Univariate Logistic Regression

Basic Ideas

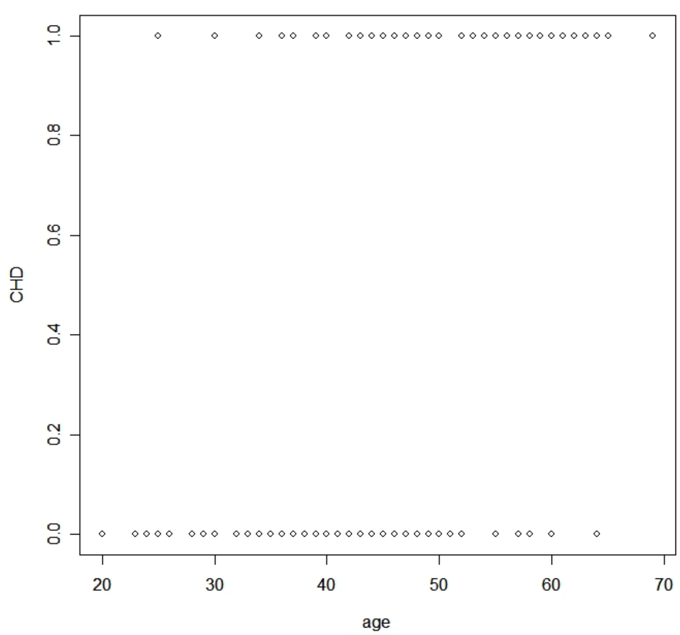
Motivation by example: Suppose we wish to examine the relationship between age and coronary heart disease (CHD). Some data relating CHD and age are (taken from Chapter 1 of Hosmer book):

Age	CHD	Age	CHD	Age	CHD	Age	CHD
20	0	35	0	44	1	55	1
23	0	35	0	44	1	56	1
24	0	36	0	45	0	56	1
25	0	36	1	45	1	56	1
25	1	36	0	46	0	57	0
26	0	37	0	46	1	57	0
26	0	37	1	47	0	57	1
28	0	37	0	47	0	57	1
28	0	38	0	47	1	57	1
29	0	38	0	48	0	57	1
30	0	39	0	48	1	58	0
30	0	39	1	48	1	58	1
30	0	40	0	49	0	58	1
30	0	40	1	49	0	59	1
30	0	41	0	49	1	59	1
30	1	41	0	50	0	60	0
32	0	42	0	50	1	60	1
32	0	42	0	51	0	61	1
33	0	42	0	52	0	62	1
33	0	42	1	52	1	62	1
34	0	43	0	53	1	63	1
34	0	43	0	53	1	64	0
34	1	43	1	54	1	64	1
34	0	44	0	55	0	65	1
34	0	44	0	55	1	69	1

While age is a continuous variable, CHD is not, so that the linear regression methods we have used so far are not appropriate.

To see why linear regression is not appropriate, let's examine a scatter plot.

