Klausur Datenanalyse/Statistik (WS 2019/20)

Matrikelnummer: Studiengang:

Aufgabe	1	2	3	4	5		$\sum$
Punkte möglich	10	9	8	16	21		64
Punkte erreicht							

Unter der Nummer

D	S	2	0	1			

werden wir ihre Ergebnisse bereitstellen.
Schreiben Sie sich diese Nummer bitte jetzt auf!

Diese Klausur wird nur dann als Prüfung gewertet, wenn Sie im Prüfungsamt angemeldet sind. Ansonsten werden die Ergebnisse nur für einen Schein gewertet.
Lesen Sie die Aufgaben genau durch. Nehmen Sie für diese Klausur grundsätzlich ein α-Niveau von 5% an.

Aufgabe 1: Daten
Wir schreiben das Jahr 2022. Der Klimawandel hat sich durch die extremen Waldbrände der letzten vier Jahre stark beschleunigt. Die Verschiebung der Klimazonen destabilisiert weltweit die Waldökosysteme und reduziert die CO_2-Speicherfähigkeit der Ozeane. Das Abschmelzen der antarktischen und nordischen Eisschilde erhöht den Meeresspiegel jährlich um ca. 4 cm. Küstenstädte wie New York, Hong Kong, Sydney und Hamburg werden trotz eifriger Dammbaumaßnahmen nur zu halten sein, wenn der CO_2-Ausstoß der Menschheit bis 2030 auf Netto 0 reduziert wird.

Sie haben aufgrund Ihrer hervorragenden Studienleistungen eine gut bezahlte Stelle am Europäischen Institut für Energieforschung (EIE) erhalten. Im Rahmen des 1 Billion Euro Sofortprogramms der EU arbeiten Sie an der Massenproduktion von Biokraftstoffen, da für eine Umstellung der Fahrzeugflotte auf E-Mobilität wegen der weltweiten Abschaltung aller mit fossilen Energieträgern betriebenen Kraftwerke einfach nicht genug Strom bereitsteht.

Es wurden 10 verschiedene Anbauregime untersucht. Dazu wurden aus den landwirtschaftlichen Nutzflächen Nord- und Mitteleuropas 100 Parzellen von jeweils 10 mal 10 Meter zufällig ausgesucht. Für jede der Parzellen wurde ein zufälliges Saatgut (`Saat`) S0-S9 einer Energiepflanze (z.B. Sizilianischer Raps) ausgewählt.

Außerdem wurde für jede Parzelle die folgenden Parameter erhoben:

- `Ertrag`: Der Energieinhalt des von den Erträgen der Parzelle generierbaren Biokraftstoffs.
- `QT`: Der Energieinhalt gemäß der schnellen Ertragsmessung mittels der sogenannten QuickTechnology.
- `Boden`: Bodentyp, eine Klassifizierung der Bodentypen nach einer EU-Bodensystematik, mit 15 verschiednen Bodentypen (B1-B15).
- `Sonne`: Der Gesamtenergiefluss an Lichtenergie während der Vegetationsperiode.
- `VegTemp`: Die durchschnittliche Temperatur während der Vegetationsperiode.
- `Season`: Die Dauer der Vegetationsperiode (in Jahren pro Jahr).

Das Ziel ist es – basierend auf genauen Klimavorhersagen – das jeweils richtige Saatgut für jedes Jahr an jedem Standort vorhersagen zu können. Dazu soll die Abhängigkeit der Energieausbeute von allen anderen Einflussgrößen untersucht werden.

```
> Biofuel[1:25,]
```

