Klausur Datenanalyse und Statistik für Naturwissenschaftler (WS 2009/10)

Aufgabe:	1	2	3	4	5	6	7	8	9	10	11	12	13
Pkt. mgl.	3	3	2	2	7	2	1	2	2	1	1	1	1
Punkte erreicht:													

Aufgabe:	14	15	16	17	18	19	20	21	22	23	24	25	$\sum$
Pkt. mgl.	4	1	2	2	2	2	2	1	2	2	2	+2	50 +2
Punkte erreicht:													

Name, Vorname: Matrikelnummer:

Fachrichtung:

Unter der folgenden Nummer finden Sie Ihr Ergebnis später im Internet:

D	S	10	1			

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 sich die Aufgaben genau durch! Nehmen Sie für diese Klausur grundsätzlich ein α**-Niveau von 5%** an. **Mehrfachantworten** sind möglich. Die in Klammern angegebene Punktzahl gibt keinerlei Auskunft über die Anzahl der richtigen Antworten, sondern nur über die relative Wichtigkeit der Frage.

Viel Erfolg!

Über den Datensatz *Firmen*

Im Jahre 1960 wurde in Deutschland eine statistische Erhebung zum Modernisierungsgrad von Firmen und dessen Auswirkungen vorgenommen. Dabei wurden aus allen deutschen Textil- und Automobilfirmen, die eine vergleichbare Größe besaßen, jeweils 20 Stück per Losverfahren ausgewählt. Der Modernisierungsgrad beschrieb die Zusammensetzung des jeweiligen Maschinenparks hinsichtlich Alter und technischer Aktualität; 100 % entsprächen einem neu errichteten state-of-the-art Maschienenpark. Im Datensatz *Firmen* ist die jeweilige Branche, der Vorjahresgewinn in Millionen D-Mark sowie der Modernisierungsgrad in Prozent (Variable *Modern*) aufgeführt. Zusätzlich wurde der Modernisierungsbedarf der Firmen auf Grundlage ihres Modernisierungsgrades als *hoch* ($Modernisierungsgrad \leq 60\%$), *moderat* ($60\% < Modernisierungsgrad \leq 80\%$), oder *niedrig* ($80\% < Modernisierungsgrad$) bewertet.

```
> Firmen
```

	Gewinn	Modern	Branche	Bedarf
1	8.504510	20.98	Automobil	hoch
2	9.195576	28.05	Automobil	hoch
3	11.465603	30.92	Automobil	hoch
4	13.963123	44.37	Automobil	hoch
5	15.175772	50.69	Automobil	hoch
6	18.279415	54.65	Automobil	hoch
7	18.472655	55.02	Automobil	hoch
8	19.408148	56.13	Automobil	hoch
9	19.468526	57.21	Automobil	hoch
10	19.532189	57.53	Automobil	hoch
11	19.598980	57.57	Automobil	hoch
12	19.682841	61.39	Automobil	moderat
13	20.485879	62.25	Automobil	moderat
14	20.723508	64.24	Automobil	moderat
15	20.912245	64.28	Automobil	moderat
16	22.739388	66.00	Automobil	moderat
17	23.437237	69.75	Automobil	moderat
18	26.871939	76.84	Automobil	moderat
19	28.858688	87.97	Automobil	niedrig
20	32.811214	90.04	Automobil	niedrig
21	-8.874056	30.03	Textil	hoch
22	-2.609382	49.80	Textil	hoch
23	-1.592875	51.35	Textil	hoch
24	-1.493212	51.96	Textil	hoch
25	0.536011	55.77	Textil	hoch

```
26 -3.388703  56.27    Textil    hoch
27 -0.579129  58.37    Textil    hoch
28 -3.154020  58.41    Textil    hoch
29 -2.404088  58.98    Textil    hoch
30 -0.495115  60.94    Textil moderat
31 -2.214522  61.53    Textil moderat
32  2.738754  63.99    Textil moderat
33  5.058967  69.47    Textil moderat
34  1.099479  70.23    Textil moderat
35  3.844241  76.33    Textil moderat
36  7.174846  77.91    Textil moderat
37  4.340434  78.52    Textil moderat
38  7.141747  80.70    Textil niedrig
39 10.878819  88.94    Textil niedrig
40  8.273973  89.65    Textil niedrig
```

Daten und Grafiken

Aufgabe 1: Welches Skalenniveau haben die folgenden Variablen? (3)

Gewinn: .....

