Assignment on R-Programming , SPSS and EXCEL

Teacher: Kajal Dihidar, Sampling and Official Statistics Unit, Indian Statistical Institute, Kolkata, India

Date :-- 09.11.2015 Last Date of Submission :-- 11.12.2015 Total Marks : 100

Consider the data file namely Birth Weight data set given below in this assignment. This data set consists of the data collected on 189 women to identify the risk factors associated with the birth of a low birth weight baby. The data was collected at the Baystate Medical Center in Springfield, Massachusetts. This data is analyzed in the book: `Applied Biostatistics for the Health Sciences' by Rossi, Rchard J.

The variables included in this data set are summarized in Table 1.

Based on this data set, answer the following questions using appropriate **R- programs**. Do the same exercises through **SPSS** and **Excel**.

Q 1. Read the data in R program, print the number of records, the number of variables, and the names of variables. Find out which variables are continuous, which are simply categorical (nominal), which are ordinal (maintaining the increasing or decreasing order of categories). Declare all the non-continuous variables as factor variables. Print the summary statistics of all the meaningful variables.

Q 2. Present the univariate frequency distribution and then the sample percentage distribution of the patients having babies with normal birth weights and low birth weights and draw the bar chart with proper title. Write comments on your results.

Q 3. Present the bivariate sample percentage distribution of the patients having babies with normal birth weights and low birth weights by smoking habit and draw the suitable bar chart with proper title. Using this bar chart give the estimates for

(a) Percentage of low-weight babies for mothers who smoked during pregnancy.

(b) Percentage of low-weight babies for mothers who did not smoke during pregnancy.

(c) Test whether the occurrences of low weight babies is independent of the smoking habit of mothers.

Q 4. Draw the boxplot of the age of the mothers in this data set with proper title. Use this boxplot to answer the following questions.

(a) Estimate the median age of the mothers.

(b) Estimate the 75-th percentile of the distribution of the variable AGE.

(c) Describe the shape of the underlying distribution of the variable AGE.

(d) Estimate the interquartile range in the distribution of the varable AGE.

(e) Are there any outliers in this data set for variable AGE.

Q 5. Draw the histogram of birth weights (BWT) in this data set for all mothers and comment on the shape of the distribution of this data. Test whether this data can be considered to come from a Normal probability model. Answer graphically as well as based on some test.

Q 6. Test whether the BWT data have come from a population having mean value greater than 2500 gm.

Q 7. Draw the boxplot of the birth weights (BWT) in this data set separated by smoking status of mothers in a single boxplot window. Comment on your results.

Q 8. Draw the histogram of the birth weights (BWT) in this data set separated by smoking status of mothers in a single plot window. Does the distribution of the variable BWT appear to have the same general shape for mothers who smoked and who did not smoke during pregnancy.

Q 9. Test whether the normal probability model can be considered as a plausible probability model for the distributions of BWT for mothers who smoked and who did not smoke. Answer graphically as well as by using testing of hypothesis.

Q 10. Estimate the mean, variance, standard deviation for BWT for mothers who smoked and did not smoke during pregnancy. Also estimate the standard error (standard deviation / sqrt(n)) for estimating mean BWT for mothers who smoked and did not smoke during pregnancy. Then compute the 95% confidence interval for the mean birth weight to mother who smoked during pregnancy and comment on your results.

Q 11. Estimate the 10^{th}, 25^{th}, 50^{th}, 60^{th}, 75^{th}, 90^{th} BWT percentiles by smoking status of the mothers.

Q 12. Test whether the sample data have come from population for which the mean birth weight for non-smoker mothers is greater than for smoker mothers.

Q 13. (a) Estimate the mean BWT for mothers of various races.

(b) Assess the normality assumption for BWT data for different races.

© Test whether the variances for BWT data for mothers of different races are equal or not.

(d) Then determine which test function is suitable to test whether the sample data have come from population for which the mean birth weight for mothers of different races are same or not.

(e) Perform the test accordingly.

(f) If the test rejects the hypothesis of equal means, then determine which pairs have different means.

Q 14. Test whether the proportion of low weight births among mothers of various races are same or not.

Q 15. Test whether the proportion of low weight births among mothers of two smoking status are same or not.

Q 16. Test whether the occurrences of low weight babies is independent of the races of mothers.

Q 17. Draw the scatter plot matrix among all the continuous variables. Print the sample correlation coefficient matrix among them. Comment on your results.

