Simulates vs Real World Data

```
library(dplyr)
library(broom)
library(ggplot2)
library(patchwork)
library(sandwich)
library(lmtest)
```

1. Simulated Data

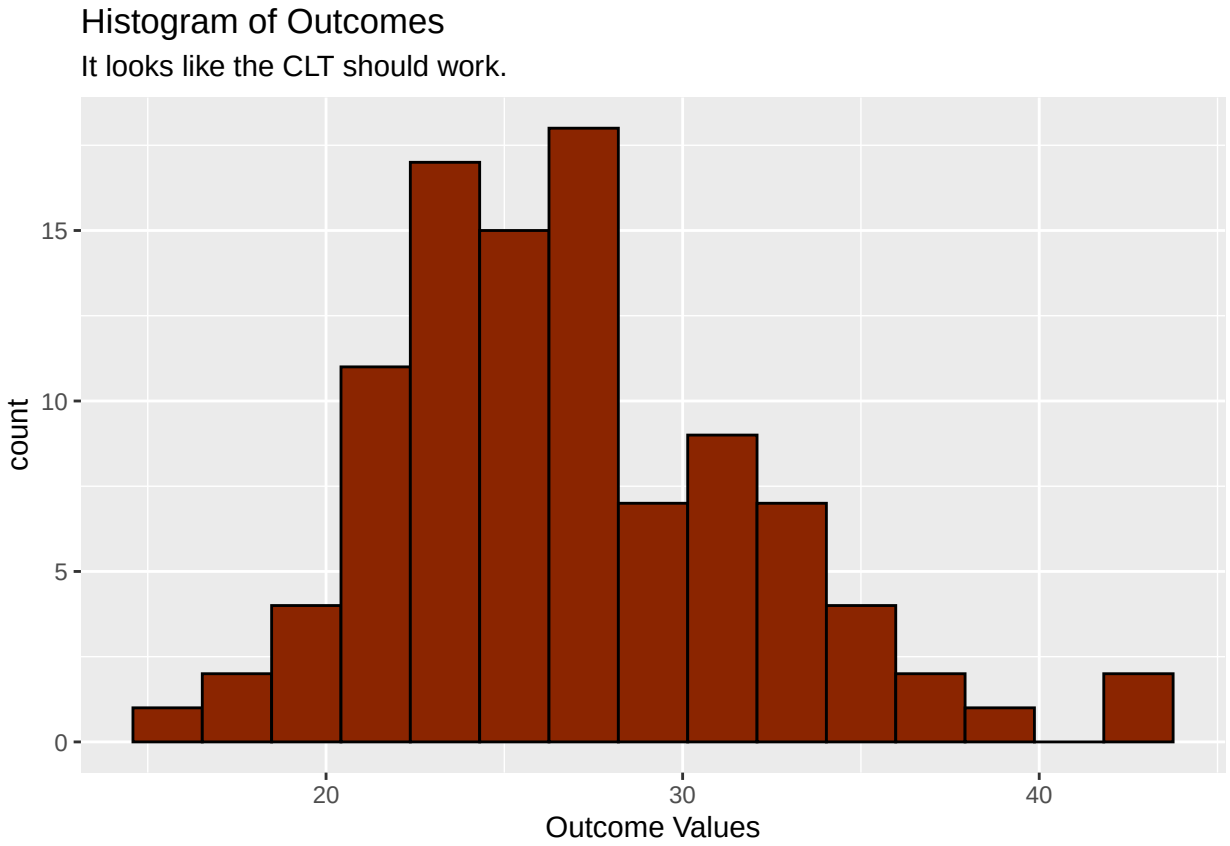
For this question, we are going to create data, and then estimate models on this simulated data. This allows us to effectively *know* the population parameters that we are trying to estimate. Consequently, we can reason about how well our models are doing.

```
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 + 2*x2 + .25*x3^2 + rnorm(n=n, mean=0, sd=1)
    )

  return(d)
}
```

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

1. Data Exploration

```
outcome_histogram <- d %>%
  ggplot() + # fill in the rest of this chunk to create a plot
  aes(x = y) +
  geom_histogram(bins = 15, fill = "orangered4", color = "black") +
  labs(
    x = "Outcome Values",
    title = "Histogram of Outcomes",
    subtitle = "It looks like the CLT should work."
  )
outcome_histogram
```



This distribution looks reasonably well behaved. Because we have more than 30 data points, I think that the CLT should work, and this data will satisfy the assumptions necessary for the large-sample model to produce consistent estimates.

