CSC 423 Homework 6

Akhil Kumar Ramasagaram

May 12, 2016

7.11 FTC cigarette study

```r
library(ggplot2)
library(gridExtra)
load("rdata/FTCCIGAR.Rdata")
summary(lm(CO ~ TAR, data = FTCCIGAR))[4]
```

```
## $coefficients
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 2.743278 0.67520594  4.062875 4.811735e-04
## TAR         0.800976 0.05032017 15.917592 6.552245e-14
```

```r
summary(lm(CO ~ NICOTINE, data = FTCCIGAR))[4]
```

```
## $coefficients
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  1.664666  0.9936018  1.675386 1.074026e-01
## NICOTINE    12.395406  1.0541519 11.758653 3.311725e-11
```

```r
summary(lm(CO ~ WEIGHT, data = FTCCIGAR))[4]
```

```
## $coefficients
##              Estimate Std. Error   t value   Pr(>|t|)
## (Intercept) -11.79527   9.721626 -1.213302 0.23732811
## WEIGHT       25.06820   9.980282  2.511772 0.01948117
```

$\beta_1 = 0.8$, $\beta_2 = 12.39$ & $\beta_3 = 25.06$. Yes, these drastic changes in beta values are result of multicollinearity problem. We can verify this by calculating the correlation matrix.

```r
cor(FTCCIGAR)
```

```
##                TAR  NICOTINE    WEIGHT        CO
## TAR      1.0000000 0.9766076 0.4907654 0.9574853
## NICOTINE 0.9766076 1.0000000 0.5001827 0.9259473
## WEIGHT   0.4907654 0.5001827 1.0000000 0.4639592
## CO       0.9574853 0.9259473 0.4639592 1.0000000
```

7.20 Log-Log transformation

```r
load("rdata/EX7_20.Rdata")
p1 <- ggplot(EX7_20, aes(x = X, y = Y)) + geom_point() +
  ggtitle("Original Scatterplot") + stat_smooth(method = "lm", se = F)
p2 <- ggplot(EX7_20, aes(x = log(X), y = log(Y))) + geom_point() +
  ggtitle("Transformed Scatterplot") + stat_smooth(method = "lm", se = F)
grid.arrange(p1,p2, ncol = 2)
```

Looking at the scatterplot of original variables, they have strong negative linear relationship with some increments in exponential order. The log transformed scatterplot is much better.

```r
summary(lm(log(Y) ~ log(X), data = EX7_20))[4]
```

```
## $coefficients
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.636378  0.6028077  17.64473 2.733350e-08
## log(X)      -2.169853  0.1614126 -13.44290 2.911084e-07
```

```r
predict(lm(log(Y) ~ log(X), data = EX7_20), newdata = data.frame("X" = 30))
```

```
##       1
## 3.25628
```

At $\alpha = 0.05$ with p-value = 2.91e-07, as p-values is less than alpha we can cocnlude that the mode is adequate. the predicted values for $X = 30$ using the transformed model is 3.25

7.21 Multicollinearity in real estate data.

```r
load("rdata/HAMILTON.Rdata")
cor(HAMILTON$X1, HAMILTON$Y)
```

```
## [1] 0.002497966
```

```r
## no their correlation is extremely small which proves there is no linear correlation among them
cor(HAMILTON$X2, HAMILTON$Y)
```

```
## [1] 0.4340688
```

```r
## the correlation is more than the above one, but looking at the values it looks like
##the variable are linearly correlated but with few influential and leverage observations.
```

Yeah, since there is no multicorrelation among variable these variables can be used to predict sale price.

```r
lm_model <- lm(Y ~ ., data = HAMILTON)
## since the model has almost near perfect r square we can conclude that above statement is true.
cor(HAMILTON$X1, HAMILTON$X2)
```

```
## [1] -0.8997765
```

```r
## the correlation implies heavy neagative linear relation among x1 & x2
```

