Unit 09 Homework

w203: Statistics for Data Science

6/28/2021

```
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
library(patchwork)
library(stargazer)

library(sandwich)
library(lmtest)
```

1. Simulated Data

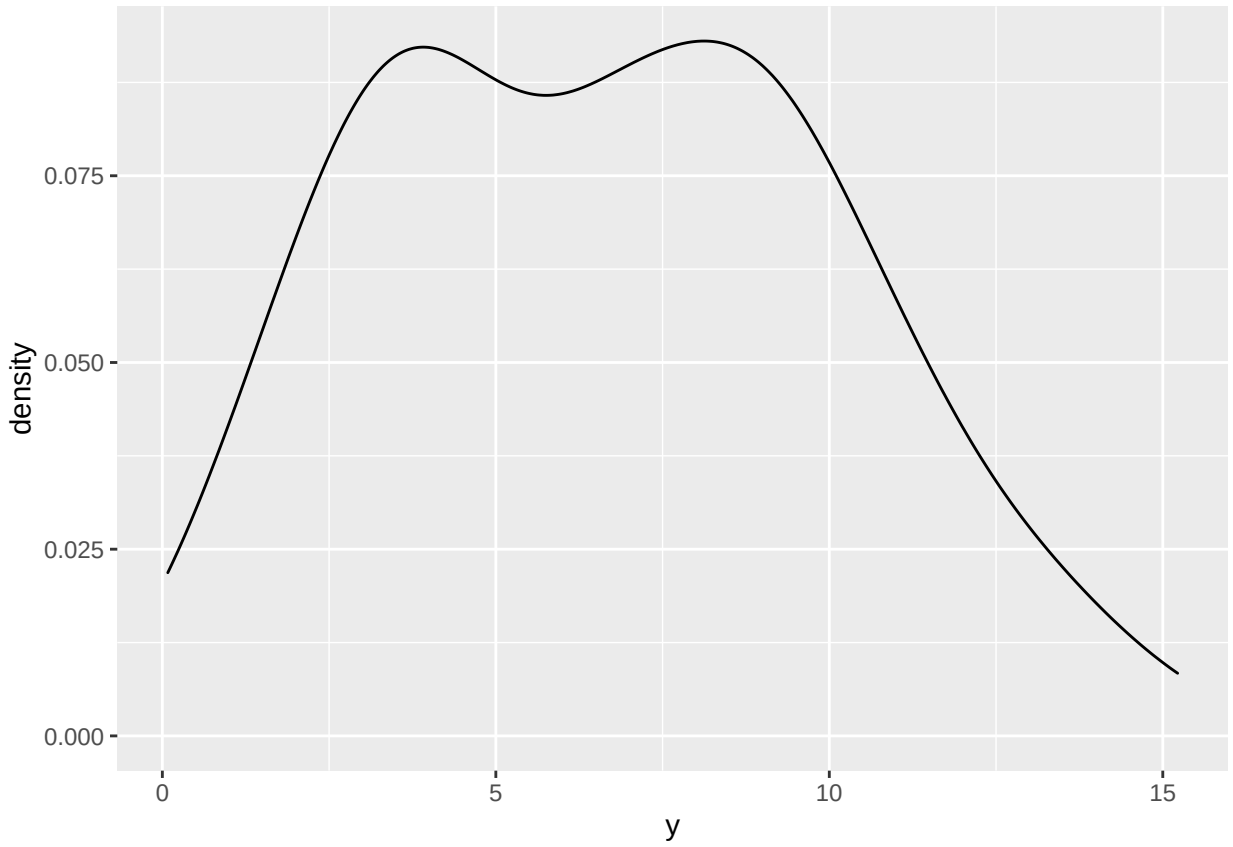
For this question, we are going to create data, and then estimate models on this simulated data. This allows us to effectively *know* the population parameters that we are trying to estimate. Consequently, we can reason about how well our models are doing.

```
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. 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?

```r
outcome_density <- d %>%
  ggplot() +
    aes(x = y) +
    geom_density()
outcome_density
```



I chose a density plot to better visualize the distribution of the data. Looking at the graph above, the outcome data seems to be slightly skewed to the left. In other words, we can see that the highest density of our data points based on the model "y" lies roughly towards the center but slightly skewed left.

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

   To apply the large-sample assumptions, we need to have greater or equal to 100 data points. We have n = 100, so therefore, we can use the Large-Sample Linear Model. As for the assumptions, there are two; I.I.D and a unique BLP should exists. In our case, IID is satisfied because the data is indepent and identically distributed. Looking at hte graph, we see a lot of similarities to normal distribution, therefore a Unique BLP exits.

3. 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)
```

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.

```
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)
}
```

```
calculate_msr(model_1)
```

```
## [1] 2.606194
```

```
calculate_msr(model_2)
```

```
## [1] 2.499543
```

```
calculate_msr(model_3)
```

```
## [1] 0.8708029
```

```
calculate_msr(model_4)
```

```
## [1] 0.8671806
```

3. Which of these models does the "best" job at estimating the population coefficients?

   Model 1.

