Milk and Money

Sakshi Bhargava, Mrunmayee Bhagwat, Ghizlaine Bennani

October 7, 2015

1. Summary

Gerard, a dairy owner in California gets paid a price known as a mailbox price every month for the milk he sells. Since the determination of this price depends on various market conditions and the demand for California dairy products, Gerard has no control over the price. His operating costs have been increasing lately and due to these price fluctuations, Gerard fears that he will incur losses.

To protect himself from these falling prices, he seeks a hedging strategy, a put option. Options on mailbox price are not available on the exchange. However, Class III, Class IV, butter and NFDM milk prices are. We found that out of these, Class III tracks the mailbox price the most closely. We have developed a linear regression model which predicts the Class III commodity price from a mailbox price to study the relation between the two. We concluded that Gerard should buy put option for Class III milk prices.

Put options gain in value as future prices fall. Since mailbox and Class III prices move together, as mailbox price increases, Gerard will incur limited losses from the put option. Whereas, when the mailbox price falls, Gerard can execute the put option in order to cover his production costs. To be ensured that the transaction doesn't lead to losses, the value of put option should be equal to the highest value of Class III price predicted for a given mailbox price. However, he should be prudent while doing so because higher strike price implies high premium costs.

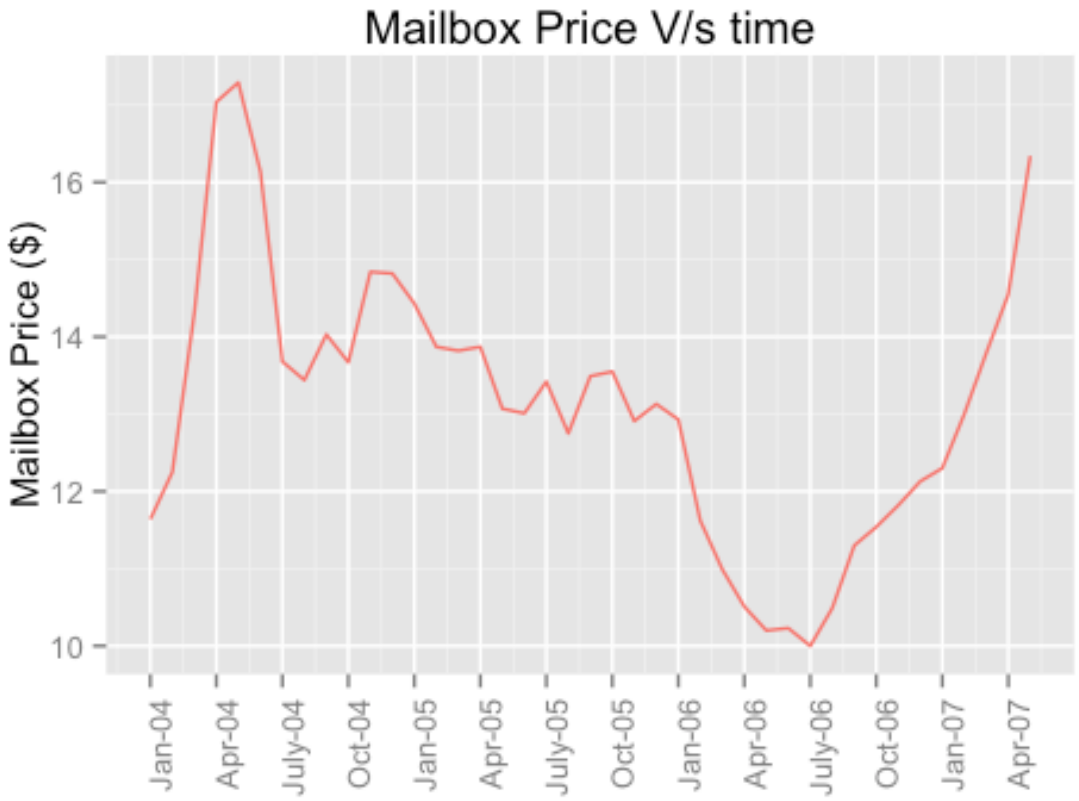
2. Background

In the state of California, the milk value chain starts from procuring the raw milk at dairy farms and shipping it to the processors. The processors then process the same yielding various value added milk products. California processors pay for milk based on its class. All payments are pooled together and paid out equitably to dairy farmers. Milk price in California, called the 'pooled price' is determined every month based on market-wide utilization of the five classes of milk:

- Class I - Fluid Milk
- Class II - Cream, Cottage, Cheese, Yogurt, sterilized products
- Class III - Ice Cream and Frozen Products
- Class IVa - Butter and Dry Milk Products
- Class IVb - Cheese other than cottage cheese

The costs associated with the milk production and manufacturing of dairy products, their demand and the prevailing dairy commodity price lead to fluctuations in tis price. The dairy

owners are paid a ‘mailbox price’ which varies depending on the quality of their milk and some complex adjustments of the ‘pooled price’.



The above graph shows the variation of mailbox price over the period January 2004 to May 2007.

Gerard, a dairy owner in California is unhappy with this milk price variation. The operation costs are increasing and the price uncertainty has started to affect his profits. He is a price taker and cannot control the "mailbox price" he receives for his milk. The falling mailbox prices is a matter of concern for Gerard and he wants to be assured that at least his production costs are covered. On his son’s suggestion, Gerard wants to trade put options on the Chicago Mercantile Exchange to hedge the risk of falling mailbox prices.

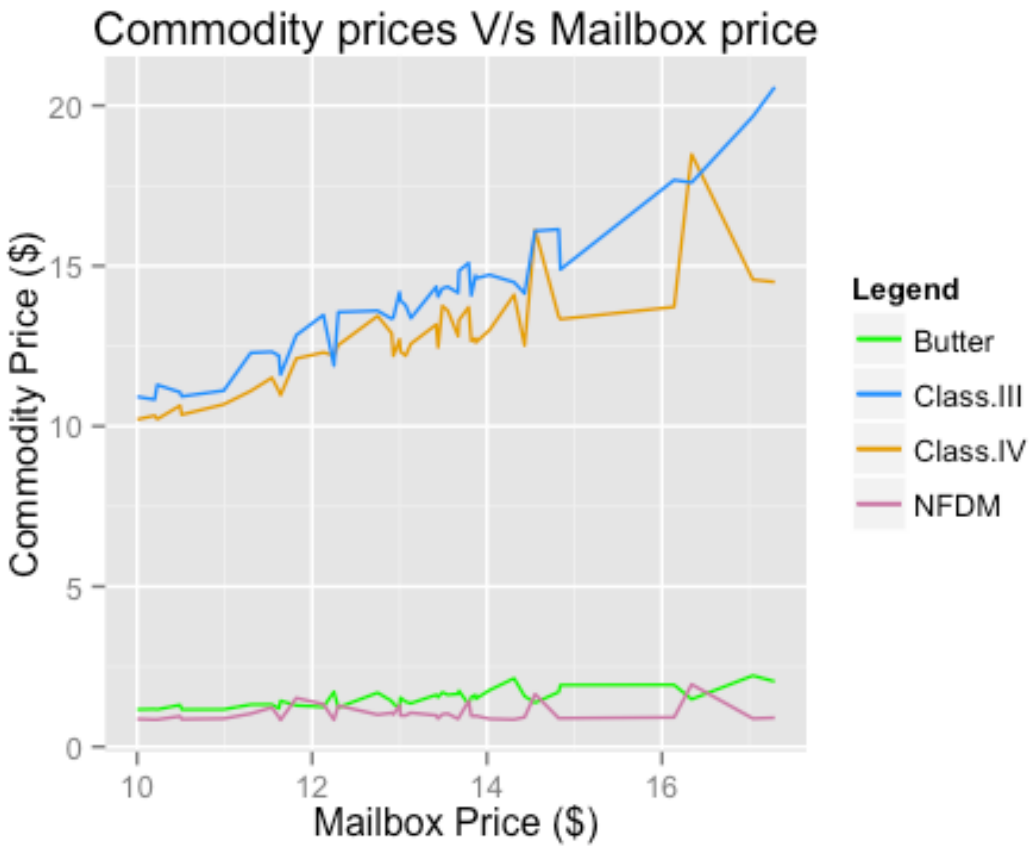
Put option gives the buyer of the contract a right but not an obligation to sell a specific amount of a commodity on a future date at a predetermined price. It is a financial tool which help in mitigating the downward risk and allow for profits. Though Gerard wants to explore how put options can hedge against the volatility of mailbox prices, the Exchange does not trade options on it. The options are available only on Class 3, Class 4, dairy and nonfat dry milk (NFDM). We are thus interested in finding which of these commodity prices behave similarly as the mailbox price to track the associated variations. The intention is to buy a put option on that commodity because it will ensure better predictability of the future milk prices for Gerard to assess the market and then place his bet wisely.

We have data for monthly mailbox and commodity prices of Class 3, Class 4, dairy and NFDM for the period January 2004 to May 2007.

