#1

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

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \text{ROE}_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + u_i$$

$l\,salary = 4.588101 + 0.2572\, l\,sales + .0112\ \text{ROE} + .158\ finance + 0.1809\ consprod - .283\ utility$

$\beta_1 = 0.2572$: If sales increase 1%, salary will increase 0.2572%

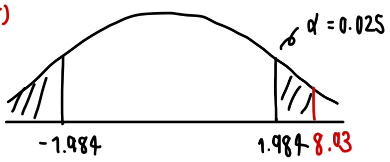
b. What is the overall significance of the regression? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State **the critical value** for hypothesis testing to receive full points.

individual → t-test
overall → F-test

Individual

T-test on sales

$H_0 : \beta_1 = 0$ (No effect)
$H_1 : \beta_1 \neq 0$ (effect)
$\alpha = 0.05$
$df = n-k = 209-6 = 203$



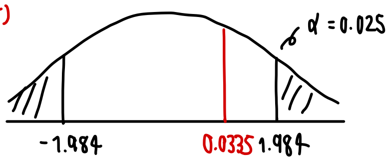
$t_{cal} = \dfrac{\hat\beta_1 - \beta_1}{se_{\beta_1}} = \dfrac{0.2572 - 0}{0.032} = 8.03$

∴ reject H_0, sale effects salary at 95% CL

$t_{cal} > t_{cri}$, t_{cal} lie beyond any boundary of CL (rejection region), we can reject null hypothesis, at significance level of 95%. In other word, we are sure that β_1 is not zero 95 out of 100 time when we sample

T-test on ROE

$H_0 : \beta_2 = 0$ (No effect)
$H_1 : \beta_2 \neq 0$ (effect)
$\alpha = 0.05$
$df = n-k = 209-6 = 203$



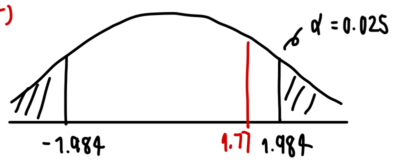
$t_{cal} = \dfrac{\hat\beta_2 - \beta_2}{se_{\beta_2}} = \dfrac{0.0112 - 0}{0.3343} = 0.0335$

∴ reject H_0, ROE no effects salary at 95% CL

$t_{cal} > t_{cri}$, t_{cal} lie within any boundary of CL (acceptance region), we can't reject null hypothesis, at significance level of 95%. In other word, we are not sure that β_2 is not zero 95 out of 100 time when we sample

T-test on finance

$H_0 : \beta_3 = 0$ (No effect)
$H_1 : \beta_3 \neq 0$ (effect)
$\alpha = 0.05$
$df = n-k = 209-6 = 203$



$t_{cal} = \dfrac{\hat\beta_3 - \beta_3}{se_{\beta_3}} = \dfrac{0.158 - 0}{0.89} = 1.77$

∴ accept H_0, work in financial industry no effects salary at 95% CL

$t_{cal} > t_{cri}$, t_{cal} lie within any boundary of CL (acceptance region), we can reject null hypothesis, at significance level of 95%. In other word, we are not sure that β_3 is not zero 95 out of 100 time when we sample

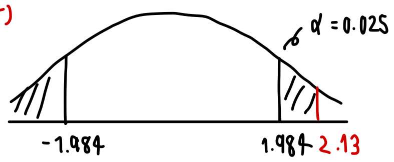
T-test on consumer production industry

$H_0 : \beta_4 = 0$ (No effect)

$H_1 : \beta_4 \neq 0$ (effect)

$\alpha = 0.05$

$df = n-K = 209-6 = 203$



$$t_{cal} = \frac{\hat{\beta}_4 - \beta_4}{se_{\beta_4}} = \frac{.1809-0}{0.0848} = 2.13$$

$\therefore$ reject H_0, work in utility industrial effects salary at 95% CL

$t_{cal} > t_{cri}$, t_{cal} lie beyond any boundary of CL (rejection region), we can reject null hypothesis, at significance level of 95%. In other word, we are sure that β_4 is not zero 95 out of 100 time when we sample

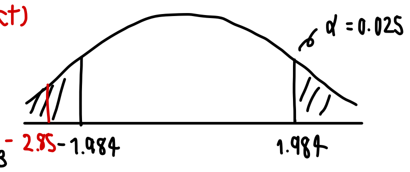
T-test on utility

$H_0 : \beta_5 = 0$ (No effect)

