Indicate the Following:

Name: Chiara Luna Pilato Moncada

SID: 3037208500

GSI: Sara Johns

Midterm 2023 - EEP/IAS 118 - Villas Boas

Instructions: Please answer questions in the boxes and spaces provided. This is an open book midterm. You must solve the exam alone. Any cooperation will be penalized with a 0 in this midterm.

To receive credit:

1. Show your work. e.g., In interpreting the regression coefficient of X on Y, via our SSS framework, if you determine that there is statistical significance at 1% significance level, demonstrate how you came to such conclusion.
2. *Do not* round your answer or intermediate steps to less than four decimal digits
3. Be sure all writing is visible and legible on the scanned pdf
4. Submit to Gradescope as one pdf; Assign all and only the appropriate pages to each question

This exam **does not require R**. R may be used as a calculator or to obtain critical values from statistical tables, but no credit will be given for the use of confidence interval or hypothesis test functions. If R is used to complete steps on a problem, make sure to include the utilized code. (For hand written exam, if you are finding critical values by R or table, either write down the R code or give us all the necessary ingredients for finding it, such as distribution assumed, alpha or alpha/2, and/or degrees of freedom.) **Final answers must be placed in the appropriate text/Markdown cell** - text/comments in coding cells will not be graded. **70 Points Total**

In [17]:
```
library(readxl)
library(tidyverse)
library(psych)
library(ggplot2)
library(haven)
library(data.table)
library(dplyr)
library(foreign)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.
3.0 ──

✔ ggplot2 3.4.0     ✔ purrr   1.0.1
✔ tibble  3.1.8     ✔ dplyr   1.1.0
✔ tidyr   1.3.0     ✔ stringr 1.5.0
✔ readr   2.1.3     ✔ forcats 1.0.0

── Conflicts ────────────────────────────────────────── tidyverse_conflicts
() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()


Attaching package: ‘psych’


The following objects are masked from ‘package:ggplot2’:

    %+%, alpha



Attaching package: ‘data.table’


The following objects are masked from ‘package:dplyr’:

    between, first, last


The following object is masked from ‘package:purrr’:

    transpose
```

Question 1 (Total: 10 Points)

We are interested in the (true) proportion of Americans that believe that gun laws should be stricter, call it p. A sample was collected in a March 2019 AP-NORC Poll with N=1,063 adults. We are given a confidence interval [0.6744, 0.7456]. Given this confidence interval, what is $\hat{p}$? What is the standard error of $\hat{p}$? What is the confidence level of the confidence interval? Show all of your work.

In [23]:
```r
# Include any code used for Q1 here only. (Coding Cell) Final answers do not

#We know what the confidence interval is,
#but we dont know the proportions of these values.


#finding p_hat

p_hat<-(0.6744+0.7456)/2
p_hat

#Calculate the Var
```

```r
var<-(p_hat*(1-p_hat))/1063
var

#Calculate the SE(p_hat)
sqrt(var)

#Finding confidence interval. I will work backwards

#and see what the right level is


#99 lower and upper
0.71-2.576*0.0139175099685722
0.71+2.576*0.0139175099685722

#95 lower and upper
0.71-1.96*0.0139175099685722
0.71+1.96*0.0139175099685722

#90 lower and upper
0.71-1.645*0.0139175099685722
0.71+1.645*0.0139175099685722
```

0.71
0.000193697083725306
0.0139175099685722
0.674148494320958
0.745851505679042
0.682721680461598
0.737278319538402
0.687105696101699
0.732894303898301

➡️ **Write Solution Here:**

- $\hat{p}$ = 0.71
- SE of $\hat{p}$ = 0.0139
- Confidence level is:
    1. 99%

➡️ Show your work here. Round answers to four decimal digits. (Markdown Cell)

We have N=1,063 & a Confidence Interval: [0.6744,07456] To get the P_hat, we just get the middle point of this interval doing (lower bound+upperbound)/2 --> (0.6744+0.7456)/2. This equation gave me an estimate of 0.71 for p_hat.

To calculate the se of p_hat we do

Var(p_hat) = (p_hat(1-p_hat))/n = (0.71(1-0.71))/1063 = 0.000193697083725306, rounded up to 4 decimal places would be 0.0002 for our var(p_hat)

se=sqrt(var(p_hat)) ->sqrt0.000193697083725306 = 0.0139175099685722, rounded up to 4 decimal places would be 0.0139 for our se(p_hat)

To calculate the confidence level, I worked backwards and used the formula to find the condfidence interval with each of the z values to see which one was the best fit for the given model.

