Likelihood ratio tests

Math 483

1 LR tests for one parameter

Likelihood methods are useful for testing hypotheses, for example

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

The inference proceeds as follows:

(a) Compute the likelihood, according to a chosen model $f(x;\theta)$, with the value of $L(\theta_0)$.

(b) Compute the MLE $\hat{\theta}$, and its likelihood, $L(\hat{\theta})$. Find *likelihood ratio* $\boxed{\Lambda = L(\theta_0)/L(\hat{\theta})}$. Note that $\Lambda \leq 1$.

(c) Let $W \equiv -2\log\Lambda = 2(\log L(\hat{\theta}) - \log L(\theta_0))$. For large sample sizes, under null hypothesis H_0, W follows chi-square distribution with one degree of freedom.

Explanation: suppose we estimate the mean of a Normal distribution $\theta = \mu$, with known variance σ^2. The MLE is known to be $\overline{X}$. Then, from the Student's theorem,

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \tag{1}$$

is $\mathcal{N}(0,1)$. On the other hand,

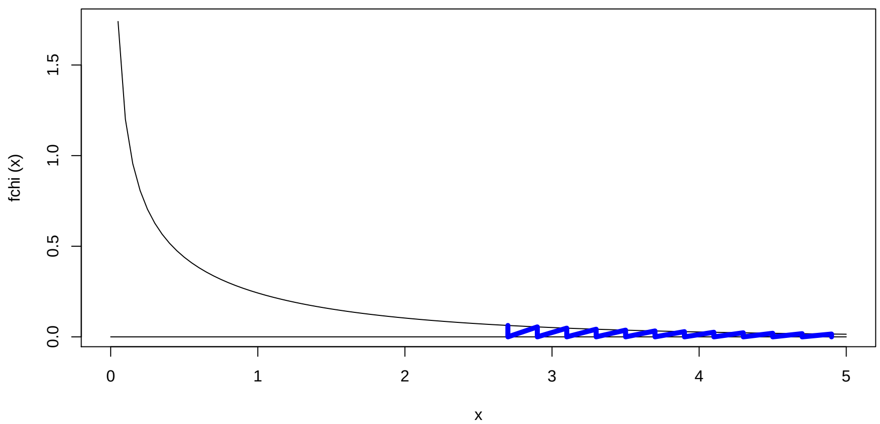
$$2(\log L(\overline{X}) - \log L(\mu)) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(X_i - \mu)^2 - \sum_{i=1}^{n}(X_i - \overline{X})^2\right].$$

After some algebra, it simplifies to $\frac{n}{\sigma^2}\left(\overline{X} - \mu\right)^2$. But this is exactly the square of (1), and is therefore $\chi^2(1)$.

In general, we can quote Theorem 6.3.1 on p.335 from the textbook.
The proof of Theorem 6.3.1 depends on the ability to approximate log likelihood with an upside-down parabola, and therefore on the finite value of Fisher information $I(\theta)$. Thus,

the regularity assumptions (R1)-(R4) are necessary.

(d) We will **reject** H_0 in favor of H_1, whenever $W > \chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}$ is the $1-\alpha$ quantile of chi-square distribution (R command `qchisq(1-α,1)`).
($\chi^2_{0.9} \approx 2.7$ is shown)



Example: Suppose $X_1, ..., X_n$ are i.i.d. from $Poisson(\theta)$. Find

$$\Lambda = \prod_{i=1}^{n} e^{-\theta_0+\hat{\theta}} \left(\frac{\theta_0}{\hat{\theta}}\right)^{X_i},$$

$$-2\log\Lambda = 2n(\theta_0 - \hat{\theta}) + 2(\log\hat{\theta} - \log\theta_0)\sum X_i \tag{2}$$

Recall that the MLE for Poisson is again $\overline{X}$. Thus, the expression (2) simplifies to

$$-2\log\Lambda = 2n[(\theta_0 - \overline{X}) + (\log\overline{X} - \log\theta_0)\overline{X}].$$

Example data: suppose that the average number of equipment failures in the past has been known to be $\theta_0 = 2.5$ failures per week. After introducing new equipment, we observed number of failures for 10 randomly chosen weeks and obtained the sample $0, 7, 5, 4, 3, 2, 0, 1, 3, 4$. Assume the Poisson model for number of failures. The sample mean is $\overline{X} = 2.9$ (later we'll see that $\overline{X}$ is a *sufficient statistic* for this model, that is, we only need to know $\overline{X}$ to draw our inferences).

