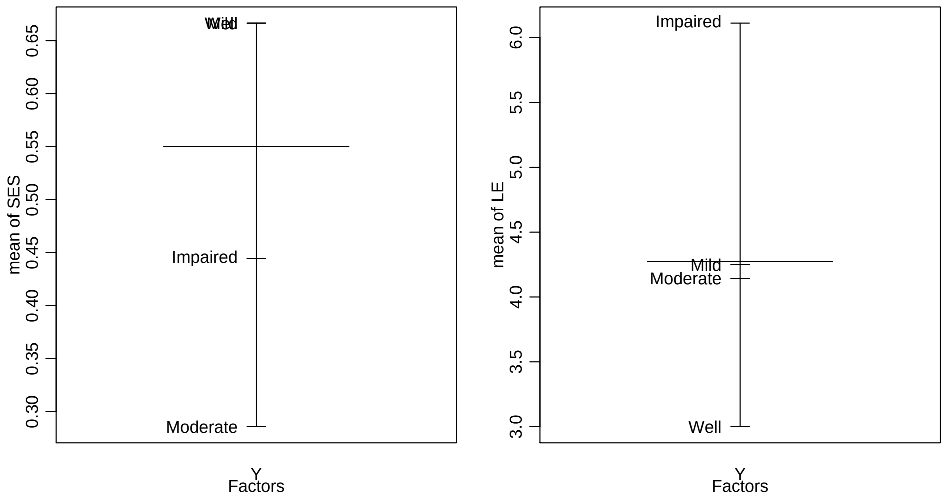
Angewandte Regression — Musterlösungen zur Serie 10

1. **a)** Ein positiver Koefffizient β_j bewirkt, dass $P[Y < k]$ sinkt falls die Variable $x^{(j)}$ wächst. In den Abbildungen ist zu erkennen, dass der Grad der Geistesschwäche bei der hohen Stufe von `SES` tiefer ist (mehr Personen mit `Well` und `Mild`). Ein hoher LE-Wert führt hingegen zu stärkerer Geistesschwäche (`Impaired`) als ein tiefer LE-Wert. Deshalb erwarten wir ein negatives β_1 und ein positives β_2.



b) Schätzung des Modelles mit R-Output von `summary(r.mental)`:

```
Re-fitting to get Hessian

Call:
polr(formula = Y ~ SES + LE, data = d.mental)

Coefficients:
         Value Std. Error   t value
SES -1.1112212  0.6108779 -1.819056
LE   0.3188592  0.1209920  2.635375

Intercepts:
                  Value   Std. Error t value
Well|Mild         -0.2819  0.6423    -0.4389
Mild|Moderate      1.2128  0.6607     1.8357
Moderate|Impaired  2.2094  0.7210     3.0644

Residual Deviance: 99.0979
AIC: 109.0979
```

Mittels `r.mental$convergence` können wir nachsehen, ob der Algorithmus konvergiert hat. Dies ist der Fall (Code 0).
Wir erhalten also die folgenden geschätzten Schwellenwerte und Koeffizienten:

$$\widehat{\alpha}_1 = -0.28\,, \quad \widehat{\alpha}_2 = 1.21\,, \quad \widehat{\alpha}_3 = 2.20\,, \quad \widehat{\beta}_1 = -1.11\,, \quad \widehat{\beta}_2 = 0.32$$

Das negative Vorzeichen von $\widehat{\beta}_1$ bedeutet, dass die Wahrscheinlichkeit von Geistesschwäche für den hohen Status ($\texttt{SES} = 1$) tiefer ist als für den niedrigen Status. Das positive Vorzeichen von $\widehat{\beta}_2$ bedeutet, dass die Wahrscheinlichkeit für einen höheren Grad von Geistesschwäche mit steigendem "life events index" LE zunimmt.
Wir haben diese Vermutung schon beim Betrachten der Abbildungen in Teilaufgabe (a) geäussert. Das *proportional-odds model* bestätigt also die dort gewonnenen Eindrücke.

