ECON30130: Group Project

GROUP MEMBERS:

RORY O BRIEN: 21356706

RORY BYRNE: 21373851

CAOIMHIN HEFFERNAN: 21387096

DAVID DOORLY: 21347063

DAIRE SWEENEY: 21383193

A

1

Y = B0 + B1X + u

Y= invited for interview (binary dependent variable)

X = whether a person is in an ethnic minority or not. X = 1 if from ethnic minority, X=0 if not from ethnic minority

u= error term

B0= likelihood of non-minority job applicant being invited for an interview (constant). i,e probability that Y=1 if X=0

B1= difference in likelihood between a minority and non-minority job applicant being invited for an interview (binary regressor). i.e difference in probabilities that Y=1 when X switches from 0 to 1

2

We are mainly interested in estimating B1 in the equation above. As stated above, it's the difference in probabilities that Y=1 when X changes from 0 to 1. Hence, B1 is the average change in probability of a job applicant being invited for an interview when the job applicant is from an ethnic minority, compared to when they aren't.

3

The random sampling assumption is that samples were collected using simple random samples of the population. Each member of the population would have an equal chance of being included in a sample. This ensures that the samples are independently and identically distributed. If the samples are collected from the same population this means that the (Xi, Yi) are identically distributed for all i= 1, . . . , n.

The conditional independence assumption is that the error term has no effect on the explanatory variables, i.e the explanatory variables are not exogenous. In most real world cases the assumption E (U|X) = 0 isn't fulfilled. It is only fulfilled in randomized experiments as if X is assigned randomly, all other individual characteristics (the characteristics that make up U) are distributed independently of X. With regards to the equation above, it is the assumption that E(u|minority) = 0. This assumes that the mean of the error term (u) is zero. For the ethnic minority variable, E(u|minority) = 0 implies that other factors which could affect the probability of being invited for an interview are not correlated with a person being an ethnic minority or not.

These assumptions are not fulfilled in this case.

For the random sampling assumption, the sample is collected from a website where jobs are advertised. This is not an accurate representation of the entire population of those who sent job applications to firms. There are many other ways of sending job applications to firms e.g. sending by post, walking into the firm's office and hand delivering the job application, the firm's own website, other job advertising websites etc. Therefore, each subject from the population (all those who sent job applications to firms) does not have an equal chance of being selected. Hence, the random sampling assumption is not fulfilled. The sample is not representative of the entire population

For the conditional independence assumption, There are likely to be many confounders in u that are both correlated with a person being an ethnic minority or not, while also having an effect on the probability of that person getting a job interview. This includes the data we have collected for yrs_educ (years of education). Corr(ethinc minority, yrs_educ) < 0, as people from ethnic minorities tend to on average have less years of education than those from non-ethnic minorities. The probability of getting an interview also depends on yrs_educ, so we can confidently say that the conditional independence assumption is not fulfilled in this case

With the conditional independence assumption and the random sampling assumption not being fulfilled, 2 of our assumptions for least squares are broken. Conditional independence being broken means we can't conclude that B1 is an unbiased estimator for the probability of a person being invited for an interview. The data not being a random sampling distribution also means that we can't conclude it vies the sampling distribution of B1. Therefore, the regression equation showing how the probability of a person getting an interview changes if a person is from an ethnic minority is more likely to be inaccurate, with these assumptions not fulfilled.

4

The conditional independence assumption is still not fulfilled, as E(u|minority) does not equal 0, due to the confounders in u which affect both the likelihood that a person will get a job interview, and the likelihood that a person is from an ethnic minority. Hence, we still can't conclude that B1 is unbiased. The conditional independence assumption is extremely difficult to fulfil in real world examples such as this. However, being female or not and years of education are two variables that plausibly effect the likelihood of getting a job interview. Hence, now B1 is likely to be less biased than before, as omitted variable bias from these variables not being included is avoided.

To reduce omitted variable bias, we would like to include information such as for example: years of work experience, whether the person has a degree or not, CV structure etc. Essentially we would want to include any other information that could affect the explanatory variable(s) in our model. We can include these in our equation as control variables where suitable. This is so we don't attribute to the existing explanatory variables the effects of these omitted variables on Y.