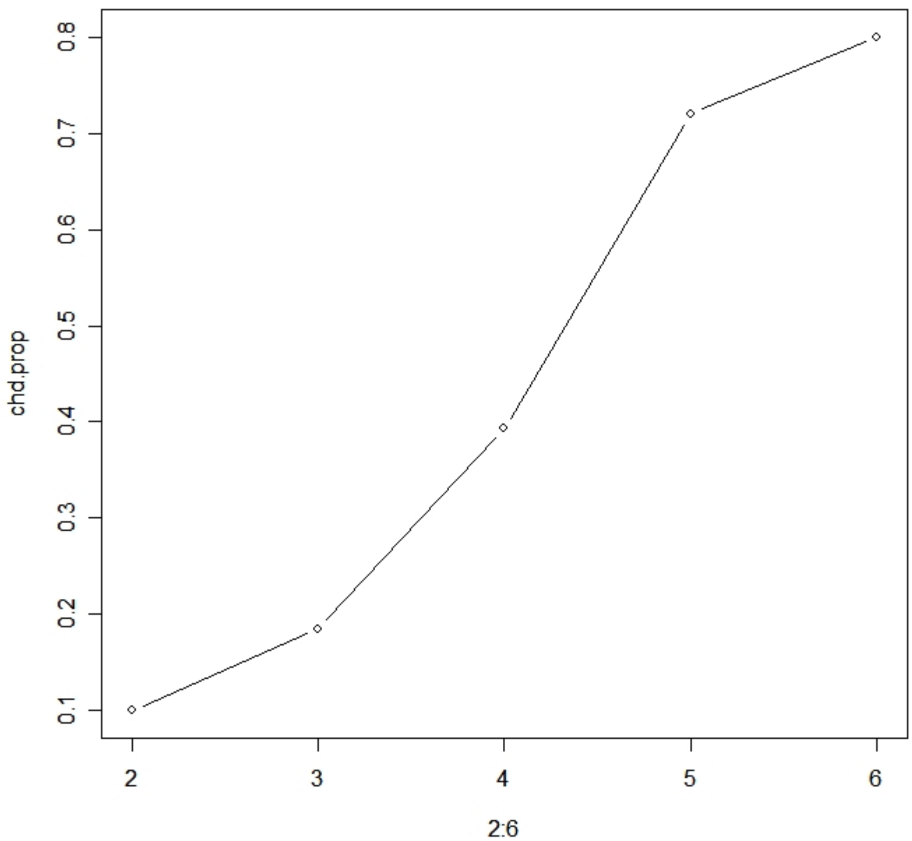
```
# Enter the data

> age <- c( 20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30,
30, 30, 30, 32, 32, 33, 33, 34, 34, 34, 34, 34, 35, 35, 36, 36, 36,
37, 37, 37, 38, 38, 39, 39, 40, 40, 41, 41, 42, 42, 42, 42, 43, 43,
43, 44, 44, 44, 44, 45, 45, 46, 46, 47, 47, 47, 48, 48, 48, 49, 49,
49, 50, 50, 51, 52, 52, 53, 53, 54, 55, 55, 55, 56, 56, 56, 57, 57,
57, 57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 62, 62, 63, 64, 64,
65, 69)

> CHD <- c( 0 , 0 , 0 , 0 , 1 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0,
1 , 0 , 0 , 0 , 0 , 0 , 0 , 1 , 0 , 0 , 0 , 0 , 0 , 1 , 0 , 0 , 1,
0 , 0  , 0 , 0 , 1 , 0 , 1 , 0 , 0 , 0 , 0 , 0 , 1 , 0 , 0 , 1 , 0,
0 , 1 , 1  , 0 , 1 , 0 , 1 , 0 , 0 , 1 , 0 , 1 , 1 , 0 , 0 , 1 , 0,
1 , 0 , 0 , 1  , 1 , 1 , 1 , 0 , 1 , 1 , 1 , 1 , 1 , 0 , 0 , 1 , 1,
1 , 1 , 0 , 1 , 1  , 1 , 1 , 0 , 1 , 1 , 1 , 1 , 1 , 0 , 1 , 1 , 1)

# Scatter plot

> plot(age, CHD)
```



While we can see some patterns using this scatter plot (for example, notice that there are increasingly more points on top compared to on bottom as age increases), it is far from optimal.

One way around this may be to group age by decades, say, and look at CHD rates within these decades.

```
#  Prepare age decade data, count how many we have in each decade:

#  Create a blank variable to be filled in later

> age.decade <- rep(NA, 5)

> for (i in 1:5) { age.decade[i]
           <- length(age[age > ( 10*(i+1) -1) & age < ( 10*(i+2))])}

> age.decade
[1] 10 27 28 25 10

#  Calculate the corresponding percentages from the CHD variable:

#  Create a blank variable to be filled in later

> chd.prop <- rep(NA, 5)

#  Create an index to sum over, based on sums in age.decade

> index.age <- c(0, 10, 37, 65, 90, 100)

> for (i in 1:5) { chd.prop[i]
        <- sum(CHD[(index.age[i]+1):index.age[i+1]])/age.decade[i]  }

> chd.prop
[1] 0.1000000 0.1851852 0.3928571 0.7200000 0.8000000

# Create a scatter plot between age as an decade and chd.prop

> plot(2:6, chd.prop, type="b")
```



