

FAKULTÄT FÜR  
WIRTSCHAFTS- UND SOZIALWISSENSCHAFTEN  
DER  
RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG



Take-Home-Exam  
**Wirtschafts- und Sozialstatistik**

– Aufgaben –

Sommersemester 2022

Dr. Eduard Brüll

Bearbeitungszeitraum:  
Montag, 3. Oktober 2022 (00:00 Uhr morgens)  
bis Montag, 10. Oktober 2022 (23:59 Uhr abends)

**Anmerkungen:**

1. Kommentieren Sie in Ihrem do-file jede Teilaufgabe mit einem Antwortsatz! (z.B. „Das arithmetische Mittel des Merkmals  $X$  beträgt ... Einheiten.“)
2. Soweit die Aufgabenstellung nicht explizit vorgibt, wie eine Berechnung durchgeführt werden soll, steht es Ihnen frei, in Stata verfügbare Befehle zu verwenden.

## Aufgabe 1

Der Datensatz `RouseCollegeDistance.dta` enthält Informationen zu Studierenden in den USA.<sup>1</sup> Dabei steht im Fokus, wie der Bildungserfolg von Schülerinnen und Schülern von ihrer Herkunft abhängt.

Variable	Beschreibung der Variable (Definition)
<code>yrsed</code>	Bildungsjahre
<code>female</code>	Geschlecht (1 = weiblich, 0 = männlich)
<code>bytest</code>	Base Year Composite Testergebnis
<code>dadcoll</code>	Bildung Vater (1= wenn Vater Collegeabschluss hat)
<code>momcoll</code>	Bildung Mutter (1= wenn Mutter Collegeabschluss hat)
<code>incomehi</code>	Familieneinkommen (1 = > \$25,000 per year)
<code>ownhome</code>	Familie besitzt Haus (1= Familie besitzt Haus)
<code>dist</code>	Distanz zum vierjährigen College (in 10.000 Meilen)

Tabelle 1: Variablen im Datensatz `RouseCollegeDistance.dta`.

- a) Importieren Sie den Datensatz in Stata und vergeben Sie Labels. Die Variablen müssen den gleichen Namen und die gleichen Labels (Beschreibung der Variable, ohne Definition) wie in Tabelle 1 haben. Achten Sie darauf, dass alle Variablen in numerischem Format sind. (3 Punkte)
- b) Benennen Sie die Variable `female` in `gender` um. Wie hoch ist der Anteil weiblicher Studierender in diesem Datensatz? (2 Punkte)
- c) Im Folgenden interessieren wir uns für die Bildungsjahre der befragten Studierenden.
  - i) Berechnen Sie die durchschnittlichen Bildungsjahre. (1 Punkt)
  - ii) Geben Sie die Spannweite, den Modalwert, den Median und den Interquartilsabstand für die Bildungsjahre an. Welche dieser Maßzahlen eignet sich am Besten, um die Verteilung der Bildungsjahre zu beschreiben? (5 Punkte)
  - iii) Erstellen Sie eine Dummy-Variable, die angibt, ob das Individuum mindestens einen Bachelorabschluss erreicht hat (=1) oder nicht (=0). Wie hoch ist der Anteil der Individuen mit mindestens einem Bachelorabschluss? (2 Punkte)  
*Hinweis: Ein Studierender erreicht einen Bachelorabschluss nach 16 Bildungsjahren.*
  - iv) Lassen Sie sich die bedingte relative Verteilung der zuvor in (iii) erstellten Variable Bachelorabschluss gegeben das Bildungsniveau des Vaters ausgeben. Interpretieren Sie die Häufigkeitstabelle. (3 Punkte)
  - v) Berechnen Sie auf Basis der Tabelle in der vorherigen Teilaufgabe den  $\chi^2$ -Wert sowie Cramer's  $V$ . Interpretieren Sie beide Werte. (4 Punkte)
- d) Im Folgenden interessieren wir uns für die Ergebnisse des Base Year Composite Test Scores.
  - i) Berechnen Sie die Varianz und den Variationskoeffizienten der Ergebnisse des Tests. (2 Punkte)
  - ii) Erstellen Sie einen Boxplot über die Ergebnisse des Tests getrennt für Schüler, deren Eltern ein Eigenheim besitzen und für Eltern, die kein Eigenheim besitzen. Interpretieren und vergleichen Sie die beiden Boxplots. (3 Punkte)

<sup>1</sup>Es handelt sich hierbei um eine bereits vereinfachte und bereinigte Version des Datensatzes zu dem Artikel von Cecilia Rouse "Democratization or Diversion? The Effect of Community Colleges on Educational Attainment," *Journal of Business and Economics Statistics* April 1995, Vol. 12, No.2, S. 217-224.

