1 Maximum likelihood estimators

Simply put, if we know the form of $f_X(x; \theta)$ and have a sample from $f_X(x; \theta)$, not necessarily random, the $ml$ estimator of $\theta$, $\theta_{ml}$, is that $\theta$ which maximizes

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta)$$

Where $f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta)$ is the joint density function of the sample, written as a function of $\theta$.

In this context, we call the joint density function of the realized sample, written as a function of $\theta$, the likelihood function. That is

$$L(x_1, x_2, \ldots, x_n; \theta) = f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta)$$

$\theta_{ml}$ is called the maximum likelihood estimate of $\theta$ because it is that estimate of $\theta$ that maximizes the likelihood of drawing the given sample, $x_1, x_2, \ldots, x_n$.

Maximum likelihood estimation is the most common estimation technique in econometrics.

We find $\theta_{ml}$ by maximizing $L(x_1, x_2, \ldots, x_n; \theta)$ with respect to $\theta$.

Maximum likelihood estimation is probably the most versatile tool in the econometrician’s tool box.

Note that one needs to assume a form for $f_X(x; \theta)$ to get the $ml$ estimator.

Note that maximum likelihood estimation does not require that one’s sample is a random sample, only that one can write down the likelihood function for the sample in question.

---

1. The sample is considered given, and the likelihood function identifies the likelihood of drawing that sample as a function of the parameter values.
2. Actually this is a lie, but only rarely a lie.
Given the likelihood function and collected data, there are numerous ways to find maximum likelihood estimates:

- One can do it the old-fashioned way: take partial derivatives of $L$ with respect to each element in $\theta$, set all of them equal to zero, solve the system for the $\theta$, and then check second-order conditions for a maximum.

- Let Mathematica, or some other such program, find those values of $\theta$ that maximize the likelihood function. These programs use search algorithms to find the parameters that maximize the function. In Mathematica use the command $\text{Min}$. Turn it into a maximization command by having it minimize $-L$. 
1.1 Look what happens when the sample is a random sample

If the sample is a random sample from \( f_X(x; \theta) \) then

\[
L(x_1, x_2, \ldots, x_n; \theta) = f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta) = f_X(x_1; \theta)f_X(x_2; \theta)\ldots f_X(x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta)
\]

because each observation is an independent draw from \( f_X(x; \theta) \).

That is, \( \theta_{ml} \) is that \( \theta \) which maximizes \( \prod_{i=1}^{n} f_X(x_i; \theta) \).

Further note that the \( \theta \) which maximizes \( \prod_{i=1}^{n} f_X(x_i; \theta) \) is also the \( \theta \) that maximizes \( \ln(\prod_{i=1}^{n} f_X(x_i; \theta)) \); that is, the \( \theta \) that maximizes \( \ln L \) is \( \theta_{ml} \).

And

\[
\ln L(x_1, x_2, \ldots, x_n; \theta) = \ln \left( \prod_{i=1}^{n} f_X(x_i; \theta) \right) = \sum_{i=1}^{n} \ln [f_X(x_i; \theta)]
\]

\( ^3 \)If \( x_1, x_2, \ldots, x_n \) is not a random sample, and if one has sufficient information about its non-randomness (a big assumption) one can determine \( f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta) \). For example imagine that one sample graduate students in Economics so that 22\% of the sample will consist of Korean students and 78\% non-Korean students, and that the sample is random within these two groups. In this case, can you write down the joint density function of the sample, if you know \( f_{X_k}(x) \) and \( f_{X_{nk}}(x) \)?
Typically, it is easier to maximize an additive function than it is to maximize a multiplicative function.\footnote{The product $\prod_{i=1}^{n} f_X(x_i; \theta)$ will often be a very small number, most of the terms will be fractions, and if, for example one has a sample of a thousand, the value of the likelihood function is the product of 1000 fractions.} \footnote{Note that $\sum_{i=1}^{n} \ln [f_X(x_i; \theta)]$ will be a negative number, bounded from above by zero.}

### 1.2 Some examples of maximum likelihood estimators

#### 1.2.1 Assume the rv $X$ has a Bernoulli distribution

That is, assume

$$f(x; p) = \begin{cases} 
  p^x(1-p)^{1-x} & \text{if } x = 0 \\
  p^x(1-p)^{1-x} & \text{if } x = 1 \\
  0 & \text{otherwise}
\end{cases}$$

where $0 < p \equiv \theta < 1$.

We want $p_{ml}$. Assume we have a random sample of 10 observations (the sample consists of zeros and ones). In this case

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{10} f_X(x_i; \theta)$$

$$= \prod_{i=1}^{10} p^{x_i}(1-p)^{1-x_i}$$

So, $p_{ml}$ is that $p$ that maximizes $p_{ml}$. $p_{ml}$ is also that $p$ that maximizes

$$\ln L(x_1, x_2, ..., x_n; \theta) = \sum_{i=1}^{10} \ln [f_X(x_i; \theta)]$$

$$= \sum_{i=1}^{10} \ln \left[p^{x_i}(1-p)^{1-x_i}\right]$$

$$= \sum_{i=1}^{10} \{x_i \ln p + (1-x_i) \ln(1-p)\}$$

$$= \sum_{i=1}^{10} x_i \ln p + \sum_{i=1}^{10} (1-x_i) \ln(1-p)$$

$$= \ln p \sum_{i=1}^{10} x_i + \ln(1-p) \sum_{i=1}^{10} (1-x_i)$$
Note that
\[ \sum_{i=1}^{10} x_i = 10 \bar{x} \]
and
\[ \sum_{i=1}^{10} (1 - x_i) = 10 - \sum_{i=1}^{10} x_i = 10 - 10 \bar{x} = 10(1 - \bar{x}) \]
so
\[
\ln L(x_1, x_2, ..., x_n; \theta) = \ln p \sum_{i=1}^{10} x_i + \ln(1 - p) \sum_{i=1}^{10} (1 - x_i) = 10 \bar{x} \ln p + 10(1 - \bar{x}) \ln(1 - p)
\]
To find \( p_{ml} \) we want to maximize \( 10 \bar{x} \ln p + 10(1 - \bar{x}) \ln(1 - p) \) with respect to \( p \).

Make up a sample with 10 observations and use the Min or Max command in Mathematica to find \( p_{ml} \).

Then do it the old fashioned way in terms of any random sample with 10 observations. That is, use calculus to maximize \( 10 \bar{x} \ln p + 10(1 - \bar{x}) \ln(1 - p) \) with respect to \( p \).

\[
\frac{d \ln L}{dp} = 10 \bar{x} \left( \frac{1}{p} \right) + 10(1 - \bar{x}) \left( \frac{1}{1 - p} \right)(-1)
\]
Set this equal to zero to find the critical point
\[
10 \bar{x} \left( \frac{1}{p} \right) - 10(1 - \bar{x}) \left( \frac{1}{1 - p} \right) = 0
\]
Solution is: \( \{ p = \bar{x} \} \). That is, \( \bar{x} \) (the sample mean) is the maximum likelihood estimate of \( p \).

