

Indicate the Following:

Name: _

SID: _

Exam Instructions

- This exam will be made available on May 11 at 9:00pm pacific time.
- It MUST be submitted to [Gradescope](#) before May 13th end of day pacific 11:59pm, NO EXCEPTIONS. Submit early enough to deal with any unforeseen technical issues that may arise while submitting. Make sure to select the page(s) that correspond to each problem part!
- This exam should be solved in two hours (three hours max). Upload it to Gradescope once you are done. Do not dither and spend too much time/energy on the exam, it will not improve your grade.
- To receive full credit, answers must include a correct answer, demonstrate all steps used to obtain the answer, and round to four decimal places unless otherwise indicated.
- Final written answers **must be placed in the indicated text/Markdown cell**. Text/comments in coding cells can easily be cut off when exported as a pdf and will not be used as answers when grading. R output provided without words or indication are not considered complete answers.
- 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/output along with your written text answers.
- We are looking for concise accurate responses and explanations. Long wishy-washy responses will not work to your advantage.
- If submitting photos of your handwritten exam, make sure the photos are clearly visible and centered on the page. Your pictures should look the same as if we had run your exam through a scanner.
- This is an open book exam; you may use all your notes and materials from the course.

Honor Code Statement:

"As a member of the UC Berkeley community I act with honesty, integrity, and respect for others."

This is an individual exam. Cases of collaboration/copying will be given scores of zero.

Please print your name below to indicate that you understand that this exam is to be completed alone, without outside assistance, and that you understand the penalty for failing to do so:

Printed Name:

Final 2021 - EEP/IAS 118 - Villas Boas

100 Points Total

Exercise 1. (26 Points)

We estimate the following model where $price_{jt}$ is the price of sandwich j at time t , and is a function of the calories of sandwich j at time t (which is measured as the difference between each sandwich's calories and 40 calories) using a dataset with 602 observations.

The OLS regression line is

$$\widehat{price_{jt}} = 3.5 + 0.185 * (Calories_{jt} - 40)$$

(0.9) (0.044)

Where standard errors of the estimated coefficients are in parentheses.

(a) (4 Points) Please construct a 95% confidence interval of the predicted average price of a sandwich with 40 calories.

→ The degree of freedom for 602 observations is $df = 602 - 2 = 600$

The z-score corresponding to the 95% confidence interval is $z_{\frac{\alpha}{2}} = 1.95996$

Therefore, the 95% confidence interval is $\hat{\mu} \pm 1.95996 * SE$

$\Rightarrow (0.185 - 1.95996 * 0.044, 0.185 + 1.95996 * 0.044) \Rightarrow (0.09876176, 0.27123824)$

In [1]:

```
0.185-1.95996*0.044
0.185+1.95996*0.044
```

0.09876176

0.27123824

(b) (5 Points) Do you find evidence that more calories are associated with higher prices at the 10 percent significance level for a one-sided alternative? Do all 5 steps in hypothesis testing.

→ Step 1: Defined Hypothesis:

$$H_0 : \beta = 0$$

$$H_1 : \beta > 0$$

Step 2: Test Statistics:

$$t = \frac{\hat{\beta} - 0}{SE} = \frac{0.185 - 0}{0.044} = 4.2045$$

Step 3: Critical Value:

The critical value of t for degree of freedom 600 and $\alpha = 0.10$ is $c = 1.282964$

Step 4: Decision:

Since, $t > c$, the null hypothesis is rejected.

Step 5: Conclusion:

It is concluded that More calories are associated with higher price.

In [2]:

```
t_statistics=0.185/0.044
t_statistics
```

4.20454545454546

(c) (5 Points) When we calculate the correlation between actual and predicted prices we obtain a value of 0.34. What is the regression's goodness of fit measure adjusted R-squared's value? Round to four decimal places.

