CWRU DSCI351-351M-453: Week00a

Roger H. French, JiQi Liu

01 November, 2018

Contents

11.1.1.1 Reading, Homeworks, Projects, SemProjects

-

11.2 Homework:

  -

-

11.3 Readings:

  -

-

11.4 Projects: We will have four 2 week EDA projects

-

11.5 451 SemProjects:

- –

- Final Exam

    - Monday December 17th, 12 noon to 3pm, Olin 313

11.5.0.1 Textbooks

- Peng: R Programming for Data Science
- Peng: Exploratory Data Analysis with R
- Open Intro Stats, v3
- Wickham: R for Data Science
- Hastie: Intro to Statistical Learning with R

11.5.0.2 Syllabus

11.5.0.3 Anscombe's Quartet of 'Identical' Simple Linear Regressions

Visualization may not be as precise as statistics,

- but it provides a unique view onto data
    - that can make it much easier to discover
    - interesting structures than numerical methods.
- Visualization also provides the context necessary
    - to make better choices
    - and to be more careful when fitting models.

Anscombe's Quartet is a case in point,

- showing that four datasets
    - that have identical statistical properties (i.e. summary statistics)
    - can indeed be very different.

11.5.0.3.1 Arguing for Graphics in 1973

In 1973, Francis J. Anscombe

- published a paper titled, Graphs in Statistical Analysis.
- The idea of using graphical methods
    - had been established relatively recently by John Tukey,
    - but there was evidently still a lot of skepticism.
- Anscombe first lists some notions
    - that textbooks were "indoctrinating" people with,
    - like the idea that "numerical calculations are exact,
    - but graphs are rough."

He then presents a table of numbers.

- It contains four distinct datasets (hence the name Anscombe's Quartet),
- each with statistical properties that are essentially identical:
    - the mean of the x values is 9.0,
    - mean of y values is 7.5,
- they all have nearly identical
    - variances,
    - correlations,

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1,**
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

Figure 1: DSCI353-453 Syllabus

– and regression lines (to at least two decimal places).

```
knitr::kable(anscombe)
```

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.10	5.39	12.50
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

11.5.0.4 Let's do the simple descriptive statistics on each data set

11.5.0.4.1 Here is mean of x and y

```
anscombe.1 <- data.frame(x = anscombe[["x1"]], y = anscombe[["y1"]], Set = "Anscombe Set 1")
anscombe.2 <- data.frame(x = anscombe[["x2"]], y = anscombe[["y2"]], Set = "Anscombe Set 2")
anscombe.3 <- data.frame(x = anscombe[["x3"]], y = anscombe[["y3"]], Set = "Anscombe Set 3")
anscombe.4 <- data.frame(x = anscombe[["x4"]], y = anscombe[["y4"]], Set = "Anscombe Set 4")

anscombe.data <- rbind(anscombe.1, anscombe.2, anscombe.3, anscombe.4)
aggregate(cbind(x, y) ~ Set, anscombe.data, mean)
```

```
##              Set x        y
## 1 Anscombe Set 1 9 7.500909
## 2 Anscombe Set 2 9 7.500909
## 3 Anscombe Set 3 9 7.500000
## 4 Anscombe Set 4 9 7.500909
```

11.5.0.5 And SD

```
aggregate(cbind(x, y) ~ Set, anscombe.data, sd)
```

```
##              Set        x        y
## 1 Anscombe Set 1 3.316625 2.031568
## 2 Anscombe Set 2 3.316625 2.031657
## 3 Anscombe Set 3 3.316625 2.030424
## 4 Anscombe Set 4 3.316625 2.030579
```

11.5.0.5.1 And correlation between x and y

```
library(plyr)

correlation <- function(data) {
    x <- data.frame(r = cor(data$x, data$y))
    return(x)
}
```

```r
ddply(.data = anscombe.data, .variables = "Set", .fun = correlation)
```

```
##              Set         r
## 1 Anscombe Set 1 0.8164205
## 2 Anscombe Set 2 0.8162365
## 3 Anscombe Set 3 0.8162867
## 4 Anscombe Set 4 0.8165214
```

As can be seen

- they are pretty much the same
- for every data set.

