BMI 541/699 Lecture 22

Where we are:

1. Introduction and Experimental Design
2. Exploratory Data Analysis
3. Probability
4. T-based methods for continous variables
5. Power and sample size for t-based methods
6. Proportions & contingency tables
7. Non-parametric hypothesis tests for one and two samples
8. Testing hypotheses about more than 2 mean(s)
9. Assessing relationships between 2 continuous variables.
10. **Regression when the response variable is binary**
    - **Logistic regression**

Logistic Regression

A Statistical Model for Bernoulli Response Variables

If we have a continuous predictor variable x and a continuous response variable y we can use linear regression to estimate the best line for predicting y from given a value of x.

- The model is
$$\hat{y} = \alpha + \beta x$$
- a range of values of x and the corresponding predicted values of y define a straight line.

What if we want to predict the probability of a success for a Bernoulli variable for a particular value of a predictor variable x?

We use logistic regression.

Example: A study of (in hospital) mortality in ICU patients

Data on 200 ICU patients

The variables are

- age: age of patient
- admit.type: type of admission (emergency or elective)
- death: in hospital mortality (yes or no)

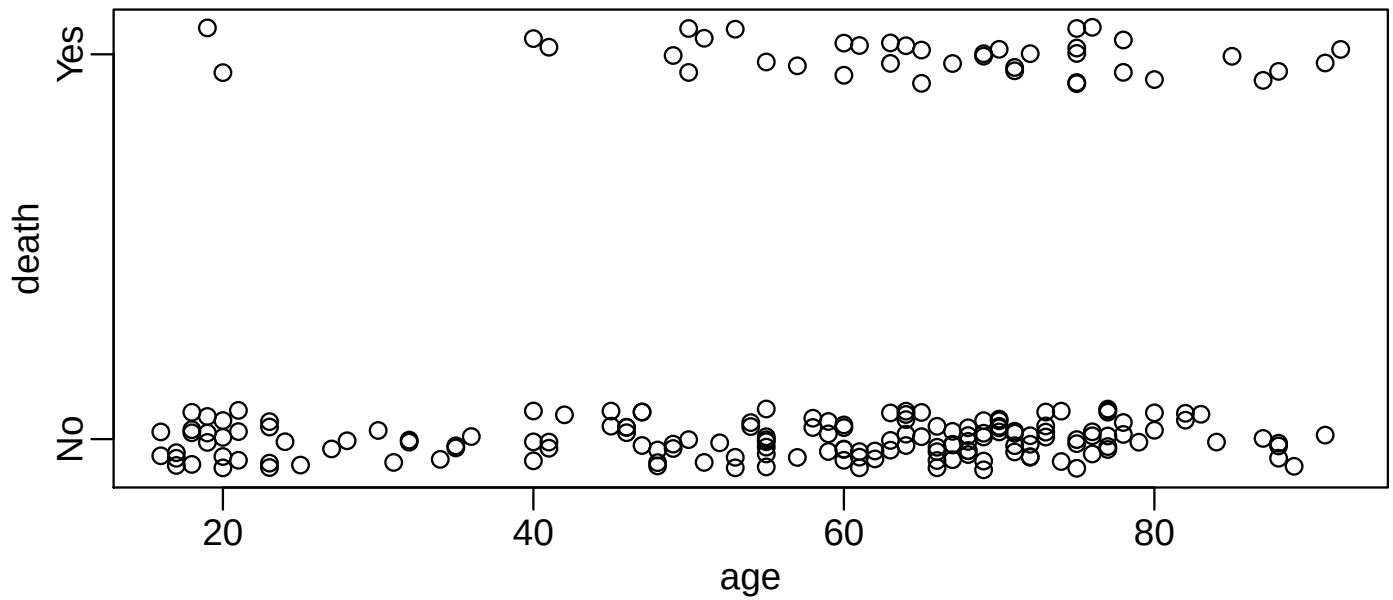
```
> icu
    death age admit.type
1      no  27  emergency
2      no  59  emergency
3      no  77   elective
4      no  54  emergency
5      no  87  emergency
⋮

197   yes  64  emergency
198   yes  60  emergency
199   yes  60  emergency
200   yes  50  emergency
>
```

We would like to use age and type of admission to predict mortality.

