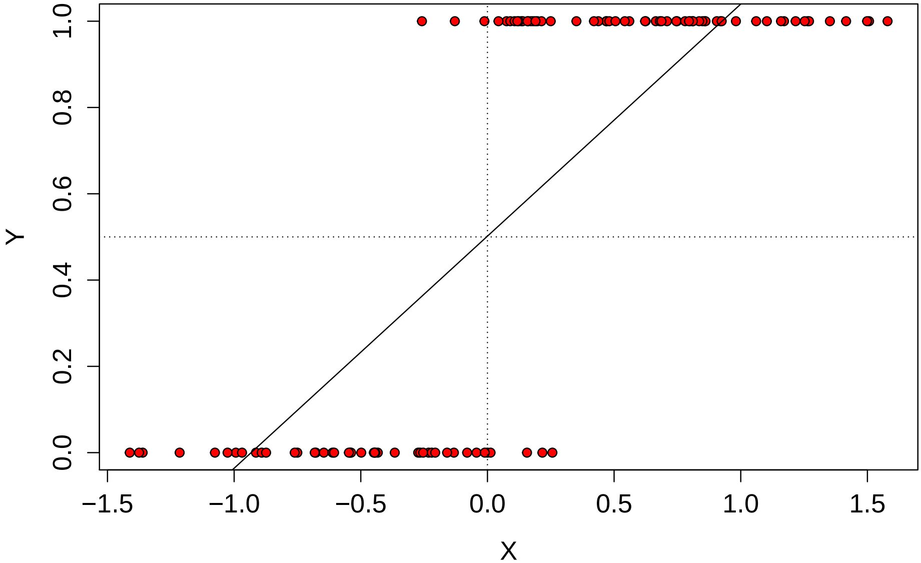
Logistic regression shares many similarities with linear regression.

But first,

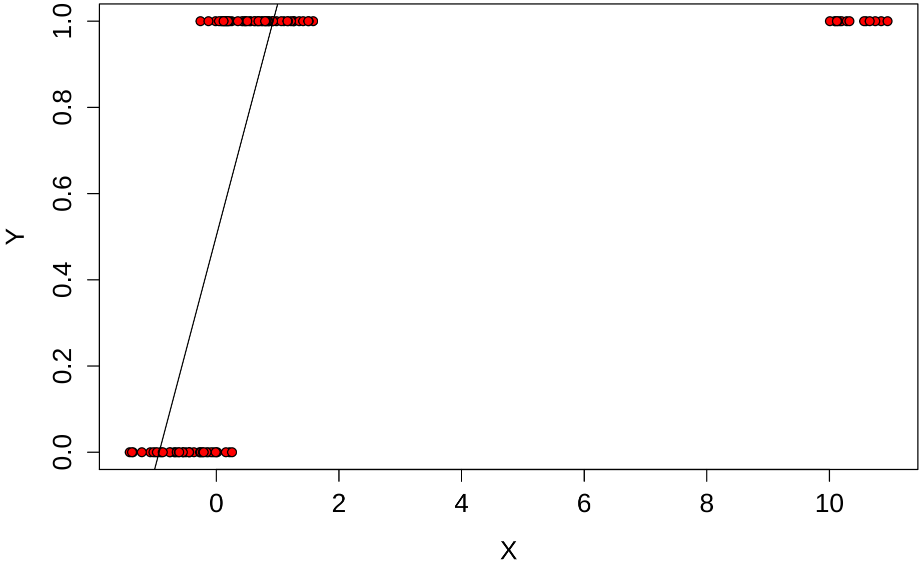
- let's look at the differences
- and understand why the linear regression approach must be modified
- but then, let's try to get back as close to linear regression as possible!


