6204640863

Source	SS	df	MS
Model	23.8109943	5	4.76219887
Residual	42.9111689	203	.211385068
Total	66.7221632	208	.320779631

Number of obs	= 209
F(5, 203)	= 22.53
Prob > F	= 0.0000
R-squared	= 0.3569
Adj R-squared	= 0.3410
Root MSE	= .45977

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
β_1 – 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
β_0 – _cons	4.588101	.2950221	15.55	0.000	4.0064	5.169801

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

1

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

- Estimated regressio : $\log(salary_i) = 4.588101 + 0.2571917 \log(sales_i) + 0.0111517\, ROE_i + 0.1579564\, finance_i + 0.1808917\, consprod_i - 0.2830015\, utility_i + u_i$ ※
- Estimated coefficient associated with $\log(salary_i)$: $\hat{\beta}_1 = 0.2571917$ which is the amount that when log of firm's sales change by \$ 1 million, the log of CEO's annual salary will change 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.

1. State the hypothesis

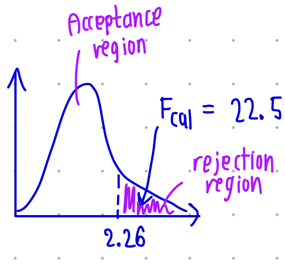
H_0 : $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_a : other wise ※

2. Use F-value method :

- $F_{cal} = \dfrac{\frac{ESS}{TSS} / (K-1)}{\frac{RSS}{TSS} / (n-k)} = \dfrac{4.76219887}{0.211385068} = 22.5285490364$ ※
- $\alpha = 0.05$
- Critical value : $F_{upper,\alpha}(5, 203) = 2.26$ ※

3.



∴ we can reject H_0. We can make sure that $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ & β_5 are not simultaneously zero. ※

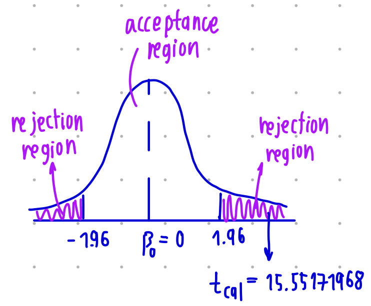
Testing individual regression coefficients

β_0

H_0 : $\beta_0 = 0$
H_a : $\beta_0 \neq 0$

$t_{cal}(\beta_0) = \dfrac{4.588101 - 0}{0.2950221} = 15.55171968$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 203$; $t_{upper} = 1.96$



∴ If t_{cal} lies beyond any boundary of CI rejection region, we can reject H_0 at significant level of 95% ↔ We are sure that β_0 is not 0 95 out of 100 times when we sample.

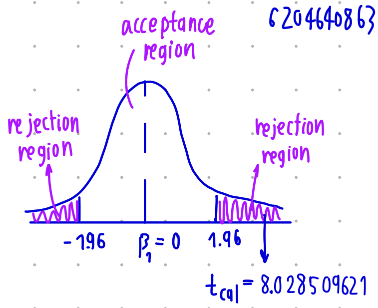
6204640863

β_1

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

$t_{cal}(\beta_1) = \dfrac{0.2571917 - 0}{0.0320348} = 8.028509621$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 203$; $t_{upper} = 1.96$



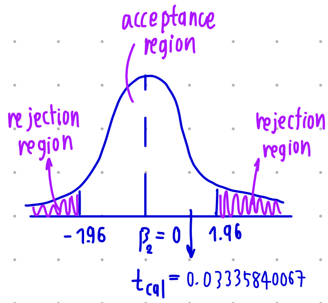
∴ If t_{cal} lies beyond any boundary of CI rejection region, we can reject H_0 at significant level of 95% ↔ we are sure that β_1 is not 0 95 out of 100 times when we sample.

β_2

$H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$

$t_{cal}(\beta_2) = \dfrac{0.0111517 - 0}{0.3342996} = 0.03335840067$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 203$; $t_{upper} = 1.96$



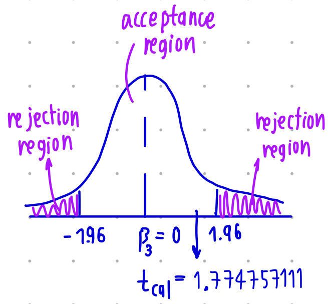
∴ If t_{cal} lies within any boundary of CI Acceptance region, we can't reject H_0 at significant level of 95% ↔ we aren't sure that β_2 is not 0 95 out of 100 times when we sample.

