PSTAT 131/231: Introduction to Statistical Machine Learning

Guo Yu

Lecture 5
More on Linear Regression

ISL Chapter 3

ESL (for 231 students) Chapter 3.1-3.2, 3.5

Quiz 1 is on this Friday Oct 7 from 12 pm to 9 pm PT

Last lecture…

Linear Regression: simple and multiple

Coefficient estimates

Assessing the accuracy of the coefficient estimates

Assessing the accuracy of the linear model

This lecture…

Other practical considerations in regression

(Hopefully) new perspectives on regression

Multiple regression

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Qualitative predictors

Extensions of the linear structures in $(X_1, \ldots, X_p)$

Linear regression diagnostics

Multiple regression

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Qualitative predictors

Extensions of the linear structures in $(X_1, \ldots, X_p)$

Linear regression diagnostics

Extensions of the linear model

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Relationship between Y and $X_1, \ldots, X_p$ is additive and linear

Additive: the effect of changes in a predictor X_j on Y is independent of the values of all other predictors

Linear: the change in the response Y due to one unit change of a predictor X_j is constant, regardless of the value of X_j

Removing the additive assumption

One way of removing the additive assumption is to include interaction term

from $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

One-unit increase in X_1 results in β_1-unit increase in Y (holding X_2 fixed)

to $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \boxed{X_1 X_2} + \varepsilon$$

interaction between X_1 and X_2

$$= \beta_0 + \boxed{(\beta_1 + \beta_3 X_2)} X_1 + X_2 \beta_2 + \varepsilon$$

One-unit increase in X_1 results in $\beta_1 + \beta_3 X_2$-unit increase in Y

the effect of X_1 on Y is no longer constant: adjusting X_2 will change the impact of X_1 on Y

Interaction: a data example

$$\textbf{sales} = \beta_0 + \beta_1 \times \textbf{TV} + \beta_2 \times \textbf{radio} + \beta_3 \times \textbf{TV} \times \textbf{radio} + \varepsilon$$

$$= \beta_0 + (\beta_1 + \beta_3 \times \textbf{radio}) \times \textbf{TV} + \beta_2 \times \textbf{radio} + \varepsilon$$

		Coefficient	Std. error	t-statistic	p-value
$\hat{\beta}_0$	Intercept	6.7502	0.248	27.23	< 0.0001
$\hat{\beta}_1$	TV	0.0191	0.002	12.70	< 0.0001
$\hat{\beta}_2$	radio	0.0289	0.009	3.24	0.0014
$\hat{\beta}_3$	TV×radio	0.0011	0.000	20.73	< 0.0001

Interpretation: an **increase** in **TV advertising** of **$1,000** is associated with $(\hat{\beta}_1 + \hat{\beta}_3 \times \textbf{radio}) \times 1000 = 19.1 + 1.1 \times \textbf{radio}$ dollars **increase** in sales

Practice: what is the effect of an increase in radio advertising of $1,000 on sales?

Extensions of the linear model

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Relationship between Y and $X_1, \ldots, X_p$ is additive and linear

Additive: the effect of changes in a predictor X_j on Y is independent of the values of all other predictors

Interaction terms!

Linear: the change in the response Y due to one unit change of a predictor X_j is constant, regardless of the value of X_j

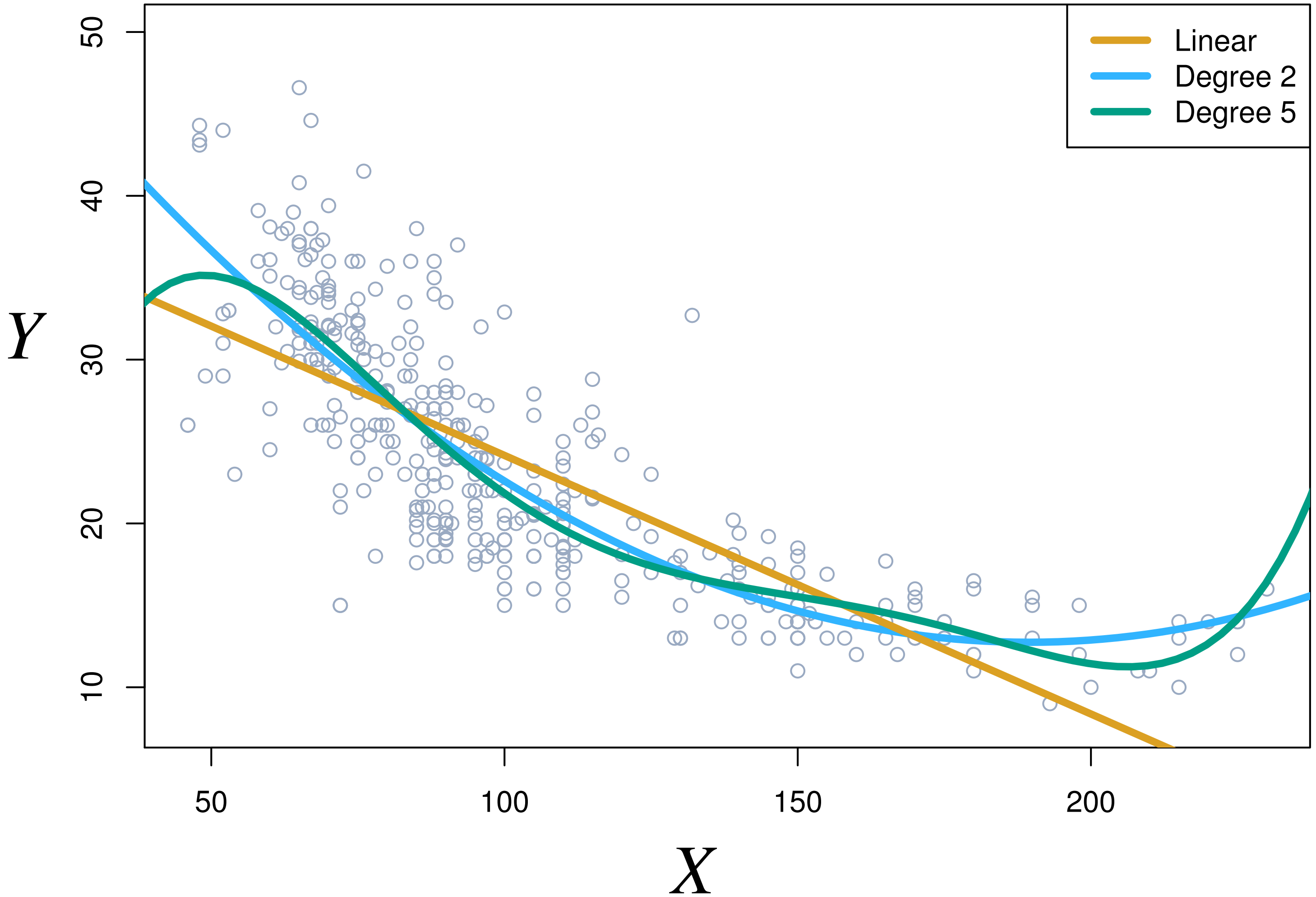
polynomial regression

Non-linear relationships

Polynomial regression of Y on X

degree

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_d X^{\boxed{d}} + \varepsilon$$



Non-linear relationships

Polynomial regression of Y on X

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_d X^d + \varepsilon$$

Y **is no longer linear in** X

But this is still a linear model!!!

Simply let $Z_k = X^k$**...**

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_2 + \ldots + \beta_d Z_d + \varepsilon$$

Y **is still linear in** $X, Z_2, \ldots, Z_d$

Practical considerations in regression

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Qualitative predictors

Extensions of the linear structures in $(X_1, \ldots, X_p)$

Linear regression diagnostics

Assumptions in MLR: model

$$\mathbf{y} = \boxed{\mathbf{X}\beta} + \varepsilon$$

Model structure: linear relationship between response and the predictors

e.g., the response does not depend on X_1^2, nor e^{X_1}, nor $\log(|X_1|)$

instead, the response depends on X_1 through β_1

Assumptions in MLR: random error

$$\mathbf{y} = \mathbf{X}\beta + \boxed{\varepsilon}$$

ε_i's are **i.i.d (unobservable) normal random errors:** $\varepsilon_i \sim N(0, \sigma^2)$

What is this assumption really about?

