Unit 09 Homework

w203: Statistics for Data Science

March 15, 2022

```
library(tidyverse)

library(patchwork)
library(stargazer)

library(sandwich)
library(lmtest)
```

Here is our intention for the homework:

1. By simulating the data in **Question 1** you should be able to observe everything thing 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

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)
d
```

```
##      id         x1        x2          x3          y
## 1     1 5.61155745  8.816511 -0.01606015  5.5088829
## 2     2 6.31219910 10.593914  3.25957711  8.3206463
## 3     3 4.79004403  8.738957 -0.50221164  5.5428095
## 4     4 8.23371354  8.867763 -0.55008359  6.9365237
## 5     5 5.00355662  9.502472  0.98213025  6.6699135
## 6     6 2.41439839 11.204018  1.50256442  3.3713501
## 7     7 2.89960418 10.024356  0.22404712  2.1272723
## 8     8 6.59458088  6.061981 -1.74543624  7.9745814
## 9     9 3.81959031 10.143620  3.73615438  7.7191800
## 10   10 8.27446564 11.525162 -1.26958947  8.9145652
## 11   11 0.72289810 11.686588 -0.18251681  0.3740936
## 12   12 5.47315720 11.585678  1.60595245  5.6918091
## 13   13 8.06705129 10.225890  0.28849161  8.8010800
## 14   14 6.40558490 12.973203  1.62913590  8.0837635
## 15   15 5.65269213 10.879882  1.65030743  6.8714794
## 16   16 1.23494717 13.022938 -1.06164695  2.6766600
## 17   17 7.46518909 11.018893  0.74627882  8.5594883
## 18   18 2.62777369 10.316134  0.61944803  4.2852721
## 19   19 3.81461910  7.737222  1.04097850  4.1538726
## 20   20 2.60461480  9.813360 -1.15409216  2.1569313
## 21   21 8.54713104 13.169946 -0.60712104 10.5065861
## 22   22 0.08267593 11.706639  1.07088715  0.7460060
## 23   23 6.08875804  8.219634  0.17107567  6.3728251
## 24   24 3.00593151 10.932514  1.70914336  5.4023924
## 25   25 9.16900265 11.063952 -0.09393189 11.4328007
## 26   26 3.50209022  6.191685  0.71001695  3.4077633
## 27   27 9.40914704 10.882639 -0.59582206  9.9205749
## 28   28 4.54628360 12.327513 -0.22432470  4.8841020
## 29   29 7.60688349 