c) Es gilt

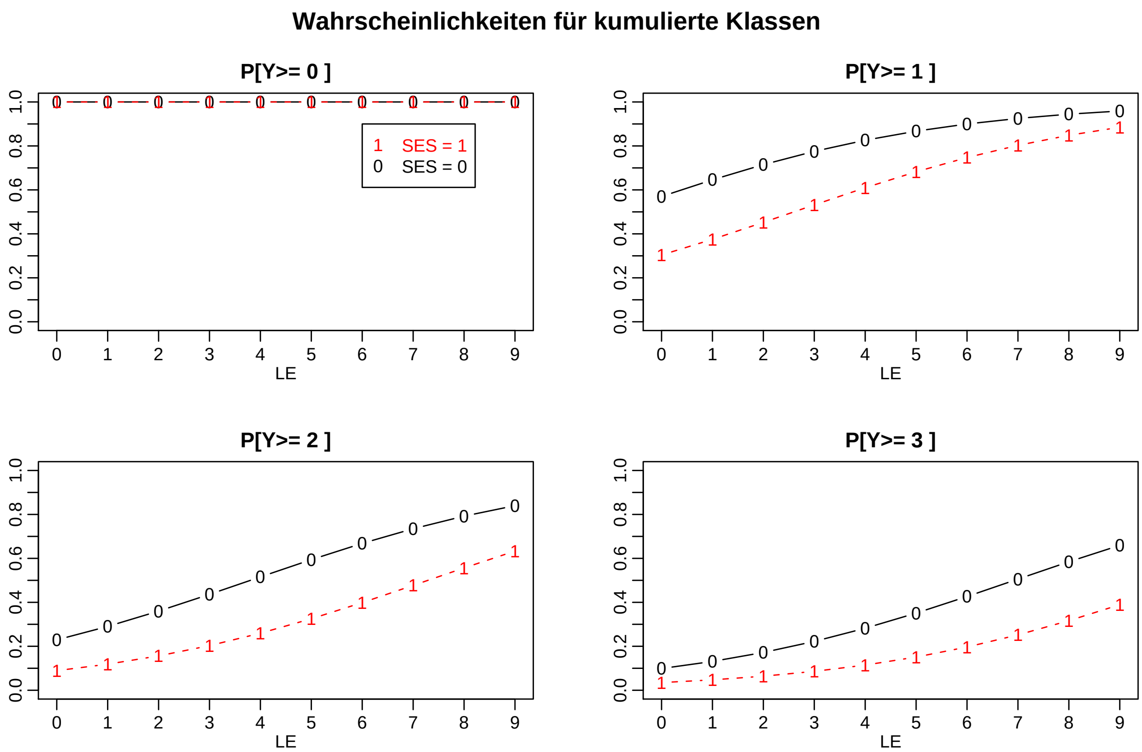
$$\frac{\text{odds}\langle Y \geq k \mid \texttt{SES} = 1\rangle}{\text{odds}\langle Y \geq k \mid \texttt{SES} = 0\rangle} = (\exp\langle\widehat{\beta}_1\rangle)^{1-0} = \exp\langle -1.11\rangle = 0.33\,.$$

