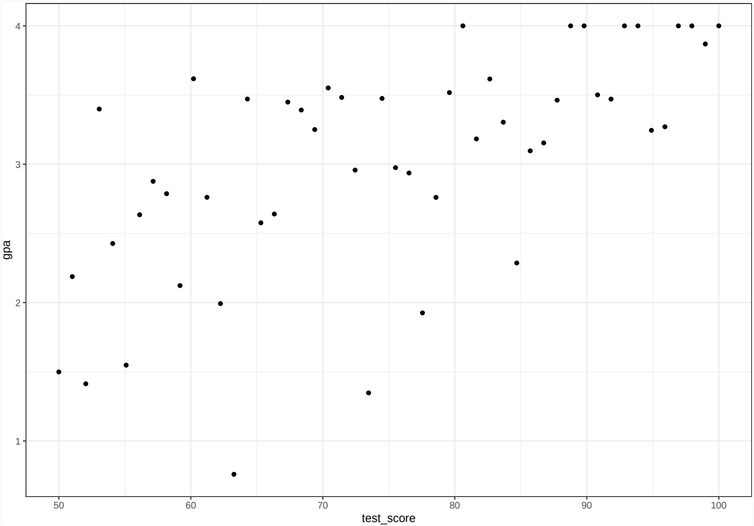
Understanding and addressing missing data

Frank Edwards

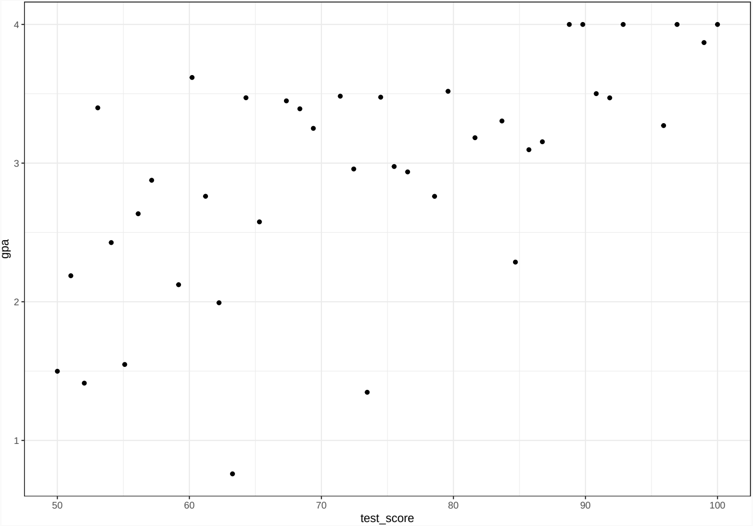
Why should we care?

- Most statistical software will conduct “complete-case analysis” by default
- This may result in throwing away a lot of perfectly good information!
- Listwise deletion understates uncertainty, may result in bias

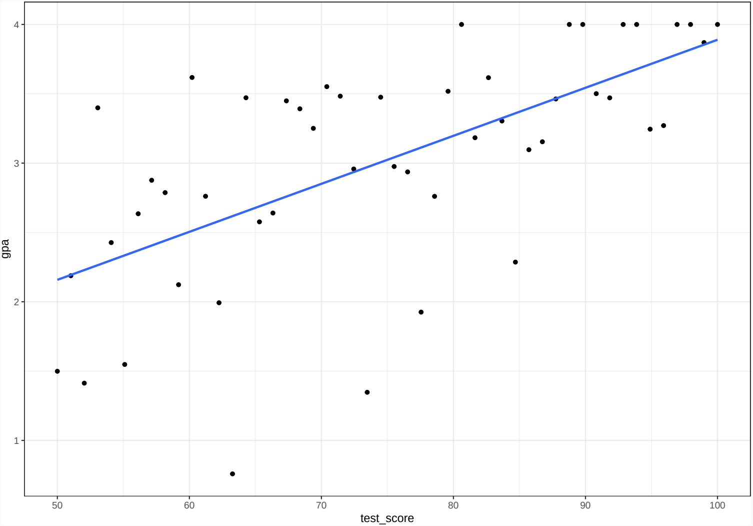
A hypothetical with missing data: predicting student GPA from a math test



