Regression Diagnostics II

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Linear Regression Assumptions

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I)$$

Assumptions

- **L**inear relationship
- **I**ndependent observations
- **N**ormally distributed residuals
- **E**qual variance across X's
- Plus need to check for influential points and outliers: one or a few observations should not dominate the model fit

Outliers

- Outliers are bizarre data points. Observations may be outlying relative only to other predictors $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ or relative to *the model*, i.e. Y_i relative to $\hat{Y}_i$.
- *Studentized deleted residuals* are designed to detect outlying Y_i observations; *leverages* detect outlying $\mathbf{x}_i$ points.
- Outliers have the potential to influence the fitted regression function; they may *strengthen* inference and reduce error in predictions if the outlying points follow the modeling assumptions and are representative.
- If not, outlying values may skew inference unduly and yield models with poor predictive properties.

Leverage

Leverage is the potential for a single observation to influence a regression statistics such a coefficient or predicted values. Leverage is determined by an observation's position in the X space.

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

$$\hat{Y}_2 = h_{21} \times Y_1 + h_{22} Y_2 + h_{23} \times Y_3 + ...$$

1	0
1	1
1	2
1	3
1	4
1	10

Table: Design Matrix

	1	2	3	4	5	6
1	0.34	0.29	0.24	0.18	0.13	-0.18
2	0.29	0.25	0.22	0.18	0.14	-0.08
3	0.24	0.22	0.19	0.17	0.15	0.03
4	0.18	0.18	0.17	0.17	0.16	0.13
5	0.13	0.14	0.15	0.16	0.17	0.24
6	-0.18	-0.08	0.03	0.13	0.24	0.87

Table: Hat matrix

Leverage

Leverage is the potential for a single observation to influence a regression statistics such a coefficient or predicted values. Leverage is determined by an observation's position in the X space.

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

$$\hat{Y}_2 = h_{21} \times Y_1 + h_{22} Y_2 + h_{23} \times Y_3 + ...$$

1	0
1	1
1	2
1	3
1	4
1	10

Table: Design Matrix

	1	2	3	4	5	6
1	0.34	0.29	0.24	0.18	0.13	-0.18
2	0.29	0.25	0.22	0.18	0.14	-0.08
3	0.24	0.22	0.19	0.17	0.15	0.03
4	0.18	0.18	0.17	0.17	0.16	0.13
5	0.13	0.14	0.15	0.16	0.17	0.24
6	-0.18	-0.08	0.03	0.13	0.24	0.87

Table: Hat matrix

We measure "leverage" by the potential for an observation to influence it's own predicted value.

Hat Matrix

Because $H = X(X'X)^{-1}X'$ is symmetric and idempotent,

$$h_{ii} = \sum_i h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \geq 0$$

$$h_{ii}(1 - h_{ii}) \geq 0$$

Hence, $0 \leq h_{ii} \leq 1$

In addition, $\sum_i h_{ii} = p$

$$\sum_i h_{ii} = \text{trace}(H) = \text{trace}(X(X'X)^{-1}X') \qquad [\text{trace}(AB) = \text{trace}(BA)]$$

$$= \text{trace}[(X'X)(X'X)^{-1}] = \text{trace}(I_p) = p$$

The rule of thumb is that any leverage h_{ii} that is larger than twice the mean leverage p/n, i.e. $h_{ii} > 2p/n$, is flagged as having "high" leverage.

Outliers & influential points

- Often outliers are "flagged" and deemed suspect as mistakes or observations not gathered from the same population as the other observations.
- Sometimes outliers are of interest in their own right and may illustrate aspects of a data set that bear closer scrutiny.
- Although an observation may be flagged as an outlier, the point *may or may not* affect the fitted regression function more than other points.
- A *DFFIT* is a measure of influence that an individual point $(\mathbf{x}_i, Y_i)$ has on the regression surface at $\mathbf{x}_i$.
- *Cook's distance* is a consolidated measure of influence the point $(\mathbf{x}_i, Y_i)$ has on the regression surface at all n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

10.2 Studentized deleted residuals

- The *standardized residuals*

$$r_i = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE(1 - h_{ii})}}$$

  have a constant variance of 1.
- Typically, $|r_i| > 2$ is considered "large." $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ is the i^{th} leverage value.
- A refinement of the standardized residual that has a recognizable distribution is the *studentized deleted residual*

$$t_i = r_i\sqrt{\frac{MSE}{MSE_{(i)}}} = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$$

  where $MSE_{(i)}$ is the mean squared error calculated from a multiple regression with the same predictors but the i^{th} observation removed.
- The studentized deleted residual t_i will be larger than a regular studentized residual r_i if and only if $MSE_{(i)} < MSE$.

