Lesson 4 Logistic Regression

Outline

- Why use a Logistic Regression Model?
- Development of Logistic Regression Equation
- Interpreting coefficients
- Estimation by maximum likelihood
- Evaluating the performance of the model

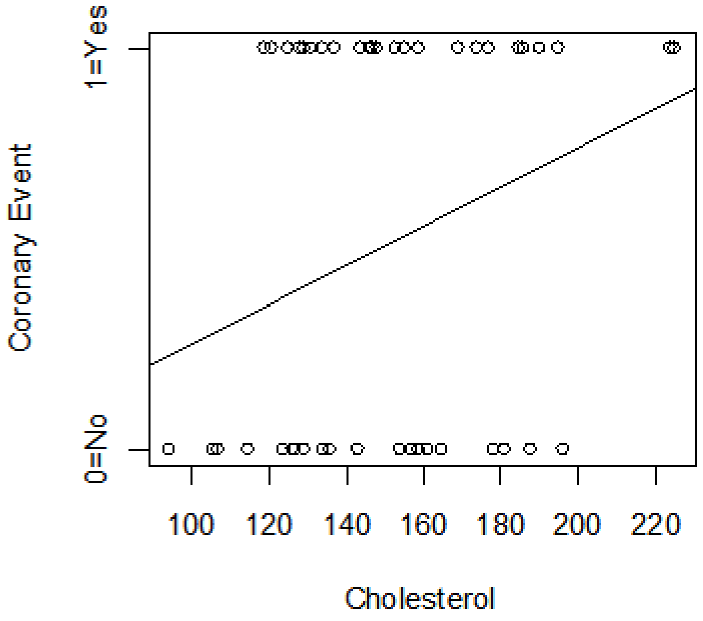
Why Use a Logistic Regression Model?

- We have seen that a linear regression model assumes that a response variable y (like "exam score") is related to an explanatory variable x (like "number of hours spent studying") in a linear fashion: $y=\beta_0+\beta_1x$. In other words, a set of data points, representing y values based on inputs x, are approximated reasonably well with a straight line. When the data involved satisfies certain conditions (absence of outliers and, roughly, that scatterplots "appear" to be approximable with a line), the use of a linear regression model is justified and predicts values reasonably well.
- An important setting in which linear regression is not appropriate arises when the outcome (y) of interest takes on just one of two possible values. Example on next slide provides a typical instance.

Why Use a Logistic Regression Model?

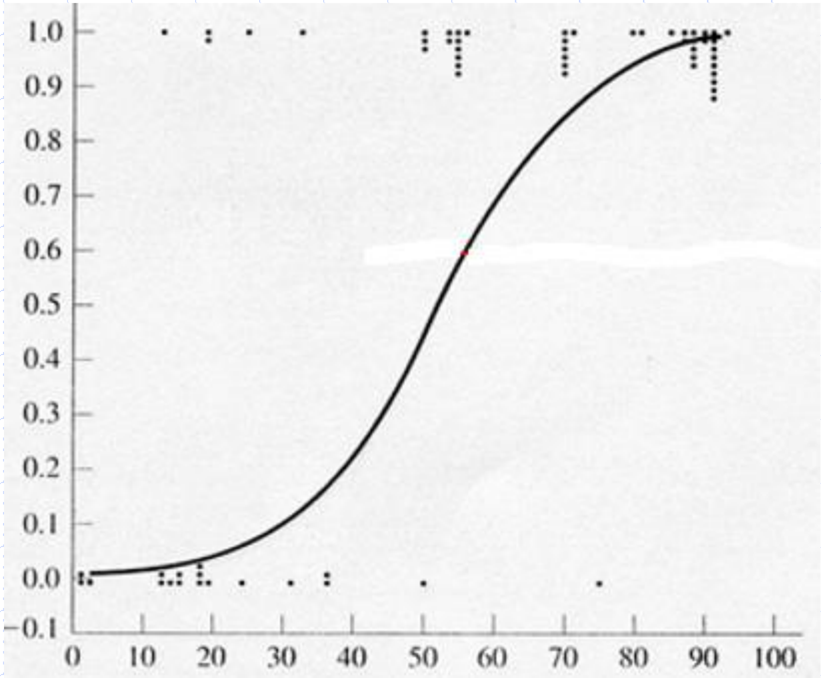
Example: Suppose we are interested in the association between cholesterol levels and having a coronary event in a high risk patient population (who have had an event in the past). We collect cholesterol data for 50 subjects and then follow each for a year to see if they have another coronary event.

