Aufgabe 05

Jonas Benischek

2022-11-24

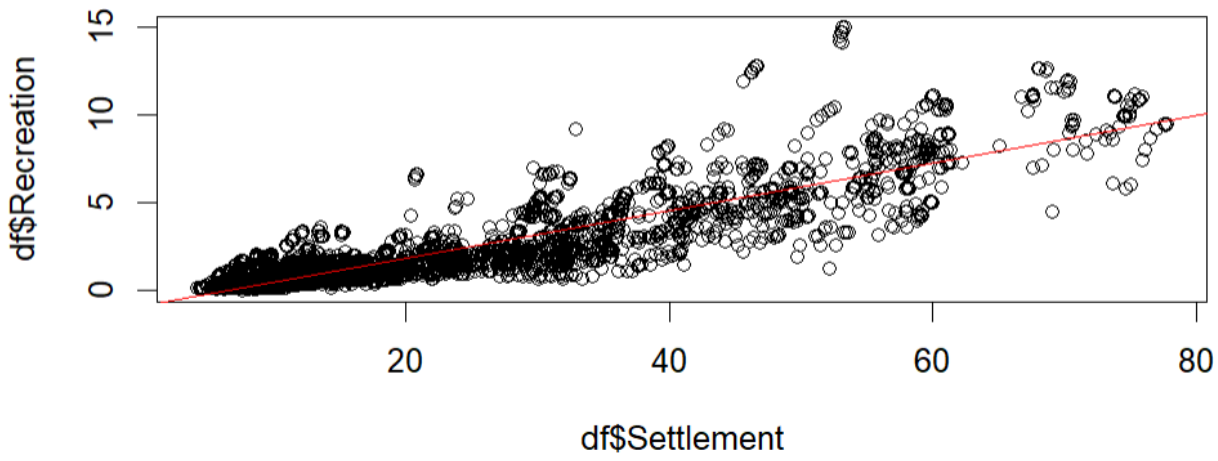
Erholung vs. Siedlung

Lineare Modellierung und Normalverteilung der Residuen

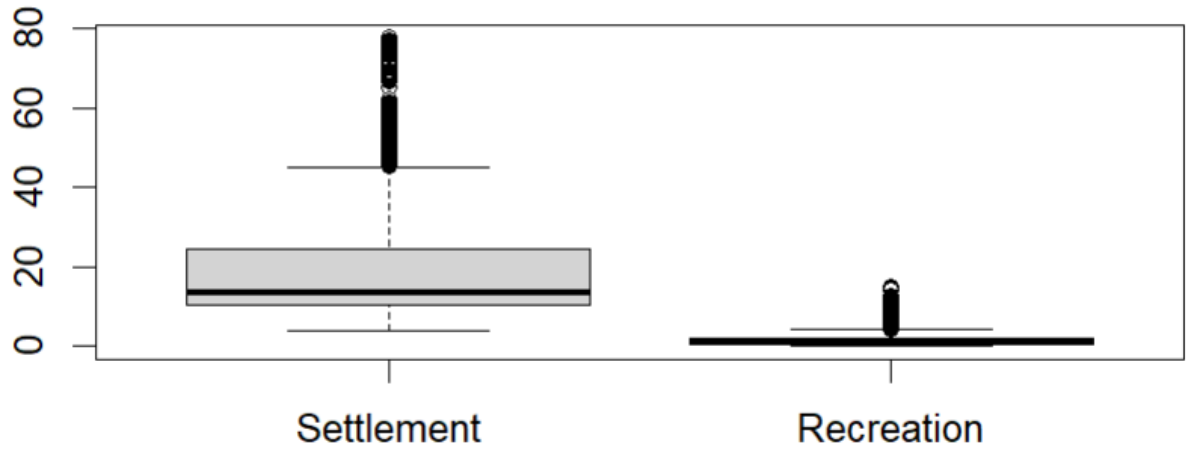
```
#-------------------------------------------------------------------------------
# Unit:5
# Abgabe: 5
# Verfasser:Jonas Benischek
# Beschreibung:Lineare Regressionsmodell berechnen
#-------------------------------------------------------------------------------

# MANDANTORY: Definieren des Stammordners Ändern Sie diese Zeile NICHT
rootDIR = "C:/Datenanalyse/UsingR/Unit5/"

#############################################################################
# Erholung vs. Siedlung
# Link: https://geomoer.github.io/moer-mpg-data-analysis/unit05/unit05-04_assignment.html
#
#########################################################
# Lineare Modellierung und Normalverteilung der Residuen
#############################################################################################
# Erholung vs. Siedlung
# Dieses Arbeitsblatt geht der Frage nach, wie der prozentuale Anteil der
# Siedlungsfläche zum Anteil der Erholungsfläche in jeder Gemeinde zusammenhängt.
#
```

```
#############################################################################
# 1. Bitte schreiben Sie ein R-Markdown-Skript, das ein lineares
# Regressionsmodell berechnet, das die prozentuale Erholungsfläche
# (y = abhängige Variable) mit dem Siedlungsgebiet in Beziehung setzt,
# und visualisieren Sie die Beziehung als Streudiagramm mit dem linearen Modell,
# das im selben Diagramm gezeichnet ist.
#############################################################################

#############################################################################################
#Datensatz Laden

df <- readRDS("lu_clean.rds")

str(df) #Ansicht in der Console

#############################################################################
#Datensatz Übersicht

summary(df)

#  Settlement               Recreation
# Median :13.40           Median :0.900
# Max.   :77.80           Max.   :15.00
```

