Applied Time Series Analysis

Vipul Bhatt

2019-08-20

Contents

Preface

These lecture notes are prepared for an upper level undergraduate course in time series econometrics. Every fall I teach a course on applied time series analysis at James Madison University. These notes borrow heavily from the teaching material that I have developed over several years of instruction of this course.

One of my main objective is to develop a primer on time series analysis that is more accessible to undergraduate students than standard textbooks available in the market. Most of these textbooks in my opinion are densely written and assume advanced mathematical skills on the part of our students. Further, I have also struggled with their topic selection and organization. Often I end up not following the chapters in order and modify content (by adding or subtracting) to meet my students needs. Such changes causes confusion for some students and more importantly discourages optimal use of the textbook. Hence, this is an undertaking to develop a primer on time series that is accessible, follows a more logical sequencing of topics, and covers content that is most useful for undergraduate students in business and economics.

Note: These notes have been prepared by me using various sources, published and unpublished. All errors that remain are mine.

Chapter 1

Introduction to Forecasting

1.1 Time Series

A time series is a specific kind of data where observations of a variable are recorded over time. For example, the data for the U.S. GDP for the last 30 years is a time series data.

Such data shows how a variable is changing over time. Depending on the variable of interest we can have data measured at different frequencies. Some commonly used frequencies are intra-day, daily, weekly, monthly, quarterly, semi-annual and annual. Figure 1.1 below plots data for quarterly and monthly frequency.

The first panel shows data for the real gross domestic product (GDP) for the US in billions of 2012 dollars, measured at a quarterly frequency. The second panel shows data for the advance retail sales (millions of dollars), measured at monthly frequency.

Formally, we denote a time series variable by y_t, where $t = 0, 1, 2, .., T$ is the observation index. For example, at $t = 10$ we get the tenth observation of this time series, y_{10}.

1.2 Serial Correlation

Serial correlation (or auto correlation) refers to the tendency of observations of a time series being correlated over time. It is a measure of the temporal dynamics of a time series and addresses the following question: what is the effect of past realizations of a time series on the current period value? Formally,

$$\rho(s) = Cor(y_t, y_{t-s}) = \frac{Cov(y_t, y_{t-s})}{\sqrt{\sigma^2_{y_t} \times \sigma^2_{y_{t-s}}}} \tag{1.1}$$

where $Cov(y_t, y_{t-s}) = E(y_t - \mu_{y_t})(y_{t-s} - \mu_{y_{t-s}})$ and $\sigma^2_{y_t} = E(y_t - \mu_{y_t})^2$

Here, $\rho(s)$ is the serial correlation of order s. For example, $s = 1$ implies *first order* serial correlation between y_t and y_{t-1}, $s = 2$ implies *second order* serial correlation between y_t and y_{t-2}, and so on.

Note that often we use historical data to forecast. If there is no serial correlation, then past can offer no guidance for the present and future. In that sense, presence of serial correlation of some order is the first condition for being able to forecast a time series using its historical realizations.

Now, we can either have positive or negative serial correlation in data. Figure **??** plots two time series with positive and negative serial correlation, respectively.

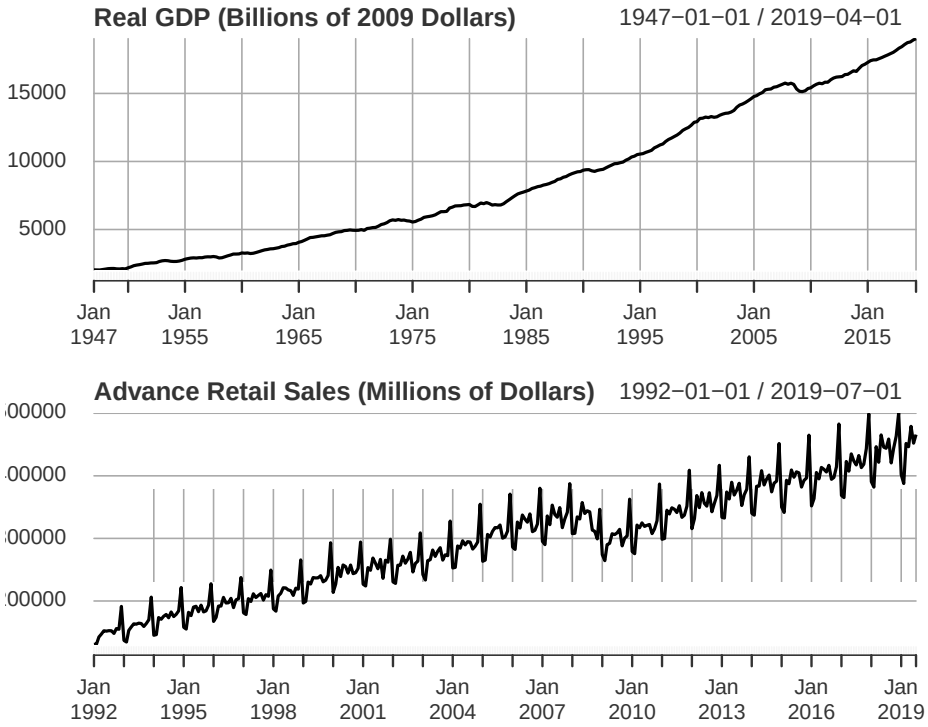


Figure 1.1: Time Series at quarterly and monthly frequency

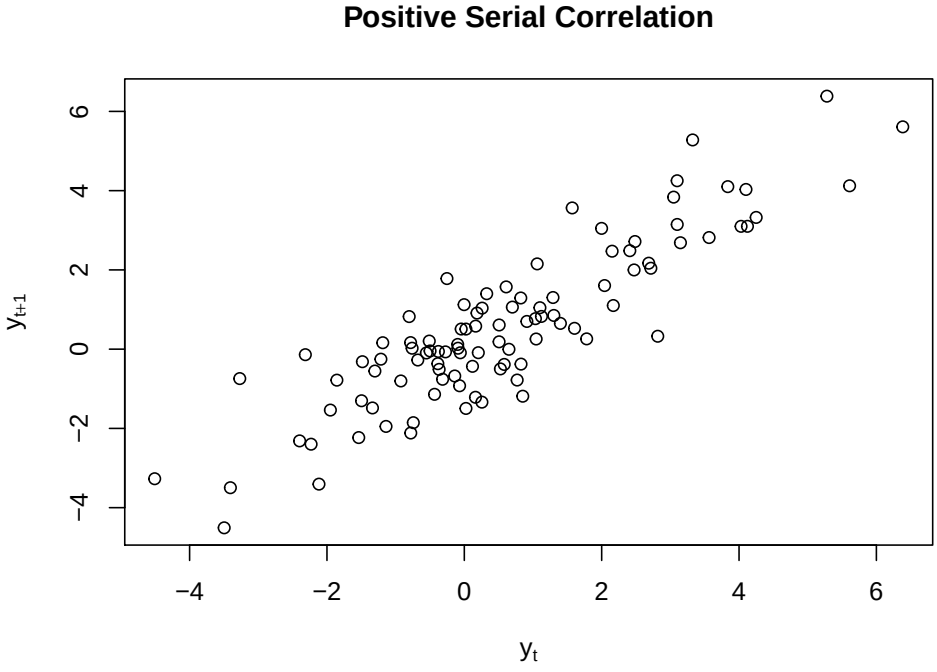


Figure 1.2: Serial Correlation

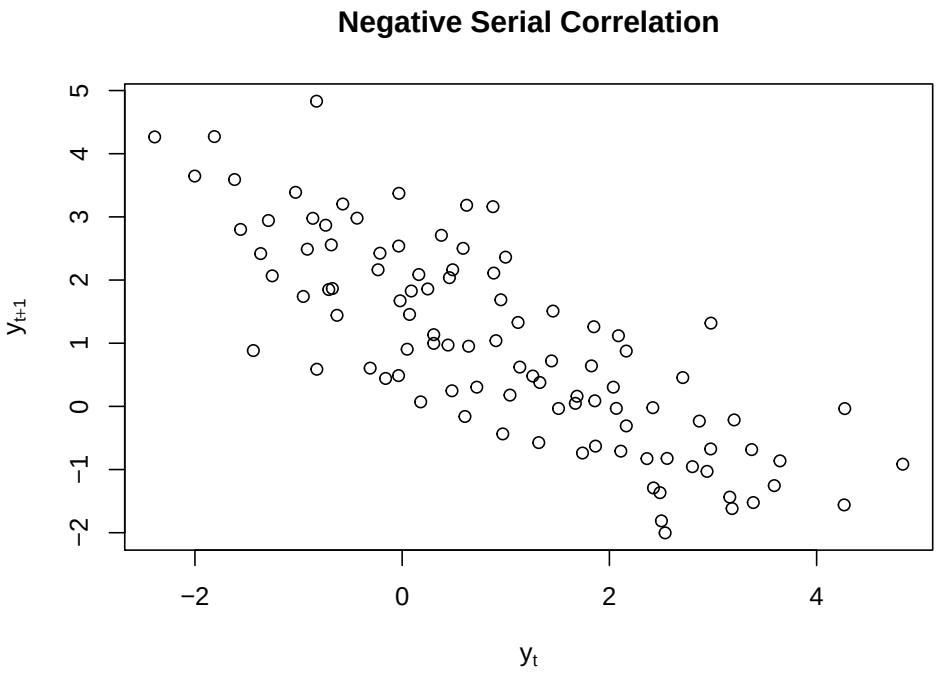


Figure 1.3: Serial Correlation

1.3 Testing for Serial Correlion

We can use a Lagrange-Multiplier (LM) test for detecting serial correlation. This test is also known as *Breuch-Godfrey* test. I will use the linear regression model to explain this test. Consider the following regression model:

$$y_t = \beta_0 + \beta_1 X_{1t} + \epsilon_t \tag{1.2}$$

Consider the following model for serial correlation of order p for the error term:

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + ... + \rho_p \epsilon_{t-p} + \nu_t \tag{1.3}$$

Then we are interested in the following test:

$$H_0 = \rho_1 = \rho_2 = ... = \rho_p = 0$$
$$H_A = Not\ H_0$$

To implement this test, we estimate the BG regression model given by:

$$e_t = \alpha_0 + \alpha_1 X_{1t} + \rho_1 e_{t-1} + \rho_2 e_{t-2} + ... + \rho_p e_{t-p} + \nu_t \tag{1.4}$$

where we replace the error term with the OLS residuals (denoted by e). The LM test statistic is given by:

$$LM = N \times R^2_{BG} \sim \chi^2_p$$

If the test statistic value is greater than the critical value then we reject the null hypothesis.

1.4 White Noise Process

A time series is a *white noise* process if it has zero mean, constant and finite variance, and is serially uncorrelated. Formally, y_t is a white noise process if:

1. $E(y_t) = 0$
2. $Var(y_t) = \sigma_y^2$
3. $Cov(y_t, y_{t-s}) = 0 \forall s \neq t$

We can compress the above definition as: $y_t \sim WN(0, \sigma_y^2)$. Often we assume that the unexplained part of a time series follows a white noise process. Formally,

$$Time\ Series\ =\ Explained\ +\ White\ Noise \tag{1.5}$$

By definition we cannot forecast a white noise process. An important diagnostics of model adequacy is to test whether the estimated residuals are white noise (more on this later).

1.5 Important Elements of Forecasting

Definition 1.1 (Forecast)**.**

A *forecast* is an *informed* guess about the unknown future value of a time series of interest. For example, what is the stock price of Facebook next Monday?

There are three possible types of forecasts:

1. *Density Forecast*: we forecast the entire probability distribution of the possible future value of the time series of interest. Hence,

$$F(a) = P[y_{t+1} \leq a] \tag{1.6}$$

give us the probability that the 1-period ahead future value of y_{t+1} will be less than or equal to a. For example, the future real GDP growth could be normally distributed with a mean of 1.3% and a standard deviation of 1.83%. Figure 1.4 below plots the density forecast for real GDP growth.

2. *Point Forecast*: our forecast at each horizon is a single number. Often we use the expected value or mean as the point forecast. For example, the point forecast for the 1-period ahead real GDP growth can be the mean of the probability distribution of the future real GDP growth:

$$f_{t,1} = 1.3 \tag{1.7}$$

3. *Interval Forecast*: our forecast at each horizon is a range which is obtained by adding *margin of errors* to the point forecast. With some probability we expect our future value to fall withing this range. For example, the 95% interval forecast for the next period real GDP growth is (-2.36%,4.96%). Hence, with 95% confidence we expect next period GDP to fall between -2.36% and 4.96%.

Definition 1.2 (Forecast Horizon)**.**

Forecast Horizon is the number of periods into the future for which we forecast a time series. We will denote it by h. Hence, for $h = 1$, we are looking at 1-period ahead forecast, for $h = 2$ we are looking at 2-period ahead forecast and so on.

Formally, for a given time series y_t, the h-period ahead unknown value is denoted by y_{t+h}. The forecast of this value is denoted $f_{t,h}$.

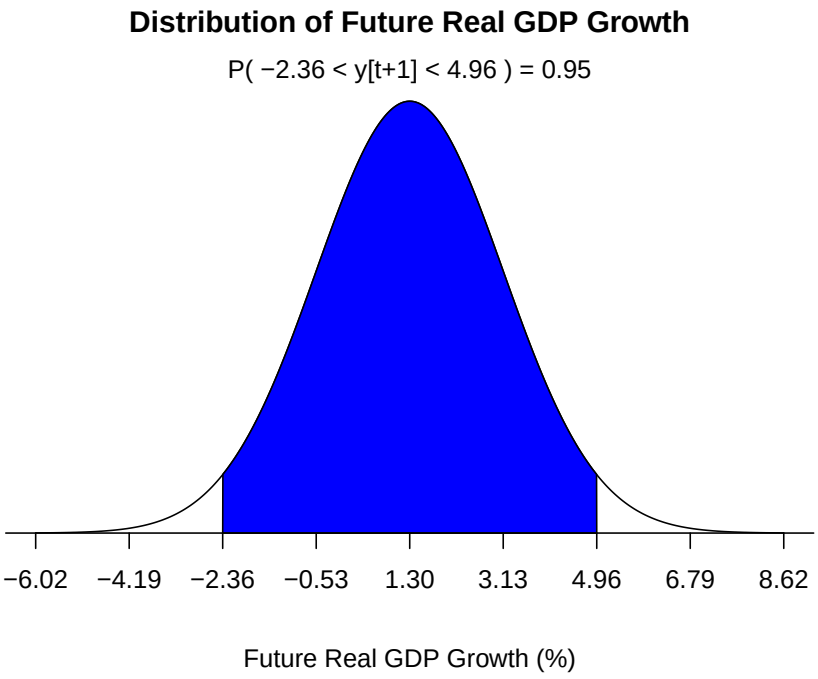


Figure 1.4: Density Forecast for Future Real GDP Growth

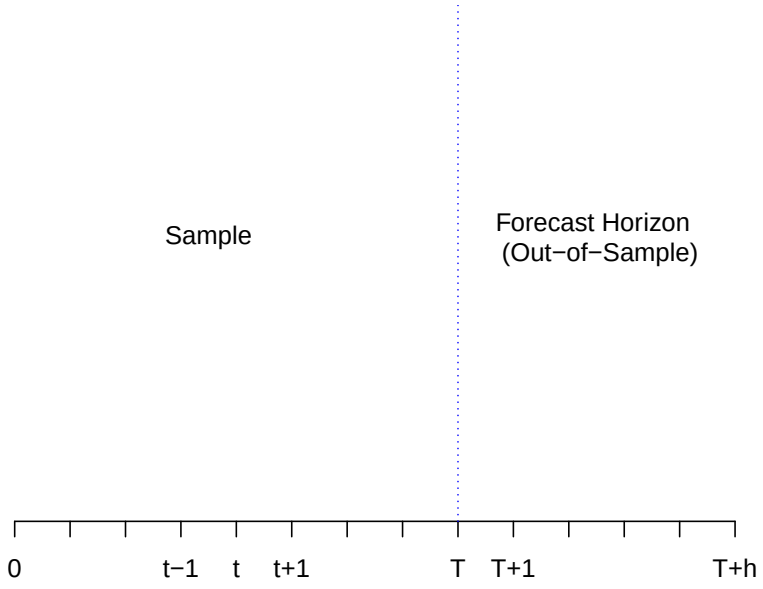


Figure 1.5: Forecast Horizon

Definition 1.3 (Forecast Error)**.**

A *forecast error* is the difference between the realization of the future value and the previously made forecast. Formally, the h-period ahead forecast error is given by:

$$e_{t,h} = y_{t+h} - f_{t,h} \tag{1.8}$$

Hence, for every horizon, we will have a forecast and a corresponding forecast error. These errors can be negative (indicating over prediction) or positive (indicating under prediction).

Definition 1.4 (Information Set)**.**

Forecasts are based on *information* available at the time of making the forecast. *Information Set* contains all the relevant information about the time series we would like to forecast. We denote the set of information available at time T by Ω_T. There are two types of information sets:

1. Univariate Information set: Only includes historical data on the time series of interest:

$$\Omega_T = \{y_T, y_{T-1}, y_{T-2}, ..., y_1\} \tag{1.9}$$

2. Multivariate Information set: Includes historical data on the time series of interest as well as any other variable(s) of interest. For example, suppose we have one more variable x that is relevant for forecasting y. Then:

$$\Omega_T = \{y_T, x_T, y_{T-1}, x_{T-1}, y_{T-2}, x_{T-2}....., y_1, x_1\} \tag{1.10}$$

1.6 Loss Function and Optimal Forecast

Think of a forecast as a solution to an *optimization* problem. When forecasts are wrong, the person making the forecast will suffer some *loss*. This loss will be a function of the magnitude as well as the sign of the *forecast error*. Hence, we can think of an *optimal forecast* as a solution to a minimization problem where the forecaster is minimizing the loss from the forecast error.

Definition 1.5 (Loss Function)**.**

A *loss* function is a mapping between forecast errors and their associated losses. Formally, we denote the h-period ahead loss function by $L(e_{t,h})$. For a function to be used as a loss function, three properties must be satisfied:

1. $L(0) = 0$
2. $\frac{dL}{de} > 0$
3. $L(e)$ is a continuous function.

Two types of loss functions are:

- Symmetric Loss Function: both positive and negative forecast errors lead to same loss. See Figure 1.6. A commonly used loss function is *quadratic loss function* given by:

$$L(e_{t,h}) = e_{t,h}^2 = (y_{t+h} - ft, h)^2 \tag{1.11}$$

- Asymmetric Loss Function: loss depends on the sign of the forecast error. For example, it could be that positive errors produce greater loss when compared to negative errors. See the function below and Figure 1.7 that attaches a higher loss to positive errors:

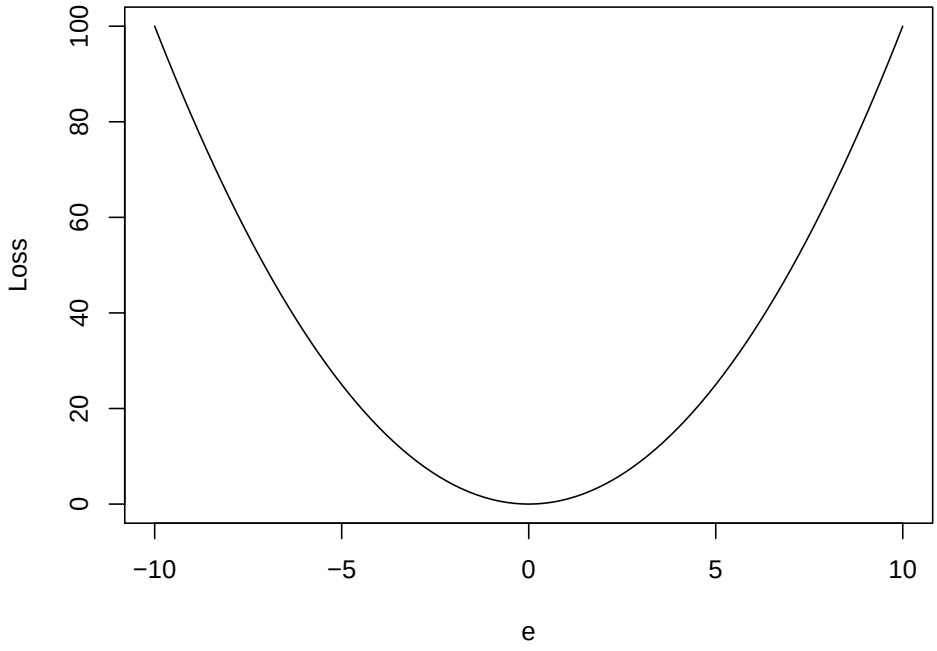


Figure 1.6: Quadratic Loss Functions

$$L(e_{t,h}) = e_{t,h}^2 + 4 \times e_{t,h} \tag{1.12}$$

Once we have chosen our loss function, the optimal forecast can be obtained by minimizing the expected loss function.

Definition 1.6 (Optimal Forecast)**.**

An *optimal forecast* minimizes the expected loss from the forecast, given the information available at the time. Mathematically, we denote it by $f_{t,h}^*$ and it solves the following minimization problem:

$$min_{f_{t,h}} E(L(e_{t,h})|\Omega_t) \tag{1.13}$$

In theory we can assume any functional form for the loss function and that will lead to a different *optimal forecast*. An important result that follows from a specific functional form is stated as Theorem 1.1.

Theorem 1.1. *If the loss function is quadtratic then the optimal forecast is the conditional mean of the time series of interest. Formally, if* $L(e_{t,h}) = e_{t,h}^2$ *then,*

$$f_{t,h}^* = E(y_{t+h}|\Omega_t) \tag{1.14}$$

Note that $E(e_{t,h}^2)$ is known as *mean squared errors (MSE)*. Hence, the expected loss from a quadratic loss function is the same as the MSE. In this course, we assume that the forecaster faces a quadratic loss function and hence based on Theorem 1.1, we will learn different models for estimating the conditional mean of the future value of the time series of interest, i.e., $E(y_{t+h}|\Omega_t)$.

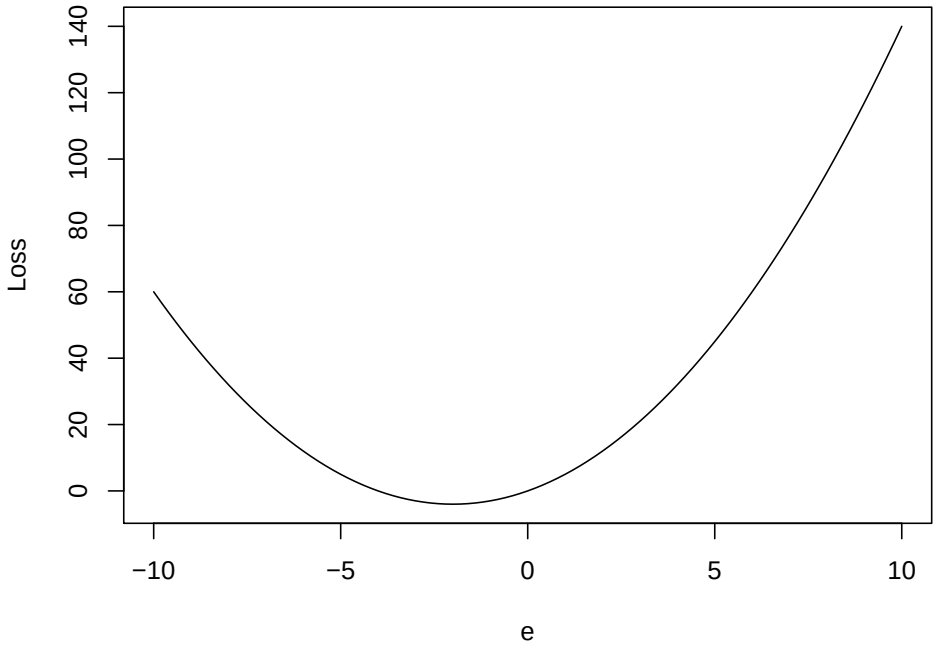


Figure 1.7: Asymmetric Loss Function

Chapter 2

Regression-based Forecasting

One way to compute the conditional expectation is the linear regression model. Here, our information set contains data on all relevant explanatory variables available at the time of forecast, i.e,

$$\Omega_t = X_{1t}, X_{2t}, ...X_{Kt} \tag{2.1}$$

Hence, we get the following equality:

$$E(y_t|\Omega_t) = E(y_t|X_{1t}, X_{2t}, X_{3t}, ..., X_{Kt}) \tag{2.2}$$

The right hand side of the above equation is the multiple regression model of the form:

$$y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + .. + \beta_K X_{Kt} + \epsilon_t \tag{2.3}$$

We can easily estimate the above model using Ordinary Least Squares (OLS) and compute the *predicted value* of y:

$$\widehat{y}_t = \widehat{\beta_0} + \widehat{\beta_1} X_{1t} + \widehat{\beta_2} X_{2t} + ... + \widehat{\beta_k} X_{Kt} \tag{2.4}$$

The above equation can be used to compute the optimal forecast. Suppose, we are interested in computed the h period ahead forecast for y. Then, using the above equation we get:

$$\widehat{y}_{t+h} = \widehat{\beta_0} + \widehat{\beta_1} X_{1t+h} + \widehat{\beta_2} X_{2t+h} + ... + \widehat{\beta_k} X_{Kt+h} \tag{2.5}$$

2.1 Scenario Analysis and Conditional Forecasts

One way to use a regression model to produce forecasts is called *scenario analysis* where we produce a different forecast for the dependent variable under each possible scenario about the future values of the independent variables. For example, what will be the forecast for inflation if the Federal Reserve Bank raises the interest rate? Would our forecast differ depending on the size of the increase in the interest rate?

2.2 Unconditional Forecasts

An alternative is to separately forecast each independent variable and then compute the forecast for the dependent variable. Yet another alternative is to use lagged variables as independent variables. Depending on the number of lags, we can forecast that much ahead into future (see Distributed Lag Section for details).

2.3 Some practical issues

1. To forecast the dependent variable we first need to compute a forecast for the independent variable. Errors in this step induce errors later.
2. *Spurious regression*: It is quite possible to find a strong linear relationship between two completely unrelated variables over time if they share a common time trend.
3. *Model Uncertainty*: We do not know the true functional form for the regression model and hence our estimated model is only a proxy for the true model.
4. *Parameter Uncertainty*: This kind of forecast uses regression coefficients that are computed using a fixed sample. Over time with new data, there will be changes in these coefficients.

2.4 Distributed Lag Regression Models

Consider the following simple regression model:

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \tag{2.6}$$

Here, if want to forecast y_{t+1} then we must either consider different scenarios for x_{t+1} or independently forecast x_{t+1} first, and then use it to compute forecast for y_{t+1}. An alternative is to estimate the following lagged regression model:

$$y_t = \beta_0 + \beta_1 x_{t-1} + \epsilon_t \tag{2.7}$$

Note that by estimating the above model we get the following predicted value equation for $t+1$:

$$\widehat{y_{t+1}} = \widehat{\beta_0} + \widehat{\beta_1} x_t \tag{2.8}$$

Hence, we can easily produce 1-period ahead forecast from this model. In order to produce forecast farther into future we would need to add more lags of the independent variable to the model. A generalized model of this kind is called *distributed lag model* and is given by:

$$y_t = \beta_0 + \sum_{s=1}^{p} \beta_s x_{t-s} + \epsilon_t \tag{2.9}$$

The number of lags to include can be determined using some kind of goodness of fit measure.

2.4.1 Dynamic Effect of X on Y

A very useful benefit of estimating a distributed lag model is that it allows us to measure how changes in x in the current period can impact the dependent variable over time. Consider a simple distributed lag model with two lags:

$$y_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \epsilon_t \tag{2.10}$$

In this model the lag structure implies that any change in x will persist for two periods in terms of its effect on y. In fact we now have to consider the *dynamic* effect of x on y. Formally, there are two types of effects:

1. *dynamic effect* of x on y given by:

$$\frac{\partial y_{t+s}}{\partial x_t} \quad s = 0, 1, 2, ...$$

In our example, the sequence of dynamic effects are:

$$\frac{\partial y_t}{\partial x_t} = 0;\ \frac{\partial y_{t+1}}{\partial x_t} = \beta_1;\ \frac{\partial y_{t+2}}{\partial x_t} = \beta_2;\ \frac{\partial y_{t+s}}{\partial x_t} = 0\ \forall\ s > 2 \tag{2.11}$$

2. *long run effect* of x on y given by:

$$\sum_{s=0}^{p} \frac{\partial y_{t+s}}{\partial x_t} \tag{2.12}$$

In our example, the long run effect is:

$$\beta_1 + \beta_2$$

2.4.2 Model Selection Criterion

Most often we compare models that have different number of independent variables. For example, in our application, in order to select the number of lags for output and capital stock, we will essentially compare models with different number of independent variables. In such cases we must account for the trade-off between goodness of fit and degrees of freedom. Increasing the number of independent variables will:

1. lower the MSE and hence leads to better fit.
2. lowers the degrees of freedom

Two commonly used measures based on MSE incorporate this trade-off:

1. Akaike Information Criterion (AIC):

$$AIC = MSE \times e^{\frac{2k}{T}}$$

where k is the number of estimated parameters, T is the sample size. Then, K/T is the number of parameters estimated per observation and $e^{\frac{2k}{T}}$ is the *penalty factor* imposed on adding more variables to the model. As we increase k, this penalty factor will increase exponentially for a given value of T.

2. Bayesian Information Criterion (BIC):

$$BIC = MSE \times T^{\frac{k}{T}}$$

Lower values of either AIC or BIC indicates greater accuracy. So we select a model with lower value of either of these two criteria. Note that the penalty imposed by BIC is harsher and hence it will typically select a more parsimonious model (Figure 2.1).

2.5 Application: A Model of Investment Expenditure

2.5.1 A Multiple Regression Model of Invesment Expenditure

Suppose have annual data on private investment, private sector output, and capital stock. Our model specification is given by:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \epsilon_t \tag{2.13}$$

We can estimate the above model using OLS and then conduct scenario-based forecasting. For ease of interpretation, we will convert all variables in natural logarithms.

Table 2.1 below presents the estimated coefficients of our regression model. Higher output and capital stock leads to greater investment expenditure.

