Problem 5

HW1

Hanyuan Chi(chixx105), Zhi Shen(shenx704)

April 10, 2017

```
suppressPackageStartupMessages({
  library(purrr)
  library(broom)
  library(tidyr)
  library(ggplot2)
  library(dplyr)
})

set.seed(12345)
```

Monte-Carlo Simulation for Measurement Errors

Question 1

Please simulate 1 sample containing $N = 100L$ observations for the following linear regression model

$$y = 0.2 + 0.5 \cdot x + \varepsilon$$

where $\varepsilon \sim N(0, 1)$.

- Hint:
  - Assume X_t is picked uniformly randomly in $[-1, 1]$ interval
- Assume that the true value of `X` is called `Xt` but the data scientist only observes the "noisy" version called `X`
  - such that $X = X_t + \eta$ where $\eta \sim N(0, 0.25)$ i.i.d
  - in other words, the data is generated with `Xt` as a regressor but is estimated with `X` as a regressor.
- Output:
  - Please create a *data.frame* `df1` that contains numeric vector `df1$X` contains the generated `X` variable, `df1$Y` contains the generated dependent variables and `df1$e` contains the generated disturbances.

Important: $N(0, 0.25)$ means that the mean is 0 and the variance is 0.25. Please use `rnorm` function carefully as it is asking for standard deviation!

```
N <- 100L
set.seed(12345)

# These are true population coefficients
b0 <- 0.2
b1 <- 0.5

df1 <- data.frame(X_t = runif(N,-1,1),
                  e = rnorm(N,0,1),
                  eta = rnorm(N,0,0.5))%>%
  mutate(X = X_t + eta,Y = b0 + b1*X_t + e) %>%
  select(X,e,Y)
```

Question 2

Please use regular OLS model to estimate the coefficients b from that sample. Please report these coefficients as well as the standard error estimates and 95% confidence interval.

- Hint:
  - Use `lm()` for linear model and `summary()` for display purposes

```
# Please write your code below
lm_model <- lm(Y~X, data = df1)

summary(lm_model)
```

```
##
## Call:
## lm(formula = Y ~ X, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.40120 -0.98139  0.05106  0.68885  2.34589
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3479     0.1157   3.007  0.00336 **
## X             0.1778     0.1556   1.143  0.25572
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.153 on 98 degrees of freedom
## Multiple R-squared:  0.01316,    Adjusted R-squared:  0.003092
## F-statistic: 1.307 on 1 and 98 DF,  p-value: 0.2557
```

```
confint(lm_model)
```

```
##                  2.5 %    97.5 %
## (Intercept)  0.1182659 0.5774793
## X           -0.1308519 0.4865236
```

```
tidy(lm_model,conf.int = TRUE)
```

```
##          term  estimate std.error statistic     p.value   conf.low
## 1 (Intercept) 0.3478726 0.1157019  3.006628 0.003355751  0.1182659
## 2           X 0.1778358 0.1555519  1.143257 0.255716985 -0.1308519
##   conf.high
## 1 0.5774793
## 2 0.4865236
```

Question 3

Please revise your code from Question 1 to generate $R = 2000L$ independent samples with $N = 100L$ observations each

- Hint:
  - Try to avoid using loops. Use `dplyr`.
  - Think very carefully about which elements you need to resample and which elements you *do not need* to resample
  - To answer the above, please remember the actual assumptions of an ordinary linear regression
- Output:
  - Please create a *data.frame* `df3` that contains numeric vector `df3$X` contains the generated `X` variable, `df3$Y` contains the generated dependent variables and `df3$e` contains the generated disturbances, `df3$id` contains the id of the sample

```
set.seed(12345)

R <- 2000L

# Please write your code below

df3 <- data.frame(X_t = rep(runif(N,-1,1),times = R),
                  id = rep((1:R),each = N),
                  e = rnorm(N*R,0,1),
                  eta = rnorm(N*R,0,0.5))%>%
  mutate(X = X_t + eta,Y = b0 + b1*X_t + e) %>%
  select(X,e,Y,id)
```

Question 4

Please revise your code from Question 2 to estimate R coefficients b from each of those samples. This implies that you should generate a set of R coefficient estimates.

