Methods of Applied Stats, Generalized Linear Models

Patrick Brown, University of Toronto and St Mike's Hospital

Sept to Dec 2020

Outline

- Examples
- Generalized Linear Models
- Likelihood-based inference
- Applications

Motivating example: Shuttle data

```
> data("shuttle", package = "SMPracticals")
> rownames(shuttle) = as.character(rownames(shuttle))
> shuttle[1:4, ]
```

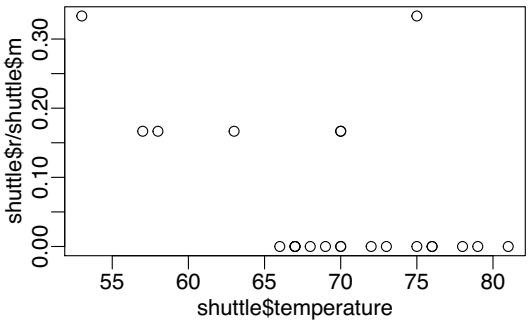
```
  m r temperature pressure
1 6 0          66       50
2 6 1          70       50
3 6 0          69       50
4 6 0          68       50
```

- m: number of rings
- r: number of damaged rings

Questions and models

- Are shuttle rings more likely to get damaged in cold weather?
- Pressure is a confounder
- The data aren't Gaussian
- Binomial distribution r failures from m trials
- Model failure probability as a function of temperature, pressure

```
> plot(shuttle$temperature,
+      shuttle$r/shuttle$m)
```



Motivating example: Fiji birth data

- Fiji Fertility Survey, 1974 opr.princeton.edu/archive/wfs/FJ.aspx
- pbrown.ca/teaching/appliedstats/data/fiji.RData created from pbrown.ca/teaching/appliedstats/data/fiji.R

```
> fiji[1:4, ]

  age ageMarried monthsSinceM failedPreg pregnancies children sons
1  25     18to20           72          0           0        0    0
2  31     15to18          184          0           6        6    2
3  40     15to18          269          0           2        2    1
4  46     15to18          206          0           9        9    6
  firstBirthInterval residence literacy ethnicity
1             60-Inf     rural      yes    fijian
2              12-23     rural      yes    fijian
3                0-7     rural      yes    fijian
4                lt0     rural      yes    fijian
```

Questions and models

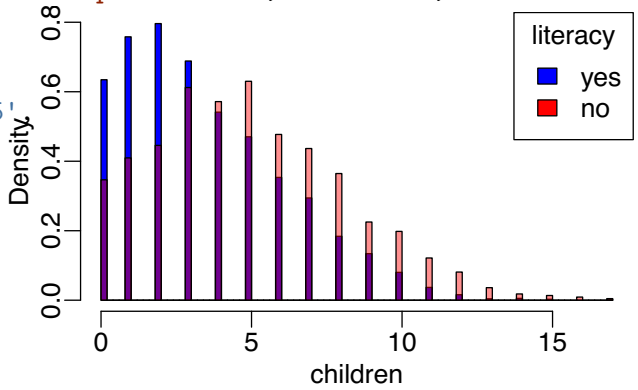
- Do literate women tend to have smaller families?
- Is this only because illiterate women marry earlier?

```
> table(fiji$ageMarried=='0to15',
+   fiji$literacy)
```

	yes	no
FALSE	3483	778
TRUE	323	333

- children $\sim$ Poisson

```
> literate = fiji$literacy == "yes"
> hist(fiji[literate, "children"], breaks = 1
+     prob = TRUE, main = "", xlab = "childre
```



Motivating example: Smoking

- 2014 American National Youth Tobacco Survey

```
> smoke[1:4, c("Age", "Sex", "Race", "RuralUrban", "state",
+   "ever_cigarettes")]
```

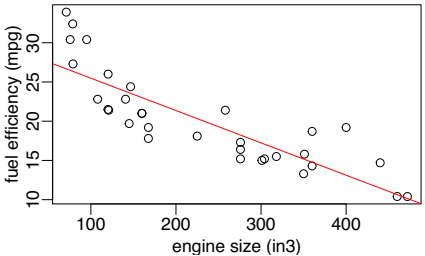
```
  Age Sex     Race RuralUrban state ever_cigarettes
1  13   M hispanic      Urban    AZ           FALSE
2  12   F hispanic      Urban    AZ           FALSE
3  14   M   native      Urban    AZ           FALSE
4  13   M hispanic      Urban    AZ           FALSE
```

```
> table(smoke$ever_cigarettes, smoke$Race)
```

```
       white black hispanic asian native pacific
FALSE   7576  2626     4543   844    250      56
TRUE    2243   731     1397   118     83      26
```

Ordinary Least Squares

Car data



- Y_i: fuel efficiency of car i
- $X_i = (X_{i1}, X_{i2})$: covariates
  - $X_{i1} = 1$, intercept
  - X_{i2} engine size

The wrong way

$$Y_i = X_i^\mathsf{T}\beta + Z_i$$
$$Z_i \sim \mathsf{N}(0, \sigma^2)$$

The right way

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = X_i^\mathsf{T}\beta$$

Parameters: β, σ^2.

OLS in R

```
> data("mtcars", package = "datasets")
> mtcars[1:3, c("mpg", "disp")]
                mpg disp
Mazda RX4      21.0  160
Mazda RX4 Wag  21.0  160
Datsun 710     22.8  108
> mtFit = lm(mpg ~ disp, data = mtcars)
> knitr::kable(summary(mtFit)$coef, digits = 3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	29.600	1.230	24.070	0
disp	-0.041	0.005	-8.747	0

$$Y_i \sim N(\beta_0 + X_i\beta_1, \sigma^2)$$

- `mpg` is the *response variable*
- `disp` is the *explanatory variable* or *covaraite*

Maximum Likelihood Estimation

- Model parameters: β, σ
- Likelihood function:

$$L(\beta, \sigma|Y) = \pi(Y; \beta, \sigma)$$

- MLE's:

$$(\hat{\beta}, \hat{\sigma}) = \text{argmax}_{\beta,\sigma} L(\beta, \sigma|Y)$$

- Here `myLik` calculates $-\log L(\beta, \sigma|Y)$
- and `optim` minimizes it
- `lm` knows the exact answer is

$$\hat{\beta} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} y$$

