ECON 3412 HW 3

David Chen, dc3451

October 21, 2020

Problem 1

Note: throughout this problem, I refer to the share of income of the bottom quintile. What I *really* mean is the proportional rate at which that share of income varies with national income, but that is a long phrase to write, so I just put the shorter "share of income" to represent that, even though different income distributions can still admit the same α_1, as long as they increase the income of the bottom quintile in the same proportion as national income.

a

The corresponding $t-$statistic here is

$$t = \frac{\hat{\alpha_1} - 1}{SE(\hat{\alpha_1})} = \frac{1.072 - 1}{0.025} = 2.88$$

Then, the corresponding $p-$value is $\approx 2(1 - \Phi(2.88)) = 0.00397 < 0.05$, so we can actually reject that $\alpha_1 = 1$.

b

The 95% confidence interval is $\hat{\alpha_1} \pm \Phi^{-1}(0.975)SE(\hat{\alpha_1}) = 1.072 \pm 1.96(0.025) = (1.023, 1.121)$.

c

So, we have that there are multiple things that are correlated with the per capita income of a country that affect the distribution of income within the country. For example, poorer countries might expect to see higher levels of government corruption, and as a result more of national income accrues to government officials and their friends, which diminishes the

share of income accruing to the bottom quintile. Another example could be something like the fact that poorer countries tend to have labor intensive sectors, meaning that potentially low-income laborers might take a larger share of the income than in richer nations, with their share rising faster than the national income.

d

- Trade Volumes: a proxy of openness, which might see trade income which probably accrues in the hands of wealthier people, although high trade volumes might create more jobs for lower income manufacturing/farming/etc, so the actual effect of trade volume is hard to say theoretically.
- Inflation: might be a proxy of government economic policy stability (and economic stability in general), and as a result countries experiencing high inflation means that the poor in the country feel it the most and thus lower the income of the bottom quintile.
- Government Consumption: higher government consumption is roughly direct to how much the government takes in taxes, which diminishes the income of lower-income people.
- Financial Development: a proxy for how well connected a nation is to international markets, and how efficient the nation's financial market is. In particular, we might expect that more developed financial markets would result in higher incomes for everyone.
- Rule of Law: poorer countries might expect to see higher levels of government corruption, and as a result more of national income accrues to government officials and their friends, which diminishes the share of income accruing to the bottom quintile.

e

We have that in this case $t = \frac{1.140-1}{0.101} = 1.386 < 1.96$, so we can no longer reject the null that $\alpha_1 = 1$ at the 95% significance level.

f

Plenty: amount of the population in manufacturing/farming industries, main type of exports, religious extremity, capital intensity. These all can determine the proportional income of the bottom quintile relative to the overall country income. For example, countries with high amounts of capital intensity might expect to see larger amount of income accrue to the wealthier classes, with little of the national income going to the bottom quintile. Similarly,

highly religious nations might see that religious minorities are extremely poor, decreasing the share of income of the bottom quintile.

Problem 2

If the $t-$statistics are independent of each other, then the odds of rejecting any given one of $\beta_i = 0$ is the likelihood that a $t-$statistic exceeds 1.96 given the real mean is 0 is $2(1-F_d(1.96))$ where F_d is the CDF of the t-distribution with d degrees of freedom. In particular, if d is large, then $F \approx \Phi$, the CDF of the standard normal distribution. In this case, we can compute $2(1 - \Phi(1.96)) = 0.05$.

Then, the odds of the null being rejected is the complement of the probability that no t-statistic exceeds 1.96. This latter quantity, the probability that no t-statistic exceeds 1.96, is $(1 - 2(1 - F_d(1.96)))^3$, which evaluates to $(1 - 0.05)^3 = 0.857$ if d is large. Then, the odds of the null being rejected is $1 - (1 - 2(1 - F_d(1.96)))^3$, which comes out to 0.142 if d is large.

Problem 3

A control variable is a variable included in the regression to account for omitted causal factors. Usually, these omitted causal factors are things that are hard to measure, and the control is something that is correlated with these causal factors, but is not causal by itself. This removes the omitted variable bias when done properly, ensuring that $E(u_i \mid X) = 0$ where X is some causal variable of interest.

