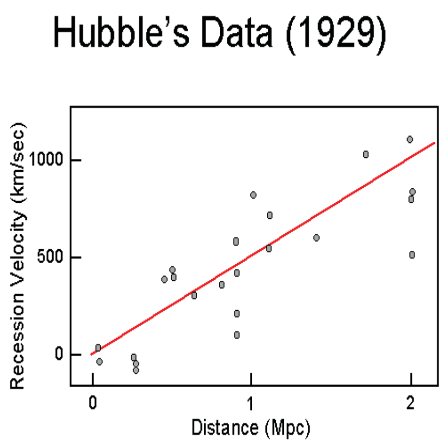
Logistic Regression

Ariel Alonso Abad

Catholic University of Leuven

Linear regression

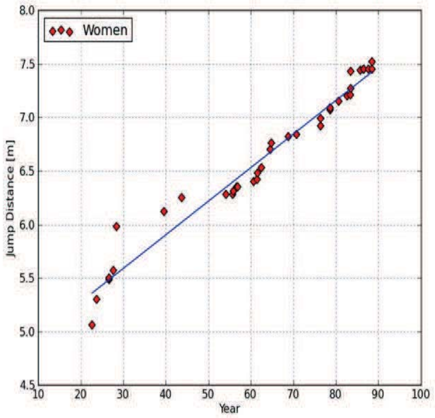
Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$



- Hubble's law

Linear regression

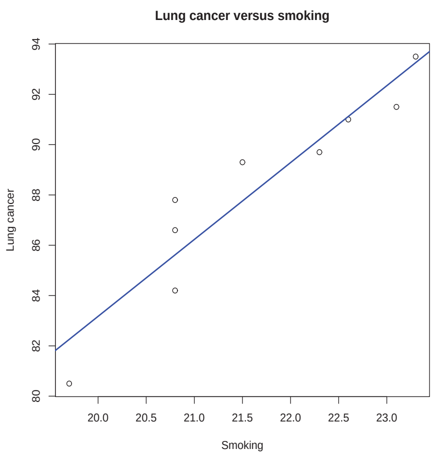
Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$



- Hubble's law
- Sport

Linear regression

Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

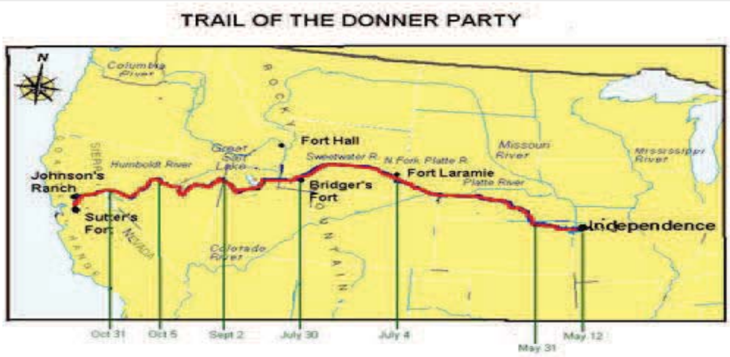


- Hubble's law
- Sport
- Smoking and lung cancer
- Linear regression versus Logistic regression

Legends of America: Donner party data

Donner party expedition

In 1846, the Donner and Reed families left Illinois for California, a 2500 mile journey that would become one of the greatest tragedies in USA history. Stranded in Sierra Nevada by a series of snowstorms, they were rescued in April of the following year. 40 members died, some (or perhaps all) of those that survived did so by resorting to cannibalism.



Donner party data

- 88 persons
- Variables: survival, gender and age
- Taking into account age, are the chances of survival larger for women than for men?

Donner party data

Name	Sex	Age	Surviva
Antoine	Male	23	No
Breen, Mary	Female	40	Yes
Breen, Patrick	Male	40	Yes
Burger, Charles	Male	30	No
Denton, John	Male	28	No
Dolan, Patrick	Male	40	No
Donner, Elizabeth	Female	45	No
Donner, George	Male	62	No
Donner, Jacob	Male	65	No
Donner, Tamsen	Female	45	No
Eddy, Eleanor	Female	25	No
Eddy, William	Male	28	Yes
Elliot, Milton	Male	28	No
Fosdick, Jay	Male	23	No
Fosdick, Sarah	Female	22	Yes
Foster, Sarah	Female	23	Yes
Foster, William	Male	28	Yes
Graves, Eleanor	Female	15	Yes
Graves, Elizabeth	Female	47	No
Graves, Franklin	Male	57	No
Graves, Mary	Female	20	Yes
Graves, William	Male	18	Yes
Halloran, Luke	Male	25	No
Hardkoop, Mr.	Male	60	No
Herron, William	Male	25	Yes
Noah, James	Male	20	Yes
Keseberg, Lewis	Male	32	Yes
Keseberg, Phillipine	Female	32	Yes
McCutcheon, Amanda	Female	24	Yes
McCutcheon, William	Male	30	Yes
Murphy, John	Male	15	No
Murphy, Lavina	Female	50	No
Pike, Harriet	Female	21	Yes
Pike, William	Male	25	No
Reed, James	Male	46	Yes
Reed, Margaret	Female	32	Yes
Reinhardt, Joseph	Male	30	No
Shoemaker, Samuel	Male	25	No
Smith, James	Male	25	No
Snyder, John	Male	25	No
Spitzer, Augustus	Male	30	No
Stanton, Charles	Male	35	No
Trubode, J.B.	Male	23	Yes
Williams, Baylis	Male	24	No
Williams, Eliza	Female	25	Yes

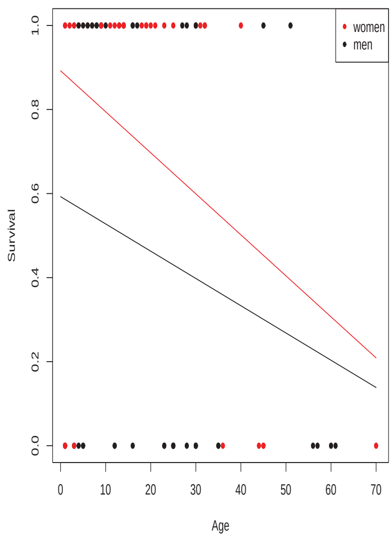
- Dependent variable is binary:

$$Y = \begin{cases} 1 & \text{Survived,} \\ 0 & \text{Died.} \end{cases}$$

- Independent predictors:

$$age, fem = \begin{cases} 1 & \text{for women,} \\ 0 & \text{for men.} \end{cases}$$

Exploring the data



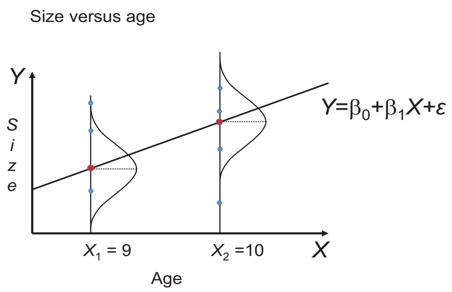
- Survived $= 1$, Died $= 0$
- Graph is not as informative as in linear regression

