diamonds

Chris Parrish

January 8, 2016

diamonds

references:
- Cannon, et al., Stat2, chapter 03, examples 3.11, 3.15

Import the data.

```
data <- read.csv("Diamonds.csv", header=TRUE)
head(data)
```

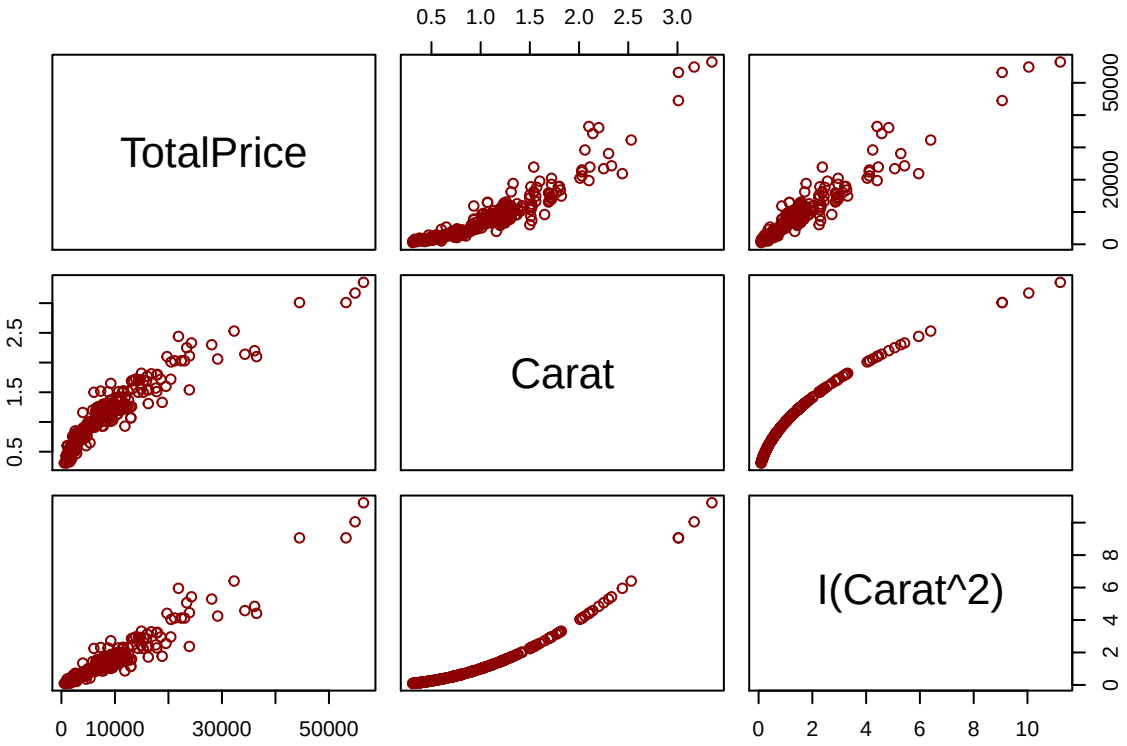
```
##   Carat Color Clarity Depth PricePerCt TotalPrice
## 1  1.08     E     VS1  68.6     6693.3     7228.8
## 2  0.31     F    VVS1  61.9     3159.0      979.3
## 3  0.31     H     VS1  62.1     1755.0      544.1
## 4  0.32     F    VVS1  60.8     3159.0     1010.9
## 5  0.33     D      IF  60.8     4758.8     1570.4
## 6  0.33     G    VVS1  61.5     2895.8      955.6
```

```
dim(data)
```

```
## [1] 351   6
```