The only two OLS assumptions are (i) iid data which is satisfied by problem setup; and (ii) reasonably well-behaved data so that the CLT works. These are both met in this data. 1. Estimate two models, called `model_1` and `model_2` that have the following form (The only difference is that the second model has squared $x3$):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{1}$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^2 + \epsilon \tag{2}$$

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

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

  msr <- mean(resid(model)^2)

  return(msr)
}
calculate_msr(model_1)
```

```
## [1] 2.611361
```

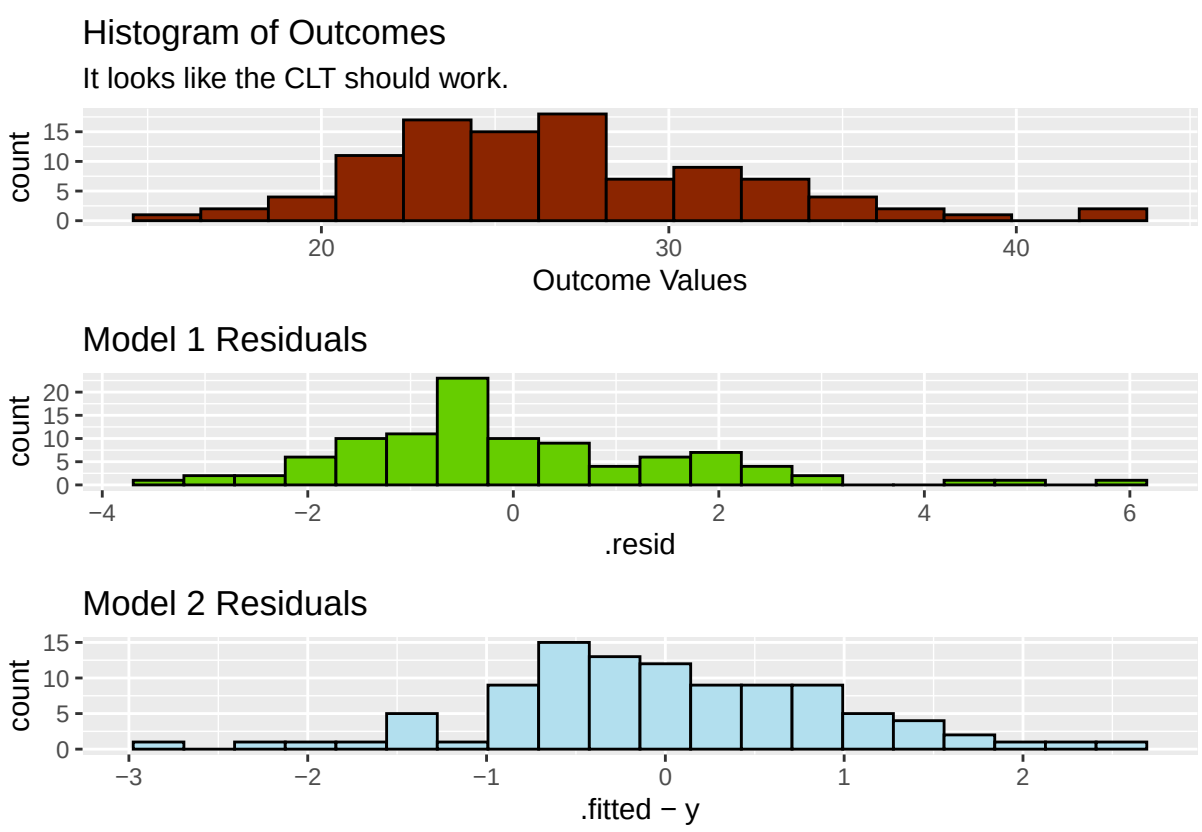
```
calculate_msr(model_2)
```

```
## [1] 0.9054898
```

The second model is doing a better job, the model that has the correctly specified x_3^2 term is fitting better. It has a lower MSE.

