STA302/STA1001, Weeks 4–5

Mark Ebden, 3–5 October 2017

With grateful acknowledgment to Alison Gibbs and Becky Lin

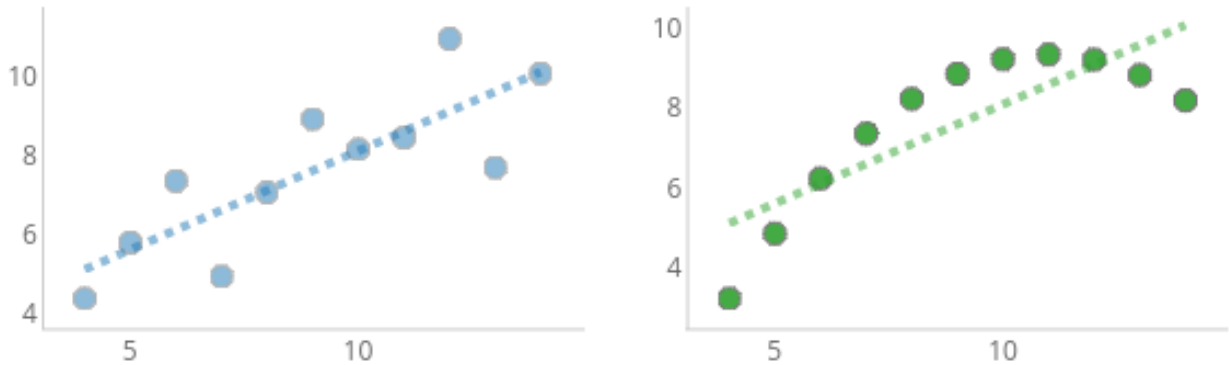
This week's content

- Tuesday: Correlation continued
- Regression diagnostics
- Remedial measures
- Reference: Simon Sheather §§3.1 & 3.2

So we have a fit, but...

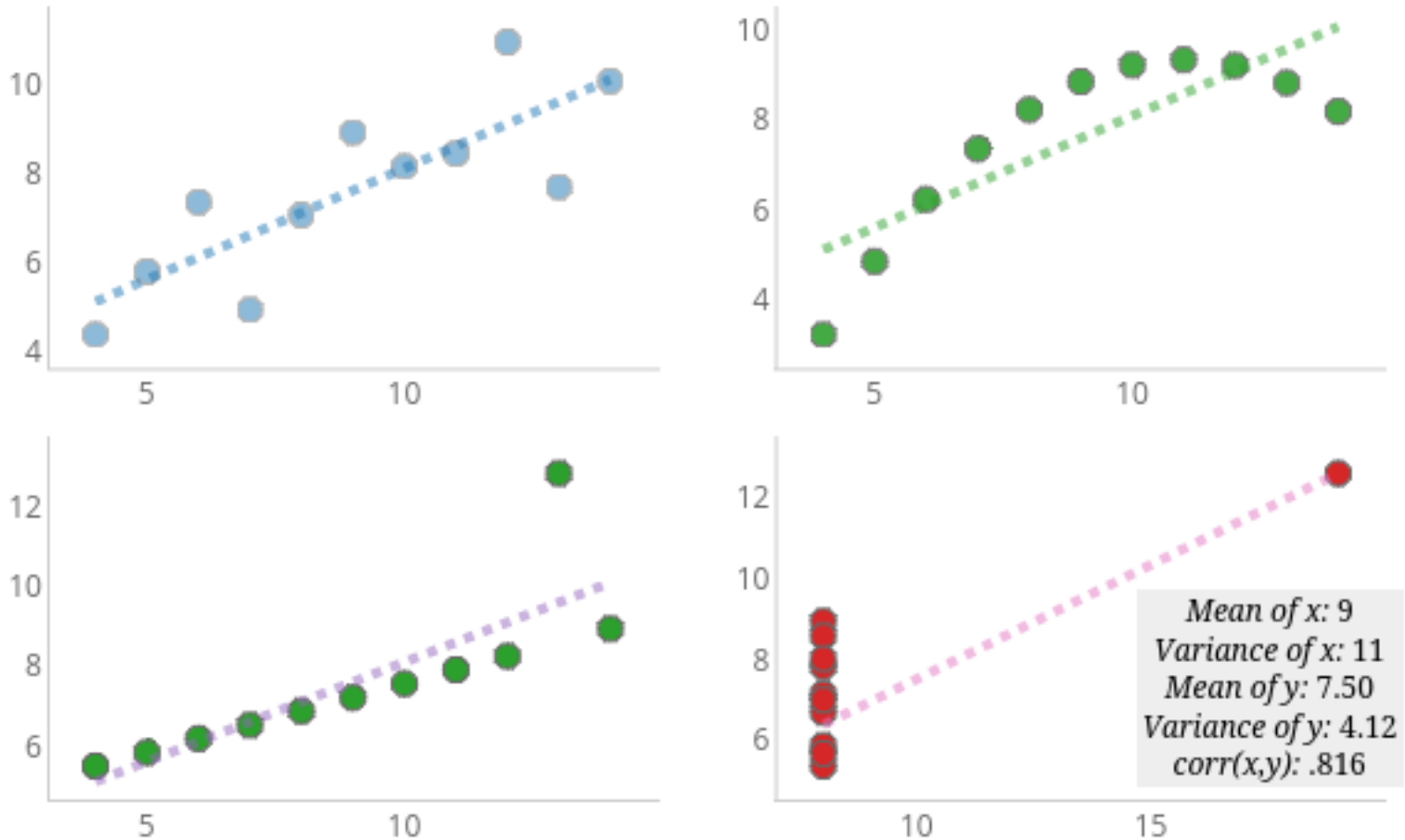
Is our linear model appropriate for the data?

$$E(Y|X = x) = \beta_0 + \beta_1 x \quad \text{and} \quad \text{var}(Y|X = x) = \sigma^2$$



We can use *regression diagnostics* to check our assumptions.

Francis Anscombe's Quartet (1973)



The statistics in the lower-right graph apply to all four graphs. Also $\hat{y} = 3 + 5x$ throughout.

Taking stock: Our assumptions so far

1. That the right form of the model is: $Y_i = \beta_0 + \beta_1 X_i + e_i$, $i \in \{1, \ldots n\}$. And we fit a least squares line: $y_i = b_0 + b_1 x_i$.

2. Gauss-Markov conditions:
   - $E(e_i) = 0$
   - $\text{var}(e_i) = \sigma^2 \ \forall i$
   - e_i, e_j uncorrelated, $i \neq j$

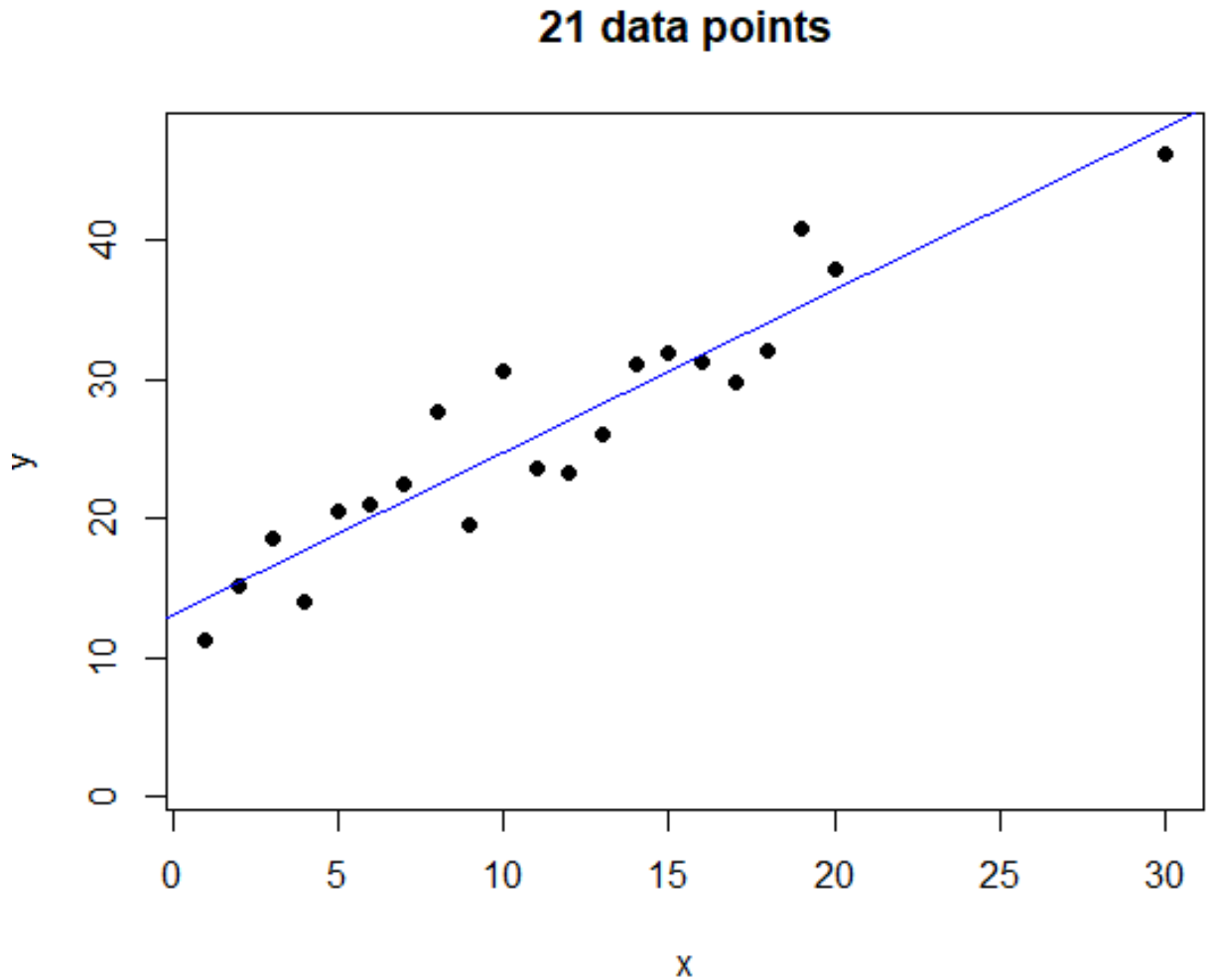
3. The e_i's are normally distributed

Regression diagnostics: The Seven C's (7 Checks)

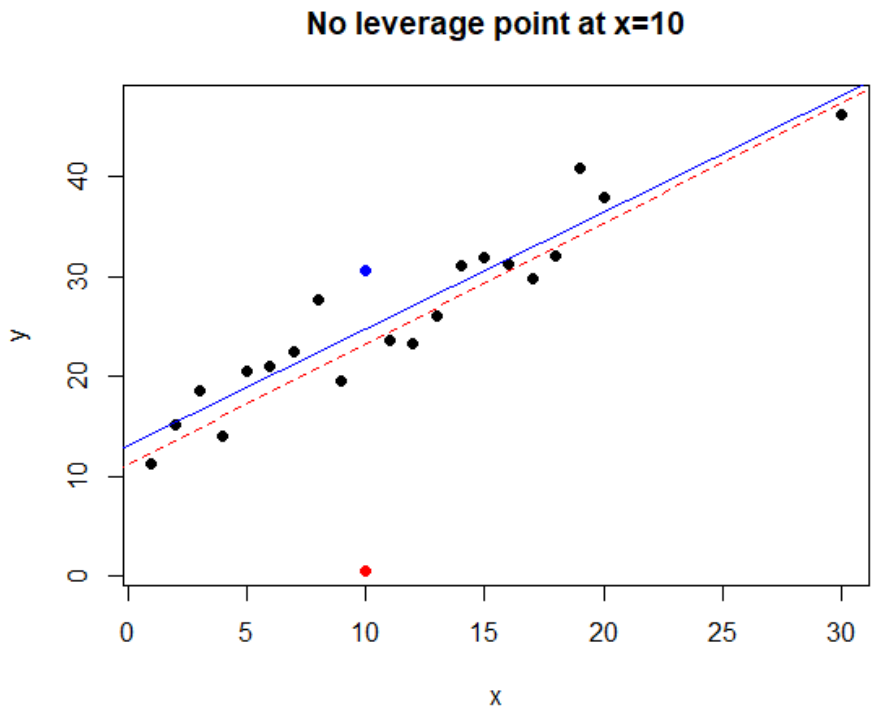
1. Plot *standardized residuals* to help determine whether the proposed regression model is a valid model
2. Identify any *leverage points*
3. Identify any *outliers*
4. Identify any *influential points*
5. Assess the assumption of *error homoscedasticity*
6. For time series: examine whether the data are *correlated over time*
7. Assess the assumption of *normal errors*

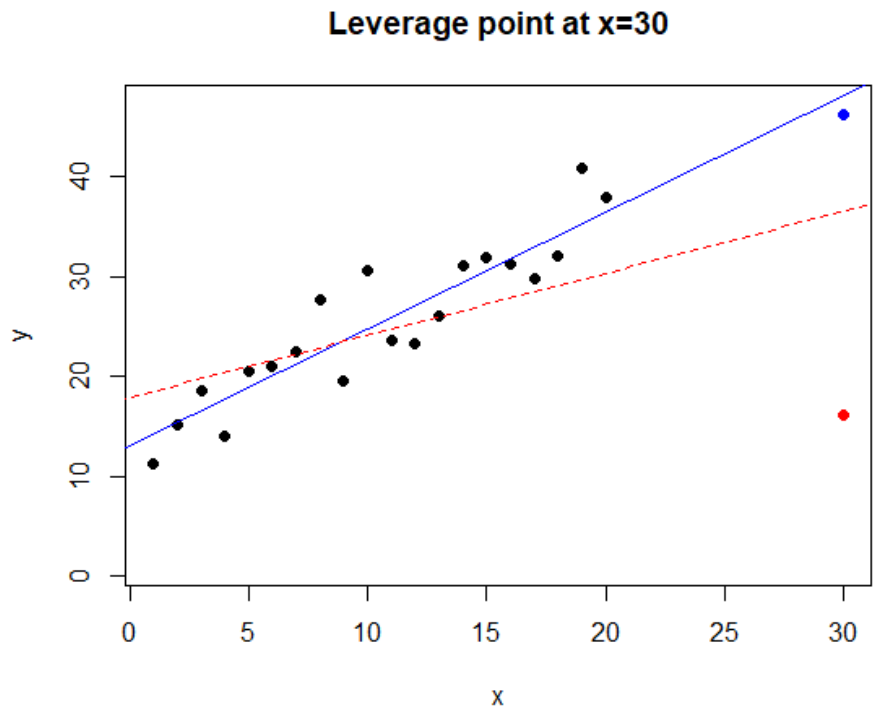
Check 2: Leverage Points

A leverage point is a data point which plays an important role when fitting the model, by having an x-value distant from the other x-values.

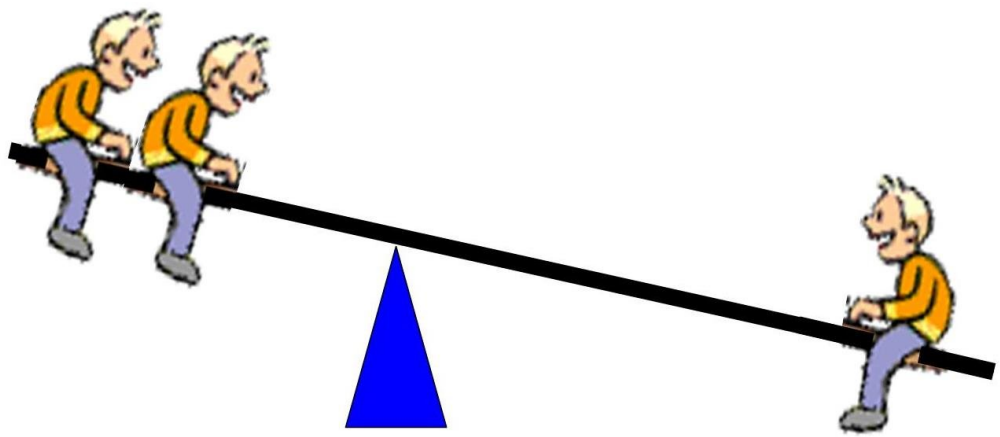


Leverage Points





The seesaw is an example of a first-class lever



Expressing $\hat{y}$ in terms of y

$$
\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \\
&= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\
&= \frac{1}{n} \sum_{j=1}^{n} y_j + \sum_{j=1}^{n} \frac{x_j - \bar{x}}{S_{xx}} y_j (x_i - \bar{x}) \\
&= \sum_{j=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right] y_j
\end{aligned}
$$

and we have expressed $\hat{y}$ in terms of linear combinations of y

What do the coefficients in the linear combination represent?

$$\hat{y}_i = \sum_{j=1}^{n} \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right] y_j = \sum_{j=1}^{n} h_{ij}\, y_j$$

where we have renamed the quantity in [ ] as h_{ij}. Key resulting equations are

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \qquad \text{and} \qquad \hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

We call h_{ii} the **leverage** of the ith data point. A **leverage point** is a data point with high leverage.

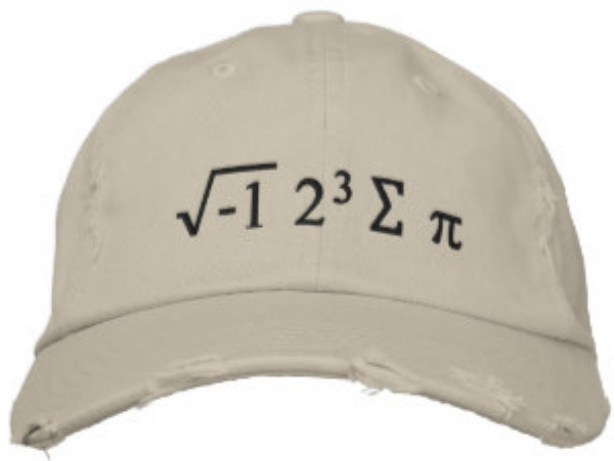
Note that in the extreme, as $h_{ii} \to 1$, the other h_{ij} terms tend to zero, so $\hat{y}_i \approx (1) y_i + \text{small terms} \approx y_i$. So for a leverage point, the line of best fit will fall close to/be attracted to the actual value, y_i.