An example of a control variable might be something like including multiple dummy variables indicating race in an regression that seeks to determine something like the causal impact of an individual's neighborhood's population density on SAT math scores. In particular, being black or hispanic or Asian or white or Amerindian etc. doesn't inherently make you worse at standardized testing (in particular, probably not on the math score; in other sections, things literature excerpts being white-skewed might actually do something like that), but it is a proxy for racial discrimination in education. Further, since black and hispanic people are more concentrated in urban areas, it is likely that the magnitude of the coefficient of the population density will be overestimated relative to its true causal effect. In this case, since it is hard to actually quantify racial discrimination in education succinctly, we can simply introduce dummy variables for race to control for that omitted causal factor.

Problem 4

a

We get that

```
> summary(lm(course_eval ~ beauty, data = read_dta("TeachingRatings.dta")))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.99827    0.02535 157.727  < 2e-16 ***
beauty       0.13300    0.03218   4.133 4.25e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5455 on 461 degrees of freedom
Multiple R-squared:  0.03574,        Adjusted R-squared:  0.03364
F-statistic: 17.08 on 1 and 461 DF,  p-value: 4.247e-05
```

Then, we have that $R^2 = 0.03574, \overline{R}^2 = 0.03364$, so we have that this regression explains 3.574% of the variance that we see in course evaluations.

b

We get that

```
> eval_beauty_age <- lm(course_eval ~ beauty + age,
                        data = read_dta("TeachingRatings.dta"))
> waldtest(eval_beauty_age, vcov = vcovHC(eval_beauty_age, type = "HC1"))
Wald test

Model 1: course_eval ~ beauty + age
Model 2: course_eval ~ 1
  Res.Df Df      F    Pr(>F)
1    460
2    462 -2 8.8965 0.0001619 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We will use $F_{d,\infty}$ for the rest of the problem, since the amount of observations is large. Then, we have that $F = 8.8965$, and the critical value for $F_{2,\infty}$ is (in `R`) `qf(0.99, 2, Inf) = 4.60`, so we reject the null that the coefficients are jointly insignificant at the 1% level.

c

We get that

```
> eval_beauty_age_min <- lm(course_eval ~ beauty + age + minority,
                            data = read_dta("TeachingRatings.dta"))
> waldtest(eval_beauty_age_min, vcov = vcovHC(eval_beauty_age_min, type = "HC1"))
Wald test

Model 1: course_eval ~ beauty + age + minority
Model 2: course_eval ~ 1
  Res.Df Df      F    Pr(>F)
1    459
2    462 -3 6.5368 0.0002457 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Then, we have that $F = 6.5368$, and the critical value for $F_{3,\infty}$ is (in R) `qf(0.99, 3, Inf) = 3.78`, so we reject the null that the coefficients are jointly insignificant at the 1% level.

d

Testing in R,

```
> eval_beauty <- lm(course_eval ~ beauty, data = read_dta("TeachingRatings.dta"))
> eval_beauty_age_min <- lm(course_eval ~ beauty + age + minority,
                            data = read_dta("TeachingRatings.dta"))
> waldtest(eval_beauty,
           eval_beauty_age_min,
           vcov = vcovHC(eval_beauty_age_min, type = "HC1"))
Wald test

Model 1: course_eval ~ beauty
Model 2: course_eval ~ beauty + age + minority
  Res.Df Df      F Pr(>F)
1    461
2    459  2 1.3309 0.2652
```

Then, we get that since the critical value is 4.605 as in part b, we have that we cannot reject the null that minority and age are jointly insignificant.

Alternatively, we have that the R^2 is 0.03574 when regressing only on beauty and is 0.04263 when regressing as in part c. Then, we have that $F = \frac{(0.04263-0.03574)/2}{(1-0.04263)/457} = 1.644 < 4.605$.

This is different F-statistic that before, which makes sense, as we have that the first is heteroskedastic.

e

Regressor (SE below)	(1)	(2)	(3)	(4)
beauty	0.13406 (0.03186)	0.13421 (0.03169)	0.13512 (0.03157)	0.15920 (0.03068)
age	2.87e-4 (2.545e-3)	-1.959e-4 (2.521e-3)	3.5037e-5 (2.495e-3)	-1.954e-3 (2.621e-3)
minority		-0.1338 (0.0820)	-0.07205 (0.08435)	-0.16942 (0.06789)
nnenglish			-0.30299 (0.09651)	-0.24384 (0.09589)
intro				0.00794 (0.05654)
onecredit				0.63300 (0.10776)
female				-0.18323 (0.05219)
intercept	3.9844 (0.1241)	4.0262 (0.1229)	4.0249 (0.1211)	4.16853 (0.13903)
F-statistics ($p-$value below)				
beauty, age	8.8965 (1.619e-4)	9.0582 (1.386e-4)	9.1864 (1.226e-4)	13.827 (1.48e-6)
beauty, age, minority		6.5368 (2.457e-4)	6.208 (3.859e-4)	11.078 (4.957e-7)
beauty, age, minority, nnenglish			8.0394 (2.599e-6)	10.888 (1.94e-8)
intro, onecredit				22.743 (3.856e-10)
minority, age				3.3325 (0.03658)
intro, age				0.3031 (0.7387)
Regression summary statistics				
$\overline{R}^2$	0.03157	0.03637	0.04991	0.1426
R^2	0.03576	0.04262	0.05814	0.1556
Regression RMSE	0.544	0.542	0.538	0.509
n	463	463	463	463