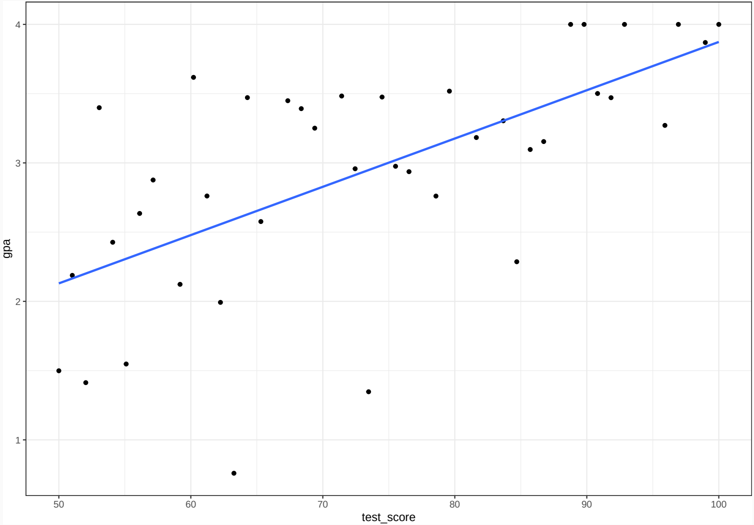
Missing observations



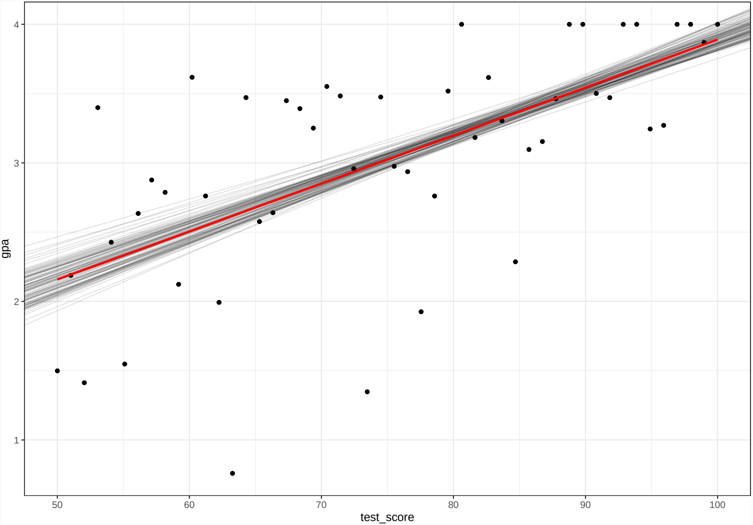
Best fit line under complete data



Best fit line under missing data



100 hypothetical lines with different sets of 10 cases missing completely at random



Three general causes of missing data: MCAR

- **Missing completely at random (MCAR)**: The probability of a value being missing is the same for all observations in the data.
- Potential MCAR mechanisms: survey non-response due to exogenous factors: e.g. lost mail, bad weather, software errors.
- Can be verified by comparing group means of missing and non-missing data on observables: for large N, values are equal

MCAR results in unbiased Beta estimates, but increases standard errors and uncertainty

```r
### true values
tidy(lm(gpa ~ test_score, data = sim))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic     p.value
##   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)   0.427    0.462       0.924 0.360
## 2 test_score    0.0346   0.00604     5.73  0.000000642
```

