Student Name: ID:

McGill University

Faculty of Science

Department of Mathematics and Statistics

Statistics Part A Comprehensive Exam

Methodology Paper

Date: Friday, May 16, 2014 Time: 13:00 17:00

Instructions

Answer only two questions out of Section L. If you answer more than two questions, then only the FIRST TWO questions will be marked.

Answer only two questions out of Section G. If you answer more than two questions, then only the FIRST TWO questions will be marked.

Questions	Marks
L1	
L2	
L3	
G1	
G2	
G3	

This exam comprises the cover page and 16 pages of questions.

Section L (Linear Regression Models)

Answer only two questions out of L1–L3

L1.

(a) (The use of indicator variables in linear regression analysis):

Consider a simple linear regression model that is used to study the relationship between a response variable y and a covariate x. The analysis is based on n independent observations

$$(x_i, y_i), \quad i = 1, 2, \ldots, n.$$

Suppose these n observations are divided into M groups each having n_m observations, such that $\sum_{m=1}^{M} n_m = n$. Assume that the groups are independent from each other. The most general simple linear regression model that can be used to analyze such data is

$$y = \beta_{0m} + \beta_{1m} x + \varepsilon, \quad m = 1, 2, \ldots, M$$

where $\varepsilon \sim N(0, \sigma^2)$. That is, one can fit M separate (different) simple linear regression models to the data, corresponding to the M independent groups of observations.

Using indicator variables,

Grade Point Averages		
Lower class	Middle class	Upper class
2.87	3.23	2.25
2.16	3.45	3.13
3.14	2.78	2.44
2.51	3.77	2.54

Denote

$_0$: grand mean of the grade point average for a college freshman.

$_i$: the effect of the i-th socioeconomic class on the grade point average, for $i \quad 1; 2; 3$.

$_i \quad _0 \quad _i$: mean of the grade point average for a college freshman in the i-th socioeconomic class, for $i \quad 1; 2; 3$.

In the above notation $i \quad 1; 2; 3$, are corresponding to the three socioeconomic classes mentioned in the table, respectively.

(i) Write down a one-way analysis of variance

L2.

An inverter is an electrical device that converts direct current (DC) to alternating current (AC). The data in this question are on measurement of the transient points of an electronic inverter. A portion of the data is given in the following Table. There are 24 observations.

Table 1:

	x_1	x_2	x_3	x_4	Y
1	3.00	3.00	3.00	3.00	0.79
2	3.00	6.00	6.00	6.00	1.71
⋮	⋮	⋮	⋮	⋮	⋮
23	2.00	3.00	8.00	6.00	1.51
24	3.00	3.00	8.00	8.00	0.75

The variables of interest are:

Y: Transient point (volts) of PMOS-NMOS inverters.
x_1: Width of the NMOS device.
x_2: Length of the NMOS device.
x_3: Width of the PMOS device.
x_4: Length of the PMOS device

Question L2 is continued on the next page.

L3.

Consider the multiple linear regression model

$$Y \quad X \quad "$$

where p $_0$; $_1$; $_2$; : : : ; $_k$q is the unknown vector of regression parameters, Y is the n 1 dimensional vector of observations of the response variable y, X is the n p k 1q dimensional design matrix; " is the n 1 dimensional vector of errors "$_i$, and " N p0; ^{2}V q, where V

