Predicting Weekly Sales of Retail Company

Key Questions to address:
1. What are the most important factors that impact weekly sales?
2. How useful are these factors in predicting weekly sales?

```
library(tidyverse)
library(corrplot)
library(fastDummies)
```

```
# read in data
setwd("G:/My Drive/Semester 1/Statistics/R Programs/Data Files")
features <- read.csv('Features data set.csv')
sales <- read.csv('sales data-set.csv')
stores <- read.csv('stores data-set.csv')
```

About the Data

The data was collected from the folloing kaggle site.
https://www.kaggle.com/manjeetsingh/retaildataset

We have three data sets that consist of weekly sales for 45 different stores across 143 weeks. For a detailed description of the data set, please refer to the Introduction section of the statistical report provided in this repository.

Below is a quick glance at each data set.

```
head(features)
```

```
##   Store       Date Temperature Fuel_Price MarkDown1 MarkDown2 MarkDown3
## 1     1   5/2/2010       42.31      2.572        NA        NA        NA
## 2     1  12/2/2010       38.51      2.548        NA        NA        NA
## 3     1 19/02/2010       39.93      2.514        NA        NA        NA
## 4     1 26/02/2010       46.63      2.561        NA        NA        NA
## 5     1   5/3/2010       46.50      2.625        NA        NA        NA
## 6     1  12/3/2010       57.79      2.667        NA        NA        NA
##   MarkDown4 MarkDown5      CPI Unemployment IsHoliday
## 1        NA        NA 211.0964        8.106     FALSE
## 2        NA        NA 211.2422        8.106      TRUE
## 3        NA        NA 211.2891        8.106     FALSE
## 4        NA        NA 211.3196        8.106     FALSE
## 5        NA        NA 211.3501        8.106     FALSE
## 6        NA        NA 211.3806        8.106     FALSE
```

```
head(sales)
```

```
##   Store Dept       Date Weekly_Sales IsHoliday
## 1     1    1   5/2/2010     24924.50     FALSE
## 2     1    1  12/2/2010     46039.49      TRUE
## 3     1    1 19/02/2010     41595.55     FALSE
## 4     1    1 26/02/2010     19403.54     FALSE
## 5     1    1   5/3/2010     21827.90     FALSE
## 6     1    1  12/3/2010     21043.39     FALSE
```

```
head(stores)
```

```
##   Store Type   Size
## 1     1    A 151315
## 2     2    A 202307
## 3     3    B  37392
## 4     4    A 205863
## 5     5    B  34875
## 6     6    A 202505
```

Data Preparation

We merged these three data sets on the store ID and date, grouping our sales by store type and date. We chose not to analyze the markdown attributes since a majority of this data was missing. Our final data set consisted of weekly sales, temperature, fuel price, CPI, unemployment, holiday, store type, and store size. Our final data set allowed us to look into three areas that impact weekly sales: store level attributes, macro-economic variables, and external variables.

```
# group sales
sales_grouped <- sales %>% group_by(Store, Date) %>%
  summarise(sum_weekly_sales = sum(Weekly_Sales))
```

```
## `summarise()` has grouped output by 'Store'. You can override using the `.groups` argument.
```

```
# join grouped sales
merge <- merge(sales_grouped, features, by=c('Store', 'Date'))

# join to get store type
all_data <- merge(merge, stores, by='Store')

# select only the predictors we want
all_data <- all_data %>%
  select(c('sum_weekly_sales', 'Temperature', 'Fuel_Price', 'CPI', 'Unemployment', 'IsHoliday', 'Type',

# convert categorical variables to factors
all_data$IsHoliday <- factor(all_data$IsHoliday)
all_data$Type <- factor(all_data$Type)
```

Exploring the Data

We first looked at store type to better understand this attribute and found that store type could be correlated with store size, with Type A stores having the highest size.

```
# understanding store types
store_grouped <- stores %>% group_by(Type) %>%
  summarise(abg = mean(Size),
            max = max(Size),
            min = min(Size),
            median = median(Size))
store_grouped
```

```
## # A tibble: 3 x 5
##   Type      abg    max   min median
##   <chr>   <dbl>  <int> <int>  <dbl>
## 1 A     177248. 219622 39690 202406
## 2 B     101191. 140167 34875 114533
## 3 C      40542.  42988 39690  39910
```

```
# Type A stores have highest size
```

