

# Which marketing action does it?

Data inspection, a little something about R,  
linear regression and problems with multicollinearities

Inholland University of Applied Sciences  
International Week 2014

Stefan Etschberger  
Augsburg University of Applied Sciences

# Who is talking to you?

- ▶ Stefan Etschberger



## Who is talking to you?

- ▶ Stefan Etschberger
- ▶ University degree in mathematics and physics



## Who is talking to you?

- ▶ Stefan Etschberger
- ▶ University degree in mathematics and physics
- ▶ Worked as an engineer in semiconductor industry



## Who is talking to you?

- ▶ Stefan Etschberger
- ▶ University degree in mathematics and physics
- ▶ Worked as an engineer in semiconductor industry
- ▶ Back to university as a researcher: doctoral degree in economic science



## Who is talking to you?

- ▶ Stefan Etschberger
- ▶ University degree in mathematics and physics
- ▶ Worked as an engineer in semiconductor industry
- ▶ Back to university as a researcher: doctoral degree in economic science
- ▶ Research focus: marketing research using data analysis



## Who is talking to you?

- ▶ Stefan Etschberger
- ▶ University degree in mathematics and physics
- ▶ Worked as an engineer in semiconductor industry
- ▶ Back to university as a researcher: doctoral degree in economic science
  - ▶ Research focus: marketing research using data analysis
  - ▶ Professor of Mathematics and Statistics since 2006



## Who is talking to you?

- ▶ Stefan Etschberger
- ▶ University degree in mathematics and physics
- ▶ Worked as an engineer in semiconductor industry
- ▶ Back to university as a researcher: doctoral degree in economic science
  - ▶ Research focus: marketing research using data analysis
  - ▶ Professor of Mathematics and Statistics since 2006
  - ▶ at University of Applied Sciences Augsburg since 2012





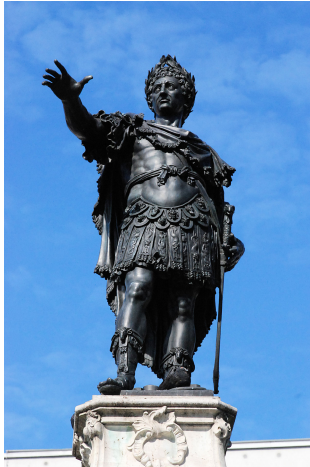
# Where am I from?

## ► City of Augsburg



## Where am I from?

- ▶ City of Augsburg
- ▶ Almost (OK, 2nd place) oldest city in Germany (15 b.C.)



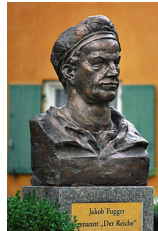
## Where am I from?

- ▶ City of Augsburg
- ▶ Almost (OK, 2nd place) oldest city in Germany (15 b.C.)
- ▶ Famous for its renaissance architecture



# Where am I from?

- ▶ City of Augsburg
- ▶ Almost (OK, 2nd place) oldest city in Germany (15 b.C.)
- ▶ Famous for its renaissance architecture
- ▶ and the oldest social housing project in the world (1521)



## Where am I from?

- ▶ City of Augsburg
- ▶ Almost (OK, 2nd place) oldest city in Germany (15 b.C.)
- ▶ Famous for its renaissance architecture
- ▶ and the oldest social housing project in the world (1521)
- ▶ A lot of university students (25.000)



## Where am I from?

- ▶ City of Augsburg
- ▶ Almost (OK, 2nd place) oldest city in Germany (15 b.C.)
- ▶ Famous for its renaissance architecture
- ▶ and the oldest social housing project in the world (1521)
- ▶ A lot of university students (25.000)
- ▶ And a business school at the Augsburg University of Applied Science



# Data analysis, Regression and Beyond: Table of Contents

- 1 Introduction
- 2 R and RStudio
- 3 Revision: Simple linear regression
- 4 Multicollinearity in Regression



- 1 Introduction
  - Mr. Maier and his cheese
  - Mr. Maier and his data



- ▶ After his bachelor's degree in marketing Mr. Maier took over a respectable cheese dairy in Bavaria



## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides





- ▶ After his bachelor's degree in marketing Mr. Maier took over a respectable cheese dairy in Bavaria
- ▶ Regularly he does marketing focused on distinct towns



## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

## R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides

- ▶ After his bachelor's degree in marketing Mr. Maier took over a respectable cheese dairy in Bavaria
- ▶ Regularly he does marketing focused on distinct towns
- ▶ He uses the phone, e-mail, mail and small gifts for his key customers



## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides

- ▶ After his bachelor's degree in marketing Mr. Maier took over a respectable cheese dairy in Bavaria
- ▶ Regularly he does marketing focused on distinct towns
- ▶ He uses the phone, e-mail, mail and small gifts for his key customers
- ▶ And he collected data about his spendings per marketing action and his revenues for 30 days after the action took place



action	revenue	telephone	e-mail	mail	gift
1	10193.70	186.20	158.60	26.90	11.10
2	4828.20	470.30	55.00	14.40	20.30
3	11139.30	41.80	154.70	20.90	12.40
4	5030.10	530.10	79.80	21.70	17.00
⋮					

## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

## R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



action	revenue	telephone	e-mail	mail	gift
1	10193.70	186.20	158.60	26.90	11.10
2	4828.20	470.30	55.00	14.40	20.30
3	11139.30	41.80	154.70	20.90	12.40
4	5030.10	530.10	79.80	21.70	17.00
⋮					

- ▶ Goal: Getting to know interesting structure hidden inside data

## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

## R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



action	revenue	telephone	e-mail	mail	gift
1	10193.70	186.20	158.60	26.90	11.10
2	4828.20	470.30	55.00	14.40	20.30
3	11139.30	41.80	154.70	20.90	12.40
4	5030.10	530.10	79.80	21.70	17.00
⋮					

- ▶ Goal: Getting to know interesting structure hidden inside data
- ▶ Maybe: Forecast of his revenue as a model dependent of the spendings for his marketing actions

## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

## R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



action	revenue	telephone	e-mail	mail	gift
1	10193.70	186.20	158.60	26.90	11.10
2	4828.20	470.30	55.00	14.40	20.30
3	11139.30	41.80	154.70	20.90	12.40
4	5030.10	530.10	79.80	21.70	17.00
⋮					

- ▶ Goal: Getting to know interesting structure hidden inside data
- ▶ Maybe: Forecast of his revenue as a model dependent of the spendings for his marketing actions
- ▶ Data has been sent the data from his external advertising service provider inside an Excel-file.

## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

## R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



action	revenue	telephone	e-mail	mail	gift
1	10193.70	186.20	158.60	26.90	11.10
2	4828.20	470.30	55.00	14.40	20.30
3	11139.30	41.80	154.70	20.90	12.40
4	5030.10	530.10	79.80	21.70	17.00
⋮					

- ▶ Goal: Getting to know interesting structure hidden inside data
- ▶ Maybe: Forecast of his revenue as a model dependent of the spendings for his marketing actions
- ▶ Data has been sent the data from his external advertising service provider inside an Excel-file.
- ▶ Mr. Maier runs his data analysis software....

## Introduction

Mr. Maier and his cheese

Mr. Maier and his data

## R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



# Data analysis, Regression and Beyond: Table of Contents

- 1 Introduction
- 2 R and RStudio
- 3 Revision: Simple linear regression
- 4 Multicollinearity in Regression



- 2 R and RStudio  
What is R?  
What is RStudio?  
First steps

# What is R and why R?

- ▶ R is a **free** Data Analysis Software



[Introduction](#)

[R and RStudio](#)

[What is R?](#)

[What is RStudio?](#)

[First steps](#)

[Simple linear  
regression](#)

[Multicollinearity](#)

[Supplementary slides](#)

# What is R and why R?

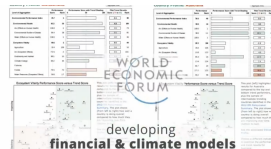
- ▶ R is a **free** Data Analysis Software
- ▶ R is very powerful and **widely used** in science and industry (in fact far more widely than SPSS)



- The average data miner reports using 4 software tools.
- R is used by the most data miners (47%).
- STATISTICA is the primary data mining tool chosen most often (17%).



source: <http://goo.gl/axhGhh>



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

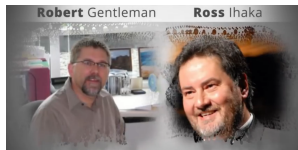
Simple linear  
regression

Multicollinearity

Supplementary slides

# What is R and why R?

- ▶ R is a **free** Data Analysis Software
- ▶ R is very powerful and **widely used** in science and industry (in fact far more widely than SPSS)
- ▶ **Created in 1993** at the University of Auckland by Ross Ihaka and Robert Gentleman



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

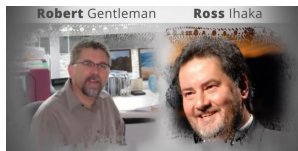
Simple linear  
regression

Multicollinearity

Supplementary slides

# What is R and why R?

- ▶ R is a **free** Data Analysis Software
- ▶ R is very powerful and **widely used** in science and industry (in fact far more widely than SPSS)
- ▶ **Created in 1993** at the University of Auckland by Ross Ihaka and Robert Gentleman
- ▶ Since then: A lot of people improved the software and wrote **thousands of packages** for lots of applications



## Introduction

### R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

Multicollinearity

Supplementary slides

# What is R and why R?

- ▶ R is a **free** Data Analysis Software
- ▶ R is very powerful and **widely used** in science and industry (in fact far more widely than SPSS)
- ▶ **Created in 1993** at the University of Auckland  
by Ross Ihaka and Robert Gentleman
- ▶ Since then: A lot of people improved the software and wrote **thousands of packages** for lots of applications
- ▶ Drawback (at first glance): No point and click tool

```
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+           data=diamonds, color=clarity,
+           xlab="Carat", ylab="Price",
+           main="Diamond Pricing")
> format.plot(p, size=24)
> |
```



## Introduction

### R and RStudio

#### What is R?

#### What is RStudio?

#### First steps

#### Simple linear regression

#### Multicollinearity

#### Supplementary slides



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

Multicollinearity

Supplementary slides

- ▶ R is a **free** Data Analysis Software
- ▶ R is very powerful and **widely used** in science and industry (in fact far more widely than SPSS)
- ▶ **Created in 1993** at the University of Auckland by Ross Ihaka and Robert Gentleman
- ▶ Since then: A lot of people improved the software and wrote **thousands of packages** for lots of applications
- ▶ Drawback (at first glance): No point and click tool
- ▶ Major advantage (at second thought): No point and click tool

- ▶ RStudio is a **Integrated Development Environment** (IDE) for using R.



[Introduction](#)

[R and RStudio](#)

[What is R?](#)

[What is RStudio?](#)

[First steps](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Supplementary slides](#)



- ▶ RStudio is a **Integrated Development Environment** (IDE) for using R.
- ▶ Works on OSX, Linux and Windows



Free & Open-Source IDE for R



[Introduction](#)

[R and RStudio](#)

[What is R?](#)

[What is RStudio?](#)

[First steps](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Supplementary slides](#)

- ▶ RStudio is a **Integrated Development Environment** (IDE) for using R.
- ▶ Works on OSX, Linux and Windows
- ▶ It's free as well



Free & Open-Source IDE for R



[Introduction](#)

[R and RStudio](#)

[What is R?](#)

[What is RStudio?](#)

[First steps](#)

[Simple linear regression](#)

[Multicollinearity](#)

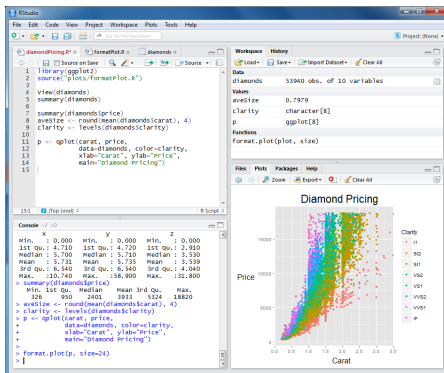
[Supplementary slides](#)

# What is RStudio?

- ▶ RStudio is a **Integrated Development Environment (IDE)** for using R.
- ▶ Works on OSX, Linux and Windows
- ▶ It's free as well
- ▶ Still: You have to write commands



Free & Open-Source IDE for R



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

Multicollinearity

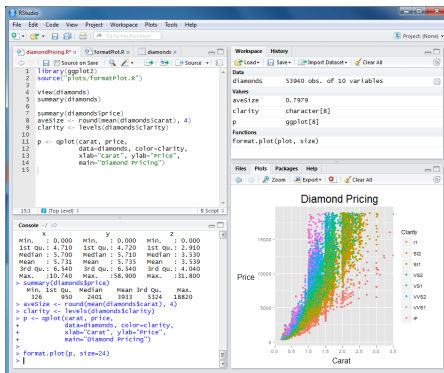
Supplementary slides

# What is RStudio?

- ▶ RStudio is a **Integrated Development Environment (IDE)** for using R.
- ▶ Works on OSX, Linux and Windows
- ▶ It's free as well
- ▶ Still: You have to write commands
- ▶ But: RStudio supports you a lot



Free & Open-Source IDE for R



Data analysis,  
Regression and  
Beyond  
Stefan Etschberger



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

Multicollinearity

Supplementary slides



## Getting to know RStudio

### ► Code

The screenshot shows the RStudio environment with the following components:

- Source Editor:** Contains R code for loading the 'diamonds' dataset, summarizing it, and creating a scatter plot titled 'Diamond Pricing' where 'Price' is on the y-axis and 'Carat' is on the x-axis. Points are colored by 'Clarity'.
- Console:** Shows the output of the R commands, including summary statistics for 'diamonds' and 'diamonds\$price'.
- Workspace:** Lists the loaded data object 'diamonds' (53940 observations) and the 'p' plot object.
- Plots:** Displays the 'Diamond Pricing' scatter plot with a legend for Clarity levels: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

```
1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 view(diamonds)
5 summary(diamonds)
6
7 summary(diamonds$price)
8 aversez <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
10
11
12 p <- ggplot(carat, price,
13            data=diamonds, color=clarity,
14            xlab="Carat", ylab="Price",
15            main="Diamond Pricing")
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

```
Summary statistics for diamonds:
  Min. : 0.000  Min. y : 0.000  Min. z : 0.000
  1st Qu.: 4.710  1st Qu.: 4.720  1st Qu.: 2.910
  Median : 5.700  Median : 5.710  Median : 3.530
  Mean   : 5.731  Mean   : 5.735  Mean   : 3.539
  3rd Qu.: 6.540  3rd Qu.: 6.540  3rd Qu.: 4.040
  Max.   :10.740  Max.   :18.900  Max.   :31.800

Summary statistics for diamonds$price:
  Min. 1st Qu.  Median  Mean 3rd Qu.  Max.
  326   950   2403   3933  5324 18820

R commands:
> aversez <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- ggplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
> format.plot(p, size=24)
> |
```

Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

Multicollinearity

Supplementary slides



## Getting to know RStudio

- ▶ Code
- ▶ Console

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading the 'diamonds' dataset, summarizing it, and creating a scatter plot of Price vs. Carat, faceted by clarity.
- Console:** Shows the execution of the code, including summary statistics for 'diamonds' and 'diamonds\$price', and the execution of the plotting functions.
- Workspace:** Lists the loaded data (diamonds) and the created plot object (p).
- Plots Panel:** Displays a scatter plot titled 'Diamond Pricing' showing Price on the y-axis (log scale) and Carat on the x-axis. Points are colored by clarity, with a legend on the right showing categories: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear regression

Multicollinearity

Supplementary slides



## Getting to know RStudio

- ▶ Code
- ▶ Console
- ▶ Workspace

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading the 'diamonds' dataset, summarizing it, and creating a scatter plot of Price vs. Carat, faceted by clarity.
- Console:** Shows the execution of the code, including summary statistics for 'diamonds' and 'diamonds\$price', and the successful execution of the plot command.
- Workspace:** Lists the loaded data object 'diamonds' (53940 observations) and the plot object 'p'.
- Plots Panel:** Displays a scatter plot titled 'Diamond Pricing' showing Price on the y-axis (0 to 15000) and Carat on the x-axis (0.0 to 3.5). Points are colored by clarity, with a legend on the right showing categories: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear regression

Multicollinearity

Supplementary slides



## Getting to know RStudio

- ▶ Code
- ▶ Console
- ▶ Workspace
- ▶ History

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading data, summarizing it, and creating a scatter plot. The code is:

```
1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 view(diamonds)
5 summary(diamonds)
6
7 view(diamonds$price)
8 aversez <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
10
11 p <- ggplot(carat, price,
12            data=diamonds, color=clarity,
13            xlab="Carat", ylab="Price",
14            main="Diamond Pricing")
15
```
- Console:** Shows the execution of the code, including summary statistics for the 'diamonds' dataset and the 'price' variable.