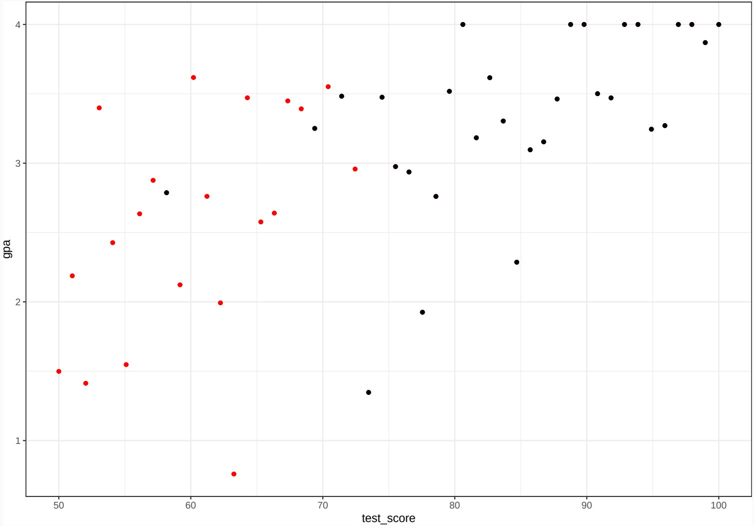
```r
### with missing data
tidy(lm(gpa ~ test_score, data = sim_mcar))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   0.386    0.516       0.749 0.458
## 2 test_score    0.0349   0.00688     5.07  0.0000107
```

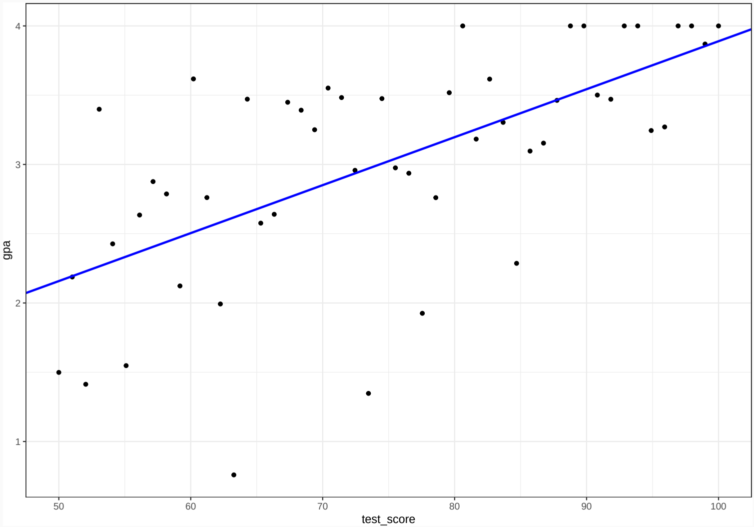
Three general causes of missing data: Missing at random

- **Missing at random (MAR)**: The probability of a value being missing is *not* completely at random (I know)
- The probability of a value being missing is determined by other variables in the data
- After controlling for other values in the data, missingess is random
- Potential MAR mechanisms: people with high income less likely to report total wealth; places with high poverty less likely to submit voluntary administrative data; news reports unlikely to identify other characteristics of child victims of crime / violence

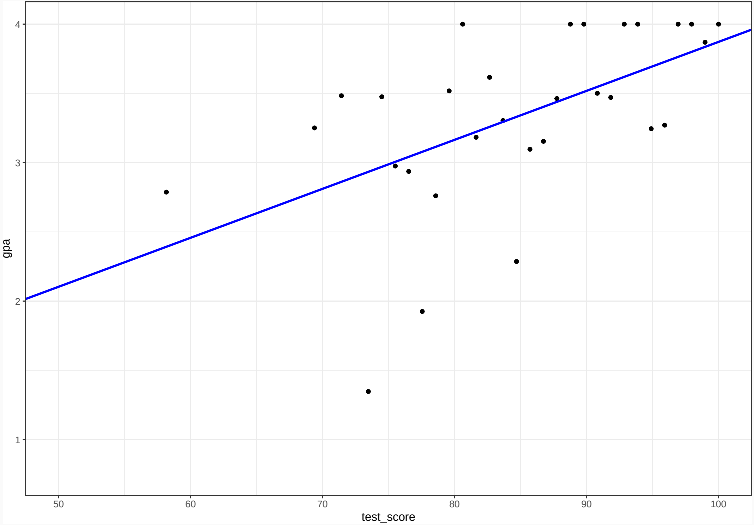
What if students with low GPAs were more likely to miss school on test day?



