APSTA-GE 2003: Intermediate Quantitative Methods

Sample Solution - Assignment 4

Created on: 11/09/2020

Modified on: 11/10/2020

Part 1

In this part, you will conduct some simulation exercises using the Shiny App.

https://tongj.shinyapps.io/2003iqm/

Question 1

Conduct a simulation study using the following values of the population parameters:

- Intercept: 1
- Slope: -1
- Variance of the error term: 9
- Sample size: 100

Report the estimated intercept (beta0): ____

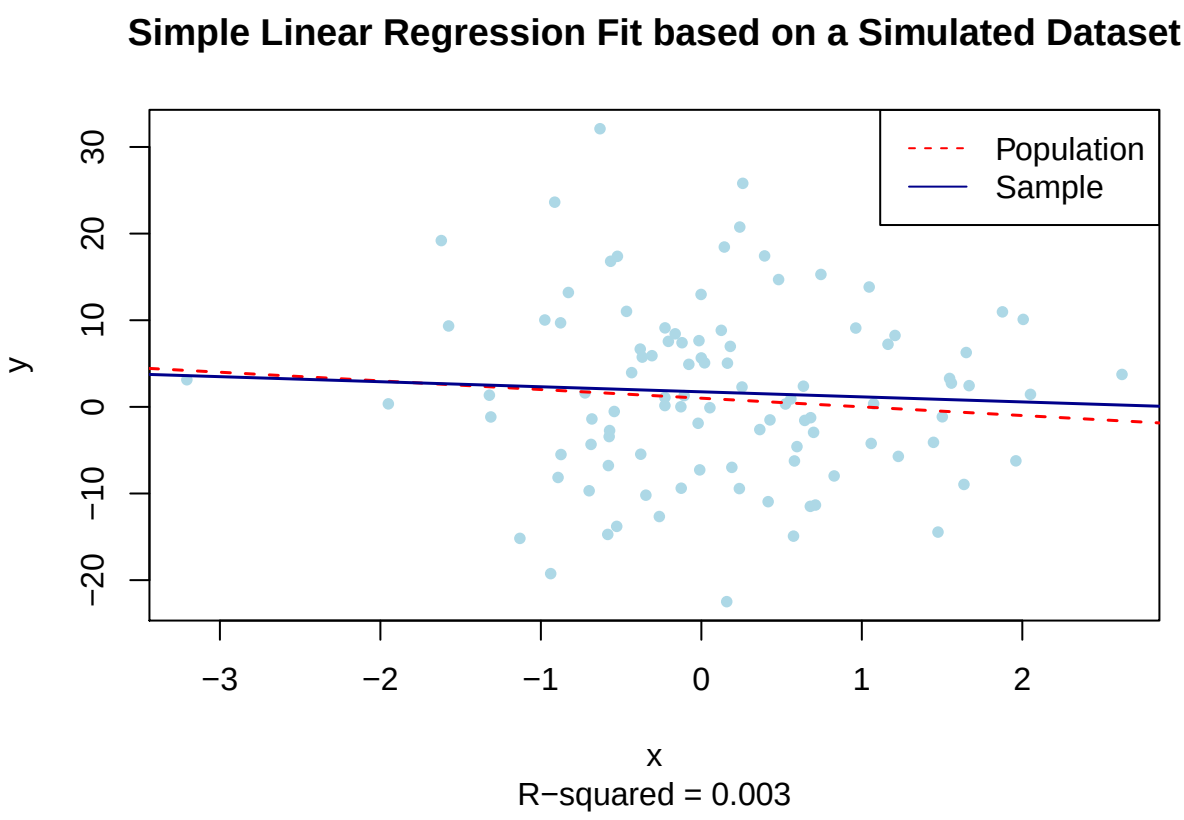
Report the estimated slope (beta1): ____