Linear regression

- Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- The expected value (average) of Y is modeled as a linear function of the predictors

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Simple linear regression



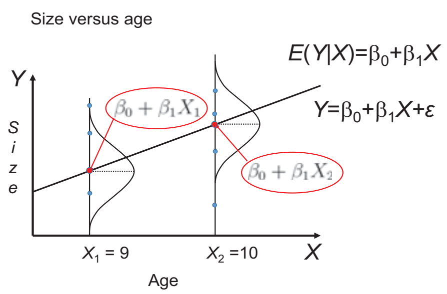
- $Y = \beta_0 + \beta_1 X + \varepsilon$

Linear regression

- Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- The expected value (average) of Y is modeled as a linear function of the predictors

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Simple linear regression



- $Y = \beta_0 + \beta_1 X + \varepsilon$
- $E(Y|X) = \beta_0 + \beta_1 X$

Binary outcome: Problem

The expected value (average) of a binary variable is a **probability**

$$E(Y|X_1, X_2, \ldots, X_p) = P(Y = 1|\boldsymbol{X})$$

where $P(Y = 1|\boldsymbol{X})$ gives the probability as a function of the covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$

$$\boxed{P(Y = 1|\boldsymbol{X})} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$0 \leq P(Y = 1|\boldsymbol{X}) \leq 1$$

but $\eta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ is not always between 0 and 1.

Binary outcome: Problem

The expected value (average) of a binary variable is a **probability**

$$E(Y|X_1, X_2, \ldots, X_p) = P(Y = 1|\boldsymbol{X})$$

where $P(Y = 1|\boldsymbol{X})$ gives the probability as a function of the covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$

$$\boxed{P(Y = 1|age, fem)} = \beta_0 + \beta_1 age + \beta_2 fem$$

$$0 \leq P(Y = 1|age, fem) \leq 1$$

but $\eta(age, fem) = \beta_0 + \beta_1 age + \beta_2 fem$ is not always between 0 and 1.

Ariel Alonso Logistic Regression 9 / 72

Binary outcome: Problem

Problem with the linear model

$$P(Y = 1|fem) = \beta_0 + \beta_1 fem$$

Assume that $\beta_0 = 0.5$ en $\beta_1 = -1$. What is the survival probability for a woman?

$$P(Y = 1|fem) = 0.5 - fem$$

For women: $fem = 1$ thus

$$\begin{aligned} P(Y = 1|fem = 1) =& 0.5 - fem \\ =& 0.5 - 1 = -0.5 \end{aligned}$$

A probability should always be between 0 and 1!

Binary outcome: Solution

- Transform $P(Y = 1|\boldsymbol{X})$:

$$\text{logit}\,[P(Y = 1|\boldsymbol{X})] = \ln \frac{P(Y = 1|\boldsymbol{X})}{P(Y = 0|\boldsymbol{X})} = \ln\left(\frac{P(Y = 1|\boldsymbol{X})}{1 - P(Y = 1|\boldsymbol{X})}\right)$$

- $-\infty \leq \text{logit}\,[P(Y = 1|\boldsymbol{X})] \leq \infty$
- Model:

$$\text{logit}\,[P(Y = 1|\boldsymbol{X})] = \eta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Transform back:

$$P(Y = 1|\boldsymbol{X}) = \frac{e^{\eta(\boldsymbol{X})}}{1 + e^{\eta(\boldsymbol{X})}}$$

Binary outcome: Solution

- Transform $P(Y=1|\boldsymbol{X})$:

$$\text{logit}\left[P(Y=1|\boldsymbol{X})\right] = \ln\frac{P(Y=1|\boldsymbol{X})}{P(Y=0|\boldsymbol{X})} = \ln\left(\frac{P(Y=1|\boldsymbol{X})}{1-P(Y=1|\boldsymbol{X})}\right)$$

- $-\infty \leq \text{logit}\left[P(Y=1|\boldsymbol{X})\right] \leq \infty$
- Model:

$$\text{logit}\left[P(Y=1|\boldsymbol{X})\right] = \eta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Transform back:

$$P(Y=1|\boldsymbol{X}) = \frac{e^{\eta(\boldsymbol{X})}}{1+e^{\eta(\boldsymbol{X})}} = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- Now is $P(Y=1|\boldsymbol{X})$ always between 0 en 1

Binary outcome: Solution

- Transform $P(Y=1|\boldsymbol{X})$:

$$\text{logit}\left[P(Y=1|\boldsymbol{X})\right] = \ln\frac{P(Y=1|\boldsymbol{X})}{P(Y=0|\boldsymbol{X})} = \ln\left(\frac{P(Y=1|\boldsymbol{X})}{1-P(Y=1|\boldsymbol{X})}\right)$$

- $-\infty \leq \text{logit}\left[P(Y=1|\boldsymbol{X})\right] \leq \infty$
- Model:

$$\text{logit}\left[P(Y=1|age, fem)\right] = \eta(age, fem) = \beta_0 + \beta_1 age + \beta_2 fem$$

- Transform back:

$$P(Y=1|age, fem) = \frac{e^{\eta(age, fem)}}{1+e^{\eta(age, fem)}} = \frac{e^{\beta_0+\beta_1 age+\beta_2 fem}}{1+e^{\beta_0+\beta_1 age+\beta_2 fem}}$$

- Now is $P(Y=1|age, fem)$ always between 0 en 1

Binary outcome: Solution

Logistic Model

$$P(Y=1|fem) = \frac{e^{\beta_0+\beta_1 fem}}{1+e^{\beta_0+\beta_1 fem}}$$

Assume that $\beta_0 = 0.5$, $\beta_1 = -1$. What is the survival probability for a woman?

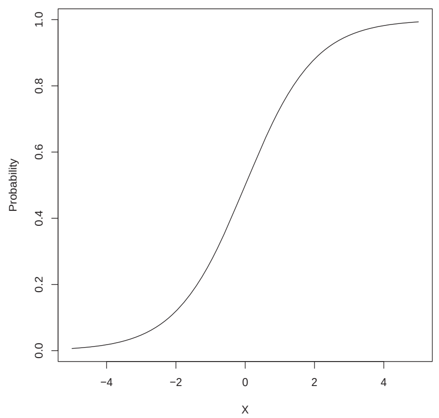
$$P(Y=1|fem) = \frac{e^{0.5-fem}}{1+e^{0.5-fem}}$$

For women: $fem = 1$ thus

$$P(Y=1|fem=1) = \frac{e^{0.5-1}}{1+e^{0.5-1}} = \frac{e^{-0.5}}{1+e^{-0.5}} = 0.378$$

Now the the survival probability for a woman is between 0 en 1!