- iii) Erstellen Sie ein Histogramm der Variable `bytest`. Färben Sie das Histogramm in rot ein und markieren Sie den Median und das arithmetische Mittel mit jeweils einer vertikalen Linie im Histogramm. Begründen Sie kurz, ob es sich um eine symmetrische, links- oder rechtsschiefe Verteilung handelt. (5 Punkte)  
*Hinweis: Verwenden Sie den `twoway`-Befehl.*

## Aufgabe 2

Der Datensatz `Logistik.dta` enthält folgende Informationen zum Umsatz der 20 von größten Logistikunternehmen in Deutschland im Jahr 2019 in Millionen Euro.<sup>2</sup>

- a) Importieren Sie den Datensatz in Stata. (1 Punkt)
- b) Welches Logistikunternehmen hatte 2019 den höchsten Umsatz? Lassen Sie sich das Unternehmen per Befehl ausgeben. (1 Punkt)
- c) Berechnen Sie den Median der Umsätze der Logistikunternehmen. (1 Punkt)
- d) Erstellen Sie die Variable  $B$  (=Merkmalssumme), die den gesamten Umsatz der Unternehmen im Datensatz angibt. (1 Punkt)
- e) Berechnen Sie die Merkmalsanteile an der Merkmalssumme für die Logistikumsätze 2019. Lassen Sie sich den Wert für UPS ausgeben und interpretieren Sie diesen. (2 Punkte)
- f) Sortieren Sie die Umsätze absteigend und bilden Sie die kumulierte Anzahl der Merkmals-träger  $i$  sowie die kumulierten relativen Umsätze  $C$ . Versehen Sie die Variablen mit den entsprechenden Labeln. (4 Punkte)
- g) Berechnen Sie den Herfindahl-Index für die Umsätze in der Logistikbranche **ohne Verwendung** des `hhi`-Befehls, d.h. verwenden Sie die Formel aus der Vorlesung. Interpretieren Sie den ermittelten Wert inhaltlich und gehen Sie dabei auf den Wertebereich des Herfindahl-Index im vorliegenden Beispiel ein. (4 Punkte)
- h) Berechnen Sie die hypothetischen Umsätze der Unternehmen, wenn der Herfindahl-Index den Wert  $H = 0,05$  annehmen würde. (2 Punkte)
- i) Nehmen Sie an, dass im Jahr 2020 die Umsätze aller Logistikunternehmen um 20% gestiegen sind. Erstellen Sie eine Variable für die fiktiven Umsätze 2020 und berechnen Sie den Herfindahl-Index erneut **unter Verwendung** des `hhi`-Befehls. Begründen Sie kurz, warum Sie den Wert erhalten und welche Unterschiede es zum Ergebnis in Teilaufgabe (h) gibt. (2,5 Punkte)
- j) Nehmen Sie an, dass im Jahr 2020 die Umsätze aller Logistikunternehmen um 38 Millionen Euro gestiegen sind. Erstellen Sie eine Variable für die fiktiven Umsätze 2020 und berechnen Sie den Herfindahl-Index erneut **unter Verwendung** des `hhi`-Befehls. Erläutern Sie kurz, warum Sie den Wert erhalten und welche Unterschiede es zum Ergebnis in Teilaufgabe (h) gibt. (2,5 Punkte)
- k) Berechnen Sie den Gini-Koeffizient für die Umsätze 2019 **ohne Verwendung** des `fastgini`-Befehls, d.h. nutzen Sie die Formel aus der Vorlesung. Interpretieren Sie diesen Wert in Bezug auf dessen Wertebereich und erläutern Sie, was dieser in dieser Anwendung aussagt. (6 Punkte)
- l) Berechnen Sie den Gini-Koeffizienten erneut **unter Verwendung** des `fastgini`-Befehls für die fiktiven Situationen in den Teilaufgaben (i) und (j). Erläutern Sie kurz, warum Sie den Wert erhalten und welche Unterschiede es zum Ergebnis in Teilaufgabe (k) gibt. (4 Punkte)
- m) Erstellen Sie ein Kreisdiagramm mit Beschriftung, welches die Verteilung der Umsätze in der Logistikbranche veranschaulicht. Gruppieren Sie zur Veranschaulichung die fünf kleinsten Logistikunternehmen als ein 'Sonstiges' Logistikunternehmen. (4 Punkte)

---

<sup>2</sup>Quelle: Statista

### Aufgabe 3

Der Datensatz `urban_gdp.xlsx` enthält Informationen zu den Bruttoinlandsprodukten pro Kopf (GDP per capita) verschiedener Länder und dem Anteil der Bevölkerung, der in dem jeweiligen Land in urbanen Räumen lebt.<sup>3</sup>