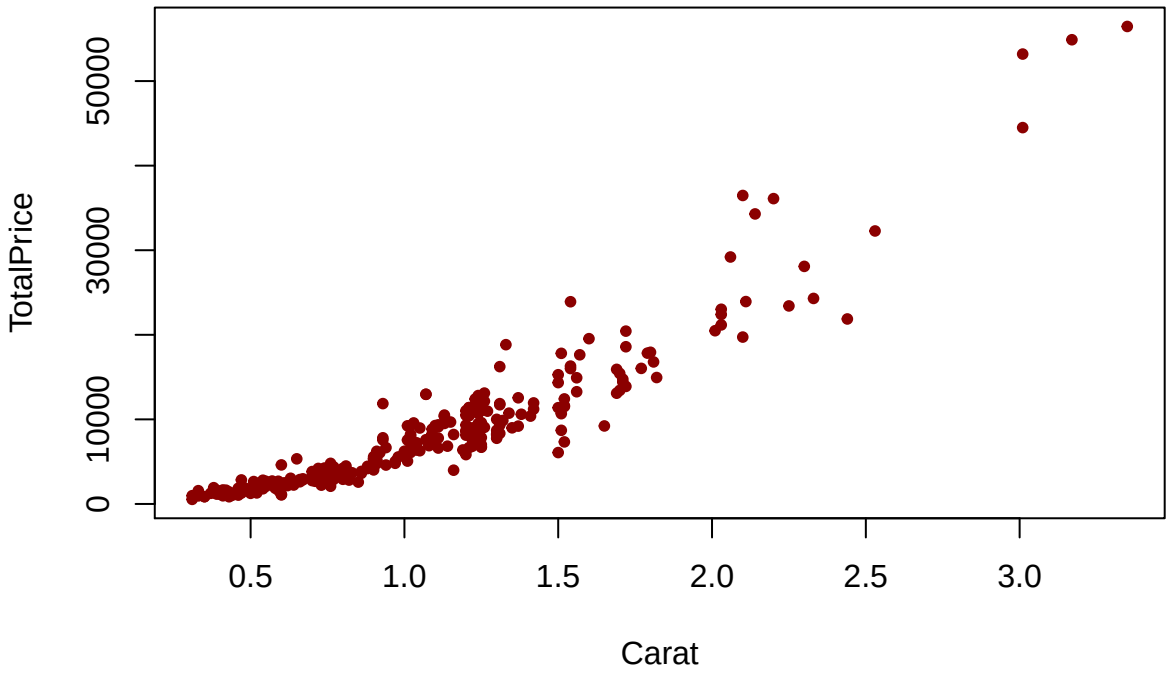
Scatterplot matrix.

```
pairs(~ TotalPrice + Carat + I(Carat^2), data=data, col="darkred")
```



Price ~ Carat

```
plot(TotalPrice ~ Carat, data=data,
     pch=20, col="darkred")
```



Quadratic linear model.

```
diamonds.lm <- lm(TotalPrice ~ Carat + I(Carat^2), data=data)
```