	Saat	Ertrag	QT	Boden	Season	Sonne	VegTemp
1	S0	186.06089	195.69008	B15	0.5863014	294.9896	24.51479
2	S3	2037.21700	2015.03621	B15	0.5917808	270.1011	23.60717
3	S9	2915.69393	2924.70595	B7	0.5808219	304.3406	27.17418
4	S0	175.27924	147.65022	B6	0.5808219	282.0890	24.67192
5	S4	9663.72216	9654.44024	B2	0.5917808	431.2796	25.46931
6	S0	250.40347	220.19403	B5	0.5917808	224.8966	23.39919
7	S5	1284.23322	1247.46685	B8	0.5808219	348.1301	26.36727
8	S3	2331.99608	2306.64107	B12	0.5972603	231.5519	24.69991
9	S3	2294.41033	2301.91295	B8	0.6082192	238.4848	24.47043
10	S5	528.07164	523.21843	B9	0.5863014	318.2036	25.82059
11	S6	53.48699	37.85617	B15	0.5917808	305.6510	24.07709
12	S3	1848.06231	1846.44304	B6	0.5808219	328.6425	26.09258
13	S7	1824.62510	1880.43344	B14	0.6082192	279.4001	22.75624
14	S8	2765.38646	2770.35888	B4	0.5972603	244.2337	25.03641
15	S3	2083.00960	2107.09051	B5	0.5863014	255.4093	25.16602
16	S8	7204.05199	7183.17783	B10	0.5972603	208.6217	25.76073
17	S1	3009.92596	2983.69239	B2	0.5917808	179.1553	24.68738
18	S9	3471.18731	3466.10147	B9	0.5808219	372.2437	25.36003
19	S5	445.98918	460.68942	B6	0.6027397	367.1993	23.50601
20	S2	1346.02017	1303.03375	B3	0.5917808	194.1603	25.77363
21	S1	2963.56289	2961.92954	B4	0.5917808	362.4976	24.62688
22	S0	221.85170	205.38580	B13	0.5753425	292.6997	24.63123
23	S9	1930.11730	1939.75629	B13	0.5917808	318.2792	23.91176
24	S3	1905.60738	1885.48588	B15	0.5753425	323.7297	27.00284
25	S7	834.29481	829.10999	B12	0.5753425	306.2635	24.90389

(1) Um welche Darstellungsform der Daten handelt es sich hier? (1)

(2) Was ist die Grundgesamtheit für diesen Datensatz? (2)

(3) Ist die Probennahme bezüglich dieser Grundgesamtheit repräsentativ? Warum? (2)

(4) Warum kann mit diesem Datensatz zwar eine kausale Aussage über die Wirkung des gewählten Saatguts auf den Ertrag gemacht werden, jedoch nicht über die Wirkung des vorliegenden Bodentyps auf den Ertrag? (2)

(5) Welches Skalenniveau haben diese Variablen? (3)

- `Saat`:
- `Ertrag`:
- `Season`:

Aufgabe 2: Methoden auswählen
Welche statistische Methodik sollte man wählen, ...

(1) ... um die Variable `Boden` grafisch darzustellen? (2)

---

(2) ... um die Variablen `Saat` und `Boden` gemeinsam darzustellen? (1)

---

(3) ... um zu widerlegen, dass der `Ertrag` normalverteilt ist? (1)

---

(4) ... um nachzuweisen, dass der `Ertrag` von der Temperatur (`VegTemp`) abhängt, wenn wir davon ausgehen, dass dieser Zusammenhang nichtlinear ist? (1)

---

(5) ... um Ausreißer zu erkennen? (1)

---

(6) ... um die Abhängigkeit des Ertrages (`Ertrag`) vom gewählten Saatgut (`Saat`) grafisch darzustellen? (1)

---

(7) ... um die Repräsentativität der Daten sicherzustellen? (1)

---

(8) ... um die Abhängigkeit des Ertrages von allen anderen Größen gleichzeitig zu untersuchen? (1)

