CSC591: Foundations of Data Science HW5: Bayesian Inference, Missing Data Analysis
Released: 11/25/15 Due: **12/04/15 (23:55pm);** (One day late: -25%; -100% after that).

Student Name: Parth Satra
Student ID: pasatra (200062999)

Q1. Bayesian Inferencing (30 points)

(a) Following table give prior distribution for the proportion of defective parts produced by a machine

p (proportion of defects)	0.2	0.3
f(p): prior probability	0.7	0.3

Let x denote the number of defectives among a random sample of size 2.

(1) Find the posterior probability distribution of p, given that x is observed. (**10 points)**

Answer

The random variable x follows a binomial distribution.
Thus f(x|p) = b(x; 2; p). since x is random sample of size 2.

Now the marginal distribution = f(x|0.2)f(0.2) + f(x|0.3)f(0.3)
Thus marginal distribution = b(x; 2; 0.2) * 0.7 + b(x; 2; 0.3) * 0.3 for x = 0,1,2

F(0) = ${}^2C_0 * (0.2)^0 * (0.8)^{(2-0)} * 0.7 + {}^2C_0 * (0.3)^0 * (0.7)^{(2-0)} * 0.3 =$ **0.595**
F(1) = ${}^2C_1 * (0.2)^1 * (0.8)^{(2-1)} * 0.7 + {}^2C_1 * (0.3)^1 * (0.7)^{(2-1)} * 0.3 =$ **0.35**
F(2) = ${}^2C_2 * (0.2)^2 * (0.8)^{(2-2)} * 0.7 + {}^2C_2 * (0.3)^2 * (0.7)^{(2-2)} * 0.3 =$ **0.055**

Now calculating the posterior probability.
f(0.2|x) = f(x|0.2)f(0.2) / F(x)
f(0.3|x) = 1 – f(0.2|x)

Thus,
f(0.2|0) = $({}^2C_0 * (0.2)^0 * (0.8)^{(2-0)} * 0.7) / 0.595 =$ **0.753**
f(0.3|0) = 1 – 0.753 = **0.247**

f(0.2|1) = $({}^2C_1 * (0.2)^1 * (0.8)^{(2-1)} * 0.7) / 0.35 =$ **0.64**
f(0.3|1) = 1 – 0.64 = **0.36**

f(0.2|2) = $({}^2C_2 * (0.2)^2 * (0.8)^{(2-2)} * 0.7) / 0.055 =$ **0.51**
f(0.3|2) = 1 – 0.51 = **0.49**

Time taken = 30min

(2) Estimate the proportion of defectives being produced by the machine if the random sample of size 2 yields 2 defects. **(10 points)**

Answer

As described in the part 1. The estimation of the proportion of defectives is as follows:

$f(0.2|2) = ({}^{2}C_{2} * (0.2)^{2} * (0.8)^{(2-2)} * 0.7) / 0.055 =$ **0.51**

$f(0.3|2) = 1 – 0.51 =$ **0.49**

Thus the proportion from "0.2" is higher but it might not be significantly higher from "0.3". Hence we cannot say with certainty what's the proportion of defect.

Time taken = 15min

(b) One of the standard measures (say "acceleration") reported for cars is the time (in seconds) required to reach 0-60 mph. A company determines that the acceleration for their new car is a normal r.v. with a s.d of 0.8 sec. Assume a normal prior distribution, N(8, 0.2). If 10 of the production cars are tested and determined that the average acceleration is 9 sec, then find the 95% Bayesian interval for μ. **(10 points)**

Answer

Given that the acceleration of the new car is normal with s.d. = 0.8

$\mu_0 = 8$

$\sigma_0 = 0.2$

Number of samples = 10

Sample mean = 9

Thus μ^* can be calculated as follows:

$$\mu^* = \frac{(10)(9)(0.2)^2+(8)(0.8)^2}{(10)(0.2)^2+ (0.8)^2} = \frac{3.6 +5.12}{1.04} = 8.385$$

and standard deviation as

$$\sigma^* = \sqrt{\frac{(0.2)^2(0.8)^2}{(10)(0.2)^2+ (0.8)^2}} = \sqrt{\frac{0.0256}{1.04}} = 0.1569$$