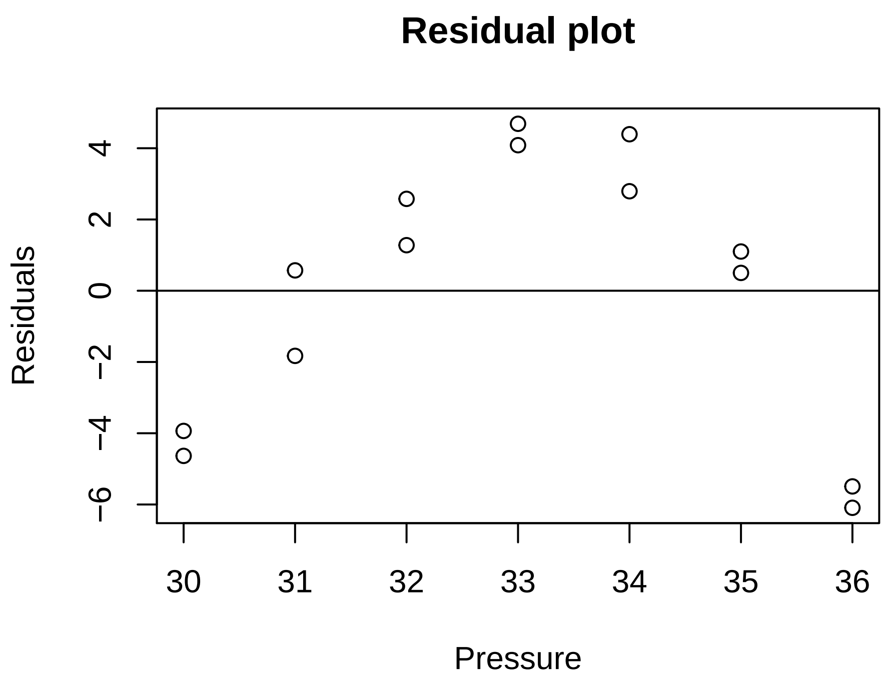
from my point of view, i wouldn't remove the other variable, i would have removed if the correlation is even higher and the fact that i could be overfitting, we need to test on more data.

8.3 New tire wear test.

```r
load("rdata/TIRES.Rdata")
lm_model <- lm(MILEAGE_Y ~ PRESS_X, data = TIRES)
summary(lm_model)[8]
```

```
## $r.squared
## [1] 0.01292991
```

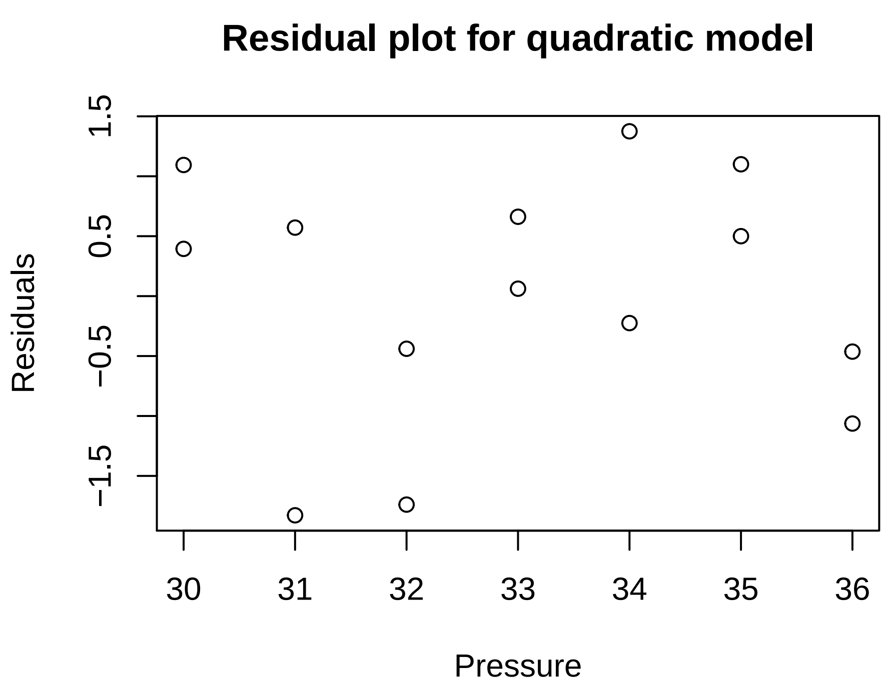
```r
rs <- resid(lm_model)
plot(TIRES$PRESS_X, rs, xlab = "Pressure", ylab = "Residuals", main = "Residual plot")
abline(h = 0)
```




```
## There is parabolic pattern in the residuals.
lm_model <- lm(MILEAGE_Y ~ poly(PRESS_X,2), data = TIRES)
summary(lm_model)[8]
```

```
## $r.squared
## [1] 0.9277414
```

