Lecture 3: Beginning statistical simulations

A new question

In STA 112, you learned about the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Question: What assumptions does this model make?

- Normality of ε_i
- constant variance of ε_i
- independence of ε_i
- linearity (shape)
- ε_i have mean 0
- randomness

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

A new question

In STA 112, you learned about the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Question: How important is it that $\varepsilon_i \sim N(0, \sigma^2)$? Does it matter if the errors are *not* normal?

Activity

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Activity: With a neighbor, brainstorm how you could use simulation to assess the importance of the normality assumption (you do not need to write code!).

- How would you simulate data?
- What result would you measure for each run of the simulation?

Activity

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

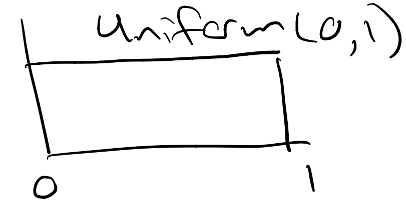
How would you study the importance of the normality assumption?

95% CI : $\hat{\beta}_1 \pm t^*_{n-2} SE(\hat{\beta}_1)$

depends on $\varepsilon_i \sim$ Normal

- do the simulations many times
- simulate data with different distributions for ε_i
  - e.g. Poisson distribution, binomial distribution, exponential, χ^2, Normal
- compare performance for different distributions of ε_i
  - e.g. · SSE
  - · confidence intervals for β_1
- construct 95% CIs for β_1 and check whether they capture the β_1 95% of time in practice

Simulating data



To start, simulate data for which the normality assumption holds:

```
n <- 100 # sample size
beta0 <- 0.5 # intercept
beta1 <- 1 # slope

x <- runif(n, min=0, max=1)
noise <- rnorm(n, mean=0, sd=1)
y <- beta0 + beta1*x + noise
```

β_0

β_1

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Can have whatever distribution we want

- `runif(n, min=0, max=1)` samples X_i uniformly between 0 and 1

number of samples

- `rnorm(n, mean=0, sd=1)` samples $\varepsilon_i \sim N(0, 1)$

simulate n observations from Uniform(0,1)

runif

"random"

abbrev. for distribution name

other dists: rexp, rchisq, rpois, etc.

Fit a model

```
n <- 100 # sample size
beta0 <- 0.5 # intercept
beta1 <- 1 # slope

x <- runif(n, min=0, max=1)
noise <- rnorm(n, mean=0, sd=1)
y <- beta0 + beta1*x + noise

lm_mod <- lm(y ~ x)
lm_mod
```

generating Xs & Ys

fit linear regression model

response

explanatory

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.2971       1.3073
```

$\hat{\beta}_0$ $\hat{\beta}_1$

The value for $\beta_1 = 1$

compare $\hat{\beta}_1$ to β_1

compare a CI for β_1 to the value of β_1

Calculate confidence interval

```
lm_mod <- lm(y ~ x)

ci <- confint(lm_mod, "x", level = 0.95)
ci
```

calculate a CI

95% CI

```
      2.5 %   97.5 %
x 0.6777885 1.936822
```

vector

fitted model

coefficient of interest (β_1)

- **Question:** How can we check whether the confidence interval contains the true β_1?

CI = (0.68, 1.94) contains $\beta_1 = 1$

0.68 < 1 & 1.94 > 1 (TRUE)

ci[1] < 1 & ci[2] > 1

Calculate confidence interval

```
lm_mod <- lm(y ~ x)

ci <- confint(lm_mod, "x", level = 0.95)
ci
```

```
      2.5 %   97.5 %
x 0.6777885 1.936822
```

- **Question:** How can we check whether the confidence interval contains the true β_1 ?

```
ci[1] < 1 & ci[2] > 1
```

```
[1] TRUE
```

Repeat!

```r
nsim <- 1000
n <- 100 # sample size
beta0 <- 0.5 # intercept
beta1 <- 1 # slope
results <- rep(NA, nsim)

for(i in 1:nsim){
  x <- runif(n, min=0, max=1)
  noise <- rnorm(n, mean=0, sd=1)
  y <- beta0 + beta1*x + noise

  lm_mod <- lm(y ~ x)
  ci <- confint(lm_mod, "x", level = 0.95)

  results[i] <- ci[1] < 1 & ci[2] > 1
}
mean(results)
```

set up parameters

results vector to store results

sample data

fit model & calculate 95% CI

check if CI contains β_1 store result

observed coverage

- What fraction of the time should the confidence interval contain β_1?

expect ≈ 0.95

Repeat!

```
nsim <- 1000
n <- 100 # sample size
beta0 <- 0.5 # intercept
beta1 <- 1 # slope
results <- rep(NA, nsim)

for(i in 1:nsim){
  x <- runif(n, min=0, max=1)
  noise <- rnorm(n, mean=0, sd=1)
  y <- beta0 + beta1*x + noise

  lm_mod <- lm(y ~ x)
  ci <- confint(lm_mod, "x", level = 0.95)

  results[i] <- ci[1] < 1 & ci[2] > 1
}
mean(results)
```

```
[1] 0.948
```

← next step: change distribution of ε_i

- What should we do next?

Class activity

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

That is, how important is the assumption that $\varepsilon_i \sim N(0, \sigma^2)$?

Continue simulation from last time, but experiment with different values of n and different distributions for the noise term.

https://sta279-s24.github.io/class_activities/ca_lecture_3.html