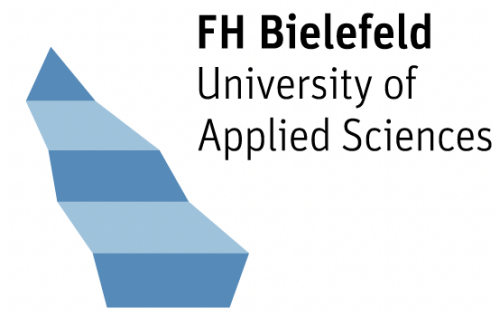
Klausur **Statistik zur Datenanalyse** 04.03.2023

Bearbeitungszeit: 90 Minuten
Zugelassene Hilfsmittel: nicht-programmierbarer Taschenrechner, handbeschriebener DIN A4 Zettel Vorder- und Rückseite



Name:

Matrikelnummer:

Ich versichere, dass ich gesundheitlich in der Lage bin, die Klausur zu schreiben.
Unterschrift:

Bevor Sie beginnen:

- Kontrollieren Sie Ihre Aufgabenblätter auf Vollständigkeit.
- Lesen Sie sich alle Aufgaben durch und verschaffen Sie sich einen Überblick über den Umfang der Aufgaben.
- Stellen Sie sicher, dass Ihre Antworten vollständig und lesbar sind. Ergänzen Sie Ihre Erläuterungen ggf. durch Skizzen.
- Schreiben Sie nur auf dem ausgeteilten Papier. Eigene zusätzliche Blätter dürfen nicht benutzt werden. Werden zusätzliche Arbeitsblätter benötigt, melden Sie sich bei der Klausuraufsicht. Die Heftung der Klausurblätter ist nicht zu lösen.
- Die Verwendung von Telekommunikationsmitteln wie z.B. Handy, Smartphone, Smart Watch, etc. ist während der der Bearbeitungszeit nicht erlaubt. Jeder Täuschungsversuch führt ohne vorhergehende Warnung zur sofortigen Abgabe der Klausur und wird mit der Note 5.0 bewertet.

Viel Erfolg!
Punktetabelle:

Aufgabe	1 (20 P.)	2 (10 P.)	3 (20 P.)	4 (20 P.)	5 (20 P.)	Summe
Punkte						

```
Call:
lm(formula = Volume ~ Diameter + Height, data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.182027 -0.074603 -0.004944  0.061530  0.240610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.64166    0.24524  -6.694 2.89e-07 ***
Diameter     5.25283    0.29572  17.763  < 2e-16 ***
Height       0.03144    0.01212   2.593   0.0149 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1102 on 28 degrees of freedom
Multiple R-squared:  0.9477,    Adjusted R-squared:  0.9439
F-statistic: 253.5 on 2 and 28 DF,  p-value: < 2.2e-16
```

Abbildung 1: 1. Aufgabe: Ausgabe Schätzergebnis aus der Statsitik-Software **R**.

1. (20 Punkte) Gegeben ist ein Datensatz mit 31 Beobachtungen von Obstbäumen (Schwarzkirsche). Es sind die drei Variablen gegeben:

   - Volumen (Volume) des Baumes, gemessen in Kubikmeter (m^3).
   - Durchmesser (Diameter) des Baumes, gemessen in Meter (m).
   - Höhe (Height) des Baumes, gemessen in Meter (m).

   Für das multiple lineare Regressionsmodell

   $$Volume_i = \beta_0 + \beta_1 Diameter_i + \beta_2 Height_i + u_i$$

   erhalten Sie das Schätzergebnis in der Statistik-Software **R**, das in Abbildung 1 zu sehen ist.

   (a) Machen Sie bitte eine Aussage zur Güte des Modells.

   (b) Interpretieren Sie den quantitativen Effekt von $Diameter$

   (c) Stellen Sie die Nullhypothese und Alternativhypothese auf für den t-Test auf statistische Signifikanz von $Height$. Geben Sie die Prüfgröße an und interpretieren Sie das Ergebnis.

```
Call:
lm(formula = log(Volume) ~ log(Diameter) + log(Height), data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.169537 -0.048572  0.004428  0.063542  0.129237

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.57663    0.69137   -2.28   0.0304 *
log(Diameter)  1.98371    0.07522   26.37  < 2e-16 ***
log(Height)    1.11439    0.20485    5.44 8.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08154 on 28 degrees of freedom
Multiple R-squared:  0.9776,    Adjusted R-squared:  0.976
F-statistic: 610.4 on 2 and 28 DF,  p-value: < 2.2e-16
```

Abbildung 2: 2. Aufgabe: Ausgabe Schätzergebnis aus der Statsitik-Software **R**.

2. (10 Punkte) Mit den Daten aus Aufgabe 1 ist ein weiteres Modell geschätzt worden.

$$\log(Volume_i) = \beta_0 + \beta_1 \log(Diameter_i) + \beta_2 \log(Height_i) + u_i$$

Die Ergebnisse sind in Abbildung 2 zu sehen.

   (a) In der Praxis werden häufig Variablentransformationen durchgeführt (z.B. durch den Einsatz von log()), bevor lineare Modelle geschätzt werden. Warum wird das gemacht?

   (b) Machen Sie bitte eine Aussage, ob und wie die beiden Modelle aus Aufgabe 1 und Aufgabe 2 miteinander vergleichbar sind.

3. (20 Punkte) In der multiplen Regression ist der Forward-Algorithmus eine Möglichkeit, um systematisch, insbesondere bei einer hohen Anzahl an erklärenden Variablen, ein Modell mit hoher Güter aus der Vielzahl an möglichen Modellen zu bestimmen. Erklären Sie die dafür notwendigen Schritte. Gehen Sie insbesondere auf die Verwendung und Interpretation des korrigierten R^2 ein.

4. (20 Punkte) Empirische Korrelation
   (a) Was misst die empirische Korrelation?
   (b) Wie ist die empirische Korrelation zu interpretieren?
   (c) Wie wird die empirische Korrelation von zwei metrischen Variablen X und Y berechnet?
   (d) Die Variablen X und Y besitzen eine metrische Skala. Was ist die Besonderheit einer Ordinalskala?
   (e) Wie kann eine Korrelation berechnet werden, wenn zwei ordinal-skalierte Variablen Z und V vorliegen? Was misst diese berechnete Korrelation?

5. (20 Punkte) In einer Qualitätskontrolle wird überprüft, ob ein Produktionsprozess eines Bauteils noch gemäß der Annahme einer Normalverteilung mit einem Sollwert von $\mu_x = 15mm$ und einer Standardabweichung von $\sigma_x = 0,5mm$ produziert. Dafuür wird in regelmäßigen Abständen eine Stichprobe der Größe $n = 25$ von den Bauteilen aus der Produktion gezogen. Der Mittelwert der Stichprobe liegt bei $15,25mm$.
   (a) Weicht die Länge des Bauteils statistisch signifikant zum Niveau $\alpha = 0,05$ vom Sollwert ab? Es gilt für die Verteilungsfunktion $\Phi()$ der Standardnormalverteilung:
   $$\Phi(-1,96) = 0,025, \Phi(0,015) = 0,506, \Phi(1,65) = 0,95$$
   (b) Was sind die Fehler 1. Art und 2. Art bei statistischen Hypothesentests?