Testat

Statistik für Geographen (SS 04)

Name: Matrikelnummer:

Lesen Sie und antworten dann. Zum Bestehen benötigen Sie 16 Punkte. Die erreichbare Punktzahl ist bei allen Teilaufgaben in Klammern () angegeben. Nehmen Sie für dieses Testat grundsätzlich ein α-Niveau von **5**% an. **Soweit nicht anders angegeben** genügt eine Antwort in **Stichworten**. Ein Teil der Daten dieses Testats sind frei erfunden, andere nicht.

Aufgabe 1: Umfrage zum Urlaubsverhalten
Der folgende Datensatz zeigt einen Ausschnitt (der Zeilen und der Spalten) einer Umfrage mit 4000 deutschen Haushalten zum Thema Tourismus. Die Umfrage wurde mit Hilfe der sogenannten zufälligen digitalen Wahl erhoben, bei der zunächst eine zufällige Telefonnummer ausgewürfelt wird und dann falls der Anschluß exisitiert und einem Privathaushalt zugeordnet ist, eine erwachsene Person dieses Haushalts per Telefon nach ihrer letzten Urlaubsreise befragt wird.

Name	**Alter**	**Urlaubstyp**	**Urlaubsort**	**Urlaubskosten**	**Dauer**	...
Huber	34	Kultur	Europa	1400	14	...
Nolte	24	Erlebnis	Europa	1000	10	...
Seibel	50	Wellness	Deutschland	1500	8	...
Meier	42	Bade	Europa	1400	14	...
Nüremberg	35	Wander	Heimat	140	7	...
Patozzi	18	Sport	Europa	500	7	...
Grauner	27	Kultur	Übersee	7600	28	...
Busch	23	Erlebnis	Deutschland	700	7	...
Niemeier	70	Kultur	Übersee	4000	22	...
Grube	45	Wellness	Europa	1100	8	...
Schäfer	36	Bade	Europa	1400	14	...
Maier	45	Wander	Heimat	500	7	...
Myer	67	Wander	Europa	1800	7	...
Mayer	29	Kultur	Europa	2000	14	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Beschreibung: Die Variable `Name` gibt den Familiennamen des Befragten an. Die Variable `Alter` gibt das Alter des Befragten an. Die Variable `Urlaubstyp` bezeichtnet den Typ des Urlaubs (Bade-, Wander-, Kultur-, Erlebnis-, Sport-, Wellnessurlaub). Der `Urlaubsort` gibt eine Einteilung nach zunehmender Entfernung wieder: (Heimat-, Deutschland-, Europa-, Überseeurlaub). Die `Urlaubskosten` bezeichnen die vom Befragten geschätzen Urlaubskosten pro Person in Euro. `Dauer` gibt die Anzahl der Urlaubstage an.

- Geben Sie die Skalenniveaus zu allen 6 Variablen *detailliert* an: (4)

- Kommentieren Sie die Representativität der Umfrage. Gehen Sie dazu auf die angestrebte Grundgesamtheit, eventuelle Bedenken gegen die Representativität und Chancen es besser zu machen ein. **Antworten Sie in ganzen Sätzen.** (4)

•Durch welche statistische Graphik läßt sich die Variable `Urlaubstyp` am besten darstellen.(1)

•Was sollte man bei der graphischen Darstellung der Variable `Urlaubsort` beachten.(1)

•Durch welche statistische Graphik läßt sich die Variable `Urlaubskosten` am besten darstellen.(1)

•Was passiert, wenn man `Dauer` durch ein Punktdiagramm darstellt und was kann man dagegen tun?(2)

•Durch welche statistische Graphik kann man die Abhängigkeit der Kosten für die Reise von der Reisedauer darstellen?(1)

•Durch welche statistische Graphik kann man die Abhängigkeit der Kosten für die Reise vom `Urlaubstyp` besonders gut darstellen?(1)