10.472801  0.26503598  7.7293001
## 30   30 3.23058144 11.246286 -2.45257348  5.9477798
## 31   31 4.92511394  9.471951 -2.21468095  6.0950521
## 32   32 4.34011128 12.661587  0.20126052  6.2286656
## 33   33 2.05391731 11.020193  2.89015824  5.0103556
## 34   34 0.21156359 11.046157 -1.14626283  0.9328698
## 35   35 0.22011505 10.582862  3.31033274  4.0219622
## 36   36 8.61843819  7.663276  3.36811614 11.9764086
## 37   37 1.82342421  7.059676 -3.47132185  5.3099919
## 38   38 9.42471461  8.303951 -3.48089439 12.3657282
## 39   39 3.34723286  7.669199 -2.61232296  5.2470140
## 40   40 2.37610674 14.147268  0.43386199  3.8287317
## 41   41 1.22802234  8.378942  1.03149428  1.9511662
## 42   42 6.53667523  7.778914 -1.25219640  6.2354763
## 43   43 2.03006285  8.214281 -0.91364687  4.3140486
## 44   44 7.64330685 13.749588 -5.11093144 15.9397344
## 45   45 9.70777926 10.621827 -3.48458089 12.5859658
## 46   46 5.34059291 11.454723  1.16567376  6.5789123
## 47   47 4.27817112 10.860965 -1.19149639  4.4993402
## 48   48 4.14689496  7.560129  1.43266333  5.3105045
## 49   49 5.15611092 11.525396  3.49267896  9.0563075
## 50   50 9.78406707 14.309255 -0.11779410 10.4349214
## 51   51 8.43345002 12.606873 -0.70606665  9.1909146
## 52   52 2.14342962 10.902641 -0.35751311  2.3520923
## 53   53 2.05617203 10.542451  3.63751586  5.9369484
## 54   54 7.82167466 12.319814 -2.76971885  8.4814461
## 55   55 6.17913882  8.556703 -0.22744740  7.8630115
## 56   56 9.43559950  7.476229  1.69597767 12.5515401
## 57   57 4.67138377 12.606611  2.06116205  5.7891453
## 58   58 8.51142158 10.280861  0.50597052  7.8139100
## 59   59 7.42777466  8.006083  4.26570434 13.9295813
## 60   60 0.15291855  5.231747 -0.96260175  0.4529808
## 61   61 5.70614396  8.399235 -0.02000583  6.2242521
## 62   62 1.85873154 11.394814 -0.21411596  0.2903629
## 63   63 6.86828172 13.761137 -1.74328232  9.2745871
## 64   64 2.23067293  8.981215 -1.87377417  4.6489587
## 65   65 6.74984352  9.572978  1.33425220  7.2374056
## 66   66 9.25697536  