The 95% Bayesian interval with $\alpha = 1.96$ for μ is given by

$8.385 – 1.96*0.1569 < \mu < 8.385 + 1.96 * 0.1569$

Thus the interval for μ is

$8.0775 < \mu < 8.6925$

Time taken = 70min

Q2. The following table summarizes two exam scores. Left half of the table gives complete scores and right half gives an example of missing data. (35 points)

Complete Data		Missing Data	
Mt1	Mt2	Mt1	Mt2
74	66	74	66
70	58	70	58
66	74	66	74
55	47	55	47
52	61	52	61
47	38	47	38
45	32	45	
38	46	38	
33	41	33	41
28	44	28	

Answer the following (using data given in the above table):

(1) Based on the missing data, determine missing data pattern and justify your answer (5 points)

Answer

The data given above has a Univariate missing pattern, because the missing values are isolated to a single variable and are relatively rare.

Time taken = 10min

(2) Compute Mean and Standard Error (SE) for (i) complete data, and (ii) missing data using list-wise deletion. (10 points)

Answer

(i) Complete data

Computing the mean for Mt1

Mean = (74 + 70 + 66 + 55+ 52 + 47 + 45 + 38 + 33 + 28) / 10 = 50.8

74	23.2	538.24
70	19.2	368.64
66	15.2	231.04
55	4.2	17.64
52	1.2	1.44
47	-3.8	14.44
45	-5.8	33.64
38	-12.8	163.84
33	-17.8	316.84

28	-22.8	519.84
	Sum	2205.6

SD = $\sqrt{\frac{(x-mean)^2}{n-1}}$

Thus SD = 15.65

SE(of the mean) = $\frac{SD}{\sqrt{n}}$

Thus SE = $\frac{15.65}{\sqrt{10}}$ = 4.949

Computing the mean for Mt2

Mean = (66 + 58 + 74+ 47 + 61 + 38 + 32 + 46 + 41 + 44) / 10 = 50.7

66	15.3	234.09
58	7.3	53.29
74	23.3	542.89
47	-3.7	13.69
61	10.3	106.09
38	-12.7	161.29
32	-18.7	349.69
46	-4.7	22.09
41	-9.7	94.09
44	-6.7	44.89
	Sum	1622.1

SD = $(1622.1 / 9)^{1/2}$ = 13.43

SE = SD / $(10)^{1/2}$ = 13.43 / $(10)^{1/2}$ = 4.245

(ii) Missing data

Computing the mean Mt1 using list elimination

Mean = (74 + 70 + 66 + 55 + 52 + 47 + 33) / 7 = 56.71

Mt1		
74	17.29	298.9441
70	13.29	176.6241
66	9.29	86.3041
55	-1.71	2.9241
52	-4.71	22.1841
47	-9.71	94.2841

33	-23.71	562.1641
	Sum	1243.4287

SD = $(1243.4287 / 6)^{1/2}$ = 14.40
SE = 14.40 / $(7)^{1/2}$ = 5.443

Computing the mean Mt2 using list elimination
Mean = (66 + 58 + 74 + 47 + 61 + 38 + 41) / 7 = 55

Mt2		
66	11	121
58	3	9
74	19	361
47	-8	64
61	6	36
38	-17	289
41	-14	196
	Sum	1076

SD = $(1076 / 6)^{1/2}$ = 13.39
SE = 13.39 / $(7)^{1/2}$ = 5.062

Time taken = 75min

(3) Comment on bias of the estimates of (2.ii) as compared to estimates from complete data (2.i). (5 points)

Answer

The following table summarizes the Mean and Standard Error for each of the attributes for complete and missing data set. As seen from the table the mean and standard errors have increased since the number of sample have decreased due to loss from missing data. As seen in the table, the standard error has increased for both the attributes hence there is a chance that the missing data is related to measured data values in some way.

