The Titanic was a passenger liner that sank in the North Atlantic Ocean on 15 April 1912 after colliding with an iceberg. The sinking of Titanic caused the deaths of 1,514 people out of 2,223 passengers and crew. Due to outdated maritime safety regulations, she carried only enough lifeboats for 1,178 people–one-third her total passenger and crew capacity (`http://en.wikipedia.org/wiki/RMS_Titanic.`)

The outcomes (survival or death) of 1316 passengers are shown in Table 1. In summary, 37.9% of the accounted-for passengers were survivors. As survival depended on being admitted on to a lifeboat, the selection of individuals and their attributes (age, gender, and ticket class) provides some insight into social stratification of the early twentieth century. The table suggests that each factor is associated with survival but it's difficult judge relative importance of the factors and whether the factors interact if we regard survival as a response variable and the three factors as explanatory variables.

12 observations b/c 12 categories.
↳ binomial

Table 1: Summary of mortality on the Titanic.

Age	Gender	Ticket class	Survived	Died	Total
Adult	Male	1st	57	118	175
		2nd	14	154	168
		3rd	75	387	462
	Female	1st	140	4	144
		2nd	80	13	93
		3rd	76	89	165
Child	Male	1st	5	0	5
		2nd	11	0	11
		3rd	13	35	48
	Female	1st	1	0	1
		2nd	13	0	13
		3rd	14	17	31

– Mi; far F/3rd

Two regression approaches can be taken for the analysis of survival: logistic regression (viewing each passenger outcome as a single observation), or, preferably, binomial regression wherein the data are viewed as 12 observations on a binomial random variable. Ramsey and Schafer refer to this second approach as logistic regression for binomial response variables.

Logistic regression for binomial response variables

The binomial random variable describes the number of successes and the probability of each when observing a sequence of independent Bernoulli trials.

Always do binomial when you can and have Bernoulli data

Suppose that m independent Bernoulli trials each with probability of success π are observed and Y counts the number of successes in the m trials. The random variable Y has a *binomial distribution*. The model parameters are m and π. The binomial denominator m is always known whereas π, the probability of success on any particular trial, is usually unknown.

m_{11}	m_{12}	m_{13}
m_{21}	m_{22}	m_{23}

In the Titanic data, there are $n = 12$ binomial responses: $y_1 = 57, m_1 = 175, \ldots, y_{12} = 17, m_{12} = 31$ where the binomial response variable counts the number of survivors. Ignoring the structure of the data, the elementary estimator of π_i is the sample proportion y_i/m_i; for example $\widehat{\pi}_1 = 57/175 = .326$.

The binomial distribution

Let Y count the number of successes in m independent Bernoulli trials, each with a common probability of success π. The possible realizations of Y are $0, 1, 2, \ldots, m$, where m is the number of trials and the binomial denominator. The probability of a particular realization can be computed using the following formula.

$$P(Y = y) = \binom{m}{y} \pi^y (1 - \pi)^{m-y}, \quad y = 0, 1, \ldots, m,$$

m = each passage = an observation

$$\text{where } \binom{m}{y} = \frac{m!}{y!(m - y)!},$$

and

$$y! = \begin{cases} y(y-1)\cdots 2 \cdot 1 & \text{if } y > 0 \\ 1 & \text{if } y = 0. \end{cases}$$

The expected value and standard deviation of Y are:

expected # of survivals

$$\mu = E(Y) = m\pi$$

and

$$\sigma = \sqrt{m\pi(1 - \pi)}.$$

The logistic regression model for binomial responses

A regression model of $E(Y|x) = m\pi$ for a Binomial random variable should not produce fitted values or predictions that are outside the range of Y, that is outside the interval $[0, m]$. Just as with Bernoulli responses ($m = 1$), a linear regression model of $E(Y|x)$ will produce fitted values that are inconsistent with the response variable.

The solution is to model the logit transformation of $\mu = E(Y|x)$ using a linear model.

