Linear Regression: Visualizations of Predictions and Residuals

Yogindra Raghav

November 9, 2018

1. NOTE: Only the first 10 rows from adding predictions are shown for the sake of saving space since the data set is long.

```
library(modelr)

library(mosaicData)

library(dplyr)

library(ggplot2)

mod1 <- lm(volume ~ hightemp, data = RailTrail)
RailTrail %>% add_predictions(mod1) %>% head(10)
```

```
##    hightemp lowtemp avgtemp spring summer fall cloudcover precip volume
## 1        83      50    66.5      0      1    0        7.6   0.00    501
## 2        73      49    61.0      0      1    0        6.3   0.29    419
## 3        74      52    63.0      1      0    0        7.5   0.32    397
## 4        95      61    78.0      0      1    0        2.6   0.00    385
## 5        44      52    48.0      1      0    0       10.0   0.14    200
## 6        69      54    61.5      1      0    0        6.6   0.02    375
## 7        66      39    52.5      1      0    0        2.4   0.00    417
## 8        66      38    52.0      1      0    0        0.0   0.00    629
## 9        80      55    67.5      0      1    0        3.8   0.00    533
## 10       79      45    62.0      0      1    0        4.1   0.00    547
##    weekday dayType     pred
## 1     TRUE weekday 456.1766
## 2     TRUE weekday 399.1578
## 3     TRUE weekday 404.8597
## 4    FALSE weekend 524.5991
## 5     TRUE weekday 233.8034
## 6     TRUE weekday 376.3503
## 7     TRUE weekday 359.2447
## 8    FALSE weekend 359.2447
## 9    FALSE weekend 439.0710
## 10    TRUE weekday 433.3691
```

2. Resulting data set has more points than the originial RailTrail dataset.

```
grid2 <- RailTrail %>% data_grid(weekday, hightemp, cloudcover, precip)

grid2
```

```
## # A tibble: 95,040 x 4
##    weekday hightemp cloudcover  precip
##    <lgl>      <int>      <dbl>   <dbl>
##  1 FALSE         41          0 0
##  2 FALSE         41          0 0.01000
##  3 FALSE         41          0 0.0200
##  4 FALSE         41          0 0.0300
##  5 FALSE         41          0 0.120
##  6 FALSE         41          0 0.140
##  7 FALSE         41          0 0.150
##  8 FALSE         41          0 0.160
##  9 FALSE         41          0 0.170
## 10 FALSE         41          0 0.200
## # ... with 95,030 more rows
```

```
nrow(grid2)
```

```
## [1] 95040
```

```
nrow(RailTrail)
```

```
## [1] 90
```

3. Visualization of predictions and residuals

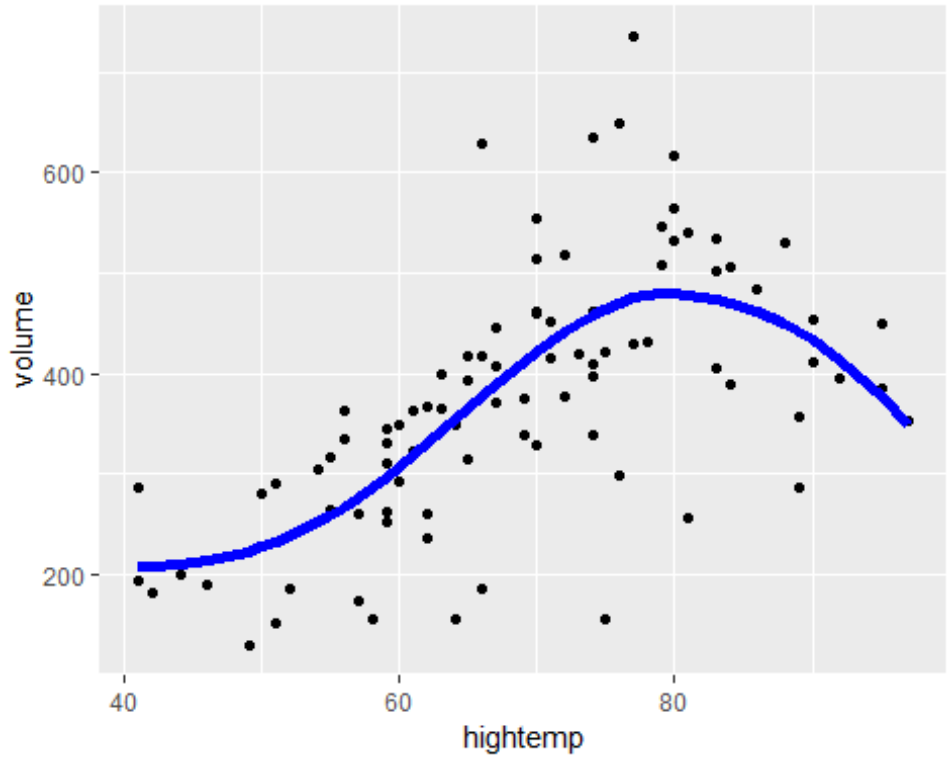
```
mod2 <- loess(volume ~ hightemp, data = RailTrail)
mod2
```

```
## Call:
## loess(formula = volume ~ hightemp, data = RailTrail)
##
## Number of Observations: 90
## Equivalent Number of Parameters: 4.92
## Residual Standard Error: 94.38
```

```
hightemp_grid = RailTrail %>% data_grid(hightemp)

hightemp_grid = hightemp_grid %>% add_predictions(mod2)
```

```
ggplot(RailTrail, aes(hightemp))+ geom_point(aes(y=volume))+ geom_line(aes(y
= pred), data = hightemp_grid, colour = "blue", size =2)
```




```
resid_railtrail = RailTrail %>% add_residuals(mod2) %>% select(hightemp, volu
me, resid)

head(resid_railtrail)
```

```
##   hightemp volume      resid
## 1       83    501  27.289840
## 2       73    419 -30.082510
## 3       74    397 -59.688367
## 4       95    385   8.250826
## 5       44    200  -9.159302
## 6       69    375 -35.727228
```

```
ggplot(resid_railtrail, aes(hightemp, resid))+ geom_ref_line(h =0)+ geom_poin
t()
```