

Assignment #2

Instructions:

- For all questions, answer up to 4 decimal places.
 - This assignment is due on **Thursday, May 20, 2021 before 23.59.**
 - Write your answer in either digital or ordinary paper. For digital paper, export pages into a single PDF file. For ordinary paper, take photos of your writing and convert them into a single PDF file as well.
 - There is no need to rewrite the question. Assign number item, i.e. 1 a., clearly before your answer is sufficient.
 - Submit your assignment into Moodle.
 - Name your file as StudentID_Nickname (in Thai) such as 123456789_ปั๋. **Please follow this instruction strictly since it will help me a lot with file management.**
-

Question 1. The data set CEOSAL1.DTA contains information on 209 CEOs for the year 1990; these data were obtained from Business Week (5/6/1991). To study effect of firm performances and types of industry where CEOs work on CEO compensation, the CEO salary regression is proposed as follows:

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \log(\text{sales}_i) + \beta_2 \text{ROE}_i + \beta_3 \text{finance}_i + \beta_4 \text{consprod}_i + \beta_5 \text{utility}_i + u_i$$

where

- $\log(\text{salary}_i)$ = logarithm of CEO annual salary (in 1,000 USD)
- $\log(\text{sales}_i)$ = logarithm of firms' sale (in 1 million USD)
- ROE_i = average return on equity for the CEO's firm for the previous three years (Return on equity is defined in terms of net income as a percentage of common equity)
- finance_i = 1 if in financial industry, = 0 otherwise
- consprod_i = 1 if in consumer product industry, = 0 otherwise
- utility_i = 1 if in utility industry, = 0 otherwise

(finance_i , consprod_i , and utility_i are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

Using STATA, the estimation result is shown below. Answer the following questions.

Source	SS	df	MS	Number of obs = 209		
ESS RSS TSS Model	23.8109943	5	4.76219887	F(5, 203)	=	22.53
Residual	42.9111689	203	.211385068	Prob > F	=	0.0000
Total	66.7221632	208	.320779631	R-squared	=	0.3569
				Adj R-squared	=	0.3410
				Root MSE	=	.45977

lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lsales	.2571917	.0320348	8.03	0.000	.0194282	.3203553
roe	.0111517	.3342996	2.59	0.010	.0026742	.0196293
finance	.1579564	.0890017	1.77	0.077 *	-.0175299	.3334426
consprod	.1808917	.0847683	2.13	0.034	.0137524	.3480311
utility	-.2830015	.0992337	-2.85	0.005	-.4786624	-.0873405
_cons	4.588101	.2950221	15.55	0.000	4.0064	5.169801

- Write out the estimated regression equation for $\log(\text{salary}_i)$. Interpret the estimated coefficient associated with $\log(\text{sales}_i)$.
- What is the overall significance of the regression? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State **the critical value** for hypothesis testing to receive full points.
- Compute the approximate percentage difference in estimated salary between the utility and transportation sector, holding sales_i and ROE_i fixed.
- Why can't we put all the sector dummies (i.e. finance_i , consprod_i , utility_i and transport_i) in the equation? What would happen if we put all the sector dummies in the equation and use STATA run the regression anyway?
- In the above model, is there any benefit if we add interaction terms between roe and sector dummies, i.e. $\text{ROE}_i * \text{finance}_i$ and/or $\text{ROE}_i * \text{consprod}_i$ and/or $\text{ROE}_i * \text{utility}_i$?

a. $\log(\text{salary}) = 4.588101 + 0.2571917 \log(\text{sales})$

positive coefficient of $\log(\text{sales})$ or β_1 . It means that when $\log(\text{sales})$

increase by 1%. $\log(\text{salary})$ will increase by 0.2571917%.

- b. What is the overall significance of the regression? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State **the critical value** for hypothesis testing to receive full points.