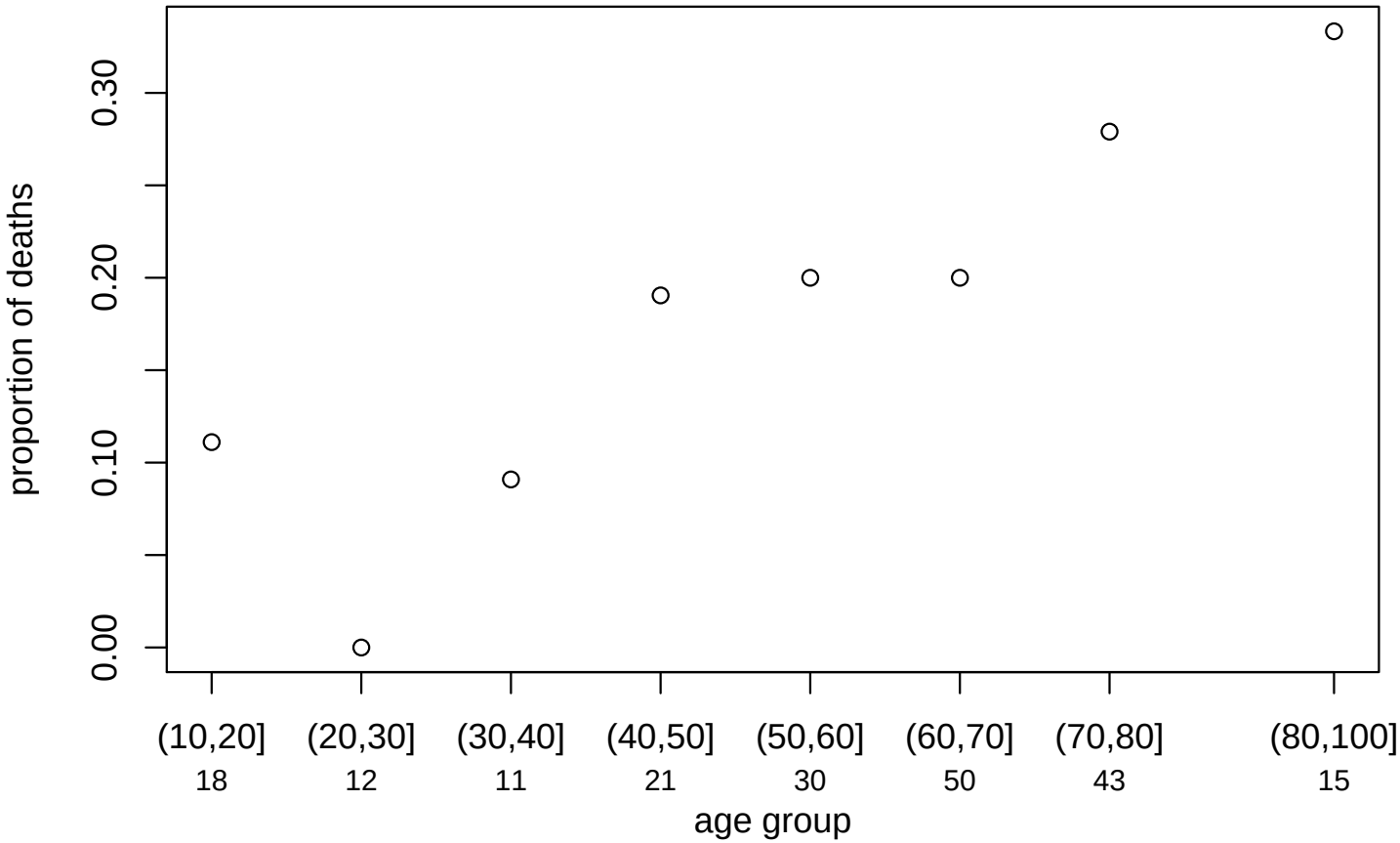
First, let's consider predicting the Bernoulli variable death using the continuous variable age.

