

# Unit 09 Homework

w203: Statistics for Data Science

March 5, 2025

```
library(tidyverse)

library(patchwork)
library(stargazer)

library(sandwich)
library(lmtest)

library(knitr)
library(kableExtra)
```

Here is our intention for the homework:

1. By simulating the data in **Question 1** you should be able to observe everything that is happening in a series of models.
2. Then, by reading, replicating, and writing about the results of a very well done analysis in **Question 2, Part (C)** you have the ability to estimate regressions with a known goal.
3. If you have time, you might be interested in working on **Question 2, Part (B)** but we have made this optional.

In this homework, there are known values that you are trying to reproduce. As we move forward into future weeks, we will no longer have a known value that we are trying to replicate. Instead, we are going to have to make choices that can be justified, but without having an “answer sheet” that we can compare to.

To complete this homework, please write code that replaces the 'fill this in' tags. Please store your results in the objects that we have created for you, this will help as graders to evaluate what you've written.

## Contents

<b>1 Simulated Data</b>	<b>3</b>
1.1 Plot of outcome data . . . . .	3
1.2 Evaluate large sample assumptions . . . . .	5
1.3 Estimate four models . . . . .	5
1.4 Evaluate model performance . . . . .	5
1.5 Consider the first model . . . . .	6
1.6 Best of the best . . . . .	6

1.7	Conduct two tests about $x_2$ .	6
1.7.1	t-test	6
1.7.2	f-test	7
1.7.3	Reason about tests	7
<b>2</b>	<b>Real-World Data</b>	<b>8</b>
2.1	Read the data	8
2.2	(Optional). Conduct an F-test	8
2.2.1	(Optional) State the null	9
2.2.2	(Optional) Why would you reject?	9
2.2.3	(Optional) Conduct an f-test	10
2.2.4	(Optional) What do you conclude?	10
2.2.5	(Optional) Interpret your conclusions	11
2.3	Reproduce Table 4	11
2.3.1	Evaluate the large sample assumptions	11
2.3.2	Conduct these regressions	11
2.3.3	Do covariates improve model fit?	13
2.3.4	Robust standard errors	14
2.3.5	State your null hypotheses	15
2.3.6	Which tests reject the null hypothesis	15
2.3.7	Interpret the effect of being sent a reminder	16
2.3.8	Interpret the coefficient on <b>age</b>	16
2.3.9	Interpret the coefficient on <b>highschool_completed</b>	16
2.3.10	Print a whole table.	16

# 1 Simulated Data

For this question, we are going to create data, and then estimate models on this simulated data. This allows us to effectively *know* the population parameters that we are trying to estimate. Consequently, we can reason about how well our models are doing.

```
create_homoskedastic_data <- function(n = 100) {  
  
  d <- data.frame(id = 1:n) %>%  
    mutate(  
      x1 = runif(n=n, min=0, max=10),  
      x2 = rnorm(n=n, mean=10, sd=2),  
      x3 = rnorm(n=n, mean=0, sd=2),  
      y = .5 + 1*x1 + 0*x2 + .25*x3^2 + rnorm(n=n, mean=0, sd=1)  
    )  
  
  return(d)  
}
```

```
d <- create_homoskedastic_data(n=100)
```

## 1.1 Plot of outcome data

Produce a plot of the distribution of the **outcome data**. This could be a histogram, a boxplot, a density plot, or whatever you think best communicates the distribution of the data. What do you note about this distribution?

```
min_out = round(min(d$y),2)  
max_out = round(max(d$y),2)  
  
outcome_histogram <- d %>%  
  ggplot(aes(y)) +  
  geom_histogram(aes(y = ..density..),  
    fill = "#0099F8") +  
  geom_density(color = "#00558A", fill = "#71C9FF", alpha = 0.6) +  
  xlim(trunc(min_out)-1, trunc(max_out) + 1) +  
  geom_vline(xintercept = min_out, linetype = "dashed", color = "#001F82") +  
  annotate(geom='text', x= min_out + 1, y=0.2, label= "Min",  
    parse=TRUE, color="#001F82", size = 4) +  
  annotate(geom='text', x= min_out + 1, y=0.19, label= min_out,  
    parse=TRUE, color="#001F82", size = 4) +  
  annotate(geom='text', x= max_out - 1, y=0.19, label= max_out,  
    parse=TRUE, color="#001F82", size = 4) +  
  annotate(geom='text', x= max_out - 1, y=0.2, label= "Max",  
    parse=TRUE, color="#001F82", size = 4) +  
  geom_vline(xintercept = max_out, linetype = "dashed", color = "#001F82") +  
  labs(  
    title = "Outcome (y) Distribution",  
    x = "Outcome (y)",  
    y = "Density"  
  ) +  
  theme_classic() +
```

```

theme(
  plot.title = element_text(color = "#0099F8",
                             size = 17,
                             face = "bold"),
  axis.title = element_text(color = "#969696",
                             size = 10,
                             face = "bold"),
  axis.text = element_text(color = "#969696", size = 10),
  axis.line = element_line(color = "#969696")
)

```

outcome\_histogram

```

## Warning: The dot-dot notation ('..density..') was deprecated in ggplot2 3.4.0.
## i Please use 'after_stat(density)' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.

