Homework 5

Ze Yang (zey@andrew.cmu.edu)

Due Thursday, October 5 at 3:00 PM

You should submit the Rmd file for your analysis. Name the file as `YOURANDREWID_HW5.Rmd` and submit it via Canvas. Also submit the `.pdf` file that is produced.

i Go to the website https://www.kaggle.com/sohier/interest-rate-records and download the

data set of daily interest rates found there. These data come from the Federal Reserve, see https://www.federalreserve.gov/releases/h15/ fsor more details.

There are clearly strong relationships among these interest rates; for example, there is a strong correlation between the prime rate and the federal funds rate. Here we are interested in the following question:

> Describe some ways (if there are any) that the relationships among these rates have changed over the time frame of this data set, i.e., since 1954.

You are to present analyses that explore this question. You have two tools at your disposal: visualization and simple linear regression.

Your response should include at least five plots, and there should be at least three regression models fit.

Your final analysis should involve **at least** three different rates.

Please note that I am **not** asking you try to find **every** way in which these relationships have changed. Further, a reasonable conclusion is that no changes are found. If that is your conclusion, there still should be figures and models to back this up.

You will be graded based on the quality and clarity of your figures, the relevance of the models fit, and manner in which you can present a coherent description of the conclusions that you drew from the figures and models.

You will **not** be rewarded for excessive figures, models, and/or text. Present your case **clearly and concisely**.

The six best responses, as judged by myself and the TAs, will receive two bonus points on the final exam. (The final will have 100 possible points.)

An Explorative Analysis On U.S. Yield Curve and Business Cycle

In this analysis we focus on the evolution of U.S. yield curve, its shape and relationship to the business cycle. We mainly address two questions:

- How does the shape of yield curve change over time?
- What is the relationship, if there is any, between the shape of yield curve and business cycle?

To answer these two questions we will use part of the Federal Reserve data set. We use the group of rates: `X_[T]_treasury_constant_maturity`, i.e. the treasury constant maturity rates with different maturities to construct the yield curve. Denote the treasury (spot) rates as R, then the yield curve is a mapping from time to maturity to the corresponding spot rate:

$$f_t : T \mapsto R_t(T)$$

Where the subscript t denotes the current time, i.e. f_t is the yield curve observed at time t. Capital T denotes the maturity of spot rate: $R_t(T)$ is the T-year spot rate observed at time t. The unit of T is *year*. There are 11 anchor rates provided in the data frame: $T \in \{\frac{1}{12}, \frac{1}{4}, \frac{1}{2}, 1, 2, 3, 5, 7, 10, 20, 30\}$. Of course we can interpolate between the anchors to approximate an continuous yield curve. For our purposes in this analysis, and for simplicity, we will just look at the discrete yield curve at those anchors. To aviod cumbersome notations, we denote the constant maturity rates as `R.[T]` in the data frame.

1. Import and Clean Rates Data

```
rates = read.csv('rates.csv')
colnames(rates) = c(
  'date', 'fed.fund', 'nf.1m', 'nf.2m', 'nf.3m',
  'f.1m', 'f.2m', 'f.3m',
  'prime', 'disc', 'tbill.4w', 'tbill.3m',
  'tbill.6m', 'tbill.1y', 'R.1m', 'R.3m',
  'R.6m', 'R.1y', 'R.2y', 'R.3y', 'R.5y',
  'R.7y', 'R.10y', 'R.20y', 'R.30y',
  'itcm.5y', 'itcm.7y', 'itcm.10y', 'itcm.20y',
  'itcm.30y', 'inflation')
tb.cols = c('R.1m', 'R.3m',
            'R.6m', 'R.1y', 'R.2y', 'R.3y', 'R.5y',
            'R.7y', 'R.10y', 'R.20y', 'R.30y')
rates[rates==-9999] = NA
rates$date = as.Date(rates$date)
tb.rates = rates[,c('date', tb.cols)]
str(tb.rates)
```

```
## 'data.frame':    23179 obs. of  12 variables:
##  $ date : Date, format: "2017-08-09" "2017-08-08" ...
##  $ R.1m : num  1.01 1 0.99 NA NA 1 1 1.02 1 1 ...
```

```
##  $ R.3m : num  1.06 1.06 1.02 NA NA 1.08 1.08 1.08 1.08 1.07 ...
##  $ R.6m : num  1.15 1.16 1.14 NA NA 1.14 1.13 1.15 1.15 1.13 ...
##  $ R.1y : num  1.21 1.24 1.22 NA NA 1.23 1.22 1.24 1.22 1.23 ...
##  $ R.2y : num  1.33 1.36 1.36 NA NA 1.36 1.34 1.36 1.34 1.34 ...
##  $ R.3y : num  1.5 1.53 1.52 NA NA 1.51 1.49 1.52 1.5 1.51 ...
##  $ R.5y : num  1.81 1.84 1.81 NA NA 1.82 1.79 1.82 1.8 1.84 ...
##  $ R.7y : num  2.06 2.1 2.07 NA NA 2.08 2.05 2.08 2.07 2.11 ...
##  $ R.10y: num  2.24 2.29 2.26 NA NA 2.27 2.24 2.27 2.26 2.3 ...
##  $ R.20y: num  2.59 2.63 2.6 NA NA 2.61 2.56 2.6 2.61 2.66 ...
##  $ R.30y: num  2.82 2.86 2.84 NA NA 2.84 2.81 2.85 2.86 2.89 ...
```