We can plot the data:



It looks like there might be a higher proportion of yeses for higher ages.

If we bin the age variable we can plot the observed proportion of deaths in each age group:



(The numbers under the x axis labels are the number of patients in each age group.)

It looks like there might be a linear relationship between age and probability of death.

We might try to fit a model like:

$$\Pr(\text{death}) = \alpha + \beta\,(\text{age})$$

Unfortunately this doesn't usually work because

- the right hand side of the equation can take on any value (depending on the values of α, β, and age).
- the left hand side of the equation is constrained to be between 0 and 1

This will often lead to predicted probabilities outside the interval (0,1)

To fix this we don't use probability p as our response variable but instead we use

The logit or log odds of the probability:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

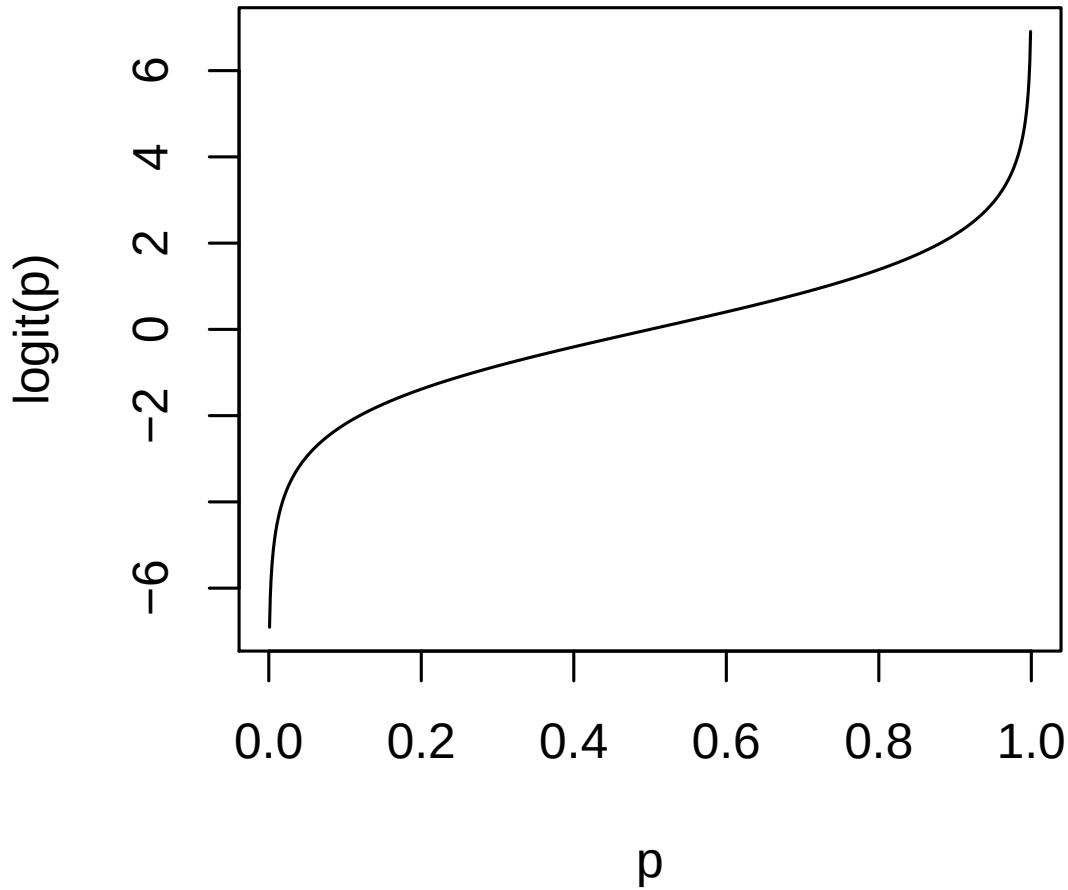
$$p = \Pr(\text{death})$$

So our model is:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta\ (\text{age})$$

As p ranges between 0 and 1 the log odds of p ranges from $-\infty$ to ∞.