10.4 DFFITS

- The i^{th} *DFFIT*, denoted $DFFIT_i$, is given by

$$DFFIT_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}},$$

  where $\hat{Y}_i$ is fitted value of regression surface (calculated using all n observations) at $\mathbf{x}_i$ and $\hat{Y}_{j(i)}$ is fitted value of regression surface *omitting the point* $(\mathbf{x}_i, Y_i)$ at the point $\mathbf{x}_j$.
- $DFFIT_i$ is standardized distance between *fitted* regression surfaces *with* and *without* the point $(\mathbf{x}_i, Y_i)$.
- Rule of thumb that $DFFIT_i$ is "large" when $|DFFIT_i| > 1$ for small to medium-sized data sets and $|DFFIT_i| > 2\sqrt{p/n}$ for large data sets. We will often just note those $DFFIT_i$'s that are considerably larger than the bulk of the $DFFIT_i$'s.

10.4 Cook's distance

- The i^{th} Cook's distance, denoted D_i, is an aggregate measure of the influence of the i^{th} observation on all n fitted values:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p(MSE)}.$$

- This is the sum of squared distances, at each $\mathbf{x}_j$, between fitted regression surface calculated with all n points and fitted regression surface calculated with the i^{th} case removed, standardized by $p(MSE)$.
- Look for values of Cook's distance significantly larger than other values; these are cases that exert disproportionate influence on the fitted regression surface as a whole.

