Further Model Checks

CHRIS SUTTON, FEBRUARY 2024

The Simple Linear Regression Model

So far we have:

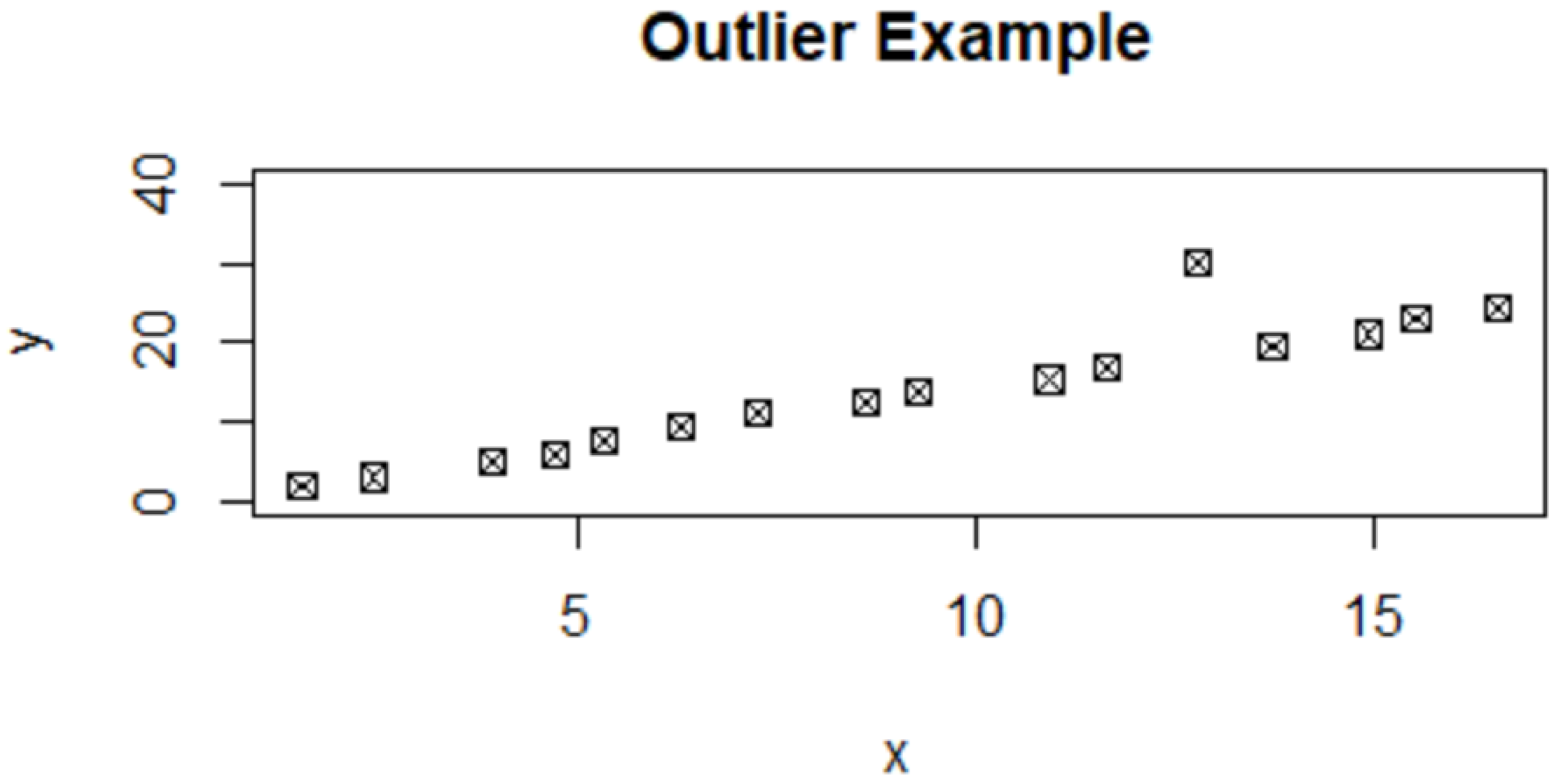
- ❑ constructed a simple linear regression model
  - ❑ least squares, `lm()` function
- ❑ analysed the output
  - ❑ `anova()`, residual plots
- ❑ made conclusions from model evidence
  - ❑ confidence intervals, test of hypotheses

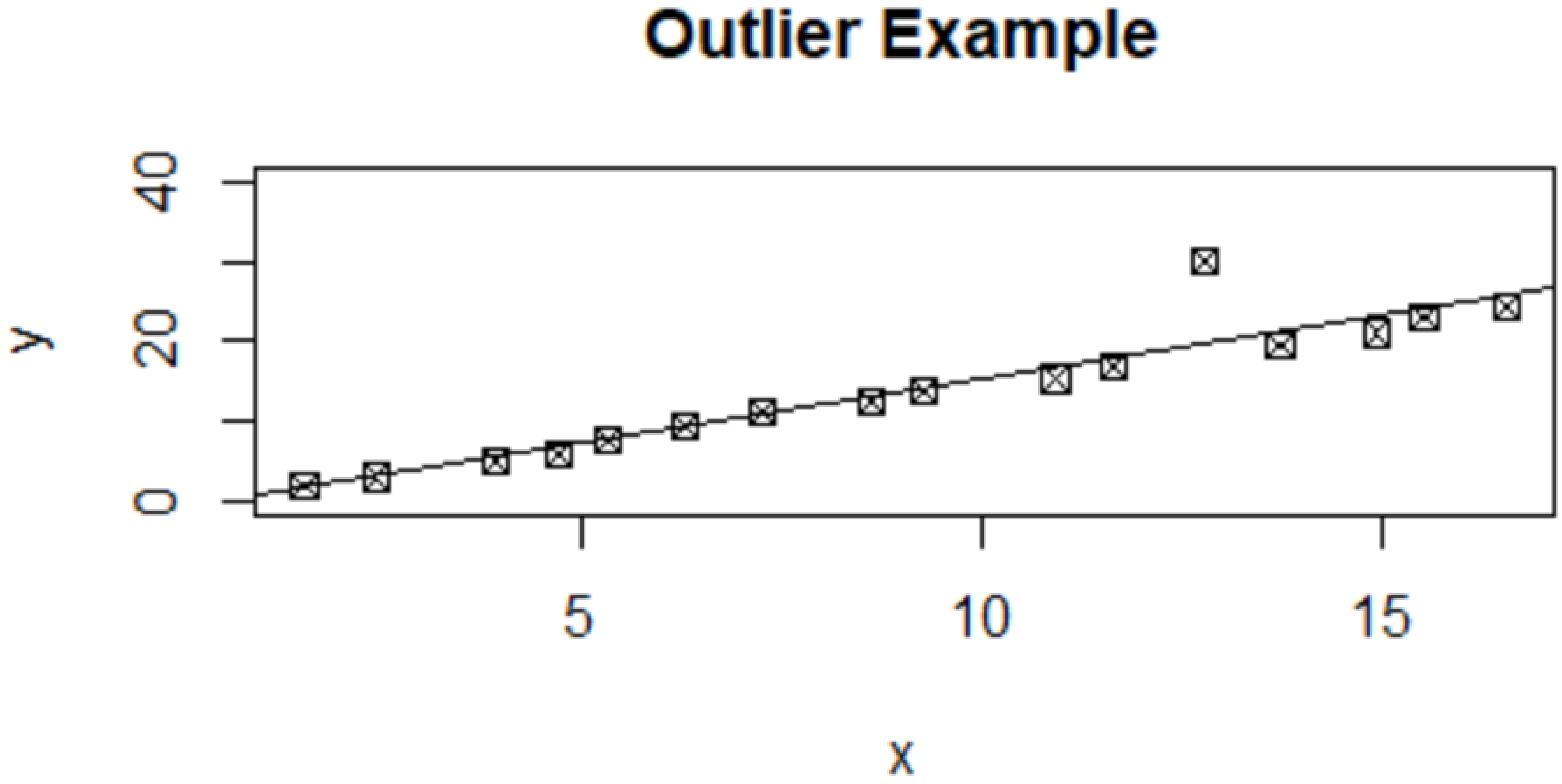
Now we will consider areas where our observed data leads to issues with using simple linear regression modelling