Model for the probability: X continuous



$$P(Y=1|X) = \frac{e^X}{1+e^X}$$

Logarithm (log)

Different notations: log

Estimating the parameters

$$\text{logit}\left[P(Y=1|\boldsymbol{X})\right] = \eta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Ordinary least squares (OLS) is not suitable
- Method of maximum likelihood (ML) is the adequate choice

Maximum Likelihood Estimation

Find the values of the parameters so that the likelihood is maximized. In other words, the ML estimates (MLE) are the values of the parameters that make the observed data most likely to have been observed. For linear models OLS and ML are equivalent.

Fitting the model in R

```
> ## Home
>
> setwd("C:\\Users\\Logistic-Regression"')
>
> ## Reading the data
>
> donner<-read.table("donner-class.txt", row.names = 1, header=TRUE)
> head(donner,10)
                         Age Outcome    Sex Family.name Status
Breen_Edward_             13       1   Male       Breen Family
Breen_Margaret_Isabella    1       1 Female       Breen Family
Breen_James_Frederick      5       1   Male       Breen Family
Breen_John                14       1   Male       Breen Family
Breen_Margaret_Bulger     40       1 Female       Breen Family
Breen_Patrick             51       1   Male       Breen Family
Breen_Patrick_Jr.          9       1   Male       Breen Family
Breen_Peter                3       1   Male       Breen Family
Breen_Simon_Preston        8       1   Male       Breen Family
Donner_Elitha_Cumi        13       1 Female    G_Donner Family
>
```

Fitting the model in R

```
> ## Keeping only the variables of interest
>
> donner.na<-na.omit(subset(donner,select=c('Age','Outcome','Sex')))
> donner.na$fem = as.numeric(donner.na$Sex=="Female")
> head(donner.na,10)
>
                         Age Outcome    Sex fem
Breen_Edward_             13       1   Male   0
Breen_Margaret_Isabella    1       1 Female   1
Breen_James_Frederick      5       1   Male   0
Breen_John                14       1   Male   0
Breen_Margaret_Bulger     40       1 Female   1
Breen_Patrick             51       1   Male   0
Breen_Patrick_Jr.          9       1   Male   0
Breen_Peter                3       1   Male   0
Breen_Simon_Preston        8       1   Male   0
Donner_Elitha_Cumi        13       1 Female   1
>
```


Fitting the model in R

```
> ## Fitting a logistic regression
>
> donner.log<-glm(Outcome ~ Age + fem,data=donner.na,family=binomial(link="logit"))
> summary(donner.log)

Call:
glm(formula = Outcome ~ Age + fem, family = binomial(link = "logit"), data = donner.na)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8828  -1.0383   0.6511   1.0261   1.7386

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.55382    0.41788   1.325   0.1851
Age         -0.03561    0.01525  -2.336   0.0195 *
fem          1.06798    0.48229   2.214   0.0268 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 108.87  on 85  degrees of freedom
AIC: 114.87

Number of Fisher Scoring iterations: 4
```

Donner party data

Model

$$\text{logit}\left[P(Y=1|age,fem)\right] = \beta_0 + \beta_1 age + \beta_2 fem$$

where $fem = 1$ for women $fem = 0$ for men. Equivalently

$$P(Y=1|age,fem) = \frac{e^{\beta_0+\beta_1 age+\beta_2 fem}}{1+e^{\beta_0+\beta_1 age+\beta_2 fem}}$$

Estimated model

$$\text{logit}\left[\hat{P}(Y=1|age,fem)\right] = 0.553 - 0.035 age + 1.067 fem$$

$$\hat{P}(Y=1|age,fem) = \frac{e^{0.553-0.035 age+1.067 fem}}{1+e^{0.553-0.035 age+1.067 fem}}$$

Interpretation of the coefficients

Linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Association between (y,x): $\mathrm{r}_{xy} = \dfrac{\mathrm{cov}(x,y)}{\sigma_y \sigma_x}$
  - Range: $-1 \leq \mathrm{r}_{xy} \leq 1$
  - Positive correlation between x en y: $\mathrm{r}_{xy} > 0$ $(\beta_1 > 0)$
  - No correlation between x en y: $\mathrm{r}_{xy} = 0$ $(\beta_1 = 0)$
  - Negative correlation between x en y: $\mathrm{r}_{xy} < 0$ $(\beta_1 < 0)$
- $\beta_1 = \mathrm{r}_{xy} \dfrac{\sigma_y}{\sigma_x}$

Association in a 2×2 cross-table

Hypothetical example

		Predictor	
		Female $fem = 1$	Male $fem = 0$
Criterium	Survived $(Y = 1)$	$P(Y=1) = \frac{2}{3}$	$P(Y=1) = \frac{1}{3}$
	Died $(Y = 0)$	$P(Y=0) = \frac{1}{3}$	$P(Y=0) = \frac{2}{3}$

Association in a 2×2 cross-table

Odds

- Odds of surviving for women:

$$\Theta_{survival|women} = \frac{P(Y=1|fem=1)}{P(Y=0|fem=1)} = \frac{2/3}{1/3} = 2$$

$\Rightarrow$ for every 2 women that survive 1 dies

- Odds of surviving for men:

$$\Theta_{survival|men} = \frac{P(Y=1|fem=0)}{P(Y=0|fem=0)} = \frac{1/3}{2/3} = \frac{1}{2} = 0.5$$

$\Rightarrow$ for every man that survives 2 die

(term comes from horse racing)

Association in a 2×2 cross-table

Odds Ratio

Ratio of odds or odds ratio = is a measure of association in 2×2 cross-tables

$$\text{Odds Ratio:} \ OR = \frac{\Theta_{survival|women}}{\Theta_{survival|men}} = \frac{2}{0.5} = 4$$

- Interpretation: the odds of survival for women are 4 times larger than the odds of survival for men

(if the odds for men are 0.5 to 1 then for women they are 2 to 1)

Association in a 2×2 cross-table

Predictor

Criterium		$X=1$ Female	$X=0$ Male
	$Y=1$	$P(Y=1)=\frac{2}{3}$	$P(Y=1)=\frac{2}{3}$
	$Y=0$	$P(Y=1)=\frac{1}{3}$	$P(Y=1)=\frac{1}{3}$

$$\Theta_{survival|women} = \frac{P(Y=1|fem=1)}{P(Y=0|fem=1)} = \frac{2/3}{1/3} = 2$$

$$\Theta_{survival|men} = \frac{P(Y=1|fem=0)}{P(Y=0|fem=0)} = \frac{2/3}{1/3} = 2$$

$$OR = \frac{\Theta_{survival|women}}{\Theta_{survival|men}} = \frac{2}{2} = 1$$

Properties of odds ratios

Odds Ratio

- $0 < OR < \infty$
- $OR = 1 \Leftrightarrow$ independence
- $OR > 1 \Leftrightarrow$ positive association
- $OR < 1 \Leftrightarrow$ negative association