To get a visual feel for what is going on, I am going to plot \( 10 \bar{x} \ln p + 10(1 - \bar{x}) \ln(1 - p) \) for three different values of \( \bar{x} \): \( \bar{x} = .9, \bar{x} = .3 \) and \( \bar{x} = .7 \).

\footnote{Note that the information in the data required to find this \( ml \) estimate is completely contained by the sample average, \( \bar{x} \). In these cases, \( \bar{x} \) is deemed a sufficient statistic because it contains sufficient information to estimate the parameter. It does not matter who was a success and who was a failure, only the proportion of successes.}
ln $L$: green $\bar{x} = .9$, red $\bar{x} = .7$, black $\bar{x} = .3$
The Bernoulli and Binomial - another way to solve the above problem

Our sample of \( n \) consists of \( n \) repeated Bernoulli trials (draws). It is well known that the number of successes (ones) in those \( n \) trials has a Binomial distribution. That is, if one has \( \sum_{i=1}^{n} x_i \) successes in \( n \) trials

\[
f(\sum_{i=1}^{n} x_i) = \binom{n}{\sum_{i=1}^{n} x_i} p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}
\]

where \( \binom{n}{\sum_{i=1}^{n} x_i} \) is the binomial coefficient. If \( n = 10 \)

\[
f(\sum_{i=1}^{10} x_i) = \binom{10}{\sum_{i=1}^{10} x_i} p^{\sum_{i=1}^{10} x_i} (1-p)^{10-\sum_{i=1}^{10} x_i}
\]

Therefore, another way to write the likelihood function (and log likelihood function) for our problem is

\[
L(x_1, x_2, \ldots, x_n; \theta) = \binom{10}{\sum_{i=1}^{10} x_i} p^{\sum_{i=1}^{10} x_i} (1-p)^{10-\sum_{i=1}^{10} x_i}
\]

and

\[
\ln L(x_1, x_2, \ldots, x_n; \theta) = \ln \left( \binom{10}{\sum_{i=1}^{10} x_i} \right) + \left( \sum_{i=1}^{10} x_i \right) \ln p + (10-\sum_{i=1}^{10} x_i) \ln(1-p)
\]

Note that the first term is not a function of \( p \). So

\[
\frac{d \ln L}{dp} = \left( \sum_{i=1}^{10} x_i \right) \frac{1}{p} - (10-\sum_{i=1}^{10} x_i) \left( \frac{1}{1-p} \right) = 10(\bar{x}) \frac{1}{p} - 10(1-\bar{x}) \left( \frac{1}{1-p} \right)
\]

Set this equal to zero and solve for \( p \) to determine that \( p_{ml} = \bar{x} \), just what we got when we did it the other way.
1.2.2 Assume the rv $X$ has a Poisson distribution

$X$ has a Poisson distribution if

$$f_X(x) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ for } x = 0, 1, 2, 3, \ldots$$

where $\lambda > 0$. The Poisson is a discrete distribution that can take only integer values. It is often, incorrectly, the distribution of choice when one wants to count something; e.g. the number of times Americans get married, or, to make a bad pun, the number of fish caught in a day of fishing.

For the Poisson\(^7\)

$$E[x] = \lambda = \text{var}[x]$$

In earlier notes we assumed that the number of marriages by individuals now dead has a Poisson distribution. This seems reasonable since the number of times one has been married is, hopefully, a nonnegative integer.

Further assume a random sample of 5 observations $(0, 0, 2, 2, 7)$.

Write down the likelihood function and the log likelihood function

$$L(x_1, x_2, \ldots, x_5; \lambda) = \prod_{i=1}^{5} f_X(x_i; \lambda) = \prod_{i=1}^{5} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

and

$$\ln L(x_1, x_2, \ldots, x_5; \lambda) = \sum_{i=1}^{5} \ln f_X(x_i; \lambda)$$

\[^7\text{It is obviously restrictive to assume that the mean and variance are equal. This restriction can be relaxed by, for example, assuming a negative binomial distribution. See, for example, MGB pages 99 and 438, and Greene pages 939 – 940.}\]
Since the term containing $x_i!$ does not contain $\lambda$, $\bar{x}$ is a sufficient statistic. Further note that the $\lambda$ that maximizes $-5\lambda + \ln(\lambda)5\bar{x} - \ln(x_i!)$ is also the $\lambda$ that maximizes $-\lambda + \ln(\lambda)\bar{x}$, so let’s find the $\lambda$ that maximizes $-\lambda + \ln(\lambda)\bar{x}$.\(^8\)

Let’s graph this function with $\lambda$ and $\bar{x}$ on the horizontal plane and $\ln L^*$ on the vertical axis.

This graph identifies $\lambda_{ml}$ for every value of $\bar{x}$ showing that $\lambda_{ml} = \bar{x}$

Now graph some slices of the above. Assuming $\bar{x} = 1$

Assuming $\bar{x} = 5$

\(^8\)Not that this function is not restricted to the negative range because it is not the $\ln$ of the likelihood function, a term in the $\ln$ of the likelihood function is omitted.
Note that the function takes positive values. Now assuming $\bar{x} = 2.2$
Now consider finding $\lambda_{ml}$ the old-fashioned way

$$\frac{d \ln L}{d \lambda} = -5 + 5\bar{x} \frac{1}{\lambda}$$

Set this equal to zero and solve for $\lambda$

$$-5 + 5\bar{x} \frac{1}{\lambda} = 0$$

Solution is: $\{\lambda_{ml} = \bar{x}\}$

Now assume the a specific random sample of 5 observations (0,0,2,2,7).

In this numerical example, $\bar{x} = 2.2$ (the average of 0, 0, 2, 2, 7) so $\lambda_{ml} = 2.2$, as indicated above. Wow.

Then I let my software (Mudpad) find the answer.

$-\lambda + \ln(\lambda)(2.2)$ Candidate(s) for extrema: $\{-0.46539\}$, at $\{\{\lambda = 2.2\}\}$

Let’s get the estimated probability associated with the first eight integer values

PoissonDen (0; 2.2) 0.1108; that is, there is a 11% chance one will not get married

PoissonDen (1; 2.2) 0.24377; that is, there is a 24% chance one will marry once

PoissonDen (2; 2.2) : 0.26814; that is, there is a 26% chance one will marry twice

PoissonDen (3; 2.2) : 0.19664; a 19% chance one will marry thrice

PoissonDen (4; 2.2) : 0.10815; that is, there is a 10% chance one will marry four times

PoissonDen (5; 2.2) : 4.7587 × 10^{-2};; that is, there is a 4% chance one will marry five times

PoissonDen (6; 2.2)
1.7448 × 10⁻²; a 1% chance one will marry six times

\[ \text{PoissonDen}(7; 2.2) \]

5.4838 × 10⁻³; a .5% chance that one will marry seven times.