Issues: As displayed in the plot on the right, the fit of the regression is not ideal. Many of the assumptions required for regression are not met in this setting (for example, data points “appear" to be approximable with a line).



The Logistic Regression Approach

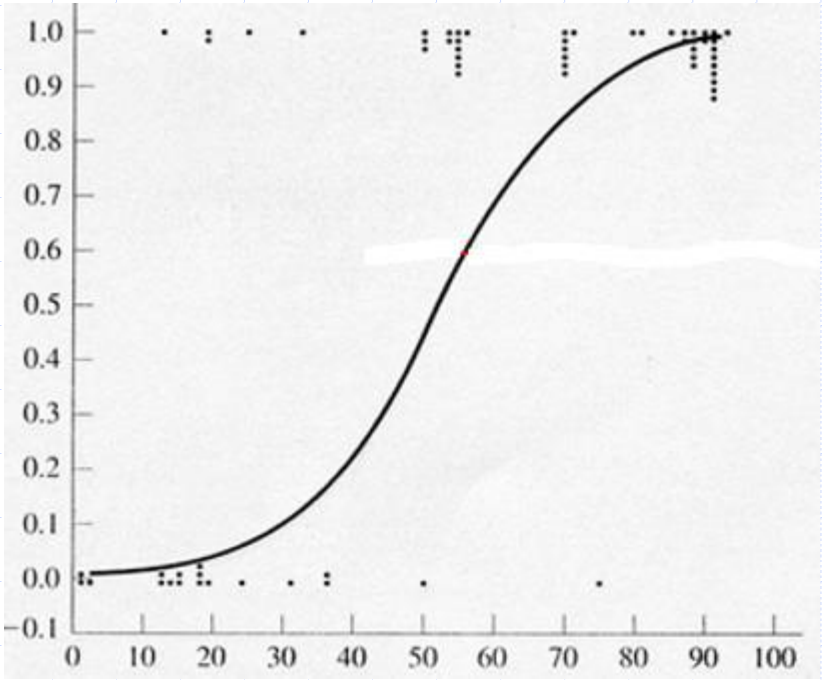
Instead of approximating data points with a line, logistic regression approximates with an "S"-curve, called a *logistic function.*



- The two possible outcomes can be viewed as values 0 or 1
  0 = no coronary event
  1 = coronary event

- The S-curve approximates well values of 0 when x is small and values of 1 when x is larger.

The Logistic Regression Approach



- The values 0 or 1 can be viewed as *probabilities* of a coronary event occurring. (1 = certain, 0 = doesn't occur)

- In logistic regression, this S curve represents the growth of probability as x increases.

- The S curve is called a *logistic function* (a special type of "sigmoid function") and has the following mathematical form:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Logistic Regression Formula

The analogue to the linear regression formula in the present context is the following, where p is the probability that the outcome (like coronary event) occurs.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The model assumes that the probability has the form of the logistic function, taking a linear argument depending on two parameters. Using two parameters $\beta_{0,}$ β_1 gives more flexibility to fit the data (instead of simply $1/(1+e^{-x})$), and linearity ($\beta_0+\beta_1 x$) makes computations as simple as possible.

