Outline

- The likelihood
- Maximization for univariate cases
- Regression
- Inference and profiling.
- Bayes connection
Estimating statistical parameters

Setup:

- A univariate probability density function, \( f_\theta(y) \)
  - depends on parameter \( \theta \)

- A sample of independent observations \( \{Y_1, Y_2, ..., Y_n\} \) from \( f_\theta(y) \)

Estimate \( \theta \) based on sample.

The plan:

- find joint density of observations
- substitute in observations to form likelihood
- maximize likelihood
- curvature of likelihood surface \( \equiv \) uncertainty of estimate
Example from exponential

\[ f_\theta(y) = \frac{1}{\theta} e^{-y/\theta} \]

Independent observations,

\[ g \text{ joint density function.} \]

\[ g_\theta(y_1, y_2, \ldots, y_n) = f_\theta(y_1) f_\theta(y_2) \ldots f_\theta(y_n) \]

\[ = \left(\frac{1}{\theta}\right) e^{-y_1/\theta} \left(\frac{1}{\theta}\right) e^{-y_2/\theta} \left(\frac{1}{\theta}\right) \ldots e^{-y_n/\theta} = \left(\frac{1}{\theta}\right)^n e^{-\sum y_i/\theta} \]

Will abuse notation: \( g_\theta(y) \) this is the probability of drawing \( y \) as a simple.
Likelihood: \( L(Y, \theta) = g_\theta(Y) \) Just substitute observations into joint density. The "likelihood" of observing this data given the parameter value \( \theta \).

**Maximize Likelihood over \( \theta \)**

Maximze

\[
L(Y, \theta) = \frac{1}{\theta^n} e^{-\left(\sum_i Y_i\right)/\theta}
\]

Maximizing likelihood is the same as maximizing log likelihood. In general the log likelihood is the natural scale to work on because it involves sums.

\[
\log L(Y, \theta) = -n \log(\theta) - \left(\sum_i Y_i\right)/\theta
\]

**Maximum Likelihood Estimator (MLE) for exponential**

\[
\hat{\theta} = \left(\sum_i Y_i\right)/n \equiv \overline{Y}
\]
How about normal?

\[ f_\theta(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \]

log likelihood for a random sample

\[
\log L(Y, \mu, \sigma) = -\log(\sqrt{2\pi\sigma}) - \sum_i \frac{(Y_i - \mu)^2}{2\sigma^2}
\]

\[ = -\log(\sqrt{2\pi}) - \log(\sigma) - \sum_i \frac{(Y_i - \mu)^2}{2\sigma^2} \]
Take partials w/r to $\mu$ and $\sigma$ set to set zero and solve

\[
\frac{\partial \log L(Y, \mu, \sigma)}{\partial \mu} = \sum_i (Y_i - \mu) \sigma^2 = 0
\]

\[
\frac{\partial \log L(Y, \mu, \sigma)}{\partial \mu} = -(1/\sigma) + \frac{\sum_i (Y_i - \mu)^2}{\sigma^3} = 0
\]

- From first equation $\hat{\mu} = \bar{Y}$ for all $\sigma$

- From second equation solution is simplified by solving for $\sigma^2$

\[
\hat{\sigma}^2 = (1/n) \sum_i (Y_i - \hat{\mu})^2
\]

**MLE for $\sigma$ is** $\sqrt{(1/n) \sum_i (Y_i - \hat{\mu})^2}$

**MLEs are transform invariant**
- the maximizer is always the maximizer!

i.e. MLE for $\omega(\theta)$ is just $\omega(\hat{\theta})$
Approximate inference

Key is to examine $2L(Y, \theta)$ as a function of parameter values and compare to maximum at MLE. Expect a quadratic surface about MLE.

**Normal example:** $(n = 75, \mu = 20, \sigma = 3)$, MLEs = (19.9, 3.05)

Contours are the difference between log likelihood and the maximum.
Approximate confidence sets are found by the contour that is offset by a chi square value from the maximum.

(1 − α)% *Confidence set:*

\[
\left\{ \theta : 2 \log L(Y, \hat{\theta}) - 2 \log L(Y, \theta) < \chi^2_p(\alpha) \right\}
\]

\( p = \text{number of parameters} \)

i.e. All parameters with lnLike large enough.

For normal example \( \chi^2_p(.05) = 5.99 \) (In R: \( \text{qchisq(.95, 2)} \)).

Grey contour (5.99/2) in previous figure.
Regression models

This is mainly to review using vector/matrix notation in likelihoods.

A linear model:

\[ y = Xd + e \]

\( \{e_1, e_2, \ldots, e_n\} \) are independent, normal mean 0 and variance \( \sigma^2 \). \( X \) a known matrix of covariates and \( d \) are the parameters.

\( y_i \) are normal mean \( [Xd]_i \) and variance \( \sigma^2 \)

\[ g(y) = f_1(y_1)f_2(y_2) \ldots f_n(y_n) \]

\[ = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_1-[Xd]_1)^2}{2\sigma^2}} \ldots \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_n-[Xd]_n)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi\sigma})^n} e^{-\frac{(y-Xd)^T(y-Xd)}{2\sigma^2}} \]

log Likelihood (least squares!):

\[ -(n/2) \log(2\pi) - n \log(\sigma) - \frac{(y-Xd)^T(y-Xd)}{2\sigma^2} \]
Partial derivatives set to zero:

\[ X^T(y -Xd) = 0 \]

\[-n/\sigma - (y -Xd)^T(y -Xd) \frac{2\sigma^2}{2\sigma^2} = 0 \]

**MLEs**

\[ \hat{d} = (X^TX)^{-1}X^Ty \]

\[ \hat{\sigma}^2 = (1/n)(y -Xd)^T(y -Xd) \]

Note: this will work for a nonlinear mean function depending on parameter vector \( d \) and will lead to minimizing the sums of squares but the MLEs may not have a closed form.

(This provides a framework for fitting variograms by nonlinear regression.)
Covariance estimation

Using a fields functions to fit a Matern covariance (smoothness=1.0) using maximum likelihood.

```r
fit3<- MLE.Matern( xHW1,yHW1, Z=elevHW1, smoothness=1.0, m=2,
                   theta.grid=seq(.2,4,.15))
plot( fit3$REML.grid[,1], fit3$REML.grid[,2])
yline( max( fit3$REML.grid[,2]) - qchisq(.95,1)/2)
```

**Profile likelihood:** For each value of $\theta$ maximize over the sill and nugget ($\rho$ and $\sigma$). This is a way to explore uncertainty in estimating a specific parameter. In this case the chi square degrees of freedom is just one (1), the total number of parameters (3) minus the number of parameters being maximized (2).