Modern: .....

Bedarf: .....

Aufgabe 2: Ist der Datensatz geeignet, um der Frage nachzugehen, ob sich der Modernisierungsgrad deutscher Textilfirmen um 1960 herum von dem entsprechender Automobilfirmen unterschied? Begründen Sie ihre Antwort im Detail! (3)

Aufgabe 3: Welche Kenngröße gehört nicht in diese Aufzählung? (2)

- **O** Median
- **O** Quartil
- **O** Mittelwert
- **O** Modus
- **O** Interquartilsabstand

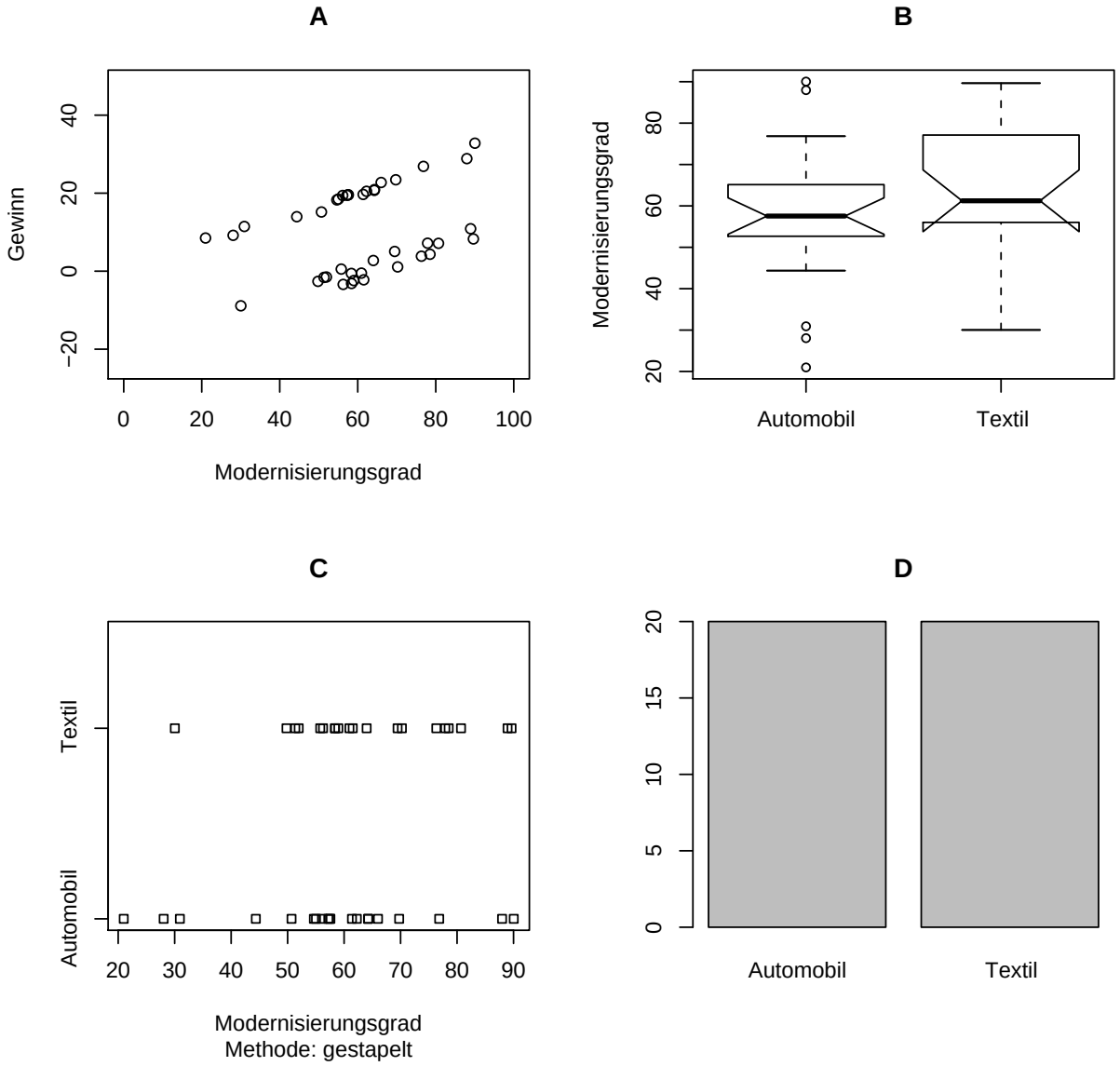

Abbildung 1: Statistische Gaphiken zum Datensatz *Firmen*

Hinweis: Die Aufgaben 4 und 5 beziehen sich auf die obigen Graphiken.