This particular formula for *p* can be derived from the assumption that the *log of the odds is linear* (next slides)

Odds

- Given some event with probability *p*, the odds of that event are given by:

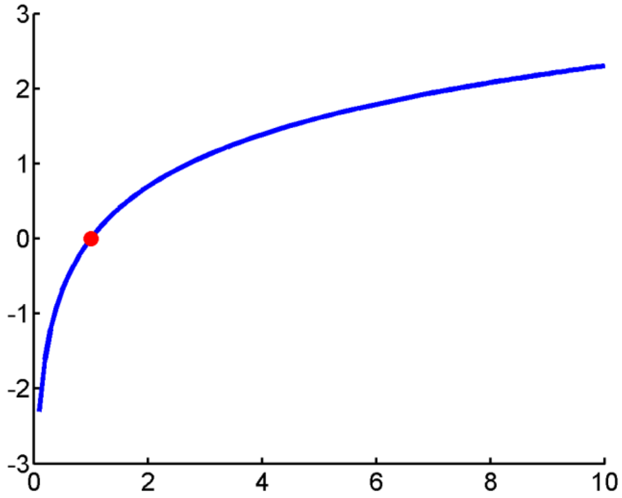
$$\boxed{\text{odds} = p\,/\,(1-p)}$$

- Example: If you roll a die, what are the odds of getting a composite number? (composite numbers are the product of two smaller numbers both > 1)
  - Solution: odds = (2/6) / (1-2/6) = 2/4

    (odds = (# ways to succeed) / (# ways to fail))

Logit Transform

- The logit is the natural log of the odds



- logit(p) = ln(odds) = ln (p/(1-p))

Logistic Regression

- In logistic regression, we seek a model:

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

- That is, the log odds (logit) is assumed to be linearly related to the independent variable x

The Logistic Regression Model

The "logit" model solves these problems:

$$\ln[p/(1-p)] = \beta_0 + \beta_1 x$$

- p is the probability of a “success”
- p/(1-p) are the odds of the event occurring
- ln[p/(1-p)] is the log odds of the event, or "logit"
- x is the explanatory variable
- β_0 is the intercept
- β_1 is the regression coefficient

Deriving the Formula for *p*

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

$$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

which gives *p* as described before, as a logistic function of x

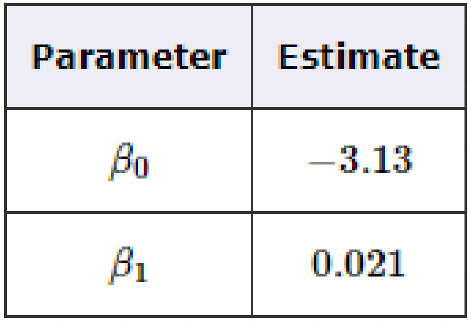
Logistic Response Function

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
  - If you let $\beta_0 + \beta_1 x = 0$, then $p = .50$
  - As $\beta_0 + \beta_1 x$ gets really big, p approaches 1
  - As $\beta_0 + \beta_1 x$ gets really small, p approaches 0

Rework the Example using Logistic Regression

- Explanatory variable: cholesterol level
- Response variable: whether or not the subject had another coronary event.
- Perform Logistic Regression

```
m<-glm(data$event ~ data$chol,
             family = binomial)
summary(m)
```



Parameter	Estimate
β_0	-3.13
β_1	0.021

chol	Event
106	0
147	1
131	1
125	1
107	0
155	1
174	1
148	1
157	0
153	1
195	1
177	1
143	0
94	0
119	1
165	0
…	…

Using the Model for Predictions

- Using the formula above, the risk of having a coronary event for a patient with a cholesterol level of 190 is predicted by the regression to be 70.3%

$$
\begin{aligned}
\hat{p} &= 1/(1 + e^{-\hat{\beta}_0 - \hat{\beta}_1 x}) \\
&= 1/(1+ e^{3.13-0.021*190}) \\
&= 0.703
\end{aligned}
$$