☐ The Coefficient of Determination, $R^2 = r^2 = 0.34^2 = 0.1156$

Adjusted R^2 is calculated using the formula:

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

Where, $N = 602$ is the sample size and $p = 1$ is the number of predictors.

$$\Rightarrow \text{Adjusted } R^2 = 1 - \frac{(1-0.1156)(602-1)}{602-1-1} = 0.1141$$

In [3]:

```
R_square=0.34**2
N=602;p=1;
AdjustedR_square=1-(1-R_square)*(N-1)/(N-p-1)
print(R_square)
print(AdjustedR_square)
```

```
[1] 0.1156
[1] 0.114126
```

(d) (5 Points) We run the regression of prices on (Calories-40) and add the level of saturated fat (measured in grams) to the regression. We note that the coefficient on saturated fat is -0.05 with a standard error of (0.01). Moreover, in that regression the estimated coefficient of (Calories-40) is now 0.06 with a standard error of (0.02). What does this tell you in terms of the correlation between the variable (Calories-40) and the variable saturated fat content of sandwiches?

☐

The sign of estimated Coefficient of Saturated Fat is negative, hence, Saturated Fat is negatively correlated with Price of Sandwich. Additionally, With the inclusion of Saturated Fat, the estimated coefficient of (Calories-40) is 0.06. When the Saturated Fat is omitted, the estimated coefficient of (Calories-40) was 0.185 (Increased).

Due to omitted variable bias, the beta coefficient of (Calories-40) is biased upwards or the bias is positive. Hence, the correlation between Saturated Fat and (Calories-40) is negative.

In general, when the omitted variable A is in negative correlation with the response variable and omission results in positive bias to a variable B, then A and B are negatively correlated.

In [4]:

```
# Include any code used for EX1-(d) here. (Coding Cell) Final answers do not belong
```

(e) (7 Points) We wish to test the null hypothesis that the coefficients on "Calories-40" is equal to 0.03 and the coefficient on "saturated fat" is equal to -0.03 at the 5% level. Please perform the 5 steps in hypothesis testing and conclude, given that the Residual standard error from the

unrestricted regression is 1.9 and the Sum of Squared Residuals (SSR) for the restricted regression is 2850.

☐ The number of Restrictions, $q = 2$

Number of Explanatory Variable, $k = 2$

Number of Observations, $N = 602$

Sum of Square of Residual for unrestricted Regression,
 $SSR_{UR} = 1.9^2 * (N - K - 1) = 1.9^2 * 599 = 2162.39$

Hypothesis Testing

Step 1: Defined Hypothesis:

$$H_0 : \beta_1 = 0.03 \text{ \& } \beta_2 = -0.03$$

$$H_1 : \beta_2 \neq 0.03 \text{ or } \beta_2 \neq -0.03$$

Step 2: Test Statistics:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N - K - 1)} = \frac{(2850 - 2162.39)/2}{2162.39/599} = 95.2368$$

Step 3: Critical Value:

The critical value of F for $\alpha = 0.05$ and degree of freedom $q = 2, N - K - 1 = 599$ is $c = 3.011$

Step 4: Decision:

Since, $F > c$, the null hypothesis is rejected.

Step 5: Conclusion:

It is concluded that the coefficient of either 'calories-40' is not 0.03 or 'Saturated Fat' is not -0.03

```
In [5]: 1.9^2*599  
(2850-2162.39)/(2*2162.39)*599
```

2162.39

95.2368421052632

Exercise 2. (24 Points)

We estimated the following model where $price_{jt}$ is the price of good j at time t , and is a function of the calories ($Calories_{jt}$) of product j at time t using a subset of the same dataset with 36 observations. We now want to control for factors specific to each product and also factors changing year to year that could affect prices common to all products. We have 3 different types of sandwiches and 2 years in the data, and these sandwiches are sold in different regions of the country. Here is the regression you would use to estimate the effect of calories on prices controlling for those factors:

$$price_{jt} = \alpha_0 + \alpha_1 type_1 + \alpha_2 type_2 + \alpha_3 type_3 + \beta Calories_{jt} + \gamma Year_2 + \varepsilon_{jt}$$

(a) (4 Points) Can you estimate all coefficients $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \beta, \gamma$ by OLS? Why or why not? Explain briefly with reference to collinearity.

