Stat 849: Linear regression diagnostics

Sündüz Keleş

Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Regression Diagnostics

- Checking the Gauss-Markov assumptions on the error vector using graphical and numerical methods.
- Checking the normality assumption.
- Weighted least squares: Heteroscedascity.
- Outlier and influential point detection.

Outliers and Influential observations

- Regression outliers.
- High leverage points.
- Influential points.

What are these and how can we measure them?

Outliers and Influential observations

- Regression outliers.
- High leverage points.
- Influential points.

What are these and how can we measure them?

One should be cautious about unusual data points in linear models since they can influence the results of the analysis, and their presence might may be a signal that the model fails to capture important characteristics of the data.

Outliers

- Outliers are extreme observations.
- It is common practice to distinguish between two types of outliers.
- Outliers in the response variable are called residual outliers.
- Outliers with respect to the predictors are called leverage points. They can affect the regression model, too. Their response variables need not be outliers. However, they may almost uniquely determine regression coefficients. They may also cause the standard errors of regression coefficients to be much smaller than they would be if the observation were excluded.

Outlier detection

- Residual outliers can be identified from (1) residual plots against X or $\hat{Y}$; (2) box plots, stem-and-leaf plots, and dot plots of the residuals.
- Plotting standardized residuals might be helpful.
- Standardized residuals: $\hat{\epsilon}_i / \sqrt{\text{var}(\hat{\epsilon}_i)}$.

$$\text{var}(\hat{\epsilon}) = ?$$

Standardized (internally studentized) residuals

$$\begin{aligned} \text{var}(\hat{\epsilon}) &= \text{var}(Y - \hat{Y}) \\ &= \text{var}(Y - PY) = \sigma^2(I - P), \end{aligned}$$

where $P = X(X^TX)^{-1}X^T$ is the projection matrix (*hat matrix H*). Then, $\text{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$ where h_{ii} is the i-th element on the main diagonal of the hat matrix, and the covariance between residuals $\hat{\epsilon}_i$ and $\hat{\epsilon}_j$ ($i \neq j$) is $\sigma^2(0 - h_{ij}) = -h_{ij}\sigma^2$, where h_{ij} is the element in the i-th row and j-th column of the hat matrix.

Note: This can only correct for the natural non-constant variance in residuals when errors ϵ_i have constant variance.

Leverage

- h_{ii} is called the leverage (in terms of the X values) of the i-th case.
- Properties of h_{ii}: $0 \leq h_{ii} \leq 1$, $\sum_{i=1}^{n} h_{ii} = p$.
- The h_{ii} values theoretically range from $1/n$ to 1. Those that exceed $2p/n$ are said to be large.
- Interpretation: It is a measure of distance between the X values for the i-th case and the means of the X values for all of the observations. It measures the degree of conformity of a single observation to the linear pattern established by the other $n-1$ observations.
- For a linear regression with model with one predictor, the leverage associated with the specific observation (x_i, y_i) is

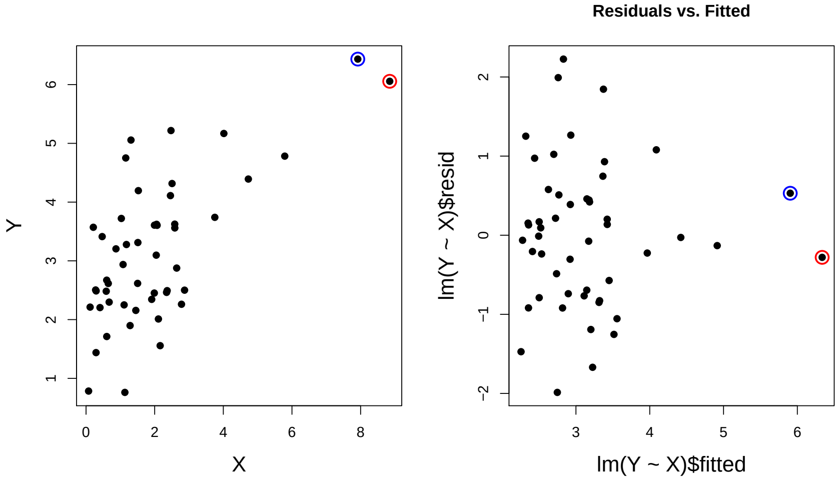
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

```
#Computing standardized  residuals
xx = cbind(1, x)
H = xx %*% solve((t(xx)%*%xx)) %*% t(xx) #Hat Matrix
dhat = (1-diag(H))


#Could also use lm.influence(lm1)$hat
lm1.stdres0 = lm1$resid/sqrt((sum(lm1$resid^2)/(n-2 ))*dhat)


#Two R functions that will give us standardized residuals
lm1.stdres2 = rstandard(lm1)
library(MASS)
lm1.stdres1 = stdres(lm1)
```

