PHP2550: Practical Data Analysis

Assignment 2: Scientific Questions and Missing Data

Antonella Basso

October 14, 2022

1. Scientific Questions

Read the paper "Association of Highly Restrictive State Abortion Policies With Abortion Rates, 2000-2014" on Canvas (called `abortion_policies.pdf`). Then, respond to the following prompts.

a. Summarize the paper in 2-3 paragraphs using your own words.

b. Write out the model that adjusts for distance in Table 3 and explain how you derived it. You may need to do some research into the methods used or make some assumptions based on what the authors write.

c. How does this model relate to the overall research question? Evaluate this approach to answering the question in 2-3 paragraphs. Be sure to mention the limitations of the approach and assumptions made.

d. If you had access to the same data, how might you extend this analysis to address the points in the previous question (1-2 paragraphs).

Solution

a. This study aims to combine county-level abortion rate data from several states where restrictive abortion laws may have been implemented to evaluate the association between statewide highly restrictive legislative climates and changes in abortion rates in the years between 2000 and 2014. Specifically, the study identified states that imposed abortion restrictions of 4 different types, categorizing a state as highly restrictive if it implemented at least 3 of them, and defined a county-level abortion rate as the total number of abortions obtained in a given year divided by the total female population (in a given county). Additionally, in light of the fact that many of these policies have resulted in clinics being shut down, thereby posing additional travel-related barriers to individuals seeking abortions, researchers sought to investigate whether the association between legislative climate and abortion rate might be mediated by distance to the nearest facility providing abortion care—estimated by the straight-line distance between a county's population centroid and that of the closest county having a high-volume facility. This study was primarily motivated by an effort to bring attention to the potential harms placed on individuals who need abortions by quantifying the degree to which restrictive abortion laws inhibit access to abortion care.

   Given the nature of the data and research endeavors, longitudinal data analysis was carried out through a series of linear regression models within a propensity score–weighted difference-in-difference framework. Due to the difficulty of assessing individual-level changes in abortion care access however, the analysis was generalized to the county-level with propensity score weights derived from county-specific demographic data, as well as state and year fixed effects to hold potentially confounding state-level differences and temporal trends constant across groups. In contradiction to prior state-specific data that suggest an association between restrictive laws and clinic closures, which increase travel distances, the study did not find a statistically significant relationship between the two. Notably however, results showed that a highly restrictive policy climate is in fact associated with a lower abortion rate. Specifically, it was estimated that highly restrictive legislative climates were associated with an abortion rate decrease of

0.48 abortions per 1000 women, when compared to a less restrictive climate. Moreover, in models that only considered states that became highly restrictive during the study period and in those that adjusted for distance to a facility, the highly restrictive legislative climate was associated with an abortion rate decrease of 0.45 and 0.44 abortions per 1000 women, respectively, with the latter estimate decreasing by an additional 0.02 abortions per 1000 women for each unit (miles) increase in distance to a facility. Given that the difference-in-difference model compares groups of counties that only vary (on average) with regards to the introduction of the investigated policies—assessing changes in abortion rates over time, with all other demographic factors constant—this result suggests that restrictive climates themselves may act as barriers to abortion care access.

b. Given the information provided in the paper and the nature of difference-in-difference models with two-way fixed effects, we follow the logic provided in Chapter 18 of *The Effect: An Introduction to Research Design and Causality*, after having referred to similar sources (e.g., *Difference-in-Differences*, *Designing Difference in Difference Studies: Best Practices for Public Health Policy Research*), to modify the standard model such that it handles multiple time periods of interest as depicted in the figure provided in the study. Particularly, as suggested in the paper and further explained in the external source linked above, this may be done by including dummy variables in the model; one for each time period of interest in relation to a baseline year, which in our case, is the year preceding an introduction of a highly restrictive climate. In accordance with the two-way fixed effects model structure implied in the paper, we identify the unit and time fixed effects as "state" and "year", respectively. Under the assumption that a "unit" corresponds to one of the two identified groups of counties, the time-invariant fixed effect of "state" removes potential confounding due to state-based differences within legislative climate groups. Similarly, the group-invariant fixed effect of "time" holds potentially confounding temporal trends constant. An attempt to recreate this model, adjusting for distance, is given below.