The probability that one will marry twelve times is

\[ \text{PoissonDen}(12; 2.2) \]

2.9736 × 10⁻⁶, which is not much.

Graphing this estimated Poisson for 0 through 7 marriages:

\[ \text{PoissonDen}(k; 2.2) = \frac{2.2^k e^{-2.2}}{k!} \]

(0, 0, 0.1108, 0, 0, 1, 0, 1, 0.24377, 1, 0, 2, 0.26814, 1, 0, 3, 0, 3, 0.19664, 3, 0, 4, 0, 4, 0.10815, 4, 0, 5, 0, 5, 4.7587 × 10⁻², 5, 0, 6, 0, 6, 1.7448 × 10⁻², 6, 0, 7, 0, 7, 5.4838 × 10⁻³, 7, 0)

That is, \( E[x] = 2.2 \). What is the maximum likelihood estimate of the variance? 2.2
Now let’s make this Poisson maximum likelihood problem a little more interesting. Assume, we know how old everyone is and that everyone in the population of interest is still alive.

\[ married = x_i \quad age_i \]

\[
\begin{array}{ccc}
0 & 12 \\
0 & 50 \\
2 & 30 \\
2 & 36 \\
7 & 97 \\
\end{array}
\]

That is, we know each individual’s age, and suspect that there might be a relationship between how many times one has been married and one’s age.

How would you change the above Poisson model to take this into account?
One could assume that $\lambda$ is a function of age; e.g.

$$\lambda_i = \lambda_0 age_i$$

Making $\lambda$ a **linear** function of age is simple by not plausible, so let's "just do it."

In which case,

$$\ln L(x_1, x_2, ..., x_5, age_1, ... age_5; \lambda_0, \ldots) = \sum_{i=1}^{5} \ln f_X(x_i; \lambda_0 age_i)$$

$$= \sum_{i=1}^{5} \ln \left[ \frac{e^{-(\lambda_0 age_i)} (\lambda_0 age_i)^{x_i}}{x_i!} \right]$$

$$= \sum_{i=1}^{5} \ln(e^{-(\lambda_0 age_i)} (\lambda_0 age_i)^{x_i}) - \ln(x_i!)$$

$$= \sum_{i=1}^{5} \ln((\lambda_0 age_i)^{x_i}) + \ln(x_i!)$$

$$= \sum_{i=1}^{5} -\lambda_0 age_i + x_i \ln(\lambda_0 age_i) - \ln(x_i!)$$

$$= \left[ -\lambda_0 \sum_{i=1}^{5} age_i + \sum_{i=1}^{5} x_i \ln(\lambda_0 age_i) - \sum_{i=1}^{5} \ln(x_i!) \right]$$

Now let’s take the partial with respect to $\lambda_0$

$$\frac{\partial \ln L(x_1, x_2, ..., x_5; \lambda_0)}{\partial \lambda_0} = \frac{\partial}{\partial \lambda_1} \left[ -\lambda_0 \sum_{i=1}^{5} age_i + \sum_{i=1}^{5} x_i \ln(\lambda_0 age_i) - \sum_{i=1}^{5} \ln(x_i!) \right]$$

$$= -\sum_{i=1}^{5} age_i + \sum_{i=1}^{5} x_i \frac{\partial \ln(\lambda_0 age_i)}{\partial \lambda_0}$$

$$= -\sum_{i=1}^{5} age_i + \sum_{i=1}^{5} x_i \frac{age_i}{(\lambda_0 age_i)}$$

$$= -\sum_{i=1}^{5} age_i + \sum_{i=1}^{5} \frac{x_i}{\lambda_0}$$

$$= -5 \bar{age} + \frac{5 \bar{x}}{\lambda_0}$$

Set this equal to zero and solve for $\lambda_0$.

$$-5 \bar{age} + \frac{5 \bar{x}}{\lambda_0} = 0$$
Solution is $\bar{x}_{age}$, average number of marriages divided by average age, and

$$\lambda_{ml} = \frac{\bar{x}}{age}$$

So,

$$\lambda_i = \lambda_0 age_i$$

$$= \frac{\bar{x} age_i}{age}$$

This is interesting, our expectation of one’s number of marriages is the sample average, $\bar{x}$, weighted by $\frac{age_i}{age}$ (the individual’s age as a proportion of the average age in the sample).

The ml estimate of $\lambda_0$ for the sample at hand is something like .0488.

Now make the problem more interesting by assuming.

$$\lambda = \lambda_0 + \lambda_1 age_i$$

That is, estimate a slope and an intercept. Try and work this out on your own.
1.2.3 Assume some random variable $X$ has a normal distribution, that is

$$f_X(x; \mu_x, \sigma^2_x) = \frac{1}{\sqrt{2\pi\sigma^2_x}} e^{-\frac{1}{2\sigma^2_x}(x-\mu_x)^2}$$

We draw a random sample from $f_X(x; \mu_x, \sigma^2_x)$ of $n$ observations from this distribution.

We want to find the $ml$ estimators of $\mu_x$ and $\sigma^2_x$.

This is the most famous $ml$ problem.
In this case,

\[
L(x_1, x_2, \ldots, x_n; \mu_x, \sigma_x^2) = \prod_{i=1}^{n} f_X(x_i; \mu_x, \sigma_x^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \sigma_x^2}} e^{-\frac{1}{2\sigma_x^2}(x_i - \mu_x)^2}
\]

because each observation is independent, and the ln of the likelihood function is\(^9\)

\[
\ln L(x_1, x_2, \ldots, x_n; \mu_x, \sigma_x^2) = \ln \prod_{i=1}^{n} f_X(x_i; \mu_x, \sigma_x^2)
= \ln \prod_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi \sigma_x^2}} e^{-\frac{1}{2\sigma_x^2}(x_i - \mu_x)^2}
= \ln \prod_{i=1}^{n} (2\pi)^{-1/2}(\sigma_x^2)^{-1/2} e^{-\frac{1}{2\sigma_x^2}(x_i - \mu_x)^2}
= \sum_{i=1}^{n} \ln((2\pi)^{-1/2}(\sigma_x^2)^{-1/2} e^{-\frac{1}{2\sigma_x^2}(x_i - \mu_x)^2})
= \sum_{i=1}^{n} \{ -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma_x^2) - (\frac{1}{2\sigma_x^2})(x_i - \mu_x)^2 \}
= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma_x^2 - (\frac{1}{2\sigma_x^2}) \sum_{i=1}^{n} (x_i - \mu_x)^2
\]

We want to maximize this with respect to \(\mu_x\) and \(\sigma_x^2\).

