6204641564 – ผิง

1) a. Write out the estimated regression equation for $\log(salary_i)$. Interpret the estimated coefficient associated with $\log(sales_i)$.

$$\log(salary_i) = 4.588101 + 0.2571917\log(sales_i) + 0.0111517\,roe + 0.1579564\,finance$$
$$+ 0.1808917\,consprod - 0.2830015\,utility$$

If sales increase by 1 percent, salary will increase by 0.2571917 percent

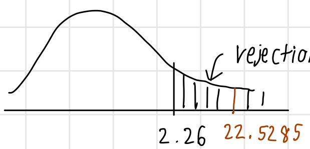
b. What is the overall significance of the regression? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State **the critical value** for hypothesis testing to receive full points.

- To test the overall significance, we use the F_{test}, $\alpha = 0.05$

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1 : otherwise

$F_{0.05(5,203)} = 2.26$



$$F_{cal} = \frac{ESS/k-1}{RSS/n-K} = \frac{23.8109943/(6-1)}{42.9111689/(209-6)} = 22.52854894$$

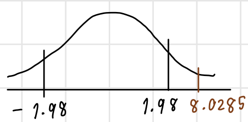
$F_{cal} > F_{critical}$: We can reject the null hypothesis and make sure that β_1, β_2, β_3, β_4, and β_5 are not simultaneously 0 at 95% confidenc level.

- Test the individual significance by using t test

1) $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$\alpha = 0.05$, $\alpha/2 = 0.025$



$t_{lower} = -1.98$

$t_{upper} = 1.98$

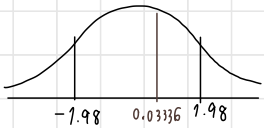
$$t_{cal}(\beta_1) = \frac{\hat{\beta}_1 - \beta_1}{Se_{\hat{\beta}_1}} = \frac{0.2571917 - 0}{0.0320348} = 8.0285$$

∴ Since t_{cal} fall into rejection region, we can reject H_0

→ β_1 is significant ($\neq 0$)

2) $H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

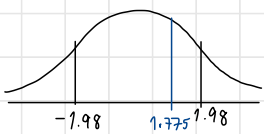


$t_{cal}(\beta_2) = \frac{0.0111517 - 0}{0.3342996} = 0.03336$

∴ since $t_{cal}(\beta_2)$ fall in the acceptance region, we can't reject the null hypothesis that $\beta_2 = 0$ at 95% confidence level

3) $H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

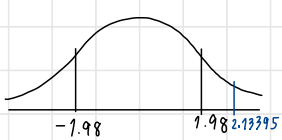


$t_{cal}(\beta_3) = \frac{0.1579564 - 0}{0.0890017} = 1.7748$

∴ since $t_{cal}(\beta_3)$ fall in the acceptance region, we can't reject the null hypothesis that $\beta_3 = 0$ at 95% confidence level

4) $H_0 : \beta_4 = 0$

$H_1 : \beta_4 \neq 0$

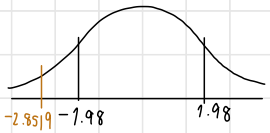


$t_{cal}(\beta_4) = \frac{0.1808917 - 0}{0.0847683} = 2.133954556$

∴ since $t_{cal}(\beta_4)$ fall in the rejection region, we can reject the null hypothesis, in other word, we can make sure that $\beta_4 \neq 0$ at 95% confidence level

5) $H_0 : \beta_5 = 0$

$H_1 : \beta_5 \neq 0$



$t_{cal}(\beta_5) = \frac{-0.2830015 - 0}{0.0992337} = -2.8518689$

∴ since $t_{cal}(\beta_5)$ fall in the rejection region, we can reject the null hypothesis.
In other word, we can make sure that β_5 to at 95% confidence level.

In this model, $\beta_1, \beta_3, \beta_4$, and β_5 are statistically significant.