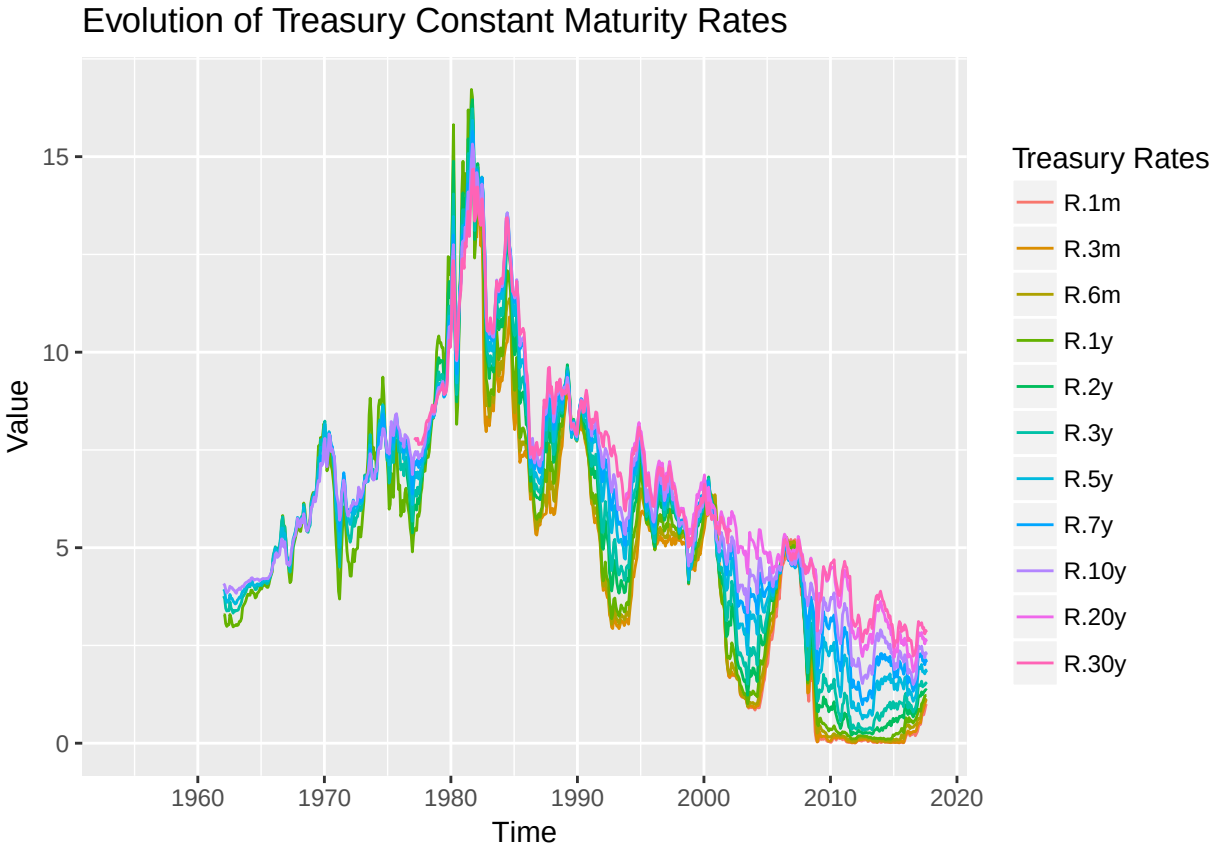
2. Yield Curve Evolution

There are many small discontinuities in the time series data of rates, due to weekends and holidays. They can be very distracting in the plots. Therefore, we choose to aggregate the rates to their monthly average to remove those small gaps, and make the plot smoother for our first analysis.

```
melt.rates = function(data, cols, idx='date') {
  #' Helper function to melt the data.
  melted = melt(
    data[,c(idx, cols)], id=idx)
  return(melted)
}

line.plot = function(melted, title='') {
  #' Helper function to produce line plots
  #' of melted data.
  g = ggplot(data=melted,
         aes(x=date, y=value,
             colour=variable)) +
    geom_line() +
    labs(x="Time", y="Value",
         colour="Treasury Rates", title=title) +
    scale_x_date(breaks = pretty_breaks(10))
  return(g)
}

# Aggregate to monthly mean.
tb.agg.ym = tb.rates %>% group_by(
  year.month=floor_date(date, "month")) %>%
  summarize_all(funs(mean(., na.rm=T)))
tb.agg.ym = as.data.frame(tb.agg.ym)
```

```
tb.melted = melt.rates(tb.agg.ym, tb.cols)
line.plot(tb.melted, 'Evolution of Treasury Constant Maturity Rates')
```

```
## Warning: Removed 2835 rows containing missing values (geom_path).
```



We have several observations:

- The long end (spot rates with longer maturity) of the yield curve is greater than the short end (spot rates with shorter maturity) in most of the time, i.e. the yield curve is usually upward sloping.
- The spread between the long and short ends differs from time to time. In some years, `2010-2016` for example, the spread is larger, which implies that the yield curve is **steeper** than other years. However, in `1980-1982` and `2008`, the spread is narrow, i.e. the yield curve is **flatter** than other year.
- In `2010-2016`, the short end rate like $R(\frac{1}{12})$ is 0.

