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

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

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) $\frac{d\,rdintens}{d\,sales} = 0.0003 - 0.000000014\,sales$

$\frac{d\,rdintens}{d\,sales} = 0$

$sales = 21428.571$

$= 21428.57(24p)$

ii) Yes, as the relationship between sales and rdintens is a curvilinear relationship. Keeping it in the quadratic term will enable use to see a clearer picture of their relationship.

iii) $\widehat{rdintens} = 2.613 + 0.0003\,sales - 0.00000000 7\,sales^2$

$\widehat{rdintens} = 2.613 + 0.0003(1{,}000\,salesbil) - 0.00000000 7((1000)^2\,salesbil^2)$

$\widehat{rdintens} = 2.613 + 0.3\,salebil - 0.007\,salesbil^2$

(0.429) (0.139) (0.00373)

iv) The equation with salebil & salebil^2 as variables is prefered.

. regress rdintens salebil salesbillsq

Source	SS	df	MS	Number of obs	=	32
				F(2, 29)	=	2.53
Model	16.1532557	2	8.07662785	Prob > F	=	0.0973
Residual	92.6802147	29	3.19586947	R-squared	=	0.1484
				Adj R-squared	=	0.0897
Total	108.83347	31	3.51075711	Root MSE	=	1.7877

rdintens	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
salebil	.3005713	.1392953	2.16	0.039	.0156805	.5854621
salesbillsq	-.0069459	.0037261	-1.86	0.072	-.0145667	.0006749
_cons	2.612512	.4294418	6.08	0.000	1.734205	3.490819

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?

iii) $H_0: \beta_4 = \beta_5 = 0$; H_a: otherwise

We need to run a regression with restricted model, β_4 & β_5.

. regress sleep totwrk educ age age2 male

Source	SS	df	MS	Number of obs	=	706
				F(5, 700)	=	19.59
Model	17092058.5	5	3418411.71	Prob > F	=	0.0000
Residual	122147777	700	174496.825	R-squared	=	0.1228
				Adj R-squared	=	0.1165
Total	139239836	705	197503.313	Root MSE	=	417.73

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
totwrk	-.1634232	.0181321	-9.01	0.000	-.199023	-.1278234
educ	-11.71332	5.866886	-2.00	0.046	-23.23212	-.1945157
age	-8.696676	11.20746	-0.78	0.438	-30.70094	13.30758
age2	.1284354	.1338954	0.96	0.338	-.1344494	.3913202
male	87.75243	34.32616	2.56	0.011	20.35786	155.147
_cons	3840.832	235.1087	16.34	0.000	3379.229	4302.435

i) Coefficient of male is 87.75, male sleeps 87.75 minutes per week more than female at 5% significant level

ii) Coefficient of totwrk is -0.163, one minute of work will result in 0.163 min of less sleep. It is significant at 1%.

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) $\log(wage) = \beta_0 + \beta_1 marij + \beta_2 educ + \beta_3 exp + \beta_4 fem + u$

ii) $\log(wage) = \beta_0 + \beta_1 marij + \beta_2 educ + \beta_3 exp + \beta_4 fem + \beta_5 marij \cdot fem + u$

$H_0: \beta_5 = 0$; $H_a: \beta_5 \neq 0$

if p value of $\beta_1 > 1.96$ or < -1.96 reject 1

iii) $\log(wage) = \beta_0 + \beta_1 light + \beta_2 moderate + \beta_3 heavy + \beta_4 edu + \beta_5 exp + \beta_6 fem + u$

iv) $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

use F-test with $q = 3$

degree of freedom: $n-6-1$

v) people may not give accurate report on marijuana usage due to the rule and regulation. Marijuana is also may be considerate as luxury good and not everyone have excess to it.

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.

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

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

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

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

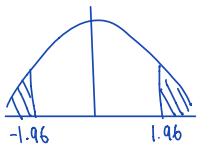
   i. Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero?

   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) the coefficient of male is 3.82, this means that if you are a male you will score 3.82 more than your female counterpart.