6.017151  0.42676197  9.7067393
## 67   67 8.97523322 11.958088  0.98235830 10.7860785
## 68   68 4.20605783  8.873559  0.37563794  6.1988389
## 69   69 9.66166681 11.260103 -0.54709329 10.6536185
## 70   70 9.87618332 12.288248 -0.47076361 12.0281047
## 71   71 9.98865681 10.029226 -1.05861588 10.2031477
## 72   72 7.16797476  8.619469 -6.06553712 17.5808419
## 73   73 2.83301205 10.295725 -1.62010782  3.1293406
## 74   74 7.12166080 12.204295  0.77787986  7.4048730
## 75   75 7.38827997  6.198928  2.16529494  8.5764342
## 76   76 7.82948171  8.674670  2.13049980  7.9040553
## 77   77 0.40272478 13.831834  0.14710233  1.1103445
## 78   78 4.13763216 10.538575  0.01434997  5.3510226
## 79   79 4.90246693  9.451379 -0.50743408  5.8916177
## 80   80 5.00949564  8.651638  0.67743121  5.9852809
## 81   81 3.92623066  8.462229 -0.88743344  3.3538556
## 82   82 8.34803257  6.772439  3.65239229 13.2756286
## 83   83 3.39560775 10.058198  1.86517336  5.1560809
## 84   84 7.95248434 10.485070  1.13020733  8.6597784
## 85   85 9.85211874 10.923434 -0.02829247 10.5622677
## 86   86 9.87120618 15.252855 -2.20246249 10.4830095
## 87   87 3.38199529  9.941705  1.38992890  4.6366683
## 88   88 1.66781045 16.594210  0.52210981  0.2714708
## 89   89 8.58786480  8.874583 -0.33033657  7.9740568
## 90   90 6.13347398 11.530429  1.67343499  5.4128535
## 91   91 6.89085925  9.841489  0.04335680  7.1528283
## 92   92 7.47771124  7.876500  0.01981807  7.7093051
## 93   93 7.16751359 11.489061  1.59367853  8.5549331
## 94   94 7.11317120  7.164669  1.65527607  7.3637543
## 95   95 8.48686868  9.177213 -0.10691489  9.0607118
## 96   96 6.34279177 10.447149  1.63591030  7.8849905
## 97   97 0.31770578 12.093937  1.35765981  0.3531779
## 98   98 8.04329850  7.964688 -2.02520601  8.9979726
## 99   99 7.26256672 10.786503  2.33453872  9.1652807
## 100 100 8.91721420  7.576989 -2.54492246 10.3047888
```

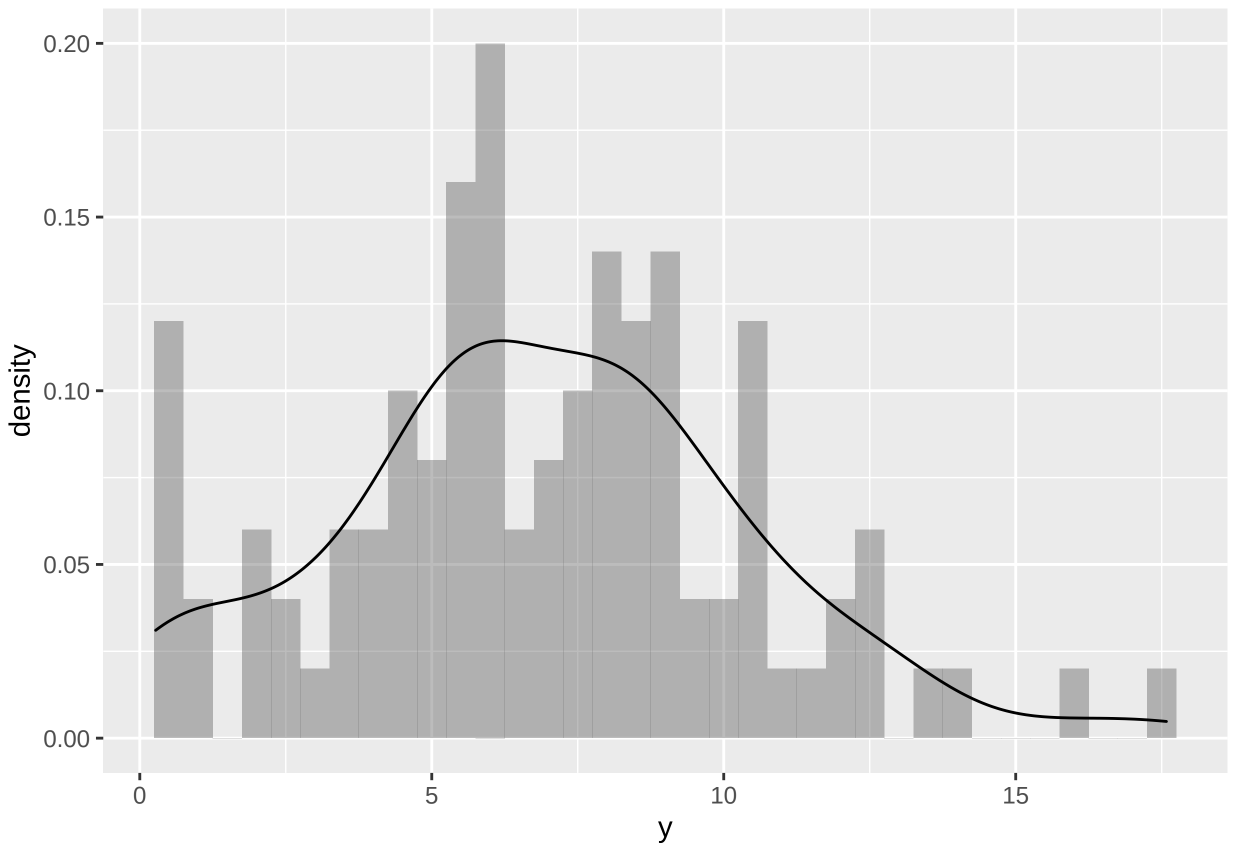
1.1 Plot of outcome data

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?

