

Empirical Methods Homework 4

Group 9: Linqi Huang, Abhesh Kumar, Yu Onohara, Maitrayee Patil, Redmond Xia

Problem 1

Question 1

Download data and compute dividend yield, excess return, term spread and default spread accordingly.

```
# Download data from CRSP and compute dividend yield
sp <- read_csv('sp500.csv')
sp <- sp %>%
  mutate(div = vwretd - vwretx) %>%
  mutate(yield_div = double(nrow(sp)))
sp$yield_div[1:12] <- NA
for (i in 13:nrow(sp)){
  sp$yield_div[i] = (sum(sp$div[(i-1) : (i-12)]) / (sp$vwretx[i]+1))
}

tb <- read_csv('t-bill.csv') %>%
  mutate(DATE = caldt)

# Compute excess return
df <- inner_join(sp, tb, by='DATE') %>%
  mutate(excess_return = log(vwretd+1) - log(t30ret+1)) %>%
  select(DATE, yield_div, excess_return) %>%
  mutate(DATE = floor_date(ymd(DATE), 'month'))

# Download data from FRED and compute spreads
t_sp <- read_csv('T10YFFM.csv') %>%
  mutate(T10YFFM = T10YFFM/100) %>%
  rename(spread_term = T10YFFM)
aaa <- read_csv('AAAFFM.csv')
baa <- read_csv('BAAFFM.csv')
d_sp <- inner_join(aaa, baa, by='DATE') %>%
  mutate(spread_default = BAAFFM/100 - AAAFFM/100)

spread <- inner_join(t_sp, d_sp, by='DATE') %>%
  select(DATE, spread_term, spread_default)

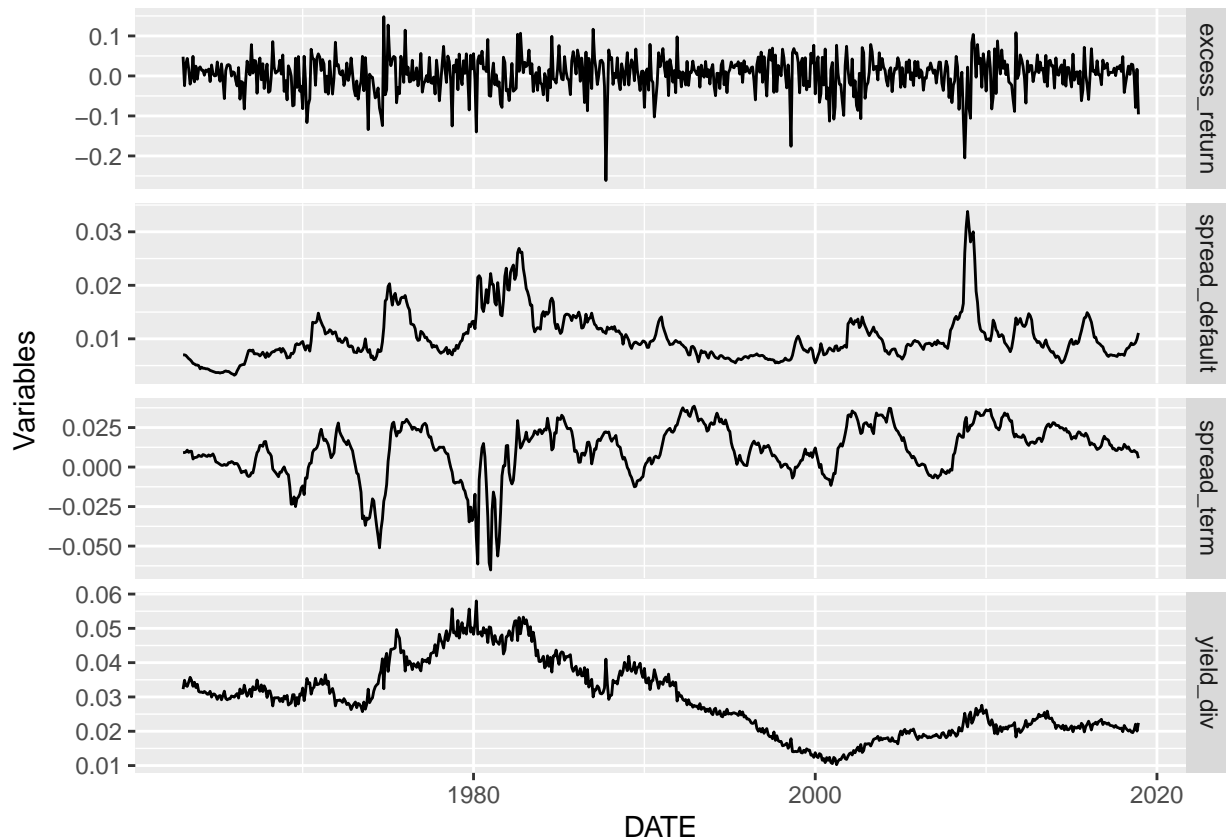
# Join dataframes to obtain all data needed for future computation
df <- inner_join(df, spread, by='DATE')
```