```

```

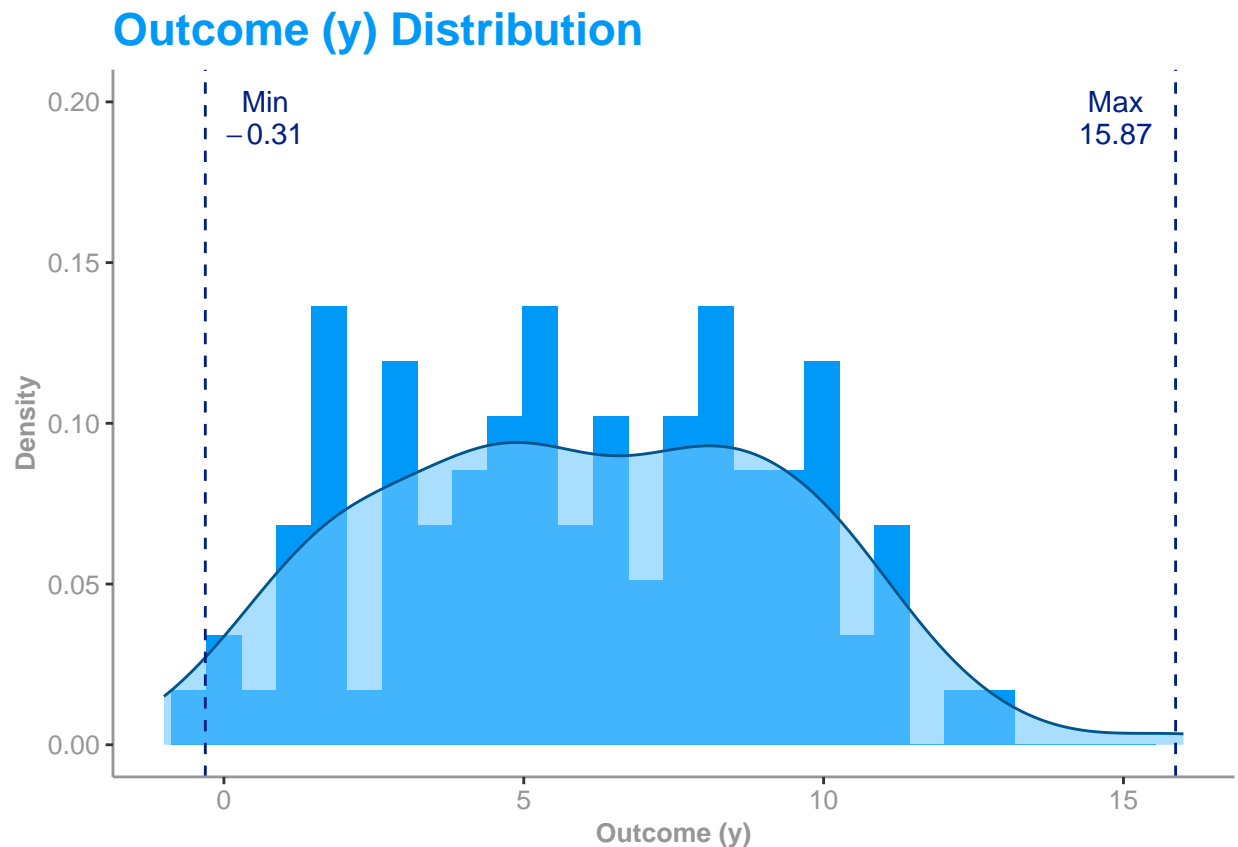
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```