B

```
library(stargazer)
library(tidyverse)
library(haven)
library(dplyr)
library(ggplot2)
library(ggthemes)
```

1

```
data <- read_dta("data_project_autumn2022.dta")
data$canadian <- ifelse(data$ethnicity == "Canada", 1, 0)
data <- as.data.frame(data)
data %>%
  select("occupation_type", "callback", "secondcallback", "female", "baquality",
         "extracurricular", "canadian") %>%
  stargazer(type= "text", summary.stat=c( "n","mean", "sd", "median", "min", "max"))
```

```
##
## ====================================================
## Statistic         N   Mean  St. Dev. Median Min Max
## ---------------------------------------------------
## callback        5,191 0.108  0.310     0     0   1
## secondcallback  5,191 0.025  0.155     0     0   1
## female          5,191 0.511  0.500     1     0   1
## baquality       5,191 0.461  0.499     0     0   1
## extracurricular 5,191 0.605  0.489     1     0   1
## canadian        5,191 0.285  0.452     0     0   1
## ---------------------------------------------------
```

Note that there are no missing observations in the dataset.

The probability of a randomly selected applicant getting a call back from the employer is 0.1080

The probability of a randomly selected applicant getting a secondcallback after being interviewed is 0.025

The probabiilty of a randomly selected applicant being female is 0.511.

The probability of a randomly selected applicant having a BA degree from a prestigious universtiy is 0.461.

The probability of a randomly selected person having extracurricular activities listed on their CV is 0.285.

2

```
data %>%
  count(nrow(data), ethnicity) %>%
  group_by(ethnicity) %>%
  mutate(prop = n/nrow(data))
```

```
## # A tibble: 5 x 4
## # Groups:   ethnicity [5]
##   `nrow(data)` ethnicity     n   prop
##          <int> <chr>     <int>  <dbl>
## 1         5191 Canada     1481 0.285
## 2         5191 Chinese    1357 0.261
## 3         5191 Chn-Cdn     504 0.0971
## 4         5191 Indian     1342 0.259
## 5         5191 Pakistani   507 0.0977
```

3

```
crosstabs<- xtabs(~type + ethnicity ,data=data)
prop.table(crosstabs, 1)
```

```
##     ethnicity
## type    Canada   Chinese   Chn-Cdn    Indian Pakistani
##    0 1.0000000 0.0000000 0.0000000 0.0000000 0.0000000
##    1 0.0000000 0.3395245 0.1597325 0.3291233 0.1716196
##    2 0.0000000 0.4036697 0.1166448 0.3709043 0.1087811
##    3 0.0000000 0.3885350 0.1184713 0.3910828 0.1019108
##    4 0.0000000 0.3517157 0.1311275 0.3786765 0.1384804
```

The result tells us that the quality of randomisation in the experiment is very poor. From the table, 100% of Canadians got the type 0 CV, meaning 0% got type 1,2,3 or 4. It is incredibly unlikely that this would occur if we randomly assigned a type of CV to each ethnicity, which is what we should do if we want our sample values to be independently and identically distributed, like is described in the experiment brief. Furthermore, 0% of all other ethnicities got type 0 CV, which is even more unlikely if type of CV was randomly assigned.

Other than this, the proportions of the other types of CV's apart from type 0 in ethnicities apart from Canada seems plausible and random, as they are all quite equal proportions for each ethnicity and type, relative to the sample proportion of each ethnicity.

Note that this problem with Canada and type 0 CV is hence likely to be an error.

