Handling Missing Data

Ashley Parker
EDU 7312

Presentation Outline

- Types of Missing Data
- Treatments for Handling Missing Data
  - Deletion Techniques
    - Listwise Deletion
    - Pairwise Deletion
  - Single Imputation Techniques
    - Mean Imputation
    - Hot Deck Imputation
  - Multiple Imputation Techniques
- Practice in R
- Simulating Missing Data and Treatments

Missing Data Assumptions

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

Missing Completely at Random (MCAR)

- **One value is just as likely to be missing as another**
- **No relationship between the missing data and the other measured variables**
- **Probability for missing data is the same across units – considered "ignorable"**
- **Many missing data techniques are valid only if the MCAR assumption holds (Allison, 2003)**
  - **Examples –**
    - Child is absent and does not receive a score on a progress monitoring assessment for the day
    - A man does not report income level because he accidentally skipped a line on the survey

Missing at Random (MAR)

- **Extent that missingness is correlated with other variables that are included in the analysis**
- **Allows missing data to depend on things that are observed, but not on things that are not observed** **(Allison, 2003)**
  - **Examples –**
    - Less educated individuals tend not to report their income, therefore the missing income values could be dependent on a person's education.
    - Women report their weight on a survey less often than men, therefore the missing value could depend on gender.

Missing Not at Random (MNAR)

- **Likelihood of a piece of data being missing is related to the value that would have been observed**
- **Most problematic type – considered "nonignorable" missing data**
  - **Examples –**
    - Individuals with low income tend to not report their income
    - Students who struggle with division are more likely to skip problems that require them to divide

Problems with Missing Data

- Can lead to bias in parameter estimates and standard errors
- Can minimize the variability in a data set
- Can lead to inefficient use of the data
- Can inflate Type 1 and Type 2 errors

Types of Missing Data Treatments

Deletion Techniques

- Listwise Deletion
- Pairwise Deletion

Single Imputation Techniques

- Mean Imputation
- Hot Deck Imputation

Multiple Imputation Techniques

Listwise Deletion

- **Simply drop all cases with missing values; if a participant is missing a data point, all of the data for that participant is deleted**
- This is the default approach in most software programs
- Also known as complete case analysis
- **Advantages in using this treatment –**
  - Easy to complete
  - Will not introduce any bias into the parameter estimates
- **Disadvantages in using this treatment –**
  - Decreases the sample size (thus the statistical power)
  - Increases the standard error and widens the confidence intervals

Example of Listwise Deletion

DV	IV1	IV2	IV3	IV4
64	38	82	74	NA
72	46	NA	71	81
83	47	27	64	92
91	24	52	77	62

DV	IV1	IV2	IV3	IV4
83	47	27	64	92
91	24	52	77	62

Pairwise Deletion

- **Only removes cases that have missing data when calculating a specific variable**
- Also known as available case analysis
- **Advantages in using this treatment –**
  - Preserves more of the data than listwise deletion
- **Disadvantages in using this treatment –**
  - Parameters in the model will be based on different sets of data – different sample sizes, different standard errors
  - Can introduce bias if data is not MCAR

Example of Pairwise Deletion

DV	IV1	IV2	IV3	IV4
64	38	82	74	NA
72	46	NA	71	81
83	NA	27	64	92
91	24	52	77	62

Removed missing data from IV2 since it is the variable being used in the analysis

DV	IV1	IV2	IV3	IV4
64	38	82	74	NA
83	NA	27	64	92
91	24	52	77	62

Single Imputation Treatments

- Imputation – substituting a missing data point with a value
- **Single Imputation – aims to replace each missing data with one plausible value**
- **Two types of Single Imputation Treatments –**
  - Mean Imputation
  - Hot Deck Imputation

Mean Imputation

- Replace a missing data point with the mean of the available data points for that variable
- Frequently used method
- **Advantage in using this treatment –**
  - Retains sample size since participants with missing data are not removed from the data set
- **Disadvantage in using this treatment –**
  - Decreases the standard deviation and standard errors; creates smaller confidence intervals

