Intermediate Econometrics HW2

Mason Ross Hayes and Vanlik TAN

10/21/2020

Question 1

We specify the model as:

$$lwage = \beta_0 + \beta_1 Male + \beta_2 Age + \beta_3 Education + \beta_4 Age^2 + \varepsilon$$

To conduct a stability test, we construct 3 models: we split the group of n observations into one group in Model 1 where $Informal = 1$ and another in Model 2 where $Informal = 0$. Model 3 is the combined model, where we let $Informal$ take any value 0 or 1.

Our hypotheses:

$$H_0 : \beta^{\mathbf{I}} = \beta^{\mathbf{F}}; H_a : \beta^{\mathbf{I}} \neq \beta^{\mathbf{F}}$$

Where the superscript I indicates informal and F indicates formal, respectively.

The stability test will then be approximated by:

$$F = \frac{SSR_3 - SSR_2 - SSR_1}{SSR_2 + SSR_1}(\frac{n-2k}{k})$$

As seen in the R script `econometrics_hw2.R` in the variable `F_stability`, this F statistic is:

$$F_{stability} = 4820 > F_{5,73405,1-\alpha} = 2.21$$

Question 2

We now consider the model:

$$lwage = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Education + \beta_4 Informal + \beta_5 m + \beta_6 u + \beta_7 ftw + \beta_8 un + u$$

where m, u, ftw, and un are dummy variables for male, urban, full time worker, and union, respectively.

Running the regression in R we obtain the following estimates for the coefficients of Age, Age^2 and $Education$, which are β_1, β_2, and β_3 respectively:

Variable:	Age	Age Squared	Education
Coefficient:	0.02693	-0.000256	0.04260

The marginal effect of a one-year increase in general education, all else equal, is a 4.26% increase in wage. The marginal effect of a one-year increase in age, all else equal, is $\beta_2 + 2\beta_3 Age = 2.69\% - 0.0511\% * Age$.

Question 3

We now consider the model

$$lwage = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Education + \beta_4 Informal + \beta_5 m + \beta_6 u + \beta_7 ftw + \beta_8 un + \beta_9 l + u$$

where m, u, ftw, un, and l are the dummy variables for male, urban, full time worker, union, and lower caste,, respectively.

We estimate this model as

```
q3_model = df %>%
  lm(formula = lwage ~ Age + age_sq + Education + Informal
  + currently_married + urban + ftw + union + lower_caste)
```

Do the marginal effects of Age and Education change?

We can easily pull the standard errors from each model with the following:

```
q3_se = coef(summary(q3_model))[c(2,4),2]
q3_var = q3_se^2
```

And we then repeat this for Model 2.

The t-statistic to test if coefficient for Age in model 2 is the same as for Age in model 3 is:

```
t_stat_age = (coef(q2_model)[2] - coef(q3_model)[2])/(sqrt(q3_var[1] + q2_var[1]))
```

Which is:

$$\hat{t}_{Age} = 0.164 < 1.96$$

So we fail to reject the null hypothesis that the coefficient of Age is the same in Model 2 and 3. We repeat the t-test for Education:

```
t_stat_educ = (coef(q2_model)[4] - coef(q3_model)[4])/(sqrt(q3_var[2] + q2_var[2]))
```

$$\hat{t}_{Education} = 2.05 > 1,96$$

So we reject the null hypothesis that the coefficient of Education is the same in Model 2 and Model 3.

Interpret the coefficient of Informal

We interpret the coefficient of Informal in q3_model as: those who work in the informal sector, all else equal, earn wages 89 percent lower than those who work in the formal sector

Question 4

We now consider the model:

$$lwage = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Education + \beta_4 Informal + \beta_5 m + \beta_6 u + \beta_7 ftw + \beta_8 un + \beta_9 l + \beta_{10} mi + u$$

where m, u, ftw, un, l, and mi are the dummy variables for male, urban, full time worker, union, lower caste, and the interaction of male and Informal, respectively.