- Hint:
  - Go for long format instead of wide format when necessary.
  - Try to avoid using loops. Use `tidyr` and `nest()`. Also, you may want to use `purrr::map()` and `broom::tidy()`.

```
# Please write your code below
df4 <- df3 %>%
  group_by(id) %>%
  nest() %>%
  mutate(estimated_model = map(data, ~lm(Y~X, data =.))) %>%
  mutate(estimated_coef = map(estimated_model, ~tidy(., conf.int = TRUE))) %>%
  unnest(estimated_coef)

head(df4)
```

```
## # A tibble: 6 x 8
##      id        term  estimate std.error statistic     p.value    conf.low
##   <int>       <chr>     <dbl>     <dbl>     <dbl>       <dbl>       <dbl>
## 1     1 (Intercept) 0.3533461 0.1153653  3.062845 0.002830673  0.12440733
## 2     1           X 0.1928755 0.1631781  1.181994 0.240068360 -0.13094617
## 3     2 (Intercept) 0.2852268 0.0963858  2.959220 0.003867066  0.09395231
## 4     2           X 0.2024213 0.1183944  1.709720 0.090482242 -0.03252855
## 5     3 (Intercept) 0.2720711 0.1017405  2.674166 0.008778385  0.07017031
## 6     3           X 0.3457340 0.1286932  2.686497 0.008482585  0.09034648
## # ... with 1 more variables: conf.high <dbl>
```

Question 5

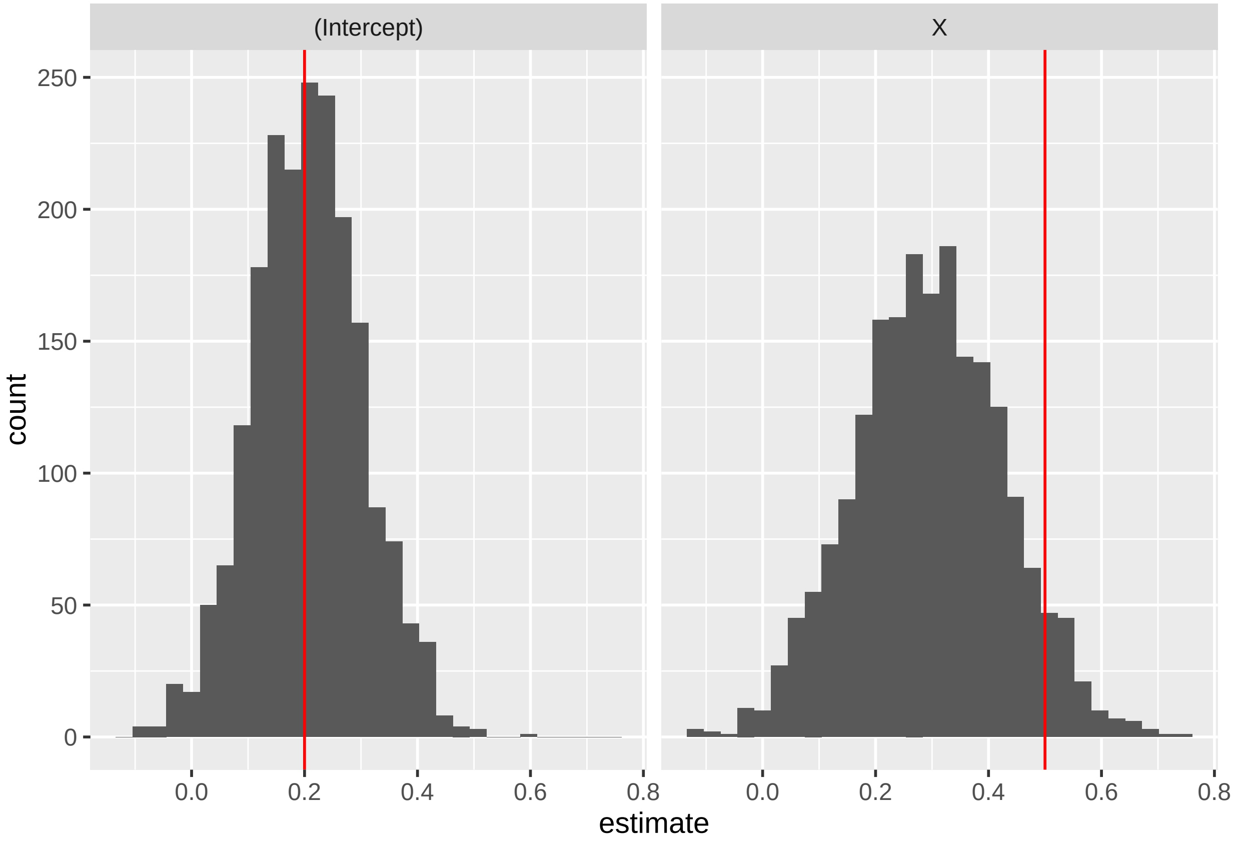
Please plot the histograms of coefficient estimates `b0` and `b1` against the true value