4. Conduct two tests about x_2.

a. 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?

```
coeftest(model_2, vcov = vcovHC(model_2, type='const'))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.877514   0.864797  3.3274  0.001244 **
## x1           1.057455   0.056220 18.8091 < 2.2e-16 ***
## x2          -0.160886   0.079494 -2.0239  0.045760 *
## x3          -0.223683   0.079611 -2.8097  0.006010 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After conducting the t-test for coefficient beta 2, we can cannot reject the null. In other words, x2 is not useful for predicting outcome for y.

b. 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     96 249.95
## 2     97 260.62 -1   -10.665 4.0961 0.04576 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. 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 part a, we are computing a test for the size of coefficients for model_2. In part b, we are conducting the the to study whether introducing a new coefficient to our nested model, affect the outcome of y. Both part a and b are essentially looking at the coefficients and whether they are affecting the outcome of y.

```
# Although it isn't necessary, we're going to remove all the data objects that
# you have created to this point so that you can start the next section with
# clear data.
rm(list = setdiff(ls(), 'calculate_msr'))
```

2. Real-World Data

"Can timely reminders *nudge* people toward increased savings?" Dean Karlan, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinnman publised 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, called `./karlan_data/karlan_2016.pdf`. 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.

A. Read the data

Read in the data using the following code.

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

B (Optional). Conduct an F-test

(This part B is optional; the later part C 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.

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

   'Fill in here: What is the null hypothesis?'

b. What criteria would lead you to reject this null hypothesis?

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

c. Using variables that indicate:

- sex (as noted in the codebook) (`female`);
- age (`age`);
- high school completion (`highschool_completed`);
- wealth (`wealthy`);
- marriage status (`married`);
- previous formal savings (`saved_formal`, which isn't in the codebook);
- weekly income (`inc_7d`);

- meeting savings goals (`saved_asmuch`)
- and, spend before saving (`spent_b4isaved`)

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$country == 3,])
long_model  <- lm(rem_any ~ female + age + highschool_completed + wealthy +
                  married + saved_formal + inc_7d + saved_asmuch +
                  spent_b4isaved, data = d[d$country == 3,])

anova(short_model, long_model, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: rem_any ~ 1
## Model 2: rem_any ~ female + age + highschool_completed + wealthy + married +
##     saved_formal + inc_7d + saved_asmuch + spent_b4isaved
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1408 206.93
## 2   1399 204.95  9    1.9763 1.4989 0.1429
```

d. Do you reject or fail to reject the null hypothesis?

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

e. 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?'

C. 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:

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

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!

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.

a. 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?

   There are two assumptions for large-sample Model. The data has to be I.I.D. and a Unique BLP should exists. The aggragate of the data is not entirely IID and the study does mention this as "Cross-site differences in setting and nonrandomized features motivate estimating site-specific treatment effects as well." However, the samples are IID when it comes to site specific samples. The sample is sufficiently large and the model being a linear model points to the data having a unique BLP. I would be comfortable running a regression on each site separately, but not the aggragate (pooled).

b. The authors have concluded that they can conduct these regressions. So, in the next code chunk, please conduct these regressions. First, estimate the model that is reported in model (3). You will have to read the notes below Table 4 to get exactly the correct covariate set that reproduces the reported estimates. Then, estimate the model that is reported in model (4).

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

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

```
library(stargazer)
## while you are writing you code, you can use 'type = 'text'' to print to the console
stargazer(mod_pooled_no_covariates,
   type = 'text', header = FALSE,
   star.cutoffs = c(0.05, 0.01, 0.001) # the default isn't in line with w203
)
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                             reached_b4goal
## ------------------------------------------------
## rem_any                        0.018*
##                                (0.009)
##
## country                       -0.149***
##                                (0.010)
##
```

```
## marketer                       -0.020***
##                                 (0.002)
##
## Constant                       0.842***
##                                 (0.020)
##
## ------------------------------------------------
## Observations                    13,560
## R2                               0.064
## Adjusted R2                      0.063
## Residual Std. Error       0.481 (df = 13556)
## F Statistic           306.410*** (df = 3; 13556)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

