PS1 Q2

f)

Using R script to run the CLRM.

R code:

library("AER")

data("TeachingRatings")

fit1<-lm(eval~beauty+age+I(age\*\*2)+log(allstudents)+gender+minority+gender:minority,data=TeachingRatings)

summary(fit1)

Running outcomes:

#Call:

#lm(formula = eval ~ beauty + age + I(age^2) + log(allstudents) +

# gender + minority + gender:minority, data = TeachingRatings)

#Residuals:

# Min 1Q Median 3Q Max

#-1.82048 -0.36540 0.07108 0.39548 1.13393

#

#Coefficients:

# Estimate Std. Error t value Pr(>|t|)

#(Intercept) 3.6978272 0.5732534 6.451 2.85e-10 \*\*\*

#beauty 0.1566171 0.0333042 4.703 3.41e-06 \*\*\*

#age 0.0345926 0.0234193 1.477 0.14034

#I(age^2) -0.0003929 0.0002403 -1.635 0.10267

#log(allstudents) -0.0885579 0.0299777 -2.954 0.00330 \*\*

#genderfemale -0.1839788 0.0568162 -3.238 0.00129 \*\*

#minorityyes -0.0164786 0.1108320 -0.149 0.88187

#genderfemale:minorityyes -0.2610383 0.1486546 -1.756 0.07976 .

#---

#Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 0.5284 on 455 degrees of freedom

#Multiple R-squared: 0.1069, Adjusted R-squared: 0.09317

#F-statistic: 7.781 on 7 and 455 DF, p-value: 6.612e-09

Thus a table describing the regression model results could be demonstrated as below:

Table 2.f

CLRM of the determinants of teaching evaluation score

|  |  |
| --- | --- |
| Variable | Coefficient  (Standard Error) |
| beauty | 0.157\*\*\*  (0.033) |
| age | 0.034  (0.023) |
| age2 | −0.0004  (0.0002) |
|  | −0.089\*\*\*  (0.030) |
| gender (female=1) | −0.184\*\*\*  (0.057) |
| minority (yes=1) | −0.016  (0.111) |
| female\_minority | −0.261\*  (0.149) |
| (constant) | 3.698\*\*\*  (0.573) |
| R2 (adjusted) | 0.093 |

Notes: \* when 0.05≤p<0.1, \*\* when 0.01≤p<0.05, and \*\*\* when p<0.01 to note significance.

Based on the regression outcome above, we can calculate the presented marginal effects:

1. an increase in the instructor’s beauty rating by one standard deviation

In this case, for each standard deviation the instructor’s beauty increased, his or her teaching evaluation score on a course would averagely increase 0.157 points out of 5.

1. the instructor being a non-minority male as opposed to non-minority female

In this case, for each instructor who is non-minority female, her teaching evaluation score on a course would averagely decrease 0.184 points out of 5 compared to an average non-minority male instructor.

On the other hand, taking the dummy checks into consideration, since the dummy of whether the individual is minority is 1 when it is true, the cross term does not affect on the discussion of non-minority partition.

1. the instructor being 60 as opposed to 50 years old

In this case, for an instructor ages 60 other than 50, his or her teaching evaluation score on a course would averagely increase 0.034×(60−50)+( −0.0004)×(602−502)=−0.1 points approximately, which means to averagely decrease about 0.1 points approximately in other words.

g)

To add another cross dummy variable *male\_minorityi* is a bad idea.

On consideration of the fact that in this case, when we confirm one instructor is female, that would definitely imply that she is not male, and vice versa. Thus, if there exists a cross dummy term *female\_minorityi* and a *male\_minorityi* simultaneously, it is inevitable that a complete collinearity (correlation=−1) would happen, violating the CLRM assumption 4, and result in it is impossible to solve out valid estimation for both variables simultaneously within one multivariate regression. That indicates it is meaningless to append this variable.

h)

Given that CLRM assumption 2 – mean-zero-error – is sufficed, which means the effect of instructors’ beauty on their evaluation scores is not endogenous, we can conclude that the effect is fairly important.