B. Use F-test

$$H_0 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$H_1 =$ Not all slope coefficients are simultaneously zero

$$F_{cal} = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} = \frac{MS(E)}{MS(R)} = \frac{4.76219887}{0.211385068} = 22.52854904$$

at $\alpha = 0.05$ $\therefore F_{cal} > F_{crit(0.05)}^{(5, 203)}$ we can reject H_0

$$F_{upper, 9}^{(5, 203)} = 2.21$$

and we can make sure that $\beta_2, \beta_3, \beta_4$ and β_5 are not simultaneously zero

\therefore According to the test, it shows that the all independent variables can significantly explain the dependent variable

$$\alpha = 0.05, \text{ d.f} = 203, t_{crit lower} = -1.96, t_{crit upper} = 1.96$$

	t_{cal}	
1. logsales) \rightarrow	8.03	\therefore Reject H_0 , significance
2. roe \rightarrow	2.59	\therefore Reject H_0 , significance
3. finance \rightarrow	1.77	\therefore cannot reject H_0 , not significance
4. consumer product \rightarrow	2.13	\therefore Reject H_0 , significance
5. utility \rightarrow	-2.85	\therefore Reject H_0 , significance
6. constant \rightarrow	15.55	\therefore Reject H_0 , significance

- c. Compute the approximate percentage difference in estimated salary between the utility and transportation sector, holding $sales_i$ and ROE_i fixed.

utility

$$\begin{aligned} \log(\text{salary}) &= \beta_0 + \beta_3 \text{finance} + \beta_4 \text{consprod} + \beta_5 \text{utility} \\ &= 4.588101 + 0.1579564(1) + 0.1808917(1) - 0.2830015(1) \end{aligned}$$

$$\log(\text{salary}) = 4.643948, 4.3051$$

Transport sector (omit industry)

$$\begin{aligned} \log(\text{salary}) &= \beta_0 + \beta_3 \text{finance} + \beta_4 \text{consprod} + \beta_5 \text{utility} \\ &= 4.588101 + 0.1579564(0) + 0.1808917(0) - 0.2830015(0) \end{aligned}$$

$$\log(\text{salary}) = 4.588101$$

$$\Delta\% = \frac{4.588101 - 4.643948}{4.643948} \times 100 = -1.203\%$$

- d. Why can't we put all the sector dummies (i.e. $finance_i$, $consprod_i$, $utility_i$ and $transport_i$) in the equation? What would happen if we put all the sector dummies in the equation and use STATA run the regression anyway?

D. if we put all dummy variables in the equation, A case of perfect collinearity will occur.

Because each variable will have exact linear relationship between them.

In stata, it will automatically reject or omit one dummy variable. Because it cannot be estimated.

e. In the above model, is there any benefit if we add interaction terms between roe and sector dummies, i.e. $ROE_i * finance_i$ and/or $ROE_i * consprod_i$ and/or $ROE_i * utility_i$?

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \log(\text{sales}_i) + \beta_2 ROE_i + \beta_3 \text{finance}_i + \beta_4 \text{consprod}_i + \beta_5 \text{utility}_i + u_i$$

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \log(\text{sales}_i) + \beta_2 ROE_i + \beta_3 \text{finance}_i + \beta_4 \text{consprod}_i + \beta_5 \text{utility}_i + \beta_6 (ROE_i * \text{finance}_i) + \beta_7 (ROE_i * \text{consprod}_i) + \beta_8 (ROE_i * \text{utility}_i) + u_i$$

$$E(\log(\text{salary}_i) | \text{finance} = 1) = \beta_0 + \beta_1 \log(\text{sales}_i) + \beta_2 ROE_i + \beta_3 (1) + \beta_4 \text{consprod}_i + \beta_5 \text{utility}_i + \beta_6 (ROE_i * 1) + \beta_7 (ROE_i * \text{consprod}_i) + \beta_8 (ROE_i * \text{utility}_i) + u_i$$

$$= (\beta_0 + \beta_3) + \beta_1 \log(\text{sales}_i) + (\beta_2 + \beta_6) (ROE_i) + \beta_4 \text{consprod}_i + \beta_5 \text{utility}_i + \beta_7 (ROE_i * \text{consprod}_i) + \beta_8 (ROE_i * \text{utility}_i) + u_i$$