A1: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **have equal variance, which is** σ^2

A2: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are normally distributed**

A3: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are independent (which implies that** $y_1, y_2, \ldots, y_n$ **are independent)**

Assumptions in MLR: data

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

All the n observations in the dataset follow this model

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{array}{c} \mathbf{x}_1 \\ \\ \\ \\ \\ \end{array} \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Assumptions in MLR: data

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

All the n observations in the dataset follow this model

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{array}{c} \\ \mathbf{x}_2 \\ \\ \\ \\ \end{array} \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Assumptions in MLR: data

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

All the n observations in the dataset follow this model

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\mathbf{x}_n$

Assumptions in MLR: model

$$\mathbf{y} = \boxed{\mathbf{X}\beta} + \varepsilon$$

Model structure: linear relationship between response and the predictors

e.g., the response does not depend on X_1^2, nor e^{X_1}, nor $\log(|X_1|)$

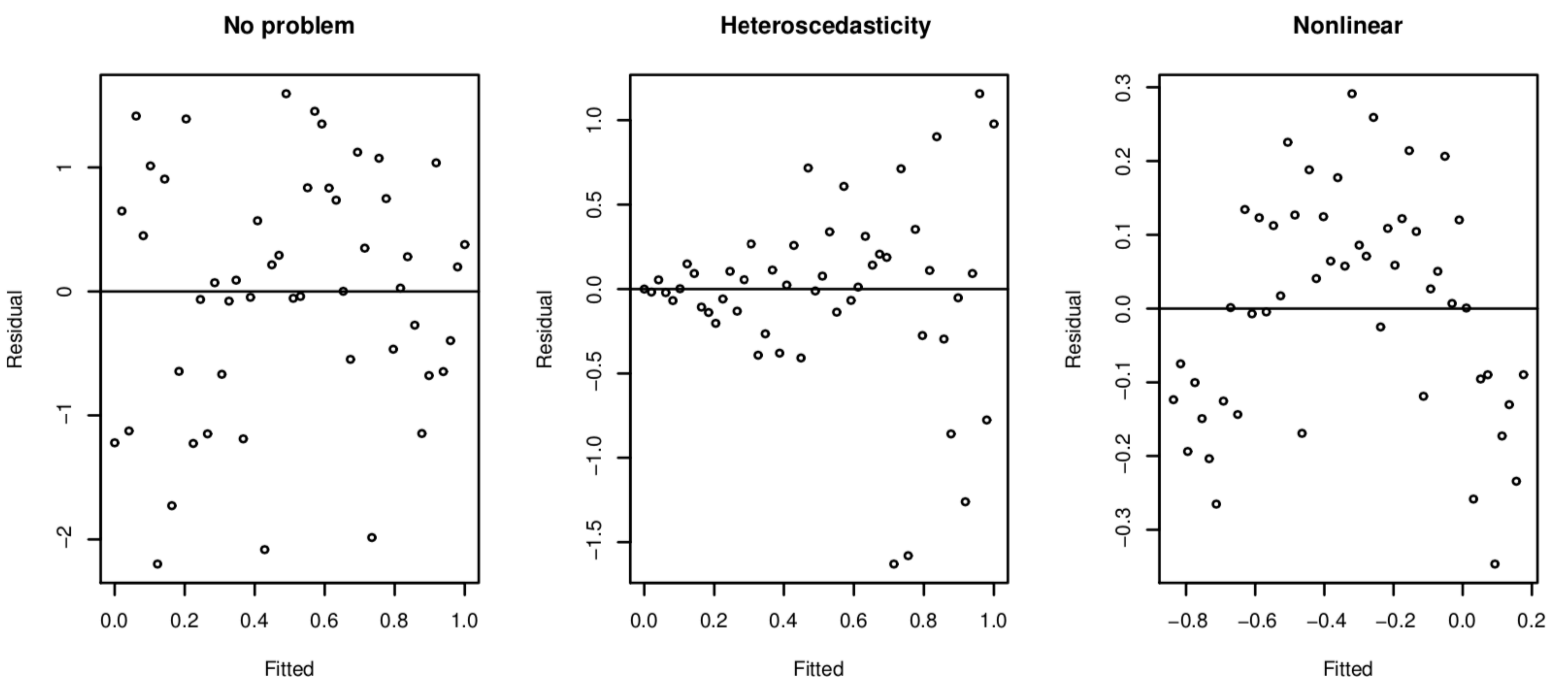
instead, the response depends on X_1 through β_1

Checking the linear relationship

We can still use the

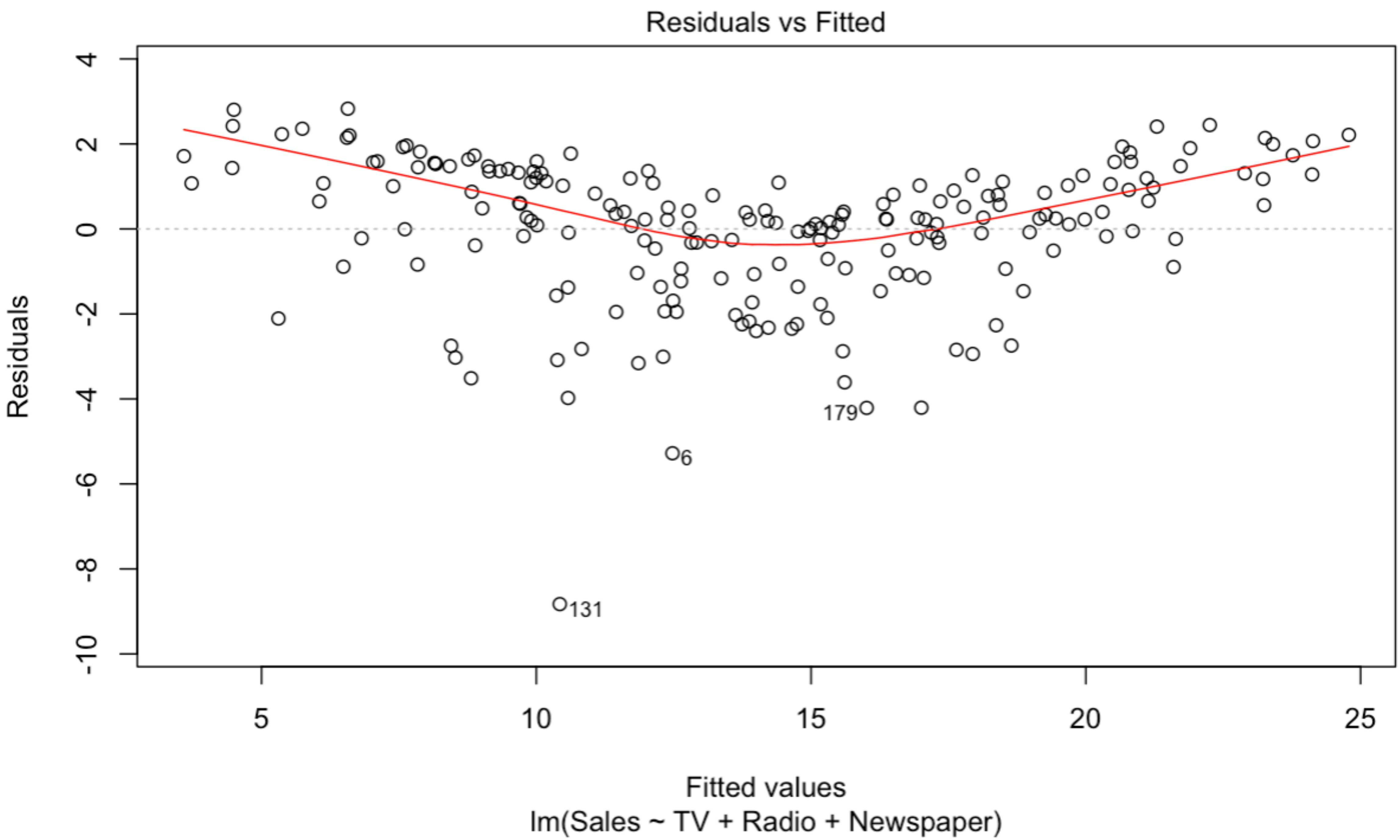
Residual plot **plot the residual $y_i - \hat{y}_i$ v.s the fitted value (prediction) $\hat{y}_i$**