```
> myLik = function(param) {
+   mu = param[1] + param[2] * mtcars$disp
+   - sum(dnorm(mtcars$mpg, mean=mu,
+         sd=param[3], log=TRUE))}
> myLik(c(1,1,1))
[1] 974915.4
> optim(c(10,0.1,1), myLik,
+   control = list( parscale = c(1,0.01,1))
+   )$par
[1] 29.58930278 -0.04117094  3.14674642
> mtFit$coef
(Intercept)        disp
29.59985476 -0.04121512
```

Generalized Linear Models

$$Y_i \sim G(\mu_i, \theta)$$
$$h(\mu_i) = X_i^\mathsf{T}\beta$$

Linear predictor

- G is the distribution of the response variable
- μ_i is a location parameter for observation i
- θ are additional parameters for the density of G.
- h is a link function
- X_i are covariates for observation i
- β is a vector of regression coefficients

↳ Non-linear parameters

— Fixed effects.

- Ch 6 of Wakefield 2013
- Ch 4,5 of Davison 2003 http://books1.scholarsportal.info/viewdoc.html?id=/ebooks/ebooks1/cambridgeonline/2012-11-08/1/9780511815850
- Ch 6 of Faraway 2005 http://www.tandfebooks.com/isbn/9780203492284

Ordinary Least Squares again

$$
\begin{aligned}
Y_i &\sim G(\mu_i, \theta) \\
h(\mu_i) &= X_i^\mathsf{T} \beta
\end{aligned}
$$

$$
\begin{aligned}
Y_i &\sim N(\mu_i, \sigma^2) \\
\mu_i &= X_i^\mathsf{T} \beta
\end{aligned}
$$

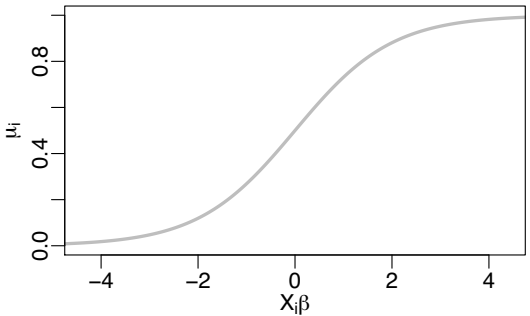
- G is a Normal distribution
- θ is the variance parameter, denoted σ^2
- h is the identity function

Binomial (or logistic) regression

$$Y_i \sim \text{Binomial}(N_i, \mu_i)$$

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = X_i^\intercal \beta$$

- G is a Binomial distribution
- ...or a Bernoulli if $N_i = 1$
- h is the logit link



- $X_i^\intercal \beta$ can be negative
- μ_i is between 0 and 1.

Are shuttle rings more likely to get damaged in cold weather?

	m	r	temperature	pressure
1	6	0	66	50
2	6	1	70	50
3	6	0	69	50
4	6	0	68	50

$$Y_i \sim \text{Binomial}(N_i, \mu_i)$$

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = X_i\beta$$

- m: number of rings, N_i
- r: number of damaged rings Y_i
- pressure, temperature: covariates X_i
- μ_i: probability of a ring becoming damaged given X_i
- $\beta_{\text{pressure}}, \beta_{\text{intercept}}$: confounders
- $\beta_{\text{temperature}}$: parameter of interest

Inference: parameter estimation

$$Y_i \overset{ind}{\sim} G(\mu_i, \theta)$$
$$h(\mu_i) = X_i\beta$$
$$\pi(Y_1 \dots Y_N; \beta, \theta) = \prod_{i=1}^{N} f_G(Y_i; \mu_i, \theta)$$
$$\log L(\beta, \theta; y_1 \dots y_N) = \sum_{i=1}^{N} \log f_G(y_i; \mu_i, \theta)$$

- The Y_i are *independently distributed*
- **Joint density** π of random variables $(Y_1 \dots Y_N)$ is the product of the marginal densities f_G
- **Likelihood function** L given observed data $y_1 \dots y_N$ is a function of the parameters
- **Maximum Likelihood Estimation**:

$$\hat{\beta}, \hat{\theta} = \text{argmax}_{\beta,\theta} L(\beta, \theta; y_1 \dots y_N)$$
$$= \underset{\beta,\theta}{\text{argmin}} - \ln L(\beta, \theta; y_1, \dots, y_N)$$

- The best parameters are those which are most likely to produce the observed data

Shuttle example in R

- `glm` works like `lm` with a `family` argument
- Binomial models in GLM require `y` to be a matrix with two columns
- ... `y` and `N-y`, which is unfortunate

```
> shuttle$notDamaged = shuttle$m - shuttle$r
> shuttle$y = as.matrix(shuttle[,c('r','notDamaged')])
> shuttleFit = glm(y ~ temperature + pressure,
+   family=binomial(link='logit'), data=shuttle)
> shuttleFit$coef

 (Intercept)  temperature     pressure
2.520194641 -0.098296750  0.008484021
```

Shuttle example the hard way

```
> logLikShuttle = function(param) {
+   muLogit = param[1] + param[2] * shuttle$temperature +
+     param[3] * shuttle$pressure
+   mu = exp(muLogit)/(1+exp(muLogit))
+   - sum(dbinom(shuttle$r, size=shuttle$m, prob=mu, log=TRUE))}
> optim(c(2,0,0), logLikShuttle,
+   control = list( parscale = c(1,0.01,0.01)))$par

[1]  2.51894107 -0.09828029  0.00848237

> shuttleFit$coef

 (Intercept)  temperature     pressure
2.520194641 -0.098296750  0.008484021
```

Efficient maximization

- Iterative Reweighted Least Squares is the 'classic' algorithm when G is in the exponential family
- ... but GLM's are easy for any density which is differentiable
- The derivatives wrt β are easy to compute with the chain rule

$$\frac{\partial}{\partial \beta_p} \log L(\beta, \theta; y_1 \dots t_N) = \sum_{i=1}^{N} \left[\frac{d}{d\mu} \log f_G(Y_i; \mu, \theta) \right]_{\mu = h^{-1}(X_i^\intercal \beta)} \left[\frac{d}{d\eta} h^{-1}(\eta) \right]_{\eta = X_i^\intercal \beta} \cdot X_{ip}$$

- Analytical expressions exist for the derivatives of $\log f_G$ and h^{-1}
- Second derivatives are also tractable
- Numerical maximization to find $\widehat{\beta}$ is fast when derivatives are available

Numerical maximizers

- There are hundreds of them
- `optim` is the standard `R` optimizer, which has 6 methods available
  - some methods will use gradients if you provide them
- `TrustOptim` uses derivatives and 'trust regions', the method used in INLA
- `ipopt` is probably the cutting edge
- Statisticians don't make enough use of of-the-shelf optimizers

Automatic differentiation

$$\sum_{i=1}^{N} \left[\frac{d}{d\mu} \log f_G(Y_i; \mu, \theta) \right]_{\mu = h^{-1}(X_i^\top \beta)} \left[\frac{d}{d\eta} h^{-1}(\eta) \right]_{\eta = X_i^\top \beta} \cdot X_{ip}$$

- Overkill for most GLM's, but infinitely extensible
- computers evaluate logs, sines, and other functions through some Taylor-series like polynomial thing.
- ... which are easy to differentiate
- AD programs can take computer code and figure out how to differentiate it
- used in Neural Nets, Hamiltonian MCMC, optimization, and many more

```
> shuttleFitAD <- glmmTMB::glmmTMB(
+   y ~ temperature + pressure,
+   family=binomial(link='logit'),
+   data=shuttle)
> shuttleFitAD$fit$par
      beta       beta       beta
 2.520195003 -0.098296758  0.008484022
```

Taylor series expansion

$$\log L(\beta, \theta; \mathbf{y}) \approx \log L(\beta_0, \theta; \mathbf{y}) + (\beta - \beta_0)^{\mathsf{T}} \left[\frac{\partial}{\partial \beta} \log L(\beta_0, \theta; \mathbf{y}) \right] + \left(\frac{1}{2} \right) (\beta - \beta_0)^{\mathsf{T}} \left[\frac{\partial^2}{(\partial \beta)^2} \log L(\beta_0, \theta; \mathbf{y}) \right] (\beta - \beta_0)$$

- Set $\beta_0 = \hat{\beta}$ and the first derivative is zero

$$L(\beta, \theta; \mathbf{y}) \approx L(\hat{\beta}, \theta; \mathbf{y}) \exp \left\{ \left(\frac{1}{2} \right) (\hat{\beta} - \beta)^{\mathsf{T}} \left[\frac{\partial^2}{(\partial \beta)^2} \log L(\hat{\beta}, \theta; \mathbf{y}) \right] (\hat{\beta} - \beta) \right\}$$

- Looks like a Normal distribution $\hat{\beta} \sim \mathrm{N} \left[\beta, I(\hat{\beta})^{-1} \right]$
- Roots of diagonal elements of $I(\hat{\beta})^{-1}$ are standard errors

Inference

- Information Matrix:

$$I(\hat{\beta}|Y) = \frac{\partial}{\partial\beta\partial\beta^{\mathsf{T}}} - \log L(\beta|Y)\Big|_{\hat{\beta}}$$

- MLE's are approximately Normal

$$\hat{\beta} \sim \text{MVN}(\beta, I(\hat{\beta}|Y)^{-1})$$

- standard errors are roots of diagonals of inverted Information Matrix

```
> (infMat = numDeriv::hessian(
+   f=logLikShuttle, x=shuttleFit$coef))
             [,1]        [,2]        [,3]
[1,]     7.81222   504.4844   1415.968
[2,]   504.48442 33096.7647  90802.259
[3,] 1415.96790 90802.2589 274388.164
> sqrt(diag(solve(infMat)))
[1] 3.486838277 0.044890480 0.007677616
```

