STA 108 Discussion 9: Outlier detection

Yuanyuan Li

Reference: textbook sec. 10.2, 10.3, 10.4; lecture summary 26.

1 Outlying cases

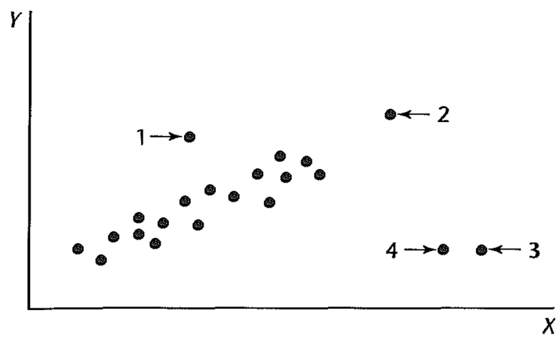


FIGURE 10.5 Scatter Plot for Regression with One Predictor Variable Illustrating Outlying Cases.

- Outlying Y observations: case 1, 3, 4.
- Outlying X observations: case 2, 3, 4.
- **Influential cases**: case 3, 4. They are outlying with regard to their X values, and their Y values are not consistent with the regression relation for the other cases.

2 Diagnostics of outliers

Simple linear regression: box plots, stem-and-Ieaf plots, scatter plots, and residual plots.

Multiple linear regression: Recall the least square estimates is $\hat{\beta} = (X'X)^{-1}X'Y$, and fitted value is $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$. Define the hat matrix as $H = X(X'X)^{-1}X'$, then $\hat{Y} = HY$.

The residulas can be expressed as $\hat{\epsilon} = (I - H)Y$, hence the variance of residual $\hat{\epsilon}_i$ is $\sigma^2(1 - h_{ii})$, where h_{ii} is the ith element on the main diagonal of the hat matrix. Using MSE as the estimator of the error variance σ^2, we get the standard error of $\hat{\epsilon}_i$ as $s.e.(\hat{\epsilon}_i) = \sqrt{MSE(1 - h_{ii})}$.

- Identifying outlying Y observations:
  - studendized residual: $r_i = \frac{\hat{\epsilon}_i}{s.e.(\hat{\epsilon}_i)} = \frac{\hat{\epsilon}_i}{\sqrt{MSE(1-h_{ii})}}$.
  - jackknife (deleted) residual: Fit the regression with the ith case deleted; let $Y_{i(-i)}$ denote the predicted value for Y_i, under this regression. $d_i = Y_i - Y_{i(-i)}$ is called the deleted residual.
  - studentized deleted residual: $t_i = \frac{d_i}{s.e.(d_i)}$, where $s.e.(d_i) = \frac{MSE_{(i)}}{1-h_{ii}}$. Under the null hypothesis that there are no outliers, $t_i \sim t_{n-p-1}$.

- Test for outliers by Bonferroni method: Reject H_0 (no outlying Y observations), if $\max_{1 \le i \le n} |t_i| > t_{n-p-1}(1 - \frac{\alpha}{2n})$.

- Identifying outlying X observations:
  - leverages: h_{ii}, which is a measure of the distance between the X values for the i th case and the means of the X values for all n cases. If $h_{ii} > 2p/n$, case i is considered outlying in X.