\(^9\)Note that the likelihood function is a \(n\)-variate normal with zero covariances.
Take the partials

\[
\frac{\partial}{\partial \mu_x} \left[ -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma_x^2 - \left( \frac{1}{2\sigma_x} \right) \sum_{i=1}^{n} (x_i - \mu_x)^2 \right] = \left[ - \frac{n\mu_x - \sum_{i=1}^{n} x_i}{\sigma_x^2} \right]
\]

and

\[
\frac{\partial}{\partial \sigma_x^2} \left[ -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma_x^2 - \left( \frac{1}{2\sigma_x} \right) \sum_{i=1}^{n} (x_i - \mu_x)^2 \right] = \frac{\sum_{i=1}^{n} (x_i - \mu_x)^2 - n\sigma_x^2}{2(\sigma_x^2)^2}
\]

Set these both equal to zero and solve for \( \mu_x \) and \( \sigma_x^2 \). Start with the first equation

\[
\left[ - \frac{n\mu_x - \sum_{i=1}^{n} x_i}{\sigma_x^2} \right] = 0
\]

Note that the \( \mu_x \) that solves this is \( \mu_x = \frac{1}{n} \sum_{i=1}^{n} x_i \). That is, the maximum likelihood estimate of \( \mu_x \) is \( \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \).

Plug this into the second partial, set equal to zero, and solve for the maximum likelihood estimate of \( \sigma_x^2 \)

\[
\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 - n\sigma_x^2}{2(\sigma_x^2)^2} = 0
\]

That is, solve \( \sum_{i=1}^{n} (x_i - \bar{x})^2 - n\sigma_x^2 \) for \( \sigma_x^2 \), which is

\[
\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

So the maximum likelihood estimate of \( \mu_x \) is \( \bar{x} \) and the maximum likelihood estimate of \( \sigma_x^2 \) is \( \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \).

Note that the first is unbiased, the second is not - both are asymptotically unbiased, a term we have not defined.
Make up some data - maybe 4, 7, 1 and find the max likelihood estimates. The log likelihood function with unnecessary terms removed is:

\[-\frac{n}{2} \ln \sigma_x^2 - \frac{1}{2\sigma_x^2} \sum_{i=1}^{n} (x_i - \mu_x)^2 = -\frac{n}{2} \ln \sigma_x^2 - \frac{1}{2\sigma_x^2} [(4 - \mu_x)^2 + (7 - \mu_x)^2 + (1 - \mu_x)^2]\]

Plotting this as a function of \( \mu_x \) and \( \sigma_x^2 \):

\[\text{Candidate(s) for extrema: } \{ \frac{-3}{2} \ln 6 - \frac{3}{2} \}, \text{ at } \{ [\sigma_x^2 = 6, \mu_x = 4] \}\]

Note that we maximize \( \text{LnL} \) by taking its derivative with respect to the parameter and searching for a local interior maximum. We could also have used a computer search algorithm such as MIN in Mathematica to find \( \hat{\mu}_x \) and \( \hat{\sigma}_x^2 \).

That is,

\[ \text{Minimize } -\ln L \]
1.3 A more general max lik problem

Consider the following problem. Assume that the \( i^{th} \) random variable \( Y_i \) is distributed\(^{10} \)

\[
f_Y (y_i : \mu_{y_i}, \sigma^2_{y_i}) \text{ where } i = 1, 2, ..., n
\]

That is, the only thing that distinguishes the distributions of \( Y_i \) and \( Y_j \) is \( \mu_{y_i} \) vs. \( \mu_{y_j} \). It might look as follows (the subscript is suppressed in this example figure). Note that we are assuming the density has two parameters, the mean and the variance.

We know the form of \( f_Y (y_i : \mu_{y_i}, \sigma^2_{y_i}) \) but not the specific value of \( \sigma^2_{y_i} \) or values of \( \mu_{y_1}, \mu_{y_2}, ..., \mu_{y_n} \). We want to estimate them.

\(^{10}\)Note that I am now naming the random variable \( Y \) rather than \( X \). This is more conventional when one assumes that the expected value of \( Y \) varies across observations as a function of one or more explanatory variables. Denoting the dependent variable \( Y \) is the convention in regression analysis.
Further assume
\[ \mu_{yi} = \alpha + \beta' x_i \] where \( i = 1, 2, \ldots, n \)
where the vector \( x_i \) is observed. For the moment, assume \( x_i \) is a scalar. In this framework, \( x_i \) is not a random variable from our perspective, rather it is assumed fixed in repeated samples. In which case, \( f_Y(y_i, \alpha + \beta x_i, \sigma_y^2) \) and the parameters are \( \alpha, \beta, \) and \( \sigma_y^2 \). Note that \( \mu_{yi} \) is a linear function of \( x_i \) and, by assumption, \( \sigma_y^2 = \sigma_y^2 \forall i \).

Imagine a random sample of \( n \) observations of \((y_i, x_i), i = 1, 2, \ldots, n\) and we want the maximum likelihood estimator of \( \alpha, \beta, \) and \( \sigma_y^2 \).

\[
L(y_1, y_2, \ldots, y_n, x_1, x_2, \ldots, x_n; \alpha, \beta, \sigma_y^2) = \prod_{i=1}^{n} f_Y(y_i, \alpha + \beta x_i, \sigma_y^2)
\]
and
\[
\ln L = \sum_{i=1}^{n} \ln f_Y(y_i, \alpha + \beta x_i, \sigma_y^2)
\]
We get the maximum likelihood estimator of \( \alpha, \beta, \) and \( \sigma_y^2 \) by maximizing \( \ln L \) with respect to these parameters.

\[11\] Given \( x_i, y_i \) is a random draw from \( f_Y(y, \alpha + \beta x_i, \sigma_y^2) \)
1.3.1 For example, if one assumes a normal distribution, which one is not forced to do

\[
f_Y(y_i, \alpha + \beta x_i, \sigma_y^2) = \frac{1}{\sqrt{2\pi \sigma_y^2}} e^{-\left(\frac{1}{2\sigma_y^2}\right)[y_i-(\alpha+\beta x_i)]^2}\]

That is (this normality assumption implies an additive normally distributed \(\varepsilon\))

\[y_i = \alpha + \beta x_i + \varepsilon_i\]

where

\[\varepsilon \sim N(0, \sigma_y^2)\]

This is the classical linear regression model (CLR model). In which case,

\[
\ln L() = \sum_{i=1}^{n} \ln f_Y(y_i, \alpha + \beta x_i, \sigma_y^2)
\]

\[
= \sum_{i=1}^{n} \ln \left[ (2\pi)^{-\frac{1}{2}} \left(\frac{1}{2\sigma_y^2}\right)^{-\frac{1}{2}} e^{-\left(\frac{1}{2\sigma_y^2}\right)[y_i-(\alpha+\beta x_i)]^2} \right]
\]

\[
= \sum_{i=1}^{n} \left[ -\frac{1}{2} \ln (2\pi) - \frac{1}{2} \ln (\sigma_y^2) - \left(\frac{1}{2\sigma_y^2}\right)(y_i-[\alpha+\beta x_i])^2 \right]
\]

\[
= -\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln (\sigma_y^2) - \left(\frac{1}{2\sigma_y^2}\right) \sum_{i=1}^{n} (y_i-[\alpha+\beta x_i])^2
\]
The maximum likelihood estimates of $\alpha, \beta,$ and $\sigma_y^2$ are those values of $\alpha, \beta,$ and $\sigma_y^2$ that maximize $\ln L()$. Let's find them.