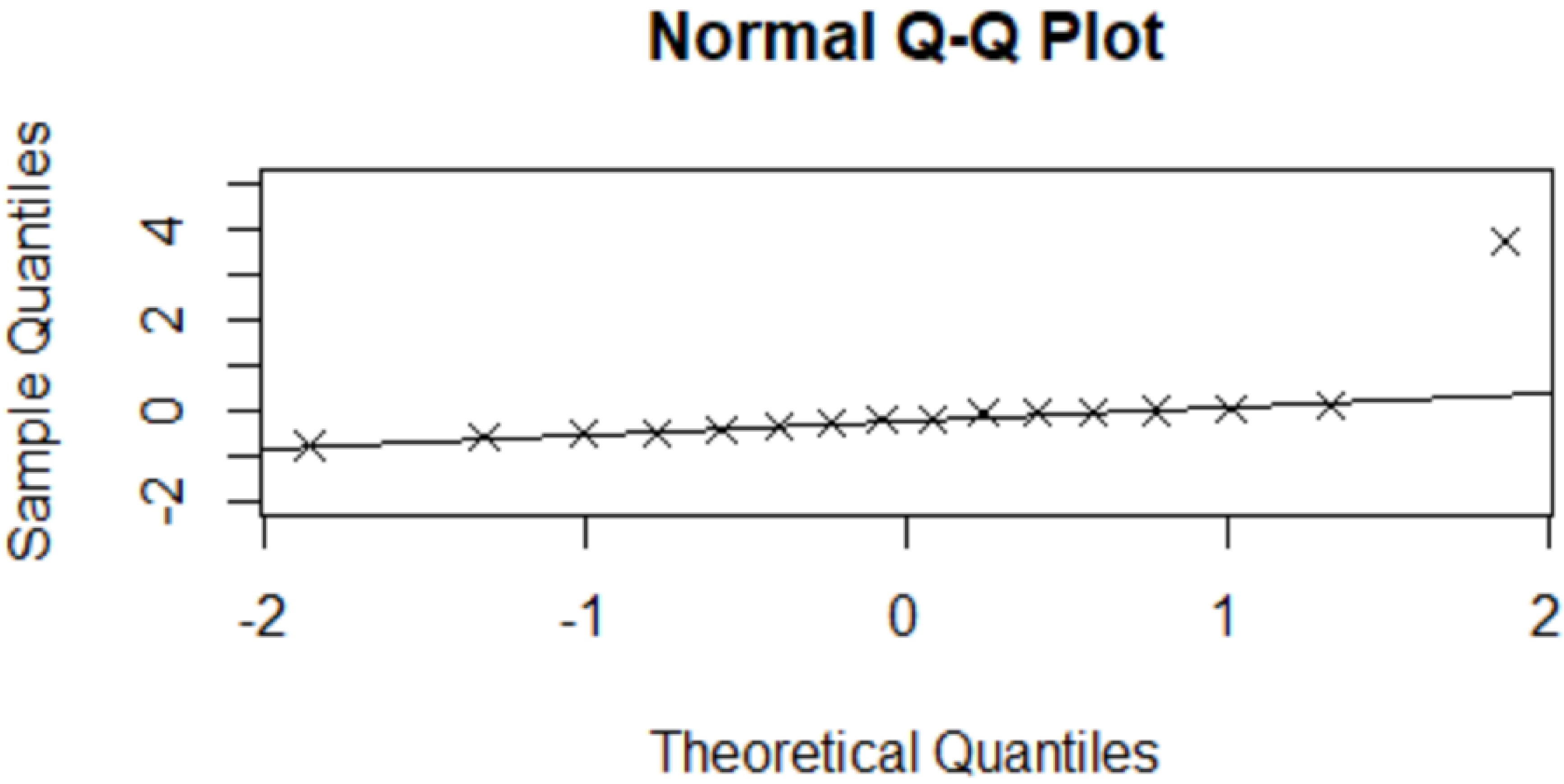
Outliers

An **outlier** is a single observation where the absolute value of the standardised residual is large compared to the rest of the observations

- Outliers are usually obvious from residual plots e.g. Q-Q plots


Outlier Example
y
0
20
40
5
10
15
x


Outlier Example
y
0
20
40
x
5
10
15


Normal Q-Q Plot
Sample Quantiles
-2
0
2
4
-2
-1
0
1
2
Theoretical Quantiles

Residuals and standardised residuals

We defined residuals e_i and standardised residuals d_i in week 2

$$e_i = y_i - \hat{y}_i$$

Which we often standardise before plotting to give a variance closer to σ^2

$$d_i = \frac{e_i}{\sqrt{s^2\left(1 - \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)}}$$

The R command to calculate a vector containing the d_i is `rstandard()` where the argument is the name we assigned to our `lm(y~x)` model

What d_i makes it an outlier?

Some books will suggest a simple rule for spotting an outlier

- e.g. | d_i | > 2

But actually what constitutes an outlier will depend on the sample size *n*

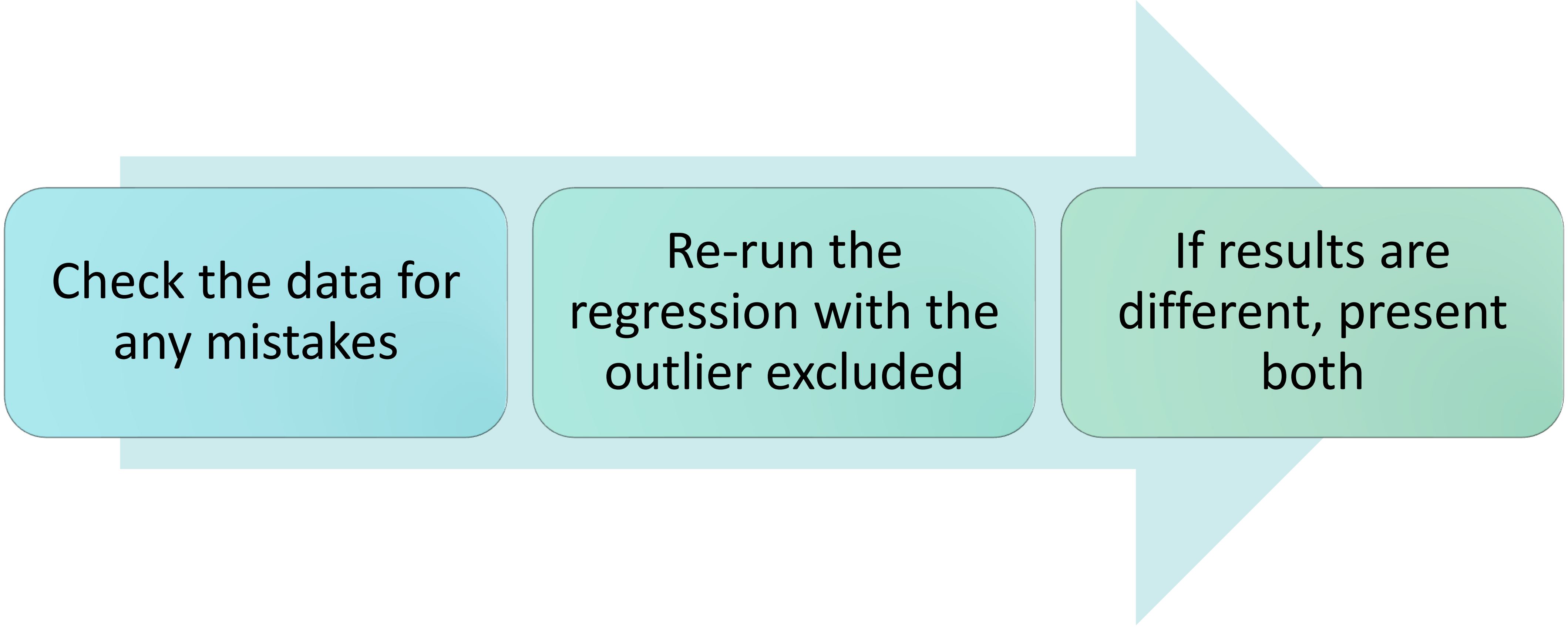
The higher *n* is, the larger the value of | d_i | needs to be before we say the observation is an outlier

We can create a table of values for | d_i | that mark the upper bound of a 95% confidence interval for d_i at different sample sizes *n*

Finding an outlier

Sample size *n*	Maximum $\|d_i\|$ at 95% significance
6	1.93
8	2.20
10	2.37
20	2.77
30	3.06
60	3.23