Aufgabe 2: Industriestandorte

Eine internationle Unternehmensberatung möchte ein neues Bewertungssystem potentieller Industriestandorte in Ländern der dritten Welt etablieren, um seinen Kunden zu Investitionen an bestimmten Orten raten zu können. Dazu wurden einige hundert Industriebetriebe in der dritten Welt zufällig ausgewählt und dann nach verschiedenen Aspekten bezüglich ihrer Standortansprüchen, ihrer Produktivität und den am Ort vorgefundenen Voraussetzungen bewertet. In dem Datensatz befinden sich viele verschiedene Variablen. Wir werden uns nur mit einigen davon beschäftigen.

Eine wichtige Determinate der Wirtschaftlichkeit sind die ortsüblichen Lohnkosten für Industriearbeiter. Da zu Neustandorten in unindustrialisierten Regionen gewöhnlich keine Ver-

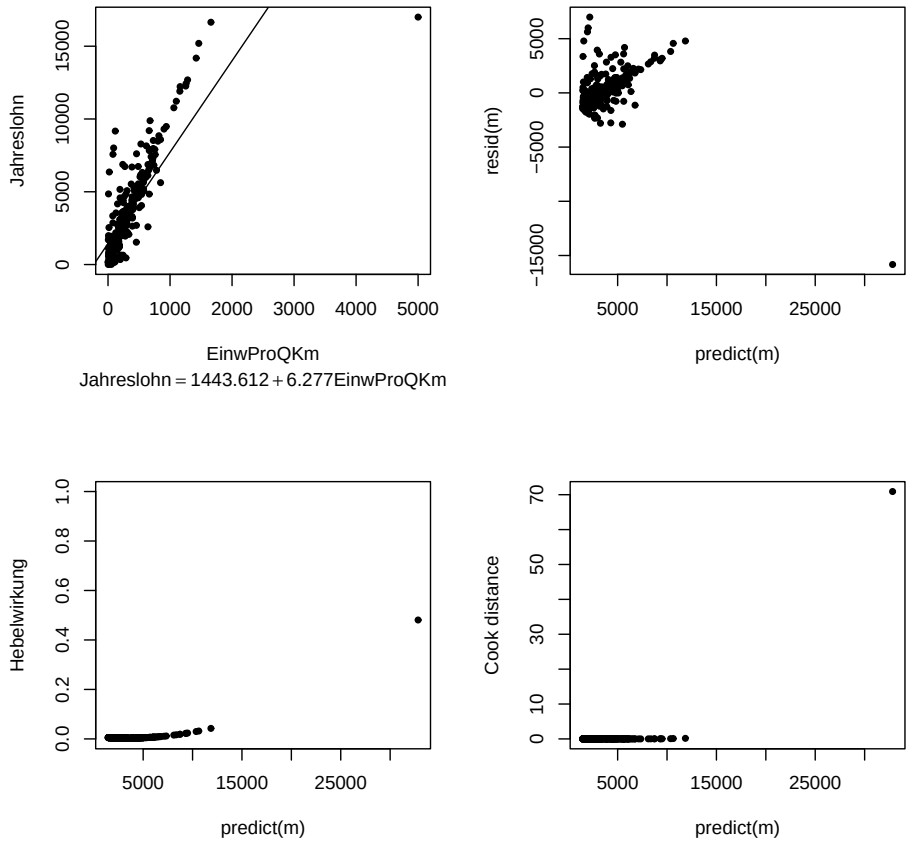

Abbildung 1: Graphiken zur Regression von Jahreslohn in Abhängigkeit von Einwohnern pro Quadratkilometer.

gleichslöhne vorliegen, sollen der Durchschnittslohn für eine Industriearbeitskraft aus sozioökonomischen Kenngrößen des Landes geschätzt werden.