$H_1 : \beta_5 \neq 0$ (effect)

$\alpha = 0.05$

$df = n-K = 209-6 = 203$



$$t_{cal} = \frac{\hat{\beta}_5 - \beta_5}{se_{\beta_5}} = \frac{-.283-0}{0.0923} = -2.85$$

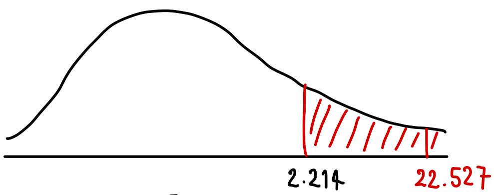
$\therefore$ reject H_0, work in utility industrial effects salary at 95% CL

$t_{cal} > t_{cri}$, t_{cal} lie beyond any boundary of CL (rejection region), we can reject null hypothesis, at significance level of 95%. In other word, we are sure that β_5 is not zero 95 out of 100 time when we sample

overall

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

H_1 : at least one parameter not equals to 0



$F_{K-1, n-K}$

$F_{5, 203}$

$F_{cri} = 2.214$

$$F_{cal} = \frac{4.7622}{.2114} = 22.527$$

$\therefore$ $t_{cal} > t_{cri}$, t_{cal} lie beyond any boundary of CL (rejection region), we can reject null hypothesis, at significance level of 95%. There is enough evidence to say that at least one parameter is not equal to 0, at $\alpha = 0.05$

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

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \text{ROE}_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + u_i$$

$lsalary = 4.588101 + 0.2572 lsales + .0112 ROE + .158 finance + 0.1809 consprod - .283 utility$

Base - transportation

$\alpha_1 = \alpha_2 = \alpha_3 = 0$: transport: $\ln salary = 4.588101 + 0.2572 lsales + 0.0112 ROE$

Base - utilities

$\alpha = \alpha_2 = 0$: utility : $\ln salaray = 4.588101 + 0.2572 lsales + 0.112 ROE - 0.2830$

$\alpha_3 = 1$

∴ average mean of l salary transportation is more than l salary utilities by 0.2830

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?

If we put all the sector dummy, it will be the dummy variable trap. We won't fall into it. There is no longer perfectly collinerity. The STATA will not work. Normally in model, the dummy equals to M-1 to keep the base when comparing with others.

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 $\text{ROE}_i * utility_i$?

First, check the real data, if there have the real relationship between ROE and sector dummies that create any benefit. Secondly, the benefit is depehs on the real interaction or not.

(#2)

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

Using individual statistically significant

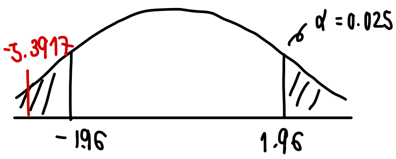
$H_0 : \beta_1 = 0$ (No effect)

$H_1 : \beta_1 \neq 0$ (effect)

$\alpha = 0.05$

$df = n-k = 1191-3 = 1188$

$t_{cri} = 1.96$



$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se_{\beta_1}} = \frac{-.5877 - 0}{.109} = -5.3917$$

$\therefore |t_{cal}| > |t_{cri}|$, reject H_0 at 95% confidenence interval. In other word, the avg. number of cigarettes the mother smoked per day while pregnant has effect on the birth weight.

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

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

note that $t_{\frac{\alpha}{2}} = t_{\frac{0.01}{2}} = t_{0.005,\,1188} = 2.581$

$df = 1191-3$

$$0.062484 - 2.581(0.0324438) \leq \beta_2 \leq 0.0624684 + 2.581(0.0324438)$$

$$-0.02126905 \leq \beta_2 \leq 0.14620585$$

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

Using individual statistically significant

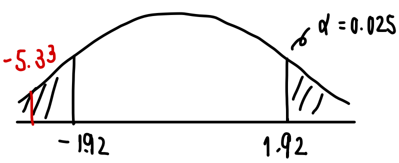
$H_0 : \beta_1 = 0$ (No effect)

$H_1 : \beta_1 \neq 0$ (effect)

$\alpha = 0.05$

$df = n-k = 1191-5 = 1186$

$t_{cri\,(0.025,\,1186)} = 1.92$



$$t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se_{\beta}} = \frac{-.5895 - 0}{.1106} = -5.33$$

So, the conclusion is not change from question a) as the $|t_{cal}| > |t_{cri}|$.