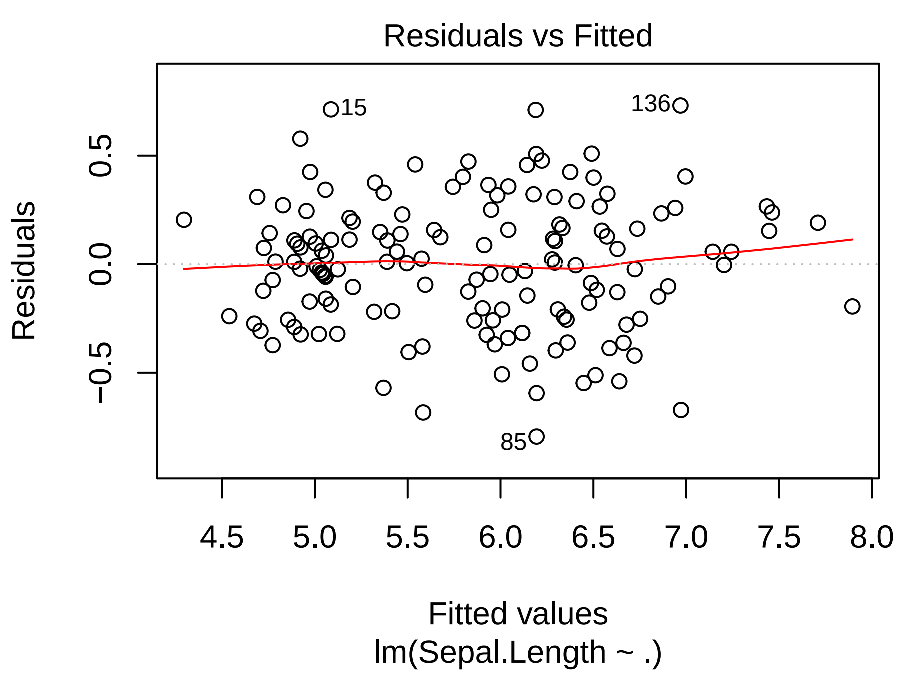
- Identifying influential cases:
  - Cook's distance: $D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{pMSE} = \frac{\epsilon_i^2}{pMSE} \frac{h_{ii}}{(1-h_{ii})^2}$. Case i is considered as influential case if $D_i > 4/n$.

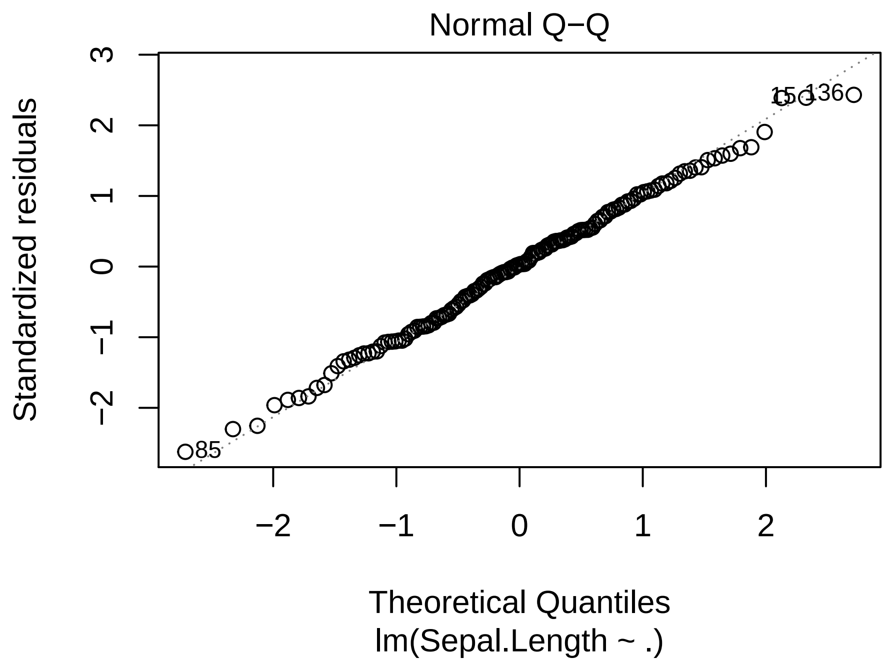
3 Example: Can you find outliers and influential cases for model below?