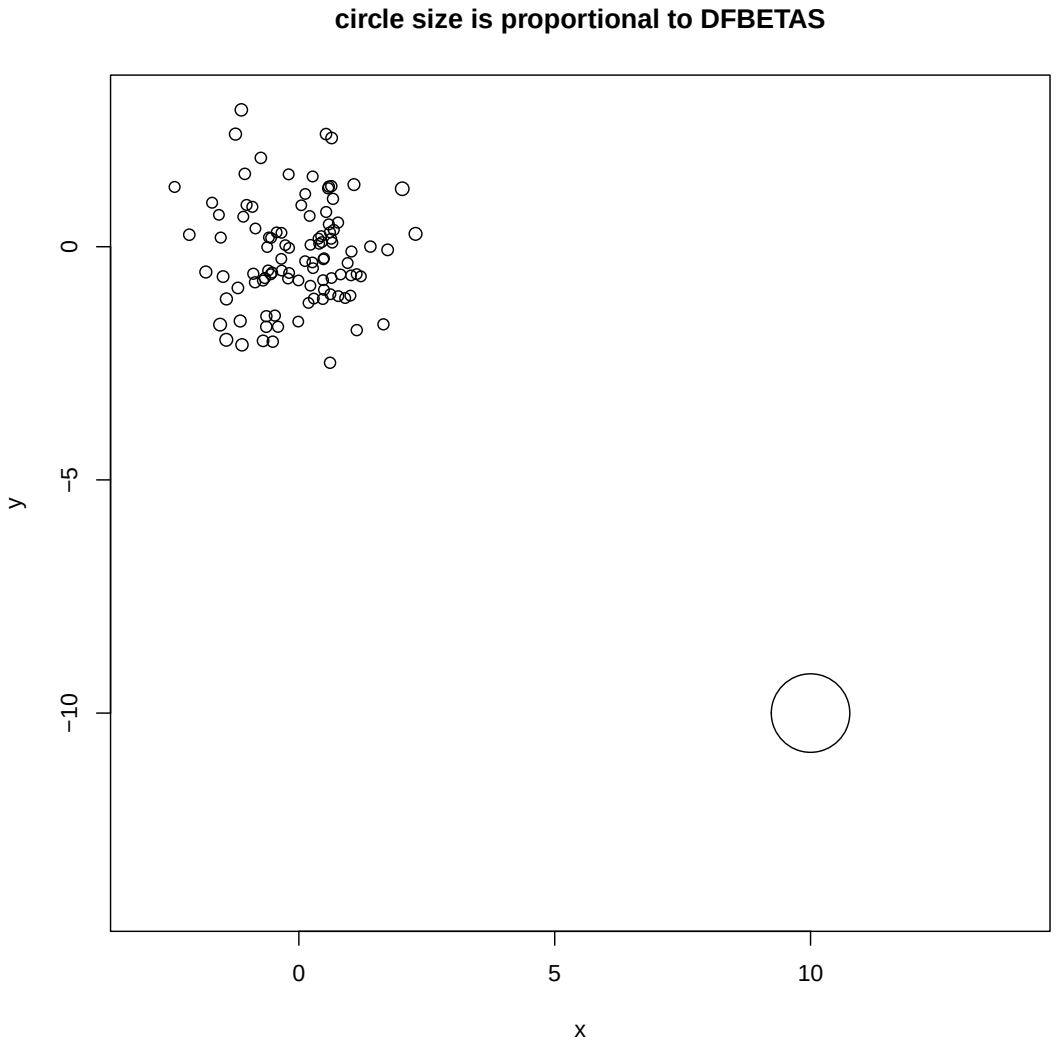
DFBETAS

Measure of how much an observation has effected the estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). Values larger than 2/sqrt(n) in absolute value are considered highly influential.

- $\text{DFBETA}_i = \hat{\beta} - \widehat{\beta_{-i}}$
- $\text{DFBETAS}_i = \frac{\hat{\beta} - \widehat{\beta_{-i}}}{se(\widehat{\beta_{-i}})}$

```
>x<-rnorm(100)
>y<-rnorm(100)
>fit<-lm(y ~ x)
>infm<-influence.measures(fit)
>infm[[1]][1:5,]
       dfb.1_        dfb.x       dffit    cov.r       cook.d        hat
1 -0.15927196  0.195599293 -0.24623646 1.003610 0.0299565058 0.02710032
2 -0.03987014 -0.003134912 -0.04020223 1.027726 0.0008151043 0.01006118
3  0.03316866  0.046934547  0.05890024 1.046815 0.0017502792 0.02739478
4  0.03891405  0.068888356  0.08091733 1.055429 0.0033016469 0.03633509
5  0.03152137  0.003616565  0.03195008 1.029053 0.0005151360 0.01012979
```

Influence Plot Using DFBETAS



Chapter 7 example: Body fat

$n = 20$ healthy females 25–34 years old.

- x_1 = triceps skinfold thickness (mm)
- x_2 = thigh circumference (cm)
- x_3 = midarm circumference (cm)
- Y = body fat (%)