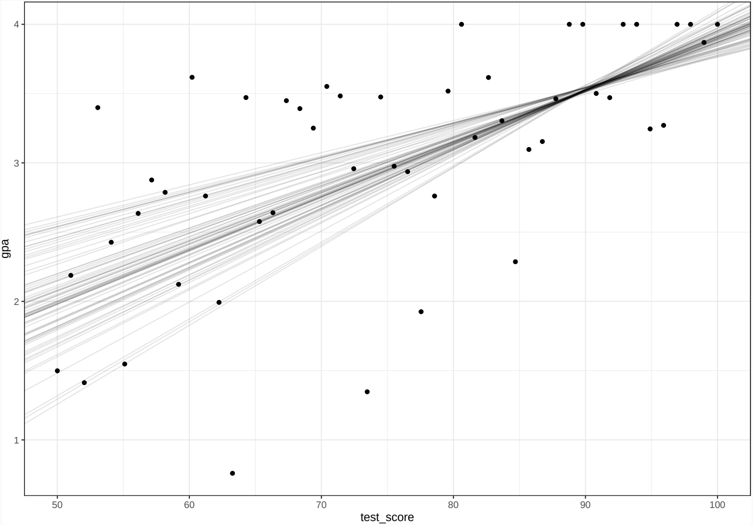
Best fit line under complete data



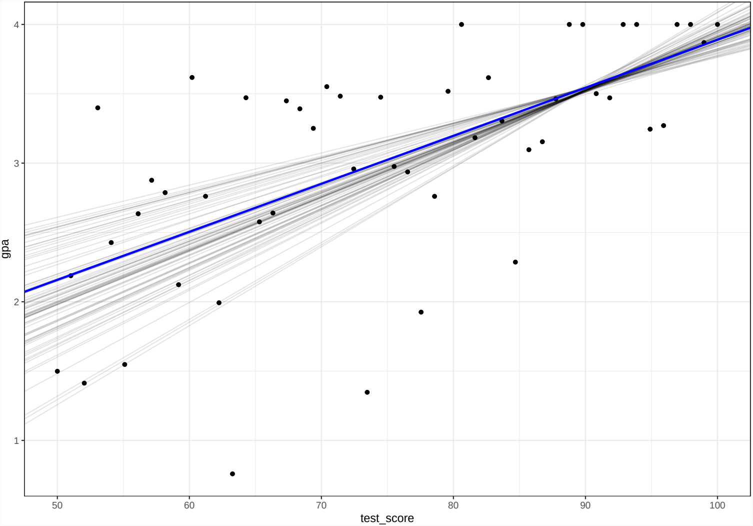
Best fit line under missing data



100 hypothetical lines with different sets of 10 cases missing at random, conditional on GPA



And with the true regression line



The probability of data being missing is conditional on GPA. If we ignore the missing data, then we will systematically understate the relationship between test score and GPA.

$$P(missing) \neq P(missing|GPA)$$

Results of regression models (spot the bias!)

```
### true values
tidy(lm(gpa ~ test_score, data = sim))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic     p.value
##   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)   0.427    0.462       0.924 0.360
## 2 test_score    0.0346   0.00604     5.73  0.000000642
```

```
### average parameter estimates for 100 simulations with missing data
tidy(lm(gpa ~ test_score, data = sim_mar))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   0.335     0.866      0.387 0.701
## 2 test_score    0.0354    0.0102     3.48  0.00164
```

Three general causes of missing data: non-random missingness

- **Missing not at random (MNAR)**: The probability of a value being missing depends on either *A)* some unobserved variable or *B)* the value itself (censorship)
- Examples: police departments with high crime may opt-out of reporting their data to the federal government; police departments with high levels of use-of-force opt-out of reporting to federal arrest-related-deaths programs; people who do not vaccinate their children opt-out of answering a survey question about vaccination
- We cannot distinguish between MAR and MNAR: you must think carefully about missing data mechanisms

Mechanisms of missing data

- Missing completely at random: missingness determined by a coin flip
- Missing (conditionally) at random: missingness on variable x determined by some other variable y
- Missing not at random: missingness on variable x depends only on variable x (or some unobserved variable z)

So what can we do?