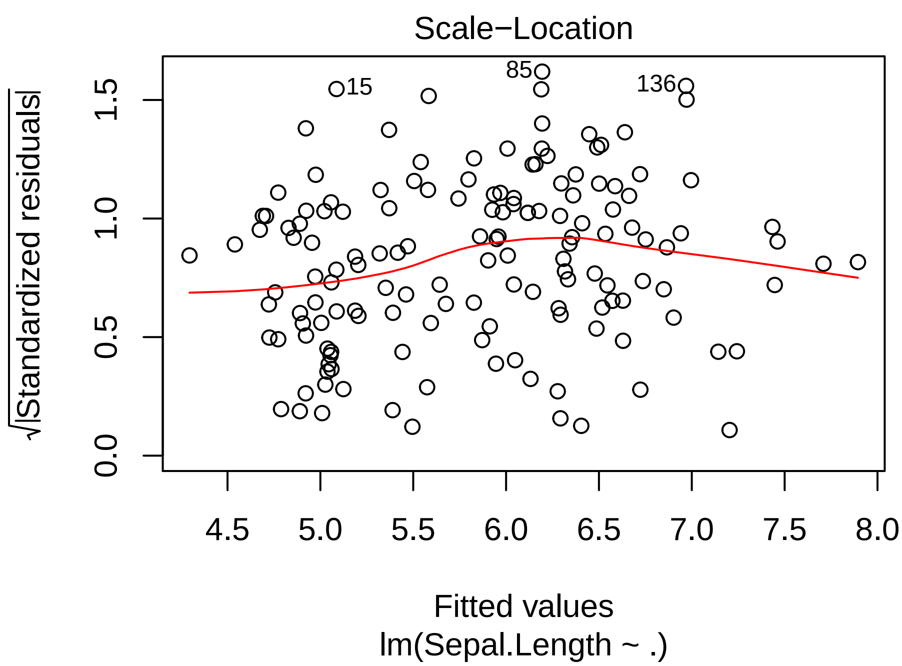
```
fit=lm(Sepal.Length~ ., data=iris)
fit
```

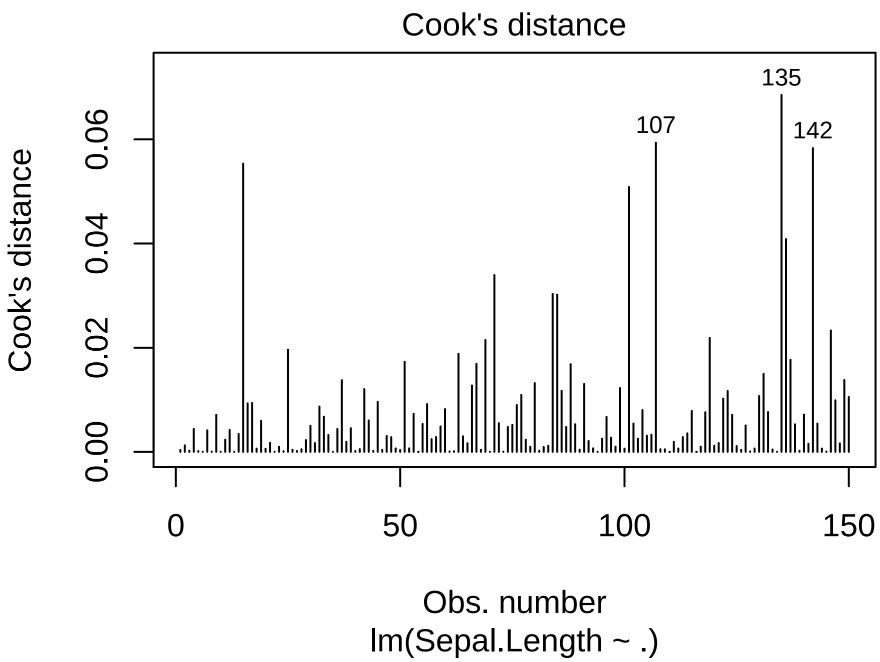
```
##
## Call:
## lm(formula = Sepal.Length ~ ., data = iris)
##
## Coefficients:
##       (Intercept)        Sepal.Width       Petal.Length        Petal.Width
##            2.1713             0.4959             0.8292            -0.3152
## Speciesversicolor   Speciesvirginica
##           -0.7236            -1.0235
```