```
15.1 | (Top Level) | R Script |
-----|-----|-----|
Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
Median : 5.700   Median : 5.710   Median : 3.530
Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
Max.   :10.740   Max.   :18.900   Max.   :31.800
> summary(diamonds$price)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 326    950    2403    3933    5324   18820
> aversez <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- ggplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
> |
```
- Workspace:** Lists the loaded data (diamonds) and the created plot object (p).
- History:** Shows the sequence of executed commands.
- Plots:** Displays a scatter plot titled "Diamond Pricing" showing Price (y-axis, 0 to 15000) versus Carat (x-axis, 0.0 to 3.5). The plot is faceted by clarity, with a legend on the right showing colors for I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

Multicollinearity

Supplementary slides





## Getting to know RStudio

- ▶ Code
- ▶ Console
- ▶ Workspace
- ▶ History
- ▶ Files

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading data, summarizing it, and creating a scatter plot. The code is:

```
1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 view(diamonds)
5 summary(diamonds)
6
7 view(diamonds$price)
8 averseize <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
10
11 p <- ggplot(carat, price,
12            data=diamonds, color=clarity,
13            xlab="Carat", ylab="Price",
14            main="Diamond Pricing")
15
```
- Console:** Shows the execution of the code, including summary statistics for the 'diamonds' dataset and the 'price' variable, and the successful execution of the plot command.

```
Min. x: 0.000 Min. y: 0.000 Min. z: 0.000
1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910
Median: 5.700 Median: 5.710 Median: 3.530
Mean : 5.731 Mean : 5.735 Mean : 3.539
3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040
Max. :10.740 Max. :18.900 Max. :31.800
> summary(diamonds$price)
Min. 1st Qu. Median Mean 3rd Qu. Max.
326 890 2403 3933 5324 18820
> averseize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- ggplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
>
```
- Workspace:** Lists the loaded data (diamonds) and the created plot object (p).
- History:** Shows the sequence of commands executed in the session.
- Plots:** Displays a scatter plot titled "Diamond Pricing" showing Price (y-axis, 0 to 15000) versus Carat (x-axis, 0.0 to 3.5). The points are colored by clarity, with a legend on the right showing categories: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

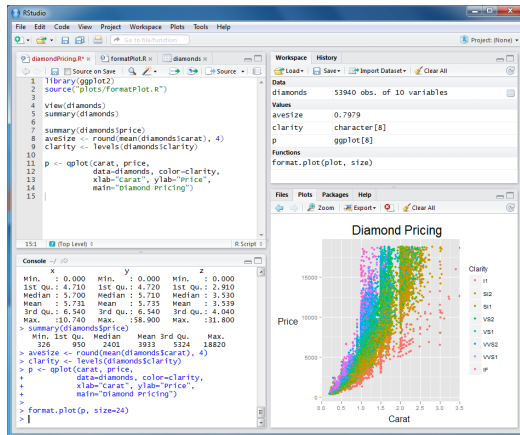
Multicollinearity

Supplementary slides



## Getting to know RStudio

- ▶ Code
- ▶ Console
- ▶ Workspace
- ▶ History
- ▶ Files
- ▶ Plots



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

Multicollinearity

Supplementary slides



## Getting to know RStudio

- ▶ Code
- ▶ Console
- ▶ Workspace
- ▶ History
- ▶ Files
- ▶ Plots
- ▶ Packages

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading data, summarizing it, and creating a scatter plot.
- Console:** Shows the execution output of the code, including summary statistics for 'diamonds\$price'.
- Workspace:** Lists the loaded data object 'diamonds' and the plot object 'p'.
- History:** Shows the sequence of executed commands.
- Plots Panel:** Displays a scatter plot titled 'Diamond Pricing' with 'Carat' on the x-axis and 'Price' on the y-axis. Points are colored by 'Clarity'.

```
1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 view(diamonds)
5 summary(diamonds)
6
7 summary(diamonds$price)
8 averseize <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
10
11 p <- ggplot(carat, price,
12            data=diamonds, color=clarity,
13            xlab="Carat", ylab="Price",
14            main="Diamond Pricing")
15
```

Console output:

```
      x      y      z
Min. : 0.000 Min. : 0.000 Min. : 0.000
1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910
Median : 5.700 Median : 5.710 Median : 3.530
Mean : 5.731 Mean : 5.735 Mean : 3.539
3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040
Max. :10.740 Max. :18.900 Max. :31.800

> summary(diamonds$price)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 326.   890   2403   3933   5324 18820

> averseize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- ggplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
>
```

Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

Multicollinearity

Supplementary slides



## Getting to know RStudio

- ▶ Code
- ▶ Console
- ▶ Workspace
- ▶ History
- ▶ Files
- ▶ Plots
- ▶ Packages
- ▶ Help

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading the 'diamonds' dataset, summarizing it, and creating a scatter plot 'p' of Price vs. Carat, faceted by Clarity.
- Console:** Shows the execution of the code, including summary statistics for 'diamonds' and 'diamonds\$price', and the creation of the plot object 'p'.
- Workspace:** Lists the loaded data object 'diamonds' (53940 observations) and the plot object 'p'.
- History:** Shows the sequence of executed commands.
- Plots Panel:** Displays the 'Diamond Pricing' scatter plot, where Price is on the y-axis (0 to 15000) and Carat is on the x-axis (0.0 to 3.5). Points are colored by Clarity, with a legend on the right showing categories: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, and IF.

### Introduction

### R and RStudio

What is R?

What is RStudio?

First steps

Simple linear regression

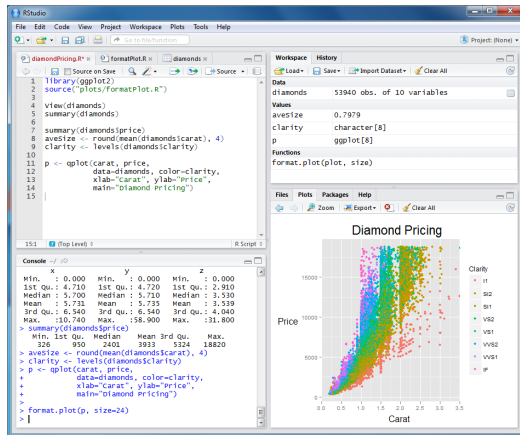
Multicollinearity

Supplementary slides



## Getting to know RStudio

- ▶ Code
- ▶ Console
- ▶ Workspace
- ▶ History
- ▶ Files
- ▶ Plots
- ▶ Packages
- ▶ Help
- ▶ Auto-Completion



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

Simple linear  
regression

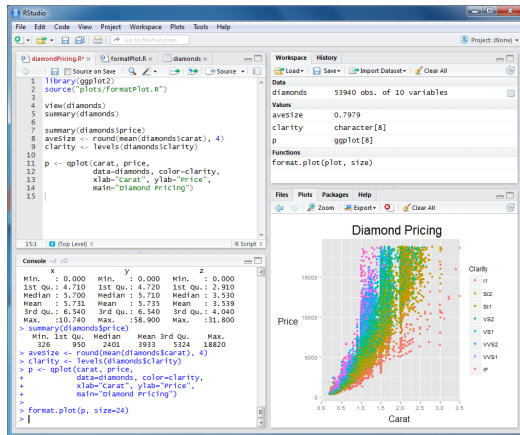
Multicollinearity

Supplementary slides



## Getting to know RStudio

- ▶ Code
- ▶ Console
- ▶ Workspace
- ▶ History
- ▶ Files
- ▶ Plots
- ▶ Packages
- ▶ Help
- ▶ Auto-Completion
- ▶ Data Import



### Introduction

### R and RStudio

What is R?

What is RStudio?

First steps

Simple linear regression

Multicollinearity

Supplementary slides



## Introduction

### R and RStudio

What is R?

What is RStudio?

First steps

### Simple linear regression

### Multicollinearity

### Supplementary slides

```
# read in data from comma-separated list
MyCheeseData = read.csv(file="Cheese.csv", header=TRUE)
# show first few lines of data matrix
head(MyCheeseData)

##   phone  gift email  mail revenue
## 1 29.36 146.1 10.32 13.36   3138
## 2  8.75 125.8 11.27 14.72   3728
## 3 36.15 124.5  8.45 17.72   3085
## 4 51.20 129.4 10.27 39.59   4668
## 5 51.36 163.4  8.19  7.57   2286
## 6 34.65 110.0  7.89 21.68   4148

# make MyCheeseData the default dataset
attach(MyCheeseData)
# how many customer data objects do we have?
length(revenue)

## [1] 80

# mean, median and standard deviation of revenue
data.frame(mean=mean(revenue),
           median=median(revenue),
           sd=sd(revenue))

##   mean median   sd
## 1 3075  3086 903.4
```



## Overview over all variables

```
summary(MyCheeseData)
```

```
##      phone      gift      email
## Min.   : 0.09   Min.   : 32.9   Min.   : 0.11
## 1st Qu.:19.41   1st Qu.: 92.1   1st Qu.: 6.62
## Median :32.16   Median :112.4   Median : 8.48
## Mean   :32.72   Mean   :114.7   Mean   : 8.40
## 3rd Qu.:48.23   3rd Qu.:134.2   3rd Qu.:10.43
## Max.   :73.59   Max.   :183.4   Max.   :16.93
##      mail      revenue
## Min.   : 1.82   Min.   : 831
## 1st Qu.:12.68   1st Qu.:2326
## Median :19.89   Median :3086
## Mean   :19.60   Mean   :3075
## 3rd Qu.:25.55   3rd Qu.:3671
## Max.   :47.47   Max.   :4740
```

[Introduction](#)

[R and RStudio](#)

[What is R?](#)

[What is RStudio?](#)

[First steps](#)

[Simple linear regression](#)

[Multicollinearity](#)

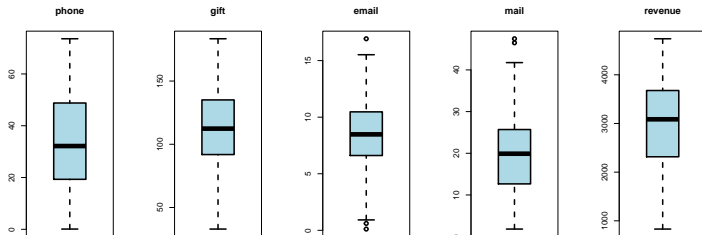
[Supplementary slides](#)





## Boxplots

```
names=names(MyCheeseData)
for(i in 1:5) {
  boxplot(MyCheeseData[,i], col="lightblue", lwd=3, main=names[i], cex=1 )
}
```



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

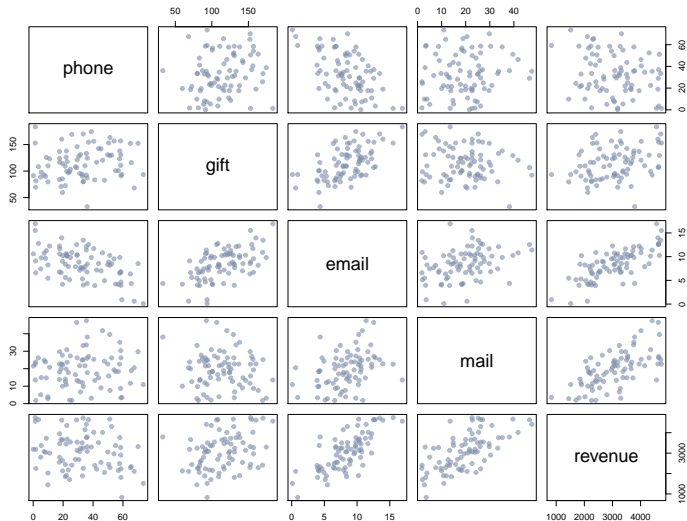
Simple linear  
regression

Multicollinearity

Supplementary slides

## Visualize pairs

```
plot(MyCheeseData, pch=19, col="#8090ADa0")
```



### Introduction

#### R and RStudio

What is R?

What is RStudio?

First steps

#### Simple linear regression

#### Multicollinearity

#### Supplementary slides



[Introduction](#)

[R and RStudio](#)

[What is R?](#)

[What is RStudio?](#)

[First steps](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Supplementary slides](#)

## List all Correlations

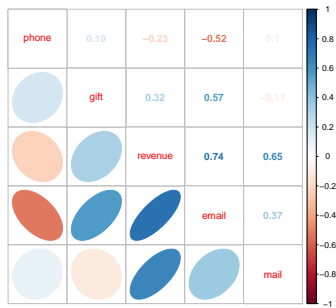
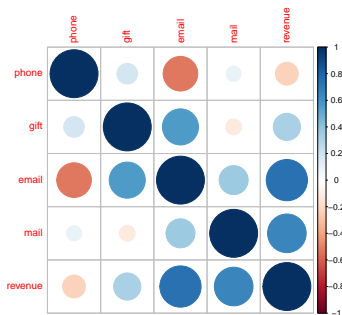
```
cor.MyCheeseData = cor(MyCheeseData)
cor.MyCheeseData
```

```
##           phone    gift  email    mail revenue
## phone  1.00000  0.1863 -0.5230  0.09869 -0.2273
## gift   0.18630  1.0000  0.5682 -0.11034  0.3220
## email  -0.52299  0.5682  1.0000  0.36645  0.7408
## mail   0.09869 -0.1103  0.3665  1.00000  0.6508
## revenue -0.22732  0.3220  0.7408  0.65076  1.0000
```



## Visualize correlation