Example of Mean Imputation

DV	IV1	IV2	IV3	IV4
64	NA	82	74	NA
72	46	NA	71	81
83	47	27	64	92
91	24	52	77	62

Means:

78	39	54	72	78

DV	IV1	IV2	IV3	IV4
64	**39**	82	74	**78**
72	46	**54**	71	81
83	47	27	64	92
91	24	52	77	62

Hot Deck Imputation

- Missing data point is filled in with a value from a similar observation in the current data set – also known as "matching"
  - If the observations have the same value for x, then the non-missing y is substituted for the missing data point
  - If multiple observations are similar, then the mean of all similar values is used to replace missing value
- **Advantage in using this treatment –**
  - Retains sample size since participants with missing data are not removed from the data set
- **Disadvantages in using this treatment –**
  - Reduces standard errors by underestimating the variability of a given variable
  - Becomes much more difficult as variables with missing data increase

Cold Deck Imputation is similar, only the data is taken from another existing data source

Example of Hot Deck Imputation

Weight (DV)	Height (IV)
260	66
190	68
NA	66
215	72
145	62
NA	62

Weight (DV)	Height (IV)
290	66
190	68
260	66
215	72
145	62
145	62

Multiple Imputation Treatments

- Each missing value is replaced with multiple plausible values to generate "complete data sets"
- R will impute multiple possible data sets, run an analysis on each data set, and pool the results to come up with one average of the estimates
- Generally, 3 – 5 imputations are sufficient
- **Advantage in using this treatment –**
  - Having multiple values reduces bias by addressing the uncertainty
- **Disadvantage in using this treatment –**
  - Highly technical and difficult to compute

Types of Multiple Imputation Treatments

- Predictive Mean Matching (pmm)
- Multivariate Imputation by Chained Equations (mice)
- Baysian Linear Regression (norm)
- Markov Chain Monte Carlo (norm)
- Logistic Regression (logreg)
- Linear Discriminant Analysis (lda)
- Random Sample (sample)
- Many others!

Comparing Bias

- Using certain data treatments to handle missing data is likely to introduce bias into your model.
- The Percent Relative Parameter Bias (PRPB) measures the amount of bias introduced under a specific set of conditions, such as a missing data treatment.
- The Relative Standard Error Bias (RSEB) is also used to calculate the bias introduced by missing data treatments, specifically the amount of bias in the standard error estimates.

Practice in R

- Create a data frame in R and name it "practice"
- Run regression with Y as the DV and X as the IV

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.3291     2.4450   2.998    0.015 *
x            -0.3249     0.3827  -0.849    0.418
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.083 on 9 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared: 0.07416,    Adjusted R-squared: -0.02871
F-statistic: 0.7209 on 1 and 9 DF,  p-value: 0.4179
```

Y	X
8	2
7	4
3	NA
6	7
9	NA
1	4
10	7
2	6
9.5	5
7.5	NA
4	9
5	3
5	8
3.3	NA
2	10

Practice in R – Listwise Deletion

- Listwise Deletion

  practicelistwise<-na.omit(practice)

- Run regression with Y as the DV and X as the IV

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.3291     2.4450   2.998    0.015 *
x            -0.3249     0.3827  -0.849    0.418
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 9 degrees of freedom
Multiple R-squared: 0.07416,   Adjusted R-squared: -0.02871
F-statistic: 0.7209 on 1 and 9 DF,  p-value: 0.4179
```

Practice in R – Mean Imputation

- Mean Imputation Code

```
library(Hmisc)

practicemean<-practice

practicemean$x<-impute(practicemean$x, mean)

practicemean$x
```

