Chapter 6

Homework 6

Vipharath Charoensukanun
6104640559

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

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

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.) $\dfrac{\partial \widehat{rdintens}}{\partial sales} = .00030 - .000000014 sales = 0$

$.000000014 sales = .00030$

$sales = \dfrac{.00030}{.000000014}$

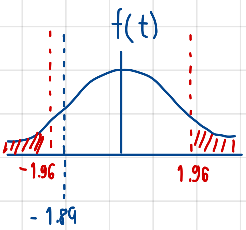
$sales = 21{,}428.571$

$\therefore$ when sales is larger than 21,428.571

ii.) $H_0: \beta_2 = 0$
$H_a:$ otherwise

$t = \dfrac{\hat{\beta}_2 - 0}{s.e._{\beta_2}} = \dfrac{-.0000000070}{.0000000037}$

$= -1.89 > -1.96$



$\therefore$ We failed to reject H_0 with 5% significant level

iii.) $salesbil = sales/1000$ $\qquad$ $salesbil^2 = sales^2/1000^2$

$sales = salesbil \cdot 1000$ $\qquad$ $sales^2 = salesbil^2/1000$

$\widehat{rdintens} = 2.613 + .0003\ sales - .000000007\ sales^2$

$= 2.613 + .0003(salesbil \cdot 1000) - .000000007(salesbil^2 \cdot 1000)^2$

$= \underset{(.429)}{2.613} + \underset{(.14)}{0.3 salesbil} - \underset{(.0037)}{0.007 salesbil^2}$

standard error also multiply by 1000 for salesbil and 1000^2 for salesbilsq

$s.e._{salesbil}$ $(.00014)(1000) = .14$

$s.e._{salesbilsq}$ $(.0000000037)(1000)^2 = .0037$

iv.) I prefer equation from (iii) since it easier to read and for calculation when it have fewer decimal points than the original one.

Chapter 7 paper

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

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

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, the coefficient of male is 87.75. Meaning if the participants are male sleep will be higher th n those who is not male by 87.75 minutes.
The evidence is very strong.

$t_{male} = \frac{87.75}{34.33} = 2.556$ which is close to 1% critical value

ii) $t_{totwrk} = \frac{-.163}{.018} = -9.056$

It have negatively impact on sleep.

$(60)(-.163) = -9.78$

$\therefore$ 1 more hour of working lessen the sleep by 9.78 minutes.

iii.) $H_0 : \beta_{age} = \beta_{age^2} = 0$

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

$H_0 : \beta_5 = 0$

iii) $\log(wage) = \beta_0 + \delta_1 nonuser + \delta_2 light + \delta_3 moderate + \delta_4 heavy + \beta_2 educ + \beta_3 exper + \beta_4 female + u$

iv.) $H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$ Test by using F-test $k=7$
D.f. $= n-k-1 = n-7-1 = n-8$; $F_{3,n-8}$

v.) law, self-selection, and social-decirability bias.

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.

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

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

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

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

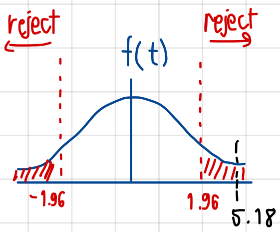
i. Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero?

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 second equation $\dfrac{\partial \widehat{score}}{\partial male} = 3.83$

It implies that male participants will have higher score by 3.83%.

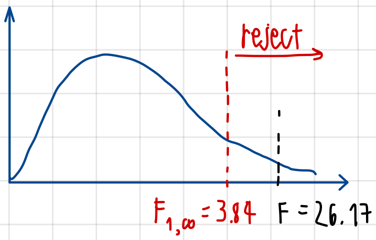


$H_0: \beta_{male} = 0$ at 95% confidence interval
H_a: otherwise

$$t_{male} = \frac{3.83}{0.74} = 5.18 > 1.96$$

∴ we reject H_0 at 95% confidence interval exclude zero.

ii) Conduct F-test
$H_0: \beta_{male} = 0$
H_a: otherwise



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

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 gender differences in score after controlling for colgpa.

iii) Since, average colgpa is 2.81, this will make colgpa − 2.81 equal to 0 and 0.479 male (0) = 0. The last term of 4th equation is deducted. It make 4th equation similar to the 2nd equation, so the value of the coefficient on male are similar.

chap 7 computer

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?

i). My expectation is $\beta_3 < 0$, the less percentile, the smarter. $\beta_4 > 0$, the more sat score, the better. and $\beta_6 < 0$ since athlete need to spare time for practice
The ones that I am unsure are β_1, β_2, and β_5
It is unclear that the size of school or gender will effecting GPA or not.

ii)

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

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

$$\widehat{colgpa} = 1.241 - .057 hsize + .005 hsizesq - .013 hsperc + .002 sat + .155 female + .169 athlete + u$$

An athlete is predicted to have higher GPA than nonathlete by .169. t = 4.00 and it is very significant

iii)

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

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

Dropped sat out of the regression, the coefficient of athlete becomes .005 with standard error of .045 which is not significant.
This implies that athlete have SAT score lower than nonathlete.

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.

iv.)

```
. generate male = female==0

. generate femaleath = female==1&athlete==1

. generate maleath = male==1&athlete==1

. generate malenonath = male==1&athlete==0
. regress colgpa hsize hsizesq hsperc sat femaleath maleath malenonath
```

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

$$\widehat{colgpa} = 1.396 - .057 hsize + .005 hsizesq - .013 hsperc + .002 sat + .175 femaleath + .013 maleath - .155 malenonath$$

The coefficient on femaleath or female athlete is .175 higher than female nonathlete.

$H_0: \beta_{femaleath} = 0$

v.)

```
gen femalesat = female*sat
. regress colgpa hsize hsizesq hsperc sat female athlete femalesat
```

Source	SS	df	MS			
				Number of obs	=	4,137
				F(7, 4129)	=	243.91
Model	524.867644	7	74.981092	Prob > F	=	0.0000
Residual	1269.32803	4,129	.307417784	R-squared	=	0.2925
				Adj R-squared	=	0.2913
Total	1794.19567	4,136	.433799728	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
femalesat	.0000512	.0001291	0.40	0.692	-.000202	.0003044
_cons	1.263743	.0974952	12.96	0.000	1.0726	1.454887

$$\widehat{colgpa} = 1.264 - .057 hsize + .005 hsizesq - .013 hsperc + .002 sat + .102 female + .168 athlete + .00005 femalesat$$

By adding femalesat to the equation with .00005 coefficient and 0.40 t-statistic. Gender rarely have effect on sat.