As p ranges between 0 and 1 the log odds of p ranges from $-\infty$ to ∞.



We wish to estimate α and β for the model:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta\,(\text{age})$$

Estimates for α and β are found done using a generalization of least squares regression (the method used for linear regression).

There is no formula for the estimates. You have to use a computer to find them.

The estimates are (from R):

	Name in R	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
$a = \hat{\alpha}$	(Intercept)	-3.05851	0.69608	-4.394	1.11e-05
$b = \hat{\beta}$	age	0.02754	0.01056	2.607	0.00913

- The z value is the estimate divided by the standard error.
- The p-values are calculated from the standard normal distribution.

So our estimated model is

$$\log\left(\frac{p}{1-p}\right) = -3.05851 + 0.02754 \text{ age}$$

We would like to be able to calculate the fitted probability of death for any age.

For example if age $= 55$ we have

$$\begin{aligned}
\log\left(\frac{p}{1-p}\right) &= -3.05851 + 0.02754 \times age \\
&= -3.05851 + 0.02754 \times 55 \\
&= -3.05851 + 1.5125 \\
&= -1.54381
\end{aligned}$$

So for age $= 55$:

$$\log\left(\frac{p}{1-p}\right) = -1.54381$$

If we can solve this for p we have the probability of death at age $=$ 55 for ICU patients.

$$\log\left(\frac{p}{1-p}\right) = -1.54381$$

for convenience let $w = -1.54381$.

$$\log\left(\frac{p}{1-p}\right) = w$$

First take the inverse log of both sides

$$\frac{p}{1-p} = e^w$$

multiply both sides by 1-p:

$$\begin{aligned} p &= (1-p)e^w \\ &= e^w - pe^w \end{aligned}$$

add pe^w to both sides

$$p + pe^w = e^w$$
$$p(1 + e^w) = e^w$$

and divide both sides by $(1 + e^w)$

$$p = \frac{e^w}{1 + e^w}$$

We started with

$$\log\left(\frac{p}{1-p}\right) = w$$

and solved for p

$$p = \frac{e^w}{1+e^w}$$

For age = 55 we had

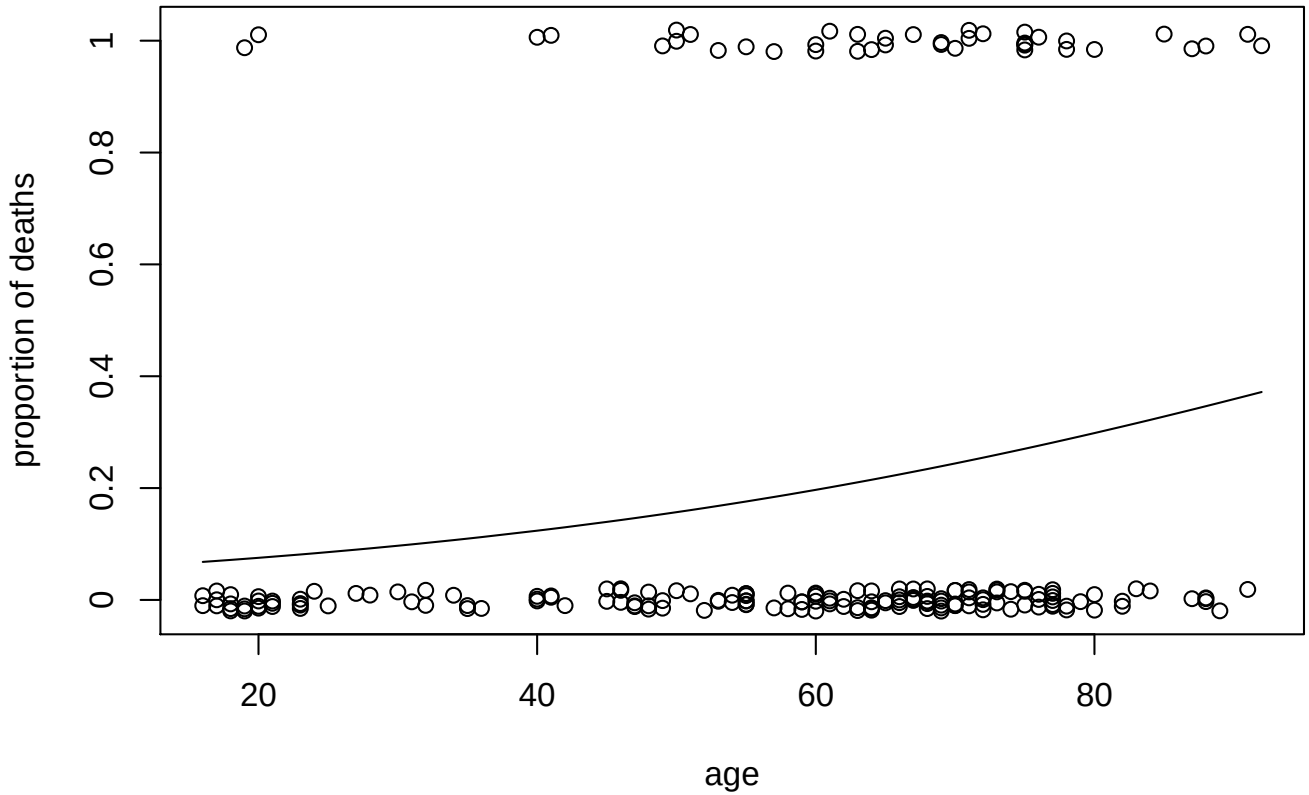
$$\log\left(\frac{p}{1-p}\right) = -1.54381$$

