Arbeitsblatt 3

Einfache lineare Regression (Diagnostik & Erweiterungen)

Aufgabe 1 (Windmühle)

Ein Entwicklungsingenieur untersucht den Einsatz von einer Windmühle zur Elektrizitätserzeugung. Dabei sammelte er unter anderem auch Daten über die Stromgewinnung (in Ampère) durch die Windmühle bei entsprechenden Windgeschwindigkeiten (in Metern pro Sekunde). Die Daten sind im File `Windmuehle.dat`.

Einlesen der Daten

```
windmill <- read.table("data/Windmuehle.dat", header=TRUE)
```

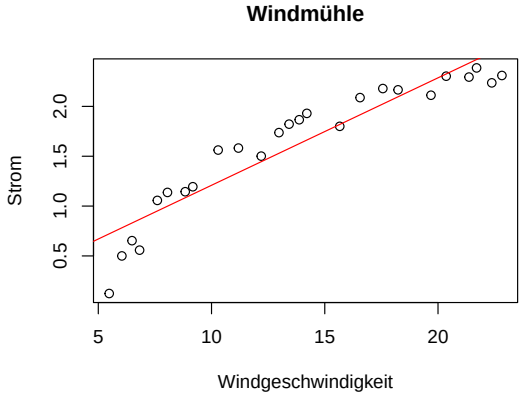
(a)

Wir betrachten zuerst das Modell

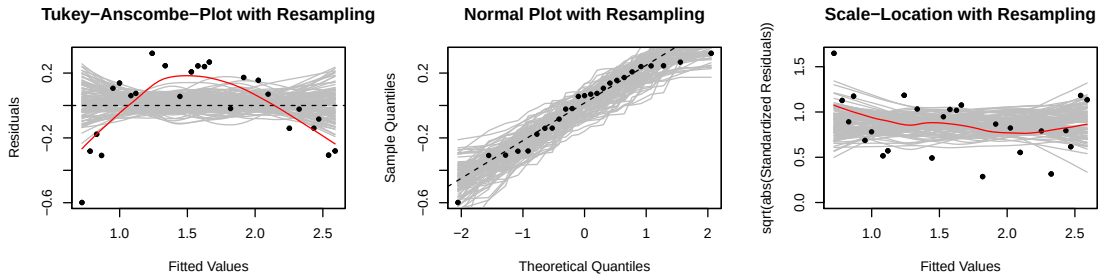
$$\text{Strom}_i = \alpha + \beta \cdot \text{Windgeschwindigkeit}_i + E_i$$

Passen Sie ein entsprechendes Regressionsmodell an und prüfen Sie mittels grafischer Residuenanalyse, ob die Voraussetzungen erfüllt sind.

```
## Anpassen des Regressionsmodells
windmill.lm1 <- lm(Strom ~ Windgeschwindigkeit, data = windmill)
par(mfrow = c(1,1))
plot(Strom ~ Windgeschwindigkeit, data = windmill, main = "Windmühle")
abline(windmill.lm1, col = "red")
```




```
## Residuen-Analyse:
load("data/resplot.rda")
par(mfrow=c(1,3))
resplot(windmill.lm1, plots = 1:3)
```



- **Tukey-Anscombe-Plot:** Die Glätter zeigt eine systematische Abweichung von der Horizontalen, d.h. der Erwartungswert ist *nicht* konstant 0 ist.
- **Normalplot**: Datenpunkte streuen recht gut um eine Gerade. Es gibt keine Hinweise, dass die Normalverteilungsannahme verletzt ist.
- **Location-Scale-Plot**: Der Glätter ist einigermassen horizontal. Die Varianz der kann als konstant angenommen werden.
- Da zeitliche Reihenfolge der Messungen unbekannt ist, daher ist es sinnlos, sich über zeitliche Korrelationen (d.h. Verletzung der Unabhängigkeit der Residuen) den Kopf zu zerbrechen.

FAZIT: Die Anpassung ist ungenügend, da systematische Abweichungen im Erwartungswert vorkommen.

