A16 - Machine learning (regression)

Bryana Benson
Conner Bryan

Read the Manheim case and then build a linear regression model for predicting the selling price of a car using http://www.richardtwatson.com/data/manheim.csv. Follow the principles for residual analysis.

Libraries

```
library(readr)
library(ggplot2)
library(dplyr)
```

Read Data

```
data <- read_csv("http://www.richardtwatson.com/data/manheim.csv")

head(data)
```

```
## # A tibble: 6 x 4
##   model price miles sale
##   <chr> <dbl> <dbl> <chr>
## 1 Y     23200 41430 Auction
## 2 Y     23100 42524 Auction
## 3 Y     23100 42692 Auction
## 4 Y     23200 39911 Auction
## 5 Y     24500 33199 Online
## 6 Y     22600 43090 Auction
```

Linear Regression

```
reg <- lm(price ~ sale + miles + model, data = data)
summary(reg)
```

```
##
## Call:
## lm(formula = price ~ sale + miles + model, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13810.0   -805.2     86.8    867.7   4461.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.289e+04  1.908e+02 119.928  < 2e-16 ***
## saleOnline   5.917e+02  1.288e+02   4.595 5.01e-06 ***
## miles       -1.272e-01  4.350e-03 -29.229  < 2e-16 ***
## modelY       5.617e+03  1.367e+02  41.103  < 2e-16 ***
## modelZ       1.224e+04  1.137e+02 107.691  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1420 on 814 degrees of freedom
## Multiple R-squared:  0.9357, Adjusted R-squared:  0.9354
## F-statistic:  2963 on 4 and 814 DF,  p-value: < 2.2e-16
```

$$\hat{Price} = \beta_0 + \beta_1(Sale) + \beta_2(Miles) + \beta_3(ModelY) + \beta_4(ModelZ)$$

From the model, the average price of a vehicle, holding all else constant, is $22,890. If the car was an online sale, the price increased by $591.70. For every additional mile on a car, the price decreases by $0.01. Relative to ModelX, ModelY cars are worth $5,617 more. A ModelZ car is worth $12,240 than the ModelX cars. Overall, this model explains about 94% of the data and is significant at the 0.01 level.