Why h?

h stands for "hat". Consider:

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

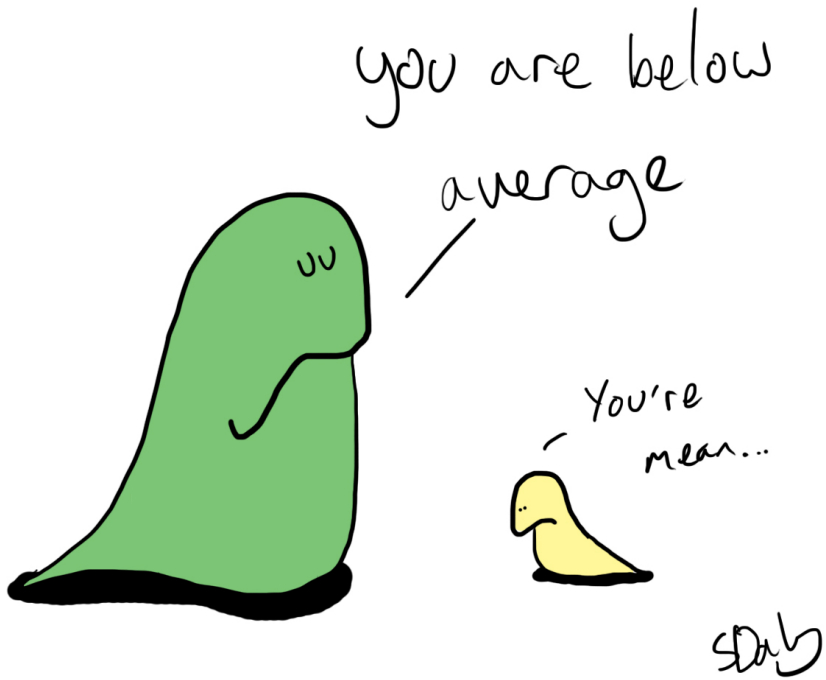
The h values are how to get from y_i's to $\hat{y}_i$'s.



Leverage points defined with maths

In practice, we don't require $h_{ii} \to 1$. Less stringently, we say a leverage point has $h_{ii} > 4/n$. The rationale is that this value of leverage is twice that of the average: $2/n$.

Points that are less than twice the average cannot count as leverage points.



Leverage exercises

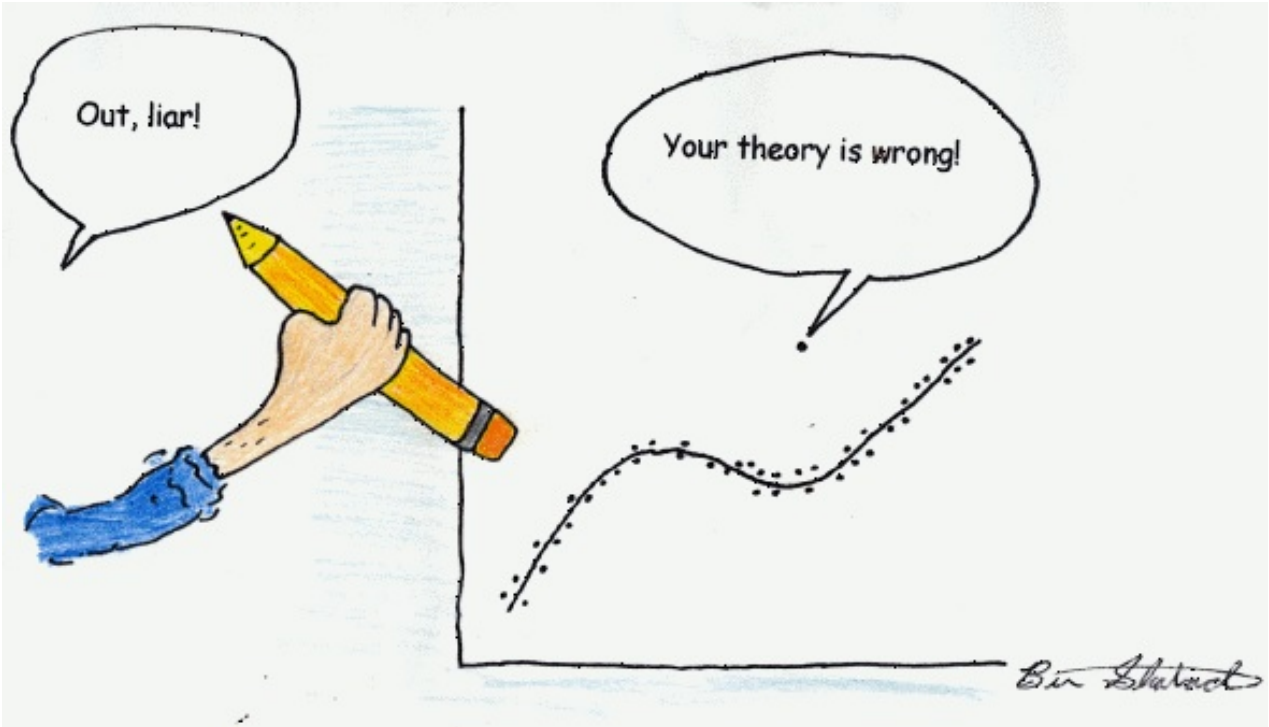
- Show why the average leverage is $2/n$
- Show $\sum_{j=1}^{n} h_{ij} = 1$
- Show $h_{ij} = h_{ji}$
- Show $\sum_j h_{ij}^2 = h_{ii}$

Regression diagnostics: The Seven C's (7 Checks)

1. Plot *standardized residuals* to help determine whether the proposed regression model is a valid model
2. Identify any *leverage points*
3. **Identify any *outliers***
4. Identify any *influential points*
5. Assess the assumption of *error homoscedasticity*
6. For time series: examine whether the data are *correlated over time*
7. Assess the assumption of *normal errors*

Check 3: Outliers

Outliers have y_i values distant from the other y_i values in some way — e.g. from what you would expect.



We'll investigate quantitative methods to identify outliers soon.

Check 4: Influential Points

Observations whose inclusion or exclusion result in substantial changes to the fitted model (estimate of coefficients, or predicted values) are said to be **influential points**.

Points can be "outlying" in any (or all or some) of the value of the:

- Explanatory variable (Check 2)
- Dependent variable (Check 3)
- Residuals (Check 1)

The combination of the first two creates an influential point. Influence can be thought of as the product of leverage and outlier-ness.

N.B. cases outlying w.r.t. the *residuals* have something in common with influential points: they represent model failure. The line doesn't fit those points adequately.

Ways to handle Influential Points

1. **Remove them**

   - But first, investigate where such points came from. Is the source of the data points unusual or different in some way from that of the rest of the data? (e.g. a different kind of financial product)
   - Only remove them if invalid; i.e. there's a good reason to say that they are unlike other points

2. **Pick a different model**

   - Maybe an incorrect theory has been applied
   - e.g. maybe a curvilinear model would be more appropriate (e.g. include extra predictor variables as polynomial terms), or you should apply a transformation (to be handled in a future lecture)
   - Different fitting methods can be used to reduce the impact of outliers: namely, robust regression

One way to measure the influence of the ith observation

DFBETA:

- Defined as the standardized change in the regression coefficient j when point i is removed
- Therefore, DFBETA is a measure of the influence of a point on a regression coefficient
- To calculate it, you take the difference in the estimate of a β with and without the ith observation:

$$\text{DFBETA}_{ik} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\text{s.e. of } \hat{\beta}_{k(i)}}$$

where $i \in \{1, \ldots n\}, \quad k \in \{0, 1\}$ for intercept and slope.

Notation: subscript (i) indicates the ith observation has been deleted from calculation. For example, $\beta_{1(i)}$ is the estimate of the slope for $n-1$ observations, with the ith observation removed. (We don't often bother with the intercept, $k = 0$.)

Another way to measure the influence of the ith observation

DFFITS

- This represents the studentized difference in fits for $\hat{y}$; it describes how the ith predicted value changes with and without the inclusion of the ith observation
- Therefore, DFFITS is a measure of the influence of a point on a fitted value

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\text{s.e. of } \hat{y}_{i(i)}}$$

Using notation similar to that on the previous slide, $\hat{y}_{i(i)}$ is predicted at x_i calculated from $n-1$ observations (ith observation excluded).

Rules of thumb for DFBETA and DFFITS

DFBETA identifies an influential point when:

- $|\text{DFBETA}_{ik}| > 2/\sqrt{n}$ for large data sets, or
- $|\text{DFBETA}_{ik}| > 1$ for small data sets, or
- $|\text{DFBETA}_{ik}|$ is separated by a large gap from the other values of $|\text{DFBETA}|$

DFFITS identifies an influential point when:

- $|\text{DFFITS}_i| > 2\sqrt{2/n}$ for large data sets, or
- $|\text{DFFITS}_i| > 1$ for small data sets, or
- $|\text{DFFITS}_i|$ is separated by a large gap from the other values of $|\text{DFFITS}|$

We'll return to influential points later.

Check 1: Standardized Residuals...

Did you know? The variance of the ith residual is actually $\text{var}(\hat{e}_i) = \sigma^2[1 - h_{ii}]$

And $\text{var}(\hat{y}_i) = \sigma^2 h_{ii}$.

Therefore, rather than handle raw residuals, it can be more informative to consider **standardized residuals**. The ith standardized residual is

$$\boxed{r_i = \frac{\hat{e}_i}{S\sqrt{1 - h_{ii}}}}$$

where $S = \sqrt{\text{RSS}/(n-2)}$ is our usual estimate of σ.

It can also be called a studentized residual, but our textbook is relaxed about this.

Note that the above is "internal studentization". Sometimes you see external studentization:

$$r_i = \frac{\hat{e}_i}{S_{(i)}\sqrt{1 - h_{ii}}}$$

where $S_{(i)}$ is the s.e. of $\hat{e}_i$ with the ith observation removed.

Proofs regarding the variances of the residuals

$\text{var}(\hat{e}_i) = \sigma^2[1 - h_{ii}]$:

$\text{var}(\hat{y}_i) = \sigma^2 h_{ii}$:

Two advantages of standardized residuals

1. Plots of standardized residuals are similar to plots of residuals, when leverage points don't exist. When leverage points do exist, plots of residuals will have nonconstant variance (even if the errors have constant variance) and thus plots of standardized residuals are more informative.
2. They tell us immediately how many estimated s.d.'s any point is away from the fitted regression model. To be an outlier, this should be >2 s.d.'s. (You can increase this number for large datasets.)

Return to Check 4, using standardized residuals

Recall the two measures of influence of the ith observation. A third is:

Cook's distance:

$$D_i = \frac{\sum_{j=1}(\hat{y}_{j(i)} - \hat{y}_j)^2}{2S^2} \qquad \text{where } S^2 = \text{MSE}$$

It can be shown that

$$D_i = \frac{r_i^2 h_{ii}}{2(1 - h_{ii})}$$

where r_i is the ith standardized residual.

Large D_i occurs when there is a large residual and high leverage.

Cook's distance refers to how far, on average, predicted $\hat{y}$-values would move if the observation in question were dropped from the data set.

Rules of thumb

Cook's distance identifies an influential point when:

- $D_i > 4/n$, or
- D_i is separated by a large gap from the other D_k's

Notes:

- Some texts tell you that points for which Cook's distance is higher than 1 are to be considered as influential
- Other texts give you a threshold of $4/(n-k-1)$ where n is the sample size and k is the number of predictor variables. e.g. $4/(n-2)$ for SLR.

Checks 1, 2, 3 $\rightarrow$ Check 4

	$h_{ii} \leq 4/n$	$h_{ii} > 4/n$
$\lvert r_i \rvert \leq 2$	(Most points)	Good leverage point
$\lvert r_i \rvert > 2$	Outlier	Influential point

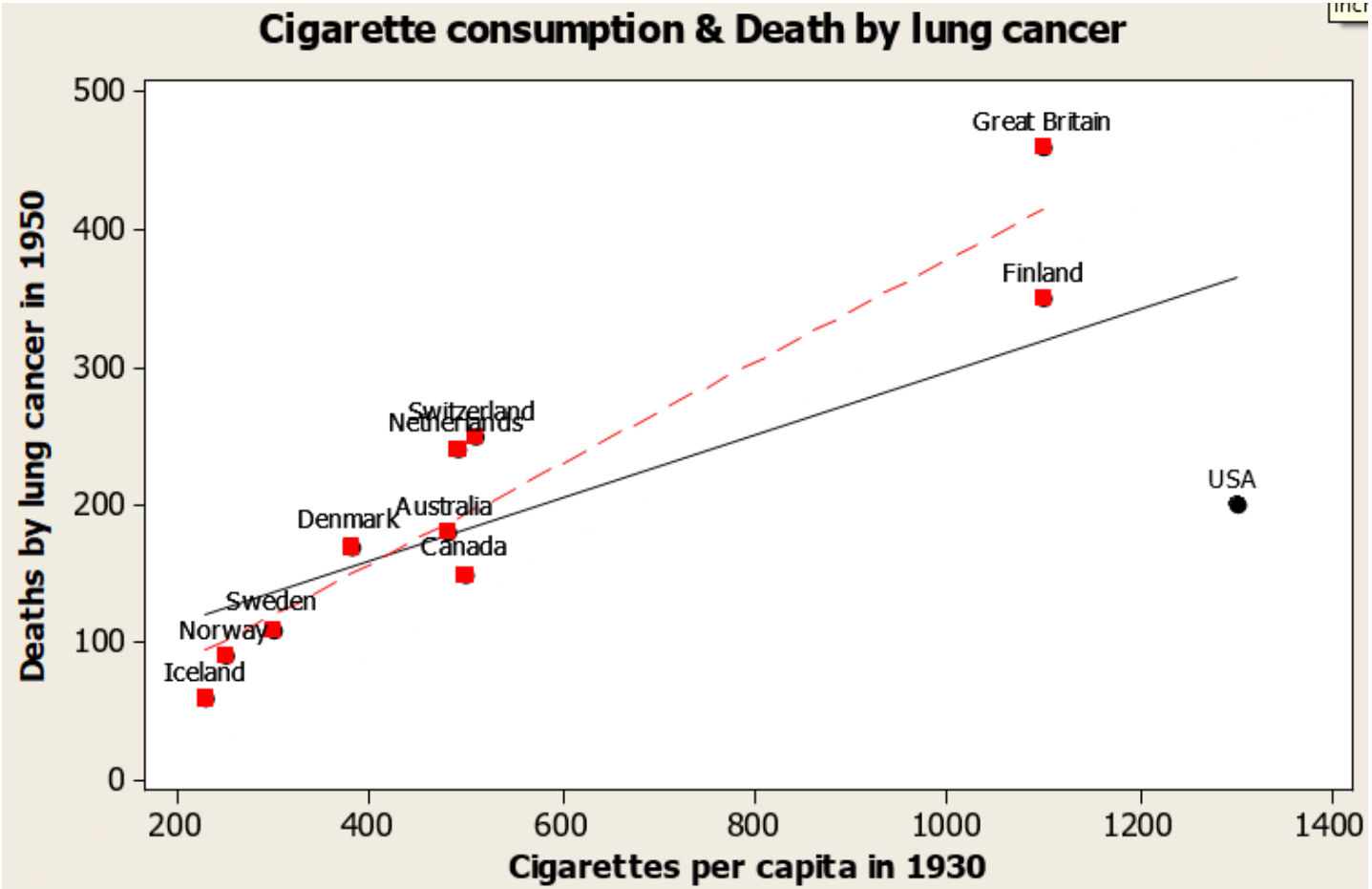
Recall that influential points are outliers and leverage points simultaneously. Simon Sheather refers to them as "bad leverage points".

