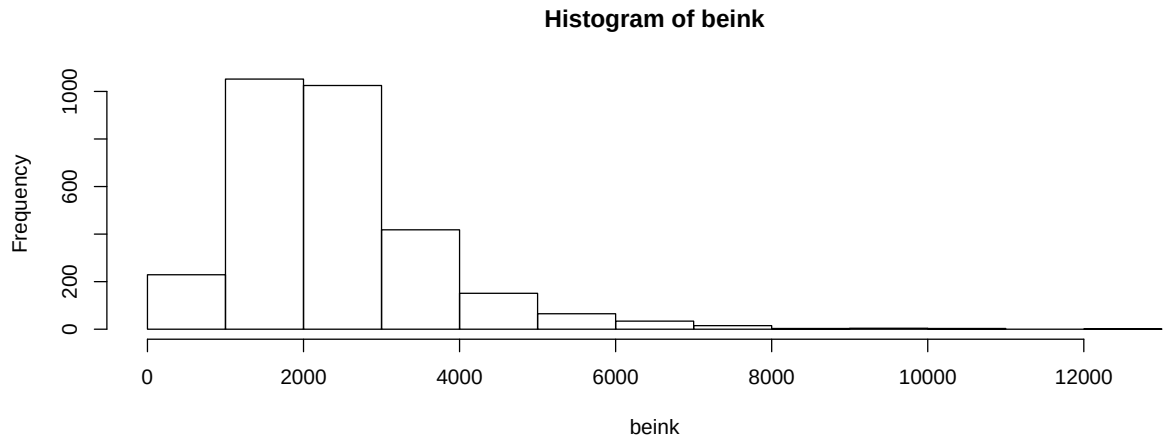
1 Generalisierte lineare Modelle (I)

Aufgabe 1

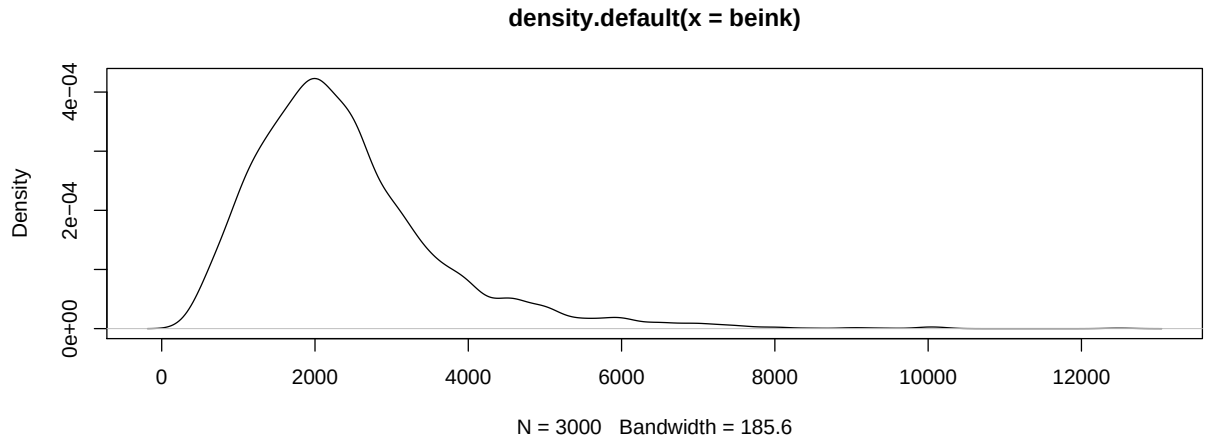
```
# Einlesen der Daten
fakesoep <- read.table("fakesoep.dat",header=T)
attach(fakesoep)
```

a) Man untersucht die Gestalt der Verteilung des Bruttoeinkommens empirisch, z.B. mit einem Histogramm, Kerndichteschätzer oder Boxplot.

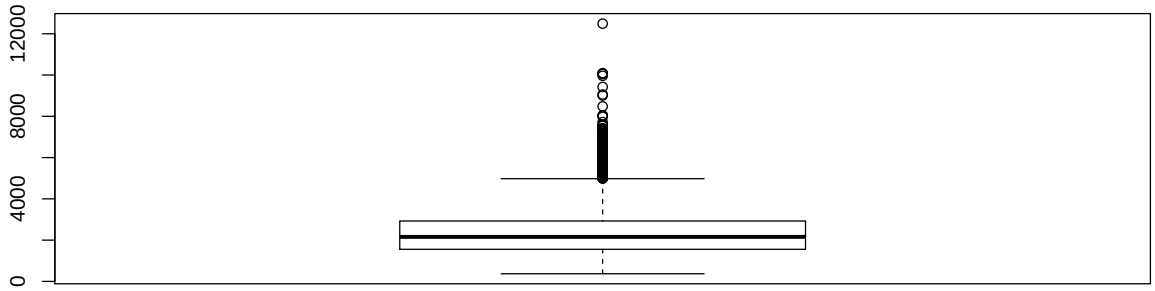
```
hist(beink)
```




```
plot(density(beink))
```



```r
boxplot(beink)
```