```
require(corrplot)
corrplot(cor.MyCheeseData)
corrplot(cor.MyCheeseData, method="number", order = "AOE", tl.pos="d", type="upper")
```



Introduction

R and RStudio

What is R?

What is RStudio?

First steps

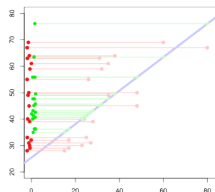
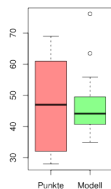
Simple linear  
regression

Multicollinearity

Supplementary slides

# Data analysis, Regression and Beyond: Table of Contents

- 1 Introduction
- 2 R and RStudio
- 3 Revision: Simple linear regression
- 4 Multicollinearity in Regression

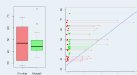


### 3 Revision: Simple linear regression

- Example set of data
- Trend as a linear model
- Least squares
- Best solution
- Variance and information
- Coefficient of determination
- $R^2$  is not perfect!
- Residual analysis

## Premier German Soccer League 2008/2009

- ▶ Given: data for all 18 clubs in the German Premier Soccer League in the season 2008/09



### Introduction

### R and RStudio

### Simple linear regression

#### Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

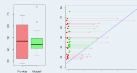
Residual analysis

### Multicollinearity

### Supplementary slides

## Premier German Soccer League 2008/2009

- ▶ Given: data for all 18 clubs in the German Premier Soccer League in the season 2008/09
- ▶ variables: **Budget** for season (only direct salaries for players)



### Introduction

### R and RStudio

### Simple linear regression

#### Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

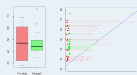
Residual analysis

### Multicollinearity

### Supplementary slides

## Premier German Soccer League 2008/2009

- ▶ Given: data for all 18 clubs in the German Premier Soccer League in the season 2008/09
- ▶ variables: **Budget** for season (only direct salaries for players)
- ▶ and: **resulting** table points at the end of the season



### Introduction

### R and RStudio

### Simple linear regression

#### Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

Residual analysis

### Multicollinearity

### Supplementary slides

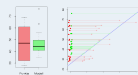


Premier German Soccer  
League 2008/2009

- ▶ Given: data for all 18 clubs in the German Premier Soccer League in the season 2008/09
- ▶ variables: **Budget** for season (only direct salaries for players)
- ▶ and: **resulting** table points at the end of the season

	Etat	Punkte
FC Bayern	80	67
VfL Wolfsburg	60	69
SV Werder Bremen	48	45
FC Schalke 04	48	50
VfB Stuttgart	38	64
Hamburger SV	35	61
Bayer 04 Leverkusen	35	49
Bor. Dortmund	32	59
Hertha BSC Berlin	31	63
1. FC Köln	28	39
Bor. Mönchengladbach	27	31
TSG Hoffenheim	26	55
Eintracht Frankfurt	25	33
Hannover 96	24	40
Energie Cottbus	23	30
VfL Bochum	17	32
Karlsruher SC	17	29
Arminia Bielefeld	15	28

(Source: Welt)



## Introduction

## R and RStudio

Simple linear  
regression

## Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

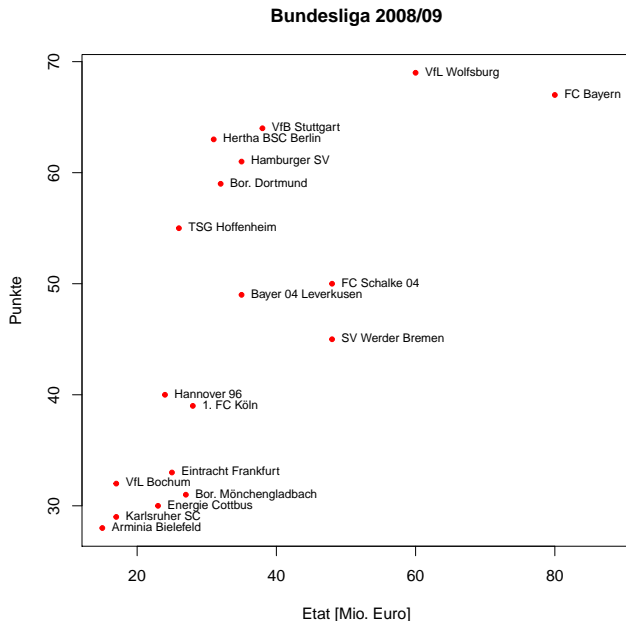
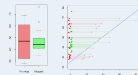
Coefficient of determination

 $R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



## Introduction

## R and RStudio

## Simple linear regression

### Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

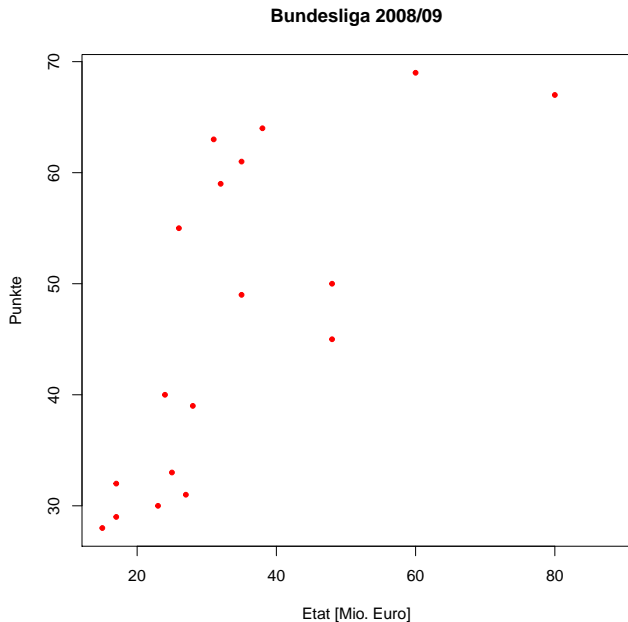
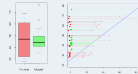
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



## Introduction

## R and RStudio

## Simple linear regression

### Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

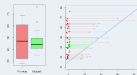
$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides

- ▶ Is it possible to find a simple function which can describe the dependency of the **end-of-season-points** versus the **club budget**?



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Example set of data](#)

[Trend as a linear model](#)

[Least squares](#)

[Best solution](#)

[Variance and information](#)

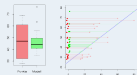
[Coefficient of determination](#)

[R<sup>2</sup> is not perfect!](#)

[Residual analysis](#)

[Multicollinearity](#)

[Supplementary slides](#)



- ▶ Is it possible to find a simple function which can describe the dependency of the **end-of-season-points** versus the **club budget**?
- ▶ In general: Description of a variable  $Y$  as a function of another variable  $X$ :

$$y = f(x)$$

- ▶ Notation:
  - $X$ : independent variable
  - $Y$  dependent variable

Introduction

R and RStudio

Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

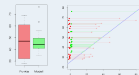
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

Multicollinearity

Supplementary slides



- ▶ Is it possible to find a simple function which can describe the dependency of the **end-of-season-points** versus the **club budget**?
- ▶ In general: Description of a variable  $Y$  as a function of another variable  $X$ :

$$y = f(x)$$

- ▶ Notation:
  - $X$ : **independent variable**
  - $Y$  **dependent variable**
- ▶ Important and easiest special case:  $f$  represents a linear trend:

$$y = a + b x$$

- ▶ To estimate using the data:  $a$  (intercept) and  $b$  (slope)
- ▶ Estimation of  $a$  and  $b$  is called: **Simple linear regression**

### Introduction

### R and RStudio

### Simple linear regression

Example set of data

### Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

Residual analysis

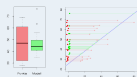
### Multicollinearity

### Supplementary slides

- ▶ using the regression model; per data object:

$$y_i = a + bx_i + \epsilon_i$$

- ▶  $\epsilon_i$  is error (regarding the population),
- ▶ with  $e_i = y_i - (\hat{a} + \hat{b}x_i)$ : deviation (**residual**) of given data of the sample und estimated values



### Introduction

### R and RStudio

### Simple linear regression

Example set of data

Trend as a linear model

### Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

Residual analysis

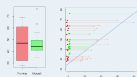
### Multicollinearity

### Supplementary slides

- ▶ using the regression model; per data object:

$$y_i = a + bx_i + \epsilon_i$$

- ▶  $\epsilon_i$  is error (regarding the population),
- ▶ with  $e_i = y_i - (\hat{a} + \hat{b}x_i)$ : deviation (**residual**) of given data of the sample und estimated values
- ▶ model works well if all residuals  $e_i$  are together as small as possible



### Introduction

### R and RStudio

### Simple linear regression

Example set of data

Trend as a linear model

### Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

Residual analysis

### Multicollinearity

### Supplementary slides

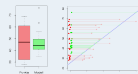


- ▶ using the regression model; per data object:

$$y_i = a + bx_i + \epsilon_i$$

- ▶  $\epsilon_i$  is error (regarding the population),
- ▶ with  $e_i = y_i - (\hat{a} + \hat{b}x_i)$ : deviation (**residual**) of given data of the sample und estimated values
- ▶ model works well if all residuals  $e_i$  are together as small as possible
- ▶ But just summing them up does not work, because  $e_i$  are positive and negative
- ▶ Hence: Sum of squares of  $e_i$
- ▶ **Ordinary Least squares** (OLS): Choose  $a$  and  $b$  in such a way, that

$$Q(a, b) = \sum_{i=1}^n [y_i - (a + bx_i)]^2 \rightarrow \min$$



### Introduction

### R and RStudio

### Simple linear regression

Example set of data

Trend as a linear model

### Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

Residual analysis

### Multicollinearity

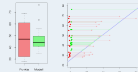
### Supplementary slides

- ▶ Best and unique solution:

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$



## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

Residual analysis

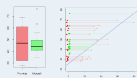
## Multicollinearity

## Supplementary slides

- ▶ regression line:

$$\hat{y} = \hat{a} + \hat{b}x$$

- ▶ Calculation of the soccer model
- ▶ With: table points  $\hat{y}$  and budget  $\hat{x}$ :



## Introduction

### R and RStudio

### Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

Residual analysis

### Multicollinearity

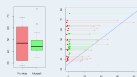
### Supplementary slides

- ▶ Calculation of the soccer model
- ▶ With: table points  $\hat{=}$  y and budget  $\hat{=}$  x:

---

$\bar{x}$	33,83
$\bar{y}$	46,89
$\sum x_i^2$	25209
$\sum x_i y_i$	31474
n	18

---



## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

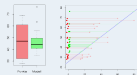
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



- ▶ Calculation of the soccer model
- ▶ With: table points  $\hat{=}$  y and budget  $\hat{=}$  x:

$\bar{x}$	33,83
$\bar{y}$	46,89
$\sum x_i^2$	25209
$\sum x_i y_i$	31474
n	18

$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33,83 \cdot 46,89}{25209 - 18 \cdot 33,83^2}$$

$$\approx 0,634$$

$$\Rightarrow \hat{a} = 46,89 - \hat{b} \cdot 33,83$$

$$\approx 25,443$$

## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

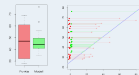
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



- ▶ Calculation of the soccer model

- ▶ With: table points  $\hat{=}$   $y$  and budget  $\hat{=}$   $x$ :

$\bar{x}$	33,83
$\bar{y}$	46,89
$\sum x_i^2$	25209
$\sum x_i y_i$	31474
$n$	18

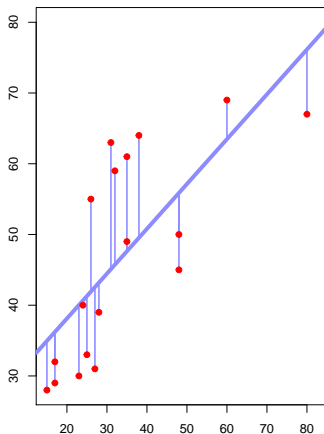
$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33,83 \cdot 46,89}{25209 - 18 \cdot 33,83^2}$$

$$\approx 0,634$$

$$\Rightarrow \hat{a} = 46,89 - \hat{b} \cdot 33,83$$

$$\approx 25,443$$

- ▶ model:  $\hat{y} = 25,443 + 0,634 \cdot x$



## Introduction

### R and RStudio

### Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

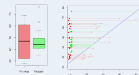
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

### Multicollinearity

### Supplementary slides



- ▶ Calculation of the soccer model

- ▶ With: table points  $\hat{=}$   $y$  and budget  $\hat{=}$   $x$ :

$\bar{x}$	33,83
$\bar{y}$	46,89
$\sum x_i^2$	25209
$\sum x_i y_i$	31474
$n$	18

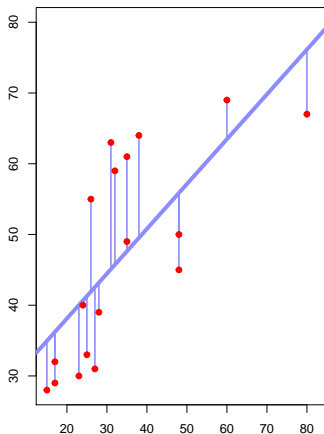
$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33,83 \cdot 46,89}{25209 - 18 \cdot 33,83^2}$$

$$\approx 0,634$$

$$\Rightarrow \hat{a} = 46,89 - \hat{b} \cdot 33,83$$

$$\approx 25,443$$

- ▶ model:  $\hat{y} = 25,443 + 0,634 \cdot x$



- ▶ prognosis for budget = 30:

$$\hat{y}(30) = 25,443 + 0,634 \cdot 30 \approx 44,463$$

## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

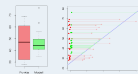
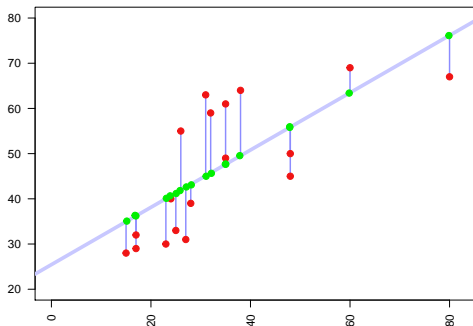
$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides

- ▶ **Variance** of data in  $y_i$  as indicator for model's **information content**
- ▶ Only a fraction of that variability can be mapped in the modeled values  $\hat{y}_i$



## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

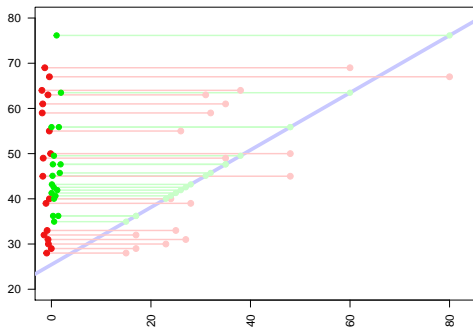
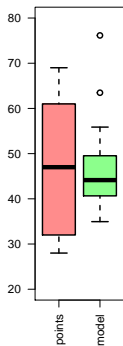
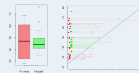
Residual analysis

## Multicollinearity

## Supplementary slides



- ▶ **Variance** of data in  $y_i$  as indicator for model's **information content**
- ▶ Only a fraction of that variability can be mapped in the modeled values  $\hat{y}_i$



## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

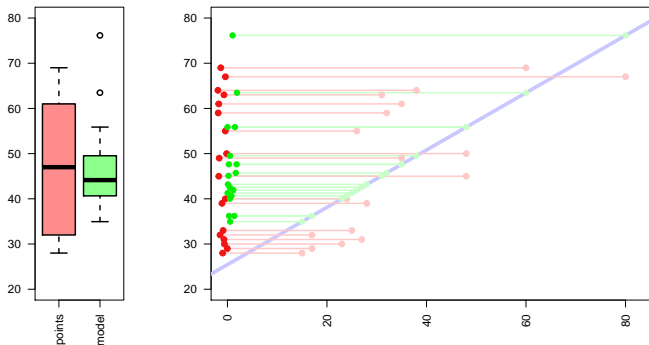
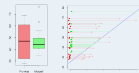
$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides

- ▶ **Variance** of data in  $y_i$  as indicator for model's **information content**
- ▶ Only a fraction of that variability can be mapped in the modeled values  $\hat{y}_i$



- ▶ Empirical variance for „red“ and „green“:

$$\frac{1}{18} \sum_{i=1}^{18} (y_i - \bar{y})^2 \approx 200,77 \quad \text{resp.} \quad \frac{1}{18} \sum_{i=1}^{18} (\hat{y}_i - \bar{y})^2 \approx 102,78$$

## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

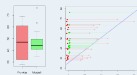
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



- ▶ Quality criterion for regression model: **Coefficient of determination:**

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^n y_i^2 - n\bar{y}^2} = r^2 \in [0; 1]$$

- ▶ Possible interpretation of  $R^2$ :  
**Proportion of total information in data which could be explained using model**
- ▶  $R^2 = 0$  , if  $X, Y$  uncorrelated  
 $R^2 = 1$  , if  $\hat{y}_i = y_i \forall i$  (every data point on regression line)
- ▶ With soccer example:

$$R^2 = \frac{\sum_{i=1}^{18} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{18} (y_i - \bar{y})^2} \approx \frac{102,78}{200,77} \approx 51,19\%$$

## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

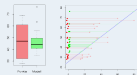
## Coefficient of determination

$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



► Famous data from the 1970ies:

$i$	$x_{1i}$	$x_{2i}$	$x_{3i}$	$x_{4i}$	$y_{1i}$	$y_{2i}$	$y_{3i}$	$y_{4i}$
1	10	10	10	8	8,04	9,14	7,46	6,58
2	8	8	8	8	6,95	8,14	6,77	5,76
3	13	13	13	8	7,58	8,74	12,74	7,71
4	9	9	9	8	8,81	8,77	7,11	8,84
5	11	11	11	8	8,33	9,26	7,81	8,47
6	14	14	14	8	9,96	8,10	8,84	7,04
7	6	6	6	8	7,24	6,13	6,08	5,25
8	4	4	4	19	4,26	3,10	5,39	12,50
9	12	12	12	8	10,84	9,13	8,15	5,56
10	7	7	7	8	4,82	7,26	6,42	7,91
11	5	5	5	8	5,68	4,74	5,73	6,89

(Quelle: **anscombe** )

## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

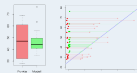
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



- ▶ in following table: results of each regression analysis
- ▶ with:  $x_k$  independent,  $y_k$  dependent variable
- ▶ Models:  $y_k = a_k + b_k x_k$

$k$	$\hat{a}_k$	$\hat{b}_k$	$R_k^2$
1	3,0001	0,5001	0,6665
2	3,0010	0,5000	0,6662
3	3,0025	0,4997	0,6663
4	3,0017	0,4999	0,6667

## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

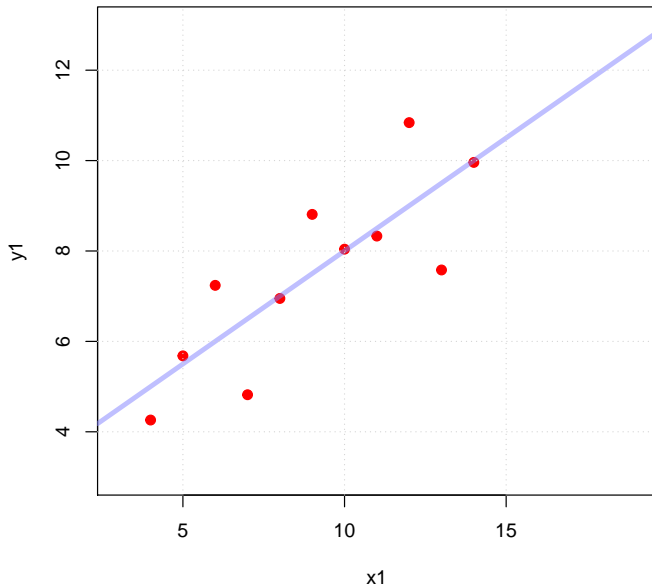
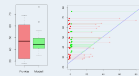
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



## Introduction

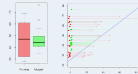
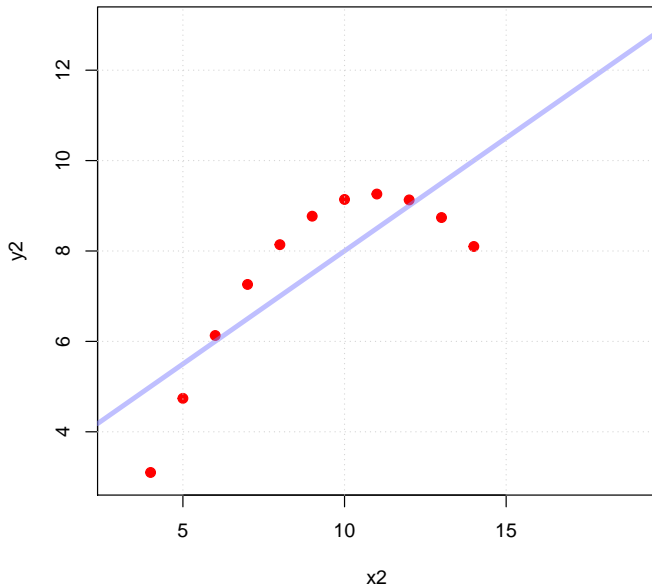
## R and RStudio

## Simple linear regression

- Example set of data
- Trend as a linear model
- Least squares
- Best solution
- Variance and information
- Coefficient of determination
- $R^2$  is not perfect!
- Residual analysis

## Multicollinearity

## Supplementary slides



## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

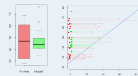
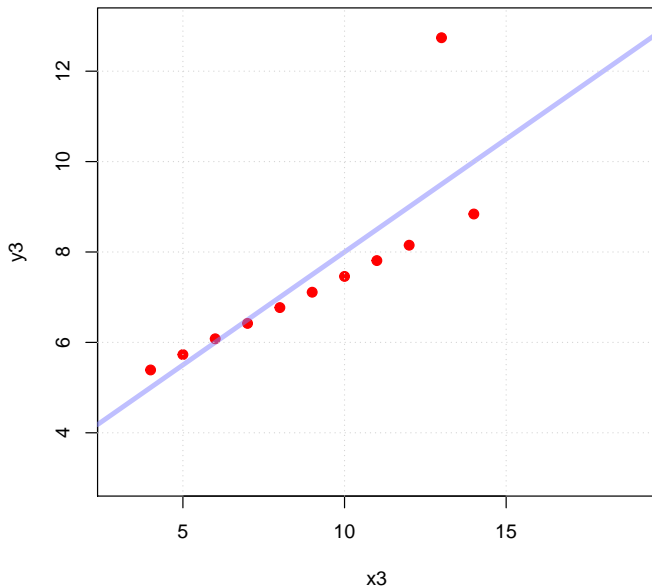
$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides

# Plot der Anscombe-Daten



## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

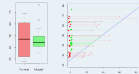
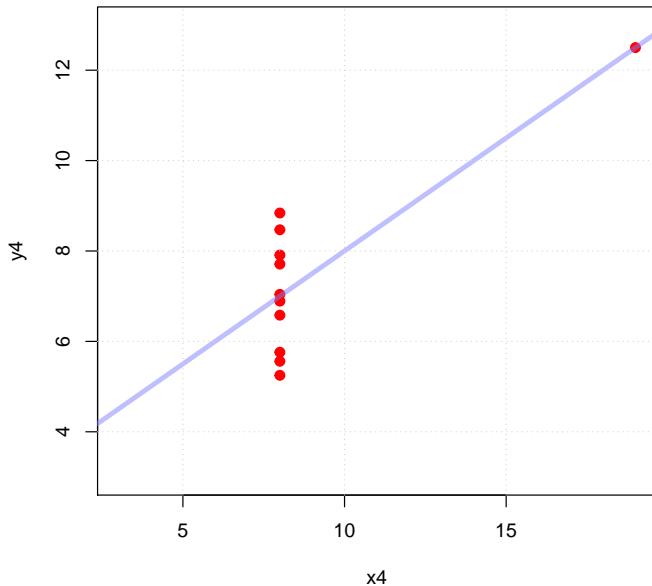
$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides





## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

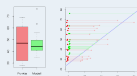
$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides

- ▶ often illuminating: distribution of **residuals**  $e_i$
- ▶ Common: graphical display of residuals
- ▶ e.g.:  $e_i$  over  $\hat{y}_i$



## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

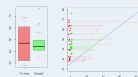
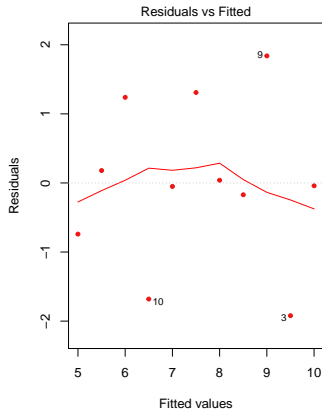
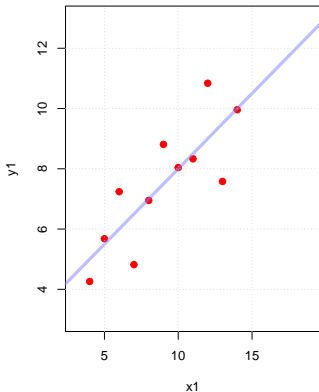
$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides

- ▶ often illuminating: distribution of **residuals**  $e_i$
- ▶ Common: graphical display of residuals
- ▶ e.g.:  $e_i$  over  $\hat{y}_i$



## Introduction

## R and RStudio

## Simple linear regression

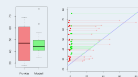
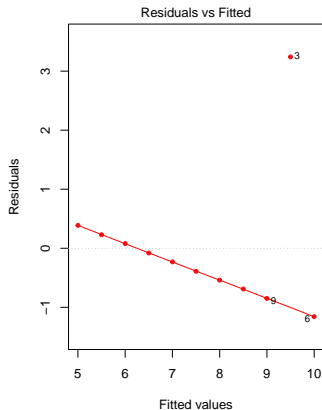
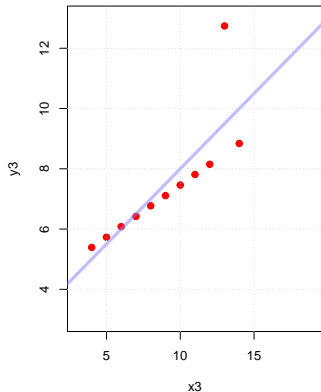
- Example set of data
- Trend as a linear model
- Least squares
- Best solution
- Variance and information
- Coefficient of determination
- $R^2$  is not perfect!

## Residual analysis

## Multicollinearity

## Supplementary slides

- ▶ often illuminating: distribution of **residuals**  $e_i$
- ▶ Common: graphical display of residuals
- ▶ e.g.:  $e_i$  over  $\hat{y}_i$



## Introduction

## R and RStudio

## Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

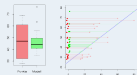
Coefficient of determination

$R^2$  is not perfect!

Residual analysis

## Multicollinearity

## Supplementary slides



## Properties of residual distribution

- ▶ Preferably **no systematic pattern**
- ▶ No change of variance dependent of  $\hat{y}_i$  (**Homoscedasticity**)
- ▶ Necessary for inferential analysis: Approximately **normal distributed** residuals (q-q-plots)

## Causality vs. correlation

- ▶ Mostly important for useful regression analysis:
- ▶ **Causal connection** between independent and dependent variable
- ▶ Otherwise: No valuable prognosis possible
- ▶ Often: **Latent variables** in the background

Introduction

R and RStudio

Simple linear regression

Example set of data

Trend as a linear model

Least squares

Best solution

Variance and information

Coefficient of determination

$R^2$  is not perfect!

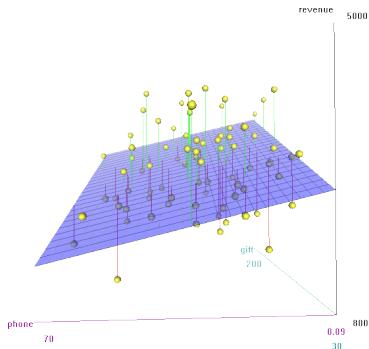
Residual analysis

Multicollinearity

Supplementary slides

# Data analysis, Regression and Beyond: Table of Contents

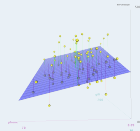
- 1 Introduction
- 2 R and RStudio
- 3 Revision: Simple linear regression
- 4 Multicollinearity in Regression



## 4 Multicollinearity in Regression

- Back to Mr. Meier
- Mr. Maier und his problem
- Vocabulary
- Geometry and Multicollinearity
- Common believe
- Solution approach
- From diagnosis to therapy
- Roundup

	phone	gift	email	mail	revenue
1	29.36	146.14	10.32	13.36	3137.85
2	8.75	125.82	11.27	14.72	3728.11
3	36.15	124.51	8.45	17.72	3084.75
4	51.20	129.36	10.27	39.59	4667.90
5	51.36	163.42	8.19	7.57	2286.41
6	34.65	110.04	7.89	21.68	4147.61
7	19.65	113.88	10.23	22.17	3648.22
8	17.51	84.04	6.79	13.82	2558.09
9	10.93	123.18	12.24	20.81	3003.83
10	1.35	152.89	15.52	22.63	4740.21
11	46.36	120.54	10.81	41.75	4014.46
12	31.61	131.27	7.69	6.72	3241.13
13	23.48	96.71	7.93	17.80	2174.79
14	70.09	152.44	8.55	29.77	3318.12
15	32.70	94.12	7.66	24.92	3504.20
		⋮			



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Meier und his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

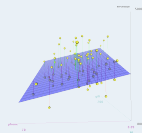
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Idea: Maybe there is a (linear) causal dependency between **revenue** and the distinct advertising actions



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and  
Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

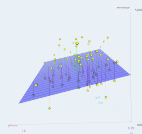
[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)



- ▶ Idea: Maybe there is a (linear) causal dependency between **revenue** and the distinct advertising actions
- ▶ In other words: How much (more) revenue in Euro do we get from investing one (more) Euro in customer gifts (mails, emails, phone calls)?



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

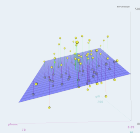
Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Idea: Maybe there is a (linear) causal dependency between **revenue** and the distinct advertising actions
- ▶ In other words: How much (more) revenue in Euro do we get from investing one (more) Euro in customer gifts (mails, emails, phone calls)?
- ▶ That means: We have to do a **Multivariate Regression** model like this:

$$\begin{aligned} Y_{\text{revenue}} = & \beta_0 + \beta_{\text{phone}} \cdot X_{\text{phone}} \\ & + \beta_{\text{gift}} \cdot X_{\text{gift}} \\ & + \beta_{\text{mail}} \cdot X_{\text{mail}} \\ & + \beta_{\text{email}} \cdot X_{\text{email}} \end{aligned}$$



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier and his problem](#)

[Vocabulary](#)

[Geometry and Multicollinearity](#)

[Common believe](#)

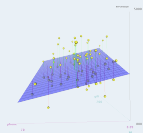
[Solution approach](#)

[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)