4

```r
t_callback<- t.test(callback~canadian,data,alternative= "two.sided",mu=0,
                    conf.level= 0.95)
t_secondcallback<- t.test(secondcallback~canadian,data,alternative= "two.sided",mu=0,
                          conf.level= 0.95)
t_baquality<- t.test(baquality~canadian,data,alternative= "two.sided",mu=0,
                     conf.level= 0.95)
t_female<- t.test(female~canadian,data,alternative= "two.sided",mu=0, conf.level= 0.95)
t_extracurricular<- t.test(extracurricular~canadian,data,alternative= "two.sided",mu=0,
                           conf.level= 0.95)

variables<-c("callback","secondcallback","baquality","female","extracurricular")
meannoncanadian<- c(t_callback[["estimate"]][["mean in group 0"]],
                    t_secondcallback[["estimate"]][["mean in group 0"]],
                    t_baquality[["estimate"]][["mean in group 0"]],
                    t_female[["estimate"]][["mean in group 0"]],
                    t_extracurricular[["estimate"]][["mean in group 0"]])
meancanadian<- c(t_callback[["estimate"]][["mean in group 1"]],
                 t_secondcallback[["estimate"]][["mean in group 1"]],
                 t_baquality[["estimate"]][["mean in group 1"]],
                 t_female[["estimate"]][["mean in group 1"]],
                 t_extracurricular[["estimate"]][["mean in group 1"]])
meandiff<- meancanadian-meannoncanadian
Pvalue<-c(t_callback[["p.value"]],t_secondcallback[["p.value"]],
          t_baquality[["p.value"]],t_female[["p.value"]],t_extracurricular[["p.value"]])
df<- data.frame( variables,meancanadian, meannoncanadian, meandiff, Pvalue)
df
```

```
##         variables meancanadian meannoncanadian     meandiff       Pvalue
## 1        callback   0.14787306      0.09191375  0.055959312 7.611869e-08
## 2  secondcallback   0.04051317      0.01832884  0.022184326 7.221456e-05
## 3       baquality   0.49291020      0.44878706  0.044123134 4.075348e-03
## 4          female   0.52059419      0.50673854  0.013855649 3.672000e-01
## 5 extracurricular   0.60364619      0.60566038 -0.002014192 8.934393e-01
```

When testing the hypothesis that the mean differences between Canadians and non-Canadians for the variables Callback, second-callback, baquality, female and extracurricular at a 95% confidence interval t-test (two sided). We can conclude with significant evidence that:

Call-back: Since (7,61eˆ-8 < 0.05) we reject H0 and conclude that there is a difference in the means, so the mean probability of a callback for Canadians is different to the probability number of a callback for non-Canadians.

Second call back: Since (7,22eˆ-5 < 0.05), we Reject H0 and conclude with significant evidence that the means differ i.e. the mean probability of a second call-back for Canadians is different to the mean probability of a second call-back for non-Canadians.

Baquality-: Since (4.08eˆ-5 < 0.05) Reject H0 and conclude with significant evidence that the means differ. i.e., the mean probability of a Canadian having a BA degrees from prestigious universities is different to that of no-Canadians.

Female: Since (0.37>0.05), we fail to reject H0 and hence fail to conclude with significant evidence that there is a difference in the mean probablity of a Canadian being female and a non-Canadian being female

Extracurricular: Since (0.89>0.05), we fail to reject H0 and hence fail to conclude with significant evidence that there is a difference in the mean probability of a Canadian having extracurricular activities listed on their CV and non-Canadians having extracurricular activities listed on their CV.