☐ **All the coefficients can not be Estimated as there is multicollinearity in the types of sandwich**

Explanation:

Each of the coefficients can be estimated if and only if there is no multi collinearity between the predictors Sandwich Type, Calories and Year in the panel data. In the given case, there are three types of Sandwich so including all the three Fixed Effect in the regression equation results in Multi Collinearity. Hence, either of the three type needs to be ommitted.

(b) (4 Points) What is the variable *type1* in this regression? Please define it and explain what it looks like in the data set.

☐

Type 1 is a dummy Variable corresponding to Sandwich Type 1. It is a fixed effect due to Sandwich Type 1. The variable represents the difference in price whether the sandwich is of type 1 or not.

If Sandwich is Type1 then $type1 = 1$ and 0 otherwise.

For parts c) through f) assume that the variable $type_3$ is the omitted group.

(c) (4 Points) Please interpret the meaning and statistical significance of the OLS estimate $\hat{\alpha}_0 = 5$ with a standard error of 1.8. (2-3 sentences max)

☐

As the variable $type_3$ is ommitted. $\hat{\alpha}_0=5$ corresponds to the fixed effect in price due to Type 3 sandwich in the year 1 i.e. $Year_2 = 0$. It is marginal effect of Type 3 sandwich on Price controlling the fixed effect due to Type1 and Type 2 sandwiches and by Controlling the effect due to Calories. Selling of Type 3 sandwich in a region results in increase in price by 5 unit. The t value is 2.77 while the critical t-value is for 30 degree of freedom and 0.05 level of confidence is 2.042272. Hence, the estimate is statistically significant.

In [6]:

```
t=5/1.8
t
```

2.777777777777778

(d) (4 Points) Please interpret the meaning and statistical significance of the OLS estimate $\hat{\alpha}_1 = -3$ with a standard error of 5. (2-3 sentences max)

☐

The estimated coefficient $\hat{\alpha}_1 = -3$ corresponds to fixed effect in price due to Type 1 Sandwich in the Year 1. For selling the type 1 sandwich by Controlling the other types and for a fixed calories and year 1, the price decreased by 3. The t value is -0.6, while the critical t-value is for 30

degree of freedom and 0.05 level of confidence is 2.042272. As $|t| < |c|$, the estimate is not significant.

In [7]:

```
T=-3/5  
T
```

-0.6

(e) (4 Points) Please interpret the meaning and statistical significance of the OLS estimate $\hat{\gamma} = 2.9$ with a standard error of 1.6. (2-3 sentences max)



The estimated coefficient $\hat{\gamma} = 2.9$ corresponds to effect in Year 2 relative to Year 1. The change in price due to the change in year from 1 to 2 controlling all other fixed effect (including sandwich type and effect due to calories). The price of the product increases by 2.9 unit. The t value is 1.812, while the critical t-value is for 30 degree of freedom and 0.05 level of confidence is 2.042272. As $|t| < |c|$, the estimate is not significant.

In [8]:

```
t=2.9/1.6  
t
```

1.8125

(f) (4 Points) Please interpret the meaning and statistical significance of the OLS estimate $\hat{\beta} = 0.02$ with a standard error of 0.002. (2-3 sentences max)



The estimated coefficient $\hat{\beta} = 0.02$ corresponds to effect due to Calories. By controlling all the fixed effect due to type of Sandwiches and by controlling the fixed effect due to year 1, the marginal effect in price of good j in time period t due to Calories is 0.02 unit.

The t value is 10, while the critical t-value is for 30 degree of freedom and 0.05 level of confidence is 2.042272. As $|t| > |c|$, the estimate is significant.

In [9]:

```
t=0.02/0.002  
t
```

10

Exercise 3. (23 Points)

Now assume there are more than 3 sandwiches in our data (using all 602 observations). We learn that a random subset of the sandwiches in our sample were affected by a quasi-experiment. On half of them a sticker was added that explained that the workers in the restaurant earn a living wage because they operate in counties that were subject to a 15-dollar hourly minimum wage regulation. The other half of the sandwiches were served in regions that did not have a minimum wage of 15 dollars an hour (and whose sandwiches did not feature stickers). Moreover, regulation happened in Year 2 and did not happen in Year 1, so we can observe the prices of sandwiches before and after the minimum wage regulation is implemented and the sticker added to the sandwich wrapper in regulated counties. We also have access to

data on the sandwiches' nutritional characteristics and restaurant and employee characteristics in year 1.