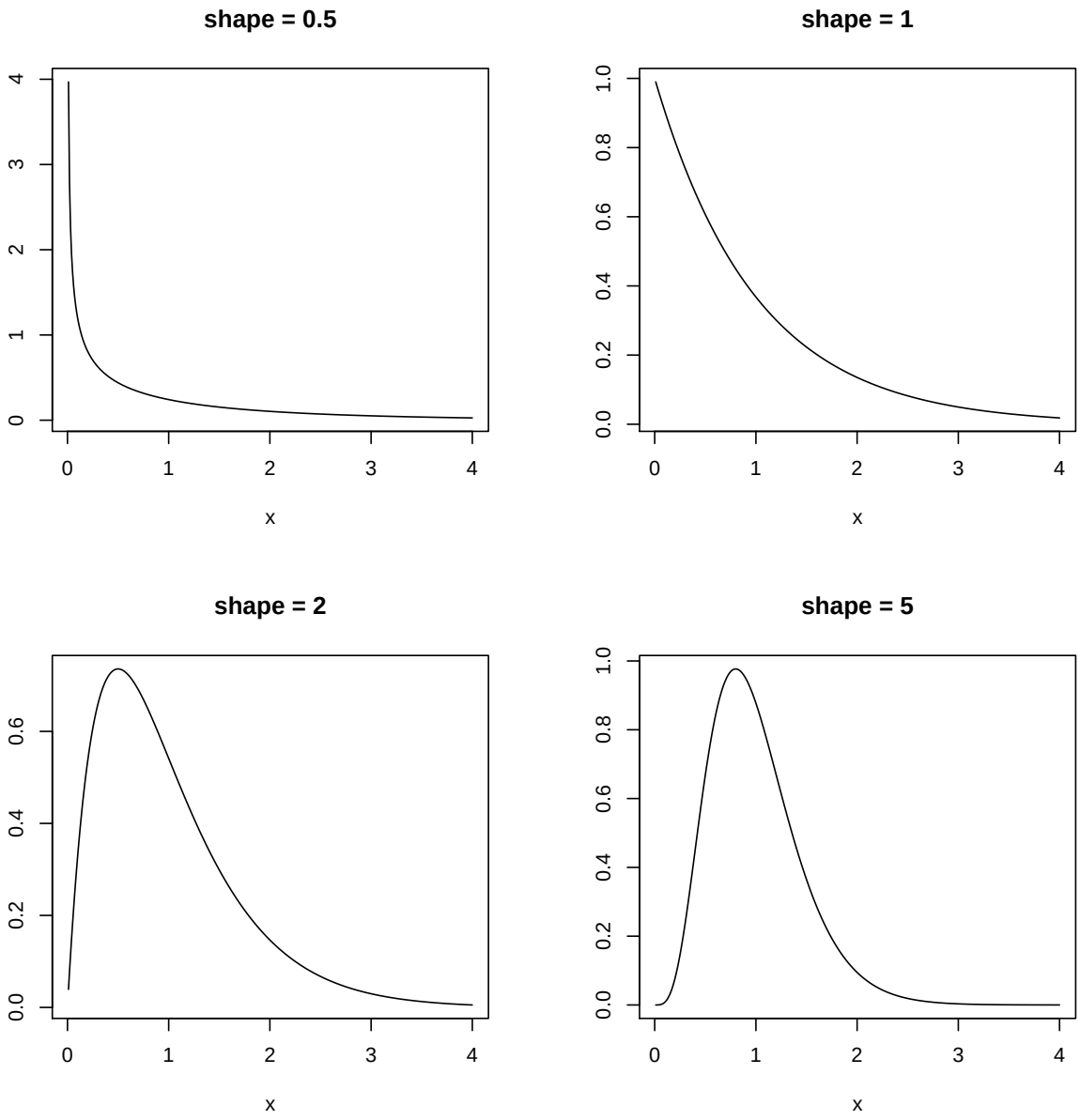



```r
summary(beink) # Median < Mean

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     371    1556    2157    2395    2926   12490
```

→Die Variable beink nimmt positive Werte auf metrischem Skalenniveau an, die (etwas) linkssteil bzw. rechtsschief verteilt sind. Dafür spricht auch die 5-Zahlen-Zusammenfassung aus

b) siehe Übungsmitschrift

c)
```r
# Dichte zeichnen
# help(dgamma)

xseq <- seq(0.01,4,by=0.01)
par(mfrow=c(2,2))
a <- 0.5
plot(xseq,dgamma(xseq,shape=a,scale=1/a),ylab="",xlab="x",
main=paste("shape =",a),type="l")
a <- 1
plot(xseq,dgamma(xseq,shape=a,scale=1/a),ylab="",xlab="x",
main=paste("shape =",a),type="l")
a <- 2
plot(xseq,dgamma(xseq,shape=a,scale=1/a),ylab="",xlab="x",
main=paste("shape =",a),type="l")
a <- 5
plot(xseq,dgamma(xseq,shape=a,scale=1/a),ylab="",xlab="x",
main=paste("shape =",a),type="l")
```



