2009	2019
$Y = \beta X + \varepsilon$	$Y = \beta X + \varepsilon$
STATISTICS	MACHINE LEARNING

#10 YEARS CHALLENGE

Yesterday we…

- Constructed our first likelihood function
  - data, model, and parameters of interest

- Maximised likelihood functions by differentiation

- Maximised likelihood functions in R using `optim()` and `optimize()`

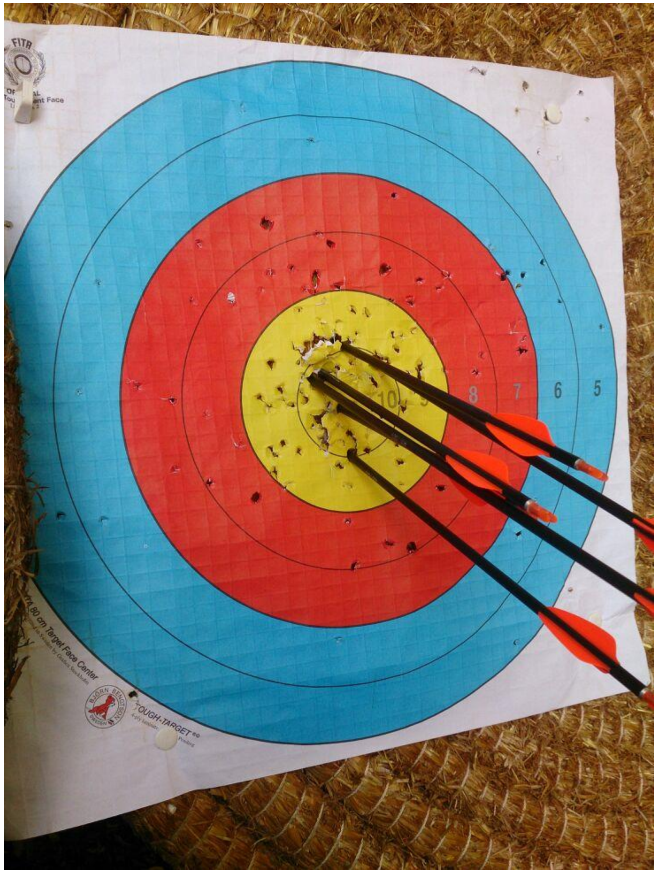
Day 3

- Properties of ML estimator
- Logistic regression example
- Likelihood-Ratio test

Properties of ML Estimator

- Asymptotically unbiased
  - on average we are hitting the target
  - $E\left[\hat{\theta}\right] \rightarrow \theta$ when $n \rightarrow \infty$

- Low variance (efficient)
  - better use of data
  - narrower confidence interval compared to other estimators



- Consistent: ML estimator converges in probability to the true parameter when $n \to \infty$

- Asymptotically normal
  - ML estimator is asymptotically distributed as normal with mean equals the true parameter value
  - remember Central Limit Theorem?
  - construction of confidence intervals (more on this later)

- Invariant principle
  - if $\hat{\theta}$ is the ML estimator for θ, then $g(\hat{\theta})$ is the ML estimator for $g(\theta)$

Example: Logistic regression

- Binary responses: dead or alive, yes or no, success or failure…
- Explanatory variable x is often called a risk factor (affect the risk/probability of "bad" outcome)
- Very common in public health/ medicine/ biology/ classification

#	State	Average cholesterol
1	Dead	5.0
2	Alive	4.4
3	Alive	3.4
4	Dead	3.7
5	Alive	3.6
6	Dead	4.7
…	…	…

- We need to find r.v. with binary outcomes to model the response variable y_i

- Bernoulli r.v.! Logistic regression assumes each response variable y_i follows a Bernoulli distribution

- Each individual will have its own p_i, which is a function of the risk factor x_i
  - x_i is the risk factor
  - $a + bx_i$ is the linear predictor

- $y_i \sim Bernoulli(p_i), where\ p_i = \eta^{-1}(a + bx_i)$, a and b are our parameters.

- What is η^{-1}?

- In logistic regression, $\eta^{-1}(a + bx_i) = \frac{e^{a+bx_i}}{1+e^{a+bx_i}}$
- η^{-1} is called "expit" transformation. The inverse of "logit" transformation
- $\eta^{-1}(a + bx_i)$ is bounded between 0 and 1 (remember, p_i is the probability of success), regardless of the values of $a + bx_i$

- Let us construct the likelihood function

- Two parameters: a and b

$$L(a,b) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} [p_i^{y_i}(1-p_i)^{1-y_i}]$$

$$= \prod_{i=1}^{n} [expit(a+bx_i)^{y_i}(1-expit(a+bx_i))^{1-y_i}]$$

- Take to log of the likelihood function

$$l(a,b) = \sum_{i=1}^{n} \{y_i \ln[expit(a+bx_i)] + (1-y_i)\ln[1-expit(a+bx_i)]\}$$

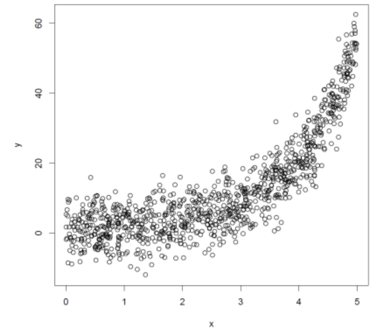
- It becomes a function of a and b only (with known y_i and x_i). We can maximise the log-likelihood function w.r.t. a and b.