	Mt1		Mt2	
	Complete Data	Missing Data	Complete Data	Missing Data
Mean	50.8	56.71	50.7	55
SE	4.949	5.443	4.245	5.062

Time taken = 30min

(4) Impute missing data using simple regression (see slide 22 from w15-c1-missing-data). (10 points)

Answer

R code to generate the regression line for missing data

```
# Regression for missing data
mt1 = c(74, 70 , 66 , 55 , 52 , 47 , 33)
mt2 = c(66 , 58 , 74 , 47 , 61 , 38 , 41)
reg_model = lm(formula = mt2 ~ mt1)
print(summary(reg_model))
```

Output of the regression model:

```
lm(formula = mt2 ~ mt1)

Residuals:
     1       2       3       4       5       6       7
 -1.359  -6.499  12.361  -6.774   9.371 -10.055   2.955

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.4516    15.5058   0.932   0.3941
mt1            0.7150     0.2662    2.686   0.0435 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.385 on 5 degrees of freedom
Multiple R-squared:  0.5907,   Adjusted R-squared:  0.5088
F-statistic: 7.216 on 1 and 5 DF,  p-value: 0.04349
```

From the above regression model summary the regression model we get is

mt2 = 0.7150mt1 + 14.4516

Thus the missing values can be calculated using this equation. The mt2 values for mt1 = 45, 38 and 28 can be calculated as follows

mt2 = 0.7150 * 45 + 14.4516 = 46.63
mt2 = 0.7150 * 38 + 14.4516 = 41.62
mt2 = 0.7150 * 28 + 14.4516 = 34.47

Thus, the following table shows all the predicted missing values.

Complete Data		Missing Data	
Mt1	Mt2	Mt1	Mt2
74	66	74	66

70	58	70	58
66	74	66	74
55	47	55	47
52	61	52	61
47	38	47	38
45	32	45	46.63
38	46	38	41.62
33	41	33	41
28	44	28	34.47

Time taken = 60min

(5) Compute Mean and SE on imputed data, and comment on bias of the estimates. (5 points)

Answer

Calculating mean for imputed data.

Mean = (66 + 58+74+47+61+38+46.63+41.62+41+34.47) / 10 = **50.77**

66	15.228	231.891984
58	7.228	52.243984
74	23.228	539.539984
47	-3.772	14.227984
61	10.228	104.611984
38	-12.772	163.123984
46.63	-4.142	17.156164
41.62	-9.152	83.759104
41	-9.772	95.491984
34.47	-16.302	265.755204
	Sum	1567.80236

SD = $(1567.80236 / 9)^{1/2}$ = 13.19

SE = $SD/(10)^{1/2}$ = $13.19/(10)^{1/2}$ = **4.17**

The statistic for complete data for the same attribute are mean = 50.7 and SE = 4.245

With the predicted value we see the that standard error for the imputed data has decreased. Thus, the missing data is related to the measured data and the estimates seem to be close to actual value. Thus reducing the standard error for mean.

Time taken = 40min

Q3. (10 points)

(a) Describe missing data patterns and missing data mechanisms (5 points)

Answer

Missing data patterns provides the configuration of the observed and the missing values in terms of its position or location in the data set. Missing pattern doesn't explain much about why the data is missing. These mainly include:
Univariate pattern: Missing data isolated to a single variable.
Unit nonresponse pattern: Same rows are missing across few attributes of data.
Monotone pattern: Associated with longitudinal study, where the missing values get accumulated over multiple attributes.
General pattern: Random pattern in missing data.
Planned missing pattern: Intentionally missing data pattern. Some attributes that don't need to have values due to other entered parameters.
Latent variable pattern: The data for a variable is missing in the entire sample.

Missing data mechanisms provide the relationship between measured variables and probability of missing data values. It describes a generic mathematical model between the observed data and the missing value.
MAR: In MAR, the probability of missing data on a particular variable is related to other measured variable(s) and not on the values of the variable with the missing value.
MCAR: In MCAR, the probability of the missing data on a particular variable is neither related to other measured variables nor the values of the variable with the missing value. Thus there is no dependency between the missing data and any other measured value.
MNAR: In MNAR, the probability of the missing data on a particular variable is only related to the measured value of the variable with the missing data and not on other measures variable(s). Thus, controlling the values of other measured variables doesn't affect the missing data.

