The Impact of Major Commodities on Crude Oil Price: An Analysis from 2000 to 2020

2025-03-06

```
library(readr)
COMMODITY_PRICE <- read_csv("C:/Users/gabri/OneDrive/Desktop/Applied Linear Model/COMMODITY_PRICE.csv")
```

```
## Rows: 246 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (16): Coal, Natural_Gas, Cocoa, Coconut_Oil, Wheat, Banana, Orange, Shr...
## date  (1): Month
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(COMMODITY_PRICE)
```

```
## # A tibble: 6 x 17
##   Month       Coal Natural_Gas Cocoa Coconut_Oil Wheat Banana Orange Shrimps
##   <date>     <dbl>       <dbl> <dbl>       <dbl> <dbl>  <dbl>  <dbl>   <dbl>
## 1 2000-01-01  22.8        3.36  918.         654  106.  0.885  0.349    14.0
## 2 2000-02-01  23.6        3.46  857.         591  108.  0.920  0.324    14.1
## 3 2000-03-01  24.4        3.51  926.         552  106.  0.864  0.299    14.4
## 4 2000-04-01  24.9        3.68  912          550  105.  0.837  0.337    14.7
## 5 2000-05-01  26.0        3.55  922.         481  112.  0.703  0.411    14.9
## 6 2000-06-01  27          3.6   948.         437  114.  0.622  0.426    14.7
## # i 8 more variables: Cotton <dbl>, Potassium <dbl>, Aluminum <dbl>,
## #   Lead <dbl>, Zinc <dbl>, Gold <dbl>, Crude_Oil <dbl>, Post_Crisis <dbl>
```

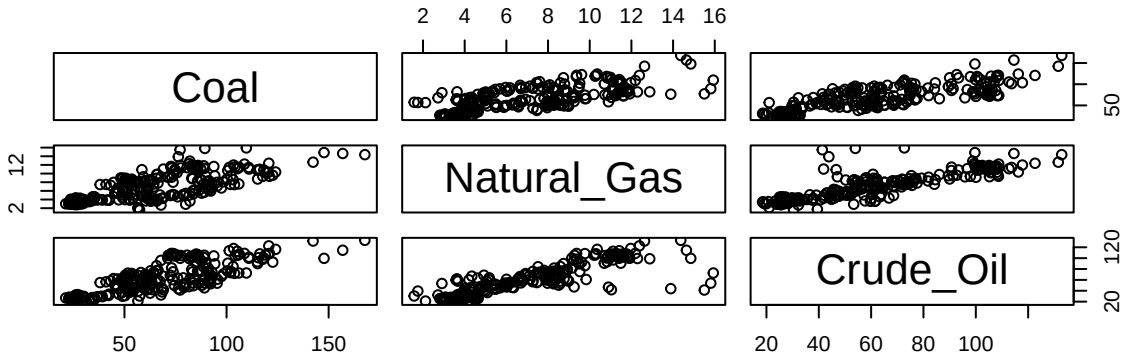
The data was taken from the World Bank website. The variables in the dataset are: - Coal (the cost in USD of coal per ton), - Natural Gas (the cost in USD of natural gas per million British thermal units (MMBtu)), - Crude Oil (the cost in USD of crude oil per barrel, this will be the response of the model), - Aluminum (the cost in USD of aluminum per ton), - Gold (the cost in USD of gold per ounce), - Lead (the cost in USD of lead per ton), - Zinc (the cost in USD of zinc per ton), - Cocoa (the cost in USD of cocoa per ton), - Coconut Oil (the cost in USD of coconut oil per metric ton), - Wheat (the cost in USD of wheat per ton), - Banana (the cost in USD of banana per kg), - Orange (the cost in USD of orange per Kg) - Shrimps (the cost in USD of orange per kg) - Potassium (the cost in USD of potassium per metric ton) - Cotton (the cost in USD of cotton per kg) - Post Crisis a categorical variable indicating whether an observation is before or after September 2008. I chose September 2008 because it is the month when the financial company Lehman Brothers declared bankruptcy. It is 0 if it's before September 2008, 1 if it's after.

Goal of the Study

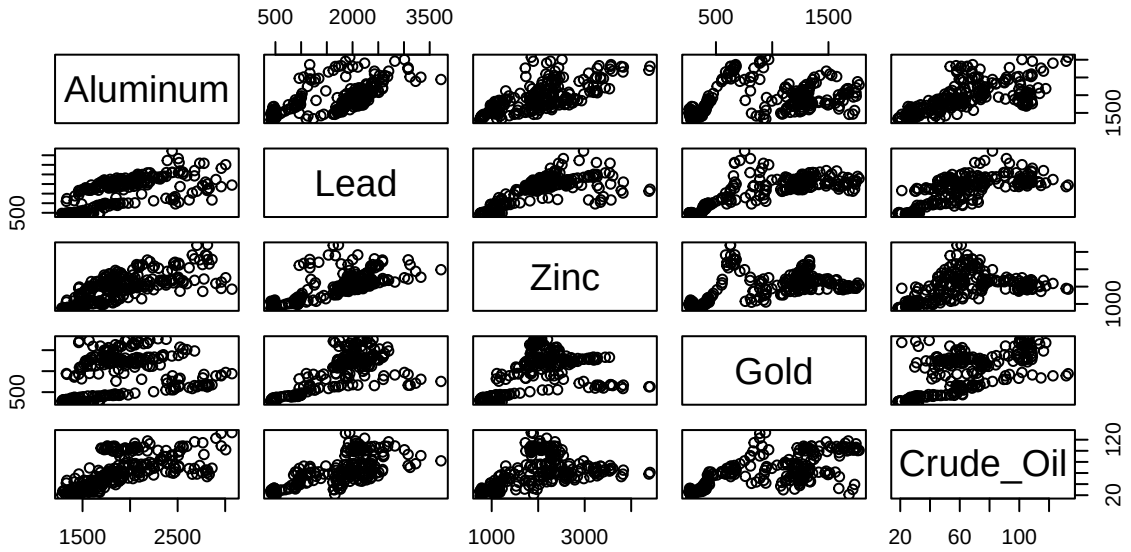
The goal of this study is to understand how some of the major commodities have influenced the price of crude oil from 2000 to 2020. Additionally, I would like to determine whether the 2008 financial crisis had an impact on commodity prices.