(a) (5 Points) Below is a table showing the average prices in Year 1 and Year 2 in areas that received the minimum wage and disclosure regulations and in non-regulated areas. Please complete all the missing pieces in this table. Show any work in the box below and **write all missing values in the table.**

Average Sandwich Prices	Regulated=0	Regulated=1	Difference (Reg-Not Regulated)
Year2 = 0	4.2	6	(iii)
Year2 = 1	5	(ii)	(iv)
Difference (Year2 – Year1)	(i)	3	(v)

(i) = 5-4.2=0.8

(ii) = 3+6=9

(iii) 6-4.2= 1.8

(iv) 9-5= 4

(v) 3-0.8= 2.2

(b) (6 Points) Below is the equation we estimate to measure the causal effect of regulation on sandwich prices.

$$\widehat{price} = \hat{\delta}_0 + \hat{\delta}_1 Year_2 + \hat{\delta}_2 Regulated + \hat{\delta}_3 Regulated * Year_2$$

What are the values of all the estimated coefficients in this price equation given what you know in (a)? Place your answers next to the coefficients in the box below.

$\hat{\delta}_0 = 4.2$

$\hat{\delta}_1 = 0.8$

$\hat{\delta}_2 = 1.8$

$\hat{\delta}_3 = 2.2$

In [10]:

Include any code used for EX3-(b) here. (Coding Cell) Final answers do not belong

(c) (2 Points) What is the impact analysis method used for causal identification in the equation in (b)? (1 sentence max)

The impact analysis method used for causal identification is **Difference in Differences Method.**

(d) (5 Points) Please test the null hypothesis that regulation has no causal effect on prices in a two-sided test at the 1 percent significance level given the provided standard errors: $\widehat{se}(\hat{\delta}_0) = 4.2$, $\widehat{se}(\hat{\delta}_1) = 0.5$, $\widehat{se}(\hat{\delta}_2) = 0.2$, $\widehat{se}(\hat{\delta}_3) = 0.5$. Use the 5 steps of hypothesis testing.

Hypothesis Testing

Step 1: Defined Hypothesis:

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

Step 2: Test Statistics for two tailed test:

$$t = \frac{\hat{\beta}_3 - 0}{SE} = \frac{2.2 - 0}{0.5} = 0.44$$

Step 3: Critical Value:

The critical value of t is $c = 2.548$ for $\alpha = 0.01$.

Step 4: Decision:

Since, $t < c$, the test fails to reject the null hypothesis.

Step 5: Conclusion:

It is concluded that the regulation has no causal effects.

In [11]:

```
# Include any code used for EX3-(d) here. (Coding Cell) Final answers do not belong
```

(e) (5 Points) You do a balance test on the characteristics of regulated and unregulated sandwiches in year 1. The p-values for the equality of averages of your observable characteristics of sandwiches and restaurants in regulated and unregulated areas in year 1 are all greater than 0.4. Are you assured that the quasi-experimental randomization is present given the method you are using to measure the causal effect of regulation on prices? Why or why not, explain briefly. (3-4 sentences max)

☐ No, It is not assured that the quasi experimental randomization is present. Explanation: The p-values for the equality of averages of your observable characteristics of sandwiches and restaurants in regulated and unregulated areas in year 1 are all greater than 0.4, shows that the test fails to reject the null hypothesis and it can be easily inferred that we do not have sufficient evidence to validate the assumption that quasi experimental randomization is present.

Exercise 4. (12 Points)

The R output below shows the summary statistics for prices of 46 footlong sandwiches and 60 panini sandwiches.

```
avg_footlong <- mean(mydata$price[which(mydata$footlong==1)])
```

```
avg_panini <- mean(mydata$price[which(mydata$panini==1)])
```

```
avg_footlong
```

```
[1] 3.44913
```

```
avg_panini
```


[1] 4.22417

```
sd_footlong <- sd(mydata$price[which(mydata$footlong==1)])
```

```
sd_panini <- sd(mydata$price[which(mydata$panini==1)])
```

```
sd_footlong
```

[1] 1.002813

```
sd_panini
```

[1] 1.242687

(a) (8 Points) Test whether the average prices of footlong (footlong=1) sandwiches and panini sandwiches (panini=1) are equal at the 5% level. Use the 5 steps of hypothesis testing and round your answer to 4 decimal places.