```

## Warning: Removed 2 rows containing missing values or values outside the scale range
## ('geom_bar()').

```



“This distribution is randomly generated so it presents a different shape every time the code runs, sometimes it seems ‘bell-shaped’, other times it looks as a ‘bimodal’ distribution, sometimes with longer tails (especially on the right side). But there are some similitudes, for example the minimum value for Y is usually somewhere around zero, and the even when the maximum can have more variation when analyzing the density we can see that usually most of the values fall between 0 and 15.”

## 1.2 Evaluate large sample assumptions

Are the assumptions of the large-sample model met so that you can use an OLS regression to produce consistent estimates?

“Since we have  $n = 100$  we could argue about having the required sample size but it is indeed in the very limit of what we can consider as ‘large’, about the assumptions, we have two of them basically, IID data, which is the case since every datapoint is randomly sampled independently from each other and from the exact same distributions. About having a unique BLP, we notice that there’s no infinite variance (or heavy tails), and because of the way we’re randomly generating our X’s from different distributions, there’s no X that could be written as a linear combination of the others. So I would say that the assumptions are satisfied indeed.”

## 1.3 Estimate four models

Estimate four models, called `model_1`, `model_2`, `model_3` and `model_4` that have the following form:

$$Y = \beta_0 + \beta_1 x_1 + 0x_2 + \beta_3 x_3 + \epsilon \quad (1)$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \quad (2)$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^2 + \epsilon \quad (3)$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3^2 + \epsilon \quad (4)$$

```
# If you want to read about specifying statistical models, you can read
# here: https://cran.r-project.org/doc/manuals/R-intro.html#Formulae-for-statistical-models
# note, using the I() function is preferred over using poly()

model_1 <- lm(y~x1+x3, data = d)
model_2 <- lm(y~x1+x2+x3, data = d)
model_3 <- lm(y~x1+x2+I(x3^2), data = d)
model_4 <- lm(y~x1+x2+x3+I(x3^2), data = d)
```

## 1.4 Evaluate model performance

Recall that *Foundations of Agnostic Statistics* used **MSE** as the evaluative criteria for population models. Use the plug-in analogue, the **Mean Squared Residual**, **MSR** in this sample to evaluate how well each of these models does at predicting outcomes.

```
calculate_msr <- function(model) {
  # This function takes a model, and uses the `resid` function
  # together with the definition of the msr to produce
  # the MEAN of the squared residuals
```

```
msr <- mean(resid(model)^2)
return(msr)
}
```

```
model_1_msr <- calculate_msr(model_1)
model_2_msr <- calculate_msr(model_2)
model_3_msr <- calculate_msr(model_3)
model_4_msr <- calculate_msr(model_4)
```

Model	MSR
Model 1	2.598015
Model 2	2.597963
Model 3	1.065054
Model 4	1.059620

## 1.5 Consider the first model

Consider, for a moment, only the first model. Is it possible to select coefficients in this model that would produce a lower mean squared residual? Why or why not?

‘It’s not possible to have a lower mean squared residual with different coefficients, because the way the coefficients are selected is specifically by minimizing the mean squared residuals.’

## 1.6 Best of the best

Which of these models does the best job, in terms of mean squared residuals, at estimating the population coefficients?

‘The model that has lowest mean squared residuals is the fourth model. Even when we know that the original outcome  $y$  only uses  $x_3^2$ , just as model\_3, we need to remember that everytime we add a new variable we can only get a lower or equal MSR (or what’s equivalent, a greater or equal  $R^2$ ), since model\_4 has all the variables model\_3 has plus an extra one, model\_4 will always have at most the same MSR as model\_3 or less.’

## 1.7 Conduct two tests about $x_2$ .

### 1.7.1 t-test

First, using model\_2 that you have estimated: conduct a wald-test (i.e. a t-test) for the coefficient  $\beta_2$ . What do you conclude from this sample about the relationship between  $x_2$  and  $y$ ?

```
coeftest(model_2,
          vcov= vcov(model_2)) # Since data is homoskedastic
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.2913475  0.9073185  1.4233  0.15790
## x1          1.0003483  0.0562959 17.7695 < 2e-16 ***
## x2          0.0038769  0.0879003  0.0441  0.96491
## x3         -0.2829848  0.0880341 -3.2145  0.00178 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

‘By looking at the summary of the model 2 we can see that the coefficient  $\beta_2$  generates a t-value=0.0441 with an associated p-value=0.9649, so even when there’s a value for that coefficient, there’s not enough evidence to suggest that we can reject our  $H_0 : \beta_2 = 0$ .’

### 1.7.2 f-test

Is there any evidence that the additional parameter that you have estimated in `model_2` makes make this second model more fully represent the true population? Conduct an F-test with the null hypothesis that `model_1` is the correct population model, and evaluate whether you should reject the null to instead conclude that `model_2` is more appropriate.

```
anova(model_2, model_1, test = 'F')
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x1 + x3
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      96 259.8
## 2      97 259.8 -1 -0.0052644 0.0019 0.9649
```

‘There’s no evidence that we should reject our  $H_0 : model_1$ , since our F-statistic= 0.0019 and its p-value = 0.9649, so it makes no sense to use the larger model and we should use the simpler reduced model instead (`model_1`)’

### 1.7.3 Reason about tests

In your own words, explain why the p-values for the tests that you have conducted in parts (a) and (b) are the same. Are these tests merely different ways of asking the same question of a model?

