

salary

Chris Parrish

December 18, 2016

salary

reference:

- Sheather, section 5.1, pp.125-130

Load package.

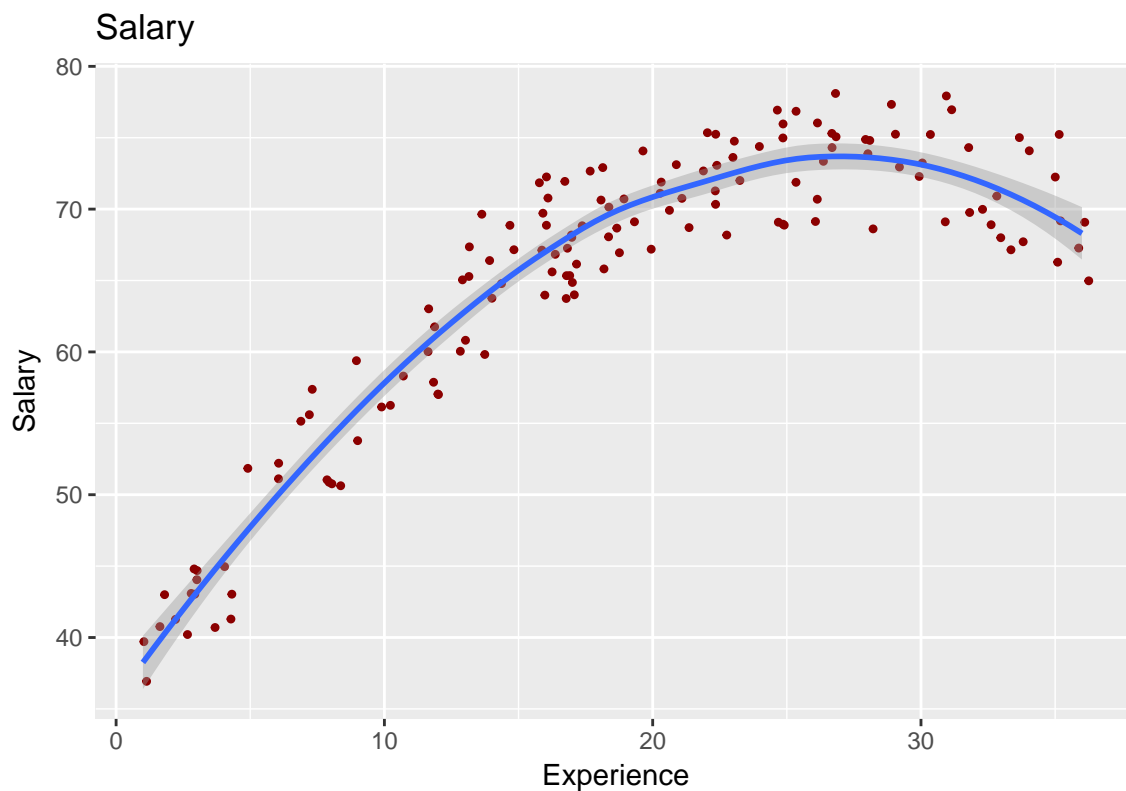
```
library(ggplot2)
```

Import the data.

```
data <- read.delim("profsalary.txt", header = TRUE)
head(data, 4)
```

```
##   Case Salary Experience
## 1     1     71         26
## 2     2     69         19
## 3     3     73         22
## 4     4     69         17
```

```
ggplot(data, aes(Experience, Salary)) +
  geom_jitter(width = 0.4, shape = 20, color = "darkred") +
  geom_smooth() +
  ggtitle("Salary")
```