Answer: Q1

First, select the `Linear Regression Simulator` tab. Then, on the left sidebar, put in parameters. Click "Draw sample and fit model".

```
set.seed(111)

rd <- function (n) {
  n <- as.numeric(n)
  round(n, digits = 3)
}

simulate_lin_mod <- function (n, beta0, beta1, var_error) {
  # Draw random samples from a population and fit a linear regression model
  # Param: n: sample size
  # Param: beta0: population intercept
  # Param: beta1: population slope
  # Param: var_error: residual variance
  pop <- rnorm(1000, mean = 0, sd = 1)
  x <- sample(pop, size = n, replace = FALSE)
  error <- rnorm(n, mean = 0, sd = var_error)
  y <- beta0 + beta1 * x + error
  dat <- data.frame(x, y)
```

```r
  lin_mod <- lm(y ~ x, data = dat)
  list(lin_mod, x, y)
}

beta0 <- 1
beta1 <- -1
var_error <- 9
n <- 100

LRS_Q1 <- simulate_lin_mod(n, beta0, beta1, var_error)
intercept_mod_Q1 <- rd(LRS_Q1[[1]]$coefficients["(Intercept)"])
slope_mod_Q1     <- rd(LRS_Q1[[1]]$coefficients["x"])
plot(LRS_Q1[[2]], LRS_Q1[[3]], pch = 16, col = "light blue", cex = 0.8, ann = FALSE)
abline(a = 1, b = -1, lty = "dashed", lwd = 1.5, col = "red")
abline(a = intercept_mod_Q1, b = slope_mod_Q1, lwd = 1.5, col = "dark blue")
legend("topright", legend = c("Population", "Sample"),
       lty = c("dashed", "solid"), col = c("red", "dark blue"))
title(main = "Simple Linear Regression Fit based on a Simulated Dataset",
      sub = paste("R-squared =", round(summary(LRS_Q1[[1]])$r.squared, digits = 3)),
      xlab = "x", ylab = "y")
```



The estimated intercept should between 0 and 2.

The estimated slope should between -2 and 1.

Extra: The `Linear Regression Simulator` randomly draws samples from a population $P(x, y)$ where P follows a normal distribution with mean at zero and standard deviation of 1:

$$P(x, y) \sim \mathcal{N}(0, 1^2)$$

The relationship between x and y in the population is linear:

$$y_P = \beta_0 + \beta_1 \cdot x_P + \varepsilon$$

where β_0 is the input `Intercept`, β_1 is the input `slope`, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is the input `Variance of the error term`.

In this case, we can write down the linear regression equation of the population:

$$y_P = -1 + 1 \cdot x_P + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, 3^2)$$

Since the sample is randomly drawed from the population and has more than 30 data points, we expect the sample regression estimation to be close to the population's regression line. However, because we only fit the model once, variance does exist. This is why our sample and population regression lines do not overlay. Therefore, for better estimation, we need to draw multiple samples, fit regression models, and compute the average.

Question 2

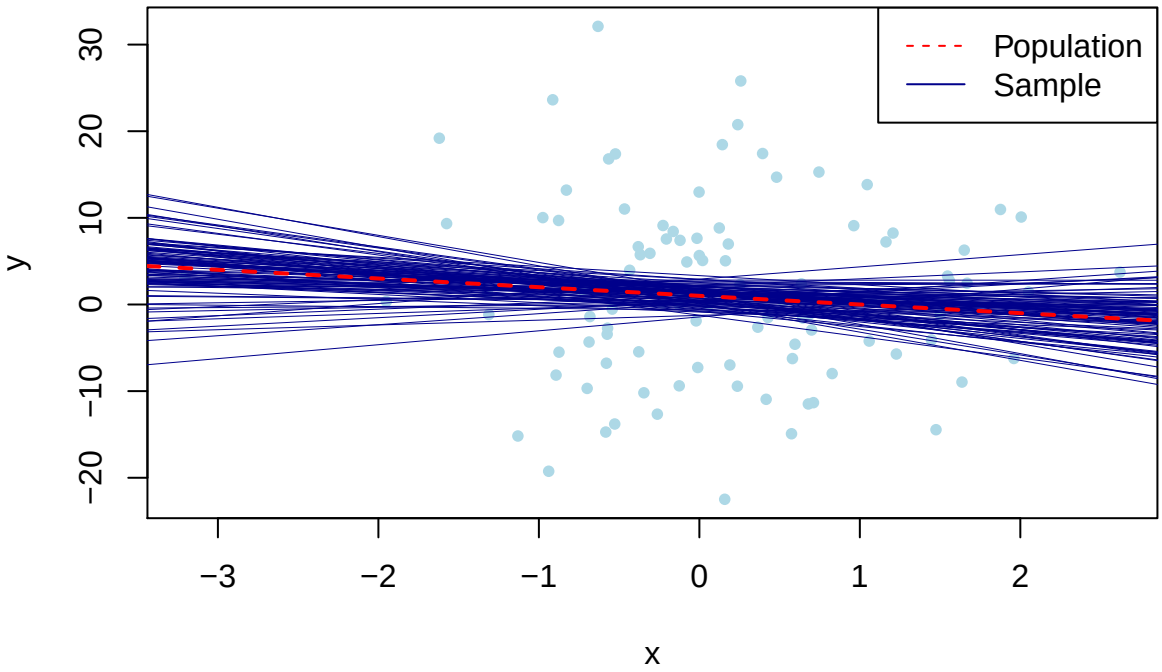
Using the same simulation set up as in 1, draw multiple samples and examine the regression outcomes across multiple samples.