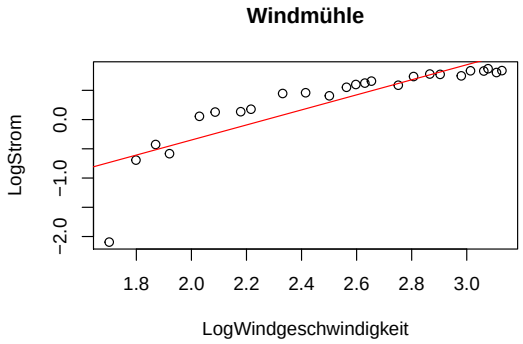
(b)

Wenden Sie die First-Aid Transformationen auf die erklärende Variable und die Zielvariable an, passen Sie das Regressionsmodell mit den transformierten Grössen an und beurteilen Sie grafisch, ob die Vorraussetzungen nun erfüllt sind.

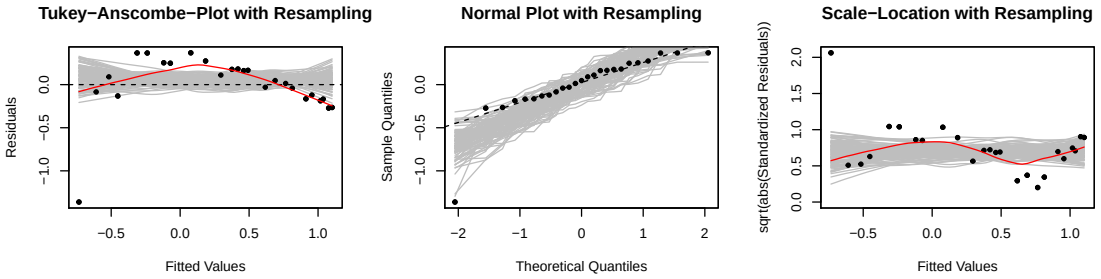
Gemäss First Aid Transformationen werden beide Variablen logarithmiert.

```
# First-Aid-Transformation
windmill$LogStrom <- log(windmill$Strom)
windmill$LogWindgeschwindigkeit <- log(windmill$Windgeschwindigkeit)

# Anpassen der Regression
windmill.lm2 <- lm(LogStrom ~ LogWindgeschwindigkeit, data=windmill)
par(mfrow = c(1,1))
plot(LogStrom ~ LogWindgeschwindigkeit, data=windmill, main = "Windmühle")
abline(windmill.lm2, col = "red")
```




```
## Residuen-Analyse:
par(mfrow=c(1,3))
resplot(windmill.lm2, plots=1:3)
```



- **Tukey-Anscombe-Plot:** Der Glätter zeigt immer noch eine "Bananen"struktur, d.h. der Erwartungswert ist nicht konstant 0 ist. Des Weiteren gibt es nun einen Ausreisser (Beobachtung 25).
- **Normalplot**: Datenpunkte liegen, abgesehen vom Ausreisser, recht gut auf einer Gerade.
- **Location-Scale-Plot**: Der Glätter verläuft wellenförmig. Es gibt Varianzschwankungen.

FAZIT: Anpassung ist ungenügend, da systematische Abweichungen im Erwartungswert vorkommen, die Varianz nicht konstant ist und ein Ausreisser sichtbar ist. Die First-Aid-Transformationen führen meist, aber eben nicht immer zu einem guten Modell.

(c)

Aus der Theorie ist folgender funktionaler Zusammenhang bekannt:

$$\text{Strom} \approx \alpha + \beta \cdot \frac{1}{\text{Windgeschwindigkeit}}$$

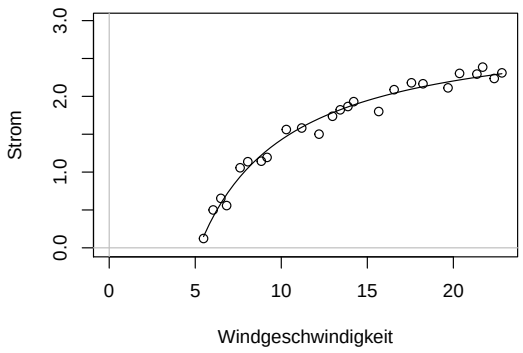
Passen Sie dieses Modell aus der Fachtheorie mittels einer Regression an (d.h. `Strom` als Zielgrösse und $x = 1/$`Windgeschwindigkeit` als erklärende Variable) und stellen Sie das Regressionsmodell in einem Streudiagramm dar.

```
# Transformierte Variable
windmill$x <- 1/windmill$Windgeschwindigkeit
# Regression
windmill.lm <- lm(Strom ~ x, data=windmill)
```

```
# Geschätzte Koeffizienten
coef(windmill.lm)
```

```
(Intercept)           x
    2.97886   -15.51546
```

```
# Grafische Darstellung
plot(Strom ~ Windgeschwindigkeit, data=windmill,
      ylim = c(0, coef(windmill.lm)[1]),
      xlim = c(0, max(windmill$Windgeschwindigkeit)))
abline(v = 0, h = 0, col = "grey") # x-/y-Achse
# Regressionsgerade
x <- seq(min(windmill$Windgeschwindigkeit),
         max(windmill$Windgeschwindigkeit), length = 50)
lines(x, coef(windmill.lm)[1] + coef(windmill.lm)[2] * (1/x))
```