It seems that beauty is always an important factor, with every combination of beauty and some other variable being jointly significant. The same holds true for minority; surprisingly, in the last regression, intro and onecredit are jointly significant (though this is likely due to onecredit likely being significant, looking at standard errors). The one that seems less important to include is age and intro as well, as they are jointly insignificant.

The above table contains entries pulled by the following `R` code:

```
library(haven)
library(sandwich)
library(lmtest)

teach_data <- read_dta("TeachingRatings.dta")

eval_1 <- lm(course_eval ~ beauty + age, data = teach_data)
cat("R^2: ", summary(eval_1)$r.squared, "\n")
cat("Adjusted R^2: ", summary(eval_1)$adj.r.squared, "\n")
cat("Regression RMSE: ", sqrt(mean(eval_1$residuals^2)), "\n")
cat("n: ", length(eval_1$residuals), "\n")
print(coeftest(eval_1, vcov = vcovHC(eval_1, "HC1")))

eval_2 <- lm(course_eval ~ beauty + age + minority, data = teach_data)
cat("R^2: ", summary(eval_2)$r.squared, "\n")
cat("Adjusted R^2: ", summary(eval_2)$adj.r.squared, "\n")
cat("Regression RMSE: ", sqrt(mean(eval_2$residuals^2)), "\n")
cat("n: ", length(eval_2$residuals), "\n")
print(coeftest(eval_2, vcov = vcovHC(eval_2, "HC1")))

eval_3 <- lm(course_eval ~ beauty + age +
               minority + nnenglish,
             data = teach_data)
cat("R^2: ", summary(eval_3)$r.squared, "\n")
cat("Adjusted R^2: ", summary(eval_3)$adj.r.squared, "\n")
cat("Regression RMSE: ", sqrt(mean(eval_3$residuals^2)), "\n")
cat("n: ", length(eval_3$residuals), "\n")
print(coeftest(eval_3, vcov = vcovHC(eval_3, "HC1")))

eval_4 <- lm(course_eval ~ beauty + age + minority + nnenglish +
               intro + onecredit + female,
             data = teach_data)
cat("R^2: ", summary(eval_4)$r.squared, "\n")
cat("Adjusted R^2: ", summary(eval_4)$adj.r.squared, "\n")
cat("Regression RMSE: ", sqrt(mean(eval_4$residuals^2)), "\n")
cat("n: ", length(eval_4$residuals), "\n")
print(coeftest(eval_4, vcov = vcovHC(eval_4, "HC1")))
```

```
print(waldtest(eval_1, vcov = vcovHC(eval_1, type = "HC1")))

print(waldtest(eval_2, lm(course_eval ~ minority, data = teach_data),
               vcov = vcovHC(eval_2, type = "HC1")))
print(waldtest(eval_2, vcov = vcovHC(eval_2, type = "HC1")))

print(waldtest(eval_3,
               lm(course_eval ~ minority + nnenglish, data = teach_data),
               vcov = vcovHC(eval_3, type = "HC1")))
print(waldtest(eval_3,
               lm(course_eval ~ nnenglish, data = teach_data),
               vcov = vcovHC(eval_3, type = "HC1")))
print(waldtest(eval_3, vcov = vcovHC(eval_3, type = "HC1")))

print(waldtest(eval_4,
               lm(course_eval ~ minority + nnenglish + intro +
                    onecredit + female, data = teach_data),
               vcov = vcovHC(eval_4, type = "HC1")))
print(waldtest(eval_4, lm(course_eval ~ nnenglish + intro +
                          onecredit + female, data = teach_data),
               vcov = vcovHC(eval_4, type = "HC1")))
print(waldtest(eval_4, lm(course_eval ~ intro + onecredit +
                          female, data = teach_data),
               vcov = vcovHC(eval_4, type = "HC1")))
print(waldtest(eval_4, lm(course_eval ~ beauty + age + minority +
                          nnenglish + female, data = teach_data),
               vcov = vcovHC(eval_4, type = "HC1")))
print(waldtest(eval_4, lm(course_eval ~ beauty + nnenglish + intro +
                          onecredit + female, data = teach_data),
               vcov = vcovHC(eval_4, type = "HC1")))
print(waldtest(eval_4, lm(course_eval ~ beauty + minority + nnenglish +
                          onecredit + female, data = teach_data),
               vcov = vcovHC(eval_4, type = "HC1")))
```