```
##  
## Call:  
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1084.8  -348.9   -46.5    333.1  1010.1   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    741.5      250.3     2.96  0.0041 **      
## phone          -68.2       34.1    -2.00  0.0494 *       
## gift           47.5       22.8     2.08  0.0408 *       
## mail          132.6       46.3     2.86  0.0054 **      
## email         -413.9      282.9    -1.46  0.1477         
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 480 on 75 degrees of freedom  
## Multiple R-squared:  0.732, Adjusted R-squared:  0.718   
## F-statistic: 51.3 on 4 and 75 DF,  p-value: <2e-16
```



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

### Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

Solution approach

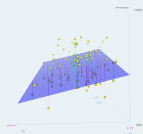
Diagnosis and therapy

Roundup

## Supplementary slides

```
##  
## Call:  
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1084.8  -348.9   -46.5    333.1  1010.1   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    741.5      250.3     2.96  0.0041 **      
## phone          -68.2       34.1    -2.00  0.0494 *       
## gift           47.5       22.8     2.08  0.0408 *       
## mail          132.6       46.3     2.86  0.0054 **      
## email         -413.9      282.9    -1.46  0.1477         
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 480 on 75 degrees of freedom  
## Multiple R-squared:  0.732, Adjusted R-squared:  0.718   
## F-statistic: 51.3 on 4 and 75 DF,  p-value: <2e-16
```

- ▶ Adjusted coefficient of determination ( $R^2$ ) 0.7179



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

Solution approach

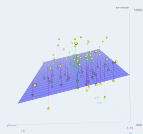
Diagnosis and therapy

Roundup

## Supplementary slides

```
##  
## Call:  
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1084.8  -348.9   -46.5    333.1  1010.1   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    741.5      250.3     2.96  0.0041 **      
## phone          -68.2       34.1    -2.00  0.0494 *       
## gift           47.5       22.8     2.08  0.0408 *       
## mail          132.6       46.3     2.86  0.0054 **      
## email         -413.9      282.9    -1.46  0.1477        
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 480 on 75 degrees of freedom  
## Multiple R-squared:  0.732, Adjusted R-squared:  0.718   
## F-statistic: 51.3 on 4 and 75 DF,  p-value: <2e-16
```

- ▶ Adjusted coefficient of determination ( $R^2$ ) 0.7179
- ▶ F-statistic: 51.2593, p-value:  $9\,9628 \times 10^{-21}$



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

### Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

Solution approach

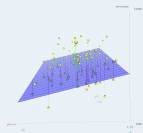
Diagnosis and therapy

Roundup

## Supplementary slides

```
##  
## Call:  
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1084.8  -348.9   -46.5    333.1  1010.1   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    741.5      250.3     2.96  0.0041 **      
## phone          -68.2       34.1    -2.00  0.0494 *       
## gift           47.5       22.8     2.08  0.0408 *       
## mail          132.6       46.3     2.86  0.0054 **      
## email         -413.9      282.9    -1.46  0.1477         
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 480 on 75 degrees of freedom  
## Multiple R-squared:  0.732, Adjusted R-squared:  0.718   
## F-statistic: 51.3 on 4 and 75 DF,  p-value: <2e-16
```

- ▶ Adjusted coefficient of determination ( $R^2$ ) 0.7179
- ▶ F-statistic: 51.2593, p-value:  $9.9628 \times 10^{-21}$
- ▶ Herr Maier is a little surprised, e.g. why email advertising seems to be this harmful.



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

### Back to Mr. Meier

#### Mr. Maier und his problem

#### Vocabulary

#### Geometry and

#### Multicollinearity

#### Common believe

#### Solution approach

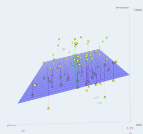
#### Diagnosis and therapy

#### Roundup

## Supplementary slides