   Let,

   - Y be the continuous outcome of interest—change in abortion rate in number of abortion cases per 1000 women between the current year and the baseline year.
   - α_s and α_t be the state and year fixed effects, respectively.
   - $T_{b\pm k}$ for $k \in [1, 6]$, be the set of binary indicator (dummy) variables for time periods k (in number of years) before (-) or after (+) the baseline year b—identified to be the year before the introduction of a highly restrictive climate.
   - X be the binary indicator for highly restrictive legislative climate.
   - D be the continuous variable for distance (in miles) to the nearest abortion facility.

$$Y = \alpha_s + \alpha_t + \beta_{b-6}T_{b-6}X + ... + \beta_{b+6}T_{b+6}X + \tau D + \epsilon$$

c. This model addresses the question of whether there exists an association between a highly restrictive legislative climate and abortion rates to the extent of statistical significance in the coefficients given the available data. Specifically, as stated in the paper, each coefficient in the model verifies whether a change in abortion rate in the years around the adoption of a highly restrictive climate are in fact due to a shift in policy. This is due to the fact that the difference-in-difference model with two-way fixed effects accounts for potentially confounding state-based differences within the two groups, as well as conflicting temporal trends in the data, thereby isolating the effect of legislative climate on abortion rates.

   Nonetheless, the strong assumptions made by this model pose a few problems for the inferential claims made in turn. The most significant among these is the *parallel trends assumption*, which holds that the two groups being compared would have shown the same trends in abortion rates had the legislative climate remained minimally or moderately restrictive in the "post-adoption" period for both groups. This suggests moreover, that the effect of legislative climate on abortion rates is constant (i.e., homogeneous), such that the adoption of highly restrictive policies is the only factor that differentiates the two groups. The problem with assuming parallel trends when it comes to internal validity and inference is that the the effect of "treatment", which in our case corresponds to a highly restrictive

legislative climate, is likely heterogeneous due to the sheer variability and randomness regarding county or "unit" characteristics that, even in the presence of additional data, may not be entirely ruled out by similar "pre-adoption" period trends or captured by state-based and temporal differences.

Researchers address the limitations related to both including a diverse set of states in their analyses, and ignoring existing differences between those included and "those that did not provide usable data", by admitting to the possibility of having obtained different results in the presence of additional state data. Although, due to the small share of observed differences in variables, they claim that this is unlikely to be a considerable source of bias. This claim however, is made on the basis of observed data and does not account for potential bias due to unmeasured confounding, making results arguably nongeneralizable to all states. Additional limitations addressed by the investigators regarding the models' ability to answer the research question(s) include: the fact that abortion rates reflect the number of abortion cases in a given period as a proportion of an entire female population, as opposed to the share of the population that is fertile; the fact that the data only covers abortion rates and policies relevant to years before 2014, not accounting for numerous restrictive laws that have been implemented in more recent years; and the inability to produce statistically significant results regarding the association between restrictive policy climate and distance to a facility.

d. With access to the same data, assuming no access to external sources, it would be difficult to address many of the limitations identified by the researchers; particularly the last three, which require more recent and more specific data on different U.S. populations (e.g., fertile females in various U.S. counties, county-level demographics across a wider set of states, etc.). However, with knowledge about the average shares of female populations that are fertile for example, it may be possible to generate more accurate estimates of abortion rates from the available data. Investigators in the study have already taken several steps to minimize potential bias and increase their chances of satisfying the parallel trends assumption. Specifically, having implemented propensity-score weighting to make counties more comparable; including fixed effects in the model to minimize potential confounding; allowing for some deviation through confidence intervals; and using cluster-robust standard errors to account for potential autocorrelation.