Good leverage points aren't exactly angels. They make you think you have a better fit than you may actually do, by:

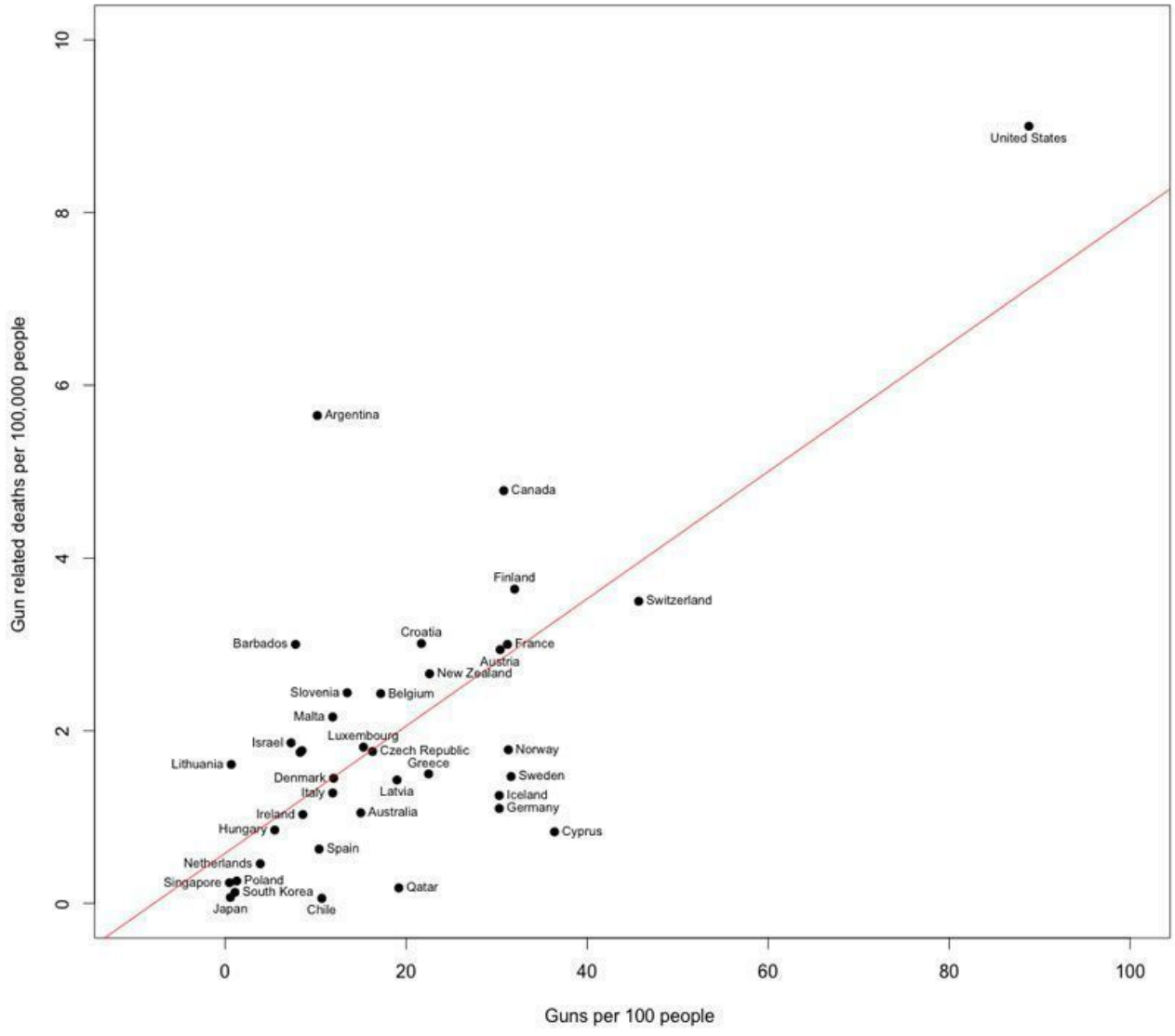
- Decreasing the s.e. of the estimate of β's, because they increase S_{xx}
- Increasing R^2

Therefore, scrutinize good leverage points: are we confident enough to keep them? Do we need a new model?

Example of a "bad" leverage point



Example of a "good" leverage point



Summary of Influence Metrics

In addition to the 2×2 matrix on slide 27, we've considered these metrics:

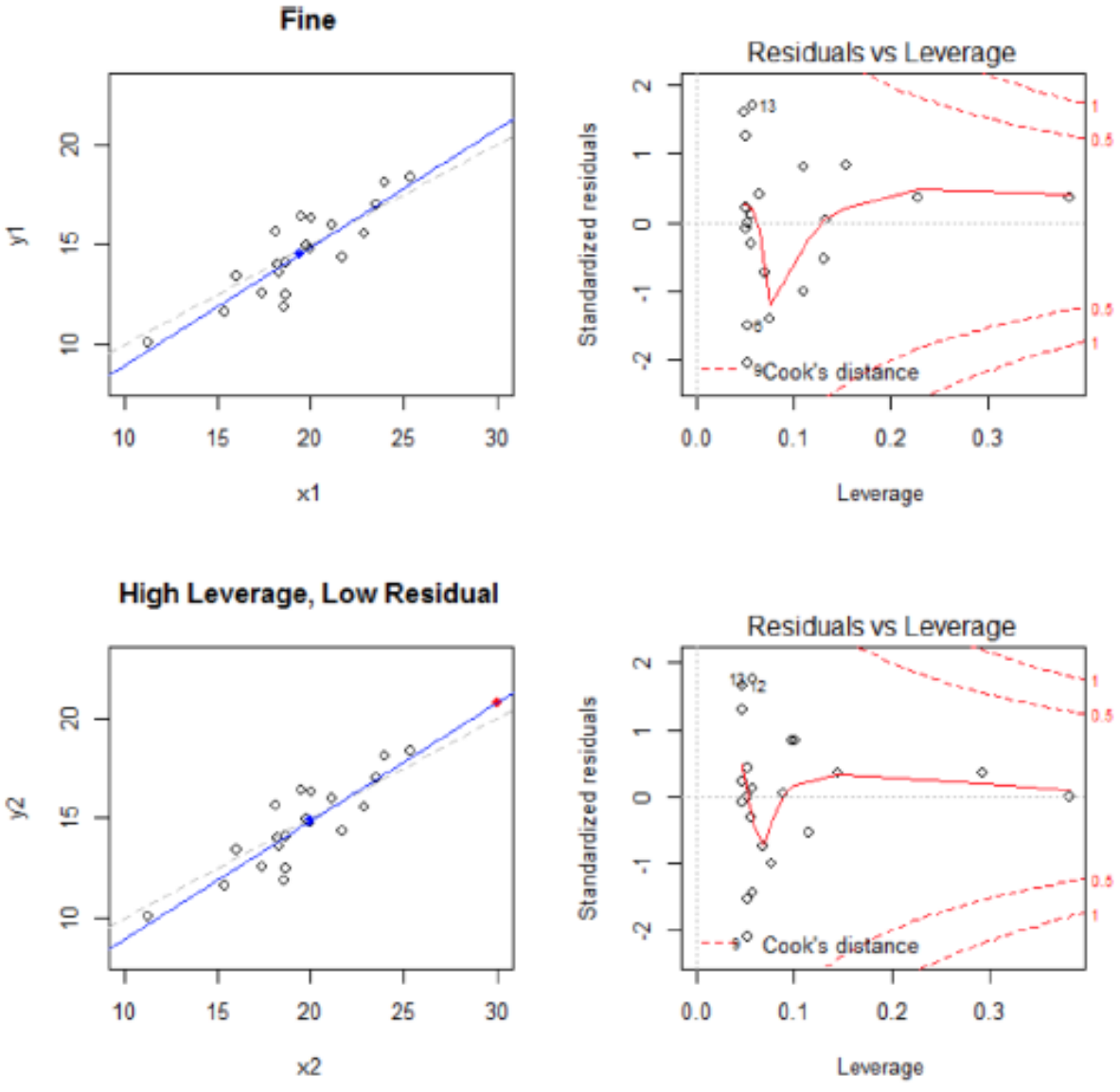
DFBETA: Measure of the influence of a point on a regression coefficient.

DFFITS: Measure of the influence of a point on a fitted value.

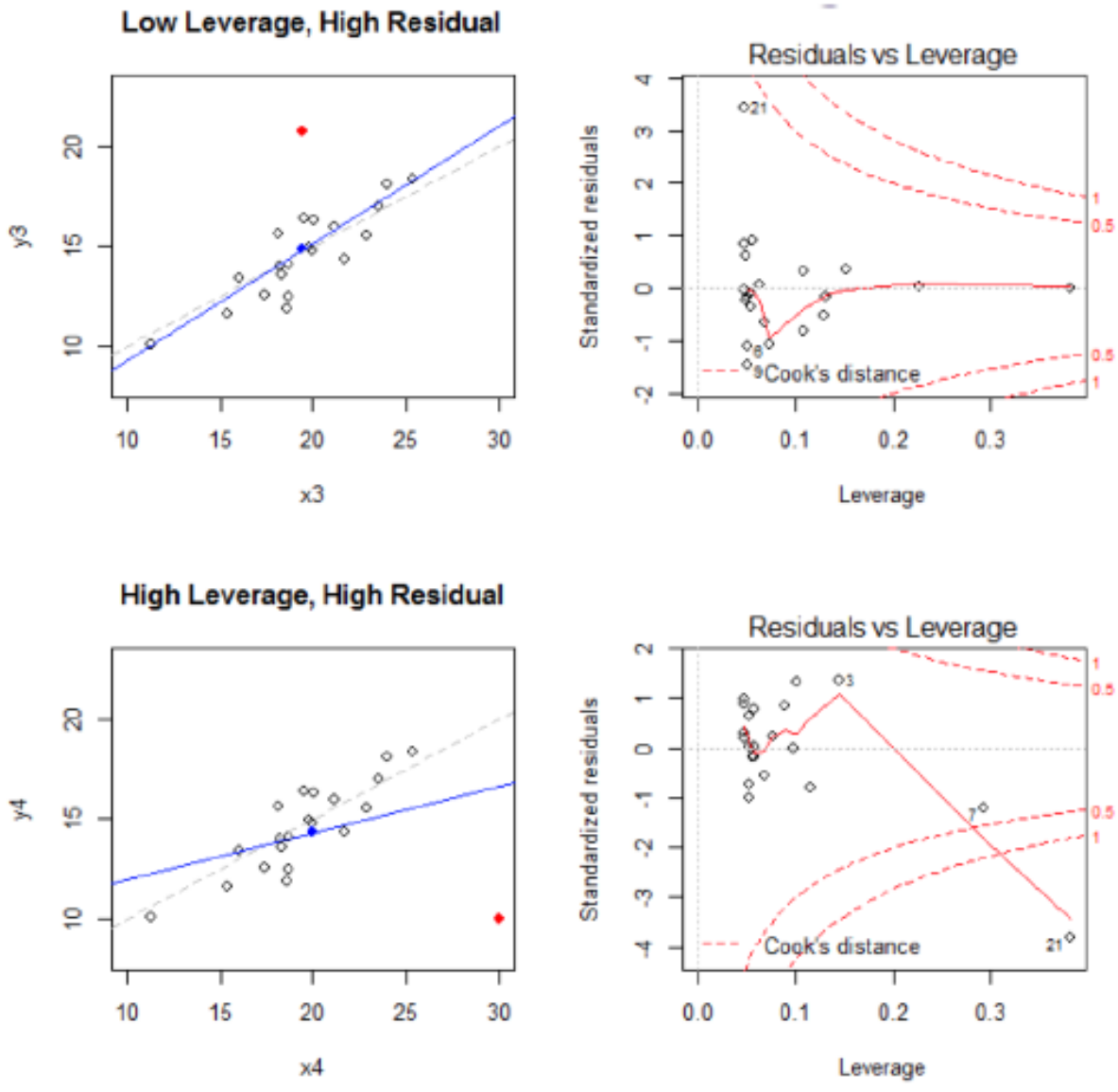
Cook's Distance: Measure of the influence of a point on fitted values. (Or, a combined measure of the leverage and outlier magnitude of a point.)

- Cook's distance is probably more important if you're doing *predictive* modelling, whereas DFBETA may become more important in *explanatory* modelling

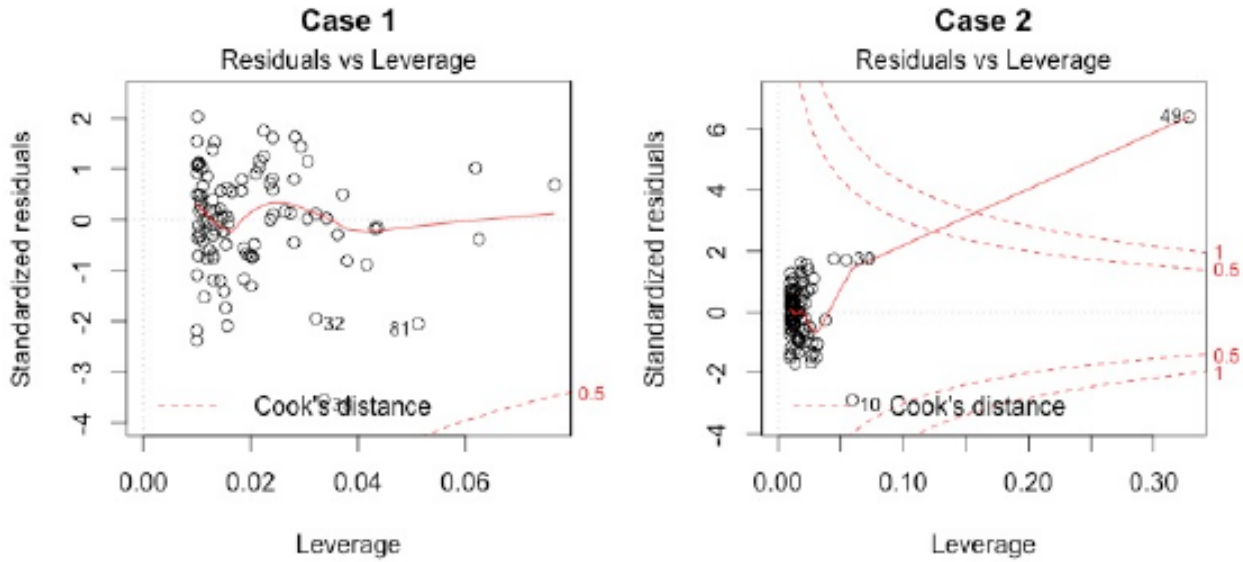
Visualization for speedy identification of influential points



Visualization for speedy identification of influential points



In a nutshell: What we're looking for in the Cook's distance plots

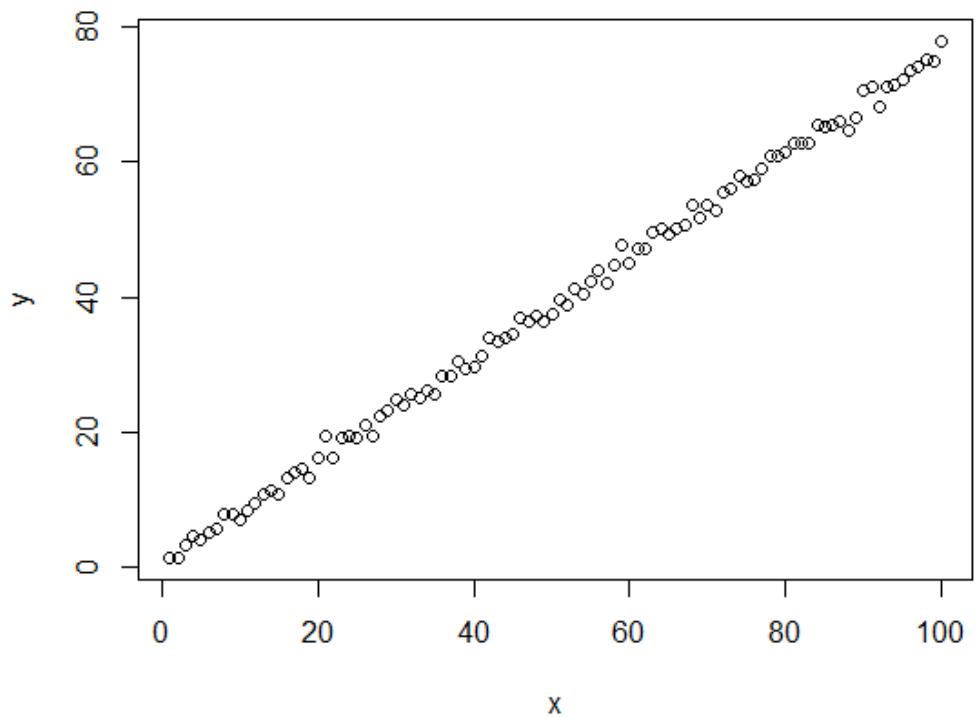


- We seek influential data points/subjects/observations/cases
- These are any points outside of the dashed line representing Cook's distance (i.e. points in the upper-right corner or lower-right corner, and beyond the dashed line). The regression results would be altered if we excluded those cases
- Case 1 (left) is typical: no influential points
- Case 2 (right) shows that subject #49 gives an influential point

Does the Cook's Distance graph require lots of programming?