3. Typical Shapes of Yield Curve

In this subsection we look at the shape of yield curve at some particular dates of interest. The dates that we picked are:

- `June 2015`, a recent date.
- `May 2007`, a date that is close to 2007-2009 great recession.
- `July 1981`, a date that is close to 1981-1982 recession induced by energy crisis.

```
# times series object of treasury rates
tb.agg.ts = xts(tb.agg.ym[,-1], order.by=tb.agg.ym$year.month)
```

```
pick.curve = function(ts, date) {
  #' Helper function to pick the yield curve
  #' at a specific date.
  curve = as.numeric(ts[date,-1])
  n.anchors = length(curve)
  curve.df = data_frame(
    maturity=c(1/12, 1/4, 1/2, 1,2,3,5,7,10,20,30),
    curve=curve, date=rep(date, n.anchors))
  return(as.data.frame(curve.df))
}

# Produce the plots of yield curve on three dates.
curve = rbind(
  pick.curve(tb.agg.ts, '2007-05-01'),
  pick.curve(tb.agg.ts, '2015-06-01'),
  pick.curve(tb.agg.ts, '1981-07-01')
)
ggplot(data=curve, mapping=aes(x=maturity, y=curve, color=date)) +
  geom_point() +
  geom_line() +
  facet_grid(.~date)
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_path).
```

The facet plots above illustrate the yield curve at three interesting dates. Each facet corresponds to the yield curve at a date, and x-axis stands for maturities (T), y-axis stands for the Treasury spot constant maturity rate with maturity T, i.e. $R(T)$.

We observe:

- In `June 2015`, the yield curve is upward sloping like it normally is.
- In `May 2007`, prior to the great recession, the yield curve is very flat compared with the `2015` shape.

- In `July 1981`, which was also an 1-year recession period, the yield curve is downward sloping.

3. Evolution of the Spread

Clearly, the yield curve is not a linear function at any time. However, we can approximately capture the evolution of the "slope" of the yield curve with the spread between long and short end rates. At time t, we define

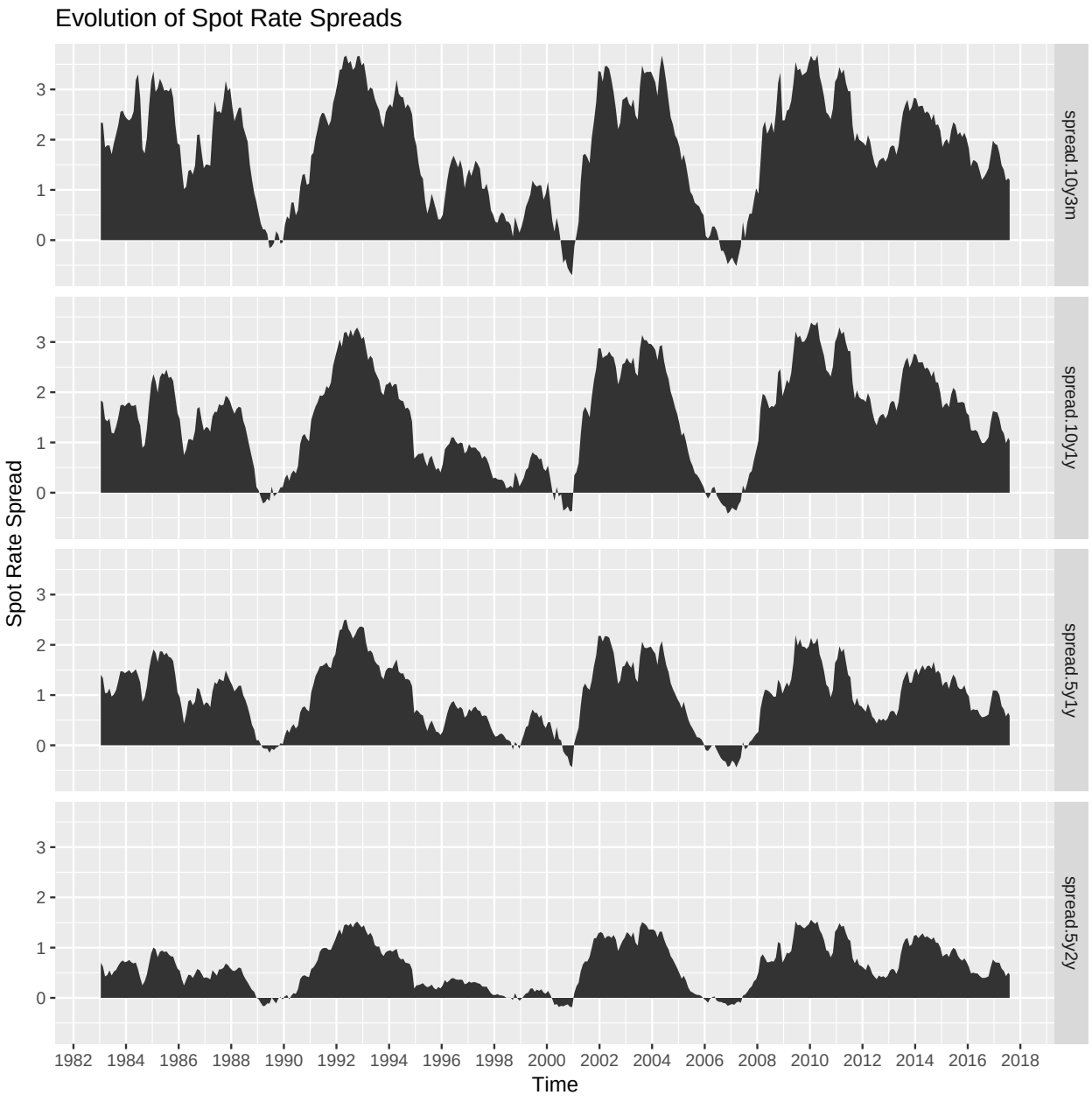
$$S_t(T_1, T_2) = R_t(T_2) - R_t(T_1) \qquad T_2 > T_1$$

