# Semesterarbeit - Statistische Datenanalyse

**CAS Statistische Datenanalyse und Datenvisualisierung, FS23**
**Fernfachhochschule Schweiz (FFHS)**

Basel, 23.06.2023

**Cyrill Martin**
Matrikel-Nr. 02-909-323
cyrill.martin@students.ffhs.ch

**Inhaltsverzeichnis**

# Einleitung

## Ausgangslage

Die vorliegende Semesterarbeit entstand im Rahmen einer wiederkehrenden Zusammenarbeit mit einem Wissenschaftsverlag, der sich auf die Veröffentlichung biomedizinischer Inhalte spezialisiert hat. Im Zuge dieser Zusammenarbeit habe ich Daten zu Forschungsprojekten verarbeitet, die durch den Schweizerischen Nationalfonds (SNF) unterstützt wurden. Zusätzlich dazu habe ich Daten zur Anzahl der Zitierungen wissenschaftlicher Publikationen besorgt, die im Rahmen dieser Projekte veröffentlicht wurden. Dadurch wurde die übergeordnete Frage aufgeworfen, ob Projekte, die eine höhere finanzielle Unterstützung erhalten, auch vermeintlich relevantere Publikationen hervorbringen, die dementsprechend häufiger zitiert werden.

## Daten

### Datenquellen und -verarbeitung

Der verwendete Datensatz basiert auf zwei verschiedenen Datenquellen:

- Die Informationen zu den Forschungsprojekten und den daraus resultierenden wissenschaftlichen Publikationen wurden vom SNF-Datenportal heruntergeladen. Der letzte Download fand am 16.05.2023 statt.
- Die Angaben zur Anzahl der Zitierungen für jede einzelne Publikation wurden über die Crossref Cited-by API bezogen. Der Datensatz wurde zuletzt in der Nacht vom 16.05.2023 auf den 17.05.2023 abgerufen.

Die Datenverarbeitung, einschließlich des Bezugs der Zitierungsinformationen, erfolgte mithilfe von R-Code, der in vier aufeinanderfolgenden Jupyter Notebooks ausgeführt wurde. Alle Verarbeitungsschritte können {{hier}} eingesehen werden.

Für die Analyse wurden die Daten auf Projekte mit bestimmten Merkmalen beschränkt:

- Es wurden nur abgeschlossene Projekte berücksichtigt.
- Es wurden nur Projekte einbezogen, deren Fördermittel ab dem Jahr 2012 bewilligt wurden.
- Es wurden nur Projekte im Bereich der biomedizinischen Forschung einbezogen.
- Es wurden nur Projekte berücksichtigt, die mindestens eine wissenschaftliche Publikation generiert haben.
- Es wurden nur Projekte einbezogen, für die ich Daten zu den Zitierungen abrufen konnte und deren wissenschaftliche Publikationen insgesamt mindestens eine Zitierung erhalten haben.

Dadurch entstand ein Datensatz mit insgesamt 1'923 Forschungsprojekten.

### Variabeln

Die wesentlichen Variablen im Datensatz, die für die Analyse relevant sind, werden in der folgenden Tabelle kurz beschrieben. Um Projekte unterschiedlicher Dauer besser vergleichen zu können, wurden für die gewährten Gelder und die Anzahl veröffentlichter Publikationen Durchschnittswerte pro Jahr berechnet. Darüber hinaus wurde für die Anzahl der Zitierungen der Durchschnitt der Zitierungen pro Publikation ermittelt.

| Variable | Beschreibung |
|---|---|

| | |
|---|---|
| GrantDurationYears | Die Dauer des Projekts in Jahren. |
| AmountGrantedAllSets | Der insgesamt gewährte Förderungsbetrag für das Projekt. |
| AmountPerYear | Der durchschnittliche Förderungsbetrag pro Jahr. |
| FundingInstrumentLevel1 | Das Förderungsinstrument, innerhalb dessen die Gelder gewährt wurden. |
| MainDiscipline_Level2 | Das Forschungsgebiet eines Projekts. |
| NrCitablePublications | Die Gesamtanzahl veröffentlichter Publikationen. |
| NrCitablePublicationsPerYear | Die durchschnittliche Anzahl veröffentlichter Publikationen pro Jahr. |
| NrCitationsTotal | Die Gesamtanzahl der Zitierungen aller Publikationen eines Projekts. |
| NrCitationsPerPublication | Die durchschnittliche Anzahl der Zitierungen pro Publikation, die ein Projekt veröffentlicht hat. |

## R-Bibliotheken

Nachfolgend werden die R-Bibliotheken geladen, die für die Auswertungen relevant sind.

In [ ]:
```
library(tidyverse)
library(gridExtra)
library(ggpubr)
library(psych)
library(car)
library(lmtest)
```

# Deskriptive Datenanalyse

Im ersten Kapitel der Arbeit sollen die Daten beschrieben und zusammengefasst werden. Es werden erste Muster, Trends und Zusammenhänge aufgezeigt, ohne dabei detaillierte statistische Auswertungen vorzunehmen.

In [ ]:
```
# Daten importieren
data <- read_delim("BiomedGrants_20230601.csv", delim = ";", col_names = TRUE, col_se
  "GrantDurationYears",
  "AmountGrantedAllSets",
  "AmountPerYear",
  "FundingInstrumentLevel1",
  "MainDiscipline_Level2",
  "NrCitablePublications",
  "NrCitablePublicationsPerYear",
  "NrCitationsTotal",
  "NrCitationsPerPublication"
))
```

In [ ]:
```
# Factors generieren für die kategorialen Variabeln
cols <- c(
    "FundingInstrumentLevel1",
    "MainDiscipline_Level2"
)

data[, cols] <- lapply(data[, cols], as.factor)
```

In [ ]:
```
# Sicherstellen, dass immer alle Spalten eines Dataframes angezeigt werden
options(repr.matrix.max.rows = Inf, repr.matrix.max.cols = Inf)
```

In [ ]:
```
# Ein Blick in den Datensatz
dim(data)
head(data)
```

A tibble: 6 × 9

| GrantDurationYears <dbl> | AmountGrantedAllSets <dbl> | AmountPerYear <dbl> | FundingInstrumentLevel1 <fct> | MainDiscipline_Level2 <fct> |
|---|---|---|---|---|
| 4.4958904 | 291191 | 64768.26 | Projects | Preventive Medicine |
| 4.9232877 | 640027 | 129999.92 | Programmes | Social Medicine |
| 0.9972603 | 42800 | 42917.58 | Careers | Basic Medical Sciences |
| 2.9972603 | 570301 | 190274.10 | Projects | Experimental Medicine |
| 2.9972603 | 593352 | 197964.79 | Careers | Clinical Medicine |
| 1.9972603 | 100000 | 50068.59 | Careers | Basic Medical Sciences |

## "Summary"

In einem ersten Schritt soll ein allgemeiner Überblick über jede Spalte im Datensatz gewonnen werden.

In [ ]:
```
summary(data)
```

```
 GrantDurationYears AmountGrantedAllSets AmountPerYear
 Min.   :0.08219    Min.   :    1500     Min.   :   4387
 1st Qu.:1.78904    1st Qu.:  131464     1st Qu.:  72529
 Median :3.00000    Median :  387326     Median : 118500
 Mean   :2.83549    Mean   :  451050     Mean   : 147089
 3rd Qu.:3.74795    3rd Qu.:  598508     3rd Qu.: 175000
 Max.   :6.91507    Max.   :10528000     Max.   :2632000
          FundingInstrumentLevel1            MainDiscipline_Level2
 Careers                : 631       Basic Medical Sciences:658
 Infrastructure         :  28       Clinical Medicine     :463
 Programmes             : 185       Experimental Medicine :586
 Projects               :1042       Preventive Medicine   :166
 Science communication:  37         Social Medicine       : 50

 NrCitablePublications NrCitablePublicationsPerYear NrCitationsTotal
 Min.   :  1.000       Min.   : 0.1714              Min.   :    1.0
 1st Qu.:  2.000       1st Qu.: 1.0000              1st Qu.:   36.0
 Median :  5.000       Median : 1.9982              Median :  121.0
 Mean   :  7.827       Mean   : 2.9914              Mean   :  336.4
 3rd Qu.: 10.000       3rd Qu.: 3.6049              3rd Qu.:  348.0
 Max.   :235.000       Max.   :58.7500              Max.   :11713.0
 NrCitationsPerPublication
 Min.   :  0.3333
 1st Qu.: 11.0000
 Median : 23.5000
 Mean   : 42.4710
 3rd Qu.: 48.9326
 Max.   :765.3333
```

Bei Betrachtung der "Summary" ist ersichtlich, dass das allgemeine Förderungsinstrument "Projects" die Daten dominiert. Zudem sind die Forschungsgebiete "Social Medicine" und "Preventive Medicine" vergleichsweise unterrepräsentiert, was darauf hindeutet, dass in diesen Bereichen weniger Forschung betrieben wird.

Die Spannweiten der metrischen Variablen sind relativ groß, wobei sich zwischen dem 3. Quantil und dem Maximalwert teilweise erhebliche Abstände befinden. Die Verteilung der Daten ist eher schief, anstatt normalverteilt. Im nächsten Abschnitt wird eine detailliertere Untersuchung der Verteilungen der Daten durchgeführt.

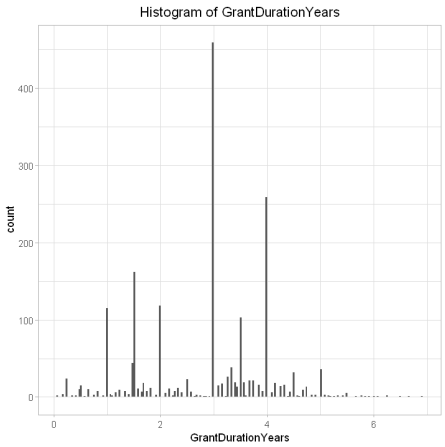
## Verteilungen

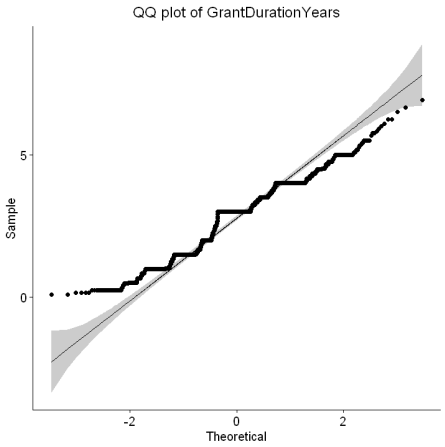
In [ ]:
```r
# Eine Funktion, um pro gewünschter Spalte ein Histogramm, ein QQ-Plot und ein Boxplo
show_plots <- function(data=data, column=column, variable=variable, bins, log=FALSE)

    if (log) {
        title <- paste("log(", variable, ")", sep = "")

        hist_plot <- ggplot(data=data, aes(x=log(!!sym(variable))))
        qq_plot <- ggqqplot(log(column), title = paste("QQ plot of", title, sep = " "
        boxplot_plot <- ggplot(data=data, aes(x="", y=log(!!sym(variable))))

        whisker_range <- 1.5 * IQR(log(column))
        lower_whisker <- quantile(log(column), 0.25) - whisker_range
        upper_whisker <- quantile(log(column), 0.75) + whisker_range

        smry <- summary(log(column))
    } else {
        title <- variable

        hist_plot <- ggplot(data=data, aes(x=!!sym(variable)))
        qq_plot <- ggqqplot(column, title = paste("QQ plot of", title, sep = " "))
        boxplot_plot <- ggplot(data=data, aes(x="", y=!!sym(variable)))

        whisker_range <- 1.5 * IQR(column)
        lower_whisker <- quantile(column, 0.25) - whisker_range
        upper_whisker <- quantile(column, 0.75) + whisker_range

        smry <- summary(column)
    }

    # Final histogram
    hist_plot <- hist_plot +
        geom_histogram(bins=bins) +
        theme_light() +
        ggtitle(paste("Histogram of", title, sep=" ")) +
        theme(plot.title = element_text(hjust = 0.5))

    # Final qq plot
    qq_plot <- qq_plot +
        theme(plot.title = element_text(hjust = 0.5))

    # Final boxplot
    boxplot_plot <- boxplot_plot +
        geom_jitter(width = 0.2, alpha = 0.25, height = 0.05) +
        geom_errorbar(aes(ymin = lower_whisker, ymax = upper_whisker), width = 0.2) +
        geom_boxplot(coef=1.5, outlier.shape = NA) +
        ggtitle(paste("Boxplot of", title, sep = " ")) +
        theme(axis.ticks.x=element_blank(),
            axis.text.x=element_blank(),
            panel.background=element_rect(fill="white"),
            panel.grid.major.y=element_line(colour="lightgrey"),
            panel.grid.minor.y=element_line(colour="lightgrey"),
            plot.title = element_text(hjust = 0.5)
           )

    # Print summary of column
    print(smry)

    options(repr.plot.width=18, repr.plot.height=6)
    grid.arrange(hist_plot, qq_plot, boxplot_plot, ncol=3)
}
```

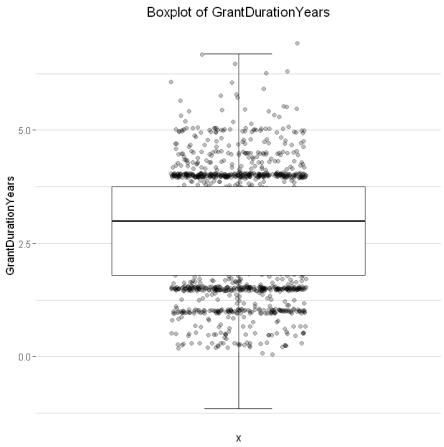
## Projektdauer in Jahren (GrantDurationYears)

In [ ]:
```r
show_plots(data=data, column=data$GrantDurationYears, variable="GrantDurationYears",
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.08219 1.78904 3.00000 2.83549 3.74795 6.91507 
```



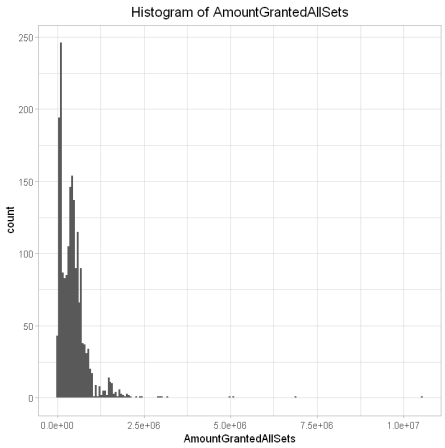



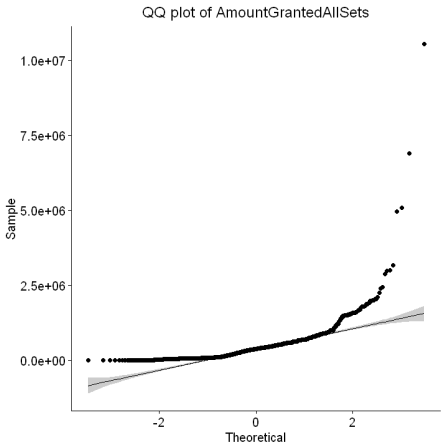

Die Daten zur Projektdauer weisen keine vollständig schiefe oder normalverteilte Verteilung auf. Die Mehrheit der Projekte hat eine Dauer von ungefähr drei Jahren. Es sind kaum signifikante Ausreißer vorhanden, abgesehen von einem einzigen Datenpunkt.