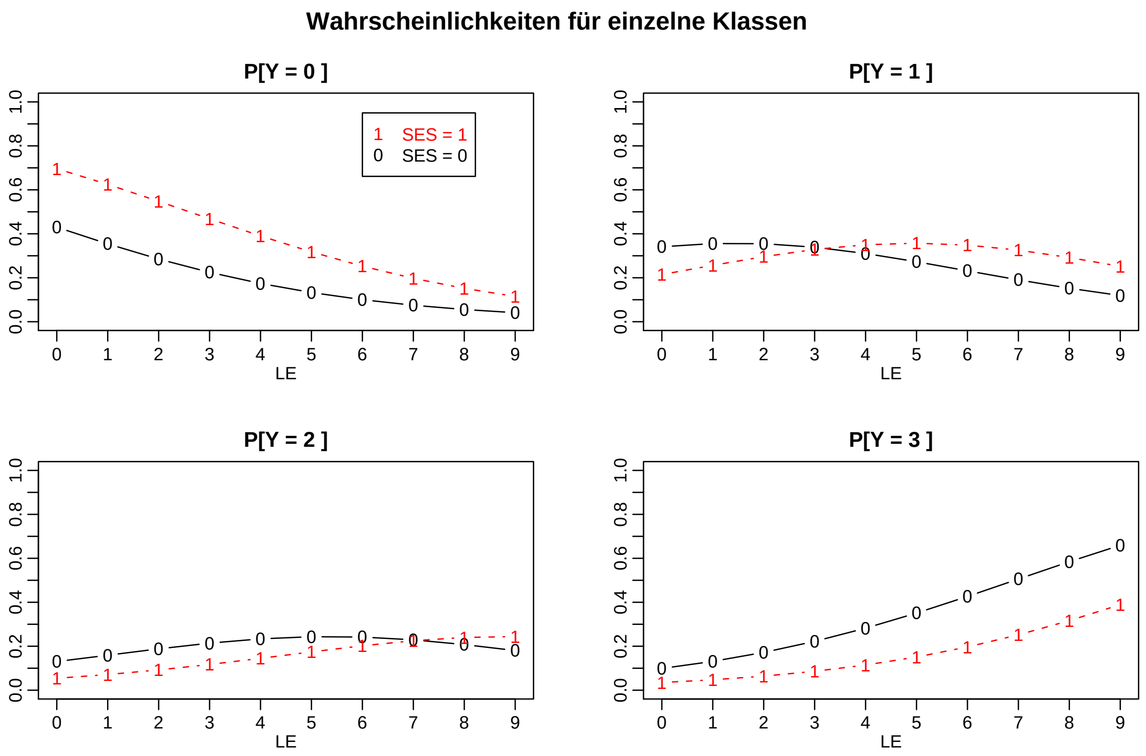
dies entspricht dem gesuchten Faktor.

d) Die Wahrscheinlichkeiten berechnen sich nach der Formel

$$P\langle Y \geq k \mid x_1, x_2\rangle = \frac{\exp\langle\beta_1\, x_1 + \beta_2\, x_2 - \alpha_k\rangle}{1 + \exp\langle\beta_1\, x_1 + \beta_2\, x_2 - \alpha_k\rangle}$$

mit $x_1 = \texttt{LE}$ und $x_2 = \texttt{SES}$. Sie sind in der folgenden Abbildung dargestellt.





Interpretation: Die Wahrscheinlichkeit, in die kleinste Kategorie zu fallen entspricht der Wahrscheinlichkeit guter geistiger Gesundheit. Sie sinkt mit wachsender Anzahl wichtiger Lebensereignisse LE. Die Kurven fallen sowohl für hohen, als auch für niedrigen sozialökonomischen Status `SES`.
Die Wahrscheinlichkeiten sind grösser für die Stufe `SES=1`. Natürlich muss es umgekehrt sein für die Klasse `k=3`. Dort sind die Wahrscheinlichkeiten für `SES=0` und wachsendem `LE` grösser. Für die mittleren Klassen gibt es überschneidende Kurven.

2. **a)** Wir erhalten einen Koeffizienten von -0.534 für die Stufen `Schule = Abitur` und `Hauptv = Staat`. Die Referenzstufe beim Faktor `Schule` ist `ungelernt` und beim Faktor `Hauptv` `Einzelne`. D.h. das Wettverhältnis

$$\frac{P\langle Hauptv = Staat \mid \underline{x}_i\rangle}{P\langle Hauptv = Einzelne \mid \underline{x}_i\rangle}$$

multipliziert sich mit dem Faktor $exp(-0.534) = 0.59$, wenn von der Stufe `ungelernt` zur Stufe `Abitur` gewechselt wird. Das Wettverhältnis wird also kleiner, d. h. Abiturienten halten den Einzelnen eher für verantwortlich als Ungelernte.