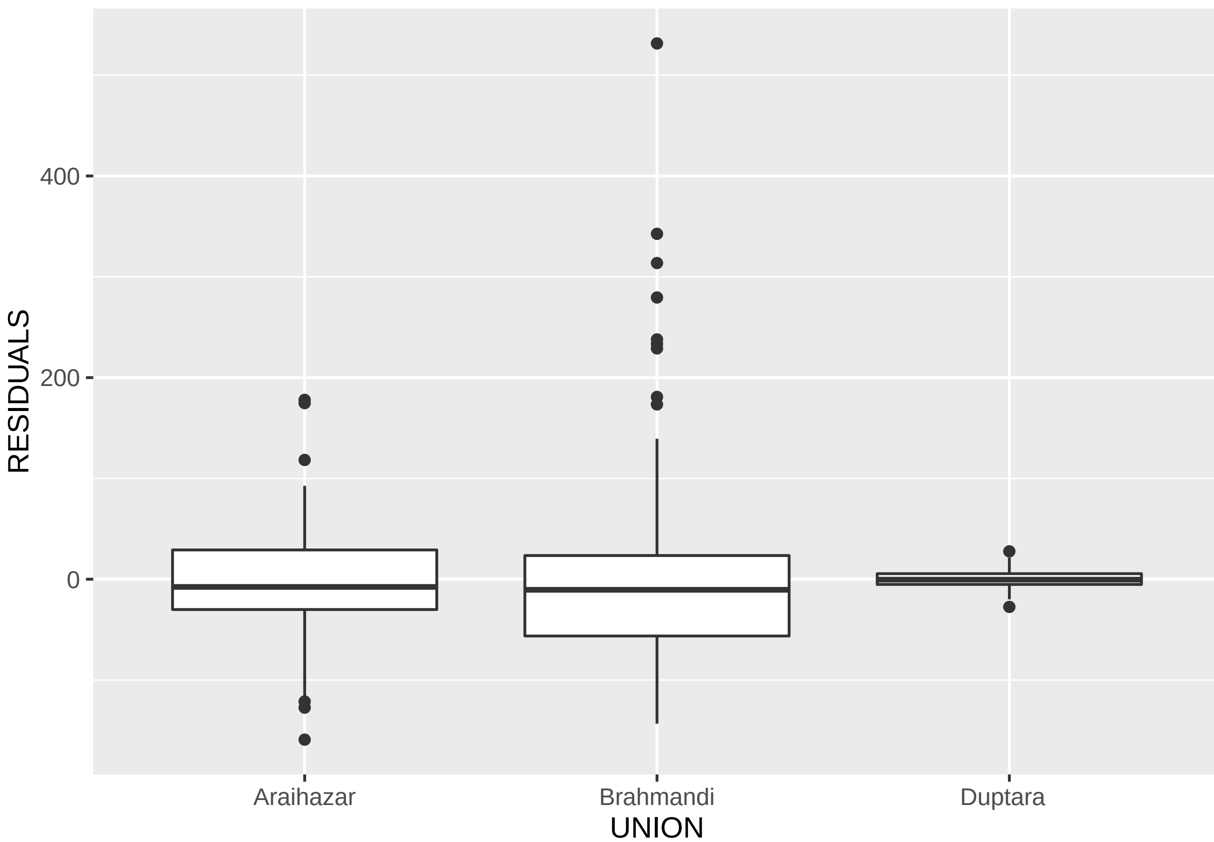
```
rs <- resid(lm_model)
plot(TIRES$PRESS_X, rs, xlab = "Pressure", ylab = "Residuals",
     main = "Residual plot for quadratic model")
```