In the CLRM outcome above, we can find the standardized beauty rating provides a strongly positive effect on the evaluation score, 0.157 improvement per SD is quite high in a 5-point scale. Meanwhile, the standard error is relatively small in this term’s estimation, result in an extreme significance degree with a confidence level far over 99.9%, indicating the strong determinacy of the positive effect of good appearance of an instructor on his or her evaluation score to be expected.

The causality of the effect is also credible given the assumption 2 holds. Since the estimation is unbiased due to the mean of errors is zero, no other confounding factors are supposed to exist.

i)

CLRM assumption 2 would fail due to endogeneity. In this empirical case, 3 common reasons of endogeneity are all possible to occur.

Hereby we provide a possible scenario of correlated unobservables to question the mean-zero-error assumption: Given that some professors have significantly bad temper. This factor could directly affect the students’ rating scores by worse feelings during the lectures, in the meantime, it could also affect these lectures’ looks permanently caused by the chronic distorting of facial muscles in frequent rages. Such unobserved factor will violate the assumption 2 on the *beautyi* term in this regression model, hence undermine the credibility of the causality concluded.

j)

1) Covariance

Using R scripts to new a vector called Residuals and calculate the covariance between *beautyi* and *εi*:

Residuals <- resid(fit1)

print(Residuals)

cov\_beauty\_resid<-cov(TeachingRatings$beauty,Residuals)

print(cov\_beauty\_resid)

Running outcome:

# [1] -1.231115e-17

If the CLRM assumption 2 holds, the mean of residuals on condition of any regressor should be zero. In this case, the *beautyi* is a standardized distributed variable, whose mean value is 0. Hence, since holds, the covariant we calculated above should be 0 because there should be no correlation. The result is −1.231×10−17 approximately. If we round up to 6 decimal place it is still −0.000000, which is highly near to 0. Thus indicates the CLRM assumption 2 is valid in a very high confidence level.

2) Residuals to logarithm of students

Using R scripts to sketch a scatter plot:

install.packages("ggplot2")

library(ggplot2)

plot3<-ggplot(TeachingRatings, aes(x=log(allstudents), y=Residuals))

plot3<-plot3+geom\_point()+labs(title='A Scatter Plot of the Residuals against ln(allstudents\_i)')

plot3

ggsave('Figures&Tables/scatterPlotResid.png',plot = plot3, width = 10, height = 7, dpi = 300)

Image outcome:

图表, 散点图

描述已自动生成

The variance of scatter points shown in this plot distributes constant over the different value of the logarithm of students generally when we observe it. Therefore, the homoskedasticity assumption is generally sufficed in this case. Additionally, we do not find any apparent feature of correlation in any direction between the two sets by observation. Above all, the CLRM assumption 3 is seemed valid generally in this case.

3) Normality

Using R scripts to sketch the density plot on comparison to the normal distribution:

plot\_temp<-ggplot(TeachingRatings)+geom\_density(aes(x=Residuals),color="black")

TeachingRatings$"normalDistribution\_x"=seq(-2,2,4.0/(nrow(TeachingRatings)-1))