Next, we compute forecast of investment expenditure under three different scenarios:

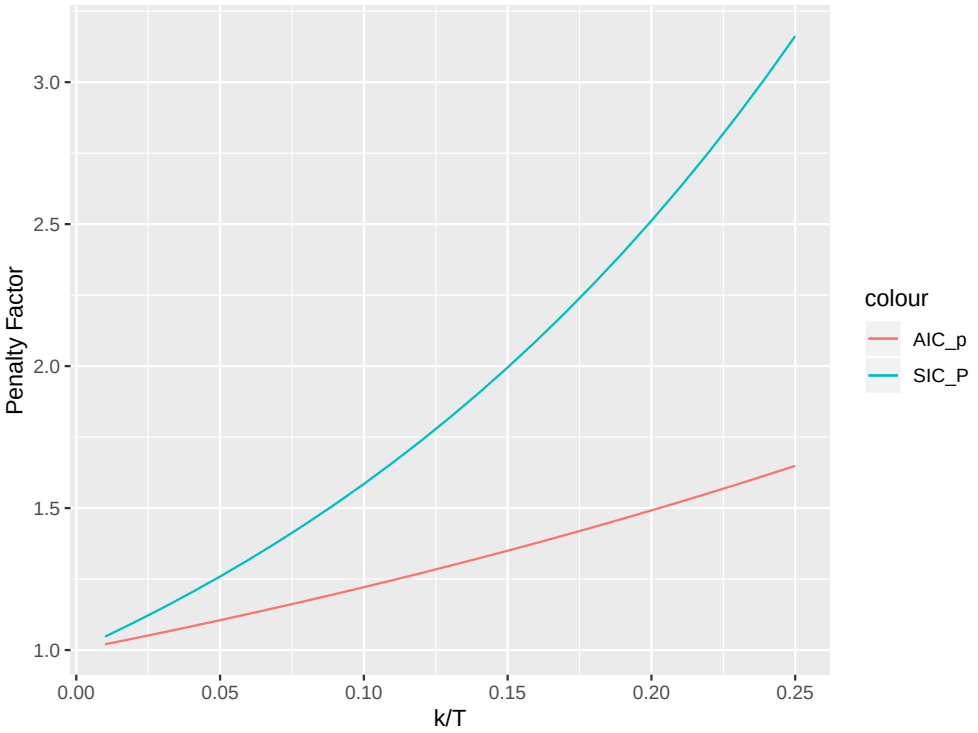


Figure 2.1: Penalty Factor of AIC and BIC

Table 2.1: A Multiple Regression Model of Investment Expenditure

	Estimated Coefficients	Std. Error	t-ratio	p-value
(Intercept)	-4.8421855	0.9623332	-5.031714	0.0000044
x1	0.9987751	0.2418282	4.130102	0.0001104
x2	0.4204833	0.3643054	1.154205	0.2528456

1. For next 3 years, both output and capital stock remain at the average of last 3 years.
2. For next 3 years, both output and capital stock remain at 1% above the average of last 3 years.
3. For next 3 years, both output and capital stock remain at 1% below the average of last 3 years.

Figure **??** below present our investment expenditure outlook under these 3 scenarios.

```
## Error in forecast(fit, newdata = newdata, level = 95): unused arguments (newdata = newdata, level = 9
## Error in forecast(fit, newdata = newdata, level = 95): unused arguments (newdata = newdata, level = 9
## Error in forecast(fit, newdata = newdata, level = 95): unused arguments (newdata = newdata, level = 9
## Error in plot(s1, include = 25, main = "Scenario 1"): object 's1' not found
## Error in plot(s2, include = 25, main = "Scenario 2"): object 's2' not found
## Error in plot(s3, include = 25, main = "Scenario 3"): object 's3' not found
```

2.5.2 A Distributed Lag Model of Investment Expenditure

In this application we will estimate a distributed lag model for investment expenditure. The idea here is that it takes time for investment to respond to output and capital stock changes. The model specification

Table 2.2: Optimal Order of the lags

Lag	AIC	BIC
1	0	0
2	0	0
3	0	0
4	0	0

we want to estimate is:

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i x_{1t-i} + \sum_{i=1}^{p} \alpha_i x_{2t-i} + \epsilon_t \tag{2.14}$$

where y denotes real investment expenditure of the private sector, x_1 denotes output of the private sector, and x_2 denotes capital stock of the private sector.

We estimate our model by first selecting the optimal lag order for each independent variable, and selecting the one with lowest value for AIC/BIC. From @??tab:ch2-table2) we find that the lowest BIC occurs at lag=2. Hence, we estimate a model with two lags for each independent variable in our model.

```
## Error in library(dynlm): there is no package called 'dynlm'
```

```
## Error in dynlm(formula = y ~ L(x1, 1:k) + L(x2, 1:k)): could not find function "dynlm"
```

Hence, our final model is given by:

$$y_t = \beta_0 + \sum_{i=1}^{2} \beta_i x_{1t-i} + \sum_{i=1}^{2} \alpha_i x_{2t-i} \tag{2.15}$$

The results of our estimation are presented below in Table **??**

```
## Error in dynlm(formula = y ~ L(x1, 1:2) + L(x2, 1:2)): could not find function "dynlm"
```

```
## Error in summary(final): object 'final' not found
```

1. Using our estimated model we can easily compute the dynamic effect as well as the long run effect of each independent variable on the dependent variable.
2. Given the lag structure of our estimated model, we can also produce forecasts for y_{t+1} by computing the following equation:

$$f_{t,1} = \widehat{y_{t+1}} = \hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{1t-1} + \hat{\alpha_1} x_{2t} + \hat{\alpha_2} x_{2t-1} \tag{2.16}$$

Chapter 3

Components of a Time Series

A given time series can have four possible components:

1. Trend: denoted by B_t captures the long run behavior of the time series of interest.
2. Season: denoted by S_t are *periodic* fluctuations over *seasons*. The period of the season is fixed and known. For example, rise in non-durable sales during Christmas.
3. Cycle: denoted by C_t are *non-periodic* fluctuations in that they occur regularly but over periods that are not fixed in duration.
4. Irregular: denoted by ϵ_t are random fluctuations, typically modeled as a white noise process.

3.1 Decomposing a time series

We can decompose any given time series into its components. There are two ways to accomplish this:

1. Additive Decomposition: Here it is assumed that all four components are added to obtain the underlying time series:
$$y_t = B_t + S_t + C_t + \epsilon_t \tag{3.1}$$
2. Multiplicative Decomposition: Here it is assumed that all four components are multiplied to obtain the underlying time series:
$$y_t = B_t \times S_t \times C_t \times \epsilon_t \tag{3.2}$$

Note that using properties of logarithms, multiplicative decomposition is the same as additive decomposition in log terms:
$$log(y_t) = log(B_t) + log(S_t) + log(C_t) + log(\epsilon_t) \tag{3.3}$$

Most statistical software can implement these decomposition using data on a time series variable as input. Typically they combine cyclical component with irregular component and provide a three-way decomposition. In Figure 3.1 I use R to decompose real GDP for the US into its components.

3.2 Uses of Decomposition of a time series

The usefulness of decomposing a time series depends on our objective.

1. It may be of interest to study each component separately or to simply improve our understanding of the temporal dynamics of a time series of interest. Decomposing it into different components is the first step towards achieving that goal.

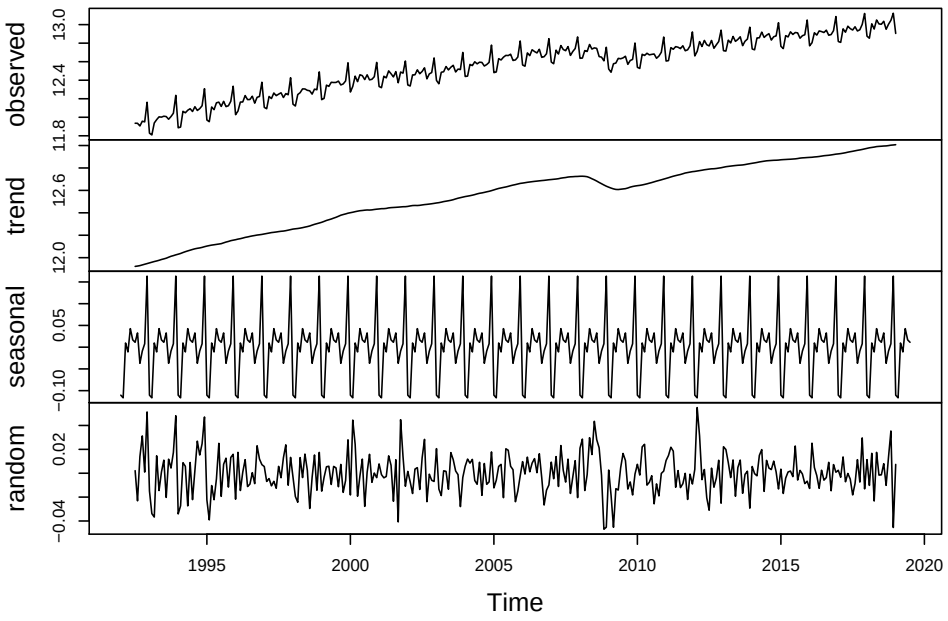


Figure 3.1: Additive Decomposition of Retail Sales

2. We can also use the decomposition to filter out components that we are not interested in studying. If for example we are only interested in modeling the cyclical component of the time series, then we can assume some kind decomposition, additive or multiplicative, and filter out the trend and seasonal component. For example, assuming additive decomposition, the filtered time series is given by:

$$Filtered\ y_t = y_t - B_t - S_t \tag{3.4}$$

We can then proceed to model the cyclical component using the filtered data.

Chapter 4

Smoothing Methods

One way to approach forecasting is to *average* out the fluctuations in the underlying time series to produce a *smoothed* data which can be extrapolated to produce forecasts. These smoothing methods are essentially *model-free* and may not even produce *optimal forecasts*. Depending on the method used one can accommodate seasonal as well as trend components of the underlying time series.

4.1 Moving Average Method

We compute an average of most recent data values for the time series and use it as a forecast for the next period.

An important parameter is the *window* over which we take the average. Let us denote this window by m, then:

$$y^s_{t+1} = \frac{\sum_{i=t-m+1}^{t} y_i}{m} \tag{4.1}$$

A larger value of m produces greater smoothing and most software have a default value of this parameter which can be changed if needed.

4.2 Simple Exponential Smoothing

In the moving average method, all observations received same weight. However, it is reasonable to argue that more recent observations may have a greater influence than those in the remote past. In this method, the weight attached to past observations exponentially decay over time. Here is the algorithm for computing the smoothed data and its forecast:

1. Initialize at t=1:
$$y^s_1 = y_1$$

2. Update:
$$y^s_t = \alpha y_t + (1-\alpha) y^s_{t-1} \quad for\ t = 2, 3, ...T$$

3: h-period ahead forecast:
$$f_{T,h} = y^s_T$$

Here the h-period ahead forecast is:

Exercise: Can you show that y^s_t is a is the weighted moving average of all past observations? Use backward substitution method.

Here $\alpha \in (0, 1)$ is the smoothing parameter, with smaller value indicating greater smoothing.

4.3 Holt-Winters Smoothing

We add trend component to the simple exponential smoothing. In step 2 the equation we use to update the smoothed data is given by:

$$\begin{aligned} y_t^s &= \alpha y_t + (1-\alpha)(y_{t-1}^s + B_{t-1}) \\ B_t &= \beta(y_t^s - y_{t-1}^s) + (1-\beta)B_{t-1} \end{aligned} \tag{4.2}$$

We now have an additional parameter β that is the trend parameter. Here the h-period ahead forecast is:

$$f_{T,h} = y_T^s + h \times B_T \tag{4.3}$$

4.4 Holt-Winters Smoothing with Seasonality

We now add seasonal component along with trend. Assuming multiplicative seasonality with period n:

$$y_t^s = \alpha \frac{y_t}{S_{t-n}} + (1-\alpha)(y_{t-1}^s + B_{t-1}) \tag{4.4}$$

$$B_t = \beta(y_t^s - y_{t-1}^s) + (1-\beta)B_{t-1} \tag{4.5}$$

$$S_t = \gamma \frac{y_t}{y_t^s} + (1-\gamma)S_{t-n} \tag{4.6}$$

The h-period ahead forecast is given by:

$$f_{T,h} = (y_T^s + h \times B_T) \times S_{T+h-n} \tag{4.7}$$

4.5 Application

We use R to implement a 12-period ahead forecast for new housing starts for the U.S. The data is at monthly frequency from Jan 1959 through March 2019. The resulting forecasts are plotted in Figure **??**.

```
## Error in forecast(s_exp1, h = 12): unused argument (h = 12)
## Error in forecast(s_exp2, h = 12): unused argument (h = 12)
## Error in forecast(s_exp3, h = 12): unused argument (h = 12)
## Error in plot(f_exp1, include = 24, main = "Simple Exponential Smoothing"): object 'f_exp1' not found
## Error in plot(f_exp2, include = 24, main = "Holt-Winters with Trend"): object 'f_exp2' not found
## Error in plot(f_exp3, include = 24, main = "Hold-Winters with Trend and Season"): object 'f_exp3' not
```

Chapter 5

Modeling Trend and Seasonal Components

5.1 Trend Estimation

An important component of a time series is *trend* that captures the long run evolution of the variable of interest. There are two types of trends:

1. Deterministic Trend: the underlying trend component is a *known* function of time with *unknown* parameters.
2. Stochastic Trend: the trend component is random.

In this note we will focus on estimating and forecasting deterministic trend models. We will come back to stochastic trend later when we talk about stationarity property of a time series.

5.1.1 Parametrizing a deterministic trend

Whether or not there is deterministic trend in the data can be typically gleaned by simply plotting the time series over time. For example, Figure @ref(fig: ch5-figure1) below plots real GDP for the US at quarterly frequency. We can observe a positive time trend with real GDP increasing with time. In this section we will learn to *fit* a function that captures this relationship accurately.

Note: The variable time is denoted by t and it is artificially created to take value of 1 for the first period, 2 for the second period and so on.

There are two commonly used functional forms for capturing a deterministic trend:

1. Polynomial Trend: We fit a polynomial of appropriate order to capture the time trend. For example, A. Linear trend:

$$y_t = \beta_0 + \beta_1 t + \epsilon_t \tag{5.1}$$

B. Quadratic trend:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t \tag{5.2}$$

In general, we can fit a polynomial of order q:

$$y_t = \beta_0 + \sum_{i=1}^{q} \beta_i t^i + \epsilon_t \tag{5.3}$$

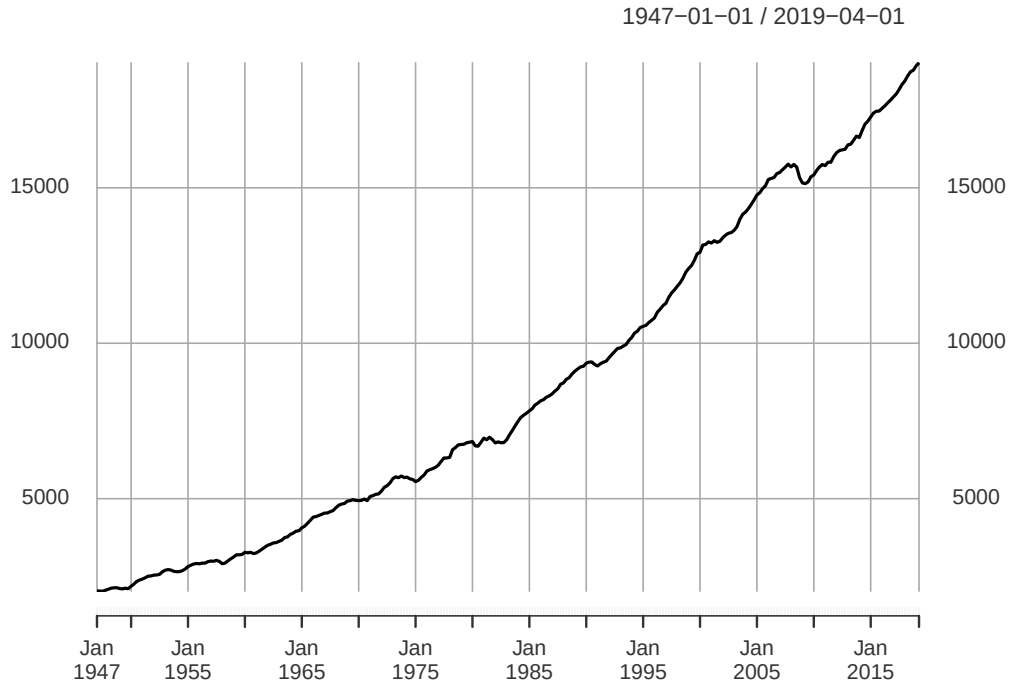


Figure 5.1: Real GDP (2012 Chained Billions of Dollars)

We can estimate this model using the OLS. One of the key component here is to determine the *right* order of the polynomial. We can begin with a large enough number for q and then select the appropriate order using AIC or BIC criterion.

2. Exponential or log-linear trend: In some cases we may want to use an exponential trend or equivalently a log-linear trend.

$$y_t = e^{(\beta_0+\beta_1 t+\epsilon_t)} \tag{5.4}$$

$$equivalently \tag{5.5}$$

$$log(y_t) = \beta_0 + \beta_1 t + \epsilon_t \tag{5.6}$$

Again we can estimate the above model using OLS.

5.1.2 Uses of the Deterministic Trend Model

Once we have finalized our deterministic trend model i.e., either a polynomial of a specific order or log-liner trend, we can use the estimated model for the following two purposes:

1. Detrending our data: Suppose we would like to eliminate trend from our data. The residual from our final trend model is the *detrended* time series.

2. Forecasting: We can also forecast our time series based on the estimated trend. For example, suppose our final model is a quadratic trend. The predicted value is given by:

$$\widehat{y_t} = \widehat{\beta_0} + \widehat{\beta_1} t + \widehat{\beta_2} t^2 \tag{5.7}$$

Then, the 1-period ahead forecast for y_{t+1} can be obtained by solving:

$$\widehat{y_{t+1}} = \widehat{\beta_0} + \widehat{\beta_1}(t+1) + \widehat{\beta_2}(t+1)^2 \tag{5.8}$$

Table 5.1: Optimal Order of the Polynomial

order	AIC	BIC
1	4813.417	4824.426
2	4199.200	4213.879
3	4169.934	4188.283
4	4089.930	4111.949

Table 5.2: Regression Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1769.417	82.765	21.379	0
trend	38.885	3.927	9.903	0
I(trend^2)	-0.262	0.055	-4.779	0
I(trend^3)	0.002	0.000	8.758	0
I(trend^4)	0.000	0.000	-9.651	0

5.1.3 Application: Estimating a polynomial trend for U.S. Real GDP

We will now fit a polynomial trend to the US real GDP data that was presented in Figure **??**. We first estimate polynomials of different orders and select the optimal order determined by the lowest possible AIC/BIC. Table 5.1. shows these statistics for up to 4th order polynomial. We find that the lowest value occur at $q = 4$.

Hence, our final trend model is:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \epsilon_t \tag{5.9}$$

The estimated trend model is presented in Table 5.2.

Using the estimated model, we can compute the detrended data as the residual and also forecast y_t. Figure 5.2 below plots the detrended real GDP obtained as a residual from our trend model.

Figure **??** shows the forecast of real GDP for next 8 quarters along with the 95% confidence bands.

```
## Error in forecast(fit, h = 8): unused argument (h = 8)
```

```
## Error in plot(fcast, include = 24, main = ""): object 'fcast' not found
```

5.2 Seasonal Model

We now focus on the *seasonal* component of a time series, i.e., that is periodic fluctuations that repeat themselves every season. For example, increase in ice cream sales during summer season. Just like trend component, such seasonal pattern could be *deterministic* or *stochastic*. In this chapter we will focus on estimating deterministic seasonal component.

In Figure 5.3 we plot housing starts in the U.S. The data is at monthly frequency and we can see a clear seasonal pattern. Housing starts seem to increase in spring and summer months. This is followed by a decline in fall and winter months.

One option to deal with seasonality is to obtain seasonally adjusted data (or deseasonalized data) from the source itself. Alternatively, we can use decomposition method and appropriately filter out the seasonal component. However, if our objective is to explicitly model the seasonal component of a time series then we must work with non-seasonally adjusted data.

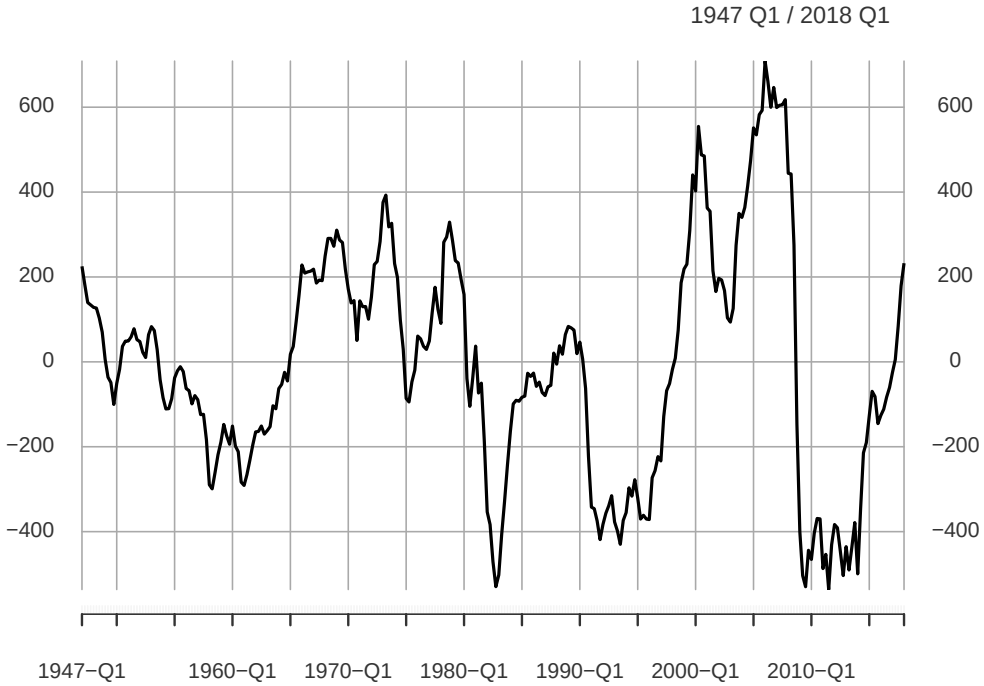


Figure 5.2: Detrended Real GDP

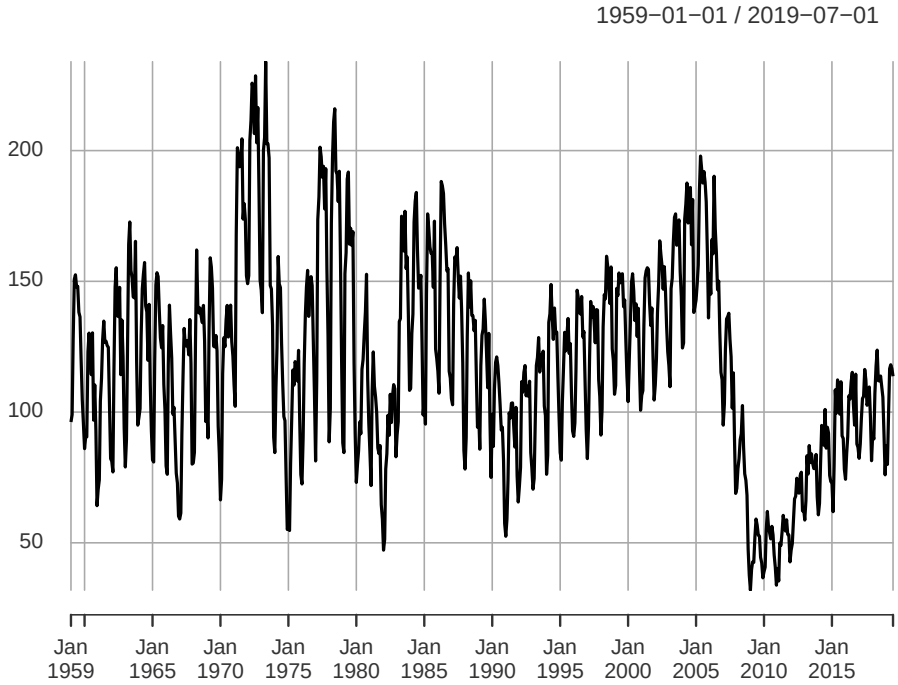


Figure 5.3: Housing Starts in U.S.

Table 5.3: Regression Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	86.925	4.320	20.119	0.000
season2	2.938	6.110	0.481	0.631
season3	32.016	6.110	5.240	0.000
season4	48.190	6.110	7.887	0.000
season5	53.364	6.110	8.734	0.000
season6	52.320	6.110	8.563	0.000
season7	46.630	6.110	7.632	0.000
season8	44.849	6.135	7.310	0.000
season9	38.210	6.135	6.228	0.000
season10	42.242	6.135	6.885	0.000
season11	21.104	6.135	3.440	0.001
season12	4.797	6.135	0.782	0.435

5.2.1 Regression Model with Seasonal Dummy Variables

One way to account for seasonal patterns in data is to add dummy variables for season. To avoid perfect multicollinearity, is there are s seasons, we can include $s-1$ dummy variables. For example, for quarterly data, $s = 4$ and hence we need $s - 1 = 3$ dummy variables in our regression model. Formally, for quarterly data, the seasonal regression model is given by:

$$y_t = \beta_0 + \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \epsilon_t \tag{5.10}$$

In the above regression model, D_1, D_2, and D_3 are dummy variables that capture first three quarters of the year. For example, $D_1 = 1$ for the first quarter and $D_1 = 0$ otherwise. Similarly, $D_2 = 1$ for the second quarter and $D_2 = 0$ otherwise. In this example, we use the fourth quarter as the *base group*.

The above model can be estimated using OLS. Again, we can use the residual from our estimated model as a measure of *deseasonlized* data. We can also forecast the dependent variable based on the seasonal component only.

5.2.2 Application: Seasonal Model of Housing Starts

We now estimate a seasonal regression model for the housing starts data presented in Figure 5.3. The data is at monthly frequency which implies we can have 12 possible seasons and hence would need 11 dummy variables in our regression model. Formally, we use January as the base group and include dummy variables for the last 11 months of the year:

$$y_t = \beta_0 + \sum_{i=2}^{12} \beta_i D_{it} + \epsilon_t \tag{5.11}$$

Table 5.3 presents the estimation results for this exercise. In Figure **??** we plot the forecast of housing starts for next 12 months using our estimated model, along with 95% confidence bands.

```
## Error in forecast(fit, h = 12): unused argument (h = 12)
```

```
## Error in plot(fcast, include = 24, main = ""): object 'fcast' not found
```

Chapter 6

Modeling Cycle

In this chapter we will focus on the cyclical component of a time series and hence focus on data that either has no trend and seasonal components, or data that is filtered to eliminate any trend and seasonality. One of the most commonly used method to model cyclicality is the *Autogressive Moving Average (ARMA)*. This model has two distinct components:

1. *Autoregressive (AR) component*: the current period value of a time series variable depends on its past (lagged) observations. We use p to denote the **order** of the AR component and is the number of lags of a variable that directly affect the current period value. For example, a firm's production in the current period maybe impacted by past levels of production. If last year's production exceeded demand, the stock of unsold goods may be used to meet this period demand first, hence lowering the current period production.

2. *Moving average (MA) component*: the current period value of a time series variable depends on current period **shock** as well as past shocks to this variable. We use q to denote the **order** of the MA component and is the number of past period shocks that affect the current period value of the variable of interest. For example, if the Federal Reserve Bank raises the interest in 2016, the effects of that policy shock may impact investment and consumption spending in 2017.

Before we consider these time series model in details it is useful to discuss certain properties of time series that allow us a better understanding of these models.

6.1 Stationarity and Autocorrelation

6.1.1 Covariance Stationary Time Series

Definition 6.1 (Covariance Stationary Time Series)**.**

A time series $\{y_t\}$ is said to be a *covariance stationary process* if:

1. $E(y_t) = \mu_y \quad \forall \quad t$
2. $Var(y_t) = \sigma_y^2 \quad \forall \quad t$
3. $Cov(y_t, y_{t-s}) = \gamma(s) \quad \forall \quad s \neq t$

One way to think about stationarity is *mean-reversion*, i.e, the tendency of a time series to return to its *long-run* unconditional mean following a shock (or a series of shock). Figure @**??**fig:ch6-figure1) below shows this property graphically.

In practice however, you will not be able to visualize a mean-reverting stationary process this clearly. For example, in Figure 6.2 we plot real GDP growth for the U.S. which is a stationary process with a mean of

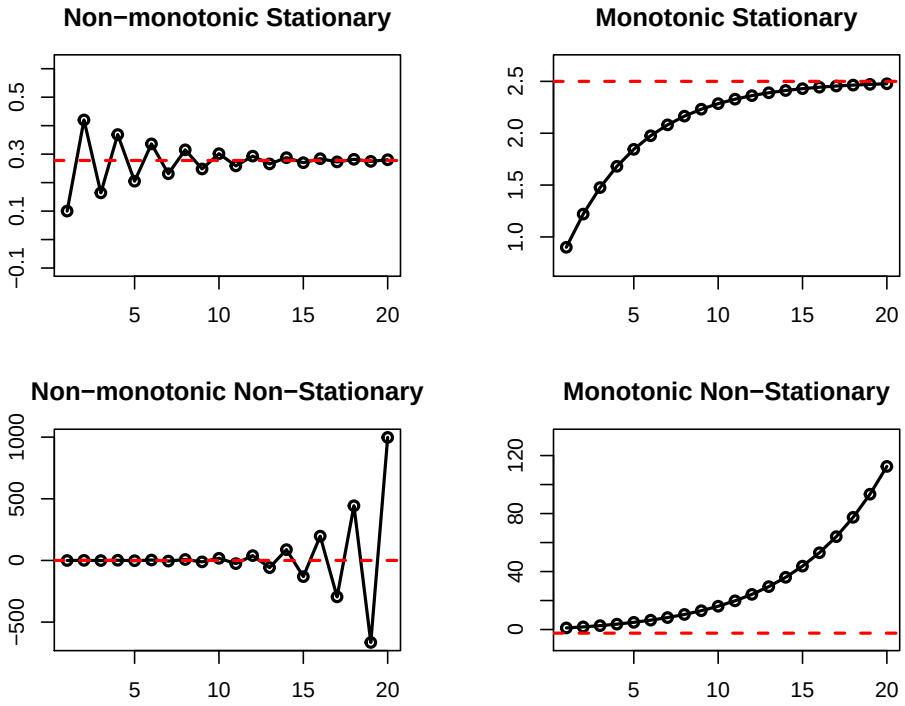


Figure 6.1: Reversion to mean

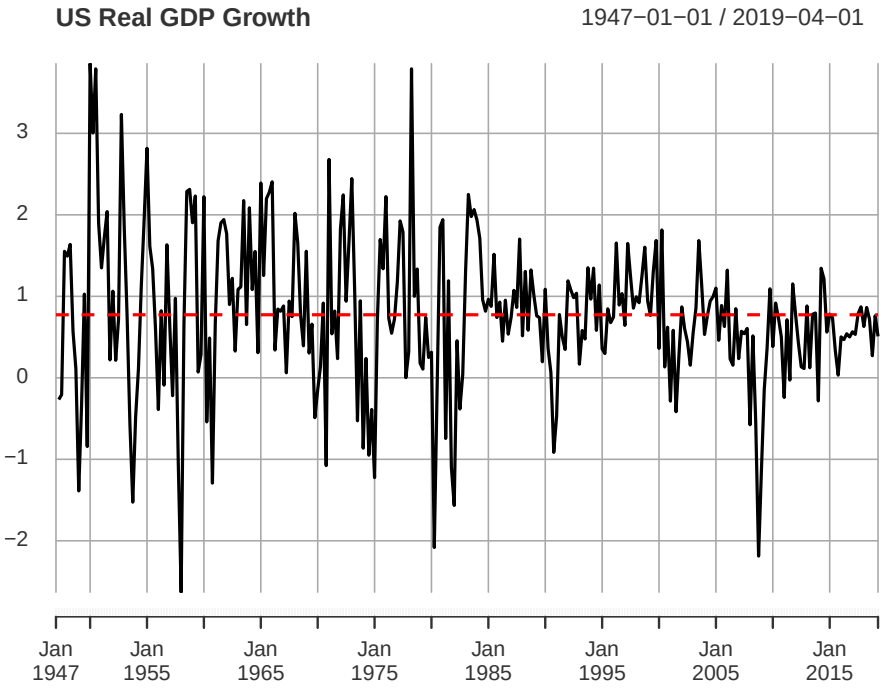


Figure 6.2: Reversion to mean in practice

0.7%. In this chapter we will only consider stationary time series data. Later on we will learn how to work with non-stationary data.