Females working in the informal sector receive 127 percent lower wages than females in the formal sector. Males working in the informal sector earn 45 percent more than females working in the informal sector, but 82 percent less than males working in the formal sector.

Question 5

With the following code we store the log of the squared residuals from the Question 4 regression, then we regress the log squared residuals on age, education, their squared terms and their interaction term.

```
log_q4_sq_residuals = log(q4_model$residuals^2)

df2 = cbind(df, log_q4_sq_residuals) %>%
  mutate(educ_sq = Education^2) %>%
  mutate(educ_times_age = Education * Age)

q5_model = df2 %>%
  lm(formula = log_q4_sq_residuals ~ Age + Education
  + age_sq + educ_sq + educ_times_age)
```

And this model `q4_model` has an F-statistic of 42.99 with $df = (5, 73409)$. Since $42.99 > 2.21$, we reject the null hypothesis of homoskedasticity.

Question 6

We are given the skedastic form:

$$Var(\varepsilon \mid X) = exp(\gamma_0 + \gamma_1 Age + \gamma_2 Education + \gamma_3 Age^2 + \gamma_4 Education^2 + \gamma_5 Age * Education + 1)$$

We estimate this model in R in the following way:

1. Store the residuals from the OLS model in Q4 (`q4_model` in our code)
2. We use this estimate of $\hat{\varepsilon}$ to regress $log(\hat{\varepsilon}^2)$ on Age, Education, and their square terms and interaction term, plus the constant 1. The exponential of these fitted values will then give us a consistent estimate of σ_i^2. We then run GLS with using the weights $\frac{1}{\hat{\sigma}_i^2}$.

See Estimated result @ref(fig: table1)

Table 2: Regression results

	Dependent variable:	
	lwage	
	(1)	(2)
Age	0.034^{***}	0.034^{***}
	(0.001)	(0.001)
age_sq	-0.0003^{***}	-0.0003^{***}
	(0.00002)	(0.00002)
Education	0.039^{***}	0.038^{***}
	(0.0005)	(0.0005)
Informal	-1.267^{***}	-1.269^{***}
	(0.009)	(0.009)
currently_married	0.044^{***}	0.042^{***}
	(0.007)	(0.007)
urban	-0.001	-0.002
	(0.006)	(0.006)
ftw	0.252^{***}	0.253^{***}
	(0.017)	(0.017)
union	0.176^{***}	0.177^{***}
	(0.007)	(0.007)
lower_caste	-0.114^{***}	-0.109^{***}
	(0.006)	(0.006)
male_informal	0.452^{***}	0.461^{***}
	(0.007)	(0.007)
Constant	6.481^{***}	6.487^{***}
	(0.031)	(0.031)
Observations	73,415	73,415
R^2	0.533	0.513
Adjusted R^2	0.533	0.513
Residual Std. Error (df = 73404)	0.661	2.131
F Statistic (df = 10; 73404)	$8{,}375.963^{***}$	$7{,}726.402^{***}$

Note: *p<0.1; **p<0.05; ***p<0.01
Column (1) represents coefficient estimated using model in question 4
While column (2) provides coefficient estimated using model in question 6

Question 7

The regression in Question 6 has some problems; for example, do we know how does marriage relate to wages? Are people with higher wages more likely to be married for reasons unrelated to their wages – for example, family background, education, etc?

There are also issues with defining what is an "informal" worker – we have followed the definition given that an informal worker is any worker who does not have access to social security benefits. This might be a good definition in that the probability of a worker being in the informal sector given that he has no access to social security benefits may be high; but what about the probability that a worker has no access to social security benefits given that he is in the (informal) formal sector? In other words, it's possible we are (underestimating) overestimating the number of workers in the informal sector which could entirely change the model.

We have to be confident that we can overcome these types of problems in order to claim that this is a good model of the wages of workers in India.