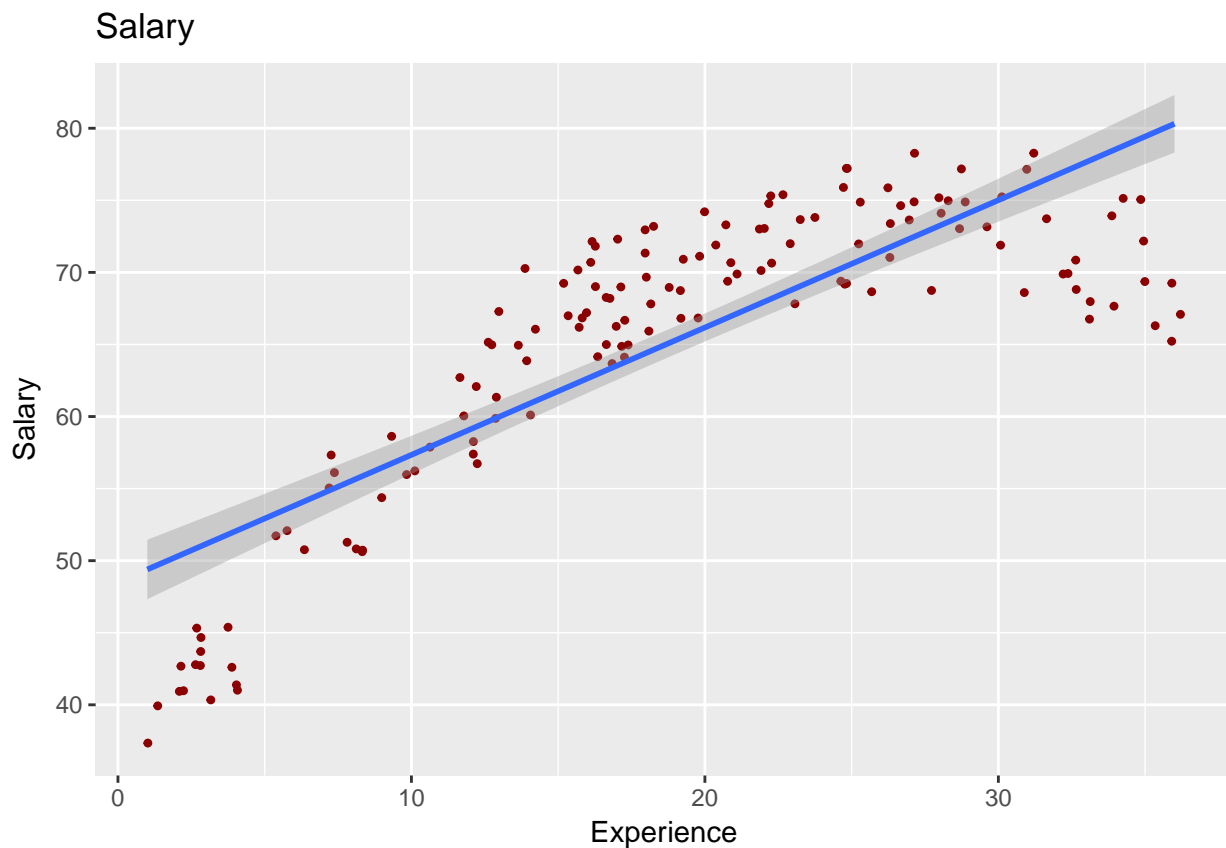


Linear model.

```
salary.lm1 <- lm(Salary ~ Experience, data = data)
options(show.signif.stars = FALSE)
summary(salary.lm1)
```