5

```
regcallCan <- lm(callback ~ canadian, data = data)
stargazer(regcallCan, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              callback
## -----------------------------------------------
## canadian                     0.056***
##                               (0.010)
##
## Constant                     0.092***
##                               (0.005)
##
## -----------------------------------------------
## Observations                   5,191
## R2                             0.007
## Adjusted R2                    0.006
## Residual Std. Error      0.309 (df = 5189)
## F Statistic          34.657*** (df = 1; 5189)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The value of B1 is 0.056. This cannot be interpreted as a slope as there are dummy variables on the left and right. Instead, it means that if a person is Canadian, the probability of them getting a callback on average increases by 5.6 percentage points. B1 is therefore the difference in probabilities that a person gets a callback when they are canadian vs when they are not canadian.

This corresponds to the t-test above in part 4 where we rejected H0 and concluded at a 95% confidence level that the mean number of callbacks for Canadians differs to that of non-Canadians. i.e. Canadians have a greater chance of being called back.

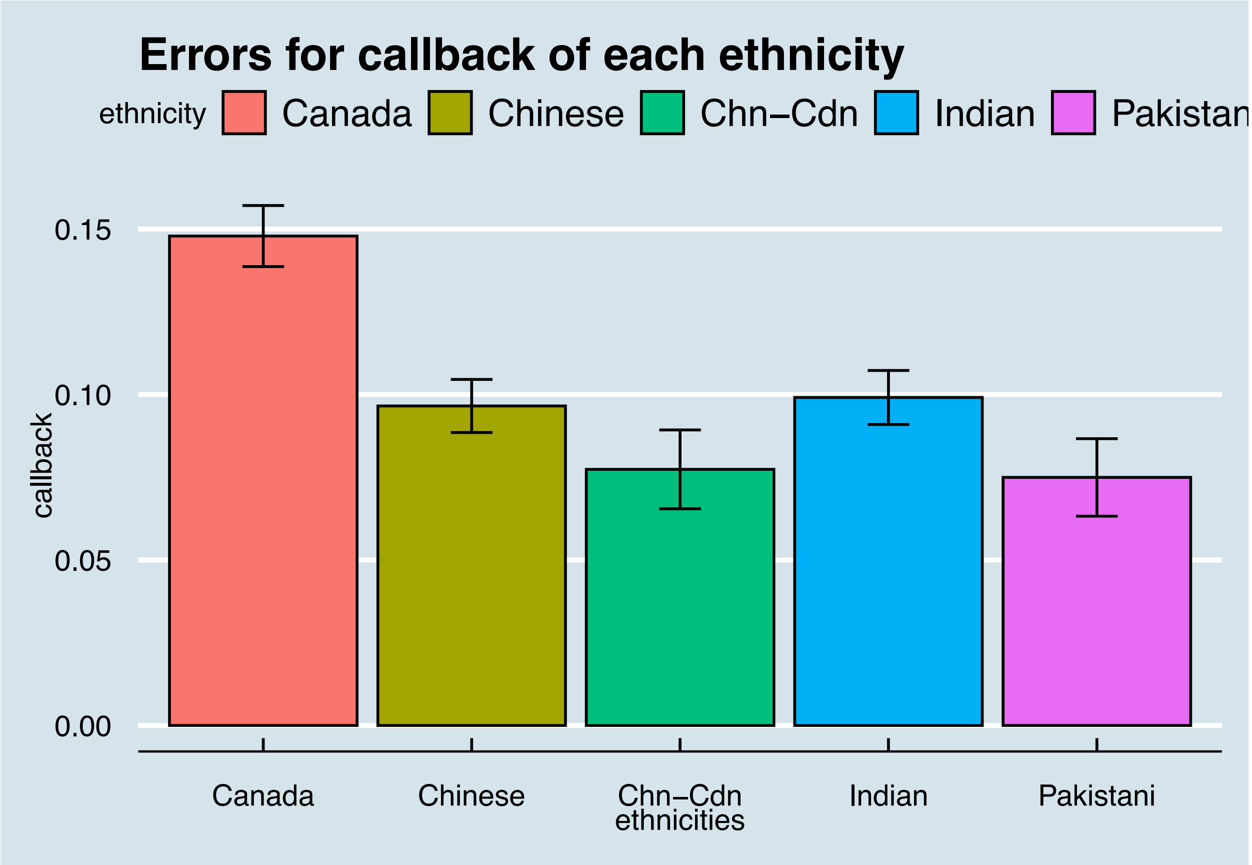
B1 is significant at the 1% level, meaning there's a 99% likelihood that B1 is different from 0. We are 99% sure that being canadian does have an effect on the likelihood of getting a callback

The constant value (BO) is 0.092. This corresponds to the probability that a non-canadian person from the sample on average gets a callback. This is a 9.2% likelihood.

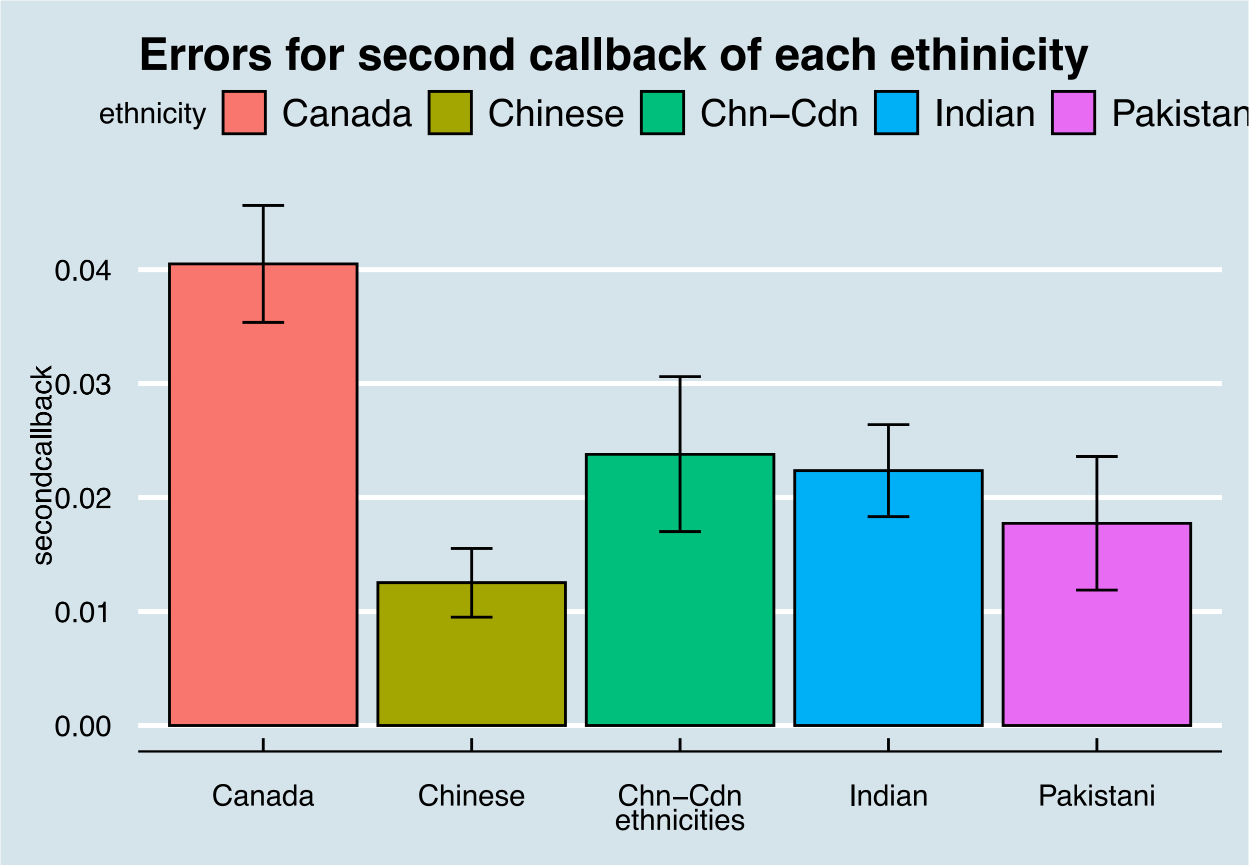
B0 is also significant at the 1% level, meaning there's a 99% likelihood that B0 (the constant) is different from 0. We are 99% sure that the probability of a non-canadian person from the population of job applicants getting a callback is greater than 0. This is consistent with mean value of non-canadians getting a callback in Q4 being 0.09, which is different from 0.

6