- Run regression with Y as the DV and X as the IV

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.4067     2.2923   3.231  0.00656 **
x            -0.3249     0.3659  -0.888  0.39068
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.948 on 13 degrees of freedom
Multiple R-squared:  0.05719,   Adjusted R-squared: -0.01534
F-statistic: 0.7885 on 1 and 13 DF,  p-value: 0.3907
```

Practice in R – Hot Deck Imputation

- Hot Deck Imputation Code

```
install.packages("rrp", repos="http://R-Forge.R-project.org")
library(rrp)
practicehd<-rrp.impute(practice)
practicehd1<-practicehd$new.data
```

- Run regression with Y as the DV and X as the IV

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.2715     1.7536   4.717 0.000403 ***
x            -0.4590     0.2646  -1.735 0.106360
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.736 on 13 degrees of freedom
Multiple R-squared: 0.188,    Adjusted R-squared: 0.1256
F-statistic:  3.01 on 1 and 13 DF,  p-value: 0.1064
```

Practice in R – Multiple Imputation

- Multiple Imputation Code

```
library(mice)
practicemi<-mice(practice, meth=c(" ","pmm"), maxit=1)
practicemi2<-with(practicemi, lm(y~x))
practicepooled<-pool(practicemi2)
pool.r.squared(practicemi2)
```

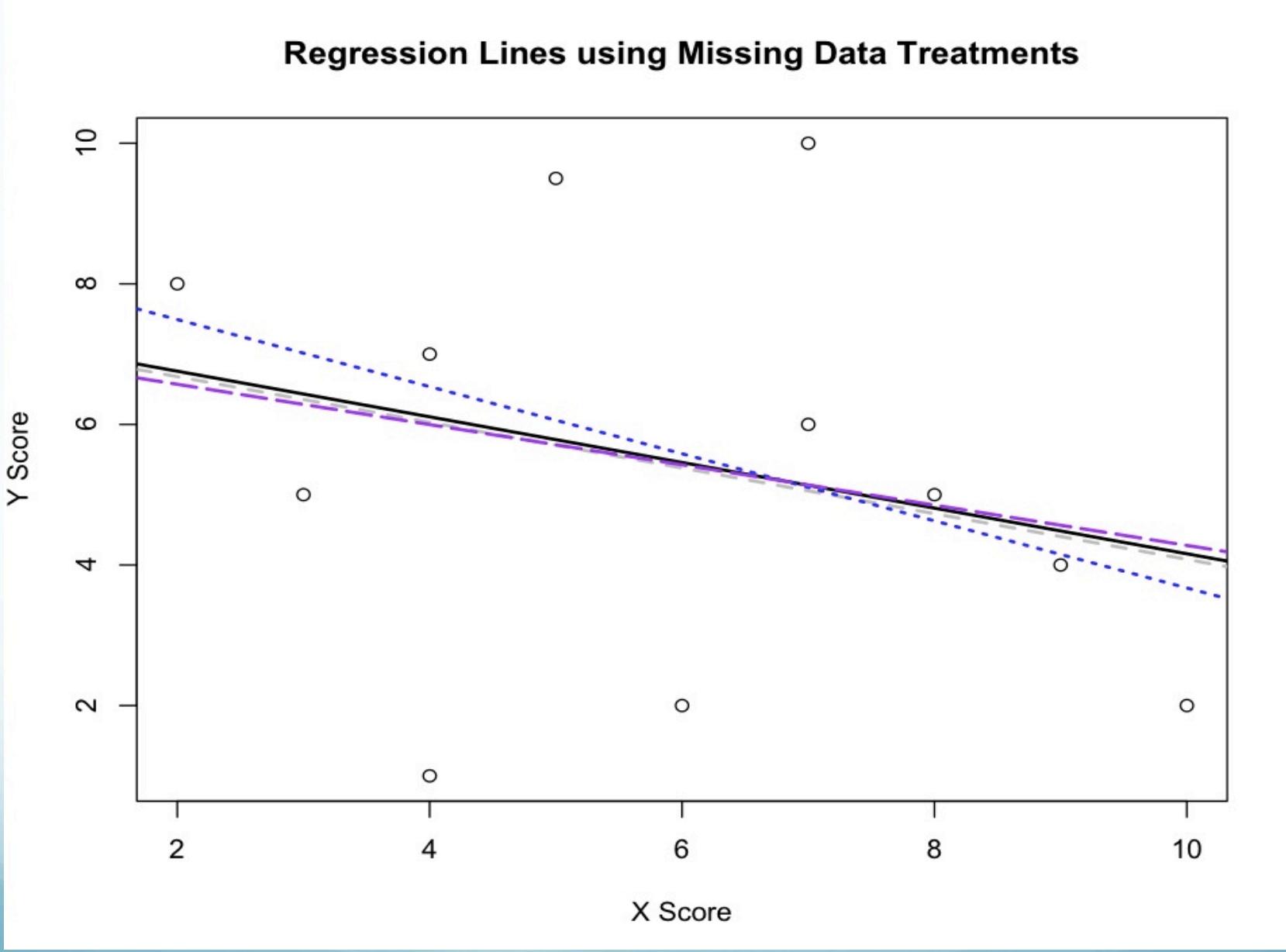
- Run regression with Y as the DV and X as the IV

```
Pooled coefficients:
(Intercept)           x
  6.9616439  -0.2487871

Fraction of information about the coefficients missing due to nonresponse:
(Intercept)           x
  0.4727954   0.4919883
```

```
          est     lo 95    hi 95       fmi
R^2 0.08738677 0.1562605 0.597364 0.4284726
```