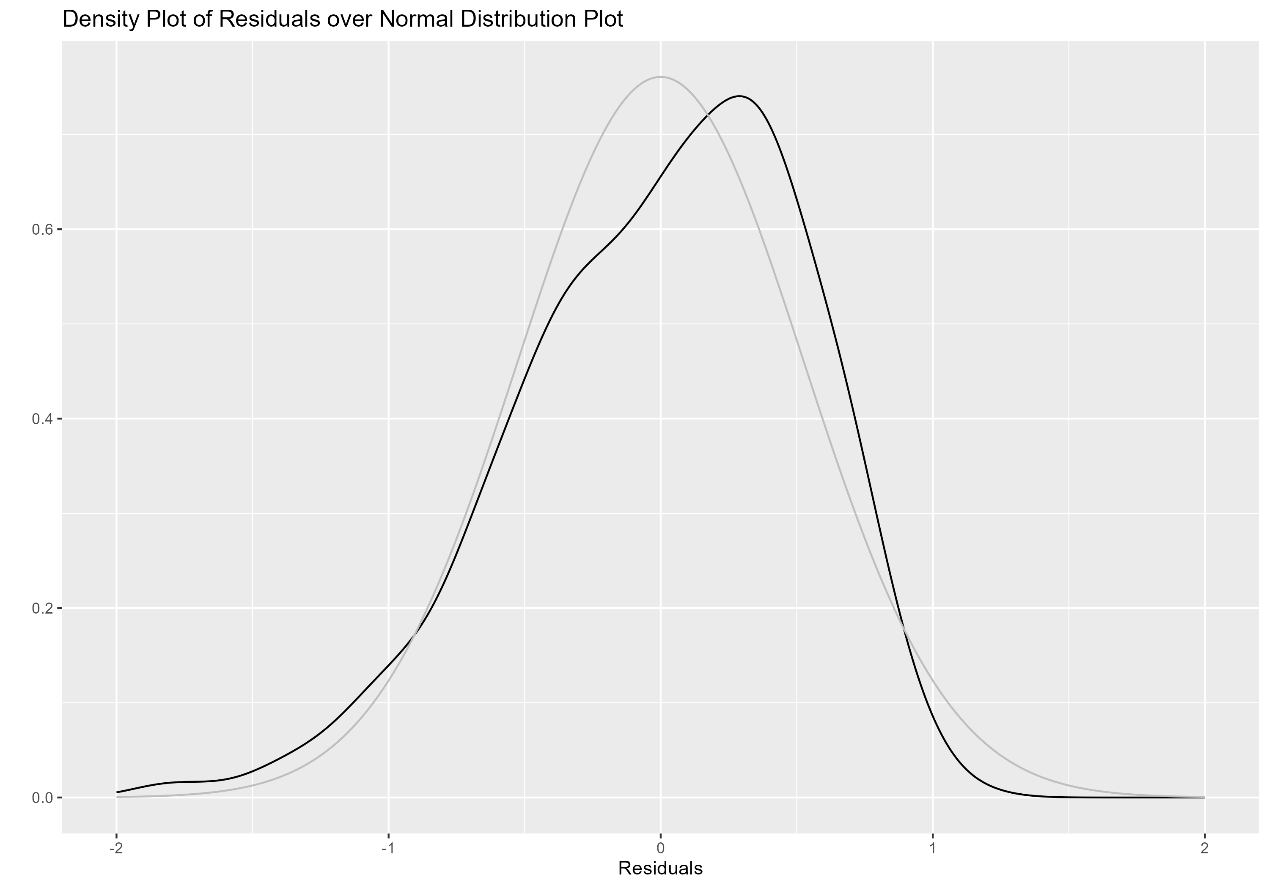
TeachingRatings$"normalDistribution\_y"=dnorm(seq(-2,2,4.0/(nrow(TeachingRatings)-1)),mean=mean(Residuals),sd=sd(Residuals))

plot4<-plot\_temp+geom\_line(aes(x=normalDistribution\_x,y=normalDistribution\_y),color="grey")+labs(x='Residuals',y='',title='Density Plot of Residuals over Normal Distribution Plot')

plot4

ggsave('Figures&Tables/densityOfResidualsOverANormalDistribution.png',plot = plot4, width = 10, height = 7, dpi = 300)

Image outcome:



Notes:

# Here I chose a different method to plot the standard normal distribution from the hint in the problem set. I directly plot a line chart for each (x,y) should appears in the density curve, at a resolution of each observation have a uniformly-distributed point, within the [-2,2] interval.

# If we follow the hint in the problem set, we can call the rnorm() method to draw *N* randomized observations from a SND, then use ggplot.geom\_density() method with these observations above to sketch the SND density curve. It will get a very similar density plot as the picture shown above.

The black curve sketched on the plot above refers the density of the actual residuals in this LM, while the light grey “background” curve refers the standard normal distribution density. We can find that the actual distribution of residuals is slightly skewed left in comparison to the SND by observing, but not too much. So, we can conclude that the normality assumption of the residual distribution, which is the CLRM assumption 5, is valid in general.