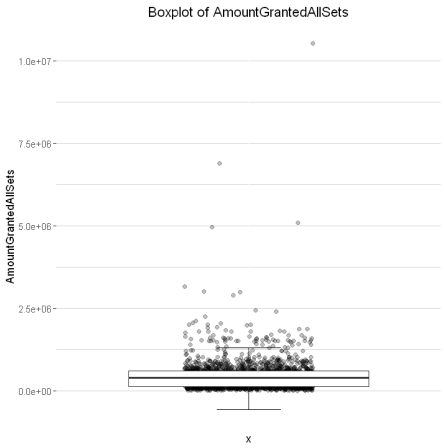
## Totaler Förderungsbetrag (AmountGrantedAllSets)

In [ ]:
```
show_plots(data=data, column=data$AmountGrantedAllSets, variable="AmountGrantedAllSet
```

```
   Min.  1st Qu.    Median      Mean  3rd Qu.      Max. 
   1500   131464    387326    451050    598508 10528000 
```



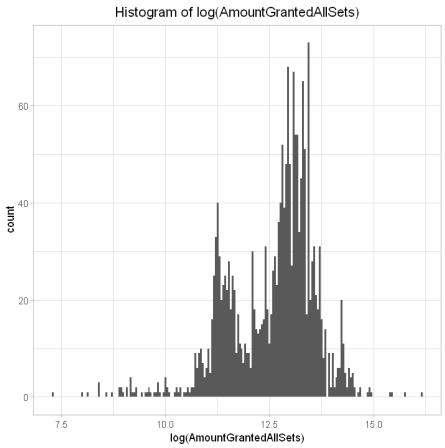



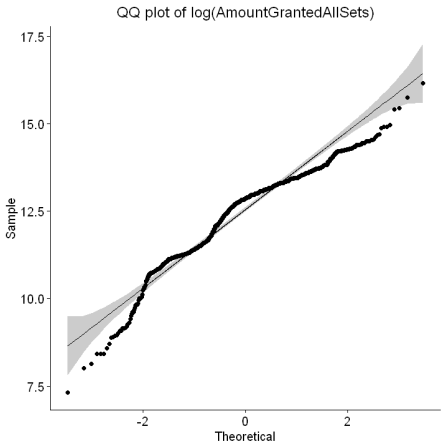



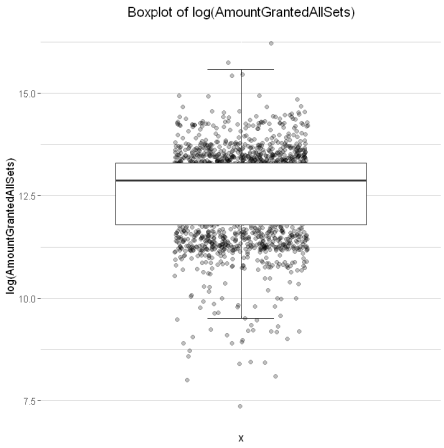

Mit log-Transformation:

In [ ]:
```
show_plots(data=data, column=data$AmountGrantedAllSets, variable="AmountGrantedAllSet
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  7.313  11.786  12.867  12.587  13.302  16.170 
```



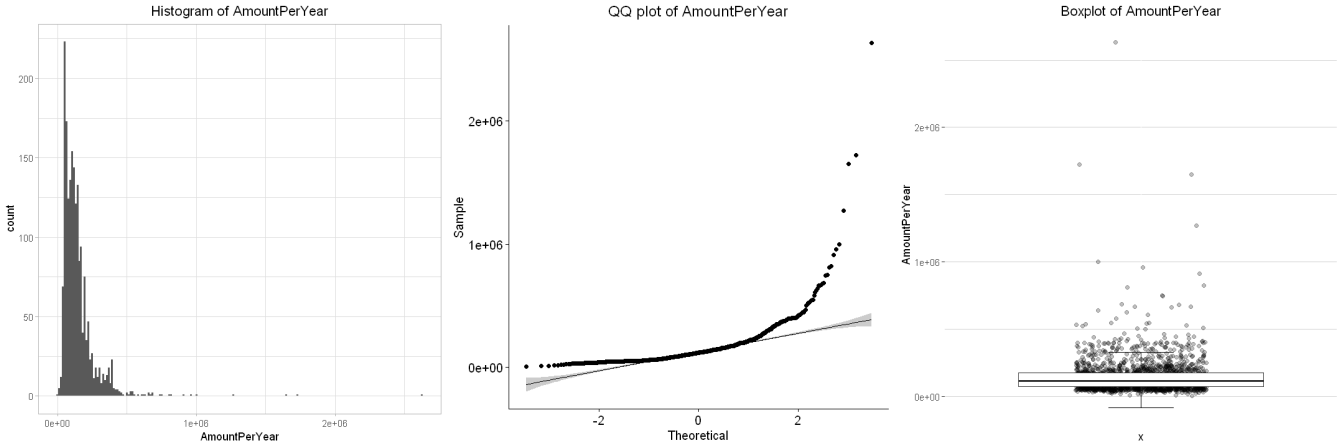




Die log-transformierten Daten weisen eine etwas gleichmäßigere Verteilung auf, obwohl sie immer noch nicht normal verteilt sind. Nach der Transformation sind immer noch einige Ausreißer vorhanden, insbesondere im Bereich der niedrigen Werte.

## Förderungsbetrag pro Jahr (AmountPerYear)

In [ ]:
```
show_plots(data=data, column=data$AmountPerYear, variable="AmountPerYear", bins=200,
```

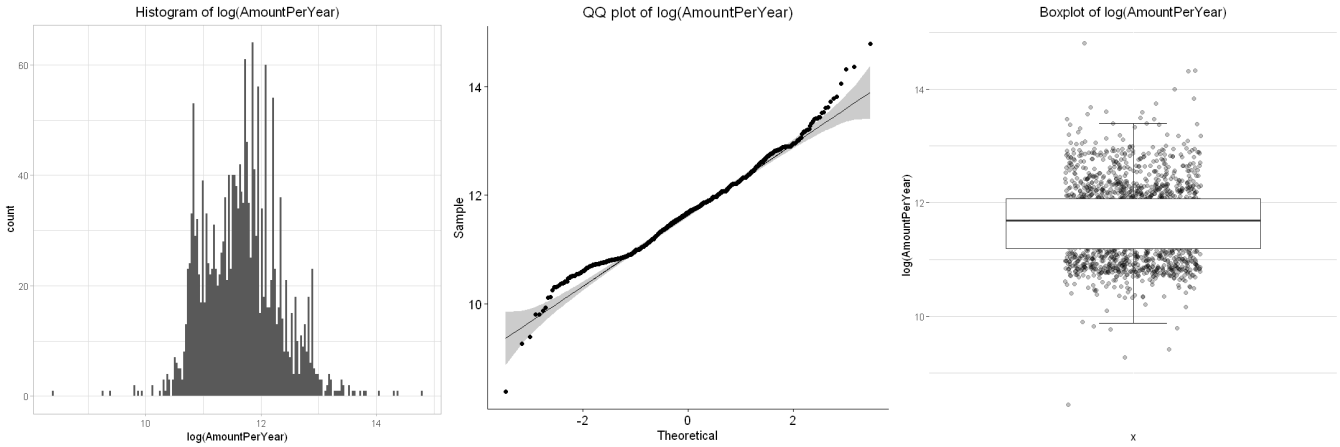
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   4387   72529  118500  147089  175000 2632000 
```




Mit log-Transformation:

In [ ]:
```
show_plots(data=data, column=data$AmountPerYear, variable="AmountPerYear", bins=200,
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  8.386  11.192  11.683  11.675  12.073  14.783 
```

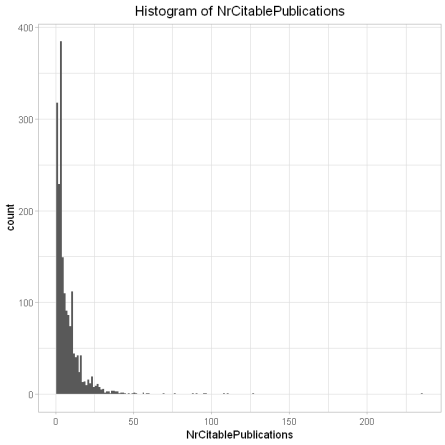


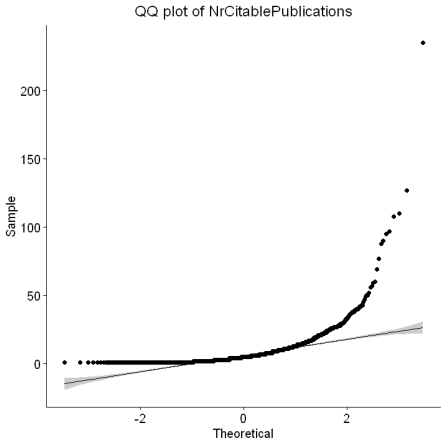

Die log-transformierten Beträge pro Jahr zeigen eine verbesserte Verteilung, weichen jedoch in Richtung der Ränder von einer optimalen Normalverteilung ab. Es sind einige Ausreißer sowohl nach oben als auch nach unten vorhanden.

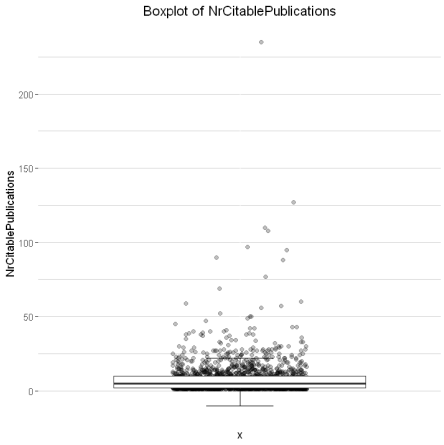
## Totale Anzahl Publikationen (NrCitablePublications)

In [ ]:
```
show_plots(data=data, column=data$NrCitablePublications, variable="NrCitablePublicati
```

```
   Min. 1st Qu.   Median    Mean 3rd Qu.    Max. 
  1.000   2.000    5.000   7.827  10.000 235.000 
```



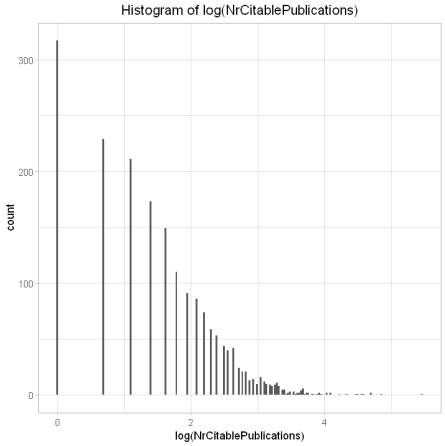



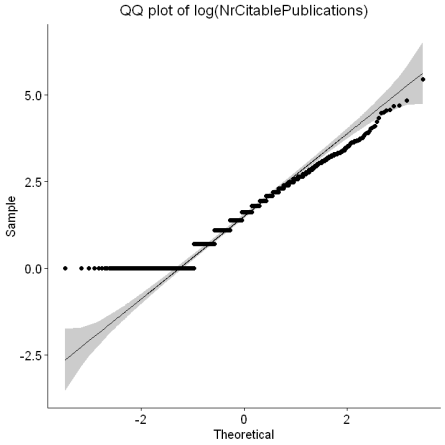



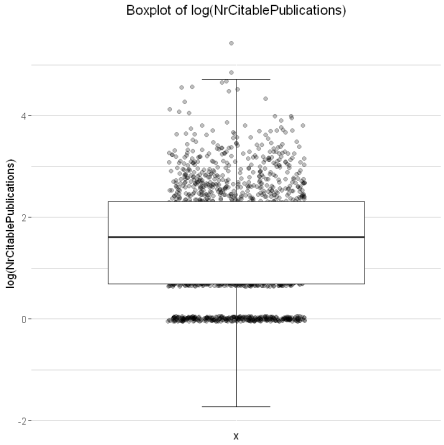

Mit log-Transformation:

In [ ]:
```
show_plots(data=data, column=data$NrCitablePublications, variable="NrCitablePublicati
```

```
   Min. 1st Qu.   Median    Mean 3rd Qu.    Max. 
 0.0000  0.6931  1.6094  1.5347  2.3026  5.4596 
```



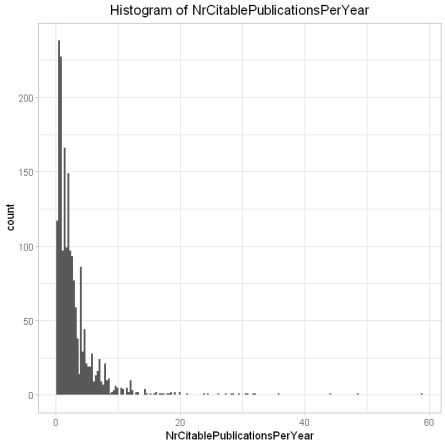



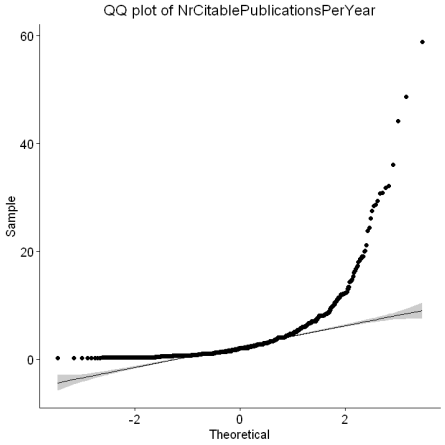

Die Daten zur Gesamtanzahl der Publikationen sind stark schief verteilt. Es scheint, dass viele Projekte letztendlich nur sehr wenige Ergebnisse veröffentlichen. Zwei Datenpunkte liegen über dem oberen Whisker im Boxplot.

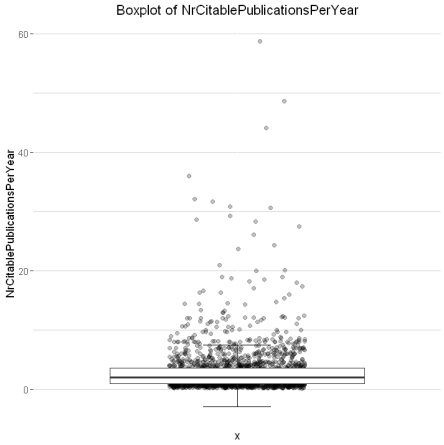
## Anzahl Publikationen pro Jahr (NrCitablePublicationsPerYear)

In [ ]:
```
show_plots(data=data, column=data$NrCitablePublicationsPerYear, variable="NrCitablePu
```