d) siehe Übungsmitschrift und

```
# Natürlicher Link
gammaNat <- glm(beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur,
family=Gamma())
summary(gammaNat)


##
## Call:
## glm(formula = beink ~ groesse + alter + dauer + verh + geschl +
##     deutsch + abitur, family = Gamma())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.46133  -0.31502  -0.05706   0.19162   1.91254
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.064e-03  8.659e-05  12.287  < 2e-16 ***
## groesse     -3.287e-06  4.680e-07  -7.024 2.66e-12 ***
## alter       -2.452e-07  3.966e-07  -0.618    0.536
## dauer       -3.839e-06  3.701e-07 -10.372  < 2e-16 ***
```

```
## verh        -7.339e-06  7.111e-06  -1.032    0.302
## geschl       1.419e-04  9.549e-06  14.855  < 2e-16 ***
## deutsch     -1.065e-05  1.210e-05  -0.880    0.379
## abitur      -1.318e-04  6.220e-06 -21.196  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1748559)
##
##     Null deviance: 757.46  on 2999  degrees of freedom
## Residual deviance: 507.22  on 2992  degrees of freedom
## AIC: 49212
##
## Number of Fisher Scoring iterations: 5
```

```
# Log Link
gammaLog <- glm(beink ~ groesse +alter + dauer + verh + geschl + deutsch + abitur,
family=Gamma(link=log))
summary(gammaLog)
```

```
##
## Call:
## glm(formula = beink ~ groesse + alter + dauer + verh + geschl +
##     deutsch + abitur, family = Gamma(link = log))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.47399  -0.31490  -0.05961   0.18374   1.94176
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.2202325  0.2182771  28.497  < 2e-16 ***
## groesse      0.0082530  0.0011930   6.918 5.58e-12 ***
## alter        0.0001786  0.0009345   0.191    0.848
## dauer        0.0119425  0.0009816  12.166  < 2e-16 ***
## verh        -0.0178813  0.0173453  -1.031    0.303
## geschl      -0.3179608  0.0216168 -14.709  < 2e-16 ***
## deutsch      0.0050755  0.0279452   0.182    0.856
## abitur       0.3466434  0.0167621  20.680  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1747557)
##
##     Null deviance: 757.46  on 2999  degrees of freedom
## Residual deviance: 502.50  on 2992  degrees of freedom
## AIC: 49184
##
## Number of Fisher Scoring iterations: 5
```

→Da man für beide Modelle die gleiche Verteilungsannahme tätigt, kann man die Anpassungsgüte anhand der Devianz, $Dev = -2\phi \sum_i (l_i(\hat{\mu}_i) - l_i(y_i))$, vergleichen. Da das Modell mit log-Link mit 502.50 den kleineren Wert aufweist (nat. Link: 507.22), ist dieses zu bevorzugen. Vergleich über AIC führt (hier selbstverständlich) zum selben Ergebnis.
→Man beachte, dass hier der Dispersionsparameter gemäß der Formel $\hat{\phi} = 1/(n-p) \sum_i (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i^2$ mitgeschätzt wird.
→Ein weiterer Vorteil des log-Links ist, dass die Erwartungswerte aufgrund von $\mu = exp(\eta)$ nur positive Werte annehmen können. Das ist beim natürlichen (inversen) Link mit $\mu = \eta^{-1}$ nicht der Fall, d.h. gegebenenfalls werden Restriktionen an den Parametervektor β erforderlich, um $\mu > 0$ zu garantieren.

Vorsicht bei der Interpretation der Parameter: bei Verwenden des inversen Links gilt $\mu = \eta^{-1}$, d.h. negative Werte für Koeffizienten von β indizieren einen steigenden Erwartungswert.

```
cbind(gammaNat$coefficients, gammaLog$coefficients)
```