What to do with outliers?

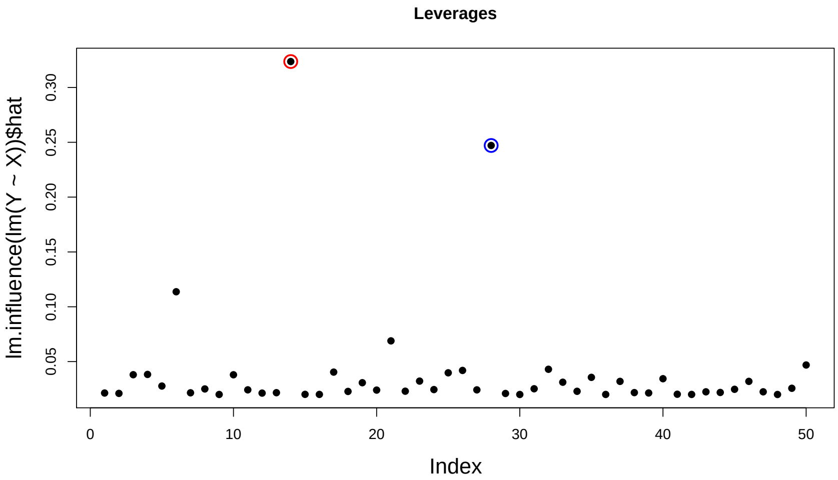
Should only discard these if there is a direct evidence that it represents an error in recoding, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance. Sometimes detection of an outlier itself might be of interest.

Example 1



Residuals vs. Fitted
Y
X
0
2
4
6
8
1
2
3
4
5
6
lm(Y ~ X)$resid
lm(Y ~ X)$fitted
3
4
5
6
−2
−1
0
1
2

Example 1



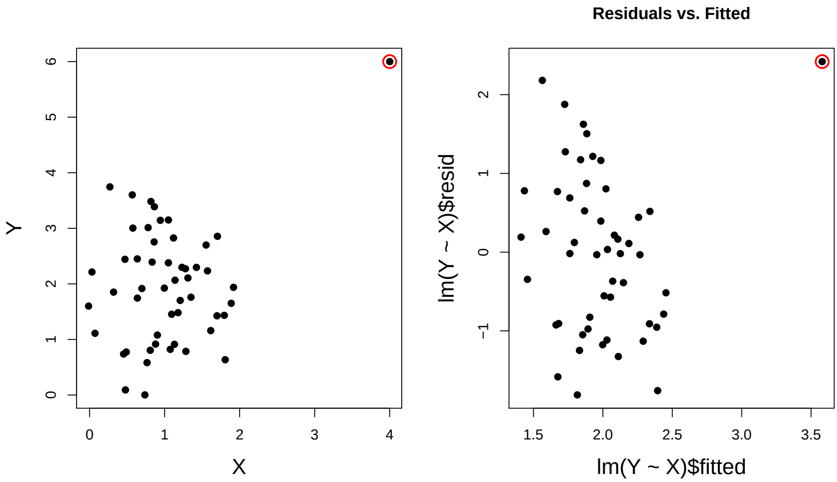
Leverages
lm.influence(lm(Y ~ X))$hat
0.05
0.10
0.15
0.20
0.25
0.30
0
10
20
30
40
50
Index

