Bachelorprüfung WS 2022

Fach: Data Science: Ökonometrie

Prüferin: Prof. Regina T. Riphahn, Ph.D.

Vorbemerkungen:

Anzahl der Aufgaben: Die Klausur besteht aus 3 Aufgaben, die alle bearbeitet werden müssen. **Es wird nur der Lösungsbogen eingesammelt.** Angaben auf dem Aufgabenzettel werden nicht gewertet.

Bewertung: Es können maximal 60 Punkte erworben werden. Die maximale Punktzahl ist für jede Aufgabe in Klammern angegeben. Sie entspricht der für die Aufgabe empfohlenen Bearbeitungszeit in Minuten.

Erlaubte Hilfsmittel:
- Formelsammlung (ist der Klausur beigefügt)
- Tabellen der statistischen Verteilungen (sind der Klausur beigefügt)
- Taschenrechner
- Fremdwörterbuch

Wichtige Hinweise:
- Sollte es vorkommen, dass die statistischen Tabellen, die dieser Klausur beiliegen, den gesuchten Wert der Freiheitsgrade nicht ausweisen, machen Sie dies kenntlich und verwenden Sie den nächstgelegenen Wert.
- Sollte es vorkommen, dass bei einer Berechnung eine erforderliche Information fehlt, machen Sie dies kenntlich und treffen Sie für den fehlenden Wert eine plausible Annahme.

Aufgabe 1: **[19 Punkte]**

Sie möchten die Jahresgehälter der Basketballspieler in der NBA prognostizieren. Folgende Variablen stehen Ihnen aus der Saison 2021/22 für 160 Basketballspieler zur Verfügung:

$salary_i$	= Jahresgehalt von Spieler i in Mio. US-Dollar
PPG_i	= Durchschnittliche Anzahl erzielter Punkte pro Spiel von Spieler i
age_i	= das Alter von Spieler i in Jahren
MPG_i	= Durchschnittliche Anzahl gespielter Minuten pro Spiel von Spieler i
APG_i	= Durchschnittliche Anzahl Vorlagen pro Spiel von Spieler i
PPG_age_i	= Interaktion zwischen PPG und age von Spieler i

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.32379   10.36696   3.118  0.00217
PPG          0.70837    0.82471   0.859  0.39170
age         -1.36718    0.42020  -3.254  0.00140
PPG_age      0.02941    0.03236   0.909  0.36477
---

Residual standard error: 9.107 on 156 degrees of freedom
Multiple R-squared:  0.5533, Adjusted R-squared:  0.5447
```

Sie schätzen folgendes Regressionsmodell (1) und erhalten obenstehenden Output:

$$salary_i = \beta_0 + \beta_1 \cdot PPG_i + \beta_2 \cdot age_i + \beta_3 \cdot PPG\_age_i + u_i \tag{1}$$

Runden Sie alle Zahlenangaben auf die <u>dritte</u> Nachkommastelle.

a) Luka ist 24 Jahre alt. Er erzielt durchschnittlich 36,8 Punkte. Berechnen Sie sein erwartetes Jahresgehalt. (2 Punkte)

b) Leiten Sie den marginalen Effekt der Variable *PPG* auf das Jahresgehalt her. Berechnen Sie den marginalen Effekt von *PPG* für Luka aus der Aufgabe a) und interpretieren Sie das Ergebnis. (3 Punkte)

c) Leiten Sie den marginalen Effekt von Alter her und berechnen Sie auf dieser Basis, bei welchem Wert von *PPG* das Jahresgehalt maximal ist. (4 Punkte)

d) Eine Kommilitonin besteht darauf, das Modell durch die Variablen MPG_i und APG_i zu ergänzen. Sie nehmen diese zwei Variablen ins Model auf und schätzen erneut (Modell 2).

$$salary_i = \beta_0 + \beta_1 \cdot PPG_i + \beta_2 \cdot age_i + \beta_3 \cdot PPG\_age_i + \beta_4 \cdot MPG_i + \beta_5 \cdot APG_i + u_i \tag{2}$$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.73583   10.08398   2.552  0.01168
PPG          2.20419    0.89165   2.472  0.01452
age         -0.85799    0.42764  -2.006  0.04657
MPG         -0.47286    0.16301  -2.901  0.00427
APG          1.33969    0.42612   3.144  0.00200
PPG_age     -0.02170    0.03364  -0.645  0.51981
---
Residual standard error: 8.736 on 154 degrees of freedom
Multiple R-squared:  0.5943,       Adjusted R-squared:  0.5811
```