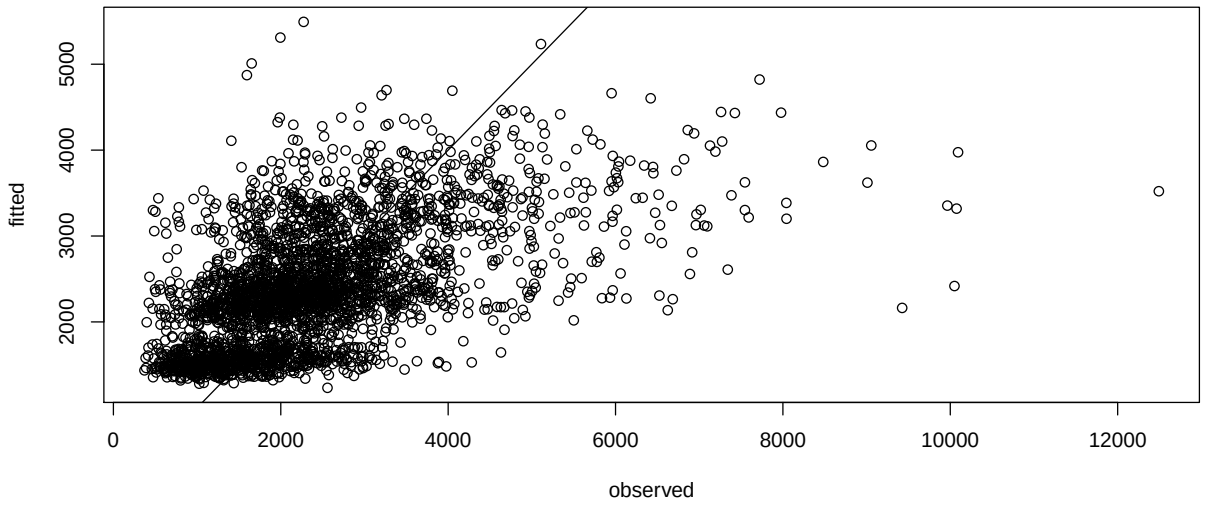
```
##                       [,1]          [,2]
## (Intercept)  1.063881e-03  6.2202325333
## groesse     -3.287389e-06  0.0082529622
## alter       -2.452122e-07  0.0001785662
## dauer       -3.839142e-06  0.0119424909
## verh        -7.338988e-06 -0.0178813399
## geschl       1.418518e-04 -0.3179607922
## deutsch     -1.064851e-05  0.0050755384
## abitur      -1.318256e-04  0.3466434243
```

→Die Effekte der Kovariablen auf das Bruttoeinkommen weisen in beiden Modellen (in fast allen Fällen) in die gleiche Richtung (positiv: groesse, alter, dauer, deutsch, abitur; negativ: geschl), zu beachten ist, dass der Effekt von verh beim natürlichen Link in die andere Richtung weist als beim log-Link. Der Effekt ist allerdings auch nicht signifikant. Das selbe gilt für alter und deutsche Staatsangehörigkeit. Bzgl. Tests sei aber auf die kommende Übung verwiesen.
Für eine tiefergehende inhaltliche Interpretation der Parameter wäre die Einbeziehung von Interaktionstermen (z.B. verh und geschl) wünschenswert. also etwa (Zugabe, keine Aufgabenstellung aber interessant!):

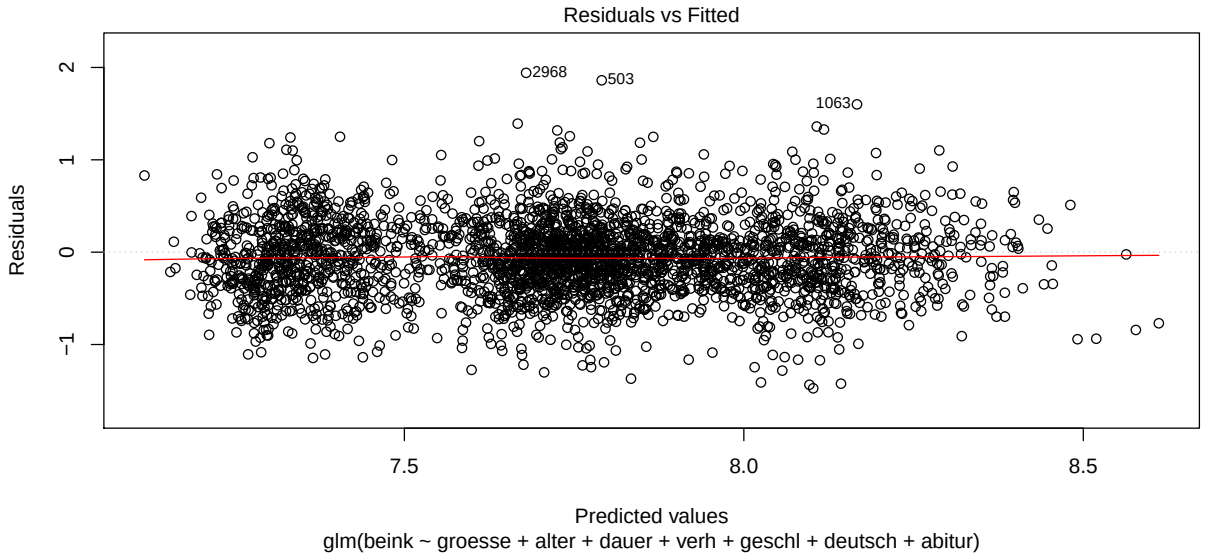
```
gammaLog2 <- update(gammaLog, . ~ . + verh:geschl)
summary(gammaLog2)
```