What to do if you find an outlier



Check the data for any mistakes

Re-run the regression with the outlier excluded

If results are different, present both

Baseball crowds

Modelling question

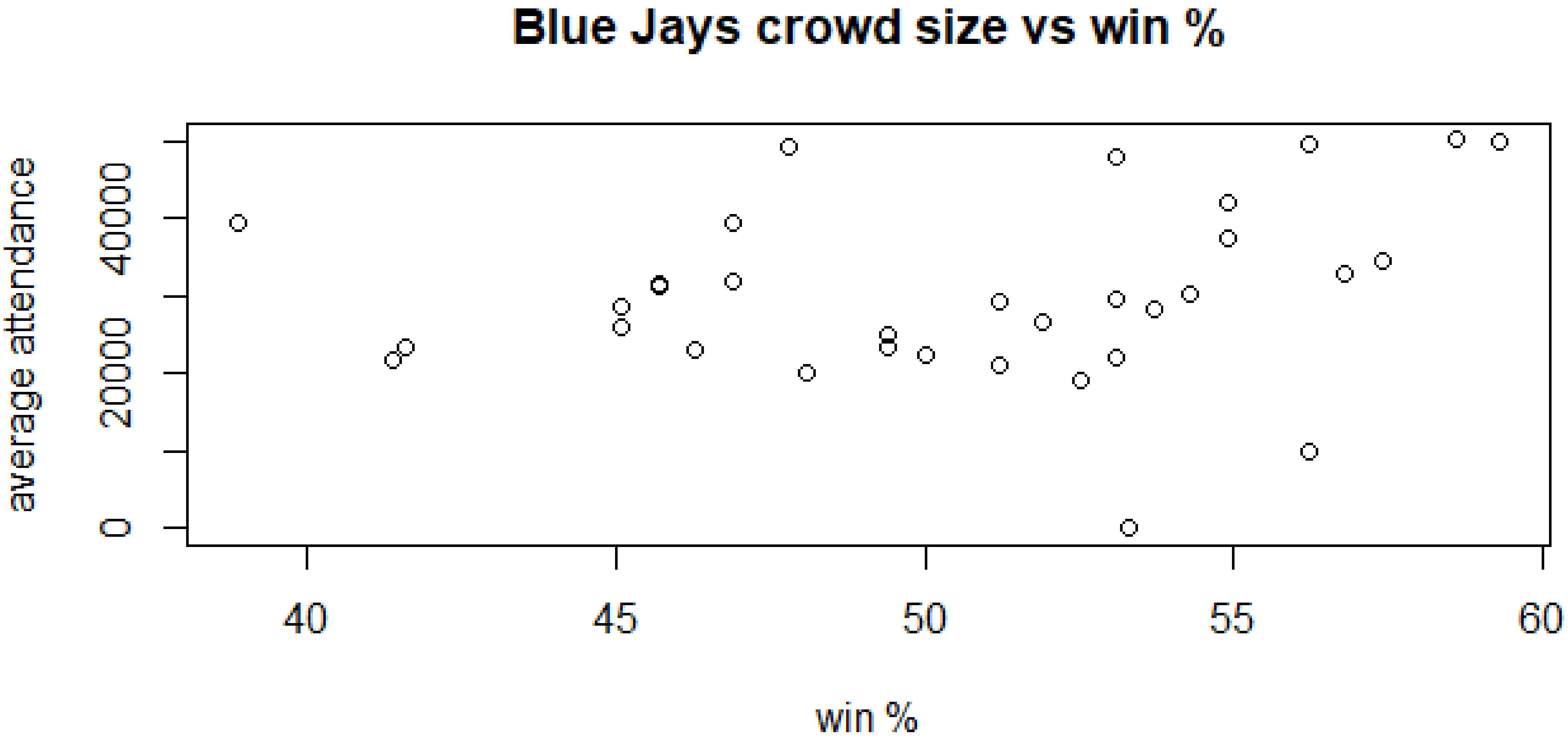
Do more people come to watch the Toronto Blue Jays at the Rogers Centre in years when the team are winning more?

source: Baseball Reference https://www.baseball-reference.com/teams/TOR/attend.shtml

For years i = 1990 to 2023

x_i = win percentage (games won / games played in the season)

y_i = average crowd size per home game


Blue Jays crowd size vs win %
average attendance
0
20000
40000
40
45
50
55
60
win %

crowd <- lm(y~x)

```
lm(formula = y ~ x)
Residuals:
   Min     1Q Median     3Q    Max
-31403  -4990  -1431   4795  20378
Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept)   7240.2    19622.6   0.369    0.715
x              453.3      385.7   1.175    0.249
Residual standard error: 11270 on 32 degrees of freedom
Multiple R-squared:  0.04138,  Adjusted R-squared:  0.01142
F-statistic: 1.381 on 1 and 32 DF,  p-value: 0.2486
```

anova(crowd)

```
Analysis of Variance Table
Response: y
          Df     Sum Sq   Mean Sq F value Pr(>F)
x          1  175507783 175507783  1.3812 0.2486
Residuals 32 4066216051 127069252
```

This does not look like a linear model with explanatory power

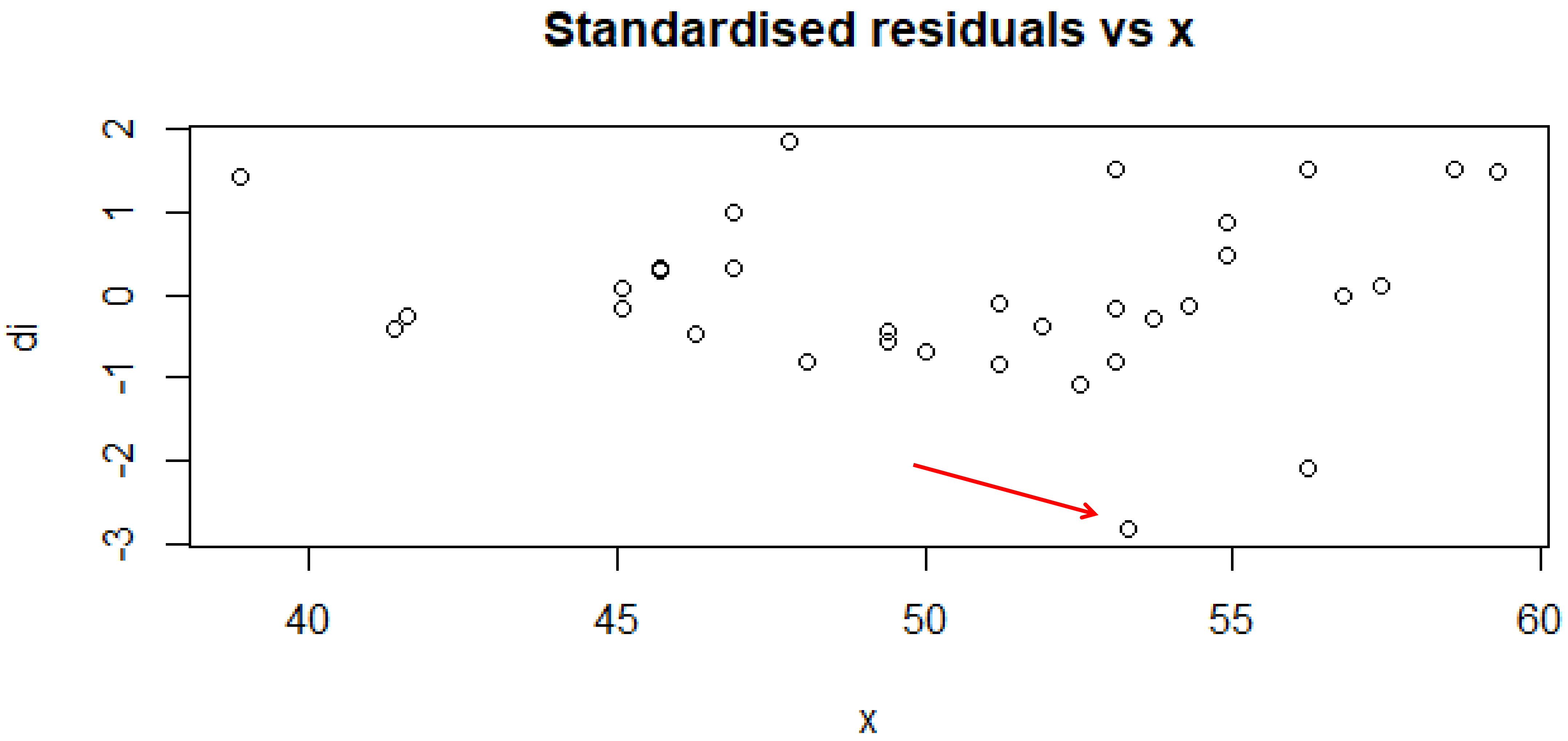
$\widehat{\beta_1} = 453$ estimated increase in crowd size for 1% increase in win %