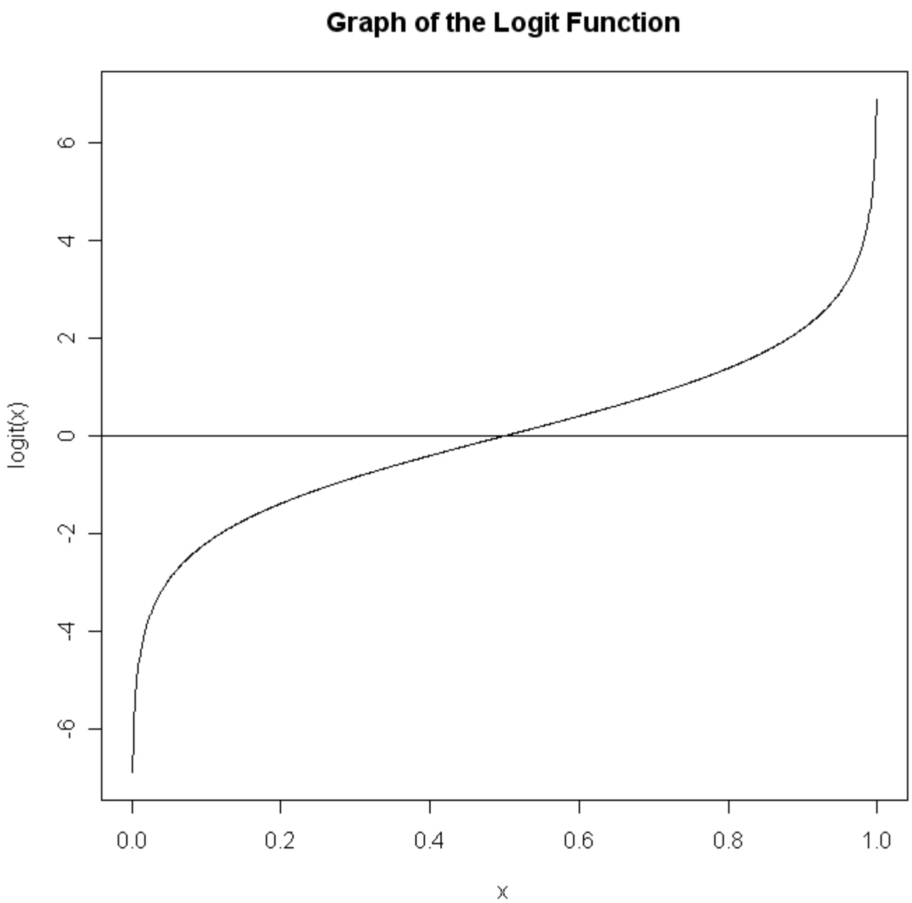
This plot is more usable than the first scatter plot, but is wasteful of information, as detailed ages are lost. Still it indicates a general trend that CHD rates increase with age, and is a useful type of plot for descriptive purposes when beginning to model.

Looking at the plot may also remind us of the shape of the inverse logit function.