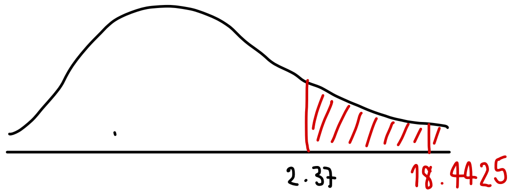
The conclusion would still be the same as there is enough evidence to support that the avg. number of cigarettes the morther smoked per day while pregant impact to the infant's birth weigth at 95% CL.

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.

→ overall significance of regression → Using F-test to test.

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

H_1: at least one parameter not equals to 0



$F_{K-1, n-K}$

$F_{5-1, 1191-5}$

$F_{4, 1186}$ $\alpha = 0.05$

$F_{cri} = 2.37$

$F_{cal} = \dfrac{7268.47691}{394.115941} = 18.4425$

∴ $t_{cal} > t_{cri}$, t_{cal} lie beyond any boundary of CL (rejection region), we can reject null hypothesis, at significance level of 95%. There is enough evidence to say that at least one parameter is not equal to 0, at $\alpha = 0.05$

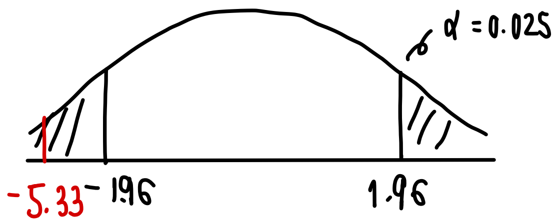
Individual

$H_0: \beta_1 = 0$ (No effect)

$H_1: \beta_1 \neq 0$ (effect)

$\alpha = 0.05$

df = n-K = 1191-5 = 1186



$t_{cal} = \dfrac{\hat{\beta}_1 - \beta_1}{se_{\beta_1}} = \dfrac{-.5895 - 0}{.1106} = -5.33$

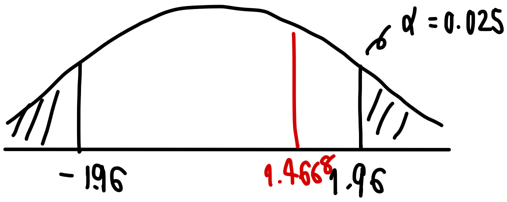
∴ reject H_0, cigarett has effect on birth weight

$H_0 : \beta_2 = 0$ (No effect)

$H_1 : \beta_2 \neq 0$ (effect)

$\alpha = 0.05$

$df = n-k = 1191-5 = 1186$



$$t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{se_{\beta_2}} = \frac{0.5383-0}{0.0367} = 1.4668$$

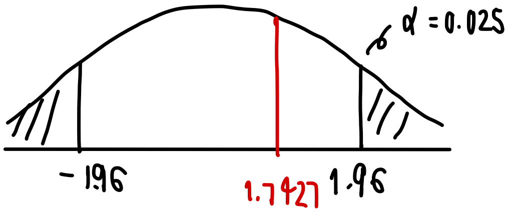
∴ accept H_0, fam inc has no effect on birth weight

$H_0 : \beta_3 = 0$ (No effect)

$H_1 : \beta_3 \neq 0$ (effect)

$\alpha = 0.05$

$df = n-k = 1191-5 = 1186$



$$t_{cal} = \frac{\hat{\beta}_3 - \beta_3}{se_{\beta 3}} = \frac{.4937-0}{.2833} = 1.7427$$

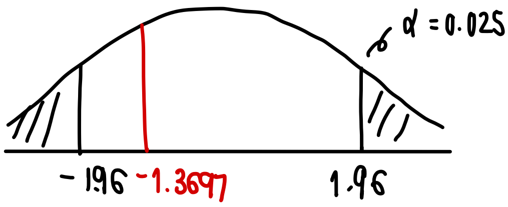
∴ accept H_0, father's years of edu has no effect on birth weight

$H_0 : \beta_4 = 0$ (No effect)

$H_1 : \beta_4 \neq 0$ (effect)