```
   Min. 1st Qu.   Median    Mean 3rd Qu.    Max. 
 0.1714  1.0000  1.9982  2.9914  3.6049 58.7500 
```



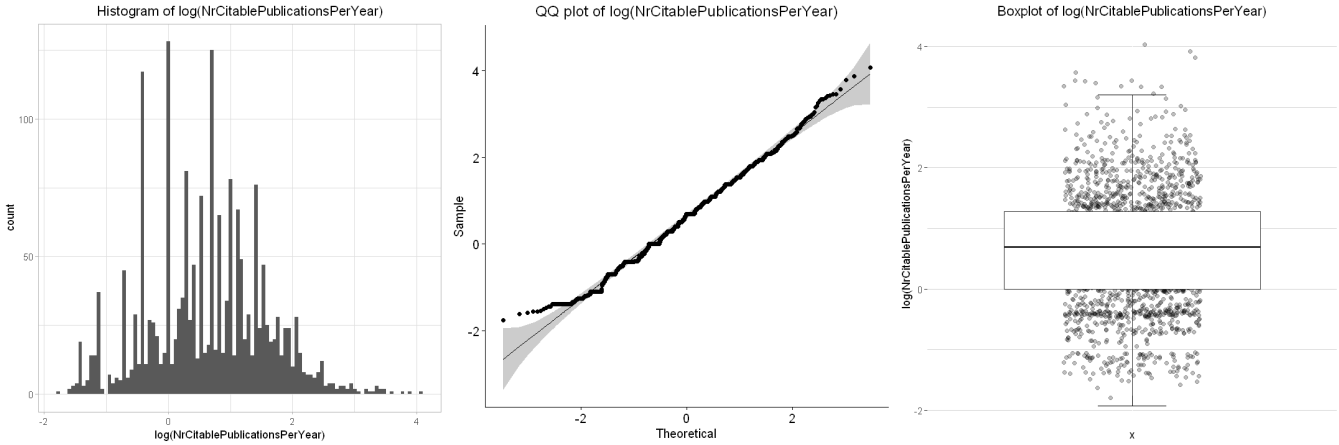




Mit log-Transformation:

In [ ]:
```
show_plots(data=data, column=data$NrCitablePublicationsPerYear, variable="NrCitablePu
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-1.7635  0.0000  0.6922  0.6227  1.2823  4.0733 
```

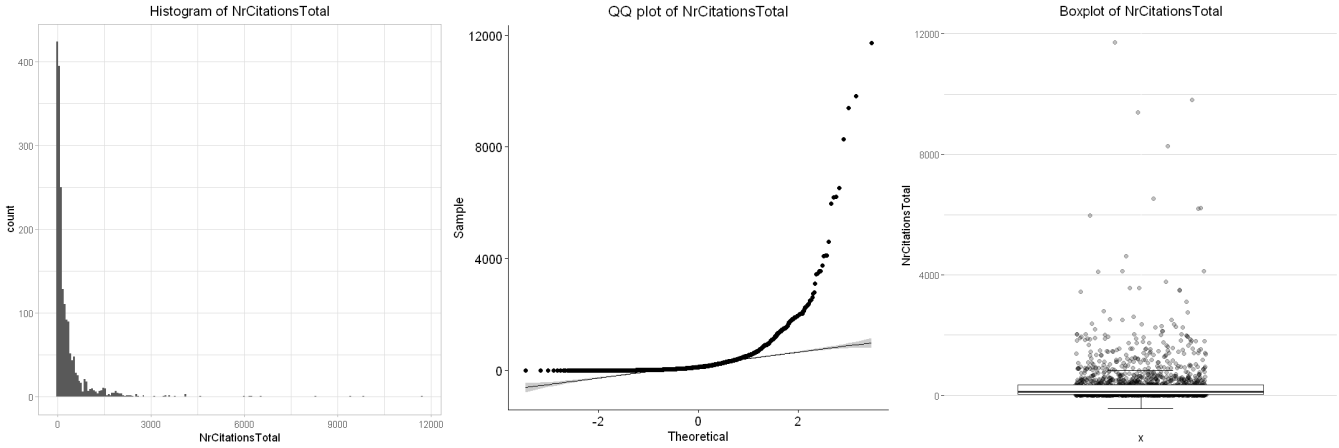



Die log-transformierten Daten zur durchschnittlichen Anzahl von Publikationen pro Jahr weisen eine annähernd normale Verteilung auf. Es gibt jedoch einige Projekte, die einen sehr hohen Wert aufweisen und als Ausreißer betrachtet werden können.

## Totale Anzahl Zitierungen (NrCitationsTotal)

In [ ]:
```
show_plots(data=data, column=data$NrCitationsTotal, variable="NrCitationsTotal", bins
```

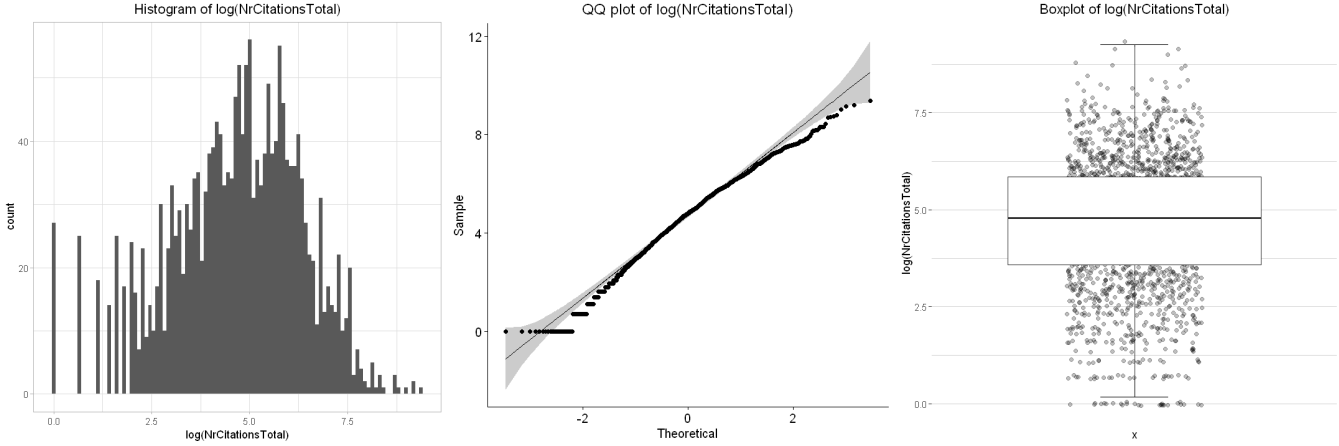
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    1.0    36.0   121.0   336.4   348.0 11713.0 
```




Mit log-Transformation:

In [ ]:
```
# Mit log-Transformation
show_plots(data=data, column=data$NrCitationsTotal, variable="NrCitationsTotal", bins
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   3.584   4.796   4.649   5.852   9.368 
```



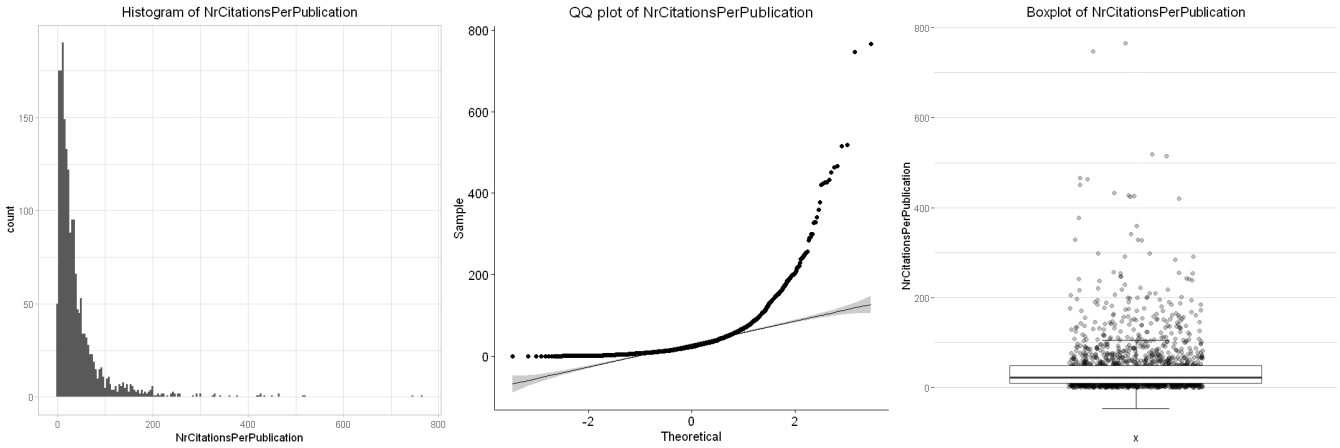

Die log-transformierten Daten zur Gesamtanzahl der Zitierungen eines Projekts weisen eine verbesserte Verteilung auf. Es gibt jedoch vor allem im unteren Bereich übermäßig viele Projekte, bei denen nur

sehr wenige Zitierungen insgesamt vorliegen.

## Anzahl Zitierungen pro Publikation (NrCitationsPerPublication)

In [ ]:
```
show_plots(data=data, column=data$NrCitationsPerPublication, variable="NrCitationsPer
```

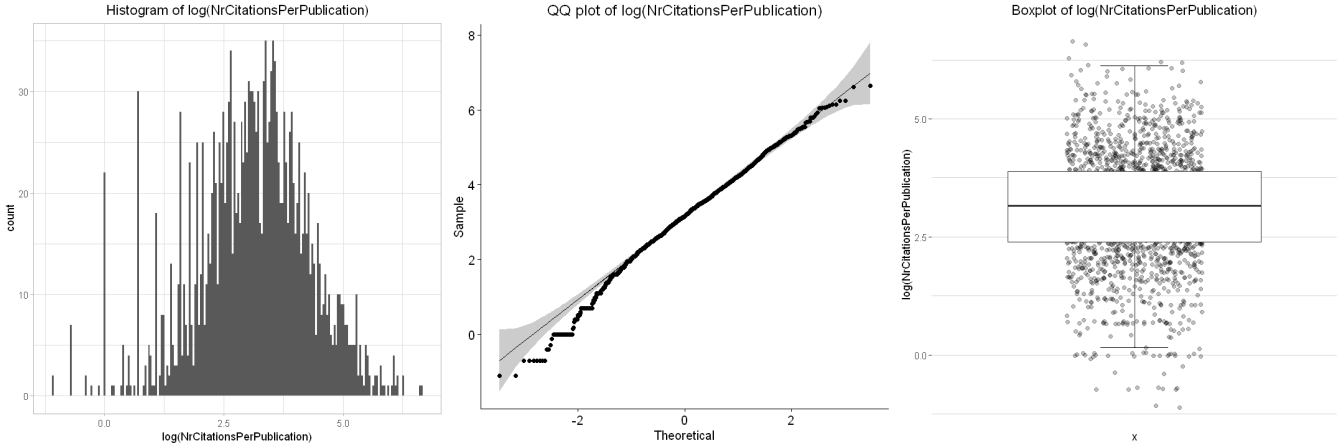
```
   Min.  1st Qu.    Median      Mean  3rd Qu.      Max. 
 0.3333  11.0000  23.5000  42.4710  48.9326 765.3333 
```




Mit log-Transformation:

In [ ]:
```
show_plots(data=data, column=data$NrCitationsPerPublication, variable="NrCitationsPer
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-1.099   2.398   3.157   3.114   3.890   6.640 
```




Auch bei der durchschnittlichen Anzahl von Zitierungen pro Publikation zeigen die log-transformierten Daten eine deutlich verbesserte Verteilung. Es sind jedoch Abweichungen von der Normalverteilung zu beobachten, insbesondere im Bereich der niedrigen Werte.

## Genereller Umgang mit Ausreissern

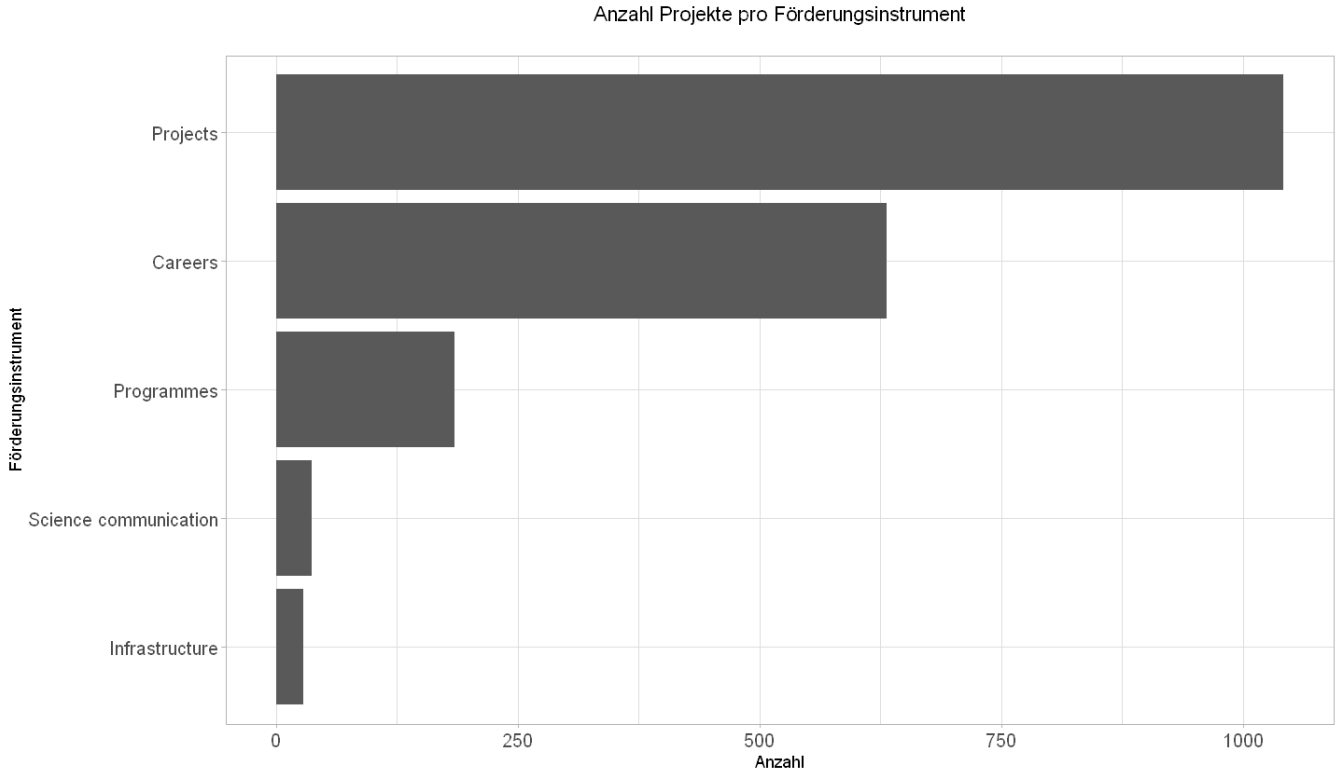
Die Ausreisser gemäss den Boxplots werden als echte und valide Datenpunkte betrachtet. Sie beruhen nicht auf Fehlern in den Daten und sind für die Zielgruppe der Analyse durchaus von Interesse.