b) Wie man sieht, sind die Doppelverhältnisse für alle Stufen der Beeinträchtigung ungefähr gleich. Die Wettverhältnisse *Einzelne* gegen *Staat und beide* verkleinern sich mit einem Faktor von etwa $1/0.83$, wenn man statt Studierende Personen mit anderen Schulabschlüssen befragt.

```
> t.lor <- NULL
> for (l.b in levels(t.dt$Beeintr)) {
    l.pp <- t.p[t.dt$Beeintr==l.b,]
    t.lor <- c(t.lor, log(l.pp[5,1]/sum(l.pp[5,2:3]))-
        log(sum(l.pp[1:4,1])/sum(l.pp[1:4,2:3])) ) }
> names(t.lor) <- levels(t.dt$Beeintr)

## log-odds-ratio
> t.lor
     nicht      etwas   ziemlich       sehr
-0.1470052 -0.1793573 -0.1773485 -0.2119173

## odds-ratio
> exp(t.lor)
    nicht     etwas  ziemlich      sehr
0.8632895 0.8358072 0.8374878 0.8090316
```

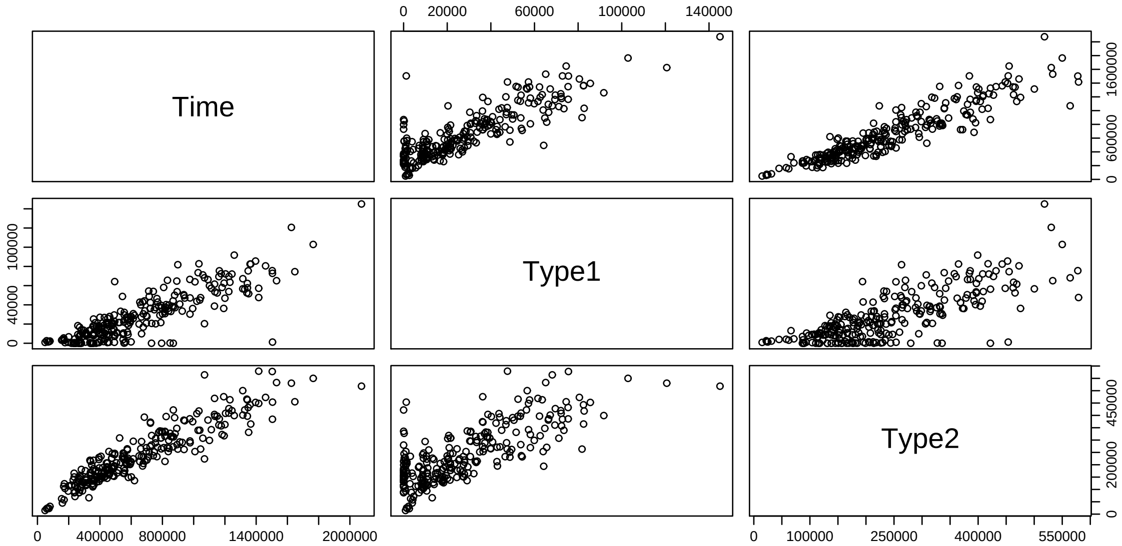
c) Da die Funktion `step` hier nicht funktioniert, müssen wir die stepwise backward Prozedur von Hand durchführen. Mit `drop1` lässt sich die am wenigsten benötigte Variable eruieren, und mit `anova` lässt sich das reduzierte Modell mit dem aktuellen vergleichen. Wir führen diese Prozedur so lange durch, bis der Modellvergleich eine signifikante Verschlechterung des Modells anzeigt.
Das Schlussmodell sieht dann so aus:

`Hauptv` $\sim$ `Geschlecht + Schule + Ortsgroesse + Beeintr`

Nach dem AIC-Kriterium ist das das beste Modell. Der P-Wert von Geschlecht ist allerdings mit 0.059 noch knapp nicht signifikant.

```
t.rm <- multinom(Hauptv ~ Alter + Geschlecht + Schule + Wohnlage +
                 Ortsgroesse + Partei + Beeintr, data=t.d)
drop1(t.rm)
t.rm1 <- update(t.rm,~.-Partei)
drop1(t.rm1)
anova(t.rm1,t.rm)
t.rm2 <- update(t.rm1,~.-Wohnlage)
drop1(t.rm2)
anova(t.rm2,t.rm1)
t.rm3 <- update(t.rm2,~.-Alter)
drop1(t.rm3)
anova(t.rm3,t.rm2)

t.rm4 <- update(t.rm3,~.-Geschlecht)
drop1(t.rm4)
anova(t.rm4,t.rm3)
```

3. a) Der Zusammenhang zwischen der Zielvariablen `Time` und den erklärenden Variablen `Type1` und `Type2` ist linear.