Problem 5

a

There is perfect co-linearity, as we have that we fall into the dummy variable trap. Indeed, `R` gives that one variable is not defined because of singularities. In particular,

```
> summary(lm(salary ~ north + south + east + west, data=read_dta("LAWSCH85.DTA")))

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    39935       1991  20.061   <2e-16 ***
north          -1568       2926  -0.536   0.5929
south          -5594       2815  -1.987   0.0488 *
east            2306       2671   0.864   0.3893
west              NA         NA      NA       NA
```

so we cannot properly do the regression this way.

b

We can simply remove the intercept. In that case, we ge that

```
> summary(lm(salary ~ north + south + east + west + 0, data=read_dta("LAWSCH85.DTA")))

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
north    38366       2145   17.89   <2e-16 ***
south    34341       1991   17.25   <2e-16 ***
east     42241       1780   23.73   <2e-16 ***
west     39935       1991   20.06   <2e-16 ***
```

c

The coefficient of east here is simply that law graduates from eastern schools will make on average \$42241 in salary (whatever that means, the data columns aren't documented).

Problem 6

log is the natural logarithm.

a

We can log transform the equation, such that

$$Q = \lambda K^{\beta_1} L^{\beta_2} M^{\beta_3} e^u \implies \log(Q) = \log(\lambda) + \beta_1 \log(K) + \beta_2 \log(L) + \beta_3 \log(M) + u$$

Given some dataset, we can apply the logarithms to the values, and no do normal OLS regression to estimate the production parameters. In particular, $\log(\lambda)$ is the intercept,

and β_i slopes in the multivariate regression, representing elasticities of labor, capital, and materials.

b

It turns out that Cobb-Douglas production functions are homogenous, such that

$$Q(cK, cL, cM) = \lambda(cK)^{\beta_1}(cL)^{\beta_2}(cM)^{\beta_3}e^u = (c^{\beta_1+\beta_2+\beta_3})\lambda K^{\beta_1}L^{\beta_2}M^{\beta_3}e^u = c^{\beta_1+\beta_2+\beta_3}Q(K, L, M)$$

Now, we can see that there are constant returns to scale under $\beta_1 + \beta_2 + \beta_3 = 1$. This suggests that we can test the hypothesis that $\beta_1 + \beta_2 + \beta_3 = 1$, where rejecting means that there are non-constant returns to scale. In particular, we can test that that $\beta_1 + \beta_2 + \beta_3 < 1$ is diminishing returns to scale and > 1 is increasing returns to scale.

To do this, we can reparameterize such that we define a new coefficient $\gamma = \beta_1 + \beta_2 + \beta_3$, such that

$$\log(Q) = \log(\lambda) + \gamma \log(K) + \beta_2(\log(L) - \log(K)) + \beta_3(\log(M) - \log(K)) + u$$

and test $\gamma = 1$ in the new regression.

c

We can do the same trick of reparameterizing, with $\beta_1 + \beta_2 + \beta_3 = 1$, such that

$$\begin{aligned}
\log(Q) &= \log(\lambda) + \beta_1 \log(K) + \beta_2 \log(L) + \beta_3 \log(M) + u \\
&= \log(\lambda) + \beta_1 \log(K) + \beta_2 \log(L) + (1 - \beta_1 - \beta_2) \log(M) + u \\
&= \log(\lambda) + \beta_1(\log(K) - \log(M)) + \beta_2(\log(L) - \log(M)) + \log(M) + u \\
\log(Q) - \log(M) &= \log(\lambda) + \beta_1(\log(K) - \log(M)) + \beta_2(\log(L) - \log(M)) + u
\end{aligned}$$

and test for β_1, β_2 in the new regression and calculate $\beta_3 = 1 - \beta_1 - \beta_2$.