```
clean_data <- data %>%
  group_by(ethnicity) %>%
  summarize(mean_callback = mean(callback), sd_callback = sd(callback),
            count = n(), se_callback = (sd_callback / sqrt(count)), mean_secondcallback =
              mean(secondcallback), sd_secondcallback = sd(secondcallback),
            se_secondcallback = (sd_secondcallback/sqrt(count)))

lab1plot <- ggplot(clean_data, aes(x = ethnicity, y = mean_callback, fill = ethnicity)) +
  geom_bar(stat = 'identity', colour = "black",
           position = position_dodge()) +
  theme_economist() +

  geom_errorbar(aes(ymin = mean_callback - se_callback,
                    ymax = mean_callback + se_callback), width = .2) +
  labs (x = "ethnicities", y = "callback") +
  ggtitle("Errors for callback of each ethnicity")
lab1plot
```



```
lab2plot <- ggplot(clean_data, aes(x = ethnicity, y = mean_secondcallback,
                                   fill = ethnicity)) +
  geom_bar(stat = 'identity', colour = "black",
           position = position_dodge()) +
  theme_economist() +

  geom_errorbar(aes(ymin = mean_secondcallback - se_secondcallback,
                    ymax = mean_secondcallback + se_secondcallback), width = .2) +
  labs (x = "ethnicities", y = "secondcallback") +
  ggtitle("Errors for second callback of each ethinicity")

lab2plot
```



7

```
regcanbaqex <- lm(canadian ~ baquality + extracurricular, data=data)
stargazer(regcanbaqex, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              canadian
## -----------------------------------------------
## baquality                    0.036***
##                               (0.013)
##
## extracurricular               -0.002
##                               (0.013)
##
## Constant                     0.270***
##                               (0.012)
##
## -----------------------------------------------
## Observations                   5,191
## R2                             0.002
## Adjusted R2                    0.001
## Residual Std. Error      0.451 (df = 5188)
## F Statistic           4.161** (df = 2; 5188)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The value of B0 (the constant) is equal to 0.27. This means that a person with a CV that doesn't list extracurricular activities, and who doesn't have a BA degree from a prestigious university has on average a 27% chance of being canadian