3. Method/Results

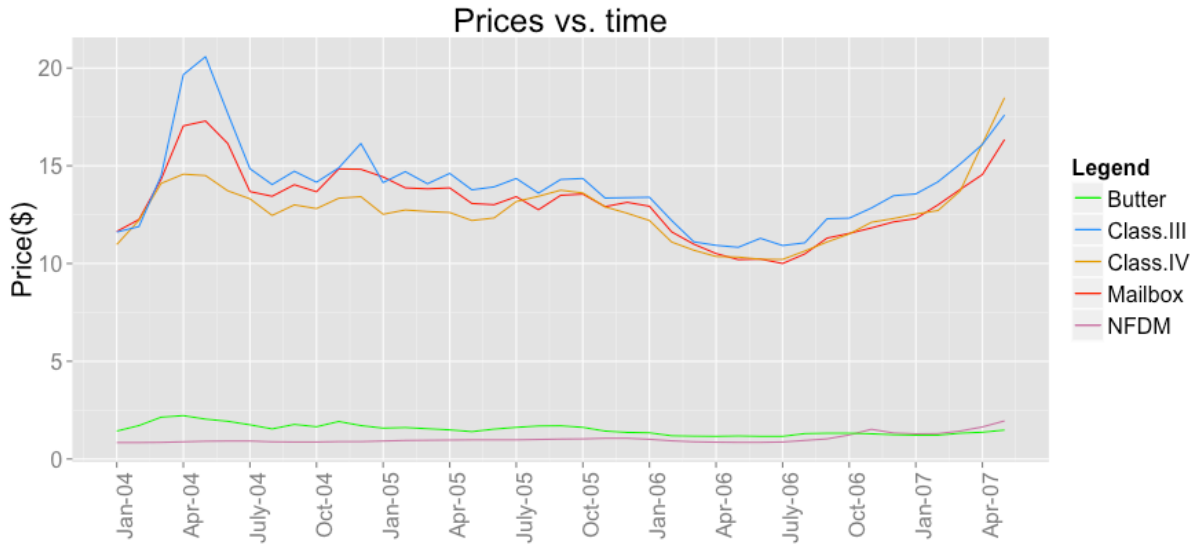
To assess the similitude between commodities into consideration and mailbox price, we compared their behavior over the given time period.

All the dollar values in the analysis are $ per cwt (Costs per hundredweight).



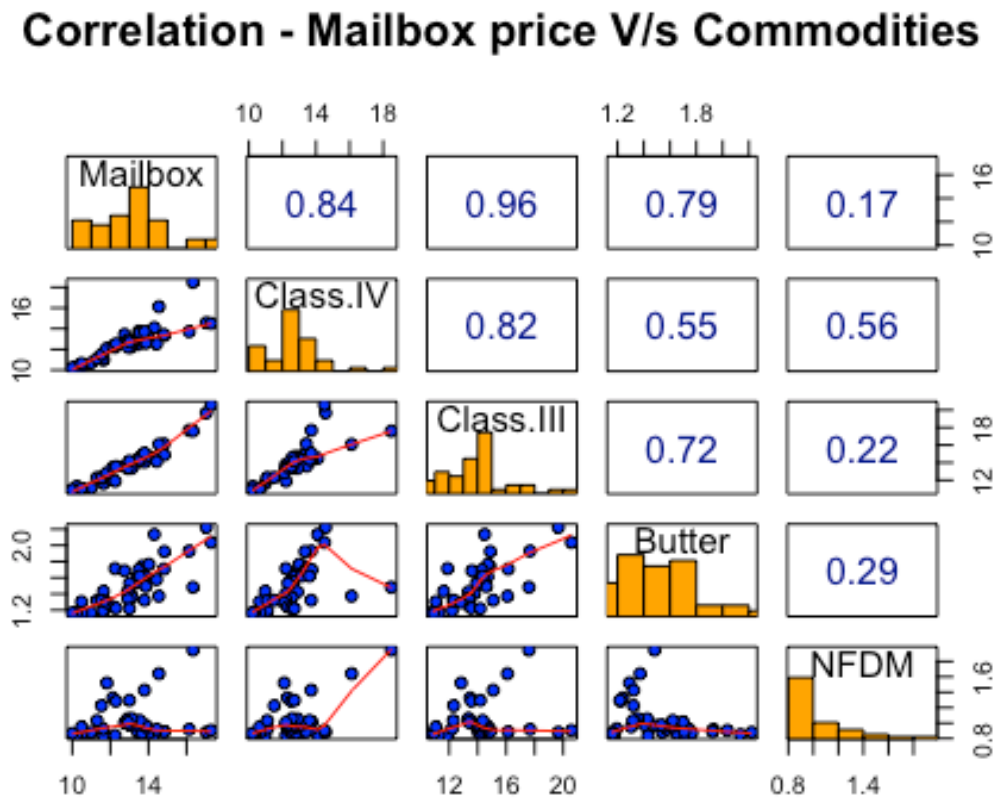
The above graph plots the commodity prices with respect to the mailbox price. As seen from the graph, both Class III and Class IV prices follow the mailbox price quite closely. It can be seen that there is a positive linear relation between them.

For the variation of mailbox price between the given values, the commodities butter and NFDM do not show much variation.



The above graph shows how the commodity price and the mailbox price move together with time.

This was further investigated by seeking statistical evidence and computing correlation between the commodity and the mailbox prices.



It is seen that Class III commodity price has the highest positive correlation with the mailbox price equal to 0.96. The commodities Class IV and butter are also positively correlated with the mailbox price. The commodity NFDM, however, has a very weak positive correlation with the mailbox price.

High correlation coefficient implies strong linear relation between two variables. We also built linear regression models to find which of the Class 3 milk, Class 4 milk, butter and NFDM milk prices are most related to the mailbox price. To come up with the best model that predicts a commodity price given a mailbox price, we had to go through several model iterations. The details of these iteration are given in the Appendix. The final models chosen are as given in the following table:

Commodity	**Linear Regression Model**	**Model Name**
Class III	Class III = 12.58526 − (0.96945 ∗ mailbox) + (0.08040 ∗ (mailbox ^ 2))	A
Class IV	Class IV = (0.03398 − (0.003416 ∗ mailbox) + (9.903e − 05 ∗ (mailbox ^ 2))) ^ (−1/2)	B
butter	butter = (1.949155 − (0.141375∗ mailbox) + 0.003159 * (mailbox ^ 2)) ^ (−1 / 1.152)	C
NFDM	For NFDM, the model iterations can be seen in the Appendix. However, none of the models we tested can be used to predict NFDM price. The models violate one or more assumptions of a linear regression model and its failure can be attributed to endogenity.	-

Using these models, following are the predictions of the commodity prices for Class III, Class IV, butter and NFDM products, for a given mailbox price of $12.50 per cwt:

Mailbox price ($)	Class III ($)	Class IV ($)	butter ($)	NFDM ($)
12.5	13.03	12.17	1.41	None

All the chosen final models pass homoscedasticity and normality tests. Transformations are used on the predictor, mailbox price to explain as much variation as can be explained in the commodity prices.

In addition to the mailbox price, we also introduced a qualitative variable for the various quarters to check if seasonality affects the commodity prices. However, the hypothesis could not be proven and we had to drop that model.

On comparing models A, B and C, we found that Model A was the best fit one. The Model A gives the highest R-squared value of .951 and an adjusted R-squared value of .9484. Thus, mailbox price can explain 95.1% variation in the price of commodity. Thus, we again reach to conclusion that commodity prices are most closely related to Gerard's mailbox prices.

Besides the fact that Class III milk prices is most closely related with mailbox price, with Model A we are able to eliminate the unpredictability in Class III prices or in other words have been successful in explaining the variation in it to the maximum (because of highest R squared seen) compared to other commodity prices. These reasons substantiate that mailbox price provides the estimation of commodity price the best among other commodities. Hence, becomes the obvious choice to buy put option on and Gerard must track this price in order to protect himself from the falling mailbox prices.

Thus, for further analysis, the following linear regression model is used.

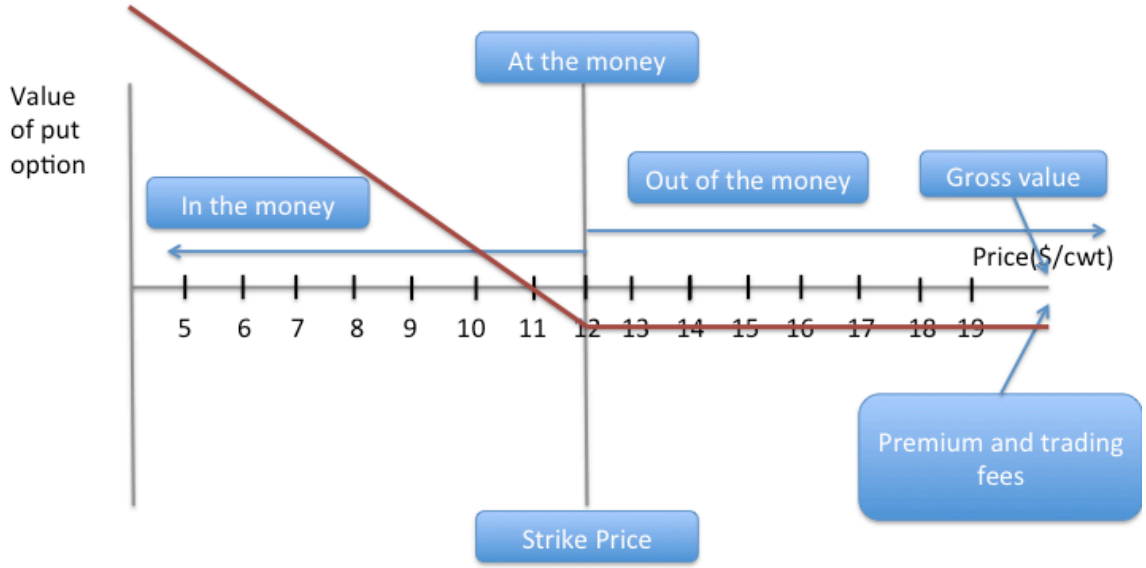
Model A:

$$\text{ClassIII} = 12.58526 - (0.96945 * \text{mailbox}) + (0.08040 * (\text{mailbox}^2))$$

Model A is explained in detail in the Appendix.