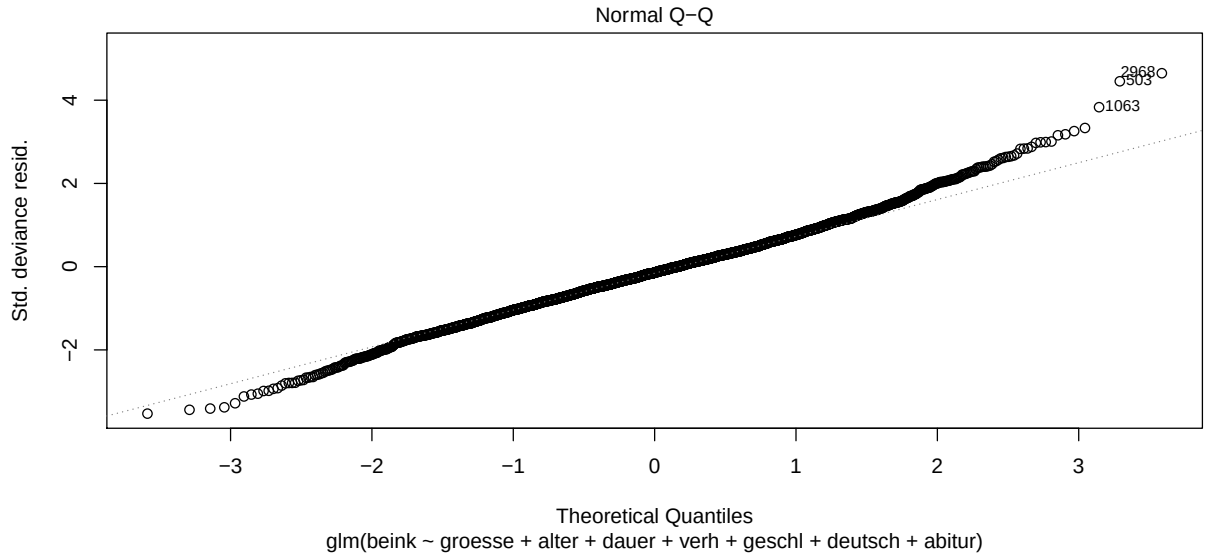
```
##
## Call:
## glm(formula = beink ~ groesse + alter + dauer + verh + geschl +
##     deutsch + abitur + verh:geschl, family = Gamma(link = log))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.41195  -0.30952  -0.05603   0.18303   1.86195
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.142e+00  2.150e-01  28.562  < 2e-16 ***
## groesse      8.106e-03  1.174e-03   6.906 6.06e-12 ***
## alter        7.459e-05  9.195e-04   0.081 0.935351
## dauer        1.186e-02  9.658e-04  12.277  < 2e-16 ***
## verh         1.292e-01  2.252e-02   5.737 1.06e-08 ***
## geschl      -1.157e-01  2.986e-02  -3.875 0.000109 ***
## deutsch      1.254e-02  2.751e-02   0.456 0.648615
## abitur       3.415e-01  1.650e-02  20.698  < 2e-16 ***
## verh:geschl -3.142e-01  3.204e-02  -9.805  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1691591)
##
##     Null deviance: 757.46  on 2999  degrees of freedom
## Residual deviance: 486.27  on 2991  degrees of freedom
## AIC: 49084
##
## Number of Fisher Scoring iterations: 5
```

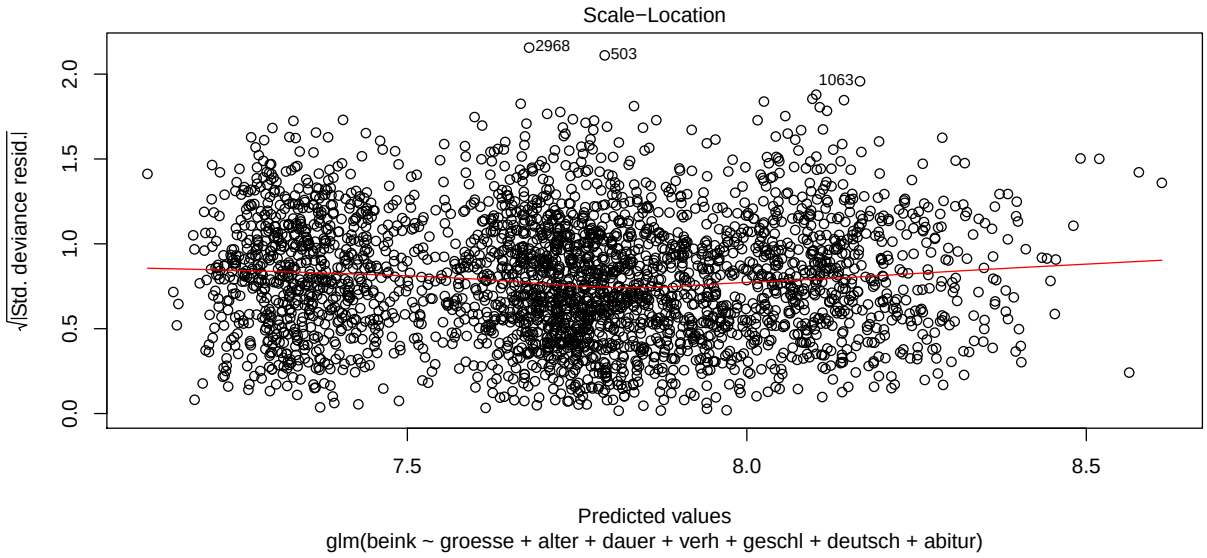
e)
```r
# Graphik-Parameter zurück
par(mfrow=c(1,1))

# Plot: Beobachtete gegen geschätzte Werte
plot(beink,gammaLog$fitted,xlab="observed",ylab="fitted")
abline(0,1)
```