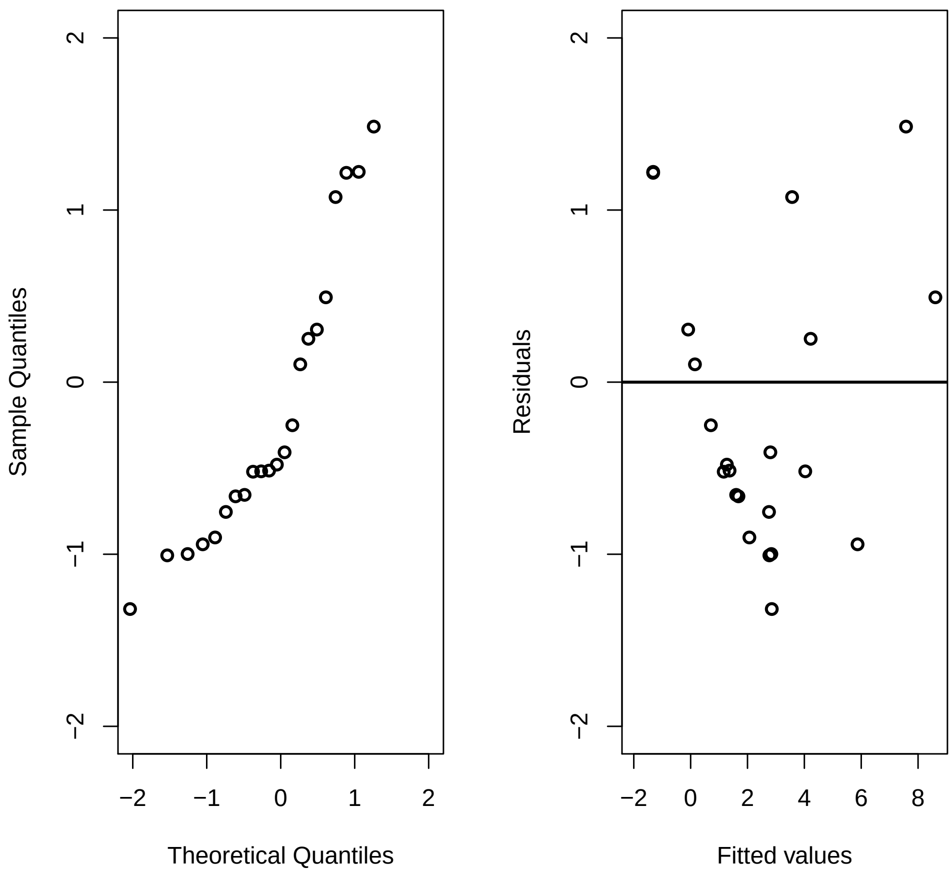
2

the

R-Code for Question L2.

```
> D1<-read.table("Q2_data.txt",header=T)
> attach(D1)
> fit1<-lm(y .,data=D1)
> summary(fit1)

Call:
lm(formula = y   ., data = D1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4894 -0.9324 -0.6098  0.7224  3.3659

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.52430    1.02317   1.490 0.152692
x1          -0.30606    0.07736  -3.956 0.000847 ***
x2           0.37439    0.05820   6.433 3.63e-06 ***
x3           0.44957    0.12354   3.639 0.001746 **
x4          -0.46557    0.13750  -3.386 0.003102 **
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 1.445 on 19 degrees of freedom
Multiple R-squared: 0.8044,     Adjusted R-squared: 0.7632
F-statistic: 19.53 on 4 and 19 DF,  p-value: 1.604e-06

> anova(fit1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1  11.989  11.989  5.7385  0.027056 *
x2         1 111.745 111.745 53.4885 6.185e-07 ***
x3         1  15.523  15.523  7.4304  0.013418 *
x4         1  23.951  23.951 11.4645  0.003102 **
Residuals 19  39.694   2.089
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
```