We use this model to estimate the ClassIII commodity price from a given mailbox price and explore how a put option can prevent Gerard from incurring losses due to falling milk prices.

What is a put option? How can a put option help Gerard achieve price stability from the falling mailbox prices?



Dairy futures contract is a risk management tool which guarantees the milk producers a fixed price for their milk and dairy products in the future. It insures them against the price decreases in the market. A put option gives the buyer of the contract a right but not an obligation to sell a specific amount of a commodity on a future date at a predetermined price. The predetermined price is known as the 'strike price'. Put option comes at a price. The buyer has to pay a premium on the put option and the trading fees which includes processing fees, commission, etc.
Gerard should use put options so that he is at least ensured that his production costs for a given month are covered. On the expiration date of the put option, there will be one of the following scenarios as shown in the above figure:

('Mailbox price' refers to the mailbox price in the put option expiration month + basis from previous month which Gerard will receive from the processor. 'Strike price' is after the effect of premium and trading fees. For further analysis, we have considered zero basis.)

1. **Strike price > Mailbox price**

The put option is said to be in the money. Gerard can execute the put option and be assured that his costs are covered.

2. **Strike price < Mailbox price**

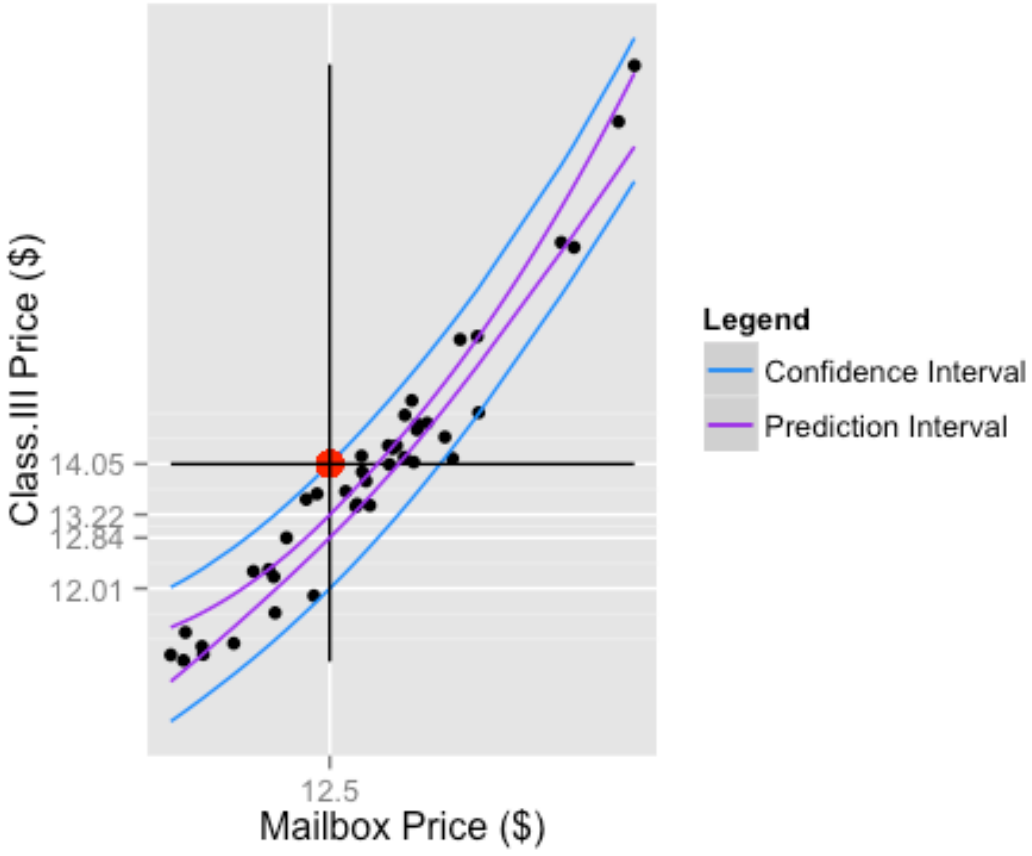
The put option is said to be out of the money. Executing the put option is worthless. As long as the difference between the mailbox price and Gerard's production costs is higher than the premium plus trading fees he paid to buy the put option, Gerard doesn't lose from the transaction.

3. Strike price = Mailbox price

The put option is said to be at the money. Gerard can either execute or relinquish the put option. As long as the mailbox price less Gerard's costs is equal or above the put option's premium plus trading fees, Gerard will not incur loss in this transaction.

Gerard wants to be ensured that he receives a price protection if the mailbox price falls below $12.50. We use Model A to find a point estimate or mean value for the put option when the mailbox price is equal to $12.50. It comes out to be $13.03 excluding the premium and trading fees. However, it cannot be assumed that the Class III commodity price will always be exactly $13.03. Actually 95% of the times, the price will lie in the range $12.01 and $14.05.

The above range is calculated using the prediction interval for the Class III price. For a given mailbox price, the Class III commodity's price is estimated from its distribution. The point estimate or the mean of the distribution of Class III commodity price is the most probable estimated value. The confidence interval gives the range in which these mean values for multiple samples can lie for a given mailbox price. However, when the model is estimating the price of Class III commodity, it may not be a mean value and thus, its range spans over the prediction interval. The prediction interval is larger than the confidence interval because it represents the distribution of any value pulled from the distribution of Class III prices, not just a mean value.



The graph shows the confidence and prediction interval for the Class III commodity price when the mailbox price is equal to $12.5. The intervals are built with a confidence level of 95%. The red point indicates the computed strike price for the Class III put option given a mailbox price of $12.5.

If Gerard's put option is in the money (strike price > mailbox price), the deal will be a profitable one for him. Considering the worst case scenario that the commodity price on the day the option expires will be the highest i.e. $14.05 (upper band of the prediction interval), Gerard must buy a put option for at least $14.05 to ensure profit.

The exchange trades put options at $0.25 intervals for values $13.75, $14.0, $14.25, $14.50, etc. It is recommended that Gerard buys a put option for $14.25. If he goes for anything below $14.05, it cannot be said with a confidence level of 95% that the option will be in the money. On the other hand, a put option of $14.25 increases this confidence interval by a small percent (as $14.25 > $14.05) and any option above this price is further going to reduce Gerard's risk of losses incurred due to falling mailbox price and increase the chances of him making a profit out of the deal, however, at the cost of put option premium and trading fees (not considered in this case).

We now address and show how Gerard can achieve price stability from buying a put option.

Let us assume that Gerard buys a put option worth $14.25 to protect himself if the mailbox price falls below $12.5. If the mailbox price falls below $12.50 to $11.50, for $11.50, **point estimate for the Class III commodity price from Model A is $12.07.**

Thus, without considering the premium and transaction cost on the put option, Gerard can obtain a profit equal to strike price - mean predicted Class III price for mailbox price of $11.50. i.e. 14.25 - 12.07 = $2.18. The mean price is indicative of the future market price of Class III commodity.

From the Model A, without considering the premium and transaction cost on the put option, if Gerard executes the put option, the 95% range of net value is calculated as follows:

	Lower bound ($)	Upper bound ($)
Put option price	14.25	14.25
Predicted Class III price	(11.04)	(13.10)
Payoff from the option (In the money put option)	3.21	1.15
Mailbox price	11.5	11.5
Net value	14.71	12.65

It can be seen that, since both the net values $12.65 and $14.71 are above $12.5, even if the mailbox price falls below $12.5 to $11.5, Gerard doesn't lose anything. There is a 95% chance that with the execution of the put option, the net selling price for him would be in the range of $12.65 and $14.71. It means that for a mailbox price of $11.5, a put option for $14.25 is in

the money 95% of the times or Gerard can be 95% assured that his net price would exceed $12.5.

Though premium and trading fees on the put option are ignored till now, they are important to determine the net value of a put option.

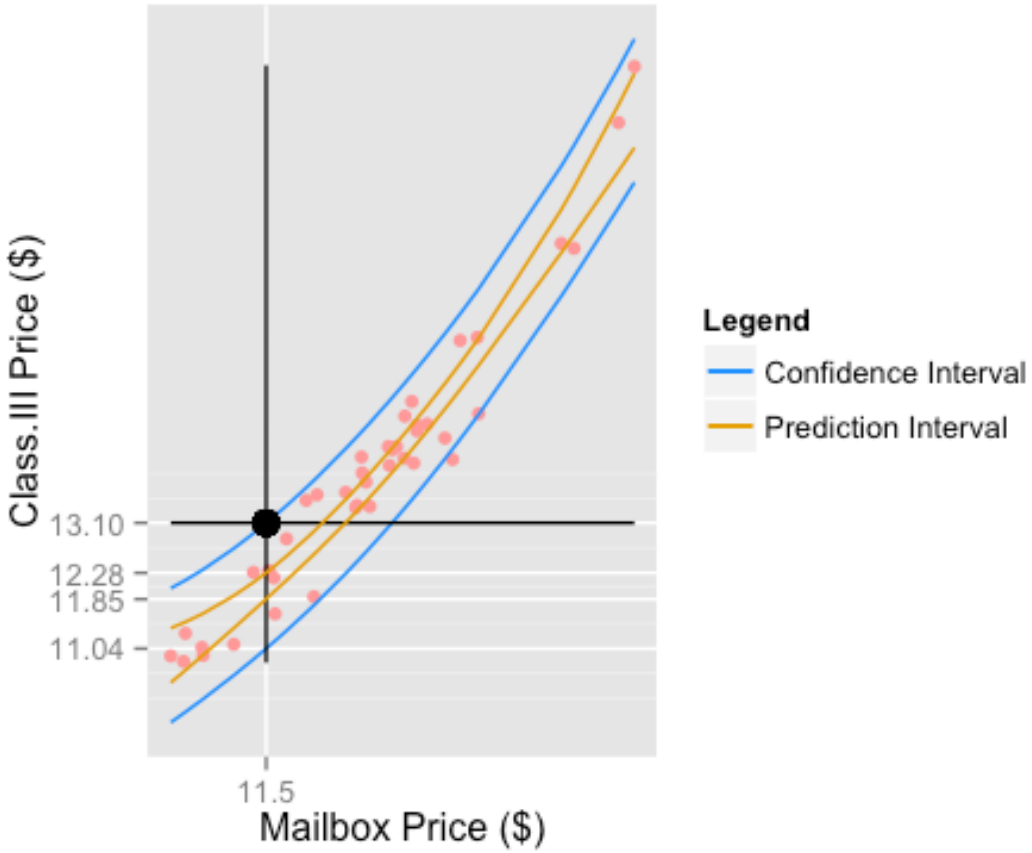
Gerard has to pay premium and trading fees which includes the processing fees, commission, etc. on the put options.