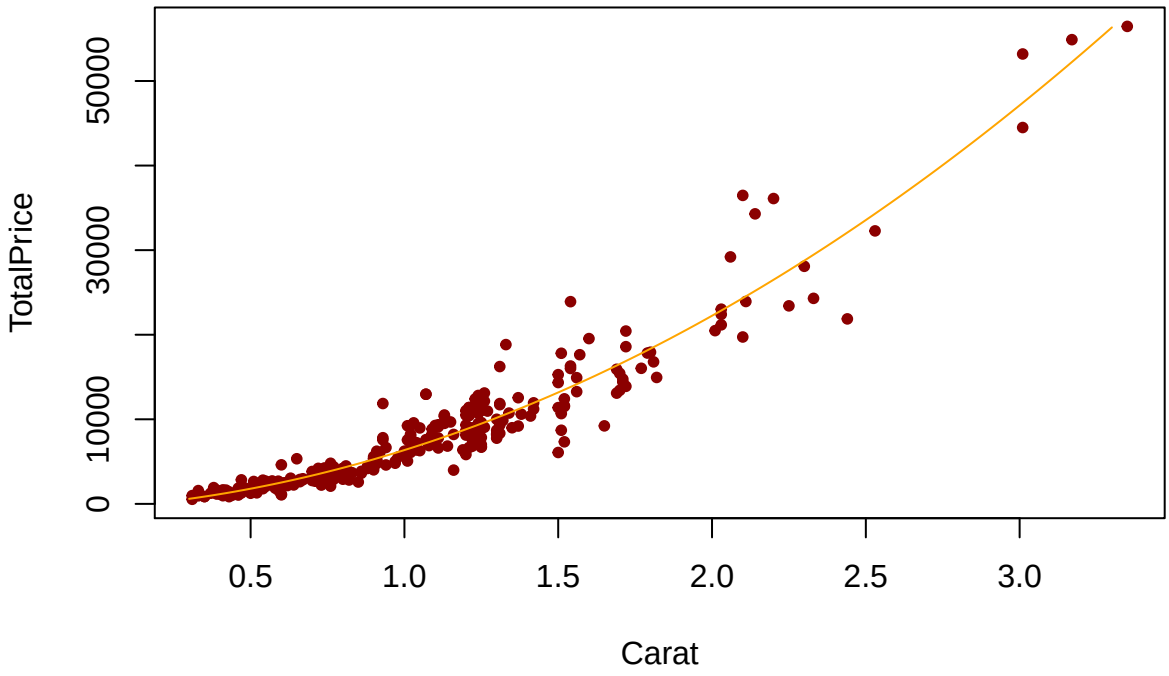
$\widehat{TotalPrice} = -522.702 + 2385.986\ Carat + 4498.206\ Carat^2$

```
options(show.signif.stars=FALSE)
summary(diamonds.lm)
```

```
##
## Call:
## lm(formula = TotalPrice ~ Carat + I(Carat^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10207.4   -711.6   -167.9    355.0  12147.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -522.7      466.3  -1.121  0.26307
## Carat         2386.0      752.5   3.171  0.00166
## I(Carat^2)    4498.2      263.0  17.101  < 2e-16
##
## Residual standard error: 2127 on 348 degrees of freedom
## Multiple R-squared:  0.9257, Adjusted R-squared:  0.9253
## F-statistic:  2168 on 2 and 348 DF,  p-value: < 2.2e-16
```

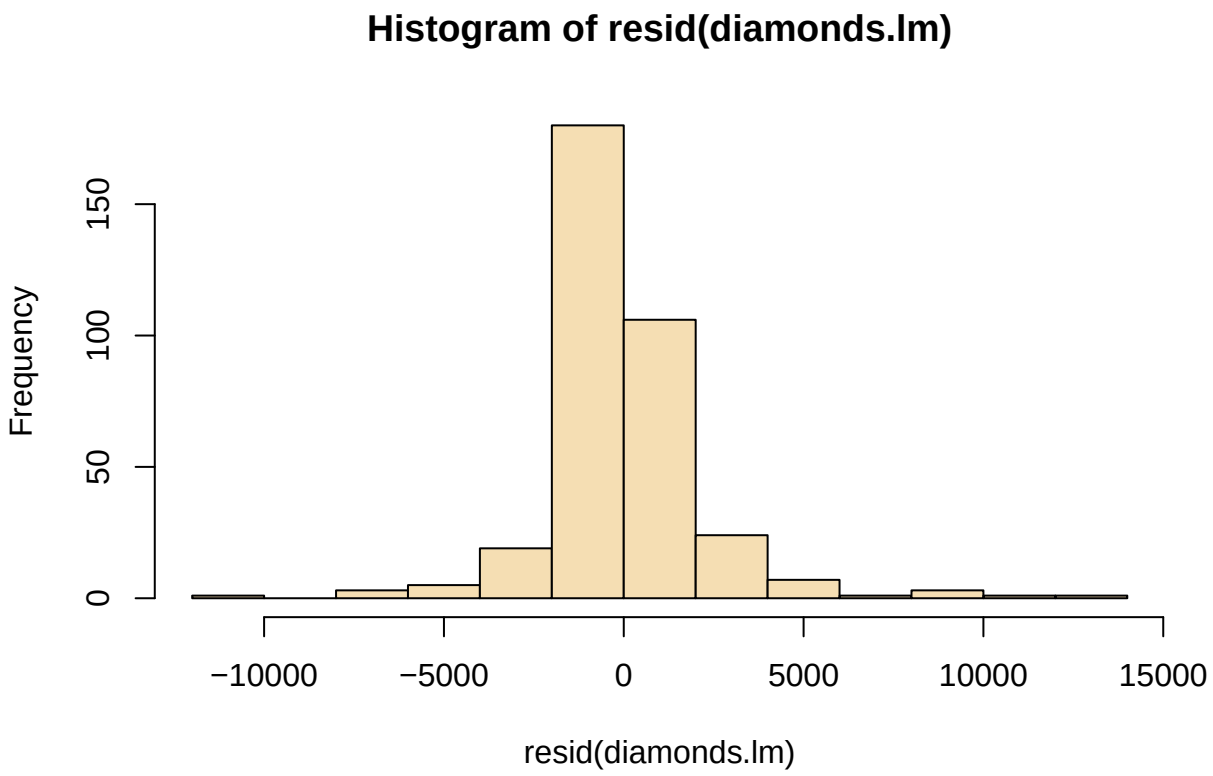
Illustration.

```
plot(TotalPrice ~ Carat, data=data,
     pch=20, col="darkred")
diamondPrice <- function(carat){
  a <- -522.7
  b <- 2386.0
  c <- 4498.2
  price <- a + b * carat + c * carat^2
  return(price)
}
curve(diamondPrice, from=0.3, to=3.3,
     col="orange", add=TRUE)
```