(a) Zunächst wird der Versuch unternommen die Abhängigkeit des Jahreslohns von der Einwohnerzahl pro Quadratkilometer mittels einer linearen Regression zu bestimmen. Die diagnostischen Graphiken befinden sich in Abbildung 1. Hier ist die Computerausgabe zu dieser Regression:

```
Korrelation= 0.8111174  R^2= 0.6579114

Analysis of Variance Table
Response: Jahreslohn
            Df     Sum Sq    Mean Sq F value    Pr(>F)
EinwProQKm   1 1810760137 1810760137  575.04 < 2.2e-16 ***
Residuals  299  941525666    3148915
```

Mit dieser Regression gibt es ein deutliches Problem.
Was ist das Problem? Antworten Sie in einem ganzen Satz.(1)

In welcher Graphik kann man das Problem erkennen?(1)

Woran kann man es in der Graphik erkennen?(1)

(b) Ab jetzt untersuchen wir die Abhängigkeit des Jahreslohns vom pro Kopf erziehlten Bruttoinlandsprodukt des Landes. Die diagnostische Graphiken und das Konfidenzintervall für die Vorhersage zusätzlicher Datenpunkte befinden sich in Abbildung 2. Die Computerausgabe lautet wie folgt:

```
Korrelation= 0.9945172  R^2= 0.9890645

Analysis of Variance Table
Response: Jahreslohn
                            Df     Sum Sq    Mean Sq F value    Pr(>F)
ProKopfBruttoinlandsprodukt   1 2722188196 2722188196   27043 < 2.2e-16 ***
Residuals                   299   30097606     100661
```


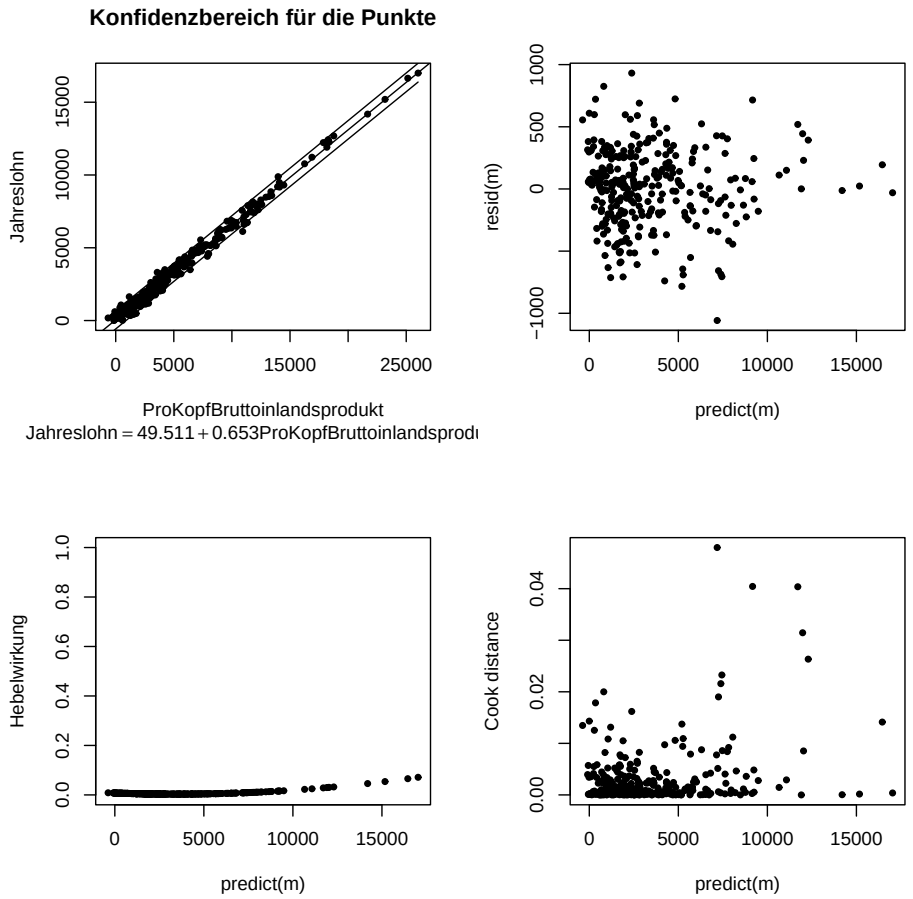

Abbildung 2: Graphiken zur Regression von Jahreslohn in Abhängigkeit von Bruttoinlandsprodukt.