(d)

Was bedeuten die beiden Parameter α und β im Modell aus der Fachtheorie?

Durch die Transformation der erklärenden Variable `Windgeschwindigkeit` verändert sich die Bedeutung von α und β:

1. α entspricht der maximalen Stromproduktion.
2. $((-1) \cdot \beta/\alpha)$ entspricht das Minimum an Wind, so dass überhaupt Strom produziert wird.

```
# Maximale Stromproduktion
coef(windmill.lm)[1]
```

```
(Intercept)
    2.97886
```

```
## Wie viel Wind braucht es, damit überhaupt Strom produziert wird?
-coef(windmill.lm)[2]/coef(windmill.lm)[1]
```

```
       x
5.208521
```

Herleitungen:

1. Die maximale Stromproduktion einer Windmühle wird bei sehr grossen Windgeschwindigkeiten erreicht. Wenn man die Windgeschwindigkeit in der Regressionsgleichung ganz gross werden lässt, dann wird die transformierte erklärende Variable $\texttt{x} = 1/\texttt{Windgeschwindigkeit}$ sehr klein und damit ist $\texttt{Strom} \approx \alpha$. Oder mathematisch:

$$\lim_{\text{Windgeschwindigkeit}\to\infty} \left[\alpha + \beta \cdot \left(\frac{1}{\text{Windgeschwindigkeit}}\right)\right] = \alpha$$

2. Wir suchen den Punkt, an welchen die Regressionskurve die x-Achse schneidet, also Strom $= 0$. An diesem Punkt gilt:

$$0 = \alpha + \beta \cdot \left(\frac{1}{\texttt{Windgeschwindigkeit}}\right)$$

   Diese Gleichung wird nun nach `Windgeschwindigkeit` aufgelöst. Dazu wird auf beiden Seiten der Gleichung α subtrahiert und anschliessend durch β geteilt. Dies ergibt

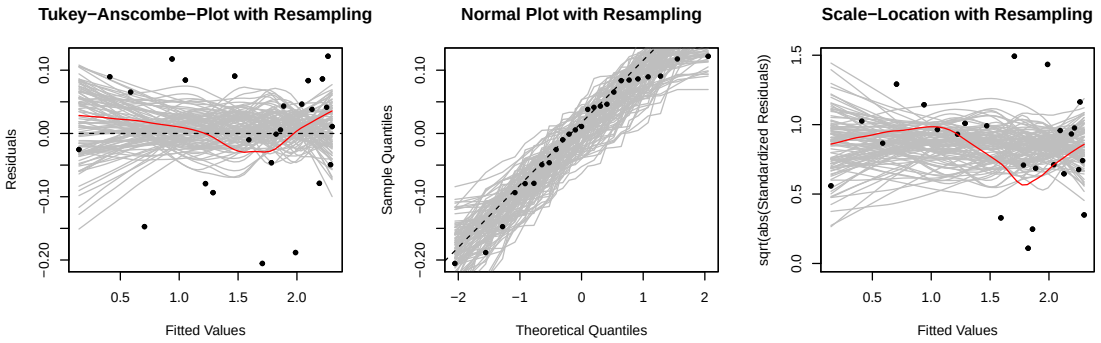
$$-\alpha/\beta = \left(\frac{1}{\texttt{Windgeschwindigkeit}}\right)$$

   Zuletzt nehmen wir die Inverse auf beiden Seiten und erhalten: Windgeschwindigkeit $= (-\beta/\alpha)$. Ab einer Windgeschwindigkeit von $(-\beta/\alpha)$ wird also Strom produziert.