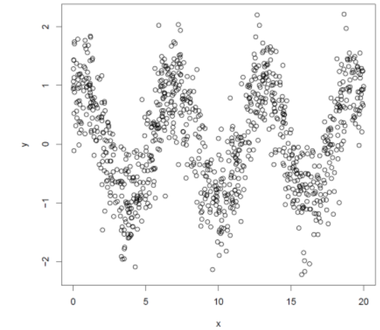
Non-standard regression

- Learning MLE means you can build your own statistical models

- Especially for non-standard cases where no "instant meals" are available

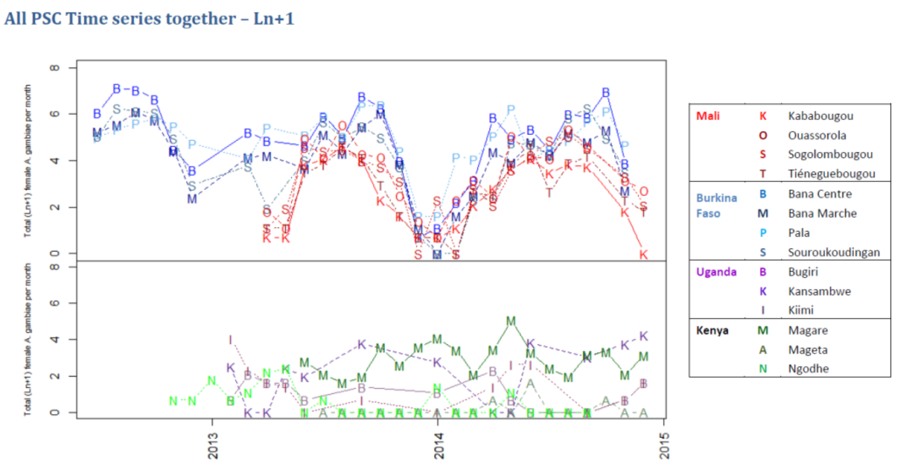
- $y_i = \exp(mx_i + b) + \epsilon_i$



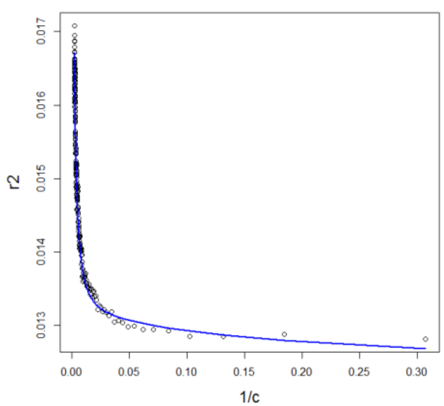


- Over-dispersed data

- Seasonal data



- State-space model



$$
\begin{array}{ccccccccccc}
p_0 & \rightarrow & p_1 & \rightarrow & p_2 & \rightarrow & \dots & \rightarrow & p_{t-1} & \rightarrow & p_t \\
\downarrow & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
x_0 & & x_1 & & x_2 & & \dots & & x_{t-1} & & x_t
\end{array}
$$

Likelihood-Ratio Test

- Hypothesis testing

- Let M1 and M2 be two models, and that M1 is nested in M2. If M2 has $d2$ parameters and M1 has $d1$ parameters $(d2 > d1)$, then $D = 2 * (\ln(L2) - \ln(L1))$ follows approximately a chi-square distribution with $(d2 - d1)$ degrees of freedom.
  - D is the LRT statistic

- The procedure is as follows:
  - Fit M1 to the data, record the maximised log-likelihood value $\ln(L1)$
  - Fit M2 to the data, record the maximised log-likelihood value $\ln(L2)$
  - Compute the likelihood-ratio statistic $D = 2 * (\ln(L2) - \ln(L1))$
  - Look up χ^2_{d2-d1} table for critical value. Accept M1 as the simplified model if D is smaller than the critical value

- Rationale:
  - The larger the log-likelihood value the better fit the model
  - M2 fits the data better with more parameters, thus yields a larger maximised log-likelihood value
  - M1 is the simplified model who has less explanatory power than M2 and therefore a smaller maximised log-likelihood value
  - D measures the difference in ‘explanatory power’
  - If the parameters dropped by M1 are unimportant, then there will only be a small decrease in explanatory power, hence a small D statistic
  - Dropping unimportant terms means we tend to accept M1 as the simplified model

Linear regression: test for intercept

- In yesterday's `recapture.csv`, we may think (biologically) that the intercept should be zero, because if a rabbit falls back to the trap "within zero days", then there should be no difference in its body length

- We let M1 be a linear regression model without an intercept i.e. $y_i = bx_i + \varepsilon_i$ (Two parameters)

- We let M2 be the full linear regression model we fitted yesterday i.e. $y_i = a + bx_i + \varepsilon_i$ (Three parameters)

- Clearly M1 is a special case of M2 with $a = 0$. We say M1 is nested in M2.