Interpretation of the coefficients

Dichotomous predictor ($X = 0$ of 1 like gender) probability model

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \begin{cases} \dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} & X = 1, \text{female} \\ \dfrac{e^{\beta_0}}{1+e^{\beta_0}} & X = 0, \text{male} \end{cases}$$

Predictor

		$X = 1$ Female	$X = 0$ Male
Criterium	$Y = 1$ Survived	$P(Y=1) = \dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$P(Y=1) = \dfrac{e^{\beta_0}}{1+e^{\beta_0}}$
	$Y = 0$ Died	$P(Y=0) = \dfrac{1}{1+e^{\beta_0+\beta_1}}$	$P(Y=0) = \dfrac{1}{1+e^{\beta_0}}$

Computing the odds ratio

		Predictor	
		$X = 1$ Female	$X = 0$ Male
Criterium	$Y = 1$	$\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\dfrac{e^{\beta_0}}{1+e^{\beta_0}}$
	$Y = 0$	$\dfrac{1}{1+e^{\beta_0+\beta_1}}$	$\dfrac{1}{1+e^{\beta_0}}$

$$OR = \frac{\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \dfrac{1}{1+e^{\beta_0+\beta_1}}}{\dfrac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \dfrac{1}{1+e^{\beta_0}}}$$

$$= e^{\beta_1}$$

$$\begin{aligned} OR &= e^{\beta_1} \\ \ln(OR) &= \beta_1 \end{aligned}$$

Donner party data

- $\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035 age + 1.067 fem$
- Gender: The odds of survival for a woman are 3 times larger than the odds of survival for a man of the same age: $\hat{OR} = e^{1.067} \approx 3$

Age=25

Age=25

$$\Theta_{survival|women} = 3 \cdot \Theta_{survival|men}$$

Donner party data

- $\text{logit}(\hat{P}(Y=1|age, fem)) = 0.553 - 0.035 age + 1.067 fem$
- Gender: The odds of survival for a woman are 3 times larger than the odds of survival for a man of the same age: $\hat{OR} = e^{1.067} \approx 3$

Age=50

Age=50

$$\Theta_{survival|women} = 3 \cdot \Theta_{survival|men}$$

Interpretation of the coefficients

Logistic regression model

$$\text{logit}\left[P(Y=1|X)\right] = \beta_0 + \beta_1 X$$

- Association between (y, x): OR
  - Range: $0 < OR < \infty$
  - Positive association between x en y: $OR > 1$ $(\beta_1 > 0)$
  - No association between x en y: $OR = 1$ $(\beta_1 = 0)$
  - Negative association between x en y: $OR < 1$ $(\beta_1 < 0)$
- $\beta_1 = \ln(OR)$

Continuous predictor X

- The interpretation is analogous to the one given for dummy predictor
- For instance, consider two ages
  - A1: Age=X
  - A2: A year older, Age=$X+1$

$$\Theta_{survival|X+1} = \frac{P(Y=1|X+1)}{P(Y=0|X+1)} \qquad \Theta_{survival|X} = \frac{P(Y=1|X)}{P(Y=0|X)}$$

$$\Theta_{survival|X+1} = e^{\beta_1} \cdot \Theta_{survival|X}$$

Donner party data

- $\text{logit}(\hat{P}(Y=1|age, fem)) = 0.553 - 0.035 age + 1.067 fem$
- Age: $\hat{\beta}_1 = -0.0356 \Rightarrow \hat{OR} = e^{-0.0356} = 0.965 \approx 0.96$, an increase of one year in age is associated with a 4% decrease in the odds of survival

Age=26

Age=25

$$\Theta_{survival|26,men} = 0.96 \cdot \Theta_{survival|25,men}$$

Donner party data

- $\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035 age + 1.067 fem$
- Age: $\hat{\beta}_1 = -0.0356 \Rightarrow \hat{OR} = e^{-0.0356} = 0.965 \approx 0.96$, an increase of one year in age is associated with a 4% decrease in the odds of survival

Age=51

Age=50

$$\Theta_{survival|51,men} = 0.96 \cdot \Theta_{survival|50,men}$$

Continuous predictor X

- A one unit change in X is not always meaningful
- For instance, consider two ages
  - A1: Age=X
  - A2: c years older, Age=$X + c$

$$\Theta_{survival|X+c} = \frac{P(Y = 1|X + c)}{P(Y = 0|X + c)} \qquad \Theta_{survival|X} = \frac{P(Y = 1|X)}{P(Y = 0|X)}$$

$$\Theta_{survival|X+c} = e^{\beta_1 \cdot c} \cdot \Theta_{survival|X}$$

Donner party data

- $\text{logit}(\hat{P}(Y=1|age, fem)) = 0.553 - 0.035 age + 1.067 fem$
- Age: $\hat{\beta}_1 = -0.0356 \Rightarrow \hat{OR} = e^{-0.0356 \cdot 10} = 0.70$, a 10 years increase in age is associated with a 30% decrease in the odds of survival

Age=30

Age=20

$$\Theta_{survival|30,women} = 0.70 \cdot \Theta_{survival|20,women}$$

Donner party data

- $\text{logit}(\hat{P}(Y=1|age, fem)) = 0.553 - 0.035 age + 1.067 fem$
- Age: $\hat{\beta}_1 = -0.0356 \Rightarrow \hat{OR} = e^{-0.0356 \cdot 10} = 0.70$, a 10 years increase in age is associated with a 30% decrease in the odds of survival

Age=60

Age=50

$$\Theta_{survival|60,women} = 0.70 \cdot \Theta_{survival|50,women}$$

More predictors

- The same approach as above: change in logit cause by increasing predictor X_j by 1 unit and keeping all the other predictors fixed is back transformed into a change in odds ratio
- Often called "adjusted odds ratio" ("adjusted" by other predictors)

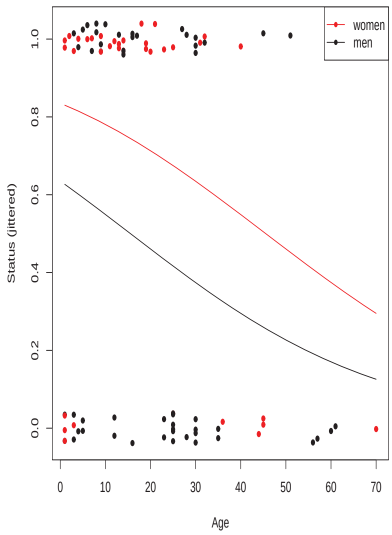
Donner party data

- Estimated model

$$\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035age + 1.067fem$$

- Gender: The survival odds for a woman are 3 times larger than the survival odds for a man of the same age: $\hat{OR} = e^{1.067} \approx 3$
- Age: A 10 year increase in age is associated with a 30% decrease in the survival odds for both men and women ($\hat{OR} = e^{-0.0356 \cdot 10} = 0.70$)