Practice in R – Comparing Methods



Listwise = Grey Mean Imputation = Black Hot Deck = Blue Multiple Imputation = Purple

Simulation in R

- Population = 100,000
- Variables = DV and IV
- Randomly generated 5 subsets, n= 5,000
- Created 3 datasets from each subset with 1%, 5%, and 10% missingness in the IV
- Performed listwise deletion, mean imputation, hot deck imputation, and multiple imputation on each dataset (15 total datasets x 4 treatments = 60 outputs)
- Compared intercept and slope for each treatment in each data set

Simulation in R

1% Missingness in Each Subset

New Data 1.1

Method	Intercept	Slope	R^2
None Missing	-114.1440	1.9885	.7949
Listwise	-117.2278	2.0278	.8021
Mean Imp.	-117.3723	2.0278	.795
Hot Deck	-117.3723	2.0296	.8035
Multiple Imp.	-117.1425	2.0268	.8020

New Data 3.1

Method	Intercept	Slope	R^2
None Missing	-115.003	2.0012	.8047
Listwise	-116.2192	2.0153	.7971
Mean Imp.	-116.2075	2.0153	.7954
Hot Deck	-116.4186	2.0179	.8028
Multiple Imp.	-116.1000	2.0138	.8011

New Data 5.1

Method	Intercept	Slope	R^2
None Missing	-115.7178	2.0081	.8051
Listwise	-114.8106	1.9979	.7981
Mean Imp.	-114.8080	1.9979	.7923
Hot Deck	-114.9122	1.9992	.7991
Multiple Imp.	-114.7392	1.9970	.7976

5% Missingness in Each Subset

New Data 1.5

Method	Intercept	Slope	R^2
None Missing	-114.1440	1.9885	.7949
Listwise	-117.2089	2.0275	.8020
Mean Imp.	-117.2170	2.0275	.7577
Hot Deck	-118.3151	2.0414	.8107
Multiple Imp.	-117.6190	2.0330	.8025

New Data 3.5

Method	Intercept	Slope	R^2
None Missing	-115.003	2.0012	.8047
Listwise	-116.4058	2.0171	.8023
Mean Imp.	-116.3857	2.0171	.7618
Hot Deck	-117.8368	2.0354	.8082
Multiple Imp.	-116.3259	2.0161	.8014

