Abstract: A comparison study was undertaken to compare listwise deletion versus multiple imputation. The dataset used had 89 total observations of which 63 were missing. Using SAS, the missing data pattern was found to be arbitrary, i.e. non-monotone therefore MCMC was used as the imputation method. The number of imputations was set at 5. Each imputed dataset was then fitted with a linear regression model. These regression models used all 89 observations. The results of the regression across the imputations were then combined, and compared to the original model. One difference between the regression models is that DeptS would be in the original regression however this variable would be excluded from the imputed model. Next steps include attempting to replicate the case study using R.

Introduction: There are many different approaches with dealing with missing values. A common approach, called listwise deletion, is to remove observations that having missing data. This reduces sample size which potentially could affect a model. An alternative, called imputation, is to replace a missing data point with an estimate. There are many imputation formulas, and on that has gained accolades in academia is called multiple imputation (MI). In MI, the original dataset is cloned, and then N different imputed datasets are created. The formula used to calculate the imputed values depends on the pattern of missing data. The imputed datasets are then combined to make an average imputed dataset that is then used for modeling. This case study will compare linear regression models created by both listwise deletion and multiple imputation.

Literature Review: The main asynchronous class lectures were given by Alan Elliot, who has authored a book on SAS programming[1]. These lectures covered the general concepts of multiple imputation and SAS Code.

Method: The steps used for this analysis were: (1) linear regression was run with listwise deletion, (2) the missing data pattern was reviewed (3) multiple imputation was performed to replace the missing values (4) regression was ran with each imputed data, (5) the multiple imputation results were combined, and (6) the results of listwise deletion and multiple imputation were compared.

Results:

(1) **Liner Regression Using Dataset with Missing Values.** The SAS code used for running linear regression was:

```
PROC REG DATA = WORK.IMPORT1;
        MODEL TOTBPT = SEXP DEPTP ANXTP DEPTS ANXTS;
RUN;
```

The main observation after running the regression analysis was that ~70% observations (63 out of 89) were automatically removed. Only 26 observations were used for the model. Hence, there was a concern that this low sample size might not produce an accurate predictive model.[2] The output from this regression will be compared to the imputed regressions later in this study.

[1] Elliot, A. (2015), SAS Essentials: A Guide to Mastering SAS, 2nd Edition, Wiley

[2] Please note that a subset of predictor variables was used in the predictive model

Number of Observations Read	89
Number of Observations Used	26
Number of Observations with Missing Values	63

(2) **Review the Missing Data Pattern:** The SAS code for obtaining the missing data pattern was:

```
ODS SELECT MISSPATTERN;
PROC MI DATA=WORK.IMPORT1 NIMPUTE=0;
VAR SEXP DEPTP ANXTP GSITP DEPTS ANXTS GSITS SEXCHILD TOTBPT;
RUN;
```

The missing data pattern shows to be arbitrary, or non-monotone.

Missing Data Patterns											
Group	**SexP**	**DeptP**	**AnxtP**	**GSItP**	**DeptS**	**AnxtS**	**GSItS**	**SexChild**	**Totbpt**	**Freq**	**Percent**
1	X	X	X	X	X	X	X	X	X	26	29.21
2	X	X	X	X	X	X	X	.	.	26	29.21
3	X	X	X	X	.	.	.	X	X	11	12.36
4	X	X	X	X	.	.	.	.	.	16	17.98
5	X	.	.	.	X	X	X	.	.	1	1.12
6	X	.	.	.	.	.	.	.	.	2	2.25
7	.	.	.	.	X	X	X	X	X	4	4.49
8	.	.	.	.	X	X	X	.	.	3	3.37

Furthermore, based on this pattern of missingness, the recommend imputation method was MCMC full-data imputation.[3]

(3) **Perform Multiple Imputation to Replace the Missing Values:** The SAS code used for performing the multiple imputation is shown below. Please note that MCMC was the default method therefore that line of code was not needed. The number of imputation was set at 5. Also, the output was placed into MIOUT which will be used in subsequent steps. The seed number allows for the analysis to be replicated such that the results produced in this case study are the same as those presented in the class lectures.

[3]Source: Table 56.5 Imputation Methods in PROC MI
https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect019.htm

```
PROC MI DATA = WORK.IMPORT1
OUT = MIOUT seed = 35399 NIMPUTE=5;
VAR SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild Totbpt;
RUN;
```

As shown in the below SAS output table, the multiple imputation was successful executed.