## Förderungsinstrument

In [ ]:
```
options(repr.plot.width=12, repr.plot.height=7) # Grösse der Plots festlegen
```

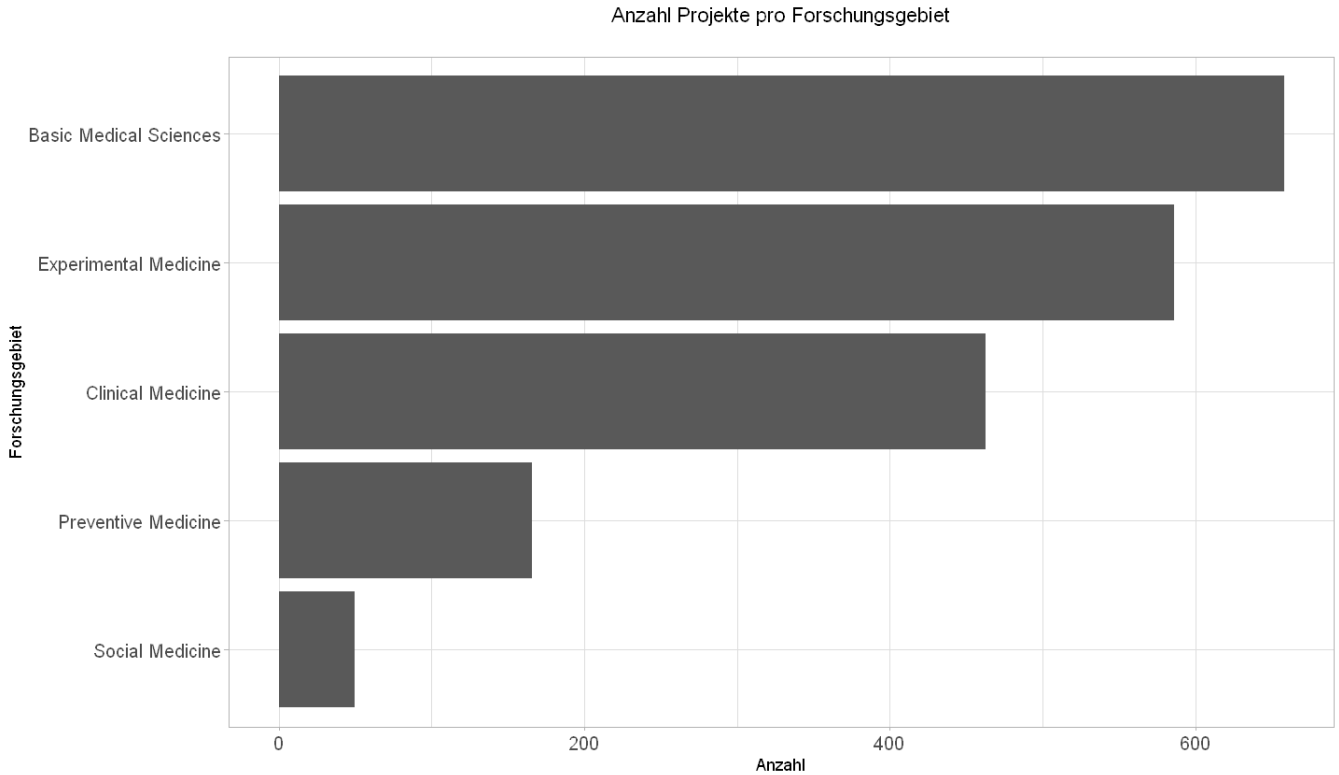
In [ ]:
```
ggplot(data,
  aes(x=reorder(FundingInstrumentLevel1, FundingInstrumentLevel1, function(x) length(
  geom_bar() +
  coord_flip() +
  theme_light() +
  ggtitle("Anzahl Projekte pro Förderungsinstrument\n") +
  theme(plot.title = element_text(hjust = 0.5)) +
```

```
  labs(x="Förderungsinstrument", y="Anzahl") +
  theme(axis.text = element_text(size = 12))
```




## Forschungsgebiete

In [ ]:
```
ggplot(data,
  aes(x=reorder(MainDiscipline_Level2, MainDiscipline_Level2, function(x) length(x)))
  geom_bar() +
  coord_flip() +
  theme_light() +
  ggtitle("Anzahl Projekte pro Forschungsgebiet\n") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x="Forschungsgebiet", y="Anzahl") +
  theme(axis.text = element_text(size = 12))
```




## "Summary" der log-transformierten Daten

Wir werden im weiteren Verlauf für einige Variabeln in der Regel mit den log-transformierten Daten arbeiten. Daher folgt hier erneut eine entsprechende "Summary".

In [ ]:
```
# Numerische Daten und log-Transformation der relevanten Variabeln
data_numeric <- data %>% dplyr::select(
                            GrantDurationYears,
                            AmountGrantedAllSets,
                            AmountPerYear,
                            NrCitablePublications,
                            NrCitablePublicationsPerYear,
                            NrCitationsTotal,
                            NrCitationsPerPublication
                            ) %>%
                        mutate(
                            logAmountGrantedAllSets=log(AmountGrantedAllSets),
                            logAmountPerYear=log(AmountPerYear),
                            logNrCitablePublications=log(NrCitablePublications),
                            logNrCitablePublicationsPerYear=log(NrCitablePublications
                            logNrCitationsTotal=log(NrCitationsTotal),
                            logNrCitationsPerPublication=log(NrCitationsPerPublicatio
                        ) %>%
                        dplyr::select(
                            GrantDurationYears,
                            logAmountGrantedAllSets,
                            logAmountPerYear,
                            logNrCitablePublications,
                            logNrCitablePublicationsPerYear,
                            logNrCitationsTotal,
                            logNrCitationsPerPublication
                        )
```

In [ ]:
```
# Summary der metrischen, log-transformierten Variabeln
summary(data_numeric)
```

```
 GrantDurationYears logAmountGrantedAllSets logAmountPerYear
 Min.   :0.08219    Min.   : 7.313          Min.   : 8.386
 1st Qu.:1.78904    1st Qu.:11.786          1st Qu.:11.192
 Median :3.00000    Median :12.867          Median :11.683
 Mean   :2.83549    Mean   :12.587          Mean   :11.675
 3rd Qu.:3.74795    3rd Qu.:13.302          3rd Qu.:12.073
 Max.   :6.91507    Max.   :16.170          Max.   :14.783
 logNrCitablePublications logNrCitablePublicationsPerYear logNrCitationsTotal
 Min.   :0.0000           Min.   :-1.7635                 Min.   :0.000
 1st Qu.:0.6931           1st Qu.: 0.0000                 1st Qu.:3.584
 Median :1.6094           Median : 0.6922                 Median :4.796
 Mean   :1.5347           Mean   : 0.6227                 Mean   :4.649
 3rd Qu.:2.3026           3rd Qu.: 1.2823                 3rd Qu.:5.852
 Max.   :5.4596           Max.   : 4.0733                 Max.   :9.368
 logNrCitationsPerPublication
 Min.   :-1.099
 1st Qu.: 2.398
 Median : 3.157
 Mean   : 3.114
 3rd Qu.: 3.890
 Max.   : 6.640
```

## Korrelationsmatrix

Auch für die Korrelationsmatrix verwenden wir, wenn relevant, die log-transformierten Daten.

In [ ]:
```
cor(data_numeric)
```

A matrix: 7 × 7 of typ

| | GrantDurationYears | logAmountGrantedAllSets | logAmountPerYear | logNrCitab |
|---|---|---|---|---|
| **GrantDurationYears** | 1.00000000 | 0.78097080 | 0.4114829 | |

| | | | |
|---|---|---|---|
| **logAmountGrantedAllSets** | 0.78097080 | 1.00000000 | 0.8584575 |
| **logAmountPerYear** | 0.41148294 | 0.85845747 | 1.0000000 |
| **logNrCitablePublications** | 0.38714592 | 0.46883065 | 0.4068491 |
| **logNrCitablePublicationsPerYear** | -0.16396099 | -0.01483358 | 0.1692715 |
| **logNrCitationsTotal** | 0.20921489 | 0.34826596 | 0.3469307 |
| **logNrCitationsPerPublication** | -0.03049333 | 0.09952490 | 0.1506438 |

Einige hochkorrelierte Zusammenhänge sind wenig überraschend und werden nicht weiter verfolgt. Dazu gehören die Korrelationen zwischen den Gesamtwerten und den entsprechenden Durchschnittswerten pro Jahr (logAmountGrantedAllSets und logAmountPerYear, logNrCitablePublications und logNrCitablePublicationsPerYear, logNrCitationsTotal und logNrCitationsPerPublication).

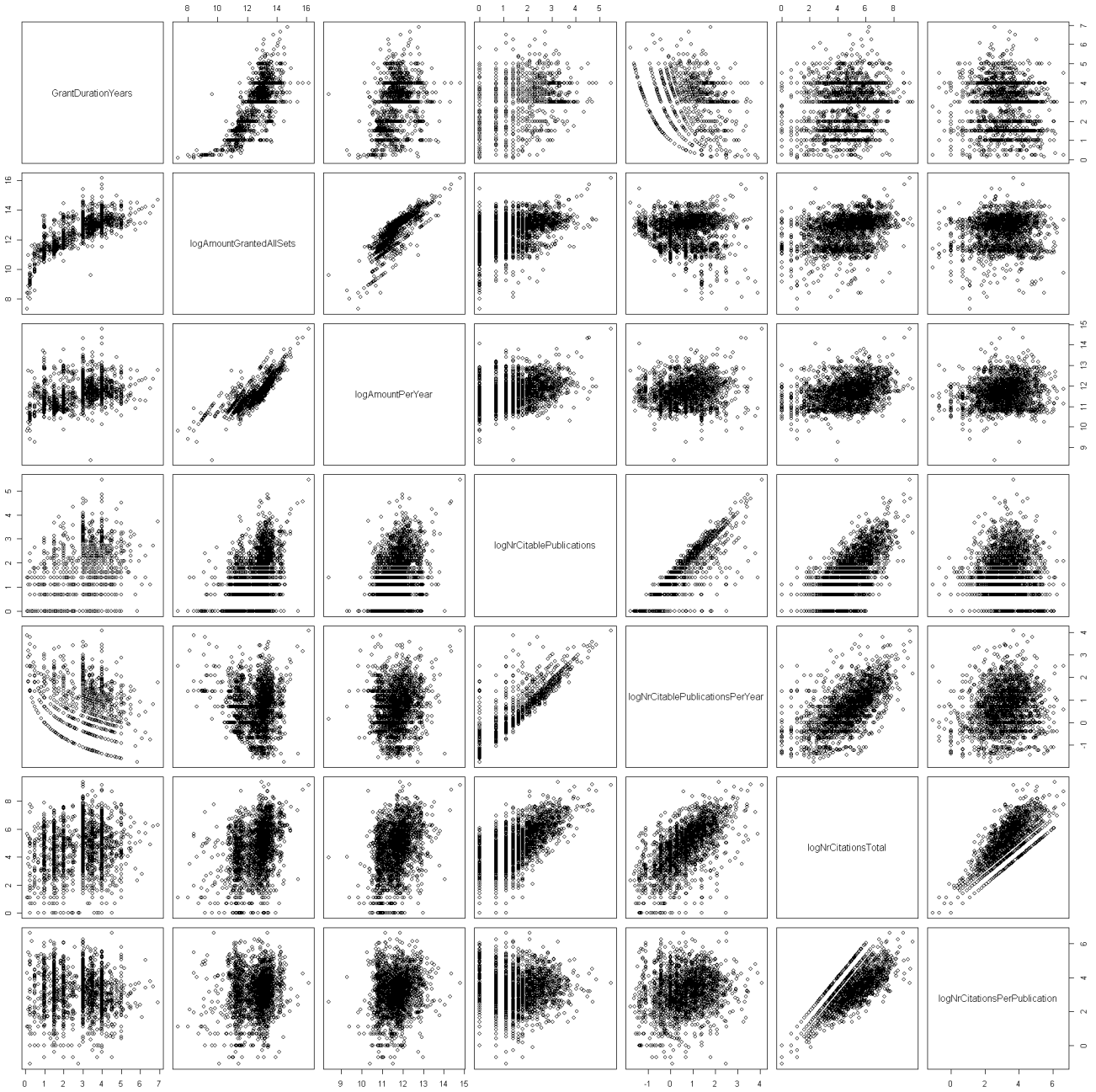
Auch die vergleichsweise hohe Korrelation zwischen der Projektdauer und dem Gesamtförderungsbetrag (GrantDurationYears und logAmountGrantedAllSets) ist wenig überraschend, da längere Projekte erwartungsgemäss mehr Förderung benötigen.

Es gibt keine besonders starke Korrelation zwischen dem Gesamtförderungsbetrag und der Gesamtzahl der Zitierungen (logAmountGrantedAllSets und logNrCitationsTotal). Ebenso besteht keine hohe Korrelation zwischen dem jährlichen Förderungsbetrag und der durchschnittlichen Anzahl der Zitierungen pro Publikation (logAmountPerYear und logNrCitationsPerPublication).

Die Korrelation zwischen der durchschnittlichen Anzahl von Publikationen pro Jahr und der Gesamtzahl der Zitierungen ist nicht völlig überraschend, setzt jedoch voraus, dass die Publikationen tatsächlich zitiert werden. Wir werden diesen Zusammenhang im weiteren Verlauf der Semesterarbeit noch näher betrachten.

## Pairs

In [ ]:
```
options(repr.plot.width=16, repr.plot.height=16)
pairs(data_numeric, cex.labels = 1.2)
```




Es sind keine offensichtlich linearen Zusammenhänge zwischen den Förderungsbeträgen (gesamt und pro Jahr) und der Gesamtanzahl der Zitierungen bzw. der Anzahl der Zitierungen pro Publikation erkennbar.

Die deutlichste Linearität besteht zwischen der Anzahl der Publikationen pro Jahr (logNrCitablePublicationsPerYear) und der Gesamtanzahl der Zitierungen (logNrCitationsTotal). Basierend auf dieser Erkenntnis soll im nächsten Kapitel überprüft werden, ob ein entsprechendes Modell generiert werden kann.

# Multiple lineare Regression

## Theoretische Fundierung

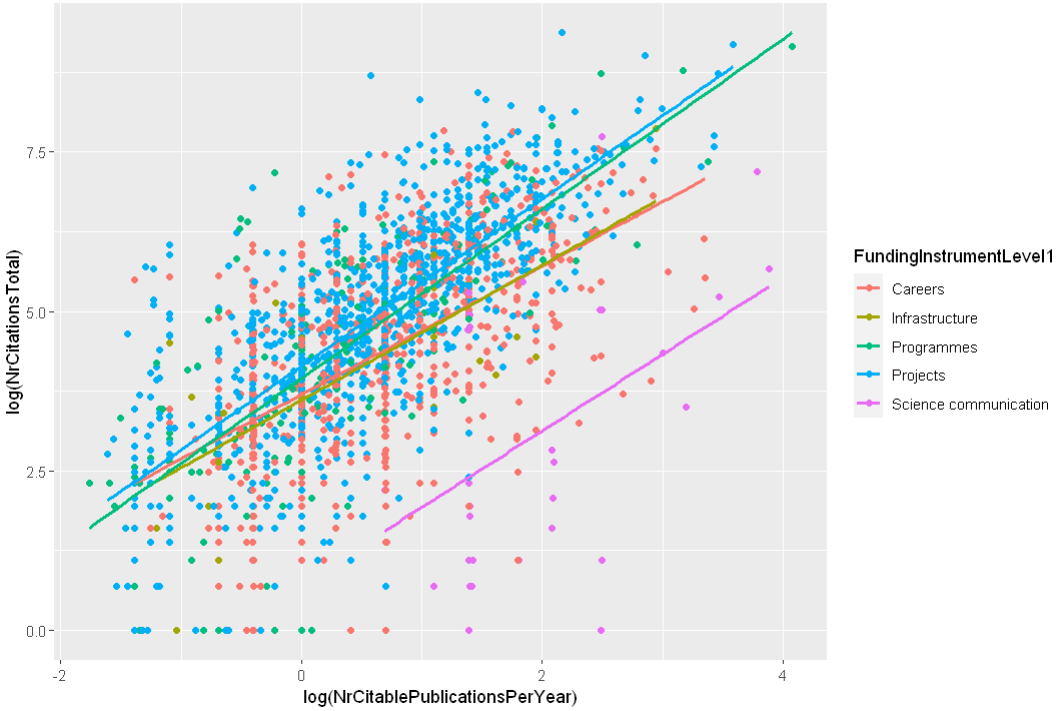
Wie im vorherigen Kapitel erwähnt, werden wir in diesem Kapitel prüfen, ob wir eine multiple lineare Regression für die Gesamtanzahl der Zitierungen erstellen können.