Donner party data



$$\hat{P}(Y=1|age, fem) = \frac{e^{(0.55-0.04age+1.07fem)}}{1+e^{(0.55-0.04age+1.07fem)}}$$

Donner party data: Model fit

- Model fit

Parameter	Estimate	Std. Error	z value	p-value
(Intercept)	0.553	0.417	1.325	0.1850
Age	-0.035	0.015	-2.336	0.0195
fem	1.067	0.482	2.214	0.0268

- 95% confidence interval for effect: Age, β_1

$$\left[\hat{\beta}_1 - 1.96 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \cdot SE(\hat{\beta}_1)\right]$$

- 95% BI for β_1: $-0.035 \pm 1.96 \cdot 0.015 \Rightarrow (-0.067, -0.006)$

Donner party data: Model fit

- Model fit

Parameter	Estimate	Std. Error	z value	p-value
(Intercept)	0.553	0.417	1.325	0.1850
Age	-0.035	0.015	-2.336	0.0195
fem	1.067	0.482	2.214	0.0268

- 95% confidence interval for odds ratio: Age, $OR = e^{\beta_1}$

$$\left[\exp\left(\hat{\beta}_1 - 1.96 \cdot SE(\hat{\beta}_1)\right), \exp\left(\hat{\beta}_1 + 1.96 \cdot SE(\hat{\beta}_1)\right)\right]$$

- 95% BI for $OR = e^{\beta_1}$: $(e^{-0.067}, e^{-0.006}) = (0.934, 0.993)$

Estimating odds ratios in R

```
> ## Odds ratios
>
> exp(donner.log$coefficients)
>
(Intercept)         Age         fem
  1.7398953   0.9650211   2.9094868

> exp(confint(donner.log))
>
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.7748972 4.0431170
Age         0.9348223 0.9930661
fem         1.1543365 7.7529827

> exp(cbind(OR =donner.log$coefficients, confint(donner.log)))
>
Waiting for profiling to be done...
                   OR     2.5 %    97.5 %
(Intercept) 1.7398953 0.7748972 4.0431170
Age         0.9650211 0.9348223 0.9930661
fem         2.9094868 1.1543365 7.7529827
>
```

Estimating odds ratios in R

```
> ## Odd ratio for Survival after 10 years increased
>
> exp(donner.log$coefficients*10)
 (Intercept)          Age          fem
2.542325e+02 7.004356e-01 4.346742e+04

> exp(c(OR =donner.log$coefficients[2]*10, confint(donner.log)[2,]*10))
>
Waiting for profiling to be done...
   OR.Age     2.5 %    97.5 %
0.7004356 0.5096720 0.9327850
>
```


Plotting the survival probabilities in R

```
## Plotting the logit curve
>
> logit<-function(x){log(x/(1-x))}
> ilogit<-function(x,a,b){exp(a+b*x)/(1+exp(a+b*x))}
>
> ## Plotting survival for men versus women
>
> cl=coef(donner.log)
> plot(donner.na$Age,jitter(donner.na$Outcome,.2),col=cols,pch=20,
+      cex=1.2,xlab="Age",ylab="Status (jittered)")
> curve(ilogit(cl[1]+cl[2]*x+cl[3]*0,0,1),add=T)
> curve(ilogit(cl[1]+cl[2]*x+cl[3]*1,0,1),add=T,col="red")
> legend("topright",pch=20,lty="solid",col=c("red","black"),c("women","men"))
>
```

Predicting the outcome

- One can use the model to predict the outcome of certain groups of interest
- For instance, in the Donner party study one may want to predict
  - The survival probability of a man with an average age (20.22 years)
  - The survival probability of a woman with an average age (20.22 years)

Predicting the outcome

Donner party example:

$$\text{logit}\left[\hat{P}(Y=1|age, fem)\right] = 0.553 - 0.035 age + 1.067 fem$$

with $fem = 1$ for women and $fem = 0$ for men

a) Survival probability for a man with the average age 20.22 ($fem = 0$)

$$\hat{P}(Y=1|20.22, man) = \frac{e^{(0.553-0.035\cdot 20.22+1.067\cdot 0)}}{1+e^{(0.553-0.035\cdot 20.22+1.067\cdot 0)}} = \frac{e^{-0.1547}}{1+e^{-0.1547}}$$

$$= \frac{0.8566}{1.8566} = 0.4614$$

Predicting the outcome

Donner party example:

$$\text{logit}\left[\hat{P}(Y=1|age, fem)\right] = 0.553 - 0.035 age + 1.067 fem$$

with $fem = 1$ for women and $fem = 0$ for men

b) Survival probability for a woman with the average age 20.22 ($fem = 1$)

$$\hat{P}(Y=1|20.22, woman) = \frac{e^{(0.553-0.035\cdot 20.22+1.067)}}{1+e^{(0.553-0.035\cdot 20.22+1.067)}}$$

$$= \frac{e^{0.9123}}{1+e^{0.9123}} = \frac{2.49}{3.49} = 0.7134$$

Predicting the outcome in R

```
> ## Predicted probabilities of survival
>
> newdata2<-data.frame(fem=1, Age=mean(donner.na$Age))
> newdata2$greP<-predict(donner.log,newdata=newdata2,type="response")
> newdata2
>
  fem      Age     greP
1   1 20.22727 0.711279
>
> newdata3<-data.frame(fem=0, Age=mean(donner.na$Age))
> newdata3$greP<-predict(donner.log,newdata=newdata3,type="response")
> newdata3
>
  fem      Age      greP
1   0 20.22727 0.4585025
>
> newdata4<-data.frame(fem=c(0,1),Age=mean(donner.na$Age))
> newdata4$greP<-predict(donner.log,newdata=newdata4,type="response")
> newdata4
>
  fem      Age      greP
1   0 20.22727 0.4585025
2   1 20.22727 0.7112790
>
```

Model building and model selection

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
>
> George E. P. Box

- Until now we have pretended that the relevant covariates and the structure of the model are both known
- In reality the situation is often more complex: frequently neither the relevant covariates nor the structure of the model are known before hand

Model building and model selection

- Thus, in practice scientists often consider several models (theories) to describe and explain reality
- For instance, in the Donner party example one may wonder if the effect of gender on survival varies across age or not, i.e., one may want to consider the model

$$
\begin{aligned}
\text{logit}\left[\hat{P}(Y=1|age, fem)\right] =& \beta_0 + \beta_1 age + \beta_2 fem + \beta_3 age \cdot fem \\
=& \beta_0 + \beta_1 age + (\beta_2 + \beta_3 \cdot age) fem
\end{aligned}
$$