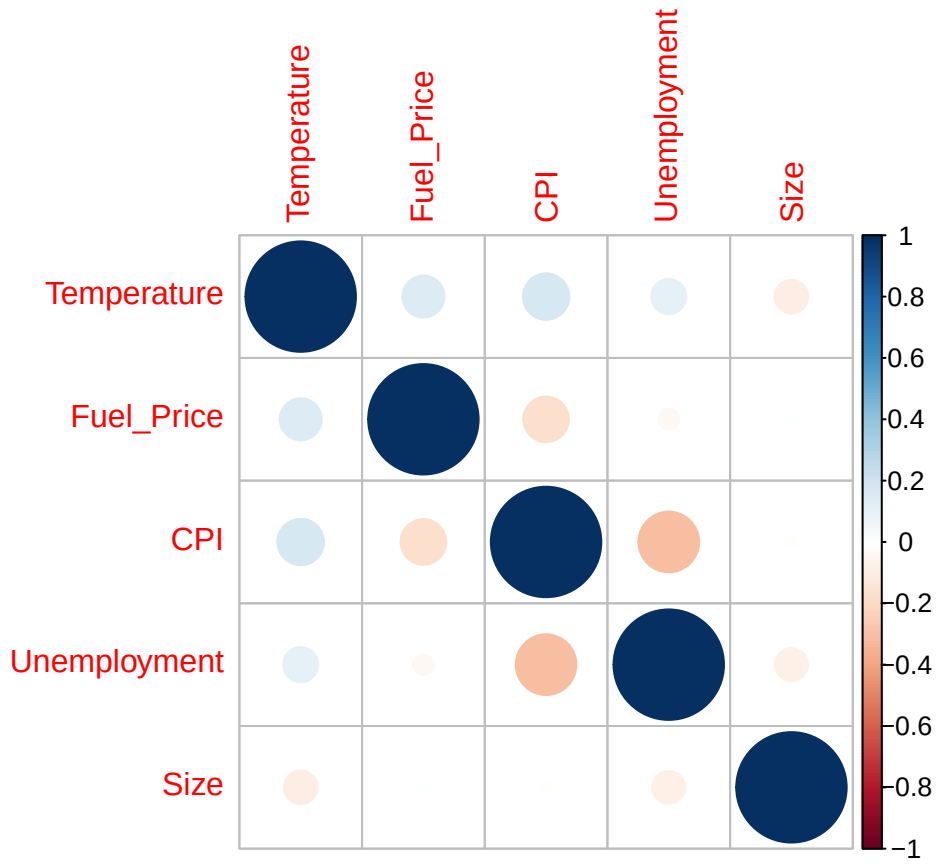
To be sure that store type and store size are correlated, we performed an anova test. The results shown below do suggest that there is a relationship between store size and store type. This would mean that if our model included store type and store size, we would not be able to interpret the beta values

```
# COLINEARITY
# size and type are highly correlated
aov_type_size <- aov(Size ~ Type, all_data)
summary(aov_type_size)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## Type            2 1.591e+13 7.953e+12    5260 <2e-16 ***
## Residuals    6432 9.725e+12 1.512e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

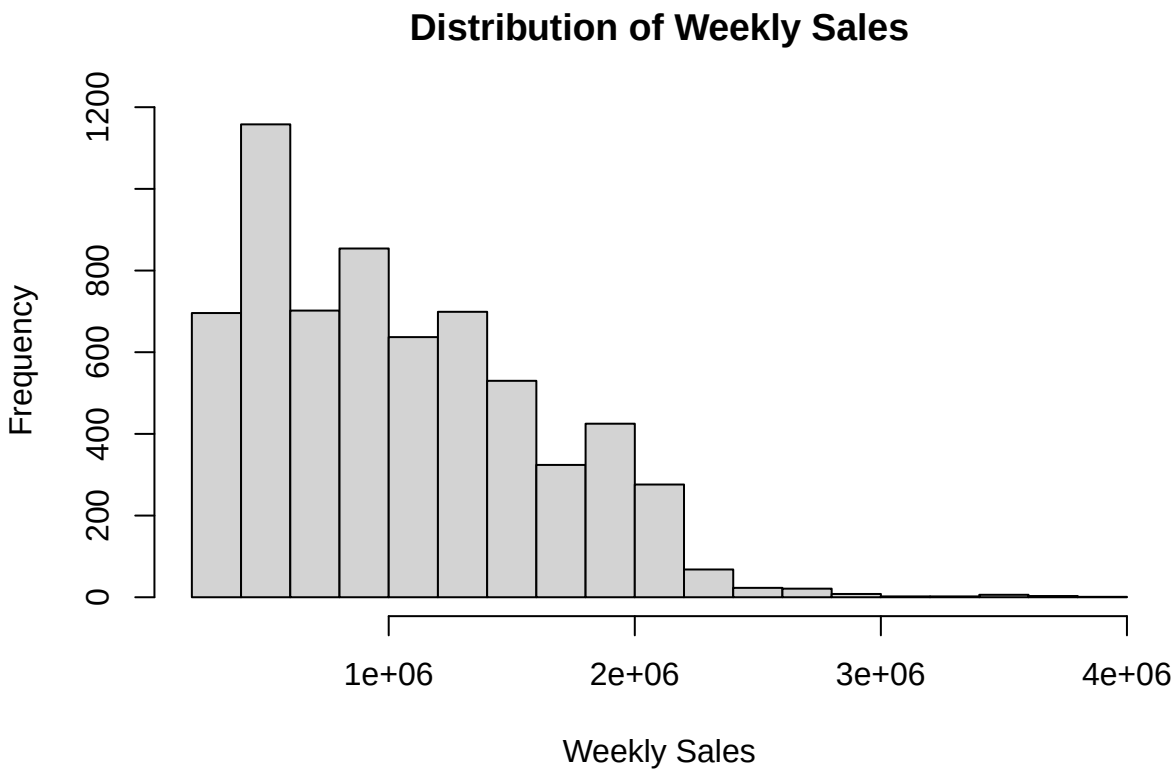
We also looked at the colinearity between our numeric predictors and found that none were highly colinear. This means that colinearity will not affect our linear regression model, meaning we will still be able to interpret the coefficients of each attribute in our model because these attributes are not colinear.

```
# COLINEARITY
# colinearity of numeric predictors
x <- cor(all_data[,c(2:5,8)])
corrplot(x)
```



In addition to looking at colinearity, we found that weekly sales are highly skewed. While having a highly skewed dependent variable does not violate an assumption, it may make OLS regression inappropriate. OLS regression models the mean weekly sales and the mean is not a good measure of central tendency in a skewed distribution

```
hist(all_data$sum_weekly_sales, main='Distribution of Weekly Sales', xlab='Weekly Sales')
```



Modeling

We tested out eight models before deciding on our final model that will predict weekly sales. Below is the code for each model.

Determining which store type to use as a base for future models

```
# type A as base level
model1 <- lm(sum_weekly_sales ~ Temperature+Fuel_Price+CPI+Unemployment+Size+IsHoliday+Type,
             all_data)
summary(model1)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Temperature + Fuel_Price + CPI +
##     Unemployment + Size + IsHoliday + Type, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -639487 -250029  -19155  151713 2699483
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.127e+05  5.002e+04   8.250  < 2e-16 ***
```

```
## Temperature    1.262e+03  2.354e+02   5.361 8.54e-08 ***
## Fuel_Price    -2.643e+04  9.145e+03  -2.890  0.00387 **
## CPI           -1.379e+03  1.132e+02 -12.176  < 2e-16 ***
## Unemployment  -2.505e+04  2.342e+03 -10.692  < 2e-16 ***
## Size           7.926e+00  1.045e-01  75.855  < 2e-16 ***
## IsHolidayTRUE  9.362e+04  1.601e+04   5.847 5.24e-09 ***
## TypeB          4.824e+04  1.192e+04   4.046 5.27e-05 ***
## TypeC          1.948e+05  1.915e+04  10.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322800 on 6426 degrees of freedom
## Multiple R-squared:  0.6733, Adjusted R-squared:  0.6729
## F-statistic:  1656 on 8 and 6426 DF,  p-value: < 2.2e-16
```