```
> knitr::kable(summary(shuttleFit)$coef, digits = 4)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5202	3.4868	0.7228	0.4698
temperature	-0.0983	0.0449	-2.1897	0.0285
pressure	0.0085	0.0077	1.1051	0.2691

```
> knitr::kable(summary(shuttleFitAD)$coef$cond, digits = 4)
```

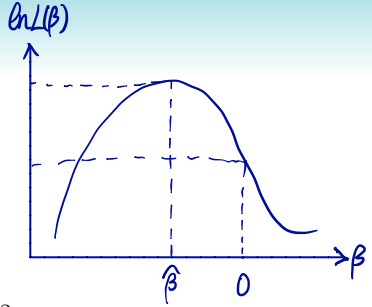
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5202	3.4753	0.7252	0.4684
temperature	-0.0983	0.0448	-2.1948	0.0282
pressure	0.0085	0.0077	1.1076	0.2681

```
> sqrt(diag(solve(infMat)))
[1] 3.486838277 0.044890480 0.007677616
```

Some notes

- Don't confuse
  - *Models*,
  - *Inference methodologies*, and
  - *Algorithms*.
- **Models**: Generalized linear models, ARIMA time series, population sampling
- **Inference methodologies**: Frequentist or Likelihood-based, Bayesian, Method-of-moments, partial likelihood.
- **Algorithms**: Least Squares, IRWLS, Markov Chain Monte Carlo, INLA, the Lasso

Likelihood ratio tests



$$2[\log L(\hat{\beta}; \mathbf{y}) - \log L(\beta; \mathbf{y})] \sim \chi^2_P$$

where P is the number of parameters in β

Note 1. $L(\hat{\beta}; \vec{y}) \geq L(\beta; \vec{y})$

2. $\hat{\beta} \approx \beta$, estimation is good $\Rightarrow \ln L(\hat{\beta}) \approx \ln L(\beta)$.

Comparing nested models

- H_0: $\beta_k = C_k$ for all $k \in \Omega$
  - $\Omega \subset \{1 \dots P\}$
- H_1: β unconstrained.
- **Nested**: H_0 is a special case of H_1
- Write $\hat{\beta}^{(C)}$ as the constrained MLE's under H_0

e.g., $\beta_1 = \beta_2 = 0$, β_0 unconstrained. $\Omega = \{1, 2\}$.

$\hat{\beta}^{(C)} = \underset{\beta_k = C_k,\ k \in \Omega}{\operatorname{argmax}} \ln L(\beta; \vec{y})$

$$2[\log L(\hat{\beta}; \mathbf{y}) - \log L(\hat{\beta}^{(C)}; \mathbf{y})] \sim \chi^2_{|\Omega|}$$

Note. $\ln L(\hat{\beta}) \geq \ln L(\hat{\beta}^{(C)})$

What about θ?

$$Y_i \sim G(\mu_i, \theta)$$
$$h(\mu_i) = X_i^\mathsf{T}\beta$$

- When G is Poisson or Binomial, θ isn't used
- When G is exponential family θ factors out
  - e.g. σ^2 for Gaussian models

- Suppose G is Weibull and θ is a shape parameter?
- Derivatives of $\log L$ wrt θ are still straightforward
  - but more complicated than wrt β
- The Taylor series expansion and information matrix are still valid
  - but not used much in practice
- Most software packages estimate a $\hat{\theta}$ with numerical optimization
  - and treat it as 'known' and fixed when making inference on β

What you need to know

- No closed form MLE's for GLM's
- Derivatives are easy so maximization is quick
- There are nice parameters and nasty parameters
  - Easy standard errors for $\widehat{\beta}$ based on Normal/2nd order Taylor approximation.
  - The θ parameters are non-linear and more challenging.

Interpreting logistic models

$$Y_i \sim \text{Binomial}(N_i, \mu_i)$$

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{p=1}^{P} X_{ip}\beta_p$$

$$\left(\frac{\mu_i}{1-\mu_i}\right) = \prod_{p=1}^{P} \exp(\beta_p)^{X_{ip}}$$

- μ_i is a probability
- $\log[\mu_i/(1-\mu_i)]$ is a log-odds
- $\mu_i/(1-\mu_i)$ is an odds
- If $\mu_i \approx 0$, then $\mu_i \approx \mu_i/(1-\mu_i)$

Suppose $X_{1p} = X_{2p}$ for all p except $X_{2q} = X_{1q} + 1$

One-unit more q, holding other variables constant.

$$\beta_q = \log\left(\frac{\mu_2}{1-\mu_2}\right) - \log\left(\frac{\mu_1}{1-\mu_1}\right)$$

$$\exp(\beta_q) = \left(\frac{\mu_2}{1-\mu_2}\right) \Big/ \left(\frac{\mu_1}{1-\mu_1}\right)$$

- β_q is the log-odds ratio
- $\exp(\beta_q)$ is the odds ratio
- $\exp(\text{intercept})$ is baseline odds, when $X_{i2} \ldots X_{iP} = 0$.

Centring parameters

```
> quantile(shuttle$temperature)
  0%  25%  50%  75% 100%
  53   67   70   75   81
> quantile(shuttle$pressure)
  0%  25%  50%  75% 100%
  50   75  200  200  200
```

- Currently the intercept is log-odds when temperature $= 0$ and pressure $= 0$
- centre the covariates so the intercept refers to
  - temperature $= 70$ (degrees Farenheit)
  - pressure $= 200$ units of something

```
> shuttle$temperatureC = shuttle$temperature - 70
> shuttle$pressureC = shuttle$pressure - 200
> shuttleFit2 = glm(y ~ temperatureC + pressureC, family = "binomial",
+   data = shuttle)
```

Pmisc

```
> install.packages("Pmisc", repos = "http://r-forge.r-project.org")
```

Shuttle odds parameters

Produces a matrix suitable for multiplying by results of summary or predict functions to give CIs of a desired quantile.

```
> (theCiMat = Pmisc::ciMat(0.95))
           est       2.5      97.5
Estimate     1  1.000000 1.000000
Std. Error   0 -1.959964 1.959964
> parTable = summary(
+   shuttleFit2)$coef[,
+   rownames(theCiMat)] %*% theCiMat
> rownames(parTable)[1]= "Baseline"
```

CI for $\exp(\beta)$

```
> knitr::kable(exp(parTable),
+   digits=3)
```

	est	2.5	97.5
Baseline	0.070	0.028	0.176
temperatureC	0.906	0.830	0.990
pressureC	1.009	0.993	1.024

Table 1: MLE's of baseline odds and odds ratios, with 95% confidence intervals.

$\frac{\frac{\mu_i}{1-\mu_i}}{1+\frac{\mu_i}{1-\mu_i}} = \mu_i$.