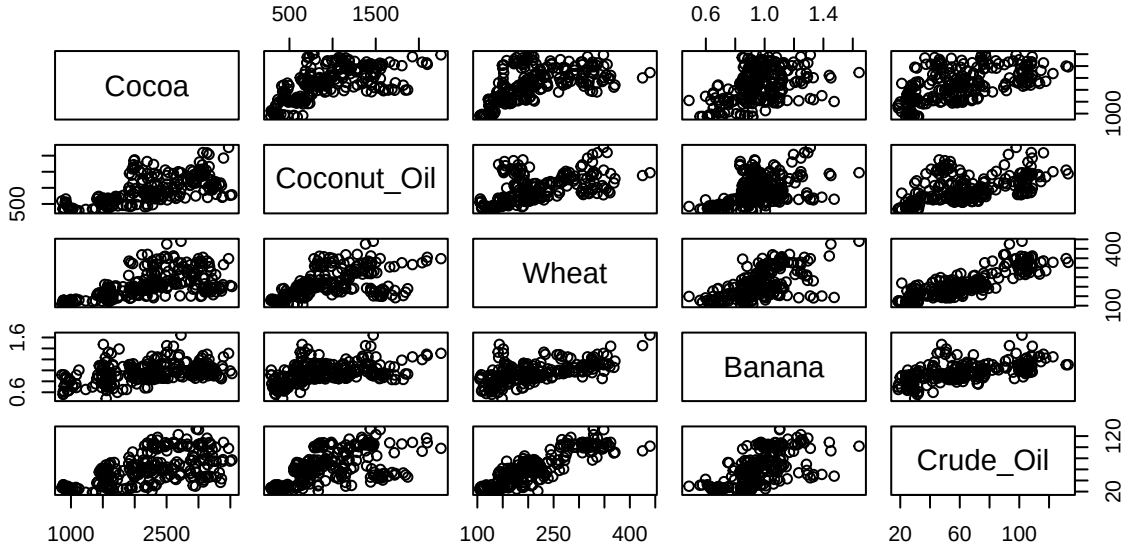
Representation of data

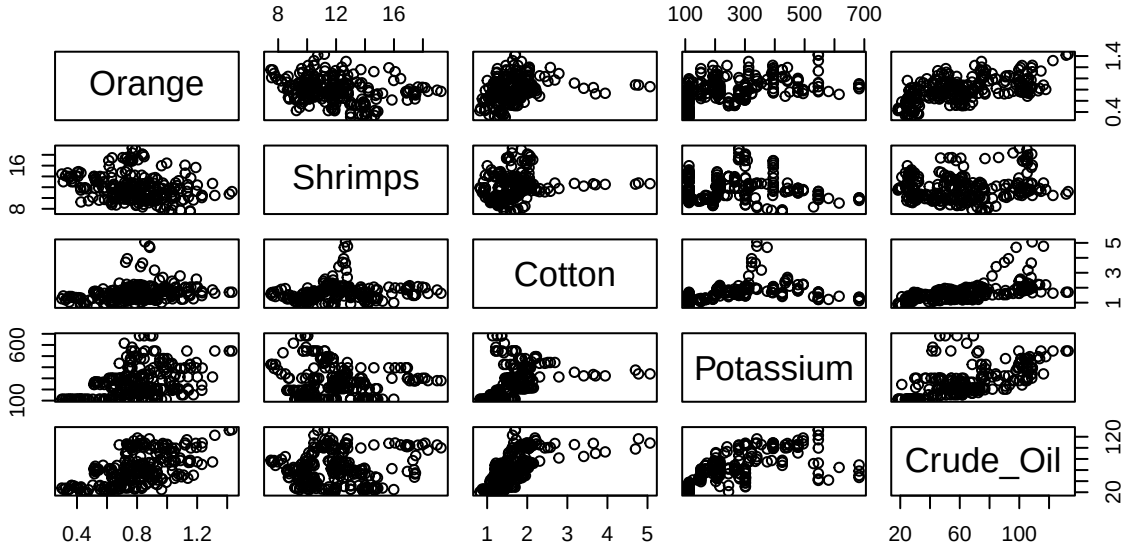
To better understand how the price of crude oil depends on the other variables, the following graphs will be very helpful. I will create four separate plots: the first one will focus on crude oil and energy related variables, the second will explore crude oil and metals, and the third and fourth will examine crude oil and the remaining



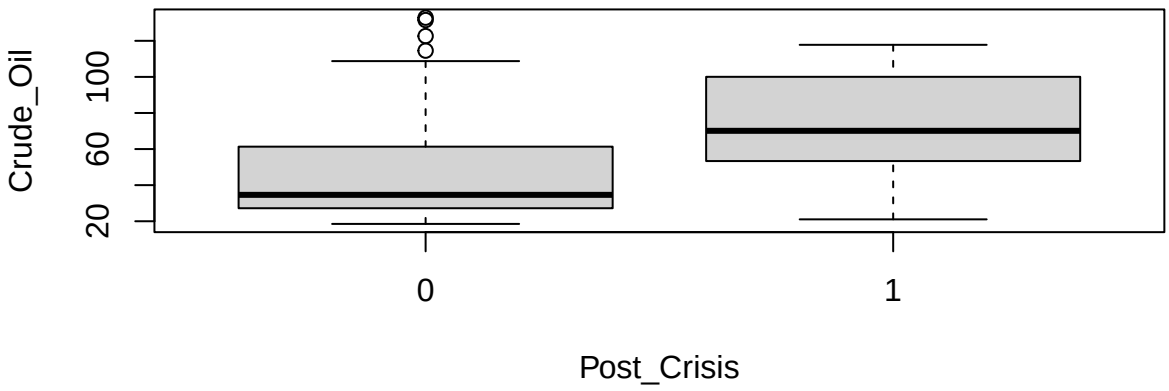
commodities.







As expected crude oil has a more linear relationship with other energy commodities compared to other types, as they share common price drivers such as energy demand, geopolitics, and substitutability between sources. It would be also interesting to see how the price of oil changed before and after the 2008 crisis.



It can be observed that the median crude oil prices are visibly higher in the post-crisis period compared to the pre-crisis period. This indicates an overall increase in prices after the crisis. Before fitting the model, it is useful to center all the covariates in order to better interpret the model's intercept.