If the linear assumption holds, then the plot will NOT show discernible pattern



Checking the linear relationship in R

```
mod <- lm(formula = Sales ~ TV + Radio + Newspaper, data = ad.data)
plot(mod)
```



Assumptions in MLR: random error

$$\mathbf{y} = \mathbf{X}\beta + \boxed{\varepsilon}$$

ε_i's are **i.i.d** **(unobservable) normal random errors:** $\varepsilon_i \sim N(0, \sigma^2)$

What is this assumption really about?

A1: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **have equal variance, which is** σ^2

A2: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are normally distributed**

A3: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are independent (which implies that** $y_1, y_2, \ldots, y_n$ **are independent)**

What could go wrong in ε?

$$\mathbf{y} = \mathbf{X}\beta + \boxed{\varepsilon}$$

ε_i's are i.i.d (unobservable) normal random errors: $\varepsilon_i \sim N(0, \sigma^2)$

how can we tell if ε_i's have a constant variance?

A1: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have equal variance, which is σ^2

how can we tell if ε_i follow normal distribution?

A2: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are normally distributed

how can we tell if ε_i are independent (or at least uncorrelated)?

A3: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent (which implies that $y_1, y_2, \ldots, y_n$ are independent)

What could go wrong in ε?

$$\mathbf{y} = \mathbf{X}\beta + \boxed{\varepsilon}$$

ε_i's are i.i.d (unobservable) normal random errors: $\varepsilon_i \sim N(0, \sigma^2)$

Wait… the random errors ε_i's are not observable, how can we check them?

We can instead examine the residuals

$$\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

where $\boxed{\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \in \mathbb{R}^{n \times n}}$ hat matrix

Technically, the residuals and random errors are not interchangeable

But diagnostics can reasonably be applied to the residuals

What could go wrong in ε?

$$\mathbf{y} = \mathbf{X}\beta + \boxed{\varepsilon}$$

ε_i's are **i.i.d** (unobservable) **normal** random errors: $\varepsilon_i \sim N(0,\sigma^2)$

how can we tell if ε_i's have a constant variance?

A1: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **have equal variance, which is** σ^2

A2: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are normally distributed**

A3: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are independent (which implies that** $y_1, y_2, \ldots, y_n$ **are independent)**

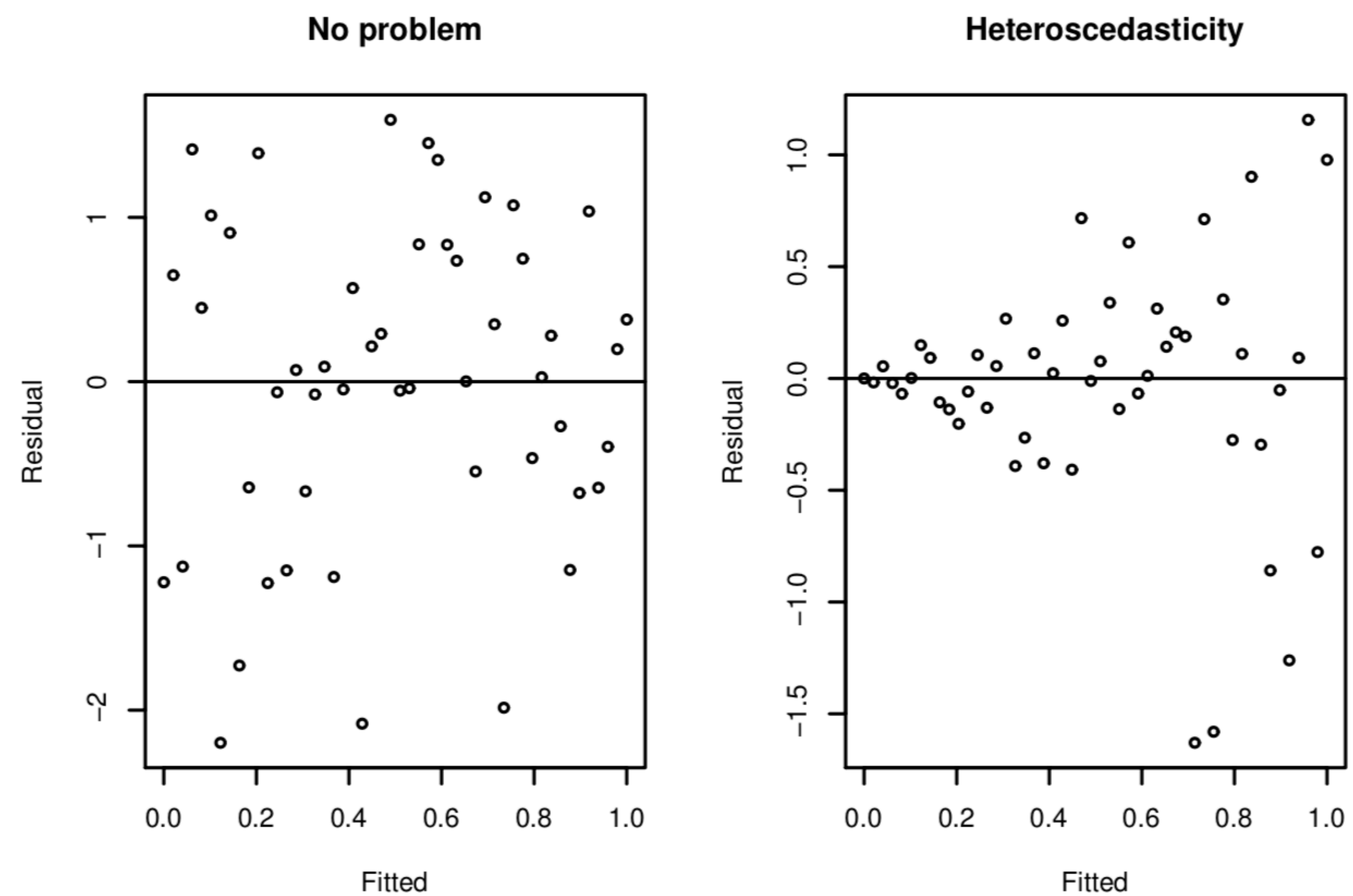
Checking constant variance

A1: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have equal variance, which is σ^2

The most commonly used diagnostic is a plot of residuals against fitted values ($\hat{y}$)