```r
library(stargazer)
## while you are writing you code, you can use 'type = 'text'' to print to the console
stargazer(mod_pooled_with_covariates,
   type = 'text', header = FALSE,
   star.cutoffs = c(0.05, 0.01, 0.001) # the default isn't in line with w203
)
```

== Dependent variable:
———————— reached_b4goal
———————————————— rem_any 0.028**
(0.009)

country -0.162***
(0.022)

marketer -0.010***
(0.003)

female 0.018*
(0.009)

age 0.001***
(0.0003)

highschool_completed -0.002
(0.009)

wealthy -0.065***
(0.016)

married 0.036**
(0.014)

saved_formal -0.026*
(0.012)

inc_7d -0.0002
(0.0002)

saved_asmuch -0.015
(0.030)

spent_b4isaved -0.008
(0.032)

depart -0.006*
(0.003)

provincia -0.022
(0.014)

branch -0.021
(0.015)

joint -0.028
(0.031)

joint_single -0.068*
(0.030)

dc -0.003
(0.030)

highint -0.036
(0.031)

rewardint 0.011
(0.031)

Constant 0.847***
(0.044)

Observations 13,560
R2 0.070
Adjusted R2 0.069
Residual Std. Error 0.480 (df = 13539)
F Statistic 51.293*** (df = 20; 13539)
========== Note: *p<0.05;* ***p<0.01;*** p<0.001

c. 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   <- mean(resid(mod_pooled_no_covariates)^2)
mean_squared_residual_with_covariates <- mean(resid(mod_pooled_with_covariates)^2)
mean_squared_residual_no_covariates
```

```
## [1] 0.2314617
```

```
mean_squared_residual_with_covariates
```

```
## [1] 0.2297488
```

The mean squared residuals of the short model are, 0.2314617. The mean squared residuals of the long model are 0.2297488. In the next chunk, we test whether the MSRs of the models are different.

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

```
## Analysis of Variance Table
##
## Model 1: reached_b4goal ~ rem_any + country + marketer + female + age +
##     highschool_completed + wealthy + married + saved_formal +
##     inc_7d + saved_asmuch + spent_b4isaved + depart + provincia +
##     branch + joint + joint_single + dc + highint + rewardint
## Model 2: reached_b4goal ~ rem_any + country + marketer
##   Res.Df    RSS  Df Sum of Sq     F    Pr(>F)
## 1  13539 3115.4
## 2  13556 3138.6 -17   -23.228 5.938 6.808e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It looks like the introduction of covariates does improve the fit of the model. My null hypothesis was that introducing the covariates does not affect the outcome, however running the F test and looking at the p-value, we can reject the null hypothesis.

d. 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 line to conduct a test with robust standard errors
#
coeftest(mod_pooled_no_covariates, vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.8423585  0.0193621  43.5055  < 2e-16 ***
## rem_any      0.0183253  0.0089205   2.0543  0.03997 *
## country     -0.1490599  0.0093892 -15.8757  < 2e-16 ***
## marketer    -0.0197339  0.0016261 -12.1360  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(mod_pooled_with_covariates, vcovHC)
```

```
##
## t test of coefficients:
##
##                         Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)           0.84677858  0.04295916 19.7112 < 2.2e-16 ***
## rem_any               0.02827780  0.00910189  3.1068 0.0018951 **
## country              -0.16247643  0.02150501 -7.5553 4.447e-14 ***
## marketer             -0.00952299  0.00232313 -4.0992 4.170e-05 ***
## female                0.01798648  0.00877009  2.0509 0.0402971 *
```

```
## age                   0.00129571  0.00033757  3.8384 0.0001244 ***
## highschool_completed -0.00195829  0.00937264 -0.2089 0.8345004
## wealthy              -0.06460989  0.01498812 -4.3107 1.639e-05 ***
## married               0.03579722  0.01325761  2.7001 0.0069399 **
## saved_formal         -0.02618312  0.01175742 -2.2269 0.0259673 *
## inc_7d               -0.00017738  0.00018353 -0.9665 0.3338184
## saved_asmuch         -0.01498470  0.02537600 -0.5905 0.5548610
## spent_b4isaved       -0.00838451  0.02752293 -0.3046 0.7606470
## depart               -0.00623297  0.00264641 -2.3553 0.0185243 *
## provincia            -0.02238394  0.01486235 -1.5061 0.1320692
## branch               -0.02082471  0.01362809 -1.5281 0.1265178
## joint                -0.02793139  0.02735135 -1.0212 0.3071747
## joint_single         -0.06761308  0.02562860 -2.6382 0.0083446 **
## dc                   -0.00264231  0.02529684 -0.1045 0.9168121
## highint              -0.03630226  0.02571193 -1.4119 0.1580071
## rewardint             0.01137899  0.02677846  0.4249 0.6708939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e. 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 is that each of those coefficients have no effect on the outcome of our model. We can see from our table that we can reject the null for some and fail to reject others.

f. 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
`marketer`	Significant
Lives in Bolivia	Not Significant
Lives in Peru	Not Significant
`female`	Significant
`age`	Significant
`highschool_completed`	Not Significant
`wealth`	Significant
`married`	Significant
`saved_formal`	Significant
`inc_7d`	Not Significant
`saved_asmuch`	Not Significant
`spent_b4isaved`	Not Significant

g. Interpret the meaning of the coefficient estimated on the `rem_any`. (We will talk about this more in a later unit, but this is the treatment effect from this experiment).

rem_any means “randomly assigned to any mail or SMS message”. This coefficient means that randomly assigning to any email or SMS does effect the outcome of our model due to it having a significant p-value.

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

Age coefficient means that age does effect the outcome of our model due to it have a significant p-value.

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

This coefficient means having completed the high school does not effect the outcome of our model due to it having a NOT significant p-value.