36-402 Homework 3

Eu Jing Chua
eujingc

February 4, 2019

#Question 1

Q1 a)

Table 1: Summary of `MAPE`

Var1	Freq
Min.	4.785
1st Qu.	11.708
Median	15.947
Mean	16.554
3rd Qu.	19.959
Max.	44.196
NA's	120.000

There are exactly 120 NAs as the column `Earnings_10MA_back` has exactly 120 NAs too.

Q1 b)

Table 2: Coefficients of linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.138	0.003	46.286	0
MAPE	-0.005	0.000	-26.567	0

Q1 c)

Table 3: MSE of linear model (5-fold CV)

x
0.00187

Question 2

Q2 a)

$$
\begin{align}
Y &= X + \epsilon_t, \text{ where} \tag{1}\\
Y &= R_t \tag{2}\\
X &= \frac{1}{M_t} \tag{3}\\
&\epsilon_t \text{ is the irreducible noise} \tag{4}
\end{align}
$$

In the form of this basic linear regression model, we can see that there is a fixed slope of 1 and fixed intercept of 0.

Q2 b)

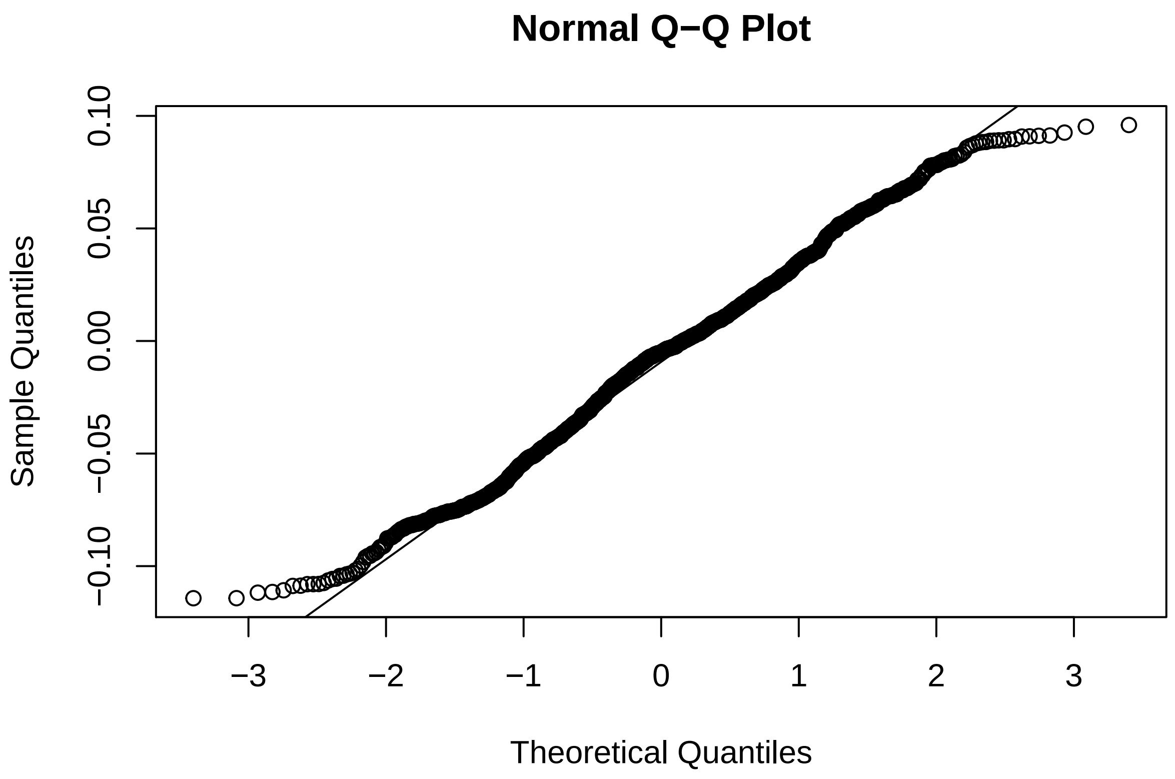
Table 4: In-sample MSE

x
0.0019

Q2 c)

In this model, the slope and intercept are fixed. This means that our fixed parameters are not a function of the finite data sample we have seen. Thus, our model will perform the same regardless of where the data is from, whether in-sample or out-of-sample.