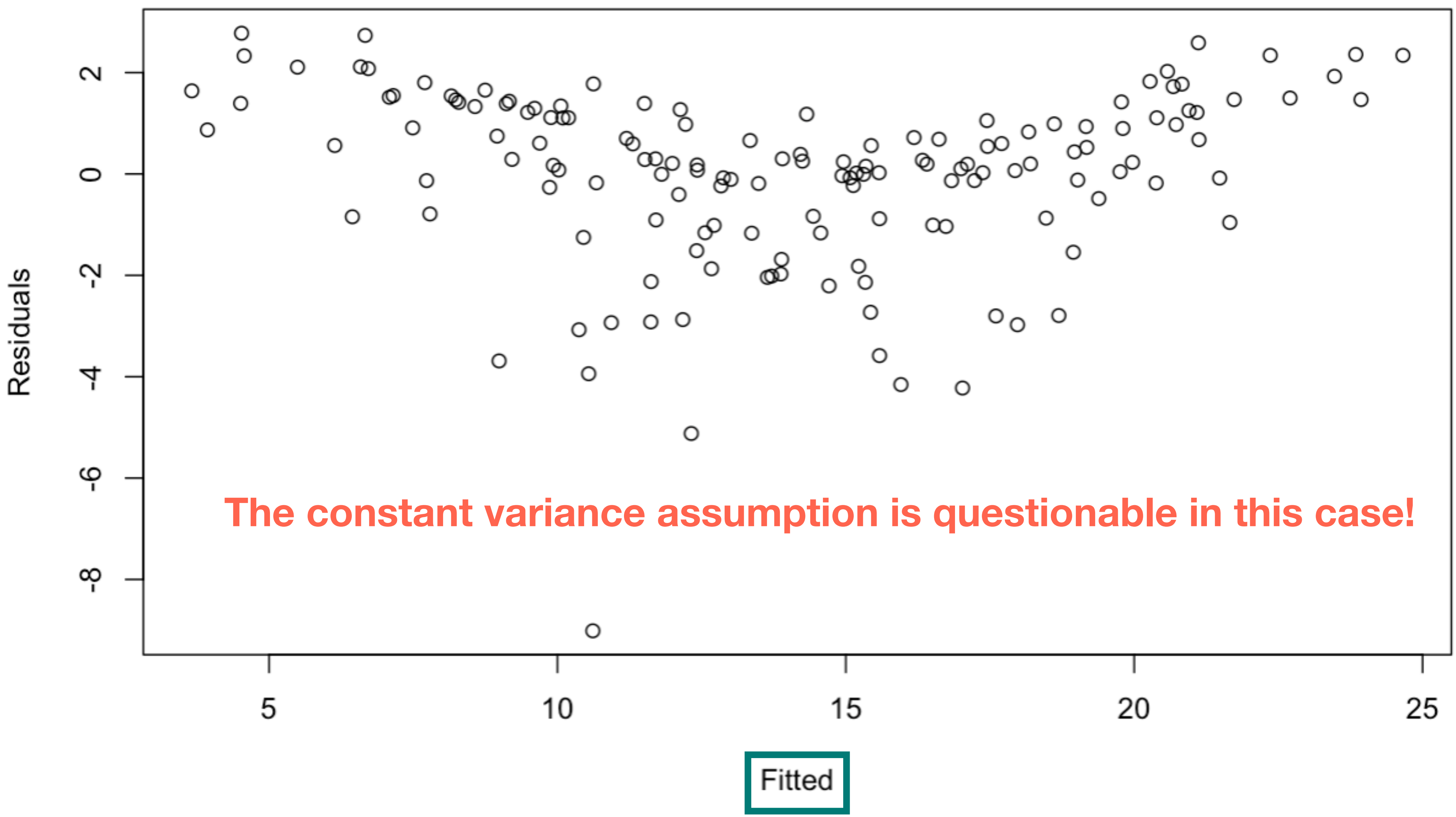
If the constant variance assumption holds

we should observe constant symmetrical variation (homoscedastic)



Checking constant variance in R

```
mod <- lm(formula = Sales ~ TV + Radio + Newspaper, data = ad.data)
plot(fitted(mod), residuals(mod), xlab = "Fitted", ylab = "Residuals")
```



The constant variance assumption is questionable in this case!

What could go wrong in ε?

$$\mathbf{y} = \mathbf{X}\beta + \boxed{\varepsilon}$$

ε_i's are i.i.d (unobservable) normal random errors: $\varepsilon_i \sim N(0, \sigma^2)$

A1: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **have equal variance, which is** σ^2

how can we tell if ε_i follow normal distribution?

A2: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are normally distributed**

A3: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are independent (which implies that** $y_1, y_2, \ldots, y_n$ **are independent)**

Checking normal distribution

A2: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are normally distributed**

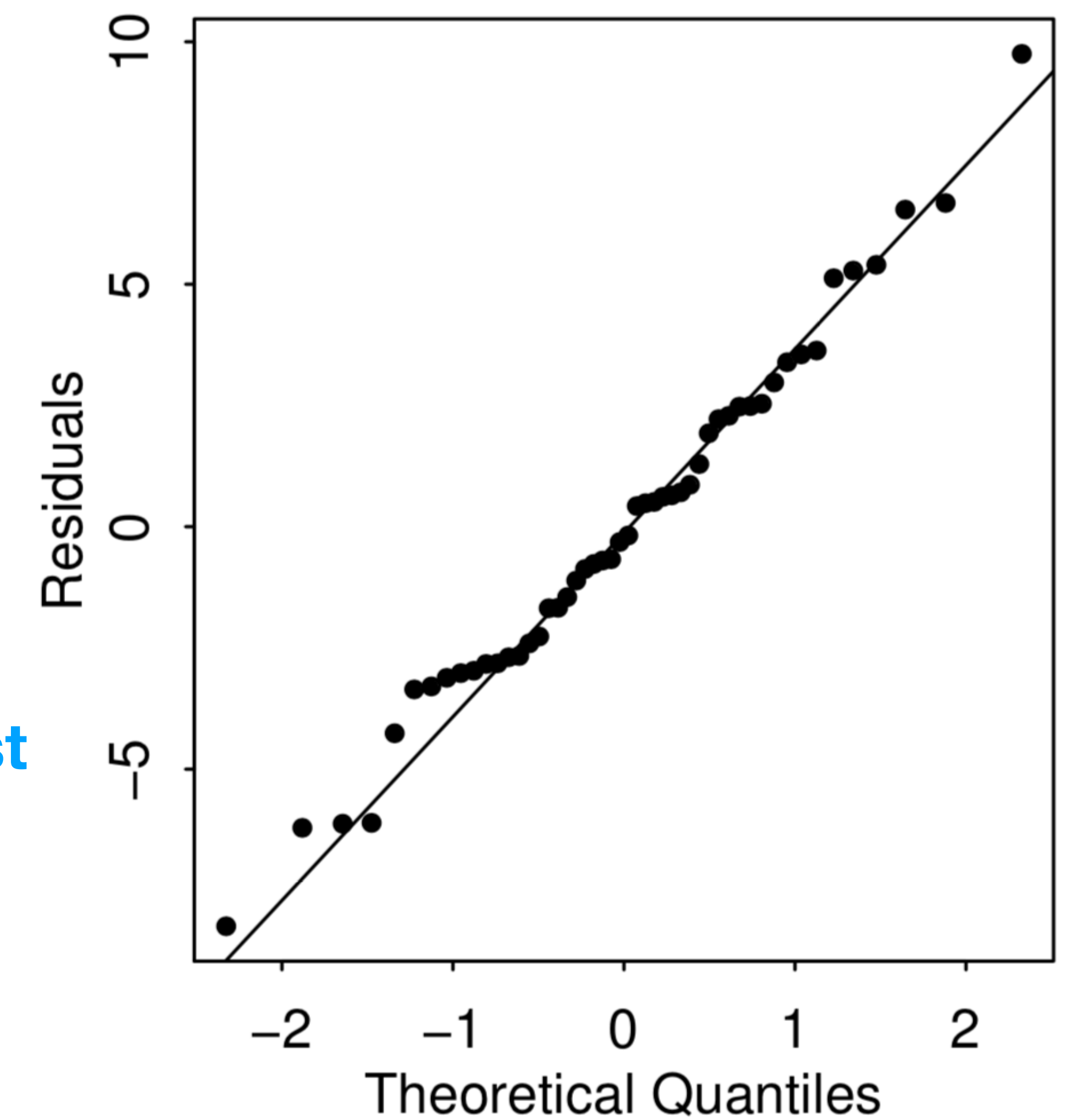
The most commonly used diagnostic is a Q-Q plot