Ist die Abhängigkeit des Jahreslohnes vom Bruttoinlandsprodukt statistisch signifikant nachgewiesen? Woran erkennt man das? (2)

Ist dieser Zusammenhang stark ausgeprägt? An welcher ausgegeben Zahl können Sie das erkennen? (2)

Ihr Kunde möchte in einem Land mit einem pro Kopf Bruttoinlandsprodukt von 5000 Dollar investieren. Können Sie (mit einer Fehlerwahrscheinlichkeit von $\leq 5\%$) garantieren, daß der Jahreslohn unter 5000 Dollar liegt?(1)

Woran erkennt man das?(1)

Aufgabe 3: Old Faithfull

Der Geisir Old Faithfull brach über viele Jahre mit großer Regelmäßigkeit aus. In zwei – mehrere Jahre auseinanderliegenden – Zeitperioden wurden jeweils die Wartezeiten zwischen zwei aufeinanderfolgenden Ausbrüchen notiert. Uns stehen diese Datenreihen zur Verfügung und wir wollen überprüfen, ob sich die Verteilung der Wartezeiten zwischen diesen beiden Perioden geändert hat.

(a) Sie überprüfen die Gleichheit der Wartezeitenverteilung mit einem Kolmogorov-Smirnov auf Gleichheit zweier Verteilungen. Das ist ein Test, der mit dem besprochenen Kolmogorov-Smirnov Test für eine einzelne Stichprobe verwandt ist. Hypothese und Alternative lauten:

$$H_0 : \text{Die Verteilungen beider Stichproben sind gleich}$$

vs.

$$H_1 : \text{Die Verteilungen beider Stichproben sind verschieden}$$

```
> ks.test(oldfaithful$WT1,oldfaithful$WT2)

        Two-sample Kolmogorov-Smirnov test

data:  oldfaithful$WT1 and oldfaithful$WT2
D = 0.1, p-value = 0.6994
alternative hypothesis: two.sided
```


Wurde die Hypothese angenommen oder abgelehnt?(1)

Nehmen wir an die Voraussetzungen des Tests waren erfüllt. Was haben Sie dann mit diesem Test **nachgewiesen**?(2)

Dieser Test ist ein nichtparametrischer Test. Nennen Sie die beiden typischen Voraussetzungen, die praktisch alle nichtparametrischen Tests wie z.B. der Wilcoxen Rangsummentest haben (2).

(b) Prof. Vierziger hat diesen Datensatz ebenfalls untersucht und 40 verschiedene Tests durchprobiert, um die beiden Meßreihen zu vergleichen: Den Kolmogorov-Smirnov Test auf gleiche Verteilung, den Quantilskorrelationstest auf gleiche Quantile, den t-test auf gleiche Mittelwerte, den Kruskal-Wallis-Test auf gleiche Mediane, den Wilcoxen-Test auf gleiche Lage, den F-Test auf gleiche Streuung und so weiter. Der 40ste Test, der Formtest auf gleiche Kurtosis, lehnt endlich die Hypothese gleicher Verteilungsform signifikant ab und Prof. Vierziger will damit nachgewiesen haben, daß sich die Verteilung der Ausbruchzeiten geändert hat. Hat er recht? Begründung? (2)

Wie geht man korrekt damit um, wenn man tatsächlich viele Tests an einem Datensatz durchführen muß? (1)