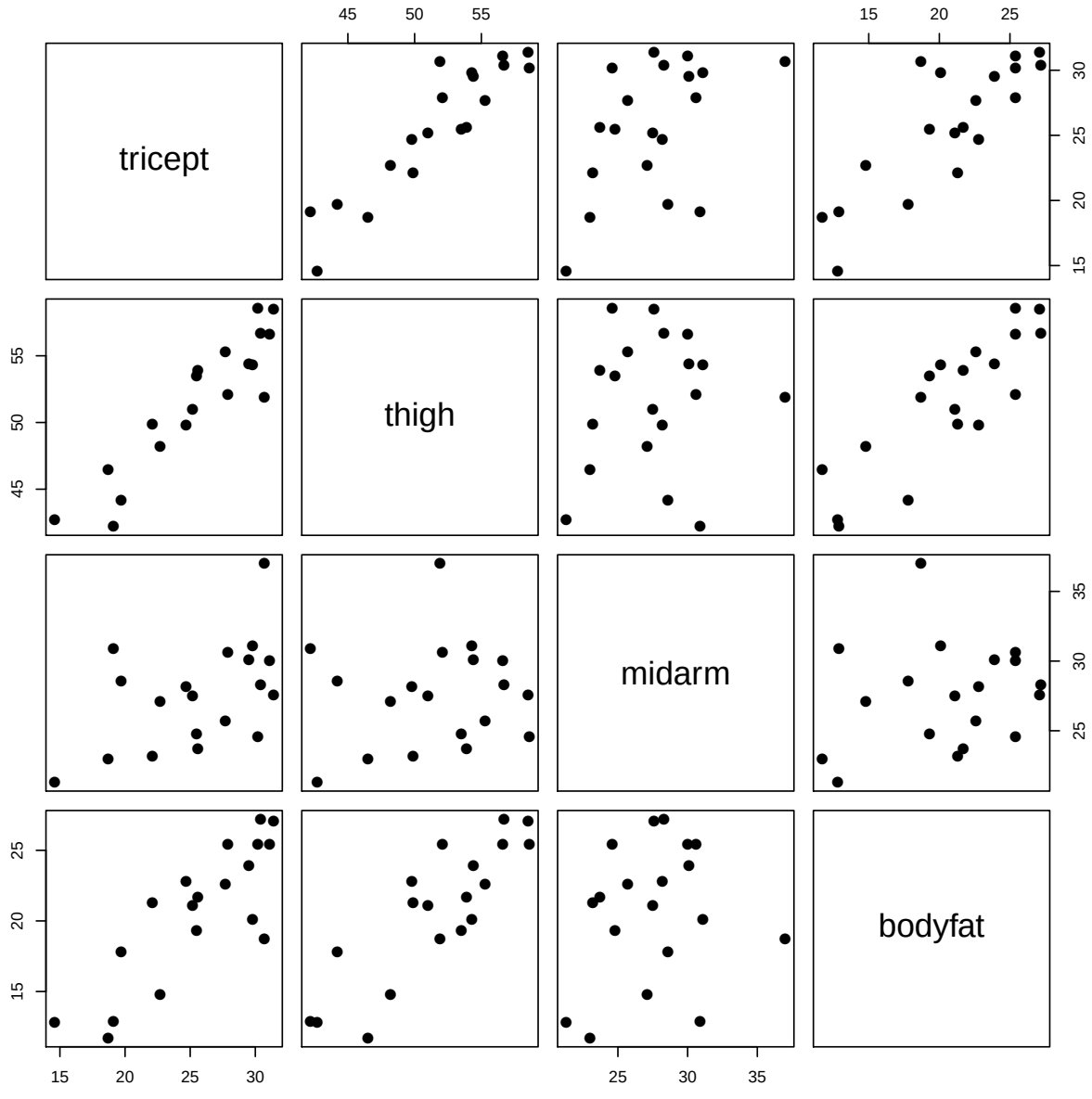
Obtaining Y_i, the percent of the body that is purely fat, requires immersing a person in water. Want to develop model based on simple body measurements that avoids people getting wet.

Full model

```
> fit<-lm(bodyfat ~ tricept + thigh + midarm, data=bodyfat)
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  189.944    107.067   1.774   0.0963 .
tricept        6.453      3.212   2.009   0.0629 .
thigh         -4.741      2.770  -1.711   0.1076
midarm        -3.267      1.688  -1.935   0.0721 .
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.383 on 15 degrees of freedom
Multiple R-squared:  0.7987,Adjusted R-squared:  0.7584
F-statistic: 19.84 on 3 and 15 DF,  p-value: 1.768e-05
```

Two of the three regression effects are negative. Holding midarm and triceps constant, increasing the thigh circumference 1 mm decreases bodyfat. Does this make sense?


tricept
thigh
midarm
bodyfat
45
50
55
15
20
25
30
35

Full model

```
> cordata<-cor(bodyfat)
> cordata
          tricept     thigh    midarm   bodyfat
tricept 1.0000000 0.9201905 0.5050211 0.8314028
thigh   0.9201905 1.0000000 0.1286432 0.8586697
midarm  0.5050211 0.1286432 1.0000000 0.1945733
bodyfat 0.8314028 0.8586697 0.1945733 1.0000000
```

There is high correlation among the predictors. For example $r = 0.92$ for triceps and thigh. These two variables are *essentially carrying the same information*. Maybe only one or the other is really needed.