```
# type B as base level
all_data_relevel <- within(all_data, Type <- relevel(Type, ref = 'B'))
model2 <- lm(sum_weekly_sales ~ Temperature+Fuel_Price+CPI+Unemployment+Size+IsHoliday+Type,
             all_data_relevel)
summary(model2)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Temperature + Fuel_Price + CPI +
##     Unemployment + Size + IsHoliday + Type, data = all_data_relevel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -639487 -250029  -19155  151713 2699483
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.609e+05  4.768e+04   9.667  < 2e-16 ***
## Temperature    1.262e+03  2.354e+02   5.361 8.54e-08 ***
## Fuel_Price    -2.643e+04  9.145e+03  -2.890  0.00387 **
## CPI           -1.379e+03  1.132e+02 -12.176  < 2e-16 ***
## Unemployment  -2.505e+04  2.342e+03 -10.692  < 2e-16 ***
## Size           7.926e+00  1.045e-01  75.855  < 2e-16 ***
## IsHolidayTRUE  9.362e+04  1.601e+04   5.847 5.24e-09 ***
## TypeA         -4.824e+04  1.192e+04  -4.046 5.27e-05 ***
## TypeC          1.466e+05  1.458e+04  10.055  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322800 on 6426 degrees of freedom
## Multiple R-squared:  0.6733, Adjusted R-squared:  0.6729
## F-statistic:  1656 on 8 and 6426 DF,  p-value: < 2.2e-16
```

```
# type C as base level
all_data_relevelC <- within(all_data, Type <- relevel(Type, ref = 'C'))
model3 <- lm(sum_weekly_sales ~ Temperature+Fuel_Price+CPI+Unemployment+Size+IsHoliday+Type,
             all_data_relevelC)
summary(model3)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Temperature + Fuel_Price + CPI +
##     Unemployment + Size + IsHoliday + Type, data = all_data_relevelC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -639487 -250029  -19155  151713 2699483
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.075e+05  4.900e+04  12.397  < 2e-16 ***
## Temperature    1.262e+03  2.354e+02   5.361 8.54e-08 ***
## Fuel_Price    -2.643e+04  9.145e+03  -2.890  0.00387 **
## CPI           -1.379e+03  1.132e+02 -12.176  < 2e-16 ***
## Unemployment  -2.505e+04  2.342e+03 -10.692  < 2e-16 ***
## Size           7.926e+00  1.045e-01  75.855  < 2e-16 ***
## IsHolidayTRUE  9.362e+04  1.601e+04   5.847 5.24e-09 ***
## TypeA         -1.948e+05  1.915e+04 -10.171  < 2e-16 ***
## TypeB         -1.466e+05  1.458e+04 -10.055  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322800 on 6426 degrees of freedom
## Multiple R-squared:  0.6733, Adjusted R-squared:  0.6729
## F-statistic:  1656 on 8 and 6426 DF,  p-value: < 2.2e-16
```

We will use Type A as the base level in all of our models because it has the lowest coefficient.

Model 1: Weekly Sales across all predictors

```
model1 <- lm(sum_weekly_sales ~ Temperature+Fuel_Price+CPI+Unemployment+Size+IsHoliday+Type,
             all_data)
summary(model1)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Temperature + Fuel_Price + CPI +
##     Unemployment + Size + IsHoliday + Type, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -639487 -250029  -19155  151713 2699483
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.127e+05  5.002e+04   8.250  < 2e-16 ***
## Temperature    1.262e+03  2.354e+02   5.361 8.54e-08 ***
## Fuel_Price    -2.643e+04  9.145e+03  -2.890  0.00387 **
## CPI           -1.379e+03  1.132e+02 -12.176  < 2e-16 ***
## Unemployment  -2.505e+04  2.342e+03 -10.692  < 2e-16 ***
## Size           7.926e+00  1.045e-01  75.855  < 2e-16 ***
## IsHolidayTRUE  9.362e+04  1.601e+04   5.847 5.24e-09 ***
## TypeB          4.824e+04  1.192e+04   4.046 5.27e-05 ***
```

```
## TypeC          1.948e+05  1.915e+04  10.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322800 on 6426 degrees of freedom
## Multiple R-squared:  0.6733, Adjusted R-squared:  0.6729
## F-statistic:  1656 on 8 and 6426 DF,  p-value: < 2.2e-16
```

Model 2: Weekly Sales with Size^2 and all Factors

```
model2 <- lm(sum_weekly_sales ~ Size + I(Size^2) + Temperature+Fuel_Price+CPI+Unemployment+IsHoliday+Ty
             all_data)
summary(model2)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Size + I(Size^2) + Temperature +
##     Fuel_Price + CPI + Unemployment + IsHoliday + Type, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -657419 -227991  -33493  134360 2719744
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.518e+05  5.273e+04  12.361  < 2e-16 ***
## Size           2.510e+00  4.312e-01   5.822 6.11e-09 ***
## I(Size^2)      2.277e-05  1.760e-06  12.937  < 2e-16 ***
## Temperature    1.067e+03  2.329e+02   4.583 4.67e-06 ***
## Fuel_Price    -2.461e+04  9.030e+03  -2.725  0.00644 **
## CPI           -1.590e+03  1.130e+02 -14.073  < 2e-16 ***
## Unemployment  -2.566e+04  2.313e+03 -11.094  < 2e-16 ***
## IsHolidayTRUE  9.167e+04  1.581e+04   5.799 6.98e-09 ***
## TypeB          1.472e+05  1.404e+04  10.483  < 2e-16 ***
## TypeC          1.865e+05  1.892e+04   9.858  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 318700 on 6425 degrees of freedom
## Multiple R-squared:  0.6816, Adjusted R-squared:  0.6812
## F-statistic:  1528 on 9 and 6425 DF,  p-value: < 2.2e-16
```

Model 3: Removing Fuel Price from Model 2