(e)

Prüfen Sie mittels Residuenanalyse, ob die Voraussetzung für das Modell aus der Fachtheorie erfüllt sind.

```
## Residuen-Analyse:
par(mfrow=c(1,3))
resplot(windmill.lm, plots=1:3)
```



- **Tukey-Anscombe-Plot**: Die Punkte sind viel gleichmässiger um die horizontale Nulllinie gestreut. Glätter zeigt zwar noch eine ganz leichte Bananenform. Diese ist aber vernachlässigbar.
- **Normalplot**: Datenpunkte streuen bis auf das rechte Ende gut um eine Gerade. Das rechte Ende ist ein Hinweis auf Kurzschwänzigkeit, welche aber für die Regressionsergebnisse ungefährlich ist.
- **Scale-Location-Diagramm**: Der Glätter zeigt eine kleine Delle. Da die Punkte aber sehr weit verstreut liegen, befindet sich diese Delle innerhalb der stochastischen Fluktuation.

FAZIT: Die Anpassung ist ok.

(f)

Gibt es einen signifikanten Zusammenhang zwischen Windgeschwindigkeit und der Stromgewinnung aufgrund des Modell aus der Fachtheorie? Führen Sie einen geeigneten Test auf dem 5% Niveau durch.

Die Frage kann mit einem t-Test zur Nullhypothese: $H_0 : \beta = 0, \quad H_A : \beta \neq 0$ beantwortet werden. Aus dem `summary()`-Output erhält man:

```
     Estimate    Std. Error       t value      Pr(>|t|)
-1.551546e+01  4.618775e-01 -3.359215e+01  4.742557e-21
```

Der t-Test verwirft die Nullhypothese, da der p-Wert < 0.05. Es gibt somit einen auf 5% signifikanten Einfluss von Windgeschwindigkeit auf die Stromgewinnung

(g)

Was sind plausible Werte für die maximale Stromgewinnung der Windmühle? Geben Sie ein 95% Vertrauensintervall an.

Der Parameter α (Achsenabschnitt) beschreibt die maximale Stromproduktion. Gesucht ist also das 95% Vertrauensintervall für den Achsenabschnitt.

```
confint(windmill.lm, parm = 1, level = 0.95)
```

```
               2.5 %   97.5 %
(Intercept) 2.885973 3.071748
```

Die maximale Stromproduktion der Windmühle liegt mit 95% Wahrscheinlichkeit zwischen 2.89 und 3.07 Ampère.

(h)

Machen Sie eine Vorhersage, wie gross die erwartete Stromproduktion bei einer Windgeschwindigkeit von 15 m/s ist. Geben Sie ein 95%-Prognoseintervall an.

```
new.x <- data.frame(x = 1/15)
predict(windmill.lm, newdata = new.x, interval = "prediction")
```

```
       fit      lwr     upr
1 1.944496 1.744763 2.14423
```

Die erwartete Stromgewinnung beträgt 1.94 Ampère und mit 95% Wahrscheinlichkeit wird ein Strom zwischen 1.74 Ampère und 2.14 Ampère erzielt.

Aufgabe 2 - Highway in Texas

Auf der Strasseoberfläche akkumulieren sich in trockenen Zeiten sehr viele Schmutzstoffe. Diese werden während dem Regen zum grössten Teil abgewaschen. In Austin (Texas) wurde an einer bestimmten Stelle am Highway die Abflussmenge bei Regen bestimmt. Die Forscher wollen nun den Zusammenhang zwischen dem Regenfallvolumen und der Abflussmenge bestimmen. Daraus können sie dann vorhersagen, wie gross die Menge an Schadstoffen ist, die in die Natur gelangt. Die gemessenen Daten befinden sich im File `highway.csv`.

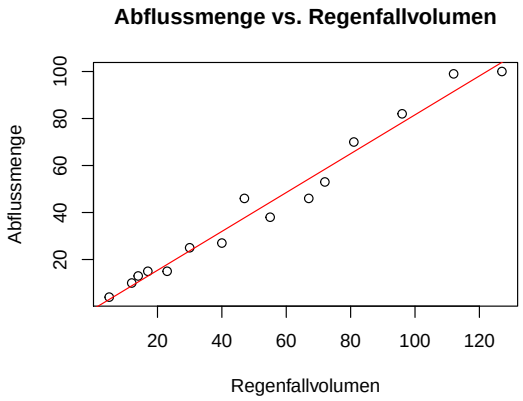
Einlesen der Daten

```
highway <- read.csv("data/highway.csv", header=TRUE)
```