We can now calculate

$$\begin{aligned} p &= \frac{e^{-1.54381}}{1+e^{-1.54381}} \\ &= \frac{0.2136}{1+0.2136} = 0.176 \end{aligned}$$

So the estimated probability of death for individuals age 55 in the ICU is 0.176.

We can calculate predicted probabilities for all ages in the range of our data and plot the result with the data



The reason the line is curved instead of straight is because the fitted values are in the probability scale whereas the linear model is in the logit scale

Assumptions for logistic regression

- The observations are a simple random sample at each value of the predictor variable.
- The model fits the data.
- Sample sizes need to be large enough so that there are enough positive and negative responses (intentionally vague).

Logistic Regression with a categorical predictor

We are also interested in the association between admission type and ICU mortality.

Both are categorical variables so we can use contingency table methods.

In R Commander: Contingency tables → Two-way table

Row variable: admit.type
Column variable: death

```
Frequency table:
           death
admit.type   no yes
  elective   51   2
  emergency 109  38

  Pearson's Chi-squared test

data:  .Table
X-squared = 11.8663, df = 1, p-value = 0.0005716
```

R Commander doesn't compute the odds ratio for us but we can do it by hand.

If we name the counts in the cells:

	death	
	no	yes
elective admission	A	B
emergency admission	C	D

then

$$OR = \frac{AD}{BC} = \frac{51 \times 38}{2 \times 109} = 8.8899$$

We can also fit this model using logistic regression:

	Estimate	Std. Error	z value	Pr($>|z|$)
a	-3.2387	0.7206	-4.495	6.97e-06
b	2.1849	0.7448	2.934	0.00335

This says that the model for the probability of death is:

$$\begin{aligned} \text{logit}(p) &= a + b\,\mathcal{I}_{emergency} \\ &= -3.2387 + 2.1849\,\mathcal{I}_{emergency} \end{aligned}$$

where

$\mathcal{I}_{emergency} = 0$ if admit type $=$ elective

$\mathcal{I}_{emergency} = 1$ if admit type $=$ emergency

So for admit type = elective

$$\begin{aligned} \text{logit}(p) &= -3.2387 + 2.1849\,\mathcal{I}_{emergency} \\ &= -3.2387 + 2.1849 \times 0 \\ &= -3.2387 \end{aligned}$$

and

$$p = \frac{e^{-3.2387}}{1 + e^{-3.2387}} = 0.0377$$

For admit type = emergency

$$\begin{aligned} \text{logit}(p) &= -3.2387 + 2.1849 \times 1 \\ &= -1.0538 \end{aligned}$$

and

$$p = \frac{e^{-1.0538}}{1 + e^{-1.0538}} = 0.2585$$

We can also calculate the odds ratio for the probability of death for emergency with respect to elective from the logistic regression estimates.