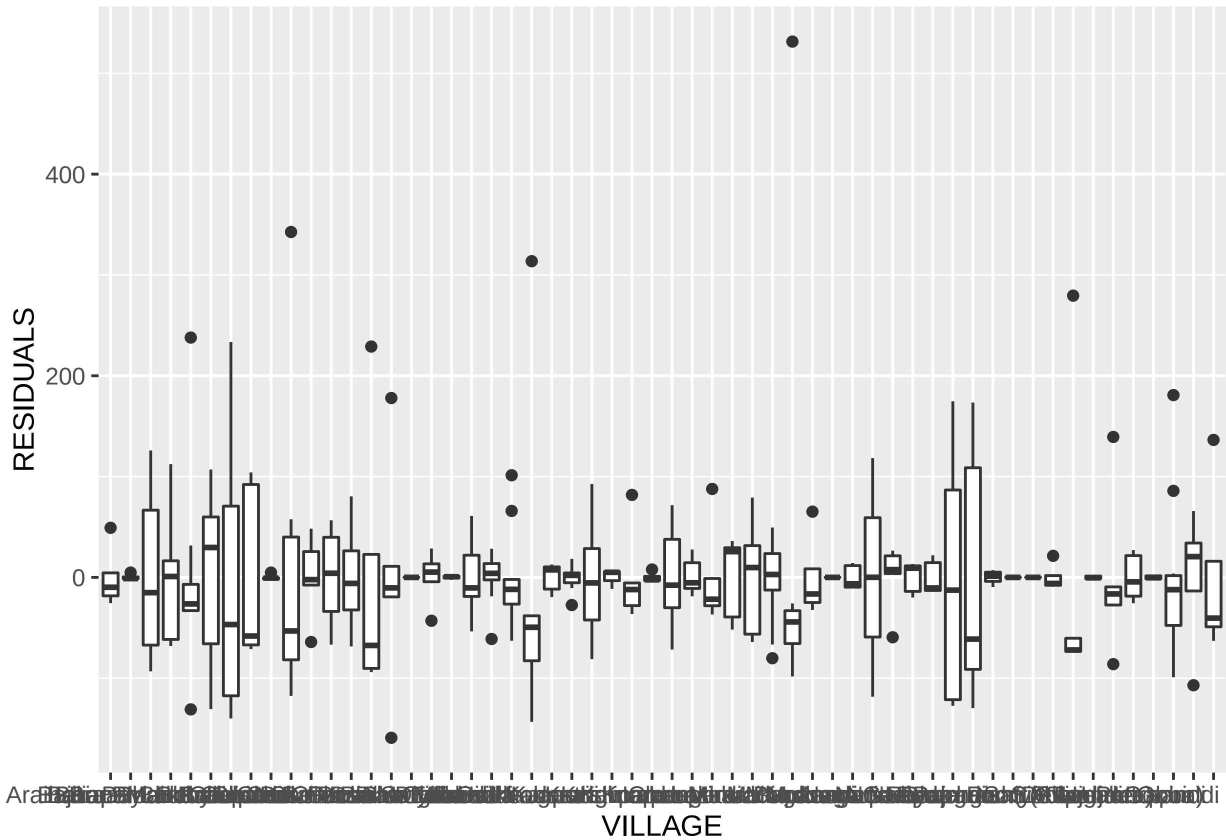
```
## As we can see from the model r square, the quadratic model has far more better accuracy.
```