```
model3 <- lm(sum_weekly_sales ~ Size + I(Size^2) + Temperature+CPI+Unemployment+IsHoliday+Type,
             all_data)
summary(model3)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Size + I(Size^2) + Temperature +
##     CPI + Unemployment + IsHoliday + Type, data = all_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -652677 -227472  -34519  134539 2725680
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.621e+05  4.122e+04  13.636  < 2e-16 ***
## Size           2.479e+00  4.313e-01   5.748 9.43e-09 ***
## I(Size^2)      2.285e-05  1.761e-06  12.974  < 2e-16 ***
## Temperature    9.432e+02  2.285e+02   4.128 3.70e-05 ***
## CPI           -1.522e+03  1.103e+02 -13.805  < 2e-16 ***
## Unemployment  -2.489e+04  2.297e+03 -10.836  < 2e-16 ***
## IsHolidayTRUE  9.371e+04  1.580e+04   5.932 3.15e-09 ***
## TypeB          1.456e+05  1.403e+04  10.374  < 2e-16 ***
## TypeC          1.846e+05  1.892e+04   9.757  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 318800 on 6426 degrees of freedom
## Multiple R-squared:  0.6813, Adjusted R-squared:  0.6809
## F-statistic:  1717 on 8 and 6426 DF,  p-value: < 2.2e-16
```

Model 4: Removing Temperature from Model 3

```
model4 <- lm(sum_weekly_sales ~ Size + I(Size^2) +CPI+Unemployment+IsHoliday+Type,
             all_data)
summary(model4)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Size + I(Size^2) + CPI + Unemployment +
##     IsHoliday + Type, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -677180 -228546  -34995  136665 2727891
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.046e+05  3.997e+04  15.128  < 2e-16 ***
## Size           2.334e+00  4.304e-01   5.424 6.05e-08 ***
## I(Size^2)      2.330e-05  1.760e-06  13.243  < 2e-16 ***
## CPI           -1.434e+03  1.083e+02 -13.240  < 2e-16 ***
## Unemployment  -2.352e+04  2.276e+03 -10.335  < 2e-16 ***
## IsHolidayTRUE  8.305e+04  1.560e+04   5.322 1.06e-07 ***
## TypeB          1.422e+05  1.403e+04  10.137  < 2e-16 ***
## TypeC          1.843e+05  1.894e+04   9.732  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319200 on 6427 degrees of freedom
## Multiple R-squared:  0.6804, Adjusted R-squared:  0.6801
## F-statistic:  1955 on 7 and 6427 DF,  p-value: < 2.2e-16
```

Model 5: Removing isHoliday from Model 4:

```
model5 <- lm(sum_weekly_sales ~ Size + I(Size^2) +CPI+Unemployment+Type,
             all_data)
summary(model5)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Size + I(Size^2) + CPI + Unemployment +
##     Type, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -600121 -228545  -36972  137986 2721975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.093e+05  4.004e+04  15.217  < 2e-16 ***
## Size          2.334e+00  4.313e-01   5.411 6.48e-08 ***
## I(Size^2)     2.330e-05  1.763e-06  13.215  < 2e-16 ***
## CPI          -1.434e+03  1.086e+02 -13.205  < 2e-16 ***
## Unemployment -2.339e+04  2.280e+03 -10.255  < 2e-16 ***
## TypeB         1.421e+05  1.406e+04  10.113  < 2e-16 ***
## TypeC         1.841e+05  1.898e+04   9.701  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 319900 on 6428 degrees of freedom
## Multiple R-squared:  0.679,  Adjusted R-squared:  0.6787
## F-statistic:  2266 on 6 and 6428 DF,  p-value: < 2.2e-16
```

Model 6: Removing Type from Model 5

```
model6 <- lm(sum_weekly_sales ~ Size + I(Size^2)+Unemployment+CPI,
             all_data)
summary(model6)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Size + I(Size^2) + Unemployment +
##     CPI, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -598852 -240946  -31486  144381 2752950
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.187e+05  3.764e+04  19.093   <2e-16 ***
## Size          2.981e+00  3.353e-01   8.890   <2e-16 ***
## I(Size^2)     1.692e-05  1.322e-06  12.796   <2e-16 ***
## Unemployment -2.063e+04  2.273e+03  -9.077   <2e-16 ***
## CPI          -1.455e+03  1.095e+02 -13.297   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 323000 on 6430 degrees of freedom
## Multiple R-squared:  0.6726, Adjusted R-squared:  0.6724
## F-statistic:  3302 on 4 and 6430 DF,  p-value: < 2.2e-16
```

Model 7: Remove Unemployment from Model 6

```
model7 <- lm(sum_weekly_sales ~ Size + I(Size^2) +CPI,
             all_data)
summary(model7)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Size + I(Size^2) + CPI, data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -630118 -235518  -34639  141048 2741872
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.797e+05  2.707e+04  17.722   <2e-16 ***
## Size         3.319e+00  3.353e-01   9.897   <2e-16 ***
## I(Size^2)    1.579e-05  1.325e-06  11.917   <2e-16 ***
## CPI         -1.142e+03  1.045e+02 -10.928   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 325100 on 6431 degrees of freedom
## Multiple R-squared:  0.6684, Adjusted R-squared:  0.6682
## F-statistic:  4320 on 3 and 6431 DF,  p-value: < 2.2e-16
```

Model 8: Removing CPI from Model 6:

```
model8 <- lm(sum_weekly_sales ~ Size + I(Size^2)+Unemployment,
             all_data)
summary(model8)
```

```
##
## Call:
## lm(formula = sum_weekly_sales ~ Size + I(Size^2) + Unemployment,
##     data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -633413 -228937  -43845  163423 2794873
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.499e+05  2.579e+04  13.565  < 2e-16 ***
```