Q2 d)



Q2 e)

The residuals look roughly Gaussian, but it seems that they have thinner tails than what would have been expected if the distribution were Gaussian.

Question 3

Q3 a)

Table 5: Slope of generalized basic model

	$\hat{\beta}_1$
Estimate	0.996
Std. Error	0.037
t value	27.275
Pr(>|t|)	0.000

Q3 b)

Table 6: MSE of generalized basic model (5-fold CV)

x
0.00184

This generalized basic model has a lower estimated MSE than both the first model and the basic model.

Question 4

Q4 a)

Table 7: Coefficients of linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.138	0.003	46.286	0
MAPE	-0.005	0.000	-26.567	0

Since the coefficient of `MAPE` has a p-value very close to 0, it is statistically significant.

Q4 b)

Table 8: Coefficients of generalized basic model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.008	0.003	-2.661	0.008
I(1/MAPE)	0.996	0.037	27.275	0.000

Since the coefficient of `1/MAPE` has a p-value very close to 0, it is statistically significant.

Q4 c)

Table 9: Coefficients of combined linear models

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.058	0.009	6.143	0
MAPE	-0.002	0.000	-7.288	0
I(1/MAPE)	0.591	0.066	8.937	0

Both `MAPE` and `1/MAPE` have coefficients that have p-values very close to 0, so both coefficients are statistically significant.

Q4 d)

Table 10: Coefficients of MAPE, 1/MAPE and MAPEˆ2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.026	0.026	0.976	0.329
MAPE	0.000	0.002	-0.141	0.888
I(1/MAPE)	0.736	0.127	5.801	0.000
I(MAPEˆ2)	0.000	0.000	-1.336	0.182

The only coefficient that is statistically significant in this model is `1/MAPE`, with a p-value very close to 0.

Q4 e)

As we start including more forms of `MAPE` in our linear models, we introduce more correlation between each term in ourgression. This tends to affect the significance of each variable, where higher correlation results in less statistically significant coefficients.

Thus, significance testing is not a viable way of selecting variables for a model as the significance of a single variable in the model is affected by factors such as correlation with other variables, variance of the variable and sample size, all of which have nothing to do with how well the variable can help in predicting the response.