Sie möchten nun testen, ob die geschätzten Parameter $\hat{\beta}_4$ und $\hat{\beta}_5$ gemeinsam statistisch signifikant sind. Benennen Sie das Testverfahren und formulieren Sie Null- und Alternativhypothese, berechnen Sie die Teststatistik und bestimmen Sie den kritischen Wert. Kann die Nullhypothese auf dem 1%-Signifikanzniveau abgelehnt werden? (6 Punkte)

e) Interpretieren Sie $\hat{\beta}_5$ statistisch und inhaltlich. (2 Punkte)

f) Welchen Wert würde $\hat{\beta}_0$ im Modell (2) annehmen, wenn $salary_i$ nicht in Mio. US-Dollars sondern in 1000 US-Dollars gemessen worden wäre? (2 Punkte)

Aufgabe 2: **[18 Punkte]**

Sie interessieren sich für das Sparverhalten von Individuen. Sie verfügen über Querschnittsdaten von 830 Personen aus dem Jahr 2020. Ihr Datensatz enthält folgenden Informationen:

$savings_i$	= jährliche Ersparnis von Person i in Euro
$income_i$	= Jahreseinkommen von Person i in Euro
$educ_i$	= Ausbildung von Person i in Jahren
age_i	= Alter von Person i in Jahren

Sie schätzen folgendes Regressionsmodell:

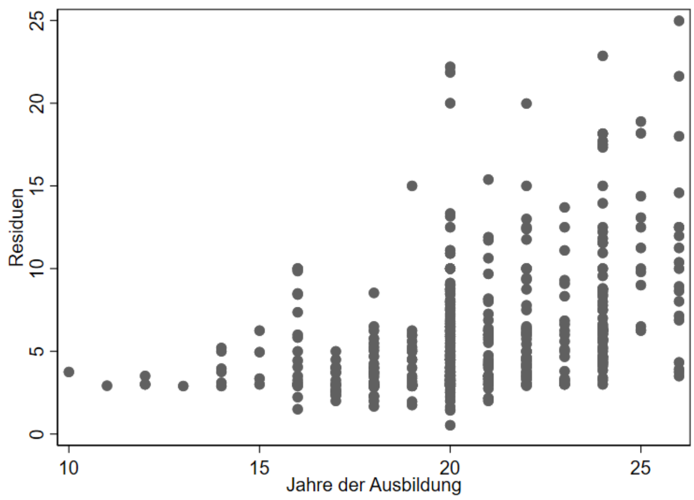
$$savings_i = \beta_0 + \beta_1 \cdot income_i + \beta_2 \cdot educ_i + \beta_3 \cdot age_i + u_i$$

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Interpretieren Sie $\hat{\beta}_1$ inhaltlich und statistisch. *Hinweis:* $\hat{\beta}_1 = 0,117$ und $p = 0,001$ (2 Punkte)

b) Wie lautet Ihre inhaltliche Interpretation von $\hat{\beta}_1$, wenn Sie statt *savings* die logarthmierte Ersparnis $ln(savings)$ als abhängige Variable nutzen. *Hinweis:* $\beta_1^{neu} = 0,005$. (1 Punkt)

c) Ihr Kommilitone rät Ihnen, die Variable *birth* mit in das Modell aufzunehmen. Diese Variable enthält Informationen über das Geburtsjahr von Person i. Sollten Sie die Variable *birth* in das Modell aufnehmen? Welche Konsequenzen hätte die Aufnahme für Ihre Schätzung? (4 Punkte)



d) Der Graph zeigt die Verteilung der Residuen für *educ*. Ist die Gauß-Markov Annahme MLR.5 hier verletzt? Erläutern Sie kurz Ihre Antwort. Welche Konsequenz hat dies für den KQ-Schätzer (3 Punkte)

e) Ihr Kommilitone meint, Sie hätten eine relevante Variable vergessen. Erklären Sie das Problem ausgelassener Variablen und welche Konsequenzen dies prinzipiell für Schätzungen hat. Erläutern Sie ein Beispiel für das Problem ausgelassener Variablen in der vorliegenden Schätzung. Begründen Sie kurz Ihre Antwort. (4 Punkte)

f) Berechnen Sie das R^2 der Schätzung und interpretieren Sie dieses. *Hinweis:* Das adjustierte Bestimmtheitsmaß beträgt 0,094. Nennen Sie einen Nachteil des R^2. (4 Punkte)

Aufgabe 3: **[23 Punkte]**

Sie interessieren sich für die Determinanten von Wohneigentum. Es steht Ihnen dafür ein Datensatz mit 5299 Personen zur Verfügung. Sie beobachten den folgenden Satz von Variablen:

$eigenheim_i$	=1, bei Eigentum von Haus oder Wohnung; =0, sonst
$alter_i$	= Alter von Person i in Jahren
$hheink_i$	= Haushaltsjahreseinkommen von Person i in Tausend Euro
$bildung_i$	= Bildungsjahre von Person i

Sie schätzen das folgende lineare Regressionsmodell und erhalten untenstehenden Output:

$$eigenheim_i = \beta_0 + \beta_1 \cdot alter_i + \beta_2 \cdot hheink_1 + \beta_3 \cdot bildung_i + u_i$$

```
              Estimate Std. Error  t value  Pr(>|t|)
(Intercept)   -1.842     0.457    -4.031   0.000
alter          0.044     0.004    11.000   0.000
hheink         0.014     0.001     ???      ???
bildung       -0.017     0.002    -8.500   0.000
---

Residual standard error: 0.461 on 5295 degrees of freedom
Multiple R-squared:  0.09813, Adjusted R-squared:  0.09762
F-statistic:   192 on 3 and 5295 DF,  p-value: < 2.2e-16
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Interpretieren Sie $\hat{\beta}_1$ inhaltlich und statistisch. (2 Punkte)

b) Berechnen und interpretieren Sie das 99%-Konfidenzintervall für den geschätzten Koeffizienten der Variable *hheink*. Gehen Sie darauf ein, ob der Koeffizient statistisch signifikant von Null verschieden ist. (5 Punkte)

c) Wie hoch ist die Wahrscheinlichkeit, dass eine 31 Jahre alte Person, mit einem Haushaltsjahreseinkommen von 55.000 Euro und 12 Jahren Schulbildung Wohneigentum besitzt? (3 Punkte)

d) Sie überlegen, dass β_2 unterschätzt sein könnte, da die Höhe des Einkommens im Zusammenhang mit der regionalen Bevölkerungdichte steht. Unter welchen Bedingungen würde sich die Vermutung bestätigen? (5 Punkte)

e) Nennen Sie zwei Nachteile des linearen Wahrscheinlichkeitsmodells. (2 Punkte)

f) Sie vermuten, dass sich die Parameter für ländliche und städtische Gebiete unterscheiden und führen einen Chow-Test auf Strukturbruch am 1%-Niveau durch. Geben Sie Hypothesen, Teststatistik, kritischen Wert und Ihre Testentscheidung an. Wird Ihre Vermutung bestätigt? (6 Punkte)
Hinweise: $SSR_{pooled} = 320$, $SSR_1 = 158$ (für *ländlich=0*), $SSR_2 = 160$ (für *ländlich=1*)

Formelsammlung – Praxis der empirischen Wirtschaftsforschung

Kapitel 1:

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_i(y_i-\bar{y})$$

$$=\sum_{i=1}^{n}(x_i-\bar{x})y_i$$

$$=\sum_{i=1}^{n}x_iy_i-n\bar{x}\cdot\bar{y}$$

$$E\left(\sum_{i=1}^{n}a_iX_i\right)=\sum_{i=1}^{n}a_iE(X_i)$$

$$E\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}E(X_i)$$

$Var(aX + bY) = a^2\,Var(X) + b^2\,Var(Y) + 2ab\,Cov(X,Y)$

Für identisch und unabhängig verteilte Zufallsvariablen Y_i:

$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_i\right)^2$$

Kapitel 2:

$y_i = \beta_0 + \beta_1 x_i + u_i$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\sum_{i=1}^{n}\hat{u}_i=0 \qquad \sum_{i=1}^{n}x_i\hat{u}_i=0$$

$$SST\equiv\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$SSE\equiv\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$SSR\equiv\sum_{i=1}^{n}\hat{u}_i^2=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

$$R^2=\frac{SSE}{SST}=1-\frac{SSR}{SST}\ ,\ 0\le R^2\le 1$$

$$E\left(\hat{\beta}_0\right)=\beta_0 \qquad E\left(\hat{\beta}_1\right)=\beta_1$$

$$Var(\hat{\beta}_1)=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sigma^2}{SST_x}$$

$$Var(\hat{\beta}_0)=\frac{\sigma^2\,\frac{1}{n}\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\hat{\sigma}^2=\frac{1}{(n-2)}\sum_{i=1}^{n}\hat{u}_i^2=\frac{SSR}{(n-2)}$$

Regression durch den Ursprung:

$$\tilde{\beta}_1=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}$$

Kapitel 3:

$$R^2=\frac{SSE}{SST}=\frac{\left(\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2\right)}$$

Wenn $y=\beta_0+\beta_1x_1+\beta_2x_2+u$