6.1.2 Correlation vs Autocorrelation

In statistics, correlation is a measure of relationship between two variables. In the time series setting, we can think of the current period value and the past period value of a variable as two **separate** variables, and compute correlation between them. Such a correlation, between current and lagged observation of a time series is called **serial correlation** or **autocorrelation**. In general, for a time series, $\{y_t\}$, the autocorrelation is given by:

$$Cor(y_t, y_{t-s}) = \frac{Cov(y_t, y_{t-s})}{\sqrt{\sigma^2_{y_t} \times \sigma^2_{y_{t-s}}}} \tag{6.1}$$

where $Cov(y_t, y_{t-s}) = E(y_t - \mu_{y_t})(y_{t-s} - \mu_{y_{t-s}})$ and $\sigma^2_{y_t} = E(y_t - \mu_{y_t})^2$

For a stationary time series, using the three conditions the **Autocorrelation Function (ACF)** denoted by $\rho(s)$ is given by:

$$ACF(s) \text{ or } \rho(s) = \frac{\gamma(s)}{\gamma(0)} \tag{6.2}$$

Non-zero values of the ACF indicates presences of serial correlation in the data. Figure 6.3 shows the ACF for a stationary time series with positive serial correlation. If your data is stationary then the ACF should eventually converge to 0. For a non-stationary data, the ACF function will not decay over time.

6.1.3 Partial Autocorrelation

Definition 6.2 (Partial Auto Correlation Function (PACF))**.**

The ACF captures the relationship between the current period value of a time series and all of its past observations. It includes both direct as well as indirect effects of the past observations on the current period value. Often times it is of interest to measure the direct relationship between the current and past observations, **partialing** out all indirect effects. The *partial autocorrelation function (PACF)* for a stationary time series y_t at lag s is the direct correlation between y_t and y_{t-s}, after filtering out the linear influence of $y_{t-1}, \ldots, y_{t-s-1}$ on y_t. Figure 6.4 below shows the PACF for a stationary time series where only one lag directly affects the time series in the current period.

6.1.4 Lag operator

A **lag operator** denoted by L allows us to write ARMA models in a more concise way. Applying lag operator once moves the time index by one period; applying it twice moves the time index back by two period; applying it s times moves the index back by s periods.

$$\begin{aligned} Ly_t &= y_{t-1} \\ L^2 y_t &= y_{t-2} \\ L^3 y_t &= y_{t-3} \\ &\vdots \\ L^s y_t &= y_{t-s} \end{aligned}$$

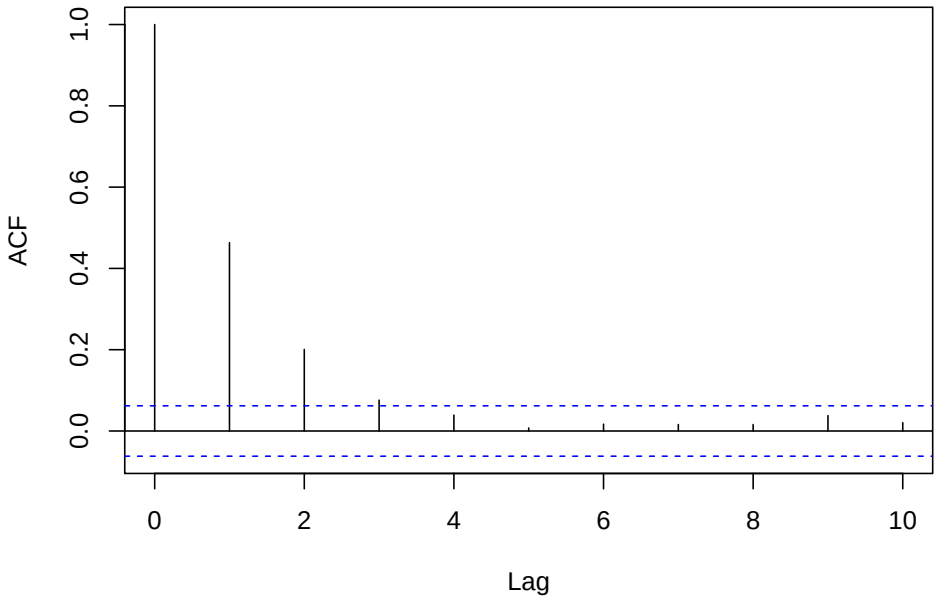


Figure 6.3: ACF for a Stationary Time Series

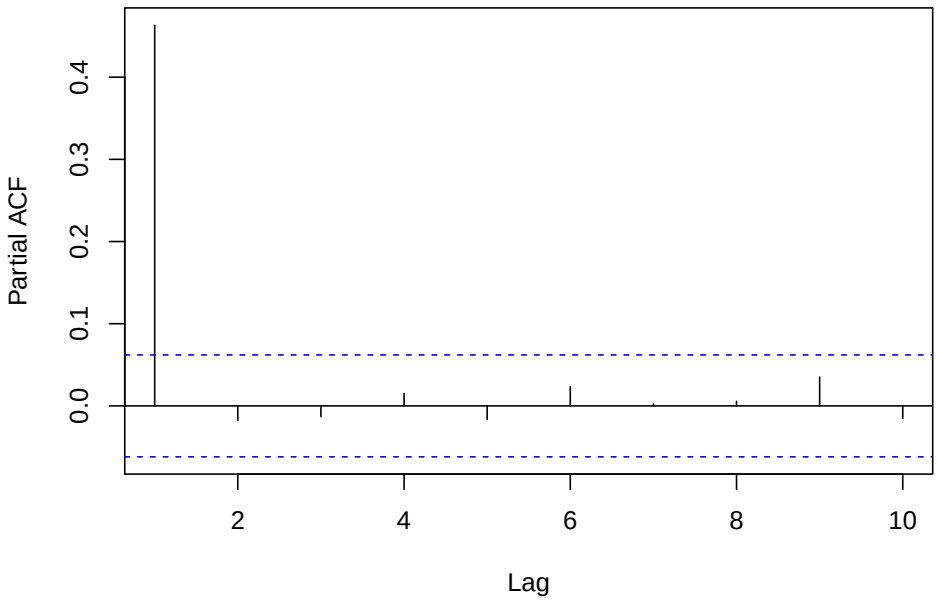


Figure 6.4: PACF for a Stationary Time Series

6.2 Autoregressive (AR) Model

A *stationary*time series $\{x_t\}$ can be modeled as an AR process. In general, an AR(p) model is given by:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t \tag{6.3}$$

Here ϕ_i captures the effect of y_{t-i} on y_t. The order of the AR process is not known apriori. It is common to use either AIC or BIC to determine the optimal lag length for an AR process.

Using the Lag operator, we can rewrite the above AR(p) model as follows:

$$\Phi(L)y_t = \phi_0 + \epsilon_t$$

where $\Phi(L)$ is a polynomial of degree p in L:

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - ..... - \phi_p L^p$$

For example, an AR(1) model can be written as:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t \Rightarrow \Phi(L)y_t = \phi_0 + \epsilon_t$$

where,

$$\Phi(L) = 1 - \phi_1 L$$

Characteristic equation: A characteristic equation is given by:

$$\Phi(L) = 0$$

The roots of this equation play an important role in determining the dynamic behavior of a time series.

6.2.1 Unit root and Stationarity

For a time series to be stationary there should be no **unit root** in its *characteristic equation*. In other words, all roots of the characteristic equation must fall outside the unit circle. Consider the following AR(1) model:

$$\Phi(L)y_t = \phi_0 + \epsilon_t$$

The characteristic equation is given by:

$$\Phi(L) = 1 - \phi_1 L = 0$$

The root that satisfies the above equation is:

$$L^* = \frac{1}{\phi_1}$$

For no unit root to be present, $L^* > |1|$ which implies that $|\phi_1| < 1$.

Typically, for any AR process to be stationary, some restrictions will be imposed on the values of $\phi_i's$, the coefficients of the lagged variables in the model.

6.2.2 Properties of an AR(1) model

A stationary AR(1) model is given by:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t \quad ; \; \epsilon_t \sim WN(0, \sigma_\epsilon^2) \; and \; |\phi_1| < 1$$

1. ϕ_1 measures the persistence in data. A larger value indicates shocks to y_t dissipate slowly over time.
2. Stationarity of y_t implies certain restrictions on the AR(1) model.
   i. Constant long run mean: is the unconditional expectation of y_t:

$$E(y_t) = \mu_y = \frac{\phi_0}{1 - \phi_1}$$

   ii. Constant long run variance: is the unconditional variance of y_t:

$$Var(y_t) = \sigma_y^2 = \frac{\sigma_\epsilon^2}{1 - \phi_1^2}$$

   iii. ACF function:

$$\rho(s) = \phi_1^s$$

   iv. PACF function:

$$PACF(s) = \begin{cases} \phi_1 & \text{if s=1} \\ 0 & \text{if s>1} \end{cases}$$

6.3 Estimating an AR model

When estimating the AR model we have two alternatives:

1. OLS: biased (but consistent) estimates. Also, later on when we add MA components we cannot use OLS.
2. Maximum Likelihood Estimation (MLE): can be used to estimate AR as well as MA components

6.3.1 Maximum Likelihood Estimation (MLE)

- MLE approach is based on the following idea:

what set of values of our parameters maximize the likelihood of observing our data if the model we have was used to generate this data.

Likelihood function: is a function that gives us the probability of observing our data given a model with some parameters.

6.3.1.1 Likelihood vs Probability

Consider a simple example of tossing a coin. Let X denotes the random variable that is the outcome of this experiment being either heads or tails. Let θ denote the probability of heads which implies $1 - \theta$ is the probability of obtaining tails. Here, θ is our parameter of interest. Suppose we toss the coin 10 times and obtain the following data on X:

$$X = \{H, H, H, H, H, H, T, T, T, T\}$$

Then, the probability of obtaining this sequence of X is given by:

$$Prob(X|\theta) = \theta^6(1 - \theta)^4$$

This is the probability distribution function the variable X. As we change X, we get a different probability for a given value of θ.

Now let us ask a different question. Once we have observed the sequence of heads and tails, lets call it our data which is fixed. Then, what is probability of observing this data, if our probability distribution function is given by the equation above? That gives us the likelihood function:

$$L(\theta) = Prob(X|\theta) = \theta^6(1-\theta)^4$$

Note that with fixed X, as we change θ the likelihood of observing this data will change.

This is an important point that distinguishes likelihood function from the probability distribution function. Although both have the same equation, the probability function is a function of the data with the value of the parameter fixed, while the likelihood function is a function of the parameter with the data fixed.

6.3.1.2 Maximum Likelihood Estimation

Now we are in a position to formally define the likelihood function.

Definition 6.3. Let X denotes a random variable with a given probability distribution function denoted by $f(x_i|\theta)$. Let $D = \{x_1, x_2, \ldots, x_n\}$ denote a sample realization of X. Then, the likelhood function, denoted by $L(\theta)$ is given by:

$$L(\theta) = f(x_1, x_2, \ldots, x_n|\theta)$$

If we further assume that each realization of X is independent of the others, we get:

$$L(\theta) = f(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)$$

A mathematical simplification is to work with natural logs of the likelihood function, which assuming independently distributed random sample, gives us:

$$lnL(\theta) = ln(f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)) = \sum_{i=1}^{N} ln(f(x_i|\theta))$$

Definition 6.4. The maximum likelihood estimator, denoted by $\hat{\theta}_{MLE}$, maximizes the log likelihood function:

$$\hat{\theta}_{MLE} \equiv arg\max_{\theta} lnL(\theta)$$

Example 6.1. Compute maximum likelihood estimator of μ of an indpendently distributed random variable that is normally distributed with a mean of μ and a variance of 1:

$$f(y_t|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_t-\mu)^2}$$

Solution: The log likelihood function is given by:

$$lnL = -Tln2\pi - \frac{1}{2}\sum_{t=1}^{T}(y_t - \mu)^2$$

From the first order condition, we get

$$\frac{\partial LnL}{\partial \mu} = \sum_{t=1}^{T}(y_t - \mu) = 0 \Rightarrow \hat{\mu}_{MLE} = \frac{\sum_{t=1}^{T} y_t}{T}$$

6.3.2 MLE of an AR(p) model

One complication we face in estimating an AR(p) model is that by definition the realizations of the variable are not independent of each other. As a result we cannot simplify the likelihood function by multiplying individual probability density functions to obtain the joint probability density function, i.e.,

$$f(y_1, y_2, \ldots, y_T|\theta) \neq f(y_1|\theta) \times f(y_2|\theta) \times \cdots \times f(y_T|\theta)$$

Furthermore, as the order of AR increases, the joint density function we need to estimate becomes even more complicated. In this class we will focus on the method that divides the joint density into the product of conditional densities and density of a set of initial values. The idea comes from the conditional probability formula for two related events A and B:

$$P(A|B) = \frac{P(\text{A and B})}{P(B)} \Rightarrow P(\text{A and B}) = P(A|B) \times P(B)$$

In the time series context, I will explain this for a stationary AR(1) model. We know that in this model only last period observation directly affects the current period value. Hence, consider the first two observations of a stationary time series: y_1 and y_2. Then the joint density of these adjacent observations is given by,

$$f(y_1, y_2; \theta) = f(y_2|y_1; \theta) \times f(y_1; \theta)$$

Similarly, for the first three observations we get:

$$f(y_1, y_2, y_3; \theta) = f(y_3|y_2; \theta) \times f(y_2|y_1; \theta) \times f(y_1; \theta)$$

Hence, for T observations we get:

$$f(y_1, y_2, y_3, ..., y_T; \theta) = f(y_T|y_{T-1}; \theta) \times f(y_{T-1}|y_{T-2}; \theta) \times .... \times f(y_1; \theta)$$

The log-likelihood function is given by:

$$ln\ L(\theta) = ln\ f(y_1; \theta) + \sum_{t=2}^{T} ln\ f(y_t|y_{t-1}; \theta)$$

We can then maximize the above likelihood function to obtain an MLE estimator for the AR(1) model.

6.3.3 Selection of optimal order of the AR model

Note that apriori we do not know the order of the AR model for any given time series. We can determine the optimal lag order by using either AIC or BIC. The process is as follows:

1. Set $p = p_{max}$ where p_{max} is an integer. A rule of thumb is to set

$$p_{max} = integer \left[12 \times \left(\frac{T}{100} \right)^{0.25} \right]$$

2. Estimate all AR models from $p = 1$ to $p = p_{max}$.
3. Select the final model as the one with lowest AIC or lowest BIC.

6.3.4 Forecasting using AR(p) model

Having estimated our AR(p) model with the optimal lag length, we can use the conditional mean to compute the forecast and conditional variance to compute the forecast errors. Consider an AR(1) model:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$$

Then, the 1-period ahead forecast is given by:

$$f_{t,1} = E(y_{t+1}|\Omega_t) = \phi_0 + \phi_1 y_t$$

Similarly, the 2-period ahead forecast is given by:

$$f_{t,2} = E(y_{t+2}|\Omega_t) = \phi_0 + \phi_1 E(y_{t+1}|\Omega_t) = \phi_0 + \phi_1 f_{t,1}$$

In general, we can get the following recursive forecast equation for h-period's ahead:

$$f_{t,h} = \phi_0 + \phi_1 f_{t,h-1}$$

Correspondingly, the h-period ahead forecast error is given by:

$$e_{t,h} = y_{t+h} - f_{t,h} = \epsilon_{t+h} + \phi_1 e_{t,h-1}$$

Theorem 6.1. *The h-period ahead forecast converges to the unconditional mean of y_t, i.e.,*

$$\lim_{h\to\infty} f_{t,h} = \mu_y = \frac{\phi_0}{1-\phi_1}$$

Theorem 6.2. *The variance of the h-period ahead forecast error converges to the unconditional variance of y_t, i.e.,*

$$\lim_{h\to\infty} Var(e_{t,h}) = \sigma_y^2 = \frac{\sigma_\epsilon^2}{1-\phi_1^2}$$

6.4 Moving Average (MA) Model

Another commonly used method for capturing the cyclical component of the time series is the **moving average (MA)** model where the current value of a time series linearly depends on current and past shocks. Formally, a *stationary* time series $\{y_t\}$ can be modeled as an MA(q) process:

$$y_t = \theta_0 + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q} \tag{6.4}$$

Using lag operator, we can write this in more compact form as:

$$y_t = \theta_0 + \Theta(L)\epsilon_t$$

where $\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q$ is lag polynomial of order q.

Note that because each one of the current and past shocks are white noise processes, an MA(q) model is always stationary.

6.4.1 Invertibility of an MA process

Consider the following MA(1) process with$\theta_0 = 0$ for simplicity:

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1}$$

Using the lag operator we can rewrite this equation as follows:

$$y_t = (1 + \theta_1 L)\epsilon_t \Rightarrow y_t(1 + \theta_1 L)^{-1} = \epsilon_t$$

Note that if $|\theta_1| < 1$, then we can use the Taylor series expansion centered at 0 and get:

$$(1 + \theta_1 L)^{-1} = 1 - \theta_1 L + (\theta_1 L)^2 - (\theta_1 L)^3 + (\theta_1 L)^4 - .....$$

Hence, an MA(1) can be rewritten as follows:

$$y_t(1 - \theta_1 L + (\theta_1 L)^2 - (\theta_1 L)^3 + (\theta_1 L)^4 - .....) = \epsilon_t$$

$$\Rightarrow y_t - \theta_1 y_{t-1} + \theta_1^2 y_{t-2} - \theta_1^3 y_{t-3}.... = \epsilon_t$$

Rearranging terms, we get the $AR(\infty)$ representation for an invertible MA(1) model:

$$y_t = -\sum_{i=1}^{\infty} (-\theta_1)^i \; y_{t-i} + \epsilon_t$$

Definition 6.5. An MA process is invertible if it can be represented as a stationary $AR(\infty)$.

6.4.2 Properties of an invertible MA(1)

An invertible MA(1) model is given by:

$$y_t = \theta_0 + \epsilon_t + \theta_1 \epsilon_{t-1} \quad ; \; \epsilon_t \sim WN(0, \sigma_\epsilon^2) \; and \; |\theta_1| < 1$$

1. Constant unconditional mean of y_t:
$$E(y_t) = \mu_y = \theta_0$$

2. Constant unconditional variance of y_t:
$$Var(y_t) = \sigma_y^2 = \sigma_\epsilon^2(1 + \theta_1^2)$$

3. ACF function:
$$ACF(s) = \begin{cases} \frac{\theta_1}{1+\theta_1^2} & \text{if s=1} \\ 0 & \text{if s>1} \end{cases}$$

4. PACF function: using the invertibility it is evident that PACF of an MA(1) decays with s.

6.4.3 Forecast based on MA(q)

Like before, the h-period ahead forecast is the conditional expected value of the time series. Consider an MA(1) model:

$$y_t = \theta_0 + \epsilon_t + \theta_1 \epsilon_{t-1}$$

Then, the 1-period ahead forecast is given by:

$$f_{t,1} = E(y_{t+1}|\Omega_t) = \theta_0 + \theta_1 \epsilon_{t-1}$$

The h-period ahead forecast for $h > 1$ is given by:

$$f_{t,h} = E(y_{t+h}|\Omega_t) = \theta_0$$

In general, for an MA(q) model, the forecast for $h > q$ is the long run mean θ_0. This is why we say that an MA(q) process has a memory of q periods.

6.5 ARMA(p, q)

An ARMA model simply combines both AR and MA components to model the dynamics of a time series. Formally,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} \tag{6.5}$$

Note that:

1. Estimation is done by maximum likelihood method.
2. Optimal order for AR and MA components is selected using AIC and/or BIC.
3. The forecast of y_t from an ARMA(p,q) model will be dominated by the AR component for $h > q$. To see this consider the following ARMA(1,1) model:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}$$

Then, the 1-period ahead forecast is:

$$f_{t,1} = E(y_{t+1}|\Omega_t) = \phi_0 + \phi_1 y_t + \theta_1 \epsilon_{t-1}$$

Here both MA and AR component affect the forecast. But now consider the 2-period ahead forecast:

$$f_{t,2} = E(y_{t+2}|\Omega_t) = \phi_0 + \phi_1 f_{t,1}$$

Hence, no role is played by the MA component in determining the 2-period ahead forecast. For any $h > 1$ only the AR component affects the forecast from this model.

6.6 Integrated ARMA or ARIMA(p,d,q)

Thus far we have assumed that our data is stationary. However, often we may find that this assumption is not supported in practice. In such a case we need to tranform our data appropriately before estimating an ARMA model. The procedure can be summarized as follows:

1. Determine whether there is a unit root in data or not. Presence of unit root indicates non-stationarity. We will use Augmented Dickey-Fuller (ADF) test for this purpose.
2. If data is non-stationary, then we need to appropriately transform our data to make it stationary.
3. Once we have obtained a stationary transformation of our original data, we can proceed and estimate the ARMA model as before.

6.7 Trend Stationary vs Difference Stationary Time Series

There are two types of time series we often encounter in real world:

1. Trend-stationary: a time-series variable is non-stationary becuase it has a deterministic trend. Once we detrend our data then it will become stationary. In this case the appropriate transformation is to estimate a trend model and then use the residual as the detrended stationary data. For example, suppose our data has a linear trend given by:

$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

Then the OLS residual from this model, $e_t = y_t - \hat{y_t}$ is the detrended y_t which will be stationary. Hence, we will estimate an ARMA(p,q) model using this detrended variable.

2. Difference-stationary: a time-series variable is non-stationary because it contains a stochastic trend. Here, the transformation requires us to difference the orignial data until we obtain a stationary time series. Let d denote the minimum number of differences needed to obtain a stationary time series:

$$\Delta_d \; y_t = (1 - L)^d \; y_t$$

In this case, we say that y_t is intergrated of order d or more formally, y_t is an $I(d)$ process. Hence, for $d = 1$ we obtain an $I(1)$ process implying that:

$$\Delta_1 \; y_t = (1 - L)^1 \; y_t = y_t - y_{t-1} \quad \text{is stationary}$$

In otherwords, the first difference of an I(1) process is stationary. Similarly for $d = 2$, we obtain an $I(2)$ process where second difference will be stationary and so forth.

6.8 Testing for a unit root

Consider the following AR(1) model with no trend and intercept:

$$y_t = \phi_1 y_{t-1} + \epsilon_t \; quad; \epsilon_t \sim WN(0, \sigma_\epsilon^2)$$

We know that if $\phi_1 = 1$ we have a unit root in this data. Lets subtract y_{t-1} from both sides and rewrite this model as:

$$y_t - y_{t-1} = (\phi_1 - 1) y_{t-1} + \epsilon_t$$

Define $\rho = phi_1 - 1$. Then, we get:

$$\Delta y_t = \rho \; y_{t-1} + \epsilon_t$$

We can now estimate the above model and carry out the following test known as the Dickey-Fuller (DF) test:

$$H_0 : \rho = 0$$

$$H_A : \rho < 0$$

If the null hypothesis is not rejected, then we do not have sample evidence against the statement that $\rho = 0 \Rightarrow \phi_1 = 1$. Hence, we conclude that there is no evidence against the statement that there is unit root in the data. In contrast, if we reject the null hypothesis, then we can conclude that there is no unit root and hence the data is stationary.

The t- statistic is for the above test is denoted by τ_1 and is given by:

$$\tau_1 = \frac{\hat{\rho}}{se(\hat{\rho})}$$

Under the null hypothesis this test statistic follows the DF-distribution and the critical values are provided in most statistical softwares. Given that this is a left-tail test, the decision rule is that if the test statistic is less than the critical value then we reject the null hypothesis.

There are two issues we face when implementing this test in practice:

1. First, the above procedure assumes that there is no intercept and trend in the data. In real world, we cannot make that assumption and must extend the test procedure to accomodate a non-zero intercept and trend. Hence, we have the following two additional versions of the DF test:

   i. Constant and no trend model: Here our AR(1) model is

   $$\Delta y_t = \phi_0 + \rho \; y_{t-1} + \epsilon_t$$

   Now we can do two possible tests. The first test is that of the unit root:

   $$H_0 : \rho = 0$$

   $$H_A : \rho < 0$$

   The t statistic for this test is denoted by τ_2 and is given by:

   $$\tau_2 = \frac{\hat{\rho}}{se(\hat{\rho})}$$

   If the test statisitic is less than the critical value, we reject the null.

   The second test we can do is:

   $$H_0 : \rho = \phi_0 = 0$$

   $$H_A : Not \; H_0$$

   The test statistic for this test is denoted by ϕ_1. If the test statistic exceeds the critical value then we reject the null.

   ii. Constant and linear trend model: Here our AR(1) model is

$$\Delta y_t = \phi_0 + \beta\ t + \rho\ y_{t-1} + \epsilon_t$$

Now we can do three possible tests. The first test is that of the unit root;

$$H_0 : \rho = 0$$

$$H_A : \rho < 0$$

The t statistic for this test is denoted by τ_3 and is given by:

$$\tau_3 = \frac{\hat{\rho}}{se(\hat{\rho})}$$

If the test statisitic is less than the critical value, we reject the null.

The second test is:

$$H_0 : \rho = \phi_0 = \beta = 0$$

$$H_A : Not\ H_0$$

The test statistic for this test is denoted by ϕ_2. If the test statistic exceeds the critical value then we reject the null.

Finally the third test is:

$$H_0 : \rho = \beta = 0$$

$$H_A : Not\ H_0$$

The test statistic for this test is denoted by ϕ_3. If the test statistic exceeds the critical value then we reject the null.

2. Second, we only have allowed for AR(1). We need to extend the above testing procedure for higher order AR models. The Augmented DF (ADF) test allows for higher order lags in testing for a unit root. For example, the model with an intercept, trend, and p lags is given by:

$$\Delta y_t = \phi_0 + \beta\ t + \rho\ y_{t-1} + \sum_{i=2}^{p} \delta_i y_{t-i} + \epsilon_t \quad where\ \rho = \sum_{i=1}^{p} \phi_i - 1$$

6.8.1 Testing for unit root in USD/CAD exchange rate

In this application we will test for unit root in US-Canada exchange rate. For this purpose we work with monthly data from Jan 1971 through Oct 2018. Below I show the results for 3 models using the **urca** package in R.

```
##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression none
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.069468 -0.008833  0.000404  0.010503  0.125044
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## z.lag.1    0.0001481  0.0005784   0.256    0.798
## z.diff.lag 0.2764474  0.0399962   6.912 1.27e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01717 on 579 degrees of freedom
## Multiple R-squared:  0.0766, Adjusted R-squared:  0.07341
## F-statistic: 24.02 on 2 and 579 DF,  p-value: 9.555e-11
##
##
## Value of test-statistic is: 0.2561
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau1 -2.58 -1.95 -1.62

##
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression drift
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.067572 -0.009509 -0.000595  0.010042  0.123448
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.010623   0.005316   1.999   0.0461 *
## z.lag.1     -0.008398   0.004315  -1.946   0.0521 .
## z.diff.lag   0.279651   0.039925   7.004 6.92e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01712 on 578 degrees of freedom
## Multiple R-squared:  0.08215,    Adjusted R-squared:  0.07898
## F-statistic: 25.87 on 2 and 578 DF,  p-value: 1.74e-11
##
##
## Value of test-statistic is: -1.9462 2.03
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau2 -3.43 -2.86 -2.57
## phi1  6.43  4.59  3.78

##
```

```
## ###############################################
## # Augmented Dickey-Fuller Test Unit Root Test #
## ###############################################
##
## Test regression trend
##
##
## Call:
## lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.067581 -0.009494 -0.000606  0.009985  0.123397
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.062e-02  5.320e-03   1.997   0.0463 *
## z.lag.1     -8.460e-03  4.447e-03  -1.903   0.0576 .
## tt           2.565e-07  4.366e-06   0.059   0.9532
## z.diff.lag   2.797e-01  3.998e-02   6.997 7.27e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01714 on 577 degrees of freedom
## Multiple R-squared:  0.08216,    Adjusted R-squared:  0.07739
## F-statistic: 17.22 on 3 and 577 DF,  p-value: 1.016e-10
##
##
## Value of test-statistic is: -1.9025 1.3521 1.8924
##
## Critical values for test statistics:
##       1pct  5pct 10pct
## tau3 -3.96 -3.41 -3.12
## phi2  6.09  4.68  4.03
## phi3  8.27  6.25  5.34
```

For the first model with no constant and trend, the test statistic $\tau_1 = 0.2469$ and the 5% critical value is -1.95. For the second model with a constant, the test statistics is $\tau_2 = -1.9315$ and the 5% critical value is -2.86. Finally, for the third model with a trend, the test statistic $\tau_3 = -1.8839$ and the 5% critical value is -3.41. In each, because the test statistic is greater than the critical value, we do not reject the null hypothesis and conclude there is unit root.

6.9 Box-Jenkins Method for estimating ARIMA(p,d,q)

Box-Jenkins is a three-step procedure for finding the best fitting ARIMA(p,d,q) for a non-statinary time series.

1. Model identification: here we determine the order of integration d, and the optimal number of AR and MA components, p and q respectively.
    i. To determine d, we conduct ADF test on successive differences of the original time series. The order of integration is the number times we difference our data to obtain stationarity.
    ii. This is followed by estimating ARMA model for different combinations of p and q. The optimal structure is chosen using either AIC or BIC.

2. Parameter estimation: we estimate the identified model from the previous step using ML estimation.
3. Model Evaluation: mostly showing that the residuals from the optimal model is a white noise process. We can do this by using the Breusch-Godfrey LM test of serial correlation for the residuals. If residuals from the final model are white noise then there should be no serial correlation.