- a) In diesem Aufgabenteil verschaffen Sie sich einen ersten Überblick über die Daten und bereiten den Datensatz für die Analyse vor.
- i) Importieren Sie den Datensatz `urban_gdp.xlsx`. Löschen Sie alle Beobachtungen, bei denen Daten fehlen. Achten Sie darauf, dass die Variablen in numerischem Format sind. (2 Punkte)
  - ii) Fügen Sie den jeweiligen Variablen die passenden Label hinzu. (1 Punkt)
  - iii) Berechnen Sie die Mittelwerte der beiden Variablen und speichern Sie die Werte für weitere Berechnungen. (2 Punkte)
  - iv) Berechnen Sie den Korrelationskoeffizienten nach Pearson per Hand anhand der Formel

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

und interpretieren Sie den berechneten Wert unter Berücksichtigung des Wertebereichs. (3 Punkte)

*Hinweis: Berechnen Sie zuerst den Zähler, dann den Nenner von  $r_{XY}$  und speichern Sie die Werte.*

- v) Berechnen Sie den Rang-Korrelationskoeffizienten nach Spearman ( $r_{XY}^R$ ) unter Verwendung des `spearman`-Befehls und vergleichen Sie den Wert mit der vorherigen Teilaufgabe. (2 Punkte)
- b) Nehmen Sie für die folgende Aufgabe an, dass der Zusammenhang zwischen dem Bruttoinlandsprodukt ( $Y$ ) und dem Anteil der Bevölkerung, der in urbanen Räumen lebt ( $X$ ) folgendermaßen dargestellt werden kann:

$$y_i = a + b \cdot x_i + u_i$$

- i) Erstellen Sie ein Streudiagramm mit  $X$  auf der horizontalen Achse und  $Y$  auf der vertikalen Achse. Ändern Sie die Farbe der Punkte im Streudiagramm in lila und beschriften Sie die Punkte mit dem Namen des Landes. Welches Vorzeichen vermuten Sie für den Koeffizienten  $b$  der Regressionsgeraden? (3 Punkte)
- ii) Bestimmen Sie die Werte von  $a$  und  $b$  gemäß der Methode der kleinsten Quadrate gemäß der folgenden Formel

$$b = \frac{s_{xy}}{s_x^2} \text{ und } a = \bar{y} - b \cdot \bar{x}$$

und interpretieren Sie die Koeffizienten. (4 Punkte)

- iii) Fügen Sie die geschätzte Regressionsgerade in das Streudiagramm ein. (1 Punkt)
- iv) Geben Sie anhand des Modells an, wie groß das prognostizierte BIP pro Kopf in Afghanistan ist. Berechnen Sie die Differenz zwischen dem erwarteten und dem tatsächlichen Wert und interpretieren Sie den berechneten Wert inhaltlich. (3 Punkte)

---

<sup>3</sup>Quelle: Weltbank

- c) In den nächsten Teilaufgaben beschäftigen Sie sich mit der Beurteilung der Aussagefähigkeit des in (b) geschätzten Modells.
- i) Berechnen Sie den vom geschätzten Modell vorhergesagten Wert für das Land mit dem geringsten Anteil der Bevölkerung, die in urbanen Räumen lebt. Was fällt Ihnen auf? *(2 Punkte)*
  - ii) Bestimmen Sie das Bestimmtheitsmaß für die Regression aus Teilaufgabe (b). *(1 Punkt)*
  - iii) Treffen Sie anhand der letzten beiden Teilaufgaben eine Aussage zur Erklärungskraft des Modells. Interpretieren Sie hierzu die berechneten Werte und fügen Sie diese für eine konsistente Begründung zusammen. *(3 Punkte)*
- d) Im Folgenden modifizieren wir das Modell und nehmen an, dass der Zusammenhang wie folgt beschrieben wird

$$\log(y_i) = a + b \cdot x_i + u_i$$

- i) Erstellen Sie eine Variable `logGDP`, die das logarithmierte GDP enthält. *(1 Punkt)*
- ii) Schätzen Sie nun das Modell mit `log(Y)` anstelle von `Y`. Bestimmen Sie die Werte von `a` und `b` mit dem `regress`-Befehl. *(1 Punkt)*
- iii) Nutzen Sie das Bestimmtheitsmaß aus dem vorherigen Output und vergleichen Sie dieses mit dem Wert in (c,ii). *(2 Punkte)*
- iv) Erstellen Sie ein Streudiagramm mit `log(Y)` und `X`. Fügen Sie die geschätzte Regressionsgerade in das Streudiagramm ein. *(2 Punkte)*  
*Hinweis: Nutzen Sie hierfür den `twoway`-Befehl und den `lfit`-Befehl.*
- v) Erläutern Sie anhand des Streudiagramms, warum sich das  $R^2$  verändert hat und welchen Effekt dies auf die Erklärungskraft des Modells hat. *(2 Punkte)*