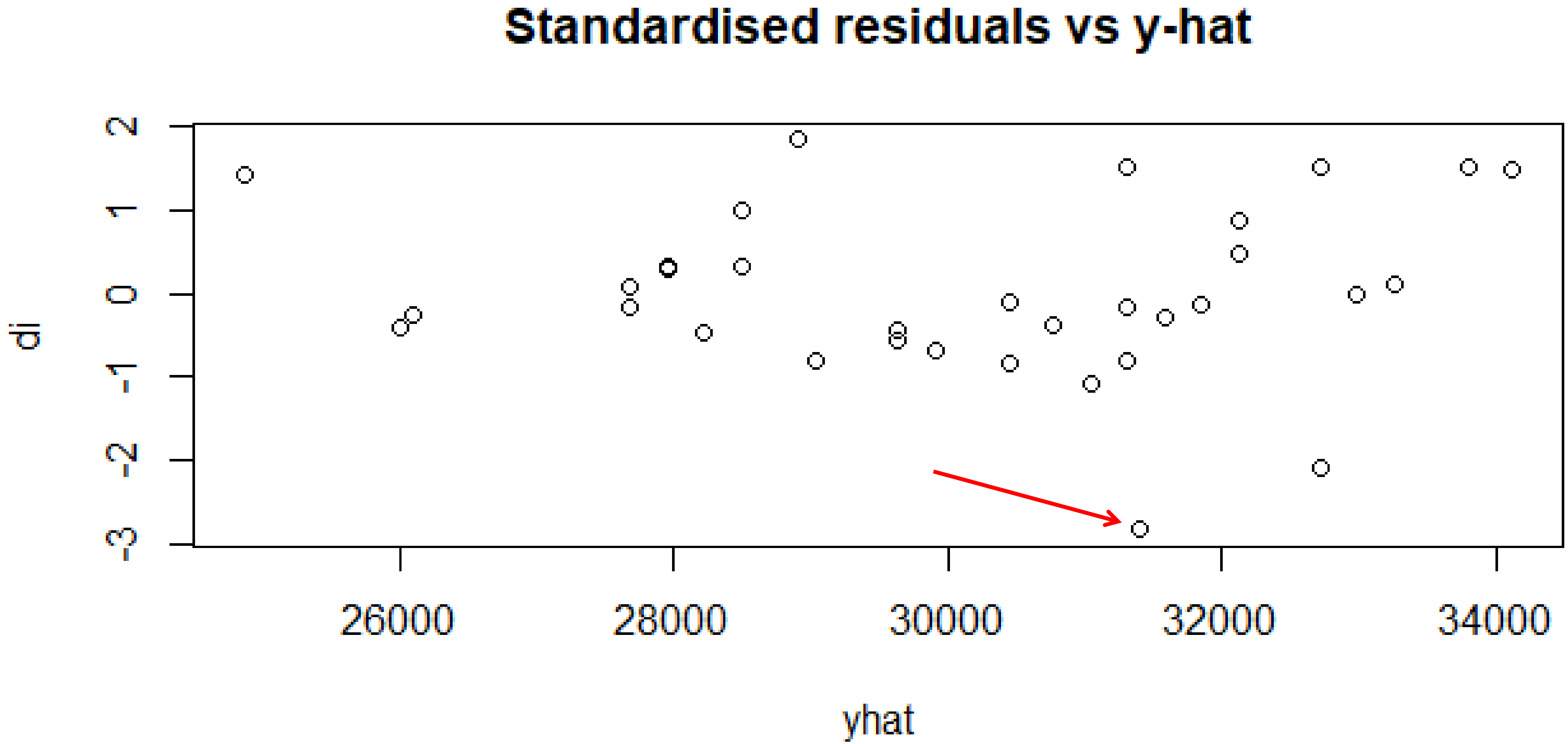
We cannot reject H_0: $\widehat{\beta_1} = 0$ at 95% significance

Variance Ratio = 1.38 < $F^1_{32}(0.05) = 4.15$

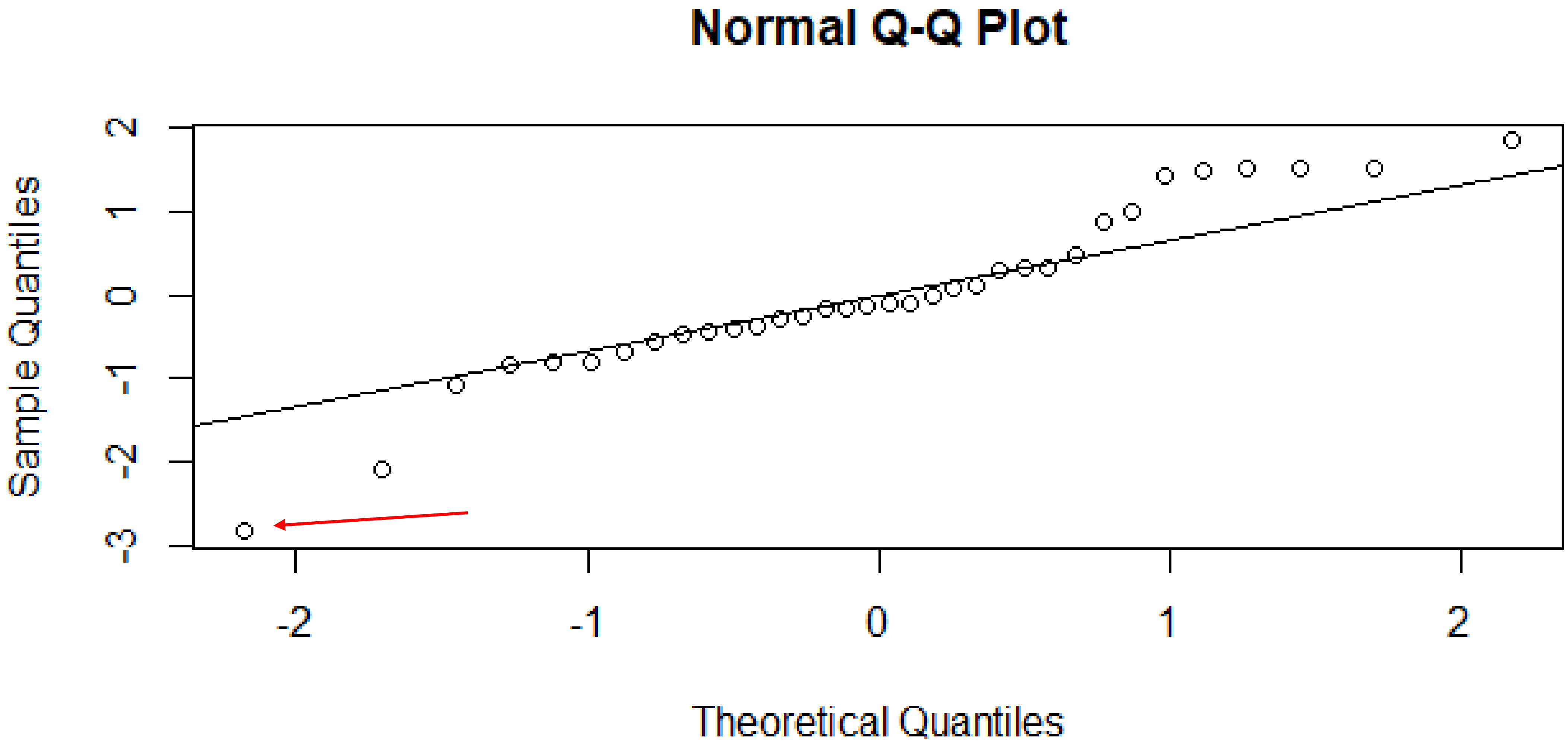
$R^2 =$ 4% virtually none of the variability in crowd size is explained by win rate

But …


Standardised residuals vs x
di
2
1
0
-1
-2
-3
40
45
50
55
60
x



Standardised residuals vs y-hat


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
2
1
0
-1
-2
-3
-2
-1
0
1
2

What happens if we remove Covid data?