Aufgabe 4: Benennen Sie die Graphiken A und D! (2)

A: .....

D: .....

Aufgabe 5: Welche der folgenden Aussagen sind richtig? Nutzen Sie die Kreise zum ankreuzen! Tragen Sie **ausschließlich** dort, wo Sie ein Kreuz gemacht haben, den zugehörigen Buchstaben der statistischen Graphik ein, aus der der jeweilige Sachverhalt ihrer Meinung nach ersichtlich ist! (7)

O	☐	es gibt in den Daten Ausreißer
O	☐	es gibt in den Daten keine Ausreißer
O	☐	Gewinn und Modernisierungsgrad scheinen zusammenzuhängen
O	☐	Gewinn und Branche scheinen zusammenzuhängen
O	☐	die Modernisierungsgrade folgen Normalverteilungen
O	☐	es gibt in den Daten Bindungen
O	☐	es gibt in den Daten keine Bindungen
O	☐	die Modalwerte der Modernisierungsgrade sind in etwa gleich groß
O	☐	die Mediane der Modernisierungsgrade könnten übereinstimmen

Aufgabe 6: Welche Größen werden in einem QQ-Plot gegeneinander abgetragen? (2)

Tests

Aufgabe 7: Welcher Test wird üblicherweise verwendet, um zu überprüfen, ob eine Merkmal einer Normalverteilung folgt oder nicht? (1)

Aufgabe 8: Die Nullhypothese eines solchen Tests wird angenommen. Was ist dann zutreffend? (2)

- **O** der p-Wert ist kleiner als das Signifikanzniveau
- **O** der p-Wert ist größer als das Signifikanzniveau
- **O** das Merkmal ist nachweislich normalverteilt
- **O** das Merkmal ist nachweislich nicht normalverteilt

Aufgabe 9: Ordnen Sie den folgenden Tests ihre möglichen Alternativhypothesen zu! (2)

a:	Shapiro-Wilk-Test	☐	Mittelwerte sind ungleich
b:	Wilcoxen Rang Summen Test	☐	Lage der Verteilungen ist ungleich
c:	Zwei-Stichproben-t-Test	☐	X ist nicht normalverteilt

Hinweis: Nutzen Sie im folgenden Abschnitt sowohl die Liste mit statistischen Tests (siehe Seite 8 - 12) als auch ihr bisheriges Wissen über den Datensatz!

Man interessierte sich dafür, ob die Modernisierungsgrade in der Textilbranche statistisch signifikant von denen in der Automobilbranche abwichen...

Aufgabe 10: Welcher Aspekt (oder: Art von Parameter) soll getestet werden? (1)

Aufgabe 11: Was für eine Stichprobensituation liegt vor? (1)

Aufgabe 12: Welchen Test wenden Sie an? (1)

Aufgabe 13: Wie lautet die Nullhypothese dieses Tests? (Formel oder Prosa) (1)

Aufgabe 14: Welche Voraussetzungen stellt der von Ihnen verwendete Test allgemein an Datensätze? (4)

Aufgabe 15: Wie lautet der p-Wert zu diesem Test? (1)

Aufgabe 16: Was schlussfolgern Sie aus dem Testergebnis in Bezug auf die ursprüngliche Fragestellung? Formulieren Sie so, dass Sie jemand versteht, der keine Ahnung von (noch Interesse an) Statistik besitzt! (2)

Aufgabe 17: Welcher Problematik tritt man mit der Bonferroni-Korrektur entgegen? (2)