We compare the residuals to the actually normally distributed observations

If the normal assumption holds,

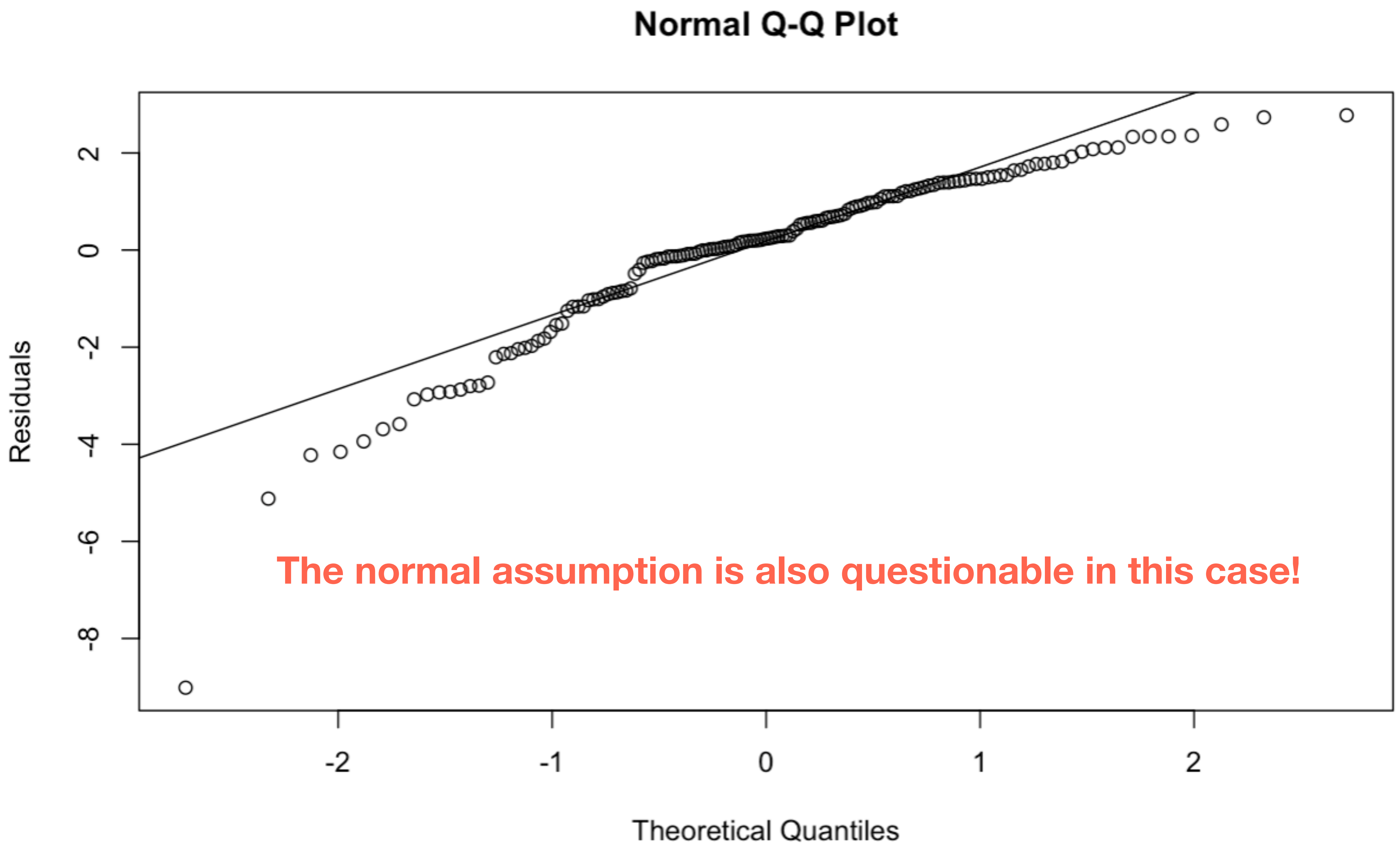
we should observe the dots follow the line

A formal test for normality is Shapiro-Wilk test



Checking normal distribution in R

```
mod <- lm(formula = Sales ~ TV + Radio + Newspaper, data = ad.data)
qqnorm(residuals(mod), ylab = "Residuals")
qqline(residuals(mod))
```



Checking normal distribution in R

```
> shapiro.test(residuals(mod))

        Shapiro-Wilk normality test

data:  residuals(mod)
W = 0.89811, p-value = 1.035e-08
```

The null hypothesis is that the residuals are normally distributed

Since p-value is extremely small, we reject the null hypothesis

What could go wrong in ε?

$$\mathbf{y} = \mathbf{X}\beta + \boxed{\varepsilon}$$

ε_i's are **i.i.d** (unobservable) **normal** random errors: $\varepsilon_i \sim N(0, \sigma^2)$

A1: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **have equal variance, which is** σ^2

A2: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are normally distributed**

how can we tell if ε_i are independent (or at least uncorrelated)?

A3: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are independent (which implies that** $y_1, y_2, \ldots, y_n$ **are independent)**

Checking correlation structures

A3: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ **are independent (which implies that** $y_1, y_2, \ldots, y_n$ **are independent)**

Difficult to check, since there are too many possible patterns of correlation

Some types of data have specific structure of correlation

e.g., spatial or temporal data

Then what should we do?

When problems are seen in diagnostic plots

Some modification of the model is suggested

If the problem is on non-constant variance

Consider doing (variance stabilizing) transformation of the response

If the problem is on correlated errors

Directly build the correlation into the model: generalized least squares

If the problem is on non-normal random errors

Usually less concerning, and could be results of other violations of model assumptions

Assumptions in MLR: data

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

$\boxed{\text{All the } n \text{ observations } (y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n) \text{ follow this model}}$

Essentially three types of observations could break the assumption:

High Leverage Points: unusual values for $\mathbf{x}_i$

Outliers: unusual values for y_i given $\mathbf{x}_i$

Influential observations: substantially change the model fit

High leverage points

Recall the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \in \mathbb{R}^{n \times n}$$

A symmetric matrix with many special properties

1. $\mathbf{H}\mathbf{X} = \mathbf{X}$
2. $\mathbf{H}\mathbf{H} = \mathbf{H}$

Diagonal elements in $\mathbf{H}$ are defined as the leverages

$$\mathbf{H}_{ii} \text{ is the leverage of } \mathbf{x}_i$$

$\mathbf{H}_{ii}$ measures the distance between the $\mathbf{x}_i$ and the average of all $\mathbf{x}$'s in the dataset

Why do we care about leverage?

One can show that

$$\mathrm{Cov}(\mathbf{y} - \hat{\mathbf{y}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$$

which implies that

$$\mathrm{Var}(y_i - \hat{y}_i) = \sigma^2(1 - \mathbf{H}_{ii})$$

A large $\mathbf{H}_{ii}$ will make the i-th residual to have a very small variance

No matter what value of y_i is observed for the i-th observation

we are nearly certain to get a fixed value of residual

The effect of $\mathbf{x}_i$ is overwhelming the effect of y_i

How do we find high leverage points?

