

VI. Likelihood and Maximum Likelihood Estimation

- a. The Likelihood Function for Regression Models
- b. Method of Maximum Likelihood
- c. Properties of MLEs
- d. ARCH-M Likelihood and Example
- e. Conclusions Regarding Maximum Likelihood

a. Likelihood

Regression models can be very useful and what we have learned about them provides a strong intuition that will help you learn and interpret more advanced methods.

However, there is one more concept which is important and will enable you to access a vast array of methods.

This concept is called the **likelihood function**.

The best way to understand this idea is to return to the regression model and give it a slight different interpretation.

a. Likelihood

What is the probability distribution of all of the sample data? Given the x values, each observation is independent. This means that the distribution of the data is MVN with a special and simple Variance-Covariance matrix.

$$y_i = x_i' \beta + \varepsilon_i \text{ where } \varepsilon_i \sim \text{iidN}(0, \sigma^2)$$

is the same as saying

$$y_i | x_i \sim N(x_i' \beta, \sigma^2)$$

a. Likelihood

Another way of looking at this is that we have specified a specific MVN distribution for our data.

$$y \sim \text{MVN}(X\beta, \sigma^2 I_n)$$

What then is the probability density of the data?

$$\begin{aligned} p(y | X) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} (y_i - x_i'\beta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i'\beta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)\right\} \end{aligned}$$

a. Likelihood

Viewed as a function of the model parameters, this function is called the likelihood function.

$$\ell(\beta, \sigma^2 \mid y, X) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta) \right\}$$

The **Likelihood Principle** tells us that this function “preserves” or tells everything the data has to say about the model parameters.

How does the data drive the shape of the likelihood function?

a. Likelihood

The only place the data has any effect is in the exponent. If we expand out the inner product of two vectors, we see that the likelihood function is driven by the data only thru sums, sums of squares, and cross-products.

$$(y - X\beta)'(y - X\beta) = y'y - 2y'X\beta + \beta'X'X\beta$$

This says that two datasets with the same values of

$$y'y, y'X, \text{ and } X'X$$

will have the same likelihood function.

b. The Method of Maximum Likelihood

The **Method of Maximum Likelihood** says that we should choose estimators of the unknown parameters so as to maximize the likelihood function.

$$\hat{\beta}_{\text{MLE}}, \hat{\sigma}_{\text{MLE}}^2 = \operatorname{argmax} \left(\ell(\beta, \sigma^2) \right)$$

Why do I care?

1. This is recipe that always works. Write down a model, maximize the likelihood and you are done!
2. MLE's have “good” sampling properties (hard to beat in larger samples).

b. Maximum Likelihood for Regression

Let's find the MLEs for regression model.

How do we maximize the likelihood function?

$$\ell(\beta, \sigma^2 | y, X) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta) \right\}$$

For any value of sigma-sq, the best we can do for beta is to maximize the exponent or minimize

$$(y - X\hat{\beta}_{MLE})' (y - X\hat{\beta}_{MLE})$$

Which means to minimize the sum of squares or least squares is an MLE.

$$\hat{\beta}_{MLE} = b = (X'X)^{-1} X'y$$

b. Maximum Likelihood for Regression

What is the Maximum Likelihood estimator for the variance of the regression errors? Stick the MLE for beta in and then maximize this.

$$\ell(\beta = \hat{\beta}_{\text{MLE}}, \sigma^2 | y, X) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (y - Xb)' (y - Xb) \right\}$$

or

$$\ln(\ell) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} e'e$$

Taking the derivative and setting to zero, we find

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} e'e$$

$$s^2 = \frac{e'e}{n-k}$$

Which is different than the “unbiased” estimator, s^2 , we have been using.