- Hint:
  - Please use `ggplot()`
  - Use `geom_histogram()` to plot the histogram
  - Use `geom_vline(..., color = "red")` to display the true mean
  - Use `facet_grid()` to display them side by side
- Answer the following questions:
  - Is the estimation of true value indeed unbiased?

ANS: The estimation for the intercept(b0) is unbiased, but the estimation for the X coefficient(b1) is biased.

```
# Please write your code below
true_df <- data.frame(term = c("(Intercept)","X"),
                      true_value = c(b0, b1),
                      stringsAsFactors = FALSE
                      )
p5 <- ggplot(df4) +
  geom_histogram(aes(estimate), bins=30) +
  geom_vline(aes(xintercept = true_value),
             color = "red",
             data = true_df) +
  facet_grid(~term)

p5
```



Question 6

Please estimate the true standard deviation of coefficients `b0` and `b1` and compare it to the estimate you obtained in Question 2.

- Answer the following questions:
  - Did Question 2 produce a good estimate of true variability across different samples?

ANS: The standard error of the intercept in Q2 is a good estimate of its true variablity across different samples but the standard error of the X coefficient in Q2 is not.

```
# Please write your code below
df4 %>%
  group_by(term) %>%
  summarise(mean(estimate), sd(estimate))
```

```
## # A tibble: 2 x 3
##          term mean(estimate) sd(estimate)
##         <chr>          <dbl>        <dbl>
## 1 (Intercept)      0.2063660   0.09680971
## 2           X      0.2954248   0.13244665
```

Question 7

Please count how often the 95% confidence interval contains true value for each `b0` and `b1` (separately)

- Hints:
  - Join with true values first, then count
- Answer the following questions:
  - Did 95% confidence interval contain the true value in approximately 95% of cases?

ANS: After trying diffrent seeds, the 95% confidence interval for the intercept indeed contains the true value for approximately 95% of the samples, but the 95% CI for the X coefficient contains the true value much less often, so the 95% CI for the X coefficient is way overconfident.

```
# Please write your code below
df4 %>%
  inner_join(true_df, by="term") %>%
  group_by(term) %>%
  mutate(contains = ifelse(conf.low <= true_value &
                             true_value <= conf.high,
                           1L,0L)) %>%
  summarise(mean(contains))
```

```
## # A tibble: 2 x 2
##          term mean(contains)
##         <chr>          <dbl>
## 1 (Intercept)         0.9565
## 2           X         0.6565
```

