cleaning

Chris Parrish

December 12, 2016

Contents

cleaning

reference:
- Sheather, A Modern Approach to Regression with R, chapter 3, pp.71-76

Load package.

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

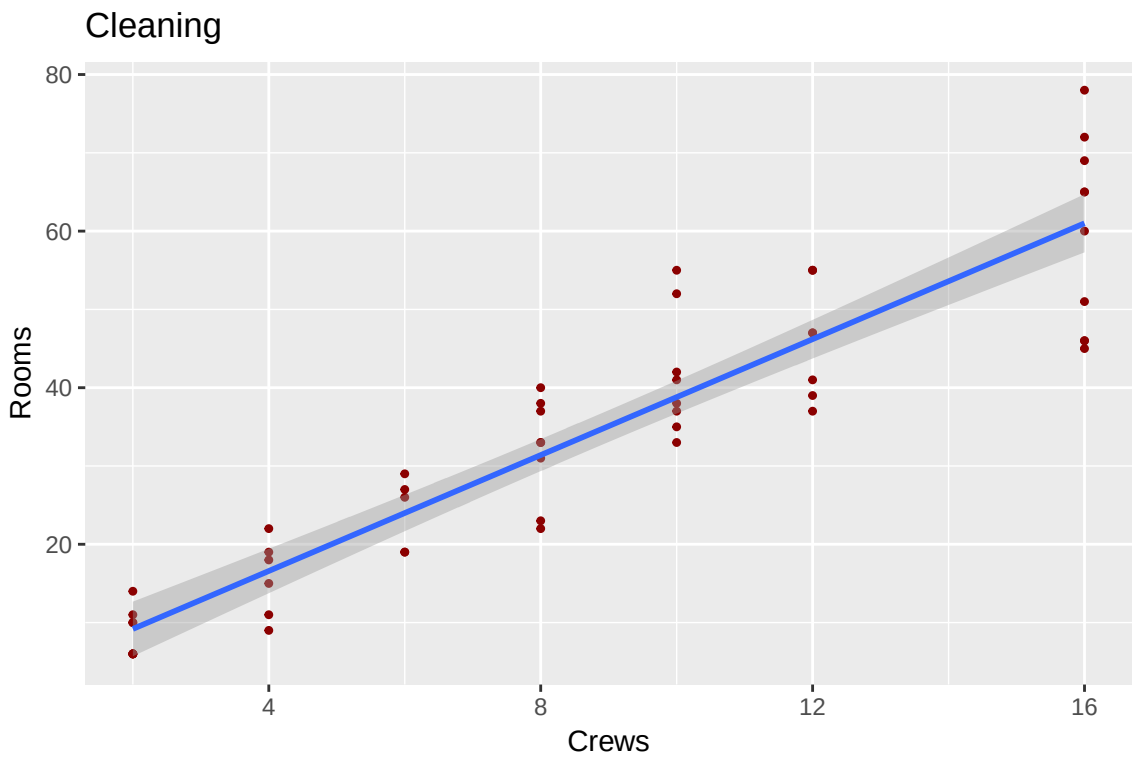
Import the data.

```
data <- read.delim("cleaning.txt", header=TRUE)
head(data)
```

```
##   Case Crews Rooms
## 1    1    16    51
## 2    2    10    37
## 3    3    12    37
## 4    4    16    46
## 5    5    16    45
## 6    6     4    11
```

linear model

```
ggplot(data, aes(Crews, Rooms)) +
  geom_point(shape = 20, color = "darkred") +
  geom_smooth(method = "lm") +
  ggtitle("Cleaning")
```




```
cleaning.lm <- lm(Rooms ~ Crews, data=data)
summary(cleaning.lm)
```

```
##
## Call:
## lm(formula = Rooms ~ Crews, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.9990  -4.9901   0.8046   4.0010  17.0010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7847     2.0965   0.851    0.399
## Crews         3.7009     0.2118  17.472   <2e-16
##
## Residual standard error: 7.336 on 51 degrees of freedom
## Multiple R-squared:  0.8569, Adjusted R-squared:  0.854
## F-statistic: 305.3 on 1 and 51 DF,  p-value: < 2.2e-16
```