Fits with and without data points 14 and 28 (Example 1)

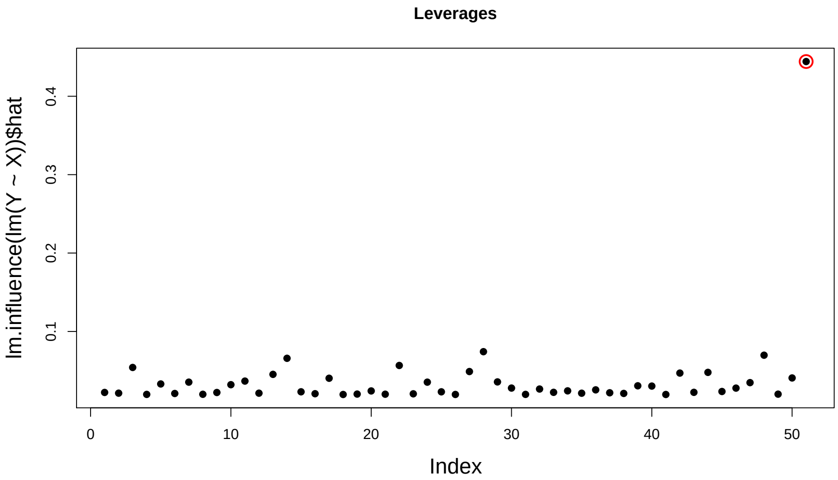
```
summary(lm(Y ~ X))$coef
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 2.2213052   0.196018 11.33215 3.610852e-15
X           0.4650999   0.074137  6.27352 9.567487e-08
```

```
summary(lm(Y[-which(X >= 7)] ~ X[-which(X >= 7)]))$coef
                    Estimate Std. Error  t value     Pr(>|t|)
(Intercept)        2.2477769  0.2349439 9.567291 1.630211e-12
X[-which(X >= 7)] 0.4464636  0.1127122 3.961095 2.573293e-04
```

Example 2


Y
0
1
2
3
4
5
6
0
1
2
3
4
X
Residuals vs. Fitted
lm(Y ~ X)$resid
−1
0
1
2
1.5
2.0
2.5
3.0
3.5
lm(Y ~ X)$fitted

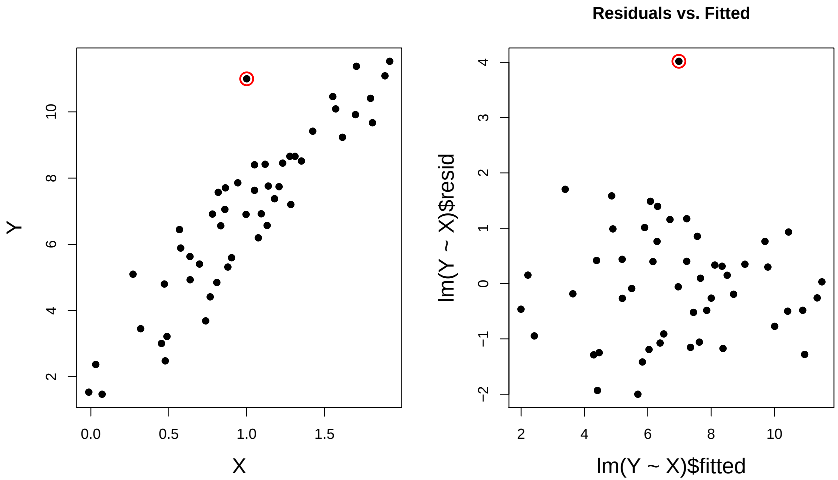
Example 2


Leverages
lm.influence(lm(Y ~ X))$hat
0.1
0.2
0.3
0.4
0
10
20
30
40
50
Index

Fits with and without data point 51 (Example 2)