Question 5

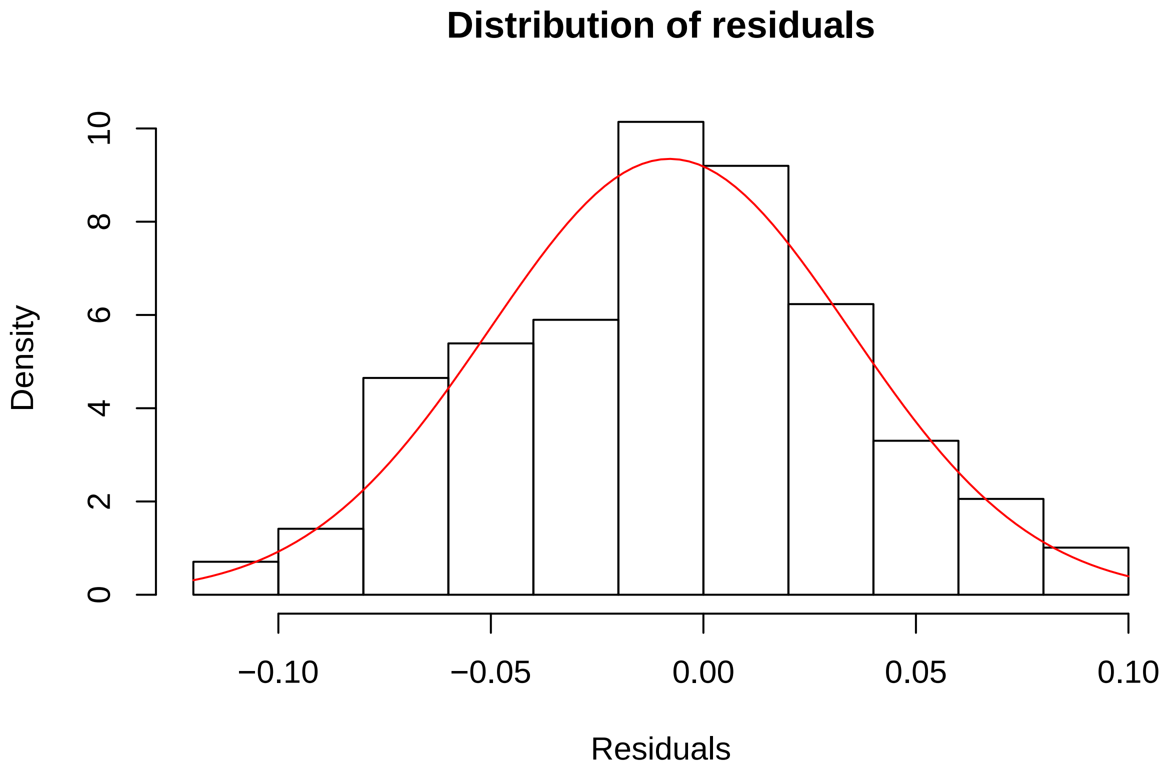
Q5 a)

TODO

Q5 b)

TODO

Q5 c)



Question 6

Q6 a)

```
# Simulates the basic model that R_t = 1/M_t + noise,
# where the noise is t-distributed with params from the input.
# Arguments:
#   MAPE: Vector of M_t
#   t.params: Named vector with m, s, and df representing the mean,
#             sample standard deviation, and degrees of freedom of
#             the t-distribution of the noise
# Returns:
#   Dataframe with a MAPE column and a predicted Return_10_fwd using
#   the basic model above
sim.basic.model <- function(MAPE, t.params) {
    n <- length(MAPE)
    m <- t.params["m"]
    s <- t.params["s"]
    df <- t.params["df"]

    # Generate noise from the input
    noise <- rt(n, df) * s + m

    results <- data.frame(
        MAPE = MAPE,
        Return_10_fwd = 1/MAPE + noise)

    return(results)
}
```

Q6 b)

```
# Runs a linear regression of R_t against 1/M_t, or Return_10_fwd
# against 1/MAPE and returns the slope
# Arguments:
#   data: data frame with Return_10_fwd and MAPE columns
# Returns:
#   slope: Coefficient of the 1/MAPE term in the linear regression
sim.get.slope <- function(data) {
    lm.fit <- lm(Return_10_fwd ~ I(1/MAPE), data = data)
    return(coef(lm.fit)[2])
}
```

Q6 c)

$P(\mid \tilde{\beta}_1 - 1 \mid \geq \mid \hat{\beta}_1 - 1 \mid)$
0.928

Q6 d)

$$H_0 : \beta_1 = 1 \tag{5}$$

$$H_a : \beta_1 \neq 1, \text{ where the test statistic is as follows:} \tag{6}$$

$$t_{score} = \frac{|\tilde{\beta}_1 - 1|}{SE_{\tilde{\beta}_1}} \sim T_{df}, \text{ where } df \text{ is obtained by fitting a t-distribution to the residuals} \tag{7}$$

The p-value of this test is then 0.928, so there is insufficient evidence to reject the null hypothesis. Thus, we conclude at 0.05 significance that the slope is exactly 1.0.