‘In this specific case, it is indeed a different way to ask the same question, in the case of the t-test we’re asking if the coefficient  $\beta_2$  seems to be different than zero (adding some weight to  $x_2$  in our model), and in the second test (the F-test) we’re basically asking if it makes sense to include  $x_2$  (multiplied by certain coefficient  $\beta_2$ ) to our model. So, saying that there’s no evidence to conclude that  $B_2 \neq 0$  is equivalent to say that there’s no evidence that adding  $x_2$  will do any good to our model. It’s not always like that, since an F-Test is not always about including just one single extra variable to our large model, there can be more of them.’

## 2 Real-World Data

“Can timely reminders *nudge* people toward increased savings?”

Dean Karlan, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinnman published a paper in 2016 examining just this question. In this research, the authors recruited people living in Peru, Bolivia, and the Philippines to be a part of an experiment. Among those recruited, a randomly selected subset were sent SMS messages while others were not sent these messages. The authors compare savings rates between these two groups using OLS regressions.

Please, take the time to read the following sections of the paper. We are asking you to read this to provide context and understanding for the data analysis task. Please, read briefly (take no more than 15-20 minutes for this reading).

1. The *Abstract*
2. The first five paragraphs of the *Introduction* (the last paragraph to read begins with, “Although the full pattern of our empirical results suggests...” )
3. Section 2: *Experimental Design* so you have a sense for where and how these experiments were conducted
4. Table 2(a), 2(b), and 2(c) so you have a sense for what the SMS messages said to participants.

The core results from this study are reported in Table 4. You can read this now, or when you are doing the data work to reproduce parts of Table 4 later in this homework.

### 2.1 Read the data

Read in the data using the following code:

- This code is using the `haven` package, and then the `read_dta` function within that package to load data that is stored in a proprietary STATA format.
- For a description of the meaning of these variables, you can see the documentation in this repository.

```
d <- haven::read_dta(file = './karlan_data/analysis_dataallcountries.dta')
glimpse(d)
```

### 2.2 (Optional). Conduct an F-test

**(This section about conducting an F-test is optional! The next section is required!)**

One of the requirements of a data science experiment is that treatment be randomly assigned to experimental units. One method of assessing whether treatment was randomly assigned is to try and predict the treatment assignment. Here’s the intuition: *it should not be possible to predict something random*.

The specifics of the testing method utilize an F-test. Here is how:

- The data scientist first estimates a model that regresses treatment using only a regression intercept,  $rem\_any \sim \beta_0 + \epsilon_{short}$ . In `lm()`, you can estimate this by writing `lm(rem_any ~ 1)`.
- Then, the data scientist estimates a model that regresses treatment using all data available on hand,  $rem\_any \sim \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon_{long}$ , where  $x_1 + \dots$  index all the additional variables to be tested.

To test whether the long model has explained more of the variance in *rem\_any* than the short model, the data scientist then conducts an F-test for the long- vs. short-models.

```
# Taking out variables with NA's and creating a new df
na_count <- sapply(d, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
na_count$row <- seq.int(nrow(na_count))
v <- na_count$row[na_count$na_count>0]
d1 <- d[,c(-1,-v)]

# Creating intercept-model and saturated model
model_short <- lm(rem_any~1, data=d1)
model_long <- lm(rem_any~., data=d1)

#Comparing using F-test
ftest<- anova(model_short, model_long, test = 'F')
ftest
```

```
## Analysis of Variance Table
##
## Model 1: rem_any ~ 1
## Model 2: rem_any ~ depart + quant_saved + age + saved_formal + female +
##   married + highschool_completed + reached_b4goal + bolivia +
##   provincia + wealthy + puzzle_ica + foto + rem_motive + peru +
##   branch + marketer + joint + joint_single + dc + highint +
##   rewardint + inc_7d + hyperbolic + saved_asmuch + spent_b4isaved +
##   philippines + country + late_rem_any + log_quant_saved +
##   rem_any_peru + rem_any_phil + rem_any_boli + missing_female +
##   missing_age + missing_highschool_completed + missing_married +
##   missing_saved_formal + missing_inc_7d + missing_wealthy +
##   missing_hyperbolic + missing_saved_asmuch + missing_spent_b4isaved +
##   missing_number_account + ivhquant_saved + logalt_quant_saved +
##   gain_rem + loss_rem + noincentive + rem_no_motive + masterid
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1  13559 3221.2
## 2  13518   0.0 41    3221.2 1.287e+28 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 2.2.1 (Optional) State the null

What is the null hypothesis for this F-test between a short- and long-model?

‘ $H_0$ : The long model is explaining things as well as the short model, meaning that *rem\_any* is generated at random’

### 2.2.2 (Optional) Why would you reject?

What criteria would lead you to reject this null hypothesis?

‘Having an F-statistic that produces a P-value  $< 0.05$ , which is the case’

### 2.2.3 (Optional) Conduct an f-test

Using variables that indicate:

- sex (as noted in the codebook) (`female`);
- age (`age`);
- high school completion (`highschool_completed`);
- wealth (`wealthy`);
- marriage status (`married`);
- weekly income (`inc_7d`);
- discount preferences (`hyperbolic`);
- and, spend before saving (`spent_b4isaved`);
- meeting savings goals (`saved_asmuch`);
- missingness of covariates indicator (`missing_female`, `missing_age`, ...)

your team has conducted an F-test to evaluate whether there is evidence to call into question whether respondents in the *Philippines* were randomly assigned to receive any reminder (`rem_any`).