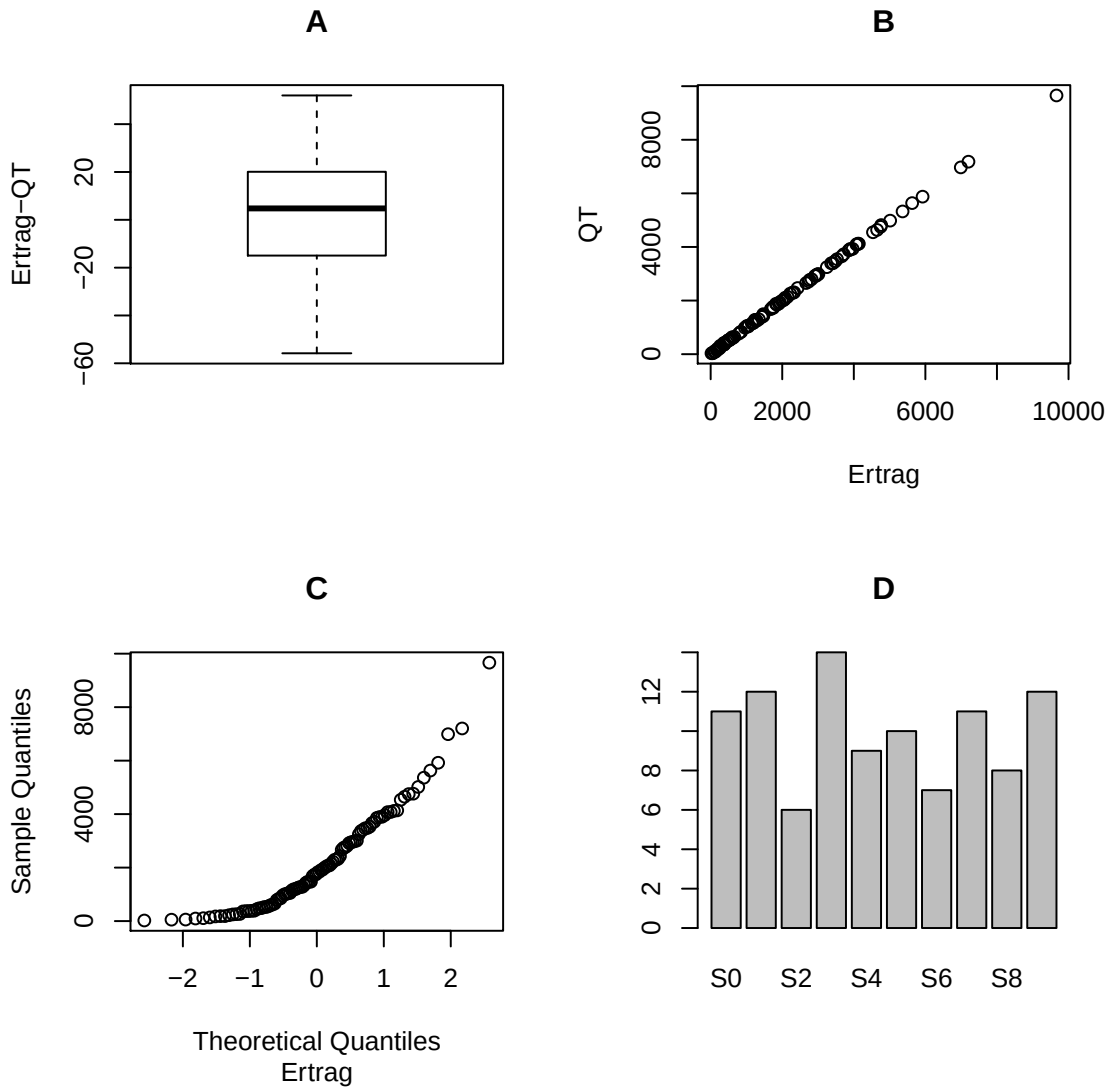

Abbildung 1: Grafiken zum Datensatz.

Aufgabe 3: Grafiken

In Abbildung 1 finden Sie eine Reihe von Grafiken zum Datensatz.

(1) Grafik A:
Wie heißt diese Grafik? (1)

Geben Sie eine Schlussfolgerung aus dieser Grafik wieder. (1)

---

(2) Grafik B:
Wie heißt diese Grafik? (1)

Geben Sie eine Schlussfolgerung aus dieser Grafik wieder. (1)

---

(3) Grafik C:
Wie heißt diese Grafik? (1)

Beschreiben Sie die Verteilung bezüglich ihrer Schiefe. (1)

---

(4) Grafik D:
Wie heißt diese Grafik? (1)

Geben Sie eine Schlussfolgerung aus dieser Grafik wieder. (1)

