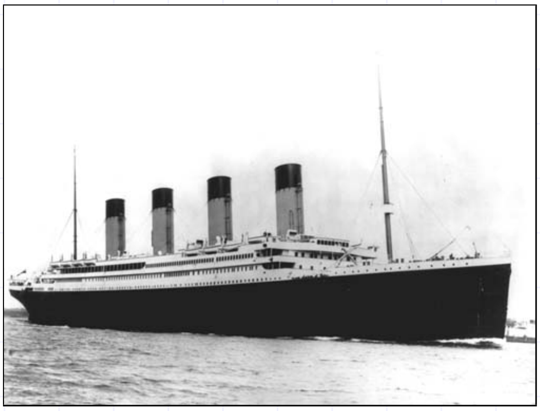
Regression Models
Lecture VII: Generalised Linear Models

DT9002: Postgraduate Certificate in Applied Statistics

Dr Joe Condon

School of Mathematical Sciences
Technological University Dublin

Logistic Regression Motivating Example



- Set sail 11 April 1912 from Cobh (Queenstown).
- Approx. 2200 people on board but only lifeboats for 1200.
- On 14 April she hit an iceberg and sank.
- Approx. 700 survived, the other 1500 perished.

Titanic Data

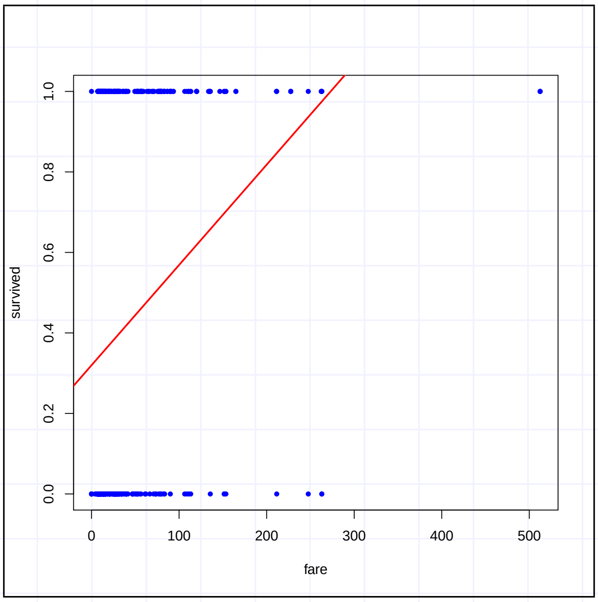
	survived	pclass	sex	age	sibsp	parch	fare	embarked
1	0	3	male	22	1	0	7.2500	S
2	1	1	female	38	1	0	71.2833	C
3	1	3	female	26	0	0	7.9250	S
4	1	1	female	35	1	0	53.1000	S
5	0	3	male	35	0	0	8.0500	S
7	0	1	male	54	0	0	51.8625	S
8	0	3	male	2	3	1	21.0750	S
9	1	3	female	27	0	2	11.1333	S
10	1	2	female	14	1	0	30.0708	C
11	1	3	female	4	1	1	16.7000	S

We will use a representative sample of the data referring to 892 passengers.

Consider the relationship between chances of survival and fare.

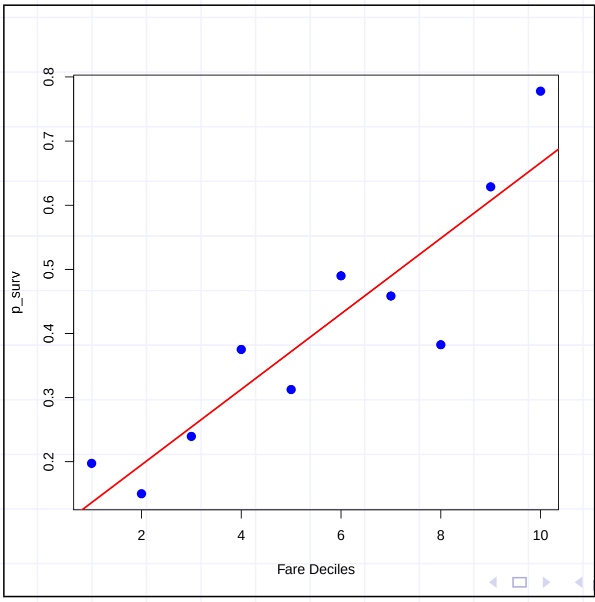
Why not use Linear Regression?

```
plot(fare,survived,pch=20,col='blue')
fitx=lm(survived~fare,data=titanic);abline(fitx,col='red',
    lwd=2)
```



Trying to make linear regression work?

```
group=cut(fare,quantile(fare,probs=c(0,seq(.1,.9,by=.1),1)))
p_surv=by(survived,group,
      function(x) {prop.table(table(x))[2]})
plot(1:10,p_surv,pch=20,cex=2,col='blue',xlab='Fare Deciles'
    );abline(lm(p_surv~I(1:10)),col='red',lwd=2)
```



Problems?