We know that

Admit type	$\log(p/(1-p))$	odds $= p/(1-p)$
elective	-3.2387 = a	$e^{-3.2387} = 0.0392$
emergency	-1.0538 = a + b	$e^{-1.0538} = 0.3486$

The odds ratio is

$$\frac{\text{odds for emergency}}{\text{odds for elective}} = \frac{0.3486}{0.0392} = 8.8899$$

which matches what we got from the 2x2 table.

also since

$$\text{odds for emergency admits} = e^{a+b} = e^a e^b$$

$$\text{odds for elective admits} = e^a$$

the odds ratio is

$$\frac{\text{odds for emergency}}{\text{odds for elective}} = \frac{e^a e^b}{e^a} = e^b = e^{2.1849} = 8.8899$$

So, we can calculate the odds ratio for a binary predictor variable as e^b where b is that variables logistic regression coefficient.

The results from logistic regression are often reported as odds ratios rather than the original coefficients.

logistic regression with a categorical predictor - R Commander

To fit the model

$$\text{logit}(p) = a + b\,\mathcal{I}_{emergency}$$

in R Commander use: Statistics $\rightarrow$ Fit models $\rightarrow$ Generalized linear model...

The required input looks complex.

- At the top choose a name for your model fit (optional)
- under "Variables (double click to formula)"
  - double click the response variable (death)
    - it will show up on the left hand side of the model
  - double click the predictor variable (admit.type)
    - it will show up on the right hand side of the model
- click OK

You can ignore the other options.

```
> GLM.3 <- glm(death ~ admit.type, family=binomial(logit), data=icu)

> summary(GLM.3)

Call:
glm(formula = death ~ admit.type, family = binomial(logit), data = icu)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7734  -0.7734  -0.7734  -0.2774   2.5601

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.2387     0.7206  -4.495 6.97e-06 ***
admit.type[T.emergency]   2.1849     0.7448   2.934  0.00335 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...
```

Hypothesis Testing and Confidence Intervals

From the R output:

	Estimate	Std. Error	z value	Pr($>|z|$)
(Intercept)	-3.2387	0.7206	-4.495	6.97e-06
admit.type[T.emergency]	2.1849	0.7448	2.934	0.00335

The fitted model is

$$\text{logit}(p) = -3.24 + 2.18\,\mathcal{I}_{emergency}$$

Test of the hypothesis that ICU death is associated with type of admission:

$$H_0 : b = 0 \text{ vs. } b \neq 0$$

is in the second line of the table.

p $= 0.00335$ so we have strong evidence against H_0.

For the test of the hypothesis that death is not associated with type of admission, the null and alternative hypothesis are

$$H_0 : b = 0 \text{ vs. } H_A : b \neq 0$$

which is equivalent to

$$H_0 : e^b = 1 \text{ vs. } H_A : e^b \neq 1$$

or

$$H_0 : OR = 1 \text{ vs. } H_A : OR \neq 1$$

From the R output we know that the b is significantly different from zero, p = 0.00335.

We can also conclude that the odds ratio is significantly different from 1, p = 0.00335.

Calculating the odds ratio and 95% CI in R

Use menus: Models → Confidence intervals...

```
> Confint(GLM.3, level=0.95, type=''LR'')
                          Estimate      2.5 %    97.5 % ...
(Intercept)             -3.238678 -5.0492011 -2.070861 ...
admit.type[T.emergency]  2.184916  0.9486731  4.024670 ...


                        ... exp(Estimate)       2.5 %     97.5 %
(Intercept)             ...     0.0392157 0.006414456  0.1260771
admit.type[T.emergency]...     8.8899054 2.582280986 55.9618254
```

The coefficients and confidence intervals come out in the original scale and in the odds ratio scale (transformed using the exp function).