Verzeichnis der Tests

a) Shapiro-Wilk-Test für den Modernisierungsgrad
b) Shapiro-Wilk-Test für Modernisierungsgrade: Automobilbranche
c) Shapiro-Wilk-Test für Modernisierungsgrade: Textilbranche
d) Shapiro-Wilk-Test für Differenzen der Modernisierungsgrade
e) Ein-Stichproben t-Test: Automobilbranche
f) Ein-Stichproben t-Test: Textilbranche
g) Varianzanalyse (ANOVA)
h) Welchs t-Test
i) Zwei-Stichproben t-Test
j) t-Test für gepaarte Stichproben
k) Wilcoxen-Rang-Summen-Test
l) Wilcoxon-Vorzeichen-Rang-Test
m) Fligner-Test
n) Bartlett-Test

a) **Shapiro-Wilk-Test für den Modernisierungsgrad**

```
> shapiro.test( Modern )

Shapiro-Wilk normality test

data:  Modern
W = 0.9512, p-value = 0.08324
```

b) **Shapiro-Wilk-Test für Modernisierungsgrade: Automobilbranche**

```
> shapiro.test( Modern[Branche == "Automobil"] )

Shapiro-Wilk normality test

data:  Modern[Branche == "Automobil"]
W = 0.946, p-value = 0.3105
```

c) **Shapiro-Wilk-Test für Modernisierungsgrade: Textilbranche**

```
> shapiro.test( Modern[Branche == "Textil"] )

Shapiro-Wilk normality test

data:  Modern[Branche == "Textil"]
W = 0.9587, p-value = 0.5189
```

d) **Shapiro-Wilk-Test für Differenzen der Modernisierungsgrade**

```
> shapiro.test( Modern[Branche == "Automobil"] - Modern[Branche == "Textil"] )

Shapiro-Wilk normality test

data:  Modern[Branche == "Automobil"] - Modern[Branche == "Textil"]
W = 0.853, p-value = 0.005978
```

e) **Ein-Stichproben t-Test: Automobilbranche**

```
> t.test( Modern[Branche == "Automobil"] , mu = 0)

One Sample t-test

data:  Modern[Branche == "Automobil"]
t = 14.7254, df = 19, p-value = 7.615e-12
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 49.57932 66.00868
sample estimates:
mean of x
   57.794
```

f) **Ein-Stichproben t-Test: Textilbranche**

```
> t.test( Modern[Branche == "Textil"], mu = 0)

One Sample t-test

data:  Modern[Branche == "Textil"]
t = 19.6922, df = 19, p-value = 4.221e-14
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 57.6065 71.3085
sample estimates:
mean of x
  64.4575
```

g) **Einfache Varianzanalyse (ANOVA)**

```
> anova( lm( Modern ~ Branche, data = Firmen ) )
Analysis of Variance Table

Response: Modern
          Df Sum Sq Mean Sq F value Pr(>F)
Branche    1  444.0   444.0  1.7001 0.2001
Residuals 38 9924.9   261.2
```

h) **Welchs t-Test**

```
> t.test( Modern[Branche == "Automobil"], Modern[Branche == "Textil"] )

Welch Two Sample t-test
data:  Modern[Branche == "Automobil"] and Modern[Branche == "Textil"]
t = -1.3039, df = 36.813, p-value = 0.2004
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.020318   3.693318
sample estimates:
mean of x mean of y
  57.7940   64.4575
```

i) **Zwei-Stichproben t-Test**

```
> t.test( Modern[Branche == "Automobil"], Modern[Branche == "Textil"]
          , var.equal = TRUE )

Two Sample t-test
data:  Modern[Branche == "Automobil"] and Modern[Branche == "Textil"]
t = -1.3039, df = 38, p-value = 0.2001
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.009352   3.682352
sample estimates:
mean of x mean of y
  57.7940   64.4575
```

j) **t-Test für gepaarte Stichproben**

```
> t.test( Modern[Branche == "Automobil"], Modern[Branche == "Textil"]
        , paired=TRUE )

Paired t-test
data:  Modern[Branche == "Automobil"] and Modern[Branche == "Textil"]
t = -4.9381, df = 19, p-value = 9.13e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.487822 -3.839178
sample estimates:
mean of the differences
                -6.6635
```

k) **Wilcoxen-Rang-Summen-Test**

```
> wilcox.test( Modern[Branche == "Automobil"], Modern[Branche == "Textil"] )

Wilcoxon rank sum test

data:  Modern[Branche == "Automobil"] and Modern[Branche == "Textil"]
W = 158, p-value = 0.2648
alternative hypothesis: true mu is not equal to 0
```