Recall that the logit function is defined by

$$f(x) = logit(x) = \log\left(\frac{x}{1-x}\right)$$

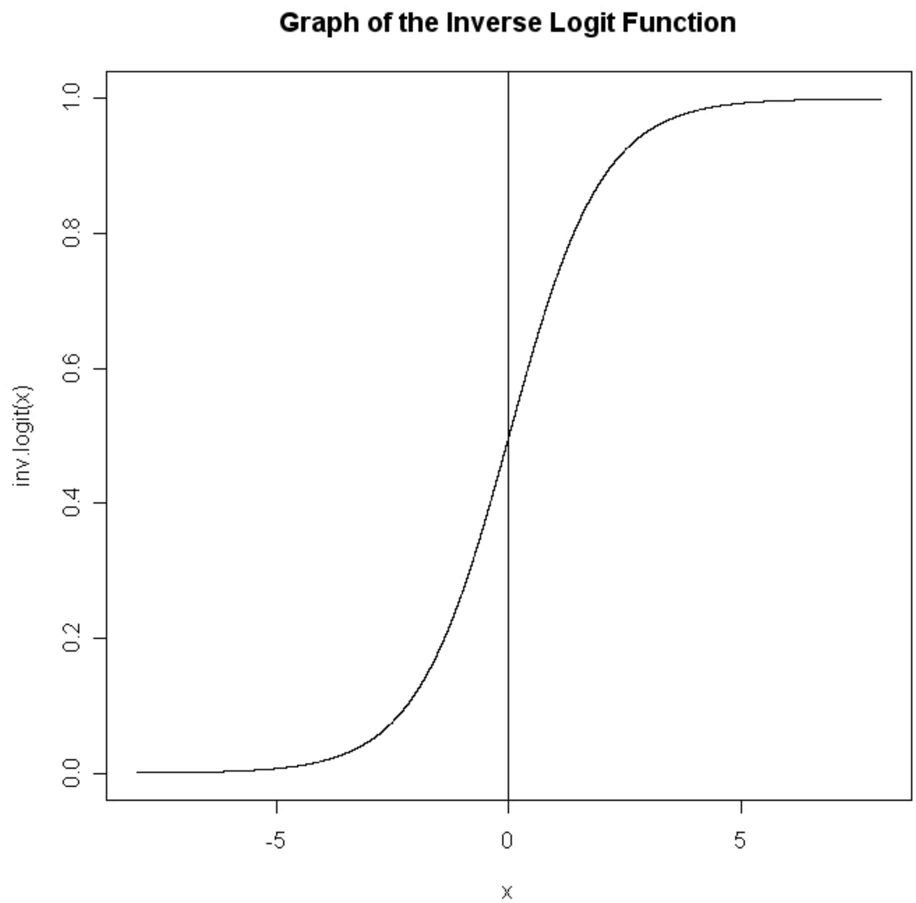
with graph



Also recall that the inverse logit function is given by

$$f(x) = inv.logit(x) = \frac{\exp(x)}{1 + \exp(x)}$$

with graph



It therefore looks reasonable to use logistic regression to model the effect of age on CHD rates. In general, logistic regression is a "first-line" model for dichotomous outcome data, just as linear regression is used for continuous outcomes or Poisson regression for count outcomes. Other options not discussed in this course includes probit models.

To use the logistic model, we need to decide what "x" needs to be in the equations for the logit and inv.logit functions.

The inverse.logit function will give us the probabilities of events (e.g. CHD) we need, while the logit function will give us the linear function that relates outcomes to the covariates. In general, the equations are:

Let $\pi(x)$ represent the probability of an event (e.g. the dependent variable, CHD) that depends on a covariate (e.g. independent variable, age). Then, using an inv.logit formulation for modeling the probability, we have:

$$\pi(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

To obtain the logit function from this, we calculate:

$$\begin{aligned}
\text{logit}[\pi(x)] &= \ln\left[\frac{\pi(X)}{1-\pi(X)}\right] \\
&= \ln\left[\frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}\right] \\
&= \ln\left[\frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{1}{1+e^{\beta_0+\beta_1 X}}}\right] \\
&= \ln\left[e^{\beta_0+\beta_1 X}\right] \\
&= \beta_0+\beta_1 X
\end{aligned}$$

To summarize, the two basic equations of logistic regression are:

$$\pi(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

which gives the probabilities of outcome events given the covariate value X, and

$$\text{logit}[\pi(X)] = \beta_0+\beta_1 X$$

which shows that we are really dealing with a standard linear regression model, once we transform the dichotomous outcome by the logit transform. This transform changes the range of $\pi(x)$ from 0 to 1 to $-\infty$ to $+\infty$, as usual for linear regression.

Similar to linear regression, the above equation represents the mean or expected probability, $\pi(X)$, given X. As this is an average, we expect an error. Again analogously to linear regression, we have an error distribution, but rather than a normal distribution, we use a binomial distribution, to match the dichotomous outcomes . The mean of the binomial distribution is $\pi(X)$, and the variance is $\pi(X)(1-\pi(X))$ (recall the properties of the binomial distribution).

Interpretation of the coefficients β_0 and β_1 in logistic regression

Interpretation of the intercept, β_0: If $X = 0$, then we have

$$\pi(x) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

Therefore, β_0 sets the event rate, through the above function, when the covariate value is equal to zero.

For example, if $\beta_0 = 0$, then

$$\pi(x) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} = \frac{e^0}{1 + e^0} = \frac{1}{1 + 1} = 0.5$$

So, positive values of β_0 give "probability intercepts" greater than 0.5, while negative values of β_0 give "probability intercepts" less than 0.5

Interpretation of the slope, β_1: Consider the effect on the probability of an event as X changes by one unit. Suppose in particular that X changes from X_0 to $X_0 + 1$.