```
## Size          3.853e+00  3.333e-01  11.561  < 2e-16 ***
## I(Size^2)     1.355e-05  1.315e-06  10.298  < 2e-16 ***
## Unemployment -1.110e+04  2.186e+03  -5.078 3.91e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 327400 on 6431 degrees of freedom
## Multiple R-squared:  0.6636, Adjusted R-squared:  0.6634
## F-statistic:  4228 on 3 and 6431 DF,  p-value: < 2.2e-16
```

After comparing the p-values for each coefficient, the R^2 values, the standard errors, and the number of attributes in each model, we believe that Model 6 is the best model in predicting sales. We have chosen model 6 because 1) it is simple, 2) it explains a majority of the variation in sales compared to other models, 3) it has a low amount of unexplained variance in sales compared to other models, and 4) it is interpretable in the business context.

Now, we will check the assumptions for this model.

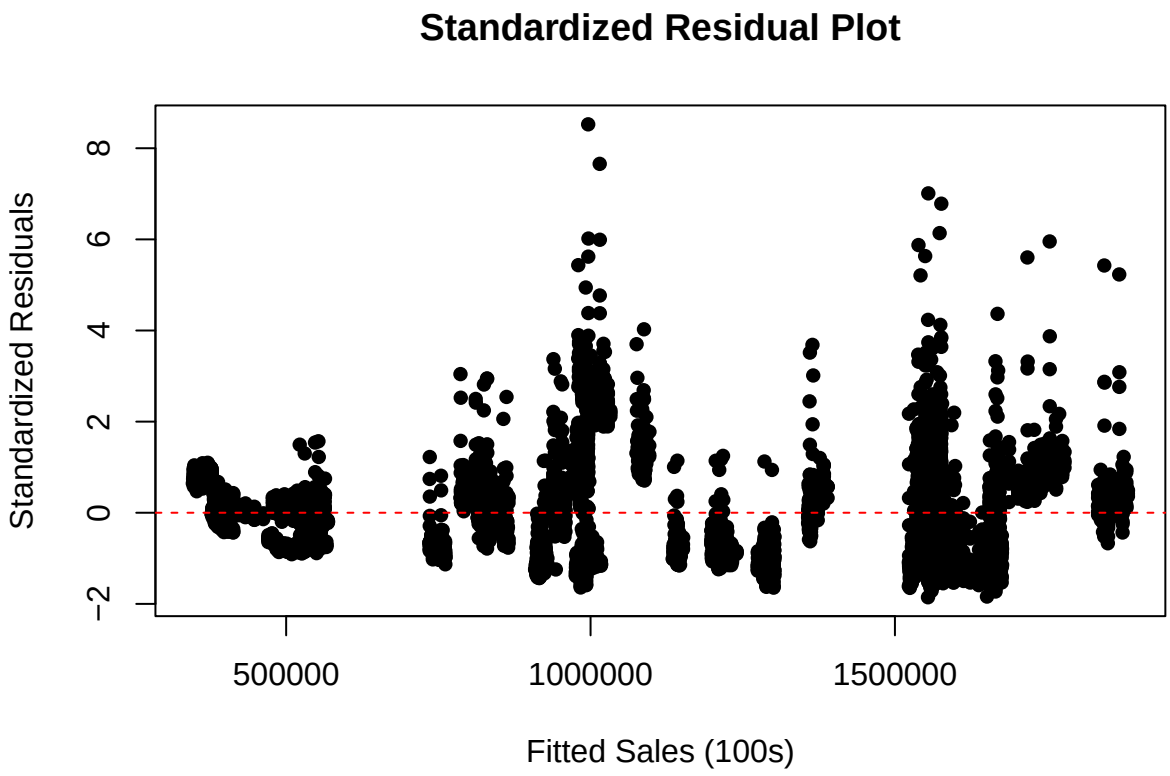
The three assumptions needed for a linear regression model are 1) Random sampling, 2) Stability over time, and 3) Error terms are normally distributed with a mean at 0 and a constant standard deviation across all possible values of the predictors.

The first two assumptions are obviously true, so we will check the third assumption by doing the following three steps:

- Check the mean and constant variance across fitted values
- Check of Mean and constant variance across all attributes
- Check the normality of the residuals with histogram distribution and Normal QQ plot. The shapiro-wilk test was not conducted because the size of data set is too large.

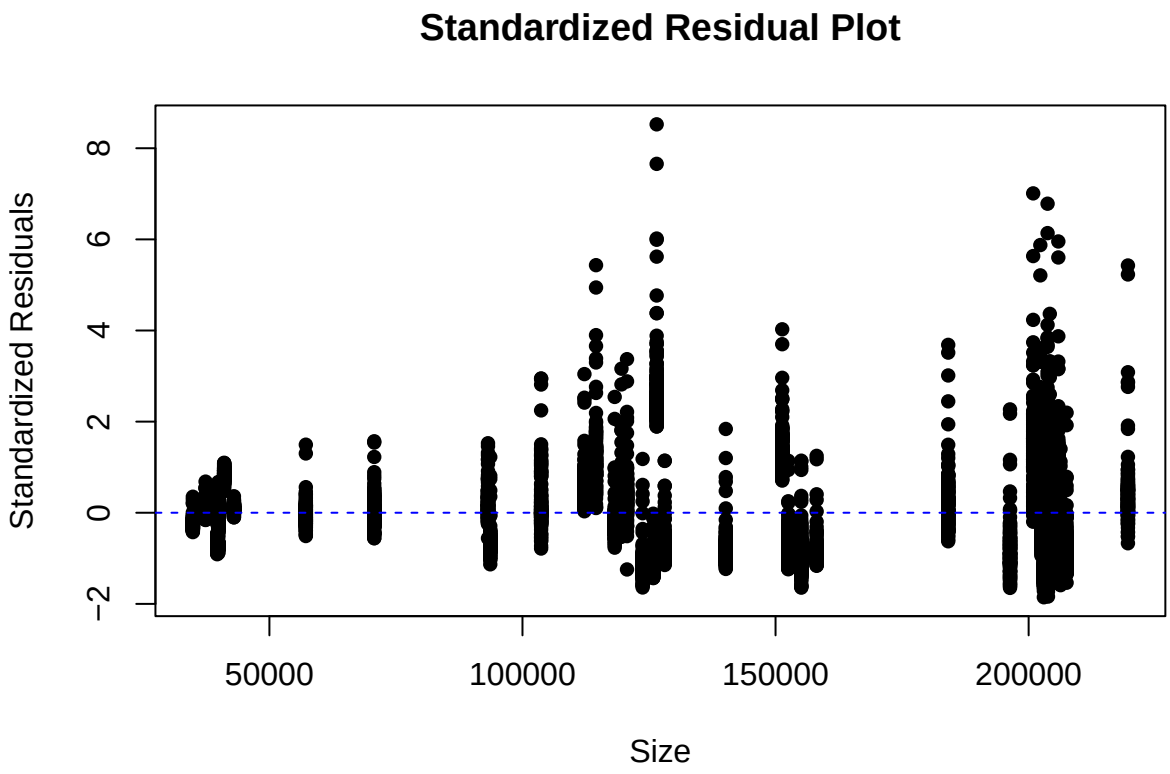
The below graph shows us that the standardized residuals over fitted values seems to have a mean at zero, but the standard deviation of the residuals seems to be inconsistent.