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

utility sector ⇒ $\log(salary) = 4.588101 + 0.2571917 \log(sales) + 0.0111517 roe - 0.2830015(1)$

$= 4.3050995 + 0.2571917 \log(sales) + 0.0111517 roe$

transportation sector ⇒ $\log(salary) = 4.588101 + 0.2571917 \log(sales) + 0.0111517 roe$

$4.588101 - 4.3050995 = 0.2830015$

∴ transportation & utility sector estimated salary are different by 0.2830015 percent

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?

Because there will be a collinearity problem and STATA will report "omitted" in one dummy.

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

we need to check for the marginal contribution of the new term.
If the newterms has the marginal contribution for the model, then the interection term is beneficial and should be added.

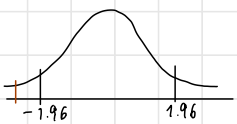
2)

a. Based on **Model 2.1**, test whether smoking has an impact on birth weight. Show your work. (use $\alpha = 0.05$)

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$\alpha = 0.05$, $df = 1191 - 3 = 1188$



$t_{lower} = -1.96$

$t_{upper} = 1.96$

$t_{cal}(\beta_1) = \frac{\hat{\beta}_1 - \beta_1}{se\hat{\beta}_1} = \frac{-0.587695 - 0}{0.1090181} = -5.390080$

∴ Since $t_{cal}(\beta_1)$ fall into rejection region, $\beta_1 \neq 0$ and smoking has and impact on the birth weight.

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

$\hat{\beta}_2 - t_{\frac{\alpha}{2}} se(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\frac{\alpha}{2}} se(\hat{\beta}_2)$

$0.062484 - 2.326(0.0324438) \leq \beta_2 \leq 0.062484 + 2.326(0.0324438)$

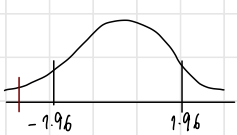
$-0.01298 \leq \beta_2 \leq 0.14621$

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

$H_0 : \beta_1 = 0$

$H_0 : \beta_1 \neq 0$

$\alpha = 0.05$, $df = 1191 - 5 = 1186$



$t_{cal} = \frac{-0.5894954 - 0}{0.1106172} = -5.3291478$

∴ since t_{cal} fall in the rejection region, we can make sure that $\beta_1 \neq 0$, hence, the conclusion from a) isn't 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.

- F_{test} → test the overall significance

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1 : otherwise

$F_{0.05(4,1186)} = 2.37$

$F_{cal} = \frac{15827.6593/4}{168209.738/1186} = 10.0231$

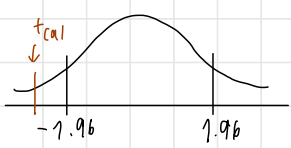
$F_{cal} > F_{critical}$, we can reject the null hypothesis and make sure that $\beta_1, \beta_2, \beta_3$, and β_4 are not simultaneously 0

- Individual test (t test)

1) $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$\alpha = 0.05$, $df = 1191 - 5 = 1186$



$t_{lower} = -1.96$

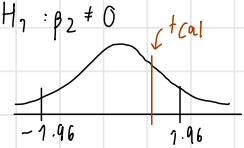
$t_{upper} = 1.96$

$t_{cal}(\beta_1) = \frac{-0.5894954 - 0}{0.1106172} = -5.3291$

∴ since t_{cal} is in the rejection region, we can reject the null hypothesis and make sure that β_1 is not 0 at the significant of 95%.