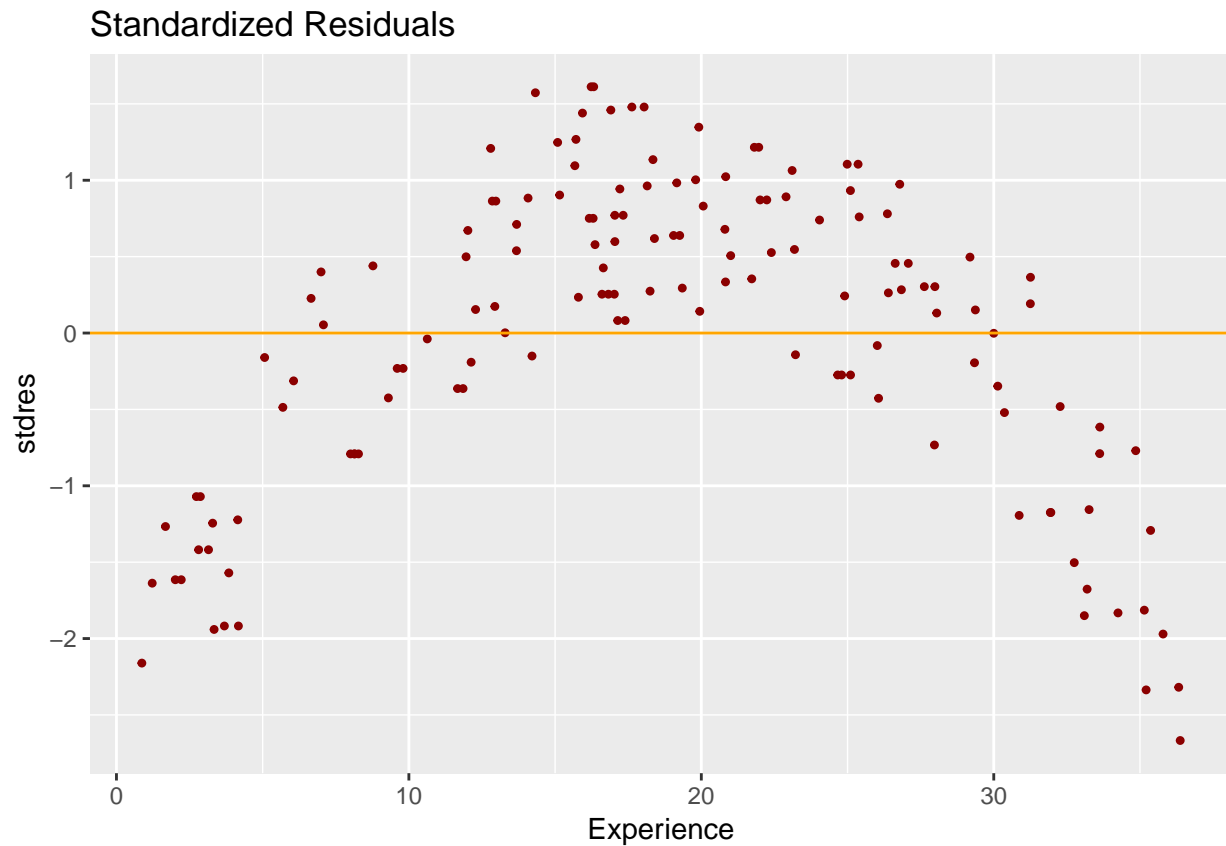
```
##
## Call:
## lm(formula = Salary ~ Experience, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.310  -3.893   1.408   4.442   9.359
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.50593    1.08810   44.58  <2e-16
## Experience   0.88345    0.05158   17.13  <2e-16
##
## Residual standard error: 5.828 on 141 degrees of freedom
## Multiple R-squared:  0.6754, Adjusted R-squared:  0.6731
## F-statistic: 293.3 on 1 and 141 DF,  p-value: < 2.2e-16
```

```
ggplot(data, aes(Experience, Salary)) +
  geom_jitter(width = 0.4, shape = 20, color = "darkred") +
  geom_smooth(method = "lm") +
  ggtitle("Salary")
```



Residuals

```
data$stdres <- rstandard(salary.lm1)
ggplot(data, aes(Experience, stdres)) +
  geom_jitter(width = 0.4, shape = 20, color = "darkred") +
  geom_hline(yintercept = 0, color = "orange") +
  ggtitle("Standardized Residuals")
```



Quadratic model.

```
salary.lm2 <- lm(Salary ~ Experience + I(Experience^2), data = data)
summary(salary.lm2)
```