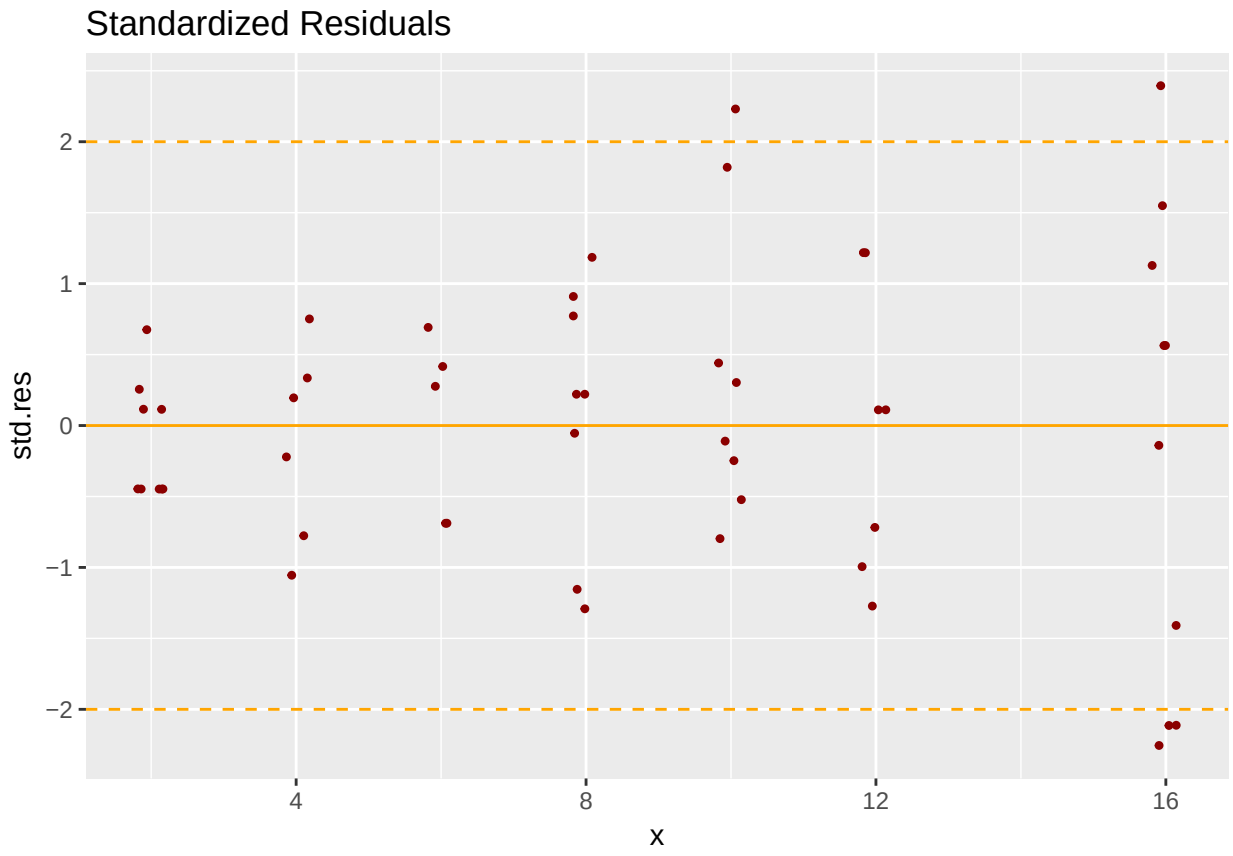
prediction intervals

```
new.data = data.frame(Crews = c(4, 16))
predict(cleaning.lm, new.data, interval="prediction")
```

```
##        fit      lwr      upr
## 1 16.58827  1.58941 31.58713
## 2 60.99899 45.81025 76.18773
```

standardized residuals

```
lm.data <- data.frame(x = data$Crews,
                      std.res = rstandard(cleaning.lm))
ggplot(lm.data, aes(x, std.res)) +
  geom_jitter(shape = 20, color = "darkred", width = 0.2) +
  geom_hline(yintercept = 0, color = "orange") +
  geom_hline(yintercept = c(-2, 2), color = "orange", lty = 2) +
  ggtitle("Standardized Residuals")
```