Interpreting shuttle parameters

	est	2.5	97.5
Baseline	0.070	0.028	0.176
temperatureC	0.906	0.830	0.990
pressureC	1.009	0.993	1.024

- The odds of a ring being damaged when temperature = 70 and pressure = 200 is 0.0697, which corresponds to a probability of

```
> signif(exp(parTable[1,'est']) / (1+exp(parTable[1,'est'])), 3)
[1] 0.0651
```

- Each degree increase in temperature (in Yankee units) decreases the odds of damage by (in percent)

```
> signif(100 * (1 - exp(parTable[2, "est"])), 3)
[1] 9.36
```

Inter-quartile range

```
> quantile(shuttle$temperature)
  0%  25%  50%  75% 100%
  53   67   70   75   81
> quantile(shuttle$pressure)
  0%  25%  50%  75% 100%
  50   75  200  200  200
```

- A one-unit change in pressure means less than a one-unit change in temperature
- **Inter-quartile range**: odds ratio between 75th and 25th percentiles of a variable

```
> shuttleIQR = apply(shuttle[,c('temperatureC','pressureC')], 2,
+         function(xx) diff(quantile(xx, probs=c(0.25, 0.75))) )
> shuttleIQR

temperatureC    pressureC
           8          125
```

Odds ratios for interquartile ranges

```
> parTableIQR = exp(
+   diag(c(1,shuttleIQR[c('temperatureC','pressureC')]) ) %*% parTable)
> parTableIQR[1,] = parTableIQR[1,]  / (1+parTableIQR[1,])
> rownames(parTableIQR) = gsub("C$", "", rownames(parTable))
> Pmisc::mdTable(parTableIQR, digits=3, caption = "MLEs of baseline probabi
```

Table 6: MLEs of baseline probabilities and odds ratios for interquartile ranges, with 95pct confidence intervals

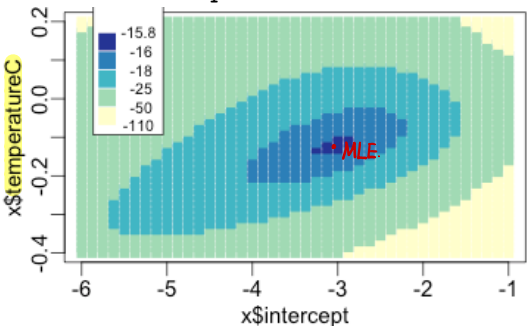
	est	2.5	97.5
Baseline	0.065	0.027	0.149
temperature	0.455	0.225	0.921
pressure	2.888	0.440	18.942

Note. When the temperature increases by 8 units (IQR), the odds are much smaller.

Plotting the likelihood function

```
> x = expand.grid(intercept = seq(-6, -1, len=41),
+   temperatureC = seq(-0.4, 0.2, len=41))
> x$lik = mapply(function(intercept, temperatureC) {
+   meanLogit = intercept + shuttle$temperatureC * temperatureC
+   meanProb = exp(meanLogit)/(1+exp(meanLogit))
+   sum(dbinom(shuttle$r, shuttle$m, meanProb, log=TRUE))
+ }, intercept = x$intercept, temperatureC = x$temperatureC)
```

```
> head(x)
  intercept temperatureC       lik
1    -6.000          -0.4 -28.65212
2    -5.875          -0.4 -28.55750
3    -5.750          -0.4 -28.53083
4    -5.625          -0.4 -28.57420
5    -5.500          -0.4 -28.68961
6    -5.375          -0.4 -28.87898
```

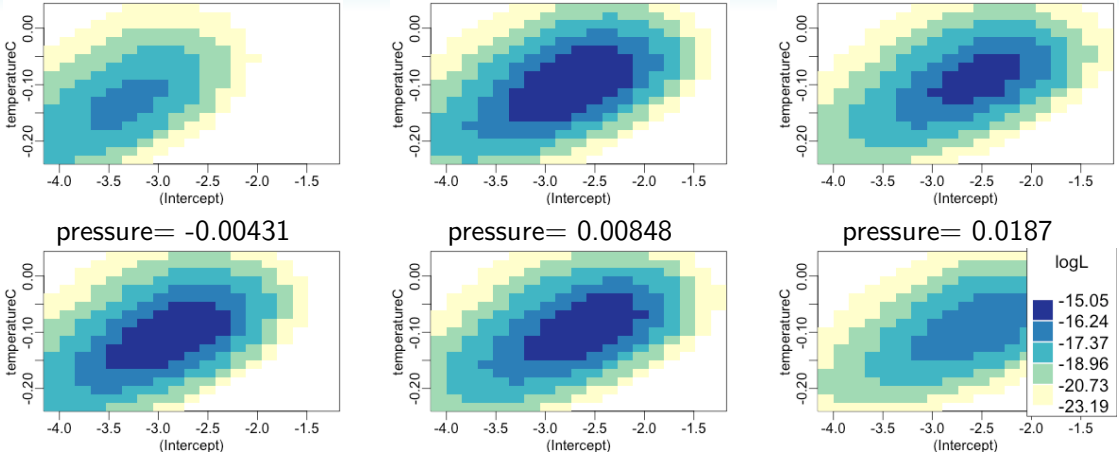


Likelihood function 3d

Evaluate the log-likelihood surface in the vicinity of the MLE

```
> shuttleL = Pmisc::likSurface(shuttleFit2)
>
> dim(shuttleL$logLik)

[1] 19 19 19

> head(shuttleL$parameters)

     (Intercept) temperatureC    pressureC
[1,]   -4.079471   -0.2329679 -0.014547274
[2,]   -3.922171   -0.2180044 -0.011988241
[3,]   -3.764872   -0.2030410 -0.009429208
[4,]   -3.607572   -0.1880775 -0.006870175
[5,]   -3.450272   -0.1731140 -0.004311143
[6,]   -3.292972   -0.1581506 -0.001752110
```

Shuttle likelihood function



pressure= -0.00431 | pressure= 0.00848 | pressure= 0.0187

pressure= 0.00337 | pressure= 0.0136 | pressure= 0.0238

Likelihood as a function of the intercept and temperature coefficient, for different values of the pressure coefficient.

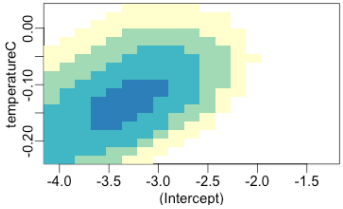
Shuttle LR test

are neither temperature nor pressure important?

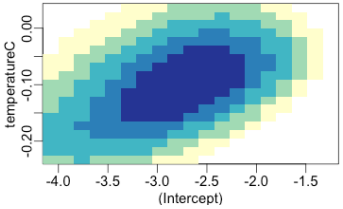
```
> shuttleFitNoCovariates = glm(y ~ 1, family = "binomial",
+   data = shuttle)
> lmtest::lrtest(shuttleFit, shuttleFitNoCovariates)
```

```
Likelihood ratio test

Model 1: y ~ temperature + pressure
Model 2: y ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -15.053
2   1 -18.895 -2 7.6847    0.02144 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

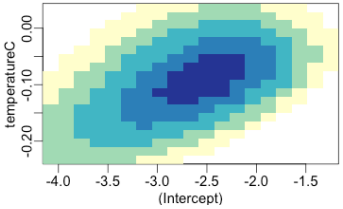
Shuttle likelihood function revisited



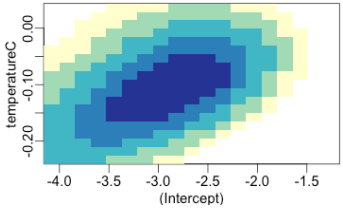
pressure= -0.00431



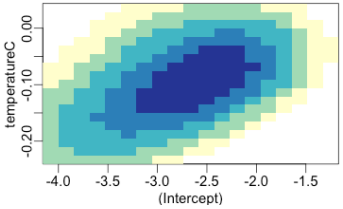
pressure= 0.00848



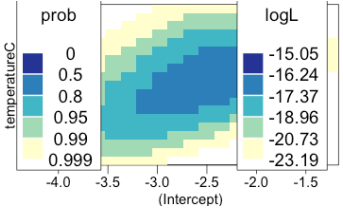
pressure= 0.0187



pressure= 0.00337



pressure= 0.0136



pressure= 0.0238

Colours are differences of χ^2_3 quantiles from the maximum (-15.0529).

Profile likelihood

- The 'pressure' covariate is a confounder (or a nuisance variable).
- It isn't significant, but we'll keep it in the model to be conservative
- Suppose we want to test $\beta_{\text{intercept}} = \text{logit}(0.1) \approx -2.2$ and $\beta_{\text{temperature}} = 0$?
- **Unconstrained model**: maximize $L(\beta_{\text{intercept}}, \beta_{\text{temperature}}, \beta_{\text{pressure}})$
- **Constrained model**: maximize $L(-2.2, 0, \beta_{\text{pressure}})$
- **Profile likelihood function** of the intercept and temperature:

$$\tilde{L}(a, b) = \max_{\beta_{\text{pressure}}} L(a, b, \beta_{\text{pressure}})$$

↳Not interested, but let it be the best one.

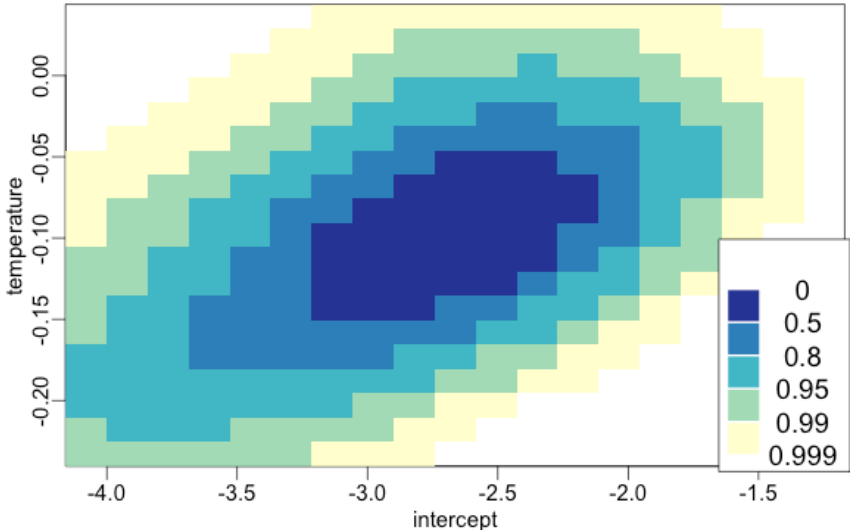
- More generally, $\theta = \{\theta_1 \dots \theta_P\}$ and $C \subset \{1 \dots P\}$ ↱ interested

$$\tilde{L}_C(\{\theta_q; q \in C\}) = \max_{\theta_r; r \notin C} L(\{\theta_q; q \in C\} \cup \{\theta_r; r \notin C\})$$

Profile likeihood for the shuttle data