Aufgabe 4: Misst die schnelle Ertragsmessung im Mittel richtig?
Im Rahmen der weiteren Untersuchungen wurde festgestellt, dass die Abhängigkeiten relativ kompliziert sind und für eine genaue Auswahl des richtigen Saatguts mehr Daten benötigt werden. Um Kosten zu sparen, soll dabei die schnellen Ertragsmessung mittels der sogenannten QuickTechnology verwendet werden. Daher soll überprüfen werden, ob die schnelle Ertragsmessung im Mittel die gleichen Ergebnisse liefert wie die genaue Untersuchtung mittels der aufwendigen vollständigen Aufarbeitung des Materials.

(1) Hier finden Sie ein paar Tests zum Datensatz:

```
> shapiro.test(Ertrag)

        Shapiro-Wilk normality test

data:  Ertrag
W = 0.89826, p-value = 1.172e-06

> shapiro.test(QT)

        Shapiro-Wilk normality test

data:  QT
W = 0.89916, p-value = 1.292e-06

> shapiro.test(Ertrag-QT)

        Shapiro-Wilk normality test

data:  Ertrag - QT
W = 0.97633, p-value = 0.06855

> t.test(Ertrag,QT)

        Welch Two Sample t-test

data:  Ertrag and QT
t = 0.0073994, df = 198, p-value = 0.9941
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -510.5577  514.4035
sample estimates:
mean of x mean of y
 2169.716  2167.793
```

```
> t.test(Ertrag,QT,paired=TRUE)

        Paired t-test

data:  Ertrag and QT
t = 0.79055, df = 99, p-value = 0.4311
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.903462  6.749340
sample estimates:
mean of the differences
               1.922939

> t.test(Ertrag,QT,var.equal=TRUE)

        Two Sample t-test

data:  Ertrag and QT
t = 0.0073994, df = 198, p-value = 0.9941
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -510.5577  514.4035
sample estimates:
mean of x mean of y
 2169.716  2167.793

> wilcox.test(Ertrag,QT)

        Wilcoxon rank sum test with continuity correction

data:  Ertrag and QT
W = 5006, p-value = 0.9893
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Ertrag,QT,paired=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  Ertrag and QT
V = 2838, p-value = 0.2826
alternative hypothesis: true location shift is not equal to 0

> fligner.test(list(Ertrag,QT))

        Fligner-Killeen test of homogeneity of variances

data:  list(Ertrag, QT)
Fligner-Killeen:med chi-squared = 0.0010688, df = 1, p-value = 0.9739
```

Kreuzen Sie unten die Tests an, deren Voraussetzungen erfüllt sind und schreiben Sie in der Zeile darunter jeweils eine stichpunktartige Begründung:
Bei erfüllten Voraussetzungen: Woher wissen Sie das?
Bei nicht erfüllter Voraussetzung: Welche Voraussetzung ist nicht erfüllt? Woher wissen Sie das? (6)

O `shapiro.test(Ertrag)`

O `shapiro.test(Ertrag-QT)`

O `t.test(Ertrag,QT,paired=TRUE)`

O `t.test(Ertrag,QT,var.equal=TRUE)`

O `wilcox.test(Ertrag,QT)`

O `fligner.test(list(Ertrag,QT))`

---

(2) Beschäftigen wir uns nun konkret mit der Frage, ob die schnelle Ertragsmessung im Mittel richtig misst. Beschreiben Sie die Testsituation bezüglich der folgenden Merkmale (4):
Anzahl der beteiligten Merkmale

Anzahl der beteiligten Stichproben

Zu testende Größe

Verteilungsvoraussetzungen

---

(3) Welchen der folgenden Tests sollten wir für diese Frage einsetzen? (2)

- **O** a) `shapiro.test(Ertrag)`
- **O** b) `shapiro.test(QT)`
- **O** c) `shapiro.test(Ertrag-QT)`
- **O** d) `t.test(Ertrag,QT)`
- **O** e) `t.test(Ertrag,QT,paired=TRUE)`
- **O** f) `t.test(Ertrag,QT,var.equal=TRUE)`
- **O** g) `wilcox.test(Ertrag,QT)`
- **O** h) `wilcox.test(Ertrag,QT,paired=TRUE)`
- **O** i) `fligner.test(list(Ertrag,QT))`