\[
\frac{d\ln L}{d\alpha} = 2 \left( -\frac{1}{2\sigma_y^2} \right) \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) (-1)
= \left( \frac{1}{\sigma_y^2} \right) \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) \text{ set } = 0 \tag{1}
\]

\[
\frac{d\ln L}{d\beta} = 2 \left( -\frac{1}{2\sigma_y^2} \right) \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) (-x_i)
= \frac{1}{\sigma_y^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) (x_i)
= \frac{1}{\sigma_y^2} \sum_{i=1}^{n} (y_i x_i - \alpha x_i - \beta x_i^2) \text{ set } = 0 \tag{2}
\]

\[
\frac{d\ln L}{d\sigma_y^2} = \left( \frac{n}{2} \right) \frac{1}{\sigma_y^2} + \left( \frac{1}{2\sigma_y^4} \right) \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2
= \left( \frac{1}{2\sigma_y^2} \right) \left[ -n + \frac{1}{\sigma_y^4} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right]
= \left( \frac{1}{2\sigma_y^2} \right) \left[ \frac{1}{\sigma_y^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 - n \right] \text{ set } = 0 \tag{3}
\]

There are three equations in three unknowns $(\alpha, \beta, \sigma_y^2)$. Solve for $\alpha_m, \beta_m, \sigma_y^m$.

Assuming $\sigma_y^2 > 0$, from the first equation we know that

\[
\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0
\]

but

\[
\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = -n\alpha + \sum_{i=1}^{n} (y_i - \beta x_i)
= -n\alpha + \sum_{i=1}^{n} y_i - \beta \sum_{i=1}^{n} x_i
\]

Noting that $\sum_{i=1}^{n} y_i = n\bar{y}$ and $\sum_{i=1}^{n} x_i = n\bar{x},$

\[
\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = -n\alpha + n\bar{y} - \beta n\bar{x} = 0
= -\alpha + \bar{y} - \beta \bar{x} = 0 \text{ implying}
\alpha = \bar{y} - \beta \bar{x} \tag{4}
\]
Plug this result into the second equation

\[
\frac{d \ln L}{d \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i x_i - \alpha x_i - \beta x_i^2) = 0
\]  

(5)

to obtain

\[
\frac{1}{\sigma^2} \sum_{i=1}^{n} \left( y_i x_i - (\bar{y} - \beta \bar{x}) x_i - \beta x_i^2 \right) = 0
\]

\[
\sum_{i=1}^{n} \left( y_i x_i - \bar{y} x_i + \beta \bar{x} x_i - \beta x_i^2 \right) = 0
\]

\[
\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i + \beta \bar{x} \sum_{i=1}^{n} x_i - \beta \sum_{i=1}^{n} x_i^2 = 0
\]

\[
\sum_{i=1}^{n} y_i x_i - \bar{y} n \bar{x} + \beta \bar{x} n \bar{x} - \beta \sum_{i=1}^{n} x_i^2 = 0
\]

\[
\begin{align*}
\beta n \bar{x}^2 - \beta \sum_{i=1}^{n} x_i^2 & = n \bar{y} \bar{x} - \sum_{i=1}^{n} y_i x_i \\
\beta \left( n \bar{x}^2 - \sum_{i=1}^{n} x_i^2 \right) & = n \bar{y} \bar{x} - \sum_{i=1}^{n} y_i x_i
\end{align*}
\]  

(6)

implying

\[
\hat{\beta}_{ml} = \frac{n \bar{y} \bar{x} - \sum_{i=1}^{n} y_i x_i}{n \bar{x}^2 - \sum_{i=1}^{n} x_i^2}
\]
There are many common ways of expressing $\hat{\beta}_{ml}$. If one multiplies the numerator and denominator by $-1$, one obtains

$$\beta_{ml} = \frac{\sum_{i=1}^{n} y_i x_i - ny \bar{x}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2}$$

To obtain another common form of $\beta_{ml}$ note that

$$\sum (y_i - \bar{y})(x_i - \bar{x}) = \sum [y_i x_i - \bar{y}x_i - \bar{y}x_i + \bar{y}x]$$

$$= \sum y_i x_i - \bar{x} \sum y_i - \bar{y} \sum x_i + n\bar{y}\bar{x}$$

$$= \sum y_i x_i - \bar{y}n\bar{x} - \bar{y} \sum x_i + n\bar{y}\bar{x}$$

$$= \sum y_i x_i - \bar{y} \sum x_i$$

$$= \sum y_i x_i - n\bar{y}\bar{x}$$

, which is the numerator in the above expressions for $\beta_{ml}$.

And

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - x_i\bar{x} - \bar{x}x_i + \bar{x}^2)$$

$$= \sum x_i^2 - \bar{x} \sum x_i - \bar{x} \sum x_i + n\bar{x}^2$$

$$= \sum x_i^2 - \bar{x}n\bar{x} - \bar{x}n\bar{x} + n\bar{x}^2$$

$$= \sum x_i^2 - n\bar{x}^2 - n\bar{x}^2 + n\bar{x}^2$$

$$= \sum x_i^2 - n\bar{x}^2$$

which is the denominator in the above expressions for $\beta_{ml}$. So,

$$\beta_{ml} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(MGB, p. 499). Or in terms of the deviations around the means

$$\tilde{y}_i \equiv y_i - \bar{y}$$

and

$$\tilde{x}_i \equiv x_i - \bar{x}$$

$$\hat{\beta}_{ml} = \frac{\sum \tilde{y}_i \tilde{x}_i}{\sum \tilde{x}_i^2}$$
Now let's calculate $\hat{\alpha}_{ml}$. Recall that

$$\alpha = \bar{y} - \beta \bar{x}$$

Plug in $\beta_{ml}$.

$$\alpha_{ml} = \bar{y} - \beta_{ml} \bar{x}$$

to obtain

$$\alpha_{ml} = \bar{y} - \bar{x} \sum \frac{\bar{y}_i \bar{x}_i}{\bar{x}_i^2}$$

Looking ahead, the maximum likelihood estimates of $\alpha$ and $\beta$, assuming