```
> dim(shuttleL$logL)

[1] 19 19 19

> shuttleProfL = apply(shuttleL$logL, 1:2, max)
> dim(shuttleProfL)

[1] 19 19

> breaks2df = max(shuttleL$logL) - qchisq(shuttleL$breaks,
+   df = 2)/2
> image(shuttleL$parameters[, 1], shuttleL$parameters[, 2],
+   shuttleProfL, breaks = breaks2df, col = shuttleL$col,
+   xlab = "intercept", ylab = "temperature")
> mapmisc::legendBreaks("bottomright", shuttleL, inset = 0)
```

Profile likelihood for the shuttle data



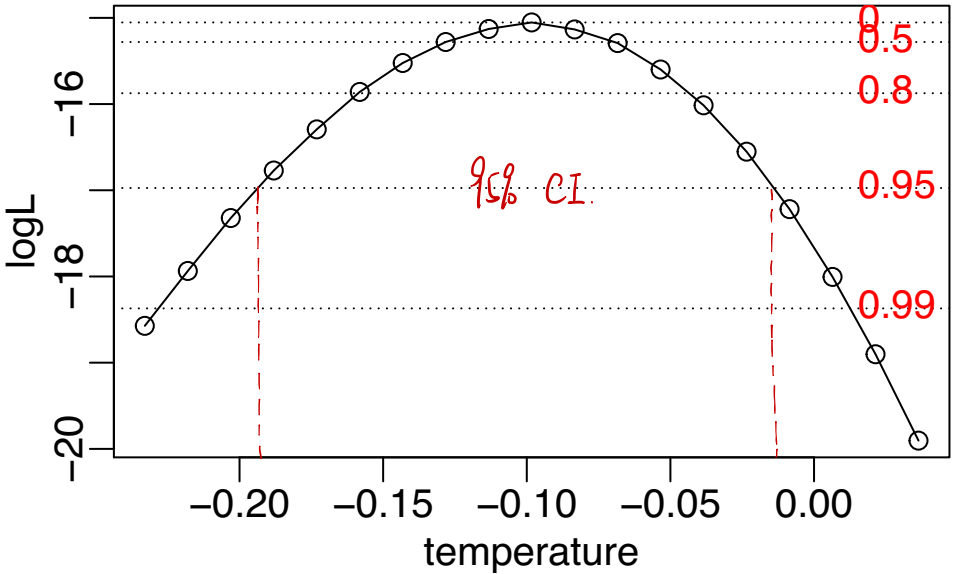
Profile likelihoods

- The `apply` function is used to find the maximum of the 19 likelihoods computed for each pair of intercept and temperature variables.
- Strictly speaking, a numerical optimizer should be used, varying the the pressure parameter continuously.
- The ξ^2 χ^2 quantiles are computed for 2 degrees of freedom (intercept and temperature parameters)
- A 1-D profile likelihood for the temperature parameter takes the maximum of the $19 \cdot 19$ likelihoods computed for each temperature parameter value.

1-D Profile likelihood

```
> dim(shuttleL$logL)

[1] 19 19 19

> colnames(shuttleL$parameters)

[1] "(Intercept)"  "temperatureC" "pressureC"