(4) Welche Hypothese wurde bei diesem Test angenommen? (1)

---

(5) Wurde damit ein statistischer Nachweis geführt? (1)

---

(6) Mit welchem Wert haben Sie dabei den p-Wert verglichen? Begründen Sie unter Bezugnahme auf die BONFERRONI-Korrektur. (1)

---

(7) Kann die schnelle Ertragsmessung mittels der sogenannten QuickTechnology als Alternative zur aufwendigen Ertragsmessung eingesetzt werden? (1)

Aufgabe 5: Lineare Modelle

In dieser Aufgabe wollen wir nun eine Methode entwickeln, um den Ertrag aus Saatgut, Boden und Klimaverhältnissen vorhersagen zu können. Diagnostische Grafiken zu einem der Modelle finden Sie in den Abbildung 2 auf der letzten Seite.

```
> xanova <- function(mod) {
+   print(anova(mod))
+   cat("R^2=",var(predict(mod))/var(predict(mod)+resid(mod)),"\n")
+   mod
+   }
> logErtrag <- log(Ertrag)
```

```
> M01 <- xanova(lm(logErtrag~Boden))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value Pr(>F)
Boden     14  22.598  1.6141  1.1311 0.3438
Residuals 85 121.297  1.4270
R^2= 0.1570444

> M02 <- xanova(lm(logErtrag~Saat))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value    Pr(>F)
Saat       9 121.241 13.4712  53.517 < 2.2e-16
Residuals 90  22.655  0.2517
R^2= 0.8425617

> M03 <- xanova(lm(logErtrag~Sonne))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value  Pr(>F)
Sonne      1   4.321  4.3209  3.0338 0.08468
Residuals 98 139.574  1.4242
R^2= 0.03002783

> M04 <- xanova(lm(logErtrag~VegTemp))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value Pr(>F)
VegTemp    1   3.688  3.6879  2.5777 0.1116
Residuals 98 140.207  1.4307
R^2= 0.02562923

> M05 <- xanova(lm(logErtrag~Saat+Boden))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value    Pr(>F)
Saat       9 121.241 13.4712 107.679 < 2.2e-16
Boden     14  13.147  0.9390   7.506  1.42e-09
Residuals 76   9.508  0.1251
R^2= 0.9339241
```

```
> M06 <- xanova(lm(logErtrag~Saat+Sonne))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value    Pr(>F)
Saat       9 121.241 13.4712  59.087 < 2.2e-16
Sonne      1   2.363  2.3635  10.367  0.001792
Residuals 89  20.291  0.2280
R^2= 0.8589867

> M07 <- xanova(lm(logErtrag~Saat+VegTemp))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value Pr(>F)
Saat       9 121.241 13.4712 52.9554 <2e-16
VegTemp    1   0.014  0.0141  0.0556 0.8142
Residuals 89  22.640  0.2544
R^2= 0.8426599

> M08 <- xanova(lm(logErtrag~Saat+Boden+Sonne))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value    Pr(>F)
Saat       9 121.241 13.4712 146.199 < 2.2e-16
Boden     14  13.147  0.9390  10.191 2.510e-12
Sonne      1   2.597  2.5973  28.188 1.083e-06
Residuals 75   6.911  0.0921
R^2= 0.951974

> M09 <- xanova(lm(logErtrag~Saat+Boden+VegTemp))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq  F value    Pr(>F)
Saat       9 121.241 13.4712 106.5780 < 2.2e-16
Boden     14  13.147  0.9390   7.4293 1.926e-09
VegTemp    1   0.028  0.0282   0.2231     0.638
Residuals 75   9.480  0.1264
R^2= 0.9341201
```

```
> M10 <- xanova(lm(logErtrag~Saat+Boden+Sonne+VegTemp))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value    Pr(>F)
Saat       9 121.241 13.4712 144.282 < 2.2e-16
Boden     14  13.147  0.9390  10.058 3.957e-12
Sonne      1   2.597  2.5973  27.818 1.275e-06
VegTemp    1   0.002  0.0016   0.017    0.8967
Residuals 74   6.909  0.0934
R^2= 0.951985

> M11 <- xanova(lm(logErtrag~Boden+Sonne+Saat))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value    Pr(>F)
Boden     14  22.598  1.6141  17.518 < 2.2e-16
Sonne      1   5.202  5.2025  56.461 9.982e-11
Saat       9 109.184 12.1316 131.660 < 2.2e-16
Residuals 75   6.911  0.0921
R^2= 0.951974

> M12 <- xanova(lm(logErtrag~Saat+Sonne+Boden))

Analysis of Variance Table

Response: logErtrag
          Df  Sum Sq Mean Sq F value    Pr(>F)
Saat       9 121.241 13.4712 146.199 < 2.2e-16
Sonne      1   2.363  2.3635  25.650 2.844e-06
Boden     14  13.380  0.9557  10.372 1.683e-12
Residuals 75   6.911  0.0921
R^2= 0.951974

> M13 <- xanova(lm(logErtrag~Saat*Boden+Sonne))

Analysis of Variance Table

Response: logErtrag
           Df  Sum Sq Mean Sq  F value    Pr(>F)
Saat        9 121.241 13.4712 167.8535 < 2.2e-16
Boden      14  13.147  0.9390  11.7007 3.988e-07
Sonne       1   2.597  2.5973  32.3627 1.012e-05
Saat:Boden 53   5.145  0.0971   1.2096    0.3192
Residuals  22   1.766  0.0803
R^2= 0.9877298
```

```
> coef(M12)
```

```
 (Intercept)       SaatS1       SaatS2       SaatS3       SaatS4       SaatS5
 5.021357327  1.930235660  1.817875681  2.232145779  2.548882064  1.004511291
      SaatS6       SaatS7       SaatS8       SaatS9        Sonne     BodenB10
-1.182335329  0.899792381  2.660440928  2.241850395  0.002973387  0.160521611
    BodenB11     BodenB12     BodenB13     BodenB14     BodenB15      BodenB2
-0.155385090  0.076306791 -0.508731843  0.633925322 -0.668208499  0.460581024
     BodenB3      BodenB4      BodenB5      BodenB6      BodenB7      BodenB8
-0.205213062 -0.092275471 -0.131321823 -0.773898376 -0.325752618  0.160046405
     BodenB9
-0.244963080
```