Interpretation of $\hat{\beta}_1$ and odds ratio

- The interpretation of the regression coefficient(s) of logistic regression are generally based on odds ratios. Consider the odds ratio of an event for a given value of x=x_1 versus a given value of x=x_2
- Let:
  - odds1 = The estimated odds of x_1 = $e^{\hat{\beta}_0 + \hat{\beta}_1 x_1}$
  - odds2 = The estimated odds of x_2 = $e^{\hat{\beta}_0 + \hat{\beta}_1 x_2}$
- Odds ratio is given by:

$$\frac{odd1}{odd2} = e^{\hat{\beta}_1 (x_1 - x_2)}$$

Interpretation of $\hat{\beta}_1$ and odds ratio

- Odds ratio:

$$\frac{odd1}{odd2} = e^{\hat{\beta}_1(x_1 - x_2)}$$

- The resulting quantity can be interpreted as follows: The odds of the event are $e^{\hat{\beta}_1(x_1 - x_2)}$ higher for every $x_1 - x_2$ unit increase in x.
- In particular, if $x_1 - x_2 = 1$, $e^{\hat{\beta}_1(x_1 - x_2)} = e^{\hat{\beta}_1}$ which can be interpreted as the relative increase in odds for every 1 unit increase in x.

Maximum Likelihood Estimation

- MLE is a statistical method for estimating the coefficients for the logistic model.
- The likelihood function (L) is the product of computed probabilities q_i of the data points occur in the sample:

  $L = q_1 * q_2 * \dots * q_n$

$$q_i = \begin{cases} p_i, & if\ ith\ data\ point\ indicates\ coronary\ event \\ 1 - pi, & if\ data\ point\ does\ not\ indicate\ this \end{cases}$$

In our earlier example(cholesterol level v.s. coronary event), q_i is the probability that ith data point has coronary event if the true outcome is that it has coronary event; (similarly, q_i is the probability that ith data point does not have coronary event if the true outcome is it does not have.)

Maximum Likelihood Estimation (MLE)

- MLE involves finding the coefficients that makes the likelihood function as large as possible. (Statisticians like to maximize log of the likelihood. Either way works.)

$$L = q_1 * q_2 * \ldots * q_n$$

$$\text{Log } L = \text{Log } (q_1 * q_2 * \ldots * q_n)$$

$$= \log q_1 + \log q_2 + \ldots \log q_n$$

How good is the model?

- Software helps to compute the best fit line, but how do we know how good the model is?
- In other words, how do we calculate R^2?
  - Answer: No consensus on how to calculate it for Logistic Regression. There are more than 10 different ways to do it!

McFadden's pseudo-R squared

- Here is the definition:

$$R^2_{\text{McFadden}} = 1 - \frac{log(L_c)}{log(L_{\text{null}})}$$

where L_c denotes the (maximized) likelihood value from the current fitted model, and L_{null} denotes the likelihood value from the bad fit model where it does not take the predictor into account.

- L_c is a measure of a good fit and L_{null} is a measure of a bad fit.

How to compute $\log(L_{null})$?

Solution:

Let a = number of people having coronary event

b = number of people who do not have coronary event

$p = a/(a+b)$. (p is the probability of sample data having coronary event.)

$\log(L_{null})$ = the log-likelihood of the data given the overall probability p

$= a * \log p + b * \log(1-p)$

Demo: compute McFadden's pseudo-R squared in R.

```
nullm <- glm(data$event ~ 1, family="binomial")
rsquare = 1-logLik(m)/logLik(nullm)
rsquare
```

McFadden's pseudo-R squared-Interesting result

- In most empirical research typically one could not hope to find predictors which are strong enough to give predicted probabilities so close to 0 or 1, and so one shouldn't be surprised if one obtains a value of R^2_{McFadden} which is not very large.
- Demo in R: even when the predictor is strong, (e.g. x as a predictor correctly predicts y values 90% of the time) McFadden's pseudo-R squared is not very large.
- Interesting article about R squared in logistic regression: http://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/

Multiple Logistic Regression

The equation:

$$\ln[p/(1-p)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- p is the probability of a “success”
- p/(1-p) are the odds of the event
- ln[p/(1-p)] is the log odds of the event, or "logit"
- $x_1, x_2, \ldots, x_k$ are the explanatory or independent variables
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients

Demo: Multiple Logistic Regression in R.