Time taken = 30min

(b) Describe various traditional methods for dealing with missing data, highlight advantage and disadvantages of each method (5 points)

Answer

Traditionally, missing data can be dealt in multiple ways. One of the ways of dealing with missing data is to ignore the missing data entries and only considering the available data for the further analysis. The other way of dealing with missing value is to predict the missing values and then use the predicted values for further analysis. The selection of any of these techniques depends on the type of missing data pattern. If the missing data is related to any other attributes then it can be predicted but in case of random pattern it's better to ignore the missing values. For each of the approaches some of the techniques are discussed below.

Listwise Deletion: In listwise deletion the entire having any missing attribute is deleted from further analysis. This is useful in case the missing values are rare or spread across multiple attributes. If the number of missing values are large then there will be significant loss of sample data and hence the further analysis will be impacted significantly. However the advantage of using this technique is that it stays consistent since we are using the subset of the sample and hence most of the properties should be available in the new sample.

Pairwise Deletion: Pairwise deletion only ignores the missing value instead of deleting the complete row of data. Thus the sample size doesn’t shrink in this case. So even in case of large number of missing values the data size remains the same. Not since different attributes can have different number of entries in the sample data, the further analysis can be biased.

Single Imputation Method: In this method, we do not delete or ignore the missing value, instead we predict the missing value and try to use this predicted value in future analysis of the sample. This method mainly consists of arithmetic mean imputation, regression imputation and stochastic regression imputation. In arithmetic mean imputation the missing value is replaced with the mean value of the attribute. In regression imputation the missing value is replaced with the value generated from the regression equation generated using the linear relation between the measured values of other attributes or the attribute itself. Finally the stochastic regression imputation is similar to regression model but additionally adds a normally distributed residual term randomly to each missing value. The advantage of using the imputation is it explores the relationship between the missing data and the existing measured data rather than ignoring it and thus might impact the accuracy of the future analysis. On the other hand it is important that such a relationship exists. If no pattern exists in the missing data and the predictive model cannot be easily constructed then predicting the missing data with incorrect values might lead to inconsistency in sample data.

Time taken = 45min

Q4. Bayes Networks (25 points)

Part 1

For each independence statement, state whether the independence is implied by the given Bayes net. If there are any active trails between the given variables, name one.

Answer

1. $A \perp C$
   D-separated. Has no active trails
2. $A \perp C \mid E$
   D-separated. Has no active trails
3. $A \perp C \mid I$
   Not D-separated. Has active trails. One of the active trail is A-D-G-I-F-C
4. $D \perp I$
   Not D-separated. Has active trails. One of the active trail is D-G-I
5. $D \perp I \mid G$
   Not D-separated. Has active trails. One of the active trail is D-F-I
6. $D \perp I \mid F, G, J$
   Not D-separated. Has active trails. One of the active trail is D-A-E-J-I
7. $D \perp I \mid F, G, J, A$
   D-separated. Has no active trails.
8. $F \perp H$
   Not D-separated. Has active trails. One of the active trail is F-D-H
9. $F \perp H \mid A$
   Not D-separated. Has active trails. One of the active trail is F-D-H
10. $F \perp H \mid D$
    D-separated. Has no active trails
11. $F \perp H \mid D, J$
    Not D-separated. Has active trails. One of the active trail is F-I-J-E-H
12. $F \perp H \mid D, I, J$
    D-separated. Has no active trails

Time taken = 75min

Part 2

Compute:

a) P(A=1,B=0,C=0,D=1,E=0,F=1)

Answer

By chain rule the joint probability can be given by

P(A, B, C, D, E, F) = P(A)P(B| A)P(C| A, B)P(D| A, B, C)P(E| A, B, C, D)P(F| A, B, C, D, E)