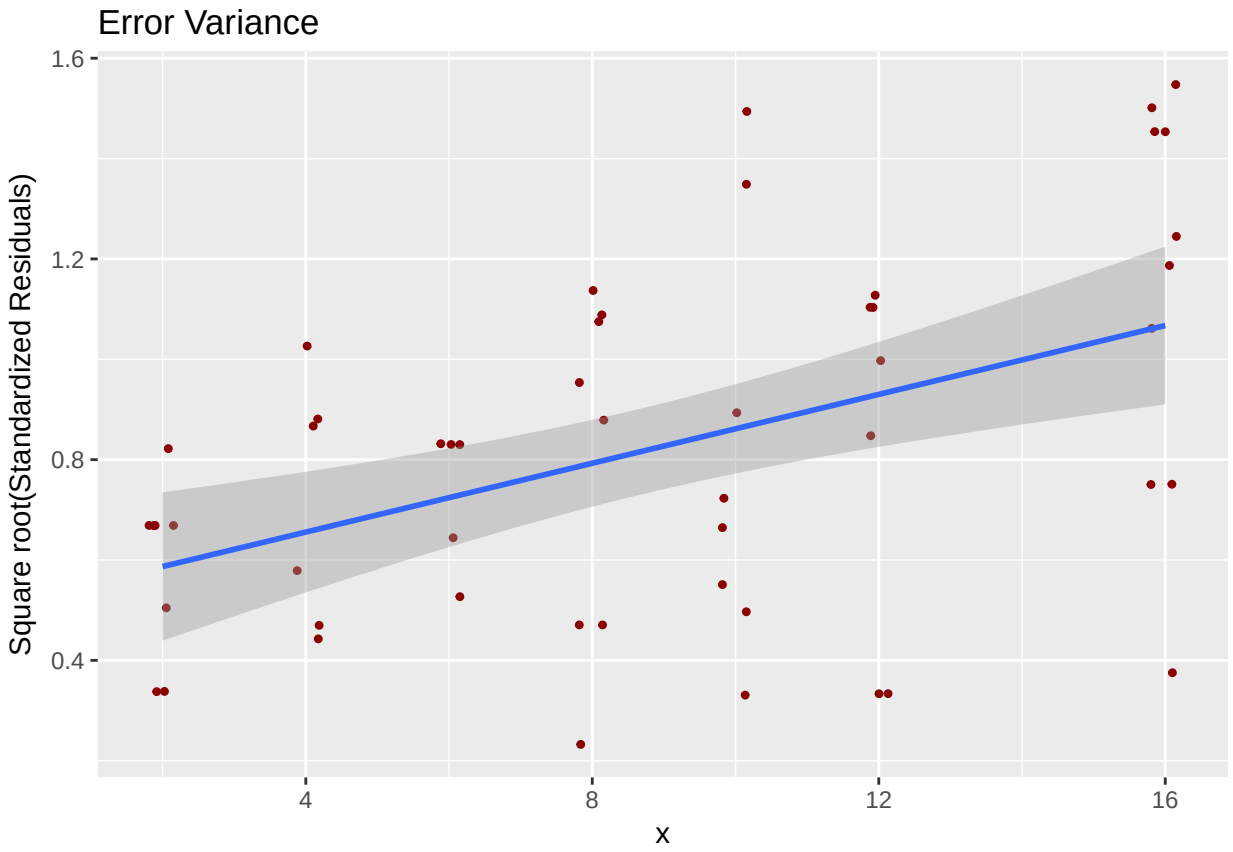


error variance

```
lm.data$sabs <- sqrt(abs(lm.data$std.res))
head(lm.data)
```

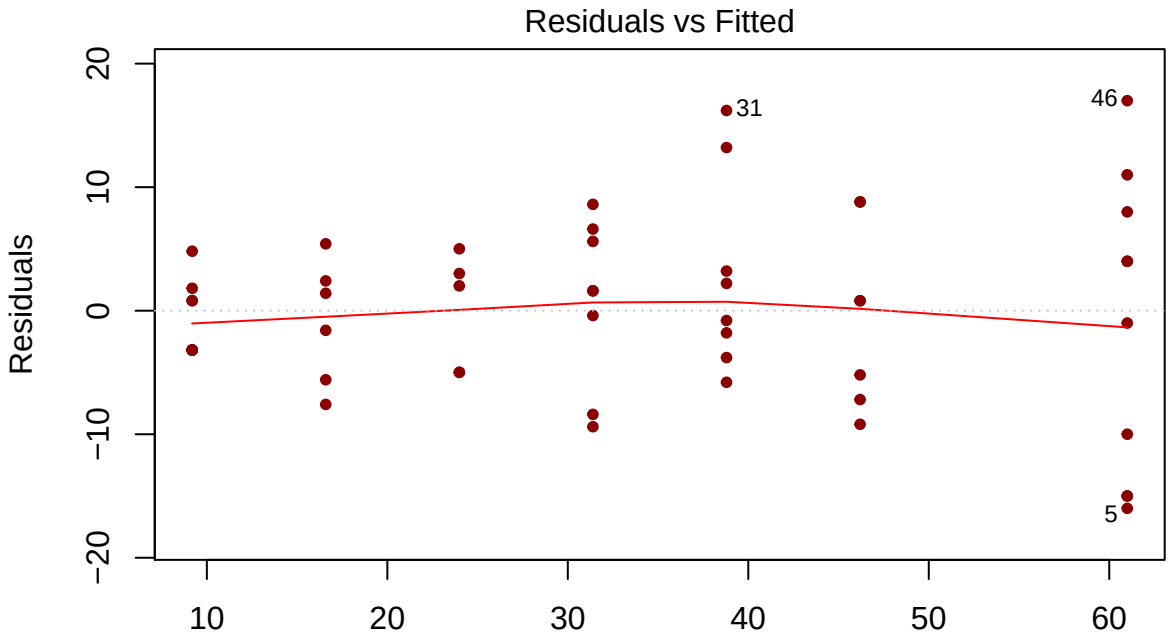
```
##    x    std.res      sabs
## 1 16 -1.4084554 1.1867836
## 2 10 -0.2470142 0.4970052
## 3 12 -1.2713996 1.1275636
## 4 16 -2.1127540 1.4535316
## 5 16 -2.2536137 1.5012041
## 6  4 -0.7762860 0.8810709
```