Recall that $\sum_{i=1}^{n} \mathbf{H}_{ii} = p + 1$

The average leverage for all the n observations is $$\frac{p+1}{n}$$

We should suspect an observation with a leverage that greatly exceeds $(p+1)/n$

The rule of thumb: examine any observations with 2-3 times greater than $(p+1)/n$

Finding high leverage points in R

the *hatvalues* function calculates the leverages of all observations

```
mod <- lm(formula = Sales ~ TV + Radio + Newspaper, data = ad.data)
lev <- hatvalues(mod)
```

alternatively, we can directly calculate the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and take its diagonal elements

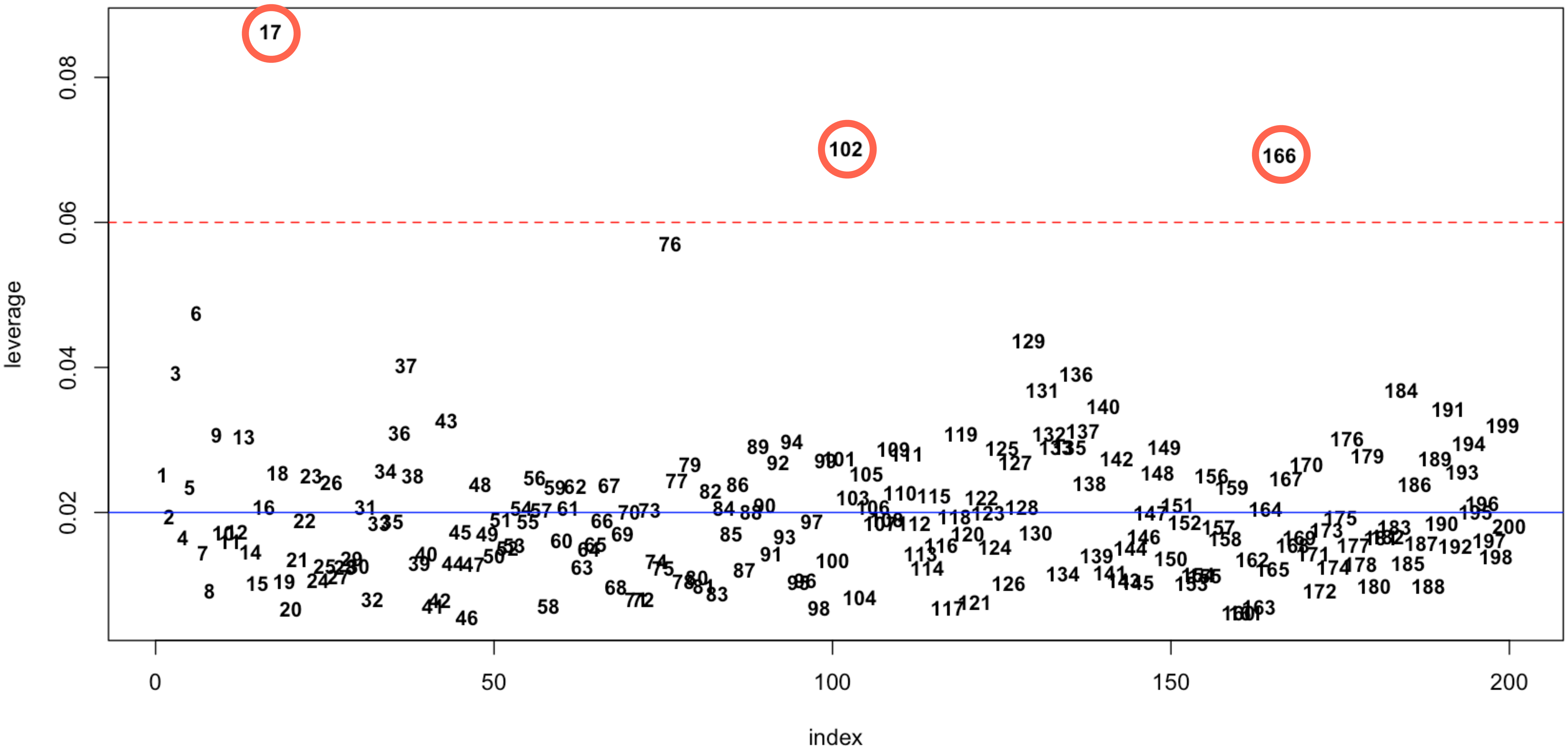
```
x <- model.matrix(mod)
H <- x %*% solve(crossprod(x), t(x))
lev_equivalent <- diag(H)
```

These two approaches give the identical results

```
> all.equal(lev, lev_equivalent)
[1] TRUE
```

Finding high leverage points in R

```r
n <- nrow(ad.data)
p <- 3
dat <- data.frame(index = seq(length(lev)), leverage = lev)
plot(leverage ~ index, col = "white", data = dat, pch = NULL)
text(leverage ~ index, labels = index, data = dat, cex=0.9, font=2)
abline(h = (p + 1) / n, col = "blue")
abline(h = 3 * (p + 1) / n, col = "red", lty = 2)
```



Assumptions in MLR: data

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

All the n observations $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)$ follow this model

Essentially three types of observations could break the assumption:

High Leverage Points: unusual values for $\mathbf{x}_i$

Outliers: unusual values for y_i given $\mathbf{x}_i$

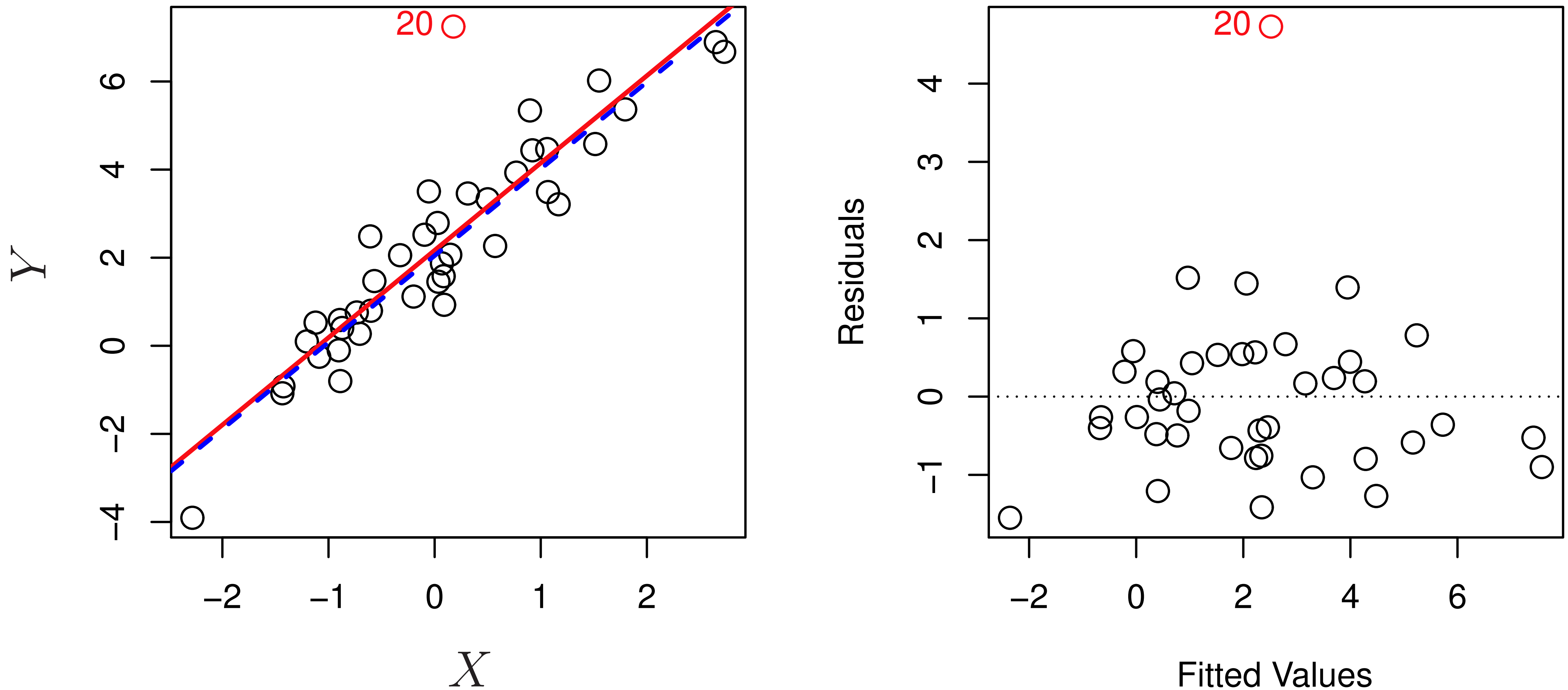
Influential observations: substantially change the model fit

Outliers

Outlier: a data point for which y_i is far from $\hat{y}_i$ given by the model

Outliers can arise for multiple reasons

e.g., incorrect recording of an observation during data collection



How do we find outliers?

