Simulation based inference activity: solutions

STAT3530 Applied Linear Models

Solutions to the two exercises in `scripts/11-simulation-inference.R`. The test being reproduced is $H_0 : \beta_{\text{dpi}} = \beta_{\text{ddpi}} = 0$.

Starting point:

```
m1 <- lm(sr ~ ., data = savings)               # full model
m0 <- lm(sr ~ pop15 + pop75, data = savings)   # model under H0

estimand <- function(y)
  anova(lm(y ~ pop15 + pop75, data = savings),
        lm(y ~ pop15 + pop75 + dpi + ddpi, data = savings))$F[2]

sim.data <- function(fit) fitted(fit) + sample(resid(fit))

set.seed(1)
null <- replicate(2000, {
  yb <- sim.data(m0)
  Fb <- estimand(yb)
  Fb
})

p.rand <- mean(null >= estimand(savings$sr))
p.rand
```

```
[1] 0.087
```

1. Change the simulation method

Only `sim.data()` changes.

```
sim.data <- function(fit) fitted(fit) + rnorm(n = 50, mean = 0, sd = sigma(fit))

set.seed(1)
null <- replicate(2000, {
  yb <- sim.data(m0)
  Fb <- estimand(yb)
  Fb
})

p.boot <- mean(null >= estimand(savings$sr))
p.boot
```

```
[1] 0.0805
```

All three agree:

```
c(parametric    = anova(m0, m1)$Pr[2],
  randomization = p.rand,
  bootstrap     = p.boot)
```

```
   parametric randomization     bootstrap
   0.08470885    0.08700000    0.08050000
```

The residuals of `m0` pass a normality check ($p = 0.99$), so shuffling them and drawing fresh ones from $N(0, \hat{\sigma}^2)$ do nearly the same thing, and the parametric F is the right reference. Either way, $p \approx 0.08$ is weak evidence at best.

2. Change the estimand

The statistic is now R^2, and the simulation runs from `m1`.

```
estimand <- function(y) {
  summary(lm(y ~ pop15 + pop75 + dpi + ddpi, data = savings))$r.squared
}

estimand(savings$sr)
```

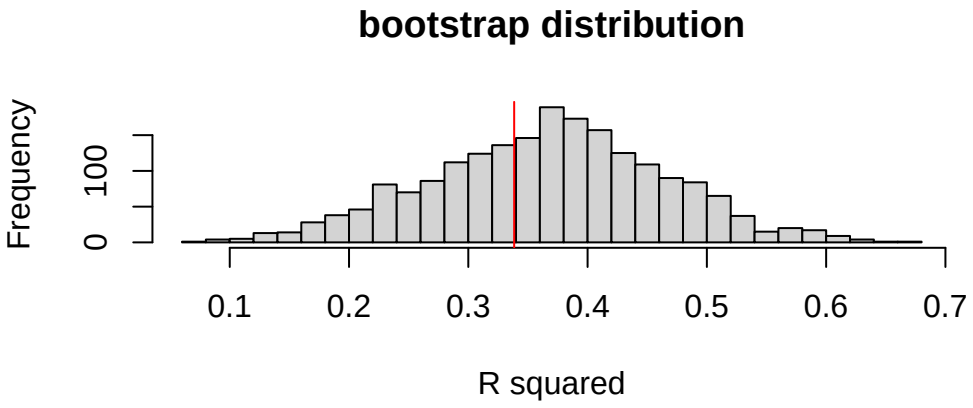
```
[1] 0.3384564
```

Bootstrap, not randomization: there is no null to impose, so simulate from the full fit.

```
sim.data <- function(fit) fitted(fit) + rnorm(n = 50, mean = 0, sd = sigma(fit))

set.seed(1)
null <- replicate(2000, {
  yb <- sim.data(m1)
  estimand(yb)
})
```

```
hist(null, breaks = 40, xlab = 'R squared', main = 'bootstrap distribution')
abline(v = estimand(savings$sr), col = 'red')
```



Three intervals:

```
# percentile
ci1 <- quantile(null, c(0.025, 0.975))

# symmetric, half width from the bootstrap
me  <- abs(null - estimand(savings$sr)) |> quantile(0.95)
ci2 <- estimand(savings$sr) + c(-1, 1)*me

# symmetric, bootstrap standard error
bse <- sd(null)
ci3 <- estimand(savings$sr) + c(-1, 1)*1.96*bse

round(rbind(percentile = ci1,
            half.width = ci2,
            boot.se    = ci3), 3)
```

```
            2.5% 97.5%
percentile 0.174 0.561
half.width 0.143 0.534
boot.se    0.146 0.531
```

All three are wide: with $n = 50$ and four predictors, R^2 is not pinned down. The two symmetric intervals agree because the bootstrap distribution is nearly symmetric. The percentile interval sits about 0.03 higher at both ends because R^2 is biased upward — the bootstrap distribution centers at 0.366, not at the observed 0.338.