Chapter 7

Modeling Volatility

Consider a stock market analyst who is interested in finding out whether to invest in a particular stock or not. What kind of variables will govern her choice? If we focus on modern portfolio theory (MPT), a benchmark model in finance, a rational investor only cares about two variables:

1. expected return from a financial asset such as a stock
2. risk or volatility underlying this asset

Let P_t denotes the adjusted closing price for a stock traded in the market. The continually compounded return of this stock is given by:

$$y_t = [log(P_t) - log(P_{t-1})] \times 100$$

According to the MPT as an investor you should care about:

1. $E(y_{t+h}|\Omega_t)$: this is the expected return on the asset, given information at time t
2. $Var(y_{t+h}|\Omega_t)$: this is the expected volatility or risk underlying this asset, given the information at time t.

Note that even though this example is using stock market as the motivation, the same is true for any other time series of interest. I use stock market as an example because here the variance has an intuitive appeal as risk underlying an asset.

Thus far our discussion has focused on modeling the conditional mean of a given time series. For example, consider the following $ARMA(1,1)$ model for the stock market return:

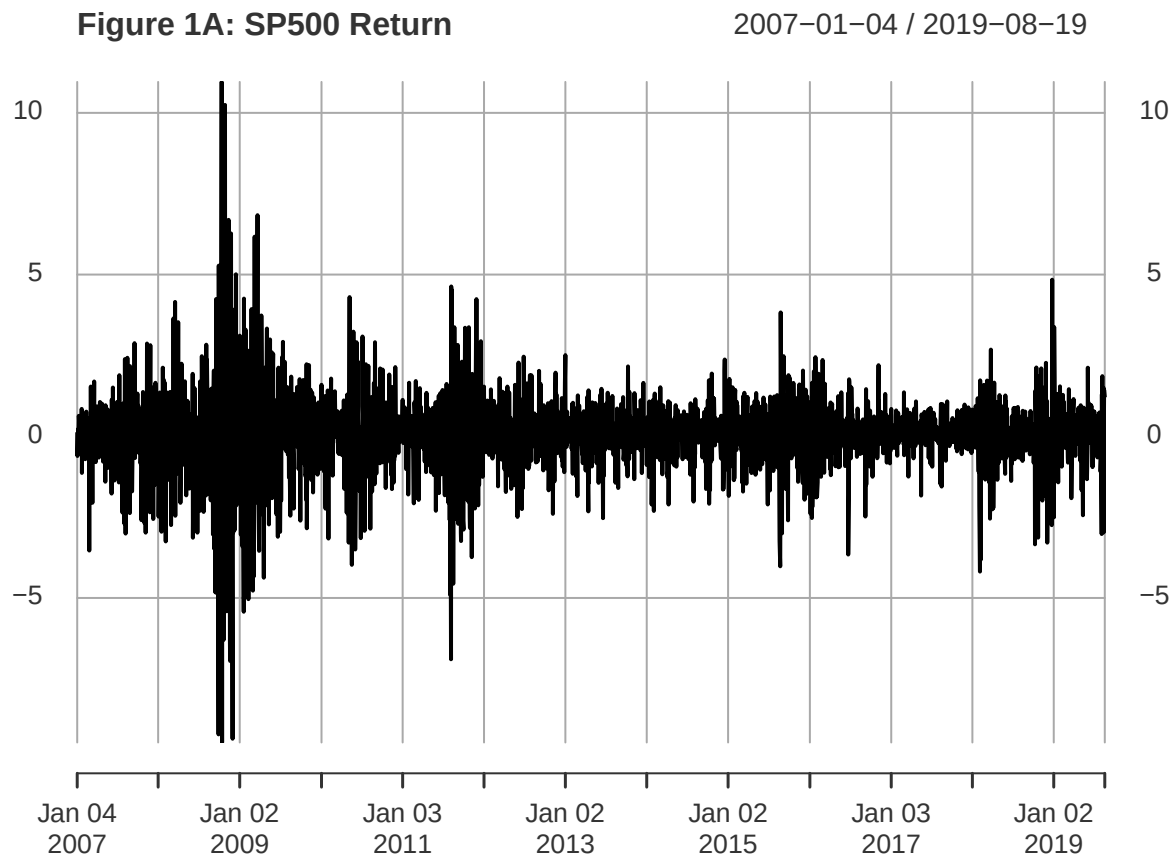
$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t; \quad \epsilon_t \sim WN(0, \sigma_\epsilon^2)$$

From this model, we can obtain the forecast for y_{t+h} as the conditional mean of y_t given the information available at time t:

$$f_{t,h} = E(y_{t+h}|\Omega_t)$$

In this sense ARMA models are inherently models for the conditional mean of a stationary time series. Going back to our stock market example, this ARMA model can give us information about how the expected return on a stock will evolve over time.

An important assumption we make in estimating this model is that the error term is homoscedastic, i.e, the error term has a constant variance across observations. However, often this assumption will not be satisfied in data. More importantly in some cases, such as our current example of stock market, making



this assumption is conceptually incorrect. This is because in our example, assuming constant variance is equivalent to assuming constant risk underlying a stock. Such an assumption is clearly not desirable for an investor.

Hence, we now need a class of models where $Var(y_t|\omega_t)$ are not constant over time. Formally, our object of interest in this topic is the conditional variance of a time series, y_t:

$$\sigma_t^2 = Var(y_t|\Omega_t)$$

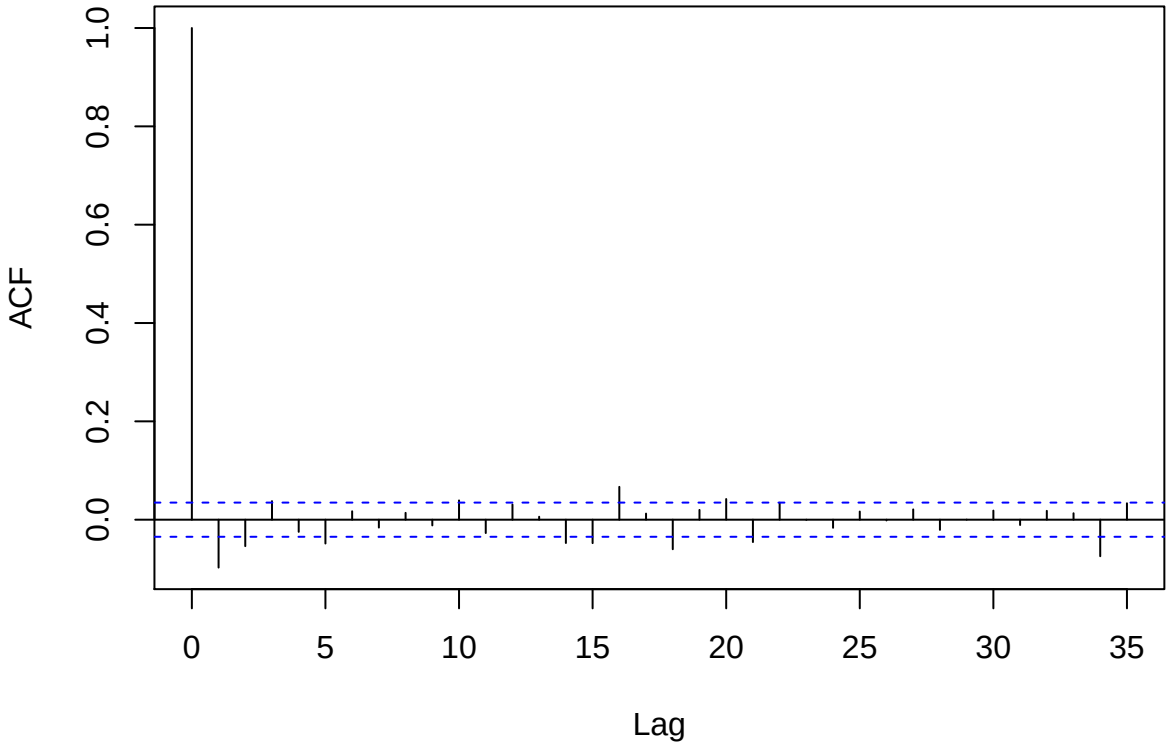
7.1 Some stylized facts about stock market volatility

Before proceeding with formally modeling the conditional variance of a time series, let us establish some stylized facts about financial assets, such as a stock. Below I plot the daily return for SP500 along with its ACF (see Figure 1A and 1B). Using squared returns as a proxy for variance, I also plot squared returns for SP500 and its ACF (see Figure 2A and 2B).

Focusing on the volatility, the plot of squared returns and its ACF establishes the following stylized facts:

1. From Figure 2A, we observe that large values of squared returns cluster together, and small values of squared returns cluster together. That is periods of high volatility are followed by periods of high volatility and periods of low volatility are followed by periods of low volatility. This phenomenon is known as **volatility persistence** or **volatility clustering** in the fields of economics and finance.

Figure 1B: ACF of SP500 Return



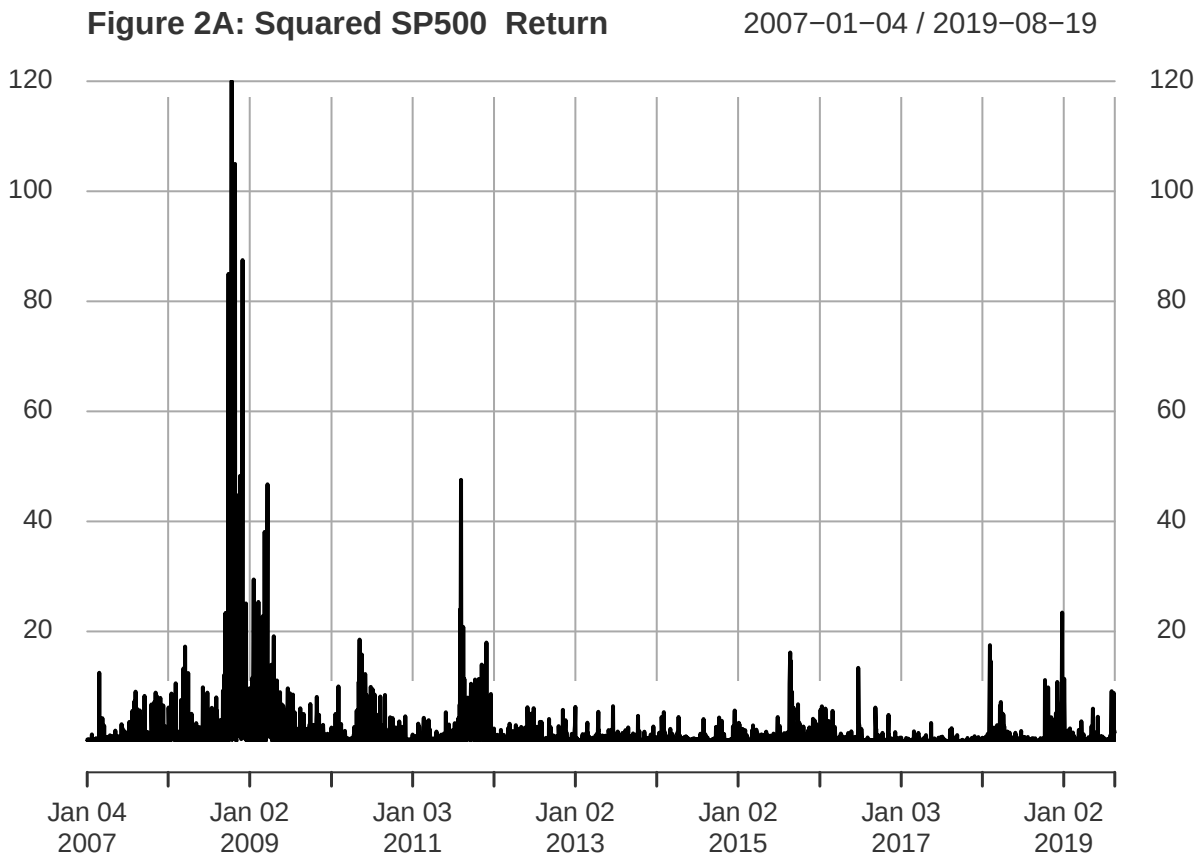
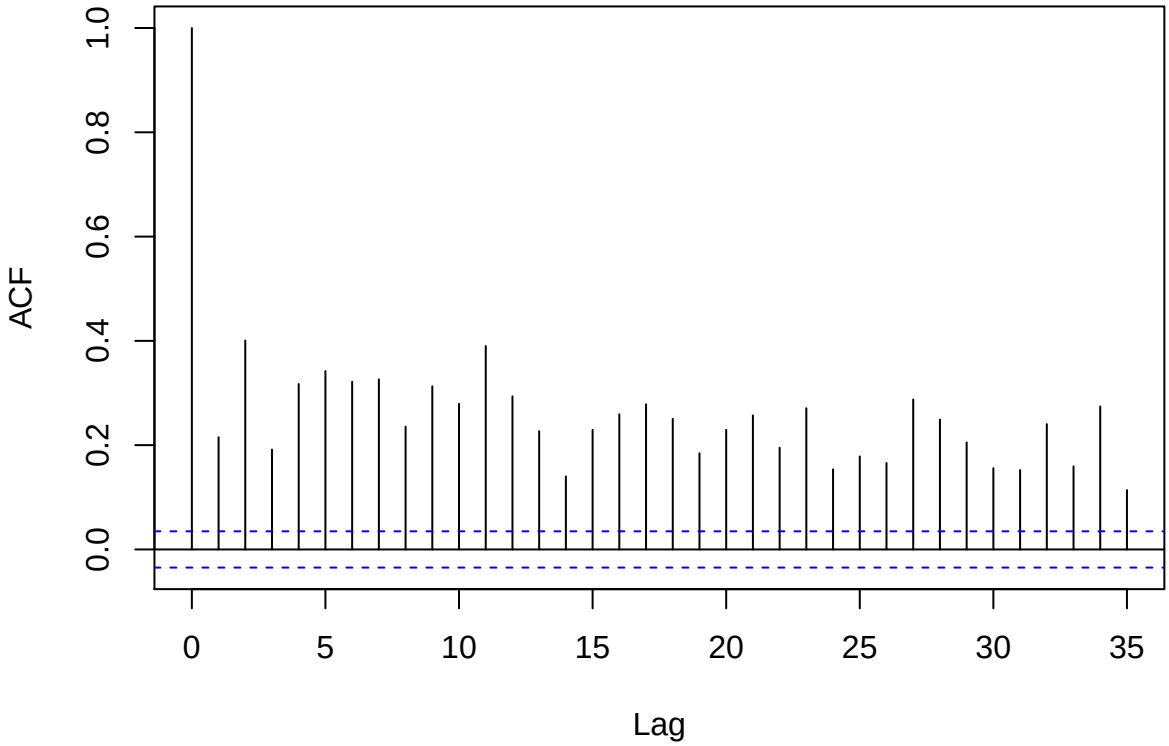


Figure 2A: Squared SP500 Return
2007−01−04 / 2019−08−19
120
100
80
60
40
20
Jan 04
2007
Jan 02
2009
Jan 03
2011
Jan 02
2013
Jan 02
2015
Jan 03
2017
Jan 02
2019

Figure 2B: ACF of Squared SP500 Return



2. A more direct evidence for volatility persistence can be inferred from the ACF plot of squared returns. From Figure 2B, we observe a strong positive serial correlation in squared returns.

Hence, it is reasonable to assume that the variance of a financial time series may not be constant over time. Next we learn two classes of models that have been suggested to model conditional variance of a time series.

7.2 ARCH(q): Autoregressive Conditional Heteroscedasticiy of order q

Engle (1982) proposed a non-linear model for the conditional variance of a stationary time series where past squared shocks affect current volatility. For simplicity, let us assume that we are not interested in modeling the mean of the time series. Hence, our model for the mean is a constant value, μ. Then, an $ARCH(1)$ model can be specified as follows:

$$\text{Mean Model:} \quad y_t = \mu + \epsilon_t \quad \text{where } \epsilon_t = \nu_t \sigma_t \text{ and } \nu_t \sim N(0,1)$$

$$\text{Variance Model:} \quad \sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 \text{where } \omega > 0 \text{ and } \alpha_1 > 0$$

In this model the unconditional variance of the time series is constant, but the conditional variance depends on the past squared error term. The variance model can be easily generalized to include q past shocks which gives us $ARCH(q)$

$$\text{Variance Model:} \quad \sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \alpha_3 \epsilon_{t-3}^2 + ... + \alpha_q \epsilon_{t-q}^2$$

$$\text{where } \omega > 0 \text{ and } \alpha_i > 0 \; \forall \; i$$

7.3 GARCH(p,q): Generalized Autoregressive Conditional Heteroscedasicity of order p and q

A generalization of the above ARCH model was proposed by Bolerslev (1986) where the conditional variance in the current period depend on past squared shocks as well as past observations of the conditional variance. This model is known as $GARCH$ which stands for generalized ARCH model. Formally, the variance equation of a $GARCH(1,1)$ is given by:

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \quad \text{where } \omega > 0, \; \alpha_1 > 0, \text{ and } \beta_1 > 0$$

In practice most models of volatility now use some version of GARCH(1,1) as it provides a more parsimonious model of volatility when compared to an ARCH(q) model. Following are few important properties of a standard GARCH(1,1) model:

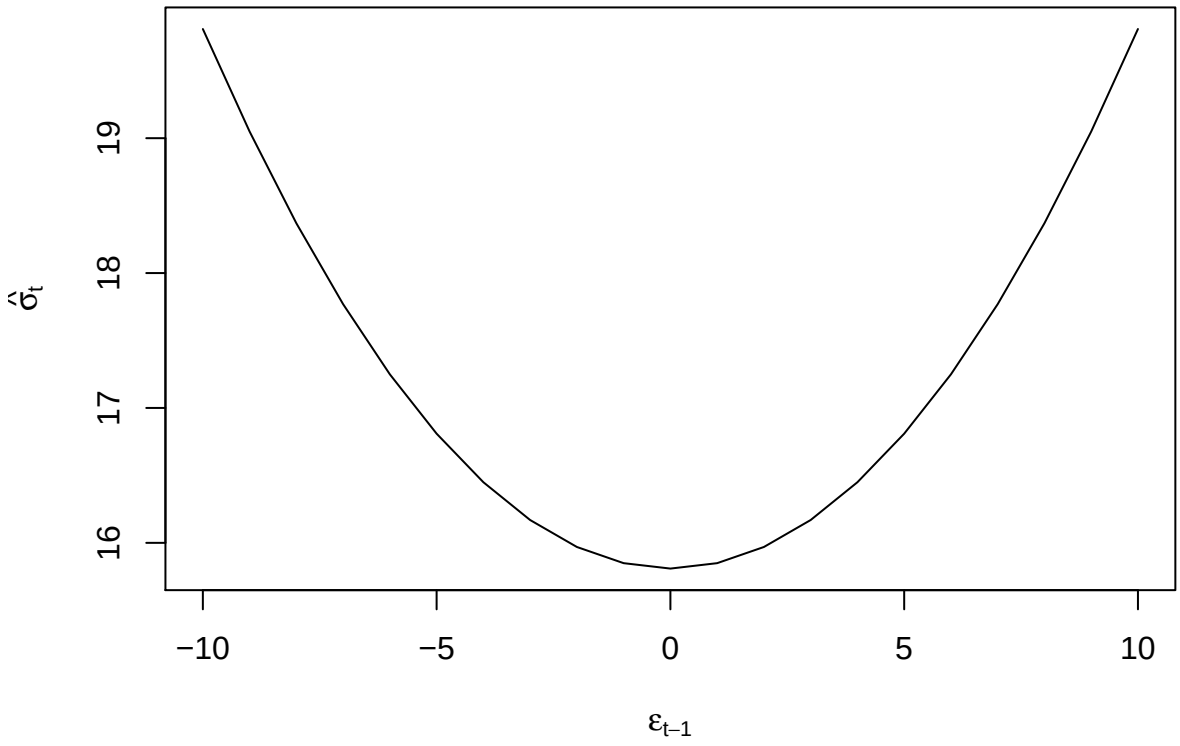
i. Unconditional variance: The unconditional variance of y_t is still constant due to stationarity and is given by:

$$\sigma_y^2 = \frac{\omega}{1 - \alpha_1 - \beta_1}$$

ii. Volatility persistence: the persistence is given by $\alpha_1 + \beta_1$.

iii. Half-life measure: R is the number of periods it takes for the estimated volatility to converge to half of the unconditional variance of the time series.

Figure 3A: Symmetric News Impact Curve



7.4 Extensions of standard GARCH model

There are two issues with the standard GARCH model that merits more discussion:

1. In the standard GARCH framework, the effect of past shocks on volatility is symmetric. That is whether the observed shock is **negative** or **positive**, the effect on the volatility is the same. However, in the financial market another stylized fact is the asymmetric response of volatility to news. For instance, it is quite common to find that negative news increases volatility more than the reduction in volatility caused by positive news. One analytical tool to illustrate this point is the **news impact curve** where we plot the shock on the x-axis and estimated/predicted volatility from our GARCH model on the y-axis. Below I plot three types of news impact curves. In the first case, the curve is symmetric: negative and positive shocks have identical effect on volatility. In the second case, negative news has a bigger effect on volatility, and finally in the third case, positive news has a bigger effect on volatility.

2. Because these models are models of variance which cannot be negative, the estimation of these models imposes non-negativity condition on all estimated parameters. For instance, in $GARCH(1,1)$ we assume that $\omega, \alpha_1,$ and β_1 are all positive.

7.4.1 GJR-GARCH(1,1)

Glosten, Jagannathan, and Runkle (1993) proposed a variation of the standard GARCH model that incorporates the asymmetry of the impact of news on volatility. The resulting model is called GJR-GARCH and it addresses the first of the issues listed above. Formally, the variance equation of the GJR-GARCH(1,1) is

Figure 3B: Asymmetric News Impact Curve−Negative news has bigger e

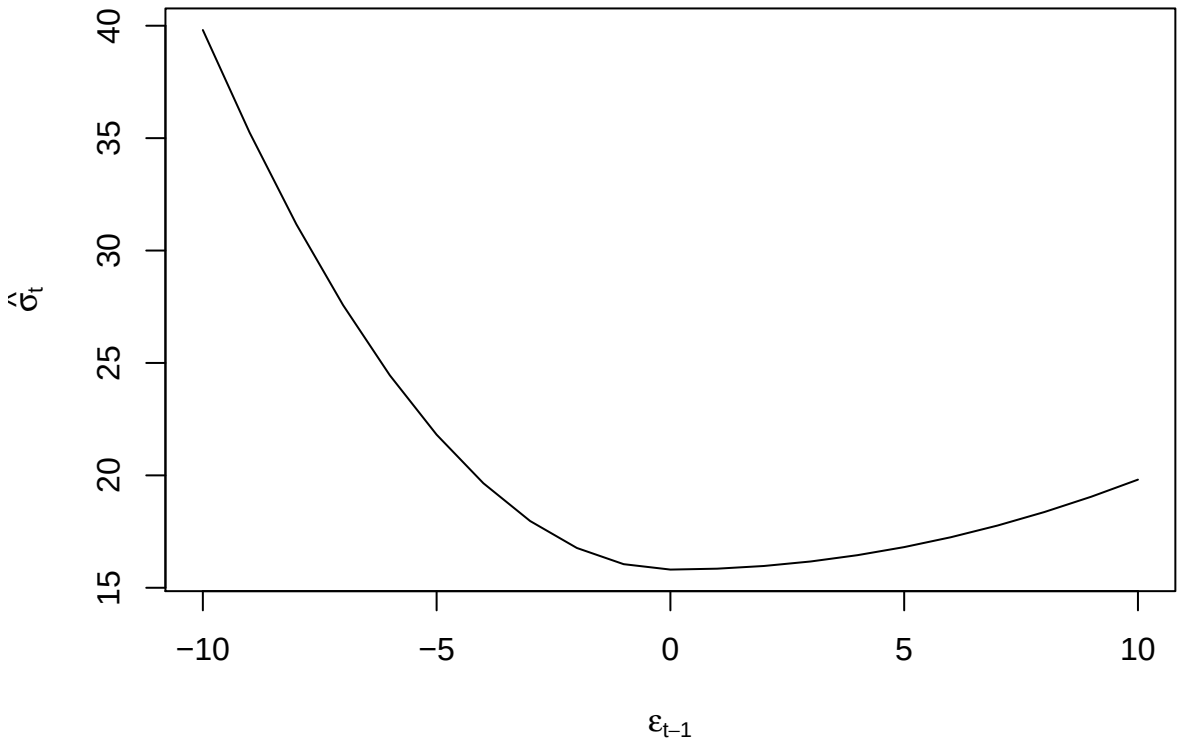
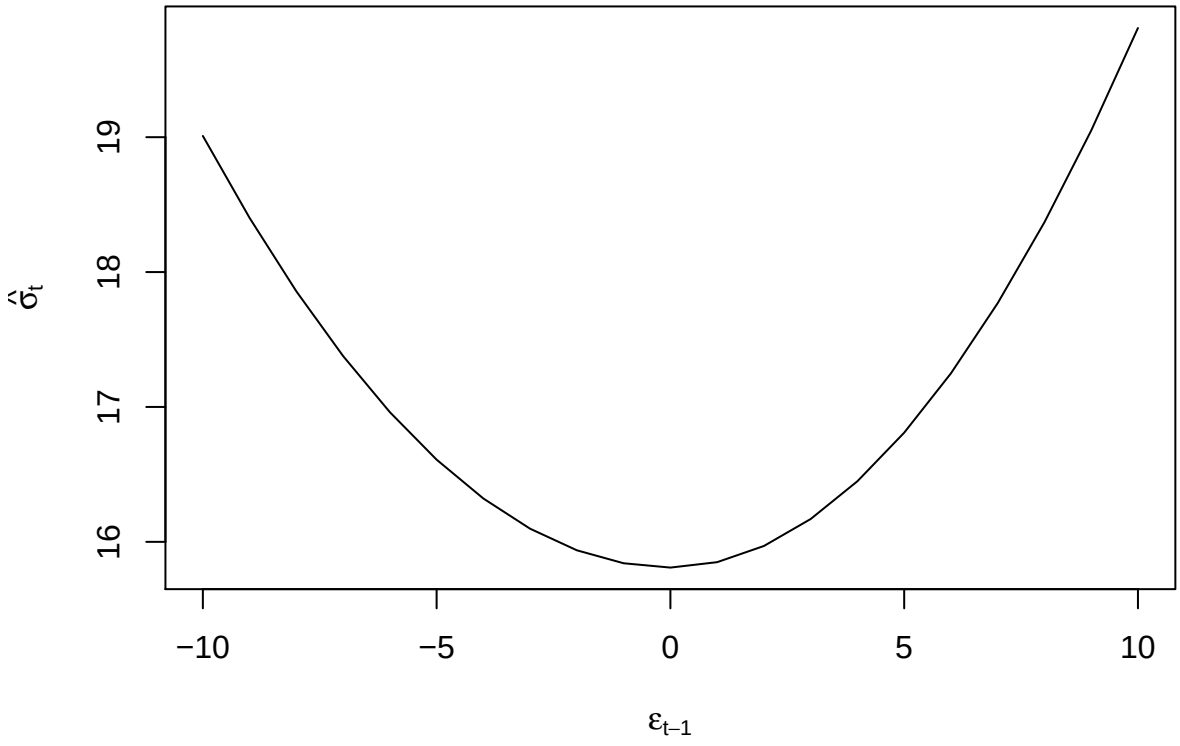


Figure 3C: Asymmetric News Impact Curve−Positive news has bigger e



given by:

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma_1 D_{t-1} \epsilon_{t-1}^2$$

where $D_{t-1} = 1$ if $\epsilon_{t-1} < 0$ and 0 otherwise. Hence, now the effect of ϵ_{t-1}^2 on volatility is $\alpha_1 + \gamma_1$ for negative shocks and α_1 for positive shocks. The news impact curve from this model will be asymmetric. The persistence from this model will also be affected by γ_1. Specfically for the GJR-GARCH(1,1) model,

1. Persistence: $\alpha_1 + \beta_1 + \frac{\gamma_1}{2}$
2. Unconditional variance:
$$\sigma_y^2 = \frac{\omega}{1 - \alpha_1 - \beta_1 - \frac{\gamma_1}{2}}$$

7.4.2 Exponential GARCH or EGARCH(1,1)

Exponential GARCH model by Nelson (1991) accounts for both the issues outlined above. It allows for asymmetric effects of shocks and also does not require the non-negativity constraints. More importantly, it also accounts for different effect of shocks of different magnitude, and hence provides an estimate of the **size effect**. Formally, the variance equation for EGARCH(1,1) is given by:

$$ln(\sigma_{t-1}^2) = \omega + \alpha_1 z_{t-1} + \gamma_1 (|z_{t-1}| - E|z_{t-1}|) + \beta_1 \sigma_{t-1}^2$$

where $z_t = \frac{\epsilon_t}{\sigma_t}$ is the standardized error term. Here, α_1 captures the sign effect and γ_1 captures the size effect. A common finding is that negative news and larger shocks have bigger effect on volatility. Accordingly, we often find in empirical applications of EGARCH that $\alpha_1 < 0$ and $\gamma_1 > 0$. For this model, we have

1. Persistence: β_1
2. Unconditional variance:
$$\sigma_y^2 = \frac{\omega}{1 - \beta_1}$$

7.5 Application of GARCH model: Estimating volatility of SP500 return

In this application, we will estimate the volatility underlying SP500 returns (Figure 1A and Figure 2A). The first step is to test whether squared returns have ARCH effects i.e, whether there is any evidence for time-varying volatility in our data. For this purpose we will use Engle's ARCH test. Consider our constant mean model:

$$y_t = \mu + \epsilon_t$$

We can estimate the above model by OLS and obtain residuals $e_t = y_t - \hat{y}_t$. The test for ARCH effects is based on the idea that if there is conditional hetroscedasiticity in our data then the squared residuals will have serial correlation. The ARCH test involves estimating the following regression:

$$e_t^2 = \beta_0 + \beta_1 e_{t-1}^2 + \beta_2 e_{t-2}^2 + ... + \beta_p e_{t-p}^2 + u_t$$

Then, the test for ARCH effects is given by:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \ \Rightarrow \text{no ARCH effects}$$

$$H_A : \text{Not } H_0$$

Table 7.1: (A) Estimated GARCH(1,1)

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.068	0.014	4.984	0
omega	0.028	0.004	6.907	0
alpha1	0.137	0.013	10.316	0
beta1	0.843	0.013	63.877	0

In R, we use a package called **aTSA** to implement this test.The function is called **arch.test()**. Figure 7.1 below shows the results of this test for our data.

```
library(aTSA)
library(forecast)

fit=arima(y, c(0,0,0))

arch.test(fit)
```

```
## ARCH heteroscedasticity test for residuals
## alternative: heteroscedastic
##
## Portmanteau-Q test:
##      order   PQ p.value
## [1,]     4 1097       0
## [2,]     8 2317       0
## [3,]    12 3633       0
## [4,]    16 4240       0
## [5,]    20 4963       0
## [6,]    24 5602       0
## Lagrange-Multiplier test:
##      order   LM p.value
## [1,]     4 2247       0
## [2,]     8  709       0
## [3,]    12  412       0
## [4,]    16  279       0
## [5,]    20  220       0
## [6,]    24  175       0
```

We find strong evidence for ARCH effects in our data as the null hypothesis of no ARCH effects is rejected at different orders of serial correlation in squared residuals. Next we estimate three types of GARCH(1,1) models using the **rugarch** package. Tables 7.1A-7.1D below show the estimated parameters of these three models. The news impact curve for these 3 classes of GARCH model are plotted in Figure 2-4 below. Finally, the estimated conditional volatility from the three models are plotted along with the return in Figures 5-7. We find that there is strong evidence for the sign effect with negative news having a bigger effect on volatility as indicated by the positive value for γ_1 in GJR-GARCH and negative value of α_1 in EGARCH. Further, in the EGARCH model we find evidence for the size effect as indicated by positive value for γ_1. These findings are confirmed by the asymmetric news impact curves for the GJR-GARCH and EGARCH models.

