Lösung zu Aufgabe 24

Die Beobachtungen aus der Tabelle werden als Realsierungen von Zufallsvariablen U_i, D_i interpretiert. Die Schätzung von a wird durch KQ-Ansätze gewonnen, wobei mehrere Möglichkeiten denkbar sind:

Erster Optimierungsansatz: Weil $U/d = \pi$, könnte man die Konstante a an die Verhältnisse $U_1/D_1, \ldots, U_n/D_n$ anpassen, d.h. den Ausdruck $\sum\limits_{i=1}^{n}(\frac{U_i}{D_i} - a)^2$ in a minimieren. Aufgrund der Optimalitätseigenschaften des arithmetischen Mittels (DuW) ist die KQ-Lösung das arithmetische Mittel

$$\hat{a}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{U_i}{D_i}$$

Aus der Tabelle der Aufgabenstellung erhält man $\hat{a}_1 \approx 3,133$.

Zweiter Optimierungsansatz: Passe ad an U an, d.h. minimiere

$$\sum_{i=1}^{n}(U_i - aD_i)^2 = \sum_{i=1}^{n}(U_i^2 - 2aU_iD_i + a^2D_i^2) = \sum_{i=1}^{n}U_i^2 - 2a\sum_{i=1}^{n}U_iD_i + a^2\sum_{i=1}^{n}D_i^2$$

Ableiten und gleich Null setzen ergibt

$$-2\sum_{i=1}^{n}U_iD_i + 2a\sum_{i=1}^{n}D_i^2 = 0 \Leftrightarrow a = \frac{\sum_{i=1}^{n}U_iD_i}{\sum_{i=1}^{n}D_i^2}$$

also den Schätzer

$$\hat{a}_2 = \frac{\sum_{i=1}^{n}U_iD_i}{\sum_{i=1}^{n}D_i^2}$$

Die betreffenden Produkt- und Quadratsummen bekommt man z.B. mit den Statistikfunktionen eines Taschenrechners und es ergibt sich die Schätzung

$$\hat{a}_2 = \frac{573740}{182254} \approx 3,14802$$

Weil ein anderer Optimierungsansatz vorliegt, erhält man - nicht unerwartet - ein anderes Ergebnis.

Wir haben hier also - je nach Optimierungsansatz - zwei mögliche Schätzer:

1. $\hat{a}_1 = \frac{1}{n}\sum\limits_{i=1}^{n}\frac{U_i}{D_i}$
2. $\hat{a}_2 = \frac{\sum\limits_{i=1}^{n}U_i}{\sum\limits_{i=1}^{n}D_i}$

Die Aufgabe ist durch Angabe eines der beiden Schätzer eigentlich gelöst, aber es bleibt die Frage, welcher dieser beiden Schätzer der geeignetere für das Problem ist.

Ohne genauere Untersuchung, insbesondere ohne weitere Modellannahmen an die den Daten zugrundeliegenden Zufallsvariablen ist es nicht möglich, dies zu beurteilen.

Mit den folgenden Zusatzannahmen kann man zumindest hinsichtlich Erwartungstreue und Schätzerrisiko einen Vergleich vornehmen. Es wird dabei angenommen, dass das Modell von 9-12 aus der Vorlesung mit $f(d) = ad$ gilt, d.h. dass

1. für die Beobachtungen U_i des Umfanges und D_i des Durchmessers der Zusammenhang $U_i = aD_i + V_i$ gilt und
2. die Fehlervariable V_i die Eigenschaften $E(V_i) = 0$ und $E(V_i^2) = var(V_i) = \sigma^2$ hat und zusätzlich
3. $D_1, \ldots, D_n, V_1, \ldots, V_n$ st.u. sind (diese Annahme ist etwas stärker als auf Folie 9-12 angenommen, dort sollten nur beobachtungsweise Regressoren und Fehler st.u. sein) und schließlich
4. $D_1, \ldots, D_n$ identisch verteilt sind und $E(D_1^2)$ existiert.

Die Ergebnisse der Analyse lauten vorwegggenommen:

1. $\hat{a}_1$ ist erwartungstreu für a.
2. $\hat{a}_2$ ist erwartungstreu für a.
3. $\hat{a}_1$ hat das Schätzerrisiko $E((\hat{a}_1 - a)^2) = \sigma^2 \cdot E(\frac{1}{nD_1^2})$
4. $\hat{a}_2$ hat das Schätzerrisiko $E((\hat{a}_1 - a)^2) = \sigma^2 \cdot E(\frac{1}{\sum\limits_{i=1}^{n} D_i^2})$
5. Es gilt $E(\frac{1}{\sum\limits_{i=1}^{n} D_i^2}) \leq E(\frac{1}{nD_1^2})$, also hat $\hat{a}_2$ gleichmäßig kleineres Schätzerrisiko als $\hat{a}_1$.