Model Information	
Data Set	WORK.IMPORT1
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	35399

(4) **Regression With Each Imputed Data:** The SAS code used for performing regression on each imputed dataset was as follows:

```
PROC REG data = MIOUT outest = outreg covout;
Model Totbpt = SexP DeptP AnxtP DeptS AnxtS;
by _Imputation_;
RUN;
```

Please note that MIOUT from the prior multiple imputation step was the input data. The BY statement was used to run regression for each of the 5 imputed datasets. Some of the SAS output is shown below. Whereas the original dataset only used 26 observations, the imputed datasets used all 89 observations.

Number of Observations Read	89
Number of Observations Used	89

The parameter estimates for each of the five models are shown in the below table. In the next step, these values will be averaged together.

Variable	Original	1	2	3	4	5
Intercept	-2.93924	-1.66002	-3.10217	-12.77419	-6.25011	-4.28616
Sep.	-3.76858	-4.87081	-3.68446	-4.02550	1.29085	-5.64106
Depp	0.88845	0.89628	0.81093	0.76927	0.73950	0.61804
AnxtP	-0.06422	-0.20472	-0.13401	0.10668	0.13852	0.20884
DeptS	-0.35464	-0.23504	-0.57784	-0.39213	-0.29757	-0.20440
AnxtS	0.60805	0.69262	0.97745	0.76526	0.39837	0.55903

(5) **Multiple Imputation Results Combined:** The SAS command for combing the outputs of the regression analysis is PROC MIANALYZE. The code used to combine the 5 different imputations is shown below.

```
PROC MIANALYZE data = outreg;
   MODELEFFECTS SexP DeptP AnxtP DeptS AnxtS Intercept;
RUN;
```

The below table summarizes some of the key metrics from both the original regression with the missing values and from the multiple imputations. The original regression was based on 26 observations whereas the multiple imputation allowed for use of all 89 observations. One major difference between the regression models is that based upon a cut-off of 0.05 for Pr > |t| the original regression would include DeptS as part of its final model however this variable would be excluded from the imputed model.

	Parameter Estimates		Standard Errors		Pr > |t|	
Variable	**Original**	**Imputations**	**Original**	**Imputations**	**Original**	**Imputations**
Intercept	-2.93924	-5.614531	12.00345	8.520291	0.8091	0.5136
SexP	-3.76858	-3.386193	2.80346	3.596210	0.1939	0.3725
DeptP	0.88845	0.766804	0.20224	0.156718	0.0003	0.0002
AnxtP	-0.06422	0.023061	0.16908	0.226807	0.7081	0.9219
DeptS	-0.35464	-0.341397	0.15540	0.195687	0.0336	0.1194
AnxtS	0.60805	0.678545	0.16604	0.259533	0.0015	0.0426

Discussion & Future Work: Multiple imputation allowed for observations with missing data to be used as otherwise this data would be deleted during regression analysis. This introductory case study helped to show the general process. SAS MI procedures make it easy to perform multiple imputation analysis. Next steps include attempting to replicate the case study using R.

Appendix: SAS Code

```
PROC CORR DATA=WORK.IMPORT1 PLOTS=MATRIX(HISTOGRAM NVAR=9);
VAR SEXP DEPTP ANXTP GSITP DEPTS ANXTS GSITS SEXCHILD TOTBPT;
RUN;

PROC REG DATA = WORK.IMPORT1;
MODEL TOTBPT = SEXP DEPTP ANXTP DEPTS ANXTS;
RUN;

ODS SELECT MISSPATTERN;
PROC MI DATA=WORK.IMPORT1 NIMPUTE=0;
VAR SEXP DEPTP ANXTP GSITP DEPTS ANXTS GSITS SEXCHILD TOTBPT;
RUN;

PROC MI DATA = WORK.IMPORT1
OUT = MIOUT seed = 35399 NIMPUTE=5;
VAR SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild Totbpt;
RUN;

PROC REG data = miout outest = outreg covout ;
Model Totbpt = SexP DeptP AnxtP DeptS AnxtS;
by _Imputation_;
RUN;

PROC MIANALYZE data = outreg;
MODELEFFECTS SexP DeptP AnxtP DeptS AnxtS Intercept ;
RUN;
```