11.5.0.6 Let's perform linear regression model for each

```r
model1 <- lm(y ~ x, subset(anscombe.data, Set == "Anscombe Set 1"))
model2 <- lm(y ~ x, subset(anscombe.data, Set == "Anscombe Set 2"))
model3 <- lm(y ~ x, subset(anscombe.data, Set == "Anscombe Set 3"))
model4 <- lm(y ~ x, subset(anscombe.data, Set == "Anscombe Set 4"))
```

11.5.0.6.1 Here are the summaries

```r
summary(model1)
```

```
## 
## Call:
## lm(formula = y ~ x, data = subset(anscombe.data, Set == "Anscombe Set 1"))
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.92127 -0.45577 -0.04136  0.70941  1.83882 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   3.0001     1.1247   2.667  0.02573 * 
## x             0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295 
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```r
summary(model2)
```

```
## 
## Call:
## lm(formula = y ~ x, data = subset(anscombe.data, Set == "Anscombe Set 2"))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.9009 -0.7609  0.1291  0.9491  1.2691 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    3.001      1.125   2.667  0.02576 *
## x              0.500      0.118   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

```r
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ x, data = subset(anscombe.data, Set == "Anscombe Set 3"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1586 -0.6146 -0.2303  0.1540  3.2411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0025     1.1245   2.670  0.02562 *
## x             0.4997     0.1179   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

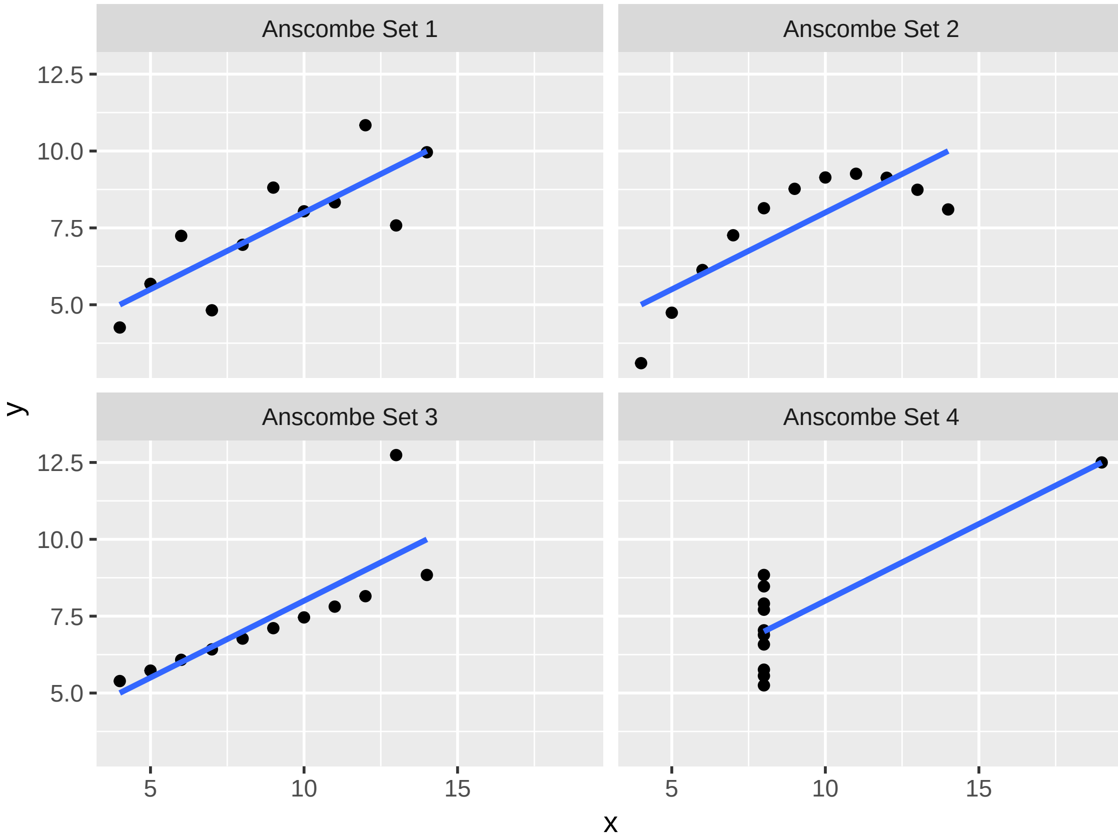
```r
summary(model4)
```

```
##
## Call:
## lm(formula = y ~ x, data = subset(anscombe.data, Set == "Anscombe Set 4"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0017     1.1239   2.671  0.02559 *
## x             0.4999     0.1178   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

11.5.0.7 Now, do what you should have done in the first place: EDA PLOTS