Variable selection

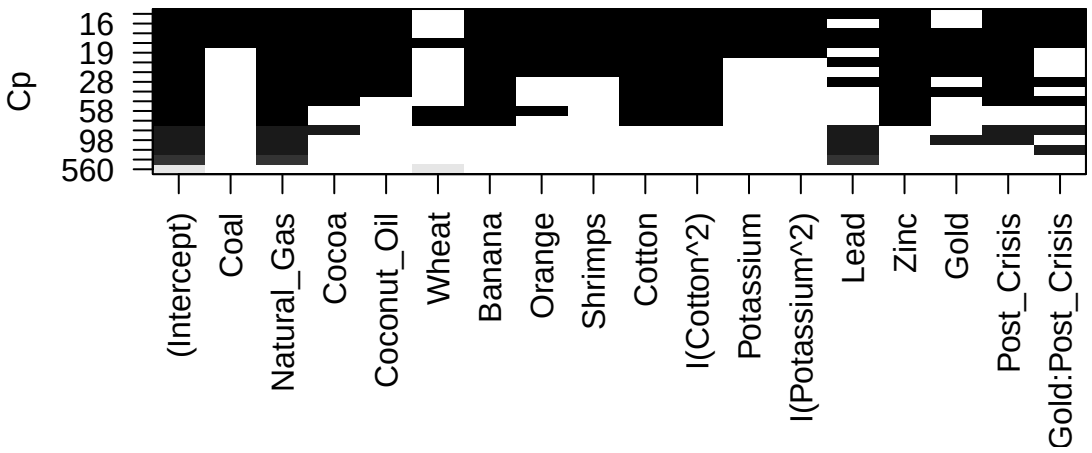
Based on the previous scatterplots, I observed that the response variable, crude oil, might have a quadratic relationship with potassium and cotton. Let's perform variable selection using best subset selection, forward selection and backward selection

```
library(leaps)
```

```
## Warning: il pacchetto 'leaps' è stato creato con R versione 4.4.2
```

```
Best_Subset_Selection = regsubsets(log(Crude_Oil) ~ Coal + Natural_Gas + Cocoa
                                   + Coconut_Oil+ Wheat + Banana + Orange
                                   + Shrimps + Cotton + I(Cotton^2) + Potassium
                                   + I(Potassium^2) + Lead + Zinc
                                   + Gold+ Post_Crisis + Gold * Post_Crisis,
                                   data = COMMODITY_PRICE, nvmax = 18)
summ = summary(Best_Subset_Selection)
Backward_Selection = regsubsets(Crude_Oil ~ Coal + Natural_Gas + Cocoa + Coconut_Oil + Wheat +
                          Banana + Orange + Shrimps + Cotton + I(Cotton^2) + Potassium +
                          I(Potassium^2) + Aluminum + Lead + Zinc + Gold + Post_Crisis +
                          Gold * Post_Crisis, data = COMMODITY_PRICE, nvmax = 18,
                        method = "backward")
summ_backward = summary(Backward_Selection)
Forward_Selection = regsubsets(Crude_Oil ~ Coal + Natural_Gas + Cocoa + Coconut_Oil + Wheat
                           + Banana + Orange + Shrimps + Cotton + I(Cotton^2) + Potassium +
                             I(Potassium^2) + Aluminum + Lead + Zinc + Gold + Post_Crisis +
                             Gold * Post_Crisis, data = COMMODITY_PRICE, nvmax = 18,
                           method = "forward")
summ_forward = summary(Forward_Selection)
```

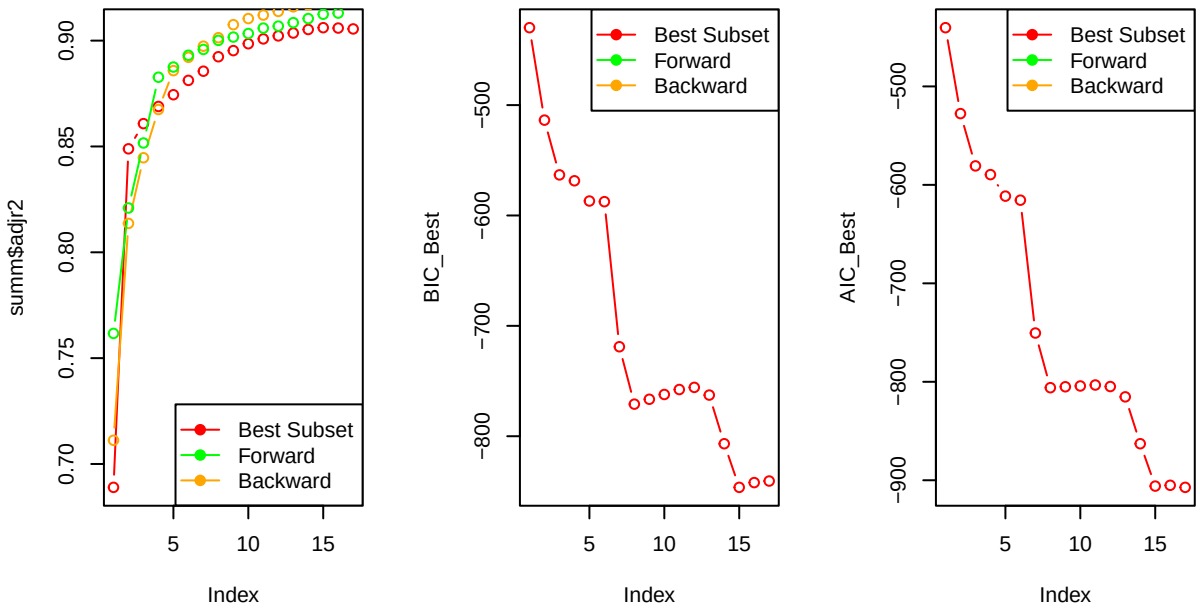
```r
plot(Best_Subset_Selection, scale = "Cp")
```




```r
n = nrow(COMMODITY_PRICE)
RSS = Best_Subset_Selection$rss[-1]
AIC_Best = c(0)
for (i in 1:length(RSS)){
  AIC_Best[i] = n*log(RSS[i]/n) + 2*(i + 2)

}
RSS = Best_Subset_Selection$rss[-1]
BIC_Best = c(0)
for (i in 1:length(RSS)){
  BIC_Best[i] = n*log(RSS[i]/n) + log(n)*(i + 2)
}
```