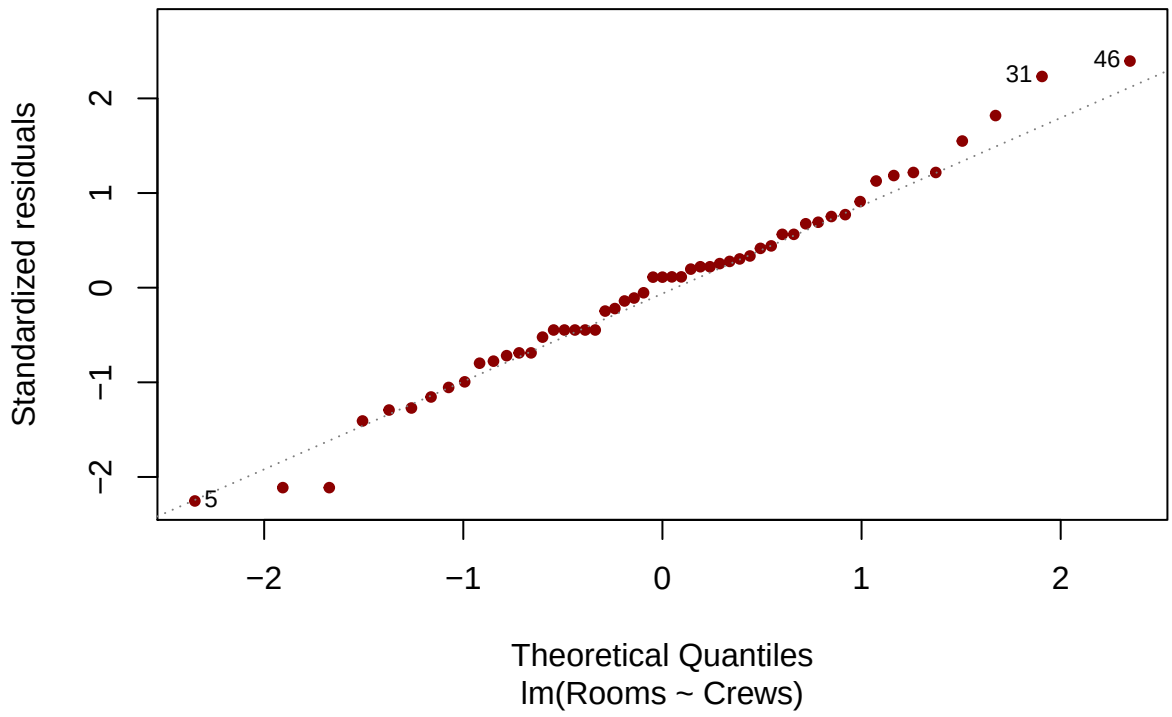
```
ggplot(lm.data, aes(x, sabs)) +
  geom_jitter(shape = 20, color = "darkred", width = 0.2) +
  geom_smooth(method = "lm") +
  ylab("Square root(Standardized Residuals)") +
  ggtitle("Error Variance")
```