Der zweite Schätzer ist unter den Modellannahmen deshalb zu bevorzugen. Um wieviel besser der zweite Schätzer ist, hängt von den weiteren Verteilungseigenschaften der D_i ab, zu denen man hier aber nichts sagen kann.

Die Aussagen werden nun nacheinander nachgerechnet.

1. Erwartungstreue von Schätzer 1: Für die weitere Rechnung ist die folgende Darstellung des Schätzers hilfreich, welche die Modellgleichung $U_i = aD_i + V_i$ verwendet:

   $$\hat{a}_1 = \tfrac{1}{n} \sum_{i=1}^{n} \frac{U_i}{D_i} = \tfrac{1}{n} \sum_{i=1}^{n} \frac{aD_i + V_i}{D_i} = \tfrac{1}{n} \sum_{i=1}^{n} \frac{aD_i}{D_i} + \tfrac{1}{n} \sum_{i=1}^{n} \frac{V_i}{D_i} = a + \tfrac{1}{n} \sum_{i=1}^{n} \frac{V_i}{D_i}.$$

   Hinweis: Diese Darstellung kann natürlich nicht verwendet werden, um den Schätzer konkret auszurechnen, denn weder a noch $V_1, \ldots, V_n$ sind beobachtbar.

   Der Schätzer ist erwartungstreu, denn

   $$E(\hat{a}_1) \quad = \quad E(a + \tfrac{1}{n} \sum_{i=1}^{n} \frac{V_i}{D_i}) = a + E(\tfrac{1}{n} \sum_{i=1}^{n} \frac{V_i}{D_i}) = a + \tfrac{1}{n} \sum_{i=1}^{n} E(\frac{V_i}{D_i}) = a$$

   Denn für jeden Summand gilt

   $$E(\frac{V_i}{D_i}) = E(V_i \cdot \frac{1}{D_i}) \overset{*}{=} E(V_i) \cdot E(\frac{1}{D_i}) \overset{**}{=} 0 \cdot E(\frac{1}{D_i}) = 0$$

   Dabei wird in $*$ ausgenutzt: V_i und D_i sind stochastisch unabhängig, also auch V_i und $\frac{1}{D_i}$. Bei st.u. ZV X, Y gilt aber $E(XY) = E(X)E(Y)$. Beachte: Hier wird nicht $E(X/Y) = E(X)/E(Y)$ verwendet, diese Aussage ist auch falsch! In $**$ wird ausgenutzt, dass $E(V_i) = 0$ nach Modellannahme für den Fehler.

2. Erwartungstreue von Schätzer 2: Auch dieser Schätzer kann mittels $U_i = aD_i + V_i$ umgeschrieben werden:

$$\hat{a}_2 = \frac{\sum\limits_{i=1}^{n} U_i D_i}{\sum\limits_{i=1}^{n} D_i^2} = \frac{\sum\limits_{i=1}^{n} (aD_i + V_i) D_i}{\sum\limits_{i=1}^{n} D_i^2} = \frac{\sum\limits_{i=1}^{n} aD_i^2 + \sum\limits_{i=1}^{n} V_i D_i}{\sum\limits_{i=1}^{n} D_i^2} = a + \frac{\sum\limits_{i=1}^{n} V_i D_i}{\sum\limits_{i=1}^{n} D_i^2}$$

Damit gilt dann

$$E(\hat{a}_2) = E\Big(a + \frac{\sum\limits_{i=1}^{n} V_i D_i}{\sum\limits_{i=1}^{n} D_i^2}\Big) = a + E\Big(\frac{\sum\limits_{i=1}^{n} V_i D_i}{\sum\limits_{i=1}^{n} D_i^2}\Big) = a$$

denn der zweite Summand ist (gleiche Argumente $*$,$**$ wie bei $\hat{a}_1$)

$$E\Big(\sum_{i=1}^{n} V_i \cdot \frac{D_i}{\sum_{i=1}^{n} D_i^2}\Big) = \sum_{i=1}^{n} E\Big(V_i \cdot \frac{D_i}{\sum_{i=1}^{n} D_i^2}\Big) \overset{*}{=} \sum_{i=1}^{n} E(V_i) \cdot E\Big(\frac{D_i}{\sum_{i=1}^{n} D_i^2}\Big) \overset{**}{=} \sum_{i=1}^{n} 0 \cdot E\Big(\frac{D_i}{\sum_{i=1}^{n} D_i^2}\Big) = 0$$