Basic approaches to missing data

- Listwise deletion (complete case analysis)
  - Appropriate for data with very few missing observations, and when missingness is completely at random
- Using alternative information on known or stable variables (e.g. imputing age based on information from prior survey wave)
- Imputation of missing values (deterministic, stochastic)

Basic approaches to missing data, deterministic

- Missing value is generated by a fixed (non-random) procedure
- Many examples: linear interpolation, last observed, regression imputation
- This is generally a bad idea.

Basic approaches to missing data, stochastic

- Missing value is generated through random sampling
- Many approaches, but multiple imputation has become widely used

Multiple imputation

- Iterative modeling of all missing outcomes/predictors in model
- Produces series of fake datasets where missing values are predicted with from regression model (with error)
- Allows you to estimate uncertainty generated by missing data
- Does not recover “true” values
- Under missing at random assumption, generates unbiased parameter and variance estimates

What muliple imputation does:

- Has two effects on model uncertainty
  - Increases your N because we aren't deleting data (pushes standard errors downward)
  - Adds in appropriate noise due to uncertainty around where missing values are (pushes standard errors upward)
- If missingess is associated with observables and we have enough data, MI can correct bias in parameter estimates

My preferred approach

- Understand your data!
    - Read the documentation
    - Do plenty of exploratory data analysis (cross tabs, data visuals, descriptives, look at the raw data)
    - Develop an understanding of the mechanisms of missing data in each dataset you use
    - Test your ideas for mechanisms of missing data when feasible

My preferred approach

- Use available information
    - Borrow data from other observations when possible
    - Some variables are time-stable (age) and can be borrowed from prior observations - but remember cautions against deterministic imputation and inducing bias

My preferred approach

- If MAR is a reasonable assumption (it often is), conduct multiple imputation
    - Because MAR is conditional on observables, including many variables in imputation models is often a good idea
- Apply preferred final model / analysis over each imputed dataset, combine with Rubin's rules (mice::pool), report revised estimates.

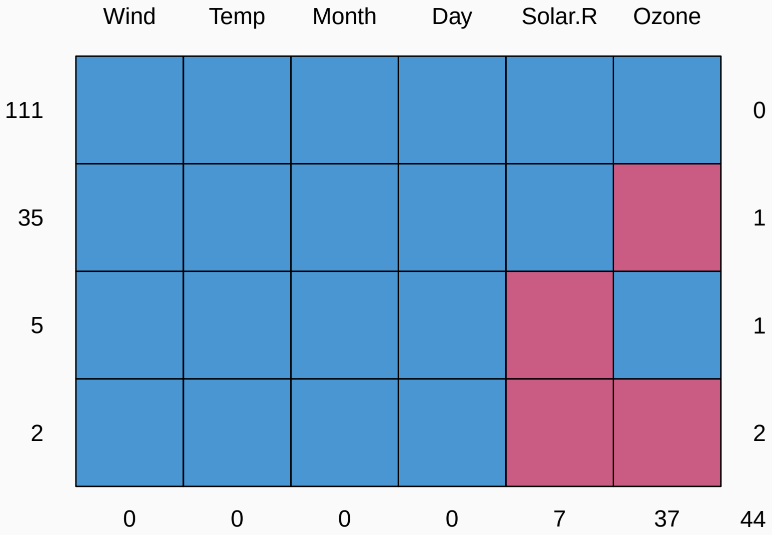
With simple data

```r
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
## 
```

Visualize missingness

```r
md.pattern(airquality)
```




```
##     Wind Temp Month Day Solar.R Ozone
## 111    1    1     1   1       1     1  0
```