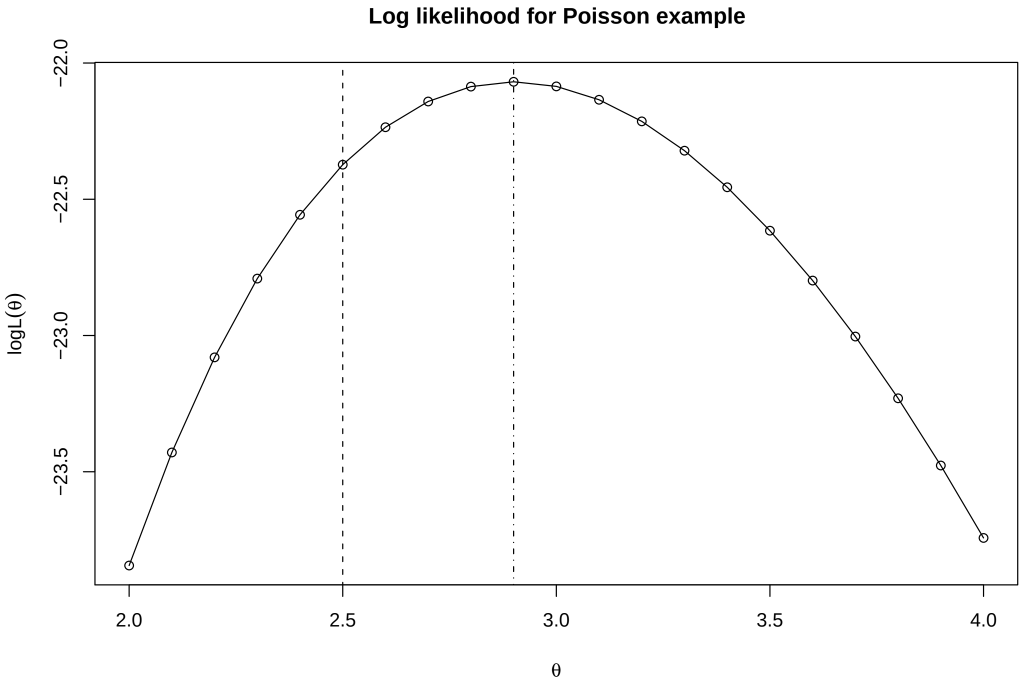
Set the hypotheses

$H_0 : \theta = 2.5$ "`average number of failures remained the same`"

$H_1 : \theta \neq 2.5$ "`average number of failures has changed`"

We then compute $W = -2\log\Lambda = 0.608$. If we used the critical region approach, we'd reject H_0 at the level $\alpha = 0.05$ whenever $W > \chi^2_{0.95}(1) = 3.84$. Thus, we accept H_0(no change).

If we use the p-value approach, we can find p-value$=P(\chi^2 > W) = 1 - \texttt{pchisq}(0.608, 1) = 0.4355$.



Looking at the plot, the difference in heights between the two values is not large enough to be significant.
Question: would we have rejected $H_0 : \theta = 2.0$ based on this data?

```
R code:

X <- c(0, 7, 5, 4, 3, 2, 0, 1, 3, 4)
lpois <- function(th){
     sum(log(dpois(X,th)))
}
xc <- seq(2,4, 0.1)
yc <- xc
nx <- length(xc)
for (i in 1:nx){
  yc[i] <- lpois(xc[i])
}
plot(xc,yc, type="o", main="Log likelihood for Poisson example", xlab=expression(theta),
   ylab = expression(logL(theta)))
 lines(c(2.9,2.9), c(-30,-20), lty = 4)
 lines(c(2.5,2.5), c(-30,-20), lty = 2)
```