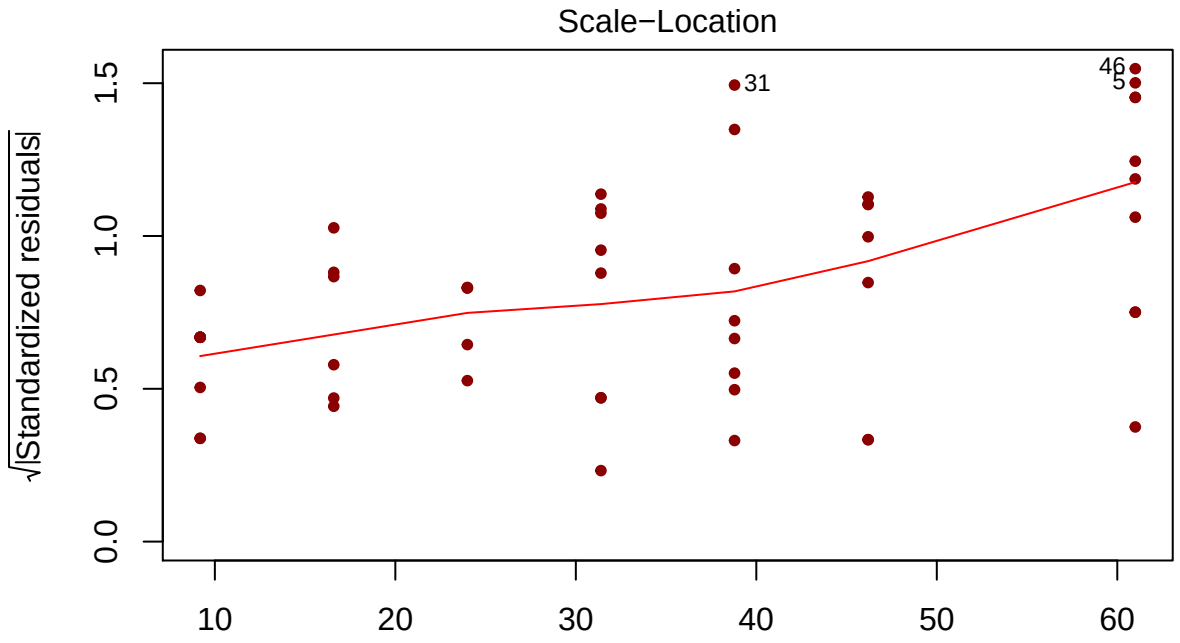


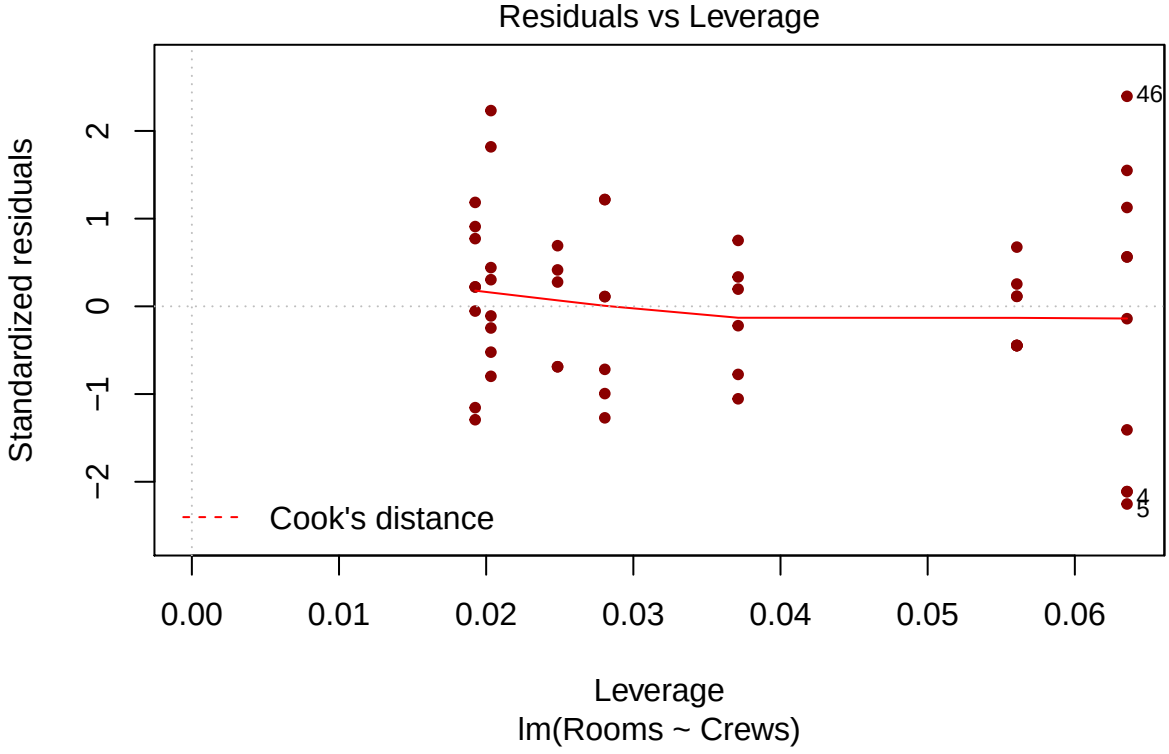
regression diagnostics

```
plot(cleaning.lm,
     pch=20, col="darkred")
```






Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
20
30
40
50
60
31
46
5
Fitted values
lm(Rooms ~ Crews)


Residuals vs Leverage
Standardized residuals
−2
−1
0
1
2
0.00
0.01
0.02
0.03
0.04
0.05
0.06
46
4
5
Cook's distance
Leverage
lm(Rooms ~ Crews)

standard deviation of $Y \mid X = x$

```
sd.data <- data %>%
  group_by(Crews) %>%
  summarise(n = n(),
            sd = sd(Rooms))
sd.data
```

```
## # A tibble: 7 × 3
##   Crews     n        sd
##   <int> <int>     <dbl>
## 1     2     9  3.000000
## 2     4     6  4.966555
## 3     6     5  4.690416
## 4     8     8  6.642665
## 5    10     8  7.927123
## 6    12     7  7.289915
## 7    16    10 12.000463
```

```
ggplot(sd.data, aes(Crews, sd)) +
  geom_point(shape = 20, color = "darkred") +
  geom_smooth(method = "lm") +
  ylab("Standard deviation(rooms cleaned)") +
  ggtitle("Standard Deviation of Y Given X")
```