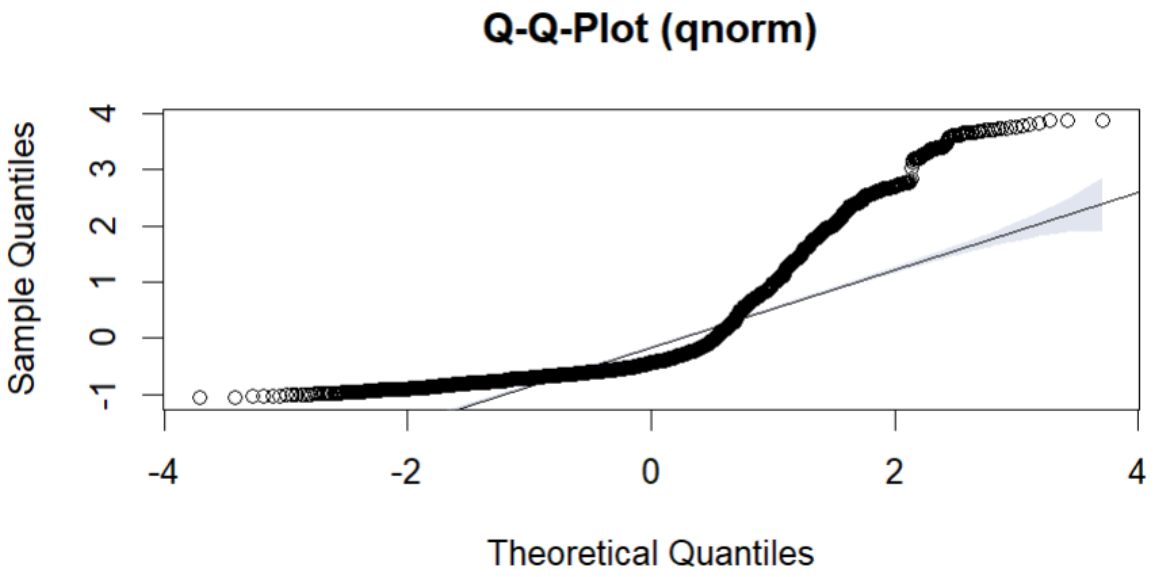
```r
################################################
#visualisieren Sie die Beziehung als Streudiagramm

#recreation Erholung/ prozentual (y = abhängige Variable)

#settlement Siedlungsgebiet (x = unabhängige Variable)

plot(df$Settlement, df$Recreation)

########################################
# Ein lineares Regressionsmodell berechnen

plot(df$Settlement, df$Recreation)
model <- lm(Recreation~Settlement, data = df)

abline(model, col="red")

summary(model)
```




```r
#############################################################################
# 2. Zusätzlich berechnen Sie eine Visualisierung, die die Beurteilung der
# Heteroskedastizität und der Normalverteilung der Residuen ermöglicht.
#############################################################################

print(df) #Zeilen 6 und 7  müssen ausgewählt werden

boxplot(df[, 6:7]) #Visualisierung der Verteilung im Boxplot


# Heteroskedastizität
model1 <- lm(Recreation~Settlement, data = df)
model2 <- lm(Recreation~Settlement + Agriculture + Forest, data = df)
summary(model)

install.packages("DescTools")

library(DescTools)
```