The logit function is

$\mu = m \cdot \pi$

$$\text{logit}(\mu) = \log\left(\frac{\mu}{m-\mu}\right) = \log\left(\frac{m\pi}{m-m\pi}\right) = \log\left(\frac{\pi}{1-\pi}\right).$$

Usually, logistic regression for binomial data is discussed in terms of modeling π since the model is of π and the transformation of π to μ requires only multiplication of π by the binomial denominator. The linear portion of the regression model is

$$\text{logit}(\pi|x) \quad = \quad \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1}.$$

Parameter estimates are computed using the method of maximum likelihood and model assessment is based on the observed likelihood function. The interpretation of model coefficients is the same as with logistic regression with binary response variables. Hypothesis testing is much the same as with logistic regression with binary response variables except that the Wald statistic is sufficiently accurate to rely upon it for hypothesis tests involving a single parameter. The R function calls used to fit the model are

survive – died

```
resp <- cbind(y,m-y)
summary(glm(resp~Age+Class+Sex,family=binomial))
```

The response variable is a matrix consisting of two columns: the left column is $y_1, \ldots, y_{12}$ and the right column is $m_1 - y_1, \ldots, m_{12} - y_{12}$.[1] If the data are not aggregated as 12 binomial observations but remain as the original 1316 binary responses, the model also can be fit (producing the same parameter estimates). There are however, important opportunities to examine model fit to a greater extent than when the data are binary response variables.

Table 2 shows the fitted additive model for the titanic data.

Table 2: Coefficients and standard errors obtained from a logistic regression of survival on age (child or adult), ticket class, and gender. Residual deviance is 110.84 on 7 degrees of freedom, $n = 12$.

Variable	Coefficient	Std. Error	Wald statistic (Z)	$P(Z > \|z\|)$
Intercept	2.006	.169	11.8	< .0001
Age (Child)	1.05	.243	4.35	< .0001
Class (2nd)	−1.01	.195	−5.18	< .0001
Class (3rd)	−1.76	.171	−10.3	< .0001
Gender (Male)	−2.37	.145	−16.3	< .0001

gives survival
prob of reference
levels
Adult
Female
1st

It's clear from the Wald statistics that all of the variables are important. To determine if the effect of gender depends on age, a model was fit including the interaction of gender and

[1] There are other syntaxes that can be used to fit the model.

$e^{1.05} = 2.3$ = child 2.3 times greater surviving than adult of same gender and class

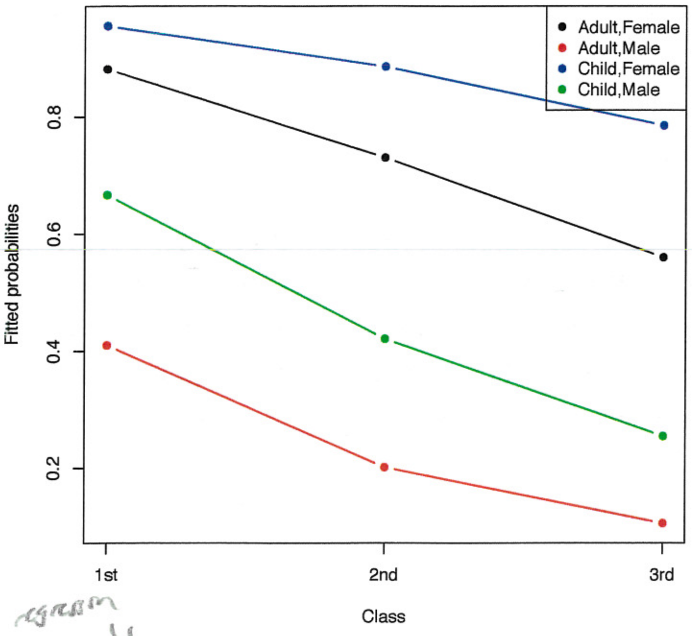
b/c replicates in each cell can use wald stat age. The drop-in-deviance test yields the likelihood ratio statistic

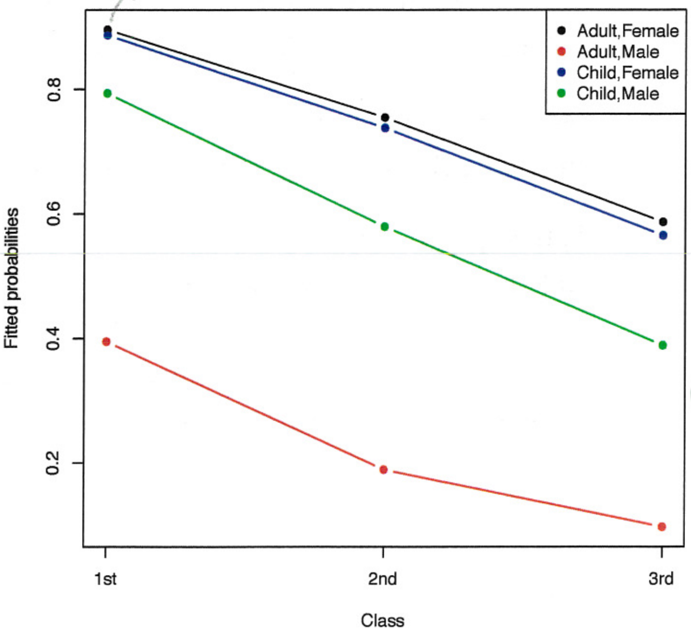
$$\begin{aligned}\text{LRT} &= \text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}}\\ &= 110.84 - 93.67\\ &= 17.17.\end{aligned}$$