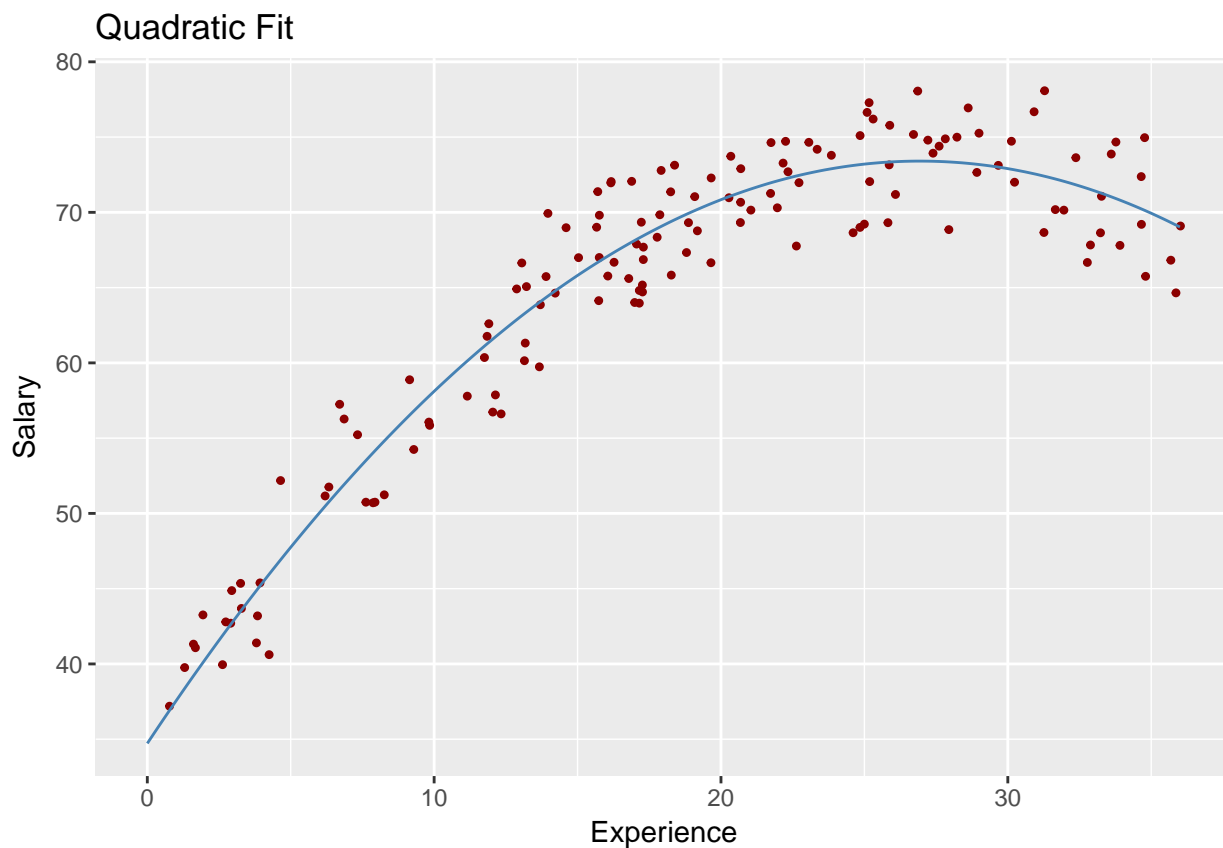
```
##
## Call:
## lm(formula = Salary ~ Experience + I(Experience^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.5786 -2.3573  0.0957  2.0171  5.5176
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.720498  0.828724   41.90 <2e-16
## Experience     2.872275  0.095697   30.01 <2e-16
## I(Experience^2) -0.053316  0.002477  -21.53 <2e-16
##
## Residual standard error: 2.817 on 140 degrees of freedom
```

```
## Multiple R-squared:  0.9247, Adjusted R-squared:  0.9236
## F-statistic: 859.3 on 2 and 140 DF,  p-value: < 2.2e-16

xs <- seq(0, 36, by = 0.2)
quadratic.data <- data.frame(Experience = xs,
                             quadratic = predict(salary.lm2, data.frame(Experience = xs)))
head(quadratic.data)
```