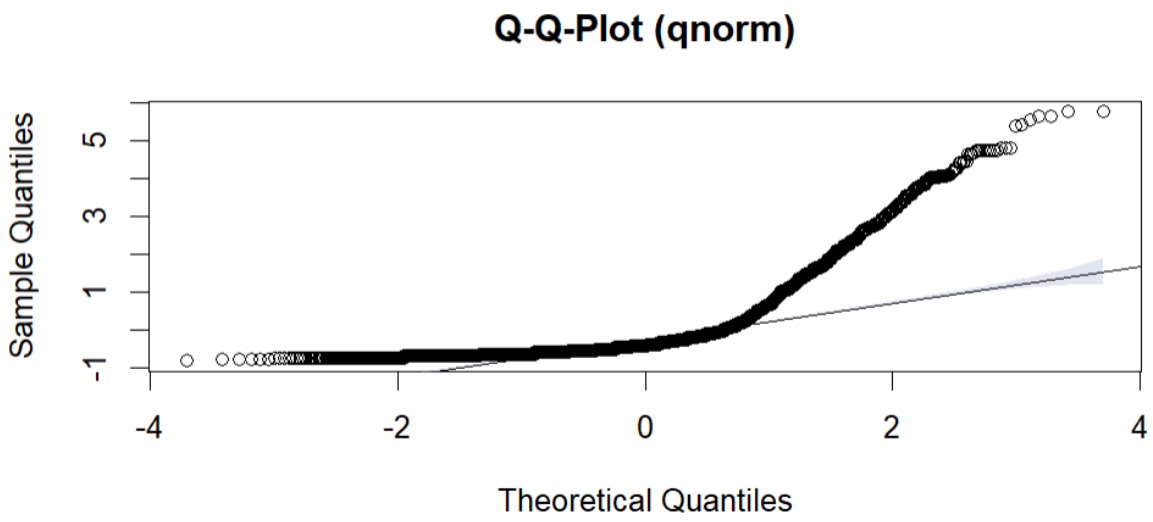



```
#########################
#Q-Q Diagramm / Settlement

df$zSettlement <- scale(df$Settlement)

qqnorm(df$zSettlement)
qqline(df$zSettlement)

PlotQQ(df$zSettlement)#Konfidenzintervall
```



Jonas/2022-11-28

```
#########################################################################
#Q-Q Diagramm Recreation

df$zRecreation <- scale(df$Recreation)

qqnorm(df$zRecreation)
qqline(df$zRecreation)

PlotQQ(df$zRecreation)#Konfidenzintervall
```



Jonas/2022-11-28

```
#############################################################################
# 3. Bitte schreiben Sie genau einen Satz, der die Kernaussage sowohl des
# Heteroskedastizitäts- als auch des Normalverteilungsdiagramms zusammenfasst.
#############################################################################

#Settlement: Die Residuen liegen fast immer ausserhalb von dem
# Konfidenzintervall daher nicht Normalverteilt

#Recreation: Die Residuen liegen fast immer ausserhalb von dem
# Konfidenzintervall daher nicht Normalverteilt

# Es zeigen sich große Abweichungen der Verteilung im Q-Q-Plot mit
# Konfidenzintervall.

#############################################################################
# 4. Nachdem Sie Ihre Daten beschrieben haben, schauen Sie sich bitte die
# Normalverteilungsbewertung genauer an.
#############################################################################
# 4.a) Bewerten Sie dazu bitte, wie oft ein Normalverteilungstest für die Residuen
# seine Nullhypothese ablehnen würde, wenn ein Regressionsmodell nicht für den
# gesamten Datensatz berechnet wird, sondern 100 Regressionsmodelle für
# 100 Teilstichproben des Datensatzes.Jede der 100 Teilstichproben sollte 50
# zufällig ausgewählte Wertepaare aus dem gesamten Datensatz enthalten.
#############################################################################
```

```r
lu_rs <- df[,6:7]
head(lu_rs)


lu_sample <- list() # Liste für die sub samples definieren
vector <- vector() # Definition eines Vektors für die Interpretation der Ergebnisse der Normalitätsbewertung
p_values <- vector() #  Definieren eines Vektors für den p-values

for(i in seq(1:100)){
  lu_sample[[i]] <- lu_rs[sample(nrow(lu_rs), 50),] # Erstellen einer Unterstichprobe mit 50 Wertepaaren und Speichern in der Liste
  lu_dep <- lu_sample[[i]][[1]] # Definition abhängiger Variable
  lu_ind <- lu_sample[[i]][[2]] # Definition unabhängige Variable
  lu_lmod <- lm(lu_dep ~ lu_ind) # Berechnung des linearen Regressionsmodells der Teilstichprobe
  shap <- shapiro.test(lu_lmod$residuals) # Shapiro-Wilk-Test
  if (shap$p.value > 0.05) { # Schreiben der Interpretationsergebnisse der p-Werte in den ersten Vektor
    vector <- append(vector,"normality")
  } else
  {vector <- append(vector,"no normality")
  }
  p_values <- append(p_values, shap$p.value) # p-values Test in dem zweiten vector
}

#####################################################
# Überprüfung der Anzahl der Nullhypothesenbestätigungen
print(vector)
```

```r
##################################
# no normality
sum(vector == "no normality")

##################################
# max. p_values
max(p_values)


#############################################################################################
# 4. b) Geben Sie das Ergebnis an oder visualisieren Sie es und diskutieren
# Sie es in Bezug auf das Probendesign und die Zuverlässigkeit der
# Testergebnisse (maximal drei Sätze).
#############################################################################################


breaks = c(0, 0.05, 0.1, 0.15, 0.20, 0.25, 0.3, 0.35, 0.4, 0.45, 0.50, 0.55, 0.6, 0.65, 0.7)
hist(p_values , breaks = breaks)


# Es zeigt sich sehr deutlich, dass die Residuen von mehr als 80 der 100
# Teilbeimodelle keine Normalverteilung aufzeigen.
```