Project 1 Econ 104

Aiden Mayhood, Han Zhang, Tiantian Zhao, Megan Oh

2022-10-14

```
library('AER')
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

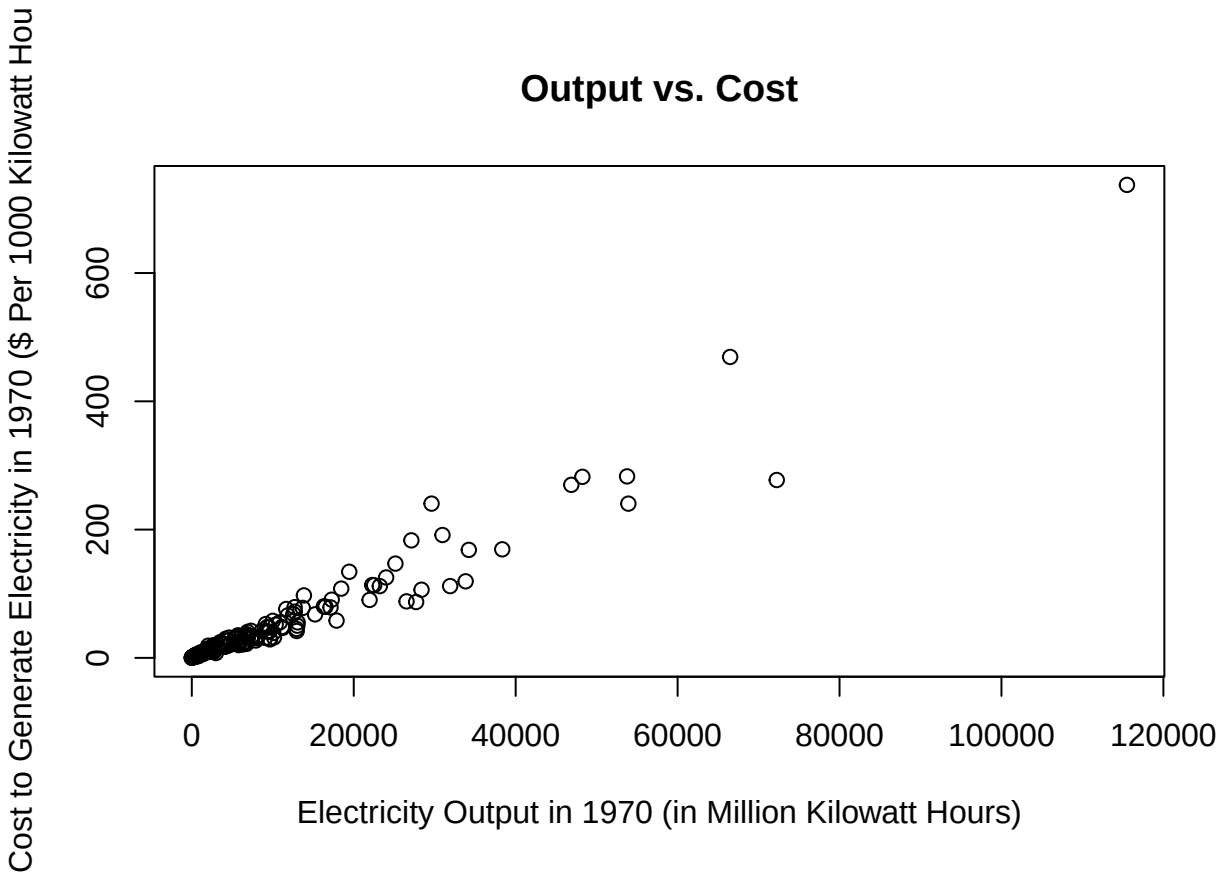
```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
data("Electricity1970")
attach(Electricity1970)
View(Electricity1970)
cost <- Electricity1970$cost
output <- Electricity1970$output
labor <- Electricity1970$labor
laborshare <- Electricity1970$laborshare
capital <- Electricity1970$capital
capitalshare <- Electricity1970$capitalshare
fuel <- Electricity1970$fuel
fuelshare <- Electricity1970$fuelshare
```

Above, we created vectors of cost, output, labor, labor share, capital, capital share, fuel, and fuel share from the 'Electricity1970' data from the AER library. The data consists of the total cost of electricity in 1970, the total output of electricity in 1970, the wage rate in 1970, the capital price index, the cost share for labor in 1970, the cost share for capital in 1970, the fuel price in 1970, and the cost share for fuel in 1970 for 158 firms. The data was collected in an attempt to estimate economies of scale for U.S. firms producing electric power. The observations collected in the data are cross-sectional, as it consists of multiple firms over one period of time.

```
plot(x = output, y = cost, xlab = "Electricity Output in 1970 (in Million Kilowatt Hours)", ylab = "Cos
```



Above, we plotted electricity output against cost to generate electricity. The output of electricity will be a contributing factor in the cost of producing the electricity, as costs will vary depending on how much electricity is produced.

This scatterplot shows a strong, positive, linear association between electricity output and the cost to generate electricity. There appears to be one potential outlier in the top right of the scatterplot. Without the outlier, it may appear to be more of a more clustered data set.

```
plot(x = labor, y = cost, xlab = "Wage Rate in 1970 (in $)", ylab = "Cost to Generate Electricity in 19
```

Above, we plotted the wage rate against cost to generate electricity. The cost of generating electricity will vary depending on how much employees of energy firms are getting paid.

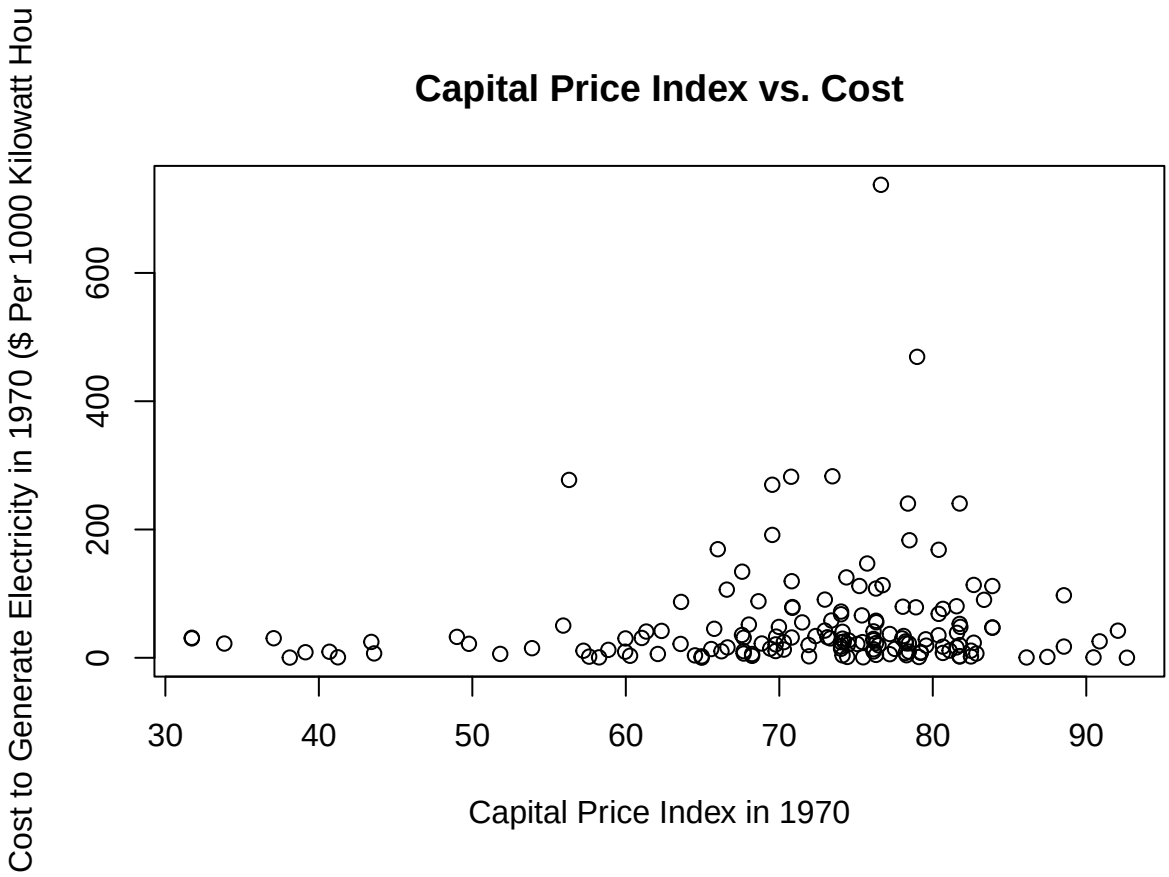
This scatterplot shows a weak, flat, linear association between the wage rate and the cost to generate electricity. It appears that as the wage rate increases, the cost to generate electricity does not increase. There are about 15 or so points on the scatterplot that don't appear to follow this trend, causing a bell-curve like shape in the distribution.

