Week 2: Linear Regression

Peiran Chen

4/6/2022

Linear Regression

$$y = X\beta + \varepsilon\hat{\beta} = (X^TX)^{-1}X^Ty$$

Both X, β has p+1 terms. y have n terms.

$$RSS = \sum(y_i - \hat{y}_i)^2 TSS = \sum(y_i - \bar{y})^2$$

We have null $H_0 : \beta_1 = ... = \beta_p = 0$. And alternative$H_1$: at least one $\beta_j \neq 0$.

$$F = \frac{(TSS - RSS)/p}{RSS/n - p - 1}$$

If F-statistic is large, we have more evidence to reject H_0. Notice we need $n - p - 1 > 0$.

```r
library(ISLR2)
fit <- lm(Sales ~ Age + Price + CompPrice + Population + Income:Advertising, data = Carseats)
summary(fit)
```

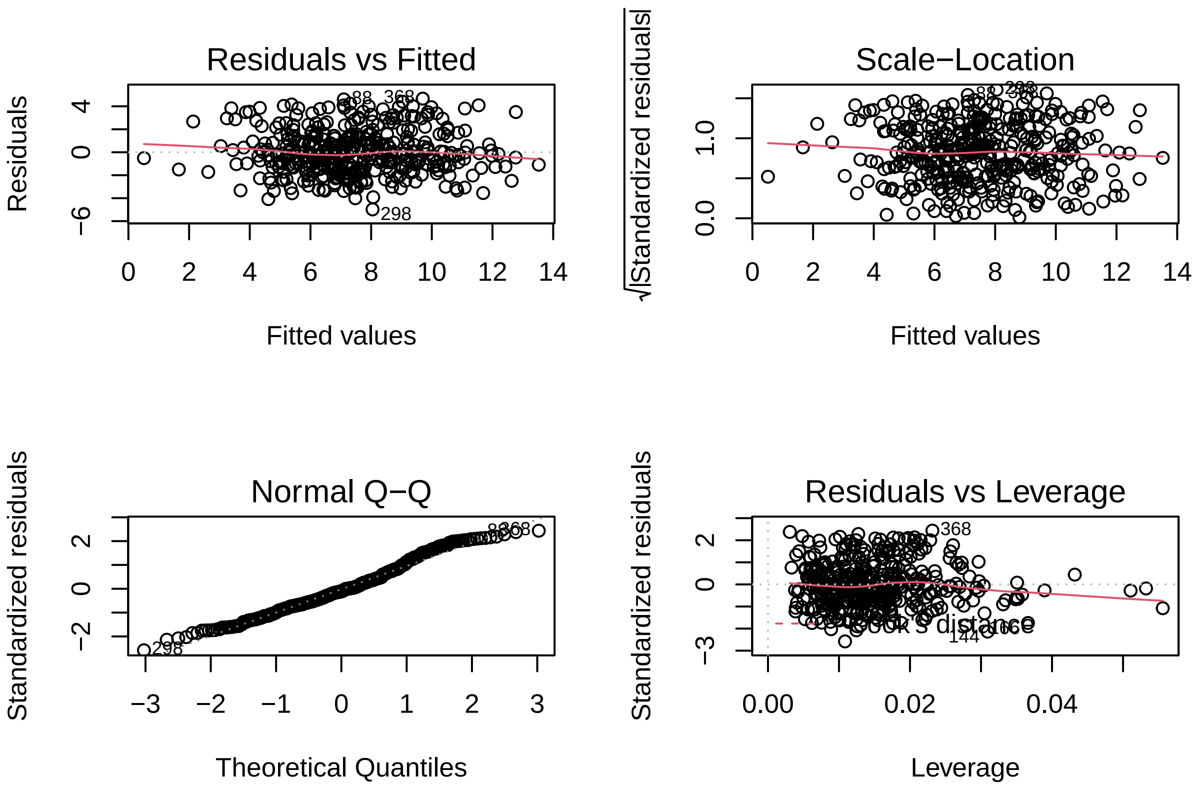
```
##
## Call:
## lm(formula = Sales ~ Age + Price + CompPrice + Population + Income:Advertising,
##     data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9767 -1.3119 -0.2094  1.2252  4.6683
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.916e+00  9.368e-01   8.450 5.68e-16 ***
## Age                -4.355e-02  6.040e-03  -7.211 2.87e-12 ***
## Price              -9.204e-02  5.072e-03 -18.144  < 2e-16 ***
## CompPrice           9.409e-02  7.876e-03  11.948  < 2e-16 ***
## Population         -4.293e-05  6.826e-04  -0.063     0.95
## Income:Advertising  1.740e-03  1.856e-04   9.376  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.939 on 394 degrees of freedom
## Multiple R-squared:  0.5345, Adjusted R-squared:  0.5286
## F-statistic: 90.49 on 5 and 394 DF,  p-value: < 2.2e-16
```

```r
fit_no_pop <- lm(Sales ~ Age + Price + CompPrice + Income:Advertising, data = Carseats)
anova(fit, fit_no_pop)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Age + Price + CompPrice + Population + Income:Advertising
## Model 2: Sales ~ Age + Price + CompPrice + Income:Advertising
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    394 1481.2
## 2    395 1481.2 -1  -0.01487 0.004 0.9499
```

Diagnostic plots:

```r
layout(matrix(c(1,2,3,4),2,2))
plot(fit)
```



Variable Selection:

If p is large, looking at individual $p-$values is misleading

- p-value < 0.05: When $\beta_j = 0$, we have less than 5% probability for it to be significant.

F-test does not suffer from this problem, but require $n > p$.

Potential Problems with LR:

- Assuming linear relationship
- Independence of errors
- Constant variance of errors, or heteroscedasticity