The degrees of freedom associated with the LRT statistic are 1 and so p-value = $P(\text{LRT} \geq 17.17) < .0001$.

The figures below show the additive model (left) and the interaction model (right). The interaction model leads to the inference that the status of females is the same for both adults and children whereas there is a distinction between ages for males.

no age distinction b/t females





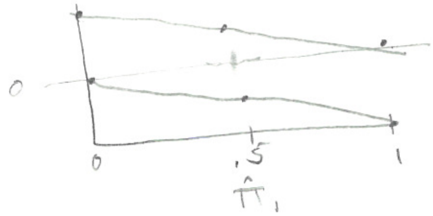
Logistic regression useless residuals b/c binary

Residuals

Logistic regression with binary responses does not lead to useful residuals since the observed values are 0 or 1. A residual of the form $y_i - \widehat{y}_i$ will take on one of two values, specifically,

$$y_i - \widehat{y}_i = \begin{cases} 1 - \widehat{\pi}_i, & \text{if } y_i = 1\\ -\widehat{\pi}_i, & \text{if } y_i = 0.\end{cases}$$

Consequently, plotting the residuals against the fitted values $\widehat{y}_i = \widehat{\pi}_i$ produces a figure with points falling on two parallel lines (one corresponding to responses with value $y_i = 1$ and the second corresponding to responses with values $y_i = 0$). There's no real information about outliers or unusual responses in the plot. When the binomial denominator m_i is larger than



if $y_i = 1$

if $y_i = 0$

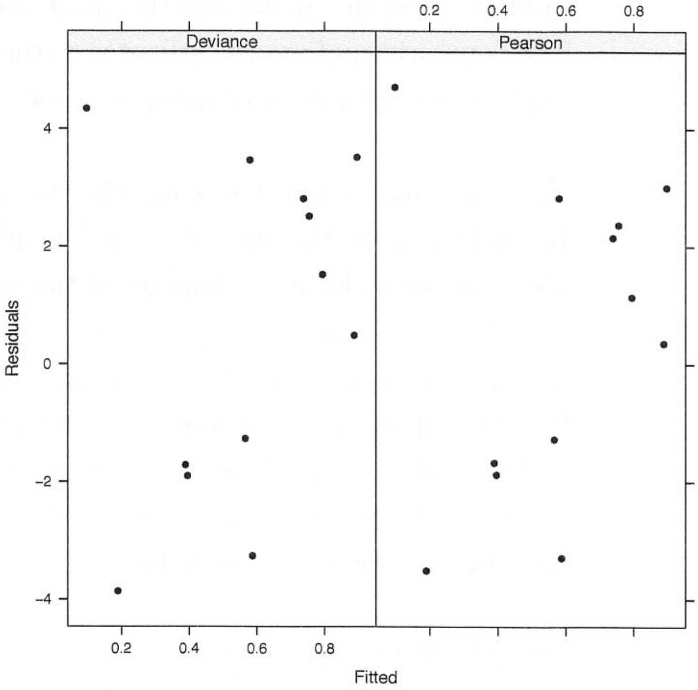
1 for most of responses, then the Pearson residuals can be used to analyze the fit of the model. The Pearson residual associated with the response pair (y_i, m_i) is

$$\mathrm{Pres}_i = \frac{y_i - m_i\widehat{\pi}_i}{\sqrt{m_i\widehat{\pi}_i(1-\widehat{\pi}_i)}.}$$

diff b/t fit $\hat{E}(y|x)$
and actual y
residual

estimate of
variability

$-m_i \le y_i - m_i\hat{\pi}_i \le m_i$



The Pearson residual is an analogue of the standardized residual[2] of multiple linear regression since numerator of Pres_i is the difference between the observed and fitted value, and the denominator is the estimated standard deviation of y_i.