```
# ?broom::augment # if you would like to view the documentation
d_model_1 <- augment(model_1)
d_model_2 <- augment(model_2)
```

```
model_1_residuals_histogram <- d_model_1 %>%
  ggplot() +
  aes(x = .resid) +
  geom_histogram(bins = 20, fill = "chartreuse3", color = "black") +
  labs(title = 'Model 1 Residuals')
model_2_residuals_histogram <- d_model_2 %>%
  ggplot() +
  aes(x = .fitted - y) +
  geom_histogram(bins = 20, fill = "lightblue2", color = "black") +
  labs(title = 'Model 2 Residuals')
outcome_histogram /
  model_1_residuals_histogram /
  model_2_residuals_histogram
```



One thing that I notice right away is that both of the models have much smaller variance in the residuals than in the outcome overall. This is good! It means that the model is working to predict outcomes. The other thing that I notice is that the residuals seem to be centered at zero, while the outcome is not. This is a guarantee that comes from Theorem 4.1.5, and something that I

was hoping to see in this data.

Finally, I notice that when I compare Model 2 residudals to Model 1 residuals, there seems to be lower variance/dispersion in the Model 2 residuals. We saw this above, when we calcluated the MSE for the two models, but this is anothe way to see this result: Although we have a guarantee that each of these etimates are the BLP given a certain set of data, the model that more closely matches the underlying form of the data will be able to predict with smaller residuals.

Beause the question asked: Mean of Y: 26.8076811 Mean of Model 1 residuals: $-5.7021556 \times 10^{-15}$ Mean of Model 2 residuals: $7.5850251 \times 10^{-15}$

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.

A. Read the data

Read in the data using the following code.

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

```
## Rows: 13,560
## Columns: 62
## $ motivo                    <chr> "COMPRA DE TERRENO", "OTROS", "EMERGENCIA~
## $ depart                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ quant_saved               <dbl> 13.09080, 39.27240, 294.54300, 58.90860, ~
## $ age                       <dbl> 28, 0, 0, 0, 0, 43, 52, 52, 34, 46, 0, 26~
## $ saved_formal              <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1,~
## $ incentive                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ female                    <dbl> 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1,~
## $ married                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,~
## $ highschool_completed      <dbl> 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,~
## $ reached_b4goal            <dbl> 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1,~
## $ bolivia                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ rem_any                   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1,~
## $ provincia                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0,~
## $ wealthy                   <dbl> 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,~
## $ puzzle_ica                <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0,~
## $ foto                      <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ rem_motive                <dbl> 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,~
## $ peru                      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0,~
## $ branch                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 3,~
## $ marketer                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 2,~
## $ joint                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
## $ joint_single              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ dc                        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ highint                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,~
## $ rewardint                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_inc               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA~
## $ inc_7d                    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,~
```