   Given that the parallel trends assumption is strong, untestable, and not easily satisfiable however, an additional way of limiting bias would be to extend the analysis to include a step for analyzing groups. Specifically, instead of making groups comparable on account of all covariates, we may choose to condition on different subsets/combinations of covariates to identify differences in average effects of legislative climate and hence potential sources of bias related to the way the groups were originally formed ("treatment assignment"). Generating different groups for comparison in this way may help explain some of these "less obvious", but nonetheless relevant, differences that can't otherwise be accounted for solely through additive fixed effects. Additionally, given the fact that confounding can't be ruled out, in tandem with the significant lack of information on all states, it seems crucial to emphasize that results are statistically significant for the states and counties included in the data, as the degree to which they generalize to the entire nation remains unknown by reason of insufficient data.

2. Missing Data

The pain clinic data we saw in class has follow-up information on a subset of patients. If we were interested in analyzing the change in pain over time, it would be important to think about the missing data due to loss to follow-up.

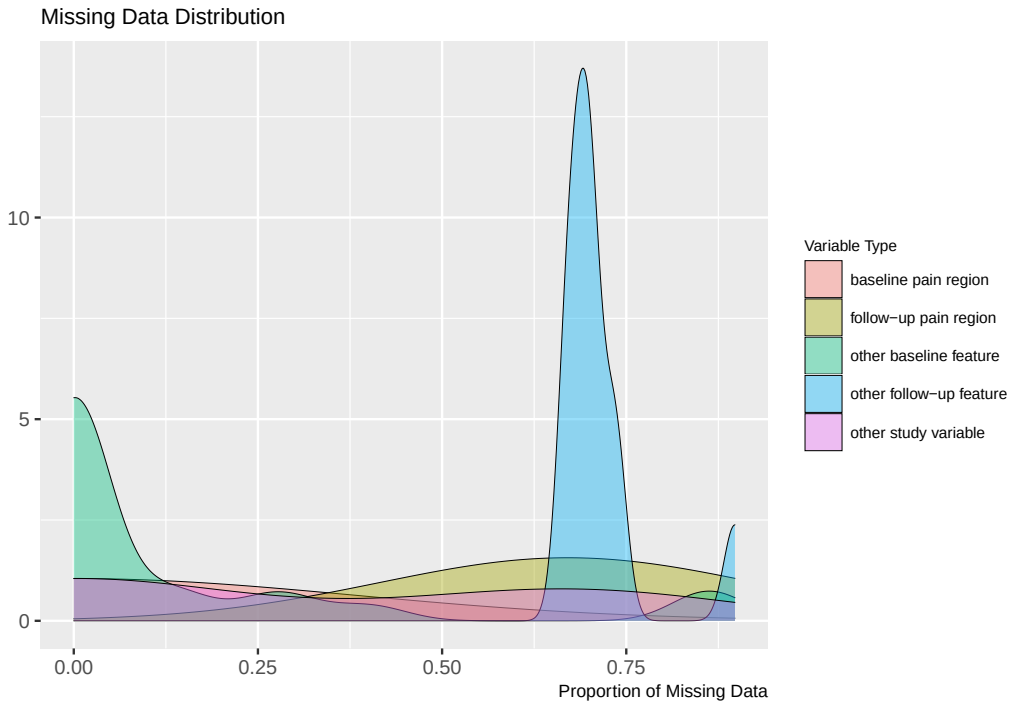
a. First, describe the patterns of missing data observed in the data set overall.

