

POL 213, Spring 2024

Midterm Exam

Instructors: Lauren Peritz and Yu-Shiuan (Lily) Huang

May 16, 2024

Instructions:

- The midterm exam is in class, closed book / note.
 - Write your answers directly onto the exam.
 - You have class time (12:10-3:00PM) to complete the exam.
 - The test is worth 15% of your course grade.
 - Attempt all questions.
-

1 (10pt)

Consider the bivariate linear regression of Y on X such that the best fit line has an equation $Y = A + BX$. For any point i in n observations, the regression equation is: $Y_i = A + BX_i + E_i$ where E_i is the residual. Using the least squares criterion, analytically derive the A and B to minimize the sum of squared residuals, $\sum E_i^2$. That is, find the values of A and B that minimize:

$$S(A, B) = \sum_{i=1}^n E_i^2 = \sum (Y_i - A - BX_i)^2$$

Explain every step of the derivation. Your final answers should be a formula the least squares coefficients A and B where \bar{Y} denotes the average of Y_i observations and \bar{X} denotes the average of X_i observations.

Because some students may be unfamiliar with calculus, I am giving you the following step. Make sure to conceptually explain how we reach this step in the derivation, even if you do not know how to execute the calculation yourself. Then use this step to derive the *normal equations*.

$$\begin{aligned}\frac{\partial S(A, B)}{\partial A} &= \sum (-1)(2)(Y_i - A - BX_i) \\ \frac{\partial S(A, B)}{\partial B} &= \sum (-X_i)(2)(Y_i - A - BX_i)\end{aligned}$$

[Q1 continued.]

2 (15pt)

The Gauss Markov Theorem establishes that under four key assumptions, of all linear unbiased estimators, the least-squares estimators are the most efficient (best linear unbiased estimator). The key assumptions concern the behavior of the errors ε_i :

- Linearity
 - Constant variance
 - Independence
 - X are fixed or, if random, are independent of errors.
- a. For each item, write the requirement specifically in mathematical notation.¹ Then explain conceptually what the assumption means and why a violation would render the estimator no longer BLUE.
- b. We often add a fifth assumption beyond the Gauss Markov theorem. Write it explicitly in mathematical notation. Then explain conceptually what the assumption means and what happens if there is a violation.

¹We will accept equations written either using scalar (one X) or matrix (multivariate \mathbf{X}) notation.

[Q2 continued.]

3 (10pt)

Suppose that the “true” model generating a set of data is $Y = \alpha + \beta_1 X_1 + \varepsilon$ where the error ε conforms to the usual linear-regression assumptions. A researcher fits the model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ which includes the irrelevant explanatory variable X_2 – that is, the true value of β_2 is 0. Had the researcher fit the (correct) simple-regression model, the variance of B_1 would have been $V(B_1) = \sigma_\varepsilon^2 / \sum (X_{i1} - \bar{X}_1)^2$

- a. Is the model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ wrong? Is B_1 for this model a biased estimator of β_1 ?
- b. The variance of B_1 in the multiple-regression model is:

$$V(B_1) = \frac{1}{1 - r_{12}^2} \times \frac{\sigma_\varepsilon^2}{\sum (X_{i1} - \bar{X}_1)^2}$$

What, then, is the cost of including the irrelevant explanatory variable X_2 ? How does this cost compare to that of *failing* to include a relevant explanatory variable?

[Q3 continued.]

4 (15pt)

One of the assumptions of the regression model is that the variation of the response variable around the regression surface is everywhere the same (homoskedasticity):

$$V(\varepsilon) = V(Y|x_1, \dots, x_k) = \sigma_\varepsilon^2$$

Conversely, nonconstant error variance is termed heteroskedasticity.

- a. Under heteroskedasticity, is the least squares estimator β unbiased and consistent? Is the least squares estimator efficient? Why or why not?
- b. Based on your answer in (a.) draw a conclusion about heteroskedasticity and whether the linear regression estimator is BLUE.