Aufgabe 4: Termiten

Der folgende Datensatz wurde erhoben, um die Wirkung eines natürlichen Insektenvertilgungsmittel (Resin) zu testen. Dazu wurden in 16 Versuchsschalen jeweils 25 Termiten gesetzt. In den ersten 8 Versuchsschalen gab man 5mg Resin hinzu und in den zweiten 8 Versuchsschalen 10mg. An jedem Tag, den der Forscher Zeit hatte, wurden die noch lebenden Termiten in jeder Schale gezählt.

```
# Daten:
   dish dose day1 day2 day3 day4 day5 day6 day7 day8 day9 day10 day11 day12 day13 day14
1     1    5   25   24    *   22   18   17   15   14    *    13    13    12    11    11
```

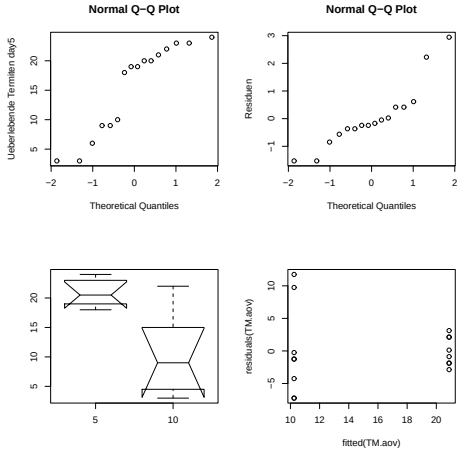

Abbildung 3: Bilder zur Aufgabe mit den Termiten

2	2	5	25	25	*	25	21	20	17	15	*	15	15	12	11	11
3	3	5	25	25	*	24	23	21	20	19	*	16	13	13	13	13
4	4	5	25	25	*	25	24	22	22	21	*	20	18	13	13	13
5	5	5	25	25	*	22	19	18	17	14	*	13	11	11	8	8
6	6	5	25	25	*	23	20	17	16	15	*	15	13	11	11	10
7	7	5	25	25	*	24	23	22	20	19	*	16	15	11	9	7
8	8	5	25	25	*	23	19	17	16	14	*	12	12	11	11	11
9	1	10	25	24	*	23	22	21	19	18	*	18	18	18	17	17
10	2	10	25	25	*	23	20	19	18	18	*	17	17	16	15	14
11	3	10	25	24	*	12	6	5	4	2	*	2	1	1	1	1
12	4	10	25	24	*	14	10	7	4	3	*	3	2	2	2	1
13	5	10	25	24	*	16	9	6	5	1	*	0	0	0	0	0
14	6	10	25	24	*	7	3	1	1	1	*	0	0	0	0	0
15	7	10	25	18	*	4	3	1	1	0	*	0	0	0	0	0
16	8	10	25	21	*	17	9	7	7	7	*	5	4	3	3	3

```
# Varianzanalyse fuer Anzahl der Ueberlebenden Termiten am 5-ten Tag
# in Abh"angigkeit von der Variable Dosis:
#
            Df Sum Sq Mean Sq F value   Pr(>F)
TM$dose      1 451.56  451.56   16.03 0.001306 **
Residuals   14 394.38   28.17
```



```
# R^2 der Varianzanalyse betraegt 0.5338013

# Die Gruppenmittelwerte:
     dose5       dose10
     20.875      10.250
# Graphiken:
# Es werden 4 Graphiken erzeugt:
# QQ-Plot der Zahl der am 5.ten Tag noch lebenden Termiten
# QQ-Plot der Residuen der Varianzanalyse
# Gekerbter Boxplot der der Zahl der am 5.ten Tag aufgeteilt nach der Dosis
# Diagnostische Graphik: Residuen in Abhaengigkeit von den Gruppenmittelwerten
```

Machen Sie vier aus den Informationen oder den Graphiken ersichtliche Aussagen über den Datensatz oder seine statistische Analyse, die für eine Auswertung im Sinne der Fragestellung relevant sein könnten.(4)