3. Zum Schätzerrisiko von $\hat{a}_1$: Aufgrund der Erwartungstreue ist das Risiko gleich der Schätzervarianz, diese ist $var(\hat{a}_1) = E(\hat{a}_1^2) - (E(\hat{a}_1))^2 = E(\hat{a}_1^2) - a^2$ und es ist noch zu berechnen

$$\begin{aligned} E(\hat{a}_1^2) &= E\Big(\big(a + \tfrac{1}{n}\sum_{i=1}^{n} \frac{V_i}{D_i}\big)^2\Big) \\ &= E\Big(a^2 + 2a\tfrac{1}{n}\sum_{i=1}^{n} \frac{V_i}{D_i} + \big(\tfrac{1}{n}\big)^2\big(\sum_{i=1}^{n} \frac{V_i}{D_i}\big)^2\Big) \\ &= a^2 + 2aE\Big(\tfrac{1}{n}\sum_{i=1}^{n} \frac{V_i}{D_i}\Big) + \tfrac{1}{n^2}\sum_{k,\ell=1}^{n} E\Big(\frac{V_k V_\ell}{D_k D_\ell}\Big) \end{aligned}$$

Dabei gilt,

a) wie oben bei der Erwartungstreue nachgerechnet, $E(\frac{1}{n}\sum\limits_{i=1}^{n} \frac{V_i}{D_i}) = 0$ und

b) für $k \neq \ell$: $E(\frac{V_k V_\ell}{D_k D_\ell}) = E(V_k)E(V_\ell)E(\frac{1}{D_k D_\ell}) = 0$ und

c) für $k = \ell$: $E(\frac{V_k V_\ell}{D_k D_\ell}) = E(\frac{V_k^2}{D_k^2}) = E(V_k^2)E(\frac{1}{D_k^2}) = \sigma^2 E(\frac{1}{D_k^2}) = \sigma^2 E(\frac{1}{D_1^2})$

also bleiben von den n^2 Summanden nur die n Stück mit $\ell = k$ übrig; es folgt

$E(\hat{a}_1^2) = a^2 + \frac{1}{n^2}\sum\limits_{k=1}^{n} \sigma^2 E(\frac{1}{D_k^2}) = a^2 + \sigma^2 \frac{1}{n^2}\sum\limits_{k=1}^{n} E(\frac{1}{D_k^2}) = a^2 + \sigma^2 \frac{1}{n^2} \cdot n \cdot E(\frac{1}{D_1^2}) = a^2 + \sigma^2 E(\frac{1}{nD_1^2})$

denn die D_i sind identisch verteilt. Es folgt $var(\hat{a}_1) = \sigma^2 \cdot E(\frac{1}{nD_1^2})$

Weiter lässt sich der Ausdruck nicht vereinfachen.

4. Zum Schätzerrisiko von $\hat{a}_2$: Auch hier ist wegen der Erwartungstreue Schätzerrisiko gleich Schätzervarianz und diese ist $var(\hat{a}_2) = E(\hat{a}_2^2) - a^2$ und der erste Term ist

$$\begin{aligned} E(\hat{a}_2^2) &= E\Big(\Big(a + \frac{\sum_{i=1}^{n} V_i D_i}{\sum_{i=1}^{n} D_i^2}\Big)^2\Big) = E\left(a^2 + 2a \cdot \frac{\sum_{i=1}^{n} V_i D_i}{\sum_{i=1}^{n} D_i^2} + \left(\frac{\sum_{i=1}^{n} V_i D_i}{\sum_{i=1}^{n} D_i^2}\right)^2\right) \\ &= a^2 + E\left(\left(\frac{\sum_{i=1}^{n} V_i D_i}{\sum_{i=1}^{n} D_i^2}\right)^2\right) \end{aligned}$$

denn der mittlere Summand ist Null (wie bei der Erwartungstreue von $\hat{a}_2$). Der rechte Summand ergibt aber nun $\sigma^2 \cdot E(\frac{1}{\sum_{i=1}^n D_i^2})$ wie folgt (die Stellen $*$,$**$ wieder mit den gleichen Argumenten wie oben)