New Data 5.5

Method	Intercept	Slope	R^2
None Missing	-115.7178	2.0081	.8051
Listwise	-114.7534	1.9973	.7993
Mean Imp.	-114.7593	1.9973	.762
Hot Deck	-115.8060	2.0108	.8052
Multiple Imp.	-114.5879	1.9951	.7981

10% Missingness in Each Subset

New Data 1.10

Method	Intercept	Slope	R^2
None Missing	-114.1440	1.9885	.7949
Listwise	-117.3031	2.0287	.8028
Mean Imp.	-117.3605	2.0287	.7205
Hot Deck	-119.4152	2.0554	.8169
Multiple Imp.	-117.3125	2.0286	.8036

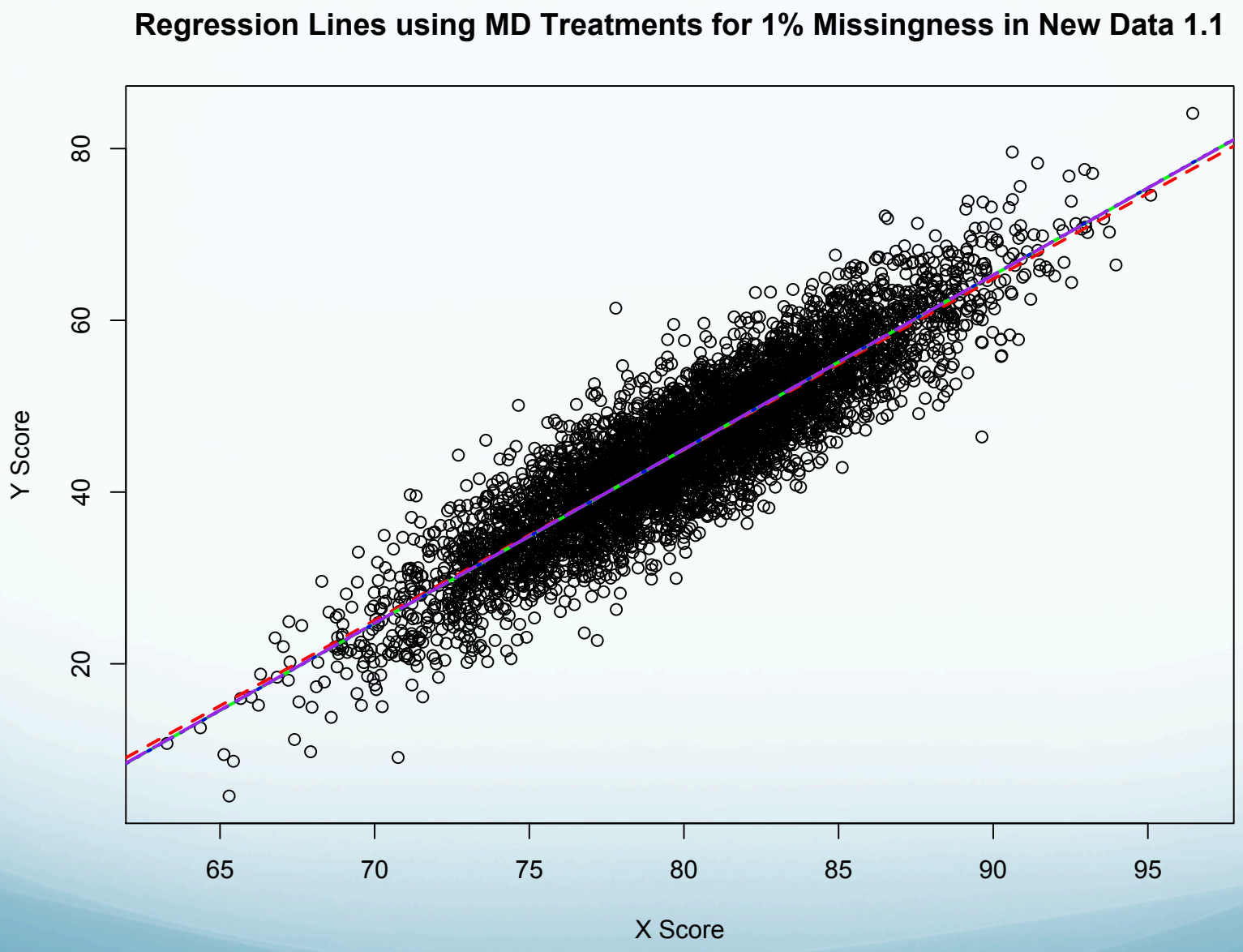