> shuttleProfTemp = apply(shuttleL$logL, 2, max)
> breaks1df = max(shuttleL$logL) - qchisq(shuttleL$breaks,
+   df = 1)/2
> plot(shuttleL$parameters[, 2], shuttleProfTemp, type = "o",
+   xlab = "temperature", ylab = "logL")
> abline(h = breaks1df, lty = 3)
> text(par("usr")[2] - 4 * par("cxy")[1], breaks1df, shuttleL$breaks,
+   pos = 4, col = "red")
```

1-D Profile likelihood



Information-based v profile likelihoods

- Profile likelihoods are not often used in Applied Statistics
- Is this for historical reasons, or are they not very useful?
- Information-based confidence regions are regarded as good enough.
- Joint confidence regions from 2-D profile likelihoods are interesting (?)

Where we are

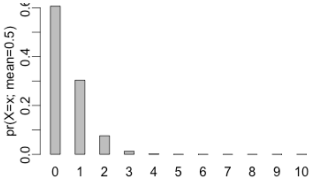
- discussed applied statistics
- reviewed GLM's
- profile likelihoods and LR tests
- saw exaples in R

Next

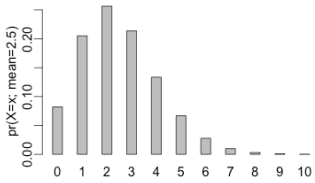
- GLM's with continuous variables
- conclusions
- practical in R

The Poisson Distribution

$$X_i \sim \text{Poisson}(\lambda_i)$$
$$pr(X_i = x) = \lambda_i^x \exp(-\lambda_i)/x!$$
$$X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$$



$\lambda = 0.5$



$\lambda = 2.5$

Poisson regression

$$Y_i \sim \text{Poisson}(\mu_i)$$
$$\log(\mu_i) = X_i\beta$$

- G is a Poisson distribution
- h is the log link

$$Y_i \sim G(\mu_i, \theta)$$
$$h(\mu_i) = X_i^\mathsf{T}\beta$$

Infinitely Divisible Laws

Normal : Continuous
Real

$$Y_i \sim \text{N}(\mu_i, \sigma^2)$$
$$Y_i = \sum_{k=1}^{K} X_{ik}$$
$$X_{ik} \sim \text{N}(\mu_i/K, \sigma^2/K)$$

Poisson: Discrete
Non-negative counts

$$Y_i \sim \text{Poisson}(\lambda_i)$$
$$Y_i = \sum_{k=1}^{K} X_{ik}$$
$$X_{ik} \sim \text{Poisson}(\lambda_i/K)$$

Gamma : Continuous
Positive

$$Y_i \sim \text{Gamma}(\alpha, \beta_i)$$
$$Y_i = \sum_{k=1}^{K} X_{ik}$$
$$X_{ik} \sim \text{Gamma}(\alpha/K, \beta_i)$$

- When our observed Y_i are actually the sum of many unobserved X_{ij} ...
  - Number of cigarettes smoked last month is sum of daily counts
  - Height of a child is sum of inches grown each month
- ... we're justified in considering only distributions which are infinitely divisible

Not infinitely divisible

Uniform

$$X_{ik} \sim \text{Unif}(a_i, b_i)$$

- $\sum_k X_{ij}$ isn't uniform.
- Can't construct a uniform Y_i from a sum.

Binomial

$$Y_i \sim \text{Binom}(\rho_i, N)$$

$$Y_i = \sum_{k=1}^{K} X_{ik}$$

$$X_{ij} \sim \text{Binom}(\rho_i, N/K)$$

- Not true for all K
- Only valid when N is a multiple of K

Log-Normal

$$X_{ik} \sim \text{LN}(\mu_i/K, \sigma/K)$$

$$\log(X_{ik}) \sim \text{N}(\mu_i/K, \sigma^2/K)$$

- $\sum_k X_{ik}$ isn't log-Normal
- but

$$\prod_k X_{ik} \sim \text{LN}(\mu_i, \sigma)$$

- a good model when Y_i is a share price?

Do literate Fijians tend to have fewer children?

$$Y_i \sim \text{Poisson}(O_i \mu_i) \qquad (1)$$
$$\log(\mu_i) = X_i \beta$$

or

$$Y_i \sim \text{Poisson}(\rho_i) \qquad (2)$$
$$\log(\rho_i) = X_i \beta + \log(O_i)$$

	monthsSinceM	literacy	children
1	72	yes	0
2	184	yes	6
3	269	yes	2
4	206	yes	9

- Y_i: number of children of women i
- μ_i: rate of children born, per month
- O_i: **Offset term**, number of months since first married
- X_i: intercept, literacy

- Write (1) when communicating with humans.
- Write (2) when communicating with computers.

$\rho_i = e^{X_i^T \beta} e^{\ln O_i} = O_i e^{X_i^T \beta} = O_i \mu_i$.

Poisson Processes in time

- Suppose P_i are a sequence of event times

$$\{P_1 \ldots P_N\} \sim \text{Poisson process}(\lambda)$$

Definition

$\{P_1 \ldots P_N\}$ is a homogeneous Poisson process with intensity λ if

1. $|\{P_i \in A\}| \sim \text{Poisson}(\lambda|A|)$ for any time interval A
2. Given $P_i, P_j \in A$, P_i and P_j are independent and uniformly distributed in A

A property

$|\{i; P_i \in A_k\}| \sim \text{Poisson}(\lambda|A_k|)$ for any non-overlapping regions A_1 and A_2.

Fijian births as a Poisson process

- μ_i is the intensity, per month, of children born from woman i
- $O_i = |A_i|$ where A_i is the time interval since first married to the date the survey was given
- $Y_i \sim \text{Poisson}(\mu_i O_i)$
- **Infinitely divisible**: divide A_i into trillions of one nanosecond long time intervals δ_{ik}
- X_{ik} is the number of children born to woman i in interval δ_{ik}

$$X_{ik} \sim \text{Poisson}\left(\mu_i \frac{1}{10^6 \cdot 60 \cdot 60 \cdot 24 \cdot 30}\right)$$

$$\sum_k X_{ik} \sim \text{Poisson}(\mu_i O_i)$$

Is this a good model

Yes

- The data look Poisson: non-negative counts, right skewed
- μ_i can be interpreted independently of O_i
- It's simple because it's a GLM
- other model could be constructued, but this one has a rigorous mathematical foundation
- A good balance betwen being right and being useful
- ... at least as a starting point

No

- Dividing our data into nanosecond-long intervals isn't necessary
- **Dependence**: births can't be closer than about 8 months apart
- **Overdispersion**: fertility and desired family size vary from person to person
- **Inhomogeneity**: fertility decreases with age

The log-log link function

- Suppose shuttle rings during launch i break following a Poisson process with intensity λ_i.
- During launch i, ring j has failure times P_{ijk} for $k = 1 \dots K_{ij}$
- $K_{ij} \sim$ Poisson$(\delta\lambda_i)$ where δ is the duration of a mission.
- We only observe $Z_{ij} = 0$ if $K_{ij} = 0$ and ~~$Z_{ij} = 0$~~ if $K_{ij} > 0$

$Z_{ij} = 1$

Note. $P(Z_{ij}=1) = 1 - P(K_{ij}=0)$

$= 1 - \exp(-\delta\lambda_i)$

$= 1 - \exp(-\delta\exp(X_i^T\beta))$

$$\{P_{ijk}; k = 1 \dots K_{ij}\} \sim \text{Poisson process}(\lambda_i)$$

$$\log(\lambda_i) = X_i\beta$$

$$pr(K_{ij} = 0) = \exp(-\delta\lambda_i) \leftarrow \text{Poisson}(0)$$

$$pr(Z_{ij} = 1) = \mu_i$$

$$\mu_i = 1 - \exp[-\delta\exp(X_i\beta)]$$

$$\log[-\log(1-\mu_i)] = \log(\delta) + X_i\beta$$

Notes on the log-log link

$$\log[-\log(1-\mu_i)] = \log(\delta) + X_i\beta$$

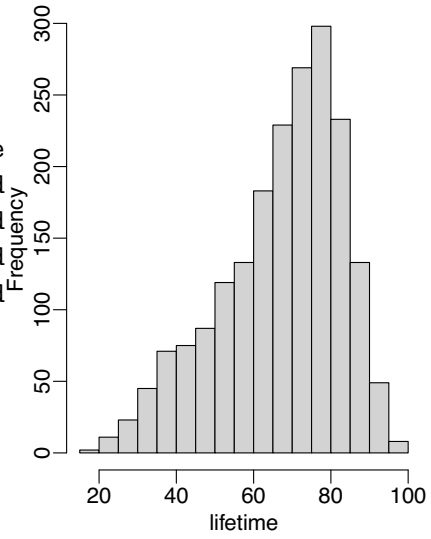
- If $Y_i = \sum_{j=1}^{N_i} Z_{ij}$, then Y_i is Binomial(N_i, μ_i).
- If we ignore δ, the intercept parameter becomes $\beta_0 + \log(\delta)$
- β parameters with log-log links are rate ratios rather than odds ratios
- The logit link function is the 'standard' approach. In practice, there would need to be a reason given when using a log-log link.
- It is rarely possible for the data to tell you which link function provides the better fit.

Cricket data

```
> data('cricketer', package='DAAG')
> cricketer[4000:4003,]
      left year life dead acd kia inbed cause
4073 right 1888   27    1   1   1     0   acd
4074 right 1890   86    1   0   0     1 inbed
4075 right 1896   90    1   0   0     1 inbed
4076 right 1916   72    1   0   0     1 inbed
```

Remove cricketers born after 1890 and those killed as soldiers.