(1) Bei den Modellen in dieser Aufgabe wird an Stelle des Ertrags der logarithmierte Ertrag als Zielgröße verwendet. Warum ist das sinnvoll? (1)

---

(2) Betrachten Sie Abbildung 2. Gibt es Probleme bei der Anwendung dieses Modells? Welche Probleme liegen vor? (3)

---

(3) Welches der Modelle `M01`-`M13` sollte aus statistischer Sicht für die Beschreibung des Sachverhalts gewählt werden? (2)

Warum? (3)

---

(4) Warum wäre es trotz des höheren R^2 nicht besser `M13` einzusetzen? (2)

---

(5) Warum wäre es nicht besser `M02` einzusetzen? (2)

---

(6) Welches Saatgut liefert die höchste Energieausbeute. Woran erkennen Sie das? (2)

---

(7) Welchen Ertrag würden Sie gemäß Model `M12` für den folgenden Standort voraussagen, wenn tatsächlich das beste Saatgut gewählt worden wäre? Geben Sie dazu eine Formel mit eingesetzten Zahlen. Es ist nicht notwendig, die Werte zu berechenen. (3)

```
> Biofuel[1,]

  Saat   Ertrag       QT Boden    Season    Sonne  VegTemp
1   S0 186.0609 195.6901   B15 0.5863014 294.9896 24.51479
```

---

(8) Was sagt uns der R^2 Wert von Model `M12`? (1)