It takes just one or two lines of R code. First let's create a fresh dataset:

```
set.seed(1000)
x = 1:100; y=1+0.75*x+rnorm(100,0,1) # generate x and y
plot(x,y)
```

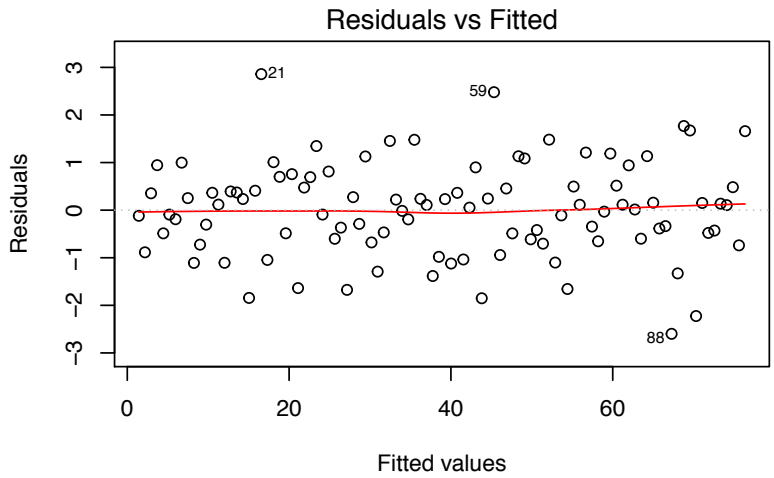


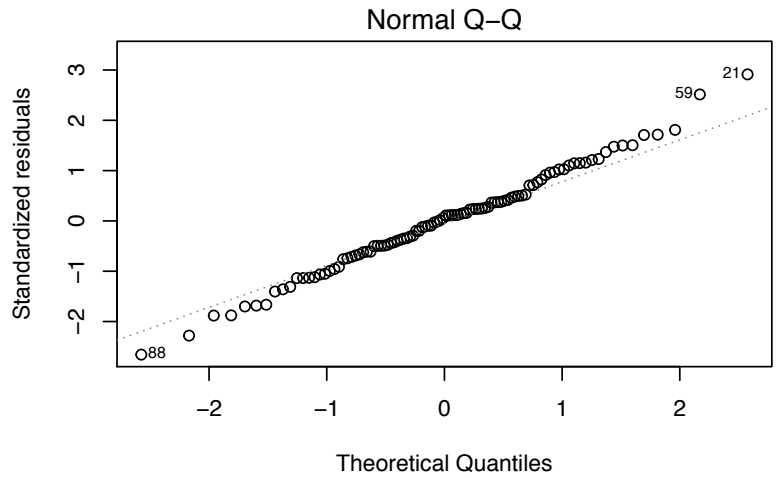
R code for Diagnostics Visualizations

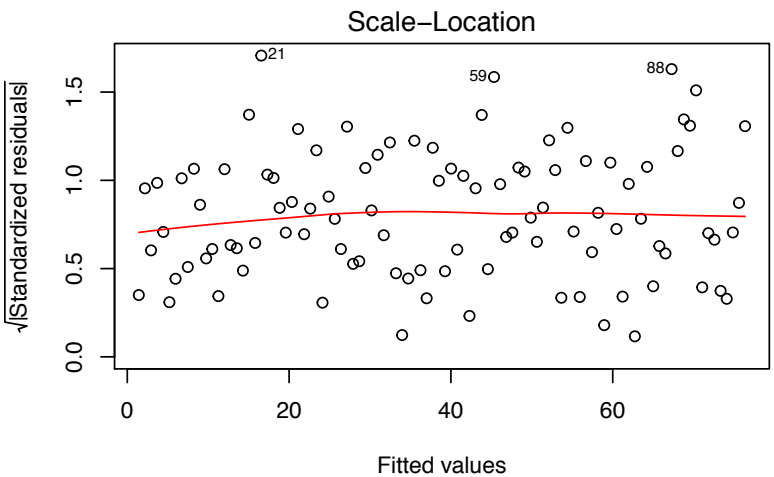
On the next slide, we'll run:

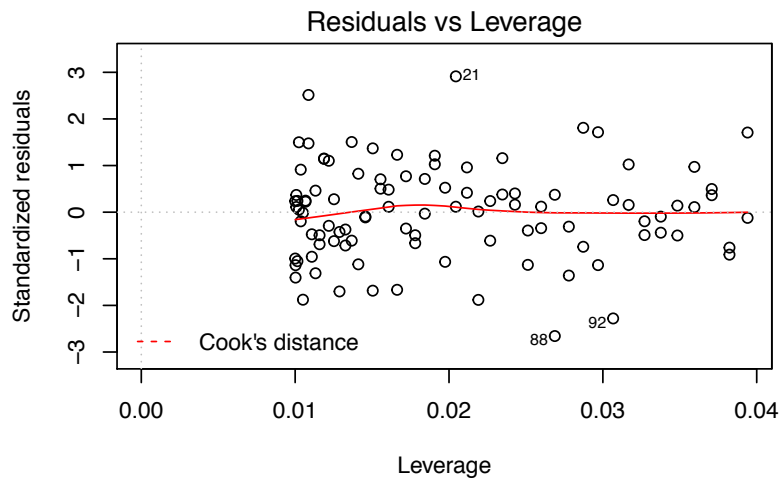
```
par(mfrow=c(2,2)) # split the panel as 2 by 2
plot(lm(y~x))
```

Simulation study









R code for diagnostics

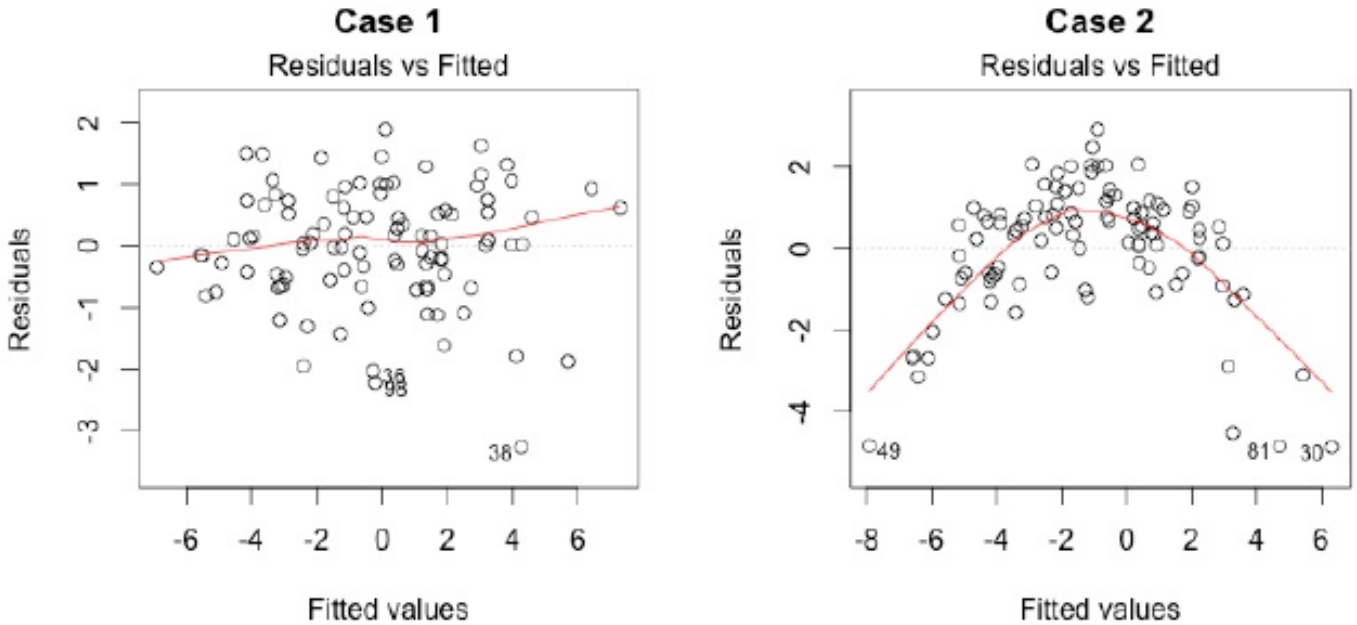
```
influence.measures(lm(y~x))$infmat[1:7,]
```

```
##         dfb.1_       dfb.x       dffit    cov.r       cook.d       hat
## 1 -0.02468749  0.02132701 -0.02468841 1.062272 0.0003078536 0.03940594
## 2 -0.18151974  0.15600696 -0.18154782 1.043384 0.0165085793 0.03822982
## 3  0.07106854 -0.06075510  0.07109404 1.057165 0.0025497843 0.03707771
## 4  0.18737375 -0.15929894  0.18749711 1.038503 0.0175878587 0.03594959
## 5 -0.09465330  0.08001053 -0.09475381 1.052175 0.0045238242 0.03484548
## 6 -0.01779918  0.01495622 -0.01782729 1.056196 0.0001605292 0.03376538
## 7 -0.03571379  0.02982398 -0.03579311 1.054417 0.0006469247 0.03270927
```

- `hat` refers to h_{ii}
- `cov.r` is beyond the scope of this course

What else do residuals tell us? (1)

They can help us check for linearity, when plotted versus $\hat{y}$



- No distinctive pattern in Case 1
- A clear pattern (a parabola) in Case 2. Nonlinearity exists

Ways to check for linearity

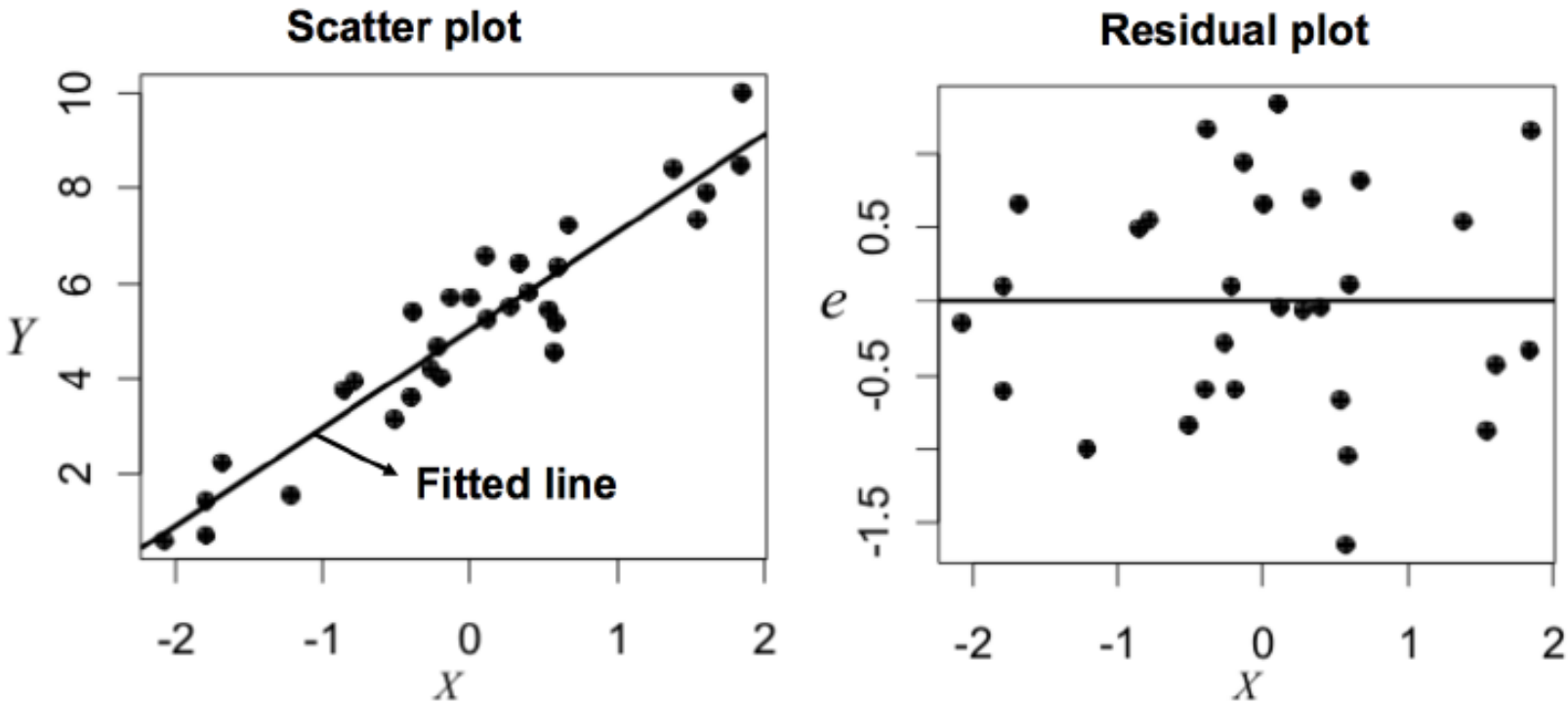
1. Plot residuals vs predictor variable ($\hat{e}_i$ vs X_i), or
2. Plot residuals vs fitted values ($\hat{e}_i$ vs $\hat{Y}_i$), or
3. A scatter plot (Y vs X) can also be used.

But residual plots are preferred because:

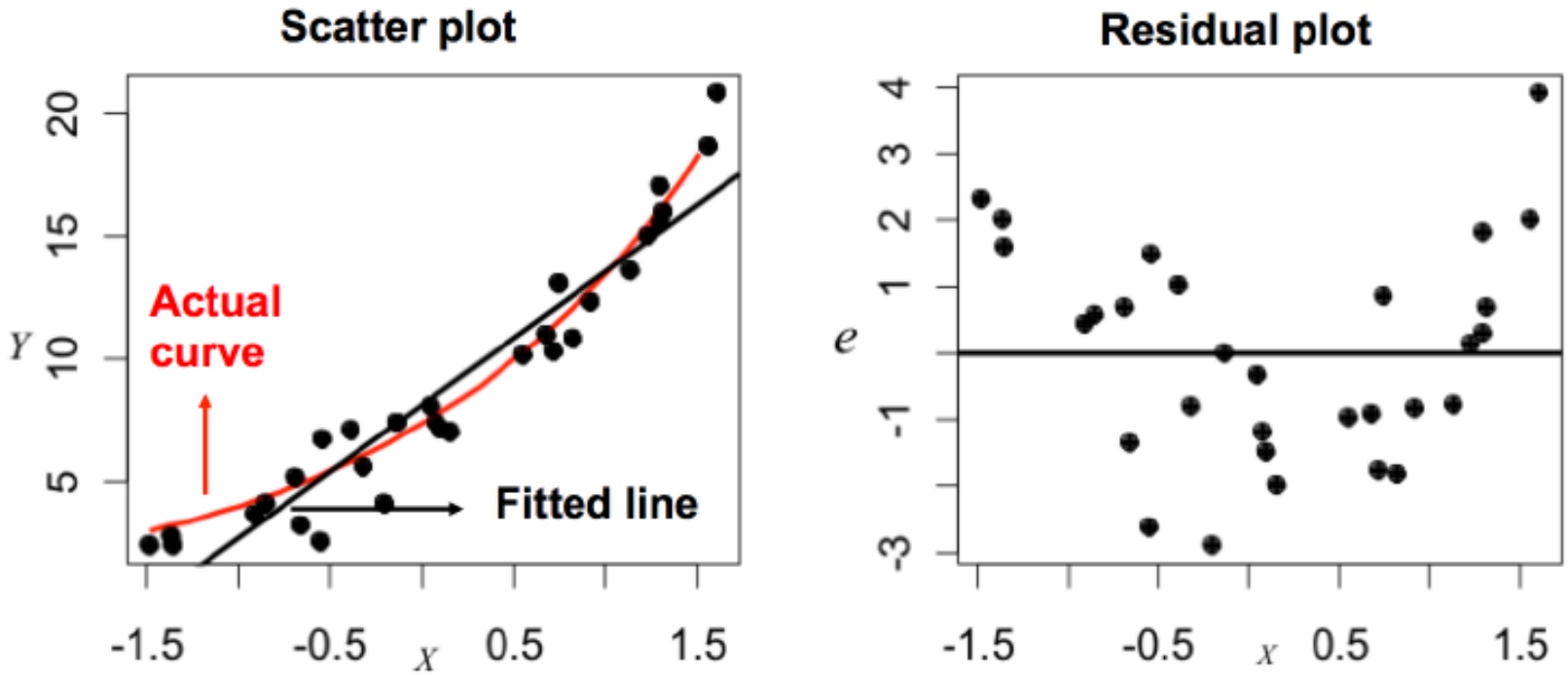
- Can spot nonlinearity more easily
- Can check other assumptions, e.g. constant variance etc
- Applicable to multiple linear regression (*)

(*) In SLR, $\hat{e}_i$ vs X_i gives information equivalent to $\hat{e}_i$ vs $\hat{Y}_i$. Not so in multiple linear regression.

Example of a linear relationship



Example of a nonlinear relationship



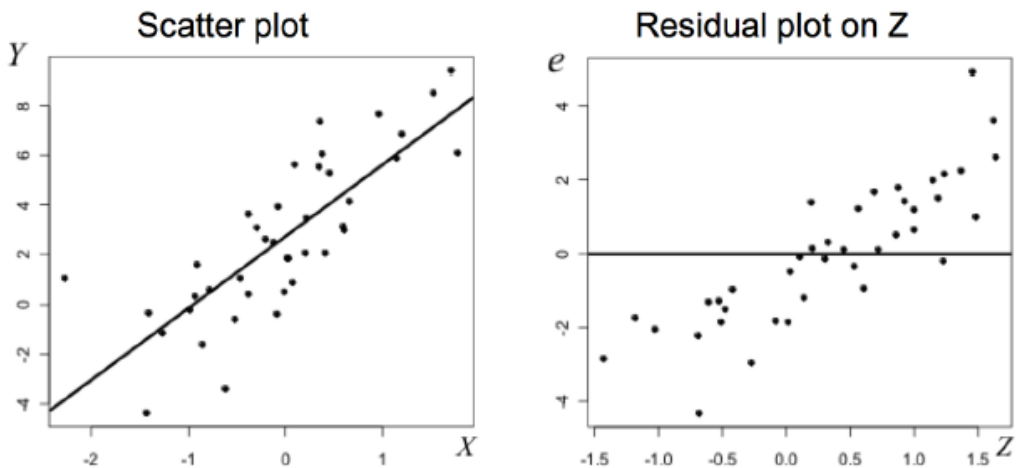
Remedial measure: If the regression function isn't linear,

- In some cases, a variable transformation can make the data "more linear"
- Otherwise, a different (e.g. nonlinear) model might be better

And what else do residuals tell us? (2)

Residuals can help us check whether important predictor variables are missing from our model. Method:

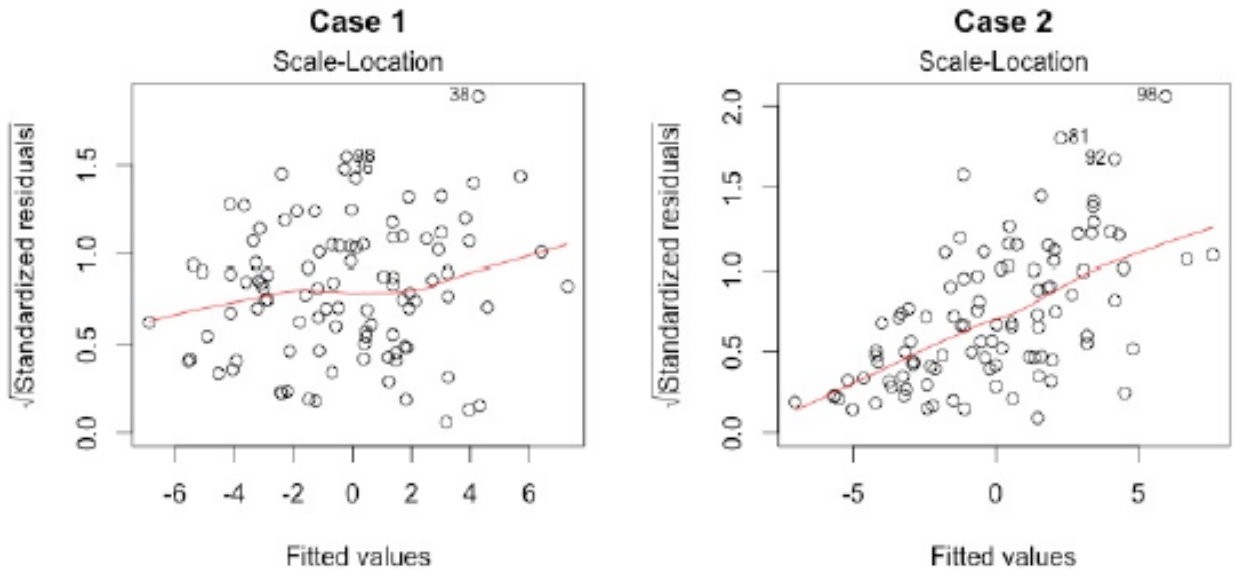
- Fit $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- For another potential predictor variable Z, plot $\hat{e}$ vs Z



Remedial measure: multiple linear regression.