```
## $ hyperbolic                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_savamt               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA~
## $ missing_havesaved            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA~
## $ missing_savedwhere           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA~
## $ missing_savwhynot            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA~
## $ missing_satisfied            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA~
## $ missing_satisfy              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA~
## $ missing_regretspend          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA~
## $ saved_asmuch                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
## $ spent_b4isaved               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,~
## $ philippines                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,~
## $ country                      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1, 3,~
## $ late_rem_any                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1,~
## $ log_quant_saved              <dbl> 2.645522, 3.695666, 5.688814, 4.092820, 4~
## $ rem_any_peru                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0,~
## $ rem_any_phil                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,~
## $ rem_any_boli                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_female               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_age                  <dbl> 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0,~
## $ missing_highschool_completed <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_married              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_saved_formal         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_inc_7d               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_wealthy              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_hyperbolic           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_saved_asmuch         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_spent_b4isaved       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ missing_number_account       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ ivhquant_saved               <dbl> 3.266513, 4.363831, 6.378575, 4.769207, 5~
## $ logalt_quant_saved           <dbl> 2.572673, 3.670777, 5.685459, 4.076157, 4~
## $ gain_rem                     <dbl> 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1,~
## $ loss_rem                     <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ noincentive                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ rem_no_motive                <dbl> 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0,~
## $ masterid                     <dbl> 1246, 2229, 1490, 943, 1922, 1931, 212, 2~
```

B. F-test

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: *because treatment was assigned at random, it should not be possible to predict something random with other data.*

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

- First estimates a model that regresses treatment using only a regression intercept, $Y \sim \beta_0 + \epsilon_{short}$. In `lm()`, you can estimate this by writing `lm(outcome_variable ~ 1)`, where `outcome_variable` is the outcome that you're actually testing.
- Then estimates a model that regresses treatment using all data available on hand, $Y \sim \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_{long}$.

To test whether the long model has explained more of the variance in Y than the short model, we conduct an F-test for the long- vs. short-models.

> The null hypothesis is that there is no difference in the residual sum of squares between the two models. That is, the longmodel with the additional estimates will not perform any better than

the short model where these estimates are restricted to be zero.

If the p-value for an F-test were to be lower than 0.05, I would reject the null hypothesis.

c. Using variables that indicate:

d. sex (as noted in the codebook);

ii. age;
iii. high school completion;
iv. wealth;
v. marriage status;
vi. previous formal savings (`saved_formal`, which isn't in the codebook);
vii. weekly income;

viii. meeting savings goals (`saved_asmuch`)
ix. and, spend before saving

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`). To do so we filter/subset the data to include only individuals who live in the Philippines, and then estimate a long- and a short- model.

```
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(long_model, short_model, test = 'F')
```

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

I fail to reject the null hypothesis because the p-value for this test is larger than the criteria I set set for rejecting.

At stake was whether the randomization was conducted correctly. Because I fail to reject the null hypothesis for this F-test, I ultimately conclude that there is no evidence that this randomization was performed incorrectly. Notice, that this doesn't necessarily mean that it **was** conducted correctly, just that there isn't any evidence that it was done improperly.

C. Reproduce OLS regression estimates

First, we reproduce the OLS regression estimates

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 are 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.

> One of the difficulties of this data is that the outcome is a binary indicator for whether the individual met (or did not meet) their commitments. At first blush, this doesn't seem like it meets the requirements of the large-sample model – after all, a bernoulli RV is *certainly* not normally distributed. But, in the large sample there is actually no requirement that the data be normally distributed. Only, that it is sampled iid and that the CLT might work. > In section 2 of this paper the authors note that individuals were recruited via door-to-door canvassing in the Phillippines; in Peru individuals were recruited via TV and radio ads; and in Bolivia the product was again marketed on TV and radio ads.
>
> And so, we have a rather careful statment that we have to make about this data. On the one hand, the people who are influenced by this marketing, and thereby come to the bank, are probabilty not either (a) independent – they might come with family members or friends, or see the ads at a bar; and (b) nor identically distributed to the rest of the population who do not come to the bank to sign up.
>
> However, because of the random assignment into treatment conditions we have an iid sample, within a subset of the population. Namely, within the population who come to the bank to sign up for a savings product. b. The authors have concluded that they can conduct these regressions.