Q 18. Perform the multiple linear regression model for relating baby's birth weight (BWT) to mothers age (AGE), mother's weight at the last menstrual period (LWT), and also to the categorical variables : smoking status of mothers during pregnancy (SMOKE), History of premature labor (PTL), History of hypertension (HT), Presence of uterine irritability (UI), Number of physician visits during the first trimester (FTV).

Extract the model residuals. Check the normality assumption of model residuals by testing as well as by normal q-q plot.

Extract also the fitted values. Check the constant variance assumption of the model residuals by plotting the model residuals Vs model fitted values.

Check whether all the explanatory variables have significant effect on the BWT data or not. Also write down the value of the coefficient of determination (R-square = Sum of squares due to Reg / Total sum of squares) and comment on it.

Q 19. From the above regression output, exclude the unimportant explanatory variables and perform the regression analysis again. Print the output of new regression analysis, and check now whether all explanatory variables have significant effect or not and also check that R^2-adjusted is improved (increased) and choose the model having largest R^2-adjusted value and fewest explanatory variables. Write down the final model.

For this model also check the normality and constant variance assumption of the residuals.

Q 20. Consider the binary response variable of having low weight birth (LOW= 0 for no and = 1 for yes). Fit a logistic regression model for this variable of having or not having low weight baby based on the explanatory variables mother's age, mother's weight at last menstrual period and the smoking status of mother during pregnancy.

Print the output of this analysis and remark on this.

Q 21. In above logistic regression output, check whether all the explanatory variables have significant effect on having the low weight baby.

If not, exclude from the model the unimportant variables and re-perform the logistic regression analysis.

Comment on the final model and write down the model.

Q 22. From model of previous question, extract the estimates and compute the Odds Ratios as the exponential of the coefficients. Comment on the Odds Ratios.

Table 1 : Description of the variables in the Birth Weight data set

Variable	Description	Code / Values	Name
1	Identification code	ID number	ID
2	Low birth weight	1 = BWT <= 2500 g 0 = BWT > 2500 g	LOW
3	Age of mother	Years	AGE
4	Weight of mother at last menstrual period	Pounds	LWT
5	Race	1 = White 2 = Black 3 = Other	RACE
6	Smoking status during pregnancy	0 = No 1 = Yes	SMOKE
7	History of premature labor	0 = None 1 = One 2 = Two, etc.	PTL
8	History of hypertension	0 = No 1 = Yes	HT
9	Presence of uterine irritability	0 = No 1 = Yes	UI
10	Number of physician visits during the first trimester	0 = None 1 = One 2 = Two, etc.	FTV
11	Birth weight	Grams	BWT

Birth Weight Data Set

ID	LOW	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FTV	BWT
85	0	19	182	2	0	0	0	1	0	2523
86	0	33	155	3	0	0	0	0	3	2551
87	0	20	105	1	1	0	0	0	1	2557
88	0	21	108	1	1	0	0	1	2	2594
89	0	18	107	1	1	0	0	1	0	2600
91	0	21	124	3	0	0	0	0	0	2622
92	0	22	118	1	0	0	0	0	1	2637
93	0	17	103	3	0	0	0	0	1	2637
94	0	29	123	1	1	0	0	0	1	2663
95	0	26	113	1	1	0	0	0	0	2665
96	0	19	95	3	0	0	0	0	0	2722
97	0	19	150	3	0	0	0	0	1	2733
98	0	22	95	3	0	0	1	0	0	2750
99	0	30	107	3	0	1	0	1	2	2750
100	0	18	100	1	1	0	0	0	0	2769
101	0	18	100	1	1	0	0	0	0	2769
102	0	15	98	2	0	0	0	0	0	2778
103	0	25	118	1	1	0	0	0	3	2782
104	0	20	120	3	0	0	0	1	0	2807
105	0	28	120	1	1	0	0	0	1	2821
106	0	32	121	3	0	0	0	0	2	2835
107	0	31	100	1	0	0	0	1	3	2835
108	0	36	202	1	0	0	0	0	1	2836
109	0	28	120	3	0	0	0	0	0	2863
111	0	25	120	3	0	0	0	1	2	2877
112	0	28	167	1	0	0	0	0	0	2877
113	0	17	122	1	1	0	0	0	0	2906
114	0	29	150	1	0	0	0	0	2	2920
115	0	26	168	2	1	0	0	0	0	2920
116	0	17	113	2	0	0	0	0	1	2920
117	0	17	113	2	0	0	0	0	1	2920
118	0	24	90	1	1	1	0	0	1	2948
119	0	35	121	2	1	1	0	0	1	2948
120	0	25	155	1	0	0	0	0	1	2977
121	0	25	125	2	0	0	0	0	0	2977
123	0	29	140	1	1	0	0	0	2	2977
124	0	19	138	1	1	0	0	0	2	2977
125	0	27	124	1	1	0	0	0	0	2992
126	0	31	215	1	1	0	0	0	2	3005
127	0	33	109	1	1	0	0	0	1	3033
128	0	21	185	2	1	0	0	0	2	3042
129	0	19	189	1	0	0	0	0	2	3062
130	0	23	130	2	0	0	0	0	1	3062
131	0	21	160	1	0	0	0	0	0	3062