Wenn wir das Modell

$$\texttt{Time}_i = \beta_0 + \beta_1\ \texttt{Type1}_i + \beta_2\ \texttt{Type2}_i + E_i$$

annehmen (ev. ohne β_0), bedeutet dies, dass die Total-Zeit für die Transaktionen additiv von den Anzahlen abhängt.

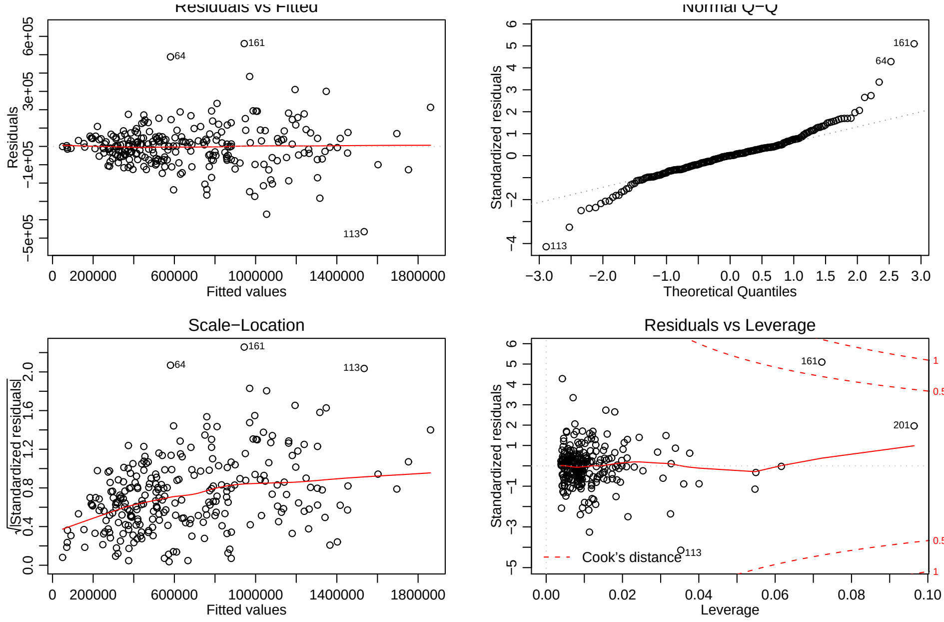
b) Wir passen das Modell

$$E\langle\texttt{Time}_i\rangle = \mu_i = \beta_0 + \beta_1\ \texttt{Type1}_i + \beta_2\ \texttt{Type2}_i$$

mit $\texttt{Time}_i \sim \mathcal{N}(\mu_i, \sigma^2)$ an.

```
Call:
lm(formula = Time ~ Type1 + Type2, data = d.trans)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.446e+04  1.705e+04   0.848    0.397
Type1       5.463e+00  4.332e-01  12.609   <2e-16 ***
Type2       2.034e+00  9.433e-02  21.567   <2e-16 ***
---
Residual standard error: 114200 on 258 degrees of freedom
Multiple R-Squared: 0.9091,    Adjusted R-squared: 0.9084
F-statistic:  1290 on 2 and 258 DF,  p-value: < 2.2e-16
```

Bemerkung: Der Koeffizient β_0 (`(Intercept)`) ist nicht signifikant.