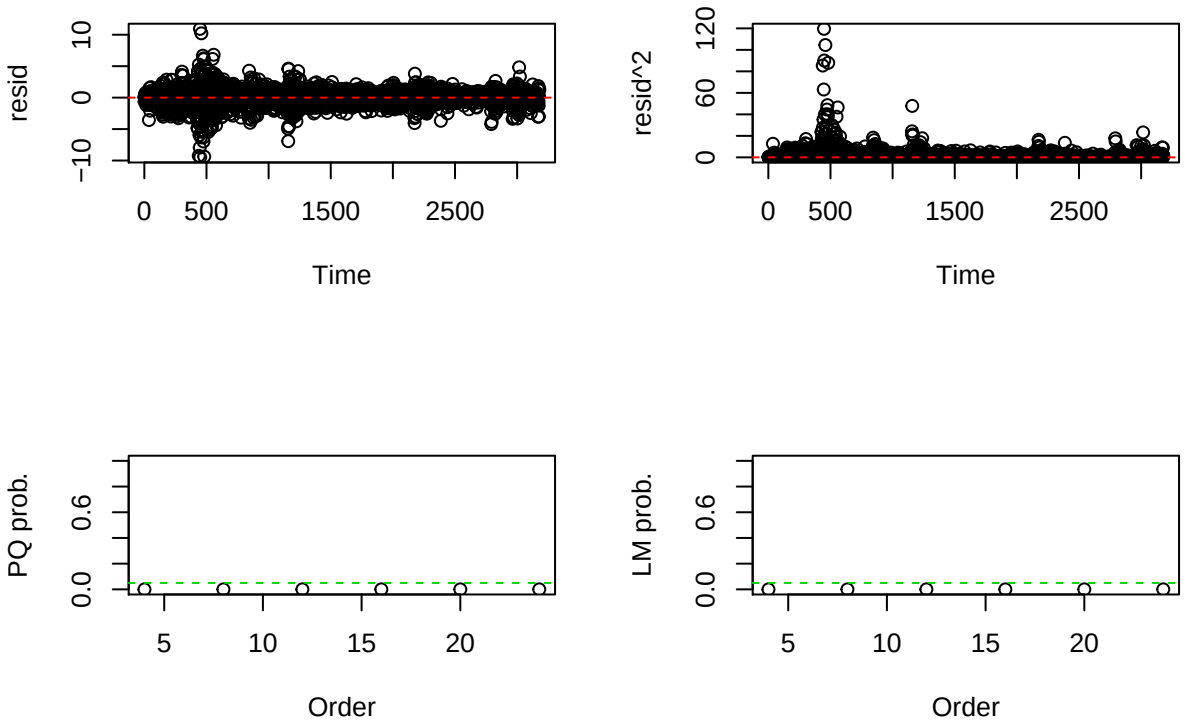


Figure 7.1: ARCH LM test

Table 7.2: (B) Estimated GJR-GARCH(1,1)

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.028	0.014	2.082	0.037
omega	0.028	0.003	8.234	0.000
alpha1	0.000	0.013	0.000	1.000
beta1	0.861	0.013	66.481	0.000
gamma1	0.225	0.023	9.631	0.000

Table 7.3: (C) Estimated EGARCH(1,1)

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.031	0.012	2.499	0.012
omega	0.000	0.003	-0.118	0.906
alpha1	-0.182	0.013	-13.795	0.000
beta1	0.968	0.004	266.919	0.000
gamma1	0.160	0.015	10.974	0.000

Figure 2: News Impact Curve for GARCH(1,1)

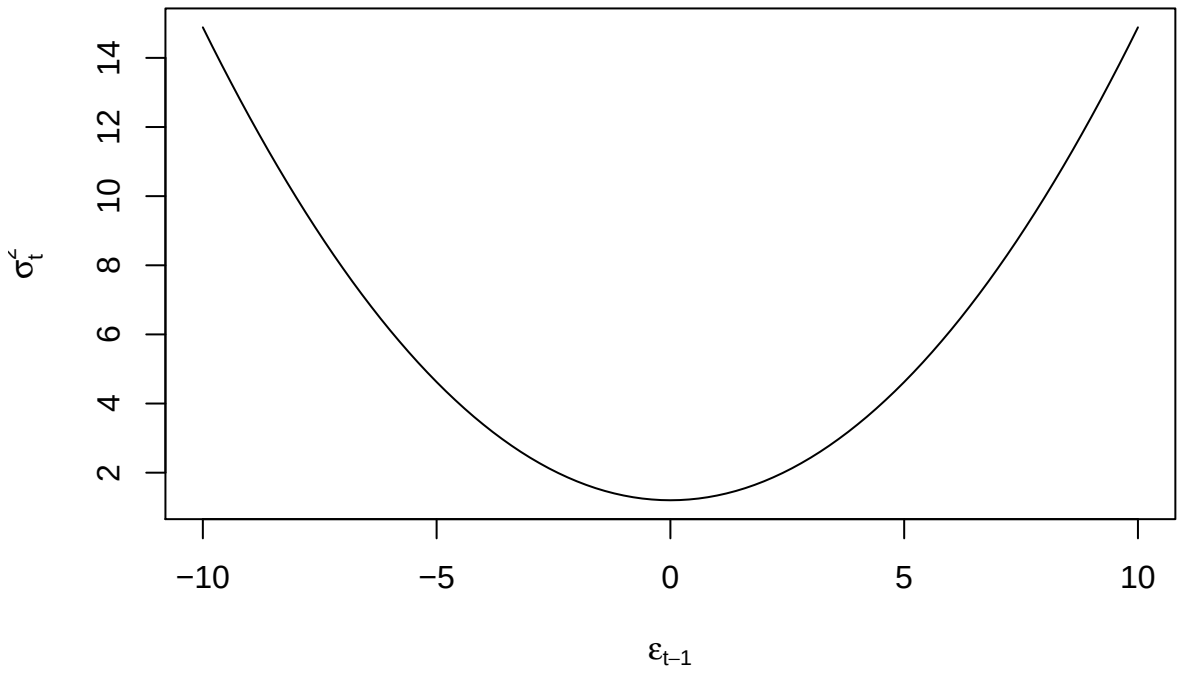


Table 7.4: (D) Persistence, Unconditional Variance, and Half-life

	GARCH(1,1)	GJR-GARCH(1,1)	EGARCH(1,1)
Persistence	0.980	0.974	0.968
Unconditional Variance	1.393	1.058	0.987
Half-life	34.552	26.109	21.172

Figure 3: News Impact Curve for GJR−GARCH(1,1)

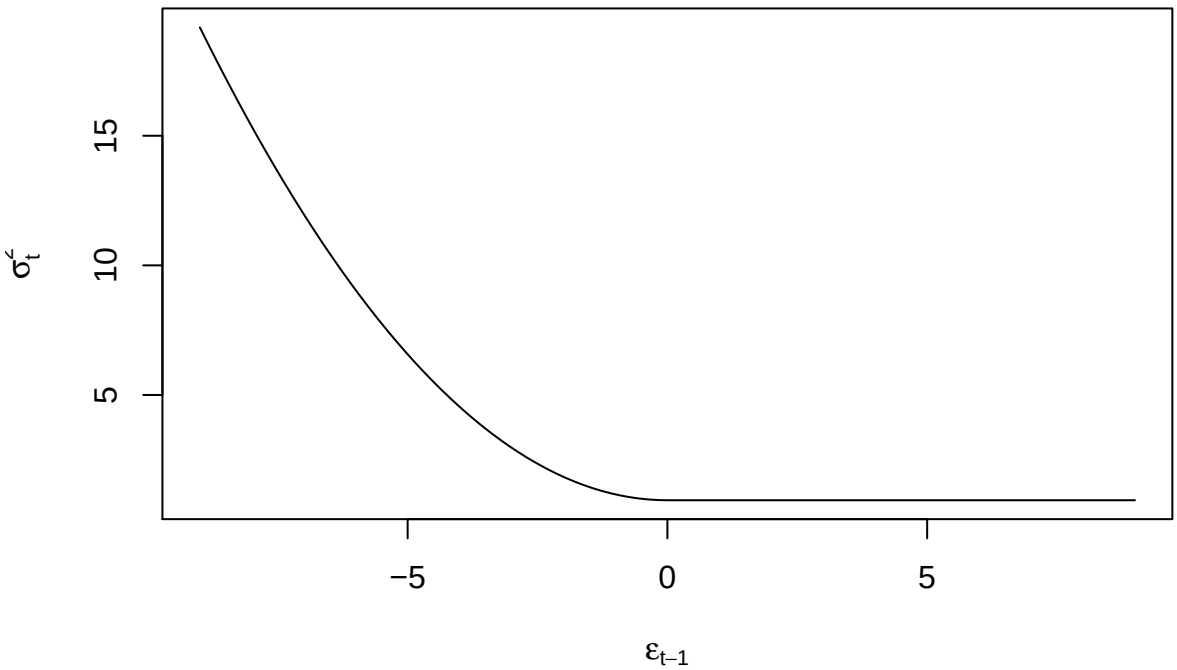


Figure 4: News Impact Curve for EGARCH(1,1)

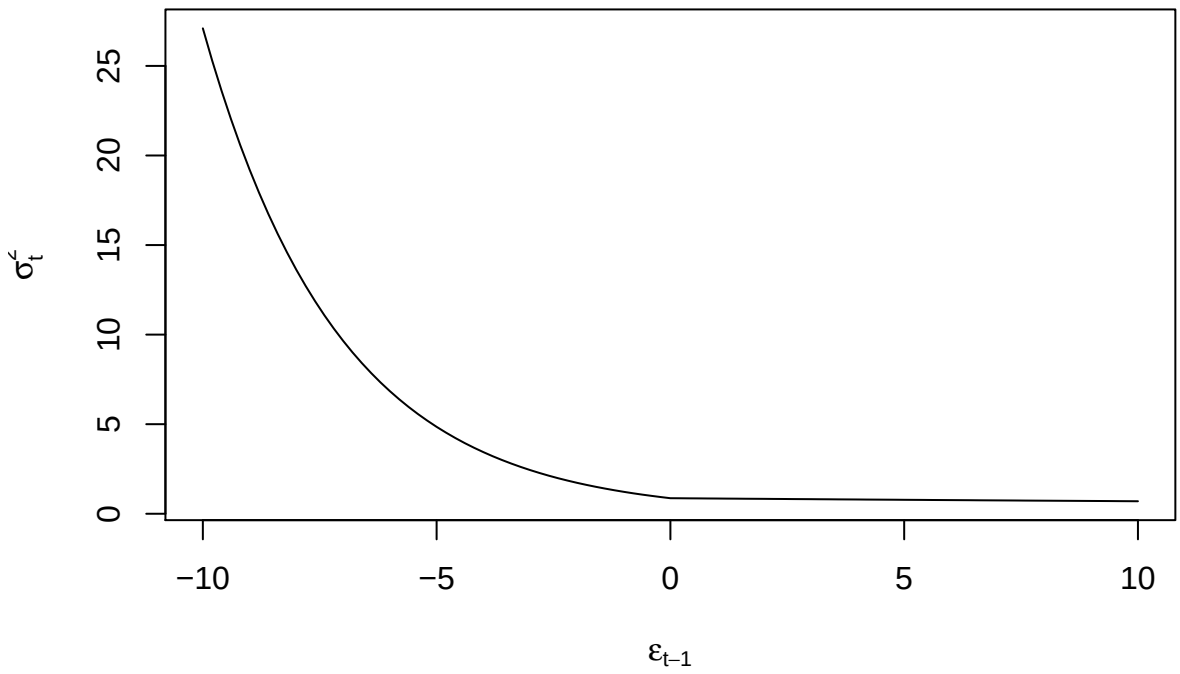


Figure 5: Estimated Conditional Volatility from GARCH(1,1)

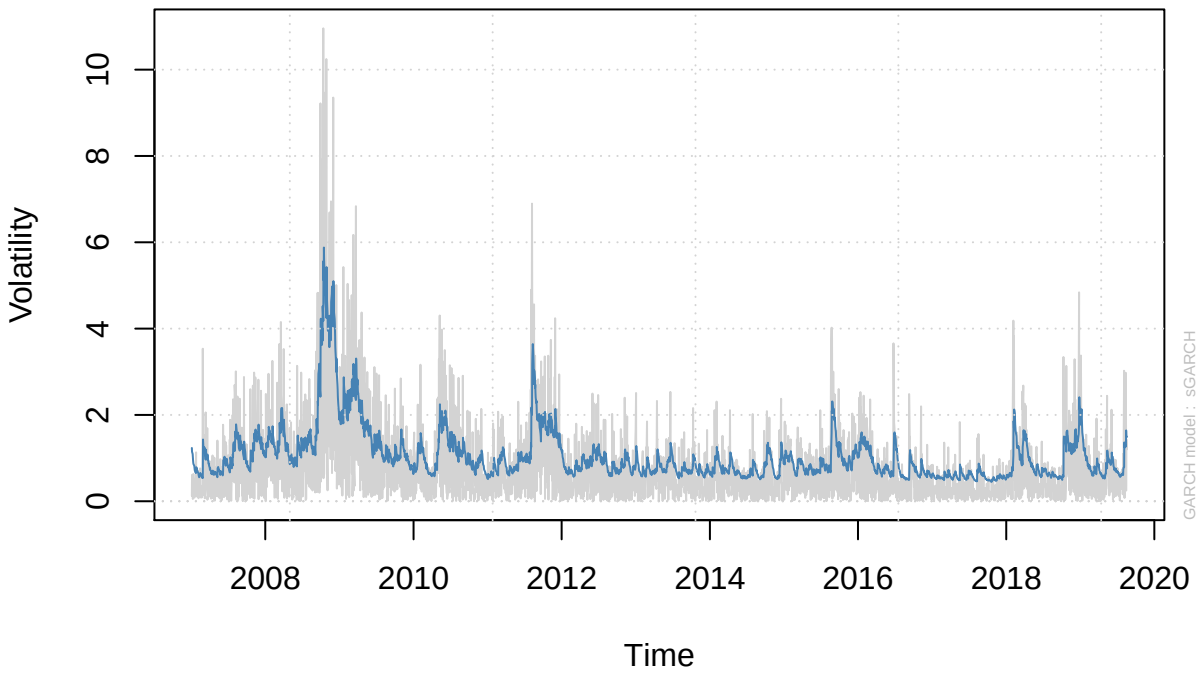


Figure 6: Estimated Conditional Volatility from GJR−GARCH(1,1)

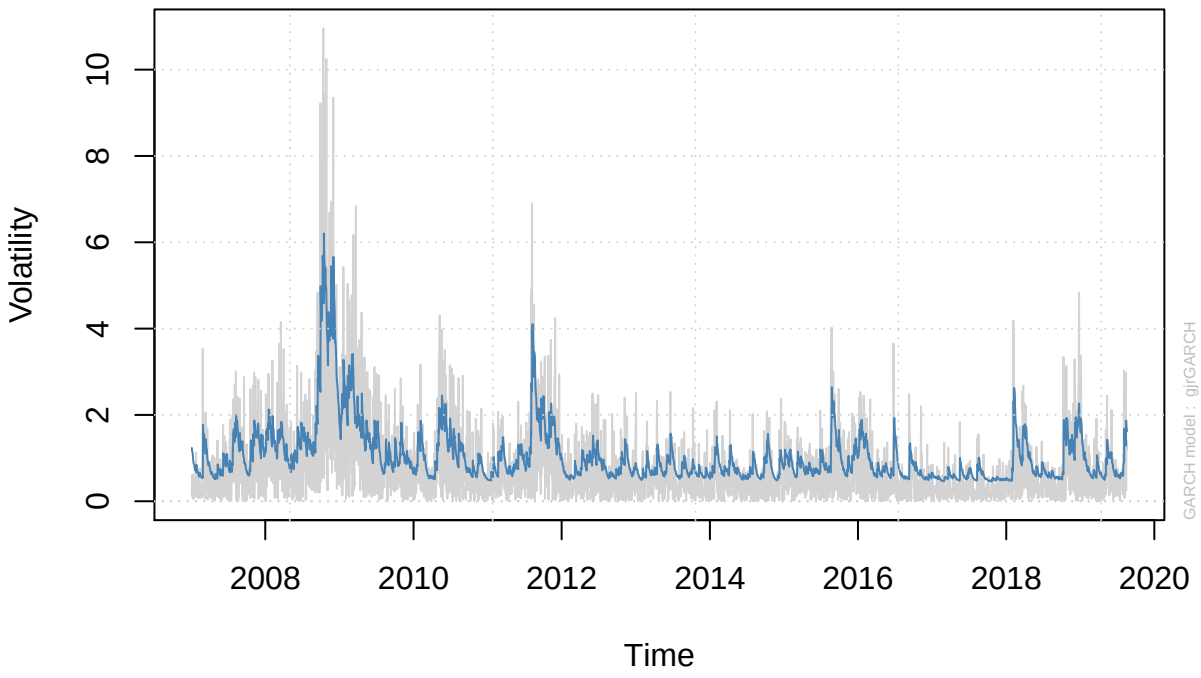
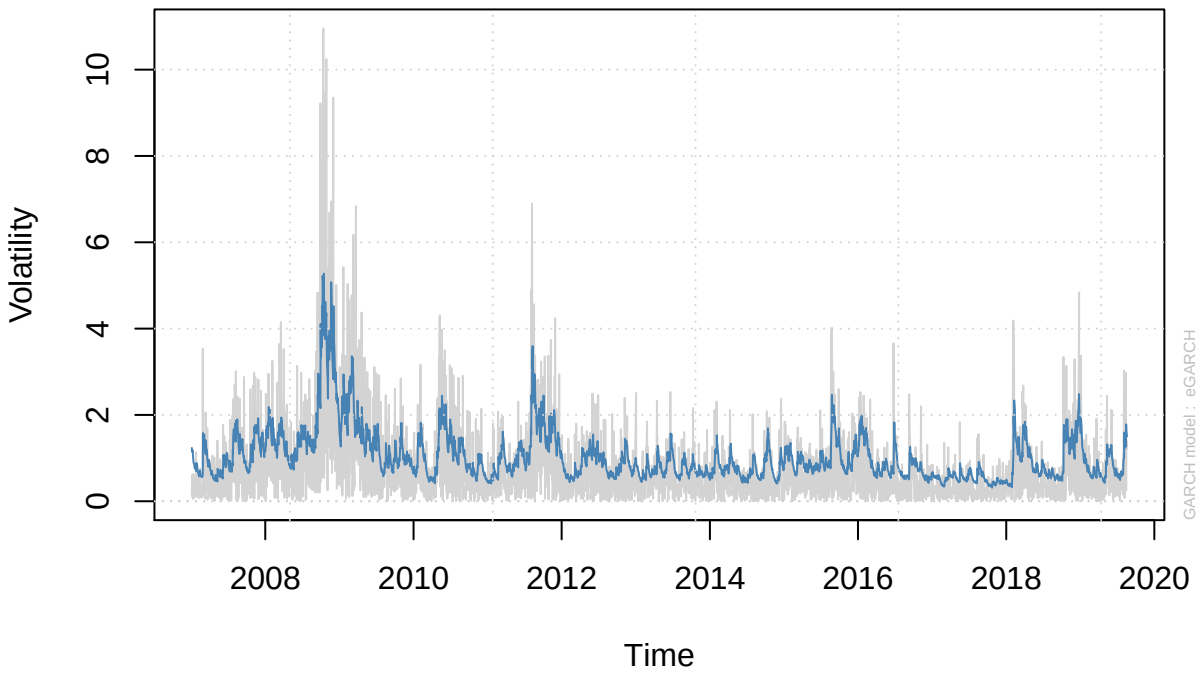


Figure 7: Estimated Conditional Volatility from EGARCH(1,1)



Finally, these models can also be used to forecast the volatility of SP500. Figure 8 below shows this forecast for the next 7 days from the EGARCH(1.1) model.

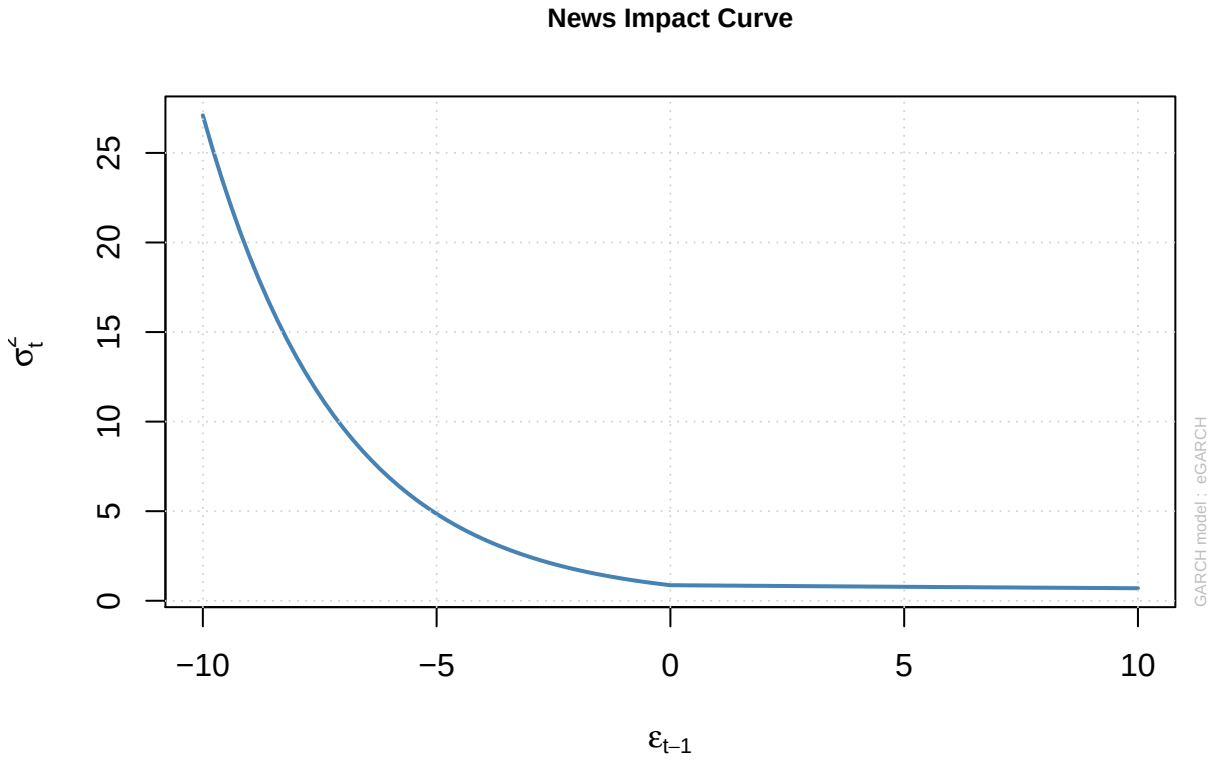
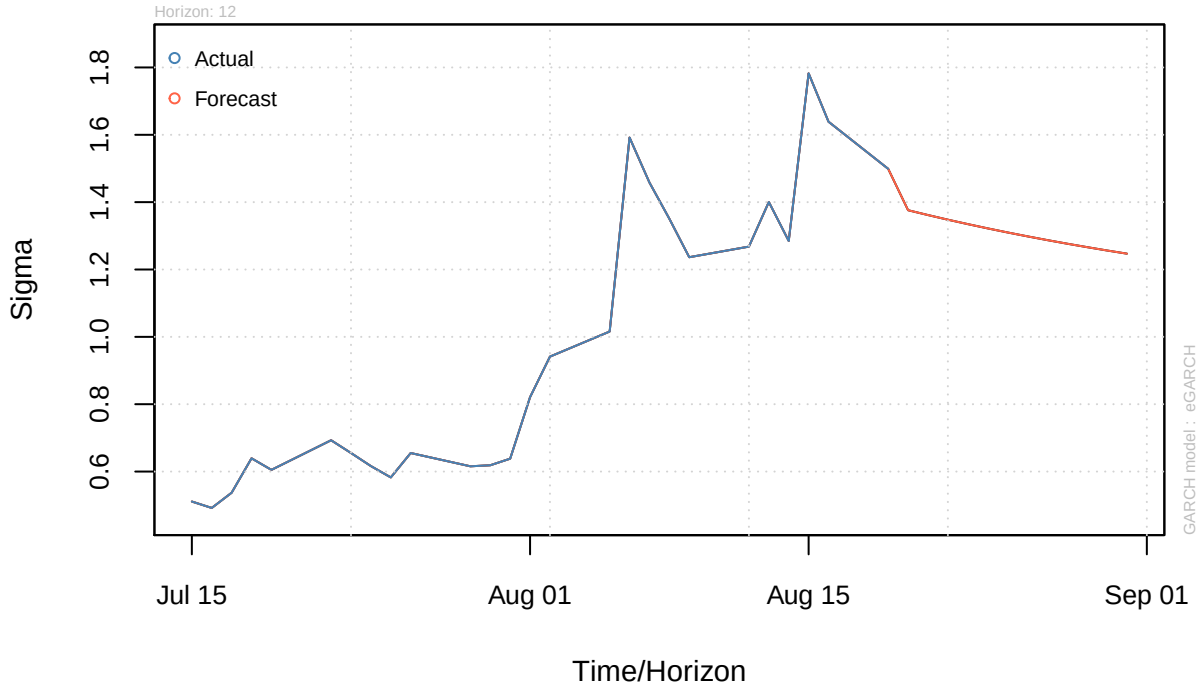

News Impact Curve
σ_t^2
0
5
10
15
20
25
−10
−5
0
5
10
ε_{t-1}
GARCH model : eGARCH

Figure 8: Forecasted Conditional Volatility from EGARCH(1,1)



Appendix A

Review of Differential Calculus and Optimization

Given that all students must have taken a course in calculus before enrolling for this class, it is assumed that everyone in the class is comfortable with concepts such as derivatives, partial derivatives, and optimization. In this chapter, I will provide a brief review of some concepts that are most pertinent for Econometrics. I strongly encourage that you read your lecture notes for Calculus if you find it difficult to follow the material presented in this chapter.

A.1 Derivative of a single variable function

Definition A.1 (Derivative of a function)**.** Consider the following function, $y = f(x)$. The *derivative* of this function measures the rate of change in y caused by a change in x.

There are two alternative notations for the derivative of y with respect to x: $f'(x)$ or $\frac{dy}{dx}$.

The derivative of a function is very closely related to the concept of *slope* of a function. Let Δ denotes change in a variable. Then, by definition, the slope of y with respect to x is given by:

$$slope = \frac{\Delta y}{\Delta x}$$

The derivative of y with respect to x is the limit value of the slope as $\Delta x \to 0$. Hence,

$$\frac{dy}{dx} \text{ or } f'(x) = \lim_{\Delta x \to 0} \left(\frac{\Delta y}{\Delta x} \right)$$

A.1.1 Rules of Differentiation

1. Derivative of a constant is 0.
2. Derivative of a function multiplied by a constant is constant times the derivative of the function:

$$\frac{d}{dx}[\ a \times f(x)] = a \times f'(x)$$

where it is assumed that a is an constant.

3. Addition rule:

$$\frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$$

4. Subtraction rule:

$$\frac{d}{dx}[f(x) - g(x)] = f'(x) - g'(x)$$

5. Product rule:

$$\frac{d}{dx}[f(x) \times g(x)] = f(x) \times g'(x) + g(x) \times f'(x)$$

6. Quotient rule:

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{f'(x) \times g(x) - g'(x) \times f(x)}{g(x)^2}$$

7. Chain rule:

$$\frac{d}{dx}[f(g(x))] = f'(g(x)) \times g'(x)$$

8. Derivative of some common functions:
   a. Power function: $f(x) = x^a$. Then,
   $$f'(x) = a \times x^{a-1}$$
   b. Natural log function: $f(x) = ln(x)$. Then,
   $$f'(x) = \frac{1}{x}$$
   c. Exponential function: $f(x) = e^x$
   $$f'(x) = e^x$$

A.2 Second derivative and non-linearity

Definition A.2 (Second derivative of a function)**.** Consider the following function, $y = f(x)$. The *second derivative* of this function measures the change in the rate of change of this function. Formally it is denoted by $f''(x)$ or $\frac{d^2y}{dx^2}$.

The second derivative measures the *curvature* of the function and hence can be used to distinguish a *linear* function from a *non-linear* function. By definition, a linear function has a constant slope implying the its second derivative must be zero.