Evaluating the distributions of means across missing and non-missing values

```r
airquality %>%
    group_by(is.na(Ozone), is.na(Solar.R)) %>%
    summarize(across(everything(), mean))
```

```
## # A tibble: 4 x 8
## # Groups:   is.na(Ozone) [2]
##   `is.na(Ozone)` `is.na(Solar.R)` Ozone Solar.R  Wind  Temp Month   Day
##   <lgl>          <lgl>            <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 FALSE          FALSE             42.1    185.  9.94  77.8  7.22  15.9
## 2 FALSE          TRUE              42.8     NA   8.14  79.6  6.8    6.4
## 3 TRUE           FALSE             NA      190. 10.2   79.1  6.43  16.7
## 4 TRUE           TRUE              NA       NA  11.2   56.5  5     16
```

Let’s get set up

```
library(mice)
# initiate an empty object, maxit = 0 prevents it from running
airquality_impTemp <- mice(airquality, maxit = 0)
```

Key components: Predictor matrix

```
airquality_impTemp$predictorMatrix
```

```
##         Ozone Solar.R Wind Temp Month Day
## Ozone       0       1    1    1     1   1
## Solar.R     1       0    1    1     1   1
## Wind        1       1    0    1     1   1
## Temp        1       1    1    0     1   1
## Month       1       1    1    1     0   1
## Day         1       1    1    1     1   0
```

Predictor matrices

- Columns indicate variables to be imputed
- Rows indicate predictors to include in imputation model
- Typically, we want to include as many predictors as is possible
- Let's disable `Day`

```
predMat <- airquality_impTemp$predictorMatrix
predMat[, "Day"] <- 0
predMat
```

```
##         Ozone Solar.R Wind Temp Month Day
## Ozone       0       1    1    1     1   0
## Solar.R     1       0    1    1     1   0
## Wind        1       1    0    1     1   0
## Temp        1       1    1    0     1   0
## Month       1       1    1    1     0   0
## Day         1       1    1    1     1   0
```

Key components: Imputation method

```
meth <- airquality_impTemp$method
meth
```

```
##    Ozone Solar.R    Wind    Temp   Month     Day
##    "pmm"   "pmm"      ""      ""      ""      ""
```

- Partial mean matching (pmm) is the general default for mice. It uses a bootstrap-like algorithm to impute missing values based on similarity to other cases in the data
- Other methods are available, but pmm is often best
- We can swap methods easily; see `https://www.gerkovink.com/miceVignettes/Convergence_pooling/Convergence_and_pooling.html`

Running the imputation

M controls the number of imputations, maxit controls the number of iterations of the sampler run per imputation.

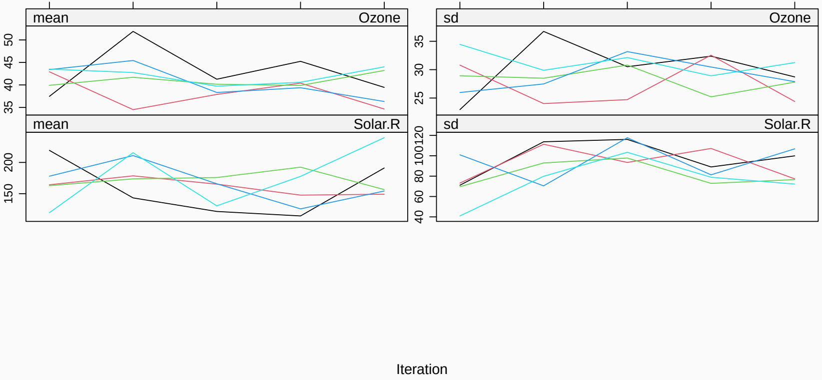
```
airquality_imp <- mice(airquality, predictorMatrix = predMat, method = meth, m = 5,
    maxit = 5)
```

```
##
##  iter imp variable
##   1   1  Ozone  Solar.R
##   1   2  Ozone  Solar.R
##   1   3  Ozone  Solar.R
##   1   4  Ozone  Solar.R
##   1   5  Ozone  Solar.R
##   2   1  Ozone  Solar.R
##   2   2  Ozone  Solar.R
##   2   3  Ozone  Solar.R
##   2   4  Ozone  Solar.R
##   2   5  Ozone  Solar.R
##   3   1  Ozone  Solar.R
##   3   2  Ozone  Solar.R
##   3   3  Ozone  Solar.R
##   3   4  Ozone  Solar.R
##   3   5  Ozone  Solar.R
##   4   1  Ozone  Solar.R
##   4   2  Ozone  Solar.R
##   4   3  Ozone  Solar.R
##   4   4  Ozone  Solar.R
##   4   5  Ozone  Solar.R
```