β_3

$H_1 : \beta_3 = 0$
$H_a : \beta_3 \neq 0$

$t_{cal}(\beta_3) = \dfrac{0.1579564 - 0}{0.0890017} = 1.774757111$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 203$; $t_{upper} = 1.96$



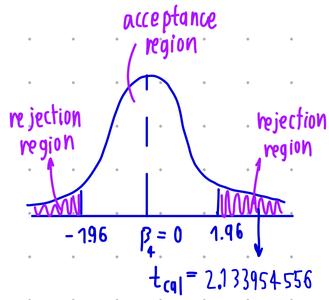
∴ If t_{cal} lies within any boundary of CI Acceptance region, we can't reject H_0 at significant level of 95% ↔ we aren't sure that β_3 is not 0 95 out of 100 times when we sample.

β_4

$H_1 : \beta_4 = 0$
$H_a : \beta_4 \neq 0$

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

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 203$; $t_{upper} = 1.96$



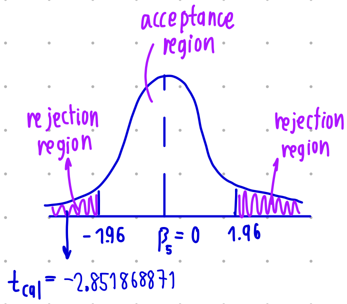
∴ If t_{cal} lies beyond any boundary of CI rejection region, we can reject H_0 at significant level of 95% ↔ we are sure that β_4 is not 0 95 out of 100 times when we sample.

β_5

$H_1 : \beta_5 = 0$
$H_a : \beta_5 \neq 0$

$t_{cal}(\beta_5) = \dfrac{-0.2830015 - 0}{0.0992337} = -2.851868871$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 203$; $t_{upper} = 1.96$



∴ If t_{cal} lies within any boundary of CI Acceptance region, we can't reject H_0 at significant level of 95% ↔ we aren't sure that β_5 is not 0 95 out of 100 times when we sample.

6204640863

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

(utility industry)
$\log(salary_i) = 4.588101 + 0.2571917 \log(sales_i) + 0.0111517\,ROE_i - 0.2830015(1)$

(transportation sector)
$\log(salary_i) = 4.588101 + 0.2571917 \log(sales_i) + 0.0111517\,ROE_i$

$\frac{2}{4} = \frac{1}{2} = 0.5$

$\frac{1}{2} = 0.5$

$\Rightarrow \ln(salary_{utility}) - \ln(salary_{trans.}) = -0.2830015$

$\ln\left(\frac{salary_{utility}}{salary_{trans.}}\right) = -0.2830015$

$\frac{salary_{utility}}{salary_{trans}} = e^{-0.2830015}$

$= 0.7535186576 \xrightarrow{-1} = -0.2464813424$

Utility < transportation = − 24.64813424% #

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?

> we omitted the transportation sector because it is the baseline, and we try to exclude the non-significant variable which is not impact on our model.
> If we put all the sector dummies, the regression analysis will fail. Because, for instance, if you have 3 dummy variable, and we put the third dummy into the modle, it will cause the perfect multicollinearity.

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 $\text{ROE}_i * utility_i$?

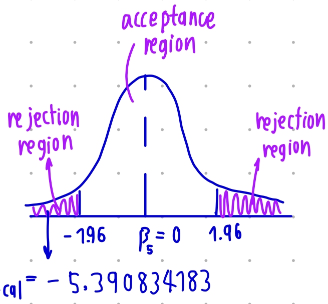
- If we add the interaction term, it will create the additional effect.
- The additional effect will be benefitcial, if the t or p-value is given.
- In this case, there is no benefit of adding the interaction dummy.

6204640863

2

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

1. $H_0 : \beta_1 = 0$
   $H_a : \beta_1 \neq 0$



2. $t_{cal} = \dfrac{-0.5876985}{0.1090181} = -5.390834183$

$\alpha = 0.05$; $df = 1188$

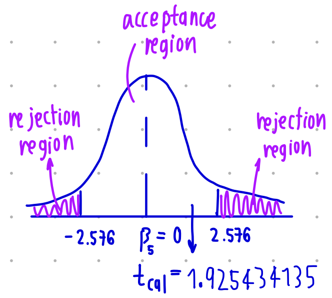
$t_{lower} = -1.96$; $t_{upper} = 1.96$

∴ If t_{cal} lies beyond any boundary of CI rejection region, we can reject H_0 at significant level of 95% ↔ we are sure that β_1 is not 0 95 out of 100 times when we sample.

⇒ smoking has impact on birth weight //

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

1. $H_0 : \beta_2 = 0$
   $H_a : \beta_2 \neq 0$



2. $t_{cal} = \dfrac{0.0624684}{0.0324438} = 1.925434135$

$\alpha = 0.01$; $df = 1188$

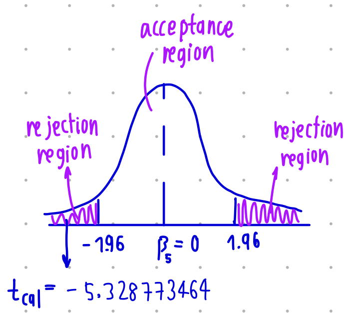
$t_{lower} = -2.576$; $t_{upper} = 2.576$

∴ If t_{cal} lies within any boundary of CI Acceptance region, we can't reject H_0 at significant level of 99% ↔ we aren't sure that β_2 is not 0 99 out of 100 times when we sample.