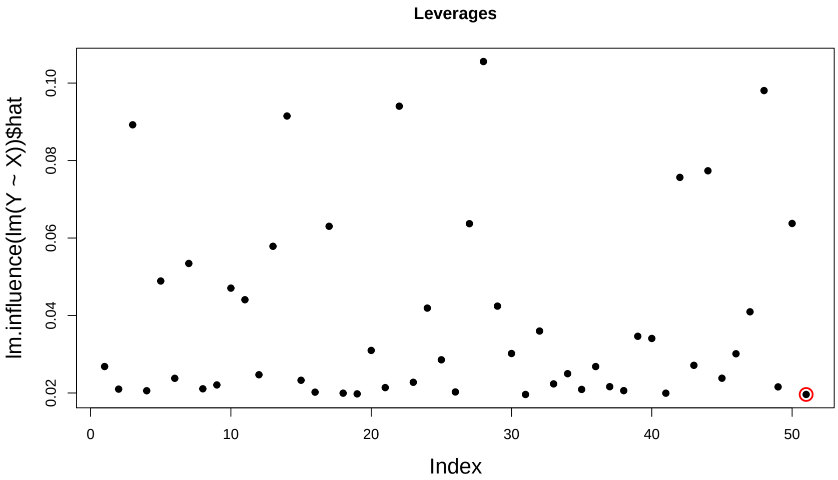
```
summary(lm(Y~X))$coef
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 1.417814  0.2813514 5.039298 6.759613e-06
X           0.540498  0.2299715 2.350282 2.282995e-02
summary(lm(Y[-51]~X[-51]))$coef
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept)  1.9872932  0.3036230  6.5452648 3.665170e-08
X[-51]      -0.0856566  0.2763996 -0.3099014 7.579775e-01
```

Example 3



Y
X
2
4
6
8
10
0.0
0.5
1.0
1.5
Residuals vs. Fitted
lm(Y ~ X)$resid
lm(Y ~ X)$fitted
−2
−1
0
1
2
3
4
2
4
6
8
10

Example 3


Leverages
lm.influence(lm(Y ~ X))$hat
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
50
Index

Fits with and without data point 51 (Example 3)

```
summary(lm(Y~X))$coef
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 2.063155  0.3531323  5.842443 4.089197e-07
X           4.918900  0.3220109 15.275571 2.884107e-20
summary(lm(Y[-51]~X[-51]))$coef
            Estimate Std. Error   t value    Pr(>|t|)
(Intercept) 1.987293  0.3036230  6.545265 3.66517e-08
X[-51]      4.914343  0.2763996 17.779852 9.29556e-23
```

R commands (Example 2)