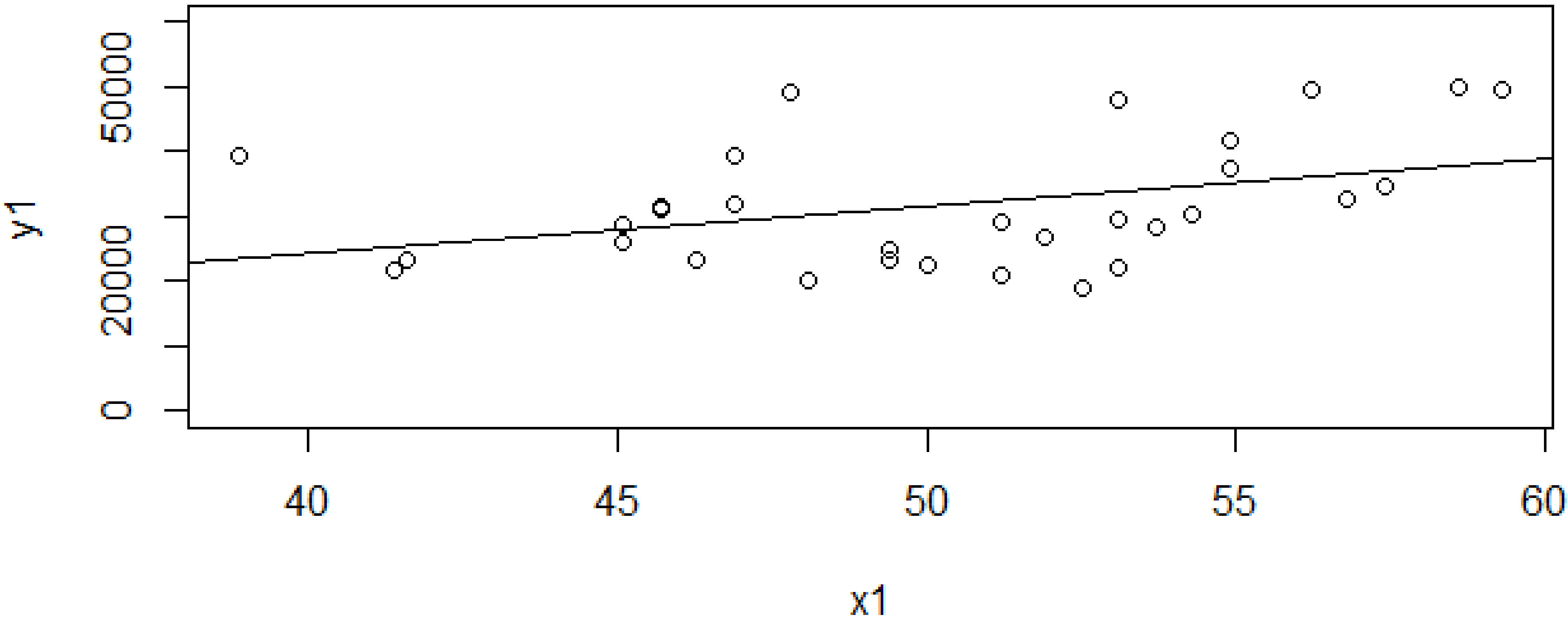
Due to COVID-19 restrictions no crowd was allowed at any games in 2020

(and restrictions still in 2021)

BlueJays had winning years

What effect do these two observations have on our model?

Crowd model without 2020-21 data



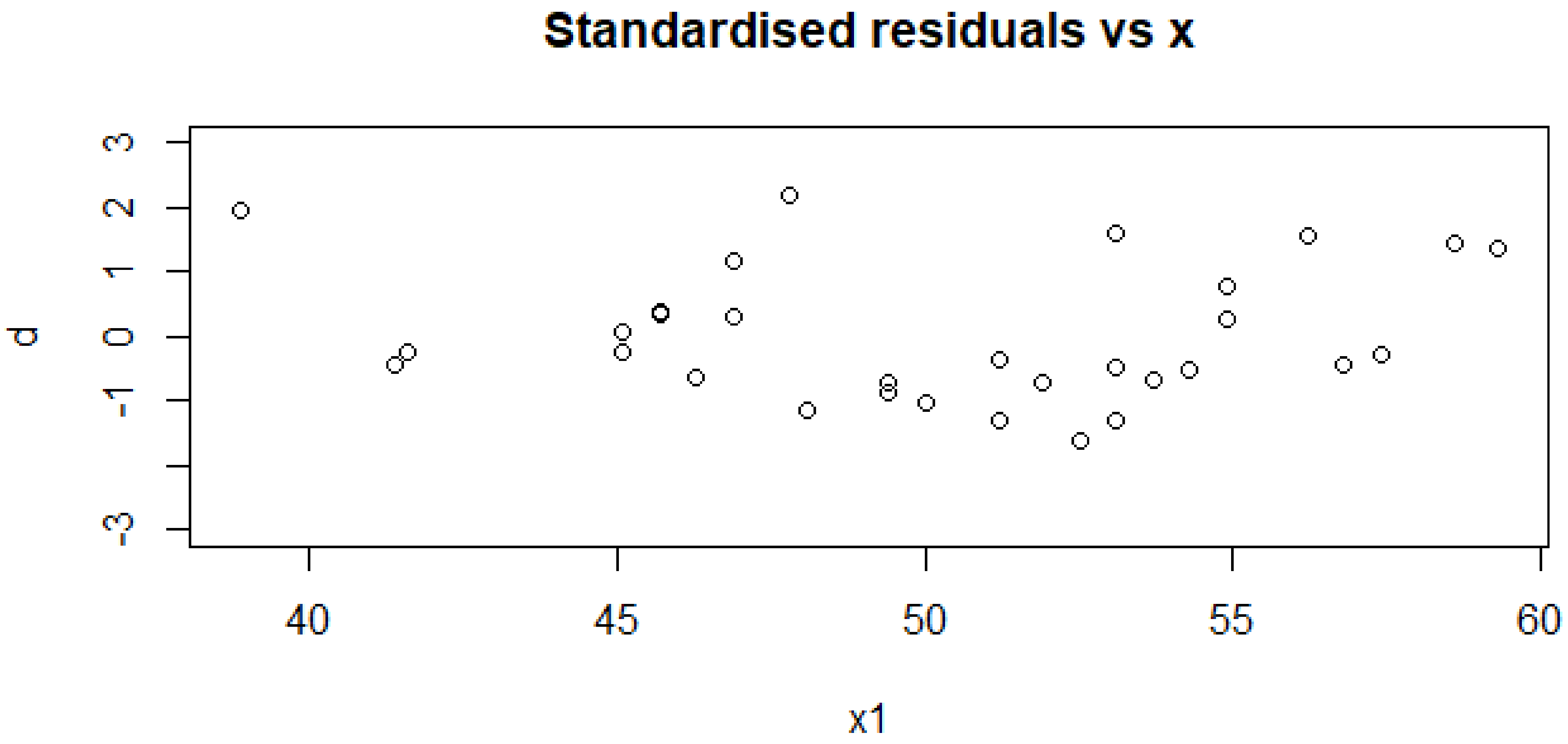
The model is still not great, but it’s better