```r
library(ggplot2)
```

```r
gg <- ggplot(anscombe.data, aes(x = x, y = y))
gg <- gg + geom_point(color = "black")
gg <- gg + facet_wrap(~Set, ncol = 2)
gg <- gg + geom_smooth(formula = y ~ x, method = "lm", se = FALSE, data = anscombe.data)
gg
```



11.5.0.7.1 Review each dataset

While dataset I

- appears like many well-behaved datasets
  - that have clean and well-fitting linear models,
- the others are not served nearly as well.

Dataset II does not have a linear correlation;

Dataset III does,

- but the linear regression is thrown off by an outlier.
- It would be easy to fit a correct linear model,
  - if only the outlier were spotted
  - and removed before doing so.

Dataset IV, finally,

- does not fit any kind of linear model,
- but the single outlier keeps the alarm from going off.

11.5.0.7.2 How do you find out which model can be applied?

Anscombe's answer is to use graphs:

- looking at the data immediately reveals a lot of the structure,
  - and makes the analyst aware of "pathological" cases like dataset IV.
- Computers are not limited to running numerical models, either.

A computer should make both calculations and graphs.

- Both sorts of output should be studied;
- each will contribute to understanding.

11.5.0.8 What is an Outlier?

In addition to showing how useful a clear look onto data can be,

Anscombe also raises an interesting question:

- what, exactly, is an outlier?
- He describes a study on education,
  - where he studied per-capita expenditures for public schools
  - in the 50 U.S. states and the District of Columbia.
- Alaska is a bit of an outlier,
  - so it moves the regression line away from the mainstream.
- The obvious response would be to remove Alaska from the data
  - before computing the regression.
- But then, another state will be an outlier.
- Where do you stop?

Anscombe argues that the correct answer

- is to show both the regression with Alaska,
- but also how much it contributes