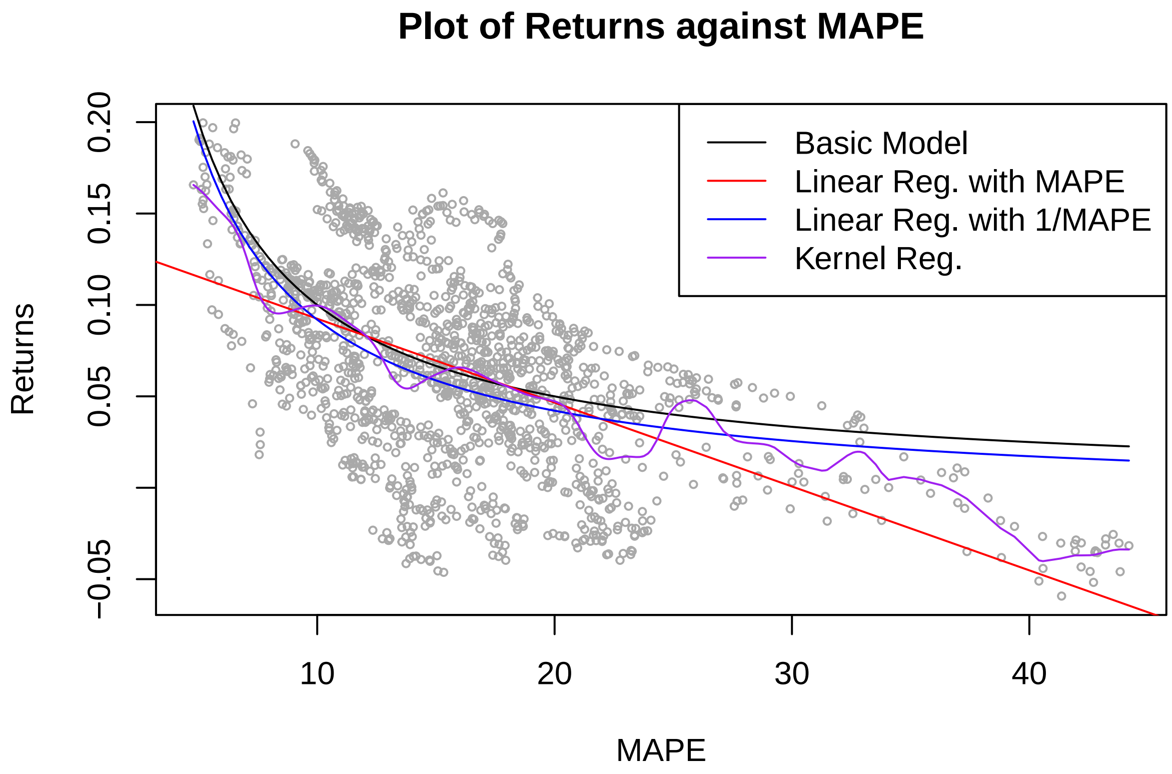
Question 7

Bandwidth
0.58051

CV MSE
0.00169

The kernel regression model has a lower CV MSE, and hence better predictive accuracy, in comparison to the other models considered so far.