In general, one predictor may be essentially perfectly predicted by the remaining predictors (a high "partial correlation"), and so would be unnecessary if the other predictors are in the model.

Detecting multicollinearity

A formal method for determining the presence of multicollinearity is the *variance inflation factor* (VIF). VIF's measure how much variances of estimated regression coefficients are inflated when compared to having uncorrelated predictors. We will start with the standardized regression model of Section 7.5.

$$\text{Let } Y_i^* = \frac{1}{\sqrt{n-1}} \frac{Y_i - \bar{Y}}{s_Y} \text{ and } x_{ij}^* = \frac{1}{\sqrt{n-1}} \frac{x_{ij} - \bar{x}_j}{s_j},$$

where $s_Y^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$,
$s_j^2 = (n-1)^{-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$, and $\bar{x}_j = n^{-1} \sum_{i=1}^n x_{ij}$.

VIF's

These variables are centered about their means and "standardized" to have Euclidean norm 1.

For example, $||\mathbf{x}_j^*||^2 = (x_{1j}^*)^2 + \cdots + (x_{nj}^*)^2 = 1$.

Note that in general $(\mathbf{Y}^*)'(\mathbf{Y}^*) = (\mathbf{x}_j^*)'(\mathbf{x}_j^*) = 1$ and

$(\mathbf{x}_j^*)'(\mathbf{x}_s^*) = corr(\mathbf{x}_j, \mathbf{x}_s) \stackrel{def}{=} r_{js}$.

VIF's

Consider the *standardized regression model*

$$Y_i^* = \beta_1^* x_{i1}^* + \cdots + \beta_k^* x_{ik}^* + \epsilon_i^*.$$

Define the $k \times k$ sample correlation matrix $\mathbf{R}$ for the standardized predictors, and the $n \times k$ design matrix $\mathbf{X}^*$ to be:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{21} & \cdots & r_{k1} \\ r_{12} & 1 & \cdots & r_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & \cdots & 1 \end{bmatrix}, \quad \mathbf{X}^* = \begin{bmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1k}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2k}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^* & x_{n2}^* & \cdots & x_{nk}^* \end{bmatrix}.$$

Since $(\mathbf{X}^*)'(\mathbf{X}^*) = \mathbf{R}$, the least-squares estimate of $\beta^* = (\beta_1^*, \ldots, \beta_k^*)'$ is given by $\mathbf{b}^* = \mathbf{R}^{-1}(\mathbf{X}^*)'\mathbf{Y}^*$. Hence $Cov(\mathbf{b}^*) = \mathbf{R}^{-1}(\sigma^*)^2$.

VIF's

Now note that if *all predictors are uncorrelated* then $\mathbf{R} = \mathbf{I}_k = \mathbf{R}^{-1}$. Hence the ith diagonal element of $\mathbf{R}^{-1}$ measures how much the variance of b_i^* is *inflated* due to correlation between predictors. We call this the ith *variance inflation factor*: $VIF_i = (\mathbf{R}^{-1})_{ii}$. Usually the largest VIF_i is taken to be a measure of the seriousness of the multicollinearity among the predictors.

Detecting multicollinearity

Predictor x_j has a *variance inflation factor* of

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is R^2 from regressing x_j on the remaining predictors $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$.

High R_j^2 (near 1) $\Rightarrow$ x_j is linearly associated with other predictors $\Rightarrow$ high VIF_j.

- $VIF_j \approx 1 \Rightarrow x_j$ is not involved in any multicollinearity.
- $VIF_j > 10 \Rightarrow x_j$ is involved in severe multicollinearity.