I used the formula:

CI=p_hat +- critical value *(se(p_hat))

Then I plugged in the z value for 99%, 95% and 90% confidence.

To calculate the #99 lower and upper 0.71-2.576*0.0139175099685722 for lower bound and 0.71+2.576*0.0139175099685722 for upper bound

For 99% I got (0.6741, 0.7458), which is right and matches the interval we have up in the question

To calculate the #95 lower and upper 0.71-1.96*0.0139175099685722 for lower bound and 0.71+1.96*0.0139175099685722 for upper bound

For 95% I got (0.6827,0.7373), which is not right because it doesnt match the interval given in the question

To calculate the #90 lower and upper 0.71-1.645*0.0139175099685722 for lower bound and 0.71+1.645*0.0139175099685722 for upper bound

For 90: I got intervals of (0.6871,0.7329) which were not right because it doesnt match the interval given in the question

Given these numbers, I said that the confidence level is 99% with an interval of (0.6741, 0.7458)

Question 2 (Total: 30 Points)

In question 2, we employ a dataset that records the following variables for each of the 51 states in January 2023:

Variable	Variable Description (Units)
regularPrice (regular)	Price of regular gasoline in the state (USD)
winteroxygenregul	Indicator variable for whether the state has winter environmental regulation
	= 1, if state has winter environmental regulation
	= 0, otherwise

Variable	Variable Description (Units)
CarsPc (pccars)	Number of cars per capita in the state
Percentleanrep2022	Percent of population leaning republican in the state (%)

Test the null hypothesis that the marginal effect of having environmental winter regulation is larger than 17 cents per gallon for regular gasoline prices, that is $\beta_1 > 0.17$, at the 5% significance level.

Show your work by explicitly following the five steps in your hypothesis testing for a one-sided alternative.

(a) (10 points) We regress the price of regular gasoline on a constant and an indicator of whether the state has winter environmental regulation, using the following commands, and obtain the following output:

STATA: reg regularPrice winterRegulation

R: lm(regularPrice ~ winterRegulation, df = dataset)

N=51	*Rest of Regression Table Deleted*					
	Coef	**Std. Error**	**t**	**P > |t|**	**[95% Conf. Interval]**	
winterRegulation	.1810714	Redacted	1.68	0.100	-.0358157	.3979585
Constant	3.198929	.0724784	44.14	0.000	3.053278	3.344579

Test the null hypothesis that the marginal effect of having environmental winter regulation is larger than 17 cents per gallon for regular gasoline prices, that is $\beta_1 > 0.17$, at the 5% significance level.

Show your work by explicitly following the five steps in your hypothesis testing for a one-sided alternative.

In [19]:
```
# Include any code used for Q2-(a) here. (Coding Cell) Final answers do not
#Calculating our statistics

#Calculating the std error given the t value
std_error<-(0.1810714-0.17)/1.68

print(paste0("standard error is ", std_error))

#Calculating t given the previous standard error under the null
t<-(0.1810714-0.17)/std_error
print(paste0("t value under null hypothesis is ", t))

#Significance level of 0.05

#Our Degrees of freedom are N-1(constant)-k
51-2


qt(0.95,49)
```

```
[1] "standard error is 0.00659011904761904"
[1] "t value under null hypothesis is 1.68"
```

49

1.67655089261685

For the 5 steps of the hypothesis testing we have: 1) Establishing the null and alternative hypothesis. Then we have step 2) Getting the test statistic. Then in step 3)Choosing a significance level and getting the critical value. Step 4)Is to see if we reject or fail to reject. Finally step 5) would be to analyse our results and conclude.

In this case, because we know we have to do a one sided alternative, as its given to us in the question. So our Null Hypothesis would be that beta1=17 and our alternative would be that beta1> 0.17 . We make the null that Beta1 is the same as 0.17 because we are looking for evidence against the fact that the marignal effect of having winter regulation is larger than 17 cents per gallon in regular gasoline prices.

