Coal Assignment

Green Team

10/11/2016

Gathering and organizing Data on Coal

```
Coal <- read.csv("https://www.eia.gov/totalenergy/data/browser/csv.cfm?tbl=T06.01")
CoalProduction <- Coal %>% filter(MSN == "CLPRPUS")
CoalProduction <- CoalProduction %>% filter(str_sub(as.character(YYYYMM),5 ) == "13")
CoalProduction <- CoalProduction %>% select(YYYYMM, Value)
names(CoalProduction) <- c("RawYear", "ProductionKShortTon")
CoalProduction$ProductionKShortTon <- as.numeric(as.character(CoalProduction$ProductionKShortTon))
CoalProduction <- CoalProduction  %>% mutate(Year = as.numeric(str_sub(as.character(RawYear),0,4 )))
Prices <- Quandl("EPI/152")
Prices$Year <- as.numeric(str_sub(Prices$Year,0,4))
names(Prices) <- c("Year", "PriceMBTU")
CoalMarket <- inner_join(Prices, CoalProduction, by ="Year")
```

The above was covered in class, but it is gathering the data on coal production and prices and putting them together in a single dataset.

Simplifying Units

Since the production is in short tons, the units need to first be converted for ease of understanding on the graphs. Since a short ton is roughly 20 MBTU, multiplying the Production in Short tons by that amount should give us MBTUs for both the price and production quantity. Since all the values are over 1,000,000, Production will also be divided by 1,000,000 in order to keep the numbering simple.

```
CoalMarket$ProductionMMBTU <- ((CoalMarket$ProductionKShortTon * 20) / 1000000)
```

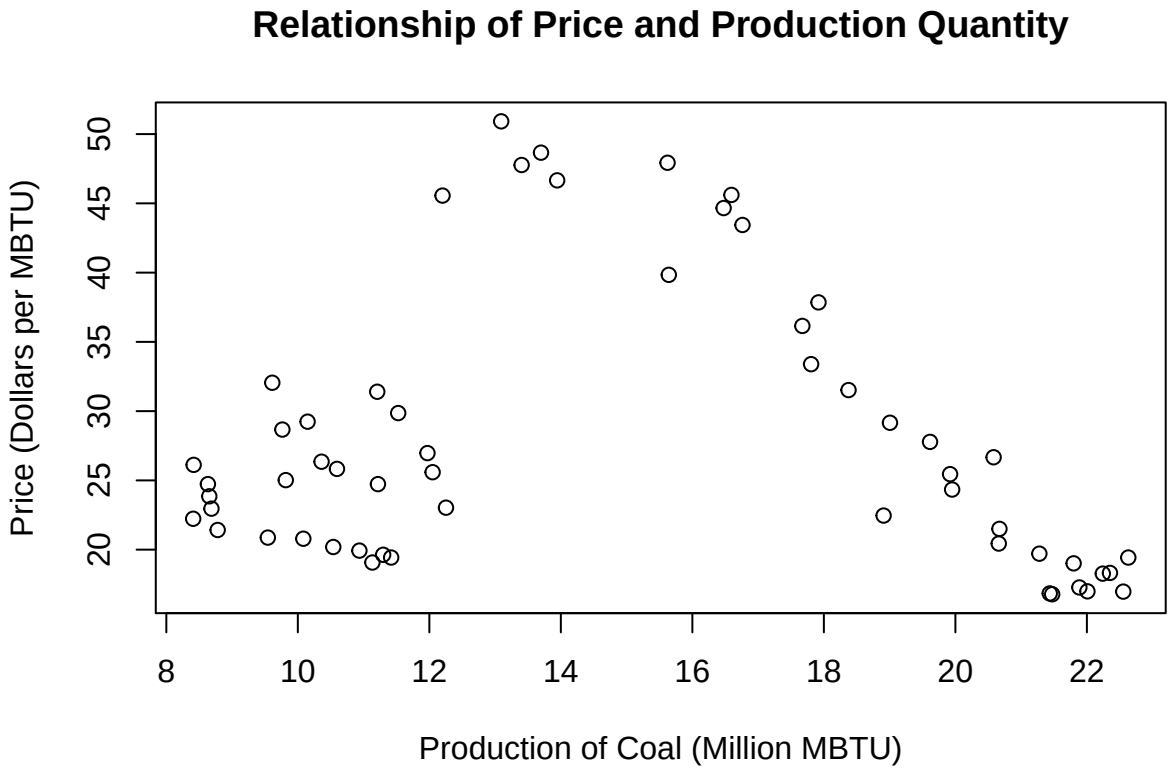
Now, the dataset CoalMarket has a number of different variables that are unnecessary for the purposes of this analysis. The following chunk will create a new Dataset called AdjCoalMkt (Adjusted Coal Market) with just year, price, and production quantity.

```
AdjCoalMkt <- CoalMarket %>% select(Year,PriceMBTU,ProductionMMBTU)
names(AdjCoalMkt) <- c("Year","PriceMBTU","ProdQtyMMBTU")
```