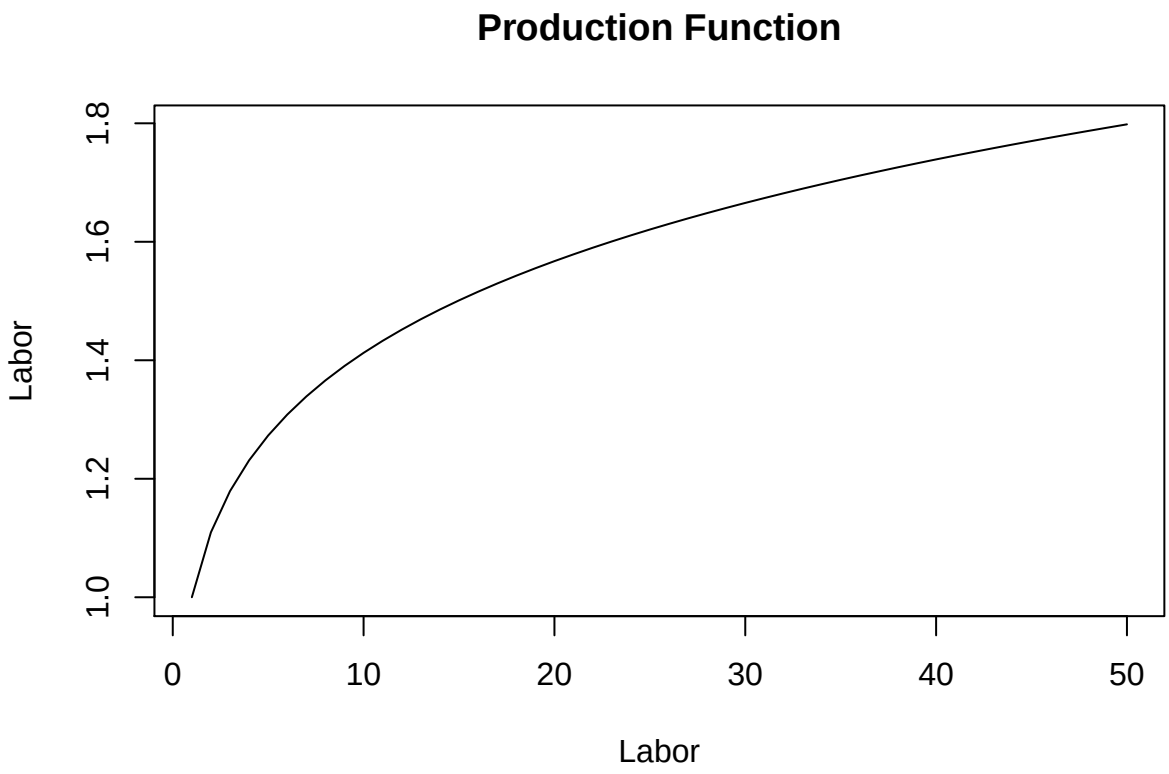
Example A.1. For example, consider the following linear function:

$$f(x) = mx + b$$

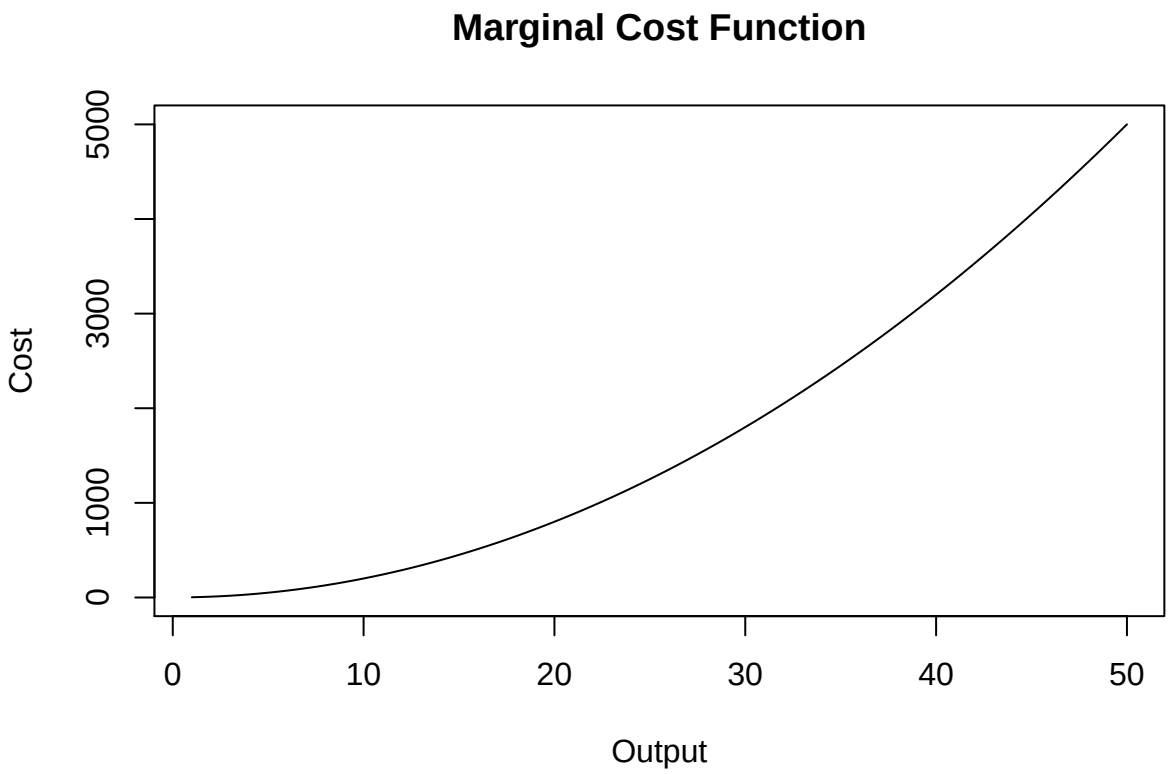
Here$f'(x) = m$ and $f''(x) = 0$.

A non-linear function will have a non-zero second derivative. There are only two possibilities:

1. $f''(x) < 0$. In this case we have a concave relationship. An example from economics is the production function where the relationship between output and input is concave.



2. $f''(x) >$. In this case we have a convex relationship. An example from economics is the marginal cost function where the relationship between cost of production and level of output can be convex.



A.3 Partial derivatives: Multi-variable functions

Ceteris paribus aka *holding other things equal* is one of the key concepts used in Economic analysis. A *partial derivative* is a mathematical counterpart of this assumption.

Definition A.3 (Partial Derivative)**.** Consider a function of n-variables given by $y = f(x_1, x_2, x_3, ..., x_n)$. Then, there are n partial derivatives of this function that can be obtained by taking derivative with respect to one of the x-variables, holding all other constant. Formally, the partial derivative of y with respect to x_i is denoted by f_{x_i}or$\frac{\partial y}{\partial x_i}$.

Example A.2. Consider the following 3-variable function:

$$y = ln(x_1) + x_1 \times x_2 + 3x_2^2 + x_1 \times x_3 + ln(x_3)$$

Then we can compute three partial derivatives of this function:

- Partial derivative of y with respect to x_1, treating x_2 and x_3 as constants:

$$\frac{\partial y}{\partial x_1} = \frac{1}{x_1} + x_2 + x_3$$

- Partial derivative of y with respect to x_2, treating x_1 and x_3 as constants:

$$\frac{\partial y}{\partial x_2} = x_1 + 6x_2$$

- Partial derivative of y with respect to x_3, treating x_1 and x_2 as constants:

$$\frac{\partial y}{\partial x_3} = x_1 + \frac{1}{x_3}$$

Example A.3 (Cobb-Douglas Production Function)**.** One of the most used functional form for the production function is the Cobb-Douglas production function. Suppose you have two inputs: labor (L) and capital (K). Let Y denotes output. Then, the Cobb-Douglas production function is given by:

$$Y = L^{\beta_1} K^{\beta_2}$$

Now, output can change because we change our labor input or our capital input. In each case, we are thinking about a change in output caused by change in one input, holding the other input constant. This is exactly what a partial derivative captures! In what follows next we will use two mathematical concepts to further our understanding of economics of production:

1. Change in natural logs of a variable approximates percent change in that variable. Formally, $\Delta ln(x) \times 100 \approx$ % change in x. Hence, it is often useful to express economic relationships in natural logs. The Cobb-Douglas production function in natural logs is given by:

$$ln(Y) = \beta_1 \times ln(L) + \beta_2 \times ln(K)$$

2. The partial derivative of the above equation gives us **elasticity of output** with respect to each input.

   a. Output elasticity of Labor:

$$\frac{\%\text{ change in Y}}{\%\text{ change in L}} = \frac{\partial ln(Y) \times 100}{\partial ln(L) \times 100} = \beta_1$$

   b. Output elasticity of Capital:

$$\frac{\%\text{ change in Y}}{\%\text{ change in K}} = \frac{\partial ln(Y) \times 100}{\partial ln(K) \times 100} = \beta_2$$

Note that we can also infer whether production is subject to increasing, decreasing, or constant returns to scale from the numerical values assigned to β_1 and β_2. Returns to scale is simply the sum of output elasiticities with respect to labor and capital:

$$\text{Returns to scale} = \frac{\%\text{ change in Y}}{\%\text{ change in L}} + \frac{\%\text{ change in Y}}{\%\text{ change in K}} = \beta_1 + \beta_2$$

Hence, we obtain constant returns to scale as long as $\beta_1 + \beta_2 = 1$. We get decreasing returns to scale if $\beta_1 + \beta_2 < 1$. Finally, increasing returns to scale require $\beta_1 + \beta_2 > 1$.

A.4 Optimization

In Economics it is often assumed that rational individuals *optimize*. For instance, firms seek to maximize profits (or minimize costs) and households seek to maximize utility. Mathematically, this is equivalent to finding **extreme** values of an **objective function**.

Example A.4. Consider a firm that is choosing a level of output (q) to maximize its profits. By definition, profits are total revenue $R(q)$ minus total cost $C(q)$. The resulting profit function $\pi(q)$ is the firm's objective function and q is the control variable:

$$\pi(q) = R(q) - C(q)$$

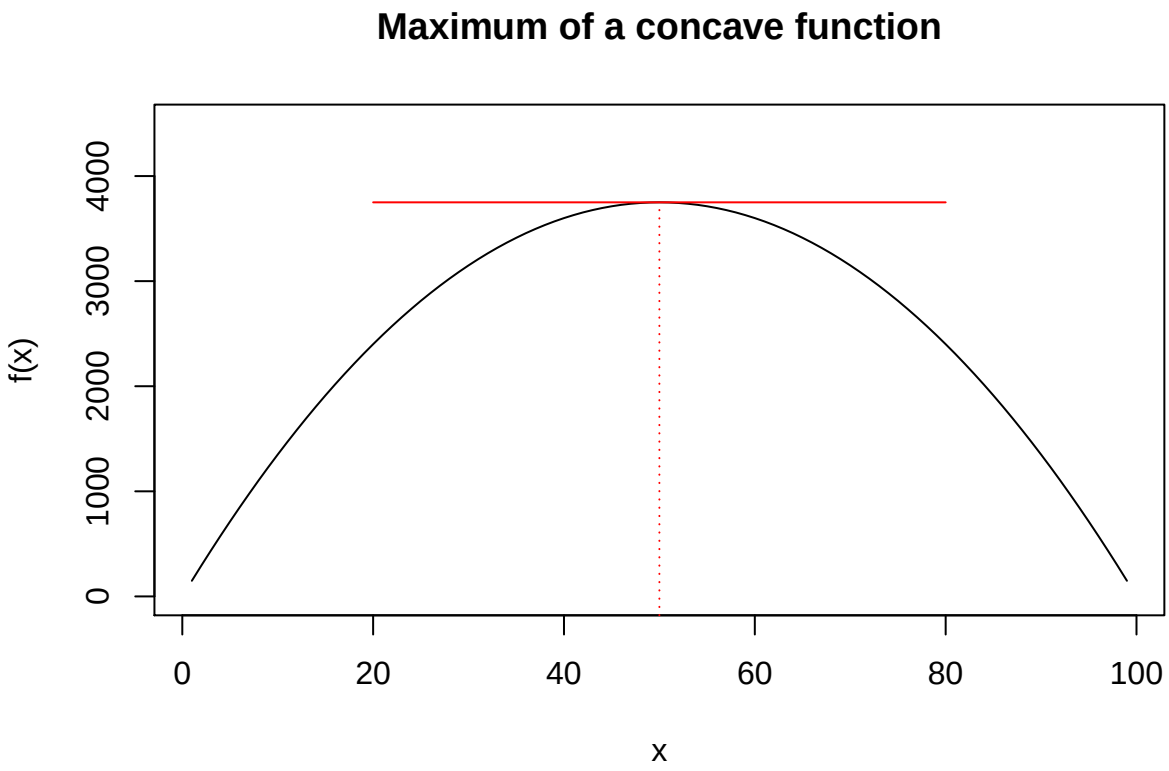
The firm will choose a value of q that will maximize its profits. Mathematically, this can be written as:

$$\max_q \pi(q)$$

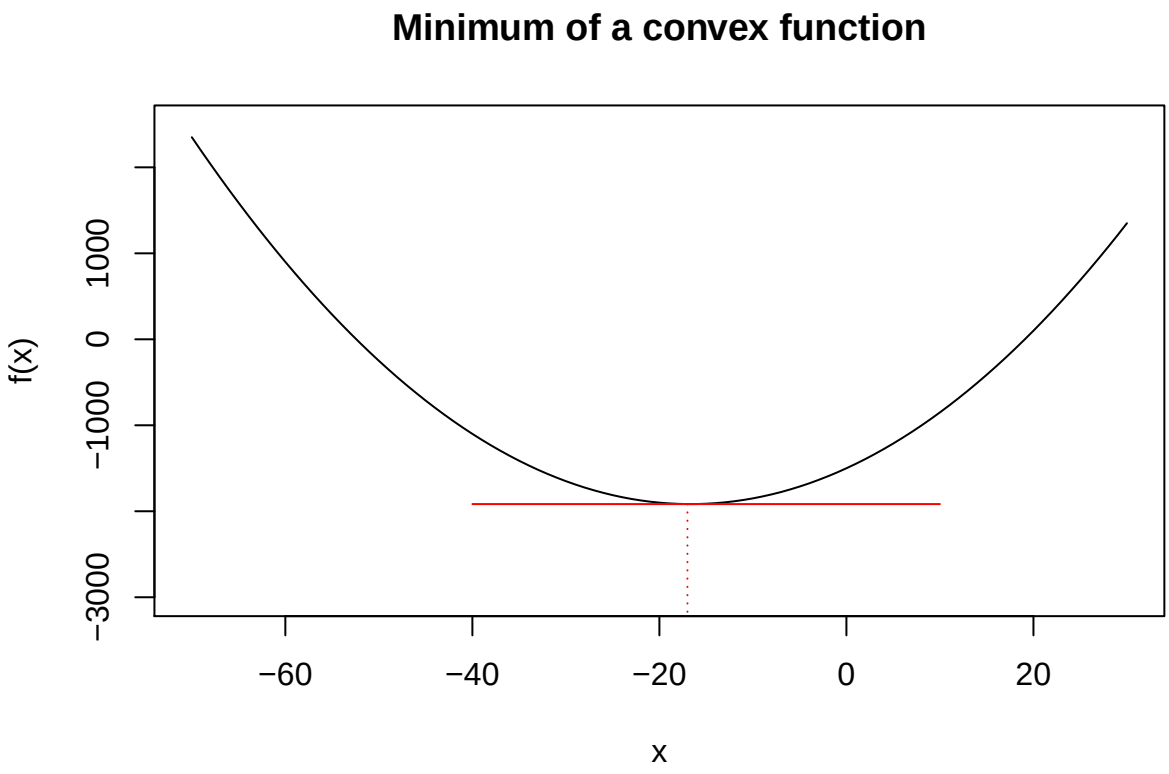
One way to solve this problem, is to assume a functional form for profits and evaluate this function for all possible values of q. Then, select the value of q that yields highest value for profits. This approach is called **numerical optimization** and is often used for complicated objective functions. But in many cases, we can use calculus and obtain an *analytical* solution for the optimization problem.

Formally, suppose the objective function is denoted by $f(x)$ and assume that this function is continuous and twice differentiable. Then,

1. x^* is a maximizer if $f(x^*) \geq f(x)$ for all $x \neq x^*$. Note that at this point the slope of the tangent to the function is 0, i.e., $f'(x^*) = 0$. This is the **first order condition (foc)** for obtaining a maximum. The graph below illustrates the maximum of a generic function. Note that the slope of the function changes sign from positive to negative around x^*. This will give us the **second order condition** for obtaining a maximum.



2. x^* is a minimizer if $f(x^*) \leq f(x)$ for all $x \neq x^*$. Note that at this point the slope of the tangent to the function is 0, i.e., $f'(x^*) = 0$. This is the **first order condition (foc)** for obtaining a minimum. The graph below illustrates the minimum of a generic function. Note that the slope of the function changes sign from negative to positive around x^*. This will give us the **second order condition** for obtaining a minimum.



Note for a maximum, Similarly, . We can now outline the steps for computing a maximum or minimum of a given function.

1. First-order condition: Compute the first derivative of the function and equate it to 0. The solution to this equation gives us x^*:

$$f'(x^*) = 0$$

2. Second-order condition: Compute the second derivative of the function and evaluate it at x^*.
   a. If $f''(x^*) < 0$, then x^* is a maxmizer.
   b. If $f''(x^*) > 0$, then x^* is a minimizer.

Example A.5 (Single variable optimization example)**.** Consider a firm that produces a single good q and sells it at a price of \$10 per unit. The cost of production is given by:

$$C(q) = 2q + 5 + 0.1q^2$$

At what level of output would profits be maximized?

Solution. The profit of a firm is revenue minus cost:

$$\pi(q) = R(q) - C(q) = 5q - 2q - 5 - 0.1q^2 = 2q - 5 - 0.1q^2$$

Hence, we want to solve the following problem:

$$\max_q \pi(q)$$

The first order condition is given by:

$$\pi'(q) = 0 \Rightarrow 2 - 0.2q = 0 \rightarrow q^* = 10$$

The second order condition is given by:

$$\pi''(q) = -0.2 < 0$$

Hence, $q^* = 10$ maximizes the profits. The maximum level of profits is given by $\pi(q^*) = 2 \times 10 - 5 = 0.1 \times 10 = 5$.

Note that the above process can be easily applied to multivariable functions. In that case there will be one first order condition for every control variable.

Example A.6 (Multi-variable optimization example)**.** Consider a two-variable function:

$$f(x_1, x_2) = 2x_1x_2 + \frac{100}{x_1} - 4x_2^2$$

Solve the following minimization problem:

$$\min_{x_1,x_2} f(x_1, x_2)$$

Solution. Now we have two first order conditions:

$$f_{x_1}(x_1, x_2) = 0 \Rightarrow 2x_2 - \frac{100}{x_1^2} = 0$$

$$f_{x_2}(x_1, x_2) = 0 \Rightarrow 2x_1 - 8x_2 = 0$$

So we have two equations in two unknowns. You can show that $x_1* = 5.84$ and $x_2^* = 1.46$. The minimum of this function is given by $f(x_1^*, x_2^*) = 2 \times 5.84 * 1.46 + \frac{100}{5.84} - 4 \times 1.46^2 = 25.65$.

Problems

Exercise A.1. Compute the derivative of the following functions.

a. $f(x) = 2x^2$

b. $f(x) = 2x^2 + ln(x)$

c. $f(x) = e^{ax}$

d. $f(x) = (2x + x^2)^3$

e. $f(x) = ln(5x + x^2)$

f. $f(x) = \dfrac{x + ln(x)}{x^3}$

Exercise A.2. Compute the second derivative of each function given in Exercise 2.1.

Exercise A.3. Compute the partial derivative for each variable for the following functions:

a. $f(x_1, x_2, x_3) = 4x_1^3x_2 - e^{x_3}x_1 + 3x_2$

b. $f(x_1, x_2) = \dfrac{2x_1 + 3x_2}{4x_1^3 - 7x_1x_2}$

c. $f(x, y) = ln(y^2) - ln(x) + 2ln\left(\dfrac{x}{y}\right)$

d. $f(x, y) = 2x^{0.4}y^{0.8} + 2x$

Exercise A.4. Solve the following optimization problems. In each case compute the maximizer(s) (or minimizer(s)) for the function as well as the optimum value of the function.

a. $\max\limits_{x} f(x) = 3ln(x) - 0.5x + 4$

b. $\min\limits_{x,y} f(x, y) = 2xy + \dfrac{2000}{x} + \dfrac{2000}{y}$

c. $\max\limits_{x} f(x) = ax^{0.5} - bx + 4$

Appendix B

Review of Probability and Statistics

Given that all students must have taken a course in statistics before enrolling for this class, it is assumed that everyone in the class is comfortable with concepts such probability, expected value, measures of central tendency, hypothesis testing etc. In this chapter, I will provide a brief review of some concepts that are most pertinent for Econometrics. I strongly encourage that you read your lecture notes for Statistics if you find it difficult to follow the material presented in this chapter.

B.1 Probability

We begin with a brief review of probability thoery. To define probability we first need to develop an understanding of what we mean by *experiment*, *sample space*, and *event* in statistics.

Definition B.1 (Experiment)**.** An experiment is a process with an uncertain observable outcome. e.g. Toss of a coin can have two possible outcomes, heads or tails.

Definition B.2 (Sample Space)**.** The sample space is the set of all possible outcomes of an experiment. I will denote it by S. If we toss a coin then $S = \{Heads, Tails\}$.

Definition B.3 (Event)**.** An event is a subset of the sample space. I will denote it by E. If we toss a coin and Heads shows up then $E = Heads$.

Now, we can define probability, which is a function that assigns a numerical value to the chance of an event occuring among all possible events in the sample space.

Definition B.4 (Probability)**.** A function P is called a probability function if:

1. For any given event, E, $0 \leq P(E) <\leq 1$.
2. Suppose there are N possible events in S, i.e., $S = \{E_1, E_2, E_3, .., E_N\}$. Then,

$$P(E_1) + P(E_2) + P(E_3) + ... + P(E_N) = 1$$

3. Consider an event E. Then,

$$P(\neg E) = 1 - P(E)$$

3. If we have two disjoint events A and B, then:
   a. $P(A \cup B) = P(A) + P(B)$
   b. $P(A \cap B) = 0$

4. If we have two non-disjoint events A and B, then:
   a. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   b. $P(A \cap B) = P(A) \times P(B|A)$
5. If we have two independent events A and B, then:
$$P(A \cap B) = P(A) \times P(B)$$
6. Bayes rule:
$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

where $P(B) = P(B|A) \times P(A) + P(B|\neg A) \times P(\neg A)$

One such probability function is:
$$P(E) = \frac{\text{Number of outcomes in E}}{\text{Number of outcomes in S}} \tag{B.1}$$

Example B.1. Consider a fair six-sided dice. The probability of obtaining an odd number if this dice is rolled once is given by 0.5. To see this, note that the event here is obtaining an odd number when a dice is rolled. Hence, $E = \{1, 3, 5\}$. Also, $S = \{1, 2, 3, 4, 5, 6\}$. Using this, we get:
$$P(E) = \frac{3}{6} = 0.5 \tag{B.2}$$

B.2 Random Variable

One of the most important applications of statistics is to resolve the randomness that is inherent in most economic choices. For example, the outcome of your college major is a random variable with many possible values. Most economic variables can be thought of as **random variables** that have many possible values which are unknown until they are realized. We will begin by formally defining a random variable.

Definition B.5 (Random Variable)**.** A random variable is a numerical representation of outcomes of an experiment. For example, in the example of a toss of a coin, suppose you win $10 if heads shows and you lose $5 if tails shows. In this case, tossing the coin was the experiment, and winnings from this game is the random variable with two possible values: $10 and -$5.

There are two types of random variables.

1. Discrete random variable: takes finite number of values. e.g. GPA points earned in Econ 385.
2. Continuous random variable: can take any value on the number line. e.g. GDP in the last quarter of 2019.

B.3 Probability distribution

By definition a random variable can take many *possible values*. In statistics a function that provides the probabilties of different realizations of a random varuable is called its **probability distribution**.

B.3.1 Probability distribution of a discrete random variable

For a discrete random variable the probability distribution is simply the list of all possible values this variable can and their corresponding probabilities. Let X be a discrete random variable with n possible values give by $\{x_1, x_2, x_3, .., x_n\}$. Let p_i denotes that probability that $X = x_i$. Then, the probability distribution function of this random variable is given by:

X	p(X)
x_1	p_1
x_2	p_2
x_3	p_3
$\vdots$	$\vdots$
x_n	p_n

Example B.2 (Grade Distribution)**.** A typical grade distribution is an example of a discrete random variable. Consider the following grade distribution:

GPA	Percent of Students
0	10%
1	20%
2	40%
3	20%
4	10%

Note that every GPA point corresponds to a letter grade. From the perspective of the student, X is the random variable that is his letter grade, and the above distribution gives the probability of obtaining a particular letter grade. We can plot this simple probability distribution as follows:

We can use the probability distribution of a discrete random variable in two different ways.

1. We can compute the probability of the random variable taking an exact value. This is known as the **probability mass function (p.m.f)** and is denoted by $f(x)$:

$$f(x) = P(X = x)$$

For example, the probability of obtaining a letter grade of C or $P(X = 2)$ is 0.4 or 40%.

2. We can also infer the probability that a discrete random variable will be less than or equal to a certain value by cumulatively adding the probabilities. Formally, we can compute the **cumulative probability distribution (c.d.f)** which is denoted by $F(x)$:

$$F(x) = P(X \leq X)$$

Going back to our grade distribution example, we can add the column of cumulative probabilities to obtain the $c.d.f$:

Grade	Percent of Students	$F(x)$
0	10%	10%
1	20%	30%
2	40%	70%
3	20%	90%
4	10%	100%

So for example, we can infer that the probability of obtaining the letter grade of C or lower i.e, $P(X \leq 2)$ is 0.7 or 70% which is obtained by adding the probabilities of obtaining letter grades of C, D, and F, respectively.

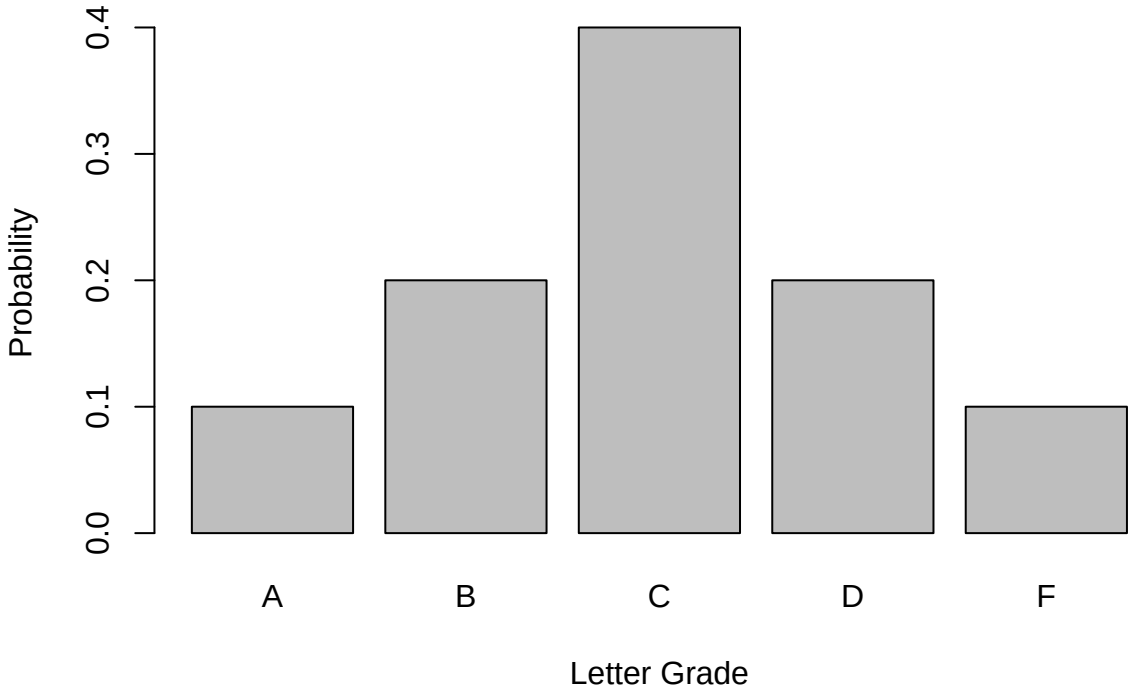


Figure B.1: Probability Distribution of Letter Grades

Example B.3 (Bernoulli Random Variable)**.** When a random variable is binary then we call it a **Bernoulli** random variable and its probability distrubition is called **Bernoulli** distribution. Consider a random variable that can only take two values, say, 0 or 1. It is common to think of these two values as coding a set criterion with 1 typically assigned if the criterion is met and 0 is assigned for failing to meet the criterion. For example, X could be whether you will get a job right after graduation. If you do then $X = 1$ and if you do not then $X = 0$. Let p denotes the probability that you will get a job. Then, the $p.m.f.$ of the Bernoulli distribution is given by:

$$f(x) = \begin{cases} p & if\ X = 1 \\ 1 - p & if\ X = 0 \end{cases}$$

The $c.d.f$ of the Bernoulli distribution is given by:

$$F(x) = \begin{cases} 0 & if \quad X < 0 \\ 1 - p & if \quad 0 \leq X < 1 \\ p & if \quad X \geq 1 \end{cases}$$

B.3.2 Probability distribution of a continuous random variable

In economics a large majority of variables of interest in theory are continous random variables. For example, the change in the price of Apple stock between two time periods is the return on Apple stock. If you are a trader in the NYSE then the stock return on Apple is a continuous random variable that can take any value on an interval. In such a case we cannot obtain the probability of the random variable taking an exact value. But we can only compute the probability that this random variable will fall in a given interval. So at best we can determine the probability that GDP growth for the US next quarter will be between say 1/% and 2%. This probability is obtained by computing the area under the **probability density function (p.d.f)**. Let X denote a continuous random variable and $f(x)$ denotes the p.d.f. Then,

1. The probability that X takes value over the interval $\{a, b\}$ is given by:

$$P(a \leq X \leq b) = \int_a^b f(x)\ dx$$

2. The c.d.f (the probability that $X \leq x$) is given by:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\ dx$$

Below I plot the empirical c.d.f for Apple's stock return. Let X denotes this stock return. From Fig 3.2 we can infer that $P(X \leq 0) = 0.47$ and $P(X \leq 3) = 0.95$.

Figure 3.3 below presents the $p.d.f$ of the daily stock return that corresponds to the $c.d.f$ plotted in Figure 3.2. Using this we can work the probability of stock returns falling in any given interval. For instance, the probability that Apple stock return will fall between 0 and 3% is the area under the p.d.f. between these two values. Figure 3.3 higlights this area and we can see that this probability is equal to 0.47.

B.4 Moments of a probability distribution function

The information contained in a probability distribution can be meaningfully summarized into measures that are called **moments** of that distribution. There are three moments we often use in economics:

1. Center of the distribution: this is first moment of a given probability distribution and it gives us the most likely value of the random variable. It can be measured by mean, median or mode. We will use mean as a measure of the center of the distribution.