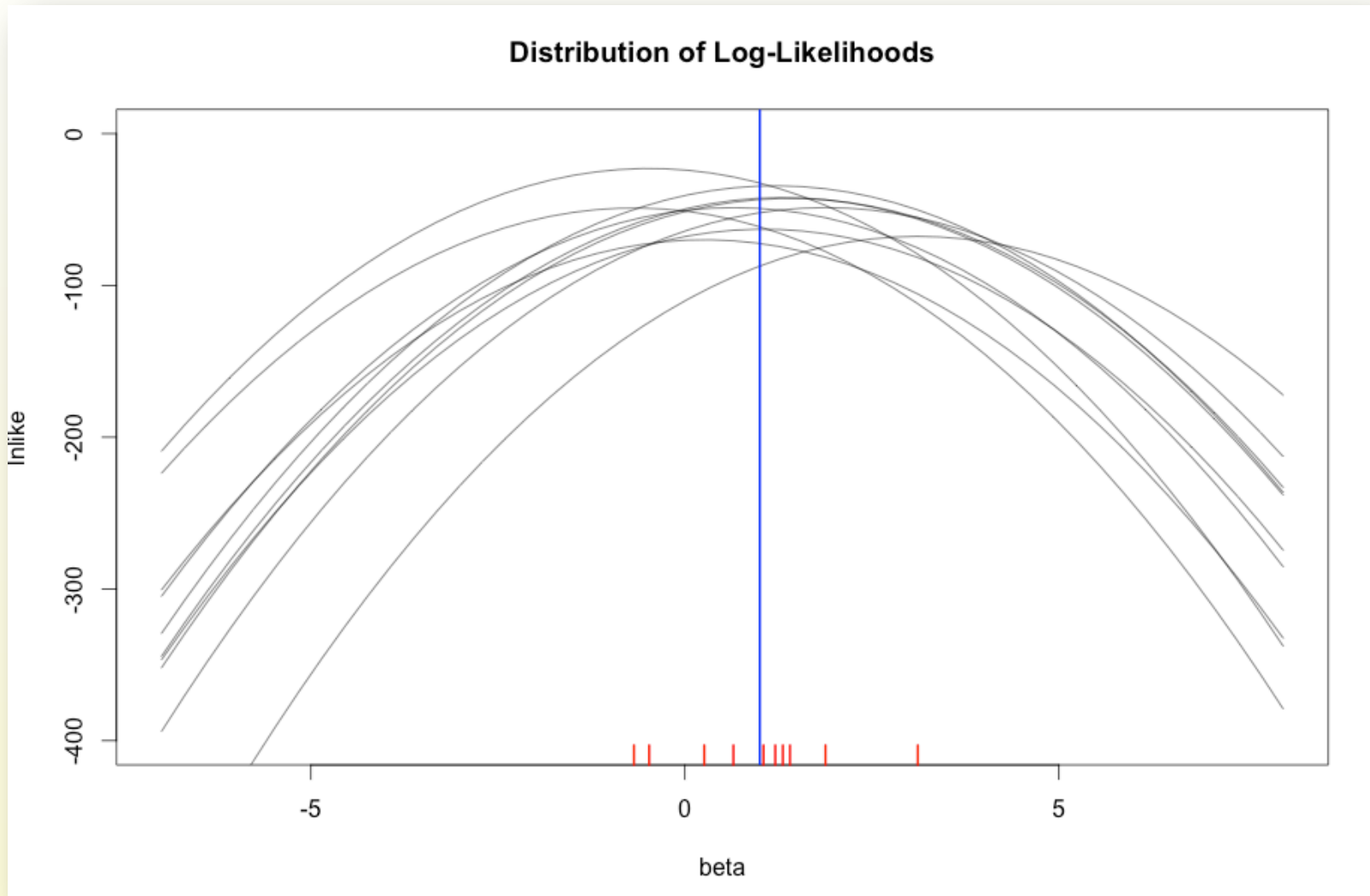
b. Distribution of Maximum Likelihood Estimates

Since the likelihood function depends on the data, the MLEs are functions of the data and therefore are RVs. Just as we can talk about any estimator, the MLEs have a sampling distribution.

Let's consider a simple example where we draw 10 samples from the same population regression model and plot the likelihoods. We will see that each likelihood is centered on the least squares estimate for that sample. We will use a simple model

$$y_i = 1 \times x_i + \varepsilon_i \quad \varepsilon_i \sim N(0, 3^2) \\ i = 1, \dots, 10$$

b. Distribution of Maximum Likelihood Estimates



c. Properties of MLEs

Some natural questions to ask:

1. Are MLE's worth considering? Do they “learn” from the data? That is, if we get more and more data, does the MLE get “closer” to the true value.
2. Are MLE's optimal (they can be biased, but who cares?)?
3. What is the sampling distribution of the MLE?
4. How can we compute MLEs and whatever is necessary to assess the sampling distribution?