Figure 1: Residuals plots for Model 1 .



```
> fit2<-lm(new_y .,data=D1)
> summary(fit2)

Call:
lm(formula = new_y   ., data = D1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36567 -0.13258 -0.04697  0.10125  0.40247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07485    0.21386   -0.35     0.73
new_x1     -1.24866    0.07939  -15.73 2.38e-12 ***
new_x2      1.58623    0.07811   20.31 2.41e-14 ***
new_x3      0.93365    0.10515    8.88 3.44e-08 ***
new_x4     -1.34004    0.11749  -11.41 6.08e-10 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

Residual standard error: 0.2086 on 19 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9724
F-statistic: 203.7 on 4 and 19 DF,  p-value: 2.57e-15

> anova(fit2)
Analysis of Variance Table

Response: new_y
          Df  Sum Sq Mean Sq F value    Pr(>F)
new_x1     1  3.2553  3.2553  74.832 5.149e-08 ***
new_x2     1 26.0375 26.0375 598.539 7.940e-16 ***
new_x3   1  0.4960  0.4960  11.402  0.003167 **
new_x4     1  5.6591  5.6591 130.089 6.080e-10 ***
Residuals 19  0.8265  0.0435
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
```

Note: The response and covariates inModel 2 are called new in the output.

Section G (Generalized Linear Models)

Answer only two questions out of G1 G3

G1. The table below summarizes some data collected on traffic accidents and seat belt usage collected in the state of Florida in 1988:

Safety Equipment In Use (S)	Whether Ejected (E)	Injury (I) Nonfatal	Fatal
Seat belt	Yes	1105	14
	No	411,111	483
None	Yes	4624	497
	No	157,342	1008

Table 2: Source: Florida Department of Highway Safety and Motor Vehicles.

(a) Describe briefly why one cannot test for goodness of fit of the three-way interaction model, S:E:I , in 3 sentences or fewer. (2 marks)

(b) Consider the homogeneous association or no three-way interaction Poisson log-linear model, S:E E:I S:I. Describe, in words, what this model assumes about the conditional independence structure of the data. (2 marks)

(c) Identify the model (`gmodel1a`, `gmodel1b`, or `gmodel1c`) in the R output on pages 14 and 15 which fits the model of homogeneous association to the Florida seat belt data. Assess the goodness of fit of the model you've identified using a significance level 0.01. (8 marks)

(d) Identify the model (`gmodel1a`, `gmodel1b`, or `gmodel1c`) in the R output on pages 14 and 15 which fits the following model: S:E E:I . Assess the goodness of fit of the model you've identified using a significance level 0.01. (8 marks)

Question G1 is continued on the next page.

(e) Consider now the two models that you selected in parts (c) and (d) on the previous page. Which is the more appropriate model? Explain your reasoning. (6 marks)

(f) Another public safety researcher modeled the data in Table 2 using the following logistic regression model:

$$\log \left(\frac{Pr(I = \text{Nonfatal} \mid S, E)}{Pr(I = \text{Fatal} \mid S, E)} \right) = \beta_0 + \beta_S I(S = \text{Yes})$$

Determine which of the three Poisson loglinear models on pages 14 and 15 yields equivalent inference to this logistic regression model and prove that the models are equivalent. (8 marks)

(g) Compute the maximum likelihood parameter estimates for $\hat{\beta}_0$ and $\hat{\beta}_S$ for the logistic regression model in part (f) using the equivalent Poisson loglinear model that you identified in part (f). (6 marks)

G2.

(a) Suppose in an independent, two-sample problem that Y_i is Poisson with $g(\mu_i) = \alpha + \beta x_i$, where $x_i = 1$ for $i = 1, \ldots, n_A$ for the observations from group A and $x_i = 0$ for $i = n_A + 1, \ldots, n$ for the observations from group B. Show that for any link function g, the score equations imply that fitted means $\hat{\mu}_A$ and $\hat{\mu}_B$ equal the sample means. (10 marks)

(b) In a GLM, suppose that $Var(Y) = v(\mu)$ for $\mu = E(Y)$. Show that the link g satisfying $g'(\mu) = [v(\mu)]^{-1/2}$ has the same weight matrix $W^{(t)}$ at each cycle of the Fisher scoring algorithm. Show that this link function for a Poisson GLM is $g(\mu) = 2\sqrt{\mu}$.