Second Step: Calculating the statistics under the null. We know t=(beta1_hat-beta1(null))/se(beta1_hat) From looking at the table we know that beta1_hat=0.1810714. We also know from our null that beta1(null) would be 0.17. To calculate the standard error we can manipulate the t equation and solve for std error instead of t (because here we have a t). Solving for std error we get (0.1810714-0.17)/1.68 = 0.00659011904761904. Manipulating that same equation and solving for t (t=0.1810714-0.17)/std_error) we get that our t value is equal to 1.68 (also on the table)

Third Step: Significance Level and Critical Value. We have the signficance level of 5% (alpha=0.05) from the question. Then looking at the t-table for a one sided tests we get that the c value is around 1.67. Its around there because there is no 49 df in the table, so we estimate from the two possible: 1.684 and 1.671 (for df 40,60) Additionally, doing the function qt as above (qt(0.95,49)) we get that c is 1.6765.

Fourth Step: Rejecting or failing to reject. We reject the null hypothesis if our t>c and we fail to reject the null if t<c.

In our case we have 1.68 > 1.67 (t>c) so we reject the null.

Fifth Stept In this case, we reject the null hypothesis. We have statistical evidence at the 5% significance level that the marginal effect of having environmental winter regulation is larger than 17 cents per gallon of regular gas prices.

(b) (10 points) Refer to the output below (not from other parts). Can you reject the null at 5% significance level that the Number of cars per capita (**CarsPc**) in a state does not affect regular gasoline prices holding winter regulation constant?

$$\widehat{regularPrice} = \underset{(0.25)}{3.36} \underset{(0.24)}{- 0.16\ \mathbf{carsPc}} + \underset{(0.08)}{0.16\ \mathbf{winteroxygenregul}} \qquad R\ squared=0.06$$

Give the shortest answer being as correct and complete as possible.

YES, because... *or* NO, because...

➡️ Type your answer / steps for Q2-(b) here. (Markdown Cell)

Here our t value for carsPc is calculated by -0.16/0.24 = -0.666. Going to our T value table for a 5% significance level we get -2.021.

In two tailed model, we reject if |t|>|c|, and we have |-0.666| and |-2.021| ---> 0.666 and 2.021, so 0.666 < 2.021. In this case we fail to reject the null.

To answer the question:

No, because as seen in the code above, we fail to reject at a 5% significance level that cars per capital in a state do no affect the regular gasoline prices holding winter regulation constant. This is because the t value that we got for it was -0.666, and our critical value was -2.021, which once put into the absolute value equation gave us that our t value was greater than our critical value. Because of this, we reject this.

(c) (10 points) Given the two regression outputs below, would you conclude that the regular gasoline price in states is influenced by the number of Cars per capita and the percent leaning republican in the state? Show your work using the five steps in hypothesis testing. Use a significance level of 10 percent, that is $\alpha = 0.10$.

```
. reg regular winteroxygenregul percentleanrep2022 pccars

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(3, 47)        =      5.13
       Model |  1.88134592         3  .627115307   Prob > F        =    0.0037
    Residual |  5.73993734        47  .122126326   R-squared       =    0.2469
-------------+----------------------------------   Adj R-squared   =    0.1988
       Total |  7.62128326        50  .152425665   Root MSE        =    .34947
```

```
. reg regular winteroxygenregul

      Source |       SS           df       MS      Number of obs   =        51
-------------+----------------------------------   F(1, 49)        =      2.81
       Model |  .414014495         1  .414014495   Prob > F        =    0.0998
    Residual |  7.20726876        49  .147087118   R-squared       =    0.0543
-------------+----------------------------------   Adj R-squared   =    0.0350
       Total |  7.62128326        50  .152425665   Root MSE        =    .38352
```

➡️ Type your answer / steps for Q2-(c) here. (Markdown Cell)

For the 5 steps of the hypothesis testing we have: 1) Establishing the null and alternative hypothesis. Then we have step 2) Getting the test statistic. Then in step 3)Choosing a significance level and getting the critical value. Step 4)Is to see if we reject or fail to reject. Finally step 5) would be to analyse our results and conclude.

Step 1) Stating hypothesis null hypothesis H0= Coefficients effect on regular gasoline price =0 alternate hypothesis H1= Coefficients effect on regular gasoline price =/= 0 (not zero)

Second Step: Calculating the statistics under the null. When looking at F tests, we are already given the statistic, we see the F value is 5.14

Third Step: Significance Level and Critical Value. We have the signficance level of 10% (alpha=0.010) from the question. Then looking at the F-table for (3, 47) we see the value is around between 2.84 and 2.76. Its around there because there is no 49 df in the table, so we estimate from the two possible by using the function qf given to us in class: (p,df1,df2) = qf(0.9,3,47,lower.tail=T). The code gave me 2.2041, so that is the value I will be using.