  - and what happens when it is removed.

The tool here, again, are graphical representations.

- Not only the actual data needs to be shown,
  - but also the distances from the regression line (the residuals),
  - and other statistics that help judge how well the model fits.
- It seems like an obvious thing to do,
  - but presumably was not the norm in the 1970s,
  - and I can imagine that it still not always is.

It can be seen both graphically

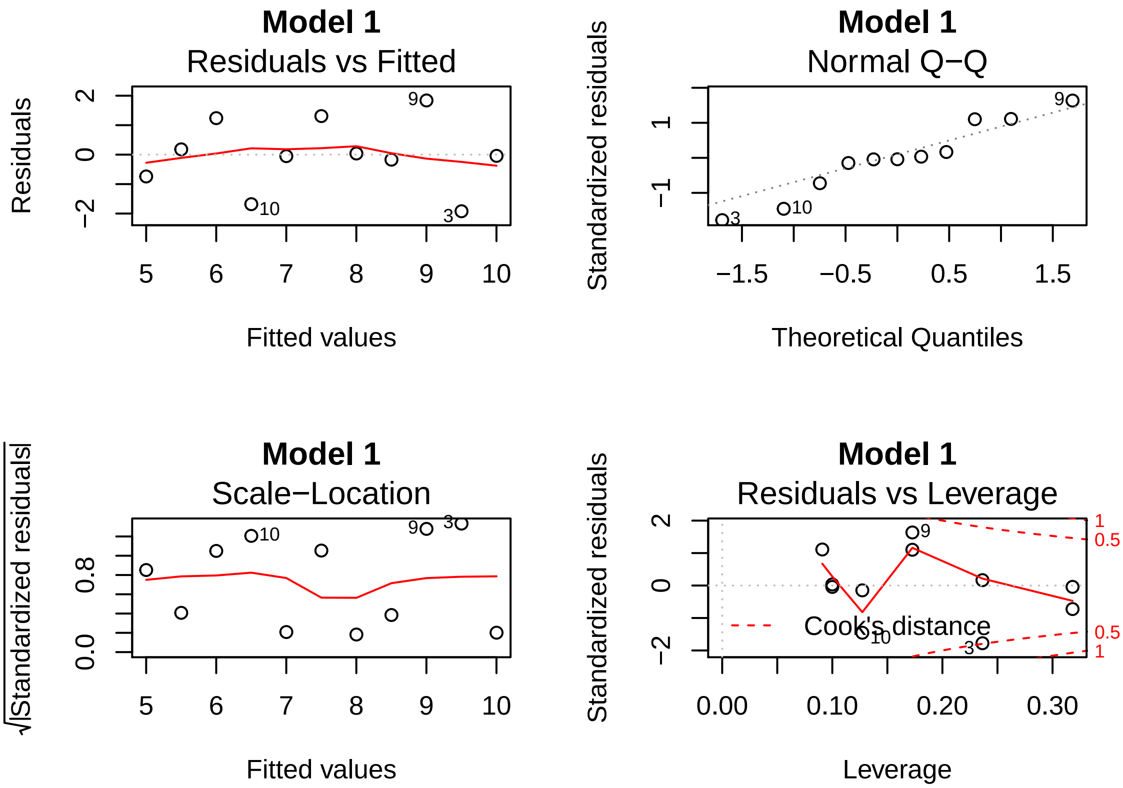
- and from regression summary
  - that each data set resulted in same statistical model!
- Intercepts,
- coeficients
  - and their p values are the same.
- SEE (standard error of the estimate, or SD of residuals),
- F-value -and it's p values
- are the same.

11.5.0.9 Conclusion:

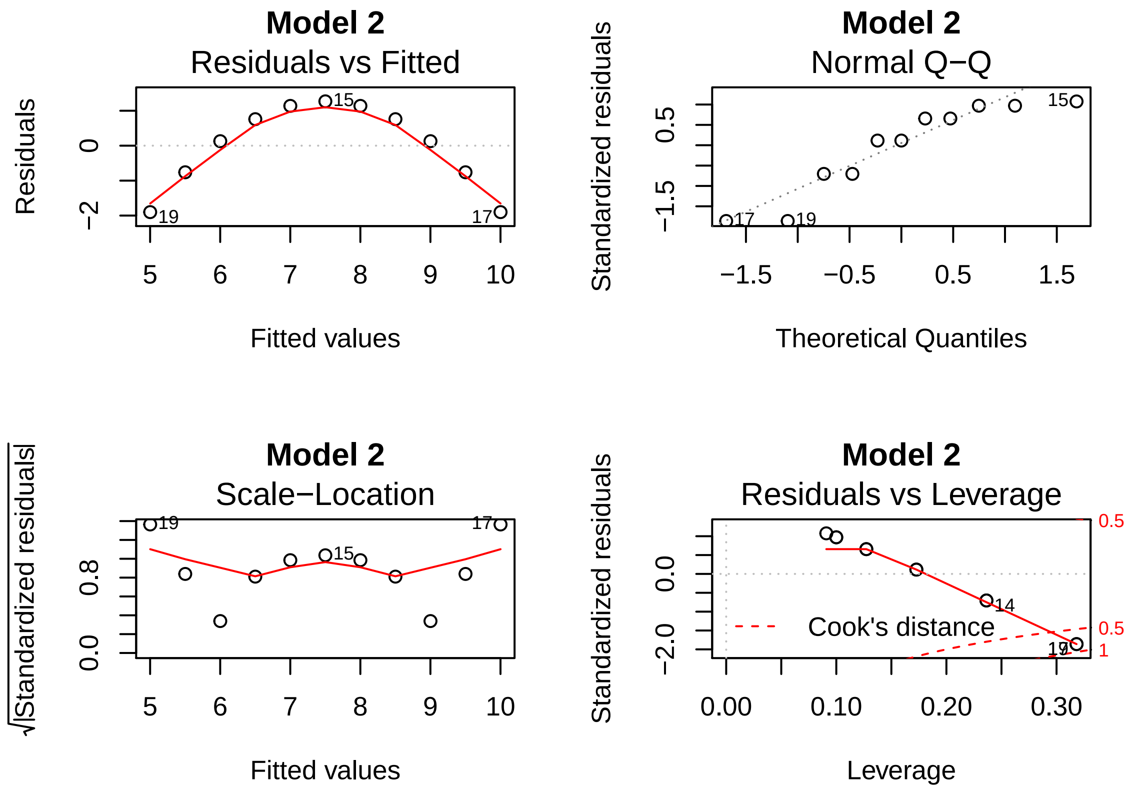
ALWAYS plot your data!

- And always do model diagnostics by plotting the residuals.

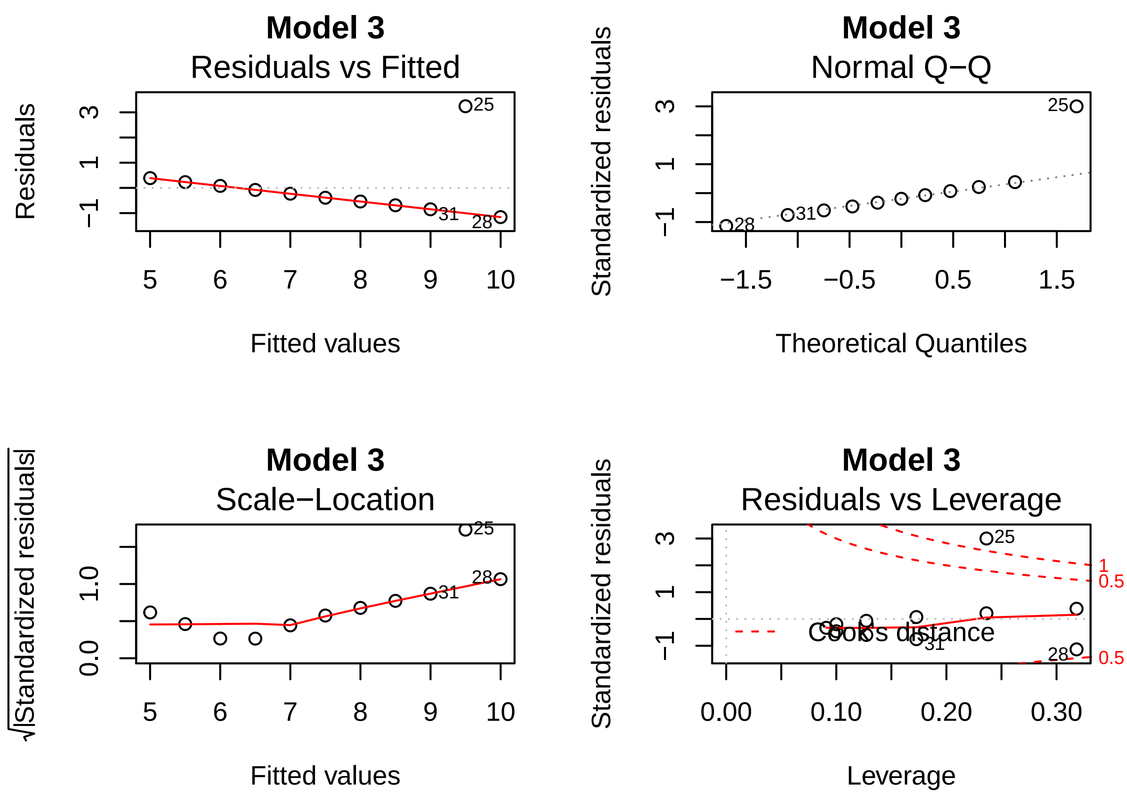
```
par(mfrow = c(2, 2))
plot(model1, main = "Model 1")
```