The value of B1 (corresponding to variable baquality) is equal to 0.036. This means that the probability of a person being canadian increases on average by 3.6 percentage points when we change them from not having to having a BA degree from a prestigious university, controlling for them having extracurricular activities listed on their CV or not

The value of B2 (corresponding to variable extracurricular) is -0.002. This means that the probability of a person being canadian, controlling for them holding a BA degree from a prestigious university or not, falls on average by 0.2 percentage points if they list extracurricular activities on their CV.

If the experiment did as it said and randomised the ethnicity and educational credentials for each job application, the probability of a person being Canadian should not depend on whether they have a BA degree from a prestigious university or whether they have extracurricular activities listed on their CV. B1 should equal 0. Here, the t-statistic for baquality (B1) is statistically significant at the 99% level of confidence, so we can conclude that the average probability of being canadian given a change in the baquality dummy variable from 0 to 1, is different from 0. Hence, our experiment is poorly designed. This is consistent with our conclusion in Q3 above

8

```r
regcalcanfem <- lm(secondcallback ~ canadian + female, data=data)
stargazer(regcalcanfem, type="text")
```

```
## 
## ===============================================
##                          Dependent variable:    
##                      ---------------------------
##                            secondcallback       
## -----------------------------------------------
## canadian                     0.022***           
##                               (0.005)           
##                                                 
## female                         0.003            
##                               (0.004)           
##                                                 
## Constant                     0.017***           
##                               (0.003)           
##                                                 
## -----------------------------------------------
## Observations                   5,191            
## R2                             0.004            
## Adjusted R2                    0.004            
## Residual Std. Error      0.155 (df = 5188)       
## F Statistic          11.046*** (df = 2; 5188)   
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The value of B0 (the constant) is 0.017. This means that the probability of getting a callback of a person who is non-canadian and non-female is on average 1.7%. This is statistically significant at the 1% level, meaning we are 99% sure that the value of B0 is different from 0 i.e the probability of a person who is non-canadian and non-female getting a second call back is greater than 0

The value of B1 (corresponding to the variable canadian) is 0.022. This means that if a person is canadian, the probability of them gettting a second call back, controlling for being a female or not, increases on average by 2.2 percentage points. This is statistically significant at the 1% level, meaning there is a likelihood less than 1% that this value B1 is equal to 0, which therefore means that being canadian does have an effect of getting a secondcallback, controlling for being a female or not.

The value of B2 (corresponding to the variable female) is 0.003. This means that if a person is female, controlling for being canadian or not the probability of them getting a call back increases on average by 0.3 percentage points. This is not statistically significantly, meaning we can conclude that being a female does not have an effect on the likelihood getting a second call back, controlling for being canadian or not.

The estimates are likely to be biased as there are likely to be other factors that are both correlated to our explanatory variables and affect the dependent variable. This means that the expected value of the explanatory variables given the error term won't be 0, i.e E(u|X) won't equal 0. It is extremely difficult to account for all of the factors that determine whether a person gets a callback(s) for an interview, that are unrelated to any of the explanatory variables already listed. For example, a person knowing the manager/existing employees of a company may have an effect on their likelihood of getting a callback, and this would also effect the likelihood of them being Canadian, as the population in this experiment is applying to companies in Toronto and Montreal, and Canadians are more likely to know each other.

C

```r
set.seed(8)
rep <- 500
n <- 100
```

1

```r
ybar <- numeric(rep)