When $X = X_0$, we have:

$$\text{logit}[\pi(X_0)] = \beta_0 + \beta_1 X_0$$

On the other hand, when $X = X_0 + 1$, we have:

$$\text{logit}[\pi(X_0 + 1)] = \beta_0 + \beta_1 (X_0 + 1)$$

Subtracting the above two terms, we have:

$$\text{logit}[\pi(X_0 + 1)] - \text{logit}[\pi(X_0)] = \beta_0 + \beta_1(X_0 + 1) - \beta_0 + \beta_1(X_0) = \beta_1$$

From the definition of the logit function, we have:

$$\begin{aligned}
\text{logit}[\pi(X_0 + 1)] - \text{logit}[\pi(X_0)] &= \beta_1 \\
\log[\frac{\pi(X_0 + 1)}{1 - \pi(X_0 + 1)}] - \log[\frac{\pi(X_0)}{1 - \pi(X_0)}] &= \beta_1 \\
\log\left[\frac{\frac{\pi(X_0+1)}{1-\pi(X_0+1)}}{\frac{\pi(X_0)}{1-\pi(X_0)}}\right] &= \beta_1 \\
\log[OR] &= \beta_1
\end{aligned}$$

The steps above follow from the definition of the logit function and the definition of an odds ratio. The term OR represents the odds ratio for a change of one unit in the independent X variable.

Taking the exponential of both sides of the equation, we get:

$$\exp(\log[OR]) = exp(\beta_1)$$

which implies

$$OR = exp(\beta_1) = e^{\beta_1}$$

Basic result:

The coefficient β_1 is such that e^{β_1} is the odds ratio for a unit change in X.

If we change X by two units, then the OR for a two unit change is $e^{2\beta_1} = \left(e^{\beta_1}\right)^2$, and so on. In general, for a change of z units, the $OR = e^{z\beta_1} = \left(e^{\beta_1}\right)^z$.

Estimating β_0 and β_1 given a data set

As discussed above, the distribution associated with logistic regression is the binomial. For a single subject with covariate value x_i, the likelihood function is:

$$\pi(x_i)^{y^i}(1 - \pi(x_i))^{1-y^i}$$

For n subjects, the likelihood function is:

$$\prod_{i=1}^{n} \pi(x_i)^{y^i}(1 - \pi(x_i))^{1-y^i}$$

To derive estimates of the unknown parameters β_0 and β_1, we need to maximize this likelihood function. We follow the usual steps, including taking the logarithm of the likelihood function, taking partial derivatives with respect to β_0 and β_1, and setting these two equations equal to zero, to form a set of two equations in two unknowns. Solving this system of equations gives the maximum likelihood equations.

We omit the details here (no easy closed form formulae), and will rely on statistical software to find the maximum likelihood estimates for us.

Inferences typically rely on SE formulae for confidence intervals, and likelihood ratio testing for hypothesis tests. Again, we will omit the details, and rely on statistical software.

Example: The effect of age of CHD event rates

Let's see how we can draw inferences about logistic regression parameters using R:

```
> output <- glm(CHD ~ age, family=binomial)
> summary(output)

Call:
glm(formula = CHD ~ age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
age          0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4
```

Once again, the standard R glm function does not provide confidence intervals by default, so we will create our own function:

```
logistic.regression.with.ci <- function(regress.out, level=0.95)
{
#################################################################
#                                                               #
#  This function takes the output from a glm                    #
#  (logistic model) command in R and provides not               #
#  only the usual output from the summary command, but          #
#  adds confidence intervals for all coefficients.              #
#                                                               #
#  This version accommodates multiple regression parameters     #
#                                                               #
#################################################################
usual.output <- summary(regress.out)
z.quantile <- qnorm(1-(1-level)/2)
number.vars <- length(regress.out$coefficients)
temp.store.result <- matrix(rep(NA, number.vars*2), nrow=number.vars)
for(i in 1:number.vars)
{
     temp.store.result[i,] <- summary(regress.out)$coefficients[i] +
      c(-1, 1) * z.quantile * summary(regress.out)$coefficients[i+number.vars]
}
  intercept.ci <- temp.store.result[1,]
  slopes.ci <- temp.store.result[-1,]
  output <- list(regression.table = usual.output, intercept.ci = intercept.ci,
              slopes.ci = slopes.ci)
return(output)
}

#  Test out the function on our output:

> logistic.regression.with.ci(output)
$regression.table

Call:
glm(formula = CHD ~ age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
age          0.11092    0.02406   4.610 4.02e-06 ***
---
```