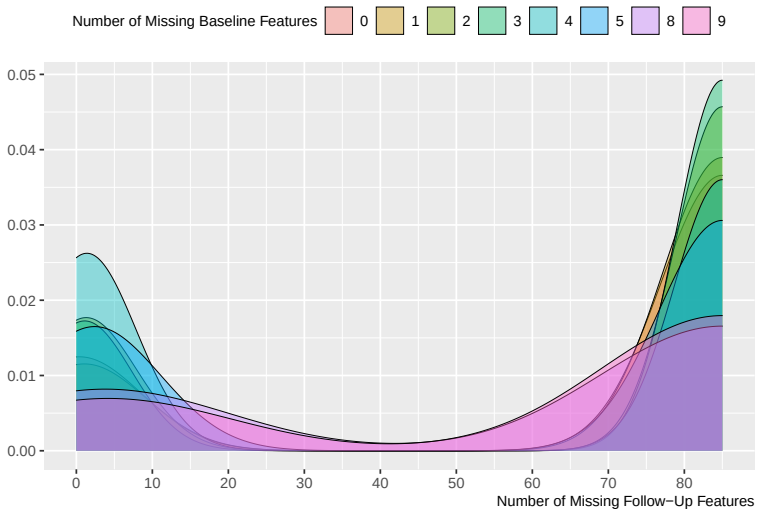
b. Compare the baseline characteristics between those with and without follow-up information. Comment on your results and discuss whether you think the data is MCAR, MAR, or MNAR.

c. Suppose we wanted to fit a model to assess risk of having worse pain at follow-up. To address the missing information, we are considering either using inverse probability weighting or multiple imputation. Explain briefly the steps we would take in each case and discuss the benefits and drawbacks to each approach. Which would you apply in this case?

Solution

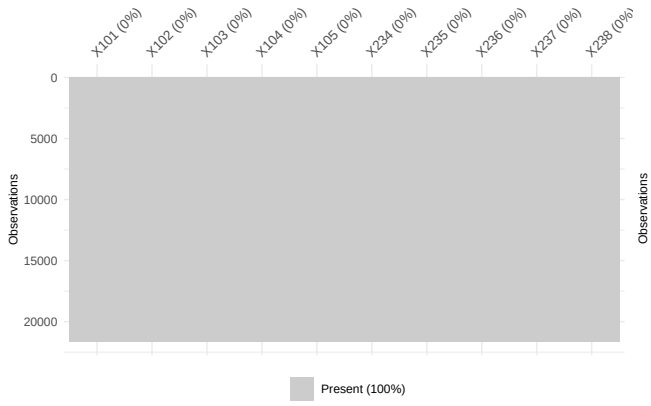
a. Without taking into account patient study indicators (`PATIENT_NUM`), log transformations of the baseline and follow-up `BODYREGIONSUM` features, or the single missing observation across all variables in the data—likely due to a file conversion error—we find that, in total, approximately 33.23% of the dataset is missing. Moreover, given the considerable number of features reflected in the pain clinic data, we group variables by "type" to produce a more generalized and meaningful analysis. Specifically, the first plot below gives a comparison of the distributions of missing data (in proportions) for each of the five chosen categories of variables. Conversely, the subsequent three plots provide the densities of missing baseline and follow-up features across observations, broken up by the number of missing baseline and , and other study features, respectively. Noticeably, the middle graph shows us that the study variables we define as "other" are likely tied to the follow-up features given their apparent proportionality. Note that a log transformation was used for visualization purposes in this particular case given the wide range of possible missing follow-up values due to pain region indicators. While both of these approaches were effective in showing the primary sources of missingness and their relative trends in the data, the six graphs that follow provide an even more robust visualization of these patterns.