```
plot(model2, main = "Model 2")
```



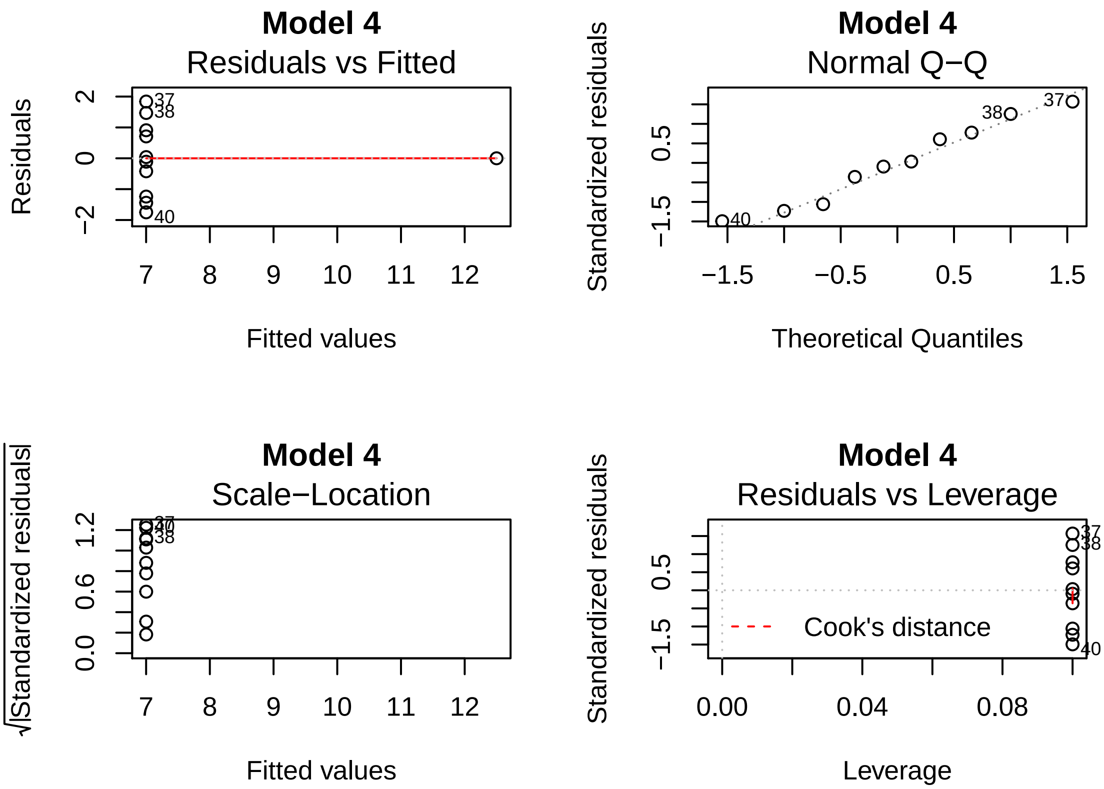

```r
plot(model3, main = "Model 3")
```



```r
plot(model4, main = "Model 4")
```

```
## Warning: not plotting observations with leverage one:
##   8

## Warning: not plotting observations with leverage one:
##   8
```



References:

Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.

What is Anscombe's Quartet and why is it important? - by Mladen Jovanovic

Anscombe's Quartet - by Robert Kosara

11.5.0.10 Links

[Anscombe's Quartet of 'Identical' Simple Linear Regressions]

11.5.0.11 References