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

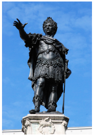
Stefan Etschberger
Augsburg University of Applied Sciences

Who is talking to you?

- Stefan Etschberger
- 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
- 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



- 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

Mr. Maier's data

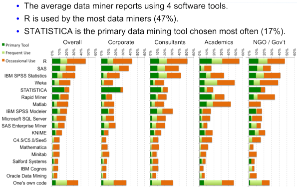
action	revenue	telephone	e-mail	mail	gift
1	10193.70	186.20	158.60	26.90	11.10
2	4828.20	470.30	55.00	14.40	20.30
3	11139.30	41.80	154.70	20.90	12.40
4	5030.10	530.10	79.80	21.70	17.00
⋮					

- 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....

Data analysis, Regression and Beyond: Table of Contents

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
- Major advantage (at second thought): No point and click tool



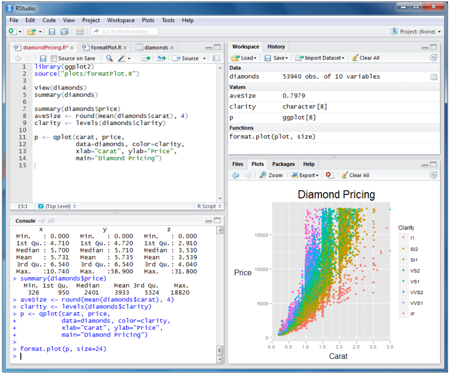
source: `http://goo.gl/axhGhh`

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

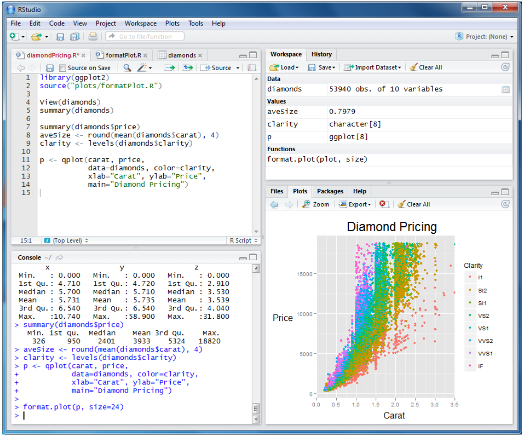
RStudio™

Free & Open-Source IDE for R



Getting to know RStudio

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



data inspection

```
# read in data from comma-seperated 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
```

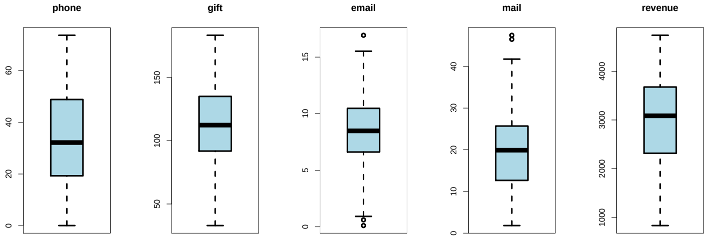
Overview over all variables

```
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  
```

