Report for Assignment 1

Li Mingming

The report chooses female, age, age square and highest education as the predictors, and uses lnw as the target variable, which could be shown like below:

$$\ln w = \beta_0 + \beta_1 * female \tag{1.1}$$

$$\ln w = \beta_0 + \beta_1 * female + \beta_2 * age \tag{1.2}$$

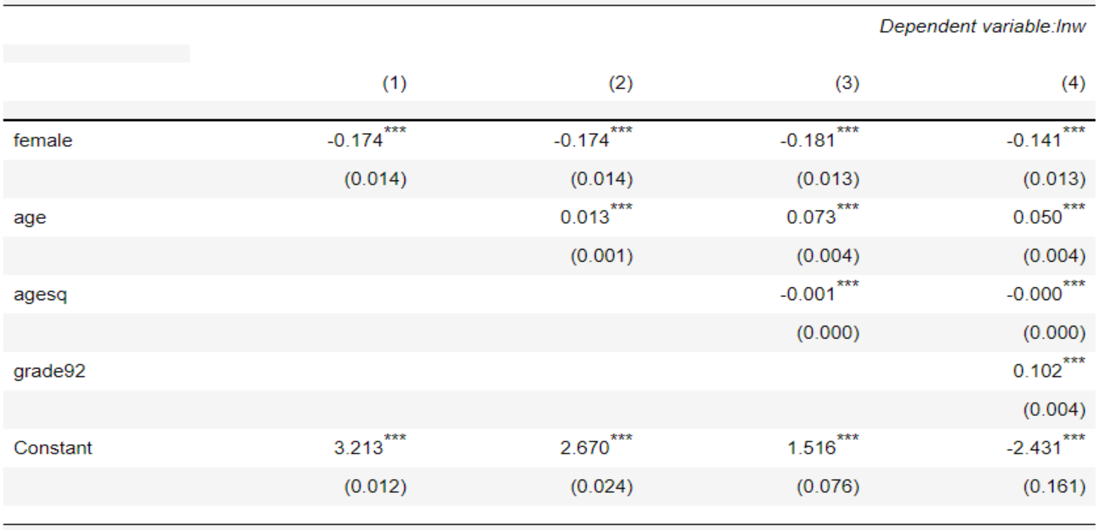
$$\ln w = \beta_0 + \beta_1 * female + \beta_2 * age + \beta_3 * age^2 \tag{1.3}$$

$$\ln w = \beta_0 + \beta_1 * female + \beta_2 * age + \beta_3 * age^2 + \beta_4 * grade92 \tag{1.4}$$

The reason of taking the logarithm of *w*: reduce heteroscedasticity and get normal distribution

The reason of taking these factors: female (productivity and discrimination), age (potential experience and physical situation); age square (The effect of age on wage growth is inverted U-shaped); grade92 (education has positive effect on productivity).

Table 1 Linear Regression Result



	Dependent variable:lnw			
	(1)	(2)	(3)	(4)
female	-0.174^{***}	-0.174^{***}	-0.181^{***}	-0.141^{***}
	(0.014)	(0.014)	(0.013)	(0.013)
age		0.013^{***}	0.073^{***}	0.050^{***}
		(0.001)	(0.004)	(0.004)
agesq			-0.001^{***}	-0.000^{***}
			(0.000)	(0.000)
grade92				0.102^{***}
				(0.004)
Constant	3.213^{***}	2.670^{***}	1.516^{***}	-2.431^{***}
	(0.012)	(0.024)	(0.076)	(0.161)

All the predictors are significant. Female is a dummy variable, which represent the female will get 14%-17% less than the male. One more year older will bring 1.3%-5% increase in hour-earning.Age square seems makes no difference though it is significant. Education is important and one-level increase will promote around 10% percent arise of the earning. R-square is increasing from the simplest to the fourth model, though it is not shown here for no enough space here.

Table 2 Compare Model Performance

	Model 1	Model 2	Model 3	Model 4
RMSE in the full sample	0.57274	0.55159	0.54330	0.52008
Cross-Validated RMSE	0.57287	0.55179	0.54353	0.52034
BIC in the full sample	14284.918	13672.081	13430.427	12716.558

We need to find the regression that would produce the smallest RMSE, Cross-Validated RMSE and lowest BIC model. From model 1 to model 2, it can be seen that both RMSE and BIC have dropped a lot. Model 3 has a small drop on the basis of model 2. After adding the highest degree of education information, model 4 performs best among all models, which also shows the appropriately increasing the complexity of the model is helpful for better prediction.