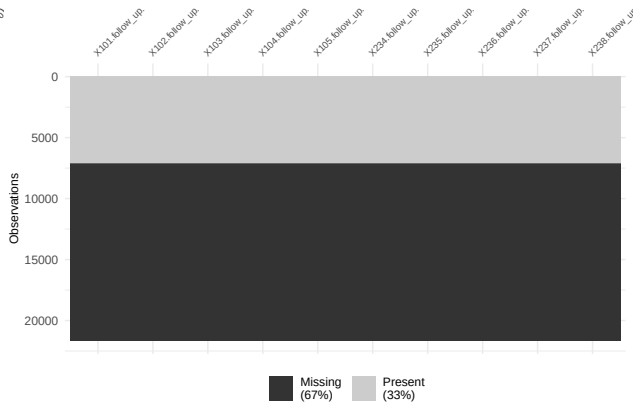



Number of Missing Baseline Features
0
1
2
3
4
5
8
9
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60
70
80
Number of Missing Follow−Up Features

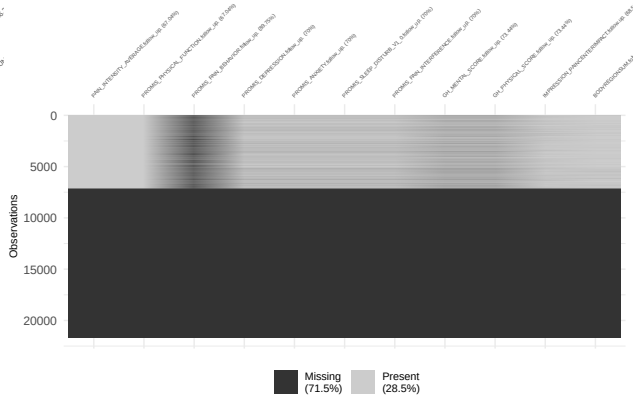

Number of Other Missing Study Features
0
1
3
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
Number of Missing Follow−Up Features
Log Transformation

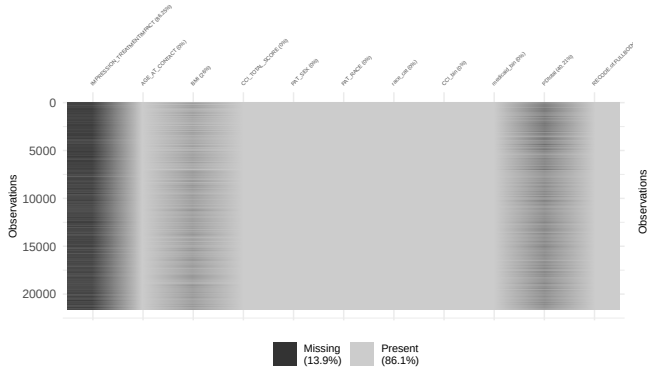

Number of Other Missing Study Features
0
1
3
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
6
7
8
9
Number of Missing Baseline Features

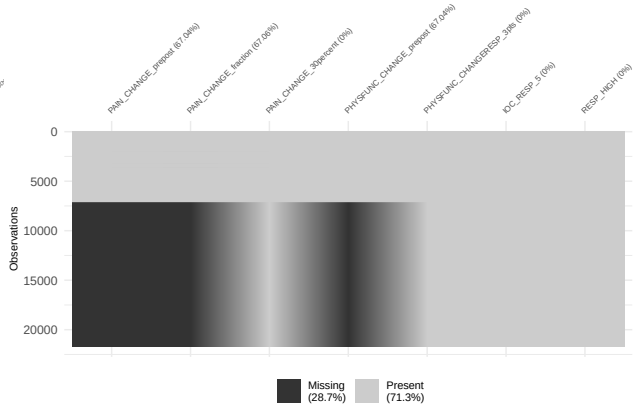

X101 (0%)
X102 (0%)
X103 (0%)
X104 (0%)
X105 (0%)
X234 (0%)
X235 (0%)
X236 (0%)
X237 (0%)
X238 (0%)
Observations
0
5000
10000
15000
20000
Present (100%)


Observations
0
5000
10000
15000
20000
Missing (67%)
Present (33%)