```
# ASSUMPTION CHECKING
attach(all_data)
# Check of Mean 0 and constant variance across all X
# Standardized residual plot - on fitted values
model6.stres <- rstandard(model6)
plot(model6$fitted.values, model6.stres, pch = 16,
     main = "Standardized Residual Plot",
     xlab = "Fitted Sales (100s)",
     ylab = "Standardized Residuals")
abline(0,0, lty=2, col="red")
```



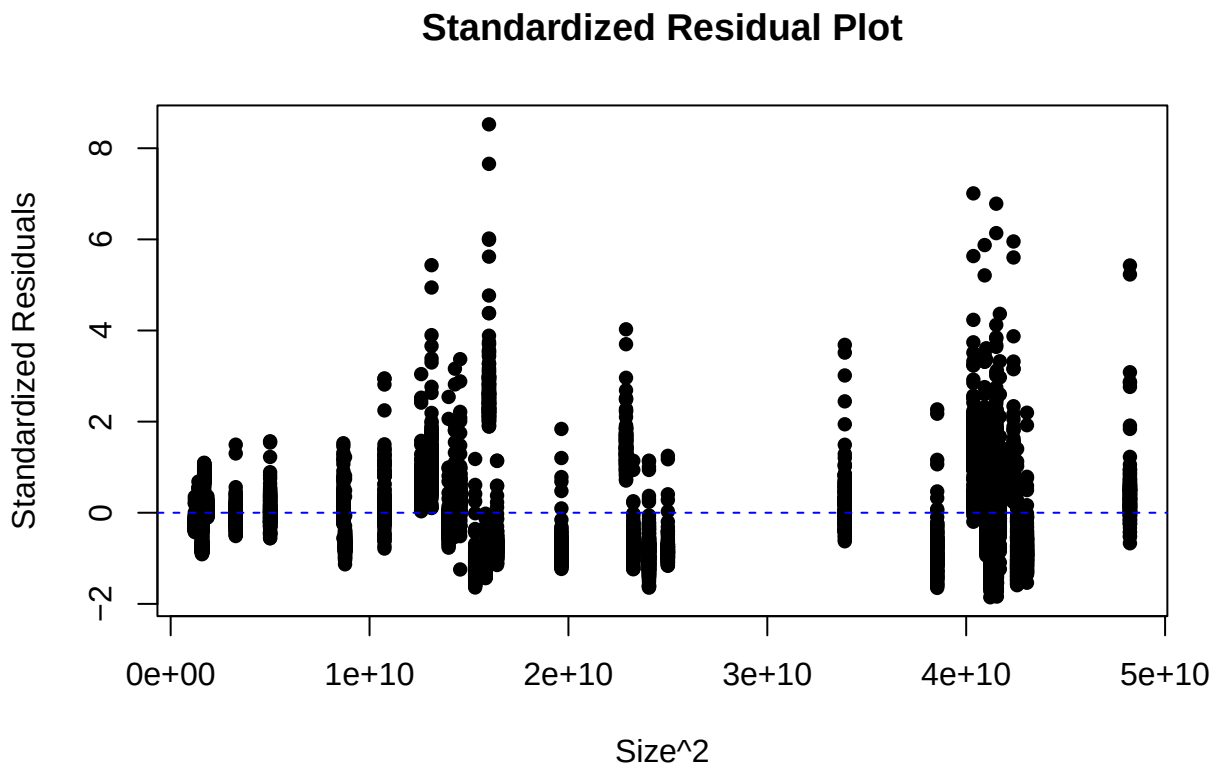
The scatter plot below shows us that the standardized residuals over Size seems to have a mean at zero, but the standard deviation of the residuals seems to be inconsistent.

```
# Individual scatter plots against St Resids
# standardized residual plot - on Permits
plot(Size, model6.stres, pch = 16,
     main = "Standardized Residual Plot",
     xlab = "Size",
     ylab = "Standardized Residuals")
abline(0,0, lty=2, col="blue")
```



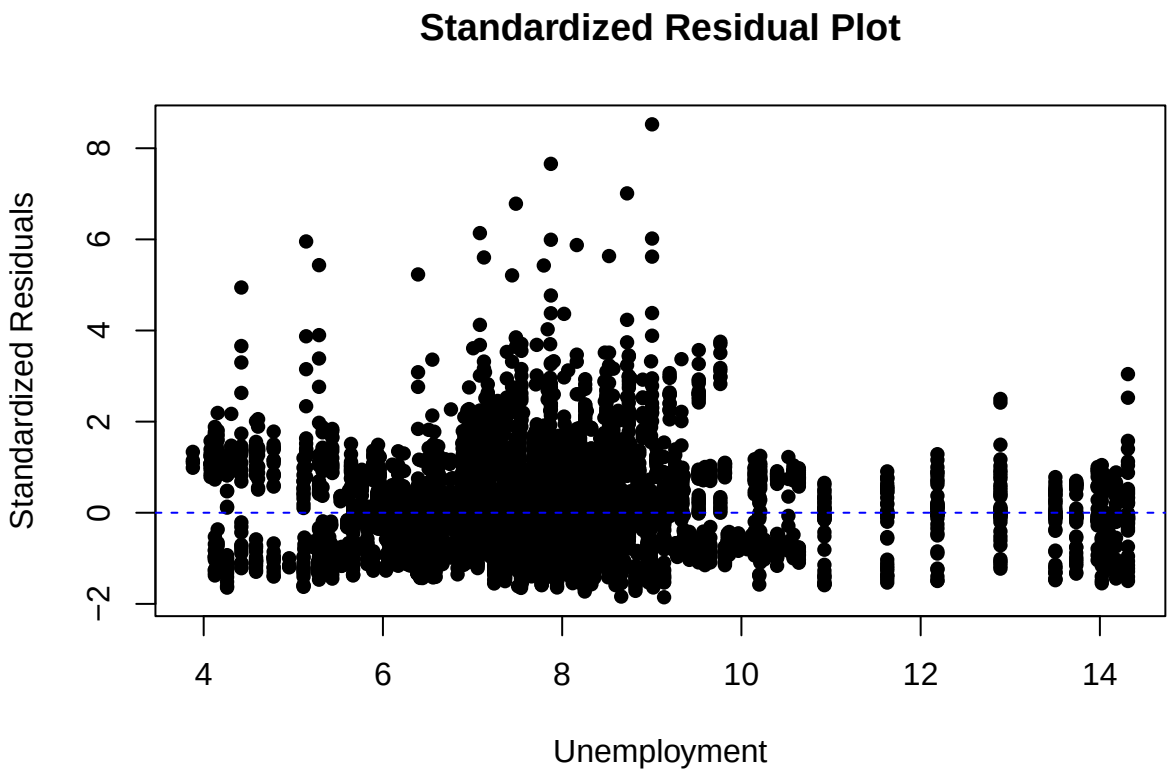
Below we see that standardized residuals over Size^2 seems to have a mean at zero, but its standard deviation seems to be inconsistent.