l) **Wilcoxon-Vorzeichen-Rang-Test**

```
> wilcox.test( Modern[Branche == "Automobil"], Modern[Branche == "Textil"]
              , paired = TRUE )

Wilcoxon signed rank test

data:  Modern[Branche == "Automobil"] and Modern[Branche == "Textil"]
V = 1, p-value = 3.815e-06
alternative hypothesis: true mu is not equal to 0
```

m) **Fligner-Test**

```
> fligner.test( list( Modern[Branche == "Automobil"]
                , Modern[Branche == "Textil"]  ))

Fligner-Killeen test for homogeneity of variances

data:  list(Modern[Branche == "Automobil"], Modern[Branche == "Textil"])
Fligner-Killeen:med chi-squared = 0.0631, df = 1, p-value = 0.8017
```

n) **Bartlett-Test**

```
> bartlett.test( list( Modern[Branche == "Automobil"]
                     , Modern[Branche == "Textil"]  ))

Bartlett test for homogeneity of variances

data:  list(Modern[Branche == "Automobil"], Modern[Branche == "Textil"])
Bartlett's K-squared = 0.6067, df = 1, p-value = 0.436
```

Lineare Regression

Im Folgenden finden Sie das Ergebnis einer linearen Regression, mit der man versuchte, den Gewinn der Unternehmen aus ihrem Modernisierungsgrad abzuleiten...

```
> regmod <- lm( Gewinn ~ Modern , data = Firmen )
> regmod

Call:
lm(formula = Gewinn ~ Modern, data = Firmen)

Coefficients:
(Intercept)       Modern
    -1.7061       0.1972


> anova(regmod)
Analysis of Variance Table

Response: Gewinn
          Df Sum Sq Mean Sq F value  Pr(>F)
Modern     1  403.1   403.1  3.7573 0.06003
Residuals 38 4077.2   107.3


> cat("R^2 =", (var(Gewinn) - var(resid(regmod))) / var(Gewinn), "\n" )

R^2 = 0.08998015
```

Aufgabe 18: Geben Sie die Regressionsgleichung mit den geschätzten Parameterwerten an! (2)

```
> par( mfrow=c(1,2) )
> plot( Firmen$Modern, Firmen$Gewinn ,asp=1 , xlim=c(0,100), pch=3
       ,xlab="Modernisierungsgrad", ylab="Gewinn")
> points( Firmen$Modern, predict(regmod) ,pch=2,cex=1.5)
> legend("bottomleft", legend=c("Daten","Vorhersage"), pch=c(3,2), bty="n")
> plot( predict(regmod), resid(regmod), ylab="Residuen" )
```