$$E(\log(\text{salary}_i) | \text{consprod} = 1) = \beta_0 + \beta_1 \log(\text{sales}_i) + \beta_2 ROE_i + \beta_3 \text{finance}_i + \beta_4 (1) + \beta_5 \text{utility}_i + \beta_6 (ROE_i * \text{finance}_i) + \beta_7 (ROE_i * (1)) + \beta_8 (ROE_i * \text{utility}_i) + u_i$$

$$= \beta_0 + \beta_4 + \beta_1 \log(\text{sales}_i) + (\beta_2 + \beta_7) (ROE_i) + \beta_3 \text{finance}_i + \beta_5 \text{utility}_i + \beta_6 (ROE_i * \text{finance}_i) + \beta_8 (ROE_i * \text{utility}_i) + u_i$$

$$E(\log(\text{salary}_i) | \text{utility} = 1) = \beta_0 + \beta_1 \log(\text{sales}_i) + \beta_2 ROE_i + \beta_3 \text{finance}_i + \beta_4 \text{consprod}_i + \beta_5 (1) + \beta_6 (ROE_i * \text{finance}_i) + \beta_7 (ROE_i * \text{consprod}_i) + \beta_8 (ROE_i * (1)) + u_i$$

$$= \beta_0 + \beta_5 + \beta_1 \log(\text{sales}_i) + (\beta_2 + \beta_8) (ROE_i) + \beta_3 \text{finance}_i + \beta_4 \text{consprod}_i + \beta_6 (ROE_i * \text{finance}_i) + \beta_7 (ROE_i * \text{consprod}_i) + u_i$$

∴ when we add interaction term between roe and sector dummies. The intercept and slope of this regression will change. y-intercept will increase and slope will be steeper or flatter is depends on roe and sector dummies that will significantly different from zero or not

Question 2. Birth weight has been used by officials as one of the main determinants of health. Data set BWGHT.DTA contains data on infant birth weights in ounces ($bwght_i$), average number of cigarettes mother smoked per day during pregnancy ($cigs$), family income ($faminc_i$), father's year of education ($fatheduc_i$), and mother's year of education ($motheduc_i$). The following two regressions were estimated using data on $n = 1191$ births:

Model 2.1: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + u_i$

regress bwght cigs faminc					
Source	SS	df	MS		
Model	14536.9538	2	7268.47691	Number of obs =	1191
Residual	468209.738	1188	394.115941	F(2, 1188) =	18.44
Total	482746.692	1190	405.669489	Prob > F =	0.0000
				R-squared =	0.0301
				Adj R-squared =	0.0285
				Root MSE =	19.852
bwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
cigs	-.5876985	.1090181			
faminc	.0624684	.0324438			
_cons	118.5568	1.234278			

Omitted for the purpose of this exam.

Model 2.2: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 fatheduc_i + \beta_4 motheduc_i + u_i$

regress bwght cigs faminc fatheduc motheduc					
Source	SS	df	MS		
Model	15827.6593	4	3956.91482	Number of obs =	1191
Residual	466919.033	1186	393.69227	F(4, 1186) =	10.05
Total	482746.692	1190	405.669489	Prob > F =	0.0000
				R-squared =	0.0328
				Adj R-squared =	0.0295
				Root MSE =	19.842
bwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
cigs	-.5894954	.1106172			
faminc	.0538254	.0366502			
fatheduc	.4936695	.2832896			
motheduc	-.4379234	.3197377			
_cons	118.0741	3.500291			

Omitted for the purpose of this exam.

- where $bwght_i$ = birth weight, ounces
- $cigs_i$ = average number of cigarettes the mother smoked per day while pregnant
- $faminc_i$ = 1988 family income, \$1000s
- $fatheduc_i$ = father's years of education
- $motheduc_i$ = mother's years of education

Answer the following questions.