In general, the amount of a premium is a function of:
• The amount of time until the option expires
The longer the time until an option expires, the higher the premium. This reflects the fact that the option owner has more flexibility in exercising the option compared to one which is closer to expiration. There is a greater probability of the option being in the-money some time in the future.

• The strike price in relation to the current futures price
The difference between the strike price and the associated futures contracts current settle price is the option's intrinsic value. A higher intrinsic value translates into a higher premium. For a put option, the higher the futures price in relation to a strike price, the lower the intrinsic value. This implies a lower premium.
• The volatility of the underlying futures contract

If a market is more prone to sudden sharp price changes, it also has a greater chance of coming in-the-money than if prices are not volatile. The option's premium is equal to its intrinsic value plus any time (volatility) value.

Net value of put option = Cost of put option – Premium - Trading fees



In this case, the premium is assumed to be $0.40 and the processing fees $0.05. Considering the above analysis, now with the premium and processing fees, for the point estimate, Gerard can obtain a profit of 14.25 – 12.07 – (0.40 + 0.05) i.e. $1.73 for mailbox price equal to $11.5

Considering that 95% interval for the commodity price,

	Lower bound ($)	Upper bound ($)
Put option price	14.25	14.25
Premium	(0.40)	(0.40)
Trading fees	(0.05)	(0.05)
	13.80	13.80
Predicted Class III price	(11.04)	(13.10)
Payoff from the option (In the money put option)	2.76	0.70

Mailbox price	11.50	11.50
Net value	14.26	12.20

The upper bound net value $12.20 is less than $12.50, and thus renders the put option useless. Gerard loses the premium and trading fees in this transaction i.e. $0.45. The put option is out of money when the value of Class III commodity rises to 13.10. Thus, to effect for the premium and trading fees Gerard should go for a higher value put option.

It can be seen that the next available price $14.5 also leads the put option worthless for the upper bound of the interval of Class III commodity price. Thus, Gerard must buy an option valued at $14.75. Since the strike price now is higher, the premium will increase. It is assumed to be $0.44 now while the trading fees remain $0.05.

Considering that 95% interval for the commodity price,

	Lower bound ($)	Upper bound ($)
Put option price	14.75	14.75
Premium	(0.44)	(0.44)
Trading fees	(0.05)	(0.05)
	14.26	14.26
Predicted Class III price	(11.04)	(13.10)
Payoff from the option (In the money put option)	3.22	1.16
Mailbox price	11.50	11.50
Net value	14.72	12.66

It can be seen that the net value for the 95% range of commodity price is above $12.50. Thus, with a put option of $14.75, Gerard can be 95% sure that he can well manage the risk of mailbox price falling below $12.50.

Highest hedged profits can be achieved at higher values of put option, however, the premium and trading fees must be taken into account.

4. Conclusion/ Recommendations

As seen before, put option can help Gerard achieve price stability. However, there are some points to be noted.

1. Model A is based on the monthly mailbox and Class III commodity prices available during the period January 2004 and May 2007. When data for future months is available, it is important to reconstruct the model taking into consideration the new data to ensure that Class III commodity still tracks the mailbox price.
2. Though put options are a low maintenance risk management instruments, Gerard must not ignore the fact that they come with a price. In order to achieve mailbox price stability, he should not end up paying a lot as premium and incur losses if the put option is out of money.
3. In the future, the available put option price may change and the analysis will change accordingly.
4. Similarly, if the production costs change for Gerard, the analysis will have to be done with new values.
5. Given basis, productions costs, trading costs and commissions a recommendation engine can be built to assist Gerard in hedging his risks. The price forecasting model can include other potential factors affecting milk prices such as climate data, make allowance for various dairy commodities, economic factors etc.

5. Appendix

To predict the prices of the commodities Class III, Class IV, butter and NFDM, we built various models and tested them for various diagnostics. The best models chosen for each commodity met the following criteria:

1. Behavior of the residuals align with the basic assumptions of linear regression.

We performed the following tests on the regression model

	Purpose	Decision (Significance level 0.05)
Shapiro-Wilk normality test	To test the normality of the residuals	p-value > 0.025, the residuals are normally distributed.
Non-constant Variance Score Test	To test the homoscedasticity of the residuals	p-value > 0.025, the residuals are homoscedastic i.e. they have a constant variance.
R-squared and adjusted R-squared	To explain the variation in the response variable when the predictor varies.	High value of R-squared and adjusted R-squared means a strong relation between the response variable and the

		predictors. is expected to be as high as possible.
Coefficient estimates	To test the significance of coefficients	p-value < 0.025 which means that the coefficient is not statistically significantly different from zero and thus can be included in the model.

2. Maximum possible variation in the response variable is explained by the predictor mailbox price or by its relevant transformations.

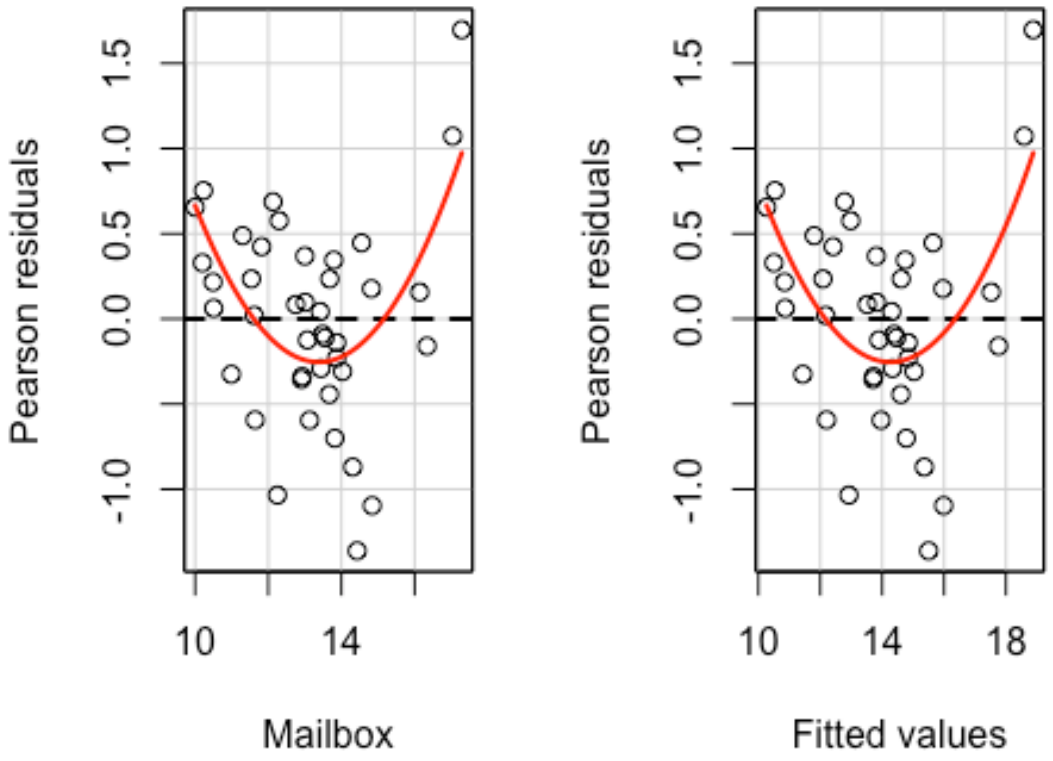
Following are the model iterations for different commodities:

Class III

Model 1

The first model includes regressing Class III commodity prices on the mailbox price in raw untransformed form.

```
lm_cls3 <- lm(Class.III ~ Mailbox, data = mnm)


Call:
lm(formula = Class.III ~ Mailbox, data = mnm)

Residuals:
     Min       1Q   Median       3Q      Max
-1.36177 -0.32504  0.04224  0.34483  1.69718

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.55720    0.70445  -2.211    0.033 *
Mailbox      1.18219    0.05319  22.226   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5989 on 39 degrees of freedom
Multiple R-squared:  0.9268,    Adjusted R-squared:  0.925
F-statistic:   494 on 1 and 39 DF,  p-value: < 2.2e-16
```




```
           Test stat Pr(>|t|)
Mailbox        4.329        0
Tukey test     4.329        0


    Shapiro-Wilk normality test

data:  lm_cls3$residuals
W = 0.98341, p-value = 0.8022
```

It can be seen from the summary that the coefficients are not statistically significantly different from zero. It is also observed that the residuals are normally distributed. However, the residuals plot suggests that the residuals spread out as the mailbox price increases. Since the relation between ClassIII price and mailbox price is positive, we can say that the error variance is larger for higher mailbox price. Similar can be inferred from the plot of fitted Class III values. As we have shown that the error terms are normally distributed, we have performed ncvTest to statistically prove the visible heteroscedasticity of the error terms.