Y
1.0
0.8
0.6
0.4
0.2
0.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X

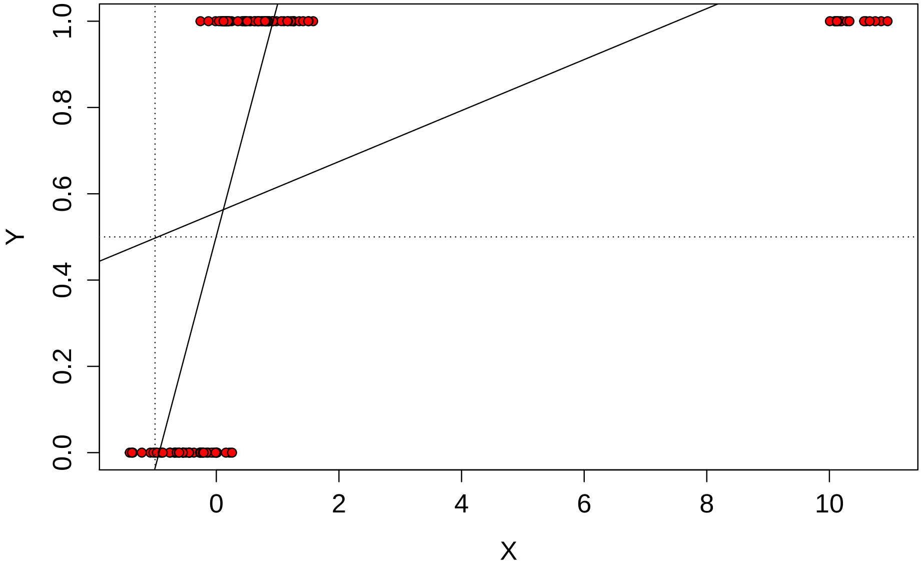

1.0
0.8
0.6
0.4
0.2
0.0
Y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X



Y
1.0
0.8
0.6
0.4
0.2
0.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X



Y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
X


0.0
0.2
0.4
0.6
0.8
1.0
Y
0
2
4
6
8
10
X

Problem: The target is not continuous.
Idea: Model probability, not the value.

Instead of modeling the binary target directly, ($Y = 1$ or $Y = 0$), model the conditional probability:

$$\pi = P(Y = 1|X)$$

Problem: $\pi \in [0, 1]$.

- Most obvious idea: model π as linear function of X. Every increment in X would add (subtract) to (from) the probability:

$$\pi = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

  But, linear functions are unbounded and $\pi \in [0, 1]$.

- Empirically, we see “diminishing returns”
  - changing π by the same amount requires a bigger change in X when π is already large (or small) than when π is close to 0.5.
  - Linear models can’t do this.

Idea: Model the odds, not the probability.

what are odds?

The odds of success are defined as the ratio of the probability of success over the probability of failure.

Odds Example:

- Probability of success is 0.8; so, probability of failure is 0.2
- Odds of success are: $\frac{0.8}{0.2} = 4$
- i.e, the odds are 4 to 1

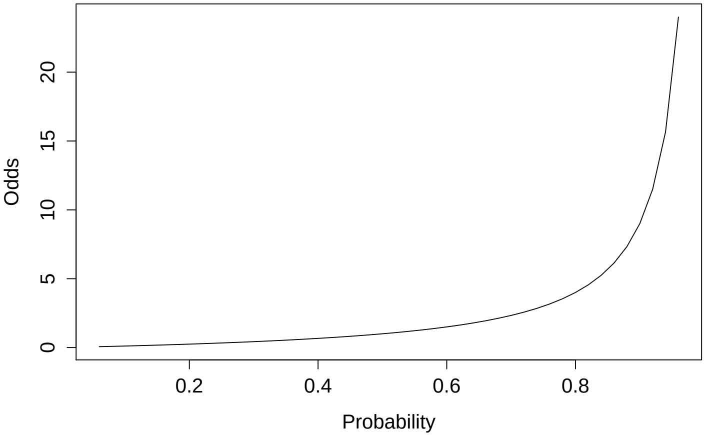
The transformation from probability to odds is a monotonic transformation.

Probability ranges from 0 to 1. Odds range from 0 to ∞.

Let's look at the previous example statement,
"changing π by the same amount requires a bigger change in X when π is already large (or small) than when π is close to 0.5."

- Odds for 50% and 51% are 1 and 1.04
  - a 4% increase in odds
- Odds for 98% and 99% are 49 and 99
  - a 102% increase in odds
- Odds for 99% and 99.9% are 99 and 999
  - a 909% increase in odds
- Odds for 99% and 100% are 99 and NAN
  - an infinite increase in odds

probability	odds
.001	.0010
.01	.0101
.1	.1111
.25	.3333
.3	.4286
.4	.6667
.5	1
.75	3
.9	9
.999	999
.9999	9999

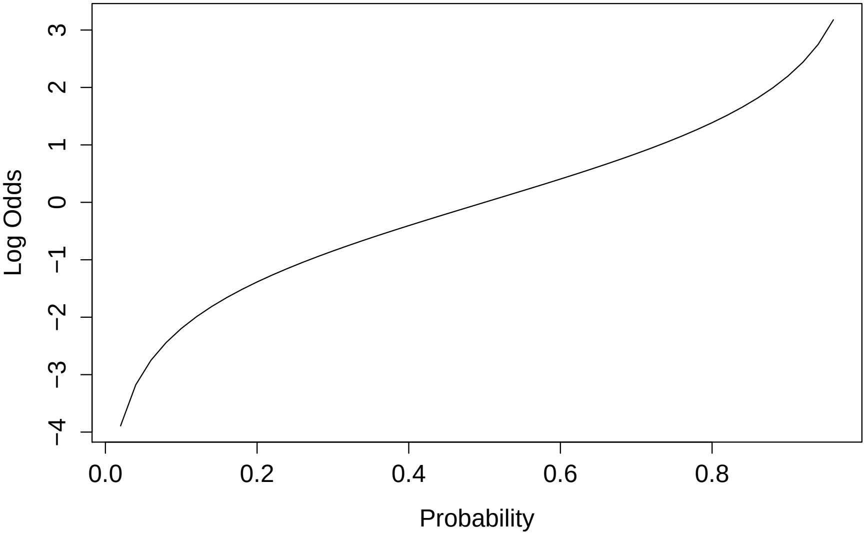


Problem: odds are bounded at 0.
Solution: Model the log of the odds:

$$\log(\text{odds}) = \log \frac{\pi}{1 - \pi}$$

- This maps the probability ranging between 0 and 1 to log odds ranging from negative infinity to positive infinity.
- Another reason is that among all of the infinitely many choices of transformation, the log of odds is one of the easiest to understand and interpret.
- This transformation is called logit transformation.

probability	odds	log(odds)
.001	.0010	-6.9068
.01	.0101	-4.5951
.1	.1111	-2.1973
.25	.3333	-1.0986
.3	.4286	-0.8473
.4	.6667	-0.4055
.5	1	0
.75	3	1.0986
.9	9	2.1972
.999	999	6.9068
.9999	9999	9.2102


Log Odds
3
2
1
0
−1
−2
−3
−4
0.0
0.2
0.4
0.6
0.8
Probability

logistic regression model

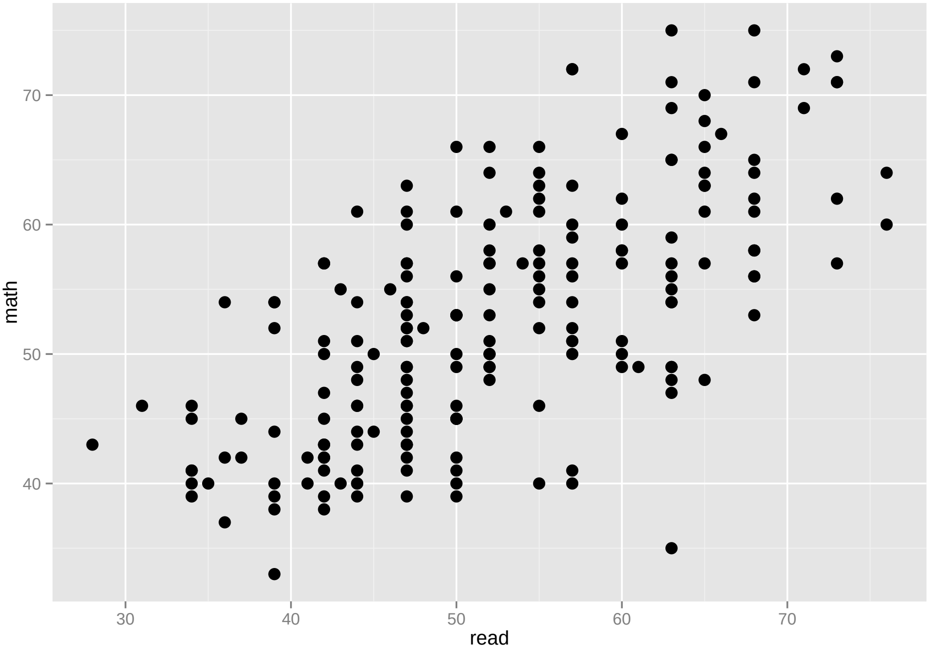
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

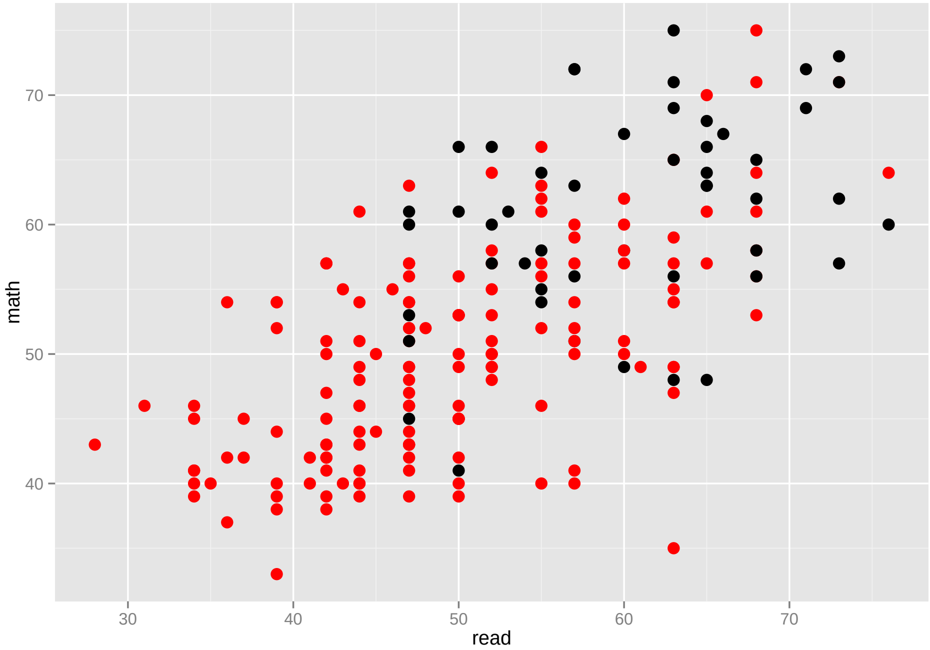
$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}$$