2 Multi-dimensional case

Suppose now that the likelihood depends on p parameters, $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$. Let the null hypothesis be given in terms of q independent constraints:

$$H_0 : g_1(\boldsymbol{\theta}) = a_1, ..., g_q(\boldsymbol{\theta}) = a_q$$

$$H_1 : \text{not all} \quad g_j(\boldsymbol{\theta}) = a_j$$

As an example, consider a goodness-of-fit test (Section 5.7). Let $\theta_1, ..., \theta_p$ be proportions of observations in each of p categories. This will correspond to a classification table with $p+1$ cells, since θ_{p+1} is uniquely determined, $\theta_{p+1} = 1 - \sum_{j=1}^{p} \theta_j$.
Then, the null hypothesis that all of the proportions are equal can be expressed as

$$H_0 : \theta_1 = 1/(p+1), ..., \theta_p = 1/(p+1), \tag{3}$$

that is, for functions $g_j(\boldsymbol{\theta}) = \theta_j$.

Now, consider *constrained likelihood*, that is, find $\boldsymbol{\theta}^*$ that solves

$$\text{maximize} \quad l(\boldsymbol{\theta})$$

$$\text{subject to} \quad g_1(\boldsymbol{\theta}) = a_1, ..., g_q(\boldsymbol{\theta}) = a_q$$

The procedure of testing H_0 is then:

- Compute constrained $\boldsymbol{\theta}^*$ and unconstrained MLE $\hat{\boldsymbol{\theta}}$.
- Consider likelihood ratio $\Lambda = L(\boldsymbol{\theta}^*)/L(\hat{\boldsymbol{\theta}})$ (it's always ≤ 1).
- $W = -2\log\Lambda$ has, for large n, approximate chi-square distribution with q degrees of freedom.
- Reject H_0 at the level α whenever $W > \chi^2_{1-\alpha}(q)$. The p-value=$P(\chi^2 > W) = 1 - \texttt{pchisq}(\texttt{W}, \texttt{q})$.

Example: linear regression. Global warming. Suppose that our parameters are $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma^2)$ in the linear regression model with *quadratic* trend:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \quad i = 1, ..., n \tag{4}$$

where Y_i is the *global temperature anomaly* measured for year i, and $X_i = i$. The errors ε_i are assumed to be i.i.d. $\mathcal{N}(0, \sigma^2)$.
The model contains a quadratic term $\beta_2 X_i^2$, but it's still linear as a function of parameters $\beta_0, \beta_1, \beta_2$. The maximum likelihood (here also *least squares*) estimation is performed similarly to the linear trend (β_0, β_1 only) case.

(a) *Testing for trend.*
We can test for any trend in the data, that is,

$$H_0 : \beta_1 = 0, \quad \beta_2 = 0$$

For this, we should compare the unconstrained likelihood $L(\hat{\boldsymbol{\theta}})$ with constrained likelihood, for $\boldsymbol{\theta}^* = (\beta_0, 0, 0, \sigma^2)$. The latter is just the situation of estimating the mean of a Normal distribution, so that $\hat{\beta}_0 = \overline{Y}$, $\hat{\sigma^2} = \frac{\sum(Y_i - \overline{Y})^2}{n}$.

Unconstrained estimation is done using software. Log likelihood function for Normal is

$$l(\boldsymbol{\theta}) = -n \log \sigma - \frac{1}{2\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2$$

And $l(\hat{\boldsymbol{\theta}})$ simplifies to $-n \log \hat{\sigma} - n/2$, with

$$\hat{\sigma^2} = n^{-1} \sum (Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2.$$

Applied to our data, $W = -2 \log \Lambda = 2(l(\hat{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}^*)) = 2(53.39 - 22.49) = 61.8$. This is highly significant, because $\chi^2_{0.95} = 5.99 << 61.8$, and p-value $= 3.8e-14$ (we use chi-square with $q = 2$ degrees of freedom).