c. Properties of MLEs

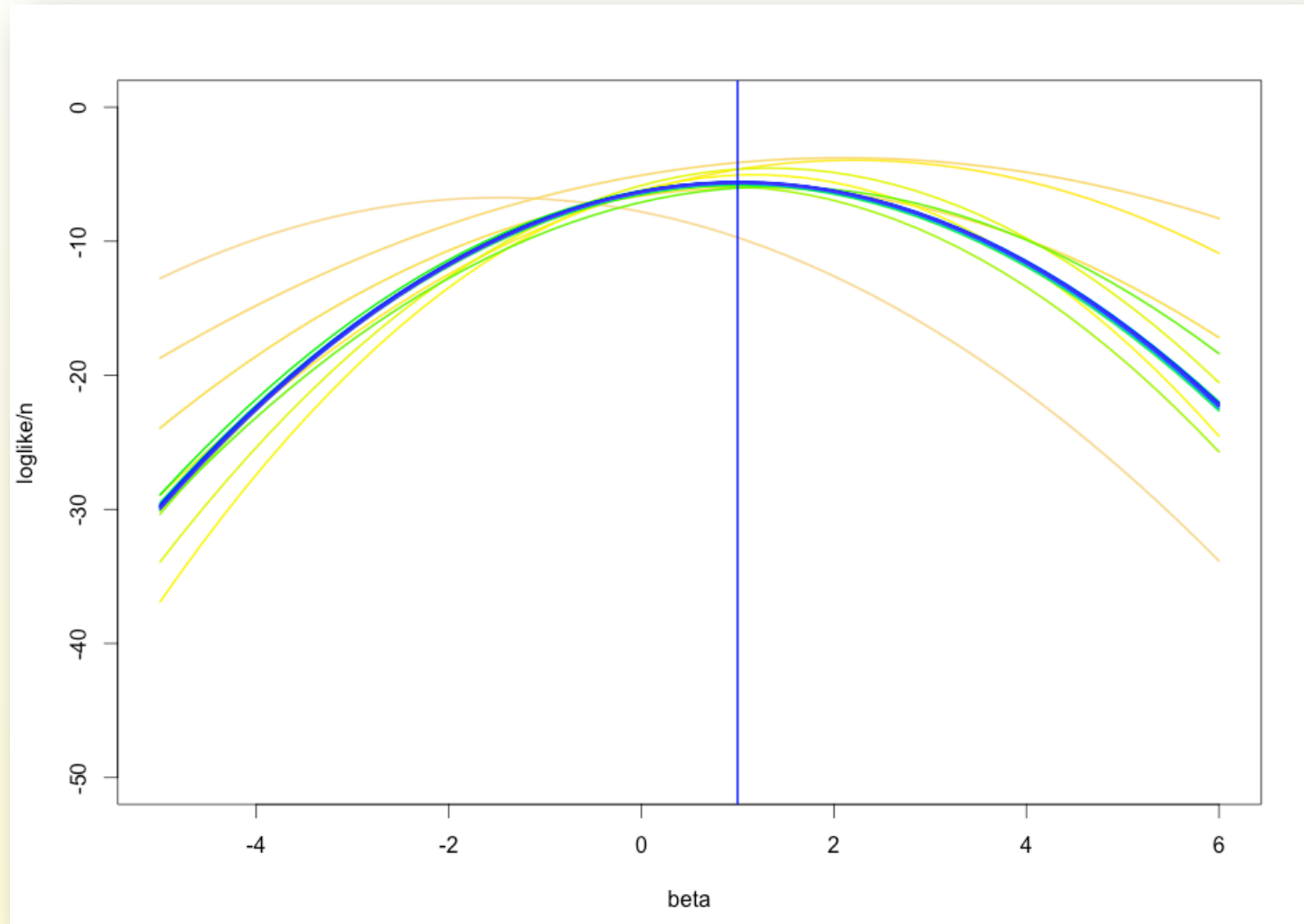
Question 1:

MLEs do learn from more and more data. In general, MLEs are consistent estimators (as n goes to infinity, the sampling distribution collapses on the true value). Let's illustrate that with our regression example.

We will take a sequence of increasing N from 10 to 500. We will then simulate samples of these sizes from the same regression model we considered before and plot the normalized log-likelihood functions for each of these samples. (as N increases the color of the plotted likelihoods transitions from yellow to green to blue – like topographical maps change in color from the lower to higher areas – but the opposite).

The log-likelihoods have maxima closer and closer to the true value of β (1).

c. Properties of MLEs



c. Properties of MLEs

Question 2:

Are MLEs optimal? To define optimal, we must adopt a criterion for optimality. Let's adopt the MSE criterion (remember that from Chapter IV),

$$\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

Then we can say

In large samples, the MLE will have the lowest MSE of all estimators. Some say that the MLE is “asymptotically efficient.” for this reason.

When we say “large” samples, we mean arbitrarily large samples (that is infinite size samples). This is called as “asymptotic” statement from the word “asymptote” which means that this is a limit as n gets large and may not be true for any fixed n .

c. Properties of MLEs

Question 3: What is the Sampling Distribution of the MLE?

Let's look at the log-likelihood for the regression model holding sigma fixed (or assuming it is known). What is the curvature of the log-likelihood surface?

$$\ell(\beta, \sigma^2 | y, X) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta) \right\}$$

$$L(\beta | y, X, \sigma^2) = \ln(\ell) \propto -\frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta)$$

$$(y - X\beta)' (y - X\beta) = (y - Xb)' (y - Xb) + (\beta - b)' X'X(\beta - b)$$

The curvature of a function is the second derivative. In this case we have a matrix of second derivatives (often called the Hessian).

c. Properties of MLEs

The curvature of a function is the second derivative. In this case we have a matrix of second derivatives (often called the Hessian).

$$H = \begin{bmatrix} \frac{\partial^2 L}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 L}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 L}{\partial \beta_1 \partial \beta_k} \\ \frac{\partial^2 L}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 L}{\partial \beta_2 \partial \beta_2} & \cdots & \frac{\partial^2 L}{\partial \beta_2 \partial \beta_k} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 L}{\partial \beta_k \partial \beta_1} & \frac{\partial^2 L}{\partial \beta_k \partial \beta_2} & \cdots & \frac{\partial^2 L}{\partial \beta_k \partial \beta_k} \end{bmatrix}$$

c. Properties of MLEs

For the regression model,

$$H = -\frac{1}{\sigma^2} X'X$$

The matrix of second derivatives is negative definite everywhere which means that the log-likelihood of regression model is globally concave.