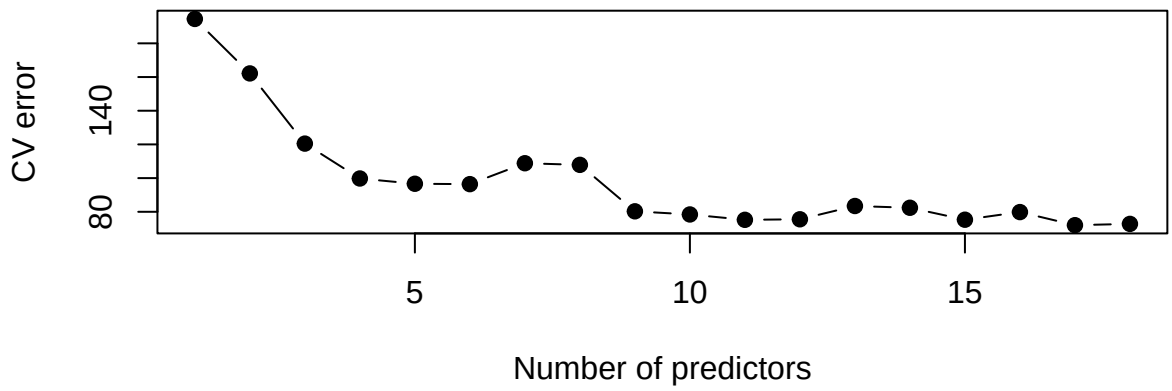
I find in the same way the AIC and BIC for Backward and Forward.




```
p = 18
k = nrow(COMMODITY_PRICE)
folds = sample (1:k,nrow(COMMODITY_PRICE),replace = FALSE)
cv.errors <- matrix (NA ,k, p, dimnames =list(NULL , paste (1:p) ))
for(j in 1:k){
  ols_best = regsubsets(Crude_Oil ~ Coal + Natural_Gas + Cocoa + Coconut_Oil
                        + Wheat + Banana + Orange + Shrimps + Cotton
                        + I(Cotton^2) + Potassium + I(Potassium^2) + Aluminum
                        + Lead + Zinc + Gold + Post_Crisis
                        + Gold * Post_Crisis,
                        data = COMMODITY_PRICE[folds !=j,], nvmax = 18)
  for(i in 1:p) {
    mat <- model.matrix(as.formula(ols_best$call[[2]]), COMMODITY_PRICE[folds==j,])
    coefi <- coef(ols_best ,id = i)
    xvars <- names(coefi )
    pred <- mat[,xvars ]%*% coefi
    cv.errors[j,i] <- mean((COMMODITY_PRICE$Crude_Oil[folds==j] - pred)^2)
  }
}
cv.mean = colMeans(cv.errors)
cv.mean
```

```
##         1         2         3         4         5         6         7         8
## 194.48584 162.15124 120.47324  99.76865  96.63727  96.40819 108.83091 107.86389
##         9        10        11        12        13        14        15        16
##  80.28015  78.41936  75.22856  75.55337  83.43728  82.38689  75.30343  79.83057
##        17        18
##  72.05676  72.80845
```

```
plot(cv.mean ,type="b",pch=19,
     xlab="Number of predictors",
     ylab="CV error")
```



I'll use also the Lasso to do variable selection.

```
library(glmnet)
```

```
## Caricamento del pacchetto richiesto: Matrix
```

```
## Loaded glmnet 4.1-8
```

```
X = model.matrix(Crude_Oil ~ . + I(Cotton^2) + I(Potassium^2)
                 + Gold * Post_Crisis, data = COMMODITY_PRICE[, -1])[, -1]
y = COMMODITY_PRICE$Crude_Oil
set.seed(1)
grid = 10^seq(8, -4, length=100)
cv.lasso = cv.glmnet(x = X, y = y, alpha = 1, nfold = 10, lambda = grid)
lambda_min = cv.lasso$lambda.min
lambda_1se = cv.lasso$lambda.1se
lambda_min
```

```
## [1] 0.01519911
```

```
lambda_1se
```

```
## [1] 0.1873817
```

```
predict(cv.lasso, type = "coefficients", s = lambda_min)
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                               s1
```

```
## (Intercept)       6.620308e+01
## Coal              1.393487e-01
## Natural_Gas       2.693511e+00
## Cocoa             6.896743e-03
## Coconut_Oil      -8.563942e-03
## Wheat             3.626202e-02
## Banana            1.188776e+01
## Orange            8.285752e+00
## Shrimps           1.761905e+00
## Cotton            1.377809e+01
## Potassium         8.469889e-02
## Aluminum          3.980177e-03
## Lead              4.242234e-03
## Zinc              3.152911e-03
## Gold             -2.700849e-02
## Post_Crisis      -1.795776e+01
## I(Cotton^2)      -3.638809e+00
## I(Potassium^2)   -1.636871e-04
## Gold:Post_Crisis  5.143175e-02
```

As we can see both the minimum lambda and the lambda + 1 standard deviation are very close to 0, so it is a very similar model to least squares. Furthermore, even if the model recommended by Lasso regression is the full model, looking at the BIC, AIC, Cp and adj R^2, we notice that in reality there are other models that are about as good as the full model but are also a bit simpler, consequently I consider as the best model (also taking into account the hierarchical principle) the one with 15 predictors, all predictors except aluminum, orange and wheat.

Collinearity

```
Best_Model = glm(Crude_Oil ~ Coal + Natural_Gas + Cocoa + Coconut_Oil
                 + Banana + Shrimps + Cotton + I(Cotton^2) + Potassium
                 + I(Potassium^2) + Lead + Zinc + Gold + Post_Crisis
                 + Gold * Post_Crisis, data = COMMODITY_PRICE)
library(car)
```

```
## Caricamento del pacchetto richiesto: carData
```

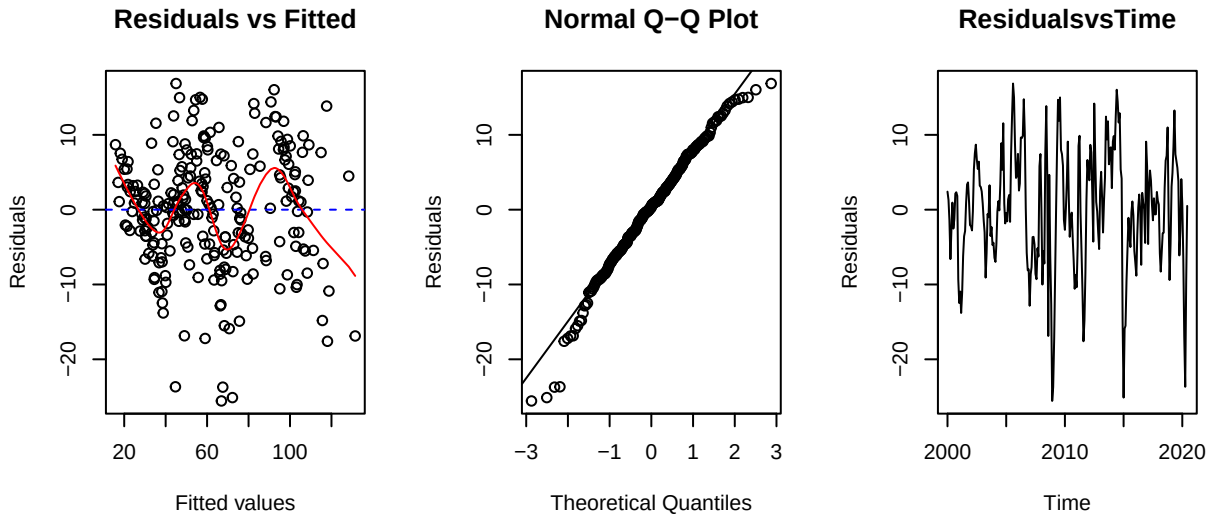
```
vif(Best_Model)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##             Coal      Natural_Gas            Cocoa      Coconut_Oil
##         6.397298         5.810127         6.059264         4.163846
##           Banana          Shrimps           Cotton      I(Cotton^2)
##         2.087358         1.614557        10.650980         4.689720
##        Potassium   I(Potassium^2)             Lead             Zinc
##        26.684708         9.101954         8.960495         4.126945
##             Gold      Post_Crisis Gold:Post_Crisis
##       121.678141        15.393841        35.151928
```

There is no collinearity problem.