The averages of estimated slope coefficients are about the same as the population slope value that they are estimating.

Answer: Q2

To do this, just keep clicking on the button to draw multiple regression lines. Record the slope for each iteration and compute the average. According to the Central Limit Theorem, for large random samples, the distribution of the sample means will be approximately normally distributed. This means that the sample means will be centered at the mean of means, which is (very close to) the population slope.

```r
# Repeat 100 times and visualize
iter <- 100
plot(LRS_Q1[[2]], LRS_Q1[[3]], pch = 16, col = "light blue", cex = 0.8, ann = FALSE)
legend("topright", legend = c("Population", "Sample"),
       lty = c("dashed", "solid"), col = c("red", "dark blue"))
title(main = "Simple Linear Regression Fit based on Many Samples",
      xlab = "x", ylab = "y")
for (i in 1:n) {
  LRS_Q2 <- simulate_lin_mod(n, beta0, beta1, var_error)
  intercept_mod_Q2 <- rd(LRS_Q2[[1]]$coefficients["(Intercept)"])
  slope_mod_Q2 <- rd(LRS_Q2[[1]]$coefficients["x"])
  abline(a = intercept_mod_Q2, b = slope_mod_Q2, lwd = 0.5, col = "dark blue")
}
abline(a = 1, b = -1, lty = "dashed", lwd = 2, col = "red")
```



Question 3

The standard deviations of the sample estimates of the regression coefficients captures the variability in these estimates. Standard errors reported in the regression are the estimates of these quantities.

Answer: Q3

This is true. Please refer to the plot in Answer: Q2.

Question 4

In this simulation study, when testing whether the slope coefficient for X is 0 or not, we already know that the null hypothesis is FALSE. Hence if we fail to reject the null hypothesis, our conclusion would be wrong.

Answer: Q4

This is true. If the null hypothesis is FALSE, the correct way is to reject it. If it is TRUE, then accept it.

Question 5

Report the power of testing H_0: β_1 (slope) = 0 using the Hypothesis Testing/many samples tab.

Answer: Q5

Click on the "Hypothesis Testing" tab. On the left sidebar, set the sample size to 100. Then, report the power.

The power is the probability of rejecting the null hypothesis. The true slope, in our case, is 1. The null hypothesis tests whether or not the true slope is 0. The power should be between 0.6 and 0.95.

Question 6

Now let's change the simulation settings:

- Intercept: 1
- Slope: -1
- Variance of the error term: 9
- Sample size: 400

The standard error of the slope coefficient (beta1) estimation is about half of the SE that was obtained from the previous simulation (when n = 100)

Answer: Q6

```
mod_Q1 <- LRS_Q1[[1]]
se_Q1 <- summary(mod_Q1)$coefficients["x", "Std. Error"]

beta0 <- 1
beta1 <- -1
var_error <- 9
n <- 400
LRS_Q6 <- simulate_lin_mod(n, beta0, beta1, var_error)

mod_Q6 <- LRS_Q6[[1]]
se_Q6 <- summary(mod_Q6)$coefficients["x", "Std. Error"]

print(c(se_Q1, se_Q6))
```

```
## [1] 1.059234 0.468807
```

This is true.

Question 7

Now let's change the simulation settings again:

- Intercept: 1
- Slope: 0
- Variance of the error term: 9
- Sample size: 400

When testing H_0: β_1 (slope) = 0, now if we reject the null hypothesis, we will make a mistake.

Answer: Q7

This is true because the true slope is 0 now. The null hypothesis is TRUE.

Question 8

Use the simulation App, answer the following TRUE/FALSE question.

For a level 5% test, the type I error rate is always around 5%, regardless the sample size.