- No guarantee the predicted response is between 0–1.
- Modelling a binary response as though it was continuous and normally distributed - there are implications for hypothesis testing and other statistical inference here.
- Grouping the data using deciles etc. introduced its own problems - e.g. how many observations are we using 800 or 10?
- If the data are really binomial (Bernoulli) then we know the variance changes with the mean (variance is no longer equal across the range of responses) - need to account for this somehow?

An alternative solution is to use a Generalised Linear Model - specifically Logistic Regression.

Formulating the Logistic regression Model

Define the linear predictor of the model just as with linear regression:

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1(\text{predictor 1}) + \beta_2(\text{predictor 2}) + \beta_3(\text{predictor 3}) + \ldots \\ &= \beta_0 + \beta_1(x_{i1}) + \beta_2(x_{i2}) + \beta_3(x_{i3}) + \ldots\end{aligned}$$

We now relate this linear predictor to the probability of the response of interest - e.g. response = surviving the Titanic disaster.

This gives:

$$f(\eta_i) = p(\text{survived}_i = 1)$$

for some mathematical function that ensures $p(\text{survived}_i = 1)$ is a number between 0 and 1.

The Bernoulli response is defined as the RV Y:

$$Y = \begin{cases} 1 & \text{if a Bernoulli success occurs} \\ 0 & \text{if a Bernoulli failure occurs} \end{cases}$$

With probability mass function:

$$p(\text{Bernoulli response} = y) = (p_i)^{y_i}(q_i)^{1-y_i}$$

Let:

$$p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \qquad \text{[called the inverse logit of } \eta_i\text{]}$$

$$\Rightarrow \log \frac{p_i}{1-p_i} = \eta_i \qquad \text{[called the logit of } p_i\text{]}$$

Example:

For the Titanic data, we postulate a model relating fare to the probability that the person survived the disaster.

Let Y_i be the Bernoulli RV for the ith individual (1=survived, 0 otherwise).

$$
\begin{aligned}
\eta_i &= \beta_0 + \beta_1(fare_i) \\
p_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}} \\
p(Y_i = y_i) &= (p_i)^{y_i}(q_i)^{1-y_i} \\
&= \left(\frac{e^{\beta_0+\beta_1(fare_i)}}{1+e^{\beta_0+\beta_1(fare_i)}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1(fare_i)}}\right)^{1-y_i}
\end{aligned}
$$

Fitting the Model: Maximum Likelihood

The values of β_0, β_1 etc. are estimated using the method of maximum likelihood.

The likelihood is the joint probability of the data, and assuming conditional independence between observations we get:

$$
L(\beta_0, \beta_1) = \prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_1(fare_i)}}{1+e^{\beta_0+\beta_1(fare_i)}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1(fare_i)}}\right)^{1-y_i}
$$

We use calculus and numerical methods to find the values of β_0 and β_1 that maximise this expression.

These are the values that make the observed data most likely - hence Maximum Likelihood.

Using the `glm(.)` function

We denote the MLE as $\hat{\beta}_0$, $\hat{\beta}_1$ etc.

```
> fit1=glm(survived~fare,family=binomial(),data=titanic)
> fit1

Call:  glm(formula = survived ~ fare, family = binomial(),
    data = titanic)

Coefficients:
(Intercept)         fare
    -0.8968       0.0160

Degrees of Freedom: 713 Total (i.e. Null);  712 Residual
Null Deviance:       964.5
Residual Deviance: 901.3   AIC: 905.3
```


Interpretation of Parameters from Logistic Regression Models

What do the parameters from a logistic regression model represent?

We get;

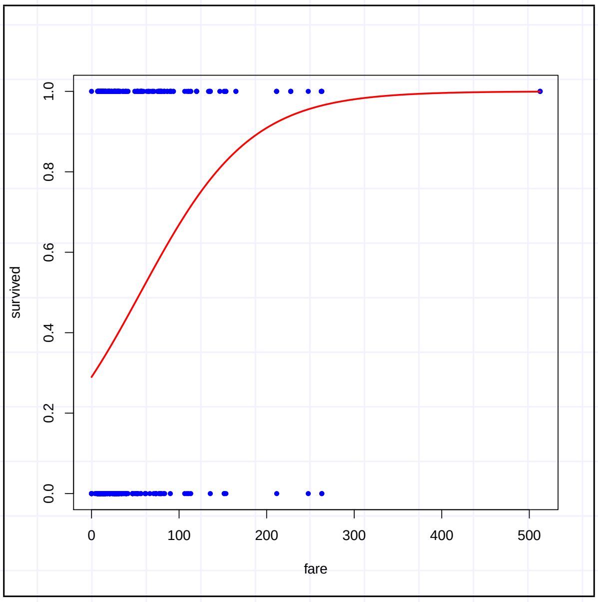
$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_1 x_i$$

So, the linear predictor from a logistic regression model is the logit of p_i for a given x_i (i..e the fare paid by passenger i.