Question 2

The plot is shown below following sequence of excess return, default spread, term spread and dividend yield.

```
# Plot excess return, dividend yield, term spread, and default spread
gather(df, yield_div : spread_default, key='yeah', value='Variables') %>%
```

```
ggplot() +
  geom_line(aes(x=DATE, y=Variables)) +
  facet_grid(yeah~., scale='free')
```



Question 3

```
# Run regressions
lm1 <- lm(excess_return ~ lag(yield_div) + lag(spread_term) + lag(spread_default), data = df)
out1<- summary(lm1)
coef1<- as.numeric(out1$coefficients[,1])[-1]
ols1 <- as.numeric(out1$coefficients[,2])[-1]
reg1 <- c(coef1[1], ols1[1], NA, coef1[2], ols1[2], NA,
          coef1[3], ols1[3], NA, out1$r.squared)

roll_func <- function(window){
  roll = double(nrow(df))
  for (i in 1 : (nrow(df)-window)){
    roll[i] = sum(df$excess_return[(i+1) : (i+window)])
  }
  for (i in (nrow(df)-window+1): nrow(df)){
    roll[i] = NA
  }
  return (roll)
}

df$mo3 <- roll_func(3)
```

```

df$mo12 <- roll_func(12)
df$mo24 <- roll_func(24)
df$mo60 <- roll_func(60)

reg_func <- function(y, lags){
  lm <- lm(y ~ yield_div + spread_term + spread_default, data = df)
  out <- summary(lm)
  coef <- c(as.numeric(out$coefficients[,1])[-1])
  ols <- c(as.numeric(out$coefficients[,2])[-1])
  nw <- NeweyWest(lm, lag=(lags-1)*1.5, prewhite=F, adjust=T)
  nw_se <- as.numeric(sqrt(diag(nw))[-1])
  result <- c(coef[1], ols[1], nw_se[1], coef[2], ols[2], nw_se[2],
              coef[3], ols[3], nw_se[3], out$r.squared)
  return (result)
}

result <- data.frame(M1 = reg1, M3 = reg_func(df$mo3,3), M12 = reg_func(df$mo12,12),
                    M24 = reg_func(df$mo24,24), M60 = reg_func(df$mo60,60))

row.names(result) <- c('Div Yield', 'OLS_DY', 'NW_DY', 'Term Spread', 'OLS_TS', 'NW_TS',
                      'Default Spread', 'OLS_DS', 'NW_DS', 'R-squared')
kable(result, caption='Results from Regressions')

```

Table 1: Results from Regressions

	M1	M3	M12	M24	M60
Div Yield	0.2708027	0.8102389	2.5270878	3.1591194	4.2525139
OLS_DY	0.1839680	0.3292672	0.6436719	0.8788831	1.1475403
NW_DY	NA	0.5426110	1.9688428	2.9865273	3.3039729
Term Spread	0.3210713	0.8763170	3.2237231	4.1906402	7.1120431
OLS_TS	0.1062065	0.1899098	0.3699312	0.5042711	0.6542109
NW_TS	NA	0.3053231	1.0401685	1.4891624	3.0699116
Default Spread	0.1160907	0.4410030	2.3164509	3.7989955	16.9174753
OLS_DS	0.4164772	0.7446847	1.4508594	1.9815726	2.5842231
NW_DS	NA	1.6054568	3.1424198	4.6516347	8.6969069
R-squared	0.0161387	0.0390917	0.1269188	0.1217730	0.2615561

In this part, we run multiple linear regressions for three predictive variables (the lagged dividend yield, term spread, and default spread). The results from each of these regressions (regression coefficients, standard errors, and R^2) are reported as above. We will choose Newey-West standard error for all of regression models, except for the 1-month regression model. The reason is that for overlapping data with long horizons, we need Newey-West method to correct autocorrelation and heteroscedasticity within the standard error and the residuals.