```
# standardized residual plot - on Size^2
plot(I(Size^2), model6.stres, pch = 16,
     main = "Standardized Residual Plot",
     xlab = "Size^2",
     ylab = "Standardized Residuals")
abline(0,0, lty=2, col="blue")
```



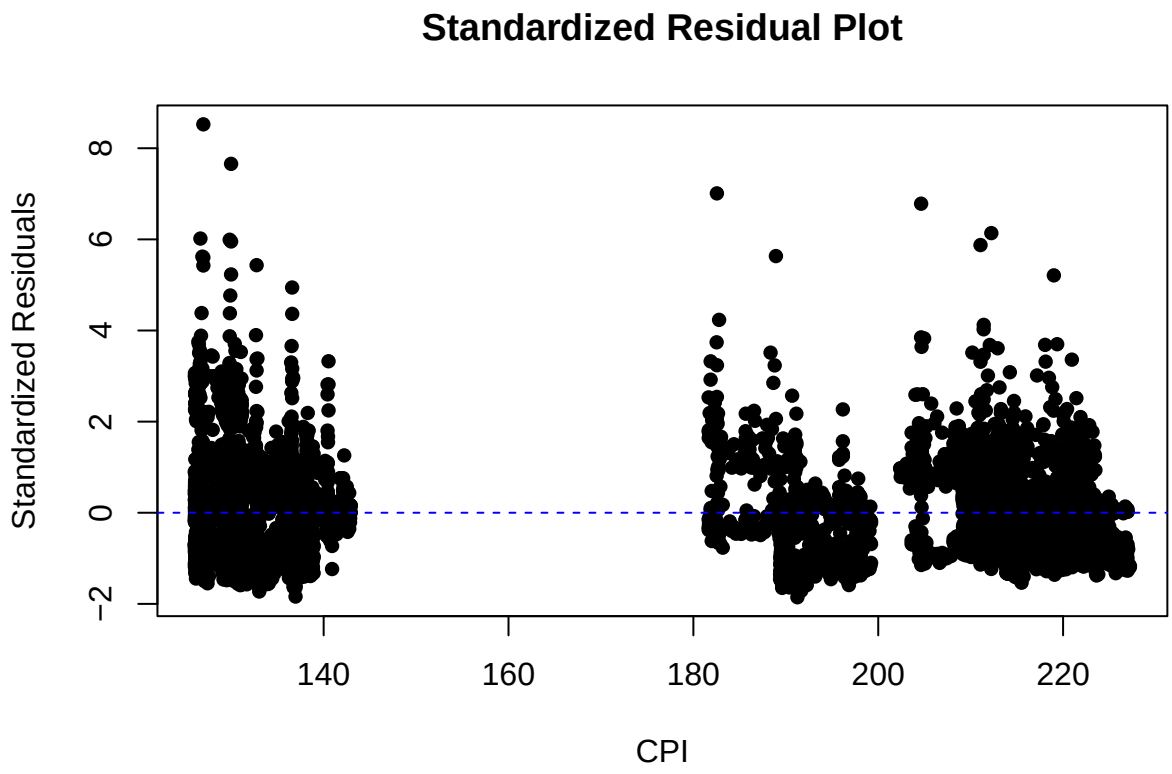
Standardized residuals over Unemployment seems to have a mean at zero and a constant standard deviation across all unemployment values.

```
# standardized residual plot - on Unemployment
plot(Unemployment, model6.stres, pch = 16,
     main = "Standardized Residual Plot",
     xlab = "Unemployment",
     ylab = "Standardized Residuals")
abline(0,0, lty=2, col="blue")
```



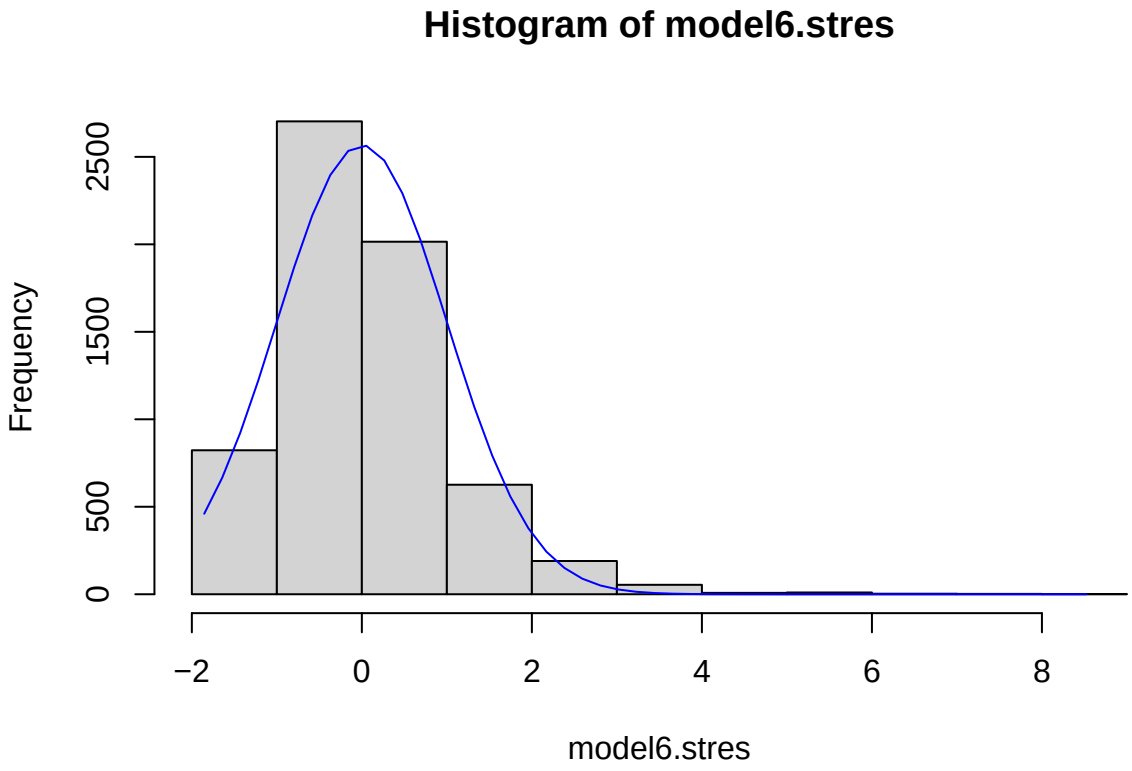
Standardized residuals over CPI seems to have a mean at zero, and a constant standard deviation across all Unemployment values.