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where

$$\epsilon_i \sim N \left(0, \sigma^2 \right)$$

are also the least square estimates. That is, for the Classical Linear Regression Model, the maximum likelihood estimates of $\alpha$ and $\beta$ are equivalent to the least squares estimates.
Now find the maximum likelihood estimates of $\sigma_y^2$. Recall the third first-order condition

$$
\left(\frac{1}{2\sigma_y^2}\right) \left[ \frac{1}{\sigma_y^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 - n \right] = 0
$$

$$
\left[ \frac{1}{\sigma_y^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 - n \right] = 0
$$

$$
\frac{1}{\sigma_y^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = n
$$

$$
\frac{1}{\sigma_y^2} = \frac{n}{\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2}
$$

$$
\sigma_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2
$$

So the maximum likelihood estimator of $\sigma_y^2$ is

$$
\sigma_{y,\text{ml}}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha_{\text{ml}} - \beta_{\text{ml}} x_i)^2
$$

In summary, we have just derived the maximum likelihood estimates of $\alpha$, $\beta$, and $\sigma_y^2$ with a random sample of size $n$ assuming in the population

$$
y_i = \alpha + \beta x_i + \varepsilon_i
$$

where

$$
\varepsilon_i \sim N(0, \sigma^2_{\varepsilon})
$$

That is, we have derived the maximum likelihood estimators for $\alpha$, $\beta$, and $\sigma_y^2$ for the classical linear regression model.
1.4 Let’s do another maximum likelihood problem: return to the Bernoulli problem

Assume two alternatives and the probability that individual $i$ chooses alternative 1 on any trial, $t$, is $p_i$.\(^{12}\) That is,

$$f_X(x_{it} : p_i) = \begin{cases} p_i^{x_{it}} (1 - p_i)^{1-x_{it}} & \text{for } x_{it} = 0 \text{ or } 1 \\ 0 & \text{otherwise} \end{cases}$$

where $i = 1, 2, \ldots, n$. $x_{it} = 1$ if individual $i$ chooses alternative 1 on trial $t$, and zero otherwise, $t = 1, 2, \ldots, T$.

Let $x_i$ be the number of times individual $i$ chooses alternative 1 in $T$ trials.

$$x_i = \sum_{t=1}^{T} x_{it}$$

In which case, we can (have) shown that

$$f_{X_i}(x_i : p_i, T) = \left( \frac{T}{x_i} \right) p_i^{x_i} (1 - p_i)^{T-x_i} \quad i = 1, 2, \ldots, n$$

Further assume\(^{13}\)

$$p_i = \alpha + \beta G_i$$

where

$$G_i = 1 \text{ if male and zero otherwise}$$

Note that the variable $G_i$, which can only take one of two values, 0 or 1. Variables with this property are typically referred to as dummy variables.

\(^{12}\)Note that we are allowing $p$ to vary across individuals, but, for a given individual, not across trials.

\(^{13}\)Note that this functional specification would allow $p_i$ to be outside of the zero to one range. I exclude that possibility below.
To make things simple, let’s say we know that \( \alpha = .10 \) (God told us). This implies that \( p_i = .1 \) if female. Also God told us that \( \beta \) equals either 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, or 0.8. (Note how God simplified the problem for us)

We have a random sample of \( n \) individuals. That is, for \( n \) independent individuals we observe the choices each makes on \( T \) independent trials. We observe \( T \) trials for each of the \( n \) individuals.

What is the maximum likelihood estimator of \( \beta? \)

First note that

\[
f_X (x_i, p_i, T) = \left( \frac{T}{x_i} \right) p_i^{x_i} (1 - p_i)^{T - x_i}
\]

\[
= \left( \frac{T}{x_i} \right) (.1 + \beta G_i)^{x_i} (1 - (.1 + \beta G_i))^{T - x_i}
\]

Therefore,

\[
L = \prod_{i=1}^{n} \left( \frac{T}{x_i} \right) (.1 + \beta G_i)^{x_i} (1 - (.1 + \beta G_i))^{T - x_i}
\]

So,

\[
\ln L = \sum_{i=1}^{n} \ln \left[ \left( \frac{T}{x_i} \right) (.1 + \beta G_i)^{x_i} (1 - (.1 + \beta G_i))^{T - x_i} \right]
\]

\[
= \sum_{i=1}^{n} \left[ \ln \left( \frac{T}{x_i} \right) + x_i \ln (.1 + \beta G_i) + (T - x_i) \ln (1 - 0.1 - \beta G_i) \right]
\]

\[
= \sum_{i=1}^{n} \left[ \ln \left( \frac{T}{x_i} \right) + x_i \ln (.1 + \beta G_i) + (T - x_i) \ln (0.9 - \beta G_i) \right]
\]

The maximum likelihood estimator of \( \beta \) is the \( \beta \) that maximizes the above equation but it is also the \( \beta \) that maximizes

\[
\sum_{i=1}^{n} [x_i \ln (.1 + \beta G_i) + (T - x_i) \ln (0.9 - \beta G_i)]
\]

for a given random sample. One would calculate this for \( \beta = 0.1, 0.2, \ldots 0.8 \) and the \( \beta_{\text{ml}} \) is the one that maximizes the function.
For example, assume $T = 2$ and $n = 3$ such that

\[
\begin{align*}
x_{11} &= 1 \\
x_{12} &= 1 \\
x_{21} &= 1 \\
x_{22} &= 0 \\
x_{31} &= 0 \\
x_{32} &= 0
\end{align*}
\]

where $x_{nt} = 1$ if individual $n$ choose alternative 1 on trial $t$.

Suppose we know that individuals 1 and 2 are males and individual 3 is a female. What is the maximum likelihood estimate of $\beta, \beta_{ml}$? You figure it out.

Remember that for maximum likelihood estimation of the above type, one needs to know the form of $f_x(x; \theta)$
1.5 why we like max lik technique

1. very general

2. Estimators have desirable asymptotic statistical properties under very general conditions.

3. doesn’t require a random sample (but having a random sample greatly simplifies the process)

4. Easy to do hypothesis testing and tests of significance (see below)

5. If \( \theta_{ml} \) is the maximum likelihood estimator of \( \theta \) and there is exists some function of \( \theta \), \( g(\theta) \), then \( g(\theta_{ml}) \) is the maximum likelihood estimator of \( g \), \( g_{ml} \). This is called the invariance principle, and it will greatly simplify your life in econometrics.

The drawback? of the ML technique is that one needs, usually, to assume knowledge of \( f_X(x; \theta) \), but mostly we do that anyway.
1.6 M.L. hypothesis testing and confidence intervals

1.6.1 the likelihood ratio test

Need to check all of the notation here for consistency and consistency with earlier sections

Imagine that one assumes some \( f_X(x; \theta) \) where \( \theta \) is a vector, \( \theta \equiv \theta_1, \theta_2, ..., \theta_k \).