for(i in 1:rep){
  B1 <- 2
  X <- rnorm(n, mean=100, sd=15)
  u <- rnorm(n, mean=0, sd=15)
  y <- B1*X+u
  ybar[i] <- mean(y)
  rm(y)
}
```

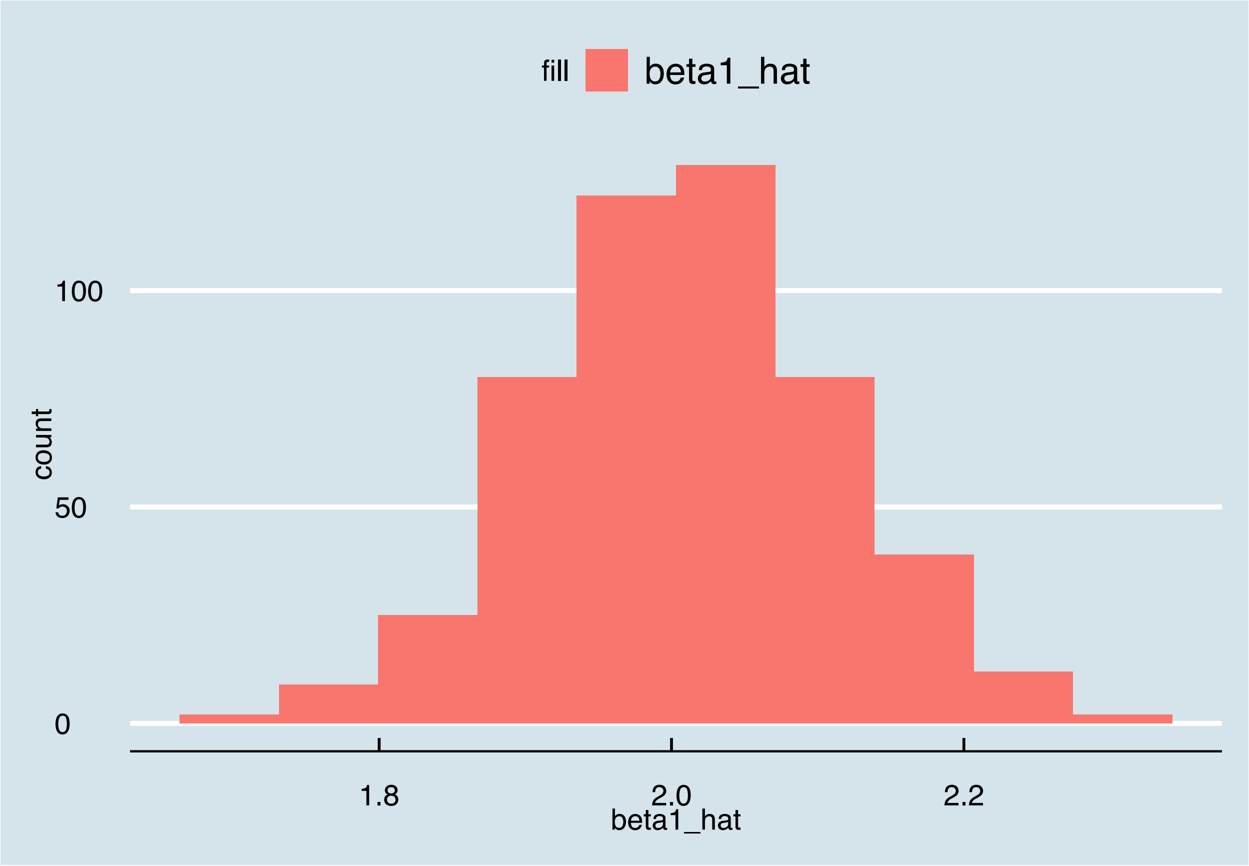
2

```
set.seed(8)
beta1_hat <-  numeric(rep)
for (i in 1:rep){
  B1 <- 2
  X <- rnorm(n, mean=100 , sd=15 )
  u <- rnorm(n, mean= 0, sd=15)
  y <- B1 * X + u
  df <- data.frame(X,y)
  reg <- lm(y~X, data=df)

  beta1_hat[i] <- coef(reg)[2]
}
beta1_hatmean <- mean(beta1_hat)
beta1_hatmean
```

```
## [1] 2.010281
```

3

```
beta <- data.frame(beta1_hat)
ggplot(data=beta, aes(x = beta1_hat, fill = "beta1_hat") ) + geom_histogram(bins=10) +
  theme_economist()
```



We obtain different estimates in different samples because each sample is a random sample, taken from a normal distribution and is identically and independently distributed. We get different values of Y for each sample, and hence we get different values for B1 in the different samples.

4

1