Residual vs fitted value plots can be used to identify outliers

But how large the residual should be before we consider the point to be an outlier?

Instead, we consider the standardized residuals

$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - \mathbf{H}_{ii}}}$$

recall that $\mathrm{Var}(y_i - \hat{y}_i) = \sigma^2(1 - \mathbf{H}_{ii})$

The rule of thumb: any observations whose absolute standardized residuals $\geq$ 3

Finding outliers in R

The *rstandard* function calculates the standardized residuals

```
mod <- lm(formula = Sales ~ TV + Radio + Newspaper, data = ad.data)
r <- rstandard(mod)
```

Alternatively, we can directly calculate the standardized residuals by definition

```
resid <- residuals(mod)
rse <- summary(mod)$sigma
r_equivalent <- resid / (rse * sqrt(1 - lev))
```

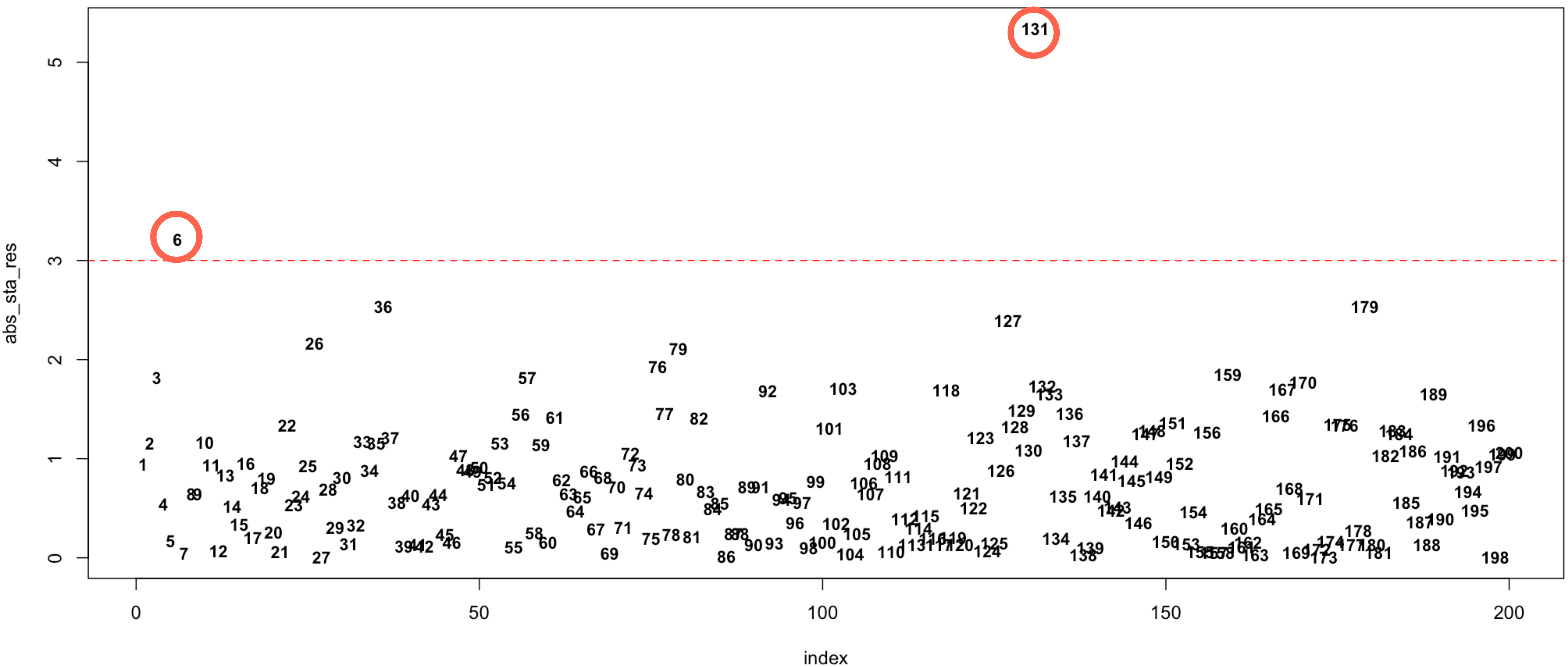
$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - \mathbf{H}_{ii}}}$$

These two approaches give the identical results

```
> all.equal(r, r_equivalent)
[1] TRUE
```

Finding outliers in R

```r
dat <- data.frame(index = seq(length(r)), abs_sta_res = abs(r))
plot(abs_sta_res ~ index, col = "white", data = dat, pch = NULL)
text(abs_sta_res ~ index, labels = index, data = dat, cex=0.9, font=2)
abline(h = 3, col = "red", lty = 2)
```



Assumptions in MLR: data

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

$\boxed{\text{All the } n \text{ observations } (y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n) \text{ follow this model}}$

Essentially three types of observations could break the assumption:

High Leverage Points: unusual values for $\mathbf{x}_i$

Outliers: unusual values for y_i given $\mathbf{x}_i$

Influential observations: substantially change the model fit

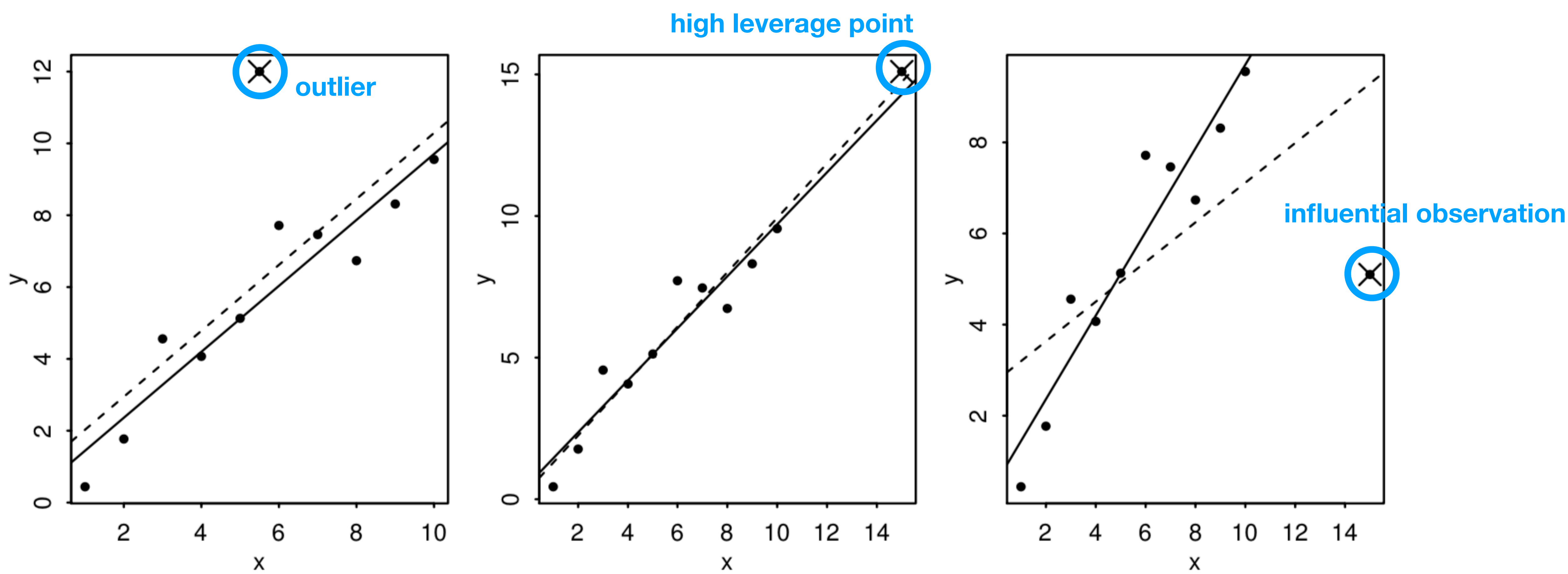
Influential observations

Observations whose removal from the dataset would cause a large change in the model fit

An influential observation may or may not be an outlier

An influential observation may or may not have large leverage

An influential observation will tend to have at least one of these two properties



How do we find influential observation?