```
plot(x = laborshare, y = cost, xlab = "Cost Share for Labor in 1970 (%)", ylab = "Cost to Generate Elec
```

Above, we plotted the cost share of generating electricity attributable to labor against the cost to generate electricity. The cost of generating electricity will depend on how much of a share of the cost is attributable to labor costs.

This scatterplot shows a weak, flat, linear association between the cost share of labor and the cost to generate electricity. It appears as the cost share attributed to labor increases, the cost to generate electricity does not increase. There are about 15 points that do not follow this trend, causing a right-tail in the distribution.

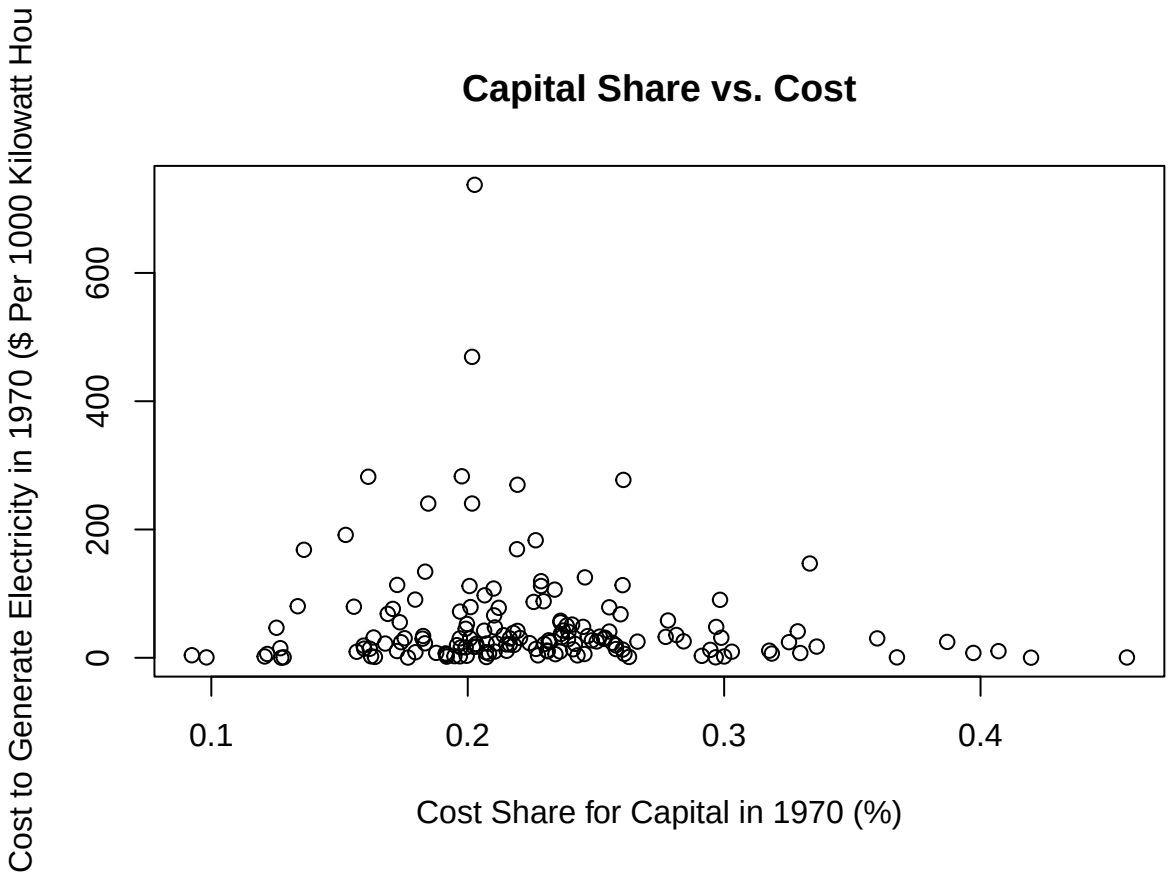
```
plot(x = capital, y = cost, xlab = "Capital Price Index in 1970", ylab = "Cost to Generate Electricity
```



Above, we plotted the capital price index against the cost to generate electricity. The cost of generating electricity will depend on the prices of fixed assets used to generate the electricity.

This scatterplot shows a weak, flat, linear association betwee the capital price index and the cost to generate electricity. It appears that as the price of capital increases, the cost to generate electricity does not increase much. There are about 15 or so points that do not follow this trend, causing a left-tail on the distribution.

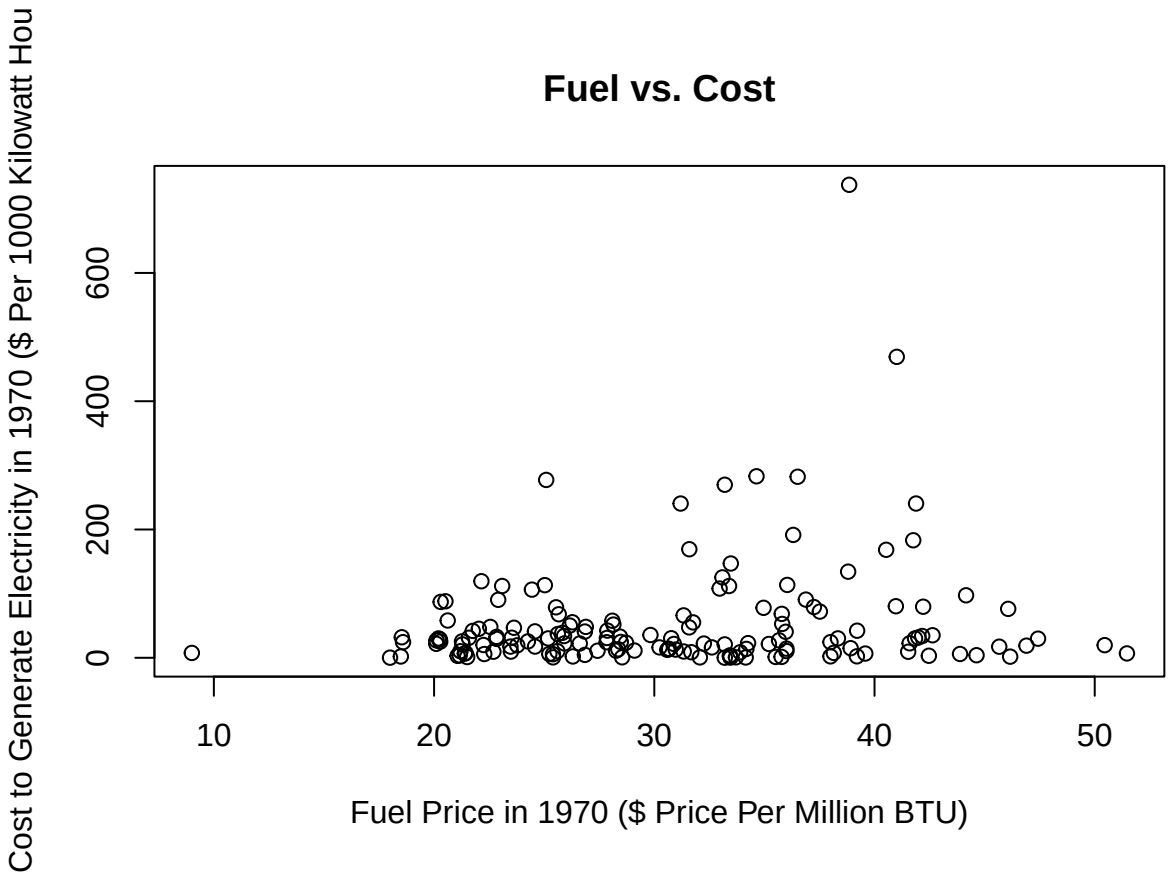
```
plot(x = capitalshare, y = cost, xlab = "Cost Share for Capital in 1970 (%)", ylab = "Cost to Generate
```



Above, we plotted the cost share of generating electricity attributable to capital against the cost to generate electricity. The cost of generating electricity will depend on how much of a share of the cost is attributable to capital costs.

This scatterplot shows a weak, flat, linear association between the cost share for capital and the cost to generate electricity. As the cost share for capital increases, there doesn't seem to be much of an increase in the cost to generate electricity. There are a few outlier points that cause the dataset to spike between the 10-30% range on the x-axis.

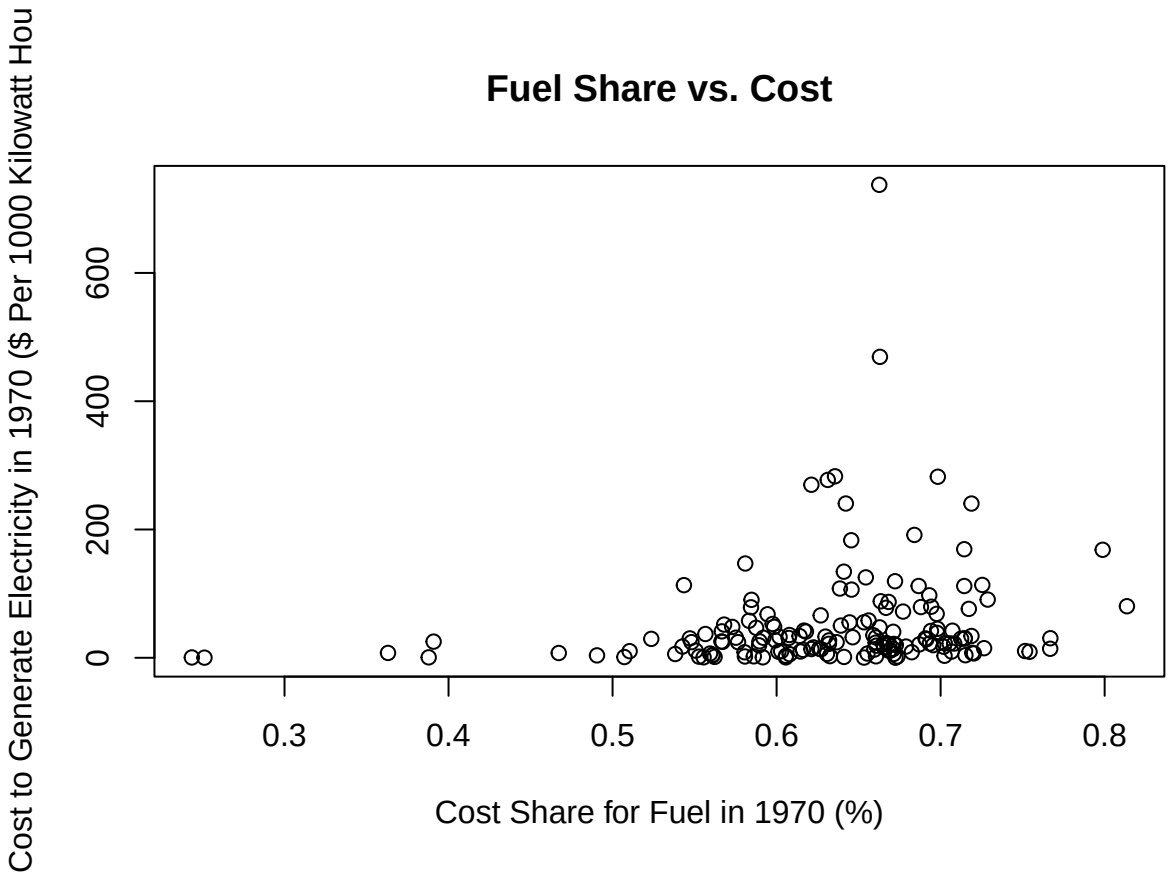
```
plot(x = fuel, y = cost, xlab = "Fuel Price in 1970 ($ Price Per Million BTU)", ylab = "Cost to Generat
```



Above, we plotted fuel price against the cost to generate electricity. The cost of producing electricity will vary depending on the cost of fuel needed to produce electricity.

This scatterplot shows a bell-curve like, flat, nonlinear association. It appears that as fuel prices increase, the cost to generate electricity does not increase or increases very little. This is a hard data distribution to describe. Most of the data follows a flat, nonlinear distribution, but many data points sit atop, causing a round top. A few points stick out, possibly indicating outliers.

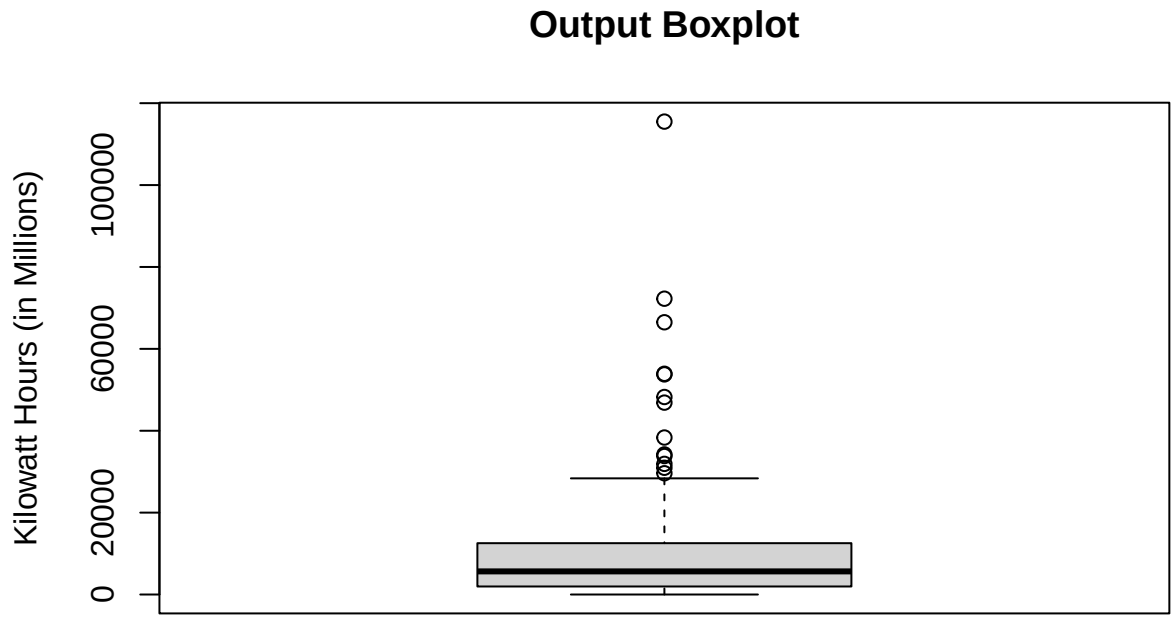
```
plot(x = fuelshare, y = cost, xlab = "Cost Share for Fuel in 1970 (%)", ylab = "Cost to Generate Electri
```



Above, we plotted the cost share of generating electricity attributable to fuel against the cost to generate electricity. The cost of generating electricity will depend on how much of a share of the cost is attributable to the price of fuel.

This scatterplot shows a weak, slightly positive, nonlinear association between the cost share for fuel and the cost to generate electricity. Most of the data is clustered between 50-70% on the x-axis, with a high spike in the 65% range. It then quickly drops back down, and the data becomes more flat again.

