Mahabhop Ananchavanich
6104640443

3. Using the data in RDCHEM, the following equation was obtained by OLS:

$$\widehat{rdintens} = 2.613 + .00030\ sales - .0000000070\ sales^2$$
$$(.429)\ \ (.00014)\qquad (.0000000037)$$
$$n = 32, R^2 = .1484.$$

i. At what point does the marginal effect of *sales* on *rdintens* become negative?

ii. Would you keep the quadratic term in the model? Explain.

iii. Define *salesbil* as sales measured in billions of dollars: $salesbil = sales/1{,}000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the *R*-squared. [*Hint*: Note that $salesbil^2 = sales^2/(1{,}000)^2$.]

iv. For the purpose of reporting the results, which equation do you prefer?

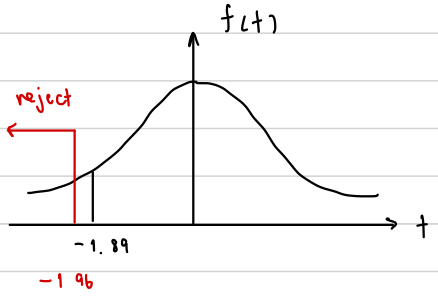
i) $\frac{\partial\, \widehat{rdintens}}{\partial\, sales} = 0.0003 - 2(0.0000000007)\, sales = 0$

$$sales = \frac{0.0003}{0.0000000014} = 21428.5714$$

marginal effect of sales on rdintens become negative when sales is larger than 21428.5714

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

$$t_{sales^2} = \frac{Coef.}{Std.err.} = \frac{-0.0000000007}{0.0000000037} = -1.89$$



$-1.98 > -1.89$, so we fail to reject H_0 at 5 % significance level.

iii) $\widehat{rdintens} = 2.613 + \frac{0.0003}{(0.00014 \times 1000)}(Salesbil \times 1000) - \frac{0.0000000007}{(0.0000000037 \times 1000^2)}(Salesbil^2 \times 1000^2)$

$$= \underset{(0.429)}{2.613} + \underset{(0.14)}{0.3}\, salesbil - \underset{(0.0037)}{0.007}\, salesbil^2$$

iv) I prefer equation in iii because it is easier to read, since it has less decimal, and easier to applied than equation in the question

1. Using the data in SLEEP75 (see also Problem 3 in Chapter 3), we obtain the estimated equation

$$\begin{aligned} \widehat{sleep} &= 3{,}840.83 - .163\, totwrk - 11.71\, educ - 8.70\, age \\ &\quad (235.11)\ (.018)\ (5.86)\ (11.21) \\ &\quad + .128\, age^2 + 87.75\, male \\ &\quad (.134)\ (34.33) \\ n &= 706, R^2 = .123, \overline{R}^2 = .117. \end{aligned}$$

The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

i. All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?

ii. Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

iii. What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

i) Yes, if the participant was male the equation will be

$\widehat{sleep} = 3840.83 - 0.163\, totwrk - 11.71\, educ - 8.7\, age + 0.128\, age^2 + 87.75\, male$

if the participant was female the equation will be

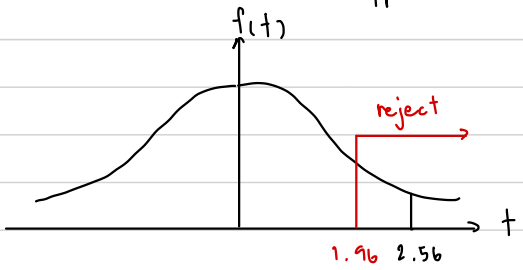
$\widehat{sleep} = 3840.83 - 0.163\, totwrk - 11.71\, educ - 8.7\, age + 0.128\, age^2$

from the equation we can see that if it is male participants, they will get 87.75 additional minutes per week spent sleeping at night

The evidence is strong since if we conduce hypothesis testing with 5% significance level,

$H_0 : \beta_{male} = 0$

$H_a : \beta_{male} \neq 0$



$t_{male} = \frac{coef}{std.\ err} = \frac{87.75}{34.33} = 2.56 > 1.96$, so we reject H_0 at 5% significance level.

ii) $\frac{\partial \widehat{sleep}}{\partial totwrk} = -0.163$, implies that if you increse 1 minute of totwrk, you will loose 0.163 minute of sleep

iii) we have to run restricted model which have totwrk, educ, male as an explanatory variable of sleep to conduct F-test

8. Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: “On how many separate occasions last month did you smoke marijuana?”

   i. Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, “Smoking marijuana five more times per month is estimated to change wage by *x*%.”

   ii. Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

   iii. Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.

   iv. Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

   v. What are some potential problems with drawing causal inference using the survey data that you collected?

i) $\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 female + u$

ii) $\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 female + \beta_5 female\_usage + u$

$H_0 : \beta_5 = 0$

H_a : otherwise

iii) $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 female + \delta_1 nuser + \delta_2 luser + \delta_3 muser + \delta_4 huser + u$

iv) $H_0 : \delta_2 = \delta_3 = \delta_4 = 0$

H_a : otherwise

F - test ~ $k = b$, $q = 3$

d.f : $n - k - 1 = n - 7$ $\Rightarrow$ $F_{3, n-7}$

v) sample is not spread enough to be a good representative of population

might not received good cooporation from public

sample selection cause nondiversified.