Using the logit inverse function and the estimates from our example we get:

$$\hat{p}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}} = \frac{e^{-0.8968+.016x_i}}{1+e^{-0.8968+.016x_i}}$$

```
nd=data.frame(fare=seq(min(fare),max(fare),len=200))
pred=predict(fit1,type='response',newdata=nd)
plot(fare,survived,pch=20,col='blue')
lines(nd$fare,pred,type='l',lwd=2,col='red',xlab='Fare',ylab
    ='p(survied)')
```



The logit is the log odds for a binary success/failure situation.

1. Probability: $\frac{\text{No of ways to success}}{\text{No of ways to success + No of ways to failure}}$, e.g. $p = .66^{\cdot}$ means in the long run $\frac{2}{3}$ experiments will result in a successes. Range = (0,1)

2. Odds: $\frac{\text{No of ways to success}}{\text{No of ways to failure}}$, e.g. odds = 2 means success is twice as likely as failure(same as probability = $0.66^{\cdot}$). Range=(0,∞).

3. Odds ratio = $\frac{\text{Odds in experiment 1}}{\text{Odds in experiment 2}}$, e.g. odds ratio = 0.5 means the odds of a success in experiment 1 is half that of experiment 2. Range=(0,∞).

Odds ratio are very popular in health information, as they tend to accentuate difference - e.g. - fictitious data -

	Cancer	No Cancer	
Smoker	1,000	2,000	3,000
Non Smoker	1,000	5,000	6,000

So, twice the probability of developing cancer if you smoke - or the odds ratio of a smoker getting cancer is 2.5 times the non smoker.

The parameter estimates from the logistic model are the log odds of success given a set of covariate values.

E.g. For the titanic data, the estimated log odds of a particular passenger who had paid a fare of £8 surviving is ;

$$-0.8968 + 0.0160(8) = -0.7688$$

From this calculate the following:

1. Odds of such a passenger surviving.
2. The probability such a passenger survives.
3. The log odds and odds ratios for a passenger paying a fare of £8 surviving over a passenger who paid £7.

In general we get the following:

$$\begin{aligned}
\text{log odds} = \eta_i &= \beta_0 + \beta_1(x_{i1}) + \beta_2(x_{i2}) + \dots \\
\text{Odds} &= e^{\eta_i} \\
p_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}} \\
\text{Odds ratio for a unit increase in } x_{ij} &= \beta_j
\end{aligned}$$

Categorical Predictors

How do we include a categorical predictor like sex?

We set up dummy variables for each category, like this:

$$d_M = \begin{cases} 1 & \text{if sex = male} \\ 0 & \text{otherwise} \end{cases} \quad d_F = \begin{cases} 1 & \text{if sex = female} \\ 0 & \text{otherwise} \end{cases}$$

Then we fit the model:

$$\begin{aligned}
\eta_i &= \beta_0 + \beta_1(d_M) + \beta_2(d_F) \\
p_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}} \\
p(Y_i = y_i) &= (p_i)^{y_i}(q_i)^{1-y_i} \\
&= \left(\frac{e^{\beta_0+\beta_1(d_M)+\beta_2(d_F)}}{1+e^{\beta_0+\beta_1(d_M)+\beta_2(d_F)}}\right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1(d_M)+\beta_2(d_F)}}\right)^{1-y_i}
\end{aligned}$$

Categorical Predictors...

It turns out that this model cannot be fitted for mathematical reasons, just like before and the solution is to drop one of the dummy variables, i.e. use a set-to-zero constraint (other constraints are possible but rarely used).

It doesn't matter which one is dropped.

Therefore we get the following actually fitted model:

$$\begin{aligned}
\eta_i &= \beta_0 + \beta_2(d_M) \\
p_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}} \\
p(Y_i = y_i) &= (p_i)^{y_i}(q_i)^{1-y_i} \\
&= \left(\frac{e^{\beta_0+\beta_1(d_M)}}{1+e^{\beta_0+\beta_1(d_M)}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1(d_M)}}\right)^{1-y_i}
\end{aligned}$$


```
> fit2=glm(survived~factor(sex),family=binomial(),data=
    titanic); fit2

Call:  glm(formula = survived ~ factor(sex), family =
    binomial(), data = titanic)

Coefficients:
    (Intercept)   factor(sex)male
          1.124            -2.478
```

Calculate the following:

1. The odds ratio that a male survives over a female.
2. The odds ratio that a female survives over a male.
3. The probability that a female/male survives.
4. The odds that a male survives.