And what else do (standardized) residuals tell us? (3)

They can help us check for homoscedasticity, through the **scale-location plot**, a.k.a. spread-location plot.

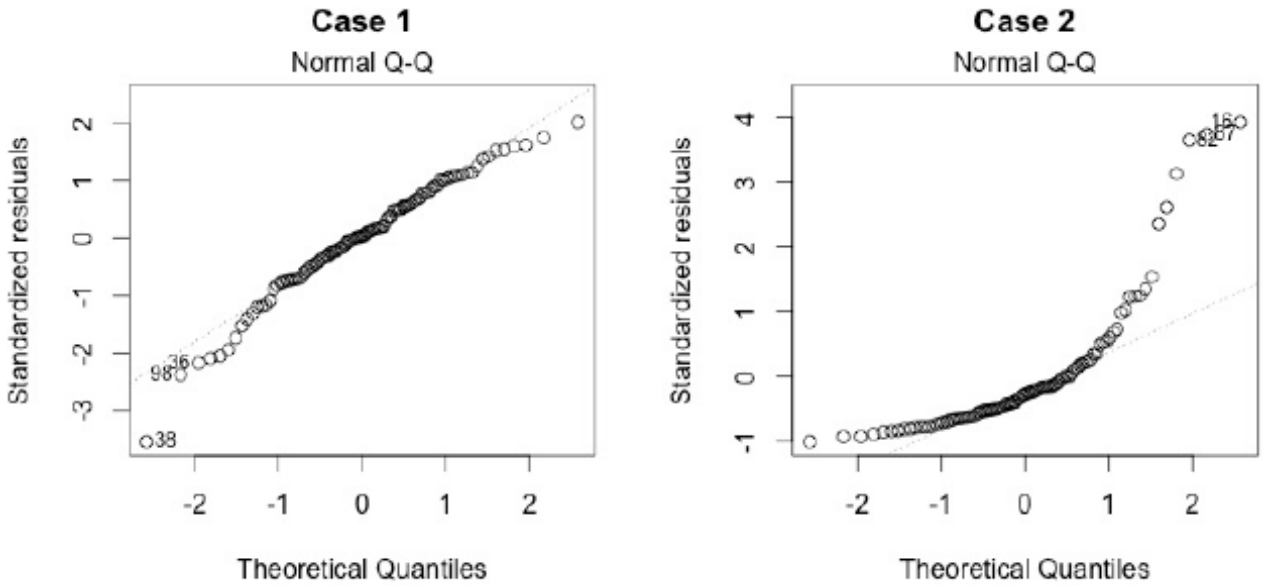


Are the residuals spread equally along the range of the predictor?

- In Case 1, the residuals appear to be randomly spread
- In Case 2, the residuals become more widely spread with x

And what else do (standardized) residuals tell us? (4)

They can help us check for normality



- In Case 1, most data points lie on a straight line and thus the normality assumption seems ok
- In Case 2, there is a heavy right tail. The data are right-skewed. (More on QQ plots soon)

Remedial measure: Often, variable transformations can help. If not, try modelling the error term with a different distribution.

Partial summary of residuals

In short, residuals can be used as a diagnostic tool to test the conditions of model that are necessary to make valid inferences.

$$\hat{e}_i = y_i - b_0 - b_1 x_i = y_i - \hat{y}_i$$

What else do we know about residuals?

- High leverage points have residuals with lower variance
- $\hat{e}_i$'s are *not* uncorrelated — we know $\sum \hat{e}_i = 0$. We can show $\text{cov}(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2$, $i \neq j$. But for large n we can ignore their correlation
- The $\hat{e}_i$'s are linear combinations of y_i's so they are normally distributed
- Their distribution can be closer to normal than that of the e_i's because of the CLT
- If there are no leverage points, i.e. h_{ii} is small, then the variances of $\hat{e}_i$'s are close — the lack of equal variance can be ignored
- Or, to adjust for unequal variance, standardize: $r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$

Summary of Residuals continued

We can use residuals to study the following types of departures from the SLR model

a) The regression function isn't linear

b) One or several important predictor variables have been omitted from the model

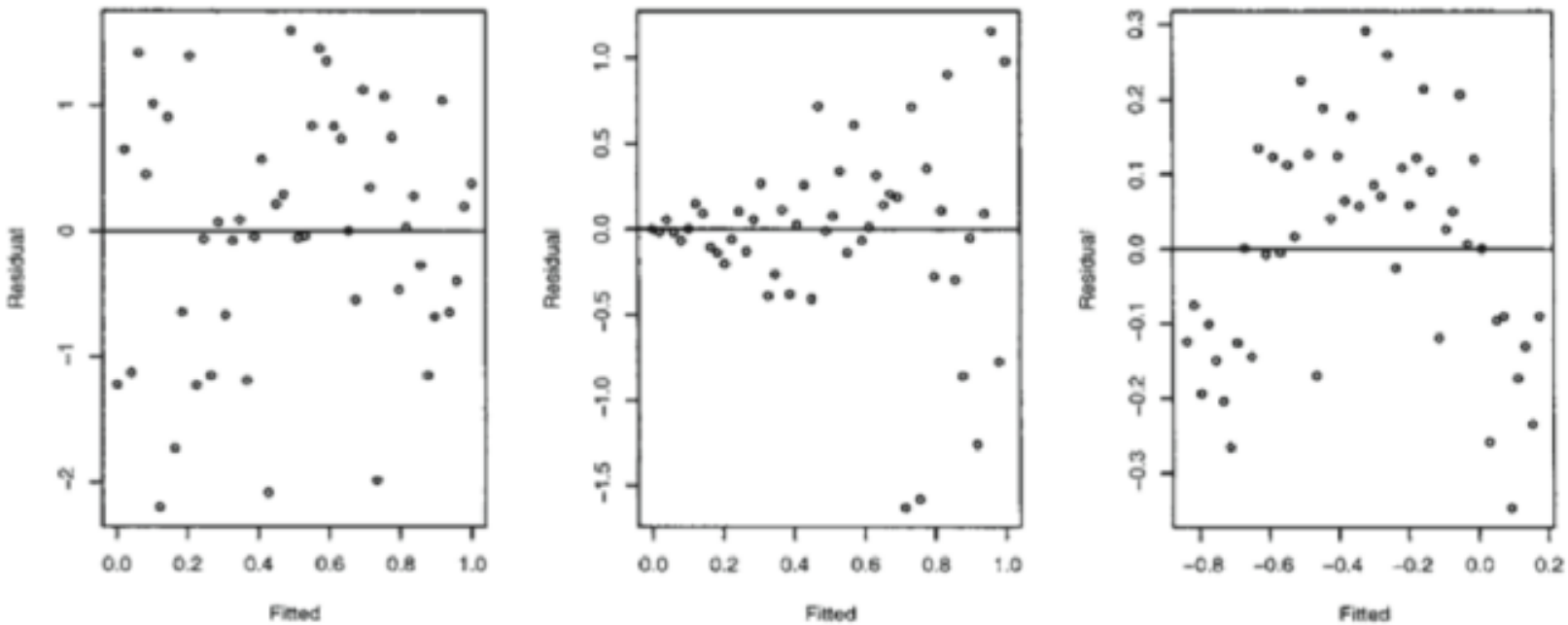
c) The model fits all points save a few outliers (Check 3)

d) The error terms don't have constant variance (Check 5)

e) The error terms aren't independent (Check 6, to come)

f) The error terms aren't normally distributed (Check 7, to come)

Three case studies: Plotting residuals versus $\hat{y}$

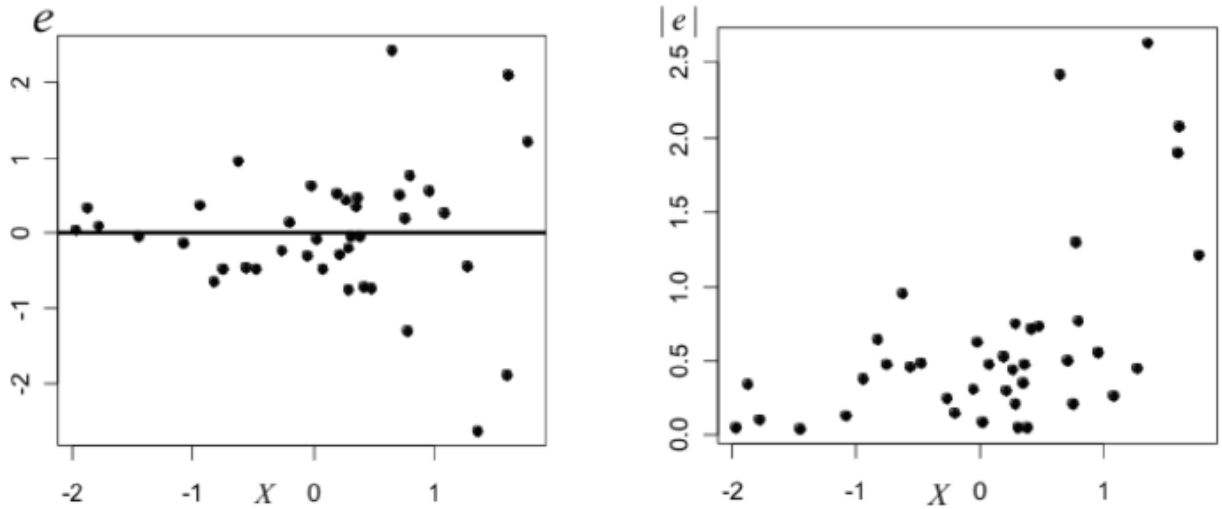


Question for you: What do the residuals tell us in each case?

Regression diagnostics: The Seven C's (7 Checks)

1. Plot *standardized residuals* to help determine whether the proposed regression model is a valid model
2. Identify any *leverage points*
3. Identify any *outliers*
4. Identify any *influential points*
5. **Assess the assumption of *error homoscedasticity***
6. For time series: examine whether the data are *correlated over time*
7. Assess the assumption of *normal errors*

Check 5: Constant variance

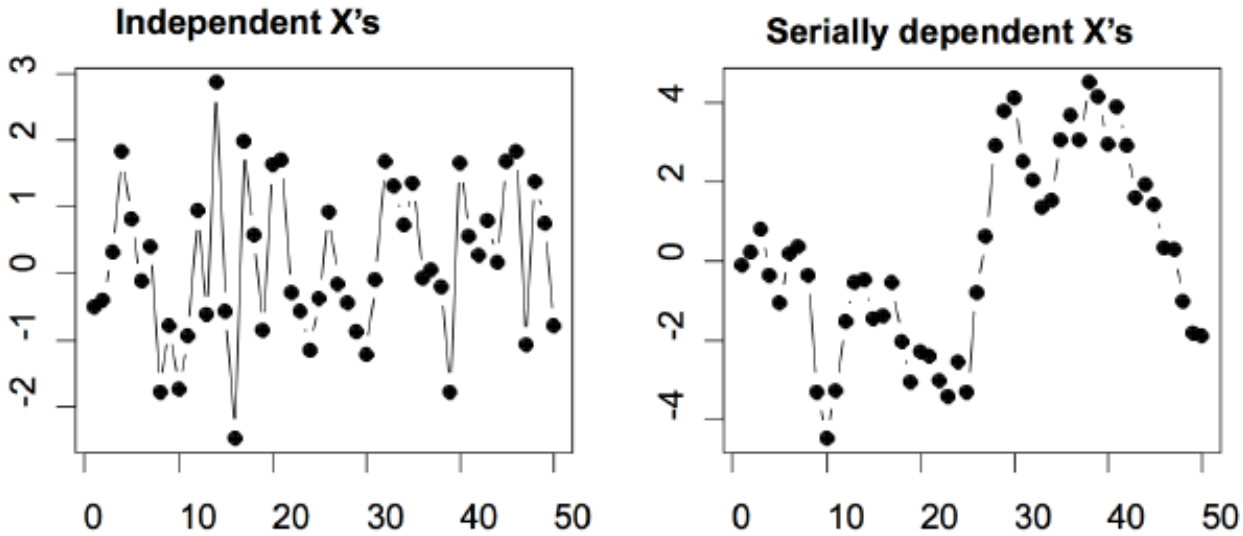


A residual plot, or absolute residual plot, should show points spread out evenly across X.

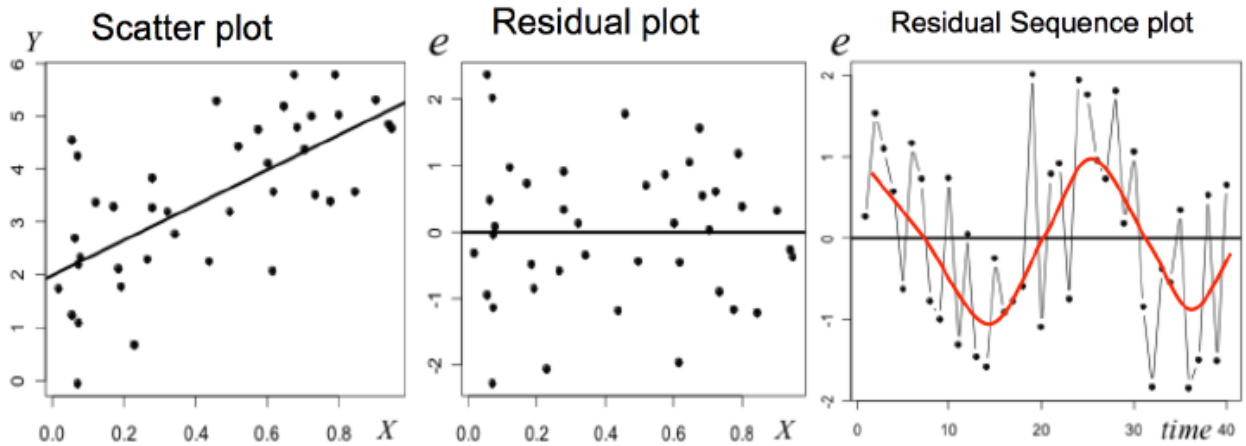
Remedial measure: if you can model variance as a function of X, a weighted regression can be used (chapter 4). Sometimes, a variable transformation can also stabilize variance (§3.3).

Check 6: Error independence

Here are two time series:



Checking for error independence



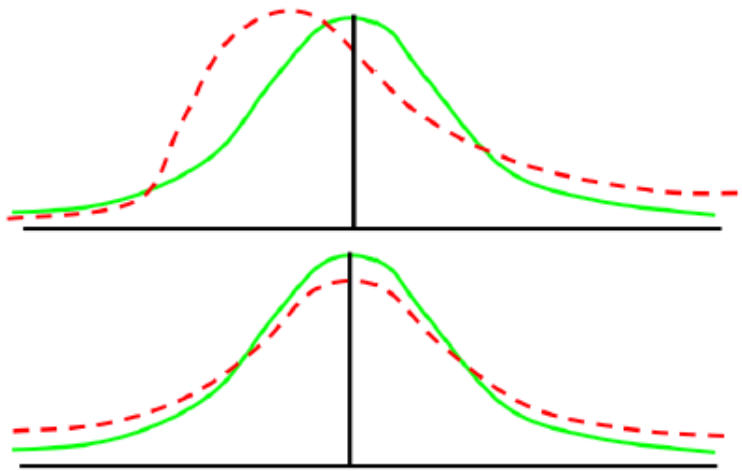
Use a residual sequence plot ($\hat{e}_i$ vs time or distance) to check for temporal- or spatial dependence

Remdial measure: If the dependence can be modelled (e.g. serially correlated e's), it can be incorporated into the regression.