Log-likelihood function for M1

```
# THE LOG-LIKELIHOOD FUNCTION FOR M1 WITHOUT AN INTERCEPT
regression.no.intercept.log.likelihood<-function(parm, dat)
{
# DEFINE THE PARAMETERS
# NO INTERCEPT THIS TIME
?????
?????

# DEFINE THE DATA
# SAME AS BEFORE
x<-dat[,1]
y<-dat[,2]

# DEFINE THE ERROR TERM, NO INTERCEPT HERE
error.term<-?????

# REMEMBER THE NORMAL pdf?
density<-dnorm(error.term, mean=0, sd=sigma, log=T)

# LOG-LIKELIHOOD IS THE SUM OF DENSITIES
return(sum(density))
}
```

Log-likelihood function for M1

```
# THE LOG-LIKELIHOOD FUNCTION FOR M1 WITHOUT AN INTERCEPT
regression.no.intercept.log.likelihood<-function(parm, dat)
{
# DEFINE THE PARAMETERS
# NO INTERCEPT THIS TIME
b<-parm[1]
sigma<-parm[2]

# DEFINE THE DATA
# SAME AS BEFORE
x<-dat[,1]
y<-dat[,2]

# DEFINE THE ERROR TERM, NO INTERCEPT HERE
error.term<-(y-b*x)

# REMEMBER THE NORMAL pdf?
density<-dnorm(error.term, mean=0, sd=sigma, log=T)

# LOG-LIKELIHOOD IS THE SUM OF THE DENSITIES
return(sum(density))
}
```

Performing likelihood-ratio test

```
# PERFORMING LIKELIHOOD-RATIO TEST
M1<-optim(par=c(1,1), regression.no.intercept.log.likelihood,
          dat=recapture.data, method='L-BFGS-B',
          lower=c(-1000,0.0001), upper=c(1000,10000),
          control=list(fnscale=-1), hessian=T)
M2<-optim(par=c(1,1,1), regression.log.likelihood,
          dat=recapture.data, method='L-BFGS-B',
          lower=c(-1000,-1000,0.0001), upper=c(1000,1000,10000),
          control=list(fnscale=-1), hessian=T)

# THE TEST STATISTIC D
D<-2*(M2$value-M1$value)
D
```

```
[1] 3.047676
```

```
# CRITICAL VALUE
qchisq(0.95, df=1)
```

```
[1] 3.841459
```

We accept the hypothesis that the intercept is zero at $\alpha = 0.05$ (Same conclusion is drawn from `lm()` using anova table)

Model selection

- AIC is a tool to determine which of two models is better by weighting the improved fit of more complex models against their larger number of parameters.

- $AIC = -2l(\hat{\theta}) + 2K$, where $l(\hat{\theta})$ is the maximised log-likelihood and K is the number of parameters in the model

- Find the model with the lowest AIC value

Exercise: Non-constant variance regression

- In `recapture.csv`, we observe that the variance of the response is increasing with `day`. (Why?)
- Can we incorporate non-constant variance in our regression?
- Not sure about how we can do it with `lm`. Transformation of variables may help, but it is relatively simple MLE.
- How about $\varepsilon_i \sim N\left(0, x_i^2 \sigma^2\right)$? The standard deviation of the error terms increases linearly with the number of days?

Log-likelihood function: non-constant variance

```
# THE LOG-LIKELIHOOD FUNCTION FOR M1 WITHOUT AN INTERCEPT
regression.non.constant.var.log.likelihood<-function(parm, dat)
{
# DEFINE THE PARAMETERS
# NO CHANGE FROM M1
b<-parm[1]
sigma<-parm[2]

# DEFINE THE DATA
# SAME AS BEFORE
x<-dat[,1]
y<-dat[,2]

# DEFINE THE ERROR TERM, NO INTERCEPT HERE
error.term<-(y-b*x)

# REMEMBER THE NORMAL pdf
density<-dnorm(error.term, mean=0, sd=x*sigma, log=T)

# THE LOG-LIKELIHOOD IS THE SUM OF INDIVIDUAL DENSITIES
return(sum(density))
}
```

```
# MAXIMISE THE LOG-LIKELIHOOD
# HOW ABOUT CALLING IT M4?
M4<-optim(par=c(1,1), regression.non.constant.var.log.likelihood,
        dat=recapture.data, method='L-BFGS-B',
        lower=c(-1000,0.0001), upper=c(1000,10000),
        control=list(fnscale=-1))
M4
```

```
> M4
$par
[1] 3.483407 1.149874

$value
[1] -60.62583

$counts
function gradient
      25       25

$convergence
[1] 0

$message
[1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
```

This afternoon…

- Free ☺