Note that $-H^{-1} = \sigma^2 (X'X)^{-1}$ which is the var-covariance matrix of the least squares estimators.

This suggests a more general property of MLEs:

c. Properties of MLEs

Question 4: how to compute MLE and it's approximate distribution

$$\hat{\theta}_{\text{MLE}} \approx N\left(\theta, -H^{-1} \Big|_{\theta=\hat{\theta}_{\text{MLE}}}\right)$$

We compute the MLE by using an optimizer based and then use the Hessian to compute standard errors.

Optimizers will require a function to evaluate the log-likelihood of the data and will often return both the parameter vector which maximizes the likelihood as well as an estimate of the Hessian.

d. ARCH Models

Let's try using these ideas for an important class of time series models called, ARCH models (Auto-Regressive Conditional Heteroskedasticity).



These models were invented by Rob Engle to capture predictability not in the level or mean of a time series but predictability in the conditional variances. See

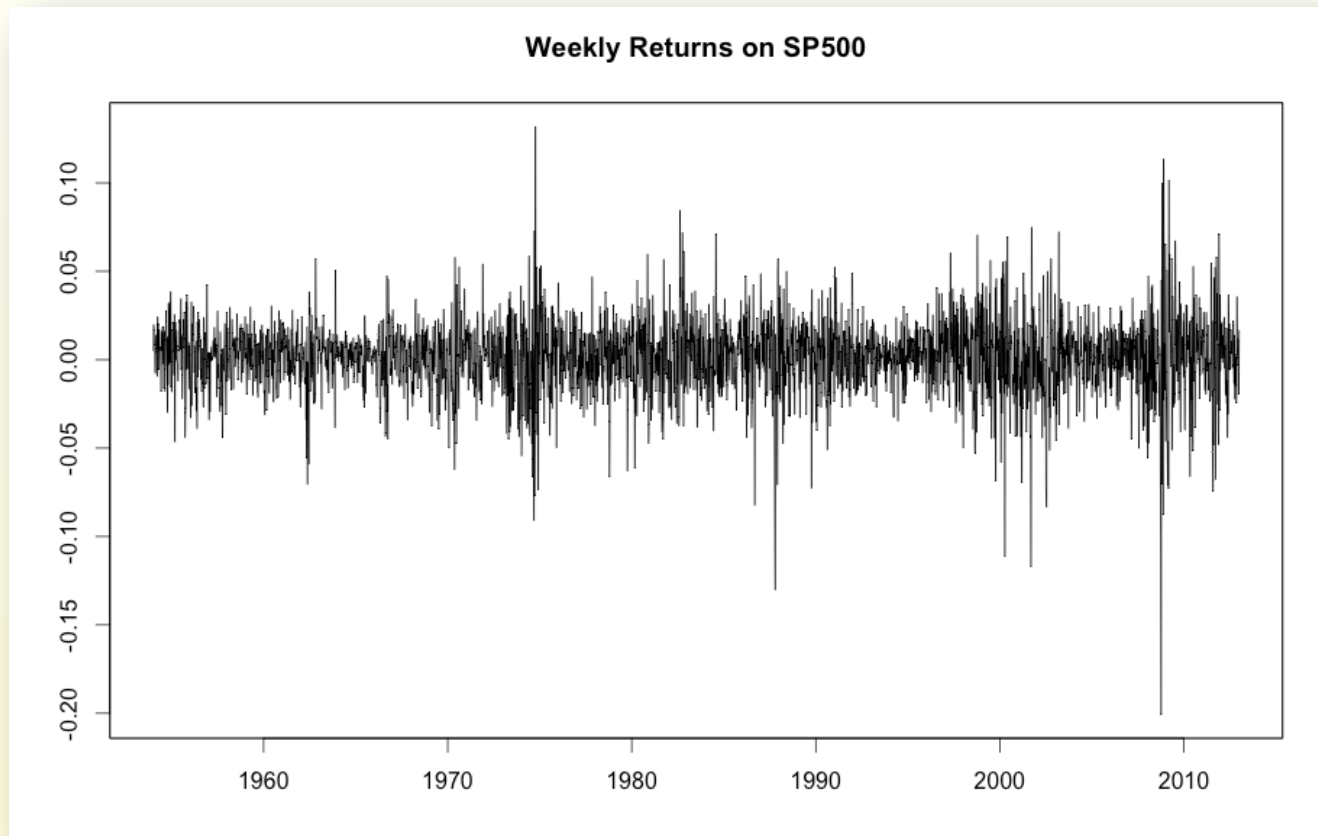
<http://www.r-bloggers.com/a-practical-introduction-to-garch-modeling/>

for more on ARCH modeling in R.