VIF in R

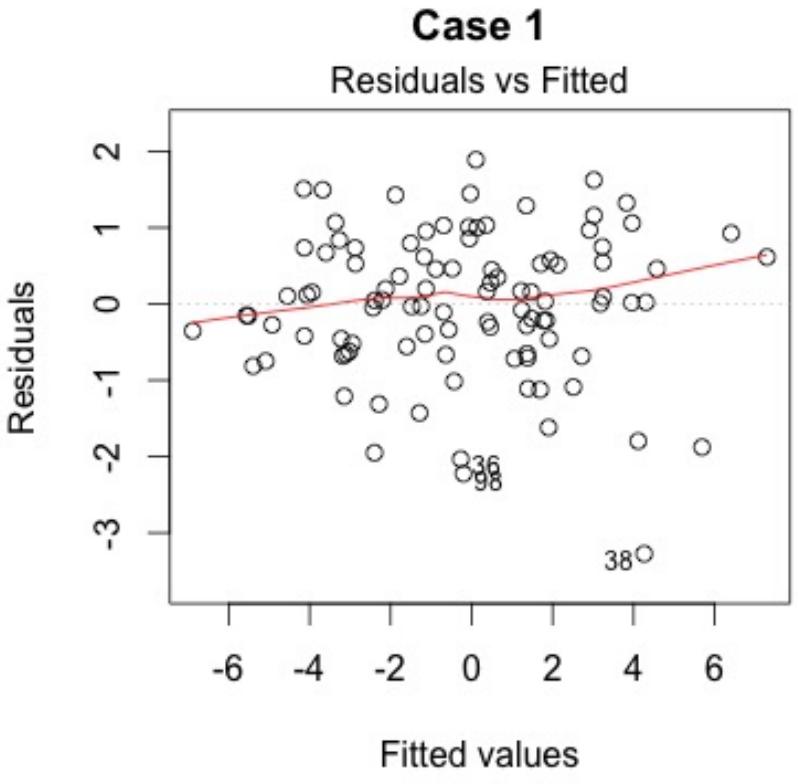
```
>vif(fit) # variance inflation factors
>tricept    thigh   midarm
>806.7198 611.0817 125.7097
```

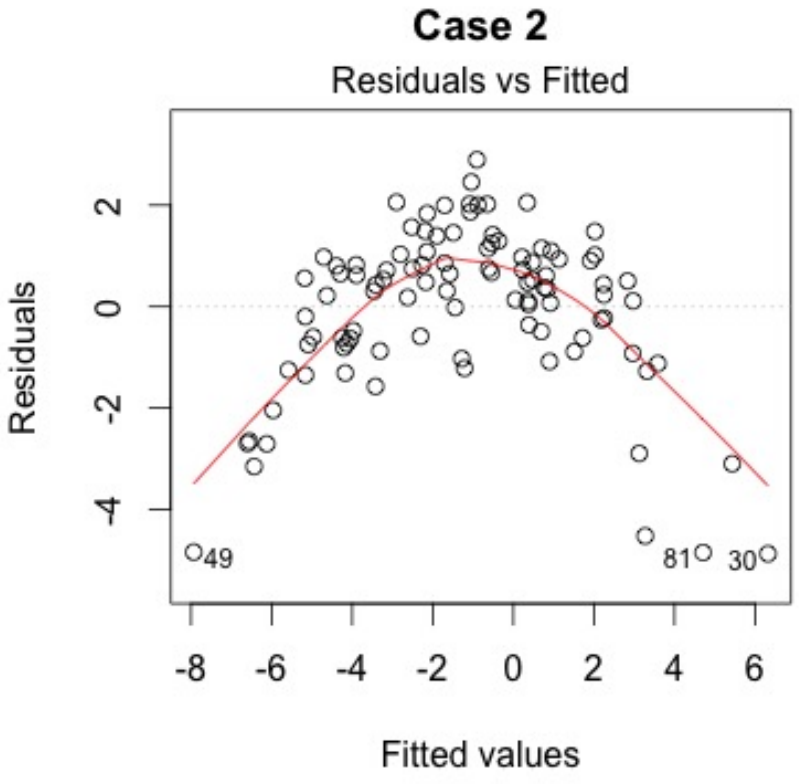
What do you conclude?