```
model_pooled_no_covariates <- lm(
  reached_b4goal ~ rem_any + highint + rewardint + joint +
    dc + joint_single + factor(country),
  data = d)
model_pooled_with_covariates <- lm(
  reached_b4goal ~ rem_any + highint + rewardint + joint +
    dc + joint_single +
    female + age + highschool_completed + wealthy + married +
    saved_formal + inc_7d + saved_asmuch + spent_b4isaved +
    missing_female + missing_age + missing_highschool_completed +
    missing_married +  missing_saved_asmuch + missing_spent_b4isaved +
    factor(depart) + factor(provincia) + factor(marketer) + factor(branch) +
    factor(country),
  data = d)
```

```
stargazer::stargazer(
  model_pooled_with_covariates, model_pooled_no_covariates,
  omit = c('depart', 'provincia', 'marketer', 'branch', 'country'),
  omit.labels = c(
    'Department Fixed Effects', 'Province Fixed Effects',
    'Marketer Fixed Effects',   'Branch Fixed Effects',
    'Country Fixed Effects'),
  single.row = TRUE,
  type = 'latex')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jan 13, 2022 - 16:53:53

```
calculate_msr(model_pooled_no_covariates)
```

```
## [1] 0.2305645
```

```
calculate_msr(model_pooled_with_covariates)
```

```
## [1] 0.219986
```

Table 1:

	Dependent variable:	
	reached_b4goal	
	(1)	(2)
rem_any	0.032*** (0.009)	0.032*** (0.009)
highint	−0.010 (0.031)	−0.002 (0.031)
rewardint	0.044 (0.031)	0.046 (0.032)
joint	0.0005 (0.031)	0.004 (0.032)
dc	0.011 (0.029)	0.012 (0.030)
joint_single	−0.038 (0.030)	−0.035 (0.031)
female	0.016* (0.009)	
age	0.002*** (0.0004)	
highschool_completed	−0.038*** (0.010)	
wealthy	−0.033* (0.017)	
married	0.018 (0.013)	
saved_formal	−0.026** (0.012)	
inc_7d	0.0001 (0.0002)	
saved_asmuch	−0.030 (0.037)	
spent_b4isaved	−0.0001 (0.034)	
missing_female	0.004 (0.145)	
missing_age	0.099** (0.039)	
missing_highschool_completed	−0.109 (0.480)	
missing_married	0.038 (0.131)	
missing_saved_asmuch	−0.084 (0.083)	
missing_spent_b4isaved		
Constant	0.250*** (0.088)	0.663*** (0.012)
Department Fixed Effects	Yes	No
Province Fixed Effects	Yes	No
Marketer Fixed Effects	Yes	No
Branch Fixed Effects	Yes	No
Country Fixed Effects	Yes	Yes
Observations	13,560	13,560
R^2	0.110	0.067
Adjusted R^2	0.106	0.067
Residual Std. Error	0.470 (df = 13503)	0.480 (df = 13551)
F Statistic	29.782*** (df = 56; 13503)	121.900*** (df = 8; 13551)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
anova(model_pooled_with_covariates, model_pooled_no_covariates, test = 'F')
```