Engle and others noted that while the ACF of many returns series is pretty small, the ACF of the squared series is much larger. Also noted was what people call “volatility clustering” meaning that these series go thru periods with high volatility followed by periods of low volatility or that there is persistence in volatility.

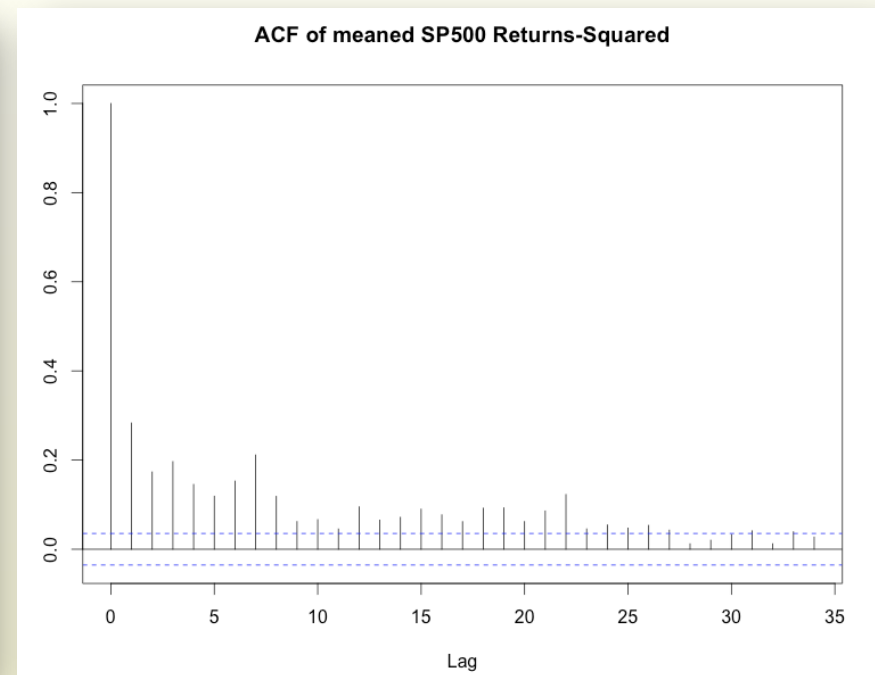
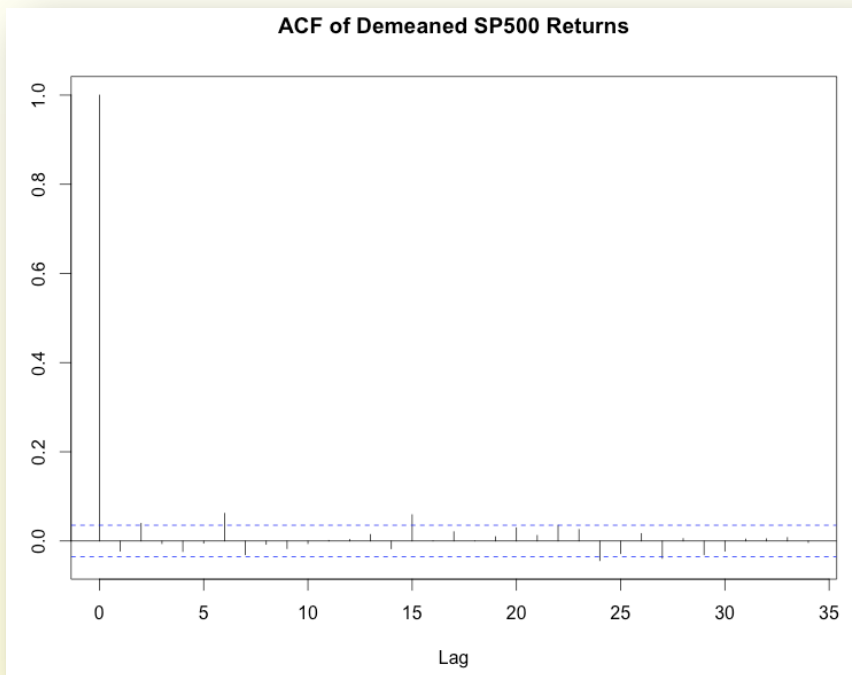
d. ARCH Models

Let's look at weekly returns on the S&P 500. See the volatility clustering!



d. ARCH Models

Now let's look at ACF of demeaned returns and demeaned returns squared.



d. ARCH Models

How do we model this? What we are seeing is the variance of series at time t , given the past is predictable.