Eine Voraussetzung für die lineare Regression besteht darin, dass der Zusammenhang zwischen der abhängigen und der unabhängigen Variable linear ist. Daher werden wir erneut das Streudiagramm der

Publikationen pro Jahr und der Gesamtanzahl der Zitierungen betrachten und die Punkte entsprechend des Förderungsinstruments einfärben.

In [ ]:
```
options(repr.plot.width=9, repr.plot.height=6) # Grösse des Plots festlegen
# Streudiagramm
ggplot(data, aes(x=log(NrCitablePublicationsPerYear), y=log(NrCitationsTotal), color=
  geom_point() +
  geom_smooth(method=lm, se=FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```




Die grünen Punkte der "Programmes" zeigen eine geringere Streuung um die Regressionsgerade. Für die multiple lineare Regression beschränken wir uns daher auf die 185 Projekte, die dem Förderungsinstrument "Programmes" zugeordnet sind. Diese Projekte sind inhaltlich besser vergleichbar, da bei der Förderung bestimmte thematische oder konzeptionell-organisatorische Rahmenbedingungen vorgegeben sind. Wir erstellen erneut einen Pair-Plot nur für die "Programmes" und sehen, dass die zuvor beobachtete Linearität zwischen der Gesamtanzahl der Zitierungen (logNrCitationsTotal) und der durchschnittlichen Anzahl der Publikationen pro Jahr (logNrCitablePublicationsPerYear) deutlich ausgeprägter ist.

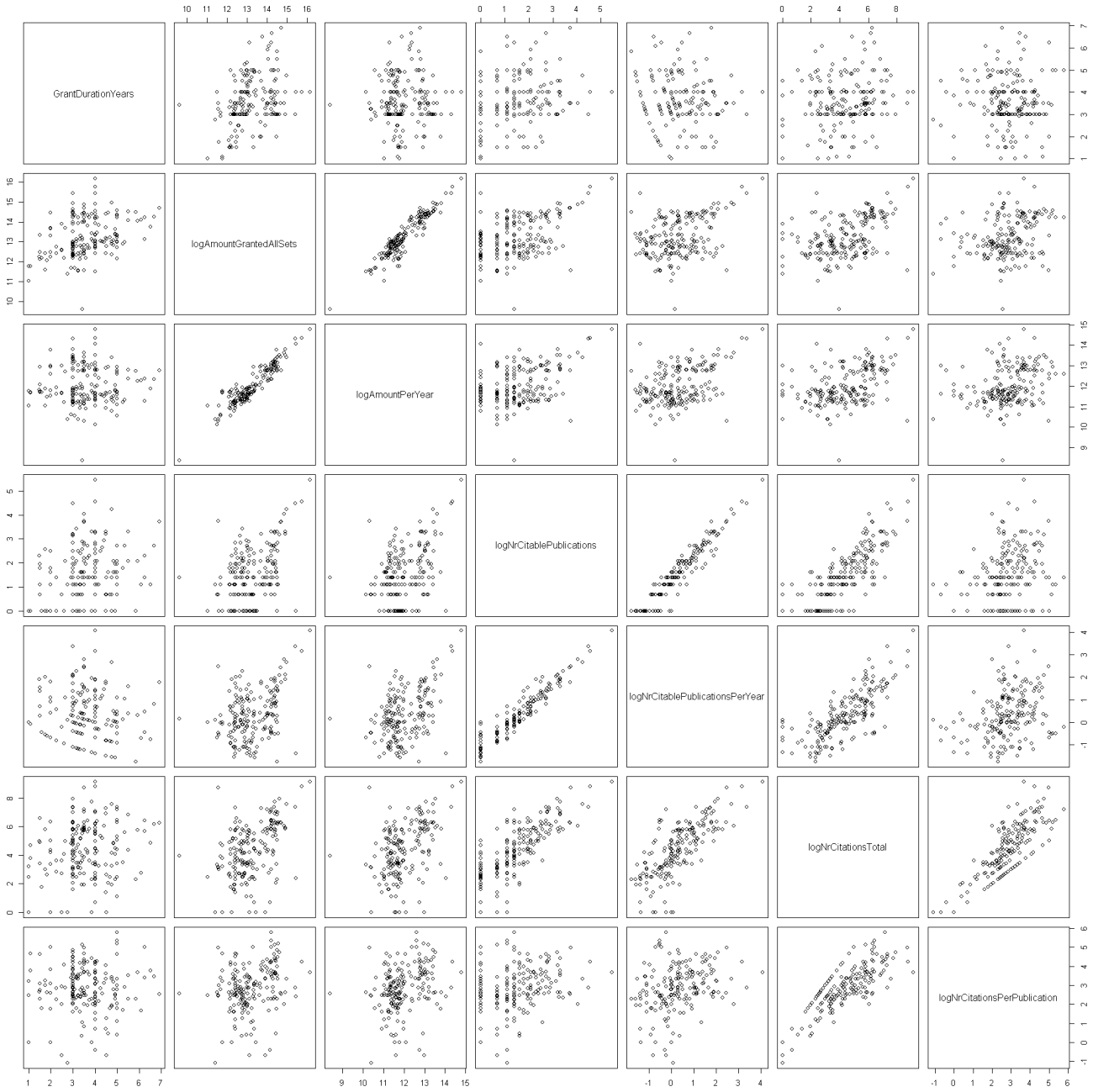
In [ ]:
```
# Nach Förderungsinstrument "Programmes" filtern
programmes <- filter(data, FundingInstrumentLevel1 == "Programmes")
dim(programmes)
```

185 · 9

In [ ]:
```
# Numerische Daten und log-Transformation der relevanten Variabeln
data_numeric.2 <- programmes %>% dplyr::select(
                            GrantDurationYears,
                            AmountGrantedAllSets,
                            AmountPerYear,
                            NrCitablePublications,
                            NrCitablePublicationsPerYear,
                            NrCitationsTotal,
                            NrCitationsPerPublication
                            ) %>%
                        mutate(
                            logAmountGrantedAllSets=log(AmountGrantedAllSets),
                            logAmountPerYear=log(AmountPerYear),
                            logNrCitablePublications=log(NrCitablePublications),
                            logNrCitablePublicationsPerYear=log(NrCitablePublications
```

```
        logNrCitationsTotal=log(NrCitationsTotal),
        logNrCitationsPerPublication=log(NrCitationsPerPublicatio
    ) %>%
    dplyr::select(
        GrantDurationYears,
        logAmountGrantedAllSets,
        logAmountPerYear,
        logNrCitablePublications,
        logNrCitablePublicationsPerYear,
        logNrCitationsTotal,
        logNrCitationsPerPublication
    )
```

In [ ]:
```
options(repr.plot.width=16, repr.plot.height=16)
pairs(data_numeric.2, cex.labels = 1.2)
```

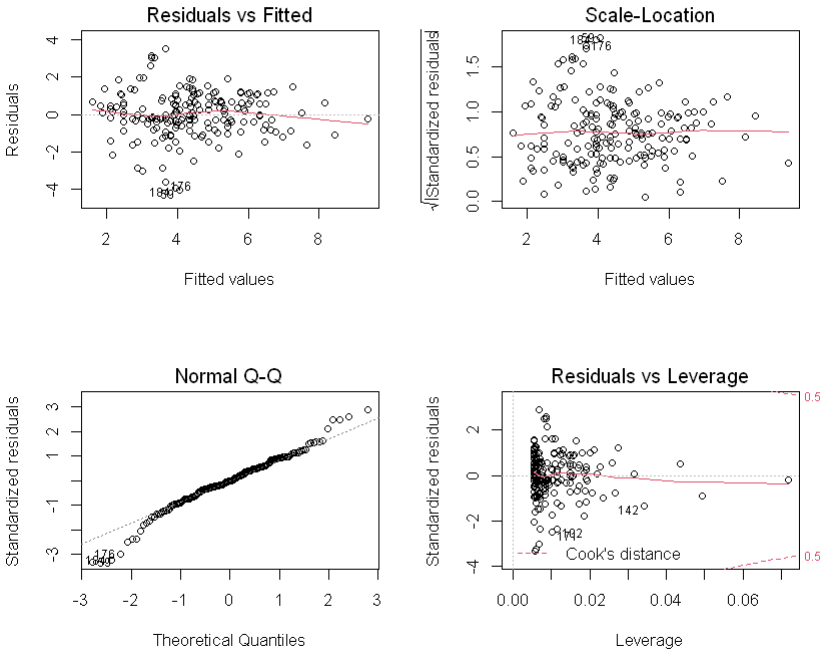



Wir beginnen mit der einfachen linearen Regression.

In [ ]:
```
lm.1 <- lm(log(NrCitationsTotal) ~ log(NrCitablePublicationsPerYear), data=programmes
summary(lm.1)

options(repr.plot.width=7, repr.plot.height=6) # Grösse der Plots festlegen

layout(matrix(c(1,2,3,4),2,2))
plot(lm.1)
```

```r
# Studentized Breusch-Pagan test
bptest(lm.1)
```

```
Call:
lm(formula = log(NrCitationsTotal) ~ log(NrCitablePublicationsPerYear),
    data = programmes)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0663 -0.6945 -0.0172  0.7120  3.5283

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.95226    0.09390   42.09   <2e-16 ***
log(NrCitablePublicationsPerYear)  1.32960    0.08402   15.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.224 on 183 degrees of freedom
Multiple R-squared:  0.5778,    Adjusted R-squared:  0.5755
F-statistic: 250.5 on 1 and 183 DF,  p-value: < 2.2e-16
        studentized Breusch-Pagan test

data:  lm.1
BP = 3.5183, df = 1, p-value = 0.06069
```




- Der Intercept und der Koeffizient des Modells sind signifikant (α = 0.05).
- Das einfache Modell erklärt etwa 58% der Variabilität der Daten.
- Der Plot "Residuals vs. Fitted" zeigt keine eindeutige Heteroskedastizität. Ein formaler Breusch-Pagan-Test bestätigt dies. Die Nullhypothese des Tests besagt, dass Homoskedastizität vorliegt. Mit einem p-Wert von 0.06 ist der Test knapp nicht signifikant genug, um die Nullhypothese abzulehnen.
- Der QQ-Plot zeigt jedoch, dass die Residuen nicht ideal normalverteilt sind. Sie weisen längere Tails auf beiden Seiten auf als in einer normalverteilten Verteilung üblich.

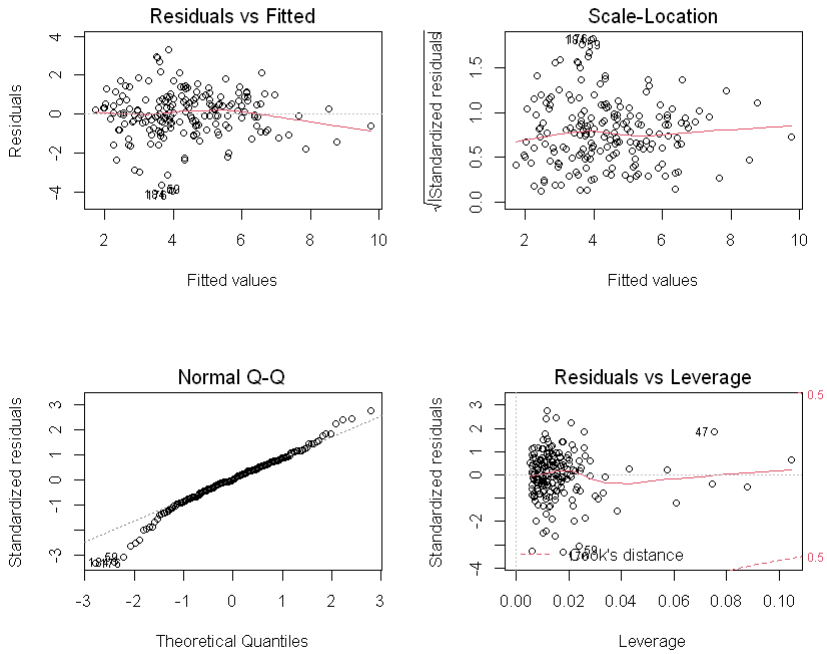
Im Folgenden werden wir versuchen, ein verbessertes Modell für die Gesamtanzahl der Zitierungen zu finden, das auf mehreren unabhängigen Variablen basiert. Bei der Bewertung des Modells werden wir die gleichen Kriterien wie bei der einfachen linearen Regression verwenden. Wir gehen davon aus, dass die Residuen möglichst normalverteilt sind und keine Heteroskedastizität vorliegt. Ausserdem sollten die unabhängigen Variablen nicht stark miteinander korrelieren (keine Multikollinearität).

## Auswertung

Da die Zielvariable die Gesamtanzahl der Zitierungen ist, könnte es naheliegend sein, auch die Projektdauer in das Modell einzuschließen. Der Pair-Plot zeigt jedoch keine deutliche Linearität zwischen der Projektdauer und der Anzahl der Zitierungen. Allerdings ist eine geringfügige Linearität zwischen dem Gesamtförderungsbetrag (gesamt und pro Jahr) und der Gesamtanzahl der Zitierungen erkennbar. Da wir bereits den durchschnittlichen Publikationswert pro Jahr im Modell haben, werden wir als nächstes den durchschnittlichen Förderungsbetrag pro Jahr in das Modell aufnehmen.

In [ ]:
```r
lm.2 <- lm(log(NrCitationsTotal) ~ log(NrCitablePublicationsPerYear) + log(AmountPerY
summary(lm.2)

options(repr.plot.width=7, repr.plot.height=6) # Grösse der Plots festlegen

layout(matrix(c(1,2,3,4),2,2))
plot(lm.2)

# Studentized Breusch-Pagan test
bptest(lm.2)
```

```
Call:
lm(formula = log(NrCitationsTotal) ~ log(NrCitablePublicationsPerYear) +
    log(AmountPerYear), data = programmes)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9758 -0.6175  0.0227  0.7208  3.3127

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.76735    1.27260   0.603    0.547
log(NrCitablePublicationsPerYear)  1.23308    0.09132  13.503   <2e-16 ***
log(AmountPerYear)                0.26793    0.10677   2.509    0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.207 on 182 degrees of freedom
Multiple R-squared:  0.5919,    Adjusted R-squared:  0.5874
F-statistic:   132 on 2 and 182 DF,  p-value: < 2.2e-16
        studentized Breusch-Pagan test

data:  lm.2
BP = 5.1023, df = 2, p-value = 0.07799
```



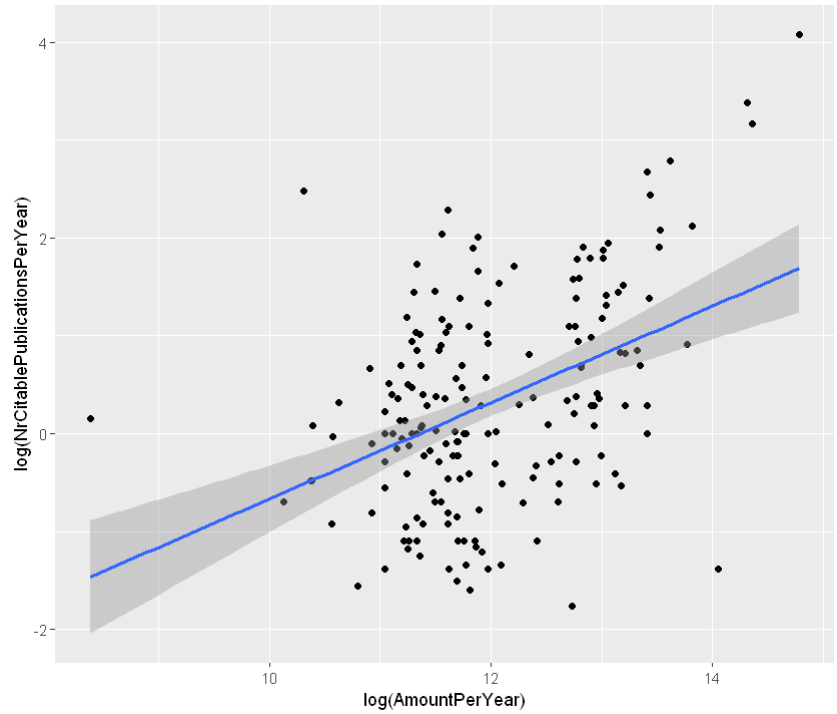

Wir betrachten noch den Zusammenhang zwischen den beiden unabhängigen Variabeln.

In [ ]:
```
cor(log(programmes$NrCitablePublicationsPerYear), log(programmes$AmountPerYear))
```

0.421194088360786

In [ ]:
```
ggplot(programmes, aes(x=log(AmountPerYear), y=log(NrCitablePublicationsPerYear))) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