```
## Analysis of Variance Table
##
## Model 1: reached_b4goal ~ rem_any + highint + rewardint + joint + dc +
##     joint_single + female + age + highschool_completed + wealthy +
##     married + saved_formal + inc_7d + saved_asmuch + spent_b4isaved +
##     missing_female + missing_age + missing_highschool_completed +
##     missing_married + missing_saved_asmuch + missing_spent_b4isaved +
##     factor(depart) + factor(provincia) + factor(marketer) + factor(branch) +
##     factor(country)
## Model 2: reached_b4goal ~ rem_any + highint + rewardint + joint + dc +
##     joint_single + factor(country)
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1  13503 2983.0
## 2  13551 3126.4 -48   -143.44 13.528 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including these extra "baseline" covariates is useful at improving the model performance. The null hypothesis is that there is not difference in the RSS of the two models; a null hypopthesis that I would reject if the p-value were lower than 0.05 for an F-test. Here, the p-value is lower than 0.05, and so I reject. This means, that the additional covariates are improving the model performance. 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.

```
coeftest(model_pooled_no_covariates, vcov. = vcovHC(model_pooled_no_covariates, type = 'HC1'))
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)        0.6628472  0.0119336  55.5445 < 2.2e-16 ***
## rem_any            0.0318551  0.0091531   3.4802 0.0005025 ***
## highint           -0.0016874  0.0256233  -0.0659 0.9474939
## rewardint          0.0461544  0.0269002   1.7158 0.0862280 .
## joint              0.0036252  0.0273634   0.1325 0.8946029
## dc                 0.0119754  0.0250280   0.4785 0.6323148
## joint_single      -0.0354484  0.0255492  -1.3875 0.1653255
## factor(country)2 -0.1143957  0.0107312 -10.6601 < 2.2e-16 ***
## factor(country)3 -0.4870514  0.0261962 -18.5924 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(model_pooled_with_covariates, vcov. = vcovHC(model_pooled_with_covariates, type = 'HC1'))[1:16
```

```
##                            Estimate   Std. Error      t value     Pr(>|t|)
## (Intercept)           2.498270e-01 0.0672883049  3.712784240 2.058162e-04
## rem_any               3.183629e-02 0.0089423905  3.560154798 3.719099e-04
## highint              -9.970223e-03 0.0256875905 -0.388133816 6.979231e-01
## rewardint             4.358375e-02 0.0268423804  1.623691624 1.044650e-01
## joint                 4.820782e-04 0.0273952056  0.017597173 9.859605e-01
## dc                    1.080298e-02 0.0249442351  0.433085133 6.649598e-01
## joint_single         -3.773627e-02 0.0254452916 -1.483035342 1.380884e-01
```

```
## female                1.579625e-02 0.0086248161  1.831488053 6.704975e-02
## age                   1.737994e-03 0.0003594338  4.835364702 1.343665e-06
## highschool_completed -3.769042e-02 0.0096628562 -3.900546853 9.644067e-05
## wealthy              -3.257150e-02 0.0162773598 -2.001030857 4.540904e-02
## married               1.771074e-02 0.0132454866  1.337115295 1.812075e-01
## saved_formal         -2.565272e-02 0.0115623416 -2.218644370 2.652749e-02
## inc_7d                8.667443e-05 0.0001693684  0.511751018 6.088336e-01
## saved_asmuch         -3.031777e-02 0.0318521013 -0.951829370 3.412006e-01
## spent_b4isaved       -1.141808e-04 0.0289497825 -0.003944099 9.968531e-01
```

For each of the tests, the null hypothesis is that the coefficient = 0. The alternative hypothesis is that the coefficient does **not** equal zero (<.05).

Model 1: Individuals who receive any reminder, and those who live in Bolivia or the Phillippines (compared to Peru) are statistically significantly different from zero. **Model 2**: The coefficients that were significant in Model 1 continue to be significant. Also, Females were more likely to meet their commitment, so are older individuals. Wealthy individuals are less likely to meet their commitments, so too are individuals who have previously saved formally. Finally, those who are married are more likely to meet their commitments. | **Variable** | **Determination** | |————————|————————-| | `rem_any` | Significant | | `marketer` | Not Significant | | `Bolivia` | Significant | | `Peru` | 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 |

Individuals who were randomly assigned to receive a reminder message were 3.2 percentage points more likely to meet their savings goals than those who were not assigned to receive such messages. This effect is statistically significant. Because these reminder messages were experimentally assigned, it is very, very easy to satisfy the "ceteris peribus" nature of the interpretation of this coefficient. For every year older was a participant, they are 0.0012, or 0.12 percentage points, more likly to meet their commitments. Compared to someone who is 18, someone who is 68 (i.e. 50 years older) is about 6 percentage points more likely to meet their savings goal. However, this requires a really strong assumption that these two individuals are otherwise identical in terms of their covariate profiles. i. Interpret the meaning of the coefficient estimated on `highschool_completed`.

The rather crass interpretation is that there is no evidence that completing highschool has any effect on meeting one's savings goals. To state this more precisely, for two individuals with the same covariate profile, the model predicts that someone who has completed highschool will be 0.002 more likely to meet their savings goal than someone who has not completed highschool. This result is not statistically significant.