Interaction model in R

```
> ## Interaction model
>
> m4<-glm(Outcome ~ Age*fem,data=donner.na,family=binomial(link="logit"))
> summary(m4)

Call:
glm(formula = Outcome ~ Age * fem, family = binomial(link = "logit"),
    data = donner.na)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9888  -1.0532   0.5961   1.0727   1.6317

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.39779    0.48139   0.826    0.409
Age         -0.02789    0.01911  -1.460    0.144
fem          1.47859    0.82469   1.793    0.073
Age:fem     -0.01977    0.03166  -0.624    0.532
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 108.47  on 84  degrees of freedom
AIC: 116.47

Number of Fisher Scoring iterations: 4
>
```


Interaction model

$$\text{logit}\left[\hat{P}(Y=1|age, fem)\right] = 0.398 - 0.028 age + 1.478 fem - 0.020 age \cdot fem$$

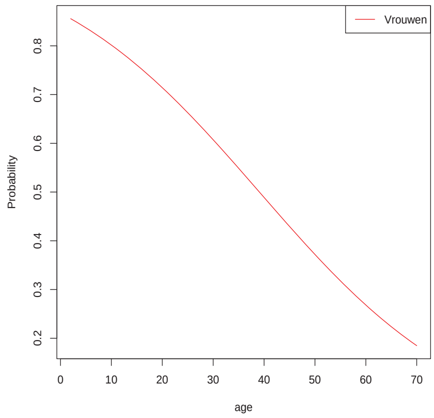
For women $fem = 1$

$$\begin{aligned}\text{logit}\left[\hat{P}(Y=1|age, women)\right] &= 0.398 - 0.028 age + 1.478 - 0.020 age \\ &= 1.876 - 0.048 age\end{aligned}$$

$$\hat{P}(Y=1|age, women) = \frac{e^{1.876-0.048 age}}{1+e^{1.876-0.048 age}}$$

Interaction model

For women $fem = 1$



$$\hat{P}(Y = 1|age) = \frac{e^{1.876-0.048age}}{1 + e^{1.876-0.048age}}$$

Interaction model

$$\text{logit}\left[\hat{P}(Y = 1|age, fem)\right] = 0.398 - 0.028age + 1.478fem - 0.020age \cdot fem$$

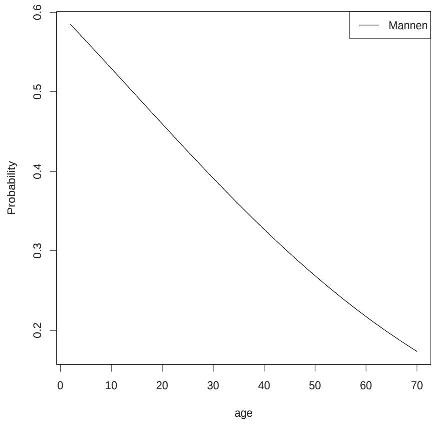
For men $fem = 0$

$$\text{logit}\left[\hat{P}(Y = 1|age, men)\right] = 0.398 - 0.028age$$

$$\hat{P}(Y = 1|age, men) = \frac{e^{0.398-0.028age}}{1 + e^{0.398-0.028age}}$$

Interaction model

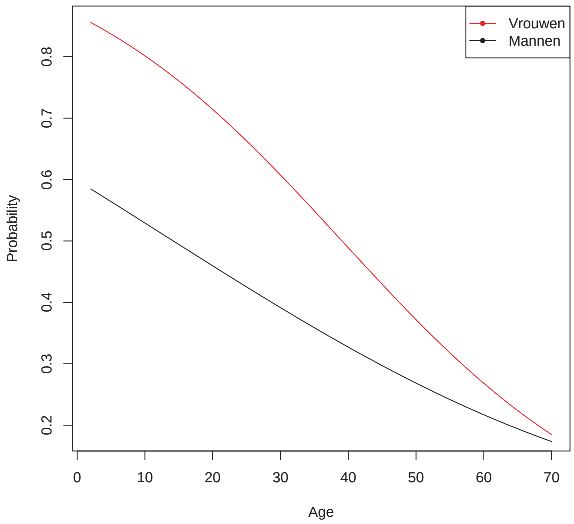
For men $fem = 0$



$$\hat{P}(Y = 1|age) = \frac{e^{0.398-0.028age}}{1 + e^{0.398-0.028age}}$$

Interaction model

Women versus men



Donner party data

- How do the odds of survival of a woman compare to those of a 10 years younger woman?
- For women: logit $\left[\hat{P}(Y=1|age, women)\right] = 1.876 - 0.048 age$
- $OR(age+10, age) = \text{Exp}(\hat{\beta}_1 \cdot 10) = e^{-0.48} = 0.62$

Age=30

Age=20

$$\Theta_{survival|30,women} = 0.62 \cdot \Theta_{survival|20,women}$$

Donner party data

- How do the odds of survival of a woman compare to those of a 10 years younger woman?
- For women: logit $\left[\hat{P}(Y=1|age, women)\right] = 1.876 - 0.048 age$
- $OR(age+10, age) = \text{Exp}(\hat{\beta}_1 \cdot 10) = e^{-0.48} = 0.62$

Age=60

Age=50

$$\Theta_{survival|60,women} = 0.62 \cdot \Theta_{survival|50,women}$$

Donner party data

- How do the odds of survival of a woman compare to those of a 10 years younger woman?
- For women: logit $\left[\hat{P}(Y = 1|age, women)\right] = 1.876 - 0.048age$
- $OR(age + 10, age) = \text{Exp}(\hat{\beta}_1 \cdot 10) = e^{-0.48} = 0.62$

$$OR(age + 10, age) = \frac{Odds(survival, age + 10)}{Odds(survival, age)} = \text{Exp}(\hat{\beta}_1 \cdot 10) = e^{-0.48} = 0.62$$

Interpretation

A 10 years increase in age is associated with a 40% (30% in the model without interaction) decrease in the odds of survival

Where do the models come from?

- Sometimes a set of models is provided based on subject-matter theory, the so-called mechanistic models. One example is the the PK/PD models used in pharmacokinetic/pharmacodynamics
- In practice good theory is very rare. Most often some simple restrictions are placed on the behavior one expects to find, for example, linear models, factorial models with limited interactions, etc. These models are sometimes called empirical models
- Nowadays model classes are available that can approximate many data generating mechanism. Furthermore, the computational resources to fit such models are rapidly increasing
- Model building and model selection

Model building: General principles

- **Goal**: To find a model that fits the data reasonably well without unnecessary complexity
- Model building is art and science: There are no clear, defined and fixed rules that you can automatically follow, but just general principles

P1 Use your **previous** scientific knowledge

- What are the research questions?
- What does the theory say?
- Are there results from previous studies?
- What does the *common sense* suggest?