Diagnostics: trace plots

First, we want to check for convergence. We are looking for the absence of patterns here

```r
plot(airquality_imp)
```

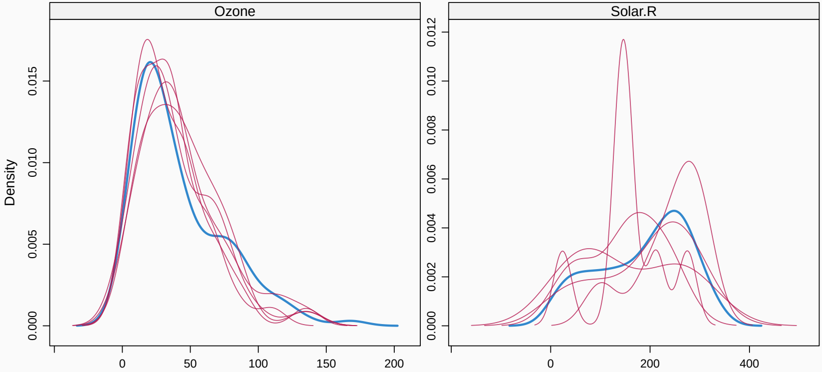


This looks fine

Diagnostics: posterior distributions

Blue line = observed; red line = imputed. Look for generally similar patterns. This looks fine.

```r
densityplot(airquality_imp)
```



Post processing: creating an imputed data frame

```
airquality_imputed <- mice::complete(airquality_imp, action = "long")
nrow(airquality)
```

```
## [1] 153
```

```
nrow(airquality_imputed)
```

```
## [1] 765
```

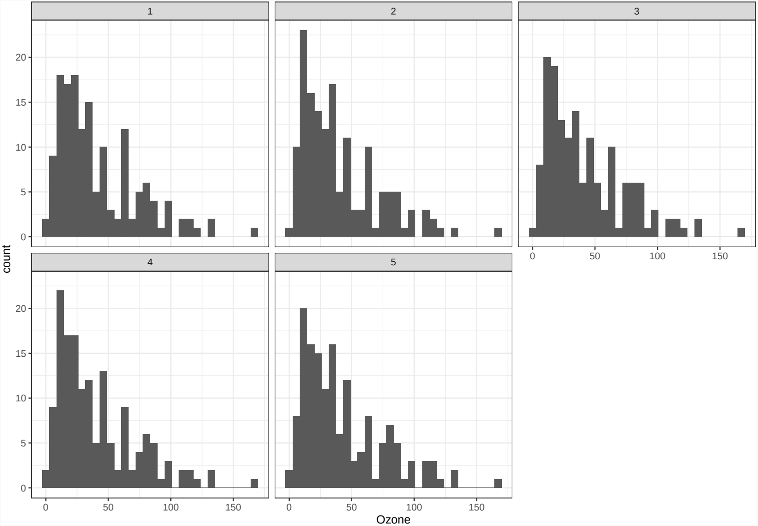
What it looks like

```
head(airquality_imputed)
```

```
##   Ozone Solar.R Wind Temp Month Day .imp .id
## 1    41     190  7.4   67     5   1    1   1
## 2    36     118  8.0   72     5   2    1   2
## 3    12     149 12.6   74     5   3    1   3
## 4    18     313 11.5   62     5   4    1   4
## 5    32     252 14.3   56     5   5    1   5
## 6    28     175 14.9   66     5   6    1   6
```