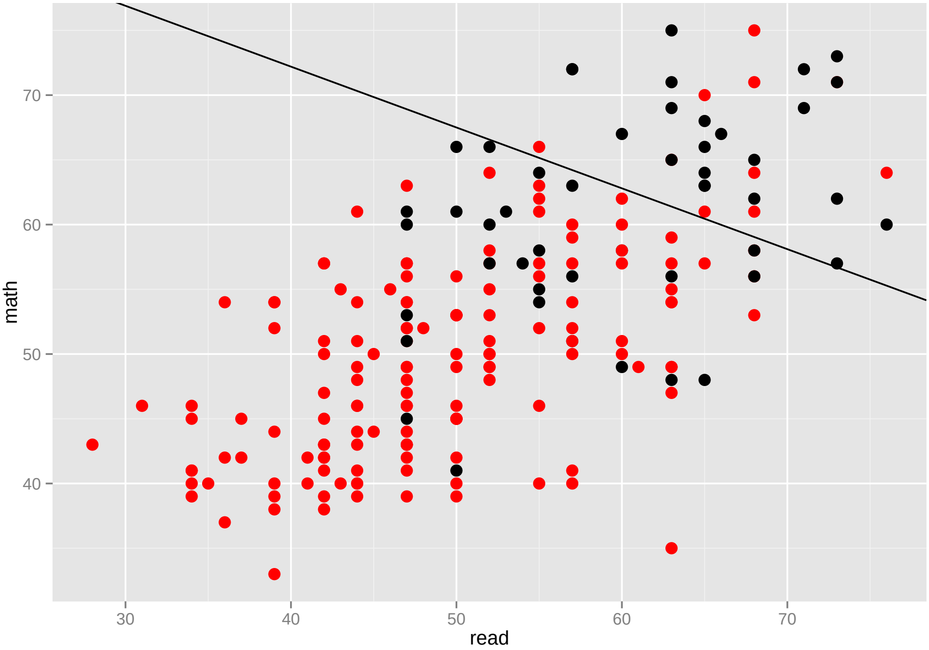
$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

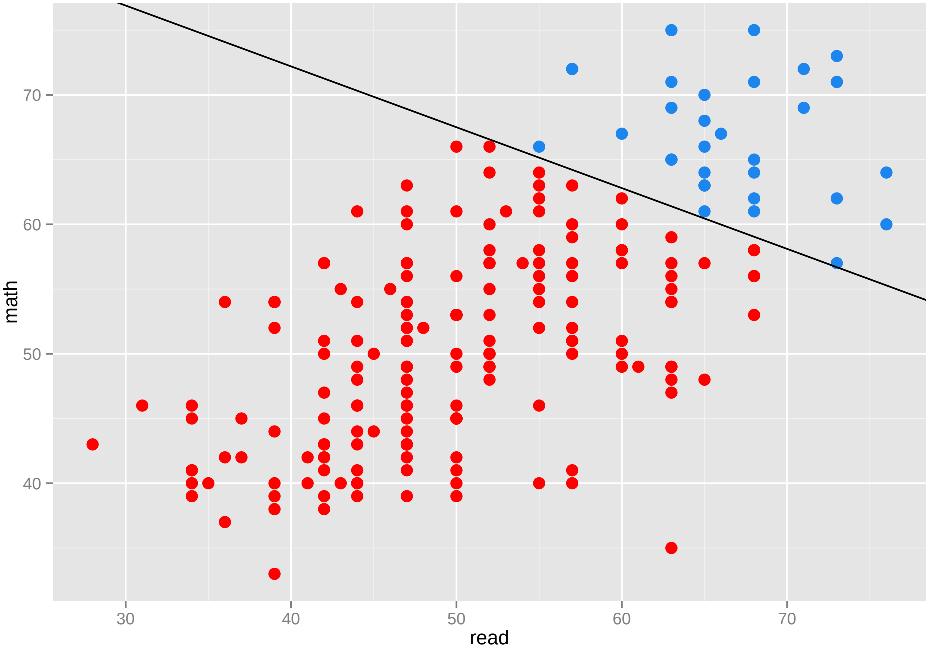
classification rule

- To minimize the mis-classification rate, predict: $Y = 1$ when $\pi \geq 0.5$; and $Y = 0$ when $\pi < 0.5$.
- That is, choose 1 whenever $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ is non-negative, and 0 otherwise.
- This means that logistic regression gives us a linear classifier.
- The decision boundary separating the two predicted classes is the solution of $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = 0$, which is a hyperplane.