As the **spread** between spot rate with maturity T_2 and that with maturity T_1. For fixed long and short anchors T_2, T_1, positive spread corresponds to the upward sloping curve. Since the x-difference $T_2 - T_1$ is fixed, the magnitude of spread is also correlated with the steepness of the curve.

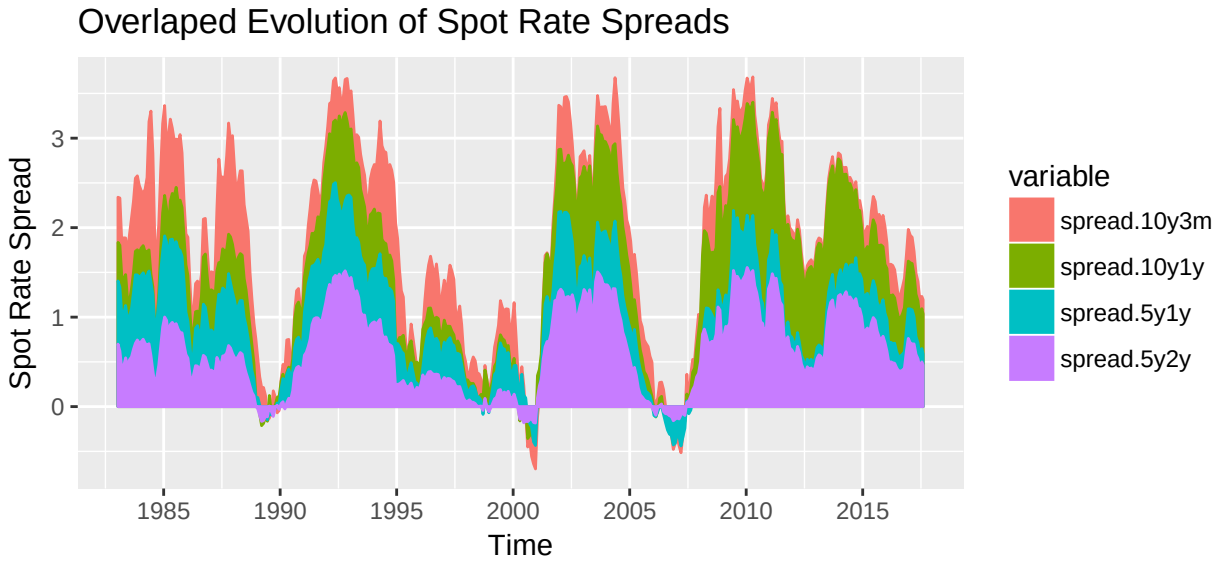
We calcute the spread for difference choices of (T_1, T_2) pairs. Suppose at time t, there are 4 anchors on the yield curve: $T_1 < T_2 < T_3 < T_4$. We expect

If the curve is upward or downward sloping everywhere, i.e. there is no "twisted" or "butterfly" structure.

```
# Calculate the spread for different choices of (T1, T2) anchor.
tb.agg.ym$spread.10y3m = tb.agg.ym$R.10y - tb.agg.ym$R.3m
tb.agg.ym$spread.10y1y = tb.agg.ym$R.10y - tb.agg.ym$R.1y
tb.agg.ym$spread.5y1y = tb.agg.ym$R.5y - tb.agg.ym$R.1y
tb.agg.ym$spread.5y2y = tb.agg.ym$R.5y - tb.agg.ym$R.2y

# Create melted data frame
spread.names = c('spread.10y3m', 'spread.10y1y',
                 'spread.5y1y', 'spread.5y2y')
# truncate the data to remove NA periods
spread.melted = melt.rates(
  tb.agg.ym[tb.agg.ym$date>'1983-01-01',], spread.names)

# Ribbon plot
ggplot(spread.melted, aes(x=date)) +
  geom_ribbon(aes(ymin=0,ymax=value)) +
  labs(x="Time", y="Spot Rate Spread",
       title='Evolution of Spot Rate Spreads') +
  facet_grid(variable~.) +
  scale_x_date(breaks = pretty_breaks(20))
```




```
# overlaped ribbon plot
ggplot(spread.melted, aes(x=date, colour=variable)) +
  geom_ribbon(aes(ymin=0,ymax=value, fill=variable)) +
  labs(x="Time", y="Spot Rate Spread",
       title='Overlaped Evolution of Spot Rate Spreads') +
  scale_x_date(breaks = pretty_breaks(10))
```



We present the evolution of spot rate spreads with different choices of (T_1, T_2) in the facet and overlapped ribbon plot above. We selected four (T_1, T_2) pairs, namely:

- `spread.10y3m` : $T_1 = 1/4$, $T_2 = 10$.
- `spread.10y1y` : $T_1 = 1$, $T_2 = 10$.
- `spread.5y1y` : $T_1 = 1$, $T_2 = 5$.
- `spread.5y2y` : $T_1 = 2$, $T_2 = 5$.