Fourth Step: Rejecting or failing to reject. We reject the null hypothesis if our calculated f>f table value (critical) and we fail to reject the null if our calculated f < f table value

In our case we have have a calculated f calculated value of 5.13 and a f table value of 2.20 so we reject the null hypothesis.

Fifth Stept In this case, we reject the null hypothesis. We have statistical evidence at the 10% significance level that the marginal effect of having percentleanrep and ppcars is larger 0 on regular gasoline prices, when holding all else constant. This makes sense, as the percent of republicans always affects the gas prices, because of political processes. Whatever the republicans wanted to push forwards (either higher or lower prices) will affect the prices. Addtionally, the amount of cars on the streets also affects the price (as it affects the demand and supply levels). This agrees with the R^2 valye of the table aswell, as as we include these two variables we see we are able to explain more of the variation in our model, the R^2 went from 0.0543 to 0.2469

In [20]:
```
# Include any code used for Q2-(c) here. (Coding Cell) Final answers do not

qf(0.9,3,47,lower.tail=T)
```

2.20418239110227

Question 3 (Total: 30 Points)

You are interested in buying a used phone, but before you make a purchase you want to understand what factors determine the price in this market to make sure you get the best deal possible. Luckily, you found a dataset that contains a random sample of used phone sales in your area including the sales price and the characteristics of the phone sold. As a first pass at understanding phone pricing, you run the regression where you regress the price measured in dollars on: (standard errors in parenthesis)

$$\widehat{Price} = \underset{(90)}{235} + \underset{(2.5)}{10}\, memoryGB + \underset{(30)}{100}\, iPhone - \underset{(102)}{300}\, pastDamage \qquad R^2 = 0.37,\ N = 1000$$

(a) (5 points) Interpret the coefficient on the memory in Giga bytes (memoryGB). Remember to discuss Sign, Size, and Significance.

➡️ Type your answer / steps for Q3-(a) here. (Markdown Cell)

As seen in the code above, the coefficient for memory in Gigabytes is 10 (a positive)

Using the SSS model we have:

Sign: The coefficient on memoryGB is positive, a positive 10. Which makes sense, because people want phone with lots of storage and would be willing to pay more for it. As in, higher storage raises the value of phones, because people are able to do and store more things on their phone; hence the higher price the more the memory Gigabytes. Basically, memoryGB has a positive estimated effect on our esitmation of price

Size: This tells us that for every additional gigabyte of memory in a phone, the model predicts that average price for phones increases by $10, holding all else (Iphone and

pastDamage) constant. This is not really surprising as, gigabytes (storage) for a phone is something people really want and look forward to when buying phones, so the more storage gigabytes that a phone has, the more expensive it will be.

Significance: At a 1% significance level, the critical value is 2.581. At a 5% significance level, the critical value is 1.962. Then at the 10% significance level, the critical value is 1.646. This is all assuming that our null hypothesis is memoryGB=0 (H0) and our alternate is that memoryGB=/=0 (H1) We also looked at the table for the two tailed and got our c

Based on our test statistic of 10/2.5=4, we can say that our statistic is statistically significant (greater than 0) at the 1%, 5% and 10% levels. This means B_hatmemoryGB significant and its not zero. We can reject the null hypothesis that the effect of memoryBG is =0 (Its not zero, so our null is not true, so its statistically significant) Our B_hatmemoryGB is statistically greater than 0.

In [24]:
```
# Include any code used for Q3-(a) here. (Coding Cell) Final answers do not

#For significance we use p value,
#which we need the t value for aswell
1000-4

#making the t value

10/2.5
```

996

4

(b) (10 points) Run a two-sided hypothesis test at the 5% significance level that having had past damages (pastDamage=1) does not matter for the prices of a used phone.

➡️ Type your answer / steps for Q3-(b) here. (Markdown Cell)

For the 5 steps of the hypothesis testing we have: 1) Establishing the null and alternative hypothesis. Then we have step 2) Getting the test statistic. Then in step 3)Choosing a significance level and getting the critical value. Step 4)Is to see if we reject or fail to reject. Finally step 5) would be to analyse our results and conclude.