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4


$intercept.ci
[1] -7.531374 -3.087533

$slopes.ci
[1] 0.06376477 0.15807752
```

But this is not really quite enough, because we are usually interested not only in the coefficients, but also the odds ratios. So, we add an extra line to the function:

```
logistic.regression.or.ci <- function(regress.out, level=0.95)
{
################################################################
#                                                              #
#  This function takes the output from a glm                   #
#  (logistic model) command in R and provides not              #
#  only the usual output from the summary command, but         #
#  adds confidence intervals for all coefficients and OR's.    #
#                                                              #
#  This version accommodates multiple regression parameters    #
#                                                              #
################################################################
usual.output <- summary(regress.out)
z.quantile <- qnorm(1-(1-level)/2)
number.vars <- length(regress.out$coefficients)
OR <- exp(regress.out$coefficients[-1])
temp.store.result <- matrix(rep(NA, number.vars*2), nrow=number.vars)
for(i in 1:number.vars)
{
      temp.store.result[i,] <- summary(regress.out)$coefficients[i] +
      c(-1, 1) * z.quantile * summary(regress.out)$coefficients[i+number.vars]
}
  intercept.ci <- temp.store.result[1,]
  slopes.ci <- temp.store.result[-1,]
```

```
  OR.ci <- exp(slopes.ci)
  output <- list(regression.table = usual.output, intercept.ci = intercept.ci,
              slopes.ci = slopes.ci, OR=OR, OR.ci = OR.ci)
return(output)
}

#  Run the function for our data:

>  logistic.regression.or.ci(output)
$regression.table

Call:
glm(formula = CHD ~ age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
age          0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4


$intercept.ci
[1] -7.531374 -3.087533

$slopes.ci
[1] 0.06376477 0.15807752

$OR
     age
1.117307

$OR.ci
```

```
[1] 1.065842 1.171257
```

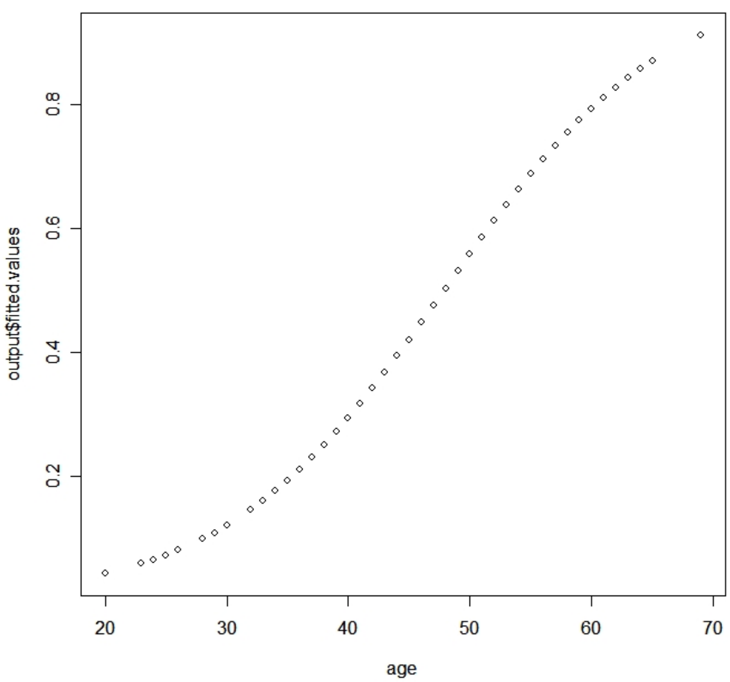
So, for each change of one year in age, there is an odds ratio of 1.117, with 95% CI (1.066, 1.171). So, for a ten year change in age, for example, we raise each of these values to the power of ten, getting an OR per 10 year change of $1.117307^{10} = 3.03$, with 95% CI of (1.89, 4.86). This is clearly a very clinically important effect.

Predictions from logistic regression models

As with linear regression, once we fit a logistic regression model, we can make predictions using the fitted equation. To get point estimates, we simply need to plug the relevant X values into the inv.logit equation, but again we will rely on R:

```
# First, let's check what types of outputs are available once
# we have run a logistic regression (which recall we saved
# in the object "output"):

> names(output)
 [1] "coefficients"      "residuals"         "fitted.values"     "effects"
 [5] "R"                 "rank"              "qr"                "family"
 [9] "linear.predictors" "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"     "df.residual"
[17] "df.null"           "y"                 "converged"         "boundary"
[21] "model"             "call"              "formula"           "terms"
[25] "data"              "offset"            "control"           "method"
[29] "contrasts"         "xlevels"

#  See R help on GLM to define all of these, we will see just one here:

#  Make predictions for each subject in the data set:

> output$fitted.values
         1          2          3          4          5          6          7          8
0.04347876 0.05962145 0.06615278 0.07334379 0.07334379 0.08124847 0.08124847 0.09942218
         9         10         11         12         13         14         15         16
0.09942218 0.10980444 0.12112505 0.12112505 0.12112505 0.12112505 0.12112505 0.12112505
        17         18         19         20         21         22         23         24
0.14679324 0.14679324 0.16123662 0.16123662 0.17680662 0.17680662 0.17680662 0.17680662
        25         26         27         28         29         30         31         32
0.17680662 0.19353324 0.19353324 0.21143583 0.21143583 0.21143583 0.23052110 0.23052110
        33         34         35         36         37         38         39         40
```

```
0.23052110 0.25078125 0.25078125 0.27219215 0.27219215 0.29471199 0.29471199 0.31828021
        41         42         43         44         45         46         47         48
0.31828021 0.34281708 0.34281708 0.34281708 0.34281708 0.36822381 0.36822381 0.36822381
        49         50         51         52         53         54         55         56
0.39438351 0.39438351 0.39438351 0.39438351 0.42116276 0.42116276 0.44841400 0.44841400
        57         58         59         60         61         62         63         64
0.47597858 0.47597858 0.47597858 0.50369030 0.50369030 0.50369030 0.53137935 0.53137935
        65         66         67         68         69         70         71         72
0.53137935 0.55887652 0.55887652 0.58601724 0.61264546 0.61264546 0.63861714 0.63861714
        73         74         75         76         77         78         79         80
0.66380304 0.68809096 0.68809096 0.68809096 0.71138714 0.71138714 0.71138714 0.73361695
        81         82         83         84         85         86         87         88
0.73361695 0.73361695 0.73361695 0.73361695 0.73361695 0.75472490 0.75472490 0.75472490
        89         90         91         92         93         94         95         96
0.77467399 0.77467399 0.79344462 0.79344462 0.81103299 0.82744940 0.82744940 0.84271622
        97         98         99        100
0.85686593 0.85686593 0.86993915 0.91246455

#  So plot age versus fitted value for age:

> plot(age, output$fitted.values)
```



Next, we will extend these methods to more than one independent variable.

A Note On Study Designs

Random Sampling: So far, we have assumed our data have arisen from a simple random sample. In this case, logistic regression models and their inferences follow immediately.

Cohort Studies: Cohort studies typically select subjects to follow at random, and so are an example of random sampling. We select subjects (and hence their covariates, like age) at the start of the study, and follow them to see if they eventually have an event (like CHD). So, as in a random sample, the logistic regression model and its inferences follow immediately.

Case-Control Studies: Here we select cases and controls first (for example, find subjects with and without CHD), and the "randomness" is not in the eventual outcome (CHD), but in what their covariates are (e.g., age). So, this design is "backwards" from a standard cohort study. Nevertheless, it can be shown that standard logistic regression models and the usual inferences can be used, the theory involving two consecutive applications of Bayes' Theorem. See Chapter 7 of Hosmer and Lemeshow for full details

Matched Studies: Logistic regression can once again be used, but with no intercept, and with the data being manipulated such that each matched pair becomes a single data point. The matched pairs are only useful if one of the subjects has an event, and the other does not. See Chapter 7 of Hosmer and Lemeshow for full details (this material is beyond the scope of this introductory course). In short, logistic regression can be used here as well, but for data that are manipulated, and with no intercept in the model.