```
> dat = cricketer[cricketer$year < 1890 &
+    cricketer$kia == 0,]
> hist(dat$life, xlab='lifetime', main='')
```



The problem

- Doug Altman and Martin Bland (2005). "Do the left-handed die young?" In: *Significance* 2.4, pp. 166–170. DOI: 10.1111/j.1740-9713.2005.00130.x
- Do left handed people live less long than right-handers?
- The hypothesis is yes, they die in accidents using right-handed equipment.
- This is really a 'survival analysis' problem requiring a proportional-hazards model, but we'll use GLM's for illustrative purposes.

Gamma distribution

$$X \sim \text{Gamma}(\phi, \nu)$$

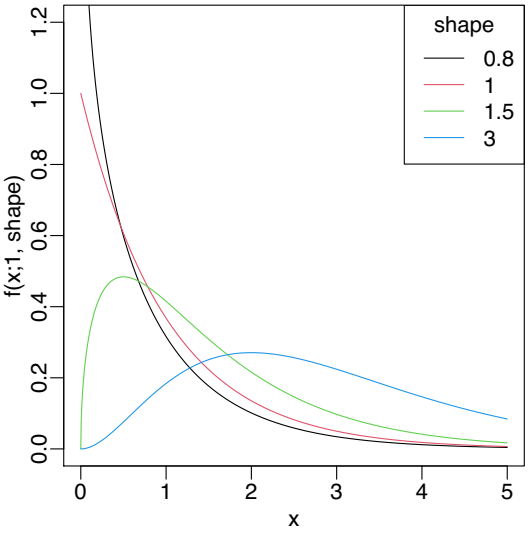
$$f(x; \phi, \nu) = \frac{(x/\phi)^{\nu-1} \exp(-x/\phi)}{\Gamma(\nu)\phi}$$

- ϕ is the range parameter
- ν is the shape parameter

$$\mathrm{E}(X) = \phi\nu, \ \mathrm{var}(X) = \phi^2\nu$$

- Coefficient of variation:

$$1/\sqrt{\nu} = \mathrm{sd}(X)/\mathrm{E}(X)$$



Gamma regression

$$Y_i \sim \text{Gamma}(\mu_i/\nu, \nu)$$
$$\log(\mu_i) = X_i^T\beta$$

- $\mathrm{E}(Y_i) = \mu_i$
- `glm` reports a 'dispersion' parameter $1/\nu$.
- and the log-link isn't the default

Note. Defaults link is inverse link. $\frac{1}{\mu_i} = X_i^T\beta$ (but cannot guarantee positivity)

```
> dat$decade = (dat$year - 1850)/10
> cFit = glm(life ~ decade + left, data=dat, family=Gamma(link='log'))
> knitr::kable(rbind(summary(cFit)$coef,
+   shape=c(1/summary(cFit)$dispersion, NA, NA, NA)), digits=4)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.1751	0.0085	490.3097	0.0000
decade	0.0242	0.0038	6.3081	0.0000
leftleft	-0.0162	0.0136	-1.1901	0.2341
shape	18.7245	NA	NA	NA

Is this a good model?

- Let's see if a histogram of the data looks Gamma distributed.
- Confine ourselves to right-handers born near 1850 to limit the effect of year and handedness on the distribution
  → similar μ_i.
- and overlay the density of the fitted Gamma distribution
- for individuals in the baseline group, lifetimes are Gamma distributed with the following parameters

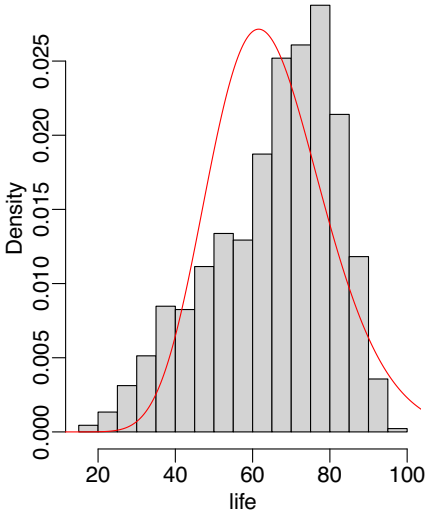
```
> shape = 1/summary(cFit)$dispersion
> scale = exp(cFit$coef["(Intercept)"])/shape
```

↓ near 1850.

Modelled and empirical distribution

```
> hist(dat$life[dat$left == "right" &
+     abs(dat$decade) < 2], prob = TRUE,
+     main = "", xlab = "life")
> xSeq = seq(0, 120, len = 1000)
> lines(xSeq, dgamma(xSeq, shape = shape,
+     scale = scale), col = "red")
```

- the `prob=TRUE` argument plots empirical densities instead of frequencies.
- `xSeq` is a vector of ages
- `dgamma` is the density of a Gamma distribution
- the Gamma hasn't capture the right-skewness.

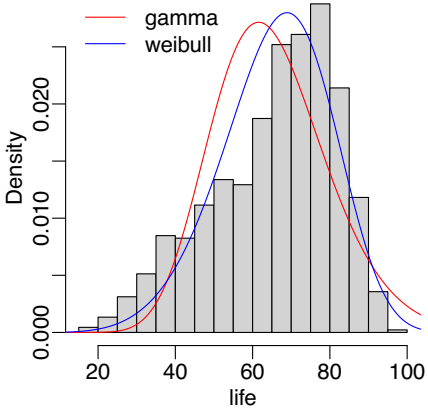


Weibull regression

Not exponential family.

```
> library('survival')
> cFitS = survreg(Surv(life) ~ decade + left,
+   data=dat, dist='weibull')
> knitr::kable(summary(cFitS)$table,digits=2)
```

	Value	Std. Error	z	p
(Intercept)	4.27	0.01	605.35	0.00
decade	0.02	0.00	5.20	0.00
leftleft	-0.01	0.01	-0.65	0.51
Log(scale)	-1.68	0.02	-90.20	0.00

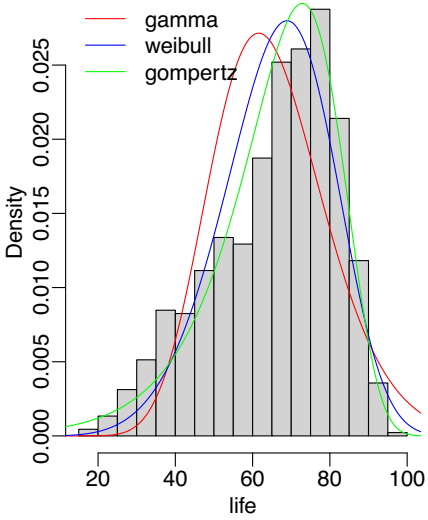


- Weibull is the 'standard' for event times
- survreg's scale is the Weibull shape parameter
- shape = 0.187, right skewed.

Gompertz regression

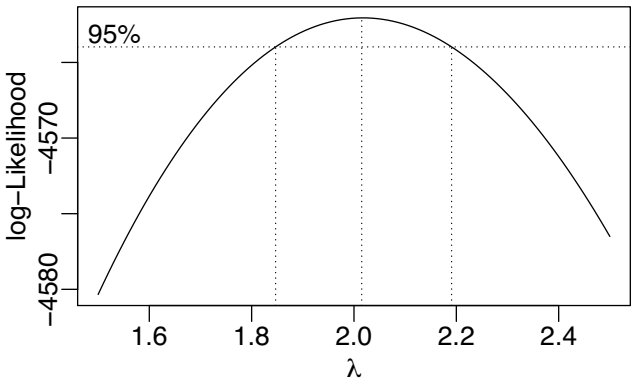
```
> library('flexsurv')
> cFitG = flexsurvreg(Surv(life)~decade+left,
+   data = dat, dist = 'gompertz')
> knitr::kable(cFitG$res, digits = 3)
```

	est	L95%	U95%	se
shape	0.079	0.076	0.082	0.001
rate	0.000	0.000	0.000	0.000
decade	-0.092	-0.125	-0.060	0.017
leftleft	0.033	-0.082	0.149	0.059

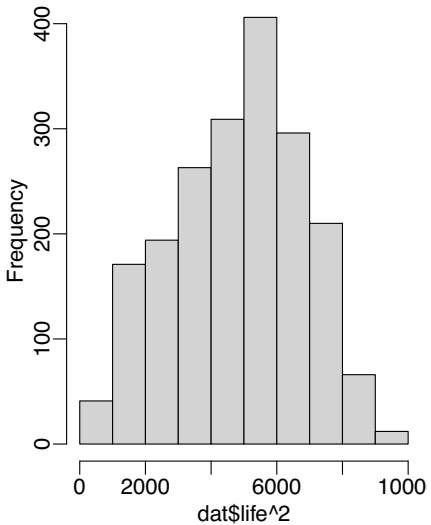


Box-Cox transformation

```
> library("MASS")
> boxcox(life ~ decade + left, data = dat,
+        lambda = seq(1.5, 2.5, len = 20))
```




```
> hist(dat$life^2, main = "")
```