```
boxplot(output, main = "Output Boxplot", ylab = "Kilowatt Hours (in Millions)")
```

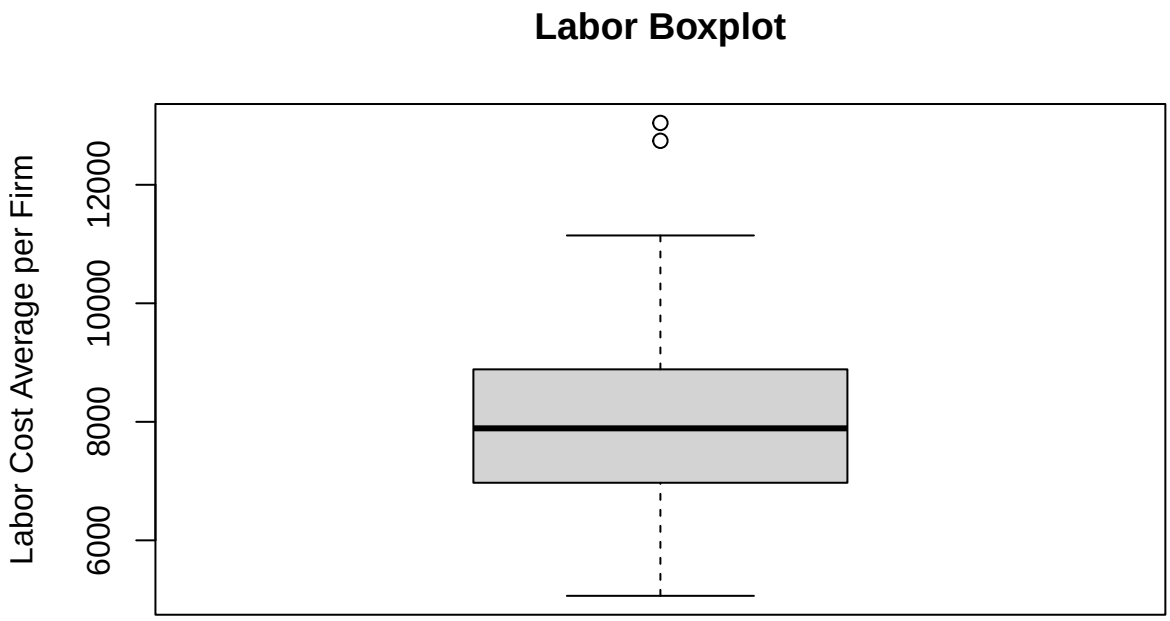



```
summary(output)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       4    1971    5646   10469   12366  115500
```

Above, the boxplot has a minimum of 4, a maximum of 115,500, a median of 5,646, a mean of 10,469, a 1st quartile of 1971, and a 3rd quartile of 12,366. The whiskers extend from 0 to 27,958.8. Observations above 27,958.8 are potential outliers in the data.