(10 marks)

(c) Assume that Y is an ordinal response variable and x is a continuous predictor. Show that for the cumulative logit model,

$$\text{logit}[Pr(Y \le j)] = \alpha_j + \beta_j x$$

cumulative probabilities may be misordered for some x values. (10 marks)

(d) Assume that X and Y are both ordinal categorical variables and Z

G3. An often used data set in statistics courses contains observations of female horseshoe crabs. For this problem, Y_i is the binary response variable of interest is whether the i th horseshoe crab has at least one other male living outside of her nest (a satellite) or not. The only covariate of interest (X_i) for this problem is the weight (in kg) of the female crab. The goal of the analysis is to characterize the relationship between weight of the female horseshoe crabs and the presence of at least one male satellite crab.

(a) A marine biologist collaborator tells you that their past experience indicates that the probability of the presence of at least one satellite crab varies linearly with the weight of the female crab. Your collaborator suggests that one should use simple linear regression with Y as the response and X as the sole covariate. Give at least two reasons, without referring to the output, why a simple linear regression model would not necessarily be appropriate here. 4 marks

(b) You suggest that a GLM with a logistic link would be a better choice. Your collaborator is concerned that this model would not respect the assumption that the probability of a satellite was approximately linear in the weight. Again without referring to the R output, explain why using a generalized linear model with a logistic link would still be reasonable if the association were truly linear over the support of X. Hint: CXC lo1 1G8c Sectiab v50ion with

k q r f o f . s 4 marks(

```
> summary(gmodel1b)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.026472   0.025986 231.916   <2e-16 ***

```

Table of the Chi-squared distribution

Entries in table are ^{2} p q; the tail quantile of Chi-squared p q distribution
given in columns, given in rows.

	Left-tail					Right-tail				
	0.99500	0.99000	0.97500	0.95000	0.90000	0.10000	0.05000	0.02500	0.01000	0.00500
1	0.00004	0.00016	0.00098	0.00393	0.01579	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	3.48954	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	17.29188	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	40.25602	43.77297	46.97924	50.89218	53.67196
40	20.70654	22.16426	24.43304	26.50930	29.05052	51.80506	55.75848	59.34171	63.69074	66.76596
50	27.99075	29.70668	32.35736	34.76425	37.68865	63.16712	67.50481	71.42020	76.15389	79.48998
60	35.53449	37.48485	40.48175	43.18796	46.45889	74.39701	79.08194	83.29767	88.37942	91.95170
70	43.27518	45.44172	48.75756	51.73928	55.32894	85.52704	90.53123	95.02318	100.42518	104.21490
80	51.17193	53.54008	57.15317	60.39148	64.27784	96.57820	101.87947	106.62857	112.32879	116.32106
90	59.19630	61.75408	65.64662	69.12603	73.29109	107.56501	113.14527	118.13589	124.11632	128.29894
100	67.32756	70.06489	74.22193	77.92947	82.35814	118.49800	124.34211	129.56120	135.80672	140.16949