PAIN_INTENSITY_AVERAGE (0%)
PROMIS_PHYSICAL_FUNCTION (0%)
PROMIS_PAIN_BEHAVIOR (59.41%)
PROMIS_DEPRESSION (6.4%)
PROMIS_ANXIETY (5.4%)
PROMIS_SLEEP_DISTURB_V1_0 (0.4%)
PROMIS_PAIN_INTERFERENCE (0.7%)
GH_MENTAL_SCORE (13.8%)
GH_PHYSICAL_SCORE (13.6%)
Observations
0
5000
10000
15000
20000
Missing (13.2%)
Present (86.8%)


Observations
0
5000
10000
15000
20000
Missing (71.5%)
Present (28.5%)


Observations
0
5000
10000
15000
20000
Missing (13.9%)
Present (86.1%)


PAIN_CHANGE_preposit (67.04%)
PAIN_CHANGE_fistclinn (67.06%)
PAIN_CHANGE_3bpercent (0%)
PHYSFUNC_CHANGE_preposit (67.04%)
PHYSFUNC_CHANGEBPESP_3186 (0%)
IDC_RESP_5 (0%)
RESP_HIGH (0%)
Observations
0
5000
10000
15000
20000
Missing (28.7%)
Present (71.3%)

`The six missing data pattern plots above are arranged for comparison&mdash;the first two plots giving ba`

b. In terms of the identifying the type of missingness reflected, it could be argued that, given the clear split between missing and non-missing data in the follow-up period shown in the right-hand column of pattern plots above, data are MNAR, at least in this regard, as this missingness appears more deliberate and systematic. Generally however, it is safe to conclude that data are most likely MAR. Particularly, since there is an association between missingness in follow-up features and baseline features we can assume a potential underlying reason behind participants' tendencies to drop out of the study or fail to report certain follow-up information.

   To investigate this baseline and follow-up relationship and see whether it missingness in the follow-up data can be attributed to any observed variable(s), we fit a linear regression model on a new variable; `followup_missing` giving the number of missing follow-up features for each observation; which, after backward selection, showed statistical significance for the following covariates:

   - `baseline_missing`: continuous variable for the number of missing baseline features
   - `other_missing`: continuous variable for the number of missing other study features
   - `PROMIS_PHYSICAL_FUNCTION`[1]: continuous variable for the PROMIS physical function T-score
   - `IOC_RESP_5`: binary indicator for whether `IMPRESSION_PAINCENTERIMPACT` "Very Much Improved" at follow-up

   This suggests that lack of responses at follow-up may be related individuals' levels of physical functionings and whether or not they saw significant improvements in pain impacts (on their everyday lives). Notably, since `IOC_RESP_5` had no missing values and a statistically significant coefficient that was negative and smaller for significant improvement compared to minor or no improvement with regards to pain impact, it is possible that individuals are more likely to report follow-up information (and have less missing follow-up features) when they've seen significant improvements in the impact pain has had on their lives. This indicates and further supports the claim that data are MAR[2].

c. If our goal is to model the risk of displaying worse pain in the follow-up period from the available data and simultaneously limit bias in our estimates, it is crucial to implement a more substantive approach to addressing the amount of missing data since it is too large to justifiably restrict our analysis to complete cases. In the scenario that we choose to implement inverse probability weighting (IPW), our focus would be on giving more weight to individuals in the study who appear more likely to have missing observations, but in fact show the contrary. In other words, we want to weight individuals according to the inverse probability of not having any missing values, that is, being a "complete case", such that those who are twice as likely to have incomplete data, but do not, count for two observations. In this way, we appeal to what our data would have been like, had it not been subject to missingness. Specifically, we many obtain such IP weights via logistic regression—modeling the probability of being a complete case, given all other features in the data. In turn, we may use them in tandem with our data to estimate one's risk of increased pain while minimizing potential bias. In the case of multiple imputation (MI), our focus would be on imputing the missing values for each variable by way of sampling from a joint probability distribution (data likelihood) n times. This would produce n different realizations of our dataset from which to obtain n risk estimates of increased pain that may be averaged out. Both of these methods have their own benefits and limitations to consider. For our purposes however, it would be best to adopt the latter approach of multiple imputation—namely, given the nature of the data (e.g., no one observation is missing all relevant features, variables are relatively easy to model, etc.) and the prevalence of missingness we observe, this method would yield the most accurate results.