```
boxplot(labor, main = "Labor Boxplot", ylab = "Labor Cost Average per Firm")
```

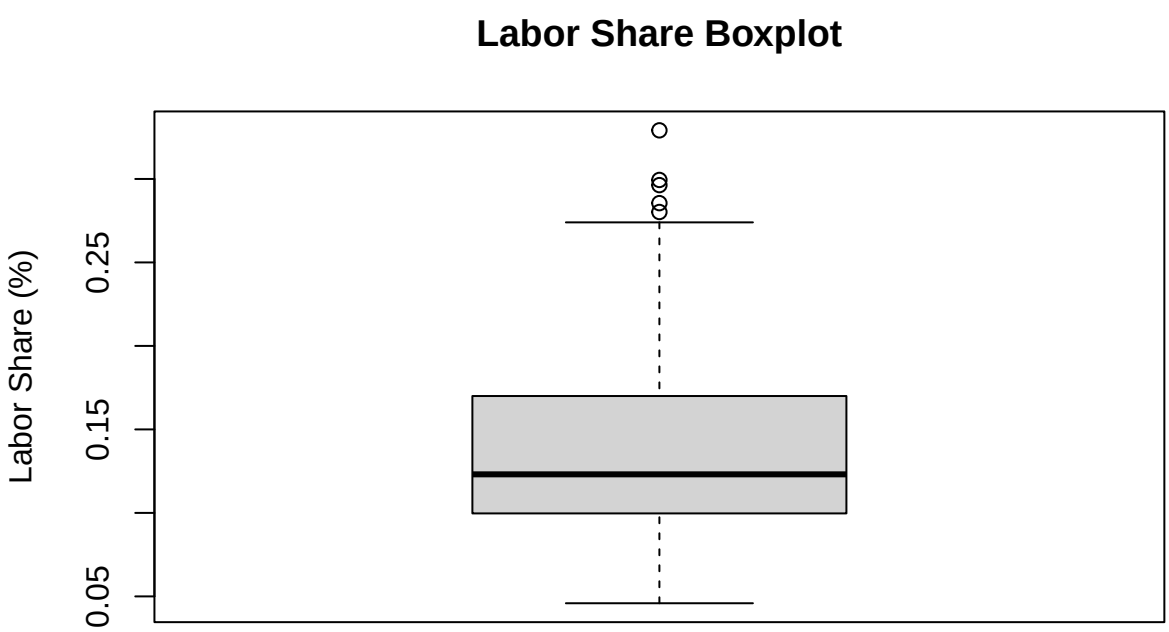



```
summary(labor)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5063    6975    7890    8002    8855   13044
```

Above, the boxplot has a minimum of 5063, a maximum of 13,044, a median of 7,890, a mean of 8,002, a 1st quartile of 6,975, and a 3rd quartile of 8,855. The whiskers extend from 4,155 to 11,675. Observations below and above these whiskers are potential outliers in the data, but there are only potential outliers above the 11,675 value. Two of them are observable in the data.

```
boxplot(laborshare, main = "Labor Share Boxplot", ylab = "Labor Share (%)")
```

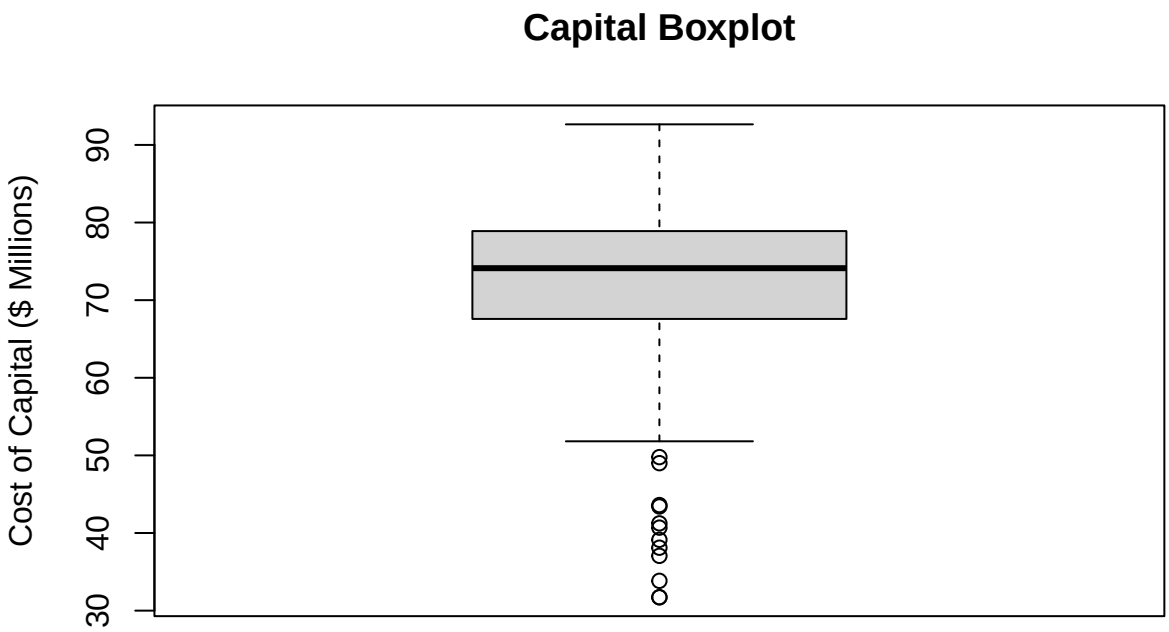



```
summary(laborshare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04590 0.09972 0.12310 0.13897 0.16980 0.32910
```

Above, the boxplot has a minimum of 0.0459, a maximum of 0.32910, a median of 0.12310, a mean of 0.13897, a 1st quartile of 0.09972, and a 3rd quartile of 0.16980. The whiskers extend from 0 to 0.27492. Observations below 0 and above 0.27942 are potential outliers in the data. There seems to be five potential outliers in the data.

```
boxplot(capital, main = "Capital Boxplot", ylab = "Cost of Capital ($ Millions)")
```

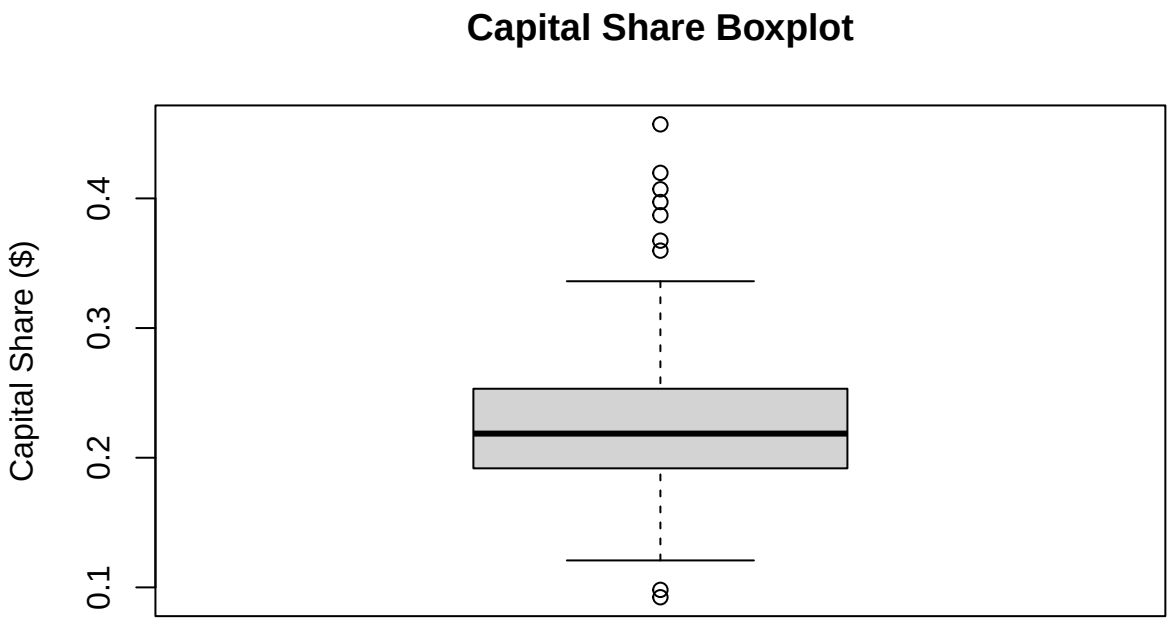



```
summary(capital)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31.73   67.61   74.12   71.42   78.79   92.65
```

Above, the boxplot has a minimum of 31.73, a maximum of 92.65, a median of 74.12, a mean of 71.42, a 1st quartile of 67.61, and a 3rd quartile of 78.79. The whiskers extend from 50.84 to 95.56. Observations below 50.84 and above 95.56 are potential outliers in the data. The only visible potential outliers in the data are below the 50.84 whisker. There appears to be 12 potential outliers.