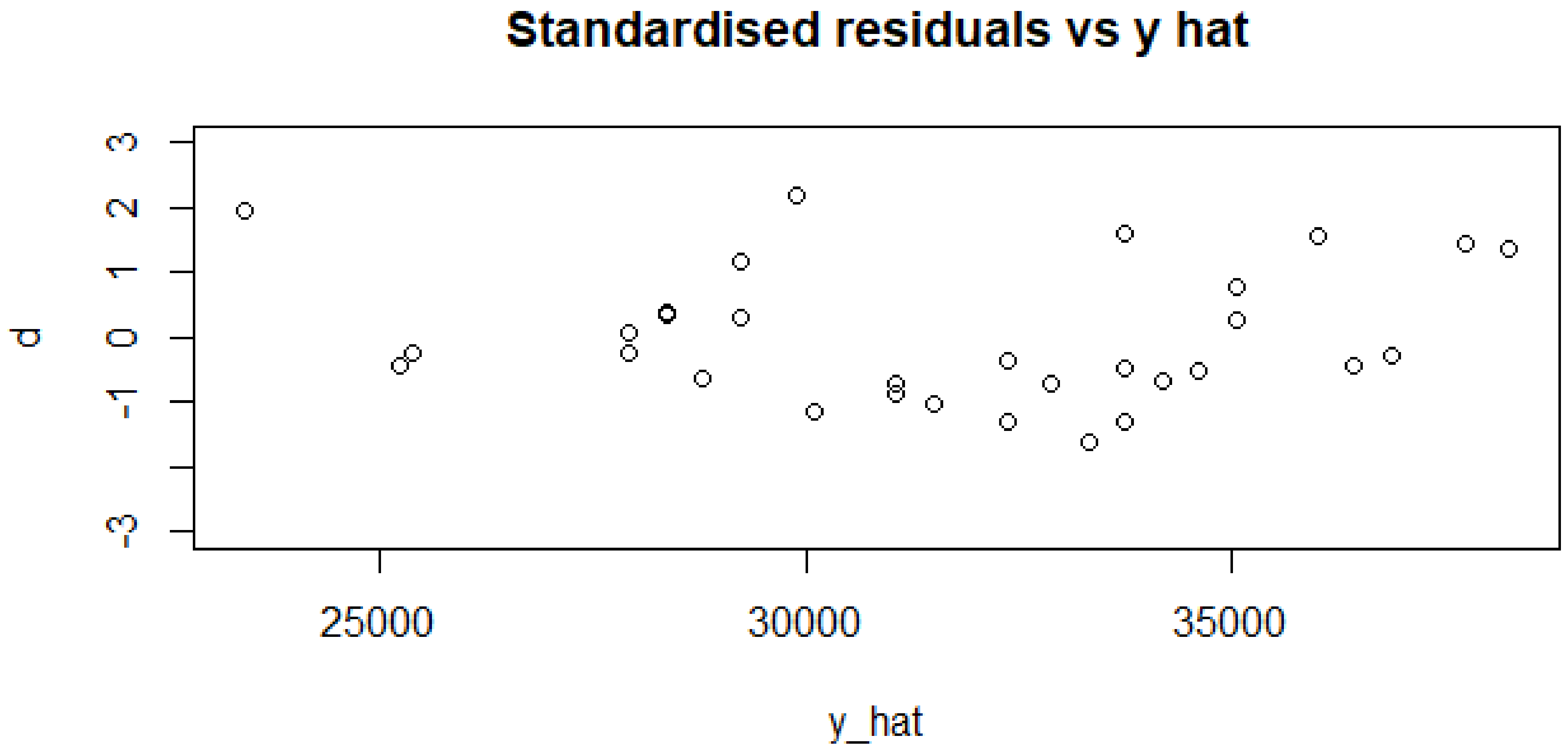
Now $\widehat{\beta_1} = 729$ extra fans per 1% win rate rise

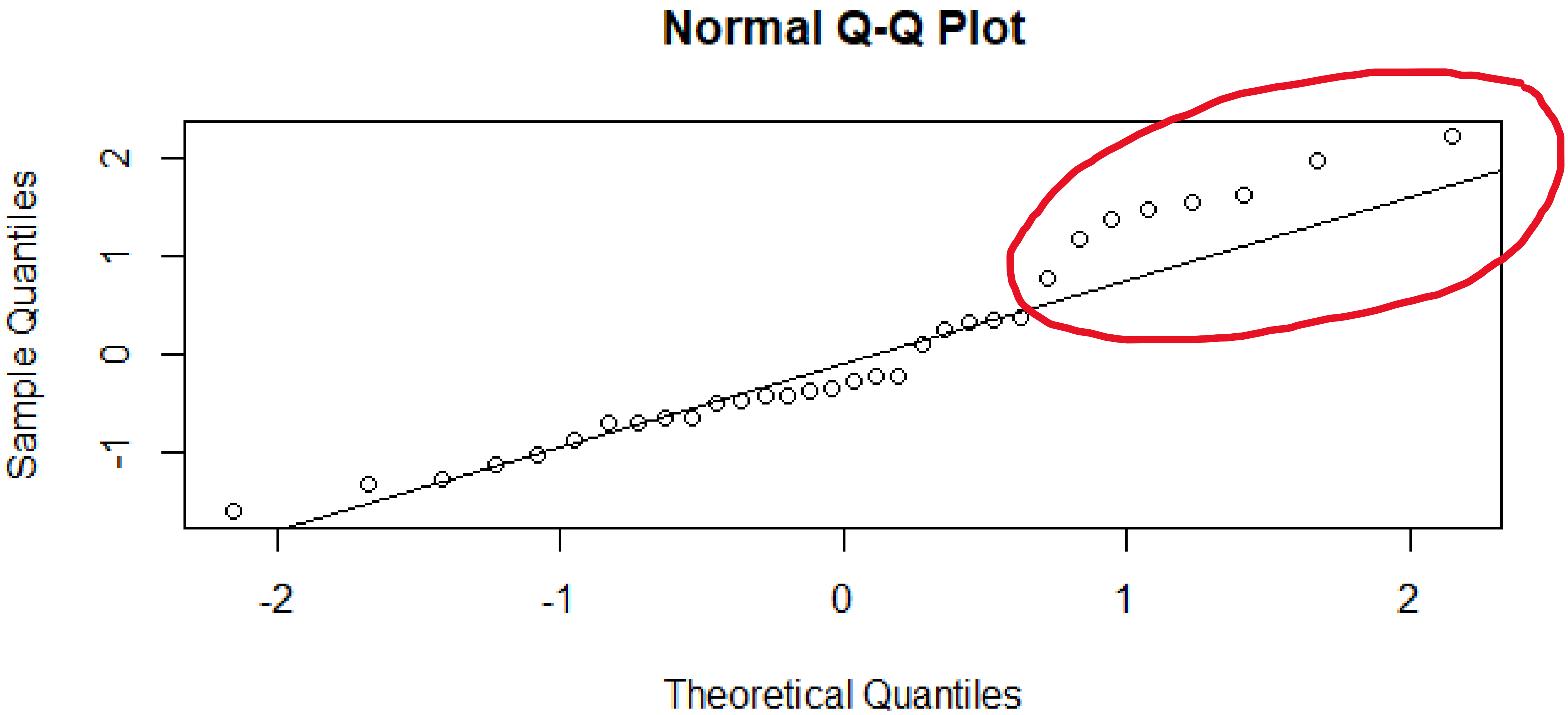
And we can reject H_0: $\widehat{\beta_1} = 0$ at 95% significance (but not at 99%)

Variance Ratio = 5.37 > $F^1_{30}(0.05) = 4.17$

$R^2 = 15\%$ little of the variability in crowd size is explained by win rate


Standardised residuals vs x
d
3
2
1
0
-1
-3
40
45
50
55
60
x1


Standardised residuals vs y hat
d
3
2
1
0
-1
-3
25000
30000
35000
y_hat


Normal Q-Q Plot
Sample Quantiles
-1
0
1
2
-2
-1
0
1
2
Theoretical Quantiles

The Q-Q plot is useful here

This second QQ plot is perhaps the most useful diagnostic tool for what is going on here

It looks as though the Normal distribution assumption holds very well for a large part of the data set