```
##  
## Call:  
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1084.8  -348.9   -46.5    333.1  1010.1   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    741.5      250.3     2.96  0.0041 **      
## phone          -68.2       34.1    -2.00  0.0494 *       
## gift           47.5       22.8     2.08  0.0408 *       
## mail          132.6       46.3     2.86  0.0054 **      
## email         -413.9      282.9    -1.46  0.1477         
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 480 on 75 degrees of freedom  
## Multiple R-squared:  0.732, Adjusted R-squared:  0.718   
## F-statistic: 51.3 on 4 and 75 DF,  p-value: <2e-16
```

- ▶ Adjusted coefficient of determination ( $R^2$ ) 0.7179
- ▶ F-statistic: 51.2593, p-value:  $9.9628 \times 10^{-21}$
- ▶ Herr Maier is a little surprised, e.g. why email advertising seems to be this harmful.
- ▶ But we know that numbers don't lie...



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

### Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

Solution approach

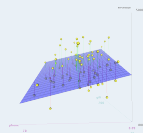
Diagnosis and therapy

Roundup

## Supplementary slides

- ▶ Calculation of phone spendings was slightly incorrect...
- ▶ ...and has been corrected

	phone.old	phone.new
1	29.36	29.36
2	8.75	13.75
3	36.15	36.15
4	51.20	56.20
5	51.36	56.36
6	34.65	39.65
7	19.65	24.65
8	17.51	22.51
9	10.93	15.93
10	1.35	6.35
11	46.36	51.36
12	31.61	31.61
13	23.48	23.48
14	70.09	75.09
15	32.70	32.70
	⋮	



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

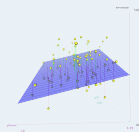


## ► Model from corrected data

```
##  
## Call:  
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1187.4  -301.7   -75.9    384.1   1083.8   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    784.2      253.4     3.09  0.0028 **      
## phone          -24.3       17.8    -1.37  0.1757       
## gift           18.5       12.2     1.52  0.1334       
## mail           73.9       25.0     2.96  0.0041 **      
## email          -49.8      147.6    -0.34  0.7369       
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 486 on 75 degrees of freedom  
## Multiple R-squared:  0.725, Adjusted R-squared:  0.71  
## F-statistic: 49.4 on 4 and 75 DF,  p-value: <2e-16
```

## ► Model of the original data

```
##  
## Call:  
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1084.8  -348.9   -46.5    333.1   1010.1   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    741.5      250.3     2.96  0.0041 **      
## phone          -68.2       34.1    -2.00  0.0494 *       
## gift           47.5       22.8     2.08  0.0408 *       
## mail           132.6      46.3     2.86  0.0054 **      
## email          -413.9     282.9    -1.46  0.1477       
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 480 on 75 degrees of freedom  
## Multiple R-squared:  0.732, Adjusted R-squared:  0.718  
## F-statistic: 51.3 on 4 and 75 DF,  p-value: <2e-16
```



MyCheeseData)

Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

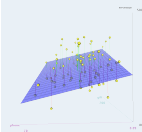
► Model from corrected data

► Model of the original data

```
##
## Call:
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1187.4   -301.7   -75.9    384.1   1083.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    784.2      253.4     3.09  0.0028 **
## phone          -24.3       17.8    -1.37  0.1757
## gift            18.5       12.2     1.52  0.1334
## mail           73.9       25.0     2.96  0.0041 **
## email          -49.8      147.6    -0.34  0.7369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 486 on 75 degrees of freedom
## Multiple R-squared:  0.725, Adjusted R-squared:  0.71
## F-statistic: 49.4 on 4 and 75 DF,  p-value: <2e-16
```