```
boxplot(capitalshare, main = "Capital Share Boxplot", ylab = "Capital Share ($)")
```

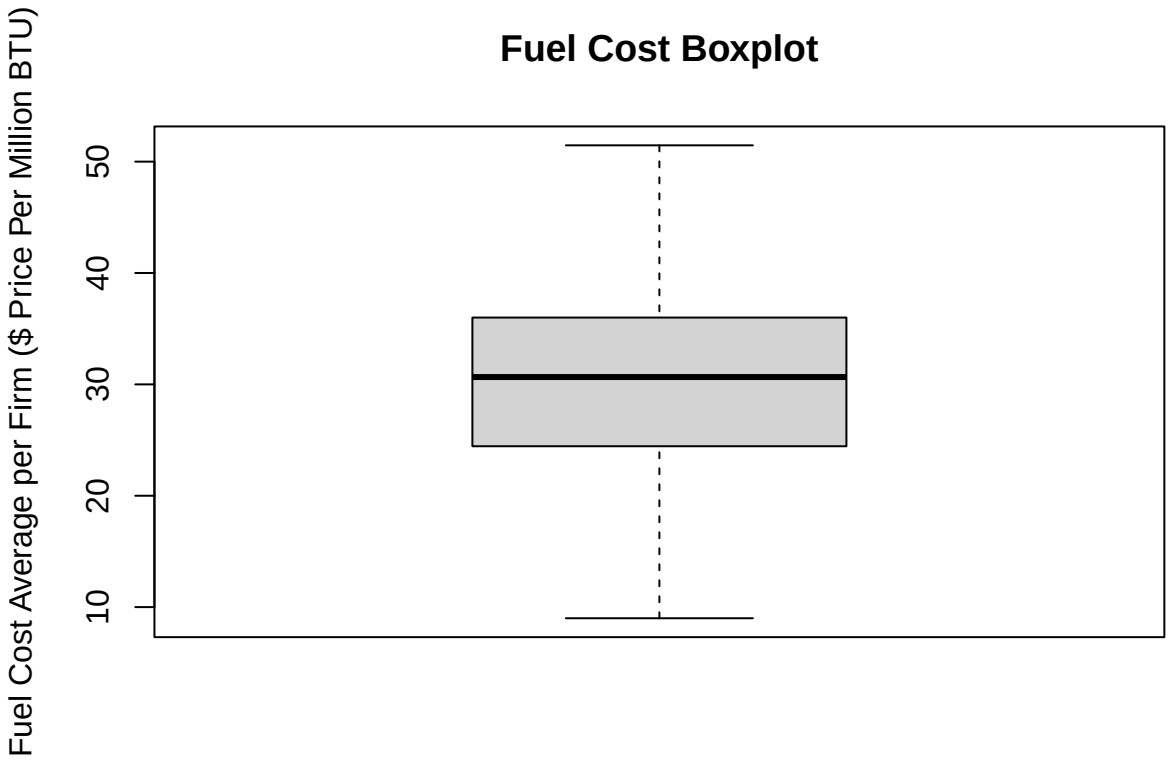



```
summary(capitalshare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0924  0.1925  0.2186  0.2264  0.2528  0.4571
```

Above, the boxplot has a minimum of 0.0924, a maximum of 0.4571, a median of 0.2186, a mean of 0.2264, a 1st quartile of 0.1925, and a 3rd quartile of 0.2528. The whiskers extend from 0.10205 to 0.30605. Observations below 0.10205 and above 0.30605 are potential outliers in the data. There are two potential outliers below 0.10205 and six potential outliers above 0.30605.

```
boxplot(fuel, main = "Fuel Cost Boxplot", ylab = "Fuel Cost Average per Firm ($ Price Per Million BTU)"
```

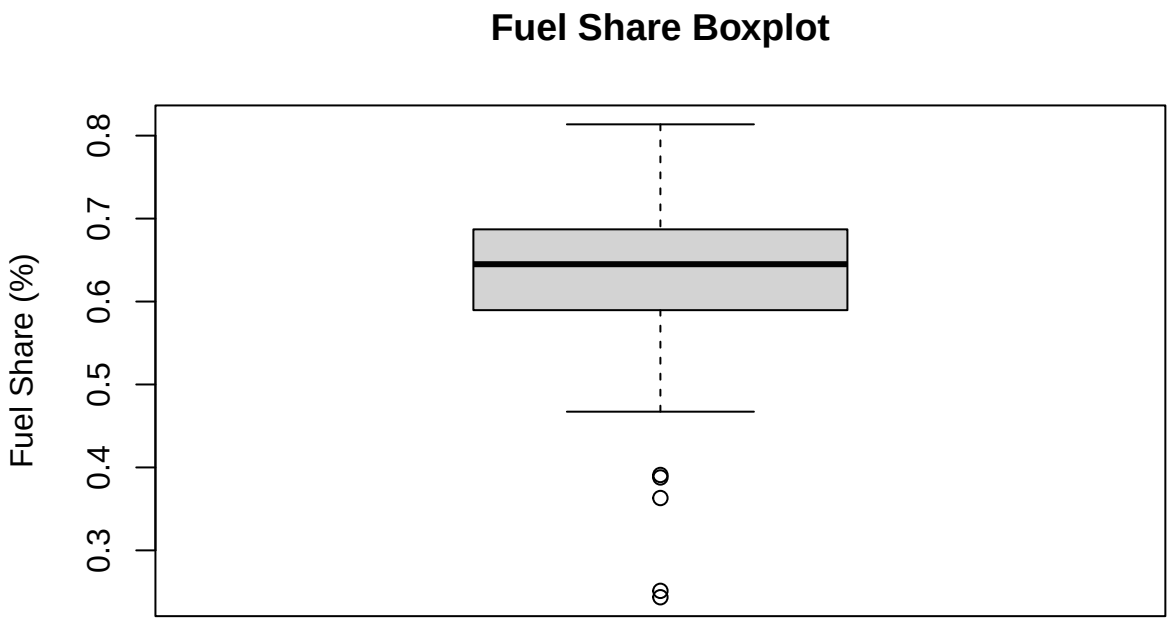



```
summary(fuel)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   24.48   30.66   30.75   36.00   51.46
```

Above, the boxplot has a minimum of 9, a maximum of 51.46, a median of 30.66, a mean of 30.75, a 1st quartile of 24.48, and a 3rd quartile of 36. The whiskers extend from 0 to 53.28. Observations below 0 and above 53.28 are potential outliers in the data. There doesn't seem to be any potential outliers.

```
boxplot(fuelshare, main = "Fuel Share Boxplot", ylab = "Fuel Share (%)")
```

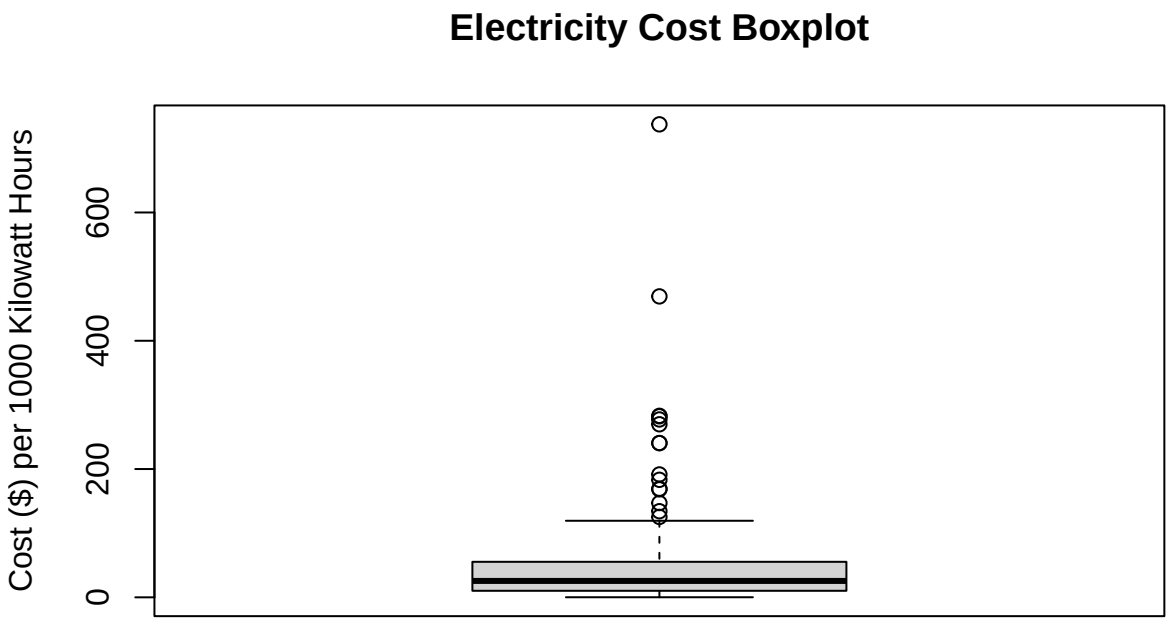



```
summary(fuelshare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2435  0.5901  0.6450  0.6324  0.6869  0.8136
```

Above, the boxplot has a minimum of 0.2435, a maximum of 0.8136, a median of 0.6450, a mean of 0.6324, a 1st quartile of 0.5901, and a 3rd quartile of 0.6869. The whiskers extend from 0.4449 to 0.8321. Observations above 0.8321 and below 0.4449 are potential outliers in the data.There are five potential outliers below 0.4449.

```
boxplot(cost, main ="Electricity Cost Boxplot", ylab = "Cost ($) per 1000 Kilowatt Hours")
```




```
summary(cost)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.1304  10.2210  25.5454  53.2700  55.3159 737.4088
```

Above, the boxplot has a minimum of 0.1304, a maximum of 737.4088, a median of 25.5454, a mean of 53.2700, a 1st quartile of 10.2210, and a 3rd quartile of 55.3159. The whiskers extend from 0 to 122.95825. Observations below 0 and above 122.95825 are potential outliers in the data. There appears to be 13 potential outliers above 122.95825.