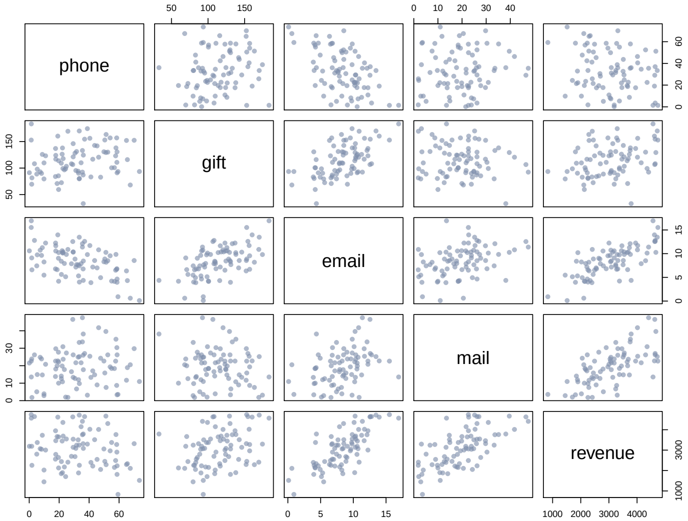
Boxplots

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



Visualize pairs

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



data inspection

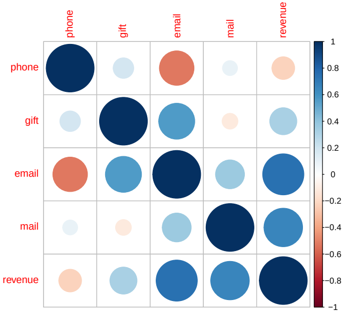
List all Correlations

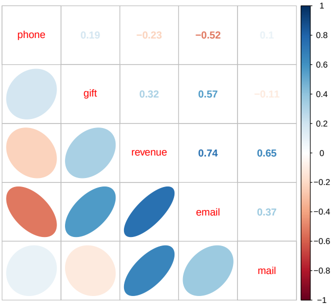
```
cor.MyCheeseData = cor(MyCheeseData)
cor.MyCheeseData

##             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
```

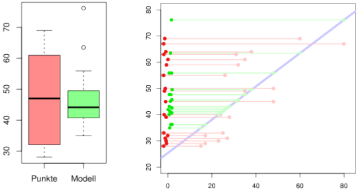
Visualize correlation

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





Data analysis, Regression and Beyond: Table of Contents



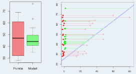
3 Revision: Simple linear regression

Premier German Soccer League 2008/2009

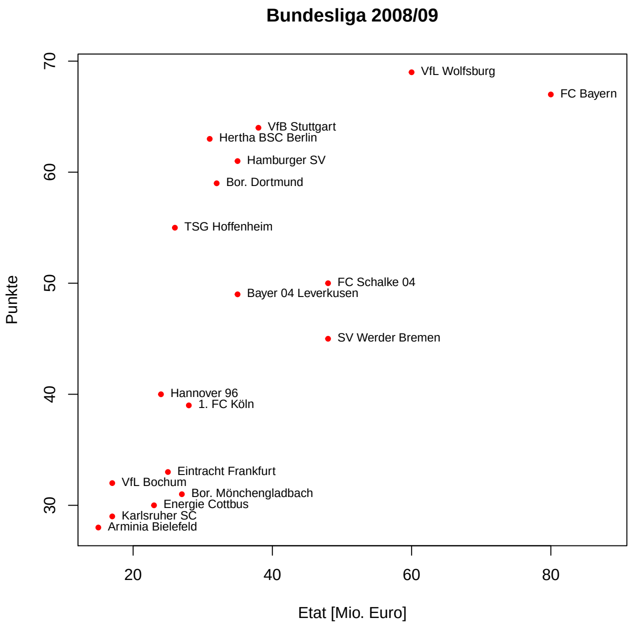
- Given: data for all 18 clubs in the German Premier Soccer League in the season 2008/09
- variables: **Budget** for season (only direct salaries for players)
- and: **resulting** table points at the end of the season

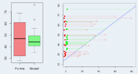
	Etat	Punkte
FC Bayern	80	67
VfL Wolfsburg	60	69
SV Werder Bremen	48	45
FC Schalke 04	48	50
VfB Stuttgart	38	64
Hamburger SV	35	61
Bayer 04 Leverkusen	35	49
Bor. Dortmund	32	59
Hertha BSC Berlin	31	63
1. FC Köln	28	39
Bor. Mönchengladbach	27	31
TSG Hoffenheim	26	55
Eintracht Frankfurt	25	33
Hannover 96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

(Source: Welt)



data in scatter plot





Trend as a linear model

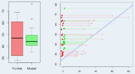
- Is it possible to find a simple function which can describe the dependency of the **end-of-season-points** versus the **club budget**?
- In general: Description of a variable Y as a function of another variable X:

$$y = f(x)$$

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

$$\boxed{y = a + b\,x}$$

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



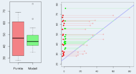
Sum of error squares

- using the regression model; per data object:

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

- ϵ_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

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



Best solution

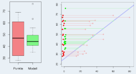
- Best and unique solution:

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

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

- regression line:

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



Soccer example

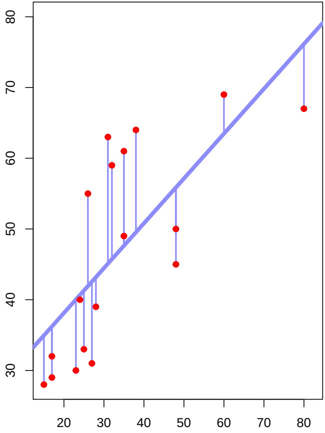
- Calculation of the soccer model
- With: table points $\hat{=}\, y$ and budget $\hat{=}\, x$:

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

$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33{,}83 \cdot 46{,}89}{25209 - 18 \cdot 33{,}83^2} \approx 0{,}634$$

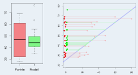
$$\Rightarrow \hat{a} = 46{,}89 - \hat{b} \cdot 33{,}83 \approx 25{,}443$$

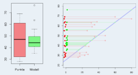
- model: $\hat{y} = 25{,}443 + 0{,}634 \cdot x$



- prognosis for budget $= 30$:

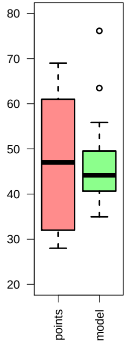
$$\hat{y}(30) = 25{,}443 + 0{,}634 \cdot 30 \approx 44{,}463$$

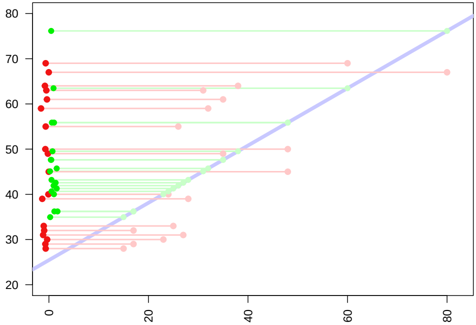




Variance and information

- **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":

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

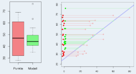
Coefficient of determination

- Quality criterion for regression model: Coefficient of determination:

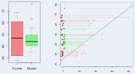
$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2} = r^2 \in [0; 1]$$

- 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:

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



Regression: Four one-dimensional examples



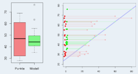
- Famous data from the 1970ies:

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**)

Residual analysis

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

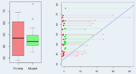


Properties of residual distribution

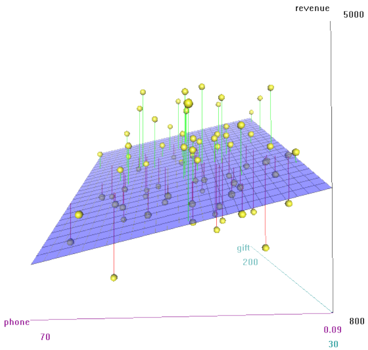
- Preferably no systematic pattern
- No change of variance dependent of $\hat{y}_i$ (Homoscedasticity)
- Necessary for inferential analysis: Approximately normal distributed residuals (q-q-plots)

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

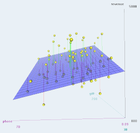


Data analysis, Regression and Beyond: Table of Contents



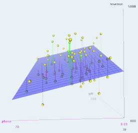
Back to Mr. Meier's data

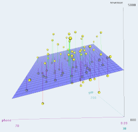
	phone	gift	email	mail	revenue
1	29.36	146.14	10.32	13.36	3137.85
2	8.75	125.82	11.27	14.72	3728.11
3	36.15	124.51	8.45	17.72	3084.75
4	51.20	129.36	10.27	39.59	4667.90
5	51.36	163.42	8.19	7.57	2286.41
6	34.65	110.04	7.89	21.68	4147.61
7	19.65	113.88	10.23	22.17	3648.22
8	17.51	84.04	6.79	13.82	2558.09
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
		⋮			