Hint: Type I error is defined as the percent of mistakenly rejecting the null hypothesis when the null is true (i.e. true beta is the same as the hypothesized beta).

Answer: Q8

This is true because p-value is the probability of mistakenly reject the null hypothesis when it is true. Therefore, p-value is equal to Type I error rate.

Question 9

When model variance increases, the SE of beta also increases.

Answer: Q9

This is false. Model variance measures model complexity. The higher the complexity, the higher the variance, the lower the standard error of coefficient estimation.

Question 10

The distribution of the p-values over multiple replications of tests is a t distribution.

Answer: Q10

This is false. The p-value is the area under the curve (AUC) of the t distribution curve on both tails (for two-tail). The distribution of T scores is a t distribution.

Part 2

In this part, you will conduct a multiple regression analysis using the dataset: `toy_example2.csv`

```
dat <- read.csv("../data/toy_example2.csv")
str(dat)
```

```
## 'data.frame':    500 obs. of  5 variables:
##  $ id         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Sex        : Factor w/ 2 levels "F","M": 1 2 1 2 1 2 1 1 2 2 ...
##  $ Birth.Order: int  1 2 1 2 4 2 4 1 2 1 ...
##  $ Height     : num  72.9 67.7 68 67.9 64.2 ...
##  $ Weight     : num  123 166 118 120 125 ...
```

Question 1

Conduct a multiple regression, using `weight` as D.V., `height`, Sex and `birth.order` as I.V.

Instead of using "Sex" directly, generate a dummy variable "`female`":

`female` = 1 if Sex = "F"

`female` = 0 if Sex = "M"

The mean of this new dummy variable "female" is 0.514.

Answer: P2Q1

```
dat$female <- ifelse(dat$Sex == "F", yes = 1, no = 0)
mod_P2Q1 <- lm(Weight ~ Height + female + Birth.Order, data = dat)

mean(dat$female)
```

```
## [1] 0.514
```

This is true.

Question 2

At level 5%, which ones of the following independent variables are significant in the above regression model? (Choose all that apply)

Answer: P2Q2

```
summary(mod_P2Q1)
```

```
##
## Call:
## lm(formula = Weight ~ Height + female + Birth.Order, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.879 -12.899   0.154  13.019  59.782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.5126    15.5221   4.027 6.53e-05 ***
## Height        1.1546     0.2185   5.283 1.90e-07 ***
## female       -9.1028     1.9312  -4.714 3.17e-06 ***
## Birth.Order  -0.2727     1.0110  -0.270    0.788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.56 on 496 degrees of freedom
## Multiple R-squared:  0.1498, Adjusted R-squared:  0.1446
## F-statistic: 29.12 on 3 and 496 DF,  p-value: < 2.2e-16
```

`Female` and `Height`.

Question 3

The R-square of this model is ____ .

Answer: P2Q3

```
summary(mod_P2Q1)$r.square
```

```
## [1] 0.149752
```

Question 4

What's the expected weight of a female who is height is 65 inches, and is the first child (birth.order=1)?

Answer: P2Q4

$$\text{Weight} = 62.5126 + 1.1546 \cdot 65 - 9.1028 \cdot 1 - 0.2727 \cdot 1 = 128.1861$$

Question 5

In one sentence, interpret the regression coefficient for "`female`" in this model.

Answer: P2Q5

Controlling for height and birth order, on average, female is 9 pounds lighter than male.

Question 6

If we were to replace the dummy variable "female" by the dummy variable "male" in this regression model, only the intercept and coefficient for male will be different, the other coefficients will be the same. The R squared of the model will also remain the same.

Answer: P2Q6

```
dat$male <- ifelse(dat$female == 1, yes = 0, no = 1)

mod_P2Q6 <- lm(Weight ~ Height + male + Birth.Order, data = dat)

coefficients(mod_P2Q1)
```

```
## (Intercept)      Height      female Birth.Order
##  62.5126286   1.1546139  -9.1028327  -0.2726655
```

```
coefficients(mod_P2Q6)
```

```
## (Intercept)      Height        male Birth.Order
##  53.4097959   1.1546139   9.1028327  -0.2726655
```

```
summary(mod_P2Q1)$r.square
```

```
## [1] 0.149752
```

```
summary(mod_P2Q6)$r.square
```

```
## [1] 0.149752
```

This is true.

END: Sample Solution - Assignment 4