$$\begin{aligned}
E(\left(\frac{\sum\limits_{i=1}^n V_i D_i}{\sum\limits_{i=1}^n D_i^2}\right)^2) &= E(\frac{\sum\limits_{k,\ell}^n V_k V_\ell D_k D_\ell}{(\sum\limits_{i=1}^n D_i^2)^2}) \\
&= E(\frac{\sum\limits_{k=1}^n V_k^2 D_k^2 + \sum\limits_{k\neq\ell}^n V_k V_\ell D_k D_\ell}{(\sum\limits_{i=1}^n D_i^2)^2}) \\
&= E(\sum_{k=1}^n V_k^2 \frac{D_k^2}{(\sum\limits_{i=1}^n D_i^2)^2} + \sum_{k\neq\ell}^n V_k V_\ell \frac{D_k D_\ell}{(\sum\limits_{i=1}^n D_i^2)^2}) \\
&\overset{*}{=} \sum_{k=1}^n E(V_k^2) \cdot E(\frac{D_k^2}{(\sum\limits_{i=1}^n D_i^2)^2}) + \sum_{k\neq\ell}^n \underbrace{E(V_k)E(V_\ell)}_{=0\cdot 0} E(\frac{D_k D_\ell}{(\sum\limits_{i=1}^n D_i^2)^2}) \\
&\overset{**}{=} \sum_{k=1}^n E(V_k^2) \cdot E(\frac{D_k^2}{(\sum\limits_{i=1}^n D_i^2)^2}) = \sum_{k=1}^n \sigma^2 \cdot E(\frac{D_k^2}{(\sum\limits_{i=1}^n D_i^2)^2}) \\
&= \sigma^2 \cdot E(\frac{\sum\limits_{k=1}^n D_k^2}{(\sum\limits_{i=1}^n D_i^2)^2}) \\
&= \sigma^2 E(\frac{1}{\sum\limits_{i=1}^n D_i^2})
\end{aligned}$$

Insgesamt folgt $var(\hat{a}_2) = a^2 + \sigma^2 E(\frac{1}{\sum\limits_{i=1}^n D_i^2}) - a^2 = \sigma^2 E(\frac{1}{\sum\limits_{i=1}^n D_i^2})$

5. Zum Schätzervergleich anhand der Schätzervarianz:

Betrachte die Funktion $h(x) = \frac{1}{x}$ für $x > 0$. h ist konvex, d.h. für alle $x, y \in \mathbb{R}$ und $\alpha \in [0;1]$ gilt $h(\alpha x + (1-\alpha)y) \leq \alpha h(x) + (1-\alpha)h(y)$. Dies kann auf n Werte $x_1, \ldots, x_n > 0$ und Faktoren $\alpha_1, \ldots, \alpha_n \in [0;1]$ mit $\alpha_1 + \cdots + \alpha_n = 1$ übertragen werden, d.h. es gilt

$$h(\alpha_1 x_1 + \cdots + \alpha_n x_n) \leq \alpha_1 h(x_1) + \cdots + \alpha_n h(x_n)$$

Speziell gilt mit $\alpha_1 = \cdots = \alpha_n = \frac{1}{n}$

$$h(\frac{x_1 + \cdots + x_n}{n}) \leq \frac{h(x_1) + \cdots + h(x_n)}{n}$$

Setzt man nun $x_i = D_i^2$, so folgt nach Einsetzen von $h(x) = \frac{1}{x}$

$$\frac{1}{\frac{1}{n}\sum_{i=1}^n D_i^2} \leq \frac{1}{n}\sum_{i=1}^n \frac{1}{D_i^2}$$

Wendet man den Erwartungswert an, so folgt

$$E(\frac{1}{\frac{1}{n}\sum_{i=1}^n D_i^2}) \leq E(\frac{1}{n}\sum_{i=1}^n \frac{1}{D_i^2}) = \frac{1}{n}\sum_{i=1}^n E(\frac{1}{D_i^2}) = E(\frac{1}{D_1^2})$$

und nach Umstellen von $\frac{1}{n}$ die gesuchte Ungleichung $E(\frac{1}{\sum_{i=1}^n D_i^2}) \leq E(\frac{1}{nD_1^2})$.