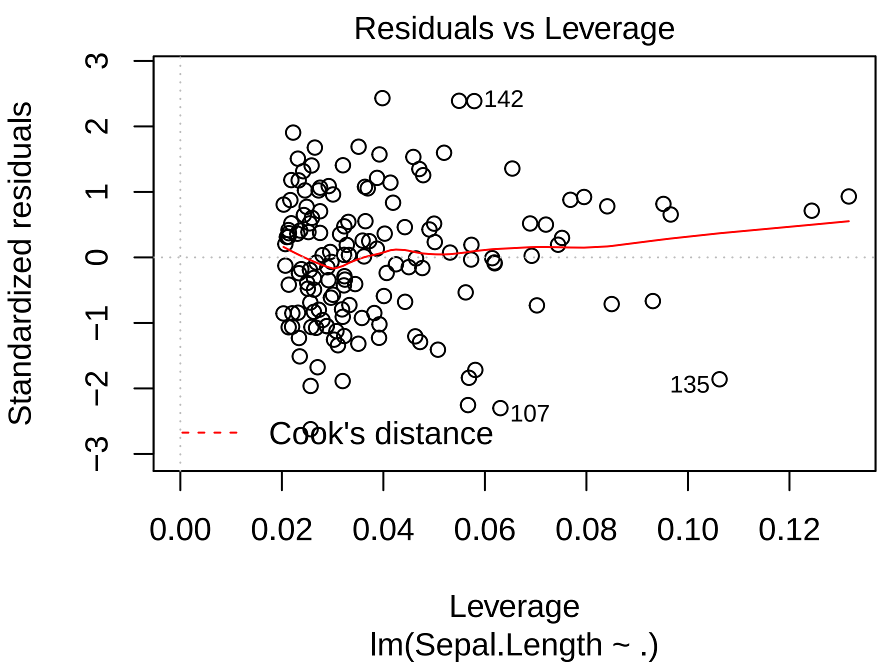
```
plot(fit,which=1:6)
```

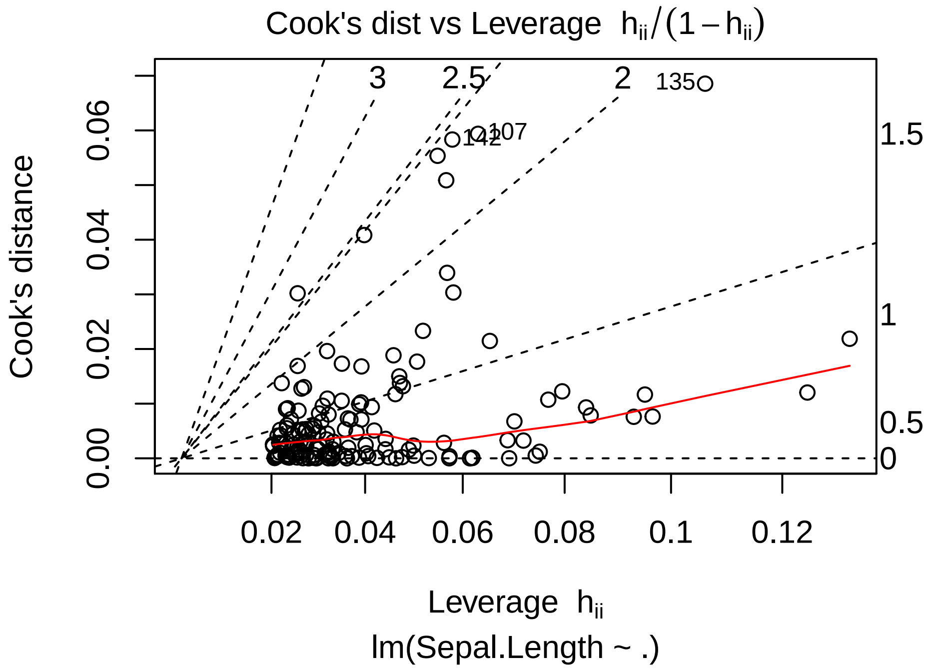



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Sepal.Length ~ .)
85
15
136


Scale−Location
√|Standardized residuals|
Fitted values
lm(Sepal.Length ~ .)
15
85
136


Cook's distance
Cook's distance
0.00
0.02
0.04
0.06
0
50
100
150
107
135
142
Obs. number
lm(Sepal.Length ~ .)


Residuals vs Leverage
Standardized residuals
−3
−2
−1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10
0.12
142
107
135
Cook's distance
Leverage
lm(Sepal.Length ~ .)


Cook's dist vs Leverage $h_{ii}/(1-h_{ii})$
Cook's distance
0.00
0.02
0.04
0.06
Leverage h_{ii}
0.02
0.04
0.06
0.08
0.1
0.12
3
2.5
2
135
142
107
1.5
1
0.5
0
lm(Sepal.Length ~ .)