The deviance residual is also used. This residual is the square root of the contribution of the ith observation towards the deviance, the measure of model (lack-of) fit. The motivation for this residual is that the squared standardized residuals used in multiple linear regression sum to the residual sum-of-squares. The deviance residual is

$$\mathrm{Dres}_i = \mathrm{sign}(y_i - m_i\widehat{\pi}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{m_i\widehat{\pi}_i}\right) + (m_i - \widehat{\pi}_i)\log\left(\frac{m_i - y_i}{m_i - m_i\widehat{\pi}_i}\right)\right]}.$$

The deviance associated with a particular model is $\sum_i^n \mathrm{Dres}_i^2$. The Figure above shows the Pearson residuals on the left and the deviance residuals on the right. The distributions of residuals are remarkably similar (usually true) and so there is little practical difference in the choice of residuals. There are no unusual departures from a random scatter of points. Comparing the Pearson residuals to 2 and −2 reveals several residuals that are unusually large in magnitude. Two obvious residuals originate from children (males and females) in 2nd class, all of whom were survivors. More generally, it's not reasonable to suppose that three factors can possibly adequately explain survivorship in such a chaotic situation.

The R functions for computing the Pearson and deviance residuals from a fitted model `glm.1` are `residuals(glm.1, "pearson")` and `residuals(glm.1, "deviance")`. The fitted values can be extracted using the function call `fitted(glm.1)`.

[2] The standardized residual is

$$\mathrm{res}_i = \frac{y_i - \widehat{\mu}_i}{\widehat{\sigma}(\widehat{\mu}_i)}.$$

Goodness of fit

Look for pattern

For simple linear regression problems, if there were sufficient numbers of replicate observations at most or all of the explanatory variable values, then a lack-of-fit test could be executed. The test compared the model estimates of the response to the averages of all response values associated with a particular treatment combination or group. The difference in fit between the model and the group mean estimates was a measure of fit, or lack thereof. The comparison of model estimates to the group mean estimates was formalized by computing the the extra-sums-of-squares F-test.

There are also replicates when the binomial denominators are greater than 1 since $Y_i \sim \text{Binom}(m_i, \pi_i)$ is the sum of m_i independent Bernoulli responses. The deviance goodness-of-fit test provides a comparison of the logistic regression model to a model with separate, completely unconstrained probabilities for each $Y_i, i = 1, \ldots, n$. This alternative model is usually called the *saturated model* since the saturated model estimates can be computed either by computing the sample proportions (or empirical probabilities) Y_i/m_i for $i = 1, \ldots, n$ or by constructing a logistic regression model with an indicator variable identifying each binomial observation. Since this model demands one parameter for each observation, the model is saturated with variables.

The `glm` R function computes the goodness-of-fit test for every fitted model (even when the test is completely inappropriate). The test is inappropriate if the m_i's are mostly or all ones. It's reported as `Residual deviance`; specifically,

$$\texttt{Residual deviance} = -2\left[\log(L_{\text{constrained}}) - \log(L_{\text{saturated}})\right]$$

where $L_{\text{constrained}}$ is the likelihood function evaluated using the model of interest with parameters $\beta_0, \ldots, \beta_{p-1}$ to estimate $\pi_i, i = 1, \ldots, n$, and $L_{\text{saturated}}$ is the likelihood function computed using the empirical probabilities or sample proportions as estimators of $\pi_i, i = 1, \ldots, n$. For example, the goodness-of-fit test is reported in the caption of Table 2 as `Residual deviance` = 110.84 on on 7 degrees of freedom. The p-value is approximately the upper tail area of a chi-square distribution with 7 degrees of freedom. The R function call `1-pchisq(110.84,7)` yields p-value $< .0001$, and so there is convincing evidence of lack of fit associated with this model. The residual deviance for the second model with the interaction term between age and gender is 93.6 on 6 degrees of freedom, and so this model too, does not fit the data well. It's not possible to obtain a small goodness of fit statistic with these data.

In book saturated = 1 parameter per observation so exact fit b/t fitted and saturated values

In situations in which the model fails to fit the data, quasi-likelihood methods are used to adjust significance tests. The topic is delayed until the discussion of Poisson regression.

$m_i > 5$ for most cells to compute p-value

* p-value large = good fit of model b/c Ho: model fit, Ha: model doesn't fit