```
##
## Call:
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1084.8   -348.9   -46.5    333.1   1010.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    741.5      250.3     2.96  0.0041 **
## phone          -68.2       34.1    -2.00  0.0494 *
## gift            47.5       22.8     2.08  0.0408 *
## mail           132.6       46.3     2.86  0.0054 **
## email          -413.9      282.9    -1.46  0.1477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 480 on 75 degrees of freedom
## Multiple R-squared:  0.732, Adjusted R-squared:  0.718
## F-statistic: 51.3 on 4 and 75 DF,  p-value: <2e-16
```

- Model seems to be very unstable
- Small changes in data have a dramatic effect to the model's parameters
- Causal analysis is necessary!



MyCheeseData)

Introduction

R and RStudio

Simple linear regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and Multicollinearity

Common believe

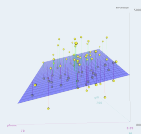
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Linear regression: models the relationship between a dependent variable  $y$ , independent variables  $x_1, \dots, x_m$  with the help of parameters  $\beta_0, \dots, \beta_m$



[Introduction](#)

[R and RStudio](#)

[Simple linear  
regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and  
Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

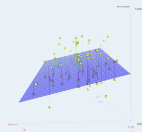
[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)

- ▶ Linear regression: models the relationship between a dependent variable  $y$ , independent variables  $x_1, \dots, x_m$  with the help of parameters  $\beta_0, \dots, \beta_m$
- ▶ in general:  $y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_m \cdot x_m + u$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = X \cdot \beta + u$$



Introduction

R and RStudio

Simple linear regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

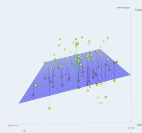
Roundup

Supplementary slides

- ▶ Linear regression: models the relationship between a dependent variable  $y$ , independent variables  $x_1, \dots, x_m$  with the help of parameters  $\beta_0, \dots, \beta_m$
- ▶ in general:  $y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_m \cdot x_m + u$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = X \cdot \beta + u$$

- ▶ The error term  $u$  is the portion of the data which can not be described by the model



Introduction

R and RStudio

Simple linear regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

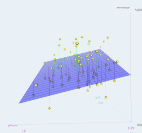
Supplementary slides

- ▶ Linear regression: models the relationship between a dependent variable  $y$ , independent variables  $x_1, \dots, x_m$  with the help of parameters  $\beta_0, \dots, \beta_m$
- ▶ in general:  $y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_m \cdot x_m + u$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = X \cdot \beta + u$$

- ▶ The error term  $u$  is the portion of the data which can not be described by the model
- ▶ Typical: Estimation of the „best“ model parameters  $\hat{\beta}_0, \dots, \hat{\beta}_m$  using a least-square analysis:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_m \end{pmatrix} = (X^T X)^{-1} X^T y$$



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

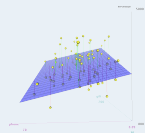
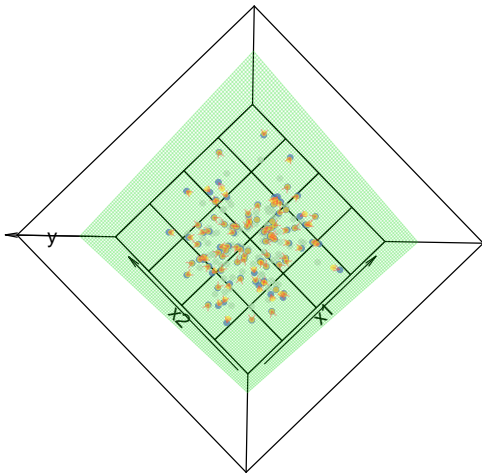
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

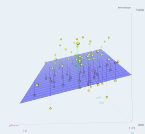
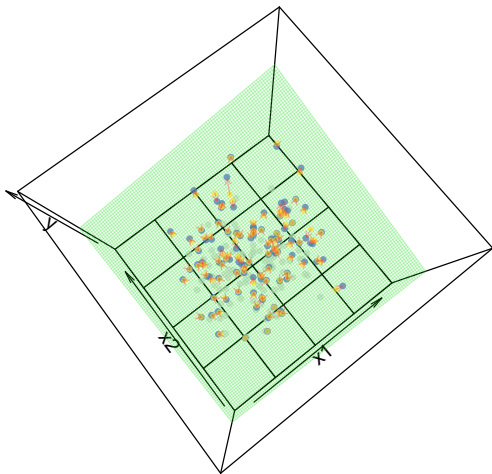
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

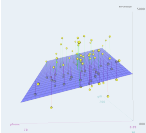
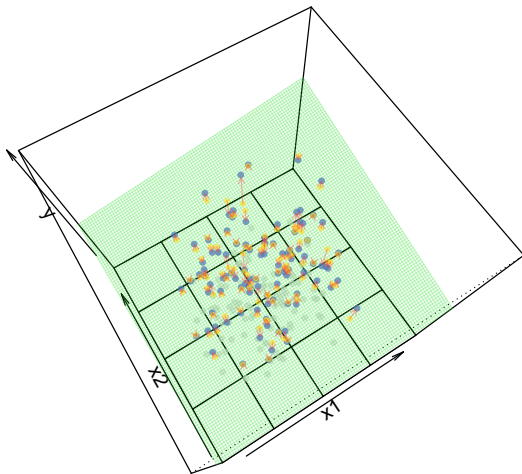
Roundup

Supplementary slides



# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

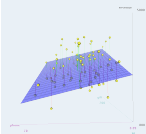
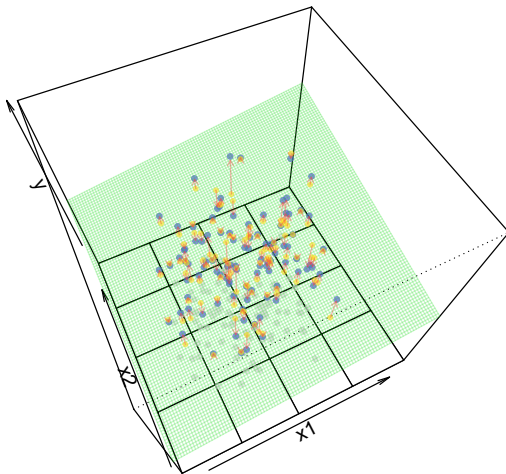
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

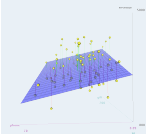
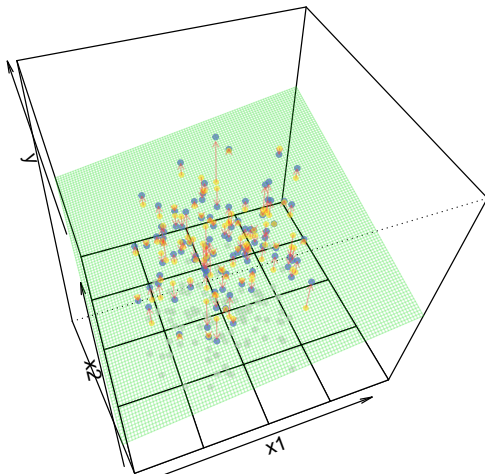
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

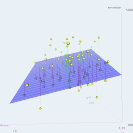
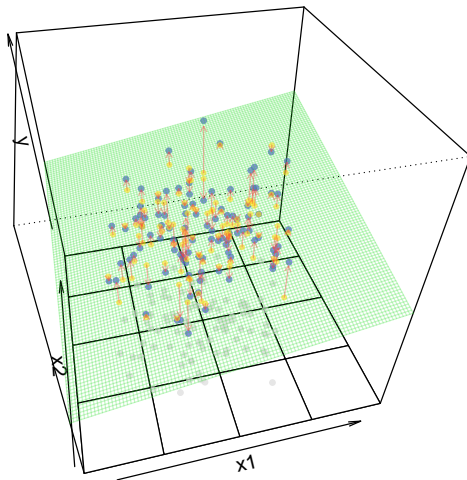
Diagnosis and therapy

Roundup

## Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

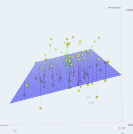
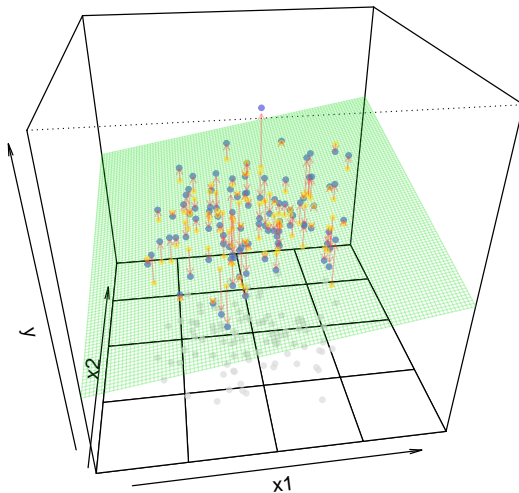
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

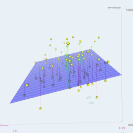
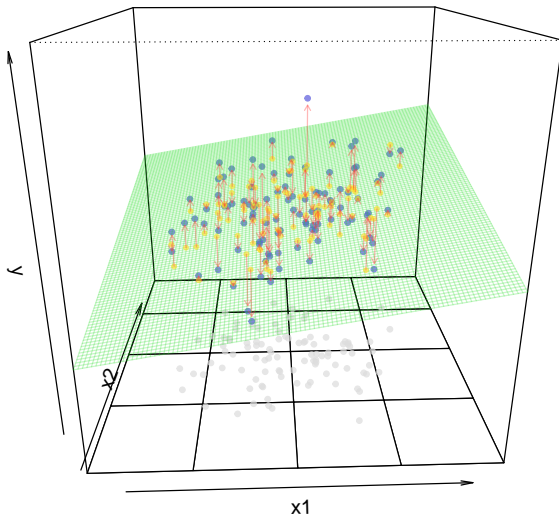
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

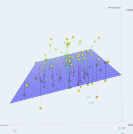
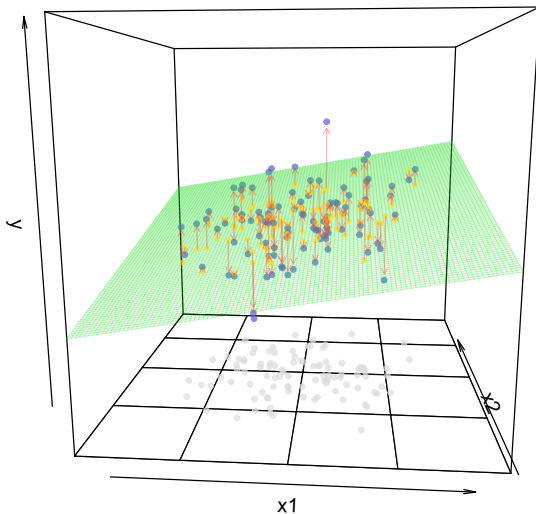
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

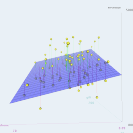
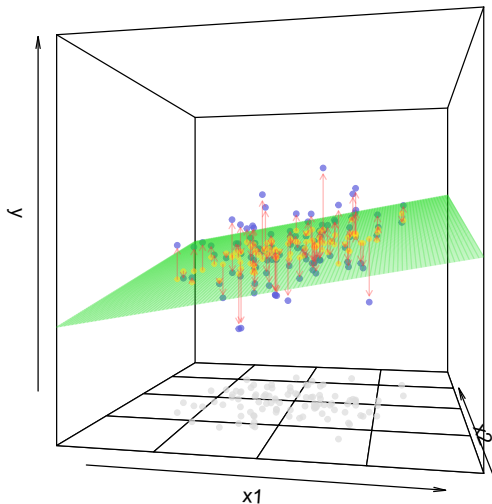
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

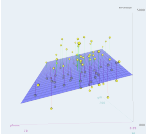
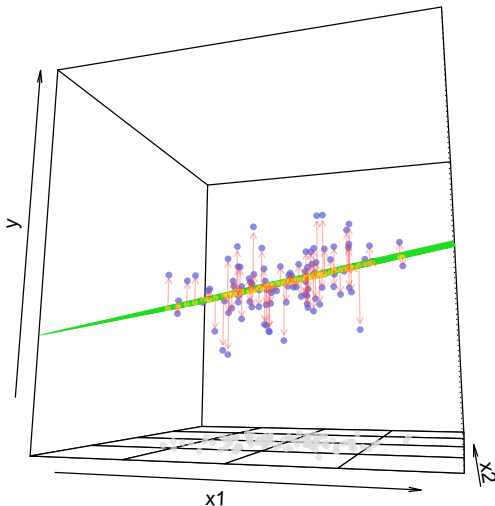
Roundup

Supplementary slides



# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

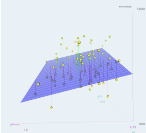
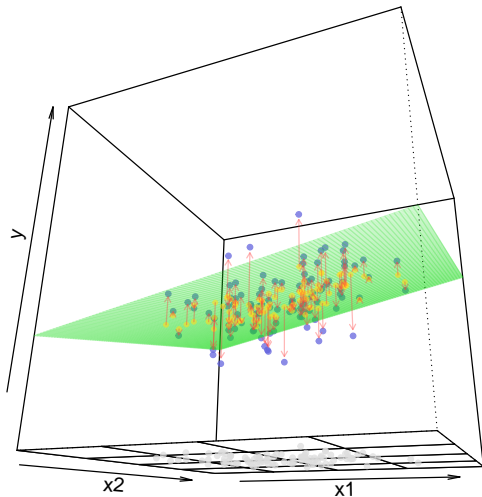
Diagnosis and therapy

Roundup

## Supplementary slides

# Two independent variables

- ▶ two-dimensional example
- ▶ stable model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

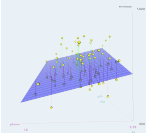
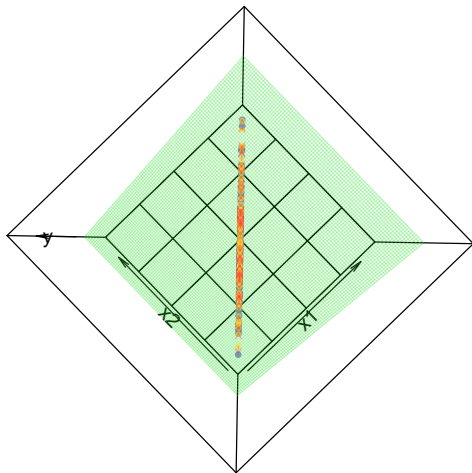
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

## ► Perfect multicollinearity



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

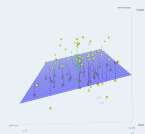
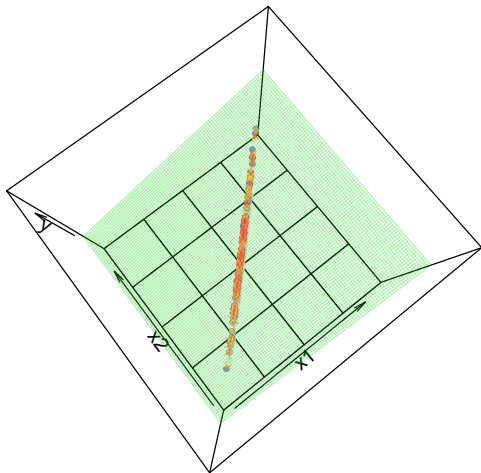
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

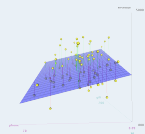
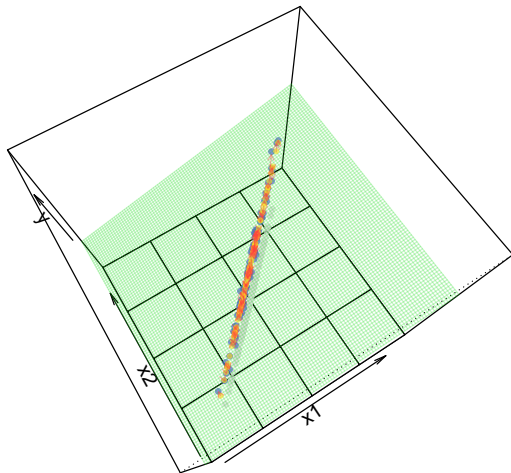
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

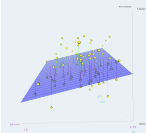
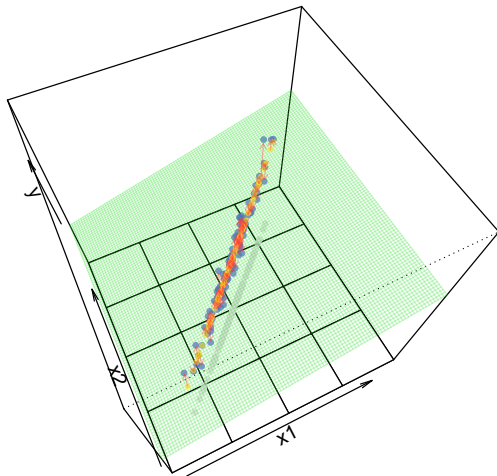
Diagnosis and therapy

Roundup

Supplementary slides

## Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



Introduction

R and RStudio

Simple linear  
regression

**Multicollinearity**

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

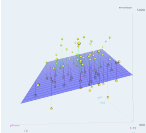
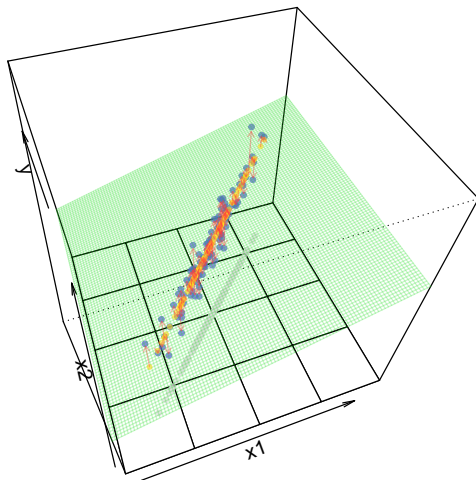
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

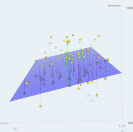
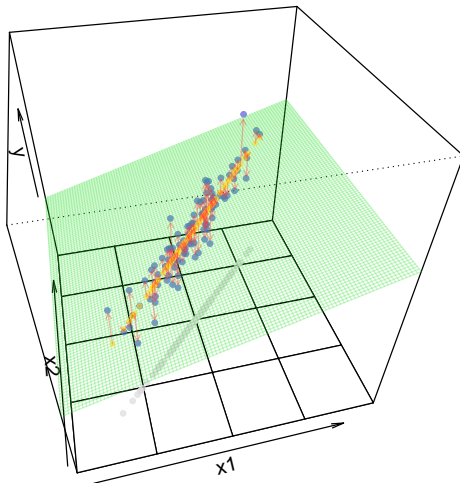
Diagnosis and therapy

Roundup

## Supplementary slides

# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

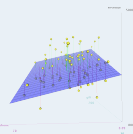
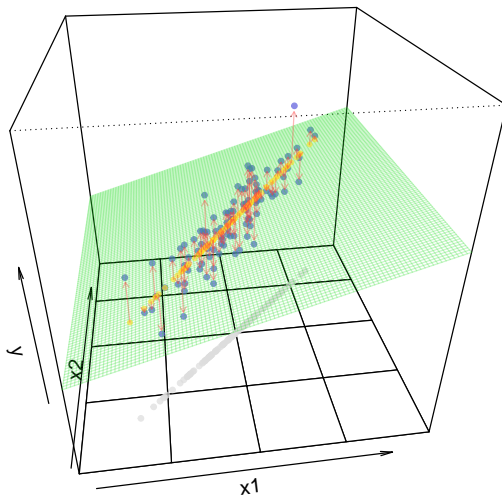
Roundup

## Supplementary slides



# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

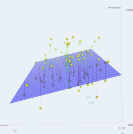
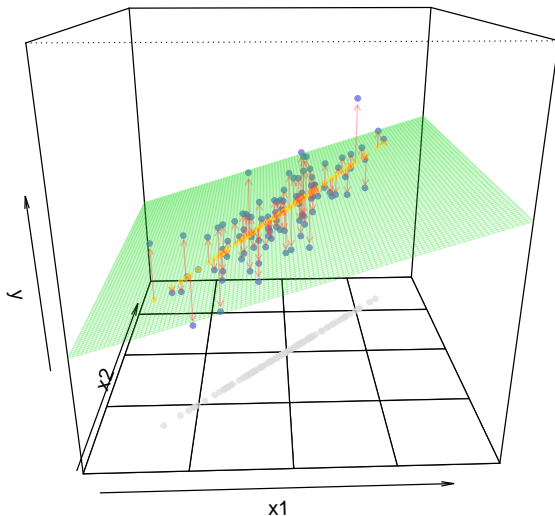
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

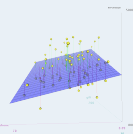
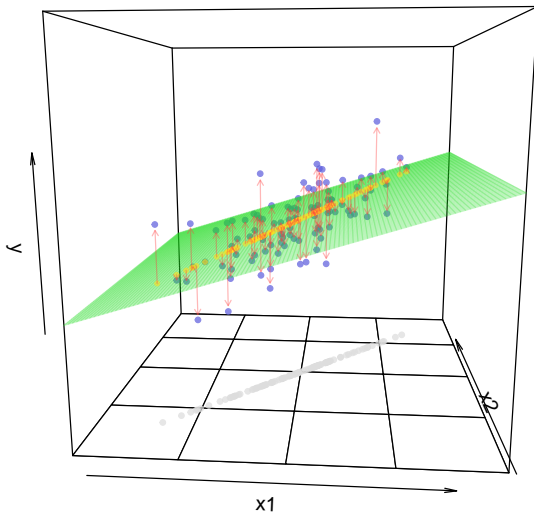
Diagnosis and therapy

Roundup

## Supplementary slides

# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

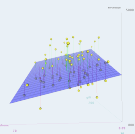
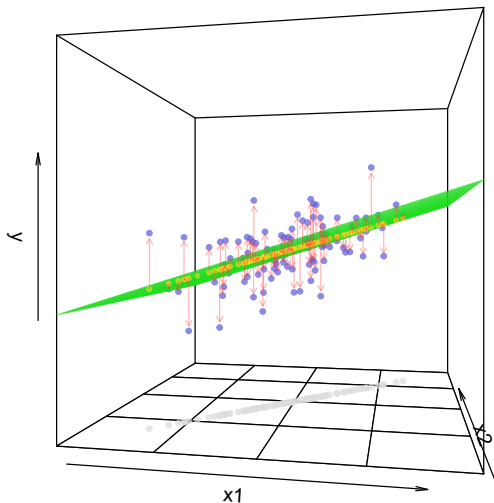
Diagnosis and therapy

Roundup

## Supplementary slides

## Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

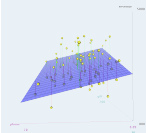
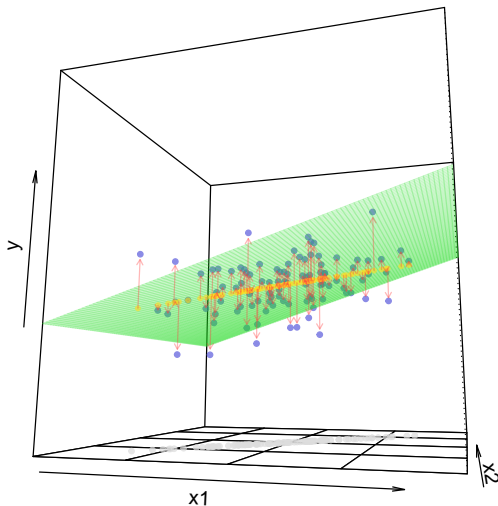
Diagnosis and therapy

Roundup

Supplementary slides

# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

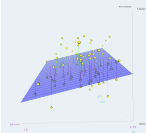
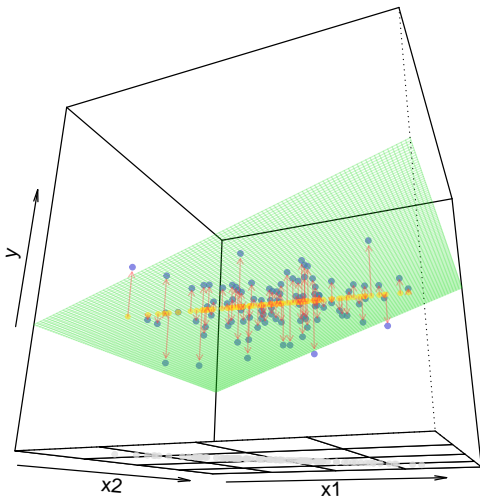
Diagnosis and therapy

Roundup

## Supplementary slides

# Two independent variables

- ▶ Perfect multicollinearity
- ▶ no regression model possible



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

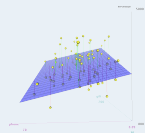
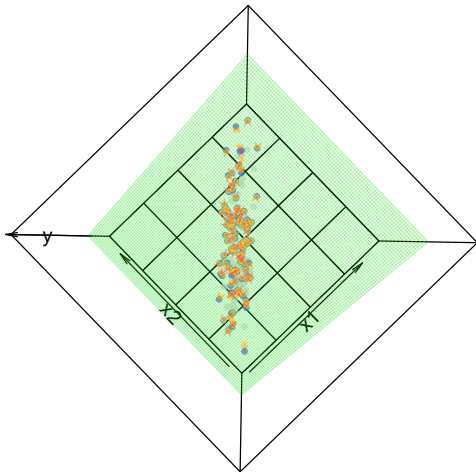
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

## ► Strong Multicollinearity



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

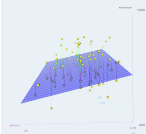
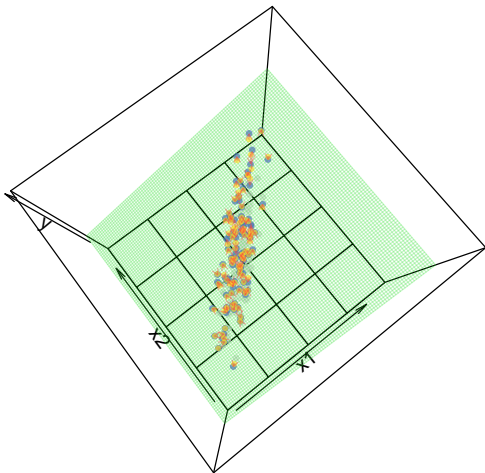
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

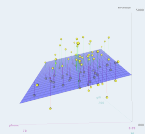
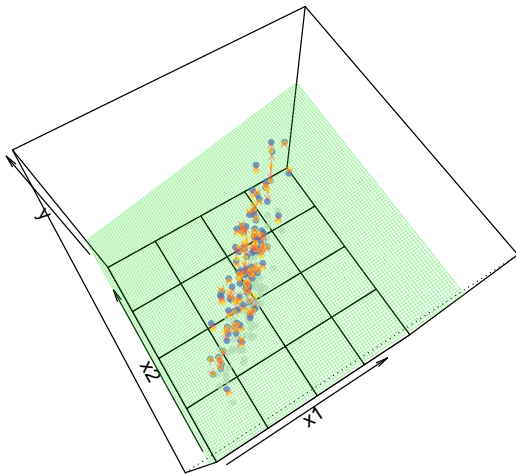
Diagnosis and therapy

Roundup

Supplementary slides



- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

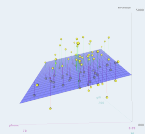
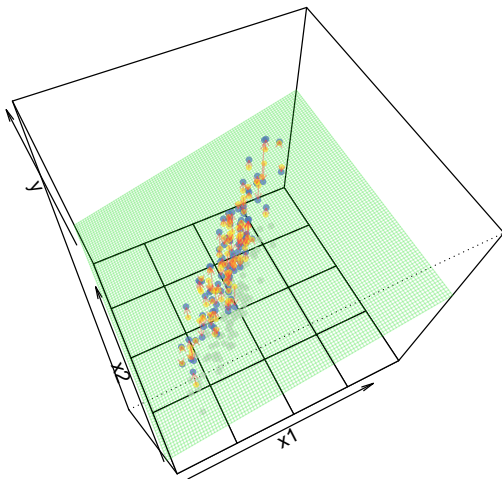
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

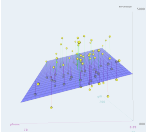
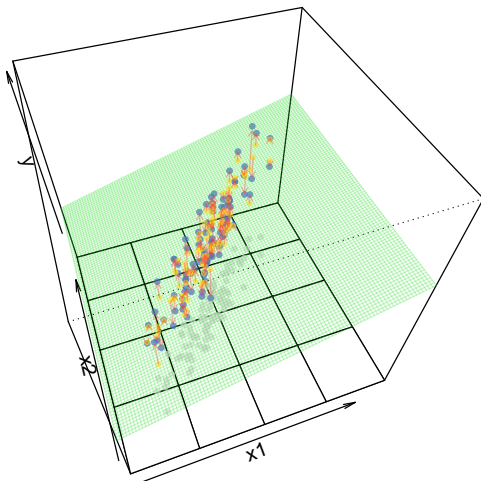
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

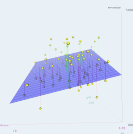
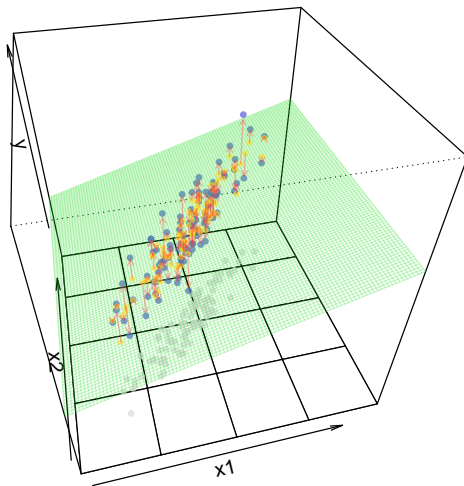
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

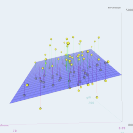
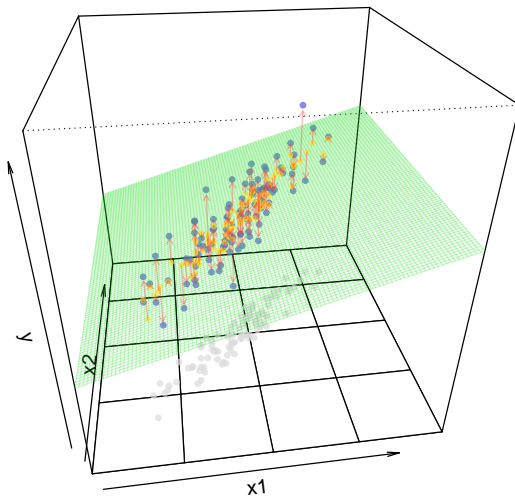
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

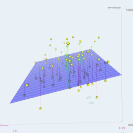
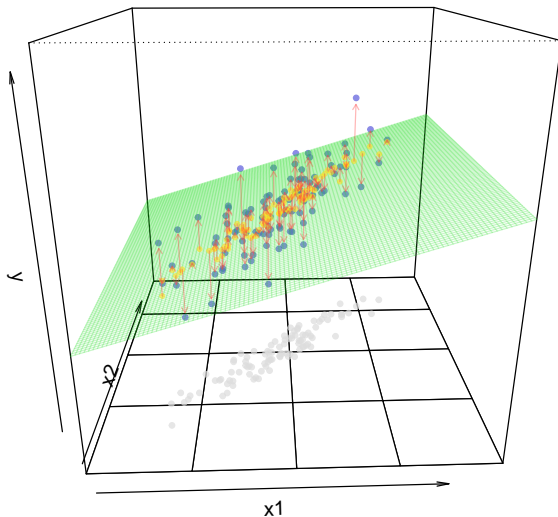
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

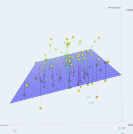
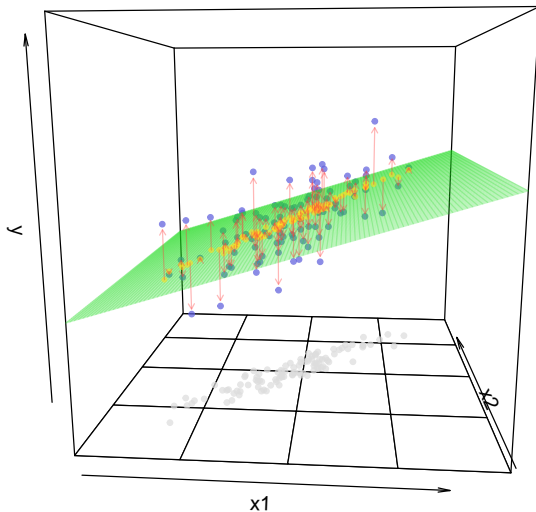
Diagnosis and therapy

Roundup

Supplementary slides

## Two independent variables

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

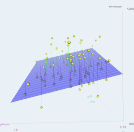
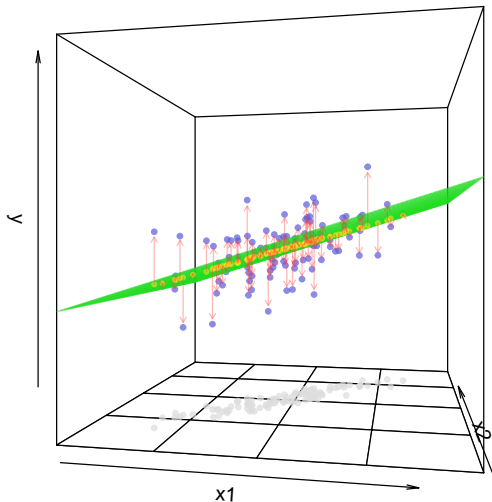
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

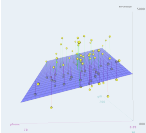
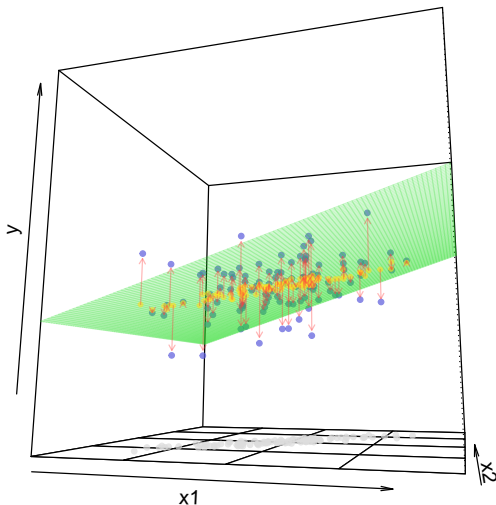
Roundup

Supplementary slides



# Two independent variables

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

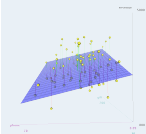
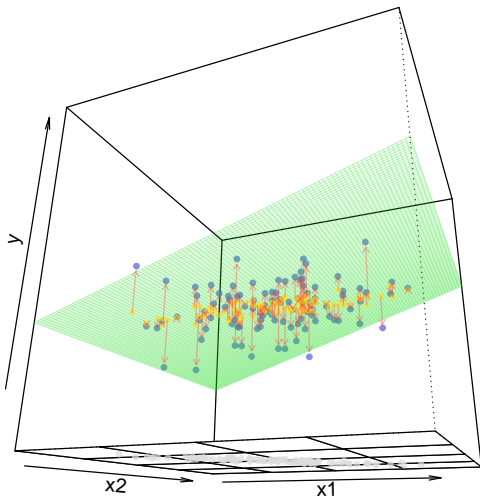
Solution approach

Diagnosis and therapy

Roundup

## Supplementary slides

- ▶ Strong Multicollinearity
- ▶ All parameters of the regression model are unstable



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

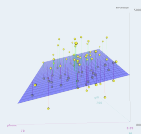
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Maybe the correlation of the independent variables is a good measure?



[Introduction](#)

[R and RStudio](#)

[Simple linear  
regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier and his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

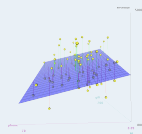
[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)

- ▶ Maybe the correlation of the independent variables is a good measure?
- ▶ But: Perfect multicollinearity between three or more vectors (which are not pairwise correlated)
- ▶ Simple Example:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

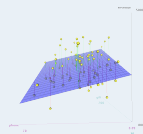
Common believe

Solution approach

Diagnosis and therapy

Roundup

Supplementary slides



## Idea for diagnosis: Correlation?

- ▶ Maybe the correlation of the independent variables is a good measure?
- ▶ But: Perfect multicollinearity between three or more vectors (which are not pairwise correlated)
- ▶ Simple Example:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Mr. Maier's data: Correlation matrix from independent variables:

	phone	gift	email	mail
phone	1.00	0.19	-0.52	0.10
gift	0.19	1.00	0.57	-0.11
email	-0.52	0.57	1.00	0.37
mail	0.10	-0.11	0.37	1.00

[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)

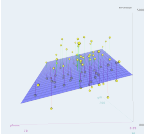
- ▶ Maybe the correlation of the independent variables is a good measure?
- ▶ But: Perfect multicollinearity between three or more vectors (which are not pairwise correlated)
- ▶ Simple Example:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Mr. Maier's data: Correlation matrix from independent variables:

	phone	gift	email	mail
phone	1.00	0.19	-0.52	0.10
gift	0.19	1.00	0.57	-0.11
email	-0.52	0.57	1.00	0.37
mail	0.10	-0.11	0.37	1.00

- ▶ Correlation is sufficient, but not necessary for multicollinearity



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and  
Multicollinearity](#)

[Common believe](#)

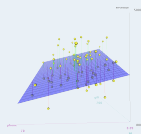
[Solution approach](#)

[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)

- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$



[Introduction](#)

[R and RStudio](#)

[Simple linear  
regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and  
Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

[Diagnosis and therapy](#)

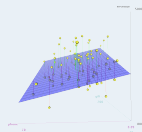
[Roundup](#)

[Supplementary slides](#)

- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

$$v_0x_0 + \dots + v_mx_m = Xv = a \approx 0$$

(If not all scalars  $v_0, \dots, v_m$  are 0)



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

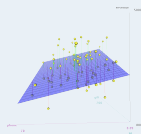


- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

$$v_0x_0 + \dots + v_mx_m = Xv = a \approx 0$$

(If not all scalars  $v_0, \dots, v_m$  are 0)

- ▶ Therefore wanted: vector  $v$  with normed length (e.g. 1), such that  $|a|$  becomes small



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

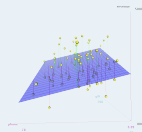
- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

$$v_0x_0 + \dots + v_mx_m = Xv = a \approx 0$$

(If not all scalars  $v_0, \dots, v_m$  are 0)

- ▶ Therefore wanted: vector  $v$  with normed length (e.g. 1), such that  $|a|$  becomes small
- ▶ Solution: Smallest eigenvalue  $\lambda_0$  (with its corresponding eigenvector  $v_0$ ) from  $X^T X$  shows strongest nearly linear dependency

▶ Details



Introduction

R and RStudio

Simple linear regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

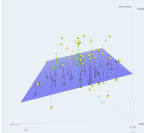
$$v_0x_0 + \dots + v_mx_m = Xv = a \approx 0$$

(If not all scalars  $v_0, \dots, v_m$  are 0)

- ▶ Therefore wanted: vector  $v$  with normed length (e.g. 1), such that  $|a|$  becomes small
- ▶ Solution: Smallest eigenvalue  $\lambda_0$  (with its corresponding eigenvector  $v_0$ ) from  $X^T X$  shows strongest nearly linear dependency
- ▶ Proportion of largest and smallest Eigenvalue as per

$$\kappa(X) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

is called **condition number**



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier and his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

[Diagnosis and therapy](#)

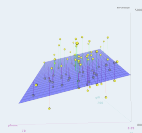
[Roundup](#)

[Supplementary slides](#)

[▶ Details](#)

- ▶ To create a benchmark out of condition numbers: Standardise variables with their standard deviation; then:

Condition number	amount of multicollinearity
$< 10$	weak
$> 30$	middle to strong



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

[Diagnosis and therapy](#)

[Roundup](#)

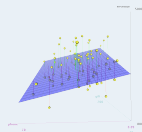
[Supplementary slides](#)

- ▶ To create a benchmark out of condition numbers: Standardise variables with their standard deviation; then:

Condition number	amount of multicollinearity
< 10	weak
> 30	middle to strong

- ▶ **condition index**  $\eta_k$  for all **eigenvalues**  $\lambda_k$ :

$$\eta_k = \sqrt{\frac{\lambda_{\max}}{\lambda_k}}$$



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

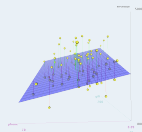
- ▶ To create a benchmark out of condition numbers: Standardise variables with their standard deviation; then:

Condition number	amount of multicollinearity
< 10	weak
> 30	middle to strong

- ▶ **condition index**  $\eta_k$  for all **eigenvalues**  $\lambda_k$ :

$$\eta_k = \sqrt{\frac{\lambda_{\max}}{\lambda_k}}$$

- ▶ One High condition index: One (nearly) multicollinear relationship



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

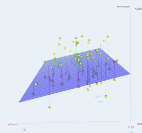
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- ▶ Which variables (including the constant) are involved with detected multicollinear relationship?



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

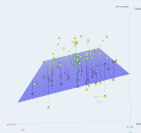
[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)

- ▶ Which variables (including the constant) are involved with detected multicollinear relationship?
- ▶ Necessary: decompose the sensitivity (variance) of the model's parameters to changes

▶ Details



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

[Diagnosis and therapy](#)

[Roundup](#)

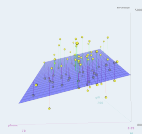
[Supplementary slides](#)



- ▶ Which variables (including the constant) are involved with detected multicollinear relationship?
- ▶ Necessary: decompose the sensitivity (variance) of the model's parameters to changes
- ▶ Result:

▶ Details

$$\pi_{jk} = \frac{\lambda_j^{-1} v_{kj}^2}{\sum_{i=0} \lambda_i^{-1} v_{ki}^2}$$



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

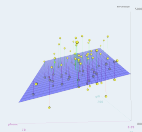
Supplementary slides

- ▶ Which variables (including the constant) are involved with detected multicollinear relationship?
- ▶ Necessary: decompose the sensitivity (variance) of the model's parameters to changes
- ▶ Result:

▶ Details

$$\pi_{jk} = \frac{\lambda_j^{-1} v_{kj}^2}{\sum_{i=0} \lambda_i^{-1} v_{ki}^2}$$

- ▶ With:



Introduction

R and RStudio

Simple linear regression

Multicollinearity

Back to Mr. Meier

Mr. Maier and his problem

Vocabulary

Geometry and

Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

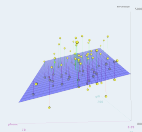
Roundup

Supplementary slides

- ▶ Which variables (including the constant) are involved with detected multicollinear relationship?
- ▶ Necessary: decompose the sensitivity (variance) of the model's parameters to changes [▶ Details](#)
- ▶ Result:

$$\pi_{jk} = \frac{\lambda_j^{-1} v_{kj}^2}{\sum_{i=0} \lambda_i^{-1} v_{ki}^2}$$

- ▶ With:
  - k: index of regression parameter  $\beta_k$



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier and his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

[Diagnosis and therapy](#)

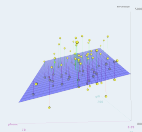
[Roundup](#)

[Supplementary slides](#)

- ▶ Which variables (including the constant) are involved with detected multicollinear relationship?
- ▶ Necessary: decompose the sensitivity (variance) of the model's parameters to changes [▶ Details](#)
- ▶ Result:

$$\pi_{jk} = \frac{\lambda_j^{-1} v_{kj}^2}{\sum_{i=0} \lambda_i^{-1} v_{ki}^2}$$

- ▶ With:
  - k: index of regression parameter  $\beta_k$
  - j: index of (large) condition index  $\eta_j$



[Introduction](#)

[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier and his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

[Solution approach](#)

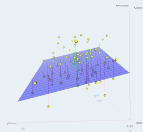
[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)

## ► Comparison of

- the condition indices (1st column)
- and the variance proportions of  $\beta_k$



[Introduction](#)

[R and RStudio](#)

[Simple linear  
regression](#)

[Multicollinearity](#)

[Back to Mr. Meier](#)

[Mr. Maier und his problem](#)

[Vocabulary](#)

[Geometry and](#)

[Multicollinearity](#)

[Common believe](#)

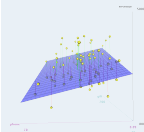
[Solution approach](#)

[Diagnosis and therapy](#)

[Roundup](#)

[Supplementary slides](#)

- ▶ Comparison of
  - the condition indices (1st column)
  - and the variance proportions of  $\beta_k$
- ▶ Result:



## Introduction

## R and RStudio

## Simple linear regression

## Multicollinearity

- Back to Mr. Meier
- Mr. Maier und his problem
- Vocabulary
- Geometry and Multicollinearity
- Common believe
- Solution approach
- Diagnosis and therapy
- Roundup

## Supplementary slides

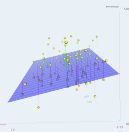
	cond.index	intercept	phone	gift	email	mail
1	1.00	0.00	0.00	0.00	0.00	0.00
2	4.05	0.00	0.01	0.00	0.00	0.00
3	5.11	0.01	0.00	0.00	0.00	0.04
4	11.35	0.99	0.01	0.00	0.00	0.00
5	83.93	0.00	0.98	0.99	0.99	0.96

- ▶ Comparison of
  - the condition indices (1st column)
  - and the variance proportions of  $\beta_k$
- ▶ Result:



	cond.index	intercept	phone	gift	email	mail
1	1.00	0.00	0.00	0.00	0.00	0.00
2	4.05	0.00	0.01	0.00	0.00	0.00
3	5.11	0.01	0.00	0.00	0.00	0.04
4	11.35	0.99	0.01	0.00	0.00	0.00
5	83.93	0.00	0.98	0.99	0.99	0.96

- ▶ **Diagnosis:** Look at the lines with high condition indices ( $> 30$ ); if there are two variance proportions  $> 0,5$ ...
- ▶ ...there is probably a dangerous multicollinearity caused by the involved variables



### Introduction

### R and RStudio

### Simple linear regression

### Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

Solution approach

Diagnosis and therapy

Roundup

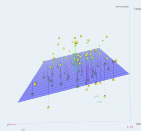
### Supplementary slides



- ▶ Comparison of
  - the condition indices (1st column)
  - and the variance proportions of  $\beta_k$
- ▶ Result:

	cond.index	intercept	phone	gift	email	mail
1	1.00	0.00	0.00	0.00	0.00	0.00
2	4.05	0.00	0.01	0.00	0.00	0.00
3	5.11	0.01	0.00	0.00	0.00	0.04
4	11.35	0.99	0.01	0.00	0.00	0.00
5	83.93	0.00	0.98	0.99	0.99	0.96

- ▶ **Diagnosis:** Look at the lines with high condition indices ( $> 30$ ); if there are two variance proportions  $> 0,5...$
- ▶ ...there is probably a dangerous multicollinearity caused by the involved variables
- ▶ Here: all 4 variables build a dangerous multicolliearity situation which results in a condition index of 83,93
- ▶ **Therapy:** Elimination of one of these variable reduces the condition number to values  $< 15$



Introduction

R and RStudio

Simple linear regression

Multicollinearity

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and Multicollinearity

Common believe

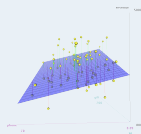
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides





## Results

- ▶ Multicollinearity is a dangerous effect if undetected
- ▶ But can be handled using
  - **condition indices** and
  - **variance decomposition proportions**
- ▶ Major data analysis software packages all support this technique  
(R, SPSS, SAS)

Introduction

R and RStudio

Simple linear  
regression