```
outcome_histogram <- d %>%
    ggplot(aes(x = y)) +
    geom_histogram(aes(y= ..density..), alpha = 0.4, binwidth = 0.5) +
    geom_density()
    labs(title = "Density Histogram",x = "Density",y = "Variable Y")
```

```
## $x
## [1] "Density"
##
## $y
## [1] "Variable Y"
##
## $title
## [1] "Density Histogram"
##
## attr(,"class")
## [1] "labels"
```

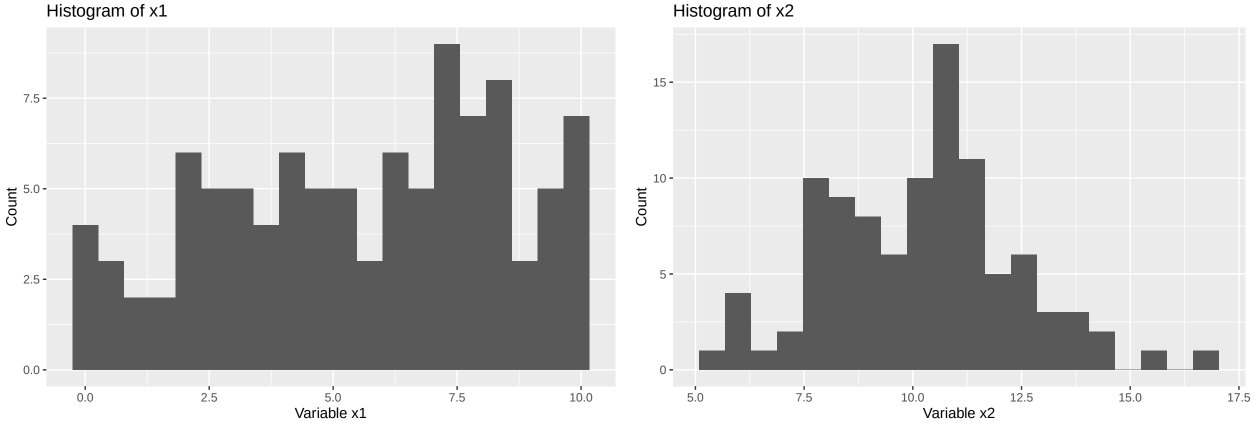
```
    # fill in the rest of this chunk to create a plot
    # you will need aes layers (to map data into the plot)
    # and geom_* layers to draw the plot. You can delete these
    # comments if you like.
outcome_histogram
```

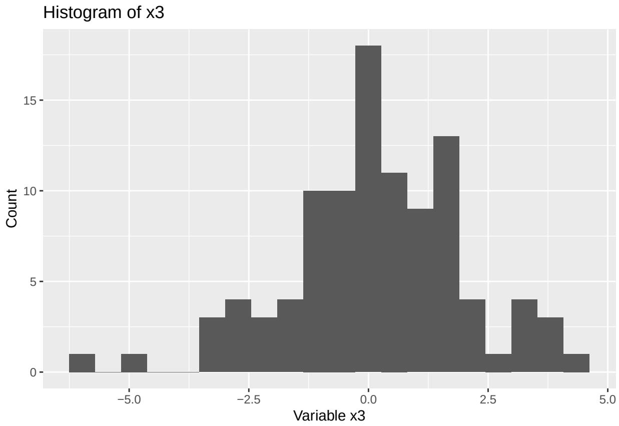


Answer: Looking at the this distribution you notice it is not normal distribution.

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?





Answer: Since we have a sample size of 100 we can use the assumptions for the large-sample model which require that the data be IID and that there is a single unique BLP. It is safe to assume there is one unique BLP and that data is independent since it is a random generation. However the data does not appear to be identically distributed.

1.3 Estimate four models

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

$$Y = \beta_0 + \beta_1 x_1 + 0x_2 + \beta_3 x_3 + \epsilon \tag{1}$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{2}$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^2 + \epsilon \tag{3}$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3^2 + \epsilon \tag{4}$$

```r
# 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.

```r
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)
```

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?

Answer: Yes but only slightly because the variance of the residual is constant so it will not vary much as the value of the predictor variable changes. Looking at model 1 vs model 2 you can see only the tiniest of differences.

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?

Answer: The smaller the residual sum of squares, the better your model fits making model 4 the best of the best at estimating the population coefficients.

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 β_2. What do you conclude from this sample about the relationship between x_2 and y?

```
waldtest(model_2, model_1)
```

```
## Wald test
##
## Model 1: y ~ x1 + x2 + x3
## Model 2: y ~ x1 + x3
##   Res.Df Df      F Pr(>F)
## 1     96
## 2     97 -1 0.2357 0.6284
```

Answer: The results from the Wald-test indicated a large P value for the coefficient β_2. Because of this we cannot reject the null hypothesis and the variables in model 1 will need to be included to find a good model fit.