[1] Significant at the 0.1 level.

[2] It should be noted that while `IOC_RESP_5` had no missing values, the related `IMPRESSION_PAINCENTERIMPACT` follow-up feature had a considerable amount of data missing. The reason for this discrepancy is not known, but should be further investigated as it poses a problem for our regression-based assumption of data being MAR.

Code Appendix

```
## Libraries

library(tidyverse)
library(visdat)
library(naniar)
```

```
## Data

pain <- read.csv("/Users/antonellabasso/Desktop/PHP2550/Data/pain.csv")
head(pain)
```

```
## Initial Data Exploration

# general
dim(pain)
sum(is.na(pain$PATIENT_NUM)) # 1 missing patient ID
sum(length(unique(pain$PATIENT_NUM))) # 21659-1=21658 unique patient IDs

# 1 row/observation completely missing
pain[is.na(pain$PATIENT_NUM), ]
# error due to (possibly) file conversion
# seen previously with the same dataset

# removing missing observation (not relevant for our analysis)
pain <- pain[!is.na(pain$PATIENT_NUM), ]
```

```
## Missing Data Exploration

# excluding patient ID, and log of body region sums

# percentage of total missing data (33.23%)
sum(is.na(pain[,-c(1, 190:191)]))/prod(dim(pain[,-c(1, 190:191)]))

# BY VARIABLES

# proportions of observed data for each variable
apply(pain[,-c(1, 190:191)], 2, function(x){return(sum(!is.na(x))/length(x))})

# proportions of missing data for each variable
# grouping by variable type
# 5 variable categories: pain region/other by baseline/follow-up & other variables
missing_cols <- data.frame(variable=as.vector(names(pain[,-c(1, 190:191)])),
                           type=c(rep("baseline pain region", 74),
                                  rep("other baseline feature", 14),
                                  rep("follow-up pain region", 74),
                                  rep("other follow-up feature", 10),
                                  "other baseline feature",
                                  "other follow-up feature",
                                  rep("other baseline feature", 7),
                                  rep("other study variable", 7)),
                           missing=as.vector(apply(pain[,-c(1, 190:191)], 2,
                                                   function(x){
                                                     return(sum(is.na(x))/length(x))})))
```

```r
# density plot grouping by variable type
md_plot1 <- ggplot(missing_cols, aes(x=missing, fill=type)) +
  geom_density(alpha=0.4, lwd=0.2) +
  labs(title="Missing Data Distribution",
       x="Proportion of Missing Data",
       fill="Variable Type") +
  theme(plot.title=element_text(size=10),
        axis.title.x=element_text(size=8, hjust=1),
        axis.title.y=element_blank(),
        legend.title=element_text(size=7),
        legend.text=element_text(size=7))


# BY OBSERVATIONS

# introducing new missingness variables
pain_miss <- pain[,-c(1, 190:191)]

# each observation's number of missing baseline, follow-up, and other features
pain_miss$baseline_missing <- as.vector(rowSums(is.na(
  pain_miss[, missing_cols[grep("baseline",
                                missing_cols$type), 1]])))
pain_miss$followup_missing <- as.vector(rowSums(is.na(
  pain_miss[, missing_cols[grep("follow-up",
                                missing_cols$type), 1]])))
pain_miss$other_missing <- as.vector(rowSums(is.na(
  pain_miss[,missing_cols[grep("study",
                               missing_cols$type), 1]])))

# each observation's total missing features
pain_miss$all_missing <- as.vector(rowSums(is.na(pain_miss)))