Note that the four ranges are chosen such that: $(2, 5) \subset (1, 5) \subset (1, 10) \subset (1/4, 10)$. From the plots we can conclude that:

- In most of the time, all the four spreads are positive, i.e. the yield curve is upward sloping.
- There are some periods in which the spread between long and short ends is negative. In these periods, the yield curve is **inverted** (downward sloping). The periods of inverted yield curve coincide with the periods of recession in the United States. These periods are: `1989-1990`, `2001`, and `2007-2008` (see https://en.wikipedia.org/wiki/List_of_recessions_in_the_United_States).
- Given anchors $T_1 < T_2 < T_3 < T_4$, In most of the time, $S_t(T_1, T_4)$ is strongly positively correlated with $S_t(T_2, T_3)$. And the spread on the narrower range is "included" within the spread on the wider range (also smaller in magnitude. I.e.

$$S_t(T_1, T_4) \propto S_t(T_2, T_3)$$

$$|S_t(T_1, T_4)| > |S_t(T_2, T_3)|$$

4. Correlation Between the Spreads

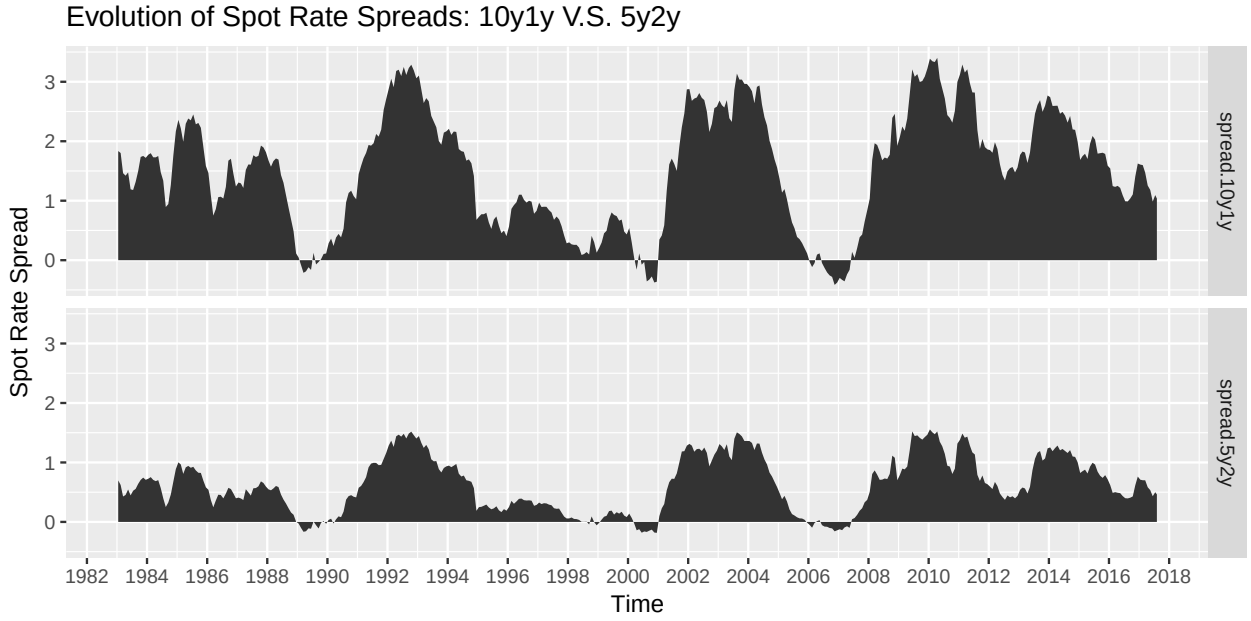
In this subsection we try to quantify the correlation between the spot rate spreads. In last subsection we have already seen that given anchors $T_1 < T_2 < T_3 < T_4$, the narrower and wider spreads are stongly correlated if the narrower is contained within the wider.

$$S_t(T_1, T_4) \propto S_t(T_2, T_3)$$

We consider another case: what if the two ranges are disjoint? Is there any correlation between the spread on the short end and that on the long end? Using same notation:

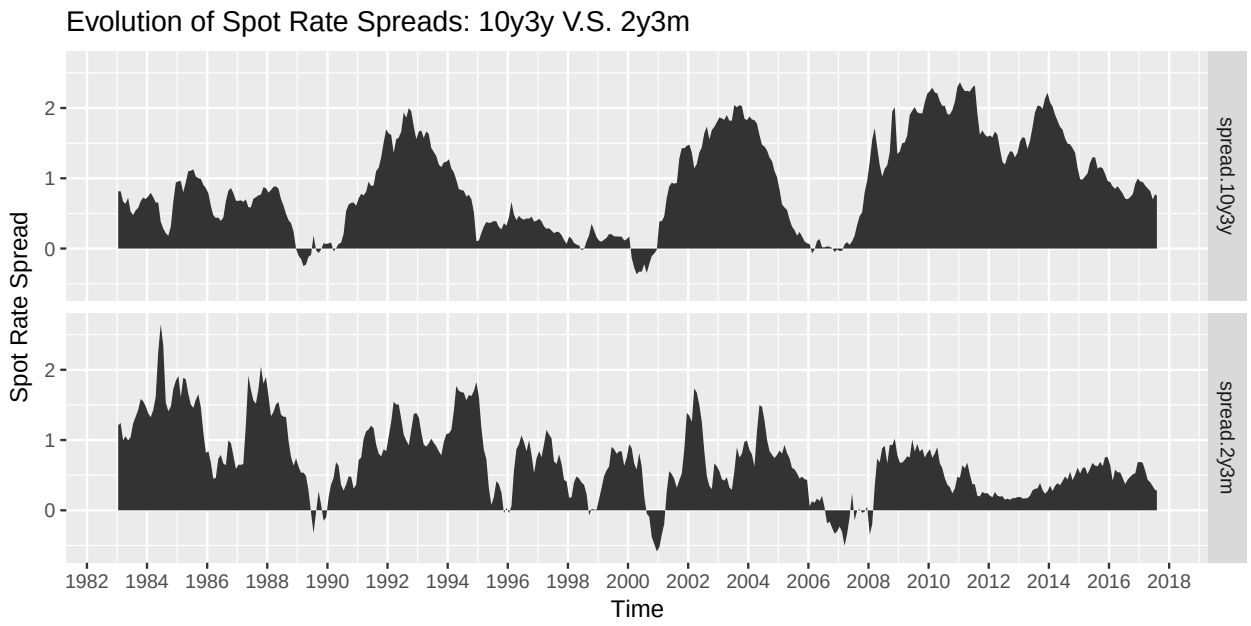
$$S_t(T_1, T_2) \quad \text{and} \quad S_t(T_3, T_4)$$

```
# Create melted data frame
spread.names.3 = c('spread.10y1y', 'spread.5y2y')
spread.melted.3 = melt.rates(
  tb.agg.ym[tb.agg.ym$date>'1983-01-01',], spread.names.3)

# Ribbon plot
ggplot(spread.melted.3, aes(x=date)) +
  geom_ribbon(aes(ymin=0,ymax=value)) +
  labs(x="Time", y="Spot Rate Spread",
       title='Evolution of Spot Rate Spreads: 10y1y V.S. 5y2y') +
  facet_grid(variable~.) +
  scale_x_date(breaks = pretty_breaks(20))
```




```
# Calculate the spread for the second case
tb.agg.ym$spread.10y3y = tb.agg.ym$R.10y - tb.agg.ym$R.3y
tb.agg.ym$spread.2y3m = tb.agg.ym$R.2y - tb.agg.ym$R.3m


# Create melted data frame
spread.names.2 = c('spread.10y3y', 'spread.2y3m')
spread.melted.2 = melt.rates(
  tb.agg.ym[tb.agg.ym$date>'1983-01-01',], spread.names.2)

# Ribbon plot
ggplot(spread.melted.2, aes(x=date)) +
```

```
  geom_ribbon(aes(ymin=0,ymax=value)) +
  labs(x="Time", y="Spot Rate Spread",
       title='Evolution of Spot Rate Spreads: 10y3y V.S. 2y3m') +
  facet_grid(variable~.) +
  scale_x_date(breaks = pretty_breaks(20))
```



For the first case, we analyze

- `spread.10y1y` : $T_{short} = 1$, $T_{long} = 10$.
- `spread.5y2y` : $T_{short} = 2$, $T_{long} = 5$.

Note that we have $(2, 5) \subset (1, 10)$.

For the second case, we analyze

- `spread.10y3y` : $T_{short} = 3$, $T_{long} = 10$.
- `spread.2y3m` : $T_{short} = 1/4$, $T_{long} = 2$.

Note that we have $(1/4, 2)$ disjoint to $(3, 10)$.

From the ribbon plots, it seems that the correlation between two spreads in the second case is weaker than the first case. We run linear regression to further investigate.

```
df = tb.agg.ym[,c('date', spread.names.2, spread.names.3)]
df = as.data.frame(df[df$date>'1983-01-01',])
model.1 = lm(spread.10y1y~spread.5y2y, df)
summary(model.1)
```

```
##
## Call:
## lm(formula = spread.10y1y ~ spread.5y2y, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.43576 -0.11991 -0.00998  0.08148  0.46244
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.23265    0.01466   15.87   <2e-16 ***
## spread.5y2y  2.06020    0.01882  109.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1802 on 414 degrees of freedom
## Multiple R-squared:  0.9666, Adjusted R-squared:  0.9665
## F-statistic: 1.198e+04 on 1 and 414 DF,  p-value: < 2.2e-16
```

```
model.2 = lm(spread.10y3y~spread.2y3m, df)
summary(model.2)
```

```
##
## Call:
## lm(formula = spread.10y3y ~ spread.2y3m, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2622 -0.5513 -0.1420  0.5623  1.4931
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.79279    0.05507  14.397  < 2e-16 ***
## spread.2y3m  0.17437    0.06198   2.813  0.00514 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6785 on 414 degrees of freedom
## Multiple R-squared:  0.01876,    Adjusted R-squared:  0.01639
## F-statistic: 7.915 on 1 and 414 DF,  p-value: 0.005137
```

The R^2 quantity for the first case is 0.9666, while for the second case it's only 0.01876. In both cases, the coefficent β_1 is positive, which implies a positive linear correlation.

The result of linear regression supports our observation from the plots. The correlation between the two spreads in the first case (wider includes narrower) is very strong. However, the correlation between the two spreads in the first case (short end disjoint to long end) is weak.