Lösung zu Aufgabe 25

Zwei Rechenwege:

1) Mit den Formeln aus der Vorlesung. Ergänze die Tabelle um zwei Spalten:

	M	F	E	$x = \ln(\frac{M}{E})$	$y = \ln(\frac{F}{E})$	x^2	xy
1	250	500	4	4.14	4.83	17.10	19.07
2	125	250	2	4.14	4.83	17.10	19.97
3	75	200	1	4.32	5.30	18.64	22.88
4	250	250	6	3.73	3.73	13.91	13.91
5	250	250	4	4.14	4.14	17.10	17.10
6	500	1000	6	4.42	5.12	19.56	22.63
7	110	125	3	3.60	3.73	12.97	13.43
$\sum$				28.48	31.67	116.39	129.88

– $a = \dfrac{S_{xy} - n\bar{x}\bar{y}}{S_{x^2} - n\bar{x}^2} = \dfrac{129.88 - 7 \cdot (28.48/7) \cdot (31.67/7)}{116.39 - 7 \cdot (28.48/7)^2} = \dfrac{1.028}{0.517} \approx 1.98$

– $b = \bar{y} - a\bar{x} = \dfrac{31.67}{7} - 1.98 \cdot \dfrac{28.48}{7} \approx -3.53$

Die Regressionsfunktion lautet also $y \approx 1.98x - 3.53$.

2) Mit Hilfe von Matrizen (als Spezialfall der multiplen linearen Regression) Hier mit den auf zwei Nachkommastellen gerundeten Werten x,y (so war es in der Klausur vorgegeben)

$$X = \begin{pmatrix} 1 & 4.14 \\ 1 & 4.14 \\ 1 & 4.32 \\ 1 & 3.73 \\ 1 & 4.14 \\ 1 & 4.42 \\ 1 & 3.60 \end{pmatrix}, \ Y = \begin{pmatrix} 4.83 \\ 4.83 \\ 5.30 \\ 3.73 \\ 4.14 \\ 5.12 \\ 3.73 \end{pmatrix}, \ (X^TX) = \begin{pmatrix} 7.00 & 28.49 \\ 28.49 & 116.49 \end{pmatrix}, \ X^TY = \begin{pmatrix} 31.68 \\ 130.00 \end{pmatrix}$$

(die letzten beiden Ausdrücke beinhalten Summen, Quadratsummen und Produktsummen, die der Taschenrechner ebenfalls liefert)

Mit $(X^TX)^{-1} = \begin{pmatrix} 31.0360 & -7.5905 \\ -7.5905 & 1.8650 \end{pmatrix}$

ergibt sich

$$\begin{pmatrix} b \\ a \end{pmatrix} = (X^TX)^{-1}(X^TY) = \begin{pmatrix} -3.5331 \\ 1.9800 \end{pmatrix}$$

Die Regressionsfunktion lautet also $y \approx 1.98x - 3.5331$

Lösung zu Aufgabe 26

Mit dem Taschenrechner ergibt sich die Regressionsfunktion $y = 1,1715x + 9,3144$. Alternativlösungen

1. Auch mit dem Taschenrechner und den darin verfügbaren Summen:
   $a = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{10033-10\times 24.4\times 37.9}{6624-10\times 24.4^2} = \frac{785.4}{670.4} \approx 1.17$, $b = \bar{y} - a\cdot\bar{x} = 37.9 - 1.17\times 24.4 \approx 9.31$

2. Über die Matrizenformeln (Einträge auch mit TR oder händisch über eine Tabelle mit Saldierung, was aber recht zeitaufwändig ist)
   $\mathbf{X}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 31 & 19 & \cdots & 15 \end{pmatrix}$, $\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} 10 & 244 \\ 244 & 6624 \end{pmatrix}$, $\mathbf{X}^T Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \begin{pmatrix} 379 \\ 10033 \end{pmatrix}$
   $(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{6704}\begin{pmatrix} 6624 & -244 \\ -244 & 10 \end{pmatrix} = \begin{pmatrix} \frac{414}{419} & -\frac{61}{1676} \\ -\frac{61}{1676} & \frac{5}{3352} \end{pmatrix} = \begin{pmatrix} 0.98807 & -0.03640 \\ -0.03640 & 0.00149 \end{pmatrix}$,
   $\beta = \begin{pmatrix} b \\ a \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \frac{1}{6704}\begin{pmatrix} 6624 & -244 \\ -244 & 10 \end{pmatrix}\begin{pmatrix} 379 \\ 10033 \end{pmatrix} = \frac{1}{6704}\begin{pmatrix} 62444 \\ 7854 \end{pmatrix} = \begin{pmatrix} 9,3144 \\ 1,1715 \end{pmatrix}$