```
short_model <- lm(rem_any ~ 1, data = d[d$country == 3,])
long_model  <- lm(rem_any ~ female + missing_female + age + missing_age +
                  highschool_completed + missing_highschool_completed +
                  wealthy + missing_wealthy + married + missing_married +
                  inc_7d + missing_inc_7d + hyperbolic + missing_hyperbolic +
                  spent_b4isaved + missing_spent_b4isaved + saved_asmuch +
                  missing_saved_asmuch,
                  data = d[d$country == 3,])

# after filling in the `long_model` above,
# you should be able to conduct your test by uncommenting the line below
anova(short_model, long_model, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: rem_any ~ 1
## Model 2: rem_any ~ female + missing_female + age + missing_age + highschool_completed +
##          missing_highschool_completed + wealthy + missing_wealthy +
##          married + missing_married + inc_7d + missing_inc_7d + hyperbolic +
##          missing_hyperbolic + spent_b4isaved + missing_spent_b4isaved +
##          saved_asmuch + missing_saved_asmuch
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1    1408  206.93
## 2    1396  203.96 12     2.9752 1.697 0.06181 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 2.2.4 (Optional) What do you conclude?

Do you reject or fail to reject the null hypothesis?

‘Since we are trying to determine whether the reminder was randomly assigned for this specific country, even if the p-value is higher than 0.5 (not by far), we should take a closer look just to

be sure especially since the previous example would reject the null for all the dataframe. When looking at the RSS there is not much difference in both models so in this case, I would fail to reject the null hypothesis (even if the p-value is low).'

### 2.2.5 (Optional) Interpret your conclusions

What do you conclude from this test? Do the additional covariates increase the model's ability to predict treatment? This is an example of using a "Golem" model for a specific task.

'When adding the additional covariates it seems like our ability to predict indeed increases, but not as much as we would need to say that the reminders are not randomly assigned for this specific country.'

## 2.3 Reproduce Table 4

There is **a lot** that is happening in Table 4 of this paper. In this part of the question, you will reproduce some parts of this table. First, reproduce the OLS regression estimates that are in the upper right of Table 4. That is, estimate effects of SMS message on meeting savings goals.

In Section 3.1 of the included paper, the authors describe the OLS model that they estimate:

$$Y_i = \alpha + \beta R_i + \gamma Z_i + \epsilon_i$$

For the upper right panel that you are estimating, the outcome,  $Y_i$  is a binary indicator for whether the individual met their savings goal. The indicator  $R_i$  is a binary indicator for whether the individual was assigned to receive a reminder. And,  $Z_i$  is a vector of additional features: a categorical variable for the country, and a binary indicator for whether the individual was recruited by a marketer. In the model labeled (3) only  $Y$ ,  $R$  and  $Z$  are used in the regression. In the model labeled (4) these variables are used, but so too are the other variables that you previously used in the F-test.

### 2.3.1 Evaluate the large sample assumptions

Examining the data, and any information provided by the authors in the paper, evaluate the assumptions for the large-sample linear model. Are the necessary assumptions met for this regression model to produce consistent estimates (i.e. estimates that converge in probability to the population values)? Why or why not? If you use data to evaluate these assumptions, please feel free to show your EDA and evaluation in this document.

'We have more than 13,500 observations, so we can indeed treat this as a large sample so we need to evaluate two assumptions. IID data, which could be argued since we don't really know if the samples are indeed random (looking at 2.2.1 they might not be), so this assumption, at least in the pooled sample may not be satisfied. Regarding a unique BLP, when looking at our outcome variable we don't have infinite variance nor heavy tails and actually this is a binary variable with just two possible values, so the second assumption is satisfied.'