Check 7: Normality of error

We look for two possible deviations from normality:

- Asymmetry
- Heavy tails



Testing for normality

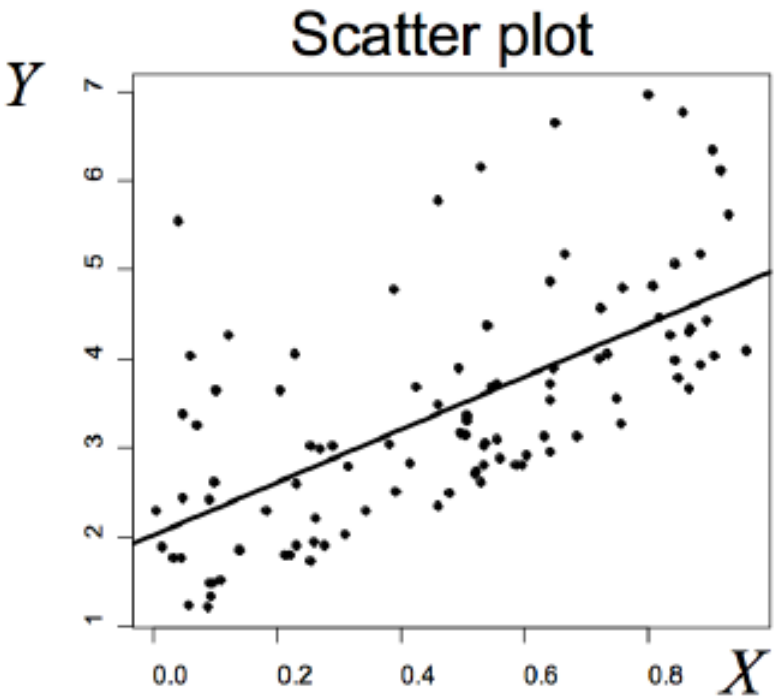
Normality is critical for inference but less important for fitting a regression line.

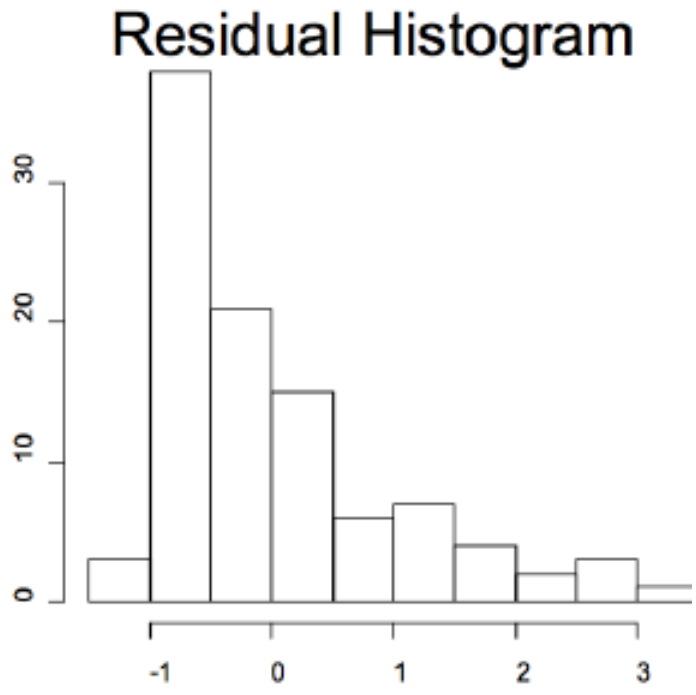
We should do Check 7 when there are small sample sizes (e.g. t-distribution based hypothesis testing and CIs), or if we want prediction intervals. Most SLR inference procedures assume the true errors are normally distributed.

One method to assess this is the *Shapiro-Wilk Test of Normality*: in R, use `shapiro-wilk()`, where H_0 : is normal.

Testing for normality

To just check symmetry (but not so much the heavy tails), you can make any plot describing the data distribution (box-plot, histogram, dot-plot...)





Normal Quantile-Quantile Plots

A better graphical tool (which checks for heavy tails as well) is the normal QQ plot. It compares the quantiles of a dataset to a set of theoretical quantiles from a normal probability distribution.

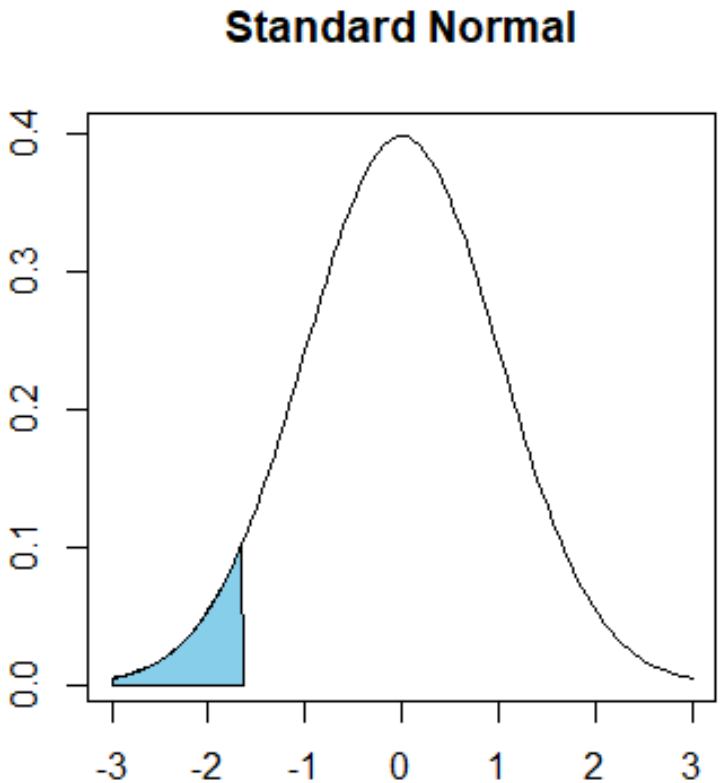
Example: The following data are the weights from 11 tomato plants.

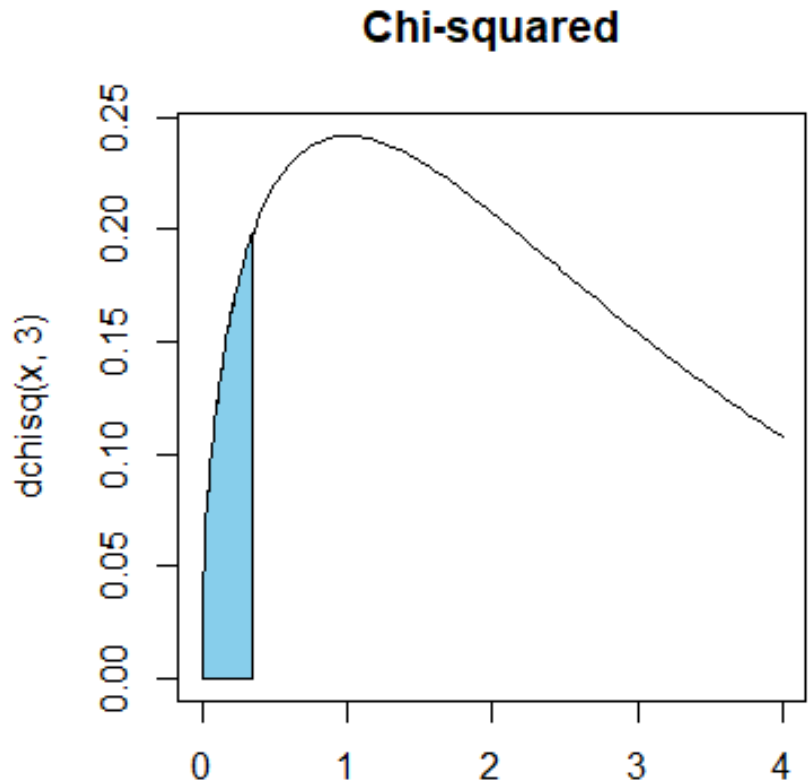
```
##  [1] 29.9 11.4 26.6 23.7 25.3 28.5 14.2 17.9 16.5 21.1 24.3
```

Do the weights follow a normal distribution?

Quantile-Quantile Plots

- In a QQ plot generally, the **quantiles** of any distribution are plotted against any other distribution (and if the dots make a straight line, the distributions are similar)
- Therefore, QQ plots can be used to investigate whether a set of numbers follows a certain distribution

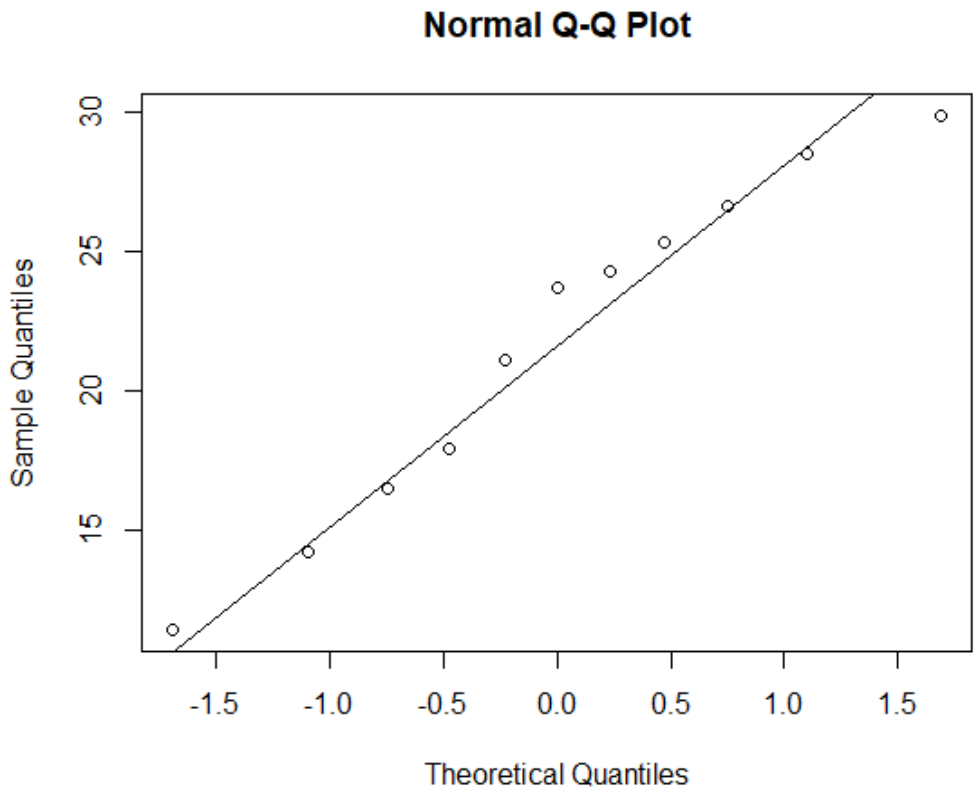




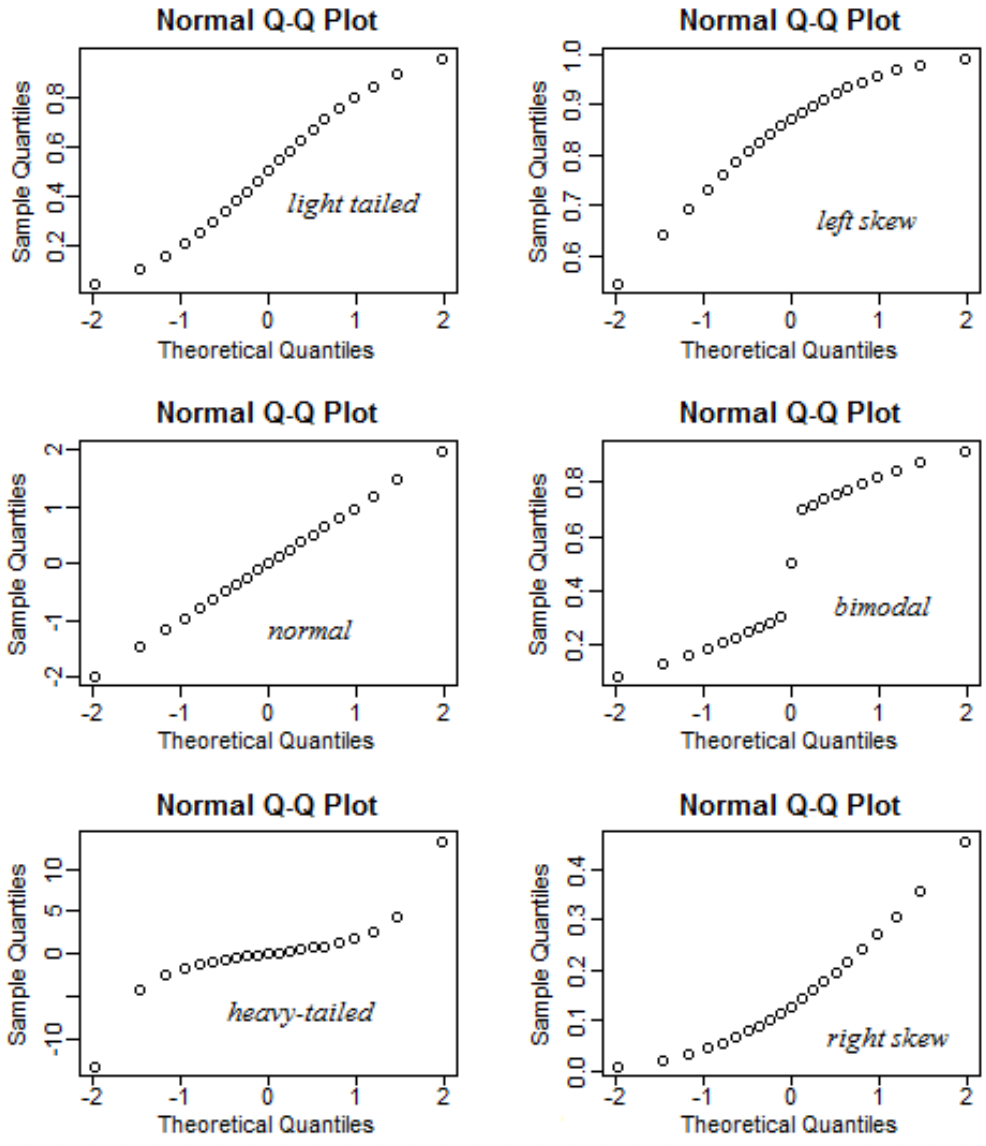
Normal Quantile-Quantile Plots

A normal QQ plot in R can be obtained using `qqnorm()` for the normal probability plot and `qqline()` to add the straight line.

```
qqnorm(tomato.data$pounds) # plot the points
qqline(tomato.data$pounds) # add line through 1st-3rd quartiles
```



Example deviations from a straight line



Note that for small samples (in the dozens) taken from a normal distribution, it's important not to read too much into every wiggle. Thanks to Glen B.

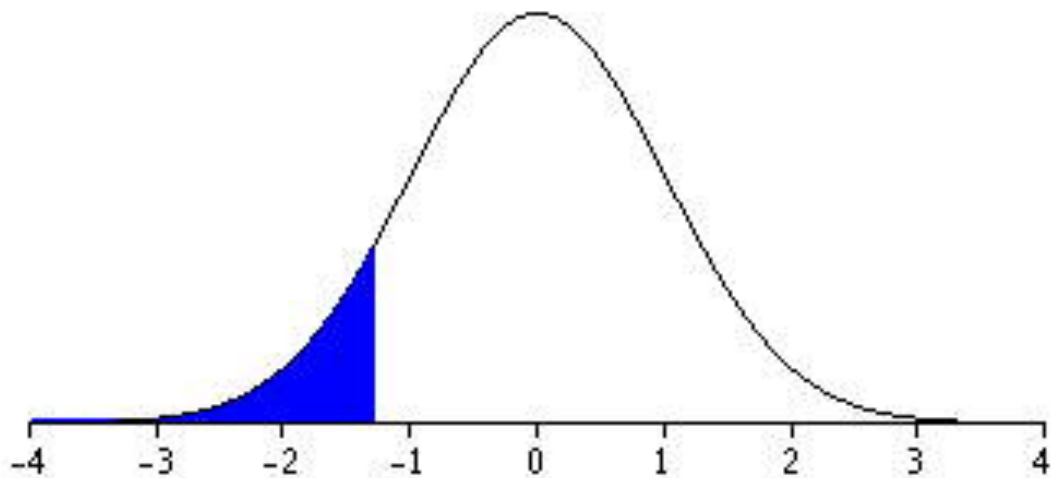
Notation recap for QQ plots

- If X is a continuous random variable with strictly increasing distribution function $F(x)$ then the pth quantile of the distribution is the value of x_p such that

$$F(x_p) = p$$

- i.e. the value such that

$$x_p = F^{-1}(p)$$



Quantile-Quantile Plots

- Suppose that we have independent observations $X_1, X_2, ..., X_n$ from a uniform distribution on $[0, 1]$, i.e. Unif$[0, 1]$
- The ordered sample values (also called the order statistics) are the values $X_{(j)}$ such that

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

- It can be shown that

$$E\left(X_{(j)}\right) = \frac{j}{n+1}$$

- This suggests that, if the underlying distribution is Unif[0,1], the following plot should be roughly linear:

$$X_{(j)} \quad \text{vs.} \quad \frac{j}{n+1}$$

Quantile-Quantile Plots

- If X is a continuous random variable with strictly increasing CDF F_X, it can be transformed to a Unif[0,1] by defining a new random variable $Y = F_X(X)$
- Suppose that it's hypothesized that X follows a certain distribution function with CDF F
- Given a sample $X_1, X_2, ..., X_n$, plot

$$F(X_{(k)}) \quad \text{vs.} \quad \frac{k}{n+1}$$

- This is equivalent to plotting the following equation:

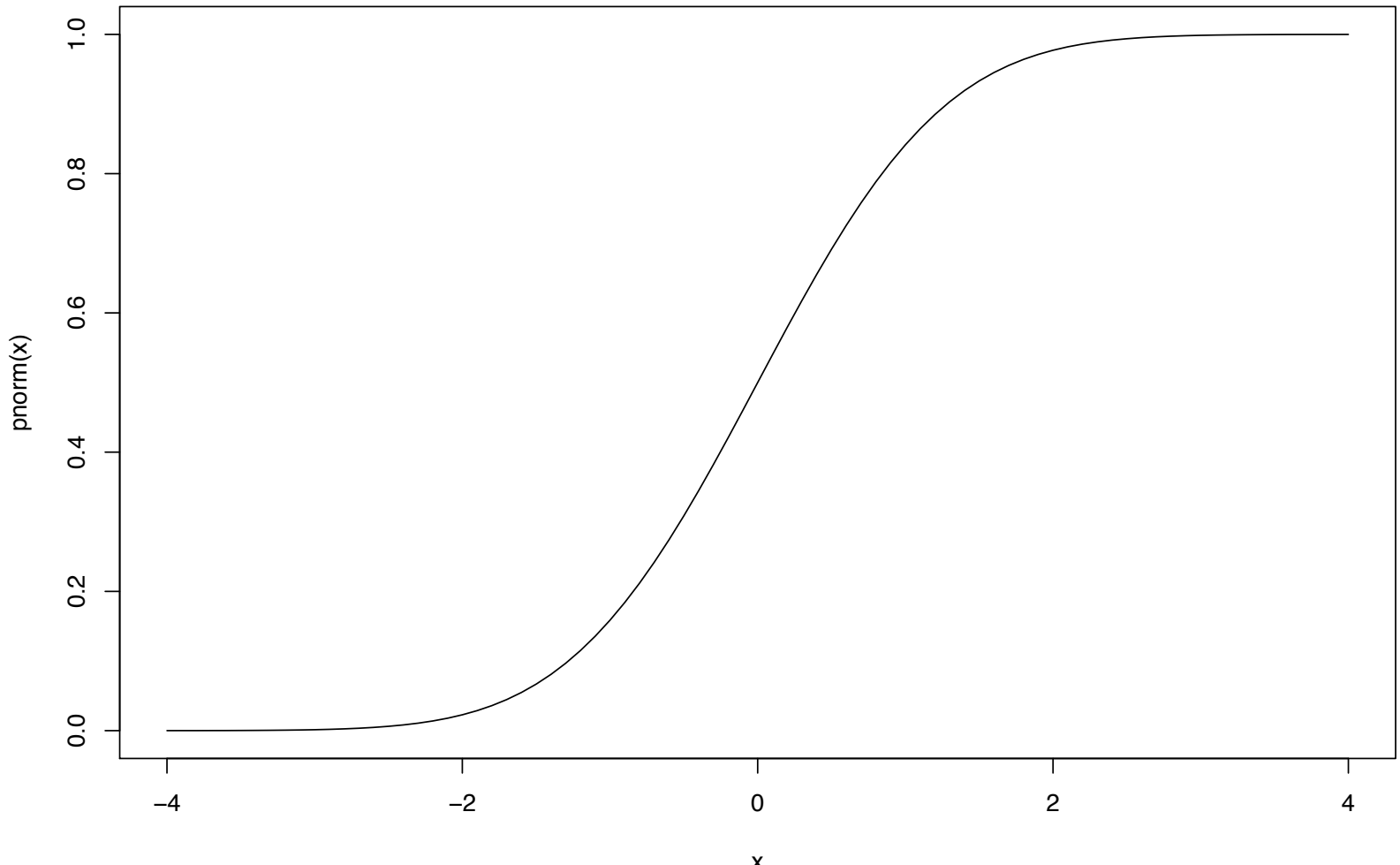
$$\boxed{X_{(k)} \quad \text{vs.} \quad F^{-1}\left(\frac{k}{n+1}\right)}$$

- $X_{(k)}$ can be thought of as empirical quantiles and $F^{-1}\left(\frac{k}{n+1}\right)$ as the hypothesized quantiles
- The quantile assigned to $X_{(k)}$ is not unique
  - Instead of assigning it $\frac{k}{n+1}$ it is sometimes assigned $\frac{k-0.5}{n}$
  - In practice, it makes little difference which definition is used

Normal Quantile-Quantile Plots

The cumulative distribution function (CDF) of the normal has an S-shape

```
x <- seq(-4,4,by=0.1)
plot(x,pnorm(x),type="l")
```



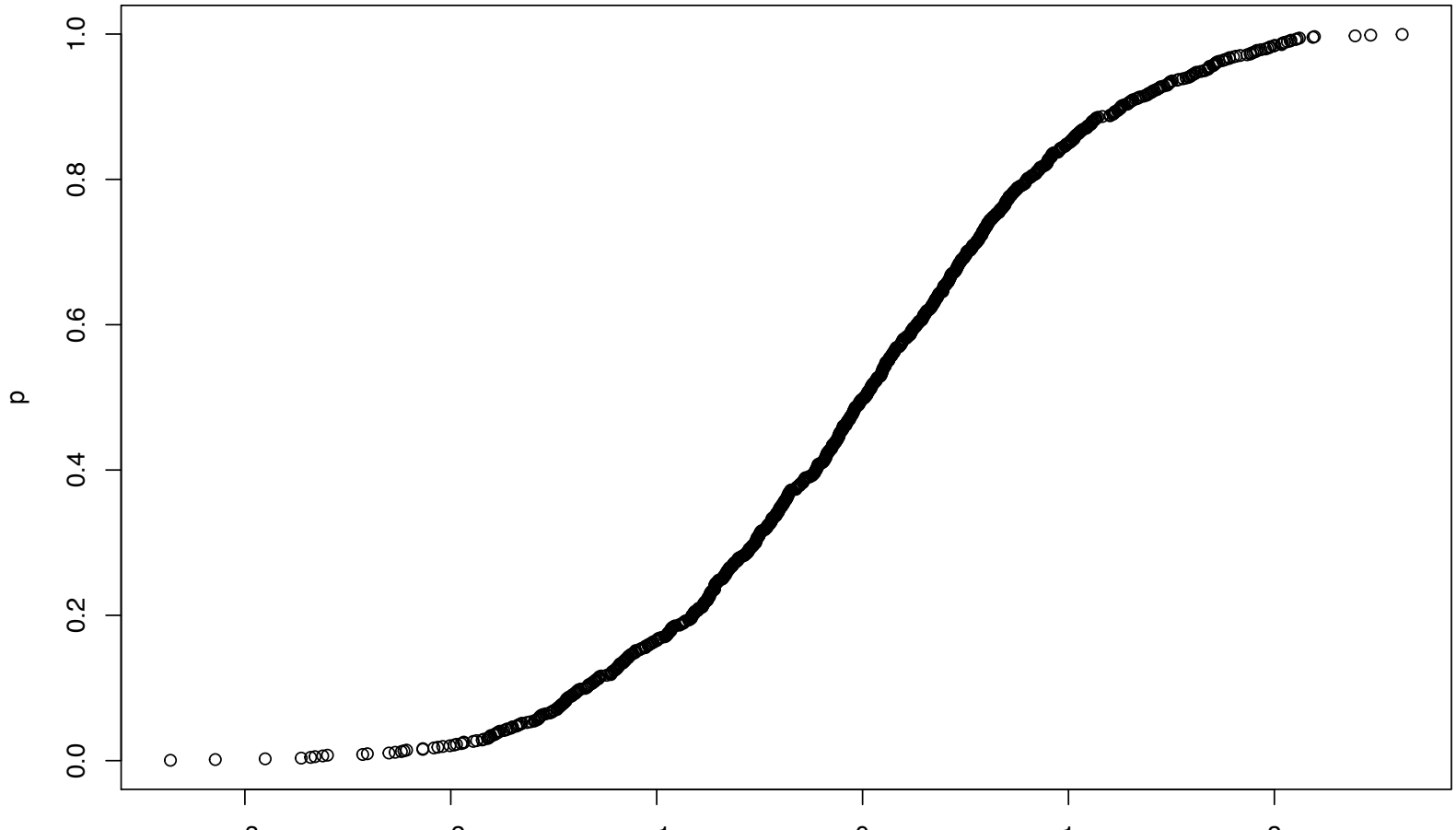
Normal Quantile-Quantile Plots

- Let $x_{(1)} < ... < x_{(N)}$ denote the ordered values of $x_1, ..., x_N$
- The normality of a set of data can be assessed by the following method.
  - Plot $p_i = (i - 0.5)/N$ versus the ordered values $x_{(i)}$ of the data
  - If the plot has the same S-shape as the normal CDF then this is evidence that the data come from a normal distribution

Normal Quantile-Quantile Plots

- Here is a plot of $p_i = (i - 0.5)/N$ vs. $x_{(i)}$, $i = 1, ..., N$ for a random sample of 1000 from a $\mathcal{N}(0, 1)$ distribution

```
N <- 1000;x <- rnorm(N);p <- ((1:N)-0.5)/N
plot(sort(x),p)
```



Normal Quantile-Quantile Plots

- It can be shown that the N points, $\{\Phi(x_i)\}$, are distributed roughly uniformly on $[0, 1]$
- This implies that $E(\Phi(x_{(i)})) = i/(N+1)$ — this is the expected value of the jth order statistic from a uniform distribution over $[0, 1]$
- This implies that the N points $(\Phi(x_{(i)}), p_i)$ should occur along a straight line
- Also the N points $(p_i, \Phi(x_{(i)}))$ should occur along a straight line, and we'll switch to this way round for now
- Now apply the Φ^{-1} transformation to the horizontal and vertical scales. The N points

$$\left(\Phi^{-1}(p_i), x_{(i)}\right)$$

  form the normal plot of $x_1, ..., x_N$
- If $x_1, ..., x_N$ are generated from a normal distribution then a plot of the points $\left(\Phi^{-1}(p_i), x_{(i)}\right), i = 1, ..., N$ should be a straight line

Normal Quantile-Quantile Plots

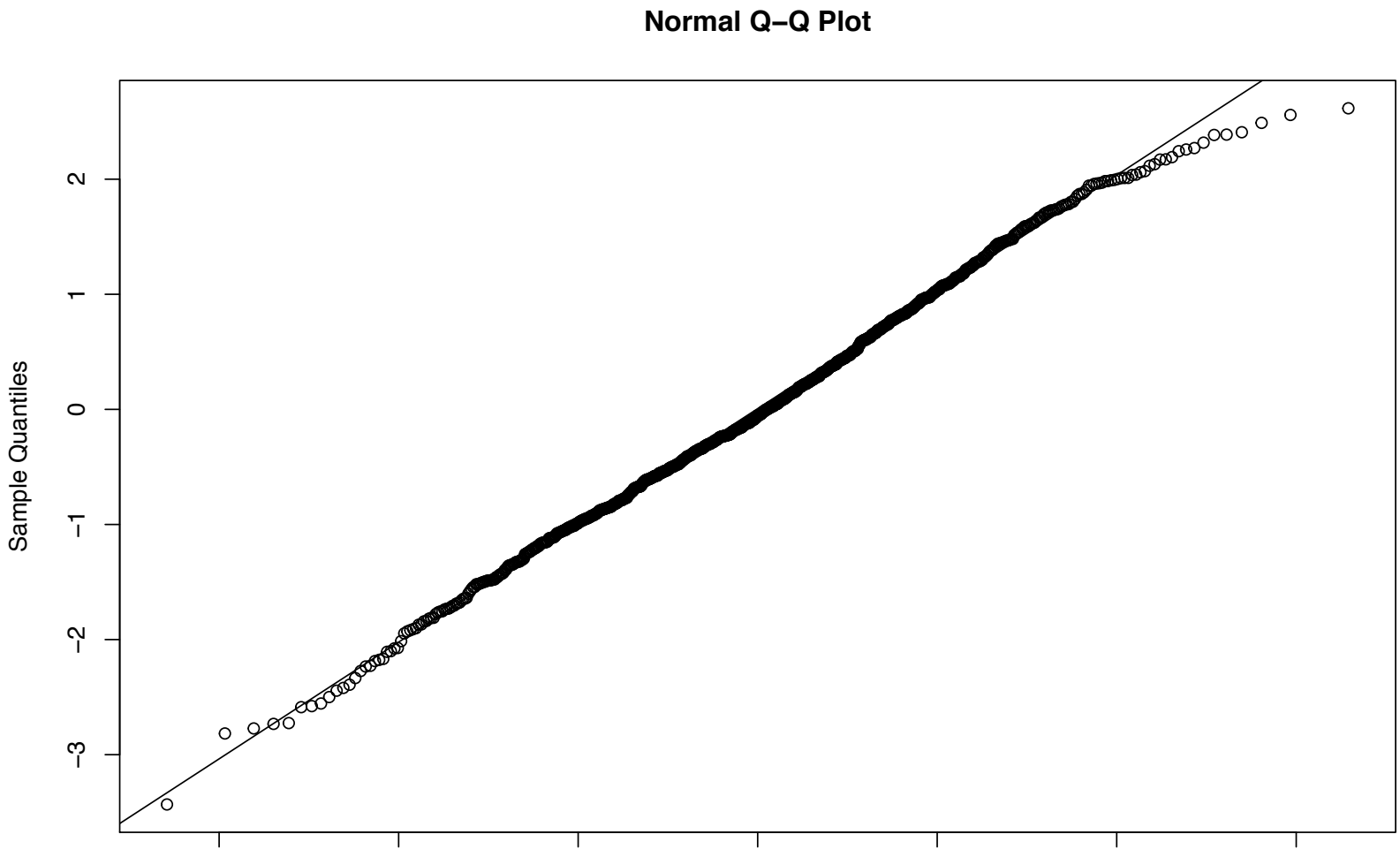
- In R, `qnorm()` is Φ^{-1} (see next slide)
- We usually use the built-in function `qqnorm()` (see two slides from now) to generate normal QQ plots. Note that R uses a slightly more general version of quantile: $p_i = (i - a)/(N + 1 - 2a)$, where

$$a = \begin{cases} 3/8 & \text{if } N \leq 10 \\ 1/2 & \text{if } N > 10 \end{cases}$$

- We also use `qqline()` to add a straight line for comparison
- A marked (systematic) deviation of the plot from the straight line would indicate that:
  - The normality assumption does not hold, and/or
  - The constant error variance assumption does not hold

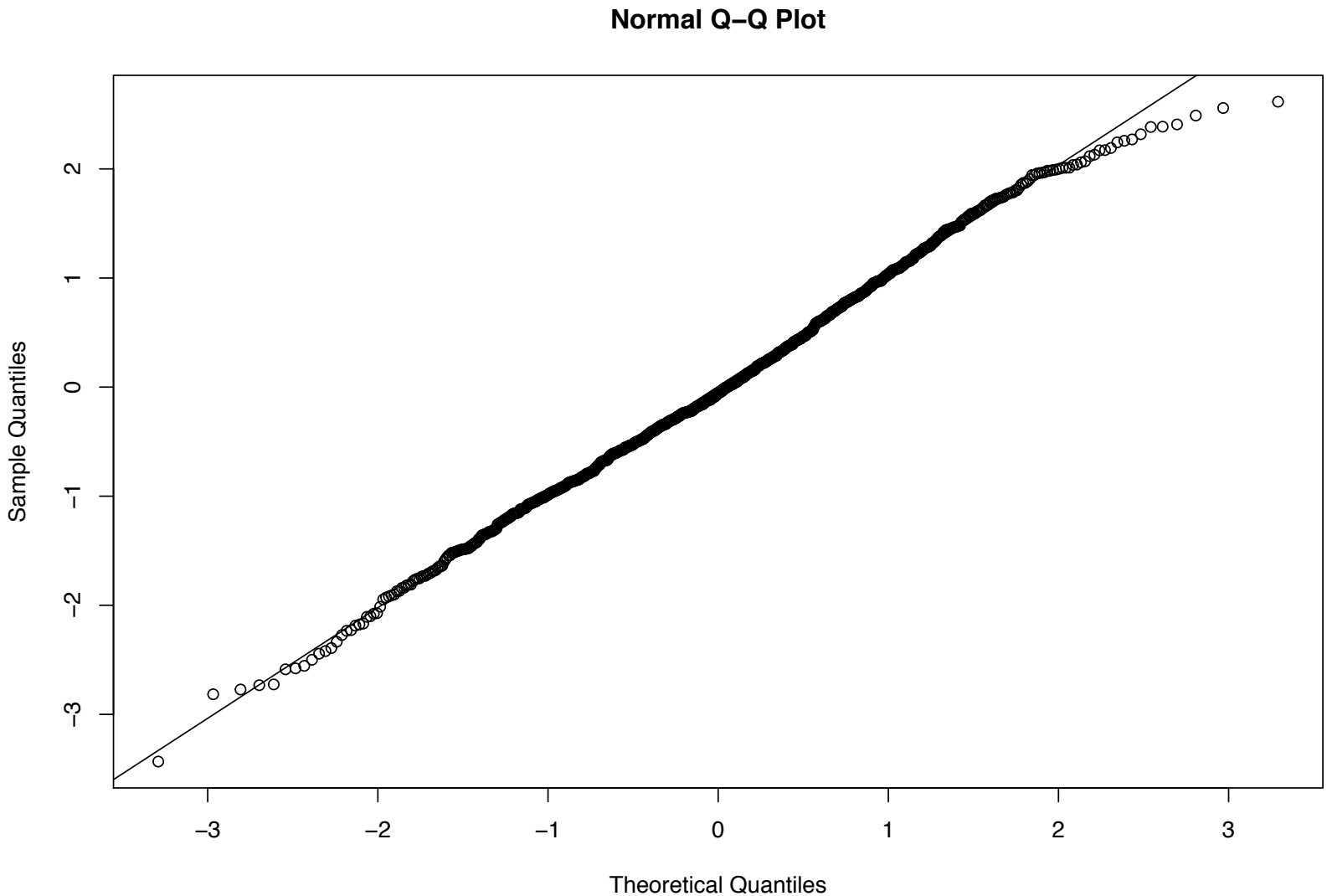
Emulating qqnorm

```
set.seed(2503)
N <- 1000
x <- rnorm(N)
p <- ((1:N)-0.5)/N # or (1:N)/(N+1)
plot(qnorm(p),sort(x)) # Empirical- vs theoretical quantiles
```