70
60
50
40
math
30
40
50
60
70
read


70
60
50
40
math
30
40
50
60
70
read


70
60
50
40
math
30
40
50
60
70
read


70
60
50
40
math
30
40
50
60
70
read

finding β

Maximum-Likelihood Estimation of Parameters

If it is assumed that the n data observations are independent; the joint probability of the *observed* values of Y is:

$$P(y_1, \ldots, y_n) = \prod_{i=1}^{n} P(Y_i = y_i | X_{1i}, \ldots, X_{pi})$$

finding β

Ignoring the constant from the binomial distribution, the *likelihood function* expresses the values for β in terms of known values of Y:

$$L(\hat{\beta}) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$$

finding β

Taking the log of the likelihood function is a convenient step:

$$
\begin{aligned}
\log L(\hat{\beta}) &= \log\left(\prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}\right) \\
&= \sum_{i=1}^{n} \left[y_i \log \pi + (1-y_i)\log(1-\pi)\right] \\
&= \sum_{i=1}^{n} \log(1-\pi) + \sum_{i=1}^{n} y_i\left(\log \pi - \log(1-\pi)\right) \\
&= \sum_{i=1}^{n} \log(1-\pi) + \sum_{i=1}^{n} y_i \log \frac{\pi}{1-\pi}
\end{aligned}
$$

finding β

$$
\begin{aligned}
\log L(\hat{\beta}) &= \sum_{i=1}^{n} \log(1-\pi) + \sum_{i=1}^{n} y_i \log \frac{\pi}{1-\pi} \\
&= \sum_{i=1}^{n} \log \frac{1}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ji}}} + \sum_{i=1}^{n} y_i \left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ji} \right) \\
&= \sum_{i=1}^{n} -\log \left(1 + e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ji}} \right) + \sum_{i=1}^{n} y_i \left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ji} \right)
\end{aligned}
$$

finding β

Take the partial derivatives of $\log L(\hat{\beta})$ with respect to each $\hat{\beta}$, and set to 0. Solve the $p+1$ system of non-linear equations.

e.g. for β_j:

$$\begin{aligned}
\frac{\partial \log L}{\partial \hat{\beta}_j} &= \sum_{i=1}^{n} \frac{-1}{1+e^{\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j X_{ij}}} \frac{\partial}{\partial \hat{\beta}_j}\left(1+e^{\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j X_{ij}}\right) + \sum_{i=1}^{n} y_i X_{ij} \\
&= \sum_{i=1}^{n} \frac{-e^{\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j X_{ij}}}{1+e^{\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j X_{ij}}} \frac{\partial}{\partial \hat{\beta}_j}\left(\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j X_{ij}\right) + \sum_{i=1}^{n} y_i X_{ij} \\
&= \sum_{i=1}^{n} \left(-\pi + y_i\right) X_{ij}
\end{aligned}$$

finding β

Important note: we cannot solve the p+1 system of equations analytically, we can solve it approximately using numeric methods.

There are many choices for numeric optimization, but the most common and important technique, the Newton-Raphson method was developed over 300 years ago (this is even pre-dates the iPhone 3!)

finding β

The Newton-Raphson method is an iterative technique that (hopefully and usually) converges to a numeric solution.

However, there are some cases in which this optimization approach will not converge. The most common reasons are: *complete separation* and *quasi-complete separation*.

See the white paper “Convergence Failures in Logistic Regression” by P. Allison (2008) for more details on both the Newton-Raphson method and separation problems.