Model building: General principles

P2 Interactions between predictors in the model should be included based on theory and plausibility

- Usually it is not necessary to evaluate all possible interactions
- Interactions between more than two predictors: very sound theoretical considerations necessary to include them

P3 There is a preference for so-called *hierarchical models* (also known as the principle of marginality)

- If the model includes an interaction, the corresponding main effects should be included as well
- If the model includes a quadratic term (x^2), a linear term should be included as well
- An intercept should be always included

Model building: General principles

P4 There is a distinction between observational and experimental studies

- Experimental research: Often limited set of factors is examined
- Model construction often less important ("true" model may be almost completely determined by the design)

P5 Importance of **replication**: A single study is **not** conclusive evidence of existence of an effect

P6 Groups or sets of predictors may belong together and, hence, move together in and out of the model

- For instance, personality can be represented using five predictors, the five factors of the Big Five
- For instance, a categorical predictor with more than two categories is included in the model using a set of dummy variables. In the final model, these dummy variables may or may not be included together

Model building: General principles

P7 Be aware of the issues associated with automatic selection procedures (stepwise, forward, backward, etc.)

- Each test is conditional on the results of the previous tests
- Distribution of these conditional statistics not fully understood
- Problems with the frequentist interpretation of α
- Multiple comparison problem
- In which sense is the final model best or optimal?
- No measure of model uncertainty

P8 Construction of a model is an iterative and creative process

Model building: General principles

P9 Inference after model construction and model selection: There is debate over which approach is correct. Active research area

P10 The objective of a study may also be the **prediction** of the criterion

- For instance, researchers may want to use a predictor(s) X to predict an outcome(s) Y
- *Understanding* is less important and, therefore, other principles can play a role

Explanation vs Prediction

- Explanation is like doing scientific research.
- Prediction is like doing engineering development. All that matters is that it works. And if the aim is prediction, model choice should be based on the quality of the predictions
- Why select a model at all?
  - It does seem a widespread misconception that model selection is about choosing **the** best model
  - For explanation one should be open to the possibility of there may be several (roughly) equally good explanatory models
  - For prediction one may want to do model averaging rather than model selection (expert opinion analogy)

Donner party data

- Reasonable models/theories

1. $\text{logit}\,[P(Y=1|age)] = \beta_0 + \beta_1 age$
2. $\text{logit}\,[P(Y=1|fem)] = \beta_0 + \beta_2 fem$
3. $\text{logit}\,[P(Y=1|age, fem)] = \beta_0 + \beta_1 age + \beta_2 fem$
4. $\text{logit}\,[P(Y=1|age, fem)] = \beta_0 + \beta_1 age + \beta_2 fem + \beta_3 fem \cdot age$

Which model should we use?

Model selection

What are you looking for?

- Model selection: One wants, given the sample, to choose a model that can describe the underlying distribution of the data
- But one only has limited information, namely the sample, and therefore one can not determine with complete certainty the underlying data generating mechanism
- Thus one looks for the most "likely" model, given your sample

- Competing models can be formally compared via
  - Nested models: Wald test, LRT
  - Nested and non-nested models: AIC (Akaike Information Criterion), BIC (Bayesian *Information* Criterion)
- Keep research question in mind!

Information criteria

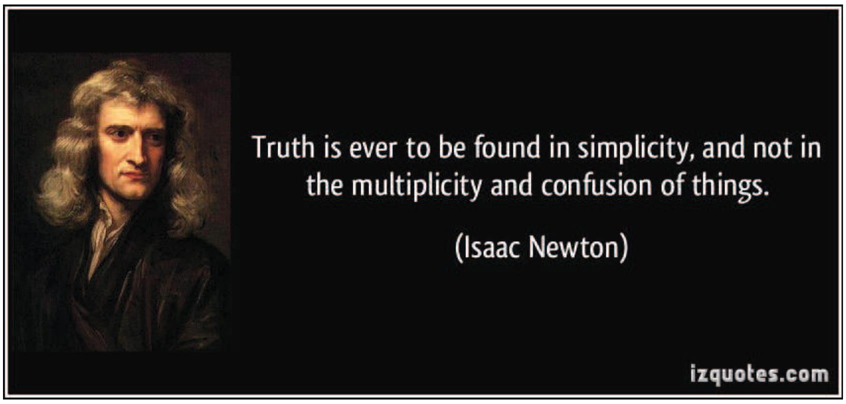
- AIC and BIC can be used to compare nested and non-nested models

$$
\begin{aligned}
\text{AIC} &= -2\log(\text{LMAX}) + 2(\#\text{ of parameters})\\
\text{BIC} &= -2\log(\text{LMAX}) + \log(n)(\#\text{ of parameters})
\end{aligned}
$$

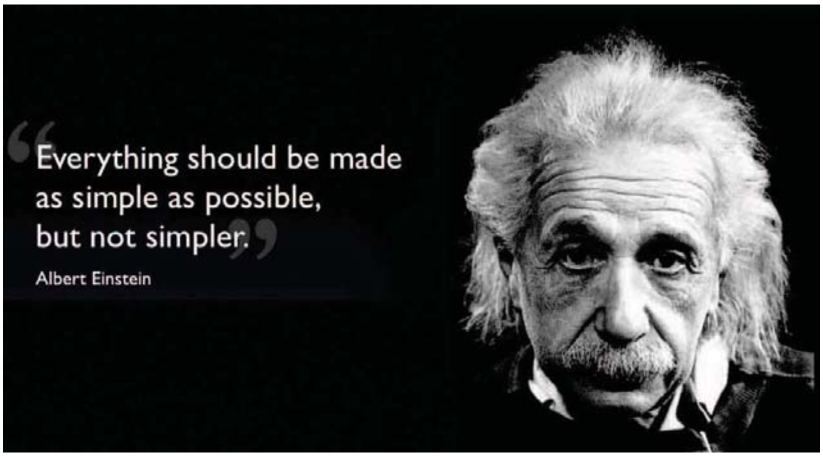
- *Penalty* for complexity, i.e., for the number of parameters used

- **Occam's razor**: Other things being equal, simpler explanations are generally better than more complex ones

Occam's razor



Occam's razor



Information criteria: AIC and BIC

- Smaller is better
- AIC and BIC are not equivalent. They have different characteristics and look for different models (different definitions of "best"). My personal choice: AIC
- AIC selects from a list of competing models the model that is "closest" to the underlying model
- "Closest" can be rigorously defined, namely, the model that minimizes the expected estimated kullback-Leibler divergence cross-entropy

Akaike information criterion: AIC

- A single AIC value is meaningless. AIC values are meaningful only when they are compared with other AIC values
- The Akaike-weights are easier to interpret: Posterior probability that the model is the "best" model in the kullback-Leibler sense
- Suppose one has a list of R competing models/theories then
  - Find the model with the smallest AIC_{min}
  - For every model i, compute $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$
  - For every model i, compute the Akaike–weights