Graphical Analysis

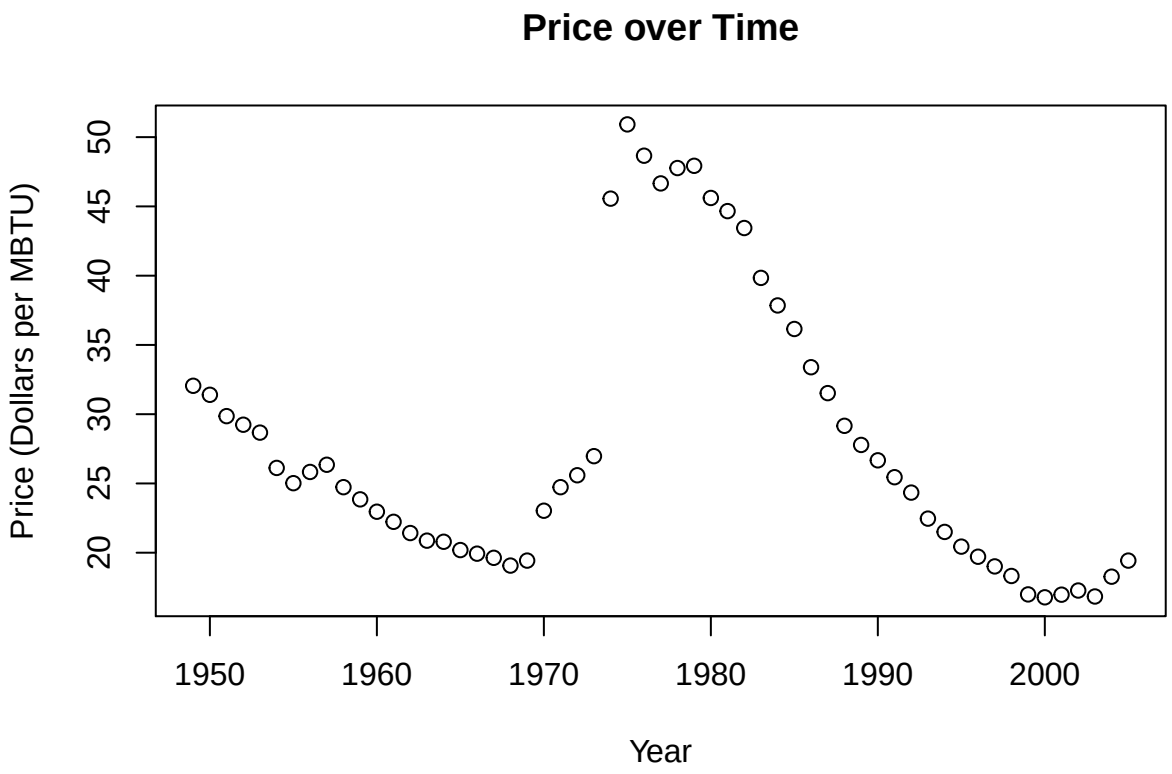
To better analyze the relationships inherent in the data, here are a few graphs detailing the relation between Price and Quantity (a demand relationship), Price over time, and Production over time.

```
plot(PriceMBTU ~ ProdQtyMMBTU, data=AdjCoalMkt, main="Relationship of Price and Production Quantity", xlab=
```



So we can see that there are almost two demand functions. One that starts at roughly 12.5 million MBTUs produced and has a pretty high correlation factor, and the one below 12.5 million MBTUs with a very low correlation factor.

```
plot(PriceMBTU ~ Year, data=AdjCoalMkt, main="Price over Time", ylab="Price (Dollars per MBTU)")
```