Outline

1. Logistic Regression Theory
2. **Logistic Regression Example**
3. Diagnostics

example problem

Honors Students

- Download data file from course website: “honors.csv”
- Download the script file: “LogRegHonors.R”
- 200 observations relating to students, including gender, test scores, and whether or not they are an “honors” student
- The target `hon` is binary: if `hon` = 1, the student is an honors student

Example from: UCLA: Statistical Consulting Group. http://www.ats.ucla.edu/stat/mult_pkg/faq/general/odds_ratio.htm/

example problem

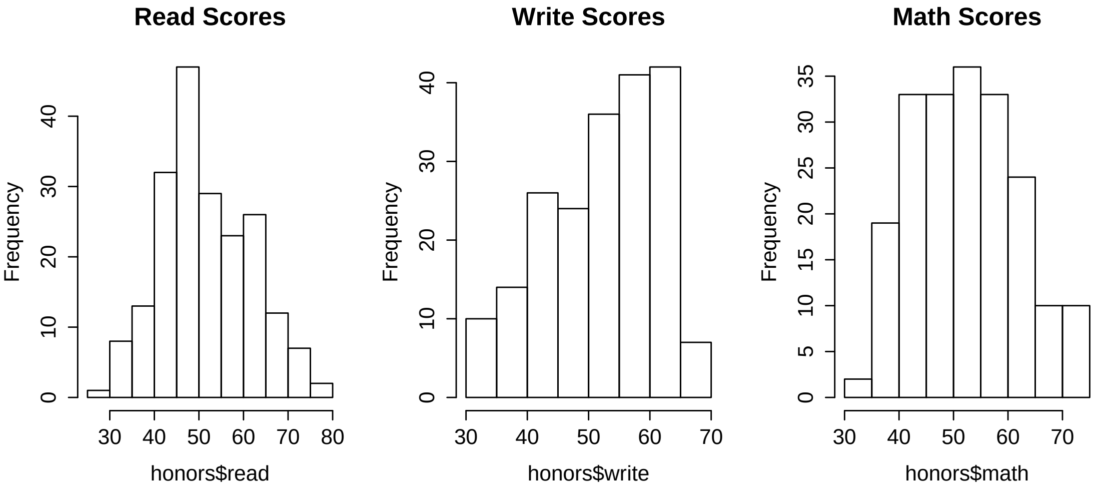
```
head(honors)
```

```
  female read write math hon
1      0   57    52   41   0
2      1   68    59   53   0
3      0   44    33   54   0
4      0   63    44   47   0
5      0   47    52   57   0
6      0   44    52   51   0
```

54.5% are female

24.5% are honors students

example problem



example: no predictors

Let’s start with the simplest logistic regression, a model without any predictor variables.

In this case, we want to model:

$$\log \frac{\pi}{1-\pi} = \text{logit}\,\pi = \beta_0$$

example: no predictors

Coefficients:

	Estimate	Std. Error	z value	$\Pr(> \|z\|)$	
(Intercept)	-1.1255	0.1644	-6.845	7.62e-12	***

$$\log \frac{\pi}{1-\pi} = \hat{\beta}_0 = -1.1255$$

hon	freq
0	151
1	49

- $\pi = \frac{49}{200} = 0.245$
- odds: $\frac{0.245}{1-0.245} = .3245$
- log of odds: $\log(0.3245) = -1.1255$

example: one nominal predictor

$$\text{logit}\ \pi = \beta_0 + \beta_1 \text{female}$$

	Estimate	Std. Error	z value	$\Pr(>\lvert z\rvert)$	
(Intercept)	-1.4709	0.2690	-5.469	4.53e-08	***
female	0.5928	0.3414	1.736	0.0825	.

let's first look at the crosstab of the **hon** with **female**:

	female		Row
hon	0	1	Total
0	74	77	151
1	17	32	49
Col Total	91	109	200

what are the odds of a male being in the honors class and what are the odds of a female being in the honors class?

example: one nominal predictor

	female		Row
hon	0	1	Total
0	74	77	151
1	17	32	49
Col Total	91	109	200

for males: odds of being in honors: $\frac{17/91}{74/91} = \frac{17}{74} = 0.2297$

for females: odds of being in honors: $\frac{32/109}{77/109} = \frac{32}{77} = 0.4155$

- Ratio of “odds for female” to “odds for male”: $\frac{32/77}{17/74} = 1.809$.
- Odds for female are about 81% higher than odds for males.

logit $\pi = \beta_0 + \beta_1$ *female*

	Estimate	Std. Error	z value	Pr($>|z|$)	
(Intercept)	-1.4709	0.2690	-5.469	4.53e-08	***
female	0.5928	0.3414	1.736	0.0825	.

$\rightarrow$ if female=0, then: logit $\pi = \beta_0$

- The intercept of -1.4709 is the log odds for males since male is the reference group (i.e. female = 0).
- That is, $e^{\beta_0} = e^{-1.471} = 0.2297 =$ odds of a male being in honors.