Diagnostics



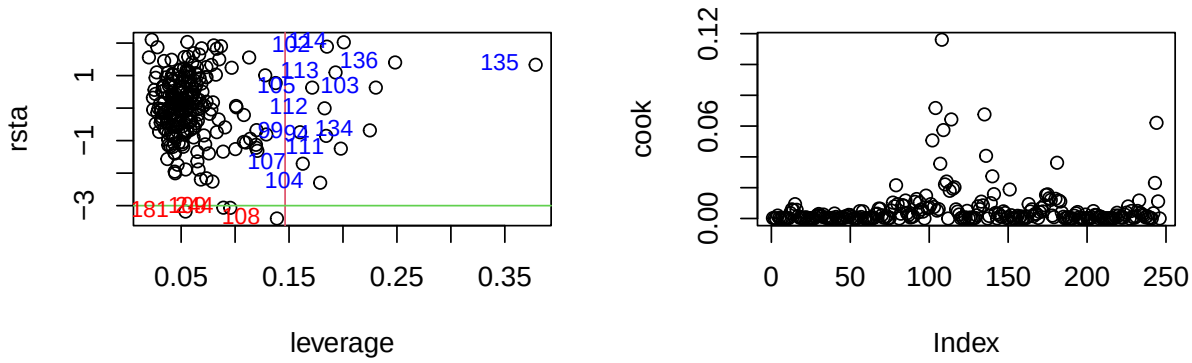

```
shapiro.test(residuals(Best_Model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(Best_Model)
## W = 0.98364, p-value = 0.006292
```

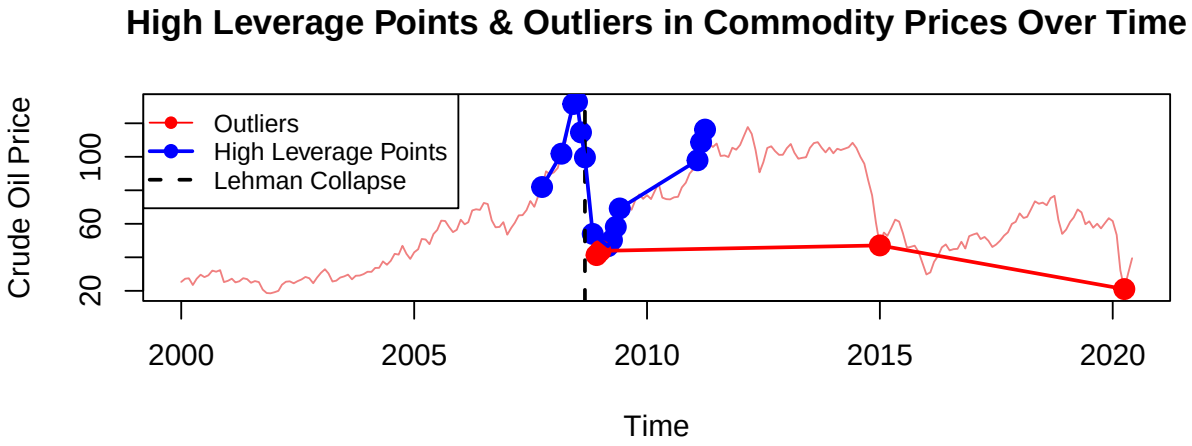
We have a problem with non constant variance and also with linearity so a transformation of the response could help. Despite the variance not being constant, I believe it is not reasonable to use Weighted Least Squares since there is no suitable variable to be used as a weight. We reject H0 with the Shapiro test, that means we reject the assumption of normality, and if we look at the Q-Q plot we can see that there is some skewness. Since we have skewness a solution could be a transformation of the response variable, such as a logarithmic transformation. The errors are a bit correlated since I'm using time series. Check High leverage points, outliers and influential points.

```
par(mfrow = c(1, 2))
rsta = rstandard(Best_Model)
infl = influence(Best_Model)
leverage = infl$hat
plot(leverage, rsta)
threshold = (17 + 1)*2/nrow(COMMODITY_PRICE)
abline(v = threshold, col = 2)
high_leverage_points = which(leverage > threshold)
text(leverage[high_leverage_points], rsta[high_leverage_points],
     labels = high_leverage_points, pos = 2, col = "blue", cex = 0.8)
abline(h=3,col=3)
abline(h=-3,col=3)
outliers = which(rsta > 3 | rsta < -3)
text(leverage[outliers], rsta[outliers], labels = outliers,
     pos = 2, col = "red", cex = 0.8)
```

```
cook = cooks.distance(Best_Model)
plot(cook)
abline(h = 0.5)
abline(h = 1)
```



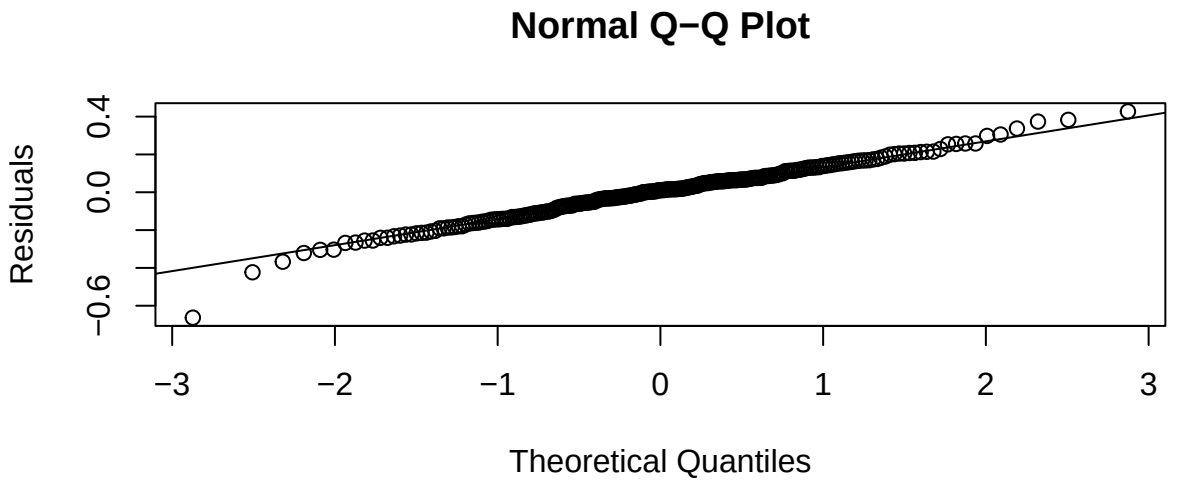
The red points are the outliers, the blue ones are the high leverage points. There are not influential points. Let's try do understand more about these high leverage points and outliers.

