25 Missing Data

- Does data go missing?
- How does that happen?
- Is missing data a problem?

We use NA for missing data in R, . in SAS.
If any predictor or the response is missing, a model fitting routine will handle it by:

25.1 Types of Missing Data

1. MCAR: completely at random.
   Someone generates iid Bernoulli(p) for each data row, omits those were we get a 1.
   Not a problem, results are unbiased.
2. MAR: at random.
   As above, but the probability of going missing, p, depends on observed covariates (reace, earnings, etc.) and can be modeled.
   Results can be adjusted to be unbiased via modeling on the covariates.
3. Not at random 1:
   As in (2), but p depends on unmeasured lurking variables.
   Results are biased, perhaps we can recover info if we can model the lurking variable.
4. Not at random 2 – No Hope:
   Missing "earnings" depends on actual earnings. Includes censoring.

25.2 Toss it out

- *Complete Cases* as in typical R, BUGs, or SAS routines leaves out rows with any missing values. Problems:
  - Omitted rows may differ in some important way from those fully observed.
  - Reduced sample size – less power.
- *Available Cases*: Say we have 2 or more responses to analyze, and missing pattern is different. We are then using a different subset of the data for each response.
  May lose the "ignorability" assumption.
- *Non-response weighting* If only one predictor has missing values, we could model missingness using the other predictors and estimate a p_i probability of nonresponse for each row, then weight each row by the inverse probability.

25.3 Simple All-Data Approaches

Danger: Single Imputation misses the variance of observations.

- *Fill in the Mean*, $\overline{x}_k$, using mean of this column. SE's are biased toward 0.
- *Carry Last Forward*, as in a time series. Not appropriate when looking for a change.
- *Fill in Predicted*, $\hat{x}_k$, where the prediction comes from a regression on the other predictors. SE's are still biased toward 0.
- *Add Indicator variable for Missingness.* Shifts all missing values the same way.
- *Logic?* If they didn't work, income must be 0.

25.4 Random Imputation - 1 variable

Repeat analysis several times with new datasets.
R packages: mi, mice, mitools, VIM (visualize patterns), tabplot, Amelia, mix, pan, norm, cat, MLmix, minnMDA, and many more.
See
CRAN Task View: Official Statistics & Survey Methodology

```
random.imp <- function(a){
    ## fill in with one of the available values
    missing <- is.na(a)
    a[missing] <- sample(a[!missing], sum(missing),
                         replace = TRUE)
    a
 }
earnings.imp1 <- random.imp(earnings)
```


Regression Approach

Set ceiling on earnings of $100K, as we only care about quantiles.

```
earnings.top <- pmin(earnings, 100)
   ## their 'topcode' function is not needed

lm.Imp1 <- lm( earnings ~ ., data = SIS,
              subset = earnings > 0)
   ## Fit to people who had earnings, but predict to all
pred.1 <- predict(lm.Imp1, newdata = SIS)

earnings.imp1 <- with(SIS,
        ifelse( is.na(earnings), pred.1, earninings))
```

Earnings are skewed, so a better approach is to use $\sqrt{\text{earnings}}$, but still $R^2 = 0.44$ is low.

Now add random reps

```
earnMiss <- is.na(SIS$earn)
pred4 <- rnorm( sum(earnMiss),
               predict(lm.Imp4, newdata = SIS[!earnMiss,]
                       sigma.hat(lm.Imp4)))
```

Still not too similar to observed earnings – see figures.
Two stage – like tobit regression: model the zero earnings, then given some earnings, model the numeric value.
Repeat Multiple times?

Cold & Hot Deck

Match each unit with missing earnings to a case with similar attributes which has an earnings.
Cold if separate data source, Hot if current data.

More than 1 predictor has missing problems

- Multivariate approach: Tends to be too packaged
- Iterative imputation: Rotate through columns of X. Models can vary in inputs, but need consistency.
  No joint likelihood?

25.6 Model-Based Imputation

BUGS wants all predictors passed in as data.
Would require a multivariate approach.