Die Residuen-Plots zeigen, dass das Modell mit normalverteilten Fehlern nicht passt. Im Tukey-Anscombe Plot ist (am Anfang) ein deutliche Trichter zu erkennen, was der Glätter des Scale-Location Plots bestätigt. Der Normal Plot zeigt, dass die Fehler E_i nicht normalverteilt sind, sondern eine etwas langschwänzigere Verteilung haben (langschwänzig auf beiden Seiten).

c) Wir passen ein verallgemeinertes lineares Modell mit Gamma-verteilter Zielvariable und der Identität als Link-Funktion (vgl. mit Tabelle 12.3. in den R-Hinweisen) an:

$$E\langle \texttt{Time}_i \rangle = \mu_i = \beta_0 + \beta_1\ \texttt{Type1}_i + \beta_2\ \texttt{Type2}_i$$

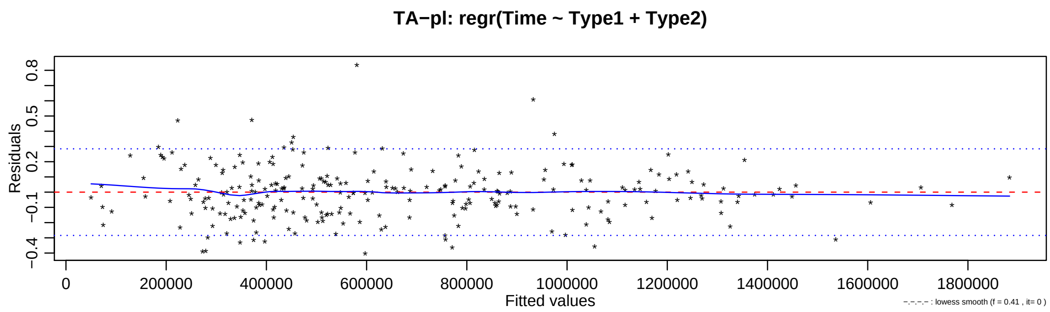
mit $\texttt{Time}_i$ Gamma$\langle \eta_i, \sigma \rangle$-verteilt.

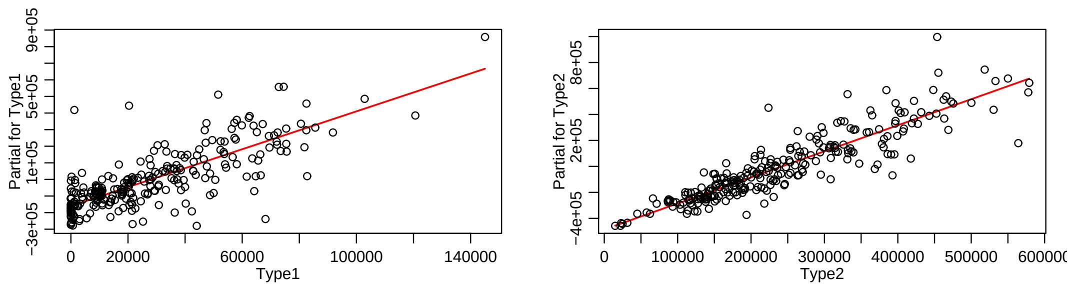
```
> r.g.regr <- regr(Time ~ Type1 + Type2, family=Gamma(link=identity),
                   data=d.trans)
> r.g.regr

Call:
regr(formula = Time ~ Type1 + Type2, data = d.trans,
     family = Gamma(link = identity))

Terms:
                    coef    stcoef    signif      R2.x df p.value
(Intercept) 15359.563935       0.0  1.505042        NA  1  0.0033
Type1           5.705443  146679.1  6.805411 0.3007046  1  0.0000
Type2           2.006855  236963.2 17.561669 0.3007046  1  0.0000

          deviance  df p.value
Model    85.123949   2       0
Residual  7.478001 258       1
Null     92.601950 260      NA

Family is Gamma.  Dispersion parameter estimated to be  0.02938966.
AIC: 6725.5
```

Bemerkung: Der Koeffizient β_0 (`(Intercept)`) ist hier signifikant, wie man mit einem `glm`-Output leicht nachprüfen kann.