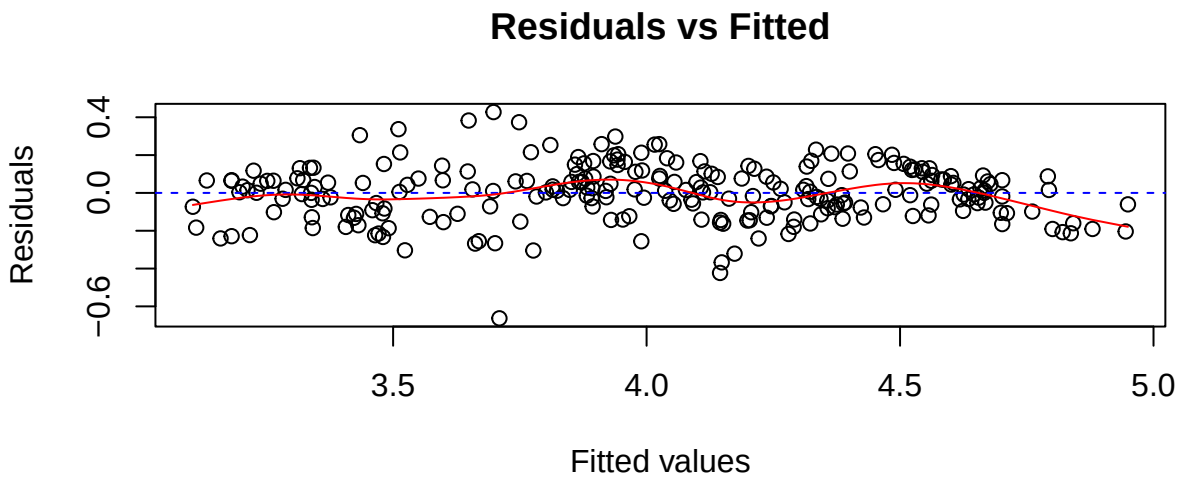


The high leverage points in the dataset correspond to the period between 2008 and 2012, which correspond with the global financial crisis. These points do not represent measurement errors or anomalies in the data but rather reflect a historical event that had extreme impacts on the global economy and, consequently, on commodity prices. A similar reasoning applies to the outliers: Between June 2014 and January 2015, the price of crude oil fall by over 50%, dropping from approximately 115 USD per barrel (June 2014) to around 46 USD per barrel (January 2015). The key cause is the excess supply, by 2014, the United States had surpassed Saudi Arabia and Russia to become the world's largest oil producer. Furthermore in November 2014, OPEC (Organization of the Petroleum Exporting Countries), in particular, Saudi Arabia and Gulf countries, refused to cut oil production. April 2020 COVID-19 Shock: Global lockdowns led to an unprecedented demand collapse, causing historic price drops. Therefore, the only modification I will make is the logarithmic transformation of the model's response variable. If we do again the variable selection, the best model will be the one with all predictors except Wheat.

```
Best_Model = glm(log(Crude_Oil) ~ Coal + Natural_Gas + Cocoa + Coconut_Oil
                 + Banana + Shrimps + Cotton + I(Cotton^2) + Potassium
                 + I(Potassium^2) + Orange + Lead + Zinc + Gold + Post_Crisis
                 + Gold * Post_Crisis, data = COMMODITY_PRICE)
```

I'll delete also Aluminum to avoid collinearity. After rechecking the outliers and high leverage points, they are still linked to the 2008 crisis and 2020, while the 2015 period is better explained by the new model. Let's see the Q-Q plot and the Fitted values vs Residuals plot.





As we can see the normality assumption is correct because just one point is far from the line. This transformation has also improved the linearity. The non constant variance problem remains but it is slightly better.

Interpretation of parameters

```
summ = summary(Best_Model)
summ
```

```
##
## Call:
## glm(formula = log(Crude_Oil) ~ Coal + Natural_Gas + Cocoa + Coconut_Oil +
##     Banana + Shrimps + Cotton + I(Cotton^2) + Potassium + I(Potassium^2) +
##     Orange + Lead + Zinc + Gold + Post_Crisis + Gold * Post_Crisis,
##     data = COMMODITY_PRICE)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.148e+00  8.735e-02  47.489  < 2e-16 ***
## Coal              1.750e-03  8.676e-04   2.017 0.044812 *
## Natural_Gas       5.110e-02  7.526e-03   6.791 9.50e-11 ***
## Cocoa             9.801e-05  3.500e-05   2.800 0.005542 **
## Coconut_Oil      -1.938e-04  4.683e-05  -4.138 4.92e-05 ***
## Banana            3.490e-01  7.986e-02   4.370 1.89e-05 ***
## Shrimps           1.843e-02  5.301e-03   3.476 0.000608 ***
## Cotton            2.809e-01  5.440e-02   5.163 5.27e-07 ***
## I(Cotton^2)      -7.495e-02  1.721e-02  -4.355 2.01e-05 ***
## Potassium         1.145e-03  3.650e-04   3.138 0.001926 **
## I(Potassium^2)   -2.780e-06  1.022e-06  -2.720 0.007036 **
## Orange            1.842e-01  6.899e-02   2.670 0.008122 **
## Lead              7.495e-05  3.924e-05   1.910 0.057399 .
## Zinc              1.275e-04  2.466e-05   5.172 5.06e-07 ***
## Gold             -1.719e-04  2.266e-04  -0.759 0.448853
## Post_Crisis      -2.438e-01  7.727e-02  -3.156 0.001816 **
## Gold:Post_Crisis  4.192e-04  2.330e-04   1.800 0.073237 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.02303627)
##
##     Null deviance: 59.9995  on 245  degrees of freedom
## Residual deviance:  5.2753  on 229  degrees of freedom
## AIC: -211.09
##
## Number of Fisher Scoring iterations: 2
```

Intercept = 4.148. When all other variables are at their mean values, the expected value of the natural logarithm of the crude oil price is 4.148. This implies that the expected crude oil price is approximately exp(4.148) = 63.3 USD per barrel. The small standard error (0.08735) suggests a reliable estimate. The p-value (< 2e-16) indicates that the intercept is highly statistically significant. **Beta.Coal = 1.750e-03**. An increase of 1 USD per ton above the mean in the price of coal results in an average increase of 0.00175 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The small standard error (0.0008676) suggests a reliable estimate. The p-value (0.044812) indicates that the effect is statistically significant ($p < 0.05$). **Beta.Natural_Gas = 5.110e-02**. An increase of 1 USD per million British thermal units (MMBtu) above the mean in the price of natural gas results in an average increase of 0.0511 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The