Under standard OLS assumptions, we can write the matrix representation for homoskedasticity as follows:

$$\mathbf{e} \stackrel{d}{\sim} N_n(0, \Sigma)$$

$$\Sigma = Var(\mathbf{e}) = E[\mathbf{e}\mathbf{e}'] = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- c. We talked about two types of error structures that violate this assumption. Two potential “fixes” are using robust and clustered standard errors. For each error structure, write a matrix Σ representing the problem and explain the problem in words. Then give an example of a real data analysis situation where you might encounter that error structure.

[Q4 continued.]

5 (10pt)

True or false and explain:

- a. Collinearity leads to bias in the OLS estimates.
- b. Collinearity leads to bias in the estimated standard errors for the OLS estimates.
- c. Collinearity leads to big standard errors for some estimates.

6 (20pt)

The `beauty.csv` file contains data from Hamermesh and Parker (2005) on student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations.

Table 1 presents the results of a multivariate linear regression analysis that examines the relationship between course evaluations (`eval`) and the main explanatory variable, `beauty`, while controlling for several other variables (instructors' sex, age, minority status, whether he/she received an undergraduate education in an English-speaking country, and whether he/she is assigned to lower-division courses). The standard error of each coefficient is reported in parentheses. Table 2 presents the analysis of variance of the multivariate linear regression model.

Table 1: Ordinary least-squares estimates of the determinants of course evaluations.

	<i>Dependent variable:</i>
	Course Evaluations
Composite Standardized Beauty	0.14 (0.03)
Female	-0.20 (0.05)
Age	-0.002 (0.003)
Minority	-0.07 (0.08)
Non-native English	-0.27 (0.11)
Lower division	0.10 (0.05)
Constant	4.19 (0.15)
Observations	463
R ²	0.10
Adjusted R ²	0.08
Residual Std. Error	0.53 (df = 456)

Note: Standard errors in parentheses.

- Perform a two-tailed hypothesis test to assess whether there is a statistically significant relationship between course evaluations ('eval') and the main explanatory variable, 'beauty', using a significance level of $\alpha = 0.05$. In your response, please clearly define the null and alternative hypotheses associated with this test and outline the steps you take to perform the test and interpret the results. (*t-distribution* table is provided on the last page).
- Conduct an F-test to evaluate the joint significance of the predictors in the multivariate linear regression model, using a significance level of $\alpha = 0.05$. In your response, please clearly define

Table 2: Anova

	df	sum of square	mean of square
beauty	1	5.0830	5.0830
female	1	4.3467	4.3467
age	1	0.2551	0.2551
minority	1	0.6398	0.6398
nonenglish	1	2.3609	2.3609
lower	1	0.9333	0.9333
residuals	456	128.6198	0.2821

the null and alternative hypotheses associated with this test and outline the steps you take to perform the test and interpret the results. (*F-distribution* table is provided on the last page)

[Q6 continued.]

[Q6 continued.]

7 (20pt)

The CIA World Factbook data set contains information about 261 countries, supranational entities (e.g., the European Union), territories (e.g., American Samoa), and some other places (e.g., the Gaza Strip). The subset of the data used here is for 134 nations with complete data on the following four variables:

- Infant mortality rate (`infant`) per 1,000 live births
- Gross domestic product (GDP) per capita (`gdp`), in thousands of U.S. dollars.
- Gini coefficient (`gini`) for the distribution of family income. The Gini coefficient is a standard measure of income inequality that ranges from 0 to 100, with 0 representing perfect equality and 100 maximum inequality.
- Health expenditures (`health`) as a percentage of GDP.

To explore the factors that may explain the variance in infant mortality rates (`infant`) across different countries, we regressed `infant` on all other variables included in the CIA World Factbook data. Before running the regression, we applied a logarithmic transformation to `infant` because its distribution is positively skewed. Table 3 presents the results of the linear regression.

```
m.cia <- lm(log(infant) ~ gdp + health + gini, data = CIA)
```

Table 3: Factors Explaining Infant Mortality Rates

	<i>Dependent variable:</i>
	Infant Mortality Rates
GDP	−0.04*** (0.004)
Gini	−0.05* (0.02)
Health expenditures	0.02*** (0.01)
Constant	3.02*** (0.29)
Observations	134
R ²	0.71
Adjusted R ²	0.71
Residual Std. Error	0.59 (df = 130)
F Statistic	107.19*** (df = 3; 130)

Note: *p<0.05; **p<0.01; ***p<0.001

The figure below displays the added variable plots from the multivariate linear regression we performed, as shown in Table 3. Table 4 presents the studentized residuals, hat values, and Cook's Distance statistics for some unusual cases.