ARCH-M

$$a_t = y_t - \mu_t$$

$$a_t = \sigma_t \epsilon_t \quad \epsilon_t \sim N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \dots + \alpha_M a_{t-M}^2$$

We will just demean the series and use this directly in modeling (we will not estimate the conditional mean).

d. ARCH Likelihood

Remember the likelihood function is the joint distribution of the observations.

$$\ell(\alpha) = p(a_1, a_2, \dots, a_T | \alpha)$$

This is not equal to the product of the densities for each observation like the regression problem because the observations are correlated in the ARCH model (thru the conditional variances). We can write down the likelihood using a fundamental idea in time series.

$$p(y_1, \dots, y_T) = p(y_1)p(y_2 | y_1)p(y_3 | y_1, y_2) \cdots p(y_T | y_{T-1}, \dots, y_1)$$

d. ARCH Likelihood

But in the case of an ARCH-M, these dependence only extends M periods into the past, e.g.

$$p(y_t | y_{t-1}, \dots, y_1) = p(y_t | y_{t-1}, \dots, y_{t-M})$$

Thus, we write the ARCH-M likelihood as

$$\begin{aligned} \ell(\alpha) = & \overset{\text{one}}{p(a_1, \dots, a_M | \alpha)} p(a_{M+1} | a_1, \dots, a_M, \alpha) \\ & \cdots p(a_T | a_{T-1}, \dots, a_{T-M}, \alpha) \end{aligned} \quad \overset{T-M}{}$$

We ignore the first terms and “condition” on the first M observations of the a series.

d. ARCH Likelihood

The ARCH-M specifies that

$$p(a_t | a_{t-1}, \dots, a_{t-M}, \alpha) \sim N(0, \sigma_t^2)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \dots + \alpha_M a_{t-M}^2$$

So the likelihood function can be written as $\ell(\alpha) = \prod_{t=M+1}^T \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{1}{2\sigma_t^2} a_t^2\right)$

Or in logs


$$L(\alpha) \propto \sum_{t=M+1}^T -.5 \log(\sigma_t^2) - .5 \frac{a_t^2}{\sigma_t^2}$$

d. ARCH Likelihood

Let's first simulate some data from an ARCH-M and try MLE on it!

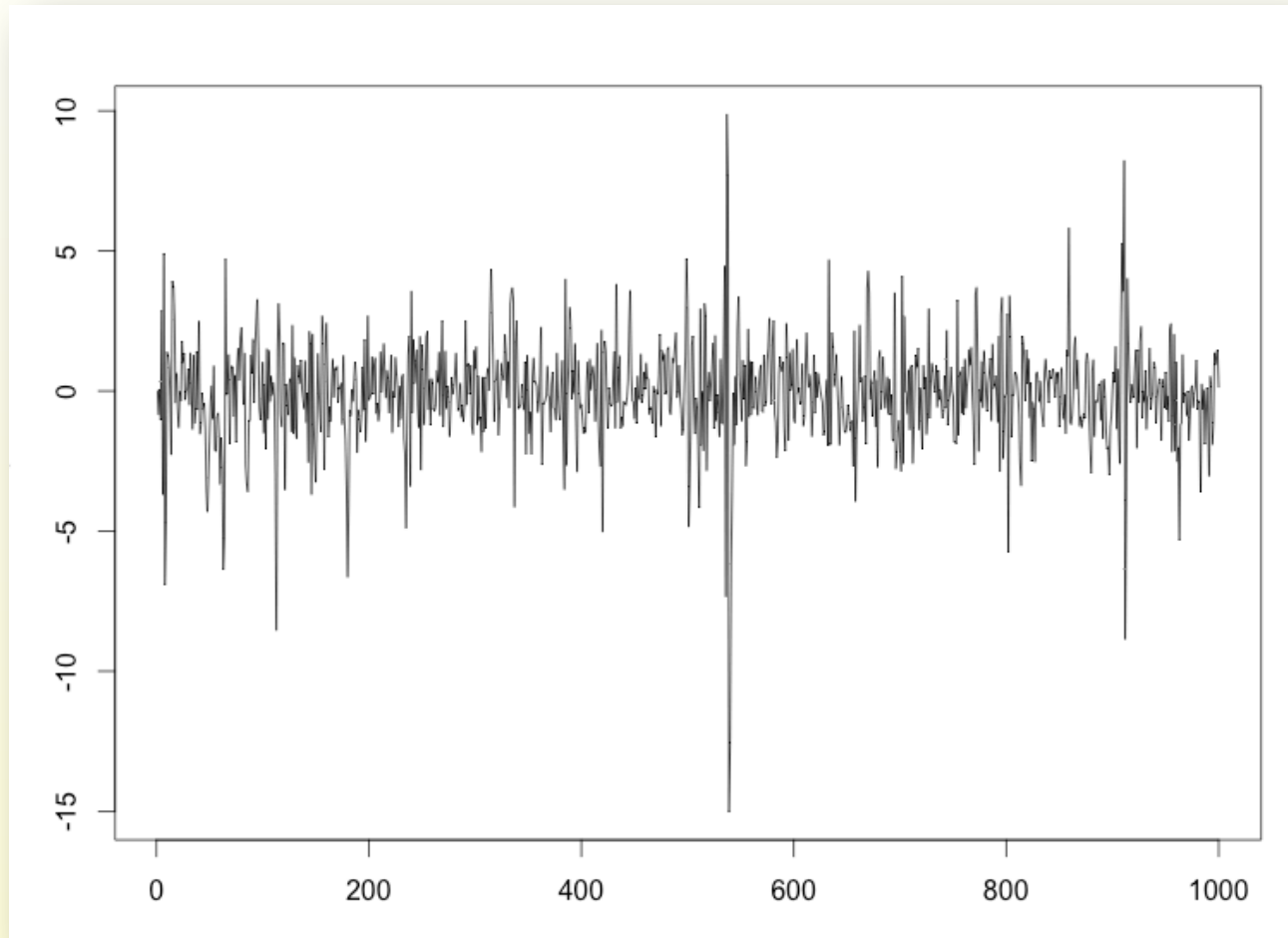
```
> sim.ARCHM=function(T, inita,alpha){
+   # function to simulate ARCHM
+   sigmasqt=function(atlag,alpha){
+     M=length(atlag)
+     return(alpha[1]+sum(alpha[2:(M+1)]*atlag**2))
+   }
+   M=length(inita)
+   a=double(T+M)
+   a[1:M]=inita
+   for(t in (M+1):(T+M)){
+     sd= sqrt(sigmasqt(a[(t-1):(t-M)],alpha))
+     a[t]=rnorm(1,sd=sd)
+   }
+   return(a[(M+1):(T+M)])
+ }
>
> alpha=c(1,.7)
> inita=c(0)
> a=sim.ARCHM(1100,inita,alpha)[101:1100]
> plot(a,type="l")
```