- Based on **Model 2.1**, test whether smoking has an impact on birth weight. Show your work. (use $\alpha = 0.05$)
- Based on **Model 2.1**, construct a 99% confidence interval for β_2 .
- Would your conclusion in a) change if you use the result from **Model 2.2**? Show your work. (use $\alpha = 0.05$)
- What is the overall significance of the regression from **Model 2.2**? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State the critical value for hypothesis testing to receive full points.
- If we are interested in testing whether “**parents’ education**” has an impact on birth weight at all, what kind of null/alternative hypothesis would we be testing? Perform the test and discuss your finding. (use $\alpha = 0.05$)

$$a. \alpha = 0.05 \quad H_0: \beta_1 = 0$$

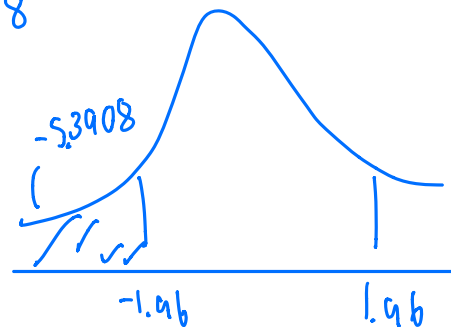
$$H_1: \beta_1 \neq 0$$

Use T-test

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se\hat{\beta}_1} = \frac{-0.5876985}{0.1090181} = -5.3908$$

$$d.f = n - 3 = 1188$$

$$t_{lower} = -1.96, \quad t_{upper} = 1.96$$



\therefore we reject H_0 . we can make sure that β_1 is not simultaneously zero.

so smoking has an impact on birth weight

b. Based on **Model 2.1**, construct a 99% confidence interval for β_2 .

$$B. \hat{\beta}_2 \pm t_{\frac{\alpha}{2}, n-3} \cdot \sigma(\hat{\beta}_2) ; t_{\frac{0.01}{2}, 1188} = 2.576$$

$$= 0.624684 \pm 2.576 \cdot (0.324438)$$

$$\text{lower} = -0.2119$$

$$\text{upper} = 1.4604$$

\therefore confidence interval for $\hat{\beta}_2$ is $\Pr(-0.2119 < \hat{\beta}_2 < 1.4604) = 0.99$

given this confidence coefficient of 99%. 99 out of 100 cases this

interval will contain the true β_2 .

c. Would your conclusion in a) change if you use the result from **Model 2.2**? Show your work. (use $\alpha = 0.05$)

$$d. \alpha = 0.05 \quad H_0: \beta_1 = 0$$

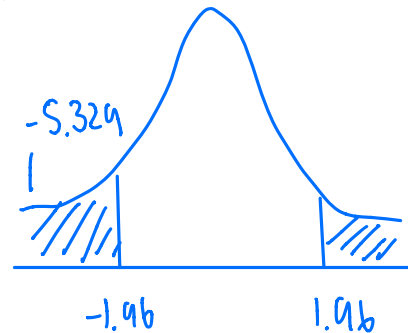
$$H_1: \beta_1 \neq 0$$

t-test

$$t_{\text{cal}} = \frac{\hat{\beta}_1 - \beta_1}{\text{se}\hat{\beta}_1} = \frac{-0.5899954}{0.1106172} = -5.329$$

$$d.f. = n - 5 = 1191 - 5 = 1186$$

$$t_{\text{lower}} = -1.96, \quad t_{\text{upper}} = 1.96$$



\therefore we reject H_0 . we can make sure that β_1 is not simultaneously zero.

so smoking has an impact on birth weight and conclusion in a. not change.

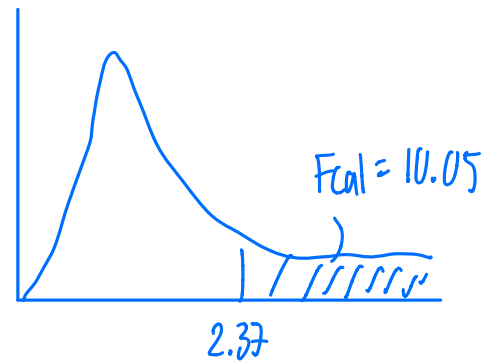
- d. What is the overall significance of the regression from **Model 2.2**? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State the critical value for hypothesis testing to receive full points.