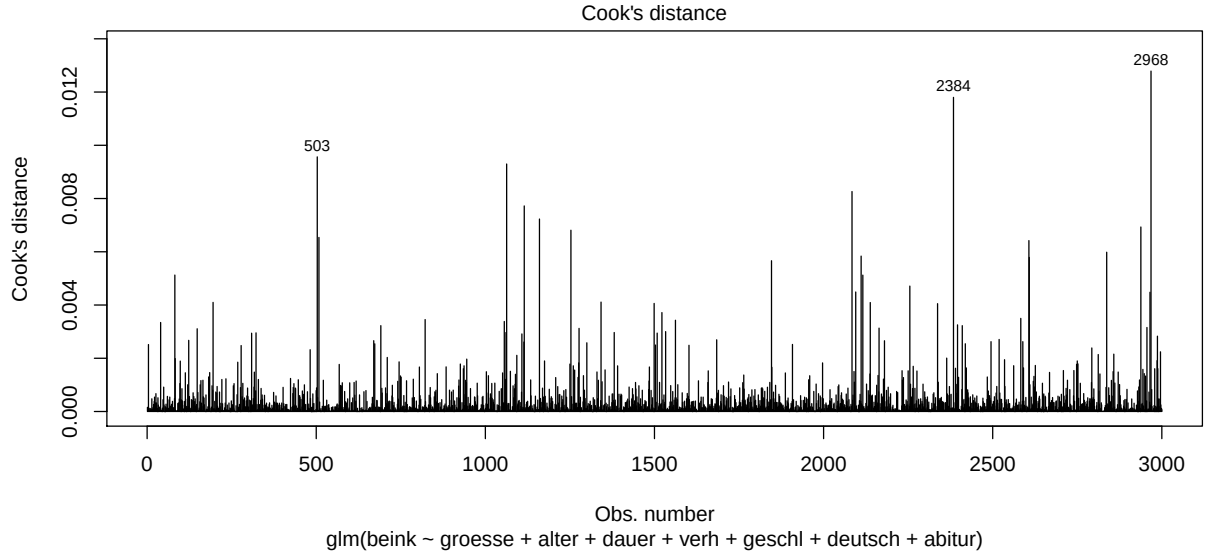


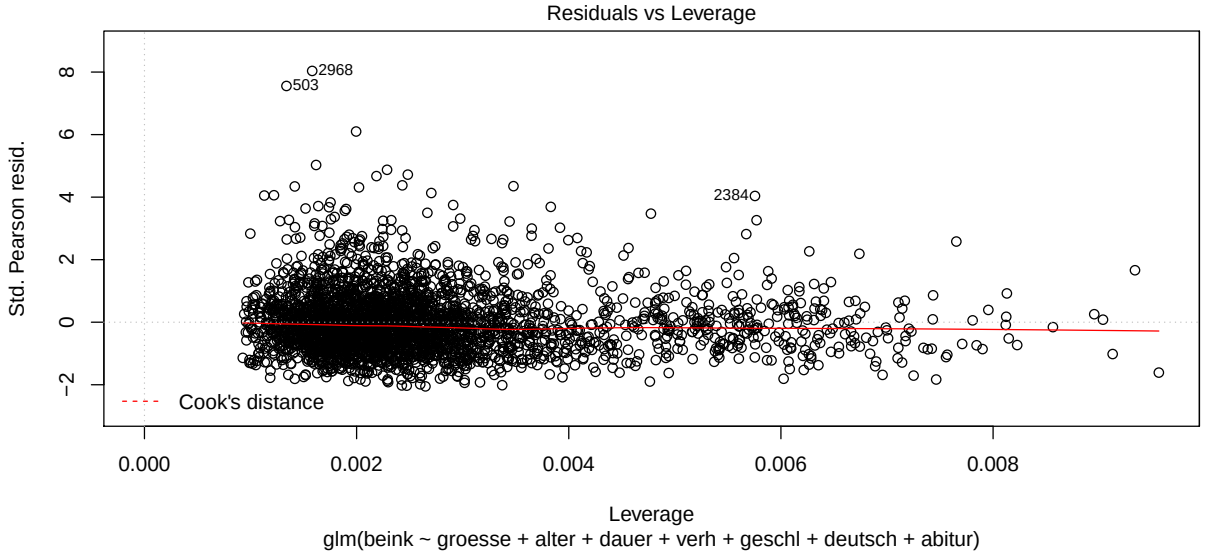

```r
# Zusatz
# points(beink[geschl==1 & abitur==0],gammaLog$fitted[geschl==1 & abitur==0],col=2)


# mit plot.glm
plot(gammaLog, which=1:5)
```




Normal Q−Q
Std. deviance resid.
2968
503
1063
−3
−2
−1
0
1
2
3
4
Theoretical Quantiles
glm(beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur)


Scale−Location
√|Std. deviance resid.|
2968
503
1063
0.0
0.5
1.0
1.5
2.0
7.5
8.0
8.5
Predicted values
glm(beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur)