---

(9) Ein Kollege äußert, dass der Datensatz mit nur 100 Parzellen ja reichlich klein sei. Er schlägt vor, das Experiment mit mindestens 10000 Parzellen zu wiederholen. Die Parzellen sollen dabei in Absprache mit den Landwirten ausgewählt werden. Zudem sollen die Landwirte jeweils das für die Parzelle geeignetste Saatgut auswählen. Kommentieren Sie diesen Vorschlag. (2)

```
> mod <- M12
> par(mfrow=c(2,2))
> plot(predict(mod),resid(mod))
> plot(predict(mod),influence(mod)$hat)
> plot(predict(mod),cooks.distance(mod))
> qqnorm(resid(mod))
```

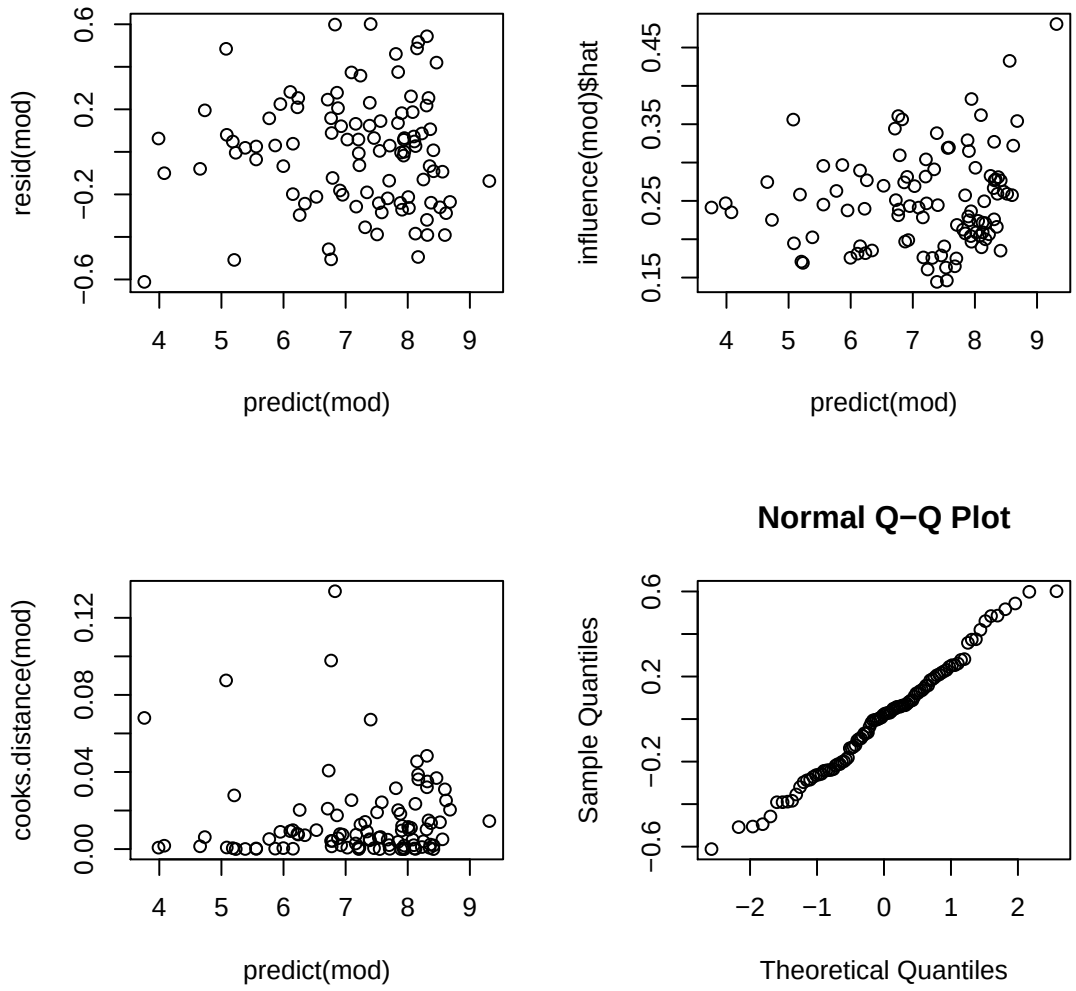

Abbildung 2: Diagnostische Grafiken zum Modell `M12`.