```
multiregbasemodel <- lm(cost ~ output + labor + laborshare + capital + capitalshare + fuel + fuelshare)
multiregbasemodel
```

```
##
## Call:
## lm(formula = cost ~ output + labor + laborshare + capital + capitalshare +
##     fuel + fuelshare)
##
## Coefficients:
##  (Intercept)        output         labor    laborshare       capital
##   -97.928494      0.005554      0.002732     74.047011      0.255396
```

```
## capitalshare          fuel     fuelshare
##    48.154956      1.406297    -18.156380
```

```
attributes(multiregbasemodel)
```

```
## $names
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
##
## $class
## [1] "lm"
```

```
summary(multiregbasemodel)
```

```
##
## Call:
## lm(formula = cost ~ output + labor + laborshare + capital + capitalshare +
##     fuel + fuelshare)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -105.088   -5.144    1.709    7.014   88.857
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.793e+01  7.552e+01  -1.297   0.1967
## output        5.554e-03  1.024e-04  54.235  < 2e-16 ***
## labor         2.732e-03  1.178e-03   2.319   0.0218 *
## laborshare    7.405e+01  7.808e+01   0.948   0.3445
## capital       2.554e-01  1.338e-01   1.909   0.0582 .
## capitalshare  4.815e+01  7.601e+01   0.634   0.5274
## fuel          1.406e+00  2.378e-01   5.914 2.17e-08 ***
## fuelshare    -1.816e+01  7.422e+01  -0.245   0.8071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.88 on 150 degrees of freedom
## Multiple R-squared:  0.9551, Adjusted R-squared:  0.953
## F-statistic: 455.4 on 7 and 150 DF,  p-value: < 2.2e-16
```

Below, we have created a multiple linear regression model that includes all the main effects only, with no interactions nor higher order terms. This is the baseline model to be adapted. Using alpha = 0.05, if the p-value for a variable is less than 0.05, the significance level, then we will deem it statistically significant. Using this criteria, we would remove the "laborshare" variable, the "capitalshare" variable, the "fuel" variable", and the "capital" variable while keeping "output", "labor", and "fuel". However, we have decided on not removing "capital" while keeping the other input variables, as its p-value is only 0.0082 above the cut-off for statistical significance at alpha = 0.05. Also, expenses on capital make sense as a determining factor of how much electricity cost will be. Some firms who are really inefficient with capital expenditures might have higher electricity costs, while others who are really efficient with capital costs might have lower electricity costs. Therefore, we have decided to include "capital" in the regression.

We have now ran a new regression excluding the cost share variables. Below is the new regression.

```
multiregbasemodel2 <- lm(cost ~ output + labor + capital + fuel)
multiregbasemodel2
```

```
##
## Call:
## lm(formula = cost ~ output + labor + capital + fuel)
##
## Coefficients:
## (Intercept)       output        labor      capital         fuel
##  -84.555714     0.005471     0.003239     0.303479     1.071896
```

```
attributes(multiregbasemodel2)
```

```
## $names
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
##
## $class
## [1] "lm"
```

```
summary(multiregbasemodel2)
```

```
##
## Call:
## lm(formula = cost ~ output + labor + capital + fuel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.438   -5.590    1.279    6.649   97.813
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.456e+01  1.511e+01  -5.595 9.90e-08 ***
## output       5.471e-03  1.036e-04  52.823  < 2e-16 ***
## labor        3.239e-03  1.207e-03   2.684  0.00807 **
## capital      3.035e-01  1.370e-01   2.215  0.02822 *
## fuel         1.072e+00  2.064e-01   5.194 6.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.59 on 153 degrees of freedom
## Multiple R-squared:  0.9507, Adjusted R-squared:  0.9494
## F-statistic: 737.3 on 4 and 153 DF,  p-value: < 2.2e-16
```

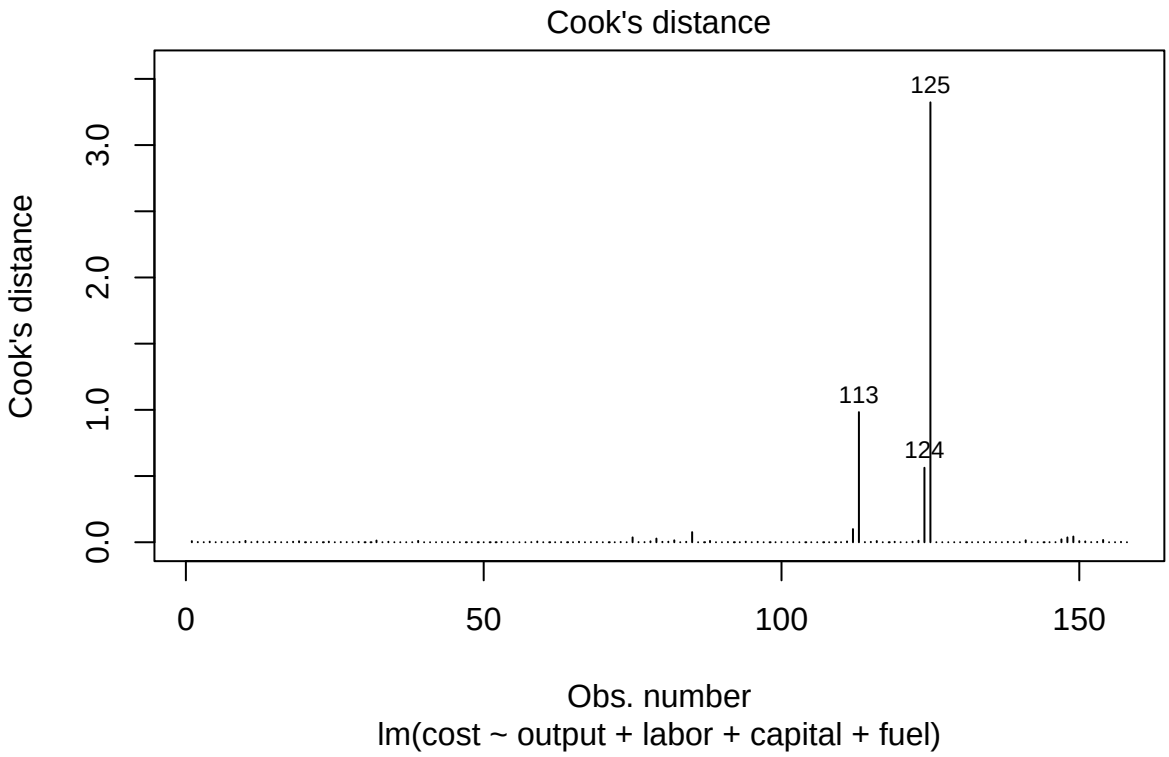
Electricity Cost (y) = -84.56 + 0.005471outputvar + 0.003239laborvar + 0.3035capitalvar + 1.072fuelvar

In the new multiregression model above, every one of our variables is statistically significant with p-values less than our alpha = 0.05. With this new model, if we are given information on how much electricity output, labor expenditure, capital expenditure, and fuel expenditure a firm had, we could predict how much it cost to produce electricity for that firm. Positive coefficients for each of our estimates shows that an increase in each independent variable leads to an increase in our dependent variable. There is a positive relationship there.

```
multiregbasemodel2
```

```
##
## Call:
## lm(formula = cost ~ output + labor + capital + fuel)
##
## Coefficients:
## (Intercept)       output        labor      capital         fuel
##  -84.555714     0.005471     0.003239     0.303479     1.071896
```

```
cooksD <- cooks.distance(multiregbasemodel2)
plot(multiregbasemodel2, which = 4)
```




```
influential_obs <- as.numeric(names(cooksD) [(cooksD > (2*(4+1)/158))])

library(car)
outlierTest(lm(cost ~ output + labor + capital + fuel, data = Electricity1970))
```

```
##      rstudent unadjusted p-value Bonferroni p
## 185  6.881391         1.4538e-10   2.2970e-08
## 173 -6.223128         4.5163e-09   7.1358e-07
## 184  5.422743         2.2617e-07   3.5735e-05
```

```
nElectricity1970 = Electricity1970[-c(113,125,124),]
newmultiregbasemodel2 <-lm(cost ~ output + labor + capital + fuel, data = nElectricity1970)
summary(newmultiregbasemodel2)
```

```
##
## Call:
## lm(formula = cost ~ output + labor + capital + fuel, data = nElectricity1970)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.788  -4.484  -0.062   4.566  77.819
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