Cook's distance
Cook's distance
503
2384
2968
0.000
0.004
0.008
0.012
0
500
1000
1500
2000
2500
3000
Obs. number
glm(beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur)



5 Plots werden von generischer plot-Fkt. ausgegeben (Diagnostik-Maße):

- Residuen gegen gefittete Werte (mit Glätter)
- QQ-Plot von standardisierten Devianzresiduen gegen theoret. Quantile der N(0,1) (vgl. später)
- sqrt(stand. Devianzresiduen) gegen gefittete Werte (mit Glätter)
- Cook's Distance (vgl. später)
- Residual gegen Leverage. Veranschaulicht Konturen mit gleichen Werten von Cook's Distance (vgl. später)

Extreme Beobachtungen werden markiert.

f) Die backward-Variablenselektion bzgl. des AIC ist in der Funktion step implementiert. Hierzu ist das glm-Objekt des vollen Modells (object) zu übergeben, sowie "backward"für das Argument method zu wählen.

→Funktionsweise: Sukzessive wird das Modell jeweils mit Weglassen einer Variable berechnet. Jene Variable, deren Weglassen das AIC am stärksten verringert, wird im nächsten Schritt entfernt. Abbruch, falls keine Verringerung des AIC mehr auftritt.

```
gammaNatsel <- step(gammaNat,method="backward")

## Start:  AIC=49212.34
## beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur
##
##           Df Deviance   AIC
## - alter    1   507.28 49211
## - deutsch  1   507.35 49211
## - verh     1   507.40 49211
## <none>         507.22 49212
## - groesse  1   515.77 49259
## - dauer    1   525.58 49315
## - geschl   1   546.71 49436
## - abitur   1   582.26 49640
##
## Step:  AIC=49210.75
## beink ~ groesse + dauer + verh + geschl + deutsch + abitur
##
##           Df Deviance   AIC
## - deutsch  1   507.43 49210
## - verh     1   507.55 49210
## <none>         507.28 49211
```

```
## - groesse  1   515.86 49258
## - dauer    1   533.59 49359
## - geschl   1   547.55 49439
## - abitur   1   583.22 49643
##
## Step:  AIC=49209.65
## beink ~ groesse + dauer + verh + geschl + abitur
##
##           Df Deviance   AIC
## - verh     1   507.69 49209
## <none>         507.43 49210
## - groesse  1   516.75 49261
## - dauer    1   533.93 49359
## - geschl   1   547.70 49437
## - abitur   1   583.64 49642
##
## Step:  AIC=49209.23
## beink ~ groesse + dauer + geschl + abitur
##
##           Df Deviance   AIC
## <none>         507.69 49209
## - groesse  1   516.82 49259
## - dauer    1   536.53 49371
## - geschl   1   549.12 49443
## - abitur   1   584.15 49643
```

```
summary(gammaNatsel)
```

```
##
## Call:
## glm(formula = beink ~ groesse + dauer + geschl + abitur, family = Gamma())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.47491  -0.31786  -0.05683   0.18823   1.92647
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.037e-03  8.108e-05  12.794  < 2e-16 ***
## groesse     -3.265e-06  4.505e-07  -7.248 5.36e-13 ***
## dauer       -4.057e-06  3.012e-07 -13.470  < 2e-16 ***
## geschl       1.425e-04  9.394e-06  15.173  < 2e-16 ***
## abitur      -1.325e-04  6.200e-06 -21.378  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.17554)
##
##     Null deviance: 757.46  on 2999  degrees of freedom
## Residual deviance: 507.69  on 2995  degrees of freedom
## AIC: 49209
##
## Number of Fisher Scoring iterations: 5
```

```
gammaLogsel <- step(gammaLog,method="backward")
```

```
## Start:  AIC=49183.51
## beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur
##
##           Df Deviance   AIC
## - deutsch  1   502.50 49182
```

```
## - alter    1   502.50 49182
## - verh     1   502.68 49183
## <none>         502.50 49184
## - groesse  1   510.86 49229
## - dauer    1   527.42 49324
## - geschl   1   539.58 49394
## - abitur   1   579.61 49623
##
## Step:  AIC=49181.55
## beink ~ groesse + alter + dauer + verh + geschl + abitur
##
##           Df Deviance   AIC
## - alter    1   502.51 49180
## - verh     1   502.69 49181
## <none>         502.50 49182
## - groesse  1   511.31 49230
## - dauer    1   527.42 49322
## - geschl   1   540.25 49396
## - abitur   1   579.82 49622
##
## Step:  AIC=49179.59
## beink ~ groesse + dauer + verh + geschl + abitur
##
##           Df Deviance   AIC
## - verh     1   502.69 49179
## <none>         502.51 49180
## - groesse  1   511.41 49229
## - dauer    1   535.67 49367
## - geschl   1   540.50 49395
## - abitur   1   580.49 49624
##
## Step:  AIC=49178.7
## beink ~ groesse + dauer + geschl + abitur
##
##           Df Deviance   AIC
## <none>         502.69 49179
## - groesse  1   511.87 49229
## - dauer    1   536.22 49369
## - geschl   1   540.81 49395
## - abitur   1   580.61 49623
```

```
summary(gammaLogsel)
```

```
##
## Call:
## glm(formula = beink ~ groesse + dauer + geschl + abitur, family = Gamma(link = log))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.46880  -0.31662  -0.05935   0.18408   1.93111
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.2008585  0.2060978   30.09  < 2e-16 ***
## groesse      0.0083614  0.0011501    7.27 4.56e-13 ***
## dauer        0.0118575  0.0008518   13.92  < 2e-16 ***
## geschl      -0.3141680  0.0211712  -14.84  < 2e-16 ***
## abitur       0.3468862  0.0167144   20.75  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1745132)
```

```
##
##     Null deviance: 757.46  on 2999  degrees of freedom
## Residual deviance: 502.69  on 2995  degrees of freedom
## AIC: 49179
##
## Number of Fisher Scoring iterations: 4
```