### 2.3.2 Conduct these regressions

The authors have concluded that they can conduct these regressions. So, in the next code chunk, would you please conduct these regressions? You will have to read the notes below Table 4 to get exactly the correct covariate set that reproduces the reported estimates. First, estimate the model that is reported in model (3); then, estimate the model that is reported in model (4). You should be able to exactly reproduce their results, including number of observations, coefficients, and standard errors.

**Table 4** Estimates of the Effect of Getting Any Reminder (vs. No Reminder)

Savings measure on LHS:	log(1 + <i>Amount saved</i> )		1 = <i>Met commitment</i>	
	(1)	(2)	(3)	(4)
Panel A: Pooled sample				
<i>Pooled sample</i>	0.059 (0.037)	0.061* (0.037)	0.032** (0.009)	0.032*** (0.009)
Baseline controls	No	Yes	No	Yes
Mean of DV	3.129	3.129	0.553	0.553
<i>N</i>	13,560	13,560	13,560	13,560
Panel B: Countries				
<i>Peru</i> ( <i>n</i> = 2,775)	0.033 (0.059)	0.023 (0.060)	0.038 (0.027)	0.034 (0.027)
<i>Bolivia</i> ( <i>n</i> = 9,376)	0.058 (0.043)	0.057 (0.042)	0.033*** (0.010)	0.032*** (0.010)
<i>Philippines</i> ( <i>n</i> = 1,409)	0.115 (0.099)	0.159 (0.098)	0.015 (0.029)	0.020 (0.028)
Baseline controls	No	Yes	No	Yes
Mean of DV	3.129	3.129	0.553	0.553
<i>N</i>	13,560	13,560	13,560	13,560
<i>P</i> -value from <i>F</i> -test of Peru = Bolivia	0.74	0.64	0.86	0.96
<i>P</i> -value from <i>F</i> -test of Peru = Philippines	0.48	0.24	0.57	0.74
<i>P</i> -value from <i>F</i> -test of Bolivia = Philippines	0.59	0.34	0.57	0.69

*Notes.* Ordinary least squares were used, with Huber–White standard errors in parentheses. *Amount saved* is the total amount of money deposited from account opening through the end of the commitment period. *Met commitment* is adhering to the term of the commitment: making all of the required deposits in Peru or Bolivia and saving the goal amount by the end of the commitment period in the Philippines. All regressions include controls for marketing offers in the Philippines (interest rate, joint/single account, deposit collection) and country fixed effects. Baseline controls include the full set of household demographics listed in Table 3 and department, province, branch, and marketer fixed effects. DV, dependent variable; LHS, left-hand side.

\* $P < 0.10$ ; \*\* $P < 0.05$ ; \*\*\* $P < 0.01$ .

Figure 1: Tables of Models to Reproduce. Students should read the caption to this table carefully, because it describes the process used to estimate this model. This style of reporting should be emulated in subsequent homework and lab work!

```
d$country <- as.factor(d$country)

mod_pooled_no_covariates <-
  lm(
    reached_b4goal ~ rem_any + country + marketer +
      joint + joint_single + dc + highint + rewardint,
    data = d
  )

mod_pooled_with_covariates <-
  lm(
    reached_b4goal ~ rem_any + country + marketer +
      joint + joint_single + dc + highint + rewardint +
      female + age +
      highschool_completed +
      wealthy + married +
      inc_7d + hyperbolic +
      spent_b4isaved + saved_asmuch,
    data = d
  )
```

### 2.3.3 Do covariates improve model fit?

Does the addition of the covariates improve the fit of the model? First, compute the MSR for each model (you can use methods from the first question, either `augment` or `resid`). Then, conduct an F-test to evaluate.

```
mean_squared_residual_no_covariates <- calculate_msr(mod_pooled_no_covariates)
mean_squared_residual_with_covariates <- calculate_msr(mod_pooled_with_covariates)
```

The mean squared residuals of the short model are, **0.2305361**. The mean squared residuals of the long model are **0.2297814**. In the next chunk, we test whether the MSRs of the models are different using an F-test.

```
f_test_of_long_vs_short <- anova(mod_pooled_no_covariates,
                                mod_pooled_with_covariates,
                                test = "F")
```

```
f_test_of_long_vs_short
```

```
## Analysis of Variance Table
##
## Model 1: reached_b4goal ~ rem_any + country + marketer + joint + joint_single +
##      dc + highint + rewardint
## Model 2: reached_b4goal ~ rem_any + country + marketer + joint + joint_single +
##      dc + highint + rewardint + female + age + highschool_completed +
##      wealthy + married + inc_7d + hyperbolic + spent_b4isaved +
##      saved_asmuch
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1  13550 3126.1
## 2  13541 3115.8   9    10.235 4.9422 1.182e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

‘We are testing whether the long model (with covariates) explains things better than the short model (with no covariates), so we’re running an F-test (besides measuring MSR for both models, since we already know that this will decrease just for adding variables). To be able to run this F-test our models should be “nested”, meaning that the short model (all its independent variables) should be “included” in the large model, and the null hypothesis for this test is  $H_0$ : The long model explains things as well as the short model (there’s no improvement)’. In this specific case we can reject the null hypothesis and choose the large model, which could imply having a more robust model (although we still need to check which coefficients are statistically significant), but it also means that we need to measure more variables which could be more expensive or difficult from a business point of view.’

### 2.3.4 Robust standard errors

The authors report that they used Huber-White standard errors. That is to say, they used robust standard errors. Use the function `vcovHC` – the variance-covariance matrix that is heteroskedastic consistent – from the `sandwich` package, together with the `coefTest` function from the `lmtest` package to print a table for each of these regressions.

