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)

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

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)

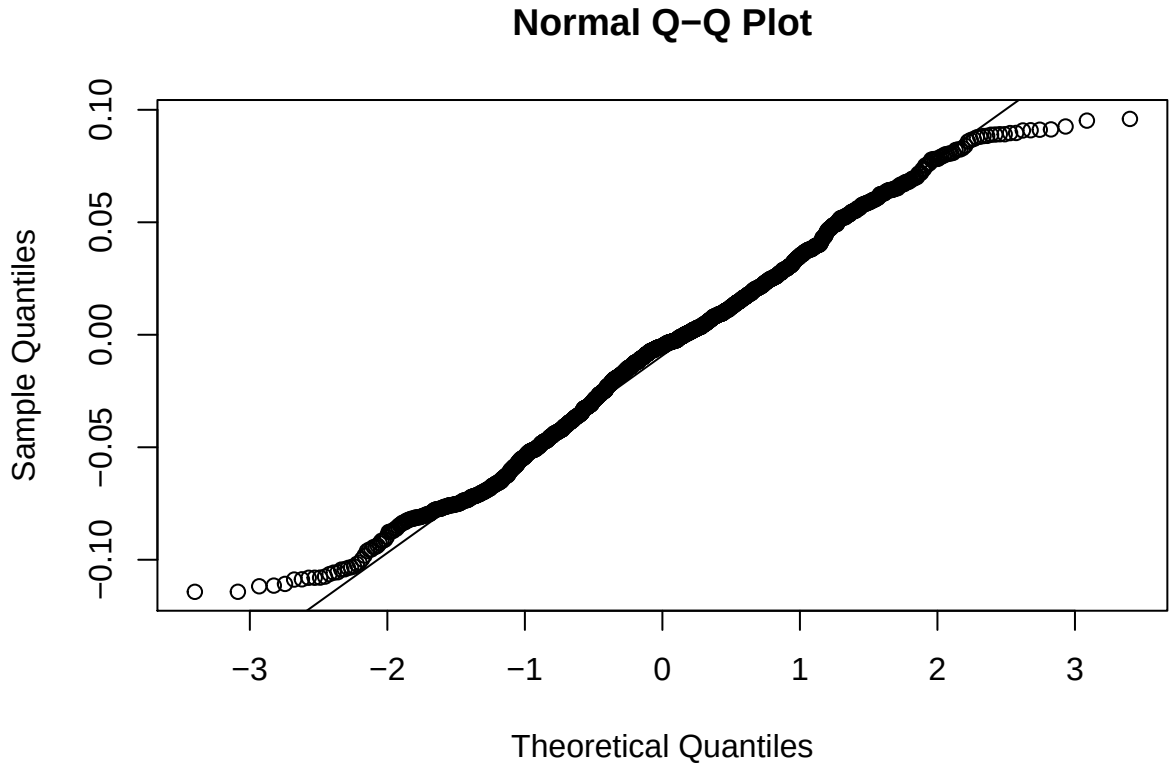
Table 4: In-sample MSE

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
0.00183

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 MAPE^2

	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(MAPE^2)	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

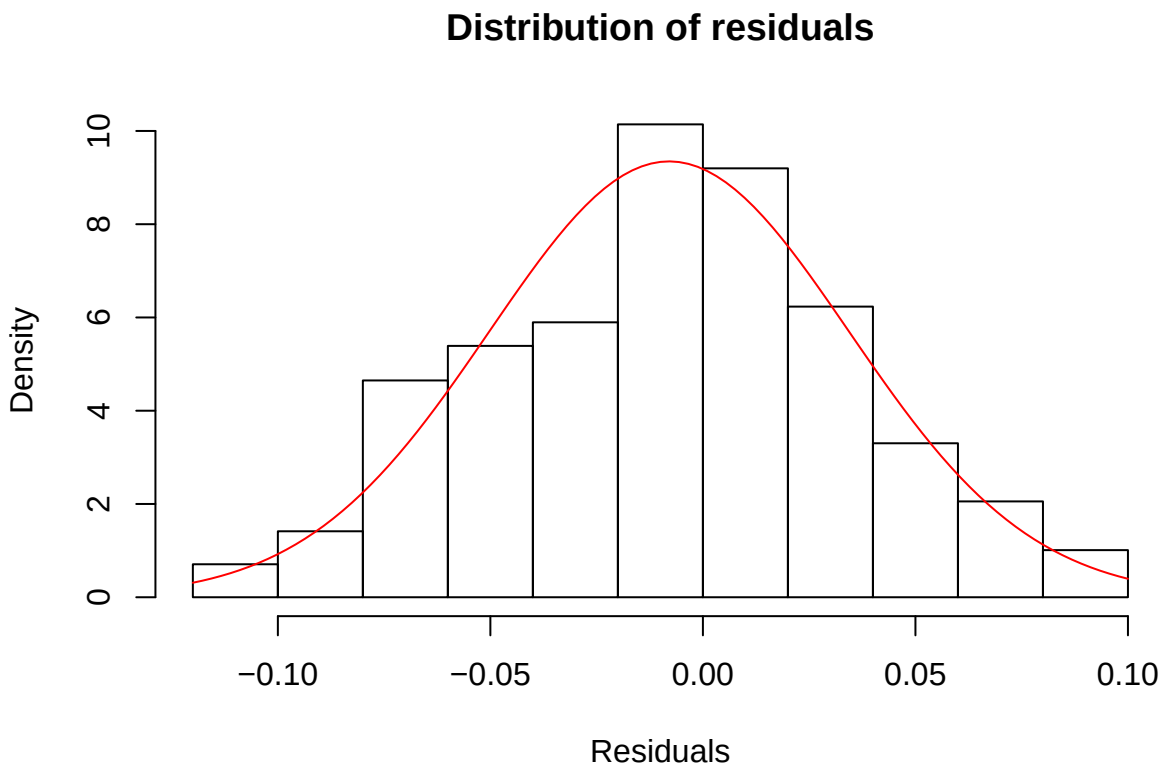
Q5 a)

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)
}
```