8.29 Arsenic in groundwater

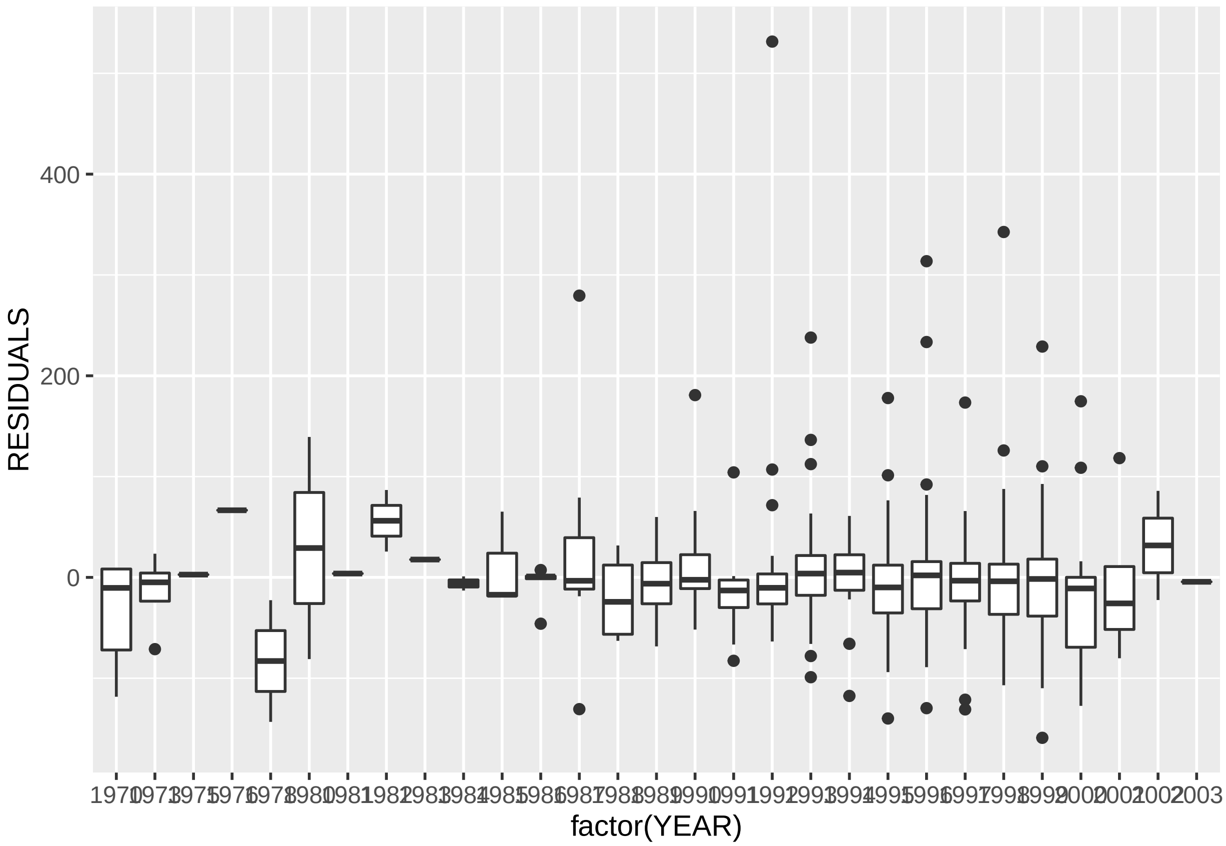
```r
load("rdata/ASWELLS.Rdata")
ASWELLS <- na.omit(ASWELLS)
lm_model <- lm(ARSENIC ~ . - WELLID, data = ASWELLS)
ASWELLS$RESIDUALS <- resid(lm_model)
ggplot(ASWELLS, aes(x = UNION, y = RESIDUALS)) + geom_boxplot()
```