Question 8

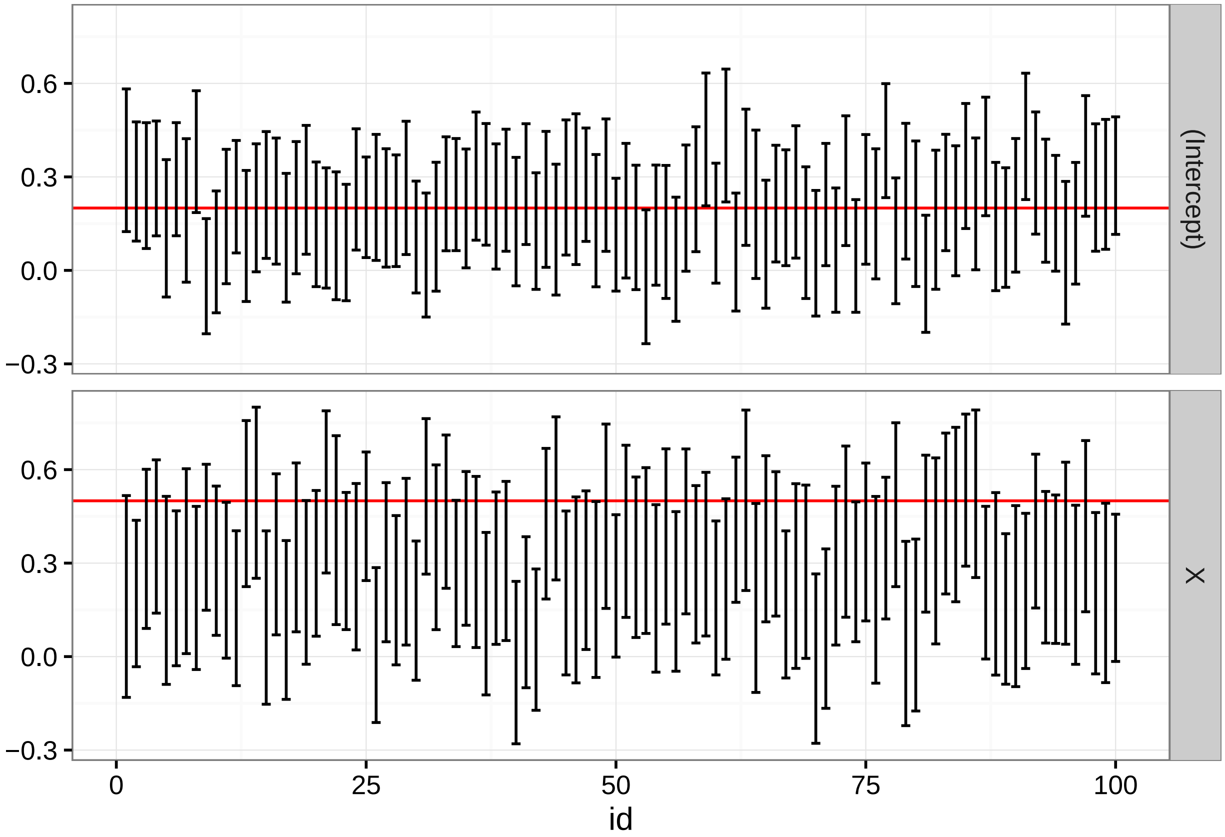
Please plot the first 100 of confidence intervals for both `b0` and `b1`, also please plot the true values

- Hints:
  - Use `geom_errorbar(aes(x=...,ymin=...,ymax=...))` for confidence intervals

    - Use `geom_hline(...)` for true values
    - Use `facet_grid()` for vertical positioning instead of horizontal

```
# Please write your code below
p8 <- ggplot(df4 %>% filter(id<=100)) +
  geom_hline(aes(yintercept = true_value),
             color = "red",
             data = true_df) +
  geom_errorbar(aes(x=id, ymin=conf.low, ymax=conf.high)) +
  facet_grid(term~.) +
  theme_bw()

p8
```



Question 9

Please write down a short summary of the results.

- What kind of a problem will you experience in ordinary linear regression estimation if your regressors have noise in them?

Please be very precise in terms of what is biased and what is not – you need to mention which estimator is biased and which is not biased. You also need to comment on 95% confidence intervals. if you fail to mention some of these, or say, things that are not correct, this will be points off.

(Please do not talk about efficiency of any estimators here as we have no basis to decide it based on these simulations)

Please write your answer below: In ordinary linear regression estimation, if the regressors have noise in them, 1)the estimates for the intercept is unbiased, but the estimate for the X coefficient is biased. 2)The standard error of the intercept is unbiased as a representative of its true variablity across different samples but the standard error of the X coefficient is biased. 3)The 95% confidence interval for the intercept is unbiased because it indeed contains the true value for approximately 95% of the samples, but the 95% CI for the X coefficient contains the true value much less often, so the 95% CI for the X coefficient is way overconfident.