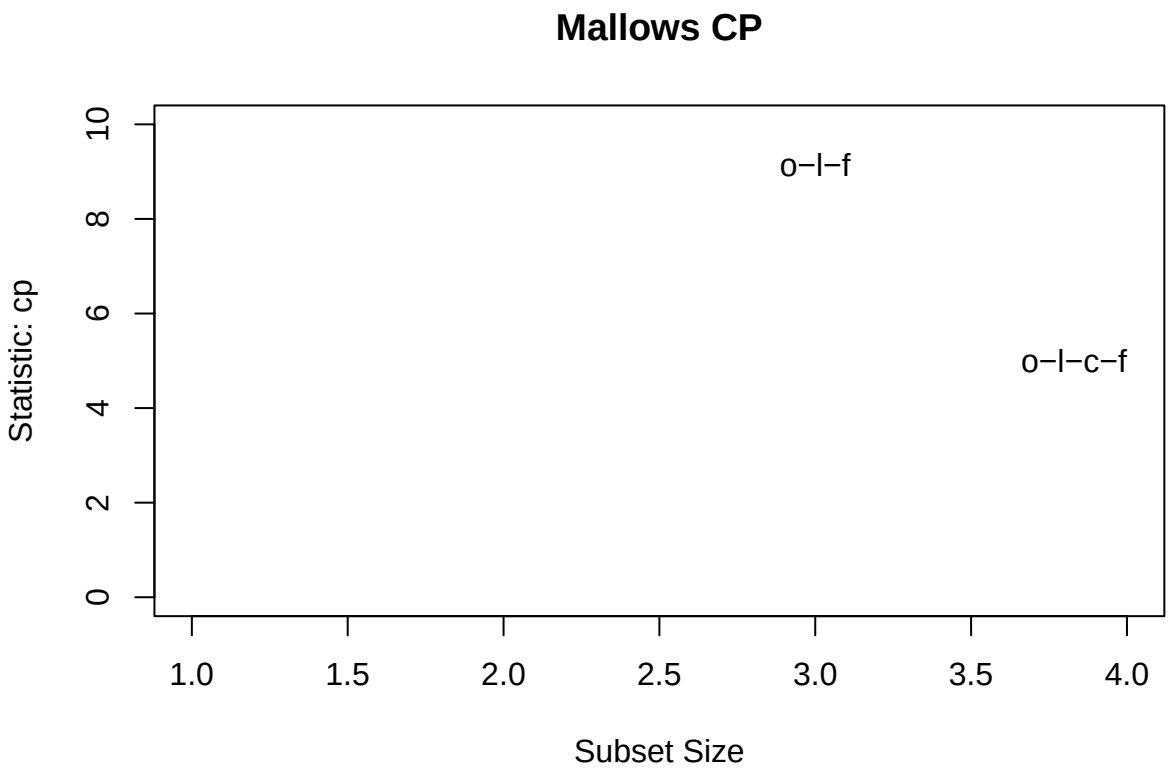
```
## (Intercept) -6.574e+01  1.004e+01  -6.548 8.72e-10 ***
## output       4.964e-03  9.679e-05  51.280  < 2e-16 ***
## labor        2.885e-03  7.944e-04   3.632 0.000385 ***
## capital      2.254e-01  9.086e-02   2.481 0.014202 *
## fuel         8.630e-01  1.361e-01   6.339 2.57e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.84 on 150 degrees of freedom
## Multiple R-squared:  0.9492, Adjusted R-squared:  0.9479
## F-statistic: 701.4 on 4 and 150 DF,  p-value: < 2.2e-16
```

After using Cook's distance to discover potential outliers, there are three data points that are highly influential in our data set. These are observations #113, #124, and #125. These observations are above the 0.063 threshold we determined in the Cook's distance function (2*(5/148)). We have removed these outliers from our new model "newmultiregbasemodel2", as the Cook's distance values are significant. These are natural observations and do not seem to be mistakes in the data, but with high residuals, they are impacting the prediction power of our regression model.

```
library(devtools)
```

```
## Loading required package: usethis
```

```
library(car)
library(AER)
library(broom)
library(PoEdata)
library(leaps)

ss = regsubsets(cost ~ output + labor + capital + fuel, method = c("exhaustive"), nbest = 3, data = nEl
subsets(ss, statistic = "cp", legend = F, main = "Mallows CP", col = "steelblue4", ylim = c(0,10))
```



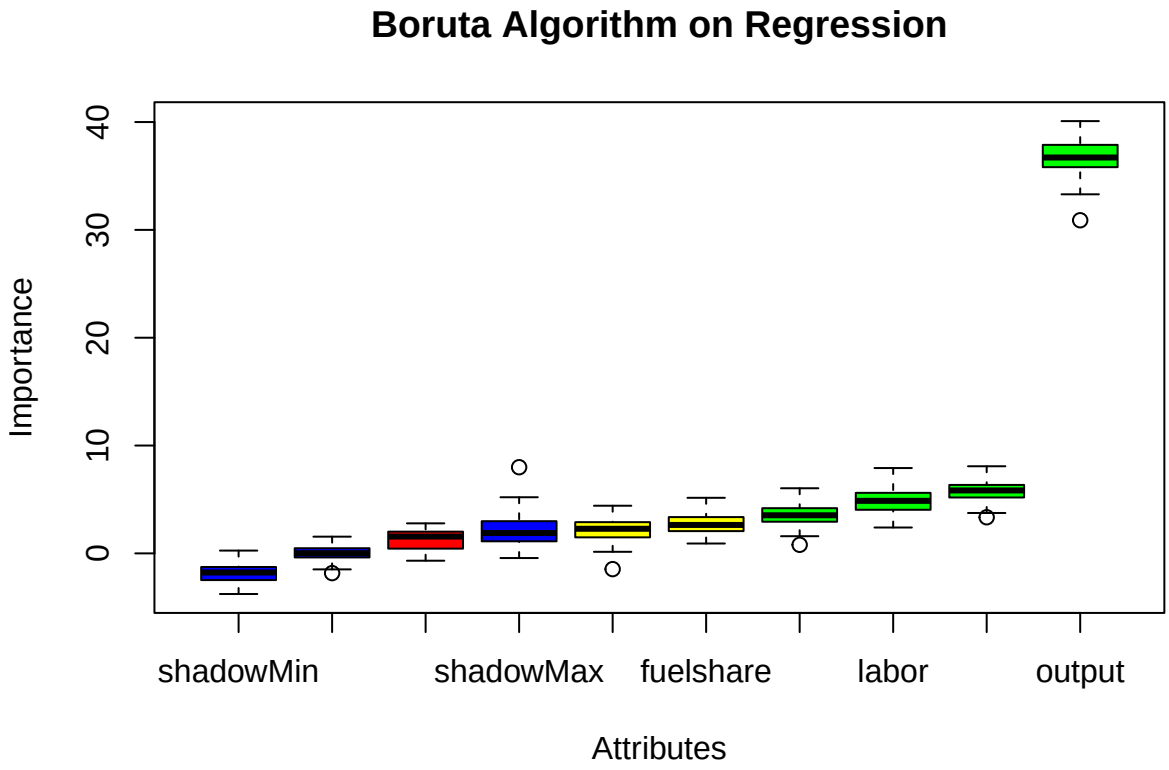

```
##         Abbreviation
## output             o
## labor              l
## capital            c
## fuel               f
```

```
library(Boruta)
Bor.res <- Boruta(cost ~., data = nElectricity1970, doTrace = 1)
```

```
## After 10 iterations, +0.62 secs:

##  confirmed 2 attributes: laborshare, output;

##  still have 5 attributes left.

## After 14 iterations, +0.84 secs:

##  rejected 1 attribute: capitalshare;

##  still have 4 attributes left.

## After 29 iterations, +1.5 secs:
```

```
##  confirmed 1 attribute: labor;

##  still have 3 attributes left.

## After 53 iterations, +2.5 secs:

##  confirmed 1 attribute: fuel;

##  still have 2 attributes left.
```

```
plot(Bor.res, main = "Boruta Algorithm on Regression")
```



The lowest Mallows CP is the one which includes capital in the model. Therefore, we will keep all four explanatory variables (output,labor, capital, and fuel) in the model.

The Boruta algorithm tells us that output is a variable of great importance for our model. We assume that electricity output is probably so important because if output increases drastically, costs to create such output increase drastically in the same scale. Fuel, labor, and capital costs do as well, but they are less predictive and may change due to circumstances outside of an increase in energy output, like economic recessions or pandemics, for instance.

```
n4Electricity1970 <- nElectricity1970[, -c(4,6)]
VIFReg <- lm(cost~., data = n4Electricity1970)
summary(VIFReg)
```