132	0	18	90	1	1	0	0	1	0	3076
133	0	18	90	1	1	0	0	1	0	3076
134	0	32	132	1	0	0	0	0	4	3080
135	0	19	132	3	0	0	0	0	0	3090
136	0	24	115	1	0	0	0	0	2	3090
137	0	22	85	3	1	0	0	0	0	3090
138	0	22	120	1	0	0	1	0	1	3100
139	0	23	128	3	0	0	0	0	0	3104
140	0	22	130	1	1	0	0	0	0	3132
141	0	30	95	1	1	0	0	0	2	3147
142	0	19	115	3	0	0	0	0	0	3175
143	0	16	110	3	0	0	0	0	0	3175
144	0	21	110	3	1	0	0	1	0	3203
145	0	30	153	3	0	0	0	0	0	3203
146	0	20	103	3	0	0	0	0	0	3203
147	0	17	119	3	0	0	0	0	0	3225
148	0	17	119	3	0	0	0	0	0	3225
149	0	23	119	3	0	0	0	0	2	3232
150	0	24	110	3	0	0	0	0	0	3232
151	0	28	140	1	0	0	0	0	0	3234
154	0	26	133	3	1	2	0	0	0	3260
155	0	20	169	3	0	1	0	1	1	3274
156	0	24	115	3	0	0	0	0	2	3274
159	0	28	250	3	1	0	0	0	6	3303
160	0	20	141	1	0	2	0	1	1	3317
161	0	22	158	2	0	1	0	0	2	3317
162	0	22	112	1	1	2	0	0	0	3317
163	0	31	150	3	1	0	0	0	2	3321
164	0	23	115	3	1	0	0	0	1	3331
166	0	16	112	2	0	0	0	0	0	3374
167	0	16	135	1	1	0	0	0	0	3374
168	0	18	229	2	0	0	0	0	0	3402
169	0	25	140	1	0	0	0	0	1	3416
170	0	32	134	1	1	1	0	0	4	3430
172	0	20	121	2	1	0	0	0	0	3444
173	0	23	190	1	0	0	0	0	0	3459
174	0	22	131	1	0	0	0	0	1	3460
175	0	32	170	1	0	0	0	0	0	3473
176	0	30	110	3	0	0	0	0	0	3475
177	0	20	127	3	0	0	0	0	0	3487
179	0	23	123	3	0	0	0	0	0	3544
180	0	17	120	3	1	0	0	0	0	3572
181	0	19	105	3	0	0	0	0	0	3572
182	0	23	130	1	0	0	0	0	0	3586
183	0	36	175	1	0	0	0	0	0	3600
184	0	22	125	1	0	0	0	0	1	3614
185	0	24	133	1	0	0	0	0	0	3614