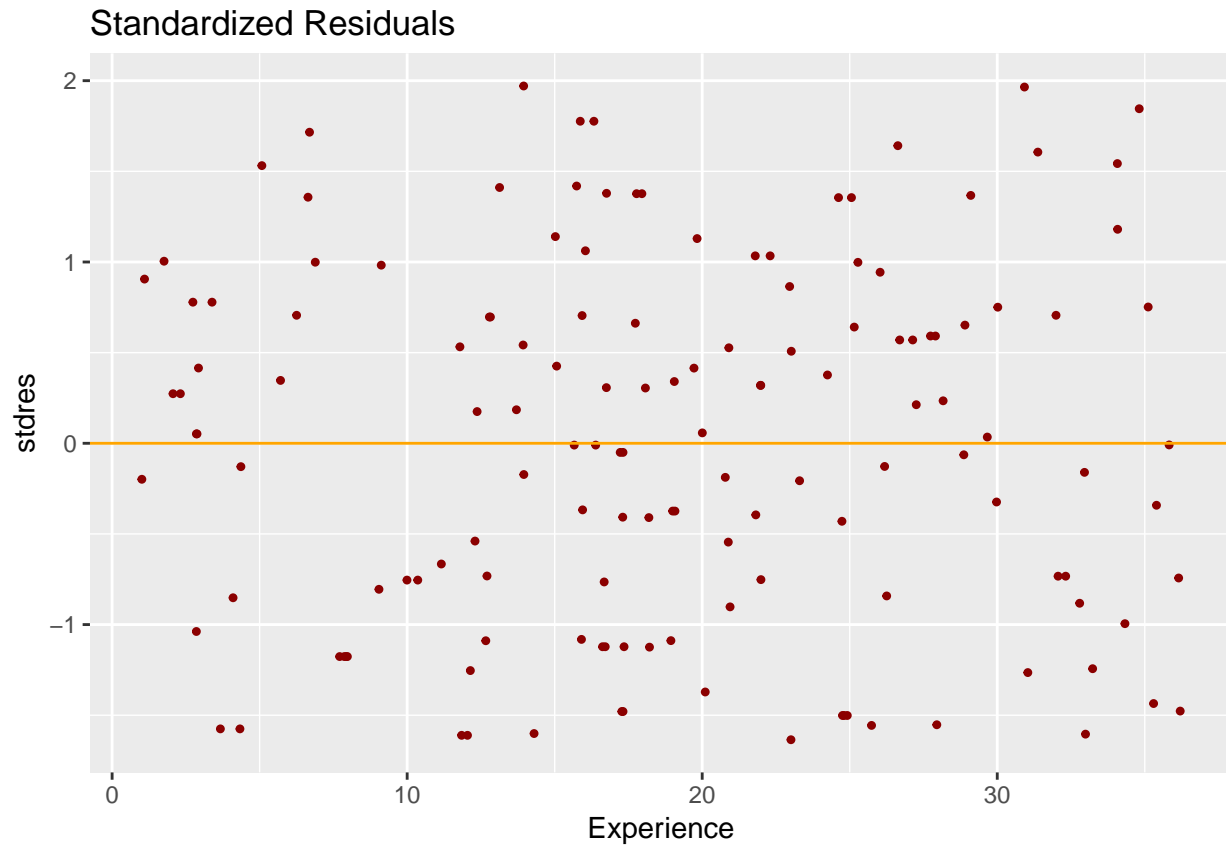
```
##   Experience quadratic
## 1         0.0 34.72050
## 2         0.2 35.29282
## 3         0.4 35.86088
## 4         0.6 36.42467
## 5         0.8 36.98420
## 6         1.0 37.53946
```

```
ggplot(data, aes(Experience, Salary)) +
  geom_jitter(width = 0.4, shape = 20, color = "darkred") +
  geom_line(data = quadratic.data,
            aes(Experience, quadratic), color = "steelblue") +
  ggtitle("Quadratic Fit")
```