Consider the \( ml \) estimator of the vector \( \theta \) with and without constraints/restrictions on that vector. For example, compare estimation where you assume \( \theta_2 = 0 \) with estimates where you do not assume, \( \theta_2 = 0 \). In the one case, you estimate \( \theta_1, \theta_2, ..., \theta_k \), in the other you estimate \( \theta_1, \theta_3, ..., \theta_k \) with \( \theta_2 \) set equal to 0, so estimate one less parameter.\(^{14}\) Refer to the former vector as \( \theta^1 \) and the latter, more restrictive vector, as \( \theta^0 \).

Estimation of \( \theta^0 \) is more constrained that estimation of \( \theta^1 \) and this has important implications. When one does \( ml \) estimation one maximizes the likelihood function, and when one is more constrained, the maximum cannot be greater than when it is less constrained: adding constraints cannot increase the maximum value of the likelihood function.

In more detail: We have defined the likelihood function in general as \( L(X_1, X_2, ... X_n : \theta) \), where the \( X \) are not random variables; they are the sample. If one evaluate it with a specific sample one has \( L(x_1, x_2, ... x_n : \theta) \) which is a function of only the \( \theta \) since the \( x_i \) are numbers. If one plugs in estimates of the \( \theta \), \( \hat{\theta} \), \( L(x_1, x_2, ... x_n : \hat{\theta}) \) is a number. If one plugs in \( \theta_{ml} \), \( L(x_1, x_2, ... x_n : \theta_{ml}) \) is the maximum number \( L() \) can take for this sample; that is why the \( \theta_{ml} \) are the \( ml \) estimates.

Compare \( L(x_1, x_2, ... x_n : \theta^0_{ml}) \) and \( L(x_1, x_2, ... x_n : \theta^1_{ml}) \) where the \( \theta^0 \) is more restricted/constrained than \( \theta^1 \). By the logic above, \( L(x_1, x_2, ... x_n : \theta^0_{ml}) \leq L(x_1, x_2, ... x_n : \theta^1_{ml}) \), the former typically a larger negative number.

The likelihood ratio test: Put simply, if \( L(x_1, x_2, ... x_n : \theta^0_{ml}) \) is sufficiently smaller than \( L(x_1, x_2, ... x_n : \theta^1_{ml}) \) we conclude that the restrictions imposed in \( \theta^0 \) were not appropriate: we reject the null hypothesis that \( \theta = \theta^0 \).

To simplify the notation, let \( L_{\theta_0} = L(x_1, x_2, ... x_n : \theta^0_{ml}) \) and \( L_{\theta_1} = L(x_1, x_2, ... x_n : \theta^1_{ml}) \). So \( \frac{L_{\theta_0}}{L_{\theta_1}} \leq 1 \), and we decide the restriction that \( \theta = \theta^0 \) is a bad idea if \( \frac{L_{\theta_0}}{L_{\theta_1}} \) is sufficiently less than one.

\(^{14}\)For example, one assumes \( N(x : \mu, \sigma^2) \). One could use maximum likelihood to find estimates of \( \mu \) and \( \sigma^2 \), or one might simplify the problem by assuming \( \sigma^2 = 1 \), reducing by one the number of parameters to estimate.
But how much less than one must it be before one rejects the null hypothesis that $\theta = \theta^0$? That is up to you.

To answer this question statistically, we need to investigate the distribution of $\frac{L_{\theta^0_{m1}}}{L_{\theta^1_{m1}}}$; it is a random variable that varies across samples. I don’t know the distribution of $\frac{L_{\theta^0_{m1}}}{L_{\theta^1_{m1}}}$ but I do know the distribution of $C = -2(\ln L_{\theta^0_{m1}} - \ln L_{\theta^1_{m1}})$. $C$ is a r.v. with a Chi-Squared distribution, and its parameter ("degrees of freedom") equals to the number of restriction in $\theta^0$ that are not restrictions in $\theta^1$.

One sets a critical level for $C$ and then sees whether the realized $c$ is greater than this critical level. If it is, one reject the null hypothesis that $\theta = \theta^0$. If not, one fails to reject the null hypothesis.

For example if I wanted to reject $\theta = \theta^0$ only 5% of the time when it is true ($\theta = \theta^0$), I would choose $c_{crit}$ such that only 5% of the Chi-squared density lies to the right of $c_{crit}$. If, for example, the degrees of freedom (number of additional restrictions) is 1, $c_{crit-.05} = 3.84$. If the degrees of freedom (number of additional restrictions) is 3, $c_{crit-.05} = 7.81$.

So, imagine I $ml$ estimated my model with and without the single restriction that $\theta_2 = 0$. I would then calculate $-2(\ln L_{\theta^0_{m1}} - \ln L_{\theta^1_{m1}})$. If is more that 3.84, I would reject the null hypothesis that $\theta_2 = 0$. If it is less than 3.84 I would fail to reject this null hypothesis.

![Chi-squared distribution with one degree of freedom (param=1)](image)
Note that the likelihood ratio test just presented is very general: 
\(-2(\ln L_{\theta_0} - \ln L_{\theta_1})\) has a Chi-square distribution no matter the form of \(f_X(x; \theta)\) or the number of restrictions, all this is required is that \(\theta^0\) is a special case (restricted case) of \(\theta^1\). Many Chi-squared tests in the literature, while not appearing to be likelihood ratio tests, are special cases of the likelihood ratio test.

For example: assume the rv \(X\) has a Bernoulli distribution and in a random sample of 10, \(\bar{x} = .3\) In which case, \(\ln L(x_1, x_2, \ldots, x_n; p) = 10(\cdot3) \ln p + 10(1 - \cdot3) \ln(1 - p)\). Graphing this

\[\ln L: \bar{x} = .3, \ln L_{\theta_1}^* = -6.1086, \ln L_{\theta_0}^* = -6.9315\]

\(p_{ml} = .3\) and the maximum value of \(\ln L_{\theta_0}^*\) \(-6.1086\) is at \(p_{ml} = .3\)

Now consider the null hypothesis that \(p = .5\), a random allocation.\(^{15}\) In which case \(\ln L_5 = 10(\cdot3) \ln(5) + 10(1 - \cdot3) \ln(1 - (5)) = -6.9315\), the black, dotted horizontal line on the above graph.

Is \(\ln L_5 = -6.9315\) sufficiently less than \(\ln L_{\theta_0}^* = -6.1086\) to reject the null hypothesis that \(p = 5\)?

\(^{15}\)Note that there are no estimated parameters in the restricted model.
Calculate the statistic \(-2(\ln L_5 - \ln L_6) = -2(-6.9315 - (-6.1086)) = 1.6458 < 3.84 \equiv \chi^2_{\alpha, 1\text{-df}}\); so in this example, one fails to reject the null hypothesis that \(p = .5\).

Most of the hypothesis tests I do are likelihood-ratio tests.

This has been a brief but important introduction to hypothesis testing.

1.6.2 Using the likelihood ratio statistic to construct a confidence interval

It is easy, at least conceptually, to construct a confidence interval using the likelihood ratio statistic.