Question 4

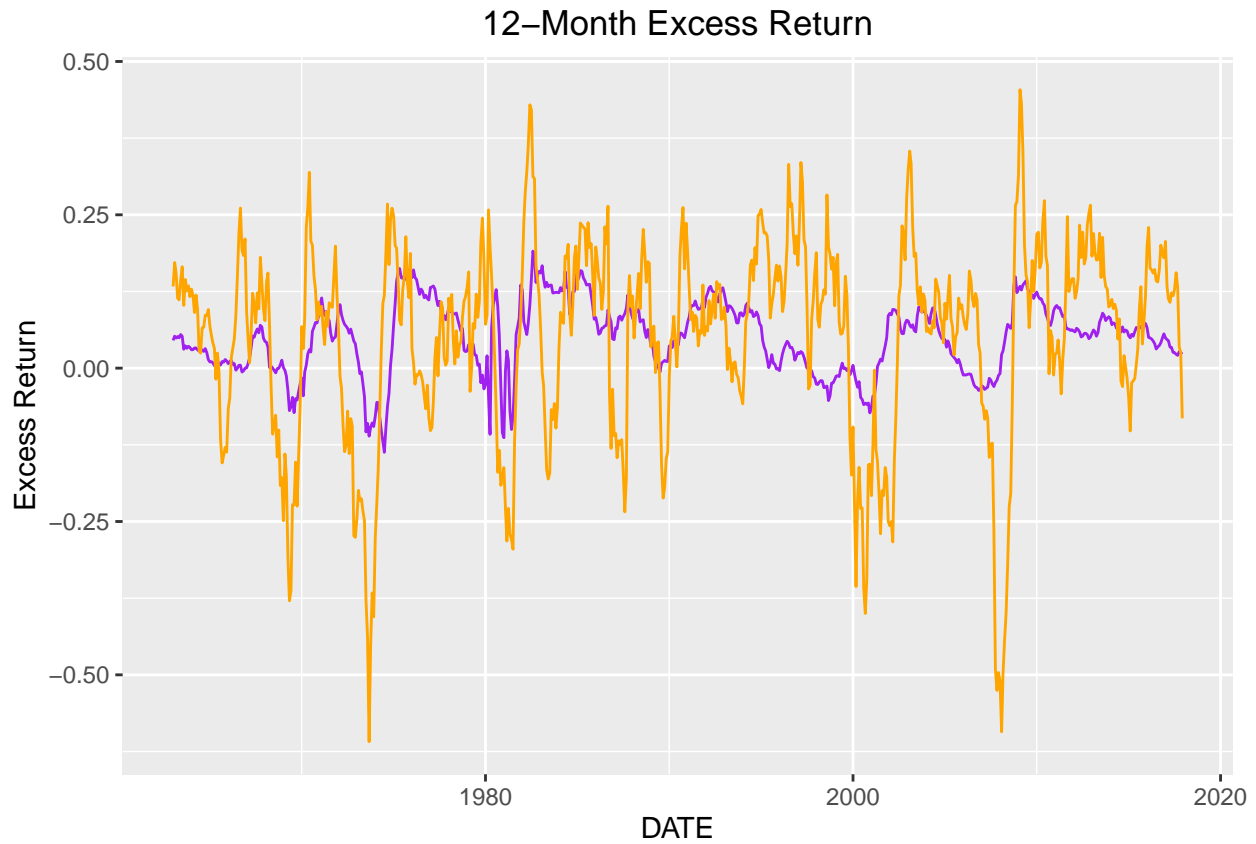
```

# Use 12-month regression model to obtain forecasting 12-month excess return
lm12 <- lm(mo12 ~ yield_div + spread_term + spread_default, data = df)
pred12 <- df %>%
  select(DATE, mo12, yield_div, spread_term, spread_default)
pred12 <- pred12 %>%
  mutate(pred = predict(lm12, pred12[,3:5])) %>%