$\rightarrow$ if female=1, then: logit $\pi = \beta_0 + \beta_1$

- So, $e^{\beta_0+\beta_1} = e^{-1.4709+0.5928} = 0.4155 =$ odds of female in honors
- The coefficient for `female` is the log of odds ratio between the female group and male group: log(1.809) = .5928
- So we can get the *odds ratio* by exponentiating the coefficient for female.

example: one continuous predictor

$$\text{logit } \pi = \beta_0 + \beta_1 \textit{math}$$

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$	
(Intercept)	-9.79394	1.48174	-6.610	3.85e-11	***
math	0.15634	0.02561	6.105	1.03e-09	***

- Here $\hat{\beta}_0$ is the log odds of a student with math score = 0 being in honors
- The odds then of being in an honors class when math = 0 is $e^{-9.79394} = .0000558$
- No student in the sample has math score lower than 30.
- The intercept in this model corresponds to the log odds of being in an honors class when math is at the *hypothetical* value of zero.

example: one continuous predictor

How do we interpret the coefficient for **math**?

$$\text{logit}\left(\pi\right) = -9.79394 + 0.15634 * \textbf{math}$$

To explain, let's fix **math** at some value, e.g. 54.

- The conditional logit of being in honors when the math score is held at 54 is: $\text{logit}(\pi)_{\text{math}=54} = -9.79394 + 0.15634 * \mathbf{54}$
- We can examine the effect of a one-unit increase in math score. When the math score is held at 55, the conditional logit of being in honors class is: $\text{logit}(\pi)_{\text{math}=55} = -9.79394 + 0.15634 * \mathbf{55}$
- Taking the difference of the two equations: $\text{logit}(\pi)_{\text{math}=55} - \text{logit}(\pi)_{\text{math}=54} = 0.15634$
- That is, for a one-unit increase in the math score, the expected change in log odds is 0.15634.
- In terms of odds: $e^{\hat{\beta}_1} = e^{0.15634} = 1.16922$

  $\longrightarrow$ for any one unit increase in math, there is about a 17% increase in the odds of being in honors.

example: multiple predictors

For multiple predictors, such as the following model:

$$\text{logit}\,\pi = \beta_0 + \beta_1 * \textit{math} + \beta_2 * \textit{female} + \beta_3 * \textit{read}$$

- Each estimated coefficient is the expected change in the log odds of being in an honors class for a unit increase in the corresponding predictor variable *holding the other predictor variables constant*.

example: multiple predictors

	Estimate	Std. Error	z value	Pr($>|z|$)	
(Intercept)	-11.77025	1.71068	-6.880	5.97e-12	***
math	0.12296	0.03128	3.931	8.44e-05	***
female	0.97995	0.42163	2.324	0.0201	*
read	0.05906	0.02655	2.224	0.0261	*

- For the same **math** and **reading** scores, the odds of honors for females (**female** = 1) over the odds of honors for males (**female** = 0) is $e^{0.97995} = 2.66$.
- That is, the odds for females are 166% higher than the odds for males.
- The coefficient for **math** says that, holding **female** and **read** at a fixed value, we will see 13% increase in the odds of getting into an honors class for a one-unit increase in math score since $e^{0.12296} = 1.13$.

example: interaction term

$$\text{logit}\,\pi = \beta_0 + \beta_1 * \mathit{math} + \beta_2 * \mathit{female} + \beta_3 * \mathit{female} * \mathit{math}$$

- Without interaction terms: the regression coefficient of a variable corresponds to the change in log odds and its exponentiated form corresponds to the odds ratio.
- When a model has interaction term(s), it attempts to describe how the effect of a predictor variable depends on the level/value of another predictor variable.

example: interaction term

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$	
(Intercept)	-8.74584	2.12913	-4.108	4e-05	***
math	0.12938	0.03588	3.606	0.000312	***
female	-2.89986	3.09418	-0.937	0.348657	
math:female	0.06700	0.05346	1.253	0.210139	

- In the presence of the term **female** × **math**, we can no longer talk about the effect of **female**, holding all other variables fixed.

- Here, we actually have two equations:

males: $\text{logit}(\pi) = \beta_0 + \beta_1 * math$

females: $\text{logit}(\pi) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) * math$

example: interaction term

	Estimate	Std. Error	z value	Pr($>\|z\|$)	
(Intercept)	-8.74584	2.12913	-4.108	4e-05	***
math	0.12938	0.03588	3.606	0.000312	***
female	-2.89986	3.09418	-0.937	0.348657	
math:female	0.06700	0.05346	1.253	0.210139	

i. For males: logit$(\pi) = \beta_0 + \beta_1 * math$

- $\hat{\beta}_1$ for **math** is the effect of **math** when **female** = 0
- i.e. for male students, a one-unit increase in math score yields a change in log odds of 0.13
- the odds ratio is $e^{0.13} = 1.14$ for a one unit increase in math.