186	0	21	134	3	0	0	0	0	2	3629
187	0	19	235	1	1	0	1	0	0	3629
188	0	25	95	1	1	3	0	1	0	3637
189	0	16	135	1	1	0	0	0	0	3643
190	0	29	135	1	0	0	0	0	1	3651
191	0	29	154	1	0	0	0	0	1	3651
192	0	19	147	1	1	0	0	0	0	3651
193	0	19	147	1	1	0	0	0	0	3651
195	0	30	137	1	0	0	0	0	1	3699
196	0	24	110	1	0	0	0	0	1	3728
197	0	19	184	1	1	0	1	0	0	3756
199	0	24	110	3	0	1	0	0	0	3770
200	0	23	110	1	0	0	0	0	1	3770
201	0	20	120	3	0	0	0	0	0	3770
202	0	25	241	2	0	0	1	0	0	3790
203	0	30	112	1	0	0	0	0	1	3799
204	0	22	169	1	0	0	0	0	0	3827
205	0	18	120	1	1	0	0	0	2	3856
206	0	16	170	2	0	0	0	0	4	3860
207	0	32	186	1	0	0	0	0	2	3860
208	0	18	120	3	0	0	0	0	1	3884
209	0	29	130	1	1	0	0	0	2	3884
210	0	33	117	1	0	0	0	1	1	3912
211	0	20	170	1	1	0	0	0	0	3940
212	0	28	134	3	0	0	0	0	1	3941
213	0	14	135	1	0	0	0	0	0	3941
214	0	28	130	3	0	0	0	0	0	3969
215	0	25	120	1	0	0	0	0	2	3983
216	0	16	95	3	0	0	0	0	1	3997
217	0	20	158	1	0	0	0	0	1	3997
218	0	26	160	3	0	0	0	0	0	4054
219	0	21	115	1	0	0	0	0	1	4054
220	0	22	129	1	0	0	0	0	0	4111
221	0	25	130	1	0	0	0	0	2	4153
222	0	31	120	1	0	0	0	0	2	4167
223	0	35	170	1	0	1	0	0	1	4174
224	0	19	120	1	1	0	0	0	0	4238
225	0	24	116	1	0	0	0	0	1	4593
226	0	45	123	1	0	0	0	0	1	4990
4	1	28	120	3	1	1	0	1	0	709
10	1	29	130	1	0	0	0	1	2	1021
11	1	34	187	2	1	0	1	0	0	1135
13	1	25	105	3	0	1	1	0	0	1330
15	1	25	85	3	0	0	0	1	0	1474
16	1	27	150	3	0	0	0	0	0	1588
17	1	23	97	3	0	0	0	1	1	1588
18	1	24	128	2	0	1	0	0	1	1701

19	1	24	132	3	0	0	1	0	0	1729
20	1	21	165	1	1	0	1	0	1	1790
22	1	32	105	1	1	0	0	0	0	1818
23	1	19	91	1	1	2	0	1	0	1885
24	1	25	115	3	0	0	0	0	0	1893
25	1	16	130	3	0	0	0	0	1	1899
26	1	25	92	1	1	0	0	0	0	1928
27	1	20	150	1	1	0	0	0	2	1928
28	1	21	200	2	0	0	0	1	2	1928
29	1	24	155	1	1	1	0	0	0	1936
30	1	21	103	3	0	0	0	0	0	1970
31	1	20	125	3	0	0	0	1	0	2055
32	1	25	89	3	0	2	0	0	1	2055
33	1	19	102	1	0	0	0	0	2	2082
34	1	19	112	1	1	0	0	1	0	2084
35	1	26	117	1	1	1	0	0	0	2084
36	1	24	138	1	0	0	0	0	0	2100
37	1	17	130	3	1	1	0	1	0	2125
40	1	20	120	2	1	0	0	0	3	2126
42	1	22	130	1	1	1	0	1	1	2187
43	1	27	130	2	0	0	0	1	0	2187
44	1	20	80	3	1	0	0	1	0	2211
45	1	17	110	1	1	0	0	0	0	2225
46	1	25	105	3	0	1	0	0	1	2240
47	1	20	109	3	0	0	0	0	0	2240
49	1	18	148	3	0	0	0	0	0	2282
50	1	18	110	2	1	1	0	0	0	2296
51	1	20	121	1	1	1	0	1	0	2296
52	1	21	100	3	0	1	0	0	4	2301
54	1	26	96	3	0	0	0	0	0	2325
56	1	31	102	1	1	1	0	0	1	2353
57	1	15	110	1	0	0	0	0	0	2353
59	1	23	187	2	1	0	0	0	1	2367
60	1	20	122	2	1	0	0	0	0	2381
61	1	24	105	2	1	0	0	0	0	2381
62	1	15	115	3	0	0	0	1	0	2381
63	1	23	120	3	0	0	0	0	0	2395
65	1	30	142	1	1	1	0	0	0	2410
67	1	22	130	1	1	0	0	0	1	2410
68	1	17	120	1	1	0	0	0	3	2414
69	1	23	110	1	1	1	0	0	0	2424
71	1	17	120	2	0	0	0	0	2	2438
75	1	26	154	3	0	1	1	0	1	2442
76	1	20	105	3	0	0	0	0	3	2450
77	1	26	190	1	1	0	0	0	0	2466
78	1	14	101	3	1	1	0	0	0	2466
79	1	28	95	1	1	0	0	0	2	2466

81	1	14	100	3	0	0	0	0	2	2495
82	1	23	94	3	1	0	0	0	0	2495
83	1	17	142	2	0	0	1	0	0	2495
84	1	21	130	1	1	0	1	0	3	2495