Prüfung

CAS Datenanalyse Modul B1

Stefan Schmidt

27.05.2020

Aufgabe 1

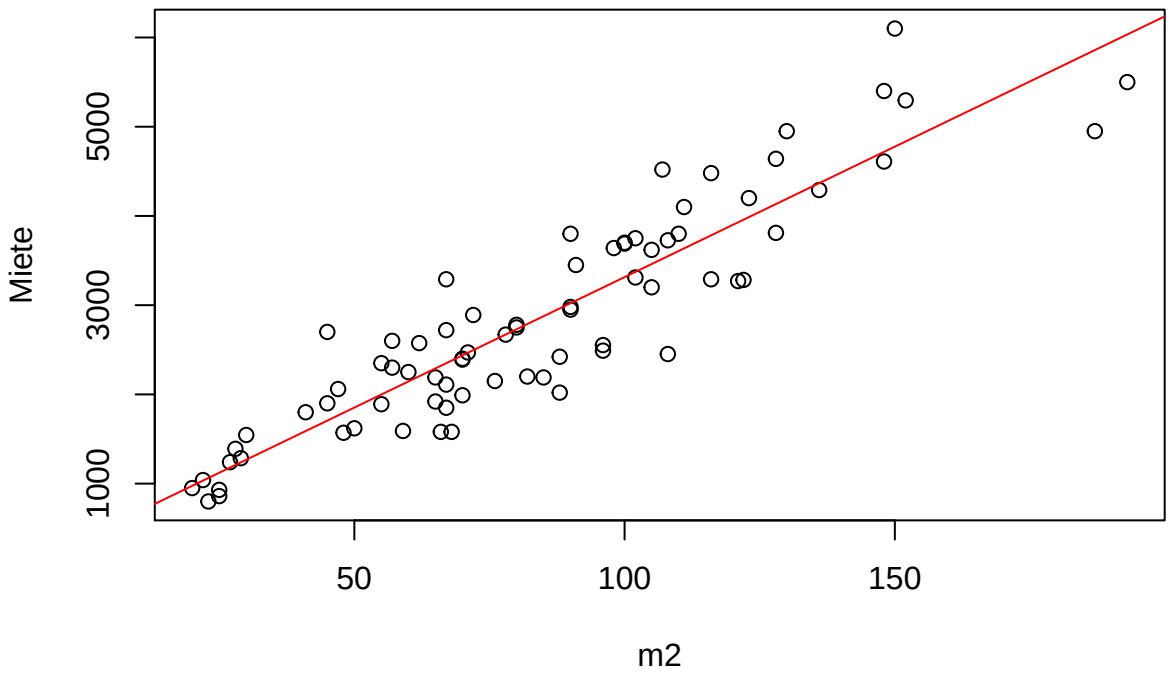
(a)

```
# Modell
fit.zh <- lm(Miete ~ m2, data = zueri)

# Koeffizienten
coef(fit.zh)
```

```
## (Intercept)          m2
##   390.08631    29.24366
```

```
# Plot
par(mfrow = c(1, 1))
plot(Miete ~ m2, data = zueri)
abline(fit.zh, col = 'red')
```



Der geschätzte Achsenabschnitt ist 390.09 und die geschätzte Steigung ist 29.24. Dies bedeutet, dass eine Wohnung mit 0 m2 CHF 390 kosten wuerde und pro hinzukommendem Quadratmeter mit einer zusaetzlichen

Miete von CHF 29.24 zu rechnen ist. Die Miete fuer eine 0 m2 Wohnung ist unsinnig und haengt damit zusammen, dass auf diesem Mietpreis-Niveau keine Wohnungen angeboten werden und keine Daten vorhanden sind.

(b)

```
summary(fit.zh)
```

```
##
## Call:
## lm(formula = Miete ~ m2, data = zueri)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1096.40  -335.19     1.26   370.89  1323.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  390.086    140.148   2.783  0.00682 **
## m2            29.244      1.527  19.157  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 502.9 on 74 degrees of freedom
## Multiple R-squared:  0.8322, Adjusted R-squared:  0.8299
## F-statistic:   367 on 1 and 74 DF,  p-value: < 2.2e-16
```

Ja, der generelle F-Test (letzte Zeile) zeigt dass grundsaetzlich ($p < 0.05$), dass ein Zusammenhang zwischen mindestens einer erklaerenden Variablen und er Zielvariablen besteht.

Im t-Test ist der p-Wert fuer m2 mit 2e-16 auch kleiner als 0.05. Somit ist die Nullhypothese (beta fuer m2 = 0) abzulehnen. Es besteht also ein signifikanter Zusammenhang zwischen Miete Anzahl der m2.

(c)

```
x0 <- data.frame(m2 = 100)
m100 <- predict(fit.zh, newdata = x0, interval = "prediction", level = 0.90)
```

Dem Freund muesste man sagen dass der Mietpreis vorraussichtlich CHF 3314.45 betragen wird und mit 90% Wahrscheinlichkeit zwischen CHF 2470.33 und CHF 4158.58 liegt.