Beim Modell mit natürlichem Link sowie beim log-Link wird das Modell ohne deutsch, alter und verh gewählt.Nach wie vor ist das Modell mit Log-Link zu bevorzugen. Wie zuvor ist hier die Devianz bzw. das AIC niedriger im Vergleich zum Modell mit natürlichem Link.

Aufgabe 2

a) siehe Übungsmitschrift

b)
```
# Modell mit natürlichem Link:
help(inverse.gaussian)
invgNat <- glm(beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur,
family=inverse.gaussian())
```

```
## Error: no valid set of coefficients has been found: please supply starting values
```

Der Schätzer mit natürlichem Link kann nicht berechnet werden, da R hier keine Startparameter für den Schätzalgorithmus findet, die die durch den natürlichen Link induzierten Restriktionen erfüllen.

```
# Modell mit log-Link:
invgLog <- glm(beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur,
family=inverse.gaussian(link=log))
#Genau wie im Gamma-Modell unterliegt der Parametervektor auch bei log-Link
#keinen Restriktionen

# Summary
summary(invgLog)
```

```
##
## Call:
## glm(formula = beink ~ groesse + alter + dauer + verh + geschl +
##     deutsch + abitur, family = inverse.gaussian(link = log))
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.040178  -0.007013  -0.001256   0.003663   0.035412
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.3768957  0.2197381  29.020  < 2e-16 ***
## groesse      0.0077052  0.0012067   6.385 1.97e-10 ***
## alter       -0.0006482  0.0009130  -0.710  0.47780
## dauer        0.0136441  0.0010259  13.300  < 2e-16 ***
## verh        -0.0547059  0.0174229  -3.140  0.00171 **
## geschl      -0.3297230  0.0216232 -15.249  < 2e-16 ***
## deutsch     -0.0096808  0.0276394  -0.350  0.72617
## abitur       0.3432797  0.0181570  18.906  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for inverse.gaussian family taken to be 8.03145e-05)
##
##     Null deviance: 0.37536  on 2999  degrees of freedom
## Residual deviance: 0.26421  on 2992  degrees of freedom
## AIC: 49412
##
## Number of Fisher Scoring iterations: 6
```

```
# zum Vergleich
summary(gammaLog)
```

```
##
## Call:
## glm(formula = beink ~ groesse + alter + dauer + verh + geschl +
##     deutsch + abitur, family = Gamma(link = log))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.47399  -0.31490  -0.05961   0.18374   1.94176
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.2202325  0.2182771  28.497  < 2e-16 ***
## groesse      0.0082530  0.0011930   6.918 5.58e-12 ***
## alter        0.0001786  0.0009345   0.191    0.848
## dauer        0.0119425  0.0009816  12.166  < 2e-16 ***
## verh        -0.0178813  0.0173453  -1.031    0.303
## geschl      -0.3179608  0.0216168 -14.709  < 2e-16 ***
## deutsch      0.0050755  0.0279452   0.182    0.856
## abitur       0.3466434  0.0167621  20.680  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1747557)
##
##     Null deviance: 757.46  on 2999  degrees of freedom
## Residual deviance: 502.50  on 2992  degrees of freedom
## AIC: 49184
##
## Number of Fisher Scoring iterations: 5
```

Übergibt man als Startwerte die Parameterschätzer aus dem Gamma-Modell, beginnt R zwar mit der Berechnung, der Algorithmus konvergiert allerdings nicht:

```
invgNat <- glm(beink ~ groesse + alter + dauer + verh + geschl + deutsch + abitur,
family=inverse.gaussian(),start=gammaNat$coef)
```

```
tail(warnings(),8)
```

```
## Warnmeldungen:
## 1: In sqrt(eta) : NaNs produced
## 2: step size truncated due to divergence
## 3: In sqrt(eta) : NaNs produced
## 4: In sqrt(eta) : NaNs produced
## 5: In sqrt(eta) : NaNs produced
## 6: In sqrt(eta) : NaNs produced
## 7: In sqrt(eta) : NaNs produced
## 8: In sqrt(eta) : NaNs produced
```