⇒ Family income doesn't have impact on birth weight //

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

1. $H_0 : \beta_1 = 0$
   $H_a : \beta_1 \neq 0$



2. $t_{cal} = \dfrac{-0.5894954}{0.1106172} = -5.328773464$

$\alpha = 0.05$; $df = 1186$

$t_{lower} = -1.96$; $t_{upper} = 1.96$

∴ If t_{cal} lies beyond any boundary of CI rejection region, we can reject H_0 at significant level of 95% ↔ we are sure that β_1 is not 0 95 out of 100 times when we sample.

⇒ smoking has impact on birth weight ; the result will be the same ///

6204640863

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.

1. State the hypothesis

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_a :$ other wise ※

2. Use F-value method :

- $F_{cal} = \dfrac{ESS/(k-1)}{RSS/(n-k)}$

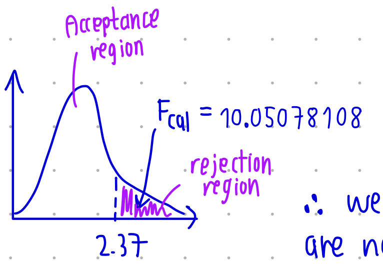
$= \dfrac{3956.91482}{393.69227} = 10.05078108$ ※

- $\alpha = 0.05$

- Critical value :

$F_{upper,\alpha}(4, 1186) = 2.37$

3.



∴ we can reject H_0. We can make sure that $\beta_0, \beta_1, \beta_2, \beta_3$ & β_4 are not simultaneously zero. ※

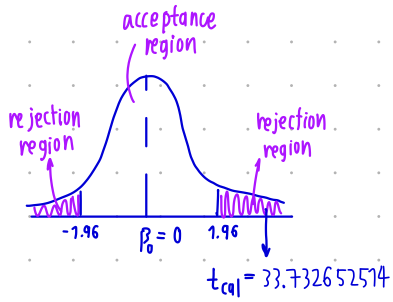
Testing individual regression coefficients

β_0

$H_0 : \beta_0 = 0$
$H_a : \beta_0 \neq 0$

$t_{cal}(\beta_0) = \dfrac{118.0741 - 0}{3.500291} = 33.732652514$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 1186$ $\quad t_{upper} = 1.96$



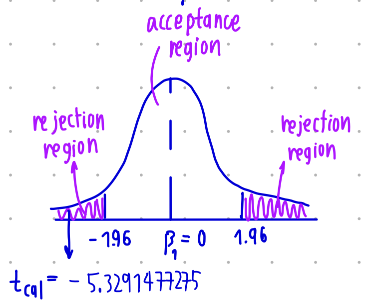
∴ If t_{cal} lies beyond any boundary of CI rejection region, we can reject H_0 at significant level of 95% ↔ we are sure that β_0 is not 0 95 out of 100 times when we sample.

β_1

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

$t_{cal}(\beta_1) = \dfrac{-0.5894954 - 0}{0.1106172} = -5.3291477275$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 1186$ $\quad t_{upper} = 1.96$



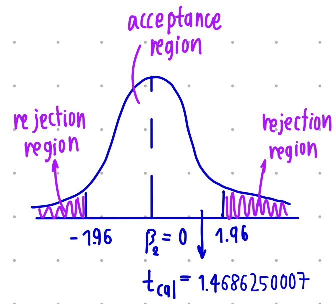
∴ If t_{cal} lies beyond any boundary of CI rejection region, we can reject H_0 at significant level of 95% ↔ we are sure that β_1 is not 0 95 out of 100 times when we sample.

β_2

$H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$

$t_{cal}(\beta_2) = \dfrac{0.0538254 - 0}{0.0366502} = 1.4686250007$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 1186$ $\quad t_{upper} = 1.96$



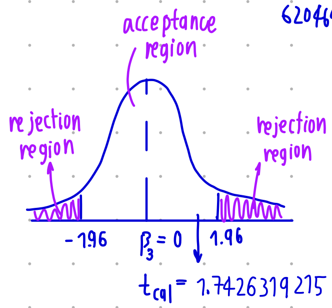
∴ If t_{cal} lies within any boundary of CI Acceptance region, we can't reject H_0 at significant level of 95% ↔ We aren't sure that β_2 is not 0 95 out of 100 times when we sample.