To start simply, assume the ln likelihood function is a function of only one parameter, \(\beta\), so the value of the likelihood function evaluated at \(\beta = \beta_{ml}\) is \(L_{\beta_{ml}} \equiv L(x_1, x_2, \ldots, x_n : \beta_{ml})\), a number.

The likelihood function might look as follows, where the blue line represents the level \(L_{\beta_{ml}}\).
Recollect that $-2(\ln L - \ln L_{\beta_m})$ has a Chi-square distribution and note that after estimation $\ln L_{\beta_m}$ is a number. Choose a critical value for the confidence interval; for example, with one degree of freedom, for a 95\% confidence interval $\chi^2_{c_{crit-.05}, df=1} = 3.84$.

So $\beta$ will be in the confidence interval if $-2(\ln L - \ln L_{\beta_m}) \leq 3.84$. To find the interval, one simple finds all those values of $\beta$ in this range.

One constructs the interval by numerically finding all of the $\beta$ for which $-2(\ln L - \ln L_{\beta_m}) \leq 3.84$. The restriction $-2(\ln L - \ln L_{\beta_m}) \leq 3.84$ implies $\ln L \geq \ln L_{\beta_m} - 1.92$.\footnote{$-2(\ln L - \ln L_{\beta_m}) \leq 3.84$ implies $-(\ln L - \ln L_{\beta_m}) \leq 1.92$ implies $(\ln L - \ln L_{\beta_m}) \geq -1.92$ implies $\ln L \geq -1.92 + \ln L_{\beta_m}$}

The following is an example graph of $\ln L$ as a function of $\beta$ where the blue line represents $\ln L_{\beta_m}$ and the orange line $\ln L = \ln L_{\beta_m} - 1.92$.

In this example, $\ln L = \ln L_{\beta_m} - 1.92 = -6.1086 - 1.92 = -8.0286$

Note that the confidence interval for $\beta$ is the projection onto the $\beta$ axis of the orange line between the red lines.

Note that this technique works no matter the number of parameters; the numerical finding of the confidence interval just becomes more difficult.
If someone would like to do a replacement assignment, they can come up with a two-parameter example where you show the confidence "interval" as a two-dimensional eclipse in $\beta_1, \beta_2$ space. That is, specify a density function with two random variables. Assume a sample. Derive the m.l. estimates for your sample, derive the joint confidence interval on $\beta_1, \beta_2$.

Further note that this simple method of finding confidence intervals for vectors of parameters is typically not the method used in most software packages.
See my notes on chi-square tests in the hypothesis testing section

1.6.3 example: a confidence interval for our Bernoulli

Consider the Chi-square statistic \(-2(\ln L_p - \ln L_p^*)\) where, from above, \(\ln L_{p_{ml}} = -6.1086\). So the statistic, written as a function of only \(\ln L_p\) is 
\[
-2(\ln L_p - 6.1086) = 2(\ln L_p^* - 6.1086) \tag{38}
\]

One will not reject the null hypothesis if 
\[
-2(\ln L_p + 6.1086) \leq 3.84 \equiv \chi^2_{crit, .05, df=1}
\]

Solve \(-2(\ln L_p + 6.1086) \leq 3.84\) for \(\ln L_p\): The solution is \(\ln L_p \geq -8.0286\). The confidence interval for \(p\) is simple those \(p\) consistent with \(L_p^*\) is

In our example, \(\ln L_p = 10(.3) \ln(p) + 10(1-.3) \ln(1-(p))\), so the restriction on the confidence interval is \(\ln L_p = 3 \ln(p) + 7 \ln(1-(p)) \geq -8.028\)

Doing a simple grid search over values of \(p\)

\[
\begin{align*}
L_p^* = 3 \ln(p) + 7 \ln(1-(p)) & \quad \text{whether } p \text{ is in the confidence interval} \\
3 \ln(.05) + 7 \ln(1-(.05)) & = -9.3462 \quad p = .05 \text{ is not in the confidence interval} \\
3 \ln(.075) + 7 \ln(1-(.075)) & = -8.3165 \quad p = .075 \text{ is not in the confidence interval} \\
3 \ln(.08) + 7 \ln(1-(.08)) & = -8.1609 \quad p = .08 \text{ is not in the confidence interval} \\
3 \ln(.085) + 7 \ln(1-(.085)) & = -8.0171 \quad p = .085 \text{ is in the confidence interval} \\
3 \ln(.1) + 7 \ln(1-(.1)) & = -7.6543 \quad p = .1 \text{ is in the confidence interval} \\
3 \ln(.2) + 7 \ln(1-(.2)) & = -6.3903 \quad p = .2 \text{ is in the confidence interval} \\
3 \ln(.3) + 7 \ln(1-(.3)) & = -6.1086 \quad p = .3 \text{ is in the confidence interval} \\
3 \ln(.4) + 7 \ln(1-(.4)) & = -6.3247 \quad p = .4 \text{ is in the confidence interval} \\
3 \ln(.5) + 7 \ln(1-(.5)) & = -6.9315 \quad p = .5 \text{ is in the confidence interval} \\
3 \ln(.6) + 7 \ln(1-(.6)) & = -7.9465 \quad p = .6 \text{ is in the confidence interval} \\
3 \ln(.61) + 7 \ln(1-(.61)) & = -8.0741 \quad p = .61 \text{ is not in the confidence interval} \\
3 \ln(.65) + 7 \ln(1-(.65)) & = -8.6411 \quad p = .65 \text{ is not in the confidence interval} \\
3 \ln(.7) + 7 \ln(1-(.7)) & = -9.4978 \quad p = .7 \text{ is not in the confidence interval} \\
\end{align*}
\]

So, the confidence interval is \(.085 \leq p \leq .6\). Cool.

How do we interpret this confidence interval? The confidence interval is a random "variable": it varies from sample to sample; \(.085 \leq p \leq .6\) is the estimated confidence interval for one particular sample. Ninety-five percent of these intervals will contain the true \(p\).

Note again that this procedure generalizes to the case where \(\theta\) is a vector of parameters. Put simply, the maximum likelihood confidence interval for
the parameter vector $\theta$ are those $\theta$ consistent with the restriction $-2(\ln L_\theta - \ln L_{\theta_{ml}}) \leq \chi^2_{crit}$ where $\theta_{ml}$ are the $ml$ parameter estimates.

Note that while the likelihood-ratio based method for hypothesis testing and constructing confidence intervals has many charms, is not the only method available for these tasks. Other methods include Wald statistics and Score statistics. Wald methods are often computationally simpler, particular when $\theta$ has many dimensions. The different methods do not always give the same results.

Confidence intervals generated by the Wald method are the ones typically crunched out by statistical software packages.

A later section of the course more generally discusses interval estimation and hypothesis testing.

1.7 Non-parametric maximum likelihood

see the review questions for max. lik.