Question 8



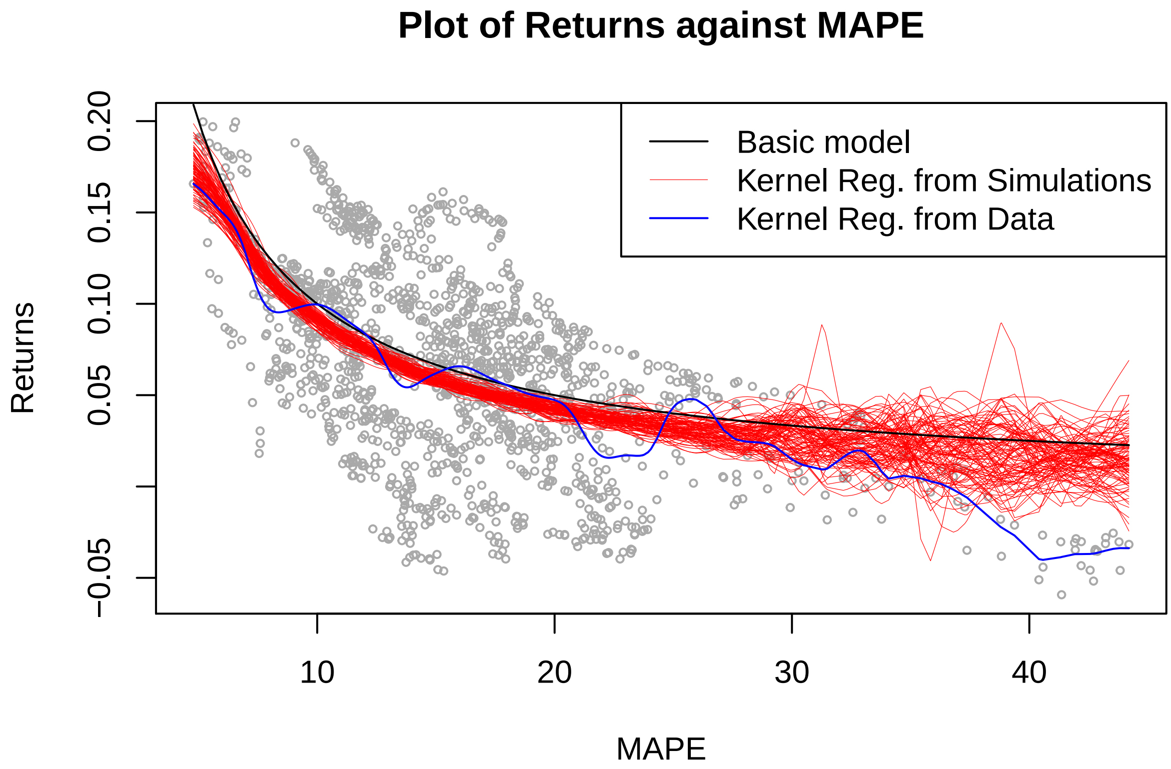
The kernel regression seems to most resemble the functions of `1/MAPE`. It seems to model the overall decaying pattern of the data, but also has finer variations than the `1/MAPE` functions.

Question 9

Q9 a)

```r
# Runs a kernel regression of Return_10_fwd against MAPE and returns the fitted values
# Arguments:
#   data: Data frame with columns Return_10_fwd and MAPE
# Returns:
#   Vector of fitted values from the kernel regression
kernel.smooth <- function(data) {
    npreg.fit <- npreg(Return_10_fwd ~ MAPE, data = data)
    return(fitted(npreg.fit))
}
```

Q9 b)



The kernel regressions from the data and the simulations do differ, in that the ones from the data are all quite smooth for values of `MAPE` below 30, tracking the function of the basic model quite well, before increasing in variability for values above 30. However, the kernel regression for the actual data is not as smooth throughout, and is consistently lower than most of the regressions of simulations for values around above `MAPE` of 35.

Q9 c)

This is a sign that the basic model might not be an accurate model of the underlying process. When running the simulations to produce new data that hopefully looks like the real data, we run the same regressions on both the synthetic data and real data, and then realize that the regressions are not so similar. This implies that the real data does not look like a run of the simulation assuming the basic model, hence the basic model might not be accurate in modelling the actual process.

Question 10

The results from Q3 b), Q6 d) seem to indicate that the expected returns are roughly inversely proportional to MAPE, as the general estimate of $\hat{R}_t = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{M_t}$ has better prediction accuracy than the basic model, from the MSEs of their 5-fold cross-validations. However, the results from Q9 c) also shows that the data generated from the basic model where expected returns are exactly inversely proportional to MAPE is not really similar to the real data collected. Hence, there does seem to be a rough inversely proportional relationship, but it is not exact in nature.