The odds ratio is 8.8899 (the same as we got for the 2×2 table) with 95% confidence interval (2.58, 55.96)

logistic regression with a continuous predictor - R Commander

We can also fit our first model in R Commander:

$$\text{logit}(p) = a + b\,age$$

```
> GLM.2 <- glm(death ~ age, family=binomial(logit), data=icu)
> summary(GLM.2)
Call:
glm(formula = death ~ age, family = binomial(logit), data = icu)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9536  -0.7391  -0.6145  -0.3905   2.2854

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05851    0.69608  -4.394 1.11e-05 ***
age          0.02754    0.01056   2.607  0.00913 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...
```

Interpretation of regression coefficients: Here,

$$a = -3.058 \quad \text{and} \quad b = 0.0275$$

The fitted model is

$$\text{logit}(p) = -3.058 + 0.0275\,\text{age}$$

Test of the hypothesis that ICU death is associated with age

$$H_0 : b = 0 \text{ vs. } b \neq 0$$

The test statistic reported in the R output is :

$$z = \frac{b}{\text{sd}(b)} = \frac{0.0275}{0.0106} = 2.607$$

and the P-value is

$$2 \times \Pr\{Z > 2.607\} = 0.0091$$

There is strong evidence against the null hypothesis that age is not related to in hospital death.

As before, We can obtain the confidence intervals for the coefficients

```
> Confint(GLM.2, level=0.95, type=''LR'')
               Estimate        2.5 %      97.5 % ...
(Intercept) -3.05851323 -4.545965431 -1.79733399 ...
age          0.02754261  0.007910569  0.04960994 ...


            ... exp(Estimate)      2.5 %    97.5 %
(Intercept) ...    0.04695746 0.01060992 0.1657402
age         ...    1.02792541 1.00794194 1.0508611
```

The estimates and CIs are given in both the original scale and the odds ratio scale.

So we know the odds ratio for age is 1.0279 with confidence interval (1.008, 1.051)

How do we interpret the odds ratio?

We don't have a 2x2 table since age is a continuous predictor.

In this case e^b is equal to

$$\frac{\text{odds of death at age x+1}}{\text{odds of death at age x}}$$

and this is true for any value of x.

For continuous predictors the odds ratio is the odds of success when the predictor is increased by one unit compared to the odds of success when the predictor is fixed at the original value.

The odds ratio for a continuous predictor will change if the units change.

logistic regression with multiple predictors

What happens if we put both age and admit.type in the model?

```
> GLM.4 <- glm(death ~ age + admit.type, family=binomial(logit), data=i

> summary(GLM.4)

Call:
glm(formula = death ~ age + admit.type, family = binomial(logit),
    data = icu)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1946  -0.7752  -0.4184  -0.2270   2.5181

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -5.50876    1.03351  -5.330 9.81e-08 ***
age                     0.03402    0.01069   3.181  0.00147 **
admit.type[T.emergency] 2.45354    0.75257   3.260  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> Confint(GLM.4, level=0.95, type=''LR'')
                              ... exp(Estimate)        2.5 %      97.5 %
(Intercept)                   ...   0.004051118 0.0004121375  0.02575755
age                           ...   1.034601333 1.0142304820  1.05797632
admit.type[T.emergency] ...  11.629385723 3.3146544903 73.91097653
```

- The coefficients in the logistic regression are estimated after adjusting for the other predictors in the model.
- The coefficients and confidence intervals are not affected by the order that the terms are put in the model.
- The coefficients and odds ratios are not the same as in the single predictor models

	OR	
predictor	1 predictor in model	2 predictors in model
age	1.027	1.035
admit.type	8.890	11.629

Logistic Regression Models: Summary

- Logistic regression provides a framework for relating the log odds, or indirectly, the probability of some event, to predictor variables.
- For a continuous variable, e^b represents the odds ratio for one additional unit of the predictor.
- For a predictor variable with 2 levels, e^b represents the odds ratio for the category listed in the output vs the other category.
- The hypothesis test of $H_0 : b = 0$ is equivalent to testing $H_0 : OR = 1$
- When more than one predictor is included in the model the parameter estimates (and odds ratios) are adjusted for the other predictors.