β_3

$H_1 : \beta_3 = 0$
$H_a : \beta_3 \neq 0$

$t_{cal}(\beta_3) = \dfrac{0.4936695 - 0}{0.2832896} = 1.7426319215$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 1186$; $t_{upper} = 1.96$



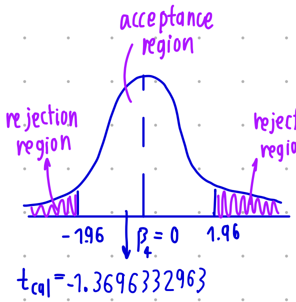
∴ If t_{cal} lies within any boundary of CI Acceptance region, we can't reject H_0 at significant level of 95% ↔ We aren't sure that β_3 is not 0 95 out of 100 times when we sample.

β_4

$H_1 : \beta_4 = 0$
$H_a : \beta_4 \neq 0$

$t_{cal}(\beta_4) = \dfrac{-0.4379234 - 0}{0.3197377} = -1.3696332963$

$\alpha = 0.05$; $t_{lower} = -1.96$
$df = 1186$; $t_{upper} = 1.96$



∴ If t_{cal} lies within any boundary of CI Acceptance region, we can't reject H_0 at significant level of 95% ↔ We aren't sure that β_4 is not 0 95 out of 100 times when we sample.

β_3, β_4

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 $\alpha = 0.05$)

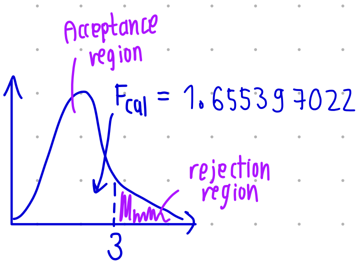
- The marginal contribution

1. H_0 : parent's education has no marginal contribution to model
   H_a : otherwise

2. $F_{cal} = \dfrac{(0.0328 - 0.0301)/(4-2)}{(1 - 0.0328)/(1191 - 5)} = \dfrac{0.00135}{0.000815514339} = 1.655397022$

   $\alpha = 0.05$

   $F_{upper,\alpha}(2, 1186) = 3$



∴ we can't reject H_0
→ The addition of variable fathereduc$_i$ or mothereduc$_i$ have no marginal contribution to the model ※

6204640863

3

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

- In this model, we have 7 variables ; $k = 7$.
- $df(\text{Model}) = k-1 = 7-1 = 6$ ※

$df(\text{Risidual}) = n-k = 428-7 = 421$ ※

$df(\text{Total}) = 421+6 = 427$ ※

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

MSS of RSS

- $\frac{RSS}{421} = 0.446526442$; $RSS = 187.98763208$ ※

TSS

- $223.327441 = ESS + 187.98763208$; $ESS = 35.33980892$ ※
- MSS of ESS $= \frac{ESS}{df} = \frac{35.33980892}{6} = 5.8899681533$ ※
- MSS of TSS $= \frac{TSS}{df} = \frac{223.327441}{427} = 0.5230150843$ ※

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

$$\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-k} \quad ; \quad R^2 = \frac{ESS}{TSS} = 0.1582421254$$

$$= 1-\left(1-\frac{35.33980892}{223.327441}\right)\left(\frac{427}{421}\right) = 0.1462455761 \text{ ※}$$

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 $F_{1,421} = 3.84$)

- Let β_8 is the excluded variable
- The result of hypothesis testing of excluded variable will reject (significant) the Null hypothesis or $\beta_8 \neq 0$.

⇒ Therefore $F_{cal} > F_{upper,\alpha}(1,421)$

Model 3.2 : $k = 8$, $n-k = 428-8 = 420$

$$F_{cal} = \frac{(R^2_{new} - R^2_{old}) / \#\text{ of new regressors}}{1-R^2_{new}/(n-k_{new})}$$

$$3.84 < \frac{(R^2_{new} - 0.1582421254)/1}{1-R^2_{new}/(428-8)} \quad ; \quad F_{cri} < F_{cal}$$

$$3.84 < (R^2_{new} - 0.1582421254)\cdot\left(\frac{420}{1-R^2_{new}}\right)$$

; $0 \leq R^2 \leq 1$, but in this case $R^2 \neq 1$

so $\frac{420}{1-R^2_{new}}$ is positive value

$$1-R^2_{new} < \frac{420R^2_{new} - 66.46169267}{3.84}$$

$$1-R^2_{new} < 109R^2_{new} - 17.30773247$$

$\Rightarrow 18.30773247 < 110R^2_{new}$

$0.1573430224 < R^2_{new} < 1$; max $R^2_{new} < 1$ ※

6204640863

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, age can't determine how much hourly wage would be because when people are getting old they will work more slowly. Therefore, it is not right that when people who are aging will earn more wage as in a modle.

The equation of wage and age must be like this

$\Rightarrow \hat{wage}_i = \beta_0 + \beta_1 age_i - \beta_2 age_i^2$