D. F-test

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0 \quad ; \quad \alpha = 0.05$$

H_1 : Not all slope coefficients are simultaneously zero

From stata $F_{(4, 1186)}^{cal} = 10.05$



$$F_{crit(4, 1186)} = 2.37$$

\therefore we can reject H_0 and make sure that β_2, β_3 and β_4 are not simultaneously zero.

$$F_{cal} > F_{crit(4, 1186)}$$

T-test : $\alpha = 0.05$; $t_{lower} = -1.96$, $t_{upper} = 1.96$

$$H_0: \beta_0 = 0$$

$$H_1: \beta_0 \neq 0$$

$$t_{cal} = \frac{\hat{\beta}_0 - \beta_0}{se\hat{\beta}_0} = \frac{118.0741}{3.500291} = 33.7326 \therefore \text{Reject } H_0, \text{ significance}$$

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se\hat{\beta}_1} = \frac{-0.5894954}{0.1106172} = -5.329 \therefore \text{Reject } H_0, \text{ significance}$$

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

$$t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{se\hat{\beta}_2} = \frac{0.538254}{0.366502} = 1.4686 \therefore \text{cannot reject } H_0, \text{ not significance}$$

$$H_0: \beta_3 = 0$$

$$H_1: \beta_3 \neq 0$$

$$t_{cal} = \frac{\hat{\beta}_3 - \beta_3}{se\hat{\beta}_3} = \frac{0.4936695}{0.2832896} = 1.7426 \therefore \text{cannot reject } H_0, \text{ not significance}$$

$$H_0: \beta_4 = 0$$

$$H_1: \beta_4 \neq 0$$

$$t_{cal} = \frac{\hat{\beta}_4 - \beta_4}{se\hat{\beta}_4} = \frac{-0.4379234}{0.3197377} = -1.3696 \therefore \text{cannot reject } H_0, \text{ not significance}$$

- e. If we are interested in testing whether “**parents’ education**” has an impact on birth weight at all, what kind of null/alternative hypothesis would we be testing? Perform the test and discuss your finding. (use $\alpha = 0.05$)

$$E. \quad H_0 : \beta_3 = \beta_4 = 0$$

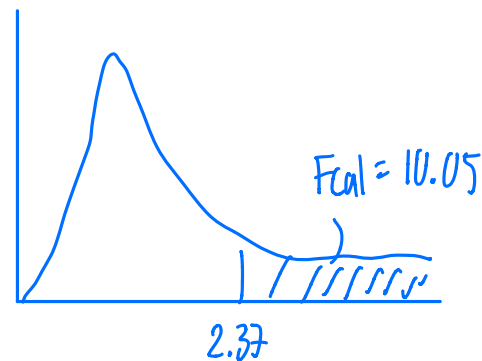
H_1 : Not all slope coefficients are simultaneously zero

$$\alpha = 0.05 \quad d.f = 4, n = 1186$$

$$F_{cal} = \frac{MSCE}{MSCR} = \frac{3956.91982}{393.69227} = 10.05$$

$$F_{crit}(4, 1186) = 2.37$$

$\therefore F_{cal} > F_{crit}(4, 1186)$ we can reject H_0 and we know that both β_3 and β_4 has an impact on birth weight.



Question 3. A model of wage equation is given by

$$lwage_i = \beta_1 + \beta_2 exp_i + \beta_3 expsq_i + \beta_4 educ_i + \beta_5 age_i + \beta_6 kid6_i + \beta_7 kid18_i + u_i$$

- where $lwage_i$ = natural log of hourly wage of married women
 exp_i = years of experience
 $expsq_i$ = years of experience squared
 $educ_i$ = years of education
 age_i = age
 $kid6_i$ = number of children aged 0-6 in a household
 $kid18_i$ = number of children aged 6-18 in a household

4 5

The regression result from OLS is shown in the table below and answer the following questions.