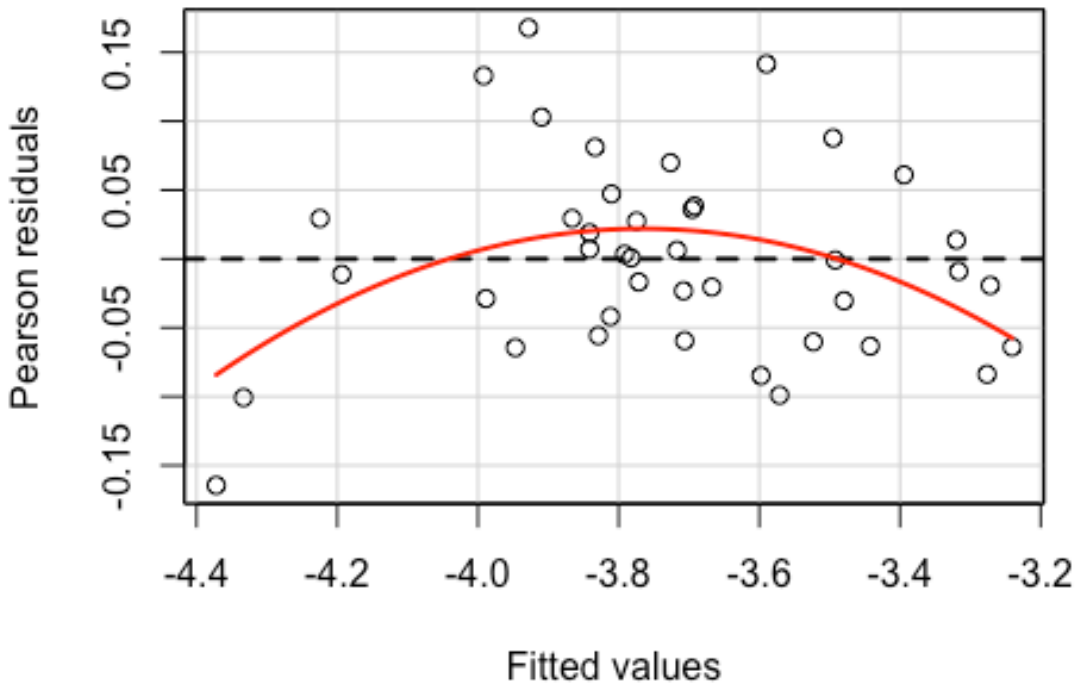
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 9.102608    Df = 1     p = 0.002552453
```

With a p-value of 0.002(< 0.025), we reject the null hypothesis and conclude that heteroscedasticity exists.

Model 2

To get rid of heteroscedasticity, we then transformed the response variable i.e. the Class III commodity. The power of transformation(λ) as determined by the box-cox transformation is -0.424. Hence to improve the model, negative square-root transformation was performed on Class III price(Y).

```
lm_cls3 <- lm(-sqrt(Class.III) ~ Mailbox, data = mnm)


Call:
lm(formula = -sqrt(Class.III) ~ Mailbox, data = mnm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16451 -0.05581 -0.00088  0.03606  0.16779

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.688444   0.083524  -20.21   <2e-16 ***
Mailbox     -0.155209   0.006306  -24.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.071 on 39 degrees of freedom
Multiple R-squared:  0.9395,    Adjusted R-squared:  0.938
F-statistic: 605.7 on 1 and 39 DF,  p-value: < 2.2e-16
```



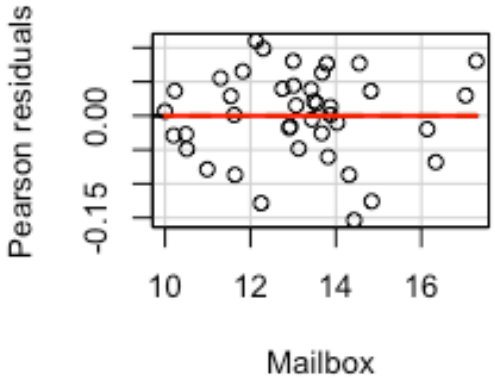
After transformation, R^2 and R^2_{adj} improved to 0.9395 and 0.938 respectively. However, the behavior of the residuals as seen in the residuals plot still shows very high variance. The p-value of 0.007 from the Lack of fit test indicates that we can include a quadratic term of mailbox price to improve the model further.

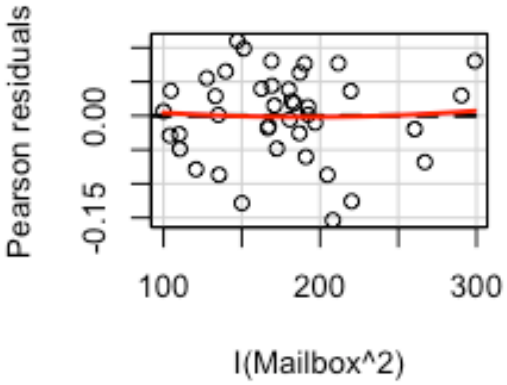
Model 3

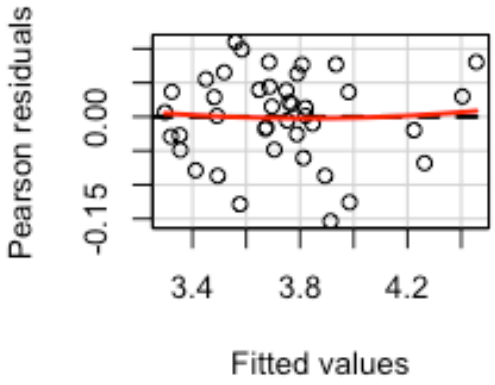
On transforming the mailbox price and the response variable and following the heirarchial principal, we get the equation as follows:

```
lm_cls3 <- lm(sqrt(Class.III) ~ Mailbox + I(Mailbox^2), data = mnm)


Call:
lm(formula = sqrt(Class.III) ~ Mailbox + I(Mailbox^2), data = mnm)

Residuals:
      Min        1Q    Median        3Q       Max
-0.153493 -0.029073  0.006495  0.039620  0.110069

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.910530   0.436587   6.667 6.97e-08 ***
Mailbox      -0.030720   0.065643  -0.468  0.64247
I(Mailbox^2)  0.006947   0.002443   2.844  0.00714 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.06532 on 38 degrees of freedom
Multiple R-squared:  0.9501,    Adjusted R-squared:  0.9475
F-statistic: 361.9 on 2 and 38 DF,  p-value: < 2.2e-16
```







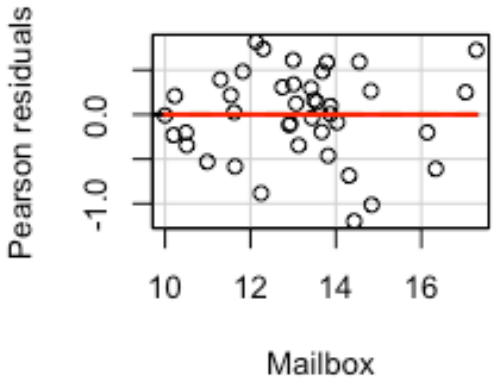

```
             Test stat Pr(>|t|)
Mailbox         -1.376    0.177
I(Mailbox^2)     1.958    0.058
Tukey test       1.970    0.049


    Shapiro-Wilk normality test

data:  lm_cls3$residuals
W = 0.97123, p-value = 0.3779

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.4323654    Df = 1     p = 0.5108305
```

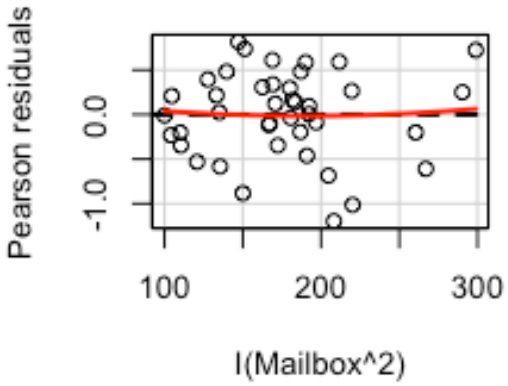
We now have statistically significant model that passes all tests. The p-value for the coefficient of mailbox goes high because of multicollinearity playing its role between mailbox and square of mailbox price. However, noting the exceptionally high p-value of mailbox, the variable may simply have too much noise and might do a poor job in explaining the variation in the ClassIII price.

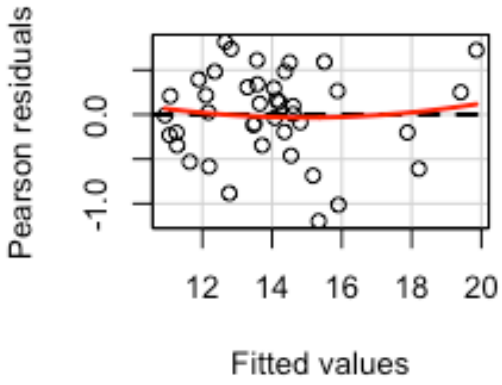
Final Model – Model A

On transforming mailbox price and following the heirarchial principal, we get the equation as:

```
lm_cls3 <- lm(Class.III ~ Mailbox + I(Mailbox^2), data = mnm)


Call:
lm(formula = Class.III ~ Mailbox + I(Mailbox^2), data = mnm)

Residuals:
     Min       1Q   Median       3Q      Max
-1.19635 -0.23117  0.02446  0.30601  0.81509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.58526    3.31849   3.792 0.000520 ***
Mailbox      -0.96945    0.49895  -1.943 0.059454 .
I(Mailbox^2)  0.08040    0.01857   4.329 0.000105 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4965 on 38 degrees of freedom
Multiple R-squared:  0.951, Adjusted R-squared:  0.9484
F-statistic: 368.7 on 2 and 38 DF,  p-value: < 2.2e-16
```







```
             Test stat Pr(>|t|)
Mailbox         -1.434    0.160
I(Mailbox^2)     2.337    0.025
Tukey test       2.393    0.017


    Shapiro-Wilk normality test

data:  lm_cls3$residuals
W = 0.97432, p-value = 0.4715

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.51042    Df = 1     p = 0.219075
```

We now have statistically significant model that passes all initial diagnostic tests. The p-value for mailbox is acceptable(although multicollinearity is still playing a role) and the R^2 is almost the same. Above all, this model is a **parsimonious version** of the previous model.

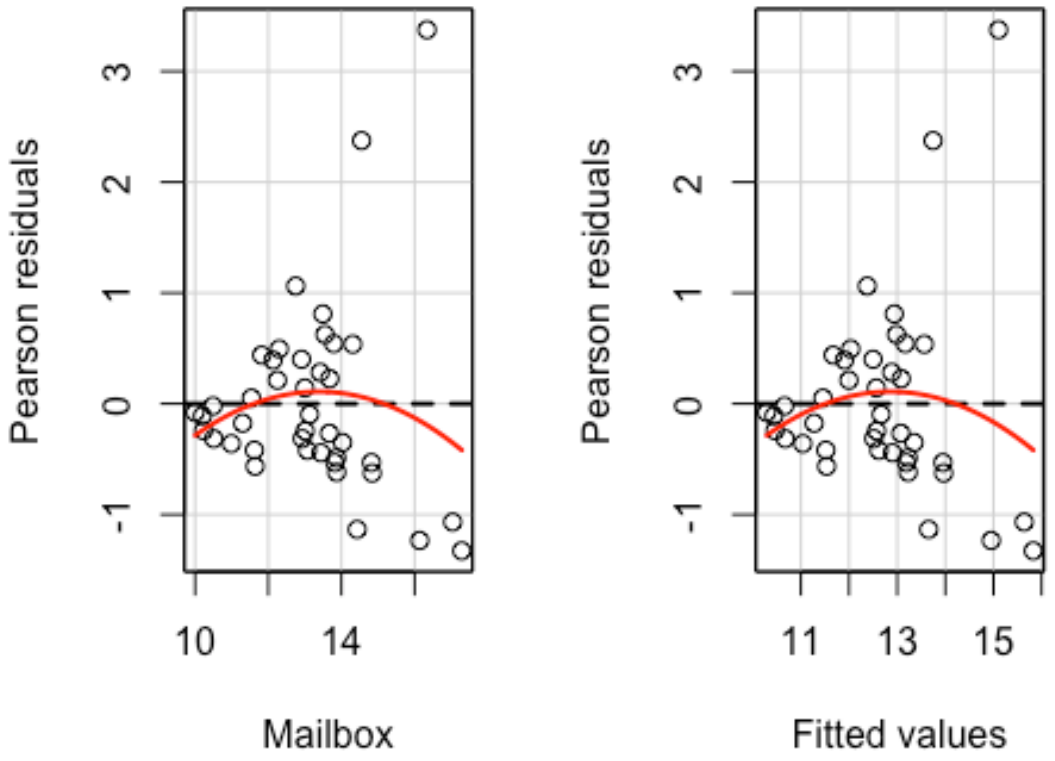
We choose this as the best among the models tested to predict the Class III commodity price from the mailbox price.

Class IV

Model 1

The first model includes regressing Class IV prices on the mailbox price in raw untransformed form.

```
lm_cls4 <- lm(Class.IV ~ Mailbox, data = mnm)


Call:
lm(formula = Class.IV ~ Mailbox, data = mnm)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3272 -0.4417 -0.1756  0.3937  3.3747

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.68914    1.02458   2.625   0.0123 *
Mailbox      0.75986    0.07736   9.822 4.24e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.871 on 39 degrees of freedom
Multiple R-squared:  0.7121,    Adjusted R-squared:  0.7048
F-statistic: 96.48 on 1 and 39 DF,  p-value: 4.24e-12
```




```
           Test stat Pr(>|t|)
Mailbox       -1.061    0.295
Tukey test    -1.061    0.289


    Shapiro-Wilk normality test

data:  lm_cls4$residuals
W = 0.8413, p-value = 4.611e-05
```

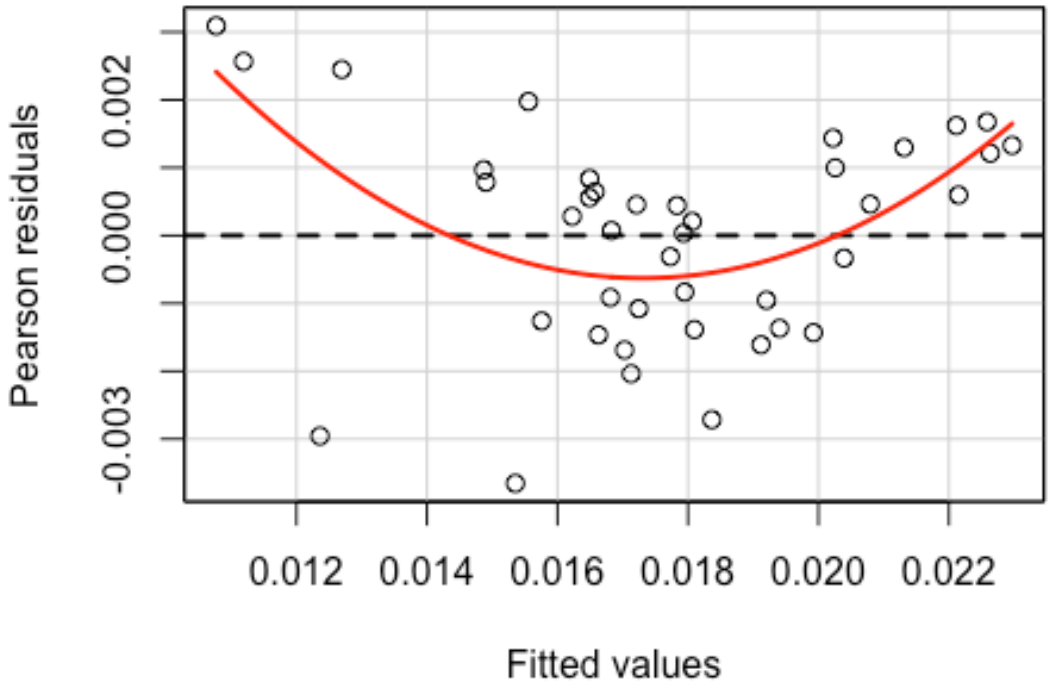
One of the coefficients of this model is significant. However, the residuals plot suggests heteroscedasticity and dominant outliers. As seen from Shapiro Wilk test, the residuals are not normal, which violates one of the assumption of linear regression model. Thus, this model cannot be selected.

Model 2

To deal with non-normality, we transform the Class IV commodity price. The power of transformation(λ) is determined by box-cox transformation and is equal to -1.6.

```
lm_cls4 <- lm(I(Class.IV ^ -1.6) ~ Mailbox, data = mnm)


Call:
lm(formula = I(Class.IV^-1.6) ~ Mailbox, data = mnm)

Residuals:
```

```
       Min         1Q     Median         3Q        Max
-0.0036545 -0.0012609  0.0002824  0.0009980  0.0030910

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0397017  0.0018457   21.51  < 2e-16 ***
Mailbox     -0.0016733  0.0001394  -12.01 1.13e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001569 on 39 degrees of freedom
Multiple R-squared:  0.7871,    Adjusted R-squared:  0.7816
F-statistic: 144.2 on 1 and 39 DF,  p-value: 1.13e-14
```



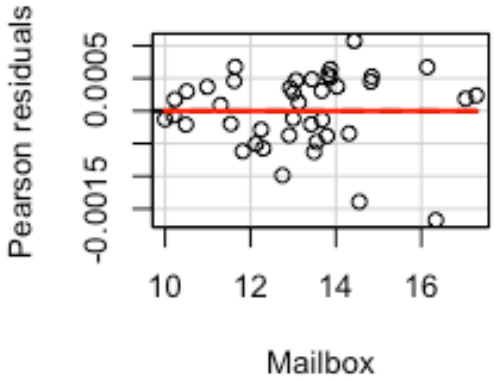
After transformation, R^2 and R^2_{adj} improve to 0.7871 and 0.7816 respectively. However, the behavior of the residuals as seen in the residuals plot still shows variance. The p-value from the significance of Lack of fit test indicates that we can include a quadratic term of Class IV to further improve the model.

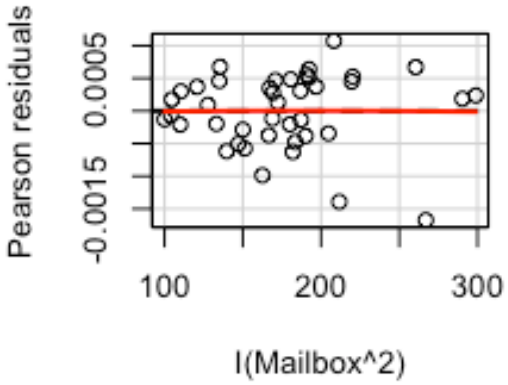
Final Model – Model B

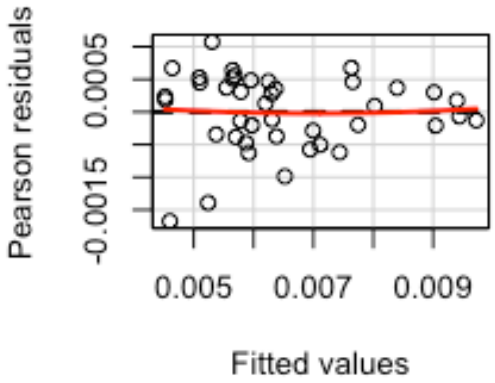
On transforming the mailbox price as well as the response variable and following the hierarchical principal, we get the regression model as follows:

```
lm_cls4 <- lm(I(Class.IV ^ -2) ~ Mailbox + I(Mailbox ^ 2), data = mnm)
```

```
Call:
lm(formula = I(Class.IV^-2) ~ Mailbox + I(Mailbox^2), data = mnm)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0016734 -0.0003449  0.0001295  0.0004488  0.0010731

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.398e-02  3.904e-03   8.705 1.39e-10 ***
Mailbox      -3.416e-03  5.869e-04  -5.821 1.00e-06 ***
I(Mailbox^2)  9.903e-05  2.184e-05   4.534 5.63e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.000584 on 38 degrees of freedom
Multiple R-squared:  0.8626,    Adjusted R-squared:  0.8554
F-statistic: 119.3 on 2 and 38 DF,  p-value: < 2.2e-16
```








```
             Test stat Pr(>|t|)
Mailbox         -0.945    0.351
I(Mailbox^2)    -0.371    0.713
Tukey test       0.441    0.660


    Shapiro-Wilk normality test
```

```
data:  lm_cls4$residuals
W = 0.94345, p-value = 0.04134

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 6.823706    Df = 1     p = 0.008995571
```

We now have a statistically significant model that passes all diagnostic tests. The amount of variation in the Class IV commodity price as explained by the regressors, R^2 improved drastically.

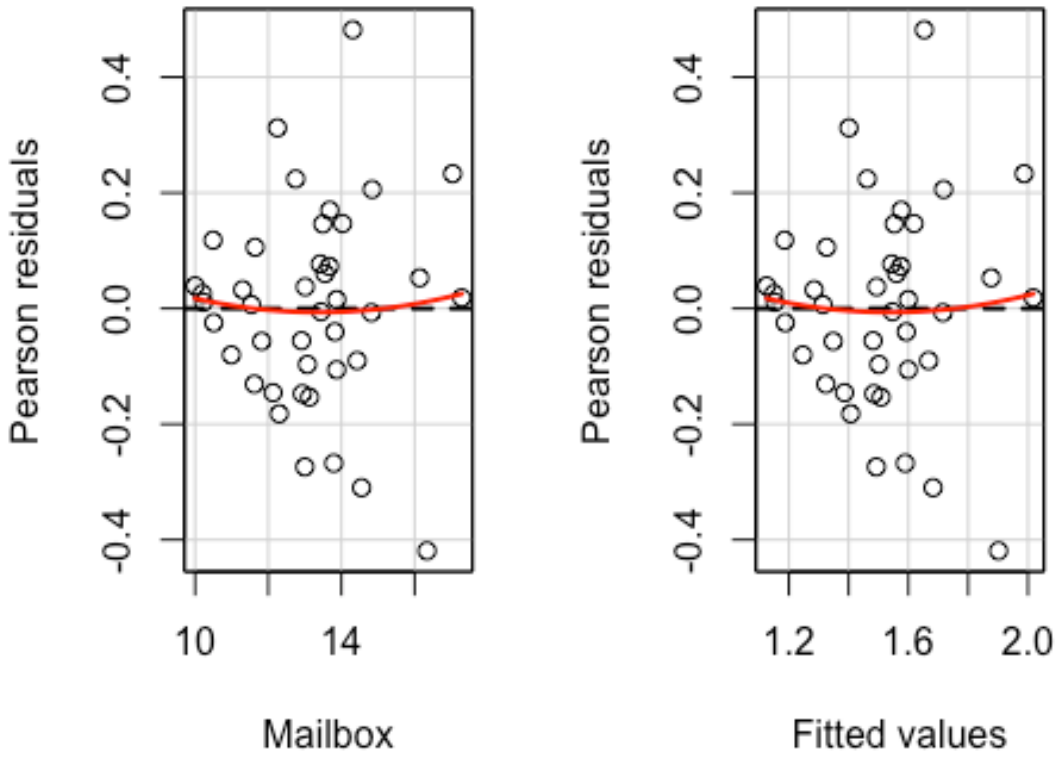
We choose this as the best among the models tested to predict the Class IV commodity price from the mailbox price.

Butter

Model 1

On regressing butter prices on the mailbox price in untransformed form.

```
lm_Butter <- lm(Butter ~ Mailbox, data = mnm)


Call:
lm(formula = Butter ~ Mailbox, data = mnm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41908 -0.09683  0.01078  0.07692  0.48196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09895    0.20423  -0.484    0.631
Mailbox      0.12243    0.01542   7.940 1.14e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1736 on 39 degrees of freedom
Multiple R-squared:  0.6178,    Adjusted R-squared:  0.608
F-statistic: 63.04 on 1 and 39 DF,  p-value: 1.14e-09
```




```
           Test stat Pr(>|t|)
Mailbox        0.319    0.752
Tukey test     0.319    0.750

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 5.931632    Df = 1     p = 0.01487146
```

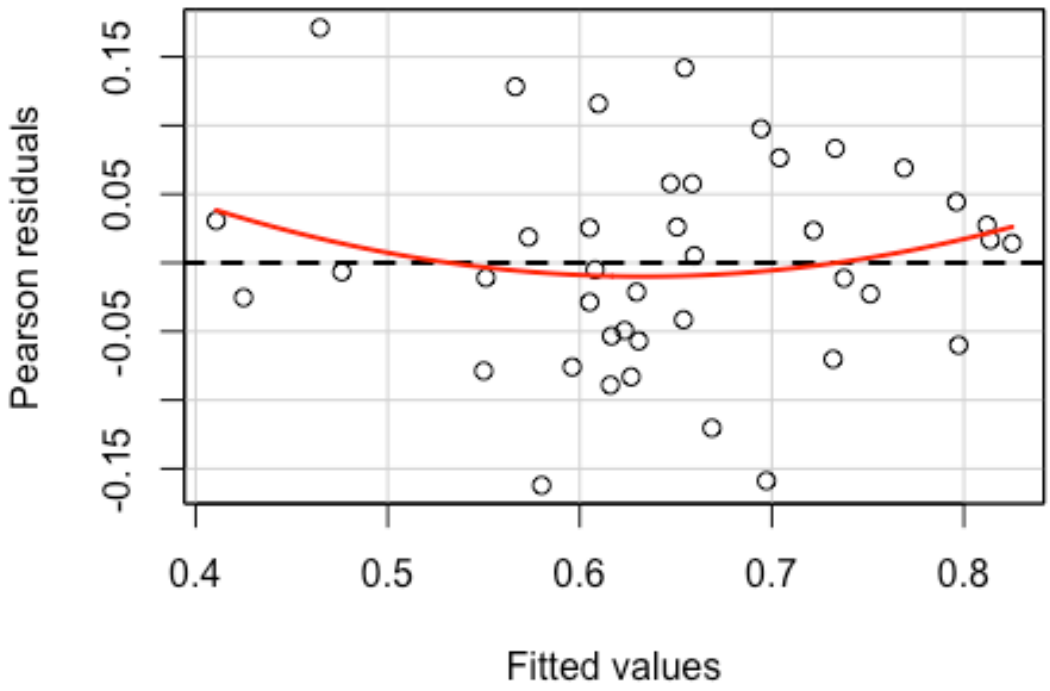
One of the coefficients of the regression model is significant (p-value > 0.025). Also, the model fails ncvTest for homoscedasticity.

Model 2

We transform the butter prices to treat heteroscedasticity. The power of transformation(λ) as determined by box-cox transformation is -1.15.