Now by local dependencies we have

$B \perp A, C$

$C \perp B \mid A$

$D \perp C, A \mid B$

$E \perp A, B \mid C, D$

$F \perp A, B, C, D \mid E$

Thus the equation now becomes
P(A, B, C, D, E, F) = P(A)P(B)P(C| A)P(D| A, B)P(E| C, D)P(F| E)
Thus
P(A=1, B=0, C=0, D=1, E=0, F=1)
= P(A=1)P(B=0)P(C=0| A=1)P(D=1| A=1, B=0)P(E=0| C=0, D=1)P(F=1| E=0)
= 0.7 * 0.6 * 0.3 * 0.6 * 0.4 * 0.9
= **0.027216**

Time taken = 30min

b) P(A=1,E=0)
Use the elimination ordering F, D, C, B (F is the innermost sum)
Answer
From the above part
P(A, B, C, D, E, F) = P(A)P(B)P(C| A)P(D| A, B)P(E| C, D)P(F| E)
Now, P(A=1, E=0)
= $\Sigma_B\Sigma_C\Sigma_D\Sigma_F$ P(A=1)P(B)P(C| A=1)P(D| A=1, B)P(E=0| C, D)P(F| E=0)
= $\Sigma_B\Sigma_C\Sigma_D$ P(A=1)P(B)P(C| A=1)P(D| A=1, B)P(E=0| C, D) Σ_F P(F| E=0)
= $\Sigma_B\Sigma_C\Sigma_D$ P(A=1)P(B)P(C| A=1)P(D| A=1, B)P(E=0| C, D)
Since Σ_F P(F| E=0) = 1
Now expanding for D we get
= $\Sigma_B\Sigma_C$ P(A=1)P(B)P(C| A=1) [P(D=0| A=1, B)P(E=0| C, D=0) + P(D=1| A=1, B)P(E=0| C, D=1)]

Now expanding for C we get
= Σ_B P(A=1)P(B) [P(C=0| A=1)P(D=0| A=1, B)P(E=0| C=0, D=0) + P(C=0| A=1)P(D=1| A=1, B)P(E=0| C=0, D=1) + P(C=1| A=1)P(D=0| A=1, B)P(E=0| C=1, D=0) + P(C=1| A=1)P(D=1| A=1, B)P(E=0| C=1, D=1)]
= Σ_B P(A=1)P(B) [0.3 * P(D=0| A=1, B)* 0.2+ 0.3 * P(D=1| A=1, B) * 0.4 + 0.7 * P(D=0| A=1, B) * 0.9 + 0.7 * P(D=1| A=1, B) * 0.5]
= Σ_B P(A=1)P(B) [0.06 * P(D=0| A=1, B) + 0.12 * P(D=1| A=1, B) + 0.63 * P(D=0| A=1, B) + 0.35 * P(D=1| A=1, B)]

Now expanding for B we get
= P(A=1)[P(B=0) * 0.06 * P(D=0| A=1, B=0) + P(B=0)*0.12 * P(D=1| A=1, B=0) + P(B=0)*0.63 * P(D=0| A=1, B=0) + P(B=0) * 0.35 * P(D=1| A=1, B=0) + P(B=1) * 0.06 * P(D=0| A=1, B=1) + P(B=1) *0.12 * P(D=1| A=1, B=1) + P(B=1)*0.63 * P(D=0| A=1, B=1) + P(B=1) * 0.35 * P(D=1| A=1, B=1)]
= 0.7 [0.6 * 0.06 * 0.4 + 0.6 * 0.12 * 0.6 + 0.6 * 0.63 * 0.4 + 0.6 * 0.35 * 0.6 + 0.4 * 0.06 * 0.8 + 0.4 * 0.12 * 0.2 + 0.4 * 0.63 * 0.8 + 0.4 * 0.35 * 0.2]
= 0.7 [0.0144 + 0.0432 + 0.1512 + 0.126 + 0.0192 + 0.0096 + 0.2016 + 0.028]
= 0.7 [0.3348 + 0.2584]

= **0.41524**

Time taken = 50min