Fit a model with `pclass` as a categorical predictor and interpret the parameters.

Fit a model with fare, age, sex, and pclass as predictors and find the following:

1. Find the estimated odds ratio that a female survives over a male - all other variables being equal.
2. Predict the probability that an passenger with the following values of the predictors survives. Sex= female, age = 18 and fare= 16, pclass=2.
3. Calculate the odds that a 50 year old male in second class with a fare of £25 survives.
4. What is the intercept estimating for this model?
5. Give the estimated odds ratio for a 2nd class passenger over a 3rd class passenger.

Hypothesis Testing & CI Estimation

We can derive a practical hypothesis testing procedure and method for calculating CIs from the output from fitted model- the Wald test and interval.

To test $H_0 : \beta_j = \beta_j^0$ versus $H_a : \beta_j \neq \beta_j^0$ use,

$$z = \frac{\hat{\beta}_j - \beta^0}{se(\hat{\beta}_j)} \sim N(0,1) \qquad \text{Or} \qquad \chi^2 = \frac{(\hat{\beta}_j - \beta_j^0)^2}{[se(\hat{\beta}_j)]^2} \sim \chi_1^2 \quad (1)$$

A Wald based CI may be derived as;

$$\text{Wald based CI:} \qquad \hat{\beta}_j \pm z_{1-\alpha/2} se(\hat{\beta}_j) \qquad (2)$$

Testing the null hypothesis:

$$H_0 : \beta_j = 0 \qquad H_a : \beta_j \neq 0$$

can be interpreted as testing if there is statistically significant evidence in the data that the predictor attached to β_j is related to the probability of the response.

A failure to reject the null hypothesis might lead to the conclusion that the predictor is probably not related to the probability of the response.

A rejection of the null hypothesis would lead to the opposite conclusion.

The standard errors used the Wald formulae are found from the variance-covariance matrix of the parameter estimates.

```
> fit4=glm(survived~age+factor(sex),family=binomial(),data=
    titanic)
> fit4

Coefficients:
    (Intercept)              age   factor(sex)male
       1.277273        -0.005426         -2.465920

> vcov(fit4)
                 (Intercept)           age  factor(sex)male
(Intercept)      0.052977676  -1.133223e-03    -1.893426e-02
age             -0.001133223   3.981503e-05    -6.286915e-05
factor(sex)male -0.018934255  -6.286915e-05     3.436704e-02
```

This is all available automatically using the `summary(.)` function.

CI for output statistics

We can get the following:

CI for log odds ratio:

$$\hat{\beta}_j \pm z_{1-\alpha/2} se(\hat{\beta}_j) = (\beta_{j,lower},\ \beta_{j,upper})$$

CI for odds ratio:

$$\left(e^{\beta_{j,lower}},\ e^{\beta_{j,upper}}\right)$$

CI for estimated odds and probabilities

This is a two stage process:

(1) Calculate the CI for the log odds:

$$\text{Let } \hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1(x_{i1}) + \hat{\beta}_2(x_{i2}) + \ldots$$

Find the $se(\hat{\eta}_i)$ from the software. This can be done in a few ways, here are tw0:

(a) estimated standard error using predict function:

```
> predict(fit, newdata=nd,se.fit=T)
```

(b) Using a L matrix and the GLHT function from the multcomp library:

```
> summary(glht(fit,linfct=L))
```

See example below.

CI log odds:

$$\eta_i \pm z_{1-\alpha/2} se(\hat{\eta}_i) = (\eta_{i,lower}, \eta_{i,upper})$$

CI for odds:

$$(e^{\eta_{i,lower}},\ e^{\eta_{i,upper}})$$

CI for probability:

$$\left(\frac{e^{\eta_{i,lower}}}{1 + e^{\eta_{i,lower}}},\ \frac{e^{\eta_{i,upper}}}{1 + e^{\eta_{i,upper}}}, \right)$$


```
fit3=glm(survived~fare+age+factor(sex)+factor(pclass),
family=binomial(),data=titanic)
```

For this model fitted to the Titanic data; answer the following:

1. Discuss the evidence for the variable 'age' being related to the response.
2. Predict the probability that a passenger with the following values of the predictors survives. fare= 82, age = 34 and , sex= female, pclass= 1. Use R to calculate a 95% confidence interval for this fitted probability.
3. Give a 95% CI for the odds ratio for a person with pclass=1 over pclass=2 surviving, all other variables being equal.