5. Rolling Correlation and Linear Regression Between Spreads

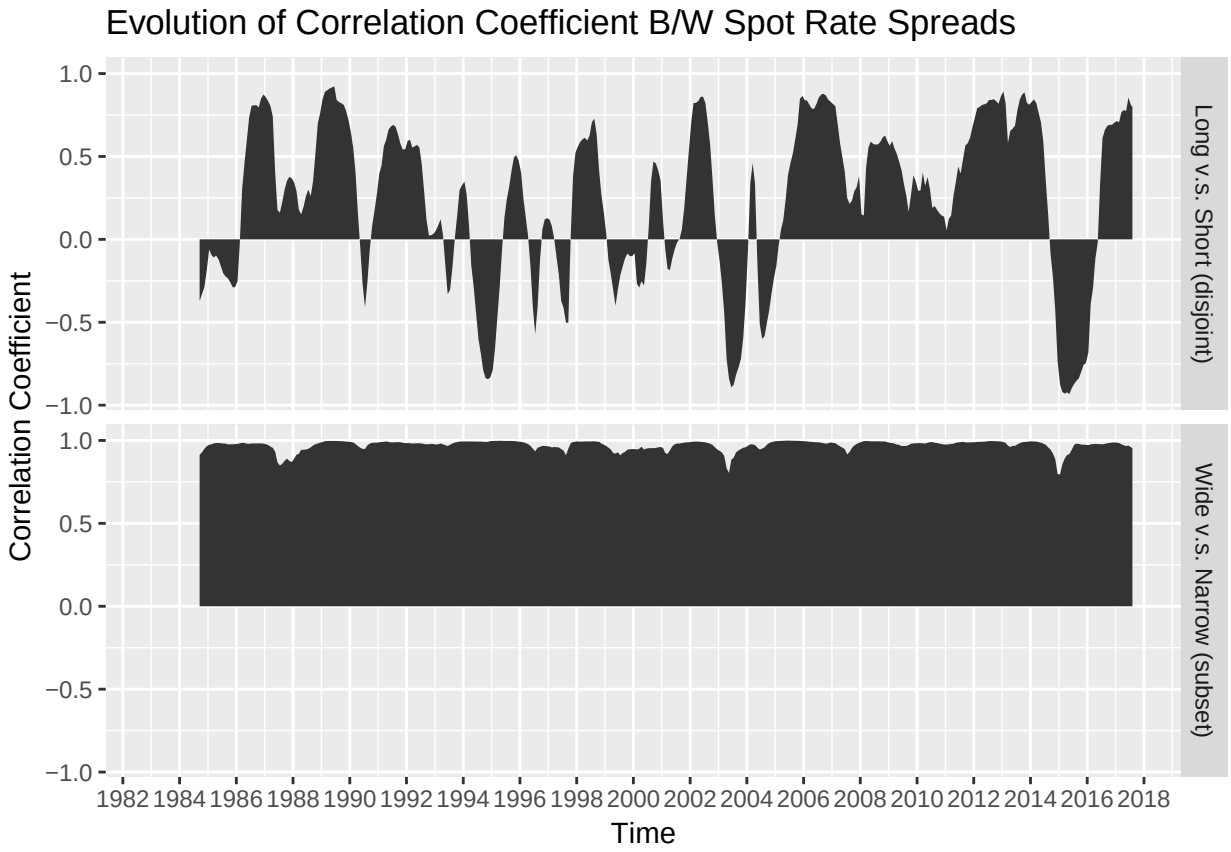
We perform rolling linear regression and rolling estimation of correlation coefficient over time to investigate how the corrlation between two spreads changes on the time horizon.

We firstly choose a fixed window width w, then loop for every $w < t \leq N$. At time t, we estimate a linear model with the data ranges from $[t - w, t]$, and also calcuate sample correlation in this manner.