(d)

```
fit.zh2 <- lm(Miete ~ m2 + Zimmer, data = zueri)
summary(fit.zh2)
```

```
##
## Call:
## lm(formula = Miete ~ m2 + Zimmer, data = zueri)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1126.90  -330.68    -3.04   344.55  1322.00
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  365.650    145.390   2.515   0.0141 *
## m2            26.947      3.775   7.138 5.76e-10 ***
## Zimmer        67.320    101.139   0.666   0.5078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 504.8 on 73 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8286
## F-statistic: 182.3 on 2 and 73 DF,  p-value: < 2.2e-16
```

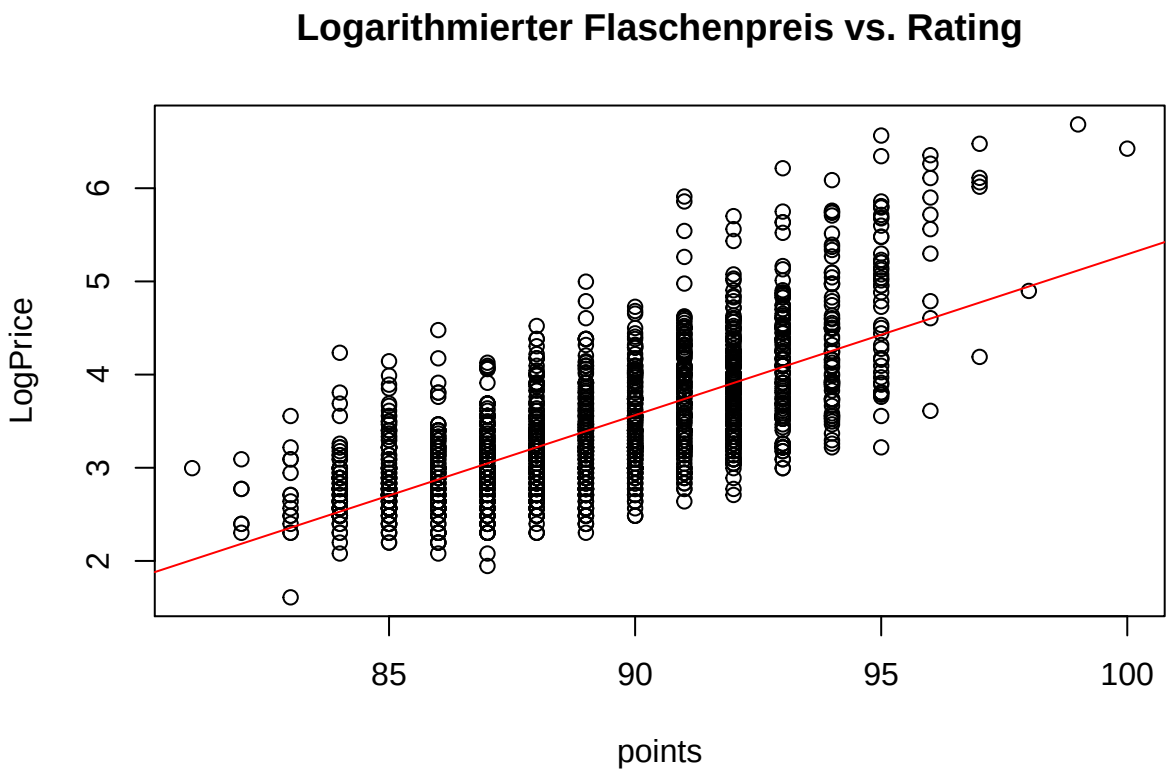
Problem: Die Anzahl der Zimmer scheint das Modell nicht zu verbessern (t-Test fuer Zimmer mit p-Wert >0.05).

Grund koennte sein, dass Flaeche (m2) und Anzahl der Zimmer miteinander korellieren. Wir also eine zirkulaere oder duplizierte Variable eingefuehrt haben.

Aufgabe 2

(a)