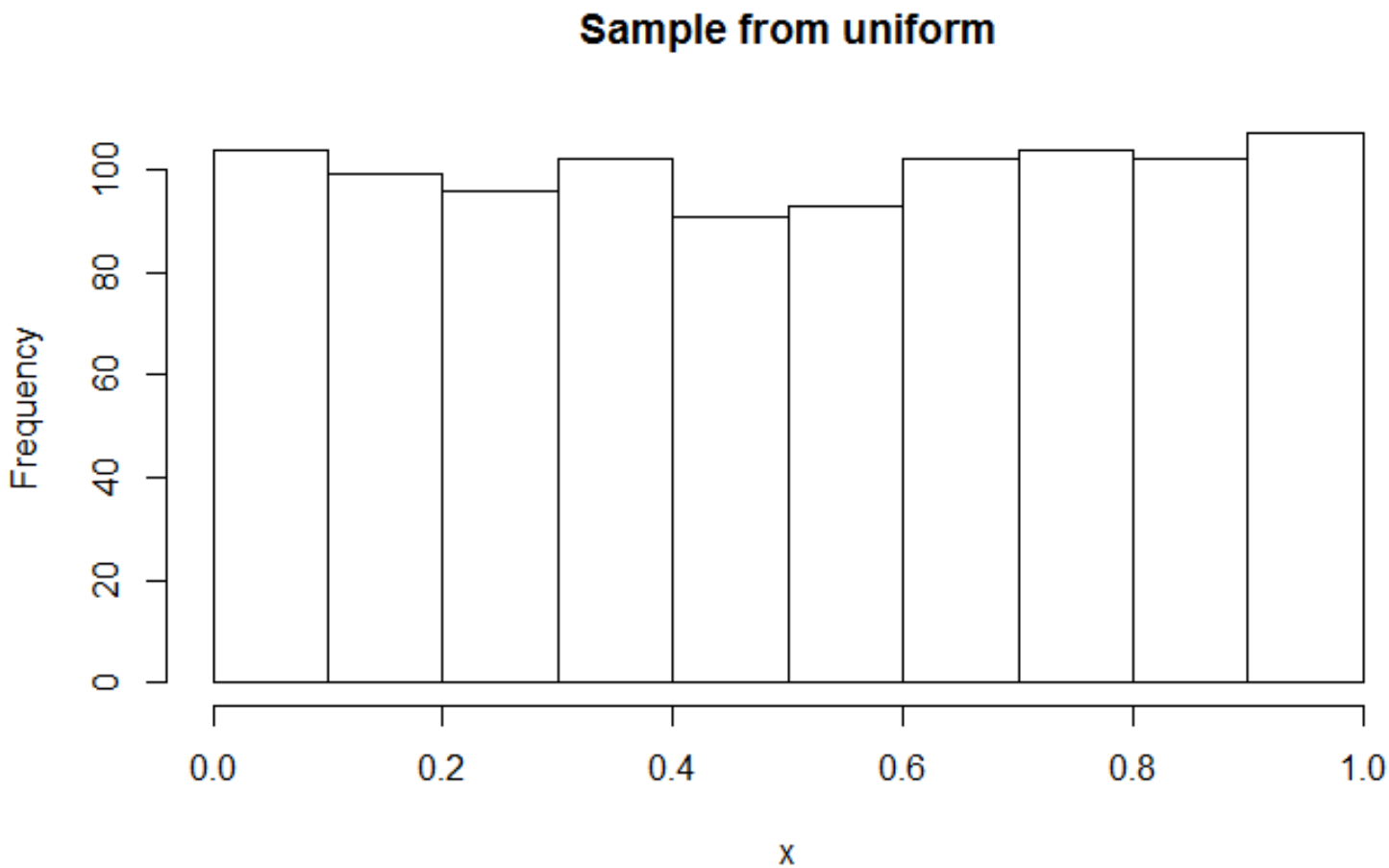
Normal Quantile-Quantile Plots (Ex. 1)

```
qqnorm(x);qqline(x)
```



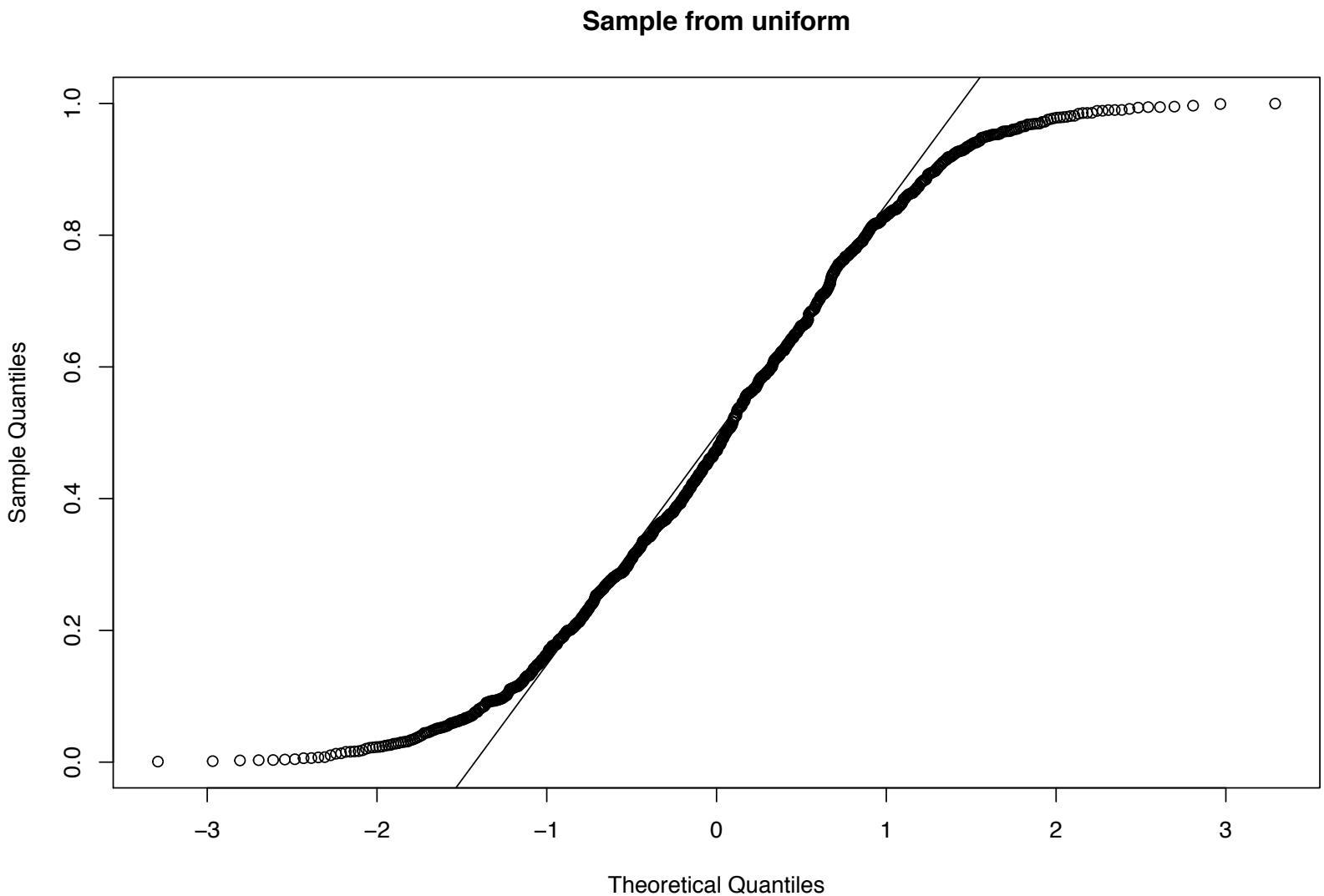
Normal Quantile-Quantile Plots (Ex. 2)

```r
x <- runif(1000)
hist(x,main = "Sample from uniform")
```

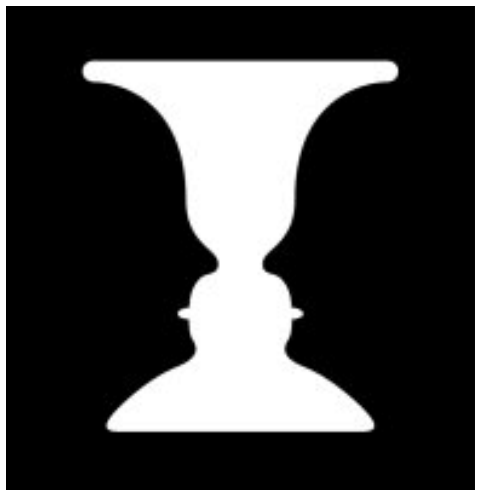


Normal Quantile-Quantile Plots

```
x <- runif(1000)
qqnorm(x,main = "Sample from uniform"); qqline(x)
```

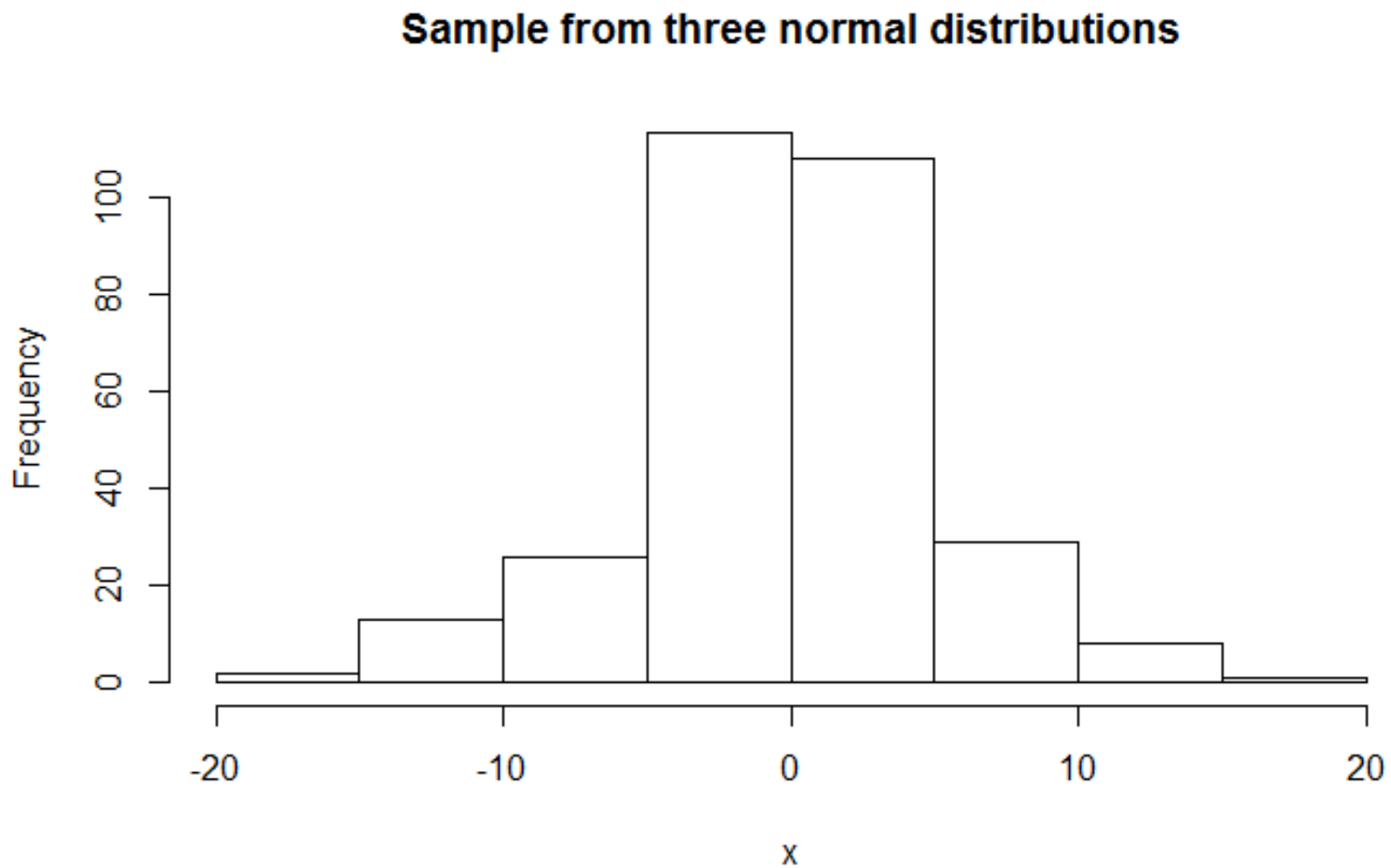


A Rubin's Vase



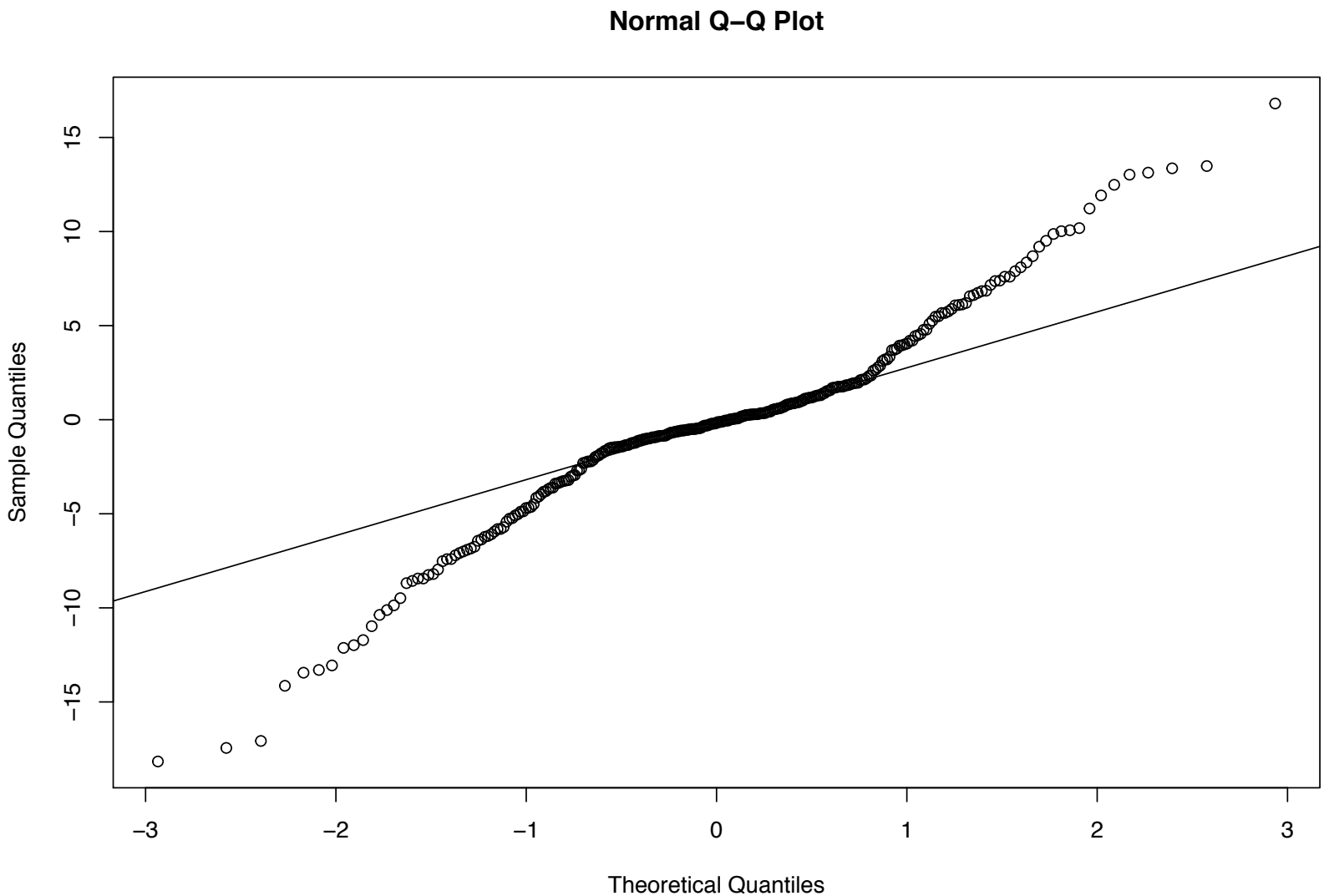
Example of Systematic Deviation (Ex. 3)

```
x1 <- rnorm(100,mean = 0,sd = 1); x2 <- rnorm(100,mean = 0,sd = 5)
x3 <- rnorm(100,mean = 0,sd = 8); x <- c(x1,x2,x3)
hist(x,main = "Sample from three normal distributions")
```



Non-constant Variance

```
x1 <- rnorm(100,mean = 0,sd = 1); x2 <- rnorm(100,mean = 0,sd = 5)
x3 <- rnorm(100,mean = 0,sd = 8); x <- c(x1,x2,x3)
qqnorm(x);qqline(x)
```



Regression diagnostics: You've Sailed the Seven C's

1. Plot *standardized residuals* to help determine whether the proposed regression model is a valid model
2. Identify any *leverage points*
3. Identify any *outliers*
4. Identify any *influential points*
5. Assess the assumption of *error homoscedasticity*
6. For time series: examine whether the data are *correlated over time*
7. Assess the assumption of *normal errors*

Next steps

Try these questions in Chapter 3:

- 1
- 3 part A
- 4(a)

Next week, among other things we'll review the suggested solutions to Chapter 2's questions 1 (b), (d).

The midterm will only cover up to §3.2