Step 1) Hypothesis:

Ho Null hypothesis: Betapastdamage=0

H1 alternate hypothesis: Betapastdamage=/=0 (not equal to 0)

Step 2) Test statistic We do (-300-H0)/se(beta_hatpastDamage) (-300-0)/10 2= -2.941176

Step 3) Critical Value for two sided test We have significance level of 5% (0.05) for a two tailed test, or we could also say 0.025 in each tail. Looking at the table we get that c=1.960

Step 4) Reject or fail to reject For two tailed testing we say that we reject the null if |t|>|c|. Here we have |-2.9411|>|-1.960| ->2.9411 > 1.960 Because of this, we reject the null hypothesis

Step 5)Conclude We can say that at a 5% significance level we reject the null hypothesis that having past damages does not affect the prices for the used phones, holding all memoryGB and Iphone constant. As in, having past damages in the phone does negatively correlate with the price, in our model, assuming its eather pastDamage=1 or pastDamage=0 (as in phone is damaged or not) we say that if there is damage, the price of phone will decrease.

As extra though, I also think that in real life this pastDamage would probably work on a scale, as in, the more damaged the phone is the less price of the phone.

In [22]:
```
# Include any code used for Q3-(b) here. (Coding Cell) Final answers do not
(-300-0)/-2.941176

lowerbound<- (-300-1.96*0.24)
upperbound<- (-300+1.96*0.24)

lowerbound
upperbound
```

```
102.000016320003
-300.4704
-299.5296
```

(c) (5 points) Suppose after running this analysis you obtain an additional 2,000 random sample observations on used phone sales. You decide to add them to your analysis. How would you expect this new data to change the hypothesis test that you ran in part b)? Back up your claim with reasoning (i.e. relevant equation).

➡️ Type your answer / steps for Q3-(c) here. (Markdown Cell)

Our initial sample size was of 1000, so getting an additional 2,000 random sampes brings up our sample size N to 3,000. Anytime that our sample size increases or decreases, our statistical tests about it will change, Adding more samples and increasing N is always helpful towards our model and help us specify it and make it more precise.

Having bigger sample sizes leads to us having smaller variance in our beta estimates. When estimating our OLS estimators Beta1_hat and beta0_hat for our minimizing of RRS, we will get more accurate results.

σ^2_u=1/n-2(sum of u_hat_i^2)

var(beta0_hat)= σ^2_u/SST_x(sum of x_1/n)

var(beta1_hat)= σ^2_u/SST

As seen in all those equations, n plays a role in the denominator, so increasing that number will reduce the size of σ_hat, and decrease the variance of our beta_hat estimates. This however also adds a larger variance in the x's (which is good for our model)

Because the se(beta1_hat) is also the sqrt of var(beta1_hat), and that is equal to σ_u/sqrt((n-1)s^2_x)), increasing our sample size will also affect the standard error of our beta1_hat

We see the formula var(y)=(yi-y(average))^2/(N-1) or just (SST/(N-1) // Here we see how if N were to increase, the denominator would become bigger, hence making the variance in y smaller.

Also, because our sample size is part of our standard errors formula (s_x/sqrtn), having a smaller sample size decreases the se, which in turn changes our results for our confidence intervals. This is because for calculating confidence intervals we do our sample average +- critical value *standard error

Laslty, I can also see how the increased sample size would affect our t values, as n is part of the equation. t=(sample average-null hypothesis)/sd/sqrtn. Hence, bigger n would lead to a bigger t

Specially in this case where our sample size more than doubled (from 1,000 from 3,000), these changes throughout our hypothesis testing process would be very visible.

(d) (5 points) After some thought, you realize that you left out an important variable: the age of the phone in the number of years since the model was first released (**Age_Years**). Assuming that buyers are willing to pay **less** for phones that have **high age**, how do you expect the inclusion of the variable **Age_Years** to change the coefficient on the memory of the phone (**memoryGB**) from the output in part a), given what you expect the "likely" sign of the correlation (**Age_Years , memoryGB**) to be? Explain your answer.

➡️ Type your answer / steps for Q3-(d) here. (Markdown Cell)

The likely sign of correlation between these two (Age_Years and memoryGB) would be negative, because as the phone gets older, its used more and there is less memory Gigabytes.

Looking at the ommitted variable bias formula E(Beta_memoryGB)=Beta_memoryGB+ (rho*beta_age_years)

rho is the correlation between age_years and memoryGB. Beta_age_years would be the parameter we get on this new included age_years variable part of the properly specified population model. Looking at this formula, we can just start estimating what the signs will look like. Because we dont have any specific numbers we cant give a concrete answer of how the coefficient of memoryGB to change.

Including the age_years would probably make the coefficient of memoryGB to decrease (become less negative), as now there is a new variable taken our of the u.
We then could analyze depending the specifics of this new variable age_years if it causes any bias (positive or negative).