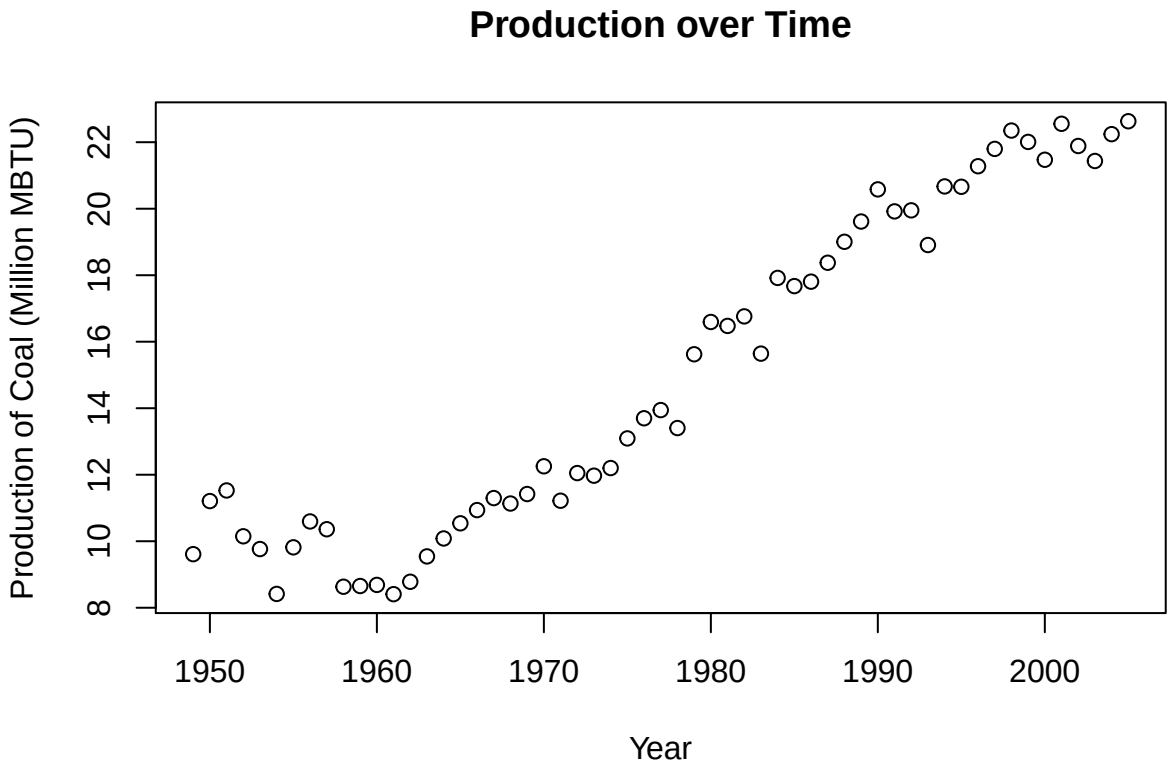


The relationship between price over time seems like it has a very high correlation, although again there are two very different lines, split around the mid-1970s.

What was behind the price spike in 1974?

The OPEC oil embargo impacted the entire energy sector, including coal. The high prices for coal, stemming from the embargo, were themselves a market incentive for additional investments into yet more coal production (in the US and internationally). After the mid-1970s, the increased production, worldwide, appears to have created a fair degree of downward price pressure on coal. (Wikipedia, https://en.m.wikipedia.org/wiki/History_of_coal_mining_in_the_United_States)

```
plot(ProdQtyMMBTU ~ Year, data=AdjCoalMkt, main="Production over Time", ylab="Production of Coal (Million M
```



Why the low in the 1950s?

Several market changes drove down demand (to its 1954 low), including railroads fuel-switching from coal to diesel, home heating and cooking switching from coal to natural gas, and a general trend of industrial production shifting from coal-fed manufacturing processes to electricity-fed production. (Andrew Needham, Power Lines: Phoenix and the Making of the Modern Southwest, Princeton University Press, 2014)

What was behind the steady increase in production, sustained since the 1950s?

The uneven but steady increase in coal production, after the lows of the 1950s, was largely driven by an equally steady increase in demand, particularly from the expansion of coal-fired power plants-a trend that was sustained into the first decade of the 21st century.

The same increase in production was also enabled by significant technology changes associated with mining-especially those technologies associated with surface mining and mechanical extraction (e.g. conveyer belts and mechanical loaders). (G.E. Harding, "American Coal Production and Use," Economic Geography, Vol. 22, No. 1 (Jan. 1946)).

