Estimating Synthetic Cases through Regression

There are two ways to estimate missing outcomes. Question 4 asks you to do so using marginal information. The following shows you how to do the same using regression. It is important that you learn how to do this using regression as well as the question will definitely be part of final exam:

Step 1 is to prepare the data for regressing the odds of mortality in 6 months on various features:

```
DROP TABLE #Data
SELECT [Column 0] as ID,[Column 1] as Age ,[Column 2] as Gender
     ,[Column 3] as Num_Assessment,[Column 4] as Days_Followed,[Column 5] as First_Assess
      ,[Column 6] as Last_Assess,[Column 7] as Unable_Eat,[Column 8] as Unable_Transfer
     ,[Column 9] as Unable_Groom,[Column 10] as Unable_Toilet,[Column 11] as Unable_Bathe
      ,[Column 12] as Unable_Walk,[Column 13] as Unable_Dress,[Column 14] as Unable_Bowel
      ,[Column 15] as Unable_Urine,[Column 16] as Survival,[Column 17] as Assessment_Num
         ,[Column 18] as Dead6Months
INTO #Data
FROM [mfh].[dbo].[Assessments]
Where [Column 18] <> 'Null'
-- (1105296 rows affected)
Select
age, Unable_Eat, Unable_transfer, Unable_Groom, unable_Bathe, Unable_dress
, unable_Bowel, Unable_Urine, Unable_Toilet, Unable_Walk
, (Sum(CAST(Dead6Months AS Float))/Cast(count(Dead6Months) AS float))
/(1-(Sum(CAST(Dead6Months AS Float))/Cast(count(Dead6Months) AS float))) AS Odds
FROM #Data
GROUP BY age, Unable_Eat, Unable_transfer, Unable_Groom, unable_Bathe, Unable_dress
, unable_Bowel, Unable_Urine, Unable_Toilet, Unable_Walk
Having count(ID) >9 and (Sum(CAST(Dead6Months AS Float))/Cast(count(Dead6Months) AS
float))<1
```

In step 2, transfer the data to Excel to regress the odds of mortality on patient features. This is how the first 4 lines of data looks like:

age	Unable Eat	Unable transfer	Unable Groom	Unable Bathe	Unable Dress	Unable Bowel	Unable Urine	Unable Toilet	Unable Walk	Odds
66	0	0	1	1	0	0	1	1	0	0.15
81	1	1	1	1	1	0	1	1	1	0.47
80	0	0	1	1	0	1	0	0	1	0.46
95	0	0	1	1	1	0	0	0	1	0.48

To do the regression in Excel you go to Data and pick Data Analysis and pick Regression. You regress Odds on all remaining variables. The resulting regression looks like this:

SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.532362869

R Square	0.283410225
Adjusted R Square	0.281816386
Standard Error	0.430085711
Observations	4507

ANOVA

	df	SS	MS	F
Regression	10	328.9132	32.89132	177.8162
Residual	4496	831.6418	0.184974	
Total	4506	1160.555		

	Coefficients	Standard Error	t Stat	P-value
Intercept	-0.62	0.03	-18.78	0.00
age	0.01	0.00	27.03	0.00
Unable Eat	0.38	0.02	23.42	0.00
Unable transfer	0.04	0.02	2.38	0.02
Unable Groom	0.10	0.02	6.34	0.00
Unable Bathe	0.02	0.02	1.00	0.32
Unable dress	-0.04	0.01	-2.79	0.01
Unable Bowel	0.12	0.01	8.33	0.00
Unable Urine	-0.17	0.01	-12.53	0.00
Unable Toilet	-0.01	0.02	-0.89	0.37
Unable Walk	0.14	0.01	9.99	0.00

In the final step we calculate the missing value by evaluating the regression at the features of the patient: 80 years, unable to walk, unable to toilet but able to do everything else. The function sumproduct can be used to do so:

Coefficients	Resident's Description
-0.62	1.00
0.01	80.00
0.38	0.00
0.04	0.00
0.10	0.00
0.02	0.00
-0.04	0.00
0.12	0.00
-0.17	0.00
-0.01	1.00
0.14	1.00

Predicted Odds 0.396316
Predicted Prob 0.28383

The missing value is estimated to be 0.28. Previously using marginals we had estimated it to be 0.25. The two estimates are close.