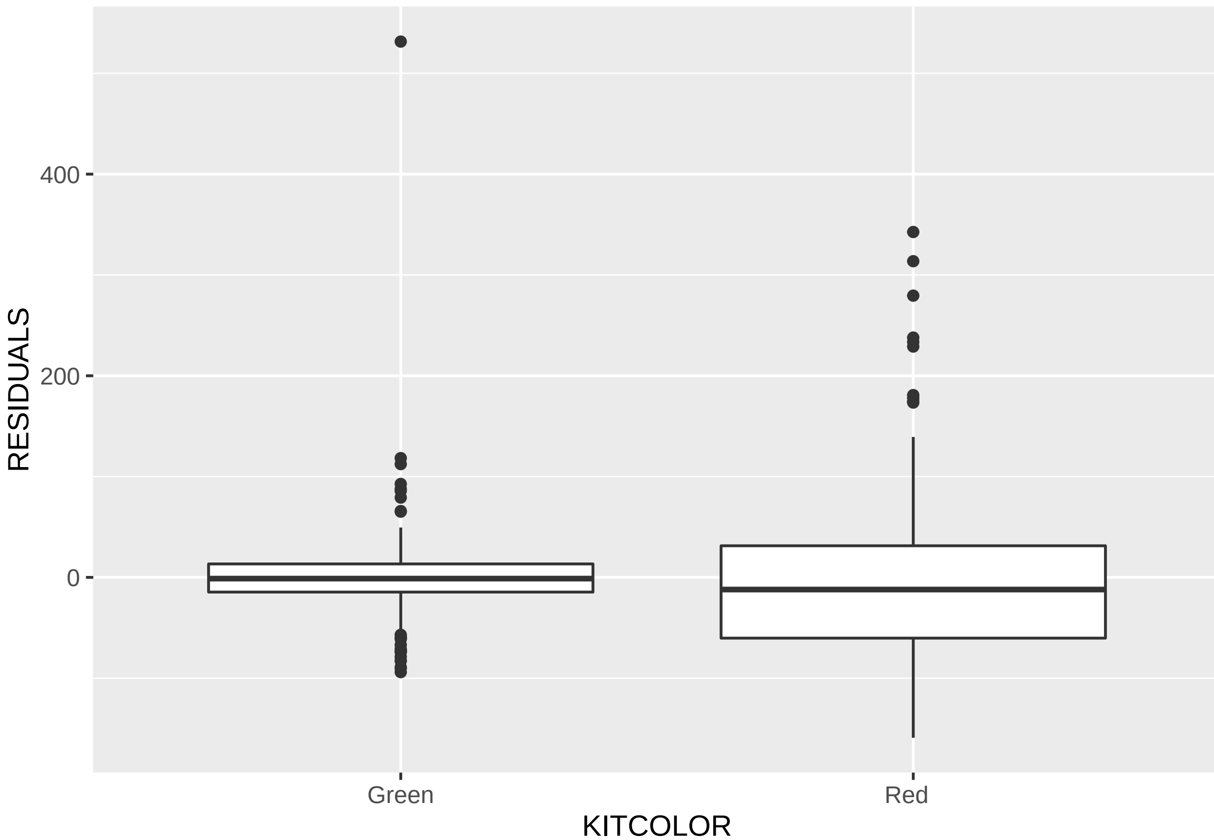

```r
ggplot(ASWELLS, aes(x = VILLAGE, y = RESIDUALS)) + geom_boxplot()
```