Before and After the Embargo

In order to get a proper look at the market, we need to add a dummy variable called "Embargo" to our dataset which will include two values: Before and After. This will be helpful to give us a solid relation between price and production

quantity. As we saw from our Price/Quantity graph, the correlation didn't start until about 12.5 million MBTUs produced. If we look at our production over time graph, we can see that production hit that mark in the mid-1970s.

```
AdjCoalMkt$Embargo <- as.factor(ifelse(AdjCoalMkt$Year < 1974, "Before", "After"))
```

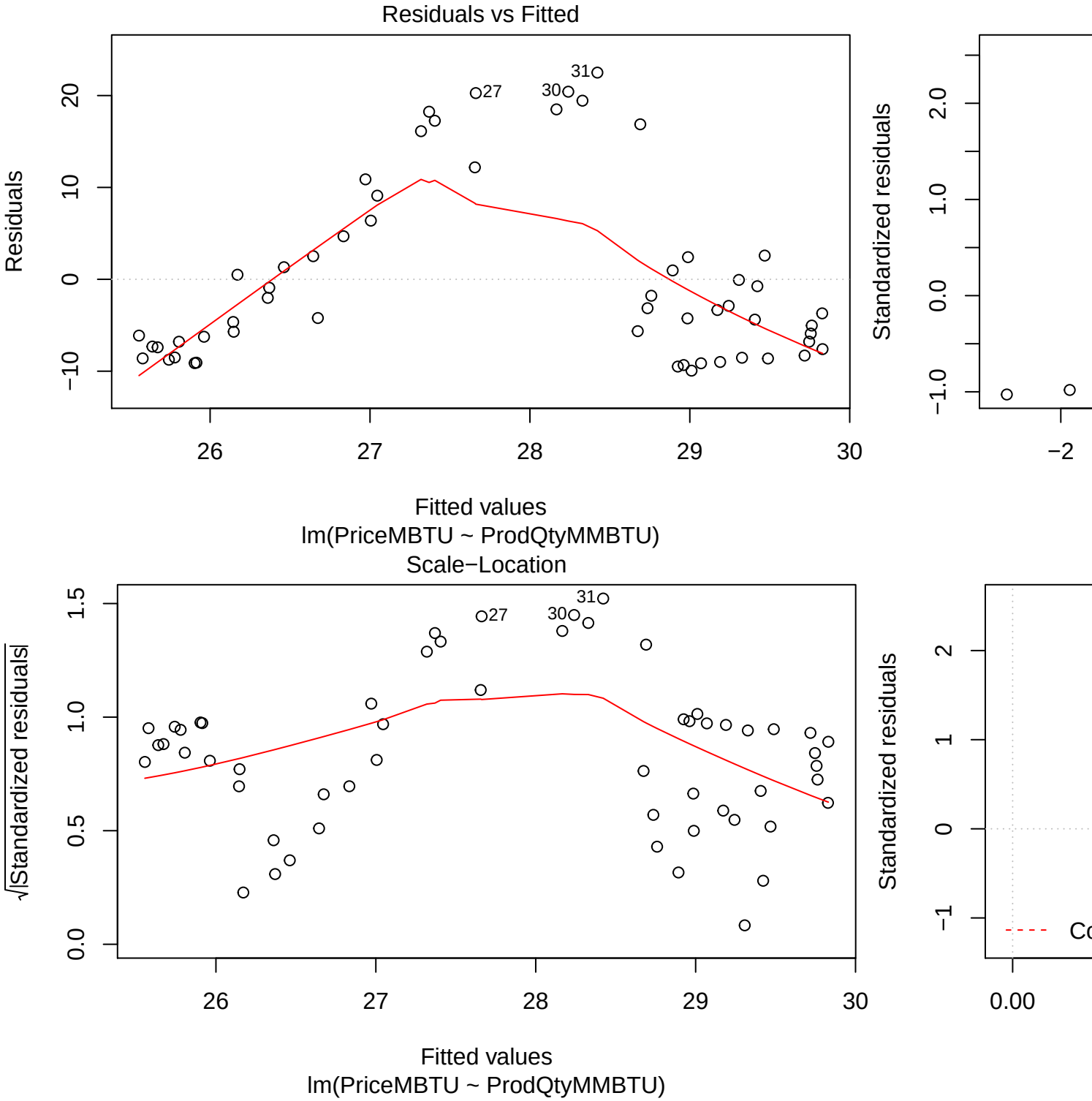
Demand Regression

First, let's take a regression without respects to the Embargo, to compare the results. We'll look at the summary as well as the plots of that regression.

```
LinModel1 <- lm(PriceMBTU ~ ProdQtyMMBTU, data=AdjCoalMkt)
summary(LinModel1)
```

```
##
## Call:
## lm(formula = PriceMBTU ~ ProdQtyMMBTU, data = AdjCoalMkt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.941 -7.402 -3.708  2.582 22.498
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.3573     4.2325   7.645 3.31e-10 ***
## ProdQtyMMBTU  -0.3006     0.2685  -1.120    0.268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.813 on 55 degrees of freedom
## Multiple R-squared:  0.02228,    Adjusted R-squared:  0.004507
## F-statistic: 1.254 on 1 and 55 DF,  p-value: 0.2677
```