True
Values



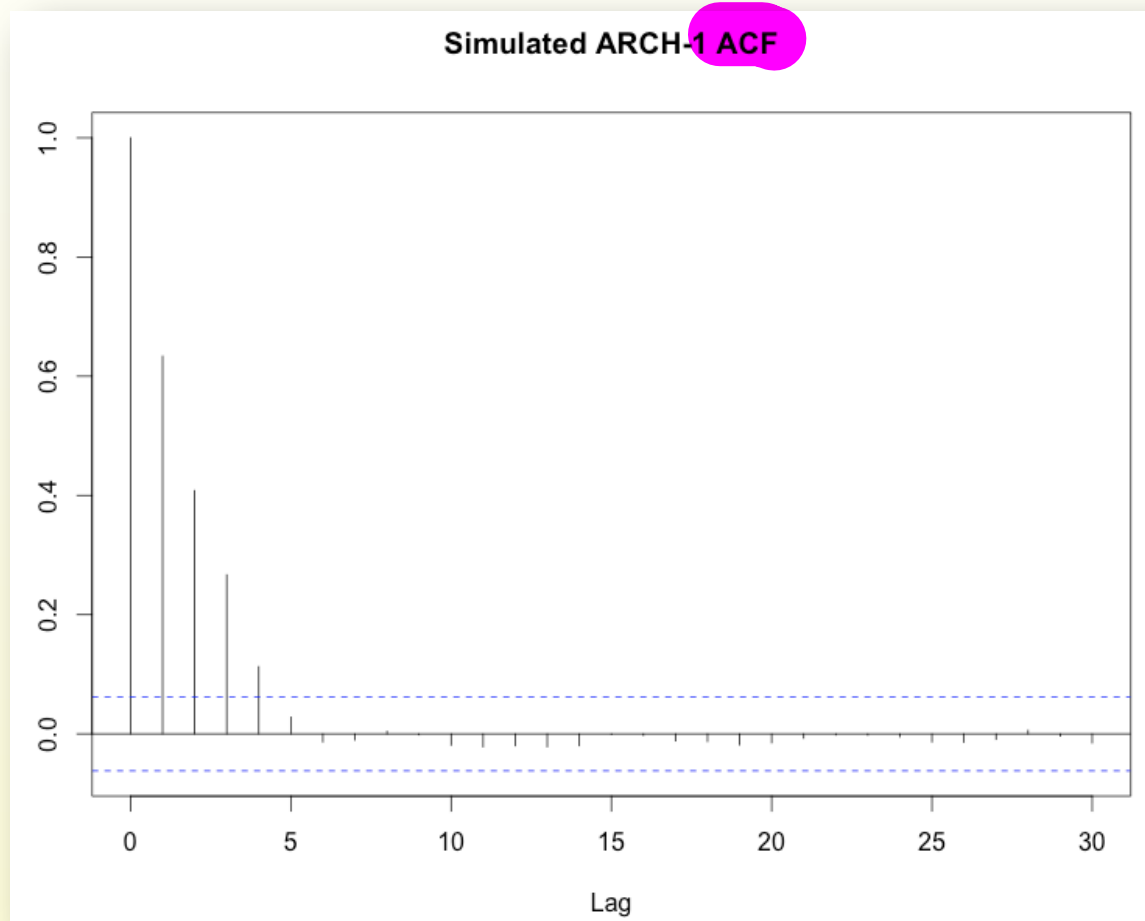
d. ARCH Likelihood

The results clearly show volatility clustering!