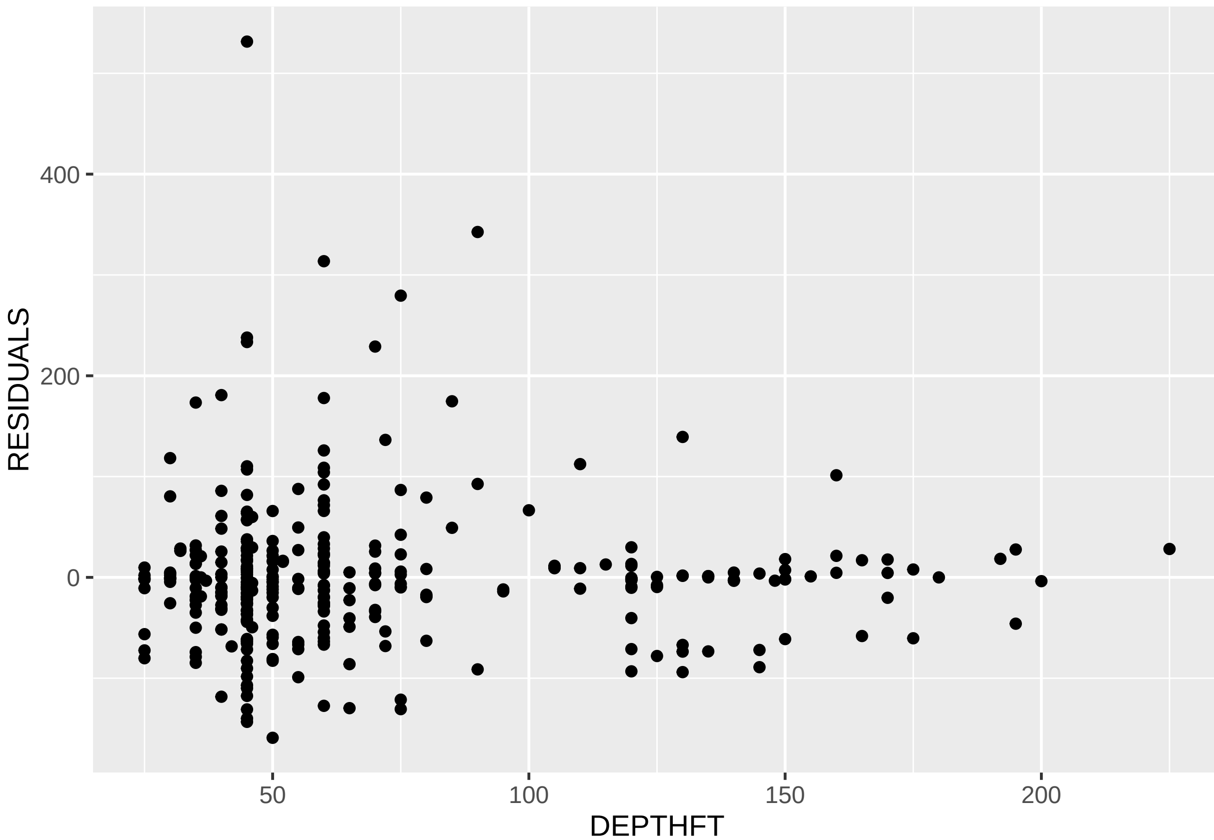

```r
ggplot(ASWELLS, aes(x = factor(YEAR), y = RESIDUALS)) + geom_boxplot()
```




```
ggplot(ASWELLS, aes(x = KITCOLOR, y = RESIDUALS)) + geom_boxplot()
```




```
ggplot(ASWELLS, aes(x = DEPTHFT, y = RESIDUALS)) + geom_point()
```




```r
good_data <- ASWELLS[ASWELLS$ARSENIC > 8.75 & ASWELLS$ARSENIC < 131.75,]
summary(lm(ARSENIC ~ . - WELLID, data = good_data))[8]
```

```
## $r.squared
## [1] 1
```

None of these outliers are influential data points, i have built two model one with original data and one without outliers. Both have near perfect r square.