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

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x3
## Model 2: y ~ x1 + x2 + x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     97 307.60
## 2     96 306.84  1   0.75349 0.2357 0.6284
```

Answer: Comparing the variability of Model 1 to Model 2 we see that the rss is the same, with a large P value and a small f value. Mean model 2 is more appropriate and we can reject the null.

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?

Answer: The F test and Wald_test are asking the same question of the model. The Wald-test shows that if the parameters for certain explanatory variables are zero, you can remove the variables from the model. With the F test we are asking if the means between two models are significantly different.

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, $rem\_any \ \sim \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_{long}$, where $x_1 + \ldots$ 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.

2.2.1 (Optional) State the null

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

> ‘Fill in here: What is the null hypothesis?’

2.2.2 (Optional) Why would you reject?

What criteria would lead you to reject this null hypothesis?

> 'Fill in here: What would lead you to reject this null hypothesis?'

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`).

```r
short_model <- lm(rem_any ~ 1, data = d[d$country == 3,])
long_model  <- 'fill this in'

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

2.2.4 (Optional) What do you conclude?

Do you reject or fail to reject the null hypothesis?

> 'Fill in here: Do you reject or fail to reject the null hypothesis?'

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.

> 'Fill in here: What do you conclude?'

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

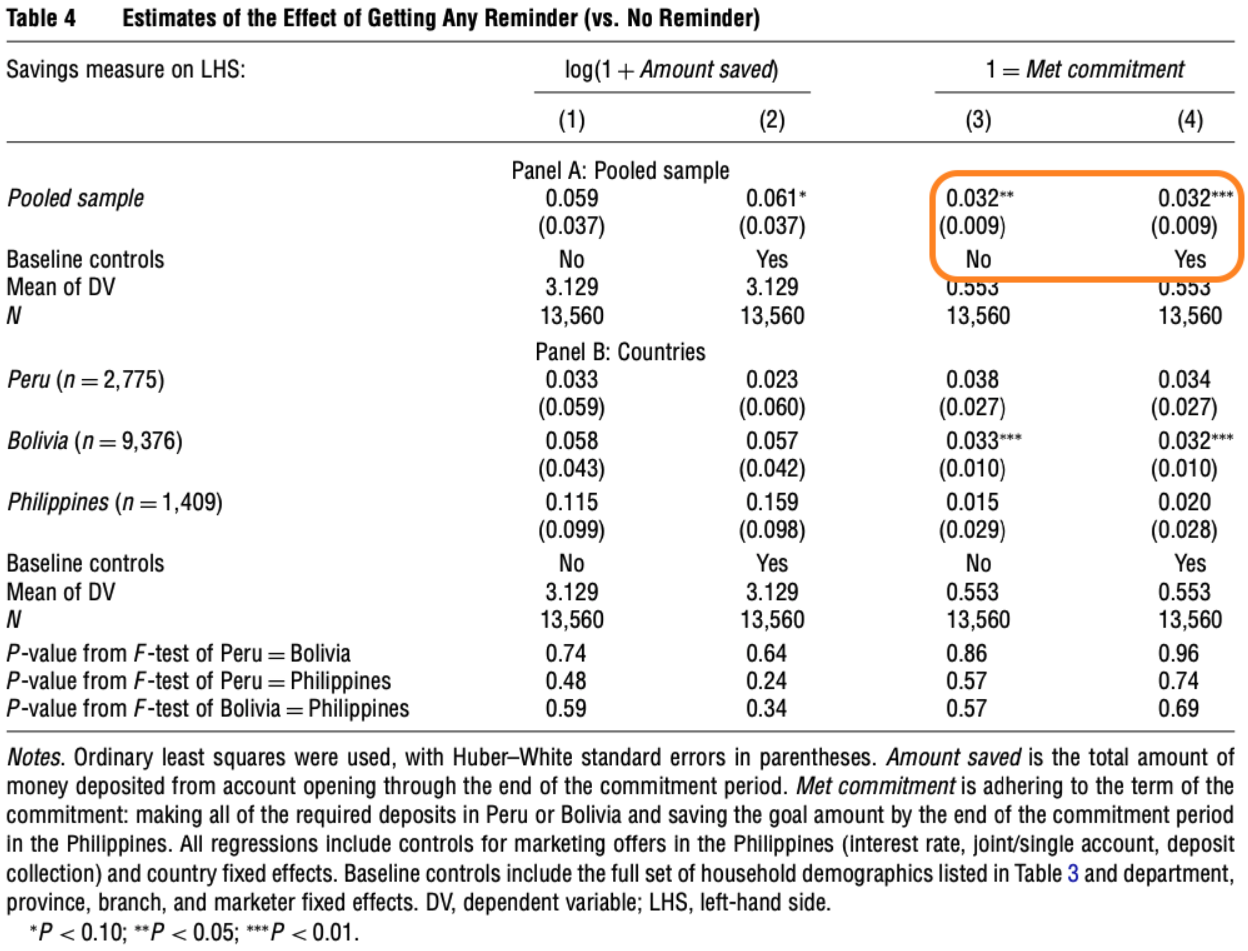


Table 4 Estimates of the Effect of Getting Any Reminder (vs. No Reminder)

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

(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.

'Fill in here: What are the large sample assumptions, and are the assumptions 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.

```
mod_pooled_no_covariates   <- lm(reached_b4goal ~ rem_any, data =d)
mod_pooled_with_covariates <- lm(reached_b4goal ~ rem_any + age + female + highschool_completed + marri
mod_pooled_no_covariates
```

```
##
## Call:
## lm(formula = reached_b4goal ~ rem_any, data = d)
##
## Coefficients:
## (Intercept)      rem_any
##     0.53845      0.02432
```

```
mod_pooled_with_covariates
```

```