(a)

Passen Sie eine Regressionsgerade (`runoff` als Zielgrösse und `rain` als erklärende Grösse). Visualisieren Sie die geschätzte Gerade.

```
# Anpassen der KQ-Geraden
fit.rain <- lm(runoff ~ rain, data = highway)
# Scatterplot
plot(runoff ~ rain, data = highway, main = "Abflussmenge vs. Regenfallvolumen",
     xlab = "Regenfallvolumen", ylab = "Abflussmenge")
abline(fit.rain, col="red")
```



(b)

Welcher Anteil der beobachteten Variation in der Abflussmenge kann mit dem einfachen linearen Regressionsmodell erklärt werden?

Hier wird nach dem Multiple R-Squared gefragt. Wie man dem Summary-Output entnehmen kann, beträgt der erklärte Anteil 97.53%.

(c)

Besteht ein auf 5% signifikanter linearer Zusammenhang zwischen Abflussmenge und Regenfallvolumen? Geben Sie auch die anschauliche Interpretation des Regressionskoeffizienten an.

```
summary(fit.rain)
```

```
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -1.1283048 2.36778251 -0.4765238 6.416111e-01
rain         0.8269731 0.03652408 22.6418585 7.896130e-12
```

Ja, der Zusammenhang ist hochsignifikant. Der Test für die Nullhypothese $H_0 : \beta = 0$ wird mit einem sehr kleinen p-Wert verworfen (p-Wert < 0.05). Es ist also statistisch gesichert, dass ein linearer Zusammehang zwischen Abflussmenge und Regenfallvolumen existiert.

Als Schätzungen für den Achsenabschnitt erhalten wir $\hat{\alpha} = -1.12$, d.h. wenn es nicht regnet, gibt es eine negative Abflussmenge. Dies kann natürlich nicht sein. Da wir keine Beobachtungen an der

Stelle $x = 0$ haben, ist die Interpretation des Achsenabschnitts eine Extrapolation und somit mit Vorsicht zu betrachten. Immerhin ist die Schätzung des Achsenabschnitts auf 5% nicht signifikant ausgefallen (p-Wert > 0.05).

Die Steigung wird mit $\hat{\beta} = 0.83$ geschätzt. Der Wert ist kleiner als 1 (sogar signifikant kleiner, da 1 nicht im das 95%-Vertrauensintervall für β).

```
confint(fit.rain, parm = 2, level = 0.95)
```

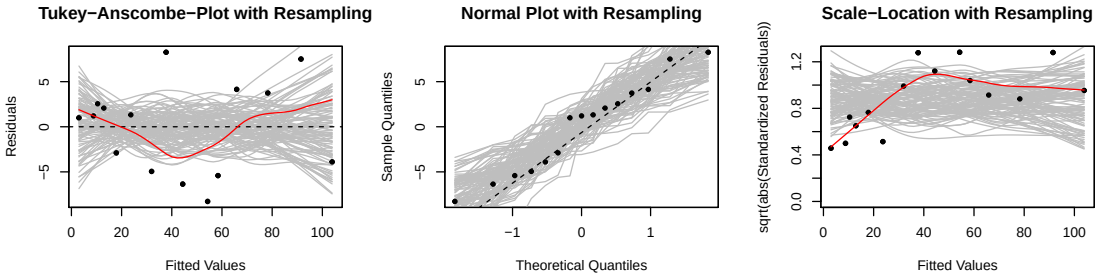
```
         2.5 %    97.5 %
rain 0.7480677 0.9058786
```

Somit ist statistisch gesichert, dass nicht der ganze Regen über die Kanalisation abfliesst. Dies ist durchaus plausibel, da ein Teil verdunstet, ein anderer liegen bleibt oder abseits der Kanalisation abfliesst.

(d)

Überprüfen Sie die für das Regressionsmodell getroffenen Annahmen. Können wir Ihren Schlussfolgerungen aus (c) vertrauen?

```
par(mfrow = c(1,3))
resplot(fit.rain, plots = 1:3)
```



- **Tukey-Anscombe-Plot:** Trotz hohem multiplem R-squared ist der Erwartungswert der Fehler nicht gleich null. Der Glätter weicht auf jeden Fall systematisch von der Horizontalen ab.
- **Normalplot**: Die Normalverteilung der Fehler passt recht gut.
- **Location-Scale-Plot**: Die Streuung der Fehler nimmt mit dem Regenfallvolumen zu.
- **Unabhängigkeit**: In den Daten gibt es keine Angaben über die zeitliche Verteilung der Messpunkte. Wir müssen hier einfach davon ausgehen, dass die Forscher die Messungen unabhängig voneinander durchgeführt haben.