New Data 3.10

Method	Intercept	Slope	R^2
None Missing	-115.003	2.0012	.8047
Listwise	-116.4706	2.0180	.8027
Mean Imp.	-116.3864	2.0180	.7172
Hot Deck	-119.2585	2.0539	.8161
Multiple Imp.	-116.4768	2.0182	.8042

New Data 5.10

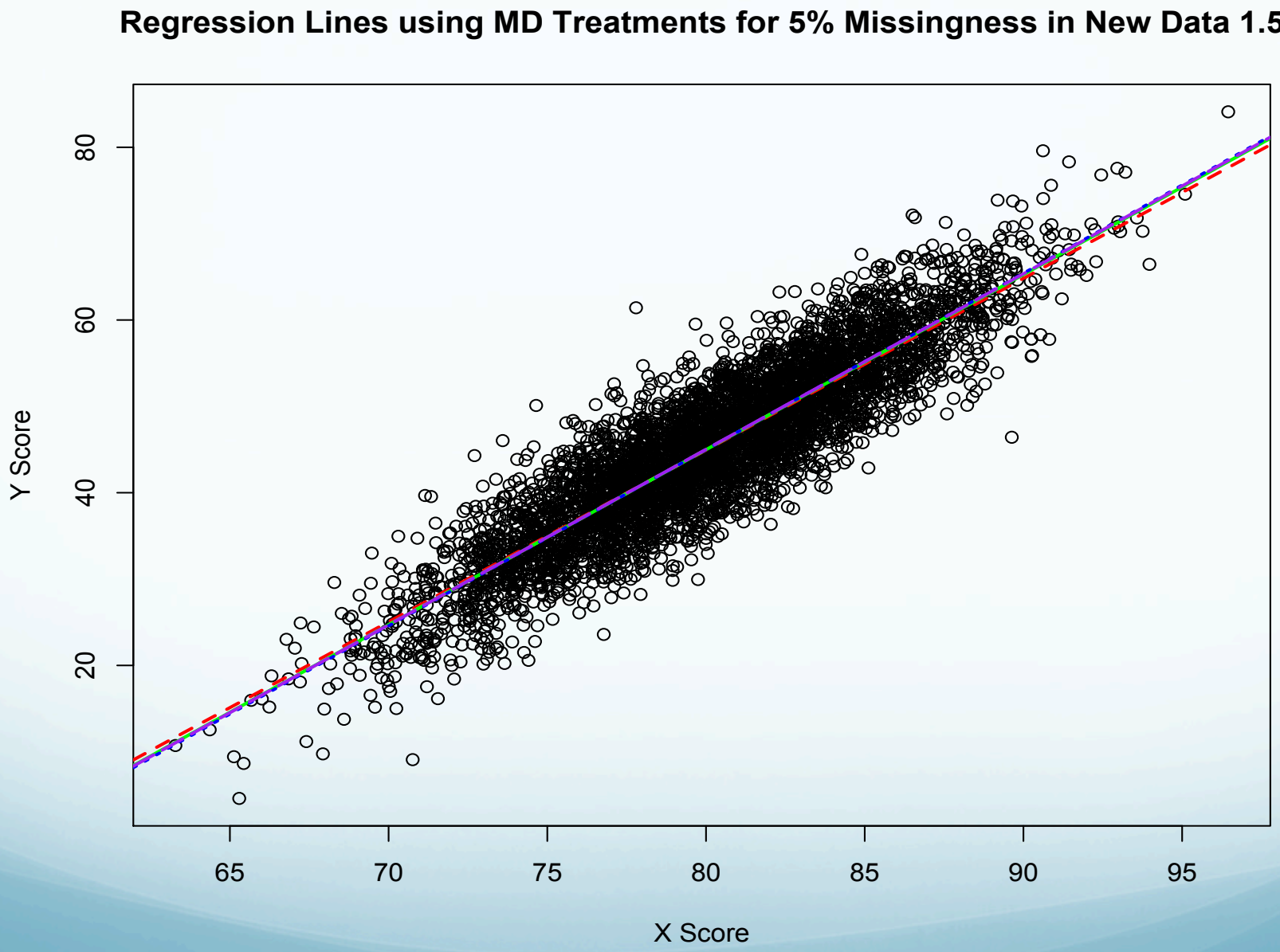
Method	Intercept	Slope	R^2
None Missing	-115.7178	2.0081	.8051
Listwise	-114.5603	1.9944	.7975
Mean Imp.	-114.5335	1.9944	.7122
Hot Deck	-117.8440	2.0368	.8103
Multiple Imp.	-114.8241	1.9977	.7992