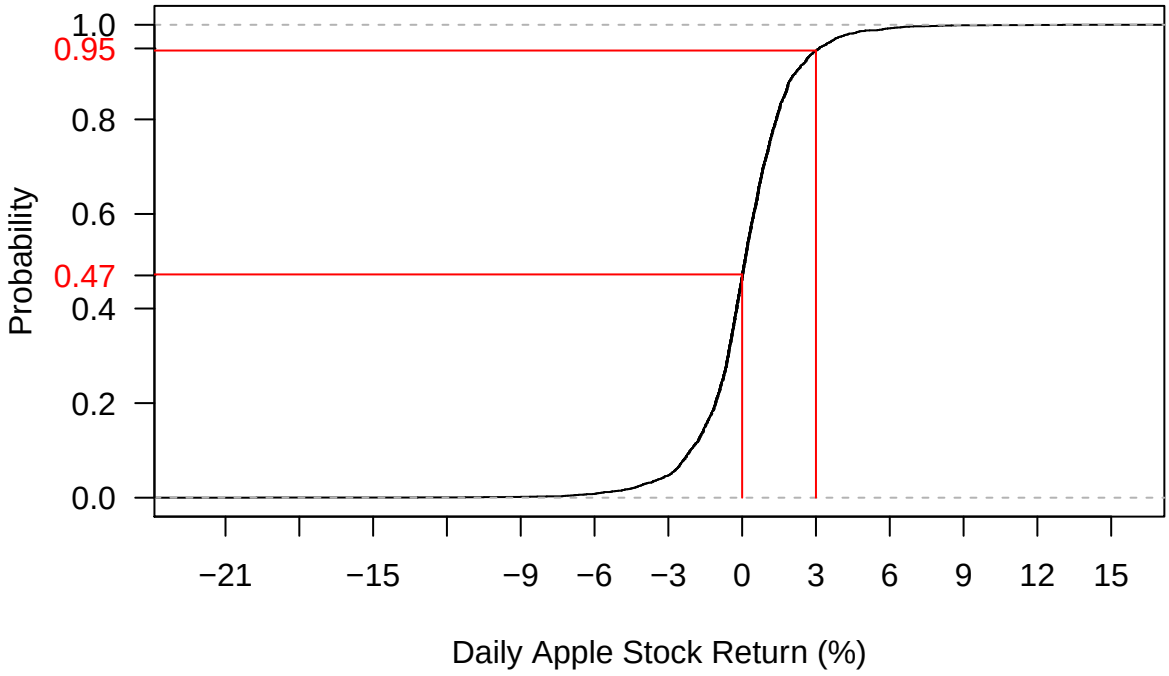


Figure B.2: Empirical c.d.f of daily Apple Stock Return (2007-2019)

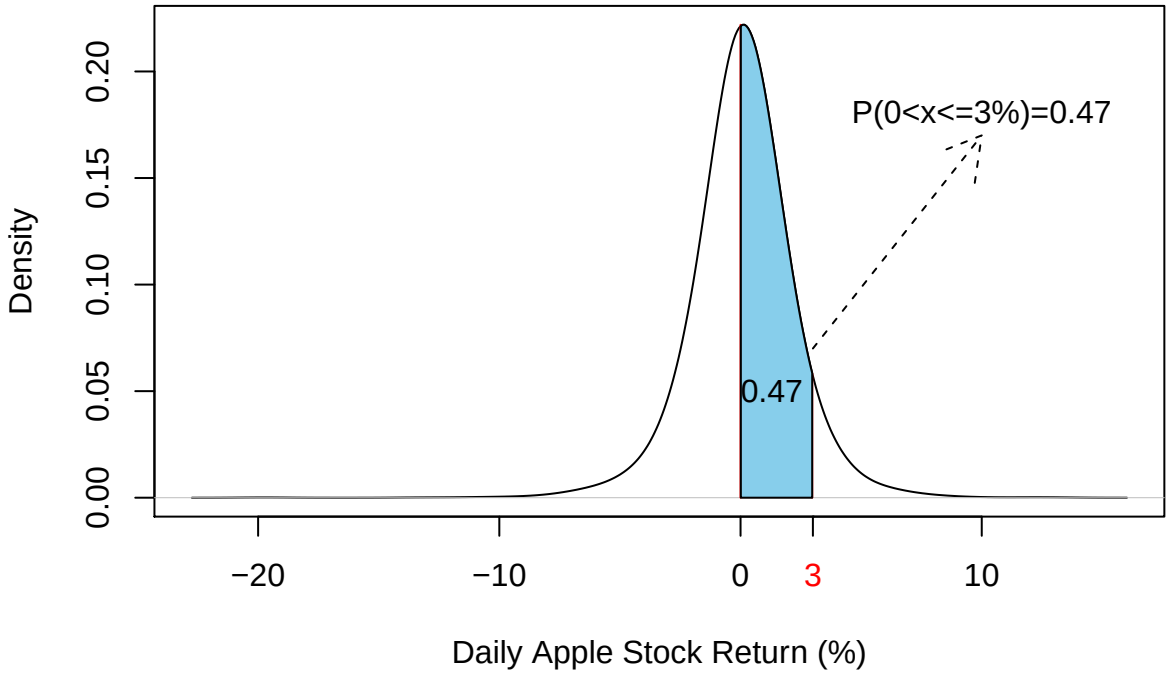


Figure B.3: Empirical p.d.f of daily Apple Stock Return (2007-2019)

2. Width of the distribution: this is the second moment and it measures the average distance from mean under a given probability distribution. We will use standard deviation as a measure of the width of the distribution.
3. Shape of the distribution: this feature relates to role played by **tail events** , i.e., events that have very low probability of happening under a given probability distribution. Two relevant measures are Skewness and Kurtosis

B.4.1 First moment of a probability distribution: Expected value

What is the most likely value of a random variable? To answer that we often compute **expected value** of the random variable which gives us the center (or peak) of the underlying probability distribution. We will use **E** to denote expected value. So $E(X)$ is the expected value of a random variable and we will use μ_X to denote the mean or average value of X.

Definition B.6 (Expected Value)**.** Consider a discrete random variable X that can take n possible values and has the following probabilty distrubution:

X	p(X)
x_1	p_1
x_2	p_2
x_3	p_3
$\vdots$	$\vdots$
x_n	p_n

Then, the expected value of X is given by:

$$E(X) = x_1p_1 + x_2p_2 + ... + x_np_n = \sum_{i=1}^{n} y_ip_i$$

Hence, expected value is a probability-weighted average of all possible values of a random variable.

Example B.4. Suppose you toss a fair coin and receive $10 if tails shows and receive 0 if heads shows. What is the expected value of the winnings from a single toss of this coin?

Solution. Let X denotes winnings from this game. It can take a value of $10 with a probability of 0.5 and 0 with a probability of half. So the expected value of X is:

$$E(X) = x_1p_1 + x_2p_2 = 10 \times 0.5 + 0 \times 0.5 = \$5$$

Example B.5. Suppose you can invest $10,000 in a mutual fund after 1 year can earn a return of 10% with a probability of 0.1 or a return of 2% with a probability of 0.5 or a loss of 5% with a probability of 0.4. What is the expected return of investing $10,000 in this mutual fund?

Solution. Let X denotes expected return in dollars. It can take 3 possible values: $1000 with a probability of 0.1, $200 with a probability of 0.5, and -$400 with a probability of 0.4 The expected value is given by:

$$E(X) = 1000 \times 0.1 + 200 \times 0.5 - 400 \times 0.4 = \$40$$

As mentioned earlier, the first moment of the probability distribution (i.e., the expected value) gives us the most likely value of the random variable. How useful is this knowledge will depend on how far any realiaztion of the random variable can be from its expected value. The average distance from the average measures the width of the distribution. Wider the distribution, less useful is the knowledge of the expected value.

B.4.2 Second moment of the distribution.

To determine the width or *dispersion* of a probability distribution we use **variance** or **standard deviation**. The variance is the expected value of the squared deviation of each realization of the random variable from its average. We will denote the variance by $Var(X)$ or σ_X^2:

$$Var(x) = \sigma_X^2 = E[(X - \mu_x)^2] = (x_1 - \mu_X)^2 \times p_1 + (x_2 - \mu_X)^2 \times p_2 + ... + (x_n - \mu_X)^2 \times p_n = \sum_{i=1}^{n} (x_i - \mu_x)^2 p_i$$

The standard deviation is simply the square root of the variance and is in the same units as the random variable. This allows easy comparison of the width and the center of the distribution. We will denote standard deviation by σ_X.

Example B.6. Using the mutual fund example, the variance will measure **riskiness** of the investment. It is given by:

$$Var(X) = (1000 - 40)^2 \times 0.1 + (200 - 40)^2 \times 0.5 - (400 - 40)^2 \times 0.4 = 53120$$

Because variance is in square units and hence hard to interpret, we can easily compute the standard deviation as the square root of the variance:

$$\sigma_X = \sqrt{53120} = \$230.47$$

Hence, even though the average return from this investment is \$40, you can be \$230 above or below this average.

How much can we say about a random variable if we only know its mean and the standard deviation? That depends on the type of distribution the random variable follows. One of the most commonly used distribution in statistic is the **Normal Distribution** or the **Gaussian Distribution**. A random variable that follows normal distribution has a bell-shaped probability distribution with a given mean and standard deviation. One of the most useful features of such a distribution is that knowledge of the first two moments alone is sufficient to characterize the entire probability distribution. Figure 3.4 below shows a normal distribution with a mean of 5 and a standard deviation of 2.

Key features of the normal distribution that are very useful for us:

a. 95% of the values fall within 1.96 times the standard deviation of the mean:

$$P(\mu_X - 1.96\sigma_X \leq X \leq \mu_X + 1.96\sigma_X) = 0.95$$

b. Tail events (low probability events) on either side of the mean are equally unlikely.

c. Central limit theorem: The distribution of sample means calculated from repeated random sampling from a given population approaches a normal distribution as the sample size approaches ∞.

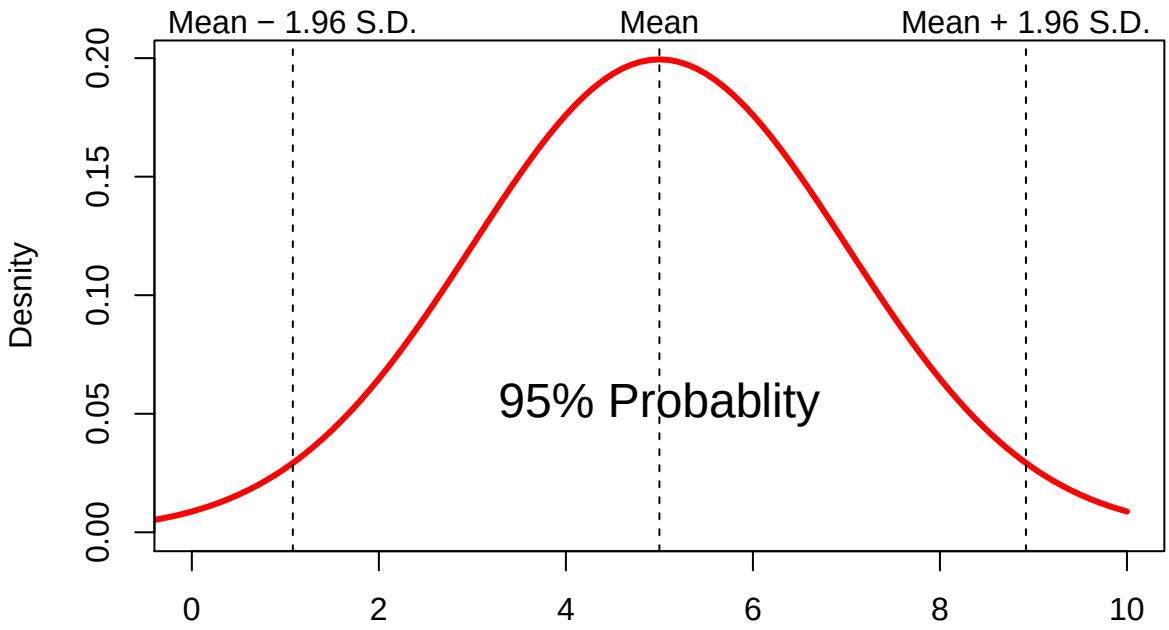


Figure B.4: Normal distribution with mean=5 and s.d.=2

B.4.3 Third and Fourth Moments: Skewness and Kurtosis

In many cases, the distribution of a random variable is not normal and in such cases higher moments provide useful information about the shape of such probability distribution. The shape of the probability distribution plays an important role in many economic and financial applications. There are two measures of shape that are of interest:

1. **Skewness**: this is the third moment of the distribution and it measures how skewed a distribution is. The formula for skewness is given by:

$$Skewness = \frac{E[(X - \mu_X)^3]}{\sigma_X^2}$$

A normal distribution has a skweness of zero. There are two possible types of skewed distributions:

a. A positively skewed distribution will have a long right tail implying lower probability of very large values relative to the mean.

b. A neagtively skewed distribution will have a long left tail implying lower probablity of very small values relative to the mean.

Figure 3.5 shows three probability distributions. For the left-skewed distribution, a longer left tail indicates low probability of obtaining values below the mean. Similarly, for the right-skewed distribution, a longer right tail indicates low probability of obtaining a value above the mean. For a normal distribution, the probability of obtaining a value above the mean is the same as the probability of obtaining a value below the mean.

2. Kurtosis: this is the fourth moment of the distribution that captures the peakedness of the distribution (or thickness of the tail), i.e., how many observations fall on the extreme ends of a given probability distribution. As a result it tells us the role played by extreme values in driving the variance of a random variable. The formula is given by:

$$Kurtosis = \frac{E[(X - \mu_X)^4]}{\sigma_X^4}$$

A normal distribution has a Kurtosis of 3. A value that is above or below 3 will give us excess or deficient Kurtosis. Two possiblities are:

a. Leptokurtic distribution: has a Kurtosis value greater than three. Such a distribution will have fat tails compared to a normal distribution indicating greater area under the tails.

b. Platykurtic distribution: has a Kurtosis value less than 3. Such a distribution will have thin tails compared to a normal distribution.

Fig 3.6 shows three types of distribution based on their Kurtosis. The leptokurtic distribution has a Kurtosis value of greater than 3 and is more **heavy-tailed** or **peaked** than a norma distribution.

B.5 Useful probability distributions

Using the normal distribution we can derive a few useful probability distributions that are utilized in hypothesis testing.

1. Standard Normal Distribution: A random variable that follows normal distribution with a mean of 0 and standard deviation of 1.

2. Chi-square distribution: is obtained by squaring and adding indpendent standard normal distribution. For example, is X and Y are two standard normal random variables, then $Z = X^2 + Y^2$ follows a Chi-square distribution with two degrees of freedom.

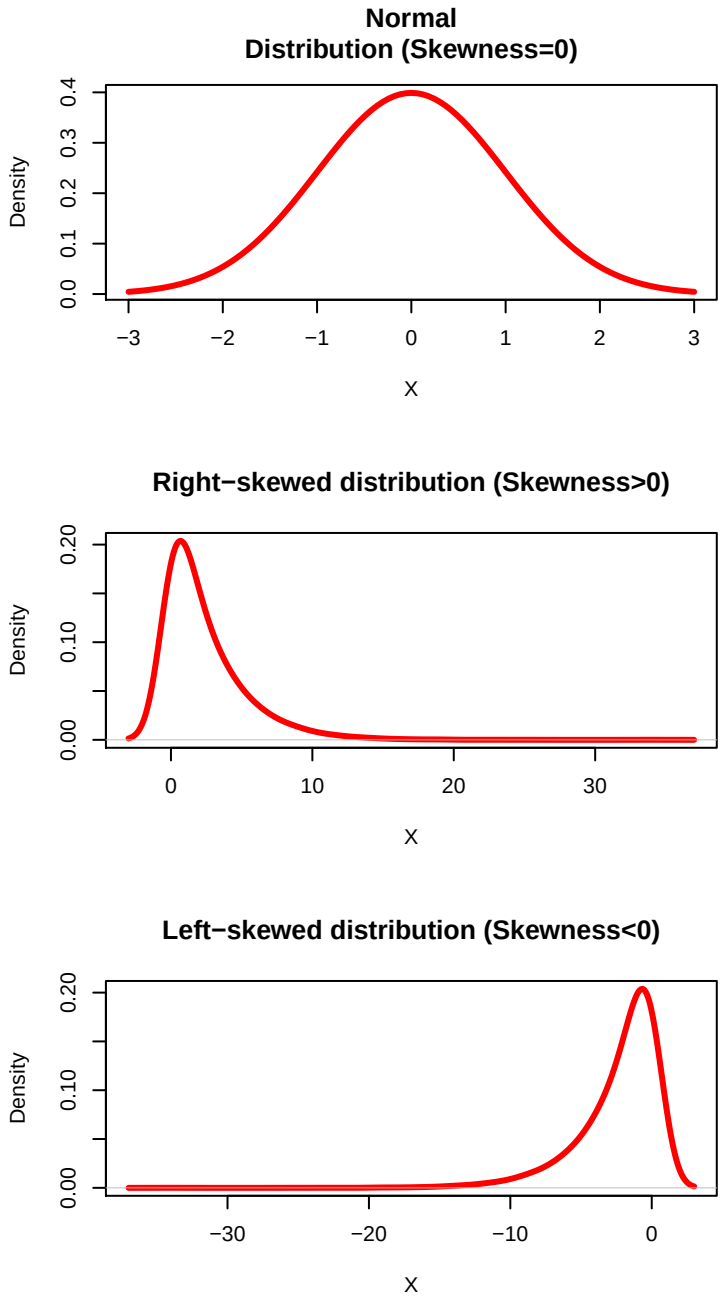


Figure B.5: Skewness of a Probablity distribution

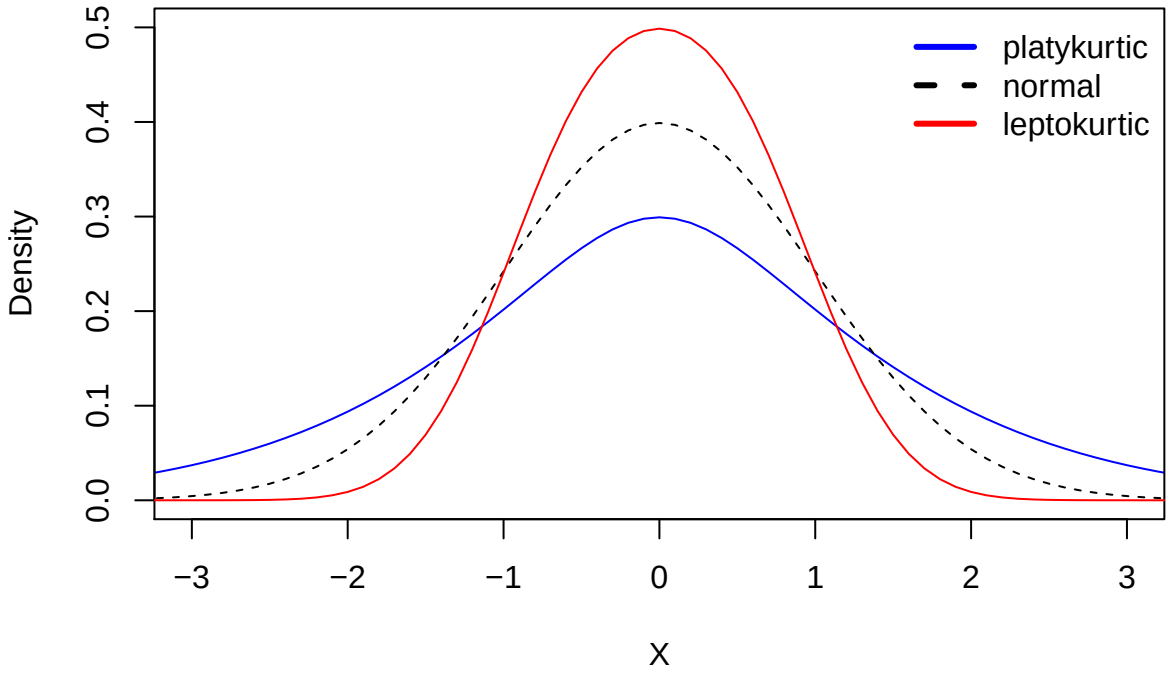
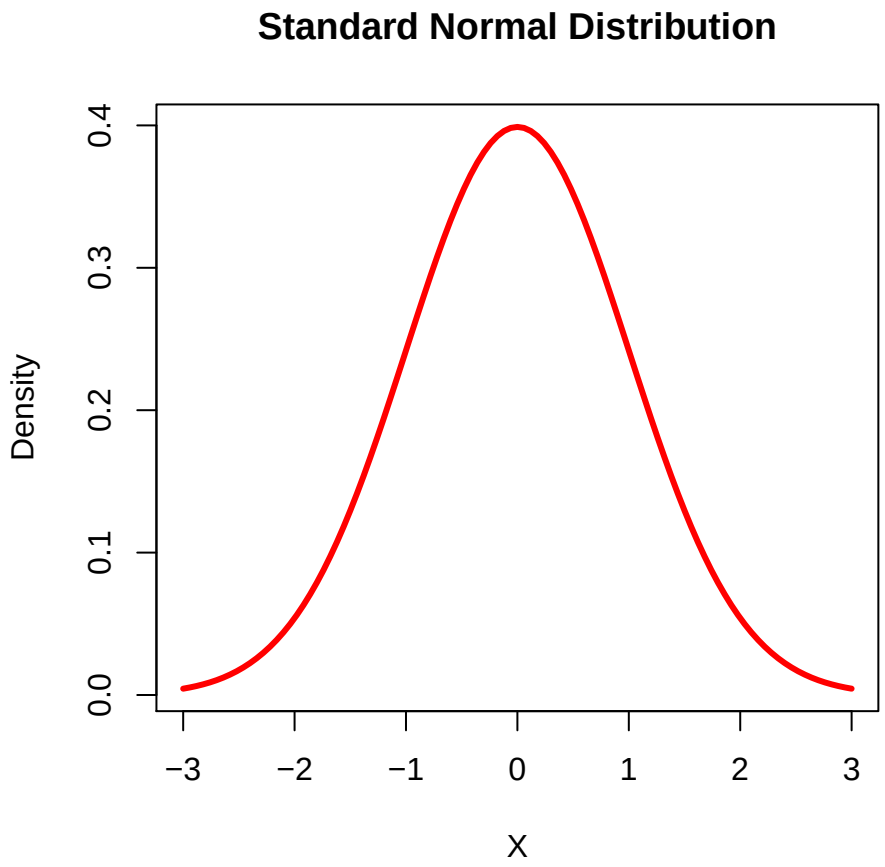
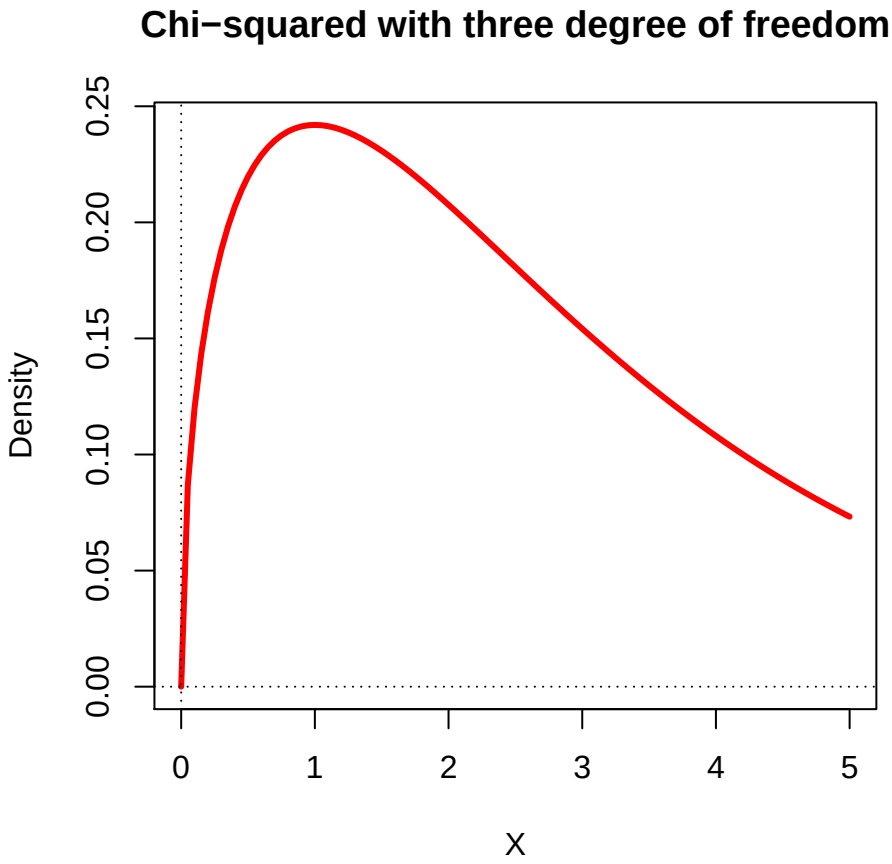


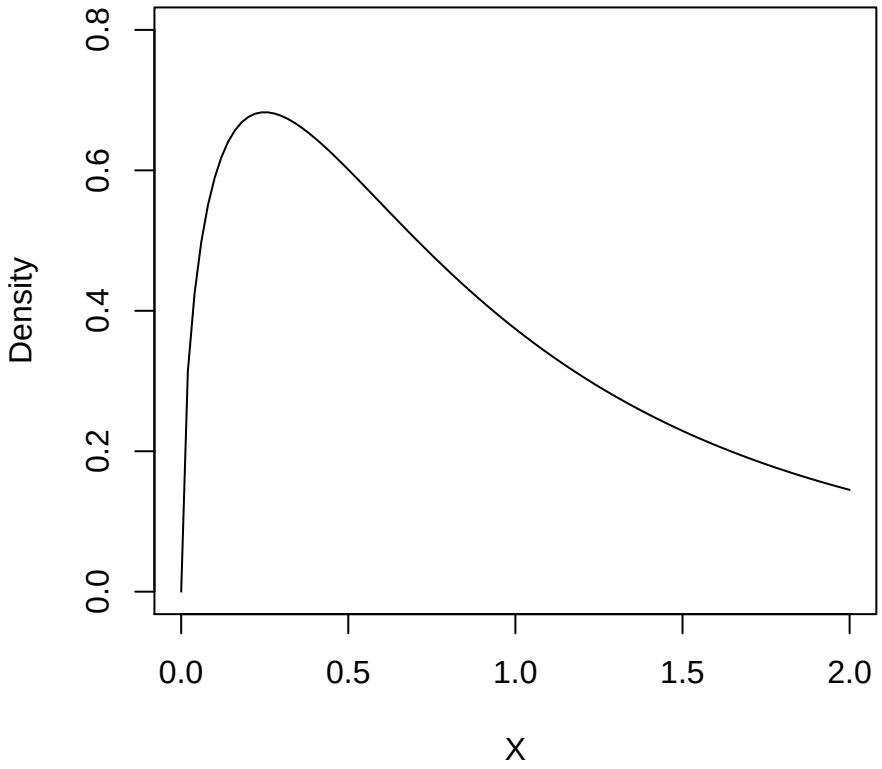
Figure B.6: Kurtosis of a Probablity distribution

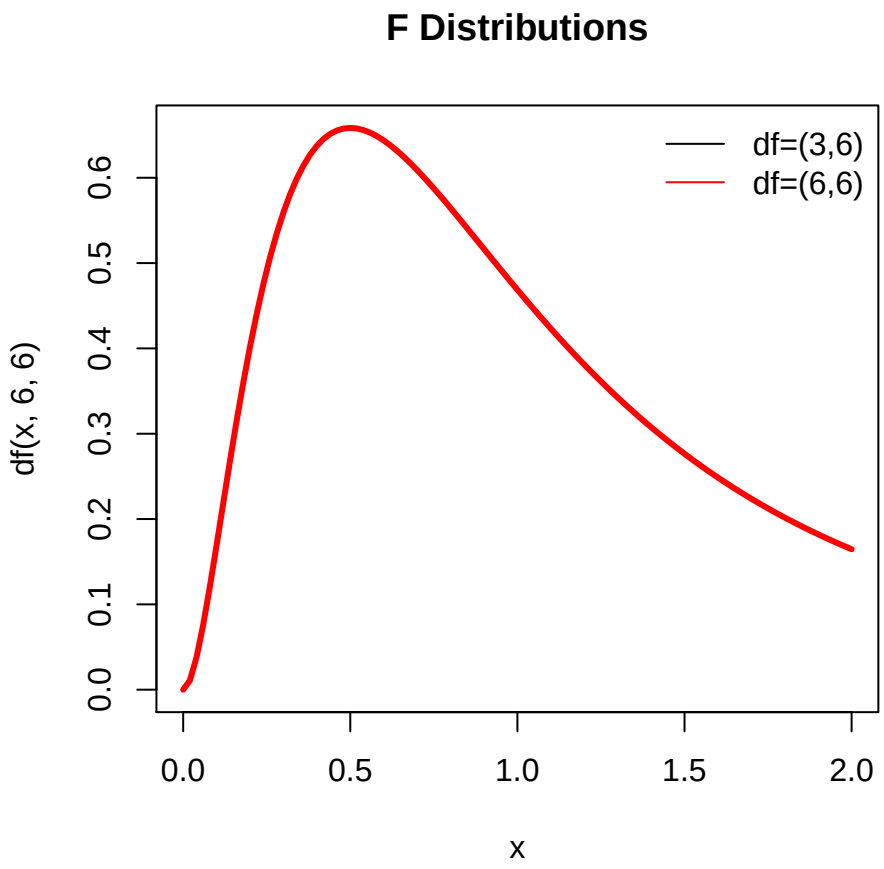
3. F-distribution: is obtained by taking a ratio of two chi-square distribution. For example, if X is Chi-square with v_1 degrees of freedom and Y is a Chi-quare with v_2 degress of freedom, then $Z = \frac{X}{Y}$ follows F-distribution with v_1 and v_2 degrees of freddom.

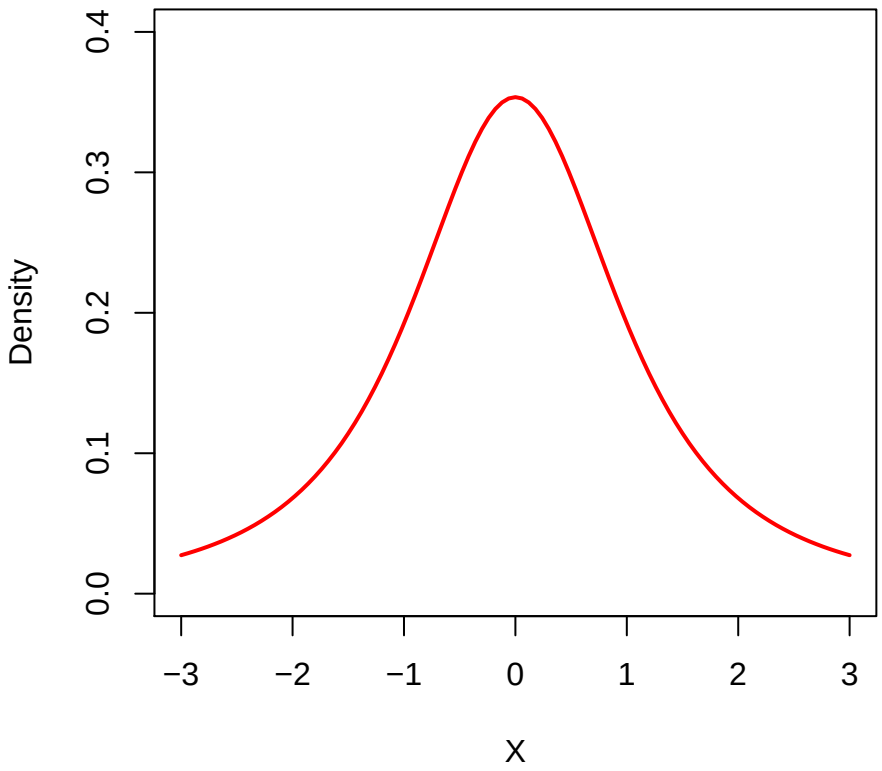
4. t-distribution: Student's t-distribution is obtained by taking a ratio of a standard normal and the square root of a Chi-square random variable. For example, if X is a standard normal and Y is a Chi-square with m degrees of freedom, then $Z = \frac{X}{\sqrt{Y/m}}$ follows t-distribution with m degrees of freedom. t-distribution has fatter tails when compared to normal.