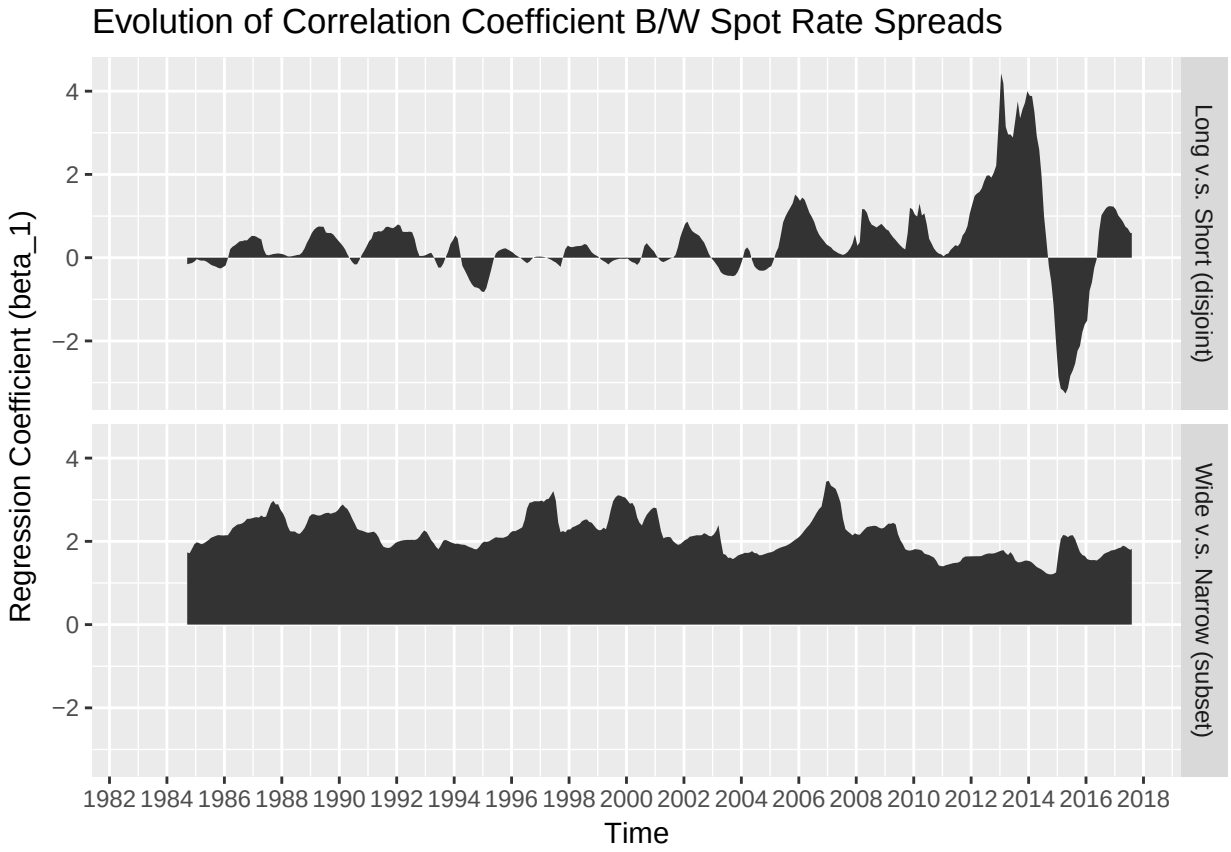
```
rolling.lm = function(ts.data, window, label) {
  # rolling simple linear regression estimate
  est = rollapply(ts.data,
            width = window,
            FUN=function(X) {
              X = as.data.frame(X)
              model = lm(formula=X[,1] ~ X[,2],
                         data=X)
              return(model$coef)
            },
            by.column=FALSE)
  est.df = data.frame(date=index(est),
                      est=as.vector(est[,2]))
  est.df$window = rep(window, length(est.df$est))
  est.df$label = rep(label, length(est.df$est))
  return(est.df)
}

rolling.corr = function(ts.data, window, label) {
  cor = rollapply(ts.data, width=window ,
                  FUN= function(x) {
                    cor(x[,1],x[,2])
                  }, by.column=FALSE)
  cor.df = data.frame(date=index(cor),
                      est=as.vector(cor[,1]))
  cor.df$window = rep(window, length(cor.df$est))
  cor.df$label = rep(label, length(cor.df$est))
  return(cor.df)
}

tb.agg.trunc = tb.agg.ym[tb.agg.ym$year.month>'1983-01-01',-1]
tb.agg.ts = xts(tb.agg.trunc[,-1],
                order.by=tb.agg.trunc$date)
spread.ts.2 = tb.agg.ts[,c(spread.names.2)]
spread.ts.3 = tb.agg.ts[,c(spread.names.3)]
```

```
cor.2 = rolling.corr(spread.ts.2, 20, 'Long v.s. Short (disjoint)')
cor.3 = rolling.corr(spread.ts.3, 20, 'Wide v.s. Narrow (subset)')
cor = rbind(cor.2, cor.3)

ggplot(cor, aes(x=date)) +
  geom_ribbon(aes(ymin=0,ymax=est)) +
  labs(x="Time", y="Correlation Coefficient",
```

```
       title='Evolution of Correlation Coefficient B/W Spot Rate Spreads') +
  facet_grid(label~.) +
  scale_x_date(breaks = pretty_breaks(20))
```




```
est.2 = rolling.lm(spread.ts.2, 20, 'Long v.s. Short (disjoint)')
est.3 = rolling.lm(spread.ts.3, 20, 'Wide v.s. Narrow (subset)')
est = rbind(est.2, est.3)

ggplot(est, aes(x=date)) +
  geom_ribbon(aes(ymin=0,ymax=est)) +
  labs(x="Time", y="Regression Coefficient (beta_1)",
       title='Evolution of Correlation Coefficient B/W Spot Rate Spreads') +
  facet_grid(label~.) +
  scale_x_date(breaks = pretty_breaks(20))
```



As we observed from the plots, the correlation between `wide` and `narrow` is strong and stable over time. One plausible reason is that some spot rates are interpolated from other nodes, as a result, the narrower spread must be governed by wider spread, if the narrower segment is interpolated from the wider segment.

In contrast, the correlation between `long` and `short` is weak and very unstable over time.

Final Conclusion

- The shape of the U.S. yield curve is different from time to time. In most of the time it's upward sloping, but in some periods it can be flat or inverted.
- We can use the spread between long end spot rate and short end spot rate as a proxy of yield curve shape. The spreads are positive in most of the time. However, in some periods the spreads are negative, which coincide with the recession periods in the history.
- The correlation between `wide` and `narrow` spread is strong and stable over time, which is consistent with the fact that some narrow segments of the yield curve are interpolated from the wide segments.
- The correlation between spreads at the long and short ends of the curve is weak and unstable, because the yield curve is not linear and its shape changes from time to time.