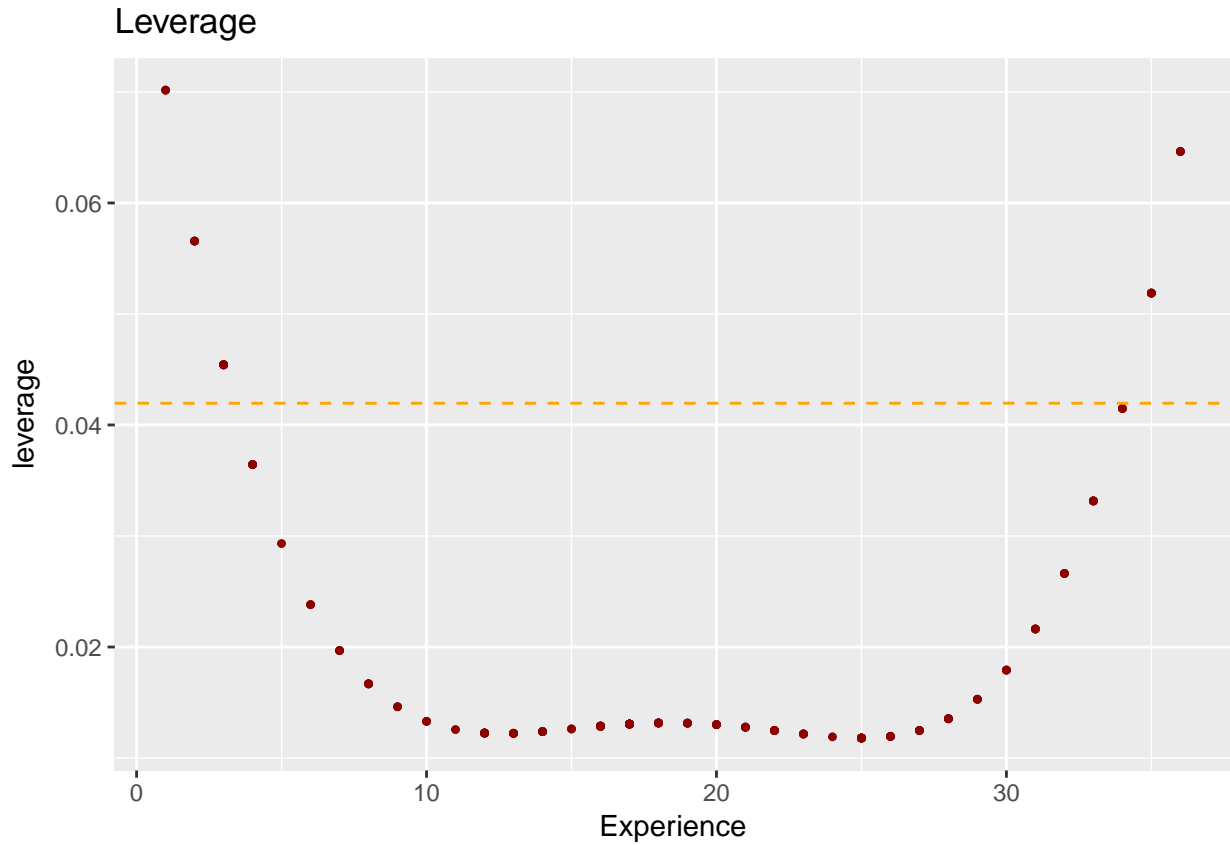
Residuals

```
data$stdres <- rstandard(salary.lm2)
ggplot(data, aes(Experience, stdres)) +
  geom_jitter(width = 0.4, shape = 20, color = "darkred") +
  geom_hline(yintercept = 0, color = "orange") +
  ggtitle("Standardized Residuals")
```



Leverage.

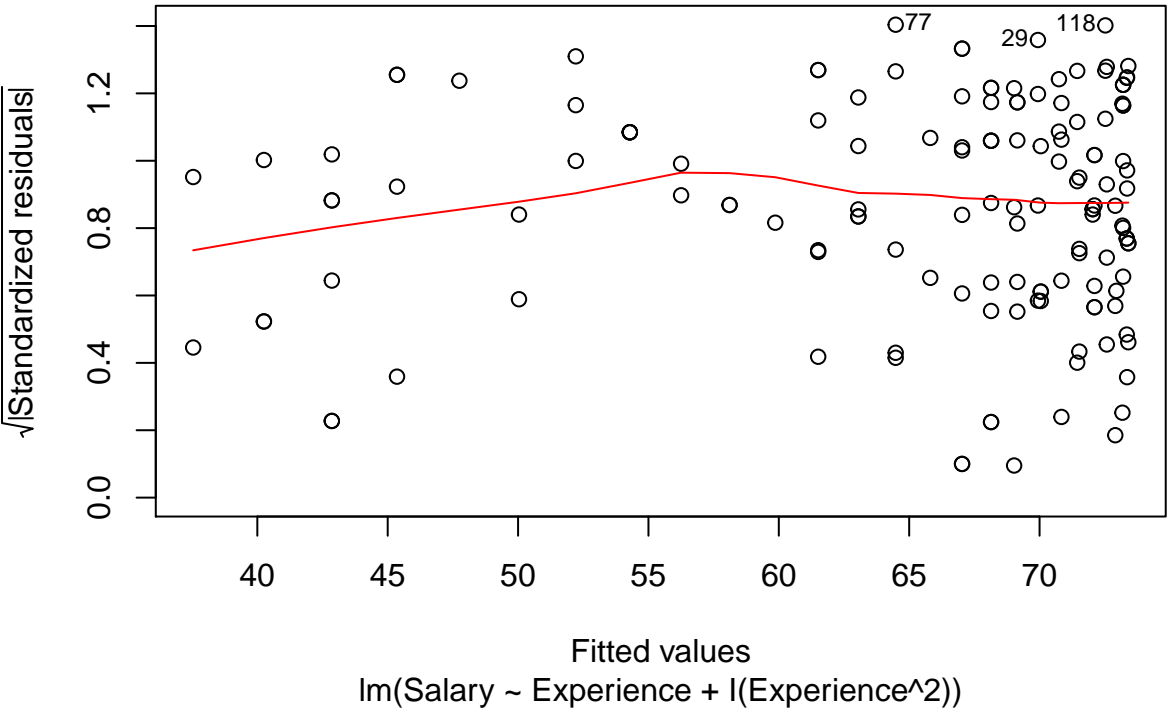
```
data$leverage <- hatvalues(salary.lm2)
p <- 2 # predictors
n <- nrow(data) # cases
cutoff <- 2 * (p + 1) / n # cutoff for points of high leverage
ggplot(data, aes(Experience, leverage)) +
  geom_point(shape = 20, color = "darkred") +
  geom_hline(yintercept = cutoff, color = "orange", lty = "dashed") +
  ggtitle("Leverage")
```