$\alpha = 0.05$

$df = n-k = 1191-5 = 1186$



$$t_{cal} = \frac{\hat{\beta}_4 - \beta_4}{se_{\beta 4}} = \frac{-.4379-0}{.3197} = -1.3697$$

∴ accept H_0, mother's years of edu has no effect on birth weight

→ only β_1 : the average number of cigarrettes the mother smoked per day while pregant effect the 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 $\alpha = 0.05$)

faminc	|	.0538254✓	.0366502
fatheduc	|	.4936695✓	.2832896
motheduc	|	-.4379234✓	.3197377
cons	|	118.0741	3.500291

To testing parents' education has the impact on the birth weight or not? we use the individual testing to test at 95% CL.

T-test on father edu

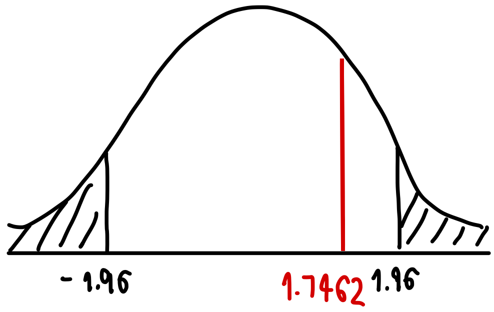
$H_0: \beta_3 = 0$

$H_1: \beta_3 \neq 0$

$\alpha = 0.05$, df = 1191 - 4 = 1186

$t_{cri} = t_{0.025, 1186} = 1.96$

$t_{cal} = \dfrac{\hat{\beta}_3 - \beta_3}{se\,\beta_3} = \dfrac{.4936695 - 0}{.2832896} = 1.7462$



∴ $|t_{cal}| < |t_{cri}|$: fail to reject H_0 : we have enough evidence to said that father's eduction does not impact on the birth weight at 95% CL

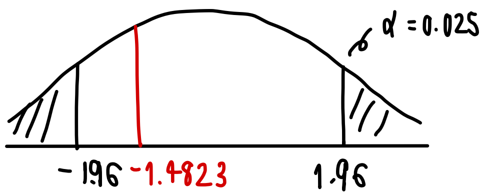
T-test on mother edu

$H_0: \beta_4 = 0$ (No effect)

$H_1: \beta_4 \neq 0$ (effect)

$\alpha = 0.05$

df = n-k = 1191-5 = 1186



$t_{cal} = \dfrac{\hat{\beta}_4 - \beta_4}{se\,\beta_4} = \dfrac{-.4379234 - 0}{.3197377} = -1.3696$

∴ $|t_{cal}| < |t_{cri}|$: fail to reject H_0 : we have enough evidence to said that mother's eduction does not impact on the birth weight at 95% CL

#3

Source	SS	df	SS/df MS
Model	ESS 35.33980892	K-1 6	5.889968153 MSE
Residual	RSS 187.9876321	n-K 421	.446526442 MSR
Total	TSS 223.327441	427	6.336494595

Number of obs	=	428 K-7
F(6 , 421)	=	13.19
Prob > F	=	0.0000
R-squared	=	0.1582
Adj R-squared	=	0.1462
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

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

Model df = K-1 = 7-1 = 6

Residual df = n-K = 428-7 = 421

Total df = Model df + Residual df = 6 + 421 = 427

$F_{(6,421)}$

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

RSS = MS × df = .446526442

ESS = TSS − RSS = 35.33980892

mean square = (Root MSE)^2 = MSR = 0.4465

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

$$\bar{R}^2 = 1-(1-\bar{R}^2)\left(\frac{n-1}{n-k}\right) = 0.1462$$

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 $F_{1,421} = 3.84$)

$$F = \frac{(R^2_{new} - R^2_{old}) \,/\, \#\text{ of new regressor}}{(1-R^2_{new}) \,/\, df_{(n-K_{new})}}$$

note

$K_{new} > K_{old}$

$R_{new} = 3.1$ $R_{old} = 3.2$

$$3.84 = \frac{(0.1582 - R^2_{3.2})\,/\,1}{(1-0.1582)\,/\,(428-7)}$$

$$R^2_{3.2} = 0.15052$$

✓

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?

Yes, this is make sense that the age doesn't determine how much hourly wage would be. As the age increase, the wage is not always increase. It depends on the work experience and education. If you have the high education, this garuntee that you will have the higher income.