```
# you can uncomment the following lines to conduct and report a test with robust standard errors
# notice that you are not storing the results of this test, and instead simply printing to the screen
#
coefTest(mod_pooled_no_covariates, vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)   0.6629276  0.0119354  55.5429 < 2.2e-16 ***
## rem_any       0.0317635  0.0091548   3.4696 0.0005229 ***
## country2     -0.1144302  0.0107317 -10.6628 < 2.2e-16 ***
## country3     -0.4560764  0.0326176 -13.9825 < 2.2e-16 ***
## marketer     -0.0037840  0.0024279  -1.5586 0.1191261
## joint         0.0038143  0.0275066   0.1387 0.8897150
## joint_single -0.0351725  0.0256539  -1.3710 0.1703861
## dc            0.0138276  0.0251819   0.5491 0.5829397
## highint      -0.0019591  0.0257396  -0.0761 0.9393310
## rewardint     0.0465328  0.0270142   1.7225 0.0849967 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coefTest(mod_pooled_with_covariates, vcov = vcovHC(mod_pooled_with_covariates, type="HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)   6.3341e-01  2.1777e-02  29.0862 < 2.2e-16 ***
## rem_any       3.2097e-02  9.1408e-03   3.5114 0.0004472 ***
## country2     -1.4425e-01  1.5278e-02  -9.4420 < 2.2e-16 ***
## country3     -4.7686e-01  3.8249e-02 -12.4671 < 2.2e-16 ***
## marketer     -3.8531e-03  2.4068e-03  -1.6009 0.1094142
## joint         1.3816e-03  2.7265e-02   0.0507 0.9595869
```

```
## joint_single      -3.9615e-02  2.5529e-02  -1.5517  0.1207467
## dc                1.1133e-02  2.4979e-02   0.4457  0.6558296
## highint           -3.6289e-03  2.5586e-02  -0.1418  0.8872173
## rewardint         4.2501e-02  2.6886e-02   1.5808  0.1139531
## female            1.8845e-02  8.7695e-03   2.1489  0.0316593 *
## age               1.2333e-03  3.3689e-04   3.6609  0.0002523 ***
## highschool_completed 2.3075e-03  9.2715e-03   0.2489  0.8034521
## wealthy           -4.2974e-02  1.5811e-02  -2.7180  0.0065762 **
## married           3.7883e-02  1.3249e-02   2.8593  0.0042528 **
## inc_7d            -5.4915e-05  1.7896e-04  -0.3068  0.7589628
## hyperbolic        8.4034e-05  2.9681e-02   0.0028  0.9977410
## spent_b4isaved    -3.5909e-03  2.7329e-02  -0.1314  0.8954645
## saved_asmuch      1.5735e-02  2.5494e-02   0.6172  0.5371183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 2.3.5 State your null hypotheses

For each of the coefficients in the table you have just printed, there is a p-value reported: This is a p-value for a hypothesis test that has a null hypothesis. What is the null hypothesis for each of these tests?

‘The null Hypothesis for each one of the coefficients is that the coefficient is equal to zero (it wouldn’t be statistically significant).  $H_0 : \beta_k = 0$  for every coefficient in our model’

### 2.3.6 Which tests reject the null hypothesis

Suppose that your criteria for rejecting the null hypothesis were: “The p-value must be smaller than 0.05”. Then, which of these coefficients rejects that null hypothesis? (Keep only one of the options in the “Determination” column of the table below.)