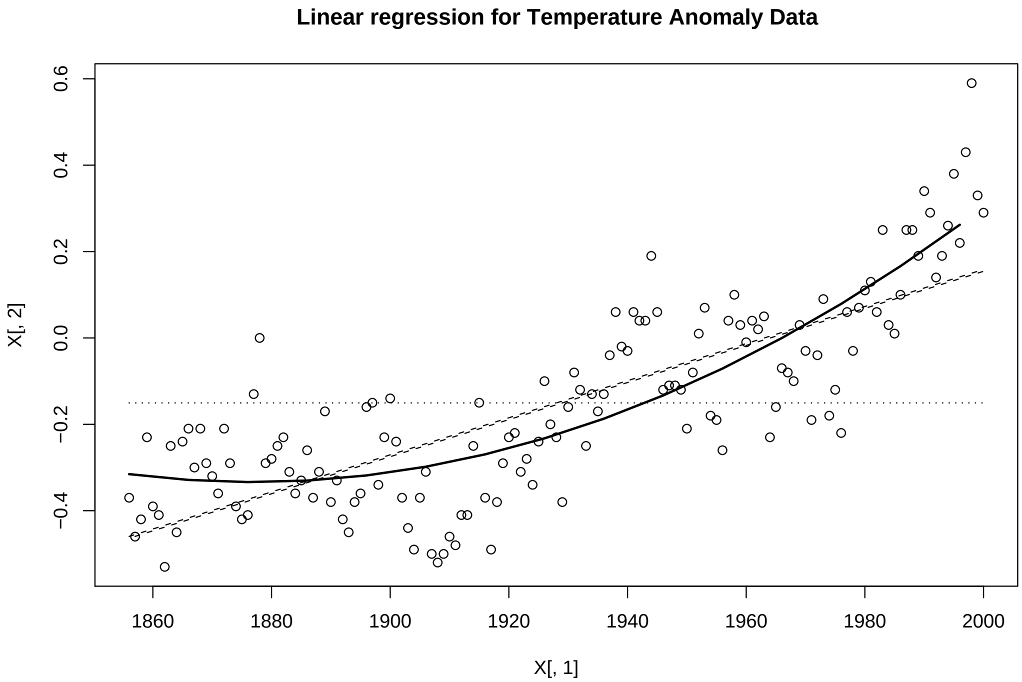
```
 Output for quadratic model:

              Estimate Std. Error t value Pr(>|t|)
beta0       -3.136e-01  3.125e-02 -10.033  < 2e-16
beta1       -1.840e-03  9.883e-04  -1.862   0.0646
beta2        4.200e-05  6.557e-06   6.406 2.05e-09

Output for linear model:
              Estimate Std. Error t value Pr(>|t|)
beta0       -0.4638247  0.0233692  -19.85   <2e-16
beta1        0.0042924  0.0002777   15.46   <2e-16


   lNorm(beta0, 0, 0)   ## log likelihood for "no trend" model
[1] 22.48718
>   lNorm(-0.463824713, 0.004292356, 0)   ## log likelihood for linear trend model
[1] 47.04444
>   lNorm(-3.135716e-01, -1.840424e-03, 4.200534e-05 )     ## for quadratic
[1] 53.39109
```

Note: T distributions appear again in this setting. However, we will not discuss t-tests in this example. T and F are more efficient tests, based on assumption of normality of ε_i. `R` reports F-test p-value less than `2.2e-16`.



(b) *Testing for quadratic.* Is quadratic term in the model (4) really necessary (despite what we see by naked eye)?

$$H_0 : \beta_2 = 0$$

Now, there is one constraint, and the constrained problem is to estimate the optimal $\boldsymbol{\theta}^* = (\beta_0, \beta_1, 0, \sigma^2)$. That is, the linear trend model discussed earlier!
For this comparison, $W = 2(53.59 - 47.04) = 13.1$, compared with $\chi^2_{0.95}(1) = 3.84$ is still highly significant. Therefore, quadratic term brings significant improvement.

There is another complication in our model. It turns out that ε_i are not really independent, but rather have a positive autocorrelation, that is, ε_i and ε_{i+1} are positively correlated. This is an often missed aspect of these data. Accounting for autocorrelation is a topic for another course, however. For example, Math 586.