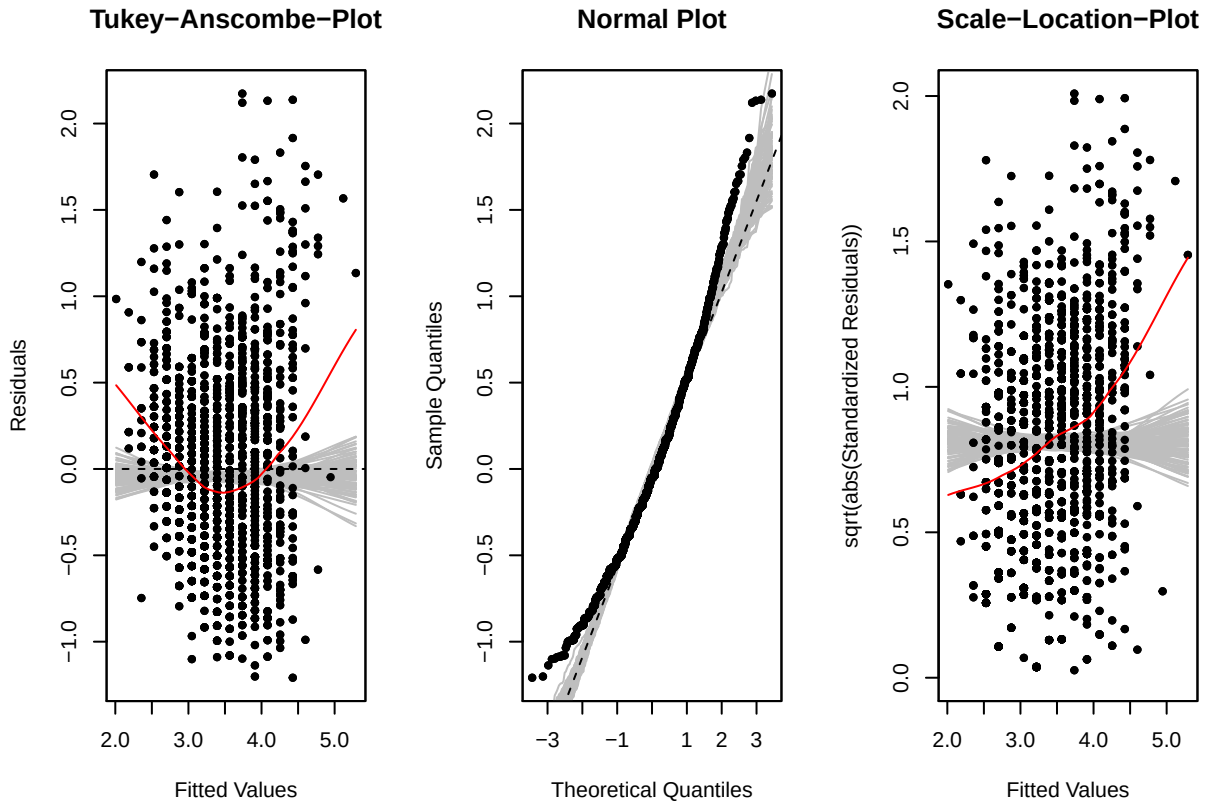
```
wine$LogPrice <- log(wine$price)
par(mfrow = c(1, 1))
plot(LogPrice ~ points, data = wine, main = "Logarithmierter Flaschenpreis vs. Rating")
fit.wine <- lm(LogPrice ~ points, data = wine)
abline(fit.wine, col = 'red')
```



Der Zusammenhang zwischen der Qualitaet (points) und dem logarithmierten Preis scheint linear zu sein. Das scheint schluessig, schliesslich ziehen hoch bewertete Weine die Aufmerksamkeit einer Kaeuferschaft aus der ganzen Welt auf sich, was die Preise exponentiell steigen laesst.

(b)

```
par(mfrow = c(1, 3))
#plot(fit.wine, 1:3)
load(paste0(data.path, "resplot.rda"))
resplot(fit.wine, 1:3)
```



Tukey-Anscombe-Plot: Die Glätter zeigt eine systematische Abweichung von der Horizontalen, d.h. der Erwartungswert ist nicht konstant 0.

Normalplot: Datenpunkte weichen links und rechts stark von der Geraden ab. Ein Hinweis auf Langschwäzigkeit. Die Daten scheinen nicht normalverteilt zu sein.

Location-Scale-Plot: Der Glätter ist nach rechts stark ansteigend. Die Varianz ist also nicht konstant.

Unabhaengigkeit: Da zeitliche Reihenfolge der Messungen unbekannt sind Aussagen zur zeitlichen Unabhaengigkeit nicht moeglich.

Fazit: Die Anpassung ist ungenuegend.

(c)

Um der langschwaenzigen Verteilung entgegenzuwirken koennte man zu einer **robusten Regressionsmethode** uebergehen. Die fehlende Linearitaet im Tukey-Anscombe-Plot koennte darauf hinweisen, dass **wichtige erklaerende Variablen fehlen** . Man koennte diese in das Modell mit aufnehmen.

(d)

```
wine$lprice <- log(wine$price)
wine$qpoints <- wine$points^2
fit.wine2 <- lm(lprice ~ points + taster_name + qpoints +
                  variety + country, data = wine)
# Schnaeppchenwein
wine[residuals(fit.wine2) == min(residuals(fit.wine2)), ]
```

```
##      country province  variety taster_name price points LogPrice   lprice
## 7654  France   Alsace Riesling  Roger Voss    37     96 3.610918 3.610918
##      qpoints
## 7654    9216
```

Den o.g. Riesling aus dem Elsass mit 96 Punkten und einem Preis von 37 (CHF?) wuerde ich auf Grund der Residuen als Schlaeppchen empfehlen.

(f)

```
x0 = data.frame(points=92, taster_name="Roger Voss", country="Austria", variety="Riesling", qpoints=92^1
exp(predict(fit.wine2, newdata = x0, interval = 'prediction', level = 0.90))
```

```
##        fit      lwr     upr
## 1 48.01583 23.57004 97.8157
```

Der mittlere vorhergesagte Preis liegt bei 48.02 Franken, der wahre Preis liegt mit 90% Wahrscheinlichkeit zwischen 23.57 und 97.81 Franken.