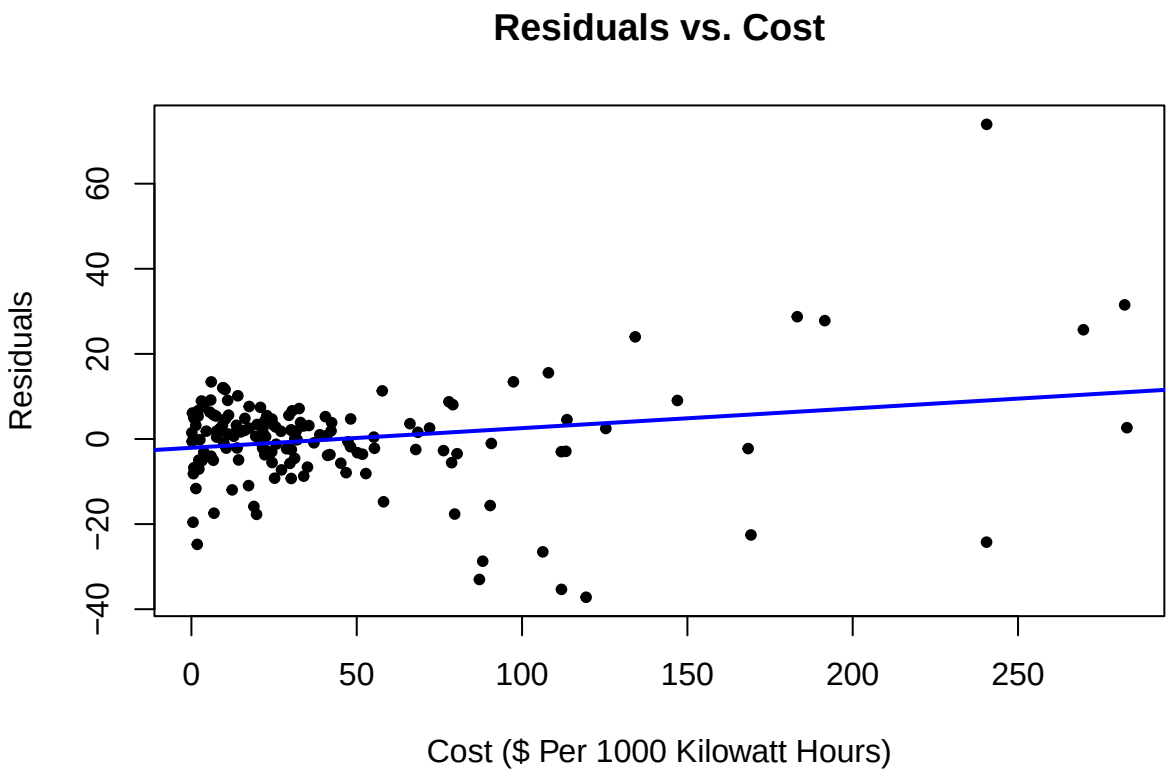
```
##
## Call:
## lm(formula = cost ~ ., data = n4Electricity1970)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.190  -3.942   0.995   4.719  73.940
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.569e+01  1.242e+01  -2.874 0.004641 **
## output       5.070e-03  9.675e-05  52.400  < 2e-16 ***
## labor        2.504e-03  7.671e-04   3.265 0.001359 **
## capital      1.857e-01  8.762e-02   2.119 0.035726 *
## fuel         1.144e+00  1.496e-01   7.645 2.37e-12 ***
## fuelshare   -5.337e+01  1.396e+01  -3.824 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.3 on 149 degrees of freedom
## Multiple R-squared:  0.9538, Adjusted R-squared:  0.9522
## F-statistic:   615 on 5 and 149 DF,  p-value: < 2.2e-16
```

```
vif(VIFReg)
```

```
##    output     labor   capital      fuel fuelshare
##  1.108766  1.188537  1.127805  1.435819  1.401913
```

The variance inflation factors from the VIF test do not exceed 5, so there is likely no multicollinearity.

```
plot(n4Electricity1970$cost, resid(VIFReg), main = "Residuals vs. Cost" ,xlab = "Cost ($ Per 1000 Kilowa
abline(lm(resid(VIFReg) ~ n4Electricity1970$cost), lwd = 2, col = "blue")
```



Above, the cost to generate electricity is plotted on the x-axis, and the residuals are plotted on the y-axis. As the electricity generation cost starts to increase around $50 per 1000 kilowatt hours, the residuals begin to spread out more in a cone-like shape, hinting that there may be heteroskedasticity present.

```
resettest(VIFReg, power = 2:3, type = "fitted")
```

```
##
##  RESET test
##
## data:  VIFReg
## RESET = 5.9336, df1 = 2, df2 = 147, p-value = 0.003325
```

```
resetmodel <- lm(cost ~ output + labor + capital + fuel^2)
summary(resetmodel)
```

```
##
## Call:
## lm(formula = cost ~ output + labor + capital + fuel^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -101.438   -5.590    1.279    6.649   97.813
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.456e+01  1.511e+01  -5.595 9.90e-08 ***
## output       5.471e-03  1.036e-04  52.823  < 2e-16 ***
## labor        3.239e-03  1.207e-03   2.684  0.00807 **
## capital      3.035e-01  1.370e-01   2.215  0.02822 *
## fuel         1.072e+00  2.064e-01   5.194 6.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.59 on 153 degrees of freedom
## Multiple R-squared:  0.9507, Adjusted R-squared:  0.9494
## F-statistic: 737.3 on 4 and 153 DF,  p-value: < 2.2e-16
```

```
summary(VIFReg)
```

```
##
## Call:
## lm(formula = cost ~ ., data = n4Electricity1970)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.190  -3.942   0.995   4.719  73.940
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.569e+01  1.242e+01  -2.874 0.004641 **
## output       5.070e-03  9.675e-05  52.400  < 2e-16 ***
## labor        2.504e-03  7.671e-04   3.265 0.001359 **
## capital      1.857e-01  8.762e-02   2.119 0.035726 *
## fuel         1.144e+00  1.496e-01   7.645 2.37e-12 ***
## fuelshare   -5.337e+01  1.396e+01  -3.824 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.3 on 149 degrees of freedom
## Multiple R-squared:  0.9538, Adjusted R-squared:  0.9522
## F-statistic:   615 on 5 and 149 DF,  p-value: < 2.2e-16
```

With alpha = 0.05, our p-value in the Ramsey RESET test is statistically significant. Therefore, adding a 2nd or 3rd degree term may improve our model. However, when compared to the VIFReg model, the VIFReg model has a higher R^2 without added exponents or manipulations. Therefore, we conclude that we will keep our model unchanged, even with the results of the RESET test.

```
bptest(VIFReg)
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  VIFReg
## BP = 32.259, df = 5, p-value = 5.278e-06
```

```
summary(VIFReg)
```

```
##
## Call:
## lm(formula = cost ~ ., data = n4Electricity1970)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.190  -3.942   0.995   4.719  73.940
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.569e+01  1.242e+01  -2.874 0.004641 **
## output       5.070e-03  9.675e-05  52.400  < 2e-16 ***
## labor        2.504e-03  7.671e-04   3.265 0.001359 **
## capital      1.857e-01  8.762e-02   2.119 0.035726 *
## fuel         1.144e+00  1.496e-01   7.645 2.37e-12 ***
## fuelshare   -5.337e+01  1.396e+01  -3.824 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.3 on 149 degrees of freedom
## Multiple R-squared:  0.9538, Adjusted R-squared:  0.9522
## F-statistic:   615 on 5 and 149 DF,  p-value: < 2.2e-16
```

```
library(lmtest)
library(sandwich)
coeftest(VIFReg, vcov = vcovHC(VIFReg, "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -3.5687e+01  1.0464e+01 -3.4102 0.0008354 ***
## output       5.0698e-03  1.8660e-04 27.1692 < 2.2e-16 ***
## labor        2.5044e-03  6.9406e-04  3.6083 0.0004201 ***
## capital      1.8569e-01  6.2985e-02  2.9481 0.0037129 **
## fuel         1.1436e+00  1.8677e-01  6.1229 7.775e-09 ***
## fuelshare   -5.3374e+01  1.5189e+01 -3.5140 0.0005848 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With alpha = 0.05, our p-value in the Breusch-Pagan test is statistically significant. Therefore, we conclude that there most likely heteroskedasticity in the present model. Comparing our heteroskedsaticity-consistent standard errors (robust standard errors) with our standard errors in the VIFReg model, the standard errors barely change. The standard errors in the model are near zero. We haven't addressed that our least squares estimators are no longer BLUE. But, since we have a sample size of 158 observations, our standard errors should be precise and the variance of our estimators should still allow us to get precise estimates from our model.

```
AIC(multiregbasemodel, multiregbasemodel2, VIFReg)
```

```
## Warning in AIC.default(multiregbasemodel, multiregbasemodel2, VIFReg): models
## are not all fitted to the same number of observations
```

```
##                     df      AIC
## multiregbasemodel    9 1386.646
## multiregbasemodel2   6 1395.342
## VIFReg               7 1225.621
```

```
BIC(multiregbasemodel, multiregbasemodel2, VIFReg)
```

```
## Warning in BIC.default(multiregbasemodel, multiregbasemodel2, VIFReg): models
## are not all fitted to the same number of observations
```

```
##                     df      BIC
## multiregbasemodel    9 1414.209
## multiregbasemodel2   6 1413.718
## VIFReg               7 1246.925
```

Given the lowest Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) is with our VIFReg model, having additional variables in our model gives a penalty to "multiregbasemodel" and "multiregbasemodel2". Seeking to minimize SSE, we conclude that VIFReg is the best model compared to multiregbasemodel and multiregbasemodel2.

Skip Question 10

We began with a dataset to predict the costs of producing electricity for a given firm. The explanatory variables to explain the cost were output, labor, labor share, capital, capital share, fuel, and fuel share. The labor share, capital share, and fuel share explanatory variables were statistically insignificant and were removed from the model. Further, after looking at boxplots and conducting a Cook's Distance test, we determined there were three outliers to remove from the model. We tested for multicollinearity and determined it was not present. After plotting the residuals against cost, the data showed that heteroskedasticity may be present. We performed the Breusch-Pagan test and confirmed it may be present, but using robust standard errors with a high number of observations in our data, we concluded that the variance of our estimators should still allow us to get precise estimates from our model. Using the RESET test, we also determined that higher order terms may improve our model, but our original model had a higher R^2 without the higher order terms. Lastly, we conducted an AIC and BIC test to compare our models throughout the exercise. Our final model had the lowest AIC and BIC, so we finally concluded that it was the best model.