d. ARCH Likelihood

The results clearly show volatility clustering!



d. ARCH Likelihood

Now write a function to evaluate log-like for ARCH-M and test on simulated data.

```
L.ARCHM=function(alpha,a,M){
  sigmasqt=function(atlag,alpha){
    M=length(atlag)
    return(alpha[1]+sum(alpha[2:(M+1)]*atlag**2))
  }
  T=length(a)
  # conditional log-like for ARCH(M)
  # a_t=sigma_t x N(0,1)
  # sigma_t=alpha0 + sum(alpha_i*a_t**2
  #
  L=0
  for(t in (M+1):T){
    atlag=a[(t-1):(t-M)]
    stsqt=sigmasqt(atlag,alpha)
    if(stsqt <= 0) {stsqt=.000000001}
    L=L-.5*log(stsqt)-.5*(a[t]**2/stsqt)
  }
  return(L)
}
```

d. ARCH Likelihood

To run this, we need to use an optimizer. Let's use the R, optim() function.

```
> alpha=c(.5,.5)
>
> mle = optim(alpha, L.ARCHM, a=a, M = 1, method = "BFGS", hessian = TRUE,
+           control = list(fnscale = -1))
>
```

```
> mle
```

```
$par
```

```
[1] 1.0180217 0.7033198
```

```
$value
```

```
[1] -864.141
```

```
$counts
```

```
function gradient
```

```
39      10
```

```
$convergence
```

```
[1] 0
```

Converged!

```
$message
```

```
NULL
```

```
$hessian
```

```
      [,1]      [,2]
[1,] -219.3003 -102.1634
[2,] -102.1634 -254.5774
```

```
> std_err_Hess=sqrt(diag(chol2inv(chol(-mle$hessian))))
> std_err_Hess
[1] 0.07488977 0.06950757
```

The MLES are
(1.02, .70) with
Std Errors
(.075,.070)

Pretty,
Pretty,Pretty
Good as Larry
David says

d. ARCH Likelihood

Now, let's try it out on the SP500 data. Let's try a moderately large value of M. M too small? See GARCH model!

```
> alpha=c(1,.1,.1,.1)
>
> mle = optim(alpha, L.ARCHM, a=ret500_demeaned*100, M = 3, method = "BFGS", hessian = TRUE,
+           control = list(fnscale = -1))
>
> mle
$par
[1] 1.9910117 0.2291934 0.1402737 0.1933998

$value
[1] -3568.495

$counts
function gradient
      92      24

$convergence
[1] 0

> std_err_Hess=sqrt(diag(chol2inv(chol(-mle$hessian))))
> std_err_Hess
[1] 0.10885540 0.02654079 0.02591075 0.02748666
```


d. Conclusions Regarding MLEs

The Method of Maximum Likelihood is a very powerful method that will deliver relatively good results for any model which can be written down as specifying the joint density of the observations.

Caution: some likelihoods may be very hard to maximize! Some have multiple maxima. Others have various singularities, poles and directions of recession that may make optimization a challenge.

There are some models whose likelihoods exist but are computationally very challenging to evaluate!

Always use an industrial strength optimizer. If possible, provide the gradient of the objective function analytically.

Continuing the Learning Experience

Our course has given you the large tool set, intuition and knowledge of how to apply the tools.

You have a resource base that you can return to via the course notes and code snippets

You have a strong intuitive basis for learning new methods and techniques

My commitment to you:

I am always willing to act as a resource for problems you meet on the job.

Your commitment to yourself:

I will keep learning!

Fini

You are now *certified* to do regression analysis at a very high level of proficiency and competence!



Important Equations

$$\hat{\theta}_{MLE} \approx N\left(\theta, -H^{-1} \middle|_{\theta=\hat{\theta}_{MLE}}\right)$$

Distribution of the
MLE

$$a_t = y_t - \mu_t$$

$$a_t = \sigma_t \epsilon_t \quad \epsilon_t \sim N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \dots + \alpha_M a_{t-M}^2$$

ARCH-M model

definition of
standardized
residual

Glossary of R Commands

- `optim(parameter, objective_function, ...)`: non-linear optimizer built-in to R.