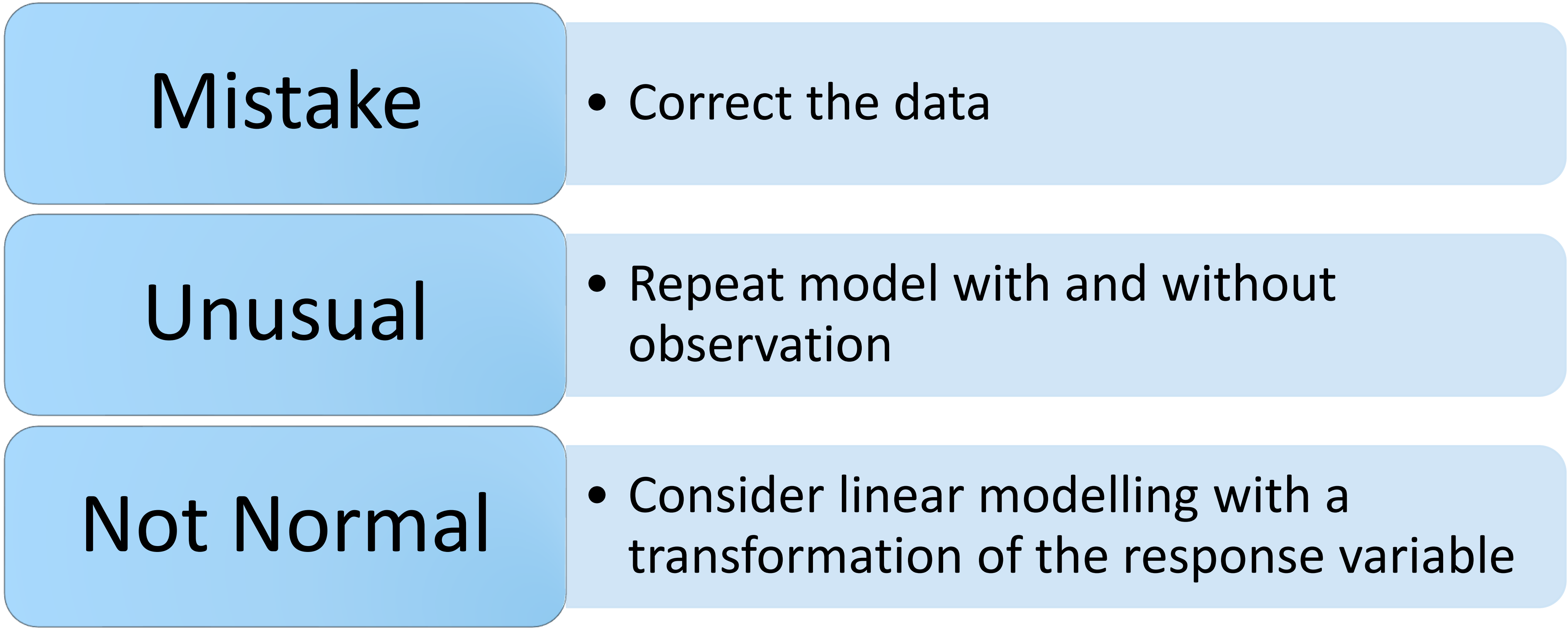
However there is a distinct set of (8) high (positive) residuals which are not what we would expect under the Normal distribution assumption

These are years where the fitted $\hat{y}$ underestimates the observed y

3 reasons why residual plots may give concern

- Mistakes in data entry
- Observation under different conditions from others
- Situations where distribution of residuals not Normal

Requires three different responses



Mistake

- Correct the data

Unusual

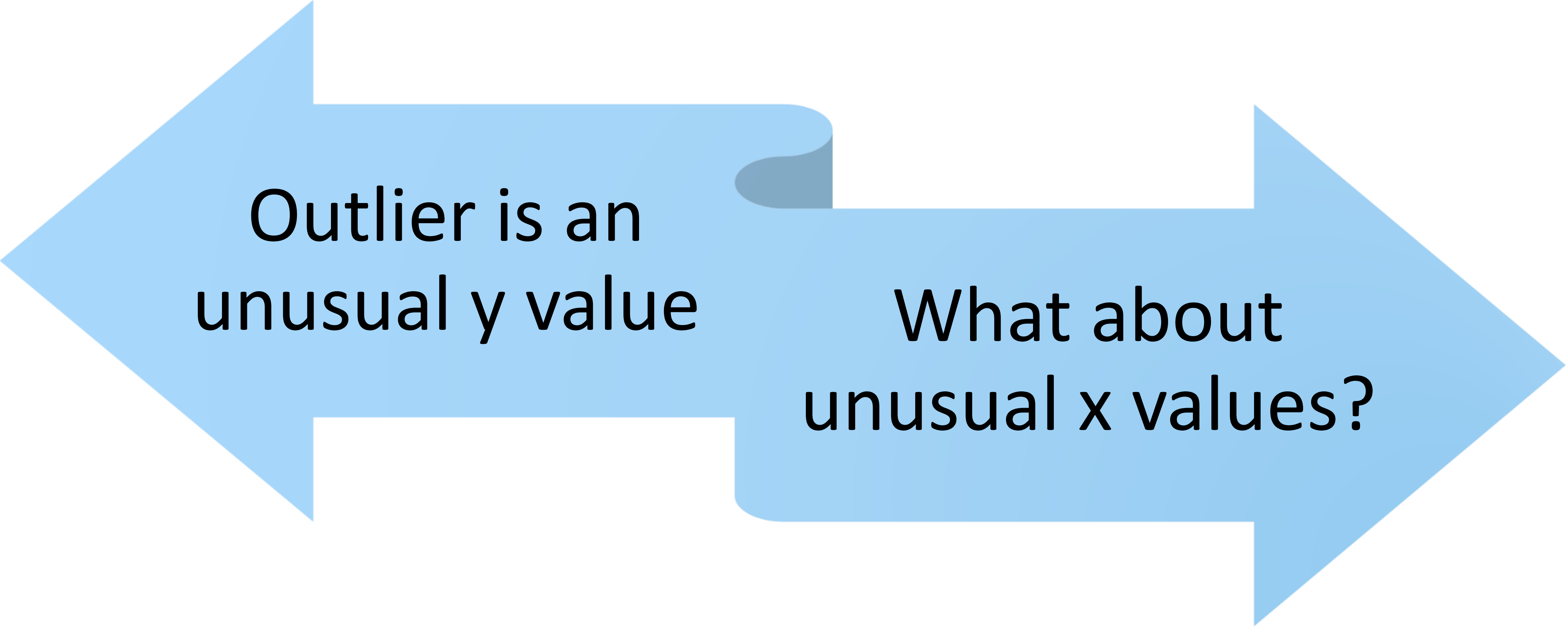
- Repeat model with and without observation