11. The following equations were estimated using the data in ECONMATH, with standard errors reported under coefficients. The average class score, measured as a percentage, is about 72.2; exactly 50% of the students are male; and the average of *colgpa* (grade point average at the start of the term) is about 2.81.

$$\widehat{score} = 32.31 + 14.32\, colgpa \qquad \text{R}$$
$$(2.00)\quad (0.70)$$
$$n = 856, R^2 = .329, \overline{R}^2 = .328.$$

$$\widehat{score} = 29.66 + 3.83\, male + 14.57\, colgpa \qquad \text{UR}$$
$$(2.04)\ (0.74)\qquad (0.69)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .348.$$

$$\widehat{score} = 30.36 + 2.47\, male + 14.33\, colgpa + 0.479\, male \cdot colgpa$$
$$(2.86)\ (3.96)\qquad (0.98)\qquad (1.383)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .347.$$

$$\widehat{score} = 30.36 + 3.82\, male + 14.33\, colgpa + 0.479\, male \cdot (colgpa - 2.81)$$
$$(2.86)\ (0.74)\qquad (0.98)\qquad (1.383)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .347.$$

i. Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero?

ii. In the second equation, how come the estimate on *male* is so imprecise? Should we now conclude that there are no gender differences in *score* after controlling for *colgpa*? [*Hint*: You might want to compute an *F* statistic for the null hypothesis that there is no gender difference in the model with the interaction.]

iii. Compared with the third equation, how come the coefficient on *male* in the last equation is so much closer to that in the second equation and just as precisely estimated?

i) from 2nd equation; $\frac{\partial \widehat{score}}{\partial male} = 3.83$,

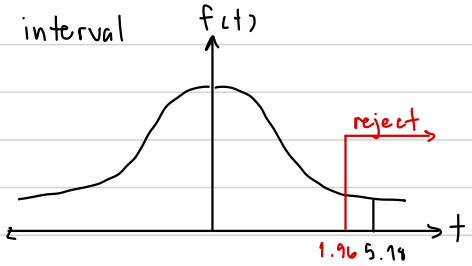
means that if it is male participants, the score will increase by 3.83 %.

Construct a hypothesis testing at 95% confidence interval

$H_0: \beta_{male} = 0$
$H_a: \beta_{male} \neq 0$

$t_{male} = \frac{Coef}{std.err} = \frac{3.83}{0.74} = 5.18$

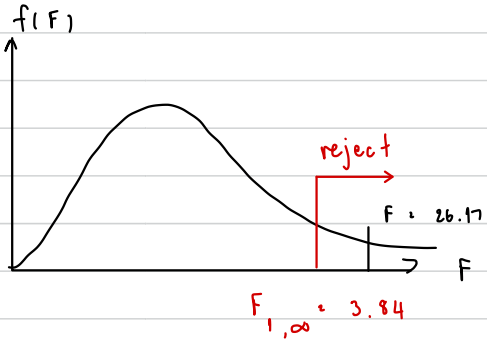
$t_{male} = 5.18 > 1.96$, so we reject H_0 at 95% confidence interval, confidence interval exclude zero.



ii) Conduct F-test

$H_0: \beta_{male} = 0$
$H_a:$ otherwise is true

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/n-k-1} = \frac{(0.348 - 0.328)}{(1 - 0.348)/853}$$
$$= \frac{0.02}{0.00076} = 26.17$$



since $F = 26.17 > 3.84$, we reject H_0 at 5% significance level, male have joint impact on score, we cannot conclude that there are no gender differences in score after controlling for colgpa

iii) as it stated in the first paragraph, average of colgpa is about 2.81, so it will make the last terms of last equation becomes zero, so, we can see that 4th equation is very similar to 2nd equation, that is why the all the betas is so close between two.

C4. Use the data in GPA2 for this exercise.

i. Consider the equation

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

where *colgpa* is cumulative college grade point average; *hsize* is size of high school graduating class, in hundreds; *hsperc* is academic percentile in graduating class; *sat* is combined SAT score; *female* is a binary gender variable; and *athlete* is a binary variable, which is one for student-athletes. What are your expectations for the coefficients in this equation? Which ones are you unsure about?

ii. Estimate the equation in part (i) and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?

iii. Drop *sat* from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).

iv. In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no ceteris paribus difference between women athletes and women nonathletes.

v. Does the effect of *sat* on *colgpa* differ by gender? Justify your answer.