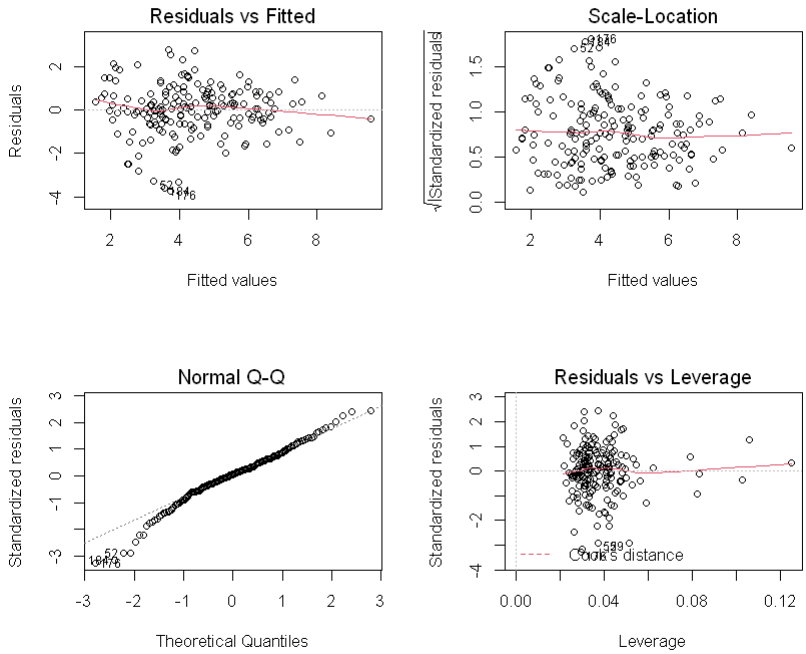

- Beide Koeffizienten sind signifikant (α = 0.05).
- Der Intercept ist nicht mehr signifikant. Dies ist akzeptabel, da wir davon ausgehen müssen, dass ein Projekt ohne Publikationen und ohne Förderungsbetrag keine signifikante Gesamtanzahl an Zitierungen generieren kann.
- Das Gütemass "Adjusted R-squared" hat sich minimal verbessert. Das Modell erklärt etwa 59% der Variabilität in den Daten.

- Es ist keine ausgeprägte Heteroskedastizität erkennbar, was durch den formalen Breusch-Pagan-Test bestätigt wird (der p-Wert ist zu hoch, um die Nullhypothese zu verwerfen).
- Der QQ-Plot zeigt eine leichte Verbesserung in der Verteilung der Residuen, aber die Tails sind immer noch etwas länger als normal, insbesondere bei niedrigen Werten.
- Obwohl eine gewisse Korrelation zwischen log(NrCitablePublicationsPerYear) und log(AmountPerYear) besteht, betrachten wir sie nicht als extrem genug, um von Multikollinearität zu sprechen.

Im nächsten Schritt schliessen wir noch die kategoriale Variable des Forschungsgebiets mit in das Modell ein. Dabei legen wir "Social Medicine" als Referenz fest.

In [ ]:
```
# "Social Medicine" als Referenz-Forschungsgebiet festlegen
programmes$MainDiscipline_Level2 <- relevel(programmes$MainDiscipline_Level2, ref = "

lm.3 <- lm(log(NrCitationsTotal) ~ log(NrCitablePublicationsPerYear) + log(AmountPerY
summary(lm.3)

layout(matrix(c(1,2,3,4),2,2))
plot(lm.3)

# Studentized Breusch-Pagan test
bptest(lm.3)
```

```
Call:
lm(formula = log(NrCitationsTotal) ~ log(NrCitablePublicationsPerYear) +
    log(AmountPerYear) + MainDiscipline_Level2, data = programmes)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7395 -0.6049  0.0683  0.7099  2.7716

Coefficients:
                                          Estimate Std. Error t value
(Intercept)                                0.86838    1.24064   0.700
log(NrCitablePublicationsPerYear)          1.26149    0.08882  14.203
log(AmountPerYear)                         0.22029    0.10607   2.077
MainDiscipline_Level2Basic Medical Sciences  0.67381    0.29817   2.260
MainDiscipline_Level2Clinical Medicine       0.28798    0.29491   0.977
MainDiscipline_Level2Experimental Medicine   1.07148    0.30190   3.549
MainDiscipline_Level2Preventive Medicine     0.10872    0.30467   0.357
                                          Pr(>|t|)
(Intercept)                               0.484875
log(NrCitablePublicationsPerYear)          < 2e-16 ***
log(AmountPerYear)                        0.039252 *
MainDiscipline_Level2Basic Medical Sciences 0.025046 *
MainDiscipline_Level2Clinical Medicine      0.330128
MainDiscipline_Level2Experimental Medicine  0.000495 ***
MainDiscipline_Level2Preventive Medicine    0.721641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.156 on 178 degrees of freedom
Multiple R-squared:  0.6336,    Adjusted R-squared:  0.6212
F-statistic: 51.29 on 6 and 178 DF,  p-value: < 2.2e-16
        studentized Breusch-Pagan test

data:  lm.3
BP = 10.007, df = 6, p-value = 0.1243
```




- Die Koeffizienten im zweiten Modell sind weiterhin signifikant (α = 0.05).
- Der Intercept bleibt nicht signifikant, was wir weiterhin akzeptieren.
- Die Forschungsgebiete "Experimental Medicine" und "Basic Medical Sciences" haben beide einen signifikanten Einfluss auf die Gesamtanzahl der Zitierungen.
- Das Gütemass "Adjusted R-squared" hat sich weiter verbessert. Das Modell erklärt nun etwa 62% der Variabilität in den Daten.
- Es ist kein eindeutiges Muster im Residuals-Plot (Residuals vs. Fitted) erkennbar. Wir können weiterhin von Homoskedastizität ausgehen.
- Die Verteilung der Residuen hat sich weiter in Richtung einer Normalverteilung verbessert, insbesondere für höhere Werte. Bei niedrigen Werten folgen die Residuen jedoch weiterhin nicht einer idealen Normalverteilung. Wir müssen akzeptieren, dass das Modell für niedrige Werte nicht besonders zuverlässig ist.

## Interpretation

Aufgrund des dritten Modells können wir festhalten, dass zur Vorhersage der Gesamtanzahl der Zitierungen, welche die Publikationen eines "Programmes"-Projekts erhalten, die folgende Basisformel verwendet werden kann. Die Basisformel gilt sowohl für Projekte aus dem Forschungsgebiet "Social Medicine" als auch für Projekte aus den Forschungsgebieten "Clinical Medicine" und "Preventive Medicine".

Basisformel:

```
log(NrCitationsTotal) = 0.86838
  + (1.26149 * log(NrCitablePublicationsPerYear))
  + (0.22029 * log(AmountPerYear))
```

- Das bedeutet, dass eine Erhöhung der Anzahl der Publikationen pro Jahr (NrCitablePublicationsPerYear) um 1% zu einer Steigerung der Gesamtanzahl der Zitierungen (NrCitationsTotal) um 1.26 % führen würde ( `(1.01^1.26149 - 1) * 100` ). Eine Erhöhung der jährlichen Publikationen um 10% würde zu einer Steigerung von etwa 12.78 % der Zitierungen führen ( `(1.01^0.22029 - 1) * 100` ).
- Eine Erhöhung des jährlichen Förderungsbetrags um 1% lässt auf eine Steigerung der Gesamtanzahl der Zitierungen um 0.22 % schließen ( `(1.01^0.22029 - 1) * 100` ).

Unter Einbeziehung der Forschungsgebiete "Experimental Medicine" und "Basic Medical Sciences" kann die Basisformel wie folgt erweitert werden:

```
log(NrCitationsTotal) = 0.86838
  + (1.26149 * log(NrCitablePublicationsPerYear))
  + (0.22029 * log(AmountPerYear))
  + (0.67381 * MainDiscipline_Level2Basic Medical Sciences)
  + (1.07148 * MainDiscipline_Level2Experimental Medicine)
```

- Bei ansonsten gleichbleibenden Werten bezüglich der Anzahl jährlicher Publikationen und des Förderungsbetrags erhöht sich die Gesamtanzahl der Zitierungen bei einem Projekt aus dem Forschungsgebiet "Experimental Medicine" um den Faktor 2.92 ( `exp(1.07148)` ), im Vergleich zu Projekten aus den Gebieten "Social Medicine", "Clinical Medicine" oder "Preventive Medicine".
- Bei einem Projekt aus dem Gebiet "Basic Medical Sciences" erhöht sich die Gesamtanzahl der Zitierungen immerhin noch um den Faktor 1.96 ( `exp(0.67381)` ), im Vergleich zu Projekten der Gebiete "Social Medicine", "Clinical Medicine" oder "Preventive Medicine".

Das Modell legt nahe, dass das Forschungsgebiet einen deutlichen Einfluss auf die Zitierungen hat, der weitaus größer ist als der finanzielle Aspekt des Förderungsbetrags. Im nächsten Kapitel werden wir daher genauer auf die durchschnittliche Anzahl der Zitierungen pro Publikation und die Unterschiede zwischen den einzelnen Forschungsgebieten eingehen.

# Varianzanalyse

## Theoretische Fundierung

Wir prüfen, ob das Forschungsgebiet einen signifikanten Einfluss ($\alpha = 0.05$) auf die durchschnittliche Anzahl der Zitierungen pro Publikation hat. Dafür betrachten wir wieder alle Projekte und führen eine Varianzanalyse (ANOVA) durch. Die ANOVA setzt folgende Voraussetzungen voraus, um eine statistisch aussagekräftige Aussage treffen zu können:

- Die abhängige Variable ist intervallskaliert/metrisch.
- Es besteht Varianzhomogenität, das heisst, die Varianz ist innerhalb der verschiedenen Gruppen etwa gleich.
- Die Residuen sind normalverteilt.

Die abhängige Variable, die durchschnittliche Anzahl der Zitierungen pro Publikation, erfüllt die metrische Anforderung. Die anderen beiden Bedingungen werden während der anschließenden Auswertung getestet.

## Auswertung

Zunächst betrachten wir erneut die statistischen Werte für jedes Forschungsgebiet.

In [ ]:
```
describeBy(log(data$NrCitationsPerPublication), data$MainDiscipline_Level2)
```

```
 Descriptive statistics by group
group: Basic Medical Sciences
   vars   n mean   sd median trimmed  mad   min  max range  skew kurtosis   se
X1    1 658 3.19 1.14   3.22    3.23 1.05 -0.69 6.24  6.94 -0.35     0.26 0.04
------------------------------------------------------------
group: Clinical Medicine
   vars   n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 463 2.74 1.14   2.77    2.79 1.05 -1.1 6.25  7.35 -0.36      0.7 0.05
------------------------------------------------------------
group: Experimental Medicine
```

```
   vars   n mean   sd median trimmed  mad   min  max range  skew kurtosis   se
X1    1 586  3.5 1.07   3.53    3.52 1.03 -0.29 6.62   6.9 -0.16     0.19 0.04
------------------------------------------------------------
group: Preventive Medicine
   vars   n mean   sd median trimmed  mad   min  max range  skew kurtosis  se
X1    1 166 2.68 1.26   2.81    2.71 1.09 -0.69 6.64  7.33 -0.19     0.33 0.1
------------------------------------------------------------
group: Social Medicine
   vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 50 2.47 1.25   2.85    2.53 0.98 -1.1 5.11  6.21 -0.55      0.1 0.18
```

Im Durchschnitt erhalten Projekte aus dem Forschungsgebiet "Experimental Medicine" die höchste Anzahl an Zitierungen pro Publikation, gefolgt von "Basic Medical Sciences". Das verhält sich also ähnlich wie die Gesamtzahl an Zitierungen für die Projekte des "Programmes"-Förderungsinstruments. Projekte aus den anderen Forschungsgebieten erhalten im Vergleich dazu relativ weniger Zitierungen pro Publikation.

Die Standardabweichungen scheinen nicht deutlich unterschiedlich zwischen den Gruppen zu sein. Dennoch werden wir die Homogenität der Varianzen mit dem formalen Levene-Test überprüfen. Die Nullhypothese des Tests besagt, dass eine Varianzhomogenität besteht.

In [ ]:
```
leveneTest(log(NrCitationsPerPublication) ~ MainDiscipline_Level2, data = data)
```

A anova: 2 × 3

| | Df | F value | Pr(>F) |
|---|---|---|---|
| | <int> | <dbl> | <dbl> |
| **group** | 4 | 1.488188 | 0.2031694 |
| | 1918 | NA | NA |

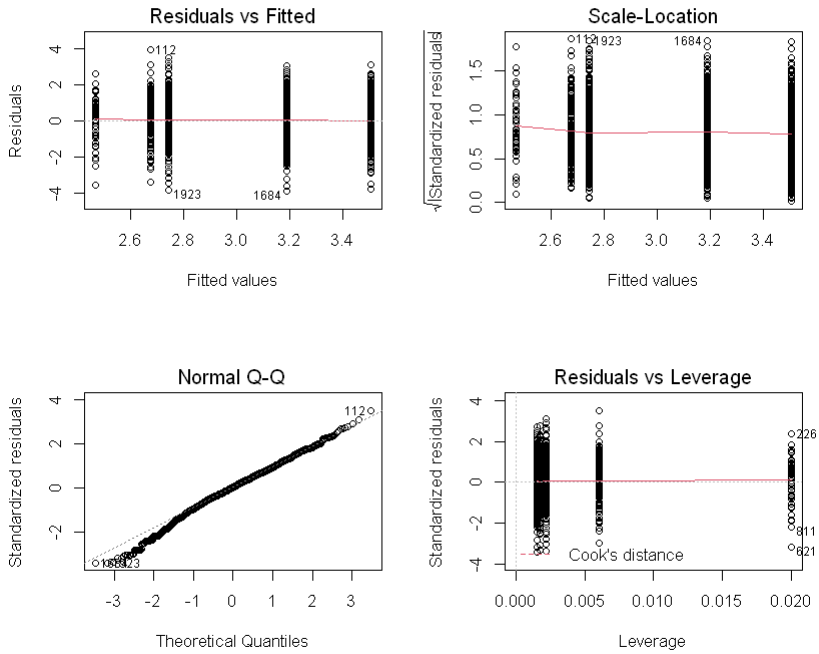
Der p-Wert des Tests ist nicht signifikant bei einem Signifikanzniveau von 5% (α = 0.05), daher können wir die Nullhypothese nicht ablehnen und von einer Homogenität der Varianzen ausgehen.