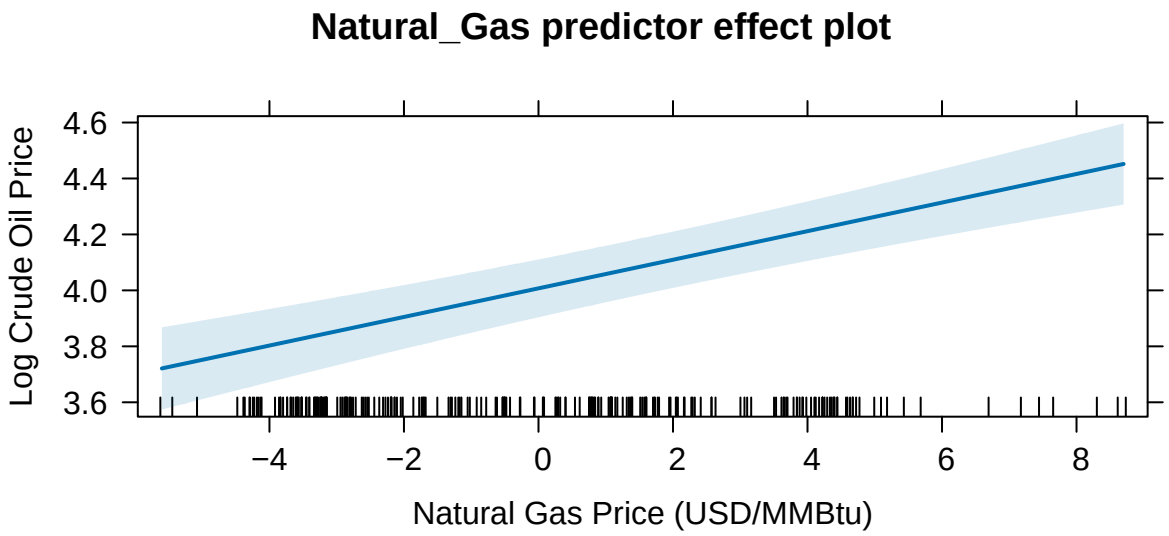
small standard error (0.007526) suggests a reliable estimate. The p-value (9.5e-11) indicates that the effect is highly statistically significant ($p < 0.001$). **Beta.Cocoa = 9.801e-05**. An increase of 1 thousand USD per ton above the mean in the price of cocoa results in an average increase of 0.00009801 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The small standard error (0.000035) suggests a reliable estimate. The p-value (0.005542) indicates that the effect is highly statistically significant ($p < 0.01$). **Beta.Coconut_Oil = -1.938e-04**. An increase of 1 USD per metric ton above the mean in the price of coconut oil results in an average decrease of -0.0001938 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The small standard error (0.00004683) suggests a reliable estimate. The p-value (4.92e-05) indicates that the effect is statistically significant ($p < 0.01$). **Beta.Banana = 3.490e-01**. An increase of 1 USD per kilogram above the mean in the price of bananas results in an average increase of 0.349 in the natural logarithm of the price of crude oil, keeping all other variables at their mean.The small standard error (0.07986) suggests a reliable estimate.The p-value (1.89e-05) indicates that the effect is statistically significant ($p < 0.001$). **Beta.Orange = 1.842e-01**. An increase of 1 USD per kilogram above the mean in the price of oranges results in an average increase of 0.1842 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The small standard error (0.06899) suggests a reliable estimate. The p-value (0.008122) indicates that the effect is statistically significant ($p < 0.01$). **Beta.Shrimps = 1.843e-02**. An increase of 1 USD per kilogram above the mean in the price of shrimps results in an average increase of 0.01843 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The small standard error (0.005301) suggests a reliable estimate. The p-value (0.000608) indicates that the effect is statistically significant ($p < 0.001$). **Beta.Cotton = 2.809e-01**. An increase of 1 USD per kilogram above the mean in the price of cotton results in an average increase of 0.2809 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The small standard error (0.05440) suggests a reliable estimate. The p-value (5.27e-07) indicates that the effect is statistically significant ($p < 0.001$). **Beta.Cotton^2 = -7.495e-02**. The negative parameter for the quadratic relationship between crude oil and cotton suggests that as the price of cotton increases, its positive effect on crude oil prices diminishes. The small standard error (0.01721) suggests a reliable estimate. The p-value (2.01e-05) indicates that the effect is statistically significant ($p < 0.001$). **Beta.Potassium = 1.145e-03**. An increase of 1 USD per metric ton above the mean in the price of potassium results in an average increase of 0.001145 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The small standard error (0.0003650) suggests a reliable estimate. The p-value (0.001926) indicates that the effect is statistically significant ($p < 0.01$). **Beta.Potassium^2 = -2.780e-06**. he negative parameter for the quadratic relationship between crude oil and potassium suggests that as the price of potassium increases, its positive effect on crude oil prices diminishes. The small standard error (1.022e-06) suggests a reliable estimate. The p-value (0.007036) indicates that the effect is statistically significant ($p < 0.01$). **Beta.Zinc = 1.275e-04**. An increase of 1 USD per ton above the mean in the price of zinc results in an average increase of 0.0001275 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The small standard error (0.00002466) suggests a reliable estimate. The p-value (5.06e-07) indicates that the effect is highly statistically significant ($p < 0.001$). **Beta.Lead = 7.495e-05**. An increase of 1 USD per ton above the mean in the price of lead results in an average increase of 0.00007495 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The standard error (0.00003924) is quite big. The p-value (0.057399) indicates that the effect is not statistically significant. **Beta.Gold = -1.719e-04**. An increase of 1 USD per ounce above the mean in the price of gold, before the 2008 crisis, resulted in an average decrease of 0.0001719 in the natural logarithm of the price of crude oil, keeping all other variables at their mean. The standard error (0.0002266) is quite big. The p-value (0.448853) indicates that the effect is not statistically significant. **Beta.Post_Crisis = -2.438e-01**. The parameter associated with the qualitative variable Post_Crisis indicates that, on average, the log-price of crude oil is 0.2438 lower after the 2008 financial crisis compared to before the crisis, keeping all other variables at their mean. The small standard error (0.07727) suggests a reliable estimate. The p-value (0.001816) indicates that the effect is statistically significant ($p < 0.01$). **Beta.Gold:Post_Crisis = 4.192e-04**. The parameter associated with the interaction between Gold and Post_Crisis indicates that the effect of gold prices on crude oil prices changed after the 2008 financial crisis. Specifically, if the price of an ounce of gold increases by 1 USD during the post-crisis period, the natural logarithm of crude oil prices increases by Beta_Gold + Beta_Gold*Post_Crisis = (- 0.0001719 + 0.0004192). The standard error (0.0002330) is quite big. The p-value (0.073237) indicates that the effect is not statistically significant. In

order to visualize the effect, of for instance Natural Gas, we can use the following graph.

```
library(effects)
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
plot(predictorEffects(Best_Model, ~ Natural_Gas),
     xlab = "Natural Gas Price (USD/MMBtu)",
     ylab = "Log Crude Oil Price")
```



This graph shows that the expected effect of a higher Natural Gas Price is an increase in Crude Oil Price (log-transformed). The shaded area represents a 95% pointwise confidence interval for the fitted values.