i) $\beta_3 < 0$, the smaller, the better the student
$\beta_4 > 0$, the higher SAT score, the better the student
β_1, β_2, β_5 and β_6 is unclear

ii)

```
. regress colgpa hsize hsizesq hsperc sat female athlete
```

Source	SS	df	MS		
Model	524.819305	6	87.4698842	Number of obs	4,137
Residual	1269.37637	4,130	.307355053	F(6, 4130)	284.59
				Prob > F	0.0000
Total	1794.19567	4,136	.433799728	R-squared	0.2925
				Adj R-squared	0.2915
				Root MSE	.5544

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0568543	.0163513	-3.48	0.001	-.0889117	-.0247968
hsizesq	.0046754	.0022494	2.08	0.038	.0002654	.0090854
hsperc	-.0132126	.0005728	-23.07	0.000	-.0143355	-.0120896
sat	.0016464	.0000668	24.64	0.000	.0015154	.0017774
female	.1548814	.0180047	8.60	0.000	.1195826	.1901802
athlete	.1693064	.0423492	4.00	0.000	.0862791	.2523336
_cons	1.241365	.0794923	15.62	0.000	1.085517	1.397212

Colgpa = 1.241 − 0.0569 hsize + 0.00468 $hsize^2$ − 0.0132 hsperc + 0.00165 sat + 0.155 female + 0.169 athlete
(0.079) (0.1635) (0.00225) (0.0006) (0.00007) (0.0187) (0.042)

holding other constant, athlete will have higher Colgpa than nonathlete by 0.169 points
the $t_{athlete} = \frac{0.169}{0.42} = 4.02$ which is very significant

iii)

```
. regress colgpa hsize hsizesq hsperc female athlete
```

Source	SS	df	MS		
Model	338.217123	5	67.6434247	Number of obs	4,137
Residual	1455.97855	4,131	.35245184	F(5, 4131)	191.92
				Prob > F	0.0000
Total	1794.19567	4,136	.433799728	R-squared	0.1885
				Adj R-squared	0.1875
				Root MSE	.59368

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0534038	.0175092	-3.05	0.002	-.0877313	-.0190763
hsizesq	.0053228	.0024086	2.21	0.027	.0006007	.010045
hsperc	-.0171365	.0005892	-29.09	0.000	-.0182916	-.0159814
female	.0581231	.0188162	3.09	0.002	.0212333	.095013
athlete	.0054487	.0447871	0.12	0.903	-.0823582	.0932556
_cons	3.047698	.0329148	92.59	0.000	2.983167	3.112229

now, estimate effect of being athlete is 0.005.
it is because we do not control SAT score, nonathlete score higher on average than athlete but in ii) we control SAT score, athlete do better than nonathletes

we choose female nonathlete as base group
Colgpa of femath will be higher than that of female nonathlete by 0.175 points. for hypothesis is testing by using t statistic on femath, t = 2.08, which is statistically significance at 5% confidence interval

iv)

```
. regress colgpa hsize hsizesq hsperc sat femath maleath malenonath
```

Source	SS	df	MS		
Model	524.821272	7	74.9744674	Number of obs	4,137
Residual	1269.3744	4,129	.307429015	F(7, 4129)	243.88
				Prob > F	0.0000
Total	1794.19567	4,136	.433799728	R-squared	0.2925
				Adj R-squared	0.2913
				Root MSE	.55446

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0568006	.0163671	-3.47	0.001	-.0888889	-.0247124
hsizesq	.0046699	.0022507	2.07	0.038	.0002573	.0090825
hsperc	-.0132114	.000573	-23.06	0.000	-.0143349	-.012088
sat	.0016462	.0000669	24.62	0.000	.0015151	.0017773
femath	.1751106	.0840258	2.08	0.037	.0103748	.3398464
maleath	.0128034	.0487395	0.26	0.793	-.0827523	.1083591
malenonath	-.1546151	.0183122	-8.44	0.000	-.1905168	-.1187133
_cons	1.39619	.0755581	18.48	0.000	1.248055	1.544324

v)

```
. regress colgpa hsize hsizesq hsperc sat female athlete femsat
```

Source	SS	df	MS		
Model	524.867644	7	74.981092	Number of obs	4,137
Residual	1269.32803	4,129	.307417784	F(7, 4129)	243.91
				Prob > F	0.0000
Total	1794.19567	4,136	.433799728	R-squared	0.2925
				Adj R-squared	0.2913
				Root MSE	.55445

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0569121	.0163537	-3.48	0.001	-.0889741	-.0248501
hsizesq	.0046864	.0022498	2.08	0.037	.0002757	.0090972
hsperc	-.013225	.0005737	-23.05	0.000	-.0143497	-.0121003
sat	.0016255	.0000852	19.09	0.000	.0014585	.0017924
female	.1023066	.1338023	0.76	0.445	-.1600179	.3646311
athlete	.1677568	.0425334	3.94	0.000	.0843684	.2511452
femsat	.0000512	.0001291	0.40	0.692	-.000202	.0003044
_cons	1.263743	.0974952	12.96	0.000	1.0726	1.454887

if we add femsat = female × sat to the equation in part (ii) we will get coefficient about 0.000051 and when we see t-statisic which is about 0.40, there is very not significance that the effect of SAT differs by genders