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{i=1}^{R}\exp(-\frac{1}{2}\Delta_i)}$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	*age* · *fem*	116.47	1.60	0.251

$$w_T = \exp\left(-\frac{0.00}{2}\right) + \exp\left(-\frac{1.60}{2}\right) + \exp\left(-\frac{3.15}{2}\right) + \exp\left(-\frac{4.01}{2}\right) = 1.7909$$

$$w_1 = \frac{\exp\left(-\frac{0.00}{2}\right)}{w_T} = \frac{1}{1.7909} = 0.559$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	$age \cdot fem$	116.47	1.60	0.251

$$w_T = \exp\left(-\frac{0.00}{2}\right) + \exp\left(-\frac{1.60}{2}\right) + \exp\left(-\frac{3.15}{2}\right) + \exp\left(-\frac{4.01}{2}\right) = 1.7909$$

$$w_2 = \frac{\exp\left(-\frac{1.60}{2}\right)}{w_T} = \frac{0.4493}{1.7909} = 0.251$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	$age \cdot fem$	116.47	1.60	0.251

$$w_T = \exp\left(-\frac{0.00}{2}\right) + \exp\left(-\frac{1.60}{2}\right) + \exp\left(-\frac{3.15}{2}\right) + \exp\left(-\frac{4.01}{2}\right) = 1.7909$$

$$w_3 = \frac{\exp\left(-\frac{3.15}{2}\right)}{w_T} = \frac{0.2070}{1.7909} = 0.115$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	$age \cdot fem$	116.47	1.60	0.251

$$w_T = \exp\left(-\frac{0.00}{2}\right) + \exp\left(-\frac{1.60}{2}\right) + \exp\left(-\frac{3.15}{2}\right) + \exp\left(-\frac{4.01}{2}\right) = 1.7909$$

$$w_4 = \frac{\exp\left(-\frac{4.01}{2}\right)}{w_T} = \frac{0.1347}{1.7909} = 0.075$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	$age \cdot fem$	116.47	1.60	0.251

- Two models/theories, 1 and 2, seem to have some degree of support
- Some non-negligible level of model uncertainty
- A framework for scientific discussion: Which theory is more biologically plausible?

Akaike–weights in R

```
> ## Akaike weights
>
> ## Fitting the models
>
> m1<-glm(Outcome ~ Age,data=donner.na,family=binomial(link="logit"))
> m2<-glm(Outcome ~ fem,data=donner.na,family=binomial(link="logit"))
> m3<-glm(Outcome ~ Age + fem,data=donner.na,family=binomial(link="logit"))
> m4<-glm(Outcome ~ Age*fem,data=donner.na,family=binomial(link="logit"))
>
## Preparing a data set for the AIC
>
> aics<-data.frame(paste("m",1:4,sep=""),c(m1$aic,m2$aic,m3$aic,m4$aic),
+ row.names=NULL)
> colnames(aics)<-c("model","AIC")
> aics<-aics[order(aics$AIC),]
> for(i in 1:dim(aics)[1]){
> aics$diff[i]<-aics$AIC[i]-aics$AIC[1]}
> aics$wi<-exp(-0.5*aics$diff)
> aics$aic.weights<-aics$wi/sum(aics$wi)
> aics
>
```


Akaike–weights in R

```
> aics
>
  model      AIC     diff        wi aic.weights
3    m3 114.8674 0.000000 1.0000000  0.55860257
4    m4 116.4699 1.602437 0.4487818  0.25069065
1    m1 118.0183 3.150870 0.2069175  0.11558464
2    m2 118.8801 4.012646 0.1344823  0.07512213
>
```

Akaike–weights in R

```
> ## Akaike weights
> install.packages("qpcR")
> library(qpcR)
>
> ## Akaike weights with qpcR
>
> modList <- list(m1,m2,m3,m4)
> aics2 <- sapply(modList, function(x){AIC(x)})
> akaike.weights(aics2)
>
$deltaAIC
[1] 3.150870 4.012646 0.000000 1.602437

$rel.LL
[1] 0.2069175 0.1344823 1.0000000 0.4487818

$weights
[1] 0.11558464 0.07512213 0.55860257 0.25069065
>
```


Akaike–weights in R

```
> aics
>
  model      AIC     diff        wi aic.weights
3    m3 114.8674 0.000000 1.0000000  0.55860257
4    m4 116.4699 1.602437 0.4487818  0.25069065
1    m1 118.0183 3.150870 0.2069175  0.11558464
2    m2 118.8801 4.012646 0.1344823  0.07512213
>
```

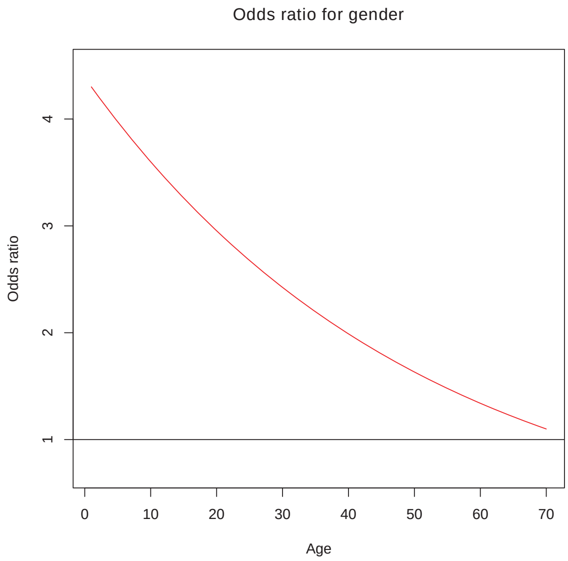
Donner party data: Conclusions

- **Main effect model**: Based on the data the estimated odds of survival for a woman is 3 ($\hat{OR} = e^{1067} \approx 3$) times larger than the corresponding odds of survival for a man of the same age, with a 95% CI for the odds ratio $(1.15, 7.75)$, approximately
- There is *model selection uncertainty*: **Interaction model** has a relatively large Akaike-weight
- However, both models indicate that the odds of survival are larger for women than for men

Odd ratio with the interaction model in R

```
> ## Gender odd ratio in the interaction model
>
> x=seq(1,70,0.01)
> y=exp(coef(m4)[3]+coef(m4)[4]*x)
>
> plot(x,y, type = "n", ylim=c(0.7, 4.5), xlab = "Age", ylab = "Odds ratio",
+ main="Odds ratio for gender")
> lines(x,y, lty = 1, col="red")
> abline(h=1)
>
```

Odd ratio in the interaction model



Donner party data: Conclusions

- It is important to note that this is an observational study (causal interpretations are therefore not justified)
- The sample was not drawn at random (inferences to a larger population are not strictly justified)