Remedies for multicollinearity

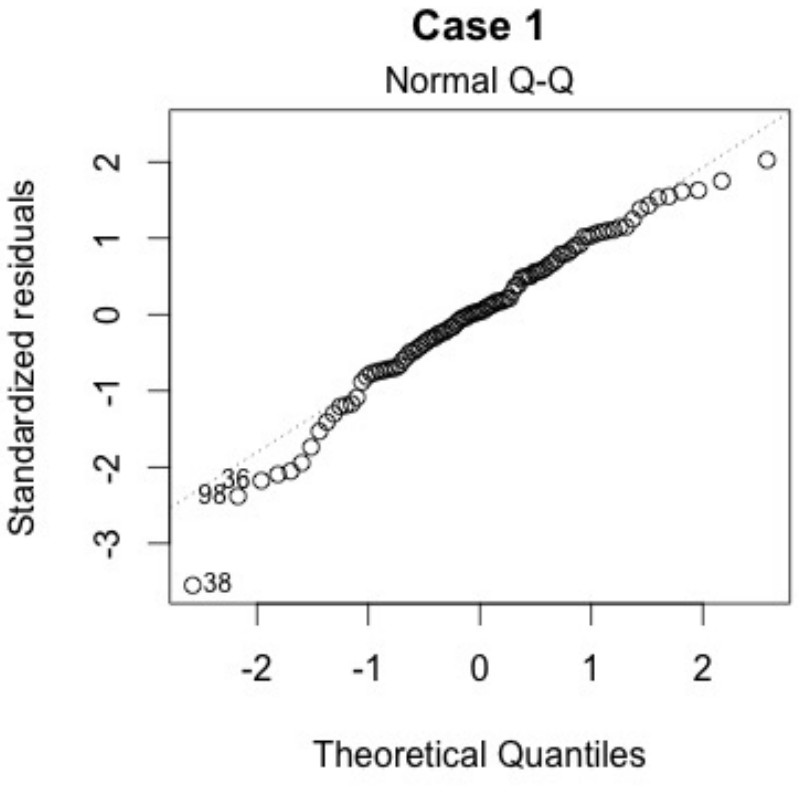
- Drop one or more predictors from the model (Chapter 9).
- More advanced: **principal components regression** uses indexes (new predictors) that are linear combinations of the original predictors as predictors in a new model. The indexes are selected to be uncorrelated. Disadvantage: the indexes might be hard to interpret.
- More advanced: **ridge regression** (Section 11.2).
- More advanced: **ensemble methods**