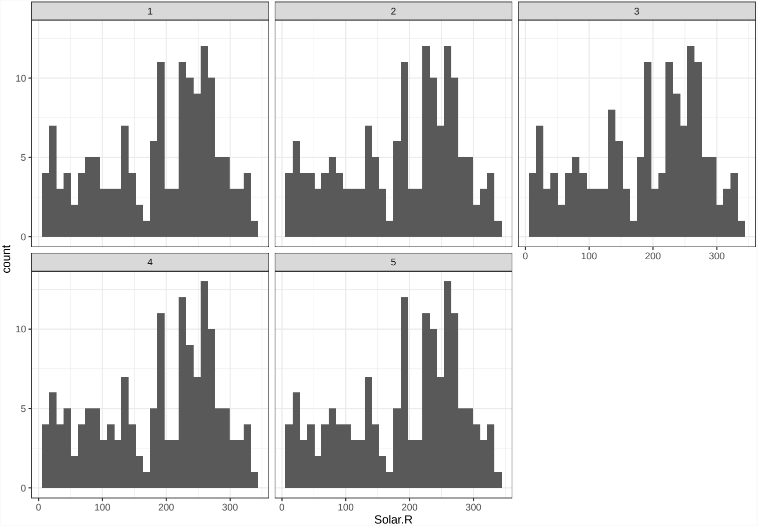
Visualize

```
ggplot(airquality_imputed, aes(x = Ozone)) + geom_histogram() + fac
```



Visualize

```
ggplot(airquality_imputed, aes(x = Solar.R)) + geom_histogram() + f
```



Model estimation

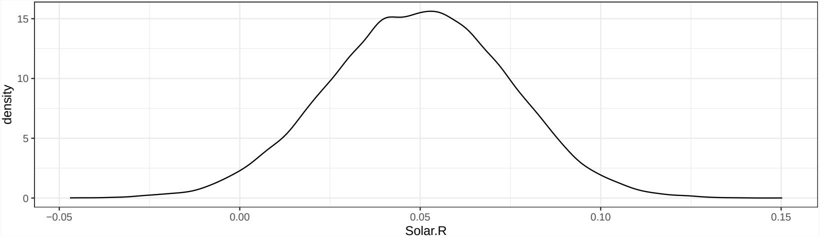
Estimate a regression model over *each* imputed dataset

```
m1 <- stan_glm(Ozone ~ Solar.R + Temp, data = airquality_imputed %>%
    filter(.imp == 1), refresh = 0)
m2 <- stan_glm(Ozone ~ Solar.R + Temp, data = airquality_imputed %>%
    filter(.imp == 2), refresh = 0)
m3 <- stan_glm(Ozone ~ Solar.R + Temp, data = airquality_imputed %>%
    filter(.imp == 3), refresh = 0)
m4 <- stan_glm(Ozone ~ Solar.R + Temp, data = airquality_imputed %>%
    filter(.imp == 4), refresh = 0)
m5 <- stan_glm(Ozone ~ Solar.R + Temp, data = airquality_imputed %>%
    filter(.imp == 5), refresh = 0)
```

With Bayesian models, we can just pool the posterior distributions

```r
posteriors <- bind_rows(data.frame(m1), data.frame(m2), data.frame(m3), data.frame(m4),
    data.frame(m5))

ggplot(posteriors, aes(x = Solar.R)) + geom_density()
```



With frequentist models, we can pool using Rubin’s Rules for combination

```r
m_out <- with(airquality_imp, lm(Ozone ~ Solar.R + Temp))
summary(pool(m_out))
```

```
##          term     estimate   std.error statistic        df      p.value
## 1 (Intercept) -133.1033176 16.16542251 -8.233829 130.84161 1.615342e-13
## 2     Solar.R    0.0496836  0.02580318  1.925484  34.01661 6.256202e-02
## 3        Temp    2.1236321  0.21946188  9.676542  97.62848 6.312213e-16
```

We’ll practice in lab on Wednesday