- 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:

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





Result: Model

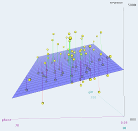
```
##
## Call:
## lm(formula = revenue ~ phone + gift + mail + email, data = MyCheeseData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1084.8  -348.9   -46.5   333.1  1010.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    741.5      250.3    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...

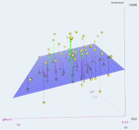
Small corrections to data

- 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
	⋮	



Corrected model



- Model from corrected data

```
##
## Call:
## lm(formula = revenue ~ phone + gift + mail + email, dat
##
## 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
```

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

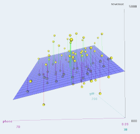
A few technical terms

- 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: $y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_m \cdot x_m + u$

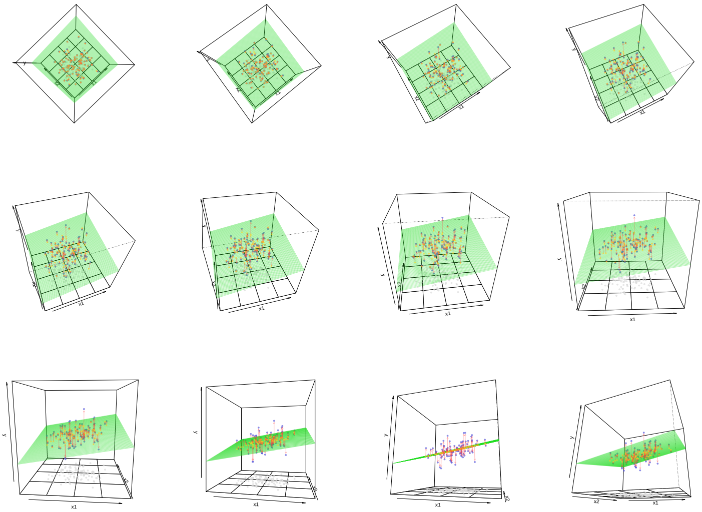
$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}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

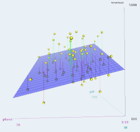
- The error term u is the portion of the data which can not be described by the model
- Typical: Estimation of the „best“ model parameters $\hat{\beta}_0, \dots, \hat{\beta}_m$ using a least-square analysis:

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



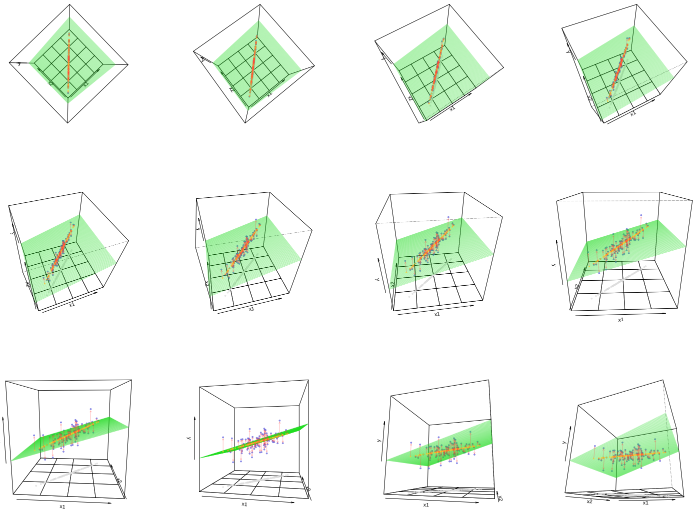
- two-dimensional example
- stable model possible

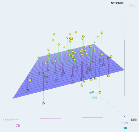


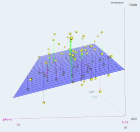


Two independent variables

- Perfect multicollinearity
- no regression model possible

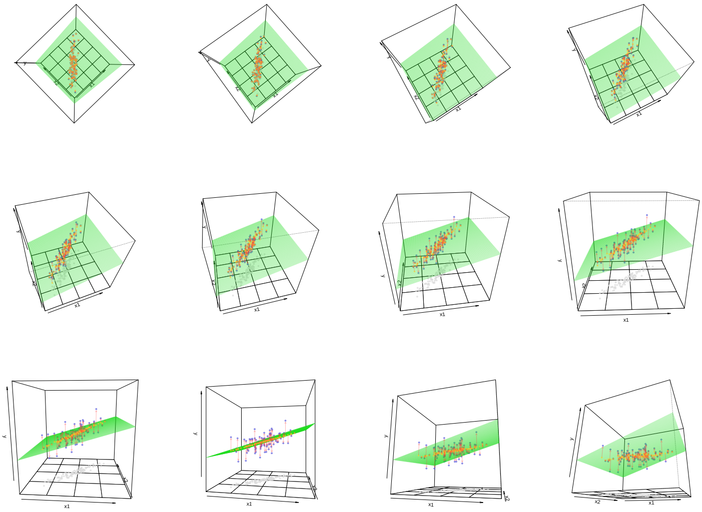






Two independent variables

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



Idea for diagnosis: Correlation?

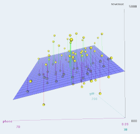
- Maybe the correlation of the independent variables is a good measure?
- But: Perfect multicollinearity between three or more vectors (which are not pairwise correlated)
- Simple Example:

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

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

- Correlation is sufficient, but not necessary for multicollinearity



Technical Consideration

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

$$v_0 x_0 + \ldots + v_m x_m = Xv = a \approx 0$$

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

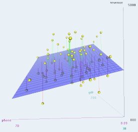
- Therefore wanted: vector v with normed length (e.g. 1), such that $|a|$ becomes small
- Solution: Smallest eigenvalue λ_0 (with its corresponding eigenvector v_0) from $X^T X$ shows strongest nearly linear dependency

Details

- Proportion of largest and smallest Eigenvalue as per

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

is called condition number



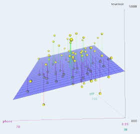