example: interaction term

	Estimate	Std. Error	z value	Pr($>|z|$)	
(Intercept)	-8.74584	2.12913	-4.108	4e-05	***
math	0.12938	0.03588	3.606	0.000312	***
female	-2.89986	3.09418	-0.937	0.348657	
math:female	0.06700	0.05346	1.253	0.210139	

ii. For females: $\text{logit}(\pi) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) * math$

- A one-unit increase in math score yields a change in log odds of (0.13 + .067) = 0.197
- and thus, the odds ratio is $e^{0.197} = 1.22$
- The ratio of these two *math* odds ratios $\left(\frac{1.22}{1.14} = 1.07\right)$ is the exponentiated coefficient of the interaction term *female* $\times$ *math*: $e^{0.067} = 1.07$.

logistic regression in R

Logistic regression is performed using the `glm` function:

```
glm(data=dataframe, formula, family="binomial")
```

Confidence intervals for coefficients are available:

```
confint(glm object)
```

Exponentiated coefficients (odds ratios) are available using:

```
exp(coef(glm object))
```

Odds ratio confidence intervals:

```
exp(confint(glm object))
```

```
fit <- glm(data=honors, hon ~ math * female , family="binomial")
```

```
confint(fit)
```

	2.5%	97.5%
(Intercept)	-13.43535	-4.96760
math	0.06463	0.20750
female	-9.11491	3.23626
math:female	-0.03802	0.17506

exp(coef(fit))

(Intercept)	math	female	math:female
0.000159	1.138120	0.055031	1.069290

exp(confint(fit))

	2.5%	97.5%
(Intercept)	1.46253e-06	0.00696
math	1.06677e+00	1.23060
female	1.10013e-04	25.43828
math:female	9.62698e-01	1.19131

Outline

1. Logistic Regression Theory
2. Logistic Regression Example
3. Diagnostics

evaluating a logistic regression model

Two types of evaluation for the logistic regression model:

1. Diagnostics associated with residuals, influence, variance inflation, etc.
2. Diagnostics associated with overall fit and classification quality

notation

Cases: $i = 1, \ldots, n$

Target: $y_i \in \{0, 1\}$

Predictors: $X_i = (X_{1i}, \ldots, X_{pi})'$

True values: $\pi = P(Y = 1|X)$

Fitted values:

$$\hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p}}$$

residuals

Two main types of residuals for logistic regression:

- Pearson residuals
- Deviance residuals

Pearson residuals

Pearson residuals are based off the idea of z-score transformations (mean-centering and scaling by standard deviation)

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i (1 - \hat{\pi}_i)}}$$

Deviance residuals

Deviance residuals are related to the contribution of each point to the likelihood.

Recall: $\log L(\hat{\beta}) = \sum_{i=1}^{n} \left[y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)\right]$

The deviance residuals are defined as:

$$d_i = s_i \sqrt{-2\left[y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)\right]}$$

where $s_i = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = 0 \end{cases}$

deviance residuals

y_i	$\hat{\pi}_i$	$y_i \log \hat{\pi}$	$(1 - y_i) \log(1 - \hat{\pi}_i)$	d_i
0	0.12	0	-0.056	-0.333
1	0.12	-0.921	0	1.357
0	0.95	0	-1.301	-1.613
1	0.95	-0.022	0	0.211

new overall measures of fit

These two definitions of residuals can be squared and summed to create an RSS-like statistic

- With the deviance residuals, we produce the *deviance*:

$$D = \sum_{i=1}^{n} d_i^2$$

- With the Pearson residuals, we produce the *Pearson statistic*:

$$X^2 = \sum_{i=1}^{n} r_i^2$$

leverage

For logistic regression we can compute something very similar to the hat matrix **H** from OLS. This new matrix is called... the hat matrix.

The diagonal elements of **H** are again called *leverages* and are used to help standardize the residuals:

$$r_{\text{standard},i} = \frac{r_i}{\sqrt{1 - h_{ii}}}$$

$$d_{\text{standard},i} = \frac{d_i}{\sqrt{1 - h_{ii}}}$$

jackknifing

Recall from OLS, there are several measures related to evaluating the model fit without the i^{th} observation:

- studentized residuals
- dffits
- dfbetas

All of these are available (based on an approximation technique) for logistic regression and have the same interpretation.

influence and inflation

Fortunately, more measures from linear regression also apply:

- Cook's D
  - based on the modified hat matrix and residuals
  - same interpretation as from OLS
- Variance Inflation Factor (VIF)
  - based only on the input data $\mathbf{X}$
  - therefore it is exactly the same as in OLS

examples

See the code “LogRegHonorsEval.R” and the data file “honors.csv” in the course website for examples of logistic regression diagnostics w.r.t residuals, influence, etc.

In the code we will also look at assessing the overall fit and classification quality using measures we have seen before such as accuracy, kappa, sensitivity, specificity, ROC, AUC, concordance; but we will also look at few new ones too!