```
# standardized residual plot - on CPI
plot(CPI, model6.stres, pch = 16,
     main = "Standardized Residual Plot",
     xlab = "CPI",
     ylab = "Standardized Residuals")
abline(0,0, lty=2, col="blue")
```



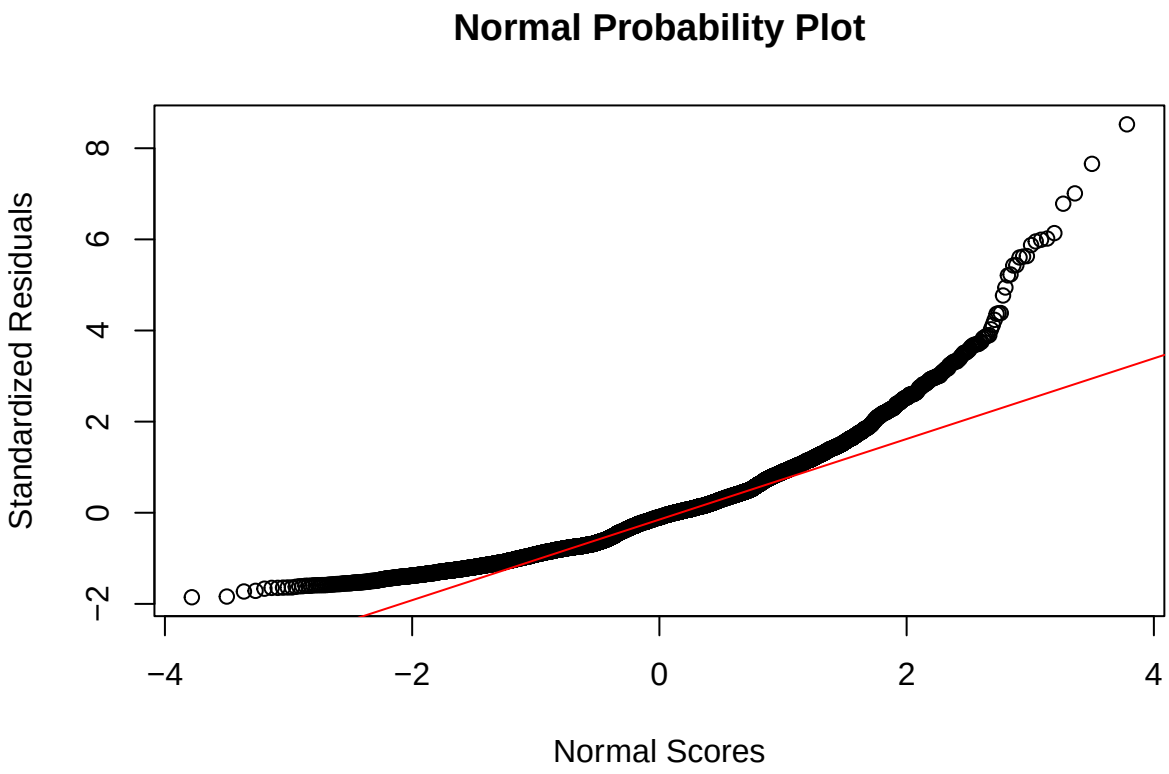
The below graph shows that the residuals have a smaller variance than a typical normal distribution.

```
# Normality checking
# Histogram with normal curve
h <- hist(model6.stres)
x <- model6.stres
xfit <- seq(min(x), max(x), length = 50)
yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue")
```



The graph above shows that the residuals have a different distribution from a typical normal distribution.

```
# Normal probability plot
qqnorm(model6.stres,
       main = "Normal Probability Plot",
       xlab = "Normal Scores",
       ylab = "Standardized Residuals")
qqline(model6.stres, col = "red")
```



Thus, our third assumption does not hold, and we will conclude that error terms are not normally distributed with a constant standard deviation across all possible values of the predictors. This is one limitation of our model and additional analyses should be done in the future to follow up with this issue.

Final Model Interpretations

Our final model found that the size of a store, unemployment rate, and the CPI value are the most important factors that impact the weekly sales. The model explains 68% of the weekly sales. Our degree of confidence of the relationship between Size, CPI and Unemployment with sales in the model is high.

Based on our model, we know that there is a positive curvilinear relationship between size and weekly sales. There is a negative relationship between CPI and sales: the weekly sales decreases by $19,930 for one unit increase in unemployment index when size and CPI stay unchanged. Finally, there is a negative relationship between unemployment and sales: the weekly sales decreases by $1,409 for one unit increase in CPI when size and unemployment stay unchanged.

This model can be used to benchmark a store's weekly sales. For example, the company can examine a particular store's size, the unemployment rate in that store's region, and the CPI value to understand how a store should be performing in terms of weekly sales. Additionally, the model can be utilized to scope out new regions to open stores in. By looking into a particular region's unemployment rate and CPI, a manager can understand what a store's weekly sales might look like, depending on the size of the store, if XYZ were to open a store in that region.