Variable	Determination
rem_any	Significant
joint	Not Significant
joint_single	Not Significant
dc	Not Significant
highint	Not Significant
rewardint	Not Significant
Lives in Bolivia	Significant
Lives in Peru	Significant
female	Significant
age	Significant
highschool_completed	Not Significant
wealthy	Significant
married	Significant
inc_7d	Not Significant
saved_asmuch	Not Significant
spent_b4isaved	Not Significant

### 2.3.7 Interpret the effect of being sent a reminder

Interpret the meaning of the coefficient estimated when individuals are sent any reminder, which is encoded on the `rem_any` variable. We will talk about this more in a later unit, but this is the treatment effect from this experiment. As you are interpreting this coefficient, keep in mind the nature of the `rem_any` variable – how many levels are there in this variable? How is this variable encoded? What does a one-unit change on this variable mean? As well, keep in mind that the outcome variable measures whether the individual met their commitment. How is this variable encoded and what does a coefficient mean in this context?

‘This variable is binary with 1 = receiving a reminder. What this coefficient tells us is that we would expect an increase of  $\beta_{rem\_any} = 3.21\%$  in the clients’ “chances” to meet their goal if a reminder is received (everything else equal).’

### 2.3.8 Interpret the coefficient on age

Interpret the meaning of the coefficient estimated on `age`.

‘When talking about age, using this model, we would expect a positive relationship between age and meeting the saving goals, specifically an increase of  $\beta_{age} = 0.123\%$  in their chances to meet their goal for each additional year in the clients’ age.’

### 2.3.9 Interpret the coefficient on `highschool_completed`

Interpret the meaning of the coefficient estimated on `highschool_completed`.

‘This coefficient would imply that we could expect an increased  $\beta_{highschool} = 0.231\%$  chance from those who finished high school, although this coefficient has a very high p-value (0.803) which makes it not significant (statistically) and obtaining it could be just a matter of our sample (luck) rather than having a true relationship.’

### 2.3.10 Print a whole table.

Finally, produce a legible regression table using the `stargazer` package that summarizes the work that you have just done. This regression table should

- Contain the model without covariates as well as the model with covariates.
- Contain the coefficients that you have estimated for the model.
- Contain the standard errors you have estimated for the model.
- Contain labels that are written in English (i.e. not variable names) that describe each concept tested on each row

Once you have estimated these models, you can print them to the screen using the `stargazer` package.

```
## while you are writing your code, you can use `type = 'text'` to print to the console
## when you compile your PDF to submit, if you like you can format this as a latex table. to do so:
## 1. change the `type = 'text'` to be `type = 'latex'` in the stargazer function call; and,
## 2. pass an argument into the chunk declaration (i.e. after `warning = FALSE` above), that is `result`

stargazer(
  mod_pooled_no_covariates,
```

```

mod_pooled_with_covariates,
no.space=TRUE,
title="Results",
dep.var.labels=c("Met Commitment"),
column.labels=c("No Covariates","With Covariates"),
order="Constant",
covariate.labels=c("(Intercept)", "Reminder","Lives in Bolivia", "Lives in Phillipines", "Marketer Con
                    "Single Account", "Deposit Collection", "High Int Rate", "Reward Int Rate", "Gender
                    "Age", "High School","Wealthy","Married","Income (7 days)", "Hyperbolic Pref", "Sper
                    "Saved as Wanted")
)

```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: mié., mar. 05, 2025 - 09:46:54 p. m.

Table 2: Results

	<i>Dependent variable:</i>	
	Met Commitment	
	No Covariates	With Covariates
	(1)	(2)
(Intercept)	0.663*** (0.012)	0.633*** (0.022)
Reminder	0.032*** (0.009)	0.032*** (0.009)
Lives in Bolivia	-0.114*** (0.011)	-0.144*** (0.016)
Lives in Phillipines	-0.456*** (0.039)	-0.477*** (0.046)
Marketer Contact	-0.004 (0.003)	-0.004 (0.003)
Joint Account	0.004 (0.032)	0.001 (0.032)
Single Account	-0.035 (0.031)	-0.040 (0.031)
Deposit Collection	0.014 (0.030)	0.011 (0.030)
High Int Rate	-0.002 (0.031)	-0.004 (0.031)
Reward Int Rate	0.047 (0.032)	0.043 (0.032)
Gender (F = 1, M = 0)		0.019** (0.009)
Age		0.001*** (0.0003)
High School		0.002 (0.009)
Wealthy		-0.043** (0.017)
Married		0.038*** (0.014)
Income (7 days)		-0.0001 (0.0002)
Hyperbolic Pref		0.0001 (0.035)
Spent B4 Save		-0.004 (0.032)
Saved as Wanted		0.016 (0.030)
Observations	13,560	13,560
R <sup>2</sup>	0.067	0.070
Adjusted R <sup>2</sup>	0.067	0.069
Residual Std. Error	0.480 (df = 13550)	0.480 (df = 13541)
F Statistic	108.546*** (df = 9; 13550)	56.886*** (df = 18; 13541)

Note:

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01