und $\tilde{y}=\tilde{\beta}_0+\tilde{\beta}_1x_1$

dann $\tilde{\beta}_1=\hat{\beta}_1+\hat{\beta}_2\tilde{\delta}_1$ mit $x_2=\tilde{\delta}_0+\tilde{\delta}_1x_1$

Allgemein für j = 1, 2, …, k:

$$Var\left(\hat{\beta}_j\right)=\frac{\sigma^2}{SST_j\left(1-R_j^2\right)}$$

$$SST_j=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-k-1}=\frac{SSR}{n-k-1}$$

$$se(\hat{\beta}_j)=\frac{\hat{\sigma}}{\left[SST_j(1-R_j^2)\right]^{\frac{1}{2}}}$$

MLR.1: Modell der Grundgesamtheit
$y = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \ldots + \beta_k x_k + u$

MLR.2: Zufallsstichprobe der Größe n folgt dem Bevölkerungsmodell.

MLR.3: Keine unabhängige Variable ist konstant. Keine perfekte Kollinearität.

MLR.4: $E\left(u\,|\,x_1,x_2,\ldots,x_k\right)=0$

MLR.5: $Var\left(u\,|\,x_1,x_2,\ldots,x_k\right)=\sigma^2$

MLR.6: u ist von x_1, x_2, …, x_k unabhängig und u ~ Normal (0, σ^2).

Kapitel 4:

$$(\hat{\beta}_j - \beta_j)\big/se(\hat{\beta}_j) \sim t_{n-k-1}$$

$$-t_{\frac{\alpha}{2},n-k-1} \leq \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \leq t_{\frac{\alpha}{2},n-k-1}$$

$$\hat{\beta}_j - c \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c \cdot se(\hat{\beta}_j)$$

$$F \equiv \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)}$$

$$F = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)}$$

Kapitel 5:

$$\lim_{n\to\infty} P(\left|\hat{\beta}_1 - \beta_1\right| > \varepsilon) \to 0$$

$$\text{plim}\left(\hat{\beta}_1\right) = \beta_1$$

Kapitel 6:

Standardisierung:

$$\frac{y_i - \overline{y}}{\hat{\sigma}_y} = \hat{\beta}_1 \left(\frac{\hat{\sigma}_1}{\hat{\sigma}_y}\right)\left(\frac{x_{i1} - \overline{x}_1}{\hat{\sigma}_1}\right) + \hat{\beta}_2 \left(\frac{\hat{\sigma}_2}{\hat{\sigma}_y}\right)\left(\frac{x_{i2} - \overline{x}_2}{\hat{\sigma}_2}\right)$$
$$+ \ldots + \hat{\beta}_k \left(\frac{\hat{\sigma}_k}{\hat{\sigma}_y}\right)\left(\frac{x_{ik} - \overline{x}_k}{\hat{\sigma}_k}\right) + \frac{\hat{u}_i}{\hat{\sigma}_y}$$

Semielastizität:
$$\%\Delta\hat{y} = 100 \cdot [\exp(\hat{\beta}_j \Delta x_j) - 1]$$

$$\overline{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - \frac{\hat{\sigma}^2}{SST/(n-1)}$$

$$\overline{R}^2 = 1 - (1-R^2)\frac{n-1}{n-k-1}$$

$$P[\hat{y}^0 - t_{\frac{\alpha}{2},n-k-1} \cdot se(\hat{e}^0) \leq y^0 \leq \hat{y}^0 + t_{\frac{\alpha}{2},n-k-1} \cdot se(\hat{e}^0)] = 1\text{-}\alpha$$

$$\widehat{\log y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \ldots + \hat{\beta}_k x_k$$
$$E(y|\mathbf{x}) = \exp(\sigma^2/2) \cdot \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

Kapitel 7:

Regression nach Gruppen
- Modell gepoolt: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$

Chow-Test (mit $SSR_P = SSR_{\text{gepooltes Modell}}$):

$$F = \frac{(SSR_P - (SSR_1 + SSR_2))/(k+1)}{(SSR_1 + SSR_2)/\left(n - 2(k+1)\right)}$$

TABLE G.2

Critical Values of the t Distribution

		Significance Level				
	1-Tailed:	.10	.05	.025	.01	.005
	2-Tailed:	.20	.10	.05	.02	.01
Degrees of Freedom	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

Examples: The 1% critical value for a one-tailed test with 25 *df* is 2.485. The 5% critical value for a two-tailed test with large (> 120) *df* is 1.96.

Source: This table was generated using the Stata® function invttail.

TABLE G.3c

1% Critical Values of the F Distribution

	Numerator Degrees of Freedom									
Denominator Degrees of Freedom	1	2	3	4	5	6	7	8	9	10
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Example: The 1% critical value for numerator $df = 3$ and denominator $df = 60$ is 4.13.
Source: This table was generated using the Stata® function invFtail.