Hypothesis Testing

Assuming Equal Variance

Defined Hypothesis:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Step 2: Test Statistics for the two tailed test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = -3.55241$$

Degree of Freedom $df = 46 + 60 - 2 = 104$

Step 3: Critical Value:

The critical value of t for 104 degree of freedom and for $\alpha = 0.05$ is $c = 1.983$

Step 4: Decision:

Since, $|t| > |c|$, the null hypothesis is rejected.

Step 5: Conclusion:

It is concluded that the the average prices of footlong Sandwich and Panini Sandwich are significantly different.

In [12]:

```
#Pooled Variance:
n1=46; n2=60;
s1=1.002813; s2=1.242687;

xbar1=3.44913; xbar2=4.22417;

t=(xbar1-xbar2)/(sqrt(s1^2/n1+s2^2/n2))
print(t)
```

```
df=(46+60-2)
print(df)
```

```
[1] -3.55241
[1] 104
```

(b) (4 Points) Construct a 99 percent confidence interval for the mean of prices for panini sandwiches. Round your answer to 4 decimal places.

☐ The z-score for 99% confidence interval is 2.576. Therefore, the 99% Confidence Interval for the mean of prices for panini sandwiches is:

$$(4.22417 - 2.576 * \frac{1.242687}{\sqrt{60}}, 4.22417 + 2.576 * \frac{1.242687}{\sqrt{60}})$$

$$\Rightarrow (3.8109, 4.6374)$$

In [13]:

```
# Include any code used for EX4-(b) here. (Coding Cell) Final answers do not belong

4.22417-2.576*1.242687/sqrt(60)
4.22417+2.576*1.242687/sqrt(60)
```

```
3.81090180003024
```

```
4.63743819996976
```

Exercise 5. (15 points)

We next want to understand the probability of individuals buying sandwiches made only from organic ingredients.

The R output below corresponds to the linear probability model of whether individuals bought organic sandwiches (`choseOrganic` = 1 or 0) as a function of whether individuals have children less than 6 years (`kidslt6` = 1 or 0) and the level of the individual's education (`educ`) and family income (`faminc`).

```
> reg1Linprobmodel<-lm(choseOrganic~educ+kidslt6+faminc, mydata2)
> summary(reg1Linprobmodel)
```

Call:

```
lm(formula = choseOrganic ~ educ + kidslt6 + faminc, data = mydata2)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.8432	-0.5156	0.2111	0.3944	0.9329

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.309e-02	9.467e-02	0.561	0.575
educ	4.490e-02	8.188e-03	5.484	5.7e-08 ***
kidslt6	-2.230e-01	3.324e-02	-6.708	3.9e-11 ***
faminc	7.204e-07	1.523e-06	0.473	0.636

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4735 on 749 degrees of freedom

Multiple R-squared: 0.09084, Adjusted R-squared: 0.0872

F-statistic: 24.94 on 3 and 749 DF, p-value: 2.167e-15

```
> coeftest(reg1Linprobmodel, vcov = vcovHC(reg1Linprobmodel, "HC1"))
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.3091e-02	9.0014e-02	0.5898	0.5555
educ	4.4900e-02	7.7069e-03	5.8260	8.431e-09 ***
kidslt6	-2.2296e-01	3.0003e-02	-7.4315	2.935e-13 ***
faminc	7.2038e-07	1.5364e-06	0.4689	0.6393

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(a) (2 Points) What is the estimate of the robust standard error for the less than 6 years children (kidslt6) parameter?

→ The estimate of the robust standard error for the parameter kidslt6 is **0.030003**

(b) (3 Points) Please interpret whether education level significantly affects the probability of choosing organic, controlling for an individual having young (less than 6 years old) children and income.