Usually use Cook's distance

$$D_i = \frac{1}{p+1} \boxed{r_i}^2 \frac{\mathbf{H}_{ii}}{1 - \mathbf{H}_{ii}}$$

standardized residual

The rule of thumb: any observations with $D_i > \frac{4}{n}$

Finding influential observations in R

The *cooks.distance* function calculates the Cook's distance

```
mod <- lm(formula = Sales ~ TV + Radio + Newspaper, data = ad.data)
d <- cooks.distance(mod)
```

Alternatively, we can directly calculate the Cook's distance by definition

```
p <- 3
r <- rstandard(mod)
d_equivalent <- r^2 * lev / (1 - lev) / (p + 1)
```

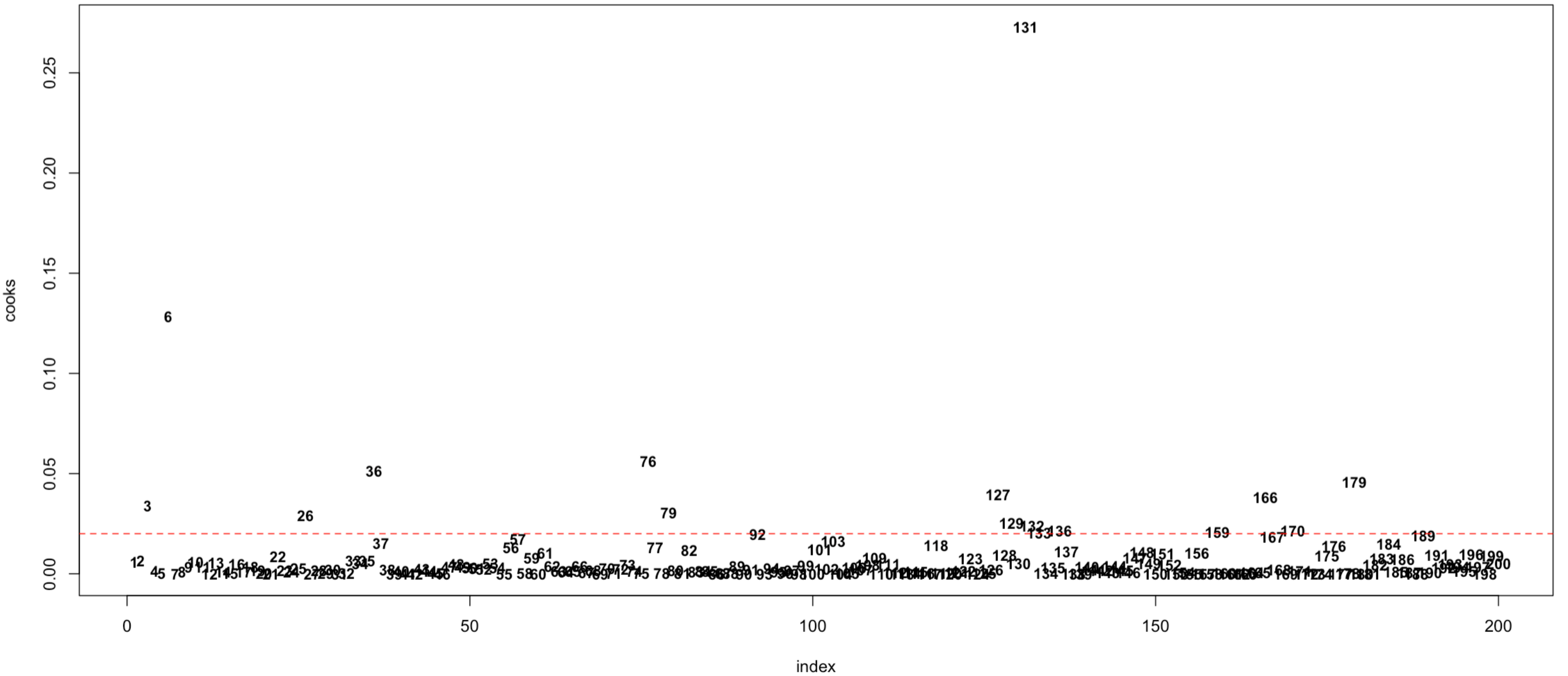
$$D_i = \frac{1}{p+1} r_i^2 \frac{\mathbf{H}_{ii}}{1 - \mathbf{H}_{ii}}$$

These two approaches give the identical results

```
> all.equal(d, d_equivalent)
[1] TRUE
```

Finding influential observations in R

```r
dat <- data.frame(index = seq(length(d)), cooks = d)
plot(cooks ~ index, col = "white", data = dat, pch = NULL)
text(cooks ~ index, labels = index, data = dat, cex=0.9, font=2)
abline(h = 4 / n, col = "red", lty = 2)
```



To consider in diagnostics of data

A high-leverage point / outlier / influential observation in one model may not be

a high-leverage point / outlier / influential observation in another model

What should we do once we find such observations?

1. **Check if there is data-entry error**

2. **Exclude the points**

3. **Try re-including them later if the model is changed**

Collinearity

Collinearity: two or more predictors are closely related to one another

collinear

If two predictors tend to increase or decrease together, it can be difficult to determine how each one is associated with the response

The variance of the estimates increase

How to detect collinearity?

Approach 1: look at correlation matrix of $X_1, \ldots, X_p$

Approach 2: compute the variance inflation factor

How to handle collinearity between, say, X_1 and X_2?

Approach 1: drop one of X_1, X_2 in regression model

Approach 2: combine X_1 and X_2 (hard to interpret)

This lecture…

Other practical considerations in regression

New perspectives on regression

Linear regression vs K-NN regression

in the general regression setting $Y = f(X) + \varepsilon$

Parametric approach	Non-parametric approach
Assume that $f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$	$f(X)$ can have any function form
No need to tune the model	Tuning parameter: K
Performs well when the true $f(X)$ is close to linear	Much more general-purpose
Interpretability, statistical inference…	Not very interpretable
Can be extended to work when p is very large ridge regression, lasso …	Curse of Dimensionality

Bias-Variance tradeoff in linear regression

Assume that $Y = f(X) + \varepsilon = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

For $\hat{f} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_{01} + \ldots + \hat{\beta}_p \mathbf{x}_{0p}$**, where** $\hat{\beta}_0, \ldots, \hat{\beta}_p$ **are least-squares estimates**

Property 1: Unbiased, i.e., $\mathrm{Bias}(\hat{f}(\mathbf{x}_0)) = \mathrm{E}[\hat{f}(\mathbf{x}_0)] - f(\mathbf{x}_0) = 0$

Property 2: Least-squares has the **smallest** expected test error among all **unbiased linear** estimates (**Gauss-Markov Theorem**)

Modern regression methods can **outperform** least-squares in terms of expected test MSE, by **having small bias** but **having much smaller variance**

In summary

Practical considerations in regression

- **Qualitative predictors**
- **Extensions of the linear structures in $(X_1, \ldots, X_p)$**
- **Linear regression diagnostics**

New perspectives on regression

- **Compare linear regression with K-NN regression**
- **Bias-variance tradeoff of linear regression**

Next…

Linear Classification method: logistic regression

Quiz 1 tomorrow!