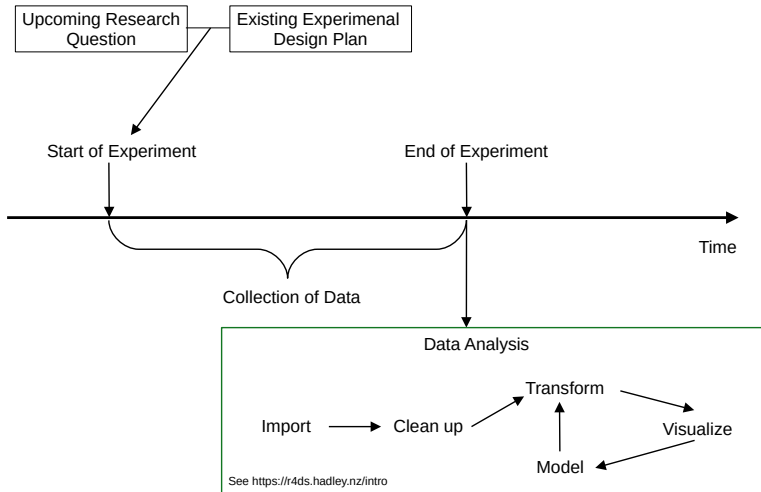# Simulations
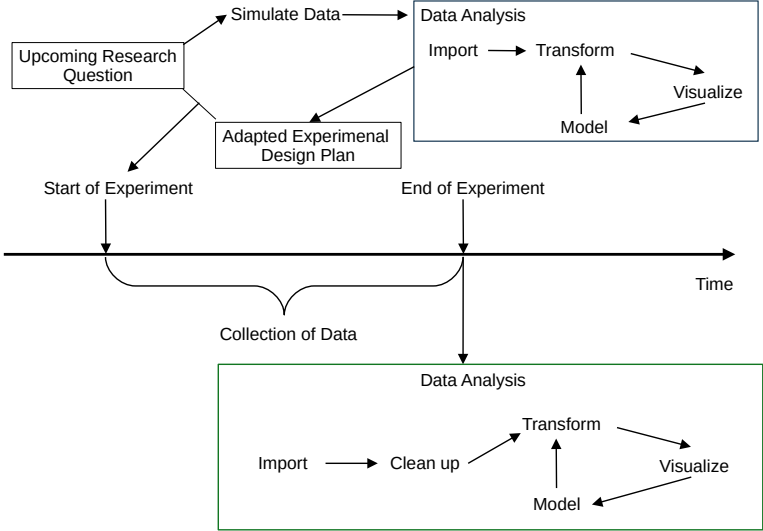
Peter von Rohr

2024-04-22

Timing of Data Analysis

When is the best time to think about data analysis?

Problems

- Dataset is given
- Analysis comes after data collection
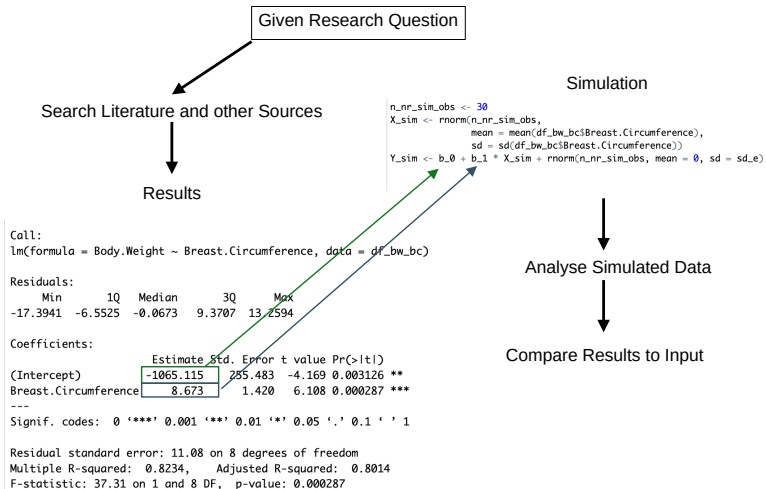- Structure of data cannot be changed

Improvement

Solution

- Plan data analysis before data collection
- Experimental design
- Use simulation . . .

Simulation

What is simulation?

Given Research Question

Search Literature and other Sources

Results

Simulation

```
n_nr_sim_obs <- 30
X_sim <- rnorm(n_nr_sim_obs,
               mean = mean(df_bw_bc$Breast.Circumference),
               sd = sd(df_bw_bc$Breast.Circumference))
Y_sim <- b_0 + b_1 * X_sim + rnorm(n_nr_sim_obs, mean = 0, sd = sd_e)
```

Analyse Simulated Data

Compare Results to Input

```
Call:
lm(formula = Body.Weight ~ Breast.Circumference, data = df_bw_bc)

Residuals:
    Min      1Q  Median      3Q     Max
-17.3941 -6.5525 -0.0673  9.3707 13.2594

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1065.115    255.483  -4.169 0.003126 **
Breast.Circumference    8.673      1.420   6.108 0.000287 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.08 on 8 degrees of freedom
Multiple R-squared:  0.8234,    Adjusted R-squared:  0.8014
F-statistic: 37.31 on 1 and 8 DF,  p-value: 0.000287
```

Procedure

- Given there is an effect between characteristics of variables
- Generate data that contains the effects
- Re-analyse generated data
- Can effects be found?

First Example

Body Weight on Breast Circumference

```
# read data
df_bw_bc <- read.table(s_bw_bc_url, header = T, sep = ",")
# fit regression
lm_bw_bc <- lm(Body.Weight ~ Breast.Circumference,
               data = df_bw_bc)
smry_bw_bc <- summary(lm_bw_bc)
```

First Example II: Simulation

```r
# parameter from results
b_0 <- smry_bw_bc$coefficients["(Intercept)", "Estimate"]
b_1 <- smry_bw_bc$coefficients["Breast.Circumference",
                                "Estimate"]
sd_e <- smry_bw_bc$sigma
# simulate data
set.seed(2204)
n_nr_sim_obs <- 30
X_sim <- rnorm(n_nr_sim_obs,
               mean = mean(df_bw_bc$Breast.Circumference),
               sd = sd(df_bw_bc$Breast.Circumference))
Y_sim <- b_0 + b_1 * X_sim +
  rnorm(n_nr_sim_obs, mean = 0, sd = sd_e)
```

First Example III: Data Analysis

```
df_sim <- data.frame(X.sim = X_sim, Y.sim = Y_sim)
lm_sim <- lm(Y.sim ~ X.sim, data = df_sim)
summary(lm_sim)
```

```
##
## Call:
## lm(formula = Y.sim ~ X.sim, data = df_sim)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -19.1698  -4.7765   0.4339   7.0468  17.1173
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1294.5609   127.0850  -10.19 6.39e-11 ***
## X.sim           9.9453     0.7107   13.99 3.66e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

More Examples

- Differences between feeds ...
- Differences between treatment and controls ...
- Differences betwween healthy and diseased ...