Source	SS	df, $K-1$	MS	Number of obs =	428
ESS Model	35.3304012	6	5.8884002	F(6 , 421) =	13.19
RSS Residual	187.997304	421	.446526442	Prob > F =	0.0000
TSS Total	223.327441	427	0.52301509	R-squared =	0.1582
				Adj R-squared =	0.1462
				Root MSE =	.66823

	lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
1	exper	.039819	.013393	2.97	0.003	.0134936 .0661444
2	expsq	-.0007812	.0004022	-1.94	0.053	-.0015718 9.37e-06
3	educ	.1078319	.0144021	7.49	0.000	.079523 .1361409
4	age	-.0014653	.0052925	-0.28	0.782	-.0118682 .0089377
5	kidslt6	-.0607106	.0887626	-0.68	0.494	-.2351836 .1137625
6	kidsge6	-.014591	.0278981	-0.52	0.601	-.069428 .0402459
7	_cons	-.4209078	.316905	-1.33	0.185	-1.043821 .2020053

- Figure out all the degrees of freedom in this model.
- Figure out all the sum of squares (ESS and RSS) and mean squares in this model.
- Figure out the adjusted R-squared (\bar{R}^2)
- Given that the model above is called 'Model 3.1', there is another competing model called 'Model 3.2' which an explanatory variable is excluded, compared to 'Model 3.1'. Though the result of estimating 'Model 3.2' is not shown here, what is the maximum value of R^2 from 'Model 3.2' which will make you conclude that the excluded variable has a significant contribution in 'Model 3.1', at the significance level of 0.05. (Hint: the critical value of the F-test at the significance level of 0.05 is $F_{1,421} = 3.84$)
- As you can see from the result, age is not significantly different from zero. In other words, age does not determine how much hourly wage would be. Does this make economic sense in your opinion? What do you think cause this insignificance?

9. $df = 6, 421, 427$

b) Figure out all the sum of squares (ESS and RSS) and mean squares in this model.

$$B. R^2 = \frac{ESS}{TSS}$$

$$0.1582 = \frac{ESS}{223.327441}$$

$$ESS = 35.3304012$$

$$RSS = TSS - ESS$$

$$RSS = 223.327441 - 35.3304012$$

$$RSS = 187.99704$$

c) Figure out the adjusted R-squared (\bar{R}^2)

$$d. \bar{R}^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k}$$

$$\bar{R}^2 = 1 - \frac{(1 - 0.1582)427}{421} = \bar{R}^2 = 0.1462$$

d) Given that the model above is called 'Model 3.1', there is another competing model called 'Model 3.2' which an explanatory variable is excluded, compared to 'Model 3.1'. Though the result of estimating 'Model 3.2' is not shown here, what is the maximum value of R^2 from 'Model 3.2' which will make you conclude that the excluded variable has a significant contribution in 'Model 3.1', at the significance level of 0.05. (Hint: the critical value of the F-test at the significance level of 0.05 is $F_{1,421} = 3.84$)

$$D. F_{(1,421)} = \frac{\frac{R^2_{3.1} - R^2_{3.2}}{\text{number of new regressors}}}{\frac{1 - R^2_{3.1}}{n - k_{3.2}}} = \frac{0.1582 - R^2_{3.2}}{1 - 0.1582} = \frac{1}{(428-7)}$$

$$3.84 = \frac{(0.1582 - R^2_{3.2})(421)}{0.8418}$$

$$R^2_{3.2} = 0.1505$$

- e) As you can see from the result, age is not significantly different from zero. In other words, age does not determine how much hourly wage would be. Does this make economic sense in your opinion? What do you think cause this insignificance?

E, In terms of economic sense, we would say NO, because in reality age is one of factor that will determines wage. For example, people with higher age may have more work experience than people with lower age. so wage must be higher than those people with lower age.

on the other hand, age variable is insignificant in this model because this term may have multicollinearity with other terms. therefore, age variable will be omitted because it unnecessary to have this variable in this model.