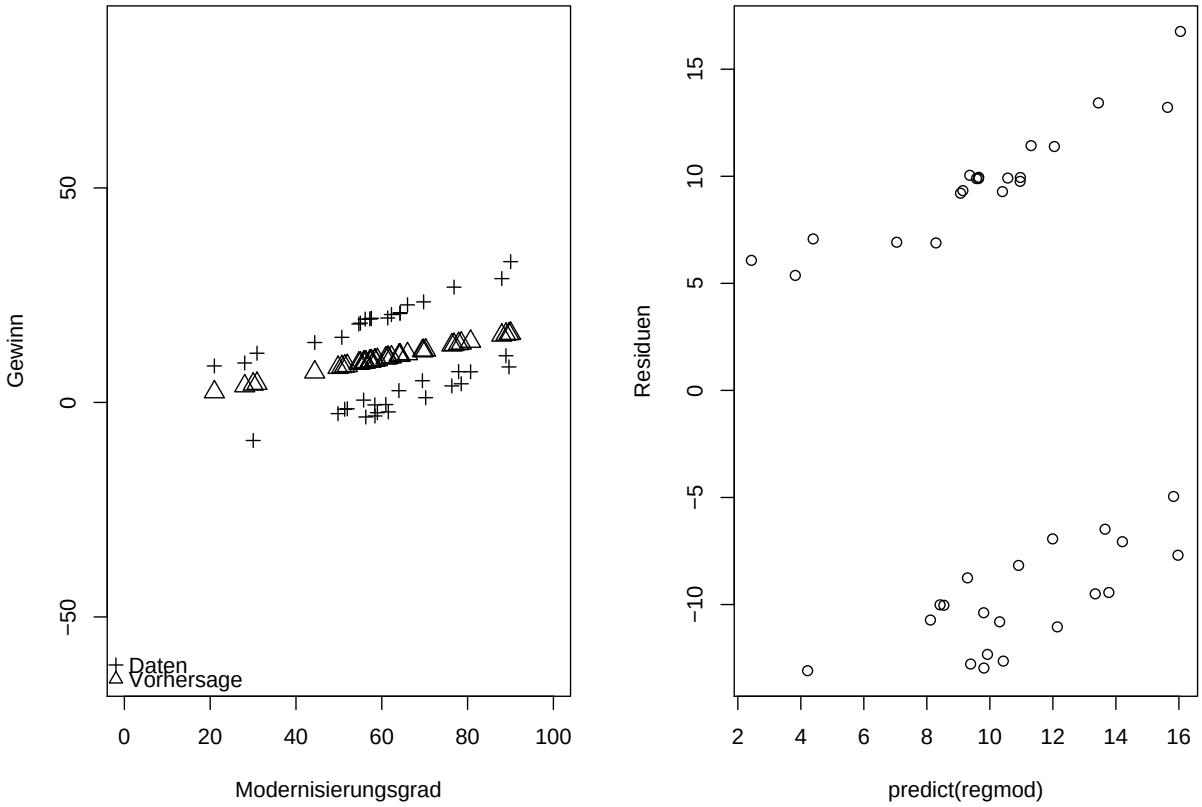

Abbildung 2: Statistische Graphiken zur linearen Regression

Aufgabe 19: Sehen die Residuen des Modells *regmod* so aus, wie man sich das wünschen würde? Erläutern Sie! (2)

Aufgabe 20: Was gibt die Hebelwirkung an? (2)

Lineares Modell

Da die lineare Regression nicht die gewünschten Ergebnisse erzielte, versuchte man nun die verschiedenen Branchen mit zu berücksichtigen:

```
> mod <- lm( Gewinn ~ Modern + Branche + Modern * Branche, data = Firmen)
> mod

Call:
lm(formula = Gewinn ~ Modern + Branche + Modern * Branche, data = Firmen)

Coefficients:
         (Intercept)               Modern         BrancheTextil
            -0.13783              0.33943             -18.83020
Modern:BrancheTextil
            -0.02633

> anova(mod)
Analysis of Variance Table

Response: Gewinn
               Df Sum Sq Mean Sq   F value Pr(>F)
Modern          1  403.1   403.1  205.4571 <2e-16 ***
Branche         1 4004.9  4004.9 2041.0550 <2e-16 ***
Modern:Branche  1    1.7     1.7    0.8481 0.3632
Residuals      36   70.6     2.0

> cat("R^2 =", (var(Gewinn) - var(resid(mod))) / var(Gewinn), "\n" )

R^2 = 0.9842338
```

Aufgabe 21: Kreuzen Sie an, welcher Formel das Modell *mod* entspricht: (2)

- **O** $Gewinn = a * Modern + b_{Branche} * Modern$
- **O** $Gewinn = a + b * Branche + c_{Modern} + d_{Modern} * Branche + \epsilon$
- **O** $Gewinn = a + b * Modern + c_{Branche} * Modern + \epsilon$
- **O** $Gewinn = a + b * Modern + c_{Branche} + d_{Branche} * Modern + \epsilon$
- **O** $Gewinn = a + b * Modern + c_{Branche} + \epsilon$
- **O** $Gewinn = a * Modern + b_{Branche} + c_{Branche} * Modern + \epsilon$
- **O** $Gewinn = a + b * Branche + c_{Modern} + d_{Modern} * Branche$

Aufgabe 22: Wie groß ist der Anteil der Varianz der abhängigen Variable, der durch das Modell *mod* erklärt wird ? (1)

Aufgabe 23: Welchen Gewinn würden Sie nach dem Modell *mod* für ein Textilunternehmen mit einem Modernisierungsgrad von 60% vorhersagen? Es genügt, die entsprechende Berechnungsformel mit den eingesetzten Werten anzugeben. (2)

Aufgabe 24: Würden Sie den Gewinn mit dem Modell *mod* vorhersagen, oder würden Sie empfehlen, das Modell noch abzuändern? Begründen Sie in jedem Fall! (2)

Aufgabe 25: (Zusatzaufgabe) Warum kann die Hinzunahme des Merkmals *Bedarf* zum Modell *mod* oder auch zum Modell *regmod* nicht zu einer wesentlichen Verbesserung der Schätzung führen? (2)
