STAT 408
Applied Regression Analysis

Nan Miles Xi

Department of Mathematics and Statistics
Loyola University Chicago

Fall 2022

Missing Data

Types of Missing Data

- In general, there are three types of missing data

- Missing cases: We fail to observe both x and y for one or multiple observations

  If the reason for this failure is unrelated to what would have been observed, then we simply have a smaller sample and can proceed as normal

- Incomplete values: We fail to observe y in time-sensitive study

  For example, in clinical trials, patient final outcomes are not known

- Missing values: We only observe some predictors or response in one or multiple observations; In this chapter, we will focus on missing values

Missing Mechanism

- <u>Missing Completely at Random</u>

  The probability that a value is missing is the same for all observations. If we simply delete all observations with missing values, we will cause no bias, only lose some information

- <u>Missing at Random</u>

  The probability of missing depends on a known mechanism. For example, in social surveys, certain groups are less likely to provide information than others;

  We can delete these missing observations but include group membership as a predictor in the regression mode

- <u>Missing not at Random</u>

  The probability that a value is missing depends on some unobserved variable. For example, people who have something to hide are less likely to provide information

Simple Deletion

- Let’s see one example of deleting observations with missing values
- Dataset “chredlin” contains information about the FAIR insurance plan (Fair Access to Insurance Requirements)
- A 1970's study on the relationship between insurance redlining in Chicago and racial composition, fire and theft rates, age of housing and income in 47 zip codes

`race`
: racial composition in percent minority

`fire`
: fires per 100 housing units

`theft`
: theft per 1000 population

`age`
: percent of housing units built before 1939

`involact`
: new FAIR plan policies and renewals per 100 housing units

`income`
: median family income in thousands of dollars

	race	fire	theft	age	involact	income
60626	10.0	6.2	29	60.4	0.0	11.744
60640	22.2	9.5	44	76.5	0.1	9.323
60613	19.6	10.5	36	73.5	1.2	9.948
60657	17.3	7.7	37	66.9	0.5	10.656
60614	24.5	8.6	53	81.4	0.7	9.730

Simple Deletion

- We randomly remove entries to create missing values and creates dataset "chmiss"

  1 – mean(complete.cases(chmiss))

- 42.6% observations have missing values

- We summarize the dataset to check the number of missing values per variable

  summary(chmiss)

```
      race             fire            theft             age            involact           income
 Min.   : 1.00   Min.   : 2.00   Min.   :  3.00   Min.   : 2.00   Min.   :0.0000   Min.   : 5.583
 1st Qu.: 3.75   1st Qu.: 5.60   1st Qu.: 22.00   1st Qu.:48.30   1st Qu.:0.0000   1st Qu.: 8.564
 Median :24.50   Median : 9.50   Median : 29.00   Median :64.40   Median :0.5000   Median :10.694
 Mean   :35.61   Mean   :11.42   Mean   : 32.65   Mean   :59.97   Mean   :0.6477   Mean   :10.736
 3rd Qu.:57.65   3rd Qu.:15.10   3rd Qu.: 38.00   3rd Qu.:78.25   3rd Qu.:0.9250   3rd Qu.:12.102
 Max.   :99.70   Max.   :36.20   Max.   :147.00   Max.   :90.10   Max.   :2.2000   Max.   :21.480
 NA's   :4       NA's   :2       NA's   :4        NA's   :5       NA's   :3        NA's   :2
```

Simple Deletion

- We regress involact on all other predictors on chmiss and chredlin

summary(lm(involact~., data = chredlin))

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.608979	0.495260	-1.230	0.225851	
race	0.009133	0.002316	3.944	0.000307	***
fire	0.038817	0.008436	4.602	4e-05	***
theft	-0.010298	0.002853	-3.610	0.000827	***
age	0.008271	0.002782	2.973	0.004914	**
income	0.024500	0.031697	0.773	0.443982	

summary(lm(involact~., data = chmiss))

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.116483	0.605761	-1.843	0.079475	.
race	0.010487	0.003128	3.352	0.003018	**
fire	0.043876	0.010319	4.252	0.000356	***
theft	-0.017220	0.005900	-2.918	0.008215	**
age	0.009377	0.003494	2.684	0.013904	*
income	0.068701	0.042156	1.630	0.118077	

- The lm function automatically drops observations with missing values (sample size 47->27)
- The model fitted on missing data generates large standard error for each parameter due to small sample size; the p-values are also larger

Missing Value Imputation

- Simply removing observations with missing values will waste the information of non-missing entries in the same observation
- A better solution is to "impute" the missing values
- The simplest imputation method is to impute the missing values by the mean of that variable

```
means <- colMeans(chmiss, na.rm = T)
chmiss.impute <- chmiss
for(i in 1:6){
  chmiss.impute[is.na(chmiss.impute[,i]), i] <- means[i]
}
```

- The for loop went through each column, identified all missing values as a True/False vector, and imputed by the corresponding variable means

Missing Value Imputation

- We compare the linear model fitted on the original true data and the mean-imputed data

summary(lm(involact~., data = chredlin))

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.608979	0.495260	-1.230	0.225851	
race	0.009133	0.002316	3.944	0.000307	***
fire	0.038817	0.008436	4.602	4e-05	***
theft	-0.010298	0.002853	-3.610	0.000827	***
age	0.008271	0.002782	2.973	0.004914	**
income	0.024500	0.031697	0.773	0.443982	

summary(lm(involact~., data = chmiss.impute))

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.128243	0.503088	0.255	0.80007	
race	0.006400	0.002627	2.436	0.01927	*
fire	0.027691	0.009266	2.989	0.00472	**
theft	-0.003065	0.002716	-1.128	0.26570	
age	0.006573	0.003091	2.126	0.03954	*
income	-0.029704	0.031333	-0.948	0.34867	

- The mean-imputation improved the standard error of estimated parameters due to increased sample size
- However, it makes theft insignificant, and moves parameter estimation close to zero

Missing Value Imputation

- The mean-imputation is a naive imputation method
  - It essentially use a null linear model to predict missing values

- The inaccurate imputation can be substantial and may not be compensated by the reduction in variance

- A better imputation method is to utilize the relationship among predictors
  - Fit a linear model to predict missing values using non-missing entries

Missing Value Imputation

- Suppose we want to impute the missing values for income

- We fit a linear model with income as response and other predictors on non-missing observations

- We impute missing income by the prediction of this model

```
lm.impute <- lm(income~fire+theft+age+race, data = chmiss)
chmiss[is.na(chmiss$income),]
predict(lm.impute, chmiss[is.na(chmiss$ income),])
```

Missing Value Imputation

- We can compare the imputed race and ground-truth

  chredlin$income[is.na(chmiss$income)]

ground-truth	imputation
8.33 11.26	6.34 10.19

- We can repeat the same process for all predictors with missing values

- Note that we <u>should not</u> use Y's information in the whole imputation process