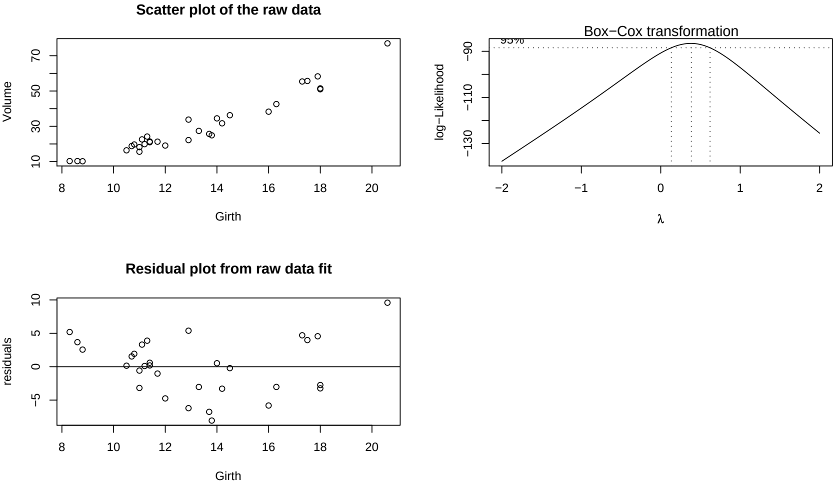
```
> cbind(lm1$resid, rstandard(lm1), rstudent(lm1),
cooks.distance(lm1))
          [,1]        [,2]        [,3]         [,4]
1   0.16501135  0.16007840  0.15847797 2.910475e-04
2   0.87295674  0.84646206  0.84397335 7.793663e-03
3  -0.78706606 -0.77629531 -0.77310187 1.722916e-02
4  -0.55512537 -0.53784507 -0.53390689 2.909499e-03
. . .
49  0.03401324  0.03295962  0.03262192 1.109980e-05
50  0.51771742  0.50705673  0.50317786 5.453846e-03
51  2.42019451  3.11442047  3.44190961 3.877808e+00

> 4/(51-2) #Adhoc threshold for Cook's distance
[1] 0.08163265

#Also check out outlier.test() in the car library.
```

Box-Cox Transformation

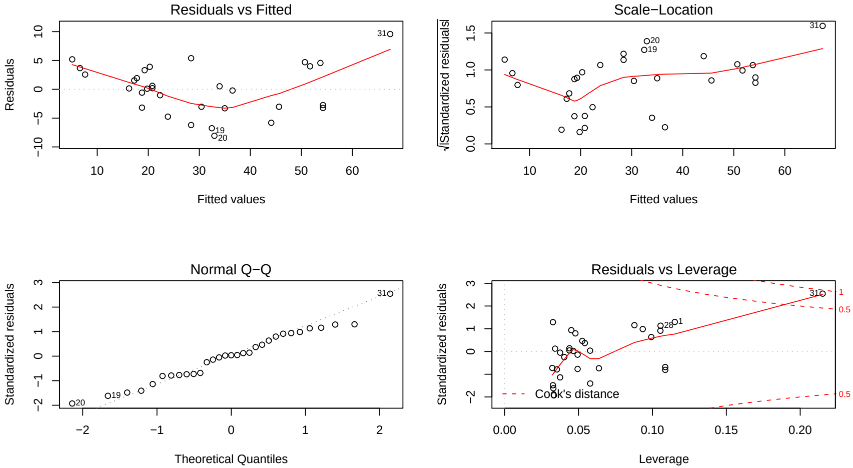
```
library(MASS)
data(trees)
attach(trees)
```

```
#This data set provides measurements of the girth, height and
#volume of timber in 31 felled black cherry trees.  Note that
#girth is the diameter of the tree (in inches) measured at 4
#ft 6 in above the ground.


lmtree = lm (Volume ~ Girth, data = trees)

par(mfcol = c(2, 2))
plot(Girth, Volume, main = "Scatter plot of the raw data")
plot(Girth, lmtree$resid, ylab = "residuals", main = "Residual
plot from raw data fit")
abline(h = 0)
boxcox(lmtree)
mtext("Box-Cox transformation")
```

boxcox()


Scatter plot of the raw data
Volume
10
30
50
70
8
10
12
14
16
18
20
Girth
Box−Cox transformation
95%
log−Likelihood
−90
−110
−130
−2
−1
0
1
2
λ
Residual plot from raw data fit
residuals
10
5
0
−5
8
10
12
14
16
18
20
Girth