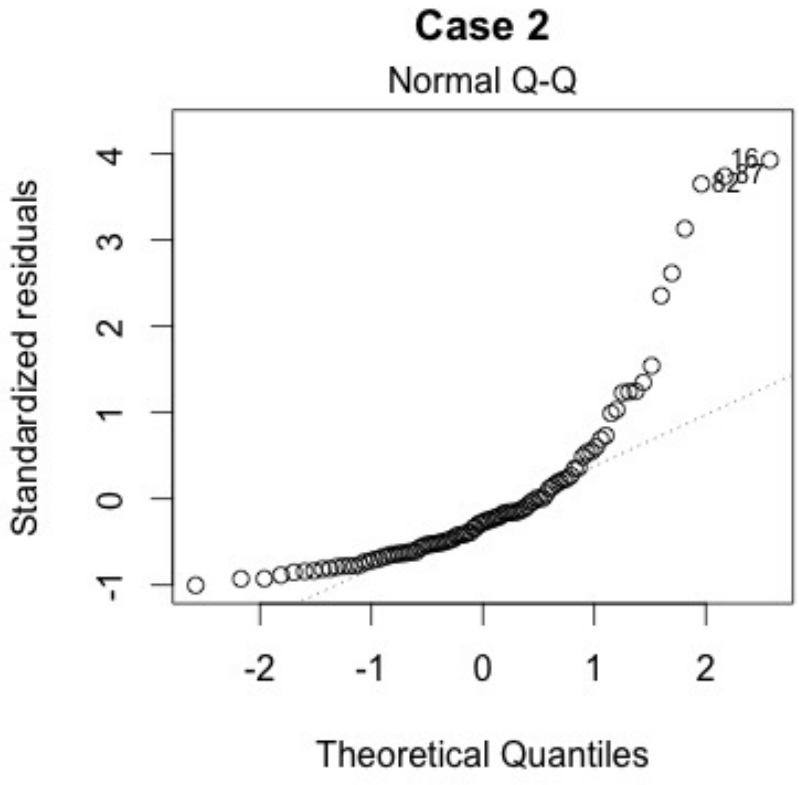
Standard Diagnostics Plots in R



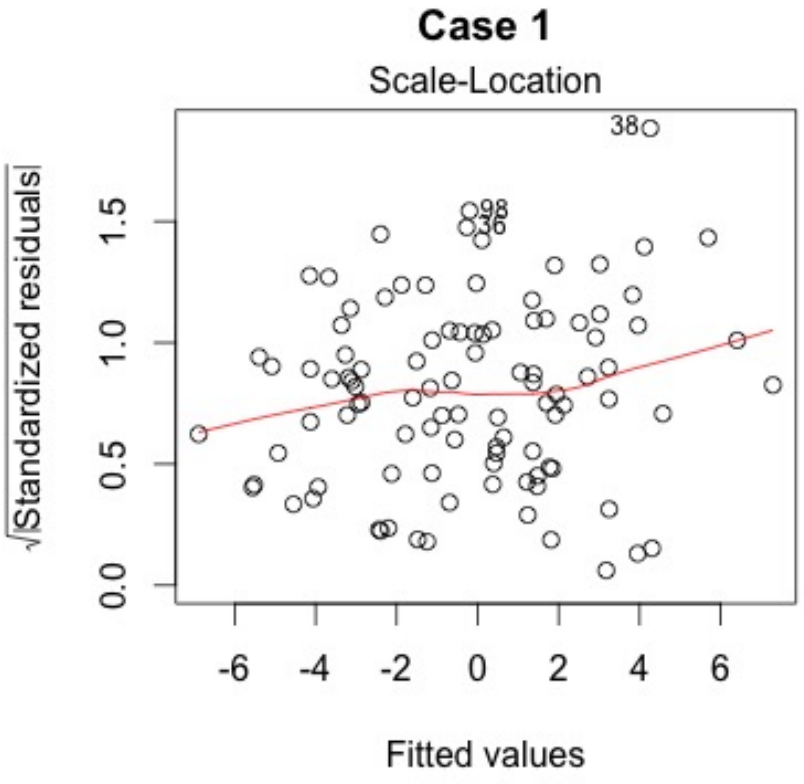


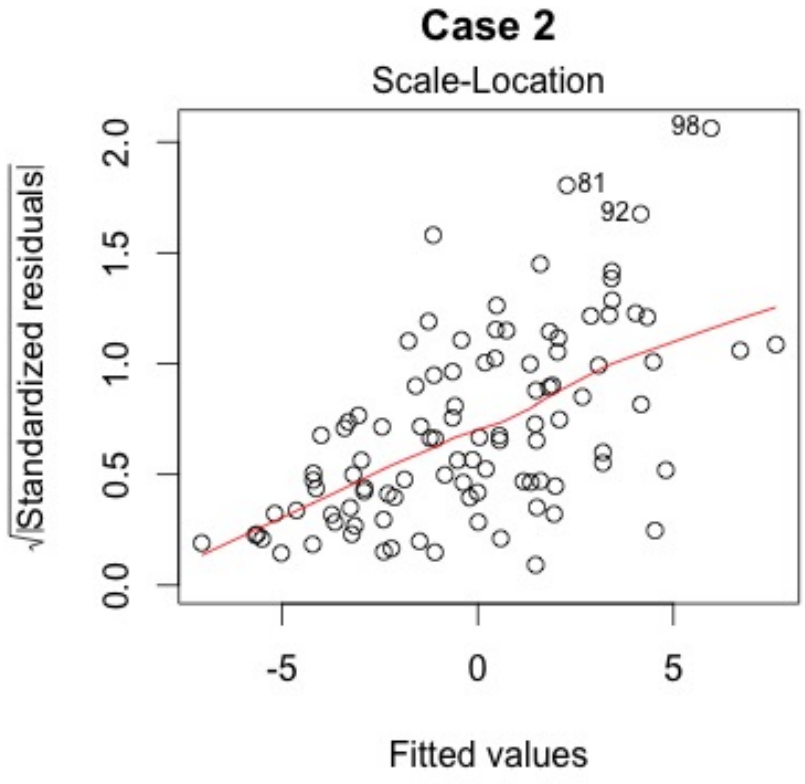
Standard Diagnostics Plots in R





Standard Diagnostics Plots in R





Standard Diagnostics Plots in R