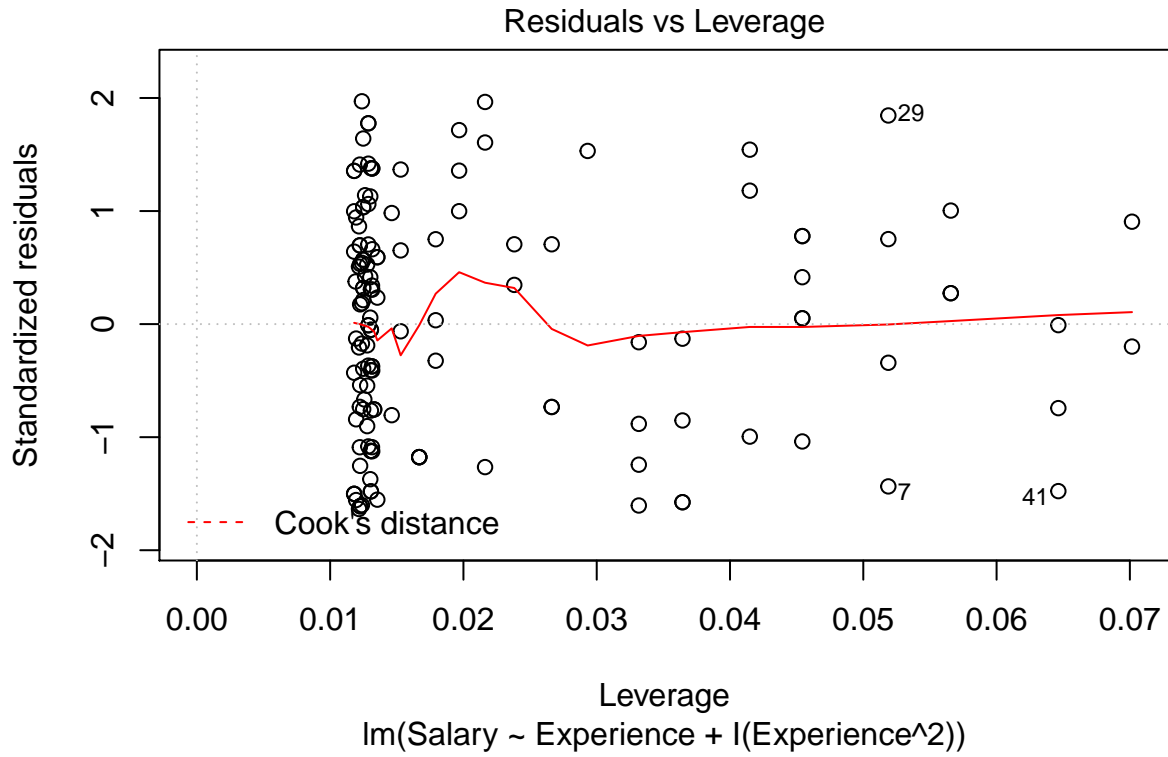


Diagnostics.

```
plot(salary.lm2)
```







Prediction interval.

```
new.data <- data.frame(Experience = 10)
predict(salary.lm2, new.data, interval = "prediction")
```

```
##      fit      lwr      upr
## 1 58.11164 52.50481 63.71847
```