Visual Inspection – Graphs



No Missingness = Red **Listwise = Grey** **Mean Imputation = Green** **Hot Deck = Blue** **Multiple Imputation = Purple**

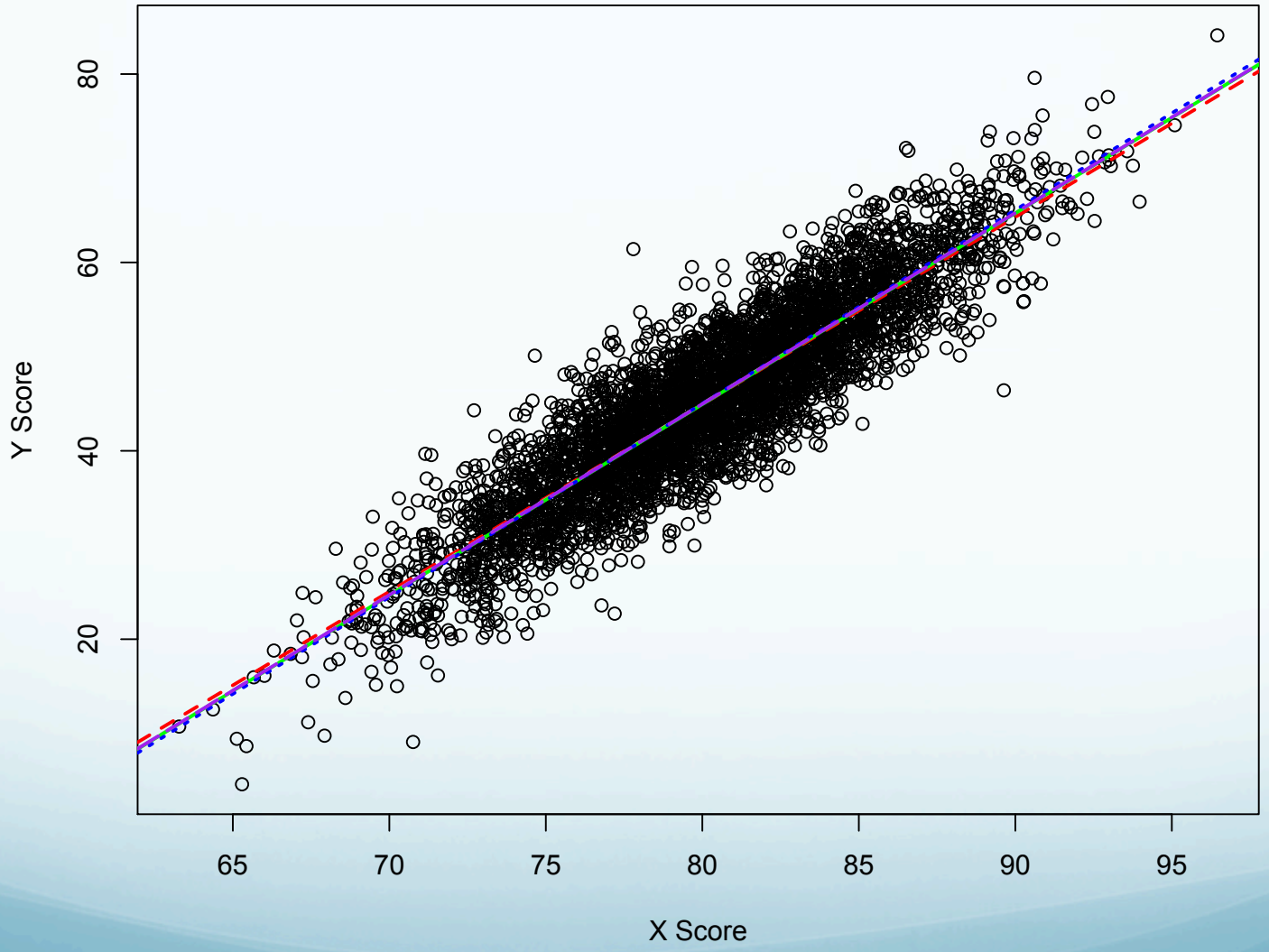
Visual Inspection – Graphs



No Missingness = Red **Listwise = Grey** **Mean Imputation = Green** **Hot Deck = Blue** **Multiple Imputation = Purple**

Visual Inspection – Graphs

Regression Lines using MD Treatments for 10% Missingness in New Data 1.10



No Missingness = Red **Listwise = Grey** **Mean Imputation = Green** **Hot Deck = Blue** **Multiple Imputation = Purple**

Conclusions

- Important to deduce why data is missing in order to choose a correct treatment
- Avoid missing data if at all possible
- There isn't a magic way to solve the NA's, therefore listwise deletion appears to be best in most scenarios (but sample size is important!)
- Wad of Gum and Open Face Reel Analogies

References

- Allison, P.D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545–557.
- Batista, G.E.A.P.A. & Monard, M.C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence, 17*(5), 519–533.
- Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York, NY: Cambridge University Press.