→ The education level significantly affects the probability of choosing Sandwich made of only organic ingredients. The probability of choosing Organic sandwich increases by $0.0449 \times 4 = 18\%$ if the education level of individual increases to 4.

Next, please consider the output below that estimates a logit model and the corresponding marginal effects.

```
> reg2Logit <- glm(choseOrganic~educ+kidslt6+faminc, mydata2, family = binomial(link = "logit"))
> summary(reg2Logit)
```

Call:
glm(formula = choseOrganic ~ educ + kidslt6 + faminc, family = binomial(link = "logit"), data = mydata2)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8598	-1.1974	0.6977	0.9909	2.0502

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.053e+00	4.442e-01	-4.622	3.79e-06 ***
educ	2.041e-01	3.882e-02	5.258	1.45e-07 ***
kidslt6	-1.004e+00	1.630e-01	-6.156	7.48e-10 ***
faminc	3.145e-06	7.018e-06	0.448	0.654

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1029.75 on 752 degrees of freedom
Residual deviance: 958.14 on 749 degrees of freedom
AIC: 966.14

Number of Fisher Scoring iterations: 4

```
> margins <- margins(reg2Logit)
> summary(margins)
```

factor	AME	SE	z	p	lower	upper
educ	0.0455	0.0081	5.6456	0.0000	0.0297	0.0613
faminc	0.0000	0.0000	0.4485	0.6538	-0.0000	0.0000
kidslt6	-0.2237	0.0328	-6.8190	0.0000	-0.2880	-0.1594

(c) (4 Points) Please construct a 90 percent confidence interval for the average marginal effect (AME) of having children less than 6 years old on the probability of choosing organic, controlling for education and income.

☞ Written Solutions for Ex5(c)

The z score for 90% confidence interval is 1.645. Therefore, 90% confidence interval of Children less than 6 year old on the probability of choosing organic is:

$$(-0.2237 - 1.645 * 0.0328, -0.2237 + 1.645 * 0.0328)$$

$$\Rightarrow (-0.277656, -0.169744)$$

In [14]:

```
# Include any code used for EX5-(c) here. (Coding Cell) Final answers do not belong  
-0.2237-1.645*0.0328  
-0.2237+1.645*0.0328
```

-0.277656

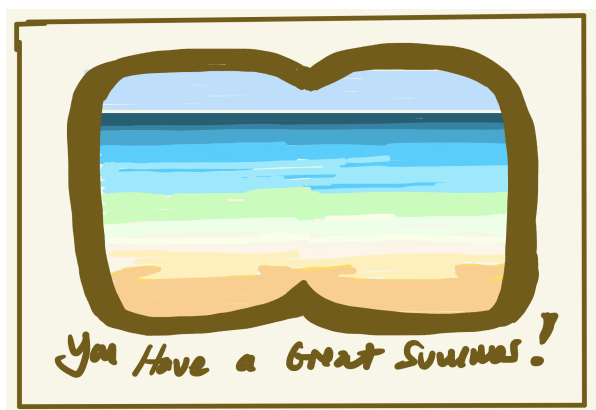
-0.169744

(d) (3 Points) Why is this interval different from the 95 percent confidence interval reported above in the corresponding columns lower and upper? Explain briefly.

☞ The interval is different because the calculated value is for 90% confidence level and the provided result is for the 95% confidence level. As the confidence level decreases, the confidence interval width also decreases which can be seen from the result displayed and calculated confidence interval.

(e) (3 Points) Does a one year increase in education significantly affect the probability of choosing organic in the logit model, all else equal? Explain.

☞ One year increase in education level significantly increases the probability of choosing organic as the probability will increase by 4.55%. Additionally, the p-value is 0 which shows that education significantly affect the probability of choosing organic.



It was a pleasure to teach you this Spring and we value your hard work and focus during this remote semester of econometrics in EEP 118. We really appreciate you filling out the evaluations and giving us feedback on things that worked well and what we can improve.

We hope you have a good end of semester and hope to meet you one day in person in the future.

All the best

Sofia, James, and Sung

From your EEP 118 Spring 2021 team