##
## Call:
## lm(formula = reached_b4goal ~ rem_any + age + female + highschool_completed +
##     married + wealthy + inc_7d + hyperbolic + spent_b4isaved +
##     saved_asmuch, data = d)
##
## Coefficients:
##          (Intercept)               rem_any                   age
##             0.464602              0.045537              0.002201
##               female  highschool_completed               married
##             0.017870              0.013966             -0.030578
##              wealthy                inc_7d            hyperbolic
##             0.011682             -0.001293             -0.122270
##       spent_b4isaved          saved_asmuch
##            -0.010259             -0.305196
```

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   <- 'fill this in'
mean_squared_residual_with_covariates <- 'fill this in'
```

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

```
f_test_of_long_vs_short <- 'fill this in'
```

> ’Fill in here: What were you testing? What was your null hypothesis? What test did you choose to conduct? What has to be true of the data in order to conduct this test (i.e. what are the assumptions)? What kind of evidence would lead you to reject that null hypothesis? What do you conclude? What are the research / business implications for this statistical conclusion?

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)
# coeftest(mod_pooled_with_covariates, vcovHC)
```

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?

> ‘Fill in here: What is the null hypothesis that is at stake for each of these coefficients?’

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 OR Not Significant
`joint`	Significant OR Not Significant
`joint_single`	Significant OR Not Significant
`dc`	Significant OR Not Significant
`highint`	Significant OR Not Significant
`rewardint`	Significant OR Not Significant
Lives in Bolivia	Significant OR Not Significant
Lives in Peru	Significant OR Not Significant
`female`	Significant OR Not Significant
`age`	Significant OR Not Significant
`highschool_completed`	Significant OR Not Significant
`wealthy`	Significant OR Not Significant
`married`	Significant OR Not Significant
`saved_formal`	Significant OR Not Significant
`inc_7d`	Significant OR Not Significant

Variable	Determination
`saved_asmuch`	Significant OR Not Significant
`spent_b4isaved`	Significant OR 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?

> ’Fill in here: Interpret `rem_any`.

2.3.8 Interpret the coefficient on `age`

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

> ‘Fill in here: Interpret `age`.’

2.3.9 Interpret the coefficient on `highschool_completed`

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

> ‘Fill in here: Interpret `highschool_completed`.’

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 you 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 `resul

# stargazer(
#  'fill this in'
#  )
```