**Multicollinearity**

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and

Multicollinearity

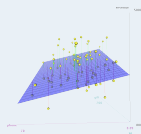
Common believe

Solution approach

Diagnosis and therapy

Roundup

Supplementary slides



## Results

- ▶ Multicollinearity is a dangerous effect if undetected
- ▶ But can be handled using
  - **condition indices** and
  - **variance decomposition proportions**
- ▶ Major data analysis software packages all support this technique  
(R, SPSS, SAS)

Thanks for your attention!

Questions?

Introduction

R and RStudio

Simple linear  
regression

**Multicollinearity**

Back to Mr. Meier

Mr. Maier und his problem

Vocabulary

Geometry and  
Multicollinearity

Common believe

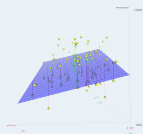
Solution approach

Diagnosis and therapy

Roundup

Supplementary slides

- 1 Columnwise standardisation of design matrix  $X$



Introduction

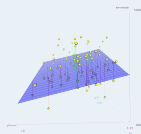
R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides

- 1 Columnwise standardisation of design matrix  $X$
- 2 Calculate eigenvalues  $\lambda_k$  und eigenvectors  $v_k$  from  $X^T X$



[Introduction](#)

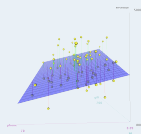
[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Supplementary slides](#)

- 1 Columnwise standardisation of design matrix  $X$
- 2 Calculate eigenvalues  $\lambda_k$  und eigenvectors  $v_k$  from  $X^T X$
- 3 Calculate the condition number  $\kappa(X)$ .



[Introduction](#)

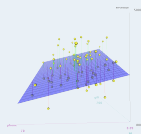
[R and RStudio](#)

[Simple linear regression](#)

[Multicollinearity](#)

[Supplementary slides](#)

- 1 Columnwise standardisation of design matrix  $X$
- 2 Calculate eigenvalues  $\lambda_k$  und eigenvectors  $v_k$  from  $X^T X$
- 3 Calculate the condition number  $\kappa(X)$ .
- 4 If  $\kappa(X) \geq 30$ : Calculate condition indices  $\eta_j$  and decompose the variance through  $\pi_{jk}$



Introduction

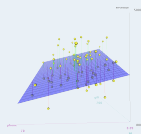
R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides

- 1 Columnwise standardisation of design matrix  $X$
- 2 Calculate eigenvalues  $\lambda_k$  and eigenvectors  $v_k$  from  $X^T X$
- 3 Calculate the condition number  $\kappa(X)$ .
- 4 If  $\kappa(X) \geq 30$ : Calculate condition indices  $\eta_j$  and decompose the variance through  $\pi_{jk}$
- 5 Write down all  $\eta_j$  and all  $\pi_{jk}$



Introduction

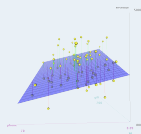
R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides

- 1 Columnwise standardisation of design matrix  $X$
- 2 Calculate eigenvalues  $\lambda_k$  und eigenvectors  $v_k$  from  $X^T X$
- 3 Calculate the condition number  $\kappa(X)$ .
- 4 If  $\kappa(X) \geq 30$ : Calculate condition indices  $\eta_j$  and decompose the variance through  $\pi_{jk}$
- 5 Write down all  $\eta_j$  and all  $\pi_{jk}$
- 6 An  $\eta_j > 30$  together with at least two  $\pi_{jk} > 0,5$  indicates dangerous multicollinearity



Introduction

R and RStudio

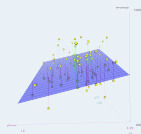
Simple linear  
regression

Multicollinearity

Supplementary slides



- 1 Columnwise standardisation of design matrix  $X$
- 2 Calculate eigenvalues  $\lambda_k$  und eigenvectors  $v_k$  from  $X^T X$
- 3 Calculate the condition number  $\kappa(X)$ .
- 4 If  $\kappa(X) \geq 30$ : Calculate condition indices  $\eta_j$  and decompose the variance through  $\pi_{jk}$
- 5 Write down all  $\eta_j$  and all  $\pi_{jk}$
- 6 An  $\eta_j > 30$  together with at least two  $\pi_{jk} > 0,5$  indicates dangerous multicollinearity
- 7 Eliminate one of the causing variables



Introduction

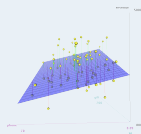
R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides

- 1 Columnwise standardisation of design matrix  $X$
- 2 Calculate eigenvalues  $\lambda_k$  und eigenvectors  $v_k$  from  $X^T X$
- 3 Calculate the condition number  $\kappa(X)$ .
- 4 If  $\kappa(X) \geq 30$ : Calculate condition indices  $\eta_j$  and decompose the variance through  $\pi_{jk}$
- 5 Write down all  $\eta_j$  and all  $\pi_{jk}$
- 6 An  $\eta_j > 30$  together with at least two  $\pi_{jk} > 0,5$  indicates dangerous multicollinearity
- 7 Eliminate one of the causing variables
- 8 Back to 1.



Introduction

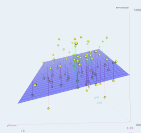
R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides

- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$



[Introduction](#)

[R and RStudio](#)

[Simple linear  
regression](#)

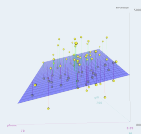
[Multicollinearity](#)

[Supplementary slides](#)

- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

$$v_1 x_1 + \dots + v_m x_m = Xv = a \approx 0$$

(If not all scalars  $v_1 \dots v_m$  are 0)



Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

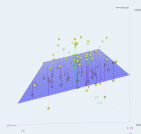
Supplementary slides

- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

$$v_1 x_1 + \dots + v_m x_m = Xv = a \approx 0$$

(If not all scalars  $v_1 \dots v_m$  are 0)

- ▶ Therefore wanted: vector  $v$  with definit length (e.g. 1), such that  $|a|$  becomes small



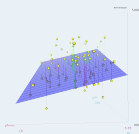
Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

$$v_1 x_1 + \dots + v_m x_m = Xv = a \approx 0$$

(If not all scalars  $v_1 \dots v_m$  are 0)

- ▶ Therefore wanted: vector  $v$  with definit length (e.g. 1), such that  $|a|$  becomes small
- ▶ That leads to a minimisation problem:

$$\min_v |a|^2 = \min_v a^T a = \min_v v^T X^T X v \quad \text{with} \quad |v|^2 = v^T v = 1$$

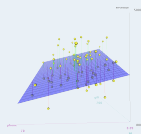
Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

$$v_1 x_1 + \dots + v_m x_m = Xv = a \approx 0$$

(If not all scalars  $v_1 \dots v_m$  are 0)

- ▶ Therefore wanted: vector  $v$  with definit length (e.g. 1), such that  $|a|$  becomes small
- ▶ That leads to a minimisation problem:

$$\min_v |a|^2 = \min_v a^T a = \min_v v^T X^T X v \quad \text{with} \quad |v|^2 = v^T v = 1$$

- ▶ Lagrange multipliers:

$$L(v, \lambda) = v^T X^T X v + \lambda(1 - v^T v)$$

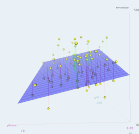
Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides



- ▶ Nearly multicollinearity: Nearly linear dependency of the columns of  $X$
- ▶  $\Rightarrow$  there is a vector  $v \neq 0$ , such that

$$v_1 x_1 + \dots + v_m x_m = Xv = a \approx 0$$

(If not all scalars  $v_1 \dots v_m$  are 0)

- ▶ Therefore wanted: vector  $v$  with definit length (e.g. 1), such that  $|a|$  becomes small
- ▶ That leads to a minimisation problem:

$$\min_v |a|^2 = \min_v a^T a = \min_v v^T X^T X v \quad \text{with} \quad |v|^2 = v^T v = 1$$

- ▶ Lagrange multipliers:

$$L(v, \lambda) = v^T X^T X v + \lambda(1 - v^T v)$$

- ▶ Derivation results in necessary condition for minimum:

$$X^T X v = \lambda v$$

Introduction

R and RStudio

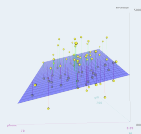
Simple linear regression

Multicollinearity

Supplementary slides



- ▶  $X^T X v = \lambda v$  is an eigenvalue problem



[Introduction](#)

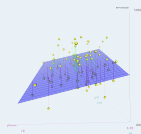
[R and RStudio](#)

[Simple linear  
regression](#)

[Multicollinearity](#)

[Supplementary slides](#)

- ▶  $X^T X v = \lambda v$  is an eigenvalue problem
- ▶ Which Eigenvalue  $\lambda$  minimises  $|a|$ ?



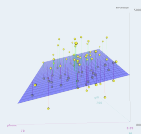
Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

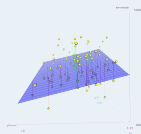
Supplementary slides



- ▶  $X^T X v = \lambda v$  is an eigenvalue problem
- ▶ Which Eigenvalue  $\lambda$  minimises  $|a|$ ?
- ▶ Trick: Multiply  $X^T X v = \lambda v$  with  $v^T$

$$\Rightarrow v^T X^T X v = \lambda v^T v \quad \Leftrightarrow \quad |a|^2 = \lambda \quad \Leftrightarrow \quad |a| = \sqrt{\lambda}$$

[Introduction](#)[R and RStudio](#)[Simple linear regression](#)[Multicollinearity](#)[Supplementary slides](#)

[Introduction](#)[R and RStudio](#)[Simple linear regression](#)[Multicollinearity](#)[Supplementary slides](#)

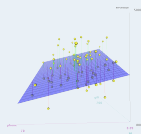
- ▶  $X^T X v = \lambda v$  is an eigenvalue problem

- ▶ Which Eigenvalue  $\lambda$  minimises  $|a|$ ?

- ▶ Trick: Multiply  $X^T X v = \lambda v$  with  $v^T$

$$\Rightarrow v^T X^T X v = \lambda v^T v \Leftrightarrow |a|^2 = \lambda \Leftrightarrow |a| = \sqrt{\lambda}$$

- ▶ Smallest Eigenvalue  $\lambda_1$  for Eigenvector  $v_1$  minimises  $|a|$  and shows strongest (nearly-)linear dependency

[Introduction](#)[R and RStudio](#)[Simple linear  
regression](#)[Multicollinearity](#)[Supplementary slides](#)

- ▶  $X^T X v = \lambda v$  is an eigenvalue problem

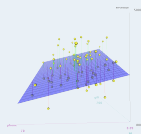
- ▶ Which Eigenvalue  $\lambda$  minimises  $|a|$ ?

- ▶ Trick: Multiply  $X^T X v = \lambda v$  with  $v^T$

$$\Rightarrow v^T X^T X v = \lambda v^T v \quad \Leftrightarrow \quad |a|^2 = \lambda \quad \Leftrightarrow \quad |a| = \sqrt{\lambda}$$

- ▶ Smallest Eigenvalue  $\lambda_1$  for Eigenvector  $v_1$  minimises  $|a|$  and shows strongest (nearly-)linear dependency

- ▶ Sort eigenvalues according to size:  $(\lambda_2, \dots)$  and Eigenvectors  $v_2, \dots$  gives the other values  $a_2, \dots$

[Introduction](#)[R and RStudio](#)[Simple linear regression](#)[Multicollinearity](#)[Supplementary slides](#)

- ▶  $X^T X v = \lambda v$  is an eigenvalue problem

- ▶ Which Eigenvalue  $\lambda$  minimises  $|a|$ ?

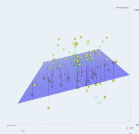
- ▶ Trick: Multiply  $X^T X v = \lambda v$  with  $v^T$

$$\Rightarrow v^T X^T X v = \lambda v^T v \Leftrightarrow |a|^2 = \lambda \Leftrightarrow |a| = \sqrt{\lambda}$$

- ▶ Smallest Eigenvalue  $\lambda_1$  for Eigenvector  $v_1$  minimises  $|a|$  and shows strongest (nearly-)linear dependency
- ▶ Sort eigenvalues according to size:  $(\lambda_2, \dots)$  and Eigenvectors  $v_2, \dots$  gives the other values  $a_2, \dots$
- ▶ Proportion of largest and smallest eigenvalue:

$$\kappa(X) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

is called **condition number**



$$\text{Var}(\hat{\beta}_k) = \sigma^2 (X^T X)^{-1}_{kk} = \sigma^2 (V \Lambda^{-1} V^T)_{kk} = \sigma^2 \sum_{j=0}^m \frac{v_{kj}^2}{\lambda_j}$$

With:

$\Lambda$  diagonal matrix of eigenvalues  $\lambda_1, \dots$  and

$V$  as matrix of eigen vectors  $v_1, \dots$

- ▶ Meaning: Small eigenvalue und large component in eigenvector (both hints for multicollinearity) result in large proportion in variance of  $\beta$ .
- ▶ Large variance of  $\beta$ : Instable model
- ▶ Weight of this variance proportion (Summanden in Formel) divided through full variance: **variance decomposition proportion**  $\pi_{jk}$

$$\pi_{jk} = \frac{\lambda_j^{-1} v_{kj}^2}{\sum_{i=0}^m \lambda_i^{-1} v_{ki}^2}$$

Introduction

R and RStudio

Simple linear  
regression

Multicollinearity

Supplementary slides