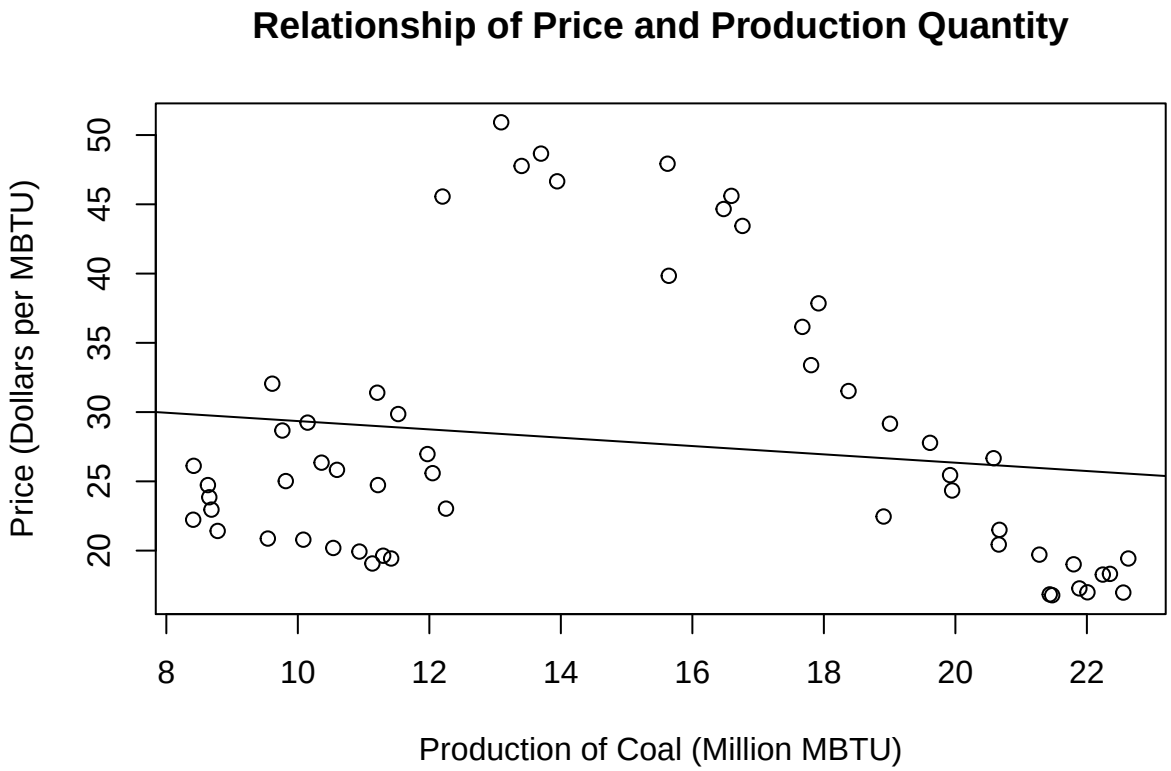
```
plot(LinModel1)
```



Taking a first glance at our Residuals vs Fitted plot graph, we can tell that this type of graph can only be fairly useful because of the upside U shape of the data points. Also, that the relationship between our two variables is not as linear as we would have hoped. As we can tell the lower the price of coal the lower the variance in our error terms. Where as we move out to the larger price points the large the variance in error terms gets and the more heteroskedasticity we see.

Just for fun, here's what that regression line would look like.

```r
plot(PriceMBTU ~ ProdQtyMMBTU, data=AdjCoalMkt, main="Relationship of Price and Production Quantity", xlab=
abline(LinModel1)
```



This really demonstrates the value of adding that dummy variable.

Now we'll take a linear regression with respects to the model.

```r
LinModel2 <- lm(PriceMBTU ~ ProdQtyMMBTU + Embargo, data=AdjCoalMkt)
summary(LinModel2)
```

```
##
## Call:
## lm(formula = PriceMBTU ~ ProdQtyMMBTU + Embargo, data = AdjCoalMkt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3179 -2.7941 -0.6247  2.4595 10.0388
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    91.8883     4.8063   19.12   <2e-16 ***
## ProdQtyMMBTU   -3.2818     0.2533  -12.96   <2e-16 ***
## EmbargoBefore -33.7456     2.4711  -13.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.693 on 54 degrees of freedom
## Multiple R-squared:  0.7805, Adjusted R-squared:  0.7723
## F-statistic: 95.98 on 2 and 54 DF,  p-value: < 2.2e-16
```

Already we see a much better result for the residuals. The following are the plots for our second Linear Model.

```
plot(LinModel2)
```