2) $H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$



$t_{cal}(\beta_2) = \frac{0.0538254 - 0}{0.0366502} = 1.468625$

∴ since t_{cal} fall in the acceptance region, we cannot say for sure that β_2 is not equals to 0 at 95% level of significant

3) $H_0 : \beta_3 = 0$

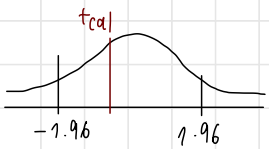
$H_1 : \beta_3 \neq 0$

$$t_{cal}(\beta_3) = \frac{0.4936695 - 0}{0.2832896} = 1.742\,68$$

∴ since t_{cal} is in the acceptance region, we cannot say for sure that β_3 is not equal to 0 at 95% level of significant

4) $H_0 : \beta_4 = 0$

$H_1 : \beta_4 \neq 0$



$$t_{cal}(\beta_4) = \frac{-0.4379234 - 0}{0.3197377} = -1.3696$$

∴ since t_{cal} lies in the acceptance region, we cannot reject the null hypothesis that $\beta_4 = 0$ at 95% significant level.

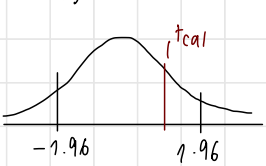
In the model 2.2, only β_1 is tested statistically significant at 5% level. (only number of cigarettes mothersmoke per day has an impact on birth weight)

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 $\alpha = 0.05$)

Testing Hypothesis of β_3, β_4, which are parents education,

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

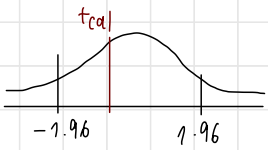


$$t_{cal}(\beta_3) = \frac{0.4936695 - 0}{0.2832896} = 1.742\,68$$

∴ since t_{cal} is in the acceptance region, we cannot say for sure that β_3 is not equal to 0 at 95% level of significant

$H_0 : \beta_4 = 0$

$H_1 : \beta_4 \neq 0$



$$t_{cal}(\beta_4) = \frac{-0.4379234 - 0}{0.3197377} = -1.3696$$

∴ since t_{cal} lies in the acceptance region, we cannot reject the null hypothesis that $\beta_4 = 0$ at 95% significant level.

Both coefficient are not statistically significant at 95% significant level. which is making sense because parents education shouldn't have any effect on the birth weight.

3)

a) Figure out all the degrees of freedom in this model.

$df_{model} = k - 1 = 8 - 1 = 7$

$df_{residual} = n - k = 428 - 8 = 420$

$df_{total} = 427$

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

$$F_{(6,421)} = \frac{MSE}{0.44652642} = 13.19$$

$MSE = 5.8897$

$$\rightarrow \frac{ESS}{df} = MSE \rightarrow \frac{ESS}{6} = 5.8897$$

$ESS = 35.3382$

$$\rightarrow \frac{RSS}{df} = MSR = \frac{RSS}{421} = 0.44652642$$

$RSS = 187.9876321$

$MS_{total} = 5.8897 + 0.44652642 = 6.33623$

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

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

$$= 1-(1-0.1582)\left(\frac{428-1}{428-7}\right)$$

$$= 0.1462028504$$

Source	SS	df	MS		
Model	35.3382	6	5.8897	Number of obs	= 428
Residual	187.9876321	421	.446526442	F(6, 421)	= 13.19
				Prob > F	= 0.0000
				R-squared	= 0.1582
				Adj R-squared	= 0.1460
Total	223.327441	427	0.336223	Root MSE	= .66823

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

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 $F_{1,421} = 3.84$)

H_0: the explanatory variable has no marginal contribution to the model
H_1: otherwise

In order to reject the null hypothesis F_{cal} must be greater than the value of $F_{critical}$ ($F_{cal} > 3.84$)

$$F_{cal} = \frac{R^2_{new} - R^2_{old} / (\text{number of new regressors})}{1 - R^2_{new}/(n-k_{new})} = \frac{0.1582 - R^2_{old} / 1}{1-0.1582/(428-7)} > 3.84$$

$R^2_{old} < 0.1505218242$

∴ maximum value of R^2 from model 3.2 = 0.150521824 2

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?

No, because age can use to determine the level of productivity of the labour (Patrick & Bruno, 2006). However, in this model age may not be significant due to the variation of the observation, for example, there may be some people who in a middle age and didn't get a job because of the bad economy.