- To create a benchmark out of condition numbers: Standardise variables with their standard deviation; then:

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

- condition index η_k for all **eigenvalues** λ_k:

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

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

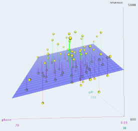


Variance decomposition proportions

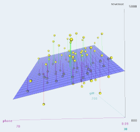
- Which variables (including the constant) are involved with detected multicollinear relationship?
- Necessary: decompose the sensitivity (variance) of the model's parameters to changes ▸ Details
- Result:

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

- With:
  - k: index of regression parameter β_k
  - j: index of (large) condition index η_j



Back to Mr. Maier's data



- Comparison of
  - the condition indices (1st column)
  - and the variance proportions of β_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

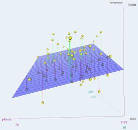
- **Diagnosis:** Look at the lines with high condition indices (> 30); if there are two variance proportions $> 0{,}5$...
- ...there is probably a dangerous multicollinearity caused by the involved variables
- Here: all 4 variables build a dangerous multicollienarity situation which results in a condition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

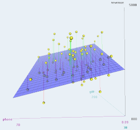
Results

- 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?





Recipe to diagnose and eliminate multicollinearities

1. Columnwise standardisation of design matrix X
2. Calculate eigenvalues λ_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 η_j and decompose the variance through π_{jk}
5. Write down all η_j and all π_{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.

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

$$\boxed{v_1 x_1 + \ldots + v_m x_m = Xv = a \approx 0}$$

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

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

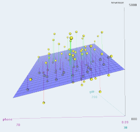
$$\min_v |a|^2 = \min_v a^T a = \min_v v^T X^T X v \quad \text{with} \quad |v|^2 = v^T v = 1$$

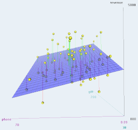
- Lagrange multipliers:

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

- Derivation results in necessary condition for minimum:

$$\boxed{X^T X v = \lambda v}$$





Theory 2

Zurück

- $X^T X v = \lambda v$ is an eigenvalue problem
- Which Eigenvalue λ 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 λ_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:

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

is called condition number

decomposition of condition indices

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

With:
Λ diagonal matrix of eigenvalues $\lambda_1, \ldots$ and

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

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

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