Not Normal

- Consider linear modelling with a transformation of the response variable

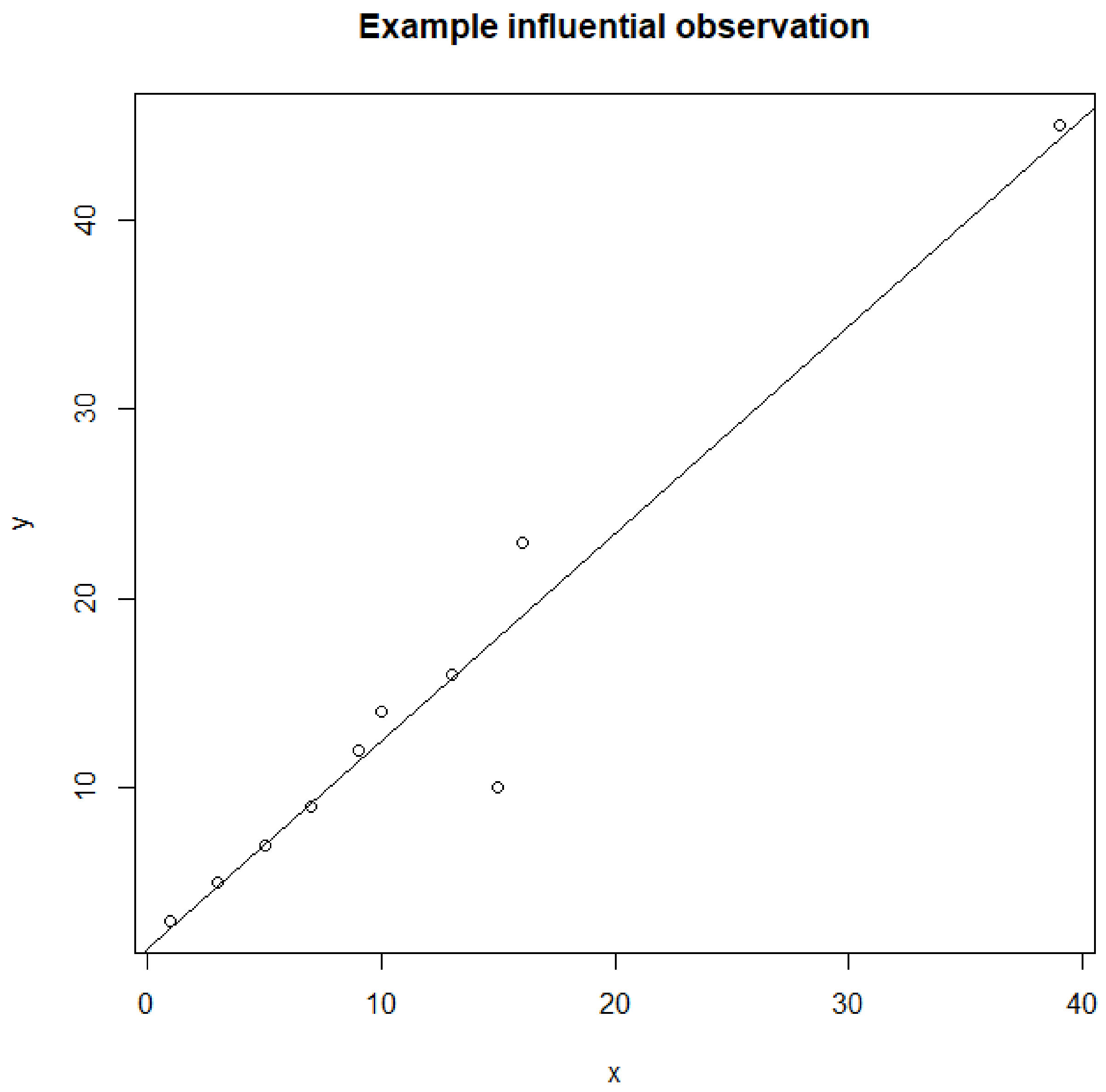
Influential Observations

Unusual *x* values



Outlier is an unusual y value

What about unusual x values?



Example influential observation

Unusual x_i value

This is different to the outlier problem

These observations are not ones we necessary want to remove from the model

- but it is good to know they are there
- and what effect they are having on the model output
- this will become an even greater issue when we consider Multiple Linear Regression models later in the module

For now we will look at how to detect so-called ***influential observations***

Recall our calculation of standardised residuals back in week 2 and 3

Because the variance and covariance of the residuals in the fitted model (e_i) do not behave in the same way as the error term in the model specification (ε_i)

It is sometimes better to work with *standardised residuals* which have

- variance closer to σ^2
- covariances closer to zero

The standardised residuals are usually written d_i

Standardised residuals

The standardised residuals are given by

$$d_i = \frac{e_i}{[s^2(1 - v_i)]^{\frac{1}{2}}}$$

where

$$v_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

We never really said anything about this quantity v_i at the time

Leverage

v_i is known as the *leverage* of an observation

$$v_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Now

$\sum_i v_i = 2$

Because each of the 2 terms in v_i sum to 1 over the *n* observations

Which means that the average leverage for an observation is $\frac{2}{n}$

What is high leverage?

Average leverage for an observation is $\frac{2}{n}$

- Leverage > $\frac{4}{n}$ (twice average) is “large leverage”
- Leverage > $\frac{6}{n}$ (three times average) is “very large leverage”

What does this mean for our model?

Large (or very large) leverage observations:

- ❑ are “influential”
- ❑ whether they are included or not causes a large change in the β parameters
- ❑ we can measure this influence using *Cook’s Statistic*
- ❑ which is usually designated D_i
- ❑ this compares the linear regression results with and without the influential observation

Cook’s Statistic

For observation *i* where *i* = 1, 2, .. n from our (x_i , y_i) observations

- first complete the linear regression as usual to obtain $\widehat{\beta_0}, \widehat{\beta_1}$ and hence the fitted $\hat{y}$ values
- then take out the one i^{th} observation
- repeat the linear regression to get new $\widehat{\beta_0}, \widehat{\beta_1}$ and hence new fitted values

which we will call $\hat{y}^{(i)}$

Cook's Statistic

Then Cook's Statistic for this i^{th} observation is

$$D_i = \frac{1}{2S^2} \sum_{j=1}^{n} (\hat{y}_j^{(i)} - \hat{y}_j)^2$$

Where there will be a separate value for D_i for each of our n observations

Now it can be shown that this statistic is related to the leverage v_i of the same observation

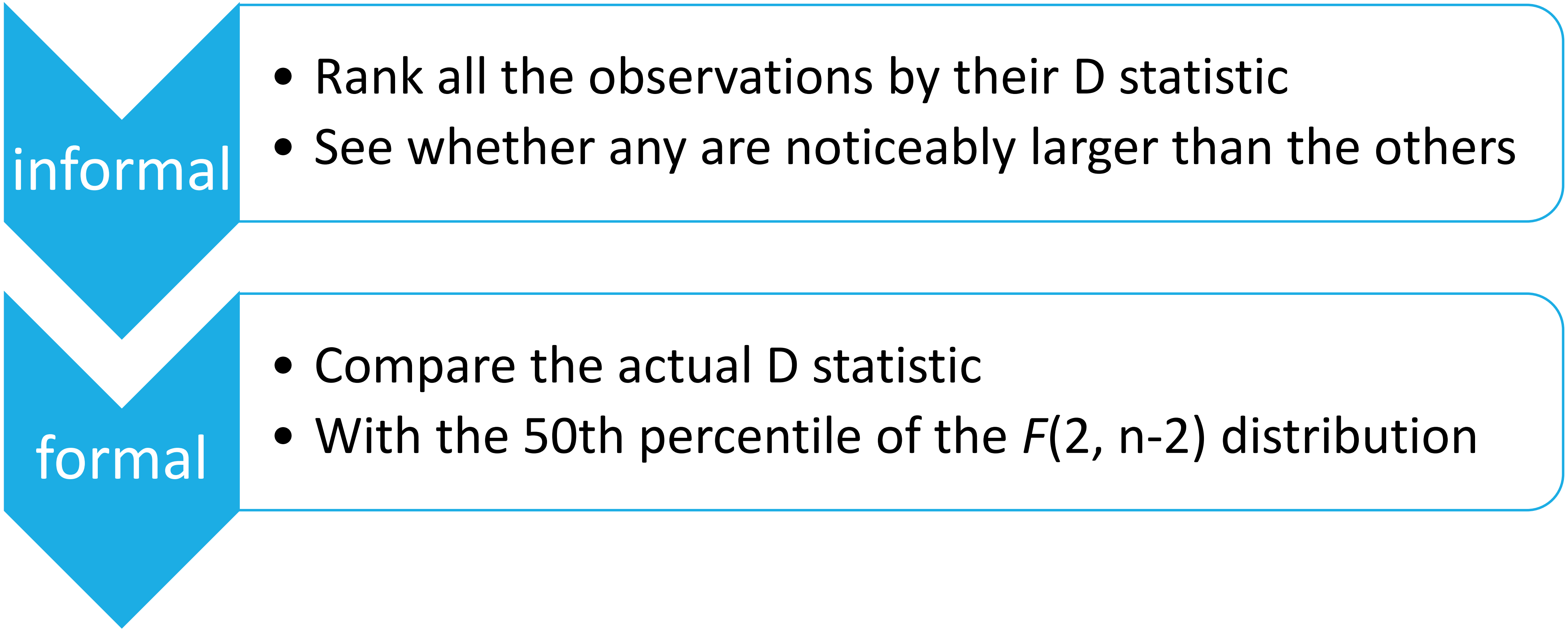
Cook and Leverage

$$D_i = \frac{1}{2} d_i^2 \frac{v_i}{1 - v_i}$$

So Cook's statistic depends on

- the standardised residual for an observation
- and its leverage

Using Cook’s Statistic



informal

- Rank all the observations by their D statistic
- See whether any are noticeably larger than the others

formal

- Compare the actual D statistic
- With the 50th percentile of the *F*(2, n-2) distribution

What to do

We don’t need to remove influential observations in same way as outliers

But when we present the results of a modelling study that includes influential observations we should

- highlight the observation(s)
- indicate how much they have affected the model output and conclusions

Transforming the response variable

Remember the residual plots in weeks 2 & 3

d_i against x_i	• Check whether a linear model is appropriate • Check the Normal assumptions
d_i *against* $\hat{y}_i$	• Check for constant variance • Called homoscedasticity
QQ plot in R	• Good first indication of Normal residuals • Looking for a straight line

What should we do if one or more of these plots shows an issue?

Transforming the response variable

If we doubt the x ➡ y relationship is linear

Or we doubt the variance of y is constant

Or we doubt the data is from a Normal distribution

Then good first thing to try is a simple transformation of the y_i

The most usual transformation (if no negative data) is $\ln y$

Common transformations

$\ln y$	where var(Y) is proportional to E(Y)^2
$\sqrt{y}$	where var(Y) is proportional to E(Y), often useful when the data is a count
$sin^{-1}(\sqrt{y})$	often useful if the data is proportions
$1/y$	