Für die eigentliche ANOVA legen wir zunächst "Social Medicine" als das Referenz-Forschungsgebiet im gesamten Datensatz fest und führen dann die ANOVA durch.

In [ ]:
```
data$MainDiscipline_Level2 <- relevel(data$MainDiscipline_Level2, ref = "Social Medic
```

In [ ]:
```
anova <- aov(log(NrCitationsPerPublication) ~ MainDiscipline_Level2, data = data)
summary(anova)

layout(matrix(c(1,2,3,4),2,2))
plot(anova)
```

```
                      Df Sum Sq Mean Sq F value Pr(>F)
MainDiscipline_Level2    4  208.7   52.18   40.64 <2e-16 ***
Residuals             1918 2462.3    1.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




- Die ANOVA bestätigt, dass das Forschungsgebiet über den gesamten Datensatz hinweg einen signifikanten Einfluss (α = 0.05) auf die durchschnittliche Anzahl der Zitierungen pro Publikation hat.
- Der QQ-Plot zeigt, dass die Residuen nicht wesentlich von einer Normalverteilung abweichen. Daher können wir die Aussage der ANOVA als statistisch valide anerkennen.

Im Rahmen der multiplen linearen Regression hatte nicht jedes Forschungsgebiet einen signifikanten Einfluss auf die Gesamtanzahl der Zitierungen. Um herauszufinden, welche Forschungsgebiete sich hinsichtlich der durchschnittlichen Anzahl der Zitierungen pro Publikation signifikant unterscheiden, führen wir einen Pairwise t-Test durch. Die Nullhypothese dieses Tests besagt, dass kein signifikanter Unterschied zwischen zwei Gruppen besteht. Bei einem Signifikanzniveau von 5% bedeutet das, dass wir die Nullhypothese verwerfen können, wenn wir p-Werte kleiner als 0,05 erhalten. In diesem Fall können wir davon ausgehen, dass der Unterschied zwischen den beiden Gruppen nicht zufällig ist.

In [ ]:
```
options(width = 120)
pairwise.t.test(log(data$NrCitationsPerPublication), data$MainDiscipline_Level2, p.ad
```

```
        Pairwise comparisons using t tests with pooled SD

data:  log(data$NrCitationsPerPublication) and data$MainDiscipline_Level2

                     Social Medicine Basic Medical Sciences Clinical Medicine Exper
imental Medicine
Basic Medical Sciences 0.00016         -                    -                 -
Clinical Medicine      1.00000         1.2e-09              -                 -
Experimental Medicine  7.4e-09         1.0e-05              < 2e-16           -
Preventive Medicine    1.00000         2.2e-06              1.00000           1.9e-
15

P value adjustment method: bonferroni
```

## Interpretation

Der Pairwise t-Test zeigt, dass die durchschnittliche Anzahl der Zitierungen pro Publikation bei Publikationen aus Projekten der Forschungsgebiete "Basic Medical Sciences" und "Experimental Medicine" signifikant höher ist als bei Publikationen aus Projekten anderer Forschungsgebiete. Darüber hinaus ist auch der Unterschied zwischen "Basic Medical Sciences" und "Experimental Medicine" signifikant. Das bedeutet, dass die Publikationen der Projekte im Forschungsgebiet "Experimental Medicine" im Allgemeinen die höchste durchschnittliche Anzahl an Zitierungen erhalten.

# Logistische Regression

## Theoretische Fundierung

Nachdem wir festgestellt haben, dass Publikationen von Projekten mit dem Schwerpunkt "Experimental Medicine" im Durchschnitt die höchste Anzahl an Zitierungen pro Publikation erhalten, möchten wir in diesem Kapitel prüfen, ob wir mithilfe einer logistischen Regression ein Modell entwickeln können, das vorhersagt, ob die Publikationen eines "Experimental Medicine"-Projekts insgesamt überdurchschnittlich viele Zitierungen erhalten.

Bei der logistischen Regression sollten folgende Bedingungen erfüllt sein:

- Die abhängige Variable ist binär.
- Es liegen keine extremen Ausreißer für die unabhängigen Variablen vor.
- Es besteht keine Multikollinearität zwischen den unabhängigen Variablen.
- Es besteht ein linearer Zusammenhang zwischen dem Logit der abhängigen Variable und den jeweiligen unabhängigen Variablen.

Für die binäre abhängige Variable werden wir den Datensatz um eine Variable ergänzen, die angibt, ob die Gesamtzahl der Zitierungen über dem Durchschnitt liegt oder nicht. Als unabhängige Variablen verwenden wir die durchschnittliche Anzahl der Publikationen pro Jahr und den jährlichen Förderungsbetrag. Die anderen Bedingungen werden wir im Rahmen der folgenden Auswertung überprüfen.

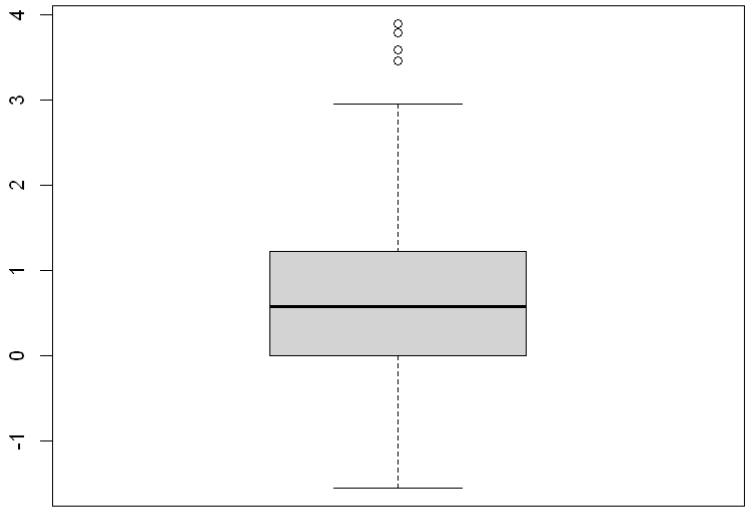
## Auswertung

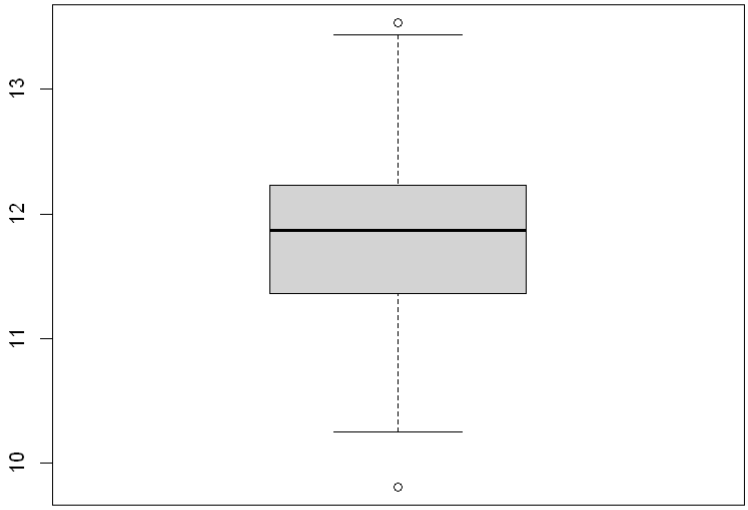
In [ ]:
```
# Nach Projekten mit Fokus "Experimental Medicine" filtern
exp_medicine <- data %>% filter(MainDiscipline_Level2 == "Experimental Medicine")
dim(exp_medicine)
```

586 · 9

Wir überprüfen das Vorhandensein von Ausreißern für die beiden unabhängigen Variablen gemäß der "1.5 * IQR"-Regel.

In [ ]:
```
boxplot(log(exp_medicine$NrCitablePublicationsPerYear))
boxplot(log(exp_medicine$AmountPerYear))
```




Für die logistische Regression werden gemäss der angewandten Regel betfeffen Ausreisser insgesamt sechs Beobachtungen entfernt.

In [ ]:
```
# Ausreisser entfernen
exp_medicine <- exp_medicine %>%
  filter(log(NrCitablePublicationsPerYear) >= quantile(log(NrCitablePublicationsPerYe
  filter(log(NrCitablePublicationsPerYear) <= quantile(log(NrCitablePublicationsPerYe
  filter(log(AmountPerYear) >= quantile(log(AmountPerYear), 0.25) - 1.5 * IQR(log(Amo
  filter(log(AmountPerYear) <= quantile(log(AmountPerYear), 0.75) + 1.5 * IQR(log(Amo
```

Um die binäre Zielvariable abzuleiten, berechnen wir den Mittelwert für die Gesamtanzahl der Zitierungen. Anschliessend fügen wir dem Daten-Subset die binäre Variable "HighlyCited" hinzu. Der Wert dieser Variable ist 1, wenn die beobachtete Gesamtanzahl der Zitierungen höher als der zuvor berechnete Mittelwert ist, andernfalls ist der Wert als 0 codiert.

In [ ]:
```
mean <- mean(log(exp_medicine$NrCitationsTotal))
mean
```

5.00406434814307

In [ ]:
```
exp_medicine$HighlyCited <- ifelse(log(exp_medicine$NrCitationsTotal) > mean, 1, 0)
```

Wir erstellen das Modell für die logistische Regression.

In [ ]:
```
glm.1 <- glm(HighlyCited ~ log(NrCitablePublicationsPerYear) + log(AmountPerYear), da
summary(glm.1)
```

```
Call:
glm(formula = HighlyCited ~ log(NrCitablePublicationsPerYear) +
    log(AmountPerYear), family = binomial, data = exp_medicine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2556  -0.6962   0.2161   0.7206   2.5110

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -7.2742     2.0901   -3.48 0.000501 ***
log(NrCitablePublicationsPerYear)  1.9746     0.1711   11.54  < 2e-16 ***
log(AmountPerYear)                 0.5358     0.1768    3.03 0.002448 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 802.70  on 579  degrees of freedom
Residual deviance: 536.58  on 577  degrees of freedom
AIC: 542.58

Number of Fisher Scoring iterations: 5
```

Die Koeffizienten beider Variablen sind signifikant (α = 0.05), wobei es den Anschein hat, dass die Anzahl der Publikationen pro Jahr einen deutlich stärkeren Einfluss auf die Wahrscheinlichkeit einer überdurchschnittlichen Gesamtzahl an Zitierungen hat.

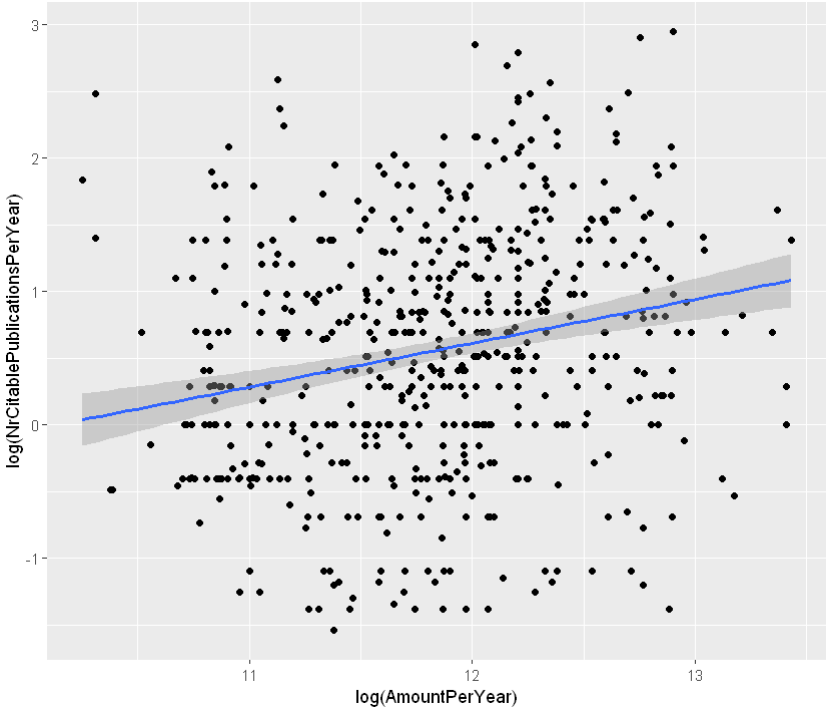
Wir prüfen nun eher informell, ob zwischen den unabhängigen Variablen eine hohe Korrelation besteht.

In [ ]:
```
cor(log(exp_medicine$NrCitablePublicationsPerYear), log(exp_medicine$AmountPerYear))
```

0.225148406706681

In [ ]:
```
ggplot(exp_medicine, aes(x=log(AmountPerYear), y=log(NrCitablePublicationsPerYear)))
  geom_point() +
  geom_smooth(method = "lm")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




- Der Korrelationskoeffizient zwischen den unabhängigen Variablen ist nicht besonders hoch.
- Dies wird auch durch das Streudiagramm bestätigt. Es gibt keine Anzeichen für eine extreme Multikollinearität zwischen den unabhängigen Variablen.

Basierend auf dem erstellten Modell berechnen wir die Wahrscheinlichkeit einer überdurchschnittlich hohen Anzahl von Zitierungen für jede Beobachtung in unserem Daten-Subset.

In [ ]:
```
probabilities <- glm.1 %>% predict(exp_medicine, type = "response")
```

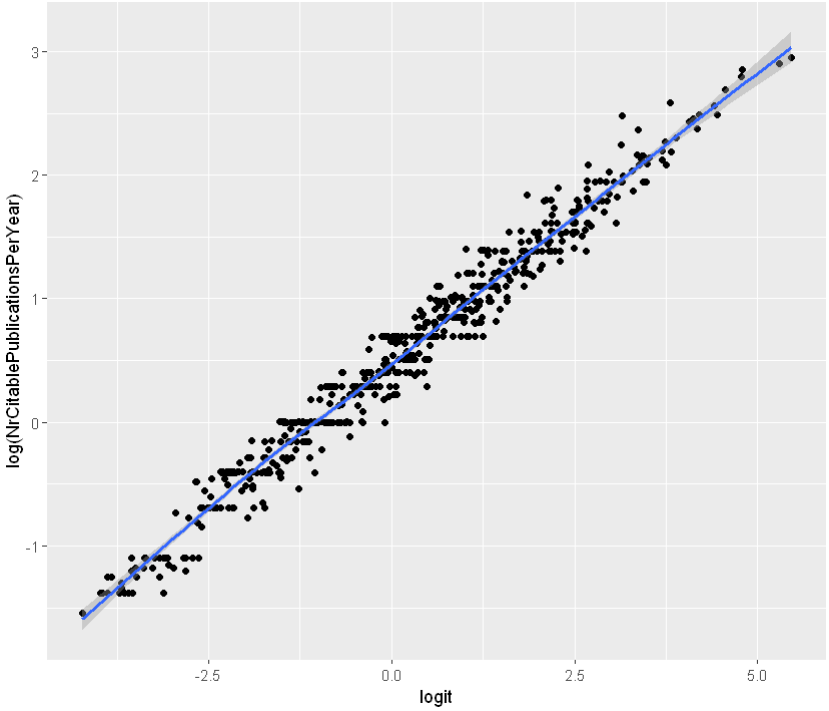
Wir ergänzen die Daten um eine Variable für den Logit jeder Beobachtung, wobei der Logit der natürliche Logarithmus für das Verhältnis der Wahrscheinlichkeit für HighlyCited == 1 zur Wahrscheinlichkeit HighlyCited == 0 ist.