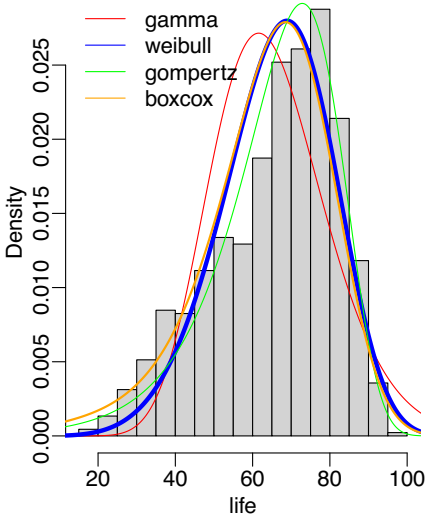


Ordinary Least Squares with transformed data

```
> dat$lifesq = dat$life^2
> fitBC = lm(lifesq~decade+left, data=dat)
> densBc = dnorm(xSeq^2,
+        mean = fitBC$coef['(Intercept)'],
+        sd = summary(fitBC)$sigma
+ ) / c(0,diff(sqrt(xSeq)))
> knitr::kable(summary(fitBC)$coef, digits=2)
```

Jacobian.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4488.34	72.20	62.17	0.00
decade	208.48	32.50	6.41	0.00
leftleft	-135.52	115.52	-1.17	0.24



Notes on GLM's with continuous data

- It's not easy to test which distribution is best, since they're not nested.
- GLM's are not often used with continuous data
  - they're almost always Binomial or Piosson
  - with the notable exception of the Weibull for event times
- 'standard practice' is to transform continuous data to normality (logs, Box-Cox)
- The squared transform model is simplest (?)
- The Weibull is easier to interpret. The percentage change in expected lifetime for left-handers (with 95% confidence interval) is

```
> 100*(exp(cFitS$coef['leftleft'] +
+     c(0, -2, 2) * summary(cFitS)$table['leftleft','Std. Error']) - 1)

[1] -0.7163102 -2.8803498  1.4959491
```

↳ Lifetime will go down by 0.72%.

Diagnostics for GLM's

- Assessing model fit is difficult for binary and count data!
- residuals don't have nice properties
- Histograms can be useful
- as can exploratory plots
- For the homework question on fruit flies, it will suffice to show that (or if) the Gamma is a good fit.

Where we are

- discussed applied statistics
- binary, Poisson and continuously-valued GLM's

Next

- contrasts

A more complex glm

interaction.

```
> cFitCause = glm(life~decade+left*cause,
+   data = dat, Gamma(link='log'))
> summary(cFitCause)$coef[,1:2]
                         Estimate  Std. Error
(Intercept)           3.81442882 0.057493774
decade                0.02379132 0.003789957
leftleft             -0.15479897 0.103518957
causeinbed            0.36433991 0.057388596
leftleft:causeinbed   0.14360209 0.104423075
> table(dat$cause)
alive   acd inbed
    0    23  1945
```

The cause and effect for left to right.

$\beta_0 + \beta_1$ decade $+ \beta_2$ left $+ \beta_3$ cause $+ \beta_4$ left · cause

$$Y_i \sim \text{Gamma}(\mu_i/\nu, \nu)$$
$$\log(\mu_i) = X_i\beta$$
$$\text{E}(Y_i) = \mu_i$$

- β_1 = time trend
- β_2 = left v right contrast
- β_3 = right (in bed v accident contrast)
- β_4 = contrast of left cause v right cause
- β_0 = intercept 1850, right, accidental death

Looking inside

```
> cFitCause$model[c(1,2,15,604),]
     life decade  left cause
2576   43   -0.4 right   acd
2577   47    3.7 right inbed
2606   68   -0.1  left inbed
3689   40    1.0  left   acd
> model.matrix(cFitCause$formula, cFitCause$model)[c(1,2,15,604),]
     (Intercept) decade leftleft causeinbed leftleft:causeinbed
2576           1   -0.4        0          0                   0
2577           1    3.7        0          1                   0
2606           1   -0.1        1          1                   1
3689           1    1.0        1          0                   0
```

On the natural scale

$$f(Y_i;\phi,\nu) = \frac{(x/\phi)x^{\nu-1}\exp(-x/\phi)}{\Gamma(\nu)\phi}$$

$$Y_i \sim \text{Gamma}(\mu_i/\nu,\nu)$$

$$\log(\mu_i) = X_i\beta$$

$\mu_i = E(Y_i)$: Expected life time.

	Estimate
(Intercept)	3.81442882
decade	0.02379132
leftleft	-0.15479897
causeinbed	0.36433991
leftleft:causeinbed	0.14360209

- when i is a lefty, j is a righty, $X_{ip} = X_{jp}$ for $p \neq 2$

$$\exp(\beta_2) = \mu_i/\mu_j$$

- k died in bed, ℓ accidental, both righties, same birth year

$$\exp(\beta_3) = \mu_k/\mu_\ell$$

- m, n are lefties died in bed and accidentally respectively.

$$\exp(\beta_4) = \frac{\mu_m/\mu_n}{\mu_k/\mu_\ell}$$

$\exp(\beta_3+\beta_4)$ → Effect for lefties

→ Effect for righties

- β_2 is the *contrast* between log-expected lifetimes of a lefty and righty

Notes

1. Baseline: $X_i = (1 \quad 0 \quad 0 \quad 0)^T$ Born in 1850, right handed, died in an accident.
   $\Rightarrow \mu_i = \exp(\beta_0)$

2. Suppose person 1 are 10 years older than person 2, holding others constant.
   $X_1 = (1 \quad ? \quad a \quad b)^T$, $X_2 = (1 \quad ?+1 \quad a \quad b)^T$
   $\Rightarrow \frac{\mu_2}{\mu_1} = \exp(\beta_1)$
   $\hat{\beta}_1 \approx 0.02$, $\exp(\hat{\beta}_1) \approx 1.02 \Rightarrow$ Every decade that you are born later, you live 2% longer.

3. $\exp(\hat{\beta}_2) \approx 0.86 \Rightarrow$ Left handed people live about 14% shorter lives than right handed people.

Suppose I want to report different contrasts?

```
                          Estimate
(Intercept)             3.81442882
decade                  0.02379132
leftleft               -0.15479897
causeinbed              0.36433991
leftleft:causeinbed     0.14360209
```

- Rate ratio for cause, with left-handers
- $\mu_m/\mu_n = \exp(\beta_3 + \beta_4)$
- What's the standard error for this thing?
- Use $\hat{\beta} \sim \mathrm{N}\left[\beta, I(\hat{\beta})^{-1}\right]$

$$\beta_3 + \beta_4 = A^T\beta$$

$$\beta_3 + \beta_4 \sim \mathrm{N}\left[A\beta, AI(\hat{\beta})^{-1}A^T\right]$$

$\beta_3 + \beta_4$

```
> Avec = c(0, 0, 0, 1, 1)
> c(est=crossprod(Avec, cFitCause$coef),
+   stderr=sqrt(crossprod(Avec, summary(cFitCause)$cov.scaled) %*% Avec))

       est     stderr
0.50794200 0.08722727
```

Conclusions

- GLM's are easy
- Before 1990 one could be forgiven for modelling r/m as Gaussian with the shuttle data
- ...now there's no excuse.

Exercise: Fiji data

- glm with 'family='poisson'
- create a variable 'logMonthsSinceM'
- put `offset(logMonthsSinceM)` in the model formula

References I

- Altman, Doug and Martin Bland (2005). "Do the left-handed die young?" In: *Significance* 2.4, pp. 166–170. DOI: 10.1111/j.1740-9713.2005.00130.x.
- Davison, A.C. (2003). *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. URL: http://books1.scholarsportal.info/viewdoc.html?id=/ebooks/ebooks1/cambridgeonline/2012-11-08/1/9780511815850.
- Faraway, J.J. (2005). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. URL: http://www.tandfebooks.com/isbn/9780203492284.
- Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics. Springer New York. DOI: 10.1007/978-1-4419-0925-1.