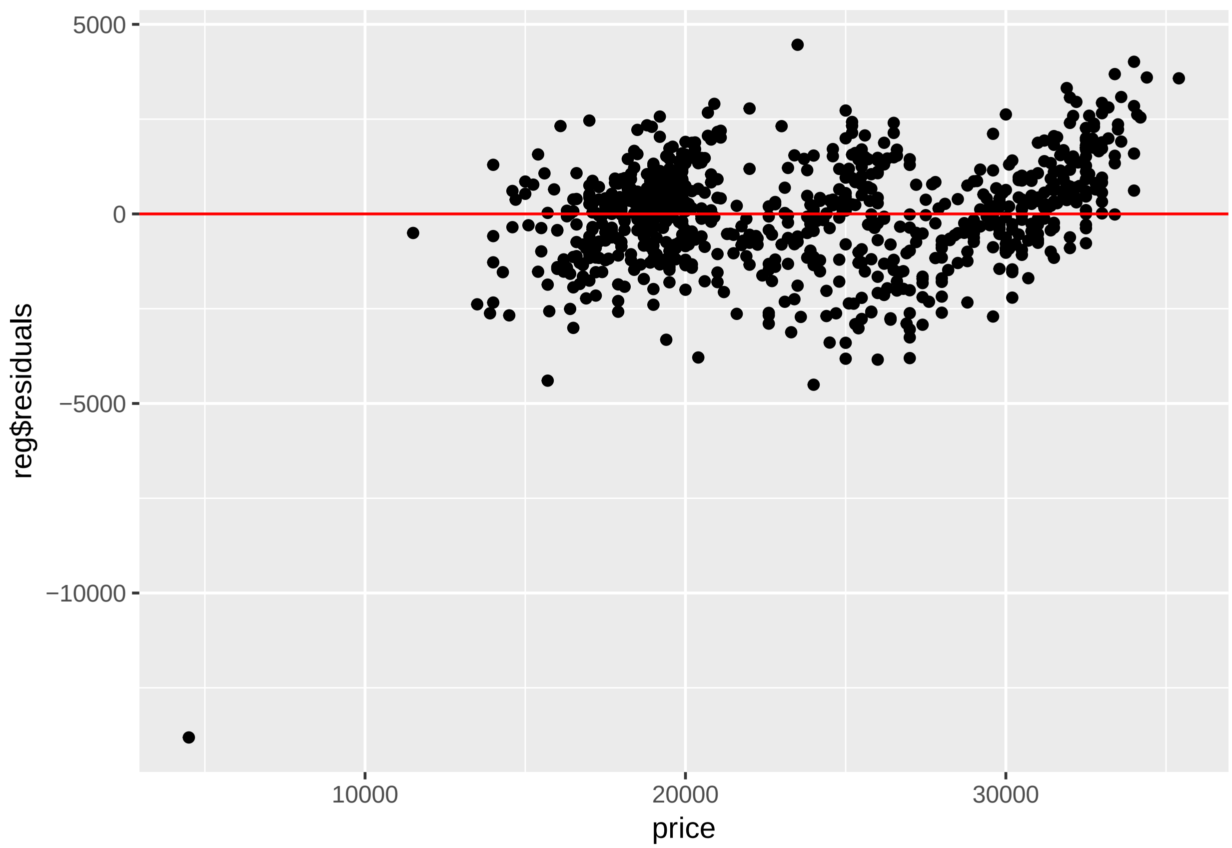
Residual Analysis

Add residuals to dataset

```r
data$residuals <- reg$residuals
```

Plot residuals

```r
ggplot(data,aes(price,reg$residuals)) +
geom_point() +
geom_hline(yintercept = 0, color='red')
```



Run Shapiro Test for Outliers

```r
shapiro.test(reg$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reg$residuals
## W = 0.9434, p-value < 2.2e-16
```

P-value is less than 0.05, so the null hypothesis that the data is normally distributed is rejected. There is evidence of outliers in the data.

Remove Outliers and re-run Shapiro-Wilk normality test

```r
st <- shapiro.test(data$residuals)

while(st$p.value < .05) {
data <- data %>% filter((data$residuals) < max(data$residuals))

mod <- lm(data$price ~ data$miles + data$model + data$sale)

data$residuals <- mod$residuals

st <- shapiro.test(data$residuals)
}

st
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$residuals
## W = 0.93159, p-value = 0.09459
```

P-value is greater than 0.05, so the null hypothesis that the data is normally distributed is not rejected. The data is more normally distributed.

Re-run Regression Model

```r
reg <- lm(price ~ sale + miles + model, data = data)
summary(reg)
```

```
##
## Call:
## lm(formula = price ~ sale + miles + model, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -785.86 -172.24   31.41  188.12  364.13
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.139e+03  4.719e+02  15.126 2.06e-12 ***
## saleOnline   1.347e+04  3.263e+02  41.287  < 2e-16 ***
## miles       -7.329e-02  1.029e-02  -7.125 6.65e-07 ***
## modelY       5.799e+03  1.833e+02  31.646  < 2e-16 ***
```

```
## modelZ       1.228e+04  1.864e+02  65.858  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292.6 on 20 degrees of freedom
## Multiple R-squared:  0.9981, Adjusted R-squared:  0.9978
## F-statistic:  2693 on 4 and 20 DF,  p-value: < 2.2e-16
```

This model is run with the normalized data that excludes outliers. As a result, the estimates are no longer impacted by those values.

$$\hat{Price} = \beta_0 + \beta_1(Sale) + \beta_2(Miles) + \beta_3(ModelY) + \beta_4(ModelZ)$$
$$\hat{Price} = 7,129 + 1,347(Sale) + .073(Miles) + 5,799(ModelY) + 12,280(ModelZ)$$