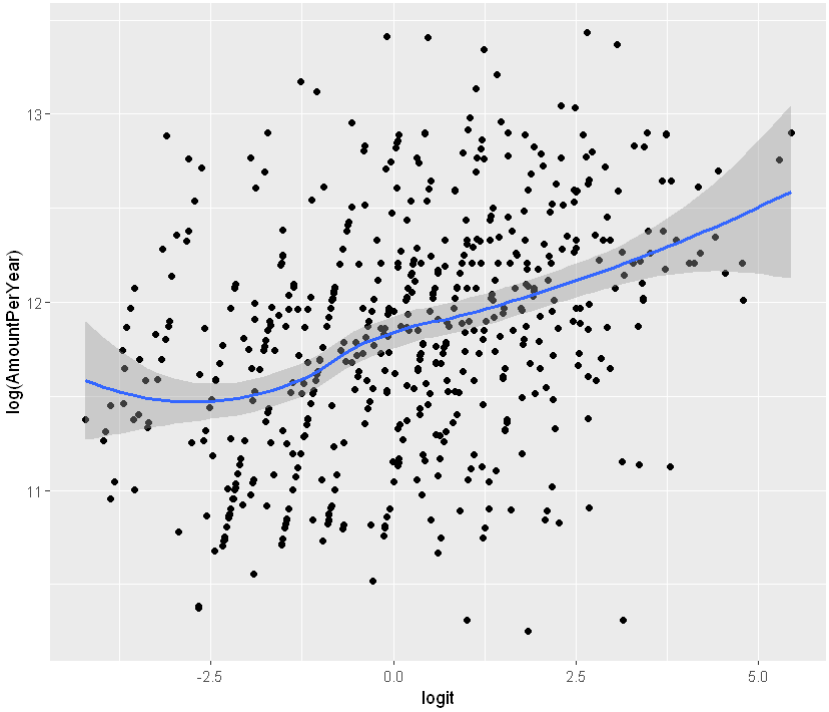
In [ ]:
```
exp_medicine <- exp_medicine %>%
  mutate(logit = log(probabilities / (1 - probabilities)))
```

Um die Annahme der Linearität zwischen den unabhängigen Variablen und dem Logit zu überprüfen, erstellen wir zwei Streudiagramme zur visuellen Analyse.

In [ ]:
```
ggplot(exp_medicine, aes(x=logit, y=log(NrCitablePublicationsPerYear))) +
  geom_point() +
  geom_smooth(method = "loess")

ggplot(exp_medicine, aes(x=logit, y=log(AmountPerYear))) +
  geom_point() +
  geom_smooth(method = "loess")
```

```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```




Die Linearität zwischen dem Logit und der durchschnittlichen Anzahl von Publikationen pro Jahr ist deutlich erkennbar. Hingegen lässt sich kein deutlicher linearer Zusammenhang zwischen dem Logit und dem Förderungsbetrag pro Jahr feststellen. Angesichts der geringeren Einflussstärke des Förderungsbetrags in unserem Modell, entfernen wir diese Variable und führen eine einfache logistische Regression durch, die nur die Anzahl von Publikationen pro Jahr berücksichtigt.

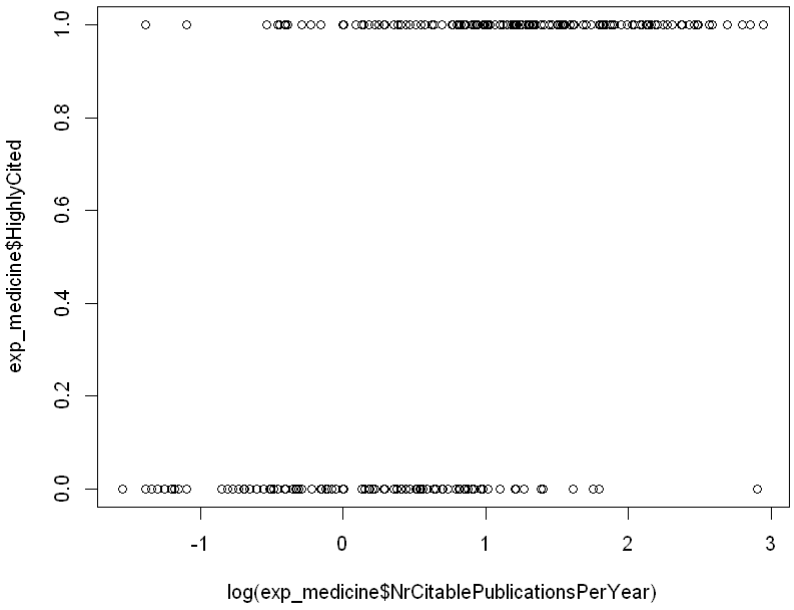
In [ ]:
```
glm.2 <- glm(HighlyCited ~ log(NrCitablePublicationsPerYear), data = exp_medicine, fa
summary(glm.2)
```

```
Call:
glm(formula = HighlyCited ~ log(NrCitablePublicationsPerYear),
    family = binomial, data = exp_medicine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1474  -0.7996   0.2226   0.7625   2.7660

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -0.9763     0.1372  -7.118  1.1e-12 ***
log(NrCitablePublicationsPerYear)  2.0393     0.1707  11.949  < 2e-16 ***
```

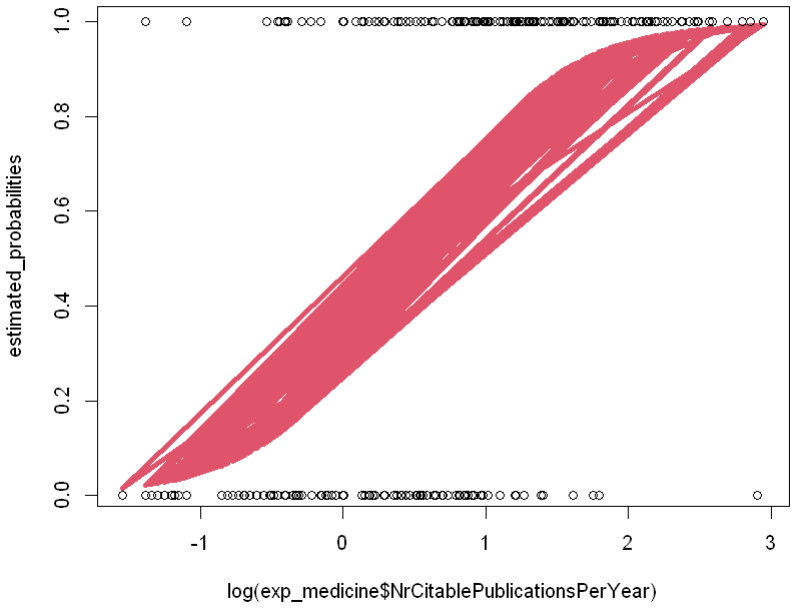
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 802.70  on 579  degrees of freedom
Residual deviance: 545.92  on 578  degrees of freedom
AIC: 549.92

Number of Fisher Scoring iterations: 5
```

Bei der einfachen logistischen Regression besteht die Möglichkeit, die Daten und das Modell auch visuell darzustellen.

In [ ]:
```
plot(log(exp_medicine$NrCitablePublicationsPerYear), exp_medicine$HighlyCited)
```




In [ ]:
```
estimated_probabilities <- glm.2$fitted.values

plot(log(exp_medicine$NrCitablePublicationsPerYear), estimated_probabilities, type="l
lines(log(exp_medicine$NrCitablePublicationsPerYear), exp_medicine$HighlyCited, col=1
```



Die mehrfachen Regressionslinien deuten darauf hin, dass bestimmte Werte der unabhängigen Variable mehrfach auftreten und unterschiedliche Werte der abhängigen Variable aufweisen. Um die logistische Regression visuell ansprechend darzustellen, behalten wir nur eindeutige Werte der unabhängigen Variable und berechnen den Mittelwert der abhängigen Variable für jeden Wert. Basierend darauf erstellen wir ein drittes Modell.

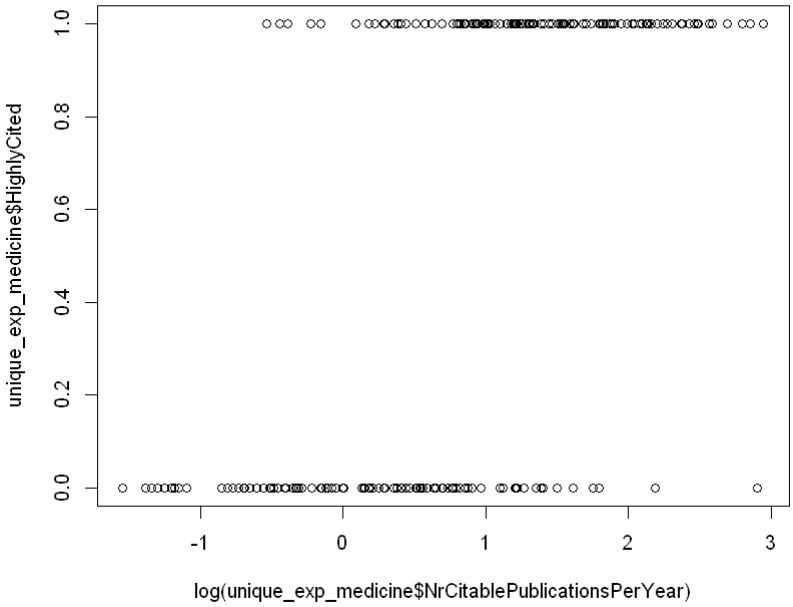
In [ ]:
```
# Duplikate entfernen und Mittelwerte berechnen
unique_exp_medicine <- exp_medicine %>%
  group_by(NrCitablePublicationsPerYear) %>%
  summarize(meanNrCitationsTotal=mean(NrCitationsTotal))

# Mittelwert für Erstellen der binären Variable berechnen
mean <- mean(log(unique_exp_medicine$meanNrCitationsTotal))

# Binäre Variable erstellen
unique_exp_medicine$HighlyCited <- ifelse(log(unique_exp_medicine$meanNrCitationsTota

dim(unique_exp_medicine)
```

266 · 3

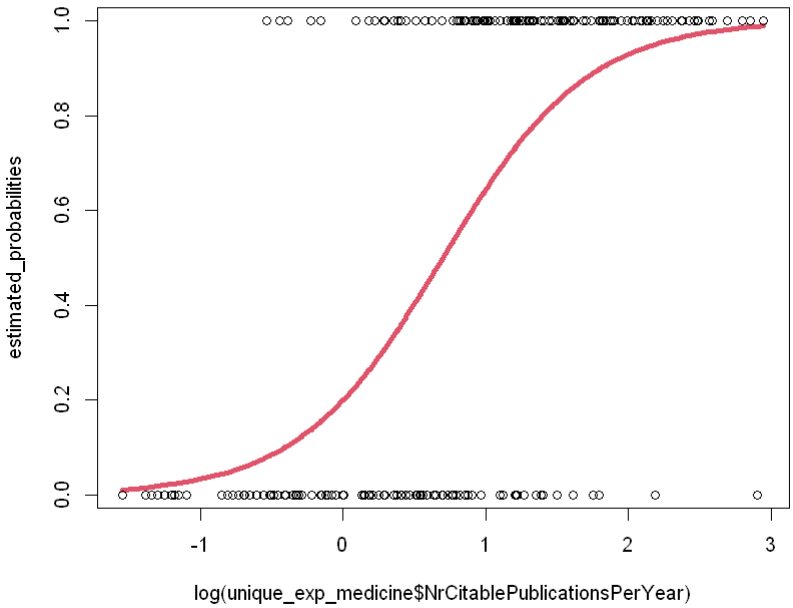
In [ ]:
```
glm.3 <- glm(HighlyCited ~ log(NrCitablePublicationsPerYear), data = unique_exp_medic
summary(glm.3)
```

```
Call:
glm(formula = HighlyCited ~ log(NrCitablePublicationsPerYear),
    family = binomial, data = unique_exp_medicine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9602  -0.6688   0.2495   0.7284   2.2497

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.3851     0.2475  -5.597 2.18e-08 ***
log(NrCitablePublicationsPerYear)   1.9813     0.2439   8.122 4.58e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 367.25  on 265  degrees of freedom
Residual deviance: 244.48  on 264  degrees of freedom
AIC: 248.48

Number of Fisher Scoring iterations: 5
```

In [ ]:
```
plot(log(unique_exp_medicine$NrCitablePublicationsPerYear), unique_exp_medicine$Highl
```




In [ ]:
```
estimated_probabilities <- glm.3$fitted.values

plot(log(unique_exp_medicine$NrCitablePublicationsPerYear), estimated_probabilities,
lines(log(unique_exp_medicine$NrCitablePublicationsPerYear), unique_exp_medicine$High
```




Wir überprüfen den Wert von log(NrCitablePublicationsPerYear), der gerade über der berechneten Wahrscheinlichkeit von 0.5 für eine überdurchschnittliche Anzahl an Zitierungen liegt.

In [ ]:
```
perdicted_probabilities <- predict(glm.3, type = "response")
threshold_value <- with(unique_exp_medicine, log(NrCitablePublicationsPerYear)[perdic
threshold_value
```

0.734397348925319

## Interpretation

Das finale Modell zeigt, dass eine Erhöhung um eine Einheit in der logarithmierten Anzahl der Publikationen pro Jahr (log(NrCitablePublicationsPerYear)) die Wahrscheinlichkeit, eine überdurchschnittliche Anzahl an Zitierungen zu erhalten, um das Siebenfache erhöht ( `exp(1.9813)`

`= 7.2522`). Bezogen auf die Originaldaten bedeutet dies, dass zusätzliche 2.7 Publikationen pro Jahr (exp(1) = 2.7183) erforderlich sind, um die Wahrscheinlichkeit für eine überdurchschnittliche Anzahl an Zitierungen entsprechend zu erhöhen.

Darüber hinaus zeigt sich, dass Projekte mit Schwerpunkt "Experimental Medicine" mindestens 2.08 Publikationen pro Jahr benötigen (`exp(0.7344)`), um insgesamt eine überdurchschnittliche Anzahl an Zitierungen zu erzielen.

## Zusammenfassung

Der generierte und analysierte Datensatz ermöglicht folgende Schlussfolgerungen:

- Der Förderungsbetrag der Projekte spielt eine eher untergeordnete Rolle für die Zitierungen, welche die Publikationen der Projekte erhalten. Dennoch ist er besonders im Kontext der Projekte des Förderungsinstruments "Programmes" und der Gesamtanzahl der Zitierungen nicht völlig zu vernachlässigen.
- Die wesentlichen Faktoren, welche die Anzahl der Zitierungen beeinflussen, sind einerseits die durchschnittliche Anzahl der jährlich veröffentlichten Publikationen und andererseits das Forschungsgebiet. Publikationen im Forschungsgebiet "Experimental Medicine" erhalten die meisten Zitierungen.
- Innerhalb der Projekte mit Schwerpunkt "Experimental Medicine" erhalten die Projekte insgesamt überdurchschnittlich viele Zitierungen, wenn sie mindestens 2.7 Publikationen pro Jahr veröffentlichen.

Abschliessend lässt sich sagen, dass der geringe Einfluss des Förderungsbetrags möglicherweise auch durch das Forschungsgebiet erklärt werden kann. Es ist denkbar, dass Forschung im Bereich "Experimental Medicine" teurer ist und gleichzeitig die meisten Zitierungen generiert werden. Um dies genauer zu untersuchen, sind weitere Analysen erforderlich.