FAZIT: Die Annahmen sind verletzt. Den Aussagen aus (c) kann man nicht hunderprozentig vertrauen.

(e)

Warum könnte hier eine Logarithmus Transformation der beiden Variablen weiterhelfen?

Die Abflussmenge und das Regenfallvolumen sind beide leicht rechtsschief verteilt. Weiter können beide Grössen nur positive Werte annehmen. Den First-Aid Transformationen zufolge könnte Logarithmus hier weiterhelfen.

(f)

Passen Sie eine Log-Log Regression an. Zeichnen Sie die neue Regressionsbeziehung in ein Streudiagramm auf der Originalskala ein. Prüfen Sie erneut die Stärke des linearen Zusammenhangs und die Signifikanz zwischen Abflussmenge und Regenfallvolumen auf dem 5% Niveau.

```
# Anpassen der Log-Log Regression
highway$LogRain <- log(highway$rain)
highway$LogRunoff <- log(highway$runoff)
fit.rain2 <- lm(LogRunoff ~ LogRain, data = highway)
summary(fit.rain2)
```

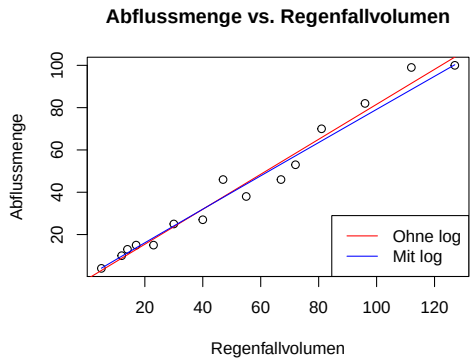
```
Call:
lm(formula = LogRunoff ~ LogRain, data = highway)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20980 -0.10952  0.02828  0.08727  0.20388

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.18369    0.13803  -1.331    0.206
LogRain      0.98917    0.03676  26.908 8.75e-13 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1293 on 13 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.981
F-statistic:   724 on 1 and 13 DF,  p-value: 8.748e-13
```

```
# Streudiagramm
plot(runoff ~ rain, data = highway, main = "Abflussmenge vs. Regenfallvolumen",
     xlab = "Regenfallvolumen", ylab = "Abflussmenge")
abline(fit.rain, col="red")
x <- seq(min(highway$rain), max(highway$rain), by = 1)
lines(x, exp( coef(fit.rain2)[1] + coef(fit.rain2)[2] * log(x)), col = "blue")
legend("bottomright", c("Ohne log", "Mit log"),
       lty = 1, col = c("red","blue"))
```



Der Zusammenhang war schon ohne die Transformation sehr stark und ist wieder hochsignifikant, d.h. p-Wert ist sehr viel als 0.05. Das Bestimmtheitsmass ist nach der Transformation sogar noch etwas besser: $R^2 = 0.982$. Hinzu kommt, dass das Modell mit Transformation auch aus praktischen Gründen zu bevorzugen ist. Nun kann das Modell keine unsinnigen, negativen Abflusswerte erzeugen.

(g)

Berechnen Sie eine Vorhersage, wie gross die erwartete Abflussmenge bei einer Regenfallmenge von 50 ist. Erzeugen Sie ein 95%-Prognoseintervall für beliebige Regenfallmengen und zeichnen dieses als sogenanntes Prognoseband in das Streudiagramm ein.

```
x0 <- data.frame(LogRain = log(50))
pred.y <- predict(fit.rain2, newdata = x0, interval = "prediction")
exp(pred.y)
```

```
       fit      lwr      upr
1 39.88372 29.86744 53.25903
```

Zu beachten ist, dass die Punkt-Vorhersage nicht der Erwartungswert der Zielgrösse ist, sondern nur deren Median. Wenn wir den erwarteten Wert vorhersagen wollen, so müssen wir noch speziell rücktransformieren.