```
lm_butter <- lm(I(Butter^ -1.15) ~ Mailbox , data=mnm)


Call:
lm(formula = I(Butter^-1.15) ~ Mailbox, data = mnm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16211 -0.05332 -0.00507  0.04430  0.17112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.393528   0.091072  15.301  < 2e-16 ***
```

```
Mailbox     -0.056842   0.006876  -8.266 4.19e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07742 on 39 degrees of freedom
Multiple R-squared:  0.6366,    Adjusted R-squared:  0.6273
F-statistic: 68.33 on 1 and 39 DF,  p-value: 4.192e-10
```



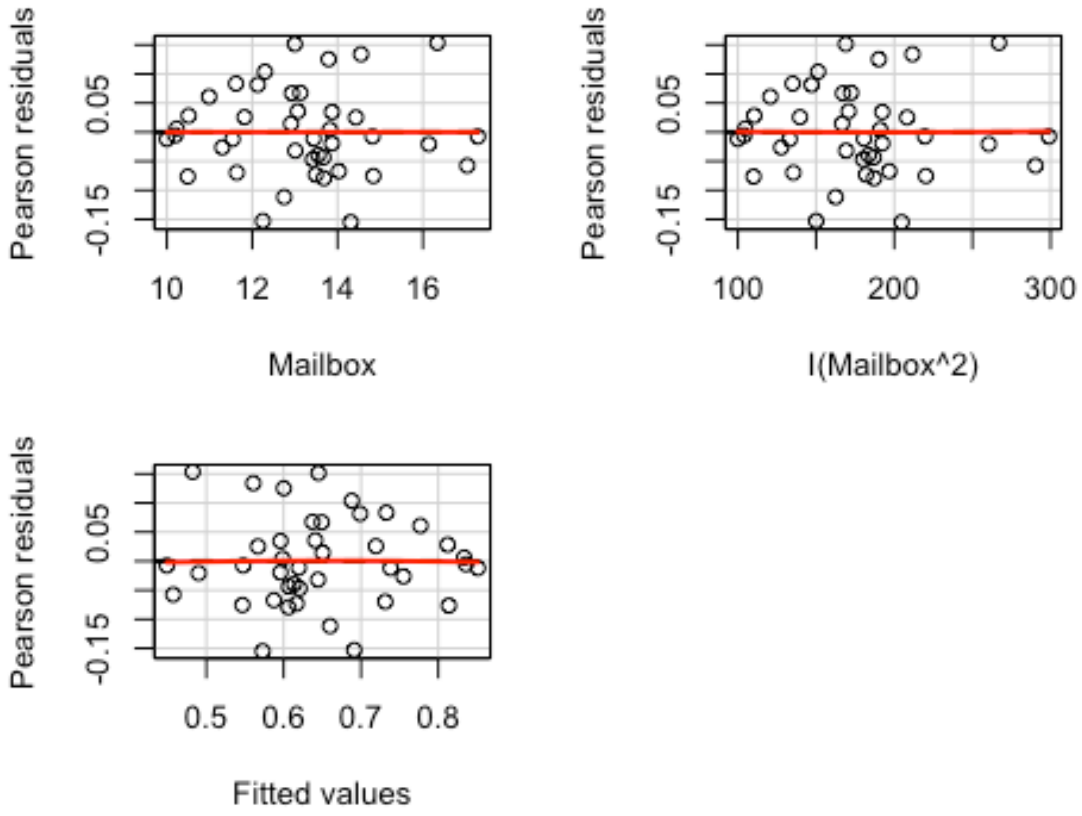
The model passes the homoscedasticity and normality tests. However, the residual plots show variance which means that a better model can be obtained if we include transformations.

Model 3

On transforming mailbox price and following the heirarchial principal, we get the equation as:

```
lm_butter <- lm(I(Butter^ -1.15) ~ Mailbox + I(Mailbox ^ 2), data = mnm)


Call:
lm(formula = I(Butter^-1.15) ~ Mailbox + I(Mailbox^2), data = mnm)

Residuals:
      Min        1Q    Median        3Q       Max
```

```
-0.154860 -0.046962 -0.007613  0.035740  0.153444

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.949155   0.516194   3.776 0.000546 ***
Mailbox      -0.141375   0.077612  -1.822 0.076400 .
I(Mailbox^2)  0.003159   0.002889   1.093 0.281075
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07723 on 38 degrees of freedom
Multiple R-squared:  0.6477,    Adjusted R-squared:  0.6292
F-statistic: 34.93 on 2 and 38 DF,  p-value: 2.459e-09
```




```
             Test stat Pr(>|t|)
Mailbox         -2.129    0.040
I(Mailbox^2)     0.245    0.808
Tukey test      -0.316    0.752
```

We know that when polynomial terms are included in the model, multicollinearity comes into play. The higher order term needs to be statistically significantly different from zero to be included in the model (although lower order terms can have a p-value slightly higher than the significance level). In this case, Mailbox^2 is insignificant and thus cannot be included in the model. Also, though R^2 increases, R^2_{adj} almost remains the same and the gap

between the two has increased. This implies that the polynamial is not adding significantly to the model.

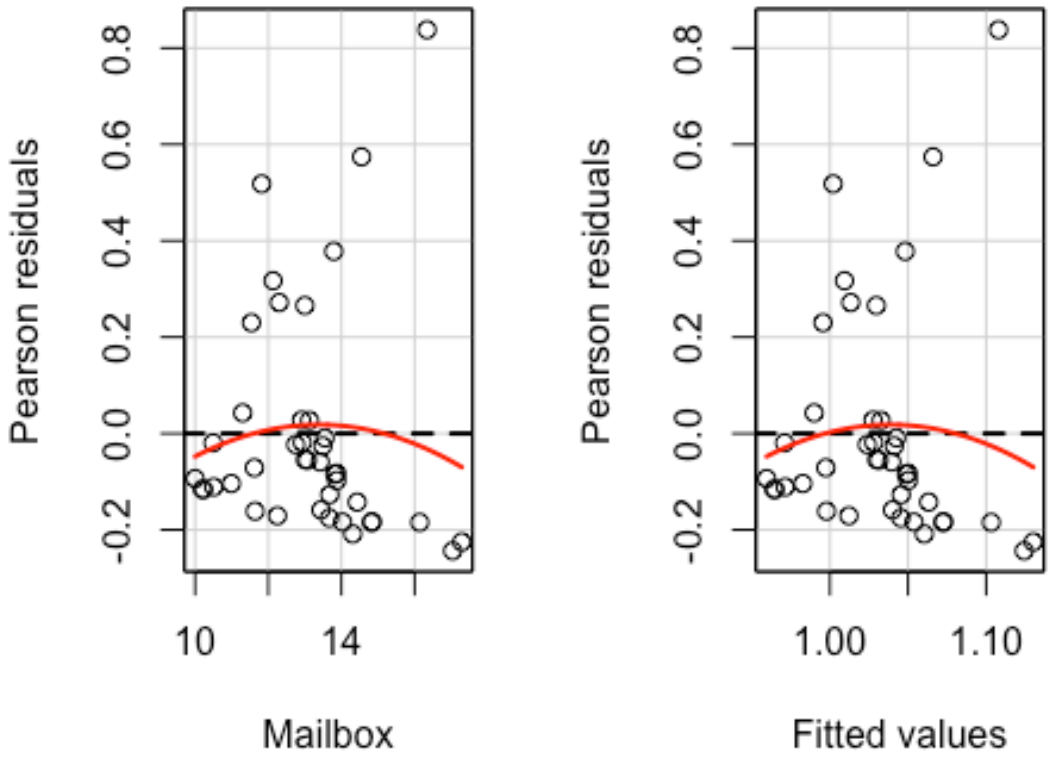
We thus take Model 3 as our final model for predicting butter prices.

NFDM

Model 1

On regressing NFDM prices on the mailbox price in untransformed form.

```
lm_nfdm <- lm(NFDM ~ Mailbox, data = mnm)


Call:
lm(formula = NFDM ~ Mailbox, data = mnm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24355 -0.15862 -0.08269  0.02694  0.83795

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.72517    0.28441   2.550   0.0148 *
Mailbox      0.02343    0.02147   1.091   0.2820
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2418 on 39 degrees of freedom
Multiple R-squared:  0.02961,   Adjusted R-squared:  0.004729
F-statistic:  1.19 on 1 and 39 DF,  p-value: 0.282
```




```
           Test stat Pr(>|t|)
Mailbox       -0.633    0.530
Tukey test    -0.633    0.526
```

One of the coefficients of the regression model is statistically significantly equal to zero. The value of R^2 is 0.02961 which indicates that mailbox price can't much explain the variation in the NFDM price.

Model 2

We transform NFDM prices using box-cox transformation. The value of lambda is -2.

```
lm_nfdm <- lm(I(NFDM ^ -2) ~ Mailbox, data = mnm)


Call:
lm(formula = I(NFDM^-2) ~ Mailbox, data = mnm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70888 -0.14497  0.04774  0.25737  0.36102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.32914    0.37100   3.583 0.000932 ***
Mailbox     -0.02180    0.02801  -0.778 0.441191
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3154 on 39 degrees of freedom
Multiple R-squared:  0.01529,   Adjusted R-squared:  -0.009961
F-statistic: 0.6055 on 1 and 39 DF,  p-value: 0.4412
```

It can be seen that there is still no improvement in the model.

Model 3

On transforming the mailbox price and following the hierarchical principal, we get the model as:

```
lm_nfdm <- lm(NFDM ~ Mailbox + I(Mailbox ^ 2), data = mnm)


Call:
lm(formula = NFDM ~ Mailbox + I(Mailbox^2), data = mnm)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22185 -0.15526 -0.07551  0.01053  0.87024

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.290268   1.628609  -0.178    0.859
Mailbox       0.177915   0.244868   0.727    0.472
I(Mailbox^2) -0.005772   0.009114  -0.633    0.530

Residual standard error: 0.2437 on 38 degrees of freedom
Multiple R-squared:  0.03975,   Adjusted R-squared:  -0.01079
F-statistic: 0.7865 on 2 and 38 DF,  p-value: 0.4627
```

It can be seen that the R^2 has improved even after transformation. All of the coefficients of the regression model have their p-vales greater than 0.025. Thus, this model cannot be used.

Even after multiple iterations and using different transformations it was not possible to arrive at a model that can predict NFDM commodity price from the mailbox price. This is also evident from the coefficient of correlation between the two which is very low, 0.17. The regression coefficients of the model remain significant due to endogenity as we are missing on the potential factors which explain the variation in the NFDM price. We can say that mailbox price does not explain much variation in NFDM and hence any of the transformations of the same won't add much in explaining the variation.