```

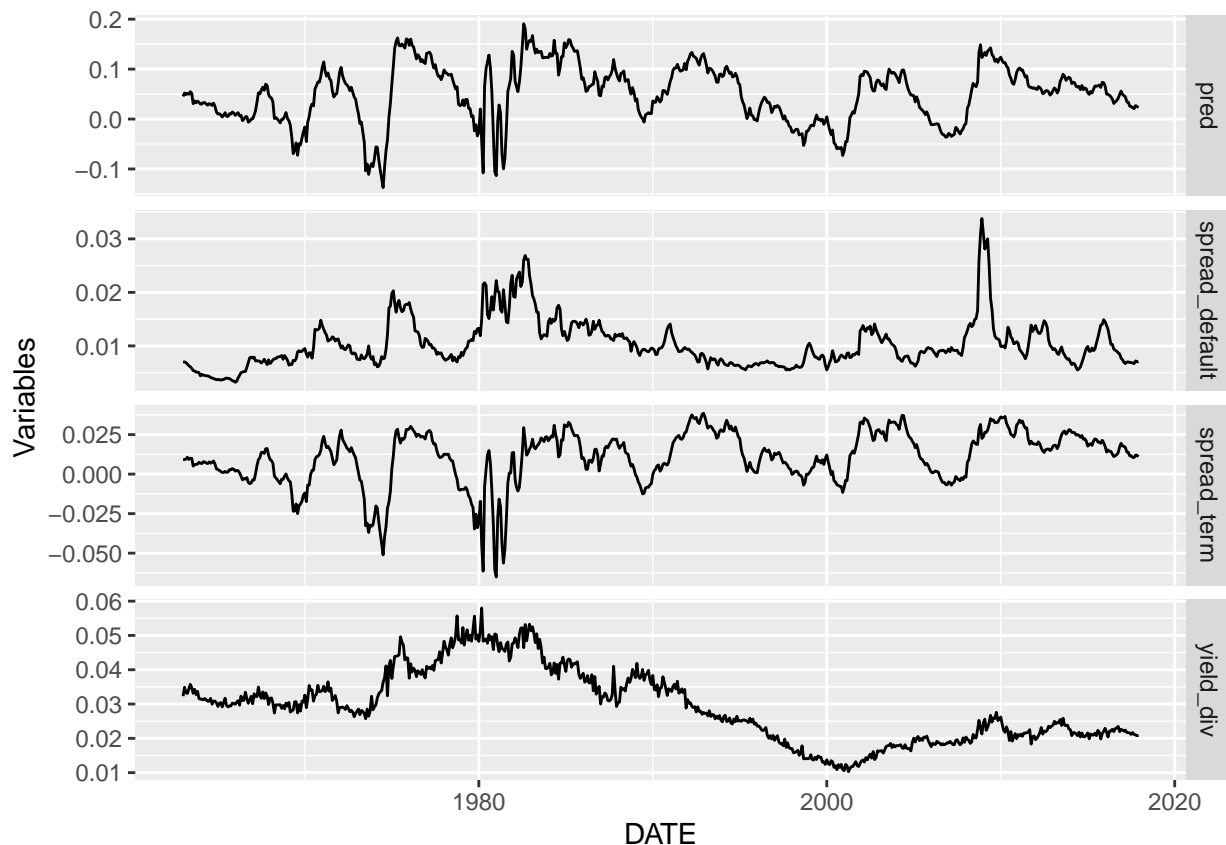
```
drop_na()

# Plot 12-Month Excess Return and predicted 12-Month Excess Return
ggplot(pred12) +
  geom_line(aes(DATE, pred), colour = 'purple') +
  geom_line(aes(DATE, mo12), colour = 'orange') +
  labs(y = 'Excess Return', title = '12-Month Excess Return') +
  theme(plot.title = element_text(hjust = 0.5))
```



12-Month Excess Return and predicted 12-Month Excess Return are plotted above, with 12-Month Excess Return in orange and predicted 12-Month Excess Return in purple.

```
gather(pred12, yield_div : pred, key='yeah', value='Variables') %>%
  ggplot() +
  geom_line(aes(x=DATE, y=Variables)) +
  facet_grid(yeah~., scale='free')
```



We also plot predicted 12-Month Excess Return together with default spread, term spread and dividend yield. As we can see, a lower 12-month expected return is usually associated with higher default spread, term spread and dividend yield, vice versa. These patterns actually make sense. Regarding the spreads, during recession, default spread and term spread will usually widen, and stock excess return will become negative. Regarding dividend yield, stock with higher yield for a long time tend to decrease in price or increase slower in the future.

Question 5

AR(1) model is useful because after we get the regression model, we only need observation from current period to predict next period, while regressing returns at different horizons on lagged predictive variables require more previous observations to do the prediction. Besides, shock to stock returns are built into the AR model with ϵ_t , which enables the model to have degrees of uncertainty or randomness built in with real-life interpretation.