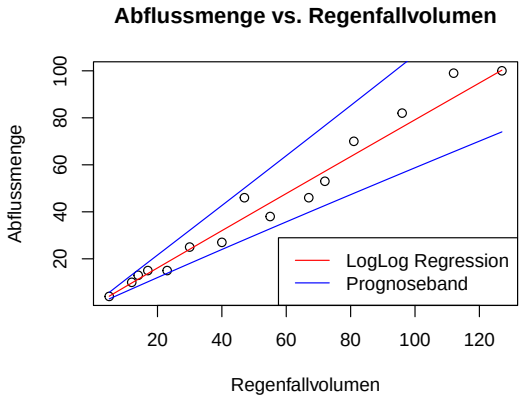
```
exp(pred.y[1] + (summary(fit.rain2)$sigma^2)/2)
```

```
[1] 40.21832
```

Die erwartete Abflussmenge bei einer Regenfallmenge von 50 beträgt 40.22. Mit 95% liegt diese zwischen 29.87 und 53.26.

Berechnung und Visualisierung des Prognosebandes:

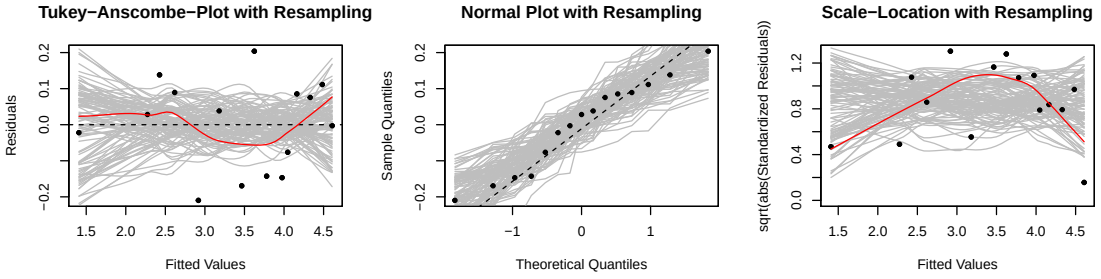
```
plot(runoff ~ rain, data = highway, main = "Abflussmenge vs. Regenfallvolumen",
     xlab = "Regenfallvolumen", ylab = "Abflussmenge")
# Hinzufügen der Log-Log Regression
x <- seq(min(highway$rain), max(highway$rain), by = 1)
h <- predict(fit.rain2, newdata = data.frame(LogRain = log(x)),interval = "prediction")
lines(x, exp(h[,"fit"]), col = "red")
# Hinzufügen des Prognosebands
lines(x, exp(h[,"lwr"]), col = "blue")
lines(x, exp(h[,"upr"]), col = "blue")
# Hinzufügen einer Legende
legend("bottomright", c("LogLog Regression", "Prognoseband"),
       lty = 1, col = c("red","blue"))
```



(i)

Führen Sie eine Residuenanalyse für das log-log Modell durch. Beantworten Sie anschliessend, welches der beiden Modelle Ihnen besser geeignet scheint, und begründen Sie Ihre Antwort.

```
par(mfrow = c(1,3))
resplot(fit.rain2, plots = 1:3)
```



- **Tukey-Anscombe-Plot:** Der Plot ähnelt sehr demjenigen ohne Transformation. Der Glätter weicht auch hier systematisch von der Horizontalen ab.
- **Normalplot**: Die Normalverteilung der Fehler scheint ziemlich gut zu passen.
- **Location-Scale-Plot**: Der Glätter weicht systematisch von der horizontalen ab.

FAZIT: Die Annahmen sind verletzt.

Zusammenfassend kann man sagen, dass zwischen den beiden Modellen nur relativ geringe Unterschiede in Bezug auf die Kenngrössen und die Residuenplots bestehen.

Das Modell mit den Transformationen ist aber zu bevorzugen. Die Residuenplots sind hier allerdings nicht der Ausschlag. Entscheidend ist die Tatsache, dass das Modell mit den log-Transformationen keine negativen Werte erzeugen kann – weder bei den angepassten Werten noch in Form des Prognoseintervalls. Zudem sagt es für keinen Regen (`rain` $= 0$) auch keinen Abfluss (`runoff` $= 0$) vorher, ebenfalls eine erwünschte Eigenschaft. Auch hat β eine Interpretation, welche für die Praxis geeigneter scheint. Das Modell mit den Transformationen ist also sachlich korrekter. Es sei aber nochmals betont, dass es keinesfalls bezüglich der Residuen besser ist.