Added-Variable Plots

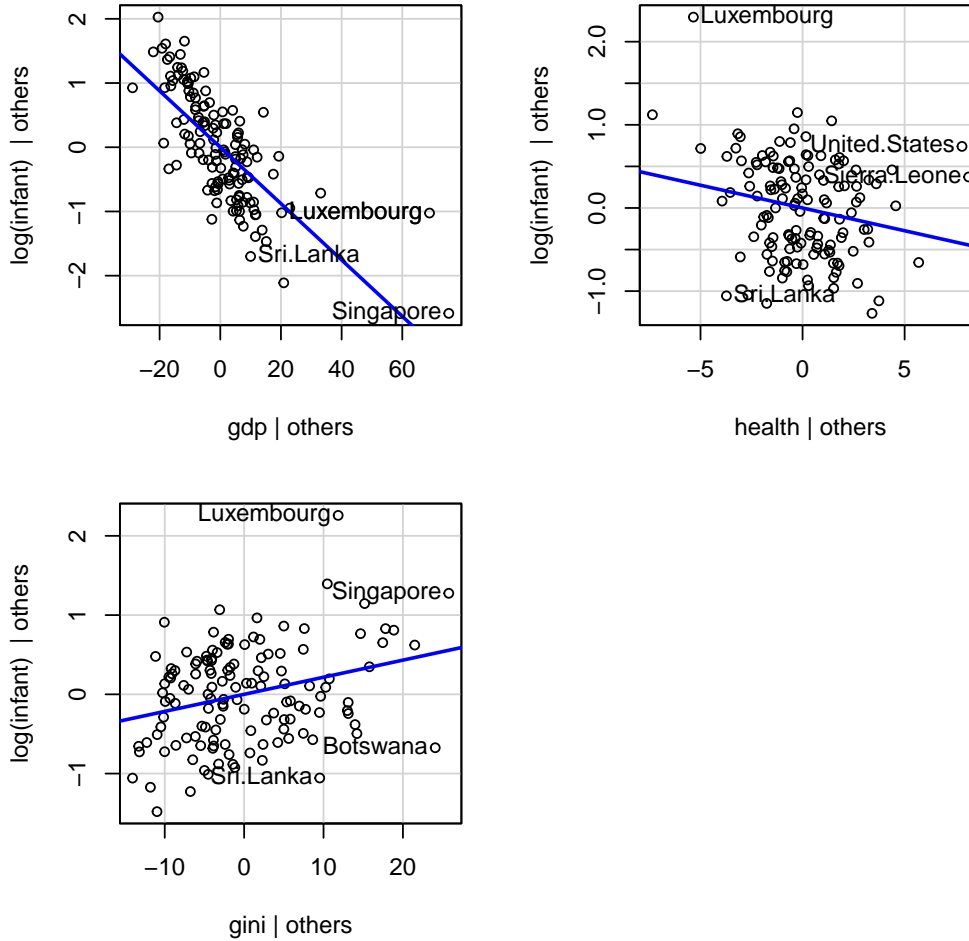


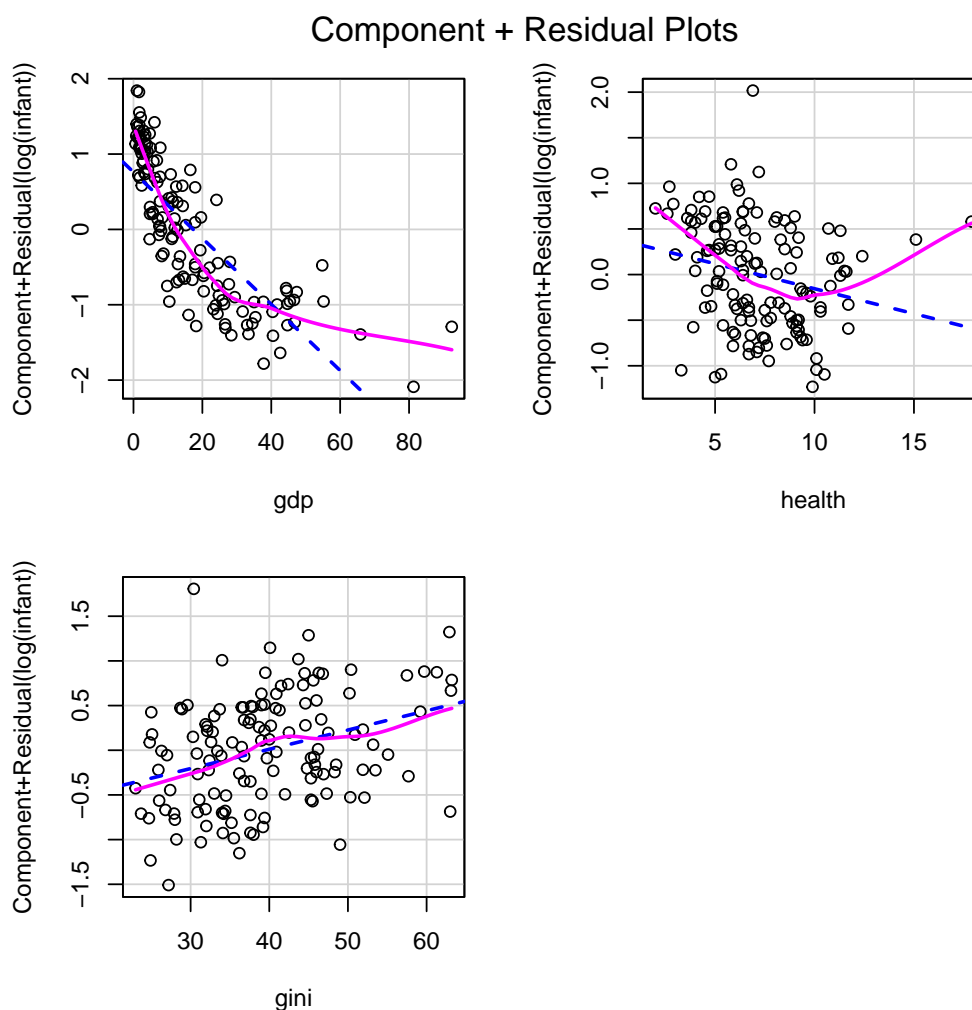
Table 4: Studentized Residuals, Hat Values, & Cook's Distance

	Studentized Residuals	Hat	Cook's D
Luxembourg	4.02	0.20	0.92
Singapore	1.41	0.24	0.16
Sri.Lanka	-2.21	0.03	0.04
United.States	2.18	0.15	0.21

- For multivariate linear regression, briefly explain why added variable plots are sometimes better for detecting influential points in the dataset compared to typical plots of residuals (either studentized or not) against predictor variables.
- In the added variable plots described above, identify some problematic cases within the

data. How do these unusual cases impact the regression coefficients? Select the two most influential cases you believe affect the regression coefficient estimates and discuss this using the information provided in Table 4. What would happen if these unusual cases were removed from the analysis?

The figure below displays the component plus residuals plots from the multivariate linear regression we performed, as shown in Table 3.



- c. Briefly discuss what each plot in the above component-plus-residual plots suggests regarding whether our model in Table 3 violates any Gauss-Markov assumptions.
- d. Following the previous question, if the analysis suggests that our model in Table 3 violates any Gauss-Markov assumptions, how can we address these violations?

[Q7 continued.]

[Q7 continued.]

Statistical Tables

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

F-table of Critical Values of $\alpha = 0.05$ for F(df1, df2)																			
	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00