- Howell, D.C. (2008). The treatment of missing data. In W. Outhwaite & S. Turner (Eds.), *Handbook of Social Science Methodology.* London: Sage. Retrieved April 26, 2013, from http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/MissingDataFinal.pdf.
- Lynch, S. (2003). Missing data (Soc 504). Princeton University Sociology 504 Class Notes. Retrieved April 23, 2013, from http://webcache.googleusercontent.com/search?q=cache:HItW60WqNdkJ:www.princeton.edu/~slynch/soc504/missingdata.pdf+Lynch,+S.+(2003).+Missing+data+(Soc+504).+Princeton+University+Sociology+504+Class+Notes.&cd=1&hl=en&ct=clnk&gl=us&client=safari.
- Scheffer, J. (2002). Dealing with missing data. *Res. Lett. Inf. Math. Sci. (2002)*3, 153-160. Retrieved April 23, 2013, from http://equinetrust.org.nz/massey/fms/Colleges/College%20of%20Sciences/IIMS/RLIMS/Volume03/Dealing_with_Missing_Data.pdf.
- Sinharay, S., Stern, H.S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*(4), 317–329.
- Su, Y.S., Gelman, A., Hill, J., & Yajima, M. (n.d.) Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*. Retrieved May 2, 2013, from http://www.jstatsoft.org.
- Van Buuren, S. & Groothuis-Oudshoorn, K. (n.d.) mice: Multivariate imputation by chained equations. *Journal of Statistical Software.* Retrieved May 2, 2013, from http://www.jstatsoft.org.