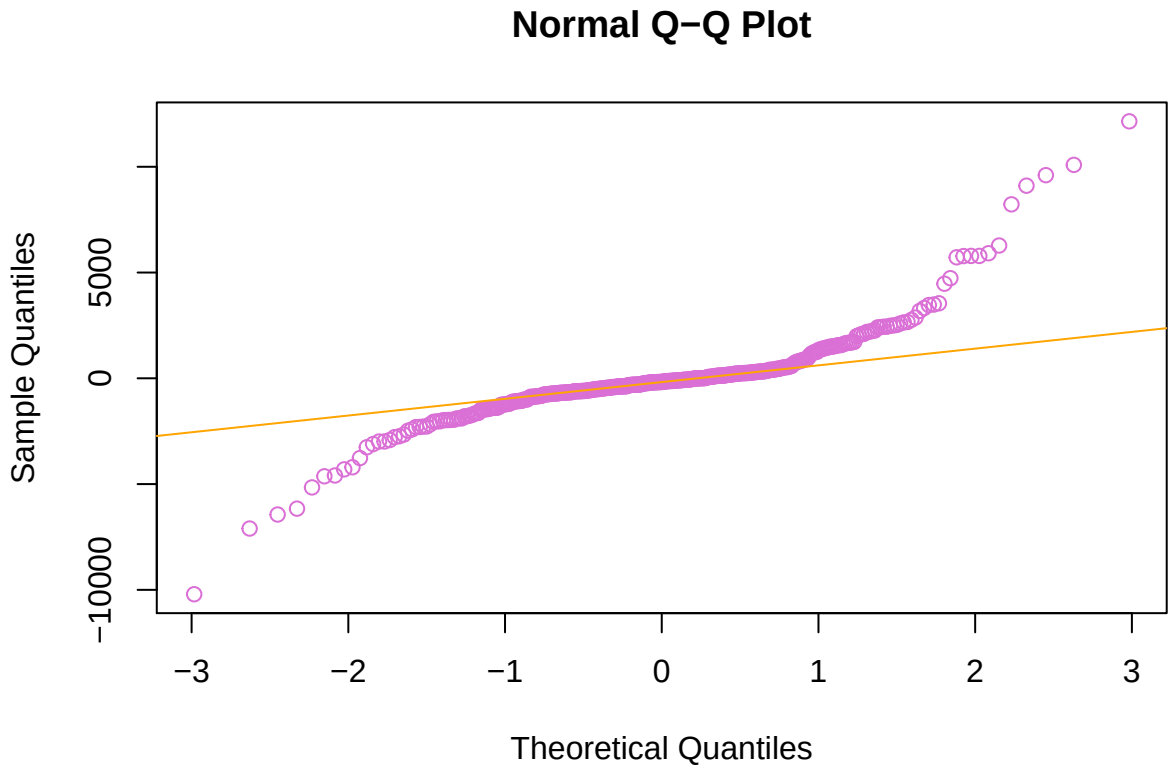


Residuals

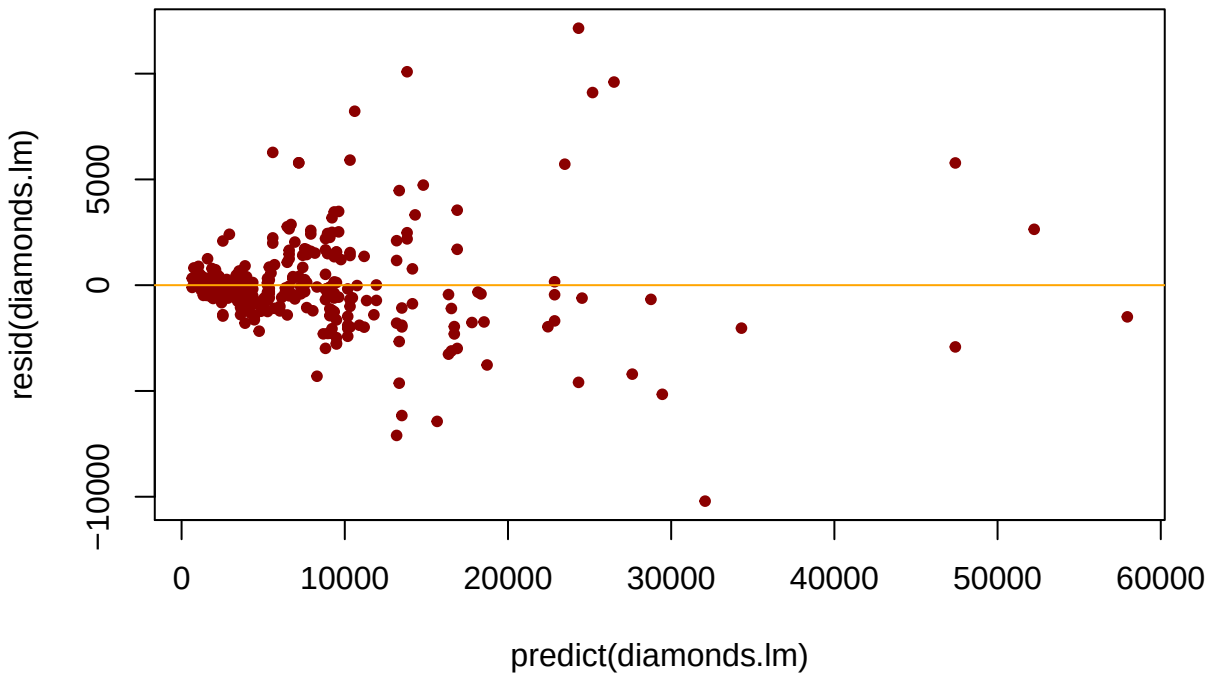
```
hist(resid(diamonds.lm),
     col="wheat")
```




```
qqnorm(resid(diamonds.lm),
       col="orchid")
qqline(resid(diamonds.lm),
       col="orange")
```



```r
plot(predict(diamonds.lm), resid(diamonds.lm),
     pch=20, col="darkred")
abline(h=0, col="orange")
```



VIF = variance inflation factor

$VIF_i > 5$ implies that $R_i^2 > 0.80$, so the ith variable is largely explained by the other variables.

```r
library(car)
diamonds.lm2 <- lm(TotalPrice ~ Carat + I(Carat^2) + Depth, data=data)
summary(diamonds.lm2)
```

```
##
## Call:
## lm(formula = TotalPrice ~ Carat + I(Carat^2) + Depth, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11166.7   -713.9    -52.7    563.9  11263.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6343.09    1436.49   4.416 1.35e-05
## Carat        2950.04     736.11   4.008 7.51e-05
## I(Carat^2)   4430.36     254.65  17.398  < 2e-16
## Depth        -114.08      22.66  -5.034 7.74e-07
##
## Residual standard error: 2056 on 347 degrees of freedom
## Multiple R-squared:  0.9308, Adjusted R-squared:  0.9302
## F-statistic:  1555 on 3 and 347 DF,  p-value: < 2.2e-16
```

```r
vif(diamonds.lm2)
```

```
##      Carat I(Carat^2)      Depth
##  10.942252  10.718736   1.117426
```