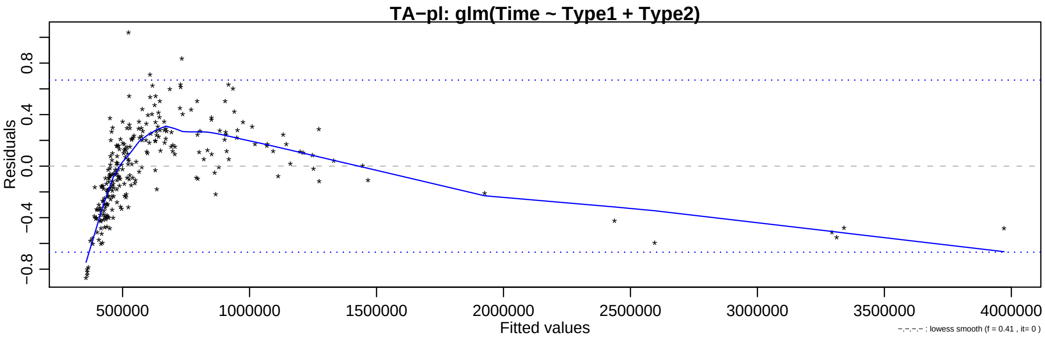
Residuenanalyse:

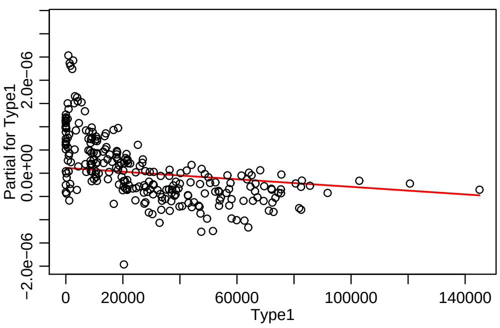
```
TA.plot(r.g.regr, labels = "*")
termplot(r.g.regr, partial=TRUE, col.res="black")
```

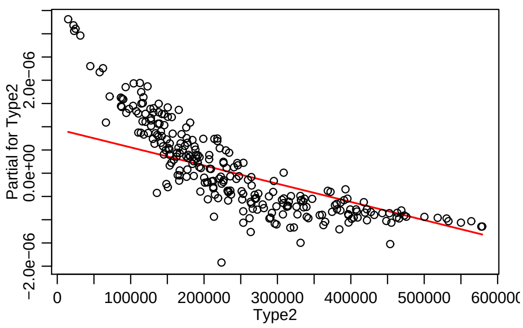




Der Tukey-Anscombe-Plot und die partiellen Residuenplots zeigen, dass dieses Modell recht gut passt. In den Plots sind jedoch Ausreisser nicht mit Sicherheit zu detektieren, da die Gamma-Verteilung viel langschwänziger als die Normalverteilung ist.
Im Vergleich zu diesem Modell mit dem Identitäts-Link ergeben sich mit den Residuen des Modelles mit der kanonischen Linkfunktion der Gammaverteilung folgende Plots, die offensichtlich machen, dass dieser Link unsinnig ist:







Übrigens ist auch das AIC des Modelles mit dem kanonischen Link (7126.6) grösser als jenes des Modelles mit dem Identitätslink (6725.5). Das wichtigste Argument gegen die Verwendung des kanonischen Links $\frac{1}{\mu}$ ist, dass durch diesen Link der additive Zusammenhang zwischen Zeit und Anzahl Transaktionen verloren geht.

Bemerkung:

Die Residuen-Plots (insbesondere der Tukey-Anscombe Plot) in Teilaufgabe b lassen erwarten, dass man mit einer Logarithmus- oder einer Wurzeltransformation der Zielvariablen ($\log\langle \texttt{Time}_i \rangle \sim \mathcal{N}(\mu_i, \sigma^2)$ oder $\sqrt{\texttt{Time}_i} \sim \mathcal{N}(\mu_i, \sigma^2)$) eine Verbesserung erreicht. Die Residuen-Plots für diese Modelle (nicht abgebildet) zeigen aber deutlich, dass keine Verbesserung, sondern eher eine Verschlechterung erreicht wird.
Zudem geht durch diese (nichtlineare) Transformation der Zielvariablen der additive Zusammenhang zwischen der Zeit und der Anzahl Transaktionen verloren.