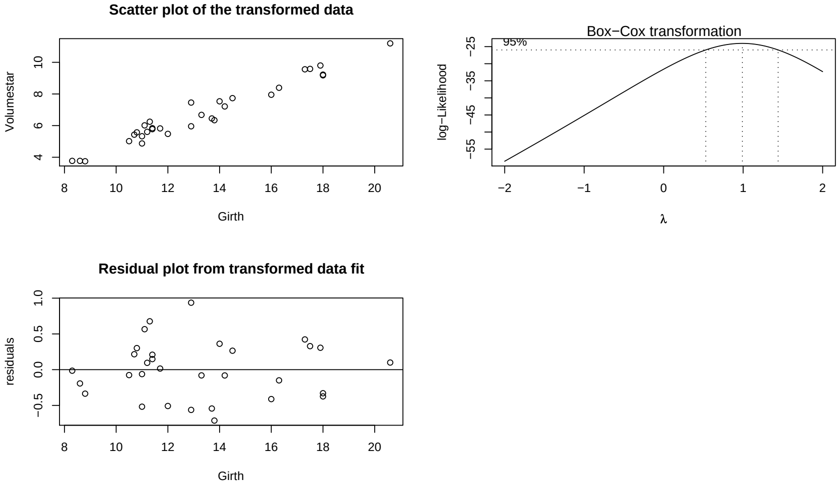
Diagnostic plots before `boxcox()`



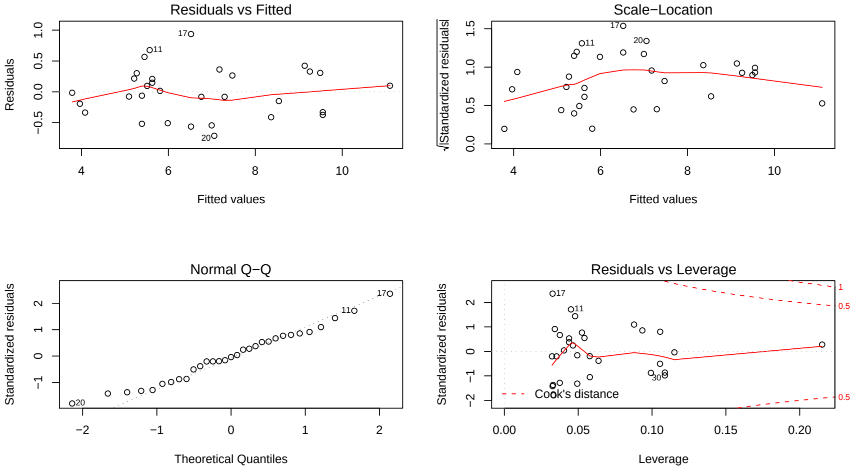
boxcox()

```
names(boxcox(lmtree)) #Find the index of lambda with the highes
                      #log-likelihood value
rev(order(boxcox(lmtree)$y))[1]
[1] 60
> boxcox(lmtree)$x[60]
[1] 0.3838384
lambda = boxcox(lmtree)$x[60]

Volumestar = (Volume^(lambda)-1)/lambda
newlmtree = lm (Volumestar ~ Girth, data = trees)


par(mfcol = c(2, 2))
plot(Girth, Volumestar, main = "Scatter plot of the
transformed data")
plot(Girth, newlmtree$resid, ylab = "residuals",
main = "Residual plot from transformed data fit")
abline(h = 0)
boxcox(newlmtree)
mtext("Box-Cox transformation")
```

boxcox()


Scatter plot of the transformed data
Volumestar
4
6
8
10
8
10
12
14
16
18
20
Girth
Box−Cox transformation
95%
log−Likelihood
−25
−35
−45
−55
−2
−1
0
1
2
λ
Residual plot from transformed data fit
residuals
−0.5
0.0
0.5
1.0
8
10
12
14
16
18
20
Girth

Diagnostic plots after boxcox()



Residuals vs Fitted
Residuals
Fitted values
−0.5
0.0
0.5
1.0
4
6
8
10
17
11
20
Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
4
6
8
10
17
11
20
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
0
1
2
−2
−1
0
1
2
20
11
17
Residuals vs Leverage
Standardized residuals
Leverage
−2
−1
0
1
2
0.00
0.05
0.10
0.15
0.20
17
11
30
Cook's distance
1
0.5
0.5