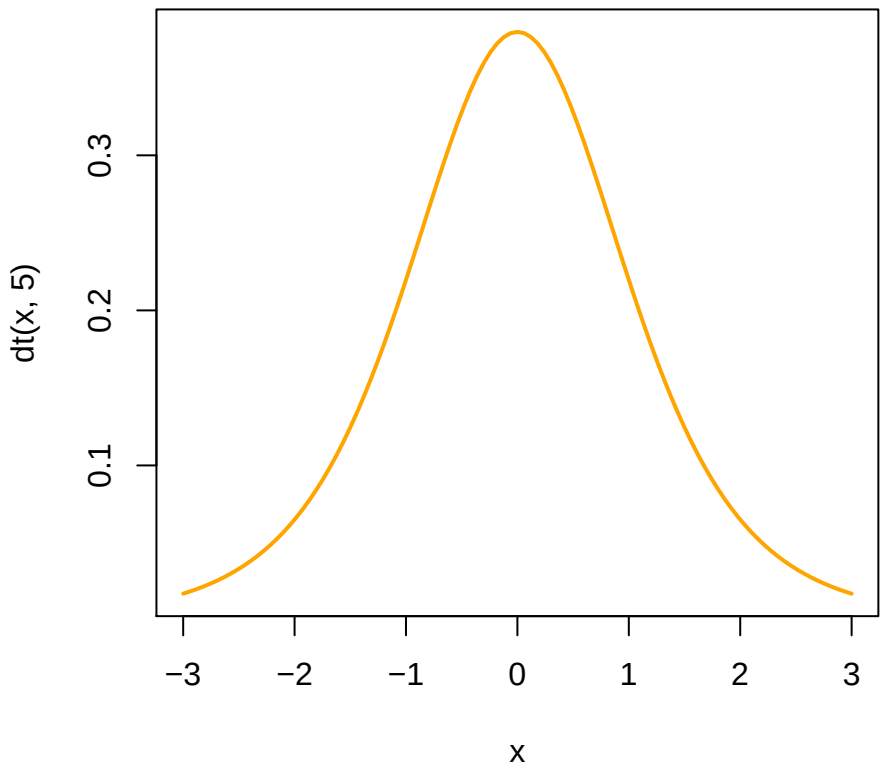


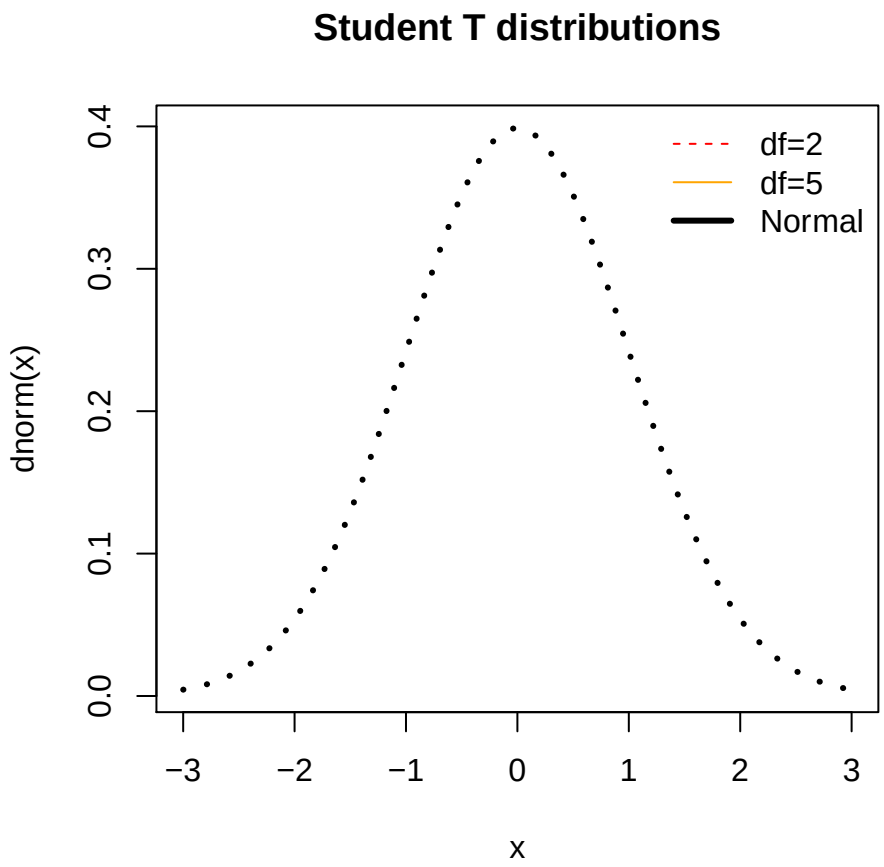

Chi−squared with three degree of freedom
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
1
2
3
4
5
X


Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
X


F Distributions
df=(3,6)
df=(6,6)
df(x, 6, 6)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.5
1.0
1.5
2.0
x


Density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
X


dt(x, 5)
0.1
0.2
0.3
−3
−2
−1
0
1
2
3
x



B.6 Joint Probability Distribution

In economics, often we are interested in the relationship between a pair of variables. For example, how does interest rate affects consumption spending? Or how does education affect wages? In order to statistically answer such questions, we need to understand the meaning of statistical relationship between two or more variables. One way to move forward is to assume that both variables jointly follow some given probability distribution which can be used to infer their relationship with one another.

For simplicity, I will use the discrete random variables case but the concepts covered can be easily extended for the continuous random variables case.

Let X and Y denote two random variables of interest, both from a common probablity distribution denoted by $F(x, y)$. This function gives us the probability that X and Y simultaneously take on certain values:

$$F(x, y) = P(X = x, Y = y)$$

Example B.7. Suppose you are an investment banker and you are considering investment into two assets: a stock listed in NYSE (X) and a cotton futures (Y) listed in Chicago Mercantile Exchange. Suppose X can take three possible values: 2/%, 3/%, or 4/%. Similarly Y can take three possible values given by 6/%,4/%, or 1/%. The value will depend on the state of the economy. Suppose there are three possiblities for the economy next year: boom, expansion, and status quo. The joint probability distribution for X and Y is given by:

State of Economy	X/Y	6	4	1	Total
Recession	2	0.15	0.2	0.1	0.45
Expansion	3	0.1	0.1	0.2	0.4
Status quo	4	0.1	0.05	0	0.15
	Total	0.35	0.35	0.3	1

So in a recession, the probability of obtaining a return of 2/% on the stock and 6/% return on the commodity, i.e, $P(X = 2, Y = 6)$, is 0.15. Using the above joint probabity distribution of X and Y we can compute two related distributions for each random variable:

1. Marginal distribution: For each random variable, we can extract its own probability distribution from the joint probability distribution. This is done by simply adding probabilities of all possible outcomes for a particular value of a given random variable. For example, the marginal distribution for X is given by:

$$P(X = x) = \sum_{i=1}^{n} P(X = x, Y = y_i)$$

Hence, in our example, the marginal distribution of X is given by the last column, called Total in the table. For Y it is the row called Total. We can use the marginal distribution to compute the unconditional expected value of each random variable. For example,

$$E(Y) = 6 \times P(Y = 6) + 4 \times P(Y = 4) + 1 \times P(Y = 1) = 3.8$$

2. Conditional distribution: For each random variable, we can also compute its probability distribution conditional on the other variable taking on a specific value. For exampl,e the conditional distribution of Y given that $X = x$ is given by:

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

From our example, what is the probability of obtaining 4% return on commodity under status quo if the return on the stock is 4%? So here we are interested in finding out:

$$P(Y = 4|X = 4) = \frac{P(X = 4, Y = 4)}{P(X = 4)} = \frac{0.05}{0.15} = 0.33$$

To see this, note that from the table that P(X=4, Y=4) under status quo is given by 0.05. Also, using the definition of marginal distribution, we know that P(X=4)=0.15.

The conditional distribution of a random variable is a first step toward understanding the statistical relationship between two or more random variables. Just like the probability distribution of a random variable has a mean and a variance, the conditional distribution can similarly be characterized by conditional mean and conditional variance:

1. Conditional expected value ($E(Y|X)$): Using the conditional distribution we can now compute the expected value of a random variable, given the value of another random variable. This is denoted by $E(Y|X)$ and can be computed as follows:

$$E(Y|X) = y_1 \times P(Y = y_1|X = x) + y_2 \times P(Y = y_2|X = x) + ... + y_n \times P(Y = y_n|X = x)$$

As we can see, this expected value will be a function of X. Depending on the realization of X our expectation of Y would change. In economics, we can imagine many such examples. For example, given our education

level our expected wage will change. Similarly, given expenditure on advertising, expected sales will change. Hence, conditional expected value goes a long way in establishing statistical relationship between economic variables.

Going back to our example, let us compute the expected return on the commodity Y conditional on the information that the return on X is 3%:

$$E(Y|X) = 6 \times P(Y = 6|X = 3) + 4 \times P(Y = 4|X = 3) + 1 \times P(Y = 1|X = 3)$$

Here, $P(Y = 6|X = 3) = \frac{0.1}{0.4} = 0.25$, $P(Y = 4|X = 3) = \frac{0.1}{0.4} = 0.25$ and $P(Y = 1|X = 3) = \frac{0.1}{0.2} = 0.5$. Hence, $E(Y|X = 3) = 3\%$. Contrast this to the uncondtional expected value of Y of 3.8% we computed earlier.

2. Conditional variance ($Var(Y|X)$): Now even the variance of a random variable can be affected by another random variable. Here, we are interested in deviations of the random variable from its conditional mean:

$$Var(Y|X) = (y_1 - E(Y|X))^2 \times P(Y = y_1|X = x) + (y_2 - E(Y|X))^2 \times P(Y = y_2|X = x) + ... \\ +(y_n - E(Y|X))^2 \times P(Y = y_n|X = x) \tag{B.3}$$

B.7 Measures of statistical association

We can now define two measures of statistical relationship. The first one is called **Covariance** and the second is **Correlation**.

1. Covariance is a measure of association that captures how deviations from mean of one random variable are related to deviations of another random variable to its respective mean. For example, if your hours of study are above average, then what is your test score relative to average? Formally,

$$Cov(X, Y) = E(Y - \mu_Y)(Y - \mu_X)$$

If the above number is positive, then there is a positive relationship between X and Y. That is, when X is above its mean then Y is also above its mean. If the number is negative then there is a negative relationship between X and Y.

Note that because X and Y are often in different units of measurement, the number we obtain for covariance has no meaning or implication for the strength of the relationship between two variables.

2. Correlation: is the value of covariance that is standardized by dividing this number by standard deviations of each random variable:

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_X \times \sigma_Y}$$

This number is unit free and falls between -1 and 1. The sign of the correlation tell us about the direction of the relationship whereas the value of the correlation gives information about the strength of the relationship. A higher absolute value indicates stronger statistical relationship between two variables.

B.7.1 Rules of expectation and variances

Here are some useful rules that are useful for our purpose:

1. $E(\beta) = \beta$ and $Var(\beta) = 0$ where β denotes a constant.

2. $E(\beta X) = \beta E(X)$ and $Var(\beta X) = \beta^2 Var(X)$ where β denotes a constant.
3. Consider two random variables X and Y, and let a and b denotes two constants. Then,
   3.1. $E(aX + bY) = aE(X) + bE(Y)$
   3.2. $E(aX - bY) = aE(X) - bE(Y)$
   3.3. $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCor(X,Y)\sqrt{Var(X)}\sqrt{Var(Y)}$
   3.4. $Var(aX - bY) = a^2 Var(X) + b^2 Var(Y) - 2abCor(X,Y)\sqrt{Var(X)}\sqrt{Var(Y)}$

B.8 Sampling and Estimation

An important distinction in statistics is between the population of interest and a sample of this population that we usually work with. Due to feasibility of data collection and cost both in terms of time and money, most real world analysis is based on a sample that is a subset of the population of interes. For example, to study how business major affects starting salary, the relevant population is all business majors from a graduating class in the U.S. in a given year. In practice however, we will most likely use a sample of this population, for example all business majors from JMU. How useful an analysis based on a sample is depends on how representative the chosen sample is of the entire population.

For our purpose, lack of data on population means that the true probablity distribution of a random variable is unknown and hence the true values of mean, variance, covarinace etc are also unknown to us. Statistics provides a way of using samples to **estimate** relevant moments of the probability distribution. The approach we take is as follows:

1. Consider the unknown moments of the true probability distribution as ** population parameters** that we would like to estimate.
2. Draw a representative sample from the population. In simple random sampling we draw n obeservations at random so that each member of the population is equally likely to be included in the sample. We can also use other complex sampling schemes where certain groups of population are more likely to be selected in the sample than others. Two examples:
   a. Suppose we are interested in finding out starting salary of CoB majors at JMU. The population will be every graduating student for a given year. However, we may work with a sample of students, where we draw randomly from every major ensuring that all graduating students have equal probability of selection.
   b. Suppose we are interested in finding out usage of food stamps in Harrisonburg area. The population of interest will be all residents of Harrisonburg who use food stample. However, we may work with a sample where a certain demographic group is more likely to be part of the sample (and hence is *oversampled*).
3. Use the sample to compute sample estimates for each population parameter of interest. For example for expected value we can use sample mean as an estimator, for variance we can use sample variance as an estimator and so on. There are following key differences between population parameters and their sample estimates:
   a. Population parameters are true but unknown values that we are interested in measuring. In contrast, sample estimates can be computed using our sample data.
   b. Population parameters are fixed whereas sample estimates change as we change our sample. For example, if we compute mean starting salary of business majors from JMU we get one number. If use data from UVA we get another number for mean starting salary.
   c. Because different samples give us different sample estimates for the same population parameter, we need to ensure that our sample estimator from one sample data is reliable.

4. Sampling distribution: Hypothetically, we can draw many samples from the same population and compute sample estimate for each sample. This will give us a distribution of for the sample estimate which will have its own mean and variance. We can use this sampling distribution to:
    a. Establish reliablity of the sample estimator. Specifically any sample estimator should be unbiased and efficient. More on this in the next section.
    b. Statsitically test hypotheses about the true population parameter ### Unbiasedness and efficiency

Let θ denote a population parameter of interest. For example, it can be the mean of the random variable of interest. Let $\widehat{\theta}$ denotes a sample estimator of θ that can be computed using sample data. Then,

1. $\widehat{\theta}$ is an **unbiased** estimator of θ if:

$$E(\widehat{\theta}) = \theta$$

The idea here is that if we repeatedly draw a sample from the same population and compute $\widehat{\theta}$ for each such sample, the average of these estimators must be equal to the true population parameter for unbiasedness. In otherwords, the center of the sampling distribution is at the true population parameter value.

We can now define **bias** of an estimator as follows:

$$Bias(\widehat{\theta}) = E(\widehat{\theta}) - \theta$$

For an unbiased estimator, $Bias(\widehat{\theta}) = 0$. If $Bias(\widehat{\theta}) > 0$ then we have an over-estimate and if $Bias(\widehat{\theta}) < 0$ then we have an under-estimate.

2. Efficiency: Unbiasedness ensure that the average of sample estimator is equal to the true population parameter. But if the standard deviation of the sample estimator is too high, then knowing that the average is close to the true value is not very useful. In statistics, we call such an estimator unbiased but **imprecise or inefficient**. To be efficient the standard deviation (or variance) of the sample estimator should be as small as possible. Between two unbiased estimators, a more efficient estimator will have a lower variance.

Example B.8. Suppose we have a random sample with n observations: $\{x_1, x_2, ..., x_n\}$ drawn from a population with a mean of μ_x. Sample mean is defined as:

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

The expected value of the sample mean is given by:

$$E(\overline{X}) = E\left(\frac{\sum_{i=1}^{N} x_i}{N}\right)$$

Using properties of the expected value, we get:

$$E(\overline{X}) = \frac{E(x_1) + E(x_2) + ... + E(x_N)}{N}$$

Note that because this is a random sample from the same population with a mean of mu_x, we get $E(x_1) = E(x_2) = .. = E(x_n) = \mu_x$. Hence,

$$E(\overline{X}) = \frac{\overbrace{\mu_x + \mu_x + ... + \mu_x}^{\text{N terms}}}{N} = \mu_x$$

As a result the sample mean is an unbiased estimator of the population mean. However, there are many other possible unbiased estimators of the population mean. We can show that among all other unbiased estimator of the population mean, sample mean has the lowest variance and hence is most efficient estimator as well.

Definition B.7 (Best Unbiased Estimator (BUE))**.** Let θ denote a population parameter of interest. Then, an sample estimator denoted by $\hat{\theta}$ is the best unbiased estimator of θ if the following two conditions are satisfied:

1. $\hat{theta}$ is an unbiased estimator, i.e., $E(\hat{\theta}) = \theta$. In this case the sampling distribution is centered at the true value of the parameter.
2. $\hat{\theta}$ is an efficient estimator, i.e., $Var(\hat{\theta}) < Var(\hat{\theta_A})$ for any other unbiased estimator denoted by $\hat{\theta_A}$. In this case the width of the sampling distribution around the mean is smallest possible.

B.9 Hypothesis testing

An important part of any statistical analysis is testing various hypotheses about population parameters of interest. This is known as *statistical inference* and here we use the sampling distribution of the estimator to formally test whether the corresponding population of interest takes a certain value or not. This is important because even with an best unbiased estimator we do not know the true value of the population parameter of interest. In this section we will look at two types of hypotheses testing procedures that are most relevant for Econometrics. The procedure for any statistical test more or less consists of the following steps:

1. Formulate a hypothesis of interest. This typically manifest as a restriction on the value of a population parameter (or a combination of multiple parameters). The goal is to test whether there is support for this restriction in our sample or not. There are two types of hypotheses that we must formulate:

   1.1. Null Hypothesis (H_0): A null hypothesis is the statement about the population parameter we assume to be true until we find evidence otherwise. For example, we can test whether the population mean of starting salary for CoB majors is \$60,000. Formally,

   $$H_0 : \mu_X = 60,000$$

   Note that the null hypothesis statement is an equality condition.

   1.2. Alternative Hypothesis (H_A): This is the logical counterpart of the null hypothesis and here we specify. There are two types of alternative hypothesis we can specify:

   a. Two-sided alternative: Here, the alternative hypothesis statement allows for both sides of the inequality. Going back to our example of starting salary, a two-sided alternative will be:

   $$H_A : \mu_X \neq 60,000$$

   b. One-sided alternative: Here, we either use a greater or less than sign for the alternative hypothesis. So for example, we can speficy the following one-sided alternative:

   $$H_A : \mu_X > 60,000$$

2. Compute the relevant test statistic that is a function of the sample data. The formula for the test statistic is a function of the sample estimator and the value of the population parameter(s) we assumed in the null hypothesis.
3. The test statistic is assumed to follow a certain probability distribution under the assumption that the null hypothesis is correct. The tails of this distribution summarizes values of the test statistic that are less likely to realize. Such a value of the test statistic provides us a threshold level, called the **critical value**, beyond which the test statistic values are less likely to realize if our hypothesis is true. The decision rule for rejecting or not rejecting the null hypothesis is based on the comparison between the computed test statistic and the associated critical value.

Note that there is always a measure of uncertainty in any hypothesis testing: we may end up making a wrong decision. There are two types of errors we can make here:

1. **Type I** error: here we reject H_0 when it is true. The probability of this type of error is denoted by α and is called the **level of significance** of a test.
2. **Type II** error: here we do reject H_0 when it is false. The probability of this type of error is related to the **power** of a test.

Ideally we would like to minimize the probability of both types of errors but we cannot do that because reducing one error comes at the cost of increasing the other. As a result, we first specificy an **acceptable** level of significance (type one error probability) and then try to minimize the probability of type two error (or maximize the power of the test). It is common to assume a level of signficance of 5% or $\alpha = 0.05$. So here we are willing to tolerate a 5% chance of falsely rejecting the null hypothesis.

Once we have fixed the level of significance, we can use the distribution table of the test-statistic to obtain the corresponding critical value(s).

B.9.1 Testing a restriction on a single population parameter

Here our goal is to develop tests for testing statements about a single population parameter of interest. So for example, we can either test a statement about a population mean or a population variance.

Example B.9 (t-test for population mean)**.** Suppose you are interested in measuring mean hourly wage of males aged 25-35. Accordingly, we collect a sample of 100 workers from the population of male in this age group with a mean of μ_X and a standard deviation of σ_X. The sample mean is $\hat{\mu}_X = \$25$ and the sample standard deviation is $\hat{\sigma}_X = \$7$. Now, suppose we want to test the following hypothesis:

$$H_0 : \mu_X = 27$$

$$H_0 : \mu_X \neq 27$$

The test statistic is given by the *t-statistic* where:

$$t = \frac{\hat{\mu}_X - \mu_X}{s.e.(\hat{\mu}_X)}$$

where $s.e.(\hat{\mu}_X) = \dfrac{\hat{\sigma}_X}{\sqrt{N}}$ is the standard error of sample mean and N denotes sample size.

If the null hypothesis is true, this test statistic follows **t-distribution** with N-1 degrees of freedom. Using the t-distribution table we can then compute the critical value which is used in formulating the decision rule. Let t_c denote this critical value from the distribution table. Then,

$$|t| > t_c \quad \Rightarrow \text{reject } H_0$$

$$|t| < t_c \quad \Rightarrow \text{do not reject } H_0$$

In our example, $N = 100$, and

$$t = \frac{25 - 27}{\frac{7}{\sqrt{100}}} = -2.86$$

The degrees of freedom is $N - 1 = 99$ and at 5% level of significance the critical value from the t-distribution table is $t_c = 1.98$. Because |t| is larger than the critical value, we reject the null hypothesis. Hence, we find evidence against the statement that the mean hourly wage of male workers is \$25.

Note that an alternative way of testing hypothesis like this is to use the **p-value** rule. The underlying idea is to find out the largest significance level at which we will fail to reject the null hypothesis. This value is called the p-value and most statistical softwares report this value. The decision-rule is then greatly simplified:

If p-value is less than the chosen level of significance (value of α) then reject H_0.

In our case, the p-value is 0.0053. Because we chose $\alpha = 0.05$, according to the p-value rule we will reject the null hypothesis.

Example B.10 (Chi-square test for population variance)**.** Using the same example, we also test a statement about the population variance. Suppose we want to test whether the variance of the hourly wage is 52.

$$H_0 : \sigma_X^2 = 52$$

$$H_0 : \sigma_X^2 > 52$$

The test statistic is given by the *V-statistic* where:

$$V = \frac{(N-1) \times \hat{\sigma_X^2}}{\sigma_X^2}$$

If the null hypothesis is true, this test statistic follows **Chi-square distribution** with N-1 degrees of freedom. Using the distribution table we can then compute the critical value which is used in formulating the decision rule. Let V_c denote this critical value from the distribution table. Then,

$$V > V_c \quad \Rightarrow \text{reject } H_0$$

$$V < V_c \quad \Rightarrow \text{do not reject } H_0$$

In our example,

$$V = \frac{(100-1) \times 7^2}{52} = 93.29$$

The degrees of freedom is $N - 1 = 99$ and at 5% level of significance the critical value from the Chi-square distribution table is $V_c = 43.77$. Because V is larger than the critical value, we reject the null hypothesis. Hence, we find evidence against the statement that the variance of the hourly wage of male workers is 52.

B.9.2 Testing a restriction on multiple population parameter

Often we are interested in testing a restriction that is a linear combination of two or more population means. Similarly, we maybe interested in comparing the variance of two different populations. In such cases we need to develop statistical tests that allow for comparison between parameters of different populations with given means and variances.

Example B.11 (t-test for comparing population mean of two populations)**.** Suppose you are interested in comparing mean weekly hours studied by Econ majors (X) and non-Econ majors in the college of business. For this purpose, you collect a sample of 25 econ majors and a sample of 30 non-econ majors. The sample mean of weekly hours studied by econ majors is 10 hours with a standard deviation of 4 hours. The sample mean of weekly hours studied by non-econ majors is 8 hours with a standard deviation of 2 hours. Also suppose that the covariance between weekly hours studied by econ and non-econ majors is 0.12. Test whether mean weekly hours studied by econ majors is more than the mean weekly hours studied by non-Econ majors.

Let X denote hours studied, N_X denotes sample size, $\hat{\mu}_X$, and $\hat{\sigma}_X$ denote sample mean and standard deviation, respectively for econ majors. Similarly, let Y denote hours studied, N_Y denotes sample size, $\hat{\mu}_Y$, and $\hat{\sigma}_Y$ denote sample mean and standard deviation, respectively for non-econ majors.

The first step, as usual, is to formulate the null and the alternative hypotheses:

$$H_0 = \mu_X - \mu_Y = 0$$
$$H_A = \mu_X - \mu_Y > 0$$

The next step is to compute the relevant test statistic, which in this case is the t-ratio given by:

$$t = \frac{(\hat{\mu_X} - \hat{\mu_Y}) - 0}{s.e.(\hat{\mu_X} - \hat{\mu_Y})}$$

Using the properties of variance, we get:

$$s.e.(\hat{\mu_X} - \hat{\mu_Y}) = \sqrt{Var(\hat{\mu_X}) + Var(\hat{\mu_Y}) - 2 \times Cor(X,Y) \times s.e.(\hat{\mu_X}) \times s.e.(\hat{\mu_Y})} = 0.84$$

So, $t = \dfrac{10-8}{0.84} = 2.38$

The sample size here is $N_X + N_Y = 55$. Using 5% level of significance and degrees of freedom of 53, the critical value from the t-distribution table for the one-sided alternative is 1.67. Because the |t| is more than 1.67, we reject the null hypothesis. We find evidence for econ majors studying more on average than non-econ majors in our sample.

Example B.12 (F-test for comparing population variance of two populations)**.** Often we may be interested in comparing the variability between two populations. Using our previous example, we may want to test whether variability in hours studied is bigger for econ majors versus non-econ majors. This can be tested by comparing the ratio of two variances against the value of 1. As before, we start by formulating the null and the alternative hypotheses:

$$H_0 : \sigma_X^2 / sigma_Y^2 = 1$$
$$H_0 : \sigma_X^2 / sigma_Y^2 > 1$$

The corresponding test statistic is the F-ratio:

$$F = \frac{\hat{\sigma_X^2}}{\hat{\sigma_Y^2}} = \frac{4^2}{2^2} = 4$$

If the null hypothesis is true, the above test statistic follows F-distribution with $N_x - 1$ degrees of freedom for the numerator and $N_y - 1$ degrees of freedom for the denominator. At 5% level of significance, the critical value for $\nu_1 = 24$ and $\nu_2 = 29$ from the F-distribution table is 3. Because the computed F-ratio exceeds the critical value we reject the null hypothesis.

B.9.3 Confidence interval and Hypothesis testing

One issue with using a sample to estimate population parameters is that by definition a sample estimator will be different for different samples. Thus, sample mean provides no information about how close this estimator is to the true population mean. This uncertainty in estimation can be summarized by computing the standard deviation, with higher value of standar deviation indicating greater uncertainty about the true population parameter. A better measure of this uncertainty is the **confidence interval**.

Definition B.8 (Confidence Interval)**.** Suppose we draw a random sample $\{x_1, x_2, ..., x_N\}$ from a normally distributed population with mean of μ_X and a standard deviation of σ_X. Let $\hat{\mu_X}$ denotes the sample mean and $\hat{\sigma_X}$ denotes sample standard deviation. Then, the 95% confidence interval for $\hat{\mu_X}$ is given by:

$$\left[\hat{\mu_X} - t_{c,2-sided} \times \frac{\hat{\sigma_X}}{\sqrt{N}}, \hat{\mu_X} + t_{c,2-sided} \times \frac{\hat{\sigma_X}}{\sqrt{N}}\right]$$

where $t_{c,2-sided}$ is the critical value that can be obtained from the t-distribution table for a given level of signicance and degrees of freedom. For example, for a 95% confidence interval we will use 5% level of significance.

Example B.13. Suppose N=20, $\hat{\mu_X} = 5$, and $\hat{\sigma_X} = 2$. Then, the 95% confidence interval for $\hat{\mu_X}$ is given by:

$$\left[5 - 2.093 \times \frac{2}{\sqrt{20}}, 5 + 2.093 \times \frac{2}{\sqrt{20}}\right] = [4.06, 5.94]$$

Hence, before we drew our sample from the population, there is a 95% chance that the true population parameter (μ_X) will fall between 4.12 and 5.94. Note that:

1. Wider the confidence interval, greater is the uncertainty about the true value of the population mean.
2. We can use the confidence interval to conduct hypothesis testing for a **two-sided** alternative hypothesis. If the null hypothesis value does not fall in the confidence interval, then with 95% confidence (or at 5% level of significance) we can reject the null hypothesis. For example, consider the following test:

$$H_0 : \mu_X = 3.8$$
$$H_A : \mu_X \neq 3.8$$

Because 3.8 is not in the confidence interval we will reject the null hypothesis at 5% level of significance. Note that we will obtain the same conclusion if we were to compute the t-ratio and compare it with the corresponding critical value from the t-distribution table.

Problems

Exercise B.1. Suppose you roll a 6-sided fair dice. If an odd number shows you win $10. If either 2 or 4 shows you lose $5. If 6 shows, you neither gain nor lose anything.

a. Denote the winnings from this game as X. Tabulate the probability distribiution of the random variable X.

b. Compute the expected value and the standard deviation for X.

Exercise B.2. Consider a population with a mean of μ and variance of σ^2. Suppose you draw a random sample $X_1, X_2, .., X_N$.

a. Show that $\hat{\mu_A} = 0.25 \times X_1 + 0.25 \times X_3 + 0.25 \times X_8 + 0.25X_20$ is an unbiased estimator of μ.

b. Show that $\hat{\mu_B} = 0.1 \times X_1 + 0.1 \times X_3 + 0.5 \times X_8 + 0.3 \times X_11$ is an unbiased estimator of μ.

c. Now compute variance of $\hat{\mu_A}$ and $\hat{\mu_B}$. Which one is more efficient estimator of μ.

Exercise B.3. Suppose you collect a random sample of 100 observations and find that sample mean is -25 and sample variance is 350.

a. Test whether the population mean is -22.

b. Test whether the population variance is 400.

Exercise B.4. Suppose you are interested in comparing performance of two different mutual funds, X and Y. Let μ_X and mu_Y denote unknown population mean returns on investment in X and Y, respectively. Suppose you collect past 20 months data for both mutual funds and find that sample mean for fund X is 2% with a standard deviation of 0.5%. In contrast, the sample mean for fund Y is 5% with a standard deviation of 2%. Suppose that the correlation between returns on these two funds is 0.2.

a. Test whether mean return on Y is greater than that on X.

b. Test whether variance of Y is greater than that of X.

c. Compute the 95% confidence interval for $\hat{\mu_X}$. Using the confidence interval, what can you say about the population mean return for fund X?