# indicator for at least 1 missing follow-up feature
pain_miss$followup_missing_bin <- as.vector(as.numeric(pain_miss$followup_missing!=0))

# distribution of missing follow-up features
# by number of missing baseline features across observations
md_plot2a <- ggplot(pain_miss, aes(x=followup_missing,
                                   fill=as.factor(baseline_missing))) +
  geom_density(alpha=0.4, lwd=0.2) +
  labs(x="Number of Missing Follow-Up Features",
       fill="Number of Missing Baseline Features") +
  theme(axis.title.x=element_text(size=9, hjust=1),
        axis.title.y=element_blank(),
        legend.title=element_text(size=9),
        legend.text=element_text(size=9),
        legend.position="top") +
  guides(fill=guide_legend(nrow=1, byrow=TRUE)) +
  scale_x_continuous(breaks=seq(0, 90, 10))

# distribution of missing follow-up features
# by number of other missing features across observations
# using log transformation for better visualization
```

```r
md_plot2b <- ggplot(pain_miss, aes(x=log(followup_missing),
                                   fill=as.factor(other_missing))) +
  geom_density(alpha=0.4, lwd=0.2) +
  labs(x="Number of Missing Follow-Up Features",
       fill="Number of Other Missing Study Features",
       caption="Log Transformation") +
  theme(axis.title.x=element_text(size=9, hjust=1),
        axis.title.y=element_blank(),
        legend.title=element_text(size=9),
        legend.text=element_text(size=9),
        legend.position="top") +
  guides(fill=guide_legend(nrow=1, byrow=TRUE))

# distribution of missing baseline features
# by number of other missing features across observations
md_plot2c <- ggplot(pain_miss, aes(x=baseline_missing,
                                   fill=as.factor(other_missing))) +
  geom_density(alpha=0.4, lwd=0.2) +
  labs(x="Number of Missing Baseline Features",
       fill="Number of Other Missing Study Features") +
  theme(axis.title.x=element_text(size=9, hjust=1),
        axis.title.y=element_blank(),
        legend.title=element_text(size=9),
        legend.text=element_text(size=9),
        legend.position="top") +
  guides(fill=guide_legend(nrow=1, byrow=TRUE)) +
  scale_x_continuous(breaks=c(0:10))
```

```r
## Missing Data Pattern Visualization

# broken up by variable type (5)

# baseline pain regions
# all the same (no missing values) -> showing first and last 5
vis_miss_pr_bl <- vis_miss(
  pain[, missing_cols[grep("baseline pain", missing_cols$type), 1][c(1:5, 70:74)]],
  warn_large_data=F)

# follow-up pain regions
# all the same -> showing first and last 5
vis_miss_pr_fu <- vis_miss(
  pain[, missing_cols[grep("follow-up pain", missing_cols$type), 1][c(1:5, 70:74)]],
  warn_large_data=F, show_perc_col=F)

# other baseline features (part 1) - for comparison with other follow-up features
vis_miss_oth_bl <- vis_miss(
  pain[, missing_cols[grep("other baseline", missing_cols$type), 1][c(1:10, 15)]],
  warn_large_data=F)

# other baseline features (part 2)
vis_miss_oth_bl2 <- vis_miss(
  pain[, missing_cols[grep("other baseline", missing_cols$type), 1][c(11:14, 16:22)]],
  warn_large_data=F)
```

```r
# other follow-up features
vis_miss_oth_fu <- vis_miss(
  pain[, missing_cols[grep("other follow-up", missing_cols$type), 1]],
  warn_large_data=F)

# all other study variables
vis_miss_all_oth <- vis_miss(
  pain[, missing_cols[grep("other study", missing_cols$type), 1]],
  warn_large_data=F)
```

```r
# modeling missing follow-up to find correlation
summary(lm(followup_missing ~ baseline_missing + other_missing +
             as.numeric(PROMIS_PHYSICAL_FUNCTION)+
             IOC_RESP_5, data=pain_miss))
```