95% CI



The 95% CI for male
$= [3.82 - 1.96(0.74), 3.82 + 1.96(0.74)]$
$= [1.3696, 5.2704]$
the confidence interval exclude '0'

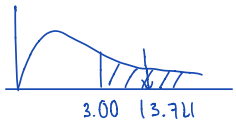
ii) This is because it has excluded many relevant variable thus making it ineffecient and imprecise

$H_0: \beta_{male} = \beta_{male.colgpa} = 0$

H_a: otherwise

$$F = \frac{SSR_{ur} - SSR_r / q}{1 - SSR_{ur} / n-k-1} = \frac{(0.549 - 0.329)/2}{1 - 0.349/856-4-1}$$

$= 13.0721$



Thus reject $H_0: \beta_{male} = \beta_{male.colgpa} = 0$, implies that there is a gender difference in score at 5% significance level.

iii)

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) β_1 is expected to be negative, colgpa would decrease as hsize increase

β_2 is expected to be negative, the higher the percentile, the lower the colgpa

β_4 is expected to be positive, the higher the SAT score, the higher the colgpa

β_6 is expected to be negative, student-athlete are expected to score lower than non-athlete

β_2 and β_5 are unsure of.

```
. regress colgpa hsize hsize2 hsperc sat female athlete
```

Source	SS	df	MS		
Model	524.819305	6	87.4698842	Number of obs	4,137
				F(6, 4130)	284.59
Residual	1269.37637	4,130	.307355053	Prob > F	0.0000
				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
hsize2	.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

ii) $\widehat{colgpa} = 1.24 - 0.569 hsize + 0.00468 hsize^2 - 0.0132 hsperc + 0.00165 sat + 0.155 female + 0.169 athlete$ the athletes are expected to score better than non athletes by 0.169 at 1% significant level

```
. regress colgpa hsize hsize2 hsperc female athlete
```

Source	SS	df	MS		
Model	338.217123	5	67.6434247	Number of obs	4,137
				F(5, 4131)	191.92
Residual	1455.97855	4,131	.35245184	Prob > F	0.0000
				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
hsize2	.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

iii) The estimated effect of being an athlete is 0.00545. The estimated is lower because we do not control SAT score anymore and the athlete score lower than nonathlete. However, we can see that it does not matter if we control the SAT score or not, athlete will still do better than nonathlete in the colgpa.

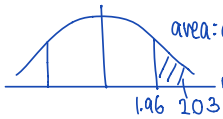
iv) femath = female * Athlete, the coefficient is 0.175, this mean that female athlete score 0.175 better than female non-athlete

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

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

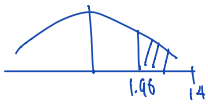
colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0568006	.0163671	-3.47	0.001	-.0888889	-.0247124
hsizesq	.0046699	.0022507	2.07	0.038	.0002573	.0090825
hsperc	-.0132114	.000573	-23.06	0.000	-.0143349	-.012088
sat	.0016462	.0000669	24.62	0.000	.0015151	.0017773
femath	.1751106	.0840258	2.08	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

$H_0: \beta_{femath} = 0$



thus reject $H_0: \beta_{femath} = 0$, there is a difference in colgpa between female athlete & nonathlete

v) $H_0: \beta_{femsat} = 0$



thus reject $H_0: \beta_{femsat} = 0$, there is effect of sat on colgpa on different gender.

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

Source	SS	df	MS		
Model	411.075682	5	82.2151365	Number of obs	4,137
				F(5, 4131)	245.55
Residual	1383.11999	4,131	.334814813	Prob > F	0.0000
				R-squared	0.2291
				Adj R-squared	0.2282
Total	1794.19567	4,136	.433799728	Root MSE	.57863

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0190793	.0051959	-3.67	0.000	-.0292661	-.0088925
hsperc	-.0156907	.0005655	-27.74	0.000	-.0167995	-.014582
femsat	.0015785	.0001058	14.92	0.000	.0013712	.0017859
female	-1.522999	.1076139	-14.15	0.000	-1.73398	-1.312018
athlete	.0226718	.0436648	0.52	0.604	-.0629347	.1082783
_cons	2.977225	.0226209	131.61	0.000	2.932876	3.021574