```
ybar <- numeric(rep)

for(i in 1:rep){
  B1 <- 2
  X <- rnorm(n, mean=100, sd=15)
  u <- rnorm(n, mean=0, sd=30)
  y <- B1*X+u
  ybar[i] <- mean(y)
  rm(y)
}
```

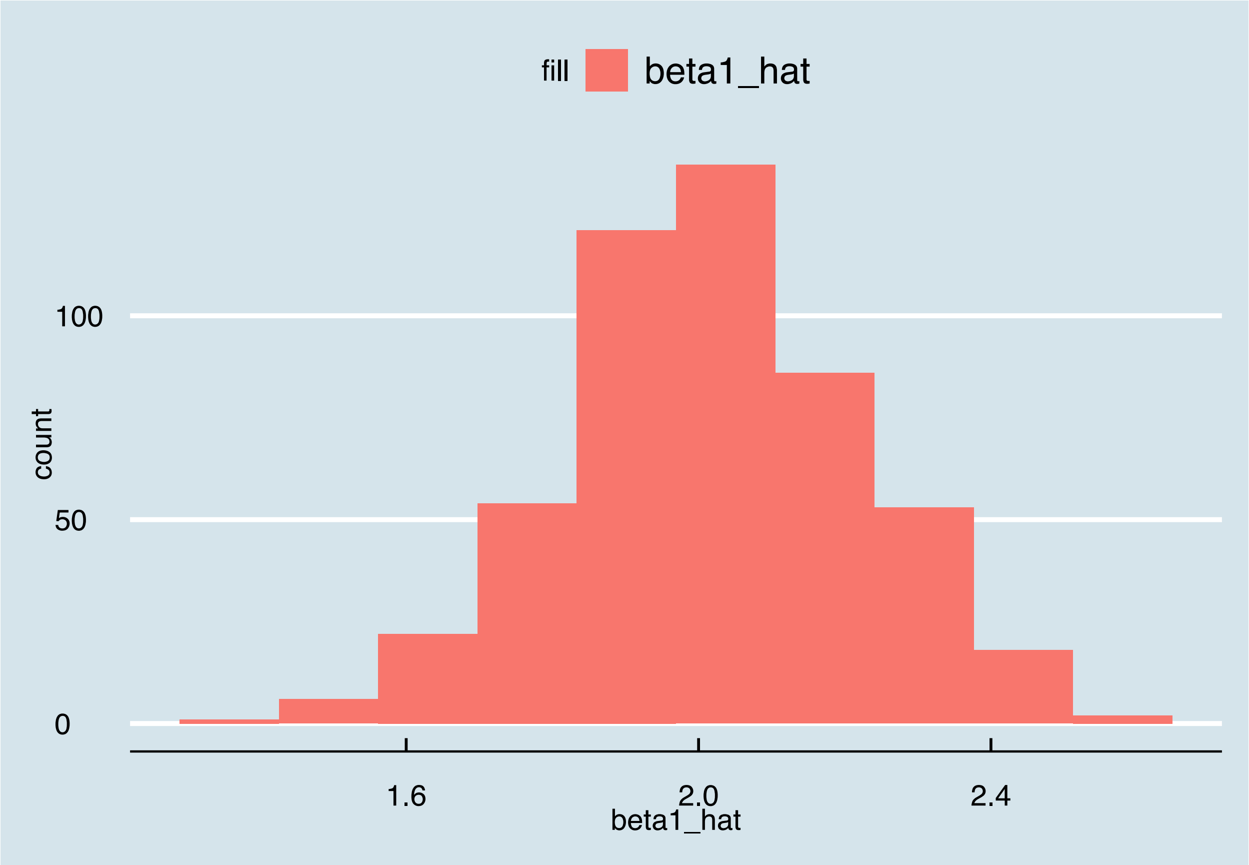
2

```
set.seed(8)
beta1_hat <-  numeric(rep)
for (i in 1:rep){
  B1 <- 2
  X <- rnorm(n, mean=100 , sd=15 )
  u <- rnorm(n, mean= 0, sd=30)
  y <- B1 * X + u
  df <- data.frame(X,y)
  reg <- lm(y~X, data=df)

  beta1_hat[i] <- coef(reg)[2]
}
beta1_hatmean <- mean(beta1_hat)
beta1_hatmean
```

```
## [1] 2.020561
```

3

```
beta <- data.frame(beta1_hat)
ggplot(data=beta, aes(x = beta1_hat, fill = "beta1_hat") ) + geom_histogram(bins=10) +
  theme_economist()
```



The distribution of B1 hat now looks different because the standard deviation, and equivalently variance of u (the error term) has increased. This means that on average, the individual sample statistics of B1 hat are more spread from the mean. This means the new distribution is now wider, and thus the predictions for y are less precise.