Table of the F-distribution distribution

Entries in table are F1 p 1; 2q the tail quantile of the F 1; 2 distribution

2z l	1	2	3	4	5	6	7	8	10	12	15	20	30	50	8
1	0.900	49.5	53.6	55.8	57.2	58.2	59.1	59.7	60.5	61.0	61.5	62.0	62.6	63.0	63.3
	0.950	199.	216.	225.	230.	234.	237.	239.	242.	244.	246.	248.	250.	252.	254.
	0.975	800.	864.	900.	922.	937.	948.	957.	969.	977.	985.	993.			
	0.990														
	0.999														
2	0.900	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.39	9.41	9.43	9.44	9.46	9.47	9.49
	0.950	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
	0.975	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
	0.990	99.0	99.2	99.2	99.3	99.3	99.4	100.	100.	100.	100.	100.	100.	100.	99.5
	0.999	999.	999.												
3	0.900	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.23	5.22	5.20	5.18	5.17	5.15	5.13
	0.950	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.70	8.66	8.62	8.58	8.53
	0.975	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.4	14.3	14.3	14.2	14.1	14.0	13.9
	0.990	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.2	27.1	26.9	26.7	26.5	26.4	26.1
	0.999	149.	141.	137.	135.	133.	132.	131.	129.	128.	127.	126.	125.	125.	123.
4	0.900	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.92	3.90	3.87	3.84	3.82	3.79	3.76
	0.950	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.86	5.80	5.75	5.70	5.63
	0.975	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.84	8.75	8.66	8.56	8.46	8.38	8.26
	0.990	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.5	14.4	14.2	14.0	13.8	13.7	13.5
	0.999	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.0	47.4	46.8	46.1	45.4	44.9	44.1
5	0.900	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.30	3.27	3.24	3.21	3.17	3.15	3.10
	0.950	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.62	4.56	4.50	4.44	4.36
	0.975	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.62	6.52	6.43	6.33	6.23	6.14	6.02
	0.990	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.1	9.89	9.72	9.55	9.38	9.24	9.02
	0.999	37.1	33.2	31.1	29.8	28.8	28.2	27.6	26.9	26.4	25.9	25.4	24.9	24.4	23.8
6	0.900	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.94	2.90	2.87	2.84	2.80	2.77	2.72
	0.950	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.94	3.87	3.81	3.75	3.67
	0.975	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.46	5.37	5.27	5.17	5.07	4.98	4.85
	0.990	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.87	7.72	7.56	7.40	7.23	7.09	6.88
	0.999	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.4	18.0	17.6	17.1	16.7	16.3	15.7
7	0.900	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.70	2.67	2.63	2.59	2.56	2.52	2.47
	0.950	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.51	3.44	3.38	3.32	3.23
	0.975	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.76	4.67	4.57	4.47	4.36	4.28	4.14
	0.990	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.62	6.47	6.31	6.16	5.99	5.86	5.65
	0.999														

Table of the F-distribution distribution

Entries in table are F_1 p 1; 2q the tail quantile of the F 1; 2 distribution

2z l	1	2	3	4	5	6	7	8	10	12	15	20	30	50	8
9	0.900	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.42	2.38	2.34	2.30	2.25	2.22	2.16
	0.950	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	3.01	2.94	2.86	2.80	2.71
	0.975	5.71	5.08	4.72	4.48	4.32	4.20	4.10	3.96	3.87	3.77	3.67	3.56	3.47	3.33
	0.990	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.26	5.11	4.96	4.81	4.65	4.52	4.31
	0.999	16.4	13.9	12.6	11.7	11.1	10.7	10.4	9.89	9.57	9.24	8.90	8.55	8.26	7.81
10	0.900	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.32	2.28	2.24	2.20	2.16	2.12	2.06
	0.950	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.84	2.77	2.70	2.64	2.54
	0.975	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.72	3.62	3.52	3.42	3.31	3.22	3.08
	0.990	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.85	4.71	4.56	4.41	4.25	4.11	3.91
	0.999	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.75	8.45	8.13	7.80	7.47	7.19	6.76
11	0.900	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.12	2.08	2.04	1.97
	0.950	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.72	2.65	2.57	2.51	2.40
	0.975	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.53	3.43	3.33	3.23	3.12	3.03	2.88
	0.990	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.54	4.40	4.25	4.10	3.94	3.81	3.60
	0.999	13.8	11.6	10.3	9.58	9.05	8.66	8.35	7.92	7.63	7.32	7.01	6.68	6.42	6.00
12	0.900	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.19	2.15	2.10	2.06	2.01	1.97	1.90
	0.950	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.62	2.54	2.47	2.40	2.30
	0.975	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.37	3.28	3.18	3.07	2.96	2.87	2.72
	0.990	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.30	4.16	4.01	3.86	3.70	3.57	3.36
	0.999	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.29	7.00	6.71	6.40	6.09	5.83	5.42
13	0.900	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.14	2.10	2.05	2.01	1.96	1.92	1.85
	0.950														