Inferences about regression coefficients

```
p = 16
t_Coal = summ$coefficients["Coal", "t value"]
p_value_Coal = 2 * (1 - pt(abs(t_Coal), df = n - p - 1))
p_value_Coal
```

```
## [1] 0.04481244
```

As can be seen from the p-values, only the coefficients associated with Lead, Gold, and the interaction term are not significantly different from zero.

Just Metals and Energy Commodities?

Let's compare the Best Model with a model with just metals and energy commodities, since metals and other energy commodities are intuitively the most connected to crude oil. To do so I'll use their deviance.

In the Null Hypothesis we state that the additional predictors in Best Model do not significantly improve model fit. That means that by removing them the deviance will not increase significantly. In the Alternative Hypothesis we state the opposite. The difference between the deviances is distributed as Chi-Squared with df_small - df_best degrees of freedom

```
Glm_Small = glm(log(Crude_Oil) ~ Coal + Natural_Gas + Zinc + Lead + Gold + Post_Crisis
                + Gold * Post_Crisis, data = COMMODITY_PRICE)
summ_small = summary(Glm_Small)
G_dif = summ_small$deviance - summ$deviance
df_dif = summ_small$df.residual - summ$df.residual
1 - pchisq(G_dif,df_dif,lower.tail=FALSE)
```

```
## [1] 0.0138139
```

As we can see the p-value is less than 0.05 that means that we can reject the Null Hypothesis: also the commodities that are not metals or related to energy are useful to the model fit.

Goodness of fit

To check how good does the model fit the data I'll use the adj-R^2, since I'm using by default the identity link function and Gaussian family in this glm model. This means that it's equivalent to ordinary least squares (OLS), just formulated as GLM.

```
RSS = summ$deviance
SSy = sum((log(COMMODITY_PRICE$Crude_Oil) - mean(log(COMMODITY_PRICE$Crude_Oil)))^2)
adj_R_2 = 1 - ((n - 1) / (n - p - 1)) * RSS/SSy
adj_R_2
```

```
## [1] 0.9059345
```

This means that the model explains the 90% of the variability of the response log(price of crude oil).

Prediction

Suppose that we have the following information about the prices of the commodities present in the model: Coal = 80.23 USD/ton Natural Gas = 11.78 USD/MMBtu Cocoa = 2438.84 USD/ton Coconut_Oil = 939.47 USD/metric ton Banana = 1.02 USD/kg Orange = 0.96 USD/kg Shrimps = 14.17 USD/kg Cotton = 1.99 USD/kg Potassium = 395 USD/metric ton Zinc = 1910.25 USD/ton Lead = 2139.79 USD/ton Gold = 1411.46 USD/ounce And it is after the Lehman Brothers bankruptcy so Post_Crisis = 1 (Since I need to subtract the mean of the corresponding commodity from these values, I have reloaded the dataset, which was initially modified to center the variables, in order to retrieve the average prices of each commodity).

```
newdata = data.frame(Coal = 67 - mean(COMMODITY_PRICE$Coal), Natural_Gas = 7 - mean(COMMODITY_PRICE$Natu
Banana = 0.95 - mean(COMMODITY_PRICE$Banana), Orange = 0.79 - mean(COMMODITY_PRICE$Orange), Shrimps = 1
Potassium = 261 - mean(COMMODITY_PRICE$Potassium), Zinc = 1945 - mean(COMMODITY_PRICE$Zinc), Gold = 952
pred = predict(Best_Model, newdata = newdata, se.fit = TRUE)
 y_hat = pred$fit
 se_y_hat = pred$se.fit
 df = n- p- 1
 15
```

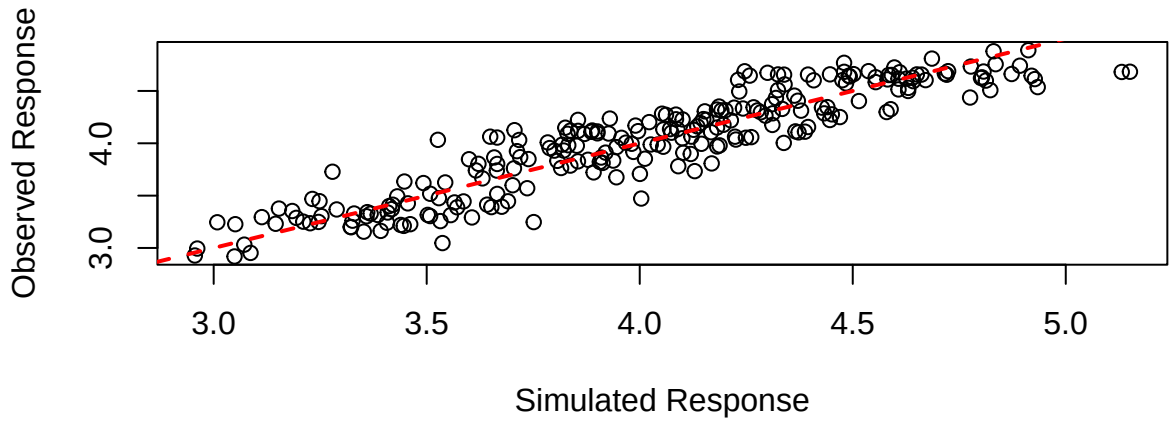
```
## [1] 15
```

```
t_value= qt(0.975,df)
 lower_bound<-y_hat-t_value *se_y_hat
 upper_bound<-y_hat + t_value *se_y_hat
 c(lower_bound,y_hat,upper_bound)
```

```
##        1        1        1
## 3.748462 3.828446 3.908429
```

(se.fit = TRUE is used to get the standard error). So the predicted price for a barrel of crude oil is exp(3.83) = 46 USD per barrel. Let's now simulate n data points from the fitted regression model, assuming the estimated parameters as the true parameters

```
RSS = summ$deviance
sigma = sqrt(RSS/(n - p - 1))
set.seed(123)
n_sim = nrow(COMMODITY_PRICE)
Y_pred = predict(Best_Model)
Y_sim = Y_pred + rnorm(n_sim, mean = 0, sd = sigma)
Y_obs = log(COMMODITY_PRICE$Crude_Oil)
plot(Y_sim, Y_obs, xlab = "Simulated Response", ylab = "Observed Response")
abline(0, 1, col="red", lwd=2, lty=2)
```



The prediction is accurate, with slight dispersion around the regression, suggesting a good model fit.